AUTOMATIC RADAR-BASED GESTURE DETECTION AND CLASSIFICATION VIA A REGION-BASED DEEP CONVOLUTIONAL NEURAL NETWORK

Yuliang Sun[⋆†] *Tai Fei*[⋆] *Shangyin Gao*[⋆] *Nils Pohl*[†]

[⋆] HELLA GmbH & Co. KGaA, Lippstadt, Germany
[†]Institute of Integrated Systems, Ruhr University Bochum, Bochum, Germany

ABSTRACT

In this paper, a region-based deep convolutional neural network (R-DCNN) is proposed to detect and classify gestures measured by a frequency-modulated continuous wave radar system. Micro-Doppler (μD) signatures of gestures are exploited, and the resulting spectrograms are fed into a neural network. We are the first to use the R-DCNN for radar-based gesture recognition, such that multiple gestures could be automatically detected and classified without manually clipping the data streams according to each hand movement in advance. Further, along with the μD signatures, we incorporate phase-difference information of received signals from an L-shaped antenna array to enhance the classification accuracy. Finally, the classification results show that the proposed network trained with spectrogram and phase-difference information can guarantee a promising performance for nine gestures.

***Index Terms*—** Faster-RCNN, FMCW radar, Gesture recognition, Micro-Doppler signature, Phase difference

1. INTRODUCTION

With the growing requirements for human-computer interface (HCI) [1], one of the emerging applications of radar sensors is recognizing human hand gesture. Unlike optical gesture recognition system, radar sensors are insensitive to the ambient light conditions; the electromagnetic waves can penetrate dielectric materials, which allows them to be embedded into devices. In addition, because of privacy-preserving reasons, radar sensors are preferable to cameras in many circumstances [2, 3]. Some research works [2–7] detected series of gestures by investigating the Doppler frequency modulation, which is called the micro-Doppler (μD) effect [8]. For example, we [3] extracted handcrafted μD features from spectrograms for gesture classification. Kim *et al.* [6] fed the spectrograms into a deep convolutional neural network, and achieved approx. 85 % classification accuracy for ten gestures. However, before they applied the trained classifier to recognize unknown gestures, they had to manually clip the data streams, such that only one single gesture is present in the clipped time slot. It means that those approaches could not automatically detect unknown gestures. Further, in most of the existing works, e.g., [3–7], the backscattered signals are received only by a single receive antenna, and the information about the direction angle of gestures have not been exploited.

In this paper, we adopt a region-based deep convolutional neural network (R-DCNN) [9–11] to simultaneously detect and classify gestures, which are measured by a 77 GHz frequency-modulated continuous wave (FMCW) radar for in-vehicle infotainment and driver monitoring systems. To the best of the authors' knowledge, the R-DCNN has not been applied to radar images for gesture recognition. In this spirit, our proposed algorithm could work fully automatic without the necessity to manually detect and clip the data streams according to each gesture as an intermediate step. In addition to μD signatures, we incorporate the phase-difference information via an L-shaped receive antenna array to enhance the classification accuracy. Accordingly, the input of our proposed network contains three channels, i.e., one spectrogram and two phase-difference channels.

The remainder of this paper is organized as follows. Section 2 introduces the FMCW radar system. Section 3 describes the spectrogram as well as the phase-difference channels. The network architecture is explained in Section 4. Experimental results using real data are presented in Section 5. Finally, conclusions are given in Section 6.

2. FMCW RADAR SYSTEM

Our 77 GHz radar system adopts the linear chirp sequence frequency modulation [12] to design the waveform. After mixing, filtering and sampling, the discrete beat signal consisting of I reflecting points of objects for K measurement-cycles from one receive antenna can be approximated as:

$$\begin{aligned} b_k(n,m) \approx \sum_{i=1}^{I} A_i \exp\left\{j2\pi\left(f_{ri}(k)nT_s - f_{Di}(k)mT_c\right)\right\}, \\ n = 0, \cdots, N-1, \quad m = 0, \cdots, M-1, \\ k = 0, \cdots, K-1, \end{aligned} \tag{1}$$

where $f_{ri}(k)$ and $f_{Di}(k)$ represent the range and Doppler frequencies, respectively, as a function of measurement-cycle

index k, T_c is the chirp duration, the complex amplitude A_i contains the phase information, N is the number of sampling points in each chirp, M is the number of chirps in each measurement-cycle, and the sampling period $T_s = T_c/N$. Furthermore, as shown in Fig. 1, to calculate the phase differences of received signals, the spatial difference between two receive antennas in elavation and azimuth directions is $\lambda/2$, where λ is the wavelength.

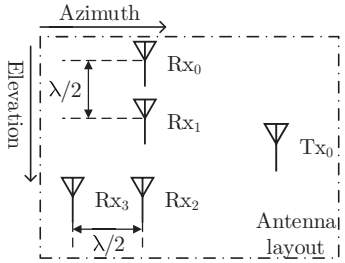

Fig. 1. Antenna layout with single transmit antenna Tx_0 and an L-shaped receive antenna array. $\{\text{Rx}_i : i = 0, 1, 2, 3\}$ denotes the i-th receive antenna. The pair of Rx_0 and Rx_1 is responsible for elevation angle, and the pair of Rx_2 and Rx_3 is used for azimuth angle calculation.

3. FRONT-END SIGNAL PROCESSING

3.1. Spectrogram Channel

A 2-dimensional finite Fourier transform (2-D FFT) is applied to process the beat signal in (1) for each measurement-cycle [13], such that the time-varying velocity information can be observed. The resulting 3-D range-Doppler-measurement-cycle array can be calculated as:

$$\begin{aligned} B(p,q,k) = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\{b_k(n,m)w(n,m)\} \\ \cdot \exp\left(-j2\pi\frac{pn}{N}\right)\cdot\exp\left(-j2\pi\frac{qm}{M}\right), \\ p = 0,\cdots,N-1, \quad q = 0,\cdots,M-1, \\ k = 0,\cdots,K-1, \end{aligned} \tag{2}$$

where $w(n,m)$ is the 2-D window function, p and q are the range and Doppler frequency indexes. Then, the spectrogram representing the μD signatures can be deduced by integrating $|B(p,q,k)|$ over range. It follows that

$$MD(q,k) = \sum_{p=0}^{N-1}|B(p,q,k)|, \tag{3}$$

and it shows the distribution of the reflected energy over velocity, as a function of measurement-cycle.

3.2. Phase-Difference Channels

Using two receive antennas that have a spatial difference of $\lambda/2$, the direction angle of an object could be estimated via the phase difference based on monopulse angle estimation principle [14]. For gesture recognition, we directly utilize the phase-difference information as a function of measurement-cycle, which contains the information of the direction angle of gestures. We first mitigate the noise influence in (2) for each cycle via the constant false alarm rate (CFAR) [15]. Then, the 3-D range-Doppler-measurement-cycle array of Rx_i can be rewritten as:

$$B_T^{(i)}(p,q,k) = \begin{cases} B^{(i)}(p,q,k), & |B^{(0)}(p,q,k)| \geq T, \\ 0, \quad \text{others}, \end{cases} \tag{4}$$

where T is the threshold obtained by the CFAR. Then, the phase difference between Rx_0 and Rx_1 in elevation and that of Rx_2 and Rx_3 in azimuth can be calculated as:

$$\begin{aligned} \Delta\psi^{(01)}(p,q,k) &= \psi\left\{B_T^{(0)}(p,q,k)\right\} - \psi\left\{B_T^{(1)}(p,q,k)\right\}, \\ \Delta\psi^{(23)}(p,q,k) &= \psi\left\{B_T^{(2)}(p,q,k)\right\} - \psi\left\{B_T^{(3)}(p,q,k)\right\}, \end{aligned} \tag{5}$$

where ψ stands for the phase of the complex value. Since the spectrogram channel shows reflected energy over velocity, as a function of measurement-cycle, we also project the phase differences $\Delta\psi^{(01)}$ and $\Delta\psi^{(23)}$ into velocity-measurement-cycle dimension. Then, the phase-difference channel of Rx_0 and Rx_1 and that of Rx_2 and Rx_3 are defined as:

$$\begin{aligned} PD^{(01)}(q,k) &= \sum_{p=0}^{N-1}|\Delta\psi^{(01)}(p,q,k)|, \\ PD^{(23)}(q,k) &= \sum_{p=0}^{N-1}|\Delta\psi^{(23)}(p,q,k)|. \end{aligned} \tag{6}$$

Hitherto, one spectrogram and two phase-difference channels are constructed as input of the proposed R-DCNN. The projection in (6) ensures that the values in both phase-difference channels have an 1-to-1 mapping relationship to the values in the spectrogram channel.

4. BACK-END SIGNAL PROCESSING

To design our radar-based gesture detection network, we follow the Faster R-CNN object detection framework [11]. It is able to detect objects with bounding boxes in red-green-blue (RGB) images and lidar point clouds for autonomous driving [16], and achieves excellent results on many popular object detection datasets, e.g., PASCAL VOC [17] and COCO [18] datasets. The network architecture of our proposed radar-based gesture detector is illustrated in Fig. 2a. Unlike object detection problems in RGB images [10,11], our network takes one spectrogram channel in (3) and two phase-difference channels in (6) as the input layer and feed them into a feature extraction network (FEN), which results into feature

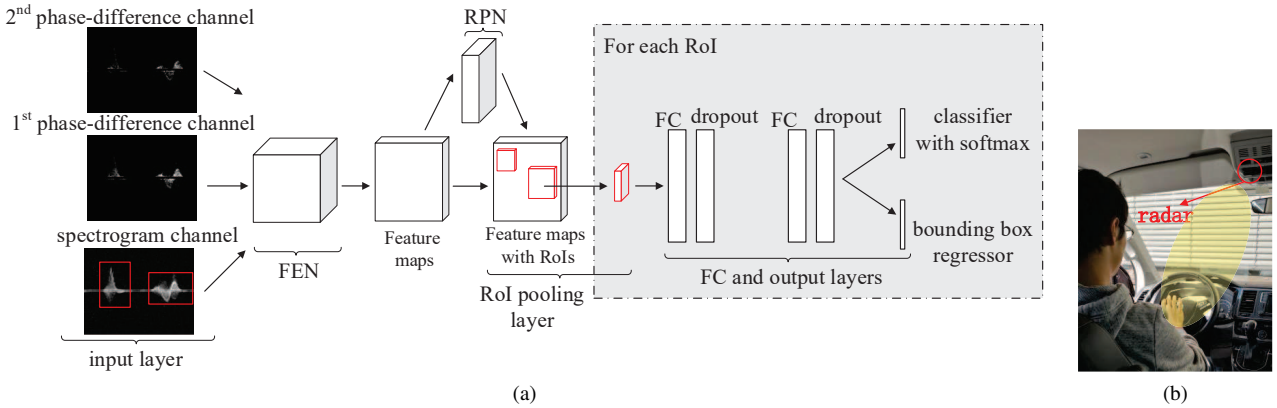

Fig. 2. (a) Network architecture of our radar-based gesture detector. (b) Radar mounting position and main beam direction.

maps. Then, a region proposal network (RPN) [11] is adopted to propose candidates of regions of interests (RoIs), which are further processed by the successive layers [10]. The output of the entire network in Fig. 2a provides us with the predicted classes and their bounding boxes positions.

4.1. Feature Extraction Network

To extract features from the spectrogram and phase-difference channels, an FEN is constructed in Fig. 3. We use 7 convolutional (Conv) layers, and each of them has a kernel size of 3×3. The kernel number of the first four Conv layers increases from 64, 128, 256 to 512, and that of Conv layer 5, 6, and 7 is 512. In each Conv layer, we use a rectified linear unit (RELU) [19] as the activation function. Besides, Conv layer 1, 2, 3 and 5 are followed by max-pooling layers with kernel size 2×2. The output of the FEN is the feature maps.

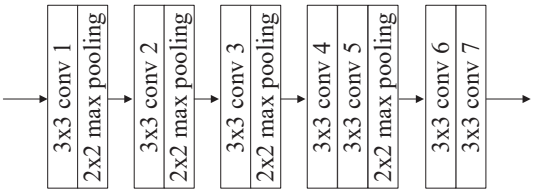

Fig. 3. Feature extraction network.

4.2. Region Proposal Network

We assume that the feature maps have a dimension of $W \times H \times 512$. In each pixel of the feature maps, we generate nine anchors using 3 scales of $8 \times 8, 16 \times 16, 32 \times 32$ and 3 aspect ratios of $1:2, 1:1, 2:1$ [11]. Then, among total $9WH$ possible anchors, the network could give several region proposals, i.e., RoIs, which are further processed by the following layers in the network.

4.3. RoI Pooling Layer

Using the region proposals acquired by the RPN, the relevant RoIs in feature maps are selected as input of the RoI pooling layer. For each RoI, the feature maps are cropped and then max-pooled to fixed-size feature maps because of size constraint in the following fully-connected (FC) layer.

4.4. FC and Output Layers

Each pooled RoI is then fed into two FC layers, either of which has 4096 hidden units and is followed by a dropout [20] layer for preventing the network from overfitting. For each RoI, the network gives two outputs using two separate output layers. The output layer followed by a softmax function gives the predicted class, and the other gives four values, which encode the bounding box position of the predicted class [10].

5. EXPERIMENTAL RESULTS

We used a 77 GHz FMCW radar mounted in the roof console of a vehicle to measure 9 classes of gestures performed by 19 human subjects for in-vehicle infotainment and driver monitoring systems. The radar has a detection range up to 3 m and an approx. 60° antenna beam width. Its mounting position and main beam direction can be seen in Fig. 2b. The 9 gestures are (a) approach steering wheel, (b) rotate clockwise, (c) rotate counter clockwise, (c) swipe from bottle right to upper left, (d) swipe down, (e) swipe left, (f) swipe right, (g) swipe up, and (i) random motion. Each subject repeated each gesture 10 times. Therefore, the total number of realizations is (9 gestures) $\times$ (19 people) $\times$ (10 times), namely 1710. In the evaluation scenario, we used the data of 15 subjects as training set, and the remaining 4 subjects are applied as test set. The network trained on known subjects is then transferred to classifying gestures from unknown humans [21].

The network was trained for 30000 iterations based on the back propagation algorithm [22] using the Adam optimizer [23] with an initial learning rate of 0.0001, which degraded by 10% after 20000 iterations. The batch size is 128.

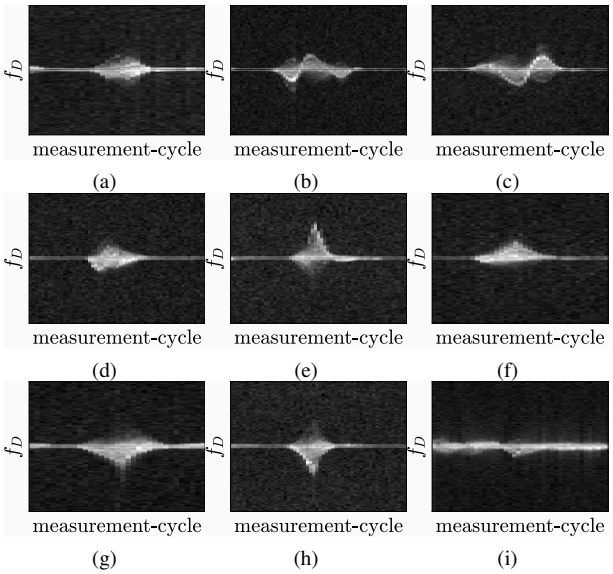

Fig. 4. (a) Approach steering wheel. (b) Rotate clockwise. (c) Rotate counter clockwise. (d) Swipe from bottle right to upper left. (e) Swipe down. (f) Swipe left. (g) Swipe right. (h) Swipe up. (i) Random motion.

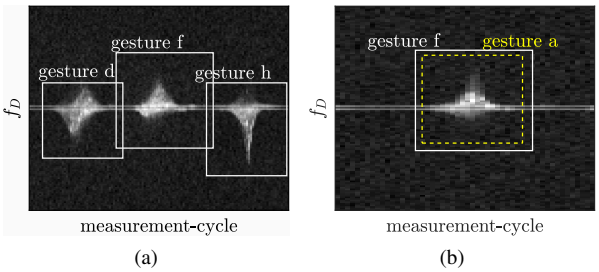

Fig. 5. (a) Three gestures are correctly detected. (b) Two bounding boxes are given to a single gesture, and that in dashed-line is false-detected.

5.1. Performance Evaluation

Fig. 4 shows the spectrograms of 9 gestures measured by the radar. Fig. 5a gives a detection example where three gestures are detected by the proposed network using three bounding boxes without manually clipping the data streams.

We define the positive predictive value (PPV) and true positive rate (TPR) as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, respectively, where TP, FP, FN denote the number of true positive, false positive, and false negative estimates. In the test set, we have 40 realizations for each gesture. It means that $\mathrm{TP}+\mathrm{FN}=40$. The PPVs and TPRs acquired by the network trained with three channels (i.e., one spectrogram and two phase-difference channels) are presented in Table 1. As a benchmark for comparison, the PPVs and TPRs obtained by the network using only one spectrogram channel are given in Table 2. Note that the average PPV (TPR) is the average value of the PPV (TPR) across the nine classes. In both tables, the number of predicted bounding boxes (PBBs) for each ground truth class (the true PBBs) can be calculated as the summation of the values in each column (a) to (i), which is not always 40. As shown in Fig. 5b, two overlapping bounding boxes are given to a single gesture, and consequently the number of true PBBs for this ground truth gesture could be larger than 40. Moreover, the non-maximum suppression algorithm [24] in the networks ensures that both bounding boxes in Fig. 5b can not be assigned to the same class. Thus, no single gesture is recounted in performance evaluation. As shown in Table 1, due to the main beam direction of the radar and individuality of each subject, gesture (f) is sometimes confused with other gestures, such as (a), (b) and (e). Further, the proposed network trained with three channels in Table 1 achieved higher average PPV and TPR, namely 95% and 96%, and outperforms the network using only one spectrogram channel in Table 2, which reaches 90% (92%) average PPV (TPR).

Table 1. Results of the network with three channels

Predicted \ True PBBs	a	b	c	d	e	f	g	h	i	PPV	TPR
a	40	0	0	0	0	3	0	0	0	0.93	1
b	0	40	0	0	0	6	0	0	0	0.86	1
c	0	1	40	0	0	1	0	0	0	0.95	1
d	0	0	0	40	0	0	0	2	0	0.95	1
e	0	0	0	0	40	3	0	0	0	0.93	1
f	0	0	0	0	0	27	0	0	0	1	0.67
g	0	0	0	1	0	0	40	0	0	0.97	1
h	0	0	0	0	0	0	0	40	0	1	1
i	0	0	0	0	0	1	0	0	40	0.97	1
Average										0.95	0.96

Table 2. Results of the network with spectrogram channel

Predicted \ True PBBs	a	b	c	d	e	f	g	h	i	PPV	TPR
a	39	0	0	0	1	5	0	0	0	0.86	0.97
b	0	40	0	0	0	2	0	0	0	0.95	1
c	0	0	40	0	0	0	2	0	0	0.95	1
d	1	0	0	33	0	0	1	6	0	0.80	0.82
e	0	0	0	0	40	5	0	0	0	0.88	1
f	0	0	0	1	0	32	0	0	0	0.96	0.80
g	0	0	0	5	0	0	37	2	0	0.84	0.92
h	0	1	0	0	0	0	1	36	0	0.94	0.90
i	1	0	0	0	0	0	0	0	37	0.97	0.92
Average										0.90	0.92

6. CONCLUSIONS

An automatic radar-based gesture detector based on the R-DCNN is developed. It could detect gestures without manually clipping the data streams according to each gesture in advance. In addition to the spectrogram channel, we incorporate the phase-difference information via an L-shaped receive antenna array to enhance the detection performance. The experimental results show that our proposed network could achieve 95% (96%) average PPV (TPR) for nine gestures.

7. REFERENCES

[1] S. Mitra and T. Acharya, "Gesture recognition: a survey," *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)*, vol. 37, no. 3, pp. 311-324, May. 2007.

[2] J. Lien *et al.*, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Trans. Graphics, vol. 35, no. 4, article 142*, July 2016.

[3] Y. Sun, T. Fei, F. Schliep, and N. Pohl, "Gesture classification with handcrafted micro-Doppler features using a FMCW radar," in *Proceedings IEEE MTT-S Int. Conf. Microw. for Intell. Mobility*, Munich, Germany, Apr. 2018.

[4] G. Li, R. Zhang, M. Ritchie, and H. Griffiths, "Sparsity-based dynamic hand gesture recognition using micro-Doppler signatures," in *Proceedings IEEE Radar Conf.*, Seattle, USA, June 2017.

[5] Z. Zhou, Z. Cao, and Y. Pi, "Dynamic gesture recognition with a terahertz radar based on range profile sequences and Doppler signatures," *Sensors*, vol. 18, no. 2, pp. 10, Dec. 2017.

[6] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-Doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125-7130, Nov. 2016.

[7] B. Dekker, S. Jacobs, A.S. Kossen, and M. Geurts, "Gesture recognition with a low power FMCW radar and a deep convolutional neural network," in *Proceedings IEEE Eur. Radar Conf.*, Nuremberg, Germany, Oct. 2017.

[8] V. C. Chen, F. Li, S. Ho, and H. Wechsler, "Micro-Doppler effect in radar: phenomenon, model, and simulation study," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 42, no. 1, pp. 2-21, Jan. 2006.

[9] V. Sze, Y. Chen, T. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295-2329, Dec. 2017.

[10] R. Girshick, "Fast R-CNN," in *Proceedings IEEE Int. Conf. Comput. Vision*, Santiago, Chile, Dec. 2015.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137-1149, June 2017.

[12] M. Kronauge and H. Rohling, "New chirp sequence radar waveform," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 4, pp. 2870-2877, Oct. 2014.

[13] Y. Sun, T. Fei, and N. Pohl, "Two-dimensional subspace-based model order selection methods for FMCW automotive radar systems," in *Proceedings Asia-Pacific Microw. Conf.*, Kyoto, Japan, Nov. 2018.

[14] S. Sharenson, "Angle estimation accuracy with a monopulse radar in the search mode," *IRE Trans. Aerosp. Navig. Electron*, vol. ANE-9, no. 3, pp. 175-179, Sep. 1962.

[15] H. Rohling, "Radar CFAR thresholding in clutter and multiple target situations," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-19, no. 4, pp. 608-621, Jul. 1983.

[16] D. Feng, L. Rosenbaum, and K. Dietmayer, "Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3D vehicle detection," in *Proceedings Int. Intell. Transp. Syst.*, Hawaii, USA, Nov. 2018.

[17] M. Everingham *et al.*, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vision*, vol. 111, no. 1, pp. 98-136, Jan. 2015.

[18] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proceedings Eur. Conf. Comput. Vision*, Zurich, Switzerland, 2014.

[19] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings Int. Mach. Learn.*, Haifa, Isreal, June 2010.

[20] 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.

[21] Y. Kim and H. Ling, "Human activity classification based on micro-Doppler signatures using a support vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 5, pp. 1328-1337, May 2009.

[22] Y. LeCun, L. Bottou, G. B. Orr, and K. Mueller, "Efficient backprop," *Neural networks: tricks of the trade*, 2nd ed. Springer, vol. 7700, pp. 9-48, 2012.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings Int. Learn. Representations*, San Diego, USA, May 2015.

[24] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proceedings Int. Conf. Pattern Recognition*, Hong Kong, China, Aug. 2006.