# Sensor Fusion for Radar Detection

Subjects: Computer Science, Artificial Intelligence

Contributor: Yi Zhou , Yutao Yue

Sensor fusion can be considered as the mapping of different modalities into a common latent space where different features of the same object can be associated together. Sensor fusion frameworks are classified into four categories: input fusion, ROI fusion, feature map fusion, and decision fusion.

automotive radars radar signal processing object detection multi-sensor fusion deep learning

## 1. Overview

Different sensors observe and represent an object with different features. Sensor fusion can be considered as the mapping of different modalities into a common latent space where different features of the same object can be associated together. The conventional taxonomy of fusion architectures into early (input), middle (feature), and late (decision) fusion is ambiguous for neural-network-based detection. For example, in the definition of late fusion, we cannot distinguish between ROI-level (without category information) fusion and object-level (with category information) fusion. Therefore, fusion methods are explicitly classified here according to the fusion stage. This is beneficial because different fusion stages correspond to different levels of semantics, i.e., the classification capabilities. As shown in **Figure 1**, fusion architectures are classified into four categories: input fusion, ROI fusion, feature map fusion, and decision fusion.

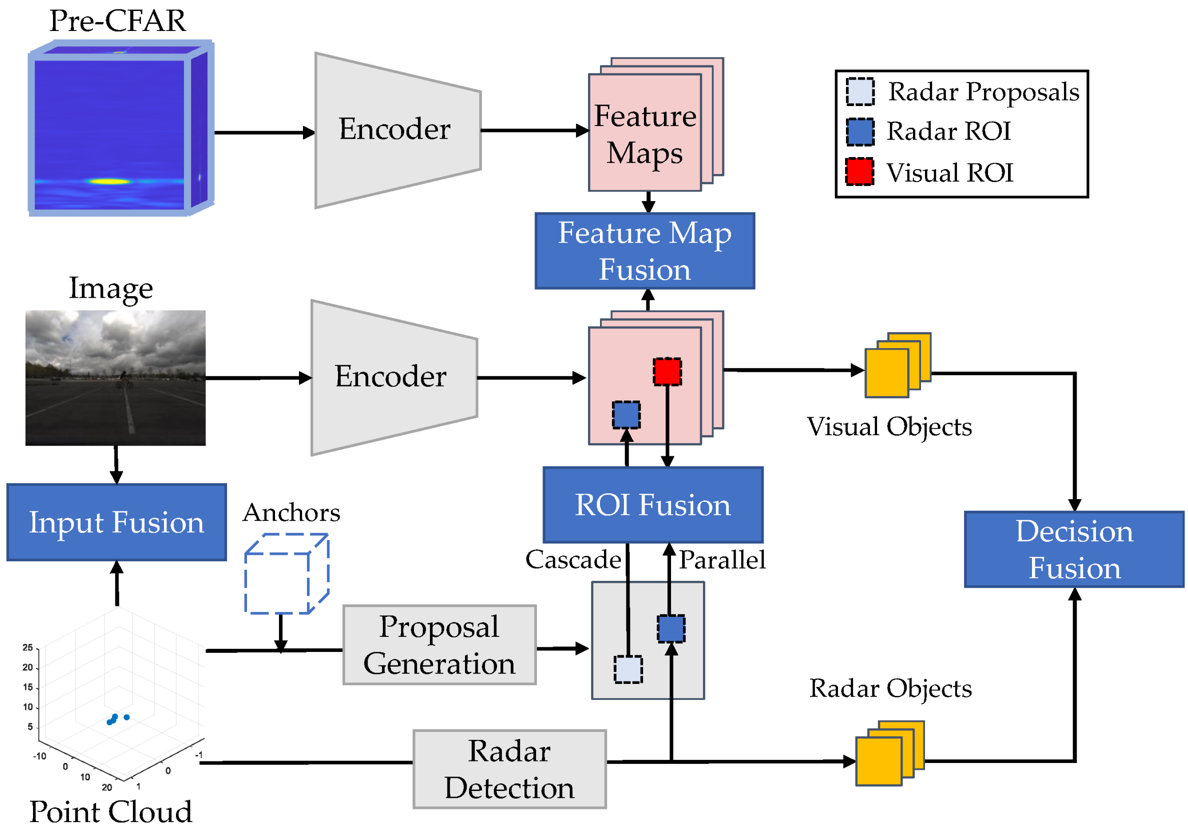

**Figure 1.** Overview of radar and camera fusion frameworks. The fusion frameworks are classified into input fusion, ROI fusion, feature map fusion, and decision fusion. For ROI fusion, two architectures are further investigated: cascade fusion, which projects radar proposals to image view, and parallel fusion, which fuses radar ROIs and visual ROIs.

## 2. Input Fusion

Input fusion is applied to the radar point cloud. It projects radar points into a pseudo-image with the range, velocity, and RCS as channels [1][2]. Then, similar to an RGB-depth image, the radar pseudo-image and the visual image are concatenated as a whole. Finally, a visual detector can be applied to this multi-channel image for detection. Input fusion does not make independent use of the detection capability of radar. In other words, the radar and vision modalities are tightly coupled. Assuming good alignment between modalities, it makes it easier for the network to learn joint feature embeddings. However, an obvious disadvantage is that the architecture is not robust to sensor failures.

The fusion performance depends on the alignment of radar detections with visual pixels. The difficulties lie in three aspects: Firstly, the radar point cloud is highly sparse. Many reflections from the surface are bounced away due to specular reflections. As a result, the detected points are sparsely distributed over the object. In addition to the sparsity, the lateral imprecision of radar measurements leads to further difficulties. The radar points can be out of the visual bounding box. The imprecision comes from different aspects, e.g., imprecise extrinsic calibration, multi-path effects, and low angular resolution. The third limitation is that low-resolution radar does not provide height information. To address these difficulties, some association techniques are required. Relying on the network to implicitly learn association is a hard task, because the network tends to simply ignore the weak modality, such as radar. The expansion methods described previously can be applied as a preprocessing stage for input fusion. However, object detection does not require such a strict association as depth completion, so some of the expansion methods are too costly for real-time processing. Nobis et al. [1] utilised the lightweight height extension as preprocessing. Both Chadwick et al. [2] and Yadav et al. [3] added a one-layer convolution to radar input before concatenation. This convolutional layer can be considered as a lightweight version of the association network. Radar detections at different ranges require different sizes of the receptive field for association. Therefore, Nobis et al. [1] concatenated the radar pseudo-image with image feature maps at multiple scales.

## 3. ROI Fusion

ROI fusion is adapted from the classical two-stage detection framework [4]. Regions of interest (ROIs) can be considered as a set of object candidates without category information. The fusion architecture can be further divided into cascade fusion and parallel fusion. In cascade fusion, radar detections are directly used for region proposal. Radar points are projected into image view as the candidate locations for anchors. Then, the ROI is determined with the help of visual semantics. In the second stage, each ROI is classified and its position is refined. Nabati et al. [5] adopted two techniques to improve the anchor quality. They added offsets to anchors to model the positional imprecision of radar detections. To mitigate the scale ambiguity in the image view, they rescaled the anchor size according to the range measurements. In their following work [6], they directly proposed 3D bounding boxes and then mapped these boxes to the image view. In this way, the rescaling step can be avoided. It is also possible to propose the region on the radar point cloud using visual ROIs. For example, CenterFusion [7] proposed a frustum-based association to generate radar ROI frustums using visual bounding boxes.

Cascade fusion is particularly well suited for low-resolution radars, where the radar point cloud has a high detection recall, but is very sparse. However, there are two potential problems with the cascade structure. Firstly, the performance is limited by the completeness of the proposed ROIs in the first stage. In other words, if an object is missed, we cannot recover it in the second stage. The second problem is that the cascade structure cannot take advantage of modality redundancy. If the radar sensor is nonfunctional, the whole sensing system will fail. Therefore, it is necessary to introduce a parallel structure to ROI fusion. Nabati et al. [6] adopted a two-branch structure for ROI fusion. The radar and visual ROIs are generated independently. Then, the fusion module merges radar ROIs and visual ROIs by taking a set union, while the redundant ROIs are removed through NMS. To enable the adaptive fusion of modalities, Kim et al. [8] proposed a gated region of interest fusion (GRIF) module for ROI fusion. It first predicts a weight for each ROI through a convolutional sigmoid layer. Then, the ROIs from radar and vision are multiplied by their corresponding weights and elementwise added together.

## 4. Feature Map Fusion

Feature-map fusion leverages the semantics from both radar and images. High-resolution radars can provide sufficient semantic cues for classification. Therefore, feature map fusion utilises two encoders to map radar and images into the same latent space with high-level semantics. The detection frameworks are flexible, including one-stage methods [9][10] and two-stage methods [11][12][13]. The one-stage methods leverage two branches of neural networks to extract feature maps from radar and images, respectively, and then concatenate the feature maps together. The two-stage fusion methods are adapted from the classical fusion architecture AVOD [14]. They firstly fuse the ROIs proposed from the radar and image in the first stage. In the second stage, the fused ROIs are projected to the radar and visual feature maps, respectively. The feature maps inside the ROIs are cropped and resized to an equal-sized feature crop. The feature crop pairs from the radar and image are then fused by the elementwise mean and sent to a detection head. Generally speaking, the two-stage method has better performance, but it is much slower than the one-stage method. Anchor-free methods [15][16] further avoid the complicated computation related to anchor boxes, such as calculating the IOU score during training.

Feature map fusion allows the network to flexibly combine radar and visual semantics. However, the fusion network may face the problem of overlooking weak modalities and modality synergies [17]. Some training techniques are needed to force the network to learn from radar input. Nobis et al. [1] adopted a modalitywise dropout approach that randomly deactivates the image branch during training. Lim et al. [9] used a weight freezing strategy to fix the weights of the pre-trained feature extractors when training the fusion branch. Experiments show that freezing only the image branch works best. However, the fusion of multiple modalities is not guaranteed to always be better than using a single modality. Sometimes, we want the network to lower the weight of the radar branch if it gives noisy inputs. To achieve adaptive fusion, Cheng et al. [18] adopted self-attention and global channel attention [19] in their network. The self-attention is used to enhance real target points and weaken clutter points. Then, the global attention module is applied to estimate modality-wise weights. Bijelic et al. [20] estimated the sensor entropy as the modality weight. For each modality, the entropy was evaluated pixel-wise as a weight mask. Then, these weight masks are multiplied with the corresponding feature maps at each fusion layer.

## 5. Decision Fusion

Decision fusion assumes that objects are detected independently by different modalities and fuses them according to their spatial-temporal relationships. This structure realises sensing redundancy at the system level and is therefore robust to modality-wise error. Due to the low resolution of radar, most existing studies do not explicitly consider the category information estimated by radar. In other words, they only fuse the location information from radar and vision branches, while retaining the category information estimated by vision. Since the next-generation 4D radar can provide classification capabilities, it is expected that future fusion frameworks should consider both location and category information.

The location can be optimally fused in a tracking framework. Different objects are first associated and then sent to a Bayesian tracking module for fusion. Due to the low resolution of radar, association is difficult to achieve in some scenarios, e.g., a truck splitting into two vehicles or two close objects merging into one. Such association ambiguity can be mitigated using a track-to-track fusion architecture [21]. By estimating tracks, temporal information can be leveraged to filter out false alarms and interpolate missed detections. Some researchers exploit deep learning to make a better association between radar and other modalities. RadarNet [15] proposed an attention-based late fusion to optimise the estimated velocity. Firstly, they trained a fiver-layer MLP with softmax to estimate the normalised association scores between each bounding box and its nearby radar detections. Then, they predicted the velocity by weighted averaging of the radar-measured velocities using the association scores. AssociationNet [22] attempts to map the radar detections to a better representation space in the contrastive learning framework. It first projects radar objects and visual bounding boxes to the image plane as pseudo images. To utilise the visual semantics, they concatenated these pseudo images with the original image. Next, the concatenated images are sent to an encoder–decoder network to output a feature map. Representation vectors are extracted from the feature map according to the locations of radar detections. A contrastive loss is designed to pull together the representation vectors of positive samples and push away the representation vectors of negative examples. During inference, they compute the Euclidean distance between the representation vectors of all possible radar–visual pairs. The pairs with a distance below the threshold are considered associative.

Category information, especially the conflict in category predictions, is difficult to handle in sensor fusion. BayesOD [23] proposes a probabilistic framework for fusing bounding boxes with categories. The locations of bounding boxes are modelled by Gaussian distributions. The category prior is modelled as a Dirichlet distribution, thereby allowing a Dirichlet posterior to be computed in closed form. Then, the bounding box with the highest categorical score is considered as the cluster centre, while the other bounding boxes are treated as measurements. Finally, Bayesian inference is used to optimally fuse the location and category information of these bounding boxes. Probabilistic methods have their inherent shortage in modelling the lack of knowledge [24]. For example, a uniform distribution brings confusion if either the network has no confidence in its prediction or the input is indeed ambiguous for classification. In contrast, set-based methods have no such problem. Chavez et al. [25] leveraged the evidential theory to fuse the LiDAR, camera, and radar. They considered the frame of discernment, i.e., the set of mutually exclusive hypotheses, as $\mathbf{\Omega} = \{pedestrians(p), bikes(b), cars(c), truck(t)\}$ , and assigned each possible hypothesis, i.e., a subset of Ω, with a belief. In the case of object detection, possible hypotheses are selected according to sensor characteristics. For example, a car is sometimes confused as part of a truck. Thus, if a car is detected, evidence should be also put into the set {c,t} and the set of ignorance Ω. Accordingly, we can assign the belief *m* to a car detection as

$$m\left(\{c\}\right) = \gamma_c \alpha_c, \quad m\left(\{c,t\}\right) = \gamma_c \left(1 - \alpha_c\right), \quad m\left(\mathbf{\Omega}\right) = 1 - \gamma_c \qquad (1)$$

where $\gamma_c$ is a discounting factor to model the uncertainty of misdetection and $\alpha_c$ is the accurateness, i.e., the rate of correct predictions in car detecting. Suppose there are two sources of evidence S1 and S2 from different modalities. Each of these sources provides a list of detections as A={$a_1$,$a_2$,…,$a_m$} and B={$b_1$,$b_2$,…,$b_n$}. Then, three propositions can be defined regarding the possible association of two detections $a_i$ and $b_j$ as:

- {1} if $a_i$ and $b_j$ are the same object;
- {0} if $a_i$ and $b_j$ are not the same object

- {0,1} for the ignorance of association.

The belief of association can be determined according to both location and category similarities. The evidence for location similarity is defined according to the Mahalanobis distance as

$$\begin{aligned} &m^p_{a_i,b_j}(\{0\}) = \alpha\left(1 - f\left(d_{a_i,b_j}\right)\right) m^p_{a_i,b_j} \\ &m^p_{a_i,b_j}(\{1\}) = \alpha f\left(d_{a_i,b_j}\right) \quad m^p_{a_i,b_j}(\{1,0\}) = 1 - \alpha \end{aligned} \tag{2}$$

where $f\left(d_{a_i,b_j}\right) = \exp\left(-\lambda d_{a_i,b_j}\right) \in [0,1]$ measure the similarity with respect to the Mahalanobis distance $d_{ai,bj}$ and a scaling factor λ and α is an evidence discounting factor. For the category similarity, two detections belonging to the same category is too weak to provide evidence that they are the same object. However, if two detections are of different categories, it is reasonable to assign evidence to the proposition that they are not the same object. Accordingly, the evidence for category similarity is given by

$$\begin{aligned} &m^c_{a_i,b_j}(\{0\}) = \sum_{A\cap B=\emptyset} m^c_{a_i}(A) m^c_{b_j}(B) \quad \forall A, B \subset \mathbf{\Omega} \\ &m^c_{a_i,b_j}(\{1\}) = 0, \quad m_{a_i,b_j}(\{0,1\}) = 1 - m^c_{a_i,b_j}(\{0\}) \end{aligned} \tag{3}$$

where the mass evidence is fused if no common category hypothesis is shared, i.e., A∩B=∅. The rest of the evidence is placed in the ignorance hypothesis. Finally, for each detection pair, the category similarity and the location similarity are fused according to Yager’s combination rule [26]. Evidential fusion provides a reliable framework for information fusion. However, it cannot be directly applied to neural-network-based detectors that make predictions on a single hypothesis. To address this problem, conformal prediction [27] can be used to generate confidence sets from a trained network using a small amount of calibration data.

## References

1. Nobis, F.; Geisslinger, M.; Weber, M.; Betz, J.; Lienkamp, M. A deep learning-based radar and camera sensor fusion architecture for object detection. In Proceedings of the 2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF), Bonn, Germany, 15–17 October 2019; pp. 1–7.

2. Chadwick, S.; Maddern, W.; Newman, P. Distant vehicle detection using radar and vision. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 8311–8317.

3. Yadav, R.; Vierling, A.; Berns, K. Radar+ RGB Fusion For Robust Object Detection In Autonomous Vehicle. In Proceedings of the 2020 IEEE International Conference on Image

Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 1986–1990.

4. 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.
5. Nabati, R.; Qi, H. Rrpn: Radar region proposal network for object detection in autonomous vehicles. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3093–3097.
6. Nabati, R.; Qi, H. Radar-camera sensor fusion for joint object detection and distance estimation in autonomous vehicles. arXiv 2020, arXiv:2009.08428.
7. Nabati, R.; Qi, H. Centrefusion: Centre-based radar and camera fusion for 3d object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 1527–1536.
8. Kim, Y.; Choi, J.W.; Kum, D. GRIF Net: Gated Region of Interest Fusion Network for Robust 3D Object Detection from Radar Point Cloud and Monocular Image. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; pp. 10857–10864.
9. 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 (NeurIPS Workshop), Vancouver, BC, Canada, 8–14 December 2019; Volume 2.
10. Zhang, J.; Zhang, M.; Fang, Z.; Wang, Y.; Zhao, X.; Pu, S. RVDet: Feature-level Fusion of Radar and Camera for Object Detection. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 2822–2828.
11. Qian, K.; Zhu, S.; Zhang, X.; Li, L.E. Robust Multimodal Vehicle Detection in Foggy Weather Using Complementary Lidar and Radar Signals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 444–453.
12. Kim, J.; Kim, Y.; Kum, D. Low-level Sensor Fusion Network for 3D Vehicle Detection using Radar Range-Azimuth Heatmap and Monocular Image. In Proceedings of the Asian Conference on Computer Vision (ACCV), Virtual, 30 November–4 December 2020.
13. Meyer, M.; Kuschk, G. Deep learning based 3d object detection for automotive radar and camera. In Proceedings of the 2019 16th European Radar Conference (EuRAD), Paris, France, 2–4 October 2019; pp. 133–136.
14. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.L. Joint 3d proposal generation and object detection from view aggregation. In Proceedings of the 2018 IEEE/RSJ International Conference

on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1–8.

15. Yang, B.; Guo, R.; Liang, M.; Casas, S.; Urtasun, R. Radarnet: Exploiting radar for robust perception of dynamic objects. In Proceedings of the 2020 European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 496–512.
16. Shah, M.; Huang, Z.; Laddha, A.; Langford, M.; Barber, B.; Zhang, S.; Vallespi-Gonzalez, C.; Urtasun, R. Liranet: End-to-end trajectory prediction using spatio-temporal radar fusion. arXiv 2020, arXiv:2010.00731.
17. Liu, Y.; Fan, Q.; Zhang, S.; Dong, H.; Funkhouser, T.; Yi, L. Contrastive multimodal fusion with tupleinfonce. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 754–763.
18. Cheng, Y.; Xu, H.; Liu, Y. Robust Small Object Detection on the Water Surface Through Fusion of Camera and Millimeter Wave Radar. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 15263–15272.
19. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
20. Bijelic, M.; Gruber, T.; Mannan, F.; Kraus, F.; Ritter, W.; Dietmayer, K.; Heide, F. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 11682–11692.
21. Matzka, S.; Altendorfer, R. A comparison of track-to-track fusion algorithms for automotive sensor fusion. In Multisensor Fusion and Integration for Intelligent Systems; Springer: Berlin/Heidelberg, Germany, 2009; pp. 69–81.
22. Dong, X.; Zhuang, B.; Mao, Y.; Liu, L. Radar Camera Fusion via Representation Learning in Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 1672–1681.
23. Harakeh, A.; Smart, M.; Waslander, S.L. Bayesod: A bayesian approach for uncertainty estimation in deep object detectors. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Virtual, 31 May–31 August 2020; pp. 87–93.
24. Hüllermeier, E.; Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Mach. Learn. 2021, 110, 457–506.
25. Chavez-Garcia, R.O.; Aycard, O. Multiple sensor fusion and classification for moving object detection and tracking. IEEE Trans. Intell. Transp. Syst. 2015, 17, 525–534.

26. Florea, M.C.; Jousselme, A.L.; Bossé, É.; Grenier, D. Robust combination rules for evidence theory. Inf. Fusion 2009, 10, 183–197.
27. Angelopoulos, A.N.; Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv 2021, arXiv:2107.07511.

Retrieved from https://encyclopedia.pub/entry/history/show/57377