PeakConv: Learning Peak Receptive Field for Radar Semantic Segmentation

Liwen Zhang*[1] Xinyan Zhang*[2] Youcheng Zhang[1] Yufei Guo[1] Yuanpei Chen[1] Xuhui Huang[1] Zhe Ma†[1]
[1]Intelligent Science & Technology Academy of CASIC, Beijing 100144, China
[2]Faculty of Computing, Harbin Institute of Technology, Heilongjiang 150001, China
lwzhang9161@126.com zxy20020109@gmail.com mazhe_thu@163.com

Abstract

The modern machine learning-based technologies have shown considerable potential in automatic radar scene understanding. Among these efforts, radar semantic segmentation (RSS) can provide more refined and detailed information including the moving objects and background clutters within the effective receptive field of the radar. Motivated by the success of convolutional networks in various visual computing tasks, these networks have also been introduced to solve RSS task. However, neither the regular convolution operation nor the modified ones are specific to interpret radar signals. The receptive fields of existing convolutions are defined by the object presentation in optical signals, but these two signals have different perception mechanisms. In classic radar signal processing, the object signature is detected according to a local peak response, i.e., CFAR detection. Inspired by this idea, we redefine the receptive field of the convolution operation as the ***peak receptive field (PRF)*** *and propose the* ***peak convolution operation (PeakConv)*** *to learn the object signatures in an end-to-end network. By incorporating the proposed PeakConv layers into the encoders, our RSS network can achieve better segmentation results compared with other SoTA methods on a multi-view real-measured dataset collected from an FMCW radar. Our code for PeakConv is available at https://github.com/zlw9161/PKC.*

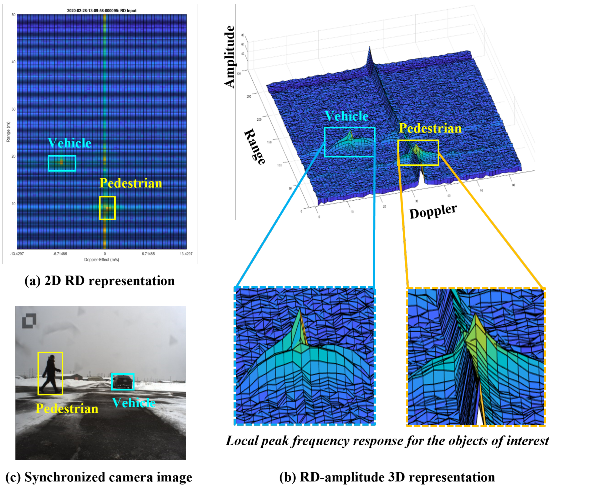
1. Introduction

Radar is a remote sensor, which usually uses modulated electromagnetic signals to detect the objects of interest through directional transmitting antennas in a specific effective working field [22]. As an active detection device, radar is more robust to extreme weather (*e.g.*, haze, rain or snow) than other active detection device such as LiDARs [2], and it is also not susceptible to dim light condition and sun glare, as the passive optical sensors are [19]. In addition to the real-world location information, it can also tell the velocity of the moving objects thanks to the Doppler effects. Due to these advantages, radar sensors have played an irreplaceable role for many automotive security and defense applications, *e.g.*, autonomous safety driving or UAV early warning.

Conventional radar detection mostly relies on the peak detection algorithm following constant false alarm rate (CFAR) [22, 23] principle. Taking frequency modulated continuous wave (FMCW) radar as example, the raw radar echos are first converted as multi-domain united frequency representations, *e.g.*, range-Doppler (RD) and range-angle (RA) maps, through a series of cascading fast Fourier transformations (FFTs). Then for each cell under test (CUT) in the input RD/RA map, the CFAR detector will determine whether it contains moving object information according to an estimated detection threshold, which fully considers the characteristics of the radar signal itself. However, to obtain good effect in practical application, it is necessary to manually fine-tune various hyper-parameters including the thresholding factor, sizes and shapes of the local scope (*i.e.*, the bandwidths of reference and guard units). Beyond that, conventional radar detection cannot give category information of the object. These two inconveniences hinder the conventional detection method from automatic semantic radar scene understanding.

Encouraged by the success of modern deep learning techniques in computational perception, especially the object detection [8, 15, 20, 21, 29] and semantic segmentation [5, 11, 16, 24, 28] in computer vision, some efforts had been made recently for better automatic radar scene interpretation. These efforts evolve the target-clutter binary hypothesis of conventional radar testing into target semantic characterization of modern machine learning, *i.e.*, radar object detection (ROD) [10, 17, 27] and radar semantic segmentation (RSS) [3, 13, 18]. Most of these methods used convolution networks as backbone models, which take radar frequency representations as input, and then make predictions on RA or RD view or both two views. For example, a multi-view RSS (MVRSS) network [18] was proposed

*Equal contribution. †Corresponding author. This research is supported by Young Science Foundation of National Natural Science Foundation of China (No.62206258).



(a) 2D RD representation

Local peak frequency response for the objects of interest

(c) Synchronized camera image

(b) RD-amplitude 3D representation

Figure 1. An examplar illustration of moving object signatures/presentations in (a) the 2D RD map and the corresponding (b) RD-amplitude 3D representation of radar signals, and their (c) synchronized camera image.

to take better advantage of radar localization capability by making "unit-wise" predictions on both RD and RA frequency domains. To support the sufficient training of these deep models, a few large-scale radar datasets were also collected and created, *e.g.*, OxfordRobotCar [9], nuScenes [4], CRUW [26] and CARRADA [19].

However, the electromagnetic object signatures received by radar are not as intuitively understood as the optical ones captured by the cameras as shown in Fig. 1. With rich texture and color information in the image, the convolution operation can learn useful semantic information from a rectangular local spatial receptive field (RF). And by introducing some intuitive priors of human vision, more efficient learning mechanisms for convolution had been proposed, *e.g.*, multi-scale fusion [12, 15, 25], dilation [5, 28] and deformation [8, 29]. So far, these mechanisms are also introduced into radar data processing, such as the inception or pyramid pooling for multi-scale information, atrous convolution for larger dilated RF and deformable convolution for irregular object signature in ROD-Net [27] and MVRSS [18]. Despite the multi-scale mechanism, which is more of a modular idea, *i.e.*, the computation is decoupled from the convolution itself, other variants are actually changing the RF itself. One conclusion might be summed up that, the RF sampling/selection manner plays a very important role in convolution. While none of these RF selection manners including the regular one is proposed specifically for the radar data, thus they might not fully exploit the potential of convolutional networks in radar scene understanding. This concern motivates us to rethink the internal relation between convolution and the conventional radar detection mechanism, and try to find a more efficient and specific convolution mechanism for radar data.

To achieve our goal, we take a look inside of the conventional radar detection method and the convolution operation in deep learning. As aforementioned, the conventional detection method is a kind of CFAR-based peak detection, *e.g.*, commonly used cell averaging-CFAR (CA-CFAR) [22]. For a CUT, x_c, of the input RD representation, CA-CFAR detection can be divided into three steps: (i) averaging aggregation from reference cells $\{x_r^{(i)}\}_{i=1}^N$ around CUT, excluding the guard cells; (ii) threshold computing, $\Theta = \xi \cdot \frac{1}{N}\sum_{i=1}^N x_r^{(i)}$; (iii) decision-making by comparing x_c and Θ. It can be seen that, the decision-making basis is the difference between CUT and its threshold, *i.e.*, the weighted summation of $\{x_r^{(i)}\}_{i=1}^N$ with a shared weight, $\frac{\xi}{N}$. In another word, the key to determine whether the CUT has object for CA-CFAR is the denoised peak frequency response from an RF consisted of the CUT and its reference cells. Yet none of the convolution operators mentioned above can explicitly possess such property, *i.e.*, each output unit is actually a weighted summation of the units in a local dense/dilated rectangular or deformable RF, which does not strictly follow the guard-reference policy.

Therefore, in this work we redefine the RF of the convolution operator as the guard-reference style, and call such new type RF the peak receptive field (PRF), which consists of the center unit and its reference neighbors. Then with some simple computational designs, we present two novel convolution operations to explicitly learn the peak response from PRF, *i.e.*, PeakConvs. Compared with other convolution operations, PeakConvs explicitly possess the advantage of the conventional radar detection methods. In comparison with the conventional CA-CFAR, adaptive peak response with learnable weights and high-level semantic representation via task-driven learning paradigm can be achieved since PeakConvs maintain the computational compatibility of the regular convolution operation. The main contributions are:

- **A novel convolution computing paradigm for radar data processing**. Instead of extracting radar signature directly from RF, we propose learning peak response from redefined PRF, which is more suitable for learning tasks related to radar data.
- **Two implementations of the proposed PeakConv**. According to the participation of center unit during interference (*e.g.*, device noises and background clutters) estimation, there are two approaches of PeakConv, including vanilla-PeakConv (PKC), and response difference aware PeakConv (ReDA-PKC).
- **Well-performed multi-view RSS frameworks based on PeakConvs**: by introducing PeakConvs into encoders of the convolutional automatic-encoder-decoder (CAED) framework, two RSS networks with

multi-input and multi-output (MIMO) style are presented. Our networks can achieve SoTA performance on both RD and RA views.

2. Related works

RF-based convolution improvement. Reasonable adjustment of RF can effectively improve the expressive ability of convolution in module-level or in operation-level. Multi-scale information extraction is one of the most commonly discussed module-level strategies, *e.g.*, capturing rich features via convolution layers with different kernel sizes in a hierarchical pyramid structure [12, 15] or parallel forked form [25]. Some works focused on directly changing the shape of the dense rectangular RF of the regular convolution operation, *i.e.*, changing the sampling manner of the kernel. Dilation [28] can effectively enlarge the RF via a fixed step size-based sparse sampling strategy, and by introducing dilation into the pyramid structured convolution module, the parameter scale can be effectively reduced while maintaining the multi-scale characterization ability, *e.g.*, atrous spatial pyramid pooling (ASPP) [5]. Deformation [8, 29] is a much more flexible way of sampling, by learning the position offset for the rectangular kernel, irregular RFs can be obtained to better handle the variety of object shapes and sizes. These works are mostly motivated by the perspective of visible light signals. On the side of radar, for now, there is barely no discussion about suitable convolution mechanisms specifically for radar signal processing.

Radar semantic segmentation. Compared with detection-based method using bounding boxes with regular shapes, the segmentation-based models can provide the “pixel” (a cell/unit) wise predictions for the input radar frequency tensor, including the objects and even the background. This characteristic makes segmentation-based method more suitable for radar scene understanding. The deep radar detector (DRD) [3] is an early attempt to incorporate a segmentation approach into a two-stage radar detection workflow, where a UNet [24] style network is treated as an object proposal generator for the angle predictor. Compared with the conventional CA-CFAR and beam-forming localization [1], DRD can provide more accurate predictions and show better robustness. Recently, an end-to-end RSS framework without any post-processing techniques, RSS-Net [13] was proposed. By taking the temporal changes of multi-frames radar input and multi-scale spatial information obtained by the ASPP [5] into consideration, good RA predictions for some particular objects can be obtained. More recently, by introducing multi-perspective learning [10] strategy into the end-to-end RSS framework, the MVRSS [18] was proposed. MVRSS uses range-angle-Doppler (RAD) frequency tensors as input, and makes predictions on both RA and RD views. Such simultaneous multi-view predictions make MVRSS a good start for radar scene understanding, which is also followed by this work.

To our knowledge, all these previous methods tried to systematically solve the RSS task via rationally utilizing the advantages of existing convolutional networks. Although these existing computational module can help improve the RSS performance, they are proposed to better characterize the object from the perspective of visual perception, after all. Directly using them to process radar data is more or less computationally redundant and inefficient. To this end, we stand on the side of radar, and revisit the correlation between the modern convolution operations and the conventional radar detection method. Then we attempt to propose a more suitable convolution operator for radar signal processing without losing its original advantages, starting from RF definition to computation designs.

3. Peak Convolution

In this section, the proposed PeakConv will be introduced in details, starting from the PRF definition to the specific implementations of the PeakConv. Then a global description of the PeakConv-based RSS network is given.

3.1. Peak Receptive Field

As illustrated in Fig. 1, the object signatures in radar signals are more like a peak-shaped frequency response in a local scope of the radar representation. Without considering the resolution of the radar itself, coordinates of the analysis domain (RD or RA) in which the peak valley vertex is located will be regarded as the state of the object in the real physical world, such as range, velocity, angle relative to radar. Such characteristic is utilized by the conventional detection methods, *e.g.*, CA-CFAR [23].

To capture the peak response as robustly as possible, two basic principles are followed: (i) local search strategy and (ii) guard band mechanism. The local search makes sure the object-related peak response cannot be suppressed by the interference with stronger energy, which is similar to the scanning process of the convolution kernel. Furthermore, setting guard band around the center unit (or CUT) will ensure its energy dose not leak into the interference estimation process, so that its response can be captured more clearly. As intuitively shown in Fig. 2, to determine whether a unit x_c from the input radar frequency representation has target information, the detector will select reference units around x_c, and exclude those units closest to x_c. Those excluded ones are so-called guard units. Then the selected reference units, $\{x_r^{(i)}\}_{i=1}^{N_r}$, will be used to calculate the detection threshold, Θ, as we analyzed in Sec. 1.

As can be seen from the above analysis, if there has object in the center unit, the peak response can be determined by the response of the center unit and the average response of its reference neighbors adjusted by a coefficient, ξ. Inspired by this mechanism, we argue that, it might be more

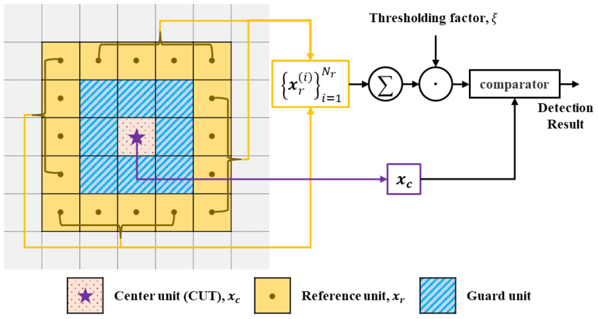


Figure 2. The illustration for capturing peak frequency response in 2D radar representation by CA-CFAR.

efficient to process the radar data, if the convolution operator in modern deep learning could capture such peak response. And as discussed before, the RF selection is a key factor for convolutions, hence we start from redefining the RF as PRF following the basic principles of the conventional CFAR method.

Given some feature point $\mathbf{x}_c \in \mathbb{R}^C$ from the input map with C channels, its PRF is consisted of $\mathbf{x}_c$ itself and its reference points, $\{\mathbf{x}_r^{(i)}\}_{i=1}^{N_r}$, where N_r is the number of the reference points. Let $\mathbf{p}_c = (p_c^x,\ p_c^y)$ and $\mathbf{p}_r^{(i)} = (p_{r,x}^{(i)},\ p_{r,y}^{(i)})$ denote the column and row coordinates of $\mathbf{x}_c$ and $\mathbf{x}_r^{(i)}$, respectively in the input map. Then the PRF, $\mathcal{R}$, for $\mathbf{x}_c$ can be defined as follows:

$$\mathcal{R} = \left\{\mathbf{x}_c,\ \left\{\mathbf{x}_r^{(i)}\right\}_{i=1}^{N_r}\right\}, \quad s.t., |\mathbf{b}_G| < |\mathbf{p}_r^{(i)} - \mathbf{p}_c| \le |\mathbf{b}_R + \mathbf{b}_G|. \tag{1}$$

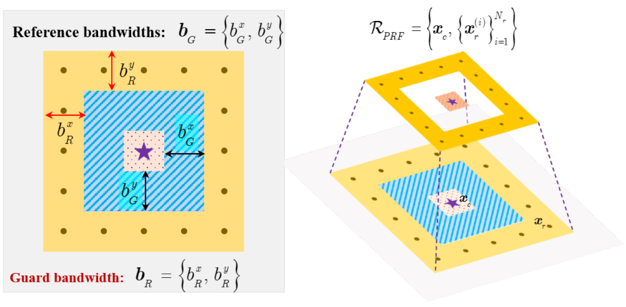
Where, $\mathbf{b}_G = (b_G^x,\ b_G^y)$ and $\mathbf{b}_R = (b_R^x,\ b_R^y)$ respectively denote the guard bandwidths and reference bandwidths in column and row as shown in Fig. 3-(a). All the coordinates and bandwidths are integers, and the number of reference points can be computed as:

$$N_r = \prod_{z\in\{x,y\}} [2(b_G^z + b_R^z) + 1] - \prod_{z\in\{x,y\}} (2b_G^z + 1). \tag{2}$$

3.2. Learning from PRF

3.2.1 Vanilla-PKC

Using the definition of PRF, we can further define the learning mechanism of the proposed PeakConv. As aforementioned, the decision-making basis of CA-CFAR is the denoised peak frequency response estimated from the CUT and its reference units, which can also be regarded as the representation of the object radar signature. Hence to embed such property in the convolution-based deep models, we define the vanilla-PKC,


Reference bandwidths: $b_G = \{b_G^x, b_G^y\}$
b_R^y
b_R^x
b_G^x
b_G^y
Guard bandwidth: $b_R = \{b_R^x, b_R^y\}$
$\mathcal{R}_{PRF} = \{x_c, \{x_r^{(i)}\}_{i=1}^{N_r}\}$
x_c
x_r

(a) Illustration of guard/reference bandwidths

(b) Illustration of Peak Receptive Field

Figure 3. Guard/reference bandwidths and PRF.

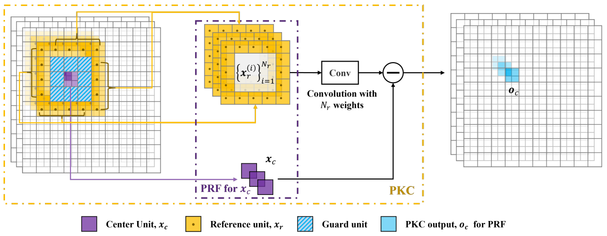

$\{x_r^{(i)}\}_{i=1}^{N_r}$
Conv
Convolution with N_r weights
x_c
PRF for x_c
PKC
o_c
Center Unit, x_c
Reference unit, x_r
Guard unit
PKC output, o_c for PRF

Figure 4. The the whole process for a PeakConv layer.

$PKC\left(\mathcal{R};\ \mathbf{W} \in \mathbb{R}^{C_{\text{in}} \times N_r \times C_{\text{out}}}\right) : \mathbb{R}^{C_{\text{in}}} \to \mathbb{R}^{C_{\text{out}}}$ as follows,

$$PKC\left(\mathcal{R};\ \mathbf{W}\right) = \mathbf{x}_c - Vec\left(\left\{\textstyle\sum_{i=1}^{N_r} \mathbf{w}_j^{(i)} * \mathbf{x}_r^{(i)}\right\}_{j=1}^{C_{\text{out}}}\right), \quad where,\ \mathbf{w}_j^{(i)} \in \mathbb{R}^{C_{\text{in}}}\ (j = 1, \cdots, C_{\text{out}})\ and\ C_{\text{in}} = C_{\text{out}}. \tag{3}$$

Where, $\mathbf{W}$ is the learning weights of the PeakConv shared by all the $\mathbf{x}_c$ in the input feature maps; $Vec(\cdot)$ and $*$ denote the vectorization and convolution operators, respectively. Compared with the fixed pre-defined global weight, $\frac{\xi}{N}$ in CA-CFAR, the PeakConv can automatically reweight the reference units in PRF under a task-driven manner. Therefore, $PKC\left(\mathcal{R};\ \mathbf{W}\right)$ can be regarded as the learnable noise suppression for each unit under test, $\mathbf{x}_c$, in which $Vec\left(\{\sum_{i=1}^{N_r} \mathbf{w}_j^{(i)} * \mathbf{x}_r^{(i)}\}_{j=1}^{C_{\text{out}}}\right)$ actually estimates the interference frequency response. The whole process for a PeakConv layer is shown in Fig. 4, which can be summarized as: (i) sampling PRFs over the input maps, *i.e.*, collecting $\{\mathbf{x}_r^{(i)}\}_{i=1}^{N_r}$ for each $\mathbf{x}_c$; (ii) learning interference frequency response from each PRF by a convolution operator with kernel size of N_r; (iii) denoising each $\mathbf{x}_c$ by eliminating its corresponding learned interference response.

3.2.2 ReDA-PKC

The PeakConv essentially attempts to characterize the response difference between each unit under test and its interference response estimated from its neighboring reference

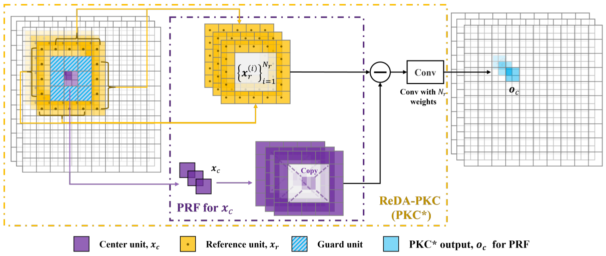


Figure 5. The the whole process for a ReDA-PKC layer.

units. Moreover, the learning process of PeakConv mentioned above is performed under the target-independent assumption, since $\mathbf{x}_c$ is excluded from the estimation. These could lead to two limitations for vanilla $PKC(\cdot)$ in Eq. (3): i) on the interference side, the overall estimation would ignore the individual diversity of interference, *e.g.*, relatively mild random noise caused by sensor or environmental background and clutters with strong energy caused by some objects of no interest, and this makes PeakConv fails to suppress noise in a more interference-driven manner; ii) on the target side, the target-independent assumption would not directly drive the PeakConv layer to carry out noise suppression or response difference characterization according to the target. To alleviate the above two limitations, we further present another variant, which is called ReDA-PKC.

ReDA-PKC directly learns the response difference between $\mathbf{x}_c$ and each of its reference units, $\mathbf{x}_r^{(i)}$. Following the definition of vanilla-PKC in Eq. (3), the ReDA-PKC, $PKC^{\star}\left(\mathcal{R};\ \mathbf{W}\in\mathbb{R}^{C_{\text{in}}\times N_r\times C_{\text{out}}}\right):\mathbb{R}^{C_{\text{in}}}\rightarrow\mathbb{R}^{C_{\text{out}}}$ for each given $\mathbf{x}_c$ can be formalized as follows,

$$PKC^{\star}\left(\mathcal{R};\ \mathbf{W}\right)=Vec\left(\left\{\textstyle\sum_{i=1}^{N_r}\mathbf{w}_j^{(i)}*\left(\mathbf{x}_c-\mathbf{x}_r^{(i)}\right)\right\}_{j=1}^{C_{\text{out}}}\right),$$
$$where,\ \mathbf{w}_j^{(i)}\in\mathbb{R}^{C_{\text{in}}}\ (j=1,\cdots,C_{\text{out}}). \tag{4}$$

It can be seen from Eq. (4) that, $PKC^{\star}(\cdot)$ directly performs on the difference between $\mathbf{x}_c$ and each $\mathbf{x}_r^{(i)}$. Such form of PeakConv would focus directly on the difference in each center-reference response, so that the diversity of noise relative to the target can be better handled by the whole ReDA-PKC layer. The learning process of ReDA-PKC is shown in Fig. 5. Besides, by encapsulating all the units of $\mathcal{R}$ into the learning part, the constraint $C_{\text{in}}=C_{\text{out}}$ can be eliminated, thus making ReDA-PKC a more flexible module compared with the vanilla version.

3.3. PeakConv-based RSS Network

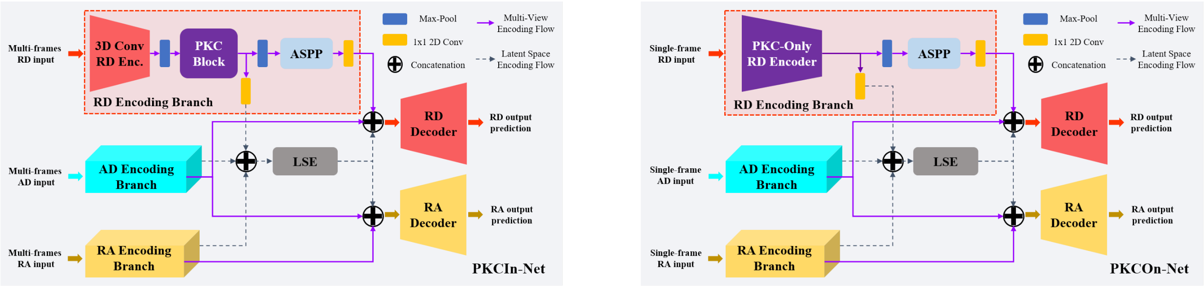
Both vanilla and ReDA versions of PKC can be easily embedded in the exiting convolutional networks as regular convolution blocks. According to different deployments of PKC layer, we further propose two PKC-based RSS frameworks following the CAED framework used in [13,18,27], *i.e.*, PKC-Inserted RSS-Net (PKCIn-Net) and PKC-Only RSS-Net (PKCOn-Net), as illustrated in Fig. 6. To achieve comprehensive radar scene understanding, both two RSS networks are designed in the multi-input-multi-output (MIMO) style, which performs on the RD, RA and AD (Angle-Doppler) radar frequency tensors and carries out semantic segmentation on both RD and RA views. Then each network is consisted of three encoding branches and two decoding branches. For PKCIn-Net, the main components of the encoder are as follows.[1]

Basic Encoder. For each single-view radar input sequence, the basic encoder is used to generate its high-level representations, and multi-view radar sequence can be handled by performing these encoders in parallel. All three single-view encoders are designed with the same structure for simplicity. To utilize temporal information, each encoder is mainly composed of two 3D convolution layers with the same kernel size of $3\times3\times3$, output channel of 128, and stride of 1 for spatial and temporal domains. To further compress the spatial size of feature maps, each convolution block follows a spatial max-pooling layer. By controlling kernel size of the pooling layer, the feature map size for each view can be unified.

Vanilla/ReDA-PKC Block. For learning the important decision-making basis, *i.e.,* peak frequency response in the radar data, two PKC layers are inserted in each single-view encoding branch. The PKC module will take the high-level representations obtained from the basic encoder as input, and generate local peak response representation on each spatial position of the input maps. According to the definition of PRF in Eq. (1), the kernel size of each PKC layer, N_r, is depend on the guard/reference bandwidths, $\mathbf{b}_{G/R}$. In basic frameworks, we set both $\mathbf{b}_G$ and $\mathbf{b}_R$ as $\{1,\ 1\}$ by default. Then N_r is equal to 16 following Eq. (2). The output channel for each PKC layer is keeping the same with the 3D convolution in the basic encoder.

Latent Space Encoder (LSE). The MIMO-style network optimization involves multi-perspective/view learning problem. The common way for addressing such problem is to learn the joint representation of the information flows from different views in a common feature space. Thus different encoding branches can interact with each other via the common feature space during learning, and the representations to be decoded can also be further enhanced with the learned joint feature. For this purpose, a multi-view shared learner, LSE is added. To form the inputs for LSE, the output maps of each single-view encoding branch will be further compressed and transformed by a 2×2 max-pooling layer and a 2D convolution layer with 1×1 kernels. Then the LSE will project those transformed features into a

[1] In PKCOn-Net, **each encoding branch is solely consisted of PKC layers**, see details in Appendix at our `code` link mentioned in Abstract.



(a) PKCIn-Net

(b) PKCOn-Net

Figure 6. The overall workflow of PeakConv-based RSS-Net.

common latent space to get a uniform embedding. For reducing the model parameters, two 1×1 2D convolution layers instead of the MLP are chosen to form the LSE.

Besides the above components, ASPP [6] module, which allows features to be jointly learned at different scales with no need for larger kernels or additional parameters [5], is also applied for enhancing the spatial representation in encoding part. This module is well suited for RSS task, since the moving objects' appearances may vary a lot [13, 18]. The decoders for RD and RA views are designed to be combination of 2D deconvolutions and convolutions like [18].

4. Experiments

The effectiveness of the proposed PeakConvs are verified under two designed RSS frameworks, *i.e.*, PKCIn-Net with temporal-style (multi-frames) input and PKCOn-Net with single-frame input. For simplicity of the tables, PKCOn and PKCOn* denotes PKCOn-Net with vanilla-PKC and ReDA-PKC, respectively, so do PKCIn and PKCIn*.

4.1. Datasets and Training Setups

4.1.1 CARRADA Dataset

The CARRADA [19] dataset is a large scale camera-radar synchronised dataset, which contains multi-view annotated radar recordings (RAD tensors) collected from a low-cost FMCW radar in various scenarios under different weathers conditions. So far, this is the only publicly available radar dataset supporting multi-view RSS task. There are 4 categories of objects: *pedestrian*, *cyclist*, *vehicle* and *background*; and multiple same/different types of moving objects would appear in a single time frame. The dimensions of RAD tensor are 256, 256 and 64, respectively. The training, validation and test subsets were split as in [18].

4.1.2 CARRADA Annotation Calibration

The RD and RA annotations of CARRADA were generated in a semi-automatic way [19]. Its key idea is to use the Mean-Shift-based tracking method [7] to associate the object information collected from the camera images and radar data into a unified physical world domain, *i.e.*, a 3D cartesian space combined with DoA (Directional of Arrival) and Doppler. However, the unreliable depth estimation results of the camera data and low angle resolution of the FMCW radar seriously affect the Mean-Shift clustering performance in terms of the centroid initialization and candidate search space. Hence, there is non-negligible deviation between the generated annotations and the ground truth locations, especially in the RA view. To this end, we further calibrate the RA annotations for CARRADA in this work, and provide a higher quality multi-view radar dataset, RA Annotation Calibrated CARRADA (CARRADA-RAC)[2]. The performance comparisons of CARRADA and CARRADA-RAC are in Sec. 4.5.

4.1.3 Training Setup

The training setup for our proposed models and other compared SoTA networks is strictly consistent as follows:

Input form: The raw 3D RAD tensor of each time frame is first compressed as 2D RA, AD and RD views with the sizes of 256×256, 256×256 and 256×64, respectively. For the networks without explicitly considering the temporal information, including RSS-Net [13] and MVA-Net [18], which use 3 frames for data augmentation, while our PKCOn-Net only depends on single frame for prediction. For the temporal encoding-based networks, more frames are used to form the input sequence, *e.g.*, 9 frames for RAMP-CNN [10], and 5 frames for TMVA-Net [18] as well as our PKCIn-Net.

Hyper-parameters: All the models were trained with Adam optimizer [14] using the default setting of hyper-parameters, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e - 8$. The initial learning rate was $1e - 4$, and decayed exponentially

[2] The details of CARRADA-RAC are in the Appendix, and the code is available at https://github.com/zlw9161/CARRADA-RAC.

with the rate of 0.9 every 20 epochs. The training epochs and mini-batch size were 300 and 6, respectively.

Evaluation metrics: All the models are evaluated with cell-wise precision/recall, Intersection over Union (IoU), and Dice score during training/validation phase. And the mean IoU (mIoU) and mean Dice (mDice) over categories are used as the principle metrics for model performance comparison. All the reported results were obtained from the test subset.

4.2. Exploration of Guard Bandwidth Setting

Different sizes of guard bandwidth would make different effects on the peak response capture. Therefore, we evaluate the RSS performance of our proposed methods under several guard bandwidth ($\mathbf{b}_G$ in Eq. 1) settings for three input views, *i.e.*, RD, AD and RA, as shown in Tab. 1.

Frameworks	B_G	**#Params**	**RD View**		**RA View**	
			mIoU	**mDice**	**mIoU**	**mDice**
PKCOn	{0, 0, 0}	4.5M	57.2%	68.9%	36.9%	44.7%
	{1, 1, 1}	5.7M	58.8%	**70.7%**	39.0%	47.7%
	{2, 2, 2}	6.9M	58.8%	70.5%	39.8%	48.5%
PKCOn*	{0, 0, 0}	4.5M	58.2%	69.7%	35.4%	42.5%
	{1, 1, 1}	5.7M	58.2%	70.1%	37.9%	46.3%
	{2, 2, 2}	6.9M	**59.1%**	70.3%	**40.2%**	**49.7%**
PKCIn	{0, 0, 0}	5.5M	58.7%	70.6%	40.4%	49.7%
	{1, 1, 1}	6.3M	60.0%	71.9%	42.5%	52.9%
	{2, 2, 2}	7.1M	60.3%	72.3%	42.7%	53.4%
PKCIn*	{0, 0, 0}	5.5M	59.2%	71.0%	41.1%	50.9%
	{1, 1, 1}	6.3M	60.7%	72.5%	42.9%	53.3%
	{2, 2, 2}	7.1M	**61.1%**	**72.9%**	**43.3%**	**53.5%**

Table 1. The effectiveness of guard bandwidths. Please note that $\mathbf{B}_G = \{\mathbf{b}_G^{\text{RD}}, \mathbf{b}_G^{\text{AD}}, \mathbf{b}_G^{\text{RA}}\}$, where $\mathbf{b}_G^x = \mathbf{b}_G^y$ by default. Bold means the best results and underline is the second best one.

As responses closer to the center unit are more likely to contain energy leaked from target, it is reasonable to shield them to estimate a more accurate interference distribution from reference units outside of the guard bands. Therefore, the performance of PeakConv-based networks with guard band mechanism, $\mathbf{B}_G = \{1, 1, 1\}$ and $\mathbf{B}_G = \{2, 2, 2\}$, are better than without, *i.e.*, $\mathbf{B}_G = \{0, 0, 0\}$. Comparing with $\mathbf{B}_G = \{1, 1, 1\}$, PeakConv with $\mathbf{B}_G = \{2, 2, 2\}$ owns a larger PRF. For the center unit with object information, larger PRF enables the interference estimation conducted via densely sampling reference units in the field further away from the peak response, thus obtains better results in both PKCOn-Net and PKCIn-Net frameworks. However, setting $\mathbf{B}_G = \{1, 1, 1\}$ is benefit to the trade-off between RSS performance and model complexity.

4.3. Exploration of Convolution Mechanism

To further explore the role of PeakConv in RSS performance improving, we take the place of it in PKCOn-Net and PKCIn-Net with regular 2D convolution (Conv), dilated convolution (DilConv) [28] and deformable convolutions (DefConv and DefConvV2) [8, 29], respectively. Experimental results shown in Tab. 2 not only reflect the performance diversity caused by different convolution mechanism, including vanilla-PKC, ReDA-PKC and other convolutions mentioned above, but also take the influence of their RFs into account. Hence convolutions with different kernel sizes, *i.e.*, 3×3 and 5×5 are investigated to compare with our PeakConvs with 16 kernel weights.

Results posted in Tab. 2 present the superiority of proposed PeakConvs among all the convolutions in the control group. Existing convolution operations attempt to extract target features from RF directly, so the design of RF is one of the key affecting model performance. Compared with regular Conv, DefConvs learn their RFs in an object-shape-driven way, which can be regarded as object feature augmentation, thus obtaining better results. However, unlike the optical images, the radar object signatures do not own clear structures or specific shapes, the advantage in object representation of DefConvs cannot be fully exerted, resulting in limited performance improvement. DilConv enlarges RF without complicating the computation by inserting dilation into convolution kernels. This sparse sampling manner indirectly forms a guard field around the center of its RF, which approximately introduces the guard band mechanism into the convolutions leading to more satisfied RSS performance. Nevertheless, with explicitly estimating and eliminating interference information from the object-interference mixed features or signals, the proposed PeakConvs yield the most encouraging RSS performance.

4.4. Comparison with State-of-The-Art (SoTA)

Our proposed RSS networks are further compared with popular visual segmentation models, including FCN [16], U-Net [24] and DeepLabv3+ [5], and SoTA RSS models, *i.e.*, RSS-Net [13], RAMP-CNN [10], MVA-Net and TMVA-Net [18]. Results in Tab. 3 show that our PeakConv-based models own the best performance. Specifically, thanks to taking the center unit into account during the estimation of interference, ReDA PKC extracts peak response in a more efficient way and obtains better performance than vanilla-PKC. As for the comparison of two PeakConv-based frameworks, PKCIn* with multi-frame input additionally providing temporal information achieves the best segmentation scores, which is significant on RA view. However, PKCOn-Net using single frame input surpasses all the existing RSS models which need to dig effective information in multi-frame (≥ 3) input. In addition, the small amount of parameters further highlights the learning efficiency of our PeakConv, which could keep a good balance of RSS effects and model complexity. Related details are illustrated in Fig. 7. All the points mentioned above validate the rationality and effectiveness of PeakConv, which learns object signature from PRF instead of adjacent field

Conv Type		Conv				DefConv				DefConvV2				DilConv				PeakConv		PeakConv*	
Kernel Size		3 × 3		5 × 5		3 × 3		5 × 5		3 × 3		5 × 5		3 × 3		5 × 5		16		16	
Frameworks		SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF
#Params		4.7M	5.6M	7.1M	7.2M	4.9M	5.7M	8.1M	8.2M	4.9M	5.8M	8.5M	8.6M	4.7M	5.6M	7.1M	7.2M	5.7M	6.3M	5.7M	6.3M
RD View	mIoU	54.0%	56.1%	55.6%	57.4%	55.5%	58.0%	55.8%	58.3%	55.4%	58.8%	56.1%	59.1%	57.1%	58.4%	57.9%	59.9%	58.8%	60.0%	**59.4%**	**60.7%**
	mDice	65.3%	68.0%	67.1%	69.2%	67.3%	69.8%	68.0%	70.2%	67.0%	70.6%	68.2%	70.8%	69.1%	70.4%	69.8%	71.9%	70.7%	71.9%	**71.2%**	**72.5%**
RA View	mIoU	36.4%	37.7%	36.4%	37.7%	38.2%	39.1%	38.4%	39.2%	38.3%	39.3%	38.6%	39.3%	37.4%	39.1%	37.7%	39.7%	**39.0%**	42.5%	38.6%	**42.9%**
	mDice	43.9%	46.2%	44.0%	46.4%	47.2%	48.1%	47.6%	48.2%	47.3%	48.6%	**47.8%**	48.6%	45.6%	48.1%	46.2%	49.3%	47.7%	52.9%	47.3%	**53.3%**

Table 2. Exploration of various convolutions. SF denotes network with single-frame input, which has the same structure with PKCOn-Net; MF denotes network with multi-frames input, which has the same structure with PKCIn-Net. Dilation step = 2 for all DilConvs.

defined for optical data processing and is more consistent with the characteristics of radar signals.

Frameworks	#Params	RD View		RA View	
	@Frames	mIoU	mDice	mIoU	mDice
FCN	134.3M@3	54.7%	66.3%	34.5%	40.9%
U-Net	17.3M@3	55.4%	68.0%	32.8%	38.2%
DeepLabv3+	59.3M@3	50.8%	61.6%	32.7%	38.3%
RSS-Net	10.1M@3	32.1%	36.9%	32.1%	37.8%
MVA-Net	4.8M@3	53.5%	65.3%	37.1%	44.8%
RAMP-CNN	106.4M@9	56.6%	68.5%	27.9%	30.5%
TMVA-Net	5.6M@5	56.1%	68.0%	37.7%	46.2%
PKCOn	5.7M@1	58.8%	70.7%	39.0%	47.7%
PKCOn*	5.7M@1	59.4%	71.2%	38.6%	47.3%
PKCIn	6.3M@5	60.0%	71.9%	42.5%	52.9%
PKCIn*	6.3M@5	**60.7%**	**72.5%**	**42.9%**	**53.3%**

Table 3. Comprehensive RSS performance comparison.

Model	Dataset	RD View		RA View	
		mIoU	mDice	mIoU	mDice
MVA-Net	CARRADA	53.5%	65.3%	37.1%	44.8%
	CARRADR-RAC	54.3%	66.1%	**43.2%**	**54.8%**
TMVA-Net	CARRADA	56.1%	68.0%	37.7%	46.2%
	CARRADA-RAC	59.7%	69.9%	**46.6%**	**57.9%**
PKCOn	CARRADA	58.8%	70.7%	39.0%	47.7%
	CARRADA-RAC	59.1%	71.0%	**46.8%**	**58.1%**
PKCOn*	CARRADA	59.4%	71.2%	38.6%	47.3%
	CARRADA-RAC	59.9%	71.8%	**46.5%**	**57.9%**
PKCIn	CARRADA	60.0%	71.9%	42.5%	52.9%
	CARRADA-RAC	60.3%	72.0%	**48.4%**	**60.1%**
PKCIn*	CARRADA	60.7%	72.5%	42.9%	53.3%
	CARRADA-RAC	61.0%	72.9%	**48.6%**	**60.3%**

Table 4. RSS performance on CARRADR-RAC.

4.5. RSS Performance on CARRADR-RAC

Through the annotation calibration mentioned in Sec. 4.1.2, we obtain a higher quality RSS dataset, CARRADR-RAC, and several models are trained and tested on this dataset. Results in Tab. 4 show that, with more reliable annotations, the RSS performance of all the mentioned models have been improved, especially on the RA view.

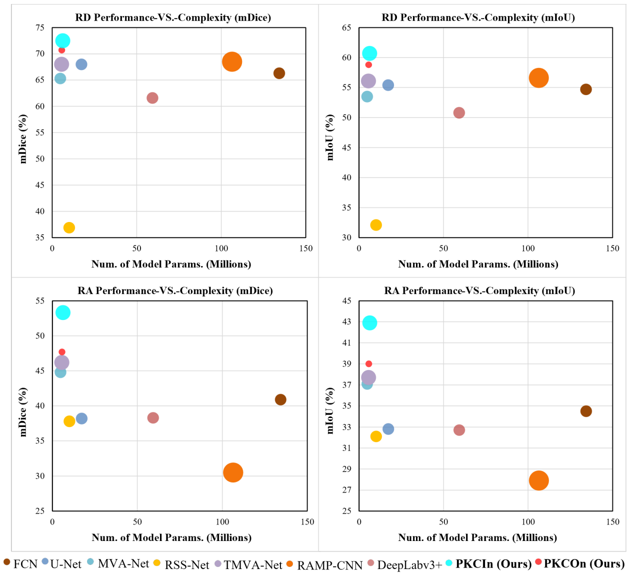


Figure 7. Performance *v.s.* Complexity.

5. Conclusion

In this paper, we propose a novel convolution operation, PeakConv, to highlight object signature from interference such as clutters/noises. PeakConv is designed specifically for radar data-related learning tasks. According to the role of center unit in interference estimation, there are two kinds of implements for PeakConv, vanilla- and ReDA-PKC. Comparing with existing RSS models, PeakConv-based networks achieve an outstanding trade-off between performance and complexity, in which PKCIn-Net achieves SoTA RSS performance and PKCOn-Net becomes sub-optimal one without additional temporal clues, *i.e.,* with single-frame input. It is obvious that the ability of PKCOn-Net to capture peak response would be further improved by introducing temporal information. Besides, in-depth optimization of PeakConv through auto-adaptive guard bandwidth is also one of our future research priorities.

References

[1] BABTLETT and S M. Smoothing periodograms from time-series with continuous spectra. *Nature*, 161(4096):686–687, 1948. 3

[2] Mario Bijelic, Tobias Gruber, and Werner Ritter. A benchmark for lidar sensors in fog: Is detection breaking down? In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 760–767, 2018. 1

[3] Daniel Brodeski, Igal Bilik, and Raja Giryes. Deep radar detector. In *2019 IEEE Radar Conference (RadarConf)*, pages 1–6, 2019. 1, 3

[4] Holger Caesar, Varun Kumar Reddy Bankiti, Alex Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 06 2020. 2

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 1, 2, 3, 6, 7

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 833–851, Cham, 2018. Springer International Publishing. 6

[7] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. 6

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 3, 7

[9] B. Dan, M. Gadd, P. Murcutt, P. Newman, and I. Posner. The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020. 2

[10] Xiangyu Gao, Guanbin Xing, Sumit Roy, and Hui Liu. Ramp-cnn: A novel neural network for enhanced automotive radar object recognition. *IEEE Sensors Journal*, 21(4):5119–5132, 2021. 1, 3, 6, 7

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, 2020. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015. 2, 3

[13] Prannay Kaul, Daniele de Martini, Matthew Gadd, and Paul Newman. Rss-net: Weakly-supervised multi-class semantic segmentation with fmcw radar. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 431–436, 2020. 1, 3, 5, 6, 7

[14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[15] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 3

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 7

[17] B. Major, D. Fontijne, A. Ansari, R. T. Sukhavasi, and S. Subramanian. Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019. 1

[18] Arthur Ouaknine, Alasdair Newson, Patrick Pérez, Florence Tupin, and Julien Rebut. Multi-view radar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15671–15680, October 2021. 1, 2, 3, 5, 6, 7

[19] Arthur Ouaknine, Alasdair Newson, Julien Rebut, Florence Tupin, and Patrick Pérez. Carrada dataset: Camera and automotive radar with range- angle- doppler annotations. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5068–5075, 2021. 1, 2, 6

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 1

[22] M. A. Richards. *Fundamentals of Radar Signal Processing, Second Edition*. Fundamentals of Radar Signal Processing, Second Edition, 2005. 1, 2

[23] Hermann Rohling. Radar cfar thresholding in clutter and multiple target situations. *IEEE Transactions on Aerospace and Electronic Systems*, AES-19(4):608–621, 1983. 1, 3

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 1, 3, 7

[25] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Los Alamitos, CA, USA, jun 2015. IEEE Computer Society. 2, 3

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

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

[28] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 1, 2, 3, 7

[29] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 3, 7