Look, Radiate, and Learn: Self-Supervised Localisation via Radio-Visual Correspondence

Mohammed Alloulah* Maximilian Arnold†
Nokia Bell Labs

Abstract

Next generation cellular networks will implement radio sensing functions alongside customary communications, thereby enabling unprecedented worldwide sensing coverage outdoors. Deep learning has revolutionised computer vision but has had limited application to radio perception tasks, in part due to lack of systematic datasets and benchmarks dedicated to the study of the performance and promise of radio sensing. To address this gap, we present MaxRay: a synthetic radio-visual dataset and benchmark that facilitate precise target localisation in radio. We further propose to learn to localise targets in radio without supervision by extracting self-coordinates from radio-visual correspondence. We use such self-supervised coordinates to train a radio localiser network. We characterise our performance against a number of state-of-the-art baselines. Our results indicate that accurate radio target localisation can be automatically learned from paired radio-visual data without labels, which is important for empirical data. This opens the door for vast data scalability and may prove key to realising the promise of robust radio sensing atop a unified communication-perception cellular infrastructure. Dataset will be hosted on IEEE DataPort.

1. Introduction

Sixth-generation (6G) wireless networks are being designed from the ground up to support sensing at the physical layer [79]. Such a brand new capability in 6G networks marks a departure from communication-only functions, and aims to supply applications with sensing primitives atop a unified communication-perception infrastructure. Concretely, dense cellular deployments in urban settings (e.g., per lamppost) would allow for unprecedented radio coverage, enabling a multitude of challenging perception tasks. Examples include around-the-corner obstacle detection in support of autonomous driving and pedestrian and drone localisation, to name a few [2].

*Correspondence to alloulah@outlook.com
†Work done whilst at Bell Labs.

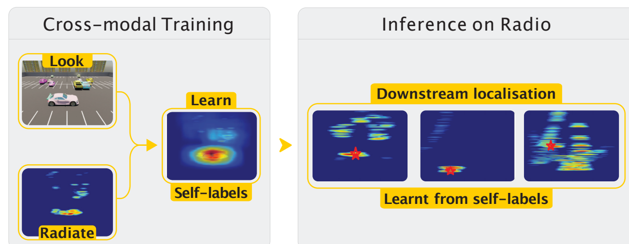

Figure 1. We train a radio localisation network by using commonalities with vision to drive spatial attention. Without laborious manual annotations, we learn to suppress clutter and localise targets in radio heatmaps.

Training perception models for radio signals is a key challenge for network infrastructure vendors. Unlike vision and audio, radio signals are hard to label manually because they are not human interpretable. Typically, sparse radio signals have been paired with a groundtruth vision modality for reliable semantic and qualitative filtration via a cross-modal annotation flow [34, 56, 77, 85]. Recently, this radio-visual pairing has been shown to work in a self-supervised fashion [10], building on a wave of progress in vision self-supervised learning (SSL) [19,22,26,40–42,45,61,62,82,84].

Computer vision has traditionally benefited from synthetic datasets for: (a) content augmentation for enhanced generalisability [54, 81], or (b) closing the learning loop on out-of-distribution failure modes [68], e.g., in the context of autonomous driving [1]. Extrapolating from vision, it is also likely that synthetic data will play an important role towards realising robust radio sensing. However, radio perception tasks have yet to benefit from such publicly available datasets.

In this work we aim to support next-gen 6G perception tasks, while championing a self-supervised radio-visual learning approach. Concretely, Fig. 1 captures the crux of our new machine learning proposition for radio sensing. We demonstrate how to automatically extract radio self-labels through cross-modal learning with vision. We then use such self-labels to train a downstream localiser network. We show that our self-supervised localiser net enhances estimation in the radio domain compared to state-of-the-art. Our contributions are:

- A synthetic dataset: We curate and synthesise radio-visual data for a new learning task designed for target detection and localisation in radio.
- A cross-modal SSL algorithm: We formulate a contrastive radio-visual objective for label-free radio localisation.
- Evaluation: We conduct numerous characterisations on synthetic and empirical data in order to validate our SSL algorithm and expose its superior performance compared to state-of-the-art.

We discuss our dataset and algorithmic findings to galvanise machine learners' interest in radio-visual learning research. We hope to both facilitate and inform future research on this new cross-modal learning paradigm.

2. Related Work

Self-supervised learning. Self-supervised learning (SSL) in its two strands (contrastive and non-contrastive) is the state-of-the-art learning paradigm for visual representations [26,42,45,82]. SSL models have progressively matched and then exceeded the performance of their fully-supervised counterparts [17, 19, 22, 40, 47, 50, 61, 62, 84], culminating recently in strong performance on uncurated billion-scale data [41]. Vision SSL relies on augmentation for semantic invariance. Differently, we deal with a new radio-visual SSL problem that relies on cross-modal correspondence [12, 13] as opposed to augmentation. Further, our work addresses SSL object detection and localisation using spatial backbone models [3, 23] rather than the prevalent object classification in vision using 1-D backbones.

Self-supervised multi-modal object detection. A related body of work leverages multiple modalities for representation learning, particularly between audio and vision [5, 11–13, 15, 16, 24, 60, 65]. Other works, also audio-visual, deal with knowledge distillation from one modality to another [4, 37]. SSL audio-visual object detectors are well researched and rely on feature attention between 1-D audio and 2-D vision [3, 5]. Differently in radio-visual, our attention (a) is complicated by a sparse radio modality which could impact the dimensional stability of cross-modal contrastive learning [51], and (b) involves a fundamentally larger feature search space between 2-D radio and 2-D vision.

Self-supervised saliency localisation. Recent works have extended visual saliency localisation [70, 86] for self-supervised systems [18]. Specifically, [59] expands class activation map (CAM) to work within an SSL network to markedly improve visual contrastive learning and mitigate against augmentation bias. While notable for vision SSL, radio-visual SSL does not suffer from the augmentation-induced geometric perturbations during training (e.g., random crop and rotation) which make accurate object localisation trickier in vision SSL.

Self-supervision with priors. Some works bake prior information back into SSL, e.g., using off-the-shelf image segmentation models [47, 48, 76]. Follow-up works replace these priors with online learning that works hand in hand with SSL [25, 49]. Our work uses priors from vision to bootstrap radio-visual SSL in a relatively small data regime. Similarly, however, radio-visual SSL could be made to work without vision priors in principle.

Radio learning. Recent works train radio models on vision-supplied labels for indoor and outdoor sensing, e.g., [34, 43, 56, 85]. SSL has also been recently applied to radio-only learning systems. [63] proposes an SSL super-resolution method that improves the angular resolution of radar antenna arrays. [38] uses radar during training as a weak supervision signal, as well as an extra input to enhance depth estimation at inference time. [55] tackles the problem of radio-only SSL for human sensing. Our work is different from the above prior art in that it neither relies on explicit supervision from vision, nor it is single-modal for radio-only learning. A recent work proposes radio-visual SSL for object classification within a distillation framework [10]. This differs from our work which (a) deals with representation learning from scratch for both radio and vision and (b) is aimed at SSL object detection and localisation using an underlying spatial backbone as opposed to standard classification.

Radio-visual datasets. A number of multi-modal datasets (with radio-visual entries) are available in the adjacent automotive literature, where it is not uncommon for datasets that are collected using fleets of cars to be quite large. Examples include CRUW [77], Carrada [64], AIODrive [78], RADIATE [72], Oxford Radar RobotCar [20]. 6G networks, on the other hand, focus on radio-visual data collected at a stationary basestation for sensing the surrounding environment. RADDet [83] and DeepSense [7] are closely related datasets. However, both are empirical datasets with low angular resolution. In contrast, our synthetic dataset has higher angular resolution and incorporates high-fidelity propagation modelling[1] and graphical rendering, which result in quality radio-visual data. This allows for much tighter characterisation and refinement of algorithms given the (a) controllability (e.g., configurability w.r.t. radio parameters, cf. Tab. 3) and (b) measurability against perfect groundtruth.

Table 1. Radio-visual datasets.†

Dataset	Automotive	6G
CRUW [77]	✓	✗
Carrada [64]	✓	✗
AIODrive [78]	✓	✗
RADIATE [72]	✓	✗
Oxford Radar RobotCar [20]	✓	✗
RADDet [83]	✗	✓
DeepSense [7]	✗	✓
MaxRay*	✗	✓

*MaxRay is the only 6G synthetic dataset.
†Refer to Tab. 3 in Appendix J for a more detailed comparison.

[1] made possible by decades of statistical radio modelling [36, 87]

3. Dataset

Figure 2. Block diagram of MaxRay.

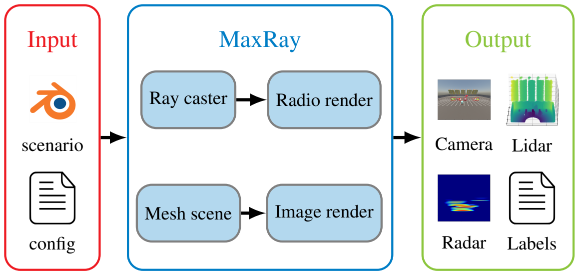

Table 2. Sensor entries in MaxRay.

Entry	Type	Label
Camera	Image	bounding box + class
Lidar	Tensor	per point material + class
Depth	Image	bounding box + class
Radar	Image	bounding box + reflectors + class
CSI	Tensor	reflectors + paths AoA*/AoD†

*AoA: angle of arrival
†AoD: angle of departure

Table 3. Radio synthesis.

Parameter	Value
RX array	16×16
Bandwidth	800MHz
Carrier	28GHz
Range-Angle Bins	480×640

Our radio-visual dataset is created using MaxRay [14]—a ray tracing tool for accurate radio propagation simulations. MaxRay also incorporates the open-source Blender engine for creating photo-realistic environments [28]. As such, we can model arbitrarily complex environments and synthesise paired responses in the vision and radio domains.

Fig. 2 depicts the tool block diagram. A Blender scenario and a configuration file (containing radio parameters such as carrier frequency and bandwidth) are inputted to MaxRay. MaxRay uses Python APIs to render responses for a variety of imaging sensors (e.g., camera, lidar, depth images) along with their labels. The rendering and label quality allow us to train an off-the-shelf Yolo v5 models [33] from scratch. The core of MaxRay leverages the ray casting capability of Blender to simulate complex radio phenomena (e.g., scattering and reflection) and calculate their propagation losses. These propagation losses are then used to create channel state information (CSI), which is in turn converted to radar heatmaps according to an orthogonal frequency-division multiplexing (OFDM) signalling architecture.

3.1. Modelling & synthesis details

Vision. We model everything in Blender. Currently, we implement five different materials (glass, wood, concrete, metal, and water), 20 different building types, and 28 unique and accurate car models. Once the dataset is open-sourced, the research community can build on our Blender models to further extend the scale and richness of our radio-visual dataset.[2] As of now, we sample from a standard normal distribution (10cm standard deviation) to randomise the location of 28 unique cars on the road, while also permuting their colour. The pose is constrained by the lane along which cars are travelling. Different scenarios are evenly distributed.

[2]See datasheet in Appendix K for further details on extension.

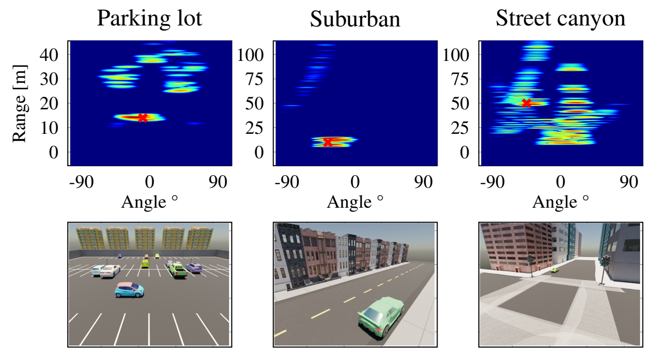

Figure 3. Example radio-visual heatmap-image pairs of three different scenarios: parking lot (left), suburban (middle), and street canyon (right)

Radio. We generate radio heatmaps that correspond to visual Blender models using ray tracing. Appendix E treats the signal processing principles of OFDM radar, which we implement in our synthesis flow. We validate our ray tracing against empirical measurements, as well as other commercial ray tracers. Recent analysis against [7] reveals effective behavioural modelling. Our current scenarios focus on cars. We plan to extend to scenarios featuring humans, which require elaborate modelling of micro Doppler effects [9].

3.2. Version 1.0

The current version of the dataset supports the sensor configurations listed in Tab. 2. All available sensors are paired and synchronised per data point, facilitating cross-modal learning. Tab. 3 lists the radio configurations used in dataset, which comply with current 5G Advanced specifications [52]. Further, the dataset has sequences of 15 data points that allow for time series modelling.

The full version of the dataset has 3 scenarios: a parking lot, a suburban street, and a street canyon. Fig. 3 depicts one example per scenario. In parking lot, one car is driven from left to right or right to left. In suburban, one car drives along the houses towards the camera or away from it. The same holds for the street canyon scenario. Note how radio heatmaps have different ranges, as well as different amount of spurious clutter. For instance, parking lot has dynamic background clutter arising from changes in the location and pose of stationary cars across data points. Groundtruth information is supplied in the form of bounding boxes for vision and target polar coordinates for radar (i.e., range and angle). In the terms of data diversity, there are 50 different cars, backgrounds and foregrounds randomised throughout the dataset. There are also portions of data that model mixed weather scenarios such as rain, snow, fog, and dust. We provide a few dataset illustrations in Appendix H.

Phase 1 of dataset release focuses on the parking lot sce-

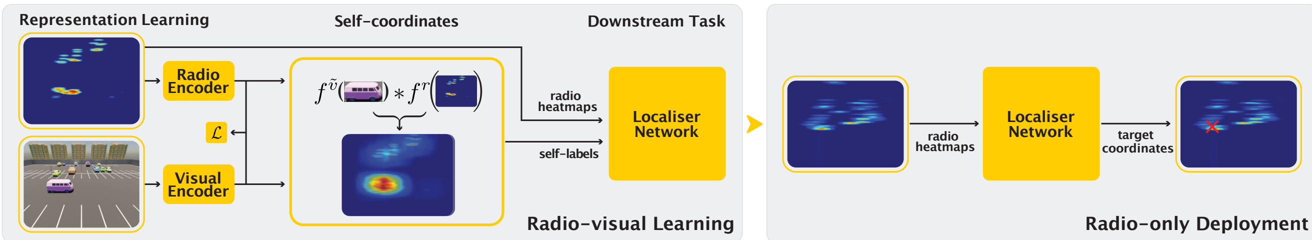

Figure 4. Radio target localisation via self-supervised radio-visual correspondence. We combine contrastive radio-visual learning with visual masking to extract radio target self-coordinates, on which we then train a radio-only localiser net.

nario only. For the remainder of this paper, we use parking lot with radar and camera entries. Specifically, parking lot has 30,000 paired radio-visual data points, split into 24k training and 6k validation sets. An additional 10k set is withheld for testing. All our results are reported on the 6k validation set.

4. Method

We aim to automatically localise a target of interest in radio by tapping into the common information radio and its paired vision capture about the physical world. In fact, barring propagation nuances, radio imaging can be thought of as a low-resolution form of vision—Appendix A justifies this view using a 1st-order analytic analysis derived from first principles. As such, jointly embedding radio and vision becomes not only a convenience, but is also naturally grounded in physics. Therefore, we would hope that the joint embedding architecture would constitute a powerful representation for building a wide variety of radio-only or combined radio-visual perception tasks. Concretely, our approach in this paper is to: ❶ learn cross-modal spatial features via radio-visual correspondence, ❷ extract self-estimates of target coordinates (i.e., pseudo labels) via cross-modal attention between the spatial features, and ❸ use the self-coordinates to train a radio-only target localiser network. Fig. 4 illustrates this three-step procedure. In what follows, we explain further ❶, ❷, and ❸.

4.1. Representation learning

We employ a flavour of contrastive learning we dub masked contrastive learning (MCL) in order to self-localise targets in radio. MCL is inspired by earlier pioneering audio-visual learning works [12, 13], as well as canonical visual contrastive learning [26, 27]. We begin by formalising MCL.

Masked contrastive learning (MCL). Let (r, v) be a radio-visual data pair, where $r \in \mathbb{R}^{1\times H\times W}$ is a radar heatmap and $v \in \mathbb{R}^{3\times H\times W}$ is a corresponding RGB image. Encode, respectively, radio and vision by two backbone nets f_{θ^r} and f_{θ^v}, and their momentum-filtered versions $f_{\bar{\theta}^r}$ and $f_{\bar{\theta}^v}$, assuming some weight parametrisation $\{\theta^r, \theta^v\}$. Each backbone net encodes per bin one C-dimensional feature vector within 2-dimensional spatial bins, i.e., $f_{\theta^r}(r), f_{\theta^v}(v) \in \mathbb{R}^{C\times h\times w}$. The spatial binning resolution $h \times w$ is generally coarser than the original image resolution $H \times W$. Denote by $f^r_n(r), f^v_n(v) \in \mathbb{R}^C$ radio and vision spatial encodings at bin $n \in \Omega = \{1, \ldots, h\} \times \{1, \ldots, w\}$. Construct a visual target mask $\gamma := [\gamma_{ij}] \in [0,1]^{H\times W}$ such that $f^{\tilde{v}}_m(\gamma \odot v) \in \mathbb{R}^C$ is defined for $m \in \tilde{\Omega} = \{1, \ldots, \tilde{h}\} \times \{1, \ldots, \tilde{w}\}$ to retain encodings for the target of interest only in the RGB image (e.g., as delineated by a bounding box), where $\odot$ is the element-wise product, $\tilde{\Omega} \subset \Omega$ is a subset of spatial locations, and $\tilde{v}$ denotes masking in vision. In practice, the target mask can either be (1) estimated using off-the-shelf vision object detectors such as Yolo [33, 67], or (2) obtained directly as groundtruth during data synthesis. Use 2-layer MLP projector heads g_{θ^r} and g_{θ^v} to collapse the spatial encodings of the backbone nets f_{θ^r} and f_{θ^v} onto vector representations as

$$
\begin{aligned}
q^r &= g_{\theta^r}(f_{\theta^r}(r)), & k^{\tilde{v}} &= g_{\theta^v}(f_{\bar{\theta}^v}(\gamma \odot v)) \\
q^{\tilde{v}} &= g_{\theta^v}(f_{\theta^v}(\gamma \odot v)), & k^r &= g_{\theta^r}(f_{\bar{\theta}^r}(r))
\end{aligned}
$$

where vectors $q^r, q^{\tilde{v}}, k^r, k^{\tilde{v}} \in \mathbb{R}^N$, superscripts r and $\tilde{v}$ denote respectively radio and masked vision, and following MoCo's query q and key k notation [45]. With each r, use $K+1$ samples of $\tilde{v}$ of which one sample $\tilde{v}^+$ is a true match to r and K samples $\{\tilde{v}^-_i\}_{i=0}^{K-1}$ are false matches—vice versa with each $\tilde{v}$, $K+1$ samples of r. The one-sided cross-modal contrastive losses that test for masked vision-to-radio and radio-to-masked vision correspondences are

$$
\mathcal{L}_c^{\tilde{v}\to r}(q^r, k^{\tilde{v}+}, \mathbf{k}^{\tilde{v}-}) = -\mathop{\mathbb{E}}_{r,v} \log \frac{e^{\mathrm{sim}(q^r, k^{\tilde{v}+})}}{e^{\mathrm{sim}(q^r, k^{\tilde{v}+})} + \sum_i e^{\mathrm{sim}(q^r, k_i^{\tilde{v}-})}}
$$

$$
\mathcal{L}_c^{r\to \tilde{v}}(q^{\tilde{v}}, k^{r+}, \mathbf{k}^{r-}) = -\mathop{\mathbb{E}}_{r,v} \log \frac{e^{\mathrm{sim}(q^{\tilde{v}}, k^{r+})}}{e^{\mathrm{sim}(q^{\tilde{v}}, k^{r+})} + \sum_i e^{\mathrm{sim}(q^{\tilde{v}}, k_i^{r-})}}
$$

where $\mathrm{sim}(x, y) := x^\top y/\tau$ is a similarity function, τ is a temperature hyper-parameter, $k^{x+/-} = g_{\theta^x}(f_{\bar{\theta}^x}(x^{+/-}))$ are encodings that denote true and false corresponding signals $x \in [r, \tilde{v}]$, and vector $\mathbf{k}^{x-} = \{k_i^{x-}\}_{i=0}^{K-1}$ holds K false encodings. Then the bidirectional masked contrastive loss[3] that incentivises cross-modal spatial attention becomes

$$
\mathcal{L}_{\mathrm{MCL}} = (\mathcal{L}_c^{\tilde{v}\to r} + \mathcal{L}_c^{r\to\tilde{v}})/2 \tag{1}
$$

After training, the visual spatial encodings of the masked target $f^{\tilde{v}}_m(\gamma \odot v)$ can be correlated against the radio spatial

[3] see Fig. 3 in Appendix C for further illustration

encodings covering the entire sensing scene $f_n^r(r)$ in order to produce an attention map (with appropriate padding)

$$h_n(r,v) = \mathrm{conv2d}\left(f_n^r(r), f_m^{\tilde{v}}(\gamma \odot v)\right), \quad n \in \Omega, m \in \tilde{\Omega} \tag{2}$$

To measure best cross-modal regional agreement, the attention map is maximised over spatial bins

$$S(r,v) = \max_{n \in \Omega} h_n(r,v) \tag{3}$$

4.2. Target self-estimation

Once the backbone networks are learnt and their spatial features are stable, we can use cross-modal attention maximisation (cf., Eqs. 2 & 3) to self-generate target coordinate estimates. This self-labelling is inherently noisy, but remarkably powerful. Particularly, a downstream localiser network is able to smooth these self-estimates when trained over a sufficiently large number of data points—as determined by the mutual information with perfect coordinates [44].

Rescaling and calibration. Target coordinate estimates are obtained in the spatial feature grid $h \times w$. We rescale to bring back to original grid $H \times W$, and perform one-off calibration for systematic offsets on entire dataset.

4.3. Localiser network

We construct the dataset $(r, \hat{y}) \in \mathcal{D}_{\text{loc}}$ from tuples of radio heatmaps r and their target self-labels $\hat{y}$. The localiser network is trained to regress $\hat{y}$ from r using a mean squared error (MSE) loss.

5. Benchmarks

We discuss our baselines and empirical evaluation. We refer the reader to Appendix D for implementation details.

5.1. Baselines

Radar target detection is a historic and thoroughly investigated topic as it pertains to many civil and military applications. The objective is to predict a target's position and velocity. However, extracting wanted information (i.e., the target) from unwanted information (i.e., clutter) is a challenging task. Due to radio propagation phenomena, both could exhibit comparable statistical behaviour. We implement expert statistical techniques used by various industries in millions of products, and designate as our first strong standard baseline. Equally, a fully-supervised localiser network trained on groundtruth coordinates naturally forms our second deep learning-based baseline. We also adapt to our dataset a third radio-visual fusion scheme called RODNet [77]. In what follows, we describe these approaches.

Statistical. Extracting information from a radio response representation is a multi-step procedure. First, radio targets in two different domains, range-angle and range-velocity, are binarised via a threshold technique (e.g., CFAR [69]) and then clustered (e.g., via DBSCAN [32]) to form one point cloud per target. Targets are then matched between the two different domains over the same and hopefully unique range. Point cloud centroids are used to track targets.

Considering such multi-step procedure, the following shortcomings come to mind. First, how should we detect the wanted target from the matched targets (e.g., how to remove clutter). Second, some information is ignored when assigning a target centroid (e.g., information from the shape of the point clouds). Third, setting the optimal thresholds, guard bands, training bands, number of points per cluster [69] is an exceedingly brittle exercise. It is our hope that end-to-end learning is able to address some of these shortcomings.

For the statistical baseline to become more competitive with learning-based approaches, we make it "Genie-aided", i.e., the peak closest to groundtruth is assigned as a target. Genie-aided algorithms are common practice in information theory literature to study upper performance bounds [30].

Supervised. In radio sensing, labelling empirical heatmaps (e.g., object type, centre, bounding box) is infeasible at scale as we cannot interpret the scene by manual inspection. However, we consider the supervised network as a useful upper bound on the performance of self-supervised localisation.

Compared to computer vision, radio imaging has no prescribed or de facto neural architectures to use for evaluation. We therefore use Microsoft's AutoML tool NNI (Neural Network Intelligence) [58] to search for strong candidate architectures. Specifically, we searched for optimizers, loss functions, learning rates, momentums, neural architectures via resolution branching, and activation functions. The performance of the supervised baseline in Sec. 6 corroborates the quality of the search. Detailed description of the architectural search space is given in Appendix I.

Radio-visual fusion. RODNet uses a student-teacher network configuration [77]. The teacher combines object detection in vision and statistical peak detection in radio to derive object class and location estimates. The radio-only student network is trained on the teacher's estimates. We use the student network as a baseline and characterise against pseudo groundtruth labels. Like our system, our RODNet implementation operates on a single heatmap snapshot without spatio-temporal convolution. Compared to statistical CFAR baseline—using a genie-aided peak selection where the peak closest to the target is always assigned—RODNet's vision+radio teacher implements a peak fusion to approximate the optimal joint camera-radio detector.

5.2. Empirical data

For further empirical validation of our radio-visual SSL algorithm, we use the parking lot scenario of the Camera-Radar of the University of Washington (CRUW) dataset [77]. Tab. 4 in Appendix J compares CRUW to MaxRay.

Pseudolabels construction. Since empirical data does not come with groundtruth correspondence labels, we employ the following pseudolabelling procedure (based on RODNet). First, we detect and segment objects from images using a Mask R-CNN object detector [46]. Second, we detect radar peaks using CFAR and cluster them into groups using

DBSCAN. Third, we perform intrinsic camera calibration to convert camera objects into x-y coordinates. Fourth, we match image segmentations to radar peak clusters.

Pseudosupervision. As a replacement for the supervised baseline on MaxRay, we use the matched range-angle pseudolabels from the above procedure to train a pseudosupervised localiser net that takes as input CRUW radar heatmaps.

5.3. Metrics

We measure location estimation performance using two error statistics: 50th and 90th percentiles (abbrv. %ile), and on the validation set unless otherwise stated. Formally, let X_e denote the location error as a random variable, x_e denote the %ile error, $\Pr$ the error probability distribution, F_{X_e} its cumulative probability distribution, and $p_e \in \{0.5, 0.9\}$ a probability value it assumes. Then $p_e := F_{X_e}(x_e) = \Pr(X_e \leq x_e)$. Throughout evaluation, we will simply quote these error statistics $(x_e|_{p_e \in \{0.5,0.9\}})$ as the 50th and 90th %ile errors.

6. Results

Having described our benchmarking setup in Sec. 5, we turn next to discussing results. For MaxRay and CRUW, results are computed on 6k and 1.8k data points of validation sets, respectively.

6.1. Localisation performance

We examine the overall performance of our MCL-based self-labelled localiser net and compare it against: fully supervised, RODNet, and statistical baselines. The self-labelled net, supervised, and RODNet share identical downstream architecture and training configurations. This common architecture is only specialised with different convolutional kernel sizes across datasets due to differences in radio configurations (cf., MaxRay vs. CRUW in Appendix J). We denote the statistical baseline by Constant False Alarm Rate (CFAR). Tab. 4 summarises the performance in terms of 50th %ile and 90th %ile localisation errors on the validation sets. Not surprisingly, the fully-supervised net performs most favourably with around 30cm and 1.4m median errors, respectively on MaxRay and CRUW. Note the drop in supervised performance between MaxRay and CRUW is largely due to the halved angular resolution of CRUW (cf., Appendices A & J). MCL comes second with approx. 0.94m and 2.5m median errors, respectively on MaxRay and CRUW. This is remarkable given that MCL has automatically learned how to localise targets by simply observing paired radio-visual data. Genie-aided CFAR performs worse with roughly 2.8× and 1.8× MCL's median errors, respectively on MaxRay and CRUW. RODNet is also worse with 3.2× and 1.3× MCL's median errors, respectively on MaxRay and CRUW. As discussed in Sec. 5.1, RODNet teacher employs a conventional radio-visual fusion scheme that relies on radio CFAR detection aided by vision. On the higher angular resolution of MaxRay, RODNet's fusion scheme seems to be far less effective than our joint embedding architecture.

Table 4. Performance summary on MaxRay and CRUW. MCL sets a new SOTA perf. for *label free* localisation.

Method	Label free	MaxRay perf. (error in m)		CRUW perf. (error in m)	
		50th %ile	90th %ile	50th %ile	90th %ile
Supervised*	✗	0.289 ±0.017	0.922 ±0.042	1.382 ±0.128	5.402 ±0.063
MCL	✓	**0.942** ±0.016	**4.681** ±0.158	**2.558** ±0.072	10.969 ±0.032
CFAR†	✓	2.709	8.062	4.659	6.161
RODNet	✓	3.012 ±0.014	8.913 ±0.107	3.281 ±0.334	7.791 ±0.355

*pseudosupervised in CRUW †Genie-aided

Table 5. Backbone training configurations: MCL, SCL, CL

Backbone	MaxRay perf. (error in m)	
	50th %ile	90th %ile
MCL	0.942 ±0.016	4.681 ±0.158
SCL	1.571 ±0.050	3.539 ±0.062
CL	3.111 ±0.358	17.498 ±0.317

Table 6. Backbones with Yolov5 bounding boxes.

Backbone	MaxRay perf. (error in m)	
	50th %ile	90th %ile
MCL^Y	1.351	7.649
SCL^Y	2.188	5.462

^{Y} using Yolov5 bounding boxes

6.2. Ablations and analysis

We now conduct experiments to better understand MCL's performance against alternatives, its dependence on masking accuracy and self-labelling density, its modelling capacity, and its sensitivity to radio-visual commonalities.

Masked contrast vs. other contrastive learning flavours. We have found MCL to be an effective radio-visual learning strategy on synthetic and noisier empirical data. We would like to understand, however, how MCL compares to other forms of contrastive learning from the literature. To this end, we first consider spatial contrastive learning (SCL) that has appeared in multiple recent works that use 2-D backbone modelling [3, 5, 80]. SCL performs contrastive learning in 2-D to incentivise cross-modal spatial attention. We adapt SCL to the radio-visual problem setting and provide formal definition and illustrative comparisons in Appendix C. We also consider vanilla contrastive learning (CL) without masking [26]. I.e., the following investigates the performance of model variants: SCL and CL. We ask: *What role does masking play during contrastive radio-visual learning?*

Tab. 5 analyses the performance of the three backbone configurations on MaxRay. We note that vanilla CL performs poorly with 3.1m median error. We attribute the high localisation error to the lack of target sensitivity of CL during training. MCL, on the other hand, is trained to attend to targets through masking and exhibits a 50th %ile error of 0.94m, interestingly 1/3rd better than SCL at 1.57m. A closer look at MCL's 90th %ile error at 4.6m reveals that it is also around 1.3× "lazier" than SCL at tracking higher %ile targets. However, we note that SCL has *failed* to train on the noisier empirical CRUW dataset, while MCL has given a new SOTA 50th %ile performance as shown in Tab. 4 (despite CRUW's low angular resolution). We conjecture that MCL's 2-layer MLP projector supports denoising—a

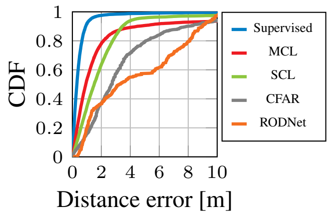

Figure 5. Localisation error CDFs of methods: Supervised, MCL, SCL, CFAR, and RODNet.

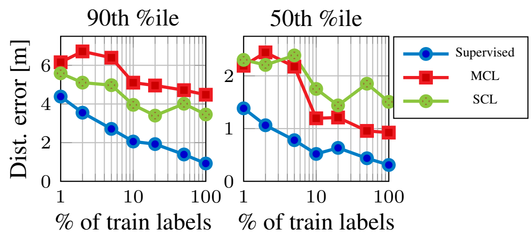

Figure 6. Effect of number of training labels on localisation performance for Supervised, MCL, and SCL.

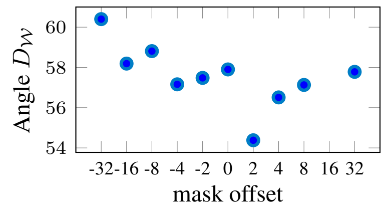

Figure 7. Effect of radio-visual mutual information on self-labelling for MCL, on MaxRay's validation set and as measured by $D_{\mathcal{W}}$.

crucial feature the more computationally efficient SCL lacks. Classically, noisy radar data could have many spurious ghost targets [69]. Hence, a parameter-heavy projector head may prove necessary to stabilise learning. Fig 5 depicts the error cumulative density functions (CDFs) of all methods.

Impact of noisy masks. Taking advantage of the controllability of MaxRay, we have so far utilised perfect masks generated during synthesis. We now investigate the impact of using noisy mask estimates during contrastive learning. To this end, Tab. 6 lists the localisation performance of MCL and SCL using masks from Yolov5, similar to how we integrated CRUW into our SSL pipeline (see Appendices G & J). We observe around 0.4m and 0.6m degradation in the median localisation performance, respectively for MCL and SCL.

Impact of label density. We investigate performance enhancements as a function of increased number of noisy labels. Using MaxRay's 24k training set, we sweep the amount of labels and self-labels used to train the localiser nets of supervised, MCL, and SCL. Then we evaluate on the validation set to gauge the localisation performance sensitivity to the amount of available training (self-)labels. We cover the training points in logarithmic steps.

With the noisy labels of MCL and SCL, it can be shown that the localiser nets learn to compensate for such noise by using sufficiently large number of data points [44]. The number of required data points is a function of the mutual information between noisy target coordinates and perfect coordinates [44]. Fig. 6 examines this effect for supervised as a reference baseline, and MCL and SCL. We observe that MCL has a better label density tolerance than SCL w.r.t. the 50th %ile performance. The opposite holds true, i.e., SCL is better than MCL w.r.t. the 90th %ile performance. This finding mirrors the localisation error analysis of Fig. 5. On MaxRay, MCL seems to be a better self-localiser of the bulk of the distribution of the validation set, while SCL seems to cope better with corner cases.

Self-labels deviation from groundtruth. We have uncovered qualitative differences between the ability of MCL and SCL to self-localise targets using cross-modal attention. We turn next to quantify how far the self-labels of MCL and SCL deviate from groundtruth labels. The following analysis sheds further light on the performance differences between masked projector contrast and spatial contrast.

We analyse the deviation of self-labels from groundtruth labels through the lens of three metrics. Let $p_{\text{gt}}(y)$ and $p_{\text{est}}(\hat{y})$ denote the distributions of groundtruth labels y and self-label estimates $\hat{y}$, respectively. The shift between p_{gt} and p_{est} can be quantified using the 1-D Wasserstein distance $D_{\mathcal{W}}(p_{\text{gt}}, p_{\text{est}}) = \int_0^1 c\big(|F_{\text{gt}}^{-1}(x) - F_{\text{est}}^{-1}(x)|\big)dx$, where F is the CDF function, and c denotes a cost function—we use a quadratic cost below. We employ $D_{\mathcal{W}}$ because of its robustness and specificity on empirical measurements [66]. We also use two more conventional information-theoretic measures: the Kullback-Leibler (KL) divergence $D_{\text{KL}}(p_{\text{gt}}(y)||p_{\text{est}}(\hat{y}))$ and mutual information (MI) $I(y;\hat{y}) = D_{\text{KL}}(p_{(\text{gt, est})}||p_{\text{gt}}\, p_{\text{est}})$ [29]. KL quantifies the shift between p_{gt} and p_{est}, similar to the Wasserstein distance. MI measures how dependent p_{est} is on p_{gt} (in nats below).

Tab. 7 evaluates numerically the three distribution deviation metrics. The metrics are computed for the training and validation sets separately. The metrics are also computed for range and angle coordinates separately. On MaxRay, we can see that SCL outperforms MCL consistently across metrics, sets, and coordinates. Specifically, both $D_{\mathcal{W}}$ and D_{KL} are lower for SCL, indicating better match to groundtruth. Similarly, MI is higher for SCL, indicating better match to groundtruth. Similar observations hold for MCL^Y and SCL^Y using mask estimates from Yolov5. For qualitative comparison of distributions, consult the empirical histograms in Fig. 5 in Appendix F.

Impact of dimensionality. We investigate if SSL backbone capacity limits performance. To do so, we train backbone configurations and measure their self-label deviation from groundtruth as a function of: (a) feature dimensionality per spatial bin (denoted by C in Sec. 4.1) for MCL and SCL, and (b) the dimensionality of the 2-layer MLP projector head for MCL. Fig. 8 depicts the Wasserstein distances across a number of (a) & (b) configurations. For SCL, doubling C up to 1024 features per spatial bin has negligible effect on range and angle self-label distances to groundtruth. For MCL, there is a mild reduction in distances as a function of feature dimensionality, and no effect for using a larger projector

Table 7. Quantifying how far self-labels deviate from groundtruth.

Config.	Training						Validation					
	Range			Angle			Range			Angle		
	$D_{\mathcal{W}}^{\downarrow}$	$D_{\text{KL}}^{\downarrow}$	$\text{MI}^{\uparrow}$	$D_{\mathcal{W}}^{\downarrow}$	$D_{\text{KL}}^{\downarrow}$	$\text{MI}^{\uparrow}$	$D_{\mathcal{W}}^{\downarrow}$	$D_{\text{KL}}^{\downarrow}$	$\text{MI}^{\uparrow}$	$D_{\mathcal{W}}^{\downarrow}$	$D_{\text{KL}}^{\downarrow}$	$\text{MI}^{\uparrow}$
MCL	28.509	7.294	0.947	59.359	5.572	1.120	27.940	5.319	0.931	57.901	3.571	1.121
SCL	19.063	7.281	1.217	40.115	5.548	1.567	18.691	5.282	1.226	39.387	3.541	1.528
MCL^{Y}	34.930	7.287	0.934	63.613	5.559	0.985	34.814	5.303	0.913	62.095	3.572	0.987
SCL^{Y}	20.867	7.273	1.117	49.472	5.566	1.512	21.035	5.289	1.124	48.213	3.562	1.476

^{Y} using pseudo bounding boxes obtained from Yolov5

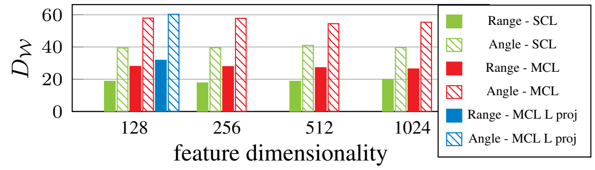

Figure 8. Effect of dimensionality on self-labelling for MCL & SCL, on MaxRay's validation set and as measured by $D_{\mathcal{W}}$. L proj in legend denotes $4\times$ projector head.

head. We, therefore, conclude that the performance of self-labels is fundamentally limited by the underlying resolution of radio imaging rather than the model's learning capacity.

Impact of cross-modal commonalities. We investigate the relationship between cross-modal commonalities and localisation performance. The InfoMin principle tells us how to "regularise" contrastive learning in order to obtain optimal downstream performance [74].[4] According to InfoMin, there are three regimes of MI captured during learning: (1) missing info, (2) sweet spot, and (3) excessive info [74]. These three regimes can be empirically observed as a U-shaped curve for a given downstream task. We control the amount of radio-visual MI through masking in the vision domain. Specifically, we train MCL model variants with the groundtruth target masks progressively enlarged or shrunk, i.e., by positively or negatively padding the masks. We then obtain self-labels for these MCL variants and measure their $D_{\mathcal{W}}$ as before. Fig. 7 depicts $D_{\mathcal{W}}$ of the angle distribution as a function of target mask offsets. We observe a U-shaped curve whose minima is at an offset of 2 pixels. This corroborates that masking in vision enhances target sensitivity (ablated in Tab. 5), and further illustrates the degradation as we increase (+ offsets) or reduce (– offsets) radio-visual MI.

7. Discussion

6G sensing. Making cellular basestations "see" the surrounding environment while sending data is a major feature in 6G networks. There are non-trivial protocol-level challenges in 6G network design in order to support sensing (see Appendix E). In this paper, we concentrate on the higher-level challenge of *automatically* building target localiser models using radio heatmaps that are accompanied by visual images, i.e., paired radio-visual data collected at a basestation equipped with a camera. Through cross-modal attention, we show how to estimate self-labels for training a downstream radio localiser network. Specifically, we demonstrate that the performance of the localiser network is *not upper bounded* by the accuracy of self-labels, and that using larger number of noisy self-labels enhances estimation. This finding is in line with prior work [44], and serves to reaffirm the paradigm of self-supervised radio-visual learning for scalable radio sensing. Our synthetic radio-visual dataset helps establish the performance trends of radio-visual SSL localisation by virtue of a controlled groundtruth. That is, out dataset is dedicated to the study and refinement of radio-visual SSL algorithms, and *not* to the production of 6G perception models. Our SSL algorithm, on the other hand, is readily applicable to empirical data with no groundtruth (cf., Tab. 4). We believe that our radio-visual SSL objective provides a viable route towards vast data scalability for 6G sensing.

Limitations. We note that radio sensing capabilities are fundamentally set by the choice of configurations in Tab. 3. We have opted to base this somewhat conservative choice on 5G Advanced specifications [52] in order to inform cellular stakeholder discussions. We would, however, note that much improved radio sensing performance can be attained through increased bandwidth and/or denser antenna arrays, such as in Terahertz or even higher Millimetre-wave bands [31, 71]. We would refine our dataset and results in light of future consensus on 6G sensing specifications.

Broader impact. Our work has a broader societal impact in that it has the potential to alleviate some of risks associated with the surveillance economy. Specifically, once trained and deployed, our radio sensing system offers a scalable alternative to pervasive vision surveillance that is inherently privacy-preserving, while achieving many of the sought-after safety and security benefits.

8. Conclusion

In this paper, we present a new radio-visual learning task for emerging 6G cellular networks. The task tackles the problem of accurate target localisation in radio, employing a novel learning paradigm that works by simply ingesting large quantities of paired radio-visual data. This is in stark difference to supervised and/or classic statistical methods whose success hinges on laborious labelling and/or modelling of empirical measurements, which are expensive to scale. We demonstrate strong label-free target localisation performance on synthetic and empirical data. Our novel target localisation paradigm is made possible by a new dataset and benchmark intended to foster future research on radio sensing for next generation cellular systems.

[4] building on earlier information bottleneck literature [6, 35, 75]

References

[1] Tesla AI day. 1

[2] A. Bourdoux, et. al. 6G White Paper on Localization and Sensing, 2020. 1

[3] Triantafyllos Afouras, Yuki M Asano, Francois Fagan, Andrea Vedaldi, and Florian Metze. Self-supervised object detection from audio-visual correspondence. pages 10575–10586, 2022. 2, 6

[4] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2143–2147. IEEE, 2020. 2

[5] 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, 6

[6] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 8

[7] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, and N. Srinivas. DeepSense 6G: Large-scale real-world multi-modal sensing and communication datasets. *to be available on arXiv*, 2022. 2, 3

[8] Mohammed Alloulah and Howard Huang. Future millimeter-wave indoor systems: A blueprint for joint communication and sensing. *Computer*, 52(7):16–24, 2019.

[9] Mohammed Alloulah, Zoran Radivojevic, René Mayrhofer, and Howard Huang. Kinphy: a kinetic in-band channel for millimetre-wave networks. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, pages 364–377, 2019. 3

[10] 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. 1, 2

[11] 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

[12] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 2, 4

[13] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision*, pages 435–451, 2018. 2, 4

[14] M Arnold, M Bauhofer, S Mandelli, M Henninger, F Schaich, T Wild, and S ten Brink. Maxray: A raytracing-based integrated sensing and communication framework. In *2022 2nd IEEE International Symposium on Joint Communications & Sensing (JC&S)*, pages 1–7. IEEE, 2022. 3

[15] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[16] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 2

[17] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019. 2

[18] Kyungjune Baek, Minhyun Lee, and Hyunjung Shim. Psynet: Self-supervised approach to object localization using point symmetric transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10451–10459, 2020. 2

[19] Adrien Bardes, Jean Ponce, and Yann LeCun. Variance-invariance-covariance regularization for self-supervised learning. *ICLR, Vicreg*, 2022. 1, 2

[20] 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. 2

[21] Martin Braun, Christian Sturm, and Friedrich K. Jondral. On the single-target accuracy of OFDM radar algorithms. In *2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 794–798, 2011.

[22] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 1, 2

[23] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014. 2

[24] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. 2

[25] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7546–7554, 2021. 2

[26] 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. 1, 2, 4, 6

[27] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 4

[28] Blender Online Community. Free and open 3D creation software. 3

[29] Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley-Interscience, 2006. 7

[30] Natasha Devroye, Patrick Mitran, and Vahid Tarokh. Achievable rates in cognitive radio channels. *IEEE Transactions on information theory*, 52(5):1813–1827, 2006. 5

[31] Mohamed Elkhouly, Jaegeun Ha, Michael J Holyoak, David Hendry, Mustafa Sayginer, Ryan Enright, Ioannis Kimionis, Yves Baeyens, and Shahriar Shahramian. Fully integrated 2D scalable TX/RX chipset for D-band phased-array-on-glass modules. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 65, pages 76–78. IEEE, 2022. 8

[32] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996. 5

[33] Glenn Jocher et. al. ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support, Oct. 2021. 3, 4

[34] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10699–10709, 2020. 1, 2

[35] Ian Fischer. The conditional entropy bottleneck. *Entropy*, 22(9):999, 2020. 8

[36] Eran Fishler, Alexander Haimovich, Rick S Blum, Leonard J Cimini, Dmitry Chizhik, and Reinaldo A Valenzuela. Spatial diversity in radars—models and detection performance. *IEEE Transactions on signal processing*, 54(3):823–838, 2006. 2

[37] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7062, 2019. 2

[38] Stefano Gasperini, Patrick Koch, Vinzenz Dallabetta, Nassir Navab, Benjamin Busam, and Federico Tombari. R4dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes. *arXiv preprint arXiv:2108.04814*, 2021. 2

[39] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[40] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 1, 2

[41] 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. 1, 2

[42] 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. 1, 2

[43] 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. 2

[44] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 5, 7, 8

[45] 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. 1, 2, 4

[46] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 5

[47] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020. 2

[48] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron Van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021. 2

[49] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. *arXiv preprint arXiv:2203.08777*, 2022. 2

[50] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 2

[51] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021. 2

[52] Younsun Kim, Youngbum Kim, Jinyoung Oh, Hyoungju Ji, Jeongho Yeo, Seunghoon Choi, Hyunseok Ryu, Hoondong Noh, Taehyoung Kim, Feifei Sun, et al. New radio (NR) and its evolution toward 5G-advanced. *IEEE Wireless Communications*, 26(3):2–7, 2019. 3, 8

[53] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[54] Philipp Krähenbühl. Free supervision from video games. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2955–2964, 2018. 1

[55] 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

[56] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. Making the invisible visible: Action recognition through walls and occlusions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 872–881, 2019. 1, 2

[57] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. *arXiv preprint arXiv:2009.09805*, 2020.

[58] Microsoft. Neural Network Intelligence, 1 2021. 5

[59] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[60] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33:4733–4744, 2020. 2

[61] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016. 1, 2

[62] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2

[63] Itai Orr, Moshik Cohen, Harel Damari, Meir Halachmi, and Zeev Zalevsky. Coherent, super resolved radar beamforming using self-supervised learning. *arXiv preprint arXiv:2106.13085*, 2021. 2

[64] Arthur Ouaknine, Alasdair Newson, Julien Rebut, Florence Tupin, and Patrick Perez. 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. 2

[65] 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. 2

[66] Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020. 7

[67] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 4

[68] Sebastian Risi and Julian Togelius. Increasing generality in machine learning through procedural content generation. *Nature Machine Intelligence*, 2(8):428–436, 2020. 1

[69] Hermann Rohling. Radar CFAR thresholding in clutter and multiple target situations. *IEEE transactions on aerospace and electronic systems*, (4):608–621, 1983. 5, 7

[70] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[71] Shahriar Shahramian, Michael J Holyoak, Amit Singh, and Yves Baeyens. A fully integrated 384-element, 16-tile, w-band phased array with self-alignment and self-test. *IEEE Journal of solid-state circuits*, 54(9):2419–2434, 2019. 8

[72] Marcel Sheeny, Emanuele De Pellegrin, Saptarshi Mukherjee, Alireza Ahrabian, Sen Wang, and Andrew Wallace. Radiate: A radar dataset for automotive perception in bad weather. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2021. 2

[73] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021.

[74] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020. 8

[75] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 8

[76] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10052–10062, 2021. 2

[77] Yizhou Wang, Zhongyu Jiang, Yudong Li, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. Rodnet: A real-time radar object detection network cross-supervised by camera-radar fused object 3D localization. *IEEE Journal of Selected Topics in Signal Processing*, 15(4):954–967, 2021. 1, 2, 5

[78] Xinshuo Weng, Yunze Man, Dazhi Cheng, Jinhyung Park, Matthew O'Toole, and Kris Kitani. All-In-One Drive: A Large-Scale Comprehensive Perception Dataset with High-Density Long-Range Point Clouds. *arXiv*, 2020. 2

[79] Thorsten Wild, Volker Braun, and Harish Viswanathan. Joint design of communication and sensing for beyond 5G and 6G systems. *IEEE Access*, 9:30845–30857, 2021. 1

[80] 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. 6

[81] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. In *European Conference on Computer Vision*, pages 775–791. Springer, 2020. 1

[82] 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. 1, 2

[83] Ao Zhang, Farzan Erlik Nowruzi, and Robert Laganiere. Raddet: Range-azimuth-doppler based radar object detection for dynamic road users. In *2021 18th Conference on Robots and Vision (CRV)*, pages 95–102. IEEE, 2021. 2

[84] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 1, 2

[85] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018. 1, 2

[86] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2

[87] Thomas Zwick, Christian Fischer, and Werner Wiesbeck. A stochastic multipath channel model including path directions for indoor environments. *IEEE journal on Selected Areas in Communications*, 20(6):1178–1192, 2002. 2