Appendix for "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

A. Appendices

Following parts are introduced in this appendices: i) the input streams for our network, *i.e.*, radar signal processing for multi-view range-angle-Doppler (RAD) tensors; ii) more detailed descriptions about the proposed PCKOn-Net; iii) detailed analysis and implementation of the proposed annotation calibration pipeline for CARRADA-RAC; iv) additional comparative analysis between existing convolutions and our PeakConv.

A.1. Radar Signals to Multi-View Representations

As mentioned in Sec. 1, most radar detection methods use FFTs as signal processing front-end. Despite the loss of fine-grained temporal information, FFT enables pure time domain radar echoes to be expressed in spatial (range and angle) and Doppler. At the same time, it can provide more intuitive and structured input to the model. Therefore, we also use such classic signal processing to aquire the multi-view RAD tensors.

Considering one-frame radar signals received from an FMCW radar, it is composed of multiple chirps from multiple antennas and can be denoted as $\{\mathbf{Chirp}_i^{(j)}\}_{i=1,j=1}^{N_D,N_A}$, where N_D and N_A denote the numbers of chirps and antennas, respectively. Then Range-FFT, $\mathcal{F}_R(\cdot)$, is conducted on each chirp, $\mathbf{Chirp}_i^{(j)}$, to obtain the DFT (Discrete FT) results, then we can get DFT tensor as follows:

$$\{\mathcal{F}_R(\mathbf{Chirp}_i^{(j)})\}_{i=1,j=1}^{N_D,N_A} = \{\mathbf{M}_R^{(j)}\}_{j=1}^{N_A} \in \mathbb{R}^{N_R \times N_D \times N_A}. \tag{1}$$

Where $\mathbf{M}_R^{(j)}$ denotes the Range-DFT matrix for j-th antenna chirps, N_R is the sampling number of each chirp. Please note that we only consider the real part of FFT results for simplification. Then for each row of $\mathbf{M}_R^{(j)}$, the Doppler-FFT, $\mathcal{F}_D(\cdot)$, is conducted to get the Doppler-DFT matrix as follows:

$$\{\mathcal{F}_D(\mathbf{M}_R^{(j)}[k,:])\}_{k=1}^{N_R} = \mathbf{M}_{RD}^{(j)} \in \mathbb{R}^{N_R \times N_D}. \tag{2}$$

Then we can group these DFT matrixes of all antennas to form the second DFT tensor, $\{\mathbf{M}_{RD}^{(j)}\}_{j=1}^{N_A} = \mathbf{M}_{RD}$. Finally, the Angle-FFT, $\mathcal{F}_A(\cdot)$, is performed on $\mathbf{M}_{RD}$ along the antenna dimension to get final RAD tensor:

$$\{\mathcal{F}_A(\mathbf{M}_{RD}[i,k,:])\}_{i=1,k=1}^{N_R,N_D} = \mathbf{M}_{RAD} \in \mathbb{R}^{N_R \times N_D \times N_A}. \tag{3}$$

In this work, $\{N_R,\ N_A,\ N_D\} = \{256,\ 256,\ 64\}$. Obviously, even a single frame of $\mathbf{M}_{RAD}$ is also quite dense and bulky for deep models. Further compression is needed to obtain affordable input streams in different frequency domains. Therefore a 2D-based multi-view compressing method is adopted in [3], *i.e.*, averaging over different frequency domains. Taking angle frequency domain as an example, 3D RAD tensor would be compressed as 2D RD view representation as follows:

$$\mathbf{X}_{RD}[i,k] = 10 * \log\left(\frac{1}{N_A}\sum_{j=1}^{N_A}|\mathbf{M}_{[i,j,k]}|^2\right). \tag{4}$$

Using such processing, the multi-view representations of our models, $\{\mathbf{X}_{RD},\ \mathbf{X}_{AD},\ \mathbf{X}_{RA}\}$ of one-frame radar tensor can be obtained. And the input scale is agressively reduced from $256\times256\times64$ to $256\times64+256\times64+256\times256$.

A.2. Additional Descriptions of PKCOn-Net

As we mentioned in Sec. 3.3, both PKCIn-Net and PKCOn-Net share the similar MIMO structure, where the encoding branches take multi-view radar tensor (*i.e.*, RD, AD, and RA) as input, and then the decoding branches make predictions on both RD and RA views. Furthermore, LSE is used to learn unified representation for the three encoding streams in a common latent space and the ASPP module aims to enhance the representation by injecting multi-scale spatial information extracted from each single view. Fig. 1 gives an intuitive illustration of the overall structure of our PKCOn-Net. Compared with the PKCIn model, its encoding branch for each single view is totally consisted of

*Equal contribution. †Corresponding author.

PeakConv layers, and only depends on single-frame input to make prediction. Such design makes sure that, the object signature characterization ability of our PeakConv can be fully verified in the ablation study.

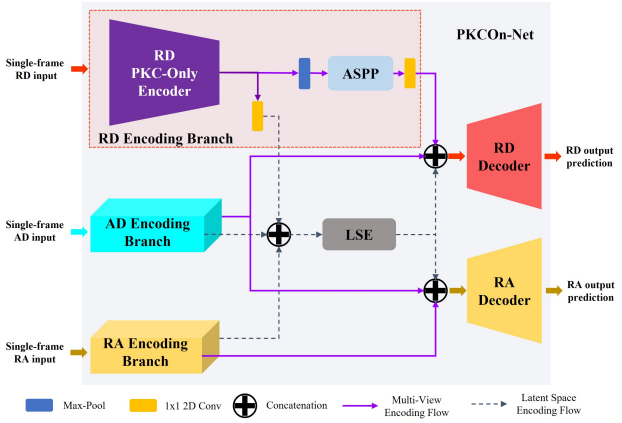


Figure 1. The overall framework of PKCOn-Net. Details of the encoding branch for RD-view is given as an example.

A.3. In-depth Analysis of Original CARRADA Annotation Generating Pipeline (AGP)

A.3.1 Original CARRADA AGP

Unlike the optical images which contain rich details for the observed objects, object signatures in radar signals are difficult to interpret, *i.e.*, it is hard to obtain the object semantics in the form of radar representation. Hence whether there is manual intervention or not, it is challenging to generate high-quality RSS annotations, *e.g.*, masks with category information. In order to automatically produce semantic labels on multi-view radar frequency representations, a visual perception integrated AGP was proposed for CARRADA [4].

On the basis of synchronized-collected radar-camera data streams, CARRADA AGP uses visual information of images to obtain physical prior knowledge, including rough positions, radial velocities and quite accurate categories of the objects. Then these information is fused with the filtered radar detection results based on CFAR [5] in the DoA-Doppler space, where the Mean-Shift [1] clustering-based temporal tracking is used to generate the semantic labels on the RAD tensor. As shown in Fig. 2-(a) the CARRADA AGP can be described as follows:

- **Camera images processing.** i) *Vision semantic segmentation*: applying pre-trained Mask-RCNN [2] to produce pixel region and category information of the object on images; ii) *Pixel position to real world coordinates*: transforming the obtained pixel-wise object positions to real world coordinates with the help of internal and external camera parameters; iii) *Velocity estimation*: using the changes of object position between two consecutive images to calculate the velocity vector, which is then converted to Doppler vector according to the complementary angle of radar relative angle. In this way, object's category and RD information could be estimated from camera images.

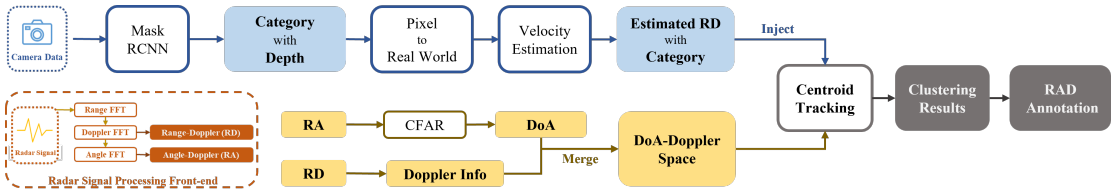
- **Radar data processing.** i) *Direction of Arrival (DoA) representation generation*: converting the polar representation of radar on RA view to Cartesian coordinate system; ii) *CFAR filtering*: filtering the obtained DoA representation through CFAR algorithm to get sparse candidate points of the moving objects; iii) *Doppler information compensation*: injecting Doppler information into the filtered DoA representation and then DoA-Doppler representation is presented.

- **Tracking-based annotation generation.** After respectively gathering object information from camera images and radar data, a tracking method based on Mean-Shift is used to fuse these information to mark final annotation: i) *Tracking centroid initialization*: the object information extracted from camera image is first projected to the DoA-Doppler representation, and then the projected DoA-Doppler point is taken as the initial centroid for tracking. ii) *Tracking association*: starting from the initial centroid, Mean-Shift clustering is conducted on the DoA-Doppler point cloud, where the clustering results are measured by JS divergence. In the following steps of object tracking and association, new centroid for the adjacent frames (front and back) would be initialized by the optimized clustering results associated with RD points extracted from the camera image. iii) *RD and RA annotation generation*: the final clustering results are respectively projected to RD and RA maps, in which the resolution of radar sensor would be taken into consideration. Finally, multi-view RSS annotations are presented.

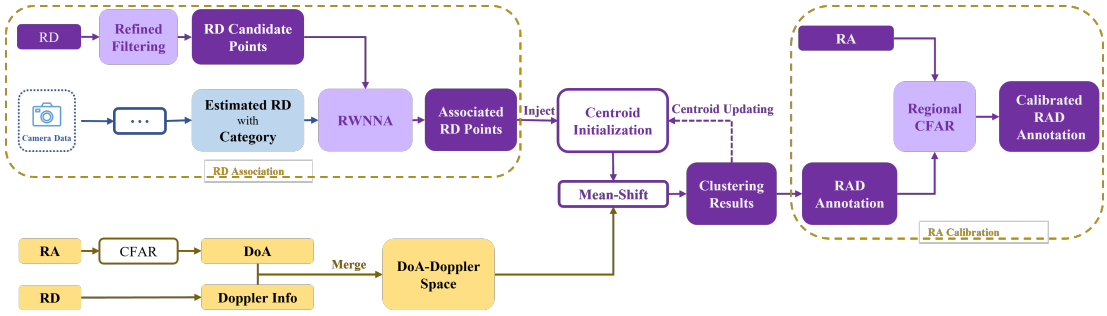
A.3.2 The Defects of CARRADA AGP

CARRADA AGP can provide relatively accurate RD annotations on semantic level. However, the annotations on RA view are less satisfactory, as shown in Fig. 3. Through in-depth analysis of CARRADA AGP, we found two reasons for its inaccurate RA annotations:

- **Low quality of cluster centroid initialization.** Mean-Shift-based tracking association is the key for generating annotations, and its clustering effects is highly sensitive to the initial centroid selection. Better centroid initialization will help the algorithm converge to a more accurate target region. However, original



(a) CARRADA-AGP



(b) CARRADA-RAC-AGP

Figure 2. The illustrations of annotation generating pipeline for CARRADA [4] and our improved pipeline for CARRADA-RAC.

AGP relies heavily on the results estimated from camera images, *i.e.*, the RD-category information obtained in image is considered entirely credible and directly projected to the DoA-Doppler space. But in fact, the vision position estimation often has a large deviation from the actual situation as there is no depth information in images, which results in the low quality of initial centroid. On the other hand, the DoA results used to conduct clustering are filtered by CFAR with high false alarm rate, which makes converging to the real object position rather difficult. Consequently, further correction of initial centroid is necessary, so as to provide a better starting point for the tracking association algorithm.

- **Low angle resolution of low-cost FMCW radar.** It can be observed from Fig. 3 that object signature has serious tailing in the angle dimension, which would deteriorate the angle estimation. On the one hand, this phenomenon is caused by the low angle resolution of the low-cost radar; on the other hand, it is also because that the acquisition of RA map requires multiple FFT operations, which may cause the loss of original radar signals accumulate in angle domain. Therefore, more refined processing methods are needed to better estimate the object location on RA maps, so as to obtain more compact DoA-Doppler candidate points, and thus providing the tracking association algorithm with a search space closer to the real object location.

A.4. Calibration of CARRADA-RAC

To cope with the issues analyzed above, the following two improvements have been made to the original CARRADA AGP: i) refined RD association strategy, which conducts on the estimated results of vision and radar to obtain better initial centroids; ii) regionalized CFAR, which could mitigate the inaccurate filtering of object candidates caused by serious distortion of object signature in RA representation. The improved pipeline for CARRADA-RAC is illustrated in Fig. 2-(b), and the details are in the following parts of this section.

A.4.1 RD Association for Centroid Initialization

In order to obtain better initialized centroid, a response weighted nearest neighbor association (RWNNA) method is proposed. To better perform RWNNA, a series of filtering steps including CFAR with low false alarm rate, zero frequency elimination (suppressing the clutters with zero Doppler) and threshold filtering (top-k magnitude as the threshold) are conducted on the original RD map to get more compact and accurate object candidate points. Then RWNNA will associate these processed RD point clouds with the RD-Category results obtained by the vision estimation. That is, for each vision candidate RD point, RWNNA will find its nearest neighbor from the candidates RD points filtered from the radar data, and assign its category to this neighbor point.

However, due to the serious deviation of vision depth estimation, more radar information should be taken into account in finding the nearest neighbor. Therefore, the ampli-

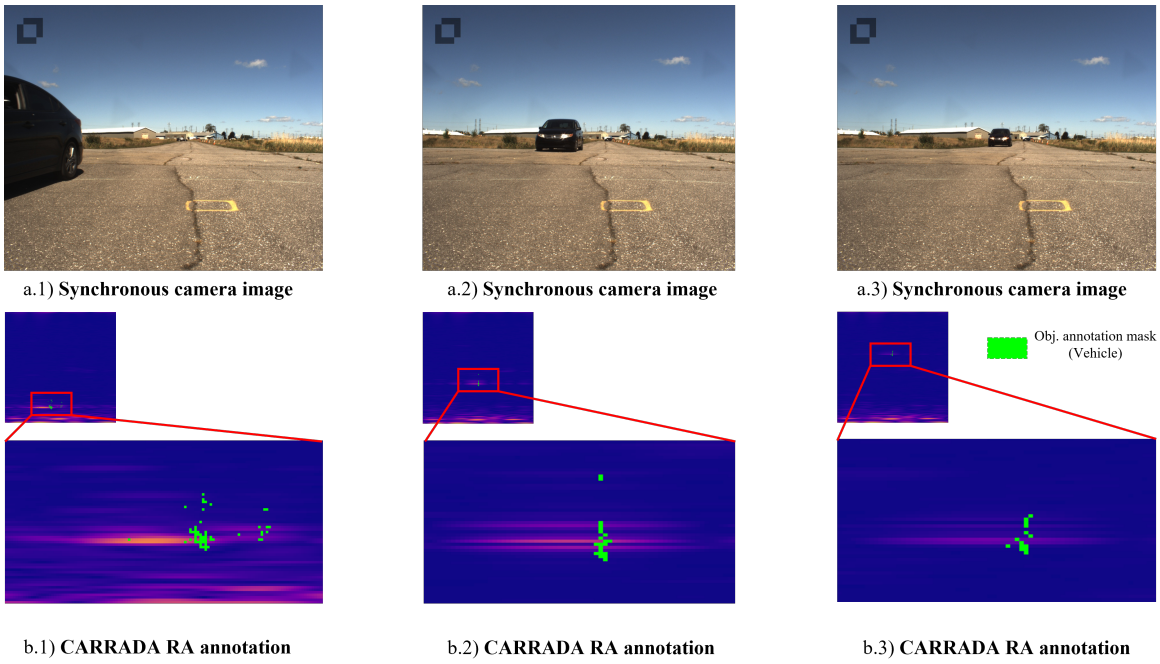


Figure 3. The illustration of RA-view annotations in CARRADA.

tude intensity for each RD point filtered from radar data is considered as the distance definition for RWNNA, i.e., the response weighted score. Given some candidate RD point $\mathbf{p}_j = \{p_j^R,\ p_j^D\}$ from vision estimation and its synchronous filtered RD point cloud $\Omega = \{\mathbf{r}_i = \{r_i^R,\ r_i^D\}\}_{i=1}^N$, the response weighted score, S can be defined as:

$$S_{i,j} = \frac{A_i}{\|\mathbf{r}_i - \mathbf{p}_j\|_2}. \tag{5}$$

Where $S_{i,j}$ denotes the response weighted score between $\mathbf{p}_j$ and $\mathbf{r}_i \in \Omega$, and A_i is the amplitude value of $\mathbf{r}_i$. The goal of RWNNA is to find $\mathbf{r}_k = \arg\max_{\mathbf{r}_i \in \Omega} S_{i,j}$ for $\mathbf{p}_j$, and assign the category information of $\mathbf{p}_j$ to $\mathbf{r}_k$. Then the following tracking method will use $\mathbf{r}_k$ with category information for centroid initialization instead of using $\mathbf{p}_j$ like CARRADA AGP. It is clear to see that, compared with the original pipeline where centroid initialization is solely depend on the vision estimation, high-quality centroid with more correct RD coordinate can be obtained by our proposed RD association strategy.

A.4.2 Regional CFAR for RA Calibration

For low-cost FMCW radar, the resolution in angle dimension is usually much worse than other dimensions, *e.g.*, range and Doppler. Going back to the signal processing front-end of AGP, as shown in Fig. 2-(a), the frequency response in angle domain is obtained from the last FFT operation, *i.e.*, angle domain would suffer the most serious information loss. Therefore, the processing of angle domain information requires extra care, which inspires us to come up with the regional CFAR, a simple but effective CFAR algorithm specific to RA view. The basic idea of this algorithm is to perform CFAR under the constraint of more reliable range information, while considering the tailing effect of target signature in angle domain. The details of the regional CFAR is shown in Algo. 1. As illustrated in Fig. 4, our RA annotations are obviously more consistent with the characteristics of object signatures in RA maps.

A.5. Additional Contrastive Analysis of PeakConv and Others

To more intuitively show the performance of various types of convolutions on RSS tasks, we further plot Fig. 5. Four conclusions can be obtained from the results:

- i. **Larger receptive field leads to better performance.** For the same convolution type, 5×5 kernel usually obtains better RSS performance than 3×3. We argue that the improvement may not only because more parameters are introduced during network learning, but also because the increased probability that interference and object information are simultaneously sampled.
- ii. **Band-pass filtering mechanism is more suitable for radar data processing.** Among all the convolutions, DilConv and ours always achieve better perfor-

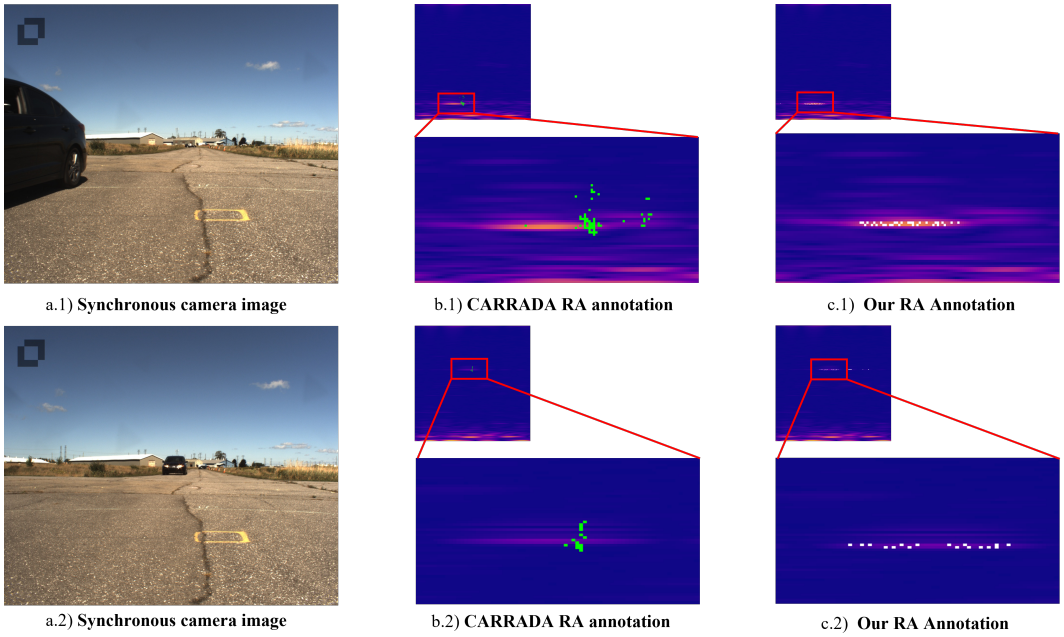
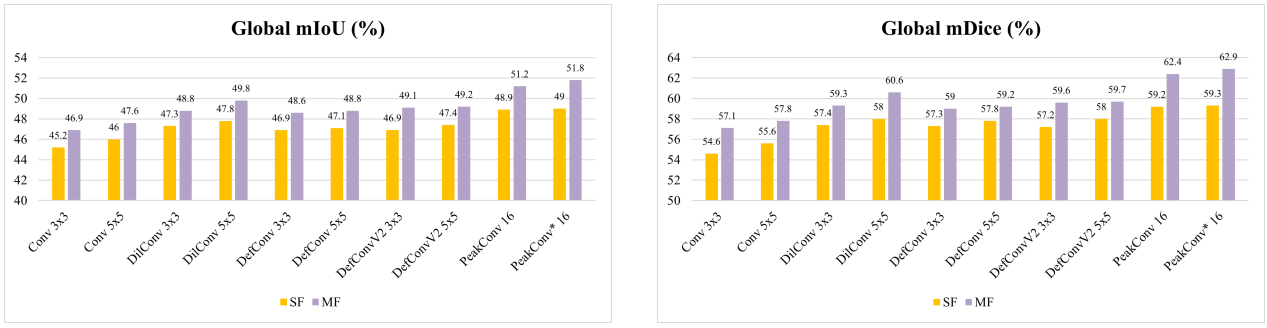


Figure 4. The Comparison between RA-view annotations of CARRADA and CARRDA-RAC.



(a) Global mIoU comparison

(b) Global mDice comparison

Figure 5. The global performance (the average performance of RD and RA views) comparison of our PeakConvs and other existing convolutions.

mance. This encourages us to rethink the difference between these two convolutions and others, *i.e.*, both of them can achieve band-pass filtering with the help of dilation and guard band mechanism, respectively.

- iii. **Interference (clutter/noise) energy/power estimation is important for radar data processing.** Although the dilation of DilConv can realize guard band effect similar to our PeakConvs to some extend, DilConv fails to exceed our PeakConvs in RSS performance without explicitly considering the interference estimation.

- iv. **Temporal information can effectively improve model capability.** The MF models always achieve better performance than their SF counterparts. It clearly indicates that, temporal information should not be ignored for learning tasks related to radar data. Therefore, PeakConvs with temporal encoding capability is worthy of in-depth research in our future work.

References

[1] 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. 2

[2] 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. 2

[3] Arthur Ouaknine, Alasdair Newson, Patrick Pérez, Florence Tupin, and Julien Rebut. Multi-view radar semantic segmentation. In *Proceedings of the IEEE/CVF International Confer-*

Algorithm 1: Regional CFAR

Input: Point set of input RA map: Ω_{RA};
RA point set filtered by Mean-Shift tracking: Ω_{DoA}.

Initialization:
Filtered RA point set: $\Omega^{\star}_{RA} = \emptyset$;
Size of guard unit: N_G;
Size of reference unit: N_R;
CFAR energy: $A^{\star} = 0$;
Detection threshold: λ.
for $\mathbf{p} = \{p_R,\ p_A\}$ *in* Ω_{RA} **do**
  if $\mathbf{p} \in \Omega_{DoA}$ **then**
    $G^{p}_{R} \leftarrow$ Getting reference field centered on $\mathbf{p}$ according to N_G and N_R;
    $\Omega^{\star} \leftarrow$ Getting reference points in G^{p}_{R};
    $N^{\star} \leftarrow$ Getting the number of points in $\Omega^{\star}$;
    for $\mathbf{p}^{\star}$ *in* $\Omega^{\star}$ **do**
      $a^{\star} \leftarrow$ Getting amplitude value of $\mathbf{p}^{\star}$;
      $A^{\star} += a^{\star}$;
    end
    $M^{\star} = A^{\star}/N^{\star}$;
    if $M^{\star} > \lambda$ **then**
      $\Omega^{\star}_{RA} = \Omega^{\star}_{RA} \cup \mathbf{p}^{\star}$;
    end
    $A^{\star} = 0$;
  end
end

Output: Return filtered RA point set: $\Omega^{\star}_{RA}$.

ence on Computer Vision (ICCV), pages 15671–15680, October 2021. 1

[4] 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. 2, 3

[5] Hermann Rohling. Radar cfar thresholding in clutter and multiple target situations. *IEEE Transactions on Aerospace and Electronic Systems*, AES-19(4):608–621, 1983. 2