HRSpecNET: A Deep Learning-Based High-Resolution Radar Micro-Doppler Signature Reconstruction for Improved HAR Classification

Sabyasachi Biswas, *Graduate Student Member, IEEE*,
Ahmed Manavi Alam, *Graduate Student Member, IEEE*,
and Ali C. Gurbuz, *Senior Member, IEEE*

***Abstract*— Micro-Doppler signatures (μ-DSs) are widely used for human activity recognition (HAR) using radar. However, traditional methods for generating μ-DS, such as the short-time Fourier transform (STFT), suffer from limitations, such as the tradeoff between time and frequency resolution, noise sensitivity, and parameter calibration. To address these limitations, we propose a novel deep learning (DL)-based approach to reconstruct high-resolution μ-DS directly from a 1-D complex time-domain signal. Our DL architecture consists of an autoencoder (AE) block to improve signal-to-noise ratio (SNR), an STFT block to learn frequency transformations to generate pseudo spectrograms, and, finally, a U-Net block to reconstruct high-resolution spectrogram images. We evaluated our proposed architecture on both synthetic and real-world data. For synthetic data, we generated 1-D complex time-domain signals with multiple time-varying frequencies and evaluated and compared the ability of our network to generate high-resolution μ-DS and perform in different SNR levels. For real-world data, a challenging radar-based American Sign Language (ASL) dataset consisting of 100 words was used to evaluate the classification performance achieved using the μ-DS generated by the proposed approach. The results showed that the proposed approach outperforms the classification accuracy of traditional STFT-based μ-DS by 3.48%. Both synthetic and experimental μ-DSs show that the proposed approach learns to reconstruct higher resolution and sparser spectrograms.**

***Index Terms*— American Sign Language (ASL), autoencoder (AE), HRSpecNet, human activity recognition (HAR), micro-Doppler signature (μ-DS), radar, short-time Fourier transform (STFT), time–frequency analysis (TFA), U-Net.**

I. Introduction

RECENT advancements in affordable solid-state transceivers, computationally efficient graphics processing units (GPUs), and innovative deep learning (DL) techniques have significantly expanded the practicality of radio frequency (RF) sensors across a growing range of applications involving human activity recognition (HAR) [1], [2], defense and security [3], [4], [5], mini-UAV classification [6], advanced driver assistance systems (ADASs) [7], [8], [9], indoor monitoring [10], [11], anomaly detection [12], and health monitoring [13], [14]. These advancements have ushered in a new era of RF sensor utility, enabling their integration into an ever-widening array of practical applications for HAR.

For radar-based HAR, time–frequency (TF) analysis is crucial as it captures essential kinematic information about dynamic activities, enabling accurate and efficient classification [15]. While some machine learning (ML)-based approaches can do classification directly from the complex RF data stream [16], [17], [18], [19], most conventional approaches for radar-based HAR require a two-level process. First, TF analysis (TFA) or other radar signal processing techniques are applied to generate 2-D (or higher 3-D and 4-D) radar data representations, such as time-varying range-Doppler (RD) or range-angle (RA) maps or micro-Doppler signature (μ-DS) [20]. While there have been some studies that propose joint domain classification [21], [22], most studies utilize the μ-DS for HAR [23], [24], [25], [26].

The most commonly used method of TFA is the short-time Fourier transform (STFT), which allows the decomposition of signals into their constituent frequencies at different time windows, providing a representation that highlights time-varying characteristics. In addition, other techniques, such as the Gabor transform, the wavelet transform, and the Wigner–Ville distribution, have also been employed [27], [28], [29]. These methods offer different insights into the TF domain, and their choice depends on the specific application and the desired level of detail in the μ-DS analysis.

Although the STFT represents a potent TFA tool, it is accompanied by specific tradeoffs and limitations.

1) *Frequency Resolution Versus Time Resolution:* One significant tradeoff in STFT is the balance between frequency and time resolution. On the one hand, a smaller window size provides better time resolution but poorer frequency resolution, making it difficult to distinguish between closely related frequencies. On the other hand,

Manuscript received 1 November 2023; revised 11 January 2024 and 14 March 2024; accepted 20 April 2024. Date of publication 2 May 2024; date of current version 17 May 2024. This work was supported in part by the National Science Foundation (NSF) under Award 1931861 and Award 2047771, and in part by the U.S. Engineer Research and Development Center under Grant W912HZ-21-2-0053. *(Corresponding author: Ali C. Gurbuz.)*

The authors are with the Department of Electrical and Computer Engineering and the Information Processing and Sensing Laboratory, Mississippi State University, Starkville, MS 39672 USA (e-mail: gurbuz@ece.msstate.edu).

Digital Object Identifier 10.1109/TRS.2024.3396172

a larger window size provides better frequency but poorer time resolution [30].

2) *Resolution Limits:* STFT has inherent resolution limits due to the fixed window size. This means that it may not capture fine details in signals with rapidly changing frequencies.
3) *Parameter Tuning:* Properly selecting the window length, overlap size, and number of frequency bins in STFT can be challenging, as these parameters need to be adjusted based on the characteristics of the signal [31]. This exhaustive parameter tuning can be time-consuming and may not always yield optimal classification results for different signal sizes and types.

Hence, there has been some interest in the development of DL methods for high-resolution frequency estimation and TF representation. Izacard et al. [32] proposed Deepfreq, a deep neural network (DNN) architecture to estimate the frequency of each component from a multisinusoidal complex-valued signal. Here, the real and imaginary channels are concatenated side by side to feed into the neural network model creating a 1-D frequency spectrum. In a related context, Pan et al. [33] proposed Cresfreq, a complex-valued neural network (CVNN) architecture to generate a 1-D frequency representation of a multisinusoidal complex-valued signal. These 1-D methods for high-resolution frequency estimation can also be leveraged to create 2-D TF representations, following a similar approach to how the fast Fourier transform (FFT) is employed in the STFT method.

However, the drawback of these DL modules is their computational inefficiency when applied repeatedly to produce μD spectrograms. They are limited by their fixed signal length and lack the flexibility to adjust window lengths, which is essential for achieving an optimal tradeoff between time resolution and frequency resolution. Moreover, these models are less robust to noise, resulting in poor estimations at lower signal-to-noise ratio (SNR) levels due to the fact that they are trained with only clean, noiseless signal samples. In response to these limitations, Pan et al. [34] introduced TFA-net, another CVNN architecture, aimed at directly generating high-resolution TF representations for complex multisinusoidal signals. Nevertheless, a notable concern with TFA-net is that the time index of the TF representations it generates aligns with the length of the input signal. This design choice imposes a substantial computational burden, especially for higher length signals, which are highly common in radar due to high sampling rates. In addition, CVNN architectures generally demand more computational time than their real-valued neural network counterparts. Consequently, while TFA-net delivers high-resolution TF representations, it does so at the cost of significantly higher computational latency compared to conventional methods such as STFT. Section II-B2 provides a detailed and comprehensive explanation of these models.

Building upon the inspirations drawn from Deepfreq, Cresfreq, and TFA-net, this article introduces HRSpecNet, an end-to-end DNN architecture designed to produce high-resolution 2-D TF representations from 1-D complex-valued signals with less computational complexity and better noise robustness. The proposed HRSpecNet architecture takes the 1-D complex data as input in two channels and comprises three distinct blocks. First, the 1-D convolutional autoencoder (AE) block is employed for effective noise reduction from the input signal, enhancing the quality of the data. Subsequently, the STFT block, which learns several convolutional filters resulting in proxy frequency domain representations, is incorporated. In this block, the weights are adaptively learned, enabling the representation of instantaneous frequency (IF) changes within the feature maps. Finally, the U-Net block utilizes these feature maps to construct a clean and precise high-resolution TF representation of the input signal. The HRSpecNet is more computationally efficient than its DNN-based predecessors due to its updated ground label generation and creating an output size as STFT. In addition, having an AE and a U-Net block together within the architecture suppresses the noise in two stages, producing much cleaner TF representations. The contributions of this article can be summarized as follows.

1) A comprehensive end-to-end neural network architecture has been developed that employs a three-stage architecture with AEs, convolutional STFT, and image reconstruction networks together with a novel weighted loss function to force the network to learn to effectively suppress noise and construct precise TF representations from 1-D noisy complex-valued signals.
2) A novel ground truth labeling process is proposed, which enforces higher resolution and has the same shape as STFT.
3) An investigation into the properties of the HRSpecNet model has been conducted, highlighting its comparative advantage over traditional STFT and showcasing its ability to generate high-quality TF representations without the need for extensive parameter tuning.
4) A rigorous evaluation of HRSpecNet has been provided, comparing it both quantitatively and qualitatively with STFT and other ML-based techniques such as Deepfreq, Cresfreq, and TFA-net. The findings of the evaluation reveal the performance of HRSpecNet in generating high-resolution spectrograms across various SNR levels.
5) To demonstrate the real-world applicability of HRSpecNet, an assessment of its generalization capability on real-life radar data has been performed. Specifically, the performance of HRSpecNet has been evaluated on a dataset comprising 100 American Sign Language (ASL) words. A standard 2-D convolutional neural network (CNN) model has been employed to classify the μD signatures generated by HRSpecNet. The results of the evaluation demonstrate that the classification of the μD spectrograms generated by HRSpecNet provides higher accuracy compared to other approaches.
6) HRSpecNet generates enhanced TF representations while maintaining very high computational efficiency compared to other ML-based techniques.

The organization of this article is as follows. The radar signal model and the existing classical and ML-based techniques to generate TF representations are summarized in Section II. Section III discusses the proposed HRSpecNet

architecture along with the process of dataset and ground truth generation. The properties of the generated TF representations and comparison with other approaches over both synthetic and real-world data are discussed in Section IV. Finally, conclusions are drawn and future directions are discussed in Section V.

II. Theoretical Background

A. Radar Signal Model

Frequency-modulated continuous-wave (FMCW) radar systems transmit linearly swept RF signals to measure both range and velocity [35]. The instantaneous frequency of the chirp signal can be modeled as

$$f_i(t) = f_0 + \frac{B}{\tau}t, \quad 0 \le t \le \tau \tag{1}$$

where f_0 is the initial frequency at time $t = 0$, B is the bandwidth, and τ represents the sweep time in seconds. The received signal is reflected back from a target with time delay T_d, mixed with a copy of the transmitted signal, and then passed through a low-pass filter (LPF) to obtain the IF signal. The IF signal can be modeled as

$$s_{\text{IF}}(t) = A\exp\left(2\pi\left(f_0 T_d + \frac{B}{\tau}T_d t - \frac{B}{2\tau}T_d^2\right)\right) \tag{2}$$

where A represents the amplitude of the signal. After sampling the IF signal, the FMCW radar stores the received data in a 3-D radar data cube (RDC) of size $P \times N \times M$, where P is the number of fast-time samples, N is the number of slow-time samples, and M is the number of receiver channels. Different RF data representations can be computed from the RDC, such as RD, RA, and μ-DS.

B. TF Representation for Radar Signals

Radar can be used to observe moving objects by measuring the change in frequency of the reflected radar signals. This change in frequency, called the Doppler shift, is caused by even the small movements of the target relative to the radar [36]. Changing target movements in time creates a signature in the frequency domain that can be observed in the TF representation of the radar signals.

1) Short-Time Fourier Transform: One of the most commonly used TF transforms for visualizing the μ-DS of a target is the spectrogram, which estimates the instantaneous μ-D frequency as a function of time by computing the square modulus of the windowed STFT across the slow-time radar data $x(t)$ as

$$S(k,\omega) = \left|\int_{-\infty}^{\infty} h(t-k)x(t)e^{-j\omega t}dt\right|^2 \tag{3}$$

where $h(t)$ is a windowing function, such as rectangular, Hamming, or Hanning window.

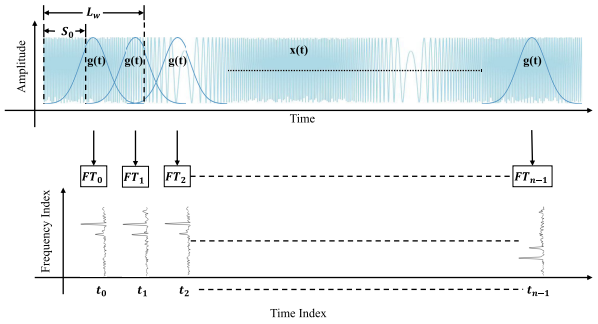


Fig. 1. Illustration of generating 2-D TF representation on a test signal $x(t)$. $g(t)$ is the Gaussian window and FT block denotes the 1-D frequency transformation architecture such as Deepfreq or Cresfreq models.

2) ML Techniques to Generate TF Representation: DL-based technique with the objective of precisely estimating the frequencies of multisinusoidal signals from a finite set of noisy samples is introduced first in the DeepFreq [32]. A novel neural network architecture is introduced in the study that outperforms existing methods such as FFT in terms of frequency estimation in high SNR scenarios. The DeepFreq framework combines two essential modules: one for enhancing frequency representation and another for automatic frequency count estimation for a fixed signal length. The DeepFreq framework is one of the very first DL-based approaches for frequency estimation, but it has several limitations. It is sensitive to noise in the input signal, which can lead to inaccurate frequency estimates. In addition, the DeepFreq framework cannot differentiate the amplitude of each frequency component within a signal. Finally, the DeepFreq framework has a high computational complexity compared to the existing FFT method.

Inspired by Deepfreq, a CVNN, Cresfreq, designed for high-resolution frequency estimation in 1-D complex signals has been introduced in [33]. The network learns complex-valued basis vectors and employs convolutional kernels for noise suppression. Subsequently, real-valued residual blocks enhance the frequency representation resolution. The Cresfreq addresses some limitations of Deepfreq such as performance in the low SNR scenarios and estimating amplitudes of the frequency components. However, it has a greater computational latency than Deepfreq.

By employing these 1-D frequency estimation techniques within segmented windows across the entire signal, we can generate 2-D TF representations as shown in Fig. 1, where L_w signifies the window length and S_0 represents the nonoverlap size. This is similar to the application of the STFT discussed in Section II-B1, replacing conventional frequency transformation with 1-D DL-based architectures. In order to perform the FT operation in Fig. 1, either Deepfreq or Cresfreq models can be utilized.

A data-driven 2-D TF representation has been introduced in [34], where a DL-based TFA model named TFA-Net is proposed. The network consists of two key modules. First, TFA-Net learns complete basis functions to obtain various TF characteristics of time series and then uses 2-D filter kernels for energy concentration to produce a TF representation of the

time-domain signal. Unlike the STFT method, it eliminates the need for window length adjustments. However, there are two significant computational challenges associated with TFA-Net. First, the TF representation's time index is aligned with the length of the input signal. This can lead to substantial computational demands, especially for signals with a large number of time indices, such as 1-D radar range profile signals with high pulse repetition frequency (PRF). This can significantly prolong the time required to generate the TF representation. In addition, TFA-Net involves complex-valued operations, which inherently consume more computational resources when compared to standard DNN models. This complexity adds to the computational overhead of the model.

While existing data-driven TF reconstruction approaches do not require adjustment of window lengths and perform better for multicomponent signals with closely adjacent IFs, they have low noise robustness and high computational complexity. As a solution, we introduce a novel DL architecture, HRSpecNet. In the following, we will delve into the specifics of our proposed approach.

III. Proposed Method

This section provides the architecture and details for the proposed ML-based generation of high-resolution μ-DS. We delve into the dataset generation process for input complex signals and corresponding ground truth TF representations. We elaborate on the HRSpecNet architecture that we have designed to robustly reconstruct high-resolution TF representations and its training process with a weighted loss function.

A. Dataset Generation

The process of dataset creation can be divided into two distinct phases. Initially, we generated a set of multicomponent 1-D complex time-domain signals, which will be the inputs to the proposed architecture. Subsequently, we produced the corresponding label TF images, representing the TF representation of the input signals.

1) 1-D Multicomponent Complex Signal Generation: We utilize multicomponent sinusoidal frequency modulated (FM) signals denoted as $s(k)$ as the dataset inputs. These signals and their corresponding IFs are expressed as follows:

$$s(k) = \sum_{q=1}^{Q} A_q \exp\big(2\pi f_q k\big) \times \exp\big(j2\pi B_q \times \sin\big(2\pi\left(a_q k^2 + b_q k + \theta_q\right)\big)\big) \quad (4)$$

$$\mathrm{IF}_q(k) = f_q + 2\pi B_q\big(2a_q k + b_q\big) \times \sin\big(2\pi\left(a_q k^2 + b_q k + \theta_q\right)\big). \quad (5)$$

In the context provided, the component number Q adheres to a discrete uniform distribution $\mathcal{U}(1, 10)$. The intensity of the qth component, denoted as A_q, is calculated as $0.5 + 8|\sigma_q|$, where σ_q follows a uniform distribution $\mathcal{U}(0, 1)$. The vibration amplitude of the qth component, represented by B_q, is sampled from $\mathcal{U}(0.2, 16)$. Both a_q and b_q fall within the ranges of $[-4, 4)$ and $[-2.4, 2.4)$, respectively. The parameter θ_q is drawn from a uniform distribution $\mathcal{U}(0, 2\pi)$, and the Doppler shift f_q follows a uniform distribution $\mathcal{U}(-1000, 1000)$. In order to limit the computational challenges while training the DL model, the signal length L is set to 1600 with a sampling frequency of 3200. The sampling frequency is set based on the pulse repetition interval (PRI) of our experimental radar system. Finally, the number of frequency bins N_f is established at 256. It is crucial to emphasize that while the model is trained with these specified parameters, the model architecture has been thoughtfully designed to ensure that the trained model can be effectively tested to generate TF representations for any input lengths as well as a wide range of sampling frequencies.

2) Ground Truth TF Representations: To generate the ground truth data, we begin by capturing each of the IF signal components as described in (5). Subsequently, we apply a moving average operation over each of them with a window size L_w and shift S_o to determine the label frequencies. This relationship can be expressed as follows:

$$\widehat{\mathrm{IF}}_q(i) = \frac{1}{L_w} \sum_{j=S_o * i}^{S_o * i + (L_w - 1)} \mathrm{IF}_q(j); \quad i = 0, 1, 2, \ldots, L_t. \quad (6)$$

This operation bears resemblance to the STFT process, resulting similar size time index for the labeled frequencies as in the STFT process.

Given that we have set the number of frequency bins to N_f, the resulting shape of the initial ground truth TF will be a matrix of size $N_f \times L_t$, where $L_t = (\lfloor (L - L_w/S_o) \rfloor + 1)$. This initial 2-D ground truth matrix, denoted as $\mathrm{GT}_{\mathrm{initial}}(f, i)$, can be represented as

$$\mathrm{GT}_{\mathrm{initial}}(f, i) = \begin{cases} A_q, & (\exists q)\big(f\Delta f \le \big(\widetilde{\mathrm{IF}}_q(i) < (f+1)\Delta f\big) \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where the time index, denoted as i, spans the range from 0 to $L_t - 1$, and the frequency index, denoted as f, ranges from 0 to $N_f - 1$. In this context, $\Delta f = (f_s/N_f)$ represents the frequency interval, with f_s being the sampling frequency. Finally, $\widetilde{\mathrm{IF}}_q(i) = \mathrm{mod}(\widehat{\mathrm{IF}}_q, f_s/2)$ signifies the modulo operation of $\widehat{\mathrm{IF}}_q$ with respect to $f_s/2$. Afterward, we convolve the 2-D $\mathrm{GT}_{\mathrm{initial}}$ with a 1-D Gaussian kernel with a kernel size of 3 and a standard deviation of 1 along the frequency dimension in order to compensate for averaging effects and smooth out our final ground truth, denoted as $\mathrm{GT}_{\mathrm{final}}$. Fig. 2 shows the flow diagram from example IF signals to the corresponding final ground truth TF image.

B. Proposed HRSpecNet Architecture

In this segment, we introduce a new DL network architecture, HRSpecNet, for high-resolution TF representation. The proposed architecture is illustrated in Fig. 3. The overall framework comprises three primary component networks, beginning with an AE block, followed by a convolutional network block resembling the STFT operation, and concluding with a U-Net block for the generation of high-resolution 2-D TF representation of the input signal. The input of the model takes complex signals with $C \times L$, where $C = 2$ accommodates

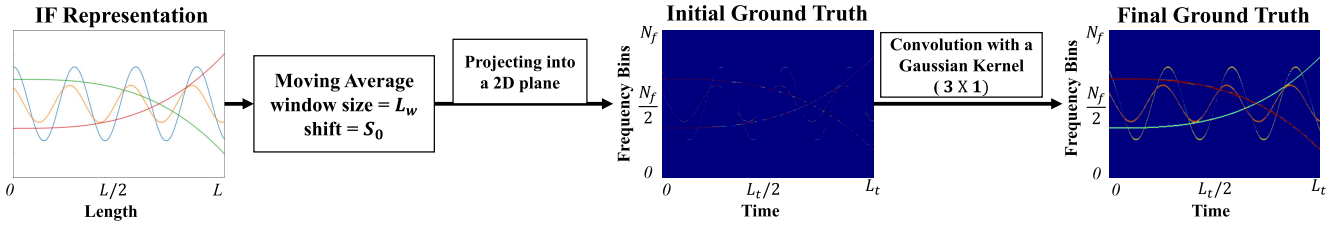


Fig. 2. Flow diagram of ground truth TF-image generation of the example signal defined in (9).

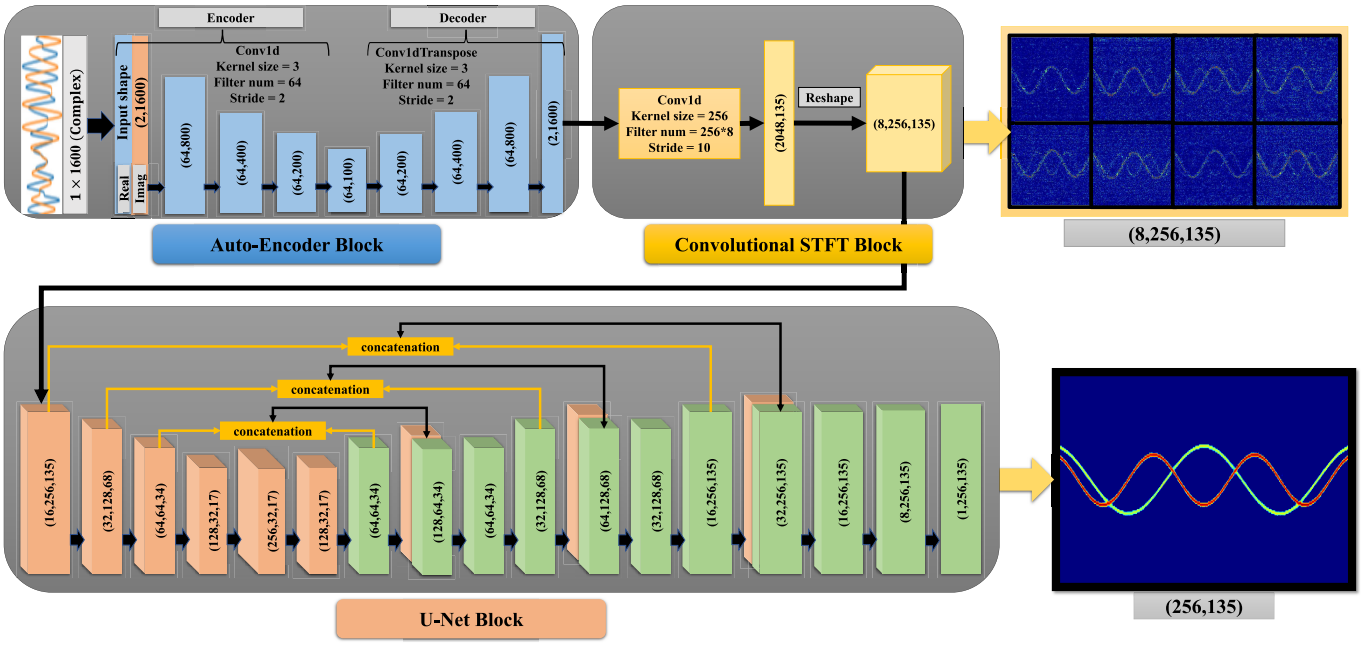


Fig. 3. HRSpecNet architecture.

the real and imaginary parts of the signals and L depends on the signal length. During the training phase, the DL framework is fed with a signal deliberately infused with noise. The primary purpose of leveraging the AE block is to mitigate noise and subsequently increase the network's overall performance. The AE's output is passed into the convolutional STFT block, where the network's convolutional filter can calculate multiple Fourier-like transformations, effectively depicting various proxy TF representations of the original IFs. Following the reshaping process, these proxy TF feature data are then passed through the U-Net block, which is harnessed for the generation of a high-resolution and focused TF representation. The whole network is trained with a weighted loss guiding the model to both learn to generate outputs close to the high-resolution labeled TF images while also learning to reduce the noise and generate a cleaner version of the noisy input signal. A more comprehensive elaboration of each block is presented subsequently.

1) AE Module: The AE architecture employed for the goal of noise reduction consists of multiple Conv1D layers. These layers are pivotal in capturing intricate patterns within the input data while progressively reducing the dimensions of the feature space. Specifically, each Conv1D block is configured with 64 filters, a kernel size of 3, and a stride of 2. These parameters enable the model to perform local convolution operations, extracting salient features with reduced spatial dimensions as the signal passes through the layers. Ultimately, this hierarchical encoding process helps the AE learn a compact representation of the input data, making it highly effective in reducing noise and enhancing the overall quality of the final TF representation. During training, the DL architecture takes input data with a shape of 2×1600, and the AE block preserves the same output shape as the input. During training, the AE module assesses its output against a noise-free version of the same input signal using a sum of squared error (SSE).

2) Convolutional STFT Module: Noise-reduced output from the AE block is directly fed into the convolutional STFT block. This block consists of a 1-D convolutional unit that helps to create several intermediate TF representations. The hyperparameters of this convolutional layer, such as kernel size and stride, represent the window size and shifting in a typical STFT operation. The number of filters in this layer helps in determining the number of intermediate TF representations that will be fed into the U-Net block. For an input signal of length L, the size and number of filter kernels used in the STFT module are, respectively, L_w and $N_f N_{\mathrm{TF}}$. Shifting of convolution operation is conducted through stride of n_0 and the output feature maps of the STFT layer become $N_f N_{\mathrm{TF}} \times L_t$.

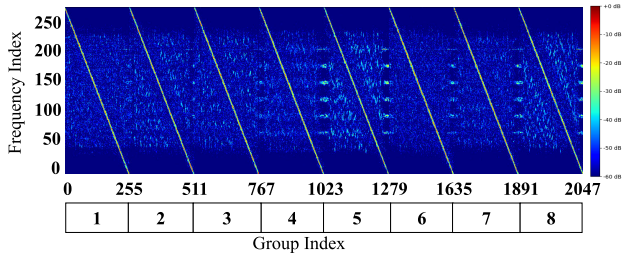


Fig. 4. Visualization of the frequency spectra learned by the convolutional STFT module.

These feature maps will be reshaped within the STFT block to produce $N_{\text{TF}} \times N_f \times L_t$ to illustrate N_{TF} number of intermediate TF representations of the original signal. For clarity, in Fig. 3, an example signal with a length of 1600 samples is illustrated. The real and imaginary parts of the signal are divided into two-channel configurations, as described in Section III-B2. In the STFT block, kernel size, filter numbers, and stride are taken as 256, 256 × 8, and 10, respectively. After reshaping, the output feature maps of the convolutional STFT block become 8 × 256 × 135. In the output shape, 8 is the number of proxy TF representations, 256 is the number of frequency bins, and 135 is the time index L_t. As an example, each crude TF representation from the convolutional STFT block is also illustrated in Fig. 3. It is crucial to note that the entire framework is not reliant on a specific signal length, and the time index L_t within the convolutional STFT block is directly determined by the number of samples provided at the input layer. Hence, the trained model for the given parameters can be tested to generate TF representations for varying input lengths.

The shape of the feature map obtained from the convolutional STFT module is $N_f N_{\text{TF}} \times 2 \times L_w$, where 2 represents the real and imaginary parts. Fig. 4 illustrates the frequency spectra learned from the weights of the convolutional STFT module, revealing the presence of eight distinct feature sets. During training, to generate the output of the convolutional STFT module closer to the labeled TF images, a weighted SSE loss function, L_2, was used, as shown in Fig. 5. Each of the outputs of this block provides a rich set of frequency features to the next U-Net block allowing it to produce a high-resolution TF representation as its output.

3) U-Net Module: The primary purpose of the U-Net module is to fuse multiple TF feature maps from the convolutional STFT module into a high-resolution 2-D TF representation. In Fig. 3, the overall framework of the U-Net module is illustrated. The U-Net architecture comprises an encoder–decoder network designed for the high-resolution TF representation task. In the encoder, a series of convolutional layers with increasing channels performs hierarchical feature extraction, capturing details at different scales. Batch normalization and rectified linear unit (ReLU) activation functions enhance training stability and nonlinearity. The decoder, on the other hand, consists of transposed convolutional layers that upsample the feature maps to recover the spatial resolution. Skip connections achieved through concatenation play a pivotal role, facilitating the fusion of low- and high-level features to produce the detailed TFA. The final convolutional layer produces a single-channel output representing the 2-D TF representation.

4) Training the Proposed Architecture: Our training dataset is constructed using 2×10^5 noisy signals, with uniformly random generated SNR levels ranging from 0 to 15 dB. For validation purposes, we employed a different set of 2×10^4 noisy signals, with SNR levels following the same variation. Losses L_2 and L_3 are computed as the SSE loss between the convolutional STFT and U-Net module outputs and the label TF images, respectively. These losses are then combined with the loss in the AE block L_1. The total loss utilized in training of the whole model is given as

$$L_{\text{total}} = (\lambda \times L_1) + (\alpha \times L_2) + L_3 \tag{8}$$

where L_{total} denotes the aggregate training loss, while parameters λ and α are hyperparameters of the model and are systematically tuned to a value of 3 and 0.1, respectively, through an iterative process, aimed at achieving optimal performance. Specifically, we tested the SSE loss term between the output of the convolutional STFT block and the labeled TF images to observe if forcing the input of the U-Net block to be closer to the final label enhanced the final output. Fig. 6(a)–(c) represents the normalized mean square error (NMSE), structural similarity index measurement (SSIM), and peak SNR (PSNR) between the model output and ground truth TFs as a function of SNR, respectively, by varying α from 0 to 10. We observed the best results when $\alpha = 0.1$. The flow diagram of the final proposed model is shown in Fig. 5 for better visualization.

IV. Performance Analysis of the Proposed Method

For a comprehensive evaluation of the performance of the proposed HRSpecNet model, this section provides results on the following:

1) the performance of the HRSpecNet evaluated through various simulation settings;
2) qualitative comparisons with existing methods in terms of resolution and SNR performance;
3) quantitative comparisons, the effect of generated μ-DS with the proposed method on classification performance tested over a challenging experimental RF dataset consisting of classifying 100 ASL signs.

In Sections IV-A and IV-B, for comparison with the STFT, Deepfreq, and Cresfreq models, the window size, L_w, and the shift size, S_0, are set to be 200 and 10, respectively. In addition, the number of frequency bins is set to 1000.

A. Simulation Validation

1) μ-DS Generation: To illustrate the performance of the proposed approach, an example signal consisting of four frequency components is considered. The signal and the corresponding IF expressions are given as follows:

$$\begin{aligned} s(t) = {} & 1.5\exp(j2\pi(80\sin(6\pi t))) \\ & + 2\exp(j2\pi(40\sin(6\pi t + 0.5))) \\ & + 4\exp(j2\pi(1000t - 500t^4)) \\ & + \exp(j2\pi(-1000t + 500t^4)) \end{aligned} \tag{9}$$

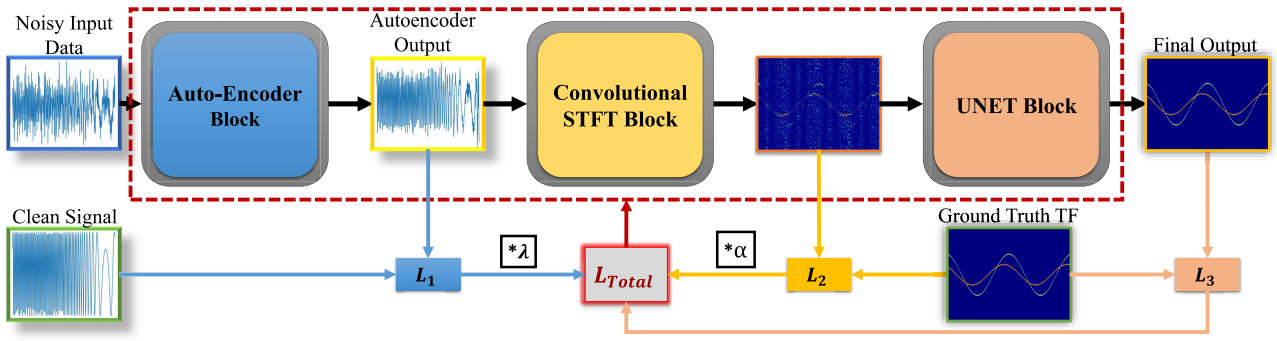


Fig. 5. Flow diagram of the proposed architecture.

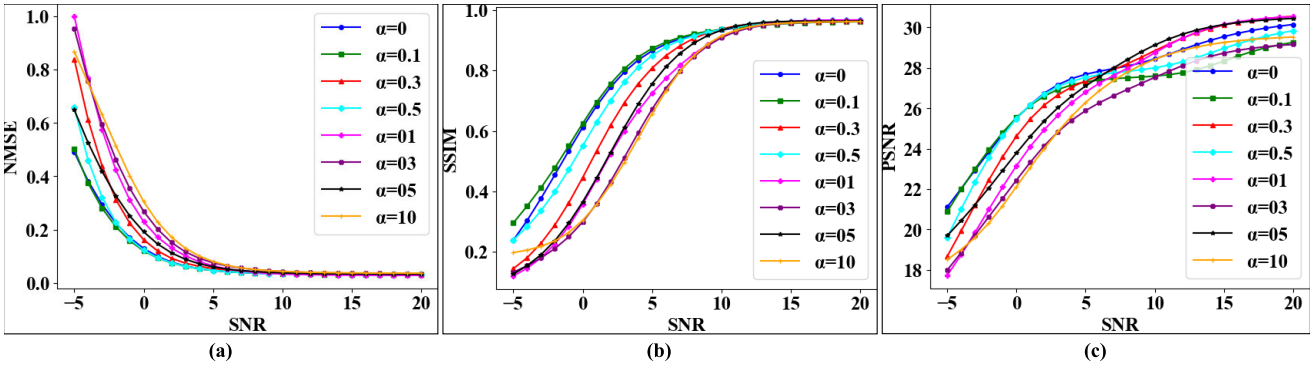


Fig. 6. Comparison between the final output and the ground truth TF with varying the weight of L_2 loss function (α) in terms of (a) NMSE, (b) SSIM, and (c) PSNR versus SNR.

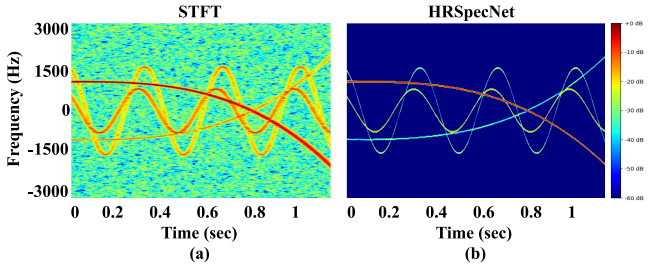


Fig. 7. TF representation of signal (9) by (a) STFT and (b) HRSpecNet.

$$\begin{aligned} \text{IF}_1 &= 480\pi\cos(6\pi t) \\ \text{IF}_2 &= 240\pi\cos(6\pi t + 0.5) \\ \text{IF}_3 &= 1000 - 2000\ t^3 \\ \text{IF}_4 &= -1000 + 2000\ t^3. \end{aligned} \tag{10}$$

In this experiment, the SNR was 10 dB, the sampling frequency was 6000 Hz, and the time duration was 1.15 s. First, the STFT of the data is computed using a window length of 256 and a nonoverlap length of 10. Fig. 7(a) and (b) compares the TF representations of STFT and HRSpecNet, respectively, with the same colormap. The STFT shows a considerable amount of noise, while the HRSpecNet gives a higher resolution, sparser, and cleaner representation similar to the ground truth. In addition, the proposed HRSpecNet method was able to detect the intersection points of IF signals much more clearly than the STFT. Finally, the amplitude information of each IF signal component was preserved in the HRSpecNet.

One important need for STFT-based TF representation is the need to select window length and shifting parameters for an optimal result. The top and bottom rows in Fig. 8 represent the TF representations using STFT with varying window sizes and the HRSpecNet for a signal with a sampling frequency of 2000 and 8000 Hz, respectively. The instantaneous frequencies have the same normalized values so that the resultant TF plot shows the same output on both occasions. The generated signal consists of two frequency components as given in the following:

$$\begin{aligned} s(t) = {} & 1.5\exp\left(j2\pi\left(\frac{1}{240}\sin\left(4\pi t^2\right)f_s\right)\right) \\ & + 4\exp\left(j2\pi\left(\frac{1}{6}f_s t - \frac{3}{10}f_s t^4\right)\right) \end{aligned} \tag{11}$$

$$\begin{aligned} \text{Normalized-IF}_1 &= \frac{1}{30}\cos\left(4\pi t^2\right) \\ \text{Normalized-IF}_2 &= \frac{1}{6} - \frac{6}{5}t^3. \end{aligned} \tag{12}$$

Here, the output shows two important properties of the proposed HRSpecNet model.

1) *More Flexible Against Parameter Tuning:* As discussed earlier, the time and frequency resolution varies with respect to the length of the window functions. The optimum length of the window function depends heavily on the signal length, sampling frequency, and the changing rate of instantaneous frequency components. As shown in Fig. 8, the spectrograms created by STFT using window lengths 32, 64, 128, and 256, show different

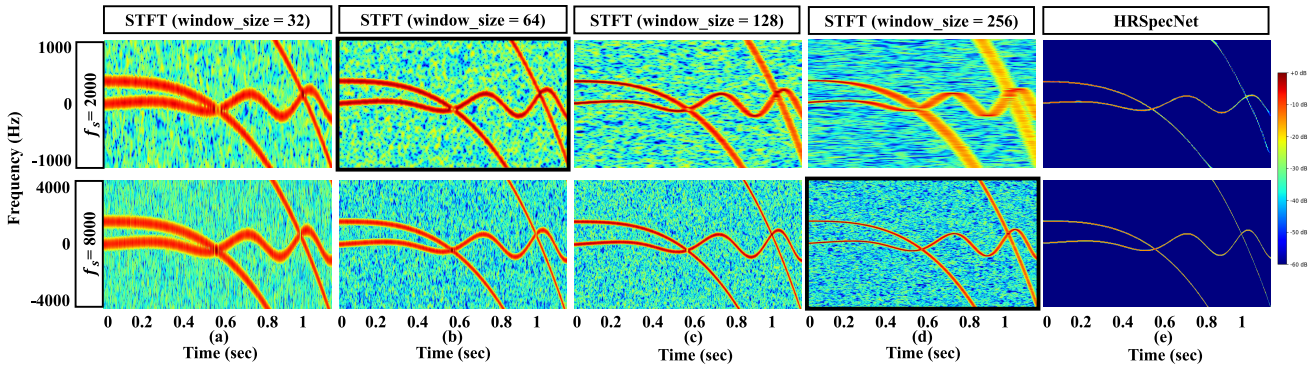


Fig. 8. TF representation of signal (11) by (a)–(d) STFT and (e) HRSpecNet. The highlighted black boxes indicate the best TF representations observed from STFT.

characteristics. Different window lengths, such as 64 and 256, were found to be better for different sampling frequency cases. On the other hand, the outputs from the HRSpecNet shown in Fig 8(e) are fairly consistent on both occasions. This shows that while the window length needs to be adjusted for STFT to get enhanced TF representations, the HRSpecNet is more robust to this exhaustive parameter tuning.

2) *Aliasing Effects:* As seen in Fig. 8, the signals contain aliasing effects as the signal component in IF_2 aliases after 0.8 s. Although the aliasing effect was not considered in the training of the model or in training dataset generation, our model was still able to reconstruct a correct TF representation for an aliased component.

2) Performance Improvements Due to AE: Some of them are listed as follows.

a) SNR improvement: As discussed in Section III-B4, the HRSpecNet model has been trained using 200 000 data samples. These samples were corrupted by Gaussian noise with the SNR levels varying from 0 to 15 dB. Fig. 9(a) shows the SNR difference between the input signal and AE output for 20 000 testing samples. Each sample duration is 1 s with a sampling frequency of 6 kHz and a maximum number of ten IFs. The test SNR levels span from -5 to 20 dB with an interval of 5. The AE's input and output SNR levels were calculated with respect to the ground truth time frequency representation (TFR). After testing the trained model, it was observed that the AE improved the SNR of the signals by an average of 2.5 dB. As shown in Fig. 9, utilizing the AE in the proposed architecture improved the SNR of the signals at all SNR levels, with the greatest improvement seen at lower SNR levels.

b) Overall performance improvement: To further evaluate the effects of the AE on the final reconstructed TF-image outputs, the NMSE, SSIM, and PSNR of the TF representation with respect to the corresponding clean signal TF representation are computed as a function of the SNR levels. Fig. 9(b)–(d) shows the NMSE, SSIM, and PSNR plots versus SNR for HRSpecNet and HRSpecNet without AE, respectively. Overall, adding the AE helped to reduce the NMSE. Furthermore, significant enhancements in SSIM and PSNR can be seen as well. These results demonstrate that the AE can improve the reconstruction performance of the U-Net module, especially at lower SNR levels. At higher SNR levels, both models showed similar performance. In short, the improvement in the SNR level due to the AE helped the U-Net module to reconstruct finer TF representations. Fig. 10(e) and (f) shows the TF representations in various SNR levels without AE and with AE block, respectively. The contribution of the utilization of the encoder on reconstructed TF images can be seen clearly.

B. Qualitative Comparison With Existing Methods

1) High-Resolution TF Spectrograms: In this section, we are going to analyze and compare how varying models perform in terms of distinguishing closely related frequency points. To test this, we chose a noiseless signal with two frequency components, with a constant frequency difference. The sampling frequency and the duration are 6 kHz and 1.15 s, respectively. The signal representation is given as follows:

$$s(t) = \exp(j2\pi(20\sin(6\pi t) - 1500t)) + \exp(j2\pi(20\sin(6\pi t) - 1500t + d)) \quad (13)$$

$$\mathrm{IF}_1 = 120\pi\cos(6\pi t) - 1500$$

$$\mathrm{IF}_2 = 120\pi\cos(6\pi t) - 1500 + d. \quad (14)$$

To evaluate the frequency resolution performance of the proposed HRSpecNet model, we compared it with the standard STFT method and three other ML-based techniques, Deepfreq, Cresfreq, and TFA-NET. The IF components were separated by a constant frequency difference d as shown in (14), where d was gradually decreased from 200 to 60 Hz with an interval of 10 Hz. The window length for STFT operation was set to 256 as we observed the best TF representations using this particular window length.

As shown in Fig. 11, when $d = 200$, all models were able to clearly distinguish the two IF components. However, when $d < 150$, the STFT, Deepfreq, and Cresfreq models gradually lost their ability to distinguish the two IF components, while the TFA-NET and HRSpecNet models were still able to clearly separate them. The TFA-NET and HRSpecNet models started to lose their ability to identify the two IF signals at $d < 70$ and $d < 60$, respectively. This shows that TFA-Net and the proposed HRSpecNet demonstrate enhanced frequency resolution performance compared to STFT.

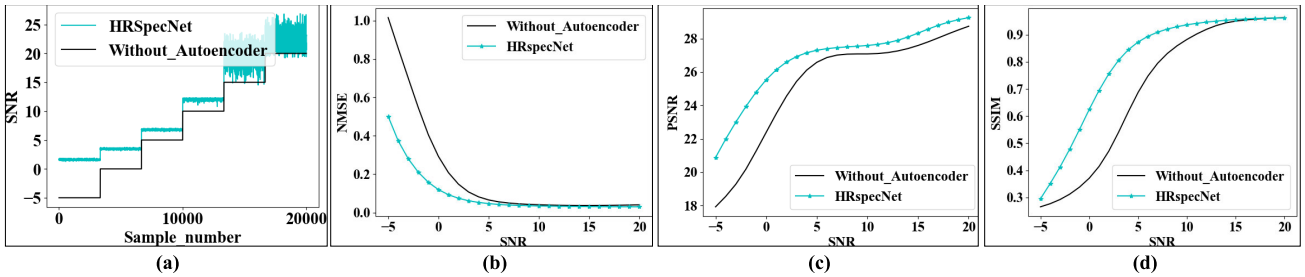


Fig. 9. Effect of AE. (a) Actual SNR levels and AE output SNR levels versus sample number. (b) NMSE at the final output. (c) PSNR at the final output. (d) SSIM at the final output.

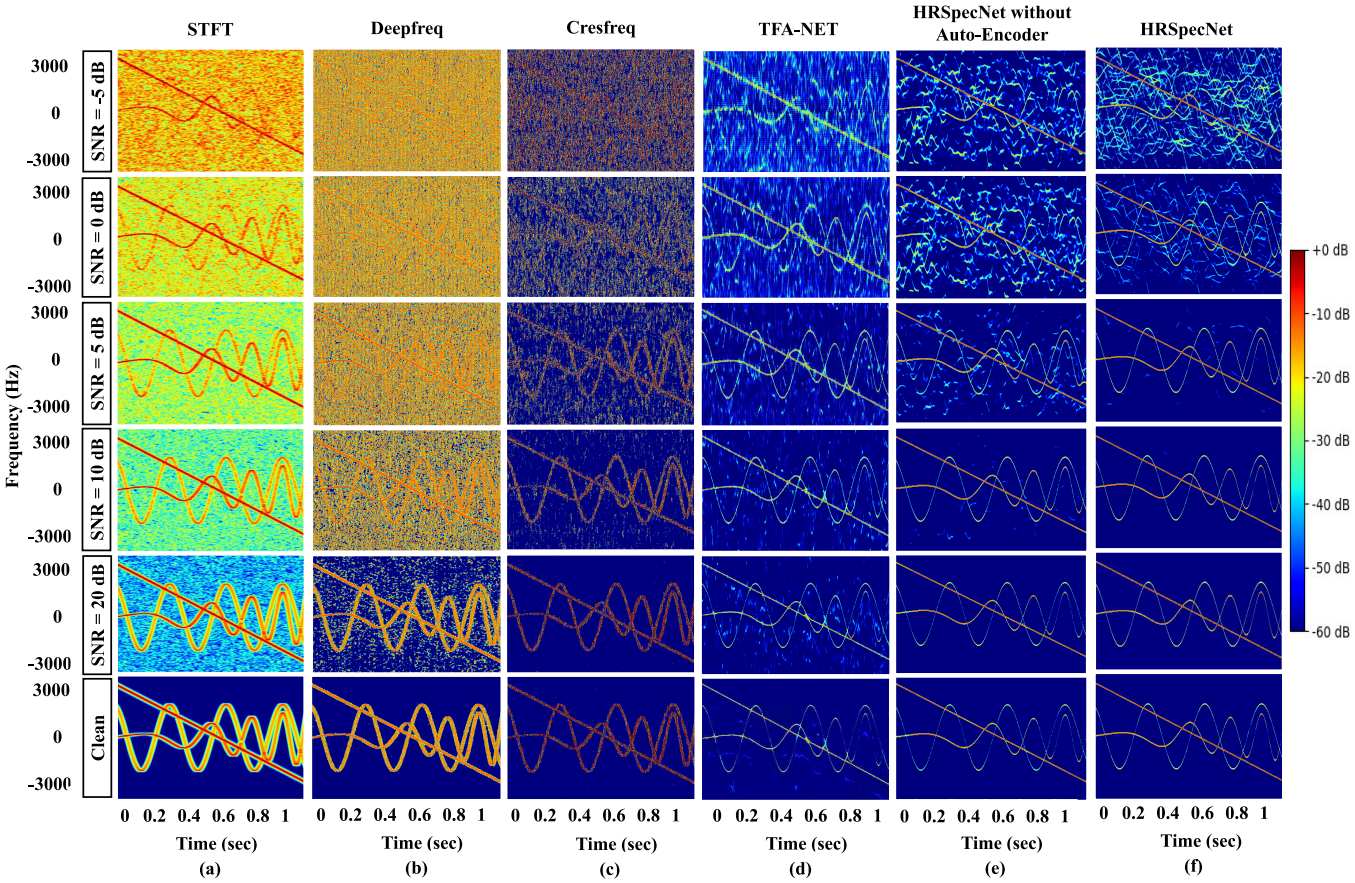


Fig. 10. TF representation of signal $s(t) = 2\exp(j2\pi(80\sin(6\pi t))) + 3\exp(j2\pi(30\sin(6\pi t^2))) + 4\exp(j2\pi(2500t - 2000\ t^2))$ in different SNR levels for (a) STFT, (b) Deep-freq, (c) Cresfreq, (d) TFA-NET, (e) HRSpecNet w/o AE, and (f) HRSpecNet.

2) Different SNR Levels: In this section, the performance of the proposed approach will be compared with the existing methods both qualitatively and quantitatively.

a) Qualitative comparison: Fig. 10 shows the TFRs generated from the proposed model along with STFT and the ML-based methods for the same signal under different SNR conditions varying from −5 to 20 dB. It is important to note that the proposed HRSpecNet model was trained with 0–15-dB SNR cases. At −5-dB SNR, all the models produce noisy TFRs. However, some of the frequency components can be seen in the TFRs from STFT, TFA-NET, and HRSpecNet. At 0-dB SNR, TFA-NET and STFT can show the frequency components in a noisy background, while HRSpecNet produces a much less noisy TFR. However, the frequency components are hardly visible in the TFRs from Cresfreq and Deepfreq. At 5-, 10-, and 20-dB SNR levels, HRSpecNet produces very clean and accurate TFRs compared to other methods. All the DL-based methods, as well as the STFT, yield accurate TF representations when applied to clean signals or signals with higher SNRs.

b) Quantitative comparison: To demonstrate a quantitative comparison, first a testing dataset was generated with SNRs varying from −5 to 20 dB with 1-dB intervals. The 1000 samples were generated for each SNR case, and thus, the testing dataset consisted of 26 000 data samples for this analysis. Each sample duration is 1 s with a sampling frequency of 6 kHz and a maximum number of ten IFs. For evaluation, we compared each of the noisy TFRs with respect to their corresponding ground truth TFRs. Then, the average of NMSE, SSIM, and PSNR metrics was calculated

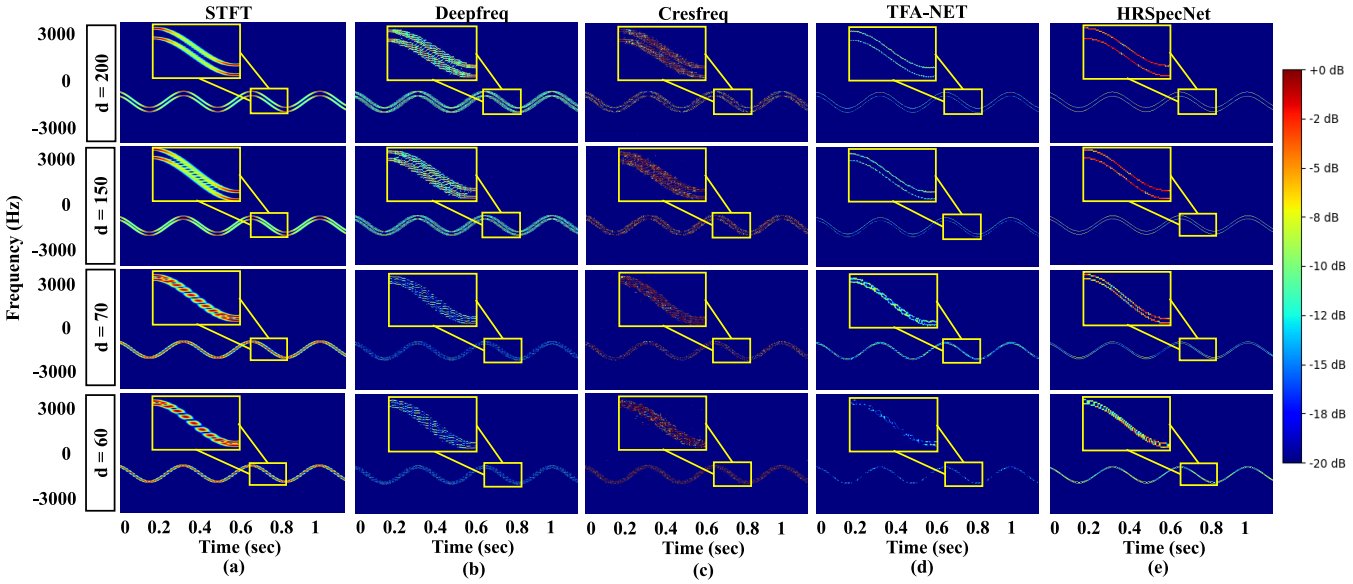


Fig. 11. TF representation of signal (13) by (a) STFT, (b) Deepfreq, (c) Cresfreq, (d) TFA-NET, and (e) HRSpecNet.

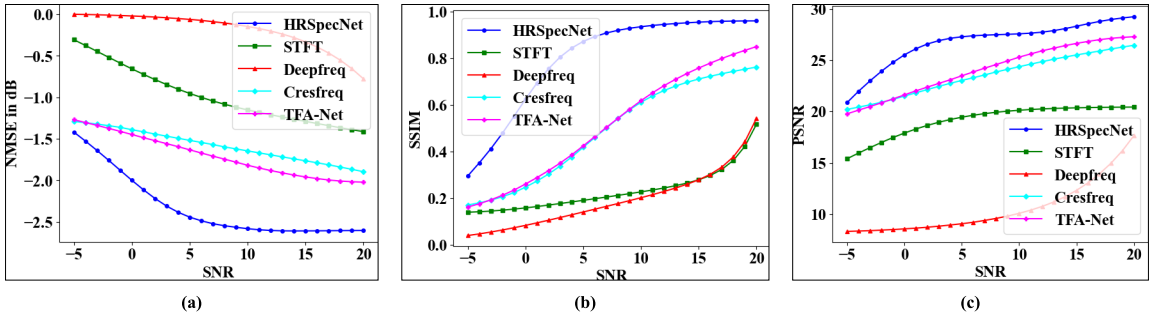


Fig. 12. Comparison between HRSpecNet, STFT, and other DL approaches. (a) NMSE. (b) SSIM. (c) PSNR versus SNR.

for each SNR level. Fig. 12(a)–(c) shows the NMSE, SSIM, and PSNR as a function of SNR for all compared approaches, respectively. The results show that the proposed approach provides enhancements compared to standard STFT and other DL-based approaches in all metrics.

This suggests that HRSpecNet could be used for applications where high-quality high-resolution TFRs are required, such as radar target recognition, speech processing, and music analysis.

C. Quantitative Comparison on Classification Performance

Although different methods can reconstruct varying TF representations, one important analysis is how the reconstructed TF representations affect the final classification performance. In order to evaluate the performance of compared methods in a real-world HAR scenario, a challenging dataset consisting of 100 ASL signs was utilized. First, the μ-DSs were generated using STFT and the trained DL-based models. Note that all data-driven approaches compared here are trained only with their respective simulated datasets. No experimental data are used in the training process, and the trained models are tested on the experimental data. In Fig. 13, μ-DSs of three ASL signs reconstructed using compared models have been given. This figure reveals that the spectrograms produced by the HRSpecNet model exhibit a higher resolution and increased ability to distinguish subtle movements, surpassing not only the other DL-based methods but also the STFT. Furthermore, the proposed approach excels in noise suppression, resulting in significantly clearer and sparse spectrograms compared to other DL-based approaches. This shows that the proposed HRSpecNet model can generate realistic spectrograms in real-world conditions even though it is only trained with synthetic signal examples. After generating the spectrograms from each method, the magnitude of the μDS is obtained and given as the input to four different DNNs for classification: deep CNN (DCNN) [37], person identification [38], VGGNet-16 [39], and a CNN model specifically developed for radar-based ASL classification, ASL-Net, based on [10] in order to test the classification performance observed over reconstructed TF images from compared techniques. Multiple different classifiers are tested to remove any bias from a specific classifier.

Next, we will discuss the experimental setup, data collection procedure, explanation of the CNN model for classification, and performance comparison with other existing methods.

1) Experimental Setup: For RF data collection, a 77-GHz TI IWR1443 automotive short-range radar was used. The

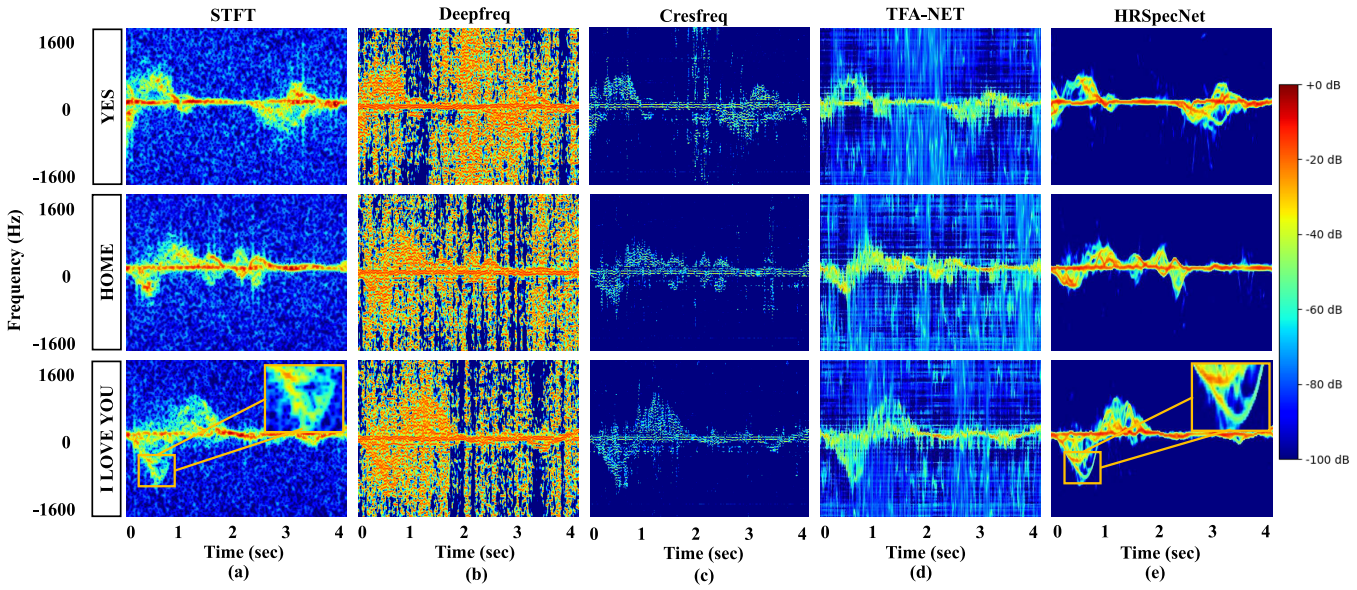
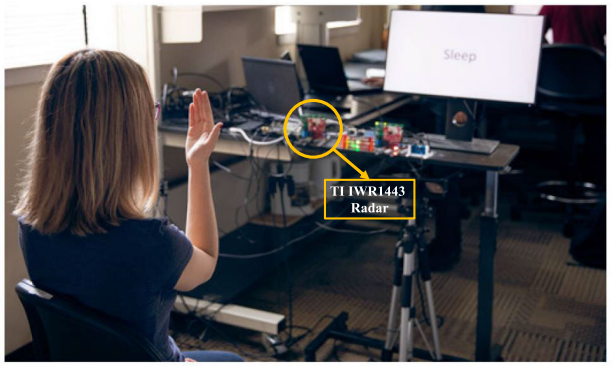


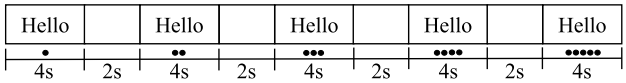
Fig. 13. Sample μD signatures for ASL signs for (a) STFT, (b) Deepfreq, (c) Cresfreq, (d) TFA-NET, and (e) HRSpecNet.

TABLE I
TI IWR1143 RADAR PARAMETERS

Parameter	Value
Number of ADC samples	256
Number of TX channels	1
Number of RX channels	1
Start frequency	77 GHz
Stop frequency	81 GHz
Bandwidth	4 GHz
RX gain	45 dB
Periodicity	40 ms
Pulse repetition interval (PRI)	312.5 μs
Pulse repetition frequency (PRF)	3200 Hz
Number of ADC samples per chirp	256
Number of Chirp loops per Frame	128
Total number of frames	700
Total number of chirps	89,600
Total time	28s



(a)



(b)

Fig. 14. Experimental setup and timing of sign articulation. (a) Experimental setup for data collection with radar [22]. (b) Example of sequential prompts given to the user.

radar has three transmitters and four receivers, but only one transmitter and one receiver were used in this experiment. The radar system parameters selected for the data collection are given in Table I.

The radar was positioned on top of a table placed against a wall in a laboratory environment at a height of 0.91 m. ASL signers were seated on a chair in front of the radar at a distance of 1.5 m. A computer monitor was placed exactly behind the radar, outside of its field of view to prevent it from reflecting the radar waves and creating noise in the signal. The monitor continuously displayed instructions to the participants about the signs they needed to articulate. This setup was designed to minimize interference from the environment and to ensure that the radar had a clear view of the signer's hands. The experimental setup is shown in Fig. 14(a).

2) Dataset: Six people participated in data collection, including four professional ASL signers, two hearing impaired people, two children of hearing impaired adults (CODAs), and two lab members. The whole data collection procedure was IRB-approved. Hundred different ASL signs were selected from the ASL-LEX2 [40] database. The signs were chosen to be high frequency and phonologically unrelated to each other in order to create a diverse dataset. More information about the dataset can be found in [41].

The dataset consists of a total of 3000 radar sign samples from six participants, with 30 samples for each class. Five repetitions of each sign were collected from each participant, with each repetition lasting 4 s followed by a 2-s interstimulus interval. In total, 28 s of data were collected for each class from each participant.

3) Classification Model: The generated μDSs from all models were saved as 128 $\times$ 128 images, which are then supplied as input to four different 2-D CNN models, including the ASL-Net architecture. The ASL-Net consists of four con-

TABLE II
PERFORMANCE OF THE COMPARED MODELS IN TERMS OF EVALUATION METRICS

Method	Proposed Classification Model Result						Average Classification Result					
	Testing Accuracy			Precision	Recall	F1 Score	Testing Accuracy			Precision	Recall	F1 Score
	Top 1	Top 3	Top 5				Top 1	Top 3	Top 5			
STFT	56.88	76.22	82.81	61	59.71	56.45	48.47	68.83	76.59	53.56	48.72	47.54
Deepfreq	35.62	58.03	68.53	44.88	36.76	36.81	29.84	50.16	59.53	34.79	30.34	29.63
Cresfreq	45.54	70.96	80.8	49.79	46.58	44.69	42.23	64.69	74.84	46.4	42.52	40.98
TFA-Net	42.01	61.5	72.75	42.7	40.96	41.57	37.64	58.93	70.34	39.94	37.31	36.21
HRSpecNet	60.63	80.89	86.44	65.53	62.89	60.22	51.95	73.66	79.19	56.71	52.77	50.92

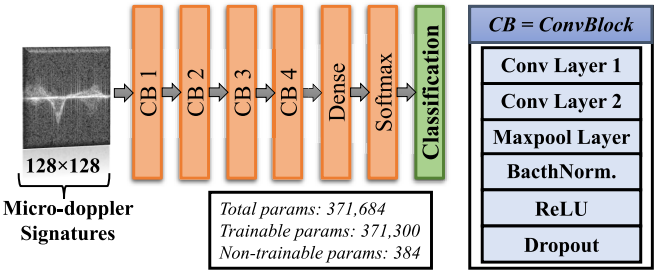


Fig. 15. CNN architecture for ASL sign classification.

TABLE III
COMPARISON OF COMPUTATIONAL EFFICIENCY

Methods	Time (Sec)
STFT	0.045
Deepfreq	2.009
Cresfreq	4.852
TFA-NET	4.112
HRSpecNet	0.081

volutional blocks, as illustrated in Fig. 15. Each convolutional block had two convolutional layers having 32 filters each in the first two blocks and 64 filters each in the rest. All the convolutional layers consist of a kernel size of 4×4. The two convolutional layers in each block were followed by 2×2 maxpooling, batch normalization, ReLU activation function, and a dropout of 0.3. After the convolutional blocks, the tensor is flattened and fed into a dense layer of size 256×1, after which a dropout of 0.3 is applied and input to a softmax classifier.

4) Classification Performance for 100 Class ASL Data: For performance evaluation, five TF-image datasets were generated based on STFT, HRSpecNet, TFA-NET, Deepfreq, and Cresfreq. All Dl-based reconstruction approaches are completely trained on simulated data only and the trained models are utilized to generate the TF images for the experimental radar data. Each reconstructed TF-image dataset was split into 80% training and 20% testing samples in the same way. All four classification models have been trained individually for a long enough time (150 epochs) to reach convergence for each model using the training portion of the generated TF-image datasets. Afterward, the trained models were tested, and confusion matrices were generated for each case. Testing accuracy, precision, recall, and $F1$ Scores were evaluated from the confusion matrices. Table II presents the obtained results for the top-performing classifier, ASL-Net, and the average results from four classification models. All tested classifiers provided the best accuracy results for the dataset generated using the proposed HRSpecNet model. In addition to outperforming compared DL-based reconstruction approaches, on average, HRSpecNet was able to give 3.14% better accuracy than the state-of-the-art STFT method. The noise-robust architecture of the proposed model with weighted loss terms, the novel labeling process, and the generation of suitable training dataset are several possible reasons for the enhanced performance.

As shown in Table II, top-3 and top-5 accuracies were also computed. Top-N accuracy indicates whether the model is able to predict the expected class within the top-N highest probabilities. Since the dataset consists of 100 classes, top-N accuracy measurement is an important performance-evaluating factor to consider. The top-3 and top-5 accuracies of HRSpecNet also surpass those of other methods.

D. Computational Efficiency

In this section, we carried out simulations to compare the computational efficiency of different methods, including the proposed HRSpecNet. For evaluation, the same testing dataset mentioned in Section IV-B2 was used. We measured the average execution time for each method, and the results are summarized in Table III. Both data processing and network training have been done using an Alienware m15 R7 laptop with an NVIDIA 3060 GPU, Intel 11th Gen CPU, and 32-GB memory.

The STFT is observed to be the least computationally intensive approach, while the HRSpecNet is the most computationally efficient method among DL-based techniques. On the other hand, TFA-NET, despite producing similar high-resolution TF representations, takes significantly more time to complete than our proposed approach. There are two main reasons for this. First, TFA-NET generates TFRs with dimensions of 256×6000, as the time index of these TFRs matches the dimension of the input signal. Second, TFA-NET involves complex-valued operations, which inherently demand more computational resources than standard DNN models. To illustrate further, even though Deepfreq has more parameters than Cresfreq, the latter, being complex-valued, ends up costing about 2.5 times more in terms of computational time.

V. CONCLUSION

An innovative DL-based architecture, named HRSpecNet, is introduced specifically to reconstruct highly concentrated TF representations of multicomponent time-varying signals.

HRSpecNet is composed of three fundamental modules: the AE, the convolutional STFT, and the U-Net modules. Notably, our proposed approach eliminates the need for exhaustive parameter tuning, a common requirement in traditional STFT methods. Furthermore, it offers the advantage of producing highly accurate high-resolution spectrograms while maintaining computational efficiency compared to other ML-based methods. One of the key strengths of our model is its noise robustness and generalization capabilities, as demonstrated by its successful application in generating μD spectrograms for a challenging experimental dataset consisting of 100 distinct ASL gestures recorded using FMCW radar. Despite being trained only on a simulated dataset, the spectrograms generated by the HRSpecNet outperformed all other methods, including STFT, in terms of classification accuracy. This result underscores the fact that HRSpecNet excels in detecting subtle micro-movements within these activities. These findings open up new avenues for applications in RF-sensing-based activity recognition.

REFERENCES

[1] S. Gurbuz, *Deep Neural Network Design for Radar Applications*. London, U.K.: IET, 2020.

[2] E. Kurtoğlu, S. Biswas, A. C. Gurbuz, and S. Z. Gurbuz, "Boosting multi-target recognition performance with multi-input multi-output radar-based angular subspace projection and multi-view deep neural network," *IET Radar, Sonar Navigat.*, vol. 17, no. 7, pp. 1115–1128, Jul. 2023.

[3] F. Fioranelli, M. Ritchie, and H. Griffiths, "Classification of unarmed/armed personnel using the NetRAD multistatic radar for micro-Doppler and singular value decomposition features," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1933–1937, Sep. 2015.

[4] Z. Ni and B. Huang, "Gait-based person identification and intruder detection using mm-wave sensing in multi-person scenario," *IEEE Sensors J.*, vol. 22, no. 10, pp. 9713–9723, May 2022.

[5] S. Björklund, T. Johansson, and H. Petersson, "Target classification in perimeter protection with a micro-Doppler radar," in *Proc. 17th Int. Radar Symp. (IRS)*, May 2016, pp. 1–5.

[6] A. Huizing, M. Heiligers, B. Dekker, J. de Wit, L. Cifola, and R. Harmanny, "Deep learning for classification of mini-UAVs using micro-Doppler spectrograms in cognitive radar," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 34, no. 11, pp. 46–56, Nov. 2019.

[7] G. Hakobyan and B. Yang, "High-performance automotive radar: A review of signal processing algorithms and modulation schemes," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 32–44, Sep. 2019.

[8] B. Debnath, I. A. Ebu, S. Biswas, A. C. Gurbuz, and J. E. Ball, "FMCW radar range profile and micro-Doppler signature fusion for improved traffic signaling motion classification," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2024, pp. 1–6.

[9] S. Biswas, J. E. Ball, and A. C. Gurbuz, "Radar–LiDAR fusion for classification of traffic signaling motion in automotive applications," in *Proc. IEEE Int. Radar Conf. (RADAR)*, Nov. 2023, pp. 1–5.

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

[11] A. Dahal, S. Biswas, and A. C. Gurbuz, "Comparison between WiFi-CSI and radar-based human activity recognition," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2024, pp. 1–6.

[12] A. M. Alam, M. Kurum, and A. C. Gurbuz, "Radio frequency interference detection for SMAP radiometer using convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 15, pp. 10099–10112, 2022.

[13] 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 Process. Mag.*, vol. 33, no. 2, pp. 71–80, Mar. 2016.

[14] F. Fioranelli and J. L. Kernec, "Contactless radar sensing for health monitoring," in *Engineering and Technology for Healthcare*. Wiley, 2021, pp. 29–59.

[15] X. Li, Y. He, F. Fioranelli, and X. Jing, "Semisupervised human activity recognition with radar micro-Doppler signatures," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5103112.

[16] S. Yang, J. L. Kernec, F. Fioranelli, and O. Romain, "Human activities classification in a complex space using raw radar data," in *Proc. Int. Radar Conf. (RADAR)*, Sep. 2019, pp. 1–4.

[17] S. Biswas, C. O. Ayna, S. Z. Gurbuz, and A. C. Gurbuz, "CV-SincNet: Learning complex sinc filters from raw radar data for computationally efficient human motion recognition," *IEEE Trans. Radar Syst.*, vol. 1, pp. 493–504, 2023.

[18] T. Stadelmayer, A. Santra, R. Weigel, and F. Lurz, "Data-driven radar processing using a parametric convolutional neural network for human activity classification," *IEEE Sensors J.*, vol. 21, no. 17, pp. 19529–19540, Sep. 2021.

[19] S. Biswas, C. O. Ayna, S. Z. Gurbuz, and A. C. Gurbuz, "Complex SincNet for more interpretable radar based activity recognition," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2023, pp. 1–6.

[20] V. Chen, *The Micro-Doppler Effect in Radar*, 2nd ed. Norwood, MA, USA: Artech House, 2019.

[21] B. Erol and M. G. Amin, "Radar data cube processing for human activity recognition using multisubspace learning," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 6, pp. 3617–3628, Dec. 2019.

[22] E. Kurtoglu, A. C. Gurbuz, E. A. Malaia, D. Griffin, C. Crawford, and S. Z. Gurbuz, "ASL trigger recognition in mixed activity/signing sequences for RF sensor-based user interfaces," *IEEE Trans. Hum.-Mach. Syst.*, vol. 52, no. 4, pp. 699–712, Aug. 2022.

[23] 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.

[24] W. Taylor, K. Dashtipour, S. A. Shah, A. Hussain, Q. H. Abbasi, and M. A. Imran, "Radar sensing for activity classification in elderly people exploiting micro-Doppler signatures using machine learning," *Sensors*, vol. 21, no. 11, p. 3881, Jun. 2021.

[25] S. Biswas, B. Bartlett, J. E. Ball, and A. C. Gurbuz, "Classification of traffic signaling motion in automotive applications using FMCW radar," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2023, pp. 1–6.

[26] 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.

[27] F. H. C. Tivive, S. L. Phung, and A. Bouzerdoum, "Classification of micro-Doppler signatures of human motions using log-Gabor filters," *IET Radar, Sonar Navigat.*, vol. 9, no. 9, pp. 1188–1195, Dec. 2015.

[28] T. Thayaparan, S. Abrol, E. Riseborough, L. Stankovic, D. Lamothe, and G. Duff, "Analysis of radar micro-Doppler signatures from experimental helicopter and human data," *IET Radar, Sonar Navigat.*, vol. 1, no. 4, pp. 289–299, Aug. 2007.

[29] B. Ozen Bozdag and I. Erer, "A comparative study on micro-Doppler signature generation methods for UAVs using rotor blade model," in *Proc. 6th Int. Conf. Electr. Electron. Eng. (ICEEE)*, Apr. 2019, pp. 298–301.

[30] S. Nisar, O. U. Khan, and M. Tariq, "An efficient adaptive window size selection method for improving spectrogram visualization," *Comput. Intell. Neurosci.*, vol. 2016, pp. 1–13, Jan. 2016.

[31] S. Kearney and S. Z. Gurbuz, "Influence of radar signal processing on deep learning-based classification," in *Proc. IEEE Radar Conf. (RadarConf23)*, May 2023, pp. 1–5.

[32] G. Izacard, S. Mohan, and C. Fernandez-Granda, "Data-driven estimation of sinusoid frequencies," in *Advances in Neural Information Processing Systems*, vol. 32. Red Hook, NY, USA: Curran Associates, 2019.

[33] P. Pan, Y. Zhang, Z. Deng, and G. Wu, "Complex-valued frequency estimation network and its applications to superresolution of radar range profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5105712.

[34] P. Pan, Y. Zhang, Z. Deng, S. Fan, and X. Huang, "TFA-Net: A deep learning-based time-frequency analysis tool," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 9274–9286, Nov. 2023, doi: 10.1109/TNNLS.2022.3157723.

[35] J. J. Lin, Y. P. Li, W. C. Hsu, and T. S. Lee, "Design of an FMCW radar baseband signal processing system for automotive application," *SpringerPlus*, vol. 5, pp. 1–16, Dec. 2016.

[36] V. C. Chen, D. Tahmoush, and W. J. Miceli, *Radar Micro-Doppler Signatures*. London, U.K.: Institution of Engineering and Technology, 2014.
[37] Y. Kim and B. Toomajian, "Application of Doppler radar for the recognition of hand gestures using optimized deep convolutional neural networks," in *Proc. 11th Eur. Conf. Antennas Propag. (EUCAP)*, Mar. 2017, pp. 1258–1260.
[38] B. Vandersmissen et al., "Indoor person identification using a low-power FMCW radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 3941–3952, Jul. 2018.
[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[40] N. K. Caselli, Z. S. Sehyr, A. M. Cohen-Goldberg, and K. Emmorey, "ASL-LEX: A lexical database of American sign language," *Behav. Res. Methods*, vol. 49, no. 2, pp. 784–801, Apr. 2017.
[41] M. M. Rahman, E. A. Malaia, A. C. Gurbuz, D. J. Griffin, C. Crawford, and S. Z. Gurbuz, "Effect of kinematics and fluency in adversarial synthetic data generation for ASL recognition with RF sensors," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 4, pp. 2732–2745, Aug. 2022.

Sabyasachi Biswas (Graduate Student Member, IEEE) received the B.Sc. degree in electrical and electronic engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 2019. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Mississippi State University, Starkville, MS, USA.

He was a Machine Learning Intern at the High-Performance Computing Collaboratory, MS, USA, in 2022. He is currently a Research Assistant with the Information Processing and Sensing (IMPRESS) Laboratory, Mississippi State University. His research interests include radar signal processing; human activity recognition using radar, camera, and lidar; and developing machine learning algorithms for activity classification using raw radar signals.

Mr. Biswas is a member of the IEEE Signal Processing Society. He was the winner and second runner-up of the Graduate Research Symposium in Fall 2022 and 2023, respectively, held at Mississippi State University.

Ahmed Manavi Alam (Graduate Student Member, IEEE) received the B.S. degree in electrical and electronic engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 2019. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Mississippi State University, Starkville, MS, USA.

He was a Machine Learning Intern at the High-Performance Computing Collaboratory, MS, USA, in Summer 2023. He is currently working as a Research Assistant at the Information Processing and Sensing (IMPRESS) Laboratory, Mississippi State University. His research focus includes algorithm development of deep learning-based inverse problems and machine learning for remote sensing and physics-aware deep learning.

Mr. Alam is a Student Member of the IEEE Geoscience and Remote Sensing Society (GRSS). He was a finalist at the IGARSS 2022 Student Paper Competition. He was a recipient of the Spring 2024 Graduate Research Symposium organized by Mississippi State University and the National Academy of Sciences (NAS) Fellowship.

Ali C. Gurbuz (Senior Member, IEEE) received the B.S. degree in electrical engineering from Bilkent University, Ankara, Turkey, in 2003, and the M.S. and Ph.D. degrees in electrical and computer engineering from Georgia Institute of Technology (Georgia Tech), Atlanta, GA, USA, in 2005 and 2008, respectively.

He held a post-doctoral position at Georgia Tech in 2009, where he researched compressive sensing-based computational imaging problems. He held faculty positions at TOBB University, Ankara, and The University of Alabama, Tuscaloosa, AL, USA, from 2009 to 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. He is currently an Assistant Professor at the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA, where he is also the Co-Director of the Information Processing and Sensing (IMPRESS) Laboratory.

Dr. Gurbuz was a recipient of the Best Paper Award for *Signal Processing* journal in 2013, the Turkish Academy of Sciences Best Young Scholar Award in Electrical Engineering in 2014, and the NSF CAREER Award in 2021. 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*.