SSDL: Sensor-to-Skeleton Diffusion Model with Lipschitz Regularization for Human Activity Recognition

Anonymous Author(s)

ABSTRACT

Human Action Recognition (HAR) involves analyzing human behavior using non-visual (e.g., sensor data) and visual data (e.g., skeleton data) and has achieved remarkable success recently. The HAR performance of existing sensor-based methods still suffers from the inherent characteristics of sensor data in terms of the lack of body pose-related 3D information, high volatility, and susceptibility to noise interference. Meanwhile, although the skeleton-based methods have achieved compelling success, benefiting from the rich, reliable spatial and temporal information of specific joint movements collected from the camera, their real-world outdoor application is unfeasible and largely limited by the strict data acquisition facilities, and the problem of occlusion. Therefore, to solve these challenges, we resort to the cross-modal generation strategy and aim to generate hard-to-collect but information-rich skeleton data conditioned on easy-to-monitor sensor data. In our work, we propose a novel Sensor-to-Skeleton Diffusion Model with Lipschitz Regularization, named SSDL. Specifically, we first design an Angular Variation module and extract angular variation information of joint movements with time information. Subsequently, consistent noise is added to the skeleton key points as well as the angular variation in the forward diffusion process. Moreover, to eliminate the challenge of noisy sensor data and enhance the stability of the training process, we integrate Lipschitz regularization with the regular training loss of the diffusion model to avoid overfitting. We verify the generalizability and effectiveness of our methods on two benchmark multimodal human action datasets: UTD-MHAD, Berkeley-MHAD, and SmartFall-MHAD dataset. Extensive results demonstrate the superiority of leveraging generated skeleton information conditioned on the sensor data for accurate human activity recognition with limited computational demands.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

Human Activity Recognition, Diffusion Models, Angular Variations, Lipschitz Continuity

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

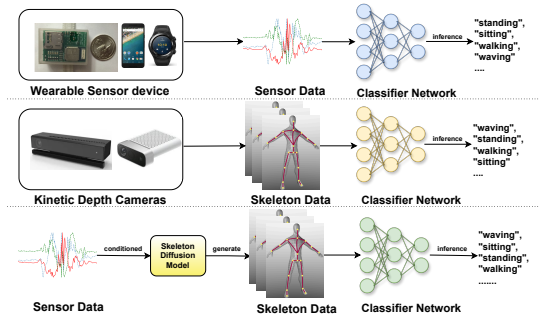

Figure 1: Comparison of various human activity recognition (HAR) systems. Form top to bottom are systems that (1) Wearable sensor-based devices are used for accelerometer/gyroscopic readings to be fed for identifying the patterns in the velocity regarding each activity, (2) the skeleton body pose of an individual is collected using Kinetic depth camera, and (3) Our proposed diffusion method where sensor data collected from wearable devices are used to generate the skeleton body pose with help of a Skeleton Diffusion Model.

ACM Reference Format:
Anonymous Author(s). 2018. SSDL: Sensor-to-Skeleton Diffusion Model with Lipschitz Regularization for Human Activity Recognition. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recently, **H**uman **A**ction **R**ecognition (HAR) task has gained increasing attention due to its vast potential in different domains like human-robot interaction, healthcare, and sports [3, 46]. According to the type of data input as shown in figure 1, HAR systems can be roughly classified into two lines: non-visual modality-based system [38] and visual modality-based one [34]. Specifically, non-visual modalities encompass sensor data such as accelerometer reading from wearable devices [42, 51], gyroscopes [72], and Wi-Fi signals [25]. Benefiting from its small footprint and widespread availability on numerous inexpensive sensors, accelerometer data collected from wearable sensor devices have propelled notable progress in the development of efficient, low-power, and cost-effective human activity recognition systems [5, 64], where potential threats and abnormal activities can be monitored in real-time and preventive measures can be taken in advance to ensure the safety and well-being of individuals. Accordingly, it gives rise to the emerging exploration of deep learning techniques to handle such time-series data, including Convolution Neural Networks (CNNs) [42, 43, 51], Recurrent Neural Networks (RNNs) [61], Long Short-Term Memory (LSTM) [13, 17, 47, 66], and Ensemble methods [35]. For instance, Mutegeki et al. [35] designed a new framework where sequential

CNN and LSTM are combined to effectively capture the spatial and temporal features. Besides, some researchers also employed an ensemble approach. For example, Guan et al. [21] developed a modified training procedure for LSTM networks, where sets of diverse LSTM learners are combined into classifier collectives through a weighted ensemble approach and outputs of multiple models are combined to enhance overall prediction accuracy. However, despite the accessibility and real-time monitoring, existing non-visual modality based systems suffer from the lack of body pose-related 3D information, high volatility and susceptibility to noise interference of sensor data, resulting in unsatisfactory performance.

In contrast, vision-based HAR systems have made significant achievements due to their remarkable advantages of rich contextual information, such as body poses and joint movements captured from camera devices. Notably, skeleton modality can eliminate the privacy concerns that occurred in the pure video-based methods [39] and encode the trajectories of human body joints. In a sense, the temporal dynamics of body parts and spatial relationships among joints can be captured with the help of a Kinetic camera [70] and the geometric 3D body movement patterns can be characterized in a continuous way. Meanwhile, the skeleton modality is not susceptible to background variations, which attracts substantial research attention to design robust systems. For example, some methods [22, 55] utilized variations of LSTM models to extract temporal information from skeleton key points. Furthermore, some researchers target at extracting both relevant spatial and temporal features with the help of transformer-based architecture [9, 28, 32, 37, 53, 59, 60, 68, 69] and Graphical Convolution Networks (GCNs). Specifically, Shi et al. [50] and Cai et al. [7] designed a two-stream graph convolution network to resolve the sparsity problem in conventional GCN-based HAR systems, and Bai et al. [4] proposed a center-connected skeleton topology to enhance the learning capability of the GCN on the potential cooperative dependencies of all joints.

Although existing sensor-based and skeleton-based efforts have achieved compelling success, their performance is still limited by the inherent characteristics of both the sensor and skeleton data. On the one hand, despite the sensor data being easily accessible using wearable devices, they only provide limited information and always lack intrusive detailed spatial information for recognizing complex real-world activities. Therefore, the activity recognition performance can be decreased, and overfitting is likely to occur during the training process. On the other hand, the acquisition of skeleton data is strict and unfeasible in outdoor setups due to the need for controlled lighting conditions and occlusion. It's not practical for real-time or on-the-go Human Activity Recognition (HAR) monitoring applications. Meanwhile, multiple frames need to be processed to gain comprehensive spatial and temporal insights from skeleton data, leading to increasing computational requirements and potential latency.

How to leverage the advantages of both sensor data and skeleton data to solve the wearable HAR problem? Our idea is to leverage the powerful cross-modal generation capabilities of generative models, where skeleton data containing rich spatial and human joint information can be generated under the condition of easily accessible sensor data. Hence, the need for detailed spatial joint movement information can be solved without complex hardware setups. Meanwhile, the diffusion model, which was initially introduced by Ho et al.[24], have succeeded in cross-modal tasks such as text-to-image and text-to-video generation[6, 20, 33, 44, 48, 58] thanks to the impressive synthesis capabilities. It involves the addition and removal of Gaussian noise to reconstruct the data. In fact, they have proven effective in generating diverse human actions based on text prompts [45, 57]. Above all, in our work, we propose a novel *Sensor-to-Skeleton Diffusion Model with Lipschitz Regularization*, named SSDL. Specifically, in the forward diffusion process, we first extract the angular information of joint movements with time information from the Angular Variation module, and then consistent noise is added to the skeleton key points as well as the angular variation step by step. Hence, our skeleton diffusion model can focus on learning joint-specific movements rather than other less relevant key points Additionally, to tackle potential noise collected from wearable devices, we introduce Lipschitz regularization as an underlying function of sensor data. The variability of the diffusion model outputs can be controlled in response to the conditioned sensor data fluctuations and the generation process can also be stabilized by compensating for noise present in the conditioning sensor data. Our main contributions can be summarized in fold:

- To the best of our knowledge, we are the first to recognize human activity based on the generated skeleton key points conditioned on the sensor data collected from the wearable devices. We used sensor data from wearable devices as a condition for the diffusion model to generate synthetic skeleton data for HAR.
- An effective loss function based on angular variations and Lipschitz regularization is added to the noise estimation loss of diffusion model to train a simple U-Net architecture, where joint-specific movements can be well learned and the generation progress can be effectively stabilized.
- Our proposed framework outperforms state-of-the-art HAR models and achieves significant performance improvement on the three MHAD datasets. As a byproduct, we will release the datasets, codes, and involved parameters to benefit other researchers and the source codes are available at here now.

2 RELATED WORK

2.1 Wearable device based HAR

Sensor devices have emerged to be one of the most convenient choices for HAR tasks. Owing to their small size, independence from environments, and consistent class-specific patterns, regardless of size. Therefore, several studies focused on optimizing human-activity recognition with the help of wearable devices [1, 2, 19, 23]. Over the last decade, the CNN-based method has also emerged to be effective in extracting the spatial features for different activities. Moreover, to account for the essential temporal dependencies in time-series sensor data, researchers have also emphasized on RNNs, variants of LSTM, and bidirectional LSTMs to focus on the temporal dynamics of HAR [12, 15, 49, 60, 66, 68, 69]. For example, Wei et al. [65] designed a framework consisting of attention mechanisms with an LSTM layer to improve the discriminatory features for HAR. However, despite the benefits mentioned above, non-visual

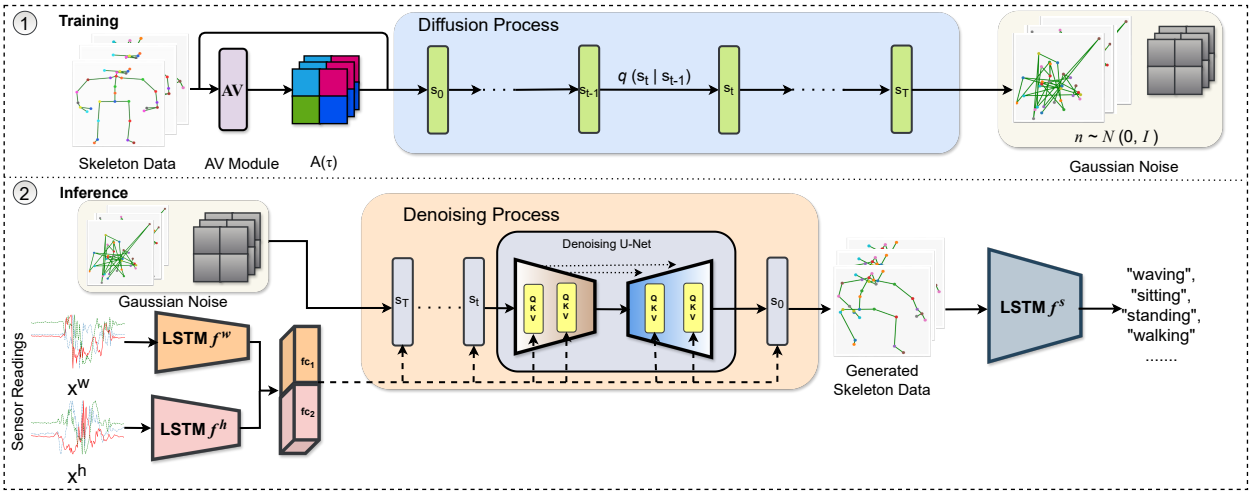

Figure 2: Overview of training approach of the SSDL model. In the first stage, we feed the skeleton data into the AV module to obtain the Angular variation module $A(\tau)$, which is then passed as an input along with the skeleton key points to the diffusion process till time step T. In the second stage, we used sensor reading from a wearable device as input to the denoising process to generate skeleton data for a specific class, which is then further used for HAR from an LSTM encoder (f^s).

sensor HAR systems lag behind visual-based modalities such as RGBD and Skeleton. This is because the sensor information only provides limited data. For example, activities such as "walking" and "running" may have similar features for certain body parts, which could negatively impact the performance of models that rely on limited information from a specific set of human parts. In this study, we aim to use important activity-related information using diffusion models to obtain skeleton data concerning the original sensor data of human activity.

2.2 Skeleton based HAR

The use of skeleton data, which contains spatiotemporal information about human motion, has gained popularity in recent years. Maxime et al. proposed using skeleton joints and their temporal variations as trajectories for Human Activity Recognition (HAR) [14]. Many conventional approaches have been proposed using handcrafted features and classifiers to improve HAR, but these methods have a limited scope for learning high-level spatiotemporal complex features. To address this issue, techniques such as CNNs and RNNs have been developed[16, 56, 60]. For example, Sun et al. [54] proposed Lattice-LSTM to integrate spatial and temporal dynamics more effectively than the traditional CNN and LSTM methods. Xu et al. [67] also proposed a 1D-CNN with an optimal training set and avoided overfitting. In addition, studies have explored the capabilities of bidirectional RNNs and LSTM to obtain better temporally aligned features from time-series data.

Furthermore, Wang et al. [62] developed an HAR system by integrating local and global attention modules in a cascade multihead attention network, leading to improved performance in temporal and spatial feature extraction. Similarly, DanHAR [18] introduced a dual attention mechanism that combines channel and temporal attention residual networks to better represent features for sensor-based human activity recognition tasks. Cha et al. [8] proposed a transformer-based approach that outperformed traditional methods by considering internal and inter-structural relationships of 3D meshes and sensor-captured skeletons. Additionally, TranSkeleton [31] introduced a unified transformer framework for modeling skeleton sequences, surpassing existing methods for skeleton-based action recognition. Although skeleton key points can enhance action recognition models, the complexity and cost of these methods may limit their practicality compared to more user-friendly options like wrist-worn devices.

2.3 Diffusion Models

Recently, a novel generative model was introduced in the form of a diffusion model, which has demonstrated remarkable performance, surpassing other generative methods such as variational auto-encoders and GANs. Ho et al. [24] proposed this synthetic data generation method using the Markov process to transform noise into data. To be precise, diffusion models consist of two consecutive stage processes, in which data are mixed with Gaussian noise in T sequential steps from s_0 to s_T, where s is the sampling latent variable and T is the number of time steps. Mathematically, each step of the forward diffusion can be described as:

$$q(s_t|s_{t-1}) := \mathcal{N}(s_t; \sqrt{1-\beta_t}s_{t-1}, \beta_t I) \quad (1)$$

where β_1 to β_T is the variance scheduler, and Equation 1 generates a noisier variation of s based on the behavior of β. If it is linear, the amount of noise added to s increases linearly with T. To obtain s at any time t, the authors proposed:

$$q(s_t|s_0) := \mathcal{N}(s_t; \sqrt{\bar{\alpha}_t}s_0, (1-\bar{\alpha}_t)I) \quad (2)$$

where $\bar{\alpha}_t := 1 - \beta_T$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. Given that it is a linear scheduler, x can be represented as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \quad (3)$$

where ϵ is from a Gaussian distribution and has the same dimensionality as the data. The reverse diffusion or denoising process aims to reconstruct the original data starting from s_T by gradually reducing the noise introduced in the forward process. This is achieved by learning a series of denoising steps that are applied

in reverse order, from step T back to step 0. Each step of reverse diffusion can be mathematically expressed as follows.

$$p_\theta(s_{t-1}|s_t) := \mathcal{N}(s_{t-1}; \mu_\theta(s_t, t), \Sigma_\theta(s_t, t)) \tag{4}$$

In Equation 4, $\mu_\theta(s_t, t)$ is the predicted mean of the original data in steps $t-1$, and $\Sigma_\theta(s_t, t)$ is the predicted covariance, both parameterized by a neural network with parameters θ. The objective is to learn θ such that $p_\theta(s_{t-1}|s_t)$ is a good approximation of the reverse of the forward process, $q(s_{t-1}|s_t)$. During training, the model parameters θ are optimized by minimizing a loss function that typically measures the difference between the model's predictions and actual data. The optimization aims to reduce the Kullback–Leibler divergence between the learned reverse and forward processes, as depicted in Equation 5.

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{q(s_1,\ldots,s_T|s_0)}\Bigg[-\log p_\theta(s_0) + \sum_{t=1}^{T} D_{\text{KL}}\big(q(s_{t-1} \mid s_t, s_0) \,\|\, p_\theta(s_{t-1} \mid s_t)\big)\Bigg] \tag{5}$$

Following the emergence of diffusion models, these advanced generative models have gained widespread adoption across a range of domains, including image, text, and time-series generation, and have demonstrated success in generating conditional data based on prompts. Notably, extensive research has also been conducted to explore the potential for generating 3D human motion from text, achieving high fidelity and diversity in the generated motions, while also exhibiting effective zero-shot learning capabilities [45, 57]. Our work utilizes the principles of conditional generation to produce 3D skeletal keypoints from wearable sensor data, effectively using accelerometer and gyroscope readings as prompts. This approach enables the model to infer activities based on body dynamics derived from data collected by wearable smartwatches on the wrist and hip, eliminating the need for a traditional setup to obtain skeleton keypoints.

3 METHODS

This section describes our proposed approach regarding the conditional generation of body points based on sensor time series data, the designed loss function involving angular variation information, and regulating the generation process under the influence of highly fluctuating sensor data with Liplitizc Regularisation (LR).

3.1 Model Architecture

In this paper, we solve the HAR task based on the generated skeleton data constructed by a conditional diffusion model (DM) where the sensor data collected from the wearable devices are regarded as conditioning. the participant wears the smartwatch on the left wrist and puts the Nexus smartphone on the right hip inside a harness. Suppose that we have a set of condition sensor data $\{(X^w, X^h)\}$, where w and h represent the wrist and hip positions.

Inspired by the huge success of representation learning, we adopt deep neural networks to obtain more powerful sensor representations. Specifically, we used two LSTM-based encoders $f^w(\Theta^w)$ and $f^h(\Theta^h)$ with parameters Θ^w and Θ^h to extract temporal features $\mathbf{A^w}$ and $\mathbf{A^h}$. Then, a simple concatenation operation is used to obtain combined features $\mathbf{Z}_0$ from $\mathbf{A^w}$ and $\mathbf{A^h}$. Furthermore, we also calculate the changes in the joint angles A using an Angular Variation (AV) module. These are then fed into the forward diffusion process along with original skeleton data (S) obtained from the Kinect camera. To improve the generation quality in terms of the skeleton structure, in the training phase, we utilize the noisy version of the original angular joints $\hat{A}_t$ at each denoising step t to calculate the angular variation loss $\mathcal{L}_c(A_t, \hat{A}_t)$. Meanwhile, we approximate the predicted noise to the added Gaussian noise with $\mathcal{L}_{\text{diff}}$ loss using a 1D-UNet [27] architecture during the reverse diffusion process. Moreover, to increase the robustness of the denoising model, we attribute a small amount of Gaussian noise δ to the conditioned embedding $\mathbf{Z}_0$ to calculate the LR loss $\mathcal{L}_{\text{reg}}$. As for the inference phase, conditioned on a pair of multi-sensor data $\{(X^w, X^h)\}$, we generate its embedding ($\mathbf{Z}_0$) from the LSTM module, which is then taken into account as conditioning for the DM to generate a specific skeleton structure, which is then fed into the classifier to recognize the specific human activity.

3.2 Angular Variations

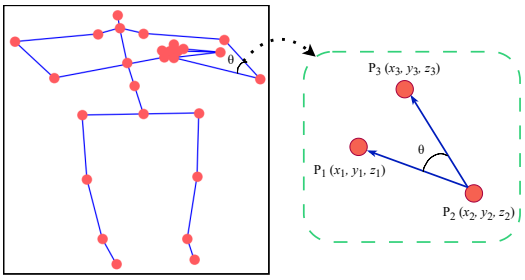

Figure 3: The group of 3-key points selection from the skeleton structure for calculating the angle between joints.

Suppose that skeleton data consists of a set of P key points, where P is determined by the the device used for capturing the skeleton data. To effectively extract the spatial and temporal features related to the evolution of joints, in the real implementation, we explore the Angular Variation (AV) module. For the given three points as shown in figure 3, the three-dimensional coordinates of three sequentially connected key points are denoted as follows,

$$P_1(x_1, y_1, z_1), \quad P_2(x_2, y_2, z_2), \quad P_3(x_3, y_3, z_3). \tag{6}$$

Then, we can define vectors $\vec{V_{2,1}}$ and $\vec{V_{2,3}}$ based on the (P_1, P_2, P_3) as follows,

$$\vec{V_{2,1}} = (x_1 - x_2, y_1 - y_2, z_1 - z_2), \tag{7}$$

$$\vec{V_{2,3}} = (x_3 - x_2, y_3 - y_2, z_3 - z_2). \tag{8}$$

Thereafter, we can obtain the angle value θ between two vectors with the help of dot product and we have,

$$\theta = \arccos\left(\frac{\vec{V_{2,1}} \cdot \vec{V_{2,3}}}{\|\vec{V_{2,1}}\|\|\vec{V_{2,3}}\|}\right), \tag{9}$$

where arccos stands for inverse cosine function. Here, θ is expressed in radians and we can convert it into degrees to obtain θ_{degree} as follows,

$$\theta_{\text{degree}} = \theta \times \left(\frac{180}{\pi}\right). \tag{10}$$

In our work, above θ_{degree} is used to obtain the angular variation information between sets of critical key points such as hands, elbow, shoulder, hip, ankle, and knee, providing crucial insights into the pose configuration in the diffusion process. Notably, we are only concerned with a subset of the angles rather than all skeleton key points to avoid the introduction of non-informative angles. Specifically, the angles between the head and neck for all frames are almost constant over time and may not provide helpful information for the discrimination of activities. Therefore, in our experiments, we consider 12 key points related to upper and lower body structure, i.e., "*shoulder->elbow->wrist*" and "*hip->ankle->knee*", and both the left and right side of the skeleton are considered. Hence, we can well capture the activity-specific features from the perspective of natural human skeleton movements. For example, the variations between the hip, knee, and foot can help to distinguish between sitting and standing actions.

3.3 Guidance through Angular Variations

After accessing the angular variations from each skeleton segment, we focused on adapting our denoising learning process to generate improved relationships between different body parts. For such, as shown in figure 2, this angular variation matrix $A(\tau)$ is used in the forward diffusion process to obtain the noisy version $\hat{A}(\tau)$ where τ represents the number of total frames. Then $\hat{A}(\tau)$ is used to evaluate the generation quality with the help of a contrastive loss strategy between the real and generated Angular Variations as shown below,

$$\mathcal{L}_c = \frac{1}{2} \left\| A(\tau) - \hat{A}(\tau) \right\|_F^2 \tag{11}$$

where $\|\cdot\|_F$ represents the Frobenius norm to measure the similarity between the generated and original angular variations across τ frames. By minimizing $\mathcal{L}_c$ loss, we encourage the model to minimize the discrepancy between the original and generated changes in the angular. In summary, we use the angular variation to guide our reverse diffusion process to generate samples following changes in key points and angular variations across time.

3.4 Lipschitzc Regularized Robust Generation

To address the issue of constant noise in the sensor data, we design a Lipschitzc Regularisation (LR) approach [36] based on the hypothesis that the generative model's response varies with noise in the input dataset. For convenience, we represent the 1D-Unet as Θ^s. More precisely, we analyze how small variations in the $\mathbf{Z_0}$ affect the generation quality. As such, we perturb the embeddings of the data with a small amount of random Gaussian noise δ. Mathematically, let $\mathbf{S}$ denote the target time-series skeleton data to be generated and $\mathbf{A^w}, \mathbf{A^h}$ denote the conditioning time-series sensor data. We include the Lipschitz regularization term with the conditional diffusion objective function to regularise the learning. The diffusion loss ensures that the diffusion model learns to predict the added noise accurately. At the same time, the LR term encourages the model to be robust to perturbations in the latent space. Therefore, our Loss function has a) Noise Prediction Loss: The noise prediction loss is the standard diffusion loss for the reverse diffusion model to estimate the amount of noise in the current sampling step. For our case, this conditioned loss function can be represented as,

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{s_0,t,\epsilon} \left[\|\epsilon - \epsilon_\theta(s_t, \mathbf{Z_0}, t)\|^2 \right], \tag{12}$$

where t is a timestep in the diffusion process, indicating the level of noise added to s_0 to obtain s_t, $\mathbf{Z_0}_t$ is embedding of the sensor data to guide the reverse diffusion process from wrist and hip and ϵ is the actual noise added to the skeleton data t-th step of the forward diffusion process. b) Lipschitzc Regularised Loss: To ensure that generated skeleton data closely match the expected output given by $\mathbf{Z_0}$, we opt to minimize the MSE loss in between the generated skeleton key points from $\mathbf{Z_0}$ and $(\mathbf{Z_0} + \delta)$.

$$\mathcal{L} = \mathbb{E}_{t,s_t,\delta} \left[\left\| f^s\left(\hat{s}(\mathbf{Z_0}, t)\right) - f^s\left(\hat{s}(\mathbf{Z_0} + \delta, t)\right) \right\|^2 \right] \tag{13}$$

where $\hat{s}_{\text{gen}}(\mathbf{Z_0}, t)$ represents the skeleton sequence generated by the model conditioned on Z_0, and $\hat{s}_{\text{gen}}(\mathbf{Z_0} + \delta, t)$ represents the skeleton sequence generated by the model conditioned on the perturbed embedding $Z_0 + \delta$. These skeleton data when passed as an input to the LSTM encoder f^s provide the output label for a specific activity. Through such, we explicitly encourage the model to maintain consistent HAR predictions for slightly varied conditioning embedding, enhancing robustness and stability in the generated outputs relative to perturbations in the sensor data embeddings.

3.5 Overall Training Objective

From the above motivation, the overall training objective of our model is a weighted combination of the loss of noise prediction for the standard diffusion model and the Lipschitz regularization term.

$$\mathcal{L}(\Phi, \Theta) = \mathcal{L}_{\text{diff}} + \lambda \cdot \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{c}} \tag{14}$$

where λ is a hyper-parameter that controls the weight of the $\mathcal{L}_{\text{reg}}$, $\mathcal{L}_{\text{c}}$ is the angular loss as explained earlier. One more advantage of our strategy is that the gradients of the model(s) can be easily computed with the help of traditional optimizers such as SGD and Adam, as shown in the equation below,

$$\Phi \leftarrow \Phi - \alpha_1 \cdot (\nabla_\Phi \mathcal{L}_{\text{diff}} + \nabla_\Phi \lambda \mathcal{L}_{\text{reg}} + \nabla_\Phi \mathcal{L}_{\text{c}}), \tag{15}$$

$$\Theta \leftarrow \Theta - \alpha_2 \cdot \nabla_\Theta \mathcal{L}_{\text{reg}}, \tag{16}$$

where α_1 and α_2 are the learning rates for specific DM and LSTM encoder branches. Moreover, during the experiments, we notice that such a joint training strategy may result in "modal collapse". Therefore, to address such challenges, we train both branches separately. Specifically, we first train the LSTM module using sensor data and then use these pre-trained embedding as conditions for *DM* to learn effective generation, and we have,

$$\nabla_\Theta \mathcal{L}\text{reg} \approx 0. \tag{17}$$

Hence, there will be no update in the sensor model's weights,

$$\begin{aligned} \Theta &\leftarrow \Theta - \alpha_2 \cdot \nabla_\Theta \mathcal{L}_{\text{reg}} \\ &\approx \Theta \end{aligned} \tag{18}$$

Thereafter, we allow our diffusion model to converge quickly without collapse of the loss function.

Algorithm 1 Training

Require: Mini-batch of data samples $(\mathbf{s}, \{(\mathbf{X}^w, \mathbf{X}^h)\})$, maximum diffusion time T, diffusion model parameters Φ, LSTM model parameters Θ, Angular Variation matrix $A(\tau)$.
Ensure: Updated parameters Φ, Θ.
1: **for** $t = 0$ to T **do**
2: Sample noise $\epsilon \sim \mathcal{N}(0, I)$ and minibatch $\{(\mathbf{s}_t^{(i)}, \mathbf{Z}_0^{(i)})\}$.
3: Compute $\mathbf{s}_{t-1}^{(i)}$ predicted by reverse diffusion:

$$\mathbf{s}_{t-1}^{(i)} = \frac{\mathbf{s}_t^{(i)} - \sqrt{1-\beta_t}\epsilon^{(i)}}{\sqrt{\beta_t}}$$

4: Calculate loss $\mathcal{L}_t(\Phi, \Theta) = \|\epsilon^{(i)} - \epsilon_\Phi(\mathbf{s}_t^{(i)}, t, \mathbf{Z}_0^{(i)})\|^2$.
5: Calculate AV $\hat{A}(\tau)$ from $\mathbf{s}_t^{(i)}$
6: Calculate $\mathcal{L}_c$ loss with equation 11
7: Add Gaussian noise δ to Z_0
8: Repeat 1-3 to generate $\hat{\mathbf{s}}_t^{(i)}$ for $(\mathbf{Z}_0 + \delta)$
9: Calculate regularization loss: L_{reg}
10: Update Φ and Θ using gradient descent on $\mathcal{L}(\Phi, \Theta)$.
11: **end for**

Algorithm 2 Inference

Require: Trained parameters Φ, Θ, initial noise level $\mathcal{N}(0, I)$, conditioning data $\{(\mathbf{X}^w, \mathbf{X}^h)\}$.
Ensure: Generated skeleton sequence $\mathbf{s}_0$.
1: Initialize $\mathbf{s}_T \sim \mathcal{N}(0, I)$.
2: **for** $t = T$ downto 1 **do**
3: Sample noise $\epsilon \sim \mathcal{N}(0, I)$ if $t > 1$, else $\epsilon = 0$.
4: Compute embeddings $\mathbf{Z}_0$ from LSTM module using $\{(\mathbf{X}^w, \mathbf{X}^h)\}$.
5: Generate $\mathbf{s}_{t-1}$ using:

$$\mathbf{s}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{s}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\Phi(\mathbf{s}_t, t, \mathbf{Z}_0)\right) + \sigma_t \mathbf{z}$$

6: **end for**
7: **return** $\mathbf{s}_0$

4 EXPERIMENTS

4.1 Dataset

In this research, we conducted experiments on the following datasets.
Berkeley MHAD [40] comprises 11 action categories performed by 12 individuals. (7 male and 5 female) in the age range between 23 to 30 years and one aged subject. All the subjects performed five repetitions of each action, generating 660 action sequences which made up 82 minutes of capturing duration.
UTD MHAD [11] contains 27 actions performed by eight subjects, with each subject repeating each Action 4 times, resulting in a total of 861 sequences. The dataset was collected using a Microsoft Kinect and wearable inertial sensors indoors. It provides multiple modalities, including RGB videos, depth maps, skeleton data, and inertial sensor data, allowing researchers to compare the performance of different methods that utilize various data modalities for human action recognition.
SmartFall MHAD is a non-public dataset collected by SmartFall Research Group at Texas State University. Data are collected from 27 participants with age greater than 60 and 12 young adults with age between 20-30. The dataset used for this paper was compiled only from 12 young adult participants (7 male and 5 female). The young adult participants performed five types of falls (front, back. left, right, and rotational) and nine pre-scribed ADLs on five repetitions each. The collection process includes the use of four types of sensors, which are three Azure Kinect cameras, a Huawei smartwatch, a Nexus smartphone, and two Meta Sensors developed by MBIENTLAB.

4.2 Training Setup

All the experiments were performed on four Nvidia GeForce RTX A6000 GPUs using PyTorch [41]. We employed a sliding window technique with a duration of three seconds for each modality, featuring a 50% overlap. For the Smaertfall dataset, we obtained 1,510 instances, categorized them into 14 distinct classes, and segregated them into training and testing sets following an 80:20 ratio. To maintain the distribution of classes, we utilized stratified sampling. Similarly, for the UTD-MHAD dataset, after removing the corrupt frames, we used 861 sequences for training and testing with a split of 80:20%. For the Berkeley dataset, we used data from the first seven participants for training and data from the remaining participants for testing.

4.3 Comparison on the HAR task

In this section, we present a comparative analysis of our method against state-of-the-art HAR approaches, as detailed in Table 1. For sensor-based HAR systems, our approach significantly outperformed most existing sensor-based models as reported by [38, 52]. This can be attributed to our modal's efficient cross-modal generation of detailed skeleton key points from sensor inputs which enhances HAR performance. Furthermore, it is noteworthy that our method demonstrates competitive performance on the UTD-MHAD dataset, comparable to that of Ni et al.'s approach [39], which utilizes cross-modality knowledge distillation for HAR. While on the Berkeley MHAD, where human motions are captured from a single sensor, our method surpasses the aforementioned method. This indicates that the effectiveness of Ni et al.'s method is easily influenced by the practical settings, while our method performs stable under various scenarios. Moreover, we also compared our method with the skeleton-based HAR methods [10, 12, 29, 63, 71]. Even though skeleton-based methods utilize skeleton data, which inherently offers precise, high-dimensional representations of human motion, our approach still our method clearly outperformed these SOTA methods, with the sensor-based modality, as presented in Table 1, validating the efficacy of our model. Therefore, our approach provides a solution to the common limitations encountered in skeleton data-capturing methods, such as the need for precise calibration and high computational costs. Our robust alternative significantly enhances both the accuracy and utility of HAR tasks. Finally, upon evaluating our method alongside RGB-based approaches, we observed enhanced performance that underscores the robustness and versatility of our method. This improvement in accuracy demonstrates that synthetic data generation can effectively address the common limitations found in RGB systems, particularly those due to poor lighting conditions and occlusions. In conclusion, our comparative findings with these SOTA methods not only validate the advantage of our approach but also potential solutions to address the persistent challenges within HAR systems.

Method	Required Modality	UTD Accuracy (%)	Berkeley Accuracy (%)	SmartFall Accuracy (%)
Hussein et al. [26]	RGB	85.5	-	N/A
Lin et al. [30]	RGB	-	96.16	N/A
Avigyan et al. [12]	Skeleton (Velocity CNN)	87.5	92.6	88.77
Avigyan et al. [12]	Skeleton (Angular CNN)	96.2	96.6	92.3
Wang et al. [63]	Skeleton	85.81	-	-
Chuankun et al. [29]	Skeleton	88.1	-	-
Zhao et al. [71]	Skeleton	92.1	-	-
Chao et al. [10]	Depth+Skeleton	93.26	-	N/A
Singh et al. [52]	Accelerometer + Gyroscope	91.4	-	89.03
Ni et al [38]	Accelerometer	95.19	94.76	94.41
Ni et al [39]	Accelerometer	**96.97**	90.18	-
Ours (SSDL)	Accelerometer	96.8	**96.9**	**94.66**

Table 1: Comparison of various models on different MHAD datasets. "-" indicates the source codes are unavailable. Meanwhile, "N/A" indicates that the method is not applicable to the dataset due to the unavailability of the testing modality.

Steps	Dataset	No Angular Variations		Angular Variations	
		Accuracy ↑	FID ↓	Accuracy ↑	FID ↓
1000	SmartFall MHAD	62.75	14.95	72.71	7.95
	Berkeley MHAD	63.24	11.2	74.91	7.28
	UTD-MHAD	62.73	15.04	77.3	6.65
4000	SmartFall MHAD	68.78	11.86	76.17	6.7
	Berkeley MHAD	68.25	9.37	78.66	5.32
	UTD-MHAD	67.84	9.11	79.6	4.67
10000	SmartFall MHAD	69.2	7.3	79.07	4.88
	Berkeley MHAD	70	9.1	84.9	4.25
	UTD-MHAD	72.64	8.1	81.11	4.60

Table 2: Ablation study of Angular Variations. The performance of all evaluated methods exhibited a consistent improvement with the inclusion of the angular variations.

4.4 Ablation study

Effect of Angular Variation. This section of the ablation study demonstrates the effect of including the Angular Variations (AV) in the diffusion model. To measure the effectiveness of our modules in the generation process, we used FID metrics, which measure the similarity between distributions generated, and a low FID score indicates better quality and closer resemblance to real samples. As presented in Table 2, we initially assessed the baseline performance of the model and observed that, regardless of the number of diffusion steps, the generation quality consistently exhibited deficiencies, as evidenced by high FID scores. Whereas, incorporating AV led to significant improvements in model performance, evidenced by increased accuracy and reduced FID scores. For instance, on the UTD-MHAD dataset, the FID score improved from 8.1 to 4.60 over 10,000 training steps, which correspondingly raised the model's accuracy. Therefore, by enabling the precise capture of joint movement dynamics through Angular Variations, the model significantly enhances its ability to differentiate subtle nuances in movements, which correspondingly increases the model's performance.

Effect of Lipschitz Regularisation To validate the effectiveness of our Lipschitz approach against noisy issues in sensor data, we monitor the change in FID score during the training. As shown in Table 3, the performance of the diffusion model step increased as shown by higher accuracy and lower FID values. Moreover, as shown in the t-SNE plot (Figure 4(c) and (d)) with the help of our LR method, the model learned distinctive features in the form of clear separation of clusters for each activity. Therefore, results show that

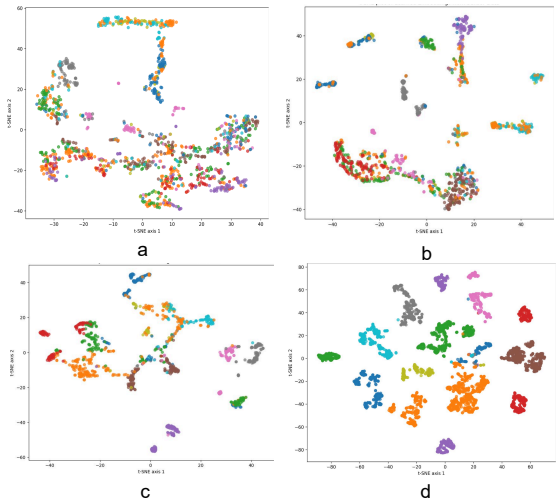

Figure 4: T-SNE analysis of the learned features on dataset SmartFall-MHAD with different training setups. a) Diffusion model and LSTM encoder are jointly learned. b) Effect of LR into the joint training setup. c) According to Equation 18, using a pre-trained model for sensor data guide the diffusion model. d) Use of the LR method clearly showed a better distinctive ability of the model to differentiate between different activities.

our method of LR allows the model to capture meaningful patterns in the data, mitigating the impact of noise and ultimately leading to improved performance and representations of human actions.

Effect of Diffusion steps. Additionally, we observed that increasing the number of steps in the diffusion model often improved its capacity to generate more realistic features. Although including angular variations hinted towards improvement in the performance of the generation quality related to each activity, it alone proved insufficient for the diffusion model to consistently converge at better solutions compared to other state-of-the-art HAR methods as shown in Table 1. Even though increasing the number of diffusion steps to improve generalization is compelling, it is computationally expensive, especially for resource-constrained devices like smartwatches.

Steps	Dataset	No Lipschitz Regularisation				With Lipschitz Regularisation			
		Joint Training		Pre-trained Embedding		Joint Training		Pre-trained Embedding	
		Accuracy ↑	FID ↓	Accuracy ↑	FID ↓	Accuracy ↑	FID ↓	Accuracy ↑	FID ↓
1000	SmartFall MHAD	77.5	4.19	83.4	2.40	76.41	4.16	84.1	1.81
	Berkeley MHAD	79.3	3.85	86.2	1.88	77.91	4.06	88.04	0.94
	UTD-MHAD	75.8	4.9	81.2	2.9	78.9	11.02	89.66	0.82
4000	SmartFall MHAD	76.4	4.65	83.16	2.35	77.23	6.45	90.44	0.85
	Berkeley MHAD	77.6	4.07	85.36	2.11	77.9	6.07	91.5	0.71
	UTD-MHAD	84.1	2.02	86.77	2.06	80.2	2.07	93.1	0.53
10000	SmartFall MHAD	80.5	2.27	85.01	2.26	77.72	4.91	94.66	0.6
	Berkeley MHAD	82	2.1	84.9	2.16	78.77	4.72	96.9	0.50
	UTD-MHAD	82.28	2.02	86.92	2.02	81.33	2.49	96.8	0.59

Table 3: Comparative Analysis of Model Performance with and without Lipschitz Regularization under Joint Training and Pre-trained Embedding Scenarios.

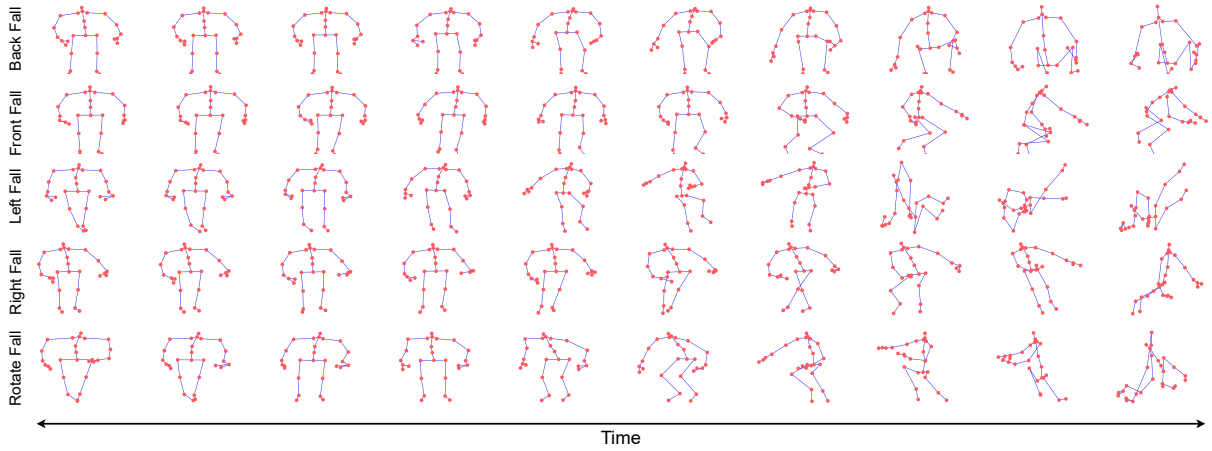

Figure 5: Synthetic set of actions generated by our SSDL diffusion-based model trained on Smartfall MHAD dataset.

Using pre-trained LSTM embedding. The results presented in Table 3 and Figure 4 indicate that the performance of the jointly trained model is relatively low, which implies that there might be an issue of potential modal collapse. To address such, we first trained the LSTM encoder and then used the pre-trained embeddings as conditioning for the reverse diffusion. The application of this approach resulted in improved model performance, as shown in Table 3. This improvement can be attributed to the reduced variation in the conditional signal to the generation process. To ensure a fair comparison, we used the same backbone architecture for all evaluations, with the only variation being the sampling time steps. The results in Table 3 affirm that using pre-trained embeddings significantly enhances the generation of skeleton data, closely mimicking authentic datasets, while the progressive increase in training steps further amplifies the model's learning capabilities.

.

Qualitative Results for the generated samples. The figure shown in Figure 5 depicts various synthetic action sequences that were generated using our proposed SSDL approach for different activities. This visualization allows us to evaluate the stability of the model in producing coherent action sets for each activity. Our SSDL method is capable of producing more naturalistic action sequences with minimal noise. As demonstrated in the figure, the synthetic sequences derived from sensor data have also improved the model's classification performance across different classes.

4.5 Inference Speed

Our SSDL model achieves 12ms on producing 90 frames with 1000 DDPM samples on a single NVIDIA A6000 since the lightweight architecture of our model. Our model inputs 75 frames for the sensor data, to which our DDPM produces its respective behavior. Our classification based on the generated skeleton data takes only 3ms to predict specific activity; a quick performance is necessary considering that activities such as "*fall*" need a quick inference time to prevent serious injuries, especially in old age people.

5 CONCLUSION

In this study, we propose a novel approach for generating skeleton critical points based on sensor data from wearable devices to solve HAR problems. The derived angular variation guided the reverse diffusion process to generate more meaningful movements. Based on our proposed framework, we generated more diverse and aligned body part movements that inherited kinetically meaningful body structures. The proposed method obtains state-of-the-art results on the two datasets and shows competitive performance in the UTD-MHAD dataset for Human Action Recognition. Our experiments further demonstrated the effectiveness of the Lipschitz Loss, which improved the performance of the diffusion model. Furthermore, the low inference time for generating specific skeleton information from sensor data validates the real-world applicability of our scheme for a robust HAR system.

REFERENCES

[1] Alexandru Iulian Alexan, Anca Roxana Alexan, and Stefan Oniga. 2024. Real-Time Machine Learning for Human Activities Recognition Based on Wrist-Worn Wearable Devices. *Applied Sciences* 14, 1 (2024). https://doi.org/10.3390/app14010329

[2] Shan E Ali, Ali Nawaz Khan, Shafaq Zia, and Mayyda Mukhtar. 2020. Human Activity Recognition System using Smart Phone based Accelerometer and Machine Learning. In *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*. 69–74. https://doi.org/10.1109/IAICT50021.2020.9172037

[3] Athanasios Anagnostis, Lefteris Benos, Dimitrios Tsaopoulos, Aristotelis Tagarakis, Naoum Tsolakis, and Dionysis Bochtis. 2021. Human Activity Recognition through Recurrent Neural Networks for Human–Robot Interaction in Agriculture. 11, 5 (2021). https://doi.org/10.3390/app11052188

[4] Zhongyu Bai, Qichuan Ding, Hongli Xu, Jianning Chi, Xiangyue Zhang, and Tiansheng Sun. 2022. Skeleton-based similar action recognition through integrating the salient image feature into a center-connected graph convolutional network. *Neurocomputing* 507 (2022), 40–53. https://doi.org/10.1016/j.neucom.2022.07.080

[5] Karim Bayoumy, Mohammed Gaber, Abdallah Elshafeey, Omar Mhaimeed, Elizabeth H. Dineen, Francoise A. Marvel, Seth S. Martin, Evan D. Muse, Mintu P. Turakhia, Khaldoun G. Tarakji, and Mohamed B. Elshazly. 2021. Smart wearable devices in cardiovascular care: where we are and how to move forward. *Nature Reviews Cardiology* 18, 8 (2021), 581–599. https://doi.org/10.1038/s41569-021-00522-7

[6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22563–22575. https://doi.org/10.1109/CVPR52729.2023.02161

[7] Jinmiao Cai, Nianjuan Jiang, Xiaoguang Han, Kui Jia, and Jiangbo Lu. 2021. JOLO-GCN: Mining Joint-Centered Light-Weight Information for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2735–2744.

[8] Junuk Cha, Muhammad Saqlain, Donguk Kim, Seungeun Lee, Seongyeong Lee, and Seungryul Baek. 2022. Learning 3D Skeletal Representation From Transformer for Action Recognition. *IEEE Access* 10 (2022), 67541–67550. https://doi.org/10.1109/ACCESS.2022.3185058

[9] Li Chao, Zhong Qiaoyong, Xie Di, and Pu Shiliang. 2017. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE International Conference on Multimedia & Expo Workshops*. 597–600. https://doi.org/10.1109/ICMEW.2017.8026285

[10] Xin Chao, Genlin Ji, and Xiaosha Qii. 2024. Multi-view key information representation and multi-modal fusion for single-subject routine action recognition. *Applied Intelligence* 54, 4 (2024), 3222–3244. https://doi.org/10.1007/s10489-024-05319-y

[11] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International Conference on Image Processing (ICIP)*. 168–172. https://doi.org/10.1109/ICIP.2015.7350781

[12] Avigyan Das, Pritam Sil, Pawan Kumar Singh, Vikrant Bhateja, and Ram Sarkar. 2021. MMHAR-EnsemNet: A Multi-Modal Human Activity Recognition Model. *IEEE Sensors Journal* 21, 10 (2021), 11569–11576. https://doi.org/10.1109/JSEN.2020.3034614

[13] Jingyang Deng, Shuyi Zhang, and Jinwen Ma. 2023. Self-Attention-Based Deep Convolution LSTM Framework for Sensor-Based Badminton Activity Recognition. *Sensors* 23, 20 (2023). https://doi.org/10.3390/s23208373

[14] Maxime Devanne, Hazem Wannous, Stefano Berretti, Pietro Pala, Mohamed Daoudi, and Alberto Del Bimbo. 2015. 3-D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold. *IEEE Transactions on Cybernetics* 45, 7 (2015), 1340–1352. https://doi.org/10.1109/TCYB.2014.2350774

[15] Yong Du, Yun Fu, and Liang Wang. 2015. Skeleton based action recognition with convolutional neural network. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. 579–583. https://doi.org/10.1109/ACPR.2015.7486569

[16] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*. 1110–1118. https://doi.org/10.1109/CVPR.2015.7298714

[17] Kavya Sree Gajjala and Basabi Chakraborty. 2021. Human Activity Recognition based on LSTM Neural Network Optimized by PSO Algorithm. In *2021 IEEE 4th International Conference on Knowledge Innovation and Invention (ICKII)*. 128–133. https://doi.org/10.1109/ICKII51822.2021.9574788

[18] Wenbin Gao, Lei Zhang, Qi Teng, Jun He, and Hao Wu. 2021. DanHAR: Dual Attention Network for multimodal human activity recognition using wearable sensors. *Applied Soft Computing* 111 (2021), 107728. https://doi.org/10.1016/j.asoc.2021.107728

[19] Daniel Garcia-Gonzalez, Daniel Rivero, Enrique Fernandez-Blanco, and Miguel R. Luaces. 2023. New machine learning approaches for real-life human activity recognition using smartphone sensor-based data. *Knowledge-Based Systems* 262 (2023), 110260. https://doi.org/10.1016/j.knosys.2023.110260

[20] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10686–10696. https://doi.org/10.1109/CVPR52688.2022.01043

[21] Yu Guan and Thomas Plotz. 2017. Ensembles of Deep LSTM Learners for Activity Recognition using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2 (jun 2017). https://doi.org/10.1145/3090076

[22] Hongling Guo, Zhitian Zhang, Run Yu, Yakang Sun, and Heng Li. 2023. Action Recognition Based on 3D Skeleton and LSTM for the Monitoring of Construction Workers Safety Harness Usage. *Journal of Construction Engineering and Management* 149, 4 (2023), 04023015.

[23] Wei Guo, Shunsei Yamagishi, and Lei Jing. 2024. Human Activity Recognition via Wi-Fi and Inertial Sensors With Machine Learning. *IEEE Access* !2 (2024), 18821–18836. https://doi.org/10.1109/ACCESS.2024.3360490

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the Advances in neural information processing systems*, Vol. 33. 6840–6851.

[25] Jinyang Huang, Bin Liu, Pengfei Liu, Chao Chen, Ning Xiao, Yu Wu, Chi Zhang, and Nenghai Yu. 2020. Towards Anti-interference WiFi-based Activity Recognition System Using Interference-Independent Phase Component. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*. 576–585. https://doi.org/10.1109/INFOCOM41043.2020.9155536

[26] Mohamed E Hussein, Marwan Torki, Mohammad A Gowayyed, and Motaz El-Saban. 2013. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Twenty-third international joint conference on artificial intelligence*.

[27] Zanobya N. Khan and Jamil Ahmad. 2021. Attention induced multi-head convolutional neural network for human activity recognition. *Applied Soft Computing* 110 (2021), 107671. https://doi.org/10.1016/j.asoc.2021.107671

[28] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. 2017. Ensemble Deep Learning for Skeleton-Based Action Recognition Using Temporal Sliding LSTM Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[29] Chuankun Li, Yonghong Hou, Pichao Wang, and Wanqing Li. 2017. Joint Distance Maps Based Action Recognition With Convolutional Neural Networks. *IEEE Signal Processing Letters* 24, 5 (2017), 624–628. https://doi.org/10.1109/LSP.2017.2678539

[30] Ji Lin, Chuang Gan, and Song Han. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 7082–7092. https://doi.org/10.1109/ICCV.2019.00718

[31] Haowei Liu, Yongcheng Liu, Yuxin Chen, Chunfeng Yuan, Bing Li, and Weiming Hu. 2023. TranSkeleton: Hierarchical Spatial–Temporal Transformer for Skeleton-Based Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 8 (2023), 4137–4148. https://doi.org/10.1109/TCSVT.2023.3240472

[32] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C. Kot. 2018. Skeleton-Based Human Action Recognition With Global Context-Aware Attention LSTM Networks. *IEEE Transactions on Image Processing* 27, 4 (2018), 1586–1599. https://doi.org/10.1109/TIP.2017.2785279

[33] Jiawei Liu, Weining Wang, Wei Liu, Qian He, and Jing Liu. 2023. ED-T2V: An Efficient Training Framework for Diffusion-based Text-to-Video Generation. In *2023 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN54540.2023.10191565

[34] L. Minh Dang, Kyungbok Min, Hanxiang Wang, Md. Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. 2020. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition* 108 (2020), 107561. https://doi.org/10.1016/j.patcog.2020.107561

[35] Ronald Mutegeki and Dong Seog Han. 2020. A CNN-LSTM Approach to Human Activity Recognition. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. 362–366. https://doi.org/10.1109/ICAIIC48513.2020.9065078

[36] Ramchandran Muthukumar and Jeremias Sulam. 2023. Adversarial robustness of sparse local lipschitz predictors. *SIAM Journal on Mathematics of Data Science* 5, 4 (2023), 920–948.

[37] Mihai Nan, Mihai Trăscău, and Adina-Magda Florea. 2024. Spatio-temporal neural network with handcrafted features for skeleton-based action recognition. *Neural Computing and Applications* (feb 2024). https://doi.org/10.1007/s00521-024-09559-4

[38] Jianyuan Ni, Anne H.H. Ngu, , and Yan Yan. 2022. Progressive Cross-modal Knowledge Distillation for Human Action Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM, Lisboa, Portugal, 10–14. https://doi.org/10.1109/ICCV51070.2023.00942

[39] Jianyuan Ni, Raunak Sarbajna, Yang Liu, Anne H.H. Ngu, and Yan Yan. 2022. Cross-Modal Knowledge Distillation For Vision-To-Sensor Action Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4448–4452. https://doi.org/10.1109/ICASSP43922.2022.9746752

[40] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*. 53–60. https://doi.org/10.1109/WACV.2013.6474999

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[42] Nabasmita Phukan, Shailesh Mohine, Achinta Mondal, M. Sabarimalai Manikandan, and Ram Bilas Pachori. 2022. Convolutional Neural Network-Based Human Activity Recognition for Edge Fitness and Context-Aware Health Monitoring Devices. *IEEE Sensors Journal* 22, 22 (2022), 21816–21826. https://doi.org/10.1109/JSEN.2022.3206916

[43] Ravi Raj and Andrzej Kos. 2023. An improved human activity recognition technique based on convolutional neural network. *Scientific Reports* 13, 1 (2023), 22581.

[44] Apoorva Rauniyar, Aryan Raj, Ashish Kumar, Ashish Kumar Kandu, Astha Singh, and Anjani Gupta. 2023. Text to Image Generator with Latent Diffusion Models. In *2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*. 144–148. https://doi.org/10.1109/CICTN57981.2023.10140348

[45] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. 2023. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096441

[46] Alina Roitberg, Alexander Perzylo, Nikhil Somani, Manuel Giuliani, Manuel Rickert, and Alois Knoll. 2014. Human activity recognition in the context of industrial human-robot interaction. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. 1–10. https://doi.org/10.1109/APSIPA.2014.7041588

[47] Pornthep Rojanavasu, Anuchit Jitpattanakul, and Sakorn Mekruksavanich. 2021. Comparative Analysis of LSTM-based Deep Learning Models for HAR using Smartphone Sensor. In *2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering*. 269–272. https://doi.org/10.1109/ECTIDAMTNCON51128.2021.9425733

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22500–22510. https://doi.org/10.1109/CVPR52729.2023.02155

[49] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1010–1019. https://doi.org/10.1109/CVPR.2016.115

[50] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[51] Amandeep Singh, Tiziana Margaria, and Florenc Demrozi. 2023. CNN-based Human Activity Recognition on Edge Computing Devices. In *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*. 1–4. https://doi.org/10.1109/COINS57856.2023.10189270

[52] Satya P Singh, Madan Kumar Sharma, Aimé Lay-Ekuakille, Deepak Gangwar, and Sukrit Gupta. 2020. Deep ConvLSTM with self-attention for human activity decoding using wearable sensors. *IEEE Sensors Journal* 21, 6 (2020), 8575–8582.

[53] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2017. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. *Proceedings of the AAAI Conference on Artificial Intelligence* 31, 1 (Feb. 2017). https://doi.org/10.1609/aaai.v31i1.11212

[54] Lin Sun, Kui Jia, Kevin Chen, Dit-Yan Yeung, Bertram E. Shi, and Silvio Savarese. 2017. Lattice Long Short-Term Memory for Human Action Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[55] Tan-Hsu Tan, Ching-Jung Huang, Munkhjargal Gochoo, and Yung-Fu Chen. 2021. Activity Recognition Based on FR-CNN and Attention-Based LSTM Network. In *2021 30th Wireless and Optical Communications Conference (WOCC)*. 146–149. https://doi.org/10.1109/WOCC53213.2021.9603203

[56] Yansong Tang, Yi Tian, Jiwen Lu, Peiyang Li, and Jie Zhou. 2018. Deep Progressive Reinforcement Learning for Skeleton-Based Action Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5323–5332. https://doi.org/10.1109/CVPR.2018.00558

[57] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=SJ1kSyO2jwu

[58] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1921–1930. https://doi.org/10.1109/CVPR52729.2023.00191

[59] Shahab Uddin, Tahir Nawaz, James Ferryman, Nasir Rashid, Md. Asaduzzaman, and Raheel Nawaz. 2024. Skeletal Keypoint-Based Transformer Model for Human Action Recognition in Aerial Videos. *IEEE Access* 12 (2024), 11095–11103. https://doi.org/10.1109/ACCESS.2024.3354389

[60] Vivek Veeriah, Naifan Zhuang, and Guo-Jun Qi. 2015. Differential Recurrent Neural Networks for Action Recognition. In *2015 IEEE International Conference on Computer Vision*. IEEE Computer Society, Los Alamitos, CA, USA, 4041–4049. https://doi.org/10.1109/ICCV.2015.460

[61] Ramya Anasseriyil Viswambaran, Gang Chen, Bing Xue, and Mohammad Nekooei. 2019. Evolutionary Design of Recurrent Neural Network Architecture for Human Activity Recognition. In *2019 IEEE Congress on Evolutionary Computation (CEC)*. 554–561. https://doi.org/10.1109/CEC.2019.8790050

[62] Jiaze Wang, Xiaojiang Peng, and Yu Qiao. 2020. Cascade multi-head attention networks for action recognition. *Computer Vision and Image Understanding* 192 (2020), 102898. https://doi.org/10.1016/j.cviu.2019.102898

[63] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li. 2016. Action Recognition Based on Joint Trajectory Maps Using Convolutional Neural Networks. In *Proceedings of the 24th ACM International Conference on Multimedia* (Amsterdam, The Netherlands). Association for Computing Machinery, New York, NY, USA, 102–106. https://doi.org/10.1145/2964284.2967191

[64] Yan Wang, Shuang Cang, and Hongnian Yu. 2019. A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications* 137 (2019), 167–190. https://doi.org/10.1016/j.eswa.2019.04.057

[65] Xiong Wei and Zifan Wang. 2024. TCN-attention-HAR: human activity recognition based on attention mechanism time convolutional network. *Scientific Reports* 14 (2024). https://doi.org/10.1038/s41598-024-57912-3

[66] Yin Xiaochun, Liu Zengguang, Liu Deyong, and Ren Xiaojun. 2022. A Novel CNN-based Bi-LSTM parallel model with attention mechanism for human activity recognition with noisy data. *Scientific Reports* 12 (2022). Issue 1. https://doi.org/10.1038/s41598-022-11880-8

[67] Zhiou Xu, Juan Zhao, Yi Yu, and Haijun Zeng. 2020. Improved 1D-CNNs for behavior recognition using wearable sensor network. *Computer Communications* 151 (2020), 165–171. https://doi.org/10.1016/j.comcom.2020.01.012

[68] Zhengyuan Yang, Yuncheng Li, Jianchao Yang, and Jiebo Luo. 2019. Action Recognition With Spatio–Temporal Visual Attention on Skeleton Image Sequences. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 8 (2019), 2405–2415. https://doi.org/10.1109/TCSVT.2018.2864148

[69] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. 2017. View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition From Skeleton Data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[70] Zhengyou Zhang. 2012. Microsoft Kinect Sensor and Its Effect. *IEEE MultiMedia* 19, 2 (2012), 4–10. https://doi.org/10.1109/MMUL.2012.24

[71] Rui Zhao, Kang Wang, Hui Su, and Qiang Ji. 2019. Bayesian Graph Convolution LSTM for Skeleton Based Action Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 6881–6891. https://doi.org/10.1109/ICCV.2019.00698

[72] Muhammad Zubair, Kibong Song, and Changwoo Yoon. 2016. Human activity recognition using wearable accelerometer sensors. In *Proceedings of 2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. 1–5. https://doi.org/10.1109/ICCE-Asia.2016.7804737