Progressive Cross-modal Knowledge Distillation for Human Action Recognition

Jianyuan Ni
j_n317@txstate.edu
Texas State University
San Marcos, TX, USA

Anne H.H. Ngu
angu@txstate.edu
Texas State University
San Marcos, TX, USA

Yan Yan*
yyan34@iit.edu
Illinois Institute of Technology
Chicago, IL, USA

ABSTRACT

Wearable sensor-based **H**uman **A**ction **R**ecognition (HAR) has achieved remarkable success recently. However, the accuracy performance of wearable sensor-based HAR is still far behind the ones from the visual modalities-based system (*i.e.,* RGB video, skeleton and depth). Diverse input modalities can provide complementary cues and thus improve the accuracy performance of HAR, but how to take advantage of multi-modal data on wearable sensor-based HAR has rarely been explored. Currently, wearable devices, *i.e.,* smartwatches, can only capture limited kinds of non-visual modality data. This hinders the multi-modal HAR association as it is unable to simultaneously use both visual and non-visual modality data. Another major challenge lies in how to efficiently utilize multi-modal data on wearable devices with their limited computation resources. In this work, we propose a novel **P**rogressive **S**keleton-to-sensor **K**nowledge **D**istillation (PSKD) model which utilizes only time-series data, *i.e.,* accelerometer data, from a smartwatch for solving the wearable sensor-based HAR problem. Specifically, we construct multiple teacher models using data from both teacher (human skeleton sequence) and student (time-series accelerometer data) modalities. In addition, we propose an effective progressive learning scheme to eliminate the performance gap between teacher and student models. We also designed a novel loss function called **A**daptive-**C**onfidence **S**emantic (ACS), to allow the student model to adaptively select either one of the teacher models or the ground-truth label it needs to mimic. To demonstrate the effectiveness of our proposed PSKD method, we conduct extensive experiments on Berkeley-MHAD, UTD-MHAD and MMAct datasets. The results confirm that the proposed PSKD method has competitive performance compared to the previous mono sensor-based HAR methods.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

Knowledge distillation, Progressive learning, Sensor-based human activity recognition, machine learning

ACM Reference Format:
Jianyuan Ni, Anne H.H. Ngu, and Yan Yan. 2022. Progressive Cross-modal Knowledge Distillation for Human Action Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548238

1 INTRODUCTION

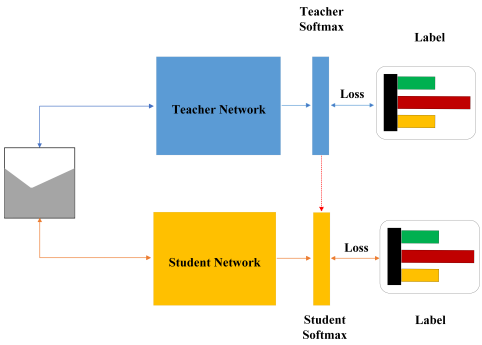
Human **A**ctivity **R**ecognition (HAR) is an active research area due to its widespread applications, for example in human-robot interaction and in smart health area [78]. There are currently two mainstreams of HAR systems: namely, visual modalities-based (*i.e.,* RGB, skeleton and depth) and non-visual modalities-based systems (*i.e.,* audio, accelerometer data, WiFi, and RFID) [70]. Among visual modalities-based systems, despite the significant achievements of video-based HAR methods had made, privacy concerns in video/image data has drawn increasing attentions recently [11, 55, 69, 94]. For instance, one of the most influential datasets in the computer vision area, ImageNet, has released an updated version that blurs people's faces for privacy protection [94]. Consequently, pure video-based approach is infeasible to be used in privacy-sensitive areas. Instead, skeleton modality can eliminate the privacy concerns while encoding the trajectories of human body joints to characterize the geometric 3D body movement patterns in a continuous way [33]. Also, skeleton modality is not susceptible to background variations and thus has attracted a lot of attentions recently [33, 93]. However, such skeleton-based systems, as well as other visual modalities-based systems, fail to be practical for real-time HAR, or anytime and anywhere HAR monitoring applications.

On the other hand, thanks to the development of Internet of Things (IoT) devices, time-series data from wearable devices has provided new opportunities in solving sensor-based HAR problem [45, 47–49, 56, 62, 80]. Currently, one of the most common sensors used in HAR problem is the accelerometer data due to its small footprint and being available on many low cost sensor devices[11]. However, the accuracy performance of sensor-based HAR system is far behind when compared to the video-based HAR system as RGB video contains richer information and can capture scene context [70]. For example, previous work demonstrated that, by using accelerometer data from a wrist-worn watch, the deep learning method for fall detection only reach 86% accuracy performance [45]. This is because the constraint of a single context from the accelerometer data lack the 3D information and can not discriminate various wrist movements when someone falls [20].

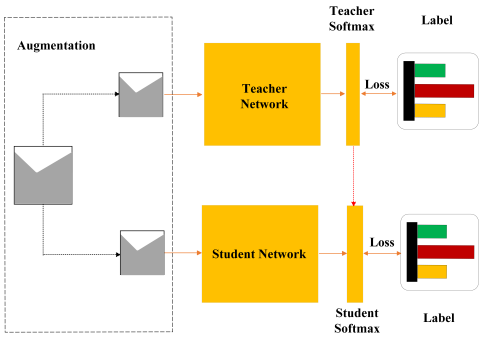
*Corresponding Author.

MM '22, October 10–14, 2022, Lisboa, Portugal

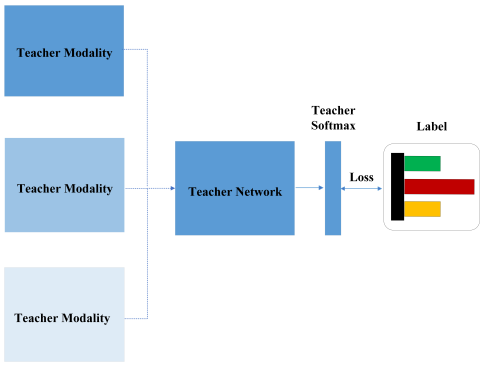
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548238



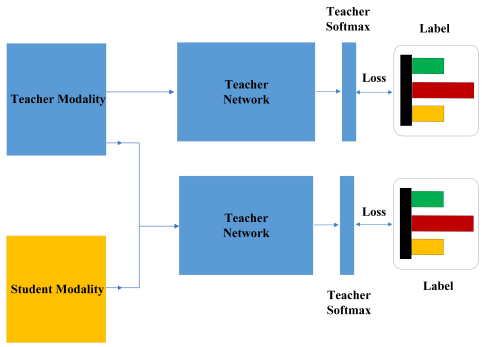
(a) Conventional two-steps knowledge distillation method with pre-trained teacher model [21, 61, 98].



(b) One-step self-distillation via augmentation [28, 92].



(c) Cross-model distillation with multi-teacher models using various modality data [42].



(d) Proposed multi-teacher method. One teacher model took teacher modality only as the input, while the other one use modality data from both teacher and student domain.

Figure 1: Comparison of various distillation methods.

In real life, humans perceive the surrounding world in a multi-modal cognition way. Similarly, multi-modal machine learning is a modeling approach trying to learn complementary features from diverse modalities of data. As a result, multi-modal approaches can often lead to more robust algorithms and better HAR performances. However, current wearable devices can only acquire certain kinds of non-visual modality data, such as accelerometer and gyroscope data [77]. This prevents the multi-modal HAR implementation on the wearable devices as it fails to use both visual and non-visual modality data simultaneously. Also, such multi-modal methods usually have complex architectures and incur high computational overheads which wearable devices can not afford. For example, previous study indicated that a smartphone (LG Nexus 5X, 1.8 GHz, Hexa-core processor with 2G of RAM) can only support a long short-term memory (LSTM) model which contains an input layer, two hidden layers, and an output layer [44].

How to leverage the advantages of advanced multi-modal methods for the wearable sensor-based HAR problem? The technique of cross-modal transfer, *i.e.,* knowledge distillation, can be one of the potential solutions. Knowledge distillation (KD) was formally popularized by distilling knowledge from a larger model (*i.e.,* teacher) into a smaller model (*i.e.,* student) as a two-steps process as shown in Figure 1a. By mimicking the pre-trained teacher model, the student model is able to retain similar accuracy as well as reducing the computation resource demand [21]. After that, the data augmentation based self-knowledge distillation methods [28, 92, 97], which adopted a consistent prediction of relevant data from the same class, *i.e.,* distorted versions of instances, has proposed to improve the performance of the student model as shown in Figure 1b. For example, Zhang *et al.* proposed a one-step self-distillation method, in which knowledge from the deeper parts of the network is distilled into its shallow sections [101]. Currently, there are only a few multi-modal KD approaches for the HAR problem [30, 42, 52, 72]. For example, Liu *et al.* [42] introduced a multi-modal KD method which integrated various sensor information to improve the vision modality as shown in Figure 1c. Instead, Ni *et al.* [52] proposed a multi-modal KD approach where the complementary information from the video domain was adaptively transferred to the sensor domain. Although those studies provide promising results on multi-modal HAR problem, there are two questions that those works have not addressed: 1) they have used a pre-trained teacher model to directly guide the student network, which is an inefficient learning process and thus contributed to the student model's performance degradation; 2) they failed to consider the fact that the student modality can also contribute to the whole KD process. We argued

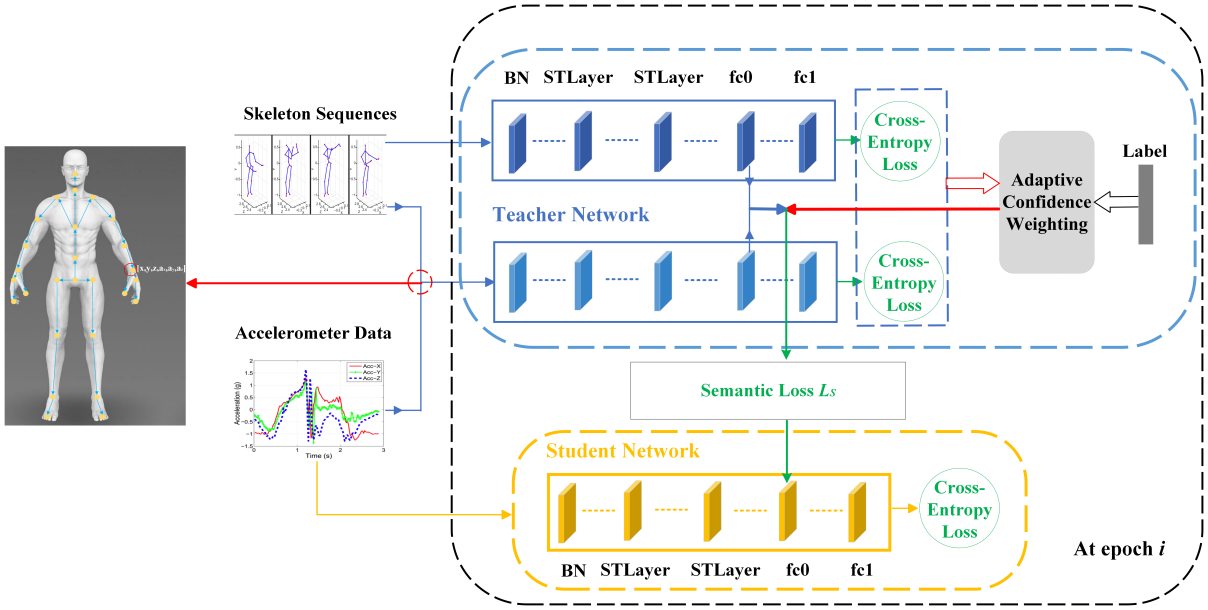


Figure 2: Schematic overview of the proposed PSKD method. Initially, multi-teacher models are constructed using both teacher and student modalities. Next, we propose a progressive learning scheme to eliminate the performance gap between teacher and student models. We also introduce a loss function to allow the student model adaptively decide either one of the teacher models or the ground-truth label it needs to mimic.

that there are two key factors in the KD process: knowledge source (*i.e.*, *teacher model*) and the distillation process. In order to increase the accuracy performance of a student model, a teacher model with higher performance (*i.e.*, *stronger teacher*) should be achieved at first. In addition, how to reduce the performance gap between teacher and student model via novel training strategy is another important step to produce a strong student model.

Driven by the aforementioned two intuitions, we proposed an end-to-end Progressive Skeleton-to-Sensor Knowledge Distillation (PSKD) for HAR recognition in this study. The overview of the proposed method is shown in Figure 2. First, we propose a new multi-teacher approach to construct multiple teacher models using skeleton (teacher) and accelerometer (student) data modalities as shown in Figure 1d. In this way, the teacher models can also understand the characteristic of the student modality data so that teacher models can generate models which are easier for student model to mimic. Next, we design an effective progressive learning (PL) scheme to eliminate the performance gap between teacher and student models. Specially, the student model will be updated after the multi-teacher models are updated every epoch to converge to the ground-truth labels. During the PL training process, a novel loss function called Adaptive-Confidence Semantic (ACS), is introduced to allow the student model adaptively decide which teacher models or the ground-truth label it needs to mimic. In summary, the contributions of this paper are summarized as follows: 1) To the best of our knowledge, this is the first study conducting the cross-modal KD model from the skeleton data domain to the wearable sensor data domain. In this PSKD model, a student model with input of accelerometer data, learns the compensatory information from multi-teacher models with both input of skeleton sequences and wearable sensor data. 2) We designed an effective PL scheme coupled with a novel loss function (ACS), which is utilized to alleviate the modality gap between the teacher and the student model. 3) We demonstrated the competitiveness performance of the proposed PSKD method on three public datasets over the previous sensor-based HAR methods.

2 RELATED WORK

We briefly review the existing studies for the skeleton and wearable sensor-based HAR problem. Multi-modal HAR and knowledge distillation work are also included in this section.

2.1 Skeleton-based HAR

Skeleton sequences encode the trajectories of human body joints, which characterize temporal contextual informative human motions over time. There are several advantages of using skeleton sequences for HAR problem, due to its informative representation and its robustness against variations of backgrounds [70]. Early skeleton-based HAR works mainly focused on extracting hand-crafted spatial and temporal features, which can be divided into joint-based [79] and body part-based [76] methods. Besides, due to the strong feature learning capability, skeleton-based deep learning methods have become one of the mainstream research in this field. Recurrent neural network (RNN) or their variants (*e.g.,* long short-term memory (LSTM)) are capable of learning the dynamic dependencies in sequential data [12, 38, 102, 103]. For example, Du *et al.* introduced an end-to-end RNN, which divided skeleton data into five body parts rather than taking the skeleton from each frame as a

whole. These five body parts were then fed to several bidirectional RNNs to generate high-level representations of the action [12]. Liu *et al.* proposed the tree structure-based skeleton traversal method to exploit the spatial information of the skeleton sequences [37]. Liu *et al.* presented an attention-based LSTM network to encode the skeleton sequence and refine the global context on the informative joints [39]. Similarly, a deep LSTM network consisted of the jointly spatial and temporal attention subnetwork was proposed to model the temporal dynamics and spatial configurations [67].

At the same time, due to the expressive power of graph structures, Graph convolutional Networks (GCNs) has been introduced to the skeleton-based HAR problem [51, 93, 104]. For example, Yan *et al.* exploited GCNs for skeleton-based HAR by introducing Spatial-Temporal GCNs (ST-GCNs) that can automatically learn both the spatial and temporal patterns for skeleton-based HAR problem [93]. Wu *et al.* adopted the ST-GCNs to learn the global information at first and then designed a residual layer to capture the spatio-temporal information of skeleton sequences [87]. Li *et al.* proposed a spatial and temporal graph router to produce new skeleton-joint-connectivity graphs. After that, this skeleton-joint-connectivity graphs was fed to the ST-GCNs for further classification [33]. Moreover, Shi *et al.* proposed a two-stream Adaptive GCN (2s-AGCN), which coupled the first-order information (coordinates of joints) with the second-order information (lengths and directions of human bones) on the skeleton-based HAR [63]. Li *et al.* proposed the Symbiotic Graph Neural Networks (Sym-GCNs) to handle both action recognition and motion prediction tasks simultaneously so that these two tasks can enhance each other [35]. In order to reduce computational costs of GCNs, Cheng *et al.* present a ShiftGCNs which applied lightweight point-wise convolutions and shift graph operations [7]. Similarly, Song *et al.* proposed a multi-stream GCNs model which fuses joint positions, motion velocities as well as bone features. Separable convolutional layers and compound scaling strategy was applied in this study to reduce the redundant training parameters [68]. In summary, the skeleton modality provides the body structure information, which is effective for HAR problem. Nevertheless, skeleton-based approaches, as well as other visual modalities-based approaches, can only be applied in a static environment where visual devices can be permanently installed. For example, visual modalities-based approaches are not suitable for outdoor HAR monitoring problem.

2.2 Wearable sensor-based HAR

Wearable sensor-based HAR methods has received huge attentions due to their robustness against occlusion and viewpoint variations [70]. Wearable sensors only includes subtle intra-class variations for the same action performance, regardless of the size of human body which varies from person to person. Therefore, wearable devices has been adopted for remote monitoring systems without worry about the privacy-safety concerns [27, 46]. Numerous Convolutional neural network (CNN) on wearable-based HAR [6, 32, 99] has been proposed. For instance, a wrist worn tri-axial accelerometer was used to perform arm movement prediction and results demonstrated the robustness of such approach [56]. RNN type of model was also suggested to deal with time-dependent input sequences [44, 45, 54, 89]. Wang *et al.* [80] integrated a CNN and bidirectional LSTM model to acquire spatial and temporal features from acceleration data. Zhao *et al.* proposed the residual bi-directional LSTM model to concatenate the forward and backward state *i.e.,* positive and negative time direction to avoids the gradient vanishing problem [106]. Wang and Liu [82] present a novel Hierarchical LSTM method to improve the system's performance. Meanwhile, some approaches also suggested converting wearable sensor sequences as images for HAR study. Zeng *et al.* [99] transformed the single-axis sensor data into one-dimensional images and then fed them to CNN for identification. Lu *et al.* [43] encoded the tri-axial acceleration data into color images, which were fed into a ResNet for HAR. However, the accuracy performance of sensor-based HAR is still far behind compared to the visual modalities-based HAR results due to the constraint of a single contextual information from accelerometer data [20]. In reality, we human, understand the surrounding environment in a multi-modal way. Hence, by utilizing the complementary information acquired from different modalities, it is possible to enrich the gained knowledge and thus enhance the sensor-based HAR performance eventually.

2.3 Multi-modal based HAR

Recently, deep learning methods have been conducted on HAR problem [1, 2, 9, 10, 59, 86]. For example, Dawar *et al.* proposed the data augmentation CNN and CNN+LSTM methods based on depth and inertial modalities, respectively. Wei *et al.* fed the 3D videos frames as well as 2D inertial images to a 3D CNN and a 2D CNN models, respectively. The score fusion strategy outperformed the feature fusion method in this study [86]. Similarly, some studies have also been conducted on the depth-inertial fusion techniques by combining two-stream CNN architectures [1, 2]. Besides that, Islam and Iqbal [24] also proposed a separate encoder to fuse RGB,skeleton and inertial modalities in a similar shaped vector representation way. Li *et al.* adopted the ST-GCN model [93] to extract the skeleton feature vector from videos and the R(2+1)D [34] model to encode the RGB videos directly. While the aforementioned multi-modal models tend to achieve better performance, one of the drawback is the high computational overhead and larger memory demand. Consequently, efficient model compression methods have emerged to build deep models with less computational resource and maintain the similar performance [19].

Knowledge distillation (KD) is one of the model compression methods which transfer knowledge from a computational expensive model into a smaller network [21]. In general, student model tried to mimic the performance from a pre-trained teacher model as shown in Figure 1a. Zagoruyko and Komodakis [98] proposed the attention information transfer method by forcing a student CNN model to mimic the attention maps from a teacher network. Park *et al.* designed the distance-wise and angle-wise distillation loss for the relational knowledge transfer in the KD process [57]. Tung *et al.* proposed a new form of KD distillation loss with the constraint that input pairs that produce similar activations in the teacher network should also produce similar activations in the student network [73]. Different from these KD methods that mainly focus on the distillation loss task, there are only a few multi-modal KD approaches for the HAR problem [16, 22, 30, 42, 52, 72]. For example, Hoffman *et al.* designed a modality hallucination architecture by using depth

as side information to guide an RGB object detection model [22]. Garcia *et al.* built a KD framework to learn representations from the depth and RGB videos, while only use RGB data at test time. Similarly, Thoker *et al.* proposed a multi-modal KD framework which used RGB videos to train the teacher network for HAR task. After that, two student networks were trained using mutual learning to improve the performance [72]. In addition, Ni *et al.* present the first multi-modal KD approach on the sensor-based HAR problem. In this study, the complementary information from the video domain was adaptively transferred to the sensor domain and improve the accuracy performance of sensor-based HAR problem [52]. However, previous works either ignored the fact that the student modality can contribute useful information to the training of the KD process or it is not efficient to let the small student network learn directly from a pre-trained teacher model. Our work, instead, utilized both teacher and student modality data to build up multiple teacher models. In this way, teacher models will tend to produce a model which is easier for the student model to understand from the human's learning analogy perspective.

3 METHODS

This section describes our proposed approach in terms of the multi-teacher models construction process, the progressive learning KD procedure, and the designed loss function used to let the student model adaptively learn from either one of these multiple teacher models or the ground-truth label directly.

3.1 Multi-teacher Construction Process

While impressive progress has been achieved under the standard teacher-student KD paradigm, the intuition that a student can learn more effectively from multiple teachers has only been investigated recently. Currently, there are several studies using multiple teacher models in the KD process [17, 40, 42, 50, 71, 81, 88, 90, 95]. However, these works ignore the following two advantages of using the teacher model that is familiar with the student's modality data: 1) when the teacher model realizes and understands the intra-modality characteristics between teacher and student domain modality data, it can generate a model which alleviate some level of difficulties when the student model tries to mimic its performance; 2) additional modality data, from both teacher and student domain, can build up teacher models with better performance.

Based on the above two perceived advantages, we constructed our multi-teacher models which use input modality from both teacher (*i.e.,* skeleton sequence) and student domain (*i.e.,* accelerometer data). Inspired by [14], we use the well established Graph Convolutional Networks (GCNs) models, Adaptive GCNs [63], to utilize the skeleton sequences as the input for the first teacher model $Teacher_{sk}$. After that, we concatenated the skeleton sequence and accelerometer data together to build another teacher model $Teacher_{fu}$ as shown in Figure 3. We briefly introduce the fusion process here: let $\mathbf{X}_{SK} \in \mathbb{R}^{(M,C_{SK},T_{SK},N_{SK})}$ be a skeleton sequence input, where M is the number of participants that are involved in an action, C_{SK} is the initial 2D joint coordinates and size T_{SK} and N_{SK} are the sequence length and number of skeleton graph nodes. For accelerometer data, the input is defined as $\mathbf{X}_{AC} \in \mathbb{R}^{(M,C_{AC},S_{AC},T_{AC})}$, where T_{AC} is the accelerometer sequence length, S_{AC} is the number of sensors and C_{AC} is the channel dimension of the accelerometer data. For example, given the accelerometer data from a smartwatch with x-, y- and z-values, the structure would be $S_{AC} = 1$ and $C_{AC} = 3$. Similar to the skeleton data, M denotes the person wearing the sensor device and $C_{AC} = C_{SK} = 3$ since skeleton sequences and accelerometer data all include x-, y- and z-axis values. As a result, if there is only one participant wearing a smartwatch during the activity performance process, the number of M_{SK} should be equivalent to that of M_{AC}. Also, a common T can be guaranteed by resampling T_{SK} and T_{AC} to the same time length. After that, the fused data which formed as $\mathbf{X} \in \mathbb{R}^{(M,C_{SK}+C_{AC},T,N_{SK})}$, can be fed into the AGCN backbone as shown in Figure 2.

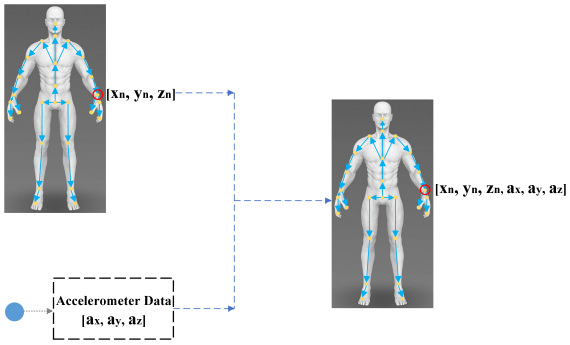


Figure 3: The fusion of skeleton sequence $[x_n, y_n, z_n]$ and accelerometer data $[a_x, a_y, a_z]$.

3.2 Progressive Learning KD Procedure

Hilton *et al.* [21] proposed the standard KD process which referred to a model-agnostic method where a small or less complex model (i.e., *student*) tried to minimize the statistical discrepancy between its predictions distributions and the predictions of a complicated model (i.e., *teacher*). However, the standard KD process cannot fully solve the performance gap between various modalities [42, 52]. We face the problem that the distribution of teacher and student modality could be very far from each other at the beginning, leading to a difficult distillation process. It has been proved that a more powerful teacher model is not a guarantee to give rise to a better student model during the KD process [60]. Also, it is hard to let the student model simply mimic a pre-trained teacher model when the capacity gap between the teacher and student is large [25, 50].

In order to solve this problem, an intermediate network has been introduced recently [50, 60, 64]. For instance, Mirzadeh *et al.* proposed the Teacher-Assistant KD method by gradually increasing the teacher size to foster the distillation process [50]. This is mainly due to the fact that the learning process can be evaluated if a suitable goal of the teacher model is set for the student model to follow. However, training intermediate networks will incur more computational cost and training time [60, 64]. Therefore, different to the original KD process where the student learns directly from the pre-trained teacher model, inspired by [64], we introduced a progressive KD learning procedure that requires the student model to learn from the teacher model step-by-step or incremental fashion

which can help the student model better mimic the performance of the teacher model. Specially, we update the student model immediately after the teacher model updates one step towards the ground-truth labels.

However, with multiple teacher models in this study it can be confusing for the student model to mimic their performances when their prediction results are inconsistent or incorrect as shown in Figure 4. For example, when both teacher models ($Teacher_{sk}$ and $Teacher_{fu}$) predict incorrectly or when only one of the teacher models ($Teacher_{sk}$ or $Teacher_{fu}$) achieve the correct prediction result, it is unrealistic for the student model to imitate their performances directly. Technically, the student model should have the knowledge to select which teacher model or the ground-truth label it needs to follow adaptively.

3.3 Adaptive-Confidence Loss Function

Currently, several multi-teacher model studies are proposed and the results demonstrated the beneficial effect of multi-teacher models on the KD process [15, 31, 88, 96]. For example, either fixed weight assignment [15, 88, 96] or other label-free schemes, such as entropy-based weight optimization method [31], has been used for the student model to learn from the multi-teacher models. However, fixed weight assignment failed to balance the importance of multi-teacher models and the other methods may misguide the student model in the presence of low-quality teacher predictions. More recently, Zhang *et al.* proposed the confidence-aware loss function which adaptively assign the sample-wise reliability for each teacher prediction based on the ground-truth labels [100]. However, they failed to realize the case where all teacher models predicted wrong. Based on these observations, we proposed a novel Adaptive-Confidence loss (AC) to let the student model adaptively emulates the best teacher model performance or the ground-truth label during the KD process. In this study, there are two teacher models, leading to four different cases we need to consider to design the AC loss intuitively: 1) when both teacher models, $Teacher_{sk}$ and $Teacher_{fu}$, all have correct prediction results, the student model should mimic the teacher model which achieved the higher prediction score; 2) when only one of the teacher models predict correctly, the student model shall mimic the one which predict correctly; 3) when both teacher models predict incorrectly, the student model will have to switch and simulate the ground-truth label instead. An illustration example of these fours cases is shown in Figure 4.

Given a teacher model T_k and a student model S_k, the soft-target $\tilde{y}^T$ produced by the teacher model is considered as high-level knowledge. The loss of KD when training a student model can be defined as:

$$\mathcal{L}_{KD} = \mathcal{L}_C(y, y^S) + \alpha \mathcal{L}_K(\tilde{y}^T, \tilde{y}^S) \tag{1}$$

$$\mathcal{L}_K = \frac{1}{m}\sum_{k=0}^{m} KL(\frac{P^{T_k}}{T}, \frac{P^{S_k}}{T}) \tag{2}$$

where y and y^S refer to the predicted labels and class probability for the student network in this study, respectively. $\tilde{y}^S$ is the soft target generated by the student model. Here $\mathcal{L}_C$ is the typical cross-entropy loss and $\mathcal{L}_K$ is the Kullback-Leibler (KL) divergence, while P^{T_k} is the class probability for the teacher network and P^{S_k} is the class probability for the student network. T represents the temperature controlling the distribution of the provability and we use $T = 4$ in this study according to [21].

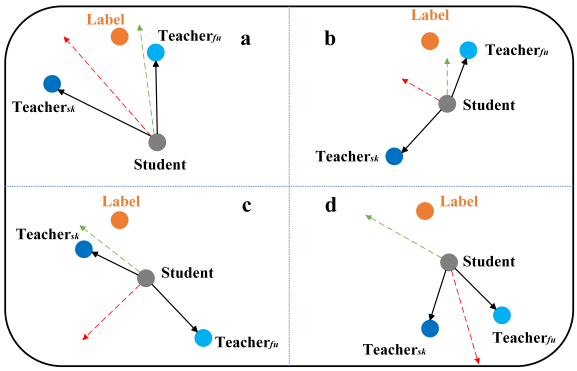


Figure 4: Comparison of the previous average weight result (red dash line) and our proposed Adaptive-Confidence weight method (green dash line).

Intuitively, we assign different weights on each teacher model by calculating the cross entropy loss between teacher predictions and the ground-truth labels:

$$\mathcal{L}^K_{CE_{KD}} = -\sum_{c=1}^{C} y^c log(\alpha(Z^c_{T_K})), \tag{3}$$

$$\omega^k_{KD} = \begin{cases} -\frac{1}{K-1}\left(1 - \frac{exp(L^k_{CE_{KD}})}{\sum_j exp(L^j_{CE_{KD}})}\right), & \mathcal{L}^K_{CE_{KD}} \leq \mathcal{L}_{KD} \\ 0, & \mathcal{L}^K_{CE_{KD}} \geq \mathcal{L}_{KD} \end{cases} \tag{4}$$

where T_K denotes the kth teacher and $\alpha(z^c)$ is the softmax function. Consequently, the overall teacher predictions are then aggregated with calculated weights:

$$\mathcal{L}_{MK} = -\sum_{k=1}^{K} \omega^k_{KD} \sum_{c=1}^{C} Z^c_{T_K} log(\alpha(Z^c_S)), \tag{5}$$

Therefore, the teacher model whose prediction is closer to the ground-truth labels will be assigned larger weight ω^k_{KD}. In addition, when all of the multi-teacher model predict incorrectly, the student model will try to follow the ground-truth label instead.

Since multi-modal action data share the same semantic content [42, 52], semantic loss is defined as:

$$\mathcal{L}_S = \frac{1}{m}\sum_{k=1}^{m}(\|H^S - H^T\|)^2_2 \tag{6}$$

where H^S and H^T represent the feature of fc0 layer from both student and teacher models. To keep H^S and H^T spatial dimensions same, we add one more fc layer (fc0) before its original fc layer (fc1) shown in Figure 2.

In summary, we use the original KD loss L_{KD} and augment it to include Adaptive-Confidence loss L_{MK} as well as the semantic loss L_S, to train the student network and the final ACS loss for the student model is defined as follow:

$$\mathcal{L} = L_{KD} + \beta L_{MK} + \gamma L_S \tag{7}$$

where α, β, γ are the tunable hyperparameters to balance the loss terms for the student network.

4 EXPERIMENTS

4.1 Dataset

In this study, three benchmark datasets were selected due to their multi-modal data forms:

Berkeley-MHAD [53]. This dataset includes 11 action classes performed by 12 participants (5 females and 7 males). There are 6 three-axis wireless accelerometers installed to measure movement at the wrists, ankles and hips. In summary, there are 3,948 accelerator samples in total. We use skeleton motion data and accelerometer data as the teacher modality and accelerator data as the student modality. In this study, we use the first 7 participants for training and the rest ones for testing mentioned in [53].

UTD-MHAD [5]. This dataset covers 27 action classes performed by 8 participants (4 females and 4 males). In this study, we use skeleton sequences and inertial data (accelerometer data) as the teacher modality and accelerometer data as the student modality. Both modalities have 861 samples and we spit them in half for training and testing mentioned in [5].

MMAct [30]. This dataset includes 37 action classes performed by 20 participants (10 females and 10 males) containing more than 36,000 trimmed clips. Since the skeleton sequences are missing in this dataset, we use OpenPose to extract them from RGB videos [4]. After that, we use skeleton sequences plus accelerometer data from watch as the teacher modality and accelerometer data from watch as the student modality. One of the various settings (cross-subject) is used to evaluate this dataset based on the train and test split strategy mentioned in [30].

4.2 Experimental Settings

All the experiments were performed on four Nvidia GeForce GTX 1080 Ti GPUs using PyTorch. To guarantee a deterministic and reproducible behavior, all training procedures are initialized with a fixed random seed. We employed the classification accuracy and F-measure as the evaluation metric to compare the performance of the PSKD model. Grid-search method [42] was conducted to evaluate the effect of hyper-parameters α, β, γ in three datasets.

4.3 Comparison to the State-of-the-Art

We compare the performance of our PSKD with state-of-the-art vision-based action recognition (VAR), multi-modal action recognition methods (MMAR), skeleton-based action recognition (SKAR), sensor-based action recognition (SAR), and knowledge distillation (KD) methods. The comparison results of three datasets are shown in Table 1, 2, and 3, respectively. In Table 1, the proposed PSKD model performs better than all the previous comparable VAR models when the RGB videos used as the input data by 0.79%-9.59% [23, 41, 65, 85, 105]. We make an improvement in the testing accuracy of 4.99% compared to the study where 16 features from accelerometer signals were captured for classification [18]. Similarly, the proposed PSKD model achieved higher accuracy performances compared to the previous MMAR and SKAR models [24, 91]. Especially, the proposed PSKD model outperforms all previous SAR methods using both accelerometer and gyroscope data as the input,

Table 1: Comparison results between our proposed method and state-of-the-art methods on UTD-MHAD dataset in accuracy performance (%). Acc. denotes accelerometer and Gyro. denotes gyroscope.

Type	Method	Testing Modality	Accuracy (%)
VAR	Hussein *et al.* [23]	RGB video	85.60
	Wang *et al.* [85]	RGB Video	85.81
	Zhao *et al.* [105]	RGB Video	92.10
	Si *et al.* [65]	RGB Video	94.40
	Liu *et al.* [41]	RGB Video	92.84
SKAR	Xiao *et al.* [91]	Skeleton	94.37
MMAR	Islam *et al.* [24]	Skeleton+RGB video	95.12
SAR	Singh *et al.* [66]	Acc. + Gyro.	91.40
	Ahmad and Khan [2]	Acc. + Gyro.	95.80
	Wei *et al.* [86]	Acc. + Gyro.	90.30
	Garcia-Ceja *et al.* [18]	Acc.	90.20
KD	Ni *et al.* [52]	Acc.	**96.97**
Proposed	PSKD	Acc.	95.19

Table 2: Comparison results between our proposed method and state-of-the-art methods on Berkeley-MHAD dataset in accuracy performance (%). Acc. denotes accelerometer.

Type	Method	Testing Modality	Accuracy (%)
VAR	Wang *et al.* [84]	RGB Video	88.19
	Zhou *et al.* [107]	RGB Video	95.32
	Lin *et al.* [36]	RGB Video	96.87
SKAR	Vantigodi *et al.* [74]	Skeleton	96.06
	Vantigodi *et al.* [75]	Skeleton	97.58
	Kapsouras *et al.* [26]	Skeleton	**98.18**
SAR	Das *et al.* [8]	Acc.(Six locations)	88.90
KD	Ni *et al.* [52]	Acc. (Left Wrist)	90.02
Proposed	PSKD	Acc. (Left Wrist)	94.76

Table 3: Comparison results between our proposed method and state-of-the-art methods on MMAct dataset in F1 socre performance (%). Acc. denotes accelerometer and Gyro. denotes gyroscope.

Type	Method	Testing Modality	Cross Subject(%)
VAR	Kong *et al.* [29]	RGB video	62.80
	Wang *et al.* [85]	RGB video	64.40
	Zhou *et al.* [107]	RGB video	66.56
	Lin *et al.* [36]	RGB video	70.12
	Kong et al. [29]	RGB video	59.10
SAR	Kong *et al.* [30]	Acc.(watch+phone)	62.67
KD	Ni *et al.* [52]	Acc. (watch)	60.14
Proposed	PSKD	Acc. (watch)	**71.42**

which validate the effectiveness of the PSKD model. However, it is worth noting that the proposed PSKD model does not perform better as compared to the previous method where the accelerometer data learn the complementary information from video domain during the KD process [52]. This degradation was mainly due to the noisy data in the skeleton domain from the UTD-MHAD dataset[13]. In general, these results demonstrated that accelerometer data in the

PSKD model can achieve competitive accuracy performances. In Table 2, even though the proposed PSKD model can not outperform the previous SKAR and VAR models where the skeleton data or RGB video was used during the testing modality [26, 36, 75], our proposed PSKD method tested with only the left wrist accelerometer data does perform better compared to the previous study where accelerometer data from six locations were used [8, 52], regardless of any data prepossessing they applied. This result sheds light on the proposed PSKD for improving sensor-based HAR. In Table 3, while accelerometer data from the watch is the only modality in the testing phase, the method achieves better F-score performance compared to [29, 30, 36, 83, 107] in which either video streams or accelerometer data from phone and watch were used in the testing phase. This result validates that accelerometer data in the PSKD model can significantly learn knowledge from skeleton data and thus effectively improve sensor-based HAR performance.

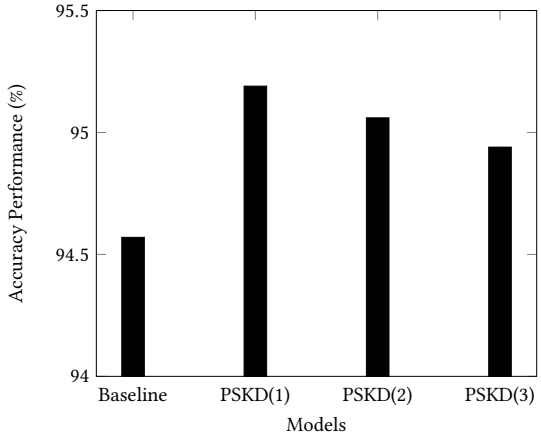


Figure 5: Accuracy performances (%) between the baseline and the proposed PSKD method on UTD-MHAD dataset [5]. The number in parenthesis means the epoch numbers which the student tries to mimic either the best teacher model or the ground-label truth iteratively.

4.4 Ablation Study

In this subsection, we design experiments to verify the effectiveness of each component in the proposed framework based on UTD-MHAD dataset [5] and try to answer the following questions:

(i) What is the effectiveness of multi-teachers progressive learning scheme in PSKD?

To evaluate the effectiveness of the proposed multi-teachers progressive learning (PL) scheme in PSKD method, we compare the PSKD with the student baseline: 1) a student baseline model which learns directly from two pre-trained teacher model ($Teacher_{sk}$ and $Teacher_{fu}$) ; 2) PSKD model with different epoch numbers which the student tries to mimic. As shown in Figure 5, the proposed PSKD model outperforms the student's baseline model by 0.37%-0.62%, proving that the PL scheme is able to reduce the performance gap and thus improve the accuracy performance of the student model. In addition, PSKD(1) achieved higher accuracy performance compared to PSKD(2) and PSKD (3), indicating that a smaller learning step can boost the learning capacity of the student model.

Table 4: Ablation study of accuracy (%) and F1 score (%) performance on UTD-MHAD dataset. Acc. denotes accelerometer and W/O denotes without. AC denotes the Adaptive-Confidence loss L_{MK}. S denotes semantic distillation loss L_S.

Method	Testing Modality	Accuracy	F1 score
Logits [3]	Acc.	93.87	94.15
Fitnet [61]	Acc.	94.03	94.27
ST [21]	Acc.	94.34	94.64
AT [98]	Acc.	94.27	94.80
RKD [57]	Acc.	95.03	95.02
SP [73]	Acc.	94.06	94.56
CC [58]	Acc.	94.12	94.72
ACS	Acc.	**95.19**	**95.67**
AC(W/O S)	Acc.	94.51	94.59
S (W/O AC)	Acc.	94.82	94.21

(ii) What are the contributions of each loss terms in the proposed ACS loss function?

To evaluate the effectiveness of the proposed loss function, we compare the ACS function with state-of-the-art KD methods [3, 21, 57, 58, 61, 73, 98]. For those methods, we use the shared codes, and the parameters are selected according to the default setting. As shown in Table 4, the proposed ACS loss function performs better than all of the comparable KD loss functions. These observations validates that our ACS can effectively transfer the knowledge from skeleton modalities to wearable sensor modalities by integrating two complementary modules, L_{MK} and L_S. In addition, the Adaptive-Confidence loss L_{MK} contributes about 0.31% to accuracy improvement as compared to semantic distillation loss L_S, which validates the assumption that distillation process is a key factor in the KD process. Also, semantic distillation loss L_S contributes 0.17% to accuracy improvements, proving that the semantic information is critical for time-series data in a KD process [52].

5 CONCLUSION

In this work, we propose a novel Progressive Skeleton-to-sensor Knowledge Distillation (PSKD) model which only needs to accept time-series data *i.e.,* accelerometer data, from a smartwatch during the testing phase. Specifically, we propose the construction of multiple teacher models using both teacher and student modalities. In addition, we design an effective progressive learning scheme to eliminate the performance gap between the teacher and the student models. After that, a novel loss function called Adaptive-Confidence Semantic (ACS), is introduced to allow the student model adaptively to select the correct teacher model or the ground-truth label it needs to mimic. Extensive experimental results on UTD-MHAD, MMAct and Berkeley-MHAD datasets confirm the effectiveness and competitive performance compared to the previous methods on the mono sensor-based HAR problem.

ACKNOWLEDGMENTS

This research is supported by NSF SCH-2123749 and SCH-2123521 Collaborative Research. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

REFERENCES

[1] Zeeshan Ahmad and Naimul Khan. 2020. CNN-based multistage gated average fusion (MGAF) for human action recognition using depth and inertial sensors. *IEEE Sensors Journal* 21, 3 (2020), 3623–3634.

[2] Zeeshan Ahmad and Naimul Mefraz Khan. 2019. Multidomain multimodal fusion for human action recognition using inertial sensors. In *BigMM*. 429–434.

[3] Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? *Advances in neural information processing systems* 27 (2014).

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*. 7291–7299.

[5] 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 *ICIP*. 168–172.

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

[7] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. 2020. Skeleton-based action recognition with shift graph convolutional network. In *CVPR*. 183–192.

[8] Avigyan Das, Pritam Sil, Pawan Kumar Singh, Vikrant Bhateja, and Ram Sarkar. 2020. MMHAR-EnsemNet: A multi-modal human activity recognition model. *IEEE Sensors Journal* 21, 10 (2020), 11569–11576.

[9] Neha Dawar and Nasser Kehtarnavaz. 2018. A convolutional neural network-based sensor fusion system for monitoring transition movements in healthcare applications. In *ICCA*. 482–485.

[10] Neha Dawar, Sarah Ostadabbas, and Nasser Kehtarnavaz. 2018. Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition. *IEEE Sensors Letters* 3, 1 (2018), 1–4.

[11] Florenc Demrozi, Graziano Pravadelli, Azra Bihorac, and Parisa Rashidi. 2020. Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey. *IEEE Access* 8 (2020), 210816–210836.

[12] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*. 1110–1118.

[13] Michael Duhme. 2021. *Multimodal Action Recognition using Graph Convolutional Neural Networks*. Master's thesis. University of Koblenz-Landau, Active Vision Group.

[14] Michael Duhme, Raphael Memmesheimer, and Dietrich Paulus. 2021. Fusion-GCN: Multimodal Action Recognition using Graph Convolutional Networks. In *DAGM German Conference on Pattern Recognition*. Springer, 265–281.

[15] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. 2017. Efficient Knowledge Distillation from an Ensemble of Teachers.. In *Interspeech*. 3697–3701.

[16] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. 2020. Listen to look: Action recognition by previewing audio. In *CVPR*. 10457–10467.

[17] Nuno C Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. 2019. Dmcl: Distillation multiple choice learning for multimodal action recognition. *arXiv preprint arXiv:1912.10982* (2019).

[18] Enrique Garcia-Ceja, Carlos E Galván-Tejada, and Ramon Brena. 2018. Multi-view stacking for activity recognition with sound and accelerometer data. *Information Fusion* 40 (2018), 45–56.

[19] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.

[20] Haodong Guo, Ling Chen, Liangying Peng, and Gencai Chen. 2016. Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble. In *UbiComp*. 1112–1123.

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[22] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. 2016. Learning with side information through modality hallucination. In *CVPR*. 826–834.

[23] Mohamed E Hussein, Marwan Torki, Mohammad A Gowayyed, and Motaz El-Saban. 2013. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *IJCAI*.

[24] Md Mofijul Islam and Tariq Iqbal. 2020. Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. In *IROS*. 10285–10292.

[25] Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. 2019. Knowledge distillation via route constrained optimization. In *ICCV*. 1345–1354.

[26] Ioannis Kapsouras and Nikos Nikolaidis. 2014. Action recognition on motion capture data using a dynemes and forward differences representation. *Journal of Visual Communication and Image Representation* 25, 6 (2014), 1432–1445.

[27] Adil Mehmood Khan, Y-K Lee, Seok-Yong Lee, and T-S Kim. 2010. Human activity recognition via an accelerometer-enabled-smartphone using kernel discriminant analysis. In *2010 5th international conference on future information technology*. 1–6.

[28] Jangho Kim, SeongUk Park, and Nojun Kwak. 2018. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems* 31 (2018).

[29] Quan Kong, Wenpeng Wei, Ziwei Deng, Tomoaki Yoshinaga, and Tomokazu Murakami. 2020. Cycle-contrast for self-supervised video representation learning. *arXiv preprint arXiv:2010.14810* (2020).

[30] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. 2019. Mmact: A large-scale dataset for cross modal human action understanding. In *ICCV*. 8658–8667.

[31] Kisoo Kwon, Hwidong Na, Hoshik Lee, and Nam Soo Kim. 2020. Adaptive knowledge distillation based on entropy. In *ICASSP*. 7409–7413.

[32] Song-Mi Lee, Sang Min Yoon, and Heeryon Cho. 2017. Human activity recognition from accelerometer data using Convolutional Neural Network. In *BigComp*. 131–134.

[33] Bin Li, Xi Li, Zhongfei Zhang, and Fei Wu. 2019. Spatio-temporal graph routing for skeleton-based action recognition. In *AAAI*, Vol. 33. 8561–8568.

[34] Jianan Li, Xuemei Xie, Qingzhe Pan, Yuhan Cao, Zhifu Zhao, and Guangming Shi. 2020. SGM-Net: Skeleton-guided multimodal network for action recognition. *Pattern Recognition* 104 (2020), 107356.

[35] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2021. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *TPAMI* (2021).

[36] Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *ICCV*. 7083–7093.

[37] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. 2016. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*. Springer, 816–833.

[38] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. 2017. Skeleton-based human action recognition with global context-aware attention LSTM networks. *IEEE Transactions on Image Processing* 27, 4 (2017), 1586–1599.

[39] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. 2017. Global context-aware attention lstm networks for 3d action recognition. In *CVPR*. 1647–1656.

[40] Linqing Liu, Huan Wang, Jimmy Lin, Richard Socher, and Caiming Xiong. 2019. Attentive student meets multi-task teacher: Improved knowledge distillation for pretrained models. (2019).

[41] Mengyuan Liu and Junsong Yuan. 2018. Recognizing human actions as the evolution of pose estimation maps. In *CVPR*. 1159–1168.

[42] Yang Liu, Keze Wang, Guanbin Li, and Liang Lin. 2021. Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition. *IEEE Transactions on Image Processing* 30 (2021), 5573–5588.

[43] Jianjie Lu and Kai-Yu Tong. 2019. Robust single accelerometer-based activity recognition using modified recurrence plot. *IEEE Sensors Journal* 19, 15 (2019), 6317–6324.

[44] Taylor Mauldin, Anne H Ngu, Vangelis Metsis, and Marc E Canby. 2020. Ensemble deep learning on wearables using small datasets. *ACM Transactions on Computing for Healthcare* 2, 1 (2020), 1–30.

[45] Taylor R Mauldin, Marc E Canby, Vangelis Metsis, Anne HH Ngu, and Coralys Cubero Rivera. 2018. SmartFall: A smartwatch-based fall detection system using deep learning. *Sensors* 18, 10 (2018), 3363.

[46] Uwe Maurer, Asim Smailagic, Daniel P Siewiorek, and Michael Deisher. 2006. Activity recognition and monitoring using multiple sensors on different body positions. In *International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06)*. 4–pp.

[47] Sakorn Mekruksavanich and Anuchit Jitpattanakul. 2020. Smartwatch-based human activity recognition using hybrid lstm network. In *2020 IEEE SENSORS*. 1–4.

[48] Sakorn Mekruksavanich and Anuchit Jitpattanakul. 2021. A Multichannel CNN-LSTM network for daily activity recognition using smartwatch sensor data. In *2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering*. 277–280.

[49] Sakorn Mekruksavanich, Anuchit Jitpattanakul, Phichai Youplao, and Preecha Yupapin. 2020. Enhanced hand-oriented activity recognition based on smartwatch sensor data using lstms. *Symmetry* 12, 9 (2020), 1570.

[50] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *AAAI*, Vol. 34. 5191–5198.

[51] Federico Monti, Karl Otness, and Michael M Bronstein. 2018. Motifnet: a motif-based graph convolutional network for directed graphs. In *DSW*. 225–228.

[52] Jianyuan Ni, Raunak Sarbajna, Yang Liu, Anne HH Ngu, and Yan Yan. 2021. Cross-modal Knowledge Distillation for Vision-to-Sensor Action Recognition. *arXiv preprint arXiv:2112.01849* (2021).

[53] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley mhad: A comprehensive multimodal human action database. In *WACV*. 53–60.

[54] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.

[55] Ayodeji Oseni, Nour Moustafa, Helge Janicke, Peng Liu, Zahir Tari, and Athanasios Vasilakos. 2021. Security and privacy for artificial intelligence: Opportunities and challenges. *arXiv preprint arXiv:2102.04661* (2021).

[56] Madhuri Panwar, S Ram Dyuthi, K Chandra Prakash, Dwaipayan Biswas, Amit Acharyya, Koushik Maharatna, Arvind Gautam, and Ganesh R Naik. 2017. CNN based approach for activity recognition using a wrist-worn accelerometer. In *EMBC*. 2438–2441.

[57] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *CVPR*. 3967–3976.

[58] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. 2019. Correlation congruence for knowledge distillation. In *ICCV*. 5007–5016.

[59] Cuong Pham, Linh Nguyen, Anh Nguyen, Ngon Nguyen, and Van-Toi Nguyen. 2021. Combining skeleton and accelerometer data for human fine-grained activity recognition and abnormal behaviour detection with deep temporal convolutional networks. *Multimedia Tools and Applications* 80, 19 (2021), 28919–28940.

[60] Mehdi Rezagholizadeh, Aref Jafari, Puneeth Salad, Pranav Sharma, Ali Saheb Pasand, and Ali Ghodsi. 2021. Pro-KD: Progressive Distillation by Following the Footsteps of the Teacher. *arXiv preprint arXiv:2110.08532* (2021).

[61] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550* (2014).

[62] Guto Leoni Santos, Patricia Takako Endo, Kayo Henrique de Carvalho Monteiro, Elisson da Silva Rocha, Ivanovitch Silva, and Theo Lynn. 2019. Accelerometer-based human fall detection using convolutional neural networks. *Sensors* 19, 7 (2019), 1644.

[63] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*. 12026–12035.

[64] Wenxian Shi, Yuxuan Song, Hao Zhou, Bohan Li, and Lei Li. 2021. Follow your path: a progressive method for knowledge distillation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 596–611.

[65] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. 2020. Skeleton-based action recognition with hierarchical spatial reasoning and temporal stack learning network. *Pattern Recognition* 107 (2020), 107511.

[66] Satya P Singh, Madan Kumar Sharma, Aimé Lay-Ekuakille, Deepak Gangwar, and Sukrit Gupta. 2020. Deep ConvLSTM with self-attention for human activity decoding using wearable sensors. *IEEE Sensors Journal* 21, 6 (2020), 8575–8582.

[67] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, Vol. 31.

[68] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. 2021. Constructing stronger and faster baselines for skeleton-based action recognition. *arXiv preprint arXiv:2106.15125* (2021).

[69] Pablo Speciale, Johannes L Schonberger, Sing Bing Kang, Sudipta N Sinha, and Marc Pollefeys. 2019. Privacy preserving image-based localization. In *CVPR*. 5493–5503.

[70] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. 2020. Human action recognition from various data modalities: A review. *arXiv preprint arXiv:2012.11866* (2020).

[71] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* 30 (2017).

[72] Fida Mohammad Thoker and Juergen Gall. 2019. Cross-modal knowledge distillation for action recognition. In *ICIP*. 6–10.

[73] Frederick Tung and Greg Mori. 2019. Similarity-preserving knowledge distillation. In *ICCV*. 1365–1374.

[74] Suraj Vantigodi and R Venkatesh Babu. 2013. Real-time human action recognition from motion capture data. In *NCVPRIPG*. 1–4.

[75] Suraj Vantigodi and Venkatesh Babu Radhakrishnan. 2014. Action recognition from motion capture data using meta-cognitive rbf network classifier. In *ISSNIP*. 1–6.

[76] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*. 588–595.

[77] Vini Vijayan, James P Connolly, Joan Condell, Nigel McKelvey, and Philip Gardiner. 2021. Review of wearable devices and data collection considerations for connected health. *Sensors* 21, 16 (2021), 5589.

[78] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. 2015. A review of human activity recognition methods. *Frontiers in Robotics and AI* 2 (2015), 28.

[79] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. 2013. Learning actionlet ensemble for 3D human action recognition. *TPAMI* 36, 5 (2013), 914–927.

[80] Jiahao Wang, Qiuling Long, Kexuan Liu, Yingzi Xie, et al. 2019. Human action recognition on cellphone using compositional bidir-lstm-cnn networks. In *CNCI*. Atlantis Press, 687–692.

[81] Kai Wang, Yu Liu, Qian Ma, and Quan Z Sheng. 2021. Mulde: Multi-teacher knowledge distillation for low-dimensional knowledge graph embeddings. In *Proceedings of the Web Conference 2021*. 1716–1726.

[82] LuKun Wang and RuYue Liu. 2020. Human activity recognition based on wearable sensor using hierarchical deep LSTM networks. *Circuits, Systems, and Signal Processing* 39, 2 (2020), 837–856.

[83] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*. Springer, 20–36.

[84] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2018. Temporal segment networks for action recognition in videos. *TPAMI* 41, 11 (2018), 2740–2755.

[85] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li. 2016. Action recognition based on joint trajectory maps using convolutional neural networks. In *Proceedings of the 24th ACM international conference on Multimedia*. 102–106.

[86] Haoran Wei, Roozbeh Jafari, and Nasser Kehtarnavaz. 2019. Fusion of video and inertial sensing for deep learning–based human action recognition. *Sensors* 19, 17 (2019), 3680.

[87] Cong Wu, Xiao-Jun Wu, and Josef Kittler. 2019. Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition. In *ICCV workshops*. 0–0.

[88] Meng-Chieh Wu, Ching-Te Chiu, and Kun-Hsuan Wu. 2019. Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks. In *ICASSP*. 2202–2206.

[89] Kun Xia, Jianguang Huang, and Hanyu Wang. 2020. LSTM-CNN architecture for human activity recognition. *IEEE Access* 8 (2020), 56855–56866.

[90] Liuyu Xiang, Guiguang Ding, and Jungong Han. 2020. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *ECCV*. Springer, 247–263.

[91] Renyi Xiao, Yonghong Hou, Zihui Guo, Chuankun Li, Pichao Wang, and Wanqing Li. 2019. Self-attention guided deep features for action recognition. In *ICME*. 1060–1065.

[92] Ting-Bing Xu and Cheng-Lin Liu. 2019. Data-distortion guided self-distillation for deep neural networks. In *AAAI*, Vol. 33. 5565–5572.

[93] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*.

[94] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2021. A study of face obfuscation in imagenet. *arXiv preprint arXiv:2103.06191* (2021).

[95] Jingwen Ye, Yixin Ji, Xinchao Wang, Kairi Ou, Dapeng Tao, and Mingli Song. 2019. Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more. In *CVPR*. 2829–2838.

[96] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1285–1294.

[97] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *CVPR*. 13876–13885.

[98] Sergey Zagoruyko and Nikos Komodakis. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928* (2016).

[99] Ming Zeng, Le T Nguyen, Bo Yu, Ole J Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. 2014. Convolutional neural networks for human activity recognition using mobile sensors. In *6th international conference on mobile computing, applications and services*. 197–205.

[100] Hailin Zhang, Defang Chen, and Can Wang. 2021. Confidence-Aware Multi-Teacher Knowledge Distillation. *arXiv preprint arXiv:2201.00007* (2021).

[101] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*. 3713–3722.

[102] Songyang Zhang, Xiaoming Liu, and Jun Xiao. 2017. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 148–157.

[103] Songyang Zhang, Yang Yang, Jun Xiao, Xiaoming Liu, Yi Yang, Di Xie, and Yueting Zhuang. 2018. Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks. *IfEEE Transactions on Multimedia* 20, 9 (2018), 2330–2343.

[104] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. 2019. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems* 21, 9 (2019), 3848–3858.

[105] Rui Zhao, Kang Wang, Hui Su, and Qiang Ji. 2019. Bayesian graph convolution LSTM for skeleton based action recognition. In *ICCV*. 6882–6892.

[106] Yu Zhao, Rennong Yang, Guillaume Chevalier, Ximeng Xu, and Zhenxing Zhang. 2018. Deep residual bidir-LSTM for human activity recognition using wearable sensors. *Mathematical Problems in Engineering* 2018 (2018).

[107] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. 2018. Temporal relational reasoning in videos. In *ECCV*. 803–818.