Convolutional Neural Network Powered Identification of the Location and Orientation of Human Body via Human Form Point Cloud

Min Chen*†, Yang Miao‡, Yi Gong†, Xingpeng Mao*§
*School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China
†University Key Laboratory of Advanced Wireless Communications of Guangdong Province,
Southern University of Science and Technology, Shenzhen, China
‡Faculty of Electrical Engineering, University of Twente, the Netherlands
§Key Laboratory of Marine Environmental Monitoring and Information Processing,
Ministry of Industry and Information Technology, Harbin, China

***Abstract*—This paper proposes a Convolutional Neural Network (CNN) based scheme using the point cloud of human body to identify its location and posture. The point cloud is randomly generated but confined within a human form. The CNN-based model is fed with point cloud for predicting mass center location and orientation of the body with help of high end graphical processing units. We propose to project the point cloud in two vertical planes to exploit the image recognition capability of CNN. The proposed method is tested with a single person for three primary postures: standing, sitting and lying, to evaluate the prediction capability. Effects of the number of points indicating point cloud density and the distance between the observation station and the target are investigated. Simulation results show a body part dependent localization accuracy smaller than 8 cm, and posture dependent success rate above 93%, validating the functionality of proposed scheme.**

***Index Terms*—Convolutional Neural Network (CNN), passive localization, passive human posture recognition, radar reflection points, point cloud**

I. Introduction

Assisted-living has been one of the inevitable trends to ensure the well-being of the aging population, in particularly in the developed regions. In order to proactively predict and prevent the alarming situations for the elderly, it is important to localize the target in 3D space, to track and predict their postures, without requiring the target wearing devices. Wearable solutions integrated with the Internet of Things [1] are promising but require attaching probes on human body as well as the careful placement, which may not be the most comfortable solution at home settings. The infrared sensor and video camera [2] are strong candidates for device-free human healthcare monitoring. As the human body emits heat in the form of infrared rays, the amount of rays impinging on the sensor array changes with the temperature on the human body surface and the surrounding background. However, sensing systems employing optical techniques only work in the line-of-sight (LOS) scenarios and have privacy concerns which further limit the applicable scenarios.

Alternatively, radio-based device-free localization and posture recognition are promising through exploiting the interactions between radio waves and human body or background objects [3], [4]. The dispersive and directive Multiple-Input Multiple-Output (MIMO) radio propagation channel properties can be exploited for its information on the 3D direction of arrival (DOA), direction of departure (DOD), time of arrival (TOA) and so on. Estimation algorithms including the multiple signal classification (MUSIC) and fast Fourier transform (FFT) have been utilized to extract the key property parameters from measurable channels. To ensure the spatial-delay resolution, [5]–[7] have adopted a wide bandwidth of about 2 GHz and used multi-antenna arrays at link ends to explore the human posture estimation. In [8], the MUSIC-based direction estimation method takes tens of seconds to accurately capture the human activity information. In [9], the FFT-based method does not provide human position information because the distance between the antenna and the target cannot be estimated. It is still challenging to estimate human posture and location using spatial-delay domain MIMO radio properties in a realistic scenario, not to mention targeting multiple persons with multiple postures.

With the development of mmWave MIMO systems, new opportunities appear by connecting the "pencil" like beam space to the "computer" vision [4]. The very narrow beam space could be used to capture the interactive points on human body to form a point cloud [10]. As a matter of fact, since the accurate estimation of human posture is very important to judge and predict the behaviors of the elderly to achieve predictive monitoring, researchers have been exploring useful methods based on computer vision techniques by identifying specific body skeleton joints [11]–[14]. Now to think one step further, the "vision" can also be provided via this body point cloud which contains essentially the interactive points of radio waves with human body; hence it is the purpose of this paper. The core focus of this paper is NOT to capture the human body point cloud using mmWave MIMO radar or monostatic, multistatic stations; the goal is to link the *point cloud in the human form* with different density/sparsity conditions to its location and posture seen from observation station, by using machine learning - the Convolutional Neural Network (CNN).

CNN has been proposed in the early stage, and the rapid developments of graphical processing units (GPUs) and high-performance-computing (HPC) in recent years have brought

this model back to spotlight. There has been fantastic progress using CNN in a variety of fields, such as image classification [15], recognition [16], and semantic segmentation [17]. In this paper, we propose to use CNN to establish the association between the data representation of point cloud in human form and the mass center location and the orientation of target seen from observation station. The point cloud in human form is assumed to be capture by radar reflection points in 3D space, and we can adjust the cloud density/sparsity. We use an offline training procedure of the CNN, and realize the target localization and posture recognition in the end. This paper serves as our first step targeting at single person.

The remaining sections of this paper are organized as follows. Section II presents the 3D data model and scenario. Section III introduces the CNN-based identification algorithm. Section IV discusses numerical investigations and Section V concludes this paper.

II. Data Representation

The scenario assumed in this paper is a confined space with a volume where a pair of mmWave multi-antenna transmitter and receiver is placed at a fixed position (x_0, y_0, z_0), as is shown in Fig. 1. A single person presents in this space and the interactive points are assumed to be captured via analyzing the spatial-delay domain of the radio link. The interactive points in this paper are randomly generated by using von Mises Fisher (VMF) distribution in a human form, as is shown in Fig. 2. The VMF distribution [18] is a common distribution on sphere and is typically used for directional data analysis, e.g., the angle of arrival of multipath channel [19], [20]. The generalized VMF distributed points in a human form are further linked to the location of the transmitter/receiver (collectively called the observation station) to represent the AoA and AoD with delay information. These points in human form contain information of the positions, orientations, and body size.

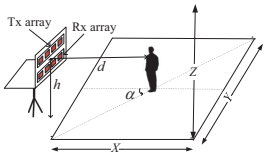

Fig. 1. Scenario

The specific human body figures used in this paper is shown in Table I. These VMF points are distributed on the surface of 6 different ellipsoids, each representing a part of the body, namely *head, torso, two legs, and two arms*. For each local ellipsoid, each point on it is represented by one azimuth and one elevation angles denoted by $\phi'_{\text{humanform}}$ and $\theta'_{\text{humanform}}$, respectively; their probability density function has the form [18]

$$f(\chi|\mu,\kappa) = \frac{e^{\kappa cos(\chi-\mu)}}{2\pi I_0(\kappa)}, \quad -\pi \le \chi < \pi, \\ \kappa > 0, \quad -\pi \le \mu < \pi, \tag{1}$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind with order 0, μ is a fixed point representing the mean direction, κ is a concentration parameter, and χ is a random point on sphere standing for $(\phi'_{\text{humanform}}, \theta'_{\text{humanform}})$. The locally generated points are further transformed to the global coordinate defined for the scenario, denoted by $(x_{\text{humanform}}, y_{\text{humanform}}, z_{\text{humanform}})$. The body orientation effect is included in the data presentation as well, where parameter α is used to indicate the facing direction of the person relative to the observation station. In addition, Rayleigh distributed power is endowed to each interactive point, representing the body absorption loss with a fixed total loss of 5 dBm over the whole body. With both the direction and power information, a standing human form model is illuminated as in Fig. 2.

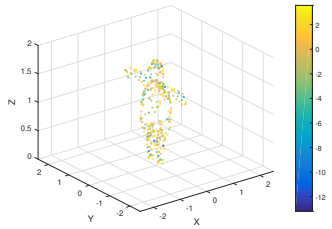

Fig. 2. 3D representation of standing human form point cloud

However, a set of point cloud is irregular and cannot be processed directly by CNN. Besides, the number of reflection points that can be detected in practice can be extremely sparse compared to the training data dimension. Thus, we project the 3D data representation onto the XY and XZ planes respectively to generate two 2D images. Considering an observation station that can detect up to 256 reflection points, a 16×16 image can be obtained with each pixel corresponding to a reflection point. These images are then used as input to a standard CNN model for localization and posture recognition.

Table I
Body size

height	1.6 m
shoulder width	0.4 m
head diameter	0.2 m
leg length	0.6 m
arm length	0.8 m

III. Proposed CNN-based localization and posture recognition approach

In recent years, CNN has emerged as one of the most popular machine learning methods in a variety of fields. The most beneficial aspect of CNN over multi-layer perception (MLP) is that it has weights of shared characteristics. The CNN-based method can not only reduce the consumption of memory, but also improve the robustness of the system. The structure of our proposed system model is described below.

Referring to the system architecture shown in Fig. 3, the proposed system model incorporates a total of two convolutional layers, each followed by a max-pooling layer with a pool size of 2, three full connectivity layers and one output layer. The input of the network is the data representation of the scenario as discussed in Section II, and their feature

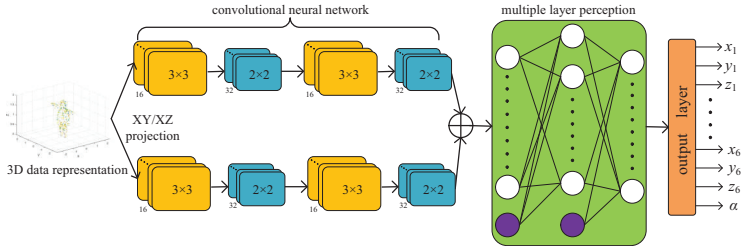

Fig. 3. Block diagram of the proposed CNN-based localization and posture recognition algorithm

representations can be automatically learned through the convolutional layers. The max-pooling layer is implemented in order to decrease the network complexity by reducing the length and width of the input of the previous layer. Pooling does not affect the number of filters. Number of filters in convolutional layers are 16 and 32 accordingly. The kernel size of the convolutional filters is 3×3, step size is 1, and padding is 1 in the same way. The design of the last convolutional layer follows a classical neural network structure, that is, to have one hidden layer for extendable flexibility in the non-linear modeling of the input-output relationship. The fully connected layers have 256, 128 and 64 units respectively. They take all neurons in the previous layer and connect them to every single neuron of current layer. To prevent network over-fitting, the network model uses dropout technique behind the first and second fully connected layers. Activation functions for all the convolutional and the first and second fully connected layers are ReLU function [21]. The ReLU introduces nonlinearity to the network and has simpler definitions in both function and gradient.

By giving a set of synthetic human form point cloud to the input layer, our goal is to estimate mass center locations and orientation of the target in the 3D space. As the target consists of 6 different body parts, each of which has a local mass center. Note that here the mass center is defined as the power weighted geometric center, which is

$$C(\phi') = \frac{\sum_{n=1}^{N} \alpha_n(\phi', \theta') \, exp(j * \phi'_n)}{\sum_{n=1}^{N} \alpha_n(\phi', \theta')}$$
$$C(\theta') = \frac{\sum_{n=1}^{N} \alpha_n(\phi', \theta') \, exp(j * \theta'_n)}{\sum_{n=1}^{N} \alpha_n(\phi', \theta')} \quad (2)$$

where $\alpha_n(\phi', \theta')$ is the power of the nth point, and N is the set of all points in the each body parts. Therefore, the output layer consists of $6 \times 3 + 1 = 19$ nodes corresponding to the 3D mass center locations (x, y, z) for 6 body parts as well as the orientation of the target. The output layer uses the softmax activation function [22] because it generates a well-performed probability distribution of the outputs. Through a large amount of simulation process, we set the learning rate and the mini-batch size of the Adam optimizer [23] to 0.001 and 100, respectively. The mean-squared error (MSE) loss function is employed to measure the difference between the estimated output and the ground truth. To avoid overfitting, the epochs of the model is set to 400.

Our task refers to use the proposed network architecture model to continuously learn the training data to find the connection between the input data and the structure of the network and finally use the learned model to classify the test data. Therefore, the proposed model is a supervised learning method. According to the above statement, our proposed algorithm includes two stages: the offline training and the online localization and posture recognition. Specifically, we create human form point cloud as input to the network, to train the network parameters in the offline phase. By minimizing the values of the loss function with the back propagation (BP) algorithm, the parameters of the proposed network can be updated by the stochastic gradient descend (SGD) method. In the online phase, the mass center locations and the orientation of the subject are predicted using the trained network and the new human form point cloud data. The signal flow graph with the input and output dimensions are shown in Table II.

Table II
SIGNAL FLOW GRAPH OF OUR PROPOSED SYSTEM MODEL

Layer Name	Input Size	Output Size
Input	$16 \times 16 \times 1$	$16 \times 16 \times 1$
Con2D_1	$16 \times 16 \times 1$	$16 \times 16 \times 16$
maxpool_1	$16 \times 16 \times 16$	$8 \times 8 \times 16$
Con2D_2	$8 \times 8 \times 16$	$8 \times 8 \times 32$
maxpool_2	$8 \times 8 \times 32$	$4 \times 4 \times 32$
FC1(full connected layer)	$4 \times 4 \times 64$	$1 \times 1 \times 256$
FC2	$1 \times 1 \times 256$	$1 \times 1 \times 128$
FC3	$1 \times 1 \times 128$	$1 \times 1 \times 64$
Output	$1 \times 1 \times 64$	$1 \times 1 \times 19$

IV. NUMERICAL INVESTIGATIONS

To verify and evaluate the performance of the algorithm, a large number of computer simulation experiments were performed on the synthetic datasets. The volume of the confined space is $5 \times 5 \times 5$ m^3. The target person is assumed to hold 3 postures: standing, sitting on a chair, and lying on the ground; the human form point clouds are generated accordingly. The sitting and lying human form point clouds are generated in a similar way to those for standing. The target faces the observation station when he is standing or sitting. The target lays down with the top of his feet pointing toward the vertical plane on which the observation station is set. In all postures, the synthetic data are split into three different parts, namely, a training part (with 3000 trials), a validation set (with

1000 trials), and a test set (with 1000 trials) for localization and posture identification. Simulation data used for training and testing the proposed framework are created according to Section II. Each reported simulation result is averaged over

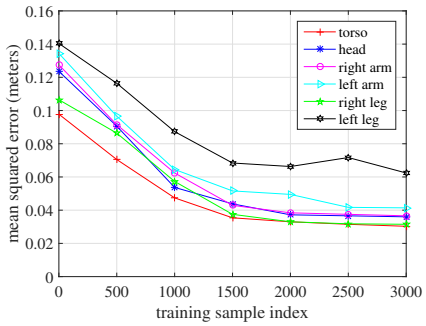

Fig. 4. The MSE of the predicted 6 mass center locations (in all three dimensions) when the subject is standing

200 Monte Carlo trials for the training set and a statistical test set is considered to improve the quality and strength of the results. Note that, our proposed scheme requires a total of 5 minutes to train the network. After the training phase, the computation time to estimate a new mass center location or orientation is measured at 0.1 seconds which is sufficiently fast to be used in practice.

The mean-squared error values of the mass center locations versus the number of samples when the target is standing is plotted in Fig. 4. Note that the abscissa in Fig. 4 refers to the number of samples NOT for the estimation phase but for the training phase; this applies for Fig. 5, Fig. 6 and Fig. 7 as well. As can be seen from Fig. 4, the torso and right leg offer better localization results than those of other parts of the subject. This can be attributed to the fact that these two parts contain dense cloud points which can be perceived completely by the observation station, the other parts, however, are occluded to a varying degree when the subject is rotated 30° in a counterclockwise direction. Moreover, the behavior of MSE for assessing the estimation accuracy of our proposed method when the target is sitting on a chair is shown in Fig. 5. From both figures, the algorithm can achieve a satisfactory estimation accuracy for the right leg. The algorithm, for the left leg, does not work as good as the right leg, mainly because of the fact that the cloud points on the left leg are occluded by the right leg when subject is standing with $\alpha = 30^\circ$. Finally, the MSE versus the number of samples when the target is lying on the ground is shown in Fig. 6. Compared with the localization results of the head in the standing and sitting postures, the algorithm yields the largest MSE for head when the person is lying. This may be due to the reflection points on the head are bit further away from the observation station. Based on the proposed method, the overall deviations for all postures are generally acceptable since most of them lie within the target's width (about 0.45m). When the target is in standing posture, Table III shows that the proposed CNN-based scheme can offer average localization errors of 3.5 cm in X, 2.6 cm in Z and 4.0 cm in Y, respectively. The results show that our model offers a better localization accuracy than RF-Pose3D [24].

Fig. 7 shows the bar diagram of the MSE based on our proposed approach for predicting the orientation of the target. It is evident that the estimated results appear approaching the ground-truth, and the errors are less than 2°. It can be observed that the MSE of all the postures reduces significantly as training sample increases, and the errors tend to be stable when the number of training samples is higher than 2000.

Table III
LOCALIZATION ACCURAY COMPARISON

	Localization Accuracy		
	X	Z	Y
RF-Pose3D	4.2 cm	4.0 cm	4.9 cm
CNN-Pose3D	3.5 cm	2.6 cm	4.0 cm

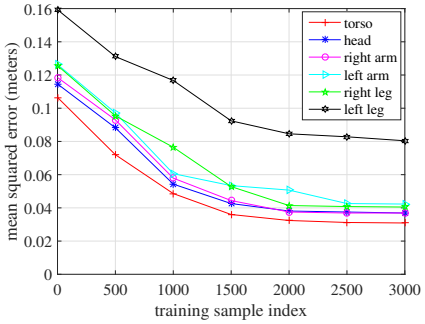

Fig. 5. The MSE of the predicted 6 mass center locations (in all three dimensions) when the subject is sitting on a chair

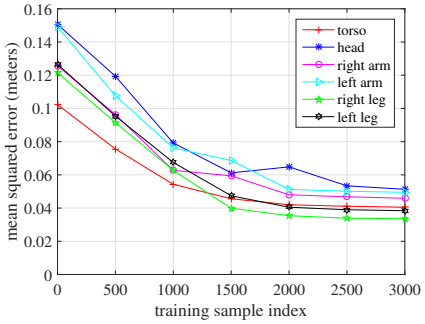

Fig. 6. The MSE of the predicted 6 mass center locations (in all three dimensions) when the subject is lying on the ground

From above, it can be confirmed that the localization method introduced in this paper works well. Besides, it is worth noting that the localization accuracy depends on the target's posture and orientation. In other words, the number of human form cloud points, as well as the distance and orientation of the target relative to the observation station have a strong impact on the localization accuracy. Table IV shows the success rate of recognizing human posture by the proposed method. As can be seen from this table, the lying posture yields the highest recognition rate as its point cloud perceived from the observation station is distinct from that of the other postures. The success rates of the standing and sitting postures are 94.2% and 93.6%, respectively. These results indicate that our

proposed CNN-based method can achieve the human posture recognition with high success probability.

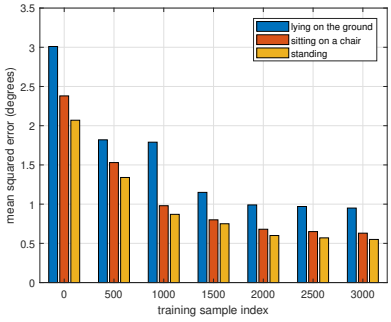

Fig. 7. The MSE of the predicted orientation of the subject for different numbers of training samples

V. Conclusion

This paper has proposed a CNN-based scheme for estimating the location and posture of a single human via the human form point cloud. The 3D points are created for 3 postures of human forms, namely standing, sitting and laying; the mass center locations and the orientation of the target person are predicted by customized CNN scheme. In terms of the estimation accuracy, the numerical results show that our proposed method works decently well given the size of the human body. Moreover, it is found that the estimation accuracy is improved 2 cm by increasing 20 cloud points and reducing 5 cm distance between the target and the assumed observation station. These results confirm that the proposed method can not only estimate human location but also identify human posture with acceptable accuracy.

For future work, we will adapt current CNN scheme to recognize dynamic postures such as walking and multi-person scenarios with sufficiently low online estimation speed.

Table IV
The correct recognition rate of the human posture

		Predicted class			
		standing	sitting on a chair	lying on the ground	correct rate
Actual class	standing	942	58	0	94.2%
	sitting on a chair	40	936	24	93.6%
	lying on the ground	0	2	998	99.8%

Acknowledgment

The work was supported by the Peng Cheng Laboratory under Grant PCL2018KP002, also supported by the Educational Commission of Guangdong Province under Grant 2017KZDXM075.

References

[1] M. Pham, D. Yang, and W. Sheng, "A sensor fusion approach to indoor human localization based on environmental and wearable sensors," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 1, pp. 339–350, 2019.

[2] H. Seke and Y. Hori, "Detection of abnormal action using image sequence for monitoring system of aged people," *IEEL Trans. Industry Applications*, vol. 122, no. 2, pp. 182–188, Feb. 2002.

[3] Y. Miao, E. Tanghe, J. Takada, and et al., "Measurement-based feasibility exploration on detecting and localizing multiple humans using MIMO radio channel properties," *IEEE Access*, vol. 8, pp. 3738–3750, 2020.

[4] A. Bourdoux, Y. Miao, and et al., "6G white paper on localization and sensing," *arXiv.org*, pp. 1–38, 2020.

[5] K. K. et al., "Experimental evaluation of estimating living-body direction using array antenna for multipath environment," *IEEE Antennas Wireless Propag. Lett.*, vol. 13, pp. 718–721, Mar. 2014.

[6] D. Sasakawa and K. K. et al., "Localizing living body using bistatic MIMO radar in multi-path environment," in *Proc. 8th Eur. Conf. Antennas Propag. (EUCAP)*, pp. 3253–3257, Apr. 2014.

[7] K. K. et al., "Localizing multiple target using bistatic MIMO radar in multi-path environment," in *Proc. IEEE Int. Workshop Electromagn., Appl. Student Innov. Competition (iWEM)*, vol. 2, pp. 90–91, Aug. 2014.

[8] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[9] T. Miwa, S. Ogiwara, and Y. Yamakoshi, "Localization of living-bodies using single-frequency multistatic Doppler radar system," *IEICE Trans. Commun.*, vol. 9, pp. 2468–2476, 2009.

[10] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmwave radars and CNNs," *IEEE Sensors J.*, vol. PP, pp. 1–1, 05 2020.

[11] E. Insafutdinov, B. Schiele, and et al., "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. Eur. Conf. on Computer Vision*, pp. 34–50, Springer, 2016.

[12] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4929–4937, 2016.

[13] G. Papandreou and K. M. et al., "Towards accurate multi-person pose estimation in the wild," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4903–4911, 2017.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[15] Y. Peng, M. Liao, Y. Song, Z. Liu, H. He, H. Deng, and Y. Wang, "FB-CNN: Feature fusion-based bilinear CNN for classification of fruit fly image," *IEEE Access*, vol. 8, pp. 3987–3995, 2020.

[16] C. S. Won, "Multi-scale CNN for fine-grained image recognition," *IEEE Access*, vol. 8, pp. 116 663–116 674, 2020.

[17] Z. Qiu, F. Yan, Y. Zhuang, and H. Leung, "Outdoor semantic segmentation for UGVs based on CNN and fully connected CRFs," *IEEE Sensors J.*, vol. 19, no. 11, pp. 4290–4298, 2019.

[18] K. V. Mardia, "Statistics of directional data," *J. Royal Statistical Society. Series B (Methodological)*, vol. 37, pp. 349–393, 1975.

[19] Y. Miao, K. Haneda, J. I. Naganawa, M. Kim, and J. I. Takada, "Measurement-based analysis and modeling of multimode channel behaviors in spherical vector wave domain," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5345–5358, 2020.

[20] Y. Miao, K. Haneda, M. Kim, and J. Takada, "Antenna de-embedding of radio propagation channel with truncated modes in the spherical vector wave domain," *IEEE Trans. Antennas Propag.*, vol. 63, no. 9, pp. 4100–4110, 2015.

[21] G. Wang, G. B. Giannakis, and J. Chen, "Learning ReLU networks on linearly separable data: algorithm, optimality, and generalization," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2357–2370, 2019.

[22] N. Ahmed, "Data-free/data-sparse softmax parameter estimation with structured class geometries," *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1408–1412, 2018.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Sci.*, pp. 7291–7299, 2018.

[24] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 7356–7365, Jun. 2018.