SENSORS

Coherent, super-resolved radar beamforming using self-supervised learning

Itai Orr[1,2]*, Moshik Cohen[2], Harel Damari[2], Meir Halachmi[2], Mark Raifel[2], Zeev Zalevsky[1]

High-resolution automotive radar sensors are required to meet the high bar of autonomous vehicle needs and regulations. However, current radar systems are limited in their angular resolution, causing a technological gap. An industry and academic trend to improve angular resolution by increasing the number of physical channels also increases system complexity, requires sensitive calibration processes, lowers robustness to hardware malfunctions, and drives higher costs. We offer an alternative approach, named Radar signal Reconstruction using Self Supervision (R2S2), which substantially improves the angular resolution of a given radar array without increasing the number of physical channels. R2S2 is a family of algorithms that use a deep neural network (DNN) with complex range-Doppler radar data as input and trained in a self-supervised method using a loss function that operates in multiple data representation spaces. Improvement of 4× in angular resolution was demonstrated using a real-world dataset collected in urban and highway environments during clear and rainy weather conditions.

INTRODUCTION

Autonomous vehicles have attracted great attention in recent years because of their tremendous influence on the economy and society (*1*) as well as their potential to save lives (*2*). The evolution from current driver assistance systems into fully autonomous vehicles requires several functionally independent sensing modalities for real-time sensing and perception (*3*). The requirement for sensing redundancy (*4*) spurred research toward more advanced camera and light detection and ranging (LiDAR)–based solutions. However, these sensing modalities suffer from inherent sensitivity to harsh weather and limited effective range, due to the electromagnetic spectrum that they use, usually 400 to 800 nm for cameras and 850 to 950 nm or 1.45 to 1.55 μm for LiDARs.

In contrast, automotive radar usually uses a frequency spectrum of 76 to 81 GHz, which offers robustness to weather conditions and longer effective range. However, utilization of radar for autonomous driving is hindered, in part, because of the relatively limited angular resolution currently provided by available commercial platforms.

The angular resolution of a radar translates to the ability to distinguish and separate between targets and is proportional to the antenna aperture. In automotive scenarios, where the environment is usually rich with objects and targets (i.e., cluttered environment), angular resolution is critical. For example, two cars driving in adjacent lanes might be miss-detected as a single object by a limited angular resolution radar. The accuracy of a radar translates to the error in measurement and is considerably smaller than the angular resolution. The relationship between accuracy and resolution is defined by $\delta\theta \cong \Delta\theta/\sqrt{2\mathrm{SNR}}$, where $\delta\theta$ is the accuracy, $\Delta\theta$ is the angular resolution, and SNR is the signal-to-noise ratio.

In a radar array, the individual antenna elements are usually positioned about $\lambda/2$ apart from each other, according to the spatial Nyquist criterion, with λ representing the central wavelength in free space; therefore, increasing the number of antenna elements should enlarge the dimensions of the physical aperture. Following this principle, an industry and academic trend has emerged to enlarge the aperture by increasing the number of physical transmitting and receiving channels (*5*). The drawbacks of this approach are a complex system architecture prone to hardware failure, difficulty to maintain phase coherence, requirement for a complex transmission scheme that achieves noninterfering transmission and robustness to Doppler ambiguity from a relatively larger number of transmit antennas, and requirements for sensitive calibration process and high costs, which hinder the adaptation of these systems in commercial applications.

An additional important factor affecting a radar's angular resolution is the algorithm used for beamforming. Fast Fourier transform (FFT) performed on the angular dimensions of a radar array is considered a conventional beamformer and sets the Fourier resolution of a radar. Super-resolution (SR) methods that aim to achieve sub-Fourier resolution include estimation of signal parameters via rotation invariance techniques (*6*) or the popular multiple signal classification (MUSIC) (*7*). MUSIC's main disadvantages are a requirement of prior information on the number of targets, an assumption on coexistent targets to be uncorrelated, and high computation costs. These limitations make its use in real-world automotive radar applications more challenging. In addition, most current SR methods usually require using several snapshots (i.e., frames) to improve the estimation of the spatial covariance matrix. This requirement is problematic in safety-critical, automotive applications because each added snapshot increases the response time of the system.

Apart from angular SR, a complementary line of work to this research is methods for super-resolving radar data in different dimensions. For example, range SR was demonstrated by sweeping over the coherence length of the transmitted signal (*8*). These methods can potentially be combined with Radar signal Reconstruction using Self Supervision (R2S2) and super-resolve a radar in multiple dimensions.

Recently, deep learning has begun to influence traditional radar signal processing, perception, and system design. Radar data were used with deep neural network (DNN) for high-resolution road segmentation (*9*), road user classification (*10*), multiclass object classification (*11*), road user detection (*12*), vehicle detection (*13*), lane detection (*14*), and semantic segmentation (*15*, *16*). Apart from perception tasks, DNNs have proven useful for cognitive antenna design in phased array radar (*17*) and enhanced radar imaging (*18*).

[1]Faculty of Engineering and the Institute for Nanotechnology and Advanced Materials, Bar Ilan University, Ramat-Gan, Israel. [2]Wisense Technologies Ltd., Tel Aviv, Israel.
*Corresponding author. Email: itaiorr@gmail.com

Another class of algorithms in radar signal processing is compressed sensing, which usually exploits sparseness in a scene to reconstruct one or more dimensions of a radar data tensor, i.e., range-Doppler-azimuth-elevation. However, this is a property that does not often occur in cluttered, urban driving scenarios. Further details on the lack of sparsity in the dataset used in this work are provided in the Supplementary Material and shown in fig. S1. Complex block sparse Bayesian learning was demonstrated for radar signal reconstruction (*19*). A spatial compressed sensing framework (*20*) was developed and evaluated by numerical simulations for five targets with constant SNR. The authors assumed that the number of targets is known and noise level is available and forewent from estimation of measurements in the range and Doppler dimensions. Iterative method with adaptive thresholding (*21*) was used for a sparse multiple in multiple out (MIMO) radar array, where the authors (*22*) conducted examinations of a corner reflector in an anechoic chamber and a single parked vehicle at a range of 4 m. Examination of compressed sensing for MIMO radar (*23*) concluded that these techniques remain valid when there are under 106 scatter points in a scene. However, in typical urban scenes that may contain many more scatter points, these methods require using a high SNR threshold to minimize the number of scatterers.

Research toward using DNNs to improve radar angular resolution is in its early stages. Radar data in range-Doppler representation were used with a generative adversarial network architecture to demonstrate SR in two specific cases (*24*): pedestrian micro-Doppler signature by collecting data of people walking on a treadmill and a staircase, which achieved angular SR with a factor of 2*x*. The authors (*24*) point out the difficulty of assembling a large manually labeled dataset in real-world scenarios for the general case of numerous types of objects, classes, materials, and shapes.

Instead of real-world data, synthetic data were used for training with a single radar snapshot as input (*25*). However, using synthetic data for training before deployment in a real-world environment usually results in relatively lower performance caused by modeling and numerical errors in the simulation used to create the synthetic data. This is also referred to as the sim-to-real adaptation challenge.

Multiple snapshots of a spatial covariance matrix were used with a convolutional neural network and a one-dimensional (1D) antenna array with simulated data for direction of arrival (DOA) estimation and SR (*26*). A single snapshot of a spatial covariance matrix was used with a fully connected model for DOA estimation and SR of a 2D antenna array with simulated data (*27*) and a 1D antenna array with both simulation and real-world data where the targets were corner reflectors (*28*). Two snapshots were used with an anechoic chamber setup to generate a dataset that was used with a fully connected model for DOA estimation (*29*).

Although shown only for simulated data or controlled scenarios with very few targets, these works show the potential that DNN has for super-resolving radar arrays in real-world environments that usually contain many targets and reflections. We hypothesize that previous methods for DNN-based radar SR failed or did not try to generalize to uncontrolled, real-world environments mainly because of a lack of a suitable training methodology.

Self-supervised learning is a young research area and is considered a part of unsupervised training. The main principle in self-supervised learning is using one part of data to predict a different part of the same data (*30*, *31*). In other words, self-supervised learning is similar to regular supervised learning but with a particular method to obtain the labels by exploiting the inner structure of the data, for example, predicting the future from the past, or by using external constraints, for example, consistency. The strength and disruptive potential of this training methodology lies in the fact that in many applications, data are in abundance; however, labeling the data, which is essential for supervised training, is a time-consuming and expensive process. Furthermore, in some applications such as image denoising (*32*), manual labeling is not a viable solution. Self-supervised techniques showed promising early results for semantic image segmentation (*33*–*35*), temporal cycle consistency to learn temporal alignment between videos (*36*), dense shape correspondence for 3D objects (*37*), and feature representation for visual tasks (*38*–*40*).

The field of image SR has also used self-supervision to create state-of-the-art results (*41*–*44*). At its fundamentals, self-supervision for image SR uses a high-resolution image that is first down-sampled to create a low-resolution image. A DNN is then trained using the low-resolution image as input and the high-resolution image as label.

Apart from computer vision, research into signal processing has also begun using self-supervised learning with other forms of data. In audio data, it was used for speech enhancement (*45*), pitch estimation (*46*, *47*), source separation (*48*), and feature representation (*49*–*51*). In electroencephalography data, self-supervision was used for representation learning (*52*), and in electrocardiogram data, self-supervised learning was used for emotion recognition (*53*).

This work proposes to leverage self-supervised learning to super-resolve a radar array. More specifically, R2S2 uses an auto-encoder trained in a self-supervised method with a diluted radar array and used to reconstruct the amplitude and phase of missing receiving channels, where the dilution of the radar array was designed to limit the resolution of the input array. In other words, we propose to use different spatial/temporal perspectives of a scene as supervision signals during training. To enforce coherence during the reconstruction process, a loss function that operates on multiple data representation spaces was used.

In contrast to several radar-based compressed sensing and SR methods, R2S2 does not require sparsity in the range-Doppler-azimuth dimensions. It can be used in highly cluttered environments, such as crowded urban streets with numerous objects and targets present in the radar field of view, and does not require prior knowledge on the number of targets in a scene. Validation was performed on a real-world dataset collected using a vehicle mounting a radar unit and driven in urban and highway environments in both clear and rainy weather conditions.

RESULTS

Data collection in unstructured environments

In this work, we target the general application of radar SR. To demonstrate our approach, a dataset was collected in uncontrolled urban and highway environments in both clear and rainy weather conditions using a vehicle mounting a temporally synced camera and radar with their field of view overlapped. The dataset was split into 118 thousand frames for training and 12 thousand frames for validation. The validation dataset was separated from the training dataset by collecting data during different dates and locations to avoid the appearance of similar frames in both datasets, which could have occurred in the case of simple random split. Samples from the training dataset are shown in Fig. 1. Note that the radar used achieves relatively high SNR and measurement accuracies, which are essential

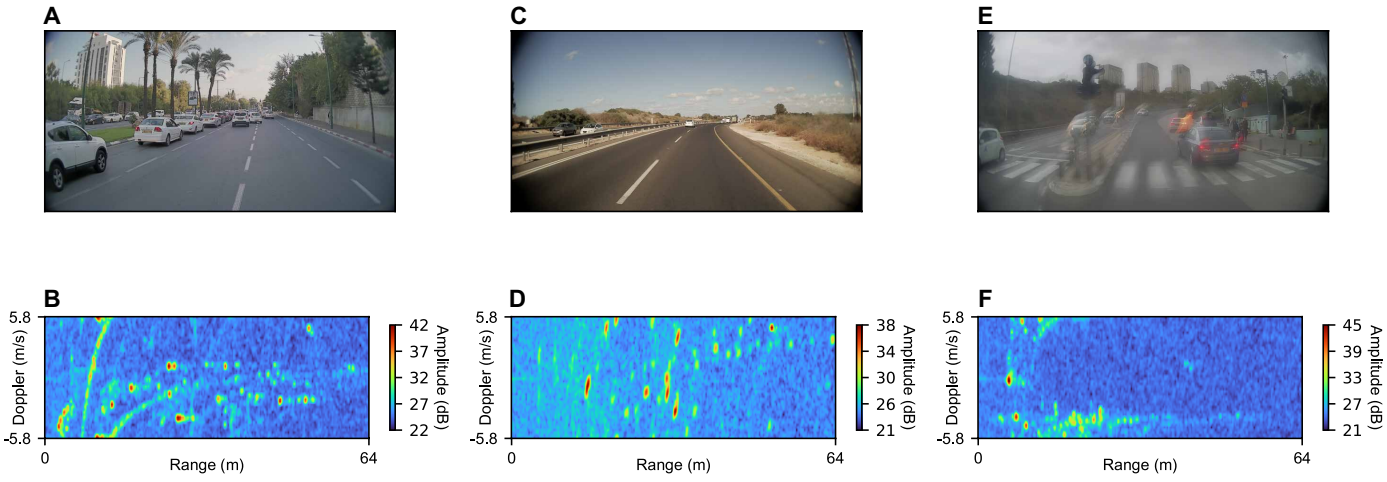

Fig. 1. Sample frames from the training dataset. (**A**, **C**, and **E**) Camera images. (**B**, **D**, and **F**) The respective range-Doppler maps. The blurry image in (E) was caused by rain droplets on the camera lens during data collection.

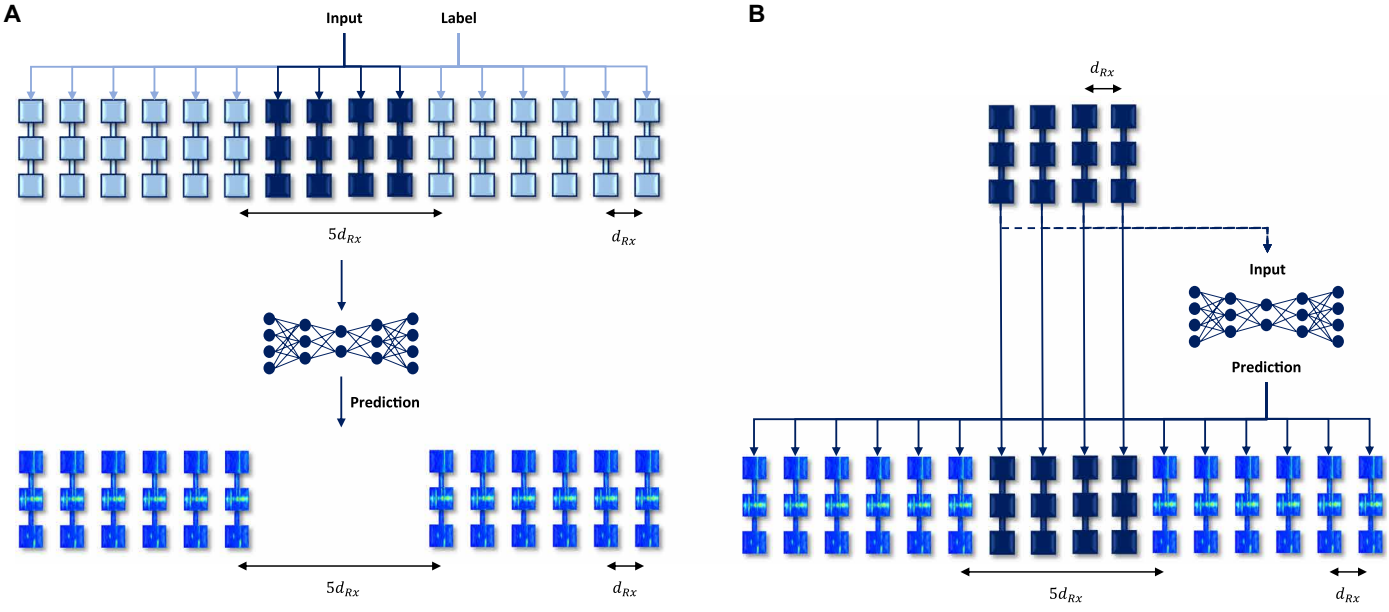

Fig. 2. Radar SR using self-supervised learning. (**A**) Training mode: In this example, 4 receiving channels are used as input (dark blue) to predict 12 receiving channels outside the original aperture (light blue). (**B**) Inference mode: The original array is used as input to a DNN, which predicts adjacent receiving channels outside the original aperture. Afterward, both input and predicted receiving channels are used for coherent beamforming. The resulting artificial array has 16 receiving channels and thus 4× improved resolution from the original array.

for next-generation safety systems. These are attributed to high-efficiency antenna and system design; both are out of scope for this work. In addition to high SNR radars, R2S2 can also be used with lower SNR systems because there is no requirement for minimum SNR.

We used a frequency modulated continuous wave (FMCW) MIMO radar with a 79-GHz carrier frequency. An FMCW radar transmits a linear chirp signal whose frequency increases linearly with time. When combined with means of signal processing (mainly FFT), it is possible to extract useful information from the raw signal such as, range, velocity, and DOA (*54*). MIMO radar is composed of multiple transmitter (Tx) and receiver (Rx) antennas. Each transmitter can transmit a waveform independently of the other transmitting antennas, whereas each of the receiving antennas can also receive these signals independently. By processing measurements from different transmitting and receiving antennas, one can create a virtual aperture whose size is larger than the physical aperture; i.e., an antenna array composed of NTx transmitters and an array of NRx receivers results in a virtual array of $NTx \times NRx$ channels. This increase in aperture size translates to improved performance such as spatial resolution, resistance to interference, and probability of detection of the targets (*55*). In this work, a collocated MIMO radar was used; however, the proposed method can be applied to a noncollocated MIMO radar as

well. In addition, although we focus on MIMO radar in this work, a similar approach can be applied with other multichannel radars.

Data preprocessing

A radar signal in its raw form contains a variety of information originating from different physical phenomena in the environment, such as targets, reflections, and electromagnetic wave propagation through the atmosphere. In addition, hardware-related effects originating from components such as the signal generator, receive/transmit chains, and antenna elements also greatly affect data fidelity. The combination of numerous, simultaneous, sometimes nonlinear, and often coupled mechanisms that affect a radar signal also makes its mathematical modeling very difficult (*56*, *57*). To differentiate between different interactions, FFT has long become a staple for radar signal processing. More specifically, FFT is used to transform a signal from its raw measurement form to different representation spaces, such as range-Doppler for example. In this work, the radar used has a uniform linear array (ULA) antenna array configuration with 16 virtual channels and angular resolution of about 6°, providing the ability to process both amplitude and phase information in three dimensions: range, Doppler, and azimuth. The waveform used was configured to 48 sweeps and 256 samples with maximum detection range of 64 m and unambiguous maximal relative velocity of 5.8 m/s. The field of view was configured to 100° horizontal. The relatively small number of sweeps and samples was chosen to allow for faster training time.

The input to the model was created by applying a window function and FFT on both sweep and sample dimensions to generate a complex data tensor with the dimensions of virtual channel, range, and Doppler. More specifically, the original signal x_{raw} has the dimensions of virtual channel, samples, and sweeps and first goes through the range processing described in Eq. 1, which includes windowing and real-to-complex FFT on the sample dimension

$$x_{\text{range}} = \mathcal{F}_{\text{sample}}(W_{\text{sample}}(x_{\text{raw}})) \quad (1)$$

where $\mathcal{F}$ is FFT and W is a window function. The transformed signal x_{range} has the dimensions of virtual channel, range, and sweeps and goes through Doppler processing, which includes windowing and complex-to-complex FFT on the sweep dimension, as described in Eq. 2

$$x_{\text{range-Doppler}} = \mathcal{F}_{\text{sweep}}(W_{\text{sweep}}(x_{\text{range}})) \quad (2)$$

The resulting signal $x_{\text{range-Doppler}}$ has the dimensions of virtual channel, range, and Doppler and is then used as input to a DNN. In our experimentation, we filtered the first few range bins to remove effects of Tx-Rx coupling. The suggested method holds several important characteristics in regard to data preprocessing, which

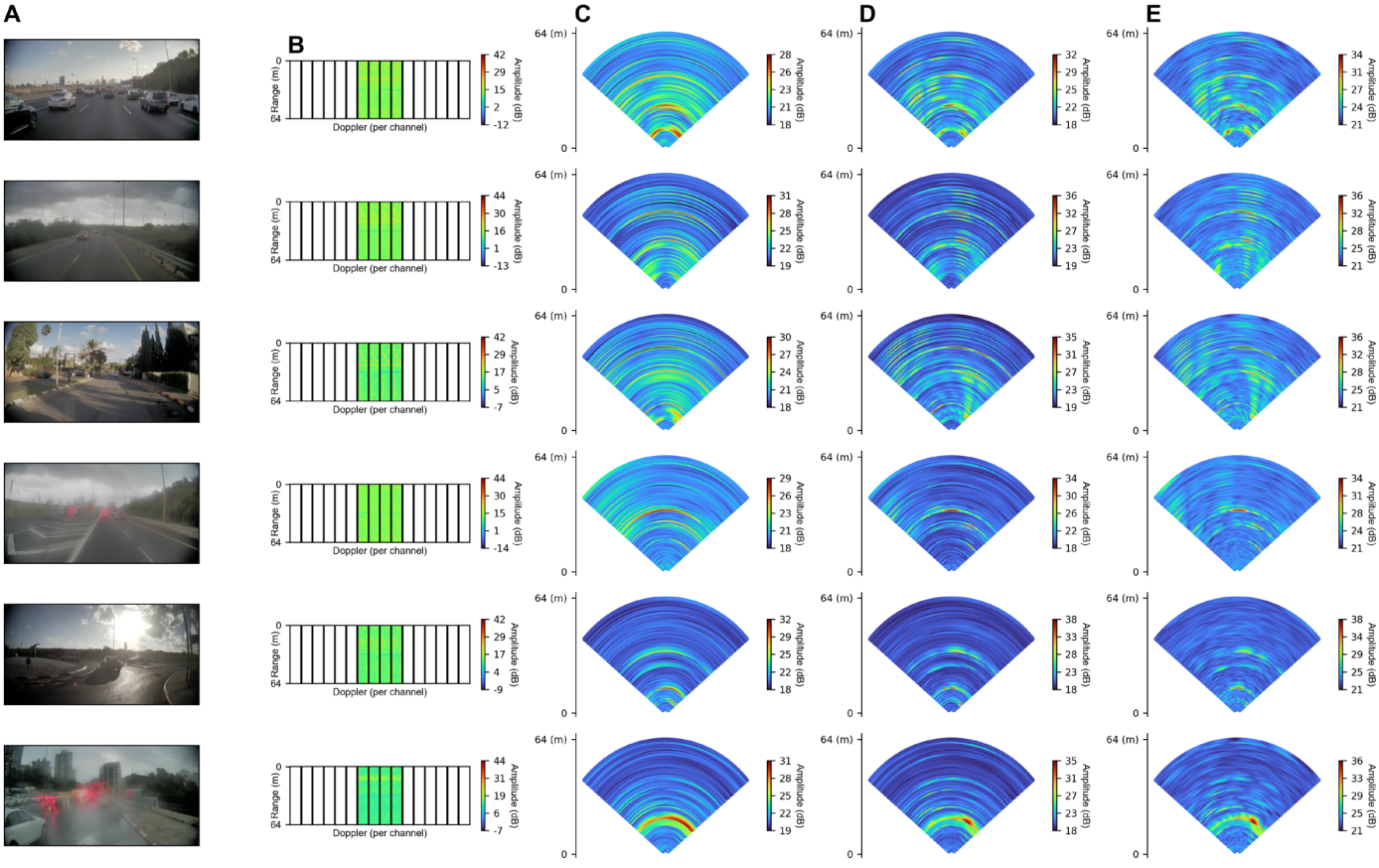

Fig. 3. Sample results for the SR configuration. (**A**) Camera images. (**B**) Input radar arrays. The empty spaces were left to orient the reader to the channel position within the radar array. (**C**) Input beamformers. (**D**) Predicted beamformers. (**E**) Label beamformers. All beamformers are displayed in Cartesian coordinates. Camera images are used for the reader's reference and were not used during training or inference. Blurry images were caused by rain droplets on the camera lens during data collection.

contribute to its generality and robustness while addressing the shortcoming of previous approaches to radar SR. Mainly, there is no requirement for specific filtering, there are no assumptions on the sparsity of the data, there is no minimum SNR threshold, no calibration is required, there is no maximum number of scatter points, and there are no requirements of prior information on the scene. In addition, to remove the requirement for complex radar signal modeling, R2S2 was designed as an end-to-end approach, forcing a DNN to implicitly learn a signal model as part of the reconstruction process. Meaning, there is no need to provide an accurate and detailed mathematical description of the signal.

SR and coherent beamforming experiments

The model was given as input a diluted 1D subarray of complex (both amplitude and phase) range-Doppler maps, whereas the remainder array was used as label. Meaning, the training is performed in a self-supervised manner. There are numerous possible permutations for the choice between input and label receiving channels; experiments were performed using an example configuration described in Materials and Methods and shown in Fig. 2, where a virtual array of 16 channels was split to 4 receiving channels, which were used as an input array with an angular resolution of about 24°, whereas the remainder 12 receiving channels were used as label,

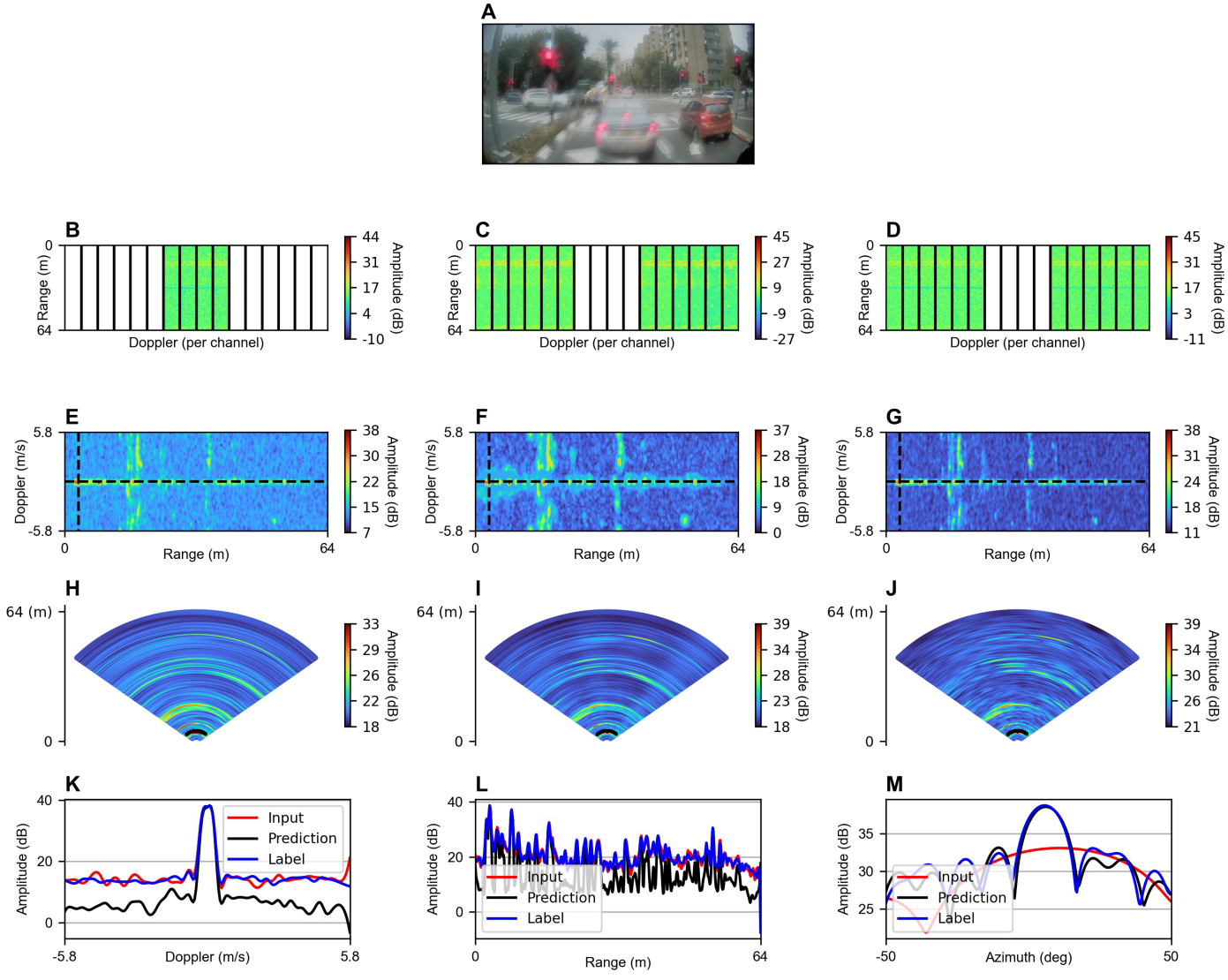

Fig. 4. Detailed results for the SR configuration. (**A**) Reference camera image. Blurry image was caused by rain droplets on the camera lens during data collection. (**B** to **D**) From left to right, input, predicted, and label radar arrays, respectively. The empty spaces were left to orient the reader to the channel position within the radar array. (**E** to **G**) From left to right, input, predicted, and label range-Doppler maps. The dotted black lines represent the maximum detection and are used for the Doppler and range cross sections in (**K**) and (**L**), respectively. (**H** to **J**) From left to right, input, predicted, and label beamformers in Cartesian top view. The dotted black line represents the maximum detection with the azimuth cross section shown in (**M**). These results show that by using R2S2, the input array is super-resolved to match the performance of the label array. In this figure, we provide a sample of a stationary scenario, meaning the radar is not moving, with similar results to samples where the radar was moving. These results suggest that the DNN is not relying solely on Doppler and micro-Doppler effects during the reconstruction process.

giving the full array an angular resolution of about 6°. Meaning, the full array has 4× improved resolution than the input array.

Sample results from the validation dataset are provided in Fig. 3, showing representative scenarios from urban and highway environments in both clear and rainy weather conditions. Also provided in Fig. 3 is a Cartesian view comparison between the input, predicted, and label beamformers, which was obtained by performing FFT on the channel dimension of the input, predicted, and label arrays as shown in Fig. 2. These results demonstrate the use of R2S2 to super-resolve a limited angular resolution radar array, thereby achieving 4× improved resolution in scenarios representing various combinations of dynamic and static objects, including vehicles, vegetation, sidewalks, poles, and structures.

The critical importance of superior angular resolution for automotive radars can be further understood by examining common everyday driving scenarios as demonstrated in Fig. 4 and figs. S2 and S3. These examples demonstrate how limited-resolution radars (i.e., the input radar array used) can falsely detect objects in front of the vehicle even though the road ahead is clear. In addition, adjacent objects can also be falsely detected as a single object. These highly undesired phenomena can be resolved by using our method to increase the angular resolution of the radar array.

Additional experiments were carried out using a stationary radar and stationary corner reflectors (i.e., point targets). The results are reported in fig. S4, where a single corner reflector was positioned at the radar's boresight at a distance of about 25 m in addition to two similar corner reflectors that were positioned at about 50 m and ±3° from the radar's boresight. The results show that the input array does not resolve between the two corner targets, whereas by using R2S2 it is possible to resolve between the two corner targets, similarly to the label array, thus demonstrating a 4× improvement in angular resolution. In addition, R2S2 is able to register a single target as such and does not "split" the target, which would have resulted in a false detection.

Further validation was performed using two evaluation metrics: L1 and peak SNR (PSNR). Both metrics were averaged over the validation dataset. Lower L1 error corresponds to improved reconstruction and was calculated using Eq. 3

$$\mathrm{L1} = \frac{1}{N_i N_j} \sum_{i,j} \frac{|\, y_{i,j}^{\mathrm{pred}} - y_{i,j}^{\mathrm{label}} \,|}{|\, y_{i,j}^{\mathrm{label}} \,|} \tag{3}$$

where L1 is the reconstruction metric. In the range-Doppler representation space, both metrics were calculated for each receiving channel separately, whereas in the beamformer representation space (i.e., range-Doppler-azimuth) the metrics were calculated globally to focus on coherence.

Because R2S2 deals with coherent reconstruction of an array's response, the important metrics are associated with the beamformer representation space and, more specifically, the combination of low L1 and high PSNR, which correlate to coherent beamforming.

Table S1 displays an ablation study performed on the loss function described in Materials and Methods and was averaged over the validation dataset. The results show that the best performances (in bold) are achieved by both parts of the loss function, suggesting that improved coherence is attained by adding the beamformer constraints to the optimization process.

To further support the general applications of R2S2, experiments were performed with a different permutation of input and label receiving channels. This configuration, which was termed "sparse array," is displayed in Fig. 5, where R2S2 is used to interpolate receiving channels between sparsely spaced input receiving channels. Sample results from the validation dataset are provided in Fig. 6 and figs. S5 and S6, where 4 uniformly spaced receiving channels are used as input and 12 receiving channels are used as label. In this configuration, the resolution of the input and label arrays is similar

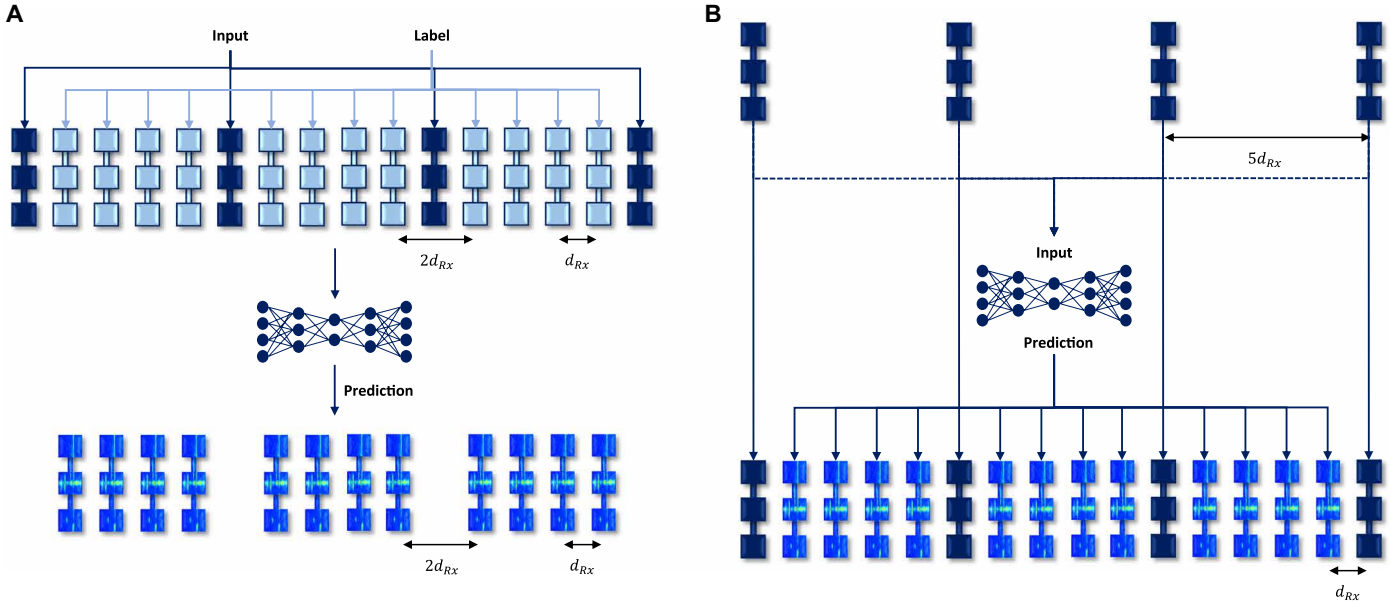

Fig. 5. Sparse array configuration. (**A**) Training mode: In this example, 4 receiving channels (dark blue) are used as input and 12 receiving channels are used as label (light blue). Together, they reconstruct a full 16 receiving channel radar array. d_{Rx} is the distance between adjacent receiving channels. (**B**) Inference mode: Input receiving channels are first used by a DNN to predict artificial receiving channels, each at specific missing locations in the full array. Afterward, both input and predicted receiving channels are used for coherent beamforming.

(they share the same aperture size); however, because of the large spacing between receiving antenna elements in the input array, the input beamformer suffers from grating lobes that severely degrade performance. By applying R2S2, we were able to coherently reconstruct the missing receiving channels and match the performance of the label array.

Experiments using corner reflectors for the sparse array configuration are provided in fig. S7. The grating lobes of the input array are clearly visible, which would cause severe ambiguity in an uncontrolled environment, rendering the usage of such array very problematic. In contrast, R2S2 is able to suppress the side lobes of the main beam, similarly to the label beamformer, which further suggests coherent reconstruction of receiving channels by our method.

Additional validation of the sparse array configuration was performed on the validation dataset and compared with linear and cubic interpolations. The interpolations were performed in the channel dimension for each range-Doppler cell. The results provided in table S2 show that linear and cubic interpolations do not enforce coherence during the reconstruction process, as evident by the high L1 score in the beamformer representation space. In contrast, R2S2 is able to reconstruct the array correctly and coherently.

Further evidence on the lack of coherence when using linear or cubic interpolations can be seen in fig. S8. There, we examine the beamformers of the input, predicted, label, linear, and cubic interpolation arrays. The results show grating lobes and amplitude bias in the linear and cubic interpolation beamformers, which are not

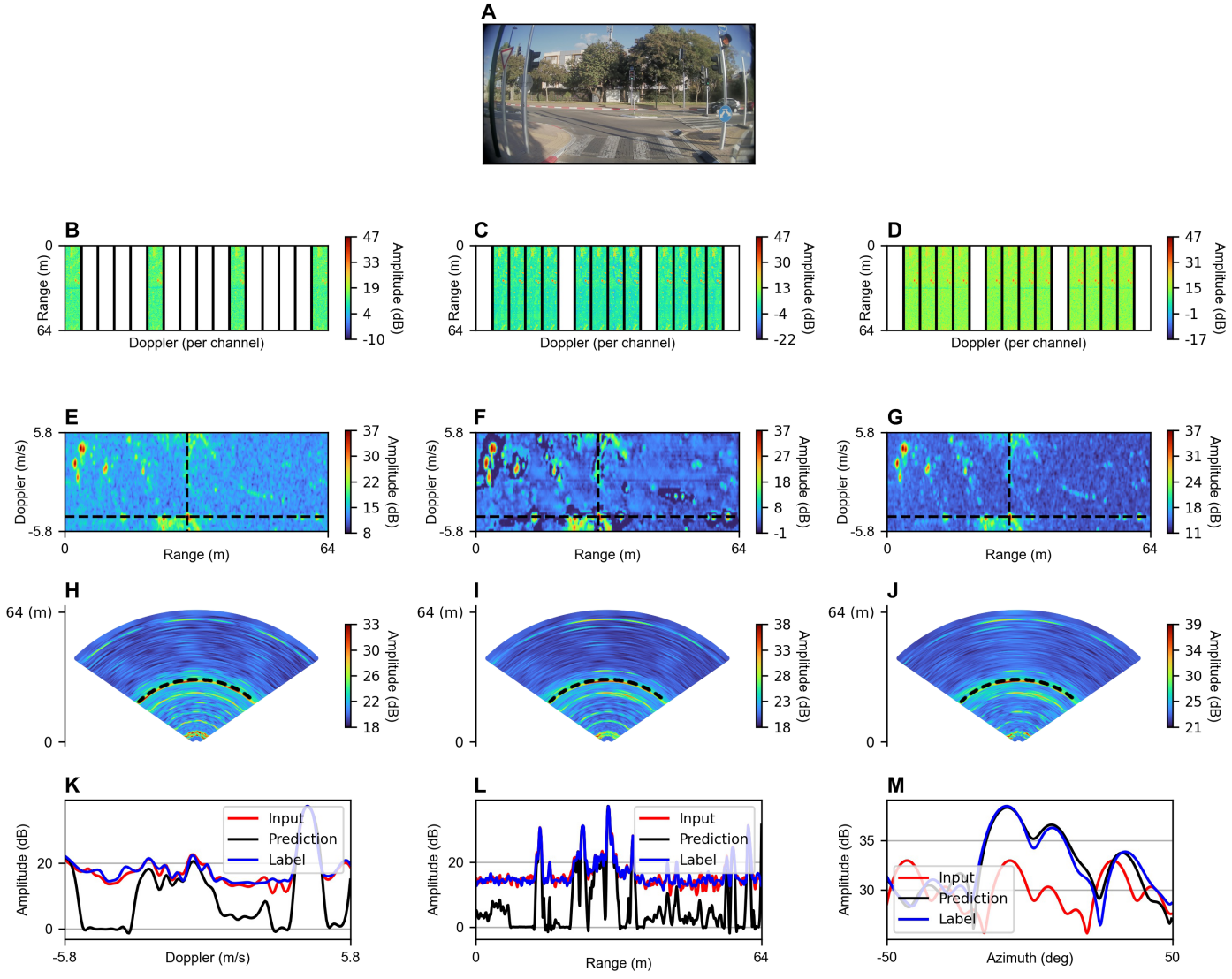

Fig. 6. Sample results for the sparse array configuration. (**A**) Reference camera image. (**B** to **D**) From left to right, input, predicted, and label radar arrays, respectively. The empty spaces were left to orient the reader to the channel position within the radar array. (**E** to **G**) From left to right, input, predicted, and label range-Doppler maps. The dotted black lines represent the maximum detection and are used for the Doppler and range cross sections in (**K**) and (**L**), respectively. (**H** to **J**) From left to right, input, predicted, and label beamformers in Cartesian top view. The dotted black line represents the maximum detection with the azimuth cross section shown in (**M**). Notice the grating lobes in (H) that are not present in (I) or in (J). These results show that beamforming on the input radar array suffers from degraded performance because of grating lobes caused by the large distance between each antenna element. By using R2S2, the gaps are filled and the performance of the predicted beamformer matches the label beamformer.

present in the label data. Combined with the quantitative results in table S2, these results suggest that these interpolation methods do not enforce coherent reconstruction. The input beamformer shows false targets due to grating lobes caused by the large distance between each antenna element (i.e., spatial undersampling), whereas the predicted beamformer shows similar results to the label beamformer, suppressing the side lobes and eliminating the false targets.

Because R2S2 uses signal reconstruction to improve resolution, it can also be used for mitigation of hardware failure. More specifically, in cases where one or more receiving channels are randomly corrupted, the suggested method can be used to replace them with artificial receiving channels. This configuration, which was termed "random missing channels configuration," is displayed in Fig. 7.

To demonstrate this approach, experiments were performed where a DNN is used to estimate random missing receiving channels. Because the number and position of the missing receiving channels can vary and are not known in advance, the DNN first needs to determine whether each receiving channel is corrupt and then coherently reconstruct it on the basis of the remaining receiving channels. Sample results from the validation dataset are provided in Fig. 8 and figs. S9 and S10, showing a detailed analysis of R2S2 for the random missing channels configuration where we observe that R2S2 can reconstruct the range-Doppler maps and create a coherent array.

To further assess the performance of this configuration, we conducted a quantitative comparison with linear and cubic interpolations with a single randomly missing receiving channel. The results are provided in table S3 and were performed using the validation dataset, with R2S2 outperforming linear and cubic interpolations, as evident in lower L1 and high PSNR in the beamformer representation space. Linear and cubic interpolations cannot estimate receiving channels at the edge of an array, whereas our method is able to extrapolate and interpolate.

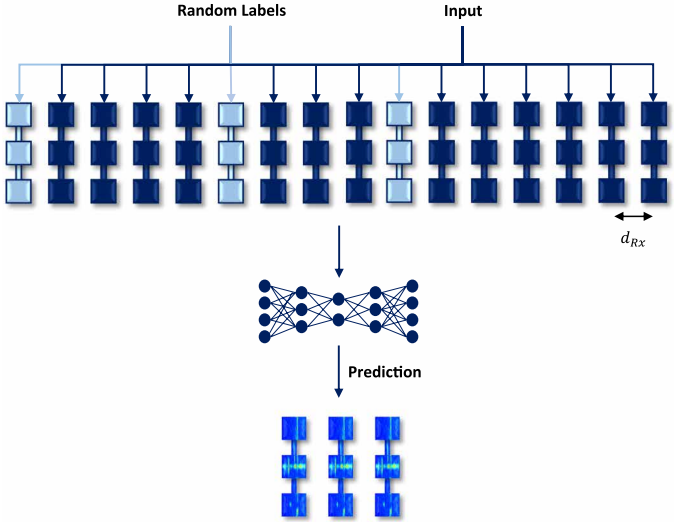

Fig. 7. Random missing channels configuration. At training, a random receiving channel is chosen and masked to be used as label. The model then identifies the missing receiving channel and predicts its measurements. Because the missing receiving channel can be located anywhere in the virtual array, this configuration combines both interpolation and extrapolation performed by the model.

A generalized version of this configuration using a single DNN trained to predict a random number of randomly positioned receiving channels was also assessed. A quantitative comparison of a DNN trained with up to eight random missing channels and validated over the validation dataset is shown in fig. S11. As expected, we observe a decrease in performance as reflected in L1 and PSNR metrics as the number of random missing channels increases. By combining this analysis with specific performance criteria, it is possible to set a maximum number of missing channels for which this configuration may be used for real-world application.

DISCUSSION

Limited angular resolution is one of the main limiting factors in automotive radar applications. An industry trend to improve angular resolution by increasing the number of physical receiving channels also increases system complexity, creates cumbersome calibration processes, adds sensitivity to hardware failure, decreases power efficiency, and drives higher cost. An alternative approach is to use SR algorithms. However, unless carefully designed and implemented, this can also introduce additional sensitivity to calibration, increase latency, and add limitations on the number of targets and, in some cases, a requirement for prior knowledge of the environment.

To address these limitations, R2S2 was designed with a single snapshot as input, which is an important property in automotive applications where reaction time is critical. Furthermore, the dataset was collected in uncontrolled urban and highway environments during both clear and rainy weather conditions and was not focused on a specific class of objects. The preprocessing stage did not contain special filtering nor required special calibration process; there was no requirement for prior knowledge on the number of targets in a scene and no minimum SNR threshold. In addition, the run time is invariant to the number of detections in a frame. Meaning, a highly cluttered scene will not cause a bottleneck in processing time, which is an important characteristic in real-time applications.

The proposed approach can replace or be used in addition to existing SR methods and uses self-supervised learning to train a DNN to predict artificial receiving channels in range-Doppler representation outside an array's aperture. The combined, original, and artificial receiving channels create a larger aperture; if coherence is maintained, the improvement of the larger array is improved angular resolution.

To enforce coherence, additional constraints were introduced during the training process. These constraints were in the form of additional loss terms operating in the beamformer representation space. Training was performed using both representation spaces (i.e., range-Doppler and beamformer representations) simultaneously. The multi-objective loss function also has drawbacks, making the fine-tuning of a model more complex with additional sensitivity to hyperparameters.

In this work, FFT was chosen as a beamformer. However, alternative beamformers can also be used. For example, the constraints introduced in the loss function as $\mathcal{L}_{\mathrm{bf}}$ can be created by applying an SR algorithm such as MUSIC. By combining the suggested approach with other SR methods, it may be possible to achieve higher improvement factors than previously achieved. Similarly to other SR methods, attention needs to be placed on errors in the prediction due to the fact that R2S2 is used to estimate and not measure the location and velocity of a target.

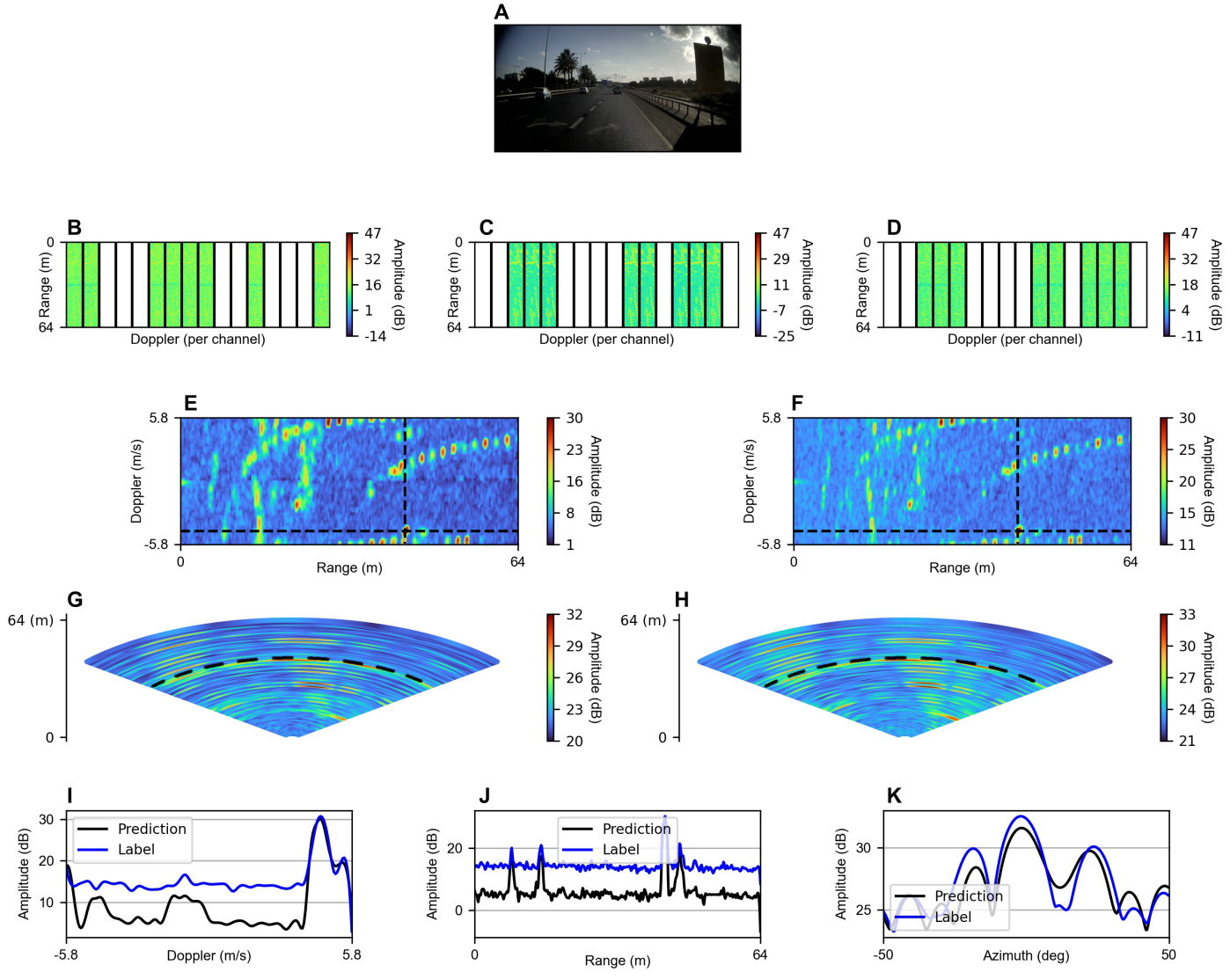

Fig. 8. Sample results for the random missing channels configuration. (**A**) Reference camera image. (**B** to **D**) From left to right, input, predicted, and label radar arrays, respectively. The empty spaces were left to orient the reader to the channel position within the radar array. In this particular example, seven input channels are missing. (**E** and **F**) From left to right, predicted and label range-Doppler maps. The dotted black lines represent the maximum detection and are used for the Doppler and range cross sections in (**I**) and (**J**), respectively. (**G** and **H**) From left to right, predicted and label beamformers in Cartesian top view. The dotted black line represents the maximum detection with the azimuth cross section shown in (**K**).

Experiments were performed with a configuration of 4 input receiving channels and 12 label receiving channels, which achieved a 4× improved angular resolution factor. However, additional permutations are also possible; for example, eight input receiving channels and eight labels receiving channels would have created a 2× improved angular resolution factor. Furthermore, given a larger original radar array, the suggested method can potentially achieve larger improvement factors. For example, an array with 64 receiving channels can be split into 8 input receiving channels and 56 label receiving channels, which results in 8× improved resolution factor.

An important observation is shown in Fig. 4 and figs. S4 and S7, which demonstrate cases where both the radar and targets are stationary, as evident by the Doppler plot centered around $v = 0$ m/s. In this case, similar qualitative results arise in comparison with cases where the radar was moving, which suggests that in contrast to previous methods (*24*), a DNN trained with R2S2 does not rely exclusively on the Doppler and micro-Doppler effects during the reconstruction process.

R2S2 can also be used with a different type of configuration, which was termed sparse array and simulates a sparse radar array. Meaning, the distance between each virtual antenna element is larger than $\lambda/2$, which is optimal in terms of grating lobes and spatial ambiguity. This allows the array to have a larger aperture size, thus improving its angular resolution. However, this enlarged element distance causes ambiguities and degraded performance in the array pattern as observed in Fig. 6 and figs. S5 to S7, where we demonstrate that when only using the sparse input receiving channels for beamforming, there is a substantial reduction in SNR and high side lobes compared with using the entire array. However, by using

R2S2, coherent artificial receiving channels are predicted to fill in the gaps, which make it possible to have a larger aperture and still maintain high performance, matching those of the full array.

Additional application of R2S2 refers to its use for mitigation in cases of corrupt receiving channels. This configuration trains a DNN to predict one or more random receiving channels from the remaining functional receiving channels. During inference, the missing receiving channels can be any receiving channel in the array, without the need to change configuration or "notify" the model which receiving channel is missing. Meaning, the model implicitly identifies which receiving channel is missing and predicts the appropriated artificial receiving channel.

This work offers an alternative approach to conventional radar beamforming and SR that challenges an industry and academic trend toward increasing the number of physical channels in radar arrays and larger physical aperture to achieve improved angular resolution. The suggested method, termed R2S2, uses a DNN trained in a self-supervised method with a diluted antenna array to super-resolve a radar by coherently predicting the amplitude and phase of receiving channels outside the physical or virtual aperture using a loss function in multiple data representation spaces. The results demonstrated robust, real-time performance and an improvement factor of 4× in cluttered scenarios by using a real-world dataset collected in urban and highway environments during clear and rainy weather conditions. In addition, R2S2 can also be used for mitigation of hardware failure, which can further increase the reliability of automotive radars.

This work suggests that learning-based methods can be combined with or replace traditional methods for radar SR in real-world applications. We hope that our method will assist to bridge the technological gap in radar angular resolution and enable radar-centric autonomous driving. In a broader sense, this work demonstrates how self-supervised learning can be used for radar signal processing, which we hope will inspire more research in this direction.

MATERIALS AND METHODS

Training methodology

A fundamental concept of self-supervised learning involves manipulating, augmenting, or masking parts of an input data and then predicting the original data, part of the original data, or which manipulation was performed. In this work, we propose to use self-supervision to predict radar data and treat the SR problem as a signal reconstruction problem while combining it with traditional beamformers. Meaning, our method can work in combination with other SR methods.

To improve a radar array's angular resolution, R2S2 uses a DNN to predict signals outside the physical or virtual array aperture. The combination of the original receiving channels and the predicted receiving channels creates an artificial radar array with a larger aperture and thus improved angular resolution. In this work, the term "artificial receiving channels" is used for the predicted receiving channels to differentiate them from virtual receiving channels created by a MIMO process.

We propose to expand a virtual MIMO array and create an artificial array composed of virtual receiving channels from the MIMO array and artificial receiving channels from the DNN's prediction. Together, all channels can then be used with a beamformer. In this work, FFT was used as beamformer; however, R2S2 can also be applied with other beamformers, for example, MUSIC. The coherence of the predicted artificial channels requires special attention. Otherwise, the resulting beamforming will not achieve SR. This task is especially difficult because it requires a DNN to extrapolate coherent data for receiving channels positioned far from the original input receiving channels.

Our method allows for flexibility in the partitioning between input and label receiving channels. For example, in a ULA with 16 channels, a partitioning of 8 input receiving channels and 8 label receiving channels will result in a 2× angular resolution improvement factor. Another example for the receiving channel partitioning is displayed in Fig. 2, where a ULA with 16 receiving channels is considered. The central 4 receiving channels are used as input to a DNN, whereas the remainder 12 receiving channels are used as label, as seen in Fig. 2A. During inference mode, the original input receiving channels are used twice, first as an input to a DNN from which 12 receiving channels are predicted. Second, they are used together with the predicted receiving channels to create an artificial array that has a total of 16 receiving channels, thus 4× improved resolution from the original 4 receiving channels input array, as seen in Fig. 2B.

Model

The design choices for model architecture are coupled with those for data representation. Meaning, different models will differ from one another based on the way the data are represented. In the case of radar data, there are multiple ways to represent a complex signal. For our experimentation, we split the data into their real and imaginary parts and concatenated the features in the channel dimension. Meaning, the original complex signal is doubled in features and transformed to a real signal with real-valued operators used for processing. The intuition behind this decision is that real and imaginary representation of complex-valued signal mixes between the amplitude and the phase with similar magnitudes and physical behavior.

Another option that we did not experiment with because of its increased complexity is to split the complex-valued signal into amplitude and phase (polar transformation). This partitioning creates very different physical information channels in the model. A logical method for this model architecture would be to create a separate encoder for the amplitude and a separate encoder for the phase without feature sharing because their physical behavior is different. After the two encoders, a main body can combine the information to create a unified prediction, thus allowing the model to learn any coupled behavior that the phase and amplitude share in the physical signal.

A third option is to use complex data structure together with complex operators. This design choice has the benefit of a smaller model on the expense of a larger number of operations (i.e., memory-compute trade-off).

The main criterion for model architecture in this work was simplicity. Meaning, we focus on the training methodology and loss function while deliberately keeping the model simple. The model used in all experiments was adapted from (*9*) and based on the encoder-decoder U-Net (*58*) model combined with position embedding and self-attention (*59*) layers working on the channel dimension to encourage learned cross-channel correlations. Additional layers used were average pooling, leaky-Relu activation, and instance normalization. All convolution and transpose convolution used a three-by-three kernel. The proposed model has about 5.7 million parameters and achieves 15-ms inference time on 2080Ti GPU, which makes the suggested approach attractive for embedded, real-time applications. Future research can explore additional data representations,

as well as transformers or different types of attention to further improve the results presented in this work.

Loss function

To coherently reconstruct a radar array's response, a loss function was constructed, which operates in two data representation spaces simultaneously. As a general partitioning, the first loss representation space was range-Doppler ($\mathcal{L}_{rd}$) and was used to reconstruct the amplitude. The second loss representation space, termed "beamformer" ($\mathcal{L}_{bf}$), was achieved by applying FFT on the channel dimension and was used mainly to reconstruct the phase while enforcing coherence throughout the array.

The loss term is a sum of range-Doppler–based and beamformer-based losses: $\mathcal{L} = \mathcal{L}_{rd} + \mathcal{L}_{bf}$. The resulting multi-objective loss function combines two different physical representations; therefore, addition of these loss terms should be done carefully. During experimentation, normalization of each loss term was examined; however, no substantial performance improvements were observed.

Both loss terms are composed of reconstruction loss and two regularization terms in the form of energy conservation and total variation. In the range-Doppler representation space, the loss function is displayed in Eq. 4:

$$\mathcal{L}_{rd} = \lambda_{rd_{rec}}\mathcal{L}_{rd_{rec}} + \lambda_{rd_{energy}}\mathcal{L}_{rd_{energy}} + \lambda_{rd_{tv}}\mathcal{L}_{rd_{tv}} \quad (4)$$

where (λ_{rdrec}, $\lambda_{rdenergy}$, λ_{rdtv}) are hyperparameters for the reconstruction, energy, and total variation losses, respectively. $\mathcal{L}_{rdrec}$ is the L2 reconstruction loss, shown in Eq. 5:

$$\mathcal{L}_{rd_{rec}} = \frac{1}{N_i N_j}\sum_{i,j}\left(y_{i,j}^{pred} - y_{i,j}^{label}\right)^2 \quad (5)$$

N_i is the number of samples, N_j is the number of receiving channels, $y_{i,j}^{pred}$ is the DNN prediction for sample i of a receiving channel j in range-Doppler representation, and $y_{i,j}^{label}$ is the associated label. $\mathcal{L}_{rdenergy}$ is a smooth L1 energy conservation loss, shown in Eqs. 6 and 7,

$$\mathcal{L}_{rd_{energy}} = \frac{1}{N_i N_j}\sum_{i,j} z_{i,j} \quad (6)$$

$$z_{i,j} = \begin{cases} 0.5\cdot\left(|y_{i,j}^{pred}| - |y_{i,j}^{label}|\right)^2 & \text{if } \|y_{i,j}^{pred}| - |y_{i,j}^{label}\| < 0.5 \\ \|y_{i,j}^{pred}| - |y_{i,j}^{label}\| - 0.5 & \text{otherwise} \end{cases} \quad (7)$$

$|y_{i,j}^{pred}|$ is the amplitude of the DNN's prediction. Displayed in Eqs. 8 and 9, $\mathcal{L}_{rdtv}$ is the total variation loss calculated over the range and Doppler dimensions:

$$\mathcal{L}_{rd_{tv}} = \frac{1}{N_i N_j}\sum_{i,j} \mathrm{tv}_{i,j} \quad (8)$$

$$\mathrm{tv}_{i,j} = \frac{1}{N_k N_l}\sum_{k,l}\|\,|y_{i,j}^{pred}(k,l)| - |y_{i,j}^{pred}(k-1,l-1)|\,\| \quad (9)$$

where (N_k, N_l) are the number of range and Doppler bins, respectively. All three loss terms were calculated per receiving channel separately to enforce tighter constraints and facilitate better reconstruction results.

The second loss term $\mathcal{L}_{bf}$, which operates in the beamformer representation space, was calculated with similar expressions for the reconstruction loss as well as the regularization terms of energy conservation and total variation. Key differences were made to encourage correct phase reconstruction. Here, the reconstruction loss $\mathcal{L}_{bfrec}$ is calculated globally to enforce coherence between the different channels as shown in Eq. 10

$$\mathcal{L}_{bf_{rec}} = \frac{1}{N_i}\sum_{i}\left(y_i^{pred} - y_i^{label}\right)^2 \quad (10)$$

In addition, energy conservation loss $\mathcal{L}_{bfenergy}$ is calculated per azimuth bin, as shown in Eq. 11:

$$\mathcal{L}_{bf_{energy}} = \frac{1}{N_i N_m}\sum_{i,m} z_{i,m} \quad (11)$$

where N_m is the number of azimuth bins and $z_{i,m}$ is described in Eq. 6. Total variation $\mathcal{L}_{bftv}$ was performed on the range and azimuth dimensions, as shown in Eq. 12 and 13:

$$\mathcal{L}_{bf_{tv}} = \frac{1}{N_i N_l}\sum_{i,l} \mathrm{tv}_{i,l} \quad (12)$$

$$\mathrm{tv}_{i,l} = \frac{1}{N_k N_m}\sum_{k,m}\|\,|y_{i,l}^{pred}(k,m)| - |\;y_{i,l}^{pred}(k-1,m-1)|\,\| \quad (13)$$

Implementation details

Training was implemented in PyTorch, the optimizer used was Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch size was 8, and learning rate used cosine decay from 3.141×10^{-4} to 3.141×10^{-7}. Training was continued until convergence and took about 25 epochs. Linear and cubic interpolations were performed using publicly available SciPy package (*60*).

Sparse array configuration

Given a radar array, R2S2 provides design flexibility in the partitioning between input and label receiving channels. As an additional example for this degree of freedom, an additional configuration is demonstrated in Fig. 5A, where an array of 16 receiving channels is split into 4 input receiving channels spread uniformly across the original array and 12 label receiving channels. Inference mode for this configuration is shown in Fig. 5B, where the 4 input receiving channels are first used with a DNN to predict 12 coherent artificial receiving channels. Afterward, both input and predicted receiving channels are arranged in their correct place in an array to allow for coherent beamforming.

This configuration is used to predict receiving channels in a MIMO virtual array based on neighboring channels. Meaning, a DNN is used to interpolate missing receiving channels in a MIMO virtual array. Performance improvement using this configuration can be achieved in two ways. First, given a specific performance metric, it is possible to decrease the number of receiving channels while still retaining high level of performance, thus saving cost and simplifying system architecture and design. Second, given a specific number of receiving channels, this configuration allows us to increase the aperture size (thus improving the angular resolution) and retain coherent beamforming with high SNR and low side lobes. This is achieved by rearranging the receiving channels and spreading them over a larger aperture size, which improves the angular resolution. However, simply increasing the distance between each receiving channel can decrease the array's performance substantially. For this end, a DNN is used to fill in the gaps with coherent artificial receiving channels and match the performance of a larger array.

Random missing channels configuration

In addition to SR, R2S2 can also be used for other purposes. In scenarios where a receiving channel becomes corrupt or exhibits performance degradation during run time operation, a DNN trained with our method can be used to replace the corrupt receiving channel with an artificial receiving channel. To accomplish this, R2S2 is used to predict random missing receiving channels from the remainder active radar array. This task is especially difficult for a DNN because the receiving channels are randomly chosen and can also be located at the edges of the array, meaning the DNN needs to extrapolate and interpolate.

To create a DNN that is invariant to the position of a missing receiving channel, a full MIMO virtual array is used as input, and randomly chosen receiving channels are masked, whereas the DNN is tasked to predict the missing receiving channels. The resulting trained DNN is invariant to the specific receiving channel missing and is able to reconstruct the data of each receiving channel individually without the need to train a separate model for each receiving channel. An illustration of the training methodology for this configuration is provided in Fig. 7.

SUPPLEMENTARY MATERIALS

www.science.org/doi/10.1126/scirobotics.abk0431
Figs. S1 to S11
Tables S1 to S3

REFERENCES AND NOTES

1. L. M. Clements, K. M. Kockelman, Economic effects of automated vehicles. *Transp. Res. Rec.* **2606**, 106–114 (2017).
2. *Road Safety Annual Report 2019* (International Traffic Safety Data and Analysis Group, 2019), pp. 1–9.
3. SAE International, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles* (SAE International, 2018), pp. 1–5.
4. International Organization for Standardization, ISO 26262-1:2018 (2018); www.iso.org/standard/68383.html.
5. O. Bialer, A. Jonas, T. Tirer, Super resolution wide aperture automotive radar. *IEEE Sens. J.* **21**, 17846–17858 (2021).
6. R. H. ROY, A. Paulraj, T. Kailath, ESPRIT—A subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE Trans. Acoust.* **34**, 1340–1342 (1986).
7. R. O. Schmidt, Multiple emitter location and signal parameter estimation. *Adapt. Antennas Wirel. Commun.* , 190–194 (1986).
8. R. Komissarov, V. Kozlov, D. Filonov, P. Ginzburg, Partially coherent radar unties range resolution from bandwidth limitations. *Nat. Commun.* **10**, 1423 (2019).
9. I. Orr, M. Cohen, Z. Zalevsky, High-resolution radar road segmentation using weakly supervised learning. *Nat. Mach. Intell.* **3**, 239–246 (2021).
10. N. Scheiner, N. Appenrodt, J. DIckmann, B. Sick, Radar-based feature design and multiclass classification for road user recognition. *IEEE Intell. Veh. Symp.* , 779–786 (2018).
11. K. Patel, K. Rambach, T. Visentin, D. Rusev, M. Pfeiffer, B. Yang, Deep learning-based object classification on automotive radar spectra, in *2019 IEEE Radar Conf. RadarConf* (IEEE, 2019).
12. A. Palffy, J. Dong, J. F. P. Kooij, D. M. Gavrila, CNN based road user detection using the 3D radar cube. *IEEE Robot. Autom. Lett.* **5**, 1263–1270 (2020).
13. B. Major, D. Fontijne, R. T. Sukhavasi, M. Hamilton, S. Lee, S. Grzechnik, S. Subramanian, Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2019).
14. Z. Feng, S. Zhang, M. Kunert, W. Wiesbeck, Applying neural networks with a high-resolution automotive radar for lane detection, in *10th GMM-Symposium, AmE 2019—Automotive meets Electronics* (VDE, 2019).
15. P. Kaul, D. De Martini, M. Gadd, P. Newman, RSS-Net: Weakly-supervised multi-class semantic segmentation with FMCW radar. arXiv:2004.03451 [cs.CV] (2 April 2020).
16. O. Schumann, M. Hahn, J. Dickmann, C. Wöhler, Semantic segmentation on radar point clouds. *2018 21st International Conference Information Fusion (FUSION)* (IEEE, 2018).
17. A. M. Elbir, K. V. Mishra, Y. C. Eldar, Cognitive radar antenna selection via deep learning. *IET Radar Sonar Navig.* , 871–880 (2019).
18. J. Gao, B. Deng, Y. Qin, H. Wang, X. Li, Enhanced radar imaging using a complex-valued convolutional neural network. *IEEE Geosci. Remote Sens. Lett.* **16**, 35–39 (2019).
19. J. Zhong, G. Wen, C. Ma, B. Ding, Radar signal reconstruction algorithm based on complex block sparse Bayesian learning, in *2014 12th International Conference Signal Processing (ICSP)* (IEEE, 2014).
20. M. Rossi, A. M. Haimovich, Y. C. Eldar, Spatial compressive sensing for MIMO radar. *IEEE Trans. Signal Process.* **62**, 419–430 (2014).
21. F. Marvasti, A. Amini, F. Haddadi, M. Soltanolkotabi, B. H. Khalaj, A. Aldroubi, S. Sanei, J. Chambers, A unified approach to sparse signal processing. *EURASIP J. Adv. Signal Process.* **2012**, 44 (2012).
22. F. Roos, H. Philipp, L. Lorraine, T. Torres, C. Knill, J. Schlichenmaier, C. Vasanelli, N. Appenrodt, Compressed sensing based single snapshot DoA estimation for sparse MIMO radar arrays, in *2019 12th German Microwave Conference (GeMiC)* (IEEE, 2019).
23. T. Strohmer, B. Friedlander, Compressed sensing for MIMO radar—Algorithms and performance, in *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers* (IEEE, 2009).
24. K. Armanious, S. Abdulatif, F. Aziz, U. Schneider, B. Yang, An adversarial super-resolution remedy for radar design trade-offs, in *2019 27th European Signal Processing Conference (EUSIPCO)* (IEEE, 2019).
25. M. Gall, M. Gardill, T. Horn, J. Fuchs, Spectrum-based single-snapshot super-resolution direction-of-arrival estimation using deep learning, in *2020 German Microwave Conference (GeMiC)* (IEEE, 2020).
26. L. Wu, Z. M. Liu, Z. T. Huang, Deep convolution network for direction of arrival estimation with sparse prior. *IEEE Signal Process. Lett.* **26**, 1688–1692 (2019).
27. M. Agatonovic, Z. Stanković, B. Milovanović, High resolution two-dimensional DOA estimation using artificial neural networks, in *2012 6th European Conference on Antennas and Propagation (EUCAP)* (IEEE, 2012).
28. J. Fuchs, R. Weigel, M. Gardill, Single-snapshot direction-of-arrival estimation of multiple targets using a multi-layer perceptron, in *2019 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)* (IEEE, 2019).
29. M. Agatonović, Z. Stanković, I. Milovanović, N. Dončov, L. Sit, T. Zwick, B. Milovanović, Efficient neural network approach for 2D DOA estimation based on antenna array measurements. *Prog. Electromagn. Res.* **137**, 741–758 (2013).
30. Y. Lecun, Self Supervised Learning—Keynote lecture. *ICLR* (2020); www.youtube.com/watch?v=8TTK-Dd0H9U&ab_channel=AIP-PursuingSoTAAIforeveryone.
31. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in *Proceedings of the 37th International Conference on Machine Learning* (PMLR, 2020).
32. S. Laine, T. Karras, J. Lehtinen, T. Aila, High-quality self-supervised deep image denoising, in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (ACM, 2019).
33. X. Zhan, Mix-and-match tuning for self-supervised semantic segmentation, in *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI, 2018).
34. S. Singh, Self-supervised feature learning for semantic segmentation of overhead imagery, in *British Machine Vision Convention* (BMVC, 2018).
35. M. Chen, T. Artières, Unsupervised object segmentation by redrawing, in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (ACM, 2019).
36. D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, A. Zisserman, G. Brain, Temporal cycle-consistency learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2019).
37. E. Rodol, A. Bronstein, R. Kimmel, Unsupervised learning of dense shape correspondence, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2019).
38. M. Noroozi, P. Favaro, Unsupervised learning of visual representations by solving jigsaw puzzles, in *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science*, B. Leibe, J. Matas, N. Sebe, M. Welling, Eds. (Springer, 2016).
39. M. Janner, J. Wu, T. D. Kulkarni, I. Yildirim, J. B. Tenenbaum, Self-supervised intrinsic image decomposition. arXiv:1711.03678 [cs.CV] (10 November 2017).
40. S. Gidaris, P. Singh, N. Komodakis, Unsupervised representation learning by predicting image rotations. arXiv:1803.07728 [cs.CV] (21 March 2018).
41. Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*, V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss, Eds. (Springer, 2018), vol. 11211.
42. X. Wang, K. C. K. Chan, C. Dong, C. C. Loy, EDVR: Video restoration with enhanced deformable convolutional networks, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (IEEE, 2019).
43. B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, Enhanced deep residual networks for single image super-resolution, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (IEEE, 2017).
44. C. Dong, C. C. Loy, Deep spatial feature transform, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2018).

45. Y.-C. Wang, S. Venkataramani, P. Smaragdis, Self-supervised learning for speech enhancement. arXiv:2006.10388 [eess.AS] (18 June 2020).
46. B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, M. Velimirovi, SPICE: Self-supervised pitch estimation, in *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (IEEE, 2020).
47. J. Engel, R. Swavely, A. Roberts, L. Hanoi, H. Curtis, Self-supervised pitch detection by inverse audio synthesis, in *ICML 2020 Workshop SAS* (ICML, 2020).
48. S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, J. R. Hershey, Unsupervised speech separation using mixtures of mixtures, in *ICML 2020 Workshop SAS* (ICML, 2020).
49. A. Saeed, D. Grangier, N. Zeghidour, Contrastive learning of general-purpose audio representations, in *ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (IEEE, 2020).
50. M. Ravanelli, Y. Bengio, U. De Montréal, Learning speaker representations with mutual information. arXiv:1812.00271 [eess.AS] (1 December 2018).
51. M. Tagliasacchi, D. Roblek, Self-supervised audio representation learning for mobile devices. arXiv:1905.11796 [eess.AS] (24 May 2019).
52. H. Banville, I. Albuquerque, A. Hyvärinen, G. Moffat, D. A. Engemann, A. Gramfort, Self-supervised representation learning from electroencephalography signals, in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)* (IEEE, 2019).
53. P. Sarkar, A. Etemad, Self-supervised learning for ECG-based emotion recognition, in *ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2020).
54. N. Long, K. Wang, R. Cheng, K. Yang, W. Hu, J. Bai, Assisting the visually impaired: Multitarget warning through millimeter wave radar and RGB-depth sensors. *J. Electron. Imaging* **28**, 013028 (2019).
55. J. Li, P. Stoica, *MIMO Radar Signal Processing* (Wiley, 2009).
56. B. Friedlander, On signal models for MIMO radar. *IEEE Trans. Aerosp. Electron. Syst.* **48**, 3655–3660 (2012).
57. H. Yang, W. Liu, W. Xie, Y. Wang, General signal model of MIMO radar for moving target detection. *IET Radar Sonar Navig.* **11**, 570–578 (2017).
58. O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. Mcdonagh, N. Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention U-Net: Learning where to look for the pancreas. arXiv:1804.03999 [cs.CV] (11 April 2018).
59. J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2019).
60. scipy.org, SciPy Version: 1.7.1.

Acknowledgments: We thank Y. Avargel, Z. Iluz, L. Korkidi, K. Twizer, H. Omer, E. Cohen, and N. Orr for their advice and help in revising the manuscript. **Funding:** This research did not receive any specific grant from funding agencies. **Author contributions:** Conceptualization: I.O. Methodology: I.O. and M.C. Data collection: H.D. Investigation: I.O., M.C., H.D., M.H., M.R., and Z.Z. Supervision: M.C. and Z.Z. Writing, review, and editing: I.O., M.C., H.D., M.H., M.R., and Z.Z. **Competing interests:** I.O., H.D., and M.C. are inventors on a patent application (US-17/205,283) held/submitted by WiSense Technologies Ltd. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 17 June 2021
Accepted 19 November 2021
Published 15 December 2021
10.1126/scirobotics.abk0431

Science Robotics

Coherent, super-resolved radar beamforming using self-supervised learning

Itai Orr, Moshik Cohen, Harel Damari, Meir Halachmi, Mark Raifel, and Zeev Zalevsky

Sci. Robot. **6** (61), eabk0431. DOI: 10.1126/scirobotics.abk0431

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.