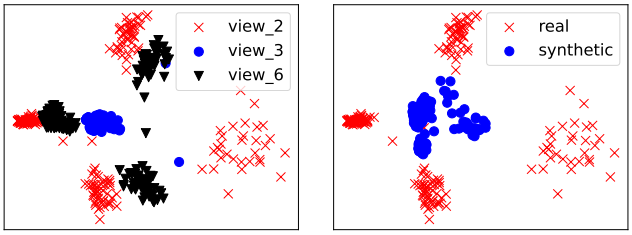
M4X: Enhancing Cross-View Generalizability in RF-Based Human Activity Recognition by Exploiting Synthetic Data in Metric Learning

Mengjing Liu, Zongxing Xie, Fan Ye
Department of Electrical and Computer Engineering, Stony Brook University
{mengjing.liu, zongxing.xie, fan.ye}@stonybrook.edu

***Abstract*—Human activity recognition provides insights into physical and mental well-being by monitoring patterns of movement and behavior, facilitating personalized interventions and proactive health management. Radio Frequency (RF)-based human activity recognition (HAR) is gaining attention due to its less privacy exposure and non-contact characteristics. However, it suffers from data scarcity problems and is sensitive to environment changes. Collecting and labeling such data is labor-intensive and time consuming. The limited training data makes generalizability challenging when the sensor is deployed in a very different relative view in the real world. Synthetic data generation from abundant videos presents a potential to address data scarcity issues, yet the domain gaps between synthetic and real data constrain its benefit. In this paper, we firstly share our investigations and insights on the intrinsic limitations of existing video-based data synthesis methods. Then we present *M4X*, a method using metric learning to extract effective view-independent features from the more abundant synthetic data despite their domain gaps, thus enhancing cross-view generalizability. We explore two main design issues in different mining strategies for contrastive pairs/triplets construction, and different forms of loss functions. We find that the best choices are offline triplet mining with real data as anchors, balanced triplets, and a triplet loss function without hard negative mining for higher discriminative power. Comprehensive experiments show that *M4X* consistently outperform baseline methods in cross-view generalizability. In the most challenging case of the least amount of real training data, *M4X* outperforms three baselines by 7.9-16.5% on all views, and 18.9-25.6% on a view with only synthetic but no real data during training. This proves its effectiveness in extracting view-independent features from synthetic data despite their domain gaps. We also observe that given limited sensor deployments, a participant-facing viewpoint and another at a large angle (e.g. 60°) tend to produce much better performance.**

I. Introduction

Human activity recognition is vital for health management, enabling remote monitoring and personalized interventions by providing insights into physical and mental well-being, facilitating chronic disease management, and contributing to overall health research [1]. The recognition of activities during the routines of our daily life can provide valuable assistance in managing diseases [2]. For instance, monitoring body movements and gait speed can facilitate the assessment of severity, progression, and medication response for Parkinson's disease in a home environment [3]. Human activity recognition using RF technologies, e.g., ultra-wideband (UWB), millimeter wave (mmWave), and Wi-Fi, has gained siginificant attention [4]–[8]. They have much less privacy exposure and pose little physical or cognitive burden compared to cameras and wearables [4], [5], [9]. However, RF-based sensing technologies suffer from generalizability issues due to environmental variations and data scarcity [6], [10], [11].



(a) Doppler distributions vary at different views (b) Domain gaps exist between real and synthetic RF data

Fig. 1: Doppler data variability across different views presents a cross-view generalizability challenge. While RF data synthesis from videos looks promising to data scarcity, it faces the issue of domain gaps between real and synthetic data.

Typically, RF-based human activity recognition (RF HAR) relies on characterizing RF signals through micro-Doppler signatures, which represent the distribution of radial velocities in space and time.[1] Such signal characterization is sensitive to environmental factors that introduce variations in the RF propagation, including the room and furniture layout, and in particular the relative view (i.e., orientation) of the sensor to the human subject. Due to cost, installation and maintenance constraints, the number of sensors deployed in a real home is often limited, leading to few and very different viewpoints. Due to the view-dependent nature of radial velocity patterns, such disparate viewpoints produce significant variations in RF signal characteristics (as illustrated in Figure 1(a)), thus divergent micro-Doppler signatures [9] that make cross-view generalizability very challenging.

Unfortunately, RF HAR generalizability is further complicated by the scarcity of open-source RF datasets, both in quantity and diversity. Collecting and labeling RF data is labor-intensive and time-consuming [12], requiring expertise for accurate annotation. This sharply contrasts the ease of

This work is supported in part by NSF grant 1652276, 1951880, 2119299.

[1]For simplicity, we interchangeably use the terms "RF data" or "Doppler data" with micro-Doppler signatures throughout this paper.

working with images or videos, where substantial benchmark datasets exist, and labeling can be done by virtually anyone, without special domain knowledge.

Despite recent work on few-shot learning and transfer learning with limited RF data [13], cross-view generalizability has not been sufficiently addressed. Such work often still requires large amounts of labeled data for model training or the generation of view-independent representations [5], [9]. Moreover, they primarily focus on hand/arm gesture recognition, rather than recognizing full human body activities, which poses more challenges due to complex combinations of movements and mutual occlusions of the torso, head, and limbs. Thus the cross-view generalizability in RF HAR remains unresolved.

To address the data scarcity issue, researchers have explored RF data synthesis using the abundantly available video data to augment RF data in both quantity and diversity [10], [14], [15]. This involves simulating the radar cross-section (RCS) reflection of human bodies using radar (e.g., the physics of Doppler effect) and computer vision (CV) techniques applied to videos of human activities. However, such simulations are inherently imperfect. We find that they heavily rely on 3D human pose/mesh estimation models, which suffer from occlusions, and have inherent depth-ambiguity problems due to unknown relative depth between body joints [16]. This results in domain gaps between synthetic and real data (as illustrated in Figure 1(b)), hindering the direct utilization of synthetic data [10].

In this paper , we aim to enhance cross-view generalizability in RF HAR under practical constraints, including: 1) a small amount of real, labeled RF data from limited viewpoints; and 2) substantial synthesized RF data of comprehensive viewpoints through physics-guided simulation from abundant labeled video data. We approach this problem through two tasks: 1) extracting view-independent features for cross-view generalizability of HAR; and 2) extracting source-independent features capitalizing abundant labeled synthetic data, compensating for the scarcity of real RF data. We consider two types of domain gaps: "view" refers to variations in RF data from different viewpoints, while "source" distinguishes the data origin, real versus synthetically generated.

To this end, we introduce ***M4X*** (as illustrated in Figure 2), a method that employs ***metric*** *learning* [17] to exploit *synthetic RF data* for enhancing ***cross-view*** *generalizability* in RF HAR. We examine two critical design issues for metric learning: the strategies for mining contrastive pairs/triplets and the selection of loss functions. We evaluate and compare ***M4X*** with representative baselines in addressing data scarcity through "synthetic-to-real" approaches. The experimental results show that under limited real data, ***M4X*** can effectively exploit abundant synthetic data to extract domain-independent representations to significantly enhance cross-view generalizability in RF HAR.

We summarize our contributions as follows:

- We conduct an in-depth investigation on video-to-RF data synthesis methods that hold promise to address the data scarcity problem. We discover their inherent limitations due to depth-ambiguity, occlusion, and inaccurate simulation, and share insights on how they impact the fidelity of synthesized data, and pose challenges for their direct utilization.
- We present ***M4X***, which employs metric learning to extract source- and view-independent features from synthetic data to enhance cross-view generalizability. We systematically examine the design choices in ***M4X***, and identify offline triplet mining with real data as anchors, and a triplet loss without "hard negatives" as optimal strategies for contrastive pairs/triplets construction and loss function selection, respectively.
- We conduct comprehensive experiments to compare ***M4X*** to three baseline methods. The results show that ***M4X*** consistently and significantly outperforms baselines by 7.9-16.5% on different views, and 18.9-25.6% on an view with only synthetic but not real data in training, proving its effectiveness in extracting view-independent features from synthetic data despite their domain gaps. We also identify the combination of viewpoints for the best performance, providing guidelines for sensor placements under practical constraints.

To the best of our knowledge, this is the first work that explores how to effectively leverage synthetic data despite their domain gaps to enhance cross-view generalizability in RF HAR.

II. Background

A. Cross-view Generalization

In the context of RF HAR, variations in RF data patterns arise when factors such as relative distance, sensor viewpoints, and room layout change [9]. There are existing works managing variations across environments or individuals by enhancing dataset diversity [4] or developing data possessing techniques, such as normalization [5] and auto-encoder decoder [14], among others. However, attaining cross-view generalizability remains a significant challenge, one that cannot be resolved solely through these methods. [2] When the viewpoint of a RF sensor changes, such as due to a different sensor deployment or the subject facing another direction, the observed radial velocities and resulting Doppler patterns are significantly affected (Figure 1(a)), posing a challenge in cross-view generalization of RF HAR.

Existing works on cross-view generalization can be primarily categorized into two types: **1) Comprehensive-view modeling.** WiAG [9] generates gesture data at all locations and orientations to train one model at each location and view. Then they select the corresponding model to recognize data based on the location and view of the data. DI-Gesture [4] proposes to augment gesture data regarding different speeds, trajectories, distances and angles so that the model is trained with diverse enough data thus independent to those factors. They use angle of arrival (AoA) information as an important feature from multiple receiving antennas. **2) View-independent feature extraction.** Widar3 [5] presents a view-independent representation called BVP (Body-coordinated velocity profile) to train

[2]We define the 'view' as the viewpoint of the RF sensor facing the human subject from a specific position and orientation.

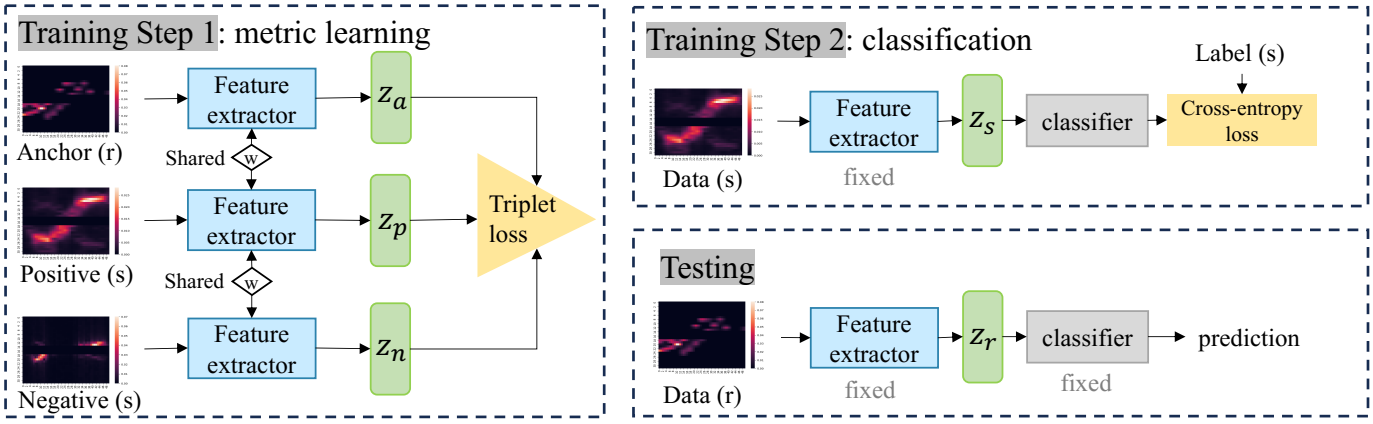


Fig. 2: Framework of our paradigm, ***M4X***, using metric learning to exploit synthetic data for enhancing cross-view generalizability in RF HAR. "(r)" denotes data in real domain, "(s)" data in synthetic domain. There are 2 steps in training: 1) train a metric learning model as a feature extractor across viewpoints and sources to get view-independent features across real and synthetic data sources. The model takes a 3-way input including anchors from real domain and positives and negatives from synthetic domain. 2) employ the trained model as a feature extractor to extract both cross-view and cross-source features from a large synthetic data set, and train a classifier for activity recognition. Although trained in synthetic domain, the classifier is effective in real domain because of the cross-source nature of the features. For testing, we employ the same feature extractor to extract features in real domain and predict the activity with the classifier obtain in the step 2.

recognition models that are orientation agnostic. However, previous works in both categories are designed for gestures-likes simple movements within a confined space, rather than whole body activities with complex occlusions in a large space. Their design and verification target only at gesture recognition, its effectiveness for human activities requiring greater full-body coverage and complex occlusions is unclear and needs in depth investigation to prove either way. Furthermore, for training they still need a large dataset, whereas one with sufficient amount and diversity for real environments may not exist.

In summary, cross-view generalizable human activity recognition under data scarcity remains unsolved.

B. RF data scarcity

Data scarcity is a serious challenge for RF-based sensing. Because RF data is not immediately understandable by humans, labeling such data is labor-intensive, time consuming, and requires respective expertise; the need for human subject protection also compounds the overhead and complexity. Widar3 [5] publicizes a wireless gesture recognition dataset collected using Wi-Fi devices incorporating Channel State Information (CSI) in various domains, accounting for the subject's location, orientation, and the room layout. As the dataset is designed for gesture recognition, activities involving larger spaces of the full human body and complex kinematics, such as 'sitting down' and 'walking,' are not included. Gurbuz et al. [18] and Rahman et al. [19] publicize RF activity recognition data sets collected with UWB, FMCW radars and NI-USRP2922 model software-defined radios. Despite the value of the datasets, their size and diversity are insufficient for complex real-life deployments, particularly given the significant variations in RF data distributions arising from differences in the data collection environment, sensor parameters, deployment setup, etc. We believe, while additional datasets with increased quantity and diversity are essential, leveraging synthetic data [20], [21] is an orthogonal approach that can greatly accelerate the development of final solutions.

C. Video-to-RF Data Synthesis

Recent efforts on video-to-RF data synthesis seek to compensate for the scarcity of RF data in both quantity and diversity [14], [15], [22]. Vid2Doppler [14] generates 3D mesh model of human body from videos, compute the radial velocity of each vertex on the mesh and synthesize the Doppler signal based on the physics of Doppler Effect [23]. SynMotion [10] employs a pictorial body model composed of primitive ellipsoids and simulates the process of signal being reflected on the body model and received. Midas [15] generates convertible radar data in both single-human and multi-human scenes, while significantly eliminating data redundancy to ensure model stability. These studies demonstrate the potential of synthesized data from videos to address the scarcity of RF data. However, we do find they suffer limitations on depth ambiguity, occlusion and simulation fidelity (Section III), thus direct utilization is not effective or even harmful. In our preliminary experiments, training an SVM classifier with a direct combination of real and synthetic RF data results in a decrease in accuracy from 72.6% to 68.5% (a degradation by 4.1%) when compared to training with real data alone. This implies that the inclusion of synthetic data does not yield immediate performance benefits, highlighting the importance of addressing the domain gap between real and synthetic data.

D. Metric Learning

Metric learning aims to learn proper distance metrics to measure the similarity among samples, and the learned optimal metric can be used for various other tasks. Siamese and Triplet networks are the most commonly used network structures [17]. Siamese network takes two-way inputs consisting of "anchor", "positive" or "negative" pairs and learns an embedding which maps "anchor" and "positive" close to each other, while "anchor" and "negative" far apart. On the other hand, Triplet network takes three-way inputs ('anchor,' 'positive,'

and 'negative' triplets) with a similar learning objective. [3] The discriminative power of the learned representation from metric learning is determined by the degree of similarity or dissimilarity among the constructed pairs/triplets [17]. There are two major issues that determines the discriminative power:

1) Contrastive pair/triplet mining: how to determine and select positive and negative samples. Among all positives/negatives, "hard" positives/negatives—those that are far from/close to the "anchor" and thus hard for models to distinguish—have significant implications for learning more discriminative representations [24]. We explore two categories of contrastive pair/triplet mining methods: 1.1) Online Mining randomly selects mini-batches during training and dynamically chooses "hard" positives and negatives within each batch, based on their distances to the anchors in the current embedding space [25], [26], [27]. 1.2) Offline Mining pre-selects fixed pairs/triplets from the entire dataset before training, without adapting to changes in the embedding space during the training process [28].

2) Loss functions define the similarity metrics between anchor, positive and negative samples and guide the model to learn an embedding with appropriate distances among them. Loss functions have been demonstrated to have a significant impact on the model performance [29], [17]. Contrastive loss and triplet loss are widely used in Siamese and Triplet networks respectively. Most of the work is in the field of computer vision, and despite the rapid development, these factors are rarely studied in RF HAR. Therefore, this work investigates the critical design aspects of metric learning for cross-view RF HAR, including contrastive pairs/triplets mining and the form of loss functions.

III. Video-to-RF Data Synthesis

Recent work studies [10], [14], [15] a novel paradigm of generating synthetic RF data from extensive video datasets. Such endeavors have shown promise for augmenting the quantity and diversity of RF data through the translation of visual information into the radio frequency domain. The key idea is simulating the radar cross-section (RCS) reflections of human bodies by estimating 3D human pose/mesh from videos using computer vision (CV) models [30]. Such simulation for activities/movements can produce micro-Doppler signatures[4] following well established radar principles [18].

We examine such video-to-RF data synthesis for an in-depth understanding of its capabilities and constraints. For the purpose of illustration, we use Vid2Doppler [14] as a representative pipeline to depict the workflow of video-to-RF data synthesis, without loss of generality (as shown in Figure 3):

1) Construct 3D human pose/mesh. In the example of Vid2Doppler, VIBE [31] is employed to produce a 3D mesh from the video based on SMPL [30], employing 6890 vertices to model the human body. The 3D coordinates of each vertex are obtained for every frame.
2) Specify the viewpoint of a virtual sensor. Notably, the sensor's viewpoint determines the radial velocity based on the velocity of the vertex. Additionally, it determines which vertices are within the view of the virtual sensor and which ones are occluded by other body parts.
3) Calculate the radial velocity of each vertex based on its translation between two consecutive frames and the frame rate.
4) Obtain the principle synthetic Doppler (Figure 4(d)) based on the radial velocity distribution of all vertices across the velocity range (e.g., -3m/s to 3m/s in our case) in a histogram format.

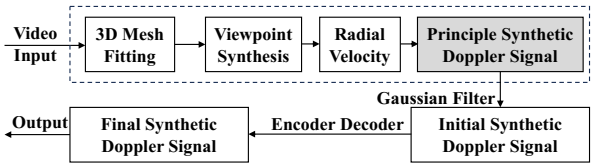


Fig. 3: Workflow of video-to-RF data synthesis.

Up to this point, we have covered the process of obtaining the principle synthetic Doppler (as illustrated in the dashed box in Figure 3). Usually additional operations will further refine the synthetic Doppler. As an example, Vid2Doppler employs a Gaussian filter and an encoder-decoder mechanism to convert the initial synthetic Doppler to real Doppler, considering noise and non-linearity in the signal characteristics.

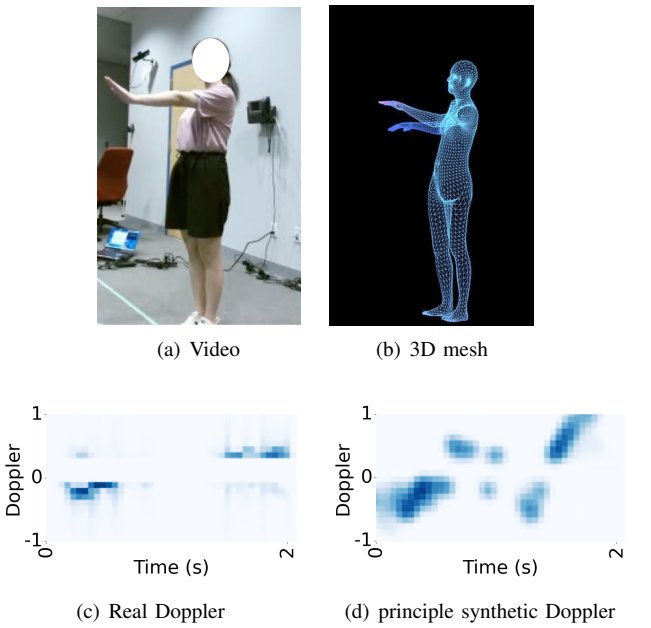
For better interpretability and explanability, we focus our discussion on the principle synthetic Doppler, the domain part impacting the fidelity of the final synthetic Doppler. It is directly derived from the 3D mesh of the human body and radial velocities based on the physics principles of RF sensing, e.g., signal propagation and reflection. Specifically, we discuss the limitations on three key aspects: depth-ambiguity, occlusion, and inaccurate simulation of signal reflection.

Depth-ambiguity. Depth-ambiguity problem exists inherently in vision-based 3D human pose estimation due to the unknown relative depth between body joints [16]. The depth estimation significantly influences the estimation of radial velocity - thus synthetic Doppler, leading to domain gaps between real and synthetic data. To illustrate, we show the Doppler data of periodical activity of "push and pull" in real and synthetic domains at three views in Figure 5. The orientations of view 4, view 6 and view 1 are shown in Figure 6. View 4 faces directly toward the person, while view 6 is 60-degree off to the right. View 1 is perpendicular to the person's facing direction, to the left of the person. The Doppler shift amplitude (i.e., the y-axis of Doppler data) depends on radial velocity, and the shade at each pixel of the Doppler data represents the signal energy reflected by the moving object at the corresponding velocity.

As shown in Figure 5, there is a clear difference in the data pattern between the real and synthetic domains in View 4 (radial view) and View 1 (tangential view). In the real domain (Figure 5(b), 5(e), 5(h)), the Doppler shift amplitudes decrease

[3] In this context, we recognize that metric learning and contrastive learning share similar underlying principles. Given that metric learning encompasses a wider range of methods, such as triplet loss, quadruplet loss, among others, beyond just contrastive loss, we adopt metric learning as the foundational concept for our study.

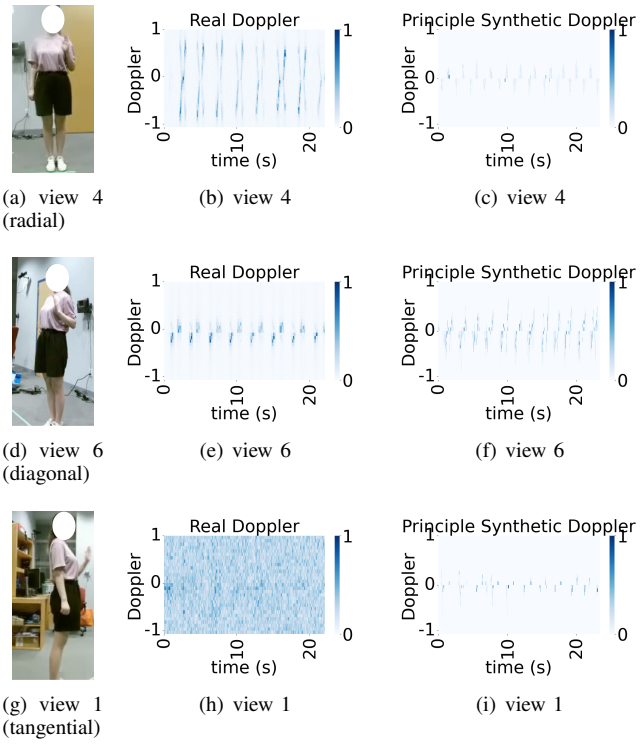
[4] For simplicity, we use "Doppler" to denote micro-Doppler signatures.



(a) Video (b) 3D mesh

(c) Real Doppler (d) principle synthetic Doppler

Fig. 4: Video image and 3D mesh, and corresponding real Doppler and principle synthetic Doppler of the same activity: push and pull. The 3D mesh has errors because of occlusion, resulting in more velocity components and larger signal energy in synthetic Doppler than the real Doppler.



(a) view 4 (radial) (b) view 4 (c) view 4

(d) view 6 (diagonal) (e) view 6 (f) view 6

(g) view 1 (tangential) (h) view 1 (i) view 1

Fig. 5: Doppler data corresponding to the "push and pull" activity at view 4, view 6 and view 1 in both real and synthetic domains. These views represent the tangential, radial, and diagonal views, respectively, with respect to the participant's facing direction. The orientations of view 4, view 6 and view 1 are shown in Figure 6. From Figure 5(e), 5(f) there are noticeable similarities in Doppler patterns between real and synthetic domains at the diagonal view (view 6). From Figure 5(b) and 5(c), Figure 5(h), 5(i), there are discrepancies in the Doppler patterns in two domains at the tangential and radial views (view 1, 4).

from view 4 to view 1, with the Doppler signal at view 1 buried in noise. This is because when the participant faces the sensor in view 4, the radial velocity is maximum. At view 6, the radial velocity is the actual velocity's projection at about 60°, much smaller than the actual velocity. At view 1, the radial velocity is nearly zero because the movement is in the tangential direction. However, in the synthetic domain (Figure 5(c), 5(f), 5(i)), the Doppler shift amplitude at view 6 is larger than that at view 4; at view 1, clear patterns of "ups and downs" persist, contradicting the ground truth. Both are due to depth ambiguity: in view 4, the depth of radial motion was estimated at much less than the actual, while much more in view 1, leading to incorrectly generated magnitudes. This demonstrates one inherent limitation from depth ambiguity, resulting in domain gaps between real and synthetic data.

Occlusion. Occlusion of body parts can significantly mislead the 3D pose estimation model [32], causing errors in 3D mesh, thus the synthetic Doppler data. As shown in Figure 4(a), the participant is pushing the left arm. However, in the generated 3D mesh in Figure 4(b), the figure appears to be pushing both arms simultaneously. The error happens because the right arm is occluded by the torso in video. This leads to the Doppler signal being generated from both arms, thus greater Doppler signal energy and more velocity components in synthetic data, as shown in Figure 4. Despite efforts on occlusion-aware pose estimation, it remains very challenging [33], [32]. Errors in 3D pose estimation inevitably leads to inaccurate calculations and synthetic Doppler data. The 3D pose models can not be perfect, their errors from occlusion will add to domain gaps.

Inaccurate Simulation of Signal Reflection. When the 3D pose model is reasonably accurate, the synthesized data appears similar to real data (Figure 5(e), 5(f)), but domain gaps still exist. This is because the simulation of signal propagation and reflection on the human body cannot be perfect, even when the 3D mesh model is accurate. This results from multiple factors:

- RF signals experience reflection, scattering, and absorption on the human body [34]. Existing works often assume that the signal is solely reflected without considering scattering or absorption effects, which diffuse and attenuate the signal energy. Also accurately simulating signal scattering and absorption on the human body is very difficult; the complex contours further compounds the challenge.
- RF signal reflection factors vary with different materials [35]. Conductors are more reflective, while insulators allow more RF signals to pass through and a smaller proportion are reflected. In addition, the reflection is also affected by the roughness of the surfaces, angle of incidence, etc. Those factors are not considered in existing data synthesis works.
- RF signal noise and environment influence (e.g. moisture, temperature, reflection on furniture) can not be perfectly simulated either.

Collectively, these issues contribute to the inherent gaps between synthetic and real RF data, even when a perfect 3D human pose/mesh model is available.

Despite the domain gaps, we observe (in Figure 5(d)) there are noticeable similarities in patterns, such as periodic "ups" and "downs" between the real and synthetic domains. While the data distributions in the two domains are not identical, the observed similarities holds promise for synthetic data to compensate for RF data scarcity. To maximize the value of synthetic RF data, next we explore metric learning to effectively bridge the domain gaps.

IV. *M4X*: Metric Learning for cross-view generalization

We present ***M4X*** (as illustrated in Figure 2) that extractx view-independent features from synthetic Doppler data to improve viewpoint generalizability under data scarcity. Firstly, we extract shared features across viewpoints and data sources to obtain view-independent features by metric learning. A small amount of real data and a large amount of synthetic data from different viewpoints are used to train a metric learning model, which will be used as a feature extractor for subsequent classification tasks. Secondly, we use the trained feature extractor to produce source- and view-independent representations as the input for training a classifier of HAR, capitalizing on substantial synthetic data. Lastly, we test the feature extractor and classifier from the two previous training steps to infer human activities on real RF data, including from views where no real data were "seen" during training.

While the overall pipeline may seem straightforward, careful consideration of design choices in metric learning is crucial for its effectiveness, especially in the context of data scarcity and mixed training sets from both domains. We conduct an in-depth investigation in two key design aspects: **Contrastive Pair/Triplet Mining** and **Loss Functions**.

A. Contrastive Pair/Triplet Mining

Metric learning models learn an embedding that maps anchors and positives closer, while anchors and negatives distant in the latent space. The effectiveness of learned features depend heavily on the construction of contrastive pairs or triples. Given the objective of cross-view robustness, the model needs to extract features that are sensitive to class labels while insensitive to changed viewpoints. Thus we determine positive and negative samples by whether the sample shares the same class label as the anchor, irrespective of data viewpoints. In addition, to extract shared features across real and synthetic data sources, we employ real data as the anchor, with synthetic data from the same class as positives and those from different classes as negatives. Given that real RF data with labels is limited, while synthetic data with labels from video can be plentiful, this allows us to create more pairs/triplets to best utilize and mine information from precious real data, and effectively utilize a large amount of synthetic data.

Next, we investigate two methods for selecting positive and negative samples: online pair/triplet mining and offline pair/triplet mining. Online mining focuses on selecting "hard" positives and negatives during training, utilizing the current embedding space, which conserves resources by avoiding "easy" positives/negatives and speeds up the training. Conversely, offline mining pre-selects fixed pairs/triplets before training, ensuring gradient stability and reducing the risk of overfitting or underfitting.

Online Mining Inspired by [29], we employ a K-P batch construction which has balanced mini-batches, each with K classes and P samples per class. We tailor the design in our hybrid training set scenario: Randomly select K classes from the entire set and randomly select P samples for each class in each domain, resulting in a mini-batch consisting of $K \times P$ samples in real domain and another $K \times P$ samples in synthetic domain. As the determination of positives and negatives relies on class labels, achieving class balance in a mini-batch ensures a balanced representation of positives and negatives for each class in each batch. This balance contributes to the stability of gradient computation during metric training.

In addition, we explore two design options with online mining: 1) Select "hard" positive/negative samples within the mini-batch based on their distance to the anchor in the embedding space. This improves the efficacy of metric learning by choosing discriminative samples, but at the cost of introducing high gradient variance. 2) Use all positive/negative samples within the mini-batch to construct learning pairs/triplets. Since the mini-batch size is usually small (e.g. $K = 4, P = 4$), it is practical to use all positive/negative samples without excessive computing overhead.

Offline mining We construct contrastive pairs/triplets offline using the whole dataset before training. We employ each instance in real dataset as an anchor, and select positives from those synthetic instances of the same class, and negatives from those synthetic ones of different classes. Specifically, we use all synthetic samples of same class in the whole dataset as positives. The number of positive samples depends on the number of samples per class in the dataset. Since there are many more other classes, the pool of potential negative samples is much bigger.

To utilize the positive and negative samples in the limited dataset efficiently, we investigate two design options for offline pair/triplet mining: 1) All Triplets: exhaust all possible positive-negative combinations to construct pairs/triplets. In this way, we can train a metric learning model using all possible triplets in the limited data set. Unlike online mining using mini-batches, when building triplets using the entire dataset, there can easily be millions of triplets even with hundreds of data samples. 2) Balanced Triplets: utilize all positive samples and an equal number of negative samples: basically we randomly select one negative sample for each (anchor, positive) pair to maintain a balanced ratio between positive and negative samples. We do this because using all negative samples can bury the most valuable information in small numbers of "hard" negatives under large amounts of "easy" negatives.

B. Loss Functions

Different loss functions are explored in recent metric learning studies [24], [29], [36], [37]. Due to the inclusion of both synthetic and real data in our training dataset, existing loss

functions developed for single-domain training data may not be suitable. Thus we firstly investigate two primary metric learning loss functions commonly used for Siamese Network and Triplet Network, and then tailor them for our scenario with 1) two domains of training data, and 2) labels in both domains.

The InfoNCE loss function [38] for unsupervised Siamese network training (with **two-way input**). In this setting, the positive sample is a transformed version of the anchor (for instance, a cropped image derived from the anchor image), and all other samples in the batch, excluding the anchor, are treated as negatives. However, since this approach operates without labels, it cannot discern whether these "negatives" belong to the same class as the anchor. Utilizing these negatives in training could potentially mislead the network.

In our context with labels, we employ a supervised transformation (abbreviated as "SUP") of contrastive loss which can handle multiple positives in the batch and eliminate false-negative problems. The loss function is shown to have the intrinsic ability to perform hard positive/negative mining for visual data [37], expressed as in Equation 1:

$$L_{SUP} = \frac{1}{|I|}\sum_{i\in I}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(z_i\cdot z_p/\tau)}{\sum_{a\in Ai}\exp(z_i\cdot z_a/\tau)} \quad (1)$$

I denotes a batch, $P(i)$ denotes all positive samples of the i-th anchor. z_i, z_p denote the embedding of anchor and positives respectively. z_a denotes the embedding of all samples except anchor in the batch. $\cdot$ denotes inner product. τ is a constant for normalization. $A(i) = I - \{i\}$. By minimizing the loss, the model learns an embedding that maximizes the ratio of the similarity between the anchor and positive samples to the similarity between the anchor and all samples.

In our scenario, with a hybrid training dataset, we have adapted the loss function, as expressed in Equation 2.

$$L_{SUP-hyb} = \frac{1}{|I^r|}\sum_{i\in I^r}\frac{-1}{|P(i)^s|}\sum_{p\in P(i)^s}\log\frac{\exp(z_i^r\cdot z_p^s/\tau)}{\sum_{a\in I^s}\exp(z_i^r\cdot z_a^s/\tau)} \quad (2)$$

Particularly, we assign the anchor from the real data source (denoted as z_i^r) and positives from the synthetic data source (denoted by z_p^s). We replace z_a with z_a^s, representing embedding of all samples in the batch from synthetic domain. By using anchors from real data sources as well as positive and negative values from synthetic data sources, we learn features that are shared across data sources.

Batch-Hard triplet loss function [29] for Triplet Network training (with **three-way input**). It takes exactly one "hardest" positive sample and one "hardest" negative sample from the mini-batch for each anchor. The "hardest" positive sample is the one that has the largest distance from the anchor among all positive samples, while the "hardest" negative sample is the one that has the smallest distance from the anchor among all negative samples. It is expressed in Equation 3:

$$L_{BH} = \frac{1}{|I|}\sum_{a\in I}[m + \max D(z_a, z_p) - \min D(z_a, z_n)]_+ \quad (3)$$

$D(\cdot)$ is the distance metric of embedding, such as Euclidean distance. z_a, z_p, z_n denote the embedding of anchor, positive and negative sample respectively. m is a marginal parameter that determines how much we want to differentiate between positive and negative instances. By minimizing the loss, the model learns an embedding where the distance between the anchor and the "hardest" positive samples is smaller than the distance between the anchor and the "hardest" negative sample by at least a margin m. It selects the most discriminative triplets, aiming at improving the correct classification of the "hardest" cases.

We tailor it into our hybrid training scenario, expressed in Equation 4. Basically, we assign anchor from real data source (denoted as z_a^r), and select the "hardest" positive and negative samples from synthetic data source (denoted as z_p^s, z_n^s). Thus the model learns features across data sources.

$$L_{BH-hyb} = \frac{1}{|I^r|}\sum_{a\in I^r}[m + \max D(z_a^r, z_p^s) - \min D(z_a^r, z_n^s)]_+ \quad (4)$$

We further design a third loss function (expressed in Equation 5) to use All triplets constructed within a mini-Batch (abbr. "BA") when calculating the loss value. It provides stability in scenarios where the "hardest" triplets may be outliers or contain noise. We compare using the "hardest" triples versus using all triples in the mini-batch to find out which can better extract view-independent features in hybrid data scenarios. Because it does not require hard positive/negative mining, it is also appropriate for offline mining where triplets are determined, selected and fixed before training.

$$L_{BA-hyb} = \frac{1}{|I^T|}\sum_{(a,p,n)\in I^T}[m + D(z_a^r, z_p^s) - D(z_a^r, z_n^s)]_+ \quad (5)$$

I^T is the set of all the constructed triplets. We explore and compare three loss functions with different pair/triplet mining strategies in our scenario to find the best design option (detailed in Section V).

V. Evaluation

A. Implementation

We employ the XeThru UWB x4m03 [39] as the RF front-end for wireless sensing of human activities. Additionally, we utilize the Kinect [40] camera to record videos of human activities for Doppler data synthesis. The UWB-based RF front-end operates at a center frequency of 8.75 GHz, and we configure its frame rate to 180 frames per second, covering a velocity range of $\pm 3m/s$. The Kinect camera operates at a frame rate of 30 frames per second. Figure 6 illustrates the arrangement of the six sensors, including the UWB sensor and Kinect camera, along with the corresponding six views. The deployment of sensors at multiple views allows for an exploration and comparison of the performance across different view combinations and validation of the cross-view generalization of activity recognition. The participant is positioned centrally (2 meters away from each sensor) and facing v_4 while engaging in various activities.

Due to the lack of publicly available RF datasets from multiple viewpoints for HAR, we collect a dataset for five activities at six distinct viewpoints to facilitate validation efforts, including gestures, torso and limb movements widely studied in activity and gesture recognition: sitting down, standing up, stepping, pushing and pulling, and drawing a circle, [5], [14]. Each activity is sampled 50 times across six different views, resulting in a total of $5 \times 50 \times 6 = 1500$ activity samples

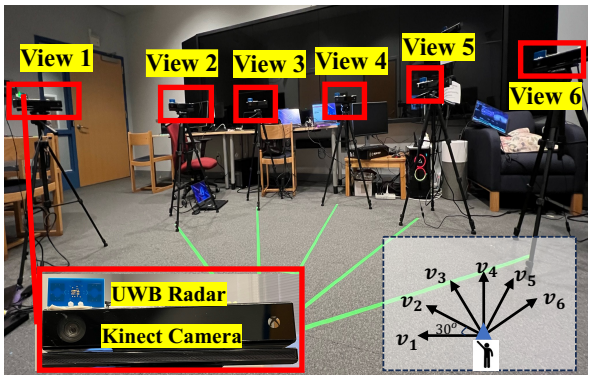


Fig. 6: The experimental setup comprises sensors deployed at six distinct views. At each view, a UWB radar and a Kinect camera are positioned at the same location, facing towards where the person stands. The arrangement ensures that adjacent views are 30 degrees apart. The participant is positioned centrally, situated 2 meters away from each sensor, and facing v_4 while engaging in activities.

for both RF and video data. In addition to the real RF data, we generate synthetic data for each view from the video data, following the implementation of Vid2Doppler [14] up to the step involving the pretrained encoder-decoder model. Consequently, we have 1500 activity samples in the synthetic domain.

We build the metric learning model comprising two Convolutional Neural Network (CNN) layers, succeeded by Max-pooling layers and fully connected layers. Relu activation is applied to each layer, while the training employs the Adam optimizer. The learning rate follows an exponential decay with an initial rate of 5×10^{-4} and a decay rate of 0.9. We use two representative types of classifiers, Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP), for comparison in the second step. For the MLP model, we use one hidden layer with 200 hidden neurons.

B. Micro-benchmarking of Different Design Choices

We conduct experiments to compare the design choices for metric learning systematically to find which options are most effective in cross-view generalizability with limited real data and a large amount of synthetic data in training. Doppler at view 4, 5, 6 is used to perform leave-one-view out test: in real domain, only two views' data is used for training, and three views' data is used for testing. We define a **"seen"** view as one in which real data is present in the view under training, and an **"unseen"** view as one in which there is no real data but only synthetic data in the view under training. The dataset is split for training and testing at 80:20 ratio, where 20% in real domain is preserved for testing. The corresponding 20% for the same activities in synthetic domain is excluded from training to prevent information leakage of the test set. This precaution is taken due to the shared information between synthetic and real data. For training, we use the remaining 80% in synthetic domain and 80% in real domain unless specified; and the same 80% in synthetic domain to train the classifier. We use different view combinations in experiments for comprehensive verification of cross-view generalizability.

Overall, experiments show that the most effective options are: to use real instead of synthetic data as anchor, offline triplet mining with balanced triplets instead of all triplets or online mining strategies, and loss function L_{BA-hyb} without hard negative mining. While using real data at view 4 and 6 for training and real data at view 4, 5, and 6 for testing, we achieve 86.0% in accuracy with the best design combination, outperform all other design choices with 17.0% improvement. For cross-view generalizability, we achieve 85.5% in accuracy at unseen views, outperform all other design choices with 25.8% improvement. In addition, we can improve accuracy at both seen and unseen views by utilizing synthetic data at unseen views. In the following, we present comparison results of different design choices and the insights from each.

1) Contrastive Pair/Triplet Mining: We compare contrastive pair/triplet mining strategies in four aspects to find which option can best utilize limited real data and a large amount of synthetic data in training.

Real versus synthetic data as anchors. Experiments show that using real data as anchors achieves an accuracy of 80.3%, with 11.3% improvement (especially 24.5% improvement at unseen views) compared to using synthetic data as anchors, showing that real data containing "authentic" Doppler information are more effective anchors. [5]

Online mining: K-P batch versus Random Batch. We compare K-P batch construction with Random Batch construction to find which leads to better convergence. For Random Batch, we randomly select $K \times P$ samples in each domain resulting a mini-batch with $2 \cdot K \cdot P$ samples. Table I shows the comparison of results with different loss functions and classifiers based on accuracy (in %). Parameters τ and m are set to 100 and 0.2, respectively. From the results, K-P batch construction outperforms random batch construction, with an average improvement of 3.5%. The most notable improvement is 15.5% when employing the L_{BA-hyb} loss function with the SVM classifier. The superiority is attributed to the class balance within each mini-batch while using K-P batch construction.

TABLE I: Comparison of different online mining strategies and loss functions

Loss/Batch	K-P batch	Random Batch
	step 2 classifier: SVM	
$L_{SUP-hyb}$	53.3	50.4
L_{BH-hyb}	59.7	58.9
L_{BA-hyb}	**65.9**	50.4
	step 2 classifier: MLP	
$L_{SUP-hyb}$	60.5	49.6
L_{BH-hyb}	58.9	59.7
L_{BA-hyb}	64.3	48.9

Offline mining: Balanced Triplets versus All Triplets. We compare these two design options using different percentages of real data in training to explore how they perform when there is more data - and therefore more triplets in training. Results show that, compared to using all triplets, using balanced triplets leads to a significant increase in accuracy—from 10.9% to 82.1% and 16.3% to 79.8%—when 32% and 16% of real data are used in training, respectively.

[5]Due to space limit, we summarize the comparison without showing the detailed numbers.

Using "all triples" (instead of "balanced triplets") yields significantly low accuracy (below 16.3%), primarily due to the inclusion of a substantial number of "easy triplets" within the entire set of "all triplets". These "easy triplets" have negatives considerably distant from the anchor (e.g., representing activities with highly distinct patterns compared to the anchor activity), with $D(z_a^r, z_n^s) \gg D(z_a^r, z_p^s)$. Thus the loss term $|m + D(z_a^r, z_p^s) - D(z_a^r, z_n^s)|_+$ easily becomes zero, contributing minimally to model training. Hard triplets have negatives closely resemble to the anchor activity as the positives, demanding the model to differentiate more carefully. The loss value is amortized to zero because of large amount of "easy triplets", burying the valuable information in "hard triplets". Thus, the significance of "hard triplets" is overshadowed by large amounts of "easy triplets", which number around 4.2 million in the experiment. As the training dataset size increases, this effect becomes more pronounced, undermining the model's ability to effectively learn from challenging instances.

Offline mining versus online mining. We compare online K-P batch mining and offline balanced-triplet mining to find which can better extract cross-view features from the hybrid training dataset. We compare two triplet mining strategies by training Triplet networks using loss function L_{BA-hyb}, which is appropriate to both online and offline mining. With offline mining, we achieve an accuracy of 87.0% (91.0% at seen views, 85.0% at unseen views) using SVM classifier and 86.0% (86.0% at seen views, 86.0% at unseen views) using MLP classifier. With online mining, we achieve an accuracy of 65.9% (84.9% at seen views, 27.9 at unseen views) using SVM classifier and 64.3% (82.6% at seen views, 27.9% at unseen views) using MLP classifier.

On average, with loss function L_{BA-hyb}, offline mining outperforms online mining on accuracy by 21.4%, showing better capability of utilizing limited real data in training. Specifically, we achieve an average accuracy of 85.5% between the two classifier at unseen views with offline mining, proving great cross-view generalizability, compared to online mining, of which the accuracy at unseen views is a mere 27.9%. This is because data at different views is not balanced in mini-batch during online batch construction, preventing the model from learning cross-view features. While with offline construction, data at different views is balanced in the whole constructed training dataset.

2) Loss functions: We compare different loss functions to find which can better extract view-independent features. Notably, the evaluation of loss functions is orthogonal to the choice of online/offline mining. Although the aforementioned findings demonstrate the superior performance of offline mining over online mining, the implementation of L_{BH-hyb} mandates online triple mining to select the "hardest" triplet when calculating the loss value. Therefore, to ensure fair comparison, we use online mining in the evaluation of all three loss functions.

Table I shows the comparison results. From the results, L_{BA-hyb} outperforms the other two loss functions with an improvement of 3.8% in accuracy. The two triplet loss functions (i.e. L_{BH-hyb} and L_{BA-hyb}) show better capability than the two-way contrastive loss function ($L_{SUP-hyb}$) on average, because triplet networks provide higher discriminative power while using intra-class and inter-class relations [17]. In addition, L_{BA-hyb} which uses all triplets outperforms L_{BH-hyb} which uses the "hardest" triplets only. This happens when the "hardest" triplets is noisy or outliers. Using only the "hardest" triplets introduces instability in the loss calculation and thus misleads training. This proves that using all triplets in the mini-batch is better with limited real data and a large amount of synthetic data.

C. Overall Comparison to Common Baselines

We compare the optimal design identified (e.g., real data as anchors, offline triplet mining, L_{our} loss function, and incorporating synthetic data at unseen views in training) against three common baselines. We use the same data setup as in Section V-B. We perform leave-one-view out experiments with different view combinations in training and show the average results over different view combinations for comprehensive verification.

We compare our method with three baselines:

- End-to-end SVM (abbr. e2e-SVM) classifier: We flatten the 2D raw Doppler data (28×52) into a 1D vector (1456 features) as input to train the end-to-end SVM classifier.
- end-to-end MLP classifier (abbr. e2e-MLP): We train the end-to-end MLP classifier using the same features used for SVM training. To keep the comparison fair, the MLP structure used in the baseline is the same as the MLP classifier used in the second step of our method.
- Fine-tune: Fine-tuning a neural network trained using synthetic data with some real labeled data is an effective solution under the "synthetic-to-real" topic [10]. We train a CNN classifier using synthetic data of $M+1$ views and fine-tune the last layer of the model using real data of M views. The CNN classifier has the same structure as the feature extractor in our metric learning model.

TABLE II: Accuracy comparison (in %) of our method versus baselines under different ratios of real to synthetic data size.

Ratio of real to synthetic	1.0	0.8	0.4
e2e-SVM	68.5	67.9	61.2
e2e-MLP	72.4	71.8	69.3
Fine-tune	71.8	65.1	60.7
M4X	**80.9**	**77.3**	**77.2**

We compare our methods with baselines under different ratios of real data to synthetic data size to explore how they perform with minimal real data in training. We use 80%, 64%, and 32% of the real data for training, achieving ratios of real to synthetic data sizes of 1.0, 0.8, and 0.4. From the results in Table II, our method consistently outperforms all baselines, irrespective of the ratios of real data. We achieve accuracy of 80.9%, 77.3%, and 77.2% with real data ratio of 1.0, 0.8, and 0.4, respectively. This translates into improvements of 7.5%-12.4%, 5.5%-12.2%, and 7.9%-16.5% compared to three baselines, with different real data ratios respectively. These results demonstrate that our method has higher cross-view capabilities compared to baselines even when using minimal real data in the training phase.

TABLE III: Accuracy (in %) at unseen views of our method versus baselines under different ratios of real to synthetic data size.

Ratio of real to synthetic	1.0	0.8	0.4
e2e-SVM	37.2	38.0	34.1
e2e-MLP	50.4	47.3	41.8
Fine-tune	62.8	46.5	39.5
M4X	**65.1**	**58.1**	**59.7**

We take a close look at the cross-view generalizability by comparing the accuracy at the unseen views. From Table III, when real to synthetic data ratio is 0.4, ***M4X*** outperforms three baselines by 17.9%-25.6%. Especially, when compared with e2e-MLP, we achieve 17.9% improvement. Since we use the same MLP classifier with e2e-MLP, the improvement attributes to the metric learning feature extractor, proving its capability of extracting cross-view features. In addition, we achieve 2.3%, 10.8% and 17.9% improvement at the unseen views compared to the best results in baselines with the ratio of real to synthetic data size of 1.0, 0.8, 0.4, respectively. These findings highlight that: 1) our method has significantly higher capability to generalize to unseen views. Because our metric learning model extracts view-independent features across real and synthetic data sources; 2) the resilience of our method to maintain cross-view performance even with minimal real data in training. Because our metric learning model can extract view-independent features efficiently when there is minimal real data and large amount of synthetic data in training.

D. Exploration of View Combinations

Given limited numbers of deployed sensors in real life, We also investigate which view combinations are optimal. While deploying sensors across multiple views can enhance performance, it also introduces significant costs and efforts for deployment and labeling. In practice, resource constraints, sensor expenses, and maintenance limitations may allow only very few deployed sensors. Therefore, finding the optimal view combinations can help achieve better performance despite limited deployment.

We employ data at view 2, 3, 4, 5, 6 for test. We exclude data at view 1 because useful information is mostly buried in noise at view 1 and it can harm the recognition, as illustrated in Figure 5(h). We consider the case where only a limited number of sensors (i.e. 1, 2, or 3 sensors) are available to deploy to get training data, and we use real data at these 5 views in test. We denote the view set by $V = \{v_1, v_2, v_3, v_4, v_5\}$. We denote the number of sensors with M, the subset of views to deploy sensors at is $VC = \{v_{i1}, v_{i2}, ..., v_{iM} | v_{ij} \in V\}$. For testing, 20% of data at 5 views' in real domain is used. For training, 1) 80% of synthetic data at 5 views' in synthetic domain is used. Because getting synthetic data does not require extra data collection and annotation effort, we can have synthetic data at all views; 2) 80% of real data at VC.

We show the results of different view combinations when $M = 2, 3$ in Table IV, V. The best results with comparable performance are highlighted in the table. When $M = 2$, in Table IV, the best two view combinations are $\{v_2, v_4\}$ and $\{v_4, v_6\}$. They achieve best performances not only at average accuracy, but also the accuracy at unseen views. This shows they have the most complementary information

TABLE IV: Accuracy of different view combinations in real domain ($M = 2$)

View Combination	Accuracy (%)		
	overall	at seen views	at unseen views
v2 + v3	63.9	90.3	51.2
v2 + v4	**77.0**	87.2	**68.6**
v2 + v5	65.4	86.0	48.6
v2 + v6	61.3	87.2	40.0
v3 + v4	47.1	95.2	24.0
v3 + v5	56.0	88.7	40.3
v3 + v6	59.2	88.7	45.0
v4 + v5	69.1	91.9	50.5
v4 + v6	**72.3**	88.4	**59.0**
v5 + v6	59.7	84.9	39.0

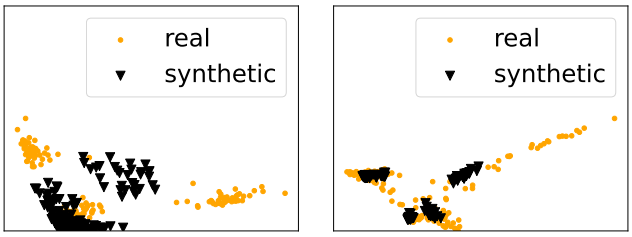
TABLE V: Accuracy of different view combinations in real domain ($M = 3$)

View Combination	Accuracy (%)		
	overall	at seen views	at unseen views
v4 + v5 + v6	**81.7**	88.4	**67.8**
v3 + v5 + v6	70.2	88.6	47.7
v3 + v4 + v6	77.5	89.5	62.8
v3 + v4 + v5	70.7	91.4	45.3
v2 + v5 + v6	70.2	86.0	37.1
v2 + v4 + v6	**82.2**	88.4	**69.4**
v2 + v4 + v5	80.6	90.7	59.7
v2 + v3 + v6	69.6	87.6	47.7
v2 + v3 + v5	72.3	88.6	52.3
v2 + v3 + v4	**81.7**	94.3	**66.3**

for cross-view generalizability. View 4 is the orientation the participant faces. View 2 and view 6 are two views more far apart from view 4 than view 3, 5; and they happen to be symmetrical about the participant's facing direction. They have the best performance because: 1) sensor at view 4 captures most velocity when the participant is facing v_4; 2) sensor at view 2 or view 6 can mostly capture the velocity perpendicular to v_4 without introducing too much noise as view 1. In Table V when $M = 3$, the best view combination is $\{v_2, v_4, v_6\}$, which is the union of the best two combinations ($\{v_2, v_4\}$, $\{v_4, v_6\}$) when $M = 2$. In addition, $\{v_4, v_5, v_6\}$ and $\{v_2, v_3, v_4\}$ achieve close accuracy (only 0.5% decrease) to the best one. These two combinations are also symmetrical about the participant's facing direction. From the observations, we draw a simple and practical guideline on view combinations of sensor deployment: 1) deploy a sensor at the participant's facing view for the maximum radial velocities; 2) deploy another sensor at some large angle (e.g., around 60 degrees to the left or right of the facing view for velocities tangential to the first one.

To better understand the reason, Figure 7 shows the Doppler space and feature space data representations (extracted from view combination v_2, v_3, v_4) at view 5. We find in Doppler space there exist clusters in real data without overlapping synthetic data, but in feature space the overlap is much more obvious except one long tail portion. This shows that our feature extractor can reduce domain gaps even the view is not seen in training.

We also quantitatively compare the domain gaps in different feature spaces trained with different view combinations. We quantify the domain gaps with mean Euclidean distance (ME) between real and synthetic features. Results show that the domain gaps in feature space at view 5 are smaller when $VC = \{v_2, v_3, v_4\}$ (ME = 0.096) compared to that when



(a) Doppler space (b) Feature space

Fig. 7: Doppler space and feature spaces at view 5. The feature space is trained using real data at view 2, 3, and 4. View 5 is excluded from the training views. In feature space, real and synthetic domain tend to have more similarity in distribution compared to Doppler space, indicating that our feature extractor can reduce domain gaps even the view is not seen in training.

$VC = \{v_2, v_3, v_6\}$ (ME = 0.126). It indicates that training view combination $\{v_2, v_3, v_4\}$ reduces domain gaps in feature space more than training view combination $\{v_2, v_3, v_6\}$. This explains why view combination $\{v_2, v_3, v_4\}$ outperforms $\{v_2, v_3, v_6\}$. The results are consistent across various view combinations — those combinations yielding higher cross-view performance typically exhibit reduced ME values in the feature space between real and synthetic sources.

VI. Discussion

Sensor Type. In this study, we use UWB sensors as the RF font-end in our system for sensing micro-Doppler signatures thus human activity recognition. It is worth noting that the concept of micro-Doppler signature is orthogonal to RF front-end. Moving forward, our research endeavors will extend to exploring alternative RF techniques, such as mmWave and FMCW, with the goal of optimizing trade-offs, including cost, power consumption, robustness, and sensing range.

Limited View Combinations. We utilize sensors at limited number of views, each separated by an angle of 30°. These sensors are strategically deployed to avoid the participant's rear view for less occlusion. Future expansions, contingent on resource availability, may include deploying additional sensors at more various viewpoints to enable more detailed observations and conclusive findings for sensor deployment.

Limited Activity Set. In this study, we exclusively focus on evaluating our design with five fundamental physical activities relevant to daily living. In future work, we plan to expand our evaluation to a wider range of activities and increase the number of data samples, particularly in real-world settings like people's homes. We will include more complex tasks such as food preparation, in order to better understand the potential and limitations of our method.

Metric Learning Design Choices. The current body of research in metric learning predominantly explores design choices specific to the computer vision domain [25]–[29]. However, there is a notable gap in the literature, and further investigation is required to identify the optimal approach specifically for scenarios involving RF data from diverse viewpoints and sources.

RF Data Synthesis. In this work, we follow the implementation of Vid2Doppler for generating RF data from videos. We have excluded the encoder-decoder model from Vid2Doppler in our study for two key reasons: 1) its performance is sub-optimal when applied to our dataset, which includes different sensors, environments, activities and so on from Vid2Doppler; 2) re-training the model demands a substantial volume of real data, which is unavailable under RF data scarcity. Future research could explore the use of diverse RF data synthesis methods within this paradigm to potentially enhance outcomes. We also plan to explore the impact of synthetic data size on the end-to-end performance.

Future Directions. In this work, we mainly work on algorithmic design to enhance cross-view generalizability for RF HAR under data scarcity. Despite achieving enhanced cross-view generalizability, the end-to-end accuracy in unseen views remains modest (60%–70%). To drive further improvement, it is important to engage the broader community in collaborative initiatives aimed at expanding data collection efforts to address the challenges associated with data scarcity. Recent advancements in home-based data collection infrastructure [12] [41] offer new possibilities to facilitate such collaborative initiatives, thereby augmenting the quantity and diversity of RF data and subsequently improving the training of RF HAR models for better generalizability. Recognizing the cost-intensive nature of data annotation, our future research will focus on investigating unsupervised learning methods to optimize the utilization of potentially extensive unlabeled datasets collected on a large scale.

VII. Conclusion

In this paper, we explore the cross-view generalizability in RF HAR under data-scarce scenario, in particular small amounts of real data yet possibly large amounts of synthesized data from videos. We firstly share our insights of the limitations of existing video-based RF data synthesis, which heavily relies on 3D pose estimation models with inherent constraints due to depth-ambiguity and occlusion problems. Together with imperfect simulation, they lead to domain gaps between real and synthetic data, limiting the benefits of utilizing synthetic data. Then we explore a method using metric learning to extract effective view-independent features from the more abundant synthetic data despite their domain gaps, thus enhancing cross-view generalizability. We systematically examine the design choices for metric learning, and identify offline triplet mining with real data as anchors, and a triplet loss without "hard negatives" as optimal strategies for contrastive pairs/triplets construction and loss function selection, respectively. Comprehensive experiments show that we consistently outperform baseline methods in cross-view generalizability. In the most challenging case of the least amount of real training data, our method outperforms three baselines by 7.9-16.5% on all views, and 18.9-25.6% on a view with only synthetic but no real data during training. This proves its effectiveness in extracting view-independent features from synthetic data despite their domain gaps. We also observe that given limited sensor deployments, a participant-facing viewpoint and another at a large angle (e.g. 60°) tend to produce much better performance. The cross-view enhancement technology developed and insights gained for

sensor deployment can be effectively utilized for recognizing activities during the routines of our daily life in home settings, offering significant benefits for health monitoring applications.

REFERENCES

[1] M. Straczkiewicz, P. James, and J.-P. Onnela, "A systematic review of smartphone-based human activity recognition methods for health research," *NPJ Digital Medicine*, vol. 4, no. 1, p. 148, 2021.

[2] D. X. Wang, J. Yao, Y. Zirek, E. M. Reijnierse, and A. B. Maier, "Muscle mass, strength, and physical performance predicting activities of daily living: a meta-analysis," *Journal of cachexia, sarcopenia and muscle*, vol. 11, no. 1, pp. 3–25, 2020.

[3] Y. Liu, G. Zhang, C. G. Tarolli, R. Hristov, S. Jensen-Roberts, E. M. Waddell, T. L. Myers, M. E. Pawlik, J. M. Soto, R. M. Wilson *et al.*, "Monitoring gait at home with radio waves in parkinson's disease: A marker of severity, progression, and medication response," *Science Translational Medicine*, vol. 14, no. 663, p. eadc9669, 2022.

[4] Y. Li, D. Zhang, J. Chen, J. Wan, D. Zhang, Y. Hu, Q. Sun, and Y. Chen, "Towards domain-independent and real-time gesture recognition using mmwave signal," *IEEE Transactions on Mobile Computing*, 2022.

[5] Y. Zhang, Y. Zheng, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Widar3. 0: Zero-effort cross-domain gesture recognition with wi-fi," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8671–8688, 2021.

[6] R. Song, D. Zhang, Z. Wu, C. Yu, C. Xie, S. Yang, Y. Hu, and Y. Chen, "Rf-url: unsupervised representation learning for rf sensing," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 282–295.

[7] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas *et al.*, "Towards environment independent device free human activity recognition," in *Proceedings of the 24th annual international conference on mobile computing and networking*, 2018, pp. 289–304.

[8] H. Xue, W. Jiang, C. Miao, F. Ma, S. Wang, Y. Yuan, S. Yao, A. Zhang, and L. Su, "Deepmv: Multi-view deep learning for device-free human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–26, 2020.

[9] A. Virmani and M. Shahzad, "Position and orientation agnostic gesture recognition using wifi," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 252–264.

[10] X. Zhang, Z. Li, and J. Zhang, "Synthesized millimeter-waves for human motion sensing," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 377–390.

[11] L. Zhang, D. Zheng, Z. Wu, M. Liu, M. Yuan, F. Han, and X.-Y. Li, "Poster: Cross labelling and learning unknown activities among multimodal sensing data," in *The 25th Annual International Conference on Mobile Computing and Networking*, 2019, pp. 1–3.

[12] M. Liu, M. Elbadry, Y. Hua, Z. Xie, and F. Ye, "Proteus: Towards a manageability-focused home-based health monitoring infrastructure," in *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2023, pp. 1–6.

[13] S. Ding, Z. Chen, T. Zheng, and J. Luo, "Rf-net: A unified meta-learning framework for rf-enabled one-shot human activity recognition," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 517–530.

[14] K. Ahuja, Y. Jiang, M. Goel, and C. Harrison, "Vid2doppler: Synthesizing doppler radar data from videos for training privacy-preserving activity recognition," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–10.

[15] K. Deng, D. Zhao, Q. Han, Z. Zhang, S. Wang, A. Zhou, and H. Ma, "Midas: Generating mmwave radar data from videos for training pervasive and privacy-preserving human sensing tasks," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 1–26, 2023.

[16] S. Zhang, C. Wang, W. Dong, and B. Fan, "A survey on depth ambiguity of 3d human pose estimation," *Applied Sciences*, vol. 12, no. 20, p. 10591, 2022.

[17] M. Kaya and H. Ş. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, no. 9, p. 1066, 2019.

[18] S. Z. Gurbuz, M. M. Rahman, E. Kurtoglu, T. Macks, and F. Fioranelli, "Cross-frequency training with adversarial learning for radar micro-doppler signature classification (rising researcher)," in *Radar Sensor Technology XXIV*, vol. 11408. SPIE, 2020, pp. 58–68.

[19] M. M. Rahman, S. Z. Gurbuz, and M. G. Amin, "Physics-aware generative adversarial networks for radar-based human activity recognition," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 3, pp. 2994–3008, 2023.

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

[21] S. Waqar and M. Pätzold, "A simulation-based framework for the design of human activity recognition systems using radar sensors," *IEEE Internet of Things Journal*, 2023.

[22] S. Bhalla, M. Goel, and R. Khurana, "Imu2doppler: Cross-modal domain adaptation for doppler-based activity recognition using imu data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–20, 2021.

[23] MediaWiKi, "Doppler effect - wikipedia," 2023. [Online]. Available: https://en.wikipedia.org/wiki/Doppler_effect

[24] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 798–21 809, 2020.

[25] M. Bucher, S. Herbin, and F. Jurie, "Hard negative mining for metric learning based zero-shot classification," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 524–531.

[26] B. Gajić, A. Amato, and C. Gatta, "Fast hard negative mining for deep metric learning," *Pattern Recognition*, vol. 112, p. 107795, 2021.

[27] J. F. Henriques, J. Carreira, R. Caseiro, and J. Batista, "Beyond hard negative mining: Efficient detector learning via block-circulant decomposition," in *proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2760–2767.

[28] M. Sikaroudi, B. Ghojogh, A. Safarpoor, F. Karray, M. Crowley, and H. R. Tizhoosh, "Offline versus online triplet mining based on extreme distances of histopathology patches," in *International Symposium on Visual Computing*. Springer, 2020, pp. 333–345.

[29] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[30] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

[31] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5253–5263.

[32] C. Xu, Y. Makihara, X. Li, and Y. Yagi, "Occlusion-aware human mesh model-based gait recognition," *IEEE transactions on information forensics and security*, vol. 18, pp. 1309–1321, 2023.

[33] Y. Yuan, U. Iqbal, P. Molchanov, K. Kitani, and J. Kautz, "Glamr: Global occlusion-aware human mesh recovery with dynamic cameras," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 038–11 049.

[34] B. O. Ayinmode and I. P. Farai, "Measurement and method in radiofrequency radiation exposure assessments," *The Pacific Journal of Science and Technology*, vol. 14, no. 2, pp. 110–118, 2013.

[35] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE access*, vol. 7, pp. 116 753–116 773, 2019.

[36] W. Deng, L. Zheng, Y. Sun, and J. Jiao, "Rethinking triplet loss for domain adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 29–37, 2020.

[37] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[38] C. Wu, F. Wu, and Y. Huang, "Rethinking infonce: How many negative samples do you need?" *arXiv preprint arXiv:2105.13003*, 2021.

[39] N. AS, "Discover uwb - novelda," 2023. [Online]. Available: https://novelda.com/

[40] Microsoft, "Kinect for windows - windows apps — microsoft learn," 2023. [Online]. Available: https://learn.microsoft.com/en-us/windows/apps/design/devices/kinect-for-windows

[41] M. Elbadry, M. Liu, Y. Hua, Z. Xie, and F. Ye, "Poster: Towards robust, extensible, and scalable home sensing data collection," in *Proceedings of the 8th ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, 2023, pp. 192–193.