Bootstrapping Autonomous Radars with Self-Supervised Learning

Yiduo Hao[2*†] Sohrab Madani[3†] Junfeng Guan[1]

Mohammed Alloulah[4] Saurabh Gupta[3] Haitham Hassanieh[1]

[1] École Polytechnique Fédérale de Lausanne (EPFL) [2] University of Cambridge

[3] University of Illinois Urbana-Champaign (UIUC) [4] Nokia Bell Labs

Abstract

The perception of autonomous vehicles using radars has attracted increased research interest due its ability to operate in fog and bad weather. However, training radar models is hindered by the cost and difficulty of annotating large-scale radar data. To overcome this bottleneck, we propose a self-supervised learning framework to leverage the large amount of unlabeled radar data to pre-train radar-only embeddings for self-driving perception tasks. The proposed method combines radar-to-radar and radar-to-vision contrastive losses to learn a general representation from unlabeled radar heatmaps paired with their corresponding camera images. When used for downstream object detection, we demonstrate that the proposed self-supervision framework can improve the accuracy of state-of-the-art supervised baselines by 5.8% *in mAP.*

1. Introduction

Millimeter-wave (mmWave) radars have received increased interest from the self-driving cars industry owing to its cost-effectiveness and its ability to operate in adverse weather conditions where cameras and lidar fail like in fog, smog, snowstorms, and sandstorms [43, 44, 68]. As such, there has been a significant amount of work, from both academia [14, 25, 56, 60] and industry [41, 42, 48, 53], on developing data-driven methods for semantic scene understanding on top of radar signals. Moreover, the advent of standard commercial automotive radars has made real-world deployments and large-scale data collection campaigns possible and several automotive radar datasets have recently been curated [14, 15, 39, 41, 45, 48, 56, 60, 70].

However, compared to de facto computer vision datasets like ImageNet, the volume of annotated open radar datasets remains very limited. This is because radar images are especially challenging for humans to interpret and thus annotate. Fig. 1 shows an example of bird's eye view (BEV) radar heatmaps and the corresponding camera images. Unlike camera images, radar heatmaps appear as blobs with no sharp boundaries or well-defined shapes for the objects present in the scene. These blobs carry little to no contextual or perceptual information and, as such, are hard to interpret by humans. Furthermore, mmWave radar signals are highly specular; meaning, mmWave signals exhibit mirror-like reflections on cars [12]. As a result, not all reflections from the car propagate back to the radar receiver, and most of the car does not appear in the image. These effects compound making it difficult even for well-trained radar imaging experts to draw precise bounding boxes of objects [25]. As a result, only a tiny fraction of radar data is typically labeled (e.g. 10%) of the hundreds of thousands of raw radar frames in open radar datasets [39]. Hence, building accurate supervised radar object detection models is extremely difficult.

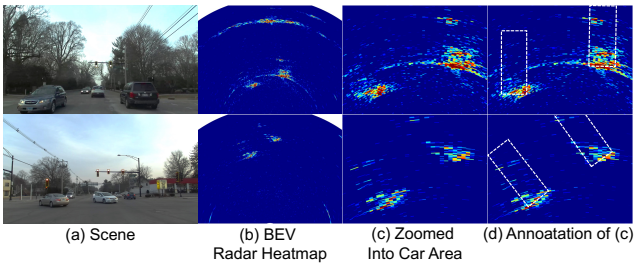

Figure 1. **Millimeter wave radar heatmaps are uninterpretable to humans and are hence difficult to annotate.**

To address the challenge of annotating radar data, prior work leverages other sensing modalities like cameras and lidar to derive labels for radar heatmaps and use these labels as groundtruth to train radar-based models [22, 35, 36, 47, 57, 59, 61]. However, different sensory modalities have different viewpoints and projection planes of the scene. For example, camera-based labels suffer from depth-unaware perspective projection onto the image plane, so they cannot provide accurate supervision along the depth axis in

*Work done while at EPFL.

†denotes co-primary first authors.

BEV radar heatmaps. Errors in viewpoint alignment between the different sensory modalities also result in highly inaccurate detection. Moreover, because radar and optical sensors (camera and lidar) operate on orthogonal portions of the electromagnetic spectrum, objects that are visible to optical sensing are not necessarily visible to radar and vice versa. Directly using lidar data to supervise the training of radar will force the radar model to focus too much on less prominent reflections in radar heatmaps, such as less-visible surfaces due to specularity. In contrast, certain materials, such as glass, are not visible to optical sensors but are visible to radars. Therefore, cross-modal supervision results in false positive (FP) and false negative (FN) detection [36]. Finally, as radar hardware continues to evolve, it requires us to keep labeling new datasets collected using new radar hardware, which is going to be very expensive in the long run.

In this paper, we aim to leverage large-scale unlabeled radar data but bypass the complexities of explicit annotations. We propose a self-supervised learning approach that uses a joint embedding architecture to pre-train a radar object detector using distillation from vision and radar itself. Learning under our cross-modal and intra-modal objectives happens at the mutual information level [4, 46], rather than explicitly annotating radar data as in prior work [22, 35, 36, 47, 57, 59, 61].

Applying self-supervised learning (SSL), which has been extensively studied in the NLP and CV communities, to the radar domain, is nontrivial because state-of-the-art self-supervised learning methods are designed for camera images. They either design pretext prediction tasks for RGB images [21, 30], or leverage camera-specific attributes to design strong augmentations to enforce semantic invariance [18, 19]. RGB augmentation methods cannot be generalized to RF sensing data, including radar. For example, radar data are natively associated with polar coordinates and hence are not invariant to transformations like translation and resizing. Previous work [37] on sensing the human pose found that directly applying popular SSL frameworks like [18, 29, 64] to radar heatmaps results in "shortcuts" in the learnt representation rather than capturing meaningful radar information.

We address these challenges by presenting *Radical*, a radar-based object detection system, that is fine-tuned on top of pre-trained radar embeddings to accurately estimate object bounding boxes from radar alone, e.g., during a snowstorm when vision and lidar fail. Our contributions are threefold:

- First, we propose a new contrastive learning framework using radar heatmaps and vision. It combines both cross-modal (radar-to-vision) and intra-modal (radar-to-radar) contrastive loss terms. The cross-modal term allows us to distill priors from vision such as object semantics in self-driving environments and the intra-modal term allows us to distill priors underlying radar structure such as sparsity and specularity.
- Second, we introduce a novel augmentation technique RMM (Radar MIMO Mask) that is tailored for state-of-the-art automotive radars. RMM leverages the fact that these radars use MIMO which combines multiple transmitters and multiple receivers. We manipulate how we combine the raw signals coming from different transmitter/receiver pairs to generate new augmented radar heatmaps. This augmentation preserves the underlying geometric structure of the scene while promoting resilience to the radar noise induced by Doppler phase distortions [26].
- Third, we conduct extensive evaluations and demonstrate significant improvements in radar-only 2D bounding box detection using our framework. Specifically, our results show that *Radical* improves the mean average precision (mAP) metric of car detection by 5.8% compared to supervised learning.

To the best of our knowledge, this is the first work on autonomous driving that uses self-supervised learning to take advantage of the vast amounts of unlabeled radar data and achieve 2D bounding box detection using radar only. Our findings may prove key in generating pre-trained models that avoid the need to annotate massive amounts of radar data and enable lifelong learning on new radar hardware and datasets. Finally, we will make our code publicly available once the paper is accepted.

2. Related Work

Self-supervised learning. SSL, in its contrastive and non-contrastive flavours, has by now become a staple of representation learning for computer vision tasks [13, 17, 18, 23, 24, 29, 46, 69]. At the core of vision SSL lies augmentation for synthetically generating positive views for enforcing semantic invariance. We build on two pioneering contrastive SSL methods for vision: SimCLR and MoCo [18, 20, 29]. SimCLR introduced the canonical contrastive architecture using in-batch negative sampling, which typically relies on a large batch size and associated memory. MoCo uses an efficient queue and momentum update, which decouples negative sampling from the batch size. Although we heavily draw on vision SSL, our work recasts recent advances within a new cross-modal learning objective for accurate vision-free bounding box estimation.

Cross-modal SSL. SSL's earlier NLP breakthroughs along with recent vision successes have spawned a plethora of new methods tackling representation learning under multi-modal settings [11], whereby paired positive views from other modalities replace or complement augmentation in vision SSL. Examples include vision and sound [2, 5–

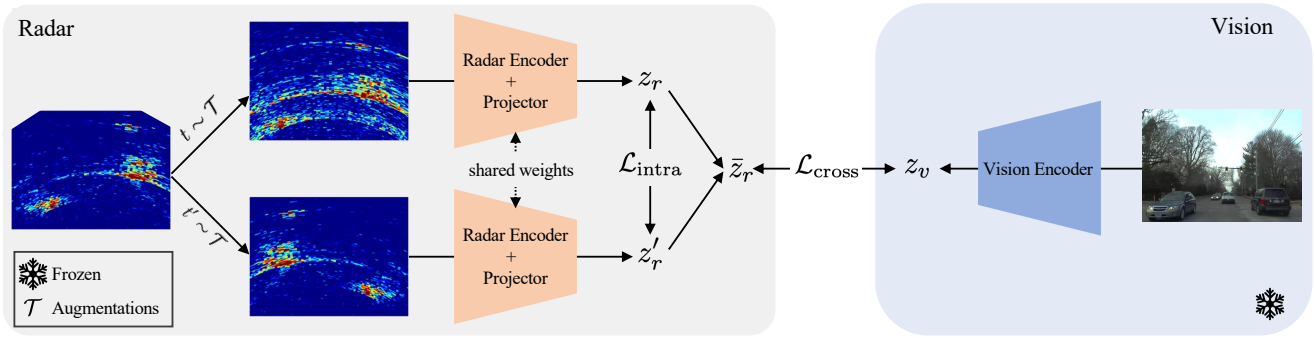

Figure 2. **Overall network of *Radical*.** Knowledge is distilled from a pretrained vision model into a radar model. A mini-batch of B radar-vision pairs flow through network, whose encodings interact locally within the radar branch and globally across the radar and vision branches. That is, *Radical* is trained using a composite contrastive loss with *intra-* and *cross-modal* terms.

9, 40, 49], vision and text [34, 52], different formats of medical imaging [62], vision and point clouds [1, 31, 66], and vision and radar [3, 4, 33, 50]. Our work expands on the early literature of radio-visual SSL, and further addresses the peculiarities of practical automotive radar, i.e., differs drastically from satellite-mounted radar for remote sensing [33, 50] while achieving accurate radio-only bounding box car detection, as opposed to simple scene classification in [4] or label-free target localization (center only) in [3].

Radio SSL. An emerging body of literature treats SSL for radio signals such as radar and WiFi [16, 37, 58, 65, 67]. Radio signals represent another SSL data domain [10] that comes with a unique set of challenges and considerations. Despite some early prior work [37, 58], there remain no mature recipes for data augmentation in the radio domain. For instance, Li et al. demonstrate that the naive application of popular contrastive learning methods to radio signals gives rise to *shortcuts* in the learned representation, and propose radio-specific transformations in mitigation [37]. Similarly, RF-URL [58] employs signal processing techniques, specific to each WiFi and radar data formats, for augmentation in order to use these radio signals within popular SSL architectures. Our cross-modal work differs from radio-only SSL literature because we also rely on vision which we argue brings robustifying and constraining priors to the much sparser radio domain. Our composite SSL loss, however, does similarly contain a radio-only term for which we devise a new augmentation scheme that we extensively characterize and benchmark.

3. Background on mmWave Radar

Millimeter-wave radars transmit FMCW (Frequency Modulated Continuous Wave) radar waveforms and receive the reflections off objects in the environment to estimate the round-trip Time-of-Flight (ToF) τ, and hence the ranges of the reflectors $\rho = \tau c/2$ (c denotes the speed of light) in the scene. Furthermore, to localize objects in the 2D range-azimuth polar coordinate (ρ, ϕ) and create a 2D bird's eye view radar heatmap, we need to use multiple receiver (RX) antennas to capture the minute ToF differences $\Delta\tau_{ij} = \tau_i - \tau_j$ between different RX. It allows us to estimate the azimuth angle (ϕ) from which the reflections arrive [32].

However, to be viable for semantic scene understanding and object detection, we must overcome the resolution limitations of radar along with a number of unique challenges. Although the wide bandwidth of mmWave radars allows us to achieve a cm-level ranging resolution, the angular resolution is bounded by the number of antenna elements and the antenna aperture size. Fortunately, the recent innovation of cascaded MIMO radars provides a much more scalable solution. It uses N TX and M RX *physical* antennas to emulate N×M *virtual* antenna links. This allows the angular resolution to scale bilinearly with the number of antennas, even though the resulting angular resolution is still no where near those of cameras and lidars.

Nevertheless, cascaded MIMO radars suffer from motion smearing in highly dynamic scenes, such as moving cars on the road, due to *Doppler-induced phase noise* [26, 39]. Consequently, radar reflections can become smeared and even appear at completely different locations. Moreover, unlike optical signals, mmWave signals are highly specular, that is, signals exhibit mirror-like reflections on cars [54]. As a result, not all reflections from the car propagate back to the mmWave receiver, and most of the car does not appear in the image, making it impossible to detect its shape [25].

Finally, radar heatmaps appear as blobs with no sharp boundaries or shapes of objects, where the voxel values represent per-voxel reflected signal energy from objects in the scene. Therefore, radar heatmaps carry little to no contextual and perceptual information and are difficult for humans to interpret and annotate.

4. Method

Our primary goal is to pretrain a radar backbone net on large-scale data in a self-supervised fashion. The learnt radar embeddings can then be employed in various down-

stream tasks. To achieve this goal, we build an SSL framework that feeds on both standalone radar and paired radar-vision data. Specifically, our *Radical* net implements a composite SSL loss with two terms: (a) intra-modal, and (b) cross-modal. The intuition is that the radar-to-radar intra-modal loss term focuses on structures specific to radar data, as we explain further in Secs. 4.2 & 4.4. The radar-to-vision cross-modal term, on the other hand, learns structures of scenes on the road where visual priors play an important role in constraining and robustifying the features of the sparser radar modality. By employing both intra-modal and cross-modal SSL, the network feeds on unlabeled radar-vision data to learn a powerful radar representation which works well on a car detection downstream task, as we demonstrate in Sec. 5. In the remainder of this section, we explain each loss term in more detail.

4.1. Distillation setup

Let $(r, v) \in \mathcal{D}$ be a radar-vision data pair in dataset $\mathcal{D}$, where $r \in \mathbb{R}^{1\times L\times A}$ is a radar heatmap and $v \in \mathbb{R}^{3\times H\times W}$ is a corresponding RGB image. Encode the radar heatmap with a backbone net f_{θ^r} then project it with an MLP head g_{ϕ^r}, assuming some weight parametrisation $\{\theta^r, \phi^r\}$, such that $z_r = g_{\phi^r}(f_{\theta^r}(r)) \in \mathbb{R}^N$. Similarly encode the paired visual image such that $z_v = f^*_{\theta^v}(v) \in \mathbb{R}^N$, with $f^*_{\theta^v}$ being a pretrained and frozen vision backbone model. Knowledge is distilled from the pretrained vision backbone $f^*_{\theta^v}$ and into the radar model f_{θ^r} by means of local interactions at the radar branch, as well as global interactions with the vision branch as depicted in Fig. 2.

4.2. Intra-modal radar learning

For radar, we aim to enrich the learnt embeddings with attributes that would enhance their discriminative power and robustness. To this end, we design a set of augmentations $\mathcal{T}$ (cf., Sec. 4.4) and formulate an intra-radar instance discrimination learning problem. Specifically as shown in the radar branch of Fig. 2, for each radar data point r, we (1) stocastically obtain two positive views of r using transformations drawn from $\mathcal{T}$, i.e., $t, t' \sim \mathcal{T}$, and (2) encode, project, and ℓ_2-normalise the positive views as $z_r = g_{\phi^r}(f_{\theta^r}(t(r)))$, $z'_r = g_{\phi^r}(f_{\theta^r}(t'(r)))$. Using a mini-batch of B samples, we then compute a contrastive loss [27, 46] for the encoded positive views of the ith sample $z_{r,i}$ and $z'_{r,i}$ against a set of negative views drawn from the mini-batch:

$$\ell_i^{r\to r'} = -\log \frac{\exp(\text{sim}(z_{r,i}, z'_{r,i}))}{\sum_{j=0}^{B} \exp(\text{sim}(z_{r,i}, z'_{r,j}))} \tag{1}$$

where $\text{sim}(x, y) := x^\top y/\tau$ is a similarity function and τ is a temperature hyper-parameter. Similarly, the encoded augmented views can be used as contrastive negatives for added efficiency, which gives us the in-batch symmetric [18] intra-radar loss function.

$$\mathcal{L}_{\text{intra}} = \frac{1}{2B} \sum_i^B (\ell_i^{r\to r'} + \ell_i^{r'\to r}) \tag{2}$$

4.3. Cross-modal radar-vision learning

As illustrated in Fig. 2, cross-modal learning uses radar and vision within a joint embedding architecture. Within this architecture, the pretrained vision model teaches the radar model how to sense and featurise the environment. Vision captures visual features from the scene in front of the vehicle. Radar data, on the other hand, is preprocessed to create 2D range-azimuth heatmaps, which represent the scene from a BEV perspective. While radar and vision operate within these different coordinate systems, their embeddings are nonetheless *aligned* via the contrastive loss.

To implement cross-modal learning, we obtain a prototype radar vector as an average of the two positive vectors $\bar{z}_r = (z_r + z'_r)/2$ following [1]. We encode and normalize the corresponding vision sample $z_v = f^*_{\theta^v}(v)$. We found it empirically beneficial to omit the MLP projector head from the frozen vision branch while keeping a projector after the radar encoder.

Similar to the radar-to-radar contrastive learning term in Eq. 1, we then compute the term $\ell_i^{\bar{r}\to v}$, where the use of the prototype $\bar{z}_r$ in radar-to-vision contrastive term is denoted by $\bar{r}$. The in-batch cross-modal contrastive loss is then given by

$$\mathcal{L}_{\text{cross}} = \frac{1}{B} \sum_i^B \ell_i^{\bar{r}\to v} \tag{3}$$

With the intra-modal and cross-modal losses defined in Eqs. 2 & 3, the overall composite loss is

$$\mathcal{L} = \mathcal{L}_{\text{intra}} + \lambda_{\text{cross}} \mathcal{L}_{\text{cross}} \tag{4}$$

where λ_{cross} is a hyper-parameter.

4.4. Augmentations

A suite of augmentations is essential to our *Radical* framework. We next treat these augmentations, as used in both intra- and cross-modal learning. We extensively compare and ablate their effectiveness in Sec. 5. Fig. 3 gives a visual intuition for all the augmentations we utilize in *Radical*.

4.4.1 Repurposed vision augmentations

Considering that BEV radar heatmaps have formats similar to camera images, a subset of standard SSL vision augmentations is potentially applicable to radar heatmaps. However, due to the different perspectives and coordinate system, most vision augmentations are not applicable or need to be carefully modified.

We conduct extensive experiments on different vision augmentations and their combinations (cf., Sec. 6). We find

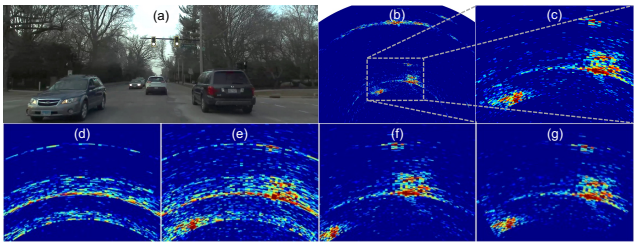

Figure 3. **Radar-specific augmentations.** (a) Scene. (b) Original radar heatmap. (c) Zoomed-in region of cars. (d) Random Phase. (e) Antenna Dropout. (f) Rotation (Polar). (g) Center Cropping (Polar).

that horizontal flip, rotation, and center cropping [18] are also suitable for radar heatmaps. We note that for radar heatmaps whose coordinates are polar, rotation and center cropping should be applied in the polar coordinates, as shown in Fig. 3(f) and (g) respectively.

4.4.2 Radar-specific augmentations

In addition to the repurposed subset of vision augmentations, we introduce and experiment with a new domain-specific augmentation for radar SSL we call Radio MIMO Mask (RMM). We briefly explain how the raw data is processed before RMM is applied.

RMM implementation. Several radar formats typically appear in related work: range-azimuth heatmaps, point clouds, or range-Doppler maps [39, 51, 55]. Differently, *Radical* uses an intermediate 3-D tensor in order to apply RMM augmentations. Specifically, consider a MIMO radar with M transmitters and N receivers. A range-azimuth heatmap $r(\rho, \phi) \in \mathbb{R}^{L \times A}$ is generated when a preceding 3-D complex tensor $S \in \mathbb{C}^{MN \times L \times A}$ is integrated noncoherently over all the antenna pairs (at the first index). RMM is applied before this integration. RMM is best presented as the composition of two operations: (1) antenna dropout, and (2) random phase noise. We further explain these below.

(1) Antenna Dropout. We leverage the reconfigurability of the virtual array emulated by the MIMO radar to design this radar-specific augmentation. We omit randomly a subset of virtual antenna elements from subsequent signal aggregation. Mathematically, we can write

$$r'(\rho, \phi) = \left| \sum_{k=1}^{MN} b_k S(\rho, \phi, k) \right|, \quad b_k \sim \text{Bernoulli}(p)$$

where $r'(\rho, \phi)$ is the augmented radar heatmap as a function of range ρ and azimuth angle ϕ, k indexes the set of $M \times N$ antenna pairs, and b_k are independent discrete random masks that nullifies the kth antenna pair with probability p. This augmentation simulates scenarios with partial sensor failure or obstructions, which promotes learning from incomplete data and improves robustness. The probability of antenna dropout $p \in [0, 1]$ is a tunable hyper-parameter.

(2) Random Phase Noise. This augmentation randomizes the phase of the received (complex) signals before their aggregation. Mathematically, we can describe this phase randomization as

$$S'_k = S_k \cdot e^{i\theta_k}, \quad \theta_k \sim U[-\alpha\pi, \alpha\pi), \quad 1 \leq k \leq MN$$

where S_k is the signal from the kth transmitter and receiver pair, S'_k is the augmented signal, and θ_k are i.i.d. phase shifts drawn from uniform distributions $\in [-\alpha\pi, \alpha\pi]$ (in radians), where $\alpha \in [0, 1)$ is a tunable hyper-parameter. This randomization mimics the phase variability introduced by environmental factors and relative motions between the radar and the scene, which is also referred to as *Doppler-induced phase noise* [26, 39]. Thus it enhances the training coverage of RF conditions likely to occur in the real-world. We note that larger α corresponds to more aggressive movements and noise in the environment.

RMM instantiation. The final RMM augmentation is the (order-invariant) composition of the two operations detailed above. We found empirically that the hyper-parameters $p = 0.9$ and $\alpha = 0.1$ lead to the best performance in our experiments (see Sec. 6.2).

4.5. Downstream fine-tuning

After pre-training, we discard the projector head and use the radar backbone only to perform downstream tasks. We fine-tune the radar backbone with a task-specific head on top. Specifically, we demonstrate *Radical* on the challenging task of bounding box detection for cars using standalone radar heatmaps. This task showcases the practical utility of our pre-training towards extending current self-driving perception stacks with weather-immune, fine-grained radar capabilities.

4.6. Implementation details

For the radar backbone, we use *Radatron* [39], which adopts an FPN-based architecture. The backbone has a two-stream architecture, which takes as inputs high- and low-resolution radar heatmaps. Though generated from the cascade MIMO capture of the same scene, the high- and low-res streams undergo different radar signal preprocessing. Specifically, each stream goes through a stem layer and then two ResNet stages, which are identical to the building blocks of ResNet50 [28]. Then the two streams are concatenated and fused in a convolutional layer. The resultant feature maps are further encoded via additional ResNet stages, and combined to create the features similar to [63]. We pre-train the backbone of the model (without the FPN and the linear regression heads) as the radar feature extractor.

Method	mAP	AP_{75}	AP_{50}
Radatron [39]	56.5 ± 0.2	64.5 ± 1.7	88.9 ± 0.4
Intra-modal	59.4 ± 1.0	66.8 ± 1.8	89.1 ± 0.5
Cross-modal	59.9 ± 0.6	66.9 ± 0.9	88.9 ± 0.2
Radical (ours)	62.3 ± 0.6	69.7 ± 1.2	89.6 ± 0.1

Table 1. **Performance of downstream bounding box detection against baselines.** Best performing model is highlighted.

Method	mAP	AP_{75}	AP_{50}
Radatron [39]	22.1 ± 0.8	17.7 ± 1.2	48.0 ± 1.5
Intra-modal	45.4 ± 0.1	48.6 ± 0.5	78.3 ± 0.1
Cross-modal	46.3 ± 0.1	49.4 ± 0.3	83.0 ± 0.1
Radical (ours)	52.6 ± 0.1	58.5 ± 0.2	86.9 ± 0.1

Table 2. **Performance of downstream bounding box detection with frozen backbone in fine-tuning.** Best performing model is highlighted. Results are averaged over 2 runs.

For the vision branch, we use a pre-trained CLIP image encoder model [52], which we freeze throughout pre-training.

5. Experiments and Evaluation

5.1. Dataset

We evaluate *Radical* on few publicly available datasets that support raw radar data format. This is because our domain-specific augmentations require raw radar format. We focus here on using Radatron dataset [39], with a total of 152K radar frames and 16K labeled frames. In addition to the requisite raw radar format, we find Radatron's size (both labeled and unlabeled frames) beneficial in the characterisation we present herein. Out of the unlabeled set, we use 63K frames for self-supervised pretraining, 13K annotated frames for supervised fine-tuning, and 3K annotated frames for testing. The train and test splits are constructed from experiments conducted on different days throughout the data collection campaign. The raw radar frames are first converted to complex, 86-channel heatmaps. These heatmaps are then fed to the network for preprocessing and stochastic augmentation.

5.2. Experiments

We pre-train *Radical* as depicted in Fig. 2 and detailed in Secs. 4. We utilize unlabeled radar-vision frames from *Radatron* as described in Sec. 5.1. We specialize the pre-trained radar embeddings for a downstream task relevant to self-driving. The task strives to detect rotated 2D bounding boxes in BEV from radar heatmaps.

During pre-training, we use a batch size of 128, learning rate of 0.05 and cosine learning rate scheduling with an SGD optimizer with momentum 0.9 and weight decay 0.0001. During fine-tuning, we adopt the same training setting as Radatron, using a batch size of 8, an SGD optimizer with learning rate of 0.01 and 25K iterations with learning rate drop at 15K and 20K iterations. We increase the weight decay to 0.001 in order to avoid overfitting problems, boosting the baseline performance.

Unless otherwise stated, our results are obtained with pre-training the backbone using 63K unlabeled frames, and fine-tuning the downstream model on 13K labeled frames. Results are averaged over 6 runs.

5.3. Baselines

We evaluate against supervised learning as well as different variants of self-supervised learning in order to expose the merit of our design choices. We denote contrastive learning by CL below.

(1) Radatron. We compare against the original implementation reported in [39] based on supervised learning.
(2) Intra-modal CL. We disable vision from contributing to the composite contrastive loss by setting $\lambda_{\text{cross}} = 0$ in Eq. 4, which results in intra-modal, radar-only CL. For this, we use the vision-based augmentations of vertical flipping and center cropping.
(3) Cross-modal CL. We disable intra-modal CL and its radar-specific augmentations, reverting to a CL configuration that is wholly reliant on cross-modal learning between radar and vision. We extend the implementations of SimCLR [18] and MoCo [29] for our cross-modal settings.

6. Results

This section presents a comprehensive analysis of *Radical*'s performance against baselines and examines the impact of various augmentations on model performance.

Evaluation metrics. Following previous radar detection work [39, 51], we use Average Precision (AP) with IoU thresholds of 0.5, and 0.75 to evaluate *Radical*'s detection performance. We also use the mean AP (mAP) of IoU values from 0.5 to 0.95 in 0.05 steps. We follow the COCO framework [38] for evaluating *Radical*.

6.1. Performance vs. baselines

We characterize *Radical*'s performance against the baselines enumerated in Sec. 5.3 and on the downstream task discussed therein. Specifically, we analyze performance by means of: (a) fine-tuning the backbone along with the task-specific head, and (b) freezing the backbone and training the task-specific head only.

Fine-tuning backbone. We pre-train *Radical* utilising our composite intra- and cross-modal CL configuration, along with the two baseline CL configurations from Sec. 5.3. We then fine-tune all pre-trained backbones along with their bounding box estimation heads. We compare these pre-trained variants to the implementation of [39] which uses random initialization. Table 1 shows the quantitative re-

Labeled data fraction	100%	30%	10%	3%	1%
Radatron [39]	56.5	51.3	44.6	38.2	32.5
Radical (ours)	62.3	54.6	48.6	44.7	40.3
Gain	+5.8	+3.3	+4.0	+6.5	+7.8

Table 3. **Label efficiency for fine-tuning.** We use all unlabeled data for self-supervised pre-training, and vary size of labeled data for fine-tuning. We use mAP as our metric.

Method	mAP	AP_{75}	AP_{50}
RMM	61.2 ± 0.5	68.1 ± 0.5	89.4 ± 0.4
Rotation	61.4 ± 0.6	68.7 ± 1.0	89.0 ± 0.6
Center Cropping	61.0 ± 1.0	67.7 ± 1.4	89.2 ± 0.7
Horizontal Flip	59.6 ± 0.8	67.5 ± 1.3	89.0 ± 0.3
Cross-modal (No Aug.)	59.9 ± 0.6	66.9 ± 0.9	88.9 ± 0.2
Threshold	59.5 ± 0.6	66.4 ± 1.1	88.7 ± 0.5
Cutout	58.4 ± 1.2	66.8 ± 1.2	89.0 ± 0.5
Vertical Flip	58.1 ± 0.6	66.6 ± 0.9	88.5 ± 0.3

Table 4. **Effect of adding one augmentation at a time to the base *Radical* net.**

sults using three metrics: mAP, AP_{50}, and AP_{75}. The mean and standard deviation of these results are obtained from 6 different runs of supervised training, while keeping the pre-trained weights the same. We see that *Radical*'s composite intra- and cross-modal configuration performs most favourably, and outperforms random initialization by 5.8% in mAP. This demonstrates the efficacy of *Radical*'s pre-training on this highly relevant downstream task. *Radical* also outperforms intra-modal CL and cross-modal CL baselines by 2.9% and 2.4% respectively. Despite good gains over random initialisation (approx. 3% each), the two CL baselines are unable to approach the performance of *Radical*'s composite CL loss.

Freezing backbone. We freeze the pre-trained weights in order to assess and compare the quality of the learnt features across our contrastive configurations. To this end, we train task-specific heads for our downstream bounding box estimation task similar to above. For Radatron, we randomly initialize its backbone and then similarly freeze it. The averages and standard deviations are listed in Table 2. We observe that *Radical* outperforms all baselines on all metrics. Random initialisation performs poorly compared to the pre-trained variants. This highlights the inadequacy of the task-specific head to perform accurate bounding box estimation without quality featurisation underneath. We also observe that the gap between *Radical* and the two CL baselines widens compared to Table 1. Without fine-tuning to compensate, this further underscores the efficacy of *Radical* compared to the CL baselines. We also observe a slight performance advantage to cross-modal over the intra-modal CL. This could point to the importance of visual priors in the training of quality radar embeddings.

Label efficiency. We investigate the impact of the number of available labeled data on performance under a fine-tuning protocol. We compare *Radical* to the random initialization of Radatron. For *Radical*, we use all unlabeled data for pre-training. Table 3 shows the mAP after full fine-tuning as a function of the fraction of labeled data used. We see consistent improvements using *Radical* pre-training over the supervised baseline across all label density regimes.

RMM	CC	HF	ROT	mAP	AP_{75}	AP_{50}
✓	✓	✓	✓	61.6 ± 0.7	68.9 ± 1.1	89.5 ± 0.5
✗	✓	✓	✓	61.3 ± 0.8	68.7 ± 1.3	89.4 ± 0.2
✓	✗	✓	✓	61.0 ± 0.8	69.1 ± 1.1	89.5 ± 0.5
✓	✓	✗	✓	61.7 ± 0.2	68.9 ± 0.3	89.8 ± 0.5
✓	✓	✓	✗	62.3 ± 0.6	69.7 ± 1.2	89.6 ± 0.1

Table 5. **Effect of removing each augmentation individually from the four best augmentations found in Table. 4.**

p	α	mAP	AP_{75}	AP_{50}
No Aug.	No Aug.	59.9 ± 0.6	66.9 ± 0.9	88.9 ± 0.2
1.0	0.3	61.0 ± 0.6	67.8 ± 1.2	89.3 ± 0.1
0.9	0.1	61.2 ± 0.5	68.1 ± 0.5	89.4 ± 0.4
0.9	0.3	60.7 ± 0.5	66.8 ± 0.9	89.0 ± 0.5
0.9	0.5	60.0 ± 0.7	67.3 ± 0.8	89.1 ± 0.2
0.8	0.3	60.3 ± 0.5	67.3 ± 0.9	88.9 ± 0.8

Table 6. **Hyper-parameters of the RMM augmentation.**

6.2. Ablating augmentations

To better understand the value of each augmentation, we dissect the contribution of individual repurposed and radar-specific augmentations, as well as effect of removing the augmentations from the best combinations.

Individual augmentations. We begin by comparing the effect of adding the individual augmentations enumerated in Sec. 4.4. We also list three other augmentations that we experimented with but did not yield beneficial results. They include two standard SSL vision augmentations: Cutout and Vertical Flip [18]. We also tested another radar-specific augmentation, Thresholding, whereby we created a binary mask by setting a power (pixel-value) threshold for the radar heatmap.

In our experiments, we pre-train the *Radical* net using one augmentation at a time, and fine-tune it with the 13K labeled dataset. The results are shown in Table 4. As a baseline for comparison, we also include the cross-modal only baseline in the table which does not use any augmentations. As seen, four out of the seven tested augmentations, namely rotation, RMM, center crop, and horizontal flip prove beneficial for pre-training. On the other hand, thresholding, cutout, and vertical flipping will on average be detrimental to performance across all three listed metrics. We make the following points. First, based on these results, we removed the three worst-performing augmentations from our final model. Second, we make the following observations regarding the effectiveness of each augmentation:

1. While radar heatmaps are symmetric along the mid-point of the azimuth axis, they are certainly not so along the range axis. This is why horizontal flip retains the underlying structure of the radar data, while vertical flip fails to do so and hurts performance.
2. While thresholding might seem an intuitive extension to

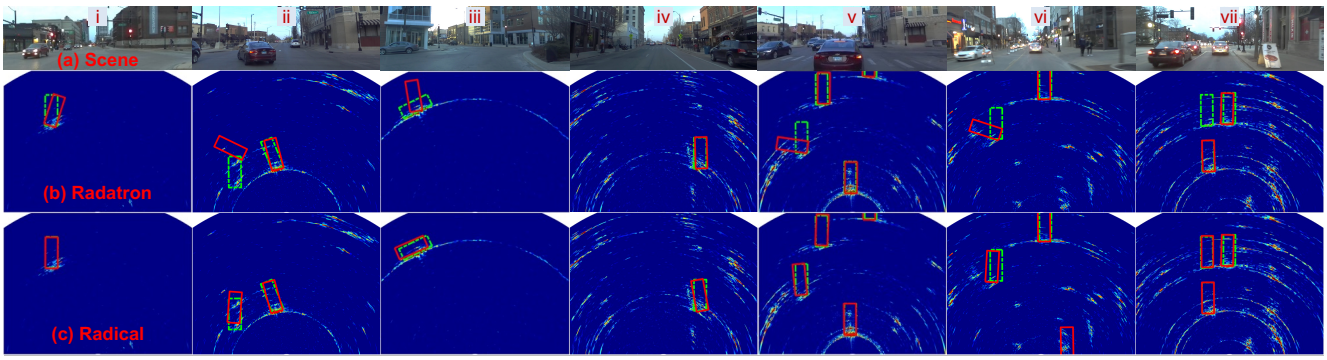

Figure 4. **Examples from our test set:** (a) Original scene. (b) *Radatron* (supervised) baseline. (c) *Radical*. Groundtruth marked in green and predictions in red.

similar quantization methods in vision, it fails to aid performance in radar data because radar data is already extremely sparse in nature.

3. Center cropping and rotation borrowed from vision seem to boost *Radical* performance. They preserve the underlying semantics of radar heatmaps.
4. RMM is a useful MIMO radar-specific augmentation for promoting robustness to Doppler-induced phase noise in MIMO radars.

Combining augmentations. Having found four individually useful augmentation in Table. 4, we next explore how to best combine them. To this end, we conduct five experiments; the first uses all four augmentations, while each subsequent experiments removes one out of four augmentations at a time. The results are shown in Table. 5. We note that all these combinations perform seemingly equally well under the AP_{50} metric. However, metrics mAP and AP_{75} reveal that the combination RMM + Center Crop + Horizontal Flip is the clear winner. We hence use it in *Radical*'s final model.

Hyper-parameters of RMM augmentation. Having just introduced RMM augmentation in this work, we next explore the configuration space of its hyper-parameters in order to identify an initial performant recipe. Table 6 sweeps p and α (cf. Sec. 4.4). To establish a comparative baseline, the first row of Table 6 shows our three AP metrics without using RMM. We observe that keeping virtual antennas with probability $p = 0.9$ and noise randomisation with $\alpha = 0.1$ results in best performance across the three metrics. This amounts to randomly omitting 10% of the antenna pairs in MIMO. While sizeable, we view this antenna masking as non-aggressive and preserving of the integrity of the radar data.

6.3. Qualitative results

We next present qualitative results and compare *Radical* to the supervised baseline Radatron. Fig. 4 shows groundtruth as dotted green bounding boxes, and model predictions in solid red. Fig. 4 consists of three rows: (1) upper row depicts front-view camera images, (2) middle row depicts Radatron's bounding box predictions overlaid on top of groundtruth in BEV, and (3) bottom row depicts *Radical*'s bounding box predictions and groundtruth. We make the following observations. First, quite a few of the failure cases—namely columns i, ii, iii, v, and vi in Fig. 4—in the baseline arise from scenarios where a car is detected, albeit its orientation and exact bounding box are missed. This is due to the low-resolution and specular nature of radar. These failures are mostly rectified by *Radical*'s network as seen in the bottom row. Second, *Radical* performs better in scenarios where a car's radar reflection might get occluded by other cars in the scene, as in Fig. 4(vii). Both these failure cases are well known in radar object detection systems, as shown by previous work [25, 39]. *Radical* overcomes these failures thanks to pre-training radars to learn radar priors like specularity and sparsity jointly with vision features, which additionally carry semantic information such as precise car location and orientations. Finally, we note that Radatron performs reasonably closer compared to *Radical* when detecting the approximate location of vehicles which is reflective of their relatively closer AP_{50} performance compared to mAP and AP_{75}. In other words, *Radical*'s strength, as noted in Sec. 5.3, lies in its more precise box detection in complex situations, illustrated in Fig.4.

7. Conclusion

In this paper, we presented a self-supervised approach to radar object detection in the context of self-driving cars, harnessing the largely untapped potential of vast quantities of unlabeled radar data. Our extensive evaluations illustrate that *Radical* achieves superior performance over supervised baselines by effectively combining intra- and cross-modal self-supervised learning, and employing radar-specific as well as vision-inspired augmentation in the context of contrastive learning. It is our hope that these contributions are followed by future advancements in the field of automative radar.

References

[1] M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, and R. Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9892–9902, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 3, 4

[2] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020. 2

[3] M. Alloulah and M. Arnold. Look, radiate, and learn: Self-supervised localisation via radio-visual correspondence. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17430–17440, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 3

[4] Mohammed Alloulah, Akash Deep Singh, and Maximilian Arnold. Self-supervised radio-visual representation learning for 6g sensing. In *ICC 2022-IEEE International Conference on Communications*, pages 1955–1961. IEEE, 2022. 2, 3

[5] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020. 2

[6] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[7] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision*, pages 435–451, 2018.

[8] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[9] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 3

[10] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023. 3

[11] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 2

[12] Kshitiz Bansal, Keshav Rungta, Siyuan Zhu, and Dinesh Bharadia. Pointillism: Accurate 3d bounding box estimation with multi-radars. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, page 340–353, New York, NY, USA, 2020. Association for Computing Machinery. 1

[13] Adrien Bardes, Jean Ponce, and Yann LeCun. Variance-invariance-covariance regularization for self-supervised learning. *ICLR, Vicreg*, 2022. 2

[14] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6433–6438. IEEE, 2020. 1

[15] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[16] Zhongping Cao, Zhenchang Li, Xuemei Guo, and Guoli Wang. Towards cross-environment human activity recognition based on radar without source data. *IEEE Transactions on Vehicular Technology*, 70(11):11843–11854, 2021. 3

[17] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 4, 5, 6, 7

[19] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 2

[20] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[22] Fangqiang Ding, Andras Palffy, Dariu M. Gavrila, and Chris Xiaoxuan Lu. Hidden gems: 4d radar scene flow learning using cross-modal supervision. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9340–9349, 2023. 1, 2

[23] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022. 2

[24] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[25] Junfeng Guan, Sohrab Madani, Suraj Jog, Saurabh Gupta, and Haitham Hassanieh. Through fog high-resolution imaging using millimeter wave radar. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11464–11473, 2020. 1, 3, 8

[26] Junfeng Guan, Sohrab Madani, Waleed Ahmed, Samah Hussein, Saurabh Gupta, and Haitham Hassanieh. Exploiting virtual array diversity for accurate radar detection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2, 3, 5

[27] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, pages 1735–1742. IEEE, 2006. 4

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 6

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[31] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021. 3

[32] Cesar Iovescu and Sandeep Rao. The fundamentals of millimeter wave sensors. *Texas Instruments*, pages 1–8, 2017. 3

[33] Umangi Jain, Alex Wilson, and Varun Gulshan. Multimodal contrastive learning for remote sensing tasks. *arXiv preprint arXiv:2209.02329*, 2022. 3

[34] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[35] Prannay Kaul, Daniele De Martini, Matthew Gadd, and Paul Newman. Rss-net: Weakly-supervised multi-class semantic segmentation with fmcw radar. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 431–436. IEEE, 2020. 1, 2

[36] Pou-Chun Kung, Chieh-Chih Wang, and Wen-Chieh Lin. Radar occupancy prediction with lidar supervision while preserving long-range sensing and penetrating capabilities. *IEEE Robotics and Automation Letters*, 7(2):2637–2643, 2022. 1, 2

[37] Tianhong Li, Lijie Fan, Yuan Yuan, and Dina Katabi. Unsupervised learning for human sensing using radio signals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3288–3297, 2022. 2, 3

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[39] Sohrab Madani, Junfeng Guan, Waleed Ahmed, Saurabh Gupta, and Haitham Hassanieh. Radatron: Accurate detection using multi-resolution cascaded mimo radar. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, page 160–178, 2022. 1, 3, 5, 6, 7, 8

[40] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33:4733–4744, 2020. 3

[41] Mohammadreza Mostajabi, Ching Ming Wang, Darsh Ranjan, and Gilbert Hsyu. High-resolution radar dataset for semi-supervised learning of dynamic objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 100–101, 2020. 1

[42] Mohammadreza Mostajabi, Ching Ming Wang, Darsh Ranjan, and Gilbert Hsyu. High resolution radar dataset for semi-supervised learning of dynamic objects. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 450–457, 2020. 1

[43] Fatemeh Norouzian, Emidio Marchetti, Edward Hoare, Marina Gashinova, Costas Constantinou, Peter Gardner, and Mikhail Cherniakov. Experimental study on low-thz automotive radar signal attenuation during snowfall. *IET Radar, Sonar & Navigation*, 13(9):1421–1427, 2019. 1

[44] Fatemeh Norouzian, Emidio Marchetti, Marina Gashinova, Edward Hoare, Costas Constantinou, Peter Gardner, and Mikhail Cherniakov. Rain attenuation at millimeter wave and low-thz frequencies. *IEEE Transactions on Antennas and Propagation*, 68(1):421–431, 2020. 1

[45] Farzan Erlik Nowruzi, Dhanvin Kolhatkar, Prince Kapoor, Fahed Al Hassanat, Elnaz Jahani Heravi, Robert Laganiere, Julien Rebut, and Waqas Malik. Deep open space segmentation using automotive radar. In *2020 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, pages 1–4. IEEE, 2020. 1

[46] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 4

[47] Itai Orr, Moshik Cohen, and Zeev Zalevsky. High-resolution radar road segmentation using weakly supervised learning. *Nature Machine Intelligence*, 3(3):239–246, 2021. 1, 2

[48] Arthur Ouaknine, Alasdair Newson, Julien Rebut, Florence Tupin, and Patrick Pérez. Carrada dataset: camera and automotive radar with range-angle-doppler annotations. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5068–5075. IEEE, 2021. 1

[49] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016. 3

[50] Jonathan Prexl and Michael Schmitt. Multi-modal multi-objective contrastive learning for sentinel-1/2 imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2135–2143, 2023. 3

[51] Kun Qian, Shilin Zhu, Xinyu Zhang, and Li Erran Li. 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*, pages 444–453, 2021. 5, 6

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[53] Julien Rebut, Arthur Ouaknine, Waqas Malik, and Patrick Pérez. Raw high-definition radar for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17021–17030, 2022. 1

[54] Giulio Reina, James Underwood, Graham Brooker, and Hugh Durrant-Whyte. Radar-based perception for autonomous outdoor vehicles. *Journal of Field Robotics*, 28 (6):894–913, 2011. 3

[55] Ole Schumann, Markus Hahn, Jürgen Dickmann, and Christian Wöhler. Semantic segmentation on radar point clouds. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 2179–2186. IEEE, 2018. 5

[56] Marcel Sheeny, Emanuele De Pellegrin, Saptarshi Mukherjee, Alireza Ahrabian, Sen Wang, and Andrew Wallace. Radiate: A radar dataset for automotive perception. *arXiv preprint arXiv:2010.09076*, 3(4):7, 2020. 1

[57] L. Sless, B. Shlomo, G. Cohen, and S. Oron. Road scene understanding by occupancy grid learning from sparse radar clusters using semantic segmentation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 867–875, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 1, 2

[58] Ruiyuan Song, Dongheng Zhang, Zhi Wu, Cong Yu, Chunyang Xie, Shuai Yang, Yang Hu, and Yan Chen. Rf-url: unsupervised representation learning for rf sensing. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, pages 282–295, 2022. 3

[59] Yizhou Wang, Zhongyu Jiang, Xiangyu Gao, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. Rodnet: Radar object detection using cross-modal supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 504–513, 2021. 1, 2

[60] Yizhou Wang, Gaoang Wang, Hung-Min Hsu, Hui Liu, and Jenq-Neng Hwang. Rethinking of radar's role: A camera-radar dataset and systematic annotator via coordinate alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2815–2824, 2021. 1

[61] Rob Weston, Sarah Cen, Paul Newman, and Ingmar Posner. Probably unknown: Deep inverse sensor modelling radar. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5446–5452, 2019. 1, 2

[62] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Self-supervised multi-modal alignment for whole body medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 90–101. Springer, 2021. 3

[63] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[64] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[65] Yashan Xiang, Jian Guo, Ming Chen, Zheyu Wang, and Chong Han. Mae-based self-supervised pretraining algorithm for heart rate estimation of radar signals. *Sensors*, 23 (18), 2023. 3

[66] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020. 3

[67] Yang Yang, Xiaoyi Yang, Takuya Sakamoto, Francesco Fioranelli, Beichen Li, and Yue Lang. Unsupervised domain adaptation for disguised-gait-based person identification on micro-doppler signatures. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6448–6460, 2022. 3

[68] Shizhe Zang, Ming Ding, David Smith, Paul Tyler, Thierry Rakotoarivelo, and Mohamed Ali 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*, 14(2):103–111, 2019. 1

[69] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2

[70] Ao Zhang, Farzan Erlik Nowruzi, and Robert Laganiere. Raddet: Range-azimuth-doppler based radar object detection for dynamic road users. *arXiv preprint arXiv:2105.00363*, 2021. 1