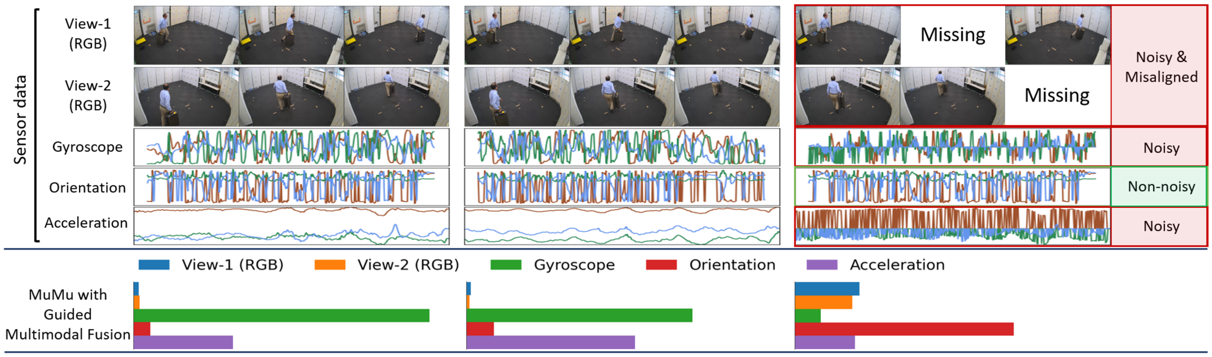
MuMu: Cooperative Multitask Learning-based Guided Multimodal Fusion

Md Mofijul Islam, Tariq Iqbal

School of Engineering and Applied Science, University of Virginia
{mi8uu,tiqbal}@virginia.edu



(a) Carry-Light (Non-noisy data) (b) Carry-Heavy (Non-noisy data) (c) Carry-Heavy (Noisy data, except Orientation sensor)

Figure 1: (a) Carry-Light and (b) Carry-Heavy activities have similar visual features. (a & b) However, these activities have distinct gyroscope and acceleration data. (a & b: bottom-row) Our proposed method, MuMu, utilizes a guided multimodal fusion approach to appropriately prioritize salient modalities (Gyroscope and Acceleration, in this case) while extracting multimodal representations. (c) MuMu can adaptively adjust attention weights when data is noisy. For example, MuMu pays more attention to the non-noisy data (Orientation) than the noisy data (Gyroscope and Acceleration) or misaligned data (View-1 & 2). (Data samples are drawn from MMAct dataset (Kong et al. 2019)).

Abstract

Multimodal sensors (visual, non-visual, and wearable) can provide complementary information to develop robust perception systems for recognizing activities accurately. However, it is challenging to extract robust multimodal representations due to the heterogeneous characteristics of data from multimodal sensors and disparate human activities, especially in the presence of noisy and misaligned sensor data. In this work, we propose a cooperative multitask learning-based guided multimodal fusion approach, MuMu, to extract robust multimodal representations for human activity recognition (HAR). MuMu employs an auxiliary task learning approach to extract features specific to each set of activities with shared characteristics (activity-group). MuMu then utilizes activity-group-specific features to direct our proposed Guided Multimodal Fusion Approach (GM-Fusion) for extracting complementary multimodal representations, designed as the target task. We evaluated MuMu by comparing its performance to state-of-the-art multimodal HAR approaches on three activity datasets. Our extensive experimental results suggest that MuMu outperforms all the evaluated approaches across all three datasets. Additionally, the ablation study suggests that MuMu significantly outperforms the baseline models ($p < 0.05$), which do not use our guided multimodal fusion. Finally, the robust performance of MuMu on noisy and misaligned sensor data posits that our approach is suitable for HAR in real-world settings.

1 Introduction

Understanding human activity ensures effective human-autonomous-system collaboration in various settings, from autonomous vehicles to assistive living to manufacturing (Sabokrou et al. 2019; Iqbal and Riek 2017; Yasar and Iqbal 2021; Iqbal et al. 2019). For example, accurate activity recognition could aid collaborative robots in assisting a worker by bringing tools or autonomous vehicles in requesting to take over the controls from a distracted driver to ensure safety (Kubota et al. 2019; Pakdamanian et al. 2020).

Human activity recognition (HAR) has been extensively studied by utilizing unimodal sensor data, such as visual (Ryoo et al. 2017; Zhang and Parker 2011; Fan et al. 2018), skeleton (Arzani et al. 2017; Ke et al. 2017; Yan, Xiong, and Lin 2018; Iqbal, Rack, and Riek 2016), and wearable sensors (Frank, Kubota, and Riek 2019; Batzianoulis et al. 2017). However, unimodal HAR methods struggle to recognize activity in various real-world scenarios for multiple reasons. First, distinct activities can be mistakenly classified as the same when relying on visual sensors (Kong et al. 2019). For example, the activities related to carrying a light and a heavy object look similar from visual modalities; however, they have distinct physical sensor data (i.e., Gyroscope & Acceleration) (Fig. 1-a & b). Second, HAR algorithms relying on unimodal sensor data may fail to recognize activities when the sensor data is noisy (Fig. 1-c). Thus, in these cases, using multiple modalities can compensate for the weaknesses of any particular modality in recognizing an activity.

Several multimodal learning approaches have been proposed to accurately recognize human activities by fusing data from multiple sensors, such as visual, motion capture, and wearable sensors (Feichtenhofer et al. 2019; Kong et al. 2019; Roitberg et al. 2015; Joze et al. 2020; Liu et al. 2019; Perez-Rua et al. 2019; Hasan et al. 2019; Islam and Iqbal 2020). Although these approaches work adequately in many scenarios, some crucial challenges remain in achieving robust recognition performance, particularly when data from multiple sensors are missing or misaligned.

First, disparate activity-groups require different modalities to accurately recognize activities (an activity-group consists of a set of activities, that exhibit similar characteristics) (Kubota et al. 2019). For example, Kubota et al. (2019) found that data from the motion capture system helps to recognize gross-motion activities involving arm and leg movements (e.g., walking). Moreover, they found that data from wearable sensors helps to recognize fine-grained motion activities involving hand or finger movements (e.g., grasping). Therefore, if a learning approach can exploit the characteristics of activity-groups while extracting the multimodal representations, then that approach can extract robust representation to improve HAR performance. Moreover, in many existing datasets, activities are grouped into major categories based on shared characteristics (Chen, Jafari, and Kehtarnavaz 2015; Kubota et al. 2019; Kong et al. 2019; Awad et al. 2018). For example, Kong et al. (2019) grouped daily human activities into three groups: complex (e.g., carrying, talking), simple (e.g., kicking, jumping), and desk (e.g., using PC). Surprisingly, apart from grouping the activities, these labels of auxiliary activity-groups have not been utilized in extracting multimodal representations.

Second, most existing multimodal learning approaches assume non-noisy and time-aligned multimodal sensor data during training and testing phases. These assumptions limit the applicability of the existing multimodal learning approaches in real-world settings, as the presence of misaligned and noisy sensor data is not uncommon due to occlusion and sensor noises (Fig. 1-c). Thus, we need to develop and evaluate the multimodal learning approaches in the presence of noisy and misaligned sensor data to ensure their applicability in real-world settings.

To address the aforementioned challenges, we propose a novel Cooperative Multitask Learning-based Guided Multimodal Fusion Approach (MuMu) for HAR. In MuMu, we have designed a multitask learning approach that involves learning two cooperative tasks: an auxiliary and a target task. First, MuMu extracts activity-group-specific features for activity-group recognition (auxiliary task). Second, the activity-group-specific features direct our Guided Multimodal Fusion Approach (GM-Fusion) to extract robust multimodal representations for recognizing activities (target task). Here, both tasks work cooperatively, where the auxiliary task guides the target task to extract complementary multimodal representations appropriately.

We compared the performance of MuMu to several state-of-the-art HAR algorithms on three multimodal activity datasets (MMAct (Kong et al. 2019), UTD-MHAD (Chen, Jafari, and Kehtarnavaz 2015) and UCSD-MIT (Kubota et al. 2019)). The results from our extensive experimental evaluations suggest that MuMu outperforms all the state-of-the-art approaches in all evaluation conditions. MuMu achieved an improvement of 4.45% and 3.61% (F1-score) on the MMAct dataset for the cross-subject and cross-session evaluation conditions, compared to the state-of-the-art approaches, respectively. Additionally, MuMu achieved an improvement of 6.86% and 2.48% (top-1 accuracy) on the UCSD-MIT and the UTD-MHAD datasets for leave-one-subject-out evaluation settings, compared to the state-of-the-art approaches, respectively. Furthermore, our qualitative analysis of multimodal attention weights suggests that our proposed guided multimodal fusion approach can appropriately prioritize the modalities while extracting complementary representations, even in the presence of noisy and misaligned sensor data (Fig. 1 & 4). Moreover, our extensive ablation study suggests that our proposed approach significantly outperforms the baseline multimodal learning approaches ($p < 0.05$), which do not use guided fusion.

2 Related Work

Multimodal Learning: Several multimodal learning approaches have been developed for various tasks (Guo, Wang, and Wang 2019; Roitberg et al. 2015), such as video classification (Feichtenhofer et al. 2019; Xiao et al. 2020), human activity recognition (Islam and Iqbal 2021; Long et al. 2018; Joze et al. 2020), and visual question answering (Lu et al. 2019; Li et al. 2019). Some of these approaches have been designed to extract representations from data of similar types of modalities (Feichtenhofer, Pinz, and Wildes 2016, 2017; Zhang et al. 2018). For example, Simonyan and Zisserman (2014) designed a two-stream CNN-based model to extract spatial and temporal features from the visual modalities. Similarly, Feichtenhofer et al. (2019) proposed a two-stream learning model to extract spatial-temporal features by varying the data sampling rate in those streams.

Other approaches have focused on extracting representations from heterogeneous modalities (Kubota et al. 2019; Kong et al. 2019; Islam and Iqbal 2020; Joze et al. 2020; Perez-Rua et al. 2019; Münzner et al. 2017; Liu et al. 2019). For example, Long et al. (2018) designed a self-attention approach to extract unimodal features from different modalities, which were then concatenated to produce multimodal representations. Several approaches have been proposed to fuse the representations at the intermediary layers of the model (Feichtenhofer et al. 2019; Xiao et al. 2020; Joze et al. 2020). For instance, Xiao et al. (2020) used a multi-stream CNN-based model to fuse representations at the intermediate layers. However, these approaches depend on human experts to determine which layers' representations should be fused. These manual fusion approaches often introduce bias in the model and produce suboptimal representations.

Multitask Learning: Several multitask learning approaches have been designed to learn various tasks by sharing their learned knowledge to improve these tasks performance (Ruder 2017; Hashimoto et al. 2016; Zhang and Yang 2017; Guo et al. 2018; Vandenhende et al. 2020; Gagné 2019; Zhou et al. 2020a). For example, Standley et al. (2020) proposed a multitask learning framework where tasks are grouped and learned by exploiting the cooperative and competitive relationships among the tasks. Similarly, Guo, Lee, and Ulbricht (2020) utilized a tree-structured design space and Gumbel-softmax (Jang, Gu, and Poole 2016; Maddison, Mnih, and Teh 2016) to determine which parts of the network can be shared or branched to maximize the parameters sharing and the tasks performance. Generally, one of the primary goals of the existing multitask learning approaches is to maximize the sharing of learning parameters or knowledge among the heterogeneous tasks (Crawshaw 2020; Søgaard and Goldberg 2016; Ruder 2017).

Additionally, multitask models have been used to learn shared representations (Ruder 2017; Xu et al. 2018; Zhou et al. 2020b; Achille et al. 2019; Zamir et al. 2018). For example, Liu, Johns, and Davison (2019) proposed a multitask attention model to extract a shared feature for learning task-specific representations. Moreover, Sun et al. (2020) designed an algorithm to learn feature sharing patterns across tasks for maximizing shared representations. The overall goal of these approaches is to compress a multitask learning model by maximizing the shared representations among the competitive heterogeneous tasks. In this work, we have designed a cooperative multitask learning approach, where the auxiliary task guides the target task to extract multimodal representations to recognize activities accurately.

3 MuMu: Multitask Learning-based Guided Multimodal Fusion Approach

3.1 Problem Formulation

We define a cooperative multitask learning problem, which involves learning the auxiliary and the target tasks cooperatively for multimodal fusion. Similar to the multi-class activity recognition problem, we aim to recognize a set of K activities, $A = (A_1, \dots, A_K)$, by extracting multimodal representations (X^c) from M heterogeneous modalities, $X^r = (X^r_1, \dots, X^r_M)$ (r stands for raw feature). We have termed this activity recognition ($A_i \in A$) as the *target task*.

Activity datasets defined activity-group in various ways. For example, UCSD-MIT uses human motion to define activity-group (gross & fine), whereas the MMAct dataset uses the complexity of the activities (complex, simple & desk). As different activity-groups share disparate characteristics, they require different modalities for recognizing activities (Kubota et al. 2019). Thus, we divide the activity set A into N activity-groups (G), where $G = (G_1, \dots, G_N)$. Here, each activity-group (G_i), consists of J_i unique activities that share similar characteristics, where $G_i = (A^i_1, \dots, A^i_{J_i})$, and $A^i_j \in A$. We have termed the activity-group recognition ($G_i \in G$) as the *auxiliary task*.

3.2 Approach Overview

Our proposed Cooperative Multitask Learning-based Guided Multimodal Fusion Approach (MuMu) consists of three learning modules (Fig. 2):

- **Unimodal Feature Encoder (UFE)** encodes modality-specific spatial-temporal features.
- **Auxiliary Task Learning (ATL)** Module extracts activity-group-specific multimodal representations.
- **Target Task Learning (TTL)** Module utilizes the activity-group-specific features from the auxiliary task as prior information to appropriately fuse and extract multimodal representations for activity recognition.

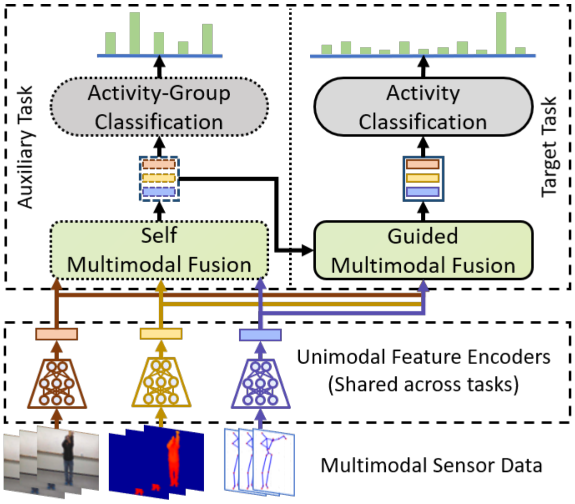


Figure 2: MuMu: Cooperative Multitask Learning-based Guided Multimodal Fusion Approach. The Unimodal Feature Encoder encodes unimodal spatial-temporal features. The Auxiliary Task module fuses the unimodal features to extract the activity-group-specific features. The activity-group features guide the Target Task module to fuse and extract complementary multimodal representations by employing a Guided Multimodal Fusion Approach. We have designed a multitask learning loss for end-to-end training.

3.3 UFE: Unimodal Feature Encoder

We have adopted the Unimodal Feature Encoder (UFE) architecture from the work by Islam and Iqbal (2020). In our implementation, UFE independently encodes salient unimodal features of each modality $m \in M$ in four steps. First, UFE segments the raw unimodal features and produces $X^r_m = (x^r_{m,1}, x^r_{m,2}, \dots, x^r_{m,S_m}) \in \mathbb{R}^{B \times S_m \times D^r_m}$, where B is the batch size, S_m is the segment size, and D^r_m is the raw feature dimension of the modality m. Second, UFE encodes the spatial features of each segment of modality $m \in M$. Third, UFE utilizes an LSTM, a variant of recurrent neural network, to encode unimodal spatial-temporal features. Fourth, a self-attention approach has been employed to extract salient unimodal features, $X^u = (x^u_1, x^u_2, \dots, , x^u_M) \in \mathbb{R}^{B \times M \times D^u}$, from the extracted spatial-temporal features (D^u is the unimodal (u) feature embedding size). Instead of utilizing a resource intensive multi-head self-attention approach (Vaswani et al. 2017), which was used by Islam and Iqbal (2020), in this work, we have adopted a lightweight self-attention model from Long et al. (2018). MuMu uses the unimodal features, X^u, in the subsequent learning modules to produce robust multimodal representations.

3.4 ATL: Auxiliary Task Learning Module

In the auxiliary task learning step, MuMu fuses the unimodal features to extract activity-group-specific multimodal representation for classifying the activity-groups in two steps:

Self Multimodal Fusion Approach (SM-Fusion): We have designed a Self Multimodal Fusion Approach (SM-

Fusion) for extracting activity-group-specific salient features. SM-Fusion assigns attention weight (α_m) to each modality for fusing unimodal features, X^u, and extracting multimodal auxiliary representation, X^{aux}. The attention weight, α_m, is calculated in the following way,

$$\gamma_m = (W^{aux})^T X^u_m \tag{1}$$

$$\alpha_m = \frac{exp(\gamma_m)}{\sum\limits_{m \in M} exp(\gamma_m)} \tag{2}$$

Here, W^{aux} is a learnable parameter. We have utilized a 1D-CNN with a filter size of 1 to calculate α_m. Finally, this weight is used to fuse the unimodal features and extract multimodal auxiliary representation, X^{aux}:

$$X^{aux} = \sum_{m \in M} \alpha_m X^u_m \tag{3}$$

Activity-Group Classification: The auxiliary representation, X^{aux}, is passed through a auxiliary task learning network, F^{aux}, to classify the activity-group:

$$y^{aux} = F^{aux}(X^{aux}) \tag{4}$$

3.5 TTL: Target Task Learning Module

In MuMu, we have designed a target task to extract multimodal representations and classify activities in two steps. First, MuMu uses activity-group features from the auxiliary task to direct our proposed Guided Multimodal Fusion Approach (GM-Fusion) to extract multimodal representations. Because activity-group features can help to prioritize the salient modalities to extract multimodal representations appropriately. Second, MuMu uses fused representations to classify the activities. In MuMu, the auxiliary and the target tasks work cooperatively to extract complementary multimodal representations for recognizing activities accurately.

Guided Multimodal Fusion Approach (GM-Fusion): GM-Fusion uses the extracted activity-group-specific features from auxiliary task as prior information, X^{aux}, to extract multimodal representations for activity recognition.

First, GM-Fusion projects the extracted unimodal features, X^u, to produce unimodal key (K^u) and value (V^u) feature vectors in the following way:

$$K^u = X^u W^K; V^u = X^u W^V \tag{5}$$

Here, W^K and W^V are learnable parameters. These unimodal key and value vectors are used to extract the multimodal representation. Second, GM-Fusion projects multimodal auxiliary representation, X^{aux}, to produce auxiliary query feature vector (Q^{aux}).

$$Q^{aux} = X^{aux} W^Q \tag{6}$$

Here, W^Q is a learnable parameter. This auxiliary query feature vector (Q^{aux}) is used as a prior to extract complementary multimodal representation, X^c, by utilizing the unimodal key (K^u) and value (V^u) feature vectors:

$$X^{c^{'}} = \sigma \left(\frac{Q^{aux} {K^u}^T}{\sqrt{D^u}} \right) V^u \tag{7}$$

$$X^c = W^o X^{c^{'}} \tag{8}$$

Here, W^o is a learnable projection parameter.

Activity Classification: Multimodal representation, X^c, is concatenated with activity-group-specific features, X^{aux}, for activity classification. X^c is passed through a target task learning network, F^t, to classify the activities:

$$X^f = W^f [X^c; X^{aux}] \tag{9}$$

$$y^t = F^t(X^f) \tag{10}$$

Here, W^f is a learnable projection parameter.

3.6 Multitask Learning Loss

We have designed a multitask learning loss for end-to-end training of MuMu. This loss is used to train the auxiliary and the target tasks jointly. First, we use cross-entropy auxiliary loss, L^{aux}, to train the auxiliary task for activity-group classification. L^{aux} enforces the auxiliary task branch to learn the activity-group-specific multimodal representations.

$$L^{aux}(y^{aux}, \hat{y}^{aux}) = \frac{1}{B} \sum_{i=1}^{B} y^{aux}_i \log \hat{y}^{aux}_i \tag{11}$$

Second, we calculate the cross-entropy loss, L^t, to train the target task for activity classification. This loss ensures that the target task learns the robust multimodal representations for activity recognition.

$$L^t(y^t, \hat{y}^t) = \frac{1}{B} \sum_{i=1}^{B} y^t_i \log \hat{y}^t_i \tag{12}$$

Finally, the auxiliary and target task losses are combined for end-to-end training of MuMu:

$$loss = L^t(y^t, \hat{y}^t) + \beta^{aux} L^{aux}(y^{aux}, \hat{y}^{aux}) \tag{13}$$

Here, β^{aux} is the weight of auxiliary task learning loss.

4 Experimental Setup

4.1 Datasets

We evaluated the performance of our proposed approach, MuMu, by applying it on three multimodal activity datasets: UCSD-MIT (Kubota et al. 2019), UTD-MHD (Chen, Jafari, and Kehtarnavaz 2015) and MMAct (Kong et al. 2019). MMAct dataset contains 37 activities which are categorized into 3 groups: 16 complex (e.g., carrying), 12 simple (e.g., kicking), 9 desk(e.g., using PCs). UCSD-MIT dataset contains nine automotive and block assembly activities from 2 groups: 4 gross-motion (e.g., attaching part), and 5 fine-motion (e.g., palmar grab). UTD-MHAD contains 27 activities which are categorized into 4 groups: 9 hand gesture (e.g., draw circle), 9 sports (e.g., bowling), 5 daily (e.g., door knock), and 4 training exercises (e.g., squat). Please check the supplementary materials for more details.

4.2 Learning Architecture Implementation

We segmented the data from visual modalities (RGB and depth) with a window size of 1 and a stride of 3. For the data from other sensor modalities, we used a window size of 5 and a stride of 5. To encode segmented spatial features, we used ResNet-50 model (He et al. 2016) for data from

Table 1: Cross-subject performance comparison (F1-Score) of multimodal learning methods on MMAct dataset

Method	F1-Score (%)
SMD (Hinton, Vinyals, and Dean 2015)	63.89
Student (Kong et al. 2019)	64.44
Multi-Teachers (Kong et al. 2019)	62.67
MMD (Kong et al. 2019)	64.33
MMAD (Kong et al. 2019)	66.45
HAMLET (Islam and Iqbal 2020)	69.35
Keyless (Long et al. 2018)	71.83
MuMu (Our method)	**76.28**

Table 2: Cross-session performance comparison (F1-Score) of multimodal learning methods on MMAct dataset

Method	F1-Score (%)
SVM+HOG (Ofli et al. 2013)	46.52
TSN (RGB) (Wang et al. 2016)	69.20
TSN (Optical-Flow) (Wang et al. 2016)	72.57
MMAD (Kong et al. 2019)	74.58
TSN (Fusion) (Wang et al. 2016)	77.09
MMAD (Fusion) (Kong et al. 2019)	78.82
Keyless (Long et al. 2018)	81.11
HAMLET (Islam and Iqbal 2020)	83.89
MuMu (Our method)	**87.50**

visual modalities (RGB and depth) and Co-occurrence approach (Li et al. 2018) for data from other sensors modalities (sEMG, Acceleration, Gyroscope, and Orientation). The unimodal feature of each modality is encoded to 128 sized feature embedding. We used two fully connected layers with Re-LU activation after the first layer for activity-group classification in auxiliary task learning. We used similar task learning architecture for the activity classification in target task learning. For more implementation and training procedure details, please check the supplementary materials.

5 Results and Discussion

5.1 Comparison with Multimodal Approaches

Results: We evaluated MuMu's performance by comparing it against the state-of-the-art HAR approaches on three datasets: MMAct, UTD-MHAD, and UCSD-MIT. For MMAct dataset, we followed originally proposed cross-subject and cross-session evaluation settings and reported F1-scores in Tables 1 & 2, respectively. The results suggest that MuMu outperforms state-of-the-art approaches on both cross-subject and cross-session evaluation settings with improvements of 4.45% and 3.61% in F1-score, respectively. For UTD-MHAD and UCSD-MIT datasets, we followed leave-one-subject-out cross-validation and reported top-1 accuracies in Tables 4 & 3, respectively. The results suggest that MuMu outperforms the best performing baselines with improvements of 6.86% and 2.48% in top-1 accuracy on UCSD-MIT and UTD-MHAD datasets, respectively.

Discussion: The experimental results on these activity datasets (Tables 1, 2, 4 & 3) suggest that MuMu outperforms all the state-of-the-art approaches in all evaluation conditions. Moreover, the results indicate that attention-based HAR methods (i.e., MuMu, Keyless (Long et al. 2018) and HAMLET (Islam and Iqbal 2020)) outperform Non-Attention-based methods (i.e., PoseMap (Liu and Yuan 2018) and TSN (Wang et al. 2016)). Unlike MuMu, the other attention-based methods do not consider the activity-group-specific information to extract multimodal representations. In our implementation, MuMu utilizes the activity-group-specific information to extract complementary multimodal representations by utilizing our proposed Guided Multimodal Fusion approach (GM-Fusion). GM-Fusion allows the prioritization of different modalities based on the activity-group information extracted by the auxiliary task learning module. Thus, the experimental results posit that incorporating activity-group information allows the extraction of complementary multimodal representations effectively to improve the HAR accuracy.

Table 3: Performance comparison (F1-Score) of multimodal learning methods on UCSD-MIT dataset (Kubota et al. 2019).

Learning Methods	Merge Types	F1-Score (%)
Non-Attention	SUM	52.35
	CONCAT	50.92
HAMLET (Islam and Iqbal 2020)	SUM	50.04
	CONCAT	48.26
Keyless (Long et al. 2018)	SUM	51.68
	CONCAT	54.48
MuMu (Our method)	-	**61.34**

Table 4: Performance comparison (top-1 accuracy) of multimodal learning methods on UTD-MHAD dataset.

Method	Accuracy (%)
MHAD (Chen, Jafari, and Kehtarnavaz 2015)	79.10
SOS (Hou et al. 2016)	86.97
S^{2}DDI (Wang et al. 2017)	89.04
DCNN (Imran and Kumar 2016)	91.20
Keyless (Long et al. 2018)	92.67
MCRL (Liu, Kong, and Jiang 2019)	93.02
PoseMap (Liu and Yuan 2018)	94.51
HAMLET (Islam and Iqbal 2020)	95.12
MuMu (Our method)	**97.60**

Although state-of-the-art multimodal HAR approaches show comparatively better performance on cross-session evaluation settings (Tables 2 & 4), the performance degrades on challenging cross-subject evaluation conditions for all evaluated baselines (Tables 1 & 3). The performance degrades because MMAct and UCSD-MIT datasets contain data samples that enforce the utilization of the wearable sensors to recognize activities accurately, where the wearable sensor data vary considerably across subjects (see Fig. 1). To address this challenge, MuMu utilizes activity-group features to guide GM-Fusion to extract salient multimodal representations for recognizing activities accurately. On the other hand, state-of-the-art approaches fused unimodal features without considering activity-group information. Additionally, in the cross-subject evaluation conditions, MuMu outperforms the F1-score of state-of-the-art approaches on MMAct and UCSD-MIT datasets with an improvement of 4.45% and 6.86%, respectively. These performance improvements indicate that MuMu can generate robust multimodal representation by prioritizing the salient modalities than other approaches.

Table 5: Performance comparison (Accuracy %) of the impact of modality changes on UTD-MHAD dataset. R: RGB, D: Depth, S: Skeleton, P: Physical Sensors.

Learning	Modality Combinations		
Methods	R+S	R+S+P	R+D+S+P
Keyless (Long et al. 2018)	90.20	92.67	83.87
HAMLET (Islam and Iqbal 2020)	95.12	91.16	90.09
MuMu	**96.10**	**97.44**	**97.60**

5.2 Impact of Supplementary Modalities

To investigate whether additional modalities help to improve the performance of learning models, we evaluated the performance of MuMu and two baseline approaches (Keyless (Long et al. 2018)) and HAMLET (Islam and Iqbal 2020)) with various combinations of modalities. We conducted this study on the UTD-MHAD dataset with RGB, Depth, Skeleton, Physical sensors modalities. The experimental results suggest that MuMu outperformed the evaluated baselines on all the combinations of modalities tested (see Table 5).

Results & Discussion: In Table 5, the results suggest that incorporating additional modalities helps MuMu to improve the HAR accuracy. However, additional modalities do not always improve the performance of two baselines. For example, incorporating the depth modality degrades the accuracy of the baseline methods, whereas the HAR accuracy of MuMu improves slightly with this additional modality.

The performance of the baselines degrades, as additional modalities may not provide salient information to recognize a set of activities accurately. For example, visual modality may not provide salient information for gesture recognition (e.g., wave, swipe), whereas physical sensors can help recognize those activities accurately. The baseline methods either concatenated or used a self-attention approach to fuse unimodal features without considering the characteristics of activity-group, which results in performance degradation with supplementary modalities. However, MuMu uses activity-group information from the auxiliary task to guide the target task for prioritizing and fusing the additional modalities to extract complementary multimodal representations for recognizing activities accurately. Therefore, it is essential to prioritize the salient modalities for extracting robust representation to recognize activities accurately.

5.3 Impact of Noisy Modalities

We conducted both quantitative and qualitative experiments to evaluate the performance of MuMu and three baselines (Non-Attention, HAMLET, and Keyless) in the presence of noisy and misaligned sensor data. We developed the Non-Attention method for evaluation purposes, where we extract unimodal features using CNN+LSTM model without using an attention mechanism. The extracted unimodal features are concatenated to classify activities.

We conducted this study in cross-subject evaluation setting on MMAct dataset with two visual modalities (RGB View 1 & 2) and three non-visual modalities (Gyroscope, Orientation & Acceleration). We randomly selected either visual or non-visual modalities with 50% probability and then dropped raw features to introduce noise. The quantitative and qualitative experimental results are presented in Table 6 and Fig 4, respectively.

Table 6: Performance comparison (F1-Score %) of the impact of noisy data on MMAct dataset. Visual: RGB (View 1 & 2), Non-visual: Gyroscope, Orientation & Acceleration.

Learning	No Noisy	Noisy Modalities	
Methods	Modality	Visual	Non-Visual
Non-Attention	68.29	66.30	66.02
HAMLET (Islam and Iqbal 2020)	69.35	64.10	67.57
Keyless (Long et al. 2018)	71.83	67.94	68.29
MuMu	**76.28**	**74.22**	**73.78**

Results & Discussion: The experimental results suggest that MuMu outperforms the evaluated baselines in the presence of noisy data (Table 6). In MuMu, our proposed Guided Multimodal Fusion Approach (GM-Fusion) appropriately prioritizes the modalities and extracts the robust multimodal representation from noisy sensor data for accurate activity recognition. However, the baseline multimodal learning approaches either use Non-Attention or self-attention based multimodal fusion, which may not effectively extract complementary multimodal representations.

Additionally, the qualitative results of multimodal attention visualization (Fig. 4-Bottom row) indicate the same phenomenon that MuMu can prioritize the salient modalities to extract complementary multimodal representations from noisy and misaligned sensor data. For example, although the gyroscope and acceleration data provide distinctive features for carry-heavy activity, MuMu adjusts the multimodal attention weights when we introduce noise in those modalities (Fig. 4-Bottom row), by paying more attention to the non-noisy modality (Orientation) and less attention to noisy modalities (Gyroscope and Acceleration), which contribute to better HAR performance on noisy data (Table 6). In Fig. 4-Center row, it can be observed that HAMLET, which uses a self-attention based fusion approach, increased the attention weight to the noisy sensor data (i.e., Acceleration in Fig 4-Right) compared to the attention weight assigned on the non-noisy data samples (Fig 4-Left). These qualitative results indicate that self-attention based fusion may not appropriately prioritize the noisy sensor data to extract robust multimodal representations (Fig. 4-Center row), which also reflects in the quantitative results in Table 6.

5.4 Ablation Study and Significance Analysis

To investigate the importance of various modules of MuMu, we developed three single-task-based baseline models by removing the auxiliary task learning branch in MuMu (Fig. 2). The Non-Attention model (B1) does not employ any attention approach in extracting unimodal or fusing multimodal features. The Unimodal Attention model (B2) employs an attention approach to extract unimodal features and concatenate multimodal features (similar to Keyless (Long et al. 2018)). The Unimodal + Multimodal Attention model (B3) uses an attention approach to extract unimodal and fuse multimodal features (similar to HAMLET (Islam and Iqbal 2020)). We trained and tested all these baselines and MuMu five times with different initialization of the learning parameters. Additionally, we conducted the significance analysis at level $\alpha = 0.05$ by following the procedure proposed by

Table 7: Ablation study of MuMu components on MMAct Dataset.

Model Type	Learning Models	Average F1-Score (%)	Standard Deviation	Significant Over [§]
Single Task	B1	68.48	1.26	None
	B2 [†]	70.52	0.98	B1 & B3
	B3 [†]	69.19	0.72	B1
Multitask	**MuMu** [*]	**75.97**	**0.29**	B1, B2 & B3

B1: Non-Attention, B2: Unimodal Attention, B3: Uni + Multimodal Attention

[†] Self-Attention based Multimodal Fusion, [*] Guided Multimodal Fusion

[§] We conduct the significance analysis at $\alpha = 0.05$ (Following Dror et al. (2019))

Dror, Shlomov, and Reichart (2019). We conducted this experimental analysis on MMAct dataset in cross-subject evaluation setting. We also included additional ablation studies in the supplementary document.

Results and Discussion: The experimental results in Table 7 suggest that the baseline B3, which uses an attention approach to prioritize the modalities, fails to outperform B2 significantly. Here, B2 uses the attention approach only to extract unimodal and concatenate the multimodal features. These results indicate that how a multimodal learning approach fuses the information is crucial in improving the HAR performance.

Moreover, the experimental results in Table 7 indicate that MuMu significantly outperforms all the baseline models and improves the HAR accuracy. The primary difference between MuMu and the baseline models is that MuMu uses activity-group features to guide the target task for extracting multimodal representations. Thus, this experimental analysis indicates that MuMu, with the help of our guided multimodal fusion approach, can appropriately fuse multimodal features to improve the HAR accuracy significantly.

5.5 Qualitative Analysis

We conducted two qualitative analyses to evaluate the effectiveness of our guided multimodal fusion approach. First, we visualized the attention weights to evaluate whether MuMu can prioritize the salient modalities (Fig. 1 & 4). Second, we visualized t-SNE embeddings of unimodal and multimodal representations obtained using MuMu (Fig. 3-Right) and HAMLET with self-attention based fusion (Islam and Iqbal 2020) (Fig. 3-Left). We conducted these studies on the MMAct dataset in cross-subject evaluation setting.

Attention Visualization: Our experimental analysis (Fig. 1 & 4) suggests that appropriately prioritizing the relevant modalities aids in improved HAR performance. The results in Fig. 1-a & b indicate that MuMu can appropriately prioritize the salient modalities (Gyroscope and Acceleration) in extracting complementary representations to distinguish visually similar activities (i.e., carry-light and carry-heavy). Additionally, when the data from these modalities are noisy, MuMu adjusts the attention weights to the non-noisy modalities (i.e., visual and orientation) to extract complementary multimodal representations (Fig. 4). These results indicate that MuMu can adjust attention weights based on the extracted unimodal features to produce complementary representations. On the other hand, the self-attention based multimodal fusion approach can not appropriately prioritize the relevant modalities (Fig. 4), which results in performance degradation (Table 7).

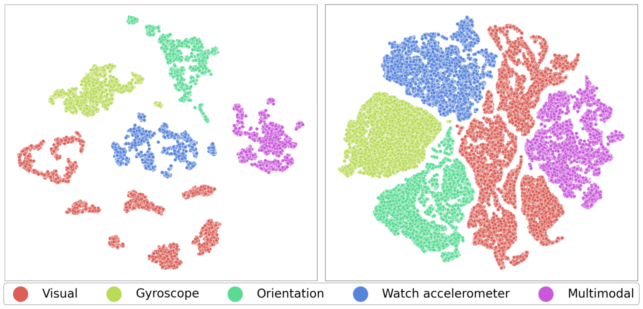
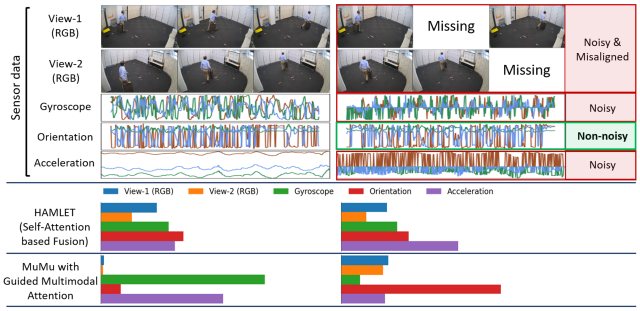


Figure 3: The t-SNE visualization of unimodal and multimodal representations. (Left) HAMLET with Self-Attention based Fusion, (Right) MuMu with Guided Multimodal Fusion.



Carry-Heavy Activity: (left) Non-noisy data (right) Noisy data (except Orientation)

Figure 4: Comparative impact of guided multimodal attention in MuMu to extract complementary multimodal representations from noisy sensor data (Multimodal attention weights visualization).

Feature Visualization (t-SNE): In Fig. 3, one can observe that the features are sparsely distributed with fractured clusters when obtained from HAMLET, whereas the features are more compact and smoothly distributed when obtained from MuMu. Specifically, for visual modalities, MuMu produces clustered representations, whereas HAMLET produces sparsely distributed representations. This visualization indicates that MuMu can extract non-overlapping distinctive representations, resulting in an improved HAR performance.

6 Conclusion

In this work, we have proposed a cooperative multitask learning-based guided multimodal fusion approach, MuMu. MuMu first extracts activity-group features for activity-group recognition (Auxiliary task). MuMu then utilizes the activity-group features in the Guided Multimodal Fusion (GM-Fusion) module to extract complementary multimodal representations for HAR (Target task). Our extensive experimental results suggest that MuMu outperforms state-of-the-art approaches on three multimodal activity recognition datasets in all evaluation conditions. Additionally, the robust performance on noisy data indicates the applicability of MuMu in real-world settings. Future work will focus on evaluating the performance of MuMu on other multimodal learning tasks, such as human motion prediction, visual-language navigation, and action or video retrieval.

References

Achille, A.; Lam, M.; Tewari, R.; Ravichandran, A.; Maji, S.; Fowlkes, C. C.; Soatto, S.; and Perona, P. 2019. Task2Vec: Task Embedding for Meta-Learning. In *ICCV*.

Arzani, M. M.; Fathy, M.; Aghajan, H.; Azirani, A. A.; Raahemifar, K.; and Adeli, E. 2017. Structured prediction with short/long-range dependencies for human activity recognition from depth skeleton data. In *IROS*.

Awad, G.; Butt, A.; Curtis, K.; Lee, Y.; Fiscus, J.; Godil, A.; Joy, D.; Delgado, A.; Smeaton, A.; Graham, Y.; et al. 2018. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*.

Batzianoulis, I.; El-Khoury, S.; Pirondini, E.; Coscia, M.; Micera, S.; and Billard, A. 2017. EMG-based decoding of grasp gestures in reaching-to-grasping motions. *RAS*.

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

Crawshaw, M. 2020. Multi-Task Learning with Deep Neural Networks: A Survey. *arXiv preprint arXiv:2009.09796*.

Dror, R.; Shlomov, S.; and Reichart, R. 2019. Deep dominance-how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2773–2785.

Fan, L.; Huang, W.; Gan, C.; Ermon, S.; Gong, B.; and Huang, J. 2018. End-to-end learning of motion representation for video understanding. In *CVPR*, 6016–6025.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast Networks for Video Recognition. In *CVPR*.

Feichtenhofer, C.; Pinz, A.; and Wildes, R. P. 2016. Spatiotemporal Residual Networks for Video Action Recognition. In *NeurIPS*.

Feichtenhofer, C.; Pinz, A.; and Wildes, R. P. 2017. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, 4768–4777.

Frank, A. E.; Kubota, A.; and Riek, L. D. 2019. Wearable activity recognition for robust human-robot teaming in safety-critical environments via hybrid neural networks. In *IROS*, 449–454. IEEE.

Gagné, C. 2019. A Principled Approach for Learning Task Similarity in Multitask Learning. In *IJCAI*.

Guo, M.; Haque, A.; Huang, D.-A.; Yeung, S.; and Fei-Fei, L. 2018. Dynamic task prioritization for multitask learning. In *ECCV*, 270–287.

Guo, P.; Lee, C.-Y.; and Ulbricht, D. 2020. Learning to branch for multi-task learning. In *International Conference on Machine Learning*, 3854–3863. PMLR.

Guo, W.; Wang, J.; and Wang, S. 2019. Deep Multimodal Representation Learning: A Survey. *IEEE Access*, 7: 63373–63394.

Hasan, M. K.; Rahman, W.; Bagher Zadeh, A.; Zhong, J.; Tanveer, M. I.; Morency, L.-P.; and Hoque, M. E. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *EMNLP-IJCNLP*, 2046–2056.

Hashimoto, K.; Xiong, C.; Tsuruoka, Y.; and Socher, R. 2016. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *NeurIPS*.

Hou, Y.; Li, Z.; Wang, P.; and Li, W. 2016. Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*.

Imran, J.; and Kumar, P. 2016. Human action recognition using RGB-D sensor and deep convolutional neural networks. In *ICACCI*.

Iqbal, T.; Li, S.; Fourie, C.; Hayes, B.; and Shah, J. A. 2019. Fast Online Segmentation of Activities from Partial Trajectories. In *ICRA*.

Iqbal, T.; Rack, S.; and Riek, L. D. 2016. Movement Coordination in Human-Robot Teams: A Dynamical Systems Approach. *IEEE Transactions on Robotics*, 32(4): 909–919.

Iqbal, T.; and Riek, L. D. 2017. Human Robot Teaming: Approaches from Joint Action and Dynamical Systems. *Humanoid Robotics*.

Islam, M. M.; and Iqbal, T. 2020. HAMLET: A Hierarchical Multimodal Attention-based Human Activity Recognition Algorithm. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10285–10292.

Islam, M. M.; and Iqbal, T. 2021. Multi-GAT: A Graphical Attention-based Hierarchical Multimodal Representation Learning Approach for Human Activity Recognition. In *IEEE Robotics and Automation Letters (RA-L)*.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Joze, H. R. V.; Shaban, A.; Iuzzolino, M. L.; and Koishida, K. 2020. MMTM: Multimodal Transfer Module for CNN Fusion. In *CVPR*.

Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; and Boussaid, F. 2017. A new representation of skeleton sequences for 3d action recognition. In *CVPR*, 3288–3297.

Kong, Q.; Wu, Z.; Deng, Z.; Klinkigt, M.; Tong, B.; and Murakami, T. 2019. MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding. In *ICCV*, 8658–8667.

Kubota, A.; Iqbal, T.; Shah, J. A.; and Riek, L. D. 2019. Activity recognition in manufacturing: The roles of motion capture and sEMG+ inertial wearables in detecting fine vs. gross motion. In *ICRA*.

Li, C.; Zhong, Q.; Xie, D.; and Pu, S. 2018. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *IJCAI*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *NeurIPS*.

Liu, G.; Qian, J.; Wen, F.; Zhu, X.; Ying, R.; and Liu, P. 2019. Action Recognition Based on 3D Skeleton and RGB Frame Fusion. In *IROS*, 258–264.

Liu, M.; and Yuan, J. 2018. Recognizing human actions as the evolution of pose estimation maps. In *CVPR*, 1159–1168.

Liu, S.; Johns, E.; and Davison, A. J. 2019. End-To-End Multi-Task Learning With Attention. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1871–1880.

Liu, T.; Kong, J.; and Jiang, M. 2019. RGB-D Action Recognition Using Multimodal Correlative Representation Learning Model. *IEEE Sensors Journal*, 19(5): 1862–1872.

Long, X.; Gan, C.; De Melo, G.; Liu, X.; Li, Y.; Li, F.; and Wen, S. 2018. Multimodal keyless attention fusion for video classification. In *AAAI*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Münzner, S.; Schmidt, P.; Reiss, A.; Hanselmann, M.; Stiefelhagen, R.; and Dürichen, R. 2017. CNN-Based Sensor Fusion Techniques for Multimodal Human Activity Recognition. In *ACM ISWC*, 158–165.

Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; and Bajcsy, R. 2013. Berkeley mhad: A comprehensive multimodal human action database. In *WACV*, 53–60. IEEE.

Pakdamanian, E.; Sheng, S.; Baee, S.; Heo, S.; Kraus, S.; and Feng, L. 2020. DeepTake: Prediction of Driver Takeover Behavior using Multimodal Data. In *CHI*.

Perez-Rua, J.-M.; Vielzeuf, V.; Pateux, S.; Baccouche, M.; and Jurie, F. 2019. MFAS: Multimodal Fusion Architecture Search. In *CVPR*.

Roitberg, A.; Somani, N.; Perzylo, A.; Rickert, M.; and Knoll, A. 2015. Multimodal Human Activity Recognition for Industrial Manufacturing Processes in Robotic Workcells. In *ICMI*.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Ryoo, M. S.; Rothrock, B.; Fleming, C.; and Yang, H. J. 2017. Privacy-Preserving Human Activity Recognition from Extreme Low Resolution. In *AAAI*.

Sabokrou, M.; PourReza, M.; Fayyaz, M.; Entezari, R.; Fathy, M.; Gall, J.; and Adeli, E. 2019. AVID: Adversarial Visual Irregularity Detection. In *Asian Conference on Computer Vision*, 488–505.

Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 568–576.

Søgaard, A.; and Goldberg, Y. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 231–235.

Standley, T.; Zamir, A.; Chen, D.; Guibas, L.; Malik, J.; and Savarese, S. 2020. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, 9120–9132. PMLR.

Sun, X.; Panda, R.; Feris, R.; and Saenko, K. 2020. AdaShare: Learning What To Share For Efficient Deep Multi-Task Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 8728–8740. Curran Associates, Inc.

Vandenhende, S.; Georgoulis, S.; Proesmans, M.; Dai, D.; and Van Gool, L. 2020. Revisiting multi-task learning in the deep learning era. *arXiv preprint arXiv:2004.13379*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 5999–6009.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 20–36. Springer.

Wang, P.; Wang, S.; Gao, Z.; Hou, Y.; and Li, W. 2017. Structured images for RGB-D action recognition. In *CVPRW*, 1005–1014.

Xiao, F.; Lee, Y. J.; Grauman, K.; Malik, J.; and Feichtenhofer, C. 2020. Audiovisual SlowFast Networks for Video Recognition. *arXiv preprint arXiv:2001.08740*.

Xu, P.; Madotto, A.; Wu, C.-S.; Park, J. H.; and Fung, P. 2018. Emo2Vec: Learning Generalized Emotion Representation by Multi-task Training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 292–298. Brussels, Belgium: Association for Computational Linguistics.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*.

Yasar, M. S.; and Iqbal, T. 2021. A Scalable Approach to Predict Multi-Agent Motion for Human-Robot Collaboration. In *IEEE Robotics and Automation Letters (RA-L)*.

Zamir, A. R.; Sax, A.; Shen, W.; Guibas, L. J.; Malik, J.; and Savarese, S. 2018. Taskonomy: Disentangling Task Transfer Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, H.; and Parker, L. E. 2011. 4-dimensional local spatio-temporal features for human activity recognition. In *IROS*.

Zhang, S.; Yang, Y.; Xiao, J.; Liu, X.; Yang, Y.; Xie, D.; and Zhuang, Y. 2018. Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks. *IEEE Transactions on Multimedia*, 20(9): 2330–2343.

Zhang, Y.; and Yang, Q. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

Zhou, F.; Shui, C.; Abbasi, M.; Robitaille, L.-É.; Wang, B.; and Gagné, C. 2020a. Task Similarity Estimation Through Adversarial Multitask Neural Network. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhou, L.; Cui, Z.; Xu, C.; Zhang, Z.; Wang, C.; Zhang, T.; and Yang, J. 2020b. Pattern-Structure Diffusion for Multi-Task Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.