

Self-Supervised Radio-Visual Representation Learning for 6G Sensing

Mohammed Alloulah^{*†}, Akash Deep Singh^{§†‡}, Maximilian Arnold[†]
[†]Bell Labs [‡]UCLA

Abstract—In future 6G cellular networks, a joint communication and sensing protocol will allow the network to perceive the environment, opening the door for many new applications atop a unified communication-perception infrastructure. However, interpreting the sparse radio representation of sensing scenes is challenging, which hinders the potential of these emergent systems.

We propose to combine radio and vision to *automatically learn* a radio-only sensing model with minimal human intervention. We want to build a radio sensing model that can feed on millions of *uncurated* data points. To this end, we leverage recent advances in self-supervised learning and formulate a new *label-free* radio-visual co-learning scheme, whereby vision trains radio via cross-modal *mutual information*. We implement and evaluate our scheme according to the common linear classification benchmark, and report qualitative and quantitative performance metrics. In our evaluation, the representation learnt by radio-visual self-supervision works well for a downstream sensing demonstrator, and outperforms its fully-supervised counterpart when less labelled data is used. This indicates that self-supervised learning could be an important enabler for future *scalable* radio sensing systems.

I. INTRODUCTION

Imagine a world where cellular basestations are not only enabling untethered connectivity everywhere, but are also actively sensing the environment. Supporting joint communication and sensing (JCAS) in next generation networks constitutes a step change in cellular systems functionality, and may well be what sixth generation (6G) networks are all about [1]–[3]. That is, ubiquitous cellular basestations will allow us to detect the presence of objects of interest, estimate their shape, location, and movement patterns, whilst relying solely on radio signals. With these newfound sensing capabilities, we will be able to gather urban analytics at unprecedented scales, e.g., for monitoring busy intersections in cities in support of safety-critical autonomous driving.

Concretely, 6G networks will enable a host of outdoor sensing applications ranging from traffic monitoring and parking spot detection, around-the-corner vehicle detection, to drone and pedestrian identification and localisation [1]. Given the privacy preserving advantage of radio sensing, and coupled with the ubiquity of cellular infrastructure, it is not unreasonable to expect that many city-scale analytics services would turn to radio spectrum for the provision of sensing primitives.

The sheer amount of sensory data produced by 6G basestations provides both opportunities and challenges for building

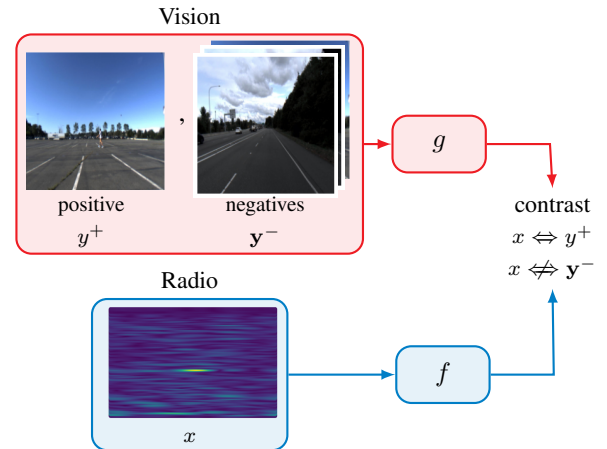


Fig. 1. Contrastive radio-visual learning. Image y^+ is a positive pair of radar heatmap x because they are recorded simultaneously, while other randomly-sampled images y^- are negative pairs.

scalable and robust sensing applications. On the one hand, ubiquitous basestations can easily collect data from diverse environmental conditions, e.g., from dense urban or sparser suburban radio channels. On the other hand, managing such data *diversity* and *scale* for producing highly performant sensing models becomes hard. Particularly, state-of-the-art deep learning (DL) sensing models have typically required manual annotation of empirical data. For example, ImageNet, the flagship 14 million images dataset for object recognition, took 22 human years to annotate [4]. Clearly, the progress of radio sensing should not hinge on the availability of *laboriously annotated* data. Alternatively, radio sensing models could be trained on synthetic RF data using ray-trace simulations [5]–[7]. Since the environment is simulated, we can specify the sensing scenario (i.e., label) first and then generate the corresponding data. However, synthetic data would only provide a 1st-order approximation of the real-world that is unlikely to meet the quality, scale, and richness requirements of highly-performant deep radio sensing models. Current synthetic data generation techniques for RF work well for simple environments, but their complexity-bounded performance degrades as we move to more realistic scenes which mirror the real-world [8]–[10].

Recent advances in self-supervised learning have demonstrated a viable alternative to label-intensive supervised learning [11]. Self-supervised methods provide two key advantages. Firstly, they do not require labelled data. Instead of solving a

*Correspondence to alloulah@outlook.com

§Work was conducted when the author was under Bell Labs employment.

specific learning task, we solve an auxiliary task that helps the model learn (by *proxy*) the underlying structure of the data. Secondly, self-supervised methods learn more granular and generalisable discriminative features than task-specific supervised methods [12]. This is because self-supervision generates higher information rate per sample than supervised learning as captured eloquently in Yann LeCun’s cake analogy [13].

In this paper, we draw inspirations from recent Natural Language Processing (NLP) breakthroughs [14], [15] and Computer Vision (CV) successes [12], [16] and propose a self-supervised method for radio representation learning. Specifically, we formulate an auxiliary prediction task that contrasts RF heatmaps to camera images. We show that such auxiliary *contrastive* prediction promotes learning powerful radio neural networks that can be specialised for specific sensing applications. We argue that self-supervision for radio signals is an important enabler for emergent 6G sensing systems.

The paper describes the following contributions:

- To the best of our knowledge, this paper is the first to demonstrate successful self-supervised representation learning for radio signals using cross-modal mutual information (MI) with camera images. A loss function bounded by MI enables such *label-free* learning.
- We specialise the self-supervised radio model for a downstream task that classifies objects of interest in scenes from empirical measurements.
- We characterise our self-supervision-based model using a number of quantitative and qualitative metrics. When evaluated against a fully-supervised baseline, we find that the self-supervised model outperforms its supervised counterpart when less labelled data is used for training.

II. OVERVIEW & INTUITION

Radio heatmaps are sparse representations of the environment. They are difficult to interpret and label by humans, because the underlying geometry of the physical space they capture is not straightforward to decipher.

Fig. 1 depicts our proposed radio-visual co-learning scheme. The idea is to (i) match a radio heatmap with its corresponding groundtruth image and (ii) contrast this true radio-vision pairing against a number of false pairings using images sampled at random. The very act of emphasising the *contrast* between positive and negative radio-visual pairings gives rise to a robust learning signal as will be formalised and explained later. Further, this simple yet powerful *contrastive* learning can be self-administered, assuming only synchronised radio-visual measurements (i.e., true pairings). Hence, the *contrastive* learning scheme is also *self-supervised* and needs no labels.

III. PRIMER

We discuss next the foundational concepts of our self-supervised radio-visual learning.

A. Self-supervised learning

Unlike supervised learning where we need to manually annotate the data, self-supervised learning leverages intrinsic

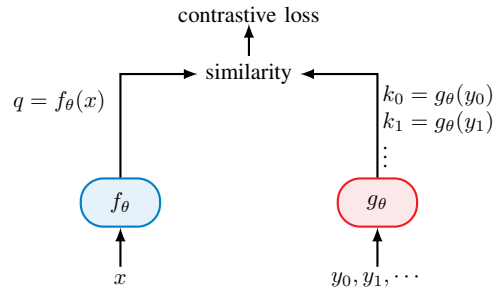


Fig. 2. Contrastive loss for self-supervised learning.

labels which can be “contrived” using the data itself. The most successful form of self-supervised learning to-date is in NLP, whereby a language model tries to predict randomly masked words in sentences. Surprisingly, this simple contrived prediction acts as a signal that forces the network to learn fundamental aspects that characterise language altogether, e.g., word association and sentence-to-sentence context. In addition to its groundbreaking applications in NLP [14], [15], self-supervised learning has also seen numerous successes in domains such as vision [17] and multi-modal learning [18]. For example, visual self-supervised learning can be made to work by predicting whether randomly sampled image patches are related or not [17].

B. Contrastive loss

For self-supervision to be effective, the signal that drives learning should also accentuate the difference (i.e. *contrast*) between true and false predictions and their data samples. The contrast between a positive datum and negative data is at the heart of learning good self-supervised models. There are many incarnations to such a contrastive loss [19]–[21]. InfoNCE is one contrastive loss designed to drive learning in a way that preserves the mutual information (MI) between inputs [19]. Concretely, referring to Fig. 2, let x and y_i be two input signals encoded by two neural networks f_θ and g_θ such that $q = f_\theta(x)$ and $k_i = g_\theta(y_i)$, assuming some weights parametrisation θ . With each x , use $K + 1$ samples of y of which one sample y^+ is a true match to x and K samples $\{y_i^-\}_{i=0}^{K-1}$ are false matches. Then the contrastive loss \mathcal{L}_c is [20]

$$\mathcal{L}_c = -\mathbb{E}_{x,y} \log \left[\frac{\exp(q \cdot k^+ / \tau)}{\exp(q \cdot k^+ / \tau) + \sum_i \exp(q \cdot k_i^- / \tau)} \right] \quad (1)$$

where \cdot is the dot product operator, $k^{+/-} = g_\theta(y^{+/-})$ are encodings that correspond to true and false y signals, vector $\mathbf{k}^- = \{k_i^-\}_{i=0}^{K-1}$ holds K false encodings, and τ is a temperature hyper-parameter.

The contrastive loss of Eq. (1) has enabled a wave of progress in self-supervised image classification that inches ever closer and at times exceeds the performance of fully supervised methods. Notably, SimCLR [22] and MoCo [16] are recent examples of bleeding-edge self-supervised vision systems.

The contrastive loss \mathcal{L}_c has information-theoretic connections to the MI between q and k (vis-à-vis x and y). This can be seen noting the following. First, the expression $\exp(q \cdot k/\tau)$ models a positive real score between q and k , which is proportional to the (unnormalised) density ratio

$$\exp(q \cdot k/\tau) \propto \frac{p(k|q)}{p(k)} \quad (2)$$

Second, Eq. (1) has the form of a $(K + 1)$ -way softmax classifier that matches q to its positive pair k^+ . Using Eq. (2) in Eq. (1), Oord et al. show that \mathcal{L}_c preserves the MI between q and k [19]. Specifically, \mathcal{L}_c becomes lower bounded by the MI between the encoded inputs¹ and that using more negative samples K enhances the learnt representation [19], [23]

$$\mathcal{L}_c(q, k^+, \mathbf{k}^-) \geq \log(K) - I(q; k) \quad (3)$$

$$\mathcal{L}_c^{\text{optimal}}(q, k^+, \mathbf{k}^-) = \log(K) - I(x; y) \quad (4)$$

where $I(q; k)$ & $I(x; y)$ denote respective MI's. For a self-contained treatment, we recapitulate the derivations of Eq. (3) & (4) in supplementary Sec. IX.

C. Cross-modal co-supervision

The notion of self-supervision can also be applied across different modalities for data with inherent multi-modal signals, e.g., video. The sheer amount of video freely available on the internet makes this proposition increasingly appealing. For instance, there is a body of literature that studies how representations for audio and vision can be learnt *jointly*, e.g., [21], [24] (and references therein). Besides learning quality audio and vision features that are on par with supervised single-modal approaches, additional cross-modal tasks can be *automatically* instantiated (from the joint embeddings), e.g., synchronisation and object localisation [21], [25].

Similar to NLP and vision, audio-visual self-supervised systems use a contrived auxiliary task to promote cross-modal representation learning. Earlier works dealt with predicting the *correspondence* of an image and an audio spectrogram snippet using a binary cross-entropy loss, e.g., [24]. Audio-visual correspondence in this context is static, i.e., from a single measurement snapshot. More powerfully, recent works extend this correspondence to capture shared temporal dynamics between a video snippet and an audio snippet, e.g., [21].

IV. RADIO-VISUAL CONTRASTIVE LEARNING

We next turn to discussing how to adapt the concepts treated in Sec. III for radio-visual self-supervised learning.

A. Pre-training

Our objective is to leverage the co-occurrence of events in the radio and vision domains and formulate a learning scheme that requires no labels. That is, the onset of an event of interest across radio and vision will give rise to MI that can be captured and used as an automatic learning signal. The judicious sampling of positive and negative data according to Eq. (1) extracts cross-modal MI.

¹which are in turn lower bounded by the MI between the raw inputs

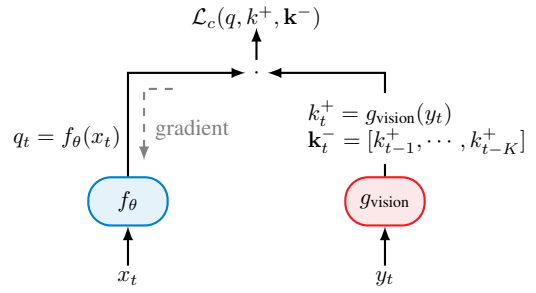


Fig. 3. Radio-visual contrastive loss.

Data collation. In practice, we operate a pair of camera and radar devices that are synchronised to continuously collect measurements. The synchronisation timestamps allow us to construct the positive and negative data sample set $S_d = \{x_t, y_t^+, y_t^-\}$ and their respective encodings $S_e = \{q_t, k_t^+, \mathbf{k}_t^-\}$, at a given time instance t . The resultant dataset can be further filtered by *light*² object detection algorithms in the vision domain to: concentrate on learning good encodings for specific objects of interest (e.g., pedestrians, cyclists, and cars), balance their distributions, etc.

Architecture. Our machine learning model consists of two parts: (i) a radar branch and (ii) a camera branch. During training, the two branches interact through the contrastive loss of Eq. (1).

Representation learning. Let (x_t, y_t) be a pair of a radio heatmap and a corresponding camera image, as depicted in Fig. 3. A neural network with two branches for radio and vision ingests (x_t, y_t) . For the vision subnetwork, we use a pretrained convolutional model to encode y_t , yielding a positive encoding $k_t^+ = g_{\text{vision}}(y_t)$. The vision subnetwork is frozen and there is no gradient flowing back to it during training. This aspect is different to self-supervised vision systems [16], [22]. The idea is to leave the vision encoder maximally consistent during training, which would significantly enhance the contrastive learning of the radar subnetwork. Under our cross-modal settings, it also makes sense to derive learning from the higher-entropy vision modality. We adopt a queue of negative encodings scheme proposed by He et al. in their momentum contrast (MoCo) system [20]. The negative encodings queue \mathbf{k}_t^- is constructed by utilising past image encodings that do not correspond to the current radar heatmap³, i.e., $\mathbf{k}_t^- = [k_{t-1}^+, \dots, k_{t-K}^+]$. The queue is a very efficient mechanism to enhance contrastive learning (cf. loss bound in Eq. (3)) under practical compute and memory constraints. Throughout training, a radar heatmap x_t is encoded according to $q_t = f_{\theta}(x_t)$, and a gradient is computed to optimise for q_t 's similarity to its positive image pair k_t^+ and dissimilarity to K negative images in \mathbf{k}_t^- . This K -way contrastive learning can be trivially extended for the mini-batch settings of stochastic gradient descent (SGD).

²i.e., without laborious annotations such as bounding boxes

³by construction of the dataset

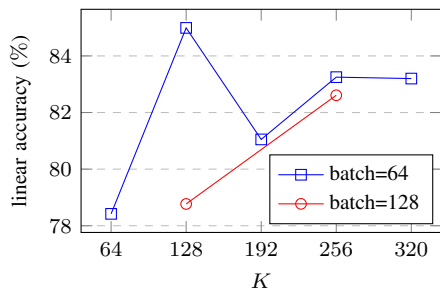


Fig. 4. Linear classification as a function of number of negatives K , for two batch sizes. Generally, linear classification is enhanced with contrastive learning against more negatives, which agrees with the theoretical MI bound.

B. Fine-tuning

Contrastive representation learning concludes by arriving at a good f_θ than can be used in a variety of sensing applications. A task-specific model is then constructed using the output of f_θ as features combined with a classifier head. Typically, sensing tasks also fine-tune (i.e., bias) these pre-trained representations towards application-specific criteria.

C. Implementation details

For both the vision and radio branches, we use pre-trained VGG-16 models [26]. The VGG-16 model takes a image of $3 \times 224 \times 224$ dimensions as input. For the radar branch, the heatmap is resampled and replicated 3 times for a valid input. The vision branch is frozen and only the radar branch is trained. We slice the VGG-16 model at the 4th layer of its classifier stage to obtain a 4096-dimensional feature vector, i.e., $q, k \in \mathbb{R}^{4096 \times 1}$. We use MoCo’s shuffling trick to mitigate against batch normalisation (BN) statistical issues [20]. BN’s statistical artefacts have also been observed in other works, e.g., [27], [28].

V. EXPERIMENTS

A. Dataset

We use the Camera-Radar of the University of Washington (CRUW) dataset [29] for evaluating our radio-visual self-supervised learning. The dataset contains around 47,000 image-heatmap pairs in a variety of settings such as empty parking lots, driving on roads, and in front of buildings. Objects present in the dataset are: pedestrians, cyclists, and cars. An example pair is depicted in Fig. 1.

Downstream task. We balance the class distribution of a subset of the dataset such that there are 4 non-overlapped classes in any given scene: empty, pedestrian, cyclist, and car. The resultant 4-category sensing task maps onto the envisioned 6G use cases, e.g., pedestrian versus car discrimination for safety applications [1]. We use a standard 80/20 train/test split.

B. Training

Similar to MoCo [16], [22], we use a stochastic gradient descent (SGD) optimiser for contrastive pre-training. The learning rate starts at 0.03 with an SGD weight decay of $1e-4$ and momentum of 0.9. The learning rate decays according to a cosine schedule. We use two mini-batch sizes: 64 and 128.

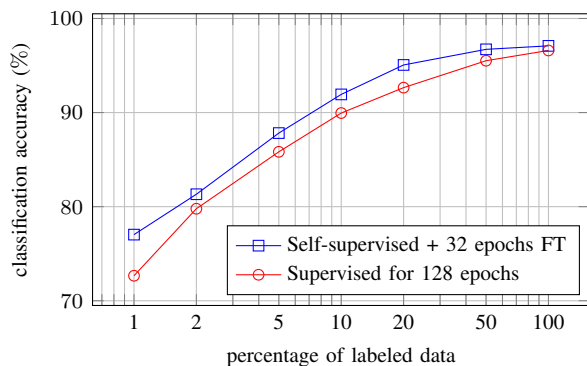


Fig. 5. Efficiency measured w.r.t. accuracy versus percentage of labels used for fine-tuning (FT). Self-supervised efficiency is compared to its fully-supervised counterpart. Self-supervised contrastive learning needs 32 epochs to exceed the level of performance of fully supervised training using 128 epochs: a factor of $4\times$ reduction in training.

We train in a distributed fashion on 4 GPUs for 400 epochs, which takes ~ 12 hours for VGG-16. For training the fully-supervised baseline, fine-tuning, and the linear classifier, we use a $1e-4$ learning rate. For the fully-supervised baseline, we train for 300 epochs. For fine-tuning and the linear classifier, we train 32 epochs only.

C. Evaluation protocol

Linear classification. We benchmark the quality of our self-supervised radio representation learning using the common linear classification protocol. Under the linear classification protocol, we: (1) pre-train using contrastive learning, (2) freeze the learnt features, and then (3) train a supervised classifier with a linear fully-connected (FC) layer whose output is normalised using a softmax.

Efficiency. We investigate the data and training efficiency of the contrastive representation f_θ . Data efficiency refers to the amount of labels required to build a downstream task using f_θ . Training efficiency refers to the number of epochs required to converge to a downstream representation with competitive performance. Both metrics are evaluated against their fully-supervised counterparts.

VI. RESULTS

A. Quantitative

We begin by examining the linear classification performance of self-supervised radio features.

Linear classification. Eq. (3) tells us that the contrastive loss benefits from increasing the number of negatives. We investigate the effect of queue size K on the linear classification accuracy in Fig. 4. For mini-batch size 64, we sweep K from 64 to 320 in 64 increments. For mini-batch size 128, $K \in [128, 256]$ only due to GPU resource constraints. Generally, apart from a blip in mini-batch 64 at $K = 128$, linear accuracy increases steadily with K for both mini-batches.

Efficiency. Fig. 5 depicts the label efficiency of self-supervised features and compares it to the fully-supervised efficiency. Using all dataset labels, both achieve a comparable accuracy

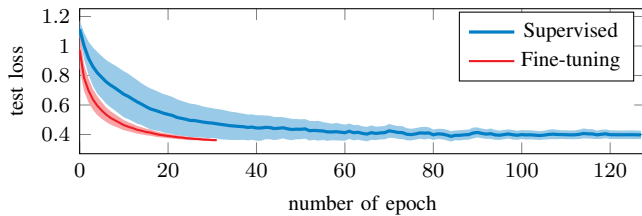


Fig. 6. Comparing the testing losses of the fully-supervised baseline and the fine-tuned model with early stopping for the downstream classification task.

of 97%. However, self-supervised features beat the supervised ones when we decrease the number of labels used for training. For instance, with 1% labels, we see a gap of about 3.5% in favour of self-supervision. This observation is generally in agreement with recent findings from the CV domain [12], which report on even more drastic advantages to self-supervision in the “low label density” regime.

In terms of training efficiency, the self-supervised features can reach the fully-supervised performance quite quickly within 32 epochs of fine-tuning. This is corroborated in Fig. 6 that analyses the testing loss during fine-tuning against its fully-supervised counterpart. This indicates that the self-supervised representation has indeed learnt good generalisable features. We expect much greater data efficiency gains under more elaborate classification settings, such as a 1000-category classification task [12]. Note that unlike fully-supervised systems, the initial training of self-supervised features would amortise over many different downstream tasks.

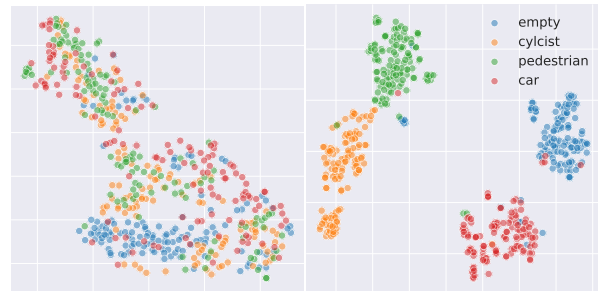
B. Qualitative

Learnt representation. As discussed in Sec. IV-C, we start the self-supervision of the radio branch from a pre-trained vision VGG network. Could this, rather than contrastive training, be responsible for the very good performance on the downstream task? After all, objects showing as blobs in radio heatmaps should in theory also benefit from many learnt vision filters, such as edge and contour detectors. To investigate this, we use t-SNE [30] to visualise the 128-dimensional features of the penultimate layer for (a) a pre-trained VGG combined with randomly initialised classification head, and (b) our fine-tuned contrastive model. Results are shown respectively in Fig. 7. Inspecting Fig. 7a, no structure in the latent space can be discerned in relation to our 4-category downstream task. In contrast, the latent space of our fine-tuned model exhibits clear clustering around the 4 categories of the downstream task as evident in Fig. 7b.

VII. RELATED WORK

Self-supervised learning, where the model is trained on implicit labels present within the dataset, has gained popularity in the fields of NLP and CV recently. In this section we highlight some key related works.

Computer Vision. Self-supervised CV models trained with simple auxiliary tasks lead to surprisingly good vision representation learning. Examples include solving puzzles [31], image colorisation [32], and predicting image rotations [33].



(a) pre-trained VGG with random head (b) fine-tuned contrastive radio model

Fig. 7. t-SNE visualisation of learnt representations.

For the latter for instance, self-supervised learning works by randomly rotating all the images in a dataset and asking the model to predict these rotations. Bleeding-edge self-supervision systems include SimCLR by Google [22] or SwAV by Facebook Research [34]. SimCLR generates two augmentations for the same image and tries to maximise the similarities between their latent representations. SwAV builds on SimCLR and adds multi-crop augmentation, i.e., two copies of an image at two different resolutions. After training on auxiliary tasks, the learnt features are then fine-tuned for a given downstream task.

RF Sensing. Self-supervised learning has recently been applied to radar systems research. In [35], the authors propose a super-resolution method called Radar signal Reconstruction using Self-supervision (R2-S2) which improves the angular resolution of a given radar array without increasing the number of physical channels. In [36], the authors propose R4Dyn which uses radars during training as a weak supervision signal, as well as an extra input to enhance the depth estimation robustness at inference time.

VIII. CONCLUSION

We asked if self-supervised learning can be used to train radio models using a radio-visual co-learning approach, and without relying on laboriously labelled data. We have found that this is indeed not only possible, but is also superior to the performance of a fully-supervised baseline, especially when less amount of labelled data is used. Our early findings are important for future 6G cellular systems where obtaining high-quality labelled data presents cost and scale challenges.

For future work, we plan to build on our early results and investigate more elaborate self-supervised formulations. Chiefly, we would like to (i) incorporate temporal mutual information in self-supervision and (ii) expand our evaluation to a larger and richer radio-visual dataset whereupon we can implement more complex downstream tasks and expose yet more compelling benefits to self-supervised learning for radio systems.

REFERENCES

- [1] T. Wild, V. Braun, and H. Viswanathan, “Joint design of communication and sensing for beyond 5G and 6G systems,” *IEEE Access*, vol. 9, pp. 30 845–30 857, 2021.

- [2] A. Bourdoux, et. al, “6G White Paper on Localization and Sensing,” 2020.
- [3] G. P. Fettweis and H. Boche, “6g: The personal tactile internet - and open questions for information theory,” *IEEE BITS the Information Theory Magazine*, pp. 1–1, 2021.
- [4] NYU, “Self supervised learning in computer vision,” <https://atcold.github.io/NYU-DLSP21/en/week10/10-1/>, 2021, accessed: 24-10-21.
- [5] A. D. Singh, S. S. Ram, and S. Vishwakarma, “Simulation of the radar cross-section of dynamic human motions using virtual reality data and ray tracing,” in *2018 IEEE Radar Conference (RadarConf18)*. IEEE, 2018, pp. 1555–1560.
- [6] A.-Y. Hsiao, C.-F. Yang, T.-S. Wang, I. Lin, and W.-J. Liao, “Ray tracing simulations for millimeter wave propagation in 5g wireless communications,” in *2017 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting*. IEEE, 2017, pp. 1901–1902.
- [7] H. T. P. da Silva, R. M. Duarte, M. S. de Alencar, and W. J. L. Queiroz, “Cell-free at millimeter wave frequency simulation using the ray tracing method,” in *2020 14th European Conference on Antennas and Propagation (EuCAP)*. IEEE, 2020, pp. 1–5.
- [8] V. Degli-Esposti, F. Fuschini, E. M. Vitucci, M. Barbiroli, M. Zoli, L. Tian, X. Yin, D. A. Dupleich, R. Müller, C. Schneider *et al.*, “Ray-tracing-based mm-wave beamforming assessment,” *IEEE Access*, vol. 2, pp. 1314–1325, 2014.
- [9] M. Lecci, P. Testolina, M. Giordani, M. Polese, T. Ropitault, C. Gentile, N. Varshney, A. Bodi, and M. Zorzi, “Simplified ray tracing for the millimeter wave channel: A performance evaluation,” in *2020 Information Theory and Applications Workshop (ITA)*. IEEE, 2020, pp. 1–6.
- [10] F. Fuschini, S. Häfner, M. Zoli, R. Müller, E. Vitucci, D. Dupleich, M. Barbiroli, J. Luo, E. Schulz, V. Degli-Esposti *et al.*, “Analysis of in-room mm-wave propagation: Directional channel measurements and ray tracing simulations,” *Journal of Infrared, Millimeter, and Terahertz Waves*, vol. 38, no. 6, pp. 727–744, 2017.
- [11] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [12] O. Henaff, “Data-efficient image recognition with contrastive predictive coding,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 4182–4192.
- [13] Y. LeCun, “Deep learning hardware: Past, present, and future. IEEE International Solid-State Circuits Conference-ISSCC,” 2019.
- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [15] T. B. Brown *et al.*, “Language models are few-shot learners,” 2020.
- [16] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” *arXiv preprint arXiv:2003.04297*, 2020.
- [17] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.
- [18] R. Arandjelovic and A. Zisserman, “Objects that sound,” in *Proceedings of the European conference on computer vision*, 2018, pp. 435–451.
- [19] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
- [21] T. Afouras, A. Owens, J. S. Chung, and A. 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*. Springer, 2020, pp. 208–224.
- [22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [23] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, “On variational bounds of mutual information,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 5171–5180.
- [24] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617.
- [25] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman, “Localizing visual sounds the hard way,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16867–16876.
- [26] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [27] A. Labatie, D. Masters, Z. Eaton-Rosen, and C. Luschi, “Proxy-normalizing activations to match batch normalization while removing batch dependence,” *arXiv:2106.03743*, 2021.
- [28] M. Allouah, M. Arnold, and A. Isopoussu, “Deep inertial navigation using continuous domain adaptation and optimal transport,” *arXiv:2106.15178*, 2021.
- [29] Y. Wang, Z. Jiang, Y. Li, J.-N. Hwang, G. Xing, and H. 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*, vol. 15, no. 4, pp. 954–967, 2021.
- [30] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [31] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in *European conference on computer vision*. Springer, 2016, pp. 69–84.
- [32] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *European conference on computer vision*. Springer, 2016, pp. 649–666.
- [33] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” *arXiv preprint arXiv:1803.07728*, 2018.
- [34] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” *arXiv preprint arXiv:2006.09882*, 2020.
- [35] I. Orr, M. Cohen, H. Damari, M. Halachmi, and Z. Zalevsky, “Coherent, super resolved radar beamforming using self-supervised learning,” *arXiv preprint arXiv:2106.13085*, 2021.
- [36] S. Gasperini, P. Koch, V. Dallabetta, N. Navab, B. Busam, and F. Tombari, “R4dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes,” *arXiv preprint arXiv:2108.04814*, 2021.

IX. CONNECTION BETWEEN \mathcal{L}_c AND MI

As per Eq. (2) in Sec. III-B, the score function between q and k models an unnormalised density ratio. The optimal value of this score function is when it captures the underlying density ratio between the unencoded signals x and y (see [19])

$$\exp(q \cdot k / \tau) \Big|_{\text{optimal}} = \frac{p(y|x)}{p(y)} \quad (5)$$

Using Eq. (5) in Eq. (1), and denoting a set of K negative y signals by $Y^- = \{y_i^-\}_{i=0}^{K-1}$, we obtain

$$\begin{aligned} \mathcal{L}_c^{\text{optimal}} &= -\mathbb{E}_{x,y} \log \left[\frac{\frac{p(y|x)}{p(y)}}{\frac{p(y|x)}{p(y)} + \sum_{y_i \in Y^-} \frac{p(y_i|x)}{p(y_i)}} \right] \\ &= -\mathbb{E}_{x,y} \log \left[\frac{1}{1 + \frac{p(y)}{p(y|x)} \sum_{y_i \in Y^-} \frac{p(y_i|x)}{p(y_i)}} \right] \\ &= \mathbb{E}_{x,y} \log \left[1 + \frac{p(y)}{p(y|x)} \sum_{y_i \in Y^-} \frac{p(y_i|x)}{p(y_i)} \right] \\ &\approx \mathbb{E}_{x,y} \log \left[1 + \frac{p(y)}{p(y|x)} (K-1) \frac{\mathbb{E}_{y_i} p(y_i|x)}{p(y_i)} \right] \\ &= \mathbb{E}_{x,y} \log \left[1 + \frac{p(y)}{p(y|x)} (K-1) \right] \\ &\geq \mathbb{E}_{x,y} \log \left[\frac{p(y)}{p(y|x)} K \right] \\ &= -I(x; y) + \log(K) \quad \blacksquare \end{aligned}$$