electronics

MDPI

Article

Deep Feature-Level Sensor Fusion Using Skip Connections for Real-Time Object Detection in Autonomous Driving

Vijay John * and Seiichi Mita

Research Center for Smart Vehicles, Toyota Technological Institute, Nagoya 468-8511, Japan; smita@toyota-ti.ac.jp
* Correspondence: vijayjohn@toyota-ti.ac.jp

Abstract: Object detection is an important perception task in autonomous driving and advanced driver assistance systems. The visible camera is widely used for perception, but its performance is limited by illumination and environmental variations. For robust vision-based perception, we propose a deep learning framework for effective sensor fusion of the visible camera with complementary sensors. A feature-level sensor fusion technique, using skip connection, is proposed for the sensor fusion of the visible camera with the millimeter-wave radar and the thermal camera. The two networks are called the RV-Net and the TV-Net, respectively. These networks have two input branches and one output branch. The input branches contain separate branches for the individual sensor feature extraction, which are then fused in the output perception branch using skip connections. The RVNet and the TVNet simultaneously perform sensor-specific feature extraction, feature-level fusion and object detection within an end-to-end framework. The proposed networks are validated with baseline algorithms on public datasets. The results obtained show that the feature-level sensor fusion is better than baseline early and late fusion frameworks.

Keywords: deep sensor fusion; intelligent vehicles

Citation: John, V.; Mita, S. Deep Feature-Level Sensor Fusion Using Skip Connections for Real-Time Object Detection in Autonomous Driving. *Electronics* **2021**, *10*, 424. https://doi.org/10.3390/electronics10040424

Academic Editors: Dong Seog Han, Kalyana C. Veluvolu and Takeo Fujii
Received: 13 January 2021
Accepted: 8 February 2021
Published: 9 February 2021

1. Introduction and Literature Review

Environment perception tasks, such as object detection, is an important research topic in autonomous driving and advanced driving assistant system [1,2]. The results obtained from perception are used by the autonomous driving systems for navigation and control. A similar application of the perception task is found in wireless sensor networks [3–5], where the perception obtained from the sensor nodes are sent to actuator nodes for intelligent traffic monitoring.

Research Problem: For object detection, typically, an array of different sensors such as visible camera, stereo camera, thermal camera and radar are used. Among these different sensors, the visible camera is widely used. However, the visible camera-based object detection is affected by environmental and illumination variations.

In this paper, we address the limitation associated with the visible camera-based object detection by proposing novel sensor fusion frameworks using deep learning and complementary sensors. The proposed deep learning frameworks simultaneously perform sensor-specific feature extraction, feature-level sensor fusion and object detection in an end-to-end manner. Two deep learning frameworks called the RVNet for radar-visible camera-based object detection and the TVNet for thermal-visible camera-based object detection are presented in this paper.

The radar and thermal camera are complementary sensors used by the RVNet and TVNet for object detection. A comparison between the visible camera and the complementary sensors are illustrated in Tables 1 and 2 and Figures 1 and 2. As illustrated, the sensor pairs have complementary characteristics, and their fusion will enhance the perception robustness in varying conditions. We next review the literature before highlighting the contribution of our work.

Figure 1. Scenes from the Nuscenes dataset, where radar reflection are obtained in varying conditions. The camera obtains the object boundaries in the same conditions.

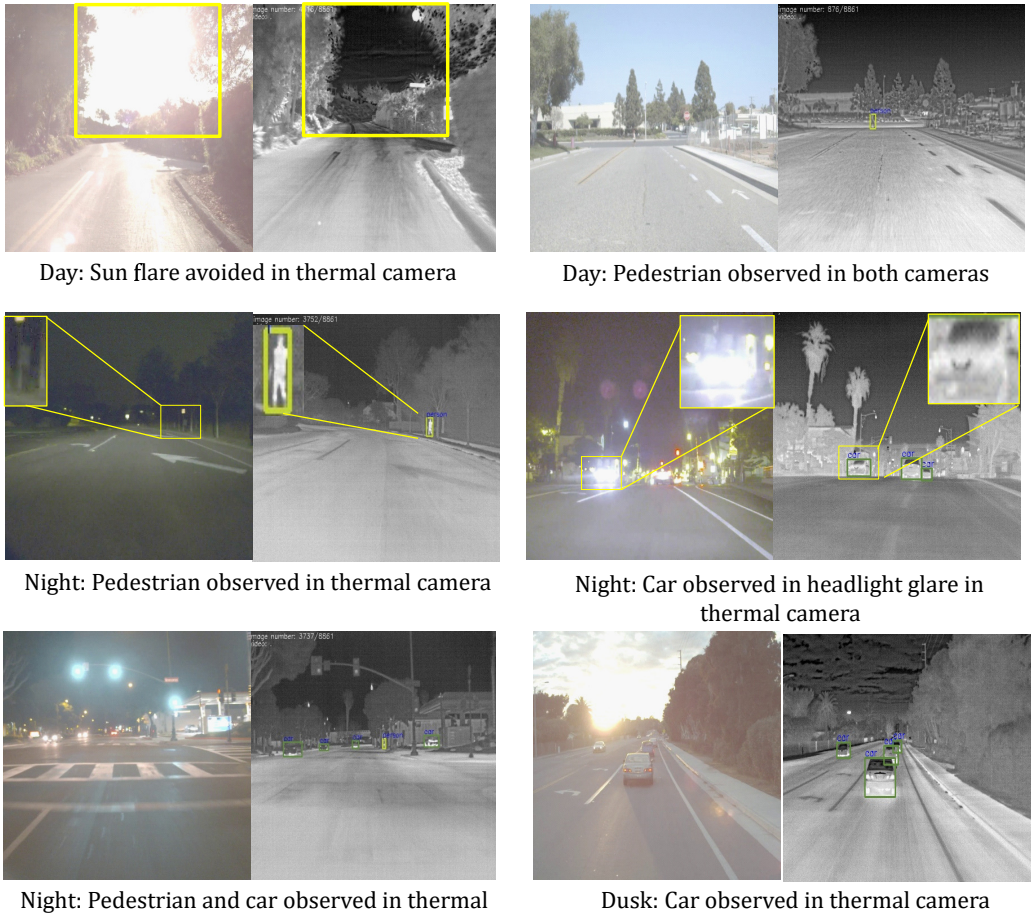

Figure 2. Scenes from the FLIR dataset, where thermal camera images are compared with visible camera images.

Literature Review: In case of the sensor fusion of the radar and visible camera, the sensor fusion techniques are classified as early fusion-based obstacle detection [6–8], late fusion-based obstacle detection [9–11] and feature fusion-based obstacle detection [12,13]. In early fusion, first, a set of candidate obstacles detected using the radar are identified in the visible camera. Subsequently, camera-based algorithms [14,15] are used to prune the candidates and identify the final set of obstacles. Bombini et al. [6] and

Sugimoto et al. [7] adopt such a strategy, where the radar object candidates are transferred to the camera coordinate for vision-based object detection. In late fusion-based obstacle detection, radar-based obstacle detection and camera-based obstacle detection are performed independently. The individual detection results are heuristically combined and the final set of obstacles are detected [9,10]. In feature fusion-based object detection, the radar and vision features are fused in a deep learning-based framework [12,13,16], where simultaneous sensor fusion and obstacle detection is performed. Chadwick et al. [12] propose a deep learning framework based on the SSD to detect obstacles in the visible image using both radar and visible camera features.

Compared to the visible-radar literature, the RVNet framework improves the state-of-the-art detection accuracy with real-time computational complexity. Additionally, the RVNet proposes a novel radar feature map called the sparse radar image for object detection.

Unlike radar and vision-based obstacle detection, the focus of the thermal and visible camera-based obstacle detection is different. Here, researchers focus on generating an enhanced multimodal visible-thermal image by fusion, instead of enhancing the object detection algorithm. Researchers report improved detection accuracy by enhancing the input to the object detection framework, without any change to the detection framework. The generated multimodal image contains the visible camera's rich appearance information as well as the thermal camera's heat signature information [17–22]. The multi-modal images are obtained using multi-resolution schemes [17–20], local neighborhood statistics [23] and learning methods [24–26]. Shah et al. [17] generate the multi-modal image using wavelet analysis and contourlet transform. The pixel value in the fused image is obtained by the weighted average of the thermal and visible pixels. The weights are calculated using the significance of the source pixels. For example, Shah et al. [23] use the local neighborhood eigenvalue statistics to calculate the significance of the source pixels. Liu et al. [18] propose a multi-resolution fusion scheme to generate the multimodal image. The visual image is enhanced using contextual information derived from the thermal images. Zhang et al. [21] generate the image using quadtree decomposition and Bezier interpolation to obtain an enhanced thermal image with distinct thermal-based features. This enhanced image is added to the visible image to generate the multimodal image. John et al. [22] generate a multi-modal image using pixel-level concatenation, and perform pedestrian detection.

In recent years, with the advent of deep learning framework, researchers have also used deep learning for this task. Shopovska et al. [24] generate the multimodal image using deep learning, where the fused image is visually similar to a color image, but contains important features in the pedestrian regions. In the work by Ma et al. [25], the FusionGAN, a GAN framework, is used to generate the multimodal image. An extension of the FusionGAN is proposed by Zhao et al. [26]. The generated multimodal images are given to object detection algorithms for robust perception [22].

Compared to the existing thermal-visible camera literature, the TVNet framework does not focus on generating the multimodal image. Instead the TVNet is a sensor fusion framework that simultaneously performs sensor-specific feature extraction, feature-level sensor fusion and also object detection within an end-to-end framework. The proposed formulation is shown to improve the detection accuracy with improved robustness.

Proposed Framework: Both the RVNet and the TVNet have similar deep learning architectures. Both the networks contain two input branches and one output branch. The input branches contain separate branches for the individual sensor feature extraction. The input branches extract the sensor-specific features. These features are transferred to the output branch using the skip connection, where the feature-level fusion and object detection is performed. The output branch performs single shot object detection and is based on the YOLO network [27]. An overview of the architecture is shown in Figure 3.

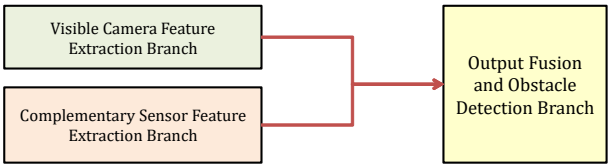

Figure 3. An overview of the skip connection-based fusion and obstacle detection architecture.

The networks are validated on the Nuscenes [28] and FLIR public dataset [29] and compared with state-of-the-art baseline algorithms. The experimental results show that the proposed feature-level fusion is better than baseline sensor fusion strategies.

Key Contributions: To the best of our knowledge, the main contributions of the proposed framework are as follows:

- The TVNet is a novel end-to-end deep learning framework which simultaneously extracts sensor specific features, performs feature-level fusion and object detection for the thermal and visible cameras.
- Comparative experimental analysis of early, late and feature-level fusion of visible and thermal camera for obstacle detection.

Table 1. Comparative study of the properties of the radar and visible camera.

Sensor	Milliwave Radar	Visible Camera
Weather	Not affected by rain, snow and fog [30]	Affected by rain, snow and fog
Illumination	Not affected	Affected
Data density	Sparse appearance	Dense appearance
Object boundary	No	Yes
Object Speed	Yes	No

Table 2. Comparative study of the properties of the thermal and visible camera.

Sensor	Thermal Camera	Visible Camera
Weather	Not affected by rain, snow and fog [30]	Affected by rain, snow and fog
Illumination	Not affected by low-light and low-visibility	Affected by low-light and low-visibility
Lens flare	Not affected by direct sunlight and headlight	Affected by direct sunlight and headlight

The remainder of the paper is structured as follows. The skip connection networks are presented in Section 2. The experimental analysis is presented in Section 3. Finally, the paper is concluded in Section 4.

2. Deep Learning Framework

Our proposed deep learning frameworks simultaneously performs sensor specific feature extraction, feature-level fusion and object detection in an end-to-end learning framework.

The vision-based fusion with radar is performed by the RV-Net, while the fusion with the thermal camera is performed by the TVNet. Both the RV-Net and the TV-Net contains two input branches, and one output branch. The independent input branches extract sensor specific features relevant for object detection. The features extracted are transferred to the output branch for feature fusion and obstacle detection using the skip connections [31]. The output branch is a single shot object detection head, based on the tiny YOLO network [27] detecting all the objects in the driving environment within a binary classification framework.

2.1. RVNet: Radar and Visible Camera Fusion and Object Detection

The RVNet is a deep learning framework with two input modules and one output module as shown in Figure 4. The two independent input branches extract visible camera I features and millimeter wave radar S features, respectively. The radar specific features are extracted from the radar input. The radar input to the RVNet, is not the raw radar data, but rather a reformulated sparse radar image.

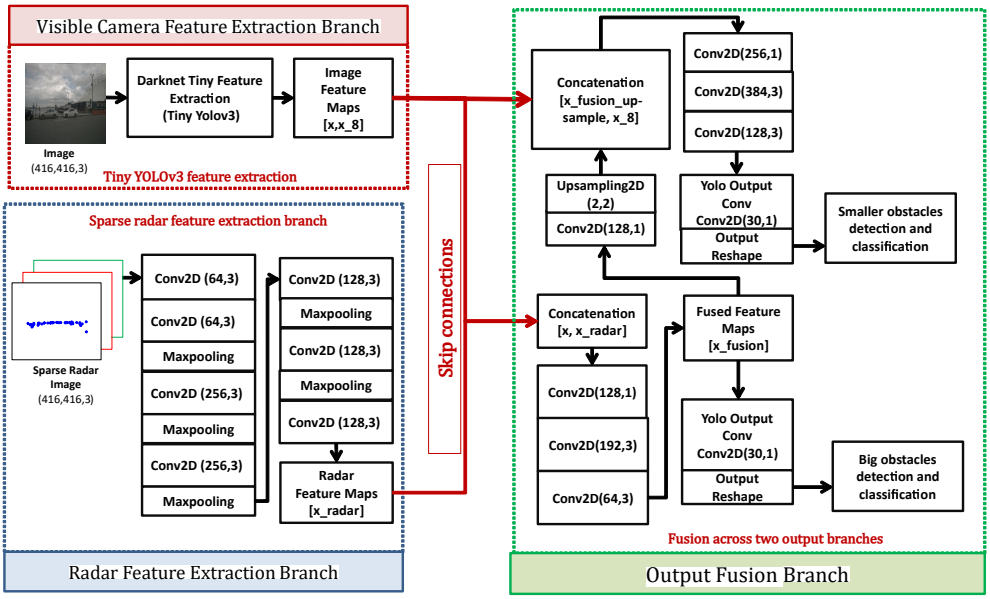

Figure 4. An overview of the RVNet architecture.

Sparse Radar Image: The sparse radar image is a highly sparse image with pixel-level radar information, corresponding to depth, lateral velocity and longitudinal velocity. To generate the sparse radar image, the raw radar points are first transformed from the radar coordinate system to the camera's coordinate system using the extrinsic calibration parameters. Subsequently, using the camera's intrinsic calibration parameters, the transformed raw radar points are formulated as the sparse radar image S. The sparse radar image's size is same as the image (416×416), and contains 3-channels corresponding to the radar features.

Input Branches: The visible camera feature extraction branch is based on the pre-trained Darknet model used in the tiny Yolo3 model [27]. The Darknet model is pre-trained on the Pascal VOC dataset [32].

The radar feature extraction branch is formulated to extract features from the highly sparse radar image S. Multiple 2D convolution filters with a single stride are used to account for every pixel-level radar feature. To reduce the dimensionality of the feature maps, maxpooling 2D is used.

Output Branch: The features extracted from the sparse radar image and the visible camera are transferred to the output branch using skip connections. The individual feature maps are concatenated with each other at various levels in the output branch. The output branch based on the tiny Yolov3 model contains two sub-branches. The first sub-branch detects small and medium-sized obstacles, and the second sub-branch detects big obstacles. The output branch uses multiple 2D convolution filters and reshapes layers.

2.2. TVNet: Thermal and Visible Camera Fusion and Object Detection

Similar to the RVNet, the TVNet also contains two input modules and one output module as shown in Figure 5. The two independent input branches extract visible camera image, I, features and thermal camera image, T, features, respectively.

As a pre-processing step, the thermal image is registered with the visible camera coordinate system. The thermal camera image is single channeled (416×416) image, which is the same size as the visible image.

Input Branches: Similar to the RVNet, the visible camera feature extraction branch in the TVNet is also based on the Darknet model used in the tiny Yolo3 model [27] to use the pre-trained weights of the model previously trained on the Pascal VOC dataset [32].

In case of the thermal camera feature extraction branch, the layer architecture is similar to the Darknet model used in the tiny Yolo3 model. However, owing to the input thermal camera image being single channeled, the pre-trained weights cannot be used for the thermal camera feature extraction branch.

Output Branch: The features extracted from the thermal and visible images are transferred to the output branch using skip connections. The architecture of the output branch is similar to the RVNet's output branch, including the concatenation of the individual feature maps. Similar to the RVNet, the output layer of the TVNet is also trained with the YOLO loss.

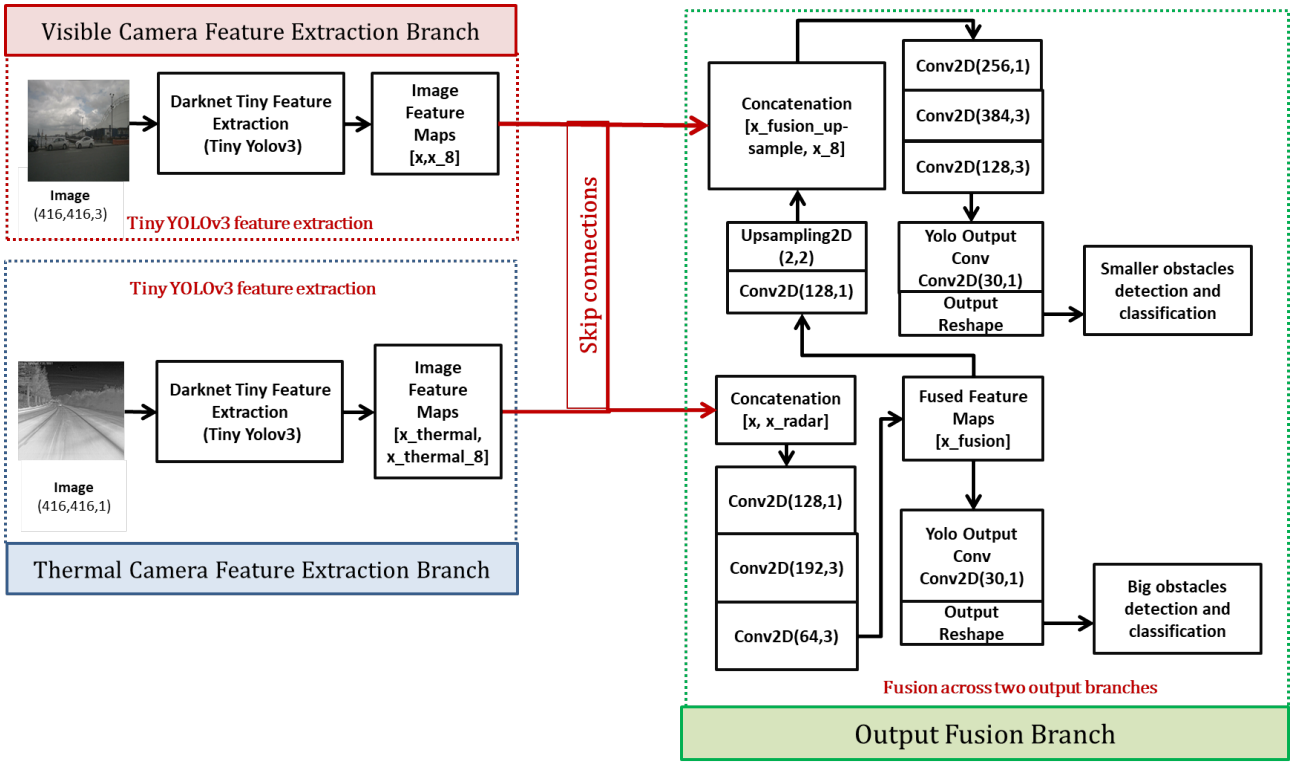

Figure 5. An overview of the TVNet architecture.

2.3. Training

Both the RVNet and the TVNet are trained as binary obstacle detectors on public datasets. The RVNet is trained with the nuscenes dataset, and the vehicles, motorcycles, cycles, pedestrians, movable objects and debris classes are considered to be obstacles [28]. On the other hand, the TVNet is trained with the FLIR dataset, where car, pedestrians, dogs and bicycles are considered to be obstacles [29].

In both the RVNet and TVNet, the pretrained weights of the feature extraction layers of the tiny YOLOv3 network trained on the Pascal VOC dataset [32] are used to initialize the weights of the image feature extraction branch.

On the other hand, the radar and thermal feature extraction and the output branches are initialized with random weights and trained without any fine-tuning. The networks are trained with an Adam optimizer and learning rate of 0.001.

3. Experimental Results and Discussion

Dataset: The RVNet was validated on the nuscenes dataset and the TVNet was validated on the FLIR dataset. The nuscenes dataset contains 3200 training samples and 1000 testing samples. The dataset contains scenes from daytime, nighttime and rainy weather (Figure 1).

The FLIR dataset contains a day subset and a night subset. The day subset contains 3714 training and 1438 test samples. The night subset contains 3670 training and 1482 testing samples. The dataset contains scenes from daytime and nighttime (Figure 2).

RVNet Baseline Algorithms: Two algorithms based on the tiny YOLOv3 network [27] are used as the baseline algorithms. The first baseline algorithm is the original Tiny Yolov3 network. which is trained on the visible camera image alone. However, unlike the original network, the baseline tiny Yolov3 networks is trained as a binary classifier.

The second baseline algorithm is a late fusion-based object detection network, where the results obtained from independent radar and vision object detection pipelines are combined in the final stage. The visible camera pipeline is a tiny Yolov3 network modified as a binary classifier. The radar pipeline is a rule-based classifier, where radar's velocity is used to identify the centroids of obstacles. In the absence of obstacle bounding box information, a pre-defined bounding box is generated around the detected centroids. The difference between the baseline and the RVNet is summarized in Figure 6.

Algo	Sensor Fusion	Feature Extraction	Feature Fusion	Object Detection	Notes
TinyYolov3 (Visible)	No	Visible camera feature	No	Visible camera features used	Visible camera-based object detection
Late Fusion	Yes	Visible camera and radar specific features	No	Visible camera and radar features used *independently*	Detection results from two independent TinyYolov3 combined in final step
RVNet	**Yes**	**Visible camera and radar specific features**	**Yes**	**Visible camera and radar features used *jointly***	**Simultaneous feature extraction, feature fusion and joint object detection**

Figure 6. RVNet compared with the baseline algorithms. Contributions to literature highlighted.

TVNet Baseline Algorithms: Multiple algorithms based on the tiny YOLOv3 network [27] are used as the baseline algorithms.

The first and second baseline algorithms are the original Tiny Yolov3 network trained on the visible camera image and thermal camera image, respectively.

The third baseline algorithm is an early fusion-based object detection framework, where a multimodal thermal-visible camera image is given as input to the tiny YOLOv3 architecture. The multimodal image is generated by the concatenation of registered thermal and visible camera images.

The final baseline algorithm is a late fusion-based object detection framework, where the results obtained from independent thermal and vision object detection pipelines are combined in the final stage. The independent pipelines are the tiny YOLOv3 networks, and the results obtained from the pipelines are combined. A non-maximum suppression is performed on the combined results to obtain the final result. The difference between the baseline and the TVNet is summarized in Figure 7.

Algorithm Parameters: The performance of the networks are reported using the Average Precision (AP) with IOU (intersection over threshold) of 0.5.

Algo	Sensor Fusion	Feature Extraction	Feature Fusion	Object Detection	Notes
TinyYolov3 (Visible)	No	Visible camera feature	No	Visible camera features used	Visible camera-based object detection
TinyYolov3 (Thermal)	No	Thermal camera feature	No	Thermal camera features used	Thermal camera-based object detection
Early Fusion	Yes	Multimodal features	Yes	Multimodal features used	Concatenated multimodal visible-thermal image used for object detection
Late Fusion	Yes	Visible and thermal camera specific features	No	Visible and thermal camera features used *independently*	Detection results from two independent TinyYolov3 combined in final step
TVNet	**Yes**	**Visible and thermal camera specific features**	**Yes**	**Visible and thermal camera features used *jointly***	**Simultaneous feature extraction, feature fusion and joint object detection**

Figure 7. TVNet compared with the baseline algorithms. Contributions to literature highlighted.

3.1. Results

3.1.1. RVNet

The performance of the RVNet and the baseline algorithms are tabulated in Table 3. The results show that the RVNet is better than the baseline algorithms. Results obtained are shown in Figure 8.

Table 3. Comparative Analysis of the RVNet.

Algorithms	Average Precision	Computing Time (ms)
Tiny Yolov3 [27]	0.40	10
Late Fusion	0.40	14
RVNet	0.56	17

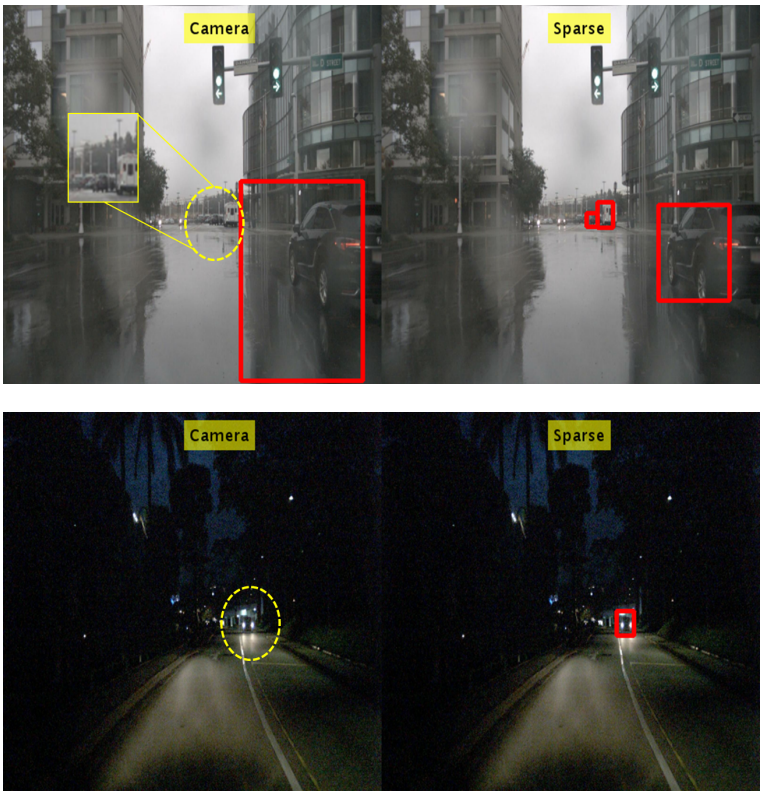

Figure 8. Detection result for the RVNet. The red bounding boxes and yellow circle indicate the true positives and false negatives, respectively.

3.1.2. TVNet

The performance of the TVNet and the baseline algorithms are tabulated in Tables 4 and 5. The TVNet performs the best among the different algorithms for the day time subset, while reporting optimal performance for the night time subset. Results are shown in Figure 9.

Table 4. Comparative Analysis of the TVNet on the daytime subset.

Algorithms	Average Precision	Computing Time (ms)
Tiny Yolov3 (Visible) [27]	0.59	10
Tiny Yolov3 (Thermal) [27]	0.58	10
Early Fusion	0.55	10
Late Fusion	0.56	17
TVNet	0.61	17

Table 5. Comparative Analysis of the TVNet on the nighttime subset.

Algorithms	Average Precision	Computing Time (ms)
Tiny Yolov3 (Visible) [27]	0.21	10
Tiny Yolov3 (Thermal)	0.61	10
Early Fusion	0.59	10
Late Fusion	0.60	20
TVNet	0.60	17

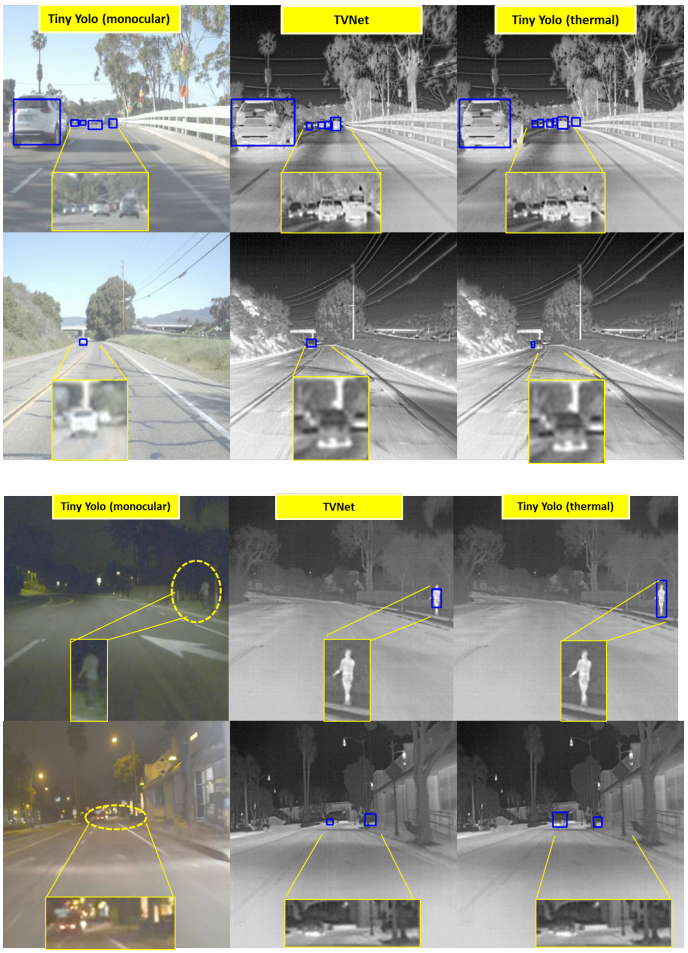

Figure 9. Detection result for the TVNet. The blue bounding boxes and yellow circle indicate the true positives and false negatives, respectively.

3.1.3. Discussion

Training: The RVNet and TVNet were trained using Keras on an Ubuntu 18.04 computer with a Nvidia Geforce 1080 GPU (NVIDIA, Santa Clara, CA, USA). The deep learning network hyperparameters and learning parameters were empirically selected considering the available computational resources, the real-time processing requirement and training accuracy. The RVNet and the TVNet were trained with batch-size 8 and 30 and 120 epochs, respectively.

During the network training process, the network weights associated with the different epochs were periodically saved. Subsequently, we computed the testing accuracy for the different network weights, and obtained the best results with 20 epochs for the RVNet and 100 epochs for the TVNet. This corresponds to an early stopping-based regularization reported in the literature [33].

RVNet: The good performance of the RVNet can be attributed to the radar's ability to identify obstacles in varying environment, and the visible camera's ability to delineate obstacles in varying environment. The advantage of the feature level sensor fusion of radar with the visible camera is clearly observed in Table 3. The RVNet reports better accuracy than the visible-camera alone Tiny YOLO (baseline-1) and the late radar-camera fusion method (baseline-2).

Sensor Fusion: The visible camera-based baseline is affected by the inherent limitations of the visible camera to low illumination and adverse weather conditions.

Feature-level Fusion: In the late fusion strategy, object detection by the two sensors are performed independently, without any feature-level fusion. Comparatively, the RVNet performs deep learning-based sensor specific feature extraction, feature-level sensor fusion and object detection within an end-to-end network. Instead of an independent object detection, which are combined in the last step (baseline-2), the RVNet uses the sensor specific features *jointly* for the object detection.

TVNet: Tables 4 and 5 compare the performance of the TVNet with the baseline algorithms on daytime and nighttime subsets. In the daytime subset, the TVNet reports the best performance among the different algorithms.

In the nighttime subset, the TVNet and the thermal camera networks report similar performance. The good performance of the thermal camera-based networks is primarily attributed to the performance of the thermal camera, which is invariant to the variations in the environment and illumination conditions.

Sensor Fusion: The advantages of feature level fusion are evident in Table 4, the TVNet reports better performance than the Tiny YOLO-based visible-alone and thermal-alone models.

Feature-level Fusion: Comparing the sensor fusion strategies, the feature level fusion strategy adopted by the TVNet is better than both the early fusion and late fusion learning frameworks (Tables 4 and 5). In the early fusion learning framework, a multimodal image is generated by thermal-visible image concatenation in the pre-processing step. No thermal and visible camera-specific feature extraction or fusion is performed in this framework.

In the late fusion learning framework, the thermal and visible camera detection pipelines are independent extracting sensor specific features. However, the features are not fused, and only the detection results are naively combined.

Comparatively, the TVNet is an end-to-end network with simultaneous camera-specific feature extraction, feature-level fusion and object detection. Consequently, the TVNet reports a better detection accuracy than the baseline algorithms.

4. Conclusions

Limitations associated with vision-based object detection are addressed using two sensor fusion networks called the RVNet and the TVNet. The RVNet and TVNet are proposed for visible camera-RADAR and visible-thermal camera-based object detection, respectively. Compared to the literature, both the proposed networks simultaneously perform sensor specific feature extraction, feature-level fusion and object detection. Both

the networks contain two independent feature extraction branches for sensor specific feature extraction. The output branch is a single shot detection framework which performs feature-level fusion as well as object detection. The proposed frameworks are validated on public datasets, and the results show that the RVNet and TVNet are better than baseline algorithms. The RVNet reports better accuracy than non-fusion frameworks and late fusion frameworks. Similarly, in the case of the TVNet, the TVNet reports better performance than non-fusion, early fusion and late fusion frameworks. Currently, the TVNet performance is slightly affected by the visible camera as observed in the nighttime sequences. In our future work, we will investigate methods to reduce the impact of the visible camera within the fusion framework for such challenging scenarios.

Author Contributions: Conceptualization, V.J., S.M.; methodology, V.J., S.M.; software, V.J.; validation, V.J.; formal analysis, V.J., S.M.; investigation, V.J., S.M.; data, V.J.; writing, V.J., S.M.; supervision, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. John, V.; Karunakaran, N.M.; Guo, C.; Kidono, K.; Mita, S. Free Space, Visible and Missing Lane Marker Estimation using the PsiNet and Extra Trees Regression. In Proceedings of the 24th International Conference on Pattern Recognition, Beijing, China, 20–24 August 2018; pp. 189–194.
2. Jazayeri, A.; Cai, H.; Zheng, J.Y.; Tuceryan, M. Vehicle Detection and Tracking in Car Video Based on Motion Model. *IEEE Trans. Intell. Transp. Syst.* **2011**, *12*, 583–595. [CrossRef]
3. Kafia, M.A.; Challal, Y.; Djenouria, D.; Abdelmadjid, M.D.; Badacheab, B.N. A Study of Wireless Sensor Networks for Urban Traffic Monitoring: Applications and Architectures. *Procedia Comput. Sci.* **2013**, *19*, 617–626. [CrossRef]
4. Nellore, K.; Hancke, G. A Survey on Urban Traffic Management System Using Wireless Sensor Networks. *Sensors* **2016**, *16*, 157. [CrossRef] [PubMed]
5. Curiac, D.-I.; Volosencu, C. Urban Traffic Control System Architecture Based on Wireless Sensor-Actuator Networks. In Proceedings of the 2nd International Conference on Manufacturing Engineering, Quality and Production Systems, (MEQAPS'10), Constanța, Romania, 3–5 September 2010.
6. Bombini, L.; Cerri, P.; Medici, P.; Alessandretti, G. Radar-Vision Fusion for Vehicle Detection. 2006. Available online: http://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.218.7866 (accessed on 5 January 2021).
7. Sugimoto, S.; Tateda, H.; Takahashi, H.; Okutomi, M. Obstacle detection using millimeter-wave radar and its visualization on image sequence. In Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, 26 August 2004; Volume 3, pp. 342–345.
8. Fang, Y.; Masaki, I.; Horn, B. Depth-based target segmentation for intelligent vehicles: Fusion of radar and binocular stereo. *IEEE Trans. Intell. Transp. Syst.* **2002**, *3*, 196–202. [CrossRef]
9. Garcia, F.; Cerri, P.; Broggi, A.; de la Escalera, A.; Armingol, J.M. Data fusion for overtaking vehicle detection based on radar and optical flow. In Proceedings of the IEEE Intelligent Vehicles Symposium, Alcala de Henares, Spain, 3–7 June 2012; pp. 494–499.
10. Zhong, Z.; Liu, S.; Mathew, M.; Dubey, A. Camera Radar Fusion for Increased Reliability in ADAS Applications. *Electron. Imaging Auton. Veh. Mach.* **2018**, *1*, 258-1–258-4. [CrossRef]
11. Wang, X.; Xu, L.; Sun, H.; Xin, J.; Zheng, N. On-Road Vehicle Detection and Tracking Using MMW Radar and Monovision Fusion. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2075–2084. [CrossRef]
12. Chadwick, S.; Maddern, W.; Newman, P. Distant Vehicle Detection Using Radar and Vision. *arXiv* **2019**, arXiv:1901.10951.
13. John, V.; Mita, S. RVNet: Deep Sensor Fusion of Monocular Camera and Radar for Image-based Obstacle Detection in Challenging Environments. In *Pacific-Rim Symposium on Image and Video Technology, Proceedings of the PSIVT, Sydney, Australia, 18–22 November 2019*; Springer: Cham, Switzerland, 2019.
14. Gaisser, F.; Jonker, P.P. Road user detection with convolutional neural networks: An application to the autonomous shuttle WEpod. In Proceedings of the International Conference on Machine Vision Applications (MVA), Nagoya, Japan, 8–12 May 2017; pp. 101–104.
15. Milch, S.; Behrens, M. Pedestrian Detection with Radar and Computer Vision. 2001. Available online: http://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.20.9264 (accessed on 5 January 2021).
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. *arXiv* **2015**, arXiv:1512.02325.
17. Shah, P.; Merchant, S.N.; Desai, U.B. Multifocus and multispectral image fusion based on pixel significance using multiresolution decomposition. *Signal Image Video Process.* **2013**, *7*, 95–109. [CrossRef]

18. Liu, Z.; Laganière, R. Context enhancement through infrared vision: A modified fusion scheme. *Signal Image Video Process.* **2007**, *1*, 293–301. [CrossRef]
19. Flitti, F.; Collet, C.; Slezak, E. Image fusion based on pyramidal multiband multiresolution markovian analysis. *Signal Image Video Process.* **2008**, *3*, 275–289. [CrossRef]
20. Shah, P.; Reddy, B.C.S.; Merchant, S.N.; Desai, U.B. Context enhancement to reveal a camouflaged target and to assist target localization by fusion of multispectral surveillance videos. *Signal Image Video Process.* **2013**, *7*, 537–552. [CrossRef]
21. Zhang, Y.; Zhang, L.; Bai, X.; Zhang, L. Infrared and visual image fusion through infrared feature extraction and visual information preservation. *Infrared Phys. Technol.* **2017**, *83*, 227–237. [CrossRef]
22. John, V.; Mita, S.; Liu, Z.; Qi, B. Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neural networks. In Proceedings of the 14th IAPR International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 18–22 May 2015.
23. Shah, P.; Srikanth, T.V.; Merchant, S.N.; Desai, U.B. Multimodal image/video fusion rule using generalized pixel significance based on statistical properties of the neighborhood. *Signal Image Video Process.* **2014**, *8*, 723–738. [CrossRef]
24. Shopovska, I.; Jovanov, L.; Philips, W. Deep Visible and Thermal Image Fusion for Enhanced Pedestrian Visibility. *Sensors* **2019**, *19*, 3727. [CrossRef] [PubMed]
25. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [CrossRef]
26. Zhao, Y.; Fu, G.; Wang, H.; Zhang, S. The Fusion of Unmatched Infrared and Visible Images Based on Generative Adversarial Networks. *Math. Probl. Eng.* **2020**, *2020*. [CrossRef]
27. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
28. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A multimodal dataset for autonomous driving. *arXiv* **2019**, arXiv:1903.11027.
29. FLIR. 2015. Available online: http://www.flir.com (accessed on 5 January 2021).
30. Manjunath, A.; Liu, Y.; Henriques, B.; Engstle, A. Radar Based Object Detection and Tracking for Autonomous Driving. In Proceedings of the IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM), Munich, Germany, 15–17 April 2018; pp. 1–4.
31. John, V.; Nithilan, M.K.; Mita, S.; Tehrani, H.; Konishi, M.; Ishimaru, K.; Oishi, T. Sensor Fusion of Intensity and Depth Cues using the ChiNet for Semantic Segmentation of Road Scenes. In Proceedings of the Intelligent Vehicles Symposium, Changshu, China, 26–30 June 2018.
32. Everingham, M.; Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
33. Prechelt, L. Early Stopping-But When? In *Neural Networks: Tricks of the Trade*; Orr, G.B., Müller, K.R., Eds.; Springer: Berlin/Heidelberg, Germany, 1996; Volume 1524, pp. 55–69.