Group Supervised Learning: Extending Self-Supervised Learning to Multi-Device Settings

Yash Jain [*1] **Chi Ian Tang** [*2] **Chulhong Min** [3] **Fahim Kawsar** [3] **Akhil Mathur** [3]

Abstract

We introduce a novel problem setting for self-supervised learning called Time-Synchronous Multi-Device Systems, which requires a solution in utilizing data from multiple data-generating devices during contrastive training. To this end, we propose a novel training setup, Group Supervised Learning (*GSL*), which is an extension of contrastive learning by contrasting time-series data gathered from different devices. GSL comprises of three main components, relating to Device Selection, Data Sampling and a novel loss function to enable contrastive learning in a group of devices. Comparisons were made between *GSL* and other semi-supervised and fully-supervised baselines, and the results demonstrated that our proposal is both data-efficient and outperforms the baselines by as high as 0.15 in micro F1-score across 2 human activity recognition datasets.

1. Introduction

Deep Learning (DL) techniques have emerged as promising alternatives for statistical feature-based Human Activity Recognition (HAR). However, the major bottleneck in all of the DL techniques in HAR is the requirement of a large labeled dataset which is expensive to collect. Semi-supervised learning has emerged as a promising avenue to counter this problem. Recently, self-supervised contrastive learning has proved to be a cornerstone approach in representational learning. A popular approach has been to learn powerful representation by augmenting unlabeled data and contrasting it with augmentations of itself or other data points (Chen et al., 2020).

In this paper, we present a novel problem setting called Time-Synchronous Multi-Device System (TSMDS) which presents an unexplored opportunity for contrastive learning in real-world settings. This problem setting is inspired by the current trend of users owning multiple sensor-enabled, data-generating devices, including smartphones and wearables. Some studies, e.g., (Safaei et al., 2017) even estimate that by 2025, each person will own 9.3 connected devices on average. An example of this trend is shown in Figure 1 – here, a user is wearing multiple accelerometer-enabled devices which are simultaneously collecting sensor data while the user is performing an activity, such as running.

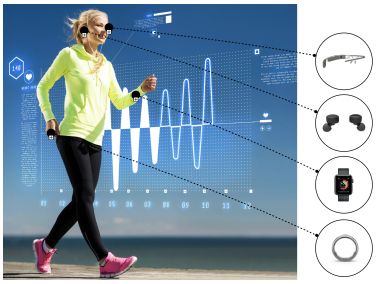

Figure 1. A user owning multiple data-generating devices.

Apart from the growing importance and practicality of this problem setting, it presents a unique opportunity for self-supervised learning. As the multiple devices are capturing the same generative process (i.e., a user's activity) from different perspectives, we hypothesize that the data captured from this group of devices have a natural affinity to each other in some latent space. More importantly, the natural affinity across this group of data samples can be leveraged to design self-supervised contrastive learning algorithms. As the supervision in this setting does not come only from 'self', but also from other devices in the group, we call this setup *Group Supervised Learning* (GSL).

Similar to contrastive learning, the intuition behind GSL is to push the feature embeddings of compatible data points from a group (called *positive group*) closer to each other while simultaneously pushing away the incompatible points (*negative group*). This way the feature extractor is able to utilize the knowledge from unlabeled data to learn a prior over the dataset, which could then be fine-tuned for specific downstream tasks using a fraction of labeled data points.

Our key contributions can be summarized as follows:

[*]Equal contribution [1]Indian Institute of Technology, Bombay [2]University of Cambridge [3]Nokia Bell Labs, Cambridge. Correspondence to: Akhil Mathur <akhil.mathur@nokia-bell-labs.com>.

ICML 2021 Workshop on Self-Supervised Learning for Reasoning and Perception.

- We propose a new problem setting, TSMDS, which exists in many domains in the real world but has not been thoroughly explored yet.
- We present a novel framework GSL addressing the TSMDS problem, utilizing the principles of contrastive learning in a group setting. We also discuss new research avenues, including device selection and data sampling in the framework.
- Our early results demonstrate that GSL outperforms supervised and semi-supervised training baselines proposed in the HAR literature by as high as 0.15 in F-1 score.

2. Related Work

Contrastive Learning of Visual Representations. Contrastive learning (Khosla et al., 2021; Chen et al., 2020; He et al., 2020; Caron et al., 2021; Harley et al., 2020) trains models to extract distinctive features for different data, by setting up a contrastive task between positive samples, those which are similar depending on the heuristics selected, and negative samples, those which are not from a similar distribution. Recent work by (Chen et al., 2020) proposed a simplistic self-supervised contrastive learning framework, SimCLR. The framework trains the feature extractor to be agnostic against transformations, by using transformed views of the same sample as positive pairs and contrasting them against other samples. The authors demonstrated that the use of embeddings from earlier layers of a contrastive model for downstream tasks increased the performance and generalizability of the embeddings. However, the simplistic sampling of negative samples in SimCLR might not provide the best supervision for training. Prior work in exploring the effect of sampling on deep embedding learning (Wu et al., 2017) has demonstrated that sampling could have a larger impact on performance compared to the design of the loss function. In this paper, we present a novel sampling method for HAR, by leveraging characteristics of synchronized data streams.

Self-supervised Learning for Human Activity Recognition. Self-supervised learning (SSL) has become an increasingly popular area of research for human activity recognition (HAR), where researchers have proposed different ways to extract supervisory signals from data (Saeed et al., 2019). Recently, the SimCLR framework has been applied in HAR (Tang et al., 2020; Saeed et al., 2019; 2020). The authors explored a set of different combinations of transformation functions that are designed for time-series data, for training feature extractors for sensor data based on the SimCLR framework. A slight improvement in performance was observed compared to other training pipelines. However, again this work focused on leveraging data from a single sensor only, and the potential for extracting stronger supervisory signals from other sensors and devices was not explored. An initial attempt to leverage multiple devices for SSL has been made for visual representation (Sermanet et al., 2018). It showed that time-synchronized visual representations can be used to provide a reward function for robot manipulation via reinforcement learning. One of its limitations is that it utilizes data from two camera views only, but our proposal explores settings with more than two data sources.

3. Method

3.1. Problem Formulation

In the Time-Synchronous Multi-Device System (TSMDS) problem setting, we are given time-aligned unlabeled data samples from K devices. Let $D^i = \{(\mathbf{x_j^i})\}_{j=1}^N$ be the unlabeled dataset from the i^{th} device, where $\mathbf{x_j^i}$ is a data sample and captured by the i^{th} device at time j. Let $D^0 \in \{D^i\}|_{i=1}^K$ be an anchor device for which we want to train and test a downstream classification model. Then the goal of GSL is to leverage the time-aligned, unlabeled multi-device datasets to learn a feature extractor F_θ that can generate effective feature representations for D^0.

3.2. Solution Framework: GSL

We propose GSL (*Group Supervised Learning*), a contrastive self-supervised learning framework which extends SSL to a setting with *groups* of time-aligned devices. The key intuition behind GSL is to take the time-aligned samples from devices similar to D^0, and pull them closer to the samples from D^0 in the embedding space. Similarly, we aim to push other samples (e.g., unaligned samples) away from D^0 in the embedding space. Our solution framework comprises of three components:

Device Selection. First, we separate the devices (other than D^0) into two groups: a 'positive group' denoted by D^+, and a 'negative group' denoted by D^-. The key idea here is that devices in the positive (or negative) group will contribute the positive (or negative) samples during contrastive learning with the anchor device D^0. The choice of device selection algorithms can depend on the problem domain and learning task. For instance, for the task of HAR using motion sensors, we can separate devices into positive and negative groups based on the similarities/dissimilarities in their degrees of freedom (DoF). E.g., for a 'chest'-mounted sensor as the anchor device, a 'back'-mounted sensor could be a positive device (similar DoF) and a 'forearm'-mounted sensor could be the negative device (different DoF). Alternatively, we can use distribution-level statistics to separate devices. In this work, we compute the Maximum Mean Discrepancy (MMD) distance between the anchor dataset D^0 and the other devices, and choose p device(s) with the smallest MMD distances as the positive device(s), and q devices(s)

with the largest MMD distances as negative devices. Here p and q are hyperparameters that influence how many and which devices will be used in contrastive learning.

Data Sampling. After separating the devices, the next question is how do we sample data from them during mini-batch training. As highlighted in prior works (Wu et al., 2017), mini-batch sampling could have a large impact on the performance of self-supervised learning and hence, this component of our framework allows for incorporating various sampling strategies for positive and negative devices.

In this paper, we propose and evaluate the idea of *synchronous positive* sampling and *asynchronous negative* sampling. While constructing the mini-batches, we take a sample $\mathbf{x_T^0}$ from the anchor device D^0, and its corresponding time-aligned samples $\mathbf{x_T^+}$ from the positive devices D^+. For the negative devices D^-, we randomly select samples $\mathbf{x_t^-}|_{\mathbf{t \neq T}}$ which are not time-aligned with the anchor sample. The key intuition here is that *synchronous sampling* from positive devices will output a sample with the same label as the anchor sample, and we can aim to pull these samples close to each other in the embedding space. Similarly, the *asynchronous negative* sampling will output non-aligned samples from the negative devices, which we will aim to push away from the anchor.

The sampled data, denoted by $\mathbf{x}^\bullet \in D^\bullet, \bullet \in \{+, -, 0\}$ is then passed to the feature extractor F_θ to obtain the feature outputs $z^\bullet$.

$$z^\bullet = F_\theta(\mathbf{x}^\bullet), \bullet \in \{+, -, 0\} \tag{1}$$

Group Supervised Contrastive Loss. Finally, we train this GSL setup using a novel loss function called Group Supervised Contrastive Loss, which is an extension of the standard contrastive loss function but compatible with multiple positive and negative samples. More specifically, we have:

$$\mathcal{L}_{GSL} = \frac{\sum_{i=0}^{|D^+|} \exp\left(\text{sim}\left(z^0, z_i^+\right)/\tau\right)}{\left(\begin{array}{l}\sum_{i=0}^{|D^+|} \exp\left(\text{sim}\left(z^0, z_i^+\right)/\tau\right) \\ \quad + \sum_{j=0}^{|D^-|} \exp\left(\text{sim}\left(z^0, z_j^-\right)/\tau\right)\end{array}\right)} \tag{2}$$

where sim(.) denotes *cosine similarity* and τ is a hyperparameter denoting temperature.

4. Evaluation

4.1. Experimental Setup

Datasets: For our experiments, we used two datasets for human activity recognition (HAR): OPPORTUNITY (Roggen et al., 2010) and REALWORLD (Sztyler & Stuckenschmidt, 2016). They contain 3-axis accelerometer and 3-axis gyroscope data from multiple on-body devices.

Method Proportion of data	GSL $\leq 75\%$	SSL $\leq 75\%$	Supervised 100%
OPP - Back	**0.769**	0.612	0.698
OPP - Left Lower Arm	**0.783**	0.736	0.756
OPP - Left Shoe	**0.732**	0.706	0.700
OPP - Right Shoe	0.722	**0.735**	0.726
OPP - Right Upper Arm	**0.831**	0.599	0.681
RW - Chest	**0.906**	0.788	0.899
RW - Forearm	**0.852**	0.839	0.833
RW - Head	**0.834**	**0.834**	0.788
RW - Shin	**0.891**	0.886	0.885
RW - Thigh	**0.899**	0.866	0.879
RW - Upper Arm	**0.876**	0.862	0.857
RW - Waist	**0.916**	0.808	0.887

Table 1. Comparison of classification performance (F1-micro scores) between *GSL* and other training pipelines on two HAR datasets (OPP - Opportunity, RW - RealWorld).

The Opportunity dataset consists of data collected from 4 participants performing activities of daily living with 17 on-body sensor devices. For the study, we used five devices deployed on back, left lower arm, right shoe, right upper arm, and left shoe, and we targeted to detect the mode of locomotion: *stand, walk, sit, and lie*. The RealWorld dataset contains accelerometer and gyroscope traces of 15 participants, sampled at 50 Hz simultaneously on 7 sensor devices mounted at forearm, thigh, head, upper arm, waist, chest, and shin. Each participant performed 8 activities: *climbing stairs down and up, jumping, lying, standing, sitting, running/jogging, and walking*.

Baselines and Evaluation Metrics: We evaluate GSL against two baselines, *Supervised* and self-supervised learning using a single device, namely *SSL*. Supervised represents the traditional supervised learning, which trains the model with all the labeled data of the anchor device. SSL follows the technique proposed in (Tang et al., 2020) and trains F_θ with unlabeled data from the anchor device using contrastive learning. The positive data for contrastive learning is generated by augmenting an anchor data point with *rotation* (simulating different sensor placements) and the negative data is chosen from other data points of the anchor device, i.e., with different timestamps.

For GSL and SSL, after training F_θ using contrastive learning, we add a classification head and train it using labeled data from D^0. We compare and report the performance (micro-averaged F_1 score) of each technique on the downstream HAR classification task on the anchor device.

Data processing: The accelerometer and gyroscope traces were segmented into time windows of 3 seconds for RealWorld and 5 seconds for Opportunity without any overlap. The whole dataset was normalized to be in the range of -1 and 1. For validation, we constructed the training and test sets by dividing the devices datasets into two parts: 60% and 40%, respectively.

Network architecture and hyperparameters: The base

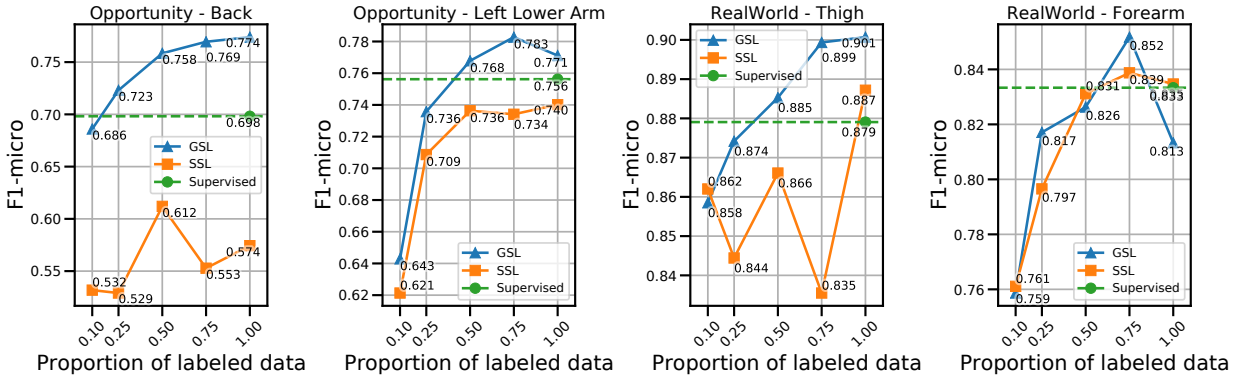

Figure 2. Assessing the classification performance of *GSL* and other training pipelines across different proportions of labeled data at four body positions. F1-micro scores were reported on two different datasets, where *GSL* outperformed the other pipelines in almost all cases.

feature extractor F_θ for GSL and SSL consists of three 1D convolutional layers with 32, 64, and 96 feature maps and kernel sizes of 24, 16, and 8 respectively. We used Dropout ($p = 0.1$) and L2 regularization, and added a global maximum pooling layer at the end. This is a design that was shown effective for HAR in previous works (Saeed et al., 2019; Tang et al., 2020). For the classification head, we use 2 dense layers and a softmax layer on top of F_θ, and the last convolutional layer of F_θ was fine-tuned together with the dense layers. For device selection, we use $p = 1$ and $q = 5$ as hyperparameters for the RealWorld dataset, i.e., we used one device with the least MMD distance from D^0 as the positive device and 5 devices with the highest MMD distance from D^0 as negative devices. For the Opportunity dataset, we use $p = 1$ and $q = 3$.

4.2. Evaluation Results

We evaluate our proposal (*GSL*) and compare its performance against baselines in the following setting: whether our proposal, with lower labeled data availability, performs on-par compared to the semi-supervised baseline, *SSL*, with the same data availability, and whether it performs better than *Supervised* with the full dataset. This setting is designed to evaluate whether our proposal outperforms the baseline methods and whether it is data-efficient, i.e., it is able to perform well with less labeled data.

To this end, we fine-tuned *GSL* and *SSL* using 10%, 25%, 50%, 75% or 100% of the labeled training data from the anchor device, and evaluated them on the test set of the anchor device. The *Supervised* model was trained using 100% of the training data from the anchor device and evaluated similarly on the test set. A hyperparameter search on training parameters were performed for all pipelines to ensure the optimal performance.

Table 1 shows the F1-micro scores of the models when evaluated at 12 anchor devices. We compare the *Supervised* model performance (trained with 100% labeled data) against the best performing *GSL* and *SSL* models trained using $\leq 75\%$ of the labeled data. The results show that the *GSL* method outperforms the other baselines in the vast majority of cases, with a performance gain compared to the second-best pipeline as high as 0.15 in F1-score. It also indicates that the *GSL* pipeline is data-efficient, by outperforming the *Supervised* baseline in all cases while using less data.

Figure 2 shows the impact on performance when the proportion of labeled data varies at four anchor devices across the two datasets (please refer to the appendix for the remaining results). We present two important findings. First, regardless of the proportion of labeled data used for fine-tuning, GSL generally outperforms SSL. This shows that our design of synchronous positive and asynchronous negative sampling of data from multiple devices contributes to enhancing the accuracy of self-supervised learning. Second, GSL outperforms *Supervised*, even with much less labeled data. For example, when evaluated at the 'back' position of the Opportunity dataset, *GSL* was able to outperform the *Supervised* baseline using a quarter of the data.

5. Concluding Remarks

Our proposed GSL framework is still a work in progress and there are a number of open research questions. First, our problem setting and solution are not limited to HAR tasks with accelerometer and gyroscope data: it can be easily extended to other time-series data such as audio or video, in multi-device settings. For example, GSL can be used to train a feature extractor for downstream tasks in a multi-view setting where the same scene is captured by different cameras, or in a multi-audio setting where different smart audio devices are recording the same speech in a room. We are currently extending our evaluation to these settings. Second, the choice of device selection and sampling algorithms can greatly affect the performance of GSL. Due to lack of space, we did not extensively compare different sampling or device selection algorithms, and it remains an important topic for future work.

References

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Harley, A. W., Lakshmikanth, S. K., Li, F., Zhou, X., Tung, H.-Y. F., and Fragkiadaki, K. Learning from unlabelled videos using contrastive predictive neural 3d mapping, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning, 2020.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning, 2021.

Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Förster, K., Tröster, G., Lukowicz, P., Bannach, D., Pirkl, G., Ferscha, A., et al. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*, pp. 233–240. IEEE, 2010.

Saeed, A., Ozcelebi, T., and Lukkien, J. 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.

Saeed, A., Ungureanu, V., and Gfeller, B. Sense and learn: Self-supervision for omnipresent sensors. *arXiv preprint arXiv:2009.13233*, 2020.

Safaei, B., Monazzah, A. M. H., Bafroei, M. B., and Ejlali, A. Reliability side-effects in internet of things application layer protocols. In *2017 2nd International Conference on System Reliability and Safety (ICSRS)*, pp. 207–212. IEEE, 2017.

Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., and Brain, G. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1134–1141. IEEE, 2018.

Sztyler, T. and Stuckenschmidt, H. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–9. IEEE, 2016.

Tang, C. I., Perez-Pozuelo, I., Spathis, D., and Mascolo, C. Exploring contrastive learning in human activity recognition for healthcare. *arXiv preprint arXiv:2011.11542*, 2020.

Wu, C.-Y., Manmatha, R., Smola, A. J., and Krahenbuhl, P. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2840–2848, 2017.

A. Evaluation results at other body positions

Figures 3 and 4 show the performances of GSL and other baselines at all remaining body positions not presented in section 4. In most cases, GSL was able to outperform the other baselines, with significant improvement in F1-scores. However, at some body positions, our proposal only achieved a slight improvement over the baselines, or performed worse. One possible reason for the variation in results is that some devices exhibit vastly different physical characteristics compared to other ones although they are the result of the same generative process. This could make contrastive learning across devices less effective, and hence the performance varies. As a future work, we plan to identify and counter such settings directly in our training framework, e.g., using the MMD distances between devices as a proxy for their utility in contrastive learning.

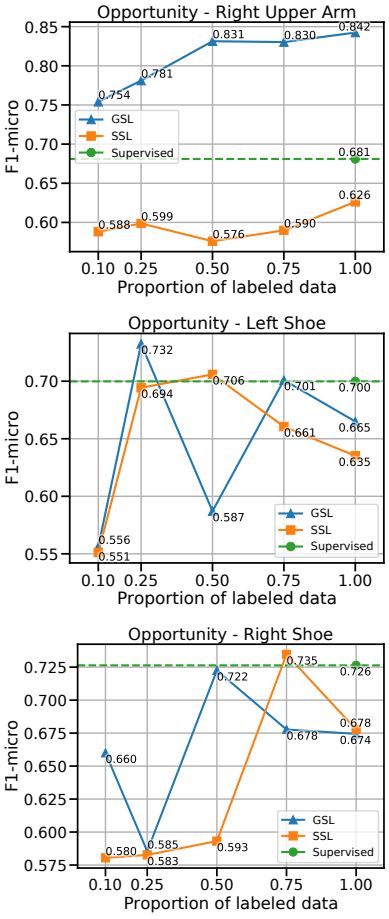

Figure 3. Assessing the classification performance of *GSL* and other training pipelines across different proportions of labeled data available at other body positions in the Opportunity dataset. F1-micro scores are reported on two different datasets.

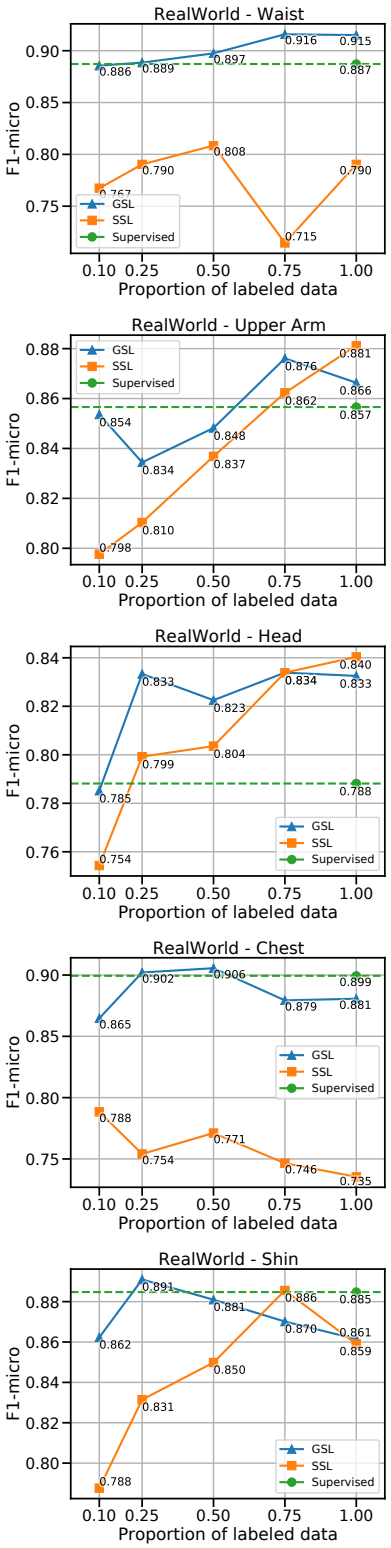

Figure 4. Assessing the classification performance of *GSL* and other training pipelines across different proportions of labeled data available at other body positions in the Opportunity dataset. F1-micro scores were reported on two different datasets.