Multi-agent Transformer Networks for Multimodal Human Activity Recognition

Jingcheng Li[*]
The University of New South Wales
Sydney, Australia
jingcheng.li@unsw.edu.au

Lina Yao
Data 61, CSIRO
The University of New South Wales
Sydney, Australia
lina.yao@unsw.edu.au

Binghao Li
The University of New South Wales
Sydney, Australia
binghao.li@unsw.edu.au

Xianzhi Wang
The University of Technology Sydney
Sydney, Australia
xianzhi.wang@uts.edu.au

Claude Sammut
The University of New South Wales
Sydney, Australia
c.sammut@unsw.edu.au

ABSTRACT

Human activity recognition has become an important challenge yet to resolve while also having promising benefits in various applications for years. Existing approaches have made great progress by applying deep learning and attention-based methods. However, the deep learning-based approaches may not fully exploit the features to resolve multimodal human activity recognition tasks. Also, the potential of attention-based methods still has not been fully explored to better extract the multimodal spatial-temporal relationship and produce robust results. In this work, we propose Multi-agent Transformer Network (MATN), a multi-agent attention-based deep learning algorithm, to address the above issues in multimodal human activity recognition. We first design a unified representation learning layer to encode the multimodal data, which preprocesses the data in a generalized and efficient way. Then we develop a multimodal spatial-temporal transformer module that applies the attention mechanism to extract the salient spatial-temporal features. Finally, we use a multi-agent training module to collaboratively select the informative modalities and predict the activity labels. We have extensively conducted experiments to evaluate MATN's performance on two public multimodal human activity recognition datasets. The results show that our model has achieved competitive performance compared to the state-of-the-art approaches, which also demonstrates scalability, effectiveness, and robustness.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Information systems** → *Data mining*; **Multimedia and multimodal retrieval**.

KEYWORDS

Activity recognition, neural networks, multi-agent reinforcement learning, multimodal learning

1 INTRODUCTION

Human activity recognition is a significant step towards human-computer interaction and enables a series of promising applications such as assisted living, skills training, health monitoring, and robotics [11]. The multimodal human activity recognition task involves processing multi-variant data modalities and correctly predicting the label of the activity.

Many existing approaches resolve the human activity recognition task by using uni-modal data, such as RGB, skeleton, and inertial data. Traditional machine learning approaches process hand-crafted features to predict the activity labels. The performance of these methods heavily relies on feature engineering, which cannot generalize well when a new task is introduced or the data quality is bad. Recent studies focus on resolving the human activity recognition task by applying deep learning methods, such as Convolution Neural Network (CNN) and Long-Short Term Memory (LSTM) Networks. These approaches are able to learn the representation of the features and predict activity labels automatically without involving human effort. However, by processing the input data as 2-D images, the CNN-based models treat the temporal dimension and the spatial dimension equally, so that they cannot fully exploit the order information, which is important in human activity recognition. While sequential models, such as LSTM networks, treat the input data as temporal sequences and extract the temporal information, they cannot exploit the spatial channel information into account. For some modalities, such as inertial and skeleton data, each channel dimension may contain salient spatial information, which is useful for identifying similar activities. Also, uni-modal approaches may not be robust enough and cannot generalize well if the input data has low quality. As a result, they cannot be widely deployed in real-world situations.

Multimodal approaches have been explored to mitigate the disadvantage of uni-modal methods. As the multimodal data can provide complementary information in the prediction-making process, these approaches are able to achieve more robust performance. However, while multimodal human activity recognition approaches can better extract the informative spatial and temporal features from multimodal data and produce better results, there remains several challenges when processing multimodal data to produce more robust and better results. First, current approaches apply modality-specific data engineering methods, such as CNN and LSTM networks, to generate embeddings of the input data. Such methods result in high complexity and cannot preserve the original structural information in the data. Second, current approaches design complex architectures to extract spatial and temporal features and generate high-level representations by using deep learning-based

[*] This is the corresponding author.

frameworks. When they try to resolve the human activity recognition task in a multimodal scenario, this further increases the computation cost. Thus, these approaches cannot work efficiently when they are deployed in the real-world environment. Third, while the multimodal data can provide comprehensive and complementary information, how to effectively extract the salient features and fuse them still need further exploration. Hence, it is essential to develop human activity recognition methods that are scalable, robust and accurate enough so that they can be widely deployed in real-life circumstances.

To address the challenges above, in this work, we propose Multi-agent Transformer Networks (MATN), a novel multimodal human activity learning approach which can extract the salient spatial and temporal features. The model is scalable, robust, and can be generalized to variant modalities as well as new subjects. MATN first separates the multimodal input data into segments and encodes them into a unified representation. Instead of using CNN or LSTM-based approaches to generate embeddings of the features, our representation learning layer preserves the original information, which can be processed in the feature learning module effectively. Also, the unified representation learning layer does not require modality-specific data engineering methods, which reduces the model complexity, thus leading to improved efficiency. MATN then applies a multimodal spatial-temporal transformer module to extract the salient spatial-temporal features of each modality and generates the high-level representation of the input data. Unlike LSTM networks, which suffer from the sequence processing problem and cannot be trained in parallel, the transformer is more scalable as it applies the self-attention mechanism, which mitigates the problem by treating the sequence as a whole. Also, the spatial-temporal transformer module is able to extract both the salient spatial and temporal features, thus preserving more information and producing more accurate results. Finally, in order to fuse the output of the multimodal streams, MATN uses a multi-agent collaboration module to select the informative modalities and generate the final prediction. While lots of existing fusion approaches simply concatenate or add the multimodal information, we use this multi-agent collaboration approach, which is a joint optimization process that can adaptively update the learning parameters and adjust the weight of each modality.

We conducted extensive experiments to evaluate MATN's performance on two public multimodal human activity recognition datasets, UTD-MHAD [9] and MMAct [29], using two subject independent evaluation protocols. The results show that our model has achieved competitive performance compared to the state-of-the-art approaches.

The key contributions in this paper are summarized as follows:

- We proposed the Multi-agent Transformer Networks for the multimodal human activity recognition task, which achieved better performance and scalability.
- We presented a unified sequence-to-sequence model that can be generalized to various modalities without requiring modality-specific encoder architecture and extra data engineering.

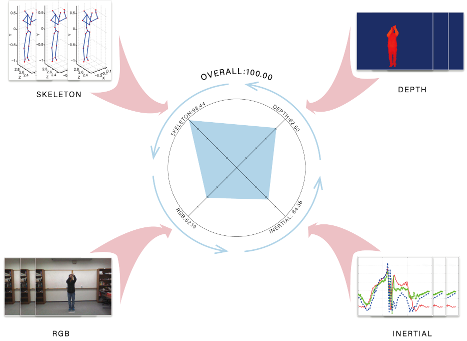


Figure 1: The illustrating scenario of how the Multi-agent Transformer Networks performs multimodal human activity recognition to the basketball shooting activity from the UTD-MHAD dataset [9]. Here, the skeleton modality and the depth modality contribute more than the RGB modality and the inertial modality when generating predictions.

- We used a multi-agent reinforcement learning based method to fuse the multimodal features and resolve the human activity recognition task.

The rest of this paper is organized as follows: Section 2 introduces the related work; Section 3 formulates the problem of multimodal human activity recognition; Section 4 presents the details of the proposed MATN model; Section 5 introduces the datasets and experimental settings; Section 6 reports the evaluation results; finally, we conclude the paper in Section 7.

2 RELATED WORK

Human Activity Recognition. Human activity recognition is a significant step towards human-computer interaction and enables a series of promising applications such as assisted living, skills training, health monitoring, and robotics [11]. Generally, the procedure of resolving the human activity recognition task can be separated into three main pathways, vision-based approaches, sensor-based approaches, and multimodal approaches.

The vision-based approaches take images or videos as the input and then perform an analysis of human behaviors. Early approaches focus on traditional machine learning models [1, 6, 40, 42, 44, 45, 47]. Those models heavily rely on the input features, and become less generalized when they are applied to different tasks. Recent methods mainly focus on further extracting the spatial-temporal relationship by utilizing deep learning approaches such as CNN, LSTM networks, and Graph Neural Networks [17, 19, 21, 30, 36, 48, 49].

Sensor-based approaches utilize on-body or ambient sensors to dead reckon people's motion details or log their activity tracks [11]. Traditional machine learning approaches [8, 15, 16, 27, 39, 52] were also explored in the early stage. Recent approaches utilized deep neural networks, such as CNN [2, 10, 26, 41, 50, 51, 54] and LSTM Networks [5, 18, 35, 38, 53], as feature extractors to learn the representation of the input sensory segments automatically, then map the representation to labels using another neural network [3].

However, approaches using single modality data may not be robust enough and usually suffers from noise or even data loss. If data quality is poor, the performance will drop significantly, prohibiting them from being widely used in real-world circumstances.

Multimodal human activity recognition approaches aim to resolve the task by extracting features from data of different modalities. Common modalities explored include RGB videos, depth videos, skeleton positions, inertial sensor signals, Wi-Fi signals, and pressure signals. In recent years, multimodal frameworks [9, 20, 24, 25, 37] have been explored to resolve the human activity recognition problem, as they can observe the same phenomenon and capture the complementary information, thus producing more robust results [4]. Chen et al. [9] developed a hybrid model which combines depth motion maps and partitioned temporal windows to perform human activity recognition on depth and inertial data. While they used different approaches for each modality, which makes feature extraction more complex, visual modalities were not considered. Memmesheimer et al. [37] proposed a novel discriminative encoding method that first transferred the skeleton and inertial data into signal images, then used EfficientNet [43] as a backbone to perform image classification. While they presented a novel way of fusing the multimodal data and achieved significant performance with lite architecture, the approach can only be applied to data that can be represented as 1-D signals, and it still lacks a way to transfer video data into the corresponding format.

The Attention Mechanism. In recent years, many works have started to explore the potential of the attention mechanism in the human activity recognition area. Recently, the proposed transformer model [46] has been widely studied and applied in many different areas. While LSTM networks can handle the long-range dependencies, the sequences must be processed token by token so that they cannot be trained in parallel. Transformers can process the sequence as a whole and integrate the information, which mitigates the long-range dependency issues and improves the scalability. However, transformers cannot capture the order information within the feature sequence, making the addition of position embeddings a necessary step to process the positional information of each token in a sequence. With the inspiration of transformer, many approaches utilized the multi-head self-attention mechanism and received SOTA results [12, 24, 25, 31, 32]. Islam and Iqbal [24] also explored the potential of the attention mechanism in the human activity recognition area by developing a multimodal hierarchical attention approach to sequentially extract the spatial and temporal features for each modality, and then fusing and passing them through a multimodal attention unit to generate predictions for human activity recognition. Their later work [25] extended [24] by adding an additional mixture-of-experts model to extract the salient features and using a cross-modal graphical attention method to fuse the features. While their work performed better in terms of extracting the salient features and pushed the state-of-the-art performance to a new level, the approach may need to use separate pre-processing methods for each modality which result in a complex architecture.

The multimodal data are able to provide complementary information, which contributes to generating more robust and accurate results. However, how to fuse the multimodal features in an efficient and effective way remains a challenge yet to resolve. For example, many existing approaches can extract salient features from specific varieties of modalities, which could perform poorly if different modalities were used, or the data contains noises. Hence, these approaches cannot be generalized to new modalities and are not robust enough. Also, many approaches mainly focus on developing more complex frameworks to extract features from each modality and use LSTM networks to generate the prediction. While this may require high time complexity, how to perform multimodal fusion is still a problem worth exploring. There still lacks a generic and scalable way to thoroughly leverage and effectively fuse the multimodal information. Moreover, in both research and real-world circumstances, it is hard to collect and annotate enough training data with high quality and diversity. The data collection step may focus on quite diverse requirements, such as high data quality, large numbers of modalities or sensors, long-term recordings, or large numbers of participants [7]. As the amount of training data can be limited, they are also easily affected by the noises. As a result, when there are only a few training examples of the activities that have not been seen before, these approaches cannot generalize well to provide suitable output. This is known as the problem of data scarcity. Currently, there are only a small number of multimodal benchmark datasets available.

Hence, to mitigate the problems above, we proposed a unified sequence-to-sequence model that can be generalized to different varieties of modalities. Also, we developed a multi-agent reinforcement learning-based method to fuse the multimodal features and resolve the human activity recognition task.

3 PROBLEM DEFINITION

Multimodal Human activity Recognition involves multiple data modalities recorded using different devices, such as Inertial Measurement Units (IMU), smartphones, smartwatches, RGB cameras and depth cameras, etc. Meanwhile, each device may record different kinds of data and each kind of data may contain multiple dimensions. Similar to the human activity recognition procedure proposed by [7], we define the multimodal human activity recognition problem as follows: Let M be the number of modalities involved and $c_{i,t}$ be the data of modality i ($1 \leq i \leq M$), then for each modality the time sequence would be $c_i = [c_{i,1}, c_{i,2}, \ldots, c_{i,t}, \ldots]$, t denotes the timestep. For each modality, we divide each data stream into segments with a fixed-length sliding window. So the segments of modality m can be represented as $Seg_m = [x_{m,1}, x_{m,2}, \ldots, x_{m,L}]$, where L denotes the segment length. Then given the input $\mathbf{X} = [Seg_1, Seg_2, \ldots, Seg_M]$, our objective is to learn a function $\mathcal{F}(Seg; \Theta)$ to correctly predict the label of the activity y, where Θ represents all the parameters to be learned during the training step.

4 PROPOSED MODEL

We propose our Multi-agent Transformer Networks (MATN) for multimodal human activity recognition. The overall framework is shown in Figure 2. Our model contains three components:

(i) a generalized representation learning layer that receives the raw data and conducts data preprocessing for further extraction.

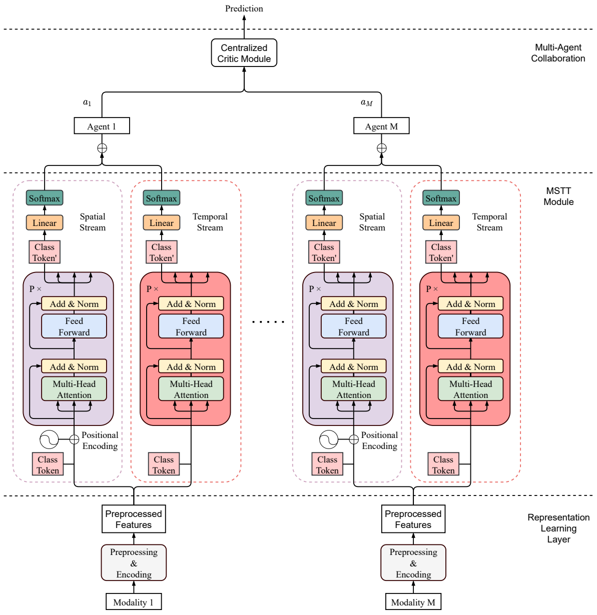


Figure 2: The Multi-agent Transformer Networks Module

(ii) a multimodal spatial-temporal transformer (MSTT) module that extracts the salient features for each modality and generates the representation of class tokens.

(iii) a multi-agent collaboration module that learns to select the informative modalities and produce the final output.

At first, the representation learning layer will preprocess and transfer the input data of each modality into a unified representation. Then in the MSTT module, we concatenate the input features with a class token for both the spatial and temporal streams. For the temporal stream, we add position encoding to the input features. The transformer encoder module then extracts the salient spatial and temporal features and outputs the representations of class tokens. In the multi-agent collaboration module, we assign an agent for each modality and aggregate the results. The model is incrementally trained with a centralized policy to predict the activity label.

In the following, we will elaborate on the preprocessing method (Section 4.1), the MSTT module (Section 4.2), the multi-agent collaboration module (Section 4.3), and the training and optimization process.

4.1 Unified Representation Learning

MATN first takes the raw multimodal data and transfers each modality into a unified representation to be fed into the MSTT module. The dataset D is a set of data records of all the modalities, where d_m^n is the n^{th} record of modality m.

$$D = \begin{bmatrix} d_1^1 & \cdots & d_M^1 \\ \vdots & \ddots & \vdots \\ d_1^N & \cdots & d_M^N \end{bmatrix}$$

For record n, $d_{m,t}^n$ denotes the record at timestep t. It is worth noting that $d_{m,t}^n$ could either be a one-dimensional vector (sensor or skeleton data) or a two-dimensional matrix (visual data, excluding the channel dimension) depending on the input data format.

$$d_m^n = \begin{bmatrix} d_{m,1}^n & \cdots & d_{m,T}^n \end{bmatrix}$$

Our method has two advantages. First, we use a common method to generate unified representations for all the modalities, which can be easily generalized to a new modality. This would not require complex encoder architecture and extra data engineering. Second, our approach could utilize a pre-trained model and be completed offline. As it does not require a sequential Neural Networks module, the computation can be done in parallel and easily scale when a new modality is introduced, which improves the computation cost during both training and testing.

As some modalities may contain high-frequency data, the adjacent frames may contain similar information, which means extracting features at the frame level would be both memory and computationally inefficient. Also, there can be fluctuation during a short period of time and the noisy data would affect the performance. Thus, we divide each data stream into segments with a fixed-length sliding window and conduct average pooling on the data within the time window. As the time to complete an activity may vary, for each modality, we keep the transferred records with the same length L_m over time to better support batch processing.

While this method works on modalities that are one-dimensional over time, it cannot be directly applied to visual data. Recently, several approaches [14, 33] developed visual transformers that can resolve image classification problems. However, they mainly work under a more static scenario and require high computation costs, thus cannot be utilized in human activity recognition tasks yet. Hence, we use an extra step to encode visual data into one-dimensional vectors over time. To resolve this problem, we apply a pre-trained ResNet50 [22] model to generate encodings for RGB and depth video data. Then, the encoded sequence of modality m can be represented as X_m with a size of $B \times L_m \times C_m$, where B denotes the batch size and C_m denotes the feature dimension.

$$X_m = \begin{bmatrix} x_{m,1}^1 & \cdots & x_{m,L}^1 \\ \vdots & \ddots & \vdots \\ x_{m,1}^N & \cdots & x_{m,L}^N \end{bmatrix}$$

4.2 Multimodal Spatial-Temporal Transformer

The multimodal spatial-temporal transformer (MSTT) module executes in a parallel stream to separately extract the salient spatial and temporal features of each modality. Unlike LSTM, which may suffer from the long-range dependency problem, self-attention-based methods can pay attention to the entire feature sequence, thus producing a more informative and robust representation of the input data. While each modality of the input data has a different representation, the spatial series and the temporal series also reveal unique information. Then we apply a multi-head transformer encoder [46] on both the spatial series and temporal series features, so the module can fully distinguish and extract the salient unimodal features over spatiality and time.

Both the spatial stream and the temporal stream contain a stack of P-layer transformers, where each layer has two sequentially connected sub-layers, a multi-head self-attention layer and a position-wise feed-forward network. A residual connection is employed around each sub-layer, followed by a layer normalization.

Taking the inspiration from BERT [13] and ViT [14], For each modality m, we add a learnable class token x_{cls} to the input sequence of each stream and use it to generate the final prediction as it can serve as an overall representation of the input features. Also, unlike CNN or RNN, the transformer model cannot take the order information of the input sequence. So, for the temporal stream, we add sinusoidal positional encoding to the input features to inject the absolute positional information of the tokens in the sequence.

We also experimented with the performance of using learned positional encoding, and just as [46] mentioned, it did not improve the overall performance. Hence, we use the absolute positional encoding to reduce the computation cost.

$$\begin{aligned} PE_{(pos,2i)} &= sin(pos/10000^{2i/d_{model}}) \\ PE_{(pos,2i+1)} &= cos(pos/10000^{2i/d_{model}}) \end{aligned} \tag{1}$$

For each modality m, we have the input segments X with a size of $L \times C$, which can be directly fed as the input of the temporal stream $X_{m,T} = [x_1, x_2, \ldots, x_L]$. A transpose operation is conducted on X to be used as the input of the spatial stream $X_{m,S} = [x_1, x_2, \ldots, x_C]$. Then the input features of the transformer H^S (spatial stream) and H^T (temporal stream) can be constructed as follows, where S denotes the spatial stream, T denotes the temporal stream, $E_{pos} \in \mathbb{R}^{(L+1)\times C}$ denotes the positional encoding.

$$H_{m,S} = X^T = [x_{S,cls}, x_1, \ldots, x_C] + E_{pos} \tag{2}$$

$$H_{m,T} = [x_{T,cls}, x_1, \ldots, xL] \tag{3}$$

For a single transformer, to extract the salient features from the input data, the queries Q_m, keys K_m and values V_m are constructed by linear projections on the input H_m.

$$Q_m = H_m W_m^Q \quad K_m = H_m W_m^K \quad V_m = H_m W_m^V \tag{4}$$

Where projections parameters $W_m^Q \in \mathbb{R}^{d_{model,m}\times d_{k,m}}$, $W_m^K \in \mathbb{R}^{d_{model,m}\times d_{k,m}}$ and $W_m^V \in \mathbb{R}^{d_{model,m}\times d_{v,m}}$. $d_{k,m}$ denotes the queries and keys dimension, $d_{v,m}$ denotes the spatial-temporal values dimension. The self-attention function computes the scaled dot products of the query with all keys to obtain the weights on the values. $d_{k,m}$ is a scaling factor to smooth the gradients in the function.

$$Attention_m(Q_m, K_m, V_m) = Softmax(\frac{Q_m K_m^{\intercal}}{\sqrt{d_{k,m}}})V_m \tag{5}$$

Multi-head attention conducts linear projects on the queries, keys and values by h times to jointly attend to information from different representation subspaces. Different projections parameters are used at each time and the outputs are then concatenated and projected to output the final values.

$$\begin{aligned} MultiHead_m(Q_m, K_m, V_m) &= [head_{m,1}, \ldots, head_{m,h}]W_m^O \\ where\ head_{m,i} &= Attention(Q_{m,i}, K_{m,i}, V_{m,i}) \end{aligned} \tag{6}$$

Where $W_m^O \in \mathbb{R}^{hd_{v,m}\times d_{model,m}}$ denotes the projection parameters used in the end. To reduce the computation cost, for all the sub-layers we use $d_{k,m} = d_{v,m} = d_{model,m}/h$.

The output of the multi-head self-attention layer is then passed through a position-wise feed-forward network with two linear layers and ReLu activation in between. Then for each transformer layer, the output would be $H_{i,m}$ and the final output would be $H_{P,m}$.

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{7}$$

For both the spatial and temporal stream, we extract the representation of the corresponding class tokens $h_{S,P,0}$ and $h_{T,P,0}$, then pass them through a linear layer. A Softmax function is applied to produce the aggregated results.

$$Y_m = Softmax(LN_S(h_{m,S,P,0})) + Softmax(LN_T(h_{m,T,P,0})) \tag{8}$$

4.3 Multi-agent Collaborative Training and Optimization

In our work, we use a multi-agent decentralized actor and centralized critic approach to predict the activity class of each input. For each MSTT module's output Y_m, we assign an agent a_m to each modality. In each episode, the model aggregates the predictions of all the modalities. Thus, the model is able to select the informative agents and maximize the reward over the episodes. The agents individually make observations o_m based on each input segment X_m and outputs an action A_m to select the class label l_i.

$$A_m \sim P_m(\cdot | f_m(X_m, o_m)) \tag{9}$$

To align the selection policy, we set a common goal for the centralized critic, which is to correctly predict the activity class after each observation and selection. Unlike recurrent networks such as LSTM, where a reward is given at each timestep t, we only consider the final outputs of the MSTT modules and assign the reward at the end of each training episode. A reward function R is used where a positive reward is assigned if a correct prediction is made, and no reward is assigned if an incorrect prediction is made.

As this is a classification problem, we add the cross-entropy loss into the loss function during the training step. Then for each modality m, the loss function is:

$$\mathcal{L}_m = -\alpha \sum_{i=1}^{Z} r(i)log(F_m(\theta; X_m)) + \sum_{i=1}^{Z} y_i log(F_m(\theta; X_m)) \tag{10}$$

Where F_m denotes the general function that generates the output based on each input X_m, α is a constant multiplication factor to adjust the balance between the reward and the cross-entropy loss, Z denotes the number of activity classes, y_i is the ground truth label and $r(i) = p(i)$ (the probability of $\hat{y}_i$) if the prediction is correct and 0 otherwise. Also, instead of simply integrating the multimodal information with the same weight, inspired by the uncertainty weighted loss [28], we also assign a weight parameter to jointly weight the loss of each agent. As a result, this joint optimization process is able to adjust the weight of each modality and select the informative representation by adaptively updating the learning parameters. The overall loss function is:

$$\mathcal{L} = \sum_{m=1}^{M} \frac{1}{\sigma_m^2} \mathcal{L}_m + \sum_{m=1}^{M} log(\sigma_m) \tag{11}$$

Therefore, the overall training and optimization process can be summarized as maximizing the reward R while minimizing the loss. During the testing step, the model assembles each MSTT module's output of modality m and outputs the prediction Y.

$$\hat{Y} = \sum_{m=1}^{M} Y_m \tag{12}$$

5 DATASETS AND EXPERIMENTS

This section reports our experimental setup on two multimodal datasets for human activity recognition. We first introduce the datasets to be used and then explain the experimental protocol and evaluation metrics to test our model's performance. Finally, we describe the experimental settings when conducting the experiments.

5.1 Datasets

We evaluate MATN's performance and compare it with multiple contemporary multimodal human activity recognition approaches on two public benchmark datasets, UTD-MHAD and MMAct. It is worth mentioning that currently, these are the only two mainstream multimodal human activity recognition datasets available.

The UTD-MHAD dataset [9] contains 27 activities performed by eight subjects, where each subject repeated the activity four times. After removing the corrupted sequences, the dataset contains 861 clips. For each activity, the modalities available are RGB, depth, skeleton and inertial sensors. The Kinect camera is used to record the RGB-D information. A wearable inertial sensor is placed on either the subject's wrist or leg, depending on which part is mostly used to perform the action. The 3-axis acceleration, the 3-axis gyroscope and the 3-axis magnetic strength information were recorded. The MMAct dataset [29] contains 35 daily life activities performed by 20 subjects, where each subject repeated the activity five times. After removing the corrupted sequences, the dataset contains 36 K clips. The dataset consists of 7 modalities, including RGB, skeleton, acceleration, gyroscope, orientation, Wi-Fi, and pressure. 4 surveillance cameras and a smart glass were used to record the RGB data from 5 different views. A smartphone was put into the pocket of the subjects' pants to record acceleration, gyroscope, orientation, Wi-Fi and pressure data. A smartwatch was also used to record additional acceleration data. We used RGB, skeleton, acceleration, gyroscope, and orientation data to conduct the experiments. It is also worth mentioning that previously, the number of activities was mistakenly reported as 37.

5.2 Evaluation Protocol

The evaluation protocol is important for developing discriminative human activity recognition approaches, especially to the models' generalization ability. While the subject-dependent protocols randomly split all the data into the training and testing sets, which leads to a selection bias where the common information of each subject is shared through the training set and the testing set. However, subject-dependent settings are unsuitable for real-life deployment because the task focuses on new users in the real-world scenario. Hence, we conduct our experiments by applying the subject-independent protocols where the data of different subjects are used in the training and testing sets.

For the UTD-MHAD dataset, we used two subject-independent protocols to conduct the experiments. First, we applied a 50-50 evaluation, where the first half of the subjects (1-4) were used for training the model and the other half (5-8) were used for testing. Also, we used a leave-one-subject-out (LOSO) protocol, where a subject's data was used for testing and the rest of the data was used for training. Instead of testing on only one selected subject, we performed a comprehensive evaluation by iteratively applying the LOSO protocol to each subject and taking the average result to reduce the bias. For the MMAct dataset, we followed the evaluation protocols proposed by the authors [29], cross-subject and cross-session. For the cross-subject setting, we used data of the first 80% of the subjects (1 to 16) as the training set and used the rest as the testing set. For the cross-session setting, we used data from the first 80% of the sessions for training and the rest as the testing set.

5.3 Experimental Settings

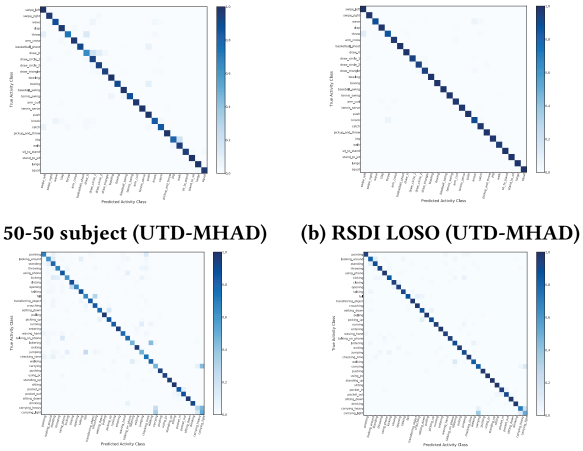
Each kind of data may contain multiple dimensions. For example, the IMU data may contain the accelerometer data, the gyroscope data and the magnetometer data. Each of them has three dimensions: x, y and z. In this work, we treated each main category as a single modality, e.g., the IMU data would contain 9 dimensions for the UTD-MHAD dataset. For each modality, we divided each data stream into segments with a fixed-length sliding window. For the RGB and depth data, we passed the segments through ResNet50 to generate the encodings. For the rest modalities, we directly passed them into the network without using extra feature extraction methods. We transposed the segments to get the input features for the spatial stream. A class token was concatenated with the spatial and temporal segments. We added sinusoidal positional encoding to the spatial stream. We initialized the parameters with uniform initialization and optimized them by using Adam optimizer with a learning rate of 0.001 for the two datasets. We applied RELU activation and dropout after each layer. We implemented the model using PyTorch and ran the experiments on an NVIDIA Titan RTX GPU. We used accuracy for experiments on the UTD-MHAD dataset and F1-score for experiments on the MMAct dataset as the evaluation metrics to measure the model's performance as suggested by the original authors.

6 RESULTS AND COMPARISONS

6.1 Overall Comparison

We evaluated the performance of MATN by conducting experiments on two multimodal HAR datasets: UTD-MHAD and MMAct. The confusion matrices are presented in Figure 3.

For the UTD-MHAD dataset, we apply both the 50-50 evaluation protocol and the LOSO evaluation protocol, and top-1 accuracy is used as the evaluation metric. The experimental results on the UTD-MHAD dataset are shown in Table 1 and Table 2. We compare our approach to the baseline approach as well as more recent multimodal approaches. Under the 50-50 protocol, the results show that MATN outperforms the other multimodal approaches by achieving



(a) 50-50 subject (UTD-MHAD)

(b) RSDI LOSO (UTD-MHAD)

(c) Cross-subject (MMAct)

(d) Cross-session (MMAct)

Figure 3: Confusion matrices for the overall experiments on the UTD-MHAD dataset and the MMAct dataset

Table 1: 50-50 subject performance comparison on the UTD-MHAD dataset. S: Skeleton, D: Depth, I: Inertial.

Method	**Modality Combination**	**Accuracy (%)**
MHAD [9]	I + D	79.10
Gimme Signals [37]	I + S	76.13
Gimme Signals [37]	I + S (data augmentation)	86.53
MATN (Our method)	I + D	81.86
MATN (Our method)	I + S	**92.72**

Table 2: LOSO performance comparison on the UTD-MHAD dataset. R: RGB, S: Skeleton, D: Depth, I: Inertial.

Method	Modality Combination				
	S + I	D + I	R + S	R + S + I	R + S + D + I
Keyless [34]	-	-	90.20	92.67	83.87
HAMLET [24]	-	-	95.12	91.16	90.09
Multi-GAT [25]	-	-	**96.27**	96.75	**97.56**
MATN (Our method)	**98.48**	**93.19**	90.37	**97.62**	97.46

Table 3: Cross-subject performance comparison on the MMAct dataset. Acc: Acceleration, Gyo: Gyroscope, Ori: Orientation.

Method	**Modality Combination**	**F1-Score (%)**
SMD [23]	Acc + RGB	63.89
Student [29]	RGB	64.44
Multi-teachers [29]	Acc + Gyo + Ori	62.67
MMD [29]	Acc + Gyo + Ori + RGB	64.33
MMAD [29]	Acc + Gyo + Ori + RGB	66.45
HAMLET [24]	Acc + Gyo + Ori + RGB	69.35
Keyless [34]	Acc + Gyo + Ori + RGB	71.83
Multi-GAT [25]	Acc + Gyo + Ori + RGB	75.24
MATN (Our method)	Acc + Gyo + Ori + RGB	**83.67**

Table 4: Cross-session performance comparison on the MMAct dataset. Acc: Acceleration, Gyo: Gyroscope, Ori: Orientation.

Method	**Modality Combination**	**F1-Score (%)**
MMAD [29]	Acc+Gyo+Ori+RGB	74.58
MMAD(Fusion) [29]	Acc+Gyo+Ori+RGB	78.82
Keyless [34]	Acc+Gyo+Ori+RGB	81.11
HAMLET [24]	Acc+Gyo+Ori+RGB	83.89
Multi-GAT [25]	Acc+Gyo+Ori+RGB	91.48
MATN (Our method)	Acc+Gyo+Ori+RGB	**91.85**

92.72% accuracy with skeleton and inertial data. Also, MATN outperforms the MHAD baseline with 81.86% using inertial and depth data. For the LOSO experimental setting, MATN outperforms the other multimodal approaches by achieving 98.48% with skeleton and inertial data. We also evaluate MATN's performance with inertial and depth data, which achieves 93.19%. Multi-GAT performs slightly better than our method when RGB, depth, skeleton, and inertial data are used.

For the MMAct dataset, we apply the cross-subject and cross-session evaluation protocol and use the F1-Score as the evaluation metric as suggested by the original paper [29]. The experimental results on the MMAct dataset are shown in Table 3 and Table 4. The results suggest that MATN outperforms the other multimodal approaches by achieving 83.67% under the cross-subject protocol and 91.85% under the cross-session protocol. MATN improves the results by 8.43% and 0.37% over the state-of-the-art multimodal human activity recognition approaches, respectively, under the cross-subject and cross-session protocol.

Many existing human activity recognition approaches aim to achieve high performance by using only one modality, while the multimodal approaches are not fully explored. MATN has shown good performance on the two datasets with different modality combinations. With the attention-based MSTT module, MATN is able to extract the salient spatial and temporal features and achieves improved results compared to the non-attention approaches. When compared to the attention-based approaches, such as HAMLET and Multi-GAT, MATN still achieves similar results with a simple and general architecture. It is worth mentioning that MATN achieves an accuracy of 98.48% by just using the skeleton and inertial data. While this significantly improves the computational efficiency. Also, one of the advantages of using non-RGB data is to mitigate the concern of privacy.

The results on the MMAct dataset show that for all the methods listed, there is a gap in performance between the cross-subject and cross-session experimental protocols. It is reasonable that MATN achieves better performance under the cross-session protocol. As both the training and testing share the same set of subjects, thus the inter-subject variation is not fully considered. However, when new subjects are completely used in the testing set, even if MATN outperforms the other multimodal approaches, the result is still 8.18% lower. This is in accordance with our discussion about the evaluation protocol above. In the real-world scenario, the human activity recognition model will be deployed to serve new users instead of the experiment participants only. If the cross-session experimental protocol was used, then it can be inappropriate and misleading, and the model would perform poorly after deployment. Hence, the future focus should be on developing robust models that can still perform well under the subject-independent scenario.

6.2 Contribution of Modalities

Multimodal human activity recognition approaches aim to resolve the problem by using data of different modalities, where the same phenomenon and complementary information are beneficial to produce more robust results. In this part, we perform further experiments to examine the contributions of different modalities to the prediction performance. The experiments are conducted on the UTD-MHAD dataset and the MMAct dataset, under the LOSO protocol and cross-subject protocol, where the RGB, depth, skeleton and inertial data are used. The results show that MATN is able to capture the common salient features as well as the complementary information, thus achieving better performance.

When MATN generates the prediction, except for the aggregated results, we also record the prediction of each agent representing a single modality. The overall performance and modality-specific performance are then evaluated across each activity class. The results in Figure 4 and Figure 5 show that, in general, the inertial stream and skeleton stream outperforms both the RGB stream and depth stream. One possible reason may be that it is easier to extract the salient features from the inertial data and skeleton data, as they have fewer dimensions and each dimension records the representative information of human activities. Also, as the RGB data and depth data are two-dimensional vectors over time, they may contain more uninformative information and the encoding step may further intensify the information loss. Further approaches are worth exploring to directly make use of the RGB and depth data.

Also, while the performance of each modality stream may vary, after collaboratively aggregating the outputs of the modalities, the overall performance is improved and becomes more robust. For example, in Figure 4, while the skeleton stream performs better than the other three modalities for class 1 (swipe left), the inertial stream outperforms the other modalities for class 3 (wave). However, after aggregating the information of all the modality streams, the overall prediction outperforms each modality stream. This shows that each modality stream may contain some modality-specific information that is not included in the other modality streams. As a result, each modality contributes some salient information that helps to produce better and more robust performance.

Table 5: Performance comparison of different self-attention architectures on the UTD-MHAD dataset

Experimental Setting	Accuracy (%)
Spatial Attention Only	77.01
Temporal Attention Only	86.08
Stacked Spatial-temporal Attention	83.46
Parallel Spatial-temporal Attention (MSTN)	**92.72**

6.3 Spatial-temporal attention architectures

The salient information in the spatial and temporal features plays an important role in human activity recognition. In this section, we conduct experiments to further investigate different architectures' ability to extract salient spatial and temporal features. As shown in the experiments, the proposed MSTN architecture can achieve a more effective and efficient performance.

We compare our MSTN module with three different self-attention architectures, spatial attention, temporal attention and stacked spatial-temporal attention in Table 5. For the spatial attention and temporal attention architectures, each block consists of either a spatial attention module or a temporal attention module, followed by an MLP layer. For the stacked spatial-temporal attention architecture, the input features are first passed through the spatial block, followed by the temporal block. We conducted the experiments on the UTD-MHAD dataset under the 50-50 subject protocol. The experimental results are shown in Table 6, demonstrating that MATN, where spatial attention and temporal attention are separately applied in a parallel way, outperforms the other three architectures. Also, while MATN achieves higher accuracy, by separating the spatial and temporal attention architecture, it could better support parallel computing, which improves efficiency.

Table 6: Performance comparison of position encoding on the UTD-MHAD dataset

Experimental Setting	Accuracy (%)
No Position Encoding	89.11
Position Encoding on Temporal Stream	86.12
Position Encoding on Spatial Stream	**92.72**
Position Encoding on Both Streams	88.35

6.4 Impact of Position Encoding

Vaswani et al. [46] introduced position encoding to mitigate the problem that transformers cannot make use of the order information of the input sequence. As our approach involves both the spatial and temporal streams, to investigate the influence of position encoding on MATN, we conduct experiments with various experimental settings on the UTD-MHAD dataset under the 50-50 subject evaluation protocol, where the inertial and skeleton data are used. The results are shown in Table 6, indicating that adding spatial position encoding can benefit MATN the most.

A common operation is to add position encoding on the temporal sequence, which is considered to be helpful for improving performance. However, this does not seem to improve the accuracy but negatively biases the prediction. This can be because dividing the data into segments would result in the loss of temporal information; thus, the position encoding cannot contribute much. Moreover, the results show that adding position encoding to the spatial stream can benefit the model a lot, which seems to be uncommon. We believe that as some dimensions are related, for example, each body joint in the skeleton data and each type of inertial data have three dimensions, there are strong connections between each other. As a result, the order information is kept and thus model can extract the salient spatial features and produce better results.

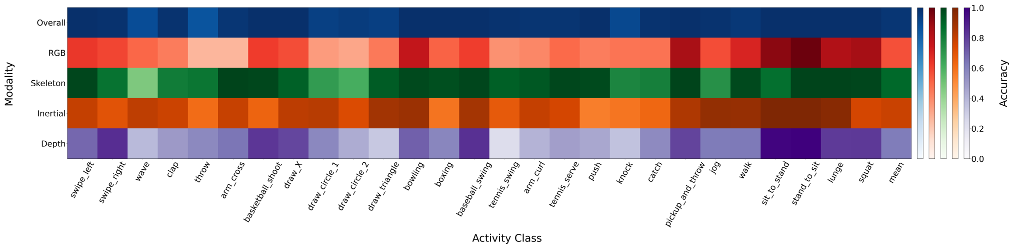


Figure 4: Performance comparison on contribution of modalities on the UTD-MHAD dataset (Top-1 Accuracy)

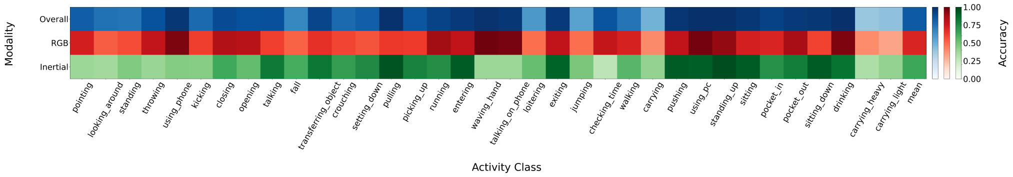


Figure 5: Performance comparison on contribution of modalities on the MMAct dataset (F1 Score)

Table 7: Ablation Study on the representation Learning Layer on the UTD-MHAD dataset

Experimental Setting	Accuracy (%)	Time per Epoch (s)
1-D CNN	91.48	1.761
2-Layer LSTM	75.28	1.771
MATN (Ours)	**92.72**	**1.706**

6.5 Effect and Efficiency of the Representation Learning Layer

We conduct an ablation study to evaluate the effectiveness and efficiency of the unified representation learning layer via comparing its performance with two other deep learning-based encoding approaches: 1-D CNN and LSTM. We use UTD-MHAD dataset and the 50-50 subject evaluation protocol, utilizing the inertial and skeleton data. Results in Table 7 show classification accuracy (Top-1 accuracy) for each variation and the time cost (second) per epoch during the training step.

We compare the proposed module with 1-D CNN and LSTM. 1-D CNN is a well-used approach to perform convolution and feature extraction on time series data. LSTM is suitable for processing data sequences and extracting temporal information. We use the same input data, with the only difference in experimental setting being the feature encoding method. Our experimental results show that MATN outperforms the 1-D CNN module and the LSTM module despite the simplicity of the unified representation learning layer. The representation learning layer also takes less time than 1-D CNN and LSTM. MATN's representation learning layer is effective and efficient when encoding data to a unified representation. This aligns with our motive that the design of a representation learning layer is to preserve spatial and temporal information without complex encoder architecture or extra data engineering.

7 CONCLUSION

Our main objective is to develop an effective and robust novel multimodal human activity recognition method which can be generalized to different varieties of modalities as well as new subjects. We present MATN, a multi-agent attention-based learning approach for multimodal human activity recognition. MATN first encodes the multimodal data through the unified representation learning layer. Then the MSTT module extracts the salient spatial-temporal features of each modality and generates the high-level representation of the input data. Finally, the multi-agent collaboration module aggregates the outputs of each agent and learns to select the informative modalities. We conduct experiments on two public multimodal datasets, UTD-MHAD and MMAct, to evaluate MATN's performance. The experimental results show that the model can extracting the salient spatial-temporal features with multimodal data streams, validating its generalization ability. We plan to advance MATN so that it can better make use of the visual data and perform well in a live human-robot interaction environment.

ACKNOWLEDGMENTS

This work was supported by the Cooperative Research Centres Projects (CRCP) Grants "DeepIoT" project.

REFERENCES

[1] Ijaz Akhter, Tomas Simon, Sohaib Khan, Iain Matthews, and Yaser Sheikh. 2012. Bilinear Spatiotemporal Basis Models. *ACM Trans. Graph.* 31, 2, Article 17 (apr 2012), 12 pages. https://doi.org/10.1145/2159516.2159523

[2] Lei Bai, Lina Yao, Xianzhi Wang, Salil S Kanhere, Bin Guo, and Zhiwen Yu. 2020. Adversarial multi-view networks for activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–22.

[3] Lei Bai, Lina Yao, Xianzhi Wang, Salil S Kanhere, and Yang Xiao. 2020. Prototype similarity learning for activity recognition. *Advances in Knowledge Discovery and Data Mining* 12084 (2020), 649.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[5] Marius Bock, Alexander Hölzemann, Michael Moeller, and Kristof Van Laerhoven. 2021. Improving Deep Learning for HAR with Shallow LSTMs. In *2021 International Symposium on Wearable Computers*. 7–12.

[6] Matthew Brand and Aaron Hertzmann. 2000. Style Machines. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., USA, 183–192. https://doi.org/10.1145/344779.344865

[7] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)* 46, 3 (2014), 1–33.

[8] KG Manosha Chathuramali and Ranga Rodrigo. 2012. Faster human activity recognition with SVM. In *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*. IEEE, 197–203.

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

[10] Kaixuan Chen, Lina Yao, Dalin Zhang, Bin Guo, and Zhiwen Yu. 2019. Multi-agent Attentional Activity Recognition. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 1344–1350. https://doi.org/10.24963/ijcai.2019/186

[11] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2021. Deep Learning for Sensor-Based Human Activity Recognition: Overview, Challenges, and Opportunities. *ACM Comput. Surv.* 54, 4, Article 77 (may 2021), 40 pages. https://doi.org/10.1145/3447744

[12] Ling Chen, Yi Zhang, and Liangying Peng. 2020. METIER: A deep multi-task learning based activity and user recognition model using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–18.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

[15] Sarah Fallmann and Johannes Kropf. 2016. Human activity recognition of continuous data using Hidden Markov Models and the aspect of including discrete data. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*. IEEE, 121–126.

[16] Lin Fan, Zhongmin Wang, and Hai Wang. 2013. Human activity recognition model based on decision tree. In *2013 International Conference on Advanced Cloud and Big Data*. IEEE, 64–68.

[17] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent Network Models for Human Dynamics. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 4346–4354. https://doi.org/10.1109/ICCV.2015.494

[18] Yu Guan and Thomas Plötz. 2017. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–28.

[19] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. 2018. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 786–803.

[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 *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 1112–1123.

[21] Michelle Guo, Edward Chou, De-An Huang, Shuran Song, Serena Yeung, and Li Fei-Fei. 2018. Neural graph matching networks for fewshot 3d action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 653–669.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. *NIPS Deep Learning and Representation Learning Workshop* (2014).

[24] Md Mofijul Islam and Tariq Iqbal. 2020. Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 10285–10292.

[25] Md Mofijul Islam and Tariq Iqbal. 2021. Multi-Gat: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition. *IEEE Robotics and Automation Letters* 6, 2 (2021), 1729–1736.

[26] Wenchao Jiang and Zhaozheng Yin. 2015. Human activity recognition using wearable sensors by deep convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1307–1310.

[27] M Humayun Kabir, M Robiul Hoque, Keshav Thapa, and Sung-Hyun Yang. 2016. Two-layer hidden Markov model for human activity recognition in home environments. *International Journal of Distributed Sensor Networks* 12, 1 (2016), 4560365.

[28] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 7482–7491.

[29] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8658–8667.

[30] Haanvid Lee, Minju Jung, and Jun Tani. 2017. Recognition of visually perceived compositional human actions by multiple spatio-temporal scales recurrent neural networks. *IEEE Transactions on Cognitive and Developmental Systems* 10, 4 (2017), 1058–1069.

[31] Bing Li, Wei Cui, Wei Wang, Le Zhang, Zhenghua Chen, and Min Wu. 2021. Two-Stream Convolution Augmented Transformer for Human Activity Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 286–293.

[32] Shengzhong Liu, Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Huajie Shao, and Tarek Abdelzaher. 2020. Giobalfusion: A global attentional deep learning framework for multisensor information fusion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–27.

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[34] Xiang Long, Chuang Gan, Gerard De Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen. 2018. Multimodal keyless attention fusion for video classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[35] Lingjuan Lyu, Xuanli He, Yee Wei Law, and Marimuthu Palaniswami. 2017. Privacy-preserving collaborative deep learning with application to human activity recognition. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1219–1228.

[36] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2891–2900.

[37] Raphael Memmesheimer, Nick Theisen, and Dietrich Paulus. 2020. Gimme signals: Discriminative signal encoding for multimodal activity recognition. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 10394–10401.

[38] Vishvak S Murahari and Thomas Plötz. 2018. On attention models for human activity recognition. In *Proceedings of the 2018 ACM international symposium on wearable computers*. 100–103.

[39] Pinky Paul and Thomas George. 2015. An effective approach for human activity recognition on smartphone. In *2015 IEEE International Conference on Engineering and Technology (ICETECH)*. IEEE, 1–3.

[40] Vladimir Pavlovic, James M. Rehg, and John MacCormick. 2000. Learning Switching Linear Models of Human Motion. In *Proceedings of the 13th International Conference on Neural Information Processing Systems* (Denver, CO) *(NIPS'00)*. MIT Press, Cambridge, MA, USA, 942–948.

[41] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. 2018. Aroma: A deep multi-task learning based simple and complex human activity recognition method using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–16.

[42] Ilya Sutskever, Geoffrey Hinton, and Graham Taylor. 2008. The Recurrent Temporal Restricted Boltzmann Machine. In *Proceedings of the 21st International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'08)*. Curran Associates Inc., Red Hook, NY, USA, 1601–1608.

[43] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.

[44] Graham W. Taylor, Leonid Sigal, David J. Fleet, and Geoffrey E. Hinton. 2010. Dynamical binary latent variable models for 3D human pose tracking. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 631–638. https://doi.org/10.1109/CVPR.2010.5540157

[45] Raquel Urtasun, David J. Fleet, Andreas Geiger, Jovan Popović, Trevor J. Darrell, and Neil D. Lawrence. 2008. Topologically-Constrained Latent Variable Models. In *Proceedings of the 25th International Conference on Machine Learning* (Helsinki, Finland) *(ICML '08)*. Association for Computing Machinery, New York, NY, USA, 1080–1087. https://doi.org/10.1145/1390156.1390292

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[47] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. 2008. Gaussian Process Dynamical Models for Human Motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 2 (2008), 283–298. https://doi.org/10.1109/TPAMI.2007.1167

[48] Xuanhan Wang, Lianli Gao, Jingkuan Song, and Hengtao Shen. 2016. Beyond frame-level CNN: saliency-aware 3-D CNN with LSTM for video action recognition. *IEEE Signal Processing Letters* 24, 4 (2016), 510–514.

[49] Xiaomin Wang, Junsan Zhang, Leiquan Wang, Philip S Yu, Jie Zhu, and Haisheng Li. 2019. Video-level Multi-model Fusion for Action Recognition. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 159–168.

[50] Hongfei Xue, Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shiyang Wang, Ye Yuan, Shuochao Yao, Aidong Zhang, and Lu Su. 2020. DeepMV: Multi-view deep learning for device-free human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–26.

[51] Jian Bo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. In *Proceedings of the Twenty-Fourth International Conference on Artificial Intelligence* (Buenos Aires, Argentina) *(IJCAI'15)*. AAAI Press, 3995–4001.

[52] Lina Yao, Feiping Nie, Quan Z Sheng, Tao Gu, Xue Li, and Sen Wang. 2016. Learning from less for better: semi-supervised activity recognition via shared structure discovery. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 13–24.

[53] Ming Zeng, Haoxiang Gao, Tong Yu, Ole J Mengshoel, Helge Langseth, Ian Lane, and Xiaobing Liu. 2018. Understanding and improving recurrent networks for human activity recognition by continuous attention. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. 56–63.

[54] Dalin Zhang, Lina Yao, Kaixuan Chen, and Sen Wang. 2018. Ready for use: subject-independent movement intention recognition via a convolutional attention model. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1763–1766.