Transfer Learning from Audio Deep Learning Models for Micro-Doppler Activity Recognition

Kimberly T. Tran*†, Lewis D. Griffin†, Kevin Chetty*,
*Department of Security and Crime Science, †Department of Computer Science,
University College London, WC1E 6BT
{kimberly.tran, l.griffin, k.chetty}@ucl.ac.uk

***Abstract*—This paper presents a mechanism to transform radio micro-Doppler signatures into a pseudo-audio representation, which results in significant improvements in transfer learning from a deep learning model trained on audio. We also demonstrate that transfer learning from a deep learning model trained on audio is more effective than transfer learning from a model trained on images, which suggests machine learning methods used to analyse audio can be leveraged for micro-Doppler. Finally, we utilise an occlusion method to gain an insight into how the deep learning model interprets the micro-Doppler signatures and the subsequent pseudo-audio representations.**

***Keywords*—Micro-Doppler signatures, activity recognition, transfer learning.**

I. Introduction

The analysis of micro-Doppler ('μ-D') signatures produced from radio waves is becoming an increasingly viable technique for the task of activity recognition. Unlike video, which relies on optimal lighting conditions and can raise concerns with respect to privacy [1], or wearable technologies, where the richness of information is dependent on levels of compliance [2], radio waves are relatively unobtrusive and undetectable by humans. Furthermore, these signals can be obtained in a passive Wi-Fi setup, which is of low-cost to implement due to its ubiquitous nature.

In addition to the Doppler information that can be extracted from the Wi-Fi signals, the μ-D effect occurs when a target interacts with a signal and produces an additional vibration or rotation [3]. As the intensity of the μ-D effect is dependent on velocity and direction, each individual movement of a target will produce a distinct μ-D signature, which can be used for activity recognition [1].

Typically μ-D signatures are analysed in the frequency domain, often through the means of a spectrogram. This is defined as the squared amplitude of the short time Fourier transform of a signal. By using this representation, the motion features of a target can be identified [3]. A human operator can perform activity recognition by looking at the spectrograms directly or by listening to the audio output of the μ-D shift, but there is room for human error [4, 5] and the μ-D shifts may be so minute that they are not identifiable by the human observer.

The consistency of patterns for individual movements makes analysing μ-D signatures through deep learning methods an attractive proposition. However, this is restricted by the lack of sufficiently large and suitably annotated μ-D signature datasets. In contrast, deep learning classification methods for audio are relatively commonplace and have achieved promising results [6]. Outside of deep learning methods, similar signal processing techniques have been used to analyse audio and μ-D signatures, primarily as different categories within each medium can vary in time depending on the target generating the signal.

This paper aims to leverage the similarities between audio and radio and the consequent machine learning techniques used on audio. We achieve this by introducing a mechanism to pre-process the μ-D signatures into a pseudo-audio representation, so that transfer learning from a deep learning model trained on audio can be successfully conducted. We also discuss an occlusion method which is used to provide further insight into the mechanisms of the deep learning model and to understand which features of the μ-D signatures appear to be most important for classification.

II. Related Work

A. Speech Processing Methods for Micro-Doppler Classification

Previous works have adapted speech recognition techniques (such as linear predictive coding, mel frequency cepstral coefficients and dynamic time warping [4, 5, 7]) for μ-D signature classification. However, the assumptions underpinning these techniques can differ greatly from the physics of radar μ-D [8]. Moreover, manual feature engineering techniques can also be slow and consequently can impact learning on larger datasets. We counteract these issues by presenting a more automated mechanism to transform the μ-D signatures directly into a representation that resembles audio and by using deep learning to extract the features.

B. Transfer Learning

Convolutional neural networks ('CNNs') are capable of achieving state-of-the-art results in both image and audio classification [6]. CNNs are often used for tasks where the input data exhibits the properties of locality (where features in the data have a local spatial support) and translation invariance (where features in the data are independent of location). This makes spectrograms a particularly suitable candidate for training through CNNs, as μ-D signatures corresponding to

different actions should be equivariant to shifts in time and frequency [9].

However, the size of labelled μ-D datasets tends to be insufficient for suitable training through a deep CNN. This can be rectified through transfer learning, where the initial learnt weights from a CNN trained on one task are extracted and frozen, and only the later layers are fine tuned [10]. The effectiveness of this knowledge transfer is dependent on the extent of similarity between the two tasks. For μ-D signature classification, only transfer learning from CNNs trained on images has been attempted [11, 12]. This paper would be the first to conduct transfer learning from a CNN trained on audio for this task.

C. Image-to-Image Translation

Image-to-image ('I2I') translation is a computer vision problem involved with learning a mapping between an input and output image. A popular approach to this is to view I2I translation as a conditional image generation problem, whereby a bijective mapping between the input and output image is learnt. This is equivalent to generating one image, conditioned on another image in a different domain. DiscoGAN, CycleGAN and DualGAN [13–15] are all examples of this approach. They are very similar in construction as they use generative adversarial networks ('GANs') to generate the images, and a reconstruction loss to maintain a bijective relationship between the two domains.

A GAN is a type of generative model that attempts to learn the underlying distribution of the input data by pitting two neural networks against each other in a minimax game. Namely, a generative model G learns to capture the underlying distribution of data whilst a discriminative model D attempts to distinguish between legitimate data samples and samples synthesised by G [16]. Letting $\mathbf{x}$ represent the input data and $\mathbf{z}$ represent noise input into G, the objective function can be summarised as follows:

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_r(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_\mathbf{z}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

The I2I networks mentioned all use two generators to represent the conditional mappings learnt and two discriminators to distinguish between real samples in each domain and synthesised samples produced by the generators. Suppose that the marginal distributions pertaining to each domain are labelled A and B. If a bijective mapping is learnt, the generators G_{AB} and G_{BA} should be inverse functions. Therefore, the difference in the reconstruction loss $d(G_{BA} \circ G_{AB}(a), a)$ (where $d(.)$ is a chosen metric) should be minimised. Letting a represent a sample in A, b represent a sample in B and λ be a chosen weighting for the reconstruction loss, the losses relating to producing a mapping from $A \to B$ can be summarised as follows (with losses relating to $B \to A$ defined similarly):

$$L_{G_{AB}} = -\mathbb{E}_{a\sim A}[\log D_B(G_{AB}(a))] + \lambda d(G_{BA} \circ G_{AB}(a), a)$$
$$L_D = -\mathbb{E}_{b\sim B}[\log D_B(b)] - \mathbb{E}_{a\sim A}[\log(1 - D_B(G_{AB}(a)))]$$

We treat the μ-D signatures as images and use a I2I network to transform them into a pseudo-audio representation.

III. Pseudo-Audio Transfer Learning

A. Transfer Learning Model

The μ-D dataset used in the experiments was obtained from a passive Wi-Fi setup detailed in [12]. The dataset comprises of 1,109 spectrograms across six movement classes: taking a bow, performing a breast-stroke motion, performing a crawl-stroke motion, performing a double punch, sitting, and standing. Each spectrogram is 51×75, where 51 corresponds to the number of frequency bins and 75 corresponds to the time steps. The 51 frequency bins cover a frequency shift range from -12.5 Hz to $+12.5$ Hz and each sample is captured over a period of 5 seconds.

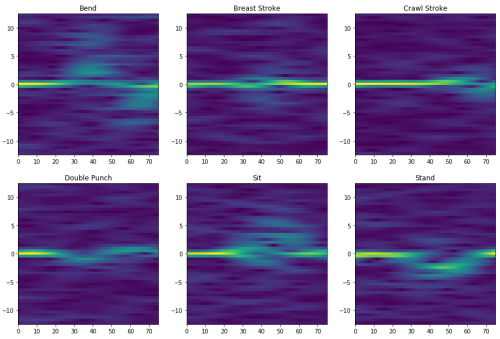


Fig. 1: Example spectrograms for each class.

For the transfer learning experiments, we chose to use a CNN trained on audio spectrograms. Using such a network has the benefit of achieving invariance to time and frequency and additionally would eliminate the requirement of converting the μ-D signatures into the time domain, which could result in loss of information.

The audio deep learning model chosen for the experiments was VGGish. This is a CNN trained on 70 million YouTube videos automatically labelled based on the metadata and image content for each video [6]. It is inspired by VGG [17], which comprises of stacks of 2D convolutional layers with increasing filters followed by max pooling layers, and three fully connected layers for classification. To appropriately process the audio clips for training, the audio from each video clip was extracted and divided into 960ms frames and transformed into log-mel spectrogram form. Each sample is 96×64, with 96 corresponding to the time steps and 64 corresponding to the number of mel-spaced frequency bins, meaning the time and frequency axes are transposed compared to the μ-D dataset.

The majority of VGGish's original architecture was retained for our experiments. All of the weights in VGGish were frozen, except for the fully connected layers which were adjusted for fine tuning. As the μ-D dataset is significantly smaller than the number of clips used to train VGGish, the two penultimate layers were reduced to 64 units and the final layer was fixed to 6 units. We also used a softmax activation function in place of the sigmoid for the final layer as each of the μ-D

signatures were labelled with one class only. Furthermore, we used random search to set batch size to 10, select the Adam optimiser and set training time to 20 epochs.

B. Pseudo-Audio Representation

Pre-processing the μ-D signatures to resemble the audio spectrograms used to train VGGish was a necessary step to ensure successful knowledge transfer. An additional consideration was to ensure the pre-processing did not remove the distinct μ-D features corresponding to each class, hence the I2I model needed to incorporate this constraint.

After rotating the axes of the μ-D signatures to match the input data for VGGish, we used a modified I2I network inspired by [13–15] for pre-processing. This architecture, labelled as 'DSGAN', used two generators to learn a bijective mapping between the presented μ-DS and audio log-mel spectrograms. To ensure a bijective relationship is maintained, we incorporated an L2 reconstruction loss. This is chosen to be weighted equally with the generator loss as this preserved more of the μ-D features in this particular dataset. We experimented with both L1 and L2 reconstruction losses, but found the L2 loss was better in retaining more of the μ-D features. We attribute this to the skewed nature of the dataset which is caused primarily by the presence of high amplitudes around zero Doppler. However, instead of using two separate discriminators to distinguish between real and synthesised samples for each domain, only one discriminator is used to distinguish between real and synthesised samples.

The discriminator's parameters were updated iteratively by presenting real and synthesised samples from the μ-D domain followed by presenting real and synthesised samples from sound domain. As the discriminator has the task of distinguishing synthesised samples from real samples from both domains, the generators should be able to deceive the discriminator by synthesising samples that are an intermediate representation between the two domains. Consequently this setup enables more of the features from the source domain to be retained.

The architecture of the two generators are identical, utilising an encoder-decoder structure similar to the implementations employed in the established I2I networks [13–15]. The encoder component is comprised of upsampled 2D convolutional layers, whilst the decoder component is comprised of downsampled 2D convolutional layers to resize the input back to the original image dimensions. Each layer uses 'same' padding, 3×3 kernels, and is activated by a ReLU activation function. Batch normalisation is applied to all layers except for the input layers. The discriminator is similarly constructed, using a stack of upsampled 2D convolutional layers. A 2D convolutional layer with one filter activated with a sigmoid function is appended at the end of the discriminator to provide a probability estimate between real and synthesised samples.

Both the μ-D signatures and sound data were resized to an even dimensionality so that the encoder-decoder structure could generate outputs that were the same size. This was conducted so that the discriminator could compare between the real and synthesised samples. This does not appear to significantly affect the quality of features, as results matching the baselines were achieved.

The sound domain dataset used for the experiments were log-mel spectrograms sampled from AudioSet [18]. This is a dataset consisting of over 2 million manually labelled 10-second sound clips from 632 audio event classes drawn from YouTube videos. DSGAN was trained by separately shuffling the μ-D data and Audioset data and grouping the random pairs for each batch fed into the model. By randomly pairing the datasets, this ensured that a more general relationship is learnt between the domains rather than collapsing to one representation.

As the μ-D dataset is small, the model was trained for varying number of epochs (1, 5, 10 and 30) to observe how the extent of training affected the learnt mapping. Following training, the μ-D data was transformed into pseudo-audio representations by feeding the entire μ-D dataset through the generator that was used to learn the mapping to sound data. These pseudo-audio representations were used to fine tune VGGish. We measured indicative performance by using five-fold stratified cross-validation to record validation accuracy. As a benchmark, we compared this to results of a one-vs-one linear support vector machine ('SVM'); pseudo-audio representations learnt by other I2I networks [13–15]; μ-DS whose mean and standard deviation were shifted to fit the Audioset samples; and VGGish fine-tuned using the unprocessed μ-D signatures. Moreover, to investigate whether transfer learning through audio was more effective than training through other modalities, we also recorded the cross-validation results following fine-tuning of VGG-16 [17], which is trained on images.

Summarising the setup of DSGAN, let R be the domain of the μ-D signatures, $r \in R$ represent a sample μ-D signature, S be the domain of the Audioset sound samples and $s \in S$ represent a sound sample, G_{RS} represent a generator learning a mapping from R to S (with G_{SR} defined similarly), and D represent the discriminator. Consequently the objective functions for each component can be described as follows:

$$\begin{aligned} L_{G_{RS}} &= -\mathbb{E}_{r\sim R}[\log D_S(G_{RS}(r))] \\ &\quad + \mathbb{E}_{r\sim R}[||G_{SR}(G_{RS}(r)) - r||_2] \end{aligned}$$

$$\begin{aligned} L_{G_{SR}} &= -\mathbb{E}_{s\sim S}[\log D_R(G_{SR}(s))] \\ &\quad + \mathbb{E}_{s\sim S}[||G_{RS}(G_{SR}(s)) - s||_2] \end{aligned}$$

$$\begin{aligned} L_D &= -\mathbb{E}_{r\sim R}[\log D_R(r) + \log(1 - D_S(G_{RS}(s)))] \\ &\quad - \mathbb{E}_{s\sim S}[\log D_S(s) + \log(1 - D_R(G_{SR}(r)))] \end{aligned}$$

C. Results and Discussion

Although pre-processing the μ-D signatures using an I2I network before fine tuning matched the results of both the SVM and pre-processing by rescaling to the sound data's

Model	Acc (%)	SD
SVM	83.54	± 4.04
Transfer learning using VGG-16 (no pre-processing)	43.01	± 11.45
Transfer learning using VGG-16 (with z-score standardisation)	68.87	± 4.59
Transfer learning using VGGish (no pre-processing)	57.61	± 16.96
Transfer learning using VGGish (rescale to sound data's mean and SD)	84.40	± 0.60

TABLE I: Validation accuracies for the baseline models.

	1 epoch		5 epochs		10 epochs		30 epochs	
Model	**Acc (%)**	**SD**	**Acc (%)**	**SD**	**Acc (%)**	**SD**	**Acc (%)**	**SD**
DSGAN	83.04	± 3.59	84.66	± 2.85	81.69	± 2.09	81.07	± 1.69
DiscoGAN (L1 loss)	78.00	± 3.05	71.31	± 3.94	79.00	± 1.79	49.94	± 3.90
DiscoGAN (L2 loss)	83.32	± 3.06	84.12	± 2.20	81.32	± 3.19	76.92	± 3.12
CycleGAN	75.13	± 3.32	74.40	± 3.53	66.29	± 3.57	61.77	± 1.63
CycleGAN (with identity loss)	76.65	± 1.71	49.33	± 2.38	77.37	± 1.10	60.42	± 1.09
DualGAN	75.20	± 2.01	55.36	± 1.82	75.11	± 2.83	56.43	± 3.02

TABLE II: Transfer learning validation accuracies after pre-processing the μ-DS using I2I networks. Default hyperparameters were implemented for all of the other I2I networks.

statistics, they performed significantly better than transfer learning without any pre-processing. This could be explained be due to the nature of the dataset itself. From the results of the SVM, it appears that the features in the dataset are primarily linearly separable but the dataset's quality could be constrained by the similarity of features between different classes. Nonetheless, the results highlight that pre-processing of the μ-D data is required to achieve successful transfer learning.

The variations in performance of the baseline models trained on unprocessed data suggest that the the transfer learning performance may primarily be affected by backpropagation issues due to the differences in amplitudes between the μ-D signatures and sound data rather than differences in the modalities themselves. This is evidenced by significant improvements to the training caused by shifting the μ-D signatures to match the sound data, which retains the distribution of intensities across each spectrogram. This suggests that μ-D signatures share more low-level features with audio spectrograms compared to other modalities.

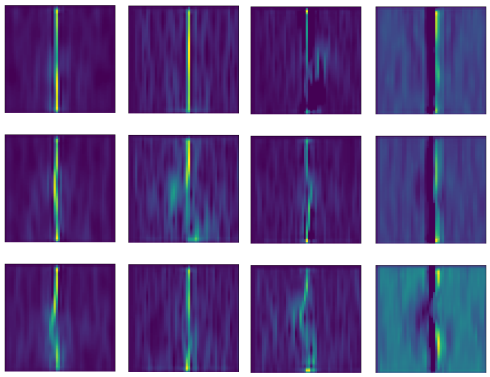


Fig. 2: Examples of a μ-D signature transformed by DSGAN (Left to right: examples after training for 1, 5, 10 and 30 epochs.)

The most consistent and best performing I2I networks for pre-processing the data appear to be DSGAN and DiscoGAN with L2 loss. The resultant spectrograms seem to retain the same amplitude intensities as the original probe data, but scaled down in a manner that better matches the reference sound data. The DualGAN paper states that L1 loss is preferred for I2I translation as it produces less blurry results [15]. However, in this circumstance, the presence of large amplitudes around zero Doppler causes the overall μ-D amplitudes to be significantly more skewed than the sound data. As the L2 loss is more sensitive to outliers, it can retain more of the features around zero Doppler. Further visual inspections show that I2I networks which use an L1 loss appear to remove the zero Doppler bin to make the spectrograms appear more like the sound data, which causes a decrease in classification accuracy. The difference in performance could also be attributed to placing an equal weighting on the generator and reconstruction loss, as opposed to placing a heavier weight on the reconstruction loss which used for CycleGAN and DUalGAN. Arguably the heavier weight makes the bijective constraint stricter, and the synthesised images look more convincingly like sound spectrograms, but this is at a cost of eliminating the features surrounding zero Doppler.

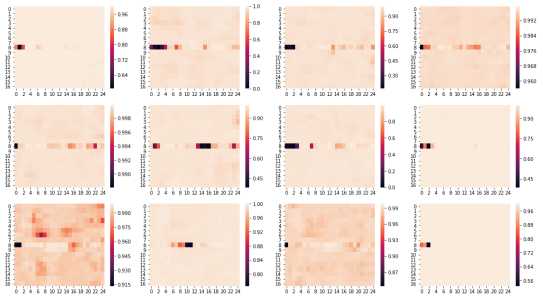
When trained for longer, DSGAN achieves a higher validation accuracy in comparison to DiscoGAN. This could be attributed to DSGAN's discriminator's participation in a more difficult minimax game which produces more intermediate representations that retain more of the source domain's features as opposed to DiscoGAN.

IV. Occlusion Experiment

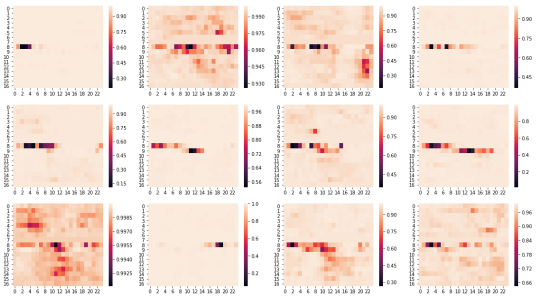
As popular CNN implementations tend to be deep, they are often treated as black box models. To improve our understanding of VGGish and to verify whether VGGish focuses on the μ-D signatures whilst training, an occlusion experiment was performed. This is a technique introduced in [19], which involves the creation of heatmaps to illustrate how prediction confidence is affected depending on which patches of the input data are occluded.

VGGish was initially fine tuned using pre-processed μ-D spectrograms without any occlusions. 3×3 patches of the spectrograms were then occluded by setting the value of the patch to be all zero. The modified spectrogram was then evaluated through VGGish. This process was repeated by moving patches across the spectrogram. The data used to fine tune VGGish was used as this would be a more accurate depiction of the features VGGish had learnt for each class. The output of the softmax belonging to the ground truth - the level of confidence that the spectrogram belonged to the correct class - was recorded for each iteration. These outputs were combined to create a 16×24 heatmap for each spectrogram.

As the large amplitudes of the unprocessed μ-D data made fine tuning VGGish more difficult, the μ-D data was pre-processed through an I2I network trained for 5 epochs and manually by matching the μ-D data's mean and standard deviation to the sampled sound data's mean and standard deviation.



(a) Data transformed by matching mean and SD to the sound data's mean and SD.



(b) Data transformed using I2I network.

Fig. 3: Heatmaps illustrating the effects of occluding different patches of one spectrogram belonging to the double punch class. Darker patches signify a decline in prediction confidence, illustrating the areas of the spectrogram which are more important for prediction. These heatmaps are a rotation of the input data used to train VGGish, as this representation illustrates the effects on classification with respect to movement position more clearly.

The main observation from both pre-processing methods is that the frequency bins around zero Doppler contained the most distinct features between classes. However, this focus around zero Doppler is less prominent for the I2I transformed data, which confirms Doppler shifts further away from zero Doppler are also important for classification. The difference in heatmaps between the I2I transformed and the manually transformed data could be attributed to the I2I network learning a non-linear transformation to transform the μ-D data into a form that resembles audio. Although from visual inspection, there does not appear to be a drastic alteration in amplitude intensities, the heatmaps indicate there is a subtle change in relative amplitude of certain frequencies.

Nonetheless, for each class there seems to be a unique pattern to the results. For example, for the double punch movement, prediction confidence tends to drop in the initial few timesteps by occluding adjacent patches near zero Doppler. For this class, the I2I network transformed data is also affected by a negative frequency shift away from zero Doppler. Although one could interpret this as a temporal component, the confidence drop does not always happen at the same time step. As CNNs are translation invariant (and so should not be influenced by changes in time), this is a demonstration of VGGish being influenced more by the change in amplitudes rather than position in the spectrogram. Furthermore, any supposed temporal factor could be explained by experimental protocol to collect the data.

V. Conclusion

The results show that provided the μ-D spectrograms were transformed into a pseudo-audio representation, transfer learning from a deep learning model trained on audio was more effective than deep learning models trained on other modalities. In this paper we introduce DSGAN, which is able to transform the μ-D spectrograms into a representation more reminiscent of audio whilst preserving the subtleties of μ-D signatures. This is achieved through formulating a more difficult minimax game for the discriminator in the architecture. Visual evaluation of the transformed data indicated that the best performing pre-processing methods preserved the amplitude intensities of the original unprocessed data, in particular the amplitudes around zero Doppler. This suggests that μ-D spectrograms are already akin to audio spectrograms and indicate that deep learning methods applied to audio may be of benefit for μ-D classification.

The occlusion experiment indicate that the I2I networks appear to perform a nonlinear transformation of the μ-D data. This change in amplitude intensities does not appear to be discernible to the human eye, which raises the possibility that the nonlinear transformation learnt by the I2I networks may prove more useful in more complicated datasets.

Furthermore, the occlusion experiment illustrated that a predominant number of distinguishing features between classes are concentrated near zero Doppler which could be indicative of the μ-D signatures. This technique may be useful in locating μ-D features where they may not be explicitly observed.

References

[1] Chen, Q. et al., 2016. Activity recognition based on micro-Doppler signature with in-home Wi-Fi. *In IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*. Munich, Germany, September 14 - 16, 2016.

[2] Tan, B. et al., 2017. Exploiting WiFi Channel State Information for Residential Healthcare Informatics. *IEEE Communications Magazine*, 56, 130-137.

[3] Chen, V. et al., 2005. Micro-Doppler Effect in Radar: Phenomenon, Model, and Simulation Study. *IEEE Transactions on Aerospace and Electronic Systems*, 42, 2-21.

[4] Smith, G. E., et al., 2010. Radar Micro-Doppler Signature Classification using Dynamic Time Warping. *IEEE Transactions on Aerospace and Electronic Systems*, 46 3, 1078-1096.

[5] Yessad, D. et al., 2011. Micro-Doppler Classification for Ground Surveillance Radar Using Speech Recognition Tools. *In CIARP: Iberoamerican Congress on Pattern Recognition*. Pucón, Chile. November 15 - 18, 2011.

[6] Hershey, S. et al. 2017. CNN Architectures for Large-Scale Audio Classification. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*. New Orleans, USA. March 5 - 9, 2017.

[7] Javier, R. J. & Kim, Y. 2014. Application of Linear Predictive Coding for Human Activity Classification Based on Micro-Doppler Signatures. *IEEE Geoscience and Remote Sensing Letters*, 11 10, 1831-1834.

[8] Erol, B. et al. 2018. Automatic Data-Driven Frequency Warped Cepstral Feature Design for Micro-Doppler Classification. *IEEE Transactions on Aerospace and Electronic Systems*, 54 4, 1724-1738.

[9] Goodfellow, I., et al., 2016, *Deep Learning*. 1st ed. Massachusetts: MIT Press.

[10] Yosinski, J. et al., 2014. How transferable are features in deep neural networks? *In Advances in Neural Information Processing Systems*. Montreal, Canada, December 8 - 13, 2014.

[11] Khanna, R. et al. 2019. Through-Wall Remote Human Voice Recognition Using Doppler Radar With Transfer Learning. *IEEE Sensors Journal*. 19 12. 4571-4576.

[12] Shi, F. et al. 2019. Passive Activity Classification Using Just WiFi Probe Response Signals. *In IEEE Radar Conference*. Boston, MA, USA. April 22 - 26, 2019.

[13] Kim, T. et al. 2017. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. *In International Conference on Machine Learning*. Sydney, Australia. August 6 - 11, 2017.

[14] Zhu, J.Y. et al. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *In International Conference on Computer Vision*. Venice, Italy. October 22 - 29, 2017.

[15] Yi, Z. et al. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. *In International Conference on Computer Vision*. Venice, Italy. October 22 - 29, 2017.

[16] Goodfellow, I. et al., 2014. Generative Adversarial Nets. *In Advances in Neural Information Processing Systems*. Montreal, Canada, December 8 - 13, 2014.

[17] Simonyan, K. & Zisserman, A. 2015. Very Deep Convolutional Networks for Large Scale Image Recognition. *In International Conference on Learning Representations*. San Diego, USA, May 7 - 9 2015.

[18] Gemmeke, J. et al. 2017. Audio Set: An ontology and human-labeled dataset for audio events. *IEEE International Conference on Acoustics, Speech and Signal Processing*. New Orleans, USA. March 5 - 9, 2017.

[19] Zeiler, M. D. & Fergus, R. 2013. Visualizing and Understanding Convolutional Networks. *In European Conference on Computer Vision*. Zurich, Switzerland. September 6 - 12 2014.