Contrastive Self-supervised Learning for Human Activity Recognition using Inertial and Skeleton Modalities

Bulat Khaertdinov*, Razvan Brinzea*, Stylianos Asteriadis

Maastricht University, Netherlands

b.khaertdinov@maastrichtuniversity.nl, r.brinzea@student.maastrichtuniversity.nl, stelios.asteriadis@maastrichtuniversity.nl

Abstract

Human Activity Recognition is a field of research where input data can take many forms. Each of the possible input modalities describes human behaviour in a different way, and each has its own strengths and weaknesses. We explore the hypothesis that leveraging multiple modalities can lead to better recognition. Since manual annotation of input data is often difficult, we focus our attention on self-supervised methods which can learn useful feature representations without any ground truth labels. We extend two recent contrastive self-supervised approaches for the task of Human Activity Recognition, leveraging inertial and skeleton data. While self-supervised learning methods have been previously applied to another set of modalities, to the best of our knowledge, this is the first work employing inertial and skeleton modalities. We evaluate the proposed frameworks on three multimodal datasets (UTD-MHAD, Berkeley-MHAD and MMAct). Our experiments show that multimodal self-supervised learning improves the quality of representations compared to the unimodal setting, and that the results are competitive even compared to supervised models.

Introduction

Recent advances in deep learning have demonstrated significant improvements in Human Activity Recognition (HAR). Modern HAR methods are based on various sources of input data. Though, the most widely used methods are based on data coming from video sensors (e.g. RGB videos, depth streams or skeletal joints) and inertial measurement units (IMU) (Wang et al. 2019). The former are also known as video-based approaches, while the latter – as sensor-based. Both modalities have their own limitations that can be minimized in multimodal settings. Nevertheless, even multimodal approaches have a significant drawback, namely, they require vast amounts of annotated data to learn robust representations.

Self-supervised learning (SSL) is a paradigm which can be used in order to solve this problem. The main idea of the paradigm is to train deep feature encoders on an auxiliary task without using actual labels. Different types of SSL methods have been successfully applied to video-based and sensor-based HAR. Though, the multimodal HAR methods proposed in the bibliography are quite limited and have not been applied to all possible modalities.

In this paper, we aim to fill the gap and address the problem of multimodal HAR in a self-supervised learning manner. Specifically, we propose two multimodal SSL approaches for HAR – the first is based on pre-training individual feature encoders, following the SimCLR (Chen et al. 2020) framework, and fusing their representations, and the second is based on contrastive multiview coding (Tian, Krishnan, and Isola 2020). We evaluate our proposed approaches using deep feature encoders for inertial and skeleton modalities on three widely used multimodal HAR datasets, in self-supervised and semi-supervised settings.

Related work

Contrastive Self-Supervised Learning

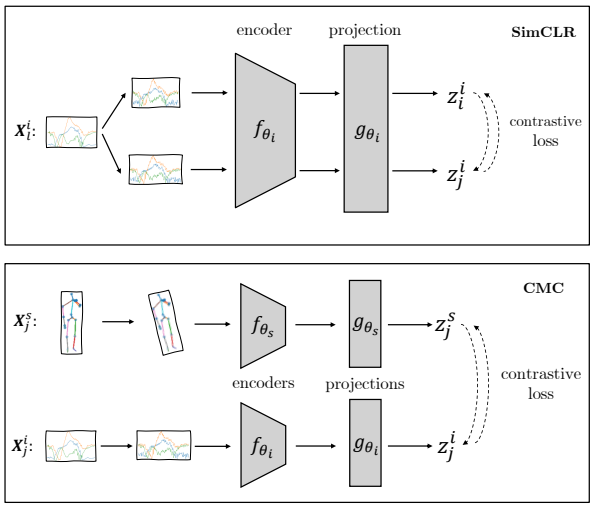
Self-supervised learning methods learn useful representations of data by means of solving a pretext task which can be formulated without the need of any ground-truth labels, and ultimately use the learnt representation to solve a downstream task. This enables researchers to use the limitless amounts of unlabeled data which are readily available for a wide number of domains.

A family of recent contrastive SSL methods has been very successful in narrowing the gap between supervised and self-supervised methods. At their core, these methods rely on generating augmented views of the data and minimizing a contrastive loss that pulls together the latent representations of similar inputs, while pushing away semantically different inputs (Chen et al. 2020). Some works improve upon this framework by means of knowledge distillation (Tian, Henaff, and van den Oord 2021), while others extend this principle to the multimodal setting, leveraging multiple visual modalities (Tian, Krishnan, and Isola 2020). In the field of HAR, contrastive SSL has only been applied on individual data sources, such as sensors (Khaertdinov, Ghaleb, and Asteriadis 2021) or skeleton data (Linguo et al. 2021).

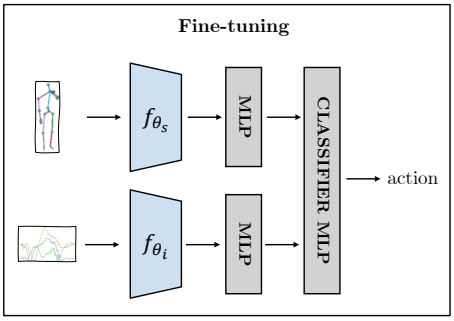
Multimodal Human Activity Recognition

To account for the significant difference between input modalities, many multimodal HAR works apply input, fea-

*Equal contribution



(a) Pre-training phase



(b) Fine-tuning phase

Figure 1: We first pre-train both of our encoders using one of the approaches shown in (a). Then, we freeze the pre-trained encoder networks and perform fine-tuning on labeled data, as shown in (b).

ture or decision fusion in various architectures comprising multiple backbone networks (Ahmad and Khan 2019), (Khaire, Kumar, and Imran 2018). More recent works propose end-to-end architectures designed specifically for multimodal HAR, and leverage multiple multi-head attention mechanisms (Islam and Iqbal 2021), or rely on knowledge distillation to transfer knowledge from teacher networks trained on one modality to student networks trained on another (Liu et al. 2021).

Although plenty of advanced supervised methods have shown impressive performance for the multimodal HAR problem, only a limited amount of works have addressed this task in a self-supervised learning manner (Akbari et al. 2021; Li et al. 2021). Moreover, to the best of the authors' knowledge, there are no works applying SSL methods for multimodal HAR using inertial and skeleton modalities. What makes it more crucial is that data labelling is a hard, expensive and time-consuming task, especially in the case when multiple sources of input data are available. That is why, in this study, we adapt the modern SSL frameworks to sensor and visual modalities and compare their performance using two evaluation scenarios.

Methodology

Multimodal Human Activity Recognition is a classification problem that can be formulated as follows: given a set of inputs $\{\boldsymbol{X}^m | m \in M\}$ from a set of modalities M, an objective is to predict a label $y \in Y$ associated with these inputs. In this paper, we used two modalities, namely inertial signals and skeleton joints.

Inertial signals might contain data from such devices as accelerometers, gyroscopes and magnetometers. They can be treated as multivariate time-series data. Precisely, at timestamp t, input signal $\boldsymbol{x_t} = [x_t^1, x_t^2, \ldots, x_t^S] \in \mathbb{R}^S$ consists of S values corresponding to sensor channels. These multichannel inputs can be merged into a matrix $\boldsymbol{X}^i = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_T}] \in \mathbb{R}^{T \cdot S}$ over T timestamps.

Skeleton data is generally provided as a set of 2D or 3D coordinates tracked over time for a number of keypoints located on the human body. For any skeleton sequence, we denote T as the number of frames in the sequence, J as the number of joints and C as the number of data channels (2 or 3). Then, a skeleton sequence $\boldsymbol{X}^s \in \mathbb{R}^{T \cdot J \cdot C}$ consists of T frames where the skeleton data for each frame is described by $\boldsymbol{P}_t = [\boldsymbol{p}_t^1, \boldsymbol{p}_t^2, \ldots, \boldsymbol{p}_t^J]$ and $\boldsymbol{p}_t^j \in \mathbb{R}^C$ is the position of joint j at frame t.

In this study, we exploit the recent advances of contrastive self-supervised learning to implement two types of multimodal SSL pre-training. First, we pre-train encoders for each modality separately. Later, we use both pre-trained encoders in a simple fusion fine-tuning scheme. In the second approach, we pre-train both encoders from scratch simultaneously using contrastive multiview coding (Tian, Krishnan, and Isola 2020). Both approaches are shown in Figure 1 and described thoroughly in the following subsections.

Feature Encoders

For the skeleton backbone encoder f_{θ_s}, we use a convolutional co-occurrence feature learning network (Li et al. 2018). The network uses a two-stream input (positions and motions) and comprises of a series of convolutional blocks, with ReLU non-linearities and max-pooling applied to certain layers. The encoder f_{θ_i} for the multivariate time-series data coming from inertial sensors is adapted from the CSSHAR paper (Khaertdinov, Ghaleb, and Asteriadis 2021). Specifically, we implement three layers of 1D convolutional blocks and two layers of multi-head self-attention.

Contrastive Pre-text Tasks

Both approaches implemented in this work are based on contrastive learning. The main idea is to group semantically similar inputs (positive samples) together in a latent space, and push away the different ones (negative samples). In SSL settings, positive samples represent different views of the same instance, while negative samples are obtained from different examples. The proposed approaches differ in the methods used to mine views and pre-train encoders for each modality.

Modality	Encoder	Method	UTD-MHAD (Acc.)	Berkeley-MHAD (Acc.)	MMAct (F1)
Inertial	Transformer	Supervised	78.84	89.82	64.99
		SimCLR	75.84	93.09	57.13
Skeleton	Co-occurrence	Supervised	94.42	98.9	82.5
		SimCLR	93.72	96.36	76.19
Multimodal	Transformer + Co-occurrence	Supervised	90.93	97.82	85.98
		SSL Simple Fusion	**94.88**	**99.27**	79.97
		CMC	93.82	98.18	**80.67**

Table 1: Performance of supervised and SSL models on test data. The highest scores for each dataset are underlined, whereas the best results for the SSL methods are highlighted in bold.

Unimodal Pre-training and Fusion. The main idea of our first approach is to pre-train encoders separately using contrastive loss and fuse representations in the fine-tuning stage. In this case, the SimCLR framework (Chen et al. 2020) has been chosen for the pre-text task. The process of pre-training is described below for the inertial modality and all the steps are the same for skeleton data. Specifically, two random transformations are applied to each input in a batch to generate two views $t_1(\boldsymbol{X}_l^i)$ and $t_2(\boldsymbol{X}_l^i)$. Later, the views are passed through the encoder and a projection head to generate feature representations $\boldsymbol{z}_i^i = g_{\theta_i}(f_{\theta_i}(t_1(\boldsymbol{X}_l^i)))$ and $\boldsymbol{z}_j^i = g_{\theta_i}(f_{\theta_i}(t_2(\boldsymbol{X}_l^i)))$. Finally, the NT-Xent loss for a positive pair z_i^i and z_j^i is computed as follows:

$$l_{i,j}^i = -log\frac{exp(s(z_i^i, z_j^i)/\tau)}{\sum_{k=1}^{2N} exp(s(z_j^i, z_k^i)/\tau)},$$

where $s(\cdot)$ is cosine similarity, N is a batch size and τ is temperature. The total loss is computed by summing across all positive pairs.

During fine-tuning, encoders pre-trained separately are fused using three MLP networks: the first two map modality-specific features into the same size, while the third one takes concatenated outputs and produces classification labels.

Contrastive Multiview Coding CMC is a contrastive self-supervised learning method which can be used to learn representations when two or more modalities are available (Tian, Krishnan, and Isola 2020). Instead of contrasting between different augmented views of the same input data, CMC contrasts the representations of the different modalities. Considering the case where inertial and skeleton modalities are available, for each sample $\{\boldsymbol{X}_j^i, \boldsymbol{X}_j^s\}$ in a training batch of size N, we use modality-specific encoders $f_{\theta_i}, f_{\theta_s}$ and projection heads $g_{\theta_i}, g_{\theta_s}$ to generate feature representations $\boldsymbol{z}_j^i = g_{\theta_i}(f_{\theta_i}(\boldsymbol{X}_j^i))$ and $\boldsymbol{z}_j^s = g_{\theta_i}(f_{\theta_s}(\boldsymbol{X}_j^s))$. These representations are then treated as a positive pair and contrasted against all other representations in the batch. Formally, the loss obtained by treating $\boldsymbol{X}_j^i$ as an anchor and enumerating over the representations of the other samples $\boldsymbol{X}_k^s$ is:

$$l_j^{i\to s} = -log\frac{exp(s(z_j^i, z_j^s)/\tau)}{\sum_{k=1}^{N} exp(s(z_j^i, z_k^s)/\tau)}$$

The total loss for the batch is computed by summing across each input sample, in both directions.

Experimental Setup

Datasets

In this paper, three multimodal datasets were used to evaluate the performance of the proposed approaches, namely UTD-MHAD (Chen, Jafari, and Kehtarnavaz 2015), Berkeley-MHAD (Ofli et al. 2013) and MMAct (Kong et al. 2019). Skeleton and inertial modalities were extracted and used from all the datasets.

UTD-MHAD. The dataset contains data collected by 10 subjects performing 27 activities, 4 trials for each. The three-dimensional joint coordinates were recorded with a Kinect camera, while the inertial data was collected using one wearable device with accelerometer and gyroscope. We follow the original evaluation protocol and use odd-numbered subjects for training and even-numbered subjects for testing.

Berkeley-MHAD. Data is recorded using 6 accelerometers installed on ankles, wrists and hips and multiple cameras. For skeleton data, we exploit the 3D keypoints obtained from the Impulse Motion Capture system. In the dataset, 11 activities are performed several times by 12 subjects. As described in the original paper, we train on the first 7 subjects, and reserve the last 5 subjects for testing.

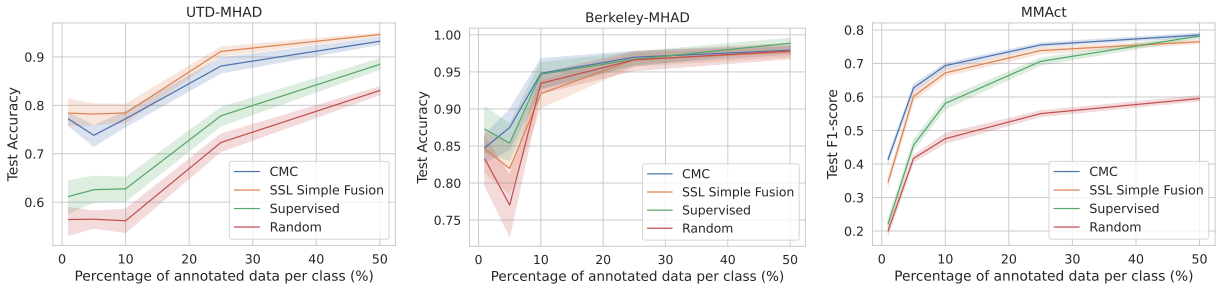
MMAct. The dataset consists of 36 activities performed by 20 subjects in different scenes. For skeleton data, we employ the 2D keypoints present in the challenge version of the dataset*. The sensor data comes from a smartwatch (accelerometer) and a smartphone (accelerometer, gyroscope, orientation) located in the subject's pocket. We follow the cross-subject evaluation protocol, using the samples from the first 16 subjects for training.

Implementation Details

All experiments were performed on an Nvidia Quadro RTX 5000 GPU with 16GB of memory, using Pytorch Lightning. In all settings, we use the Adam optimizer with a learning rate of 0.001. We tune the other model and training hyperparameters through a grid search procedure. All MLPs contain two hidden layers of 256 and 128 neurons, respectively, with ReLU applied to each hidden layer. Projection MLPs use batch normalization in the first layer.

Pre-processing. We re-sample all input sequences to 50 timesteps. Additionally, we normalize joint positions in all skeleton sequences based on the first frame of each sequence, following a standard normalization procedure (Khaire, Kumar, and Imran 2018).

*challenge dataset: https://mmact19.github.io/challenge/



(a) UTD-MHAD (b) Berkeley-MHAD (c) MMAct

Figure 2: Average values of performance metrics with 95% confidence intervals for the semi-supervised learning scenario.

Supervised Models. We train all supervised models for 100 epochs, using a batch size of 64 for all datasets. For the inertial model, the three convolutional layers consist of 32, 64, and 128 output channels, respectively. We also use 2 layers of transformer-like self-attention both containing 2 heads. For the skeleton encoder, we use the same architecture and hyperparameters as described in the original paper, removing dropout layers and adding batch normalization to each layer.

SSL Models. We pre-train individual encoders with the SimCLR framework for 300 epochs, and then fine-tune for 100 epochs. For the inertial modality, we test different sets of augmentations to be randomly applied for each dataset as proposed in (Khaertdinov, Ghaleb, and Asteriadis 2021) using the grid search strategy. The following augmentation sets have shown optimal performance: UTD-MHAD – {Jittering, Scaling, Permutation, Channel shuffle}, Berkeley-MHAD – {Jittering, Scaling, Permutation}, MMAct – {Scaling, Rotation}. The optimal batch sizes and temperature values for pre-training have also been selected individually for each dataset: UTD-MHAD – 128 and 0.05, Berkeley-MHAD – 64 and 0.1, MMAct – 64 and 0.2. When pre-training the skeleton encoder, we use the following set of augmentations which we apply randomly to the data: {Jittering, ResizedRandomCrop, Scaling, Rotation, Shearing}. The selected batch sizes and temperature values for each dataset are: UTD-MHAD – 32 and 0.5, Berkeley-MHAD – 32 and 0.15, MMAct – 128 and 0.2.

For the experiments involving the CMC framework, we train for 200 epochs and then fine-tune for 100 epochs, with a batch size of 64 for all datasets and a temperature value of 0.1. We keep the same model hyperparameters as in the previous experiment.

Evaluations

Learning Feature Representations

The first evaluation scenario aims to assess the robustness of feature representations learnt during pre-training and to compare the performance of SSL models to identical models trained in a supervised manner. In this scenario, the entire labeled training set is used to train supervised models and fine-tune models pre-trained in SSL settings. In Table 1, we present the results of this scenario for supervised and SSL models trained and evaluated on unimodal and multimodal data.

In the unimodal setting, models trained on the skeleton modality generally perform better than the ones trained on inertial data. Nevertheless, the multimodal approaches reach the highest values of performance metrics. Furthermore, for the UTD-MHAD and Berkeley-MHAD datasets, the SSL models containing encoders pre-trained on inertial and skeleton modalities separately (SSL Simple Fusion) have even outperformed the supervised approaches. For the MMact dataset, the CMC model has shown the best performance among the SSL models. However, it is still worse than the supervised multimodal model by about 4.5% F1-score. While comparing metrics obtained for multimodal SSL approaches, it can be seen that the difference in performance is marginal. However, pre-training models separately requires more time than pre-training them within the CMC framework.

Semi-supervised Learning Scenario

The second evaluation scenario is more practical and realistic as only part of data available for training is annotated. Specifically, the SSL models are pre-trained on the entire unlabeled dataset and fine-tuned using only a percentage of labeled data samples per activity $p \in \{1\%, 5\%, 10\%, 25\%, 50\%\}$. We also compare the performance of the multimodal SSL models to the identical supervised and randomly initialized frozen models trained using limited data. For each p, we repeat the experiment 10 times and compute the 95% confidence intervals. The results for this scenario are summarized in Figure 2.

What can be clearly seen is that SSL methods significantly outperform supervised models especially when a very limited data is available for the UTD-MHAD and MMAct datasets. For example, when only 1% of labeled data is available, SSL models achieve about 80% accuracy, where the performance of the supervised model is close to the performance of the random model (close to 60%). However, for the Berkeley-MHAD dataset, all the models, including the random one, perform at relatively the same level. This might be related to the low complexity of the task in the dataset, i.e. passing input data through shallow MLP networks is enough to achieve a satisfactory performance.

Conclusions and Future Work

In this paper, we proposed to use contrastive SSL techniques for the problem of multimodal Human Activity Recognition. Specifically, the SimCLR and CMC frameworks have been adapted to skeleton and inertial modalities. The extensive evaluations on three datasets have shown that the proposed multimodal SSL methods achieve performance compararable to supervised models and even outperform them when limited amounts of data are available.

Future research directions on this topic can focus on mitigating the harmful effect of negative pairs in contrastive losses (Chen and He 2021; Li et al. 2021), or could explore the impact of adding more modalities, such as RGB vision data, to the multimodal framework.

Acknowledgement

This work has been partially funded by the European Union's Horizon2020 project: PeRsOnalized Integrated CARE Solution for Elderly facing several short or long term conditions and enabling a better quality of LIFE (Procare4Life), under Grant Agreement N.875221.

References

Ahmad, Z.; and Khan, N. M. 2019. Human Action Recognition Using Deep Multilevel Multimodal (M2) Fusion of Depth and Inertial Sensors. *CoRR*, abs/1910.11482.

Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.-H.; Chang, S.-F.; Cui, Y.; and Gong, B. 2021. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Chen, C.; Jafari, R.; and Kehtarnavaz, N. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International Conference on Image Processing (ICIP)*, 168–172.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1597–1607. PMLR.

Chen, X.; and He, K. 2021. Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15750–15758.

Islam, M. M.; and Iqbal, T. 2021. Multi-GAT: A Graphical Attention-Based Hierarchical Multimodal Representation Learning Approach for Human Activity Recognition. *IEEE Robotics and Automation Letters*, 6(2): 1729–1736.

Khaertdinov, B.; Ghaleb, E.; and Asteriadis, S. 2021. Contrastive Self-supervised Learning for Sensor-based Human Activity Recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, 1–8.

Khaire, P.; Kumar, P.; and Imran, J. 2018. Combining CNN streams of RGB-D and skeletal data for human activity recognition. *Pattern Recognition Letters*, 115: 107–116. Multimodal Fusion for Pattern Recognition.

Kong, Q.; Wu, Z.; Deng, Z.; Klinkigt, M.; Tong, B.; and Murakami, T. 2019. MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding. In *The IEEE International Conference on Computer Vision (ICCV)*.

Li, C.; Zhong, Q.; Xie, D.; and Pu, S. 2018. Co-Occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, 786–792. AAAI Press. ISBN 9780999241127.

Li, L.; Wang, M.; Ni, B.; Wang, H.; Yang, J.; and Zhang, W. 2021. 3D Human Action Representation Learning via Cross-View Consistency Pursuit. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 4741–4750. Computer Vision Foundation / IEEE.

Linguo, L.; Minsi, W.; Bingbing, N.; Hang, W.; Jiancheng, Y.; and Wenjun, Z. 2021. 3D Human Action Representation Learning via Cross-View Consistency Pursuit. In *CVPR*.

Liu, Y.; Wang, K.; Li, G.; and Lin, L. 2021. Semantics-Aware Adaptive Knowledge Distillation for Sensor-to-Vision Action Recognition. *IEEE Transactions on Image Processing*, PP.

Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; and Bajcsy, R. 2013. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, 53–60.

Tian, Y.; Henaff, O. J.; and van den Oord, A. 2021. Divide and Contrast: Self-supervised Learning from Uncurated Data. arXiv:2105.08054.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Multiview Coding. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 776–794. Cham: Springer International Publishing.

Wang, J.; Chen, Y.; Hao, S.; Peng, X.; and Hu, L. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119: 3–11.