IMU2IMG: IMU in the Language of Vision Foundation Models

Sun Kyung Lee*
sklee2014@etri.re.kr
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea

Se Won Oh
sewonoh@etri.re.kr
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea

Gyuwon Jung
gwjung@etri.re.kr
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea

Hyuntae Jeong
htjeong@etri.re.kr
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea

Seungeun Chung
schung@etri.re.kr
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea

Jeong Mook Lim
jmlim21@etri.re.kr
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea

Kyoung Ju Noh
kjnoh@etri.re.kr
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea

Abstract

Motivated by the growing interest in foundation models across domains such as vision and language, this study aims to explore their potential in the field of human locomotion recognition using wearable sensor data. In response to Task 1 of Sussex-Huawei Locomotion-Transportation (SHL) recognition challenge 2025, which focuses on the foundation model based user-independent activity recognition with inertial signals, our team 'HELP' proposes IMU2IMG, an algorithmic pipeline that leverages a vision foundation model for prediction. Specifically, our method transforms multimodal inertial data into RGB images, enabling explicit alignment between sensor representations and visual tokens. Through extensive experiments, we compare IMU2IMG with conventional machine learning, deep learning, and recent foundation model-based baselines. Our results demonstrate superior performance and offer insights on how sensor-to-vision mappings can support robust and interpretable human activity recognition.

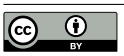
CCS Concepts

• **Computing methodologies** → **Supervised learning by classification**; • **Human-centered computing** → **Ubiquitous and mobile computing**.

Keywords

Activity recognition, Foundation model, Multimodal sensors, Human locomotion, SHL Dataset

*Corresponding author



UbiComp Companion '25, Espoo, Finland

ACM ISBN 979-8-4007-1477-1/2025/10
https://doi.org/10.1145/3714394.3756208

ACM Reference Format:
Sun Kyung Lee, Se Won Oh, Gyuwon Jung, Hyuntae Jeong, Seungeun Chung, Jeong Mook Lim, and Kyoung Ju Noh. 2025. IMU2IMG: IMU in the Language of Vision Foundation Models. In *Companion of the the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '25), October 12–16, 2025, Espoo, Finland*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3714394.3756208

1 Introduction

Human activity recognition (HAR) plays a crucial role in a wide range of applications, including personalized services, intelligent transportation systems, and health monitoring [4]. The Sussex-Huawei Locomotion-Transportation (SHL) recognition challenge series has become a benchmark for real-world HAR, utilizing multimodal sensor data collected via smartphones [8, 25]. Over the past few years [20, 23, 24], numerous studies have leveraged this dataset to develop models for activity recognition, contributing significantly to the field of wearable sensing and mobile computing.

The 2025 edition of the SHL recognition challenge introduces a novel focus, which is the application of foundation models to human activity recognition. Inspired by the rapid advances in large-scale models across domains such as language and vision, SHL recognition challenge 2025 Task 1 invites participants to leverage foundation models for transportation mode recognition using 9-axis inertial sensor data. This shift marks a critical evolution in the field, highlighting the cross-domain transferability and generalization capabilities of foundation models in mobile sensing tasks.

Several recent studies have explored the use of foundation models to tackle sensor analysis tasks, including human activity recognition. A common strategy involves translating sensor data into other modalities such as time-series [13], text [12], or image representations [15], allowing models from those domains to be applied effectively. For example, sensor-to-text translation often leverages intermediate adapters or prompting layers [9, 10], while sensor-to-image approaches typically involve manually crafted transformations such as spectrograms [5] or graph-based [16] visualizations.

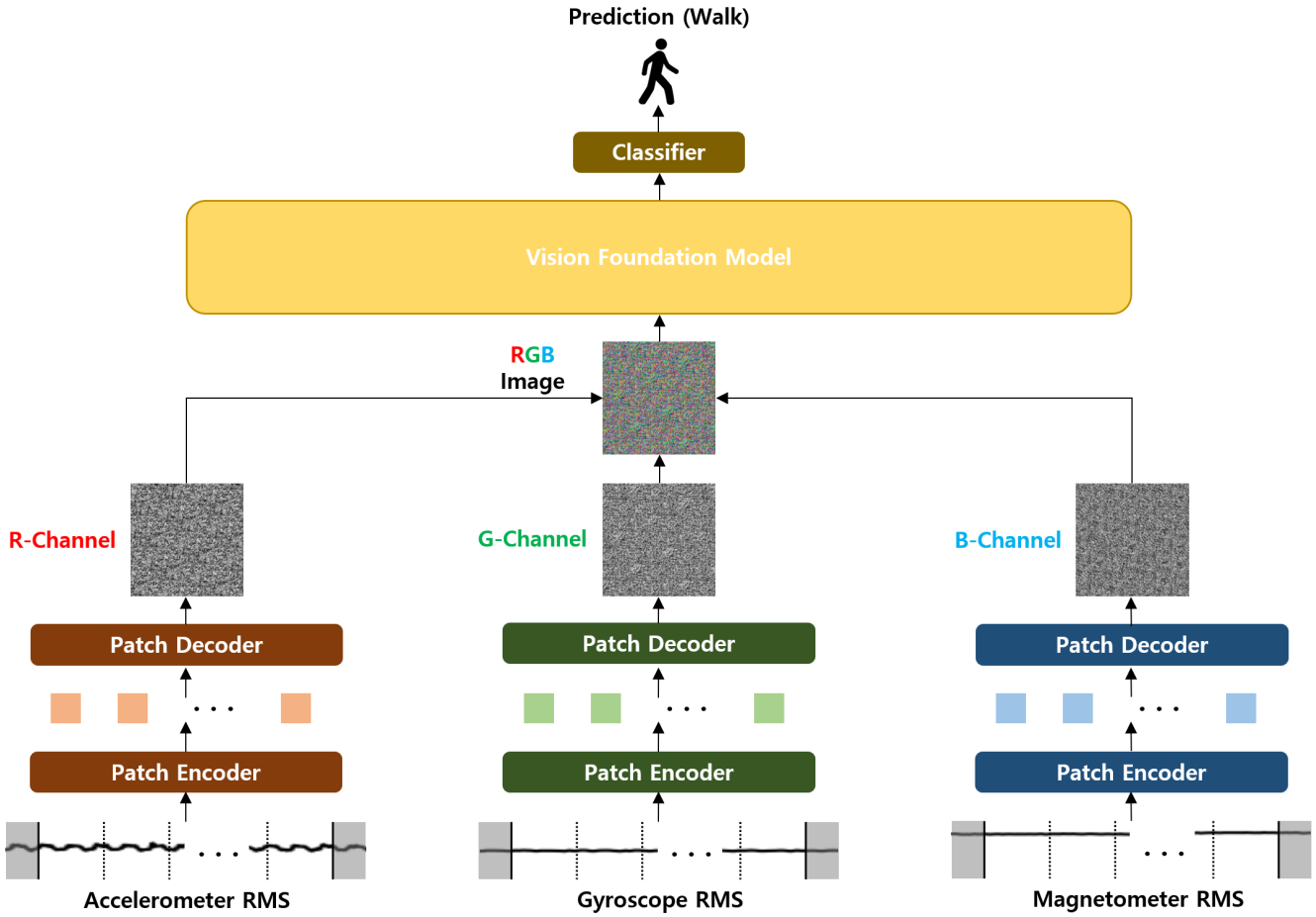


Figure 1: Proposed IMU2IMG pipeline. For each sensor, the Patch Encoder first encodes the input sensor RMS values into meaningful tokens. Then, the Patch Decoder converts the token sequences into a 2D image to represent a single channel image. The combined RGB image is then processed by the vision foundation model (ViT is adopted for our experiments) and classifier for the final prediction.

However, translating sensor data to each modality presents trade-offs. Text- and time-series-based methods can struggle with explicit interpretability and spatial structure, while image-based methods often require manual pre-processing steps that limit flexibility and scalability. Inspired by the strengths and limitations of both approaches, we propose IMU2IMG, a novel framework that introduces a sensor-to-image translation layer which automatically maps inertial data to image space. This image representation is then processed through a vision foundation model, such as a Vision Transformer (ViT), enabling rich feature extraction. Our method leverages the expressiveness of vision models while eliminating the need for handcrafted visual encoding, offering a scalable and interpretable solution for sensor-based activity recognition. The main contributions of our work can be summarized as follows:

- We propose IMU2IMG, a novel pipeline that enables an automated and explicit alignment between multimodal sensor data and visual representations by leveraging vision foundation models.
- We empirically demonstrate the superiority of IMU2IMG over strong baselines, including traditional machine learning, deep learning, and foundation model-based approaches.
- We provide visualizations and detailed analysis of the generated sensor images, offering insight into the interpretability and effectiveness of the proposed alignment process.

2 SHL recognition challenge 2025 Dataset

The SHL recognition challenge 2025 dataset is divided into three subsets: 59 days of training data, 6 days of validation data, and 28 days of test data. The training set was collected from User 1 with a smartphone placed at four body locations: bag, hand, hip, and torso. The validation and test sets were collected from Users 2 and 3. While the training and validation sets include all four smartphone locations, the hand position is excluded in the test set. With the previous observations [26] on the performance degradation when using hand location data, we only exploited data from other three locations. Therefore, the training set consists of 588,216 frames

(196,072 frames × 3 locations), the validation set contains 86,367 frames (28,789 frames × 3 locations), and the test set includes 92,726 frames in total.

Each data sample consists of a 5-second frame recorded at 100 Hz, resulting in 500 time steps per frame. For each time step, sensor readings are provided from three modalities (accelerometer, gyroscope, and magnetometer) each with three axes (X, Y, Z), yielding a 500×9 matrix. In order to minimize the efforts on data pre-processing, we kept the temporal dimension [17] and only applied a root mean square (RMS) transformation over the three axes of each sensor type. This results in a compact and more stable 500×3 representation while maintaining the underlying motion dynamics critical for locomotion classification.

Each time step in each frame is annotated with one of eight locomotion activities: Still, Walk, Run, Bike, Car, Bus, Train, and Subway. Although some training and validation frames contain dual labels due to transitions between activities, they take only a small portion (0.3% - 581 out of 196,072 in train set and 0.4% - 104 out of 28,789 in validation set) in the entire dataset. Therefore, we assign each frame a single dominant label to simplify the classification task and facilitate stable model training.

3 Model Design

Figure 1 illustrates a model design of IMU2IMG, which automatically transforms multimodal inertial signals into RGB images for interpretable locomotion classification, exploiting vision foundation models. Each sensor modality (accelerometer, gyroscope, and magnetometer) is independently encoded and decoded into single channel images corresponding to the R, G, and B channels respectively. These three images are then concatenated to a single RGB image and passed through a vision foundation model and classifier, where the final activity label is predicted.

3.1 Patch Encoder

The Patch Encoder encodes each modality (500×1) into meaningful tokens. To align with the patch size of the downstream vision model, we center-crop the input to 490×1 and divide it into 49 (representing 7×7 grid) non-overlapping patches with a window size of 10. Here, the Token Generator Unit from Mantis [7] is adopted, which provides a calibration through several techniques such as Multi-Scaled Scalar Encoder [14] and layer normalization [1]. The outputs of the Patch Encoder are token sequences in the shape of $49 \times d$, where d is an embedding dimension.

3.2 Patch Decoder

The Patch Decoder converts the token sequences into a 2D image representation for each modality. Given 49 tokens, each token is projected from d to 1024 dimensional space (i.e., 32×32) using a linear layer. Then the projected tokens can be reshaped into a single channel image of shape $(7\times32)\times(7\times32) = 224\times224$. Finally, a *tanh* activation is applied to ensure that pixel values lie in the range [-1, 1], matching the input distribution expected by the vision foundation model. This process is independently applied to each modality, resulting in three grayscale images. These are stacked along the channel dimension to form a final RGB image.

3.3 Vision Foundation Model

For the backbone, we adopt the Vision Transformer (ViT) architecture [6], a widely used vision foundation model pre-trained on large-scale image datasets. We use the ViT variant[1] with a patch size of 7×7 (resolution 32×32), which balances computational efficiency and representational capacity by processing fewer input tokens. This choice also aligns with the 7×7 patch layout used in our Patch Encoder. Furthermore, the selected ViT model expects image inputs within the range of [-1, 1]. Our use of *tanh* in the Patch Decoder ensures that the generated image values fall within the desired [-1, 1] range, facilitating a seamless integration with the pre-trained model.

3.4 Classifier

The pooler outputs of the vision foundation model are passed through a single linear layer for final 8 class classification. Cross-entropy loss is used for training the entire model.

4 Experiments

4.1 Settings

As summarized in previous section, 588,216 frames are used for training and 86,367 frames are used for validation. Hyperparameters used for training IMU2IMG are summarized in Table 1. The computational resources related to the experiments are as follows:

- CPU: Intel Core i9-9980XE
- RAM: 64GB
- GPU: NVIDIA RTX A6000
- Model Size: 340MB
- Train Time: 23 minutes/epoch
- Software: Python 3.9, Pytorch 1.12, Transformers, Scikit-learn, Matplotlib, etc.

Table 1: Parameter configuration

Parameter	Setting value
Input size	(500, 3)
Backbone Model	ViT
d	256
Optimizer	AdamW
Learning rate	0.0001
Weight decay	0.01
Epochs	10
Batch size	512

The performance is evaluated on macro-F1 and the model with the highest F1 score is chosen. Also, the baselines compared to the proposed algorithm are as follows:

- **Machine Learning Approaches (ML)**: In SHL recognition challenge 2024 [20], ML approach with carefully extracted hand crafted features (HCF) took the first rank in the challenge, making ML approaches strong baselines to compare. Random forest (RF) [2], LightGBM [11], XGBoost [3] with 99 hand crafted features [18] were used.

[1] https://huggingface.co/google/vit-base-patch32-224-in21k

- **Deep Learning based Approaches (DL)**: Instead of exploiting handcrafted features, we evaluate a lightweight three-layer convolutional neural network (CNN) which is runnable directly on the sensor data. To assess the impact of simple pre-processing, two variants of this CNN are trained. One is trained and evaluated on raw RMS data and the other on pre-processed (filtering and normalization) RMS data.
- **Foundation model based Approaches (F)**: IMU data can be easily considered as a multivariate time-series data. Here, recent time-series foundation model Mantis [7] is used. We fully fine-tune the model. Also, we compare training IMU2IMG with pre-trained knowledge of a vision foundation model. IMU2IMG (S) is trained from scratch and IMU2IMG (P) is trained starting from pre-trained ViT weights in order to experiment whether the vision knowledge actually help interpreting sensor data.

4.2 Quantitative Results

Table 2 summarizes the macro-F1 of the baselines and the proposed IMU2IMG on the validation set. IMU2IMG shows the best performance among strong baselines. And here are several detailed analysis. First, compared to ML approaches, DL and F based approaches give improved performance. As the users in train/validation sets are different, which imply domain difference, DL might learn rather more generalizable features than hand crafted features. Second, comparing pre-processing results, the ability of deep learning can be further boosted through careful calibration of sensor data. Finally, it is notable that exploiting foundation models is a powerful approach for sensor based activity classification. Furthermore, comparing IMU2IMG (S) and (P), it is obvious that the pre-trained visual knowledge help the model get better performance. Figure 2 summarizes the confusion matrix from IMU2IMG (P), which shows that the model still has most difficulty in classifying Car/Bus and Train/Subway as can be observed in previous studies [18, 19], while showing stable performance on other classes.

4.3 Qualitative Results

Figure 3 shows the qualitative results of generated image by IMU2IMG on the samples from validation set. In order to visualize, we first map the Patch Decoder outputs of [-1, 1] range to [0, 1] using a linear transformation by first multiplying 0.5 and adding 0.5 to each pixel value. This allows the proper rendering as standard RGB images. Generated images from each 8 classes are depicted. For each class, sensor RMS line plots, single channel images and concatenated RGB images are ordered from top to bottom. From left to right, accelerometer (R channel), gyroscope (G channel) and magnetometer (B channel) are ordered.

It can be observed that the sensor characteristics are translated into an image. For example, stable values from sensor data results in patch-wise patterns as can be observed in Still gyroscope. Also, periodical values from sensor data lead to diagonal patterns in the generated image as can be observed in Walk and Run accelerometer. And intense vibration in sensor data leads to a discernible lines between the patches as can be observed in Car and Bus accelerometer. Merging this, the total RGB image shows certain visual patterns according to its corresponding class. These observations indicate that IMU2IMG can successfully translate the sensor data into the language of image.

Table 2: Summary of the macro-F1 on the models for the experiments. Foundation model based approaches outperforms machine learning or deep learning based approaches, and the proposed IMU2IMG (P) shows the best performance among them.

Model	Approach	Input	F1
RF	ML	HCF	59.06
LightGBM	ML	HCF	59.85
XGBoost	ML	HCF	60.42
CNN	DL	RMS	61.45
CNN	DL	Pre-processed RMS	67.72
Mantis	F	RMS	73.66
IMU2IMG (S)	F	RMS	73.01
IMU2IMG (P)	F	RMS	**74.91**

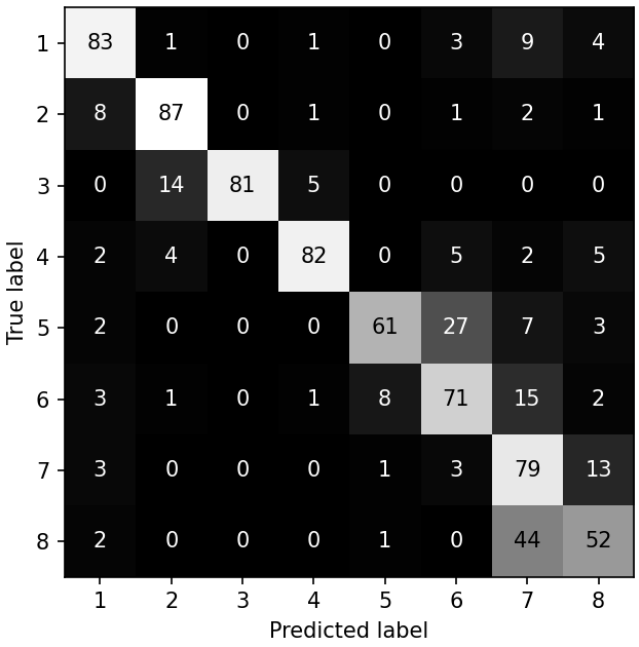


Figure 2: Confusion matrix from the proposed IMU2IMG (P)

4.4 Final Submission

To generate the final test set classifications and address variability across users for optimal recognition accuracy, we fine-tuned the weights of our pre-trained model IMU2IMG (P) using the validation set. As shown in prior work [22], this adaptation strategy can yield substantial gains. We employed this tuned model to produce the final classification outputs on the test set.

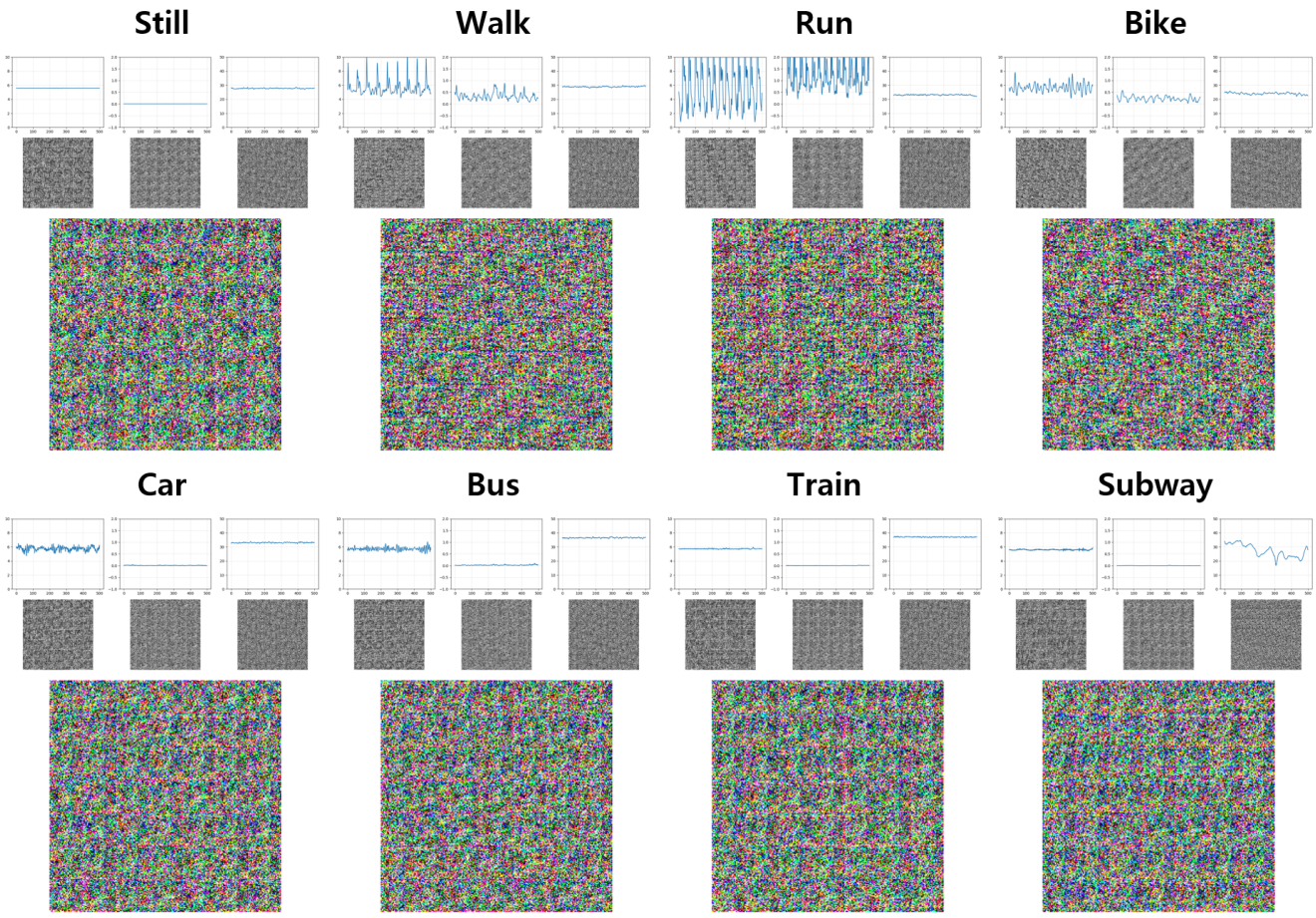


Figure 3: Qualitative results of IMU2IMG. For each class, sensor RMS line plots, generated single channel images (accelerometer - R, gyroscope - G, magnetometer - B from left to right) and combined images are depicted from top to bottom. The generated images inherit the original sensor characteristics, leading to certain visual patterns such as patch-wise or diagonal patterns in combined images according to their corresponding class.

5 Conclusion

In this work, we introduced IMU2IMG, a novel end-to-end pipeline that transforms multimodal inertial sensor signals into RGB images, enabling explicit alignment between time-series data and vision representations. By minimally pre-processing raw IMU data and leveraging a pre-trained Vision Transformer, IMU2IMG achieves superior macro-F1 performance on the SHL recognition challenge 2025 dataset when compared against strong machine learning, deep learning, and time-series foundation model baselines. Our quantitative experiments demonstrate that (1) learned representations consistently outperform handcrafted features in a user-independent setting, (2) modest pre-processing (filtering and normalization) further boosts performance even for end-to-end models, and (3) visual knowledge transferred from a vision foundation model yields clear gains over training from scratch. Beyond accuracy improvements, our qualitative analyses reveal distinct visual patterns for each locomotion class—validating that IMU2IMG not only predicts more robustly but also provides interpretable sensor-to-vision mappings. These insights confirm the value of explicit image-based representations for human activity recognition. We believe that the IMU2IMG paradigm opens a new avenue for combining the strengths of sensor processing and vision-based deep learning in a unified, interpretable framework. The recognition result for the testing dataset will be presented in the summary paper of the challenge [21].

Acknowledgments

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [25ZB1100, Core Technology Research for Self-Improving Integrated Artificial Intelligence System].

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
[2] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[3] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[4] Anusha David, Rajavel Ramadoss, Amutha Ramachandran, and Shoba Sivapatham. 2023. Activity recognition of stroke-affected people using wearable sensor. *ETRI Journal* 45, 6 (2023), 1079–1089.

[5] Satvik Dixit, Laurie M Heller, and Chris Donahue. 2024. Vision language models are few-shot audio spectrogram classifiers. *arXiv preprint arXiv:2411.12058* (2024).

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[7] Vasilii Feofanov, Songkang Wen, Marius Alonso, Romain Ilbert, Hongbo Guo, Malik Tiomoko, Lujia Pan, Jianfeng Zhang, and Ievgen Redko. 2025. Mantis: Lightweight calibrated foundation model for user-friendly time series classification. *arXiv preprint arXiv:2502.15637* (2025).

[8] Hristijan Gjoreski, Mathias Ciliberto, Lin Wang, Francisco Javier Ordonez Morales, Sami Mekki, Stefan Valentin, and Daniel Roggen. 2018. The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices. *IEEE Access* 6 (2018), 42592–42604.

[9] Sheikh Asif Imran, Mohammad Nur Hossain Khan, Subrata Biswas, and Bashima Islam. 2024. LLaSA: A Multimodal LLM for Human Activity Analysis Through Wearable and Smartphone Sensors. *arXiv preprint arXiv:2406.14498* (2024).

[10] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR)*.

[11] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[12] Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D Salim. 2024. Sensorllm: Aligning large language models with motion sensors for human activity recognition. *arXiv preprint arXiv:2410.10624* (2024).

[13] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. 2024. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 6555–6565.

[14] Chenguo Lin, Xumeng Wen, Wei Cao, Congrui Huang, Jiang Bian, Stephen Lin, and Zhirong Wu. 2023. Nutime: Numerically multi-scaled embedding for large-scale time-series pretraining. *arXiv preprint arXiv:2310.07402* (2023).

[15] Jingchao Ni, Ziming Zhao, ChengAo Shen, Hanghang Tong, Dongjin Song, Wei Cheng, Dongsheng Luo, and Haifeng Chen. 2025. Harnessing Vision Models for Time Series Analysis: A Survey. *arXiv preprint arXiv:2502.08869* (2025).

[16] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730* (2022).

[17] Se Won Oh, Hyuntae Jeong, Seungeun Chung, Jeong Mook Lim, and Kyoung Ju Noh. 2023. Multimodal sensor data fusion and ensemble modeling for human locomotion activity recognition. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*. 546–550.

[18] Se Won Oh, Hyuntae Jeong, Seungeun Chung, Jeong Mook Lim, and Kyoung Ju Noh. 2024. A hybrid algorithmic pipeline for robust recognition of human locomotion: Addressing missing sensor modalities. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 591–596.

[19] Jeroen Van Der Donckt, Jonas Van Der Donckt, and Sofie Van Hoecke. 2024. Magnitude and Rotation Invariant Detection of Transportation Modes with Missing Data Modalities. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 597–602.

[20] Lin Wang, Mathias Ciliberto, Hristijan Gjoreski, Paula Lago, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2024. Summary of SHL Challenge 2024: Motion Sensor-based Locomotion and Transportation Mode Recognition in Missing Data Scenarios. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 555–562.

[21] Lin Wang, Mathias Ciliberto, Hristijan Gjoreski, Paula Lago, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2025. Summary of SHL Challenge 2025: Locomotion and Transportation Mode Recognition Using Foundation Models. In *Proceedings of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2025 ACM International Symposium on Wearable Computers*.

[22] Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Paula Lago, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2020. Summary of the Sussex-Huawei locomotion-transportation recognition challenge 2020. In *Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers*. 351–358.

[23] Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Paula Lago, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2021. Three-year review of the 2018–2020 SHL challenge on transportation and locomotion mode recognition from mobile sensors. *Frontiers in Computer Science* 3 (2021), 713719.

[24] Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Paula Lago, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2023. Summary of SHL challenge 2023: Recognizing locomotion and transportation mode from GPS and motion sensors. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*. 575–585.

[25] Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Sami Mekki, Stefan Valentin, and Daniel Roggen. 2019. Enabling reproducible research in sensor-based transportation mode recognition with the Sussex-Huawei dataset. *IEEE Access* 7 (2019), 10870–10891.

[26] Yida Zhu, Haiyong Luo, Runze Chen, Fang Zhao, and Lin Su. 2020. DenseNetX and GRU for the sussex-huawei locomotion-transportation recognition challenge. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 373–377.