Micro-motion Target Classification Based on FMCW Radar Using Extended Residual Neural Network

Hai Le[1], Van-Sang Doan[2], Dai Phong Le[1], Thien Huynh-The[3], and Van-Phuc Hoang[1(✉)]

[1] Institute of System Integration, Le Quy Don Technical University, Hanoi, Vietnam
phuchv@lqdtu.edu.vn
[2] Faculty of Communication and Radar, Naval Academy, Nhatrang, Vietnam
[3] ICT-CRC, Kumoh National Institute of Technology, Gumi, Korea
thienht@kumoh.ac.kr

Abstract. Micro Doppler (m-D) effect is a phenomenon that provides signatures to discriminate different moving objects. Accordingly, this paper presents a novel residual convolutional neural network that can classify different moving targets based on m-D analysis of reflected frequency modulation continuous wave (FMCW) radar signals. The proposed network is optimized through the experiments of varying number of residual blocks. As a result, the proposed network yields the average classification accuracy of 93.48% with five residual blocks, 64 filters per convolution layer, and the filter size of 3×3. Moreover, thanks to the residual connection, our network remarkably outperforms two other existing networks in terms of accuracy.

Keywords: Convolution neural network · Micro Doppler · Moving target

1 Introduction

Autonomous vehicles have been yielding significant interest in the last decade, with considerable attention and investment from technology companies (such as Tesla, Waymo, and Baidu), governments, and academic research community [1]. To be able to complete driving autonomy on the road, the autonomous vehicles have to be equipped different types of sensors to provide the capability of sensing the surrounding environment and other moving objects, for example vehicles, humans, and animals. Indeed, various popular sensing technologies have been proposed, typically such as camera, LiDAR, or Radar [2]. The camera sensor [3] is used for objects classification based on color and texture signatures. They can be relatively cheap compared to the other types of sensors,

N.-S. Vo et al. (Eds.): INISCOM 2021, LNICST 379, pp. 104–115, 2021.
https://doi.org/10.1007/978-3-030-77424-0_9

however, camera devices suffer from limited depth of view, adverse weather, and light conditions. The second sensor is LiDAR (Light Detection and Ranging) [4], which uses steering laser arrays to produce an accurate 3-dimensional map of the surrounding environment around the autonomous vehicle. However, this sensor is still rather expensive and requires significant computational complexity to address the adverse effect of light and weather (for example, rainy, foggy, and snowy conditions). With above-mentioned disadvantages, cameras and LiDAR are not sufficient for a completely autonomous vehicle driving; therefore, radar sensor becomes a potential solution to overcome those disadvantage. Besides not being affected by light and weather conditions, the radar sensor can exploit the range-Doppler signature for entity classification processing [5].

Any movement of target poses a frequency shift in the radar return due to Doppler effect. Therefore, a moving target can be detected and recognized based on Doppler shift signatures. Since a target, for example, a helicopter flies, its blades rotate, or when a person walks, their arms swing naturally. These micro scale movements produce additional Doppler shifts, referred to as micro-Doppler (m-D) effects, which are useful to identify target features [6]. Furthermore, the m-D effects are modeled and simulated in some cases of micro-motion dynamics, such as vibration and rotation [7].

Recently, deep learning (DL) has been exploited to address many challenging detection and classification tasks in various applications, from computer vision [8–11] to medical informatics [12–14] and wireless communications [15–19]. For instance, Samaras et al. [20] have exploit DL for classifying different types of drone through a dataset handled from a surveillance radar. The classification method based on Deep Neural Network (DNN) was validated and reached the accuracy up to 95%. Not only drone, human motion also produces Doppler shift of reflected radar signal. Therefore, m-D radar is able to be applied to detect human motion, which is usually employed in automotive vehicle application [21]. In another scenario of parking monitoring application, Garcia et al. [22] have presented an effective convolutional neural network to classify radar images in order to detect vacant parking spaces with a 77-GHz imaging radar. Not only the convolutional neural network, the recurrent neural network is also considered for m-D target classification task, as presented in [23]. In addition, Angelov et al. [24] and Mento et al. [25] have demonstrated that the combination of convolution and LSTM (Long Short Term Memory) can facilitate the network model to be more stable.

Despite improving classification accuracy, the above-mentioned models reveal several weak points, including vanishing, over-fitting, and more computational complexity. Therefore, to effectively handle those limitations, a novel neural network for classifying m-D radar targets is proposed in this paper. Accordingly, the proposed network uses skip-connections to extract more strong features at many former layers, which can improve the classification accuracy. Additionally, the convolution layer is configured with grouped convolution operation that can noticeably reduce the number of network parameters. As a result, our network

with five residual blocks attains high classification accuracy and remarkably outperforms two other existing models.

2 Doppler Effect and Time-Frequency Spectrogram

The Doppler effect occurs if a target has relative movement to the radar. In that case, the corresponding frequency shift is described as follows:

$$f_d = f_0 \cdot \frac{2 \cdot \boldsymbol{v} \cdot \boldsymbol{r}}{c}, \tag{1}$$

where f_d is the Doppler frequency shift, f_0 is the center carrier frequency of the radar signal, $\boldsymbol{v}$ is the target velocity, $\boldsymbol{r}$ is the radial range vector from the radar to the target, and c is the speed of electromagnetic wave. For a non-rigid target (for example human and bicycle), micro-motions have mechanical vibrations and rotations which usually exist along with bulk translation. As a result, the spectrogram transformed from a radar signal of a moving target will contain different Doppler signatures unexpectedly. Therefore, the Doppler signatures of radar signals reflected from walking pedestrian and running bicycle are important information for detection and classification.

The m-D signature can be visually plotted as a spectrogram using the short-time Fourier transform (STFT) that given as follows:

$$X(\tau, \omega) = \mathrm{STFT}\{x(t)\} = \int_{-\infty}^{+\infty} x(t)\, w(t-\tau) e^{-j\omega t} dt \tag{2}$$

where $x(t)$ is the input signal of transformation and $w(t-\tau)$ is the kernel (so-called window) function. The resolution of STFT spectrogram is identified via the window function and the overlapping rate.

3 Proposed Neural Network

In this study, we propose a neural network for recognizing various types of FMCW radar target, including human, bicycle, and combination. The network model is designed with residual connection to reuse the former feature maps, which increase the classification accuracy via enhancing learning efficiency. As shown in Fig. 1a, the proposed network architecture consists of an input block, a series of residual blocks, and an output block. The blocks are built from several function layers which are connected to each other in certain flows. Generally, each layer can be formalized by a function as follows:

$$y^l = f\{x^l, w^l\} + b^l \tag{3}$$

where y^l is the output of l-th layer, $f\{\}$ denotes a layer function, x^l is the input of l-th layer, w^l is the set of weights of l-th layer, and b^l is the bias of l-th layer.

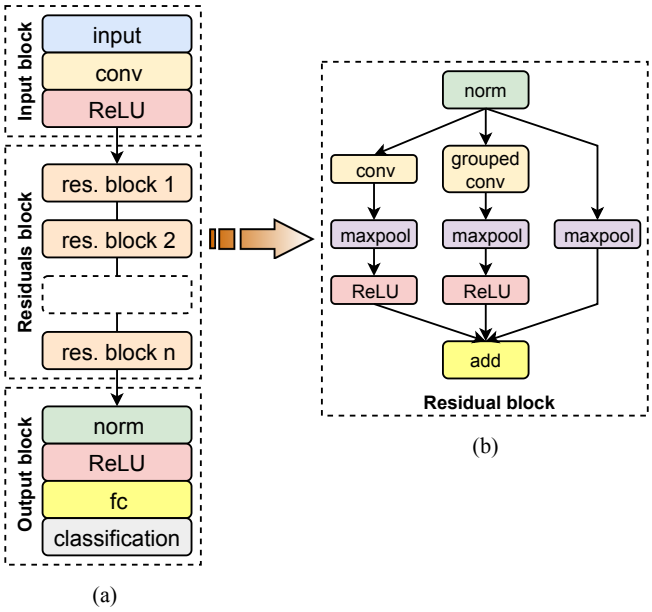

Fig. 1. Structure of the proposed network model: (a) Overall; (b) Residual block.

The input block is structured by an input layer, a convolution layer, and a ReLU (Rectified Linear Unit) activation layer. Specifically, the input layer is designated with the size of 400×144 to be appropriate to the size of spectrogram image. The input layer is followed by a convolution layer (conv) specified by 64 filters of size 1×1 to provide 64 channels at the output. Then, a ReLU activation layer is added to return the identical value with the positive input and the zero value with the negative input.

As the primary components to learn representational features at multi-scale resolutions, several residual blocks are organized in a cascade. Each residual block consists of three branches as shown in Fig. 1b. The first branch has three layers, including conv, max pooling (maxpool), and ReLU layers. The conv layer is designed by 64 filters of size 3×3. The maxpool layer follows the conv layer, which is configured by the pool size of 3×3 and the stride of (2,2). The spatial size of feature maps halves at the output of maxpool layer. The second branch also contains three layers with the same structure of the first one. However, the convolution layer of the second branch is a grouped type. The difference between the standard convolution and the grouped convolution is indicated in Fig. 2. Obviously, a grouped convolution with g groups uses g times fewer parameters and g times lower computational cost than a standard one. In particular, in this study we divide 64 filters into 8 groups with 8 filters of each. As a result, the number of learnable parameters in grouped conv layer reduces by 8 times if compared with a standard conv layer besides consuming a lower cost. The last branch is so-call a skip-connection from the output of normalization layer

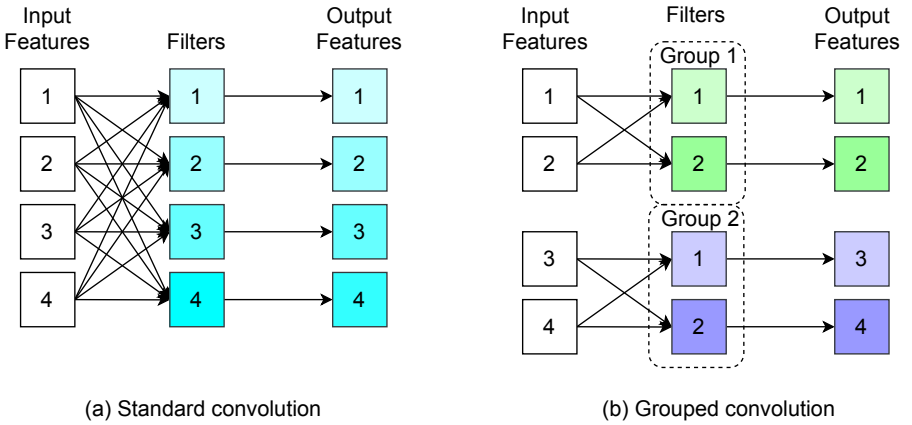

Fig. 2. Difference between (a) standard and (b) grouped convolution.

to the input of addition layer. Because the feature map size at the output of the first and second branches equal to a half of their input, a maxpool layer should be inserted in the skip-connection that produces a feature map with the same spatial size of the two other branches. Notably, each residual block is started with a normalization layer (norm) and finalized with addition layer. The normalization layer is used to accelerate the neural network training progress with a more stable weight update scheme through normalization of the input features by re-centering and re-scaling. The addition layer performs an element-wise addition operation with three inputs resulted from three branches, wherein they have the same volume size. Accordingly, the useful features can be enhanced through each residual block.

The last block of our model is output block, which contains in turn norm, ReLU activation, fully connected (fc), softmax, and classification layers. The fc layer performs flattening its input feature maps into a vector. The output of fc layer is assigned 5 classes being identical to the number of targets in a given dataset (Ped, Bic, Ped + Ped, Bic + Pred, and Bic + Bic). The output values of fc layer is transferred to the softmax layer, where the probability (score) of each class is calculated by the following equation:

$$\rho_i(\mathbf{z}) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{4}$$

where $\mathbf{z}$ is the input vector, C is the number of classes, and i and j are the indices of a element in the vector $\mathbf{z}$. At the end, the Classification layer is the last one which executes the target judgement based on the highest score provided by the softmax layer. Accordingly, the predicted target is defined as follows:

$$Target_{predicted} = \arg\{\max\{\rho_i(\mathbf{z})\}\} \tag{5}$$

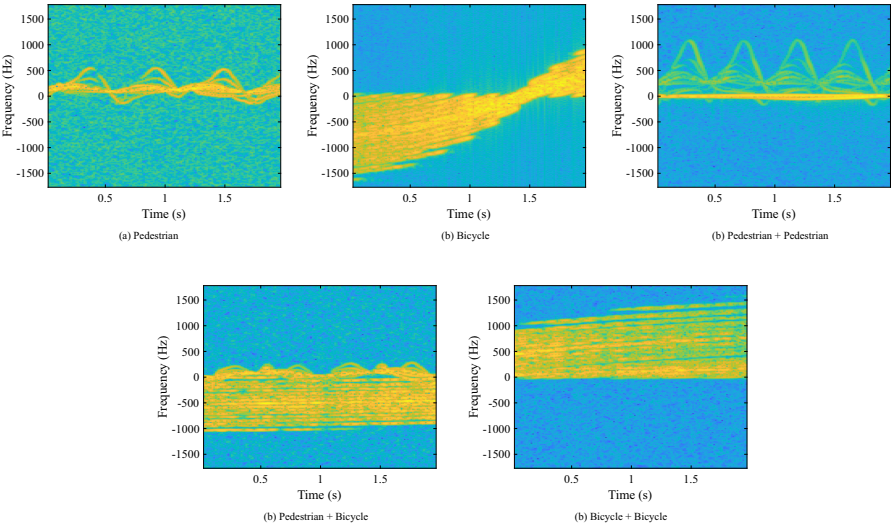

Fig. 3. m-D Spectrograms of several types of targets

4 Experiment Results and Discussion

4.1 Dataset Generation

In this study, we perform evaluation of the designed model based on the synthesized dataset, which is generated using the Matlab simulation program. Accordingly, we designed a radar simulation platform that transmits the FMCW signals to measure different types of target, including walking human (denoted Pedestrian - Ped) and cycling human (denoted Bicycle - Bic). Here, we assume that the swing movements of human hands and legs when walking produce the m-D effect in the spectrogram figure (Fig. 3a). Whereas, the cycle movement of bicycle wheels when running pose the m-D effect (Fig. 3b). The Doppler shift due to translation movements of pedestrian and bicyclist are also analyzed. In addition, several combination scenarios are taken into account to make more challenging to the network. In particular, the combinations of two walking people (denoted as Ped + Ped, Fig. 3c), walking people with running bicycle (denoted as Ped + Bic, Fig. 3d), and two running bicycles (Bic + Bic, Fig. 3e).

The radar used in the simulation platform transmits FMCW signal at 24 GHz band, the bandwidth of 250 MHz, and the waveform repetition time of 1 μs. The bicycle moves with speed less than 10 m/s. The simulation scenario is realized with varying properties, for example, bicyclists pedaling at different speeds and pedestrians with different heights walking at different speeds. As another assumption, the radar is fixed at the origin coordinate, meanwhile moving targets are uniformly distributed in a rectangular area of [5, 45] and [−10, 10] m from the origin location. Detailed other configurable parameters of dataset generation are listed in Table 1. It is noted that U{·} presents the uniformly contribution.

As a result, there are total 25,000 spectrogram images transformed from radar signals using STFT in the synthesized dataset, in which each target class has 5,000 images. We divide whole dataset into 20,000 images (80%) for training CNN model, and the remainder (20%) for testing the model performance.

Table 1. Crucial parameters of dataset generation.

Radar		Human		Bicycle	
Frequency	24 GHz	Height	U{1.5, 2} (m)	Gear rate	U{0.5, 6}
Bandwidth	250 MHz	Heading	U{−180, 180} (deg.)	Heading	U{−180, 180} (deg.)
Rep. frequency	2 μs	Speed	U{0.1, 2.8} (m/s)	Speed	U{1, 10} (m/s)

4.2 Experiment Results

The proposed model is trained in 20 epochs with the mini batch-size of 32, the initial learning rate of 0.01 with a drop factor of 0.1 after every 4 epochs. The stochastic gradient descent optimizer is applied in the training process with the computer hardware: CPU Core i5-9300H @2.4 GHz, RAM 8 GB Bus @2667 MHz, and GPU NVIDIA GeForce GTX 1660ti 6 GB.

In the first experiment, we train the network model with different residual blocks (from one to six blocks) on the training set without additive Gaussian noise. However, the trained network is then evaluated on the test set with adding the artificial Gaussian noise. The result of this experiment is shown in Fig. 4, where the classification accuracy of the network is improved along with the increment of the number of residual blocks (where the network goes deeper). Interestingly, the accuracy is significantly improved with smaller number of residual blocks (for example, two blocks is better than one block around 7.00%), while a tiny gap is deducted if increasing from five to six blocks). It is worth noting that the network complexity increases further as the number of blocks increases. Accordingly, the network with five residual blocks should be chosen for a good trade-off between the classification accuracy and the network complexity.

From the first experiment result, the network with five residual blocks is selected to classify the m-D radar targets of the test set when adding the Gaussian noise of different SNR (signal to noise ratio) levels. As a result, Fig. 5 shows the target classification accuracy of proposed network under SNRs ranging from −10 dB to 30 dB with step size of 5 dB. Obviously, the accuracy is improved with the increment of SNR, especially from 5 dB to 20 dB. The trained network yields the high accuracy for SNR $\geq$ 20 dB.

In the second experiment, we compare the performance of deep network trained on the training with and without additive noise. It should be noted that the accuracy is evaluated on the test set with noise. The comparison result is given in Fig. 6, where the network training with noise performs classification better than the network training without noise. Particularly, the network training with noise obtains the classification accuracy of $\geq$80% for SNR $\geq$ 0 dB and can

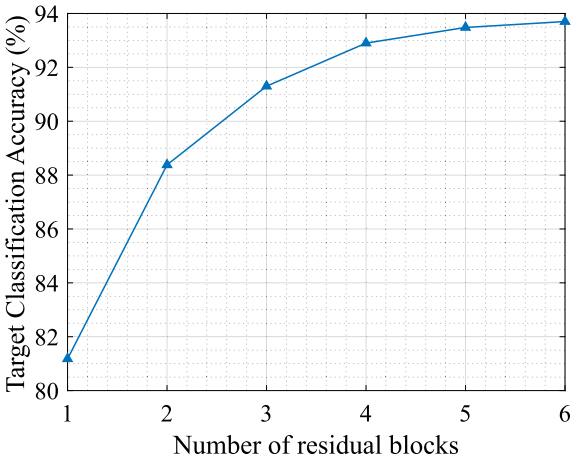

Fig. 4. The target classification accuracy of the proposed network with different number of residual blocks.

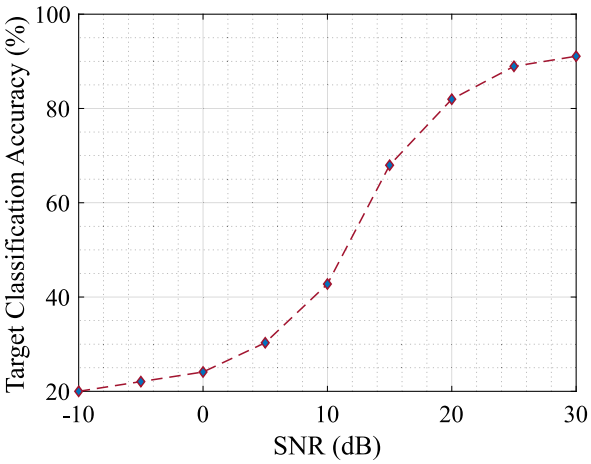

Fig. 5. The network performance in terms of classification accuracy when trained with dataset without noise.

achieve up to >90% for SNR $\geq$ 15 dB; whereas, the network training without noise yields accuracy of >80% for SNR $\geq$ 20 dB. It is observed that training with noise is significantly better than training without noise at SNR $\leq$ 20 dB. Therefore, it can be suggested that the network model should be trained with a diverse dataset to improve accuracy and prevent the over-fitting problem.

In the final experiment, we compare the proposed network of five residual blocks with two other existing models, including so-called Net01 in [26], and Net02 in [27].

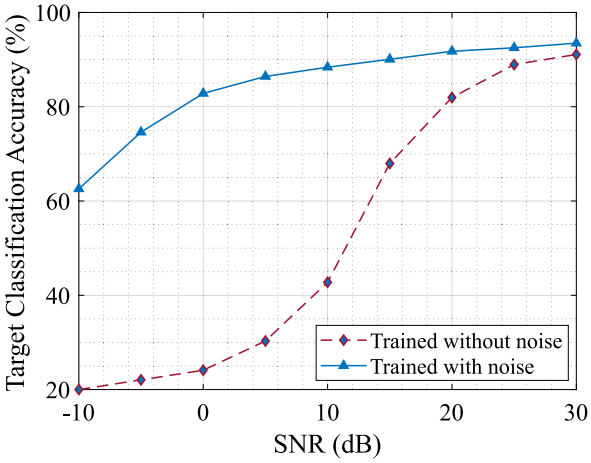

Fig. 6. Comparison of accuracy performance of proposed network when trained with different dataset options (with and without noise).

- Net01 is proposed for identifying indoor walking person using FMCW radar. The Net01 model consists of four conv layers with a filter size of 3×3 and number of filters in turn raising from 8 to 64. Each convolution layer is followed by 2×2 non-overlapping maxpool layer. Inserting to output of maxpool layer is an eLU (Exponential Linear Unit) activation layer that improves learning characteristics. Two fully connected layers are used as a classifier. The first fully connected layer is defined with output of 128, and second one is designed for five output classes. To prevent the over-fitting problem, the network is configured by a dropout layer with factor of 0.5 following the first fully connected layer. In this competition, the size of input layer of the Net01 model is modified from 256×45 to 400×144 to satisfy with the size of spectrogram images in our dataset.
- Net02 model is proposed for classifying different objects, including trolley, bike, cone, mannequin, sign and dog, based on 300 GHz radar data. The Net02 network is structured by three consecutive layer groups (where each group: conv + ReLU + maxpool). Three conv layers have the numbers of filters of 16, 32, and 64 with filter size of 5×5, 6×6, and 6×6, respectively. The maxpool layers has the pool size of 2×2. The third group is followed by a dropout layer with a factor of 0.5 that is employed to avoid the over-fitting problem in the training process. Following the dropout layer is conv and ReLU layers with 128 filters of size 3×3. The last convolution is designated by five filters of size 3×3 for being compatible with the number of classes in our dataset. In addition, the size of the input layer is modified to 400×144 for processing the spectrogram images.

Table 2. Performance comparison of the neural networks.

Networks	No. params	Accuracy (%)
Net01	418K	87.9
Net02	309K	88.2
Our model	230K	93.5

The above-mentioned networks are trained on the same dataset and the same configuration of the training process. The performance comparison of those networks is reported in Table 2, in which our proposed network remarkably outperforms two other ones. Specifically, Net01 is the largest network with the number of parameters of approximately 418K but obtains the lowest classification accuracy of 87.9%. Despite having the smallest number of learnable parameters of around 230K, our network achieves the highest accuracy due to leveraging residual connections to effectively re-usage highly meaningful features extracted from many former layers.

5 Conclusion

In this paper, we have proposed and designed a novel network architecture that is inspired by the residual convolutional neural network. The network is configured not only with skip-connection, but the grouped convolution is also employed to remarkably reduce the learnable parameters. Through experiments with different numbers of residual blocks, our network achieves the best trade-off performance with the 5-block configuration. In competition with other models, the proposed network of five residual blocks has significantly outperformed two other existing ones. For future works, we intend to design the model for more types of m-D radar targets and simultaneously improve the network performance in terms of classification accuracy and computational cost. Moreover, other signal pre-processing techniques will be taken into account to enhance the target classification accuracy.

References

1. Granath, E.: 5 Top Autonomous Vehicle Companies to Watch in 2020. Intelligent Mobility Xperience, 1 September 2020. www.intelligent-mobility-xperience.com/5-top-autonomous-vehicle-companies-to-watch-in-2020-a-958065/
2. RADAR, Camera, LiDAR and V2X for Autonomous Cars. NXP. www.nxp.com/company/blog/radar-camera-lidar-and-v2x-for-autonomous-cars:BL-RADAR-LIDAR-V2X-AUTONOMOUS-CARS
3. Rosique, F., Navarro, P.J., Fernández, C., Padilla, A.: A systematic review of perception system and simulators for autonomous vehicles research. Sensors **19**(3), 648 (2019)

4. Zhang, Y., Wang, J., Wang, X., Dolan, J.M.: Road-segmentation-based curb detection method for self-driving via a 3D-LiDAR sensor. IEEE Trans. Intell. Transp. Syst. **19**(12), 3981–3991 (2018)
5. Belgiovane, D., Chen, C.: Micro-Doppler characteristics of pedestrians and bicycles for automotive radar sensors at 77 GHz. In: 11th European Conference on Antennas and Propagation (EUCAP), Paris, pp. 2912–2916 (2017)
6. Chen, V.C.: The Micro-Doppler Effect in Radar. Artech House, Norwood (2011)
7. Chen, V.C., Li, F., Ho, S.-S., Wechsler, H.: Micro-Doppler effect in radar: phenomenon, model, and simulation study. IEEE Trans. Aerosp. Electron. Syst. **42**(1), 2–21 (2006)
8. Hua, C.-H., Huynh-The, T., Lee, S.: Convolutional networks with bracket-style decoder for semantic scene segmentation. In: Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, pp. 2980–2985 (2018)
9. Huynh-The, T., Hua, C.-H., Ngo, T.-T., Kim, D.-S.: Image representation of pose-transition feature for 3D skeleton-based action recognition. Inf. Sci. **512**, 112–126 (2020)
10. Huynh-The, T., Hua, C., Kim, D.: Encoding pose features to images with data augmentation for 3-D action recognition. IEEE Trans. Industr. Inf. **16**(5), 3100–3111 (2020)
11. Huynh-The, T., Hua, C.H., Tu, N.A., Kim, D.S.: Learning 3D spatiotemporal gait feature by convolutional network for person identification. Neurocomputing **397**, 192–202 (2020)
12. Huynh-The, T., Hua, C.H., Tu, N.A., Kim, D.S.: Physical activity recognition with statistical-deep fusion model using multiple sensory data for smart health. IEEE Internet Things J. (2020). https://doi.org/10.1109/JIOT.2020.3013272
13. Hua, C.-H., et al.: Bimodal learning via trilogy of skip-connection deep networks for diabetic retinopathy risk progression identification. Int. J. Med. Informatics **132**, 103926 (2019)
14. Hua, C.-H., Huynh-The, T., Lee, S.: DRAN: Densely reversed attention based convolutional network for diabetic retinopathy detection. In: Proceedings of the 42nd International Engineering in Medicine and Biology Conference (EMBC), Montréal, Québec, Canada, 20–24 July 2020 (2020)
15. Huynh-The, T., Hua, C., Kim, J., Kim, S., Kim, D.: Exploiting a low-cost CNN with skip connection for robust automatic modulation classification. In: Proceedings of 2020 IEEE Wireless Communications and Networking Conference (WCNC), Seoul, Korea (South), pp. 1-6 (2020)
16. Doan, V.-S., Huynh-The, T., Kim, D.-S.: Underwater acoustic target classification based on dense convolutional neural network. IEEE Geosci. Remote Sens. Lett. (2020). https://doi.org/10.1109/LGRS.2020.3029584
17. Huynh-The, T., Hua, C., Pham, Q., Kim, D.: MCNet: an efficient CNN architecture for robust automatic modulation classification. IEEE Commun. Lett. **24**(4), 811–815 (2020)
18. Doan, V.-S., Huynh-The, T., Hua, C.-H., Pham, Q.-V., Kim, D.-S.: Learning constellation map with deep CNN for accurate modulation recognition. arXiv preprint arXiv: 2009.02026 (2020)
19. Huynh-The, T., Doan, V.S., Hua, C.H., Pham, Q.V., Kim, D.S.: Chain-Net: learning deep model for modulation classification under synthetic channel impairment. arXiv preprint arXiv: 2009.02023 (2020)

20. Samaras, S., Magoulianitis, V., Dimou, A., Zarpalas, D., Daras, P.: UAV classification with deep learning using surveillance radar data. In: Tzovaras, D., Giakoumis, D., Vincze, M., Argyros, A. (eds.) ICVS 2019. LNCS, vol. 11754, pp. 744–753. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-34995-0_68
21. Ma, X., Zhao, R., Liu, X., Kuang, H., Al-Qaness, M.A.A.: Classification of human motions using micro-doppler radar in the environments with micro-motion interference. Sensors (Basel) **19**(11), 2598 (2019)
22. García, J.M., Zoeke, D., Vossiek, M.: MIMO-FMCW radar-based parking monitoring application with a modified convolutional neural network with spatial priors. IEEE Access **6**, 41391–41398 (2018)
23. Han, L., Feng, C.: Micro-doppler-based space target recognition with a one-dimensional parallel network. Int. J. Antennas Propag. **2020**, 1–10 (2020)
24. Angelov, A., Robertson, A., Murray-Smith, R., Fioranelli, F.: Practical classification of different moving targets using automotive radar and deep neural networks. IET Radar Sonar Navig. **12**(10), 1082–1089 (2018)
25. Minto, M.R.I., Tan, B., Sharifzadeh, S., Riihonen, T., Valkama, M.: Shallow neural networks for mmWave radar based recognition of vulnerable road users. In: 12th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP) (2020)
26. Vandersmissen, B., et al.: Indoor person identification using a low-power FMCW radar. IEEE Trans. Geosci. Remote Sens, **56**(7), 3941–3952 (2018)
27. Sheeny, M., Wallace, A., Wang, S.: RADIO: Parameterized generative radar data augmentation for small datasets. Appl. Sci. **10**(11), 3861–3873 (2020)