Zero-Shot Learning for IMU-Based Activity Recognition Using Video Embeddings

CATHERINE TONG*, Department of Computer Science, University of Oxford, UK
JINCHEN GE*, Department of Computer Science and Technology, University of Cambridge, UK
NICHOLAS D. LANE, Department of Computer Science and Technology, University of Cambridge, UK

The Activity Recognition Chain generally precludes the challenging scenario of recognizing new activities that were unseen during training, despite this scenario being a practical and common one as users perform diverse activities at test time. A few prior works have adopted zero-shot learning methods for IMU-based activity recognition, which work by relating seen and unseen classes through an auxiliary semantic space. However, these methods usually rely heavily on a hand-crafted attribute space which is costly to define, or a learnt semantic space based on word embedding, which lacks motion-related information crucial for distinguishing IMU features. Instead, we propose a strategy to exploit videos of human activities to construct an informative semantic space. With our approach, knowledge from state-of-the-art video action recognition models is encoded into video embeddings to relate seen and unseen activity classes. Experiments on three public datasets find that our approach outperforms other learnt semantic spaces, with an additional desirable feature of scalability, as recognition performance is seen to scale with the amount of data used. More generally, our results indicate that exploiting information from the video domain for IMU-based tasks is a promising direction, with tangible returns in a zero-shot learning scenario.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computing methodologies** → **Learning paradigms**.

Additional Key Words and Phrases: human activity recognition, zero-shot learning, cross-modal knowledge transfer

ACM Reference Format:
Catherine Tong, Jinchen Ge, and Nicholas D. Lane. 2021. Zero-Shot Learning for IMU-Based Activity Recognition Using Video Embeddings. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 180 (December 2021), 23 pages. https://doi.org/10.1145/3494995

1 INTRODUCTION

When developing a Human Activity Recognition (HAR) system for IMU data, the first step is to define a set of activity classes to recognize. This determines what data to collect, how to collect them, and importantly, what activities can be recognized at test time. Once model learning is complete, the developed system generally cannot handle challenging scenarios where instances of new activities classes are encountered, in which case the system will output an uninformative *"other"* class (if included in training) or predict a wrong label belonging to the seen classes.

*Both authors contributed equally to this research.

Authors' addresses: Catherine Tong, eu.tong@cs.ox.ac.uk, Department of Computer Science, University of Oxford, UK; Jinchen Ge, Department of Computer Science and Technology, University of Cambridge, UK; Nicholas D. Lane, ndl32@cam.ac.uk, Department of Computer Science and Technology, University of Cambridge, UK.

2474-9567/2021/12-ART180 $15.00
https://doi.org/10.1145/3494995

In reality, the scenario of encountering unseen activities is common in practical applications of HAR systems. The primary reason is that the number of human activities is large but existing datasets only cover a limited number of them [54] (most IMU datasets [44, 45, 47, 52] contain fewer than 20 activity labels), which means users are bound to perform activities which are not currently recognizable. Although it is possible to build larger IMU datasets so that the HAR systems are trained with more diverse activities in the first place, the high costs associated with data collection and annotation remain prohibitive. Further, since the types of activities performed varies with user depending on lifestyle, environment and occupation factors, it is hard to collect sufficient IMU instances for each specific activity. These limitations inhibit the use of HAR systems in the continuous monitoring of Activities of Daily Living (ADLs), which is of great interest to psychology and healthcare research communities [38, 48, 58].

Zero-Shot Learning (ZSL) [57] offers a framework to solve the problem of recognizing previously unseen activities. Its development has been especially driven by applications related to images, videos and natural language processing (NLP). The main principle behind most zero-shot learning methods is to associate seen and unseen classes through some auxiliary information, usually defined by a *semantic space*. This approach is analogous to how humans recognize unfamiliar objects – a common example given in computer vision is that we can recognize *"zebras"* without seeing one before, if we were given the information that *"zebra looks like horses with stripes"*, and that we can recognize *"horses"* and *"stripes"*. Therefore, the semantic space will need to contain rich information about both seen and unseen classes, and be related to the feature space. By using such a semantic space, ZSL methods can determine the class labels for instances of unseen classes, even when no labelled instances were available.

In the field of Activity Recognition from IMU data, there have been few prior studies on zero-shot learning [19, 36, 56, 59]. Most studies in this direction adopt a hand-engineered attribute space, one of the most widely used semantic spaces in zero-shot learning. However, while attributes can be intuitively defined for tasks like animal recognition (e.g. attributes could be colour or habitat), defining attributes for activities are difficult since the way activities are performed may change considerably across individuals and time. As a result, domain expert knowledge is required to define the attribute space, which may look entirely different under different studies. Using a learnt semantic space facilitates a principled approach to refining the zero-shot learning problem.

Although there have been attempts to employ learnt semantic spaces built from word embeddings of class labels or descriptions, they have been found to give varying performance compared to manual attributes [36, 59]; This is unsurprising given the semantic gap between words and IMU signals (since words lack motion-specific information), as well as confusions when the activities have compound names [21].

In this work, we propose a novel strategy to construct an informative semantic space for IMU-based HAR using videos of human activities. With this strategy, feature representations of videos are extracted from state-of-the-art video action recognition models trained from thousands of videos and hundreds of activity classes. These rich feature representations, which we refer to as *video embeddings*, are then used to construct a semantic space to relate seen and unseen classes. Not only does this video-based semantic space circumvent the need to manually define attributes, it also manifests the transfer of knowledge from video-based HAR models to an IMU-based HAR problem.

We conducted systematic experiments on three public IMU datasets to evaluate zero-shot learning approaches based on the three aforementioned semantic spaces: *attributes*, *word*, and *video*. Our experiments show that the video semantic space is consistently superior to text-based methods, and comparable to and sometimes even better than that of attribute-based methods. By varying the number of videos used per activity, we demonstrate that only a small number of videos are required for each activity to achieve reasonable accuracy. Our analysis suggests that a video semantic space is scalable, as the recognition performance is found to increase with the number of videos used; This is desirable property since, in contrast to IMU data, video data can be collected easily from public online repositories and curated datasets. More generally, our work highlights the value in exploiting

information and resources from other modalities (in our case, video data and video HAR models) for tackling long-standing challenges faced in ubiquitous activity recognition using IMU data.

We summarize the key contributions of this paper as follows:

- We propose a novel video-based semantic space for zero-shot learning of IMU-based activity recognition.
- We propose a practical strategy to construct the video-based semantic space by exploiting a state-of-the-art pre-trained video activity recognition model.
- We empirically show that our video-based semantic space is scalable, and achieves strong zero-shot learning performances against word-based and attribute-based approaches.
- We investigate a combined approach to leverage both video-based and word-based semantic spaces for improved zero-shot learning performance.

2 RELATED WORK

We outline connections and differences of our work to the following lines of research.

Zero-Shot Learning. Zero-shot learning describes a scenario where a classifier is asked to determine the class labels of instances that belong to unseen classes, which have no labelled instances during model learning (or adaptation) [57]. Our approach belongs to the broad group of projection-based methods, in which instances are projected from the feature space to a semantic space, where the classification is carried out using classifiers such as the nearest neighbour classifier. One popular and pioneering projection-based method is the Direct Attribute Prediction (DAP) [33], which trains multiple SVMs to predict/map each of the binary attributes in the semantic space separately, so that each SVM simply classifies one semantic attribute/property. The final predictions are then inferred using Maximum A Posteriori (MAP) estimation. However, separate classifiers mean the semantic space must be a manual attribute space because the numbers in most learnt spaces typically have no explicit meaning. Therefore, later works such as Attribute Label Embedding (ALE) [4] usually predict the entire semantic embedding at once using a projection function that is as simple as a linear matrix projection. Linear mappings are limited since usually the instances are not linearly separable, so nonlinear mappings are also introduced by many works [6, 49, 62], some of which are unsurprisingly based on neural networks. For example, Socher et al. propose [49] to use a multilayer perceptron (MLP) for nonlinear projection.

IMU-based Zero-Shot Human Activity Recognition. Few previous works which implement zero-shot learning exist for the application of IMU-based activity recognition. Cheng et al. [19] proposed a method that is similar to DAP [33], which employs separate SVMs to predict each attribute of a binary attribute semantic space. Thereafter, a typical nearest neighbour classifier is used for classification. Cheng et al. [18] later extended their work by replacing the SVMs with the conditional random field (CRF) and the nearest neighbour classifier with the junction tree algorithm [25], but the attribute semantic space remains the same. Wang et al. [56] proposed a nonlinear-compatibility-based method, which is equivalent to using an MLP to project from the feature space to the attribute semantic space. Ohashi et al.'s work [39] employs a CNN to extract features from raw IMU data and perform the projection at the same time. Their semantic space is still an attribute space, but it contains binary, discrete and continuous attributes. Moreover, they also introduced the idea of attributes' importance so that instances from different classes would have different emphasis on the attributes, but the importance table is also manually defined. The use of learnt semantic spaces, as defined by word embeddings, was not studied until recent works [36, 59]. In [36], Matsuki et al. compared the use of attribute vectors and two variations of word embedding vectors and found their performances to be similar. Most recently, Wu et al. [59] proposed an MLP with skip connections for projection, and they found in their experiments that a manually defined attribute space gives superior performance to a semantic space defined by Word2Vec [37]. While previous work predominantly utilises the manual attribute space, and even word embedding spaces have not been fully explored, our work instead explores both word embedding spaces and the novel video embedding space.

Learning from Cross-Modal Virtual IMU Embeddings. One motivation of our work is to leverage the video domain, which provides rich datasets and modelling resources, to benefit activity recognition on IMU data, where data collection costs are high. This motivation is shared by a growing body of recent works which use videos as a source domain to generate virtual embeddings for activity recognition on IMU data [31, 32, 35, 46] as well as other sensing data [2, 9]. An approach developed for tri-axial virtual IMU data generation from videos is IMUTube [31, 32], which applies a sequence of pose estimation, tracking and relevant video processing techniques to infer the 3D motion of human bodies depicted in videos, which is then used to extract virtual IMU data as if the sensors were placed on different parts of the body. Although this strategy could be used to generate IMU signals for an increased number of activity classes, this type of method still tends to require a large number of videos to start with, because from one video, one could at most extract a piece of IMU signal that has the same length as the video. In comparison, our method only requires very few videos per class (as few as 1 video per class) to enable the recognition of unseen classes at inference time. Moreover, since the goal of these methods is to synthesize virtual IMU data, they are often contingent on accurate pose tracking, which makes them highly sensitive to occlusions of the human body and thus constrained by the quality of the collected video data. In contrast, our method does not require pose tracking and employs video activity recognition models which make use of the *entire* video clip, including pixels not just from the human body, to learn informative embeddings for activity classification.

3 PROPOSED APPROACH

In this section, we detail the design of our proposed video-based zero-shot learning approach.

3.1 Problem Formulation

Zero-shot learning tackles the problem of classifying classes that have no labelled instances available. We denote the set of seen classes as $\mathcal{S} = \left\{c_i^s\right\}_{i=1}^{N_s}$ and the set of unseen classes as $\mathcal{U} = \left\{c_i^u\right\}_{i=1}^{N_u}$, where $\mathcal{S} \cap \mathcal{U} = \emptyset$. During training, we are given the training dataset $D_{train} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N} \in \mathcal{X} \times \mathcal{S}$, where N is the number of training instances, $\mathcal{X} \in \mathbb{R}^d$ is the feature space, and all instances are from the seen classes $\mathcal{S}$. During testing, we are given $D_{test} = \{\mathbf{x}_i\}_{i=1}^{N_{test}} \in \mathcal{X}$, and these test instances belong to the unseen classes $\mathcal{U}$.

Since the classifier cannot learn from any labelled instances belonging to the unseen classes, a semantic space needs to be defined to solve the zero-shot learning problem. In the semantic space, every (seen and unseen) class has a corresponding vector representation called a *class prototype*. We denote the semantic space as $\mathcal{P} = \{\mathbf{p}_k\}_{k=1}^{N_s+N_u}$, where $\mathbf{p}_k \in \mathbb{R}^F$ is the prototype for class c_k.

We adopt a projection-based approach for zero-shot learning, where the main idea is to obtain labelled instances for the unseen classes by projecting the instances from the feature space $\mathcal{X}$ onto the semantic space $\mathcal{P}$.

For an instance $\mathbf{x}_i$, the projection function $h(\cdot)$ is defined as follows:

$$\mathcal{X} \rightarrow \mathcal{P} : \mathbf{z}_i = h(\mathbf{x}_i) \tag{1}$$

After projection, classification is performed in the semantic space. Due to the limited number of labelled instances in the semantic space (the prototype is the only labelled instance of a class), it is common practice to use a nearest neighbour classifier (1NN) to output the final prediction class.

The overall workflow can be summarized into the following three stages:

(1) We use a pre-trained video HAR model $\phi(\cdot)$ to extract the prototypes $\mathcal{P}$ of all seen classes $\mathcal{S}$ and unseen classes $\mathcal{U}$ from collected video data.
(2) Given training instances $(\mathbf{x}_i, y_i) \in D_{train}$ belonging to seen classes $\mathcal{S}$, we learn the projection function $h(\cdot)$ to project IMU features $\mathbf{x}_i$ to $\mathbf{z}_i$.

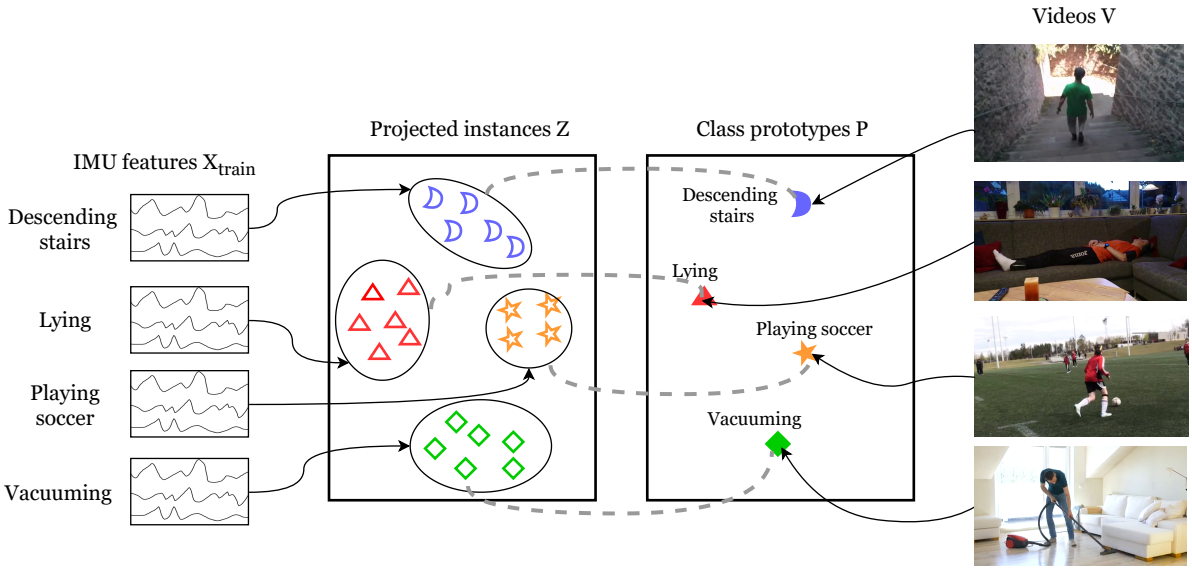

Fig. 1. Overview of our projection-based method for zero-shot learning. The right side illustrates the construction of a video semantic space, which is done by passing video data through a pre-trained I3D model to compute class prototypes. The left side shows the projection of IMU features into the video semantic space, done using a 4-layer MLP.

(3) Given testing instances $\mathbf{x}_i \in D_{test}$ belonging to unseen classes $\mathcal{U}$, we use the learnt projection function to project IMU features $\mathbf{x}_i$ to $\mathbf{z}_i$. Then, we use nearest neighbour classification to retrieve the corresponding class of the closest prototype $\mathbf{p}_k$.

In the following sub-sections, we describe our implementations of these three steps in detail.

3.2 Constructing the Video Semantic Space

We describe our strategy for constructing the video semantic space. This assumes the availability of video data belonging to activities described in the seen $\mathcal{S}$ and unseen classes $\mathcal{U}$ (we describe the collection of video data in Section 4.2.)

We extract the video embedding prototype $\mathbf{p}_k \in \mathbb{R}^F$ of an activity as the set of features of this activity's video extracted with a pre-trained video HAR model ϕ:

$$\mathbf{p}_k = \phi(\mathbf{v}_k), \tag{2}$$

where $\mathbf{v}_k$ is the raw video data of activity class c_k. If multiple videos are collected for each activity class, then we define the prototype as the mean of individual feature vectors:

$$\mathbf{p}_k = \frac{1}{N_k^{vid}} \sum_{n=1}^{N_k^{vid}} \phi(\mathbf{v}_k^n) \tag{3}$$

where N_k^{vid} is the number of available videos of activity class c_k.

In our experiments, ϕ is defined as video activity recognition model I3D [13]. Specifically, we use an I3D model pre-trained on Kinetics-400 [26]. The Kinetics-400 dataset is a commonly-used benchmark for video activity recognition, containing 306, 245 labelled videos in total for 400 activity classes, and I3D is a popular architecture

achieving state-of-the-art results on the dataset. The I3D architecture employs a 3D-CNN to treat the temporal dimension in videos as an extra spatial dimension. There are two streams in the I3D architecture, namely the RGB stream and the optical flow stream. The RGB stream inflates an Inception-v1 [51] architecture by endowing 2D filters with a third dimension and copying weights pre-trained on ImageNet [20] across this third dimension as initialization. The extra optical flow stream shares the same architecture, but in our pipeline, in the interest of efficiency, only the RGB stream is kept.

The construction of the video embedding space is represented in the right side of Figure 1. Videos of each activity class are passed into the RGB branch of the aforementioned I3D model, and a forward pass is conducted to extract features from the penultimate or last layer, which we refer to as *I3d-features* or *I3D-logits* respectively. Embedding vectors from *I3D-feature* has dimension $F = 1024$, and they can be seen as the output of the I3D feature extractor, encoding high-level features which are most useful for differentiating activities from video data. Vectors from *I3d-logits* are the outputs of the last layer before the final activation layer, thus giving the probability distribution over activities in the Kinetics 400 dataset, with dimension $F = 400$.

3.3 Projecting IMU Features into Video Semantic Spaces

In order to relate the IMU feature space and video semantic space, we project the IMU feature vectors into the extracted video embedding space. Since the focus of our study is the effectiveness of different semantic spaces, we follow closely the projection operation described in [59]. This projection operation is depicted on the left side of Figure 1.

The input to the pipeline is the IMU features $\mathbf{x}_i$ extracted from raw IMU data using the sliding window method. We define the projection function $h(\cdot)$ as a 4-layer Multilayer Perceptron (MLP) that is similar to that used in [59] and apply batch normalization [24] between all layers of the model. Each of the first 3 layers in the MLP $h(\cdot)$ is a typical fully connected layer with ReLU activation defined as

$$\mathbf{h}_l = relu(W_l \mathbf{h}_{l-1} + b_l), \tag{4}$$

where l is the layer number, W_l is the weight matrix and b_l is the bias. The MLP has two skip connections connecting from the first and second layer to the third layer, so the definition of the fourth (output) layer is

$$\mathbf{h}_4 = relu(W_4[\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3] + b_4). \tag{5}$$

The reason behind these skip connections is that the output of previous layers may also help the classification [59]. Finally, the output of the whole MLP $h(\cdot)$ is the projected instance

$$\mathbf{z}_i = h(\mathbf{x}_i) = \mathbf{h}_4. \tag{6}$$

After projection, classification is performed with a 1 Nearest-neighbour (1NN) classifier, which uses the cosine similarity as the distance metric between the projected instance $\mathbf{z}_i$ and any class prototype $\mathbf{p}_k$:

$$SIM(\mathbf{z}_i, \mathbf{p}_k) = \frac{\mathbf{z}_i \cdot \mathbf{p}_k}{\|\mathbf{p}_k\|_2}, \tag{7}$$

The Softmax probability of $\mathbf{x}_i$ belonging to a class c_k can thus be written as:

$$p\left(y_i = c_k \mid \mathbf{z}_i\right) = \frac{\exp\left(SIM\left(\mathbf{z}_i, \mathbf{p}_k\right)\right)}{\sum_{q \in \mathcal{T}} \exp\left(SIM\left(\mathbf{z}_i, \mathbf{p}_q\right)\right)}, \tag{8}$$

where $\mathcal{T}$ is $\mathcal{S}$ during training and $\mathcal{U}$ during testing.

This gives the final prediction as

$$\hat{y}_i = \arg\max_{c_k \in \mathcal{T}} p\left(y_i = c_k \mid \mathbf{z}_i\right), \tag{9}$$

3.4 Learning the Projection Function

To train the projection mapping, we adopt the Cross Loss function introduced in [59]:

$$L = \sum_{i=1}^{N} \|\mathbf{z}_i - \mathbf{p}^{y_i}\|_2 - \lambda \sum_{i=1}^{N} \sum_{k \in \mathcal{S}} y_i^k \log\left(p\left(y_i = c_k \mid \mathbf{z}_i\right)\right) \tag{10}$$

,

where $\mathbf{p}^{y_i}$ is the prototype of ground truth class y_i and y_i^k is the k^{th} entry of the one-hot representation of y_i. The terms in the loss function are explained as follows:

- The first term $\sum_{i=1}^{N} \|\mathbf{z}_i - \mathbf{p}^{y_i}\|_2$ denotes the projection error. This ensures that the learnt projection function maps the IMU features to a vector more similar to the video prototype vector.
- The second term $-\sum_{i=1}^{N} \sum_{k \in \mathcal{S}} y_i^k \cdot \log\left(p\left(y_i = c_k \mid \mathbf{z}_i\right)\right)$ denotes the cross-entropy loss on the prediction of the seen activity classes from the IMU vector.
- λ is a trade-off parameter between the projection and prediction losses.

3.5 Combining Semantic Spaces

In addition to the standard case of employing one semantic space, we may also employ multiple spaces for inference at test time, for instance by combining word and video spaces. In the following, we describe a method to combine two semantic spaces based on distance vector combinations.

First, we train two ZSL models using the two spaces separately. During inference, the test instance is passed through both models so that we would get two projections $\mathbf{z}^A$ and $\mathbf{z}^B$. Before nearest neighbour classification, in the respective semantic spaces of $\mathbf{z}^A$ and $\mathbf{z}^B$, we calculate their distances measurements to all label prototypes and get two distance vectors Δ^A and Δ^B. Then the combined distance vector for the space *(A+B)* is defined as:

$$\Delta^{comb} = (1 - \alpha)\Delta^A + \alpha\Delta^B, \tag{11}$$

where α is a parameter for adjusting which of the semantic spaces carries more weight in influencing the prediction. The combined distance vector Δ^{comb} is then used to identify the closest prototype and make a prediction.

4 DATA COLLECTION AND IMPLEMENTATION

4.1 IMU Data

Three publicly available datasets are used in our experiments, namely the PAMAP2 [44, 45] dataset, the DaLiAc [34] dataset and the UTD-MHAD [17] dataset. In each dataset, we use data from triaxial accelerometers and gyroscopes.

- **PAMAP2** The PAMAP2 [44, 45] Physical Activity Monitoring Data Set consists of IMU data of 9 subjects and 18 daily activities. Each subject wears 3 IMUs positioned on the wrist, the chest and the ankle. Before feature extraction, we remove 10 seconds from the start and end of each activity recording to make sure no noise is included. Following [45], we segment the raw data into sliding windows with a window size of 5.12 seconds and an overlap of 1 second. Within each window, we calculate the mean and standard deviation and then save them as features. In total, we extract 24222 instances, with each instance containing 36 dimensions.
- **DaLiAc** The DaLiAc (Daily Life Activities) database [34] contains IMU data collected from 19 subjects performing 13 daily life activities (e.g. *"washing dishes"*, *"vacuuming"*). We merge two classes, *"bicycling on ergometer (50 W)"* and *"bicycling on ergometer (100 W)"*, into one as it is probably unnecessary to differentiate

them and impossible to find corresponding videos for each of them. Each subject wears 4 IMU sensor nodes, located on the left ankle, the right hip, the chest, and the right ankle. We again use the sliding window method with 5.12 seconds of window size and 1 second of overlap to extract mean and standard deviation. Finally, we got 21889 instances, with each instance containing 48 dimensions.

- **UTD-MHAD** The UTD Multimodal Human Action Dataset (UTD-MHAD) [17] is very different from PAMAP2 and DaLiAc because it contains short actions (e.g. *"waving"*, *"clapping"*) and it was collected for multi-modal action recognition instead of IMU-only HAR. Therefore, apart from IMU data, UTD-MHAD also contains time-synchronized RGB videos and depth information. For IMU data collection, two IMUs were placed on the wrist and thigh of 8 subjects, and each subject performed 27 classes of actions. Since the actions in UTD-MHAD are very short, we do not use the sliding window approach, but instead treat the whole recording as one single window. This results in 861 instances, and each instance has 12 dimensions.

Train-Test Split. Following [56], we adopt a k-fold evaluation approach to split the activity classes in each dataset into train (seen) and test (unseen) portions over k disjoint folds. On PAMAP2, we adopt the same train-test split defined in [56] ($k = 5$), which leaves 3 to 4 unseen classes representing different activity types per fold. For DaLiAc, we randomly select 3 unseen classes in each of $k = 4$ folds. For UTD-MHAD, we randomly select 5 to 6 unseen classes in $k = 5$ folds. To ensure that the resulting train-test splits are balanced in the types of activities, in both cases we first manually identify the types of activities that a class belongs to, then randomly select classes from each type of activity to populate the test set of each fold. More details on train-test splits can be found in Appendix A.

4.2 Video Data

To build the video embedding space, there should be at least one video clip per (seen and unseen) activity class. In total, we collect 10 videos per class for the 18 activity classes in PAMAP2, and 13 activity classes in DaLiAc. For activity classes that appear in both datasets (e.g. *"walking"*), videos are reused for both datasets. Figure 2 demonstrates example frames of the collected videos for 6 random activities.

The data collection of videos for PAMAP2 and DaLiAc activities are guided by the given class labels. Where possible, we source annotated videos from public datasets, namely HMDB51 [30], UCF101 [50] Kinetics [12, 26] and RealWorld [52], by manually browsing videos annotated with semantically similar activities, e.g. we refer to videos annotated as *"using computer"* in Kinetics when sourcing videos for *"computer work"*. If not enough relevant videos cannot be found, we use keyword-driven search to download online videos from YouTube. As a result, the collecting videos are a mixture of amateurly or professionally recorded, with varying visual quality, camera stability, backgrounds, identity and number of human subjects. There may also be a small number of frames in each clip where the said activity cannot be identified due to occlusion, lighting conditions, or blurriness. We cap the video duration at 2 minutes with a minimum of 5 seconds. To extract video embeddings from the I3D model, we randomly select 512 consecutive frames from each clip, which represents 17 seconds assuming an average frame rate of $50Hz$; shorter videos are repeated to the 512 frames before input.

For UTD-MHAD, we directly use the videos provided by the dataset to extract the embeddings. There are 32 video clips available for each class label, as 8 subjects performed each activity 4 times. These videos have limited variability as the camera and the background are always fixed, and the subjects are always located in the centre. The videos have a frame rate of $15Hz$ so we use 256 consecutive frames to extract video embeddings from the I3D model for consistency.

4.3 Baseline Semantic Spaces

Attribute Space. We also consider attribute spaces, which act as baselines. These spaces are built with manually defined attributes. For the PAMAP2 dataset, we adopt the attributes from the work of Wang et al. [56], which is

Fig. 2. Example frames from the collected videos belonging to activity classes in the PAMAP2 or DaLiAc datasets. The source of each video frame is listed in Appendix B.

based on the movement of body (e.g. *"arms straight (or not)"*) and related objects & environment (e.g. *"indoor (or not)"*). We then manually define the attribute space for the DaLiAc dataset and the UTD-MHAD dataset using a similar set of attributes. We arrive at attribute spaces of PAMAP2, DaLiAc and UTD-MHAD with 42, 45 and 29 dimensions respectively. These are documented Appendix C.

Word Embeddings. We consider two word representation models, Word2Vec [37] and GloVe [41] to extract word embeddings. Both models leverage the co-occurrence of words in a large text dataset to measure their similarity, thereby creating a word semantic space where words with similar meanings are closer, and vice versa. Although both Word2Vec and Glove embeddings are popular in zero-shot learning problems in other domains, Word2Vec embeddings were only recently introduced in the field of IMU-based activity recognition [36, 59], so our experimentation with Glove embeddings represent a novel application altogether.

To extract the word embeddings from class labels, we use the Word2Vec model trained on the Google News dataset, as well as the GloVe model trained on the Wikipedia 2014 and the Gigaword 5. Both models produce 300-dimensional word vectors. For class labels with multiple words such as *"car driving"*, we compute the average of the feature vector of every word. For the PAMAP2 and the DaLiAc dataset, we use class labels provided by the authors. We make slight alterations to the labels of the UTD-MHAD dataset to make them more consistent and descriptive (e.g. *"catch"* to *"hand catch"*).

Random Embedding Space. We additionally consider a random embedding space as a sanity check for the effectiveness of the semantic spaces and the ZSL method. Here, each class prototype is a randomly generated 400-dimensional vector. Intuitively, a method using this random embedding space should have the accuracy of a random guess.

4.4 Model Implementations

Owing to our zero-shot learning task and k-fold cross-validation setup (Section 4.1), we keep the amount of hyperparameter tuning to a minimum so as not to be influenced by the unseen class result during model development. This is in keeping with best practices in the ZSL domain since tuning hyperparameters based on the cross-validation test results (which contains the unseen classes) would violate the zero-shot assumption [14, 61]. We therefore fix any required hyperparameters after reviewing the ranges of common settings in IMU-based activity recognition literature [43], and after a round of initial experiments which confirm that the relative performances of all methods are not changing. We arrive at the following: We employ the ADAM optimizer [28] with learning rate of 10^{-4}, adopted L_2 regularization with $\lambda = 10^{-4}$ and ran training with 15 epochs with a batch size of 64. Unless otherwise specified, we follow these configurations throughout the paper. The models and supporting functions are implemented in Python with the PyTorch [40] framework, and model training is carried out mainly on an NVIDIA Tesla P100 GPU or an NVIDIA GeForce RTX 2070 Max-Q GPU.

5 EXPERIMENTS

In this section, we describe the experiments conducted to examine the proposed zero-shot learning pipeline, with a particular focus on the effectiveness of the video-based semantic space. In each experiment, we discuss the underlying question, methods used, the results along with any implications and analysis.

5.1 Experiment I: Effectiveness of the Video Semantic Space

Question. The first question we investigate is if we are able to create a meaningful video semantic space to facilitate zero-shot learning in IMU-based activity recognition. **Does the Video Semantic Space Work?**

Method. We collect 10 videos per activity class to construct a video-based semantic space for zero-shot learning. We evaluate our pipeline on the three benchmark datasets, namely PAMAP2, DaLiAc and UTD-MHAD, using the I3D-features space, the I3D-logits space as defined in Section 3.2, as well as their combined space I3D-both (following Section 3.5). For comparison, we also perform controlled tests on semantic spaces of random, attribute and word embeddings. We use average per-class accuracy as our main evaluation metric throughout the paper as it is the most widely used metric for existing zero-shot learning methods [56]. We also report macro F1-Score here as it is a common metric used in IMU-based activity recognition.

Results. We report the performance comparison between video and other embeddings in Table 1. We see that video embeddings deliver strong ZSL performance when compared to alternate semantic spaces. Considering I3D-features and I3D-logits alone, they achieve consistently better performance than word embeddings by a large margin (accuracy improvement ranges from 13% to 33%). In addition, I3D-both is seen to give performance gains when compared to even the hand-crafted attribute space. Overall, this is a positive sign that videos of human activity, even those collected in-the-wild without much curation, encode a more informative representation for relating seen and unseen classes.

Interestingly, the difference in performance given by the two layers of the same video network (I3D-features and I3D-logits) is substantial (average difference of 12% in accuracy), and I3D-logits is seen to give better performance on PAMAP2 and UTD-MHAD. We believe this is related to the larger overlap in the nature of activities present in these datasets and Kinetics, on which the I3D model was pre-trained. Since the 400 dimensions of video-based class prototypes can be interpreted as a probability distribution over the 400 activity classes, this information may be more useful for representing the unseen classes found in PAMAP2 and UTD-MHAD (which, like Kinetics, contain many sports-related actions and activities). On the other hand, the I3D-features space extracts feature representations that are less fine-tuned to the particulars of the Kinetics label space, and may be more generally informative. The significant performance gains observed when combining both spaces into I3D-both seen on

Table 1. Comparison of zero-shot learning approaches when different semantic spaces are used. Best results amongst the individual learnt spaces are bolded. Best overall results are coloured in blue.

Semantic space		PAMAP2		DaLiAc		UTD-MHAD	
		accuracy	F1-score	accuracy	F1-score	accuracy	F1-score
Hand-crafted	Random	28.5	23.4	35.6	28.8	18.6	16.3
	Manual attributes	60.6	54.7	70.7	63.2	40.8	37.8
Learnt (Word)	Word2Vec	42.5	38.4	59.6	53.6	32.6	29.1
	GloVe	40.5	33.8	60.0	57.9	24.3	21.0
Learnt (Video)	I3D-features	49.3	45.7	65.5	**60.2**	36.4	32.8
	I3D-logits	**56.4**	**49.5**	**68.0**	58.0	**42.8**	**38.5**
	I3D-both	**65.5**	**61.2**	**71.6**	**66.2**	**43.4**	**39.5**

PAMAP2 and DaLiac (average gains of 8% in accuracy) indicates that these representations can be complementary and effectively combined to improve performance.

Visualizations. Figure 3 shows the per-activity F1 performances given by the semantic spaces of video, word and manual attributes when evaluated on the unseen classes in each fold. We see that the zero-shot learning problem is indeed a challenging problem as no single semantic space performs consistently well without fail across all activity classes, though their relative performances may reveal the limitations of each space. Word2Vec performs especially poorly in many activities which are predominantly characterised by their motions (such as *"running"* and *"playing soccer"* in PAMAP2 and *"descending stairs"* in PAMAP2 and DaLiAc), which may be related to the limited amount of motion-specific information contained in word-based spaces. Weak Word2Vec performance on some sedentary activities (such as *"watching TV"* in PAMAP2) may be attributed to the varying definitions of the composite word vectors, which may be too non-specific for activity identification. While the video-embedding space seems to perform well on most of the aforementioned activity examples, they still struggle in certain cases (e.g. confusing *"rope jumping"* with *"descending stairs"* or *"ascending stairs"*) where the motion characteristics alone may not lend enough specificity for their identification in a zero-shot learning setting.

The UTD-MHAD dataset presents the hardest scenario for all three types of semantic spaces, possibly because the activity classes here involve the most fine-grained distinctions between motions, e.g. *"swipe left"* was almost completely misclassified by all three spaces. We see word embeddings perform relatively well when class labels involve phrases that are prescriptive, such as *"pick up then throw"* and *"two hand push"* which neither video nor manual attributes spaces are able to recognize. We see poor Word2Vec performance in classes where the label is a single word, e.g. *"wave"* and *"squat"*, which suggests that common feature representations of verbs alone may not contain sufficient motion-related information needed for an IMU zero-shot learning problem. Moreover, the contrasting performance of word embeddings on *"sit then stand"* versus *"stand then sit"* suggests its difficulty in separating sequential actions; in contrast, both activities can still be classified to varying degrees by the video and attribute embedding space. In addition, we also see that video embeddings perform better in most cases where visual knowledge is expected (e.g. *"drawing circle clockwise"*).

5.2 Experiment II: Scalability of the Video Semantic Space

Question. In the previous section, we have curated 10 video clips per activity class in order to derive the video prototype vectors in PAMAP and DaLiAc. In this section, we examine how zero-shot classification performance varies with the amount of video data used. **Is the Video Semantic Space Scalable?** We believe this is an important investigation that can shed light on the trade-offs between video data curation and model performance.

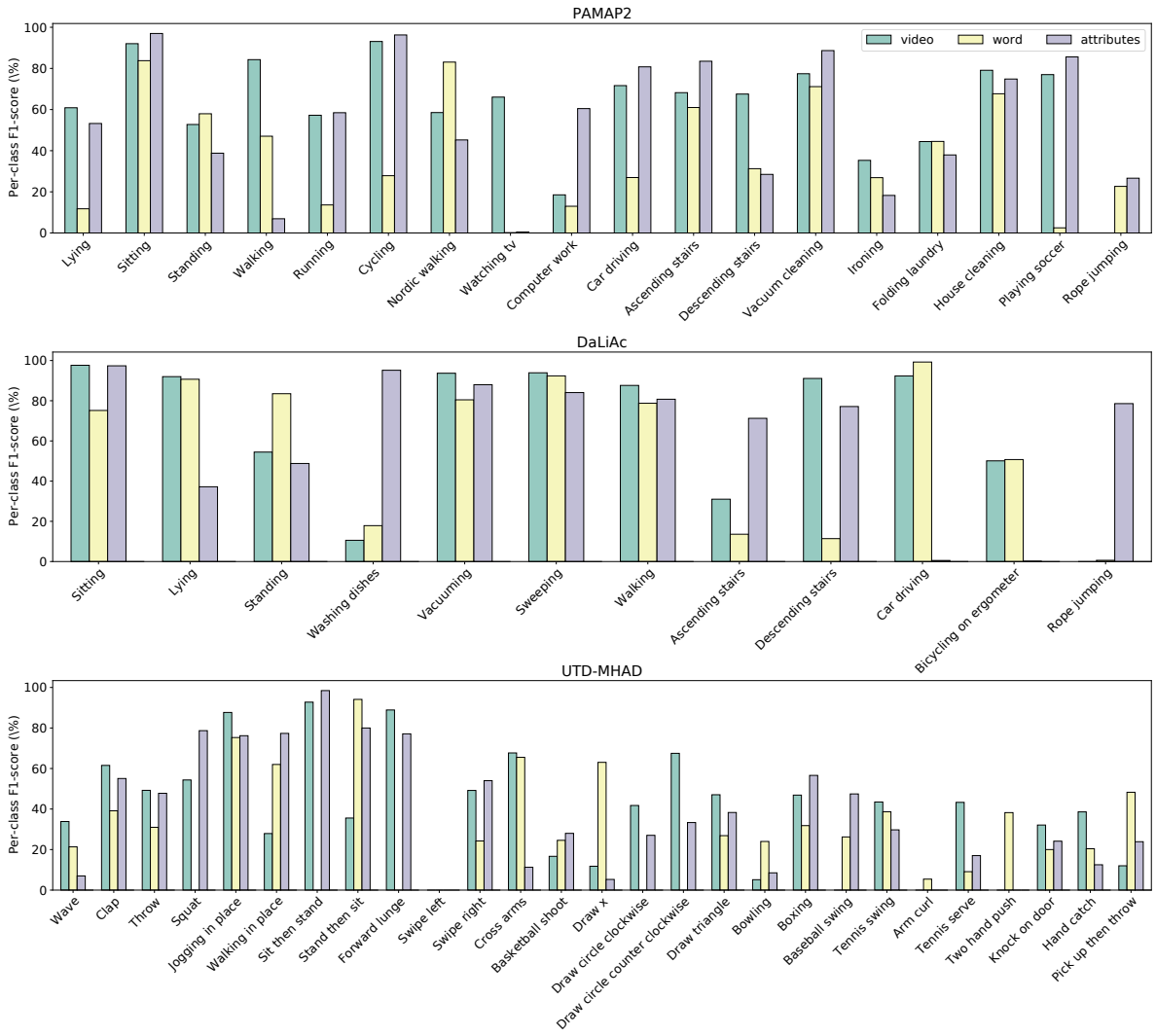

Fig. 3. Per-class F1 scores of all activities encountered in k-fold evaluation on each dataset. ZSL approaches using I3D-both, the best-performing word-based space (i.e. GloVe on DaLiAc and Word2Vec on PAMAP2 and UTD-MHAD), and manual attributes are plotted here for comparison.

If we observe that the ZSL can perform reasonably well with fewer videos, this may translate to lower data curation costs as we need not curate 10 videos per activity. Additionally, we are also interested in seeing if the model performance increases with the number of videos added, as that would suggest that we can potentially make use of the vast amount of video resources to improve performance in IMU-based HAR and open up many exciting research opportunities.

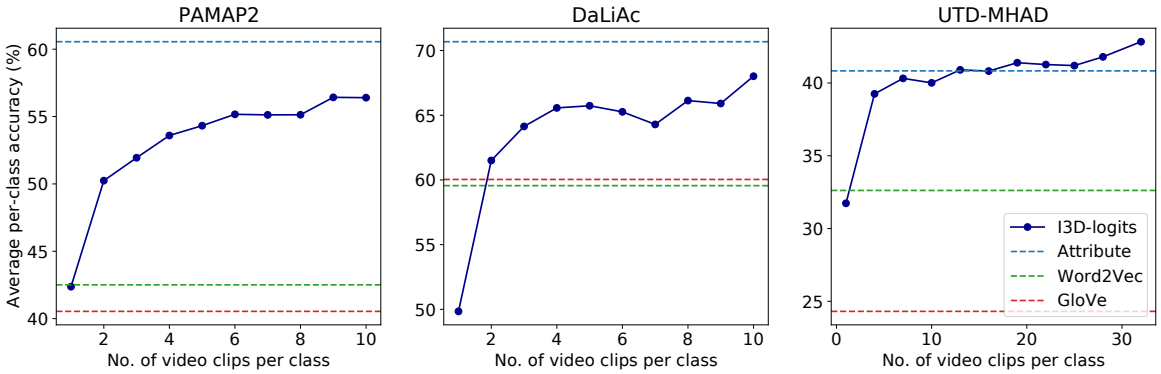

Fig. 4. The relationship between the number of videos per class and observed classification accuracy.

Method. We conduct a new set of experiments on PAMAP2, DaLiAc and UTD-MHAD by varying the number of videos per class n used to compute the video embedding vector. We randomly sample n videos from a total of $N = 10$ for PAMAP2 and DaLiAc, and $N = 32$ for UTD-MHAD. For each n, we repeat the sampling of videos 10 times and report the averaged results reached after training for 10 epochs per run. We conduct these experiments using the I3D-logits space.

Results. The experiment results are shown in Figure 4. A positive trend is observed in all three datasets, where classification accuracy is increased as the number of videos per class n is raised. The jump in accuracy is especially visible in the region of $n \leq 3$. Across the three datasets, one sees that the improvement gradually lowers as n gets closer to 10. This could indicate that 10 is already a good balance on the trade-off between accuracy and video collection, but if more accuracy is needed, there could still be some room for further growth. For the DaLiAc, it is even possible that the performance of the I3D features space could surpass that of the attribute space if after $n > 10$. In addition, it is worth noting that only $n \geq 3$ videos per class are enough for the video embedding spaces to outperform the word-based approaches.

We note that it is surprising to also observe such a positive trend on the UTD-MHAD dataset, since the 32 videos available per class (4 repeated actions by 8 subjects) have little visual difference when judging by the human eye. There is still an upward trend even when n is close to 32. Although we have adopted a simple averaging approach to aggregate multiple video embedding vectors into one class prototype, these results suggest that this simple strategy is enough to retain information from an increasing number of video data, such that the resulting increase in accuracy is still substantial.

Overall, we believe the results from this experiment reveal a desirable property for using video embeddings to provide auxiliary information for zero-shot learning, when compared to the attribute or word semantic space, which do not provide a similar route for further improvement. To increase the amounts of data, researchers can collect more videos in a similar manner to our data collection protocols described in Section 4.2, which have a much lower cost of curation than IMU data, or investigate well-established data augmentation strategies developed in computer vision (e.g. cropping, rotating). Therefore, researchers can make use of this property to carefully balance the trade-off in model performance and data curation cost according to their own needs.

Table 2. Comparison of zero-shot learning approaches when different combined semantic spaces are used. Best results are coloured in blue.

Semantic space		PAMAP2		DaLiAc		UTD-MHAD	
		accuracy	F1-score	accuracy	F1-score	accuracy	F1-score
Combined	I3D-features + Word2Vec	55.0	51.7	68.5	63.9	44.6	40.9
	I3D-features + GloVe	52.6	48.9	70.5	68.6	40.0	36.4
	I3D-logits + Word2Vec	**61.6**	**55.9**	68.9	64.4	**48.8**	**45.9**
	I3D-logits + GloVe	60.1	54.1	**74.0**	**70.2**	44.8	41.4

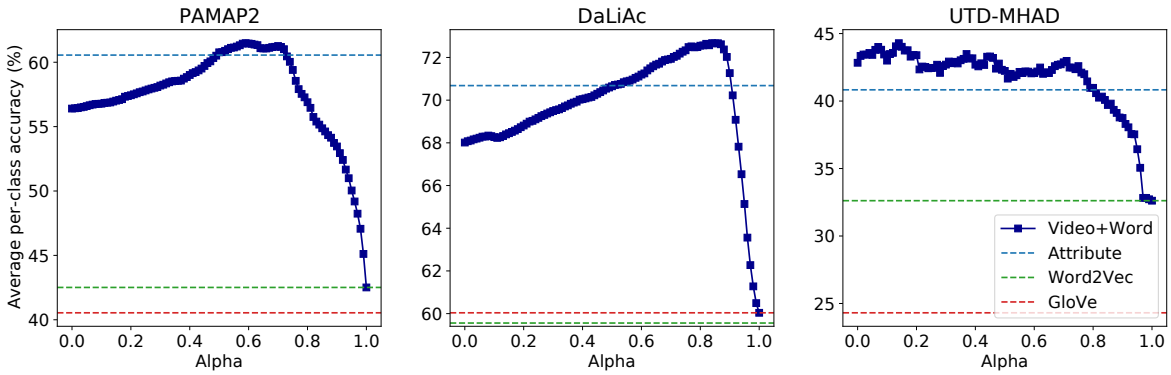

Fig. 5. Classification accuracy when α is varied from 0 to 1 in steps of 0.01. $\alpha = 0$ represents the case of learning with video-based space only, and $\alpha = 1$ represents the case of word-based space only.

5.3 Experiment III: Combining Semantic Spaces

Question. In Experiment I, we saw that the word and video semantic space each offers a different approach in creating an informative class representation for zero-shot learning. Meanwhile, previous works in ZSL for general applications [3, 5, 15, 60] and video-based HAR [29, 55] have found that performance improvements may be achieved by combining different semantic spaces; The intuition is that different semantic spaces encode complementary relationships between the classes, and such combination may also be seen as a variant of ensemble model learning. In this experiment, we investigate the combination of alternative perspectives from words and videos in constructing a joint semantic space. **Is it beneficial to combine learnt embedding space for zero-shot learning problems in IMU-based HAR?**

Method. To answer this question, we construct combined word and video spaces using the trained models presented in Section 5.1 and report the best performance achieved by varying α from 0 to 1 in steps of 0.01 according to the formulation presented in Section 3.5.

Results. Table 2 report the best ZSL performance achieved by the combined semantic spaces. Best performances are seen by combinations of I3D-logits with the best-performing word-based space on each dataset, which increases accuracy by 9% on both PAMAP2 and DaLiAc and by 5% on UTD-MHAD. Figure 5 shows how the performance accuracy changes as α is adjusted on each dataset, where a higher weighting for the video embedding

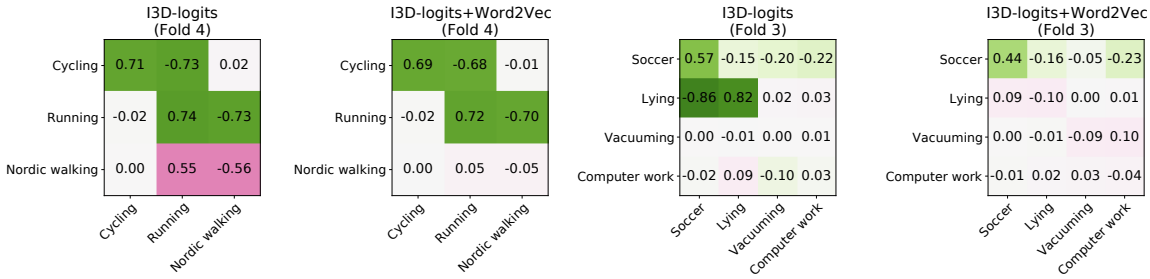

Fig. 6. Delta confusion matrices between I3D/I3D+Word2Vec and the Word2Vec on Fold 4 and Fold 3 on the PAMAP2 dataset. The depicted folds are chosen because I3D-logits achieve the best and worst accuracies on them. From left to right: Best fold (I3D-logits), best fold (I3D-logits+Word2Vec), worst fold (I3D-logits), worst fold (I3D-logits+Word2Vec). In each matrix, better and worse performance of video embeddings over word embeddings is indicated in green and red respectively; Therefore, on the main diagonal, a positive number appears green, and elsewhere, a negative number appears green.

is seen to be optimal for PAMAP2 and DaLiAc but not UTD-MHAD; this suggests that the combination of the semantic spaces must be carefully considered for each setting.

The observed performance gains could be explained by a smoothing effect brought by combining the semantic spaces, which we illustrate using delta confusion matrices on example folds of the PAMAP2 dataset in Figure 6. Each delta confusion matrix is computed by subtracting the normalised confusion matrix of Word2Vec from that of I3D-logits or I3D-logits+Word2Vec and indicates their relative performances. We see the smoothing effect leading to synergies of the two spaces in cases where I3D-logits struggle, most noticeably in the confusion between *"Nordic walking"* and *"running"*. Here, using the video embedding space alone has led to unsatisfactory results, possibly due to the visual similarity between *"Nordic walking"* and *"running"*, whereas the word embeddings of these activities are clearly different, leading to benefits when considering both information together. However, we note that the current simple combination approach does not always improve performance, as the benefits of individual embedding spaces may also be smoothed out, e.g. identifying *"lying"* in PAMAP2. Inspection of delta matrices across the datasets led to similar findings. This observation, together with the varying accuracy seen in Figure 5, both point to future work in investigating different techniques to fully exploit the benefits of combining semantic spaces.

6 DISCUSSION

There are still challenges and obstacles that could constrain the practical application of video-embedding-based ZSL. In this section, we discuss the limitations and potential future extensions of our work.

Activity Classes Found in IMU Datasets. In order to evaluate ZSL systems, there needs to be a large number of activity classes so that the training and testing portions of the datasets would contain enough seen and unseen activities respectively. Evaluation of ZSL problems for IMU HAR is currently difficult due to the limited diversity of activity classes present in public datasets. In addition, since our method assumes the availability of video data corresponding to each activity class, the mismatch in activity class labels between IMU HAR datasets and Video HAR datasets also present another limitation. For example, many IMU datasets have large amounts of data for static classes such as (e.g. *"standing"*, *"sitting"*, *"lying"*), which are rarely captured in long durations in video dataset; Even when they are captured, the available video data often focus on the transitional portion of these activities (e.g. going from sitting to standing and vice versa) instead of the static activity itself. However, we expect that the development of methods such as [31] will be helpful in addressing these issues, as these methods

also rely on video data collected from the same sources as that proposed in this paper, and simulated IMU data can be generated on demand.

Indistinguishable Videos. There are a number of activities which may look visually similar but are actually different in reality and measurable by IMU data, e.g. *"lift going up"* and *"lift going down"*. We expect that video embedding space would fail on tasks seeking to distinguish such activities. We therefore suggest that a good ZSL pipeline should take this into consideration in the definition of activity classes and reflect activities that are both visually and inertially distinguishable; one might also consider combined semantic spaces so that the class representation does not only consider videos as its source.

Impact of Changing Video Quality. We posit that our method is robust to common changes in video quality (e.g. blurriness, occlusion, varying lightness, changing camera angles) due to our use of pre-trained video activity recognition models which are robust to such variations. Specifically, we have employed the state-of-the-art I3D network, which was pre-trained Kinetics-400, a dataset consisting of at least 400 clips per activity sourced from YouTube. Thus, the resulting semantic space is given by a model which has already encountered videos which vary greatly in quality (e.g. the camera framing, viewpoint, how the action was performed, clothing, body pose), importantly these include videos which have considerable camera motion/shake, illumination variations, shadows, background clutter, etc [26]. Together with image augmentation during training, we expect the model to be able to extract embeddings while allowing for common variations in video quality. We have qualitatively confirmed these assumptions by inspecting different subsets of randomly selected videos and their performances used in Figure 4 (Section 5.2), where we observed no meaningful correlations between video quality and ZSL results. As a systematic evaluation of the empirical impact of video quality on ZSL performances would require a larger curated video dataset, we leave this investigation as a future work.

6.1 Future Work

Generalized Zero-Shot Learning. None of the previous works in IMU-based HAR have attempted a generalized zero-shot learning (GZSL) setting [7], which generalizes the zero-shot learning task to one where both seen and unseen classes are classified at test time. This is a realistic scenario for activity recognition tasks, since the seen classes (which we have training IMU data for) are often the most frequently performed activities, so it is likely that the seen activities are also targeting activities of a classifier. Since our current model has only been trained with instances belonging to the seen classes, the resulting predictions will bias towards seen classes if our model was naively applied for GZSL. There are some universal GZSL algorithms such as two-stage methods [23, 49] in which we first classify whether an instance belongs to seen classes or unseen classes, or the calibrated-stacking-based methods [8, 16, 22], in which the seen classes are deliberately disadvantaged during the final *argmax*, but these methods are not related to embedding spaces. We believe the more interesting and promising direction is the generative-based methods [42], in which a conditional generative model is trained to take the semantic prototype and synthesize an instance of the corresponding class, and then many instances of unseen classes could be generated to train a classifier. We anticipate that if the semantic prototype is from a video semantic space, it should contain much richer domain-specific information, which could lead to more realistic synthesized instances.

Data Augmentation on Videos. As shown in Section 5.2, we observe that the classification accuracy increases with the number of videos used in constructing the semantic space, even when the videos are visually similar to each other in the case of UTD-MHAD. While collecting or recording more videos may incur additional time, it can be flexibly done depending on the research resources available. Another way to artificially generate more video instances with minimum cost is to perform data augmentation on the base videos. Spatially, random cropping, flipping and rotation could be easily applied to video frames. Temporally, window slicing may also be used to break long videos into shorter pieces.

Skeleton-Based HAR Models. To further reduce the domain gap between the IMU features and the class prototypes generated from video data, we may consider other video-based models. A promising class of models are those developed for skeleton-based HAR models [1]. The intuition is that the key points in the skeleton explicitly represents the movement of body parts, and activity recognition models from these skeleton joints will only focus on the video's motion-related information, just like IMU-based HAR models. Many open-sourced toolboxes exist which allow the easy extraction of skeleton joints from RGB videos (e.g. OpenPose [10]). In addition, if researchers are considering recording their own videos, using skeleton-based data are better for preserving the identities of human subjects.

7 CONCLUSION

In this paper, we propose a strategy to exploit videos of human activities to construct an informative semantic space for zero-shot learning problems on IMU-based activity recognition. This novel video semantic space effectively facilitates the knowledge transfer from state-of-the-art video action recognition models to unseen activity recognition using IMU data. We evaluated our method on three public IMU datasets for zero-shot learning and compared the results against traditional learnt and hand-crafted embedding spaces. Our results show that our video embedding space consistently outperforms a word embedding space, and is comparable to the accuracy of the manual attribute space. With these encouraging results, our investigation suggests that exploiting videos of human activities as well as their associated models developed in the computer vision space has great potential for improving activity recognition tasks using IMU data.

REFERENCES

[1] J.K. Aggarwal and Lu Xia. 2014. Human activity recognition from 3D data: A review. *Pattern Recognition Letters* 48 (2014), 70–80. https://doi.org/10.1016/j.patrec.2014.04.011 Celebrating the life and work of Maria Petrou.

[2] Karan Ahuja, Yue Jiang, Mayank Goel, and Chris Harrison. 2021. *Vid2Doppler: Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445138

[3] Zeynep Akata, Honglak Lee, and Bernt Schiele. 2014. Zero-shot learning with structured embeddings. *arXiv preprint arXiv:1409.8403* (2014).

[4] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2013. Label-Embedding for Attribute-Based Classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 819–826. https://doi.org/10.1109/CVPR.2013.111

[5] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. 2015. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2927–2936.

[6] Jimmy Ba, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov. 2015. Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions. *CoRR* abs/1506.00511 (2015). arXiv:1506.00511 http://arxiv.org/abs/1506.00511

[7] Abhijit Bendale and Terrance E. Boult. 2016. Towards Open Set Deep Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Yannick Le Cacheux, Hervé Le Borgne, and Michel Crucianu. 2018. From Classical to Generalized Zero-Shot Learning: a Simple Adaptation Process. *CoRR* abs/1809.10120 (2018). arXiv:1809.10120 http://arxiv.org/abs/1809.10120

[9] Hong Cai, Belal Korany, Chitra R. Karanam, and Yasamin Mostofi. 2020. Teaching RF to Sense without RF Training Measurements. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 120 (Dec. 2020), 22 pages. https://doi.org/10.1145/3432224

[10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *CoRR* abs/1812.08008 (2018). arXiv:1812.08008 http://arxiv.org/abs/1812.08008

[11] Smart Floor Care. [n. d.]. *man with vacuum cleaner at home MQ3FTYV*. Youtube. https://www.youtube.com/watch?v=7ITNixZ2FDA

[12] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A Short Note about Kinetics-600. *CoRR* abs/1808.01340 (2018). arXiv:1808.01340 http://arxiv.org/abs/1808.01340

[13] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4724–4733. https://doi.org/10.1109/CVPR.2017.502

[14] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. 2016. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5327–5336.

[15] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. 2016. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5327–5336.

[16] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. 2016. An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild. *CoRR* abs/1605.04253 (2016). arXiv:1605.04253 http://arxiv.org/abs/1605.04253

[17] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 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. https://doi.org/10.1109/ICIP.2015.7350781

[18] Heng-Tze Cheng, Martin Griss, Paul Davis, Jianguo Li, and Di You. 2013. Towards zero-shot learning for human activity recognition using semantic attribute sequence model. *UbiComp 2013 - Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 355–358. https://doi.org/10.1145/2493432.2493511

[19] Heng-Tze Cheng, Feng-Tso Sun, Martin Griss, Paul Davis, Jianguo Li, and Di You. 2013. NuActiv: Recognizing unseen new activities using semantic attribute-based learning. *MobiSys 2013 - Proceedings of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, 361–374. https://doi.org/10.1145/2462456.2464438

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[21] Valter Estevam, Helio Pedrini, and David Menotti. 2021. Zero-shot action recognition in videos: A survey. *Neurocomputing* 439 (2021), 159–175. https://doi.org/10.1016/j.neucom.2021.01.036

[22] Rafael Felix, Michele Sasdelli, Ian D. Reid, and Gustavo Carneiro. 2019. Multi-modal Ensemble Classification for Generalized Zero Shot Learning. *CoRR* abs/1901.04623 (2019). arXiv:1901.04623 http://arxiv.org/abs/1901.04623

[23] Chuanxing Geng, Lue Tao, and Songcan Chen. 2020. Guided CNN for generalized zero-shot and open-set recognition using visual and semantic prototypes. *Pattern Recognition* 102 (2020), 107263. https://doi.org/10.1016/j.patcog.2020.107263

[24] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR* abs/1502.03167 (2015). arXiv:1502.03167 http://arxiv.org/abs/1502.03167

[25] David Kahle, Terrance Savitsky, Stephen Schnelle, and Volkan Cevher. 2008. Junction tree algorithm. *Stat* 631 (2008).

[26] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *CoRR* abs/1705.06950 (2017). arXiv:1705.06950 http://arxiv.org/abs/1705.06950

[27] kidsfit4u. [n. d.]. *Kidsfit Spin Bike*. Youtube. https://www.youtube.com/watch?v=zTf_YJtr8E8

[28] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (12 2014).

[29] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. 2015. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of the IEEE international conference on computer vision*. 2452–2460.

[30] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

[31] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D. Abowd, Nicholas D. Lane, and Thomas Plötz. 2020. IMUTube: Automatic Extraction of Virtual on-Body Accelerometry from Video for Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 87 (Sept. 2020), 29 pages. https://doi.org/10.1145/3411841

[32] Hyeokhyen Kwon, Bingyao Wang, Gregory D Abowd, and Thomas Plötz. 2021. Approaching the Real-World: Supporting Activity Recognition Training with Virtual IMU Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–32.

[33] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 951–958. https://doi.org/10.1109/CVPR.2009.5206594

[34] Heike Leutheuser, Dominik Schuldhaus, and Bjoern M. Eskofier. 2013. Hierarchical, Multi-Sensor Based Classification of Daily Life Activities: Comparison with State-of-the-Art Algorithms Using a Benchmark Dataset. *PLOS ONE* 8, 10 (10 2013), 1–11. https://doi.org/10.1371/journal.pone.0075196

[35] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. When Video Meets Inertial Sensors: Zero-Shot Domain Adaptation for Finger Motion Analytics with Inertial Sensors. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation* (Charlottesvle, VA, USA) *(IoTDI '21)*. Association for Computing Machinery, New York, NY, USA, 182–194. https://doi.org/10.1145/3450268.3453537

[36] Moe Matsuki, Paula Lago, and Sozo Inoue. 2019. Characterizing word embeddings for zero-shot sensor-based human activity recognition. *Sensors* 19, 22 (2019), 5043.

[37] Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *CoRR* abs/1310.4546 (2013). arXiv:1310.4546 http://arxiv.org/abs/1310.4546

[38] David C Mohr, Mi Zhang, and Stephen M Schueller. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology* 13 (2017), 23–47.

[39] Hiroki Ohashi, Mohammad Al-Naser, Sheraz Ahmed, Katsuyuki Nakamura, Takuto Sato, and Andreas Dengel. 2018. Attributes' importance for zero-shot pose-classification based on wearable sensors. *Sensors* 18, 8 (2018), 2485.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[41] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[42] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, and Xi-Zhao Wang. 2020. A Review of Generalized Zero-Shot Learning Methods. *CoRR* abs/2011.08641 (2020). arXiv:2011.08641 https://arxiv.org/abs/2011.08641

[43] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. 2018. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–27.

[44] Attila Reiss and Didier Stricker. 2012. Creating and Benchmarking a New Dataset for Physical Activity Monitoring. In *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments* (Heraklion, Crete, Greece) *(PETRA '12)*. Association for Computing Machinery, New York, NY, USA, Article 40, 8 pages. https://doi.org/10.1145/2413097.2413148

[45] Attila Reiss and Didier Stricker. 2012. Introducing a New Benchmarked Dataset for Activity Monitoring. In *2012 16th International Symposium on Wearable Computers*. 108–109. https://doi.org/10.1109/ISWC.2012.13

[46] Vitor Fortes Rey, Peter Hevesi, Onorina Kovalenko, and Paul Lukowicz. 2019. Let There Be IMU Data: Generating Training Data for Wearable, Motion Sensor Based Activity Recognition from Monocular RGB Videos. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, United Kingdom) *(UbiComp/ISWC '19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 699–708. https://doi.org/10.1145/3341162.3345590

[47] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, Jakob Doppler, Clemens Holzmann, Marc Kurz, Gerald Holl, Ricardo Chavarriaga, Hesam Sagha, Hamidreza Bayati, Marco Creatura, and José del R. Millàn. 2010. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh International Conference on Networked Sensing Systems (INSS)*. 233–240. https://doi.org/10.1109/INSS.2010.5573462

[48] Sandra Servia-Rodríguez, Kiran K Rachuri, Cecilia Mascolo, Peter J Rentfrow, Neal Lathia, and Gillian M Sandstrom. 2017. Mobile sensing at the service of mental well-being: a large-scale longitudinal study. In *Proceedings of the 26th International Conference on World Wide Web*. 103–112.

[49] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. 2013. Zero-Shot Learning through Cross-Modal Transfer. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 935–943.

[50] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR* abs/1212.0402 (2012). arXiv:1212.0402 http://arxiv.org/abs/1212.0402

[51] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. *CoRR* abs/1409.4842 (2014). arXiv:1409.4842 http://arxiv.org/abs/1409.4842

[52] Timo Sztyler and Heiner Stuckenschmidt. 2016. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 1–9. https://doi.org/10.1109/PERCOM.2016.7456521

[53] Nick Taylor. [n. d.]. *002 042212 Basic Long Passing.MTS*. Youtube. https://www.youtube.com/watch?v=4ZaWbcKDrUE

[54] Catherine Tong, Shyam A. Tailor, and Nicholas D. Lane. 2020. Are Accelerometers for Activity Recognition a Dead-end? *CoRR* abs/2001.08111 (2020). arXiv:2001.08111 https://arxiv.org/abs/2001.08111

[55] Qian Wang and Ke Chen. 2017. Zero-shot visual recognition via bidirectional latent embedding. *International Journal of Computer Vision* 124, 3 (2017), 356–383.

[56] Wei Wang, Chunyan Miao, and Shuji Hao. 2017. Zero-Shot Human Activity Recognition via Nonlinear Compatibility Based Method. In *Proceedings of the International Conference on Web Intelligence* (Leipzig, Germany) *(WI '17)*. Association for Computing Machinery, New York, NY, USA, 322–330. https://doi.org/10.1145/3106426.3106526

[57] Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. 2019. A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Trans. Intell. Syst. Technol.* 10, 2, Article 13 (Jan. 2019), 37 pages. https://doi.org/10.1145/3293318

[58] Matthew Willetts, Sven Hollowell, Louis Aslett, Chris Holmes, and Aiden Doherty. 2018. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. *Scientific reports* 8, 1 (2018), 1–10.

[59] Tong Wu, Yiqiang Chen, Yang Gu, Jiwei Wang, Siyu Zhang, and Zhanghu Zhechen. 2020. Multi-Layer Cross Loss Model for Zero-Shot Human Activity Recognition. In *Advances in Knowledge Discovery and Data Mining*, Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan (Eds.). Springer International Publishing, Cham, 210–221.
[60] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. 2016. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 69–77.
[61] Yongqin Xian, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4582–4591.
[62] Hongguang Zhang and Piotr Koniusz. 2018. Zero-Shot Kernel Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7670–7679. https://doi.org/10.1109/CVPR.2018.00800

APPENDIX

A TRAIN-TEST SPLITS

The splits of each dataset used throughout our experiments are shown in Table 3, Table 4 and Table 5.

We observe that the splits adopted by Wang et al [56] ensured that at least one activity of each type is present in the training (seen) set and the test (unseen) set, according to the activity types which we have identified in Table 6. Inspired by this, we constructe similar tables for DaLiAc (Table 7) and UTD-MHAD (Table 8) and arrive at the current splits by randomly selecting classes of different activity types into the test set.

Table 3. Train-test split of the PAMAP2 dataset from Wang et al. [56].

Fold	Test Classes
1	watching TV, house cleaning, standing, ascending stairs
2	walking, rope jumping, sitting, descending stairs
3	playing soccer, lying, vacuum cleaning, computer work
4	cycling, running, Nordic walking
5	ironing, car driving, folding laundry

Table 4. Train-test split of the DaLiAc dataset.

Fold	Test Classes
1	sitting, vacuuming, descending stairs
2	lying, sweeping, car driving
3	standing, walking, bicycling on ergometer
4	washing dishes, ascending stairs, rope jumping

Table 5. Train-test split of the UTD-MHAD dataset.

Fold	Test Classes
1	swipe left, cross arms, draw triangle, arm curl, jogging in place, pick up then throw
2	swipe right, basketball shoot, bowling, tennis serve, walking in place, squat
3	wave, draw x, boxing, two hand push, sit then stand
4	clap, draw circle clockwise, baseball swing, knock on door, stand then sit
5	throw, draw circle counter clockwise, tennis swing, hand catch, forward lunge

Table 6. Activity types identified in the PAMAP2 dataset.

Activity Types	Classes
Static activities	lying, sitting, standing
"Walking" activities	walking, Nordic walking, ascending stairs, descending stairs
House chores	vacuum cleaning, ironing, folding laundry, house cleaning
Sports	running, cycling, playing soccer, rope jumping
"Sitting" activities	watching TV, computer work, car driving

Table 7. Activity types identified in the DaLiAc dataset.

Activity Types	Classes
Static activities	sitting, lying, standing
House chores	washing dishes, vacuuming, sweeping
"Walking" activities	walking, ascending stairs, descending stairs
Other activities	car driving, bicycling on ergometer, rope jumping

Table 8. Activity types identified in the UTD-MHAD dataset.

Activity Types	Classes
One-hand activities	swipe left, swipe right, wave, throw, knock on door, hand catch
Two-hand activities	clap, cross arms, arm curl, two hand push
Drawing activities	draw x, draw circle clockwise, draw circle counter clockwise, draw triangle
Sports	basketball shoot, boxing, baseball swing, tennis swing, tennis serve
Activities involving legs	bowling, pick up then throw, jogging in place, walking in place, sit then stand, stand then sit, forward lunge, squat

B SOURCE OF VIDEOS

The following table lists the video source of the example frames depicted in Figure 1 and Figure 2.

Table 9. Video source of selected example frames

Activity Classes	Sources of Example Frames
Bicycling on ergometer	Youtube video [27]
Descending stairs	RealWorld dataset [52]
Computer work	Kinetics dataset [12]
Lying	RealWorld dataset [52]
Playing soccer	Youtube video [53]
Vacuum cleaning	Youtube video [11]

C MANUALLY-DEFINED ATTRIBUTE SPACES

The following table lists the attribute-based class prototypes manually-defined for PAMAP2, DaLiAc and UTD-MHAD respectively.

Table 10. Manual attributes for the PAMAP2 dataset from Wang et al. [56].

Attribute Aspects	Body Movements																												Related Objects and Environments													
Activity \ Attribute	motion	static	cyclic motion	intense motion	translation motion	free motion	body vertical	body incline	body horizontal	body forward	body backward	body up	body down	body in place	torso transform	arms motion	arms static	arms bent	arms straight	arms bent-straight transform	hands hold something	legs motion	legs static	legs bent	legs straight	legs bent-straight transform	legs alternate move forward	legs move up and/or down	seat	bike	poles	television	computer	car	stairs	vacuum	iron	clothes	soccer	rope	indoor	outdoor
lying	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
sitting	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	1	1	1	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1
standing	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	1	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
walking	1	0	1	0	1	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
running	1	0	1	1	1	0	1	0	0	1	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
cycling	1	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	1	1	1	0	1	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1
Nordic walking	1	0	1	0	1	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
watching TV	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	1	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0
computer work	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0
car driving	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	1	0	0	1	1	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
ascending stairs	1	0	1	0	0	0	1	1	0	1	0	1	0	0	0	1	0	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	1	1
descending stairs	1	0	1	0	0	0	1	1	0	1	0	0	1	0	0	1	0	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1
vacuum cleaning	1	0	0	0	0	1	1	1	0	1	1	0	0	0	1	1	0	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
ironing	1	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0
folding laundry	1	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0
house cleaning	1	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0
playing soccer	1	0	0	1	0	1	1	1	0	1	0	1	1	0	1	1	0	1	1	1	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rope jumping	1	0	1	1	0	0	1	0	0	0	0	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1

Table 11. Manual attributes for the DaLiAc dataset.

Attribute Aspects	Body Movements																												Related Objects and Environments															
Activity \ Attribute	motion	static	cyclic motion	intense motion	translation motion	free motion	body vertical	body incline	body horizontal	body forward	body backward	body up	body down	body in place	torso transform	arms motion	arms static	arms bent	arms straight	arms bent-straight transform	hands hold something	legs motion	legs static	legs bent	legs straight	legs bent-straight transform	legs alternate move forward	legs move up and/or down	seat	bike	poles	television	computer	car	stairs	vacuum	iron	clothes	soccer	rope	ergometer	broom	dishes	indoor	outdoor
Sitting	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	1	1	1	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Lying	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Standing	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	1	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Washing dishes	1	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	
Vacuuming	1	0	0	0	0	1	1	1	0	1	1	0	0	0	1	1	0	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	
Sweeping	1	0	0	0	0	1	1	1	0	1	1	0	0	0	1	1	0	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0
Walking	1	0	1	0	1	0	1	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
Ascending stairs	1	0	1	0	0	0	1	1	0	1	0	1	0	0	0	1	0	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	
Descending stairs	1	0	1	0	0	0	1	1	0	1	0	0	1	0	0	1	0	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	
car driving	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	1	0	0	1	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Bicycling on ergometer	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	1	1	1	0	1	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	
Rope jumping	1	0	1	1	0	0	1	0	0	0	0	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	

Table 12. Manual attributes for the UTD-MHAD dataset.

Attribute Aspects	Body Movements																						Related Objects					
Activity \ Attribute	body always vertical	body inclined	body up	body down	torso transform	arms static	only one arm	two arm	arm(s) has bent	arms straight	arm(s) periodic motion	hands hold something	drawing (incl. swipe)	drawing return to origin	drawing left first	drawing continuous	legs static	legs motion	legs(s) small bent	legs(s) large bent	leg(s) periodic motion	foot relocate	foot no relocate	basketball	baseball	tennis	chair	bowling ball	object away
swipe left	1	0	0	0	0	0	1	0	1	0	0	0	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0
swipe right	1	0	0	0	0	0	1	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0
wave	1	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
clap	1	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
throw	1	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
cross arms	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
basketball shoot	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1
draw x	1	0	0	0	0	0	1	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0
draw circle clockwise	1	0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0
draw circle counter clockwise	1	0	0	0	0	0	1	0	1	0	0	0	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0
draw triangle	1	0	0	0	0	0	1	0	1	0	0	0	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0
bowling	0	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	1	1
boxing	1	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
baseball swing	1	0	0	0	1	0	0	1	1	0	0	1	0	0	0	0	1	1	1	0	0	1	1	0	1	0	0	0	1
tennis swing	1	1	0	0	1	0	1	0	1	1	0	1	0	0	0	0	1	1	1	0	0	1	1	0	0	1	0	0	1
arm curl	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
tennis serve	1	0	0	0	1	0	0	1	1	0	0	1	0	0	0	0	1	1	1	0	0	1	1	0	0	1	0	0	1
two hand push	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
knock on door	1	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
hand catch	1	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
pick up then throw	0	1	1	1	0	0	1	1	1	0	0	1	0	0	0	0	0	1	0	1	0	1	1	0	0	0	0	0	1
jogging in place	1	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	1	0	1	0	1	0	0	0	0	0	0
walking in place	1	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	1	0	1	0	1	0	0	0	0	0	0
sit then stand	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0
stand then sit	0	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0
forward lunge	1	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0
squat	0	1	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0