Received 10 November 2025; revised 11 December 2025; accepted 25 December 2025. Date of publication 29 December 2025; date of current version 9 January 2026.

Digital Object Identifier 10.1109/OJCOMS.2025.3649235

mmGAN: Semi-Supervised GAN for Improved Gesture Recognition in mmWave ISAC Systems

NABEEL NISAR BHAT[1], SIDDHARTHA KUMAR[2], MOHAMMAD HOSSEIN MOGHADDAM[2], JAKOB STRUYE[1], JESUS OMAR LACRUZ[3] (Member, IEEE), JACOPO PEGORARO[4] (Member, IEEE), JOERG WIDMER[3] (Fellow, IEEE), RAFAEL BERKVENS[1], AND JEROEN FAMAEY[1]

[1]IDLab - imec Research Group, University of Antwerp, 2000 Antwerp, Belgium
[2]Qamcom Research and Technology AB, 41285 Gothenburg, Sweden
[3]IMDEA Networks Institute, 28918 Madrid, Spain
[4]Department of Information Engineering, University of Padua, 35122 Padua, Italy

CORRESPONDING AUTHOR: N. N. BHAT (e-mail: nabeelnisar.bhat@uantwerpen.be)

This work was supported in part by the Hexa-X-II Project, funded by the Smart Networks and Services Joint Undertaking (SNS JU) through the European Union's Horizon Europe Research and Innovation Programme under Grant 101095759; in part by the Research Foundation - Flanders (FWO) Project WaveVR under Grant G034322N; in part by the European Union's Horizon Europe Research and Innovation Programme through SNS-JU under Grant 101192521 (MultiX) and through the Marie Skłodowska-Curie Actions (UNITE) under Grant 101129618; and in part by the Comunidad de Madrid through the DISCO6G-CM Project under Grant TEC-2024/ COM-360 and through the TUCAN6-CM Project under Grant TEC-2024/COM-460, funded under ORDEN5696/2024. The work of Nabeel Nisar Bhat was supported by the Fund for Scientific Research Flanders (FWO) under Grant 1SH5X24N.

ABSTRACT Integrated sensing and communication (ISAC) has gained significant traction in recent years, primarily because it allows existing communication infrastructure to support sensing applications with minimal additional costs. In particular, millimeter-wave (mmWave) ISAC has the potential to offer improved sensing performance in applications such as pose estimation and gesture recognition. For complex sensing tasks and environments, data-driven sensing, which relies on deep learning, is becoming increasingly popular and has shown promising results. However, deep learning models for these tasks require large labeled datasets to achieve high accuracy. Dataset collection and labeling are labor-intensive and time-consuming. Consequently, there is growing interest in leveraging unlabeled data to overcome these challenges. To address this, we propose mmGAN, a semi-supervised method for ISAC-based gesture recognition. We propose a novel loss function for mmGAN based on softplus, feature matching, and manifold regularization to significantly improve gesture recognition performance. We evaluate mmGAN on a 5G Orthogonal Frequency Division Multiplexing (OFDM) mmWave dataset comprising power per beam pair measurements. When training both mmGAN and the supervised baseline with only 0.6% of the labeled data, mmGAN demonstrates up to 25 percentage points higher accuracy than the supervised baseline. Our method serves as a strong foundation for cross-subject transfer learning, demonstrating the significant value of leveraging unlabeled data to enhance cross-domain sensing performance in ISAC systems. Our results demonstrate that the proposed loss function achieves superior performance across diverse subjects. Further, mmGAN significantly narrows the performance gap between semi-supervised and fully supervised models on the publicly available Widar dataset. Moreover, we provide an interpretable analysis of mmGAN performance through saliency maps and ablation studies, revealing key insights into the model's behavior and generalization. This work is the first to evaluate gesture recognition performance in 5G OFDM mmWave ISAC systems using a semi-supervised learning approach, covering the entire pipeline from testbed implementation to model evaluation.

INDEX TERMS 5G, beam sweeping, generative adversarial networks, GAN, integrated sensing and communication, millimeter-wave, semi-supervised learning.

I. INTRODUCTION

INTEGRATED sensing and communication (ISAC) [1], [2], [3] is an emerging technology that leverages existing communication infrastructure for various sensing applications such as gesture recognition [4], [5], pose estimation [6], [7], localization [8], [9] and gait

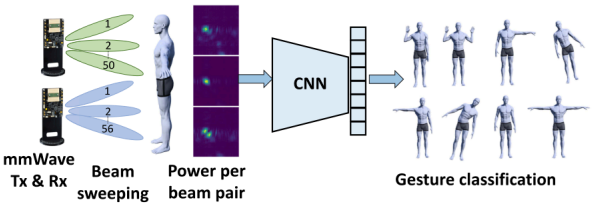


FIGURE 1. 5G OFDM mmWave ISAC gesture recognition system.

identification [10], [11]. ISAC is seen as a key enabler for the future 6G networks [2], [12]. Moreover, recent standardization efforts, such as IEEE 802.11bf for Wi-Fi, further underscore the growing momentum around ISAC [13]. Integrating sensing capabilities into the network can transform it into a "perceptive network" that gathers environmental data. This will play a crucial role in envisioned 6G applications, such as digital twins, extended reality (XR), and network-assisted mobility.

Millimeter-wave (mmWave) ISAC offers additional advantages due to its high bandwidth and the use of large antenna arrays, which enable improved sensing resolution [14]. Recent works in mmWave ISAC show promising results in the sensing applications mentioned earlier [15], [16], [17]. Deep Learning (DL) has been a major driving force behind the success of ISAC for such advanced sensing applications [4], [5]. Data-driven sensing utilizing Deep Neural Networks (DNNs) can automatically learn relevant features directly from raw data, significantly reducing or even eliminating the need for traditional signal processing. In contrast, model-based approaches based on physical theories [18] require efforts to build suitable models and find the right model parameters. Moreover, model-based approaches do not generalize well to complex and dynamic targets and environments, as the modeling effort would be too complex.

However, ISAC-based applications relying on DL face an important challenge. DL-based systems for ISAC rely on fully supervised methods, requiring large amounts of high-quality labeled data to achieve high accuracy. Collecting labeled data is time-consuming, labor-intensive, and often infeasible for applications such as gesture recognition, where each gesture must be manually annotated. Moreover, wireless signals lack cues that can be easily interpreted by humans, which are found in image-based systems, making the annotation process even more complex. As a result, obtaining large and high-quality labeled datasets for ISAC applications is challenging and costly [19], [20]. Consequently, ISAC lags behind computer vision in terms of the availability of large, publicly accessible datasets. This challenge is even more pronounced for mmWave ISAC, due to the scarcity of research-grade hardware and commercial-off-the-shelf (COTS) devices operating at such frequencies.

In contrast, unlabeled data can be collected at a low cost and continuously without the need for manual annotation. The problem lies in harnessing this vast amount of unlabeled data to improve the performance of ISAC-based systems. Recent research emphasizes the potential of utilizing unlabeled data to build more generic foundation models [21], [22], [23]. These models can leverage large volumes of unlabeled data, adapt effectively to various tasks, and reduce the risk of overfitting [24], a common challenge when working with limited labeled data. Leveraging unlabeled data helps mmWave ISAC systems overcome labeled data scarcity, making them more scalable and adaptable to diverse real-world sensing tasks.

In this work, we propose mmGAN, a Semi-Supervised Learning (SSL) framework for mmWave ISAC-based gesture recognition, focusing on in-domain scenarios. mmGAN leverages 5G mmWave OFDM signals and Generative Adversarial Network (GAN) for improved gesture recognition (cf., Figure 1). We propose a novel loss function for mmGAN based on feature matching [25], manifold regularization [26], and softplus loss, to achieve stable and effective semi-supervised training across diverse subjects and under extremely low-label settings. We evaluate mmGAN on a novel 5G mmWave OFDM dataset. Specifically, we extract power per beam pair (PPBP) measurements from 5G signals and input them into mmGAN for classification into one of eight gestures or classes. To the best of our knowledge, this represents a pioneering effort towards developing a comprehensive 5G mmWave dataset for ISAC-based gesture recognition.

We begin by establishing a supervised baseline using a custom Convolutional Neural Network (CNN). We also compare mmGAN to other baselines, such as ResNet18 and a CNN with batch normalization (BN) layers. Then we use the same CNN in the discriminator of mmGAN for a fair comparison. Traditionally, in GANs, the discriminator is a binary classifier. However, we turn the mmGAN discriminator into a multi-class classifier that has dual functionality, i.e., to assign correct labels to the data (standard classifier loss) and to distinguish between real and fake data. This duality results in a GAN loss that combines supervised loss (label loss) and unsupervised loss, allowing the discriminator to learn from both labeled and unlabeled data. We show that the proposed loss function performs better than binary cross entropy (BCE) and the loss proposed in works [25], [27]. mmGAN can therefore significantly improve gesture recognition accuracy across diverse subjects. Moreover, mmGAN beats Auto-Fi [28], a state-of-the-art self-supervised method for Wi-Fi sensing, across all percentages of labeled data. Additionally, we evaluate our method on the sub-6 GHz Widar dataset [29], a widely recognized benchmark for wireless sensing, where our approach demonstrates superior performance. Furthermore, we go beyond accuracy metrics and dig deeper to analyze mmGAN's performance. We examine its feature representations, report class-wise metrics such as F1-scores, and provide an interpretability analysis with Explainable AI (XAI) that offers novel insights into the model's behavior and generalization.

A. CONTRIBUTIONS

The contributions of this work can be summarized as follows:

1) *5G mmWave ISAC system*: We are the first to develop and evaluate a 5G OFDM mmWave ISAC system for gesture recognition covering the whole pipeline, from testbed implementation, dataset collection,[1] to model evaluation. This contrasts most prior practical ISAC works, which rely on Wi-Fi signals [4], [7], [30], [31].
2) *Semi-supervised GAN framework (mmGAN):* To leverage unlabeled data, we propose mmGAN, a GAN-based approach designed to enhance the generalization of mmWave ISAC in label-scarce scenarios. Our design enables mmGAN to utilize labeled, unlabeled, and generated data for learning. This stands in contrast with state-of-the-art GAN-based approaches, such as W-GAN [30], WiGAN [32], and CSI4Free [33], which do not support the use of unlabeled data. Leveraging unlabeled data not only improves the decision boundaries of mmGAN for classification but also encourages it to learn generalized representations, a step towards foundation learning. Moreover, the discriminator of mmGAN is a lightweight CNN that can be deployed on resource-constrained devices.
3) *Robust loss for low-label ISAC:* We propose a novel loss function for mmWave ISAC that combines the advantages of softplus, feature matching, and manifold regularization. This addresses the problem of unstable GAN training, poor feature discrimination, and generalization across diverse distributions under low-label settings. Further, the proposed loss ensures meaningful gradient updates to the generator even when the discriminator saturates, addressing the problem of vanishing gradients (such as in BCE loss). Consequently, this results in better classification performance across diverse subjects than BCE, Least Squares [27], and feature matching loss [25].
4) *PPBP-based sensing:* In terms of ISAC, our work's novelty lies in the utilization of PPBP, rather than channel state information (CSI) or micro-Doppler. PPBP is more robust (e.g., to phase noise) and practical (e.g., not requiring advanced synchronization). Moreover, the PPBP measurements can be easily extracted from the communication signals, rather than relying on dedicated sensing packets. This makes our approach especially suitable for ISAC, entailing no overhead or additional resource utilization.
5) *Comprehensive evaluation and interpretability:* We address the problem of quantitatively evaluating the quality of the generated samples, which is made more challenging by the difficulty of visualizing radio signals compared to, e.g., images. To this end, we reconsider the evaluation metric for GANs in ISAC, adopting Fréchet Classifier Distance (FCD). We dig deeper inside mmGAN and compute class-wise metrics, and examine the internal feature space to gain additional insights. Further, we provide an interpretable analysis of mmGAN's performance through XAI techniques by computing global saliency maps that reveal the model's attention, identifying the key areas of focus, and performing an ablation study to validate the importance. This provides novel insights into the model's behavior and generalization, which prior GAN works on the ISAC lack.

Several studies [30], [34], [35] have employed GANs to improve generalization. However, such works [34], [35] concentrate on cross-domain scenarios (leave-one-user-out settings). In contrast, our work focuses on in-domain scenarios. In our setting, the labeled, unlabeled, and test data belong to the same domain (same subject). Our objective is to narrow down the performance gap with a fully supervised baseline by leveraging a minimal amount of labeled data. Specifically, we target extremely low-label scenarios, such as a 5-shot setting [36], to demonstrate the effectiveness of our approach. The evaluation of diverse baselines, loss functions, and generator architectures reveals notable findings, delivering important lessons for the research community to advance data-driven generative research for ISAC in SSL settings.

B. PAPER STRUCTURE

The remainder of the paper is structured as follows: Section II explains the motivation behind SSL and the utilization of unlabeled data. Section III reviews related work on gesture recognition and SSL. Section IV provides background on SSL, GAN, and Beam Sweeping. Section V outlines the ISAC system model. In Section VI, we present our method, mmGAN. Section VII covers the experimental settings and results. Section VIII details the advanced studies. Sections IX and X present the discussion and conclusion, respectively. Finally, Section XI covers the limitations and directions for future work.

II. MOTIVATION: LEARNING FROM UNLABELED DATA

Current ISAC-based systems that utilize deep learning require substantial volumes of labeled data to achieve good performance. The process of labeling this data is both costly and labor-intensive, as each training example must be annotated [19]. Additionally, the expense of labeling data for each new user or environment presents a significant obstacle to deploying these systems. The complexity of labeling varies across different ISAC tasks. For instance, person identification involves straightforward labeling. In contrast, labeling becomes more complicated for gesture recognition, as it requires annotating multiple gestures, each repeated numerous times. Pose tracking, on the other hand, demands a camera synchronized tightly with the ISAC system to obtain the ground truth accurately.

Our mmGAN addresses this challenge by utilizing a small portion of labeled data and a large amount of unlabeled data from the same domain. This differs from works that focus on cross-domain settings [34], [35]. In the SSL setting, training

[1]The dataset is available via https://ieee-dataport.org/documents/mmhsense-multi-modal-and-distributed-mmwave-isac-datasets-human-sensing.

examples such as gestures or poses are manually labeled for a subset of the data or by employing auxiliary systems such as cameras or other motion capture systems. The ISAC system continuously gathers unlabeled data as users interact with it. This data can be leveraged to further train the system, enabling the model to learn valuable features without the need for explicit labels.

By using a small portion of labeled data, we cut down the time and expense required for annotation, making the system much more practical. Moreover, as users interact with the system, their behavior may vary over time, meaning that the gestures might be performed differently or at different paces, leading to a distribution gap between the training and test data. By occasionally collecting unlabeled data as performance declines, the system can adapt to these subtle changes, ultimately improving accuracy. Learning from unlabeled data has gained significant attention within the research community, especially in the context of foundation learning [23], where the focus is on leveraging large quantities of unlabeled data to improve model performance and adaptability to diverse downstream tasks. Learning from unlabeled data allows the model to learn something generic, and we hypothesize that our method trained in such a way can therefore serve as a strong initializer for cross-subject transfer learning [37].

III. RELATED WORK

This section covers recent developments in gesture recognition with mmWave ISAC and the use of SSL techniques in ISAC applications.

A. GESTURE RECOGNITION WITH mmWAVE ISAC

While the sub-6 GHz ISAC has advanced significantly, progress in mmWave ISAC has been relatively limited. This is primarily due to hardware costs and availability, limited large-scale public datasets, and the complexities of data collection, including high sampling rates and memory constraints. In this section, we examine the major contributions to mmWave ISAC, particularly in the context of gesture recognition. Yu et al. [16] made a pioneering contribution to mmWave Wi-Fi sensing. The authors used mid-grained beam Signal-to-Noise Ratio (SNR) for human pose classification. A single user performed a set of 8 distinct poses. The authors used a DNN to extract features from different poses and were able to achieve around 89% accuracy in pose classification. Bhat et al. [17] extracted mmWave CSI from 60 GHz Wi-Fi access points (APs). The authors collected a dataset of 3 users performing a set of 8 poses. Microsoft Kinect was used as a ground truth. Both pose regression and pose classification were considered. A CNN was used to map changes in CSI to different poses. The authors reported 93.6% accuracy in classifying poses on a combined dataset of 3 users. For pose estimation, the authors achieved an MSE of 0.0048 between the true skeletons and predicted ones using mmWave Wi-Fi. Recently, Pegoraro et al. [20] published a mmWave dataset for ISAC-based activity recognition. The authors used IEEE 802.11ay-compliant packets. Different from the previous works, a 60 GHz SDR is used to collect a dataset of channel estimates to perform human sensing tasks. The dataset consists of 7 subjects performing a set of 4 activities.

While there has been limited yet promising research in mmWave ISAC, studies such as those by Yu et al. [16], and Bhat et al. [17] have focused on a small number of participants, with only 1 and 3 users, respectively. Similarly, Pegoraro et al. [20] included 7 users but restricted their analysis to just 4 activities. In contrast, our work addresses these limitations by collecting a comprehensive mmWave dataset involving 8 users, each performing 8 distinct gestures that include both dynamic movements and static poses, covering a wide range of motion types. This provides more opportunities to study the domain gap and improve transfer learning techniques. Moreover, while the aforementioned studies focus on transmitting packets compliant with mmWave Wi-Fi standards IEEE 802.11ad and 802.11ay standards, our ISAC system utilizes the 5G NR cellular OFDM waveform. To our knowledge, this is the first study to leverage 5G OFDM mmWave communication signals for gesture recognition. This forms the basis for ISAC integration in future 6G standards. Moreover, existing methods that depend on channel impulse response (CIR) or CSI are not easily accessible on COTS devices without requiring software modifications. In contrast, we leverage PPBP robust to hardware imperfections such as phase errors. PPBP is computed in hardware with minimal complexity, is accessible to users on COTS devices, and additionally reduces the complexity of the method and the size of DNN.

B. SEMI-SUPERVISED LEARNING

In various fields, such as text, speech, images, and wireless sensing, there is a vast amount of unlabeled data available, while obtaining labeled data can be expensive. This has sparked considerable interest in SSL [38], which leverages both labeled and unlabeled data for training. Njima et al. [30] propose a semi-supervised approach to enhance indoor localization accuracy using Wi-Fi signals. They introduce two solutions: one based on pseudo-labeling and another utilizing a GAN. In the first approach, a DNN generates pseudo-labels for the unlabeled data by leveraging the available labeled data. These pseudo-labeled and real labeled datasets are then combined to train a generic localization model. Coefficient weights are applied to emphasize the most confident samples, thereby increasing their impact on the training process. In the second approach, the authors use a conditional GAN to generate synthetic Received Signal Strength Indicator (RSSI) data along with corresponding coordinates (labels). The real and synthetic data are then merged to train a localization-focused DNN, with coefficient weights employed to limit the influence of the less confident samples. The proposed methods improve localization accuracy by 10.11% and 8.53%, respectively.

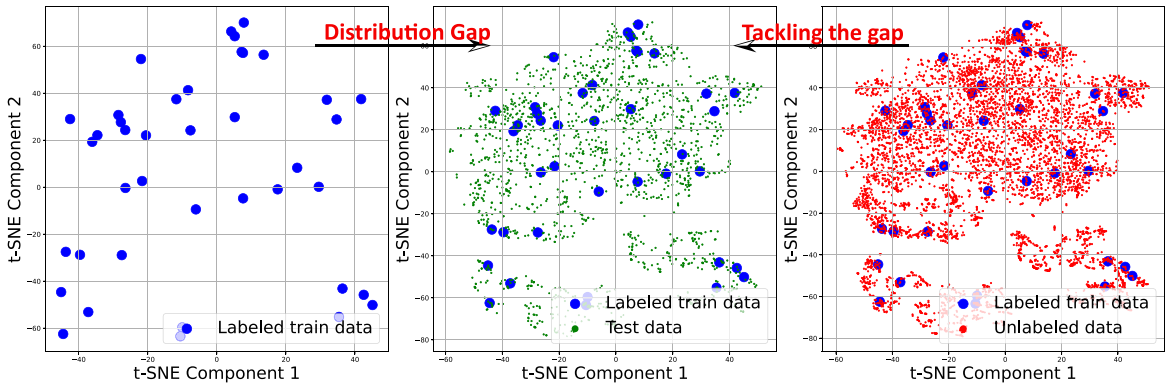


FIGURE 2. t-SNE embeddings of high-dimensional input data: In-domain distribution gap.

Recently, self-supervised learning has also been used in Wi-Fi sensing to improve the performance of the downstream classifier. Yang et al. [28] propose Auto-Fi, an annotation-efficient method for human activity recognition based on Wi-Fi signals. The Auto-Fi framework leverages contrastive learning and mutual information and employs a novel geometric structural loss, which helps Auto-Fi to enable various downstream tasks. The authors use a traditional contrastive learning approach, i.e., augmenting CSI and generating two views, and then employing a geometric structural loss. After unsupervised pre-training, a supervised calibration module can transfer the knowledge to the required downstream tasks. This method is similar to the work by Chen et al. [39]. The method outperforms other SSL methods on the UT-HAR dataset under 10-shot and 20-shot settings. Similarly, Bocus et al. [40] present a self-supervised learning framework based on contrastive learning to improve the performance of the supervised method. The authors use an Alex-Net-based encoder for unsupervised pre-training. A Normalized Temperature Cross-entropy (NT-Xent) is employed for contrastive loss. Supervised fine-tuning on a small labeled dataset follows unsupervised pre-training. The results reveal a 22% increase in macro F1 score when only 1.29% of labeled training samples are considered in the fine-tuning stage.

The existing research demonstrates the potential of SSL, particularly in enhancing the performance of supervised methods by utilizing readily available unlabeled data. Pseudo-label-based approaches [30] are highly sensitive to noise, as the generated pseudo-labels can be unreliable. Self-supervised methods [28], [39], in contrast, have shown promising results in image classification but rely heavily on strong augmentations like translation and rotation, which are unsuitable for wireless signals, as the underlying semantic content may not be preserved under these augmentations. GAN-based methods offer an alternative by generating synthetic data that mirrors the distribution of real data, thereby helping classifiers establish robust decision boundaries without relying on hand-crafted augmentations. Additionally, GANs' ability to learn complex data distributions makes them essential components in developing foundation models [23], a class of Generative AI designed to learn from vast amounts of unlabeled data for tasks ranging from text generation to image synthesis. Inspired by this, we propose mmGAN, a GAN-based method to effectively leverage both unlabeled data and a small set of labeled data for enhanced mmWave gesture recognition. We propose a novel loss function that combines the advantages of softplus, manifold regularization, and feature matching, and show its advantages against BCE and feature matching loss. To the best of our knowledge, this is the first attempt to use GANs for leveraging unlabeled data to improve ISAC-based mmWave gesture recognition performance in in-domain scenarios. Our method outperforms Auto-Fi [28] at all percentages of labeled training data. Moreover, mmGAN shows improved generalization in cross-domain transfer learning with limited data from the target domain. Additionally, the discriminator of mmGAN occupies only 0.57 MB of memory, with only 150K parameters and an average inference latency of 0.6 milliseconds, making it suitable to be deployed on resource-constrained devices.

IV. BACKGROUND

A. SEMI-SUPERVISED LEARNING

In classical machine learning problems, we have a set of M ordered labeled data points $D_M = \{(x_m, y_m)\}_{m=1}^{M}$, where each x_m is the data and y_m is its corresponding label. D_M essentially represents the labeled data. Based on this ordered pair, supervised algorithms attempt to learn a function or mapping $f : X \rightarrow Y$ such that for a given input $x \in X$, the predicted output $\hat{y} = f(x)$ is as close as possible to the true label y for any unseen data point x'. However, in many real-world problems, we also have access to U unlabeled data points, $D_U = \{x_u\}_{u=1}^{U}$ such as test data, on which we want to make predictions. Although the labels y are unknown for these data points, the features can still be useful for improving the model, especially in SSL or unsupervised learning settings.

Leveraging both labeled and unlabeled data in machine learning offers significant advantages, such as reducing the

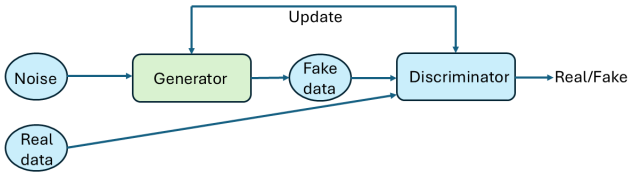


FIGURE 3. GAN architecture: Generator generates fake data while discriminator tries to identify fake and real data.

costs and time associated with data labeling, enhancing model generalization to unseen environments, and improving classification accuracy. Figure 2 shows the t-distributed stochastic neighbor embedding (t-SNE) [41] of high-dimensional PPBP data for one of the subjects, from the collected dataset. The blue dots denote labeled samples, the green dots denote the test data, and the red dots indicate unlabeled samples from the same domain (subject). A classifier trained exclusively on the limited labeled blue points is likely to exhibit poor generalization to the unseen green points, even within in-domain scenarios, due to the distribution gap. In contrast, incorporating unlabeled data that follows a similar distribution to the unseen test data can significantly improve the robustness and accuracy of the classifier. SSL attempts to construct a classifier where the number of M data points is significantly smaller than the number of U points.

B. GENERATIVE ADVERSARIAL NETWORKS (GANs)

GANs [42] consist of two neural networks, a generator, and a discriminator, competing against each other and improving at the cost of the other. In classic GANs (cf., Figure 3), the generator tries to generate high-dimensional synthetic data while the discriminator aims to separate the generated and real distribution. Based on the feedback from the discriminator, the generator improves its output and finally can generate data that is indistinguishable from the real data. The generator, $\mathcal{G}$, takes random Gaussian noise (z) as an input and generates a high-dimensional output $\mathcal{G}(z)$ such as an image, audio signal, or wireless signal. On the other hand, the discriminator, $\mathcal{D}$, has access to generated and original data and is trained explicitly using the pseudo labels (domain labels) associated with the generated and original data. The discriminator is usually a binary classifier that outputs the probability of a sample being real. Ideally, the discriminator produces an output close to 1 for real samples and near 0 for generated or synthetic ones. The generator has to generate realistic synthetic data to fool the discriminator. Once the GAN training converges, the discriminator can ideally no longer distinguish between the two data. After training, the discriminator can be discarded as it was used just to train the generator. The trained generator can then generate highly realistic and diverse synthetic data.

The discriminator is essentially a binary classifier whose goal is, for example, to produce 1 for a real sample (x) and 0 for a sample coming from the generator $\mathcal{G}(z)$. So, we can use BCE loss to train it. The standard BCE loss is defined as follows:

$$L_{\text{BCE}}(y, \hat{y}) = -\big[\, y \log(\hat{y}) + (1-y) \log(1-\hat{y}) \,\big] \tag{1}$$

where $y \in \{0, 1\}$ is the true label and $\hat{y}$ is the predicted probability. For a real sample x with label $y = 1$, the discriminator's loss becomes:

$$\begin{aligned} L_{\mathcal{D}}(x, 1) &= -\big[\, 1 \cdot \log(\mathcal{D}(x)) + (1-1) \log(1-\mathcal{D}(x)) \,\big] \\ &= -\log(\,\mathcal{D}(x)\,) \end{aligned} \tag{2}$$

So, the discriminator should minimize $\log(\mathcal{D}(x))$ which is minimized if $\mathcal{D}(x) \to 1$.

For a fake sample $G(z)$ with label $y = 0$, the discriminator loss becomes:

$$\begin{aligned} L_{\mathcal{D}}(G(z), 0) &= -\big[\, 0 \cdot \log(\mathcal{D}(G(z))) \\ &\qquad + 1 \cdot \log(1 - \mathcal{D}(G(z))) \,\big] \\ &= -\log(1 - \mathcal{D}(\mathcal{G}(z))) \end{aligned} \tag{3}$$

Here, the discriminator should minimize $-\log(1 - \mathcal{D}(G(z)))$ which is minimized if $\mathcal{D}(\,G(z)) \to 0$. Combining the two losses, the overall discriminator loss is:

$$L_{\mathcal{D}} = -\log(\,\mathcal{D}(x)) - \log(\,1 - \mathcal{D}(\mathcal{G}(z))) \tag{4}$$

The goal of the discriminator is to minimize this loss. Removing the negative sign converts the minimization into a maximization problem. Hence, the final discriminator loss for a single data point can also be written as:

$$L_{\mathcal{D}} = \max\big[\, \log(\mathcal{D}(x)) + \log(1 - \mathcal{D}(G(z))) \big] \tag{5}$$

On the other hand, the generator's loss has only one term, as it does not get to see the real samples:

$$L_{\mathcal{G}} = \min\big[\, \log(1 - \mathcal{D}(G(z))) \,\big] \tag{6}$$

The generator's goal is to fool the discriminator so that $\mathcal{D}(G(z)) \to 1$, i.e., it wants to minimize its loss by maximizing the discriminator's output for fake samples. This encourages the generator to produce samples that the discriminator classifies as real.

This optimization of GAN is a two-player min–max problem [43]. The aim is to reach the Nash equilibrium [44], a point where no player can improve by changing its weights. The overall objective of the GAN is:

$$\begin{aligned} \max_{\mathcal{D}} \min_{\mathcal{G}} V(\mathcal{D}, \mathcal{G}) = \; & \mathbb{E}_{x \sim p_{\text{data}}(x)}\big[\, \log(\mathcal{D}(x)) \,\big] \\ & + \mathbb{E}_{z \sim p_z(z)}\big[\, \log(1 - \mathcal{D}(\mathcal{G}(z))) \,\big] \end{aligned} \tag{7}$$

where $V(\mathcal{D}, \mathcal{G})$ denotes the overall loss function, $\mathbb{E}_x$ is the expected value over all real data instances, and $\mathbb{E}_z$ is the expected value over all random inputs to the generator.

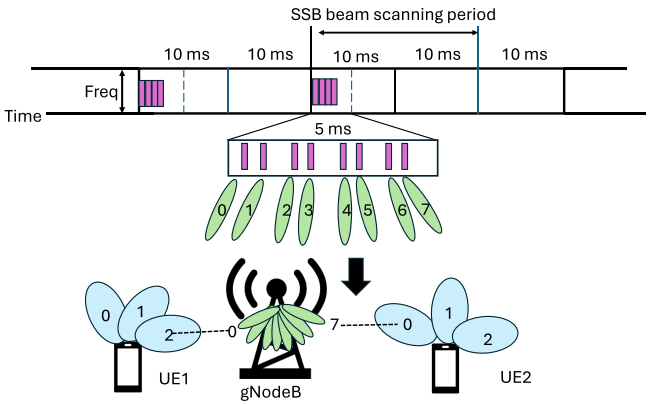


FIGURE 4. 5G OFDM radio frame and SSB Beam Sweeping.

C. BEAM SWEEPING

Large antenna arrays at 5G mmWave frequencies enable highly directional communication through narrow beams, effectively mitigating significant propagation losses [14]. To support this, mmWave systems utilize a technique called beam sweeping [45], [46], [47], where the base station periodically transmits beams in all directions. This process allows the end user to identify and connect with the strongest beam for reliable communication. Beam sweeping is an integral feature of the 5G mmWave standard, allowing access to beam signal-to-noise ratios (SNRs) without incurring additional overhead, unlike CSI or micro-Doppler measurements. These SNR values serve as reliable indicators of channel quality and can be effectively utilized for sensing applications.

Figure 4 shows the 5G OFDM radio frame, consisting of 10 milliseconds duration [48], and the synchronization signal block (SSB) beam sweeping mechanism in the 5G system, where the base station (gNodeB) transmits an SSB beam in different directions that helps the user equipment (UE) find the best beam. SSBs are sent in a time division manner, with each SSB beam having an associated unique number, i.e., SSB index. The default beam scanning period is 20 milliseconds. Each round of SSB beam scanning takes 5 milliseconds. The maximum number of beams depends on the central frequency. In the case of FR2 (24.25 GHz to 71.0 GHz), the maximum number of beams is 64. The UEs measure the SSB signal through wide beam scanning to determine an optimal beam for communication. For example, according to Figure 4, UE1 selects beam 2 and UE2 selects beam 0 as an optimal beam. This periodic beam sweeping process, designed for link establishment and beam alignment, captures rich spatial and temporal information about the propagation environment. Consequently, it can be leveraged for environmental and human sensing.

V. ISAC SYSTEM MODEL: PPBP AS A SIGNAL FEATURE

Figure 1 illustrates the proposed system pipeline for gesture classification using mmWave signals. The system consists of bi-static 5G mmWave nodes functioning as transmitter (Tx) and receiver (Rx). We use two Sivers mmWave evaluation kits (EVK06002) with a front end supporting a frequency range of 57-71 GHz. The EVKs are controlled using Radio Frequency System-on-Chip (RFSoC). The system implements a 5G-NR OFDM waveform consisting of SSB and random data over a single carrier with 792 subcarriers. The subcarriers are spaced 960 kHz apart, having a total bandwidth of approximately 760 MHz. The Tx continuously transmits a single frame of 10 milliseconds as per the 5G standard. Each frame consists of 10 subframes, and each frame is made up of 112 OFDM symbols with a Cyclic prefix (CP) of varying length (in symbols) and a 1024-point Fast Fourier Transform (FFT) block. Beam sweeping is performed across 50 Tx and 56 Rx beams, with 2 OFDM symbols transmitted per beam pair during the process.

The RFSoC has been configured to sample the received signal in such a way that exactly two OFDM symbols—minus the corresponding CPs—exist per Tx-Rx beam pair. We configure RFSoC to convolve the raw time domain Rx signal with its counterpart, the raw time domain Tx signal without the CPs. The RFSoC uses an efficient convolution implementation using FFTs; hence, it computes the convolution block-wise, with each block having a size at most equal to the FFT size (1024 samples). We use the convolved signal to compute a 50x56 power grid. This is achieved by first computing the power of the convolved signal, then averaging over the number of samples and OFDM symbols to obtain an array of average PPBP. Finally, the grid is obtained by reshaping the resulting array into a matrix of size 50×56. The time-series representation of the power grid captures the unique characteristics of the gesture, which is then provided as input to the mmGAN for gesture classification.

Since the grid is formed from beam-specific measurements, the choice of beamwidth directly impacts both performance and efficiency. The 60 GHz Sivers EVK employs a 16-element horizontal phased-array antenna (PAA). The vendor's predefined beamforming codebook includes a set of beam patterns (BPs) with an approximate half-power beamwidth of 6°. The narrower the beam, the higher the spatial resolution, but then it also means more beams are needed to cover the target area. The smaller beamwidth can result in more detailed information about the user's gesture at the cost of increased computational complexity. Further, sensing resolution is also affected by the number of Tx and Rx beams and the rate of beam sweeping. A higher sweeping rate improves temporal resolution, while a larger number of beams enhances spatial resolution, leading to a higher spatio-temporal sensing accuracy. However, this may reduce communication throughput due to the increased time spent on beam sweeping operations. We plan to quantify this trade-off in future work. In this work, our focus is on the sensing aspect of ISAC.

Most existing works [4], [20], [34] employ complicated inputs to the DNN, such as CIR, CSI, and micro-Doppler,

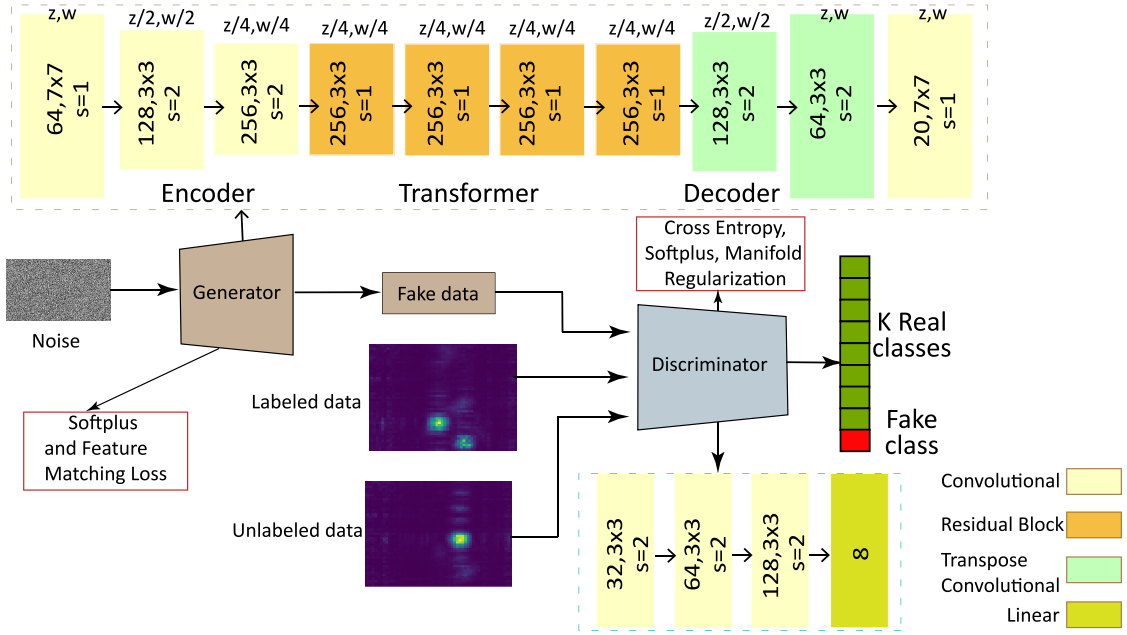


FIGURE 5. Semi-Supervised mmGAN: The generator takes random Gaussian noise as input to generate fake data. The discriminator-turned-classifier has access to three types of data. For labeled data, the discriminator classifies the input into one of the *k* real classes. For the unlabeled and fake data, the discriminator is trained to separate the two distributions apart.

which are not readily available in COTS devices and complicate the size of the DNN. While CSI is part of the Wi-Fi 802.11 standard, obtaining it often requires firmware modifications or special drivers, making it less accessible to non-expert users. Furthermore, in bi-static or multi-static configurations, such setups require tight synchronization in terms of timing, carrier frequency offset (CFO), and clock, between Tx and Rx [49], [50]. Moreover, most micro-Doppler-based approaches require fixed inter-frame spacing, which is not always feasible, or rely on sparse-based methods [51], which involve significant computational complexity. Additionally, the presence of phase noise or phase offsets in the received signal necessitates careful recalibration of the phase component. In contrast, our system relies on simple power measurements, which are computed in hardware with very low complexity and readily available to users in COTS devices, and does not require any calibration or sanitization. Finally, it results in reduced complexity of the DNN.

3GPP Release 18 introduces several enhancements in beam management and positioning that form a solid basis for power-based sensing methods. Our gesture recognition approach builds directly on these mechanisms, using only per-beam received-power measurements that are already available through standard beam-sweeping and measurement processes. Our methodology does not modify the 5G protocol stack and can operate with existing measurement reports (beam sweeping). As such, it is fully compatible with current and upcoming 5G-Advanced systems and could naturally fit within the sensing and beam-management framework being developed toward Release 19.

VI. METHODOLOGY

In this section, we present mmGAN, a method to improve the performance of a downstream classifier in label-scarce ISAC systems, specifically in in-domain scenarios, where the labeled and unlabeled data belong to the same domain (same user). Figure 5 illustrates the schematic representation of our SSL framework, where the objective is to boost the classification performance of the downstream classifier (discriminator). In this SSL setting, we transform the binary discriminator of a GAN into a multi-class classifier. Now the roles are reversed, as by the end of the training process, the generator can be discarded, having served its primary role of assisting the discriminator during training. The generator acts as an auxiliary source of information, providing the discriminator with additional unlabeled training data. Traditionally, in GANs, the discriminator's primary function is to distinguish between real and fake inputs. However, in the proposed mmGAN framework, the discriminator takes on a broader role: it not only identifies whether inputs are real or fake but also learns to predict class probabilities for the original dataset [25]. This expanded role allows the discriminator to be trained on three types of data: labeled data D_M, unlabeled data D_U, and generated data $\mathcal{G}(z)$, all of which contribute to improving classification performance.

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the University of Antwerp Ethics Committee for the Social Sciences and Humanities (EA SHW) under Application No. SHW_2023_313_2. Informed consent was obtained from the participants.

A. mmGAN

The discriminator (cf., Figure 5) is a lightweight CNN borrowed from baseline CNN (cf., Section VII-B) except for using LeakyReLU activations instead of ReLU as per standard practices for training GANs [52]. The filters double every layer to increase the capacity of the network as the spatial dimensions of the input shrink. The discriminator then outputs logits (raw scores) for each class, and an additional linear layer is used to compress this information into a single logit, with a sigmoid activation applied on top of it. This approach, based on BCE loss, does not show good classification performance. As such, we instead propose an improved solution based on the work by Salimans et al. [25]. We turn the discriminator into a 9-class classifier, 8 real classes (for 8 gestures), and one for the fake one. However, the softmax distribution in such a case is overparameterized, and the unnormalized logit (l_{fake}) for a sample to be fake (coming from the generator) can be set to 0. The probability of it being real becomes:

$$p(x) = \frac{Z(x)}{Z(x) + \exp(l_{\text{fake}})} = \frac{Z(x)}{Z(x)+1} \tag{8}$$

where $Z(x) = \sum_{k=1}^{K} \exp[l_k(x)]$ is the sum of unnormalized probabilities, x represents the input (cf., Equation (7)), and K represents the number of classes. So, after the final layer of the baseline model, we use Equation (8) as a final activation function to get the probability of a sample being real. However, keeping everything in exponential-space leads to the problem of exploding activations [53], where $Z(x)$ becomes very large and approaches ∞. As a result, $p(x)$ is no longer bounded between 0 and 1. We solve this problem by computing the activation in log-space, as it is numerically safer, and we can represent very small numbers. We can therefore compute the loss of the discriminator from Equation (7) as:

$$\begin{aligned} L_{\mathcal{D}} = &-\mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log\left(\frac{Z(x)}{Z(x)+1}\right)\right] \\ &-\mathbb{E}_{z\sim p_z(z)}\left[\log\left(1-\frac{Z(\mathcal{G}(z))}{Z(\mathcal{G}(z))+1}\right)\right] \end{aligned} \tag{9}$$

Simplifying the right-hand side of the equation further:

$$\begin{aligned} L_{\mathcal{D}} &= -\mathbb{E}_{x\sim p_{\text{data}}(x)}\big[\log(Z(x))\big] \\ &\quad +\mathbb{E}_{x\sim p_{\text{data}}(x)}\big[\log(Z(x)+1)\big] \\ &\quad -\mathbb{E}_{z\sim p_z(z)}\left[\log\left(\frac{1}{Z(\mathcal{G}(z))+1}\right)\right] \\ &= -\mathbb{E}_{x\sim p_{\text{data}}(x)}\big[\log(Z(x))\big] \\ &\quad +\mathbb{E}_{x\sim p_{\text{data}}(x)}\big[\text{softplus}(Z(x))\big] \\ &\quad +\mathbb{E}_{z\sim p_z(z)}\big[\text{softplus}(Z(\mathcal{G}(z)))\big] \end{aligned} \tag{10}$$

where $L_{\mathcal{D}}$ denotes the unsupervised discriminator loss. The softplus function has a smoother and linear gradient and does not suffer from a vanishing gradient, unlike the BCE loss. The first two terms in Equation (10) represent the discriminator loss on real unlabeled data, while the third term represents the discriminator loss on generated data. On the other hand, for the generator loss, only the second term of Equation (7) is relevant, as the generator does not get to see the real data. Therefore, the generator loss can be defined as:

$$\begin{aligned} L_{\mathcal{G}} &= \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1-\frac{Z(\mathcal{G}(z))}{Z(\mathcal{G}(z))+1}\right)\right] \\ &= \mathbb{E}_{z\sim p_z(z)}\big[-\log(Z(\mathcal{G}(z))+1)\big] \\ L_{\mathcal{G}} &= \mathbb{E}_{z\sim p_z(z)}\big[-\text{softplus}(Z(\mathcal{G}(z)))\big] \end{aligned} \tag{11}$$

$L_{\mathcal{G}}$ denotes the generator loss. Our generator architecture, Residual-G (leveraging residual connections), is an adaptation of the model from the work Zhu et al. [54]. It consists of an encoder, transformer, and decoder blocks (cf., Figure 5). The blocks depict the kernel size (A×A) and output features or channels (F_c). Initially, the input is processed through a convolutional layer, with output features $F_C = 64$, $A = 7$, and stride (s) of 2, followed by instance normalization and ReLU activation. For brevity, normalization and activations are not shown in Figure 5. The encoder block then reduces the input's spatial dimensions (z,w) while increasing the number of channels. This block consists of two convolutional layers, each followed by instance normalization and ReLU activation, which compress the input data and extract essential features. To address the loss of information during compression and reconstruction, we introduce a transformer block after the encoder. This block includes residual connections to preserve information across layers, comprising two residual blocks, each with two convolutional layers. These connections ensure that important details from the compressed data are retained and effectively utilized. Finally, the decoder block utilizes transpose convolution layers to upsample the data back to its original dimensions. The process concludes with a final convolutional layer followed by a Tanh activation function, which scales the output to the range of –1 to 1, matching the original input's scale.

Furthermore, we incorporate manifold regularization [26] and feature matching [25] to the discriminator and the generator losses, respectively. The regularization forces the discriminator to yield similar logits for nearby points in the latent space in which z resides and prevents the discriminator from overfitting. It is implemented by slightly perturbing z and computing the discriminator's and generator's output one more time. The regularization can be implemented as:

$$L_{MR} = MSE\big(\mathcal{D}(\mathcal{G}(z)),\ \mathcal{D}(\mathcal{G}(z'))\big) \tag{12}$$

where $z' = z + \mathcal{N}(\mu, \sigma)$ represents the perturbed noise, and MSE refers to the mean squared error. We perturb z using $\sigma = 10^{-4}$. Larger σ values ($> 10^{-4}$) resulted in stability issues during training. For the generator, feature matching regularization has been shown to be highly effective, especially in stabilizing training and ensuring faster convergence.

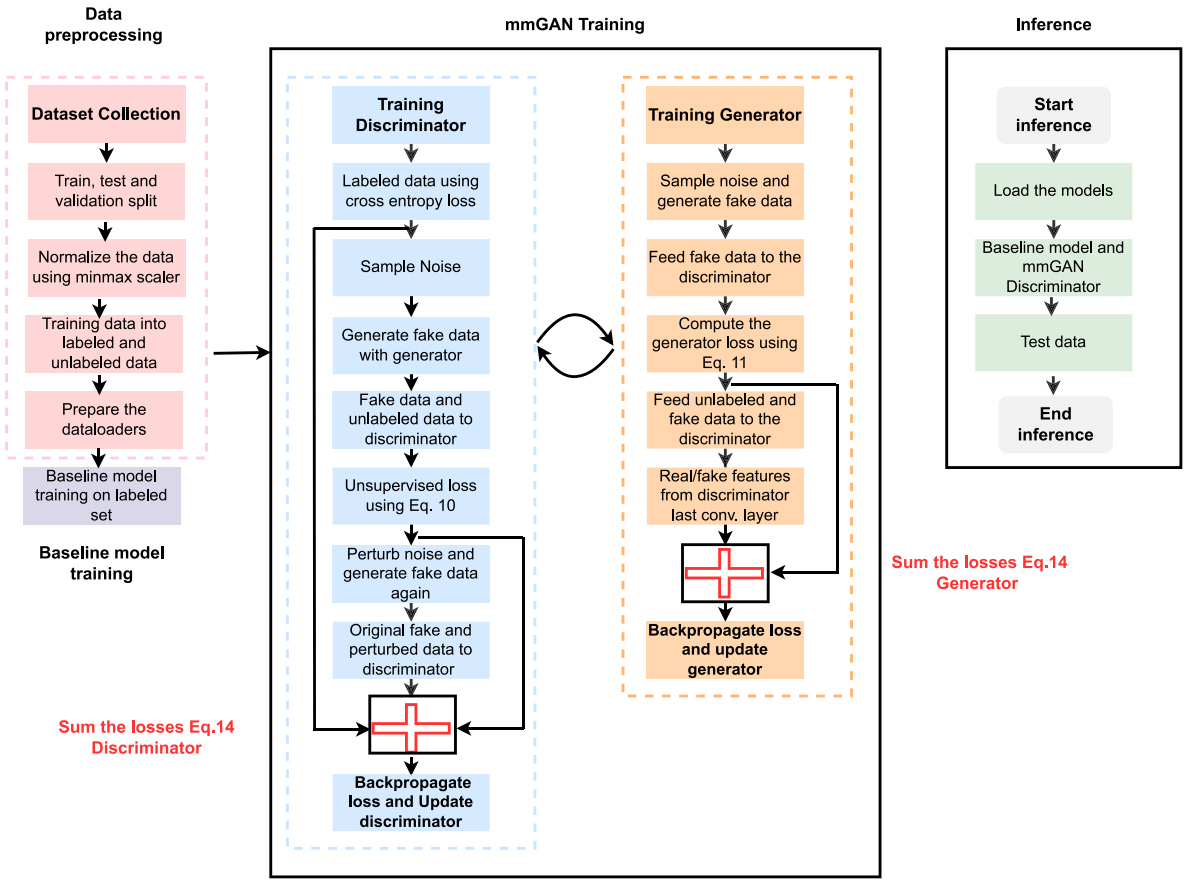


FIGURE 6. Flowchart illustrating the schematic representation of our methodology: data preparation, mmGAN training, and inference.

The feature matching objective encourages the generator to match the statistics of the real data. The generator tries to minimize the statistical difference between the features of the real images and the generated images. Feature matching can be implemented by taking the output of the final convolution layer of the discriminator (feature extractor) on fake and real data and minimizing the distance between them, usually using mean absolute error (MAE).

$$L_{FM} = MAE(\mathcal{D}_F(x), \mathcal{D}_F(\mathcal{G}(z))) \tag{13}$$

where $\mathcal{D}_F$ represents the backbone or the feature extraction layers of the discriminator. The overall loss for $\mathcal{G}$ and $\mathcal{D}$ becomes:

$$\begin{aligned} L_{\mathcal{D}'} &= \lambda L_{\mathcal{D}} + \lambda_{MR} L_{MR} + (1-\lambda) L_{CE} \\ L_{\mathcal{G}'} &= L_{\mathcal{G}} + \lambda_{FM} L_{FM} \end{aligned} \tag{14}$$

where L_{CE} is the standard cross-entropy loss, λ_{MR} is a hyperparameter controlling the weight of the regularization loss (L_{MR}), λ_{FM} represents the hyperparameter that controls the weight of the feature matching loss (L_{FM}), λ controls the weight between the unsupervised and the label loss, $L_{\mathcal{D}'}$ represents the total discriminator loss and $L_{\mathcal{G}'}$ refers to total generator loss. The combination of softplus loss, manifold regularization, and feature matching improves both training stability and generalization, leading to more effective performance across different subjects (cf., Section VIII-A).

Figure 6 shows the schematic representation of our methodology, including data preparation, mmGAN training, and inference. Data is divided into train, test, and validation sets. Afterwards, the datasets are normalized. Finally, labeled and unlabeled dataloaders are created from the training set. At inference, only the mmGAN discriminator is relevant for testing.

VII. EXPERIMENTAL RESULTS

This section details the measurement setup, neural network architectures, and supervised and SSL configurations. We compare the performance of mmGAN against Auto-Fi [28], a state-of-the-art self-supervised method for Wi-Fi-based gesture recognition. Finally, we present results related to transfer learning and an ablation study.

A. MEASUREMENT SETUP

We set up our ISAC system in an office environment. The EVKs are positioned on a cabinet in a corridor, where the subject performs various gestures at the red marked line as shown in Figure 7. Depending on the gestures the subject makes, we see unique patterns in the power grid (cf., Figure 1). Figure 1 illustrates the gestures we consider, motivated by the prior works [16], [55]. We include dynamic gestures (gestures involving continuous motion over the

TABLE 1. Description of the 8 gestures used in the experiments.

ID	Gesture Name
AU	Arms Up
AW	Arms Wide (Arms Open)
E	Empty (Standing Still)
PP	Push Pull
LHU	Left Hand Up
LL	Left Lean
RHU	Right Hand Up
RL	Right Lean

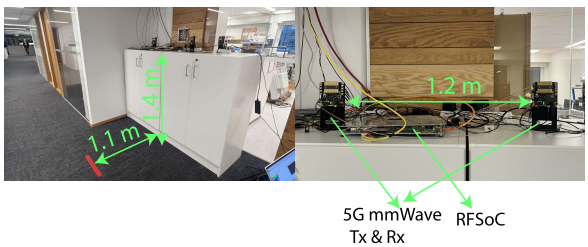


FIGURE 7. Experimental settings and environment.

10-second window) such as push-pull, arms up, left hand up, right hand up, and arms wide, along with static gestures such as right lean, left lean, and empty or standing still (cf., Table 1). Each user performs a single gesture repeatedly for a 10-second interval. Due to memory limitations in the RFSoC, extending this interval beyond was not feasible. We collect gestures from 8 distinct users spanning diverse age groups (29–50 years), height (1.73 to 1.88 meters), and body morphologies to capture realistic variability in human motion. Each participant performs gestures at a natural pace for 10 seconds, introducing natural diversity in execution speed and style. Seven such repetitions are conducted for data diversity and robustness. This ensures that the network is exposed to diverse data distributions, as each repetition exhibits distinct characteristics such as execution speed and style. The variations in body shape and gesture dynamics across different subjects influence PPBP and ultimately the downstream classifier. By incorporating this diversity, our dataset better reflects real-world heterogeneity. Moreover, the gestures considered remain representative of a broad range of Human Computer Interaction (HCI) applications.

B. EXPERIMENTS: SEMI-SUPERVISED LEARNING

1) CONSIDERED DL APPROACHES: MODEL ARCHITECTURES

We evaluate a custom CNN baseline, standard ResNet-18, and a version of the custom CNN that includes BN layers. For the custom baseline, we tune the number of layers, filter size for each layer, and other hyperparameters such as learning rate and number of epochs. After tuning, our custom baseline consists of three convolutional layers, each followed by ReLU activations. Finally, dropout with probability 0.4 and a linear layer is employed to output a score for 8 gestures (cf., Figure 5). In the custom baseline with BN, each convolutional layer is followed by a BN layer to improve performance according to standard practices [56]. The motivation to choose ResNet18 is due to the complex architecture and skip connections that counter the problem of vanishing gradient while training a deeper model [57].

We empirically determine the optimal parameters for the Residual-G, including the number of downsampling and upsampling blocks, as well as the output features. These settings are consistently applied across experiments using various percentages of labeled data. After tuning, we found that Residual-G, with 2 downsampling, residual, and upsampling blocks and 64 features, performed best on the validation data. Moreover, we compare the performance of our mmGAN with a self-supervised method, Auto-Fi [28]. In the latter, we tune the architecture of the backbone, the hidden neurons (output of the last layer of the backbone), and the complexity of the classifier. After tuning, we find the backbone architecture the same as the baseline CNN, 512 hidden neurons, and the non-linear projection head (classifier) with two linear layers performs best on the validation data.

2) FULLY-SUPERVISED TRAINING SETTING

We first train a custom CNN model in a fully supervised setting, meaning that the model has access to 100% of the labeled data. The CNN is trained subject-specific. The input data is of shape $u \times v \times z \times w$, where $u = 8608$ denotes the number of examples, $v = 20$ denotes the time samples, and $z = 50$ and $w = 56$ represent the number of Tx and Rx beams, respectively, per subject. We split the entire dataset into 70% for training, 10% for validation, and 20% for testing. We scale the data to [–1,1]. The test data is left unseen to the classifier. The model is trained for 600 epochs using the cross-entropy loss function.

3) SEMI-SUPERVISED TRAINING AND HYPERPARAMETER SETTING

In our SSL setup, we divide the 70% training data into labeled and unlabeled subsets. We create labeled data comprising 0.6% to 26% of the original training data, corresponding to 5 and 200 labeled samples per class, with the rest of the training data treated as unlabeled. Note that the labeled data is created randomly without any label selection strategies, and the same labeled data (using a fixed seed) is used for all the models, including the mmGAN. The baseline models are trained on the labeled dataset for 600 epochs, as further training did not yield improvements in validation accuracy. We use the Adam optimizer [58] with a learning rate of 3×10^{-4} for the three baselines. We monitor the validation accuracy and save the model with the highest validation performance. The best-performing model is then evaluated on unseen test data. Similarly, we train mmGAN for 600 epochs using the same labeled data as the baselines, along with the additional unlabeled data. The learning rate

TABLE 2. Key hyperparameters, search ranges, and selected values for each model.

Model	Hyperparameter	Search Range	Selected Value
CNN	#Conv layers	2–5	3
CNN	Filter sizes	16–256	[32, 64, 128]
mmGAN-G	#Down/Up blocks	1–3	2
mmGAN-G	Base features	32–128	64
mmGAN-G/D	Learning rate	10^{-3}–10^{-6}	2×10^{-4}
mmGAN	λ	{0.1, 0.3, 0.5, 0.9}	0.5
mmGAN	λ_{FM}	10^{-2}–10^{-6}	10^{-4}
mmGAN	λ_{MR}	10^{-2}–10^{-6}	10^{-3}
Auto-Fi	Learning rate (Pre-training)	10^{-4}–10^{-6}	10^{-6}
Auto-Fi	Learning rate (Fine-tuning)	10^{-4}–10^{-6}	10^{-5}
Auto-Fi	Hidden states	128–512	512

is set to 2×10^{-4} for both the generator and discriminator, with the hyperparameters λ set to 0.5, λ_{MR} to 1×10^{-3}, and λ_{FM} to 1×10^{-4}. We train our GAN subject-specific, as is the case with the baseline models. For Auto-Fi, we tune the learning rates for each training phase, setting it to 1×10^{-6} for self-supervised pre-training and 1×10^{-5} for supervised fine-tuning. The hidden states were set to 512. We also tune the number of epochs. The self-supervised stage is trained over 400 epochs, and the supervised fine-tuning stage runs for 600 epochs. All hyperparameters and the number of epochs were selected based on the best validation accuracy.

Hyperparameter-tuning strategy: Table 2 summarizes the key hyperparameters for the baseline CNN and mmGAN models, with the search ranges and the selected values. We used a random search over a predefined hyperparameter grid and selected the configuration that achieved the highest validation accuracy across diverse subjects. For the baseline custom CNN, we varied the number of convolutional layers between 2 and 5 and filter sizes for each layer, ranging from 16 to 256. For the mmGAN generator (mmGAN-G), we explored 1 to 3 upsampling and downsampling blocks, and the number of features in the first convolutional layer (base features) between 32 and 128. The learning rates for the generator and discriminator (mmGAN-G/D) were explored within the range of 10^{-3} to 10^{-6}. Additionally, the weighting parameters were tuned over ranges, $\lambda \in \{0.1, 0.3, 0.5, 0.9\}$, while λ_{FM} and λ_{MR} were varied between 10^{-2} and 10^{-6}. Similarly, we tune the learning rate for the self-supervised pre-training and supervised fine-tuning stage between 10^{-4} to 10^{-6}. The hidden states, encoding the representations learned during pre-training, were set between 128 and 512.

4) RESULTS

Figure 8 shows the performance of three baselines, a fully supervised model, Auto-Fi, and the proposed method, mmGAN. The x-axis shows the percentage of the labeled data, and the y-axis represents the model's accuracy. S1-S8 represents the 8 subjects under evaluation. A dashed horizontal red line indicates the performance of the fully supervised method. We can see that the fully supervised model achieves >96% accuracy for all 8 subjects, demonstrating the potential of 5G mmWave ISAC for gesture recognition. Additionally, we apply background subtraction (clutter removal) [59] as a pre-processing step on the data before inputting it into the CNN. The background data is collected in an empty environment and subtracted from the input data. However, we do not notice any performance improvement. While background subtraction has shown significant performance increases for signal processing-based methods, we believe that deep learning-based methods can extract relevant features without the need for this additional calibration.

Figure 8 also shows the baseline models' performance for subjects S1 to S8 in SSL settings. Generally, ResNet18 underperforms compared to both versions of the Custom CNN. Across different percentages of the labeled data, the Custom CNN without BN achieves the highest accuracy, while the version with BN exhibits intermediate performance. However, as the amount of labeled data reaches 26%, ResNet18, and the Custom CNN with BN outperform the Custom CNN without BN for most subjects. We expect this trend to persist as more labeled data becomes available, leading to reduced overfitting.

The weaker performance of ResNet18 in low-data settings can be attributed to its complexity, which leads to overfitting on limited data, reducing its generalization to test data. Moreover, our results suggest that BN layers should be avoided in low-label settings, as seen in the performance drop of the Custom CNN with BN. The latter demonstrates lower accuracy with sparse labeled data compared to a CNN without BN. This is likely due to BN memorizing training statistics rather than generalizing when training data is limited and not fully representative of the test set. However, as the amount of labeled data increases, the performance gap among baseline models narrows significantly, and both ResNet18 and the Custom CNN with BN show improved accuracy, with ResNet18 ultimately surpassing the other baselines for most subjects at the highest percentage of the labeled data (26%). To ensure a rigorous evaluation of our SSL method, we compare its performance against the best-performing baseline. This methodology allows us to benchmark our approach accurately.

The red solid curve in Figure 8 shows the performance of mmGAN. mmGAN significantly improves the performance of the baseline CNN. For example, for subject S4, mmGAN achieves 95% accuracy, only 4.5 percentage points (pp) less than the fully supervised model trained using 100% labeled data. We achieve the same performance on S4 as the custom baseline CNN while using 7 times fewer training data (at 0.6% labeled data). For S3 also, we see nearly 25 pp jump in accuracy compared to the best baseline at 0.6% of labeled data. We notice that the smaller the amount of the labeled data, the more the improvement. For all percentages of labeled data shown in Figure 8, the GAN-based method outperforms all the baselines, except for S5 at 200, the ResNet-based baseline leads by 0.6

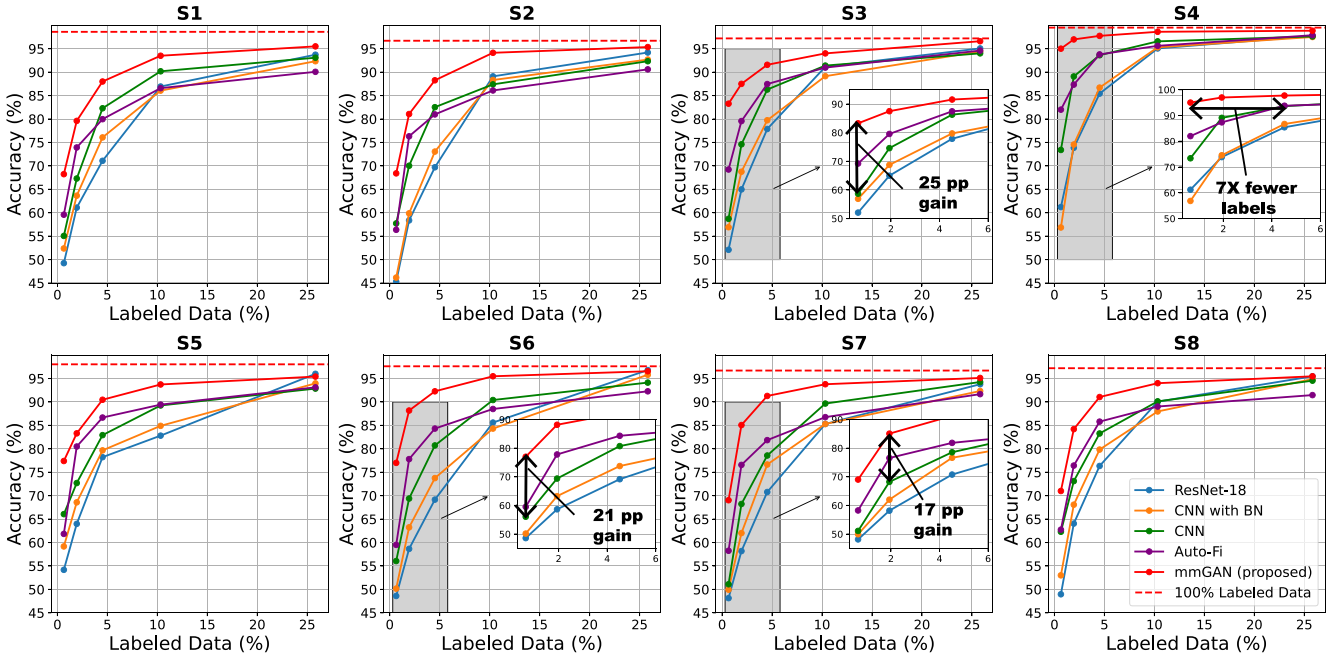


FIGURE 8. Comparison of accuracy for mmGAN, Auto-Fi, and 3 Baseline models: The blue, orange, and green curves represent three baselines: ResNet18, CNN with Batchnorm, and CNN without Batchnorm. The purple and red solid lines indicate Auto-Fi and mmGAN performance, respectively. A dashed red line represents the performance of a CNN under 100% training data.

pp, and for S8, we notice the same performance for the ResNet-based baseline and mmGAN. As expected, the gap between the best-performing baseline and mmGAN narrows down as the amount of labeled data increases. However, depending on the subject, the performance improvements vary due to the domain gap between the subjects that arises from the body characteristics and the way gestures are performed.

Additionally, the discriminator of mmGAN is a lightweight CNN occupying only 0.57 MB of memory with an average inference latency of 0.6 milliseconds. This also enhances energy efficiency, making it suitable for deployment in resource-constrained devices. Overall, mmGAN can substantially improve the performance of ISAC systems for gesture recognition. Variations in performance across subjects stem from the distribution gap, reflecting the varying complexity of the task for different individuals.

mmGAN consistently outperforms Auto-Fi. For instance, for subject S6, the improvements are 18 pp and 11 pp at 0.6% and 2% labeled data, respectively. While Auto-Fi outperforms the baseline CNN's performance at low-label percentages, its effectiveness decreases as the amount of labeled data increases, ultimately falling short of the top-performing baseline. We believe that Auto-Fi and other self-supervised learning methods require strong augmentations and a significant amount of unlabeled data [39], [60]. These augmentations are often carefully designed for specific data types and tasks. Currently, Gaussian noise is the primary augmentation applied to wireless signals, highlighting a need for further research into effective, domain-specific augmentations.

C. TRANSFER LEARNING

In this scenario, we assume access to a small fraction of labeled target data and aim to assess whether our mmGAN can generalize more effectively across different subjects. Specifically, we fine-tune two neural networks: baseline CNN and the mmGAN. For each subject, we train the two networks with 0.6% of the labeled data. Then we use the trained baseline CNN and mmGAN as a starting point and train it on every other subject (cross-domain) using 4% of the target labeled data. We fine-tune the whole network on the target labeled data for 400 epochs and then select the best model according to the validation accuracy. Figure 9 shows the results of fine-tuning the baseline CNN (best-performing) and mmGAN on cross-domain target data. The X-axis shows the subject under evaluation, and the Y-axis represents accuracy. Each boxplot then shows the distribution of accuracy for a given subject when using other subjects' data to train the two models. The fine-tuned mmGAN can generalize better. For all subjects, the median accuracy of fine-tuned mmGAN is always higher than the fine-tuned baseline. Moreover, for most subjects, the minimum accuracy of fine-tuned mmGAN is greater than the maximum of the fine-tuned baseline. Thus, mmGAN serves as a strong foundation for cross-subject transfer learning, improving the model's ability to generalize and perform well across different subjects.

VIII. ADVANCED STUDIES

This section details extended experiments such as ablation studies, performance evaluations across diverse data distributions, analysis on the Widar dataset, and performance metrics, including insights from interpretability analysis.

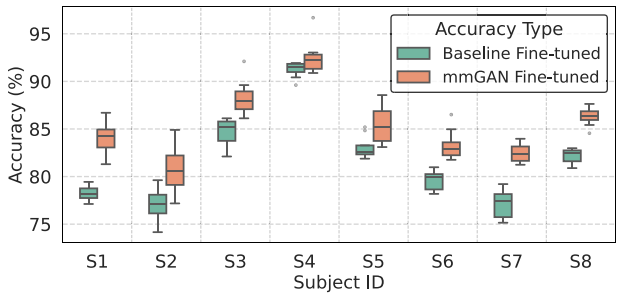


FIGURE 9. Fine-tuning on 4% of target labeled data.

TABLE 3. Comparison of accuracy (Acc) and FCD at 0.6% of the labeled data: DCGAN generator (D) vs proposed residual generator (R).

Subj.	Acc (D)	Acc (R)	FCD (D)	FCD (R)
S1	57.14%	68.23%	335	4.5
S2	54.35%	68.40%	310	45
S3	60.56%	83.27%	58.8	25.1
S4	77.93%	95.01%	357	47
S5	67.47%	77.40%	267	16.7
S6	67.04%	77.0%	131	26.33
S7	50.69%	69.03%	8.1	2.8
S8	58.76%	71.00%	113	13.2

A. ABLATION STUDY

1) ARCHITECTURE OF THE GENERATOR

Since the discriminator in the GAN shares the same architecture as the baseline Custom CNN, the role of the generator becomes critical in refining the decision boundaries. By generating diverse and realistic synthetic data, the generator challenges the discriminator, ultimately helping it to better distinguish between real and generated samples, which leads to improved classification performance. In this section, we compare the performance of Residual-G with a popular architecture choice for the generator based on DCGAN [61], [32]. We tailor the DCGAN generator to our dataset. The adapted DCGAN generator consists of a noise mapping layer, followed by a ReLU activation layer. The architecture further consists of 3 convolution transpose layers, two of them followed by BN and ReLU. The final transpose layer is followed by Tanh activation. We compare the performance of the two generators at the smallest label setting (0.6%). Table 3 shows that the proposed Residual-G (R) outperforms the DCGAN-styled generator (D) in terms of accuracy (Acc) for different subjects (Subj). This improvement is due to the more complex architecture of Residual-G and the use of residual connections, which address vanishing gradient issues, allowing for better gradient flow and feature reuse. In contrast, the simpler DCGAN generator struggles to generate realistic data.

Evaluating GANs can be particularly challenging [62], especially with wireless signals, as there is no straightforward method for visualization like with images. To compare the performance of the two generators, we calculate the Fréchet Inception Distance (FID) [62], [63] as a quantitative measure of their output quality. FID is a metric used to evaluate the realism and diversity of data generated by GANs. It measures the similarity between the distributions of real and generated data by comparing the statistics (mean and covariance) of feature representations extracted from a pre-trained neural network. A lower FID score indicates that the generated data is closer to the real data in terms of quality and variety. FID is computed as:

$$FID = \|\mu_R - \mu_G\|^2 + tr\left(\Sigma_R + \Sigma_G - 2(\Sigma_R\Sigma_G)^{\frac{1}{2}}\right) \quad (15)$$

where R and G are real and generated embeddings obtained from a pre-trained model, μ_G and μ_R are the means of the vectors R and G, respectively, and Σ_R and Σ_G represent the covariances of these vectors. For images, the embeddings are typically obtained using a pre-trained Inception network (trained on image data). In contrast, in our case, the embeddings R and G are extracted from the baseline Custom CNN model (pre-trained at 0.6% of the labeled data) by performing forward passes on the real and generated data, respectively. So, in our case, we refer to FID as FCD (Fréchet Classifier Distance). FCD is calculated for both generators across all subjects. For reference, we also compute the FCD between the unlabeled data and the test data as a sanity check, where we anticipate the FCD to be close to 0. Additionally, we assess the FCD between the unlabeled data and random data, expecting this value to be very high. For example, for S4, the FCD between the unlabeled and test data is 0.34, while the score between the random and unlabeled data is 415. Similar trends hold for other subjects.

Table 3 also shows the FCD comparison between the DCGAN generator (D) and Residual-G (R) at 0.6% of the labeled data. FCD was calculated using 10000 generated samples in all the cases. Residual-G achieves a significantly lower FCD score compared to the DCGAN generator across all subjects, demonstrating its effectiveness in preserving both the realism and diversity of the generated data. It's important to note that an FCD score close to 0 is not ideal for GANs, as it would indicate that the model is simply memorizing or copying the data, rather than learning meaningful representations. This metric provides a deeper insight into why Residual-G surpasses the DCGAN generator in performance. Training GANs poses significant challenges and often leads to instability, stemming from the adversarial dynamics between the two competing neural networks. This instability has limited the widespread adoption of GANs in ISAC research. Choosing an optimal generator architecture is essential for stable training and improving the accuracy of mmWave ISAC gesture recognition, particularly in settings with limited labeled data.

2) LOSS FUNCTION

We compare the performance of the loss function proposed in Section VI against the widely used BCE and Least Squares loss (LS-GAN) [27], [64]. The works [34], [35] primarily

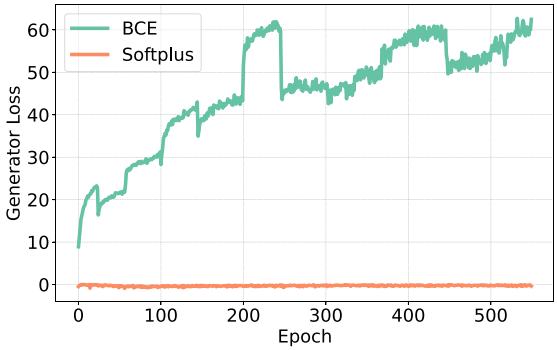


FIGURE 10. Generator Loss: Softplus vs BCE. BCE leads to exploding loss and unstable training.

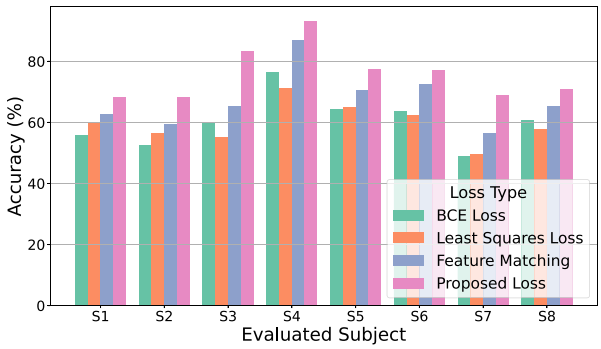


FIGURE 11. Accuracy comparison of mmGAN across different loss functions at 0.6% of labeled data.

address cross-domain settings, making a direct comparison with mmGAN infeasible. However, both approaches formulate adversarial loss using binary cross-entropy (BCE).

We utilize the same Residual-G and the same discriminator, with the only modification being the addition of a linear layer followed by a sigmoid activation, as the BCE loss function requires the input to be a probability. Consequently, both the generator and discriminator losses are formulated using BCE loss. Figure 10 shows the generator loss using BCE loss and the proposed softplus loss for one of the subjects at the smallest label setting (0.6%). The softplus generator loss is bounded. In contrast, the generator loss using BCE keeps exploding. This divergence in the loss behavior implies that the generator is struggling to produce samples that sufficiently challenge the discriminator. This is because the cost function that approximates the min-max game using BCE loss has plateaus (vanishing gradient) when the two distributions (real and generated) do not overlap. As a result, the generator does not get feedback to improve. On the other hand, using the softplus loss allows the gradients to remain informative even when the two distributions are far apart. Some studies [25], [65] suggest employing the feature matching loss in the generator to stabilize GAN training and enhance convergence speed. Based on this, we disabled the softplus loss for the generator and applied only feature matching, while keeping the discriminator loss unchanged. Figure 11 shows the accuracy of mmGAN when employing different loss functions: LS, BCE, feature matching, and the proposed loss (with 0.6% labeled data). For feature matching, a notable improvement can be seen compared to LS and BCE. However, in general, the three losses fall behind the proposed loss in terms of accuracy. The proposed loss achieves an improvement ranging from a minimum of 6 pp to a maximum of 18 pp compared to the feature matching loss. In conclusion, the proposed loss function provides the best performance.

B. PERFORMANCE UNDER DIVERSE DATA DISTRIBUTIONS

Our mmGAN is designed to be person-specific since wireless sensing systems can be fine-tuned or calibrated to the specific subject. In most practical applications, e.g., extended reality (XR) and personalized HCI systems, ISAC models should generally have a short calibration phase to be adapted to the target user to maximize performance. This process may involve collecting a small amount of labeled data from the target user, together with continuously acquired unlabeled data, to enhance generalization. Moreover, users may naturally vary their gesture execution over time, for instance, in speed or motion dynamics. Periodically incorporating new labeled and unlabeled samples allows the model to remain adaptive and robust. Consequently, mmGAN is well-suited for personalized ISAC applications, where user-specific fine-tuning is both feasible and beneficial. Note that state-of-the-art pose estimation or gesture recognition systems, such as those based on depth cameras with markers, generally also have a calibration phase.

In this section, we examine the impact of diverse data distributions on the performance of mmGAN. We conduct two experiments: in the first, data from three users is combined, and in the second, data from all users is aggregated. It is important to emphasize that in both scenarios, the validation and test sets are drawn from the same set of users as the training data. Our results highlight that mmGAN's performance tends to slightly decline as the diversity of the training data increases. Additionally, we evaluate the model's performance on unseen users, further demonstrating the challenges posed by data diversity.

1) 5-SHOT LEARNING WITH 3 SUBJECTS

We combined data from subjects S3, S4, and S6, utilizing 5 labeled examples per subject, resulting in a total of 120 ($5 \times 3 \times 8$) labeled examples, with 8 representing the number of classes. Both a baseline model and mmGAN were trained following the steps outlined earlier. On the test set consisting of the three subjects, the baseline model achieved the best accuracy of 68%, while mmGAN reached 82.18%, reflecting an improvement of 14 pp (cf., Figure 12). In theory, we expect improvement closer to 23 pp, the average performance of the mmGAN on the 3 subjects (cf., Figure 8). However, the diversity in data distribution poses

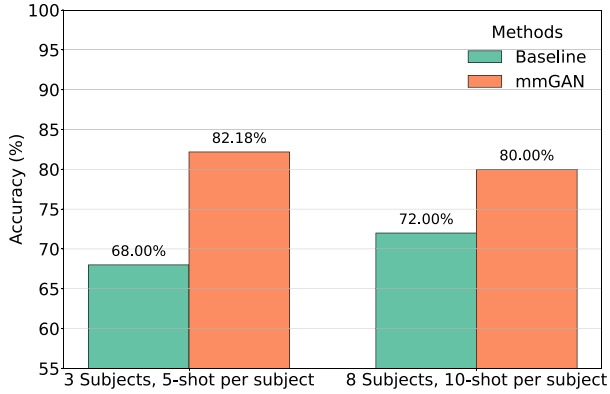


FIGURE 12. Comparison of the best Baseline Model and mmGAN on combined data.

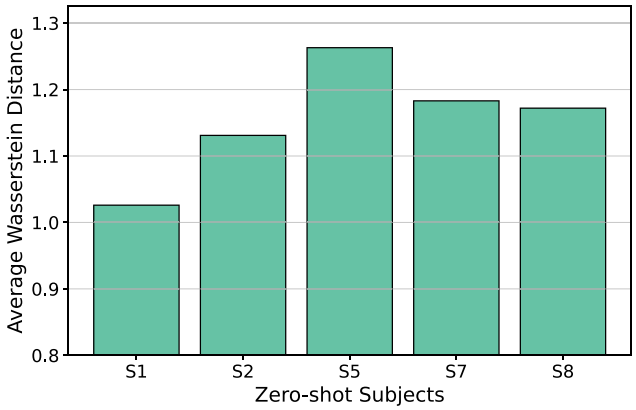


FIGURE 13. Average Wasserstein Distance on zero-shot subjects.

TABLE 4. Comparison of accuracy of the baseline model and mmGAN trained on S3, S4, and S6 and tested on the test set of unseen subjects: zero-shot settings.

Subj.	Accuracy (Baseline)	Accuracy (mmGAN)
S1	47%	60.2%
S2	45%	63%
S5	42%	48%
S7	48%	60.1%
S8	47%	60.05%

challenges for achieving greater performance improvements, which is also evident in the baseline model's reduced performance.

2) ZERO-SHOT CROSS-DOMAIN SETTINGS

We expect better performance on unseen subjects compared to the baseline, as training on unlabeled data should allow mmGAN to learn more generalized features. To evaluate this, we test its performance on unseen subjects in a zero-shot setting, where no data from the target domain is available, neither labeled nor unlabeled. Table 4 shows the performance of mmGAN tested on left-out subjects. It is evident that mmGAN consistently outperforms the baseline on unseen subjects. Except for S5, mmGAN achieves an improvement of over 12 pp on all other subjects.

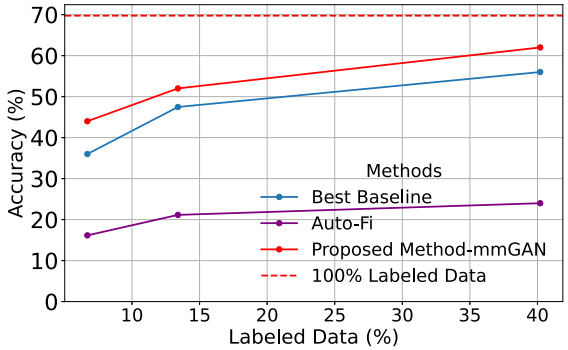


FIGURE 14. Performance of mmGAN on the Widar dataset.

The variability in performance arises due to significant differences in data distribution between training and unseen subjects. To quantify the distribution gap, we compute the Wasserstein Distance [66], which is a measure of similarity between two distributions. We compute the distance class-wise in the PPBP domain between the training and zero-shot subjects. Figure 13 shows the average Wasserstein distance for zero-shot subjects, averaged over 8 classes. We see that for subject S5, the distance is larger compared to other subjects, 19% larger than S1 and 7% larger than S8. In contrast, subjects with lower distances obtain higher accuracy. For example, S7 and S8 have similar distances and similar accuracies. While Wasserstein distance provides a quantitative measure, other factors, such as model biases, also play a role in zero-shot performance.

3) COMBINING ALL DATA, 8 SUBJECTS AND 10-SHOT PER SUBJECT

Finally, we aggregate data from all subjects and evaluate the model on the test set for the same subjects. Specifically, we select 10 labeled examples per class per subject, treating the remaining data as unlabeled. Given the increased dataset size in this setup, we apply dimensionality reduction techniques to decrease the computational complexity of GAN training, particularly for the generator. We evaluated several dimensionality reduction techniques, including linear methods like Principal Component Analysis (PCA) and nonlinear approaches such as kernel PCA [67] and Uniform Manifold Approximation and Projection (UMAP) [68]. However, these methods struggled to maintain the semantic structure of the data, likely due to inherent assumptions, such as linearity, that do not align with the underlying data distribution. Additionally, the computational complexity of UMAP makes it impractical for real-time applications. As such, we instead compute the temporal average of the PPBP (across the time index) to reduce the complexity of the data and restrict the dataset to 50 receiver beams. This preprocessing reduces the dataset dimensions to $u \times 1 \times v \times w$, where u is the number of samples, and $v = w$ represents the spatial dimensions. This significantly speeds up GAN training. We train the baseline models and mmGAN on the combined dataset. The best baseline model achieves an accuracy of 72%, whereas

mmGAN achieves an accuracy of 80% on the test set (cf., Figure 12). This shows that even in diverse data distributions, mmGAN can fairly improve the accuracy of ISAC systems for gesture recognition.

To summarize, mmGAN improves performance under diverse data distributions both in in-domain settings and cross-domain settings. We believe that data processing and dimensionality reduction can be key to improving performance under diverse data distributions, and we leave it as a part of future work.

C. EVALUATION ON WIDAR DATASET

We further evaluate mmGAN on the significantly larger multi-person, multi-environment publicly available Widar dataset [31], [69], which comprises 43,652 data points spanning 22 distinct gesture categories. To facilitate SSL, we partition the dataset into labeled and unlabeled subsets. Specifically, we allocate 100, 200, and 600 labeled samples per class, corresponding to 6.7%, 13.4%, and 40.2%, respectively, of the training set, which consists of 32,371 samples. Figure 14 shows the gesture classification accuracy of the best baseline model, AutoFi, and mmGAN on the Widar dataset for different percentages of labeled data. We also train the baseline model with 100% labeled data. From Figure 14, mmGAN significantly improves the classifier's performance for gesture recognition compared to the baseline model, while Auto-Fi struggles with the performance augmentation. This is because Auto-Fi's strengths do not align when data comes from different domains (environments). In contrast, mmGAN achieves an accuracy of 62.2% using only 40.2% of labeled data, falling just 7 pp below the performance of the baseline model trained with a fully labeled dataset (100% labeled data 69.25% [69]).

D. INSIDE mmGAN: PERFORMANCE ANALYSIS

1) STATISTICAL SIGNIFICANCE: P-VALUE

We perform a statistical significance test to reveal whether the performance difference between the models, baseline CNN and mmGAN, is due to genuine improvements or random chance. For brevity, we report results on the S4 data. However, similar results hold for other subjects. Table 5 shows the accuracy for the baseline model and mmGAN at different percentages of labeled data. μ denotes the mean and σ denotes the standard deviation over five runs. Δ indicates the difference between the accuracies. The p-value [70] quantifies the probability that the observed differences occurred purely by chance. The p-values are compared against a predefined significance threshold, commonly accepted value of 0.05 or a more stringent 0.01. From Table 5, we see that mmGAN outperforms the baseline model, particularly at the lowest label settings (0.6%). At all percentages of labeled data, the p-values are significantly smaller than the widely used 0.01 (1%) significance threshold. Therefore, the performance improvements are considered statistically significant.

TABLE 5. Baseline vs. mmGAN test accuracy on S4 (mean $\mu\pm$ standard deviation over five runs).

#Labels (%)	Baseline ($\mu \pm \sigma$)	mmGAN ($\mu \pm \sigma$)	Δ (pp)	p-value
0.6	71.70 ± 1.37	94.81 ± 0.24	23.11	0.0017
1.9	87.27 ± 1.65	96.02 ± 0.94	8.76	0.0045
4.5	92.53 ± 0.79	96.59 ± 0.76	4.06	0.0055
10.3	95.29 ± 0.87	97.98 ± 0.37	2.69	0.0048
25.8	97.38 ± 0.23	98.63 ± 0.12	1.25	0.0003

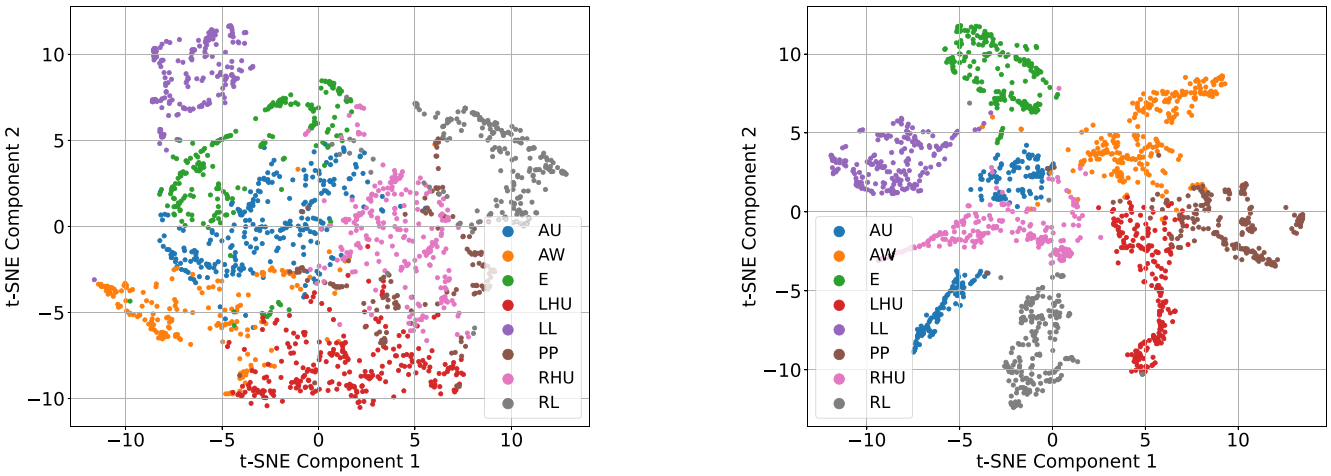
TABLE 6. Comparison of baseline vs. mmGAN across 8 gestures for S4.

Gesture	Baseline			mmGAN		
	P	R	F1	P	R	F1
AU	0.663	0.840	0.741	0.961	0.961	0.961
AW	0.715	0.619	0.664	0.843	0.960	0.897
E	0.808	0.861	0.833	0.995	0.967	0.981
LHU	0.569	0.712	0.632	0.928	0.980	0.953
LL	1.000	0.946	0.972	0.995	0.980	0.988
PP	0.738	0.452	0.561	0.975	0.848	0.907
RHU	0.562	0.576	0.569	0.945	0.974	0.959
RL	0.892	0.912	0.902	0.986	0.940	0.962
Accuracy		0.734			0.950	

2) T-SNE EMBEDDINGS AND CONFUSION MATRIX

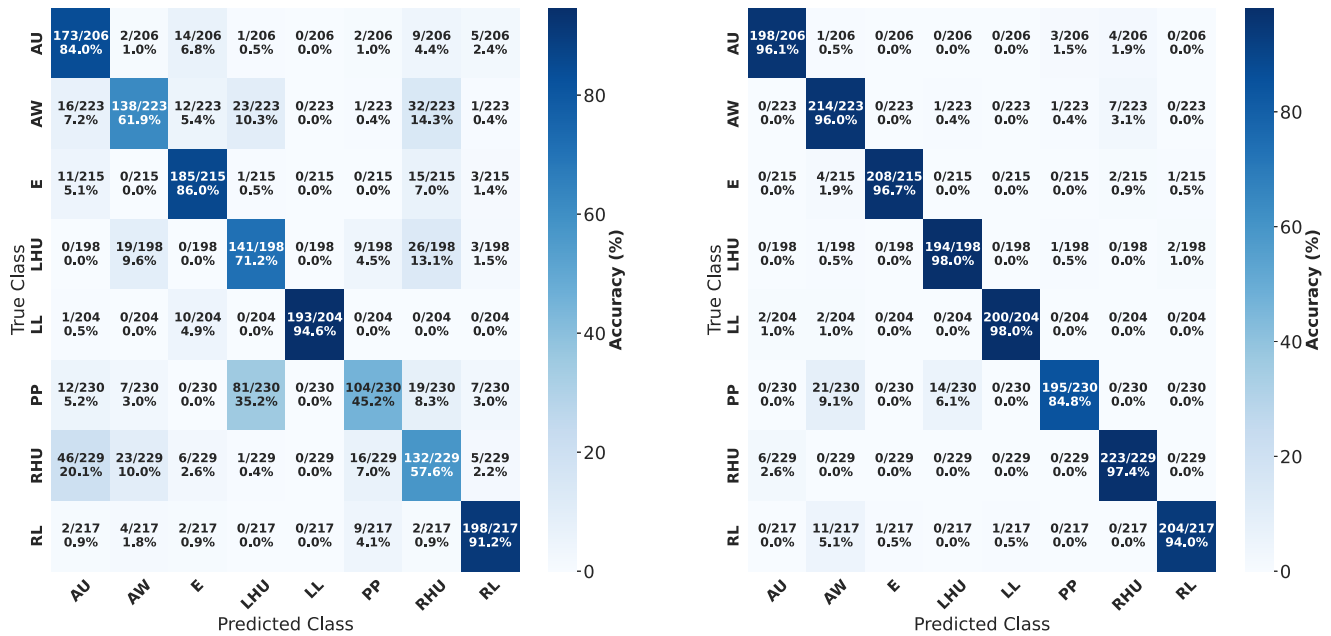
In this section, we analyze in detail the performance metrics of mmGAN against the baseline model. We report additional metrics beyond accuracy and discuss how mmGAN augments the accuracy. For brevity, we report results corresponding to Subject S4 at the lowest label setting (0.6%). We investigate metrics such as precision, recall, F1-scores, and t-SNE representations of the final layer of the mmGAN. We thoroughly compare it with the best baseline classifier.

Table 6 shows the evaluation metrics (precision P, recall R, and F1) of the baseline model against the mmGAN-aided classifier (discriminator). We see that for the 8 gestures, even though the baseline model obtains an overall accuracy of 73.4%, the class-wise F1-scores vary significantly. We see that static gestures such as E, LL, and RL achieve high accuracy scores as far as the baseline model is concerned. In contrast, dynamic gestures such as PP, LHU, RHU, AU, and AW obtain lower F1-scores, indicating that the baseline model requires more data to learn their complex temporal patterns effectively. This is because dynamic gestures exhibit a higher degree of variability in execution speed and orientation, resulting in multimodal distributions. On the other hand, mmGAN attains F1-scores above 0.89 across all gestures. This suggests leveraging unlabeled data helps the model to learn more robust and temporally consistent feature representations. Thus, mmGAN enhances feature representation and class separability. This is further illustrated with t-SNE embeddings (cf., Figure 15) and confusion matrices (cf., Figure 16) of baseline and mmGAN.



(a) Baseline Model (b) mmGAN

FIGURE 15. Comparison of t-SNE embeddings for Baseline and mmGAN on S4 test data.



(a) Baseline Model (b) mmGAN

FIGURE 16. Comparison of confusion matrices for Baseline and mmGAN on S4 test data.

Figure 15a and Figure 15b show the feature representations of the final layer for the baseline model and mmGAN, respectively. We can clearly see that the features are well separated for mmGAN, and classes can be easily distinguishable, indicating that the generator effectively aids the discriminator in learning more discriminative and refined decision boundaries. On the other hand, for the baseline model, the classes are not well separable due to limited labeled data and the inability of the model to separate clusters effectively.

Similarly, Figure 16a and Figure 16b show the confusion matrices for the baseline and mmGAN, respectively, including the support (number of examples per gesture) on unseen data. mmGAN substantially reduces misclassifications across most classes. These results highlight the benefit of mmGAN in improving feature discrimination and overall classification performance.

3) INTERPRETATION THROUGH BEAM PAIR IMPORTANCE: XAI

To further gain insights into the model's performance, we analyze the internal feature representation of the model using an XAI approach. In particular, we examine the importance of Tx-Rx beam pairs and visualize the corresponding global saliency heatmaps [71] for both models on the S4 test dataset.

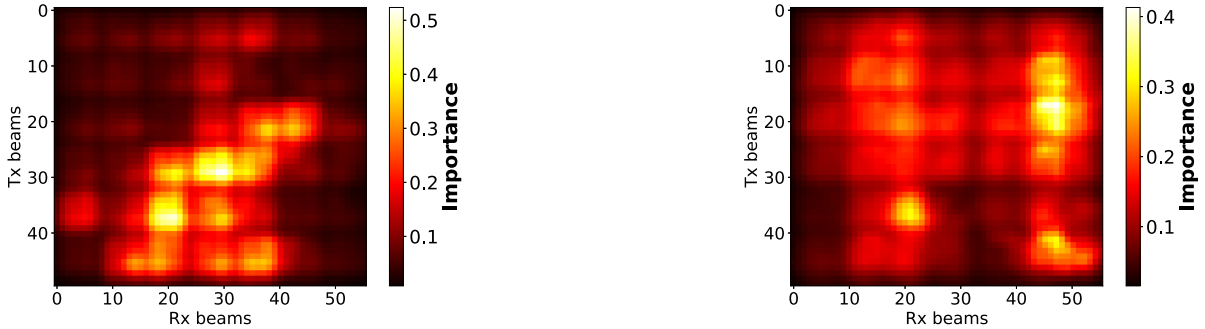


FIGURE 17. Comparison of global saliency heatmaps for the Baseline model and mmGAN on S4 test data.

The heatmap represents how selectively the model responds to particular spatial or angular regions.

Figure 17a and Figure 17b show the beam pair global (averaged over all classes) saliency maps for the baseline and mmGAN model, respectively, highlighting the regions each model relies on most. Global saliency maps are obtained by aggregating information across all gestures. These maps illustrate where the baseline and mmGAN models focus their attention, revealing the critical Tx-Rx beam pairs for their classification decisions. We can see that the baseline map is more concentrated on a few hotspots, which likely represent the line-of-sight and a few dominant non-line-of-sight components. On the other hand, mmGAN has a more distributed heatmap reflecting richer spatial information from a wide range of beam pairs. These contributions more likely come from weaker multi-path components affected by the gestures. mmGAN, therefore, has a more distributed understanding of the classification task. Thus, the generator training encourages the mmGAN model to extract relevant features from weaker paths in addition to the primary path, resulting in improved generalization.

Further, we perform a global beam pair importance ablation for both models. After computing the global saliency map, which highlights the most critical beam pairs for classification, we selectively perturb the top-$K\%$ beam pairs (with $K = 1, 2, \ldots, 50$) and record the classification accuracy. The perturbation is done by replacing such beam pairs with random uniform noise, effectively destroying the information they carry. Note that there are 2800 beam pairs (50×56), and we perturb the top 1400 (50%).

Figure 18 shows the ablation study on the beams for the two models. For the baseline model, perturbation of the top-1% beams causes a drop in accuracy to 42% from 73.4%. As the ablated beams reach 10%, the baseline accuracy further drops to 24%. At 50% ablated beams, the accuracy is around 12%. On the other hand, mmGAN sees a sharp decline in accuracy at the top-2% of the ablated beams; the accuracy drops from 95% to 18% (42% at top-1%), suggesting the fact that it relies heavily on critical beams. While mmGAN is more sensitive to perturbations, it is more interpretable. The model's decisions are highly focused on critical beams,

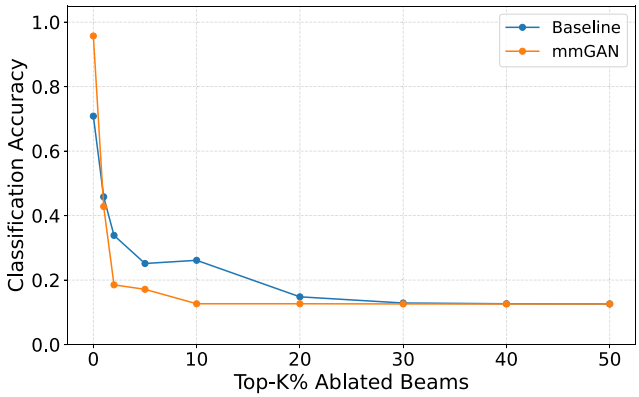


FIGURE 18. Ablation study on global beam pair importance.

allowing for easier post-hoc identification of critical beams for each gesture and providing a more transparent view of learned representations than the baseline counterpart. In general, both models experience a sudden drop in accuracy as the top-$K\%$ beams are perturbed, suggesting that the top-$K\%$ beams are correctly identified. This can result in efficient pruning techniques [72], resulting in low computational cost and latency for real-time applications.

E. TRAINING COMPLEXITY

We acknowledge that the performance improvements achieved with GANs come at the cost of increased training complexity. This is because of the adversarial training that involves two neural networks competing with each other. Further, usually the generator is more complex than the discriminator. This is because the generator's role is to generate high-quality synthetic data capable of challenging the discriminator. As a result, GAN training has a high computational cost, requires careful hyperparameter tuning and careful monitoring to ensure convergence, in comparison to standard supervised training. Moreover, due to the generator's complex architecture, the inference cost is also high in terms of memory footprint and latency, compared to simpler supervised models.

We train mmGAN using one NVIDIA A100 80GB GPU with the PyTorch deep learning framework and CUDA 12.2 on Linux 5.15, and 48 CPU cores (96 threads) and Python 3.11.8. The training process took 4 hours and 49 minutes to complete. However, we envision the GAN training process to occur in the cloud, where sufficient compute resources are available. In contrast, the model deployed on the edge device is a lightweight discriminator with only 150K parameters, making it significantly smaller than standard architectures such as ResNet-18, which has 11.2M parameters, occupies 42 MB of memory, and achieves an average inference latency of 2.2 milliseconds. Latency was determined by performing ten forward passes through the model and reporting the average execution time. Instead, our discriminator model has a memory footprint of only 0.57 MB and an average inference latency of 0.6 milliseconds, making it approximately 74 times smaller and 3.7 times faster than ResNet-18.

IX. DISCUSSION

Our extensive investigation reveals that mmGAN can significantly benefit mmWave ISAC systems by effectively leveraging labeled, generated, and unlabeled data.

Low-label performance: First and foremost, mmGAN improved the gesture recognition performance across diverse subjects in extremely low-label settings. For example, at 0.6% of labeled data, the classification accuracy for subject S3 was enhanced by 25 pp. For subject S4, mmGAN achieved the same performance as a supervised baseline but was trained with seven times less data. Similarly, for S6, the gains were 21 pp. Our results indicated that the performance gains were most significant when the fraction of labeled data is small and the amount of unlabeled data is large, and these gains gradually decreased as the difference between labeled and unlabeled data is reduced.

Subject-specific differences: However, these gains varied across subjects due to differences in their gesture execution and body characteristics, which resulted in distinct data distributions. We quantified these distributional gaps using the Wasserstein distance to explain the observed performance variations. We found that the subject with a larger Wasserstein distance resulted in lower accuracy under zero-shot settings. For example, subject S5 had lower accuracy because the distance was 19% larger than S1 and 7% larger than S8. In future work, we plan to further quantify these person-specific gaps using multiple metrics, such as the Structural Similarity Index, examining both the PPBP and feature domains.

Diverse data distributions: Our results further showed that, even under diverse data distributions and challenging scenarios such as 5-shot and 10-shot learning, mmGAN improved baseline performance by 14 pp and 8 pp for experiments involving 3 and 8 subjects, respectively. In zero-shot settings and no access to target labeled data, mmGAN also improved the gesture recognition performance (5 to 18 pp).

Transfer learning: Further, mmGAN also served as a strong initializer for cross-subject transfer learning. Both mmGAN and the baseline were fine-tuned using 4% of the target labeled data, and mmGAN adapted more rapidly to new subjects, achieving higher median accuracy than the baseline across all subjects.

Generalization and comparison: Additionally, we demonstrated that mmGAN outperforms the state-of-the-art self-supervised method, Auto-Fi, on the 5G OFDM dataset as well as the publicly available Widar dataset. On the Widar dataset, mmGAN's performance was only 7 pp lower than the baseline trained with 100% labeled data, while mmGAN used only 40% of the labeled data.

Inside mmGAN: We analyzed the mmGAN model to reveal its internal feature representations and reported detailed metrics beyond accuracy, class-wise F1 scores, along with an interpretability study that offers deeper insights into the model's behavior and generalization. We further quantified differences in classification performance between static and dynamic gestures. We found that under low-label-settings baseline model suffers from poor performance as far as dynamic gestures are concerned, whereas mmGAN achieved a substantial accuracy improvement across all gestures. mmGAN's saliency maps showed attention on distributed beam pairs, suggesting that unlabeled data helps the model to capture secondary multipath components relevant to the gestures, which the baseline model completely misses. We also identified the most critical beam pairs for the baseline and mmGAN models and performed an ablation to validate their significance. This can be used to develop efficient pruning techniques for creating lightweight models.

Edge efficiency: Further, we emphasized that GAN training can be computationally expensive. However, in our case, only the lightweight discriminator is deployed during inference. The mmGAN discriminator occupied only 0.57 MB of memory with an inference time of 0.6 milliseconds, resulting in a model that is 74 times smaller and 3.7 times faster than ResNet-18.

X. CONCLUSION

In this work, we proposed an SSL method, mmGAN, to enhance the performance of mmWave ISAC systems for gesture recognition in label-scarce settings. We proposed a loss function that integrates the advantages of softplus, feature matching, and manifold regularization to significantly improve gesture recognition performance across diverse subjects. To evaluate mmGAN, we developed a testbed with 5G mmWave OFDM signals and extracted PPBP as a sensing feature. The time series of PPBP was provided as input to the mmGAN. We benchmarked mmGAN against rigorously evaluated baseline models to ensure a fair and comprehensive performance comparison, including the state-of-the-art self-supervised method, Auto-Fi. mmGAN was evaluated on the 5G OFDM dataset and the publicly available Widar dataset. Our results showed that by leveraging unlabeled data, mmGAN improved gesture recognition performance

in both subject-specific and diverse data scenarios, reduced dependence on labeled data, and generalized effectively across multiple subjects, served as a strong initializer for transfer learning, representing a step forward toward foundation learning in mmWave ISAC systems.

XI. LIMITATIONS AND FUTURE WORK

mmGAN is trained on a subject-specific basis. In such settings, the performance improvement is substantial. However, when data from multiple subjects is combined, there is limited improvement in performance due to diverse distributions. For example, when data from all users is combined, there is only an 8 pp improvement in gesture recognition performance. This is inherently due to the complexity of the dataset. We consider addressing this challenge in future work. Additionally, GAN training can be computationally intensive, but this challenge can be mitigated by offloading the training process to the cloud. In such a setup, only the lightweight discriminator is deployed on the edge device, enabling faster inference times compared to diffusion models.

Diffusion models: Diffusion models [73] are a class of generative models designed to produce data that closely resembles the training data. They operate by progressively adding noise to the data through a forward Markov chain, effectively transforming the data into pure noise. The model then learns the reverse process, a backward Markov chain, to denoise the noisy data, reconstructing the original data distribution. This two-step approach ensures that the model captures the underlying data distribution, enabling it to generate realistic samples from noise. Diffusion models have achieved success in many applications such as computer vision, natural language processing, and multimodal modeling [73]. However, their effectiveness comes with challenges: they require large quantities of high-quality data to train effectively, and their iterative sampling process is computationally expensive, often involving thousands of steps to generate a single output. This high computational complexity demands substantial hardware resources and can lead to slower generation times compared to other generative models such as GANs or Variational Autoencoders (VAEs). Moreover, recent studies have suggested that the generated data from diffusion models leads to degraded classifier performance [74]. We consider evaluating diffusion models in future work.

Rotation and translation artifacts: We further want to extend our analysis by collecting data where the subject's orientation is varied. Unlike images, we cannot artificially apply translations or rotations to wireless data, since such operations change the semantic meaning of the PPBP measurements. Our current results already show strong performance across diverse data distributions, and we expect that such transformations would mainly create a different distribution that the generative ability of mmGAN can also handle. We leave a detailed evaluation of this aspect for future work, as this would require collecting additional data with rotated subjects.

Domain adaptation: Further, we aim to improve performance on cross-domain settings without using target-labeled data. We plan to combine our approach with adversarial domain adaptation techniques [75], such as feature alignment or pixel alignment, such as CSIGAN [34], to improve the performance on target data without using labels (unsupervised domain adaptation). However, these approaches often face challenges due to semantic mismatches across domains, for example, the source features or examples of one gesture may be mapped to the target features or examples of another [76]. Furthermore, CSIGAN [34] focuses on learning a one-to-one mapping across domains, whereas recent research highlights that distributions are inherently multimodal [77]. More research is needed in this direction, and we leave it as part of future work.

Sensing and communication trade-off: Moreover, our dataset could be used to analyze the trade-off between gesture recognition (sensing) and communication performance, specifically examining how the number of Tx and Rx beams, both in space and time, impacts the sensing performance and communication overhead in ISAC.

REFERENCES

[1] T. Wild, V. Braun, and H. Viswanathan, "Joint design of communication and sensing for beyond 5G and 6G systems," *IEEE Access*, vol. 9, pp. 30845–30857, 2021.

[2] F. Liu et al., "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1728–1767, Jun. 2022.

[3] J. A. Zhang et al., "Enabling joint communication and radar sensing in mobile networks—A survey," *IEEE Commun. Surv. Tut.*, vol. 24, no. 1, pp. 306–345, 1st Quart., 2022.

[4] F. Wang, J. Feng, Y. Zhao, X. Zhang, S. Zhang, and J. Han, "Joint activity recognition and indoor localization with Wi-Fi fingerprints," *IEEE Access*, vol. 7, pp. 80058–80068, 2019.

[5] C. Li, M. Liu, and Z. Cao, "WiHF: Enable user identified gesture recognition with Wi-Fi," in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 586–595.

[6] Y. Ren, Z. Wang, Y. Wang, S. Tan, Y. Chen, and J. Yang, "GoPose: 3D human pose estimation using Wi-Fi," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 2, pp. 1–25, 2022.

[7] W. Jiang et al., "Towards 3D human pose construction using Wi-Fi," in *Proc. 26th Annu. Int. Conf. Mobile Comput. Netw.*, 2020, pp. 1–14.

[8] R. Yang, X. Yang, J. Wang, M. Zhou, Z. Tian, and L. Li, "Decimeter level indoor localization using Wi-Fi channel state information," *IEEE Sensors J.*, vol. 22, no. 6, pp. 4940–4950, Mar. 2022.

[9] J. Ding, Y. Wang, H. Si, S. Gao, and J. Xing, "Three-dimensional indoor localization and tracking for mobile target based on Wi-Fi sensing," *IEEE Internet Things J.*, vol. 9, no. 21, pp. 21687–21701, Nov. 2022.

[10] J. Zhang et al., "Gate-ID: WiFi-based human identification irrespective of walking directions in smart home," *IEEE Internet Things J.*, vol. 8, no. 9, pp. 7610–7624, May 2021.

[11] L. Deng, J. Yang, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "GaitFi: Robust device-free human identification via Wi-Fi and vision multimodal learning," *IEEE Internet Things J.*, vol. 10, no. 1, pp. 625–636, Jan. 2023.

[12] H. Wymeersch et al., "Integration of communication and sensing in 6G: A joint industrial and academic perspective," in *Proc. IEEE 32nd Annu. Int. Symp. Personal, Indoor Mobile Radio Commun. (PIMRC)*, 2021, pp. 1–7.

[13] C. Chen, H. Song, Q. Li, F. Meneghello, F. Restuccia, and C. Cordeiro, "Wi-Fi sensing based on IEEE 802.11bf," *IEEE Commun. Mag.*, vol. 61, no. 1, pp. 121–127, Jan. 2023.

[14] C. De Lima et al., "Convergent communication, sensing and localization in 6G systems: An overview of technologies, opportunities and challenges," *IEEE Access*, vol. 9, pp. 26902–26925, 2021.

[15] N. N. Bhat, R. Berkvens, and J. Famaey, "Gesture recognition with mmWave Wi-Fi access points: Lessons learned," in *Proc. IEEE 24th Int. Symp. World Wireless, Mobile Multimedia Netw. (WoWMoM)*, 2023, pp. 127–136.
[16] J. Yu, P. Wang, T. Koike-Akino, Y. Wang, P. V. Orlik, and H. Sun, "Human pose and seat occupancy classification with commercial mmWave Wi-Fi," in *Proc. IEEE Globecom Workshops*, 2020, pp. 1–6.
[17] N. N. Bhat, J. Sameri, J. Struye, M. T. Vega, R. Berkvens, and J. Famaey, "Multi-modal pose estimation in XR applications leveraging integrated sensing and communication," in *Proc. 1st ACM Workshop Mobile Immersive Comput. Netw. Syst.*, 2023, pp. 261–267.
[18] Y. Ma, G. Zhou, and S. Wang, "WiFi sensing with channel state information: A survey," *ACM Comput. Surveys*, vol. 52, no. 3, pp. 1–36, 2019.
[19] J. Zhang et al., "Data augmentation and dense-LSTM for human activity recognition using Wi-Fi signal," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4628–4641, Mar. 2021.
[20] J. Pegoraro, J. O. Lacruz, M. Rossi, and J. Widmer, "DISC: A dataset for integrated sensing and communication in mmWave systems," *IEEE Wireless Commun. Mag.*, vol. 63, no. 10, pp. 94–100, Oct. 2025.
[21] S. Subramanian et al., "Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–21.
[22] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.
[23] J. Du, T. Lin, C. Jiang, Q. Yang, C. F. Bader, and Z. Han, "Distributed foundation models for multi-modal learning in 6G wireless networks," *IEEE Wireless Commun.*, vol. 31, no. 3, pp. 20–30, Jun. 2024.
[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[25] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2234–2242.
[26] B. Lecouat, C.-S. Foo, H. Zenati, and V. R. Chandrasekhar, "Semi-supervised learning with GANs: Revisiting manifold regularization," in *Proc. Int. Conf. Learn. Represent. Workshop*, 2018, pp. 1–6.
[27] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2794–2802.
[28] J. Yang, X. Chen, H. Zou, D. Wang, and L. Xie, "Auto-Fi: Toward automatic Wi-Fi human sensing via geometric self-supervised learning," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 7416–7425, Apr. 2023.
[29] Y. Zheng et al., "Zero-effort cross-domain gesture recognition with Wi-Fi," in *Proc. 17th Annu. Int. Conf. Mobile Syst. Appl. Services*, 2019, pp. 313–325.
[30] W. Njima, A. Bazzi, and M. Chafii, "DNN-based indoor localization under limited dataset using GANs and semi-supervised learning," *IEEE access*, vol. 10, pp. 69896–69909, 2022.
[31] Y. Zhang et al., "Widar3.0: Zero-effort cross-domain gesture recognition with Wi-Fi," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8671–8688, Nov. 2021.
[32] D. Jiang, M. Li, and C. Xu, "WiGAN: A Wi-Fi based gesture recognition system with GANs," *Sensors*, vol. 20, no. 17, p. 4757, 2020.
[33] N. N. Bhat, R. Berkvens, and J. Famaey, "CSI4Free: GAN-augmented mmWave CSI for improved pose classification," in *Proc. IEEE 4th Int. Symp. Joint Commun. Sens. (JCS)*, 2024, pp. 1–6.
[34] C. Xiao, D. Han, Y. Ma, and Z. Qin, "CsiGAN: Robust channel state information-based activity recognition with GANs," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10191–10204, Dec. 2019.
[35] S. Wang, L. Wang, and W. Liu, "Feature decoupling and regeneration towards Wi-Fi-based human activity recognition," *Pattern Recognit.*, vol. 153, Sep. 2024, Art. no. 110480.
[36] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surveys*, vol. 53, no. 3, pp. 1–34, 2020.
[37] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.
[38] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 8934–8954, Sep. 2023.
[39] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22243–22255.
[40] M. J. Bocus, H.-S. Lau, R. McConville, R. J. Piechocki, and R. Santos-Rodriguez, "Self-supervised Wi-Fi-based activity recognition," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 552–557.
[41] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100378.
[42] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
[43] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3313–3332, Apr. 2023.
[44] L. J. Ratliff, S. A. Burden, and S. S. Sastry, "Characterization and computation of local Nash equilibria in continuous games," in *Proc. 51st Annu. Allerton Conf. Commun. Control Comput. (Allerton)*, 2013, pp. 917–924.
[45] Y.-N. R. Li, B. Gao, X. Zhang, and K. Huang, "Beam management in millimeter-wave communications for 5G and beyond," *IEEE Access*, vol. 8, pp. 13282–13293, 2020.
[46] A. Abdelreheem, E. M. Mohamed, and H. Esmaiel, "Location-based millimeter wave multi-level beamforming using compressive sensing," *IEEE Wireless Commun. Lett.*, vol. 22, no. 1, pp. 185–188, Jan. 2018.
[47] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.
[48] X. Lin et al., "5G new radio: Unveiling the essentials of the next generation wireless access technology," *IEEE Commun. Stand. Mag.*, vol. 3, no. 3, pp. 30–37, Sep. 2019.
[49] J. Pegoraro et al., "Jump: Joint communication and sensing with unsynchronized transceivers made practical," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 9759–9775, Aug. 2024.
[50] S. Bartoletti et al., "Positioning and sensing for vehicular safety applications in 5G and beyond," *IEEE Commun. Mag.*, vol. 59, no. 11, pp. 15–21, Nov. 2021.
[51] J. Pegoraro, J. O. Lacruz, M. Rossi, and J. Widmer, "SPARCS: A sparse recovery approach for integrated communication and human sensing in mmWave systems," in *Proc. 21st ACM/IEEE Int. Conf. Inf. Process. Sens. Netw. (IPSN)*, 2022, pp. 79–91.
[52] T. Nguyen, T. Le, H. Vu, and D. Phung, "Dual discriminator generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 2667–2677.
[53] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," in *Proc. 2nd Int. Conf. Comput. Sci. Technol.*, New York, NY, USA, 2018, pp. 1–20.
[54] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.
[55] N. N. Bhat, "Multi-modal pose estimation in XR applications leveraging integrated sensing and communication: Dataset," in *Proc. 1st ACM Workshop Mobile Immersive Comput. Netw. Syst.*, Nov. 2023, pp. 261–267.
[56] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diega, CA, USA, 2015, pp. 1–15.
[59] W. Li, R. J. Piechocki, K. Woodbridge, C. Tang, and K. Chetty, "Passive Wi-Fi radar for human sensing using a stand-alone access point," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 1986–1998, Mar. 2021.
[60] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[61] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–16.
[62] K. Shmelkov, C. Schmid, and K. Alahari, "How good is my GAN?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 213–229.
[63] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–38.
[64] C. Yu, Z. Wu, D. Zhang, Z. Lu, Y. Hu, and Y. Chen, "RFGAN: RF-based human synthesis," *IEEE Trans. Multimedia*, vol. 25, pp. 2926–2938, 2023.
[65] S. Y. X. Bang, S. M. Raza, H. Yang, and H. Choo, "EMP-GAN: Encoder-decoder generative adversarial network for mobility prediction," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2023, pp. 1–6.
[66] K. Nguyen and N. Ho, "Energy-based sliced Wasserstein distance," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 18046–18075.
[67] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, 2010.
[68] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.
[69] J. Yang et al., "SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing," *Patterns*, vol. 4, no. 3, 2023, Art. no. 100703.
[70] M. C. Gonçalves and R. Silva, "The effect of statistical hypothesis testing on machine learning model selection," in *Proc. Brazilian Conf. Intell. Syst.*, 2023, pp. 415–427.
[71] R. Müller, "How explainable AI affects human performance: A systematic review of the behavioural consequences of saliency maps," *Int. J. Human–Comput. Interact.*, vol. 41, no. 4, pp. 2020–2051, 2025.
[72] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, Oct. 2021.
[73] L. Yang et al., "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surveys*, vol. 56, no. 4, pp. 1–39, 2023.
[74] S. Yamaguchi and T. Fukuda, "On the limitation of diffusion models for synthesizing training datasets," 2023, *arXiv:2311.13090*.
[75] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7167–7176.
[76] J. Hoffman et al., "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.
[77] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–189.