RELCON: RELATIVE CONTRASTIVE LEARNING FOR A MOTION FOUNDATION MODEL FOR WEARABLE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

We present RelCon, a novel self-supervised *Rel*ative *Con*trastive learning approach that uses a learnable distance measure in combination with a softened contrastive loss for training an motion foundation model from wearable sensors. The learnable distance measure captures motif similarity and domain-specific semantic information such as rotation invariance. The learned distance provides a measurement of semantic similarity between a pair of accelerometer time-series segments, which is used to measure the distance between an anchor and various other sampled candidate segments. The self-supervised model is trained on 1 billion segments from 87,376 participants from a large wearables dataset. The model achieves strong performance across multiple downstream tasks, encompassing both classification and regression. To our knowledge, we are the first to show the generalizability of a self-supervised learning model with motion data from wearables across distinct evaluation tasks.

1 INTRODUCTION

Advances in self-supervised learning (SSL) combined with the availability of large-scale datasets have resulted in a proliferation of foundation models (FMs) in computer vision (Oquab et al., 2023), NLP (OpenAI et al., 2023), and speech understanding (Yang et al., 2024). These models provide powerful, general-purpose representations for a particular domain of data, and support generalization to a broad set of downstream tasks without the need for finetuning. For example, the image representation contained in the DINOv2 (Oquab et al., 2023) model was trained in an entirely self-supervised way and achieves state-of-the-art performance on multiple dense image prediction tasks such as depth estimation and semantic segmentation, by decoding a frozen base representation with task-specific heads. In contrast to these advances, the times-series have not yet benefited from the foundation model approach, with a few exceptions (Abbaspourazad et al., 2024; Das et al., 2023). This is particularly unfortunate for problems in mobile health (mHealth) signal analysis, which encompasses data modalities such as accelerometry, PPG, and ECG (Rehg et al., 2017), as the collection of mHealth data from participants can be time-consuming and expensive. However, recent advances in self-supervised learning for mHealth signals (Abbaspourazad et al., 2024; Yuan et al., 2024; Xu et al., 2024) have shown promising performance, raising the question of whether it is now feasible to train foundation models for mHealth signals.

In this paper, we demonstrate, for the first time, the feasibility of adopting a foundation model approach for the analysis of accelerometry data across tasks. Accelerometry is an important mHealth signal modality that is used in human activity recognition (HAR) (Haresamudram et al., 2022), physical health status assessment (Xu et al., 2022), energy expenditure estimation (Stutz et al., 2024), and gait assessment (Apple, 2021), among many other tasks. We use a novel method for self-supervised learning combined with pretraining on a large-scale accelerometry dataset. We show that this approach yields an effective representation for accelerometry data which is useful for multiple downstream tasks, including HAR and the estimation of gait metrics such as stride velocity and double support time. Crucially, we obtain state-of-the-art performance on these tasks without finetuning and also exceed the performance of fully-supervised models trained on smaller datasets.

The key to our approach is a novel method for contrastive representation learning which significantly extends the recent REBAR (Xu et al., 2024) architecture. Contrastive learning relies on the construction of positive and negative pairs that instruct the model to learn key differences and sim-

	RelCon (ours)	**REBAR** (Xu et al., 2024)	**AugPred** (Yuan et al., 2024a)	**SimCLR** (Tang et al., 2020)
Models **Within**-User Interactions	✓	✓		
Models **Between**-User Interactions	✓	✓		✓
Resistant to **False Positives**	✓		✓	✓
Resistant To **False Negatives**	✓			

Diagrams

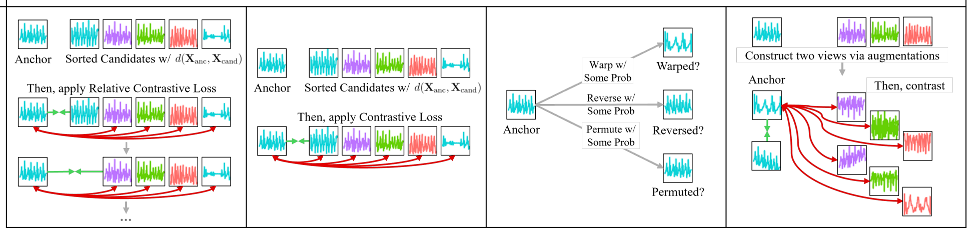

Figure 1: **Comparisons between various SOTA SSL approaches for Accelerometry data.** Each sequence color represents a different user's time-series. SimCLR resists false positives by creating positive pairs through augmentations, ensuring shared semantic meaning. It captures between-user interactions by treating other batch items as negatives, shown by differently colored sequences. RelCon captures within-user interactions by sampling from the same time-series, illustrated by the two blue candidates. Using a learnable distance function to rank within- and between-user candidates, RelCon's iterative loss function captures relative positions, offering better resistance to false positives and negatives compared to binary comparison methods like REBAR.

ilarities between selected time-series segments. Prior works have explored a variety of methods for constructing positive and negative pairs. One straightforward approach is to select segments based on temporal proximity (Abbaspourazad et al., 2024; Kiyasseh et al., 2021; Jeong et al., 2023a). Unfortunately, this heuristic can fail when segments that are far apart in time have a similar structure, as in the case for repetitive or periodic movements. Another popular approach is to construct positive pairs via data augmentation (Haresamudram et al., 2022; Matton et al., 2023; Gopal et al., 2021). This method, exemplified by SimCLR (Chen et al., 2020), has been shown to be extremely effective for computer vision tasks, where the rich invariant structure stemming from the imaging process provides many sources for augmentations. Unfortunately, it has proven to be challenging to define common augmentations for general time-series data that are relevant across diverse downstream tasks and encode meaningful invariances, without inadvertently altering important signal properties.

Our approach exploits the observation that many time-series of interest are typically composed of motifs, short distinctive temporal shapes that may approximately repeat themselves over time. For example, the upward swing of the arm, captured by a wristworn accelerometer during a bout of walking, will repeat at regular intervals depending on the cadence. Prior work has exploited this observation to learn a distance measure that captures motif similarity (Xu et al., 2024), and used this measure to identify positive and negative pseudo-labels. Since this approach forms pairs by matching based on signal shape, it's more likely to identify semantically meaningful pairs than approaches based on a heuristic like temporal distance, and since it can exploit natural variations in signal shapes across an extended time-series recording, it can identify within-person temporal dynamics, such as fatigue over time. Then, by introducing accelerometry-specific augmentations, we can improve the prior distance measure to learn to compare motifs, invariant to sensors positions or orientation. Another limitation of Xu et al. (2024) is that the contrastive loss is *hard* in the sense that all negative samples are treated equally in the loss computation. Accounting for the degree of similarity between the positive anchor sample and negative samples, via a *soft* loss, could imbue the embedding space with greater semantic meaning - for instance, having clusters of samples retain relative similarities like having 'walking' samples closer to 'running' than 'yoga' in an activity classification task. We introduce a novel RelCon method that accomplishes this goal and show that it significantly improves upon the performance of REBAR.

Our RelCon-trained motion foundation model achieves consistently strong performance across 6 different downstream datasets across 4 different types of downstream tasks for a wide set of metrics, outperforming current SOTA accelerometry SSL approaches. We are the first to show the generalizability of a self-supervised foundation model with motion data from wearables across each of these distinct evaluation tasks, ranging from activity classification to gait metric regression.

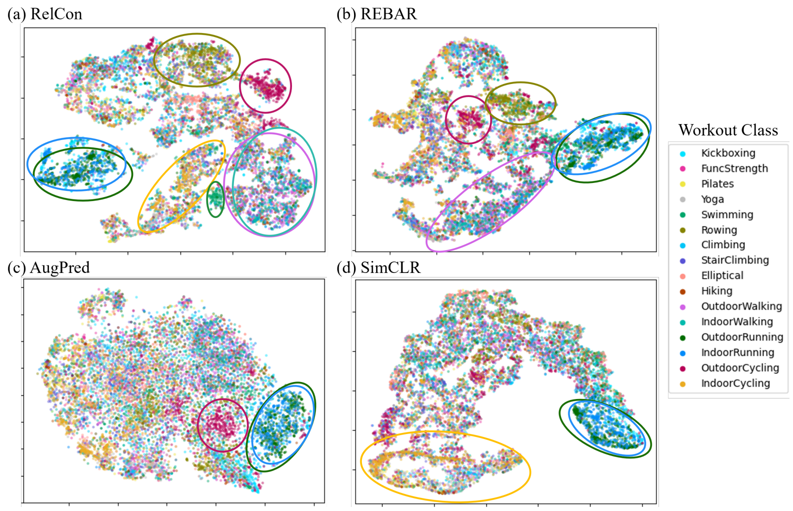

Figure 2: **t-SNE of representation spaces with perplexity=100.** Our RelCon approach has the clearest clusterings based on semantic classes, even forming a specific swimming cluster that is not seen in the other methods. RelCon can also better clearly separate between In/Outdoor Cycling.

2 Related Work

Time-series FM: We define foundation models to be representations that are pre-trained on broad-scale data and are capable of solving multiple diverse downstream tasks without fine-tuning (Bommasani et al., 2021) (e.g., each task uses frozen weights with supervised training of a light-weight prediction head). We believe we are the first to train a model for accelerometry data which meets these criteria. The closest related work is Yuan et al. (2024), which uses data from the UK Biobank to train an accelerometry representation. However, this work does not exhibit multi-task downstream performance, since they are all Human Activity Recogintion (HAR) variants, making it is less clear how well their representation captures the data domain. In contrast, we test our FM on multiple diverse tasks including HAR, workout classification, and gait analysis regression. Other notable FMs are Abbaspourazad et al. (2024); Das et al. (2023), but these do not model accelerometry.

Time-series SSL: Our RelCon architecture is a novel contrastive learning approach for time-series data. The closest related method is REBAR (Xu et al., 2024). We adopt REBAR's motif-based approach to generating positive pairs but their model was non-specific to accelerometry and thus only used a simple exact-motif-matching mechanism, which would not be invariant to changes in sensor position and other invariances. Additionally, due to its "hard" contrastive loss, it suffers from sensitivity to false positives and negatives because only one candidate is set to be positive and all other are set to be negative. Other prior works on SSL primarily adopt either data augmentations or masked auto-encoder approaches, including other prior works on HAR (Yuan et al., 2024; Haresamudram et al., 2022; 2024; Straczkiewicz et al., 2021). Additional works have addressed SSL for signals such as PPG and ECG, for applications including health condition predictions and sleep staging (Abbaspourazad et al., 2024; Song et al., 2024; Thapa et al., 2024; McKeen et al., 2024; Yuan et al., 2024; Kiyasseh et al., 2021; Diamant et al., 2022; Jeong et al., 2023b; Das et al., 2023).

In the motion wearables domain, most SSL approaches have focused on human activity recognition. Haresamudram et al. (2022) benchmarked a number of SSL approaches on accelerometer data, showing generalizability of pretraining approaches across datasets and sensor positions. Models for other motion-based tasks – including walking speed prediction (Yang & Li, 2012; He & Zhang, 2011; Shrestha & Won, 2018; Soltani et al., 2021), gait classification (Slemenšek et al., 2023; Brand et al., 2024), and health monitoring (Takallou et al., 2024) have focused on physics-based models, supervised learning, or SSL training for a single task.

Soft Label Learning Approaches: A feature of our RelCon approach is the use of soft labels to obtain more fine-grained characterizations of similarity. Most prior self-supervised contrastive methods for time-series have adopted the hard label approach, including the previous REBAR method (Xu et al., 2024) as well as (Yue et al., 2022), (Tonekaboni et al., 2021), (Zhang et al., 2022), and (Abbaspourazad et al., 2024). One exception is Lee et al. (2023), which proposed a soft contrastive learning approach for downstream classification and anomaly detection, using non-learnable time-series distance measures, including dynamic time warping (Müller, 2007).

3 METHODOLOGY

The RelCon methodology has two key components. The first includes several innovations to learn a better distance measure to capture accel semantic information in Section 3.1. The second component is a novel relative contrastive loss that encodes relative order relationships, presented in Section 3.2.

3.1 LEARNABLE DISTANCE MEASURE

Following prior work (Xu et al., 2024), we train a neural network to learn a distance measure to identify whether two sequences have similar temporal motifs (Schäfer & Leser, 2022) and are semantically similar. After training, the architecture is frozen and used as a static function to determine the relative similarities of candidate samples in the RelCon approach. The distance measure architecture is defined in Eq. 1 below:

$$d(\mathbf{X}_{\text{anc}}, \mathbf{X}_{\text{cand}}) \coloneqq \|\hat{\mathbf{X}}_{\text{anc}|\text{cand}} - \mathbf{X}_{\text{anc}}\|_2^2 \tag{1}$$

$$\hat{\mathbf{X}}_{\text{anc}|\text{cand}} = = \big((\text{CrossAttn}(\mathbf{X}_{\text{anc}}|\mathbf{X}_{\text{cand}})\mathbf{W}_o + \mathbf{b}_o) + \mu_{\text{cand}}\big)\sigma_{\text{cand}} \tag{2}$$

$$\text{CrossAttn}(\mathbf{x}_{\text{anc}}|\mathbf{X}_{\text{cand}}) = \sum_{\mathbf{x}_{\text{cand}} \in \mathbf{X}_{\text{cand}}} \underset{\mathbf{x}_{\text{cand}} \in \mathbf{X}_{\text{cand}}}{\text{sparsemax}}\Big(\text{sim}\big(f_q(\mathbf{x}_{\text{anc}}), f_k(\mathbf{x}_{\text{cand}})\big)\Big) f_v(\mathbf{x}_{\text{cand}}) \tag{3}$$

$$f_{\{q/k/v\}}(\mathbf{X}_{\{\text{anc/cand}\}}) = \text{DilatedConvNet}_{\{q/k/v\}}\left(\frac{\mathbf{X}_{\{\text{anc/cand}\}} - \mu_{\text{cand}}}{\sigma_{\text{cand}}}\right) \tag{4}$$

where $\mathbf{X} \in \mathbb{R}^{T \times D}$ and $\mathbf{x}, \mu, \sigma \in \mathbb{R}^D$ with T as the time length and D=3 for our 3-axis accelerometry signals. The distance between an anchor sequence and a candidate sequence, $d(\mathbf{X}_{\text{anc}}, \mathbf{X}_{\text{cand}})$, is defined as the reconstruction accuracy to generate the anchor from the candidate. The distance measure is strictly dependent on the motif similarities between the anchor and candidate that are captured in the dilated convolutions in $f_{\{q/k\}}$ (Xu et al., 2024).

Prior work with this distance measure only captured a simple exact-motif-matching mechanism because the original masked reconstruction training task can be solved by exact-matching the non-masked regions. To enhance the distance measure, we introduce 3 key innovations:

1. Use Accel-specific augmentations (Tang et al., 2020) during training to learn a motif-matching mechanism that is invariant to Accel-semantic-preserving transformations.

During training, we define $\mathbf{X}_{\text{cand}} \coloneqq Aug(\mathbf{X}_{\text{anc}})$, making the candidate an augmented version of the original sequence, using augmentations from Accel SimCLR (Tang et al., 2020). This helps the model learn to reconstruct the original anchor from semantically similar but altered candidates and prevents the exact-match solution, such as recognizing running signals even with changes like an upside-down wearable device or increased noise from a loose fitting device.

2. Replace the softmax in the cross-attention with a sparsemax formulation (Martins & Astudillo, 2016) in Eq. 3 to encourage precise motif comparison.

Sparsemax returns the euclidean projection of the unnormalized logits onto the probability simplex, encouraging sparsity (Martins & Astudillo, 2016). This prevents diffuse attention distributions that compare minor, irrelevant motifs within the anchor and candidate. Sparsemax ensures the model reconstructs with distinct features, enabling our measure to capture class-specific information.

3. Modify the reversible instance normalization (Kim et al., 2021) to normalize an anchor based upon the candidate, to preserve relative magnitude information.

The anchor sequence and final reconstruction output are normalized (in Eq. 4) and unnormalized (in Eq. 2) using candidate sequence statistics. When an anchor and candidate have drastically different magnitudes, it is more likely they represent different fundamental motions and should have worse reconstruction error. Reversible normalization helps preserve this effect.

3.2 RELATIVE CONTRASTIVE LOSS

ALL INSTANCES ARE POSITIVE ... BUT SOME ARE MORE POSITIVE THAN OTHERS

Normalized temperature cross entropy loss (NT-Xent is standard in contrastive learning (Eq. 5). It learns a class discriminative embedding space by pulling the anchor and positive instances closer together in the embedding space and pushing all negatives away from the anchor, as in:

$$\ell(\mathbf{X}_{\text{anc}}, \mathbf{X}_{\text{pos}}, \mathcal{S}_{\text{neg}}) = -\log \frac{\exp(\text{sim}(\mathbf{X}_{\text{anc}}, \mathbf{X}_{\text{pos}})/\tau)}{\sum_{\mathbf{X}_{\text{neg}} \in \mathcal{S}_{\text{neg}}} \exp(\text{sim}(\mathbf{X}_{\text{anc}}, \mathbf{X}_{\text{neg}})/\tau) + \exp(\text{sim}(\mathbf{X}_{\text{anc}}, \mathbf{X}_{\text{pos}})/\tau)} \tag{5}$$

Traditionally, NT-Xent uses strict labels that define absolute positive and negative sets. However, we would like to use the learned distance measure to modify the loss function to capture relative ordering among the candidates. That is, if $d(\mathbf{X}_{\text{anc}}, \mathbf{X}_i) > d(\mathbf{X}_{\text{anc}}, \mathbf{X}_j)$ the loss learns to preserve that relative ordering in the embedding space.

To do this, we redefine $\mathcal{S}_{\text{neg}} := f_{\text{neg}}$ in Eq. 6. Now the set of negatives is replaced with a function that creates different negative sets, depending on the current positive pairing. This enables us to construct our Relative Contrastive Loss function in Eq. 7. This loss function iterates across all candidates, $\mathbf{X}_{\text{i}} \in \mathcal{S}_{\text{cand}}$, and sets each candidate as positive, $\mathbf{X}_{pos} := \mathbf{X}_{\text{i}}$ before calculating the negative set. All candidates with a larger distance measure than this newly defined positive are defined to be the negative set, $\mathcal{S}_{\text{neg}} := f_{\text{neg}}(\mathbf{X}_{\text{anc}}, \mathbf{X}_{\text{i}}, \mathcal{S}_{\text{cand}})$, and then NT-Xent loss is calculated for this iteration.

$$f_{\text{neg}}(\mathbf{X}_{\text{anc}}, \mathbf{X}_{\text{pos}}, \mathcal{S}) = \{\mathbf{X} \in \mathcal{S} : d(\mathbf{X}_{\text{anc}}, \mathbf{X}) > d(\mathbf{X}_{\text{anc}}\} \tag{6}$$

$$\mathcal{L}_{\text{RelCon}} = \sum_{\mathbf{X}_{\text{i}} \in \mathcal{S}_{\text{cand}}} \ell(\mathbf{X}_{\text{anc}},\ \mathbf{X}_{pos} := \mathbf{X}_{\text{i}},\ \mathcal{S}_{\text{neg}} := f_{\text{neg}}(\mathbf{X}_{\text{anc}}, \mathbf{X}_{\text{i}}, \mathcal{S}_{\text{cand}})) \tag{7}$$

In RelCon, our pool of candidates originates from 2 sources: 1) sampling within the user, across time and 2) sampling within the batch. In (1), we choose c random subsequences from the same user as the anchor sequence to be contrasted. In our experiments, c=20. This sampling method helps capture within-person temporal dynamics, such as how a user's motion signals may indicate fatigue over time. Sampling within the batch in (2) allows the model to learn similarities and differences across other users. An visualization of our RelCon loss procedure can be found in Fig. 1.

4 Experimental Design

4.1 Foundation Model Pretraining

We trained models on Inertial Movement Unit (IMU) sensors from the Apple Heart & Movement Study (AHMS) (MacRae, 2021). AHMS is an ongoing research study designed to explore the links between physical activity and cardiovascular health, which is sponsored by Apple and conducted in partnership with American Heart Association and Brigham and Women's Hospital. To be eligible for the study, participants must — among other eligibility criteria – be at least 18 years of age (at least 19 in Alabama and Nebraska; at least 21 in Puerto Rico), reside in the United States, have access to an Apple Watch, and provide informed consent electronically in the Apple Research app.

The training data included a subset of study data with 87,376 participants recorded over one day, with a 10/3/3 train/val/test split. Following prior convention (Reyes-Ortiz et al., 2015), we use 2.56 seconds of the raw 100 Hz 3-axis x,y,z accelerometry signal of the IMU sensor in the wearable device as input to our embedding model. Each model was pre-trained with 8 x A100 GPUs for 24 hours. Models iterated over a total of 1 billion samples, and a total of $\sim$30k unique participant-days.

We evaluated RelCon along with other commonly used methods in SSL with Accel for comparison:

- *Accel SimCLR* (Tang et al., 2020; Chen et al., 2020) contrasts positive pairs formed by accel-specific augmentations (i.e. 3D rotation) and has been shown to have state-of-the-art performance in the accel domain (Haresamudram et al., 2022).
- *Augmentation Prediction* (Yuan et al., 2024; Haresamudram et al., 2022) predicts whether an accel-specific augmentation was applied to each sample.
- *Accel REBAR* (Xu et al., 2024) uses a contrastive loss where positive pairs are identified by a learned motif-similarity. Instead of using the original version of REBAR designed for generalized time-series data, we utilize it with the accel-focused innovations proposed in Section 3.1.

For each of these methods, we used a 1D ResNet-34 encoder backbone that used average pooling to generate a 256-dimensional embedding vector (3.9M parameters).

4.2 Downstream Evaluation

First, in order to assess generalizable of our learned embedding, we evaluate our models with our **Task Diversity Evaluation** with two different datasets , in which we compare the RelCon foundation model against a set of diverse downstream tasks and to self-supervised models trained on our data with three other methods that have seen strong performance accel data: SimCLR, AugPred, and REBAR. We use the embeddings with linear regression for various gait metrics, and with linear probe classification on both the subsequence and workout-level.

Next, in order to compare against prior work, we evaluate our RelCon foundation model with our **Benchmarking Evaluation** with four additional classification datasets . Specifically, we compare against another large-scale pre-trained accel foundation model (Yuan et al., 2024) and against an accel SSL benchmarking study (Haresamudram et al., 2022), with comparisons to seven distinct self-supervised learning methods . These datasets help evaluate model generalizability under data distribution shifts, including differing sensor positions and inference window lengths.

In total, we compare against 6 different downstream datasets across 4 different types of downstream tasks. We compare our RelCon foundation model against 11 models total: 3 pre-trained from scratch in the *Task Diversity Evaluation* and 8 from the prior literature in the *Benchmarking Evaluation*.

4.2.1 Task Diversity Evaluation Datasets and Set-up

Gait Metric Regression: We used a dataset including gait mat collection to evaluate models on gait metric regression (Apple, 2021). All participants completed proctored overground walking tasks with two mobile devices with IMU sensors at each side of the body in different locations (i.e. at the hip, in a front or back pocket, or in a waist bag). We used the Design Cohort A participants, which included 359 participants with an average age of 74.7, who provided informed consent. Participants were instructed to conduct 4 different walking tasks: 1) one lap at self-selected speed, 2) four laps at an instructed slow speed, and 3) as many laps as possible within 6 minutes. Each walking task was conducted along a 12-meter straight-line course, with an 8-meter pressure mat placed centrally and various statistics about each participant's gait were calculated: step count, walking speed, step length, double support time, and walking asymmetry. Models were evaluated on predicting double support time (DST) and stride velocity.

Each 2.56s subsequence was matched to the lap aggregated target (e.g. total number of steps or average walking speed). The participants were assigned into 50% train and 50% test randomly based on participant ID, where every lap for a given participant falls into the same split. The training split was used to train linear regression probes on embeddings from the self-supervised models. Metrics were selected following a prior report (Apple, 2021): mean squared error (MSE), std dev of squared error (SDSE), mean absolute error (MAE), std dev of absolute error (SDAE), and Pearson's Correlation Coefficient. Mean and std devs were calculated by aggregating predictions across all inputs of a given user and are related to bias and variance respectively. Correlation is used in order to assess how each user's average gait metric corresponds to ground truth values. Ranges for each metric were calculated by retraining the linear regression probe five times with different set seeds.

Activity Classification: We evaluated activity classification performance using self-reported activity labels gathered from data in AHMS. We used a subset of data with $\sim$2k total users across 14 workouts (full list in Table 2). The 14 selected workouts captured a range of diverse activities (e.g. kickboxing and rowing) that are non-trivial to separate (i.e. outdoor cycling vs. indoor cycling). For evaluation, workouts were class balanced so each included $\sim$22 hours of data, for a total of 310 workout hours. We used a 4/1/5 train/val/test split based on participant ID. We ensured the subset of data used for classification did not overlap with the pre-training dataset, preventing any data leakage.

Embeddings are generated for each 2.56s-long subsequence. We evaluate classification on both the *subsequence-level* as well as the *workout-level*. At the workout-level, we predict workout classes across each of 2.56s subsequences within the workout, and select the most frequently predicted workout. Including two prediction scales allow us to evaluate provides additional context about how information learned by the embedding interacts with time aggregation. We report F1, Kappa, Accuracy, and macro AUC metrics following previous work (Haresamudram et al., 2022; Yuan et al., 2024; Xu et al., 2024). Ranges per metric are obtained by retraining the probe five times and calculating the mean and standard deviation.

	Gait Metric Regression (Wrist→Waist)									
	Stride Velocity					Double Support Time				
Model	↓ MSE	↓ SDSE	↓ MAE	↓ SDAE	↑ Corr	↓ MSE	↓ SDSE	↓ MAE	↓ SDAE	↑ Corr
SimCLR	.0121 ± .0002	.0234 ± .0019	.0827 ± .0020	.0726 ± .0007	**.8454 ± .0133**	.0016 ± .0001	.0025 ± .0001	.0317 ± .0009	.0254 ± .0006	.7074 ± .0136
Aug Pred	.0144 ± .0002	.0241 ± .0003	.0940 ± .0007	.0749 ± .0005	.7950 ± .0035	.0015 ± .0000	.0025 ± .0001	.0296 ± .0002	.0251 ± .0003	.7214 ± .0029
REBAR	.0147 ± .0010	.0285 ± .0035	**.0818 ± .0042**	.0818 ± .0042	.7853 ± .0178	.0016 ± .0001	.0026 ± .0001	.0316 ± .0011	.0254 ± .0008	.6817 ± .0286
RelCon (ours)	**.0115 ± .0005**	**.0190 ± .0013**	.0833 ± .0019	**.0678 ± .0016**	.8431 ± .0039	**.0014 ± .0000**	**.0024 ± .0001**	**.0275 ± .0004**	**.0249 ± .0004**	**.7559 ± .0120**

Table 1: **Results of Gait Metric Regression**. Mean and standard deviations of Squared Error and Absolute Error were calculated by aggregating predictions across given user and are related to bias and variance respectively. Strong consistent performance of RelCon shows that our model is able to do well on capturing user-specific information for regression.

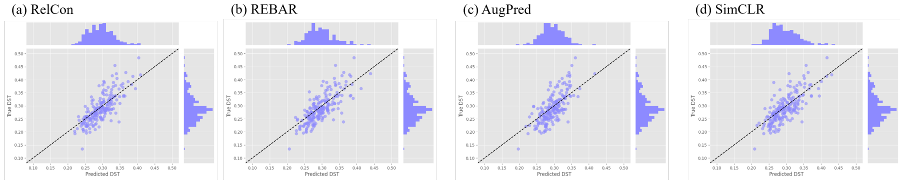

Figure 3: **Plot of correlation between predicted and true DST**. RelCon has the highest correlation.

4.2.2 Benchmarking Evaluation Datasets and Set-up

Comparison to a Large-Scale Pre-trained Accel Model: Yuan et al. (2024) proposes an Accel foundation model with large-scale pre-training on the UK BioBank (Bycroft et al., 2018) dataset. They are able to train over ∼100k unique participants for a total of ∼700k unique participant-days. Their foundation model is based upon an augmentation prediction approach with a ResNet-18 backbone, and they use it to train and embed 10s subsequences of a 3-axis accelerometry signal collected at the wrist. Please refer to the original work for more details (Yuan et al., 2024).

Following their evaluation methodology, we directly compare our RelCon foundation model against their model along three different dimensions: fine-tuning the entire model, using an MLP probe, and training from scratch. We exactly mimic their cross-validation splits on the Opportunity (Roggen et al., 2010) and PAMAP2 (Reiss, 2012) activity classification datasets and use these splits to generate the metric ranges. We also match their 10s length used for inference.

Comparison to an Accel SSL Benchmarking Study: Haresamudram et al. (2022) seeks to assess the current state of the accelerometry self-supervised learning field, in the context of activity classification. To this end, they benchmark a diverse range of self-supervised, including our aforementioned accel-specific methods, Accel SimCLR (Tang et al., 2020) and Augmentation Prediction (Saeed et al., 2019), as well as generalized SSL methods, such as SimSiam (Chen & He, 2021) and BYOL (Grill et al., 2020). Each of their SSL approaches have a unique backbone architecture that corresponds to their original work. They pre-train their SSL methods on the 3-axis wrist-mounted accelerometry signals from Capture-24 dataset (Chan Chang et al., 2021), which is composed of 151 unique participants for a total of ∼4k participant-hours.

Similar to our RelCon method, each of their SSL methods are trained on wrist data, but for evaluation, they use a different activity classification datasets that each correspond to accelerometry signals collected at different positions. We compare our RelCon FM with a downstream dataset at each benchmarked position: HHAR (Blunck et al., 2015) for Wrist, Motionsense for Waist (Malekzadeh et al., 2018), and PAMAP2 (Reiss, 2012) for leg. In this way, each of their methods, as well as our RelCon FM, will have pre-trained on wrist accelerometry data and then evaluate on potentially different sensor position. We adopt the same exact cross-validation splits in the original work and use them to generate our metric ranges. We match their 2s length used for inference.

5 Results and Discussion

5.1 Task Diversity Evaluation Results

Gait metric regression: Table 1 shows gait metric regression performance of each SSL method. RelCon had the strongest consistent performance across *both* gait metrics, double support time and stride velocity. While methods based on prior work with Accel SimCLR (Haresamudram et al., 2022; Tang et al., 2020) were not developed with a focus on gait, they still achieve strong perfor-

	AHMS Classification (Wrist→Wrist)							
	Subsequence Level				Workout Level			
Model	↑ F1	↑ Kappa	↑ Acc	↑ AUC	↑ F1	↑ Kappa	↑ Acc	↑ AUC
SimCLR	36.35 ± .42	**39.32 ± .29**	37.32 ± .29	.8372 ± .0004	50.06 ± .44	49.66 ± .33	53.09 ± .30	.8855 ± .0033
Aug Pred	34.48 ± .08	30.69 ± .12	35.02 ± .11	.8175 ± .0003	54.63 ± .79	52.44 ± .65	55.71 ± .60	.9049 ± .0010
REBAR	36.39 ± .28	32.75 ± .23	36.96 ± .21	.8334 ± .0011	50.29 ± .78	48.53 ± .91	51.94 ± .87	.8775 ± .0035
RelCon (ours)	**38.56 ± .23**	35.10 ± .20	**39.15 ± .19**	**.8417 ± .0009**	**55.28 ± .86**	**53.87 ± .71**	**56.94 ± .68**	**.9078 ± .0016**

Table 2: **Results on Activity Classification**. RelCon achieves consistently strong performance when evaluated both at the subsequence-level and workout-level. SimCLR does well at the subsequence-level, but performs much worse at the workout-level. The opposite is true for AugPred. This show that there exist nuances captured by our models at the subseq- vs. workout-level.

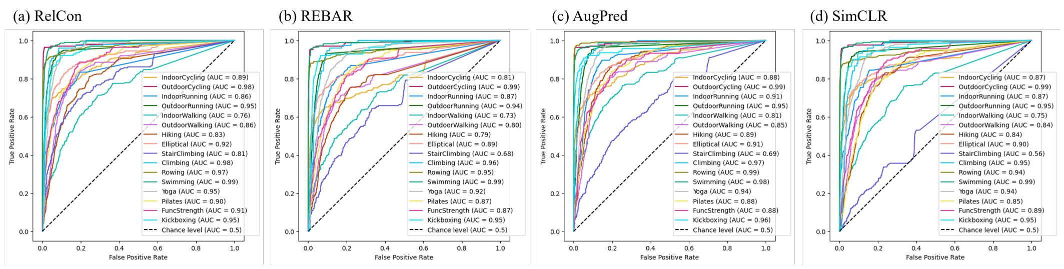

Figure 4: **ROC curves of AHMS Workout Level Classification Task**. Amongst other performance improvements, unlike the other approaches, RelCon is able to clearly better classify stair climbing.

mance, particularly for stride velocity. Although REBAR also uses a learnable distance, it generally has lower performance than RelCon. RelCon's relative constrastive loss seems to improve the model's ability to capture fine-scale differences between subsequences of differing walking speeds or DSTs.

Activity classification: Table 2 summarizes activity classification results on the activity classification dataset described in Sec. 4.2.1. RelCon embeddings had distinct cluster separability, highlighted in the its t-SNE visualization in Fig. 2, and further detailed confusion matrics can be found in Appendix A.3. SimCLR particularly had low performance in workout-level prediction of stair climbing, which was frequently confused as climbing (Fig. 4 and 5). SimCLR has worse performance in workout-level classification metrics compared to the subseq-level ones, and the opposite is true for AugPred. RelCon is consistently strong across both subseq and workout-level. RelCon is better at classifying stair climbing compared to the other methods, which will often confuse stair climbing with climbing or elliptical (Fig. 4 and 5). These three classes have slower, deliberate hand swinging motions, but we hypothesize that RelCon's resistance to false positives and negatives enables the model to capture subtle differences. Additionally, RelCon can better predict outdoor running from indoor running. We hypothesize that RelCon can better identify when the running pattern is more uniform, implying that the running is perhaps being done indoors.

			Opportunity (Wrist→Wrist)		PAMAP2 (Wrist→Wrist)	
	Eval Method	Pre-train Data	↑ F1	↑ Kappa	↑ F1	↑ Kappa
RelCon FM	MLP Probe	AHMS	**69.1 ± 8.3**	**62.1 ± 6.2**	**85.38 ± 3.6**	**84.73 ± 3.6**
Yuan et al. (2024)'s FM	MLP Probe	UKBioBank	57.0 ± 7.8	43.5 ± 9.2	72.5 ± 5.4	71.7 ± 5.7
RelCon FM	Fine-tuned	AHMS	**98.4 ± 0.9**	**97.9 ± 0.8**	**98.81 ± 1.3**	**98.6 ± 1.6**
Yuan et al. (2024)'s FM	Fine-tuned	UKBioBank	59.5 ± 8.5	47.1 ± 10.4	78.9 ± 5.4	76.9 ± 5.9
RelCon FM	From Scratch	n/a	94.0 ± 1.3	92.8 ± 1.8	97.5 ± 1.1	97.0 ± 1.3
Yuan et al. (2024)'s FM	From Scratch	n/a	38.3 ± 12.4	23.8 ± 15.4	60.5 ± 8.6	59.6 ± 8.6

Table 3: **RelCon FM compared to another Large-Scale Pre-trained Accel Model (Yuan et al., 2024)**. Although the RelCon FM embeds time-series into 256 dim vectors, smaller than Yuan et al. (2024)'s 1024 dim vector, it achieves stronger MLP probe performance. Additionally, fine-tuning our ResNet-34-based RelCon improves upon the from scratch performance, demonstrating the utility of our approach. Results on Yuan et al. (2024) are quoted from the original work.

			HHAR (Wrist→Wrist)	Motionsense (Wrist→Waist)	PAMAP2 (Wrist→Leg)
			↑ F1	↑ F1	↑ F1
		RelCon FM	**57.63 ± 3.24**	80.35 ± 0.71	**53.98 ± 0.76**
Self-supervised w/ Frozen Embedding + Linear Probe	Haresamudram et al. (2022)	Aug Pred	50.95 ± 2.70	74.96 ± 1.37	46.90 ± 1.14
		SimCLR	55.93 ± 1.75	**83.93 ± 1.78**	50.75 ± 2.97
		SimSiam	45.36 ± 4.98	71.91 ± 12.3	47.85 ± 2.48
		BYOL	40.66 ± 4.08	66.44 ± 2.76	43.89 ± 3.35
		MAE	43.48 ± 2.84	61.14 ± 3.45	42.32 ± 1.63
		CPC	56.24 ± 0.98	72.89 ± 2.06	45.84 ± 1.39
		Autoencoder	53.57 ± 1.14	55.12 ± 3.46	50.79 ± 1.09
Fully Supervised		DeepConvLSTM	54.39 ± 2.28	84.56 ± 0.85	51.22 ± 1.91
		Conv classifier	55.43 ± 1.21	**89.25 ± 0.50**	**59.76 ± 1.53**
		LSTM classifier	37.42 ± 5.04	86.74 ± 0.29	48.61 ± 1.82

Table 4: **RelCon FM compared to the Accel SSL Benchmarking Study (Haresamudram et al., 2022)**. Results show the strength of the RelCon foundation model, which outperforms the fully-supervised models when the pre-training and target domains are matched to both be wrist data. In MotionSense, there is a sizable gap between SimCLR, our FM and the remaining SSL methods. Results on Haresamudram et al. (2022) are quoted from the original work.

5.2 Benchmarking Evaluation Results

Results from Comparison to a Large-Scale Pre-trained Accel Model: Table 3 shows that RelCon achieved a stronger performance across all datasets and evaluation methods. We used a ResNet-34 based architecture with an embedding dimension of 256 and 3.96M parameters, while Yuan et al. (2024) used a ResNet-18 based architecture with an embedding dimension of 1024 and 10M parameters. The smaller and deeper architecture is more effective in this case, as seen in the "From Scratch" results in Table 3. Although our learned embedding vector had a lower dimensionality, the RelCon FM has a stronger performance with the MLP probe (69.1% vs. 57.0% and 85.38% vs. 72.5% for Opportunity and PAMAP2 respectively). This suggests that the representations learned by the RelCon approach are able to efficiently capture the most important properties of the sequences. The fine-tuning evaluation shows that the initialization provided by our RelCon pre-training is able to boost performance even further with 94.0% $\rightarrow$ 98.4% and 97.5% $\rightarrow$ 98.8% improvements for Opportunity and PAMAP2 F1 metrics. This demonstrates the value of our SSL training methodology.

Results from Comparison to an Accel SSL Benchmarking Study: Table 4 reports results for the RelCon FM alongside results from Haresamudram et al. (2022). The results show the strong generalizability of our RelCon FM, even when there is a sensor position mismatch between training and evaluation, as seen in PAMAP2. When the target dataset's sensor position matches the pre-training domain, as seen in HHAR, RelCon acheives exceptionally strong performance, even beating the fully-supervised methods. Although SimCLR outperforms RelCon in Motionsense, there is a considerable gap to the rest of the methods.

5.3 Ablation Study Results

The Importance of Accel Augmentations: Invariances can capture conceptually important information for motion data, and augmentations have been used extensively in prior approaches towards capturing these invariances (Yurtman & Barshan, 2017; Florentino-Liaño et al., 2012; Zhong & Deng, 2014). For example, learning 3D rotation invariance allows the embeddings to be robust to how different users may wear or carry an accelerometry device. In our ablation study in Table 5, the *w/o Augmentations* results were generated by removing the augmentations from the training of the learnable distance measure. Removing augmentations had a strong negative impact on results, across tasks. RelCon uses augmentations to learn a motif-similarity based distance metric that is used with a relative loss. While prior approaches using augmentations learn only invariances (or variances, in the case of augmentation prediction), RelCon incorporates learned invariances along with motifs and within-person dynamics, enabling the model to learn diverse motion properties and better generalize across tasks.

Within-Person Dynamics in Time-series: Variations in time-series signals can be related to both within-user dynamics and between-user dynamics. Comparing subsequences within a given user

	Velocity	DST	AHMS-Subseq	AHMS-Workout	HHAR	PAMAP2 MLP
	↑ Corr	↑ Corr	↑ F1	↑ F1	↑ F1	↑ F1
RelCon	0.8431	0.7559	38.56	55.28	57.63	85.38
w/o Augmentations	-5.42%	-13.88%	-12.5%	-17.93%	-8.92%	-15.12%
w/o RevIN	-11.66%	-6.01%	-3.81%	-4.11%	-1.61%	-4.65%
w/o SparseMax	-0.87%	-3.7%	-5.06%	-7.11%	-4.34%	-0.62%
w/o Sampling Within-Subject	-1.4%	-1.4%	-4.49%	-7.25%	2.24%	-0.93%

Table 5: **Ablations of RelCon components across tasks**. The decrease in performance upon applying ablations shows the importance of key components of the RelCon approach. The large decrease in performance when augmentations are removed from the metric learning stage highlights the importance of learning a motif-similarity metric that is robust to invariances.

while learning the embedding can enable the learning of important within-user dynamics, which we ablate in *w/o Sampling Within-Subject* in Table 5. Here, we no longer draw candidates from sampling within the user, across time. Instead, our candidates are now only the other members within the batch, as well as a augmented version of the anchor, which mimics the candidates available to SimCLR. Afterwards, we follow the standard RelCon procedure, ranking the candidates and applying the RelCon loss. We see that this ablation results in a 4.49% drop in subsequence-level classification (AHMS-Subseq) by 4.49%, and by an even larger 7.25% for the workout-level. This demonstrates the importance of capturing within-user dynamics over time in order to achieve better performance for the longer-scale workout-level classifications. This intuition is reflected in Table 2, in which SimCLR, which does not capture within-user interactions, has relatively stronger performance in the subsequence-level metrics compared to the workout-level metrics.

Relative Rankings Soften the Impact of False Negatives/Positives: REBAR and RelCon each use a learnable distance to act as a static pseudo-labeling function to identify positives and negatives. However, out of all of the possible candidates, REBAR only identifies one candidate as positive and all others as negative. This unfortunately leads itself to false positives and false negatives. A false positive would occur when our distance measure identifies a candidate as positive, even when it originates from a different class than the anchor. False negatives occurs when there are multiple positive candidates within the pool of candidates, but only one of them is chosen to be a positive and the rest to be negative. This is an issue even across methods. SimCLR draws its negatives from the randomly chosen batch, and there is no guarantee that items in the batch are all semantically distinct from the anchor. Fortunately, RelCon's learning approach allows us to capture how "positive" a candidate is relative to the other candidates, thereby lessening the effect of false pseudo-labels. In Tables 1 and 2, although both REBAR and RelCon used the same exact learned distance function, RelCon's relative contrastive loss approach led to consistently stronger results, demonstrating the usefulness of relative positive rankings.

Refining our Learned Distance: Table 5 shows that if we remove the reversible instance normalization (*RevIN*) component or replace the *SparseMax* with the traditional softmax operator within our learnable distance measure, then performance of our RelCon approach suffers. Removal of RevIN in particular causes a 11.66% drop in the velocity regression correlation, which shows the importance of capturing relative magnitude for predicting velocity.

6 Conclusions

RelCon FM embeddings have strong performance across diverse tasks including motion velocity, double support time, and workout classification and across datasets, as compared to benchmarked approaches and models. This consistently strong performance of the frozen embeddings with simple probes shows that RelCon embeddings capture fundamental properties of motion relevant across tasks. To capture these motion properties, we introduced a number of novel modifications to our learned motif-based distance learning, including introducing augmentations to learn sensor invariances and using sparsemax to increase precision. We contribute a methodology for using the learned distance measure with a relative contrastive loss, in order to learn fine-grained interactions between and within users. Our RelCon foundation model showed the strongest, consistent performance in across all evaluated tasks and datasets, highlighting the value of the proposed approach.

7 REPRODUCIBILITY

We will release the code for our RelCon foundation model publicly upon acceptance. This will include the RelCon training methodology, architecture, as well as the reproducible evaluation code from our “Benchmarking Evaluation” task, which utilizes public datasets. We believe that this will greatly benefit the broader community by providing code to easily re-train the RelCon approach in addition to providing a unified benchmarking task to guide model development.

REFERENCES

Salar Abbaspourazad, Oussama Elachqar, Andrew Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=pC3WJHf51j`.

Apple. Lock or unlock apple watch. URL `https://support.apple.com/guide/watch/lock-or-unlock-apple-watch-apd0e1e73b6f/watchos`.

Apple, May 2021. URL `https://www.apple.com/ca/healthcare/docs/site/Measuring_Walking_Quality_Through_iPhone_Mobility_Metrics.pdf`.

Henrik Blunck, Sourav Bhattacharya, Thor Prentow, Mikkel Kjrgaard, and Anind Dey. Heterogeneity Activity Recognition. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5689X.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Yonatan E Brand, Felix Kluge, Luca Palmerini, Anisoara Paraschiv-Ionescu, Clemens Becker, Andrea Cereatti, Walter Maetzler, Basil Sharrack, Beatrix Vereijken, Alison J Yarnall, et al. Self-supervised learning of wrist-worn daily living accelerometer data improves the automated detection of gait in older adults. *Scientific Reports*, 14(1):20854, 2024.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O’Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Rhiannon A Campbell, Elizabeth J Bradshaw, Nick Ball, Adam Hunter, and Wayne Spratford. Effects of digital filtering on peak acceleration and force measurements for artistic gymnastics skills. *Journal of Sports Sciences*, 38(16):1859–1868, 2020.

S Chan Chang, R Walmsley, J Gershuny, T Harms, E Thomas, K Milton, P Kelly, C Foster, A Wong, N Gray, et al. Capture-24: Activity tracker dataset for human activity recognition. 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Yuqing Chen and Yang Xue. A deep learning approach to human activity recognition based on single accelerometer. In *2015 IEEE international conference on systems, man, and cybernetics*, pp. 1488–1492. IEEE, 2015.

Jen-Tzung Chien and Kuan Chen. False negative masking for debiasing in contrastive learning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, 2024.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron D Aguirre, Collin M Stultz, and Puneet Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLoS computational biology*, 18(2):e1009862, 2022.

Ehab Essa and Islam R Abdelmaksoud. Temporal-channel convolution with self-attention network for human activity recognition using wearable sensors. *Knowledge-Based Systems*, 278:110867, 2023.

Blanca Florentino-Liaño, Niamh O'Mahony, and Antonio Artés-Rodríguez. Human activity recognition using inertial sensors with invariance to sensor orientation. In *2012 3rd International Workshop on Cognitive Information Processing (CIP)*, pp. 1–6. IEEE, 2012.

Sara García-de Villa, David Casillas-Pérez, Ana Jiménez-Martín, and Juan Jesús García-Domínguez. Inertial sensors for human motion analysis: A comprehensive review. *IEEE Transactions on Instrumentation and Measurement*, 72:1–39, 2023.

Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine Learning for Health*, pp. 156–167. PMLR, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

W3C Working Group. Motion sensors explainer, Aug 2017. URL `https://www.w3.org/TR/motion-sensors/`.

Harish Haresamudram, Irfan Essa, and Thomas Plötz. Assessing the state of self-supervised human activity recognition using wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6 (3), sep 2022. doi: 10.1145/3550299. URL `https://doi.org/10.1145/3550299`.

Harish Haresamudram, Chi Ian Tang, Sungho Suh, Paul Lukowicz, and Thomas Ploetz. Solving the sensor-based activity recognition problem (soar): self-supervised, multi-modal recognition of activities from wearable sensors. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, pp. 759–761, 2023.

Harish Haresamudram, Irfan Essa, and Thomas Ploetz. Towards learning discrete representations via self-supervision for wearables-based human activity recognition. *Sensors*, 24(4):1238, 2024.

Zhenyu He and Wei Zhang. Estimation of walking speed using accelerometer and artificial neural networks. In *International workshop on computer science for environmental engineering and ecoinformatics*, pp. 42–47. Springer, 2011.

Xinxin Huang, Yunan Xue, Shuyun Ren, and Fei Wang. Sensor-based wearable systems for monitoring human motion and posture: A review. *Sensors*, 23(22):9047, 2023.

Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2785–2795, 2022.

Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. Collossl: Collaborative self-supervised learning for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–28, 2022.

Hyewon Jeong, Nassim Oufattole, Aparna Balagopalan, Matthew Mcdermott, Payal Chandak, Marzyeh Ghassemi, and Collin Stultz. Event-based contrastive learning for medical time series. *arXiv preprint arXiv:2312.10308*, 2023a.

Hyewon Jeong, Collin M Stultz, and Marzyeh Ghassemi. Deep metric learning for the hemodynamics inference with electrocardiogram signals. *arXiv preprint arXiv:2308.04650*, 2023b.

Xiyuan Jin, Jing Wang, Lei Liu, and Youfang Lin. Time-series contrastive learning against false negatives and class imbalance. *arXiv preprint arXiv:2312.11939*, 2023.

Sungyeon Kim, Minkyo Seo, Ivan Laptev, Minsu Cho, and Suha Kwak. Deep metric learning beyond binary supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2288–2297, 2019.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2021.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pp. 5606–5615. PMLR, 2021.

Seunghan Lee, Taeyoung Park, and Kibok Lee. Soft contrastive learning for time series. *arXiv preprint arXiv:2312.16424*, 2023.

Calum A. MacRae. Apple heart & movement study, Jul 2021. URL `https://clinicaltrials.gov/study/NCT04198194`.

Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. Protecting sensory data against sensitive inferences. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems*, pp. 1–6, 2018.

Almontazer Mandong and Usama Munir. Smartphone based activity recognition using k-nearest neighbor algorithm. In *Proceedings of the International Conference on Engineering Technologies, Konya, Turkey*, pp. 26–28, 2018.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pp. 1614–1623. PMLR, 2016.

Katie Matton, Robert Lewis, John Guttag, and Rosalind Picard. Contrastive learning of electrodermal activity representations for stress detection. In *Conference on Health, Inference, and Learning*, pp. 410–426. PMLR, 2023.

Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model, 2024. URL `https://arxiv.org/abs/2408.05178`.

Jamie McQuire, Paul Watson, Nick Wright, Hugo Hiden, and Michael Catt. Uneven and irregular surface condition prediction from human walking data using both centralized and decentralized machine learning approaches. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1449–1452. IEEE, 2021.

Max Moebus, Lars Hauptmann, Nicolas Kopp, Berken Demirel, Björn Braun, and Christian Holz. Nightbeat: Heart rate estimation from a wrist-worn accelerometer during sleep. *arXiv preprint arXiv:2411.00731*, 2024.

Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pp. 69–84, 2007.

R OpenAI et al. Gpt-4 technical report. *ArXiv*, 2303:08774, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

James M Rehg, Susan A Murphy, and Santosh Kumar. *Mobile Health: Sensors, Analytic Methods, and Applications*. Springer, 2017. doi: 10.1007/978-3-319-51394-2.

Attila Reiss. PAMAP2 Physical Activity Monitoring. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5NW2H.

Jorge Reyes-Ortiz, Davide Anguita, Luca Oneto, and Xavier Parra. Smartphone-Based Recognition of Human Activities and Postural Transitions. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C54G7M.

Daniel Roggen, Alberto Calatroni, Ricardo Nguyen-Dinh, Long-Vanand Chavarriaga, and Hesam Sagha. OPPORTUNITY Activity Recognition. UCI Machine Learning Repository, 2010. DOI: https://doi.org/10.24432/C5M027.

Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–30, 2019.

Patrick Schäfer and Ulf Leser. Motiflets: Simple and accurate detection of motifs in time series. *Proceedings of the VLDB Endowment*, 16(4):725–737, 2022.

Ian Shapiro, Jeff Stein, Calum MacRae, and Michael O'Reilly. Pulse oximetry values from 33,080 participants in the apple heart & movement study. *NPJ Digital Medicine*, 6(1):134, 2023.

Aawesh Shrestha and Myounggyu Won. Deepwalking: Enabling smartphone-based walking speed estimation using deep learning. In *2018 IEEE global communications conference (GLOBECOM)*, pp. 1–6. IEEE, 2018.

Jan Slemenšek, Iztok Fister, Jelka Geršak, Božidar Bratina, Vesna Marija van Midden, Zvezdan Pirtošek, and Riko Šafarič. Human gait activity recognition machine learning methods. *Sensors*, 23(2):745, 2023.

Abolfazl Soltani, Nazanin Abolhassani, Pedro Marques-Vidal, Kamiar Aminian, Peter Vollenweider, and Anisoara Paraschiv-Ionescu. Real-world gait speed estimation, frailty and handgrip strength: a cohort-based study. *Scientific reports*, 11(1):18966, 2021.

Junho Song, Jong-Hwan Jang, Byeong Tak Lee, DongGyun Hong, Joon myoung Kwon, and Yong-Yeon Jo. Foundation models for electrocardiograms, 2024. URL https://arxiv.org/abs/2407.07110.

Marcin Straczkiewicz, Peter James, and Jukka-Pekka Onnela. A systematic review of smartphone-based human activity recognition methods for health research. *NPJ Digital Medicine*, 4(1):148, 2021.

Jan Stutz, Philipp A Eichenberger, Nina Stumpf, Samuel EJ Knobel, Nicholas C Herbert, Isabel Hirzel, Sacha Huber, Chiara Oetiker, Emily Urry, Olivier Lambercy, et al. Energy expenditure estimation during activities of daily living in middle-aged and older adults using an accelerometer integrated into a hearing aid. *Frontiers in Digital Health*, 6:1400535, 2024.

Sungho Suh, Vitor Fortes Rey, and Paul Lukowicz. Tasked: transformer-based adversarial learning for human activity recognition using wearable sensors via self-knowledge distillation. *Knowledge-Based Systems*, 260:110143, 2023.

Mohammad Ali Takallou, Farahnaz Fallahtafti, Mahdi Hassan, Ali Al-Ramini, Basheer Qolomany, Iraklis Pipinos, Sara Myers, and Fadi Alsaleem. Diagnosis of disease affecting gait with a body acceleration-based model using reflected marker data for training and a wearable accelerometer for implementation. *Scientific reports*, 14(1):1075, 2024.

Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. Exploring contrastive learning in human activity recognition for healthcare. *arXiv preprint arXiv:2011.11542*, 2020.

Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore IV, Gauri Ganjoo, Emmanuel Mignot, and James Y. Zou. SleepFM: Multi-modal representation learning for sleep across ECG, EEG and respiratory signals. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024. URL https://openreview.net/forum?id=cDXtscWCKC.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. *International Conference of Learning Representations*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ning Wang, Eliathamby Ambikairajah, Nigel H Lovell, and Branko G Celler. Accelerometry based classification of walking patterns using time-frequency analysis. In *2007 29th annual international conference of the ieee engineering in medicine and biology society*, pp. 4899–4902. IEEE, 2007.

Maxwell Xu, Alexander Moreno, Hui Wei, Benjamin Marlin, and James Matthew Rehg. Rebar: Retrieval-based reconstruction for time-series contrastive learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Zheng Xu, Nicole Zahradka, Seyvonne Ip, Amir Koneshloo, Ryan T Roemmich, Sameep Sehgal, Kristin B Highland, and Peter C Searson. Evaluation of physical health status beyond daily step count using a wearable activity sensor. *npj Digital Medicine*, 5(1):164, 2022.

Shu-wen Yang, Heng-Jui Chang, Zili Huang, Andy T Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, et al. A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Shuozhi Yang and Qingguo Li. Inertial sensor-based methods in walking speed estimation: A systematic review. *Sensors*, 12(5):6102–6116, 2012.

Hang Yuan, Shing Chan, Andrew P Creagh, Catherine Tong, Aidan Acquah, David A Clifton, and Aiden Doherty. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *NPJ digital medicine*, 7(1):91, 2024.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8980–8987, 2022.

Aras Yurtman and Billur Barshan. Activity recognition invariant to sensor orientation with wearable motion sensors. *Sensors*, 17(8):1838, 2017.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 35:3988–4003, 2022.

Yu Zhong and Yunbin Deng. Sensor orientation invariant mobile gait biometrics. In *IEEE international joint conference on biometrics*, pp. 1–8. IEEE, 2014.

A APPENDIX

A.1 MODEL IMPLEMENTATION DETAILS

SimCLR: We follow the approach described by Haresamudram et al. (2023), following the implementation located here: `https://github.com/ubicompsoartutorial/soar_tutorial/tree/main/simclr`. Then we use a batch-size of 64, temperature of 1, and train for 1e5 steps.

Aug Pred: We follow the approach described by Yuan et al. (2024), following the implementation located here: `https://github.com/OxWearables/ssl-wearables?tab=readme-ov-file`. Then we use a batch-size of 64 and train for 1e5 steps.

RelCon: We use the augmentations utilized by the aforementioned SimCLR model, batch-size of 64, temperature of 1, candidate set size of 20, and train for 1e5 steps.

REBAR: We follow the approach described by Xu et al. (2024), following the implementation located here: `https://github.com/maxxu05/rebar`. Then we use the prior SimCLR augmentations, a batch-size of 64, temperature of 1, candidate set size of 20, and train for 1e5 steps.

A.2 EXTRA AHMS CLASSIFICATION RESULTS

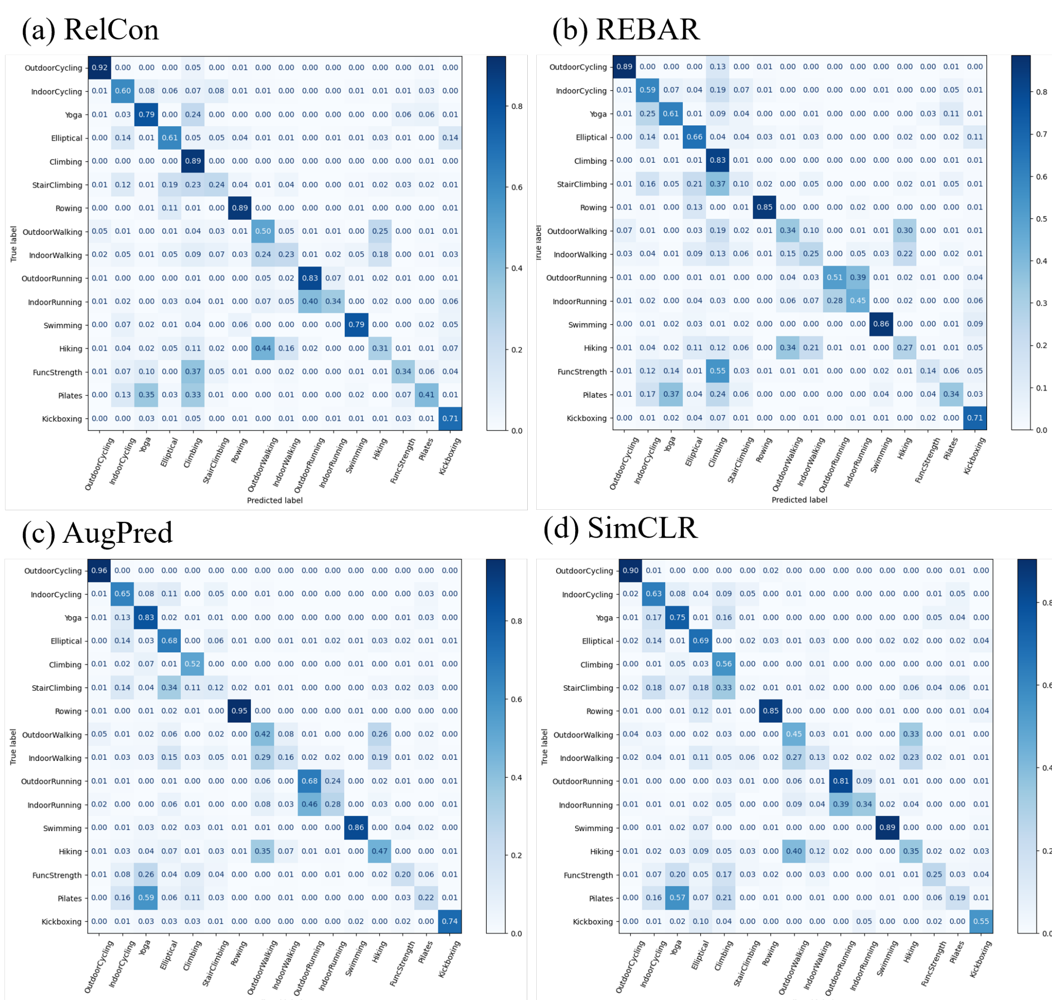

Figure 5: **Confusion Matrices for AHMS Classification at the Workout Level.** We can see RelCon has the best performance. Unlike REBAR and AugPred, RelCon can better predict Outdoor running from Indoor Running. RelCon is also able to better predict Stair Climbing unlike the others.

A.3 Details and Preprocessing on the AHMS Pre-training Dataset

Please refer to original paper on AHMS for further details on the dataset (Shapiro et al., 2023). Specifically, Figure 8 in that work includes visualizations that show general subject demographic distributions including age, body mass index, and self-reported race and ethnicity.

Preprocessing: We use the raw 100hz accelerometry data as is, without specific preprocessing techniques, such as filtering or downsampling. We choose to not filter the data as prior work has discouraged filtering in a daily monitoring setting (Campbell et al., 2020), and may prevent our deep learning model from modeling subtle nuances within the signal. Additionally, we would like to develop our methods to be robust to noise via augmentations, such as additive gaussian noise augmentations. Non-wear detection is automatically done on our device (Apple), and we purposely do not attempt to filter out periods of low accelerometry activity. This is because small, minor changes in the accelerometer signal have been shown to still be informative, being able to predict heart rate (Moebus et al., 2024).

A.4 Justification for Using Single-Sensor-Accelerometer-Only Data

Motion information can be analyzed with multiple sensor streams, whether it be through multiple accelerometer sensors strapped in different locations (Jain et al., 2022) or via extra IMU-based sensors, such as the gyroscopic sensor (García-de Villa et al., 2023). Many prior state-of-the-art human activity recognition, supervised machine learning models will exploit the full sensor suite that includes multi-modality and multi-location sensors (Essa & Abdelmaksoud, 2023; Suh et al., 2023). We recognize the value of a foundation model that can incorporate a multi-modality and/or multi-location stream of data, as it would enable for greater insights on human motion and physiology, and we are interested in investigating this in future work.

However, for our foundation model, we strive for broad generalizability in order to ensure that our learning approach and model is applicable across various settings. This includes low resource settings that only have accelerometer sensors available, as gyroscopic sensors are quite power hungry (Group, 2017). Accelerometer sensors are thus the most common sensor for monitoring human motion (Huang et al., 2023). Additionally, we would like our model to also be applicable for real-world field settings, in which multi-location sensors are uncommon for daily usage due to their bulkiness and discomfort. As such, our foundation model is able to be benchmarked against a broad range of datasets (i.e. our AHMS classification, our Gait Metric Data, HHAR, Motionsense, PAMAP2, Opportunity), which each utilize different sensor hardwares, but all include at least one 3-axis accelerometer sensor.

A.5 Elaborating on Table 1

Negatives in a SimCLR-based well-studied problem, with many recent works proposing novel methodologies to address this (Huynh et al., 2022; Jin et al., 2023; Chien & Chen, 2024). Additionally, the accelerometry SimCLR we are benchmarking (Tang et al., 2020) makes no distinction to model within-user interactions, by treating every subsequence as independent, and hence does not model within-user interactions. Both SimCLR and augmentation prediction creates instances from the original sequence via augmentations, and so they will be resistant to False Positives.

Similar to REBAR, RelCon explicitly models within and between-user interactions by explicitly comparing an anchor against candidates from within-user across time and across other users. However, unlike REBAR, we model the relative positions of our candidates, rather than a binary comparison that treats all negative instances as the same. Then, RelCon is more likely to be resistant to False positives and False negatives due to the enhanced comparison that captures the nuanced differences between candidates.

A.6 Computational Complexity of RelCon

During training, RelCon needs to compare the relative distances of each candidate from the anchor with our distance function. The distance function utilizes a highly parallelizable transformer function with complexity of $O(T^2 \times d)$(Vaswani et al., 2017) and a convolution to embed the inputs with a complexity of $O(k \times T \times d^2)$ (Vaswani et al., 2017). T=256 the subsequence length, d=64 the embedding dimension, and k=15 the kernel size. Therefore, because we calculate this distance function for every candidate, given a size of c, our total complexity during training is $O(c \times T^2 \times d + c \times k \times T \times d^2)$. In our future research, we will work on decreasing the computational cost during training.

A.7 Evaluation Dataset Descriptions

A.7.1 Comparison to a Large-Scale Pre-trained Accel Model (Yuan et al., 2024)

Yuan et al. (2024) has released their code publicly here, which contains exact data split and generations: `https://github.com/OxWearables/ssl-wearables`

Opportunity: 4-fold leave-one-subject-out cross validation with the sitting, standing, walking, and lying labels.

PAMAP2: 9-fold leave-one-subject-out cross validation with the lying, sitting, standing, walking, ascending stairs, descending stairs, vacuum cleaning, and ironing classes. This is the wrist-specific accelerometry data.

A.7.2 Comparison to an Accel SSL Benchmarking Study (Haresamudram et al., 2022)

Although Haresamudram et al. (2022) has not released their code publicly, we contacted the authors directly to ensure that we matched their exact splits and classes evaluated. Table 4 in our paper is then constructed by drawing from Table 3 and 4 in the original paper. Specifically, HHAR is drawn from Table 4, PAMAP2 from Table 3/4, and MotionSense from Table 3 (Haresamudram et al., 2022). Note that in Haresamudram et al. (2022), Augmentation Prediction is referred to as ”Multi-Task Self Supervision”.

HHAR: 5-fold leave-subject-out cross validation with the bike, sit, stairs down, stairs up, stand, and walk classes.

Motionsense: 5-fold leave-subject-out cross with the downstairs, upstairs, sitting, standing, walking, and jogging labels.

PAMAP2: 5-fold leave-subject-out cross validation with the lying, sitting, standing, walking, running, cycling, nordic walking, ascending stairs, descending stairs, vacuum cleaning, ironing, and rope jumping classes. This is the ankle-specific accelerometry data.

A.8 Generalizing to Time Lengths Beyond 2.56 seconds

2.56s is a common accelerometry subsequence size for motion tasks (Reyes-Ortiz et al., 2015; Chen & Xue, 2015; Wang et al., 2007; McQuire et al., 2021; Mandong & Munir, 2018). Additionally, we have shown that our approach can easily be used for time-series with other, differing lengths. In our comparisons against Yuan et al. (2024), we utilize a subsequence length of 10 seconds in order to match their evaluations, and we still show strong performance. In our comparisons against Haresamudram et al. (2022), we utilize a subsequence length of 2 seconds in order to match their evaluations. This highlights that our model is robust to varying input lengths and is generalizable across input data configurations, as would be desirable from a foundation model. This flexibility is enabled by the final temporal-global-average-pooling layer that we have at the end of our architecture. Additionally, in our AHMS workout-level classification task, we show how our method can be used with variable-length time-series that can last up to 10 minutes long by aggregating predictions across windows.

A.9 Comparisons of RelCon vs. Softer or Harder Loss Functions

Our relative contrastive loss in Eq. 7 is particularly interesting because the relative contrastive loss is able to have a nice balance of softness. If we increase the softness in our loss function by utilizing a metric learning loss function (Kim et al., 2019), then this will hurt performance on the gait metric regression tasks. However, if we increase the hardness to a binary contrastive loss (i.e. REBAR),

	Velocity	DST	AHMS-Subseq	AHMS-Workout
	↑ Corr	↑ Corr	↑ F1	↑ F1
RelCon	0.8431	0.7559	38.56	55.28
w/ Softer Metric Loss (Kim et al., 2019)	-3.64%	-6.11%	-1.45%	1.39%
w/ Harder Binary Contrastive Loss (Xu et al., 2024)	-6.86%	-9.82%	-5.63%	-9.03%

Table 6: Comparisons of RelCon vs. Softer or Harder Loss Functions

then this hurts performance across both types of tasks. Please see the table below, and we have added discussion to Section A.9.

A.10 Comparisons of RelCon vs. Yuan et al. (2024) with Same Encoder Backbone

We have re-trained our RelCon FM with the ResNet-18 backbone with a final encoding dimensionality of 1024, matching Yuan et al. (2024), and we show the results in the Table 7 below. Out of the 3/4 evaluations, RelCon continues to have stronger performance.

				Opportunity (Wrist → Wrist)		PAMAP2 (Wrist → Wrist)	
	Architecture	Eval Method	Pre-train Data	↑ F1	↑ Kappa	↑ F1	↑ Kappa
RelCon FM	ResNet-18	Fine-tuned	AHMS	**98.1 ± 0.1**	97.6 ± 1.2	**97.9 ± 0.5**	**97.5 ± 0.8**
Yuan et al. (2024)'s FM	ResNet-18	Fine-tuned	UKBioBank	59.5 ± 8.5	47.1 ± 10.	78.9 ± 5.4	76.9 ± 5.9
RelCon FM	ResNet-18	MLP Probe	AHMS	**65.4 ± 12.**	**55.7 ± 12.**	62.5 ± 13.	53.4 ± 6.6
Yuan et al. (2024)'s FM	ResNet-18	MLP Probe	UKBioBank	57.0 ± 7.8	43.5 ± 9.2	**72.5 ± 5.4**	**71.7 ± 5.7**

Table 7: Comparisons of RelCon vs. Yuan et al. (2024) with Same Encoder Backbone