

Dynamic Hand Gesture Classification Based on Multistatic Radar Micro-Doppler Signatures Using Convolutional Neural Network

Zhaoxi Chen and Gang Li

Dpt. Electronic Engineering
Tsinghua University
Beijing 100084, China
chenzx16@mails.tsinghua.edu.cn
gangli@tsinghua.edu.cn

Francesco Fioranelli

School of Engineering
University of Glasgow
Glasgow G12 8QQ, U.K.
francesco.fioranelli@glasgow.ac.uk

Hugh Griffiths

Dpt. Electronic & Electrical Engineering
University College London
London WC1E 7JE, U.K.
h.griffiths@ucl.ac.uk

Abstract—We propose a novel convolutional neural network (CNN) for dynamic hand gesture classification based on multistatic radar micro-Doppler signatures. The time-frequency spectrograms of micro-Doppler signatures at all the receiver antennas are adopted as the input to CNN, where data fusion of different receivers is carried out at an adjustable position. The optimal fusion position that achieves the highest classification accuracy is determined by a series of experiments. Experimental results on measured data show that 1) the accuracy of classification using multistatic radar is significantly higher than monostatic radar, and that 2) fusion at the middle of CNN achieves the best classification accuracy.

Keywords—convolutional neural network, data fusion, dynamic hand gesture classification, micro-Doppler, multistatic radar

I. INTRODUCTION

Dynamic hand gesture recognition has attracted increasing attention in recent years. Numerous noncontact approaches based on radar have been proposed [1], which cause less distraction to the user and provide a more comfortable experience than approaches using wearable sensors. Most algorithms of dynamic hand gesture recognition with radar sensors are based on micro-Doppler analysis [2-6]. In the conventional two-phase classification algorithms, features such as the empirical features [2], the principal component analysis based features [3], and the sparse features [4][5] are first extracted from the time-frequency domain and then fed into an off-the-shelf classifier, such as the nearest neighbor, the support vector machine, and the decision trees. The emerging deep neural network, including CNNs, which have enjoyed great successes in various fields [7], is regarded as another powerful tool for dynamic hand gesture classification. CNN is directly applied to the time-frequency spectrogram [6], or combined with recurrent neural networks [8] to recognize the temporal pattern of hand gestures.

In recent studies of human activity classification [5][9-11], multistatic radar and multi-angle radar systems have shown

their potential to provide higher classification accuracy than single monostatic radar sensors, thanks to the diversity of viewing angles. The key to high-accuracy classification with a multistatic radar is a fusion strategy that makes full use of the signals from different receivers. In [9], empirical features extracted from three perpendicular radar sensors are concatenated and fed into a decision tree for dynamic hand gesture classification. Similar approach is adopted in [5] where the sparse features of hand gestures from multi-angle radar receivers are concatenated before being sent to the nearest neighbor classifier. While the aforementioned studies apply fusion to the features (mid fusion), late fusion is adopted in [10][11], where the decisions, i.e., the classification results obtained from different receivers, are fused to give the final output. Binary voting [10] and weighted averaging [11] are used to fuse the classification results individually obtained from three multistatic channels for recognition of armed and unarmed personnel targets. However, the data fusion in each of the above mentioned works operates at pre-designed positions, which may not be the optimal.

In this paper, we propose a novel CNN for dynamic hand gesture recognition based on multistatic radar micro-Doppler signatures. In the proposed CNN, micro-Doppler spectrograms of different receivers are processed by a series of convolutional layers, where data fusion is carried out at an adjustable position between two consecutive layers. The effect of the fusion position on classification accuracy is further investigated and the optimal fusion position is determined by experiments. The proposed CNN is evaluated on real data collected by a multistatic radar that contains one transmitter antenna and four receiver antennas. Experimental results show that the average classification accuracy is about 63% when using only one receiver, over 83% when using two receivers, and close to 99% when using all the four receivers. This confirms the advantage of multistatic radar, and that the best classification accuracy is achieved by fusion of the mid-level features extracted by a convolutional layer in the middle of the CNN.

The remainder of this paper is organized as follows. The radar system and data collection are described in Section II. In Section III, we present the proposed CNN in detail. In Section IV, the experimental results based on measured data are demonstrated. The conclusions are given in Section V.

This work was supported in part by the National Natural Science Foundation of China under Grant 61661130158, in part by the Royal Society Newton Advanced Fellowship, in part by the IET A. F. Harvey Prize awarded to Hugh Griffiths in 2013, and in part by the Engineering and Physical Sciences Research Council under Grant EP/G037264/1. (Corresponding author: Gang Li.)

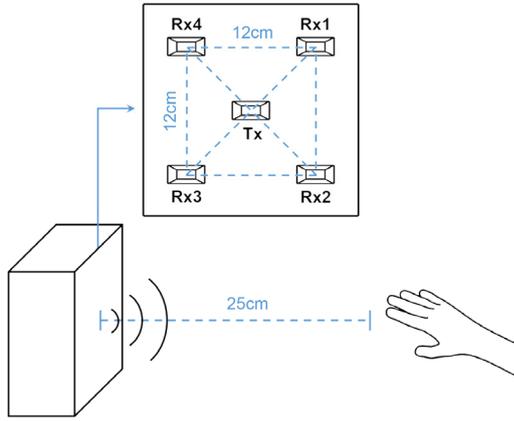


Fig. 1 The multistatic radar system used for measurement.

II. MEASUREMENTS AND DATASET

A. Measurement Setup

The data presented in this paper are collected by a Frequency Modulated Continuous Wave (FMCW) multistatic radar system operating at 24 GHz. The frequency of the transmitted signal is sawtooth modulated with a bandwidth of 500 MHz, which ensures that 1) the hand always stays within one range bin, and that 2) interference caused by the movement of the subject's chest wall can be easily removed by range bin selection. The pulse repetition frequency is set to 1 kHz, which allows the velocity of hand to be captured without any Doppler ambiguity. As shown in Fig. 1, the radar system records four coherent channels using one transmitter antenna and four identical receiver antennas co-located in a vertical plane. The transmitter antenna is located at the centroid of a 12 cm \times 12 cm square, while the four receiver antennas are at the corners. All the antennas are vertically polarized. A seated subject performs hand gestures in front of the antennas at a distance of about 25 cm. After data collection, a narrow notch filter is utilized in frequency domain to remove the static clutter.

B. Dynamic Hand Gesture Dataset

The aforementioned multistatic radar is capable of measuring three-dimensional velocity of the observed target, making it feasible to distinguish dynamic hand gestures with similar radial motions. To emphasize this advantage, six different dynamic hand gestures are considered in this paper. Generally speaking, these gestures can be divided into two groups: a) Swiping, including swiping up, down, left and right and b) Circling, including circling clockwise and counter-clockwise. In our experiments, all the above gestures are performed with fingers loosely close to each other, and the hand moves back to the starting point gently and naturally after performing each hand gesture. The readers may refer to Table I for a detailed instruction of each gesture.

Before taking a deeper look into the gesture set, we simply assume the hand to be a mass point at its centroid. Gestures within the same group will share almost the same radial motions under this assumption. To help understanding, colors that indicate radial velocity are superposed on the stokes of some gestures in Fig. 2. When the subject is performing the gesture of swiping up (Fig. 2 (a)), the radial velocity of his/her hand changes from positive to negative which is the same for the gestures of swiping down, left, and right. As for the gesture group (b), circling to either direction have ideally zero radial

TABLE I
ALL DYNAMIC HAND GESTURES USED FOR EXPERIMENT
(SHOWN IN FRONT VIEW)

(a) Swiping				(b) Circling	
(SU)	(SD)	(SL)	(SR)	(CCCW)	(CCW)
Swiping Up	Swiping Down	Swiping Left	Swiping Right	Circling Counter-clockwise	Circling Clockwise

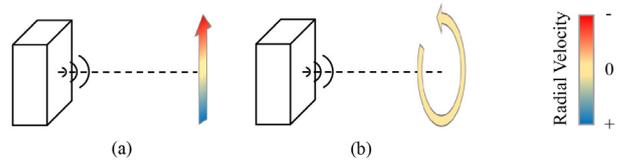


Fig. 2 Radial velocity of selected dynamic hand gestures. (a) swiping up; (b) circling counter-clockwise

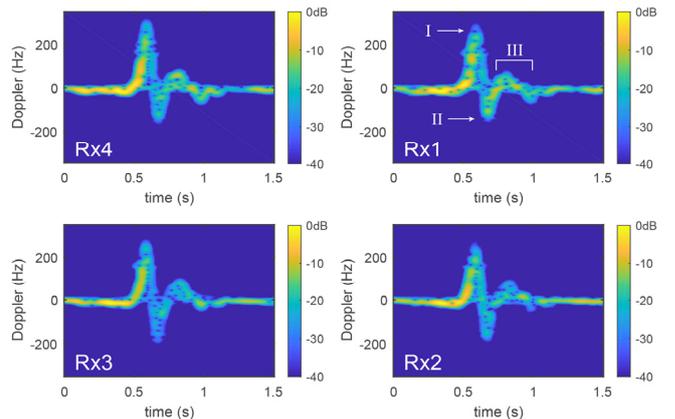


Fig. 3 STFT spectrograms of the hand gesture "swiping up" obtained at all four channels. Positive Doppler frequency indicates moving towards the radar.

velocity (Fig. 2 (b)) if the hand trajectory is perfectly symmetric. As a result, gestures within the same group are hard to separate by using a monostatic radar, which is only sensitive to radial velocity. However, in practice the hand is not a point-like target and the measured hand trajectories are not necessarily symmetric. This means some information about the radial movements of the gestures may still be measurable by using monostatic radar. Even so, the classification accuracy produced by only using monostatic radar is far from satisfactory performances for real data experiments, as shown in Section IV.

By contrast, a multistatic radar with multiple coherent receiving channels measures hand gestures from different viewing angles. The information of angular motion can be captured by the subtle differences between channels of the received signal [12]. In Fig. 3, we demonstrate a typical recording of the hand gesture "swiping up" by applying the short-time Fourier transform (STFT) to the raw signal. The Doppler peaks I and II in Fig. 3 represent the radial velocity of the up-swipe, while segment III indicates that the hand

TABLE II
SPECIFICATIONS OF LAYERS IN THE PROPOSED CNN

Layer No.	Convolution kernels	Pooling after convolution
Conv 1	7×7 @16 ^a	2×2 mean pooling
Conv 2	5×5 @32	2×2 mean pooling
Conv 3	5×5 @64	2×2 mean pooling
Conv 4	3×3 @128	2×2 mean pooling
Conv 5	3×3 @128	-
Conv 6	3×3 @128	-
Conv 7	3×3 @128	-
Avg 8	-	Global mean pooling
Fc 9	@256	-
Fc 10	@6	-
Fusion	1×1 @variable ^b	-

^a kernel size @ kernel number.

^b The kernel number is determined according to the position of the fusion layer.

gently moves back to the starting point. The four spectrograms obtained at four receiving channels have different strengths and heights of the Doppler peaks, as the evidence of the angular diversity.

The specification of the dataset is as follows. We asked 8 right-handed adults to perform each hand gestures for 30 times, generating a dataset containing $8(\text{subjects}) \times 6(\text{gestures}) \times 30(\text{repetitions}) = 1440$ recordings. We symmetrically zero-padded all the recordings to the same time duration. To make the measurement close to real life scenario, the subjects were given only necessary instructions on how to perform the gestures. Variations were even encouraged in certain gestures, that is, we encouraged the starting point of the Circling gestures to be uniformly distributed on the circle.

III. THE PROPOSED ALGORITHM

We propose a CNN architecture designed for dynamic hand gesture classification using the multistatic radar system described in section II. The STFT spectrograms of the multi-channel micro-Doppler signatures are adopted as input to determine the class label out of six different hand gestures.

A. Data Pre-processing

In our previous work [13], we pre-processed the raw signal by computing the STFT spectrograms with three different window sizes. Such multi-resolution spectrograms provide richer time-frequency information of the signal and thus contribute to higher accuracy of human gait classification. We adopt similar pre-processing approach in this paper. Three STFT spectrograms are computed for each channel using the Blackman windows of length 32, 64, and 128 time samples, respectively. The FFT length is set to 128 points, and the step of sliding window is set to 10 time samples. Three spectrograms with different window lengths are aligned and then stacked along the third dimension, and finally fed into the proposed CNN. The dimension of the input is $387(\text{time}) \times 128(\text{Doppler}) \times 3(\text{resolution})$ according to the aforementioned parameters.

B. The proposed CNN

The proposed CNN is a multiple-input-single-output network constructed by inserting a fusion layer into a sequence of layers, which are seven convolutional layers (conv1 – conv7) followed by a global mean pooling layer (avg 8) and two fully connected layers (fc9 – fc10). Specifications

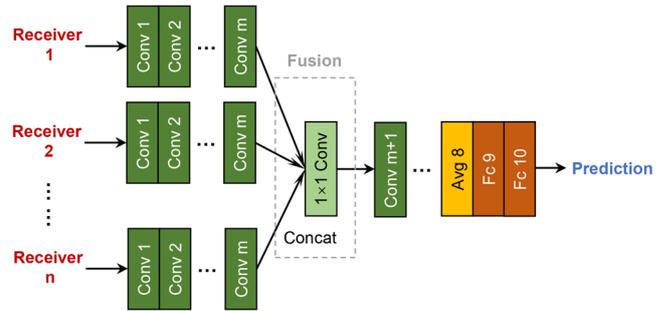


Fig. 4 The general architecture of the proposed CNN

of the layers, including kernel number, kernel size, and pooling size, are summarized in Table II. Fig. 4 demonstrates the general architecture of the proposed CNN, where the fusion layer is inserted between the m th and $(m+1)$ th convolutional layer. The 1st to the m th convolutional layers are copied n times, with shared weights, to form individual branches for inputs from different receivers, whilst the other layers are utilized to process the fused data. The position of the fusion layer is adjustable, that is, it can be inserted before any convolutional layers and avg 8. The proposed CNN processes the input signal in three steps. 1) The spectrograms of n receivers are processed by the branches. We denote the dimension of the output feature map of each branch as $W \times H \times D$. 2) These feature maps are concatenated along the third dimension (with shape $W \times H \times nD$), and then go through a 1×1 convolutional layer to reduce the dimension back to $W \times H \times D$ ¹. These two operations are carried out within the fusion layer. 3) The fused feature map is processed by the remaining layers to make the final decision. When only one input exists (i.e. monostatic radar), the fusion layer is omitted and the proposed CNN degrades to a sequential network.

We use the Rectified Linear Unit as activation function throughout the network, except for the last fully connected layer Fc 10 that needs a Softmax activation. Batch normalization [14] is also included, which prevents gradient vanishing or explosion and mitigates overfitting by standardizing the feature maps within each mini batch. We follow the approach used in resNet [15], that is, batch normalization is applied right after convolution. Moreover, mean pooling is used after the first four layers to reduce the dimension of the feature maps.

C. Implementation Details

The proposed CNN is implemented on Keras [16]. We train the network by Adam [17] optimizer for 100 epochs in total. In the first 80 epochs, the learning rate is set to 3×10^{-3} , which is decreased to 3×10^{-4} for the last 20 epochs. In addition, weight decay of 5×10^{-4} , dropout [18], and data augmentation are used in the training phase.

IV. EXPERIMENTAL RESULTS

The proposed CNN is evaluated on the measured hand gesture dataset described in Section II. All the following experimental results are based on leave-one-subject-out cross

¹ The output depth of the 1×1 convolution should be at least $3n$, which equals the total dimension of the multiple inputs, where n is the number of inputs and three is the depth of the input spectrogram. This constraint avoids dimension reduction during data processing and thus prevents potential loss of information.

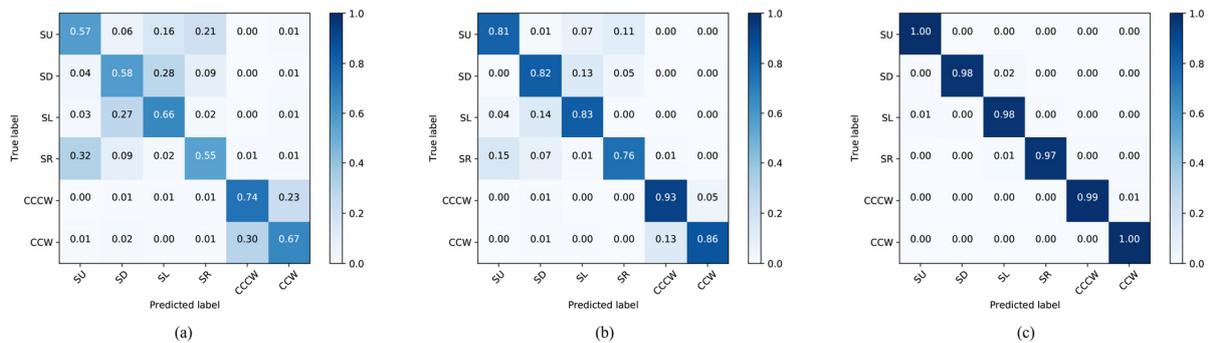


Fig. 5 Confusion matrices for (a) monostatic input, (b) multistatic input using 2 receivers, and (c) multistatic input using 4 receivers. The names of hand gestures are abbreviated and the full names could be found in Table I.

TABLE III
CLASSIFICATION ACCURACIES OF MONOSTATIC INPUT AND MULTISTATIC INPUTS

Receivers	Accuracy (%)	Average Accuracy (%)
1	61.61	62.89
2	62.45	
3	67.75	
4	59.76	
1, 3	81.19	83.32
2, 4	85.45	
1, 2, 3, 4	98.79	98.79

validation. Specifically, we choose one subject for testing and train the CNN on the remaining seven subjects. This procedure is repeated for all the subjects and the average classification accuracy is reported. The classification accuracy is defined as the ratio of correctly classified samples to all testing samples.

A. Comparison between Monostatic and Multistatic Inputs

We first evaluate the classification accuracy of the proposed network with respect to the number of receivers used for input. The following three cases are considered: 1) monostatic input: using signal from only one receiver antenna; 2) multistatic input: using the signals of two receiver antennas on the diagonal (receivers 1 & 3, or receivers 2 & 4), which has the longest baseline in all the two-receiver cases; 3) multistatic input: using signals from all the four antennas. Table III summarizes the classification accuracy of each case, which is merely 63% for the monostatic case, more than 83% for the two-receiver case and close to 99% for the four-receiver case.

To shed light on the reason for the performance gain, we plot the average confusion matrix of each case in Fig. 5. We observe a block diagonal pattern in the confusion matrices, which is the most obvious in the monostatic case and noticeable in the two-receiver case. The block pattern is caused by the difficulty in distinguishing hand gestures within the same group.

From Fig. 5 we can see that the classification accuracy in the monostatic case is unsatisfactory, which is consistent with the analysis in Section II. That is, it is difficult to distinguish the direction of gestures by only measuring radial velocity. Classification error reduces significantly in the two-receiver case, but around 17% error still remains, which indicates that two receivers are not sufficient for perfect classification. In the four-receiver case, almost every gesture is correctly

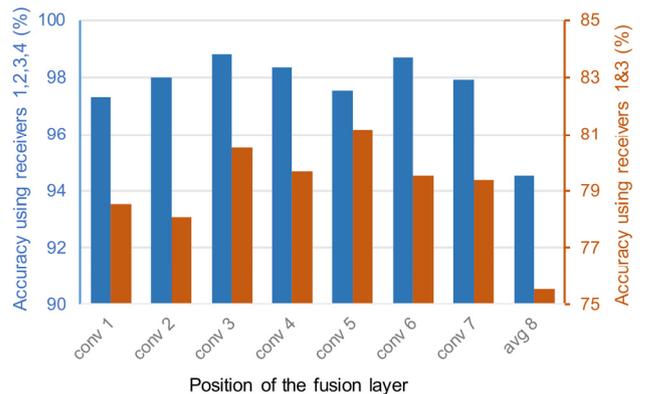


Fig. 6 Classification accuracies of the proposed CNN versus positions of the fusion layer

recognized thanks to the additional information collected by all the four receivers.

B. The effect of the fusion position

The effect of the fusion position on classification accuracy is investigated in this subsection, which is demonstrated in Fig. 6. The horizontal axis of Fig. 6 represents the layer before which the fusion layer is inserted. Results of the four-receiver case is plotted in blue on the left axis, whilst the two-receiver case (receivers 1 & 3) in orange on the right. In general, the proposed CNN achieves the highest accuracy when the fusion layer is at the middle of the CNN, though the optimal position for each case is different, which is the position right before conv3 for the four-receiver case, and conv5 for the two-receiver case. The explanation for this observation is two-fold. On the one hand, the high-level semantic features captured by the last few layers provide limited angular velocity information due to the low resolution of the feature maps. As discussed in Section II, angular information lies in the subtle differences between spectrograms from each receivers, which are concealed under the blurred high-level feature maps. On the other hand, the low-level features learned by the first several layers have weak representation capability, which is possibly insufficient to describe the complex hand gestures. As a result, fusion at a middle layer achieves a balance between these two factors and thereby the best accuracy.

V. CONCLUSION

In this paper, a novel CNN is proposed for dynamic hand gesture recognition based on multistatic radar micro-Doppler signatures. The proposed CNN enables data fusion at an adjustable position by inserting a fusion layer into different

positions among a sequence of convolutional layers. The optimal fusion position that achieves the highest classification accuracy is obtained by experiments. The best classification accuracy on the measured data using four receivers and one receiver are about 99% and 63%, respectively, indicating a substantial advantage of the multistatic radar. In future work, we will focus on more efficient, adaptive, and less human-dependent fusion strategies.

REFERENCES

- [1] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *Int. J. Mach. Learn. & Cyber.*, pp. 1–23, Aug. 2017.
- [2] G. Li, S. Zhang, F. Fioranelli, and H. Griffiths, "Effect of sparsity-aware time–frequency analysis on dynamic hand gesture classification with radar micro-Doppler signatures," *IET Radar, Sonar Navigat.*, vol. 12, no. 8, pp. 815–820, Aug. 2018.
- [3] Q. Wan, Y. Li, C. Li, and R. Pal, "Gesture recognition for smart home applications using portable radar sensors," in *36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Chicago, IL, 2014, pp. 6414–6417.
- [4] G. Li, R. Zhang, M. Ritchie, and H. Griffiths, "Sparsity-driven micro-Doppler feature extraction for dynamic hand gesture recognition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 2, pp. 655–665, Oct. 2018.
- [5] L. Yang and G. Li, "Sparsity aware dynamic gesture classification using dual-band radar," in *19th Int. Radar Symp. (IRS)*, Bonn, Germany, 2018, pp. 1–6.
- [6] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-Doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125–7130, Oct. 2016.
- [7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
- [8] S. Wang, J. Song, J. Lien, I. Poupayev, and O. Hilliges, "Interacting with Soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, Tokyo, Japan, 2016, pp. 851–860.
- [9] S. Lan, Z. He, K. Yao, and W. Chen, "Hand gesture recognition using a three-dimensional 24 GHz radar array," in *2018 IEEE/MTT-S Int. Microw. Symp.*, Philadelphia, PA, 2018, pp. 138–140.
- [10] F. Fioranelli, M. Ritchie, and H. Griffiths, "Centroid features for classification of armed/unarmed multiple personnel using multistatic human micro-Doppler," *IET Radar, Sonar Navigat.*, vol. 10, no. 9, pp. 1702–1710, 2016.
- [11] J. S. Patel, F. Fioranelli, M. Ritchie, and H. Griffiths, "Multistatic radar classification of armed vs unarmed personnel using neural networks," *Evolving Syst.*, vol. 9, no. 2, pp. 135–144, Jun. 2018.
- [12] F. Fioranelli, M. Ritchie, and Hugh Griffiths, "Aspect angle dependence and multistatic data fusion for micro-Doppler classification of armed/unarmed personnel," *IET Radar, Sonar Navigat.*, vol. 9, no. 9, pp. 1231–1239, 2015.
- [13] Z. Chen, G. Li, F. Fioranelli, and H. Griffiths, "Personnel recognition and gait classification based on multistatic micro-Doppler signatures using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 669–673, May 2018.
- [14] S. Ioffe, and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Lille, France, 2015, pp. 448–456.
- [15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, 2016, pp. 770–778.
- [16] F. Chollet, Keras [Online]. Available: <https://github.com/fchollet/keras>
- [17] D. P. Kingma, and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015, pp. 1–15.
- [18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.