# VirtualIMU: Generating Virtual Wearable Inertial Data from Video for Deep Learning Applications

Ignacio Gavier
*Manning College of Information and Computer Science*
*University of Massachusetts*
Amherst, USA
igavier@umass.edu

Yunda Liu
*Manning College of Information and Computer Science*
*University of Massachusetts*
Amherst, USA
yundaliu@umass.edu

Sunghoon Ivan Lee
*Manning College of Information and Computer Science*
*University of Massachusetts*
Amherst, USA
sunghoonlee@umass.edu

2023 IEEE 19th International Conference on Body Sensor Networks (BSN) | 979-8-3503-3841-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/BSN58485.2023.10331242

***Abstract*—In the era of deep learning, accessibility to a large amount of wearable Inertial Measurement Unit (IMU) data plays a crucial role in various biomedical and health applications. However, collection of 'big' IMU data is extremely challenging due to its cost and time requirements. To address this, researchers have explored using publicly available videos, such as those on YouTube, to extract human skeletal models and synthesize IMU data. However, existing methods for converting skeletal models to virtual IMU data are oversimplified and lack systematic data augmentation capabilities. In this study, we propose a systematic approach to synthesize realistic and diverse IMU data, including three-axis accelerometer and gyroscope measurements, from video-based skeleton representations. Through experiments involving seven healthy individuals, we demonstrate that our method can accurately synthesize accelerometer and gyroscope data with a normalized root mean square error of $14.4$ % and $16.0$ %, respectively. Furthermore, we qualitatively evaluate the algorithm's ability to generate a large volume of diverse IMU data. Our findings affirm the potential of obtaining diverse synthetic IMU data from videos, offering a promising solution to reduce the costs associated with collecting IMU data in deep learning-based applications.**

***Index Terms*—Human activity recognition, wearables accelerometer, wearable gyroscope, data augmentation.**

## I. Introduction

Wearable Inertial Measurement Units (IMUs) have emerged as a valuable tool in the domain of biomedicine and health [1]. These devices offer a wide range of applications, including monitoring human activities and tracking behavioral phenotypes in individuals with motor impairments [2]. However, in the era of deep learning, the field of wearable computing has encountered obstacles in its advancement primarily due to limited access to large amounts of data [3], which arises from the expensive and time-consuming nature of data collection.

To counteract the challenges of data collection, prior studies have explored leveraging publicly available videos, such as those on YouTube, to extract human skeletal models and synthesize IMU data [3], [4]. However, the off-the-shelf pose extractor algorithms used for extracting skeleton representations from videos do not provide explicit information about the position and orientation of the virtual sensor node. As a result, the augmentation of IMU data based on these representations is oversimplified and uniform, which does not effectively scale to support deep learning-based applications.

In this paper, we introduce a novel algorithm capable of synthesizing realistic and diverse IMU data, consisting of a three-axis accelerometer and a three-axis gyroscope, from videos containing human motion. To the best of our knowledge, this study is the first attempt to augment wearable gyroscope data. To accomplish this, we offer several key contributions. Firstly, we tackle the challenge of geometric alignment between the IMU sensors and the skeletal representation extracted from the videos. This alignment is essential to ensure the synthesis of *realistic* IMU data from video-based skeletal motion. Furthermore, we present augmentation techniques to enhance the *diversity* of the synthetic IMU data. These techniques systematically introduce variations in sensor placement and orientation, thereby expanding the range of simulated scenarios and generating a more comprehensive, realistic representation of IMU data. One significant advantage of our method lies in its fully differentiable pipeline. This end-to-end differentiability enables efficient training and optimization of downstream tasks using gradient-based methods. We believe that the proposed research herein addresses the challenges posed by data availability in human movement research, particularly in the context of deep learning by offering a comprehensive solution to generate high-quality IMU datatsets. Our code can be found at https://github.com/ignaciogavier/VirtualIMU.

## II. Background

The problem of converting video to IMU data can be described in two steps:

1) Conversion of a sequence of two-dimensional images $\mathbf{V}(t) \in \mathbb{R}^{H\times W\times 3}$ into a sequence of a skeletal representation of the human subject under analysis, denoted as $\mathbf{S}(t) \in \mathbb{R}^{J\times 3}$. Here, $(H, W, 3)$ is the size of RGB images, and $J$ is the number of 3D anatomical landmarks used in the skeletal representation.
2) Transform the skeletal representation into synthetic measures of IMU sensors, denoted as $\mathbf{I}(t) = [\mathbf{a}(t), \boldsymbol{\omega}(t)] \in \mathbb{R}^{D\times 6}$. Here, $D$ is the number of IMU sensor nodes positioned on the body, each consisting of a three-axis accelerometer and a three-axis gyroscope.

The first step, which involves extracting the human skeletal model from video, has been extensively studied in the field of

deep learning [5], [6]. However, the accuracy of the resulting skeletal representations could be compromised by various factors, including noise present in the input video, inherent errors in the pose estimation algorithms, model inaccuracies due to occlusions or complex poses, and errors in skeletal tracking across frames [7]. These factors introduce variability and distortions in the skeletal representation, which in turn can lead to inconsistencies and inaccuracies in the subsequent conversion to IMU data.

The second step, which involves synthesizing IMU data from the skeletal model, can be performed based on a geometric transformation of the position $\mathbf{p}(t)$ and orientation $\mathbf{R}(t)$ of virtual sensors within a fixed, arbitrary global coordinate system. The transformation can be expressed as:

$$\begin{cases} \mathbf{a}(t) & = \mathbf{R}^T(t)\left(\frac{d^2}{dt^2}\mathbf{p}(t) + \mathbf{g}\right) \\ \boldsymbol{\omega}(t) & = \times^{-1}\left[\mathbf{R}^T(t)\frac{d}{dt}\mathbf{R}(t)\right] \end{cases}, \tag{1}$$

where $\mathbf{g}$ is the gravity vector, and $\times^{-1}[\cdot]$ is the inverse of the skew-symmetric operator [8]. $\mathbf{a}(t)$ and $\boldsymbol{\omega}(t)$ represent the synthesized accelerometer and gyroscope data within the sensor's local coordinate system, respectively.

Accurate synthesis of IMU data requires precise determination of the position and orientation of virtual sensor units relative to the skeletal representation of the subject. In the widely-adopted conversion method proposed by Young *et al.* [9], the position $\mathbf{p}(t)$ is identified based on the location of a joint, while the orientation $\mathbf{R}(t)$ is determined based on the coordinates of predecessor joints. When considering a virtual sensor on the wrist, its position is assumed to be identical to the center of mass of the wrist joint, while its orientation is computed by linking the positions of the shoulder, elbow, and wrist. However, this inference process synthesizes only a subset of possible IMU data as it overlooks the potential variation in the sensor position and orientation on the skin. In reality, wearable sensors are located on the surface of the skin with varying orientations, and individuals have varying skin depths, leading to significant variability in IMU behaviors (see Fig. 1 for an example).

## III. Methodology

In this section, we present our approach to augment a large volume of realistic IMU data from a skeletal model. Specifically, we capitalize on the aforementioned sources of variations in IMU signal behaviors: the imperfect position and orientation of the virtual sensor on the skeleton model.

### A. Pre-processing of the Skeletal Model

The high-frequency noise introduced by the pose extractor from video footage often exhibits non-linear characteristics. For instance, the noise in skeletal movement becomes more pronounced during intense or vigorous movements by the subject. Therefore, conventional low-pass filters (e.g., Butterworth) that attenuate frequencies above a specific threshold are not well-suited for this type of noise. In this paper, we applied the adaptive median filter [10] to the skeletal model, which offers non-linear filtering capabilities that adapt to the intensity of movement. The resulting filtered skeletal coordinates are denoted as $\mathbf{S}(t)$.

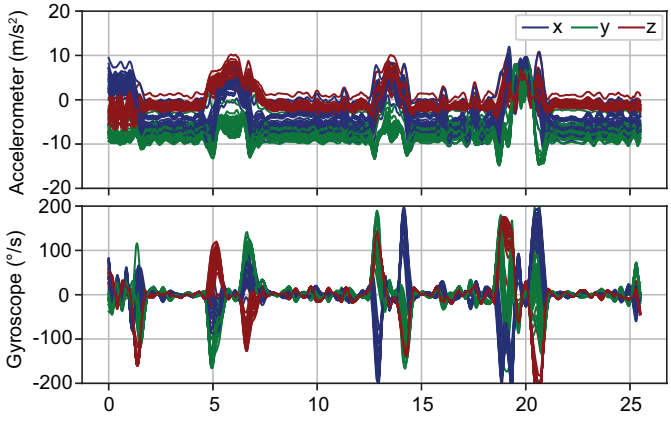


Fig. 1. Synthetic IMU sensor data generated using Monte Carlo augmentation.

### B. Alignment of the Virtual Sensor on the Skeletal Model

The position and orientation of the virtual sensor unit on the skeleton affect both the acceleration and the gyroscope measurements, as described in (1), because gravity affects each local coordinate axis in a disparate and time-varying manner. To that end, we present a systematic approach to augment $\mathbf{p}(t)$ and $\mathbf{R}(t)$ based on the sensor placement protocol, as described below, based on the study by Jiang *et al.* [11]. This section uses a virtual sensor on the left wrist as an illustrative example, without loss of generality.

Let $\mathbf{S}_f(t)$ and $\mathbf{S}_k(t)$ represent the forearm and metacarpophalangeal joint (knuckles) position vectors, respectively, which can be determined based on the position vectors of the wrist, elbow, and the $2^{\text{nd}}$ and $5^{\text{th}}$ knuckles: $\mathbf{S}_f(t) = \mathbf{S}_w(t) - \mathbf{S}_e(t)$ and $\mathbf{S}_k(t) = \mathbf{S}_{k_2}(t) - \mathbf{S}_{k_5}(t)$. Considering $\mathbf{S}_k^{\perp f}(t)$ as the component of $\mathbf{S}_k(t)$ orthogonal to $\mathbf{S}_f(t)$, the position and orientation of the left wrist sensor can then be obtained as follows:

$$\begin{cases} \mathbf{p}(t) & = \mathbf{S}_w(t) + \mathbf{R}(t)\mathbf{p}_0 \\ \mathbf{R}(t) & = \left[\frac{\mathbf{S}_f(t)}{||\mathbf{S}_f(t)||}, \frac{\mathbf{S}_k^{\perp f}(t)}{||\mathbf{S}_k^{\perp f}(t)||}, \frac{\mathbf{S}_f(t)\times\mathbf{S}_k^{\perp f}(t)}{||\mathbf{S}_f(t)\times\mathbf{S}_k^{\perp f}(t)||}\right]\mathbf{R}_0 \end{cases}, \tag{2}$$

where $\mathbf{p}_0$ accounts for the offset of the sensor position from the skeletal landmark (e.g., skin depth), and $\mathbf{R}_0$ accounts for the 3D orientation of the sensor placement.

### C. Augmentation of IMU Data using Monte Carlo Simulation

While $\mathbf{p}_0$ and $\mathbf{R}_0$ can be approximated using the population statistics [12], we exploit the potential diversity in human body morphology and variations in sensor placements to augment a wide range of realistic IMU data. To achieve this, we employ Monte Carlo simulation [13] to introduce variability in $\mathbf{p}_0$ and $\mathbf{R}_0$. This augmentation method enables us to generate and diversify the possible IMU data from a single video, thereby enhancing the robustness and generalizability of downstream deep learning models.

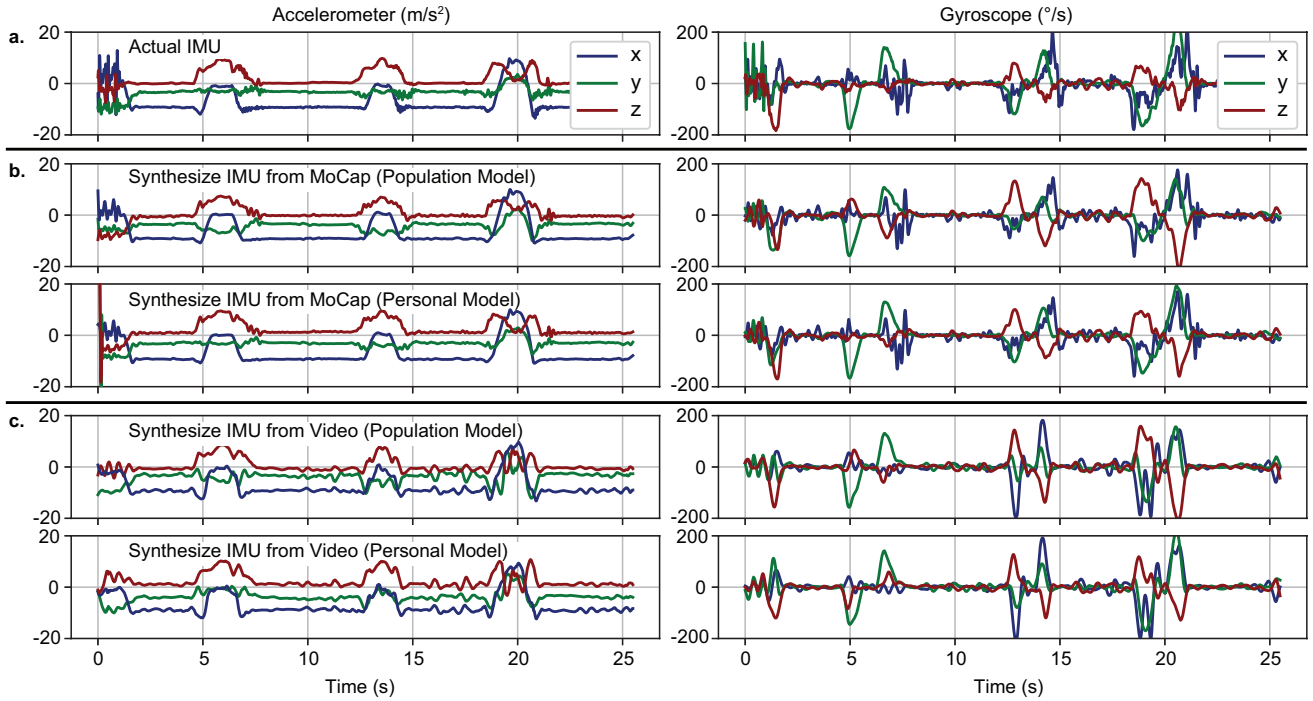


Fig. 2. Qualitative comparison between real IMU data (a), synthesized IMU from MoCap data (b) and synthesized IMU data from Video (c). For the synthesized data from MoCap and Video, both the mean population and personalized models are shown.

The placement offset along the forearm can be described as $\mathbf{p}_0 = [-p_f, 0, p_w]$, where $p_f$ represents the position along the axis of the forearm, and $p_w$ denotes the skin depth. The orientation of the virtual sensor unit on the forearm can be described as $\mathbf{R}_0 = \text{Eu}_{YXZ}(-\alpha_f, \alpha_w, 0)^T$, where $\alpha_f$ corresponds to the rotation along the forearm axis, and $\alpha_w$ represents the shape of the forearm (i.e., the slope that reflects the change in the radius of the muscles along the forearm axis). The Monte Carlo simulation is performed by sampling the input variables from a normal distribution, taking into account the statistical properties of left arm geometry [12]:

$$\begin{cases} p_f \sim \mathcal{N}(75\ mm, 4\ mm^2) \\ p_w \sim \mathcal{N}(43\ mm, 9\ mm^2) \\ \alpha_f \sim \mathcal{N}(0.8\ rad, 0.05\ rad^2) \\ \alpha_w \sim \mathcal{N}(0.6\ rad, 0.04\ rad^2) \end{cases} \tag{3}$$

Fig. 1 shows samples of 30 IMU data augmented using Monte Carlo simulation. It is noticeable that the resulting time-series predominantly align with their modal value, indicating a common trend. The introduced variability through this augmentation approach is more realistic compared to the simple addition of additive noise, such as Gaussian noise, which is frequently employed for signal augmentation [14]. This distinction is crucial as deep learning models downstream can readily abstract and disregard such additive noise [15].

## IV. Experimental Setup

Data were collected and analyzed from seven healthy subjects ($28.1 \pm 5.7$ years old, 3 males). Participants were instrumented with a six-axis IMU (Shimmer3, Shimmer) on the left wrist and five reflective markers for a motion capture system (Miqus, Qualysis). The markers were placed on the medial and lateral wrist, medial and lateral elbow, and the IMU. During the study, participants performed four different motor tasks involving different postures, including shoulder flexion, shoulder hyperflexion, and two motor tasks focused on object repositioning. These activities were recorded using the motion capture system and a static front-view camera (GoPro). The camera images were processed using MediaPipe [6] to extract the human skeletal model. It is worth noting that our IMU synthesis and augmentation method can be extended to the skeletal model captured using motion capture systems. We opted to apply our synthesis method to the skeletal model generated by the motion capture system to better understand the impact of the inherent errors in the pose estimation algorithm (i.e., MediaPipe). For the motion capture skeletal model, we estimated the $\mathbf{S}_k(t)$ in (2) using the medial and lateral markers on the wrist. In summary, we generated three different types of data from the experiment: 1) ground truth inertial data captured by a wrist-worn IMU, 2) synthetic inertial data generated based on video, and 3) synthetic inertial data generated based on the motion capture system.

## V. Results

Fig. 2 provides a qualitative comparison of the virtual IMU data generated using two different models that set the model parameters $\mathbf{p}_0$ and $\mathbf{R}_0$. The first model, namely the *population model*, determined $\mathbf{p}_0$ and $\mathbf{R}_0$ using the population statistics (i.e., generic values representing the average person) [12].

The second model, namely the *personal model*, determined the values of $\mathbf{p}_0$ and $\mathbf{R}_0$ that empirically minimize the errors between the synthetic vs. actual IMU data. To achieve this, we employ the gradient descent algorithm, leveraging the fully differentiable pipeline. This model can represent the specific data collection setup, including the individual's physical characteristics, sensor placement, and orientation. While the proposed method has the capability to augment data beyond these two models by sampling $\mathbf{p}_0$ and $\mathbf{R}_0$ from their respective distributions, we utilize these two time-series examples to showcase the effectiveness and accuracy of the proposed augmentation method. Fig. 2a represents the actual IMU data obtained from the wrist-worn IMU, while Fig. 2b and c display the synthetic data generated using the *population model* and *personal model*, respectively.

TABLE I
RMS AND NORMALIZED RMS ERROR IN ACCELEROMETER AND GYROSCOPE OF IMU DATA SYNTHESIZED FROM MOCAP AND VIDEO WITH TWO DIFFERENT FILTER TECHNIQUES, TAKING AS REFERENCE THE ACTUAL IMU DATA.

| | | Population Model | | Personal Model | |
|---|---|---|---|---|---|
| | | $\mathbf{a}$ | $\boldsymbol{\omega}$ | $\mathbf{a}$ | $\boldsymbol{\omega}$ |
| MoCap | RMSE | $2.64\ m/s^2$ | $1.77\ ^\circ/s$ | $2.18\ m/s^2$ | $1.51\ ^\circ/s$ |
| | NRMSE | $10.2\ \%$ | $8.2\ \%$ | $8.4\ \%$ | $7.0\ \%$ |
| Video | RMSE | $4.07\ m/s^2$ | $3.52\ ^\circ/s$ | $3.73\ m/s^2$ | $3.45\ ^\circ/s$ |
| | NRMSE | $15.7\ \%$ | $16.3\ \%$ | $14.4\ \%$ | $16.0\ \%$ |

Table I compares error rates between the synthetic IMU data and the actual IMU data for the two aforementioned models. The evaluation is based on the root mean square error (RMSE) and its normalized value with respect to the range of the IMU data (NRMSE). For both the population and personal models, the synthetic inertial data obtained from the motion capture skeletal model outperforms the video-based skeletal model due to the inherent noise associated with the video-based pose extractor. Moreover, the personal models exhibit better performance compared to the population model, although the difference in error rates is marginal. This outcome might be attributed to motion artifacts and the simplified process involved in estimating $S_k^{\perp f}(t)$ while computing $\mathbf{R}(t)$, which affects both $\mathbf{a}$ and $\boldsymbol{\omega}$. Nevertheless, qualitatively, our model can accurately synthesize the inertial characteristics of the movements, as shown in Fig. 2.

We further examined the effect of employing the non-linear adaptive median filter to reduce noise in the pose estimation algorithms, as compared to the commonly used Butterworth low-pass filter. The population model for video-based synthetic IMU resulted in NRSEs of $57.2\ \%$ for $\mathbf{a}$ and $18.5\ \%$ for $\boldsymbol{\omega}$, which were substantially higher than those obtained using the adaptive median filter (i.e., $15.7\ \%$ for $\mathbf{a}$ and $16.3\ \%$ for $\boldsymbol{\omega}$). This demonstrates the superior performance of the adaptive median filter in reducing noise in the synthetic data

## VI. CONCLUSIONS

In this study, we have introduced a novel method for synthesizing realistic IMU data from video-based skeleton representations. We evaluated the accuracy of the virtual IMU data against real IMU data collected from subjects performing various upper-limb movements. This method not only accurately augments equivalent IMU data from video footage, but also enriches the IMU data obtained through Monte Carlo simulation. We firmly believe that our proposed method will significantly contribute to the augmentation of diverse IMU training data for deep learning applications, offering a promising solution to reduce the costs associated with data collection while enhancing the robustness and generalizability of the trained model. Important future work involves determining an optimal, efficient sampling technique for $\mathbf{p}_0$ and $\mathbf{R}_0$ to maximize the performance of downstream deep learning models.

## REFERENCES


[1] S. Majumder, T. Mondal, and M. J. Deen, "Wearable sensors for remote health monitoring," *Sensors*, vol. 17, no. 1, p. 130, 2017.

[2] C. Adans-Dester, N. Hankov, A. O'Brien, G. Vergara-Diaz, R. Black-Schaffer, R. Zafonte, J. Dy, S. I. Lee, and P. Bonato, "Enabling precision rehabilitation interventions using wearable sensors and machine learning to track motor recovery," *NPJ digital medicine*, vol. 3, no. 1, p. 121, 2020.

[3] H. Kwon, C. Tong, H. Haresamudram, Y. Gao, G. D. Abowd, N. D. Lane, and T. Ploetz, "IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–29, 2020.

[4] V. F. Rey, K. K. Garewal, and P. Lukowicz, "Yet it moves: Learning from generic motions to generate IMU data from YouTube videos," *arXiv preprint arXiv:2011.11600*, 2020.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[6] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "MediaPipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[7] J. Song, L. Wang, L. Van Gool, and O. Hilliges, "Thin-slicing network: A deep structured model for pose estimation in videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4220–4229.

[8] S. Zhao, "Time derivative of rotation matrices: A tutorial," *arXiv preprint arXiv:1609.06088*, 2016.

[9] A. D. Young, M. J. Ling, and D. K. Arvind, "IMUSim: A simulation environment for inertial sensing algorithm design and evaluation," in *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*. IEEE, 2011, pp. 199–210.

[10] H. Hwang and R. A. Haddad, "Adaptive median filters: new algorithms and results," *IEEE Transactions on image processing*, vol. 4, no. 4, pp. 499–502, 1995.

[11] S. Jiang, B. Lv, W. Guo, C. Zhang, H. Wang, X. Sheng, and P. B. Shull, "Feasibility of wrist-worn, real-time hand, and surface gesture recognition via sEMG and IMU sensing," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3376–3385, 2017.

[12] T. Edmond, A. Laps, A. L. Case, N. O'Hara, and J. M. Abzug, "Normal ranges of upper extremity length, circumference, and rate of growth in the pediatric population," *Hand*, vol. 15, no. 5, pp. 713–721, 2020.

[13] C. Z. Mooney, *Monte Carlo simulation*. Sage, 1997, no. 116.

[14] C. I. Tang, I. Perez-Pozuelo, D. Spathis, S. Brage, N. Wareham, and C. Mascolo, "SelfHAR: Improving human activity recognition through self-training with unlabeled data," *arXiv preprint arXiv:2102.06073*, 2021.

[15] J.-Y. Franceschi, A. Fawzi, and O. Fawzi, "Robustness of classifiers to uniform $\ell_p$ and Gaussian noise," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1280–1288.