Multi-Modal Cross Learning for Improved People Counting using Short-Range FMCW Radar

Conference Paper · April 2020

DOI: 10.1109/RADAR42522.2020.9114871

4 authors, including:

Souvik Hazra
Infineon Technologies
14 PUBLICATIONS **153** CITATIONS
SEE PROFILE

Avik Santra
Infineon Technologies
71 PUBLICATIONS **484** CITATIONS
SEE PROFILE

Multi-Modal Cross Learning for Improved People Counting using Short-Range FMCW Radar

(Invited Paper)

Cem Yusuf Aydogdu, Souvik Hazra, Avik Santra
Infineon Technologies AG
Am Campeon 1-12, 85579 Neubiberg
E-mail: avik.santra@infineon.com

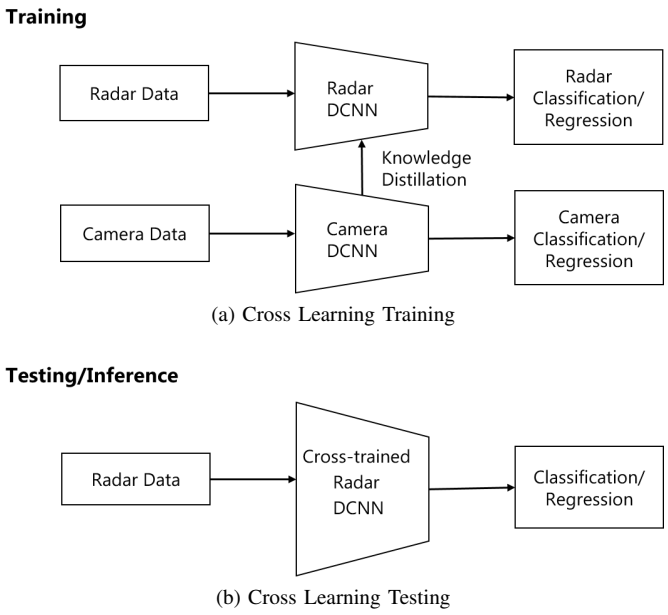
Robert Weigel
Institute for Electronics Engineering,
University of Erlangen-Nurnberg
Email: robert.weigel@fau.de

***Abstract*—Radar systems enable remote-less sensing of multiple persons in its field of view. In this paper, we propose a novel people counting system using 60-GHz frequency modulated continuous wave radar sensor. The proposed deep convolutional neural network learns from supervised radar data and also through knowledge distillation via multi-modal cross-learning of representation from a synchronized camera-based deep convolutional neural network. To overcome several shortcomings of the radar data, novel multi-modal cross learning algorithm is proposed that leverage the high-level abstractions learnt from camera modality. We also propose novel focal-regularized loss function to facilitate improved feature learning. We demonstrate the superior performance of our proposed solution in counting upto 4 people and detection of more than 4 people in indoor environment in comparison to the state-of-art radar-based unimodal learning.**

I. Introduction

People counting has been a hot topic for research in industry and academia due its wide range of use-cases. In indoors, it can be used for regulating energy consumption in an automatic manner by using the information for smart control of lighting, heating and air conditioning (HVAC) systems [1]. In scenarios that involve public gatherings such as malls, cinemas or transportation system, people counting solutions can be used to generate statistic about foot-fall which can give the administrations an edge for better understanding of conversion rates and service planning. People counting solutions also can be used as a warning or alarming system in places like elevator or any area with limitation of maximum number of people at a given time.

There have been multiple proposals of people counting solutions using sensors like cameras, infrared and radar sensors. With the use of deep learning, camera based solutions tend to be accurate and reliable even in cases with highly dense crowd or clutter environment [2][3][4][5][6][7]. However, any form of people surveillance with cameras raise privacy concerns and also underperform in indoor scenarios with limited illumination. Even though radar-based solutions [8][9][10][11][12][13] are immune to such limitations but suffer from issues such as missed detection caused due to occlusion, low resolution data which lowers the chances of target identification and time-varying radar signal strengths caused due to superposition of reflections coming from different body parts. These challenges makes the use of short-range and low-cost radars using conventional approaches for people counting solutions in dense scenarios ineffective. In this paper, a novel multi-modal cross learning approach is proposed which involves distilling high level feature extraction learned by a camera based deep convolutional neural network (DCNN) model to a radar based DCNN in such a way that the radar DCNN can perform people counting more accurately and reliably over a unimodal radar DCNN.



(a) Cross Learning Training

(b) Cross Learning Testing

Fig. 1. (a) Training methodology for multi-modal cross learning through knowledge distillation from superior sensor modality's network, and (b) Testing or Inference setup using cross-learned neural network parameters.

A learning system that involves a single modality data stream (e.g. radar) for the task of classification or regression is known as unimodal learning. Multi-modal learning on the other hand involves use of multiple sensor data streams (e.g. camera & radar) for learning to perform a common task of classification or regression and likewise the same sensor fusion setup for testing or inferring in real-time. In multi-modal cross learning, the networks for different data streams are trained independently or together for a common task and through

well-defined loss function such that the inferior network can perform as good as the superior network independently by distilling high level feature extraction and representation from the superior network. The superiority of a network may be due to availability of large dataset, high resolution and high quality attributes or just for the use of a deeper or complex network. Fig. 1 presents the high-level concept of the proposed multi-modal cross learning framework where radar DCNN learns not only from radar data but also supervised representation from camera DCNN trained on similar task.

Multiple works portraying the use of multi modal learning for merging information from multi modality through slow-fusion in initial layers or late-fusion in last layers or through concatenation of features extracted independently [14][15]. Recently, RFPose [16] was proposed wherein the authors propose a camera DCNN to act as a teacher network providing supervision to WiFi based-DCNN to learn artifacts and features that can help WiFi-based DCNN to predict human poses during inference - a task seemingly difficult for WiFi DCNN alone to achieve. In [17], a correlation-based similarity metric have been proposed to have a superior network for the task of gesture recognition by learning features from one modality to another. In this paper, we propose a novel multi-modal cross-learning framework for people counting application using frequency modulated continuous wave (FMCW) radar, which is trained not only from supervised radar data but also learns its' parameters through high-level features distilled from camera-based DCNN. We demonstrate the people counting performance of our proposed solution with upto 4 people counting and detection of more than 4 people in indoor environment, which surpasses the performance achievable through its counterpart radar DCNN exploiting supervised radar-alone data.

II. FMCW Radar System Design

The FMCW radar chipset *BGT60TR13C* from *Infineon Technologies AG* in Fig. 2 has been used for the proposed solution. The most commonly used mode for FMCW radars are to use sequence of frequency chirps with short ramp-times, delays between chirps to save power and at the end of the sequence of chirps for data acquisition and processing[8]. The frequency chirps with bandwidth of 1.0 GHz within the 60-GHz band and pulse repetition time of $400\,\mu s$ was used.

The interval between two starting times of chirp sequences is called frame time and is set to 100 ms. Number of ADC samples per chirp NTS $=$ 128 and PN $=$ 64 number of chirps per frame was used. The number of samples per chirp defines the maximum unambiguous range for the range FFT and depends on the chirp time as well as the sampling rate of the analog-to-digital converter (ADC). The ADC to digitize the data has a resolution of 12 bit and the data was sent from the evaluation board via USB to the PC working as master.

The chipset *BGT60TR13C* is configured with the system parameters and derived parameters provided in Tab I.

Fig. 2. *Infineon BGT60TR13C* radar system

TABLE I
Operating Parameters.

Parameters, Symbol	Value
Ramp start frequency, f_{min}	60.5 GHz
Ramp stop frequency , f_{max}	61.5 GHz
Bandwidth, B	1 GHz
Range resolution, δr	15 cm
Number of samples per chirp, NTS	128
Maximum range, $R_{\max}$	9.6 m
Sampling frequency, fs	2 MHz
Chirp time, T_c	64 μs
Chirp repetation time, T_{PRT}	400 μs
maximum Doppler , $v_{\max}$	3.125 m/s
Number of chirps, PN	64
Doppler resolution, δv	0.0977 m/s
Number of Tx antennas, N_{Tx}	1
Number of Rx antennas, N_{Rx}	2
Elevation θ_{elev} per radar	90°
Azimuth θ_{azim} per radar	130°

III. Data Preparation & Processing

A. Range-Angle Image

The intermediate frequency signal from a chirp with NTS $=$ 128 number of samples and PN $=$ 64 consecutive chirps are collected and arranged in the form of a 2D matrix, as PN $\times$ NTS. As a first step, a range-Doppler image (RDI) is generated by subtracting the mean along fast time, followed by 1D Fast Fourier Transform (FFT) along fast time for all the PN chirps to obtain the range transformations. Following which mean across slow-time is subtracted followed by 1D FFT to obtain the Doppler transformation for all range bins. The RDI is then processed through moving target indicator (MTI) filter to removes reflections from any static targets, such as chairs and furnitures in the room. Once the RDI across both received channel $N_{\text{Rx}} = 2$ is computed, the range-angle image (RAI) is computed through digital beam-forming algorithm utilizing the derived weights from angle model as follows:

$$z_{\text{RAI}}(r,\theta) = \sum_{\nu=-\nu_{max}}^{\nu_{max}} \sum_{j=1}^{N_{\text{Rx}}} z^{j}_{\text{RDI}}(r,\nu) e^{-j\frac{2\pi d^{j}\sin(\theta)}{\lambda}} \quad \forall -\frac{\theta_{azim}}{2} < \theta < \frac{\theta_{azim}}{2} \tag{1}$$

where θ is the estimated angle sweeped across the field of view, i.e. $-\theta_{azim}/2 < \theta < \theta_{azim}/2$, where θ_{azim} is the half-power beam width, and z^{j}_{RDI} is the complex RDI from the j^{th} receive channel across N_{Rx} receive channels. The first summation in eq. (1) transforms the RDI across each virtual

channel into RDI across the angle space, and the second summation marginalizes across Doppler bins to generate the RAI.

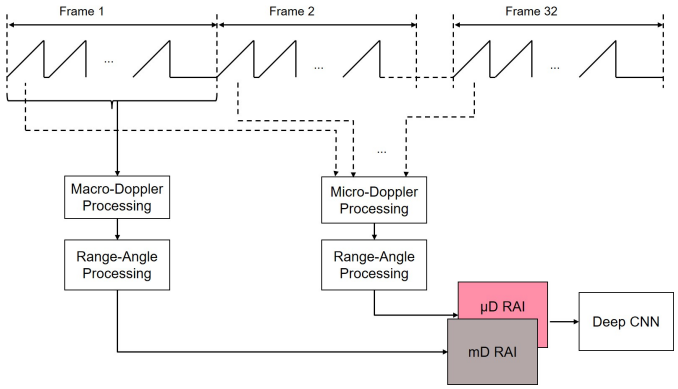
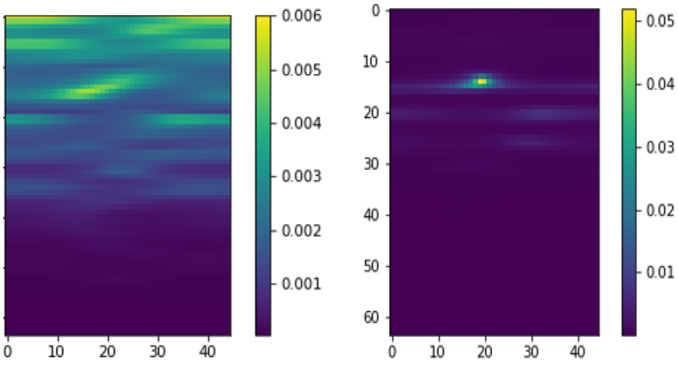


Fig. 3. Pre-processing pipeline to generate separate range-angle image using micro-Doppler and macro-Doppler components and is fed as separate channels to the deep neural net.



(a) macro-Doppler RAI (b) micro-Doppler RAI

Fig. 4. RAI generated from macro-Doppler components and micro-Doppler components for a static person.

Figure 3 presents the pre-processing step, whereby two separate range-angle images are generated using macro-Doppler (mD) processing and micro-Doppler (μD) processing by using chirps within the physical frame and extracting first chirp from 32 consecutive physical frames as explained in details in [1]. In case of static person, the human target exhibits generates micro-motion dynamics due to breathing or small body movements resulting in Doppler modulations on the returned signal [18]. Thus, mD RAI-channel capture human targets during movements and major motions, while μD RAI-channel capture human targets during static or quasi-static scenarios. Figure 4 presents the RAI generated from macro-Doppler and micro-Doppler processing pipeline. As can be seen, in case of static person the RAI generated from the micro-Doppler processing has a clear peak while that from the macro-Doppler processing doesn't, and alternately for moving humans. The two RAIs are fed as separate channels into the deep neural net.

B. Camera Processing (CSR-Net)

In the paper [6], the authors propose a network that can perform robust people counting in highly dense scenarios while also generating density heatmaps. The network uses a modified *VGG-16* network for reducing computational power requirement which consists of only the first ten layers of the original *VGG-16* network with 3 pooling layers only. Since output size of the modified network is $1/8^{\text{th}}$ of the input image, a bi-linear interpolation with a factor of 8 is performed on it to match its size to that of the input. The scaled output is then passed through a dilated convolution network, which is a very critical aspect of the architecture design due to its ability to capture information over large receptive field by using small kernels

A 2-D dilated convolution can be defined as followed:

$$y(m,n) = \sum_{i=1}^{M}\sum_{j=1}^{N} x(m + r \times i, n + r \times j)w(i,j) \quad (2)$$

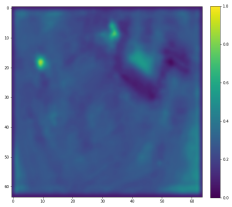
where r is the dilation rate and is a normal convolution if set to 1, $w(i,j)$ is the filter and $x(m,n)$ and $y(m,n)$ are the input and output of the dilated convolution respectively. The dilated convolution network consists of 6 dilated convolution layers that uses 3×3 kernels with dilation rate of 2 and 512 number of feature maps for the first three layers, 256 for the fourth, 128 and 64 for the last two layers respectively. The camera based density maps is generated by blurring each annotated head in an image with a Gaussian kernel where the ground truth $\delta(.)$ and target object x in it is convolved using a Gaussian kernel with a standard deviation equal to $0.3(\beta)$ times the average di of the distances from the three nearest neighbors. The kernel is defined as:

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i}(x);\ \sigma_i = \beta d_i \quad (3)$$

where the $\delta(.)$ is the ground truth and x the target object and is convoloved with a Gaussian kernel whose standard deviation is given by multiplication of $\beta = 0.3$ and the average distance d_i of three nearest neighbors.

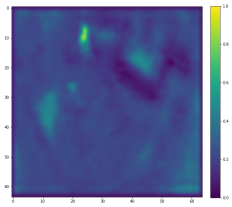
We employ 'transfer learning' by freezing all the weights, except for the last two layers and re-train the model with few supervised camera images from our setup. This process is performed to adapt the *CSR-Net* and ensure robustness of the network for our setup. The output of the adapted *CSR-Net* for a given camera image is a density map from which a people count can be done by counting pixels with a higher value than that of the defined threshold after clustering. The qualitative performance of the adapted *CSR-Net* is illustrated in Fig. 5 that depicts the density heatmap outputs of the adapted *CSR-Net* for camera image inputs for 2 and 3 person in the room.

(a) Camera Image



(b) adapted CSR-Net Heatmap

(c) Camera Image



(d) adapted CSR-Net Heatmap

Fig. 5. Camera Image of (a) 2 persons (c) 3 persons in field of view, corresponding heatmap from adapted CSR-Net in (b) and (d) respectively.

IV. Proposed Solution - Learning

In our proposed solution, a 2D CNN autoencoder is developed wherein the input is the radar RAI and the network is trained to be re-construct the density heatmap, which is generated by the adapted CSRNet from the corresponding synchronized camera image. During training, thus the idea is to learn feature embedding, which captures the representation and semantics derived from both the radar input image as well as the density maps from the camera-based DCNN. The aim of the proposed training framework is to improve the learning process by transferring or distilling the high-level knowledge abstraction (density heatmap) from the camera modality to the radar DCNN, and acts as an additional supervision, much like the teacher-student network. To achieve this knowledge distillation from high accurate network modality, we propose a novel reconstruction loss function, which is a combination of focal-regularized mean square error and cross entropy.

Figure 6(a) presents the proposed framework for training the radar RAI to learn feature embedding that is capable of reconstructing high quality density heatmap perfectly suitable for the task of people counting. Figure 6(b) presents the second step in training, the learned embedding from the autoencoder is fed into a fully-connected layer followed by softmax and only the FC-layer is re-trained through binary cross-entropy loss for people count class, while keeping the encoder weight's frozen. In this section, we outline the DCNN autoencoder architecture, the overall loss function and the final classification step for people counting through classification.

A. Convolutional Autoencoder

In our learning approach, we use a convolutional auto-encoder to be able to generate density heatmaps like the output of the adapted *CSR-Net* by feeding RAIs generated by the signal processing pipeline as input. The encoder consists of three convolutional layers with ReLu activation and a pooling layer after each convolution layer. Each of the convolution layer have 32 feature maps and a kernel size of 3×3. The decoder also consist of similar three convolution layers with upsampling layer after each of the convolution layer instead of a pooling layer. The factor for the pooling and upsampling is 2. Additionally, the decoder has a convolution layer with one feature map and kernel size of 1×1 with sigmoid activation as the last layer. Furthermore, add layers are used between first two pooling layers of both the encoder and the decoder to establish residual connections among the convolution layers. Since the supervised reconstruction image for the auto-encoder is high resolution heatmaps, the network learns feature representation based on high-level feature abstractions (heatmaps) from the camera modality.

B. Loss Function

We propose a loss function a custom loss function that combines mean square error (MSE), cross entropy (CCE) and data-dependent focal regularization. The MSE is a measure of the closeness of the input X to the output X_r in Euclidian distance domain. The MSE loss is defined as $l_{\text{MSE}}(X, X_r) = 1/N||X - X_r||^2$. The CCE calculates the number of bits of information preserved by reconstructing the density heatmap over the ground truth. The CCE can be defined as $l_{\text{CCE}}(X, X_r) = \sum_k X^k \log X_r^k + (1 - X^k) \log(1 - X_r^k)$.

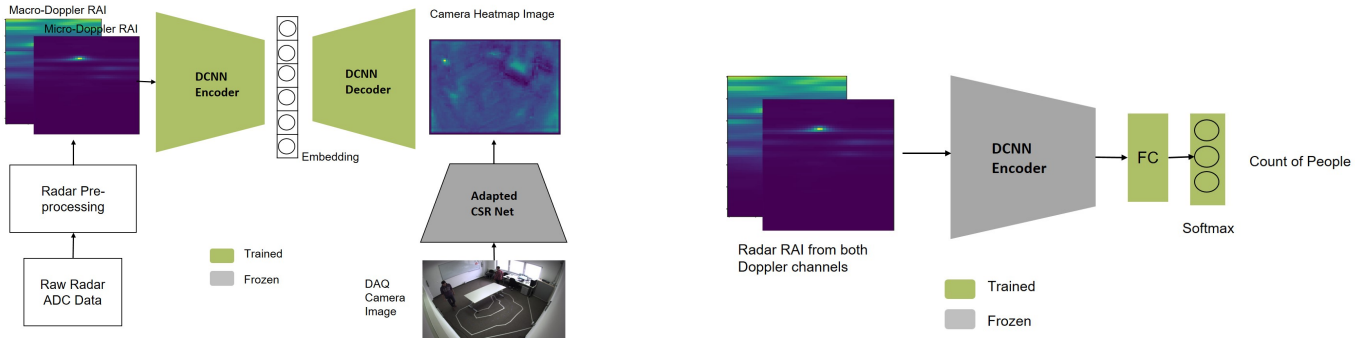
Most of the pixel values in the density heatmaps are close to zero and very few of them are close to one as very close values to one signifies a head of human target. Since this imbalance needs to be addressed for a robust learning, the MSE loss tackles it by performing a selective scaling of the cost function by thresholding the pixel values. In case of CCE, we use the power factor γ to steer the training focus on a selected class. We assign higher focal loss to pixels with a value more than that of an adaptive threshold computed by performing mean scaling per image.

$$l_{\text{MSE}} = \frac{1}{N_0} w^0 ||X^0 - X_r^0||^2 + \frac{1}{N_1} w^1 ||X^1 - X_r^1||^2 \quad (4)$$

$$l_{\text{CCE}} = \sum_{k \notin \Omega} (X^k)^{\gamma_0} \log(X_r^k) + \sum_{k \in \Omega} (1 - X^k)^{\gamma_1} \log(1 - X_r^k) \quad (5)$$

$$l_{\text{total}} = l_{\text{CCE}} + \lambda * l_{\text{MSE}} \quad (6)$$

where the selective weights for pixel values, less and greater than the adaptive mean value, are represented as w^0 and w^1 respectively. The γ_0 and γ_1 represents the power factors to steer the selective emphasize to the pixel values that are less and greater than the adaptive mean. The set for pixel values in X_r, which are more than the threshold $\eta = 0.4$ is represented by the set Ω. Since there are few cases where the pixel values cross the defined threshold, we set $w^1 > w^0$ and $\gamma_1 > \gamma_0$. The degree of contribution of MSE loss to the total loss function is controlled by the weighting hyperparameter λ whose value is computed through cross-validation.



(a) Training Stage 1

(b) Training Stage 2

Fig. 6. Proposed multi-modal cross-learning framework, (a) Stage 1 network is trained to learn camera CSR-Net heatmap, (b) Stage 2 network with fully connected layer learn to count number of people

C. People Count Classification

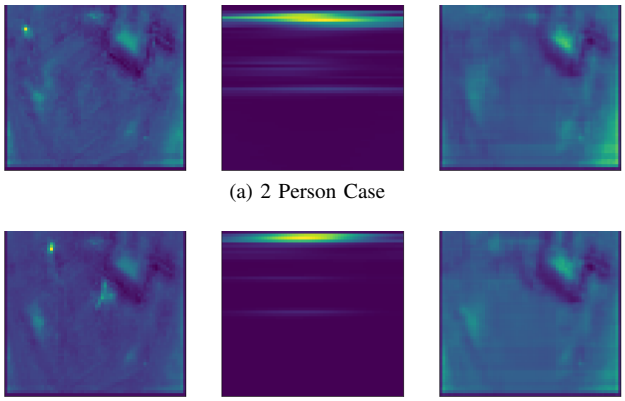
On successful training of the autoencoder to perform reconstruction of density heatmaps, the feature embeddings that are generated as outputs of the encoder for the training set is used to train a dense layer with five hidden units and softmax activation. Only the single dense layer is trained with categorical cross entropy loss and the count from the training set as targets to predict 1-4 person classes and an additional class of more than four person class. In the first step of training, since the encoder already learns to generate feature embeddings that captures the high-level abstractions and semantics that is capable of classifying people count, a fully-connected layer with softmax can readily predict the count the people. One can also train an end to end multi-task based CNN network that can generate both a density heat map and people count.

V. Experimental Results

The data was collected by syncing four cameras mounted in four corners of a room of size $8m \times 10m$ and four radars placed just below them. The frame per second was set to 10. The recording was performed with the help of 10 individuals who entered or left the room without any timely manner. However a constraint of maximum 7 number of people in the room at any given point of time was imposed. The per frame people count labelling was done in an automatic fashion by using the adapted CSR-Net. Additionally, a manual scrutiny was performed frame wise against the label.

The recording was done in 5 sets where each set was roughly 5 mins. After each set, different static objects such as tables, chairs, wall mounts etc. were added or removed. The total data set includes 59720 frames with an average of 5000 frames for each count class. A training and testing split of 80% and 20% on the total dataset was performed. Since the aim of the solution is to run on a low commodity hardware, the network architecture for the auto encoder model was found optimal based on its accuracy and model size. During inference, we just need the trained encoder block with the classification block which takes in input the raw radar data and predicts the count class. The entire inference block has a very small memory footprint of 44 kB, which is an added advantage arising from the proposed multi-modal cross learning. The use of only convolution and pooling layers instead of fully connected layers in the encoder block makes the architecture much more robust and fast.

Figure 7 presents the camera heatmap generated by the adapted *CSR-Net*, the input radar RAIs and the reconstructed heatmap at the output of the trained autoencoder for exemplary scenarios of 2 persons, and 3 persons. As can be observed from the reconstructed image, the radar autoencoder is capable of transforming the low-resolution radar RAIs into high-resolution heatmaps, which is starkingly close to the ones provided by the adapted *CSR-Net* from the camera input image. Thus, when the embedding representation learned by the trained encoder is used for classification of number of people in its field of view, the deep Net predicts quite accurate people counts.



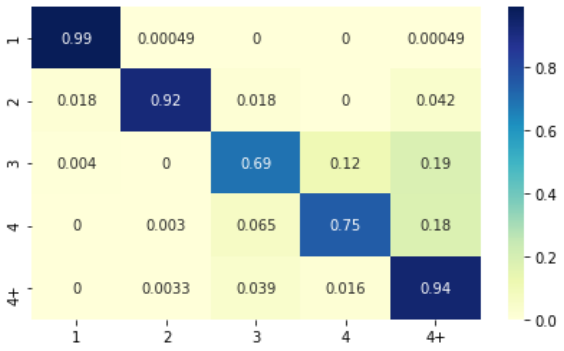
(a) 2 Person Case

(b) 3 Person Case

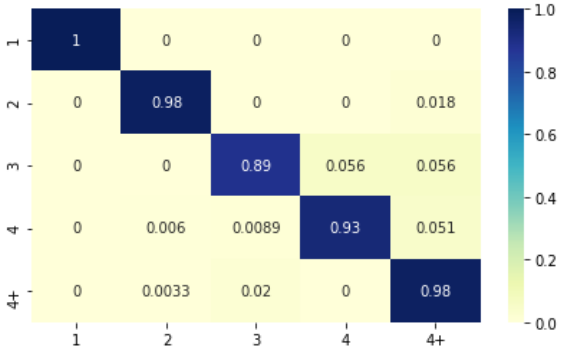
Fig. 7. Heatmap from CSR-Net, corresponding macro-Doppler RAI and reconstructed heatmap by proposed autoencoder for (a) 2 persons (b) 3 persons in the field of view.

Figure 8(a) presents the confusion matrix of count of people from 1 to 4 and detection of more than 4 person using uni-modal learning, wherein the autoencoder was trained with input from radar RAI and reconstruction also to the same radar RAI. The compact feature learned by the autoencoder in the next step is used for classifying the number of people. On the

contrary, Fig. 8(b) presents the confusion matrix of count of people using the proposed multi-modal cross learning through camera density heatmap. The same autoencoder architecture is used with different training modalities in both the cases for fair comparison. The unimodal approach is able to reach classification accuracy of 0.86, while the accuracy of the proposed solution is 0.955 on the test dataset, demonstrating the superior performance of the proposed solution.



(a) Uni-modal Learning



(b) Cross-Learning

Fig. 8. Confusion Matrix of People Counting using radar data with (a) Deep Convolutional Neural Network, whose parameters are learnt through radar-only data, (b) Deep Convolutional Neural Network whose parameters are learnt through camera-based cross learning with radar data

VI. Conclusion

In this paper, we present a novel framework for multi-modal cross learning and utilize it for developing a high accurate people counting system using 60-GHz frequency modulated continuous wave radar sensor. We presented a cross-learning solution, wherein camera processed data was used as an additional supervised information during training to improve the radar deep neural network's classification performance during training and inference. We demonstrate that the proposed solution is not only able to predict the number of people accurately but also is capable of generating high-resolution density heatmaps, which can considerably boost the detection and tracking performance of the low-cost, small form-factor radar systems. The performance of the proposed solution is demonstrated to be superior to its unimodal learning counterpart.

References

[1] A. Santra, R. V. Ulaganathan, and T. Finke, "Short-range millimetric-wave radar system for occupancy sensing application," *IEEE sensors letters*, vol. 2, no. 3, pp. 1–4, 2018.

[2] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1299–1302.

[3] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 2016, pp. 640–644.

[4] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.-M. Cheng, and G. Zheng, "Crowd counting with deep negative correlation learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5382–5390.

[5] Z. Zou, X. Su, X. Qu, and P. Zhou, "Da-net: Learning the fine-grained density distribution with deformation aggregation network," *IEEE Access*, vol. 6, pp. 60 745–60 756, 2018.

[6] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.

[7] C. Gao, P. Li, Y. Zhang, J. Liu, and L. Wang, "People counting based on head detection combining adaboost and cnn in crowded surveillance environment," *Neurocomputing*, vol. 208, pp. 108–116, 2016.

[8] C. Will, P. Vaishnav, A. Chakraborty, and A. Santra, "Human target detection, tracking, and classification using 24-ghz fmcw radar," *IEEE Sensors Journal*, vol. 19, no. 17, pp. 7283–7299, Sep. 2019.

[9] J. W. Choi, D. H. Yim, and S. H. Cho, "People counting based on an ir-uwb radar sensor," *IEEE Sensors Journal*, vol. 17, no. 17, pp. 5717–5727, 2017.

[10] J. He and A. Arora, "A regression-based radar-mote system for people counting," in *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2014, pp. 95–102.

[11] X. Yang, W. Yin, and L. Zhang, "People counting based on cnn using ir-uwb radar," in *2017 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, 2017, pp. 1–5.

[12] X. Yang, W. Yin, L. Li, and L. Zhang, "Dense people counting using ir-uwb radar with a hybrid feature extraction method," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 1, pp. 30–34, 2018.

[13] M. Stephan and A. Santra, "Radar-based human target detection using deep residual u-net for smart home applications," in *Proceedings of the 18th IEEE international conference on machine learning applications (ICMLA)*. IEEE, 2019.

[14] J. Wu, J. Cheng, C. Zhao, and H. Lu, "Fusing multi-modal features for gesture recognition," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 453–460.

[15] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2015.

[16] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7356–7365.

[17] M. Abavisani, H. R. V. Joze, and V. M. Patel, "Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1165–1174.

[18] V. C. Chen, *The micro-Doppler effect in radar*. Artech House, 2019.