Cross-View and Cross-Modal Contrastive Learning for Radar Object Detection

Qiaolong Qian, Yi Shi, Ruichao Hou, *Member, IEEE*, Haoyu Qin, and Gangshan Wu, *Member, IEEE*

***Abstract*—Frequency-modulated continuous-wave radar is a cornerstone of advanced driver assistance systems thanks to its low cost and resilience to adverse weather. Yet the absence of explicit semantics makes radar annotation difficult, and the scarcity of large-scale labeled data limits the performance of radar perception models. To address this issue, we propose a self-supervised framework for object detection directly from Range–Azimuth–Doppler (RAD) cubes that learns transferable representations from unlabeled radar data. Specifically, we introduce cross-view contrastive learning to model correspondences among complementary views of the RAD cube, encouraging the network to capture spatial structure from multiple perspectives. In addition, an auxiliary cross-modal contrastive objective distills semantic knowledge from vision into radar. The joint objective integrates cross-view and cross-modal signals to strengthen radar feature representations. We further extend the framework to cross-domain pretraining using datasets from different sources. Experimental results demonstrate that the proposed method significantly improves radar object detection performance, especially with limited labeled data.**

***Index Terms*—Radar object detection, cross-view, cross-modal, contrastive learning**

I. Introduction

FREQUENCY-modulated continuous-wave radar [1] is widely used in advanced driver-assistance systems due to its low cost and robustness in adverse weather [2], [3]. Raw ADC returns are converted via cascaded FFTs into Range–Azimuth–Doppler (RAD) tensors that encode range, azimuth, and radial velocity, enabling radar perception tasks such as object detection [4]–[11], tracking [12], and semantic segmentation [7], [13], [14].

However, radar signals are difficult for humans to interpret, making manual annotation slow and costly. The scarcity of large-scale labeled radar datasets, therefore, severely constrains performance. Existing automatic or semi-automatic labeling methods [4], [15]–[18] rely on supervision from other modalities, but face several challenges. First, the need for other sensors, such as LiDAR and stereo cameras, increases data collection costs. Second, label quality suffers from projection inaccuracies across modalities, requiring cumbersome calibration and manual correction. Finally, poor generalization across different radar setups due to hardware variation. Recently, self-supervised learning [19]–[21] has been explored for radar perception [10], [22]–[27] to reduce the reliance on manual annotations. For example, Zhuang et al. [26] adopt a masked image modeling paradigm to reconstruct masked radar heatmaps. However, most existing methods operate on a single radar view, such as bird's-eye view (BEV) [24], [28], the Range–Azimuth (RA) view [22], or the Range–Doppler (RD) view [23], and thus overlook the inherent correspondence among different views of RAD data. Moreover, Zhu et al. [29] model multi-view relationships within radar in a radar-only manner, but do not incorporate cross-modal alignment.

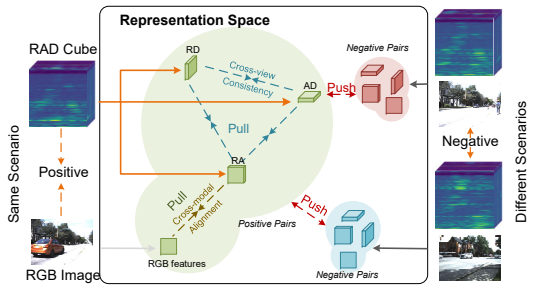

Fig. 1. Diagram of the proposed self-supervised learning method. Given pairs of RAD cube and RGB image, the features of three views of the RAD cube and RGB image are projected into a common representation space. We pull positive pairs together and push negative pairs apart by a contrastive loss.

To address these issues, we propose a cross-view and cross-modal (CVCM) contrastive learning framework that learns radar–vision aligned RAD representations.By operating at the feature level, our approach reduces dependence on labeled data and is more robust to calibration noise and temporal misalignment. As shown in Fig. 1, the framework embeds three RAD views and the corresponding RGB image into a shared feature space to learn generalizable representations from unlabeled data. We design two complementary objectives: a cross-view contrastive loss to model spatial correspondences among RAD views, and a cross-modal contrastive loss to align RAD tensors with visual semantics. The joint learning objective enables the model to capture both semantic priors from vision and spatial cues from radar. Experiments on RADDet [4] and cross-domain evaluations on CARRADA [16] and UWCR [30] demonstrate significant gains under limited supervision and strong transferability across datasets.

The main contributions can be summarized as follows:

- We propose a novel self-supervised learning framework for RAD tensors, enabling the model to learn transferable

This work was supported by the National Natural Science Foundation of China (62072232), the Key R&D Project of Jiangsu Province (BE2022138), the Fundamental Research Funds for the Central Universities (021714380026), and the Collaborative Innovation Center of Novel Software Technology and Industrialization. *(Corresponding authors: Ruichao Hou)*

Qiaolong Qian, Yi Shi, Ruichao Hou, and Gangshan Wu are with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210008, China (e-mail: {qql, yishi}@smail.nju.edu.cn; {rchou, gswu}@nju.edu.cn).

Qiaolong Qian and Haoyu Qin are also with the 8th Research Academy of CSSC, Yangzhou 225000, China (e-mail: qinhaoyu99@163.com).

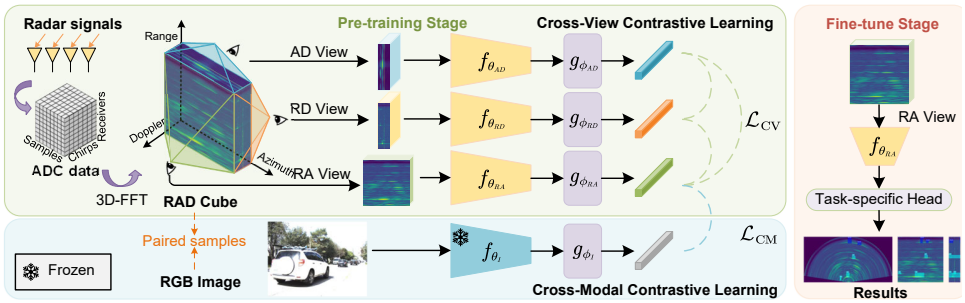

Fig. 2. The overall framework of the proposed method. The pre-training stage consists of two parts: (1) cross-view contrastive learning, which constructs correspondence between pairs of different views of the RAD cube; (2) cross-modal contrastive learning, which transfers knowledge from a pre-trained vision backbone to the RAD model. After pre-training, only the pre-trained RA view backbone and a task-specific head are used for downstream tasks.

representations from unlabeled radar data.

- We introduce a cross-view contrastive learning strategy to model the correspondence among different views of the RAD cube, facilitating spatial distribution understanding from multiple perspectives.
- We extend cross-modal contrastive learning between RAD tensors and visual data, transferring semantic priors from vision to radar.

II. Methodology

A. Overview

Suppose we are given a dataset $\mathcal{D} = \{(\mathbf{C}_i, \mathbf{I}_i)_{i=1}^{|\mathcal{D}|}\}$, where $\mathbf{C}_i \in \mathbb{R}^{R\times A\times D}$ is a RAD cube and $\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}$ is the corresponding image. The RAD cube is obtained from the raw ADC data via FFT. R, A, and D represent range, azimuth, and Doppler sample size, respectively. $\mathbf{C}_i^{RD} \in \mathbb{R}^{R\times D\times A}$, $\mathbf{C}_i^{AD} \in \mathbb{R}^{A\times D\times R}$ are obtained by transposing the RAD cube. For the purpose of pre-training a RA view feature extractor $f_{\theta_{RA}}(\cdot)$, we use three auxiliary encoders $f_{\theta_{RD}}(\cdot)$, $f_{\theta_{AD}}(\cdot)$ and $f_{\theta_I}(\cdot)$ to extract latent features of different inputs. Then there follows a projection head using multi-layer perceptron for each encoder, namely $g_{\phi_{RA}}(\cdot)$, $g_{\phi_{RD}}(\cdot)$, $g_{\phi_{AD}}(\cdot)$ and $g_{\phi_I}(\cdot)$, to project the features to a common representation space.

We learn transferable RAD representations via self-supervised pre-training on paired RAD tensors and RGB images. Features from the same sample form positives, while those from different samples serve as negatives. Following contrastive learning [19], [20], we pull positives together and push negatives apart in the embedding space. Our joint objective combines a cross-view loss to enforce consistency across three views and a cross-modal loss to align radar and image features. An overview is shown in Fig. 2.

B. Cross-View Contrastive Learning

The RA view of the RAD tensor encodes the range and azimuth of objects, but the spatial distribution of Doppler velocities is lost when processed by a 2D CNN [4]. Similarly, the RD view represents object range and speed but ignores azimuthal cues, while the AD view lacks range information. Unlike previous multi-view radar perception methods [7], [13], [14], [29], [31] that directly combine three views, our method leverages complementary cues across views to enhance RAD tensor representation learning.

Given the RA view tensor $\mathbf{C}_i^{RA}$, we construct cubes of other views $\mathbf{C}_i^{RD}$ and $\mathbf{C}_i^{AD}$ by performing transposition. $\mathbf{C}_i^{RA}$ is mapped to feature embedding space through the RA view feature extractor $f_{\theta_{RA}}$, and the resulting vector is projected to an invariant space through the projection head $g_{\phi_{RA}}$. Cubes of other views are processed in the similar way. The process can be formulated as:

$$\mathbf{z}_i^V = g_{\phi_V}\left(f_{\theta_V}\left(\mathbf{C}_i^V\right)\right), \tag{1}$$

where V denotes views of the input RAD tensor, *i.e.* RA, RD and AD, $\mathbf{z}_i^V$ denotes the projected vectors of $\mathbf{C}_i^V$.

Different views of the same cube form positive pairs of examples, and views of different cubes in the mini-batch yield negative pairs. For the positive pair of two views $\mathbf{z}_i^{V1}$ and $\mathbf{z}_i^{V2}$, we leverage NT-Xent loss proposed in SimCLR [19] to compute the contrastive loss:

$$l\left(i, V1, V2\right) = -\log \frac{\exp\left(s\left(\mathbf{z}_i^{V1}, \mathbf{z}_i^{V2}\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{i\neq k} \exp\left(s\left(\mathbf{z}_i^{V1}, \mathbf{z}_k^{V2}\right)/\tau\right)}, \tag{2}$$

where $\mathbf{1}_{i\neq k} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $i \neq k$, N is the mini-batch size, τ is the temperature coefficient, $s(\cdot)$ denotes the cosine similarity function, $V1$ and $V2$ denote views of the input RAD tensor, *i.e.* RA, RD and AD. The contrastive loss between view V1 and V2 is defined as follows:

$$\mathcal{L}_{V1\leftrightarrow V2} = \frac{1}{2N}\sum_{i=1}^{N}\left[l\left(i, V1, V2\right) + l\left(i, V2, V1\right)\right]. \tag{3}$$

The contrastive loss enforces positive pairs of examples to be pulled closer and the negative pairs to be pushed away in a mini-batch. For every pair of two views, namely RA and RD, RD and AD, and RA and AD, we compute the contrastive loss for them. The total cross-view loss is calculated as follows:

$$\mathcal{L}_{CV} = \mathcal{L}_{RA\leftrightarrow RD} + \mathcal{L}_{RA\leftrightarrow AD} + \mathcal{L}_{RD\leftrightarrow AD}. \tag{4}$$

C. Cross-Modal Contrastive Learning

The visual modality provides rich semantics such as object position and category, enabling the RAD model to learn prior knowledge from images. While several studies [22], [24], [25] have explored radar–vision contrastive learning, contrastive learning between 3D RAD tensors and 2D images remains underexplored.

Given the RA view tensor $\mathbf{C}_i^{RA}$ and the corresponding image $\mathbf{I}_i$, we leverage a pre-trained vision backbone f_{θ_I} to extract the image features. The vision backbone is frozen during pre-training of the RAD model. Then, an image projection head is used to project the feature vectors to the latent feature space. The projected image feature is defined as follows:

$$\mathbf{z}_i^I = g_{\phi_I}\left(f_{\theta_I}\left(\mathbf{I}_i\right)\right). \tag{5}$$

The RA view encodes range–azimuth, which directly describes an object's spatial location in radar coordinates and thus has a more natural correspondence to the camera's 2D spatial layout. In contrast, RD and AD lack explicit range, making their signatures more motion-dependent and less geometrically compatible with a single RGB image. Therefore, we apply cross-modal contrastive learning only between the RA-view RAD tensor and the corresponding image. Similar to the contrastive loss defined between two views, the cross-modal contrastive loss is formulated as:

$$l\left(i, RA, I\right) = -\log \frac{\exp\left(s\left(\mathbf{z}_i^{RA}, \mathbf{z}_i^I\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{i\neq k} \exp\left(s\left(\mathbf{z}_i^{RA}, \mathbf{z}_k^I\right)/\tau\right)}, \tag{6}$$

where $\mathbf{1}_{i\neq k}$, s, N, τ refers to the same parameters as in (2). The cross-modal loss function $\mathcal{L}_{CM}$ is then defined as follows:

$$\mathcal{L}_{CM} = \frac{1}{2N}\sum_{i=1}^{N}\left[l\left(i, RA, I\right) + l\left(i, I, RA\right)\right]. \tag{7}$$

The total loss function is given by:

$$\mathcal{L} = \lambda\mathcal{L}_{CV} + (1-\lambda)\mathcal{L}_{CM}. \tag{8}$$

D. Network Architectures

For the radar branch, we adopt RadResNet [4] as the backbone, which consists of several residual and downsampling blocks. For the RA view, the RadResNet backbone remains unchanged for fair comparison. For the RD and AD views, since their channel dimensions are much larger, we simplify the network by using fewer residual blocks and smaller convolution kernels to reduce computational cost. For the vision branch, we use YOLOv9-c [32] pre-trained on COCO [33] as the image backbone. A two-layer MLP projection head maps the extracted features into a 256-dimensional latent space. After pre-training, only the RA feature extractor $f_{\theta_{RA}}$ is retained for downstream tasks, while other extractors and projection heads are discarded.

III. EXPERIMENTS

A. Datasets

We conduct experiments on RADDet [4] and use CARRADA [16] and UWCR [30] for pre-training to evaluate cross-domain transferability. RADDet contains 10,158 frames of synchronized images and RAD tensors. We follow the official data split and use all unlabeled training samples for pre-training. During fine-tuning, 5%, 10 %, 20 %, 50 %, and 100 % of the labeled training data are used. CARRADA provides 7,193 frames of RA and RD maps, unlabeled RAD tensors, and images, while UWCR offers 19,800 frames of raw ADC radar data converted into RAD tensors via FFT.

B. Experimental Settings

To evaluate the effectiveness of our self-supervised pre-training, we build upon the RADDet model [4], which employs two detection heads: a RAD YOLO head for 3D RAD bounding boxes and a 2D YOLO head for Cartesian detections. RADDet first trains the RAD YOLO head, then freezes the backbone before training the 2D head. In contrast, we initialize the backbone with pre-trained weights and jointly fine-tune it with the 2D detection head to assess the impact of our pre-training strategy. Performance is evaluated using mean Average Precision (mAP) at IoU thresholds of 0.1, 0.3, 0.5, and 0.7. During pre-training, the model is trained for 150 epochs with a batch size of 16, an initial learning rate of 0.0001, and a decay rate of 0.96 every 10k steps after 60k warm-up steps. For fine-tuning, we follow the same hyperparameter settings as RADDet for fair comparison.

C. Results

To evaluate the impact of labeled data quantity, we fine-tune the model using different proportions of labeled samples (5%, 10%, 20%, 50%, and 100%) for two object detection tasks. Our proposed method is compared with training from scratch, and detailed results are shown in Table I. CVCM consistently improves performance across all label scales. When sufficient labeled data are available, the improvement is marginal, as the ground truth provides enough supervision. However, with fewer labeled samples, the performance gains become more significant, demonstrating the effectiveness of self-supervised pre-training under limited supervision. As shown in Fig. 3, models initialized with pre-trained weights converge faster and achieve higher final accuracy than those trained from scratch, indicating that pre-training enables the model to learn transferable knowledge from unlabeled data.

TABLE I
THE PERFORMANCE ON DIFFERENT FRACTIONS OF LABELED DATA.

Task	Fraction	Method	$AP_{0.1}$	$AP_{0.3}$	$AP_{0.5}$	$AP_{0.7}$
3D RAD	5%	scratch	0.405	0.241	0.082	0.012
		CVCM	**0.584**	**0.342**	**0.128**	**0.023**
	10%	scratch	0.482	0.311	0.121	0.022
		CVCM	**0.615**	**0.391**	**0.160**	**0.023**
	20%	scratch	0.589	0.400	0.165	0.030
		CVCM	**0.671**	**0.461**	**0.198**	**0.038**
	50%	scratch	0.710	0.511	0.225	0.045
		CVCM	**0.736**	**0.539**	**0.245**	**0.053**
	100%	scratch	0.764	0.563	0.251	0.059
		CVCM	**0.776**	**0.583**	**0.273**	**0.063**
2D BEV	5%	scratch	0.262	0.175	0.079	0.013
		CVCM	**0.379**	**0.275**	**0.129**	**0.027**
	10%	scratch	0.418	0.344	0.204	0.046
		CVCM	**0.543**	**0.433**	**0.255**	**0.061**
	20%	scratch	0.611	0.522	0.353	0.111
		CVCM	**0.721**	**0.628**	**0.422**	**0.142**
	50%	scratch	0.790	0.695	0.473	0.173
		CVCM	**0.833**	**0.734**	**0.516**	**0.191**
	100%	scratch	0.855	0.778	0.546	0.225
		CVCM	**0.877**	**0.791**	**0.573**	**0.250**

D. Ablation Study and Analysis

Impact of joint learning objective. We evaluate the impact of combining cross-view and cross-modal contrastive learning

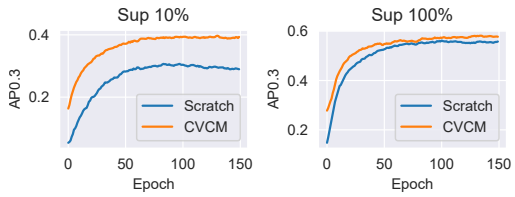

Fig. 3. Smoothed AP0.3 curves of the validation set when fine-tuning with 10% and 100% of labeled data in the 3D RAD detection task.

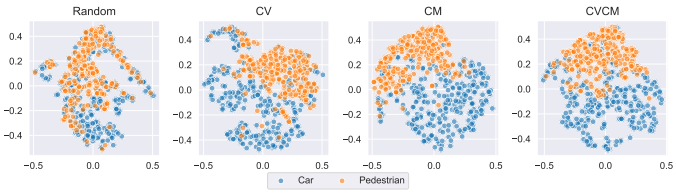

Fig. 4. t-SNE visualization of features on the test set of RADDet.

by pre-training models with each component individually, using 10% of labeled data for 3D RAD detection. As shown in Table II, the joint objective achieves better performance, where cross-view contrastive learning captures spatial relations in the Doppler dimension and cross-modal learning transfers semantic cues from vision. Fig. 4 further shows that the randomly initialized model fails to separate classes, whereas the jointly trained model yields clearer boundaries between car and pedestrian, indicating stronger feature discrimination.

TABLE II
EFFECT OF CROSS-VIEW AND CROSS-MODAL CONTRASTIVE LEARNING AND LOSS WEIGHTING.

CV	CM	λ	$AP_{0.1}$	$AP_{0.3}$	$AP_{0.5}$	$AP_{0.7}$
✗	✗	–	0.482	0.311	0.121	0.022
✓	✗	1.0	0.591	0.356	0.135	0.022
✗	✓	0.0	0.583	0.363	0.142	0.022
✓	✓	0.25	0.597	0.373	0.142	0.019
✓	✓	**0.50**	**0.615**	**0.391**	**0.160**	**0.023**
✓	✓	0.75	0.595	0.367	0.137	0.020

Combination of different views. To examine the effect of each view pair in cross-view contrastive learning, we disable $\mathcal{L}_{RA\leftrightarrow RD}$, $\mathcal{L}_{RA\leftrightarrow AD}$, or $\mathcal{L}_{RD\leftrightarrow AD}$ in Eq.(4) during pre-training, excluding cross-modal learning for clarity. The model is fine-tuned on 10% labeled data. As shown in Table III, a single view pair yields minor gains, while combining all three markedly improves representation.

TABLE III
COMPARISONS OF DIFFERENT COMBINATIONS IN CROSS-VIEW CONTRASTIVE LEARNING.

Cross-view	$AP_{0.1}$	$AP_{0.3}$	$AP_{0.5}$	$AP_{0.7}$
✗	0.482	0.311	0.121	0.022
RA↔RD	0.501	0.313	0.114	0.022
RA↔AD	0.507	0.304	0.121	0.021
RA↔RD, **RA**↔AD	0.550	0.336	0.130	**0.024**
RA↔RD, **RA**↔AD, RD↔AD	**0.591**	**0.356**	**0.135**	0.022

Impact of cross-domain pre-training. We further extend our method to pre-training on other datasets. Due to differences in radar and camera setups, cross-modal contrastive learning becomes challenging when combining datasets, so we pre-train on each dataset separately. Specifically, the model is pre-trained on CARRADA or UWCR and then fine-tuned on RADDet. As shown in Table IV, cross-domain pre-training improves performance even with few labeled samples, despite large domain gaps such as fewer object categories in CARRADA and distinct radar configurations in UWCR. However, using the same-domain data still yields the best results.

TABLE IV
THE PERFORMANCE OF CROSS-DOMAIN PRE-TRAINING.

pre-training dataset	$AP_{0.1}$	$AP_{0.3}$	$AP_{0.5}$	$AP_{0.7}$
None	0.482	0.311	0.121	0.022
CARRADA	0.541	0.337	0.133	0.024
UWCR	0.586	0.361	0.137	**0.029**
RADDet	**0.615**	**0.391**	**0.160**	0.023

E. Qualitative Evaluation

In Fig. 5, we present detection results of models fine-tuned on the full training set. In the first scene, the ground truth omits a pedestrian that our model correctly detects, while the original RADDet fails. This improvement results from cross-modal contrastive learning, which aligns radar and visual features and reduces dependence on imperfect annotations. In the second scene, our model achieves more accurate classifications as semantic cues are transferred from vision. Furthermore, cross-view contrastive learning compensates for missing spatial information in individual views, yielding more precise bounding boxes, especially along the Doppler dimension.

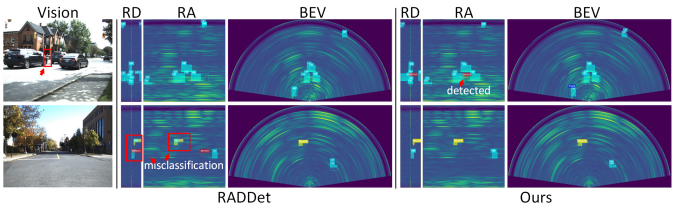

Fig. 5. Qualitative results on the test set of RADDet. Ground truth boxes are labeled with facecolors and predictions are without facecolors.

IV. CONCLUSION

In this work, we proposed a novel self-supervised method for RAD representation learning, which allows the RAD model to learn transferable representations from unlabeled data. A joint cross-view and cross-modal contrastive learning objective was introduced to facilitate representation learning for RAD. The experimental results on two downstream tasks demonstrate that our proposed method significantly boosts the detection performance with limited labeled data. Furthermore, extensive experimental results show the feasibility of transferring knowledge from different domains, opening up the possibility of few-shot learning by utilizing other datasets for radar perception tasks.

REFERENCES

[1] A. G. Stove, "Linear fmcw radar techniques," in *IEE Proceedings F (Radar and Signal Processing)*, 1992.

[2] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "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*, 2020.

[3] S. Zang, M. Ding, D. Smith, P. Tyler, T. Rakotoarivelo, and M. A. Kaafar, "The impact of adverse weather conditions on autonomous vehicles: How rain, snow, fog, and hail affect the performance of a self-driving car," *IEEE Vehicular Technology Magazine*, vol. 14, no. 2, pp. 103–111, 2019.

[4] A. Zhang, F. E. Nowruzi, and R. Laganiere, "Raddet: Range-azimuth-doppler based radar object detection for dynamic road users," in *Conference on Robots and Vision*, 2021.

[5] Y.-J. Li, S. Hunt, J. Park, M. O'Toole, and K. Kitani, "Azimuth super-resolution for fmcw radar in autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[6] H. Lai, G. Luo, Y. Liu, and M. Zhao, "Enabling visual recognition at radio frequency," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 388–403.

[7] X. Gao, G. Xing, S. Roy, and H. Liu, "Ramp-cnn: A novel neural network for enhanced automotive radar object recognition," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 5119–5132, 2020.

[8] S. Madani, J. Guan, W. Ahmed, S. Gupta, and H. Hassanieh, "Radatron: Accurate detection using multi-resolution cascaded mimo radar," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 160–178.

[9] Z. Gu, J. Ma, Y. Huang, H. Wei, Z. Chen, H. Zhang, and W. Hong, "Hgsfusion: Radar-camera fusion with hybrid generation and synchronization for 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, 2025, pp. 3185–3193.

[10] Y. Luo, R. Hoffmann, Y. Xia, O. Wysocki, B. Schwab, T. H. Kolbe, and D. Cremers, "Radler: Radar object detection leveraging semantic 3d city models and self-supervised radar-image learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4452–4461.

[11] F. Luo, A. Li, J. He, Z. Yu, K. Wu, B. Jiang, and L. Wang, "Improved multi-task radar sensing via attention-based feature distillation and contrastive learning," *IEEE Transactions on Information Forensics and Security*, 2025.

[12] P. Zhao, C. X. Lu, B. Wang, N. Trigoni, and A. Markham, "3d motion capture of an unmodified drone with single-chip millimeter wave radar," in *2021 IEEE International Conference on Robotics and Automation*, 2021.

[13] A. Ouaknine, A. Newson, P. Pérez, F. Tupin, and J. Rebut, "Multi-view radar semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[14] Y. Dalbah, J. Lahoud, and H. Cholakkal, "Transradar: Adaptive-directional transformer for real-time multi-view radar semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[15] Y. Wang, Z. Jiang, X. Gao, J.-N. Hwang, G. Xing, and H. Liu, "Rodnet: Radar object detection using cross-modal supervision," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

[16] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Pérez, "Carrada dataset: Camera and automotive radar with range-angle-doppler annotations," in *International Conference on Pattern Recognition*, 2021.

[17] J. Rebut, A. Ouaknine, W. Malik, and P. Pérez, "Raw high-definition radar for multi-task learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[18] X. Huang, Z. Xu, H. Wu, J. Wang, Q. Xia, Y. Xia, J. Li, K. Gao, C. Wen, and C. Wang, "L4dr: Lidar-4dradar fusion for weather-robust 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 3806–3814.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[21] M. Chen, Y. Liu, Z. Zhang, and W. Guo, "Rcrfnet: Enhancing object detection with self-supervised radar–camera fusion and open-set recognition," *Sensors*, vol. 24, no. 15, p. 4803, 2024.

[22] M. Alloulah, A. D. Singh, and M. Arnold, "Self-supervised radio-visual representation learning for 6g sensing," in *IEEE International Conference on Communications*, 2022.

[23] C. Decourt, R. VanRullen, D. Salle, and T. Oberlin, "Leveraging self-supervised instance contrastive learning for radar object detection," *arXiv preprint arXiv:2402.08427*, 2024.

[24] Y. Hao, S. Madani, J. Guan, M. Alloulah, S. Gupta, and H. Hassanieh, "Bootstrapping autonomous driving radars with self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 012–15 023.

[25] M. Alloulah and M. Arnold, "Look, radiate, and learn: Self-supervised localisation via radio-visual correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[26] L. Zhuang, T. Jiang, J. Wang, Q. An, K. Xiao, and A. Wang, "Effective mmwave radar object detection pre-training based on masked image modeling," *IEEE Sensors Journal*, 2023.

[27] K. Hou, X. Du, G. Cui, X. Chen, J. Zheng, Y. Rong, and W. Ma, "A hybrid network-based contrastive self-supervised learning method for radar signal modulation recognition," *IEEE Transactions on Vehicular Technology*, 2025.

[28] H. Zhao, R. Guan, T. Wu, K. L. Man, L. Yu, and Y. Yue, "Unibevfusion: Unified radar-vision bevfusion for 3d object detection," in *IEEE International Conference on Robotics and Automation*, 2025, pp. 6321–6327.

[29] H. Zhu, H. He, A. Choromanska, S. Ravindran, B. Shi, and L. Chen, "Multi-view radar autoencoder for self-supervised automotive radar representation learning," in *IEEE Intelligent Vehicles Symposium*, 2024, pp. 1601–1608.

[30] X. Gao, Y. Luo, G. Xing, S. Roy, and H. Liu, "Raw adc data of 77ghz mmwave radar for automotive object detection," 2022. [Online]. Available: https://dx.doi.org/10.21227/xm40-jx59

[31] L. Zhang, X. Zhang, Y. Zhang, Y. Guo, Y. Chen, X. Huang, and Z. Ma, "Peakconv: Learning peak receptive field for radar semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[32] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," *arXiv preprint arXiv:2402.13616*, 2024.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014.