LightHART: Lightweight Human Activity Recognition Transformer

Syed Tousiful Haque[1], Jianyuan Ni[1], Jingcheng Li[2], Yan Yan[3], and Anne Hee Hiong Ngu[1]

[1] Texas State University, San Marcos Texas, USA
{bgu9,j_n317,angu}@txstate.edu
[2] University of New South Wales, Sydney, 2052, NSW, Australia
jingcheng.li@unsw.edu.au
[3] Illinois Institute of Technology Chicago, IL, USA
yyan34@iit.edu

Abstract. Human Activity Recognition (HAR) using wearable sensors has gained significant attention due to its portability and unobtrusiveness. However, the data obtained from wearable sensors are limited to inertial data from predefined locations on the human body. In contrast, skeletal data from motion capture devices, such as the Kinect camera, offer richer information by capturing the whole body dynamics of a human action. Unfortunately, the use of skeletal data is impractical in wearable sensor-based HAR for real-world deployment. Currently, transformer neural networks, known for their self-attention mechanism, have shown effective handling of data from diverse modalities in wearable sensor-based HAR. However, the deployment of multimodal transformer on wearable devices is challenging due to their inherent large model size. We propose a Lightweight HAR Transformer (LightHART) framework that trains an unimodal Inertial Transformer (IT) network by transferring knowledge from a large multimodal transformer using a knowledge distillation approach. We evaluate the proposed framework on three public multimodal human activity datasets and compare the performance of the LightHART student model with various state-of-the-art approaches. Experimental results demonstrate that our LightHART model achieves competitive performance in terms of effectiveness and scalability with a model size of only 1.43 Mb. We are the first to deploy and validate the LightHART fall detection model on a SmartFall App running on a WearOS-compatible smartwatch showcasing its potential in advancing wearable sensor-based HAR research.

Keywords: Human Activity Recognition, Transformer, Knowledge Distillation, Multi-modal Learning, Wearable Devices

1 Introduction

A wearable sensing system that can facilitate Human Activity Recognition (HAR) utilizing information extracted from diverse visual and inertial (accelerometer,

gyroscope, etc.) modalities can have a significant societal impact. For example, HAR can improve elder care in assisted living centers from timely detection of falls and timely administration of medication. In addition, HAR can also revolutionize diverse context-aware applications like fitness tracking, health monitoring, and gesture recognition, just to name a few [16].

Human perceives the world in a multimodal view, automatically integrating information from multiple sensors like vision, sound, touch, etc. It is known that multimodal deep learning approaches can leverage information from multiple sources like accelerometers, gyroscopes, and visual inputs and alleviate the limitation regarding unimodal approaches via complementary information, reducing the ambiguity of activity recognition, and being robust against noisy data. While the multimodal learning model offers various benefits for HAR problem, implementing them in wearable devices is challenging due to hardware limitations in executing models of large size and the inability to acquire the visual modality continuously with on-body sensors without compromising users' privacy.

Knowledge Distillation (KD) is a potential solution that can leverage multimodal algorithms for wearable devices. KD was first introduced in [9] to distill knowledge from large models *i.e. teacher* into smaller models *i.e. student*. Initially, a large complex model is trained with data suitable for the task. These models typically had a large number of parameters and thus can achieve high accuracy by learning rich representations. Next, a smaller model is trained on the same dataset, but instead of using only the ground truth labels, it is trained to mimic the behavior of the teacher model. To improve the performance of deep learning models on HAR tasks involving vision modality, particularly when dealing with occlusion, the authors of [13] introduced a multimodal knowledge distillation approach that integrates diverse sensor information. A cross-modal knowledge distillation method is introduced in [23] that transfers knowledge from multimodal to unimodal networks. Though this work aimed to produce a model for wearable devices, the ResNet18 student network used in this research resulted in a complex model that is not usable in wearable devices. A small Distilled Mid-fusion Transformer student model is produced by [14], but the student model only works in the presence of multimodal data, which makes it inappropriate for use in portable wearable devices since it is not possible to acquire the visual data in real-time while being mobile and free of the burden to carry a specialized on-body visual sensor. Meanwhile, previous studies applied several fusion methods in building effective multimodal model [22, 14]. For instance, the work in [22] uses a late fusion, and the authors in [14] introduce a Temporal Mid Fusion. However, these fusions don't take the spatial and temporal features into account at the same time and thus can't produce an effective knowledge representation when transferring to student models.

To leverage multi-modal learning on wearable devices, we propose a Lightweight HAR Transformer (LightHART) framework that produces an Inertial Transformer (IT), the student model, that can learn to mimic a Spatio-Temporal ConvTransformer (STConvT) teacher model. First, we train the STConvT model with data from multiple modalities *(i.e. skeleton, inertial)* and fuse the spatial

and temporal information using Attention Feature Fusion. We then train the student model on only inertial data (unimodal) guided by the feature representation acquired in STConvT using knowledge distillation. This LightHART framework tries to minimize the distillation loss during its training. After training, LightHART's student model can achieve competitive performance on three multimodal HAR datasets with a model size of only 1.43 Mb. We further tested and deployed the LightHART fall detection model (a specific type of human activity) on a SmartFall App [21] running on a WearOS-compatible smartwatch. The contributions of this paper are summarized as follows:

- We propose LightHART that generates a lightweight transformer model running on inertial modalities only. To our knowledge, this is the first study conducting a knowledge distillation process from a skeleton-to-inertial domain using an unimodal Transformer model which is lightweight.
- We propose a STConvT model with Attention Feature Fusion that can produce better feature representation aligning both spatial and temporal information.
- We demonstrated the effectiveness and generalization ability of the proposed LightHART method on three public datasets.
- We are the first to test and deploy the LightHART fall detection model on a real-world fall detection App to demonstrate its potential in advancing wearable sensor-based HAR research.

Our paper is organized as follows. In the related work in Section 2, we describe some background work on human activity recognition and the motivation behind choosing a transformer-based architecture. Next, we present the methodology and the architecture of LightHART in Section 3. We outline the setup of the Spatial and Temporal encoder blocks and the Attention Feature Fusion strategy. In Section 4, we describe the dataset used, the experimental setup, and the evaluation protocol used. In Section 5, we compare the performance of LightHART with other SOTA approaches. In Section 6, we conduct ablation studies to showcase the effectiveness of our fusion strategy and the spatial block. Finally, we discuss the implications of our findings and future directions for our work in the conclusion section.

2 Related Work

Human Activity Recognition: HAR is used to detect and classify human activities under appropriate labels. An activity refers to the collective movement of parts of the body to complete a task. For example, moving the head in negation is a gesture, and walking, jumping, and hand waving are activities [27]. The approaches to resolving the human activity recognition task can be divided into three types: vision-based HAR [2], sensor-based HAR [8], and multimodal HAR [14]. A wide spectrum of methods, ranging from traditional machine learning, rule-based, to deep learning methods have been used for HAR over the years.

An extensive comparison among K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM) for wearable sensor-based HAR is discussed in [1]. The early traditional machine learning approaches depend on features built by domain experts and can't efficiently differentiate between very similar activities such as walking upstairs and walking downstairs [26]. RNN, LSTM, and CNN are unimodal deep learning networks that have become popular in recent years and have achieved state-of-the-art in recognizing different HAR tasks. For example, an ensemble Recurrent Neural Network (RNN) method has been used in [17] to do fall detection from wearable devices. Multiple other research works such as those in [26, 19, 25] have used LSTM and a hybrid CNN-LSTM network for HAR.

Wearable devices using unimodal data have shown the promise of bringing personalized health monitoring closer to consumers [20]. For example, smartwatches like the Apple Series, which feature built-in "hard fall" detection and ECG monitoring apps, are a viable platform for digital health applications when paired with a smartphone. However, unimodal deep learning methods using data from wearable devices have certain limitations [30, 12]. Data produced by wearable sensors can be noisy, lack contextual information, and face difficulties discriminating among activities producing similar patterns. For example, if a person is wearing a watch on the left wrist and the left wrist does not move during a fall, the fall will be missed. Video or skeleton modalities can provide complementary and contextual information to unimodal data from wearable devices for better recognition of human activities. To capture information from both spatial and temporal domains, the authors in [16] introduced a multimodal network called AttnSense. DanHar framework was proposed in [7] to blend channel attention and temporal attention with a CNN model. However, none of the above multimodal models have a model size that is small enough for real-world deployment to a wearable device.

Transformer: Deep learning methods like LSTM and CNN have some inherent problems when used for HAR. Although LSTM can handle temporal dynamics in long sequences of data from human activities, their singular perception limits them in capturing complex patterns that require multiple viewpoints. Convolutional Neural Networks (CNNs) are primarily designed to extract local spatial patterns within data. By leveraging multiple layers, they can also capture more complex and global spatial features. However, CNNs are inherently limited in their ability to process temporal information. The continuous HAR signal patterns are more distinguishable when seen from a global temporal viewpoint. Transformer [28] possesses a global viewpoint courtesy of its self-attention layer, and the multiple heads in self-attention help to create multiple viewpoints. Transformer has already been used successfully in NLP, Computer Vision, Recommendation Systems, and many others. It also has been used in HAR. For example, the authors in [31] used a two-stream Transformer network to capture both spatial and temporal features from inertial data.

However, these multimodal networks aren't suitable for deployment in wearable sensors due to the unavailability of visual modalities in real-time [23, 13, 15]. Moreover, the constraints of computation power of wearable devices preclude the deployment of the usually large multi-modal learning model.

To the best of our knowledge, only a few studies by [6, 10, 32] have conducted efficient experiments on lightweight transformer-based architecture in HAR domain.

3 Methodology

In this paper, we introduce our LightHART framework that produces a lightweight (Inertial Transformer) student model from the knowledge distillation process that only uses inertial data and still maintains similar accuracy as the multimodal teacher model. An STConvT network works as a teacher by extracting the salient spatial and temporal features and using an Attention Feature Fusion to combine features from skeleton and inertia modalities effectively.

Figure 1 gives the overview of the knowledge distillation process that distillates knowledge from a multimodal teacher model to an unimodal student model. First, we train a multimodal STConvT teacher network with skeleton and inertial data. The input from different modalities is segmented using the sliding window technique described in [34]. We then add a learnable positional embedding to each of the modalities to preserve the positional information. A Spatial Block consisting of two convolutional layers extracts accurate spatial information from the skeleton data.

The output of the Spatial Block is divided into patches and passed on to a Temporal block which leverages ViT architecture [5]. The Temporal Block consists of two Transformer Encoders that apply a multi-head self-attention mechanism [28] on the patches to extract the salient temporal features while preserving spatial information. On the other hand, inertial data is passed to a separate Temporal Block. The features from the intermediate Transformer Encoder dedicated to the skeleton and inertial data are fused using Attention Feature Fusion and passed to an MLP layer for final prediction. Finally, a knowledge distillation procedure is used to transfer the feature representation learned by the teacher module to the student's Inertial Transformer(IT) that works on inertial data only. Our IT also adopts the ViT architecture [5]. In the following, we elaborate on the framework to produce the lightweight student model using the knowledge distillation procedure.

3.1 Inertial Transformer (IT)

The original transformer model consists of an encoder and a decoder. The encoder generates embeddings from the input, while the decoder uses these embeddings to produce output in a different language. However, for activity recognition, only the encoder is needed to extract both spatial and temporal information.

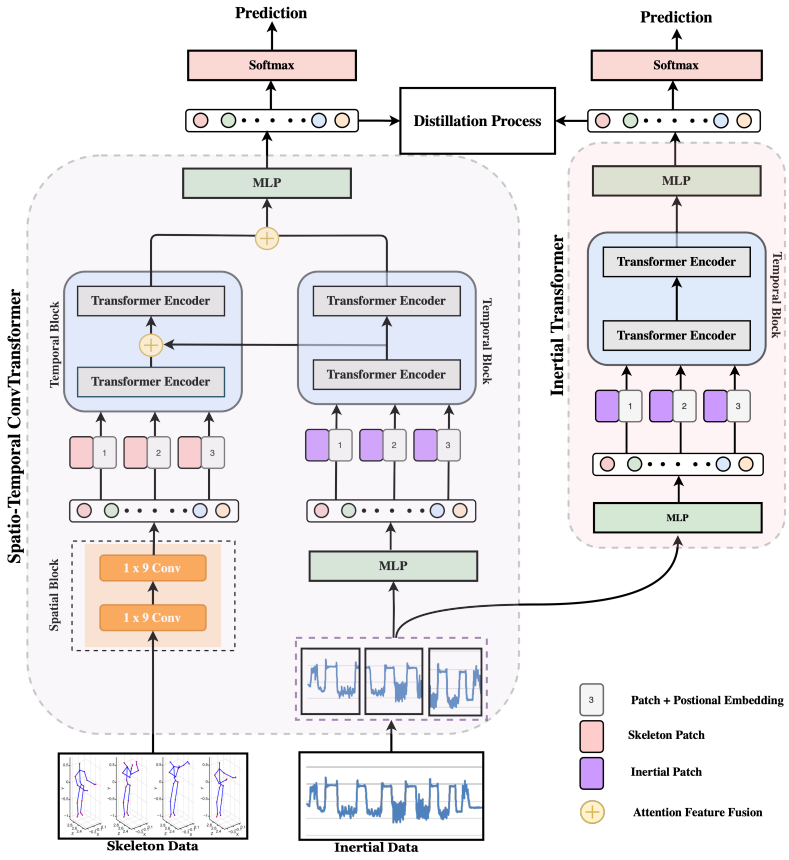


Fig. 1: LightHART framework with STConvT as teacher and IT as student.

In ViT, an image's input is first segmented into patches. We can think of the inertial data as 2-D images with shape (W, C_{iner}) where W is the window size and C_{iner} is the number of channels of inertial data. The input for IT $x \in \mathbb{R}^{(N\times(W\times C_{iner})/P)}$ is reshaped into a sequence of patches, where N is the number of patches and P is the patch size. The IT uses a constant embedding size of D through all its layers. The patches are transformed to D dimension using a linear layer (Eq. 1). A learnable class token is appended at the start of the sequence of embedding patches ($z_0^0 = x_{class}$) whose state at the output of the Transformer's encoderz_L^0 serves as the inertial data representation y. A learnable one-dimensional positional embedding E_{pos} is added to the patch embeddings. The Transformer encoder [28] comprising of interleaved layers of multiheaded self-attention (MSA) and MLP blocks is applied to the patches. Layer normalization (LN) is employed preceding each block for stabilized training, with residual connections following each block. The residual connection was used to avoid a vanishing gradient and ensure a direct flow of information. The class token of the last encoder block output is then passed to the MLP head with the softmax activation to get the final prediction.

To keep the network small, we construct it with only two Transformer encoder blocks with small embedding dimensions in the student's IT.

$$z_0 = [x_{class}; x_p^1 E; x_p^2; ...; x_p^N E;] + E_{pos} \quad E \in \mathbb{R}^{((W\times C_{iner})/P)\times D} \tag{1}$$

3.2 Spatio-Temporal ConvTransformer

The Spatio-Temporal ConvTransformer is made up of three important parts: *1. Spatial Block, 2. Temporal Blocks, and 3. Attention Feature Fusion* that helps it to analyze both spatial and temporal information effectively.

Spatial Block: The Spatial Block is depicted in Fig 1 in orange color. This module is in charge of dealing with the spatial details found in skeleton data. It uses two 2-dimensional (2D) convolution layers that could effectively extract the relationships between nearby joints. These layers have a special property called translation invariance inductive bias, making them particularly effective at processing spatial information. Let $x_{SK} \in \mathbb{R}^{(C_{SK}, J_{SK}, W_{SK})}$ is the skeleton input to the Spatial Block where C_{SK} is the channels of skeleton data, J_{SK} is the number of predefined joints and W_{SK} is the size of the window. The 2D Convolution layers in the Spatial Block take in an input of (C_{in}, H, W) where C_{in} is the number of channels, H is the height of input and W is width. To process the skeleton data with 2D Convolution Layer we set $C_in = C_{SK}$, $H = J_{SK}$, and $W = W_{SK}$. Both the convolution layers had a filter shape of $(1, 9)$ to gather spatial information from three adjacent joints. The Spatial Block (SP) produces an output s_p of shape (C_{out}, H_{out}, W), where C_{out} is the output channel size and H_{out} is the output height as of Eq. 2. The output of the Spatial Block is then reshaped to $(N, C_{out} \times H_{out} \times (W/P))$ where N is the number of patches and P

is the size of the patches.

$$s_p = \mathbf{SP}(X_{skl}) \qquad X_{skl} \in \mathbb{R}^{(C_{SK}, J_{SK}, W_{SK})} \tag{2}$$

Temporal Block: The Temporal Block has a structure that is the same as the IT. So, the sequence of skeleton patches s_p is transformed into $z_0^{skl} \in \mathbb{R}^{(B,P,D)}$ (Eq. 4) where D remains constant all across the network. Creating patches from the embedding will help the Temporal Block to process temporal information together [5]. We also process the inertial data with a Temporal Block. Let $X_{iner} \in \mathbb{R}^{(W \times C_{iner})}$ be the inertial data. This inertial is then reshaped to $i_p \in \mathbb{R}^{(N \times (W \times C_{iner})/P)}$. To match the dimension of the Transformer Encoder the input is transformed to $x \in \mathbb{R}^{(N \times D)}$ and 1-D learnable positional embedding E_{pos} and class embedding i_{class} was added (Eq. 3). The processed inertial data and output from the Spatial Block then go through the first Encoder on two different Temporal Blocks as shown in Fig 1 and produce embedding z_1^{skl} and z_1^{iner} (Eq. 5).

$$z_0^{iner} = [i_{class}; i_p^1 E; i_p^2; ...; i_p^N E;] + E_{pos} \qquad E \in \mathbb{R}^{((W \times C_{iner})/P) \times D} \tag{3}$$

$$z_0^{skl} = [s_{class}; s_p^1 E; s_p^2; ...; s_p^N E;] + E_{pos} \qquad E \in \mathbb{R}^{((C_{out} \times H_{out} \times W)/P) \times D} \tag{4}$$

$$z_1^m = \mathbf{Encoder}(z_0^m) \qquad m \in (iner, skl) \tag{5}$$

Attention Feature Fusion z_1^{skl} and z_1^{iner} are then added together to produce z_1^{comb} (Eq. 6). This fusion purpose is named as **Attention Feature Fusion(AFF)** as the output of transformer encoder layers dedicated to different modalities are fused. AFF merges complementary information from temporally aligned patches of different modalities. This fusion in terms helps the subsequent self-attention layer(MSA) in better exploring the relation between patches. For all subsequent layers, z_l^{comb} is produced by fusing z_{l-1}^{comb} and z_{l-1}^{iner} (Eq. 7). The final prediction y is generated by passing the class token $z_L^{comb_0}$ of the L-th encoder block (last) through an MLP layer (Eq. 8). A softmax function is used on the output of the MLP layer to produce the class predictions.

$$z_1^{comb} = \mathbf{Encoder}(z_1^{skl} + z_1^{iner}) \tag{6}$$

$$z_l^{comb} = \mathbf{Encoder}(z_{l-1}^{comb} + z_{l-1}^{iner}) \tag{7}$$

$$y = \text{softmax}(\mathbf{MLP}(z_L^{comb_0}) \tag{8}$$

3.3 Multimodal to Unimodal Knowledge Distillation

The knowledge distillation begins after we finish training the STConvT with skeleton and inertial data. During knowledge distillation, a teacher's STConvT takes multimodal (skeleton & inertial) data as input and the student's IT takes only the inertial data. In general, neural networks produce a class probability by

taking the logits and passing it through a softmax function $p_i = softmax(z^i)$. But, the knowledge distillation method in [9] used a soft prediction with parameter Temperature (T) (Eq. 9). The higher the temperature the softer the prediction. Both the teacher and the student produce soft predictions $P_{teacher}$ and $P_{student}$. These soft predictions are then compared using a KL-Divergence Loss (Eq. 10). The entropy between the ground truth y^{gt} and the student's (IT) final prediction y^{stud} is measured using a cross-entropy loss and added with the KL-Divergence loss to get the knowledge distillation loss $\mathcal{L}_{KD}$ (Eq. 11). The student model tries to mimic the teacher's prediction by minimizing this loss during its training.

$$p_i = \frac{e^{\frac{(z_i)}{T}}}{\sum_j e^{(\frac{z_j}{T})}} \tag{9}$$

$$\mathcal{L}_{kl}(P_{student}, P_{teacher}) = \sum_i P_{student,i} \log \frac{P_{student,i}}{P_{teacher,i}} \tag{10}$$

$$\mathcal{L}_{KD} = \mathcal{L}_{cross}(y^{gt}, y^{stud}) + \mathcal{L}_{kl}(P_{teacher}, P_{student}) \tag{11}$$

4 Experiments

4.1 Datasets

We evaluated the LightHART's performance on three human activity datasets. UTD-MHAD and Berkeley-MHAD are a few of the mainstream multimodal human activity recognition datasets publicly available. SmartFallMM is another multimodal human activity recognition dataset developed in our lab with a specific focus on fall detection.

The UTD-MHAD dataset [3] was collected using a single Kinect camera and one wearable inertial sensor. The Kinect camera captures full-body visual data during activities, while the inertial sensor records acceleration, gyroscope, and magnetometer data. The sensor was placed on the subject's right wrist or thigh, depending on whether the action primarily involved the arm or leg. The use of only a Kinect camera and inertial sensor is due to their low cost and non-intrusive nature. The dataset includes 27 actions performed by 8 subjects (4 males and 4 females), with each action repeated 4 times, resulting in 861 samples after excluding corrupted ones.

The Berkeley-MHAD dataset [24] consists of temporally synchronized and geometrically calibrated data from an optical mocap system, multi-baseline stereo cameras from multiple views, depth sensors, accelerometers, and microphones. We used the accelerometer data collected from the left wrist for our experiment. It contains 11 actions performed by 7 male and 5 female subjects in the range of 23-30 years of age except for one elderly subject. All the subjects performed 5 repetitions of each action, yielding about 660 samples which correspond to about 82 minutes of total recording time.

The SmartFallMM[1] multi-modal dataset comprises data from two distinct modalities, collected using four different types of devices. The skeleton data was gathered using three Azure Kinect cameras. Additionally, accelerometer and gyroscope data were obtained from three types of inertial sensors: Meta sensors (from MBIENT), a Huawei Smartwatch running WearOS, and a Google Nexus phone. This dataset includes a total of 14 activities, performed by 36 participants. Among these activities, 9 are Activities of Daily Life (ADL), and 5 are different fall activities, resulting in a total of 1,134 activity trials, and only 11 participants could perform fall activities. We used the accelerometer data sensed from Huawei SmartWatch and the skeleton data for our experiments.

4.2 Evaluation Protocol

For the UTD-MHAD dataset, we follow the established evaluation protocol outlined in the original paper [3]. Specifically, subjects with odd-numbered identifiers (1, 3, 5, 7) are designated for training purposes, while subjects with even-numbered identifiers (2, 4, 6, 8) are reserved for testing. Given the limited size of the dataset, this approach serves to maintain a balance between the sizes of the training and testing datasets. Moreover, the segmentation based on person IDs serves the dual purpose of preventing data leakage and ensuring the integrity of the evaluation process. We adhere to the evaluation protocol outlined in the original paper [24] for Berkeley-MHAD. The training dataset comprises of first 7 persons' data while the testing dataset consists of the last 5 persons' data.

We performed recognition of fall-related activities on SmartFallMM dataset with real-world testing and evaluation in mind, as we already have a fall detection system developed for a wearable device [21]. We used the first 9 persons' data for training and the last 2 persons' data for testing. After training an offline student IT model with LightHART, we deploy this model to a Huawei Smartwatch running the SmartFall App for real-time evaluation. Two student participants are recruited under IRB 9461 for the real-time evaluation. They performed all 9 ADLs and 5 Fall activities five times each activity wearing the smartwatch.

4.3 Experimental Setup

The inertial modality may contain multiple streams (e.g. the accelerometer and gyroscope) of data. Despite the presence of different streams, we consider them as a single modality since they are all time-series data. Skeleton data is sensed as a sequence of time-series (accelerometer) data from multiple skeletal nodes. Both skeleton and inertial data have variable lengths across activity trials and different sampling rates. To optimize training, we equalized the sampling rates and extracted synchronized windows of size 64 from both skeleton and inertial modalities, with a 10-timestamp overlap between windows. The STConvT architecture consists of 2 consecutive Convolution layers with both having a filter size of 9 to facilitate the extraction of spatial information from adjacent joints.

[1] Url: https://anonymous.4open.science/r/smartfallmm-4588

The two Temporal Blocks had two Transformer encoders each with an input dimension of 32. To optimize the model, we employed the SGD optimizer with a learning rate set at 0.0025 and utilized the knowledge distillation loss (Eq. 11) function during the training phase.

5 Studies & Results

5.1 Evaluations & Comparisons

We compared our LightHART's performance with other state-of-the-art multimodal transformers with knowledge distillation-based methods using inertial and skeleton data as input. Table 1 and 2 show the experimental results on UTD-MHAD and Berkeley-MHAD respectively. We evaluated SmartFallMM mainly for fall detection activities and is not included in this table. The inertial data from UTD-MHAD had two streams (accelerometer and gyroscope). We compared the performance of LightHART with multimodal transformer models like CrossVit [33], DMFT [14] and TokenFusion [29]. LightHART outperformed these transformer-based methods as it consecutively gains 8.67% and 14.44%, over TokenFusion [29], CrossVit [33]. Though DMFT [14] has a higher accuracy of 92.12%, it's worth mentioning that it had a complex architecture with 262.2× larger model size than the student model trained with LightHART which makes it infeasible for deployment in wearable devices. The increased accuracy of LightHART is primarily due to the knowledge distillation method. Before knowledge distillation, the accuracy of LightHART student's model was 73.618 % on UTD-MHAD dataset and the teacher Spatial-Temporal ConvTransformer had an accuracy of 89.81%.

Table 1: Performance comparison on the UTD-MHAD dataset. S: Skeleton, D: Depth, I: Inertial, aug:augumentation.

Method	**Modality Combination**	**Accuracy(%)**
UTD-MHAD [3]	I + D	81.86
Gimme Signals [18]	I + S	76.13
Gimme Signals [18]	I + S(aug)	86.53
TokenFusion [29]	I + S	78.89
CrossViT [33]	I + S	75.37
MobileHART(XS) [6]	I	77.52
DMFT [14]	I+S	**92.12**
LightHART(Teacher)	I + S	89.81
LightHART (Student)	I	73.62
LightHART(KD)	I	**87.56 (13.94 ↑)**

But after the knowledge distillation, the accuracy of the student model went up to 87.56% which is a 13.942% increase in accuracy. The LightHART student model also has an 10.037% accuracy gain over MobileHART(XS) [6] - a

lightweight Transformer model - which further supports the effectiveness of our LightHART framework.

The gap between teacher and student is as small as 2.25% which demonstrates that the STConvT supported by Attention Feature Fusion creates feature representations that the student's uni-modal IT (Inertial Transformer) can easily mimic.

Table 2: Performance comparison on the Berkeley-MHAD dataset. S: Skeleton, D: Depth, I: Inertial

Method	**Modality Combination**	**Accuracy(%)**
MMhar-Ensemblenet [4]	I + D	81.86
TokenFusion [29]	I + S	79.91
CrossVit [33]	I + S	75.37
DMFT [14]	I + S	78.18
LightHART (Teacher)	I + S	**85.69**
LightHART (student)	I	80.33
LightHART(KD)	I	81.93(1.60 ↑)

Similar trends are observed in the case of the Berkeley-MHAD dataset. The inertial modality had only the accelerometer stream for this dataset. LightHART outperformed multimodal Transformer networks like TokenFusion [29] and CrossVit [33] and DMFT [14][2] as it consecutively gains 3.04% and 6.56% and 3.75% . The knowledge distillation method effectively increased the accuracy of the student model by 1.6%. The gap between teacher and student was 3.76%. The accuracy gain after knowledge distillation was 1.6% which is lower than the UTD-MHAD dataset. This was due to the absence of a gyroscope stream in inertial data as gyroscopes provide much-needed information about angular velocity. We couldn't compare the results with MobileHART(XS) [6] as it required both gyroscope and accelerometer modalities.

Fig. 2 illustrates the performance comparison of STConvT, IT, and IT with KD on the SmartFallMM dataset. The teacher model, STConvT, achieved an accuracy of 99.75%, while the student IT model of LightHART had an accuracy of 77.0% before applying KD. By employing STConvT as the teacher during the knowledge distillation process, the accuracy of the IT model increased by 1.50% for fall detection.

Table. 3 shows the model size comparison of different multimodal Transformer models. The student model generated using LightHART had a model size of 1.43 Mb which is 262.2× smaller than DMFT [14] which has a model size of 375 Mb. The DMFT uses a ResNet50 pre-trained model size of 98 Mb. Even if they used an architecture without the ResNet50, the model size would still be

[2] DMFT wasn't originally evaluated on Berkley-MHAD datasets. We trained this model for 250 epochs for Berkley-MHAD to provide the same training time for fair comparison

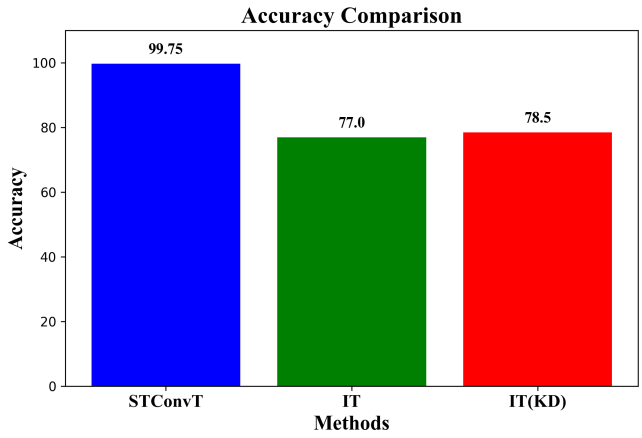


Fig. 2: Accuracy Comparison for Fall Detection Task of SmartfallMM dataset

193.7× larger than the student model generated by LightHART. CrossVit [33] and MobileHART(XS) also have 425× and 7.23× larger model sizes compared to our student model. Only TokenFusion [29] has a smaller model size than our student network. However, this smaller model size also compromises the accuracy as it drops to 78.89% for UTD-MHAD and 79.91% for Berkeley-MHAD. Overall, only our student model can maintain competitive performance while reducing the model size.

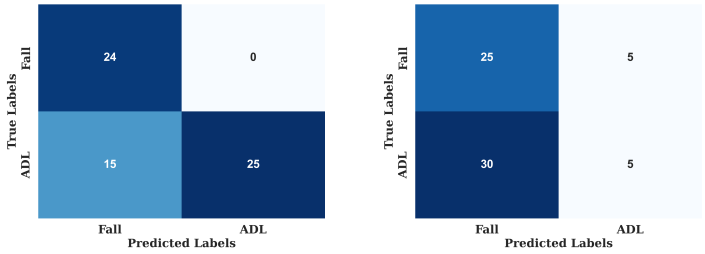
Table 3: Model Size comparison for different Transformer models

Modalities	Model	Model size(mb)
I	LightHART	1.43
I + S	TokenFusion [29]	**.68**
I + S	CrossVit [33]	608.09
I + R + S	DMFT [14]	375
I	MobileHART(XS) [6]	10.36

5.2 Performance on Wearable Devices

We ported two different IT models to run on a smartwatch, one generated by LightHART and the other purely based on uni-modal accelerometer data without knowledge distillation to observe the average inference time and performance. Both of our IT models running on the device could make an inference in .4459 ms to .8428 ms for a stream of data with a duration of 4 seconds compared to 1 to 13 ms for 2.56 seconds duration of data using MobileHART(XS) [6]. The

LightHART student IT model's performance improvement in fall detection task after training with KD can also be observed in Figure 3. Though both the models have a similar number of True Positive detection of 24 and 25, the IT model trained without KD cannot differentiate the intrinsic patterns between ADL and Fall activities as it can only detect 5 of 35 ADL activities accurately compared to 25 out of 35 of the student model trained with KD. The accuracy of the model without KD drops by 31.69% and becomes 45.31% during the on-device evaluation. The student's model trained with KD can maintain similar accuracy with on-device evaluation as its accuracy only becomes 76.56% which represents only a 1.94% drop. This on-device performance comparison shows models trained with KD can help maintain better performance.



(a) Performance of IT with KD (b) Performance of IT without KD

Fig. 3: Confusion matrices for on device performance of LightHART(student) with and without Knowledge Distillation

6 Ablation Studies

6.1 Effectiveness of Attention Feature Fusion

Table 4 shows the effectiveness of Attention Feature Fusion(AFF). For this experiment, we used the SimpleFusion [11], TokenFusion by [29], CrossView Fusion by [33] and Attention Feature Fusion(AFF) with our STConvT to observe which fusion methods have the most impact on the student model's accuracy. The result shows that the teacher model using AFF has a student model with the highest accuracy of 87.56%. Though the teacher network with CrossView Fusion had better accuracy, the representation was complex for a lightweight student model to mimic. Thus, the student had the lowest accuracy of 69.47%

6.2 Effectiveness of Convolution Spatial Block

Table 5 shows the impact of the Convolution Spatial Block. First, we changed the Spatial Block to a Transformer-like architecture with 2 encoders. The accuracy dropped to 77.12% in comparison to 89.81% for the Convolution Spatial

Table 4: Performance comparison of different fusion methods on UTD-Mhad dataset

Method	Teacher Accuracy(%)	KD Accuracy(%)
SimpleFusion [11]	87.68	84.36
TokenFusion [29]	85.00	70.04
CrossView Fusion [33]	**90.0**	69.47
AFF	89.81	**87.56**

Block. This shows that the Convolutional layers with an inductive bias for spatial information outperform vanilla transformers. On the other hand, a network without Spatial Block had an accuracy of 80.25% which is 9.56% lower than a model with Convolution Spatial Block.

Table 5: Performance comparison with and w/o Convolution Spatial Block

Method	Teacher Accuracy(%)
Transformer SB	77.12
W/O SB	80.25
Convolution SB	89.81

A supplementary study on the effectiveness of the Temporal Block is presented in Table 1 of the supplementary materials.

7 Conclusion

In this paper, we propose a LightHART network architecture to generate a lightweight transformer model (student) using unimodal inertial data that has a very small model size while retaining similar accuracy as the complex multimodal transformer (teacher) network in the case of UTD and Berkeley datasets. With SmartFallMM dataset, we show that the IT model with KD performs better than the one without. The experimental results also demonstrate that our lightweight student model with a model size of 1.43 Mb can achieve competitive performance as compared to other student models distilled from state-of-the-art multimodal learning frameworks. We further tested and deployed the LightHART student's model on a wearable smartwatch device running a fall detection App. The real-world testing of the model using two participants demonstrates the better performance of a uni-modal fall detection trained using a knowledge distillation approach. However, while we have demonstrated that a lightweight LightHART model can be deployed successfully on the device that outperforms the model without KD, there is still a considerable performance gap between the teacher and student model in LightHART which we believe can be reduced by adopting more advanced knowledge distillation methods. Furthermore, using the Smart-

FallMM dataset, the fall detection model trained with KD still needs to be optimized to reduce the high False Positive ratio for practical use.

Acknowledgment

We thank the National Science Foundation for funding the research under the NSF-SCH (21223749).

References

1. Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., Amirat, Y.: Physical human activity recognition using wearable sensors. Sensors **15**(12), 31314–31338 (2015)
2. Ben-Arie, J., Wang, Z., Pandit, P., Rajaram, S.: Human activity recognition using multidimensional indexing. TPAMI **24**(8), 1091–1104 (2002)
3. Chen, C., Jafari, R., Kehtarnavaz, N.: Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: ICIP. pp. 168–172. IEEE (2015)
4. Das, A., Sil, P., Singh, P.K., Bhateja, V., Sarkar, R.: Mmhar-ensemnet: a multimodal human activity recognition model. IEEE Sensors Journal **21**(10), 11569–11576 (2020)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. EK, S., Portet, F., Lalanda, P.: Lightweight transformers for human activity recognition on mobile devices. arXiv preprint arXiv:2209.11750 (2022)
7. Gao, W., Zhang, L., Teng, Q., He, J., Wu, H.: Danhar: Dual attention network for multimodal human activity recognition using wearable sensors. Applied Soft Computing **111**, 107728 (2021)
8. Han, J., Bhanu, B.: Human activity recognition in thermal infrared imagery. In: CVPR Workshops. pp. 17–17. IEEE (2005)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
10. Huan, S., Wang, Z., Wang, X., Wu, L., Yang, X., Huang, H., Dai, G.E.: A lightweighthtweight hybrid vision transformer network for radar-based human activity recognition. Scientific Reports **13**(1), 17996 (2023)
11. Ijaz, M., Diaz, R., Chen, C.: Multimodal transformer for nursing activity recognition. In: CVPR. pp. 2065–2074 (2022)
12. Islam, M.M., Nooruddin, S., Karray, F., Muhammad, G.: Multi-level feature fusion for multimodal human activity recognition in internet of healthcare things. Information Fusion **94**, 17–31 (2023)
13. Kong, Q., Wu, Z., Deng, Z., Klinkigt, M., Tong, B., Murakami, T.: Mmact: A large-scale dataset for cross modal human action understanding. In: ICCV. pp. 8658–8667 (2019)
14. Li, J., Yao, L., Li, B., Sammut, C.: Distilled mid-fusion transformer networks for multi-modal human activity recognition. arXiv preprint arXiv:2305.03810 (2023)

15. Liu, Y., Wang, K., Li, G., Lin, L.: Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition. TIP **30**, 5573–5588 (2021)
16. Ma, H., Li, W., Zhang, X., Gao, S., Lu, S.: Attnsense: Multi-level attention mechanism for multimodal human activity recognition. In: IJCAI. pp. 3109–3115 (2019)
17. Mauldin, T., Ngu, A.H., Metsis, V., Canby, M.E.: Ensemble deep learning on wearables using small datasets. ACM Trans. Comput. Healthcare **2**(1) (dec 2021). https://doi.org/10.1145/3428666, https://doi.org/10.1145/3428666
18. Memmesheimer, R., Theisen, N., Paulus, D.: Gimme signals: Discriminative signal encoding for multimodal activity recognition. In: IROS. pp. 10394–10401. IEEE (2020)
19. Mutegeki, R., Han, D.S.: A cnn-lstm approach to human activity recognition. In: ICAIIC. pp. 362–366. IEEE (2020)
20. Ngu, A.H., Metsis, V., Coyne, S., Chung, B., Pai, R., Chang, J.: Personalized fall detection system. In: 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). pp. 1–7. IEEE (2020)
21. Ngu, A.H., Yasmin, A., Mahmud, T., Mahmood, A., Sheng, Q.Z.: P-fall: Personalization pipeline for fall detection. In: Proceedings of the 8th ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies. pp. 173–174 (2023)
22. Ni, J., Ngu, A.H., Yan, Y.: Progressive cross-modal knowledge distillation for human action recognition. In: ACM MM. pp. 5903–5912 (2022)
23. Ni, J., Sarbajna, R., Liu, Y., Ngu, A.H., Yan, Y.: Cross-modal knowledge distillation for vision-to-sensor action recognition. In: ICASSP. pp. 4448–4452. IEEE (2022)
24. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Berkeley mhad: A comprehensive multimodal human action database. In: WACV. pp. 53–60. IEEE (2013)
25. Ordóñez, F.J., Roggen, D.: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1), 115 (2016)
26. Ronao, C.A., Cho, S.B.: Human activity recognition with smartphone sensors using deep learning neural networks. Expert systems with applications **59**, 235–244 (2016)
27. Saleem, G., Bajwa, U.I., Raza, R.H.: Toward human activity recognition: a survey. Neural Computing and Applications **35**(5), 4145–4182 (2023)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
29. Wang, Y., Chen, X., Cao, L., Huang, W., Sun, F., Wang, Y.: Multimodal token fusion for vision transformers. In: CVPR. pp. 12186–12195 (2022)
30. Wu, Q., Huang, Q., Li, X.: Multimodal human action recognition based on spatio-temporal action representation recognition model. Multimedia Tools and Applications **82**(11), 16409–16430 (2023)
31. Xiao, S., Wang, S., Huang, Z., Wang, Y., Jiang, H.: Two-stream transformer network for sensor-based human activity recognition. Neurocomputing **512**, 253–268 (2022)
32. Xu, H., Zhou, P., Tan, R., Li, M., Shen, G.: Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In: Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems. pp. 220–233 (2021)
33. Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., Schmid, C.: Multiview transformers for video recognition. In: CVPR. pp. 3333–3343 (2022)
34. Zhang, Y., Wang, L., Chen, H., Tian, A., Zhou, S., Guo, Y.: If-convtransformer: A framework for human activity recognition using imu fusion and convtransformer. IMWUT **6**(2), 1–26 (2022)