A Survey on Multimodal Wearable Sensor-based Human Action Recognition

Jianyuan Ni, *Student Member, IEEE,* Hao Tang, *Senior Member, IEEE,* Syed Tousiful Haque, *Student Member, IEEE,* Yan Yan, *Senior Member, IEEE* and Anne H.H. Ngu, *Member, IEEE*

Abstract—The combination of increased life expectancy and falling birth rates is resulting in an aging population. Changes associated with aging, can impact an individual's quality of life, potentially leading to injuries, mental health issues, or reduced physical activity. Wearable Sensor-based Human Activity Recognition (WSHAR) emerges as a promising assistive technology to the healthy living of older individuals, unlocking vast potential for human-centric applications. However, recent surveys in WSHAR have been limited, focusing either solely on deep learning approaches or on a single sensor modality. In real life, our human interact with the world in a multi-sensory way, where diverse information sources are intricately processed and interpreted to accomplish a complex and unified sensing system. To give machines similar intelligence, multimodal machine learning, which merges data from various sources, has become a popular research area with recent advancements. In this study, we present a comprehensive survey from a novel perspective on how to leverage multimodal learning to WSHAR domain for newcomers and researchers. We begin by presenting the recent sensor modalities as well as deep learning approaches in HAR. Subsequently, we explore the techniques used in present multimodal systems for WSHAR. This includes inter-multimodal systems which utilize sensor modalities from both visual and non-visual systems and intra-multimodal systems that simply take modalities from non-visual systems. After that, we focus on current multimodal learning approaches that have applied to solve some of the challenges existing in WSHAR. Specifically, we make extra efforts by connecting the existing multimodal literature from other domains, such as computer vision and natural language processing, with current WSHAR area. Finally, we identify the corresponding challenges and potential research direction in current WSHAR area for further improvement.

Index Terms—Multimodal learning, wearable device, inertial measurement units, human action recognition, survey.

1 Introduction

The global population aged 60 or over is expanding at an unprecedented rate. According to the World Population Report, life expectancy at birth is projected to increase from 71 years in 2010–15 to 77 years in 2045–50 [1]. This demographic shift poses a challenge for most societies, as they strive to ensure their health systems are equipped to adapt. Efforts have been made to maintain or improve the quality of life for older individuals. These include the development of new systems that leverage medical and assistive technologies for long-term care provision, as well as the creation of age-friendly environments. In recent years, the advancement of sensors, wireless communication, and machine learning techniques have spurred the development of assistive technologies. These technologies promote independent, active, and healthy aging [2].

Among these techniques, human action recognition (HAR) is a significant field in extracting deep insights about human behavior from raw sensor data, enabling computing systems to monitor, analyze, and assist individuals in their daily lives. As a result,various HAR systems present in many applications, including video surveillance [4], human-robot interaction [5], and healthcare for the elderly [6], [7] as shown in Figure 1.a. At present, two primary types of HAR systems are in common use: video-based systems and wearable sensor-based systems. The first type relies on visual modalities such as RGB video, skeleton, and depth data, whereas wearable sensor-based systems use sensors like gyroscope, accelerometer, etc [3]. Currently, HAR research is dominated by video-based approaches that contains richer information and can capture scene context [8]. While there have been substantial advancements in HAR methods that rely on visual modalities, they have led to increasing privacy concerns due to the utilization of video/image data. For example, purely visual-based systems may not be suitable for areas where privacy is a priority [9]. Moreover, these video-based systems are unable to detect activity if the user is beyond the camera's field of view. Other factors such as lighting conditions, clutter, and occlusion can also adversely affect recognition performance. Consequently, these systems may not be ideal for real-time HAR or for HAR applications that require continuous operation. Over the past decade, the emergence of inexpensive, energy-efficient, and compact Inertial Measurement Units (IMUs) has been a game-changer. Market analysis indicated that the global wearable devices market will reach to about $63 billion by 2025 [10]. This growth in IMUs sensor technology and ubiquitous computing has not only made wearable sensor-based human action recognition (WSHAR) approaches more common but also ensured the preservation of user privacy. Currently, sensors can be embedded in a variety of portable devices, such as smartphones, smartwatches, smart cloths and other specifically-designed devices as shown in Figure 1.b. These devices, equipped with accelerometers and gyroscopes data, enable unobtrusive tracking of human motion and activity

- *J.,Ni and A.H., Ngu are with the Department of Computer Science,Texas State University, San Marcos, TX 78666, USA. E-mail: j_n317@txstate.edu*
- *H.Tang is with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA.*
- *Y. Yan is with the Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616, USA.*

Manuscript received April, 2024

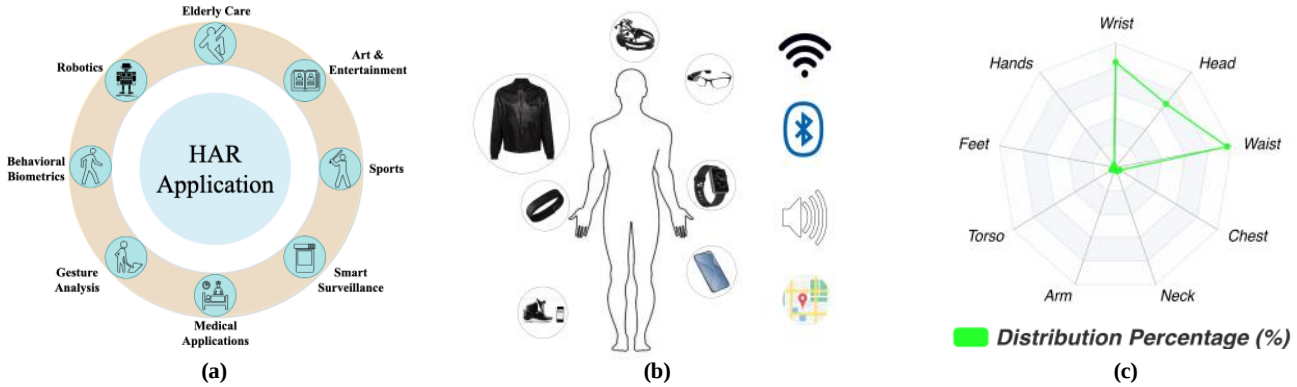

Fig. 1: (a) Applications of wearable sensor-based human activity recognition (HAR). (b) Typical wearable devices for the WSHAR problem. (c) Distribution of wearable devices placed on human body areas [3].

logging for WSHAR purposes.

Among these WSHAR systems, smartphone and smartwatch have emerged as primary platforms not only due to their embedded sensors, but also their communication, processing, and user feedback capabilities [11]. Smartwatches, in particular, may outperform smartphones in this regard because smartphones are only 'on the user' for about 23% of the time [12], and their position relative to the user's body is not fixed [13] as shown in Figure 1.c. Despite their advantages, WSHAR system also present certain challenges. Notably, these WSHAR systems tend to have significantly lower accuracy performance compared to systems based on visual modalities [14]. For example, a prior study observed that fall detection, when using accelerometer data from a wrist-worn watch and processed with deep learning (DL) methods, can only attain an accuracy rate of 86% [15]. In addition, the measurements from wearable devices are sensitive to the sensor's location on the body. Generally, augmenting the number of sensors on various body parts (*e.g.,* head, wrists, waist, legs, feet) can enhance the performance and robustness of WSHAR systems [16]. Therefore, many existing systems based on IMUs require users to wear multiple sensors on various body parts. However, the complex deployment of multiple sensors on the body could lead to higher costs, obtrusive practical implementation challenges, and hinder the user's ability to perform activities naturally, particularly for older users who are capable of living independently. Moreover, the combination of sensors across human body may still struggle to accurately recognize certain activities that exhibit similar sensor-derived characteristics, such as putting on jackets and falling [7]. Consequently, effectively addressing the challenge of HAR requires an approach that goes beyond the limitations of single-modality systems.

In daily life, we engage with our surroundings through a multimodal cognitive system. This is mainly due to the fact that different regions of our brain are responsible for processing this diverse information, each specializing in a specific type of sensory input [17]. This system allows us to perceive objects, interpret sounds, and understand text on a daily basis. For example, reading can stimulate the reconstruction of corresponding visual imagery in our minds. Therefore, utilizing multimodal data can be advantageous in interpreting complex activities, as it encompasses a wealth of semantic knowledge [8]. Meanwhile, multimodal data is a rich source of information that can reveal long-term temporal relationships between objects. These relationships are similar to the sequential order of activities within an extended sequence, much like how the human brain works [18]. This comprehensive analysis should encompass the interpretation of objects, scenes, and the temporal relationships of activities. For instance, when a person recalls a memory, it's like triggering a sequence in a long-term video. Such a strategy not only offers a more complete perspective but also improves our capacity to forecast activities over extended timeframes.

Previously, there have been several surveys which focused on the taxonomy of HAR and reviewed HAR systems implemented with conventional machine learning methods. For instance, Poppe [19] reviewed vision-based HAR research, discussing image representation and action classification methods. Aggarwal and Ryoo [20] presented a taxonomy for HAR based on their approach, discussing recognition methods for simple human actions and high-level activities. Incel *et al.* provided a smartphone activity recognition taxonomy, introducing the process and challenges of HAR on phones, and reviewing works based on location, motion, and other contextual information [21]. Lara and Labrador [22] surveyed HAR with wearable sensors, discussing different types of activities, design issues, and recognition methods. They evaluated 28 systems in terms of recognition accuracy, energy consumption, obtrusiveness, and flexibility. Bulling *et al.* [23] provided a tutorial on HAR with conventional machine learning methods based on wearable IMUs. More recently, with the development of DL methods, several state-of-the-art surveys were conducted for HAR problem [3], [8], [10], [16], [24], [25], [26], [27], [28], [29], [30], [31], [32]. However, these surveys are either focusing on advanced DL approaches [25], [27], [29], [30], [31] or only each single sensor modality [10], [16], [24], [26], [32]. While there are also a few surveys, such as [3], [8], that have summarized existing HAR methods from the perspective of data modalities, they have not emphasized how to use multimodal learning approaches to enhance WSHAR

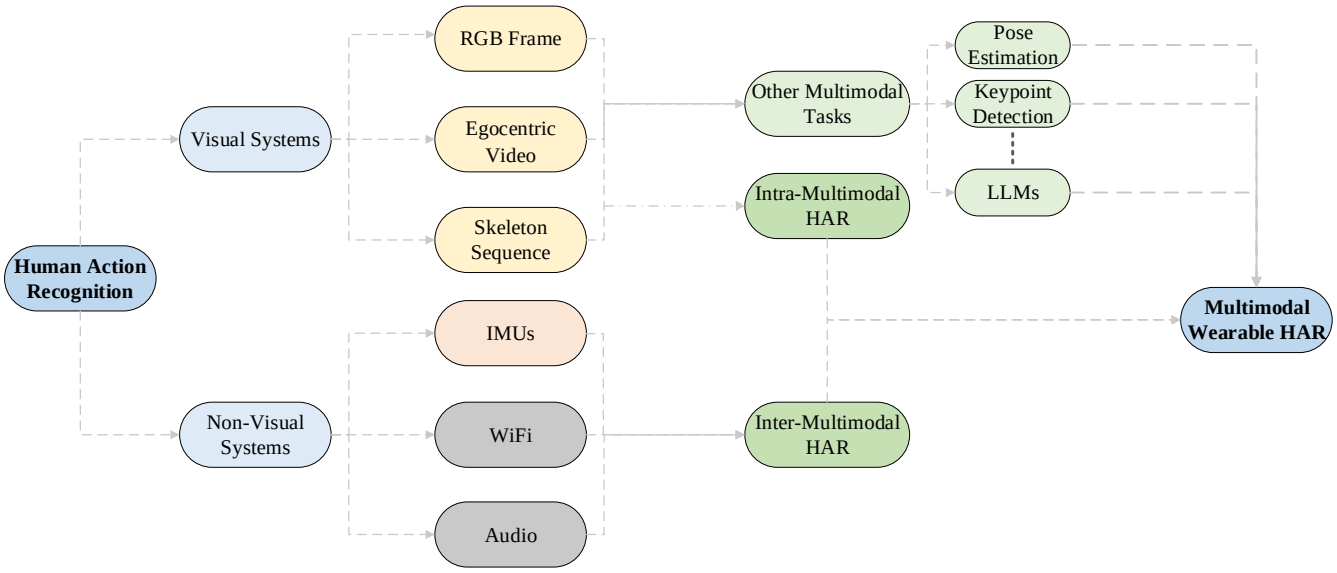

Fig. 2: Overall structure of our survey. We first present two mainstream representations available for HAR systems (Visual and Non-Visual) and their current achievements. Next, we proceed to introduce multimodal applications to emphasis on the emergence need in wearable HAR domain. We take extra efforts by combining existing multimodal studies from other tasks to form the basis for our discussions on the existing challenges and possible future directions.

performance. As mentioned earlier, each data modality has its own strengths and limitations, and understanding how to leverage these can be crucial for improving WSHAR performance. Motivated by these observations, this comprehensive survey is designed to bridge the existing knowledge gap by connecting the advanced multimodal achievements from computer vision (CV) or natural language processing (NLP) domains with current WSHAR area. It emphasizes the exploration of different modalities' strengths and how they can be leveraged to enhance WSHAR's overall performance. It serves as an invaluable resource for new researchers venturing into the field of WSHAR, providing them with a wealth of information and guidance. Those who are grappling with the choosing appropriate methods to address challenges in WSHAR field can find strong clues in this survey. The main structure is illustrated in Fig.2.

The overall structure of the survey is as follows: In Section 2, we provide an overall analysis in terms of data characteristics for HAR problem, including visual representations (RGB frame, egocentric video, and skeleton sequences) and non-visual representations (audio, WiFi, and inertial sensor data). Next, we introduce current multimodal datasets which include inertial sensor data for WSHAR problem in Section 3. We also devise the current multimodal approaches for WSHAR from two perspectives: inter-multimodal learning (modalities from visual and non-visual systems) and intra-multimodal learning (modalities from non-visual systems only). We then proceed to present how the latest multimodal approaches can be a solution to solve some common challenges in WSHAR domain in Section 4. In addition to the task-wise and technical introduction, we also make extra efforts by connecting the existing multimodal literature from other tasks to form the basis for possible future directions in Section 5. Section 6 includes the final remarks and conclusions.

2 Data Analysis for HAR

In real life, each modality, with its distinct advantages and disadvantages, contributes to our understanding of complex systems and phenomena. In this section, we mainly focus on the inherent traits of various data modalities, including visual and non-visual based HAR systems. Specifically, visual-based modalities, such as video frame, egocentric video and skeleton sequence, which are the primary source of information in the human sensory system, is introduced first. After that, we present the current achievements using non-visual modalities for HAR problem, including audio, WiFi and inertial sensor data.

2.1 Visual-based System

2.1.1 RGB Frame

In the early days, handcrafted features approaches are applied to calculate the movements and spatial changes in the video to conduct HAR problem [49]. After that, the application of DL methods on HAR has received considerable attentions. In general, video-based HAR work can be categorized into three sections: 2D Convolutional Neural Network (CNNs), Recurrent Neural Network (RNNs) and 3D CNNs-based methods [8]. For instance, Simonyan and Zisserman present a two-stream CNNs which use spatial and temporal features from individual RGB frame of video input for HAR [50]. RNNs was also used to analyze temporal sequence data due to the recurrent connections in their hidden layers [51]. Subsequently, several 3D CNNs approaches has been proposed recently. For instance, Feichtenhofer *et al.* [33] proposed a two-stream 3D CNNs framework consisting of a fast and slow pathway. Of this model, the slow pathway operate on the RGB frames at a low rate to capture semantic features, while the fast pathway work on high frame rates to extract motion features. In addition, Lin *et al.* [34]

TABLE 1: Various modality examples of visual and non-visual systems with pros and cons

Type	Modality	Mainstream Methods	Pros	Cons
Visual	RGB Frame	3D CNNs [33], [34] Transformer [35], [36]	· Rich appearance information · Easy to operate	· Viewpoint sensitive · Privacy concerns
	Egocentric Video	Transformer [37], [38]	· Active Behavior · First person view	· Motion alteration · Privacy concerns
	Skeleton Sequence	GNNs [39], [40] Transformer [41], [42]	· Simple yet informative · Viewpoint Insensitive	· No appearance information · No shape information
Non-Visual	Audio	CNNs [43], [44]	· Ubiquity · Low cost	· Noise inference · Variability
	WiFi	CNNs [45], [46]	· Wide coverage · Penetrating ability	· Robustness · Activity Variation
	Inertial Sensor	RNNs [47], [48]	· Low cost · Privacy preservation	· less expressiveness · Noisy

proposed a Temporal Shift Module (TSM), which shifts a part of the channels along the temporal dimension. The information from adjacent frames was then interacted with the current frame after shifting. More recently, Transformer architecture was applied to video-based HAR field [35], [36]. For instance, in [36], Multiview Transformers, which utilize multiple individual encoders, was proposed. Lateral connections between these individual encoders was integrated to efficiently fuse information from different representations of the input video. In summary, video modality contains rich RGB information and it is easy to collect for HAR problem. However, RGB video are often sensitive to various viewpoints and occlusions and are not privacy-preserving.

2.1.2 Egocentric Video

Egocentric videos, as opposed to third-person RGB videos, are rich in intrinsic features which are beneficial for HAR problem as these features encompass interactions with objects without occlusion [52]. Currently, there are mainly two categories for egocentric HAR research area, including object-driven and motion-driven approaches [53]. At present, object-driven approaches demonstrated that objects present in the scene and, specially, objects related to tasks are the main cues in the recognition of actions [54], [55], [56]. In addition to cues from objects, motion-related cues, such as the overall movement produced by scene objects, also play a crucial role for HAR problem [57], [58]. More recently, Transformer architecture was applied to egocentric video-based HAR field [37], [38]. For instance, Zhang *et al.* propose an object-aware decoder to enhance the performance of spatio-temporal representations for HAR using egocentric videos [37]. This approach involves augmenting object awareness during training by training the model to predict hand positions, object positions, and semantic labels of objects using paired captions when provided. During inference, the model only needs RGB frames as inputs and can effectively track and ground objects. However, one of the primary issues using egocentric videos is the substantial camera motion resulting from the wearer's movements, which can lead to inconsistent and unstable footage. Additionally, the wearer's viewpoint, which defines the field of view in egocentric videos, may not encompass the entire context of the action, particularly if it involves other people or objects outside this field. Privacy is another concern as the camera might inadvertently capture sensitive details about the wearer or others present in the video.

2.1.3 Skeleton Sequence

Skeleton sequences encode the trajectories of human body joints, which characterize informative human motions, can be another candidate for HAR problem. Previously, Various methods [59], [60] have applied RNNs and LSTMs to effectively model the temporal context information within the skeleton sequences for HAR problem. Due to the expressive power of graph neural networks (GNNs), analyzing graphs with learning models have received great attention recently [39], [40], [61]. For example, Yan *et al.* exploited GNNs for skeleton-based HAR by introducing Spatial-Temporal GCNs (ST-GCNs) that can automatically learn both the spatial and temporal patterns from skeleton data [39]. More specifically, the pose information was estimated from the input videos and then passed through the spatio-temporal graphs to achieve action representations with strong generalization capabilities for HAR. More recently, Transformer architecture was applied to HAR field using skeleton sequences [41], [42]. For instance, Qiu *et al.* introduced the SpatioTemporal Tuples Transformer (STTFormer) architecture for HAR problem. The STTFormer initially divides a skeleton sequence into non-overlapping clips and then utilizes a spatio-temporal self-attention module to capture multi-joint dependencies between adjacent frames. Finally, an inter-frame feature aggregation module aggregates sub-actions to refine the recognition process.

In summary, the skeleton modality provides the body structure information, which is simple, efficient, and informative for representing human behaviors. Nevertheless, HAR using skeleton data still faces challenges, due to its very sparse representation, the noisy skeleton information, and the lack of shape information that can be important when handling human-object interactions.

2.2 Non Visual-based System

In addition to systems based on visual data, non-visual systems have received significant attentions recently due to their robustness, ability to preserve privacy, and potential for integration into multimodal recognition systems [62]. These attributes make them a valuable complement to visual systems in the field of HAR research.

2.2.1 Audio

Sound serves as an effective medium for capturing the structural characteristics of human activities. Previously,

this approach has proven feasible across various sensing platforms and application domains, including bathroom-related activities [63] and context recognition systems [64]. Moreover, several studies applied DL methods to perform general HAR from audio signals [43], [44], [65], [66]. For example, Lane Lane *et al.* designed DeepEar, a pilot mobile application for multi-task sound-based detection [65]. Yatani and Truong [66] developed the BodyScope system, capable of detecting 12 human activities related to throat movement. Laput *et al.* also present a a plug-and-play HAR system that leverages sound features from multiple online datasets [43]. Liang and Thomaz [44] further applied a pre-trained large-scale model to extract acoustic embedding features from public YouTube video sound clips to improve HAR performance in real-world settings. This framework integrates transfer learning, oversampling, and a deep learning architecture, eliminating the requirement for feature augmentation or semi-supervised methods. However, Using audio alone for accurate HAR is uncommon due to its limited ability to provide sufficient information.

2.2.2 WiFi

WiFi is a prevalent indoor wireless signal which can be used for HAR, and even through-wall HAR, as human bodies reflect wireless signals well [67]. Most existing WiFi-based HAR methods focus on using Channel State Information (CSI), a fine-grained information derived from raw WiFi signals, for HAR tasks [8], [68]. The unique variations in CSI at the WiFi receiver are usually generated by the reflected WiFi signal of a person performing an action. Previously, LSTM networks have been used for HAR using CSI signal [45], [46]. In the work of [46], the spatial features of the CSI signal were first extracted from a fully connected layer of a pre-trained CNN. These features were then fed to a Bi-LSTM to capture the temporal information for HAR. Chen *et al.* [45] directly passed the raw CSI signal through an attention-based Bi-LSTM to predict the action class. Different from the above-mentioned works, Gao *et al.* [69] transformed the CSI signal into radio images, which were then fed to a deep sparse auto-encoder to learn features for HAR problem. However, there are still some challenges which need to be further addressed, such as how to more effectively use the CSI phase and amplitude information, and to improve the robustness when handling dynamic environments.

2.2.3 Inertial Sensor

Rapidly development of wearable devices, such as smartwatch and smartphone, make it suitable to monitor HAR problem. Previous work has been investigated on how to apply CNN on wearable-based HAR [70]. RNNs type model was then suggested to deal with such time-dependent input sequences [48], [71]. Wang *et al.* [47] integrated a CNN and Bi-LSTM model to acquire spatial and temporal features from acceleration data. Meanwhile, some approaches also suggested converting wearable sensor sequences as images for HAR study. Lu *et al.* [72] encoded the tri-axial acceleration data into color images, which were fed into a ResNet for HAR later. Jiang and Yin [73] assembled accelerometer and gyroscope sequences into an active image. After that, a CNN model was adopted to learn the optimal features from the generated active image. Currently, there is still a significant performance gap between WSHAR and visual-based HAR systems due to intra-modality variations.

3 MULTIMODAL HUMAN ACTIVITY RECOGNITION

In real life, humans often perceive the environment in a multimodal cognitive way. By leveraging the advantages and capabilities of various data modalities, multimodal machine learning can often offer more robust and accurate HAR. In this section, we summarize the latest achievements in multimodal learning designed for current WSHAR research from two perspectives: inter-multimodal systems (utilizing modalities from both visual and non-visual systems) and intra-multimodal systems (using modalities from non-visual systems only).

3.1 Inter Multimodal HAR Approach

Multimodal learning is a modeling method combining and processing compensatory information from multiple modalities. Therefore, visual and non-visual modalities can be fused to leverage their complementary discriminative capabilities for more accurate and robust HAR models. Previously, some handcrafted feature-based methods [74], [75] have exploited the fusion of acceleration and visual modalities for HAR. For example, both feature and decision level fusion was used by [74], with the combination of Kinect depth sensor and IMUs using collaborative representation classification. Their results showed 2-23% improvement in the accuracy with the combination of depth and IMUs in comparison to the situations where they are used individually. Elmadany *et al.* [75] further fused the RGB video, depth, skeleton and IMUs sensor data for addressing the HAR problem. The study employed both bimodal hybrid centroid canonical correlation analysis (BHCCCA) and multimodal hybrid centroid canonical correlation analysis (MHCCCA) to explore discriminative and informative shared spaces.

More recently, several DL methods [81], [82] have been proposed for inter-multimodal HAR systems. Combining video cameras and IMUs enhances HAR performance, as IMUs typically provide orientation and acceleration information for body segments, while videos offer positional information [83]. Marcard *et al.* [84] introduced a hybrid tracker that combined multi-view video camera and IMUs for motion capture to mitigate the limitations of each sensor type. This approach used videos to obtain drift-free body position and accurate limb orientations and robust performance during rapid motions using IMUs sensors. Li *et al.* [85] integrated incremental learning and decision table with swarm-based feature selection to achieve fast and accurate HAR performance by fusing kinect camera and wearable sensor data. Ranieri *et al.* [86] use simultaneously data from videos, wearable IMUs and ambient sensors for HAR problem and result indicated that introduction of data from ambient sensors expressively improved the accuracy results. More recently, Ijaz *et al.* [87] proposed a multimodal Transformer for nursing action recognition, where the correlations between skeleton and acceleration data are fused together. Specifically, the tokens from the temporal transformer block of the spatial-temporal skeleton model serve as queries, while the tokens from the acceleration encoder block serve as key and value pairs.

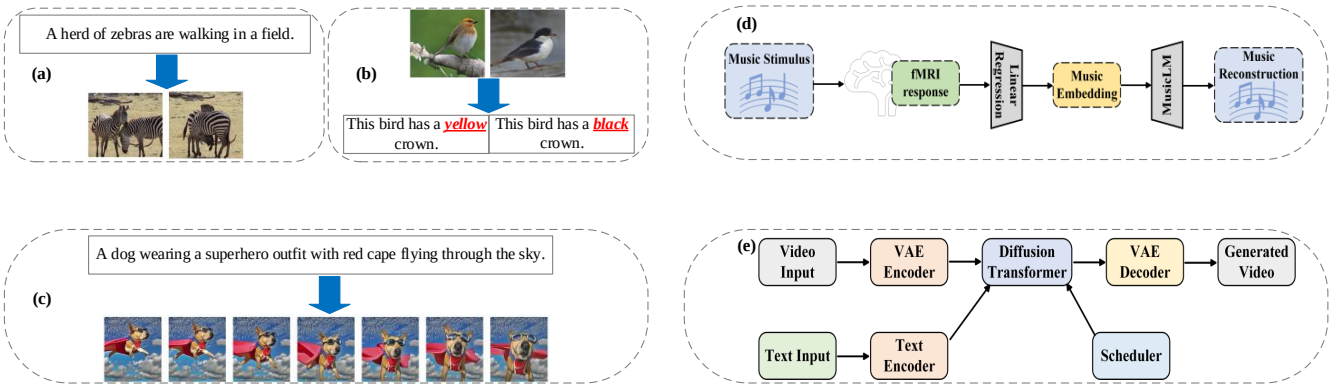

Fig. 3: Current advanced multimodal tasks for other tasks. (a) Text-to-image generation task [76]. (b) Image-to-text generation task [77]. (c) Text-to-image generation task [78]. (d) Reconstructing music from human brain activity [79]. (e) Recent Sora study using diffusion models for video generation tasks [80].

Additionally, several approaches have converted inertial signals into images to fully leverage advanced computer vision models [81], [88], [89], [90]. For instance, Dawar *et al.* [88] represented inertial signal as an image, and utilized two CNNs to fuse the depth images and inertial signals using score fusion. Wei *et al.* [81] respectively fed the 3D video frames and 2D inertial images to a 3D CNN and a 2D CNN for HAR, and the score fusion achieved better performance than feature fusion. Das *et al.* [89] present a multimodal ensemble networks which consisted of three models. CNNs are made for skeleton sequence while one CNN and one LSTM was trained for RGB images. After that, accelerometer and gyroscope data was converted to signal diagram for another CNN model. Qin *et al.* [90] also encoded time series of sensor data as images. After that, a fusion residual network is adopted by fusing these heterogeneous data with pixel-wise correspondence. In this way, WSHAR problem can be transformed as image recognition task using CV techniques.

Another type of multimodal HAR category is refereed as knowledge transfer and knowledge distillation (KD) is considered as a general technique to assist the learning process from teacher modality to student one [91]. Kong *et al.* [92] proposed a multimodal attention distillation method to model video-based HAR with the instructive side information from inertial sensor modality. Liu *et al.* [4] introduced the Semantics-aware Adaptive Knowledge Distillation Networks (SAKDN) model on HAR. In this model, the knowledge from multiple wearable sensors were adaptively transferred to video modality. More recently, Ni *et al.* present the first multimodal KD approach for the WSHAR problem [93]. In this study, an adaptive transfer of complementary information from the video domain to the sensor domain was proposed to improve the accuracy of sensor-based HAR. In order to eliminate the privacy concern, they further adopted skeleton sequence modality as the teacher model to distill knowledge to time-series modality for accurate WSHAR problem [14]. Of these frameworks, they will not only improve the accuracy performance of WSHAR, but also reduce the computation resource demands during the testing phase. However, a significant drawback is that student models, which take time-series data as input, typically exhibit lower accuracy performance compared to the pre-trained teacher model. This suggests that the KD method may not be able to fully exploit the advantages of multimodal learning for HAR problems due to the problem of the performance gap, which refers to the performance difference between the teacher and student models [94], [95], [96].

3.2 Intra Multimodal HAR Approach

Currently, some studies have also combined various sensors from non-visual systems by fusing them to increase the HAR performance. For instance, CNNs was introduced to identify human activities by gathering multi-channel time-series data by employing several IMUs [110]. Chetty *et al.* [111] presented a multimodal CNNs system using gyroscope and accelerometer data from smartphone, applying it to eHealth scenarios for the elderly and people with special needs. In [112], the authors integrated various IMUs using a feature ensemble method from multiple wearable sensors. Additionally, they introduced a layered fusion model that utilizes entropy weight to track human activities using these IMUs [113]. Yao *et al.* designed a architecture which consist of three different CNNs sequential blocks which can learn local patterns, high-level relationship as well as temporal features among input sensors, to merge multimodal data for sensor-based HAR problem [114]. After that, Sena *et al.* utilized multi-scale CNNs ensemble approach to not only extract both simple movement patterns as well as complex movements to deal with data heterogeneity problem [115].

LSTM-based model was designed in [116], where the data obtained from the gyroscope and accelerometer were first normalized. After that, the normalized data were then passed on to the stacked LSTM network for HAR problem. Yu *et al.* [117] proposed a Bi-LSTM model utilizing data from gyroscope and accelerometer sensors from a mobile phone for HAR problem. Moreover, Ihianle *et al.* [118] present a multi-channel architecture using both CNNs and BLSTM to extract features from multimodal sensing devices for activity recognition. Dua *et al.* [119] further integrated CNNs with Gated Recurrent Unit (GRU) modules to extract long-term temporal dependencies from accelerometer and gyroscope sensors data for HAR problem.

TABLE 2: Representative multimodal benchmark datasets with various data modalities for WSHAR . S: Skeleton, D: Depth, Au: Audio, Ac: Acceleration, Gyr: Gyroscope, EMG: Electromyography.

Dataset	Year	Modality	# Class	# Subject	# Sample	# Viewpoint
Gabel *et al.* [97]	2012	D,Ac	6	23	-	1
Berkeley MHAD [98]	2013	RGB,S,D,Au,Ac	12	12	660	4
Delachaux *et al.* [99]	2013	D,Ac	11	-	-	4
Liu *et al.* [100]	2014	D,Ac	6	3	-	1
UTD-MHAD [101]	2015	RGB,S,D,Ac,Gyr	27	8	861	1
Malleson *et al.* [102]	2017	RGB,Ac	-	8	-	1
Dawar *et al.* [103]	2018	D,Ac	5	12	-	1
Manzi *et al.* [104]	2018	RGB, D,Ac	10	20	-	1
MMAct [92]	2019	RGB,S,Ac,Gyr,etc.	37	20	36,764	4
EV-Action [105]	2020	RGB,S,D,EMG	20	70	7,000	9
HOMAGE [106]	2021	RGB,Ac,Gyr,etc.	75	27	1,752	2-5
Ego4D [107]	2022	RGB,S,D,Au,Ac	-	923	-	1
EPIC-KITCHENS-100 [108]	2022	RGB,Au,Ac	-	45	89,979	1
VIDIMU [109]	2023	RGB,Ac	13	54	-	1

More recently, attention mechanism was introduced for multimodal sensing HAR [120], [121], [122]. For instance, Gao *et al.* present a dual attention framework that integrates both channel and temporal attention simultaneously to capture temporal-spacial patterns from accelerometer and gyroscope for HAR task [120]. Tang *et al.* [121] further proposed a a triplet cross-dimension attention for WSHAR problem. These three attention branches were used to capture the cross-interaction between sensor dimension, temporal dimension and channel dimension from accelerometer and gyroscope sensor signals. Al-qaness *et al.* incorporated RNNs with attention module to extract time-series feature for wearable HAR problem [122].

Simultaneously, other works explore non-visual modalities, such as audio or WiFi signals, in conjunction with accelerometer data for multimodal HAR problems [29], [123], [124], [125]. For instance, Garcia *et al.* [124] treated audio and accelerometer sensor data as different views and applied stacked generalization approach to fuse them for wearable HAR problem. Siddiqui and Chan [125] further investigated the use of acoustic signals with an accelerometer and gyroscope from the human wrist for gesture recognition. Mollyn *et al.* first employed the IMU inputs to serve as a trigger for identifying activity events. Upon detection of these events, a multimodal learning model which augmented the IMU samples with sub-sampled audio data capture from a smartwatch was proposed for WSHAR problem [126]. In addition, Lin *et al.* [127] proposed a multimodal system using a smartphone with an off-the-shelf WiFi router for HAR problem. The router functions as a hotspot for transmitting WiFi packets, while the smartphone is equipped with customized firmware and developed software to capture WiFi CSI information simultaneously.

While there are numerous studies focusing on multimodal learning approaches for WSHAR problems, the current methods employed in WSHAR field are still in their infancy compared to the models used in other advanced tasks, such as text to image/video generation [76], [78], [80] or music reconstruction from brain activity [79] as illustrated in Figure 3. More recently, Large Language Models (LLMs) have gained substantial attentions since preliminary results indicated that LLMs possessed distinct capabilities in utilizing inherent world knowledge to interpret IoT sensor data and make logical deductions about physical world tasks [128]. This not only paves the way for new applications of LLMs beyond conventional text-based tasks, but also introduces innovative methods for integrating human knowledge into real-world systems. As a consequence, we highlight significant multimodal approaches addressing challenges in current WSHAR systems. Some of these solutions involve the application of multimodal learning, derived from other advanced tasks, to enhance the performance of WSHAR systems.

4 Challenges in multimodal Wearable HAR System

Previously, many studies have pointed out current challenges in the WSHAR area, including annotation scarcity, class imbalance, distribution discrepancy and computational cost [8], [26]. However, the application of multimodal learning approaches as to address these challenges in WSHAR problems has not been previously explored. In the following section, we will outline some of the existing challenges in the WSHAR domain and provide recent multimodal solutions for addressing these common challenges.

4.1 Multimodal Dataset Challenge

The dataset plays a critical key in the success of deep learning research. Consequently, numerous multimodal datasets have been collected to train and evaluate the HAR performances [8]. However, only a few of them include IMUs data, which is essential for WSHAR problem. Additionally, there are also some other challenges, such as limited labeled data and class imbalance, existing in these multimodal datasets.

4.1.1 Large-scale Dataset Scarcity

In the age of deep neural networks, the absence of large-scale labeled datasets makes it challenging to fully utilize the advantages of DL methods. As a result, the development of efficient methods for acquiring labeled datasets has been a long-standing research interest across various communities. In fact, prior research has suggested that the lack of extensive dataset is a contributing factor to the slower progress

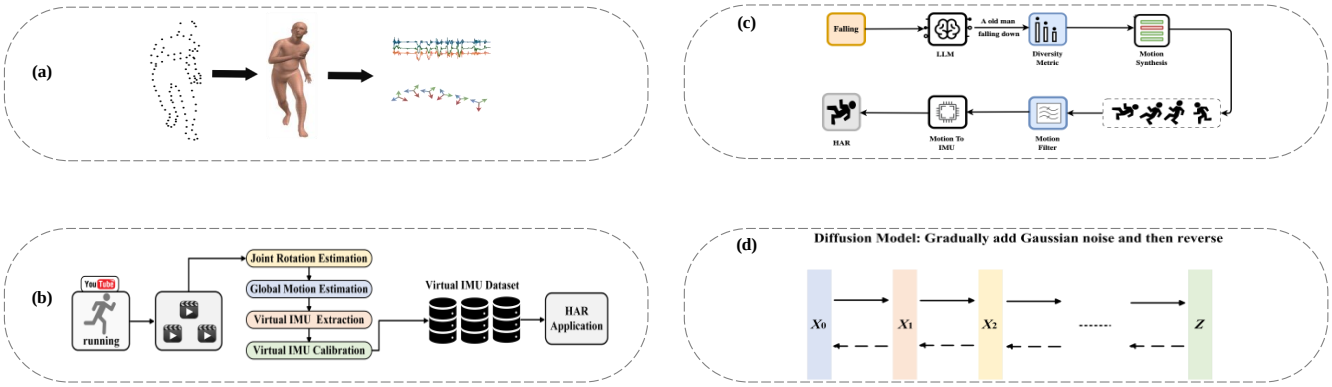

Fig. 4: Current approaches for multimodal WSHAR dataset scarcity problem. (a) Synthesized IMU data from motion capture datasets using skinned multi-person linear (SMPL) model [129]. (b) Virtual IMU data generation pipeline from LLMs domain [130]. (c) Virtual IMU data generation pipeline from video domain [131]. (d) IMU data generation using advanced diffusion model [80].

of WSHAR area compared to other fields [132]. Nevertheless, there are only a fewer multimodal datasets including IMUs signals for WSHAR problems, as shown in Table 2. In general, MMAct [92], and EPIC-KITCHENS-100 [108] are large benchmark datasets suitable for multimodal learning approaches for WSHAR field. In addition, Berkeley MHAD and UTD-MHAD [101] datasets are also popular used. Compared to benchmark datasets from other modalities domain, such as ImageNet [133], VidProM [134], Panda-70M [135], and BEHAVIOR-1K [136] in CV area, and Aya [137] in NLP domain, which contained millions of samples, current large-scale multimodal datasets for WSHAR problem are extremely under-explored. Moreover, the collection of even unlabeled sensor data presents its own set of logistical and ethical hurdles. These include the installation of sensors, the recruitment of human participants, and the need to address privacy concerns. Consequently, the development of efficient and effective methods for rapidly acquiring large-scale, fully labeled datasets would be a valuable contribution to the WSHAR research community.

Existing studies have utilized various tasks, such as human pose estimation [129], [138], [139] and image classification [140], [141] to leverage along with the WSHAR research. For instance, Huang *et al.* [129] synthesized IMU data from motion capture datasets using skinned multi-person linear (SMPL) model surface [142] as shown in Figure 4.a. Specifically, to generate synthetic IMU training data, virtual sensors are positioned on the SMPL mesh surface. Orientation readings are obtained directly through forward kinematics, while accelerations are derived using finite differences. Xia *et al.* also created a comprehensive synthetic HAR dataset using the SMPL model. This dataset includes multimodal data, such as acceleration and angular velocity, which were generated according to the forward kinematics approach [138]. Hao *et al.* also presented a high-precision virtual IMU sensor simulator from either motion capture systems or single-lens RGB cameras using the SMPL model surface [142]. In order to reduce measurement noise and calibration errors, the functional mapping from imperfect trajectory estimations was learned by a DNN model to mitigate the data scarcity problem. Additionally, Hashim and Amutha [141] transformed accelerometer and gyroscope sensor data to the visual image using novel activity image creation (NAICM) method. After that, pre-trained models on ImageNet [133] are transferred for HAR problem. Yoon [143] also converted IMU sensor data into visually interpretable spectrograms. Pre-trained representations from the ImageNet [133] dataset were employed for diverse few-shot IMU tasks by using contrastive learning.

In situations where video and IMU data streams cannot be accessed simultaneously, a straightforward solution could be to leverage large-scale datasets from existing multimodal HAR datasets. The objective here is to utilize the data from existing HAR datasets, which are extensive and modality-agnostic, to address the issue of data scarcity in the WSHAR domain. For instance, numerous studies in the CV domain have focused on advanced methods to extract skeleton joints data from video streams [144], [145]. Since accelerometer data can be regarded as the second derivative of the skeleton sequence coordinates [146], extracting accelerometer data directly from video streams is practical. Indeed, the use of such sophisticated tools has significantly advanced the field of WSHAR by generating virtual IMU data from the video domain [131], [143], [147], [148], [149], [150], [151]. For instance, Kwon *et al.* proposed an engineering pipeline, IMUTube, to generate on-body virtual sensor data using data from video modality [131] as shown in Figure 4.b. The proposed processing pipeline transforms the video data into usable virtual sensor (IMU) data. This involves extacting 2D pose information from videos, which is subsequently converted to 3D. By tracking individual joints of the resulting 3D poses, sensor data like tri-axial acceleration values is generated across various locations on the body. These values are then post-processed to align with the target application domain. They tested their approach in a realistic gym exercises scenario involving large body movements. The result demonstrated that HAR systems trained with virtual sensor data significantly outperform baseline models trained only with real IMU data [147]. Jain *et al.* [148] further evaluated the effectiveness of the IMUTube pipeline in detecting subtle motion activities, particularly focusing on eating detection tasks. They found

that IMUTube significantly improved recognition accuracy in these tasks. Leng *et al.* designed a motion subtlety index to measure local pixel movements and changes in pose at specified virtual sensor locations, exploring its relationship with the accuracy of activity recognition to evaluate the IMUTube pipeline [152]. Additionally, Fortes *et al.* first trained a general regression model for both accelerometer and gyro signals [153]. This model is then applied to video footage of specific activities, enabling the generation of synthetic IMU data that can be used to improve HAR models. In order to further improve the quality of virtual IMU data, some studies aim to produce more realistic and diverse IMU signals [149], [154]. For instance, Gavier *et al.* [149] propose a systematic approach to synthesize realistic and diverse IMU data, including three-axis accelerometer and gyroscope measurements, from video-based skeleton representations.

Meanwhile, with the recent advancements of large language models (LLMs), pre-trained LLMs models that can be adapted to solve a wide range of multimodal downstream tasks [155]. For instance, CLIP was developed to learn the association between images and their textual descriptions [156]. More recent developments, such as Next-GPT [157], have pushed these boundaries even further, enabling the integration of multiple diverse modalities. Inspired by these achievements, scholars are currently investigating the capabilities of LLMs in the field of time series analysis. In order to address the issue of scarce annotated data, Li *et al.* employed clinical reports that are automatically generated by LLMs to serve as a guide for a self-supervised pretraining framework for ECG data [158]. A trainable ECG encoder and a fixed language model were employed to embed paired ECG and automatically generated clinical reports independently. Similarly, Liu *et al.* [159] provided a systematic demonstration of how LLMs can effectively interpret numerical time series data using few-shot prompt tuning. Zhang *et al.* evaluated the proficiency of LLMs, such as Claude-2 [160], in identifying unusual patterns in mobility data [161]. The experimental observations indicated that LLMs are capable of achieving commendable performance in anomaly detection. Liu *et al.* further demonstrated that LLMs are capable of grounding time series data for activity recognition with only few-shot tuning approaches [159].

More recently, a few studies started to use LLMs to generate varied virtual IMU data for a wide range of real-world activity contexts [130], [152]. For instance, Leng *et al.* utilized LLMs to generate prompts which can subsequently processed by CLIP-based model to produce 3D human motion sequences which were converted into streams of virtual IMU data for further HAR problem [152] as shown in Figure 4.c. After that, virtual IMU data can be calculated using inverse kinematics based on these motion sequences. Leng *et al.* [130] further proposed language-based cross modality transfer system for HAR problem. Specifically, an LLMs model generate textual descriptions of activities automatically, which which are then converted into motion sequences by a motion synthesis model. After that, a motion filter was designed to screen out incorrect sequences to obtain only relevant motion sequences for virtual IMU data extraction. A diversity metric was introduced to measure the distribution shift between textual descriptions and motion sequences to determine when data generation should be stopped for most effective HAR performance.

In addition to the aforementioned approaches, another solution is to use adversarial network to synthesize data. Currently, there are many studies using VAE [162], GAN [163] and diffusion [20], [164], [165] methods for multimodal data generation purpose. For instance, Aggrawal *et al.* [165] present both offline and online approaches to generate music from video. Currently, there are several studies trying to use adversarial networks to produce synthetic time-series data for HAR problems [80], [166], [167], [168], [169], [170] as shown in Figure 4.d. For instance, GANs can augment smaller datasets by generating new, previously unseen data [170]. More recently, Ni *et al.* [146] proposed a cross-modal adversarial framework to produce corresponding synthetic skeleton joints from accelerometer data. Thus, wearable devices are capable of not only collecting time series data but also able to generate skeleton sequences for further multimodal process. This addresses the need for real-time applications of multimodal wearable HAR system that can be conducted ubiquitously. While adversarial networks have shown success in generating single time series, their potential in multimodal HAR system on wearable sensors remains largely untapped.

4.1.2 Limited Labeled Data in Existing Dataset

Unlike images or audio, where annotations can be derived directly from the raw data, annotating sensor data is a complex task for humans without the aid of post-experimental video recordings. In addition, current methods for obtaining labeled datasets often require significant human effort [174]. Therefore, current multimodal datasets for wearable HAR problem are considered as small- or medium- scale datasets. As a result, how to fully utilized such small/medium datasets is also critical for wearable HAR community.

A straightforward solution could be leveraging data augmentation approaches from existing WSHAR tasks. Data augmentation (DA) is a method used to enhance the variety of training samples without the need to gather new data. Currently, transformations in the time series domain are among the simplest and most effective data augmentation techniques for time series data. The majority of these methods directly alter the original input time series, such as time slicing window, jittering, scaling, rotation, permutation, channel permutation and so on [175]. At the same time, there are many data augmentation approaches including AutoAugment [176], MixUp [171] and CutMix [172] in CV domain as shown in Figure 5.a and 5.b. Moreover, there are many studies in multimodal learning are using data augmentation methods for various tasks [177], [178], [179], [180], [181], [182], [183]. For instance, Renduchintala *et al.* proposed a multimodal data augmentation network for automatic speech recognition. This network, comprising two distinct encoders, is capable of handling multimodal inputs, both acoustic and symbolic. It facilitates the effortless integration of extensive text datasets with considerably smaller transcribed speech corpora during the training process. This approach optimizes the use of available resources, enhancing the learning process. [177]. Xu *et al.* introduced a multimodal data augmentation framework aimed at enhancing performance in multimodal image-text classification tasks. The framework was designed to learn a

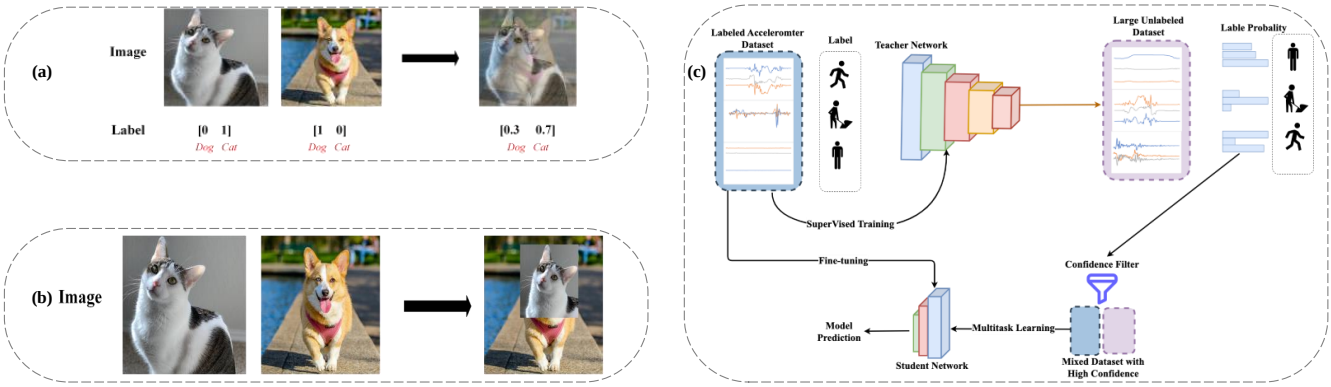

Fig. 5: Current approaches for limited labeled data problem. (a) MixUp method [171]. (b) CutMix approach [172]. (c) Self-training based SSL method [173].

cross-modality matching network, which selects image-text pairs from existing unimodal datasets to create a synthetic multimodal dataset. This dataset is then used to improve the performance of classifiers [180]. Falcon *et al.* proposed a multimodal data augmentation approach using feature space and generate new videos and captions by blending samples that are semantically alike to solve the text-video retrieval problem [178]. Liu *et al.* further present a method for learning multimodal data augmentation. This method can autonomously learn to augment multimodal data in the feature space, without any restrictions on the types of modalities or their interrelationships [181]. Oneata *et al.* employed data augmentation techniques to prompt the multimodal system to focus on visual cues. The experiment results indicated that this technique is not only conceptually more straightforward but also consistently enhances performance in a multimodal environment [179]. Hua *et al.* designed a BERT-based back-Translation Text and Entire-image multimodal model to detect fake news. The proposed framework applied data augmentation method to not only mitigate the issue of limited data, but also generates positive samples that are beneficial for the following contrastive learning module [182]. Josi *et al.* present a data augmentation method which relied on local occlusions and global modality masking methods for person re-identification problem [183]. While data augmentation have shown success in improving single modality HAR and other multimodal problems, its potential in leveraging multimodal approaches on WSHAR system remains significantly unexplored.

Another approach is to use self-supervised learning (SSL) to address the problem of limited labeled data. SSL is an effective technique where the model can learn good data representations from unlabeled data to improve downstream task performances. For instance, video data typically encompasses several modalities such as visuals, sound, and text. The temporally synchronized characteristics of these modalities offer an inherent method for deriving positive pairs that share the same time frame, eliminating the necessity for preliminary text tasks [184]. Consequently, many SSL methods have been initially developed for audio-visual learning [185], speech-visual learning [186] and other tasks [187]. For instance, Asano *et al.* introduced a clustering method that enables pseudo-labeling of a video dataset without the need for human annotations. This is achieved by capitalizing on the inherent relationship between the audio and visual modalities [185]. Alayrac *et al.* [187] proposed a multimodal versatile network which applied SSL to learn representations by leveraging visual, audio and language streams modalities from video. Specifically, they investigated the optimal way to merge these modalities, ensuring that detailed representations of both visual and audio modalities are preserved, while simultaneously incorporating text into a unified embedding. Similarly, Li *et al.* presents a distinctive cross-modal self-supervised learning technique that models the certainty of audio and visual observations by taking advantage of the complementarity and uniformity among various modalities [188].

Currently, there are also a few studies using SSL for either single modality WSHAR [173], [189], [190], [191] or multimodal [192], [193], [194], [195] HAR problem. For instance, Saeed *et al.* utilized a SSL method to learn feature from unlabeled data for HAR task. Specifically, a multi-task temporal convolutional network was trained to differentiate various transformations which performed on the raw input signals [189]. Tang *et al.* proposed a self-training based SSL method for WSHAR as shown in Figure 5.c. Initially, a teacher model was employed to distill the knowledge from labeled accelerometer data to label a large unlabeled dataset. High-confidence data points from this step were selected, and a student model was trained to discriminate these selected activities. The ground truth labels from the training labeled dataset were further used to fine-tune the student model [173]. Yuan *et al.* further leveraged SSL for one of the largest unlabeled wearable sensor dataset, UK-Biobank activity tracker dataset, for WSHAR problem [190]. Jain *et al.* presented a collaborative SSL technique which takes advantage of unlabeled data gathered from various wearable devices used by an individual to learn superior characteristics of the data [191]. In addition, Akbari *et al.* presents a Transformer-based framework which takes raw signals as input and extracts multimodal representations that are rich enough to benefit a variety of downstream tasks. This framework was trained using multimodal contrastive losses and the result demonstrated its robust performance by action recognition task [192]. After that, Choi *et al.* further proposes a hard negative sampling method for

multimodal HAR which relies on a hard negative sampling loss for skeleton and IMUs data pairs [193]. Brinzea *et al.* implements a multimodal SSL framework which can exploit modality-specific knowledge to encode inertial and skeleton data for HAR problem [194]. Deldari *et al.* introduces a cross-modal SSL which can generate masking intermediate embeddings by modality-specific encoders. After that, these embeddings can integrated into a global embedding via a cross-model aggregator for HAR task [195].

More recently, a few studies tried to apply CLIP-based framework for HAR problem [156], [196], [197]. Girdhar *et al.* present a framework to learn a joint embedding across various modalities including images, text, audio, depth, thermal and IMUs data. This framework which uses CLIP-like architecture, can extend to new modalities just by using their natural pairing with images [198]. Moon *et al.* proposed a contrastive learning approach that transforms the IMUs sensor readings and the textual annotations upon the videos of human activities into a shared embedding space to enhance information retrieval [196]. More recently, Xia *et al.* designed a cross-modal co-learning method for few-shot HAR problem. This method first utilized the semantic-rich label text to search for human activity images to form an augmented dataset consisting of partially-labeled time series and fully-labeled images. A pre-trained CLIP image encoder was used to train a time series encoder with constrative learning. After that, the feature extracted from the input time series is compared with the embedding of the pre-trained CLIP text encoder using prompt learning and the best match is output as the HAR classification results [197]. The experimental results demonstrated that the proposed method performed close to or better than the fully supervised methods even using limited labeled samples.

4.1.3 Imbalance Class Issue

The balanced distribution of classes in the training dataset is a fundamental assumption in the creation of many machine learning algorithms. Yet, this is not always the case, and an imbalance in class distribution could potentially introduce a bias that adversely affects the performance of these models. For example, it is difficult to collect datasets of actual unexpected falls due to the infrequency and varied circumstances of falls in real life. The problem is further amplified in multimodal datasets, where there could be numerous interrelations between majority and minority classes.

Generally, The issue of class imbalance in data can be tackled by data and algorithm levels. The objective of at the data level is to adjust the original datasets through resampling in order to achieve a balanced class distribution. This approach involves several different forms of resampling methods, such as undersampling the majority class or oversampling the minority class. For example, random oversampling is able to duplicate random instances until a certain class balance is reached [199]. However, this method might carry the potential risk of eliminating valuable instances or resulting in a final training dataset that is considered too small. At the same time, synthetic minority oversampling technique (SMOTE) method is another popular framework at the data level [200]. SMOTE aims to generate new instances by combining nearby instances of the same class and has demonstrated to surpass random sampling techniques in many instances [199], [201]. For instance, Junaid *et al.* compared SMOTE and its variants, such as SVMSMOTE, BorderlineSMOTE, ADASYN (Adaptive Synthetic), SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous), and SMOTEENN (Synthetic Minority Over-sampling Technique with Ensemble of Neighbors), to evaluate the effectiveness solving imbalance class problem. After conducting multiple experiments on various data modalities, they concluded that SMOTENN was the best fit for addressing the class imbalance in the training set for early detection of Parkinson's disease [201].

More recently, GAN approaches have been used as an oversampling methods [202], [203]. For instance, Lee *et al.* presented a boundary-focused GAN (BFGAN) oversampling technique aimed at selectively controlling the placement of generated samples to tackle the class imbalance problem in multimodal time series classification [203]. The proposed BFGAN incorporated a specifically designed additional label to reflect the importance of a sample's position in the data space. After considering both the multimodality and importance of a sample, the BFGAN generated synthetic samples using GAN approaches. Similarly, Li *et al.* utilized a GAN approach for synthesizing samples, even when all multimodal features are missing, to address the issue of imbalanced multimodal data [204]. Currently, there are a few studies trying to address class imbalance issue in the HAR field [205], [206]. Guo *et al.* introduce a dual-ensemble class imbalance learning method. In this method, an internal ensemble learning model which include several heterogeneous sub-classifier was designed. The one with the highest recognition accuracy is selected as the base classifier. Sequentially, multimodal evolutionary algorithms were presented to find the optimal combination that contains the smallest number of base classifiers while accurately identifying human actions [205]. Furthermore, the proposed distillation multiple choice learning framework by Garcia *et al.* addresses the HAR problem by enabling different modality networks to learn cooperatively from scratch. This cooperative learning approach leads to significantly higher accuracy compared to training the networks separately, as each modality benefits from the complementary information offered by the multimodal data [206].

4.2 Heterogeneous Feature Alignment Challenge

Typically, single modality representation involves a linear or nonlinear mapping of an individual input stream (*e.g.*, image, video, or sound, etc.) into a high-level semantic representation. Multimodal representation, however, combines the correlation power of each single modality sensation by aggregating their spatial outputs. Despite this, current DL models often struggle to accurately represent the structure and representation space of both the source and target modality. In this section, we will delve into insight solutions aimed at addressing the challenges of feature alignment in multimodal HAR systems, providing a comprehensive understanding of this complex domain.

4.2.1 Cross-model Transfer Learning

One of the solutions for feature alignment is transfer learning. However, transfer learning heavily relies on whether

the underlying domains or tasks across the source and target domains are the same [207]. Consequently, many studies worked on heterogeneous transfer learning which refers to the case where the source and target feature spaces differ for HAR problem [208], [209], [210], [211], [212], [213], [214], [215], [216], [217], [218]. For instance, Wei *et al.* proposed a co-regularized heterogeneous transfer learning model, which built a common semantic space derived from social media and labeled physical sensor data [208]. Wang *et al.* present an unsupervised source selection algorithm for HAR problem. The most similar k source domains from a list of available domains was selected first. After that, the time and spatial relationship between activities were captured using a transfer neural network to perform knowledge transfer for activity recognition [209]. Wang *et al.* designed a stratified transfer learning method which can dramatically improve the classification accuracy for cross-domain activity recognition. Specifically, it first utilizes majority voting technique to capture pseudo labels from the target domain. After that, it performed intra-class knowledge transfer iteratively to transform both domains into the same subspaces. The labels of target domain are eventually obtained via the second annotation for final transfer learning stage [210]. Qin *et al.* present an adaptive spatial-temporal transfer learning method to adaptively evaluate the relative importance between the marginal and conditional probability distributions in spatial features. It also adopted an incremental manifold learning to capture temporal features for cross-dataset activity recognition [219]. Lu *et al.* introduced an optimal transport-based method to better utilize the locality information of activity data for cross-domain HAR for accurate and efficient knowledge transfer [211]. It utilized clustering methods to capture the substructures of activities and sought the coupling of the weighted substructures between different domains.

More recently, Yuan *et al.* designed a multimodal contrastive training method for visual representation learning. It exploited intrinsic data properties within each modality and semantic information from cross-modal correlation simultaneously, hence improving the quality of learned visual representations [212]. Sung *et al.* built an adapter-based parameter-efficient transfer learning techniques for vision-and-language tasks. It employed a unified format and architecture to solve the tasks in a multi-tasking learning setup [213]. Khaertdinov *et al.* applied a dynamic time warping (DTW) algorithm in a latent space to force features to be aligned in a temporal dimension [214]. Geng *et al.* proposed a simple and scalable multimodal masked autoencoder architecture to learn a unified encoder for both vision and language data via masked token prediction. The experiment results indicated that this architecture is able to learn generalizable representations that transfer well to downstream classification tasks [215]. Lu *et al.* designed a semantic-discriminative Mixup approach which considers the activity semantic ranges to overcome the semantic inconsistency brought by domain differences for generalizable cross-domain HAR problem. In addition, they also introduced the large margin loss to enhance the discrimination of Mixup to prevent misclassification brought by noisy virtual labels [216]. Thukral *et al.* introduces a cross-domain HAR transfer learning framework which follows the teacher-student self-training paradigm to more effectively recognize activities with very limited label information. It can bridge conceptual gaps between source and target domains, including sensor locations and type of activities [217].

4.2.2 Multimodal Knowledge Distillation

Another solution to solve the heterogeneous feature alignment issue is to use KD method, which tries to transfer knowledge from a complicated pre-trained network (*i.e.*, teacher model) to a smaller network (*i.e.*, student model) by minimizing the Kullback-Leibler (KL) divergence of predictions between teacher and student models [91]. By mimicking the accuracy performance from teacher model, the student model can eventually improve its performance [91]. At present, there are several studies applying KD for multimodal learning tasks [220], [221], [222]. For instance, Hoffman *et al.* proposed a modality hallucination architecture that uses depth as side information to guide an RGB object detection model [220]. In order to facilitate feature alignment, same multi-layer CNNs networks were selected for the whole hallucination architecture as the baselines. Garcia *et al.* designed a multimodal KD framework that utilizes both depth and RGB videos to learn representations. During testing, images are simultaneously processed by both the RGB and hallucination networks to enhance detection performance. This allows the proposed method to transfer information typically derived from depth training data to a network capable of extracting similar information from RGB data [221]. The entire framework utilizes the ResNet-50-based model as the baseline architecture for each stream block within the framework to mitigate modality discrepancies. Similarly, another work learned sound presentations by transferring knowledge from video to sound modality [223]. Similar CNN architectures were employed for both the sound and video recognition networks to promote the extraction of similar features, enhancing knowledge transfer between heterogeneous modalities. Andonian *et al.* adopted progressive self-distillation and soft image-text alignments to more efficiently learn robust representations from noisy data for cross-modal contrastive learning. The framework distilled its own knowledge to dynamically generate soft-alignment targets for a subset of images and captions in every mini-batch, which were subsequently used to update its parameters [222].

More recently, Thoker *et al.* proposed a multimodal KD framework for the HAR task. They used RGB videos to train the teacher CNNs network and then trained two student CNNs networks were trained using mutual learning to improve the performance [224]. Furthermore, Quan *et al.* proposed a Semantic-aware Multimodal Transformer Fusion Decoupled Knowledge Distillation (SMTDKD) method, enhancing video data recognition by facilitating information interaction not only between different wearable sensor data but also between visual sensor data and wearable sensor data. To address modality discrepancies and promote the extraction of similar semantic features, graph cross-view attention maps were constructed across different convolutional layers to improve the feature alignment process for HAR problem [225]. More recently, Ni *et al.* present the first multimodal KD approach for the WSHAR problem [93]. In this study, to enable visual recognition of time series data

from IMU sensors, one-dimensional action data from wearable sensors were transformed into visual representations to preserve the local temporal relation. Subsequently, an effective transfer of complementary information from the video domain to the sensor domain was achieved using the same visual VGG-16 models. In order to eliminate the privacy concern from video streams, they further adopted skeleton sequence modality as the teacher model to distill knowledge to time-series modality for accurate WSHAR problem [14]. Specifically, both the teacher and student models employed identical GNN architectures to address the feature alignment problem.

4.3 Model Deployment Challenge

While deep learning approaches have demonstrated promising performance in recognizing human activities through IMUs data, they often require substantial resources. Furthermore, wearable technology typically possesses constrained computational capabilities, which obstructs the extensive implementation of models. For instance, a prior study demonstrated that a smartphone (LG Nexus 5X, 1.8 GHz, Hexa-core processor with 2G of RAM) can only sustain a long short-term memory (LSTM) model comprising an input layer, two hidden layers, and an output layer [229]. Therefore, addressing the challenge of significant computational expense is crucial for enabling instantaneous and dependable recognition of human activities on mobile devices using advanced multimodal techniques.

4.3.1 Multimodal Model Compression

At present, there exist numerous recognized methods for the purpose of model compression. These include, but are not limited to, techniques such as pruning, knowledge distillation, quantization, and low-rank factorization [226] as shown in Figure 6.a and 6.b. These approaches aim to convert large, resource-intensive models into smaller versions suitable for storage on resource-limited mobile devices. Nonetheless, present compression techniques are mainly constructed on the basis of unimodal networks or particular model architectures, making it challenging to broaden these methods to a variety of multimodal learning methods. A straightforward approach involves training the large-scale model first and then applying post-compression techniques to reduce its size [230]. For instance, Nooruddin *et al.* [230] first proposed a two-stream multi-resolution fusion architecture for HAR problem. Two quantization approaches, such as post-training quantization and quantization aware training were further introduced to optimize these models for deployment in edge devices.

Meanwhile, attention mechanism was adopted for lightweight model design process [231], [232], [233]. For example, Zhou *et al.* present a lightweight model which included a cross-chanel interaction Transformer encoder and global temporal extraction layer for HAR task [231]. The experiment results showed that the proposed model even surpassed the optimized DeepConvLSTM [234] with reduced model size by more than 93% on several datasets. Gao *et al.* present a multimodal temporal segment attention network for HAR problem using RGB video and IMU data. This network was tested on a Raspberry Pi 4B, which was equipped with a 64-bit 1.5GHz quad-core CPU and 8GB of RAM. The results of the experiment showed that the network had significantly reduced FLOPs and latency, making it more suitable for edge deployment [232]. Cai *et al.* present an adaptive compression framework to address the computational resource challenges which enable input-dependent runtime compression locally on resource-constrained embedded devices [233]. Specifically, They propose an offline model transformation module to upgrade the static network with two kinds of dynamic components to support online structural adjustment. A lightweight policy network was designed to generate multi-granularity and data-dependent compression strategies for different model parts.

Another approach is to use specialized toolsets for simplifying complicated models, such as TensorFlow Lite, Caffe2, Pytorch Mobile and TensorRT [235], [236], [237], [238]. For instance, TensorFlow Lite is a dedicated suite of tools designed for use on mobile and IoT devices. It provides post-training quantization methods which reduced the model weights using fewer bites instead of full floating-point numbers [235]. Consequently, Bursa *et al.* conducted performance evaluations on a variety of model architectures that were converted using TensorFlow Lite for HAR problem. The experiment results revealed that the application of quantization methods in TensorFlow Lite led to a substantial reduction in model sizes. Importantly, this size reduction did not compromise the accuracy of the models, demonstrating the effectiveness of TensorFlow Lite's optimization techniques [236]. Mazzia *et al.* proposed a self-attentional architecture that leverages pose representations over small temporal windows. Once converted using TensorFlow Lite, the method provides a low-latency solution that ensures both accuracy and efficiency in real-time performance [237]. In [238], to demonstrate real-world deployment and applications, the authors utilized TensorRT, an SDK for high-performance DL inference, to convert the trained GNN HAR models based on PyTorch into a TensorRT model for Jetson AGX Xavier and into CoreML for iPhone XR. Experimental results on latency showed that models compressed by TensorRT could reduce latency by approximately 10% and 50% on AGX and iPhone XR, respectively.

4.3.2 Advanced Lightweight Model

While contemporary DL methodologies have made great progress, they continue to face substantial energy requirement challenges which associated with the training and inference processes. As a result, there has been a surge of interest in developing low-power techniques [239], such as brain-inspired spiking neural networks (SNNs) as shown in Figure 6.c, which leads to high energy efficiency. This is largely due to their event-driven nature, where computations are performed only when events (or spikes) occur [227]. Currently, there are several studies using SNNs for multimodal learning tasks [240], [241], [242], [243], [244]. For instance, Sengupta *et al.* proposed an SNN to fuse temporal, spatial, and orientation data for multimodal brain data modeling. The proposed framework was assessed qualitatively and quantitatively using artificially created data to understand its behavior and its capacity to incorporate spatial, temporal, and orientation data [240]. Liu *et al.* integrated SNNs with attention mechanism to fuse visual and auditory

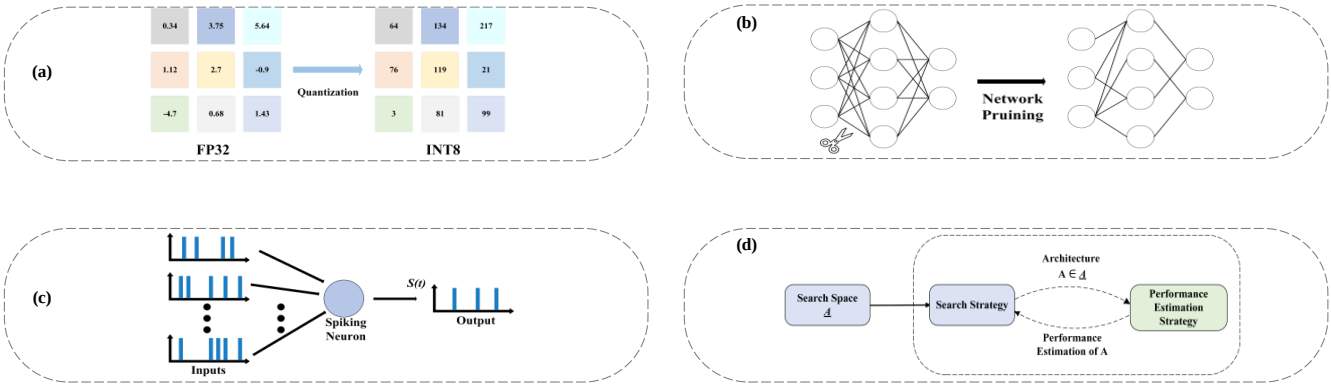

Fig. 6: Advanced approaches for model deployment problem. (a) Weight quantization approach [226]. (b) Network pruning method [226]. (c) Spiking neural network architecture [227]. (d) Neural architecture search framework [228].

data. This attention feature evaluates the importance of each modality and subsequently distributes weights between the two modalities [241]. Wang *et al.* adopted SNNs to merge RGB frames and event streams concurrently for a pattern recognition problem [242]. Wang *et al.* designed an event-enhanced multimodal spiking actor network based on deep reinforcement learning. This network combined data from both the Laser and event camera to extract and fuse more effective information [243]. Guo *et al.* proposed a framework that combined SNNs with Transformer architectures for a multimodal audiovisual classification problem. This framework achieved commendable accuracy in multimodal classification tasks while maintaining low energy usage, positioning it as an efficient and effective solution for such classification tasks [244].

Meanwhile, there are a couple of studies utilizing SNNs for HAR problem [245], [246], [247]. For example, Fra *et al.* compared optimized classifiers based on traditional DL architectures and demonstrated the efficiency of SNNs in processing time-dependent signals for HAR problems by not only yielding high performances at a low energy cost [245]. Specifically, spiking CNNs exhibited the lowest energy consumption at 5.49 μJ, nearly three orders of magnitude less than the corresponding CNN models on Intel's Loihi. Khan *et al.* integrated SNNs with LSTM networks to achieve energy efficiency and preserve privacy in HAR problem [246]. Specifically, the proposed spiking LSTM showed a significant improvement in energy efficiency of 32.30%, compared to simple LSTM. Li *et al.* further present a continuous authentication system which employed SNNs to analyze biometric behavioral patterns recorded by smartphone sensors [247]. To achieve the conversion of ANNs to SNNs, weights and activations were mapped to generate suitable spike neuron models and synaptic connections with higher accuracy performance.

4.3.3 Neural Architecture Search

Neural architecture search (NAS) is a process designed to automatically identify the most effective neural network structures that deliver optimal performance while using minimal computational resources [228], [248], [249] as shown in Figure 6.d. Currently, several studies utilize NAS for multimodal tasks [250], [251], [252], [253], [254]. For instance, Perez *et al.* introduced a generic search space that encompassed a wide range of potential fusion architectures. A sequential model-based exploration method was designed to find the optimal architecture in the proposed search space. The experiment results demonstrated the benefits of framing multimodal fusion as a problem of neural architecture search [250]. Yu *et al.* devised a generalized NAS framework across several multimodal learning tasks, including visual question answering, image-text matching, and visual grounding [251]. The framework was based on a deep encoder-decoder, where each block of the encoder or decoder corresponds to an operation selected from a pre-established operation pool. By employing a gradient-based NAS algorithm, they efficiently learned optimal architectures for different tasks. Xu *et al.* proposed a NAS algorithm to simultaneously search across multimodal fusion strategies and modality-specific architectures for electronic health records diagnosis code prediction [252]. Shi *et al.* proposed an efficient automatic speech recognition method that benefits from the natural advantage of differentiable NAS in reducing computational overheads. This differentiable architecture search method was fused with Conformer blocks to form a complete search space [253]. Si *et al.* designed a multimodal fusion architecture search framework to automatically design promising multimodal fusion architectures for violence detection tasks [254]. Specifically, multilayer neural networks based on attention mechanisms are meticulously constructed to grasp intricate spatio-temporal relationships and extract comprehensive multimodal representation.

More recently, NAS methods have been applied in HAR domain [255], [256]. For instance, Wang *et al.* [255] adopted a multi-objective NAS method to solve the tradeoff problem between high efficiency and high performance. This framework was extended to a tri-objective task where the search targets were based on the weighted F1 score, the number of FLOPs, and its memory use. Lim *et al.* proposed a mobile HAR NAS based on a differentiable neural architecture search for automatic design of the architecture of a HAR model for a mobile device [256]. Experiments were also conducted with the Galaxy A31 and Galaxy S10 smartphones as target devices. the latency of the A31-optimized model was, on average, 2% faster than that of the S10-optimized model on the A31 device.

5 Future Research Direction

Despite the efforts devoted to these above-mentioned challenges, some of them are still not fully explored, such as heterogeneous feature alignment, lightweight model deployment, and so on. While existing research may not offer complete and dependable solutions to these challenges, they do provide a solid base and valuable insights for future work. Additionally, there are several other challenges that have been scarcely investigated before and require immediate exploration. In the following section, we highlight several pivotal research directions that urgently need exploration. It is our hope that the challenges identified in this study can serve as catalysts for these future explorations.

5.1 Future Activity Prediction

Future activity prediction can be considered as an extension version of current HAR problem. Unlike activity recognition, which identifies current actions, the predictive system can anticipate user behaviors beforehand. This system plays a crucial role in understanding human intentions, thereby finding applications in smart services, crime detection, and driver behavior prediction. In some common behavior tasks, the activities are usually in a certain order. Therefore, modeling the temporal dependencies across activities is beneficial to predict future predictions. Cross-modal knowledge distillation framework [257] is suitable for such tasks. But for long-span activities captured from partial video, KD cannot achieve such long dependencies due to the limited context information. In this case, adversarial KD-based approach based on generative network can assist to solve the early action prediction task [258]. Moreover, LLMs have shown considerable potential in identifying patterns, predicting future events, and detecting anomalous behavior across diverse domains [259], [260]. Consequently, exploring the potential of leveraging LLMs for future WSHAR prediction emerges as a promising avenue for research.

5.2 Identifying Unknown Activities

Discovering unprecedented actions that remain unobserved by the models poses a substantial obstacle in HAR domain. Hence, it becomes imperative to examine the capacity of models to adjust to dynamic environments and prevent the disastrous loss of previously acquired knowledge. In fact, an effective model should have the ability to acquire new insights in an online manner and execute accurate discernment in the absence of ground truth. A promising way to enable models to continuously adapt to dynamic input data, is continue learning [261]. However, how to build models with the ability to perpetually adapt to multi-model data still an under-explored problem. Moreover, LLMs have shown significant achievements to directly comprehend visual signals. For example, LLMs can fundamentally considers images as linguistic entities, translating them into discrete words from the LLM's vocabulary [262]. More recently, the semantic space of LLMs has demonstrated to be able to guide time series embeddings by maximizing the cosine similarity in the joint space [263]. Therefore, exploring LLMs fully for identifying unseen activities can be another promising research area for HAR improvements.

5.3 New Foundation Models

Currently, Transformers have dominated the HAR domain [8]. However, one of the disadvantages of Transformer model is the computational inefficiency on long sequences data. As a result, a new model called Mamba has been proposed to solve this problem [264]. In this study, a selection mechanism to structured state space models is designed to engage in context-dependent reasoning while maintaining linear scalability in sequence length. Mamba and its variants have showcased the extensive applicability of selective state space models in modalities requiring extensive context, such as audio, image, and video [265], [266], [267]. Meanwhile, RWKV [268] integrated the efficient parallelizable training of Transformers with the effective inference capabilities of RNNs, maintaining constant computational and memory complexity during inference. RWKV and its variants have exhibited comparable performance as well as promising latency and memory utilization efficiency [269], [270], [271]. Further research into the application of these foundation models in the WSHAR domain could potentially lead to the development of more robust models.

5.4 Unified Multimodal Systems

Existing papers that apply multimodal learning approaches for wearable HAR problem mainly concentrates on the fusion of diverse inputs from different modalities and a single task at a time, such as forecasting and classification. However, these studies do not facilitate the simultaneous analysis of multimodal and multitask scenarios. In the domains of CV, NLP and audio, models such as Unified-io [272], AnyGPT [273], and UniAudio [274] have integrated multiple input modalities to support the execution of multiple tasks within a singular Transformer-based architecture. Omni-Dimensional INstance segmentation (ODIN) method was proposed to segment and label both 2D RGB images and 3D point clouds [275]. This approach utilizes a Transformer architecture that alternates between 2D within-view and 3D cross-view information fusion. Similarly, a one-for-all model, called UniST [276], was proposed to solve for urban spatio-temporal prediction problem. More recently, UniTS [277] showed its capable of handling various tasks such as forecasting and anomaly detection through a universal task specification. Further research into the application of multimodal and multitask analysis in the HAR domain could potentially lead to the development of more potent time series foundation models.

5.5 Concurrent Activity Segmentation

Human activities naturally exhibit a hierarchical structure, as seen in daily routines that include a variety of tasks such as washing, grooming, and eating. These tasks can be further divided into specific actions like washing dishes or washing hands. However, compiling a large-scale dataset of everyday activities with detailed annotations is a challenging task, mainly due to the time-consuming requirement for manual annotations. Furthermore, distinguishing between activities with comparable performance trends poses a challenge for online models in wearable HAR domain. In CV and audio domain, models such as SAM [278] and AV-SAM

[279] achieved amazing segmentation performance and how to leverage such method to wearable HAR area is a promising direction for future research. Furthermore, LLMs can be employed to generate descriptions of both overarching routines and their associated detailed actions, potentially serving as annotations for the synthesized virtual IMU data. This advanced hierarchical dataset can be further utilized to train a model capable of identifying and segmenting various activity levels, thereby enhancing the performance of WSHAR systems.

5.6 Personalization and Privacy

The majority of existing research on wearable HAR and time series analysis usually focuses on a global model for all users. However, the development of personalized models for individual users, derived from the global model, could potentially offer additional advantages and adaptability. This approach could lead to more tailored solutions that better meet the unique needs of each user. Moreover, privacy is indeed a crucial factor, particularly as a significant amount of time series data is gathered in private contexts for purposes such as clinical applications or smart home technologies. As a result, federated learning is applied to build personalized model for wearable HAR problem using single modality [280]. However, the task of utilizing federated learning frameworks in the context of multimodal HAR remains a complex and less explored area. Meanwhile, style transfer in CV area involves generating a new image by combining the content of one image with the style of another image [281]. The goal of style transfer is to create an image that preserves the content of the original image while applying the visual style of another image. As a result, style transfer approach can be a valuable direction for personalized sample regeneration in time series domain. In fact, advancing research into multimodal personalized model and user privacy preservation would broaden the scope and utility of multimodal HAR problem.

6 Conclusion

We present the first survey that systematically analyzes the WSHAR field from the perspective of multimodal learning approaches. Initially, we discuss recent advancements in sensor modalities and the latest deep learning approaches for HAR. Then, we explore recent techniques used in present multimodal systems for WSHAR, covering both inter-multimodal systems utilizing sensor modalities from visual and non-visual systems, and intra-multimodal systems using modalities from non-visual systems only. In addition to providing a comprehensive summary of existing multimodal datasets for WSHAR, we also discuss the accomplishments of multimodal approaches in addressing some of the challenges in the WSHAR field. By connecting the existing multimodal achievements from other tasks, such as CV and NLP domains, we have laid the groundwork for discussions on the existing challenges and potential future directions. This paper concludes with final remarks that encapsulate the essence of our findings and discussions. We hope that our work will inspire further research in this exciting and rapidly evolving research field.

References

[1] U. DESA, "World population prospects: Key findings and advance tables," *New York: UN DESA*, 2017.

[2] A. Kuerbis, A. Mulliken, F. Muench, A. A. Moore, and D. Gardner, "Older adults and mobile technology: Factors that enhance and inhibit utilization in the context of behavioral health," 2017.

[3] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, p. 106970, 2021.

[4] Y. Liu, K. Wang, G. Li, and L. Lin, "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *TIP*, vol. 30, pp. 5573–5588, 2021.

[5] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos, "Multimodal human action recognition in assistive human-robot interaction," in *ICASSP*. IEEE, 2016, pp. 2702–2706.

[6] M. Á. Á. de la Concepción, L. M. S. Morillo, J. A. Á. García, and L. González-Abril, "Mobile activity recognition and fall detection system for elderly people using ameva algorithm," *PMC*, vol. 34, pp. 3–13, 2017.

[7] N. Maray, A. H. Ngu, J. Ni, M. Debnath, and L. Wang, "Transfer learning on small datasets for improved fall detection," *Sensors*, vol. 23, no. 3, p. 1105, 2023.

[8] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *TPAMI*, 2022.

[9] L. Lyu, X. He, Y. W. Law, and M. Palaniswami, "Privacy-preserving collaborative deep learning with application to human activity recognition," in *CIKM*, 2017, pp. 1219–1228.

[10] S. Zhang, Y. Li, S. Zhang, F. Shahabi, S. Xia, Y. Deng, and N. Alshurafa, "Deep learning in human activity recognition with wearable sensors: A review on advances," *Sensors*, vol. 22, no. 4, p. 1476, 2022.

[11] A. Galán-Mercant, A. Ortiz, E. Herrera-Viedma, M. T. Tomas, B. Fernandes, and J. A. Moral-Munoz, "Assessing physical activity and functional fitness level using convolutional neural networks," *Knowledge-Based Systems*, vol. 185, p. 104939, 2019.

[12] K. Van Laerhoven, M. Borazio, and J. H. Burdinski, "Wear is your mobile? investigating phone carrying and use habits with a wearable device," *Frontiers in ICT*, vol. 2, p. 10, 2015.

[13] R. Rawassizadeh, B. A. Price, and M. Petre, "Wearables: Has the age of smartwatches finally arrived?" *Communications of the ACM*, vol. 58, no. 1, pp. 45–47, 2014.

[14] J. Ni, A. H. Ngu, and Y. Yan, "Progressive cross-modal knowledge distillation for human action recognition," in *ACM MM*, 2022, pp. 5903–5912.

[15] T. R. Mauldin, M. E. Canby, V. Metsis, A. H. Ngu, and C. C. Rivera, "Smartfall: A smartwatch-based fall detection system using deep learning," *Sensors*, vol. 18, no. 10, p. 3363, 2018.

[16] Y. Wang, S. Cang, and H. Yu, "A survey on wearable sensor modality centred human activity recognition in health care," *Expert Systems with Applications*, vol. 137, pp. 167–190, 2019.

[17] I. Bornkessel-Schlesewsky, M. Schlesewsky, S. L. Small, and J. P. Rauschecker, "Neurobiological roots of language in primate audition: common computational properties," *Trends in cognitive sciences*, vol. 19, no. 3, pp. 142–150, 2015.

[18] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, p. 107561, 2020.

[19] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[20] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *CSUR*, vol. 43, no. 3, pp. 1–43, 2011.

[21] O. D. Incel, M. Kose, and C. Ersoy, "A review and taxonomy of activity recognition on mobile phones," *BioNanoScience*, vol. 3, no. 2, pp. 145–171, 2013.

[22] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192–1209, 2012.

[23] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *CSUR*, vol. 46, no. 3, pp. 1–33, 2014.

[24] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimedia Tools and Applications*, vol. 79, no. 41-42, pp. 30 509–30 555, 2020.

[25] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A survey on deep learning for human activity recognition," *CSUR*, vol. 54, no. 8, pp. 1–34, 2021.

[26] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *CSUR*, vol. 54, no. 4, pp. 1–40, 2021.

[27] E. Ramanujam, T. Perumal, and S. Padmavathi, "Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review," *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13 029–13 040, 2021.

[28] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *IJCV*, vol. 130, no. 5, pp. 1366–1401, 2022.

[29] S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, and G. Fortino, "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," *Information Fusion*, vol. 80, pp. 241–265, 2022.

[30] S. G. Dhekane and T. Ploetz, "Transfer learning in human activity recognition: A survey," *arXiv preprint arXiv:2401.10185*, 2024.

[31] S. Nerella, S. Bandyopadhyay, J. Zhang, M. Contreras, S. Siegel, A. Bumin, B. Silva, J. Sena, B. Shickel, A. Bihorac *et al.*, "Transformers in healthcare: A survey," *arXiv preprint arXiv:2307.00067*, 2023.

[32] L. Yang, O. Amin, and B. Shihada, "Intelligent wearable systems: Opportunities and challenges in health and sports," *CSUR*, 2024.

[33] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *ICCV*, 2019, pp. 6202–6211.

[34] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *ICCV*, 2019, pp. 7083–7093.

[35] J. Yang, X. Dong, L. Liu, C. Zhang, J. Shen, and D. Yu, "Recurring the transformer for video action recognition," in *CVPR*, 2022, pp. 14 063–14 073.

[36] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang, C. Sun, and C. Schmid, "Multiview transformers for video recognition," in *CVPR*, 2022, pp. 3333–3343.

[37] C. Zhang, A. Gupta, and A. Zisserman, "Helping hands: An object-aware ego-centric video recognition model," in *ICCV*, 2023, pp. 13 901–13 912.

[38] T. Shiota, M. Takagi, K. Kumagai, H. Seshimo, and Y. Aono, "Egocentric action recognition by capturing hand-object contact and object state," in *WACV*, 2024, pp. 6541–6551.

[39] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[40] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019, pp. 12 026–12 035.

[41] H. Qiu, B. Hou, B. Ren, and X. Zhang, "Spatio-temporal tuples transformer for skeleton-based action recognition," *arXiv preprint arXiv:2201.02849*, 2022.

[42] H. Duan, M. Xu, B. Shuai, D. Modolo, Z. Tu, J. Tighe, and A. Bergamo, "Skeletr: Towards skeleton-based action recognition in the wild," in *ICCV*, 2023, pp. 13 634–13 644.

[43] G. Laput, K. Ahuja, M. Goel, and C. Harrison, "Ubicoustics: Plug-and-play acoustic activity recognition," in *UIST*, 2018, pp. 213–224.

[44] D. Liang and E. Thomaz, "Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos," *IMWUT*, vol. 3, no. 1, pp. 1–18, 2019.

[45] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "Wifi csi based passive human activity recognition using attention based blstm," *TMC*, vol. 18, no. 11, pp. 2714–2724, 2018.

[46] B. Sheng, F. Xiao, L. Sha, and L. Sun, "Deep spatial–temporal model based cross-scene action recognition using commodity wifi," *IoT-J*, vol. 7, no. 4, pp. 3592–3601, 2020.

[47] J. Wang, Q. Long, K. Liu, Y. Xie *et al.*, "Human action recognition on cellphone using compositional bidir-lstm-cnn networks," in *CNCI 2019*. Atlantis Press, 2019, pp. 687–692.

[48] K. Xia, J. Huang, and H. Wang, "Lstm-cnn architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56 855–56 866, 2020.

[49] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *TPAMI*, vol. 23, no. 3, pp. 257–267, 2001.

[50] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *arXiv preprint arXiv:1406.2199*, 2014.

[51] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE access*, vol. 6, pp. 1155–1166, 2017.

[52] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *CVPR*, 2015, pp. 287–295.

[53] A. Núñez-Marcos, G. Azkune, and I. Arganda-Carreras, "Egocentric vision-based action recognition: A survey," *Neurocomputing*, vol. 472, pp. 175–197, 2022.

[54] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *ICCV*. IEEE, 2011, pp. 407–414.

[55] N. Aboubakr, J. L. Crowley, and R. Ronfard, "Recognizing manipulation actions from state-transformations," *arXiv preprint arXiv:1906.05147*, 2019.

[56] T. Nagarajan, Y. Li, C. Feichtenhofer, and K. Grauman, "Ego-topo: Environment affordances from egocentric video," in *CVPR*, 2020, pp. 163–172.

[57] Y. Poleg, A. Ephrat, S. Peleg, and C. Arora, "Compact cnn for indexing egocentric videos," in *WACV*. IEEE, 2016, pp. 1–9.

[58] S. Narayan, M. S. Kankanhalli, and K. R. Ramakrishnan, "Action and interaction recognition in first-person videos," in *CVPR Workshops*, 2014, pp. 512–518.

[59] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *CVPR*, 2015, pp. 1110–1118.

[60] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," in *WACV*. IEEE, 2017, pp. 148–157.

[61] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction," *TPAMI*, 2021.

[62] F. Al Machot, M. R. Elkobaisi, and K. Kyamakya, "Zero-shot human activity recognition using non-visual sensors," *Sensors*, vol. 20, no. 3, p. 825, 2020.

[63] J. Chen, A. H. Kam, J. Zhang, N. Liu, and L. Shue, "Bathroom activity monitoring based on sound," in *PERVASIVE*. Springer, 2005, pp. 47–61.

[64] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *ICASSP*, vol. 14, no. 1, pp. 321–329, 2005.

[65] N. D. Lane, P. Georgiev, and L. Qendro, "Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *UbiComp*, 2015, pp. 283–294.

[66] K. Yatani and K. N. Truong, "Bodyscope: a wearable acoustic sensor for activity recognition," in *UbiComp*, 2012, pp. 341–350.

[67] X. Wu, Z. Chu, P. Yang, C. Xiang, X. Zheng, and W. Huang, "Tw-see: Human activity recognition through the wall with commodity wi-fi devices," *TVT*, vol. 68, no. 1, pp. 306–319, 2018.

[68] J. Wang, X. Zhang, Q. Gao, H. Yue, and H. Wang, "Device-free wireless localization and activity recognition: A deep learning approach," *TVT*, vol. 66, no. 7, pp. 6258–6267, 2016.

[69] Q. Gao, J. Wang, X. Ma, X. Feng, and H. Wang, "Csi-based device-free wireless localization and activity recognition using radio image features," *TVT*, vol. 66, no. 11, pp. 10 346–10 356, 2017.

[70] Y. Chen and Y. Xue, "A deep learning approach to human activity recognition based on single accelerometer," in *SMC*. IEEE, 2015, pp. 1488–1492.

[71] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[72] J. Lu and K.-Y. Tong, "Robust single accelerometer-based activity recognition using modified recurrence plot," *IEEE Sensors Journal*, vol. 19, no. 15, pp. 6317–6324, 2019.

[73] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *ACM MM*, 2015, pp. 1307–1310.

[74] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," *THMS*, vol. 45, no. 1, pp. 51–61, 2014.

[75] N. E. D. Elmadany, Y. He, and L. Guan, "Multimodal learning for human action recognition via bimodal/multimodal hybrid centroid canonical correlation analysis," *TMM*, vol. 21, no. 5, pp. 1317–1331, 2018.

[76] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[77] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *ICCV*, 2015, pp. 4534–4542.

[78] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *CVPR*, 2019, pp. 1505–1514.

[79] T. I. Denk, Y. Takagi, T. Matsuyama, A. Agostinelli, T. Nakai, C. Frank, and S. Nishimoto, "Brain2music: Reconstructing music from human brain activity," *arXiv preprint arXiv:2307.11078*, 2023.

[80] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *ICCV*, 2023, pp. 4195–4205.

[81] H. Wei, R. Jafari, and N. Kehtarnavaz, "Fusion of video and inertial sensing for deep learning–based human action recognition," *Sensors*, vol. 19, no. 17, p. 3680, 2019.

[82] Z. Ahmad and N. Khan, "Human action recognition using deep multilevel multimodal (M^2) fusion of depth and inertial sensors," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1445–1455, 2019.

[83] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *JBHI*, vol. 21, no. 1, pp. 4–21, 2016.

[84] T. Von Marcard, G. Pons-Moll, and B. Rosenhahn, "Human pose estimation from video and imus," *TPAMI*, vol. 38, no. 8, pp. 1533–1547, 2016.

[85] T. Li, S. Fong, K. K. Wong, Y. Wu, X.-s. Yang, and X. Li, "Fusing wearable and remote sensing data streams by fast incremental learning with swarm decision table for human activity recognition," *Information Fusion*, vol. 60, pp. 41–64, 2020.

[86] C. M. Ranieri, S. MacLeod, M. Dragone, P. A. Vargas, and R. A. F. Romero, "Activity recognition for ambient assisted living with videos, inertial units and ambient sensors," *Sensors*, vol. 21, no. 3, p. 768, 2021.

[87] M. Ijaz, R. Diaz, and C. Chen, "Multimodal transformer for nursing activity recognition," in *CVPR Workshop*, 2022, pp. 2065–2074.

[88] N. Dawar and N. Kehtarnavaz, "A convolutional neural network-based sensor fusion system for monitoring transition movements in healthcare applications," in *ICCA*. IEEE, 2018, pp. 482–485.

[89] A. Das, P. Sil, P. K. Singh, V. Bhateja, and R. Sarkar, "Mmharensemnet: a multi-modal human activity recognition model," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11 569–11 576, 2020.

[90] Z. Qin, Y. Zhang, S. Meng, Z. Qin, and K.-K. R. Choo, "Imaging and fusing time series for wearable sensor-based human activity recognition," *Information Fusion*, vol. 53, pp. 80–87, 2020.

[91] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[92] Q. Kong, Z. Wu, Z. Deng, M. Klinkigt, B. Tong, and T. Murakami, "Mmact: A large-scale dataset for cross modal human action understanding," in *ICCV*, 2019, pp. 8658–8667.

[93] J. Ni, R. Sarbajna, Y. Liu, A. H. Ngu, and Y. Yan, "Cross-modal knowledge distillation for vision-to-sensor action recognition," in *ICASSP*. IEEE, 2022, pp. 4448–4452.

[94] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *AAAI*, vol. 34, no. 04, 2020, pp. 5191–5198.

[95] Y. Tian, C. Zhang, Z. Guo, X. Zhang, and N. Chawla, "Learning mlps on graphs: A unified view of effectiveness, robustness, and efficiency," in *ICLR*, 2022.

[96] Y. Tian, S. Pei, X. Zhang, C. Zhang, and N. V. Chawla, "Knowledge distillation on graphs: A survey," *arXiv preprint arXiv:2302.00219*, 2023.

[97] M. Gabel, R. Gilad-Bachrach, E. Renshaw, and A. Schuster, "Full body gait analysis with kinect," in *EMBC*. IEEE, 2012, pp. 1964–1967.

[98] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *WACV*. IEEE, 2013, pp. 53–60.

[99] B. Delachaux, J. Rebetez, A. Perez-Uribe, and H. F. Satizábal Mejia, "Indoor activity recognition by combining one-vs.-all neural network classifiers exploiting wearable and depth sensors," in *IWANN*. Springer, 2013, pp. 216–223.

[100] K. Liu, C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of inertial and depth sensor data for robust hand gesture recognition," *IEEE Sensors Journal*, vol. 14, no. 6, pp. 1898–1903, 2014.

[101] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *ICIP*. IEEE, 2015, pp. 168–172.

[102] C. Malleson, A. Gilbert, M. Trumble, J. Collomosse, A. Hilton, and M. Volino, "Real-time full-body motion capture from video and imus," in *3DV*. IEEE, 2017, pp. 449–457.

[103] N. Dawar and N. Kehtarnavaz, "Action detection and recognition in continuous action streams by deep learning-based sensing fusion," *IEEE Sensors Journal*, vol. 18, no. 23, pp. 9660–9668, 2018.

[104] A. Manzi, A. Moschetti, R. Limosani, L. Fiorini, and F. Cavallo, "Enhancing activity recognition of self-localized robot through depth camera and wearable sensors," *IEEE Sensors Journal*, vol. 18, no. 22, pp. 9324–9331, 2018.

[105] L. Wang, B. Sun, J. Robinson, T. Jing, and Y. Fu, "Ev-action: Electromyography-vision multi-modal action dataset," in *FG*. IEEE, 2020, pp. 160–167.

[106] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, and J. C. Niebles, "Home action genome: Cooperative compositional action understanding," in *CVPR*, 2021, pp. 11 184–11 193.

[107] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *CVPR*, 2022, pp. 18 995–19 012.

[108] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *IJCV*, pp. 1–23, 2022.

[109] M. Martínez-Zarzuela, J. González-Alonso, M. Antón-Rodríguez, F. J. Díaz-Pernas, H. Müller, and C. Simón-Martínez, "Multimodal video and imu kinematic dataset on daily life activities using affordable devices," *Scientific Data*, vol. 10, no. 1, p. 648, 2023.

[110] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition." in *IJCAI*, vol. 15. Buenos Aires, Argentina, 2015, pp. 3995–4001.

[111] G. Chetty and M. Yamin, "Intelligent human activity recognition scheme for ehealth applications," *Malaysian Journal of Computer Science*, vol. 28, no. 1, pp. 59–69, 2015.

[112] H. Guo, L. Chen, L. Peng, and G. Chen, "Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble," in *UbiComp*, 2016, pp. 1112–1123.

[113] M. Guo, Z. Wang, N. Yang, Z. Li, and T. An, "A multisensor multiclassifier hierarchical fusion model based on entropy weight for human activity recognition using wearable inertial sensors," *THMS*, vol. 49, no. 1, pp. 105–111, 2018.

[114] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *WWW*, 2017, pp. 351–360.

[115] J. Sena, J. Barreto, C. Caetano, G. Cramer, and W. R. Schwartz, "Human activity recognition based on smartphone and wearable sensors using multiscale dcnn ensemble," *Neurocomputing*, vol. 444, pp. 226–243, 2021.

[116] M. Ullah, H. Ullah, S. D. Khan, and F. A. Cheikh, "Stacked lstm network for human activity recognition using smartphone data," in *EUVIP*. IEEE, 2019, pp. 175–180.

[117] S. Yu and L. Qin, "Human activity recognition with smartphone inertial sensors using bidir-lstm networks," in *ICMCCE*. IEEE, 2018, pp. 219–224.

[118] I. K. Ihianle, A. O. Nwajana, S. H. Ebenuwa, R. I. Otuka, K. Owa, and M. O. Orisatoki, "A deep learning approach for human activities recognition from multimodal sensing devices," *IEEE Access*, vol. 8, pp. 179 028–179 038, 2020.

[119] N. Dua, S. N. Singh, and V. B. Semwal, "Multi-input cnn-gru based human activity recognition using wearable sensors," *Computing*, vol. 103, pp. 1461–1478, 2021.

[120] W. Gao, L. Zhang, Q. Teng, J. He, and H. Wu, "Danhar: Dual attention network for multimodal human activity recognition using wearable sensors," *Applied Soft Computing*, vol. 111, p. 107728, 2021.

[121] Y. Tang, L. Zhang, Q. Teng, F. Min, and A. Song, "Triple cross-domain attention on human activity recognition using wearable sensors," *TETCI*, vol. 6, no. 5, pp. 1167–1176, 2022.

[122] M. A. Al-qaness, A. Dahou, M. Abd Elaziz, and A. Helmi, "Multi-resatt: Multilevel residual network with attention for human activity recognition using wearable sensors," *TII*, vol. 19, no. 1, pp. 144–152, 2022.

[123] C. Zhang, A. Waghmare, P. Kundra, Y. Pu, S. Gilliland, T. Ploetz, T. E. Starner, O. T. Inan, and G. D. Abowd, "Fingersound: Recognizing unistroke thumb gestures using a ring," *IMWUT*, vol. 1, no. 3, pp. 1–19, 2017.

[124] E. Garcia-Ceja, C. E. Galván-Tejada, and R. Brena, "Multi-view stacking for activity recognition with sound and accelerometer data," *Information Fusion*, vol. 40, pp. 45–56, 2018.

[125] N. Siddiqui and R. H. Chan, "Multimodal hand gesture recognition using single imu and acoustic measurements at wrist," *Plos one*, vol. 15, no. 1, p. e0227039, 2020.

[126] V. Mollyn, K. Ahuja, D. Verma, C. Harrison, and M. Goel, "Samosa: Sensing activities with motion and subsampled audio," *IMWUT*, vol. 6, no. 3, pp. 1–19, 2022.

[127] G. Lin, W. Jiang, S. Xu, X. Zhou, X. Guo, Y. Zhu, and X. He, "Human activity recognition using smartphones with wifi signals," *THMS*, vol. 53, no. 1, pp. 142–153, 2022.

[128] H. Xu, L. Han, M. Li, and M. Srivastava, "Penetrative ai: Making llms comprehend the physical world," *arXiv preprint arXiv:2310.09605*, 2023.

[129] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, "Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time," *TOG*, vol. 37, no. 6, pp. 1–15, 2018.

[130] Z. Leng, A. Bhattacharjee, H. Rajasekhar, L. Zhang, E. Bruda, H. Kwon, and T. Plötz, "Imugpt 2.0: Language-based cross modality transfer for sensor-based human activity recognition," *arXiv preprint arXiv:2402.01049*, 2024.

[131] H. Kwon, C. Tong, H. Haresamudram, Y. Gao, G. D. Abowd, N. D. Lane, and T. Ploetz, "Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition," *IMWUT*, vol. 4, no. 3, pp. 1–29, 2020.

[132] W. Chen, S. Lin, E. Thompson, and J. Stankovic, "Sensecollect: We need efficient ways to collect on-body sensor-based human activity data!" *IMWUT*, vol. 5, no. 3, pp. 1–27, 2021.

[133] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. Ieee, 2009, pp. 248–255.

[134] W. Wang and Y. Yang, "Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models," *arXiv preprint arXiv:2403.06098*, 2024.

[135] T.-S. Chen, A. Siarohin, W. Menapace, E. Deyneka, H.-w. Chao, B. E. Jeon, Y. Fang, H.-Y. Lee, J. Ren, M.-H. Yang *et al.*, "Panda-70m: Captioning 70m videos with multiple cross-modality teachers," *arXiv preprint arXiv:2402.19479*, 2024.

[136] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, W. Ai, B. Martinez *et al.*, "Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation," *arXiv preprint arXiv:2403.09227*, 2024.

[137] S. Singh, F. Vargus, D. Dsouza, B. F. Karlsson, A. Mahendiran, W.-Y. Ko, H. Shandilya, J. Patel, D. Mataciunas, L. OMahony *et al.*, "Aya dataset: An open-access collection for multilingual instruction tuning," *arXiv preprint arXiv:2402.06619*, 2024.

[138] S. Xia, L. Chu, L. Pei, Z. Zhang, W. Yu, and R. C. Qiu, "Learning disentangled representation for mixed-reality human activity recognition with a single imu sensor," *TIM*, vol. 70, pp. 1–14, 2021.

[139] Y. Hao, X. Lou, B. Wang, and R. Zheng, "Cromosim: A deep learning-based cross-modality inertial measurement simulator," *TMC*, 2022.

[140] B. Ataseven, A. Madani, B. Semiz, and M. E. Gursoy, "Physical activity recognition using deep transfer learning with convolutional neural networks," in *DASC/PiCom/CBDCom/CyberSciTech*. IEEE, 2022, pp. 1–6.

[141] M. Hashim and R. Amutha, "Deep transfer learning based human activity recognition by transforming imu data to image domain using novel activity image creation method," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 3, pp. 2883–2890, 2022.

[142] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

[143] H. Yoon, H. Cha, C. H. Nguyen, T. Gong, and S.-J. Lee, "Img2imu: Applying knowledge from large-scale images to imu applications via contrastive learning," *arXiv preprint arXiv:2209.00945*, 2022.

[144] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *CVPR*, 2019, pp. 11 996–12 004.

[145] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *CVPR*, 2022, pp. 2969–2978.

[146] J. Ni, H. Tang, A. H. Ngu, G. Liu, and Y. Yan, "Physical-aware cross-modal adversarial network for wearable sensor-based human action recognition," *arXiv preprint arXiv:2307.03638*, 2023.

[147] H. Kwon, B. Wang, G. D. Abowd, and T. Plötz, "Approaching the real-world: Supporting activity recognition training with virtual imu data," *IMWUT*, vol. 5, no. 3, pp. 1–32, 2021.

[148] Y. Jain, H. Kwon, and T. Ploetz, "On the effectiveness of virtual imu data for eating detection with wrist sensors," in *UbiComp/ISWC*, 2022, pp. 50–52.

[149] I. Gavier, Y. Liu, and S. I. Lee, "Virtualimu: Generating virtual wearable inertial data from video for deep learning applications," in *BSN*. IEEE, 2023, pp. 1–4.

[150] J. Li, L. Huang, S. Shah, S. J. Jones, Y. Jin, D. Wang, A. Russell, S. Choi, Y. Gao, J. Yuan *et al.*, "Signring: Continuous american sign language recognition using imu rings and virtual imu data," *IMWUT*, vol. 7, no. 3, pp. 1–29, 2023.

[151] P. S. Santhalingam, P. Pathak, H. Rangwala, and J. Kosecka, "Synthetic smartwatch imu data generation from in-the-wild asl videos," *IMWUT*, vol. 7, no. 2, pp. 1–34, 2023.

[152] Z. Leng, H. Kwon, and T. Plötz, "Generating virtual on-body accelerometer data from virtual textual descriptions for human activity recognition," *arXiv preprint arXiv:2305.03187*, 2023.

[153] V. Fortes Rey, K. K. Garewal, and P. Lukowicz, "Translating videos into synthetic training data for wearable sensor-based activity recognition systems using residual deep convolutional networks," *Applied Sciences*, vol. 11, no. 7, p. 3094, 2021.

[154] C. Xia and Y. Sugiura, "Virtual imu data augmentation by spring-joint model for motion exercises recognition without using real data," in *ISWC*, 2022, pp. 79–83.

[155] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He *et al.*, "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," *arXiv preprint arXiv:2302.09419*, 2023.

[156] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[157] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," *arXiv preprint arXiv:2309.05519*, 2023.

[158] J. Li, C. Liu, S. Cheng, R. Arcucci, and S. Hong, "Frozen language model helps ecg zero-shot learning," in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 402–415.

[159] X. Liu, D. McDuff, G. Kovacs, I. Galatzer-Levy, J. Sunshine, J. Zhan, M.-Z. Poh, S. Liao, P. Di Achille, and S. Patel, "Large language models are few-shot health learners," *arXiv preprint arXiv:2305.15525*, 2023.

[160] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[161] Z. Zhang, H. Amiri, Z. Liu, A. Züfle, and L. Zhao, "Large language models for spatial trajectory patterns mining," *arXiv preprint arXiv:2310.04942*, 2023.

[162] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[163] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," vol. 27, 2014.

[164] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," vol. 33, pp. 6840–6851, 2020.

[165] G. Aggarwal and D. Parikh, "Dance2music: Automatic dance-driven music generation," *arXiv preprint arXiv:2107.06252*, 2021.

[166] J. Wang, Y. Chen, Y. Gu, Y. Xiao, and H. Pan, "Sensorygans: An effective generative adversarial framework for sensor-based human activity recognition," in *IJCNN*. IEEE, 2018, pp. 1–8.

[167] M. R. Siyal, M. Ebrahim, S. H. Adil, and K. Raza, "Human action recognition using convlstm with gan and transfer learning," in *ICCI*. IEEE, 2020, pp. 311–316.

[168] X. Li, J. Luo, and R. Younes, "Activitygan: Generative adversarial networks for data augmentation in sensor-based human activity recognition," in *UbiComp/ISWC*, 2020, pp. 249–254.

[169] J. Wang, Y. Chen, and Y. Gu, "A wearable-har oriented sensory data generation method based on spatio-temporal reinforced conditional gans," *Neurocomputing*, vol. 493, pp. 548–567, 2022.

[170] X. Li, V. Metsis, H. Wang, and A. H. H. Ngu, "Tts-gan: A transformer-based time-series generative adversarial network," in *AIME*. Springer, 2022, pp. 133–143.

[171] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[172] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*, 2019, pp. 6023–6032.

[173] C. I. Tang, I. Perez-Pozuelo, D. Spathis, S. Brage, N. Wareham, and C. Mascolo, "Selfhar: Improving human activity recognition through self-training with unlabeled data," *IMWUT*, vol. 5, no. 1, pp. 1–30, 2021.

[174] S. Zhang and N. Alshurafa, "Deep generative cross-modal on-body accelerometer data synthesis from videos," in *UbiComp/ISWC*, 2020, pp. 223–227.

[175] G. Iglesias, E. Talavera, Á. González-Prieto, A. Mozo, and S. Gómez-Canaval, "Data augmentation techniques in time series domain: a survey and taxonomy," *arXiv preprint arXiv:2206.13508*, 2022.

[176] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[177] A. Renduchintala, S. Ding, M. Wiesner, and S. Watanabe, "Multimodal data augmentation for end-to-end asr," *arXiv preprint arXiv:1803.10299*, 2018.

[178] A. Falcon, G. Serra, and O. Lanz, "A feature-space multimodal data augmentation technique for text-video retrieval," in *ACM MM*, 2022, pp. 4385–4394.

[179] D. Oneata and H. Cucu, "Improving multimodal speech recognition by data augmentation and speech representations," *arXiv preprint arXiv:2204.13206*, 2022.

[180] N. Xu, W. Mao, P. Wei, and D. Zeng, "Mda: Multimodal data augmentation framework for boosting performance on sentiment/emotion classification tasks," *IEEE Intelligent Systems*, vol. 36, no. 6, pp. 3–12, 2020.

[181] Z. Liu, Z. Tang, X. Shi, A. Zhang, M. Li, A. Shrivastava, and A. G. Wilson, "Learning multimodal data augmentation in feature space," *arXiv preprint arXiv:2212.14453*, 2022.

[182] J. Hua, X. Cui, X. Li, K. Tang, and P. Zhu, "Multimodal fake news detection through data augmentation-based contrastive learning," *Applied Soft Computing*, vol. 136, p. 110125, 2023.

[183] A. Josi, M. Alehdaghi, R. M. Cruz, and E. Granger, "Multimodal data augmentation for visual-infrared person reid with corrupted data," in *WACV*, 2023, pp. 32–41.

[184] D. Wang and S. Karout, "Fine-grained multi-modal self-supervised learning," *arXiv preprint arXiv:2112.12182*, 2021.

[185] Y. Asano, M. Patrick, C. Rupprecht, and A. Vedaldi, "Labelling unlabelled videos from scratch with multi-modal self-supervision," vol. 33, pp. 4660–4671, 2020.

[186] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *CVPR*, 2020, pp. 9879–9889.

[187] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman, "Self-supervised multimodal versatile networks," vol. 33, pp. 25–37, 2020.

[188] Y. Li, H. Liu, and H. Tang, "Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking," in *AAAI*, vol. 36, no. 2, 2022, pp. 1456–1463.

[189] A. Saeed, T. Ozcelebi, and J. Lukkien, "Multi-task self-supervised learning for human activity detection," *IMWUT*, vol. 3, no. 2, pp. 1–30, 2019.

[190] H. Yuan, S. Chan, A. P. Creagh, C. Tong, D. A. Clifton, and A. Doherty, "Self-supervised learning for human activity recognition using 700,000 person-days of wearable data," *arXiv preprint arXiv:2206.02909*, 2022.

[191] Y. Jain, C. I. Tang, C. Min, F. Kawsar, and A. Mathur, "Collossl: Collaborative self-supervised learning for human activity recognition," *IMWUT*, vol. 6, no. 1, pp. 1–28, 2022.

[192] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," vol. 34, pp. 24 206–24 221, 2021.

[193] H. Choi, A. Beedu, and I. Essa, "Multimodal contrastive learning with hard negative sampling for human activity recognition," *arXiv preprint arXiv:2309.01262*, 2023.

[194] R. Brinzea, B. Khaertdinov, and S. Asteriadis, "Contrastive learning with cross-modal knowledge mining for multimodal human activity recognition," in *IJCNN*. IEEE, 2022, pp. 01–08.

[195] S. Deldari, D. Spathis, M. Malekzadeh, F. Kawsar, F. Salim, and A. Mathur, "Latent masking for multimodal self-supervised learning in health timeseries," *arXiv preprint arXiv:2307.16847*, 2023.

[196] S. Moon, A. Madotto, Z. Lin, A. Dirafzoon, A. Saraf, A. Bearman, and B. Damavandi, "Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text," *arXiv preprint arXiv:2210.14395*, 2022.

[197] K. Xia, W. Li, S. Gan, and S. Lu, "Ts2act: Few-shot human activity sensing with cross-modal co-learning," *IMWUT*, vol. 7, no. 4, pp. 1–22, 2024.

[198] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *CVPR*, 2023, pp. 15 180–15 190.

[199] W. C. Sleeman IV, R. Kapoor, and P. Ghosh, "Multimodal classification: Current landscape, taxonomy and future directions," *CSUR*, vol. 55, no. 7, pp. 1–31, 2022.

[200] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *JAIR*, vol. 16, pp. 321–357, 2002.

[201] M. Junaid, S. Ali, F. Eid, S. El-Sappagh, and T. Abuhmed, "Explainable machine learning models based on multimodal time-series data for the early detection of parkinson's disease," *Computer Methods and Programs in Biomedicine*, vol. 234, p. 107495, 2023.

[202] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Systems with applications*, vol. 91, pp. 464–471, 2018.

[203] H. K. Lee, J. Lee, and S. B. Kim, "Boundary-focused generative adversarial networks for imbalanced and multimodal time series," *TKDE*, vol. 34, no. 9, pp. 4102–4118, 2022.

[204] Q. Li, G. Yu, J. Wang, and Y. Liu, "A deep multimodal generative and fusion framework for class-imbalanced multimodal data," *Multimedia Tools and Applications*, vol. 79, pp. 25 023–25 050, 2020.

[205] Y. Guo, Y. Chu, B. Jiao, J. Cheng, Z. Yu, N. Cui, and L. Ma, "Evolutionary dual-ensemble class imbalance learning for human activity recognition," *TETCI*, vol. 6, no. 4, pp. 728–739, 2021.

[206] N. C. Garcia, S. A. Bargal, V. Ablavsky, P. Morerio, V. Murino, and S. Sclaroff, "Distillation multiple choice learning for multimodal action recognition," in *WACV*, 2021, pp. 2755–2764.

[207] S. G. Dhekane, H. Haresamudram, M. Thukral, and T. Plötz, "How much unlabeled data is really needed for effective self-supervised human activity recognition?" in *ISWC*, 2023, pp. 66–70.

[208] Y. Wei, Y. Zhu, C. Leung, Y. Song, and Q. Yang, "Instilling social to physical: Co-regularized heterogeneous transfer learning," in *AAAI*, vol. 30, no. 1, 2016.

[209] J. Wang, V. W. Zheng, Y. Chen, and M. Huang, "Deep transfer learning for cross-domain activity recognition," in *ICCSE*, 2018, pp. 1–8.

[210] J. Wang, Y. Chen, L. Hu, X. Peng, and S. Y. Philip, "Stratified transfer learning for cross-domain activity recognition," in *PerCom*. IEEE, 2018, pp. 1–10.

[211] W. Lu, Y. Chen, J. Wang, and X. Qin, "Cross-domain activity recognition via substructural optimal transport," *Neurocomputing*, vol. 454, pp. 65–75, 2021.

[212] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *CVPR*, 2021, pp. 6995–7004.

[213] Y.-L. Sung, J. Cho, and M. Bansal, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *CVPR*, 2022, pp. 5227–5237.

[214] B. Khaertdinov and S. Asteriadis, "Temporal feature alignment in contrastive self-supervised learning for human activity recognition," in *IJCB*. IEEE, 2022, pp. 1–9.

[215] X. Geng, H. Liu, L. Lee, D. Schuurmans, S. Levine, and P. Abbeel, "Multimodal masked autoencoders learn transferable representations," *arXiv preprint arXiv:2205.14204*, 2022.

[216] W. Lu, J. Wang, Y. Chen, S. J. Pan, C. Hu, and X. Qin, "Semantic-discriminative mixup for generalizable sensor-based cross-domain activity recognition," *IMWUT*, vol. 6, no. 2, pp. 1–19, 2022.

[217] M. Thukral, H. Haresamudram, and T. Ploetz, "Cross-domain har: Few shot transfer learning for human activity recognition," *arXiv preprint arXiv:2310.14390*, 2023.

[218] R. Bao, Y. Sun, Y. Gao, J. Wang, Q. Yang, H. Chen, Z.-H. Mao, X. Xie, and Y. Ye, "A survey on heterogeneous transfer learning," *arXiv preprint arXiv:2310.08459*, 2023.

[219] X. Qin, Y. Chen, J. Wang, and C. Yu, "Cross-dataset activity recognition via adaptive spatial-temporal transfer learning," *IMWUT*, vol. 3, no. 4, pp. 1–25, 2019.

[220] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *CVPR*, 2016, pp. 826–834.

[221] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *ECCV*, 2018, pp. 103–118.

[222] A. Andonian, S. Chen, and R. Hamid, "Robust cross-modal representation learning with progressive self-distillation," in *CVPR*, 2022, pp. 16 430–16 441.

[223] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," vol. 29, pp. 892–900, 2016.

[224] F. M. Thoker and J. Gall, "Cross-modal knowledge distillation for action recognition," in *ICIP*, 2019, pp. 6–10.

[225] Z. Quan, Q. Chen, W. Wang, M. Zhang, X. Li, Y. Li, and Z. Liu, "Smtdkd: A semantic-aware multimodal transformer fusion decoupled knowledge distillation method for action recognition," *IEEE Sensors Journal*, 2023.

[226] X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, "A survey on model compression for large language models," *arXiv preprint arXiv:2308.07633*, 2023.

[227] J. K. Eshraghian, M. Ward, E. O. Neftci, X. Wang, G. Lenz, G. Dwivedi, M. Bennamoun, D. S. Jeong, and W. D. Lu, "Training spiking neural networks using lessons from deep learning," *Proceedings of the IEEE*, 2023.

[228] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *Journal of Machine Learning Research*, vol. 20, no. 55, pp. 1–21, 2019.

[229] T. Mauldin, A. H. Ngu, V. Metsis, and M. E. Canby, "Ensemble deep learning on wearables using small datasets," *HEALTH*, vol. 2, no. 1, pp. 1–30, 2020.

[230] S. Nooruddin, M. M. Islam, F. Karray, and G. Muhammad, "A multi-resolution fusion approach for human activity recognition from video data in tiny edge devices," *Information Fusion*, vol. 100, p. 101953, 2023.

[231] Y. Zhou, H. Zhao, Y. Huang, T. Riedel, M. Hefenbrock, and M. Beigl, "Tinyhar: A lightweight deep learning model designed for human activity recognition," in *UbiComp*, 2022, pp. 89–93.

[232] Z. Gao, Y. Wang, J. Chen, J. Xing, S. Patel, X. Liu, and Y. Shi, "Mmtsa: Multi-modal temporal segment attention network for efficient human activity recognition," *IMWUT*, vol. 7, no. 3, pp. 1–26, 2023.

[233] Q. Cai, X. Liu, K. Zhang, X. Xie, X. Tong, and K. Li, "Acf: An adaptive compression framework for multimodal network in embedded devices," *TMC*, 2023.

[234] M. Bock, A. Hölzemann, M. Moeller, and K. Van Laerhoven, "Improving deep learning for har with shallow lstms," in *UbiComp*, 2021, pp. 7–12.

[235] C. Luo, X. He, J. Zhan, L. Wang, W. Gao, and J. Dai, "Comparison and benchmarking of ai models and frameworks on mobile devices," *arXiv preprint arXiv:2005.05085*, 2020.

[236] S. Ö. Bursa, Ö. Durmaz İncel, and G. Işıklar Alptekin, "Building lightweight deep learning models with tensorflow lite for human activity recognition on mobile devices," *Annals of Telecommunications*, vol. 78, no. 11, pp. 687–702, 2023.

[237] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, "Action transformer: A self-attention model for short-time pose-based human action recognition," *Pattern Recognition*, vol. 124, p. 108487, 2022.

[238] M.-S. Kang, D. Kang, and H. Kim, "Efficient skeleton-based action recognition via joint-mapping strategies," in *WACV*, 2023, pp. 3403–3412.

[239] Y.-T. Hsieh, K. Anjum, and D. Pompili, "Ultra-low power analog recurrent neural network design approximation for wireless health monitoring," in *MASS*. IEEE, 2022, pp. 211–219.

[240] N. Sengupta, C. B. McNabb, N. Kasabov, and B. R. Russell, "Integrating space, time, and orientation in spiking neural networks: a case study on multimodal brain data modeling," *TNNLS*, vol. 29, no. 11, pp. 5249–5263, 2018.

[241] Q. Liu, D. Xing, L. Feng, H. Tang, and G. Pan, "Event-based multimodal spiking neural network with attention mechanism," in *ICASSP*. IEEE, 2022, pp. 8922–8926.

[242] X. Wang, Z. Wu, Y. Rong, L. Zhu, B. Jiang, J. Tang, and Y. Tian, "Sstformer: bridging spiking neural network and memory support transformer for frame-event based recognition," *arXiv preprint arXiv:2308.04369*, 2023.

[243] Y. Wang, B. Dong, Y. Zhang, Y. Zhou, H. Mei, Z. Wei, and X. Yang, "Event-enhanced multi-modal spiking neural network for dynamic obstacle avoidance," in *ACM MM*, 2023, pp. 3138–3148.

[244] L. Guo, Z. Gao, J. Qu, S. Zheng, R. Jiang, Y. Lu, and H. Qiao, "Transformer-based spiking neural networks for multimodal audio-visual classification," *TCDS*, 2023.

[245] V. Fra, E. Forno, R. Pignari, T. C. Stewart, E. Macii, and G. Urgese, "Human activity recognition: suitability of a neuromorphic approach for on-edge aiot applications," *Neuromorphic Computing and Engineering*, vol. 2, no. 1, p. 014006, 2022.

[246] A. R. Khan, H. U. Manzoor, F. Ayaz, M. A. Imran, and A. Zoha, "A privacy and energy-aware federated framework for human activity recognition," *Sensors*, vol. 23, no. 23, p. 9339, 2023.

[247] Y. Li, X. Sun, Z. Yang, and H. Huang, "Snnauth: Sensor-based continuous authentication on smartphones using spiking neural networks," *IoT-J*, 2024.

[248] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, X. Chen, and X. Wang, "A comprehensive survey of neural architecture search: Challenges and solutions," *CSUR*, vol. 54, no. 4, pp. 1–34, 2021.

[249] Y.-T. Hsieh, K. Anjum, S. Huang, I. Kulkarni, and D. Pompili, "Neural network design via voltage-based resistive processing unit and diode activation function-a new architecture," in *MWSCAS*. IEEE, 2021, pp. 59–62.

[250] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "Mfas: Multimodal fusion architecture search," in *CVPR*, 2019, pp. 6966–6975.

[251] Z. Yu, Y. Cui, J. Yu, M. Wang, D. Tao, and Q. Tian, "Deep multimodal neural architecture search," in *ACM MM*, 2020, pp. 3743–3752.

[252] Z. Xu, D. R. So, and A. M. Dai, "Mufasa: Multimodal fusion architecture search for electronic health records," in *AAAI*, vol. 35, no. 12, 2021, pp. 10 532–10 540.

[253] X. Shi, P. Zhou, W. Chen, and L. Xie, "Efficient gradient-based neural architecture search for end-to-end asr," in *ICMI*, 2021, pp. 91–96.

[254] D. Si, Q. Ye, J. Lv, Y. Zhou, and J. Lv, "Violence-mfas: Audio-visual violence detection using multimodal fusion architecture search," in *ICONIP*. Springer, 2023, pp. 205–216.

[255] X. Wang, X. Wang, T. Lv, L. Jin, and M. He, "Harnas: human activity recognition based on automatic neural architecture search using evolutionary algorithms," *Sensors*, vol. 21, no. 20, p. 6927, 2021.

[256] W.-S. Lim, W. Seo, D.-W. Kim, and J. Lee, "Efficient human activity recognition using lookup table-based neural architecture search for mobile devices," *IEEE Access*, 2023.

[257] X. Wang, J.-F. Hu, J.-H. Lai, J. Zhang, and W.-S. Zheng, "Progressive teacher-student learning for early action prediction," in *CVPR*, 2019, pp. 3556–3565.

[258] N. Zheng, X. Song, T. Su, W. Liu, Y. Yan, and L. Nie, "Egocentric early action prediction via adversarial knowledge distillation," *TOMM*, vol. 19, no. 2, pp. 1–21, 2023.

[259] J. Su, C. Jiang, X. Jin, Y. Qiao, T. Xiao, H. Ma, R. Wei, Z. Jing, J. Xu, and J. Lin, "Large language models for forecasting and anomaly detection: A systematic literature review," *arXiv preprint arXiv:2402.10350*, 2024.

[260] D. Zhang, Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu, "Mm-llms: Recent advances in multimodal large language models," *arXiv preprint arXiv:2401.13601*, 2024.

[261] H. Gammulle, D. Ahmedt-Aristizabal, S. Denman, L. Tychsen-Smith, L. Petersson, and C. Fookes, "Continuous human action recognition for human-machine interaction: a review," *CSUR*, vol. 55, no. 13s, pp. 1–38, 2023.

[262] L. Zhu, F. Wei, and Y. Lu, "Beyond text: Frozen large language models in visual signal comprehension," *arXiv preprint arXiv:2403.07874*, 2024.

[263] Z. Pan, Y. Jiang, S. Garg, A. Schneider, Y. Nevmyvaka, and D. Song, "S^2ip-llm: Semantic space informed prompt learning with llm for time series forecasting," *arXiv preprint arXiv:2403.05798*, 2024.

[264] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
[265] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024.
[266] D. Liang, X. Zhou, X. Wang, X. Zhu, W. Xu, Z. Zou, X. Ye, and X. Bai, "Pointmamba: A simple state space model for point cloud analysis," *arXiv preprint arXiv:2402.10739*, 2024.
[267] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, "Videomamba: State space model for efficient video understanding," *arXiv preprint arXiv:2403.06977*, 2024.
[268] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, H. Cao, X. Cheng, M. Chung, M. Grella, K. K. GV *et al.*, "Rwkv: Reinventing rnns for the transformer era," *arXiv preprint arXiv:2305.13048*, 2023.
[269] Y. Duan, W. Wang, Z. Chen, X. Zhu, L. Lu, T. Lu, Y. Qiao, H. Li, J. Dai, and W. Wang, "Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures," *arXiv preprint arXiv:2403.02308*, 2024.
[270] K. An and S. Zhang, "Exploring rwkv for memory efficient and low latency streaming asr," *arXiv preprint arXiv:2309.14758*, 2023.
[271] H. Hou and F. R. Yu, "Rwkv-ts: Beyond traditional recurrent neural network for time series tasks," *arXiv preprint arXiv:2401.09093*, 2024.
[272] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Unified-io: A unified model for vision, language, and multi-modal tasks," *arXiv preprint arXiv:2206.08916*, 2022.
[273] J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li *et al.*, "Anygpt: Unified multimodal llm with discrete sequence modeling," *arXiv preprint arXiv:2402.12226*, 2024.
[274] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, X. Wu *et al.*, "Uniaudio: An audio foundation model toward universal audio generation," *arXiv preprint arXiv:2310.00704*, 2023.
[275] A. Jain, P. Katara, N. Gkanatsios, A. W. Harley, G. Sarch, K. Aggarwal, V. Chaudhary, and K. Fragkiadaki, "Odin: A single model for 2d and 3d perception," *arXiv preprint arXiv:2401.02416*, 2024.
[276] Y. Yuan, J. Ding, J. Feng, D. Jin, and Y. Li, "Unist: A prompt-empowered universal model for urban spatio-temporal prediction," *arXiv preprint arXiv:2402.11838*, 2024.
[277] S. Gao, T. Koker, O. Queen, T. Hartvigsen, T. Tsiligkaridis, and M. Zitnik, "Units: Building a unified time series model," *arXiv preprint arXiv:2403.00131*, 2024.
[278] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.
[279] S. Mo and Y. Tian, "Av-sam: Segment anything model meets audio-visual localization and segmentation," *arXiv preprint arXiv:2305.01836*, 2023.
[280] Z. Xiao, X. Xu, H. Xing, F. Song, X. Wang, and B. Zhao, "A federated learning system with enhanced feature extraction for human activity recognition," *Knowledge-Based Systems*, vol. 229, p. 107338, 2021.
[281] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.