Response to Review of Paper:

TMC-2024-07-1090: UMIMO: Universal Unsupervised Learning for mmWave Radar Sensing with MIMO Array Synthesis

Haoyu Zhang, Dongheng Zhang, Ruiyuan Song, Zhi Wu, Jinbo Chen, Liang Fang, Zhi Lu, Yang Hu, Hui Lin, and Yan Chen*, *Senior Member, IEEE*

We would like to thank all the reviewers and the editor for their thorough reviews and constructive comments. We have carefully revised the paper to clarify and address the reviewers' comments. In the document that follows, we describe changes made to the original version of the paper and address the comments of the reviewers.

Haoyu Zhang, Dongheng Zhang, Ruiyuan Song, Zhi Wu, Jinbo Chen, Liang Fang, Zhi Lu, Hui Lin, and Yan Chen are with School of Cyber Science and Technology, University of Science and Technology of China, Heifei 230026, China. (E-mail: haoyuzhang@mail.ustc.edu.cn, dongheng@ustc.edu.cn, {rysong, wzwyyx, jinbochen, fangliang}@mail.ustc.edu.cn, zhilu@ustc.edu.cn, linhui@whu.edu.cn, eecyan@ustc.edu.cn).

Yang Hu is with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230026, China (E-mail: eeyhu@ustc.edu.cn)

Corresponding author: Yan Chen (E-mail: eecyan@ustc.edu.cn)

I. Responses to Associate Editor

The authors would like to express sincere thanks to the associate editor for the thorough and constructive comments. In what follows, we present detailed comments in response to the individual points raised by the associate editor and elaborate on how the manuscript has been revised.

- **Comment 0.1:** *Overall the paper makes a relevant contribution. However, the reviewers point out several issues that should be addressed in a major revision.*

  Response 0.1: We thank the associate editor for the comment. We have revised the manuscript in response to the reviewers and elaborated on the revisions made to the manuscript.

II. Responses to Reviewer 1

The authors would like to sincerely thank the reviewer for the thorough and constructive comments. In what follows, we present detailed comments in response to the individual points raised by the reviewer.

- **Comment 1.1:** *This paper proposes a new method for contrastive unsupervised learning applied to RF signals. The approach of using different combinations of radar antennas to generate positive pairs of the same image is a very interesting idea, and it offers clear potential in the field. The paper is well-structured, the experimental results are convincing, and the contributions are clearly explained.*

  Response 1.1: We would like to express our sincere thanks to the reviewer for the recognition of our paper.

- **Comment 1.2:** *However, I have one question out of curiosity: You have increased the dataset by exploiting only the spatial dimension of the radar (i.e., the antenna combinations). Have you considered whether it would be possible to apply a similar strategy using other dimensions? For instance, could you combine different bands, waveforms, or temporal sampling (which could capture Doppler information)? If so, what do you think the potential impact on the results could be?*

  Response 1.2: We sincerely appreciate the reviewer's question about the potential impact of some similar strategies using other dimensions. Strategies involving different frequency bands, waveforms, or time sampling (to capture Doppler information) are indeed promising approaches. However, applying these strategies to a variety of scenarios and tasks may be somewhat difficult. Below we will discuss the possible impacts of these solutions.

  (i) Frequency Bands: Different frequency bands could provide complementary information due to variations in sensitivity to different materials and objects. Integrating data from different frequency bands could potentially enhance feature diversity in the learned representations. However, in practice, combining subbands can be challenging, particularly with FMCW radars, where frequency components are mixed and isolating individual subbands becomes difficult.

  (ii) Waveforms: Different radar waveforms offer additional signal characteristics, which can facilitate the extraction of more effective features. However, in most existing millimeter-wave radar sensing tasks, only a single waveform is typically employed, resulting in datasets that predominantly contain data associated with one specific waveform. This limitation poses challenges for applying such a multi-waveform approach to existing datasets. Additionally, incorporating multiple waveforms would increase system complexity and signal processing requirements.

(iii) Temporal Sampling (Doppler Information): Temporal sampling techniques that capture Doppler shifts can provide useful information for velocity-related features. Integrating Doppler information is particularly beneficial for tasks involving motion, such as human activity recognition or motion tracking. We believe this can enhance the temporal alignment performance in ECG monitoring tasks. Doppler information is typically derived from multiple chirps [1], requiring the use of a multi-chirp configuration during data collection. However, many datasets, including the contactless ECG monitoring dataset [2] and the HIBER dataset [3] used in our study, are based on single-chirp configurations and do not inherently support Doppler-based approaches.

In summary, these schemes offer valuable and insightful ideas for constructing positive samples from different perspectives, each contributing potentially useful information that could enhance the performance of contrastive learning. However, these methods are somewhat difficult to generalize across various downstream tasks. Recognizing that existing millimeter-wave radar sensing systems typically employ multi-transmitter and multi-receiver configurations, we propose to use antenna combinations to facilitate universal unsupervised learning for millimeter-wave radar. The UMIMO framework requires only a single commercial MIMO millimeter-wave radar and is compatible with the vast majority of existing millimeter-wave radar datasets, which is crucial for the approach's general applicability. Future work could explore the integration of these additional dimensions to further enhance performance.

III. Responses to Reviewer 2

The authors would like to express sincere thanks to the reviewer for the thorough and constructive comments. In what follows, we present detailed comments in response to the individual points raised by the reviewer and elaborate on how the manuscript has been revised.

- **Comment 2.1:** *The authors propose UMIMO, an unsupervised learning framework that combines the hardware nature of MIMO radar with deep learning techniques to address the issue of insufficient labeled data. Through experiments on tasks such as contactless ECG monitoring and 3D human pose estimation, the authors demonstrate that the proposed framework significantly enhances the performance of learning-based mmWave radar sensing in an unsupervised context. The manuscript is readable and it is technically sound. However, I believe that there are several things in the manuscript that need to be addressed.*

  Response 2.1: We thank the reviewer for the comments and suggestions. We have carefully revised the manuscript to address the reviewer's concerns, which will be elaborated as follows point-by-point.

- **Comment 2.2:** *The authors explained that the third (elevation) Tx antenna has a minor impact on azimuthal resolution and grating lobes in the ECG monitoring. However, they did not discuss the impact of this antenna on the other two tasks. Considering that Θ_{min} in the other two tasks is really small, the manuscript should include a detailed discussion on the impact of the elevation antenna.*

  Response 2.2: We sincerely appreciate the reviewer's feedback and acknowledge the need for a detailed discussion on the impact of the elevation antenna. As suggested by the reviewer, we have included a detailed analysis of the elevation antenna for two additional tasks in the revised manuscript.

  For the tasks of 3D pose estimation and human silhouette segmentation, we conduct experiments using the HIBER dataset [3]. This dataset includes data from the azimuth antennas of both horizontal and vertical radars. According to [3], the azimuth information is obtained by the azimuth antenna of the horizontal radar, while the elevation information is obtained by the azimuth antenna of the vertical radar. Consequently, analyzing the effect of the elevation antenna requires an analysis of the vertical radar.

  First, the Θ_{min} of the horizontal radar is also applicable to the vertical radar, as the analysis is conducted in three-dimensional space. For the GL_{min} of the vertical radar, it is determined by the subject's height and the minimum distance to the radar. Grating lobes should not appear within the range of human height. Since the HIBER dataset does not impose any restriction on the minimum distance between the subject and the radar, subjects may be positioned very close to the radar.

Therefore, grating lobes should not appear within the range of $[-90°, +90°]$, implying that the GL_{min} for the vertical radar is also $90°$. As a result, in the other two tasks, if a vertical radar is used, the corresponding parameters of the horizontal radar can be directly applied to the vertical radar.

The detailed analysis of the elevation antenna in the revised manuscript are provided as follows:

V.A.2 Θ_{min} and GL_{min} for 3D Pose Estimation

For the tasks of 3D pose estimation, we conduct experiments using the HIBER dataset [42]. This dataset includes data from the azimuth antennas of both horizontal and vertical radars. In [42], the azimuth information is obtained by the azimuth antenna of the horizontal radar, while the elevation information is obtained by the azimuth antenna of the vertical radar. Therefore, if both radars are used, the Θ_{min} and GL_{min} of the vertical radar will also need to be analyzed. Firstly, the Θ_{min} of horizontal radar is also applicable to the vertical radar, as the analysis is conducted in three-dimensional space. As for GL_{min} of vertical radar, it is determined by the subject's height and the closest distance from the radar. Grating lobes should not appear within the range of human height. Since the HIBER dataset does not impose any restriction on the minimum distance between the subject and the radar, subjects may be positioned very close to the radar. Therefore, there should be no grating lobes in the range of $[-90°, +90°]$, which means the GL_{min} for vertical radar is also equal to $90°$.

- **Comment 2.3:** *Since the HIBER dataset includes data points at different distances, the setup in Fig. 8b and the value $\Theta_{min} = 2.29°$ look very arbitrary. Clearly, at different target distances, especially those $\ll 5m$ and $\gg 5m$, Θ_{min} will be significantly different. I believe that authors should explore, or least discuss, the idea of time-varying Θ_{min}.*

Response 2.3: We thank the reviewer for the comment. We fully appreciate the concern raised about the potential for a time-varying Θ_{min} depending on target distances, particularly for distances much less than or greater than 5m.

To clarify, the value of Θ_{min} is indeed intended to be a fixed lower limit, not a time-varying parameter. This value is derived from the extreme case involving the longest possible distance (5m) between the human body and the radar, and the smallest joint distance (20cm) of the human body. Therefore, Θ_{min} serves as a definitive lower bound for the radar's resolution, ensuring effective differentiation between joints at the maximum distance in the dataset. In the subsequent filtering process, antenna arrays with a resolution greater than Θ_{min} will be excluded, as they would no longer be able to separate the closest joints in the scenario described above. Since the HIBER dataset was

collected with subjects positioned no further than 5m from the radar, the maximum distance for the dataset was set to 5m.

We apologize that this misunderstanding has been caused by the lack of a detailed explanation of the specific usage of Θ_{min} and GL_{min} in the original manuscript. We have elaborated on this aspect in the revised manuscript (see Sec. IV-A and V-A).

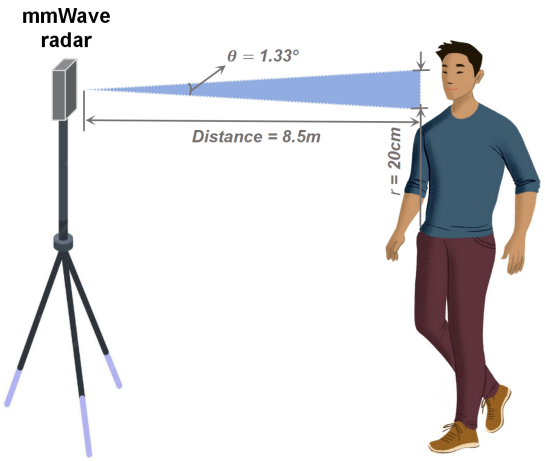

Fig. 1. The scene diagram of 3D human pose estimation.

- **Comment 2.4:** *Considering that the pose and silhouette detection are trained on the HIBER dataset, what are the limitations of UMIMO? How would the trained UMIMO system perform in a different setup, for example, in open space with larger distances. Please elaborate.*

Response 2.4: We would like to sincerely thank the reviewer for the thoughtful comment, which has helped us improve our manuscript and strengthen our work. To further explore the effect of UMIMO in different environmental settings, we collect an additional dataset in an open-space setup. This dataset includes three open-space scenes: two teaching building lobbies and a fully completed outdoor scene, encompassing a total of eleven subjects. The radar configuration and label collection process are the same as the HIBER dataset [3].

To evaluate the model's performance at greater distances and considering the angular resolution limit of the radar used (TI MMWCAS-RF-EVM FMCW radar), the maximum distance between the subject and the radar has been extended to 8m. As illustrated in Fig. 1, when all antennas are utilized, the radar achieves an angular resolution of $1.33°$ ($\theta = 2/N, N = 86$). Assuming that the closest joints on the human body are the nose and neck, which are approximately 20cm apart, we can calculate that the maximum distance at which these joints can still be distinguished is 8.5m. Therefore, to ensure

the ability to separate the joint, we set the maximum distance to 8m, slightly below the calculated threshold of 8.5m.

The dataset is divided into 30,121 training samples and 7,482 test samples. Table I shows the experimental results of the R3D-B model with 100%, 50% and 20% labeled training data. It can be seen from the results that the model pre-trained on the HIBER dataset can still improve the effect on open-space data. In the case of 100%, 50%, and 20% labeled training data, MPJPE is reduced by 7.4mm, 15.5mm and 25mm respectively. In the case of less training data, the effect of UMIMO becomes more significant. The experimental results show that even with data collected in different setups, UMIMO can enhance the performance of supervised fine-tuning, confirming that the UMIMO framework enables the backbone to extract more generalized features from RF signals. However, due to the transfer of data domains, the improvement observed is less significant compared to the results in Table II. This limitation can be addressed by using more diverse unlabeled pre-training data or by incorporating domain adaptation techniques.

The experimental results and analysis in the revised manuscript are provided as follows:

VII.C.6 Generalization to different setups

To explore the effect of UMIMO on different environmental settings, we collected an additional dataset in an open space setup. This dataset contains three open-space scenes, including two teaching building lobbies and a completed outdoor scene, with a total of eleven subjects. The radar configuration and label collection process are the same as the HIBER dataset. We divide this dataset into 30,121 training samples and 7,482 test samples. Table I shows the results of the R3D-B model with 100%, 50% and 20% labeled training data. It can be seen from the results that the model pre-trained on the Hiber dataset can still improve the effect on open space data. In the case of 100%, 50%, and 20% labeled training data, MPJPE is reduced by 7.4mm, 15.5mm, and 25mm respectively. In the case of less training data, the effect of UMIMO becomes more significant. The experimental results show that even with data collected in different setups, UMIMO can enhance the performance of supervised fine-tuning, confirming that the UMIMO framework enables the backbone to extract more generalized features from RF signals. However, due to the transfer of data domains, the improvement observed is less significant compared to the results in Table II. This limitation can be addressed by using more diverse unlabeled pre-training data or by incorporating domain adaptation techniques.

- **Comment 2.5:** *It is not clear how the authors chose the hyperparameters in Sec. V-B. Please explain the choice of these values, and possibly support it with references.*

TABLE I

MPJPE OF R3D-B MODEL ON OPEN SPACE DATASET.

Pre-training	100% labels	50% labels	20% labels
w/o	111.8	131.8	162.7
UMIMO	**104.4**	**116.3**	**137.7**

TABLE II

MPJPE OF THE R3D-B MODEL ON HIBER DATASET.

Pre-training	100% labels	50% labels	10% labels
w/o	78	105	164
UMIMO	**66**	**78**	**110**

Response 2.5: We sincerely thank the reviewer for the comments. The selection of these values was primarily guided by prior works. Based on [4], UMIMO employs an MLP with two linear layers as the projection head, interconnected by a ReLU activation function, with a hidden layer size of 256. Additionally, following [5], we chose SGD with a cosine decay schedule for optimization, with a learning rate of 0.003.

We apologize for not stating the source of the hyperparameters and have included the corresponding references in the revised manuscript:

V.B General Setup

Hyperparameters: UMIMO employs an MLP with two linear layers as the projection head, interconnected by a ReLU function, with a hidden layer size of 256 [11]. It uses SGD with a cosine decay schedule as the optimizer and a learning rate of 0.003 [8].

IV. Responses to Reviewer 3

The authors would like to express sincere thanks to the reviewer for the thorough and constructive comments. In what follows, we present detailed comments in response to the individual points raised by the reviewer and elaborate on how the manuscript has been revised.

- **Comment 3.1:** *Overall the paper is well written and in a very relevant topic. I like various parts of the design section and extensive evaluations.*

  Response 3.1: We thank the reviewer for the comments and suggestions. We have carefully revised the manuscript to address the reviewer's concerns, which will be elaborated as follows point-by-point.

- **Comment 3.2:** *What is universal about the proposed unsupervised learning approach? Is this that the proposed method can be applicable to various use cases? It is not clarified well.*

  Response 3.2: We sincerely thank the reviewer for raising this important point. The universality of UMIMO lies in its flexibility and broad applicability across various radar sensing tasks, without being limited by task type, hardware implementation, dataset, or other factors. Specifically:

  (i) Universal for a wide range of tasks: UMIMO eliminates the need for tailored signal processing pipelines for different tasks and can be applied to various MIMO radar hardware. As such, it can be easily adapted to applications of radar sensing in diverse fields, including healthcare and security systems. Our experiments demonstrate UMIMO's successful evaluation across three distinct tasks.

  (ii) Universal for different Radar Systems: UMIMO operates on raw radar data, leveraging antenna combinations to create signal views. This approach is independent of the specific radar hardware, which makes it universally applicable to any MIMO radar system with multiple transmitting and receiving antennas.

  (iii) Universal for existing datasets: Since most of the existing millimeter-wave radar sensing data is collected through MIMO configurations, UMIMO can directly use these datasets for pre-training without having to re-collect the dataset or having restrictions on the dataset.

  We apologize for not clarifying the concept of universality well in the original manuscript and we have included a more detailed clarification in the revised manuscript to provide readers with a clearer understanding of UMIMO's universal aspects as follow:

  I Introduction

  In this way, universal unsupervised learning of millimeter-wave radar signals can be achieved. The

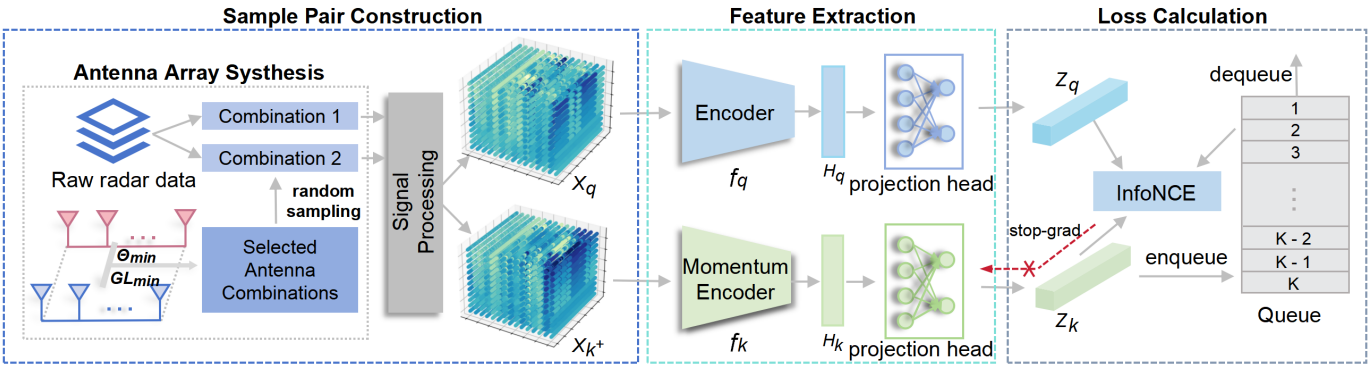

Fig. 2. An illustration of UMIMO framework for RF sensing.

universality of UMIMO is reflected in three aspects:

- **Universal for a wide range of tasks:** UMIMO eliminates the need for tailored signal processing pipelines for different tasks and can be applied to various MIMO radar hardware.
- **Universal for different radar systems:** UMIMO operates on raw radar data, utilizing antenna combinations to create signal views. This approach is independent of the specific radar hardware, which makes it universally applicable to any MIMO radar system with multi-transmitting and receiving antennas.
- **Universal for existing datasets:** Since most of the existing millimeter-wave radar sensing data is collected through MIMO configurations, UMIMO can directly use these datasets for pre-training.

- **Comment 3.3:** *Fig.5: Can you indicate how the output would contribute to a specific use case, suppose what the fig 5 would look like for the ECG monitoring use case.*

Response 3.3: We sincerely thank the reviewer for the valuable comment and apologize for not explaining clearly how to use Θ_{min} and GL_{min} in the original description. In the revised manuscript, we have included a detailed explanation and algorithm flow to demonstrate how to use Θ_{min} and GL_{min}. Additionally, we have revised the Antenna Array Synthesis part in Fig. 5 to make it easier to understand.

After obtaining Θ_{min} and GL_{min}, we use these two parameters to determine the available antenna combinations. Specifically, assuming that the radar used has M transmitting antennas and N receiving antennas. We define the following:

- $\mathbb{T}$: The set of transmitting antenna combinations corresponding to the M transmitting antennas. The number of antenna combinations, t_m, is the sum of all combination numbers of M antennas without selecting 0 elements, i.e., $t_m = \sum_{i=1}^{m} \binom{m}{i}$.

- $\mathbb{R}$: The set of receiving antenna combinations. The number of combinations, r_n, is similarly $r_n = \sum_{i=1}^{n} \binom{n}{i}$.

The set of all antenna combinations, $\mathbb{A}$, is the Cartesian product of $\mathbb{T}$ and $\mathbb{R}$, resulting in $a = t_m \times r_n$ combinations. The process of selecting available antenna combinations is outlined in Algorithm 1. We define the selected antenna combination set as $\mathbb{S}$. For each antenna combination in $\mathbb{A}$, check whether the angular resolution θ is less than Θ_{min} and whether generalized grating lobes (GL) is greater than GL_{min}. If both conditions are satisfied, add the combination to $\mathbb{S}$. Finally, the set of all available antenna combinations, $\mathbb{S}$, is obtained. As shown in Fig. 2, during the pre-training process, each training sample randomly selects two antenna combinations from $\mathbb{S}$ to form a positive sample pair, and then perform corresponding signal processing and training.

It is important to note that UMIMO requires the selection of antennas before signal processing. Unlike traditional methods, where the signal is processed first and then stored for later use during training, this approach does not allow for pre-processing the original signal and directly using the stored processed data. As a result, there is a certain time delay, which we will discuss in more detail in Comment & Response 3.8.

For the contactless ECG monitoring task, the device used (TI AWR1843 radar) has 2 transmitting antennas and 4 receiving antennas. There are 45 candidate antenna combinations in total, from which 13 antenna combinations are selected after filtering. Taking into account the elevation antenna, the number of available antenna combinations doubles to 26. For the 3D pose estimation and human silhouette generation tasks, the device used (TI MMWCAS-RF-EVM FMCW radar) has 9 transmitting antennas and 16 receiving antennas. There are 33,488,385 candidate antennas in total, and after screening, 43,732 combinations remain. Since the screening process only involves simple calculations and judgments, it is completed quickly.

The detailed explanation to demonstrate how to use Θ_{min} and GL_{min} in the revised manuscript are provided as follow:

IV.A Antenna Array Synthesis Strategy

After obtaining Θ_{min} and GL_{min}, we use these two parameters to determine the available antenna combinations. Specifically, assuming that the radar used has M transmitting antennas and N receiving antennas. We define the following:

- $\mathbb{T}$: The set of transmitting antenna combinations corresponding to the M transmitting antennas. The number of antenna combinations, t_m, is the sum of all combinations combinations of M antennas without selecting 0 elements, i.e., $t_m = \sum_{i=1}^{m} \binom{m}{i}$.

- $\mathbb{R}$: The set of receiving antenna combinations. The number of combinations, r_n, is similarly $r_n = \sum_{i=1}^{n} \binom{n}{i}$.

The set of all antenna combinations, $\mathbb{A}$, is the Cartesian product of $\mathbb{T}$ and $\mathbb{R}$, resulting in $a = t_m \times r_n$ combinations. The process of selecting available antenna combinations is outlined in Algorithm 1. We define the selected antenna combination set as $\mathbb{S}$. For each antenna combination in $\mathbb{A}$, check whether the angular resolution θ is less than Θ_{min} and whether generalized grating lobes (GL) is greater than GL_{min}. If both conditions are satisfied, add the combination to $\mathbb{S}$. Finally, the set of all available antenna combinations, $\mathbb{S}$, is obtained. During the pre-training process, each training sample randomly selects two antenna combinations from $\mathbb{S}$ to form a positive sample pair and then performs corresponding signal processing and training.

V.A.1 Θ_{min} and GL_{min} for Contactless ECG Monitoring

The TI AWR1843 radar has 2 transmitting antennas and 4 receiving antennas. So there are 45 ($\sum_{i=1}^{2} \binom{2}{i} \times \sum_{i=1}^{4} \binom{4}{i}$) candidate antenna combinations in total, from which 13 antenna combinations are selected after filtering by Algorithm 1. Taking into account the elevation antenna, the number of available antenna combinations doubles to 26.

V.A.2 Θ_{min} and GL_{min} for 3D Pose Estimation

The TI MMWCAS-RF-EVM FMCW radar has 9 transmitting antennas and 16 receiving antennas. There are 33,488,385 ($\sum_{i=1}^{9} \binom{9}{i} \times \sum_{i=1}^{16} \binom{16}{i}$) candidate antennas in total, and after filtering by Algorithm 1, 43,732 combinations remain. Since the filtering process only involves simple calculations and judgments, it will be completed quickly.

Algorithm 1 Available antenna combination selecting.

Require:

The set of transmitting antenna combinations, $\mathbb{T}$;

The set of receiving antenna combinations, $\mathbb{R}$;

The set of selected antenna combinations, $\mathbb{S}$;

Θ_{min}, GL_{min};

Ensure:

for each $t \in \mathbb{T}$ **do**
  for each $r \in \mathbb{R}$ **do**
    $\theta = Calculate\theta(t, r)$;
    $GL = CalculateGL(t, r)$;
    if $\theta < \Theta_{min}$ and $GL > GL_{min}$ **then**
      $S \leftarrow S \cup \{(t, r)\}$
    end if
  end for
end for

- **Comment 3.4:** *What do you mean by the positive sample wrt pose estimation, at this stage (Sec. IV.A) of discussion. It is advised to provide more details about the positive samples of the contrastive learning at the beginning of the paper.*

Response 3.4: We sincerely thank the reviewer for this insightful comment and for pointing out the need for further clarification. In the context of the pose estimation task, positive samples refer to different representations of the subject's pose, captured from various antenna subarrays. In contrastive learning, the goal is to train the network to extract effective features by minimizing the distance between positive samples while maximizing the distance between negative samples. Positive samples can be understood as different representations of a sample. For example, in computer vision, positive samples could be two cropped versions (v_1 and v_2) of the same RGB image. v_1 and v_2 are two different images, but the two images are similar and have the same main components. The design of positive samples is a critical factor since the backbone learns the key information in the sample by learning the components that are similar between the positive samples and different from the negative samples. However, there has been work [6] that proves that methods for constructing positive samples in computer vision such as cropping and color jittering are not suitable for RF data. This is why we propose to use antenna array synthesis to construct positive sample pairs of RF signals and use

two parameters Θ_{min} and GL_{min} to ensure that the positive samples contain information related to downstream tasks.

In Sec. IV.A, we designed several antenna arrays to illustrate that not all antenna combinations can be used to construct positive samples since some combinations will cause the samples to lose information related to downstream tasks, which will be detrimental to contrastive learning [7].

As the reviewer suggested, we have included a more detailed explanation of the positive sample of contrastive learning at the beginning of the revised manuscript:

I Introduction

This method focuses on repulsing different images (regarded as negative samples) while attracting diverse views of the same image (regarded as positive samples). Although positive samples are strictly different images, they contain common critical information in the original images. The key to contrastive learning lies in constructing diverse positive samples for the same sample to extract consistent information, which has been particularly attracting in the field of computer vision. For instance, simple data augmentations such as cropping, color jittering, Gaussian blur, and others can be applied to a single image, thereby generating valuable positive samples [10].

II Related Work

Unsupervised Contrastive Learning: With the continuous development of deep learning, unsupervised learning has gained significant attention, and within it, contrastive learning has become highly popular for effective representation learning. The core idea of contrastive learning is to attract positive sample pairs while repelling negative ones. The positive sample pairs can be understood as different representations of a sample. For example, in computer vision, positive samples could be two cropped versions (v_1 and v_2) of the same RGB image. v_1 and v_2 are two different images, but the two images are similar and have the same main components. The design of positive samples is a critical factor since it learns the key information in the sample by learning the components that are similar between the positive sample pairs and different from the negative samples.

- **Comment 3.5:** *What is the goal of this use case? what to expect if UMIMO was not invented, and how UMIMO is offering more?*

Response 3.5: We sincerely thank the reviewer for the comment. UMIMO is committed to providing a universal unsupervised learning framework that can be easily applied to various millimeter-wave radar sensing tasks without the need to design complex processing pipelines, thereby improving the

supervised learning performance of downstream sensing tasks. UMIMO was proposed primarily to address the following challenges:

(i) Supervised training is constrained by the amount of labeled data: Labeled data for radar signals is difficult to collect due to the fact that RF signals are not directly interpretable by humans. When labeled samples are scarce, supervised training suffers from performance degradation due to insufficient data. Existing methods struggle to learn robust representations when labeled data is limited.

(ii) Task-specific millimeter-wave unsupervised learning is difficult to generalize: Existing unsupervised frameworks for millimeter-wave radar (e.g., RF-URL [8]) rely on specific signal processing pipelines to generate positive sample pairs, which is not applicable to all applications. For example, RF-URL requires two signal processing pipelines, AoA-ToF and DFS, for gesture recognition. However, using the two pipelines to process signals becomes challenging in applications such as vital signs estimation. These methods are not universally applicable across all tasks due to their reliance on task-specific pipelines, which are complex to design and vary from task to task.

UMIMO solves the above problems through the following points:

(i) Universal unsupervised learning leveraging MIMO hardware features: UMIMO utilizes the inherent features of MIMO radar to construct positive samples for contrastive learning by synthesizing different antenna arrays. Unlike traditional methods, UMIMO works directly with raw radar data, eliminating the need to manually design task-specific signal representations. Commercial millimeter-wave radars currently used for sensing have multi-transmitter, multi-receiver antenna configurations, which allows UMIMO to be seamlessly used for a variety of millimeter-wave radar sensing tasks.

(ii) Improving performance with limited training samples through unsupervised learning: UMIMO pre-trains the backbone using unlabeled data through contrastive learning, enabling it to extract more generalized features. This approach enhances the performance of downstream tasks, particularly when labeled data is scarce.

- **Comment 3.6:** *How the Θ_{min}, GL_{min} of Sec. V are being used in Sec. VI, VII and VIII.*

Response 3.6: As the response to Comment 3.3, Θ_{min} and GL_{min} are used to select the available antenna combinations. We apologize for not explaining clearly how to use Θ_{min} and GL_{min} in the original description and have explained in detail how these two parameters are used in the revised manuscript.

TABLE III

EVALUATING THE EXTRACTED EMBEDDINGS OF PRE-TRAINING BY FREEZING R3D-B BACKBONE.

Approach	w/o pre-training	UMIMO	Extra time consumed
Time (min)	14.75	17.62	2.87 (19.5%)

- **Comment 3.7:** *Table VIII: Marginal performance improvement.*

  Response 3.7: We sincerely thank the reviewer for the thoughtful comments. The performance improvement of the human silhouette generation task is indeed small in the case of 100% and 50% labeled data. This is because, unlike the 3D pose estimation task, which requires regressing each 3D coordinate point, the silhouette generation task is a dense binary classification task, making it relatively simpler. 100% or 50% labeled data is already sufficient for good generalization. However, with only 10% labeled data, the performance of UMIMO remains above 0.6, while the results without pre-training drop to around 0.4 – 0.5. This demonstrates that UMIMO can effectively extract more generalized features, enabling it to maintain good performance with a reduced amount of labeled training data.

- **Comment 3.8:** *What is the trade-off or computation complexity of such a process?*

  Response 3.8: We sincerely thank the reviewer for raising this important question. Since UMIMO needs to randomly select antenna combinations before signal processing during pre-training, it cannot process and store all samples in advance like traditional methods and use them directly during pre-training. Therefore, UMIMO needs to perform signal processing during pre-training, which will result in some time delay for certain tasks. We analyze the time delay caused by UMIMO to three downstream tasks.

  For contactless ECG monitoring, the signal processing for this task involves only simple FFT and differentiation, which can be completed before the next training batch begins. Therefore, the time delay introduced by UMIMO is negligible and it does not impact the training process for this task.

  For the tasks of 3D pose estimation and human silhouette generation, we use the AoA-ToF [9] algorithm to process samples. Since this algorithm involves substantial computation, it cannot be completed quickly, leading to a training time delay. However, the signal processing of RF signals

primarily involves matrix addition and multiplication, which are highly parallelizable operations. By utilizing a GPU, we can significantly accelerate this processing. We transfer the raw data samples to the GPU and perform the signal processing on it. The results in Table III show the time consumption for one epoch of pre-training of the R3D-B model on hardware equipped with 4 × NVIDIA GeForce RTX 4090 GPUs (24GB). It can be seen that UMIMO increased the training time per epoch by 19.5%. Nevertheless, this increase in time is warranted by the versatility and effectiveness of the method, making it a reasonable trade-off.

We have included the above analysis of time trade-off in the revised manuscript:

VII.C.7 Time trade-off

Since UMIMO needs to randomly select antenna combinations before signal processing during pre-training, it cannot process and store all samples in advance like traditional methods and use them directly. Therefore, compared with traditional methods, UMIMO needs to perform signal processing during pre-training, which will cause a certain time delay. In contactless ECG monitoring, the signal processing involves only simple FFT and differentiation, the time of signal processing can be ignored. However, the AoA-ToF processing algorithm involves substantial computation, so it will bring a certain time delay. Fortunately, since it mainly involves matrix multiplication and addition operations, the GPU can be used to greatly accelerate the processing flow. We transfer the raw data samples to the GPU for processing. The results in Table III show the time consumption for one epoch of pre-training of the R3D-B model on hardware with 4× NVIDIA GeForce RTX 4090 GPUs (24GB). It can be seen that our method increased the training time per epoch by 19.5%. Nevertheless, this increase in time is warranted by the versatility and effectiveness of the method, making it a reasonable trade-off.

- **Comment 3.9:** *Conclusions: How is this applicable to the wide range of mmWAve radars? Seems a random claim.*

Response 3.9: We sincerely thank the reviewer for pointing out the need for further clarification. We apologize that the claim regarding the applicability of UMIMO to a wide range of mmWave radars seemed unclear. Since UMIMO is based on the multi-transmit and multi-receive hardware characteristics of mmWave radar, it can be seamlessly used on other tasks and existing datasets without being restricted by hardware models, datasets, etc.

As suggested by the reviewer, we have better clarified how UMIMO is applicable to various mmWave radar sensing tasks in the conclusion:

IX Conclusion

This paper proposed UMIMO, a novel unsupervised learning framework for mmWave radar sensing. UMIMO integrates the hardware configuration of multi-transmitting, multi-receiving antennas in radar with deep learning techniques, using data from various synthesized antenna arrays to construct positive samples for contrastive learning. The proposed two constraint parameters ensure the effectiveness and universality of UMIMO. Since UMIMO does not require designing the appropriate signal processing pipelines for each task and is applicable to a wide range of MIMO radar hardware, it can be seamlessly used on various radar sensing tasks and existing datasets, pushing the boundary of unsupervised learning in wireless signal processing further.

REFERENCES

[1] X. Li, X. Wang, Q. Yang, and S. Fu, "Signal processing for tdm mimo fmcw millimeter-wave radar sensors," *IEEE Access*, vol. 9, pp. 167 959–167 971, 2021.

[2] J. Chen, D. Zhang, Z. Wu, F. Zhou, Q. Sun, and Y. Chen, "Contactless electrocardiogram monitoring with millimeter wave radar," *IEEE Transactions on Mobile Computing*, 2022.

[3] Z. Wu, D. Zhang, C. Xie, C. Yu, J. Chen, Y. Hu, and Y. Chen, "Rfmask: A simple baseline for human silhouette segmentation with radio signals," *IEEE Transactions on Multimedia*, pp. 1–12, 2022.

[4] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[6] T. Li, L. Fan, Y. Yuan, and D. Katabi, "Unsupervised learning for human sensing using radio signals," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3288–3297.

[7] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" *Advances in neural information processing systems*, vol. 33, pp. 6827–6839, 2020.

[8] 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.

[9] D. Zhang, Y. Hu, Y. Chen, and B. Zeng, "Breathtrack: Tracking indoor human breath status via commodity wifi," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3899–3911, 2019.