Highlights

Optimizing End-to-End Sensor-Based Human Activity Recognition through Multi-Attention Interaction

Ying Yu,Haoran Wang,Jinwei Wang,Mingke Yan,Xuerong Han,Dongchen Wu ,Qi Shen,Hanyu Liu

- Proposed an end-to-end human activity recognition process.
- Utilized an unsupervised, guidance-driven diffusion model for data augmentation of human activity data.
- Developed the Spilt Abstraction Temporal Spatial-Dynamic Update network, which is based on the Split-Attention mechanism for multi-attentional interaction.
- Designed a method for real-time updating of multi-loss function weights according to batch performance.

Optimizing End-to-End Sensor-Based Human Activity Recognition through Multi-Attention Interaction

Ying Yu[b], Haoran Wang[c], Jinwei Wang[c], Mingke Yan[a], Xuerong Han[a], Dongchen Wu [a], Qi Shen[a] and Hanyu Liu[a,*]

[a]School of Medicine and Bioinformatics Engineering, Northeastern University, Shenyang, 110169, China
[b]School of Pharmaceutical Sciences, Liaoning University, Shenyang, 110036 , China
[c]School of Information , Liaoning University, Shenyang, 110036 , China

ARTICLE INFO

Keywords:
Human Activity Recognition
Diffusion Model
Multi-Attention Interaction
Loss Function combined

ABSTRACT

Sensor-based Human Activity Recognition (HAR) is a pivotal technology for numerous intelligent human-centric applications. However, research in sensor-based HAR is still at a nascent stage, and researchers are confronted with several unresolved challenges. These include difficulties in data augmentation due to the variability in activity data distributions and the complexity of extracting deep latent features for different activities. Addressing these issues, we propose an end-to-end HAR optimization process centered around the Multi-Attention Interaction. Our approach utilizes an unsupervised statistical feature-guided diffusion model for highly adaptive data augmentation and introduces a novel network architecture, Spilt Abstraction Temporal Spatial-Dynamic Update(SATS-DU), which enhances the capability to mine deep latent features through sequenced fusion of multi-dimensional characteristics. Furthermore, we incorporate a multi-loss function fusion strategy during the training phase, dynamically adjusting the fusion weights between batches to optimize training outcomes. Extensive testing on public datasets, including ablation studies and comparisons with state-of-the-art methods, demonstrates that our approach significantly improves HAR performance, surpassing existing technologies.

1. Introduction

Human Activity Recognition (HAR) is an emerging field with broad application prospects, aiming to identify subjects' behaviors over time using motion information [1]. Current HAR systems are predominantly video-based or sensor-based [2]. Video-based systems recognize behaviors through captured images or videos, facing societal and technical challenges, such as privacy concerns, dependency on environmental lighting, resolution constraints, and the high cost and complexity associated with video processing algorithms. These challenges have significantly impeded their widespread adoption [3]. In contrast, inertial sensors, including accelerometers, gyroscopes, and magnetometers, have emerged as an ideal solution for HAR due to their privacy-preserving and convenient nature. These sensors are often embedded in wearable devices like smartwatches or gloves, offering compactness, precision, and cost-effectiveness, thus overcoming many limitations of environmental equipment [4]. With the proliferation of sensor devices, HAR has garnered considerable attention in the domain of pervasive computing. This field employs a variety of algorithms to interpret human activities, utilizing data collected from sensors attached to different parts of the body. This burgeoning research area has propelled the development of numerous context-aware applications, including healthcare, fitness monitoring, smart home technologies, and elderly fall detection [5][6]. Considerable research has been devoted to exploring HAR, initially adopting classical machine learning methods such as Decision Trees (DT) [7], Support Vector Machines (SVM) [8], Random Forests (RF) [9], and Naive Bayes (NB) [10], favored for their low computational complexity and broad applicability. However, these methods are limited by the representational features they extract, constraining classification performance and suitability for smaller datasets. Various deep neural networks, such as Convolutional Neural Networks and Long Short-Term Memory networks, have become significant research topics in extensive HAR scenarios, exhibiting sustained superiority [11, 12]. Deep learning methods, compared to traditional machine learning techniques, can automatically extract deep feature representations from sensor signals, enhancing

*Corresponding author.
yas254540645@outlook.com (Y. Yu); w1851687@gmail.com (H. Wang); Jwei3138@outlook.com (J. Wang); 20217285@stu.neu.edu.cn (M. Yan); hxr20030512@126.com (X. Han); 1812195067@qq.com (D.W.); shenq1232019@163.com (Q. Shen); 20217237@stu.neu.edu.cn (H. Liu)
ORCID(s):

the accuracy of HAR [13]. Nevertheless, deep feature extraction for sensor-based HAR continues to pose serious challenges:

Deep feature extraction: The diversity of human activities means that the same behavior may exhibit different feature patterns across various environments and contexts. For instance, the action of jumping to hit a ball in badminton is similar to that in volleyball, yet they are distinct activities. This often leads to reduced accuracy, longer development and training costs, and decreased robustness of models when applied [14].

Class Imbalance: Class Imbalance is inherent in label data, often presenting a long-tailed distribution [15]. In the real world, some human activities occur more frequently than others, leading to class imbalance issues within datasets. This imbalance can result in poor recognition performance for minority classes, as they may be overlooked due to their smaller representation, causing the model to generalize poorly on the training set [16]. The remainder of this paper is organized as follows: Section 2 reviews some related methods in the field relevant to our work. Section 3 outlines the current workflow and discusses the implementation details of our method. Section 4 presents the experimental details. Finally, Section 5 exhibits the performance of the model in experiments and discusses the current research findings.

2. Related Work

2.1. Deep Feature Extraction

The diversity, multimodality, high dimensionality, variability, and dynamism of human activities make deep feature extraction crucial for the accurate classification and recognition of these activities. Zhang et al. [17] applied residual connections in HAR, combining spatial features extracted by 1D-CNN with bidirectional long short-term memory (BLSTM) through residual links, thus enhancing the model's capability to comprehend complex temporal patterns. The ResNet network has also been adapted as an underlying network for HAR. Ronald et al. [18] improved and applied the ResNet network to HAR tasks, outperforming other deep learning architectures previously proposed on four public datasets. However, traditional ResNet models face issues such as poor inter-channel correlation, large parameter count, and insufficient feature reuse. To address these issues, Zhang et al. [19] proposed the ResNeSt network, which quickly gained widespread application upon its introduction. The network, by redesigning the feature aggregation in residual blocks and introducing the Split-Attention module, effectively captures inter-channel relationships, thereby enhancing feature extraction Mekruksavanich et al. [20] capabilities. propose a DL network with aggregation residual transformation called the ResNeXt model that can classify human activities based on inertial and stretch sensor data with satisfactory results.

2.2. Attention Mechanisms

The introduction of attention mechanisms allows models to focus more on specific aspects, thereby improving performance. Essa et al. [21] proposed the TCCSNet architecture, where the second branch consists of a set of convolutional and self-attention blocks to capture local and temporal features in sensor data. Pramanik et al. [22]proposed an attention mechanism based on deep backward transformer to guide lateral residual features, ensuring that the model learns the optimal correlation information between spatial and temporal features. Liu et al. [13]combined a multi-scale residual network with gate mechanism and ECA attention mechanism, enhancing the capability of channel feature extraction to better differentiate various human movements in daily life.

2.3. Data Augmentation

Data augmentation plays a critical role in addressing data scarcity and enhancing model generalization in HAR. Several data augmentation techniques have been proposed for HAR. Cheng et al. [23] employed contrastive supervision, using contrastive loss to supervise the intermediate layers of deep neural networks, proving effective in learning time-series invariance and improving classification accuracy. Wang et al. [24] generated samples with varying distances, angles, and human motion velocities through operations such as distance shifting, angle rotation, and velocity simulation, thus enhancing the model's generalization across different scenarios. They further improved the accuracy of HAR tasks by combining contrastive learning with generative learning and employing automatic augmentation strategy search methods [24]. Some data augmentation techniques, such as linear combination, scaling, and jittering, have preserved accurate labels in ConvLSTM networks, improving classification accuracy [25]. Diffusion models, a novel data augmentation method proposed in recent years, generate a wide range of synthetic sensor data that precisely represent the original by imposing conditional constraints on statistical properties [26]. Today, diffusion models have been applied to data augmentation for imbalanced datasets, such as in epilepsy seizure prediction. To address imbalance

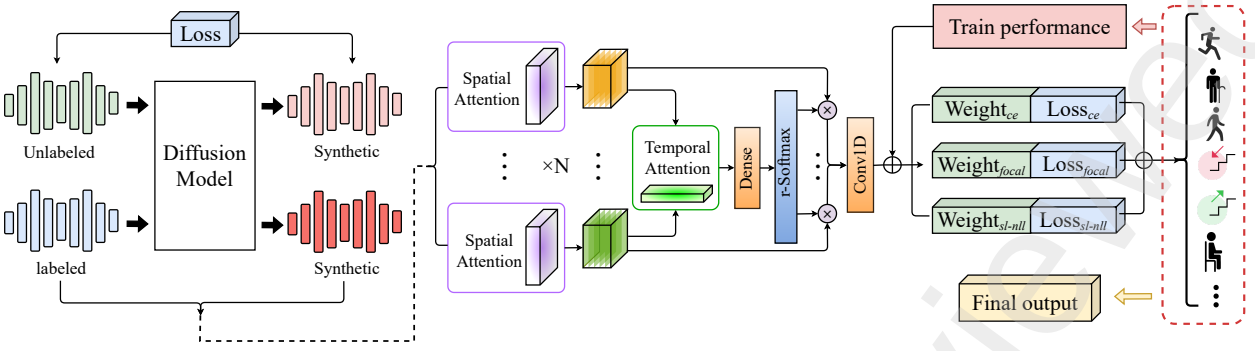

Figure 1: The total process of task.

issues, a new data augmentation method, DiffEEG, was introduced, utilizing diffusion models and demonstrating superiority over existing methods [27].

3. Methodology

3.1. Method Definition

To enhance the effectiveness of HAR, we introduce an end-to-end recognition process and present the novel SATS-DU model. During the preprocessing stage, we employ the Statistical Feature-guided Diffusion Model (SF-DM) to enrich the dataset with labeled and augmented training data, ensuring robustness and accuracy of the model. Considering the temporal and spatial correlations inherent in human activity data, we base our model on the split-attention mechanism, redefining the number of split-attention branches, channels in each branch, and the network layer structure. We integrate both temporal and spatial attention to accurately capture the interdependencies and spatiotemporal features within the data. To further enhance model performance, we devise a novel DU loss function, coupled with specific optimization strategies during the training phase, to bolster the model's learning and generalization capabilities. The synergy of these methods aims to improve the precision and efficiency of HAR tasks, offering more reliable solutions for practical applications. Figure 1 illustrates the workflow of the entire process.

3.2. Diffusion

Wearable sensor data in HAR often faces the challenges of scarce labeled training data and annotation difficulties. These challenges can compromise the accuracy and robustness of HAR models in practical applications. Hence, addressing the scarcity and complexity of labeling training data is crucial for enhancing model performance. To overcome these issues, we generate diverse synthetic sensor data using the SF-DM model proposed by Si et al. [27], which does not rely on labeled data for training, thereby improving the performance of HAR models.

Method implementation: We construct an encoder-decoder framework: the encoder consists of three convolutional layers with 9×9 kernels and a max-pooling layer with a 2×2 kernel and stride of 2, to learn features from the input. The inputs to the convolutional layers are statistical features, noise data, and embedded diffusion steps. Before entering the convolutional layer, statistical features are projected to match the shape of the noise data. The output of the diffusion steps is added to the output of the noise data, then concatenated with the output of the statistical features processed by the convolutional layer, providing additional information. The decoder comprises an upsampling layer and a convolutional layer with a 9×9 kernel. The upsampling layer is responsible for restoring resolution before matching the previous layer, while the convolutional layer transfers information contained in one data point to multiple data points. Finally, an output projection layer matches the dimensions of the output from the diffusion model to the real input data.

Method Definitions: Firstly, we extract statistical features from real data obtained from sensors. These features include the mean, standard deviation, Z-score $z = \frac{x-\mu}{\sigma}$ (where χ is the observed value, μ is the mean of all values, and σ is the standard deviation of the sample), and skewness $\gamma = \mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]$. These features have the same length as the input sequence and are fully connected as f. Then, we input the sensor data to generate noisy data $\tilde{x}$. Next, we feed $\tilde{x}$ and the

statistical features into a diffusion model, training the diffusion model to generate synthetic data by minimizing the reconstruction loss between the original real wearable sensor data and the generated data. Subsequently, we pretrain the HAR classifier using the synthetic data. Finally, we fine-tune the HAR classifier using real sensor data and export the relevant data.

The mathematical formulations of the entire process are shown in equations (1):

$$\begin{aligned} \tilde{x} &= x \times \sqrt{\beta[t]} + \in x\sqrt{1-\beta[t]} \\ L_{rec}(x, \tilde{x}, f; \theta_E) &= \frac{\sum_{l=1}^{n} |D(x_l, f_l) - x_l|}{n} \\ L_{syn}(\omega, f, y; \theta_C) &= -\sum_{l=1}^{n_c} y_l log C(E(\omega_l, f_l)) \\ L_{real}\left(x, y; \theta_C\right) &= -\sum_{l=1}^{n} y_l \log C\left(x_l\right) \end{aligned} \tag{1}$$

In the above equations, x represents the labeled real sensor data, β represents the noise level, t represents the diffusion step, and ϵ provides random noise with the same shape as the input real data. $\tilde{\chi}$ represents the input noisy data, f represents the statistical features, D represents the diffusion model with decoder and encoder, and n is the number of training samples. γ represents the corresponding class labels, E represents the pretrained diffusion model, C is the activity classifier, ω represents the random noise input to the diffusion model, and n_c represents the number of classes.

3.3. SATS

Models with split-attention can improve accuracy without increasing computational cost. Based on these advantages, our proposed SATS model improves the split-attention branch number, the number of channels in each branch and intermediate layer of each residual block, and the network layer structure, aiming to improve computational efficiency. To better capture temporal and spatial features in the data, we introduce temporal attention and spatial attention on this basis. Thus, the designed SATS model improves computational efficiency while extracting deep features from the data. Figure 2 shows the model structure of SATS.

Split Attention

First, we apply a 3×3 convolution and then use a feature map group and split attention operation to divide the features into several groups. The number of feature map groups is determined by the cardinality hyperparameter K, and another hyperparameter R represents the number of splits within each cardinality, so the total number of feature groups is $G = KR$. We can apply a series of transformations $F_1, F_2, \ldots F_G$ to each individual group, and the intermediate representation of each group is $U_i = Fi(X)$, where $i \in \{1, 2, \ldots G\}$, dispersing attention across the cardinality Then, we sum the combination of each cardinality group across multiple splits for fusion.

Temporal Attention

To address the issue of channel dependencies, we first consider the signal of each channel in the output features. Each learned filter operates with a local receptive field, so each unit of the ransformed output U cannot utilize contextual information outside that region. To alleviate this problem, we compress global spatial information into channel descriptors by generating channel-wise statistics, represented as:

$$\mathrm{AAP(U)c, j} = \frac{1}{L_{out}} \sum_{j=l}^{L_{out}} U_{c,start(i,j)} \tag{2}$$

Here, U is the input tensor, L_{out} is the output length, *start(i, j)* is the starting index of the j - th segment, fully capturing channel-wise dependencies. We choose to use a simple gating mechanism with a sigmoid activation:

$$s = F_{ex}(z, W) = \sigma\left(g\left(z, W\right)\right) = \sigma\left(W_2 \delta\left(W_1 z\right)\right) \tag{3}$$

Where δ refers to the ReLU function, $W_1 \in \mathbb{R}^{C/r \times C}$ and $W_2 \in \mathbb{R}^{C/r \times C}$. To constrain model complexity and aid generalization, we parameterize the gating mechanism by forming a bottleneck around nonlinearity with two fully

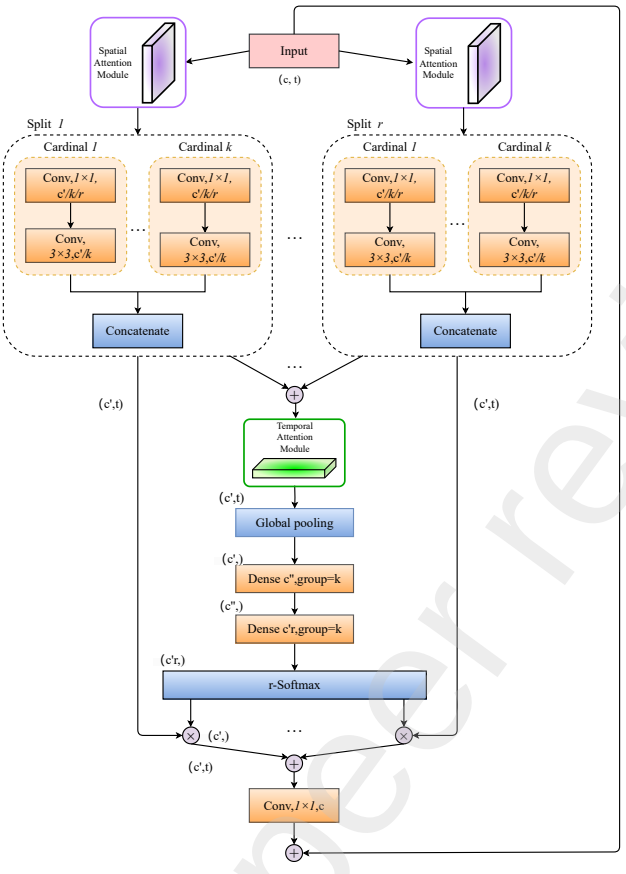

Figure 2: SATS Model

connected layers, and the final output is obtained by rescaling U using the activation s:

$$\tilde{x_c} = F_{scale}(u_c,\ s_c) = s_c \times u_c \tag{4}$$

Here, $\tilde{X} = [\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_c}]$and $F_{scale}(u_c,\ s_c)$ refers to the channel multiplication between the scalar s_c and the feature map $u_c \in \mathbb{R}^{H\times W}$.

Spatial Attention

We initiate our process by applying average pooling and max pooling operations along the temporal axis, concatenating them to forge an effective feature descriptor. Upon the concatenated features, a convolutional layer is applied to generate a spatial attention map, $M_s(F) \in R^{C\times T}$, which encodes positions to be emphasized or suppressed. Channelinformation is aggregated through the use of two pooling operations, resulting in the formation of two feature maps: $F^s_{avg} \in \mathbb{R}^{C\times T}$ and $F^s_{max} \in \mathbb{R}^{C\times T}$. Each represents the average pooling and max pooling features across channels, respectively. Subsequently, they are concatenated and convolved through a standard convolutional layer.

$$\begin{aligned} F &= ([F^s_{avg}; F^s_{max}]) \\ M_s(F) &= \sigma(f_{3\times 3} \times F)) \end{aligned} \tag{5}$$

Here, σ represents the sigmoid function, and $f_{3\times3}$ represents the convolution operation with a filter size of 3×3.

SA Block

The basic arrays of the SA Block are connected along the channel dimension:$V = \text{Concat}\{V^1, V^2, ...V^K\}$. A 3×3 convolution serves as the low-level processing unit. Before entering the split attention, we use a one-dimensional spatial attention module to focus on specific regions or elements in the input sequence. After the split attention, we reduce the number of channels and decrease the dimension of the feature maps using a 1×1 convolutional layer, similar to a standard residual block.

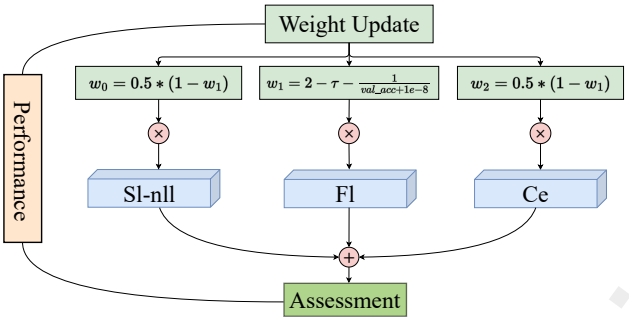

Figure 3: Loss fusion

3.4. Loss fusion

Class imbalance presents a significant challenge as it can adversely affect the classifier's training and generalization capabilities. To address this issue, this study introduces a composite loss function that integrates cross-entropy loss, focal loss, and label-smoothing regularization to enhance the model's adaptability and robustness to imbalanced data. The cross-entropy loss, a standard loss function for multi-class classification problems, optimizes the model by minimizing the disparity between the predicted probability distribution and the target distribution. The focal loss, by adjusting the parameters α(sample weight) and γ (modulating contribution of easy-to-classify examples), intensifies the model's focus on minority classes. The introduction of label-smoothing regularization aims to mitigate the model's overconfidence in its predictions, thereby improving generalization performance. The model employs the Adam optimizer for parameter updates, which incorporates the mechanisms of first-order and second-order moment estimation with bias correction, as well as a weight decay strategy. The specific expressions are as follows:

$$\begin{aligned} m_t &= \beta_l m_{t-1} + (1-\beta_l) g_t \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\ \hat{m}_t &= \frac{m_t}{r - \beta_t^t} \\ \hat{v}_t &= \frac{v_t}{l - \beta_2^t} \\ \theta_t &= \theta_{t-1} \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} (\hat{m}_t + \lambda \theta_{t-1}) \end{aligned} \tag{6}$$

where m_t and v_t represent the estimates of the first and second moments, respectively; $\hat{m}_t$ and $\hat{v}_t$ are their bias-corrected values; θ_t is the parameter vector; η is the learning rate; ϵ is a stabilizing term to prevent division by zero; and λ is the weight decay coefficient.

At the end of each training batch, the composite loss is composed of the following parts:

$$\begin{aligned} Loss_{ce} &= -\sum_{i=1} p\left(x_i\right) \log q\left(x_i\right) \\ Loss_{fl} &= -\alpha_t (1-p_t)^{\gamma} \log\left(p_t\right) \\ Loss_{sl-nll} &= -\sum_{k=1}^{K} \log p(k) \left((1-\epsilon)\delta_{k,y} + \frac{\epsilon}{K} \right) \\ Loss_{total} &= \omega_0 Loss_{sl-nll} + \omega_1 Loss_{fl} + \omega_2 Loss_{ce} \end{aligned} \tag{7}$$

Here, α_t is a weight factor in the range $[0, 1]$; γ is a modulating factor; e is a hyperparameten within $[0, 1]$; K is the number of label categories; k represents a specific label; and ω_0 , ω_1, and ω_2 are the weights of the loss functions.

To address the issue of weight selection, this study has designed an algorithm that dynamically adjusts weights based on the accuracy of the validation set during backpropagation, thereby optimizing model performance. Specifically,

Table 1
Dataset Details

Datasets	PAMAP2	WISDM
Sensor	3(40)	3
Subject	9	29
Class	9(12)	6
Window Size	171	90
Batch Size	256	512
Lr	0.001	0.005
Epoch	30	50

tasks with higher accuracy are assigned lower weights, and vice versa. The weight update formula is as follows:

$$\begin{aligned} \omega_0 &= 0.5 * (1 - \omega_1) \\ \omega_1 &= 2 - \tau - \frac{1}{acc + 1e - 8} \\ \omega_2 &= 0.5 * (1 - \omega_1) \end{aligned} \tag{8}$$

where τ is the weight factor of [0, 1], and acc is the current accuracy of the model. Ultimately, the trained model is evaluated using the validation set to calculate the model's classification accuracy and loss function value, which are used to adjust the model parameters. After model fine-tuning, the final model is assessed using the test set to determine the model's classification accuracy and the value of the composite loss function. Figure 3 shows the update of the DU loss function.

4. Experimental Design

The experiments in this study were conducted on the Kaggle platform. We utilized the default CPU, an NVIDIA P100 GPU with 16GB, and the default configurations for memory, storage, and other hardware components.

4.1. Data Sets Used in the Experiment

For the evaluation of our model's performance, we selected the WISDM and PAMAP2 datasets to facilitate an objective assessment of our method. Here are the specific details of the datasets:
PAMAP2 [28]: This dataset for HAR was released in 2012 by the Reiss and Stricker research group. It comprises data collected from 9 participants performing 18 different physical activities. Participants wore Inertial Measurement Units (IMUs) on their wrists, chest, and ankles. These IMUs include a tri-axial accelerometer, gyroscope, and magnetometer, capturing various activities such as walking, cycling, and playing soccer. Each sample was manually labeled, noting the body posture and type of activity.
WISDM [29]: This is a dataset for HAR research released by the Database Systems Research Center at Pennsylvania State University in 2010. It contains sample data of 51 volunteers performing 6 common daily activities, including standing, sitting, walking, going up and down stairs, and lying down. The data were collected using tri-axial miniature accelerometers worn on the volunteers' wrists. To match the datasets with the model test for evaluating the model's predictive outcomes, the data were preprocessed. The specific parameters of the datasets with respect to the SATS-DU model are shown in Table 1.

4.2. Evaluation Metrics

To assess the performance of the proposed HAR model, we employed multiple evaluation metrics for a comprehensive assessment of our model. Accuracy, the core metric of interest, measures the proportion of correct classifications made by the model across all samples, providing a quick overview of overall performance. F1-weighted offers a synthesis of performance for multiclass classification. Lastly, we use G-mean to examine the model's recognition effectiveness on real data after training on samples that have undergone data augmentation, ensuring robust performance across all categories. The combined use of these metrics allows for a thorough understanding of the model's strengths

Table 2
Ablation Experiment

Model		WISDM		
Main	**Additional**	**ACC**	**F1-w**	**G-mean**
SA	/	97.67	97.67	98.08
SAS		97.73	97.72	97.98
SAT		97.6	97.62	98.01
SATS	/	98.58	98.58	98.98
	DU Loss	98.45	97.31	98.25
SATS-DU	Weighted DU	98.64	98.56	98.88
	Diffusion Data	98.84	98.79	99.02

and limitations in the task of HAR.

$$\begin{aligned}
\text{Accuracy} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}} \\
\text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\
\text{F1-macro} &= \frac{2\times(\text{Precision}\times\text{Recall})}{\text{Precision}+\text{Recall}} \\
\text{F1-weighted} &= \sum_i \frac{2\times\omega_i\times(\text{Precision}_i\times\text{Recall}_i)}{\text{Precision}_i+\text{Recall}_i} \\
\text{G-mean} &= \sqrt{\frac{\text{TP}}{\text{TP}+\text{FN}}+\frac{\text{TN}}{\text{TN}+\text{FP}}}
\end{aligned} \tag{9}$$

Where TP and TN represent the number of true positives and true negatives, respectively, while FN and FP are the numbers of false negatives and false positives. Precision represents the average precision across all labels, and Recall represents the average recall rate across all labels. ω_i is the proportion of the i class samples.

5. Results and Discussion

5.1. Ablation Study

To validate the rationale behind our constructed model, we selected the WISDM dataset for our ablation study. Specifically, we first assessed the performance of the SA network with the implementation of a spatiotemporal attention mechanism. Subsequently, we validated the weight update method for multi-loss fusion. Finally, we tested the proposed data augmentation technique. The specific test of WISDM data set is shown in Table 2.

5.1.1. Module Efficacy Test

In order to better test the effectiveness of each module of the model, we built four network models: SA, SAS, SAT, and SATS. The SA network, designed to capture in-depth data features, exhibited commendable performance in terms of accuracy. The SAS network, equipped solely with spatial attention, demonstrated a marginal decline across three evaluation metrics on the WISDM dataset. We hypothesize that this decline could stem from an overemphasis on specific spatial locations or regions, neglecting the complexity of spatial relationships, consequently deteriorating all three metrics. In contrast, the SAT network, incorporating only temporal attention, showed a slight increase in accuracy and F1-W, with a minor decrease in G-mean on the WISDM dataset. We speculate that temporal attention might lead to the loss of information for certain time periods or activities, resulting in the reduced G-mean. Nevertheless, the SATS network, integrating both temporal and spatial attentions judiciously, exhibited significant improvements across all

Table 3
Comparison of related work

Model	PAMAP2(100%)		PAMAP2(50%)		WISDM	
	Accuracy	F1-weighted	Accuracy	F1-weighted	Accuracy	F1-weighted
CNN[11]	48.53	47.32	44.54	44.57	93.31	93.51
LSTM[12]	49.45	49.36	45.68	45.74	96.71	96.68
LSTM-CNN[30]	49.15	50.16	46.03	46.62	95.90	95.97
CNN-GRU[31]	51.01	50.62	47.39	47.84	94.95	96.21
SE-Res2Net[32]	53.16	53.77	48.45	49.01	95.52	95.56
ResNeXt[20]	52.11	52.67	48.73	47.85	96.67	96.66
Gated-Res2Net[33]	54.36	54.89	49.32	49.38	97.02	97.02
Rev-Attention[22]	56.42	56.87	49.79	**50.29**	97.46	97.49
MAG-Res2Net[13]	55.83	56.06	48.74	49.33	98.32	98.42
SF-DM[27]	55.4	/	**50.0**	/	/	/
SATS-DU	**56.75**	**57.32**	**49.88**	**50.29**	**98.84**	**98.79**

three evaluation metrics on the WISDM dataset, with gains of approximately. This suggests that the SATS network is adept at leveraging spatiotemporal correlation information, thereby enhancing model performance.

5.1.2. Loss Function Efficacy Test

To further enhance model performance, we introduced both the unweighted DU loss function and the weighted DU loss function into the SATS network. When evaluated with the unweighted DU loss function on the WISDM dataset, all three metrics experienced a decline, with F1-W showing a notable decrease of 1. 27%. We believe that the unweighted setting may fail to adequately balance the impact of each loss function, leading to network instability and thus affecting the performance of evaluation metrics. The introduction of an adaptively weighted DU loss function, however, resulted in improvements across all three metrics on the WISDM dataset compared to previous iterations of the network, underscoring the method's effectiveness in balancing individual loss functions and enhancing model performance.

5.1.3. Integration of SATS-DU Network with Diffusion Model

The integration of the SATS-DU network with the diffusion model yielded improvements across all three evaluation metrics compared to the aforementioned networks, marking it as the best-performing model. This robust enhancement in performance underscores the efficacy of integrating the diffusion model into the SATS-DU architecture. The refined synergy between the SATS-DU network and the diffusion model not only optimized the balance among different loss functions but also maximized the utilization of spatiotemporal information, achieving superior results in HAR tasks. This comprehensive improvement reaffirms the importance of employing advanced methodologies to elevate the state-of-the-art in activity recognition systems.

5.2. Comparison of related work

In the domain of HAR, we have contrasted our model with a spectrum of alternative approaches, which include traditional CNN and LSTM network frameworks, as well as those amalgamating state-of-the-art performance models. Further, to delineate the performance of our method with precision, networks that closely resemble our strategy were also considered for comparison, such as SE-Res2Net, Gated-Res2Net, MAGRes2Net, and SF-DM. Evaluations were carried out using the WISDM and PAMAP2 datasets. Mindful of the stringent requirements for data augmentation, we adopted a methodology inspired by our forerunners [27], selectively employing triaxial accelerometer data from the PAMAP2 dataset for both augmentation and testing, achieving laudable performance. The integration of SATS-DU network and diffusion model and the test comparison results of other network models are shown in Table 3.

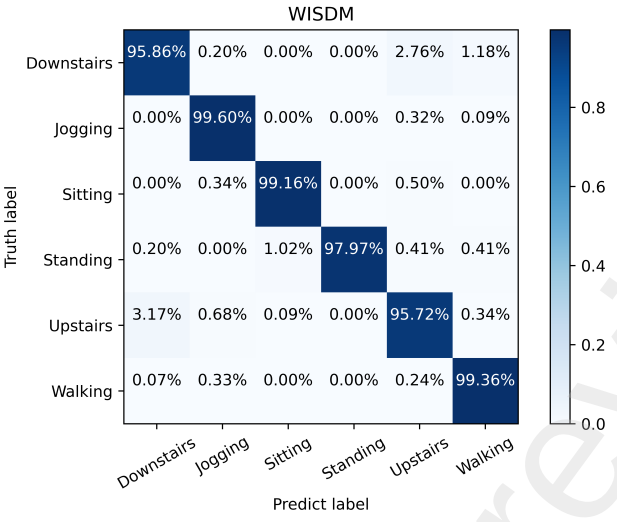

Figure 4: Confusion matrices on the WISDM

Table 4
Classification results of the WISDM dataset.

WISDM	SE-Res2Net		ResNeXt		Gated-Res2Net		Rev-Attention		MAG-Res2Net		SATS-DU	
	ACC	F1-m	ACC	F1-m	ACC	F1-m	ACC	F1-m	ACC	F1-m	ACC	F1-m
Downstairs	80.16	88.56	90.43	86.77	87.14	89.98	88.54	92.46	91.25	91.46	**95.86**	**95.91**
Jogging	99.28	99.30	99.08	99.02	98.86	99.17	99.20	99.14	99.20	98.87	**99.60**	**99.42**
Sitting	98.31	98.32	99.32	98.82	**99.66**	98.39	99.15	99.07	98.33	98.56	99.16	**99.08**
Standing	98.13	98.34	97.95	98.45	97.56	98.56	98.34	98.47	**98.74**	98.27	97.97	**98.97**
Upstairs	**96.43**	89.13	80.44	87.71	86.42	91.00	95.44	92.36	93.46	91.23	95.72	**95.56**
Walking	98.81	98.92	99.52	98.62	99.27	99.16	99.34	99.33	98.71	98.64	**99.36**	**99.43**

Notably, SATS-DU demonstrated enhanced accuracy on the PAMAP2 dataset at a 50% data volume. Although not the zenith of accuracy, the F1 weighted score of SATS-DU matched that of the previously top-performing model, Rev-Attention, yet it surpassed Rev-Attention in accuracy. With accuracies of 98. 84% on WISDM and 56. 75% on PAMAP2 (100%), SATS-DU exceeded the highest benchmarks previously set by other models. In comparison with the latest models, SATS-DU achieved an increase of 1. 35% and 0. 52% in accuracy on the PAMAP2 (100%) and WISDM datasets, respectively, with F1 weighted scores improving by 1. 26% and 0. 37%. These results validate the superior performance of our model on both straightforward and complex datasets, aptly handling tasks with class imbalances. Therefore, SATS-DU has exhibited robust capabilities in extracting deep features and discerning the spatiotemporal interrelations among them, thus affirming the accuracy of our underlying assumptions.

The confusion matrix for our model on the WISDM dataset is depicted in Figure 4. Msap-dm achieves higher accuracy across a broader spectrum of categories, effectively mitigating issues stemming from category confusion. For instance, between the "Downstairs" and "Upstairs" categories, the model exhibits a mere 2. 62% confusion rate. Additionally, in Table 4, we showcase the recognition capabilities of existing high-performance models for fine-grained categories, including both accuracy and F1 scores. It is apparent that our method maintains a performance edge in the majority of categories without incurring excessive resource consumption. Overall, our SATS-DU model can effectively handle HAR tasks in various situations and has leading performance.

6. Conclusion

In this study, we proposed a novel end-to-end optimization approach in response to the challenges faced by HAR (HAR). The unsupervised statistical feature-guided diffusion model we designed offers a highly adaptable solution for data augmentation, overcoming challenges of distributional discrepancies while ensuring sample validity, thus addressing the issue of data imbalance. Moreover, our SATS-DU model, through the Multi-Attention Interaction, effectively fused multi-dimensional features, enhancing the extraction of deep latent features and thereby improving the accuracy of activity recognition. Furthermore, our proposed multi-loss fusion strategy demonstrated its advantages during model training by dynamically adjusting loss weights, achieving an optimal combination of loss functions and further optimizing training outcomes. Experiments confirmed that our data augmentation method overcame distributional discrepancies while ensuring sample validity, addressing the data imbalance problem. The validity and performance of the proposed model were rigorously verified in the experimental section. Overall, our research not only introduces a potent method for HAR but also offers a new perspective for research in HAR methodologies. We believe these contributions will provide effective assistance for future related research.

Acknowledgment

This work was supported by National Training Program of Innovation and Entrepreneurship for Undergraduates(202310145023); National Natural Science Foundation of China (62072089); Fundamental Research Funds for the Central Universities of China (N2116016, N2104001 and N2019007).

References

[1] E. P. Ijjina, K. M. Chalavadi, Human action recognition in rgb-d videos using motion sequence information and deep learning, Pattern Recognition 72 (2017) 504–516.

[2] P. Kumar, S. Chauhan, Human activity recognition with deep learning: Overview, challenges & possibilities, CCF Transactions on Pervasive Computing and Interaction 339 (3) (2021) 1–29.

[3] P. Yang, C. Yang, V. Lanfranchi, F. Ciravegna, Activity graph based convolutional neural network for human activity recognition using acceleration and gyroscope data, IEEE Transactions on Industrial Informatics 18 (10) (2022) 6619–6630.

[4] E. Sansano, R. Montoliu, O. Belmonte Fernandez, A study of deep neural networks for human activity recognition, Computational Intelligence 36 (3) (2020) 1113–1139.

[5] S. K. Yadav, A. Luthra, K. Tiwari, H. M. Pandey, S. A. Akbar, Arfdnet: An efficient activity recognition & fall detection system using latent feature pooling, Knowledge-Based Systems 239 (2022) 107948.

[6] K. Host, M. Ivašić-Kos, An overview of human action recognition in sports based on computer vision, Heliyon (2022).

[7] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz, et al., A public domain dataset for human activity recognition using smartphones., in: Esann, Vol. 3, 2013, p. 3.

[8] C. E. Galván-Tejada, J. I. Galván-Tejada, J. M. Celaya-Padilla, J. R. Delgado-Contreras, R. Magallanes-Quintanar, M. L. Martinez-Fierro, I. Garza-Veloz, Y. López-Hernández, H. Gamboa-Rosales, et al., An analysis of audio features to develop a human activity recognition model using genetic algorithms, random forests, and neural networks, Mobile Information Systems 2016 (2016).

[9] L. Liu, L. Shao, P. Rockett, Human action recognition based on boosted feature selection and naive bayes nearest-neighbor classification, Signal Processing 93 (6) (2013) 1521–1530.

[10] F. J. Ordóñez, D. Roggen, Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition, Sensors 16 (1) (2016) 115.

[11] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, J. Zhang, Convolutional neural networks for human activity recognition using mobile sensors, in: 6th international conference on mobile computing, applications and services, IEEE, 2014, pp. 197–205.

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

[13] H. Liu, B. Zhao, C. Dai, B. Sun, A. Li, Z. Wang, Mag-res2net: A novel deep learning network for human activity recognition, Physiological Measurement 44 (11) (2023) 115007.

[14] K. Cao, Y. Chen, J. Lu, N. Arechiga, A. Gaidon, T. Ma, Heteroskedastic and imbalanced deep learning with adaptive regularization, arXiv preprint arXiv:2006.15766 (2020).

[15] S. Ahn, J. Ko, S.-Y. Yun, Cuda: Curriculum of data augmentation for long-tailed recognition, arXiv preprint arXiv:2302.05499 (2023).

[16] M. Ronald, A. Poulose, D. S. Han, isplinception: An inception-resnet deep learning architecture for human activity recognition, IEEE Access 9 (2021) 68985–69001.

[17] S. Mekruksavanich, A. Jitpattanakul, A hybrid convolution neural network with channel attention mechanism for sensor-based human activity recognition, Scientific reports (2023).

[18] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, et al., Resnest: Split-attention networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 2736–2746.

[19] Z. Xiao, X. Xu, H. Xing, F. Song, X. Wang, B. Zhao, A federated learning system with enhanced feature extraction for human activity recognition, Knowledge-Based Systems 229 (2021) 107338.

[20] S. Mekruksavanich, A. Jitpattanakul, A deep learning network with aggregation residual transformation for human activity recognition using inertial and stretch sensors, Computers 12 (7) (2023) 141.
[21] S. Agac, O. Durmaz Incel, On the use of a convolutional block attention module in deep learning-based human activity recognition with motion sensors, Diagnostics 13 (11) (2023) 1861.
[22] R. Pramanik, R. Sikdar, R. Sarkar, Transformer-based deep reverse attention network for multi-sensory human activity recognition, Engineering Applications of Artificial Intelligence 122 (2023) 106150.
[23] Z. Wang, D. Jiang, B. Sun, Y. Wang, A data augmentation method for human activity recognition based on mmwave radar point cloud, IEEE Sensors Letters (2023).
[24] C. Xu, Y. Li, D. Lee, D. H. Park, H. Mao, H. Do, J. Chung, D. Nair, Augmentation robust self-supervised learning for human activity recognition, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.
[25] S. Shao, V. Sanchez, A study on diffusion modelling for sensor-based human activity recognition, in: 2023 11th International Workshop on Biometrics and Forensics (IWBF), IEEE, 2023, pp. 1–7.
[26] Y. Wang, C. Cheng, H. Sun, J. Jin, H. Fang, Data augmentation-based statistical inference of diffusion processes, Chaos: An Interdisciplinary Journal of Nonlinear Science 33 (3) (2023).
[27] S. Zuo, V. F. Rey, S. Suh, S. Sigg, P. Lukowicz, Unsupervised statistical feature-guided diffusion model for sensor-based human activity recognition, arXiv preprint arXiv:2306.05285 (2023).
[28] A. Reiss, PAMAP2 Physical Activity Monitoring, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5NW2H (2012).
[29] J. R. Kwapisz, G. M. Weiss, S. A. Moore, Activity recognition using cell phone accelerometers, ACM SigKDD Explorations Newsletter 12 (2) (2011) 74–82.
[30] K. Xia, J. Huang, H. Wang, Lstm-cnn architecture for human activity recognition, IEEE Access 8 (2020) 56855–56866.
[31] N. Dua, S. N. Singh, V. B. Semwal, Multi-input cnn-gru based human activity recognition using wearable sensors, Computing 103 (2021) 1461–1478.
[32] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE transactions on pattern analysis and machine intelligence 43 (2) (2019) 652–662.
[33] C. Yang, M. Jiang, Z. Guo, Y. Liu, Gated res2net for multivariate time series analysis, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7.