IJWR

A Review on Transformer-Based Methods for Human Activity Recognition

Fatemeh Sadat Lasani[a]*, Ronak Fatahi[b]

[a]Department of Electrical and Computer Engineering, Qom University of Technology, Qom, Iran; lesani@qut.ac.ir

[b]Faculty of Electrical and Computer Engineering, Shahab Danesh University, Qom, Iran; ronakfatahi6@gmail.com

ABSTRACT

With the expansion of smart homes, Human Activity Recognition (HAR) has become a key challenge in artificial intelligence, enhancing not only the comfort and safety of residents but also contributing to the development of applications such as healthcare and smart surveillance. The Transformer architecture, with its ability to model long-term dependencies and process data in parallel, has made significant advancements in recognizing human activities. In addition, its multi-head attention mechanism enables the analysis of complex input data by allowing the model to focus on different parts of the input simultaneously, capturing diverse relationships and dependencies within the data. This paper examines the application of Transformers in HAR and analyzes recent studies (since 2019). In addition to investigating innovative architectures, feature extraction methods, and accuracy improvements, it also discusses the challenges and future prospects of these models in recognizing human activities. Rapid advancements in deep learning and access to extensive datasets have made Transformers a key tool for improving the accuracy and efficiency of HAR systems in smart environments.

Keywords— Human Activity Recognition, Transformer, Attention, Deep Learning, Pattern Recognition.

1. Introduction

The growth of technology and the artificial intelligence have making it a significant factor to improve the quality of human life. In recent years, one of the most popular applications of artificial intelligence is Human Activity Recognition (HAR). A HAR system automatically identifies the human activities, actions and behaviors based on the data captured by various sensors [1]. Traffic control systems, healthcare [2], elderly care [3], sports [4], security monitoring [5], emotion detection[6], and surgical activity recognition[7] systems are just a few examples of services that rely on HAR.

In fact, an HAR system identifies activities by analyzing a series of data sent from sensors over time [8]. Most of these activities are ordinary, such as walking, talking, standing, and sitting. It is also possible to perform an activity in smart place such as a smart home [9]. The stages of activity recognationusing the signals received from different sensors and based on deep learning techniques are briefly displayed in Figure 1.

Sensor data can be recorded remotely, such as through installed home cameras [10] , door sensors [11], radar, Wi-Fi [12], or other wireless methods. Alternatively, data can be directly recorded on the subject of interest, such as wearable sensors including accelerometers, gyroscopes, and magnetometers [13] or smartphones [14].

HAR involves processing large datasets, which can lead to high computational costs and increased training time for models. Traditional machine learning approaches excel at recognizing human activities through meticulous handcrafted features and algorithms; however, these methods face limitations due to their reliance on manual design and feature selection processes [15]. Conversely, deep learning models like Convolutional Neural Networks (CNN) demonstrate significant potential in HAR [16, 17], outperforming conventional techniques by automatically extracting meaningful representations directly from input data without requiring explicit feature engineering. As deep learning gains prominence within the field of machine learning and data mining, there's been an emerging shift towards more automated and self-learning systems, which promise improved

http://dx.doi.org/10.22133/ijwr.2024.485291.1244

Citation F.S. Lasani, R. Fatahi, "A Review on Transformer-Based Methods for Human Activity Recognition ", *International Journal of Web Research*, vol.7, no.4,pp.81-100, 2024, doi: http://dx.doi.org/10.22133/ijwr.2024.485291.1244.

*Coressponding Author

Article History: Received: 24 June 2024; Revised: 3 September 2024; Accepted: 15 September 2024.

performance and reduced dependence on expert-driven feature extraction [18]. The feature extraction and model building processes are often performed simultaneously in deep learning models. Features can be automatically learned through the network instead of being manually designed. Additionally, a deep neural network can extract high-level representations in the deep layers, making it more suitable for complex activity recognition tasks [15].

The transformer architecture is a deep learning model which was introduced for the first time in December 2017 in a paper titled "Attention Is All You Need", which states attention is everything you need in Google's machine translation [19] . At first, transformers were introduced to address the challenges of sequence modeling tasks, but their success in the field of natural language processing has encouraged researchers to explore various applications beyond text translation [20]. Transformers implement an attention-based encoder-decoder architecture for sequence analysis. Attention mechanisms learn to gather information from the entire sequence, thus accurately identifying behavioral patterns [21].

The core concept of self-attention in transformers has been employed in many recent methodologies, including Bidirectional Encoder Representations from Transformer (BERT), Generative Pretrained Transformer (GPT), and Vision Transformer (ViT). In addition, the transformer is utilized across various domains such as natural language processing (NLP), object detection [22], action recognition [23] , HAR [24], and computer vision (CV) [25].

Transformers, due to their specific architecture, have had various applications in HAR systems and have been able to improve some of the challenges of these systems. For example, in the data preprocessing and feature extraction [26, 27] stage, they can help reduce the data volume and have shown good accuracy in activity recognition [27].

The transformer architecture has emerged as a popular choice for recognizing human activities, making it crucial to review the research in this area. In this study, several papers published from 2019 onwards were analyzed to explore the role and application of transformers in HAR research. Our paper provides a comprehensive overview of the innovative applications of transformer architectures in the field of HAR. By systematically reviewing various types of sensors utilized in HAR, including sensor-based and vision-based technologies, we establish a foundational understanding of the data sources that drive HAR systems. The exploration of Transformer-based architectures, such as attention-based, vision-based, and hybrid models, highlights their unique capabilities in processing complex input data and capturing long-term dependencies.

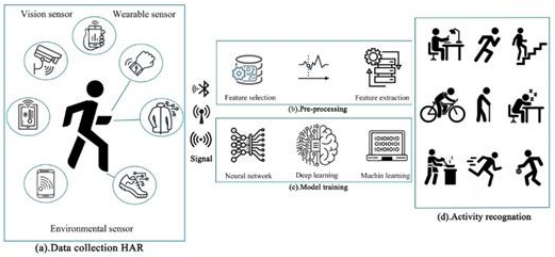

Figure. 1. Human activity recognition framework comprises of four main parts, (a) data collection for HAR using vision sensor , wearable sensor and environmental sensor; (b) data preprocessing, feature selection and feature extractin which performs essential pre-processing steps for the collected data; (c) training phase,which utilizes neural network or machine learning(ML) or deep learning approaches to learn patterns from thecollected data, and (d) activities recognition.

Furthermore, our review delves into the datasets used for HAR, emphasizing their importance in training effective models. We also address the challenges faced when implementing Transformers in HAR, offering insights into potential solutions and future directions for research.

The article is organized as follows: Section 2 introduces the types of data collection methods, and Section 3 introduces the types of transformers. Section 4 discusses the most common datasets in this field, and Section 5 reviews various articles that have used the distinctive architecture of transformers for human activity recognition. Section 6 the challenges associated with the use of transformers in HAR applications are discussed. Finally, Section 7 presents the results of the study and outlines potential future work. Figure.2 illustrates the research process of our paper.

2. Different Types of Sensors in Har

The first step in developing a HAR system is data collection. The data for a HAR system is obtained using various devices and sensors [28]. Figure 3 provides an overview of different data collection methods for a HAR system. Generally, data collection techniques can be divided into two main categories: image-based methods [29], and sensor-based methods [30].

2.1. Sensor-based HAR

A. Environmental Sensors: Environmental sensors measure environmental conditions. For example, temperature, humidity [31], light, air pressure, and air quality [32]. sensors fall into this category. These sensors can be used in smart home systems, smart city equipment, and other environmental equipment [33] .

B. Wearable Sensors: Wearable sensors are placed in devices or wearable items and measure physical activity, physiological parameters, location, and other information [34] . For example, heart rate

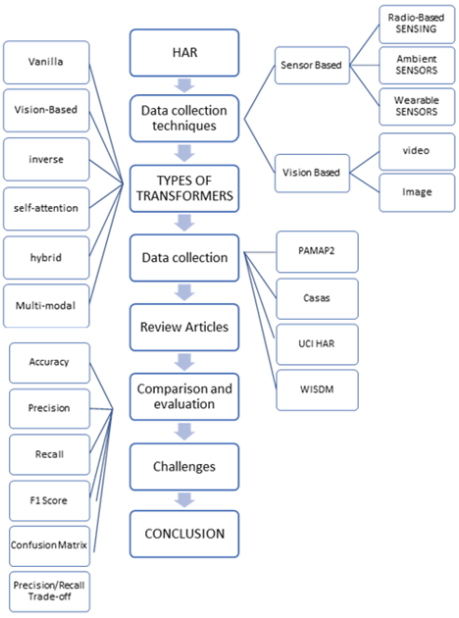

Figure. 2. research process

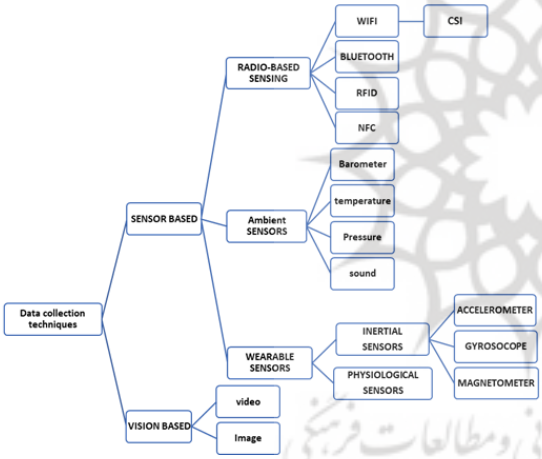

Figure. 3. Data collection techniques

sensors, accelerometers, gyroscopes, magnetometers, and GPS sensors fall into this category. These sensors are commonly used in wearable devices such as smartwatches, fitness bands, and sports equipment [33, 35] .

C: Radio sensors: Radio sensors use radio waves to detect and transmit information. For example, NFC (Near Field Communication) [36], RFID (Radio Frequency Identification) [37], and CSI [38]. are types of these sensors. They are used in various applications such as object identification, wireless payment, and device control. New optimized methods have been introduced in recent research, such as the activity detection system using CSI-based Wi-Fi reconstruction. Although Wi-Fi channels (CSI) are non-contact and low-cost, they are limited due to high computational complexity and poor cross-domain generalization performance[39].

2.2. Vision-based HAR

A. Video: Videos captured by cameras and other sensors can be processed to recognize human activites [40]. By analyzing video images, different features can be extracted for the purpose of object identification, motion detection and behavior analysis [41]. These sensors are used in many fields, including robotics [42, 43], smart cars [44] , security surveillance [45, 46] and smart city systems [47].

B. Images: This category includes sensors and devices that receive and process still images. These sensors usually capture images at a specific moment and extract information for object identification [48], face detection [49], and image analysis such as Medical Image Analysis [50]. These images can be obtained from digital cameras, CCTV cameras, mobile cameras, and other image sources. The use of images is common in fields such as medicine [51] , security [52] , aerial imaging [53], and industrial [54]imaging [55].

3. Transformer-Based Architectures

In this section, we provide an overview of different transformers architectures which are utilized in HAR research.

Transformer architectures rely on the self-attention mechanism, which offers several advantages over recurrent layers, such as better model parallelism and reduced inductive bias compared to convolution networks. This mechanism allows the model to dynamically focus on different parts of the input sequence, establish pairwise correlations, and model long-range dependencies between input data elements. In self-attention, the model calculates attention weights for each position in the sequence, reflecting the importance of each position relative to others. This enables the model to attend to various parts of the sequence based on the input. The attention module's input is processed by three distinct fully connected (FC) layers, which are trained to produce Query (Q), Key (K), and Value (V) tensors. The scaled dot-product attention (A), as described in Equ(1), is then computed [19].

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{Qk^T}{\sqrt{D_k}}\right)V \qquad (1)$$

The Query and Key are multiplied in an element-by-element manner to produce a score matrix, which is divided by $\sqrt{Dk}$, the square root of output dimensions of the Key matrix to alleviate the gradient vanishing problem. The softmax function boosts high score values and dampens lower score values. The attention score is finally obtained by

multiplying the attention and value matrix, as given in Equ(1).

3.1. Attention-Based Transformers (Vanilla Transformer)

The Vanilla Transformer is a sequence-to-sequence model composed of an encoder and a decoder, each consisting of sets of identical blocks. Each encoder block primarily comprises a multi-head self-attention module and a position-wise feedforward network (FFN). Additionally, self-attention modules in the decoder are adjusted to prevent the presence of each position in subsequent positions. [19, 56]. The overall architecture of the Vanilla Transformer is illustrated in the Figure 4.

3.2. Vision-Based Transformers

In 2019, with the introduction of the vision transformer architecture by Dosovitskiy and his colleagues [13], significant progress occurred. The aim of this new approach is to process images without relying on traditional convolutional operations commonly used in computer vision tasks. Vision transformers, using self-attention mechanisms, intended to capture relationships between different regions of an image. This advancement led to new opportunities in analyzing image data. Some sample applications of vision transformer are images classification [57], and person re-identification [58] . Figure 5 shows the architecture of vision transformer.

3.3. Inverse Transformers

The Inverse Transformer (iTransformer) simply reverses the tasks of the attention mechanism and the feed-forward network. Specifically, temporal points in time series are embedded as token signals used by the attention mechanism to capture multi-variable correlations. Meanwhile, the feed-forward network is applied to each variable token to learn non-linear representations. The iTransformer model achieves state-of-the-art performance on various real-world datasets, enhancing the Transformer family with improved performance, generalization across different variables, and better utilization of arbitrary review windows, making it a compelling alternative [2] .

3.4. Self-Attention-Based Transformers

Self-attention in the transformer model refers to the model's ability to attend to all parts of the input sequence at each prediction step. This feature allows the model to consider complete information from the input sequence and prevent the loss of some information in subsequent stages. This is in contrast to "local attention" which only attends to specific parts of the input. Self-attention fundamentally enables the model to look at the entire context of the

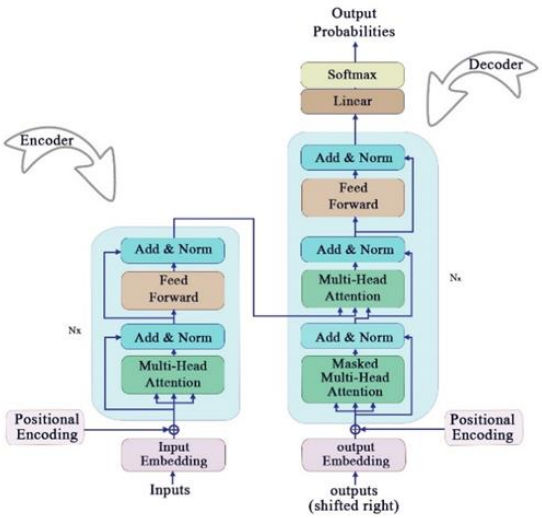

Figure. 4. Vanilla transformer architecture

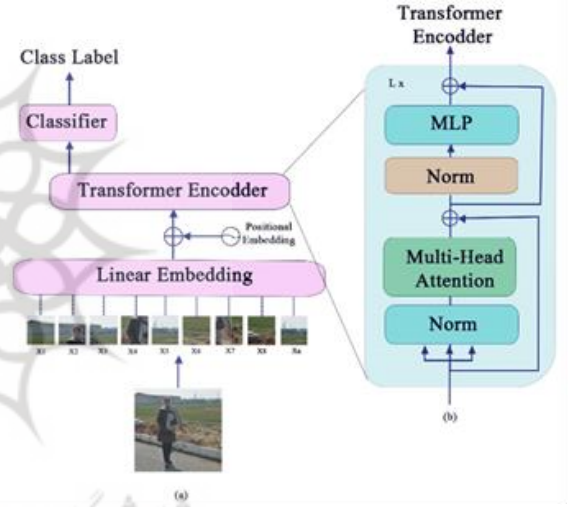

Figure. 5. The Vision Transformer architecture: (a) the main architecture of the model; (b) the Transformer encoder module.

input sequence during the encoding of each input element and not lose any information over time. This allows the model to make predictions with a complete view of the input without losing important information over time [19].

3.5. Transformer-Based Hybrid Models

Self-attention in the transformer model refers to the model's ability to attend to all parts of the input sequence at each prediction step. This feature allows the model to consider complete information from the input sequence and prevent the loss of some information in subsequent stages. This is in contrast to "local attention" which only attends to specific parts of the input. Self-attention fundamentally enables the model to look at the entire context of the input sequence during the encoding of each input element and not lose any information over time.

This allows the model to make predictions with a complete view of the input without losing important information over time [59].

The hybrid transformers are models that combine the architecture of transformers with other architectures such as CNN [60]. This combination is made to improve the performance and capabilities of neural models. In general, in hybrid transformers, parts of the model use architectures other than transformers, such as CNN, or are added to the main transformer structure with modifications. This can provide improvements in areas such as performance, learning speed, and model generalizability. Like This study proposes a hybrid SqueezeNet-vision transformer (SViT) model that combines the strengths of SqueezeNet and vision transformer (ViT) [61].

3.6. Multi-modal Transformer

Multi-modal Transformers are deep learning models based on the Transformer architecture that are capable of processing multi-modal data. The model is based on the original Transformer architecture and usually has multiple inputs, each representing a type of data, such as images and text. These models are able to combine different information from the input data using self-attention layers and extract spatial information and important features using the Transformer neural network [62].

4. Datasets of Har

This section summarizes key datasets commonly used in Human Activity Recognition (HAR) research. These datasets are critical for evaluating new methodologies and serve as benchmarks for comparing algorithm performance:

1. The PAMAP2 Dataset: This dataset includes motion and activity data captured by various sensors, including accelerometers, gyroscopes, and magnetometers. The dataset provides raw data about various activities such as walking, running, climbing stairs, and more [63].

2. The Casas Dataset: This dataset contains data related to daily activities in the home environment. It includes data such as movement, interaction with objects, use of household items, and environmental changes. This dataset is commonly used for research related to smart homes and smart living. Subsets include Casas Aruba, Milan, and Kyoto [64].

3. UCI HAR Dataset: This dataset contains motion data obtained using an accelerometer and gyroscope. It includes various activities such as walking, running, sitting, and other movements. Additionally, the data has been preprocessed and various features have been extracted from it [65].

4. WISDM Dataset: The WISDM dataset is a publicly available standard HAR dataset recorded by the Wireless Sensor Data Mining Lab. This dataset contains data of a specific set of daily activities performed by 51 individuals. Participants placed an Android mobile phone in their front pocket and performed various activities such as sitting, slow walking, up, down, standing, and walking for a certain period of time We have reviewed the general characteristics of the introduced dataset in the Table1.

5. CMDFALL Dataset: Contains data from 50 individuals wearing wrist and waist sensors, capturing normal activities (e.g., walking, sitting) and abnormal events (e.g., falls), sampled at 50 Hz [66].

6. C-MHAD Dataset: Focuses on hand gestures for smart TV control, offering 2-minute video and inertial data streams from 12 subjects [67].

7. DaLiAc Dataset: Includes 13 activities recorded from 19 subjects using IMUs (hip, chest, wrist, ankle) at 204.8 Hz [68].

8. Penn-Action Dataset: Offers 2,326 RGB video sequences across 15 action classes (e.g., pushups, baseball swings) [69].

9. NTU-RGB+D Dataset: Contains 56,880 samples across 60 action classes, providing multi-modal data (depth maps, 3D skeletons, RGB, infrared) with cross-subject and cross-view protocols [70].

10. Opportunity Dataset: Includes 6 hours of recordings from wearable and ambient sensors, featuring 113 sensor channels and annotations for posture states and gestures [71].

11. UTD-MHAD Dataset: Comprises 27 activities performed by 8 subjects, with multi-modal data (RGB, Depth, Skeleton, Inertial) across 861 samples [72].

12. MMAct Dataset: Contains 35 activities performed by 20 subjects with data from 7 modalities (e.g., RGB, Skeleton, Acceleration), recorded using cameras, smart glasses, smartphones, and smartwatches [73].

Table 1 provides a comparative overview of these datasets, highlighting sensor types, activities, participants, and environments. These datasets are foundational for training and evaluating machine learning and deep learning models in HAR, offering researchers a structured basis for algorithm benchmarking and analysis.

5. Transformer-Based Human Activity Recognition

In this section, we will explore different architectures of transformers and their applications

Table 1. Summary of Datasets used in HAR research papers

Dataset	*Year*	*Total Samples*	*Modality*	*Devices*	*Subjects*	*Features*	*Activity*	*Dataset Characteristics*
PAMAP2	2012	3,850,505	wireless inertial	Wearable	9	18	Real	Multivariate, Time-Series
Casas (17)	2011	1,045,876	Environmental	Sensor	1	11	Real	Multivariate, Time-Series
UCI HAR	2013	10,299	accelerometer and gyroscope	Wearable	30	6	-	Multivariate, Time-Series
WISDM	2012	15,630,426	accelerometer and gyroscope	Smartphone, smartwatch	51	Real	18	Multivariate, Time-Series
CMDFALL[66]	2018	20,764,515	Accelerometer, gyroscope, magnetometer	Wearable	۶	13	Real	Multivariate, Time-Series
C-MHAD [67]	2013	2,320,000	Accelerometer, RGB, depth	Kinect, wearable	12	8	Real	Multimodal, Multivariate
DaLiAc	2017	14,180,000	Accelerometer	Wearable	15	6	Real	Multivariate, Time-Series
Penn-Action [69]	2015	232,000	Video	RGB camera	232۶	15	Annotated	Single-modality, Action videos
NTU-RGB+D [70]	2016	56,880	RGB, depth, skeleton	Kinect, wearable	40	60	Annotated	Multivariate, Time-Series
Opportunity [71]	2011	4,184,000	Accelerometer, gyroscope, magnetometer	Wearable, environmental	4	113	Annotated	Multimodal, Multivariate
UTD-MHAD [72]	2015	861,888	Accelerometer, gyroscope,depth	Kinect, wearable	8	27	Annotated	Multivariate, Time-Series
MMAct [73]	2019	6,000,000	Accelerometer, gyroscope,video	Smartphone, RGB camera	50	20	Real	Multimodal, Multivariate

in HAR studies. We will delve into specific HAR applications of these transformer architectures, showcasing their effectiveness in identifying and classifying human activities in various contexts.

5.1. Attention-Based Transformers (Vanilla Transformer)

The input to HAR systems consists of datasets collected from various sensors, presented as a time sequence. These data include precise timestamps and sensor status at each moment. Unlike smartphone or smartwatch data, which are recorded at fixed intervals, environmental sensors in smart homes generate data in response to events, leading to irregular recording as an event stream.

Kwapisz et al. [74] examine the use of deep learning models, especially the transformer model, for detecting human activities using wearable sensors. In their paper, deep learning models including RNN, LSTM, BLSTM, 1D CNN, and DeepConvLSTM have been investigated. Saidani and colleagues' recommendations for improving model generalization and preventing overfitting include using data augmentation techniques such as time warping, domain adaptation, and adding Gaussian noise to training data. The composite features of this model extract information from low-level and high-level sensor data, enhancing the system's discriminative power. These features include mel-spectrogram, tonnetz, spectral contrast, chromagram, and MFCC. In the proposed model, the extracted features are used as input for the transformer model, which increases model accuracy by using fewer layers to extract long-range dependencies. Finally, metrics such as accuracy for datasets PAMAP2, UCI HAR, and WISDM are proposed for evaluating the proposed transformer model.

Overall the irregularity of sensor events poses challenges for traditional time series processing methods, which assume fixed intervals. Therefore, newer approaches like "activity-based sliding windows" are better suited for handling such data [75] .

The Sliding Window Algorithm [76] is an activity-based approach used in time series analysis and signal processing. This method moves a moving window over the data and performs operations such as feature extraction or pattern recognition at each window location. The sliding window algorithm is capable of identifying patterns and important events in time series data and continuously updating them. This method is used in various fields such as activity

recognition in sensor systems, motion analysis in videos, and detection of special patterns and events.

Transformers, known for their ability to capture complex relationships within data, combined with the sliding window method, significantly enhance activity recognition accuracy. This integration enables the model to analyze both the information within each window and the relationships across different time windows.

In a recent study by Huang et al. [24], combining activity-based sliding windows with transformer models significantly enhances the accuracy of activity recognition by effectively integrating different features of sensor data. Transformers, with their ability to process data in parallel and learn complex patterns, have proven to be powerful tools for sensor data analysis.

In summary, innovative methods like activity-based sliding windows and transformers address the challenges of processing sensor data in smart environments, enabling HAR systems to detect activities with greater accuracy and efficiency.

One key challenge in designing HAR systems is determining the optimal window size. The time window represents a sequential subset of input data presented to the model at each step. Choosing the right window size is critical—too small may lack sufficient information, while too large can reduce model performance and increase computational cost. The ideal window size depends on factors like activity type, sensor sampling rate, and model complexity.

Trung-Hieu Le and colleagues investigated the Transformer model for detecting human actions from inertial sensors [77]. This model has the capability to discover temporal correlations between features and offers advantages such as parallelization of large time series computations and the ability to learn more accurate context in long time series.

In their experiment, the impact of window size on the performance of the Transformer model has been investigated, and choosing an appropriate window size can contribute to improving detection .The results have shown that with an increase in the window size, the detection accuracy also improves. This model has been evaluated on three publicly released sensor datasets, MHAD-C, CMDFALL, and DaLiAc, demonstrating better performance compared to conventional methods, especially in the CMDFall dataset, which shows a 4.19% increase in F1 score compared to the conventional method. In the MHAD-C dataset, the accuracy has also reached 56.99%.

Ultimately, the comparison with other methods shows that the proposed approach has superior performance in gesture recognition from accelerometer and gyroscope data and has higher accuracy. The results confirm the role of Transformer models in identifying human activities.

A key challenge in HAR systems is enhancing model accuracy and generalization, especially with sensor data that includes noise or unpredictable variations. Overfitting is a significant issue, limiting models to training data and hindering performance on unseen data, particularly when data is collected at low frequency or contains high levels of noise.

Saeidnia et al. proposed a system that uses data augmentation techniques to increase model generalization and prevent overfitting [78]. Techniques such as time shifting, domain adaptation, and Gaussian noise are used to increase training data and improve model generalization. The proposed composite features of this model have the ability to extract information from low-level and high-level sensor data, which helps enhance the system's discriminative power. Some of the features used in this model include mel-spectrogram, tonnetz, spectral contrast, chromagram, and MFCC. In this model, the extracted features are used as the input to a transformer model. By using fewer layers to extract long-range dependencies, which leads to improved accuracy, the model has successfully been used to recognize complex patterns of human activities.

HAR data is usually collected from diverse sources, including video, skeleton data, and other sensors, each offering complementary information about activities. However, effectively integrating these diverse data sources remains challenging due to issues like data heterogeneity, noise, and imprecise labeling. Traditional methods often struggle with these challenges, limiting their ability to provide a comprehensive and accurate understanding of activities.

A method for multi-feature representation based on mutual learning and attention mechanism is proposed in the research [79], For cross-modal learning, a data fusion method combines spatiotemporal video cross-modal data and a skeleton, transforming them into multi-class tokens to address the challenge of integrating cross-modal movement data. The STAR-transformer introduces a novel cross-modal attention module that replaces the multi-head attention of ViT and has demonstrated outstanding performance through various experiments.

To evaluate the performance of this algorithm, two datasets have been utilized: Action-Penn and D+RGB-NTU. For comparison, previous State of the Art (SoTA) methods have also been employed.

The transformer-STAR algorithm has the capability to integrate diverse features, including RGB video frames, skeleton information, and shared trajectories, using multi-class tokens. A spatial-temporal cross-modal attention module has been employed as a fundamental tool for simultaneously understanding these features.

The transformer-STAR algorithm, utilizing the Transformer and spatial-temporal cross-modal attention module, is capable of combining information from various sources for optimal action recognition. These results demonstrate the high performance of this algorithm in action recognition tasks using both video data and skeletal information.

Another key challenge in HAR is deploying and generalizing features derived from sensor data across different locations and conditions. Sensor data, especially from wearable devices, can vary greatly depending on circumstances (e.g., location or activity type). These variations can lead to reduced accuracy and performance of HAR systems, particularly when the extracted features fail to generalize effectively across diverse conditions and subsets of the data.

The authors of paper [80] have proposed an adversarial learning-based transformer framework for HAR using wearable sensors in various locations through the TASKED KnowledgE-Self distillation. In this approach, a neural network consists of three main components: feature extractor, activity classifier, and topic discriminator. The feature extractor uses a transformer to map sensor data to a common feature space. The transformer is used as a larger architecture for modeling complex transformations and relationships over time and space.

This method employs the technique of self-knowledge distillation to enhance the deployment of extracted self-knowledge features on various thematic subsets. The topic discriminator and activity classifier are simultaneously trained to embed not only activity categorization but also thematic information into the extracted features. The paper utilizes common evaluation metrics such as accuracy, class-wise accuracy (Fw), and proposition-wise accuracy (Fm) to assess the system's performance. These evaluations are conducted based on distinctions between activities and the generalization ability of the extracted features across different subsets of data.

A major challenge in HAR is ensuring that features extracted from sensor data can generalize across different conditions and locations. Variations in sensor data, especially from wearable devices, can lead to reduced accuracy in HAR systems. To tackle this, the authors of [81] proposed a transformer-based approach integrated with adversarial learning.

The study in [81] has utilized a new approach called "Transformer Based Attention Consensus (TBAC)" to improve the detection of human actions in videos. This new approach is presented in the form of a transformer-based attention consensus module (TBAC). In this paper, the transformer has been employed as a key component in the TBAC module. This use of the transformer aims to enhance the extraction and aggregation of temporal features from videos. Additionally, a consensus decision algorithm (DC) has been used, which leverages multiple independent but related action recognition models to improve their performance.

datasets HMDB51 and HAA500 and accuracy metrics (Top-1 accuracy and Top-3 accuracy) were employed to assess the models' performance. The results indicated that the TBAC module and the Attention-based Decision Consensus (DC) algorithm have significantly improved the accuracy of motion detection.

HAR and Smartphone Localization Recognition (SLR) are key challenges in biosignal processing. Extracting complex temporal patterns from sequential data is critical to addressing these challenges. Traditional models, like CNNs, have been widely used, but they struggle to capture long-term and intricate relationships within the data effectively.

Shavit et al. [82] employ an architectural approach based on the Transformer model, which has proven effectiveness in sequence analysis. Various datasets, including SLR for mobile phone location identification, HAR and SHAR combining mobile location and human activity identification, with over 27 hours of recordings from 91 users, have been used.

The Transformer is employed to fuse sequences and extract temporal features from sequential data. This aids in identifying and integrating temporal patterns in sequential data.

The authors used three categories of data, SLR, HAR, and SHAR, to evaluate the performance of their method and employed accuracy as the main metric for evaluation. The results indicate that the proposed method (IMU-Transformer) consistently improves classification accuracy compared to a comparative method (IMU-CNN). This improvement is observed in challenging scenarios and data diversity within the datasets. Additionally, this paper examines the execution time of the models, demonstrating that the execution time of the Transformer architecture is higher compared to the CNN architecture. However, this difference in execution time is negligible compared to the expected time for classification.

Researchers in their study [83], proposed and implemented an Attention-based Transformer model as a novel architectural design. This model was compared to the conventional approach of LSTM recurrent neural networks. The Transformer operates based on attention mechanisms without the need for recurrent or convolutional layers. It demonstrates the ability to learn high-level features by training an end-to-end neural network, requiring significantly less time and computational resources for feature design. The final results indicate that the attention-based Transformer model has performed better than the LSTM model. Although overfitting has been observed in the Transformer pattern, the evaluation accuracy and attention feature performance of the Transformer-based model are higher compared to LSTM.

5.2. Vision-Based Transformers

The authors in [84], have used a method for HAR using 3D skeleton data. The method involves extracting spatiotemporal geometric features from the 3D skeletal joint information, which are then analyzed using a transformer encoder to recognize human activities. This model is solely based on the transformer encoder without convolutional or recurrent layers.

To evaluate the method, the authors utilized several well-known HAR datasets, including KARD, Florence 3D, UTKinect Action 3D, and MSR Action 3D. Various evaluation protocols were employed, such as data split into training/testing sets with different ratios and the "new-person" or "leave-one-actor-out" protocol. The results indicate that the proposed method exhibited significant improvements compared to many existing approaches and demonstrated good performance in recognizing human activities from 3D skeleton data.

Liu et al. [85] employed a method called 'Spatio-Temporal Transformer Networks (SSTNs)' for detecting key activities in videos. This approach utilizes deep neural network models to learn active attention regions in the spatial and temporal frames of videos. The paper employs the Transformer architecture to utilize attention mechanisms in the temporal space dependent on video features. This usage can be beneficial for learning precise relationships between different frames of the video and improving the network's capability to detect more accurate activities.

The model utilizes a cost function to learn attention regions in spatial and temporal features. The authors have employed supervised learning techniques to train the networks. A combination of STN architecture with regressive guiding of attention regions has also been used to enhance the learning process. To evaluate the performance of the model, the authors used the MPII Cooking Activities dataset. A common evaluation criterion of accuracy is used to evaluate the efficiency of the model.

As presented in [86], the authors introduced employed a novel approach using transformer neural networks for HAR . These transformer neural networks (TNNs) consist of two main components: the Record Transformer (ReT) and its extension called Vision Transformer (ViT). In this approach, ResNet50 is utilized as a Feature Extractor. For ReT (Record Transformer), transformer layers are employed to analyze sequences of features extracted from ResNet50. A transformer-based approach is used for feature extraction, analysis, and interpretation of the data. For performance evaluation, the authors have employed several evaluation metrics, including accuracy and runtime at certain stages of model execution. The performance of the models has been investigated on various datasets, including YouTube action, UCF50, UCF101, and HMDB51. The results demonstrate that the proposed models (ViT-ReT) offer significant improvements in accuracy and execution speed compared to contemporary models, especially in resource-constrained environments.

5.3. Inverse Transformers

Efficiently and dynamically adjusting the learning rate enhances the model's ability to utilize input data and optimize the learning process. Another key challenge in this domain is managing attention mechanisms and extracting relevant features from sensor data, particularly in complex models dealing with variable data types.

The authors of paper [87] , introduce a sensor-based HAR technique using deep learning, employing a reverse attention mechanism based on transformers. This reverse attention is calibrated throughout the learning period, regularizing attention modules and dynamically adjusting the learning rate. This approach outperforms other advanced methods on five general sensor-based HAR datasets. Additionally, an alternative architecture is introduced using Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and an inverse attention decoder. Various evaluation methods are employed in this article to assess the performance of HAR models.

5.4. Self-Attention Transformers

Real-time processing of sensor data in Human Activity Recognition (HAR) is challenging due to the presence of raw, noisy, and irrelevant information, which can degrade model performance. To address this, Lee et al. [88] propose an experimental system leveraging a Self-Attention Transformer for efficient data filtering and activity recognition.

In this article , an experimental system has been designed, consisting of a filtering network and an initial classifier, to investigate the impact of unfiltered data on the detection of HAR without the need for reclassification. The filtering network extracts important features from the data in a time window using the attention mechanism of the transformer. The extracted features are fed into an initial time-series-based classifier using LSTM. This initial classification is used to determine whether the collected information is sufficient for activity recognition or not. If it is sufficient, the system provides activity recognition results.

Additionally, the filtering model is used to mitigate the impact of unfiltered data in real-time settings. The purpose of the filter network is to learn the representation of sensor states in transmission windows and filter out irrelevant data related to the target activity. A transformer encoder, based solely on self-attention mechanisms, has been applied as a filter network. In this article, a real-time activity recognition system based on neural networks has been developed. This system uses a filter network to process unfiltered data. The filter network assists in real-time detection of user activities. The experimental results demonstrate that the use of the filter network significantly improves the performance of activity recognition and can improvement unfiltered data.

Sharifi et al. [89] have employed a deep learning architecture called BioMAT, built on the transformer model. This model is used for predicting joint kinematics from the signals of multiple inertial measurement units (IMUs), involving the segmentation of motion data into cycles of consistent length using IMUs and associated signals.

The BioMAT architecture, as a transformer, features self-attention layers that enable the consideration of global dependencies in data when processing sequences in parallel. This model includes an encoder section as attention layers and a decoder section that transforms the output from the representation vector to the target sequence.

To evaluate the model's performance, metrics such as Root Mean Square Error (RMSE), normalized Root Mean Square Error (nRMSE), and Pearson correlation coefficient (r) between the predicted and measured kinematics have been utilized. The results demonstrate that BioMAT outperforms ordinary deep neural network models with higher accuracy, even without the need for motion cycle segmentation, enhancing precision.

Jiang et al. have proposed a new method for identifying continuous human movements [90]. This method consists of two main elements:

1. Continuous Motion Identification Network Based on Transformer: The micro-Doppler time-frequency map input is first transformed into a sequential layer (feature encoding layer). This layer extracts different features from the input and maps them into a high-dimensional space acceptable to the transformer. A multi-channel attention mechanism transformer has been used to predict continuous movements over time. This transformer specifically utilizes the attention mechanism to predict movements over time.
2. A separate network for motion type recognition, without considering temporal information, is trained using an ordered feature encoding layer. The predicted motion information over time and the information of only the predicted sequence of movements before and after the movement are combined and evaluated based on the state transition graph to decide whether the continuous movement is valid or not. Two strategies are used to determine the types of movements in the dataset: changing the dataset labels for network training and rules to restrict the predicted movements over time. Accuracy and convergence metrics are used to evaluate the accuracy and performance of the networks.

this paper presents a new method for identifying and evaluating continuous movements, utilizing a transformer for predicting movements over time and an OCR network for recognizing movement types without time. The results indicate that this method has achieved high accuracy in identifying continuous human movements and successfully utilized the two mentioned strategies for movement evaluation.

Guo et al. [91], have employed a method for HAR using high-dimensional radar data. The initial approach involves using radar keypoints as input for Point Transformer (PT) attention networks for classification.

The Transformer initially generates high-dimensional radar-based point clouds. These point clouds include information such as three-dimensional spatial coordinates, Doppler, intensity, time, and other features related to the size and motion of objects. Subsequently, these point clouds are used as input for Point Transformer (PT) self-attention models. These models leverage the self-attention mechanism to extract important features from radar data. The self-attention mechanism allows the network to focus on different regions of the data and extract more critical information.

The results indicate that the use of these self-attention models for processing radar-based point clouds has significantly improved accuracy and overall performance. This method has been employed for high-precision detection of human activities, even in scenarios where activities are stationary with minimal Doppler motion.

In another research the transformer is directly employed as a time series processing model to extract features from signals received from smartphone sensors [92]. Due to its high parallelization capability and fast computation for time series, this model is introduced as an effective alternative to recurrent and convolutional networks in this domain. The vision transformer is also explored in this article.

The authors utilized the KU-HAR dataset, which contains 18 different categories of human activities, for training and evaluating their models. In the testing phase, they employed unseen examples from the test dataset to predict activities using the neural network. Attention matrices were employed to investigate the effectiveness of the transformer in studying relationships in time series. The accuracy metric of the best validation set was used as the primary criterion for evaluating the models.

5.5. Transformer-Based Hybrid Models

Yan et al. have employed a method called MM-HAT for HAR based on mmWave point cloud data [93]. This method utilizes an end-to-end Transformer network that introduces specific enhancements for mmWave point cloud data in the input embeddings of the transformers. MM-HAT consists of three main components: input embeddings, encoder, and decoder. Two types of inputs are utilized: point clouds and target data. Instead of the standard Transformer embeddings, a Point Cloud Representation Extractor (PRE) is designed to learn hidden representations of point clouds, and a Target Representation Extractor (TRE) is designed for embedding target data. Subsequently, the two types of embedded data are introduced as inputs to the Transformer encoder-decoder architecture.

The evaluation results include accuracy of activity recognition , scalability, and inference time. The conducted evaluations indicate that MM-HAT is competitive compared to existing methods and exhibits better scalability than other approaches.

Yi Liu and colleagues [94] have employed a new method called TransTM for HAR through collecting COTSRFID data using the device-free approach. This method consists of a combination of multiple layers of multi-scale transformer networks. In this method, raw RFID RSSI data is initially taken as input. Then, a TransTM model is employed, which is a combination of multi-scale transformer and convolutional networks. This model is used to learn behavioral features for the detection of various activity categories.

In this paper, the transformer is utilized as a pre-trained model and serves as a key component for understanding global information in the inputs. The performance of the model is evaluated using four evaluation metrics: recognition accuracy, F1-score, which is the harmonic mean of precision and recall, the number of floating-point operations per second (FLOPs), and the number of parameters. These metrics are employed for a comprehensive evaluation of the model's performance.

Wensel et al. [86], report the TransTM model has shown the best performance in HAR through the collection of RFID data compared to other advanced models. This model demonstrates a very high detection accuracy (99.1%) and has achieved better results compared to ordinary models.

In another study [95], researchers employed a lightweight transformer, as the main architecture for HAR classification and the TransFed model with a dual-layer transformer and a learning rate of 0.01 using the Adam optimizer. Additionally, a parameter for the number of blocks in the transformer was utilized. The learning transfer of TransFed was applied for joint learning in a federated environment. This configuration ensures that the model is collectively trained on different devices while preserving privacy.

The model's performance was evaluated through extensive experiments in two environments (union and central) using a newly created dataset by the researchers. Additionally, a public dataset called WISDM was used for model evaluation. Common evaluation metrics for machine learning models, including accuracy, precision, recall, and F1 score, were employed.

According to the results reported in the paper, the proposed lightweight transformer model has shown the best performance compared to CNN and RNN-based models in the evaluation metrics. Additionally, in the union environment, the lightweight transformer has also shown improvement in terms of privacy. In other words, this model has successfully performed in joint learning and privacy preservation for edge devices.

Chen et al. introduced a new model for detecting human movements in images using a transformer to extract and integrate multi-scale behavioral information [96]. The model is pre-trained using a Swin-Transformer base network and incorporates a feature integration module to extract and integrate multi-scale behavioral information. This model has

demonstrated high accuracy in detecting human movements in images.

The authors have utilized a Feature Fusion Module to enhance performance. This module is designed to prevent dependence on the last layers and integrate features from different stages of the image. to evaluate the model's performance, they have employed five different datasets, including Li-6, PPMI-24, Stanford-40 Action, Distracted Driver V1, and Distracted Driver V2. The results indicate that the Swin-Fusion model, utilizing a combination of the Feature Fusion Module and transformer optimization techniques, exhibits a more competitive performance compared to previous methods.

A hybrid approach has been employed combining multi-scale CNN and transformers to detect human activities and determine their start and end times [97]. In multi-scale convolution modules, transformers are employed to capture global features. This transformer is implemented with multi-scale convolutional transformer blocks.

To evaluate the model, the CSI dataset collected by Google Nexus 5 has been used. This dataset includes seven different human activities such as falling, sitting, walking, etc., performed in indoor environments. The proposed model has shown very good performance, achieving a weak micro F1 score of 98.37% and a strong micro F1 score of 92.81%. These results exhibit a significant improvement compared to the compared models (such as CNN, LSTM, ABLSTM, and ResNet18), indicating that the combination of CNN and Transformer as a hybrid approach is effective for HAR in CSI data.

5.6. Multimodal Transformer

In a recent study by Djenouri et al. a new method introduced called CVTN (Convolution Visual Transformer Network) for detecting and analyzing human activities from sensor data [98]. The CVTN method leverages the combination of two deep learning architectures, namely CNN and the Visual Transformer (VT) for image transformation.

The CVTN operates in two phases. In the first phase (Spatial Visual Learning), it focuses on learning spatial visual features from sensor data using CNN. Specific techniques are employed to transform temporal data into images. In the second phase (Temporal Learning), a transformer-based network is utilized to detect temporal dependencies in the sequence of spatial features. This paper uses the Kinetics dataset, which includes over 650,000 short videos covering 400 human activity categories.

using the accuracy metric, they represent how well the CVTN model is capable of correctly detecting human activities. In comparison to existing baseline methods such as DST-LSTM and Hybridnet, it demonstrates higher accuracy.

As described in [99], the authors have employed a novel method called the 'Two-Stream Transformer Network (TTN) for addressing HAR tasks. This approach utilizes the transformer architecture and two streams, namely Temporal and Spatial, to model temporal and spatial dependencies in multi-sensor sensory signals.

In this architecture, the transformer is used simultaneously in two streams to model temporal and spatial dependencies. This use of the transformer enables the model to effectively justify both temporal and spatial dependencies and integrate information from multiple sensors into its architecture. One important dataset in this paper includes the sensor attention block in the spatial stream, which has the ability to focus on the importance of sensor axes. It assists the model in extracting crucial information and making the best use of it.

To evaluate the performance of the proposed method, the authors used common metrics such as Macro F1-score on four different datasets. The results indicated that the proposed TTN model outperforms comparative models such as CNN, LSTM, ConvLSTM, ConvAE, and Transformer Encoder, especially in cases where information is collected from various sensors and different locations.

In another research, a method called "MATN" (Multi-Agent Attention-based Temporal Network) is employed for multi-sensor HAR [100]. The MATN method utilizes multi-agent attention to extract spatiotemporal features from multi-sensor data. The authors have utilized transformer networks to encode multi-sensor data.

1. Representation Learning Layer: A visibility learning layer is utilized to encode multi-sensor data into a unified representation.
2. MSTT (Multi-Agent Spatio-Temporal Transformer) Module: This module employs a transformer network to extract spatiotemporal features from each sensor.
3. Multi-Agent Collaboration Module: This module gathers the output from each agent (associated with each sensor) and learns to select effective sensors.

The transformer is employed as a key component of the MSTT Module. This network functions as a Spatio-Temporal Transformer model to extract spatial and temporal features from multi-sensor data.

The final results have shown that MATN performs well in extracting spatial-temporal features

from multi-sensor data and has the ability to generalize to various types of data. It demonstrates this capability even in cases where only specific sensors have been used.

Li et al. have employed a multi-modal HAR method called DMFT (Distilled Mid-Fusion Transformer) [101]. This approach utilizes multiple detection stages. A Unified Encoding Layer is used for each modality, such as RGB, Depth, Skeleton, Inertial, and Wi-Fi data. This layer provides a unified representation for input data from each modality, without the need for modality-specific encoder networks, yielding a consolidated feature representation.

The Multi-Modal Spatial-Temporal Transformer (MSTT) module utilizes the transformer encoder structure to extract spatial and temporal features for each modality. This module highlights the advantages of transformer-based structures over LSTM and employs self-attention mechanisms to enhance the extraction of salient features. The Temporal Mid-fusion Transformer Module (TMT) is a transformer module

designed to fuse multi-modal temporal features at the mid-fusion stage in the feature extraction process.

The results of this paper have been evaluated on two multi-modal datasets, UTD-MHAD and MMAct, under various settings. Different evaluation metrics have been used for each dataset, including Top-1 accuracy for the UTD-MHAD dataset and F1-Score for the MMAct dataset. The results indicate that the DMFT method outperforms other methods and has demonstrated effective performance in resource-constrained environments.

Table 2 provides a concise overview of recent studies exploring the application of Transformer models in HAR. These studies highlight the effectiveness of Transformers in modeling complex spatiotemporal relationships and demonstrate their superiority over traditional models across various tasks. Key takeaways include:

- Diverse Applications: Transformers are utilized for tasks ranging from gesture recognition and activity detection to multi-modal feature extraction and time-series analysis.
- Improved Accuracy: Most models outperform conventional methods, achieving high accuracy and robustness in different datasets and settings.
- Challenges: Limitations such as dependency on large datasets, imbalanced data, and computational complexity persist, necessitating further innovations.

The table emphasizes the potential of Transformers to revolutionize HAR, particularly with continued integration of advanced techniques and tailored model designs.

6. Challenges in using Transformers in Har

Transformers have demonstrated significant potential in HAR, but their application comes with specific challenges that require careful consideration. One major issue is data quality and noise, as HAR datasets often include noisy or overlapping activities and similar behaviors that reduce model accuracy. Transformers address this by using attention mechanisms to focus on relevant data sections, extracting key features to mitigate noise and improve performance. Another challenge is the limited availability of labeled data, as acquiring labeled HAR datasets can be both time-intensive and expensive. Techniques such as self-supervised and semi-supervised learning enable Transformers to learn effectively from sparse or unlabeled data, reducing reliance on extensive labeled datasets.

Variations in activities and environments also pose a problem, as they can affect the generalizability of models. By leveraging attention mechanisms, Transformers can simulate environment-specific features and adapt to new contexts through transfer learning, enhancing their versatility. Additionally, HAR models must cope with variable data, including sudden shifts in behavior or data patterns. Transformers excel in modeling long sequences, allowing them to adapt to both gradual and abrupt changes in data dynamics. In complex scenarios, such as those involving overlapping activities or diverse features, Transformers can model intricate relationships between features, enabling better management of overlaps and interdependencies.

However, computational demands remain a concern, especially for resource-constrained devices like IoT systems and smartphones. Optimization techniques, such as model pruning and quantization, can significantly reduce resource requirements without sacrificing performance. Privacy and security are also critical in HAR, as the data often involve sensitive personal information. Federated learning addresses this by enabling data processing on local devices, reducing the need for data transfers and enhancing privacy. Lastly, model interpretability is essential, particularly in applications like healthcare, where trust in model decisions is vital. Visualization techniques and attention maps can improve interpretability, making model decisions more transparent.

To effectively implement Transformers in HAR, especially on resource-constrained systems, it is crucial

Table 2. Summary of some recent studies on transformer-based HAR

		Name	Year	Methodology	Sensor model	Data collection	Assessment	Limitation	findings
Vanilla	1	Xinmei Huang [٢٤]	2023	Seq2Seq	environmental	Casas-Aruba	Accuracy 0.659 f1 score 0.964	Small sample size	Due to the mechanism of self-attention in sensor sequence processing and activity detection better than traditional models
	2	Trung Hieu Le[77]	2022	sliding window Diagnosing operation with a transformer	inertia	C.MHD,CMDFALL,DALiAc	Accuracy 99.56	Need for more extensive data sets	Gesture recognition based on accelerometer and gyroscope data
	3	Saidani [78]	2023	Enhanced data technique	Peripheral, wearable, wireless (mobile)	WISDM,PAMPA2,UCIHAR	Accuracy 98.2, 98.6 and 97.3	Imbalanced data set	Transformer and combined characteristics of spectrographs and...
	4	Dasom Ahn [٧٩]	2023	Transformer STAR	Video and skeleton-based frames	Penn-Action , NTU-RGB+D	Accuracy 98.7	Low number of data	Proper display of spatio-temporal features,Setting pairs of FAttn, ZAttn and BAttn modules.
	5	Sungho Suh[80]	2022	TASKED Transformer-based adversarial learning	wearable	Opportunity PAMAP2, MHEALTH و RealDISP	improved 3.08, 3.89, 3.39 percent points of accuracy, Fw, and Fm over the best state-of-the-art method	Ability to recognize mutual activities	Extractor with transformer to map sensor data to a common feature space
	6	Santosh Kumar Yadav [81]	2022	TBAC: Transformers Based Attention Consensus	video	HAA500,HMDB51	Accuracy 85.23 and 83.73	The use of pre-trained weights and the TBAC's dependency on a suitable base network limit its generalizability.	The TBAC module enhances the performance of CNNs by leveraging temporal features and balanced attention.
Vision-Based	7	Ahmed Snoun [84]	2022	Model only based on transformer encoder	3D skeleton image	KARD ،Florence 3D ،UTKinect Action 3D و MSR Action 3D	Accuracy 93%	Dependency on precise skeletal data and evaluation on small datasets limits the applicability of the method.	Extraction of spatio-temporal geometric features from 3D skeletal joint information
	8	Dichao Liu [85]	2019	esupervised spatial transformer networks (SSTNs)	the video	MPII Cooking Activities Dataset	accuracy Detail level SSTN for VGG-F,VGG-M,VGG-16 29.91,31.09,32.32%	Evaluation is limited to the MPII dataset and lacks advanced pooling methods.	the six SSTNs streams are complementary to each other, and fusing them bring better performance
	9	James Wensel [86]	2023	Combining ViT and ReT with complexity reduction.	the video	YouTube action, UCF50, UCF101, and HMDB51	Accuracy (up to52.64%) and inference time (up to 38.2% on average)	Lack of full utilization of parallel processing and weaker performance of Vision Transformer compared to ResNet50 in certain tasks.	ViT-ReT combines ViT and ReT to deliver faster and more accurate performance compared to previous models.
Inverse	10	Rishav Pramanik [87]	2023	Transformer-based deep inverse attention network	environmental	MHEALTH ،USC-HAD ،WHARF ، UTD-MHAD1, 2	Accuracy for MHEALTH,USC-HAD, WHARF, UTD-MHAD1, UTD-MHAD2 1,0.95 ,0.92,0.64,0.8421	Increasing the number of parameters due to the existence of an attention scheme requires training	During the self-calibrating learning process, it regularizes attention modules and dynamically adjusts the learning rate.
Self-Attention	11	Tae-Hoon Lee [88]	2023	A transformer encoder, based on the attention mechanism, as a filter network	environmental	Casas: milan,kyoto8, kyoto11	The best accuracy results for Milan, Kyoto8,Kyoto11 datasets,0.88,0.92,080	Low data complexity due to the small number of residents	Filter unbalanced data in the transition window and initial classifier of associated features
	12	Sharifi-Renani [89]	2023	BioMAT BERT architecture with an encoder	Wearable inertia	Lower extremity biomechanics dataset	On average, RMSE is 5.5	Model training with more diverse and newer data	It performs better without the need for segmentation and increases accuracy.
	13	Liubing Jiang [90]	2023	Detection of continuous	Radar based	Continuous motion on micro-Doppler features	Accuracy 94%	Data imbalance and reliance on augmentation limit generalizability	An OCR network for detecting motion types without time
	14	Zhongyuan Guo [91]	2023	Three models of self-care (point transformer)	Image	Experimental dataset DelftTU	F1 score 92.8%	Small dataset size and high manual labeling cost limit model accuracy and generalizability.	Radar-based point cloud processing improves accuracy and performance
	15	Luptáková [92]	2022	Parallelization and fast computation of time series	wearable (smartphone)	KU-HAR dataset	Accuracy 99.2%	Data imbalance and reliance on augmentation limit generalizability.	Time series processing for feature extraction from smartphone sensor signals

		Name	Year	Methodology	Sensor model	Data collection	Assessment	Limitation	findings
Hybri	16	Yi Liu [94]	2023	TransTM Multiscale transformer and Cnn	environmental	RFID	Accuracy 99.1%	Dependence on custom data and high computational cost limit the model's generalizability.	It receives low quality data and avoids the data cleaning process
	17	Ali Raza [95]	2021	A lightweight transformer and TransFed model	wearable	WISDM	Accuracy 98.74%	Dependency on custom data and high resource requirements limit generalizability.	The model is trained collectively on different devices while maintaining privacy.
	18	Tianshe ng Chen [96]	2023	Swin-Transformer	Image	Li-6 ،PPMI-24 ، Stanford40 ،AUC V1,AUC V2	Accuracy 100,97.69,96.24,94.41,92.33%	The smallness of the data set Li-6	Swin-Fusion combines fusion module and transformer optimization techniques
	19	Dejun Gao [97]	2022	Transformers to record global characteristics	environmental	CSI	Weak micro F1 score of 98.37% and strong of 92.81%	Dependence on WiFi data and the impact of environmental factors limit the model's applicability.	Combination of CNN and transformers to identify human activities and determine start and end times
Multi modal	20	Jingche ng Li [101]	2023	DMFT (Distilled Fusion Intermediate Transformer)	RGB ، Depth ، Inertial و Wi-Fi	UTD-MHAD و MMAct	Top-1 accuracy for UTD-MHAD more than96% and F1-Score for MMAct more than91%	Intersubject variability in the detection of multimodal human activities	Transformer to extract spatial and temporal features for each modality
	21	Shuo Xiao [99]	2022	Two current transformer network (TTN)	Multimodal sensor	USC_HAD ،Opportunity ،PAMAP2 و Skoda	Score 1F macro, 0.99 0.69, 0.56 and 0.95	Weaker performance in short-duration similar activities and instability due to reliance on positional encoding.	Transformer simultaneously in two currents to model temporal and spatial dependencies
	22	Jingche ng Li [100]	2022	MATN-A temporal network based on multifactorial attention	Multimodal sensor	UTD-MHAD , MMAct	Accuracy for UTD-MHAD,92.72%, F1-Scorefor MMAct,91.85%	Limited scalability and sensitivity to noisy data.	Extracting spatio-temporal features from multi-sensor data and the ability to generalize to all types of data

design lightweight and compact architectures. These architectures should balance minimal resource consumption with high accuracy and efficiency in activity recognition, ensuring they meet practical constraints while maintaining strong performance.

By addressing these challenges, HAR models leveraging Transformers can achieve higher accuracy, adaptability, and efficiency, ensuring robust activity recognition even in complex scenarios.

7. Conclusions

In this review, we have provided a comprehensive overview of the transformative impact of transformer-based architectures on HAR. Our analysis of recent studies since 2019 highlights the significant advancements made in leveraging transformers to enhance the accuracy and efficiency of HAR systems. The ability of transformers to model long-term dependencies and process complex input data in parallel has opened new avenues for research and application in smart environments.

We explored various innovative architectures, including attention-based, vision-based, and hybrid models, demonstrating their unique capabilities in handling diverse data types from sensor-based and vision-based technologies. Additionally, our review emphasized the critical role of datasets in training effective HAR models, underscoring the need for high-quality, diverse data to support robust model performance.

Despite the advancements in transformer architectures for HAR, several challenges must be addressed to fully leverage their potential. Key issues include the high computational resource requirements, the need for model interpretability, and the ability to generalize across diverse environments. Future research should prioritize the development of lightweight transformer models that deliver robust performance while being optimized for real-time applications on mobile and embedded devices.

Additionally, innovative techniques such as self-supervised and federated learning should be explored to mitigate reliance on labeled datasets and address privacy concerns. These methodologies can enhance the applicability of transformers in real-world scenarios, particularly in environments characterized by varying sensor data and user behaviors.

It is essential for future studies to investigate cross-domain adaptation strategies that enable models trained in one context to effectively generalize to others. As HAR increasingly incorporates diverse data sources, including wearable sensors, cameras, and environmental sensors, research should focus on multimodal fusion techniques. These approaches can improve the integration of varied inputs, thereby enhancing recognition accuracy and system robustness.

The inherent black-box nature of transformer models presents challenges in understanding their decision-making processes. Therefore, future work should emphasize developing methods that enhance model interpretability, allowing researchers and practitioners to discern how specific features contribute to activity recognition outcomes.

Moreover, optimizing transformers for real-time processing capabilities is crucial for their practical deployment in smart environments. This includes

investigating efficient training algorithms, minimizing inference latency, and implementing online learning techniques that enable models to adapt continuously to new data.

Finally, addressing ethical considerations related to privacy and data security is paramount. Future research should explore frameworks that ensure user consent, data anonymization, and secure data handling practices within HAR systems.

In conclusion, the future of transformer-based approaches in HAR is promising. By tackling these challenges and pursuing these research directions, we can significantly enhance the effectiveness of HAR systems, fostering innovative applications that improve safety, efficiency, and quality of life in smart environments.

Declarations

Funding

This paper received no specific grant from any funding agency in the public, commercial, or notfor-profit sectors.

Data availability

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Authors' contributions

FSL: Study design, conceptualization, supervision, review and approval, drafting the manuscript, revision of the manuscript;
RF: Study design, interpretation of the results, drafting the manuscript, revision of the manuscript.

Conflict of interest

We have no conflict of interest on this paper.

Reference

[1] N. A. Choudhury and B. Soni, "An Adaptive Batch Size based-CNN-LSTM Framework for Human Activity Recognition in Uncontrolled Environment," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 10, pp. 10379-10387, Oct. 2023, https://doi.org/10.1109/TII.2022.32295222023.

[2] A. Subasi, K. Khateeb, T. Brahimi, and A. Sarirete, "Human activity recognition using machine learning methods in a smart healthcare environment," in *Innovation in health informatics*: Elsevier, 2020, pp. 123-144. https://doi.org/10.1016/B978-0-12-819043-2.00005-8

[3] A. Hayat, F. Morgado-Dias, B. P. Bhuyan, and R. Tomar, "Human activity recognition for elderly people using machine and deep learning approaches," *Information*, vol. 13, no. 6, p. 275, 2022. https://doi.org/10.3390/info13060275

[4] M. A. Khatun *et al.*, "Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1-16, 2022. https://doi.org/10.1109/JTEHM.2022.3177710

[5] A. Sunil, M. H. Sheth, and E. Shreyas, "Usual and unusual human activity recognition in video using deep learning and artificial intelligence for security applications," in *2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2021: IEEE, pp. 1-6. https://doi.org/10.1109/ICECCT52121.2021.9616791

[6] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274-176285, 2020. https://doi.org/10.1109/ACCESS.2020.3026823

[7] S. N. Kabataş and D. Sarıkaya, "Surgical activity recognition with transformer networks," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*, 2021: IEEE, pp. 1-4. https://doi.org/10.1109/SIU53274.2021.9477969

[8] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern recognition letters*, vol. 119, pp. 3-11, 2019. https://doi.org/10.1016/j.patrec.2018.02.010

[9] 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. https://doi.org/10.1016/j.patcog.2020.107561

[10] S. Simonsson, F. D. Casagrande, and E. Zouganeli, "Use of Clustering Algorithms for Sensor Placement and Activity Recognition in Smart Homes," *IEEE Access*, vol. 11, pp. 9415-9430, 2023. https://doi.org/10.1109/ACCESS.2023.3239265.

[11] M. Gochoo, T.-H. Tan, S.-H. Liu, F.-R. Jean, F. S. Alnajjar, and S.-C. Huang, "Unobtrusive activity recognition of elderly people living alone using anonymous binary sensors and DCNN," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 693-702, 2018. 1 https://doi.org/0.1109/JBHI.2018.2833618

[12] Z. Gu, T. He, Z. Wang, and Y. Xu, "Device-free human activity recognition based on dual-channel transformer using WiFi signals," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1 p. 4598460, 2022. https://doi.org/10.1155/2022/4598460

[13] A. Dahou, M. A. Al-qaness, M. Abd Elaziz, and A. M. Helmi, "MLCNNwav: Multi-level Convolutional Neural Network with Wavelet Transformations for Sensor-based Human Activity Recognition," *IEEE Internet of Things Journal*, vol. 11, no. 1, pp. 820-828, 1 Jan.1, 2024, https://doi.org/10.1109/JIOT.2023.3286378.

[14] A. I. Middya, S. Kumar, and S. Roy, "Activity recognition based on smartphone sensor data using shallow and deep learning techniques: A Comparative Study," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 9033-9066, 2024/01/01 2024, https://doi.org/10.1007/s11042-023-15751-w.

[15] P. Kumar, S. Chauhan, and L. K. Awasthi, "Human Activity Recognition (HAR) Using Deep Learning: Review, Methodologies, Progress and Future Research Directions," *Archives of Computational Methods in Engineering*, vol. 31, no. 1, pp. 179-219, 2024. https://doi.org/10.1007/s11831-023-09986-x

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015, https://doi.org/10.1038/nature14539.

[17] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295-2329, 2017. https://doi.org/10.1109/JPROC.2017.2761740.

[18] S. Ramasamy Ramamurthy and N. Roy, "Recent trends in machine learning for human activity recognition—A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1254, 2018. https://doi.org/10.1002/widm.1254

[19] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. https://doi.org/10.48550/arXiv.1810.04805

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, 2020: Springer, pp. 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[22] T. Perrett, A. Masullo, T. Burghardt, M. Mirmehdi, and D. Damen, "Temporal-relational cross Transformers for few-shot action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 475-484. https://doi.org/10.48550/arXiv.2101.06184

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014. https://doi.org/10.48550/arXiv.1409.0473

[24] X. Huang and S. Zhang, "Human Activity Recognition based on Transformer in Smart Home," in *Proceedings of the 2023 2nd Asia Conference on Algorithms, Computing and Machine Learning*, 2023, pp. 520-525. https://doi.org/10.1145/3590003.3590100

[25] K. T. Chitty-Venkata, S. Mittal, M. Emani, V. Vishwanath, and A. K. Somani, "A survey of techniques for optimizing transformer inference," *Journal of Systems Architecture*, vol. 144, p. 102990, 2023. https://doi.org/10.1016/j.sysarc.2023.102990

[26] A. Hussain, T. Hussain, W. Ullah, and S. W. Baik, "Vision Transformer and Deep Sequence Learning for Human Activity Recognition in Surveillance Videos," *Computational Intelligence and Neuroscience*, vol. 2022, p. 3454167, 2022/04/04 2022, https://doi.org/10.1155/2022/3454167.

[27] G. Pareek, S. Nigam, and R. Singh, "Modeling transformer architecture with attention layer for human activity recognition," *Neural Computing and Applications*, vol. 36, pp. 5515–5528, 2024, https://doi.org/10.1007/s00521-023-09362-7.

[28] F. MortezaPour Shiri, T. Perumal, N. Mustapha, R. Mohamed, M. A. B. Ahmadon, and S. Yamaguchi, "A Survey on Multi-Resident Activity Recognition in Smart Environments," *arXiv e-prints*, p. arXiv: 2304.12304, 2023. https://doi.org/10.48550/arXiv.2304.12304

[29] D. Bouchabou, S. M. Nguyen, C. Lohr, B. LeDuc, and I. Kanellos, "A survey of human activity recognition in smart homes based on IoT sensors algorithms: Taxonomies, challenges, and opportunities with deep learning," *Sensors*, vol. 21, no. 18, p. 6037, 2021. https://doi.org/10.3390/s21186037

[30] S. Qiu *et al.*, "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. https://doi.org/10.1016/j.inffus.2021.11.006

[31] P. Wei *et al.*, "Impact Analysis of Temperature and Humidity Conditions on Electrochemical Sensor Response in Ambient Air Quality Monitoring," *Sensors*, vol. 18, no. 2, p. 59, 2018. https://doi.org/10.3390/s18020059

[32] J. Saini, M. Dutta, and G. Marques, "Sensors for indoor air quality monitoring and assessment through Internet of Things: a systematic review," *Environmental Monitoring and Assessment*, vol. 193, no. 2, p. 66, 2021, https://doi.org/10.1007/s10661-020-08781-6.

[33] F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi, "Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey," *IEEE access*, vol. 8, pp. 210816-210836, 2020. https://doi.org/10.1109/ACCESS.2020.3037715.

[34] Y. Cheng, K. Wang, H. Xu, T. Li, Q. Jin, and D. Cui, "Recent developments in sensors for wearable device applications," *Analytical and bioanalytical chemistry*, vol. 413, no. 24, pp. 6037-6057, 2021. https://doi.org/10.1007/s00216-021-03602-2

[35] Z. Zhang and H. Liu, "Application of sports wearable sensor based on edge computing in sports industry," *Measurement: Sensors*, vol. 31, p. 101008, 2024, https://doi.org/10.1016/j.measen.2023.101008.

[36] Y. Liu, C. Ouyang, Z. Wang, J. Xu, X. Mu, and A. L. Swindlehurst, "Near-Field Communications: A Comprehensive Survey," *arXiv preprint arXiv:2401.05900*, 2024. https://doi.org/10.48550/arXiv.2401.05900

[37] S. Abdullah, G. Xiao, and R. E. Amaya, "A review on the history and current literature of metamaterials and its applications to antennas & radio frequency identification (RFID) devices," *IEEE Journal of Radio Frequency Identification*, vol. 5, no. 4, pp. 427-445, 2021. https://doi.org/10.1109/JRFID.2021.3091962.

[38] P. Fard Moshiri, R. Shahbazian, M. Nabati, and S. A. Ghorashi, "A CSI-Based Human Activity Recognition Using Deep Learning," *Sensors*, vol. 21, no. 21, p. 7225, 2021. https://doi.org/10.3390/s21217225

[39] X. Chen, Y. Zou, C. Li, and W. Xiao, "A Deep Learning Based Lightweight Human Activity Recognition System Using Reconstructed WiFi CSI," *IEEE Transactions on Human-Machine Systems*, vol. 54, no. 1, pp. 68-78, 2024. https://doi.org/10.1109/THMS.2023.3348694

[40] M. Kulbacki *et al.*, "Intelligent Video Analytics for Human Action Recognition: The State of Knowledge," *Sensors*, vol. 23, no. 9, p. 4258, 2023. https://doi.org/10.3390/s23094258

[41] H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, "Video-based human action recognition using deep learning: a review," *arXiv preprint arXiv:2208.03775*, 2022. https://doi.org/10.48550/arXiv.2208.03775

[42] H. B. Mahajan *et al.*, "Automatic robot Manoeuvres detection using computer vision and deep learning techniques: a perspective of internet of robotics things (IoRT)," *Multimedia Tools and Applications*, vol. 82, no. 15, pp. 23251-23276, 2023. https://doi.org/10.1007/s11042-022-14253-5

[43] K. Okarma, "Applications of Computer Vision in Automation and Robotics," *Applied Sciences*, vol. 10, no. 19, p. 6783, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/19/6783.

[44] V. Selvaraju, N. Spicher, R. Swaminathan, and T. M. Deserno, "Face detection from in-car video for continuous health monitoring," in *Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications*, 2022, vol. 12037: SPIE, pp. 49-56. https://doi.org/10.1117/12.2612911

[45] R. Pandey, S. Saha, N. Yathiraju, I. S. Abdulrahman, R. Nittala, and V. Tripathi, "Integration of RFID and Image Processing for Surveillance ABased Security System," in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India, IEEE, 2023, pp. 380-384. https://doi.org/10.1109/ICACITE57410.2023.10182987.

[46] R. P. Singh, H. Srivastava, H. Gautam, R. Shukla, and R. K. Dwivedi, "An Intelligent Video Surveillance System using Edge Computing based Deep Learning Model," in *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, Bengaluru, India, IEEE, 2023, pp. 439-444. https://doi.org/10.1109/IDCIoT56793.2023.10053404

[47] Y. Myagmar-Ochir and W. Kim, "A Survey of Video Surveillance Systems in Smart City," *Electronics*, vol. 12, no. 17, p. 3567, 2023. https://doi.org/10.3390/electronics12173567

[48] J. Kaur and W. Singh, "A systematic review of object detection from images using deep learning," *Multimedia Tools and Applications*, vol. 83, no. 4, pp. 12253-12338, 2024, https://doi.org/10.1007/s11042-023-15981-y.

[49] M. S, G. E. R, M. N, K. J, V. N, and R. P, "Detection and Recognition of Face Using Deep Learning," in *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, Coimbatore, India, 2023, pp. 72-76, https://doi.org/10.1109/ICISCoIS56541.2023.10100435.

[50] S. M. A. Sharif, R. A. Naqvi, M. Biswas, and W. K. Loh, "Deep Perceptual Enhancement for Medical Image Analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 10, pp. 4826-4836, 2022, https://doi.org/10.1109/JBHI.2022.3168604.

[51] R. Azad *et al.*, "Advances in medical image analysis with vision transformers: A comprehensive review," *Medical Image Analysis*, vol. 91, p. 103000., 2023. https://doi.org/10.1016/j.media.2023.103000

[52] A. Singh, P. Li, K. K. Singh, and V. Saravana, "Real-time intelligent image processing for security applications," *Journal of Real-Time Image Processing*, vol. 18, no. 5, pp. 1787-1788, 2021, https://doi.org/10.1007/s11554-021-01169-w.

[53] Q. Ming, L. Miao, Z. Zhou, J. Song, Y. Dong, and X. Yang, "Task interleaving and orientation estimation for high-precision oriented object detection in aerial images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 241-255, 2023. https://doi.org/10.1016/j.isprsjprs.2023.01.001

[54] A. Ettalibi, A. Elouadi, and A. Mansour, "AI and Computer Vision-based Real-time Quality Control: A Review of Industrial Applications," *Procedia Computer Science*, vol. 231, pp. 212-220, 2024, https://doi.org/10.1016/j.procs.2023.12.195.

[55] R. Szeliski, *Computer vision: algorithms and applications*. Springer Nature, 2022.

[56] J. B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," *arXiv preprint arXiv:1911.03584*, 2019. https://doi.org/10.48550/arXiv.1911.03584

[57] K. Duan, S. Bao, Z. Liu, and S. Cui, "Exploring vision transformer: classifying electron-microscopy pollen images with transformer," *Neural Computing and Applications*, vol. 35, no. 1, pp. 735-748, 2023, https://doi.org/10.1007/s00521-022-07789-y.

[58] B. Chen, F. Zhang, X. Yang, Q. Ning, and V. C. M. Leung, "Vision transformer with multiple granularities for person re-identification," *Neural Computing and Applications*, vol. 35, no. 31, pp. 23213-23223, 2023, https://doi.org/10.1007/s00521-023-08913-2.

[59] R. Yin and J. Yin, "A Two-stream Hybrid CNN-Transformer Network for Skeleton-based Human Interaction Recognition," *arXiv preprint arXiv:2401.00409*, 2023. https://arxiv.org/html/2401.00409v1

[60] B. Liu and S. Fang, "Multi-level wavelet network based on CNN-Transformer hybrid attention for single image deraining," *Neural Computing and Applications*, vol. 35, no. 30, pp. 22387-22404, 2023, https://doi.org/10.1007/s00521-023-08899-x.

[61] G. R. Hemalakshmi, M. Murugappan, M. Y. Sikkandar, S. S. Begum, and N. B. Prakash, "Automated retinal disease classification using hybrid transformer model (SViT) using optical coherence tomography images," *Neural Computing and Applications*, vol. 36, pp. 9171–9188, 2024, https://doi.org/10.1007/s00521-024-09564-7.

[62] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12113-12132, Oct. 2023, https://doi.org/10.1109/TPAMI.2023.3275156.

[63] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *2012 16th international symposium on wearable computers*, Newcastle, UK, IEEE, 2012, pp. 108-109, https://doi.org/10.1109/ISWC.2012.13

[64] D. J. Cook, "Learning Setting-Generalized Activity Models for Smart Spaces," In *IEEE Intell Syst*, vol. 27, no. 1, pp. 32-38, Sep 9 2010, https://doi.org/10.1109/mis.2010.112.

[65] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Esann*, 2013, vol. 3, p. 3.

[66] T. H. Tran *et al.*, "A multi-modal multi-view dataset for human fall analysis and preliminary investigation on modality," in *2018 24th International Conference on Pattern Recognition (ICPR)*, Beijing, China, 2018, pp. 1947-1952, https://doi.org/10.1109/ICPR.2018.8546308.

[67] H. Wei, P. Chopada, and N. Kehtarnavaz, "C-MHAD: Continuous Multimodal Human Action Dataset of Simultaneous Video and Inertial Sensing," *Sensors*, vol. 20, no. 10, p. 2905, 2020. https://doi.org/10.3390/s20102905

[68] H. Leutheuser, D. Schuldhaus, and B. M. Eskofier, "Hierarchical, multi-sensor based classification of daily life activities: comparison with state-of-the-art algorithms using a benchmark dataset," *PloS one*, vol. 8, no. 10, p. e75196, 2013. https://doi.org/10.1371/journal.pone.0075196

[69] W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," *2013 IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, 2013, pp. 2248-2255, https://doi.org/10.1109/ICCV.2013.280.

[70] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," *2016*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 1010-1019, https://doi.org/10.1109/CVPR.2016.115

[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. https://doi.org/10.3390/s16010115

[72] 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 *2015 IEEE International conference on image processing (ICIP)*, Quebec City, QC, Canada, IEEE, 2015, pp. 168-172. https://doi.org/10.1109/ICIP.2015.7350781.

[73] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 8657-8666, https://doi.org/10.1109/ICCV.2019.00875

[74] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74-82, 2011. https://doi.org/10.1145/1964897.1964918

[75] J. Wan, M. J. O'grady, and G. M. O'Hare, "Dynamic sensor event segmentation for real-time activity recognition in a smart home context," *Personal and Ubiquitous Computing*, vol. 19, pp. 287-301, 2015. https://doi.org/10.1007/s00779-014-0824-x

[76] C. Wojek, G. Dorkó, A. Schulz, and B. Schiele, "Sliding-windows for rapid object class localization: A parallel technique," in *Joint Pattern Recognition Symposium*, 2008: Springer, pp. 71-81. https://doi.org/10.1007/978-3-540-69321-5_8

[77] T. H. Le, T. H. Tran, and C. Pham, "Human action recognition from inertial sensors with Transformer," in *2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, Phu Quoc, Vietnam, 2022, pp. 1-6, https://doi.org/10.1109/MAPR56351.2022.9924794.

[78] O. Saidani, M. Alsafyani, R. Alroobaea, N. Alturki, R. Jahangir, and L. J. Menzli, "An Efficient Human Activity Recognition using Hybrid Features and Transformer Model," *IEEE Access*, vol. 11, pp. 101373-101386, 2023, https://doi.org/10.1109/ACCESS.2023.3314492

[79] D. Ahn, S. Kim, H. Hong, and B. C. Ko, "STAR-Transformer: a spatio-temporal cross attention transformer for human action recognition," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2023, pp. 3319-3328, https://doi.org/10.1109/WACV56688.2023.00333

[80] S. Suh, V. F. Rey, and P. Lukowicz, "TASKED: Transformer-based Adversarial learning for human activity recognition using wearable sensors via Self-KnowledgE Distillation," *Knowledge-Based Systems*, vol. 260, p. 110143, 2023. https://doi.org/10.1016/j.knosys.2022.110143

[81] S. K. Yadav, S. B. Kera, R. V. Gonela, K. Tiwari, H. M. Pandey, and S. A. Akbar, "Tbac: transformers based attention consensus for human activity recognition," in *2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, 2022, pp. 1-8, https://doi.org/10.1109/IJCNN55064.2022.9892906.

[82] Y. Shavit and I. Klein, "Boosting inertial-based human activity recognition with transformers," *IEEE Access*, vol. 9, pp. 53540-53547, 2021. https://doi.org/10.1109/ACCESS.2021.3070646

[83] G. Augustinov *et al.*, "Transformer-Based Recognition of Activities of Daily Living from Wearable Sensor Data," in *Proceedings of the 7th International Workshop on Sensor-based Activity Recognition and Artificial Intelligence*, 2022, pp. 1-8. https://doi.org/10.1145/3558884.3558895

[84] A. Snoun, T. Bouchrika, and O. Jemai, "View-invariant 3D Skeleton-based Human Activity Recognition based on Transformer and Spatio-temporal Features," in *ICPRAM*, 2022, pp. 706-715. https://doi.org/10.5220/0010895300003122

[85] D. Liu, Y. Wang, and J. Kato, "Supervised Spatial Transformer Networks for Attention Learning in Fine-grained Action Recognition," in *VISIGRAPP (4: VISAPP)*, 2019, pp. 311-318. https://doi.org/10.5220/0007257803110318

[86] J. Wensel, H. Ullah, and A. Munir, "ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos," *IEEE Access*, vol. 11, pp. 72227-72249, 2023, https://doi.org/10.1109/ACCESS.2023.3293813.

[87] R. Pramanik, R. Sikdar, and R. Sarkar, "Transformer-based deep reverse attention network for multi-sensory human activity recognition," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106150, 2023. https://doi.org/10.1016/j.engappai.2023.106150

[88] T. H. Lee, H. Kim, and D. Lee, "Transformer based Early Classification for Real-time Human Activity Recognition in Smart Homes," in *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, 2023, pp. 410-417. https://doi.org/10.1145/3555776.3577693

[89] M. Sharifi-Renani, M. H. Mahoor, and C. W. Clary, "BioMAT: An Open-Source Biomechanics Multi-Activity Transformer for Joint Kinematic Predictions Using Wearable Sensors," *Sensors*, vol. 23, no. 13, p. 5778, 2023. https://doi.org/10.3390/s23135778

[90] L. Jiang, M. Wu, L. Che, X. Xu, Y. Mu, and Y. Wu, "Continuous Human Motion Recognition Based on FMCW Radar and Transformer," *Journal of Sensors*, vol. 2023, no. 1, p. 2951812, 2023. https://doi.org/10.1155/2023/2951812

[91] Z. Guo, R. G. Guendel, A. Yarovoy, and F. Fioranelli, "Point Transformer-Based Human Activity Recognition Using High-Dimensional Radar Point Clouds," in *2023 IEEE Radar Conference (RadarConf23)*, TX, USA, IEEE, 2023, pp. 1-6. https://doi.org/10.1109/RadarConf2351548.2023.10149679

[92] I. D. Luptáková, M. Kubovčík, and J. Pospíchal, "Wearable Sensor Based Human Activity Recognition with Transformer," *Sensors*, vol. 22, no. 5, p. 1911; 2022, https://doi.org/10.3390/s22051911.

[93] J. Yan, X. Zeng, A. Zhou, and H. Ma, "MM-HAT: Transformer for Millimeter-Wave Sensing Based Human Activity Recognition," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*, Rio de Janeiro, Brazil, IEEE, 2022, pp. 547-553, https://doi.org/10.1109/GLOBECOM48099.2022.10000673.

[94] Y. Liu *et al.*, "TransTM: A device-free method based on time-streaming multiscale transformer for human activity recognition," *Defence Technology*, vol. 32, pp. 619-628. 2023. https://doi.org/10.1016/j.dt.2023.02.021

[95] A. Raza, K. P. Tran, L. Koehl, S. Li, X. Zeng, and K. Benzaidi, "Lightweight transformer in federated setting for

human activity recognition," *arXiv preprint arXiv:2110.00244*, 2021. https://doi.org/10.48550/arXiv.2110.00244

[96] T. Chen and L. Mo, "Swin-fusion: swin-transformer with feature fusion for human action recognition," *Neural Processing Letters*, vol. 55, pp. 11109–11130, 2023. https://doi.org/10.1007/s11063-023-11367-1

[97] D. Gao and L. Wang, "Multi-scale Convolution Transformer for Human Activity Detection," in *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*, Chengdu, China, IEEE, 2022, pp. 2171-2175, https://doi.org/10.1109/ICCC56324.2022.10065954.

[98] Y. Djenouri and A. N. Belbachir, "A Hybrid Visual Transformer for Efficient Deep Human Activity Recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Paris, France, 2023, pp. 721-730, https://doi.org/10.1109/ICCVW60793.2023.00080.

[99] S. Xiao, S. Wang, Z. Huang, Y. Wang, and H. Jiang, "Two-stream transformer network for sensor-based human activity recognition," *Neurocomputing*, vol. 512, pp. 253-268, 2022. https://doi.org/10.1016/j.neucom.2022.09.099

[100] J. Li, L. Yao, B. Li, X. Wang, and C. Sammut, "Multi-agent Transformer Networks for Multimodal Human Activity Recognition," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 1135-1145. https://doi.org/10.1145/3511808.3557402

[101] J. Li, L. Yao, B. Li, and C. Sammut, "Distilled Mid-Fusion Transformer Networks for Multi-Modal Human Activity Recognition," *arXiv preprint arXiv:2305.03810*, 2023. https://doi.org/10.48550/arXiv.2305.03810

Fatemeh Sadat Lesani received the B.S. degree in computer engineering from the University of Qom, in 2012 and the M.S. and Ph.D. degree in information technology engineering from the University of Qom, in 2014 and 2019, respectively. Since 2023, she has been an Assistant Professor with the Department of Electrical and Computer Engineering, Qom University of Technology, Qom, Iran. Her research interests include medical image processing, pattern recognition, pervasive computing, context aware systems, and Internet of Things.

Miss Lesani's awards and honors include the National Elite Foundation Academic Award in 2016 and 2017, and the Young Researcher Award of Qom province in 2018.

Ronak Fattahi received her bachelor's degree from Kermanshah Islamic Azad University in 1400 and is a master's student in artificial intelligence at Shahab Danesh University in Qom. Interested in deep learning and machine learning research.