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

Mohammed Alloulah Maximilian Arnold
Nokia Bell Labs

Abstract

This supplementary material consists of 11 appendices. It provides expanded discussion, background details, results, illustrations, and documentation for radio-visual data and algorithms.

A. Radio-visual analytic relationship

In radio imaging, there are two main phenomena that govern our ability to resolve objects in space. First, range resolution Δr is determined by bandwidth B and obeys $\Delta r = c/2B$, where c is the speed of light. In typical millimetre-wave frequencies for 6G, $B \approx 1$GHz which gives ~ 0.3 metre resolution. Second, the angular resolution $\Delta\phi$ is considerably worse and is generally related to our ability to pack antennae in a reasonable form factor. That is, the imaging performance disparity between vision and radio is largely a function of disparities in angular resolution. To see this, let $I(x, y)$ be an image of a sensing scene, where x and y are its horizontal and vertical dimensions, respectively. Let w be the so-called beamwidth of an RF horn antenna. Then the antenna response $h(x, y) = e^{-(x^2+y^2)/(2w^2)}$ is a "distortion" function associated with RF's resolution-limited imaging of a given scene. Specifically, $h(x, y)$ will act as a blurring function that convolves with the original image according to

$$I'(x, y) = I(x, y) * h(x, y) \quad (1)$$

where I' is the degraded image and $*$ is the convolution operator.

Fig. 1 contrasts normal camera imaging against RF's resolution-limited imaging. Left-most Fig. 1a shows a grey-scale image of a factory. Assuming 1 degree angular resolution ($\Delta\phi = 1^\circ$), Fig. 1b in the middle illustrates the blurring effect of Eq. 1 on the camera image. Under higher angular resolution distortion $\Delta\phi = 10^\circ$, the right-most Fig. 1c shows significant blurring as a result of a coarser beamwidth w acting on I.

(a) grey-scale camera image (b) moderate blur (c) high blur

Figure 1. Radio-visual relationship. A grey-scale camera image (a) undergoes blurring in (b) & (c) to simulate the effect of RF's limited angular resolution when using radio to image the environment. (b) shows moderate blur while (c) shows significant blur as a result of angular resolutions $\Delta\phi = 1^\circ$ and $\Delta\phi = 10^\circ$, respectively.

B. Radio-visual subspace analysis

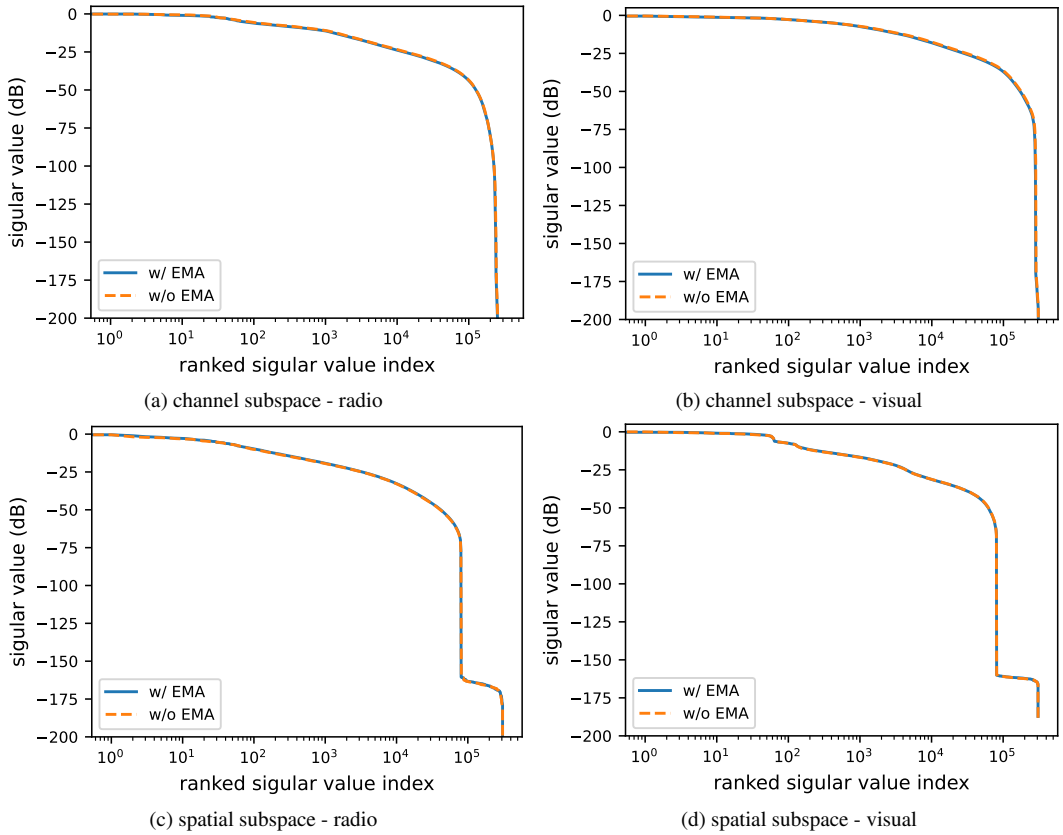
In Sec. 4.1, the spatial encoders $f_{\theta^r}(r), f_{\theta^v}(v) \in \mathbb{R}^{C\times h\times w}$ are introduced. Following the implementation conditions detailed in Appendix D, $f_{\theta^r}(r), f_{\theta^v}(v)$ are concretely $\in \mathbb{R}^{128\times 60\times 80}$. In this section, we analyse their dimensionality after 100 epochs of training on the contrastive loss of Eq. 4. To do so, we evaluate these embedding tensors for the validation set. For each channel $c \in \{1,\dots,128\}$ and spatial bin $n \in \{1,\dots,60\}\times\{1,\dots,80\}$, we compute the centred covariance matrices $\text{Cov}_c \in \mathbb{R}^{128\times 128}$, $\text{Cov}_n \in \mathbb{R}^{4800\times 4800}$ according to

$$\text{Cov}_x = \frac{1}{N}\sum_{k=0}^{N-1}(\mathbf{z}_k^x - \bar{\mathbf{z}}^x)(\mathbf{z}_k^x - \bar{\mathbf{z}}^x)^T \tag{2}$$

where $\mathbf{z}_k^x$ is the embedding vector of a channel or spatial bin[1] $x \in [c, n]$, N is the number of validation samples, and $\bar{\mathbf{z}}^x$ is the respective average. To measure subspaces dimensionality, we compute the singular value decomposition on the covariance matrix $\text{Cov}_x = U\Sigma V^T, \Sigma = diag(\sigma^k)$, following general practice in SSL theory [14, 20]. We use these subspace measurements to quantify changes in the learnt contrastive representation as a result of architectural tweaks such as EMA.

We concatenate the singular values of all channels and all spatial bins and sort them in descending order. Fig. 2 depicts on a logarithmic scale these aggregated singular values. We can readily see that EMA has little effect on the dimensionality of the learnt representation across channels and spatial bins, for both radio and vision branches. We, therefore, opt to exclude it from our experiments for efficiency.

[1] i.e., unfolding the original 2-D spatial bins into a vector of $wh = 4800$ length



(a) channel subspace - radio

(b) channel subspace - visual

(c) spatial subspace - radio

(d) spatial subspace - visual

Figure 2. Radio-visual subspace analysis w/ and w/o EMA.

C. Contrastive learning background & definitions

Contrastive learning (CL). 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 neural networks 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 $\{\theta^r, \theta^v\}$. Additionally, use projector heads g_{θ^r} and g_{θ^v} respectively, such that

$$\begin{aligned} q^r &= g_{\theta^r}(f_{\theta^r}(r)), \quad k^v = g_{\theta^v}(f_{\bar{\theta}^v}(v)), \\ q^v &= g_{\theta^v}(f_{\theta^v}(v)), \quad k^r = g_{\theta^r}(f_{\bar{\theta}^r}(r)) \end{aligned} \tag{3}$$

where vectors $q^r, q^v, k^v, k^r \in \mathbb{R}^N$, superscripts r and v denote respectively radio and vision, and following MoCo's query q and key k notation [13]. With each r, use $K+1$ samples of v of which one sample v^+ is a true match to r and K samples $\{v_i^-\}_{i=0}^{K-1}$ are false matches—vice versa with each v, $K+1$ samples of r. The one-sided cross-modal contrastive losses that test for vision-to-radio and radio-to-vision correspondences are

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

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

where $\mathrm{sim}(x, y) \coloneqq x^\top y/\tau$ 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 $x \in [r, v]$, and vector $\mathbf{k}^{x-} = \{k_i^{x-}\}_{i=0}^{K-1}$ holds K false encodings. Then the bidirectional cross-modal contrastive loss is

$$\mathcal{L}_{\mathrm{CL}} = (\mathcal{L}_c^{v\to r} + \mathcal{L}_c^{r\to v})/2 \tag{4}$$

Spatial contrastive learning (SCL). 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 neural networks f_{θ^r} and f_{θ^v}, assuming some weight parametrisation $\{\theta^r, \theta^v\}$. Each backbone network 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 $f_n^r(r), f_n^v(v) \in \mathbb{R}^C$ radio and vision spatial encodings at bin $n \in \Omega = \{1, \ldots, h\} \times \{1, \ldots, w\}$. Construct a target mask $\gamma \coloneqq [\gamma_{ij}] \in [0, 1]^{H\times W}$ such that $f_m^{\tilde{v}}(\gamma \odot v) \in \mathbb{R}^C$ 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 and $\tilde{\Omega} \subset \Omega$ is a subset of spatial locations. In practice, the target mask can either be (1) estimated using off-the-shelf vision object detectors such as Yolo [11, 18], or (2) obtained directly as groundtruth during data synthesis.

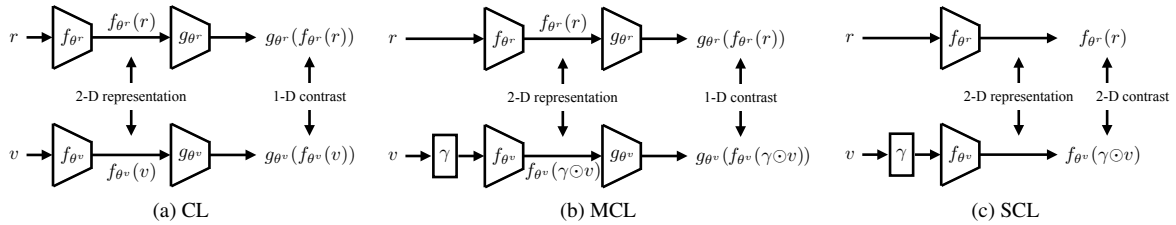
Noting attention maximisation defined earlier in main paper in Eqs. 2 & 3, spatial cross-modal contrastive losses can then be implemented during training from a batch $\mathcal{B}$ for all radio-visual pairs $(r, v) \in \mathcal{B}$ according to

$$\begin{aligned} \mathcal{L}_a^{v\to r}(\mathcal{B}) &= -\mathop{\mathbb{E}}_{\mathcal{B}} \log \frac{\exp\big(S(r, v)/\tau\big)}{\sum\limits_{i\in\mathcal{B}} \exp\big(S(r, v_i)/\tau\big)}, \\ \mathcal{L}_a^{r\to v}(\mathcal{B}) &= -\mathop{\mathbb{E}}_{\mathcal{B}} \log \frac{\exp\big(S(r, v)/\tau\big)}{\sum\limits_{i\in\mathcal{B}} \exp\big(S(r_i, v)/\tau\big)} \end{aligned} \tag{5}$$

where the one-sided loss $\mathcal{L}_a^{v\to r}$ tests for vision-to-radio correspondence, similarly $\mathcal{L}_a^{r\to v}$ tests for radio-to-vision, and τ is a temperature hyper-parameter. The bidirectional contrastive loss that incentivises cross-modal spatial attention becomes

$$\mathcal{L}_{\mathrm{SCL}} = (\mathcal{L}_a^{v\to r} + \mathcal{L}_a^{r\to v})/2 \tag{6}$$

For clarity, Fig. 3 illustrates the three contrastive learning flavours used in this work.



(a) CL (b) MCL (c) SCL

Figure 3. Three contrastive architectures that use spatial backbones: CL, MCL, and SCL. CL follows the original SimCLR architecture [9] and its accessible queue-based MoCo optimisation [10], with the addition of a spatial backbone [5, 8]. MCL is broadly similar to CL except for target masking on the vision branch, which promotes added target sensitivity. SCL does not use a projector head and instead rely on spatial contrast [1, 2, 24].

D. Implementation details

The spatial backbone of the radio and vision encoders uses an architecture similar to VGG-M [5, 8], swapping max pooling for average pooling as recommended in [1]. For standard contrastive ablation in Sec. 6.2 (Contrastive Learning (CL) & Masked Contrastive Learning (MCL)), we base our cross-modal contrastive learning on MoCo v2 and its public implementation [13]. We extend MoCo's implementation with two queues for radio and vision similar to the audio-visual active sampling work in [16]. We have found that filtering the encoders with exponential moving average (EMA) when implementing radio-visual contrast has no tangible advantage, as detailed in Appendix B.

For mask generation in vision, we rely on groundtruth bounding boxes from Blender. We also characterise downstream performance using bounding boxes estimated from off-the-shelf Yolov5 model [11]. Tab. 1 reports Yolov5's IoU-0.5 performance metric as measured on MaxRay.

We train on 640×480 resolution for both RGB images and radio heatmaps. Both radio and vision branches output $128\times80\times60$ spatial features whose dimensionality is reduced using 2-layer MLP projectors to 64-D vectors in the case of CL & MCL. For CL & MCL, we use a MoCo v2 queue whose size equals to the batch size. For CL & MCL, the temperature hyper-parameter is 0.07, whereas for SCL it is 0.1. When implementing spatial attention, we pad bounding boxes by a margin of 5 pixels, and pad a target spatial response by a margin of 1 feature. For backbone training, we use the Adam optimiser [15] with a learning rate of 10^{-5} and no schedule. For all model variants, we train for 200 epochs. We use a batch size of 32 and train in a distributed fashion on 8 GPUs. We trained experiments on two machines with GeForce RTX 2080 Ti GPUs and RTX A5000 GPUs, throughout for backbone training, supervised training, and NNI search space. Backbone training takes around 16–24 hours per experiment depending on model variant and configuration. Both the localiser network trained on self-coordinates and supervised baseline use identical architecture and training as detailed in Tab. 2. MaxRay and CRUW use different convolutional network settings due to differences in range and angular resolutions (cf., Tab. 4). The NNI search space took around 5 days. For ray tracing MaxRay, the ray casting settings of Blender greatly influence performance. We set the maximum number of interactions to 5 and the maximum length travelled to 500m. We parallelise frame creation on 3060 Ti GPU, which gives 200sec creation time per frame. This results in a total of 11.6 days of ray tracing time for the parking lot scenario of dataset.

E. OFDM radar primer

Sec. 3.1 detailed the modelling and synthesis flow MaxRay incorporates for vision and radio data. 6G network design is an active area of research whose details are in a state of flux. We, therefore, elaborate here on our radio data synthesis flow in order to enhance the clarity of MaxRay's radio modelling and assumptions.

Fig. 4 depicts a simplified block diagram of our 6G cellular system with sensing support. This 6G model consists of two simulation flows: (a) propagation via ray tracing, and (b) OFDM-based basestation signal processing.

(a) Propagation. The basestation transmits OFDM signals. These OFDM signals interact with the synthetic environment of Blender through a set of complex propagation phenomena. As such, backscatter signals captured at the basestation receiver chain enable radar detection. For synthesising these backscatter signals, MaxRay uses high-fidelity radio ray tracing. Specifically, MaxRay (i) implements geometric radio ray casting within Blender, (ii) calculates the propagation losses of these rays upon interacting with the synthetic environment model, and (iii) induces approperiate Doppler effects that correspond to moving objects (see Fig. 4). The propagation model concludes by presenting "environmentally-modulated" OFDM signals back to the basestation model.

(b) Basestation. In 6G networks, sensing is to be supported at the *physical layer*, unlike earlier attempts for opportunistically using standard wireless channel estimates for sensing [4]. For this to happen, the basestation transmits OFDM signals and then receives them back "modulated" by environmental effects. Specifically, the echoes backscattered from objects in the environments are received back at the basestation *coherently* w.r.t. the local oscillator of the receive chain. This coherent transceiver is illustrated in Fig. 4 as a *coupling* between the transmit and receive analogue chains. The modified transceiver remains compatible with standard downlink and uplink communications.

Radar processing in MaxRay is then implemented on top of OFDM communication signals. OFDM is the workhorse of modern communication systems. Using OFDM radar makes sensing much more amenable to integration in communication systems. Specifically, OFDM radar processing begins after we obtain wireless channel estimates from the OFDM demodulator as shown in Fig. 4. OFDM radar finally outputs the sensing primitives (i.e., the heatmaps) that our radio-visual SSL uses.

Note that joint communication and sensing in 6G as illustrated in Fig. 4 is non-trivial. Concretely, 6G requires (a) new hardware at cellular basestations, as well as (b) new resource allocation protocol involving space, time, frequency, and power optimisations of the network [23]. For completeness, the following describes briefly the signal processing principles of OFDM radar [7] as implemented in MaxRay.

OFDM radar signal processing. For N_{symb} known transmitted symbols $\mathbf{X}$, the channel can be estimated from the received data $\mathbf{Y}$ according to

$$\mathbf{H}^{k,n} = \frac{\mathbf{Y}^{k,n}}{\mathbf{X}^{k,n}} \tag{7}$$

where $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{N_{\text{sub}} \times N_{\text{symb}}}$, N_{sub} is the number of subcarriers, k and n are respectively subcarrier and symbol indices, and division is element-wise for efficient single tap equalisation. The signal traverses a finite number of paths L to the receiver. As such we can write the channel according to

$$\mathbf{H}^{k,n} = \sum_{\ell=0}^{L} \rho_{\text{loss}} \underbrace{e^{j2\pi n T_0 f_\ell}}_{\text{Doppler}} + \underbrace{e^{j2\pi k d_\ell / c_0 \Delta f}}_{\text{distance}} + \eta^{k,n} \tag{8}$$

where f_ℓ is the per-path Doppler-induced phase shift that modulates OFDM symbols, and T_0 is the symbol duration. The distance travelled induces another phase shift that affects OFDM subcarriers, with Δf being the subcarrier spacing, and $\eta^{k,n} \sim \mathcal{N}(0, \sigma^{k,n})$ is zero-mean Gaussian noise. Eq. (8) tells us that the phase information per path (i.e., bounced off some object) can be used to determine the relative speed and range of objects encountered during propagation. The angle of an object can also be estimated by phase processing multiple $\mathbf{H}^{k,n}$ across antennae, i.e., spatial processing. Orthogonality in OFDM allows for efficient periodogram estimation of the channel as [7]

$$\mathbf{P}^{k,n} = \left| \sum_{m=0}^{N_{\text{smyb}}-1} \left(\sum_{p=0}^{N_{\text{sub}}-1} \mathbf{H}^{p,m} e^{-j2\pi \frac{pn}{N_{\text{symb}}}} \right) e^{j2\pi \frac{mk}{N_{\text{sub}}}} \right|^2 \tag{9}$$

using the fast Fourier transform (FFT) over symbols, and the inverse FFT over subcarriers. This gives rise to peaks at the corresponding distance and speed of respective objects.

The above treatment shows that OFDM signalling for communication can be reused for implementing radar techniques for sensing. Integrating such sensing functionality alongside communications, with acceptable tradeoffs, is an active area of research for 6G networks.

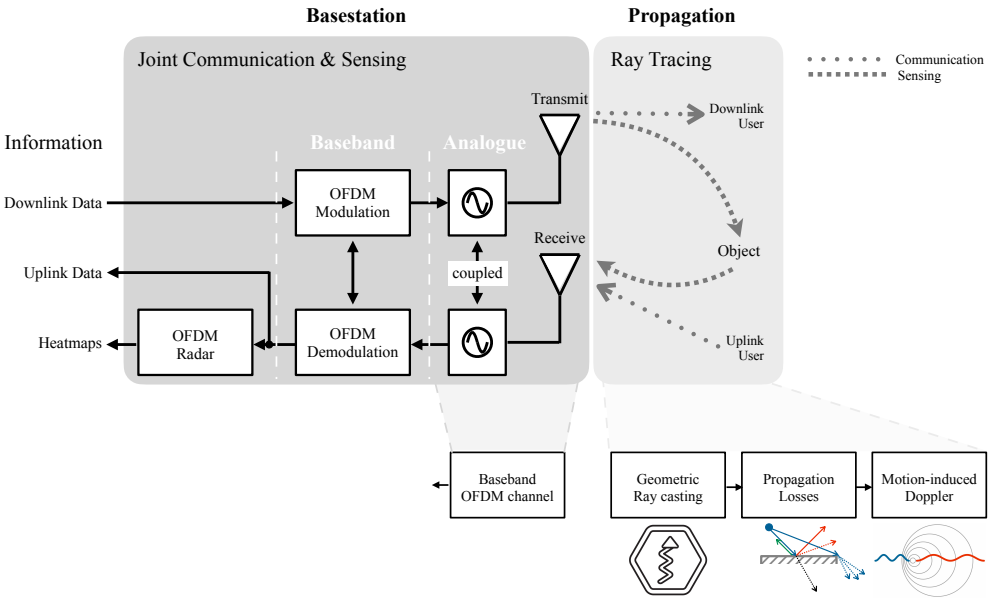


Figure 4. 6G network model with sensing support used in MaxRay. The model consists of two subsystems: (a) basestation and (b) propagation. The basestation implements OFDM radar signalling within a phase coherent signal processing architecture. We simulate the OFDM channel in baseband. Propagation simulations are performed via geometric ray tracing. We extensively model propagation losses (e.g., diffraction, backscatter, reflection, scatter, penetration, etc.) as well as Doppler effects.

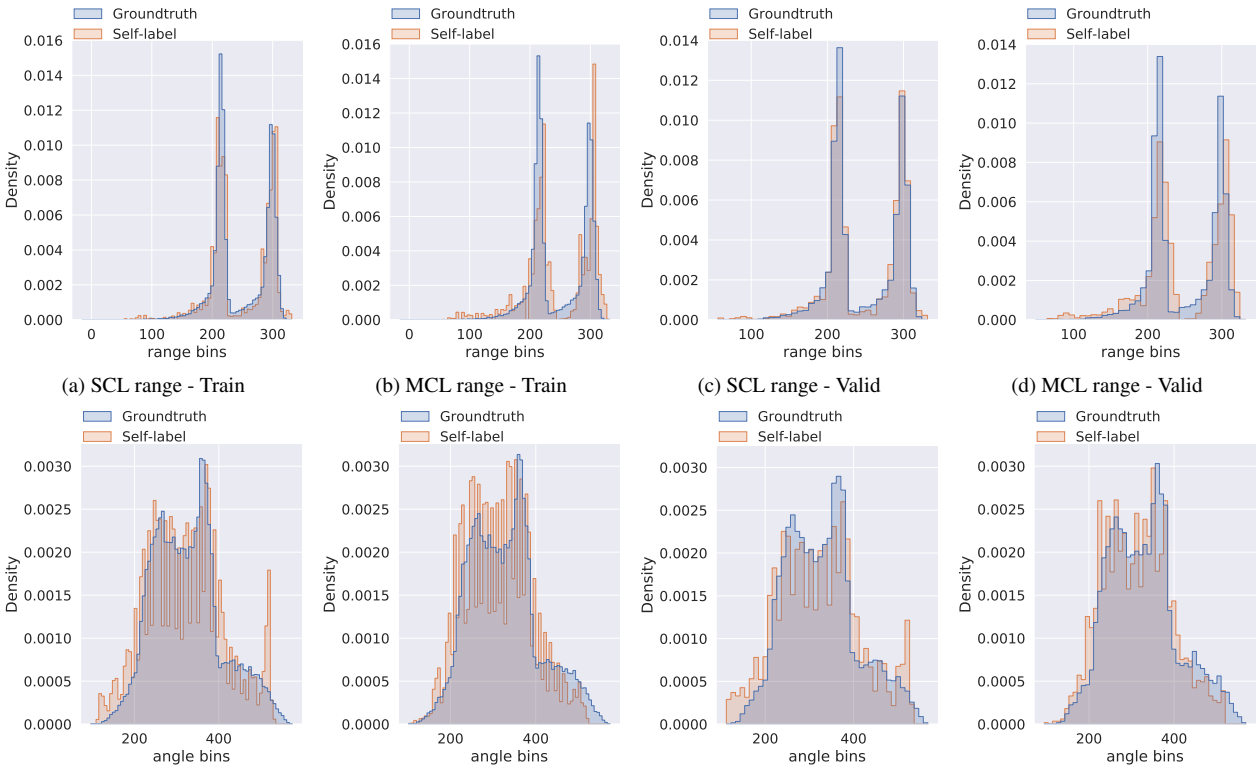
F. Self-labels analysis

Further to discussions in Sec. 6.2, Fig. 5 details the empirical histograms that characterise SCL's and MCL's self-label deviation from groundtruth labels.

G. Additional results

Table 1. Yolov5 performance on MaxRay.

	mAP_{50}	IoU-0.5
MaxRay	100	0.9374



(a) SCL range - Train (b) MCL range - Train (c) SCL range - Valid (d) MCL range - Valid

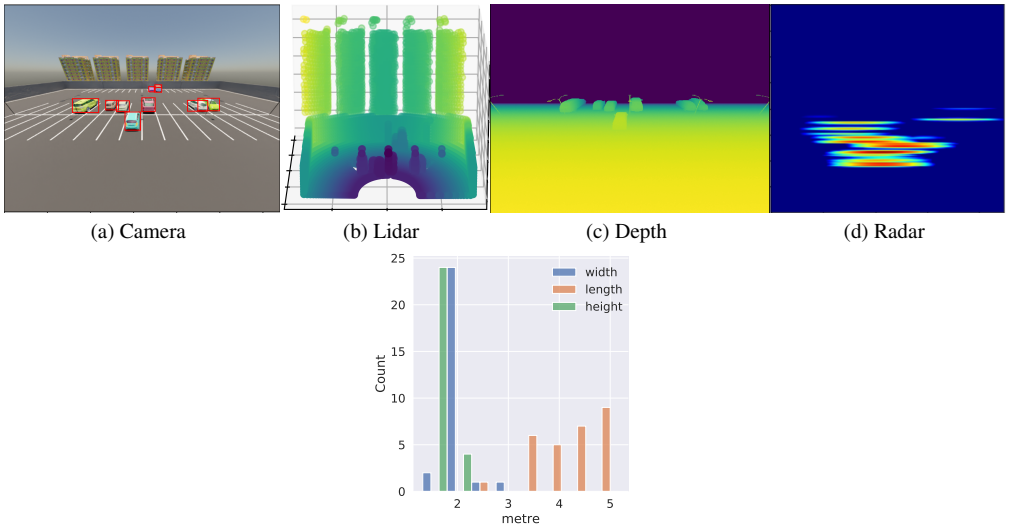
(e) SCL angle - Train (f) MCL angle - Train (g) SCL angle - Valid (h) MCL angle - Valid

Figure 5. Groundtruth label distributions for target range and angle bins, along with their respective self-label distributions overlaid. SCL and MCL behave differently in their ability to derive self-labels. Such distributions are depicted for both training and validation sets.

H. MaxRay illustrations

Further to Sec. 3, Fig. 6 shows snapshot examples of various data entries from MaxRay. The examples belong to the parking lot scenario supported in phase 1 of dataset release. There are currently up to 20 random and identically distributed cars. The statistics of the dimensions of these cars are depicted in Fig. 6e.

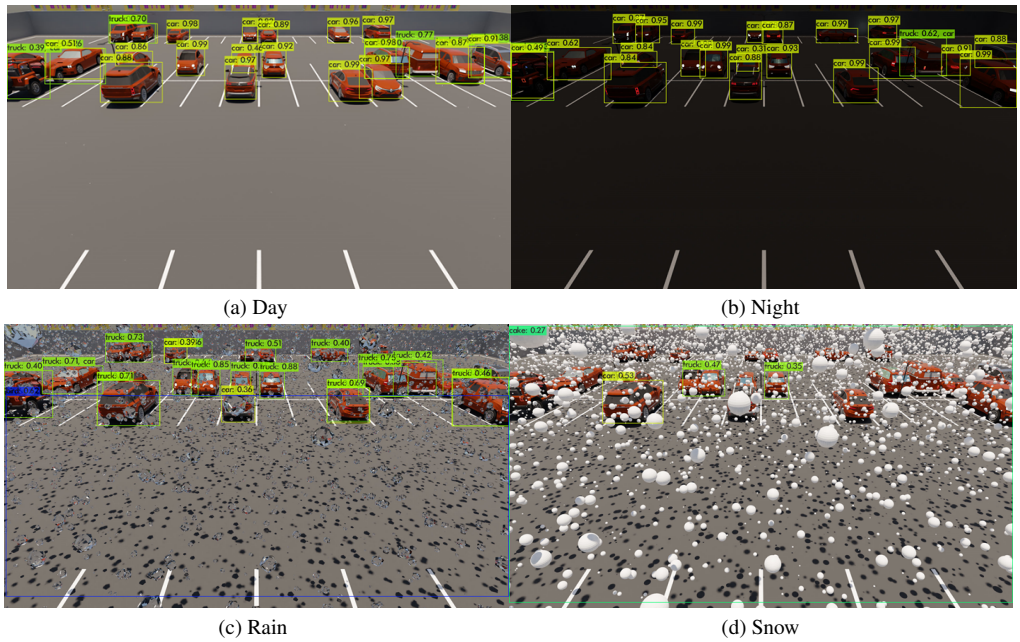
Fig. 7 shows examples of different lighting and weather conditions supported in MaxRay. Notice how the reliability of vision detection (Yolo v5 here) drops under unfavourable conditions, particularly snow as depicted in Fig. 7d.



(a) Camera (b) Lidar (c) Depth (d) Radar

(e) Car dimension statistics

Figure 6. Example of different modalities supported in MaxRay. From left to right: Camera image with bounding boxes, Lidar point cloud with object type, Depth image with range, Radar heatmap with groundtruth coordinates. Distribution of car dimensions throughout dataset is also illustrated.



(a) Day (b) Night

(c) Rain (d) Snow

Figure 7. Example of different lighting and weather conditions supported in MaxRay.

I. In-depth NNI explanation

Neural Network Intelligence (NNI) is an automatic machine learning (AutoML) tool that enables the systematic exploration of the optimisation space. We list the parameters and neural architectures we considered during AutoML optimisation in the Tab. 2. The optimal search choice is shown under the right-most column.

Fig. 8 depicts the final architecture of the supervised network.

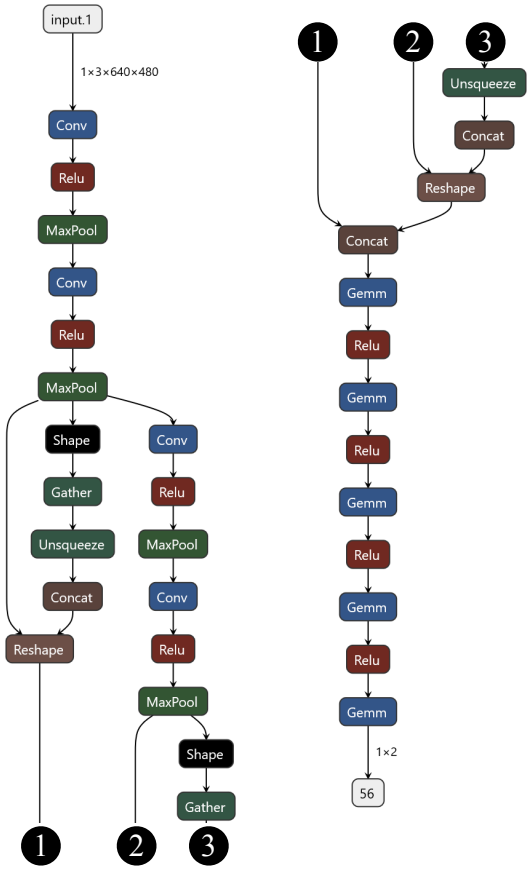


Figure 8. Final neural architecture of the supervised localiser network. ❶, ❷, and ❸ denote vertical network break and continuation.

Table 2. NNI optimisation architecture & parameters

Parameter	Explanation	Selection	Values	Best net chosen
lr	Learning rate	Choice	0.0001, 0.001, 0.01	0.001
momentum	Momentum for optimizer	Uniform	0.8, ..., 1	0.948985588
act_func	Activation function of conv layer	Choice	"ReLU", "LeakyReLU", "Sigmoid", "Tanh", "Softplus"	ReLU
optimizer	Optimizer type	Choice	"SGD", "Adam"	Adam
loss_func	Loss function for training only	Choice	"MSE", "L1"	MSE
c1_size	Convolutional kernels of c1 layer	Choice	4, 8, 16, 32, 64	8
c2_size	Convolutional kernels of c2 layer	Choice	4, 8, 16, 32, 64	16
c3_size	Convolutional kernels of c3 layer	Choice	4, 8, 16, 32, 64	8
c4_size	Convolutional kernels of c4 layer	Choice	4, 8, 16, 32, 64	32
k1_size	Kernel size of c1 layer	Choice	2, 3, 4	4
k2_size	Kernel size of c2 layer	Choice	2, 3, 4	3
k3_size	Kernel size of c3 layer	Choice	2, 3, 4	2
k4_size	Kernel size of c4 layer	Choice	2, 3, 4	4
s1_size	Stride of c1 layer	Choice	1, 2	2
s2_size	Stride of c2 layer	Choice	1, 2	2
s3_size	Stride of c3 layer	Choice	1, 2	2
s4_size	Stride of c4 layer	Choice	1, 2	1
lin1_size	Linear layer 1	Choice	128, 256, 512	128
lin2_size	Linear layer 2	Choice	16, 32, 64, 128, 256	16
lin3_size	Linear layer 3	Choice	16, 32, 64, 128, 256	64
lin4_size	Linear layer 4	Choice	64, 182, 256	64

J. Dataset comparison

Tab. 3 is a verbose version of Tab. 1 presented in Sec. 2. Further, Tab. 4 summarises the properties of CRUW and how it compares to our MaxRay dataset.

Table 3. Radio-visual datasets.

	Application		Resolution			# of data points					
Dataset	Automotive	6G	Range	Azimuth	Elevation	Total	Labelled	Frame rate	Groundtruth	Radar	Reconfigurability
CRUW [21]	✓	✗	23cm	15°	—	396k[1]	260k[1]	30	Camera	FMCW	✗
Carrada [17]	✓	✗	20cm	15°	—	12.7k	7.2k	10	Camera	FMCW	✗
AIODrive [22]	✓	✗	N/A	N/A	N/A	100k	100k	10	Synthetic	N/A	✗
RADIATE [19]	✓	✗	17.5cm	1.8°	1.8°	200k	44k	N/A	Camera	FMCW	✗
Oxford Radar RobotCar [6]	✓	✗	4.38cm	0.9°	—	240k	—	4	N/A	FMCW	✗
RADDet [25]	✗	✓	19.5cm	15°	30°	10.2k	10.2k	10	Camera	FMCW	✗
DeepSense [3]	✗	✓	60cm	15°	30°	WIP[2]	WIP[2]	10	Camera+Lidar	FMCW	✗
MaxRay*	✗	✓	18.75cm	6.75°	—	30k	30k	30	Synthetic	OFDM	✓

[1] only a fraction available publicly.
[2] work-in-progress: dataset scenarios are being released.
* MaxRay is the only 6G synthetic dataset, and is the only reconfigurable dataset.

Table 4. Comparison between MaxRay and CRUW. CRUW* requires preprocessing for integration into our radio-visual SSL algorithm.

Entry	MaxRay	CRUW	Preprocessing
Camera	30 FPS @ 640×480 pixels	30 FPS @ 1440×1080 pixels	Linear downscaling to 640×480
Radio	OFDM Radar @ 800MHz BW dense 16×16 antenna array	2× FMCW Radar @ 1250MHz BW sparse 4×2 antenna array	Radar range filtered to 5-30m, and periodogram upsampled
Range resolution	18.75cm	23cm	
Angular resolution	6.75°	15°	
Radio groundtruth	Perfect high-fidelity ray tracing	Camera-radar fusion (RODNet labels)	None
Vision groundtruth	Perfect target bounding box	Yolov5 target bounding box	None
Scenario	Parking lot (see Sec. **??**)	Parking lot (see [21])	
# of data points	30k	9k	

* https://www.cruwdataset.org

K. Datasheet

We document in Tab. 5 various aspects of our radio-visual dataset according to the specifications stipulated in [12].

Table 5. Dataset datasheet

Motivation	
For what purpose was the dataset created?	To facilitate radio-visual SSL research for 6G sensing.
Who created the dataset and on behalf of which entity?	Bell Labs Core Research (BLCR) on behalf of Nokia.
Who funded the creation of the dataset?	Nokia.
Composition	
What do the instances that comprise the dataset represent?	heatmap-image pairs sampled from a parking lot scenario.
How many instances are there in total?	30,000 labelled for parking lot.
Does the dataset contain all possible instances or is it a sample of instances from a larger set?	All.
What data does each instance consist of?	Radio heatmaps are range-azimuth description of the environment and RGB images are their visual pairs.
Is there a label or target associated with each instance?	Object groundtruth coordinates for radio and bounding boxes for vision.
Is any information missing from individual instances?	No.
Are relationships between individual instances made explicit?	Correspondence between each radio-visual pair.
Are there recommended data splits?	80:20 train-validation split for downstream regression.
Are there any errors, sources of noise, or redundancies in the dataset?	Not at the data instance level; It is a synthetic dataset. At the radio signal level, high-fidelity propagation modelling captures non-trivial sources of noise such as clutter and fading.
Is the dataset self-contained, or does it link to or otherwise rely on external resources?	Self-contained.
Does the dataset contain data that might be considered confidential?	No.
Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?	No.
Does the dataset identify any subpopulations?	No.
Is it possible to identify individuals, either directly or indirectly from the dataset?	No.
Does the dataset contain data that might be considered sensitive in any way?	No.
Collection Process	
How was the data associated with each instance acquired?	Synthesised using CAD tools.
What mechanisms or procedures were used to collect the data?	Ray-tracing for radio and rendering for vision.
If the dataset is a sample from a larger set, what was the sampling strategy?	N/A.
Who was involved in the data collection process and how were they compensated?	Nokia employees under full-time employment.
Over what timeframe was the data collected?	Data generation took several months of in-house development effort.
Were any ethical review processes conducted?	N/A.
Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?	N/A.
Were the individuals in question notified about the data collection?	N/A.
Did the individuals in question consent to the collection and use of their data?	N/A.
If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?	N/A.
Has an analysis of the potential impact of the dataset and its use on data subjects been conducted?	N/A.
Preprocessing/cleaning/labeling	
Was any preprocessing/cleaning/labeling of the data done?	No.
Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?	N/A.
Is the software that was used to preprocess/clean/label the data available?	N/A.
Uses	
Has the dataset been used for any tasks already?	Mainly radio-visual SSL research disclosed in this paper.
Is there a repository that links to any or all papers or systems that use the dataset?	N/A.
What (other) tasks could the dataset be used for?	This is a 1st radio-visual SSL work, and future research would build on our ideas and/or investigate alternative approaches, e.g., for more discriminative radio signals obtained from finer angular resolutions.
Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?	No.
Are there tasks for which the dataset should not be used?	N/A.
Distribution	
Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?	Yes.
How will the dataset will be distributed?	Hosted on a public website.
When will the dataset be distributed?	2023.

Cont. Tab. 5

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?	Yes.
Have any third parties imposed IP-based or other restrictions on the data associated with the instances?	No.
Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?	No.
Maintenance	
Who will be supporting/hosting/maintaining the dataset?	Nokia Bell Labs.
How can the owner/curator/manager of the dataset be contacted?	Email.
Is there an erratum?	No.
Will the dataset be updated?	Yes.
If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?	N/A.
Will older versions of the dataset continue to be supported/hosted/maintained?	Yes.
If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?	We will provide reference Blender files which can be modified to model different environments. Our radio raytracing is however proprietary and cannot be released. To work around this, users could licence equivalent commercial radio raytracers in order to generate paired radio heatmaps from Blender's 3D models.

End Tab. 5

References

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

[3] 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. 9

[4] Mohammed Alloulah and Howard Huang. Future millimeter-wave indoor systems: A blueprint for joint communication and sensing. *Computer*, 52(7):16–24, 2019. 4

[5] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision*, pages 435–451, 2018. 3, 4

[6] 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. 9

[7] 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. 4, 5

[8] 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. 3, 4

[9] 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. 3

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

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

[12] 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. 10

[13] 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. 3, 4

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

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[16] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. *arXiv preprint arXiv:2009.09805*, 2020. 4

[17] 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. 9

[18] 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. 3

[19] 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. 9

[20] 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. 2

[21] 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. 9

[22] 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. 9

[23] 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. 4

[24] 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

[25] 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. 9