Motion Classification Using Kinematically Sifted ACGAN-Synthesized Radar Micro-Doppler Signatures

BARIS EROL, Member, IEEE
Siemens Corporate Technology, Munich, Germany

SEVGI ZUBYEDE GURBUZ, Senior Member, IEEE
University of Alabama, Tuscaloosa, AL, USA

MOENESS G. AMIN, Fellow, IEEE
Villanova University, PA, USA

Deep neural networks have recently received a great deal of attention in applications requiring classification of radar returns, including radar-based human activity recognition for security, smart homes, assisted living, and biomedicine. However, acquiring a sufficiently large training dataset remains a daunting task due to the high human costs and resources required for radar data collection. In this article, an extended approach to adversarial learning is proposed for generation of synthetic radar micro-Doppler signatures that are well adapted to different environments. The synthetic data are evaluated using visual interpretation, analysis of kinematic consistency, data diversity, dimensions of the latent space, and saliency maps. A principle-component analysis-based kinematic-sifting algorithm is introduced to ensure that synthetic signatures are consistent with physically possible human motions. The synthetic dataset is used to train a 19-layer deep convolutional neural network to classify micro-Doppler signatures acquired from an environment different from that of the dataset supplied to the adversarial network. An overall accuracy of 93% is achieved on a dataset that contains multiple aspect angles (0°, 30°, and 45° as well as 60°), with 9% improvement as a result of kinematic sifting.

Manuscript received March 16, 2019; revised July 13, 2019 and November 23, 2019; released for publication January 6, 2020. Date of publication January 27, 2020; date of current version August 7, 2020.

DOI. No. 10.1109/TAES.2020.2969579

Refereeing of this contribution was handled by S. S. Ram.

Authors' addresses: B. Erol was with the Center for Advanced Communications, Villanova University, Villanova, PA 19085 USA. He is now with the Siemens Corporate Technology, 81739 Munich, Germany, E-mail: (baris.erol@siemens.com); S. Z. Gurbuz is with the Department of Electrical and Computer Engineering, University of Alabama, Tuscaloosa, AL 30332 USA, E-mail: (szgurbuz@ua.edu); M. G. Amin is with the Center for Advanced Communications, Department of Electrical and Computer Engineering, Villanova University, PA 19085 USA, E-mail: (moeness.amin@villanova.edu). *(Corresponding author: Sevgi Gurbuz.)*

0018-9251 © 2020 IEEE

I. INTRODUCTION

Over the past decade, radio frequency (RF) sensing has gained increased attention as its efficacy and unique advantages have been demonstrated for a variety of automotive, smart home, human computer interaction, and remote health monitoring applications [1]–[8]. Radar systems are both low cost and low power, making them a safe sensing alternative, which can operate in darkness and all weather conditions. Moreover, radar is noninvasive, and when used for monitoring, does not require an alteration in daily habits or routines. These attributes have made RF sensing popular in motion monitoring.

Meanwhile, progress in machine learning and Internet of Things is rapidly growing the expectations and performance requirements of ubiquitous sensing. Radar-based gesture recognition for man-machine interfaces requires an ability to recognize slight differences in hand motions, separating gestures intended to give commands versus daily hand movements [9]. Biomedical applications of abnormal gait analysis, fall detection, fall risk assessment, and monitoring of hip/knee operations or neuro-muscular disorders, also require high sensitivity and specificity, consistent with medical standards [10]–[13]. Thus, even slight increases in accuracy and robustness are considered significant in the advancement of indoor radar technology and its adoption in smart homes and medical diagnosis.

Deep neural networks (DNNs) have shown great potential to achieve high accuracy, even as the number of classes increases, and may well lead the way as a preferred method for motion classification in the near future [14]–[19]. However, DNN architectures in RF applications are often limited by the fact that only small datasets are available for training, as data acquisition can be time consuming, costly, and limited in terms of the scope of scenarios and targets sampled. This impacts not only DNN depth, but also the ability of the DNN to generalize across different body types, speeds, and motion classes [20], as well as adapt to different noise sources and environmental conditions.

Researchers have attempted to overcome this challenge by data augmentation, where the available radar data is modified through operations such as translation, time shifting, and segmentation [21], [22]. However, in RF applications, these approaches may not necessarily lead to statistically independent training samples that effectively span probable variations in target signatures. This is because the pixel values in the 2-D data representations, generated through time frequency (TF), analysis are related to the complex electromagnetic scattering and kinematics of the dynamic target being observed. Radar returns from a moving target include not only a central Doppler shift, resulting from translational motion, but also micro-Doppler frequencies induced by slight rotations or vibrations of parts of the target [23], [24]. In humans, micro-Doppler frequencies derive from the unique, bipedal, time-varying kinematics of human motion, and varies even for the same activity depending upon body size, speed, and individual gait style. Thus, methods for data augmentation motivated by image processing applications, such as scaling and rotating, may significantly disrupt RF data patterns by generating samples that are kinematically untenable. The inclusion of such physically impossible samples in the training data has adverse effects, and compromises rather than improves performance.

To overcome these limitations, a simulation methodology rooted in kinematic modeling [25]–[27] via motion capture was recently proposed in [28]. Instead of applying pixel-based data augmentation, transformations to the underlying skeletal model were applied to generate a large number of unique but kinematically consistent micro-Doppler signatures spanning expected target profiles. A key disadvantage, however, is that the approach does not provide a means to account for the variations in signal-to-noise ratio, artifacts of sensor-related electronic interference, signal dispersion caused by frequency dependent obstructions, like walls, or nontarget related motion (e.g., spinning ceiling fan).

Generative adversarial networks (GANs) have been proposed for synthesizing realistic images in a variety of applications [29], including synthetic aperture radar [30]. An early effort at applying adversarial learning to synthetic data generation of micro-Doppler was first proposed in 2018 [31], in which a deep convolutional GAN (DCGAN) was used to generate synthetic data that emulated the Boulic walking model. The Boulic model consists of mathematically well-defined trajectories and, therefore, does not represent the spectral richness and intricacy of actual, measured micro-Doppler signatures. By using such pristine and systematic simulated data to drive the DCGAN, replicas of the data were easily generated and nearly identical.

In another study [32], 150 simulated spectrograms were augmented with 1000 GAN-generated spectrograms to classify a test set of 50 simulated signatures comprised of three activity classes: running, walking, and jumping. A 4% increase in classification accuracy was noted. However, only a small number of samples were generated by the GAN, and the classes considered are easily identifiable so that the simulation study was not designed to vet the validity, merits, or detriments of using GANs to simulate micro-Doppler signatures.

The first study exploiting adversarial learning for the classification of real micro-Doppler data was published in [20], where Yang *et al.* evaluated the efficacy of adversarial learning for addressing the open-set problem—the case where the training dataset does not include all the classes as the test dataset. Subsequent studies in [33] and [34] utilized GANs for mitigating the problem of low sample support and reported the classification accuracy of DNNs trained with GAN-generated synthetic data for human activity recognition.

In fact, the ability of GANs to synthesize authentic radar micro-Doppler signatures is hampered by differences between radar phenomenology and optics. The values of pixels in micro-Doppler signatures relate not to physical shapes, but instead to human kinematics. It is, thus, possible for GANs to generate numerous synthetic samples that, while visually similar, are incompatible with the kinematics of human motion.

The work in this article is developed concurrently with that of [33], [34] and provides, to our knowledge, the first in-depth analysis of GAN-generated synthetic data in terms of kinematic fidelity and diversity. In particular, we propose the utilization of auxiliary classifier generative adversarial networks (ACGANs) [35], as opposed to conditional variational autoencoders (CVAEs) [36], for the generation of synthetic micro-Doppler signatures with greater diversity and sharpness. The issues of kinematic fidelity of the ACGAN-generated synthetic data are illustrated using physics-based rules applied to walking and falling motion classes. The relationship between kinematic fidelity and the dimensionality of the latent space as well as sample diversity is also examined. We propose a new technique for kinematic sifting based on principal component analysis (PCA) to eliminate the kinematically impossible samples from the synthetic training dataset and, as such, limit their corrupting effects on performance. This underlies the importance of considering kinematics when generating synthetic micro-Doppler signatures using adversarial learning. The proposed technique achieves a 9% improvement in performance over that attained if the ACGAN-generated signatures are used directly for training, without kinematic sifting.

Finally, we show the benefits of ACGAN-generated synthetic data to adaptation to different sensing locations and environments. A small number of measured radar data collected from one location (with multiple aspect angles: 0°, 30°, 45°, and 60°) is used by ACGAN to grow the synthetic dataset for training, while the test dataset is collected at a different location in a through-the-wall configuration. A 19-layer convolutional neural network (CNN) trained using the kinematically sifted data generated by the ACGAN is shown to yield a high 93% classification accuracy across different environments.

This article is organized as follows. In Section II, the experimental radar measurements conducted in two distinct locations and environments is described. In Section III, the generative model, ACGAN, is discussed in relation to an alternative generative model, CVAE. In Section IV, diversity, accuracy, and kinematic fidelity of the ACGAN-generated synthetic images are evaluated. In Section V, classification with PCA-based kinematic sifting of ACGAN-generated synthetic data for training a 19-layer CNN in a scenario involving adaptation across two distinct environments is presented. Discussion of the conclusion and future work is provided in Section VI.

II. RADAR MICRO-DOPPLER MEASUREMENTS

Commercially available continuous-wave radars are compact in size and provide a measurement of Doppler frequencies as a function of time. The radar system used in this article operates at a transmitting frequency of 25 GHz, sweep time of 10 ms, while collecting 128 samples per sweep. Thus, the received radar signal is highly oversampled at a rate of 12.8 kHz [37]. A higher sampling frequency causes the spectrum to shrink and cluster around the origin, leaving considerable vacant space in the time-frequency domain. Therefore, during preprocessing, the received radar signal is first downsampled to 1.2 kHz. The power output and antenna gain of the radar are 16 dBm and 18 dBi, respectively.

A. Time-Frequency Representation: Spectrograms

Human activity recognition is typically accomplished through identification of unique patterns in the radar micro-Doppler signature, a time-frequency representation of the radar received signal [38]. Quadratic time-frequency distributions are considered a powerful tool for the analysis of time-varying signals, with spectrograms being the simplest and most commonly used TF distribution [39]. Spectrograms are the energetic form of the short-time Fourier transform (STFT), which is obtained by splitting the time domain signal into many overlapping or disjoint consecutive segments, and then taking the Fourier transform (FT) of each segment. A spectrogram thus exposes the signal's local frequency behavior and is mathematically defined as

$$\mathrm{S}(n,k) = \left| \sum_{m=0}^{N-1} \mathrm{h}(m)\mathrm{x}(n-m)e^{-j2\pi km/N} \right|^2 \tag{1}$$

where $\mathrm{h}(m)$ is a window function, which can affect both the time and frequency resolutions. The window slides over the data to capture the instantaneous frequencies. The amount of overlap is variable, so that the window could slides one or more samples each time. At each window time-position, the local frequency behavior is emphasized through the windowed FT. The window length trades off spectral and temporal resolutions, with long windows providing high frequency resolution, whereas short windows offer high temporal resolution.

Optimal sampling frequency and STFT parameters can be found using a grid search; however, this might not lead to a global optimum since the parameter step size is determined manually. Therefore, we used data-driven optimization with genetic algorithms (GA) to determine the optimum hyperparameters of the STFT and the sampling frequency, while maximizing the classification performance achieved by generalized PCA (GPCA) and minimum distance classifier (MDC). For one set of hyperparameters, GPCA is used to reduce the dimensionality and extract features in time and frequency, which are subsequently provided to the MDC. Classification accuracy is used as the fitness function of the GA, while the GA structure was selected as NSGA-II—one of the most popular multiobjective optimization algorithms [40].

The upper and lower bound of the hyperparameters are determined as: sampling frequency 200 Hz–12 kHz, window length 64–1024 (in samples), overlapping length 64–1024 (in samples), and number of FFT points 128–4096 (in samples). Only one constraint is forced into the optimization procedure, namely, that the window length must be greater than the overlapping length.

Based on this approach, in this work, spectrograms are generated using 1024 frequency samples, a Hanning window of length 512, and an overlap of 256 samples. Note that after the spectrograms are computed, they are converted to grayscale prior to input to the ACGAN and CVAE. Before the preprocessing for the clutter mitigation, spectrograms were converted to grayscale and resized into $100 \times 100 \times 1$. This is the final dimensionality of the inputs provided to generative models and the 19-layer CNN.

B. Experimental Datasets Collected

In this article, eight different activities are considered as follows.

1) Bending—person stands and moves torso from a vertical to horizontal position, resulting in both positive and negative frequency components over the same time interval. Positive frequencies result from the forward movement of torso, coupled with negative frequencies due to the posterior moving away from the radar.
2) Falling—person falls forward onto a mattress, resulting in the shape of an upside-down bow in the signature. We only consider nonprogressive falls, which exhibit relatively high Doppler frequencies.
3) Gesturing—gross arm motion, such as by moving the arm up and down to turn the TV ON/OFF, or pointing a lamp with different orientations to turn it ON/OFF.
4) Standing—in-place motion of a person to standings up from the sitting position.
5) Kneeling—person lowers position to set one knee on the ground, as one would when tying shoelaces. This results in a distinct spike in the micro-Doppler signature.
6) Reaching—person extends torso and arms upward from a sitting position.
7) Sitting—person is standing upright, then sits on a chair.
8) Walking—micro-Doppler signature exhibits a distinct sinusoidal pattern for the strongest return caused by the slight up-and-down motion of the torso incurred as a function of time. The periodic forward-backward motion of the arms and legs results in higher amplitude, periodic oscillations modulated around the main Doppler shift. Leg motion causes the highest frequency oscillation, followed by that of the arms, which appear at distinct, midlevel frequencies.

A sample spectrogram for each class collected in two different settings is shown in Fig. 1. To create an environment for radar measurements different for training and testing data, we placed the radar in an adjacent room with obstructed line-of-sight (LOS) to the target through an interior wall. The LOS dataset was acquired from the Radar Imaging Laboratory, while the dataset associated with an

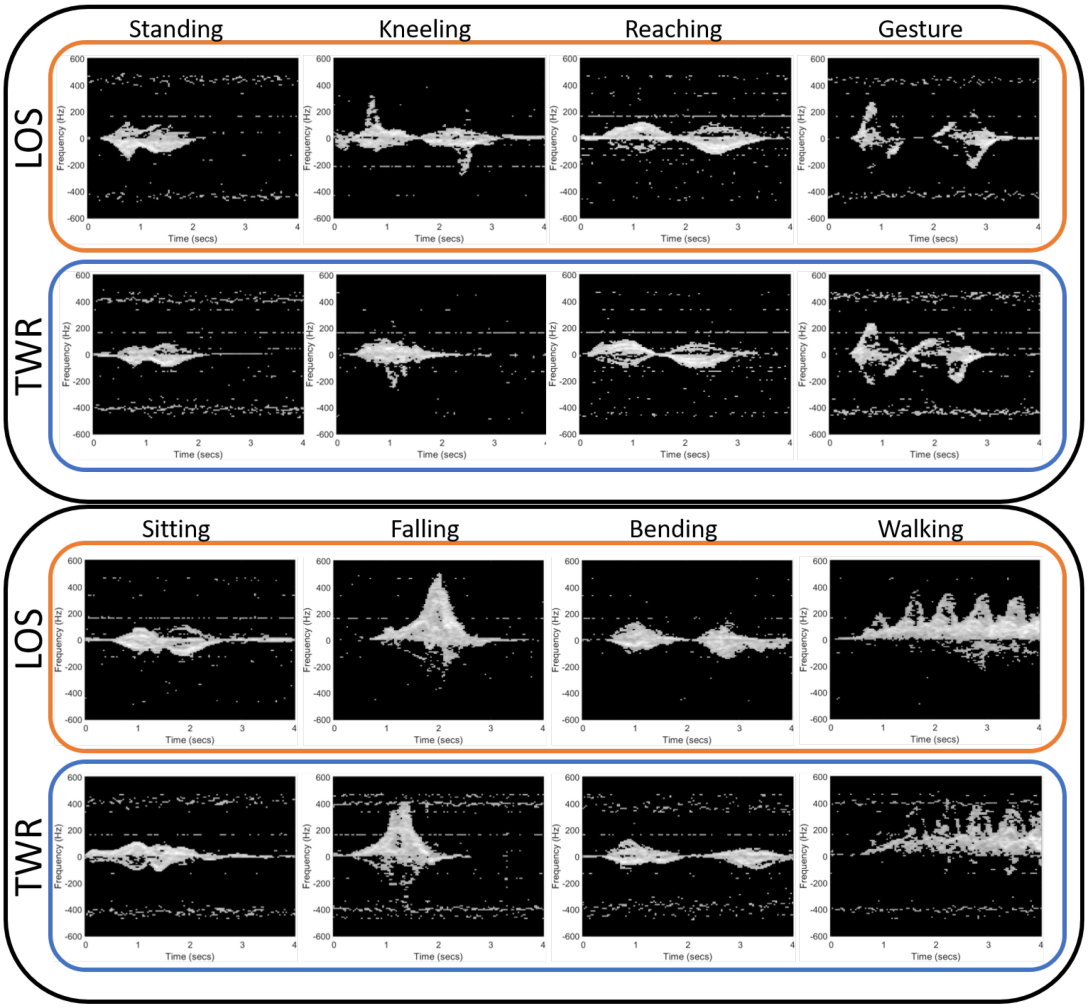

Fig. 1. Real spectrogram images (after all preprocessing) of different human activities.

obstructed LOS was acquired at the Center for Advanced Communications (CAC) conference room, both located at Villanova University. The latter sensing environment is meant to generate through-the-wall radar (TWR) dataset. The radar was placed on a table with a height of 3.2 ft for both of the locations. In the LOS experiment, a total 14 participants were involved in the data collection (12 males and two females), who had heights ranging from 5.1 to 6.3 ft, and weights ranging from 119 to 220 lbs. All activities were conducted for three different walking angles (0°, 30°, and 45°) and three different speeds (slow, typical, and fast), resulting in a dataset that covers a wide variety of motions with sufficient intra- and interclass variance. A total of 1586 samples were collected, with the number of samples per class shown in parenthesis as follows: bending (167), falling (350), kneeling (216), gesture (150), reaching (140), sitting (233), standing (130), and walking (200).

In contrast with the LOS dataset, the radar and test subject were separated by a plywood wall. Subjects started the motion 5 m away from the wall and after 4 s of data collection experiment is repeated. Experiments included both moving towards and away from the radar. The TWR dataset was conducted at four different angles, including 0°, 30°, and 45° as well as 60°. The test subject in the TWR experiments was a male participant, who was not part of the LOS data collect. A total of 387 TWR samples were collected, with bending (50), falling (72), gesture (50), kneeling (15), reaching (50), sitting (50), standing (50), and walking (50). A summary of the LOS and TWR datasets is given in Table I.

In this article, the LOS dataset was used in conjunction with the ACGAN for training data generation, while the TWR dataset was used for testing. Note that the visual similarity between 7 of 8 classes (walking is the exceptional class), inclusion of multiangle measurements, and difference in environment makes this classification problem relatively more challenging [41] in comparison to other scenarios considered in the literature.

C. Preprocessing for Clutter Mitigation

The classical signal processing approach to deal with environmental factors is to remove any clutter or unwanted artifacts using filtering. In this article, we applied an approach known as the extended CLEAN (eCLEAN)

TABLE I
Experimental Dataset Summary

	# subjects	Aspect angles	# activities	Location	# samples
LOS	14	0°, 30°, 45°	8	CAC Conf. Room	1586
TWR	1*	0°, 30°, 45°, 60°	8	Radar Imag. Lab	387

*Subject different from those in LOS dataset.

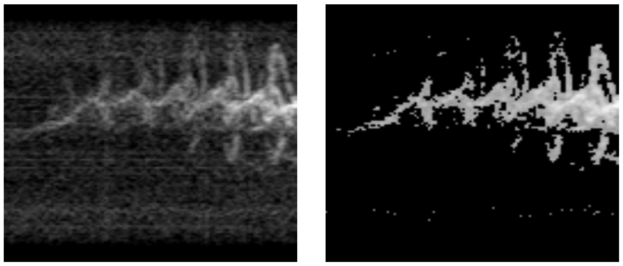

Fig. 2. Preprocessing of micro-Doppler images. (a) Predefined thresholding. (b) Proposed eCLEAN.

algorithm, which was originally designed for range-Doppler processing [42]. eCLEAN aims at suppressing unwanted distortions or noise effects while enhancing the natural structural integrity of the data. Simple predefined thresholding is the most commonly preprocessing method in the micro-Doppler processing. However, due to high variance in our data (different aspect angles, data collection environments, subjects, etc.), determining a threshold that works for every data/class is challenging. A simple example is provided in Fig. 2 for a walking micro-Doppler image filtered with simple thresholding and eCLEAN algorithm. Note that this threshold value works really well for some of the other walking micro-Doppler images, however, for this particular walking example it did not remove any of the noise components, which would degrade classification performance. On the other hand, eCLEAN removes all the noise and artifacts without needing a predefined threshold. It automatically determines the number of points, which are needed to be removed using a simple and efficient histogram-based method. eCLEAN, first, computes the 2-D histogram of the sample spectrogram, dowmsamples it and applies a normalization. Afterwards, it automatically determines the threshold where the number of counts is below 0.1. After the threshold is acquired, it slides over the time axis and examines each time column and determines the number of points should be extracted depending on the threshold. It operates on time slices and creates mask functions. This continues until all number of points are extracted. An example pseudocode of the eCLEAN is provided in Algorithm 1.

All spectrograms illustrated in Fig. 1 have had the eCLEAN algorithm applied on the data. Thus, it is important to note that clutter mitigation was not sufficient in removing all artifacts in the data, and that environmental differences remain in the two datasets despite such mitigation efforts. This point is significant because it underscores the necessity of developing DNN approaches that can overcome nontarget artifacts present in the data.

Algorithm 1: eCLEAN Algorithm.

```
Input: Training spectrogram datacube ($\mathcal{X} \in \mathbb{R}^{I_1 \text{x} I_2 \text{x} I_3}$,
    $I_1$ and $I_2$ original image sizes and $I_3$ number of
    training samples),
Output: Cleaned training spectrogram datacube ($\mathcal{X}_c \in$
    $\mathbb{R}^{I_4 \text{x} I_5 \text{x} I_3}$, $I_4$ and $I_5$ resized spectrogram
    dimensions)
    PROCESS:
    for $n = 1$ to $I_3$ do
        Y $\in \mathbb{R}^{I_1 \text{x} I_2} \leftarrow \mathcal{X}(:, :, n)$, matrix slice of tensor $\mathcal{X}$
        Normalize & resize the matrix Y and compute 2D histogram
        Find the intensity index ($\alpha_n$) where distribution starts to plateau
        for $k = 1$ to $I_2$ do
            p $\in \mathbb{R}^{I_1} \leftarrow$ Y$(:, k)$, fiber of tensor $\mathcal{X}$
            Compute 1D histogram of fiber p
            Apply $\alpha_n$ and determine the number of points (N$_s$) should be extracted from fiber p
            for $j = 1$ to $N_s$ do
                $f_j = \max_k(\mathrm{p} \in \mathbb{R}^{I_1})$,   $v_j = \arg\max_k(\mathrm{p} \in \mathbb{R}^{I_1})$
                Subtract a fraction of the point spread function centered at the peak from p.
                Record the peak amplitude and position in the cleaned vector (p$_c \in \mathbb{R}^{I_1}$)
            end for
            Store p$_c$'s in cleaned spectrogram image Y$_c \in \mathbb{R}^{I_1 \text{x} I_2}$
        end for
        Store Y$_c$'s in cleaned spectrogram datacube $\mathcal{X}_c$
    end for
```

III. GENERATIVE MODELS

The term "generative" is used in many ways in the machine learning community. Within the scope of this article, this term refers to a model that takes a training data with distribution p_{data} and seeks to learn a close estimate of it, denoted as p_{model}. More specifically, generative models attempt to predict features given a certain label, whereas, discriminative models try to predict a label of a given input data [43]–[45]. Generative models can be classified into two broad categories: explicit (VAE, PixelRNN/CNN [46], [47]) and implicit (GAN [48], Markov chain) approaches [49].

Generative models have been successfully employed in image recognition, such as performance improvement in reinforcement learning, domain adaptation, presentation, and

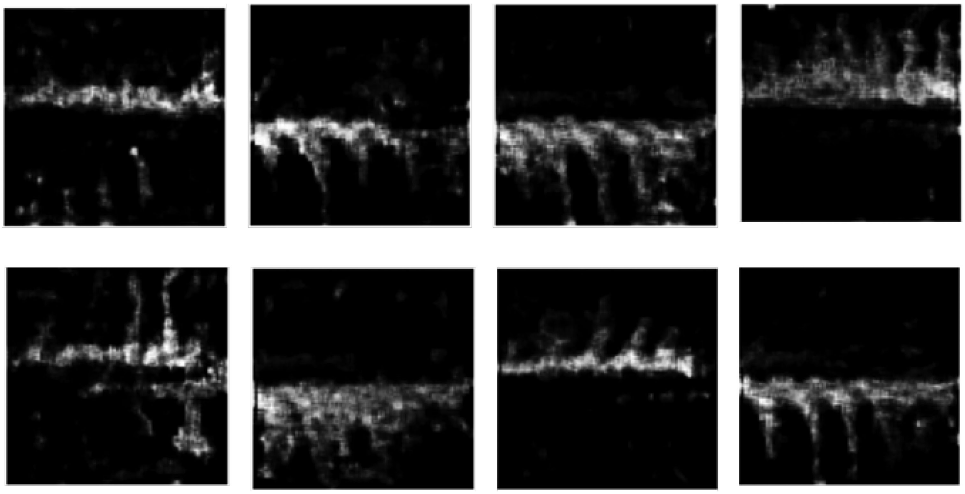

Fig. 3. Randomly chosen eight samples generated by WGAN for walking class.

manipulation of high-dimensional distributions and overcoming the problems with missing data [49]. In this article, we apply generative models in the context of human motion classification to increase the amount of training data as well as to broaden the intraclass motion diversity, while taking into account environmental factors, e.g., clutter sources, which are not included in kinematic models of human motion. The Wasserstein GAN (WGAN) is a popular variant of the GAN architecture, which employs the 1-Wasserstein distance, also known as the earth-mover distance rather than alternative metrics, such as the Kullback–Leibler (KL) divergence or the Jenson–Shannon divergence, to quantify the distance between the model and target distributions [50]. The WGAN is advantageous because it provides for a more stable training process, with proven convergence of the loss function, and is less sensitive to model architecture or hyperparameter selection.

The results of applying a WGAN to synthesize radar micro-Doppler signatures for walking is shown in Fig. 3. It may be observed that many of these samples have features that are deviant from the typical properties of walking micro-Doppler, such as high frequency components disconnected from the low-frequency micro-Doppler of the torso, negative micro-Doppler corresponding to motion in the reverse direction, and filled in regions between the peaks that would be inconsistent with the arm motion of a typical walking person.

As a result, in this article, we focus on conditional generative models, principally the CVAE and ACGAN, which allow the generative model to condition on external class labels. This has the benefit of improving the visual accuracy of the synthetic images generated. An alternative to these conditional models is to train N (number of classes) separate models. However, this has an adverse effect of causing the problem of overfitting due to the small amount of available training data, and requires high computational power. Moreover, it has been shown that forcing a model to perform additional tasks or constraints improves the performance of the original problem [35].

A. Conditional Variational Autoencoders

CVAEs are an extension of the vanilla VAE, where the input observations modulate the prior on Gaussian latent variables that generate the outputs [51]. A vanilla VAE consists of an encoder, a decoder, and a loss function. The encoder and decoder are usually designed as neural networks, and they are given the weights of θ and ϕ, respectively. The encoder takes an input image and outputs a latent representation in lower dimensions. It is important to note that the latent space is stochastic: the encoder outputs parameters to a Gaussian probability density, which can, then, be sampled to obtain noisy values of the latent representation z. Then, the decoder takes the encoded latent representation as an input and outputs parameters to the probability distribution of the data. In this article, the encoder and decoder are denoted as $q_\theta(\mathrm{z}|\mathrm{x})$ and $p_\phi(\mathrm{x}|\mathrm{z})$, respectively.

The loss function of a vanilla VAE is the negative log-likelihood with a regularizer. It can be decomposed into a single spectrogram image since there are no global connections between images. The loss function l_i for a single image x_i is defined as

$$l_i(\theta, \phi) = -\mathrm{E}_{\mathrm{z}\sim q_\theta(\mathrm{z}|\mathrm{x}_i)}[\log p_\phi(\mathrm{x}_i|\mathrm{z})] + \mathrm{KL}(q_\theta(\mathrm{z}|\mathrm{x}_i)||p(\mathrm{z})) \tag{2}$$

where the first and second term represent the reconstruction error and the regularizer, respectively. The former encourages the decoder network to learn how to reconstruct the input data, while providing the smallest error, as in basic autoencoders. If the decoder is unable to reconstruct the data well enough, then it will incur a high loss function value. The regularizer is the KL divergence, which measures how much information is lost when using $q_\theta(\mathrm{z}|\mathrm{x})$ to represent $p(\mathrm{z})$. The regularization term forces the encoder to map images from the same classes onto the same region in the latent space. Moreover, in the VAE, p is specified as the normal distribution with mean zero and variance one ($N(0, 1)$).

Similar to vanilla VAEs, a CVAE consists of an encoder, a decoder, and a loss function. However, in contrast to VAEs, CVAEs have additional input branches called conditions (external class labels) to both the encoder and decoder. Due to embedding of class labels, the encoder is conditioned on the spectrograms and corresponding class labels, whereas, the decoder is conditioned on latent variables and class labels. Other, than, conditional embeddings, CVAEs have the same principle as VAEs, where the encoder takes the spectrograms and class labels (x, y) and outputs a hidden representation z, with the attached weights (θ) and biases (ϕ). Then, the decoder takes z and y as inputs and outputs the parameters to the probability distribution of the data. The CVAE is trained to maximize the conditional log likelihood. In CVAEs, the empirical lower bound is defined as

$$L_{\text{cave}}(\mathrm{x}, \mathrm{y}; \theta, \phi) = -\mathrm{KL}(\, q_\phi(\mathrm{z}|\mathrm{x}, \mathrm{y}) \,||\, p_\theta(\mathrm{z}|\mathrm{x})) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathrm{y}|\mathrm{x}, \mathrm{z}^{(l)}) \tag{3}$$

where $\mathbf{z}^{(l)} \approx N(0, 1)$, L is the number of samples, $q_\phi(z|x, y)$ is the conditional recognition distribution, and $p_\theta(z|x)$ is the generative distribution. A more detailed theoretical background and implementation considerations on VAE and CVAE can be found in [36].

As a preprocessing step, the input spectrograms (64 × 64 × 1) are reshaped into flat vector representations of 4096 × 1 pixel values. Then, the vectorized spectrogram images and class labels are concatenated. In our case, the input size of the CVAE is 4104 × 1 (reshaped image size + number of classes). The encoder and decoder configurations used in this article consist of fully-connected (dense) layers. The encoder takes the merged data and passes it to sequential dense layers with specified neurons and activation functions: (2048 × 1) - ReLU, (1024 × 1) - ReLU, (512 × 1) - ReLU, and 10 × 1 - Linear. The encoder has a total of 11 018 762 (trainable) parameters. The final layer is responsible for the mean and standard deviation for the variational sampling that will occur from the latent space z. After sampling, the decoder reconstructs $\hat{\mathrm{x}}$ and consists of four dense layers as (512 × 1) - ReLU, (1024 × 1) - ReLU, (2048 × 1) - ReLU, and (4096 × 1) - Sigmoid. The decoder has total of 11 022 848 (trainable) parameters. We applied stochastic gradient descent with Adam optimizer [52], an adaptive moment estimation method, controlled by parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is determined as 0.0005 for 500 epochs and minibatch size of 16. A total of 40 000 synthetic spectrograms are generated using CVAE (5000 for each class).

B. Auxiliary Classifier Generative Adversarial Networks

GANs are implicit generative models that aim to learn the data distribution from a set of training samples. Due to their implicit structure, generative models do not need any intractable density functions as in CVAE. The basic idea of GANs stems from a game-theoretic approach between two players (both neural networks): generator, and discriminator. These two entities are in constant battle during training. The generator (G), seeks to generate samples that are intended to come from the same distribution of the training data. The input of the generator can be sampled from a Gaussian distribution as random noise. The generator gets samples z from the selected distribution and maps $G(z)$ to the image space. The main goal of the generator is to make the image space distribution as close as possible to the p_{data}. The second network is called discriminator and denoted as D. The role of the discriminator is to discriminate between real and fake samples generated by the generator. It takes a simple input x and outputs D(x), which is a probability of the given image is being real.

Since GANs use a game-theoretic application, the objective function can be represented as a minimax function. In essence, the discriminator tries to maximize the objective function using gradient ascent, whereas the generator tries to minimize the objective function using gradient descent. Training of these networks can be done by alternating between gradient ascent and descent. The loss function of the adversarial networks can be shown as

$$\min_G \max_D \mathrm{E}_{x\sim p_{\text{data}}} \log(D(x) + \mathrm{E}_{z\sim p_z}[\log(1 - D(G(z)))]). \tag{4}$$

In the objective function, the discriminator is trained to maximize the D(x) for images with x $\sim p_{\text{data}}$. The objective of the generator is to produce images $G(z)$ to fool D during training such that D(G(z)) $\sim p_{\text{data}}$. During training, the generator improves its ability to synthesize more realistic images while discriminator improves its ability to distinguish between real from fake images.

ACGAN is an extension of the vanilla GAN model that enables the model to be conditioned on external labels to improve the quality of the generated images. One method to produce class conditional samples can be done by supplying both generator and discriminator with class labels as in CVAE. However, the strategy behind the ACGAN is to instead of feeding the class information to the discriminator, one can task the discriminator with reconstructing the label information. This can be done by modifying the discriminator to contain an auxiliary decoder network that outputs the class labels for the training data [35]. In this respect, the objective function of the ACGAN has two parts: the log likelihood of the correct source L_{s}, and the log likelihood of the correct class L_{y}

$$L_s = \mathrm{E}\left[\log p(\mathrm{s} = \mathrm{real}|\mathrm{x}_{\text{real}})\right] + \mathrm{E}\left[\log p(\mathrm{s} = \mathrm{fake}|\mathrm{x}_{\text{fake}})\right] \tag{5}$$

$$L_{\mathrm{y}} = \mathrm{E}\left[\log p(\mathrm{Y} = y|\mathrm{x}_{\text{real}})\right] + \mathrm{E}\left[\log p(\mathrm{Y} = y|\mathrm{x}_{\text{fake}})\right] \tag{6}$$

where s are the generated images. The discriminator is trained in order to maximize the $L_s + L_Y$ whereas the generator is trained to maximize $L_Y - L_s$.

The employed ACGAN architecture consists of two different parts: generator, and discriminator. The generator takes a vector of 100 × 1 random noise (latent space) drawn from a uniform distribution ($N(0, 2)$) and class labels as inputs and outputs a spectrogram image of size 64 × 64 × 1. We used a similar generator network, as in the original

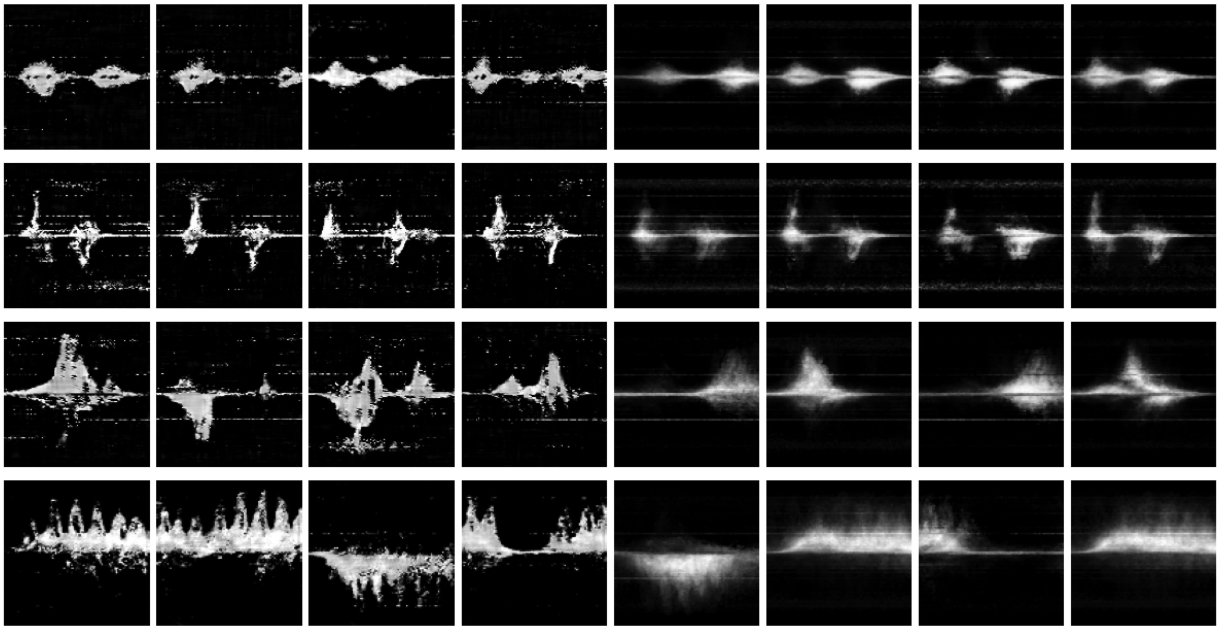

Fig. 4. Randomly chosen four samples generated by ACGAN (first four in rowwise) and CVAE (last four in rowwise). Each row represents a different class (bending, gesture, falling, and walking).

ACGAN paper, with minor modifications for generating radar spectrogram images. The generator consists of a fully connected dense layer reshaped to size 4 × 4 × 128 and four 2-D convolutional layers with 3 × 3 kernel size. Filter sizes for each convolutional layer are determined as 256, 128, 64, and 1. The last layer contains only one filter due to gray-scale channel size. Batch normalization with the momentum of 0.8 and 2-D up-sampling (kernel size 2 × 2 with strides of 2) are applied to each layer (including the dense layer) of the generator network, except for the output layer. In addition to the batch normalization, dropout of 0.15 is also applied in every even layer considering the small amount of real training data. Adding these regularizes into the generator network helps combat overfitting and mode collapsing. ReLU activation functions are applied to all convolutional layers except the output layer, which employs a tanh activation function. Discriminator structure consists of seven 2-D convolutional layers with a kernel size of 3 × 3. LeakyReLU is utilized as an activation function after every convolutional layer except for the last one (the slope of the leak was set to 0.2). Max pooling is only included in the first layer with a filter size of 2 × 2 and strides of 2. Downsampling is done in every odd convolutional layer with a stride rate of 2. Batch normalization with momentum 0.8 is utilized in every layer except for the first one. In addition to batch normalization, a dropout of 0.25 is applied in every even layer. The number of filters in each convolutional layer is determined as 64, 128, 128, 256, 256, 512, and 512. The last layer of the discriminator uses a sigmoid for the validity of the generated images and softmax for reconstruction of the class labels.

The preprocessing step for the ACGAN (also for CVAE) includes a cleaning and filtering algorithm, which is followed by scaling of the images between $(-1, 1)$ for tanh activation function. Weights are initialized with a normal distribution. An Adam optimizer is utilized with learning rate of 0.0002, $\beta_1 = 0.5$, and $\beta_2 = 0.999$ for 3000 epochs and minibatch size of 16. Some examples generated by the proposed ACGAN are depicted in Fig. 4. A total of 40 000 synthetic spectrograms are generated using ACGAN (5000 for each class).

Note that the training of CVAE and ACGAN are done offline with a PC equipped with GT 1080Ti. The computational cost of the CVAE is low due to the fast convergence (around 500 epochs) of the autoencoder topology. However, for the ACGAN, convergence takes more time due to the minmax structure of the adversarial learning. Moreover, the topologies of the generator and discriminator in the ACGAN are more complex than that of the CVAE, which results in increased computational time costs. Our experimentation shows that the ACGAN converges after around 5000 epochs.

IV. KINEMATIC EVALUATION OF SYNTHETIC SIGNATURES

Despite progress in the theoretical understanding of generative models and increased attention in GAN research, evaluating and comparing the performance of these models still remain a hard task. While several measures have been introduced, there is no consensus yet as to which measure best captures the strengths and limitations of the models and yield a fair model comparison [53]. Moreover, evaluation metrics are usually problem specific. Because the underlying physics of the problem is different, performance metrics valuable in the optical domain, such as inception score or Fréchet inception distance, do not necessarily translate to the RF domain.

In radar micro-Doppler classification, important challenges in the context of training include obtaining a sufficient amount of real data to drive synthetic data generation with GANs, while ensuring that the synthetic signatures are diverse, spanning the characteristics of all expected motion signatures, and correspond to human motion that is physically possible. In the underlying problem, the human skeleton constrains the possible variations of spectrograms corresponding to a given class—a condition we should observe to avoid erroneously training the network. Towards this end, we consider four measures to evaluate the efficacy of the generative networks: 1) visual inspection, 2) kinematic fidelity, 3) signature diversity, and 4) the dimension of the latent space.

A. Visual Inspection

A sample of some of the spectrograms generated by CVAE and ACGAN are shown in Fig. 4 for four classes: bending, gesture, falling, and walking. At the outset, it may be noticed that the CVAE-generated signatures are almost unrealistically blurry, a feature exhibited across all classes. The main reason for this blurriness stems from the challenge of fitting of the data distribution into a tractable density distribution.

1) *Bending:* Both ACGAN and CVAE capture the most essential kinematic property of bending, namely, the presence of positive and negative frequency components within the same time limit. Moreover, ACGAN was able to learn to place time separation between the first and second part of the bending motion. In some generations, the time difference between the bending down (first hump) and standing up (second hump) is close 0.2 s, whereas in some other generations it is up to 2 s.

2) *Gesturing:* For gesturing, the ACGAN generated some variations capturing the different orientations and velocities of the arm. CVAE again seizes the kinematic property of the motion; however, generated images remain blurry.

3) *Falling:* For falling, ACGAN underscored some salient features about the motion articulation. Note that, all real falling experiments in the LOS dataset were performed towards the radar, resulting in positive Doppler frequency. Interestingly, ACGAN learned how to mirror the spectrogram and generated some examples as if the subject had performed the motion in the opposite direction. Moreover, in some cases, ACGAN generated signatures that resemble "progressive falling": i.e., putting the knee first—holding onto something then falling.

4) *Walking:* In walking, again the principal kinematic properties are captured by ACGAN. The motions of the legs and arms can be seen in the example spectrograms in Fig. 4. In the final example for walking, ACGAN generated a spectrogram, which has a gap over which the micro-Doppler is nearly zero. Kinematically, this corresponds to a situation in which the subject was walking, took two steps (as evidenced by the two peaks in frequency), stopped, and then took another two steps walking forward.

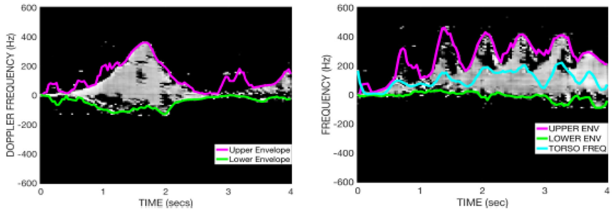

Fig. 5. Rule definition process: low-pass filtered upper/lower envelope and torso frequency extraction in the generated images (left: falling, right: walking).

B. Kinematic Fidelity

The patterns observed in radar spectrograms directly relate to kinematics of the motion being observed. For example, in the case of walking, the torso response represents the strongest return and exhibits a sinusoidal oscillation. The periodic motion of the legs causes the highest frequency oscillations around the main Doppler shift. Known as physical features, such properties have often been used in classification of micro-Doppler signatures. In this section, we consider kinematic properties of synthetic walking and falling signatures, as these are challenge cases for the ACGAN due to the great diversity within real training samples as well as the greater richness of frequencies comprising the signature.

In particular, kinematic fidelity of the synthetic signatures are evaluated by imposing upon the images three different kinematic rules.

1) Generated spectrograms should be periodic, and thus, represent the cyclic motion of the body.
2) The maximum torso frequency should be lower than the that of the legs.
3) If the generated spectrogram occupies the positive frequencies, indicating that the motion is performed towards the radar, the signature should not contain any high negative frequency components, and vice versa.

The first rule is only applicable to periodic motions, whereas the other two rules can be applied to walking and falling. Note that there is no guarantee that these kinematic rules ensure that every generated signature is fully compatible with the kinematic constraints of human motion. However, they do serve to enforce the most basic properties of the skeletal constraints on human motion, and can eliminate unrealistic or impossible synthetic signatures.

The abovementioned three rules can be tested by extracting the upper/lower envelopes and the torso frequency, as depicted in Fig. 5. To illustrate the process of sifting the data with these kinematic rules, let us randomly select 25 synthetic walking spectrograms generated by ACGAN, as shown in Fig. 6. The green labels indicate that the synthetic images passed all three kinematic rules, while orange indicates minor issues (i.e., only one or two rules failed), and red indicates that the image fails. Inspecting the two images from this random selection of 25 signatures, it may be observed that one fails because it only has a faint clutter

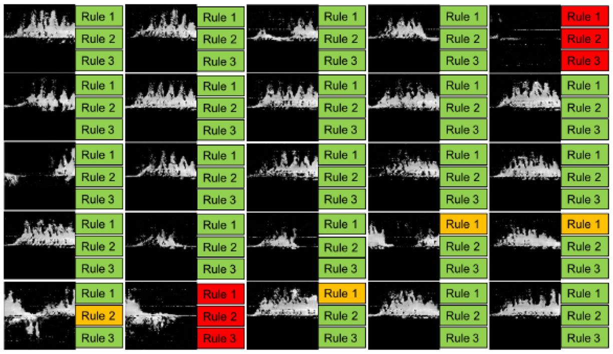

Fig. 6. Output of the kinematic sifting algorithm on the 25 randomly selected walking spectrograms generated by ACGAN. Green, orange, and red colors indicate that image passes the rule without any problems, passes the rule with minor problems, and fails the rule.

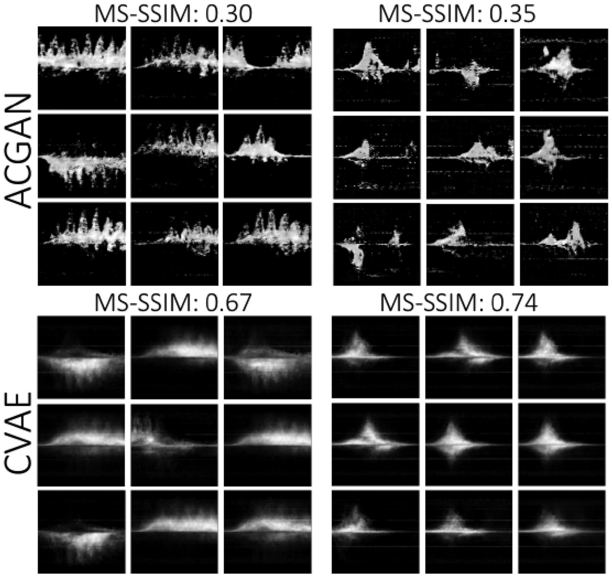

Fig. 7. MS-SSIM scores for randomly chosen 100 walking (left) and falling (right) image pairs for ACGAN (top) and CVAE (bottom).

component and essentially no target component. The other signature that failed all rules overtly has no periodicity and inconsistent distribution of positive and negative frequencies. Furthermore, the strongest response is not consistent with typical torso motion.

When these rules are applied to the 5000 synthetic walking spectrograms generated by ACGAN, 15% of the signatures failed all three of the kinematic rules. For falling, only the second and third rules were enforced, resulting in the failure of 10% of the signatures. These results mean that while ACGANs are predominantly successful in simulating human motion, there is nevertheless a significant portion of the synthetic data, which is kinematically impossible and could lead to a degradation in classification accuracy. In this article, we propose implementation of a PCA-based kinematic sifting algorithm to eliminate such undesirable synthetic samples prior to utilizing the synthetic data for training. This is discussed in more detail in Section V.

C. Synthetic Data Diversity

A generative model is considered unsuccessful if it only outputs one type of image (also known as mode collapse). This is a well-known phenomenon in GANs, where the generator will collapse and outputs a single prototype that maximally fools the discriminator [50]. To evaluate potential mode collapse, we utilized a quantitative similarity measure called MS-SSIM [54]. MS-SSIM attempts to discount aspects of an image that are not important for human perception. It assumes the values range between [0, 1] where higher values correspond to perceptually more similar images, and smaller values indicate a better diversity. Mathematically, it is defined

$$\mathrm{SSIM}(\mathrm{x}, \mathrm{y}) = [l_M(\mathrm{x}, \mathrm{y})]^{\alpha_M} \prod_{j=1}^{M} [c_j(\mathrm{x}, \mathrm{y})]^{\beta_j} [s_j(\mathrm{x}, \mathrm{y})]^{\gamma_j} \quad (7)$$

where α_M, β_j, and γ_j are used to adjust relative importance of different components. Luminance, contrast and structure comarison measures are defined as $l((\mathrm{x}, \mathrm{y}))$, $c((\mathrm{x}, \mathrm{y}))$, $s((\mathrm{x}, \mathrm{y}))$, respectively. M depicts the scale number that will be used in the iterative filtering and downsampling. x and y are defined as the compared images.

As a simple example, we randomly selected 100 image pairs from both CVAE and ACGAN-based synthetic spectrogram datasets of the walking and falling classes. Some sample images from the chosen pairs can be seen in Fig. 7. The MS-SSIM values for CVAE for walking and falling are found to be 0.67 and 0.74, respectively. These values indicate that CVAE has low diversity for the selected random pairs. For the ACGAN, MS-SSIM values are found to be 0.30 for walking, and 0.35 for falling, indicating a much higher degree of diversity among the synthetic signatures generated. For comparison, measured samples of falling and walking yielded MS-SSIM values of 0.45 and 0.40, respectively. Thus, the ACGAN generated signatures provide not only sharper images, relative to CVAE generated signatures, but also a greater degree of diversity that is comparable to that expected based on measured data.

The results of detailed analysis of the MS-SSIM values for ACGAN-generated synthetic data are given in Fig. 8, which shows a box plot of the MS-SSIM values for 100 randomly selected image pairs in each class. The walking class exhibits the most diversity, as would be expected by the possible variations of a complex motion. Gesturing and falling exhibit the next greatest levels of diversity. Considering that falling is a more or less uncontrolled motion, and that gesture is highly open to participant interpretation during enactment, these results match expectations.

The level of diversity in the synthetic dataset generated also depends upon the amount of measured data provided to ACGAN during generation. Fig. 8(b) depicts the variation of MS-SSIM values for falling and walking as a function of training data size. Juxtaposed on top of this curve is the relation between training data size and percentage of synthetic samples that pass the kinematic rules. Note that

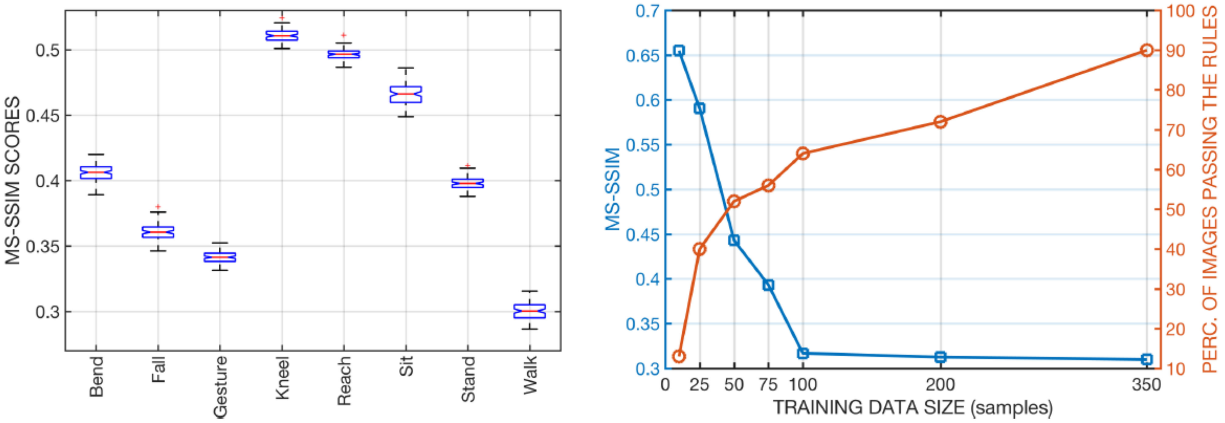

Fig. 8. Diversity measures: (a) Box plots of the intra-class diversities measured by MS-SSIM, (b) MS-SSIM diversity values and percentages of kinematically correct images as a function of the real training samples used in the training.

when a minuscule amount of measured data is used to drive the ACGAN, the network has a tendency to generate a large amount of data that is highly similar (over 0.8 MS-SSIM values), and that are also kinematically meaningless—virtually none of the synthetic signatures actually adheres to kinematic rules. It is only when at least 100 measured samples are supplied the ACGAN that signatures are obtained, which predominantly meet kinematic rules and exhibit a high degree of diversity. Further increasing the amount of measured data supplied for training the ACGAN does not significantly affect the data diversity, but does increase the kinematic fidelity of the data generated. Note that when 350 measured training samples are utilized, the percentage of data passing the kinematic rules rises to 90%—a 25% increase of the percentage attained with just 100 measured training samples.

In micro-Doppler literature, studies involving several thousand measurements are typical. The proposed method, however, requires only 350 samples to generate 40 000 synthetic samples. This represents a 10 fold decrease in data collection requirements, while enabling increased sample diversity and a 100-fold increase in the size of the training dataset. As shown in Section V, DNNs trained on this synthetic dataset outperforms DNNs trained on measured data only.

D. Strolling in the Latent Space

Analysis of the latent manifold helps us to understand the model details, indicates the signs of memorization, and shows if the latent space is hierarchically collapsed [45]. Kinematic and physical changes (such as different velocities, the orientation of the target, the direction of the motion, etc.) in the generated spectrograms while strolling through the latent space indicate that the model has learned relevant and interesting representations from the training data. As an example, consider an ACGAN retrained with a generator that has a latent size of 5×1. After the training is complete, a walking image is generated by randomly sampling the latent variables from a uniform distribution and passing it through the generator. Then, we changed the first latent variable value between -3.0 and 3.0 with linear increments of 1.5. Note that the other four latent variables are kept fixed during this operation. The resulting walking spectrograms for different latent variable values are depicted in the top row of Fig. 9.

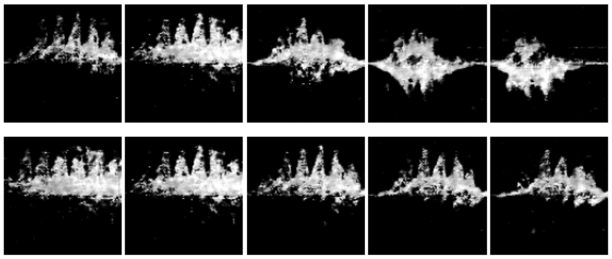

Fig. 9. Strolling on the latent space: top row and bottom rows depict the generated images by changing the first and second latent variables to (− 3.0, − 1.5, 0, 1.5, 3.0), respectively.

Examining the top row of Fig. 9, as the value of the latent variable is increased, the Doppler bandwidth is first reduced and then begins to flip, with an increasing peak in the negative Doppler frequencies. Moreover, the Doppler bandwidth does indeed vary according to the aspect angle between the radar LOS and target direction of motion. Thus, it may be deduced that the first latent variable models direction of motion.

Next, we change the second latent variable and observe the resulting changes in micro-Doppler (see second row of Fig. 9). It is evident that this variable models stride rate. This may be seen by counting the peaks in the signature. While the leftmost spectrogram has six distinct peaks, with each peak corresponding to a step, the last spectrogram on the right has only three peaks. This indicates that stride rate decreases as the second latent variable increases.

As can be seen from the above mentioned example, the dimensionality of the latent space effectively relates to how the network models the underlying representation of the data. In general, the question of how many latent variables should be used in GANs still remains unanswered. However, it is known that the real distribution arises out of lower

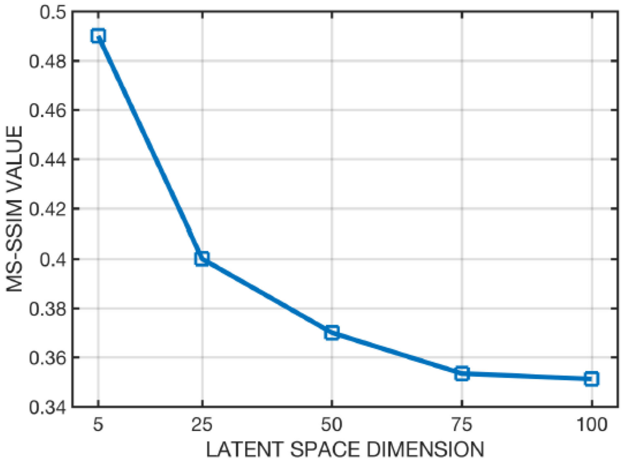

Fig. 10. ACGAN latent space analysis of falling spectrograms in terms of diversity measurements.

dimensional latent distributions. There is also a concern if a lower dimensional latent space is utilized, the GAN might not have enough information to model the data, causing the modes to collapse. A large latent space dimension, on the other hand, makes the model so complex that the training time becomes overly long. Moreover, the mapping of latent variables into spectrograms becomes difficult in high-dimensional latent space. Some earlier works have used 100 as a *de facto* value [45]. Using this as a baseline, we examine the effect of latent space dimensionality on the MS-SSIM values of resulting synthetic spectrograms. Five ACGAN models are trained with different latent space dimensions: 5, 25, 50, 75, and 100. The resulting MS-SSIM diversity metrics for each model is shown in Fig. 10 for the falling class. It may be seen that small latent space dimensions suffer from low diversity due to the limited number of combinations that can be achieved, while increasing the latent space dimension yields better diversity. However, beyond 75 latent variables, the diversity starts to plateau, indicating that increasing complexity is offering little benefit to data diversity. Thus, in this article, we elect to use 100 latent variables in our implementation of the ACGAN.

E. Evaluation of Kinematic Fidelity With PCA

In this section, we propose a kinematic sifting algorithm using generalized PCA (GPCA) [55]. Kinematic evaluation of the synthetic spectrograms in Section VI proved that some spectrograms are still kinematically not consistent with true human motion. The proposed sifting algorithm aims to eliminate some of the inconsistent images that might degrade classification performance. GPCA is first applied on the real training images for each class, $D_i, i = 1, 2, \ldots, 8$. The objective is to find a matrix subspace set $\tilde{\mathrm{U}}^{(1)}_{D_i} \in \mathbb{R}^{I_1 \times P_1}$ and $\tilde{\mathrm{U}}^{(2)}_{D_i} \in \mathbb{R}^{I_2 \times P_2}$ that project the original tensor into a low-dimensional matrix subspace $\mathrm{Y}^{D_i}_m \in \mathbb{R}^{P_1 \times P_2}$ (with $P_1 \leq I_1$ and $P_2 \leq I_2$) defined as

$$\mathrm{Y}^{D_i}_m = \mathrm{S}^{D_i}_m \times_1 \mathrm{U}^{(1)^T}_{D_i} \times_2 \mathrm{U}^{(2)^T}_{D_i} \qquad (8)$$

Algorithm 2: Data Sifting With Generalized PCA.

```
Input: Real and generated spectrograms
Output: Kinematically sifted images
for EACH CLASS do
    Read real spectrograms
    Apply GPCA subspace learning method and find the optimized subspaces, (100x100 spectrogram dimensions reduced to 2x2)
    Find the boundaries of the reduced feature space using the 4 dimensional convex hull method
    Save the convex hull parameters and optimized subspaces.
end for
for EACH CLASS do
    Read the genered spectograms
    Load the optimized subspaces and convex hull parameters
    Apply optimized subspaces and get the reduced feature space
    Check if the current generated spectrogram is within the convex hull boundaries with a tolerance
    If in keep else eliminate
end for
returnSifted images
```

where $\mathrm{S}^{D_i}_m$ is the real training spectrogram from class D_i. The objective function of the GPCA can be written as

$$(\tilde{\mathrm{U}}^{(1)}_{D_i}, \tilde{\mathrm{U}}^{(2)}_{D_i}) = \underset{\mathrm{U}^{(1)},\mathrm{U}^{(2)}}{\arg\max} \sum_{m=1}^{M} \left\| \mathrm{Y}^{D_i}_m - \overline{\mathrm{Y}^{D}_i} \right\|^2_F \qquad (9)$$

where $\overline{\mathrm{Y}} = \frac{1}{M} \sum_{m=1}^{M} \mathrm{Y}_m$. The core matrix for each m samples can be obtained by projecting the original images using optimized subspaces $\tilde{\mathrm{U}}^{(1)}_{\mathrm{D_i}}, \tilde{\mathrm{U}}^{(2)}_{\mathrm{D_i}}$ as

$$\tilde{\mathrm{Y}}^{D_i}_m = \mathrm{S}^{D_i}_m \times_1 \tilde{\mathrm{U}}^{(1)^T}_{D_i} \times_2 \tilde{\mathrm{U}}^{(2)^T}_{D_i}. \qquad (10)$$

Finally, the feature vector of a training sample for a specific class m can be constructed as $\mathrm{C}_m = \mathrm{vec}(\tilde{\mathrm{Y}}_{\mathrm{m}}), \in \mathbb{R}^{1 \times D}$, where $D = P_1 \times P_2$ and $\mathrm{vec}(\cdot)$ is the matrix columnwise vectorization operator. We defined P_1 and P_2 as 2.

Upon finding the optimized subspaces and reduced feature space for each class, the n-dimensional convex hull method is applied to determine the feature space boundaries of the each class. Falling and walking feature space boundaries (with feature space dimensions set to 3) are shown in Fig. 11. Next, the optimized subspaces on the synthetically generated images are used to reduce the dimensionality of the feature space. Finally, features are checked to ensure they fall within the specific class boundaries, as determined by real training examples. Pseudocode of the proposed method is depicted in Algorithm 2. By using the sifting method, we are able to eliminate 11% of bending, 18% of falling, 8% of gesture, 33% of kneeling, 7% of reaching, 15% of sitting, 36% of

TABLE II
Performance Comparison (Accuracy) Between TF-AlexNet, CVAE, and ACGAN, PCA-ACGAN With Tolerance 1.0 and PCA-ACGAN With Tolerance 0.5

	TF-AlexNet	TF-VGG16	CVAE	ACGAN	PCA-ACGAN-TOL-1.0	PCA-ACGAN-TOL-0.5
Accuracy	0.765	0.842	0.732	0.825	0.877	0.932

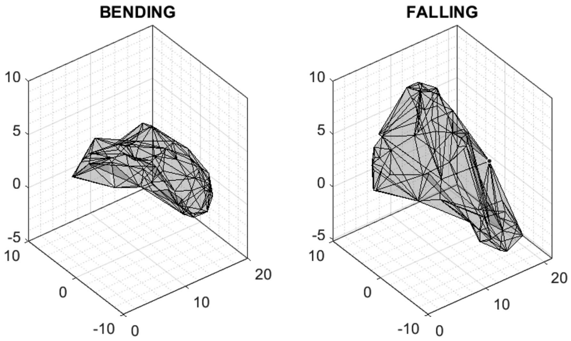

Fig. 11. Boundaries determined by n-dimensional convex hull for bending and falling.

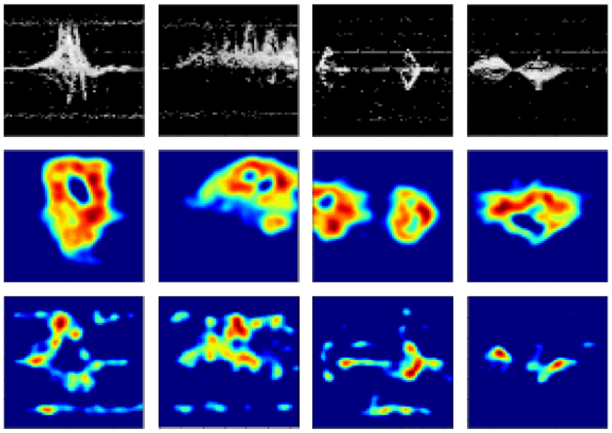

Fig. 12. Original TWR (first row) spectrograms and corresponding saliency maps achieved by the ACGAN-DCNN (second row) and DCNN (third row) for different class samples (columns left to right: falling, walking, gesture, reaching).

standing, and 22% of walking. This elimination reduces the generated dataset size from 40 000 to 31 133.

V. ACGAN WITH PCA-BASED KINEMATIC SIFTING

A. Experimental Results

1) *Classification Accuracy:* In this section, we present classification performances of transfer learning on AlexNet (TF-ALexNet) and VGGnet (TF-VGG16), DCNNs trained on the synthetic spectrograms generated by CVAE (CVAE-DCNN) and ACGAN (ACGAN-DCNN) with and without kinematic sifting at various tolerances, as shown in Fig. 13. Note that both AlexNet and VGG16 are pretrained on ImageNet. After the weights of the networks are acquired, only the last two layers are retrained using the real radar data. Moreover, the last softmax layer was also adjusted to the number of classes. Moreover, we tested two different tolerances in the proposed kinematic sifting algorithm, labeled as PCA-ACGAN-TOL-1.0 and PCA-ACGAN-TOL-0.5.

To analyze the improvement achieved by the generative models, we collected a challenging data test set in a completely different environment and configuration from the real samples (mentioned in Section II), which were used to train CVAE and ACGAN. The test performances are presented in Table II in terms of accuracy. The average test accuracies for TF-AlexNet, TF-VGG16, CVAE-DCNN, ACGAN-DCNN, PCA-ACGAN-DCNN-TOL-1.0, and PCA-ACGAN-DCNN-TOL-0.5 are determined to be 76%, 84%, 73%, 82%, 87%, and 93%, respectively. In prior studies, VGGnet is a network that has provided accuracies that have surpassed that of other pretrained networks, such as GoogleNet, as well as convolutional autoencoders (CAEs) and supervised learning with handcrafted features [56], when the amount of training data is limited [19]. Therefore, it is not surprising that VGGnet surpasses the performance of AlexNet, and even that attained by using the unsifted, initial training database generated by ACGAN without consideration of any kinematics. The low performance of the CVAE-DCNN is also expected since the generated images are blurry and unrealistic, have low diversity, and there was no fine tuning in the training.

Significantly, the performance of the DCNN *increases* as the ACGAN-generated synthetic signatures are increasingly sifted, identifying, and discarding those samples that are kinematically impossible and, therefore, unrepresentative of the related class label. The most sifting is done with the smallest tolerance, and the performance of the proposed PCA-ACGAN-DCNN-TOL-0.5 approach is drastically higher than that achieved with transfer learning, or training data that is unsifted. This result further demonstrates the need to consider physics and kinematics in the generation of synthetic training data for micro-Doppler classification.

2) *Confusion Matrix:* The test confusion matrix of the PCA-ACGAN-DCNN-TOL-0.5 is provided in Table III. It is observed that the proposed scheme provides the best test accuracy around 93%. The primary source of confusion is between bending and kneeling, as expected. These two motions have the same kinematic structure in the TF domain. Both include a positive and a negative hump adjacent to each other, which depict the motion of the upper body in forward direction for reaching, and the motion of the upper body in the downwards direction for kneeling. However, one significant difference between these activities is the motion of the knee. In some experiments the motion of the knee was very pronounced, resulting in increased confusion.

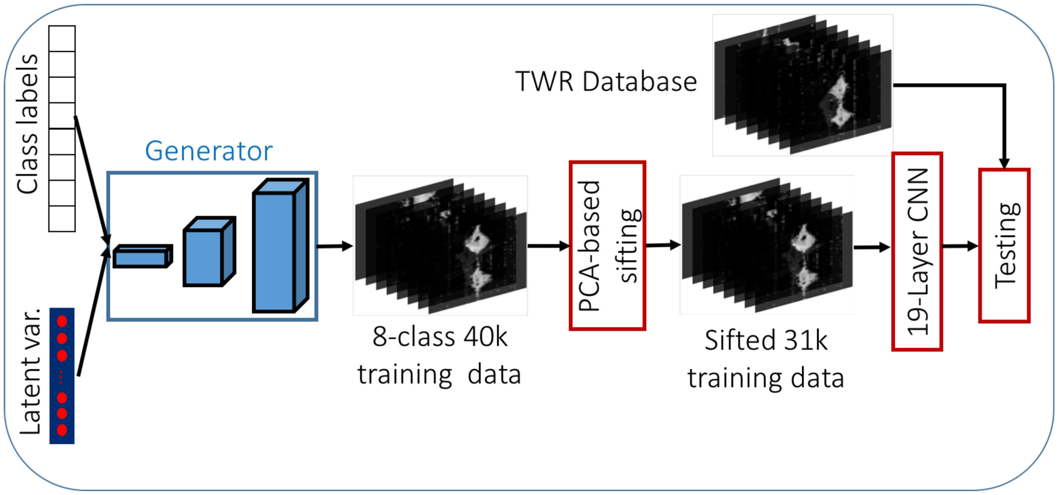

Fig. 13. Flowchart of the proposed approach with PCA sifting algorithm. Note that dataset 2 was collected in a TWR setup.

TABLE III
Test Confusion Matrix for PCA-ACGAN-DCNN With Tolerance 0.5 (Test Accuracy of 93%)

%	Bending	Falling	Gesture	Kneeling	Reaching	Sitting	Standing	Walking
Bending	100	0	0	0	0	0	0	0
Falling	0	90	3	0	2	0	0	5
Gesture	0	0	96	0	0	0	0	4
Kneeling	20	0	0	80	0	0	0	0
Reaching	0	0	0	0	84	16	0	0
Sitting	0	0	0	2	0	98	0	0
Standing	0	0	0	2	0	0	98	0
Walking	0	0	0	0	0	0	0	100

There is also a high misclassification rate between reaching and sitting. Some reaching signatures only contain the forward or backward motion of the upper body, which results in similar signatures as that for sitting. Next, there is 5% misclassification between falling and walking. This is caused by the presence of progressive falls in the training database and lower stride rates in some of the testing signatures. Some testing samples include only one or two strides meaning that subject only took 1 or 2 steps. These signatures have similar visual representation to falling signatures due to their low periodicity.

A high fall detection performance (90%) is achieved by employing the proposed method. Note that, experimental results are based on real TWR radar data obtained from multiple aspect angles (0°, 30°, 45°, 60°, 90°), which demonstrates that the proposed algorithm yields the highest overall classification accuracy among other methods.

3) *Saliency Maps:* The benefits of training the DCNN with a large synthetic training dataset can also be illustrated by examining saliency maps of the network. A saliency map is an image that shows each pixel‘s unique importance according to the DCNN’s classification declaration [57]. Saliency maps of four test spectrogram images are computed for ACGAN-DCNN and DCNN (trained with dataset 1) and are shown in Fig. 12. It may be observed that the ACGAN-DCNN places more importance on the perimeter of the actual signatures in the spectrograms, and ignores the noisy parts in the images. In contrast, a basic CNN trained on measured data focuses on some of the noisy parts and looks for specific pixels rather than the signature envelope. More specifically, the ACGAN-DCNN wraps all physical components of the falling, whereas the basic CNN puts importance on the highest frequency components. This trivialization by the basic CNN can be problematic with high aspect angle data. For the “reaching” class, the ACGAN saliency map shows that the network tries to connect two different (positive and negative) components together, whereas those components are disjoint in the basic CNN. In summary, the ACGAN-DCNN approach enables correct identification of the motion components of the spectrogram, rejecting clutter components. Saliency map observations, thus, reinforce the importance of training on a large dataset that has great diversity, while also maintaining kinematic fidelity.

VI. CONCLUSION

In this article, we proposed a novel approach for generating synthetic radar micro-Doppler signatures for human motion classification. The proposed approach leverages auxilliary conditional generative adversarial networks (ACGANs) to build a diverse dataset for training deep neural networks. However, the ACGAN-generated signatures include kinematically impossible signatures, which can degrade classification performance. To overcome this problem, a PCA-based kinematic sifting algorithm was proposed to eliminate inconsistent samples that could corrupt DNN training. A 19-layer DCNN trained on kinematically sifted ACGAN-based synthetic data was shown to be effective in classifying challenging datasets collected across different environments, as demonstrated by 93% correct classification. In our experiment, test data were collected through-the-wall, while LOS measurements were used to drive the ACGAN in training data generation. This result surpasses other previously proposed approaches, including transfer learning and ACGAN-generated data that is not kinematically sifted.

REFERENCES

[1] F. Ahmad, A. E. Cetin, K. C. D. Ho, and J. Nelson
Signal processing for assisted living: Developments and open problems [from the Guest Editors]
IEEE Signal Process. Mag., vol. 33, no. 2, pp. 25–26, Mar. 2016.

[2] L. Pallotta, C. Clemente, A. De Maio, J. J. Soraghan, and A. Farina
Pseudo-zernike moments based radar micro-doppler classification
In *Proc. IEEE Radar Conf.*, May 2014, pp. 0850–0854.

[3] P. Molchanov, J. Astola, K. Egiazarian, and A. Totsky
Ground moving target classification by using DCT coefficients extracted from micro-Doppler radar signatures and artificial neuron network
In *Proc. Microw., Radar Remote Sens. Symp.*, Aug. 2011, pp. 173–176.

[4] B. G. Mobasseri and M. G. Amin
A time-frequency classifier for human gait recognition
Proc. SPIE, Opt.Photon. Homeland Secur. V Biometric Technol. Human Identification VI, vol. 7306, May 2009.

[5] 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.

[6] Q. Wu, Y. D. Zhang, W. Tao, and M. G. Amin
Radar-based fall detection based on Doppler time-frequency signatures for assisted living
IET Radar Sonar Navigat., vol. 9, no. 2, pp. 164–172, Feb. 2015.

[7] B. Y. Su, K. C. Ho, M. J. Rantz, and M. Skubic
Doppler radar fall activity detection using the wavelet transform
IEEE Trans. Biomed. Eng., vol. 62, no. 3, pp. 865–875, Mar. 2015.

[8] M. Amin
Ed. *Radar for Indoor Monitoring: Detection, Classification, and Assessment*. Boca Raton, FL, USA: CRC Press, Sep. 2017.

[9] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges
Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum
In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, (ser. UIST '16.) New York, NY, USA, 2016, pp. 851–860. [Online]. Available: http://doi.acm.org/10.1145/2984511.2984565

[10] S. Z. Gurbuz, C. Clemente, A. Balleri, and J. J. Soraghan
Micro-Doppler-based in-home aided and unaided walking recognition with multiple radar and sonar systems
IET Radar, Sonar Navigat., vol. 11, no. 1, pp. 107–115, Jan. 2017.

[11] A. Seifert, M. G. Amin, and A. M. Zoubir
New analysis of radar micro-doppler gait signatures for rehabilitation and assisted living
In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 4004–4008.

[12] A. Seifert, M. G. Amin, and A. M. Zoubir
Detection of gait asymmetry using indoor doppler radar
In *Proc. IEEE Radar Conf.*, Apr. 2019, pp. 1–6.

[13] M. S. Seyfioglu, A. M. Ozbayoglu, and S. Z. Gurbuz
Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities
IEEE Trans. Aerosp. Electron. Syst., vol. 54, no. 4, pp. 1709–1723, Aug. 2018.

[14] B. Jokanovic and M. Amin
Fall detection using deep learning in range-doppler radars
IEEE Trans. Aerosp. Electron. Syst., vol. 54, no. 1, pp. 180–189, Feb. 2018.

[15] R. P. Trommel, R. I. A. Harmanny, L. Cifola, and J. N. Driessen
Multi-target human gait classification using deep convolutional neural networks on micro-Doppler spectrograms
In *Proc. Eur. Radar Conf.*, Oct. 2016, pp. 81–84.

[16] Y. Kim and T. Moon
Human detection and activity classification based on micro-doppler signatures using deep convolutional neural networks
IEEE Geosci. Remote Sens. Lett., vol. 13, no. 1, pp. 8–12, Jan. 2016.

[17] 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.

[18] J. Kwon and N. Kwak
Human detection by neural networks using a low-cost short-range Doppler radar sensor
In *Proc. IEEE Radar Conf.*, May 2017, pp. 0755–0760.

[19] M. S. Seyfioglu and S. Z. Gurbuz
Deep neural network initialization methods for micro-Doppler classification with low training sample support
IEEE Geosci. Remote Sens. Lett., vol. 14, no. 12, pp. 2462–2466, Dec. 2017.

[20] Y. Yang, C. Hou, Y. Lang, D. Guan, D. Huang, and J. Xu
Open-set human activity recognition based on micro-doppler signatures
Pattern Recognit., vol. 85, pp. 60–69, Jan. 2019.

[21] J. Ding, B. Chen, H. Liu, and M. Huang
Convolutional neural network with data augmentation for SAR target recognition
IEEE Geosci. Remote Sens. Lett., vol. 13, no. 3, pp. 364–368, Mar. 2016.

[22] Z. Wang, L. Du, J. Mao, B. Liu, and D. Yang
SAR target detection based on SSD with data augmentation and transfer learning
IEEE Geosci. Remote Sens. Lett., vol. 16, no. 1, pp. 150–154, Jan 2019.

[23] V. Chen
The Micro-doppler Effect in Radar (ser. Artech House radar library). Norwood, MA, USA: Artech House, 2011. [Online]. Available: https://books.google.com/books?id=eJ7eMHpxt30C

[24] V. Chen, D. Tahmoush, and W. Miceli
Radar Micro-Doppler Signatures: Processing and Applications (ser. Electromagnetics and Radar). London, U.K.: Institution Eng. Technol., 2014. [Online]. Available: https://books.google.com/books?id=qx_zAwAAQBAJ

[25] S. Sundar Ram and H. Ling
Simulation of human microdopplers using computer animation data
In *Proc. IEEE Radar Conf.*, May 2008, pp. 1–6.

[26] S. S. Ram, C. Christianson, Y. Kim, and H. Ling
Simulation and analysis of human micro-dopplers in through-wall environments
IEEE Trans. Geosci. Remote Sens., vol. 48, no. 4, pp. 2015–2023, Apr. 2010.

[27] B. Erol and S. Z. Gurbuz
A kinect-based human micro-doppler simulator
IEEE Aerosp. Electron. Syst. Mag., vol. 30, no. 5, pp. 6–17, May 2015.

[28] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin
Diversified radar micro-Doppler simulations as training data for deep residual neural networks
In *Proc. IEEE Radar Conf.*, Apr. 2018, pp. 0612–0617.

[29] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath
Generative adversarial networks: An overview
IEEE Signal Process. Mag., vol. 35, no. 1, pp. 53–65, Jan. 2018.

[30] B. Lewis, J. Liu, and A. Wong
Generative adversarial networks for sar image realism
Proc. SPIE, vol. 10647, Apr. 2018.

[31] X. Shi, Y. Li, F. Zhou, and L. Liu
Human activity recognition based on deep learning method
In *Proc. Int. Conf. Radar*, Aug. 2018, pp. 1–5.

[32] Y. Mi, X. Jing, J. Mu, X. Li, and Y. He
Dcgan-based scheme for radar spectrogram augmentation in human activity classification
In *Proc. IEEE Int. Symp. Antennas Propagat. USNC/URSI Nat. Radio Sci. Meeting*, Jul. 2018, pp. 1973–1974.

[33] B. Erol, S. Z. Gurbuz, and M. G. Amin
GAN-based synthetic radar micro-doppler augmentations for improved human activity recognition
In *Proc. IEEE Radar Conf.*, Apr. 2019, pp. 1–5.

[34] I. Alnujaim, D. Oh, and Y. Kim
Generative adversarial networks to augment micro-doppler signatures for the classification of human activity
In *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 9459–9461.

[35] A. Odena, C. Olah, and J. Shlens
Conditional image synthesis with auxiliary classifier GANs
Oct. 2016. [Online]. Available: http://arxiv.org/abs/1610.09585

[36] C. Doersch
Tutorial on Variational Autoencoders
2016. [Online]. Available: https://www.semanticscholar.org/paper/Tutorial-on-Variational-Autoencoders-Doersch/2932c27534879345a1ff9c753c95ac60f8469179

[37] SDR-KIT 2500b | Ancortek
[Online]. Available: http://ancortek.com/sdr-kit-2500b

[38] B. Boashash
Time-Frequency Signal Analysis and Processing: A Comprehensive Reference. San Francisco, CA, USA: Academic, Dec. 2015.

[39] E. Sejdic, I. Djurovic, and J. Jiang
Time frequency feature representation using energy concentration: An overview of recent advances
Digit. Signal Process., vol. 19, no. 1, pp. 153–183, Jan. 2009.

[40] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan
A fast and elitist multiobjective genetic algorithm: NSGA-II
IEEE Trans. Evolu. Comput., vol. 6, no. 2, pp. 182–197, Apr. 2002.

[41] R. M. Narayanan and M. Zenaldin
Radar micro-Doppler signatures of various human activities
IET Radar, Sonar Navigat., vol. 9, no. 9, pp. 1205–1215, Dec. 2015.

[42] B. Erol and M. G. Amin
Radar data cube analysis for fall detection
In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 2446–2450.

[43] T. Salimans *et al.*
Improved techniques for training GANs
In *Advances in Neural Information Processing Systems*, Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2016, pp. 2234–2242.

[44] L. Theis, A. v. d. Oord, and M. Bethge
A note on the evaluation of generative models
Nov. 2015. [Online]. Available: http://arxiv.org/abs/1511.01844

[45] A. Radford, L. Metz, and S. Chintala
Unsupervised representation learning with deep convolutional generative adversarial networks
Nov. 2015. [Online]. Available: http://arxiv.org/abs/1511.06434

[46] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu
Conditional image generation with PixelCNN decoders
Jun. 2016. [Online]. Available: https://arxiv.org/abs/1606.05328

[47] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu
Pixel recurrent neural networks
Jan. 2016. [Online]. Available: https://arxiv.org/abs/1601.06759

[48] I. J. Goodfellow *et al.*
Generative adversarial networks
Jun. 2014. [Online]. Available: http://arxiv.org/abs/1406.2661

[49] I. J. Goodfellow
NIPS 2016 tutorial: Generative adversarial networks
2017, *arXiv:1701.00160*. [Online]. Available: http://arxiv.org/abs/1701.00160

[50] M. Arjovsky, S. Chintala, and L. Bottou
Wasserstein GAN
2017. [Online]. Available: https://www.semanticscholar.org/paper/Wasserstein-GAN-Arjovsky-Chintala/2f85b7376769473d2bed56f855f115e23d727094

[51] K. Sohn, H. Lee, and X. Yan
Learning structured output representation using deep conditional generative models
In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2015, pp. 3483–3491.

[52] D. Kingma and J. Ba
Adam: A method for stochastic optimization
In *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, 2015.

[53] A. Borji
Pros and Cons of GAN evaluation measures
Feb. 2018 [Online]. Available: http://arxiv.org/abs/1802.03446

[54] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli
Image quality assessment: From error visibility to structural similarity
IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[55] B. Erol, M. Francisco, A. Ravisankar, and M. Amin
Realization of radar-based fall detection using spectrograms
Proc. SPIE, vol. 10658, May 2018, Art. no. 106580B.

[56] M. S. Seyfioglu, A. M. Ozbayoglu, and S. Z. Gurbuz
Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities
IEEE Trans. Aerosp. Electron. Syst., vol. 54, no. 4, pp. 1709–1723, Aug. 2018.

[57] K. Simonyan, A. Vedaldi, and A. Zisserman
Deep inside convolutional networks: Visualising image classification models and saliency maps
Dec. 2013. [Online]. Available: http://arxiv.org/abs/1312.6034

Baris Erol (Member, IEEE) received the B.S. degree in electrical and electronics engineering and the M.Sc. degree in electrical engineering from the TOBB University of Economics and Technology, Ankara, Turkey, in 2014 and 2015, respectively, and the Ph.D. degree in electrical and computer engineering from Villanova University, Villanova, PA, USA, in 2018.

He is currently a Research Scientist with the Siemens Corporate Technology in the Automation Runtime Systems research group, Princeton, NJ, USA. He has authored and coauthored more than 20 peer-reviewed articles on the topics of his research interests, which include sensor integration, radar signal processing, and machine learning.

Sevgi Zubyede Gurbuz (Senior Member, IEEE) received the B.S. degree in electrical engineering with minor in mechanical engineering and the M.Eng. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1998 and 2000, respectively, and the Ph.D. degree in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 2009.

From February 2000 to January 2004, she was a Radar Signal Processing Research Engineer with the US Air Force Research Laboratory, Sensors Directorate, Rome, NY, USA. She was an Assistant Professor with the Department of Electrical-Electronics Engineering, TOBB University, Ankara, Turkey and Senior Research Scientist with the TUBITAK Space Technologies Research Institute, Ankara, Turkey. She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Alabama at Tuscaloosa. Her current research interests include radar signal processing, machine learning and pattern recognition, cognitive radar, and sensor networks.

Dr. Gurbuz is a recipient of the 2020 SPIE Rising Researcher Award, EU Marie Curie Research Fellowship, USAF Achievement Medal, USAF Commendation Medal, AFRL Technical Achievement Award, National Defense Science and Engineering Fellowship, C.S. Draper Fellowship, and the 2010 IEEE Radar Conference Best Student Paper Award.

Moeness G. Amin (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Colorado, Boulder, CO, USA, in 1984.

Since 1985, he has been with the Faculty of the Department of Electrical and Computer Engineering, Villanova University, Villanova, PA, USA, where he became the Director of the Center for Advanced Communications, College of Engineering, in 2002. He has authored more than 800 journal and conference publications in signal processing theory and applications. He coauthored 22 book chapters and is the Editor of the three books *Through the Wall Radar Imaging* (CRC Press 2011), *Compressive Sensing for Urban Radar* (CRC Press, 2014), and *Radar for Indoor Monitoring* (CRC Press, 2017).

Dr. Amin is a fellow of the Institute of Electrical and Electronics Engineers, fellow of the International Society of Optical Engineering, fellow of the Institute of Engineering and Technology, and fellow of the European Association for Signal Processing. He is the recipient of the 2017 Fulbright Distinguished Chair in Advanced Science and Technology, 2016 Alexander von Humboldt Research Award, 2014 IEEE Signal Processing Society Technical Achievement Award, 2009 Individual Technical Achievement Award from the European Association for Signal Processing, 2015 IEEE Aerospace and Electronic Systems Society Warren D. White Award for Excellence in Radar Engineering, IEEE Third Millennium Medal, 2010 NATO Scientific Achievement Award, 2010 Chief of Naval Research Challenge Award, Villanova University Outstanding Faculty Research Award, 1997, and IEEE Philadelphia Section Award, 1997. He was a Distinguished Lecturer for the IEEE Signal Processing Society from 2003 to 2004. He was a member and later, became the Chair of the Electrical Cluster of the Franklin Institute Committee on Science and the Arts from 2000 to 2015.