Self-Supervised Learning Improves Accuracy and Data Efficiency for IMU-Based Ground Reaction Force Estimation

Tian Tan, Peter B. Shull*, *Member, IEEE,* Jenifer L. Hicks, Scott D. Uhlrich, and Akshay S. Chaudhari

Abstract—Objective: **Recent deep learning techniques hold promise to enable IMU-driven gait assessment; however, they require large extents of marker-based motion capture and ground reaction force (GRF) data to serve as labels for supervised model training. We thus propose a self-supervised learning (SSL) framework to leverage large IMU datasets for pre-training deep learning models, which can improve the accuracy and data efficiency of IMU-based vertical GRF (vGRF) estimation.** ***Methods:*** **To pre-train the models, we performed SSL by masking a random portion of the input IMU data and training a transformer model to reconstruct the masked portion. We systematically compared a series of masking ratios across three pre-training datasets that included real IMU data, synthetic IMU data, or a combination of the two. Finally, we built models that used pre-training and labeled data to estimate vGRF during three prediction tasks: overground walking, treadmill walking, and drop landing.** ***Results:*** **When using the same amount of labeled data, SSL pre-training improved the accuracy of vGRF estimation during walking compared to baseline models trained by conventional supervised learning. The correlation coefficients for vGRF estimation improved from 0.92 to 0.95 for overground waking and from 0.94 to 0.97 for treadmill walking. Also, using 1–10% of walking data to fine-tune pre-trained models yielded comparable accuracy to the baseline model that was trained on 100% of walking data.** ***Conclusion:*** **The proposed SSL framework leveraged large real and synthetic IMU datasets to increase the accuracy and data efficiency of deep-learning-based vGRF estimation, reducing the need of labels.** ***Significance:*** **This work may unlock broader use cases of IMU-driven assessment where only small labeled datasets are available.**

Index Terms**—Inertial measurement unit, machine learning, SSL, wearable sensing, kinetics.**

I. Introduction

BIOMECHANICAL assessment of human movement is important for ascertaining human health. Kinetics during locomotion can be used to predict injury risk and assess rehabilitation outcomes. For example, ground reaction forces (GRF) during a sit-to-stand task can provide insights into rehabilitation outcomes after hip injury [1], and knee adduction and flexion moments during walking have been associated with knee osteoarthritis [2]–[4]. Additionally, GRF patterns during drop landing, where subjects step off an elevated platform and land on the ground, are associated with several lower-extremity injuries [5], [6].

Despite the value of estimating kinetics, measurement of GRF and joint moments typically requires gold standard measurements from force plates and marker-based motion capture. However, these devices are expensive, require trained personnel to operate, and confine the measurement environment. Consequently, such challenges associated with acquiring kinetic data limit the accessibility of biomechanical assessment and prevent translation of biomechanical studies to larger and diverse populations.

Using wearable sensors to monitor kinetic parameters in natural environments represents a promising opportunity for moving disease characterization, prevention, and rehabilitation assessment into real-world settings [7]. Inertial measurement units (IMUs) are a practical choice for out-of-lab assessment due to their low cost and small form factor. IMUs typically measure acceleration and angular velocity (and magnetic field in some scenarios), which can directly be used to derive body segment orientation for subsequent kinematic estimation.

While estimating kinetics from IMUs and physics-based models is challenging, new research studies attempt to employ deep learning models to fill this gap [8]–[10]. To train these models, previous studies have collected synchronized IMU, GRF, and marker-based motion capture data during hundreds of minutes of walking and running on treadmills equipped with force plates. The GRF and marker data then serve as "labels" to train models using conventional supervised learning. However, collecting these "labels" in large quantities to train accurate deep learning models is not always feasible for activities beyond treadmill gait. For overground walking, since GRF data is typically collected via one to three floor-embedded force plates, each trial often yields only one or several steps with "labels". The scarcity of data also holds true for a variety of other activities including sprinting, cutting, and jumping [11].

Self-supervised learning (SSL) may be used to pre-train deep learning models to mitigate the challenge of data scarcity [12]–[14]. It may also improve models' robustness to noisy data [15] and distribution shifts where the distribution of model training data differs from that of testing data [16]. SSL aims to pre-train deep learning models by designing model pre-

Tian Tan is with the Department of Radiology, Stanford University, Stanford, CA, 94305, USA (email: alanttan@stanford.edu).

Peter B. Shull is with the State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (email: pshull@sjtu.edu.cn).

Jenifer L. Hicks and Scott D. Uhlrich are with the Department of Bioengineering, Stanford University, Stanford, CA, 94305, USA (email: jenhicks@stanford.edu; suhlrich@stanford.edu).

Akshay S. Chaudhari is with the Department of Radiology and Biomedical Data Science, Stanford University, Stanford, CA, 94305, USA (email: akshaysc@stanford.edu).

Manuscript received October 19, 2023. Corresponding author: Peter B. Shull.

TABLE I: Two datasets for model pre-training and three datasets for downstream evaluation.

Type	Ref. (name)	Motion	# of Participants[a]	# of Windows [a, b]
Pre-training Datasets	[18] (MoVi)	21 motions including walking	90	14,418
	[19] (AMASS)	100+ motions including walking	500	142,853
Downstream Datasets	[20]	Overground walking	22	510
	[21]	Treadmill walking	17	23,481
	[22]	Drop landing	15	509

[a] The pre-training datasets have more participants as well as comparable or more windows than downstream evaluation datasets.
[b] Each window is a model input segmented from original trials.

training tasks that are solely based on unlabeled data. One widely used pre-training task is masking and reconstruction, where a random portion of the input data is masked and the model is trained to learn the data representation by reconstructing the masked portion. When limited "labeled" data is available for a downstream prediction task of interest, representations learned from pre-trained models can enhance the training process, and lead to better performance than conventional supervised learning with randomly initialized parameters. SSL with masking and reconstruction has facilitated the training of powerful language models [12], [17], vision models [13], and multi-modal models [14]. However, it is unknown whether SSL with masking and reconstruction can learn IMU representations that improve performance on downstream tasks of estimating kinetics. Additionally, unlike the fields of natural language and vision, the optimal pre-training dataset and strategy for this IMU-based application remain unclear.

The aim of this paper is to introduce an SSL framework for improving the accuracy and data efficiency of an IMU-driven deep learning model. We first investigated the optimal SSL strategy using a large "unlabeled" dataset of 590 participants with real and synthetic IMU data. We then compared the performance and data efficiency of SSL pre-trained models on estimating vGRF compared to baseline models whose parameters are randomly initialized following conventional supervised learning. We first estimated vGRF on overground and treadmill walking datasets, which was within the distribution of the SSL pre-training datasets. Then, we estimated vGRF on a drop landing dataset, which is not present in the pre-training dataset. To facilitate broader usage and reproduction of work, we make our framework broadly usable by releasing our source code and trained models.

II. Methods

A. Pre-training Datasets

We used three datasets for our SSL pre-training: real IMU data, synthetic IMU data, and a fused dataset combining them both (Table I). The real IMU dataset has 17 IMUs, and we used eight of them that were placed on the trunk and pelvis, as well as both thighs, shanks, and feet. For the synthetic IMU dataset, we generated data on the same eight body segments. Each IMU has a 3-axis accelerometer and a 3-axis gyroscope. The z-axes of the IMUs were aligned with the segment surface normal, x-axes were pointing left during standing, and y-axes were perpendicular to the x- and z-axes following the right-hand rule (Fig. 1).

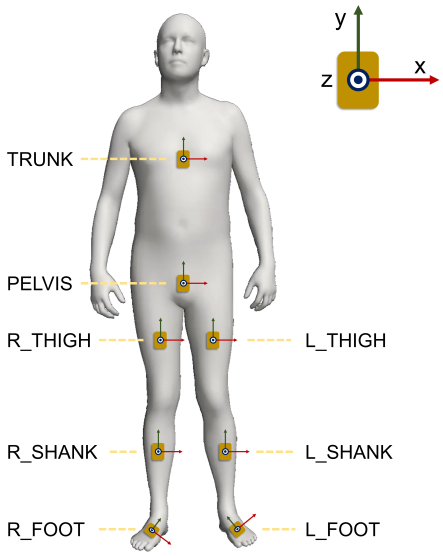

Fig. 1: Locations and orientations of eight synthetic IMUs generated from the AMASS dataset. Red arrows, green arrows, and blue dots represent x-, y-, and z-axes of IMUs, respectively.

1) Real IMU: This dataset is based on MoVi, a public human motion dataset with 90 participants who performed a collection of 21 everyday actions and sports movements such as walking, vertical jumping, and throwing [18]. Participants wore a 17-IMU suit that operated at 120Hz (Noitom, China). We linearly resampled each trial from 120Hz to 100Hz and then segmented it into windows of 128 time steps ($1.28s$ each) with 50% overlap ($0.64s$) between successive windows. We excluded 18,760 (56.5%) windows in which participants were standing or sitting still, when the average standard deviation of all accelerometer axes was smaller than $0.2m/s^2$. In total, we obtained 14,418 windows and randomized their order for pre-training.

2) Synthetic IMU: This dataset is based on AMASS, a large-scale, public human motion archive involving 500 participants [19]. AMASS unifies 23 existing motion capture datasets that have different marker sets, sampling frequencies ranging from 59–120Hz, and a variety of activities, with walking being one of the most common activities. Using kinematics to generate synthetic IMU data is a widely-used method for augmenting the volume of training data [23]–[26]. To generate synthetic IMU data, we low-pass filtered the joint angles and positions of each trial at 15Hz using a zero-

lag, fourth-order Butterworth filter and resampled the filtered data to 100Hz. Subsequently, we placed eight synthetic IMUs on their corresponding body segments (Fig. 1) and generated their positions and orientations via forward kinematics using a prior pipeline [27]. Then, synthetic angular velocities were computed by taking the first derivative of orientation, whereas synthetic accelerations were computed by taking the second derivative of positions and incorporating gravity. Similar to the MoVi dataset pre-processing, we segmented the synthetic IMU data into windows of 128 time steps with 50% window overlap. We excluded 170,568 (54.0%) windows whose average acceleration was smaller than $0.2m/s^2$. We also excluded 2,454 (0.8%) windows that contained abnormally large synthetic acceleration or angular velocity with thresholds of $160m/s^2$ and $2000deg/s$, respectively, which were determined based on measurement ranges of typical IMUs (MPU-9250, Invensense, USA and BHI-160, Bosch Sensortec, Germany). In total, we obtained 142,853 windows and randomized their order for SSL.

3) Real and Synthetic IMU: This dataset combines real IMU data and synthetic IMU data described above, resulting in a total of 157,271 windows. Such a data dataset has the potential to leverage the benefits of a substantial data scale as well as authentic sensor noise representations.

B. Downstream Datasets

We evaluated the performance of IMU representations from SSL pre-training in estimating vGRFs for three different tasks, with corresponding datasets of synchronized IMU and GRF data: overground walking, treadmill walking, and drop landing. For consistency, the vGRF of the right leg was used as the estimation target for all three tasks.

We assumed that sensors were placed to body segments close to the following alignment: one of their surfaces aligned to the body surface and one of their edges aligned with the segment's long axis. We manually verified whether the IMU-to-segment placement conventions of each dataset are consistent with the assumed alignment. If not, we rotated the IMU axes to match the assumed alignment. The proposed model uses data from eight IMUs as input (Fig. 1). For downstream tasks that used a smaller IMU set (e.g., overground walking [20]), we substituted each omitted IMU with a randomly initialized, trainable data patch that is updated during the training process.

1) Overground Walking [20]: The original dataset contains 19 participants walking on stairs, ramps, treadmills, and level ground at various speeds, but we only used level-ground walking data to build the model to estimate vGRF. The input IMU data are collected at 200Hz from four IMUs (Yost, USA) placed on the torso and right thigh, shank, and foot. GRF was measured by force plates (Bertec, USA) at 1,000Hz and low-pass filtered at 15Hz. We downsampled the synchronized IMU and GRF data to 100Hz to match that of the pre-training datasets. Each trial has three gait cycles with gold-standard GRF measurements. We extracted these gait cycles using windows of 128 time steps, starting at 40 samples before the stance phases. In total, there are 510 windows available for model training and evaluation.

2) Treadmill Walking [21]: This dataset includes 17 participants walking on an instrumented treadmill with various speeds, foot progression angles, trunk sway angles, and step widths. Eight 8 IMUs (SageMotion, USA) were placed on the trunk, pelvis, and both thighs, shanks, and feet to collect data at 100Hz. The vGRF of the right leg is measured by an instrumented treadmill (Bertec, USA) at 1,000Hz, low-pass filtered as 15Hz, and downsampled to 100Hz. For each trial, continuous walking data were segmented into windows of 128 time steps with 50% overlap. We retained all data windows because GRF is measured by treadmill-embedded force plates throughout the entire trial. In total, there were 23,481 windows available for model training and evaluation.

3) Drop Landing [22]: This dataset includes 17 participants performing double-leg drop landing trials from a 30-cm-high box with various toe-out angles. Eight IMUs (Xsens, The Netherlands) were placed on the trunk, pelvis, and both thighs, shanks, and feet to collect data at 100Hz. The vGRF of the right leg was measured by a floor-embedded force plate (AMTI, USA) at 1,000Hz, low-pass filtered at 15Hz, and downsampled to 100Hz. Each trial contains 0.8s of data starting at the moment of jumping, corresponding to 80 time steps. We appended zeros to the end of each trial to match the input length of our transformer. We retained all the trials that were successfully collected. In total, we obtained 509 windows from 509 drop landing trials for model training and evaluation.

C. Model Architecture

We built a transformer model to learn IMU representations during pre-training (Fig. 2(a)). The transformer's input is a window of IMU data with 128 samples and 48 IMU axes (eight 3-axis accelerometers and eight 3-axis gyroscopes). We first normalized each IMU axis by removing the mean and scaling to unit variance. Then, the 128 samples were sliced into patches and the values of each patch were flattened. We randomly masked a portion of patches with a trainable masking patch that is updated during the training process. The masking ratio is a critical parameter in SSL, and it depends on the patch length of transformers. We performed a grid search to jointly optimize the masking ratio (6.25%, 12.5%, 12.5%, 25%, 37.5%, 50%, and 62.5%) and the patch length (1, 2, 4, and 8). After masking, we linearly mapped each patch to a vector of length 192 and added sinusoidal positional encoding [30]. Then, the data was fed into the transformer model, generating an output with dimensions identical to the input. Each output patch was linearly mapped to the original patch size (patch length $\times$ 48 IMU axes) to reconstruct the unmasked input data. The transformer was trained to minimize the mean squared error between the reconstructed data and the unmasked input data over the masked patches. The transformer has six self-attention blocks, with a per-block configuration of 8 attention heads, 512 feedforward units, 10% dropout, 10^{-5} layer normalization, and rectified linear unit (ReLU) activation function. The total number of trainable parameters is 2 million.

The model for downstream prediction tasks (i.e., estimating vGRF) has the same architecture as the one used in pre-training (Fig. 2(b)). The output of each patch was linearly

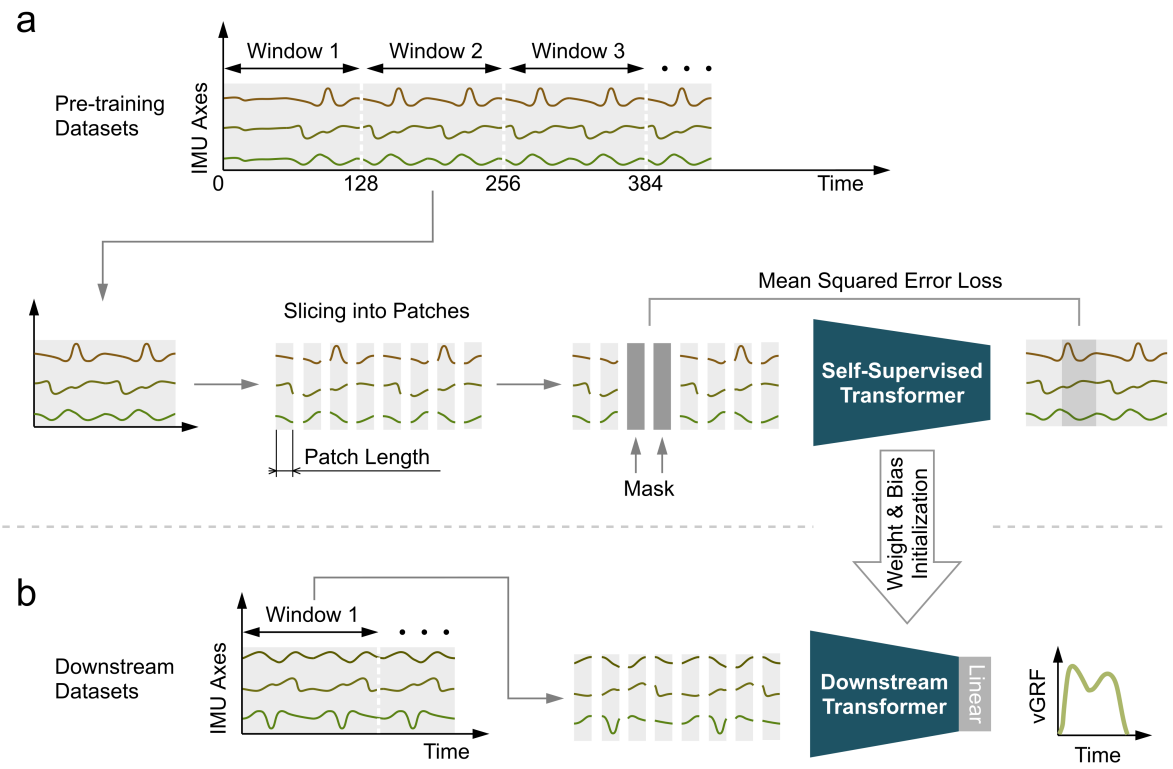

Fig. 2: Self-supervised IMU representation learning to boost downstream prediction tasks. (a) Continuous IMU data are segmented into windows, sliced into patches, randomly masked, and finally fed into a transformer model to reconstruct the original data window. (b) The weights and biases of the self-supervised transformer were copied to initialize those of the transformer for downstream evaluation. A linear layer was appended to the transformer to map the transformer outputs to the estimation target. During fine-tuning, we optimized the linear layer first and then fine-tuned the entire model to prevent distortion of pre-trained IMU representations [28], [29]. The fine-tuning data preprocessing steps are identical to those used in SSL, except for the exclusion of masking.

mapped to vGRF. The data pre-processing steps are identical to those in pre-training except that no patch is masked. For each downstream task, we investigated the performance of three pre-trained models whose parameters are inherited from SSL pre-training on the three different pre-training datasets. We also trained separate baseline models whose parameters are randomly initialized following conventional supervised learning with PyTorch's default settings. Specifically, we set linear layer weights following the Kaiming uniform method [31], attention layer weights following the Xavier uniform method [32], and model biases to zeros.

D. Training Protocols

Models were implemented in Python 3.9 with PyTorch 1.13 and Numpy 1.23. The random seeds of Numpy and PyTorch are both set to 0. One Nvidia RTX 2080 Ti graphics card was used to train and test all the models.

For pre-training, we used the AdamW optimizer [33] to minimize the mean squared error with a learning rate of 10^{-4}, a weight decay of 0.01, and a minibatch size of 64 for 10^4 backpropagation steps without early stopping. We used a linear warmup of the learning rate for the first 20% of steps, and decayed the learning rate with a cosine decay schedule without restarts [34]. We used the entire dataset for pre-training without splitting a test set because the model evaluation was performed based on downstream datasets that are different from pre-training datasets.

For downstream evaluation, we implemented a two-step fine-tuning strategy for pre-trained models, which optimizes the linear layer first and then fine-tunes the entire model to prevent distortion of pre-trained IMU representations [28], [29]. For the first step where the linear layer was trained, we used AdamW optimizer with a minibatch size of 64, a learning rate of 10^{-3}, and a weight decay of 0.01 for 300 backpropagation steps. For the second step where the entire model is fine-tuned, we used AdamW optimizer with a minibatch size of 64, a learning rate of 10^{-4}, and a weight decay of 0.01 for 300 backpropagation steps. For both steps, we used the mean squared error loss function, linearly warmed up the learning rate for the first 20% of backpropagation steps, and decayed the learning rate with a cosine decay schedule without restarts [34]. The same two-step strategy was used to train the baseline model (with no pre-training). We trained and tested models with data from different participants for each dataset using a five-fold cross-validation approach, where participants were randomly split into five groups. Four groups

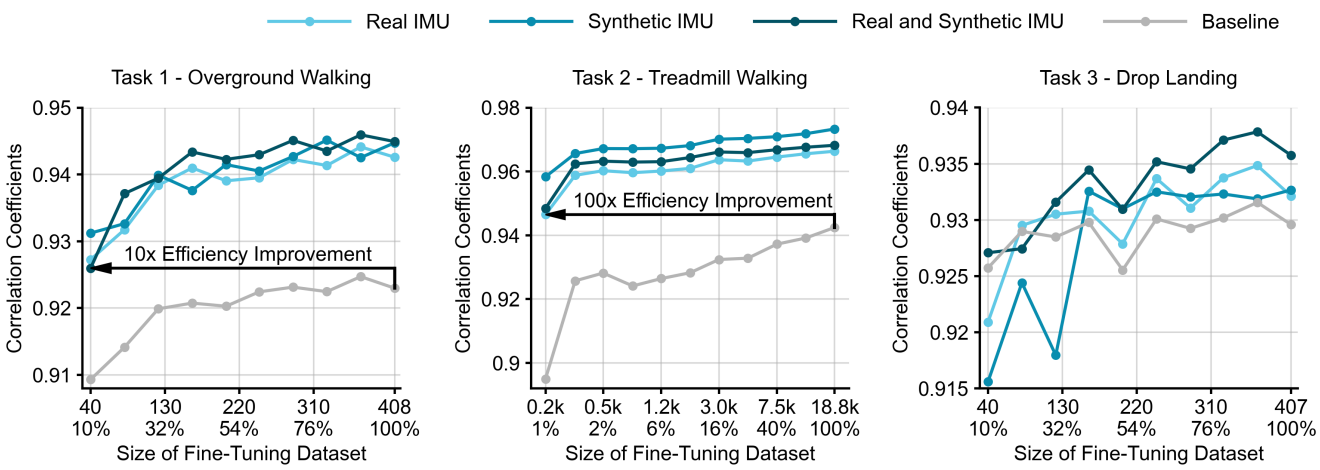

Fig. 3: Correlation coefficient between gold-standard and estimated vGRF for SSL pre-trained and baseline models when using a range of reduced downstream datasets for training. The sizes of overground walking and drop landing datasets for model fine-tuning were reduced linearly whereas the size of treadmill walking dataset was reduced exponentially due to a significantly larger dataset size. Black arrows in task 1 - overground walking and task 2 - treadmill walking indicate improvements in data efficiency.

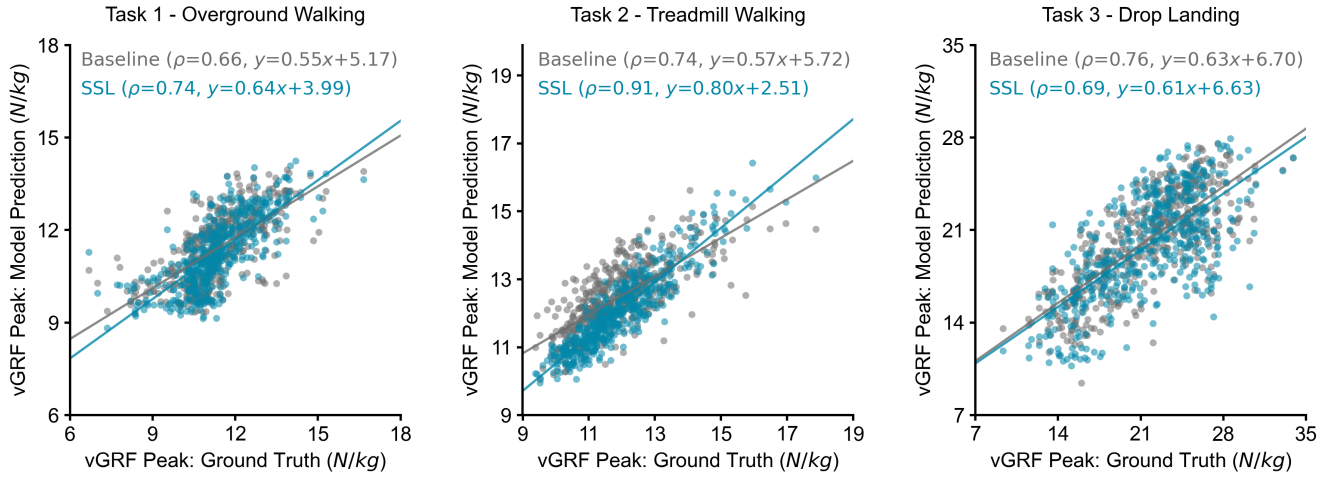

Fig. 4: Scatter plot and regression line showing gold-standard and estimated vGRF peaks for SSL pre-trained and baseline models. All the windows are shown for task 1 (overground walking) and task 3 (drop landing), whereas 500 windows were randomly selected from 23,481 windows for task 2 (treadmill walking).

were combined as the training set for each of the five iterations, and the remaining group was used as the test set.

E. Performance Evaluation

We compared different models using Pearson's correlation (ρ) between the estimated and gold-standard vGRF across all the stance phases of each subject. Swing phases where vGRF is lower than 2% of the body weight are excluded from the analysis. A one-way analysis of variance (ANOVA) was used to determine if pre-training would affect the accuracy of vGRF estimation when using the entire downstream dataset for fine-tuning. If significant differences were detected, paired t-tests with Bonferroni corrections were used to compare pairs of models. The level of significance was set to 0.05.

Beyond overall vGRF measures, we also quantified the error of our pre-trained models across a spectrum and compared it to that of baseline models. We performed spectrum analysis using Fast Fourier Transform (FFT) on the vGRF measured by the force plate as well as on the vGRF estimated by models for each window. Then, we computed the differences in FFT results and averaged the differences across all the windows.

In addition, we investigated scaling law curves for model performance by varying the downstream dataset sizes. For each fold of cross-validation, we randomly sampled a subset of the training set for model training, while used the entire test set for evaluation. We linearly reduced the sizes (100%, 90%, 80%, ..., 10%) for vGRF estimation during overground walking and drop landing, whereas exponentially reduced the sizes ($10^0 \times$

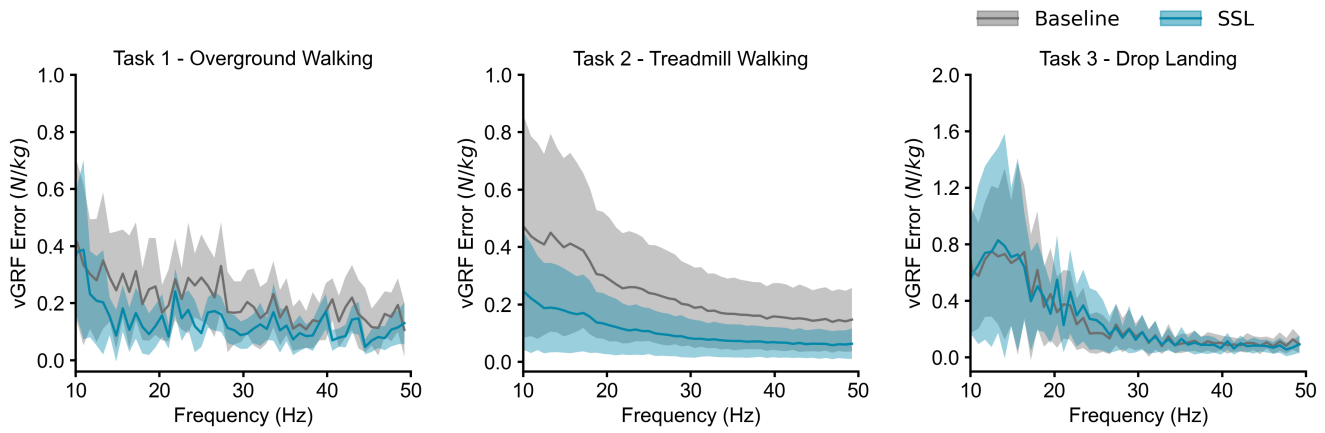

Fig. 5: The error between gold-standard measurements and model estimations across the spectrum. Pre-trained models (blue) exhibited smaller errors compared to baseline models (gray) for overground and treadmill walking.

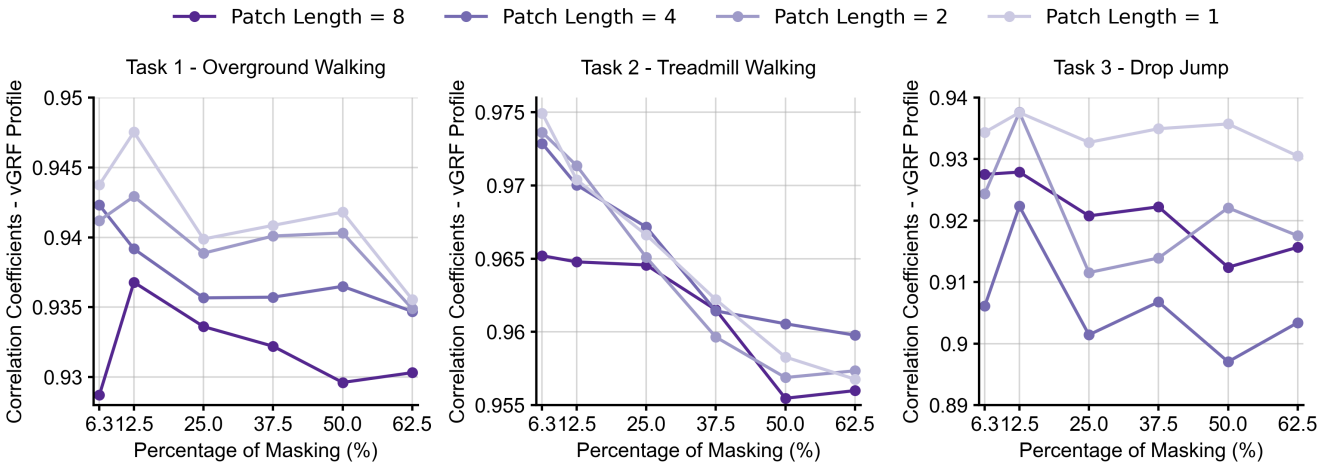

Fig. 6: Correlation coefficient between the gold-standard and estimated vGRF for a variety of models trained with patch lengths of 1–8 and with 6.25–62.5% masking during pre-traing. A patch length of 8 and 6.25–12.5% masking have the highest accuracy.

100%, $10^{-0.2} \times 100\%$, $10^{-0.4} \times 100\%$, ..., $10^{-2} \times 100\%$) for vGRF estimation during treadmill walking due to its larger base size.

III. Results

SSL pre-trained models achieved higher accuracy in estimating vGRF during overground walking and treadmill walking across various downstream dataset sizes compared to baseline models that have an identical model architecture (Fig. 3). The models pre-trained on real IMU data, synthetic IMU data, and a combination of the two were all significantly more accurate than the baseline models when using the entire overground or treadmill walking dataset for fine-tuning. No differences were observed between pre-trained and baseline models in estimating vGRF during drop landing. Using 10% of overground walking data to fine-tune the SSL pre-trained models yielded comparable accuracy to the baseline model trained on 100% of overground walking data, indicating a 10x data efficiency improvement. Similarly, using 1% of treadmill walking data to fine-tune the pre-trained models yielded comparable accuracy to the baseline models that are trained on 100% of treadmill walking data, indicating a 100x data efficiency improvement. There was no statistical difference between the three SSL pre-trained models for all three prediction tasks except that the model pre-trained on the synthetic IMU dataset was significantly more accurate than the model pre-trained on the real IMU dataset for treadmill walking.

When evaluating performance on vGRF peaks, the SSL pre-trained models also outperformed the baseline model for overground and treadmill walking (Fig. 4). SSL pre-training increased the correlation of peak vGRF estimation from 0.66 to 0.74 for overground walking and from 0.74 to 0.91 for treadmill walking. SSL pre-training reduced the correlation of peak vGRF estimation from 0.76 to 0.69 for drop landing. Consistently smaller errors were observed in pre-trained models compared to the baseline models across the spectrum for overground and treadmill walking (Fig. 5).

For drop landing, the pre-trained model exhibited comparable error to the baseline model across the spectrum.

A masking ratio of 6.25–12.5% with shorter patch lengths achieved the best performance across all three prediction tasks (Fig. 6). Thus, we implemented our self-supervised transformer models with a patch length of 1 (equivalent to 10 ms at 100Hz sampling frequency) and 12.5% masking, and all the above results are reported based on these models.

IV. Discussion

This work shows that using SSL to pre-train deep learning models that predict walking vGRF from IMUs can improve data efficiency by 10–100x. Prior deep learning models for biomechanical estimation are predominantly trained on proprietary datasets with a small number of participants. In contrast, we used open-source motion datasets with hundreds of participants to pre-train our models. This code and pre-trained model have been made publicly available for researchers interested in employing them to improve the results on new and customized downstream prediction tasks (Appendix A).

In the absence of large-scale real IMU datasets, we utilized a large-scale motion dataset (AMASS) to generate synthetic IMU data for SSL. Its combined window duration is equivalent to 56 hours of data, substantially exceeding prior kinetic estimation studies that used 1.2–6 hours of real IMU data [8]–[10] and 0.2–2 hours of synthetic IMU data [35], [36]. Performance was similar between models pre-trained on a relatively smaller real IMU dataset and a larger synthetic IMU dataset. This suggests that the size of the synthetic dataset compensated for the distribution shift between real and synthetic data that arise from soft tissue artifacts, sensor noise, and human model simplification [23]. A large dataset of real IMU data would likely improve performance compared to our results. Collecting such a large "unlabeled" IMU dataset is easier and more efficient than marker-based motion capture data because hours of data can be collected in unconstrained real-life or sports environments. Walking is involved in both real and synthetic IMU datasets for SSL pre-training, enabling pre-trained models to learn from these data for enhancing prediction tasks during walking. Although the activities in both datasets is varied, they do not include drop landings; accordingly, pre-training on these datasets did not improve vGRF estimation during drop landing. Tailoring the distribution of datasets for specific estimation tasks is a commonly used strategy for improving the performance of deep learning [37]–[39]. Thus, when collecting a large "unlabeled" IMU dataset, it should involve a wide range of motions such as cutting, jumping, and landing that are related to clinical assessments and of interest to biomechanists.

The performance of our pre-trained models is comparable to or slightly higher than those reported in prior studies that used the same downstream datasets for vGRF estimation. Specifically, a prior deep learning model achieved a correlation coefficient of 0.96 ± 0.01 for overground walking and 0.96 ± 0.01 for treadmill walking [40], while our models achieved 0.95 ± 0.02 and 0.97 ± 0.01, respectively. Another prior deep learning model achieved a correlation coefficient of 0.92 ± 0.11 [22] for vGRF estimation during drop landing, while our model achieved 0.93 ± 0.04. Unlike prior studies, we excluded swing phases during performance evaluation, which reduced correlations and increased RMSEs of our models. Also, prior studies tailored their model architectures and hyperparameters for specific prediction tasks, and thus they might not be reused for other prediction tasks. In contrast, we demonstrated that our unified model architecture can be generalized to multiple downstream prediction tasks. In addition, the RMSE of peak vGRF estimation of the pre-trained models were 0.56–0.88N/kg during treadmill and overground walking, which are comparable to the smallest detectable change of peak vGRF during walking in patients with stroke or knee osteoarthritis (0.56–0.61N/kg) [41], [42]. The RMSE of peak vGRF estimation of the pre-trained models was 3.53N/kg during drop landing, which is higher than the smallest detectable change of peak vGRF during drop landing (0.55–0.63N/kg) [43].

The optimal mask ratio is related to the information density of the input data. Previous language models used low masking ratios for natural language that is information-dense (e.g., 15% [17]), whereas vision models used high mask ratios for images that have high information redundancy (e.g., 75% [13], [14]). We observed that 6.26–12.5% masking ratios have the best performance (Fig. 6), indicating that IMU data are information-dense. Nevertheless, the optimal masking may increase when the IMU sampling frequencies are substantially higher than 100Hz due to increased information redundancy. Our proposed model used a window length of 128 samples ($1.28s$), which is longer than the typical duration of a walking gait cycle or a drop landing trial. However, we acknowledge that such a window length may limit the model performance on other tasks such as GRF estimation during balancing, which needs a longer window length to extend the temporal scope of the model.

Our code, pre-trained models, and tutorials are publicly available for researchers who are interested in using them for their downstream prediction tasks. Our models were trained with eight IMUs; however, they are robust to distribution shift caused by a smaller IMU set, as demonstrated in overground walking vGRF estimation, where only four IMUs were used. Thus, they can be used for sparse IMU configurations that are more practical for real-world measurements [26], [44], [45] We recommend placing IMUs with their z-axes aligned with the segment surface normal, x-axes pointing left during standing, and y-axes perpendicular to the x- and z-axes following the right-hand rule (Fig. 1). For a different IMU-to-segment placement configuration, the recommended configuration can typically be achieved through a post-hoc rotation of the IMU data.

V. Conclusion

SSL with large IMU datasets for model pre-training can improve data efficiency and accuracy compared to conventional supervised learning when the distribution of the downstream dataset is within that of pre-training datasets. Future studies may collect "unlabeled" IMU data under real-life or sports

environments to expand the scale and distribution of pre-training datasets, thus potentially unlocking newer and broader use cases of IMU-driven assessment where only limited gold-standard measurements are available.

Appendix A
Code and Data Availability

The trained models are available on https://github.com/StanfordMIMI/SSL_IMU. The source code is also provided in the repository for researchers who are interested in applying SSL for pre-training specialized deep learning model architectures.

Acknowledgment

This study was funded by the Joe and Clara Tsai Foundation through the Wu Tsai Human Performance Alliance, the U.S. National Institutes of Health (NIH) under grant R01 AR077604, R01 EB002524, R01 AR079431, and P41 EB027060, and the National Natural Science Foundation of China under grant 52250610217.

References

[1] J. Houck, J. Kneiss, S. V. Bukata, and J. E. Puzas, "Analysis of vertical ground reaction force variables during a sit to stand task in participants recovering from a hip fracture," *Clinical Biomechanics*, vol. 26, no. 5, pp. 470–476, 2011.

[2] D. Hurwitz, A. Ryals, J. Case, J. Block, and T. Andriacchi, "The knee adduction moment during gait in subjects with knee osteoarthritis is more closely correlated with static alignment than radiographic disease severity, toe out angle and pain," *Journal of Orthopaedic Research*, vol. 20, no. 1, pp. 101–107, 2002.

[3] L. Sharma, D. E. Hurwitz, E. J.-M. Thonar, J. A. Sum, M. E. Lenz, D. D. Dunlop, T. J. Schnitzer, G. Kirwan-Mellis, and T. P. Andriacchi, "Knee adduction moment, serum hyaluronan level, and disease severity in medial tibiofemoral osteoarthritis," *Arthritis & Rheumatism*, vol. 41, no. 7, pp. 1233–1240, 1998.

[4] T. Miyazaki, M. Wada, H. Kawahara, M. Sato, H. Baba, and S. Shimada, "Dynamic load at baseline can predict radiographic disease progression in medial compartment knee osteoarthritis," *Annals of the Rheumatic Diseases*, vol. 61, no. 7, pp. 617–622, 2002.

[5] I. Aerts, E. Cumps, E. Verhagen, J. Verschueren, and R. Meeusen, "A systematic review of different jump-landing variables in relation to injuries," *J Sports Med Phys Fitness*, vol. 53, no. 5, pp. 509–519, 2013.

[6] B. Caulfield and M. Garrett, "Changes in ground reaction force during jump landing in subjects with functional instability of the ankle joint," *Clinical Biomechanics*, vol. 19, no. 6, pp. 617–621, 2004.

[7] F. Porciuncula, A. V. Roto, D. Kumar, I. Davis, S. Roy, C. J. Walsh, and L. N. Awad, "Wearable movement sensors for rehabilitation: a focused review of technological and clinical advances," *PM&R*, vol. 10, no. 9, pp. S220–S232, 2018.

[8] F. J. Wouda, M. Giuberti, G. Bellusci, E. Maartens, J. Reenalda, B.-J. F. Van Beijnum, and P. H. Veltink, "Estimation of vertical ground reaction forces and sagittal knee kinematics during running using three inertial sensors," *Frontiers in Physiology*, vol. 9, p. 218, 2018.

[9] T. Tan, Z. A. Strout, and P. B. Shull, "Accurate impact loading rate estimation during running via a subject-independent convolutional neural network model and optimal imu placement," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1215–1222, 2020.

[10] R. S. Alcantara, W. B. Edwards, G. Y. Millet, and A. M. Grabowski, "Predicting continuous ground reaction forces from accelerometers during uphill and downhill running: a recurrent neural network solution," *PeerJ*, vol. 10, p. e12752, 2022.

[11] T. Tan, A. A. Gatti, B. Fan, K. G. Shea, S. L. Sherman, S. D. Uhlrich, J. L. Hicks, S. L. Delp, P. B. Shull, and A. S. Chaudhari, "A scoping review of portable sensing for out-of-lab anterior cruciate ligament injury prevention and rehabilitation," *NPJ Digital Medicine*, vol. 6, no. 1, p. 46, 2023.

[12] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[14] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, "Scaling language-image pre-training via masking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 390–23 400.

[15] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International Conference on Machine Learning*, 2020, pp. 9229–9248.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2019, pp. 4171–4186.

[18] S. Ghorbani, K. Mahdaviani, A. Thaler, K. Kording, D. J. Cook, G. Blohm, and N. F. Troje, "Movi: A large multi-purpose human motion and video dataset," *Plos one*, vol. 16, no. 6, p. e0253157, 2021.

[19] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5442–5451.

[20] J. Camargo, A. Ramanathan, W. Flanagan, and A. Young, "A comprehensive, open-source dataset of lower limb biomechanics in multiple conditions of stairs, ramps, and level-ground ambulation and transitions," *Journal of Biomechanics*, vol. 119, p. 110320, 2021.

[21] T. Tan, D. Wang, P. B. Shull, and E. Halilaj, "Imu and smartphone camera fusion for knee adduction and knee flexion moment estimation during walking," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1445–1455, 2022.

[22] T. Sun, D. Li, B. Fan, T. Tan, and P. B. Shull, "Real-time ground reaction force and knee extension moment estimation during drop landings via modular lstm modeling and wearable imus," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[23] T. Zimmermann, B. Taetz, and G. Bleser, "Imu-to-segment assignment and orientation alignment for the lower body using deep learning," *Sensors*, vol. 18, no. 1, p. 302, 2018.

[24] T. Tan, D. P. Chiasson, H. Hu, and P. B. Shull, "Influence of imu position and orientation placement errors on ground reaction force estimation," *Journal of Biomechanics*, vol. 97, p. 109416, 2019.

[25] E. Rapp, S. Shin, W. Thomsen, R. Ferber, and E. Halilaj, "Estimation of kinematics from inertial measurement units using a combined deep learning and optimization framework," *Journal of Biomechanics*, vol. 116, p. 110229, 2021.

[26] Y. Jiang, Y. Ye, D. Gopinath, J. Won, A. W. Winkler, and C. K. Liu, "Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation," in *SIGGRAPH Asia 2022 Conference Papers*, 2022, pp. 1–9.

[27] X. Yi, Y. Zhou, and F. Xu, "Transpose: Real-time 3d human translation and pose estimation with six inertial sensors," *ACM Transactions on Graphics*, vol. 40, no. 4, Aug. 2021.

[28] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," in *International Conference on Learning Representations*, 2022.

[29] J. Dominic, N. Bhaskhar, A. D. Desai, A. Schmidt, E. Rubin, B. Gunel, G. E. Gold, B. A. Hargreaves, L. Lenchik, R. Boutin *et al.*, "Improving data-efficiency and robustness of medical imaging segmentation using inpainting-based self-supervised learning," *Bioengineering*, vol. 10, no. 2, p. 207, 2023.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[32] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth*

International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.

[34] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2016.

[35] M. Mundt, A. Koeppe, S. David, T. Witter, F. Bamer, W. Potthast, and B. Markert, "Estimation of gait mechanics based on simulated and measured imu data using an artificial neural network," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 41, 2020.

[36] E. Dorschky, M. Nitschke, C. F. Martindale, A. J. van den Bogert, A. D. Koelewijn, and B. M. Eskofier, "Cnn-based estimation of sagittal plane walking and running biomechanics from measured and simulated inertial sensor data," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 604, 2020.

[37] I. Ktena, O. Wiles, I. Albuquerque, S.-A. Rebuffi, R. Tanno, A. G. Roy, S. Azizi, D. Belgrave, P. Kohli, A. Karthikesalingam *et al.*, "Generative models improve fairness of medical classifiers under distribution shifts," *arXiv preprint arXiv:2304.09218*, 2023.

[38] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J. M. Z. Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, and A. Chaudhari, "Roentgen: vision-language foundation model for chest x-ray generation," *arXiv preprint arXiv:2211.12737*, 2022.

[39] Y. Meng, M. Michalski, J. Huang, Y. Zhang, T. Abdelzaher, and J. Han, "Tuning language models as training data generators for augmentation-enhanced few-shot learning," in *International Conference on Machine Learning*, 2023, pp. 24 457–24 477.

[40] M. S. B. Hossain, Z. Guo, and H. Choi, "Estimation of lower extremity joint moments and 3d ground reaction forces using imu sensors in multiple walking conditions: A deep learning approach," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[41] I. Campanini and A. Merlo, "Reliabilty, smallest real difference and concurrent validity of indices computed from grf components in gait of stroke patients," *Gait & posture*, vol. 30, no. 2, pp. 127–131, 2009.

[42] D. Pacifico, R. Visscher, R. List, J. F. Item-Glatthorn, N. C. Casartelli, and N. A. Maffiuletti, "Discriminant validity and reproducibility of spatiotemporal and kinetic parameters during treadmill walking in patients with knee osteoarthritis," *Gait & posture*, vol. 80, pp. 77–79, 2020.

[43] L. Howe, J. North, M. Waldron, and T. Bampouras, "Reliability of independent kinetic variables and measures of inter-limb asymmetry associated with bilateral drop-landing performance," *International Journal of Physical Education, Fitness and Sports*, vol. 7, no. 3, pp. 32–47, 2018.

[44] M. Lueken, J. Wenner, S. Leonhardt, and C. Ngo, "Using synthesized imu data to train a long-short term memory-based neural network for unobtrusive gait analysis with a sparse sensor setup," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, 2022, pp. 3653–3656.

[45] T. Van Wouwe, S. Lee, A. Falisse, S. Delp, and C. K. Liu, "Diffusion inertial poser: Human motion reconstruction from arbitrary sparse imu configurations," *arXiv preprint arXiv:2308.16682*, 2023.