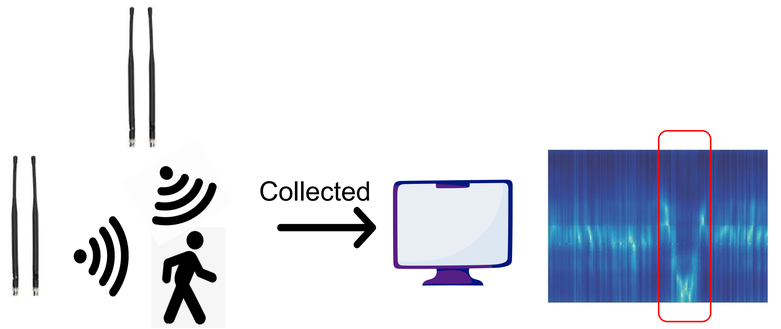
# Wavelet Transform Embedding and Masked Autoencoder Transformer for Gesture Detection

Haozhe Pang, Shucheng Yu
AISecLab, Department of Electrical & Computer Engineering, Stevens Institute of Technology
*hpang2@stevens.edu, shucheng.yu@stevens.edu*

## Background

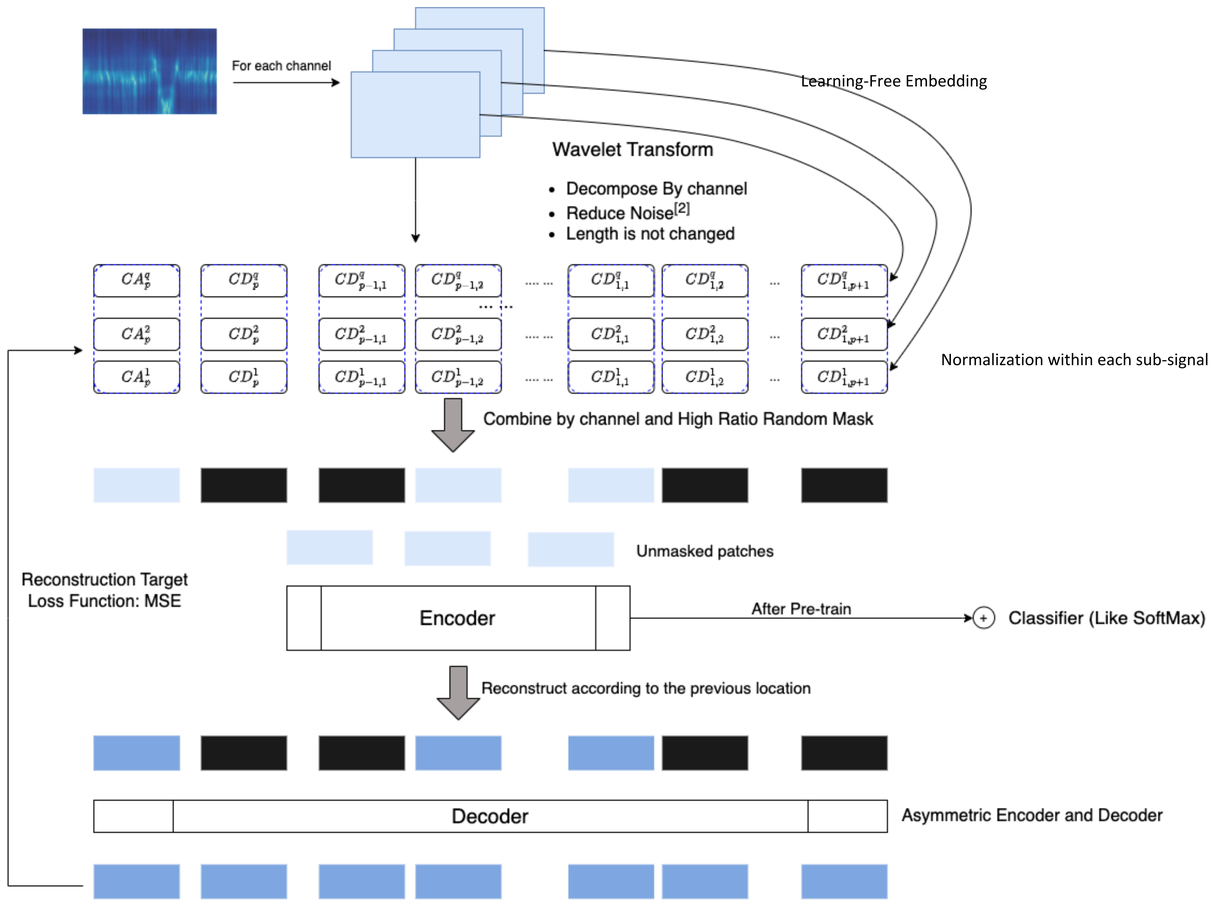
Gesture Detection by wireless physical layer data[2]




### Challenges

1. Noisy redundant background information
2. Long Series makes optimization of Transformer difficult.

## Model Structure




## Experimental Results

### Encoder Structure

| | Path Size | embedding dim | Layer Num | Head Num |
|---|---|---|---|---|
| ViT - small | 4 | 256 | 6 | 4 |
| ViT - standard | 8 | 512 | 6 | 6 |

The Decoder Structure has only two layers.

### Comparison of Accuracy

| Dataset | NTU-Fi HAR | NTU-Fi HumanID |
|---|---|---|
| OURS (Without MAE) | 98.48 | 99.64 |
| ViT | 93.75 | 76.84 |
| Other top 1 | 99.69(MLP) | 99.38(Bi-LSTM) |
| other top 2 | 99.69(Bi-LSTM) | 98.96(GRU) |

| Dataset | Widar |
|---|---|
| OURS (With MAE) | 84.07% |
| ViT | 67.22 |
| Other top 1 | 71.7 (ResNet) |
| other top 2 | 70.9(CNN 5) |

- Improve ViT results ranging from 5% to 10%
- SOTA in two Widar and NTU-Fi HumanID datasets or close to SOTA in NTU -HAR
- Other top models in different datasets have different structures

## Conclusion

- Pure Wavelet Transform as Embedding can help to reduce the noise in signal datasets.
- Masked Autoencoder structure is good at dealing with signal data with redundancy introduced from wavelets

[1]. He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[2]. Lang, Markus, et al. "Noise reduction using an undecimated discrete wavelet transform." *IEEE Signal Processing Letters* 3.1 (1996): 10-12.
[3]. Yang, Jianfei, et al. "SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing." *Patterns* 4.3 (2023).