A Sample Article Using IEEEtran.cls for IEEE Journals and Transactions

IEEE Publication Technology, *Staff, IEEE,*

***Abstract*—**

***Index Terms*—Article submission, IEEE, IEEEtran, journal, LaTeX, paper, template, typesetting.**

I. Introduction

THIS

II. Preliminaries

1) Dataset Description:

a) **HWU-USP Dataset:** The HWU-USP dataset is acquired in a smart home setting equipped with a TIAGO robot and wearable inertial measurement units (IMUs). It provides synchronized RGB video, depth maps, and IMU signals for nine daily activities, such as dishwashing, making tea, reading newspapers, and laptop usage. The dataset captures diverse user interactions in realistic conditions, making it suitable for multimodal activity recognition tasks.

b) **Berkeley-MHAD Dataset:** The Berkeley Multimodal Human Action Database (Berkeley-MHAD) contains recordings from 12 subjects (7 male and 5 female) performing 11 activities, each repeated five times, resulting in 660 sequences. The dataset offers synchronized RGB video, depth images, and inertial signals, enabling robust multimodal analysis. Actions include jumping jacks, bending, sitting, punching, throwing, waving, and clapping, providing both repetitive and dynamic movements for recognition.

c) **UTD-MHAD Dataset:** The UTD-MHAD dataset includes 27 distinct actions performed by 8 subjects (4 male and 4 female), each repeated four times. The data capture consists of RGB video, depth maps, 3D skeletal joint positions, and IMU signals from wearable sensors. Activities include sport-like gestures (tennis swing, baseball swing), fitness routines (arm curls, squats), and daily actions (picking up objects, knocking doors), supporting multimodal learning frameworks.

III. Proposed Methodology

A. Data Preprocessing

The preprocessing pipeline prepares RGB, depth, and IMU modalities for multimodal fusion while ensuring spatial and temporal alignment.

A) **RGB Modality:** For RGB data, each video sequence $\mathcal{V} = \{I_1, I_2, \ldots, I_T\}$ is uniformly sampled to obtain N_f frames according to $I'_k = I_{\lfloor k \cdot T / N_f \rfloor}$, where $k = 1, \ldots, N_f$. Each frame I'_k is resized while maintaining its aspect ratio and center-cropped to a fixed spatial resolution $(H_t, W_t) = (224, 224)$. Pixel intensities are normalized channel-wise using the ImageNet mean and standard deviation, i.e.,

$$I_k^{\text{norm}}(x, y, c) = \frac{I'_k(x, y, c)/255 - \mu_c}{\sigma_c},$$

where $\mu_c = \{0.485, 0.456, 0.406\}$ and $\sigma_c = \{0.229, 0.224, 0.225\}$ for the $c \in \{R, G, B\}$ channels. The preprocessed frames are then stacked to form an RGB tensor $\mathcal{X}_{\text{RGB}} \in \mathbb{R}^{N_f \times H_t \times W_t \times 3}$.

B) **Depth Modality:** Depth sequences $\mathcal{D} = \{D_1, \ldots, D_T\}$ undergo the same temporal sampling procedure. Each depth frame is resized to $(224, 224)$ and min–max normalized according to

$$D_k^{\text{norm}}(x, y) = \frac{D_k(x, y) - \min(D_k)}{\max(D_k) - \min(D_k)},$$

thus yielding the depth tensor $\mathcal{X}_{\text{Depth}} \in \mathbb{R}^{N_f \times H_t \times W_t \times 1}$.

C) **IMU Modality:** The IMU data, represented as a time series $S(t) \in \mathbb{R}^{N \times C}$ for C sensor channels (e.g., accelerometer and gyroscope), is first processed to replace missing values through linear interpolation. Outliers are identified using a z-score threshold, $z_c(t) = |S_c(t) - \mu_c|/\sigma_c > 3$, and replaced by averaging adjacent valid samples. To smooth high-frequency noise, a Savitzky–Golay filter is applied:

$$\hat{S}_c(t) = \sum_{k=-M}^{M} a_k S_c(t + k),$$

with a quadratic polynomial ($p = 2$) and window size $2M + 1 = 9$. Signals are then resampled to a fixed length $L_t = 180$ using linear interpolation, and normalized to zero mean and unit variance:

$$S_c^{\text{final}}(\tau) = \frac{S_c^{\text{res}}(\tau) - \bar{S}_c}{\sigma_c}.$$

To capture time–frequency characteristics, each channel undergoes a continuous wavelet transform (CWT),

$$W_c(a, b) = \int_{-\infty}^{\infty} S_c(t)\, \psi^* \left(\frac{t - b}{a} \right) dt,$$

yielding a spectrogram W_c. These spectrograms are bilinearly upsampled to match the spatial resolution of the

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

visual modalities: $W_c^{\uparrow} = \mathcal{I}_{\text{bilinear}}(W_c, (224, 224))$. The final IMU representation is thus $\mathcal{X}_{\text{IMU}} \in \mathbb{R}^{C\times 224\times 224}$.

After processing, the three modalities form the final multimodal representation

$$\mathcal{X} = \{\mathcal{X}_{\text{RGB}}, \mathcal{X}_{\text{Depth}}, \mathcal{X}_{\text{IMU}}\},$$

which is temporally aligned and normalized, providing a unified input for multimodal learning architectures.

IV. Proposed Methodology:

To classify human activities, we proposed a framework, which integrates CNN and Mamba structures. It follows a typical hierarchical design, and its basic building block is the **HMCN-FBAM Mamba block**, which unifies the **Hierarchical Multiheaded Convolution Network (HMCN)** and the **Feature-Based Attention Module (FBAM)** with the Vim (visual Mamba) block. The overall architecture of Multimodal-Mamba is shown in Fig. 3. It consists of a **Hierarchical convolution encoder**, **HMCN-FBAM Mamba blocks**, and an **activity prediction head**. Before we describe the design principles of the MultimodalMamba, it is worth introducing two important components we used in the MultimodalMamba, which are the **HMCN** and the **FBAM** block.

1) Hierarchical Multiheaded Convolution Network: After preprocessing, the multimodal data comprising RGB frames and wearable sensor signals, is transformed into a unified representation $\varphi^1 \in \mathbb{R}^{B\times C\times H\times W}$. This input is processed hierarchically through a sequence of convolutional blocks. The process begins by passing the input through a depthwise convolutional layer with a 3×3 kernel, followed by a pointwise convolution (Conv-dwp 2d), Batch Normalization (BN), and a ReLU activation function R. The resulting output, with channel width (w), is split (S) into two equal parts along the channel dimension. One part is propagated directly to the output, while the other part is concatenated with the subsequent input φ^2. The concatenated result is then passed through another depthwise convolutional layer with a 5×5 kernel, followed by pointwise convolution, BN, and ReLU activation. This hierarchical process is repeated for subsequent inputs, where the output of each layer is split, concatenated, and processed using progressively increasing kernel sizes (e.g., 7×7, 9×9), enabling multi-scale feature extraction.

$$Y_l = \begin{cases} S(R(BN(\text{Conv-dwp 2d}(\varphi_l)))), & l = 1, \\ S(R(BN(\text{Conv-dwp 2d}(Y_{l-1} \oplus \varphi_l)))), & 1 < l \leq s, \end{cases} \tag{1}$$

A skip connection is incorporated across the entire flow to stabilize training. The features from all hierarchical splits are then concatenated and passed to the subsequent attention mechanism for further processing.

2) Frequency-Based Attention Module: The frequency-based attention module using wavelets enhances the attention mechanism by applying the Discrete Wavelet Transform (DWT). The low-frequency components act like global average pooling to capture global context, while high-frequency details help refine channel importance. Building upon FcaNet, this method also incorporates spatial attention using wavelet-derived features.

For the image $f(x, y) \in \mathbb{R}^{M\times N}$, the discrete wavelet transform (DWT) coefficient using the scale function $\phi_{j_0,m,n}(x, y)$ is defined as:

$$W_\phi(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y)\, \phi_{j_0,m,n}(x, y), \tag{2}$$

where the 2D scale function is separable, satisfying:

$$\phi_{j_0,m,n}(x, y) = \phi_{j_0,m}(x)\phi_{j_0,n}(y), \tag{3}$$

Substituting the separable form into the expression for the DWT coefficient yields:

$$W_\phi(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y)\, \phi_{j_0,m}(x)\phi_{j_0,n}(y), \tag{4}$$

which is equivalent to performing convolution with scale kernels in both spatial dimensions.

In the specific case of the Haar wavelet, where $\phi_{j_0,m}(x) \cdot \phi_{j_0,n}(y) = 1$ within each local window, the DWT operation reduces to an averaging operation over subdomains. Thus, the sum of all low-frequency subband coefficients becomes:

$$\begin{aligned} \sum_{m,n=0}^{2^{j_0}-1} W_\phi(j_0, m, n) &= \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y), \\ &= \text{GAP}(f)\sqrt{MN}. \end{aligned} \tag{5}$$

where $\text{GAP}(f)$ denotes the global average pooling of the image. This shows that the total energy captured in the low-frequency Haar subband is proportional to the global average of the input image.

Inspired by this, we propose a Frequency based Attention Module, which consists of two primary components: Wavelet Channel Attention (WCA) and Wavelet Spatial Attention (WSA). Unlike conventional methods that concatenate both attentions directly, we introduce a Parallel Gated Fusion (PGF) mechanism, which adaptively adjusts the contribution of each attention component using a learned gating parameter α. This ensures that the most relevant attention information is emphasized dynamically.

a) ***Wavelet Channel Attention (WCA)****:* The WCA module utilizes the DWT to extract multi-frequency information. Given an image $f(x, y)$ of size $M \times N$, the input feature F is decomposed into four frequency subbands using:

$$\text{DWT}(F) = \{LL, LH, HL, HH\}, \tag{6}$$

where LL represents the low-frequency approximation, and LH, HL, HH capture high-frequency details. These coefficients are aggregated to extract channel-wise features:

$$W_c = \sum_{i=0}^{M/2} \sum_{j=0}^{N/2} (LL + LH + HL + HH), \tag{7}$$

where $+$ represents element-wise summation. To enhance discriminative channel attention, we apply a non-linear transformation followed by a sigmoid activation:

$$MW_c = \sigma(\text{FC}(\text{ReLU}(\text{FC}(W_c)))). \tag{8}$$

*b) **Wavelet Spatial Attention (WSA):*** The WSA module refines spatial feature learning by leveraging both low and high-frequency wavelet components. Instead of using conventional max or average pooling, we construct the spatial feature representation as:

$$W_s = \text{Cat}(LL, HL + LH + HH), \tag{9}$$

where Cat denotes concatenation. To enhance discriminative spatial attention, we apply a non-linear transformation followed by a sigmoid activation:

$$MW_s = \sigma(\text{FC}(\text{ReLU}(\text{FC}(W_s)))). \tag{10}$$

*c) **Parallel Gated Fusion of Wavelet Attention:*** To avoid a fixed combination of WCA and WSA, we employ a learned gating mechanism. The final feature representation is computed as:

$$F_{\text{final}} = \alpha \cdot MW_c + (1 - \alpha) \cdot MW_s, \tag{11}$$

where α is a learnable parameter that dynamically balances the contributions of WCA and WSA.

Finally, we concatenate F_{final} with the input Z and pass it through a fully connected layer (FC) followed by a softmax activation to obtain the output:

$$O = \text{Softmax}(\text{FC}(\text{Cat}(Z, F_{\text{final}}))). \tag{12}$$

The proposed FBAM utilizes the Discrete Wavelet Transform (DWT) with the Haar wavelet, which has a minimal filter length ($L = 2$) and a computational complexity of $\mathcal{O}(2MN)$. This lightweight transformation enables FBAM to efficiently process compressed feature maps and improves feature discrimination in complex tasks.

A. MultimodalMamba–An Efficient Activity Classification Model

To classify human activities, we propose the Multimodal-Mamba architecture, which integrates CNN-based local feature extraction with lightweight bidirectional state space modeling for efficient spatio-temporal learning. It follows a hierarchical design in which the fundamental unit is the FBAM-Mamba block, combining a Frequency-Based Attention Module (FBAM) for local channel-wise spatial enhancement and a lightweight Vim block for global temporal-spatial context modeling. Before introducing the FBAM-Mamba block, we describe the lightweight Vim block used in our architecture.

1) Lightweight Vim Block: In the proposed lightweight Vim block, the input token sequence $T_{1:l-1}$ is first normalized and projected into two vectors x and z, each of dimension d_h:

$$x, z = \text{LinearNorm}(T_{1:l-1}, d_h), \tag{13}$$

where $d_h \in \{64, 128\}$ is the reduced hidden dimension of the SSM, chosen to significantly decrease both parameter count and FLOPs compared to the original $d_h = 256$. The vector x is processed bidirectionally (forward and backward). For each direction, a depthwise 1-D convolution is applied to x to produce x'_o:

$$x'_o = \text{DWConv1D}(x, k = 3), \tag{14}$$

where DWConv1D denotes a depthwise convolution with kernel size $k = 3$, replacing the standard 1-D convolution to reduce complexity from $\mathcal{O}(C^2)$ to $\mathcal{O}(C)$. The resulting x'_o is projected into the parameters G_o, J_o, and Δ_o using low-rank linear layers:

$$\begin{aligned} G_o &= \text{LowRankLinear}(x'_o, r), \\ J_o &= \text{LowRankLinear}(x'_o, r), \\ \Delta_o &= \text{LowRankLinear}(x'_o, r), \end{aligned} \tag{15}$$

where $r \in (0, 1]$ is the projection rank ratio controlling the degree of reduction in matrix multiplications, leading to a 30%–50% saving in projection cost. The parameters Δ_o are used to update $\bar{F}_o$ and $\bar{G}_o$ of the SSM. The bidirectional SSM computations are then performed as:

$$y_{\text{forward/backward}} = \text{SSM}(\text{DWConv1D}(x)), \tag{16}$$

where the SSM follows the discretized formulation of Eqs. (11) and (12). Instead of using the fixed gating vector z from the projection step, we introduce a learnable gating mechanism derived from the forward and backward outputs:

$$g = \sigma\left(W_g[y_{\text{forward}}; y_{\text{backward}}]\right), \tag{17}$$

where $[\cdot;\cdot]$ denotes concatenation along the feature dimension, W_g is a learnable weight matrix, and $\sigma(\cdot)$ is the sigmoid activation function ensuring $g \in (0, 1)$. The final output token sequence is:

$$T_l = \text{Linear}\left(y_{\text{forward}} \odot g + y_{\text{backward}} \odot (1 - g)\right), \tag{18}$$

where $\odot$ denotes element-wise multiplication. By reducing the hidden dimension, employing depthwise convolutions, and applying low-rank projections, the lightweight Vim block achieves substantial reductions in parameters and FLOPs, making it more suitable for edge-device deployment in multimodal HAR scenarios.

2) FBAM-Mamba Block: The FBAM-Mamba block serves as the fundamental computational unit of the proposed architecture, integrating the Frequency-Based Attention Module (FBAM) with the lightweight Vim block. This integration enables the model to capture local channel-wise spatial features through FBAM while modeling global temporal-spatial dependencies via the bidirectional state space modeling of the Vim block. In contrast to conventional attention-based architectures that employ a patch embedding module to convert 2D features into 1D tokens, FBAM and the Vim block require a tailored

In Section **??**, we denote F1-Score as **Score**, Precision as **Prec.**, Recall as **Rec.**, Parameters as **Para.**, and M to represent millions.

interface because the FBAM outputs a 2D feature map whereas the Vim block accepts only 1D sequential tokens. To bridge this structural gap, the output feature map $X \in \mathbb{R}^{H \times W \times C}$ from FBAM is reshaped using a *Rearrange* operation into a token sequence of size $(H_P \times W_P, P^2 \cdot C)$, where $P \times P$ is the patch size. This rearrangement is implemented using the `einops` package. Rather than using a fully connected layer for dimensionality adjustment—which would increase parameters and computational cost—the sequence is projected to the reduced hidden dimension required by the Vim block using a lightweight depthwise `Conv1D` with kernel size 3. The transformed sequence is then processed by the lightweight Vim block, which performs bidirectional temporal modeling with a learnable gating mechanism to balance forward and backward contextual contributions. Following Vim processing, the token sequence is converted back to its spatial arrangement via an inverse `Conv1D` and *Rearrange* operation, preparing it for the next FBAM-Mamba stage. This design enables the FBAM-Mamba block to fuse local spectral-spatial emphasis with global sequence modeling while adhering to lightweight constraints for efficient edge deployment.

3) MultimodalMamba Backbone with HMCN and FBAM: The MultimodalMamba backbone adopts the FBAM-Mamba block as its core building unit within a multi-stage hierarchical structure. The network begins with a convolution encoder consisting of a Hierarchical Multi-scale Convolutional Network (HMCN) block for multi-resolution feature extraction, followed by an FBAM module for channel-wise spatial refinement, and a max-pooling layer for initial downsampling. The backbone is composed of multiple stages $S_1, S_2, \ldots, S_k$, where each stage doubles the number of channels from the previous one, and scalability can be achieved by adjusting both the number of stages and the number of FBAM-Mamba blocks within each stage. Unless otherwise specified, the patch size P and hidden state dimension D for all FBAM-Mamba blocks are fixed at $P = 7$ and $D = 256$, and the FBAM shrink rate is set to 0.25. After feature extraction, the prediction head applies global average pooling to the final feature map f_{out} to reduce its spatial dimensions, followed by a fully connected layer with Softmax activation to output the class probabilities. By combining the multi-scale convolutional capabilities of HMCN, the local channel-spatial attention of FBAM, and the lightweight global temporal modeling of the Vim block, the MultimodalMamba backbone provides a computationally efficient, hardware-friendly architecture that achieves high recognition accuracy with reduced parameter count and FLOPs, making it suitable for real-time deployment in multimodal human activity recognition on resource-constrained devices.

TABLE I
MULTIMODALMAMBA EVALUATION ON HWU-USP, BERKELEY-MHAD, AND UTD-MHAD

Model	Accuracy (%)			F1 Score (%)			FLOPs / Params		
	HWU	Berk	UTD	HWU	Berk	UTD	HWU	Berk	UTD
MultimodalMamba(1)	–	–	–	–	–	–	–	–	–
MultimodalMamba(1,1)	–	–	–	–	–	–	–	–	–
MultimodalMamba(1,2)	–	–	–	–	–	–	–	–	–
MultimodalMamba(1,2,1)	–	–	–	–	–	–	–	–	–
MultimodalMamba(1,2,2,1)	–	–	–	–	–	–	–	–	–

Description: This table compares the performance of different *MultimodalMamba* configurations across three benchmark datasets: HWU-USP, Berkeley-MHAD, and UTD-MHAD. Each configuration specifies the number of FBAM-Mamba blocks per stage (e.g., (1,2,1) means Stage 1 has one block, Stage 2 has two blocks, and Stage 3 has one block). Metrics reported include classification accuracy, F1 score, and computational complexity in terms of FLOPs and parameter count.

TABLE II
COMPARISON WITH STATE-OF-THE-ART MODELS ON HWU-USP, BERKELEY-MHAD, AND UTD-MHAD

Model	Accuracy (%)			F1 Score (%)			FLOPs / Params		
	HWU	Berk	UTD	HWU	Berk	UTD	HWU	Berk	UTD
MultimodalMamba	–	–	–	–	–	–	–	–	–
CNNGRU [5]	–	–	–	–	–	–	–	–	–
DL Framework [6]	–	–	–	–	–	–	–	–	–
Skeleton2Sensor (PSKD) [7]	–	–	–	–	–	–	–	–	–
Deep Fusion [6]	–	–	–	–	–	–	–	–	–
Ensemble RF [?]	–	–	–	–	–	–	–	–	–

Description: This table benchmarks *MultimodalMamba* against several state-of-the-art (SOTA) multimodal and unimodal HAR models, including CNNGRU, DL Framework, Skeleton2Sensor, Deep Fusion, and Ensemble RF. The comparison spans accuracy, F1 score, and model complexity. It highlights the computational efficiency of MultimodalMamba while maintaining competitive or superior recognition accuracy.

TABLE III
COMPARISON OF MULTIMODALMAMBA WITH RELATED WORK

Model	Dataset	Accuracy (%)	F1 Score (%)	FLOPs / Params	Reference
MultimodalMamba	HWU-USP	–	–	–	(This paper)
MultimodalMamba	Berkeley-MHAD	–	–	–	(This paper)
MultimodalMamba	UTD-MHAD	–	–	–	(This paper)
CNNGRU [5]	HWU-USP	–	–	–	[2]
CNNGRU [5]	Berkeley-MHAD	–	–	–	[2]
CNNGRU [5]	UTD-MHAD	–	–	–	[2]
PSKD [7]	Berkeley-MHAD	–	–	–	[45]
PSKD [7]	UTD-MHAD	–	–	–	[45]
DL Framework [6]	HWU-USP	–	–	–	[46]
DL Framework [6]	UTD-MHAD	–	–	–	[46]

Description: This table lists a dataset-wise comparison between MultimodalMamba and related HAR models, showing how each model performs on specific datasets. It helps in assessing model generalizability across different multimodal datasets and understanding performance trade-offs between architectures.

Description: This table evaluates the effect of replacing the SE (Squeeze-and-Excitation) attention module in MultimodalMamba with other visual attention mechanisms such as CoTA, CBAM, ECA, SA, and SPSA. It provides insight into how different attention designs impact classification accuracy, F1 score, and computational cost.

TABLE IV
MULTIMODALMAMBA MODELS WITH DIFFERENT VISUAL ATTENTION MODULES

Model	Accuracy (%)			F1 Score (%)			FLOPs / Params		
	HWU	Berk	UTD	HWU	Berk	UTD	HWU	Berk	UTD
MultimodalMamba (SE)	–	–	–	–	–	–	–	–	–
MultimodalMamba (CoTA)	–	–	–	–	–	–	–	–	–
MultimodalMamba (CBAM)	–	–	–	–	–	–	–	–	–
MultimodalMamba (ECA)	–	–	–	–	–	–	–	–	–
MultimodalMamba (SA)	–	–	–	–	–	–	–	–	–
MultimodalMamba (SPSA)	–	–	–	–	–	–	–	–	–

V. References Section

VI. Simple References

References

[1] Z. Quan *et al.*, "SMTDKD: A semantic-aware multimodal transformer fusion decoupled knowledge distillation method for action recognition," *IEEE Sensors J.*, vol. 24, no. 2, pp. 2289–2304, Jan. 2024.

[2] V. Martínez-Villaseñor *et al.*, "UP-Fall detection dataset: A multimodal approach," *MDPI Data*, vol. 4, no. 2, pp. 1–10, 2019.

[3] C. Chen *et al.*, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2015, pp. 168–172.

[4] F. Luo, A. Li, B. Jiang, S. Khan, K. Wu, and L. Wang, "Multimodal-Mamba: A CNN-Mamba Hybrid Neural Network for Efficient Human Activity Recognition," *IEEE Transactions on Mobile Computing*, vol. 24, no. 8, pp. 6797-6812, Aug. 2025.

[5] M. Batool, M. Alotaibi, S. A. Alotaibi, D. A. Alhammadi, M. A. Jamal, A. Jalal, and B. Lee, "Multimodal Human Action Recognition Framework Using an Improved CNNGRU Classifier," *IEEE Access*, vol. 12, pp. 158388–158406, 2024.

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

[7] J. Ni, A. H. H. Ngu, and Y. Yan, "Progressive cross-modal knowledge distillation for human action recognition," *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 5903-5912.

References