Effect of Kinematics and Fluency in Adversarial Synthetic Data Generation for ASL Recognition with RF Sensors

M.M. Rahman, *Student Member, IEEE*, E. Malaia, A.C. Gurbuz, *Senior Member, IEEE*, D.J. Griffin, C. Crawford and S.Z. Gurbuz, *Senior Member, IEEE*

***Abstract*—RF sensors have been recently proposed as a new modality for sign language processing technology. They are non-contact, effective in the dark, and acquire a direct measurement of signing kinematic via exploitation of the micro-Doppler effect. First, this work provides an in depth, comparative examination of the kinematic properties of signing as measured by RF sensors for both fluent ASL users and hearing imitation signers. Second, as ASL recognition techniques utilizing deep learning requires a large amount of training data, this work examines the effect of signing kinematics and subject fluency on adversarial learning techniques for data synthesis. Two different approaches for the synthetic training data generation are proposed: 1) adversarial domain adaptation to minimize the differences between imitation signing and fluent signing data, and 2) kinematically-constrained generative adversarial networks for accurate synthesis of RF signing signatures. The results show that the kinematic discrepancies between imitation signing and fluent signing are so significant that training on data directly synthesized from fluent RF signers offers greater performance (93% top-5 accuracy) than that produced by adaptation of imitation signing (88% top-5 accuracy) when classifying 100 ASL signs.**

***Index Terms*—radar, micro-Doppler, sign language, ASL, adversarial learning, kinematics**

I. Introduction

According to the World Federation for the Deaf (WFD), an estimated 74 million people world-wide communicate using sign language. American Sign Language (ASL) is estimated to be the primary mode of communication for over a million people in North America and Canada, based on statistics provided by Gallaudet University (the world's only university designed to be barrier-free for deaf and hard of hearing students located in Washington, D.C.). Although much research in ASL recognition has focused on translation (e.g. sign to speech) as a means to bridge the communication gap between Deaf and hearing individuals, this work aims at the development of RF sensing-enabled sign language processing (SLP) technologies in an ambient, non-invasive fashion for human-computer interaction (HCI) applications, such as assistive robots [1], [2] and smart environments [3]. RF sensors have been successfully used for remote health monitoring of vital signs [4], fall detection [5], gait analysis [6], and detection of sleep apnea [7] or sudden infant death syndrome [8]. The addition of ASL recognition capability to such systems would extend their use potential to Deaf populations, and enhance the quality of life for those who use ASL.

RF sensors offer unique advantages in that they are non-contact, not restrictive or invasive, operate at a distance, protect the privacy of the user and personal spaces, and are effective in the dark, regardless of what the individual is wearing. Thus, RF sensors can recognize signing [9], [10] in situations where other sensors, such as wearables [11], [12] or cameras [13], [14], [15], are either undesireable or ineffective. RF sensors cannot perceive hand shapes or facial expressions, but they can provide a direct measure of distance and velocity as a function of time. Velocity can be obtained via the Doppler effect; namely, the principle that the frequency shift incurred in the received RF signal is proportional to the radial velocity of an object in motion. While translational motion results in a central Doppler frequency, *micro-Doppler* [16] refers to the frequency modulations generated about the center frequency that result from vibrations or rotations. As such, the micro-Doppler signature is comprised of unique patterns directly related to the kinematics of the underlying motion, and can serve as a biometric to identify individuals [17], various activities [18], and gestures [19], [20], [21].

Deep learning has enabled great advances in the recognition capabilities for many sensing modalities, including RF sensors [22]. However, deep neural networks (DNNs) require large amounts of data to learn complex underlying data representations. In biomedical applications for human motion recognition, acquisition of adequate sample sizes can be challenging; targeted populations may be mobility-impaired or reluctant to participate. Given that the Deaf are a minority community, its members are highly sought after for involvement in a variety of research studies, and thus may be wary of frequent requests. Inclusion of Deaf researchers is critical for hearing researchers to understand Deaf cultural perspectives and incorporate Deaf experiences and knowledge of the language into all aspects of SLP technology design [23]. Our research team includes

This work was supported in part by the National Science Foundation under Grants 1932547, 1931861 and 1734938. Human studies research was conducted under University of Alabama Institutional Review Board (IRB) Protocol #18-06-1271.

M.M. Rahman and S.Z. Gurbuz are with the University of Alabama, Department of Electrical and Computer Engineering, Tuscaloosa, AL 35487 (email: mrahman17@crimson.ua.edu, szgurbuz@ua.edu).

E. Malaia is with the University of Alabama, Department of Communication Disorders, Tuscaloosa, AL 35487 (e-mail: eamalaia@ua.edu).

A.C. Gurbuz is with Mississippi State University, Department of Electrical and Computer Engineering (e-mail: gurbuz@ece.msstate.edu).

D.J. Griffin is with the University of Alabama, Department of Communication Studies, Tuscaloosa, AL 35487 (e-mail: djgriffin1@ua.edu).

C. Crawford is with the University of Alabama, Department of Computer Science, Tuscaloosa, AL 35487 (e-mail: crawford@ua.edu).

a Child-of-Deaf-Adults (CODA), who is fluent in ASL, and benefits from the involvement of community partnerships with Gallaudet University and the Alabama Institute for the Deaf and Blind (AIDB), who have provided feedback on Deaf-centric design and aided in the recruitment of Deaf participants.

Nevertheless, reliance on extensive amounts of human subject data for training deep models can result in an undue burden on Deaf participants, even if well-intentioned and for the benefit of the community. One way some researchers have addressed the need for signing data has been to utilize ASL learners or imitation signers, e.g. [24], [25], who are more easily recruitable. Imitation signing refers to the process of asking hearing sign-naive participants to replicate the signs shown in any video. However, it can take learners of sign language at least 3 years to produce signs in a manner that is perceived as fluent by fluent signers [26]. Thus, even with "training" sessions to teach participants how to articulate the signs prior to conducting the experiment with imitation signing videos, the production of hearing imitation signers is not comparable to that of fluent signing and may indeed contain significant errors in temporal dynamics and repetitions (which RF sensors easily perceive), as well as hand shape and place of articulation, i.e. position of the sign in space. In our prior work [21], [27], we found that imitation signing and fluent signing occupy different regions in the feature space, enabling machine learning to effectively distinguish between imitation signing and fluent signing.

But does this mean that imitation signing data has no value in the training of DNNs for fluent ASL recognition? Adversarial domain adaptation is an approach that has been utilized in the computer vision community for style transfer [28], [29], [30] and image-to-image translation [31], [32], [33]. One approach to the design of generative adversarial networks (GANs) for domain adaptation is to use Pix2Pix GAN [32] for image-to-image translation based on the conditional GAN, where a target image is generated, conditional on a given input image. In this case, the Pix2Pix GAN changes the loss function so that the generated image is both plausible in the content of the target domain, and is a plausible translation of the input image. Another approach is to minimize the *cycle-consistency loss* [33], [34], [35], which enforces two mappings to be the reverse of each other: $F(G(x)) = x$. As an alternative to the cycle-consistency proposed with CycleGAN [33], TravelGAN [36] has also been proposed, which instead utilizes an additional Siamese network to guide the generator in generating shared semantics, and thus learn mappings between more complex domains that have large differences beyond that of just style or texture. Thus, one approach to training deep models for ASL recognition could be to utilize adversarial domain adaptation to transform the imitation signing signatures to have greater resemblance to fluent signing signatures.

A significant challenge to the application of adversarial learning to RF datasets, however, is that the data supplied to the DNNs are not optical images, but computed images, derived from time-frequency transform of the raw complex received RF signal. Thus, the pixels in an RF micro-Doppler signature bear no relation to geometry, lighting, or perspective. In contrast, it is the kinematics of the human skeleton that determines the frequency profiles revealed in the RF signature. In prior work, we have shown that consideration of kinematics can lead to great gains in DNN training for human activity recognition:

1) Data augmentation via temporal and spatial scaling of the underlying skeletal animation [37] yields much more effective, diversified and statistically independent samples than image-based data augmentation techniques, which can corrupt the signatures and result in physically impossible samples.
2) The classification accuracy for eight different daily human activities was boosted by 10% simply by discarding 9,000 kinematically impossible samples, which were identified as outliers relative to real data samples using Principal Component Analysis (PCA) [38], from a synthetic dataset of 40,000 samples generated using an Auxilliary Conditional GAN (ACGAN).
3) Use of the signature envelope in addition to the signature itself in a multi-branch GAN (MBGAN) architecture was shown [39] to shift the distribution of synthetic human activity data so as to have greater overlap with that of real data, as visualized using t-SNE [40]. Studies of gross motor motion recognition showed that MBGAN offered improved classification accuracy [41].

Thus, an alternative approach to the transformation of imitation signing signatures could be to directly synthesize ASL signatures for training using adversarial learning.

In this work, we investigate the kinematic properties of sign production by fluent signers versus hearing imitation signers using RF sensors, as well as the impact of fluency and sign kinematics (i.e. components of sign phonology [42], [43]) on training DNNs for classification of fluent signing using synthetic signatures that are (a) transformed from imitation signing data, versus (b) directly generated from a small set of real signatures from fluent signers. In Section II, the experimental procedure for acquiring the 100-sign ASL datasets for imitation signers and fluent ASL users is presented. The kinematic and linguistic properties of these signs, as listed in the ASL-LEX2 database [44], are described. In Section III, techniques for estimating these properties from the RF micro-Doppler signature for each sign are presented. In Section IV, the adversarial networks for domain adaptation and design of a 3-branch MBGAN for ASL training data synthesis are detailed. In particular, using the estimators developed in Section III, the degree to which different adversarial networks preserve the salient properties of each sign are quantitatively evaluated and the advantages of embedding kinematics into the GAN architecture are demonstrated. The similarity of transformed imitation signing and synthesized ASL is compared with that of fluent ASL signatures for different database sizes. In Section IV, kinematically deviant signatures are sifted out from the generated data, and the resulting synthetic datasets are used to train DNNs. In this way, the efficacy of the proposed approaches to classyfing a large number of ASL signs while minimizing real human subject data requirements

is demonstrated. The paper concludes in Section V with a discussion of results and future research directions.

II. Experimental RF Datasets

A. RF Sensors and Test Environment

The RF sensor used in this work is a TI IWR1443BOOST 77 GHz - 81 GHz automotive short-range radar (SRR) sensor, which transmits a pulsed, linear frequency modulated continuous wave (FMCW) signal (a.k.a "chirp" signal). The normalized transmitted signal of the FMCW radar is [45]

$$x_{tr}(t) = exp\left\{j2\pi(f_c t + \frac{k}{2}t^2)\right\} \tag{1}$$

where t denotes the fast time within a chirp (a frequency modulation period), $T_s/2 \leq t \leq T_s/2$, f_c and $k = B/T_s$ denote the center frequency and the frequency slope of the chirp, and B and T_s denote the bandwidth and the time duration of the chirp, respectively.

The transmitted signal illuminates an ASL signer who sits 1.5 meters in front of the sensor and signs in ASL. The radar receives backscatter from the moving arms and hands, as well as reflection from static parts of the body and environment. According to geometric diffraction theory [46], when the wavelength of the incident wave is much smaller than the target size, the backscattered returns from the target can be expressed as the superposition of a set of independent scattering centers. Thus, the signal received by the receiver is a weighted summation of time-delayed, frequency-shifted versions of the transmitted signal given by the the superposition of returns from M points on the body [47]. Thus,

$$x_{rec}(t) = \sum_{i=1}^{M} a_i exp\left\{ - j\frac{4\pi f_c}{c}R_{t,i}\right\}, \tag{2}$$

where $R_{t,i}$ is the range to the i^{th} body part at time t, f_c is the transmit center frequency, c is the speed of light, and the amplitude a_i is the square root of the power of the received signal as given by the radar range equation [48]. Thus, RF sensors provide a complex-time series of measurements in the form $x[t] = I[t] + jQ[t]$.

Typically, this data stream is re-shaped into a 2D matrix for each RF receive channel, so that the columns represent *fast-time*, e.g. the analog-to-digital converter samples, and the rows represent *slow-time*, e.g. pulse number. The range (R) between the radar and any scattering point is found from the round-trip travel time (t_d) as $R = ct_d/2$. In an FMCW radar system, the travel time can be found by mixing the received signal with the transmitted signal and filtering out high frequency components to obtain the beat frequency, $f_b = f_t - f_r$, which is the difference in the instantaneous frequencies of transmit and receive signals, f_t and f_r, respectively. Since the chirp rate is $\gamma = B/\tau = f_b/t_d$, the range is found as $R = c\tau f_b/2B$, where τ is the pulse width. The radial velocity of motion, v_r, is given by the Doppler shift, $f_D = 2v_r f_t/c$, which may be found by taking the Fast Fourier Transform (FFT) across pulses for a specific range bin. The significance of these relations is that the range and velocity estimates obtained from RF sensors are independent measurements.

To enable fine grain motion recognition, it is also important for the radar to have sufficient resolution so as to distinguish the motion of the left and right hands, as well as fingers. The minimum interval between two adjacent targets that can be discriminated by the radar in the radial direction is defined as the range resolution [45], and is given by $\Delta r = c/2B$. The velocity resolution determines the minimum difference in velocity that can be perceived by the RF sensor, and, mathematically, is inversely proportional to the coherent processing interval or CPI, during which the target is illuminated. If T_f is the CPI and λ is the wavelength, then the velocity resolution [45], v_{res} is given by $v_{res} = \lambda/2T_f = \lambda/2N_p\tau$, where N_p is the number of pulses transmitted during a CPI.

With a bandwidth of $4GHz$, center frequency of $77GHz$ and a CPI of $40ms$, the RF transmit waveform offers a range resolution of $0.0375m$ and a velocity resolution of $0.0487m/s$. These numbers indicate that this sensor is capable of recognizing fine-grained motion characteristic of ASL.

B. Experimental Design

The data were collected in a laboratory setting, where the sensor was placed on a table at an elevation of 0.91 m from the ground. Participants sat on a chair directly facing a computer monitor, which was placed immediately behind the radar system. The monitor was used to relay prompts indicating the signs to be articulated. The radar system was positioned at a distance of 1.5 meters in front of the participant.

4 fluent ASL signers took part in the IRB-approved study, of whom 2 were Deaf and 2 were CODAs. The experiments included 100 ASL signs, as shown in Table 1, which were selected from the ASL-LEX2 [44] database to include signs of high frequency, but not phonologically related to ensure a diverse dataset in terms of both handshapes and sign kinematics. The participants repeated each sign 5 times. The same experiment was repeated with 12 hearing participants, who did not know sign language. These participants were shown the signs prior to the experiment to familiarize them with the task. During the experiment, immediately prior to recording, the participants were prompted with a video of each sign in isolation, and asked to repeat it. Participants were presented with a random ordering of single signs minimize coarticulation during sign production. A total of 2000 fluent sign samples and 6000 imitation signing samples were collected.

C. RF Data Pre-Processing

The kinematic properties of signing are captured by the frequency modulations in the phase of the received signal. Micro-motions [16], e.g. rotations and vibrations, result in micro-Doppler (μD) frequency modulations centered about the main Doppler shift, which is caused by translational motion. Signing results in a time-varying pattern of micro-Doppler frequencies. Each sign generates its own unique patterns, which can be revealed through time-frequency analysis. The *micro-Doppler signature*, or spectrogram, is found from the square modulus of the Short-Time Fourier Transform (STFT) of the continuous-

TABLE I
LISTING OF THE 100 ASL SIGNS UTILIZED IN EXPERIMENTS.

One-Handed ASL Signs

Words	Strokes	Words	Strokes	Words	Strokes
YOU	2	THANK YOU	3	FINE	3
YES	2	TIRED	3	SOMETHING	3
ME	2	HELLO	3	MOTHER	≥4
HOME	2	OK	3	SEE	≥4
HOLD	2	THIS	3	HOT	≥4
FATHER	2	GOOD	3	BREAKFAST	≥4
MY	2	MUST	3	WATER	≥4
MORNING	2	HE	3	EAT	≥4
THREE	2	TIME	3	OH I SEE	≥4
WRONG	2	BETTER	3	LET ME SEE	≥4
TOILET	2	TOMORROW	3	SOON	≥4
THERE	2	WHY	3	WHERE	≥4
LONG	2	LIKE	3	PLEASE	≥4
I LOVE YOU	3	YOUR	3	SHOULD	≥4
DEAF	3	ONE	3	ALWAYS	≥4
SLEEP	3	DON'T LIKE	3	TABLE	≥4

Two-Handed ASL Signs

Words	Strokes	Words	Strokes	Words	Strokes
BOOK	2	PET	3	PAPER	≥4
MORE	2	MONTH	3	WHAT	≥4
BED	2	GAS	3	TODAY	≥4
HELP	2	AGAIN	3	SCHOOL	≥4
HAVE	2	WEEK	3	COFFEE	≥4
CITY	2	GO	3	NOTHING	≥4
GO AHEAD	2	NIGHT	3	SHOP	≥4
SUMMON	2	TIE UP	3	TECHNOLOGY	≥4
LICENSE	2	CAN	3	WALK	≥4
THRILLED	2	RIGHT	3	COOK	≥4
WANT	3	FAMILY	3	DOSEN'T MATTER	≥4
TIRED	3	KITCHEN	≥4	SHOES	≥4
FRIEND	3	WINTER	≥4	TEACHER	≥4
READ	3	WORK	≥4	MAYBE	≥4
CHANGE	3	TEACH	≥4	EXCITED	≥4
READY	3	CAR	≥4	MONEY	≥4
BRING	3	EVENING	≥4	PEOPLE	≥4
-	-	EXPLANATION	≥4	-	-

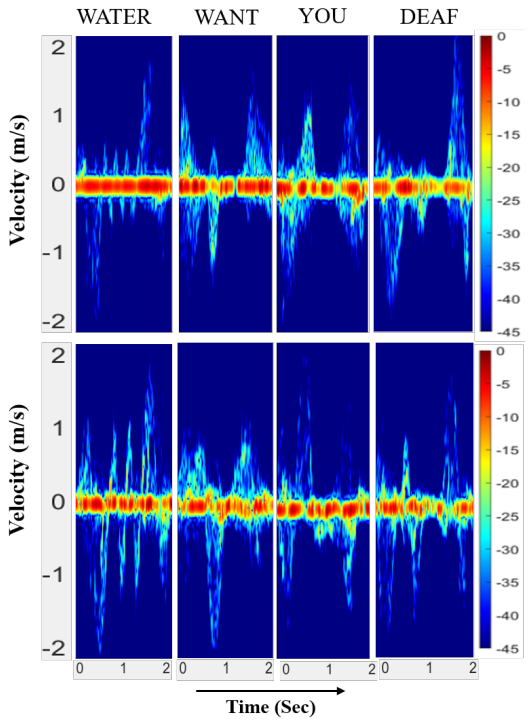


Fig. 1. Micro-Doppler signatures acquired from fluent (row 1) and imitation (row 2) signers.

time input signal $x(t)$ and can be expressed in terms of the window function, $h(t)$, as

$$S(t,\omega) = \left| \int_{-\infty}^{\infty} h(t-u)x(u)e^{-j\omega t}du \right|^2. \quad (3)$$

Ground clutter from stationary objects, such as furniture and the walls, will appear in the micro-Doppler signature as a band centered around 0 Hz. At 77 GHz, elimination of low-speed signal components during clutter filtering results in performance degradation [10], therefore no filtering was applied on the data. Samples of the micro-Doppler signatures for both fluent and imitation ASL users are shown in Fig. 1.

III. ESTIMATION OF SIGNING KINEMATICS

The most relevant kinematic information of a sign can be extracted from the motion associated with the arms and hands. Thus, the prosody of ASL is encoded in the velocity trace of the sign, since this is the simplest interpretation of the motion of the hands and arms of the signer [49]. In RF sensing, the micro-Doppler signature captures this velocity trace since the Doppler frequency is proportional to velocity. In this section, we describe the processing steps taken to extract three kinematic properties of ASL: hand speed, type of signs (one-handed vs. two-handed), and the number of directionally isolatable components of the sign with the motion toward the radar, termed strokes (including transitions toward and from initial and final handshapes).

A. Hand Speed

The speed of signing is measured by tracing the Doppler velocity of the upper and lower envelopes of the micro-Doppler signature. The upper envelope represents the radial velocity of the fastest point moving towards the radar, while the lower envelope gives the speed of the fastest point on the body moving away from the radar. Thus, envelopes provide a means for learning the speed of the hands during signing. Envelopes are extracted by using an energy-based thresholding method [5]. First, the energy corresponding to each slow time index is computed, for which the first frequency bin whose corresponding spectrogram amplitude is greater than or equal to a threshold is tagged as an envelope pixel. The threshold

is computed as the product of a pre-determined scaling factor and the energy at that slow-time index. Figure 2 shows an example of spectrograms for fluent and imitation signers, as well as their corresponding upper and lower envelopes.

The average hand speed during articulation of a sign is calculated by taking an average over each upper and lower envelope extracted from each spectrograms. Figure 3 and 4 show the average hand speed and its standard deviation for fluent and imitations signers for 100 signs. The RF measurements show that the hand speed of imitation signers, on average (0.45 m/s), is greater than that of fluent signers (0.36 m/s), while the standard deviation is much greater. This is reflective of the inconsistency between hearing participants in sign articulations. The greater speed in hand movements of imitation signers may on the one hand seem surprising, as one might think someone less fluent would be more hesitant. But, perhaps in part because hearing imitation signers perceive signing more akin to gesturing, than talking, their articulations are more rushed and sweeping. In contrast, fluent signers articulate the sign within a tighter space, i.e. traverse less distance, but with calculated, precise expression. This results in, on average, slower hand speeds.

Moreover, the RF measurements show that the average hand speed of two-handed signs are greater than that of one-handed signs. This may be in part because two-handed signs typically involve larger movements, while one-handed signs have finer-scale finger movements or hand shapes.

B. Number of Strokes

The number of strokes corresponds to the number of times the hands move towards the radar throughout the duration of the isolated sign (i.e. including transition to the initial handshape, and transition after the final handshape). In other words, number of positive peaks in the micro-Doppler signature correspond to the number of strokes. The number of strokes for a sign can be measured by applying peak detection algorithm to the upper envelope. For repetitive motion (reduplicated signs), such as in signs WALK, WATER, and SHOP, imitation signers were likely to err in production kinematics, producing an incorrect number of strokes (a typical error of early sign language learners). From Figs. 3 and 4, it may be observed that as the number of strokes in a sign increases, hand speed also increases.

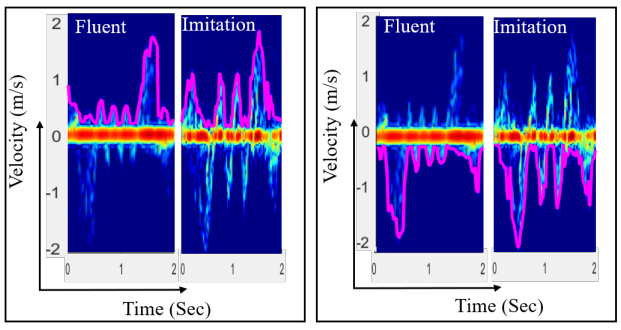


Fig. 2. Spectrograms with upper and lower velocity envelope.

C. One-Handed vs. Two-Handed Signs

Signs in ASL can be one-handed or two-handed. One-handed signs result in less backscatter than two-handed ones. Thus, the average received signal energy for a sign can be indicative of whether the sign involves one or two hands. The total energy of a spectrogram is computed by summing the energy corresponding to each slow-time index. This process is repeated for each spectrogram and then divided by the number of samples to find the algebraic mean. In this way, the total average energy for all 100 words is calculated. Figure 5 compares the average total energy for one-handed and two-handed signs. Note that energy of two-handed signs is distinctly higher than that of one-handed signs. Thus, a threshold can then be designated for categorizing whether a sign is one-handed or two-handed. we found that a threshold of 0.674 yielded a classification accuracy of 81% for one-handed versus two-handed signs.

The kinematic errors of real imitation signers can be quantitatively compared with that of fluent signers using the metrics of average speed (V_h), number of strokes (N_{str}), and handedness detection (N_{h12}), as shown in Table II. Notice that the deviation in average speeds of imitation signers is greater than that of fluent signers. In addition, the errors in repetitions during the articulation of signs, as indicated by the number of strokes, is also significantly higher for imitation signers. This is consistent with the visual observations of fluent and hearing participants during experiments.

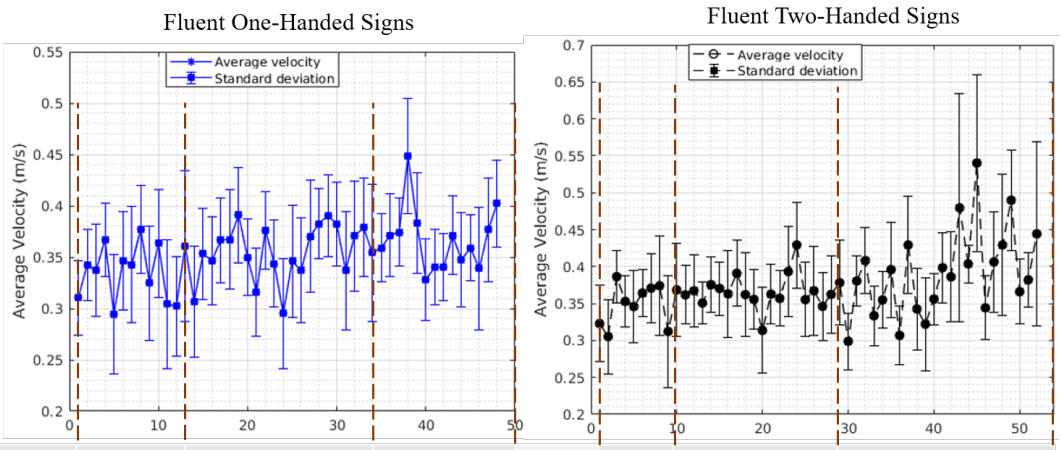
TABLE II
STATISTICS OF REAL FLUENT AND IMITATIONS SIGNER DATA.

Data Source	Number of samples	Avg Speed Deviation *Vh(m/s)*	Average Error (%)	
			N_{str}	N_{h12}
Fluent signers	2000	0.050 ± 0.02	4.27%	0%
Imitation signers	6000	0.15 ± 0.03	22.42%	1%

IV. ADVERSARIAL LEARNING APPROACHES

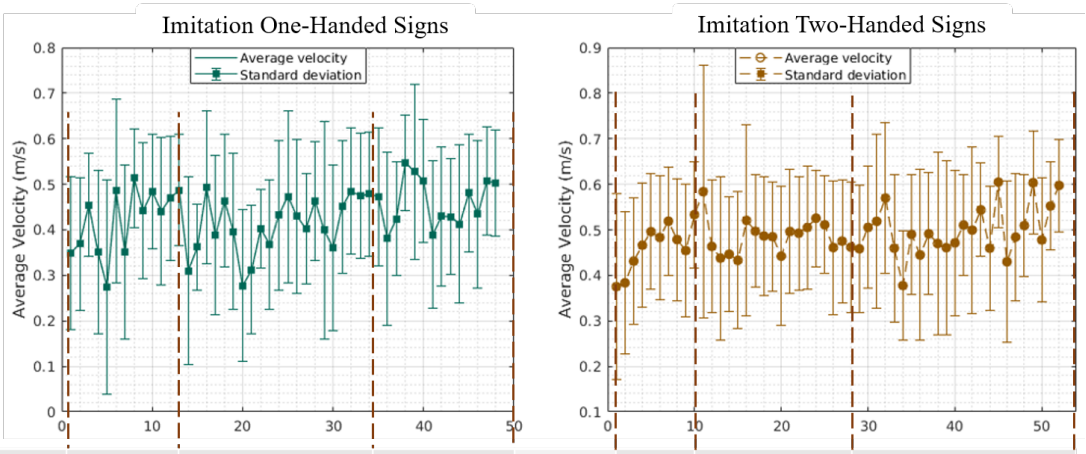
Compilation of large datasets for training state-of-the-art DNNs is difficult when human subjects are involved, due to the time spent in measuring numerous iterations of each class. In previous work [10], ASL recognition using conventional supervised machine learning was explored due to the small amount of available real data: 9 samples per class per sensor, using a total of 5 sensors. The minimum-redundancy maximum-relevance (mRMR) method was used to select 150 handcrafted features for input to a random forest classifier, resulting in a classification accuracy of 72.5% for 20 ASL signs. Later, a slightly larger dataset was acquired (on average of 40 samples per class per sensor for 3 sensors at different frequencies) to train a DNN for fusion of multi-frequency sensor data to achieve an accuracy of 95% for the same 20 ASL signs [50]. The limitations in the amount of available real training data also limited the depth and accuracy of the DNNs utilized.

One approach that has been used in some studies [24], [25] is to instead use imitation signing data for both training and testing of algorithms. However, this can lead to over-optimistic



Strokes #	2	3	≥ 4	…	2	3	≥ 4
Avg Vel (m/s)	0.336	0.355	0.367	…	0.350	0.366	0.391
Avg Std	0.049	0.0473	0.042	…	0.051	0.047	0.061

Fig. 3. Average velocity with standard deviation for one-handed and two-handed fluent ASL signing.



Strokes #	2	3	≥ 4	…	2	3	≥ 4
Avg Vel (m/s)	0.421	0.410	0.460	…	0.461	0.485	0.499
Avg Std	0.157	0.155	0.146	…	0.141	0.147	0.143

Fig. 4. Average velocity with standard deviation for one-handed and two-handed imitation ASL signing.

results [50] due to the differences in production between imitation and fluent signers, which are also captured by the RF sensor measurements as presented in Section III. This is further evidenced by the ability to distinguish between the RF data from fluent versus imitation signers using a support vector machine classifier [21]. Thus, we wish to emphasize that in this work, all DNNs have been tested on ASL signs articulated only by fluent signers.

The question of whether imitation signing data can be leveraged in any way to train DNNs for ASL recognition of fluent signers is an interesting avenue to explore. Due to the differences in data distribution, direct use of imitation signing data as training data is not effective: when a convolutional neural network (CNN) is trained on imitation signing data and tested on fluent ASL-R dataset, only 24% accuracy is attained [50]. One possible remedy could be to use domain adaptation techniques to transform imitation signing data into signatures that better match the distribution of fluent ASL data, as discussed next.

A. *Transformation of Imitation Signing Signatures*

Image translation is a class of computer vision techniques where the goal is to learn a mapping between an input and an output image. A number of image-to-image translation techniques such as Pix2Pix[32], CycleGAN[33] and TravelGAN[36] have been proposed in the literature. As CycleGAN has been shown to outperform TravelGAN on RF signatures [51], in this work we consider the efficacy of both Pix2Pix and CycleGAN for transformation of imitation signing data. The architectures of both techniques are illustrated in Figure 6.

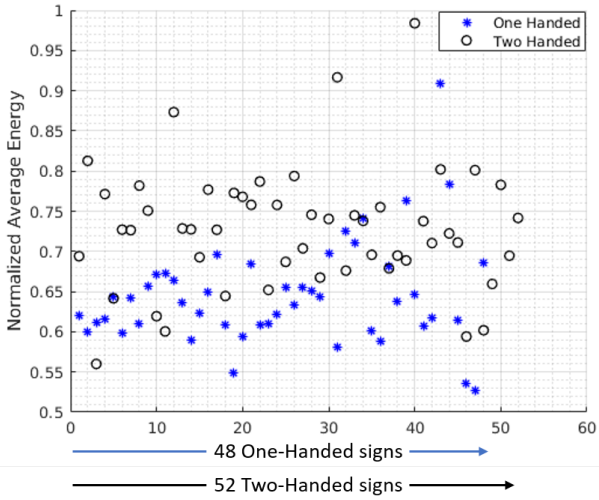


Fig. 5. Energy-based sign type identification.

1) Pix2Pix: Pix2Pix is a type of conditional GAN (cGAN), where the generation of the output image is conditioned on the input; in this case, a source image. The generator of Pix2Pix uses the U-Net [52] architecture. In general, image synthesis architectures take in a random vector as input, project it onto a much higher dimensional vector via a fully connected layer, reshape it, and then apply a series of de-convolutional operations until the desired spatial resolution is achieved. In contrast, the generator of Pix2Pix resembles an auto-encoder. The generator takes in the image to be translated, compresses it into a low-dimensional vector representation, and then learns how to upsample it into the output image. The generator is trained via adversarial loss, which encourages it to generate plausible images in the target domain. The generator is also updated via an ℓ_1-loss measured between the generated image and the expected output image. This additional loss encourages the generator model to create plausible translations of the source image.

The architecture of the discriminator is a PatchGAN / Markovian discriminator [53] that works by classifying individual ($N \times N$) patches in the image as "real vs. fake," as opposed to classifying the entire image. This enforces more constraints that encourage sharp high-frequency detail in the output images. The discriminator is provided both with a source image (in this case, an imitation signing signature) and the target image (fluent signing signature) and must determine whether the target is a plausible transformation of the source image.

One limitation of Pix2Pix is that since it is a paired image-to-image translation method, the total number of synthetic samples generated is identical to the number of real imitation signing signatures acquired. In this work, a total of 6,000 transformed signatures are synthesized using Pix2Pix.

2) CycleGAN: In constrast to Pix2Pix, CycleGAN is a GAN for unpaired image-to-image translation. Thus, a greater amount of synthetic data can be generated than the real imitation samples used at the input of the network. For two domains A and B, CycleGAN learns two mappings: $G{:}A{\rightarrow}B$ and $F{:}B{\rightarrow}A$. CycleGAN translates an image from a source domain A to a target domain B by forming a series connection between two GANs to form a "cycle": the first GAN tries to synthesize "fake fluent" ASL data from the imitation signing data, while the second GAN works to reconstruct the original sample, synthesizing "fake imitation" ASL samples. Thus, the network tries to minimize the cycle consistency loss, i.e. the difference between the input of the first GAN and the output of second GAN.

Each CycleGAN generator is comprised of three sections: an encoder, a transformer, and a decoder. The input image is fed directly into the encoder, which shrinks the representation size while increasing the number of channels. The encoder is composed of three convolution layers. The resulting activation is passed to the transformer, a series of six residual blocks. It is then expanded again by the decoder, which uses two transpose convolutions to enlarge the representation size, and one output layer to produce the final transformed image. The discriminators are comprised of PatchGANs - fully convolutional neural networks that look at a "patch" of the input image, and output the probability of the patch being "real." This is both more computationally efficient than trying to look at the entire input image, and is also more effective since it allows the discriminator to focus on more localized features, like texture.

3) Comparison of Pix2Pix and CycleGAN: Samples of Pix2Pix and CycleGAN transformed signatures are shown in Figure 6 (c). Although the general trends in the signatures are consistent, the Pix2Pix signatures have greater visual resemblance to the signature from fluent ASL users. CycleGAN signatures appear more faded and blurry, especially in regions outside the 0 Hz ground clutter returns. These differences can be quantitatively compared via the kinematic properties of ASL, which can be extracted from RF data as described in Section III. Table III lists the mean error and standard deviation of hand speed, V_h, as well as the percentage of erroneous samples of strokes, N_{str}, and handedness detections, N_h, for the number of synthetic samples, N_s. While Pix2Pix can only transform 6,000 samples, CycleGAN is used to generate both 6,000 and 50,000 samples.

Pix2Pix signatures show better adherence to the kinematic properties of fluent signing than CycleGAN. Notice that on average, the CycleGAN signatures exhibit more error in hand speed, number of strokes and detection of handedness relative to those generated by Pix2Pix. Increasing the number of

TABLE III
COMPARISON OF KINEMATIC ERRORS IN PIX2PIX AND CYCLEGAN SIGNATURES.

Data Source	Number of Samples, *Ns*	Avg Speed Error ± STD, V_h *(m/s)*	Average Error (%)	
			N_{str}	N_{h12}
Pix2Pix	6000	0.19 ± 0.14	6%	2%
CycleGAN	6000	0.26 ± 0.09	11%	7%
	50000	0.28 ± 0.10	18%	9%

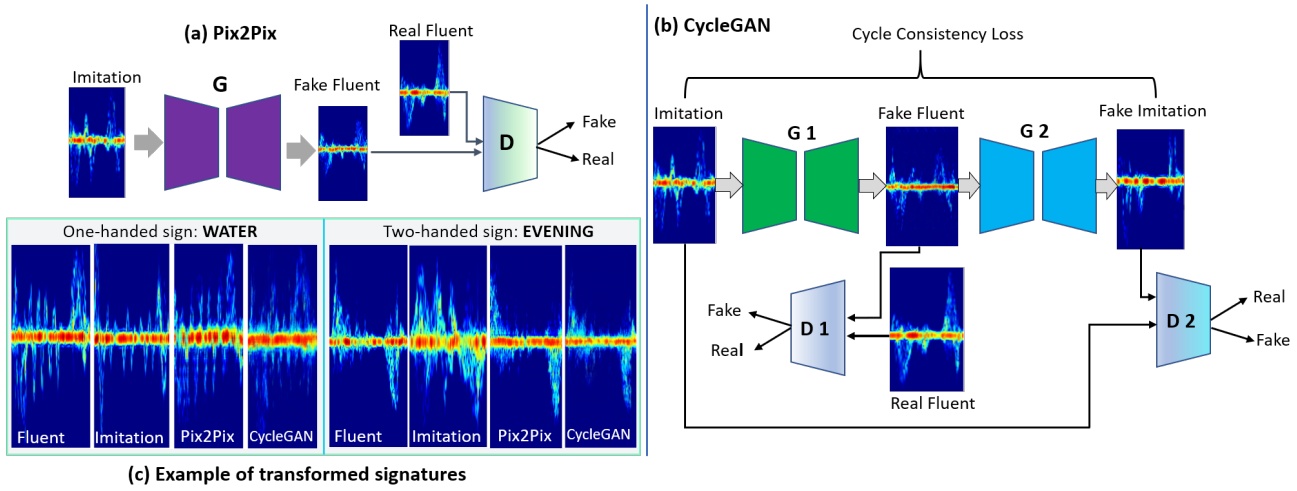


Fig. 6. Imitation to fluent sign transformation using (a) Pix2Pix and (b) CycleGAN. (c) Examples of transformed signatures.

generated CycleGAN samples has only a slight detrimental effect on hand speed, while more significant errors are induced in the number of strokes and handedness.

Note that there are two different causes for kinematic errors: first, the lack of fluency in the language, and second, the network itself. Let us first consider the reasons for why Pix2Pix significantly outperforms CycleGAN. In prior work, we showed that the generative process creates synthetic data with significant kinematic errors [38], which are described more in the next section. Networks generate kinematic errors because RF data is not naturally an image, but converted into a 2D format via time-frequency analysis (Section II-C). Hence, spatial correlations are not based on physical proximity (as in optical images), but on the distribution of velocity across the human body and the constraints imposed by the skeleton. However, the GAN architectures are not supplied with any information or metric pertaining to these constraints, resulting in synthetic samples that bear spatial resemblance, but in fact correspond to physically impossible movement. The CycleGAN architecture includes two generators, in contrast to the single generator of Pix2Pix; hence, the greater the amount of kinematic errors exhibited in the CycleGAN synthesized samples.

Moreover, imitation signing data itself has significantly more error in average speed as well as the number of strokes, as was shown in Table II. That these errors persist in the domain adapted signatures can be seen by observing that the average errors in Pix2Pix and CycleGAN synthesized data remain significantly greater than the levels observed in fluent signing data. In fact, the error in average speed of Pix2Pix data exceeds that observed even in real imitation signing data.

B. Direct Synthesis of ASL Sign Signatures

An alternative to transformation of imitation signatures is to instead use a small amount of real, fluent ASL data as input to a GAN, which generates a larger number of synthetic samples for training. In our prior work [38], [39], several different types of architectures have been explored for synthetic data synthesis, including auxiliary-conditional GANs (ACGANs), conditional variational autoencoders (CVAE) and WGANs, but all were found to generate data that exhibits significant discrepancies from that of real RF signatures. Examples include

- *Disjoint components micro-Doppler:* Real micro-Doppler signatures are connected and continuous, because all points on the human body are connected with each other, forming a continuous spread of velocities. This prevents human RF signatures from having disjoint components or regions in the signature.
- *Leakage between target and non-target components:* A benefit of GANs is that sensor-artifacts can also be synthesized, but sometimes this results in leakage (connected segments) between target movements and sensor artifacts or noise, which are not physically possible.
- *Incorrect shape of signature:* When the shape of the micro-Doppler is distorted, with additional peaks, or symmetric reflections about the x-axis, these components correspond to physically impossible movements; e.g., a person whose hand simultaneous moves towards and away from the radar, additional repetitions, or sudden motion back and forth that are not normally part of the sign.

While these erroneous components may not seem significant visually, they ultimately correspond to kinematically impossible articulations, which, when used as training data, incorrectly trains the DNN and significantly degrades classification accuracy.

One way to mitigate such problems is to design the GAN so as to enable greater emphasis on preservation of the shape of the envelope. The envelopes correspond to the maximum velocity towards/away from the radar; so, from the standpoint of hand kinematics, the synthetic signatures should conform to, and not exceed the envelope profiles of source data. In prior work [39], [41], a multi-branch GAN (MBGAN) architecture with an additional auxiliary branch in the WGAN discriminator, which took as input the upper envelope, was proposed as a means of ensuring kinematic accuracy when synthesizing

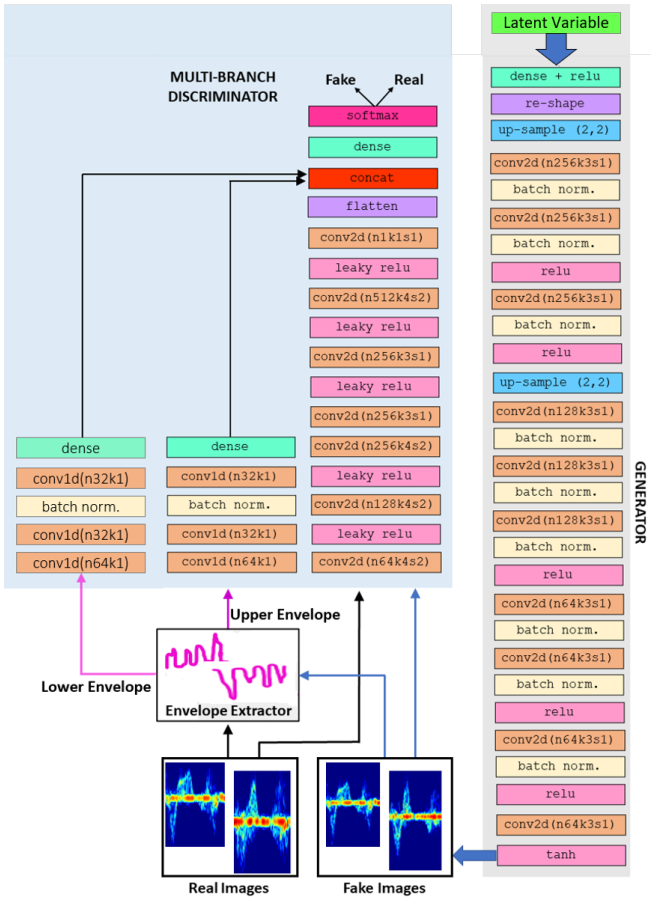


Fig. 7. Proposed 3-branch discriminator MBGAN.

micro-Doppler signatures of different ambulatory gaits, such as walking, limping, or taking short steps. However, during production of sign language, the hands may move towards or away from the radar, so both the upper and lower envelopes are important for maintaining critical kinematic features. Hence, in this work, we incorporated two additional auxiliary branches in the discriminator: one that takes the upper envelope as input, and a second that takes the lower envelope as input. The resulting MBGAN with 3-branch discriminator is shown in Figure 7. The generator is comprised of 10 convolutional layers; each layer is followed by batch normalization with 0.9 momentum and a Rectified Linear Unit (ReLU) activation function. The main branch of the discriminator is an 8-layer CNN, where each layer is followed by a Leaky-ReLU activation function. Each auxiliary branch is comprised of three 1D-convolutional layers. The outputs of the dense layers are concatenated with the flattened output of the main discriminator.

The kinematic errors incurred in the synthetic signatures generated by WGAN and MBGAN are compared in Table IV. The both networks used 75% of the fluent signing data as input during training, and the remaining are used for testing. Notice that the synthetic training data generated by WGAN or MBGAN outperform both Pix2Pix and CycleGAN with respect to generating signatures that have greater kinematic fidelity to fluent signing data. While the kinematic errors in WGAN generated signatures increase as the network generates

TABLE IV
COMPARISON OF KINEMATIC ERRORS IN WGAN AND MBGAN SIGNATURES.

Data Source	Number of Samples, *Ns*	Avg Speed Error ± STD, V_h *(m/s)*	Average Error (%)	
			N_{str}	N_{h12}
WGAN	6000	0.13 ± 0.11	0.047%	1%
	50000	0.16 ± 0.12	2.00%	2%
MBGAN	6000	0.09 ± 0.09	0.035%	1%
	50000	0.09 ± 0.08	1.80%	1%

a greater amount of synthetic samples, the errors in MBGAN signatures are fewer, and constant over sample size - only a slight drop in the accuracy in replicating the correct number of strokes is incurred, from %99.5 to %98. Visually, MBGAN signatures may be observed to have greater resemblance to fluent ASL samples in comparison with the WGAN samples, as shown in Figure 8. Note that in comparison, peaks in WGAN signatures are not as clearly constructed, slightly faded, and have envelopes whose shape has some differences from the envelope of the fluent ASL signature.

Signatures with kinematic errors may divert the classifier in the wrong direction during the feature extraction, and thereby result in poor recognition performance. Hence, it is important to identify and exclude the incorrect kinematic signatures generated in GANs synthesis. In the next section, several kinematic rules are defined and the synthesized data are sifted by these constraints.

C. Kinematic Sifting

Although the 3-branch discriminator MBGAN does have the intended effect of generating signatures with greater kinematic fidelity to fluent ASL, relative to the other networks considered, it is possible that it still generates kinematically unrealistic synthetic samples. Ideally, we wish to generate training data that is statistically independent, diverse, and representative of the range of potential variations within possible articulations of each sign. The presence of kinematically erroneous samples can have a corrupting effect that leads to confusion between different signs. Thus, we seek to remove such samples.

One way of removing outliers is to determine a boundary in feature space based on the measured, fluent ASL data acquired. Using Principal Component Analysis (PCA) or t-SNE [40], each sample can be projected to a 3-D feature space and a convex hull encompassing all samples computed. The convex hull thus forms a boundary; any synthetic samples lying beyond this boundary could be excluded from the training dataset as "erroneous." However, it is still possible for samples within the convex hull boundary to be kinematically flawed. Instead of relying on the PCA-based convex hull, instead we identify and sift flawed synthetic data based on the following kinematic properties:

- ***Rule 1 - Number of strokes:*** The number of strokes estimated using the peak detection algorithm described in Section III is compared with the number of strokes listed for each sign in Table I, as given by ASL-LEX. If

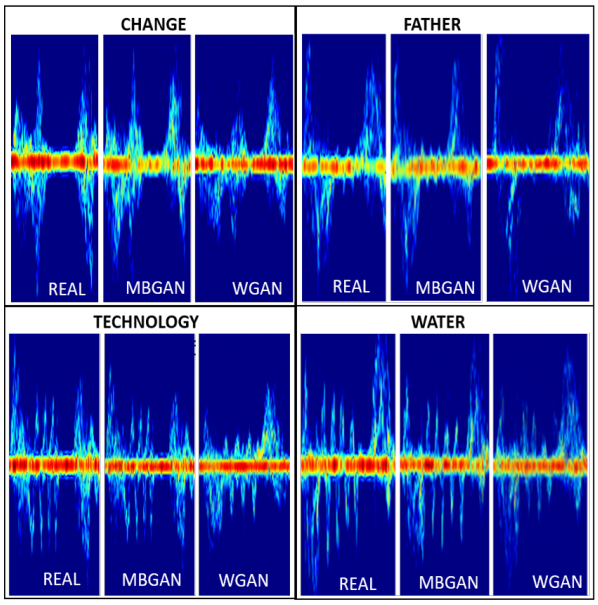


Fig. 8. Example of WGAN and MBGAN generated spectrograms.

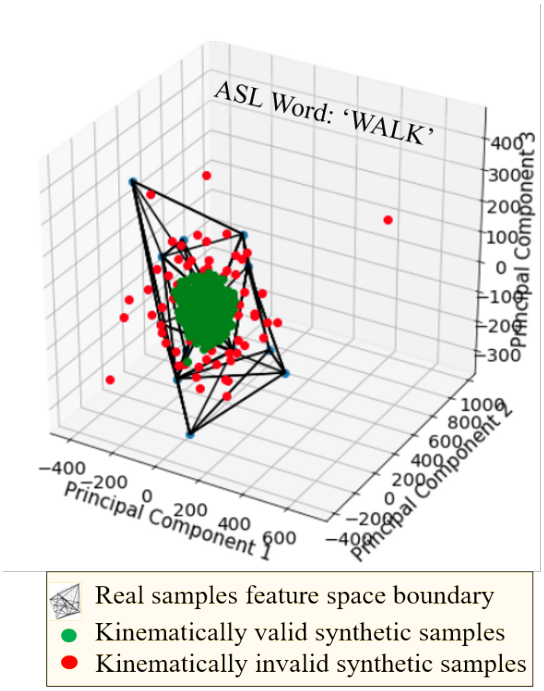


Fig. 9. Kinematically sifted synthetic samples projected on real sample's feature space boundary.

the detected number of strokes for a synthetically generated signature is incorrect, then this synthetic sample is removed from the dataset. For example, this rule will preclude signatures corresponding to a signer utilizing an incorrect number of repetitions.

- ***Rule 2 - Total Energy:*** As energy is related to whether a sign is one handed or two handed, ensuring the synthetic data lies within reasonable energy bounds, given that we know whether the sign is one or two handed, can be an effective criterion. The rule is tested by first finding the average total energy and its standard deviation from the real, fluent ASL data. Then, for each synthetic signature, the total energy is calculated and checked to see whether it falls within $\pm$ 1 standard deviation of the average total energy of the real signatures. This is tested on a class-by-class basis. If the criterion holds true then the sample is regarded as kinematically valid, otherwise it is sifted out.
- ***Rule 3 - Envelope Matching:*** The envelope of the spectograms is a time-series curve and the similarity between curves can be measured by taking into account both the location and ordering of the points along the curve [54]. Dynamic Time Warping (DTW) [55] is a commonly used curve matching technique that measures the similarity between two temporal sequences. To apply envelope matching as a kinematic metric , first, for each class, the average DTW distance and standard deviation are calculated from the combinations of all real samples. Then the DTW distance for each synthetic samples are computed with respect to each real samples, from which the average distance is found. Then this average distance is examined whether it falls within $\pm$1 Standard deviation of the average DTW distance for that class. If it is within the limit then the sample is kinematically valid; otherwise sifted out as kinematically invalid.

Kinematic sifting provides for tighter constraints than the convex hull boundary. Consider synthetic samples for the ASL sign WALK projected to a 3D space using t-SNE, as shown in Figure 9. The boundary found based on the convex hull derived from the real fluent ASL samples is shown with solid black lines. The synthetic samples are shown as dots. Notice that most fall within the convex hull boundary, while some are outliers. With statistical sifting, only the outliers outside this boundary would be sifted out of the training dataset. However, if we apply the kinematic rules described above, we may see that there are many kinematically invalid samples (shown in red) that remain within the hull. The valid samples (shown in green) form a tight nucleus within the convex hull. Hence, the kinematic rules form a more stringent constraint.

Table V provides a listing of the number of synthetic samples sifted out, N_{sft}, by kinematic rules for the Pix2Pix, CycleGAN, WGAN, and MBGAN networks, and the kinematic errors based on the sifted synthetic datasets. For all synthesis methods, comparison with error metrics reported in Tables II and III shows that the sifting process reduced the average error in the remaining data. As the number of samples generated increases, kinematic errors only slightly increase. CycleGAN appeared to be the network most prone to errors, with the greatest number of samples failing the kinematic rules, and, hence, was excluded from the final synthetic dataset. In contrast, even after sifting, the proposed MBGAN remains the network that results in synthetic signatures that exhibit the greatest kinematic fidelity to fluent signing data.

V. ASL Recognition Results

In this section, the resulting classification accuracies obtained using the various methods for synthesizing training data

TABLE V
COMPARISON OF KINEMATIC ERRORS IN SIGNATURES AFTER SIFTING.

Data Source	Number of Samples, *Ns*	Sifted Out Samples, N_{sft}	Avg Speed Error ± STD, V_h *(m/s)*	Average Error (%) N_{h12}
Pix2Pix	6000	354	0.18 ± 0.12	1%
CycleGAN	6000	1321	0.25 ± 0.08	6%
	50000	11586	0.26 ± 0.09	8%
WGAN	6000	483	0.12 ± 0.10	1%
	50000	3933	0.14 ± 0.10	1%
MBGAN	6000	221	0.07 ± 0.08	1%
	50000	1654	0.08 ± 0.075	0%

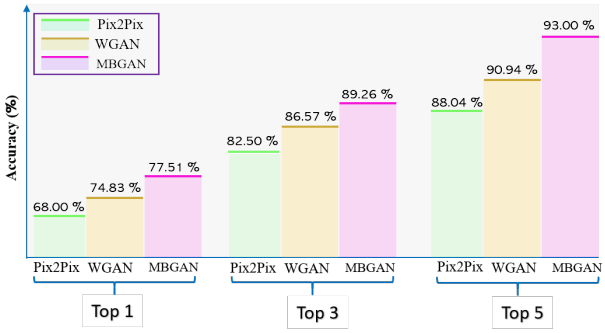


Fig. 10. Comparison of accuracies attained with Pix2Pix, WGAN, and proposed MBGAN methods for synthesizing training data.

are compared.

A. DNN Architectures

In previous work [56], Convolutional Auto-encoders (CAEs) were shown to be effective when small, yet reasonable, amounts of real data are available for training, outperforming transfer learning from weights pre-trained using ImageNet [57] for VGG[58] and Resnet [59]. Consequently, in this work, a four-layer convolutional autoencoder (CAE) has been utilized to classify the 100-sign fluent ASL dataset. CAEs use unsupervised pre-training to initialize the network near a good local minima. In each layer, a filter concatenation technique is employed, in which a filter size of 3×3 and 9×9 were concatenated to take advantage of multilevel feature extraction. After training the CAE model, the decoder was removed, and two fully connected layers with 256 neurons followed by a dropout of 0.55 were added after flattening the output of the encoder. At the output, a softmax layer with 100 nodes was employed for classification. During training, an ADAM [60] optimizer was utilized, along with a batch size of 16, learning rate of 0.0005 and 30 epochs.

B. Classification Accuracy

The classification accuracies obtained using the CAE trained on the various sources of synthetic data are compared in Table VI, while the best performing techniques are compared in Figure 10 based on the Top-1, Top-3, and Top-5 accuracies. The proposed approach of direct training data synthesis with MBGAN surpasses other conventional approaches by achieving a 77% top-1 accuracy, 89% top-3 accuracy, and 93% top-5 accuracy.

C. Implications and Discussion

The most important conclusion we may draw from these results is based on the observation that the resulting classification accuracies are inversely related to the amount of kinematic errors in the synthetic data. The greater the error, the lower the classification accuracy. For all methods, sifting out samples that fail the kinematic rules results in performance improvement.

As mentioned earlier in the paper, there are two sources of errors: namely, kinematic errors generated by DNNs used

TABLE VI
100 ASL SIGNS RECOGNITION USING CAE.

Data Source	Accuracy (%) *(No Kinematic Sifting)*			Accuracy (%) *(With Kinematic Sifting)*		
	Top -1	Top -3	Top -5	Top-1	Top -3	Top-5
Real Data	56.35	69.12	73.54	N/A	N/A	N/A
Pix2Pix	67.78	82.21	87.92	68.00	82.50	88.04
CycleGAN	61.41	75.17	81.54	63.38	79.85	84.33
WGAN	72.76	85.00	90.23	74.83	86.57	90.94
MBGAN	75.84	88.15	92.28	77.51	89.26	93.00

for adaptation and synthesis, and kinematic errors inherent to the original source data. The direct synthesis approach with GANs have the benefit of utilizing fluent signer data in the synthesis process. In contrast, the data synthesized via domain adaptation contains both sources of errors. Note that the error in average speed reflected in Pix2Pix synthesized samples is $0.19m/s$, which is greater than that computed from the real data from imitation signers ($0.09m/s$), Pix2Pix's source data. In other words, Pix2Pix cannot compensate for the imitation signing errors, and exhibits additional model-based errors as well.

In comparing Pix2Pix results with that of WGAN, readers should be reminded that domain adaptation methods predominantly utilize a PatchGAN architecture in the discriminator, which operates on localized patches in the image, while WGAN and MBGAN discriminators operate on the entire image. Both results in the literature [32] and comparisons we conducted on radar micro-Doppler signatures reveal that operations on patches are more effective than that on the entire image. For example, Pix2Pix with a PatchGAN discriminator generates much crisper and textured synthetic signatures than Pix2Pix with a discriminator operating on the entire image. This is likely because discriminators operating on the entire image cannot model the sharpness of high frequency components in the image as effectively. Modeling high frequencies requires restricting attention to the structure in local image patches through the application of penalties at a patch-scale. Despite utilization of the entire image, rather than patches, in the discriminator, WGAN synthesized signatures exhibited fewer kinematic errors than the domain adaptation networks

we considered.

Although the error in average speed in WGAN signatures ($0.13m/s$) is lower than that of Pix2Pix, it is greater than that of imitation signing. These errors are due to the generation process itself, and can be mitigated through modification of GAN architecture, such as done in the proposed MB-GAN. In future work, we plan to explore extensions of the proposed approach (e.g. modifications of GAN architecture and inclusion of envelopes as an auxiliary input, as well as modifications to the loss function to include physics-based loss regularization [41]) to adversarial domain adaptation to improve the resulting classification accuracy when imitation signing data is leveraged for model training.

Another important open question for future work that relates to Explainable AI is to better understand the physical interpretation, i.e. underlying kinematic model, and nature of the diversity seen in GAN-synthesized micro-Doppler signatures. For example, do the variations between synthetic samples correspond to plausible variations within a certain subject profile (physical or linguistic), or span all probable articulations within a class? Improvements to the generation of synthetic data for training will require a better understanding of not just the statistical properties, but also the physical and linguistic properties of the synthetic samples to ensure good model generalization.

VI. Conclusion

Although imitation signing has been used in some studies of sign language recognition, imitation signers exhibit significant differences in kinematics of sign production as compared with fluent signers. This results in substantial statistical differences between imitation and fluent ASL data, which has rendered imitation data ineffective when used to train DNNs for fluent signing recognition [50]. This work investigates the use of domain adaptation to transform imitation signing samples to have greater resemblance to fluent signing data, and compares the efficacy of this approach with direct generation of synthetic data from fluent signing data. A novel approach to synthetic RF signature generation is proposed, which is shown to generate samples with greater kinematic fidelity than conventional GANs for transformation of imitation signing samples. Proposed kinematic metrics are extracted from RF ASL signatures and used to evaluate GAN-generated synthetic data from a kinematic perspective. The classification results obtained using a CAE were found to be directly proportionate to the kinematic fidelity of the synthetic data. The proposed methods were used to achieve 77% top-1 accuracy, 89% top-3 accuracy, and 93% top-5 accuracy for the recognition of 100 ASL signs.

Acknowledgment

The authors would like to thank Dr. Caroline Kobek-Pezzarossi from Gallaudet University, Washington D.C. and Dr. Dennis Gilliam from AIDB for their support of this research.

M. Mahbubur Rahman received the B.S. degree in Electronics and Communication Engineering from Khulna University of Engineering and Technology (KUET), Bangladesh, in 2016. He is currently a Ph.D. student in Electrical and Computer Engineering at the University of Alabama (UA), Tuscaloosa, AL, USA, and a research assistant in the UA Laboratory of Computational Intelligence for Radar (CI4R). His research interests include radar signal processing, machine learning, and multi-modal sensing for fall detection and gait analysis, vehicular autonomy and human-computer interaction.

M.M. Rahman is a recipient of the UA Graduate Council Fellowship in September 2019 and 3rd place in the Best Student Paper Competition of the IEEE Radar Conference in April 2021.

Evie A. Malaia received her Ph.D. degree in Computational Linguistics from Purdue University, West Lafayette, in 2004.

Formerly a Research Scientist at Indiana University and Purdue University, and an Assistant Professor at the University of Texas at Arlington, she is currently an Associate Professor at the University of Alabama at Tuscaloosa, Department of Communicative Disorders. Her current research interests include neural and physical bases of sign language communication, classification of higher cognitive states, and neural bases of autism spectrum disorders.

Dr. Malaia is a recipient of the Ralph E. Powe Award from DOE/ORAU, EurIAS Research Fellowship, EU Marie Curie Senior Research Fellowship, and the APS Award for Teaching and Public Understanding of Psychological Science.

Ali Cafer Gurbuz received B.S. degree from Bilkent University, Ankara, Turkey, in 2003, in Electrical Engineering, and the M.S. and Ph.D. degrees from Georgia Institute of Technology, Atlanta, GA, USA, in 2005 and 2008, both in Electrical and Computer Engineering. From 2003 to 2009, he researched compressive sensing based computational imaging problems at Georgia Tech. He held faculty positions at TOBB University and University of Alabama between 2009 and 2017 where he pursued an active research program on the development of sparse signal representations, compressive sensing theory and applications, radar and sensor array signal processing, and machine learning. Currently, he is an Assistant Professor at Mississippi State University, Department of Electrical and Computer Engineering, where he is co-director of Information Processing and Sensing (IMPRESS) Lab.

Dr. Gurbuz is the recipient of The Best Paper Award for Signal Processing Journal in 2013 and the Turkish Academy of Sciences Best Young Scholar Award in Electrical Engineering in 2014. He has served as an associate editor for several journals such as Digital Signal Processing, EURASIP Journal on Advances in Signal Processing and Physical Communications.

Darrin J. Griffin received the B.S. degree in communication sciences and disorders with a focus on deaf education and the M.A. degree in communication studies in 2004 and 2007, respectively, from The University of Texas at Austin. The Ph.D. degree was completed at The University at Buffalo, SUNY in 2010 in communication with a focus on deceptive communication.

From August 2010 to current he has served as a faculty member at The University of Alabama, Department of Communication Studies where he currently teaches and conducts research as an associate professor on topics related to nonverbal communication, deceptive communication, and deafness. Dr. Griffin is fluent in American Sign Language and participates in various forms of community engagement with the Deaf community.

Dr. Griffin is recipient of the 2020 College of Communication and Information Sciences Board of Visitors Research Excellence Award; the 2018 President's Faculty Research Award at The University of Alabama; and a 2018 Premiere Award from The University of Alabama Council on Community-Based Partnerships for research that raised weather awareness and preparedness for the Deaf & hard of hearing community.

Chris S. Crawford received the Ph.D. degree in human-centered computing from the University of Florida, Gainesville, FL, USA. He is currently an Assistant Professor at the University of Alabama's Department of Computer Science. He directs the Human-Technology Interaction Lab (HTIL). He has investigated multiple systems that provide computer applications and robots with information about a user's cognitive state. In 2016, he lead the development of a BCI application that was featured in the world's first multiparty brain-drone racing event. His current research focuses on computer science education, human-robot interaction, and brain-computer interfaces.

Sevgi Z. Gurbuz (S'01–M'10–SM'17) 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 worked as a Radar Signal Processing Research Engineer with the U.S. Air Force Research Laboratory, Sensors Directorate, Rome, NY, USA. Formerly an Assistant Professor in the Department of Electrical-Electronics Engineering at TOBB University, Ankara, Turkey and Senior Research Scientist with the TUBITAK Space Technologies Research Institute, Ankara, Turkey, she is currently an Assistant Professor in the University of Alabama at Tuscaloosa, Department of Electrical and Computer Engineering. Her current research interests include physics-aware machine learning, RF sensor-enabled cyber-physical systems, radar signal processing, sensor networks, human motion recognition for biomedical, automotive autonomy, and human-computer interaction (HCI) applications.

Dr. Gurbuz is a recipient of the IEEE Harry Rowe Mimno Award for 2019, 2020 SPIE Rising Researcher Award, EU Marie Curie Research Fellowship, and the 2010 IEEE Radar Conference Best Student Paper Award.

REFERENCES

[1] J. Fasola and M. J. Mataric, "Using socially assistive human–robot interaction to motivate physical exercise for older adults," *Proceedings of the IEEE*, vol. 100, no. 8, pp. 2512–2526, 2012.

[2] N. Céspedes, M. Múnera, C. Gómez, and C. A. Cifuentes, "Social human-robot interaction for gait rehabilitation," *IEEE Trans on Neural Systems and Rehabilitation Engg.*, vol. 28, no. 6, pp. 1299–1307, 2020.

[3] R. Zhang, S. He, X. Yang, X. Wang, K. Li, Q. Huang, Z. Yu, X. Zhang, D. Tang, and Y. Li, "An EOG-Based human–machine interface to control a smart home environment for patients with severe spinal cord injuries," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 1, pp. 89–100, 2019.

[4] J. Muñoz-Ferreras, Z. Peng, R. Gómez-García, and C. Li, "Review on advanced short-range multimode continuous-wave radar architectures for healthcare applications," *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology*, vol. 1, no. 1, pp. 14–25, 2017.

[5] M. G. Amin, Y. D. Zhang, F. Ahmad, and K. C. D. Ho, "Radar signal processing for elderly fall detection: The future for in-home monitoring," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 71–80, 2016.

[6] A. K. Seifert, M. G. Amin, and A. M. Zoubir, "Toward unobtrusive in-home gait analysis based on radar micro-doppler signatures," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 9, pp. 2629–2640, 2019.

[7] Y. S. Lee, P. N. Pathirana, C. L. Steinfort, and T. Caelli, "Monitoring and analysis of respiratory patterns using microwave doppler radar," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 2, pp. 1–12, 2014.

[8] E. G. Ziganshin, M. A. Numerov, and S. A. Vygolov, "UWB baby monitor," in *2010 5th International Confernce on Ultrawideband and Ultrashort Impulse Signals*, 2010, pp. 159–161.

[9] S. Gurbuz, A. Gurbuz, C. Crawford, and D. Griffin, "Radar-based methods and apparatus for communication and interpretation of sign languages," in *U.S. Patent Application No. US2020/0334452 (Invention Disclosure filed Feb. 2018; Provisional Patent App. filed Apr. 2019.)*, October 2020.

[10] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. S. Crawford, M. M. Rahman, E. Kurtoglu, R. Aksu, T. Macks, and R. Mdrafi, "American sign language recognition using rf sensing," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3763–3775, 2021.

[11] V. E. Kosmidou and L. J. Hadjileontiadis, "Sign language recognition using intrinsic-mode sample entropy on semg and accelerometer data," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 12, pp. 2879–2890, 2009.

[12] Y. Li, X. Chen, X. Zhang, K. Wang, and Z. J. Wang, "A sign-component-based framework for chinese sign language recognition using accelerometer and semg data," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 10, pp. 2695–2704, 2012.

[13] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2306–2320, 2020.

[14] C. Sun, T. Zhang, B. Bao, C. Xu, and T. Mei, "Discriminative exemplar coding for sign language recognition with kinect," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1418–1428, 2013.

[15] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri, "A modified lstm model for continuous sign language recognition using leap motion," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 7056–7063, 2019.

[16] V. Chen, *The Micro-Doppler Effect in Radar, 2nd Ed.* Boston: Artech House, 2019.

[17] B. Vandersmissen, N. Knudde, A. Jalalvand, I. Couckuyt, A. Bourdoux, W. De Neve, and T. Dhaene, "Indoor person identification using a low-power FMCW radar," *IEEE Trans. on Geoscience and Remote Sensing*, vol. PP, pp. 1–12, 04 2018.

[18] A. Sevgi Z. Gurbuz, Carmine Clemente and John J. Soraghan, "Micro-doppler-based in-home aided and unaided walking recognition with multiple radar and sonar systems," *IET Radar, Sonar & Navigation*, vol. 11, pp. 107–115(8), January 2017.

[19] A. Arbabian, S. Callender, S. Kang, M. Rangwala, and A. Niknejad, "A 94 GHz mm-wave-to-baseband pulsed-radar transceiver with applications in imaging and gesture recognition," *Solid-State Circuits, IEEE Journal of*, vol. 48, pp. 1055–1071, 04 2013.

[20] Z. Wang, Z. Yu, X. Lou, B. Guo, and L. Chen, "Gesture-radar: A dual doppler radar based system for robust recognition and quantitative profiling of human gestures," *IEEE Trans on Human-Machine Systems*, vol. 51, no. 1, pp. 32–43, 2021.

[21] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. S. Crawford, M. M. Rahman, E. Kurtoglu, R. Aksu, T. Macks, and R. Mdrafi, "American sign language recognition using rf sensing," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3763–3775, 2021.

[22] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 16–28, 2019.

[23] D. Bragg and e. Koller, "Sign language recognition, generation, and translation: An interdisciplinary perspective," in *The 21st Int. ACM SIGACCESS Conference on Computers and Accessibility*, 2019, p. 16–31.

[24] B. Fang, J. Co, and M. Zhang, "Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, 2017.

[25] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "Signfi: Sign language recognition using wifi," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, 2018.

[26] J. S. Beal and K. Faniel, "Hearing l2 sign language learners: How do they perform on asl phonological fluency?" *Sign Language Studies*, vol. 19, no. 2, pp. 204–224, 2018.

[27] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. Crawford, M. M. Rahman, R. Aksu, E. Kurtoglu, R. Mdrafi, A. Anbuselvam, T. Macks, and E. Ozcelik, "A linguistic perspective on radar micro-doppler analysis of american sign language," in *2020 IEEE International Radar Conference (RADAR)*, 2020, pp. 232–237.

[28] R. Li, C. H. Wu, S. Liu, J. Wang, G. Wang, G. Liu, and B. Zeng, "Sdp-gan: Saliency detail preservation generative adversarial networks for high perceptual quality style transfer," *IEEE Trans. on Image Processing*, vol. 30, pp. 374–385, 2021.

[29] X. Gao, Y. Tian, and Z. Qi, "Rpd-gan: Learning to draw realistic paintings with generative adversarial network," *IEEE Trans on Image Processing*, vol. 29, pp. 8706–8720, 2020.

[30] X. Pan, M. Zhang, D. Ding, and M. Yang, "A geometrical perspective on image style transfer with adversarial learning," *IEEE Trans. on PAMI*, pp. 1–1, 2020.

[31] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *NIPS*, 2017.

[32] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *IEEE CVPR*, pp. 5967–5976, 2017.

[33] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *IEEE ICCV*, pp. 2242–2251, 2017.

[34] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *ICML*, 2017.

[35] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," *IEEE ICCV*, pp. 2868–2876, 2017.

[36] M. Amodio and S. Krishnaswamy, "Travelgan: Image-to-image translation by transformation vector learning," *IEEE CVPR)*, pp. 8975–8984, 2019.

[37] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "DNN transfer learning from diversified micro-doppler for motion classification," *IEEE Trans on AES*, vol. 55, no. 5, pp. 2164–2180, 2019.

[38] B. Erol, S. Z. Gurbuz, and M. G. Amin, "Motion classification using kinematically sifted acgan-synthesized radar micro-doppler signatures," *IEEE Trans on AES*, vol. 56, no. 4, pp. 3197–3213, 2020.

[39] ——, "Synthesis of micro-doppler signatures for abnormal gait using multi-branch discriminator with embedded kinematics," in *IEEE Radar Conf.*, 2020, pp. 175–179.

[40] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[41] M. M. Rahman, S. Z. Gurbuz, and M. G. Amin, "Physics-aware design of multi-branch gan for human rf micro-doppler signature synthesis," in *IEEE Radar Conf.*, 2021, pp. 1–6.

[42] E. Malaia and R. B. Wilbur, "Kinematic signatures of telic and atelic events in asl predicates," *Language and speech*, vol. 55, no. 3, pp. 407–421, 2012.

[43] E. A. Malaia and R. B. Wilbur, "Syllable as a unit of information transfer in linguistic communication: The entropy syllable parsing model," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 11, no. 1, p. e1518, 2020.

[44] N. Caselli, Z. Sehyr, A. Cohen-Goldberg, and K. Emmorey, "Asl-lex: A lexical database of american sign language," *Behavior Research Methods*, vol. 49, 05 2016.

[45] M. Jankiraman, B. J. Wessels, and P. van Genderen, "Design of a multi-frequency fmcw radar," in *1998 28th European Microwave Conference*, vol. 1, 1998, pp. 584–589.

[46] J. B. Keller, "Geometrical theory of diffraction∗," *J. Opt. Soc. Am.*, vol. 52, no. 2, pp. 116–130, 1962.

[47] P. van Dorp and F. Groen, "Human walking estimation with radar," *IET Radar, Sonar and Navigation*, vol. 150, pp. 356–365(9), 2003.

[48] M. Richards. McGraw-Hill Education, 2014.

[49] R. Wilbur and A. M. Martnez, "Physical correlates of prosodic structure in american sign language," *Chicago Linguistic Society*, vol. 38, pp. 693–704, April 2002.

[50] S. Z. Gurbuz, M. Mahbubur Rahman, E. Kurtoglu, E. Malaia, A. C. Gurbuz, D. J. Griffin, and C. Crawford, "Multi-frequency rf sensor fusion for word-level fluent asl recognition," *IEEE Sensors Journal*, pp. 1–1, 2021.

[51] M. Rahman, R. Mdrafi, A. Gurbuz, E. Malaia, C. Crawford, D. Griffin, and S. Gurbuz, "Word-level sign language recognition using linguistic adaptation of 77 GHz FMCW radar data," in *Proc. IEEE Radar Conference*, May 2021.

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[53] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," *CoRR*, vol. abs/1604.04382, 2016.

[54] M. G. Amin, Z. Zeng, and T. Shan, "Arm motion classification using curve matching of maximum instantaneous doppler frequency signatures," in *IEEE International Radar Conference*, 2020, pp. 303–308.

[55] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time serie," in *in KDD workshop, Seattle, WA*, 1994.

[56] M. S. Seyfioğlu and S. Z. Gürbüz, "Deep neural network initialization methods for micro-doppler classification with low training sample support," *IGERS Letters*, vol. 14, no. 12, pp. 2462–2466, 2017.

[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009, pp. 248–255.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd Int Conf on Learning Representations, ICLR 2015*, Y. Bengio and Y. LeCun, Eds.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[60] H. Sun, L. Gu, and B. Sun, "Adathm: Adaptive gradient method based on estimates of third-order moments," in *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, 2019, pp. 361–366.