MULTI-STAGE BASED FEATURE FUSION OF MULTI-MODAL DATA FOR HUMAN ACTIVITY RECOGNITION

Hyeongju Choi *Apoorva Beedu* *Harish Haresamudram* *Irfan Essa*

Georgia Institute of Technology, USA
{hchoi375, abeedu3, hharesamudram3, irfan}@gatech.edu

ABSTRACT

To properly assist humans in their needs, human activity recognition (HAR) systems need the ability to fuse information from multiple modalities. Our hypothesis is that multimodal sensors, visual and non-visual tend to provide complementary information, addressing the limitations of other modalities. In this work, we propose a multi-modal framework that learns to effectively combine features from RGB Video and IMU sensors, and show its robustness for MMAct and UTD-MHAD datasets. Our model is trained in two-stage, where in the first stage, each input encoder learns to effectively extract features, and in the second stage, learns to combine these individual features. We show significant improvements of 22% and 11% compared to video only and IMU only setup on UTD-MHAD dataset, and 20% and 12% on MMAct datasets. Through extensive experimentation, we show the robustness of our model on zero shot setting, and limited annotated data setting. We further compare with state-of-the-art methods that use more input modalities and show that our method outperforms significantly on the more difficult MMact dataset, and performs comparably in UTD-MHAD dataset.

Index Terms— Multi-modal learning, Human Activity Recognition, deep learning, Multi-modal fusion

1. INTRODUCTION

In recent years, as we have seen a rapid increase in the bodily worn sensors and it's applications from tracking GPS to automatically switching between activities e.g. triathlons, we have seen a boom in the machine learning algorithms that combine information from multiple sensors for Human Activity Recognition (HAR). With this rise of bodily worn sensors and smart home systems [1, 2], we have seen the expansion to vision-based HAR applications in human-robot interaction, senior care, healthcare, personal fitness, and surveillance [3, 4]. Thus depending on the type of system used, different input modalities such as RGB-D, IMU sensors, IoT device information etc can be used for HAR. Using only IMU data limits the complexity and variety of the actions and activities that can be detected and classified by the HAR system. For instance, the HAR system that relies only on IMU data from the inertial sensor equipped on a wrist might not be able to distinguish between waving and cleaning a window. Hence, these limitations call for a multimodal approach that can leverage other data for accurate detections.

Multimodal HAR is a paradigm that uses data from more than one modality (e.g., RGB videos, IMU sensors, Skeleton joints, or depth videos) and learns an effective way to combine these different modalities to improve the accuracy of the HAR system. With this motivation, our contributions are as follows: *(1)* We propose a HAR framework that works on multi-modal inputs such as RGB video and IMU data. *(2)* We compare single modality vs multi-modal input performance for two datasets: UTD-MHAD [5] and MMAct [6]. *(3)* Through extensive experimentation, we show the effectiveness of our framework even when limited data is available and in zero-shot setting.

2. RELATED WORK

Our work is focused on learning strong representations for the task of human activity recognition. In what follows, we divide the existing literature into two categories based on the number of input modalities used and summarize these works.

2.1. Uni-modal Human Activity Recognition

Single modality based HAR systems have been extensively studied by both ubiquitous computing and vision communities [7, 8, 9, 10, 11, 1]. Vision-based activity recognition's effectiveness mainly stems from the expanding field of representation learning for image and video features[10, 8, 12]. In addition, the abundance of annotated data from datasets like Kinetics [13], HowTo100M [8] has also benefited the vision-based HAR systems. However, the vision-based systems are typically affected by the recording conditions (camera angles, viewpoint, lighting, background variations etc.) as well as occlusions. For target scenarios where privacy is a concern, e.g., medical applications, obtaining recorded video may be very challenging, or even impossible. In such scenarios, inertial measurement unit (IMU) signals are more suitable, as they are generally available on commodity platforms including smartphones and smartwatches and can record data without disruption of user experience. Due to this, recent years have seen significant interest in developing sensor-based activity recognition systems [14, 15, 16]. Although the single

Under review

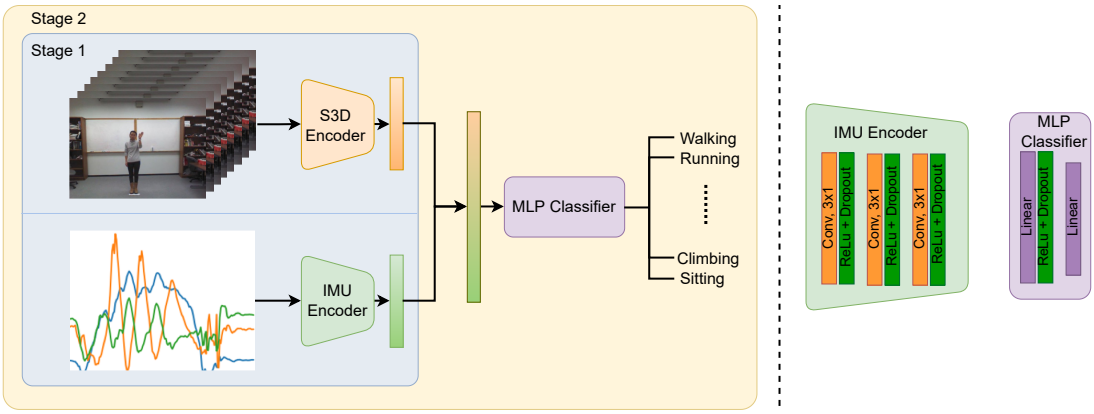

Fig. 1. ***Left:*** Overview of our framework. Videos are encoded using an 3D-CNN architecture: S3D, and IMU data are encoded using 3 blocks of 1D-CNN layers. These features are then fused and passed through a two layer MLP to predict action classes. ***Right:*** IMU Encoder with three layers of 1D convolution + Dropout + ReLu, and MLP Classifier with two linear layer, and one non-linear activation between them.

modality based approaches have demonstrated their effectiveness in recognizing activities, each modality has weaknesses that the other can complement. For example, sensitivity to occlusions can be overcome by using IMU data, and the sensitivity of IMU sensors to body-worn positions [17] can be addressed by using videos and skeleton data.

2.2. Multi-modal Human Activity Recognition

The advantages of complementary modalities have been explored in computer vision for a wide range of tasks. For example, object pose estimation methods often use RGB + Depth data [18, 19], whereas Visual Question Answering (VQA) utilizes text questions on RGB images for answering [20, 21]. Action Recognition tasks have also combined inputs from multiple modalities [22] like depth and infrared data in addition to RGB videos to estimate predefined human actions. With the introduction of Transformer [23] based methods, works such as UniVL [24], Omnivore [25], HAMLET [26] effectively learn to fuse information from multiple modalities for the task of action recognition. UniVL [24] combines information from RGB videos and textual inputs to answer questions about a video and generate captions. Omnivore [25] on the other hand learns to effectively learn joint representation from multiple input modalities such as RGB images, videos, and depth maps. This learnt shared representation is then fine-tuned for action and image classification tasks. However, the combination of Video and IMU data has seen limited exploration, though recently, techniques such as [27, 28, 29] have been introduced. However, these methods generally use skeleton data, or perform fusion and learn features in a non-synchronous way. In contrast, our framework specifically works on the synchronized RGB Video and IMU data, and is trained in a two-stage method that allows for the model to learn strong representations from individual input modalities in the first stage, and effectively fuse the information in the second stage.

3. METHODOLOGY

To address the problems of single modality HAR systems, we introduce a multimodal framework using two modalities: third-person view RGB video and IMU data. We discuss the different feature-extracting modules and fusion of these different features obtained in this section. An overview of the network is shown in Figure 1.

3.1. Feature Encoder for IMU data

IMU data is a multivariate time-series data that (generally) contains data from accelerometers, gyroscopes, and sometimes magnetometers on the x, y, and z axes. 1D Convolutional neural networks (CNNs) are known to be effective for feature learning, as they can parse data across time, thereby capturing the time-series nature of sensor data [7]. We implement a simple encoder for the IMU data, similar to [30], but comprising 3 blocks of 1D convolutional layers. The kernel size is set to 24, 16, and 8 respectively, and the architecture is shown in Figure 1. The IMU encoder is initialized with random weights during training.

3.2. Feature Encoder for Video

We use the S3D network [8] to encode the video features. The S3D network is a video classification network that projects a video into a temporal and spatial representation and is pre-trained on Kinetics400[13].

3.3. End-to-End training

Our framework is trained in two stages. In the first stage, we pre-train both the video and the IMU encoders separately as a supervised action recognition task. As a second stage, we fine-tune the *Mixed_4c* and *Mixed_5c* layers of the S3D architecture, and all the layers of the IMU encoder to predict action classes. Features from both the encoders are concatenated together and then passed through a 2 layer MLP structure to predict the final action classes.

Table 1. Training hyperparameter details

	UTD-MHAD			MMAct		
	lr	Weight decay	batch	lr	Weight decay	batch
IMU	1e-4	1e-6	128	5e-3	5e-5	256
Video	1e-3	1e-5	16	1e-3	1e-5	20
IMU + Video	5e-4	5e-6	16	1e-4	1e-6	18

The entire model is trained with a cross-entropy loss for action classification. Details of the implementation are discussed in Section 4.2.

4. EXPERIMENTS AND RESULTS

4.1. Dataset

To evaluate the performance of our proposed model in this project, UTD-MHAD [5] and MMAct [6] are used.

UTD-MHAD dataset contains a total of 861 samples collected using one wearable inertial sensor and one Kinect camera from 8 subjects, 4 males and 4 females, performing 27 different activities, repeated 4 times for each activity. The dataset consists of 4 different modalities including RGB videos, Depth videos, Skeleton positions, and inertial signals. For our study, only the RGB videos and inertial signals were used. Same as the original paper, we use odd-numbered subjects: 1, 3, 5, 7 for training and even-numbered subjects: 2, 4, 6, 8 for testing for all models.

MMAct dataset contains a total of 36764 samples from 37 activities with each activity repeated 5 times by 20 subjects in 4 different scenes including the scene of occlusion to overcome the weakness of the vision-based HAR systems. The dataset consists of 7 modalities including RGB videos and inertial signals. we follow the cross-subject evaluation protocol by using the samples from the first 16 subjects for training and the rest for testing.

4.2. Implementation

Our model was implemented using the PyTorch [31] framework. Training hyperparameters are detailed in Table 1. All the training for UTD-MHAD was done on a single A40 GPU, while MMAct was trained on 4 A40 GPUs.

Pre-Processing: Both RGB videos and IMU signals were pre-processed for feature encoders. RGB frames were sampled at 15HZ for UTD-MHAD and 30 fps for MMAct. The IMU signals were sampled at 50HZ for UTD-MHAD and a mixture of 100HZ (acceleration) and 50HZ (gyroscope, orientation) for MMAct. The signals were padded with zeros whenever a dimension mismatch occurred.

4.3. Evaluation Metrics

To compare the performance of these different HAR models we report Top 1 accuracy, Top 5 accuracy, and F1 score on the test set for each of the datasets.

4.4. Results and Discussions

We compare the effectiveness of our model with single modality inputs in Table 2. We also compare our results against HAMLET [26], MuMu [29], a contrastive learning based method CMC-CMKM [32], and VSKD [28].

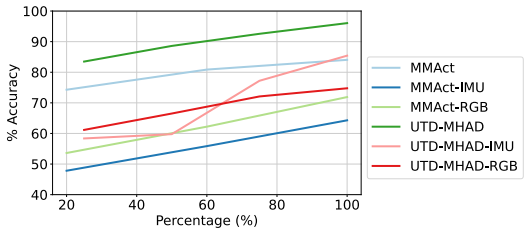

Fig. 2. Action recognition performance for different ratio of training data for MMAct and UTD-MHAD datasets.

We see that when compared against single modality frameworks, our methods perform significantly better (10.6% compared to IMU, and 21.29% compared to Video) across all metrics. When compared to other baseline methods, our method sees a slight drop in performance for UTD-MHAD dataset, but performs comparable or better for MMAct. It is worth noting that other methods use Skeleton information in addition to RGB Video and Sensor data. CMC-CMKM, a contrastive loss based framework uses Skeleton in lieu of video, however, this comes at an additional computation cost and speed, as having an accurate skeleton estimation becomes crucial to their method. As opposed to that, our methods use video features directly, without needing a post-processing framework. However, utilizing skeleton aids in the action recognition task, and we intend to explore this modality further in future work. Although [28] uses the same modalities as our framework, they propose a novel loss called DASK that is responsible for the performance improvement. Their student network that uses IMU data uses a ResNet-18 based model and performs significantly better than our 1D-CNN based model.

4.5. Effect of the size of training dataset

In this experiment, we evaluate the effect of different ratios of the training data on performance. As can be seen in Table 2, we only see a drop of about 10% on UTD-MHAD and 10% on MMAct when the training sample is reduced to 25%. Compared to single modality performances, multi-modal method performs significantly better. This further shows the effectiveness of multi-modal learning for human activity recognition.

4.6. Zero Shot setting

We also test our framework on zero shot setting, and compare against zero shot on RGB and IMU (When few actions classes in the data are hidden during pre-training). During the fine-tuning, the action classes stay hidden to ensure that no data leak happens during the training. We compare the performances on the test dataset for when single modalities are evaluated for zero shot, and when one of the modalities have seen the action classes in Figure 3 in multi-modal setting. For single modality, we see a performance drop of 2% when one action class is hidden and a drop of 14% when 5 action classes are hidden for the IMU data. Whereas, for video, we see a larger drop of 4% when one class is hidden and 12% when 5 classes are hidden. The multi-modal setup also sees a

Table 2. Action recognition performance for Top-1, Top-5 and F1 score compared with baseline methods.

		UTD-MHAD			MMAct		
Method	Data used	Top 1	Top 5	F1	Top 1	Top 5	F1
HAMLET [26]	Video + Skeleton + Sensor	95.12	-	-	-	-	-
MuMu [29]	Video + Depth + Skeleton + Sensor	97.6	-	-	-	-	76.28
CMC-CMKM [32]	Skeleton + Sensor	**97.21**	-	-	**84.05**	-	-
VSKD + DASK [28]	Video + Sensor	96.97	-	**96.38**	-	-	65.83
VSKD [28]	Sensor	94.87	-	-	-	-	61.38
Ours (IMU only)	Sensor	85.42	97.14	85.44	64.28	90.57	65.45
Ours (Video only)	Video	74.76	95.14	74.09	71.89	94.92	72.86
Ours (IMU + Video)	Video + Sensor	96.05	**100**	96.05	**84.05**	**98.96**	**84.78**

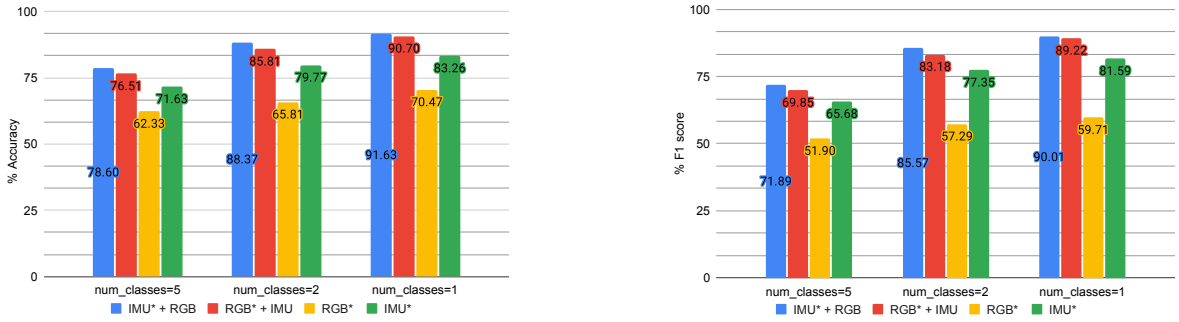

Fig. 3. Zero shot performances on UTD-MHAD dataset. * indicates the input modality from which the action classes were hidden. The four colour bars indicates different input modality combinations and the three columns represent performances when num_classes=$< x >$ were hidden from the input. *Left:* shows the % accuracy, while *Right:* shows the % F1 score.

drop in performance, but performs significantly better to their single modal counterparts. We further notice that when action classes are hidden from IMU data input, but are available in the video input, $(IMU^* + RGB)$ performs better than $(RGB^* + IMU)$ setup. This validates our hypothesis that RGB videos are better at providing complementing information to the model when the IMU input lacks it.

5. CONCLUSION

We have explored multimodal HAR using IMU signals and RGB videos, and the role of pre-training feature encoders for multimodal HAR in this project. We concluded that multimodal HAR outperforms single modal HAR significantly for all datasets. Furthermore, training in two-stages helps the individual encoders to extract relevant information which when fused performs better than other SOTA works. We show the effectiveness of the multi-modal training when smaller datasets are used in training, and also in a zero-shot setting. For the next steps of this project, we would like to explore the effectiveness of using self-attention in learning joint feature representation from multiple modalities and learn better feature fusing protocols, and also explore other modalities for efficient HAR systems.

6. REFERENCES

[1] Homay Danaei Mehr and Huseyin Polat, "Human activity recognition in smart home with deep learning approach," in *2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG)*. IEEE, 2019, pp. 149–153.

[2] Yegang Du, Yuto Lim, and Yasuo Tan, "A novel human activity recognition and prediction in smart home based on interaction," *Sensors*, vol. 19, no. 20, pp. 4474, 2019.

[3] Ming Zeng, Le T. Nguyen, Bo Yu, Ole J. Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *6th International Conference on Mobile Computing, Applications and Services*, 2014, pp. 197–205.

[4] Shugang Zhang, Zhiqiang Wei, Jie Nie, Lei Huang, Shuang Wang, and Zhen Li, "A review on human activity recognition using vision-based method," *Journal of healthcare engineering*, vol. 2017, 2017.

[5] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz, "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, 2015, pp. 168–172.

[6] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami, "Mmact: A large-scale dataset for cross modal human action understanding," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8657–8666.

[7] Federico Cruciani, Anastasios Vafeiadis, Chris Nugent, Ian Cleland, Paul McCullagh, Konstantinos Votis, Dimitrios Giakoumis, Dimitrios Tzovaras, Liming Chen, and Raouf Hamzaoui, "Feature learning for human activity recognition using

convolutional neural networks," *CCF Transactions on Pervasive Computing and Interaction*, vol. 2, no. 1, pp. 18–32, 2020.

[8] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 305–321.

[9] Hossein Rahmani and Ajmal Mian, "3d action recognition from novel viewpoints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1506–1515.

[10] Jun Liu, Amir Shahroudy, Dong Xu, Alex C Kot, and Gang Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 3007–3021, 2017.

[11] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[12] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman, "The kinetics human action video dataset," 2017.

[14] Nils Y Hammerla, Shane Halloran, and Thomas Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," *arXiv preprint arXiv:1604.08880*, 2016.

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

[16] Harish Haresamudram, Irfan Essa, and Thomas Plötz, "Contrastive predictive coding for human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, pp. 1–26, 2021.

[17] Subhas Chandra Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *IEEE sensors journal*, vol. 15, no. 3, pp. 1321–1330, 2014.

[18] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," 2019.

[19] Pengyuan Wang, HyunJun Jung, Yitong Li, Siyuan Shen, Rahul Parthasarathy Srikanth, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam, "Phocal: A multi-modal dataset for category-level object pose estimation with photometrically challenging objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21222–21231.

[20] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra, "Visual dialog," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 326–335.

[21] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[22] Sandy Ardianto and Hsueh-Ming Hang, "Multi-view and multi-modal action recognition with learned fusion," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1601–1604.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[24] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou, "Univl: A unified video and language pre-training model for multimodal understanding and generation," *arXiv preprint arXiv:2002.06353*, 2020.

[25] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra, "Omnivore: A single model for many visual modalities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16102–16112.

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

[27] Md Mofijul Islam, Mohammad SAMIN Yasar, and Tariq Iqbal, "Maven: A memory augmented recurrent approach for multimodal fusion," *IEEE Transactions on Multimedia*, 2022.

[28] Jianyuan Ni, Raunak Sarbajna, Yang Liu, Anne HH Ngu, and Yan Yan, "Cross-modal knowledge distillation for vision-to-sensor action recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4448–4452.

[29] Md Mofijul Islam and Tariq Iqbal, "Mumu: Cooperative multitask learning-based guided multimodal fusion,"," AAAI, 2022.

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

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[32] Razvan Brinzea, Bulat Khaertdinov, and Stylianos Asteriadis, "Contrastive learning with cross-modal knowledge mining for multimodal human activity recognition," *arXiv preprint arXiv:2205.10071*, 2022.