Wi-Fitness: Improving Wi-Fi Sensing With Video Perception for Smart Fitness

Mengli Wei, Daguo Zhao, Lei Zhang, *Member, IEEE*, Cheng Wang, *Senior Member, IEEE*, Yonggang Zhang, *Member, IEEE*, Qi Wang, *Member, IEEE*, Xiaochen Fan, Yaping Zhong, and Shiwen Mao, *Fellow, IEEE*

***Abstract*—With advancements in AI, smart home gyms are becoming increasingly popular for providing fitness assistance in indoor environments. In this research, we propose a layer-by-layer framework, called Wi-Fitness, which bridges video perception with Wi-Fi sensing for smart fitness. At the data preprocessing layer, the singular value decomposition-based channel state information denoising mechanism is leveraged to do the Wi-Fi data calibration. Diverse and high-quality training samples are generated by a random quantization-based data augmentation method. At the bimodal fusion layer, the heterogeneity between the Wi-Fi and video is mitigated by the local attention mechanism and the bimodal feature integration mechanism. For the video modality, the attention-based spatio-temporal graph convolutional network (AST-GCN Net) is proposed to refine spatial information. The spatio-temporal semantic alignment module is proposed to transfer spatial information from video to Wi-Fi and maintain temporal consistency across modalities. The fitness assessment layer provides exercise visualization. The generalization of Wi-Fitness is enhanced by layer-by-layer collaboration. Wi-Fitness demonstrates its effectiveness by achieving an average F1-Score of 92.68% in three typical indoor environments.**

***Index Terms*—Action recognition, channel state information (CSI), multimodal, skeleton, Wi-Fi.**

Received 6 August 2024; revised 14 September 2024; accepted 1 October 2024. Date of publication 15 October 2024; date of current version 24 January 2025. This work was supported in part by the Natural Science Foundation of Tianjin under Grant 22JCYBJC00120; in part by the Key Laboratory of Embedded System and Service Computing (Tongji University), Ministry of Education under Grant ESSCKF 2024-04; in part by Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology Shandong Academy of Sciences under Grant 2023ZD036; in part by the Fundamental Research Funds for the Central Universities, JLU under Grant 93K172022K09; and in part by the Open Project of Tianjin Key Laboratory of Optoelectronic Detection Technology and System under Grant 2024LODTS108. *(Daguo Zhao and Mengli Wei are co-first authors.) (Corresponding authors: Yaping Zhong; Shiwen Mao.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Medical Ethics Committee of Wuhan Sports University under Application No. 2023029, and performed in line with the Declaration of Helsinki.

Mengli Wei is with the Sports Big-Data Research Center, Wuhan Sports University, Wuhan 430079, China (e-mail: weimengli@whsu.edu.cn).

Daguo Zhao is with the College of Intelligence and Computing and the Tianjin Key Laboratory of Advanced Network Technology and Application, Tianjin University, Tianjin 300050, China (e-mail: iethan@tju.edu.cn).

Lei Zhang is with the College of Intelligence and Computing and the Tianjin Key Laboratory of Advanced Network Technology and Application, Tianjin University, Tianjin 300050, China, and also with Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China (e-mail: lzhang@tju.edu.cn).

Cheng Wang is with the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 200092, China (e-mail: cwang@tongji.edu.cn).

Yonggang Zhang is with the Key Laboratory of Symbolic Computation and Knowledge Engineer, Ministry of Education, Jilin University, Changchun 130012, China (e-mail: zhangyg@jlu.edu.cn).

Qi Wang is with the School of Electronic and Information Engineering and Tianjin Key Laboratory of Optoelectronic Detection Technology and System, Tiangong University, Tianjin 300387, China (e-mail: wangqitju@163.com).

Xiaochen Fan is with the Institute for Electronics and Information Technology, Tsinghua University, Tianjin 300467, China, and also with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: fanxiaochen33@gmail.com).

Yaping Zhong is with the Sports Big-Data Research Center, Wuhan Sports University, Wuhan 430079, China (e-mail: zhongyaping@whsu.edu.cn).

Shiwen Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA (e-mail: smao@ieee.org).

Digital Object Identifier 10.1109/JIOT.2024.3476291

I. Introduction

FITNESS activities can be conducted in gyms, or sports fields, with the help of the coaches. However, lots of people do not have time to go to the gyms. Therefore, they seek more convenient alternatives to do the exercise [1]. The online tutorials can provide the guidance for the fitness. Though "home gym" offers convenience and flexibility, whenever the standard guidance can be followed, the health can be improved. However, without a coach, whether the exerciser really follows the standard baselines and whether the health can be improved as well as the possible physical injuries can be prevented are the questions. With the quick advances of AI, smart fitness assistants [2], [3], [4], which can guarantee the standardization, effectiveness, and security of fitness become increasingly popular.

With the advancement of sensing technology, there are a variety of smart fitness assistants, such as exercising monitoring BikeNet [5], accelerometer sensors for motion detection [6], wearable devices for fitness coach FitCoach [7], and smartwatch for workout tracking MiLift [8]. The various devices can help users engage in effective fitness. However, these sensing devices may restrict user comfort and freedom due to their needs to be carried on the body. Therefore, video-based smart fitness assistants have become popular, such as PhyCoVIS [9]. However, these kinds of methods still have issues with sensitivity to lighting, obstacles, and privacy concerns.

Wi-Fi has widely evolved from just a communication tool. It is now used as an integrated sensing method. This highlights its potential to become a qualified technical support of smart fitness assistants. When applied to a smart fitness assistant, WiFi sensing faces challenges of environmental dependence and generalization. Due to Wi-Fi signals' low-spatial resolution, models trained in specific environments suffer significant performance degradation even with

minor changes like furniture relocation or subject orientation variations. Mitigating this dependency is crucial for enhancing model generalization across real-world scenarios.

In this study, a smart fitness assistant called Wi-Fitness is proposed. By automatically logging fine-grained fitness data and evaluating exercise performance, Wi-Fitness offers personalized fitness assistance. The workouts are analyzed by both short-duration and long-duration exercise execution and the corresponding feedback is offered. The video is introduced to help Wi-Fi improve the generalization through layer-by-layer collaboration. However, the bimodal fusion still faces the following key challenges.

1) There is a significant difference and high heterogeneity between video perception and Wi-Fi sensing. Even for the same signal, there is a predominant heterogeneity between the temporal and frequency domain features. To effectively enrich Wi-Fi spatial information with video, it is necessary to maintain the consistency of spatio-temporal information through semantic alignment. Current research [10], [11] lacks a comprehensive understanding of the adaptation between the physical mechanisms of radio signals and the existing classification models, resulting in suboptimal performance.
2) High-level semantic information provides a concise body layout that illustrates the relationships between limbs, facilitating the understanding of the exercise. Local behavior information reveals subtle behavioral details, enabling a precise grasp of human dynamics. However, the coarse-grained behavior of Wi-Fi poses challenges for accurately capturing geometric features and local information. Current Wi-Fi-based behavior recognition studies lack the exploration of spatial information connections between keypoints. Thus, establishing correlations between keypoints remains challenging [12], [13].

In the beginning, the professionals perform a series of fitness movements to make a personalized movement profile. This profile acts as a baseline for subsequent exercise assessment. There are six predefined fitness movements, including push-ups, sit-ups, lateral raises, squats, bench presses, and leg raises. During the exercise, a user can pick up any one or more fitness movement combinations from the six predefined ones and do them continuously. When a user does an exercise, Wi-Fitness continuously monitors a user's movements and makes comparisons against the established profile to detect the activity deviations from the corresponding professionals'. To our knowledge, Wi-Fitness is the first cross-layer framework to achieve the personal fitness assessment with both Wi-Fi and video bimodal for training and only Wi-Fi for testing. The key contributions of this work can be outlined as follows.

1) A random quantization-based data augmentation method is proposed to generate diverse and high-quality training data for both Wi-Fi and video.
2) A local attention mechanism is proposed to capture local features, achieving a comprehensive understanding of movement states. The combination of cross-modal and local attention allows for a more thorough comprehension of both global and local information.
3) For the two modalities, a bimodal feature integration mechanism is proposed. The features from different scales are integrated. This enables the fused features to capture both high-level semantic information as well as local detailed information.
4) An attention-based spatio-temporal graph convolutional network (AST-GCN Net) is proposed. This network extracts spatial information between key points by establishing correlations between key points. The network enhances its ability to perceive important key points.
5) A spatio-temporal semantic alignment module that minimizes the temporal alignment loss between the two modalities is proposed. The spatial information is transferred effectively from the video to the Wi-Fi while maintaining temporal consistency between the modalities.
6) Extensive experiments have validated Wi-Fitness's exceptional performance, effectiveness, and robustness.

The following sections are arranged as follows. Section II surveys related work. In Section III, the preliminary is demonstrated. In Section IV, the Wi-Fi signal processing mechanism is described. Then, a detailed framework layer by layer is provided. In Section V, experimental evaluations are presented. Section VI discusses the practicability and limitations of Wi-Fitness. Finally, in Section VII, conclusions are drawn.

II. Related Work

We discuss the recent progress in fitness assistant systems and their supporting technology through the types in this section.

A. Vision-Based

The key task in vision-based fitness assistants is learning the spatial and temporal information contained in consecutive video frames. Spatial and temporal information can be extracted in LSTM-based approach [14]. Spatial features are initially extracted using a 2-D CNN, and subsequently, LSTM is used for long-range time-dependent modeling based on high-dimensional abstracted features. However, a lot of low-dimensional information is lost in this way. Spatial and temporal features are extracted in 3-D convolutional structures to obtain the motion details encoded in adjacent frames [15], [16], [17], [18], [19]. Originally coming from computer vision, the skeleton sequence is structured as a time series of coordinates of human body joints in 3-D space constituting a time series, encoding the motion information of human body joints. It contains a high volume of information and is not easily affected by noise. Skeleton sequences are modeled as spatio-temporal graph structures and graph convolution is used in skeleton-based human behavior recognition studies [20], [21], [22], [23], [24], achieving superior performance. However, in some scenarios, the camera can violate user privacy. Therefore, fitness assistants based on video data are increasingly unable to meet the user's needs for recognition speed and privacy protection in real-world application scenarios.

B. Inertial Sensor-Based

Exercise activities are monitored by attaching sensors to the human body or fitness equipment in inertial sensor-based

systems [7], [25]. These systems exhibit greater robustness to environmental changes than vision-based systems, but the sensors are troublesome and inconvenient to carry during the exercise.

C. Wireless-Based

Wireless-based fitness assistants use various signals, such as ultrasonic signals [26], radio frequency identification (RFID) [27], and Wi-Fi for workout sensing. As one of the most common indoor wireless signals [12], Wi-Fi has several advantages, such as nonvisual, contactless, low cost, easy to deploy, not influenced by lighting conditions, and privacy protection. Channel state information (CSI) is mainly used for human behavior recognition tasks in existing Wi-Fi-based approaches [28], [29]. The behavior recognition task can be accomplished based on the connection between CSI amplitude changes and human movement state changes [29], [30]. Doppler shift features are extracted from CSI signals [31], and the relationship between Doppler shift and motion direction is modeled. The accuracy of human motion recognition is effectively improved. The correlation between CSI signal fluctuations, motion speed, and specific actions is quantified [28] to achieve finer grained human activity recognition. BVP is extracted in the human coordinate system [32], and an environment-independent gesture recognition system is implemented. Due to the bandwidth constraint and low-spatial resolution of Wi-Fi, Wi-Fi-based fitness assistants struggle with environmental dependency and generalization issues.

III. Preliminary

A. Channel State Information

Wi-Fi signal is essentially an electromagnetic wave that selectively fades when transmitted over a wireless channel, causing a change in signal strength. The multipath effect is a propagation phenomenon in which a radio signal travels from the transmitter through multiple paths to reach the receiver. CSI data can be obtained in packets received using wireless cards compatible with the IEEE 802.11a/g/n protocol [33], [34], [35], presented as follows:

$$H(f_i) = \|H(f_i)\| e^{j\angle H(f_i)} \tag{1}$$

where $H(f_i)$ denotes the CSI value for the ith subcarrier, and $\angle H(f_i)$ and $\|H(f_i)\|$ represent its phase and amplitude, respectively. According to the different effects of different propagation paths on the channel frequency response, the classification of signal propagation paths includes both dynamic paths (human reflection paths) and static paths (including LOS paths and static object reflection paths), and the corresponding channel frequency response can also be divided into dynamic channel frequency response and static channel frequency response. In the multipath environment, the channel frequency response is given by adding the dynamic channel frequency response to the static

$$\begin{aligned} H(f_i, t_i) &= e^{-j\theta_{\text{offset}}}(H_S(f_i, t_i) + H_D(f_i, t_i)) \\ &= e^{-j\theta_{\text{offset}}}\left(H_S(f_i, t_i) + A(f_i, t_i)e^{-j2\pi\frac{d(t_i)}{\lambda}}\right) \end{aligned} \tag{2}$$

where $H_D(f_i, t_i)$ is the dynamic component and $H_S(f_i, t_i)$ is the static component. The phase shift $e^{-j\theta_{\text{offset}}}$ is caused by the difference in carrier frequencies between the transmitter and receiver. The complex attenuation, phase shift, and path length of the dynamic component are represented by $A(f_i, t_i)$, $e^{-j2\pi([d(t_i)]/\lambda)}$ and $d(t_i)$, respectively. Since the position of the transmitter, receiver, and object is always fixed, the signal propagation path remains constant on the static path, and the static channel frequency response $H_S(f_i, t_i)$ is effectively constant. The dynamic channel frequency response is caused by the signal changes due to human movements.

B. CSI Ratio Model

In practical fine-grained human motion sensing, commercial WiFi devices suffer from asynchronous transmitter and receiver timing. Consequently, each CSI sample contains a random phase shift $e^{-j\theta_{\text{offset}}}$. The CSI-ratio model [36] describes the association between the target's motion. Most of the noise and time-varying phase offset in the raw CSI amplitude can be eliminated by performing.

For commercial wireless network cards, the time-varying phase offset on different antennas of the wireless network card is the same because they share the same radio frequency oscillator. For a small movement, the difference in the change of the reflection path lengths between the two nearby antennas, denoted as Δd_i, is represented by $d_2(t_i) - d_1(t_i)$. The CSI ratio is expressed as the following equation:

$$\begin{aligned} \frac{H_1(f_i, t_i)}{H_2(f_i, t_i)} &= \frac{e^{-j\theta_{\text{offset}}}\left(A_1 e^{-j2\pi\frac{d_1(t_i)}{\lambda}} + H_{S,1}\right)}{e^{-j\theta_{\text{offset}}}\left(A_2 e^{-j2\pi\frac{d_2(t_i)}{\lambda}} + H_{S,2}\right)} \\ &= \frac{A_1 e^{-j2\pi\frac{d_1(t_i)}{\lambda}} + H_{S,1}}{A_2 e^{-j2\pi\frac{\Delta d_i}{\lambda}} e^{-j2\pi\frac{d_1(t_i)}{\lambda}} + H_{S,2}} \end{aligned} \tag{3}$$

where $H_1(f_i, t_i)$, $A1$, and $H_{S,1}$ represent the CSI, complex attenuation, and static component of the first antenna, respectively. $H_2(f_i, t_i)$, $A2$, and $H_{S,2}$ represent the CSI, complex attenuation, and static component of the second antenna, respectively. Let $\alpha = A_1$, $\beta = H_{S,1}$, $\gamma = A_2 e^{-j2\pi(\Delta d_i/\lambda)}$, $\eta = H_{S,2}$, and $\mu = e^{-j2\pi([d_1(t_i)]/\lambda)}$. μ signifies a unit circle rotating clockwise with an increase in $d_1(t_i)$. The CSI ratio can be simplified as

$$\frac{H_1(f_i, t_i)}{H_2(f_i, t_i)} = \frac{\alpha\mu + \beta}{\gamma\mu + \eta} \tag{4}$$

the expression is in the form of a Mobius transformation [37], given that $\beta\gamma - \alpha\eta \neq 0$. The CSI Ratio model ingeniously combines complementary amplitude and phase in a more fine-grained manner for human motion sensing, further enhancing sensing accuracy and precision.

IV. Method

A. Overview

The framework consists of four layers: data collection layer, data preprocessing layer, bimodal fusion layer, and the fitness assessment layer, as depicted in Fig. 1. Through

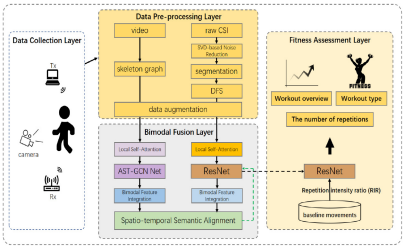

Fig. 1. Wi-Fitness framework.

layer-by-layer collaboration, the generalization of Wi-Fitness is improved. At the data collection layer, Wi-Fi and video data are acquired simultaneously. At the data preprocessing layer, singular value decomposition (SVD)-based denoising is conducted. The double sliding windows are proposed to segment the exercise. The Wi-Fi and Video data are augmented by a random quantization-based data augmentation method. At the bimodal fusion layer, Wi-Fi and video are effectively integrated. At the fitness assessment layer, users are provided with visualized exercise performance evaluations.

B. Data Preprocessing Layer

1) Spectrogram Generation From CSI Signals (Initial Filtering): Since the sampling frequencies of the transmitter and receiver are not perfectly synchronized during transmission, there are random phase errors in the original CSI, such as sampling frequency offset (SFO), carrier frequency offset (CFO), etc. In FarSense [36], CSI Ratio model is utilized to eliminate random phase offset errors. In addition, the CSI data measured by commercial Wi-Fi devices contain low-frequency interference and impulse noise, etc. [38], [39]. To filter out both high and low-frequency noise, a Butterworth bandpass filter is employed, with the cutoff frequencies configured to 10 and 80 Hz, respectively. Values of the CSI amplitude are mapped to a range between -1 and 1.

CSI Denoising Based on SVD [40]: The denoising technique based on SVD belongs to a class of subspace algorithms. We aim to decompose the vector space of the signals into two subspaces, one dominated by the activity-induced signals and the other by the noise signals. Then, by removing the components of the noisy signal vectors lying in the "noise space," the movements-induced signals can be deduced. The values in the diagonal matrix of singular values represent the significant components of the signals, which can be used to determine the main structure and noise components of the signals. Assuming the CSI matrix Y contains noisy signals, which can be expressed as $Y = X + D$. X is the matrix containing the movement-induced signal data, and D is the matrix containing the noise data. The objective is to recover the signals contained in X from the given noisy signal matrix Y. Applying SVD to the reduced-dimensional CSI signal results in decomposing the signal matrix into three parts: 1) a matrix of left singular vectors; 2) a matrix of right singular vectors; and 3) a diagonal matrix containing singular values. As shown in the following:

$$X = U_x \Sigma_x V_x = [U_{x1} U_{x2}] \begin{bmatrix} \Sigma_{x1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_{x1} \\ V_{x2} \end{bmatrix} \tag{5}$$

where U_{x1} and U_{x2} are matrices with dimensions $N \times r$ and $N \times (N - r)$, respectively. Σ_{x1} is an $r \times r$ matrix, V_{x1} is an $r \times m$ matrix, V_{x2} is an $(N - r) \times m$ matrix.

The space spanned by U_{x1} corresponds to the column space of X, referred to as the signal subspace. Using the properties of the matrices V_{x1} and V_{x2}, as well as $V_{x1}V_{x1}^H + V_{x2}V_{x2}^H = I$, which is the unitary matrix, we can rewrite the noisy signal matrix Y as follows:

$$\begin{aligned} Y &= X + D \\ &= X + D\left(V_{x1}V_{x1}^H + V_{x2}V_{x2}^H\right) \\ &= (XV_{x1} + DV_{x1})V_{x1}^H + (DV_{x2})V_{x2}^H \\ &= \left(P_1 S_1 Q_1^H\right)V_{x1}^H + \left(P_2 S_2 Q_2^H\right)V_{x2}^H \\ &= (P_1 P_2)\begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}\begin{pmatrix} Q_1^H V_{x1}^H \\ Q_2^H V_{x2}^H \end{pmatrix} \end{aligned} \tag{6}$$

where $P_1 S_1 Q_1^H$ and $P_2 S_2 Q_2^H$ represent the SVD of the matrices within the parentheses in (6), namely, $XV_{x1} + DV_{x1} = P_1 S_1 Q_1^H$ and $DV_{x2} = P_2 S_2 Q_2^H$. If $P_1^H P_2 = 0$, the column spaces of matrices P_1 and P_2, are orthogonal, then the above equation represents an effective SVD. As seen from (6), due to $P_1 \neq U_{x1}$, we cannot directly recover the signal subspace of X. Hence, we employ a low-rank model-based method, which is a least squares method to estimate the signal matrix X. Specifically, the method seeks the best rank $r (r < \text{rank}(X))$ matrix in the least squares sense, and minimizes the following squared error:

$$\min_{\widehat{X}} \|\widehat{X} - X\|_F^2 \tag{7}$$

where $\| \bullet \|_F^2$ represents the Frobenius norm. The solution for $\widehat{X}$ can be expressed as follows:

$$\widehat{X}_{LS} = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^H \tag{8}$$

where u_k and v_k represent the left and right singular vectors of the noisy matrix Y, respectively. σ_k represents the r largest singular values of Y (i.e., $\sigma_1 > \sigma_2 > \ldots > \sigma_r$). Here, we assume the effective rank of matrix X is r and $r < \text{rank}(X)$.

Performing an inverse transformation on the processed signal matrix is to restore it to the original CSI signal space. The original signal is restored with some noise removed.

Exercise Partition: In order to effectively interpret an exercise, a dynamic dual sliding window algorithm is introduced to enable an exercise partition. The purpose is to determine the beginning and end points of effective movements within a continuous exercise. In this section, local and global dual sliding windows are used to locate the duration. The variance of the amplitude is chosen as the indicator to monitor state changes.

There is a global sliding window X_1 and a local sliding window X_2. The sampling frequency is f packets/s, and the size of X_1 is set to 5 times f. There is a 2-s overlap between two adjacent global windows. Meanwhile, the local window size is set to $(1/2)f$, with a sliding frequency of $(1/4)f$.

At the i_{th} slide of X_1, the variance $\sigma_i^2(X_2^n)$(n represents the nth slide of X_2 in X_1) of time series in each X_2 is calculated, and get its average variance $\mu_i(\sigma)_i^2(X_2^n)$ and its

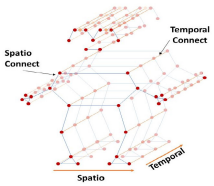

Fig. 2. Spatio-temporal graph of the skeleton sequence.

standard deviation of variance $\sigma_i(\sigma)_i^2(X_2^n)$. The segmentation threshold ϕ_i is set to be $\alpha * \mu_i((\sigma)_i^2(X_2^n)) + \beta * \sigma_i((\sigma)_i^2(X_2^n))$, here the weight values are set to be $\alpha = 2/3$ and $\beta = 1/3$ according to the experience. When $\sigma_i^2(X_2^n) > \phi_i$, the lower bound of the current interval in the local sliding window C_1 is recorded. On the contrary, $\sigma_i^2(X_2^n) < \phi_i$ the upper bound of the local sliding window C_2 is recorded.

Doppler Spectrograms: Doppler frequency shift (DFS) is a common feature that can characterize human activity and behavior because the movement of a target leads to changes in the reflected path length of signals, as well as in frequency [41].

Usually, the Doppler shift of the reflected signal caused by human movements can be expressed in (9) as

$$f_D = -\frac{1}{\lambda}\frac{d}{dt}d(t). \tag{9}$$

The short-time Fourier transform (STFT) enables time-frequency analysis of CSI by converting the waveform into a spectrogram.

2) Construction of Spatio-Temporal Skeleton Graph: MediaPose [42] human pose estimation algorithm is used for human joint point extraction, obtaining 3-D coordinates (x, y, and z), and using z-score to normalize the coordinate values of each joint point between 0 and 1. On this basis, the skeleton sequence is constructed as the undirected spatio-temporal graph $G = (V, E)$ shown in Fig. 2, with the node set $V = \{v_{ti} | t = 1, \ldots, T, i = 1, \ldots, N\}$ and the edge set E made up of E_S and E_F : $E_S = \{v_{ti}v_{tj} | (i, j) \in H\}$ is the first class of edges, which represents the connection relationship between each joint point in the same time frame, H is the natural connection of the human skeletal structure; $E_F = \{v_{ti}v_{(t+1)i}\}$ is the second class of edges, which represents the connection relationship between the same joint point in adjacent frames, called interframe edges. For a joint node i, the edges in E_F represent its motion trajectory over time. This spatio-temporal skeleton diagram contains the physical structure between human body joints in the spatial dimension and the motion information in consecutive time frames in the temporal dimension.

3) Data Augmentation: Insufficient training data leads to poor model performance. Thus, a data augmentation method based on quantized representation learning is employed. This method uses calibrated data to augment the Doppler spectrograms and the corresponding skeleton diagrams, thereby the quality of the generated data is improved.

A collection of nonoverlapping intervals $R = \{R_i = [x_i, x_{i+1})\}, i = 0, 1, \ldots, n-1$, are used to construct a quantized. n represents the total number of these intervals. All values in $[x_i, x_{i+1})$ are mapped to a scalar y_i by the quantized for the input signal s. For $s \in R_i$, the quantized is defined as $Q(s) = y_i$. As shown by the following equation:

$$Q(s) = \sum_i y_i \cdot R_i(s) \tag{10}$$

where $R_i(s)$ is the indicator function, which is 1 if $s \in R_i$ and 0 otherwise. The quantized representation uses a finite number of bits for the original signal, which introduces errors in signal recovery. Quantization includes uniform and nonuniform types. Uniform quantization involves equally spaced intervals and values, whereas nonuniform quantization features intervals or values that are not evenly spaced. The essential issue of data augmentation is to find a better tradeoff between data transmission capacity and replication errors.

The quantization is utilized as a tool for data preservation. Information is preserved within each quantization bin and across bins. By randomizing intervals, complex quantization augmentation is generated. Specifically, given $R_i = [x_i, x_{i+1})$, x_i is generated by

$$x_0, x_1, \ldots, x_{n-1} = \text{sort}\left(x_0', x_1', \ldots, x_{n-1}'\right) \tag{11}$$

$$x_i' = U(\min(s), \max(s)), i = 0, 1, \ldots, n-1 \tag{12}$$

where U represents random sampling uniformly distributed within the interval and $\min(s)/\max(s)$ denotes the minimum/maximum value of each channel s. The replication value y_i is randomly sampled within the corresponding interval

$$y_i = U(x_i, x_{i+1}). \tag{13}$$

The resulting random quantized is nonuniform. The number of quantization bins n is an augmentation hyperparameter.

Applying the aforementioned random quantization augmentation method to both modalities aims to enhance the model's generalization. By performing random quantization on Doppler spectrograms of Wi-Fi signals and skeleton maps of video data, diverse data samples are generated. The model's adaptability to noise and variations is improved and the risk of overfitting is effectively reduced.

Specifically, the random quantization augmentation method creates multiple data variants by performing uniformly distributed random sampling within the signal intervals and adjusting the replication values within each quantization bin. These variants preserve the main information of the original data but introduce randomness. Thus, the diversity and coverage of the data expanded. For Wi-Fi data, adjusting the quantization bins and replication values of Doppler spectrograms can simulate different wireless signal strengths and interference conditions. For video data, modifying the quantization parameters of skeleton maps can simulate various shooting angles and lighting conditions.

The t-SNE [43] method is used to analyze the distribution of synthesized data. Fig. 3 shows the data visualization results after the dimensionality reduction of processed Wi-Fi Doppler and video skeleton data. The experiment involves six basic activities, with points in different colors representing different activities. For each activity, the real and synthetic data are well integrated, ensuring that the virtual data distribution is

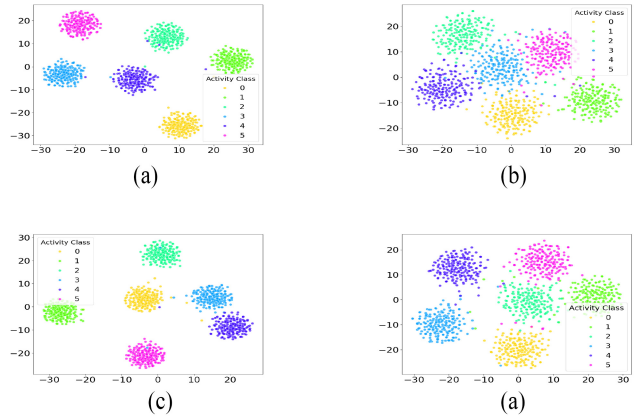

Fig. 3. (a) t-SNE visualization of doppler spectrograms (original data). (b) t-SNE visualization of doppler spectrograms (augmented data). (c) t-SNE visualization of video frames (original data). (d) t-SNE visualization of video frames (augmented data).

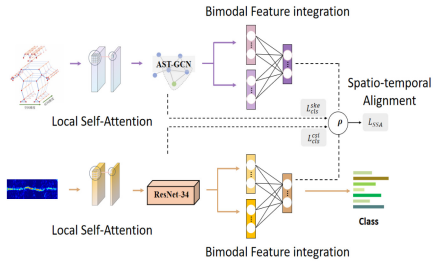

Fig. 4. Bimodal collaborative training framework.

consistent with the real data distribution, thereby enhancing data diversity.

C. Bimodal Fusion Layer

In the framework, as depicted in Fig. 4, the bimodal data are for the training phase, and Wi-Fi sensing data for the testing.

Input: The spectrograms generated from CSI signals and spatiotemporal skeleton maps constructed from the video skeleton are the input for the model training.

Training: The model includes two parts: 1) local information extraction and 2) global information fusion. Local information extraction includes a local self-attention (LSA) feature extraction module for both Wi-Fi and video modalities. The global information fusion part comprises four modules: 1) Wi-Fi feature extraction network; 2) skeleton feature extraction network; 3) bimodal feature integration module; and 4) spatiotemporal semantic alignment module. ResNet is selected as the feature extraction network for Wi-Fi modality and AST-GCN Net is the feature extraction network for video modality. The features are extracted and multilevel feature integration is performed. The high-dimensional shared semantic information in the skeleton sequence is derived through spatiotemporal semantic alignment to accomplish the knowledge migration.

Testing: The testing is conducted with only Wi-Fi sensing data. In the following, the details of the system is provided.

1) Local Self-Attention Feature Extraction Networks: To accurately capture fine-grained pose information and facilitate

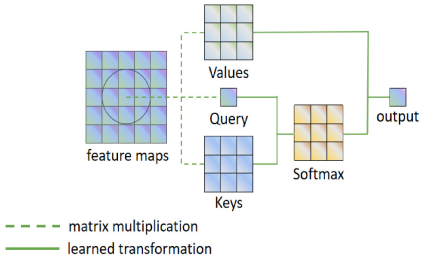

Fig. 5. LSA layer.

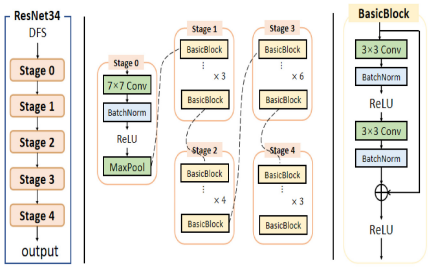

Fig. 6. ResNet34 as Wi-Fi modal feature extraction network. Input: Spectrogram.

multimodal global information sharing and fusion, preliminary feature extraction is performed on the skeletal time series and Doppler spectrogram. Since the accuracy of key point localization primarily depends on the extraction of local information, a LSA module with relative positional embedding is introduced for more precise feature extraction. For a pixel $x_{i,j}$ located at the i-th row and j-th column of the sequence graph, a 3×3 window $N_3(i, j)$ centered at $x_{i,j}$ is selected. The pixels $x_{p,q}$ within this window carry row offsets p-i and column offsets q-j, respectively, which are associated with embedding $r_{p\text{-}i}$ and $r_{q\text{-}j}$. Therefore, the spatial relative attention can be represented as

$$y_{i,j} = \sum_{p,q \in N_3(i,j)} \text{softmax}_{p,q}\left(q_{i,j}^{\top} k_{p,q} + q_{i,j}^{\top} r_{p-i,q-j}\right) v_{p,q} \quad (14)$$

where the queries $q_{i,j} = W_Q x_{i,j}$, keys $k_{p,q} = W_K x_{p,q}$, values $v_{a,b} = W_v x_{a,b}$ represent linear transformation applied to the point(i, j) and its neighbors, W_Q, W_K, and W_v are the learning transformation matrices, and $\text{Softmax}_{p,q}$ denotes the Softmax operation performed at all positions in the neighborhood $N_3(i, j)$, as depicted in Fig. 5. In the system, the pixel features of the sequences are divided into four different groups, each processed separately through a single-head attention mechanism. The results from each head are concatenated, and a fully connected layer then reconstructs this concatenated output to produce the final result.

2) Wi-Fi Modal Feature Extraction Network: 34-layer residual network (ResNet-34) [44] shown in Fig. 6 is utilized as the CSI feature extractor. The extractor consists of five stages, where Stage 0 uses a 7×7 convolution for image transformation, which can be regarded as the initial feature extraction after local information extraction from the input image, and Stage 1 through Stage 4 are composed of 3, 4, 6, and 3 BasicBlocks, respectively.

3) Skeleton Modal Feature Extraction Network: To further extract the motion-related temporal and spatial information

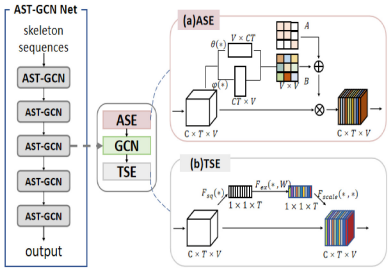

Fig. 7. Skeleton feature extraction network. (a) ASE module: Adjacency matrix attention module, (b) TSE module: Temporal attention module.

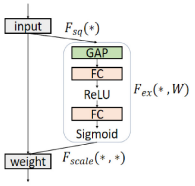

Fig. 8. TSE module.

in the skeleton graph, the attentional spatiotemporal graph convolutional network (AST-GCN Net) based on ST-GCN is proposed and implemented. The network structure is shown in Fig. 7. In the process of deep feature extraction, taking the spatio-temporal skeleton sequence after feature preextraction as input, the high-level semantic features are extracted by five attention spatio-temporal graph convolutional blocks with AGCN block, the core modules of AGCN block are the adjacency matrix attention (ASE) module and the temporal attention (TSE) module. The adaptive adjacency matrix module parameterizes the adjacency matrix and captures the connection relationships between the nodes to make the topology of the network adaptive. The temporal attention module captures the importance of different time frames during the change of motion states. Temporal attention combined with an adaptive adjacency matrix allows the network to gather information across both temporal and spatial dimensions and learn the mapping relationships between different nodes and consecutive time frames.

a) Adjacency matrix attention module: Inspired by the attention mechanism [45], the adaptive adjacency matrix attention (ASE) module is proposed. It uses the embedding function to map the input features into an adaptive matrix and combines it with the original adjacency matrix. In this way, an adaptive adjacency matrix is generated to capture the potential connection relationship between the global joint points, so that the topology of the network is adaptive and flexible. It is optimized continuously during the network training. As depicted in Fig. 7(a), the spatiotemporal skeleton diagram after initial feature extraction G is first taken as the input feature $f_{\text{in}} in R^{C\times T\times N}$. Then the inputs are mapped to different feature spaces by two embedding functions $\theta(\star)$ and $\varphi(\star)$, which are converted to $R^{N\times CT}$ and $R^{CT\times N}$ feature matrix. Then they multiply each other. The matrix elements are normalized to the values between $0 \sim 1$ by the Softmax operation. The correlation matrix B is obtained. b_{ij} represents the correlation between vertices v_i and vertices v_j, as follows:

$$B = \text{Softmax}\left(f_{\text{in}}^T W_\theta^T W_\varphi f_{\text{in}}\right) \tag{15}$$

where the embedding functions $\theta(\star)$ and $\varphi(\star)$ are implemented using two 1×1 convolutional layers, respectively, and W_θ and W_φ are arguments to the functions $\theta(\star)$ and $\varphi(\star)$, respectively.

To improve the model's flexibility without degrading its performance, a new adaptive adjacency matrix is formed by combining the correlation matrix B with the original adjacency matrix A. The resulting matrix is multiplied with the input feature f_{in} and weight W. The final output is as follows:

$$f_{\text{out}} = W f_{\text{in}}(A + B) \tag{16}$$

where W represents the parameters of 1×1 convolutional layer. It can be implied from the above that the ASE module aggregates local information as well as context information and explores the spatial information between joint points. This not only indicates whether there is a connection between each joint point but also indicates the strength of the connection. This constructs the correlation and dependence between the joint points. It plays the same role as the attention mechanism and enhances the perception ability of the key joint points in the network.

b) Temporal attention module: The contribution of the human motion state to activity recognition in different time frames is different. The channel attention is improved by adopting a time attention TSE module. This module applies channel direction attention to time direction. Inspired by the idea of compression and excitation in SE block [45], the global features are obtained through the compression operation. Then the weight of each timeframe in the feature map is obtained through the excitation operation. The different timeframes are weighted to assess the significance of features across different timeframes. Therefore, attention has been concentrated on the long-distance time dependence and the change of human movement state between successive time frames. As shown in Fig. 8, there are three steps.

The first step is the compression operation $F_{sq}(\star)$, which compresses the time dimension of the input feature with size $C\times T\times N$ to 1 through global average pooling. The feature vector $U \in R^{1\times 1\times T}$ is obtained as following:

$$U = F_{sq}(f_{\text{in}}) = \frac{1}{C \times N} \sum_{i=1}^{C} \sum_{j=1}^{N} f(i, j). \tag{17}$$

The second step is to execute the operation $F_{ex}(\star, W)$, using the fully connected layer. And the feature map with dimensions $1\times 1\times C$, and obtain the corresponding weight of each time frame. The weight represents the importance of each time frame. By fully deriving the dependencies between timeframes, the nonlinear relationship between timeframes can be learned, and all the time frame characters can be ensured. Specifically, two fully connected layers are constructed, with the first layer used to perform dimensionality reduction. Through the ReLU function and the second fully connected layer, the dimension is restored to the original time dimension, and the weight vector S of each timeframe is obtained

$$S = F_{ex}(U, W) = \sigma(W_2 \delta(W_1) U) \tag{18}$$

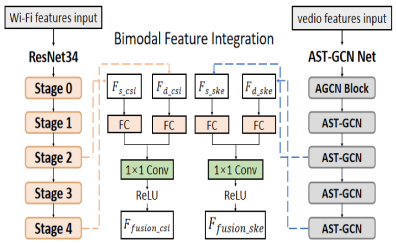

Fig. 9. Bimodal feature integration module.

where $\sigma(*)$ is the Sigmoid activation function, $\delta(*)$ indicates the ReLU function, and W_1 and W_2 represent the weights for the two fully connected layers, respectively.

Finally, in the weighting operation $F_{\text{scale}}(*,*)$, as can be seen from (19), the weight takes the value range of $0 \sim 1$, and the weight is used to scale the input feature as the contribution of each timeframe to generate the weighted output feature, as shown in

$$F_{\text{scale}}(U, S) = \sum_{t=1}^{T} u_t s_t \tag{19}$$

$$f_{\text{out}} = F_{\text{scale}}(U, S) f_{\text{in}}. \tag{20}$$

4) Bimodal Feature Integration Module: A bimodal feature integration module is proposed to enhance the model's capacity for capturing both low-level and high-level semantic information, as shown in Fig. 9. The Wi-Fi features from Stages 2 and 4 of ResNet are extracted as shallow features F_{s_csi} and deep features F_{d_csi}, respectively. The video features from the second AGCN block and the last AGCN block of the AST-GCN network are derived as shallow features F_{s_ske} and deep features F_{d_ske}, respectively. They are mapped to the same dimension using four independent fully connected layers. Following the fully connected layers are bimodal feature integration with the 1×1 convolutional layer and nonlinear activation with the ReLU activation function. The fused features can be expressed as follows:

$$F_{\text{fusion\_csi}} = \delta\left(W_{\text{conv1}} \text{cancat}(W_{fc1} F_{s\_csi}, W_{fc2} F_{d\_csi})\right) \tag{21}$$

$$F_{\text{fusion\_ske}} = \delta\left(W_{\text{conv2}} \text{cancat}(W_{fc3} F_{s\_\text{ske}}, W_{fc4} F_{d\_ske})\right) \tag{22}$$

the weight parameters of the four fully connected layers are denoted by $W_{fc1}, W_{fc2}, W_{fc3}$, and W_{fc4}. W_{Conv1} and W_{Conv2} denote the weight parameters of the two 1×1 convolutional layers. Low-level features and high-level complementary semantic features are aggregated by the multiscale feature integration module.

5) Spatio-Temporal Semantic Alignment Module: $F_{\text{fusion\_csi}}$ and $F_{\text{fusion\_ske}}$ are multilevel semantic features of Wi-Fi modality and video modality, respectively. It is assumed that the features in $F_{\text{fusion\_csi}}$ have corresponding features in $F_{\text{fusion\_ske}}$ with the same semantic meaning. Therefore, the correlation between all features can be expressed in the following correlation matrix:

$$\text{corr}(F_{csi}) = \hat{F}_{\text{fusion\_csi}} \hat{F}^T_{\text{fusion\_csi}} \in d \times d \tag{23}$$

$$\text{corr}(F_{ske}) = \hat{F}_{\text{fusion\_ske}} \hat{F}^T_{\text{fusion\_ske}} \in d \times d \tag{24}$$

where $\hat{F}_{\text{fusion\_csi}}, \hat{F}_{\text{fusion\_ske}}$ denote the normalized matrices and d denotes the matrix dimension.

Although the features of different modalities are different, their high-dimensional features have similar semantic information. Therefore, the similarity between the high-dimensional features of two modalities can be presented by the difference between the feature correlation matrices. For example, the larger the difference between the feature correlation matrices, the lower the similarity between the two high-dimensional features. On the contrary, the smaller the difference between the correlation matrices, the higher the semantic similarity between the features. Therefore, to learn the cross-modal shared knowledge, the network needs to maximize learning as the similarity increases and minimize the difference between the correlation matrices. The temporal alignment loss can be used to portray the difference between different modalities and minimize the temporal alignment loss during the training process to learn the shared knowledge [46]. L_{SSA} is used to represent the temporal alignment loss. Specifically, the process is to calculate the difference of the feature correlation matrices and take the square of the Frobenius norm of the result, multiplied by ρ, as shown in

$$L_{\text{SSA}} = \rho \| \text{corr}(F_{csi}) - \text{corr}(F_{ske}) \|_F^2 \tag{25}$$

where ρ is a regularization parameter to prevent negative migration.

Too much variation between modalities can sometimes have a negative effect on one of the modalities. This leads to performance degradation and negative migration on network performance. Classification loss can be used to measure the accuracy of the content learned by the network. When the classification loss is smaller, the network learns more accurately. We use the difference $\Delta L = L_{\text{cls}}^{\text{csi}} - L_{\text{cls}}^{ske}$ between the classification loss $L_{\text{cls}}^{\text{csi}}, L_{\text{cls}}^{ske}$ of two modalities to measure the network performance difference. When the difference ΔL is positive, it means that the current skeleton feature extraction network learns more information than the Wi-Fi feature extraction network. It can be added to the shared network. Therefore, the regularization parameter ρ is adjusted to $e^{\Delta L - 1}$. When the difference is negative, it means that the current skeleton feature extraction network learns less than the current Wi-Fi feature extraction network. If the knowledge migration is performed, there will be negative migration on the Wi-Fi feature extraction network performance. Therefore, the regularization parameter ρ is adjusted to 0. The regularization parameter ρ is adjusted based on the classification loss difference to control the contribution of shared knowledge from two modalities, as shown in

$$\rho = S\left(e^{\Delta L - 1}\right) = \begin{cases} e^{\Delta L - 1}, & \Delta L > 0 \\ 0, & \Delta L \le 0 \end{cases} \tag{26}$$

where $S(\cdot)$ is the zero threshold function.

6) Joint Loss Function: The function is composed of two components, classification loss $L_{\text{cls}}^{\text{csi}}$ and spatio-temporal alignment loss L_{SSA}. The classification loss is used to measure the accuracy of the network for the classification so that the learning result of the network gradually approximates the

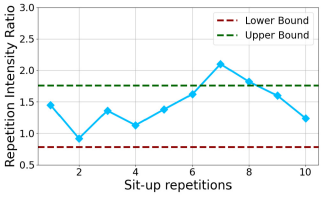

Fig. 10. RIR.

true label. The spatio-temporal alignment loss is employed to ensure the network learns multimodal complementary knowledge during the training process. It is also used to transfer knowledge between feature extraction networks of different modalities. At the training phase, the total loss function is obtained by summing the classification loss and the spatio-temporal alignment loss

$$L = L_{\text{cls}}^{\text{csi}} + \lambda \sum_{n=1}^{M} L_{\text{SSA}} \tag{27}$$

where M is a constant value of 2, representing the total number of modalities and λ is the hyperparameter.

D. Fitness Assessment Layer

This layer aims to provide users with workout assessment and analysis. Wi-Fitness analyzes and evaluates the user's exercise regularity and intensity. Specifically, based on the frequency, intensity, time, and type (FITT) principle [47], a new metric is adopted to portray the effects of the user's exercise.

The repetition intensity ratio (RIR) refers to the intensity (or speed) at which the user performs repeated exercise movements. Wi-Fitness provides each user with exercise trends based on the metric, allowing a user to adjust his exercise forms accordingly. The RIR is denoted as follows:

$$\text{RIR} = \frac{T_{\text{if}}^{k}}{T_{fi}^{k}} \tag{28}$$

where T_{if}^{k} is the duration for the kth repeated exercise movement from the initial position to the final position and T_{fi}^{k} is the duration to move from the final position back to the initial position for the kth repeated exercise movement. Wi-Fitness provides each user with exercise trends based on this metric as well as the actual duration, allowing a user to adjust his exercise accordingly.

The basic idea behind exercise evaluation is to compare and assess the exercise results for a regular user and professionals based on the metric above. Based on the evaluation results, Wi-Fitness then provides effective exercise suggestions to help a user. Specifically, Wi-Fitness not only shows the trend of the RIR metric across all repeated exercise actions but also displays the upper and lower bounds of the metric. This allows a user to see a visual representation of his exercise evaluation and get to know if his exercise actions are effective. According to the visual feedback, the user can adjust subsequent exercise activities.

Fig. 10 shows the RIR trend line chart of a user after performing ten consecutive sit-ups. It is obvious that the 7th and 8th actions exceeded the upper bound. This indicates that the user needs to increase the movement intensity from the initial to the final position. Conversely, if the intensity falls below the lower bound, the intensity of the movement from the initial to the final position needs to be reduced.

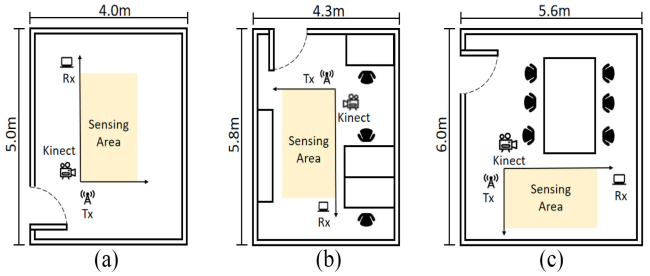

Fig. 11. Layouts. (a) Empty room. (b) Office. (c) Meeting room.

V. Experiments and Evaluation

A. Experiment Setup

1) Devices: Lenovo E73S desktop with Intel AX210 network interface cards (NICs), running PicoScense CSI Tool on Ubuntu 20.04, are used for data capture. One computer with two antennas works as the transmitter and the other computer with two antennas works as the receiver. The antennas are placed in a horizontal line. The devices are positioned 2.5-m apart at a height of 1.6 m. The system operates on a 5-GHz Wi-Fi channel (Channel 165) with a bandwidth of 20 MHz. Each receiving antenna has 57 subcarrier, a total of 114 for the data stream. The default packet transmission rate is 2000 packets/s. It transmits packets at 900 Hz, and both the transmitter and receiver are configured to a power level of 15 dBm. A Kinect V2 camera, centrally positioned between two antennas, records video at 600 × 480 pixels resolution and 30 frames/s.

To address the varying number of CSI samples due to random access protocol and packet loss, 30 CSI measurements per video frame are resampled using first-order linear interpolation to align with the timestamps. Network time protocol (NTP) is employed to synchronize all devices, achieving an average synchronization error of 5 ms.

2) Data Collection: Most Wi-Fi-based human activity recognition data sets do not include the corresponding video data. In this study, the data are collected during the experiment. Wi-Fi sensing data are collected from ten gender-balanced volunteers with diverse physical characteristics (aged 20–45 years old, weights 43–80 kg, and heights 155–185 cm). Six predefined activities are conducted, including push-ups, sit-ups, lateral raises, squats, bench presses, and leg raises. Each volunteer selects one or more activities from the six activities and conducts them. Each group of chosen activities is conducted 3 times. These activities are conducted in three distinct environments as shown in Fig. 11: an empty room [5.0 m × 4.0 m, Fig. 11(a)], an office [5.8 m × 4.3 m, Fig. 11(b)] and a conference room [6.0 m × 5.6 m, Fig. 11(c)]. The experimental setup is shown in Fig. 12.

3) Training Details: Wi-Fitness is trained and evaluated on an NVIDIA GeForce RTX 3090 GPU, utilizing Python

3.10.8 and PyTorch 1.13.1. The training process starts with an initial learning rate of 0.001, halved every ten epochs using a multistep decay strategy. After training for 80 epochs, the model with the minimum loss is selected as the best model. The computational complexity of the model is 27.7 GFLOPs, comprising 24.5 million parameters. Both training and testing are performed using the NVIDIA GeForce RTX 3090 GPU. The Adam optimizer is employed with a batch size set to 128. λ is initially set to 10 in the loss function. The data is divided into two sets: 1) 80% for training and 2) the remaining 20% for testing.

4) Evaluation Metric: The model's performance is evaluated using the following metrics.

The following parameters are introduced, TP, true positives, represents positive cases that are predicted as true (predicted as positive and actually positive). FP, false positives, represents negative cases that are predicted as true. FN, false negatives, represents positive cases that are predicted as false. TN, true negatives, represents negative cases that are predicted as false. Accuracy is used to evaluate the proportion of correctly predicted samples (including correctly predicted positives and negatives, i.e., TP and TN) to the total number of samples. Although accuracy can indicate overall correctness, it may not be a good metric in cases of imbalanced samples. High accuracy in such cases might be meaningless, rendering accuracy ineffective. Therefore, additional performance metrics are introduced.

Precision represents the ratio of TP to the sum of TP and FP. It indicates the proportion of true positives out of all predicted positives.

True Positive Rate (TPR, Recall) is defined as the ratio of TP to the total number of actual positive samples, which includes TP plus FN. It indicates the proportion of TP out of all actual positives

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{29}$$

False Positive Rate (FPR) represents the proportion of actual negative cases that are falsely identified as positive

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}. \tag{30}$$

F1-Score Precision and recall influence each other, and the F1-score takes both into account. The F1-score represents their harmonic mean, with higher values indicating better classification performance. By substituting the formulas for precision and recall, it is evident that when the F1-score is low, true positives increase relative to false positives and false negatives, thereby increasing both precision and recall. As shown in

$$F_1 - \mathrm{Score} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{31}$$

B. Models Comparison

To validate the performance of Wi-Fitness, it is compared with methods, such as HuAc [48], WiPose [49], InFit [4], and FitAssist [3]. The results indicate that Wi-Fitness consistently achieves the highest accuracy compared to the other four, as indicated in Table I. These results benefit from the proposed layer-by-layer framework.

TABLE I
METHODS COMPARISON

Method	F1-Score(%)
HuAc	85.49
WiPose	83.72
InFit	73.23
FitAssist	75.61
Wi-Fitness	92.68

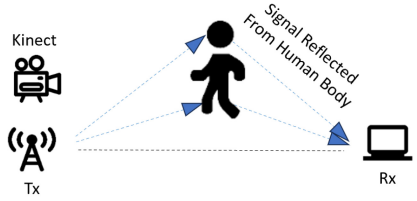

Fig. 12. Experiment setup.

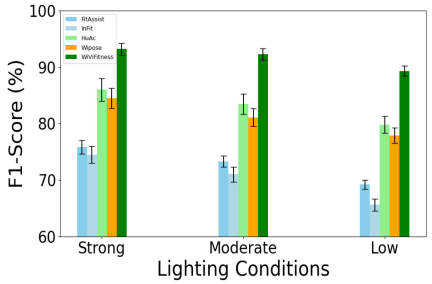

Fig. 13. Comparison in different lighting conditions.

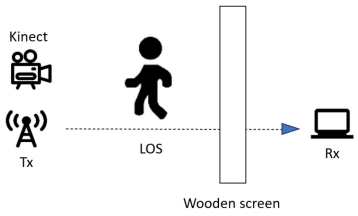

Fig. 14. Wooden screen occlusion.

This advantage stems from the enhanced data collection device (AX210) and the more efficient modality fusion techniques employed by Wi-Fitness. In the subsequent comparative experiments, we further compared Wi-Fitness with previous methods.

1) Impact of Lighting Conditions: To test the impact of lighting changes, Wi-Fitness initially trained in a well-lit environment is tested under different lighting conditions. There are three different lighting levels: 1) strong; 2) moderate; and 3) low. The results are illustrated in Fig. 13, video-enhanced systems like Wi-Fitness, WiPose, and HuAc exhibit superior robustness to lighting changes when compared to FitAssist and InFit.

2) Impact of the Occlusion: To evaluate the performance of Wi-Fitness under occlusion, a wooden screen is placed between the subject and the receiver, as shown in Fig. 14.

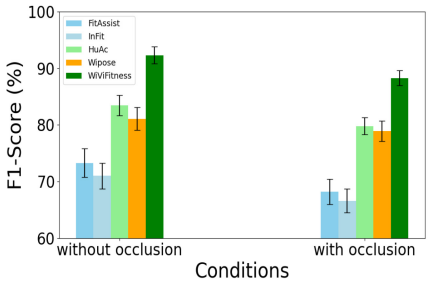

Fig. 15. Comparison under occlusion scenarios.

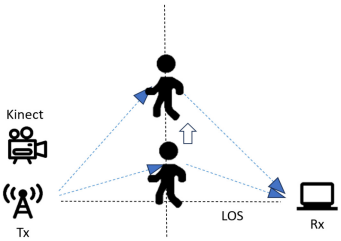

Fig. 16. Impact of subject position.

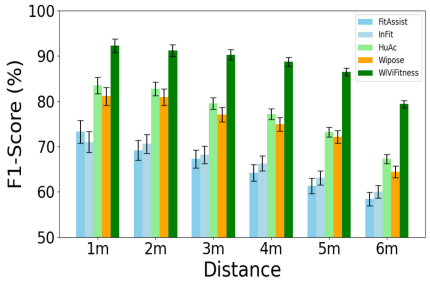

Fig. 17. Comparison at different distances.

Wi-Fitness is trained in an unobscured environment and tested in both unobscured and occluded conditions. As shown in Fig. 15, Wi-Fitness is most robust to occlusion. Despite the screen reduces the strength of the Wi-Fi signal, the majority of the fitness-related information can still be preserved. Even though obstacles cause increasing errors compared to the cases without them, Wi-Fitness can still maintain pretty good performance, highlighting its resilience.

3) Impact of Subject's Position: As shown in Fig. 16, the subject faces the Kinect and moves along the perpendicular bisecting line of the LOS path. The distance from the subject to the LOS path varies from 1 to 6 m. The model is trained in conditions where the subject is 1-m away from the LOS and tested in 1–6-m environments. The experimental result in Fig. 17 indicates that Wi-Fitness can achieve high accuracy within the 5-m × 5-m range. Introducing the video makes Wi-Fitness less sensitive to distance variations.

C. Ablation Study

Each component's impact is evaluated by the ablation study. The findings are summarized in Table II.

1) Impact of the Doppler: The Doppler feature significantly impacts the system's ability to capture fine motion details. The Doppler feature is crucial for identifying subtle movements and enhances the overall performance. If the DFS derived from the STFT is adopted, the F1-Score can reach as high as 0.9268. Without DFS, the F1-Score is 0.7282. Thus, it is necessary to keep it.

TABLE II
ABLATION STUDY ON NOVEL COMPONENTS

Method	F1-Score(%)
w/ raw CSI	72.82
w/o local self-attention module	74.10
w/o ASE and TSE	69.26
w/o bimodal feature integration module	55.48
Wi-Fitness(ours)	92.68

2) Impact of the Local Self-Attention Module: The LSA module enhances the extracted key features by employing a local attention mechanism. This module focuses on specific regions within the heatmaps and video data. It assigns appropriate weights to different parts and refines the feature extraction process. The F1-Score decreases from 0.9268 to 0.741 with and without the LSA module. This indicates the crucial role of this model. It improves the accuracy by emphasizing pertinent regions.

3) Imapact of the Adjacency Matrix Attention (ASE) Module and Temporal Attention (TSE) Module: The adjacency matrix attention (ASE) module captures potential connections between global nodes by generating an adaptive adjacency matrix, making the network topology adaptive and flexible. The ASE module aggregates not only local information but also contextual information, uncovering spatial information between connected nodes.

The temporal attention (TSE) module enhances channel attention by applying channel-wise attention in the temporal direction. This module integrates WiFi-based features with video-based features. It creates a multimodal feature learning environment. The module obtains global features through a compression operation and then derives the weights of each time frame in the feature map through an excitation operation. Different time frames are weighted to catch the importance of features over different temporal ranges.

When the corresponding module is removed, and only ST-GCN is used as the classification network, the F1-Score decreases to 0.6926. This indicates that without this module, the cohesive integration of video and Wi-Fi cannot be ensured.

4) Impact of the Bimodal Feature Integration Module: The interaction between features extracted from Wi-Fi and video is enhanced by the bimodal feature integration module. This module captures the complex details of human motion by effectively combining Doppler features with skeletal action features. The bimodal feature integration module can effectively learn cross-modal shared knowledge. When this module is removed, the F1-Score decreases by 0.372. This indicates the importance of this module in achieving global feature integration and improving accuracy.

D. Robust Evaluation

1) Impact of Different Environments: The experiments are conducted in three different environments: 1) an empty room;

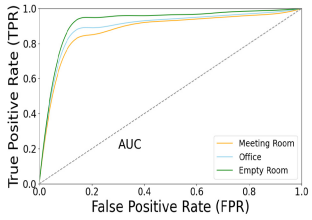

Fig. 18. Impact of different environments.

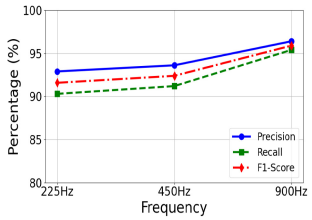

Fig. 19. Impact of different frequencies.

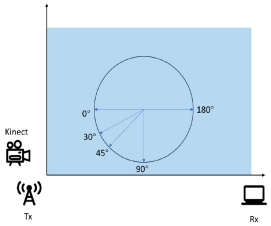

Fig. 20. User orientations.

2) an office; and 3) a conference room. Fig. 18 shows the results. Among them, the F1 scores trained with samples from the empty room and tested with samples from the office and conference room are 0.856 and 0.839, respectively. When the environment changes, the performance gets a little bit worse. But F1 scores remain above 0.8. This indicates that the framework can mitigate environmental dependencies.

2) Impact of the Transmission Rate: The system is initially configured to transmit CSI packets at a default rate of 900 Hz. The CSI packet transmission rate changes from 450 to 225 Hz in order to assess the influence of the rate. It is necessary to adjust the CNN architecture, particularly the convolutional kernel sizes, to accommodate the altered input dimensions. Fig. 19 shows that higher packet transmission rates improve system accuracy by better capturing rapid body motions. Additionally, increasing the number of packets per second can enhance the accuracy. Notably, even with reduced packet rates, Wi-Fitness maintains robust performance, highlighting its adaptability and effectiveness across various scenarios.

3) Impact of Different Subject's Orientations: We investigate the impact of user orientation. User orientation is the direction in which a user faces during exercise. As depicted in Fig. 20, the user orientation angle is the angle between the direction the user is facing and the Tx-Rx.

When a user faces the Kinect during the exercise, the user orientation angle is 0°. When a user faces LOS during the exercise, the user orientation angle is 90°. When facing the receiver, the user orientation angle is 180°. The experiment

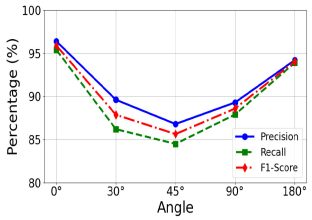

Fig. 21. Impact of different orientations.

tests the impact of user orientation. The model is trained in the 0° user orientation angle and tested in others (0°, 30°, 45°, 90°, and 180° orientations). As shown in Fig. 21, Wi-Fitness performs best when the user faces Kinect. Generally, the model performs similarly across various orientations. There is hardly a big difference, even in the user orientation angle of 180°. Benefiting from the proposed framework, the orientation impact is mitigated effectively. This verifies that the model is robust against variations in user orientation.

VI. Practicability and Limitations

Wi-Fitness can be integrated into applications, such as smart homes and art buildings. An exerciser can do the predefined activities and get feedback from Wi-Fitness which helps him to improve his fitness effectively. Although Wi-Fitness demonstrates its effectiveness in a wide range of applications, it still possesses some limitations. First, Wi-Fitness cannot achieve satisfactory performance when there are multiple exercisers within the sensing area and do the exercise together. Second, Wi-Fitness still confronts security issues. The private personal fitness information might be captured by an attacker since its plaintext transmission. Man-in-the-Middle Attack might happen as well. These issues deserve future work.

VII. Conclusion

In this article, Wi-Fitness, an advanced fitness assistant is proposed. Wi-Fitness leverages the complementary bimodal sensing of Wi-Fi and video to provide comprehensive fitness assessments and personalized workout suggestions. Through the proposed framework, the heterogeneity issue between video perception as well as wireless sensing is tackled and the generalization is improved. This is the first smart fitness assistant using both Wi-Fi and video bimodal for training and single Wi-Fi for testing.

Extensive experiments validate the effectiveness, robustness, and superior performance of Wi-Fitness. This work sets the stage for further investigation on multimodal sensing and enhances the potential of smart fitness assistants in providing safe, effective, and standardized exercise guidance.

References

[1] C. Xiang, "The development path of home fitness among Chinese citizens after COVID-19 pandemic: A perspective of network technology and information dissemination," in *Proc. Int. Conf. Inf. Technol. Cont. Sports (TCS)*, 2021, pp. 1–5.

[2] X. Guo, J. Liu, C. Shi, H. Liu, Y. Chen, and M. C. Chuah, "Device-free personalized fitness assistant using Wi-Fi," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 4, pp. 1–23, Dec. 2018. [Online]. Available: https://doi.org/10.1145/3287043

[3] Y. Zhu, D. Wang, R. Zhao, Q. Zhang, and A. Huang, "FitAssist: Virtual fitness assistant based on Wi-Fi," in *Proc. 16th EAI Int. Conf. Mobile Ubiquitous Syst., Comput., Netw. Services*, 2019, pp. 328–337.

[4] H. Li, J. Xiao, W. Wang, L. Wang, D. Zhang, and H. Jin, "InFit: Combination movement recognition for intensive fitness assistant via Wi-Fi," *IEEE Trans. Mobile Comput.*, vol. 22, no. 12, pp. 7188–7202, Dec. 2023.

[5] S. B. Eisenman, E. Miluzzo, N. D. Lane, R. A. Peterson, G.-S. Ahn, and A. T. Campbell, "BikeNet: A mobile sensing system for cyclist experience mapping," *ACM Trans. Sensor Netw. (TOSN)*, vol. 6, no. 1, pp. 1–39, 2010.

[6] K.-H. Chang, M. Y. Chen, and J. Canny, "Tracking free-weight exercises," in *Proc. Int. Conf. Ubiquitous Comput.*, 2007, pp. 19–37.

[7] X. Guo, J. Liu, and Y. Chen, "FitCoach: Virtual fitness coach empowered by wearable mobile devices," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2017, pp. 1–9.

[8] C. Shen, B.-J. Ho, and M. Srivastava, "Milift: Efficient smartwatch-based workout tracking using automatic segmentation," *IEEE Trans. Mobile Comput.*, vol. 17, no. 7, pp. 1609–1622, Jul. 2017.

[9] H. Guo, Z. Shanchen, C. Lai, and H. Zhang, "PhyCoVIS: A visual analytic tool of physical coordination for cheer and dance training," *Comput. Animat. Virtual Worlds*, vol. 32, no. 1, Nov. 2020, Art. no. e1975.

[10] J. Guo et al., "Improving human action recognition by jointly exploiting video and Wi-Fi clues," *Neurocomputing*, vol. 458, pp. 14–23, Oct. 2021.

[11] Y. Wang, L. Guo, Z. Lu, X. Wen, S. Zhou, and W. Meng, "From point to space: 3-D moving human pose estimation using commodity Wi-Fi," *IEEE Commun. Lett.*, vol. 25, no. 7, pp. 2235–2239, Jul. 2021.

[12] H. Zou, J. Yang, H. Prasanna Das, H. Liu, Y. Zhou, and C. J. Spanos, "Wi-Fi and vision multimodal learning for accurate and robust device-free human activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 426–433.

[13] L. Deng, J. Yang, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "GaitFi: Robust device-free human identification via Wi-Fi and vision multimodal learning," *IEEE Internet Things J.*, vol. 10, no. 1, pp. 625–636, Jan. 2023.

[14] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 461–470.

[15] A. Diba et al., "Temporal 3-D ConvNets: New architecture and transfer learning for video classification," 2017, *arXiv:1711.08200*.

[16] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3-D residual networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 3154–3160.

[17] S. Ji, W. Xu, M. Yang, and K. Yu, "3-D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3-D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[19] Z. Wang, Q. She, and A. Smolic, "Action-Net: Multipath excitation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13214–13223.

[20] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 183–192.

[21] K. Cheng, Y. Zhang, X. He, J. Cheng, and H. Lu, "Extremely lightweight skeleton-based action recognition with shiftgcn++," *IEEE Trans. Image Process.*, vol. 30, pp. 7333–7348, 2021.

[22] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3595–3603.

[23] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12026–12035.

[24] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.

[25] D. Morris, T. S. Saponas, A. Guillory, and I. Kelner, "RecoFit: Using a wearable sensor to find, recognize, and count repetitive exercises," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, New York, NY, USA, 2014, pp. 3225–3234. [Online]. Available: https://doi.org/10.1145/2556288.2557116

[26] Y. Xie, F. Li, Y. Wu, and Y. Wang, "HearFit: Fitness monitoring on smart speakers via active acoustic sensing," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2021, pp. 1–10.

[27] H. Ding et al., "FEMO: A platform for free-weight exercise monitoring with RFIDs," in *Proc. 13th ACM Conf. Embed. Netw. Sensor Syst.*, New York, NY, USA, 2015, pp. 141–154. [Online]. Available: https://doi.org/10.1145/2809695.2809708

[28] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial Wi-Fi devices," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1118–1131, May 2017.

[29] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-eyes: Device-free location-oriented activity identification using fine-grained Wi-Fi signatures," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, 2014, pp. 617–628.

[30] S. Tan and J. Yang, "WiFinger: Leveraging commodity Wi-Fi for fine-grained finger gesture recognition," in *Proc. 17th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2016, pp. 201–210.

[31] K. Qian, C. Wu, Z. Zhou, Y. Zheng, Z. Yang, and Y. Liu, "Inferring motion direction using commodity Wi-Fi for interactive exergames," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2017, pp. 1961–1972.

[32] Y. Zheng et al., "Zero-effort cross-domain gesture recognition with Wi-Fi," in *Proc. 17th Annu. Int. Conf. Mobile Syst., Appl., Services*, 2019, pp. 313–325.

[33] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Predictable 802.11 packet delivery from wireless channel measurements," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 159–170, 2010.

[34] E. H. Ong, J. Kneckt, O. Alanen, Z. Chang, T. Huovinen, and T. Nihtilä, "IEEE 802.11 ac: Enhancements for very high throughput WLANs," in *Proc. IEEE 22nd Int. Symp. Pers., Indoor Mobile Radio Commun.*, 2011, pp. 849–853.

[35] Y. Xiao, "IEEE 802.11n: Enhancements for higher throughput in wireless LANs," *IEEE Wireless Commun.*, vol. 12, no. 6, pp. 82–91, Dec. 2005.

[36] Y. Zeng, D. Wu, J. Xiong, E. Yi, R. Gao, and D. Zhang, "FarSense: Pushing the range limit of Wi-Fi-based respiration sensing with CSI ratio of two antennas," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–26, Sep. 2019. [Online]. Available: https://doi.org/10.1145/3351279

[37] H. Schwerdtfeger, *Geometry of Complex Numbers: Circle Geometry, Moebius Transformation, Non-Euclidean Geometry*. New York, NY, USA: Dover Publ., Inc., 1979.

[38] Y. He, Y. Chen, Y. Hu, and B. Zeng, "Wi-Fi vision: Sensing, recognition, and detection with commodity MIMO-OFDM Wi-Fi," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8296–8317, Sep. 2020.

[39] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of Wi-Fi signal based human activity recognition," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 65–76.

[40] S. K. Jha and R. Yadava, "Denoising by singular value decomposition and its application to electronic nose data processing," *IEEE Sensors J.*, vol. 11, no. 1, pp. 35–44, Jan. 2011.

[41] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using Wi-Fi signals," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 363–373.

[42] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," 1906, *arXiv:1906.08172*.

[43] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[46] M. Abavisani, H. R. V. Joze, and V. M. Patel, "Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1165–1174.

[47] A. Barisic, S. T. Leatherdale, and N. Kreiger, "Importance of frequency, intensity, time and type (FITT) in physical activity assessment for epidemiological research," *Can. J. Public Health*, vol. 102, no. 3, pp. 174–175, 2011.

[48] L. Guo, L. Wang, J. Liu, W. Zhou, and B. Lu, "HuAc: Human activity recognition using crowdsourced Wi-Fi signals and skeleton data," *Wireless Commun. Mobile Comput.*, vol. 2018, no. 1, 2018, Art. no. 6163475. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2018/6163475

[49] W. Jiang et al., "Towards 3-D human pose construction using Wi-Fi," in *Proc. 26th Annu. Int. Conf. Mobile Comput. Netw.*, New York, NY, USA, 2020, pp. 1–14. [Online]. Available: https://doi.org/10.1145/3372224.3380900

Mengli Wei received the M.S. degree in human movement science from Wuhan Sports University, Wuhan, China, in 2021, where he is currently pursuing the Ph.D. degree with Sports Big-Data Research Center.

His research interests include promoting health through exercise supported by AI.

Daguo Zhao received the B.E. degree from Tianjin University, Tianjin, China, in 2023, where he is currently pursuing the master's degree.

His research interests include wireless sensing and data mining.

Lei Zhang (Member, IEEE) received the Ph.D. degree in computer science from Auburn University, Auburn, AL, USA, in 2008.

From 2008 to 2011, she was an Assistant Professor with the Department of Computer Science, Frostburg State University, Frostburg, MD, USA. She is currently an Associate Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. Her research interests include mobile computing and data mining.

Dr. Zhang is a member of the ACM.

Cheng Wang (Senior Member, IEEE) received the M.S. degree from the Department of Applied Mathematics, Tongji University, Shanghai, China, in 2006, and the Ph.D. degree from the Department of Computer Science and Technology, Tongji University in 2011.

He is currently a Professor with the Key Laboratory of Embedded System and Service Computing, Tongji University. His research interests include cyberspace security and workplace safety.

Yonggang Zhang (Member, IEEE) received the Ph.D. degree from the College of Computer Science and Technology, Jilin University, Changchun, China, in 2005.

He was a Postdoctoral Fellow with the School of Mathematics, Jilin University from 2007 to 2009, where he is currently a Professor with the College of Computer Science and Technology, Jilin University. His current research interests include constraint programming and artificial intelligence.

Qi Wang (Member, IEEE) received the B.S. degree in automation, the M.S. degree in measurement technology and automatic devices, and the Ph.D. degree in engineering from Tianjin University, Tianjin, China, in 2007, 2009, and 2012, respectively.

Since 2012, she has been engaged in scientific research and teaching with the School of Electronics and Information Engineering, Tiangong University, where she is currently a Professor. Her major research interests include electrical measurement and intelligent information processing.

Xiaochen Fan received the B.E. degree from Beijing Institute of Technology, Beijing, China, in 2013, and the Ph.D. degree from the University of Technology Sydney, Sydney, NSW, Australia, in 2021.

He is a Postdoctoral Researcher with the Institute for Electronics and Information Technology, Tianjin, China, and the Department of Electronic Engineering, Tsinghua University, Beijing, China. His research interests include mobile and urban computing, IoT, and deep learning.

Yaping Zhong received the Ph.D. degree in computer science from Auburn University, Auburn, AL, USA, in 2003.

He was a Professor with Shandong Sports University, Jinan, China, from 2008 to 2014. He is currently a Professor with Sports Big-data Research Center, Wuhan Sports University, Wuhan, China. His research interests include intelligent sports and data mining.

Shiwen Mao (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Polytechnic University, Brooklyn, NY, USA, in 2004.

He is currently a Professor and an Earle C. Williams Eminent Scholar of Electrical and Computer Engineering with Auburn University, Auburn, AL, USA. His research interests include wireless networks, multimedia communications, and smart grid.

Prof. Mao received the IEEE ComSoc TCCSR Distinguished Technical Achievement Award in 2019, the Auburn University Creative Research and Scholarship Award in 2018, and the NSF CAREER Award in 2010. He is a co-recipient of the 2021 Best Paper Award of Elsevier/KeAi Digital Communications and Networks Journal, the 2021 IEEE Internet of Things Journal Best Paper Award, the 2021 IEEE Communications Society Outstanding Paper Award, the IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, the IEEE ComSoc MMTC 2018 Best Journal Paper Award and the 2017 Best Conference Paper Award, the Best Demo Award of IEEE INFOCOM 2022 and IEEE SECON 2017, the Best Paper Awards of IEEE ICC 2022 and 2013, IEEE GLOBECOM 2019, 2016, and 2015, IEEE WCNC 2015, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems. He is a Distinguished Lecturer of IEEE Communications Society and IEEE Council of RFID, and was a Distinguished Lecturer from 2014 to 2018 and a Distinguished Speaker from 2018 to 2021 of IEEE Vehicular Technology Society. He serves on the editorial boards for several journals, including IEEE TRANSACTIONS ON MOBILE COMPUTING.