FreqMAE: Frequency-Aware Masked Autoencoder for Multi-Modal IoT Sensing

Denizhan Kara
Tomoyoshi Kimura
University of Illinois
Urbana-Champaign, USA
kara4@illinois.edu

Shengzhong Liu*
Shanghai Jiao Tong University
Shanghai, China
shengzhong@sjtu.edu.cn

Jinyang Li
University of Illinois
Urbana-Champaign, USA
jinyang7@illinois.edu

Dongxin Liu
Meta Platforms, Inc
dxliu@meta.com

Tianshi Wang
Ruijie Wang
Yizhuo Chen
Yigong Hu
University of Illinois
Urbana-Champaign, USA
tianshi3@illinois.edu

Tarek Abdelzaher
University of Illinois
Urbana-Champaign, USA
zaher@illinois.edu

ABSTRACT

This paper presents FreqMAE, a novel self-supervised learning framework that synergizes masked autoencoding (MAE) with physics-informed insights to capture feature patterns in multi-modal IoT sensor data. FreqMAE enhances latent space representation of sensor data, reducing reliance on data labeling and improving accuracy for AI tasks. Differing from data augmentation-based methods like contrastive learning, FreqMAE's approach eliminates the need for handcrafted transformations. Adapting MAE for IoT sensing signals, we present three contributions from frequency domain insights: First, a Temporal-Shifting Transformer (TS-T) encoder that enables temporal interactions while distinguishing different frequency bands; Second, a factorized multi-modal fusion mechanism for leveraging cross-modal correlations and preserving unique modality features; Third, a hierarchically weighted loss function that emphasizes important frequency components and high Signal-to-Noise Ratio (SNR) samples. Comprehensive evaluations on two sensing applications validate FreqMAE's proficiency in reducing labeling needs and enhancing resilience against domain shifts.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Multimodal sensing, Self-supervised learning, Internet of Things

*Shengzhong Liu is the corresponding author.

WWW '24, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0171-9/24/05…$15.00
https://doi.org/10.1145/3589334.3645346

ACM Reference Format:
Denizhan Kara, Tomoyoshi Kimura, Shengzhong Liu, Jinyang Li, Dongxin Liu, Tianshi Wang, Ruijie Wang, Yizhuo Chen, Yigong Hu, and Tarek Abdelzaher. 2024. FreqMAE: Frequency-Aware Masked Autoencoder for Multi-Modal IoT Sensing. In *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3589334.3645346

1 INTRODUCTION

The paper advances the state of the art in self-supervised learning from time-series sensor data. Self-supervised learning aims to transform *unlabeled* input data into a latent space that captures data semantics, simplifying extensive downstream tasks. Two popular methods to achieve this are *contrastive learning* and *masked autoencoders*. Contrastive learning uses data augmentations, like image rotations, to train networks on identifying semantically similar items. On the other hand, MAEs don't require augmentations; they conceal parts of the input for the network to reconstruct, leveraging its grasp of higher-level semantics for precise reconstruction and obscured trait comprehension. This approach enhances latent space representation of object attributes, streamlining inference task training. As a label-free method, MAEs simplify training and boost downstream AI task accuracy with fewer samples [24].

Although MAEs excelled in vision and natural language domains [16, 28, 52], they lag behind contrastive methods in processing time-series sensor data [50]. We find that appropriately integrating insights from a conventional signal processing perspective can effectively simplify the optimization space and boost the performance of MAEs. Therefore, we introduce FreqMAE, a specialized MAE for multi-modal IoT sensing. It integrates three distinct frequency-aware insights applicable across sensing tasks, which set FreqMAE apart from standard MAEs, tailoring it for time-frequency analysis.

First, we introduce a frequency-aware Transformer, the *Temporal-Shifting Transformer (TS-T)*, tailored for sensor spectrogram encoding. Traditional Transformers [45] and Vision Transformer (ViT) encoders [19] are less effective with spectrograms due to their global attention, missing spectrogram-specific traits like frequency

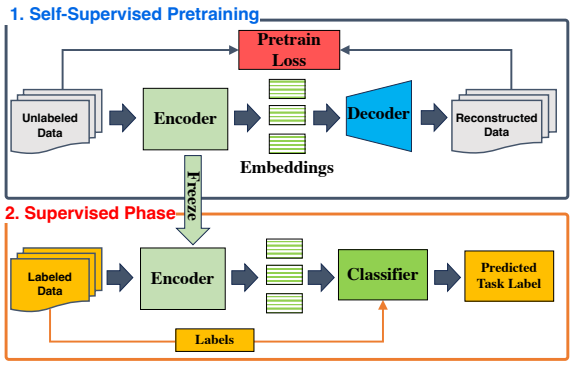


Figure 1: Masked Autoencoder (MAE) Workflow

translation and scaling invariance, and temporal shifts in amplitudes and frequencies from physical non-stationarity [32]. TS-T addresses these challenges with a local attention mechanism in the frequency domain for focused processing of localized short-time Fourier windows and by adapting attention in the temporal domain to account for frequency shifts, thus *preserving the spectrogram's spectral structure while accommodating shifting frequency behaviors.*

Second, we introduce a factorized data fusion mechanism emphasizing cross-modal correlations and modality-specific features. This approach recognizes that synchronized modalities share common information and provide unique, complementary insights [57]. By applying single masking to each modality's input, we create two post-encoding feature spaces: *(i)* a modality-specific *private space* for self-reconstruction, and *(ii)* a *shared space* capturing cross-modal information, enabling reconstruction of one modality's input with shared embeddings from others. Two specialized lightweight decoders facilitate this, *ensuring minimal overhead during inference.*

Third, we propose a hierarchically weighted loss function that prioritizes important frequency regions and high Signal-to-Noise Ratio (SNR) samples. To illustrate the benefits of weighting, we consider IoT applications, where crucial information is predominantly found in the low-frequency components, whereas high-frequency sections are mostly noise [24]. By focusing on accurately reconstructing these low-frequency areas and emphasizing high SNR samples with significant energy content, we improve the effectiveness of representation learning. For instance, in vehicle classification via audio and seismic sensors, measurements captured when vehicles are nearby are especially informative [51].

This work is driven by the rise in computational power of embedded devices and the robust modeling of deep neural networks (DNN), advancing the Artificial Intelligence of Things (AIoT) domain in areas like activity detection, vehicle tracking, and smart healthcare [3, 18, 53]. Most existing work [3, 10, 54] depends on *supervised* learning, demanding large volumes of labeled data, which is challenging to obtain for time-series in controlled environments, unlike images and text [39]. Moreover, *DNN models trained on data from limited environments often exhibit sensitivity to unforeseen changes in the actual deployment setting* [48].

By utilizing *self-supervised learning*, we train the encoder *without the need for labeled data.* Subsequently, we perform supervised fine-tuning using a limited number of data labels to train the downstream inference task. This approach is highly label-efficient and yields pretrained data encoders with enhanced robustness against environmental variations. Unlike contrastive learning frameworks [7, 11] which heavily rely on human intuition to create label-invariant transformations, FreqMAE only employs simple random masking as the preprocessing step. It also integrates physical signal knowledge that is applicable across various sensing applications as improvements, resulting in *higher automaticity and extensibility.*

We extensively evaluate FreqMAE using four datasets, demonstrating its superior performance over existing approaches in various sensing applications. The results highlight the exceptional potential of the self-supervised FreqMAE framework as a step towards building foundation models specially tailored for sensing streams and time series data. Beyond the dataset evaluations, we use a real-world case study to demonstrate the robustness of FreqMAE. One standout feature is *its exceptional performance in the face of environmental variations.* FreqMAE shows unparalleled capability in managing dynamic, real-life scenarios, affirming its utility for representing information from dynamic sensing streams.

The rest of this paper is organized as follows: Section 2 covers background information, Section 3 details FreqMAE's design, Section 4 presents experiments and findings, Section 5 reviews related work, and Section 6 discusses limitations and concludes the study.

2 PRELIMINARIES

This section outlines the foundational concepts of self-supervised learning and the inspirations behind FreqMAE's design.

2.1 Masked Autoencoders

Compared to the prevalent contrastive learning paradigm for IoT data [9, 42, 47], reliant on domain-specific augmentations, we introduce a fully automated, *augmentation-free* self-supervised MAE method [16] that *significantly reduces labeled data dependence* across sensing contexts. Figure 1 illustrates the MAE setup, including an encoder, decoder, and classifier, focusing on two-phase training: self-supervised pretraining followed by supervised fine-tuning.

Pretraining leverages unlabeled data to derive versatile representations for various tasks, by employing random masking on segments of unlabeled spectrograms. The encoder transforms masked data into low-dimensional embeddings, which the decoder uses to reconstruct masked areas. The training aims to minimize the discrepancy between the decoded results and the original data in masked areas. To encourage the model to capture overarching semantics over low-level interpolations, we apply masking at the granularity of frequency patches with a high masking ratio.

In the fine-tuning stage, we discard the decoder and directly connect the encoder to a lightweight classifier (*i.e.,* one fully connected layer). During this phase, the pretrained encoder parameters remain fixed, and the linear classifier is updated using the representations generated by FreqMAE, which are based on limited labels specific to the downstream task. This approach offers two advantages: *(i) the need for fewer labels for convergence [21]* and *(ii) faster training.*

2.2 Characteristics of IoT Sensing Data

IoT sensing data exhibit unique characteristics that set them apart from other contexts. Following common practices [22, 54], we use spectrogram data after a short-time Fourier transform (STFT) on

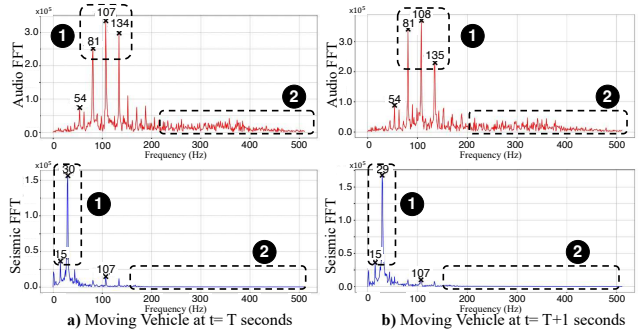


Figure 2: Audio FFT signatures for a moving vehicle. ❶ The presence of characteristic peaks in localized regions needs local harmonic associations and shift-sensitive representations. ❷ Higher frequency regions mostly contain noise.

the raw input, as the modality input. We carefully examine the fundamental properties of such spectrograms to guide the design of FreqMAE. Figure 2 presents two sensor (audio and seismic) readings from two consecutive time windows for a moving vehicle, collected as it passes by the sensors. Several observations are highlighted.

2.2.1 ***No Scale and Shift Invariance.*** While vanilla MAE employs global attention due to visual objects' invariance to translation or scaling, this assumption doesn't hold for IoT data. Here, the positioning and scaling of frequency content significantly influence semantics. Thus, global self-attention might be less effective when time-frequency information is predominantly local. For instance, only linking harmonic patches vertically through frequency (see ❶ in Figure 2) may be suboptimal due to recurring harmonics while associating the shifted harmonics horizontally through time can yield more comprehensive insights into non-stationary patterns.

2.2.2 ***Multi-Modal Fusion.*** IoT data stems from various sensors, such as accelerometers, gyroscopes, and magnetometers, each providing a distinct perspective into the observed event. By fusing information from multiple sensors, a richer understanding and increased system efficacy can be achieved [6]. Therefore, aligning with the emerging trend on multi-modal fusion [5, 26, 37, 46, 54], an effective SSL framework should support data fusion across diverse modalities and feature generalization across various sensors.

2.2.3 ***Differentiated Frequency and Sample Importance.*** Regarding the reconstruction objective in MAE, we observed that differentiated importance should be imposed locally among different frequency bands and globally among different samples. First, in physical sensing tasks, it is well-known that valuable information tends to be found in the low-frequency sections of the spectrogram [24]. Conversely, the very high-frequency sections often consist mostly of noise (*e.g.,* ❷ in Figure 2). Second, due to external factors and the nature of physical sensing data, some samples are more important than others regarding the detection of the observed phenomenon. For instance, samples with higher SNR provide more useful information than lower SNR samples that include noise.

3 FRAMEWORK

In this section, we introduce FreqMAE and its three novel components (motivated by the aforementioned characteristics).

3.1 Overview

FreqMAE generates embeddings for unlabeled time series data from multiple sensory modalities. With P modalities $\mathcal{M} = M_1, M_2, \ldots, M_P$ and N *unlabeled* training samples $\mathcal{X} = \mathbf{x}_1, \mathbf{x}2, \ldots, \mathbf{x}N$, where each $\mathbf{x}ij$ is input from modality M_j of sample $\mathbf{x}i$, the goal is $\mathbf{h}ij = E_j(\mathbf{x}ij)$, using encoders $\mathcal{E} = E_1, E_2, \ldots, E_P$ for embedding generation. Inputs are transformed into spectrograms via STFT for time-frequency analysis. Pretraining, as shown in Figure 3, segments spectrograms into patches for linear projection embeddings, excluding positional embeddings as supported by [27] (see Appendix D.5 for details).

We then randomly mask a significant portion of spectrogram patches, crucial for efficient self-supervised pretraining [16], following a Bernoulli process with each patch having a probability p of being masked, termed the *masking ratio*. Given the two-dimensional nature of spectrograms for time-frequency analysis, we compared unstructured and structured masking, finding unstructured random masking superior for pretraining (details in Appendix D.2). Similar to image processing [16], a high masking rate of 70% to 80% optimally supports representation learning.

FreqMAE utilizes Temporal-Shifting (TS) Transformer encoders for each modality, a transformer design incorporating localized attention with a spectrogram-compatible shifting mechanism inspired by the SwinTransformer[27]. The encoder-generated embeddings are merged into private and shared modality representations through the factorized fusion mechanism. Private embeddings capture modality-specific information, while shared embeddings encapsulate information common to all modalities. This approach facilitates the learning of cross-modality representations and the association of diverse information available across modalities.

Decoders, also constructed from TS-Transformers, utilize modality embeddings to reconstruct the pre-masking input. Different from prior work [16, 29], FreqMAE employs a weighted reconstruction objective, leveraging preliminary signal knowledge to prioritize important patches and samples during the pretraining. Specifically, the objective prioritizes lower-frequency areas rich in information and samples with higher Signal-to-Noise Ratios (SNRs), over their higher-frequency, noisier counterparts during pretraining.

3.2 Temporal-Shifting (TS) Transformer

The vanilla MAE [16] uses global self-attention in Transformers, ideal for visual contexts where object semantics are spatially and scale-independent. However, for time-frequency spectrograms, the importance of positions, scales, and shifts critically affects signal semantics [34], highlighting a misfit with the original design for our domain. Figure 2-(a) reveals that while lower frequency band harmonics can predict higher frequency bands vertically, they're less adept at horizontal predictions in the time domain. This is due to higher frequency harmonics shifting gradually from inherent non-stationarity in physical signals. As seen between Figure 2-(a) and (b), this shift complicates predictions using lower frequency bands. The sequence and positioning of spectrogram patches are critical for accurate signal interpretation, indicating that global attention might not be the most effective approach for spectrograms where time-frequency details are predominantly local with gradual shifts.

Inspired by SwinTransformer [27], TS-Transformer incorporates two fundamental insights: *(i)* the local nature of time-frequency

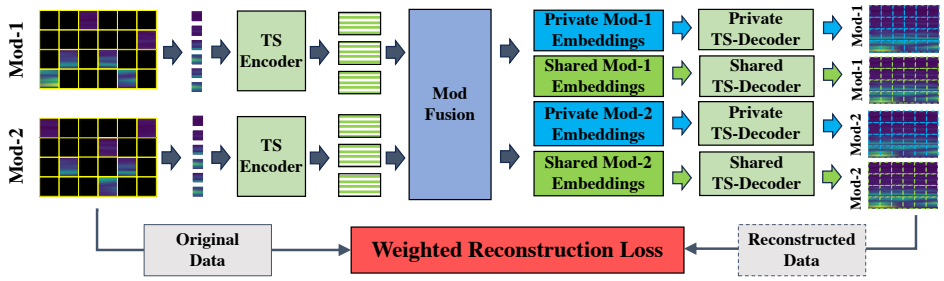


Figure 3: FreqMAE design components with self-supervised pretraining workflow.

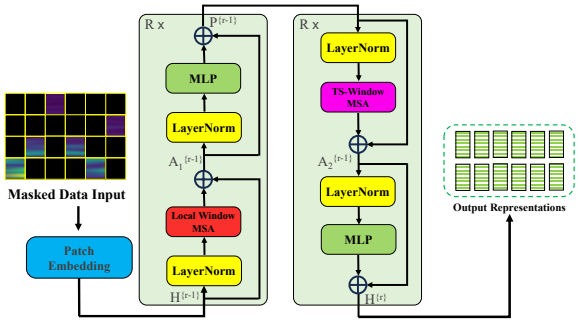


Figure 4: TS-Transformer blocks. Both the Local Window MSA and the TS-Window MSA are multi-head self-attention with local and temporally shifted windows.

components in spectrograms requiring association between local harmonics, and *(ii)* the necessity to represent shifting frequency components due to non-stationarity. Localized attention is crucial for *limited invariance*, as (slightly) shifted frequencies from non-stationarity might represent the same physical phenomenon at different times. Thus, effective learning should capture these shifts while maintaining the frequency components' position and scale.

Figure 4 illustrates the TS-Transformer design. The masked spectrograms are fed into the patch embedding layer, a convolutional layer that produces a vector embedding from the unmasked patch signals with a dimension of H_{dim}. Masked spectrograms enter a patch embedding (convolutional) layer, creating H_{dim}-dimensional embeddings from unmasked patches. The TS-Transformer has two sequential transformer blocks, processing H-dimensional embeddings over R iterations to produce representations of the same dimensionality, described by:

$$\begin{aligned}
\mathbf{A}_1^{\{r-1\}} &= \text{WMSA}\left(\text{LayerNorm}\left(\mathbf{H}^{\{r-1\}}\right)\right) + \mathbf{H}^{\{r-1\}}, \\
\mathbf{P}^{\{r-1\}} &= \text{MLP}\left(\text{LayerNorm}\left(\mathbf{A}_1^{\{r-1\}}\right)\right) + \mathbf{A}_1^{\{r-1\}}, \\
\mathbf{A}_2^{\{r-1\}} &= \text{TS-WMSA}\left(\text{LayerNorm}\left(\mathbf{P}^{\{r-1\}}\right)\right) + \mathbf{P}^{\{r-1\}}, \\
\mathbf{H}^{\{r\}} &= \text{MLP}\left(\text{LayerNorm}\left(\mathbf{A}_2^{\{r-1\}}\right)\right) + \mathbf{A}_2^{\{r-1\}},
\end{aligned}$$

where LayerNorm(·) is the layer normalization [2]. The MLP(·) comprises two fully-connected layers. Both WMSA(·) and TS-WMSA(·) are multi-head self-attention modules [45] configured with regular (Local Window MSA) and temporally shifted window (TS-Window MSA) attention settings and A attention heads, respectively.

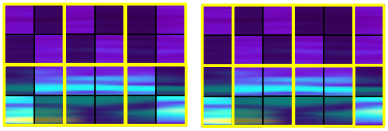


Figure 5: Local attention and temporal shifted windows.

To capture local frequency structures, we use a *local attention mechanism* targeting short frequency bands, organizing spectrogram patches into local windows in spatial dimension and applying self-attention within these windows to identify relationships among local frequencies. Furthermore, to tackle temporal non-stationarity, we introduce a *temporal shifting* procedure that associates harmonics with their temporally shifted counterparts. Figure 5 demonstrates local window attention regions and how temporally shifted windows are partitioned. Local windows move 50% horizontally (*i.e.,* in the time dimension) across layers for cross-window interactions, with no frequency dimension shifts due to the unique physical meanings of frequency bands. This approach allows for focused attention on specific frequency bands and recognizes temporal correlations between shifted harmonics, as shown in Figure 2.

3.3 Factorized Modality Fusion

Multi-modal fusion leverages the diverse and rich information provided by different modalities, each offering a unique perspective on the observed phenomenon. To effectively extract representations from multi-modal data, we emphasize the necessity for *a complementary modality fusion* approach. On one hand, it's vital to *extract shared information between collaborating modalities* to understand their semantic relationships. On the other hand, these modalities mutually enrich each other by offering unique, private information that complements the data from other modalities. A practical framework should be capable of *extracting both shared and unique patterns across modalities to enhance generalizability.*

To achieve this, we introduce a factorized fusion mechanism within FreqMAE, encompassing both modality self-reconstruction and cross-modality reconstruction. Figure 6 provides a visual explanation of this approach. After fusion, each modality's embedding space is partitioned into two subsets: private and shared spaces. Private embeddings come directly from the encoding of the current modality. Conversely, shared embeddings are generated by fusing the embeddings of other modalities through a shared fusion layer, comprising two feed-forward layers. Both private and shared embeddings are then fed into separate decoders to reconstruct the current modality. This reconstruction uses the same weighted loss

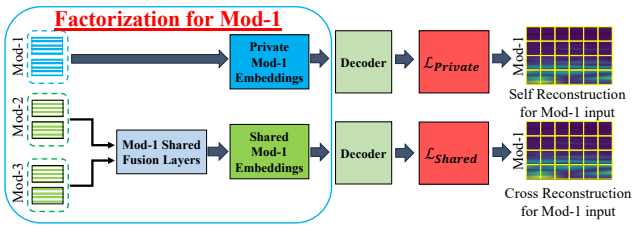


Figure 6: Factorized Fusion in FreqMAE.

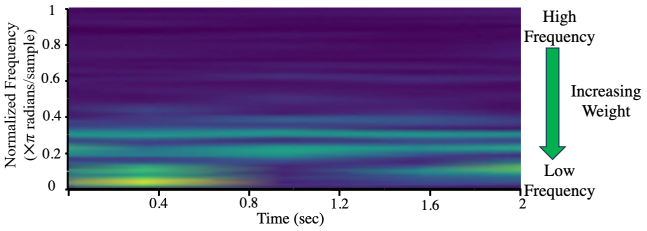


Figure 7: Weighted Mean Square Error weights.

function, resulting in two distinct reconstruction losses: $\mathcal{L}_{\text{private}}$ and $\mathcal{L}_{\text{shared}}$. The overall pretraining loss is calculated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{private}} + \gamma \mathcal{L}_{\text{shared}} \tag{1}$$

where γ is the hyperparameter that controls the weight between two loss components. Because of the asymmetric structure between the deep encoders and lightweight decoders in MAE, we will show later in the evaluation that the introduced extra decoder only incurs negligible computation overhead. Moreover, decoders are discarded after the pretraining stage, removing overhead at inference time.

The proposed factorized fusion mechanism, unique to FreqMAE, ensures encoded embeddings carry semantical information for self and peer modality input reconstruction. Experiments show that a higher γ value, favoring shared embeddings, suits datasets with many modalities (e.g., IMU data with 3+ modalities), while a lower γ benefits tasks with fewer, distinct modalities (e.g., audio-seismic pairs, with audio offering rich semantics). The impact of γ is further explored in Appendix D.3. Hence, our fusion scheme is *flexible to accommodate diverse sensor combinations and distributions, with adjustable contributions from private and shared modality information.*

3.4 Importance Weighting Loss Function

This module is motivated by two key insights. First, we should emphasize informative content within the signal samples using physical primitives that are common among the sensory data. For instance, in most physical sensing tasks, such as vehicle classification (see Figure 2) and human activity recognition, where the frequency content of most activities lie between 0 and 20 Hz [1], most of the useful information is located in the lower frequency parts of the spectrogram, while high-frequency parts are usually noise [24]. Second, an efficient pretraining objective should emphasize the signal samples containing richer information for the observed physical phenomenon without using labels. Since pretraining is performed with a large amount of unlabeled data, the inherent "class imbalance" is even more evident in such large datasets, where most of the measurements do not contain any activity or context. Devoting excessive attention to reconstructing such samples may cause the model to struggle in capturing meaningful feature patterns.

The vanilla MAE utilizes Mean-Squared Error (MSE) for reconstructing the masked patches during pretraining, defined as:

$$\text{MSE} = \frac{1}{T \times F} \sum_{t=1}^{T} \sum_{f=1}^{F} \left(\mathbf{X}(f,t) - \hat{\mathbf{X}}(f,t) \right)^2, \tag{2}$$

where $\mathbf{X}$ and $\hat{\mathbf{X}}$ refer to the original and reconstructed spectrograms and $T \times F$ represents the time-frequency dimensionality of the spectrogram. Although it is suitable for images where no preliminary knowledge about object location is known, MSE doesn't perform optimally with sensing spectrogram input. To address this, we leverage our initial insight on prioritizing lower frequency regions, and thus, define the Weighted Mean Squared Error (WMSE) as follows:

$$\text{WMSE} = \frac{1}{T \times F} \sum_{t=1}^{T} \sum_{f=1}^{F} \mathbf{W}_f \left(\mathbf{X}(f,t) - \hat{\mathbf{X}}(f,t) \right)^2, \tag{3}$$

where W_f refers to the weights of the corresponding spectrogram frequencies. As shown in Figure 7, the weight for the highest frequency is minimum and the weights linearly increase as the frequency decrease. In particular, we set

$$\mathbf{W}_F = \mathbf{W}_{\min}, \quad \mathbf{W}_1 = \mathbf{W}_{\min},$$
$$\mathbf{W}_f = \mathbf{W}_{\max} - \frac{(f-1)(\mathbf{W}_{\max} - \mathbf{W}_{\min})}{F-1}, \tag{4}$$

where we set $\mathbf{W}_{\min} = 0$ and $\mathbf{W}_{\max} = 1$ in our experiments.

Besides, in order to prioritize informative samples with movement over background samples, we calculate the mean cumulative energy of the sample across modalities M:

$$\text{E} = \frac{1}{M \times T \times F} \sum_{m=1}^{M} \sum_{t=1}^{T} \sum_{f=1}^{F} \mathbf{X}(f,t)^2, \tag{5}$$

where M is the number of modalities. Note that using the mean cumulative energy across modalities, as opposed to the energies of individual modalities, helps avoid bias towards modalities with typically higher energy content. Since our aim is to comparatively differentiate across samples, the mean energy across modalities provides fair supervision for the training objective. Inspired by the commonly used peak-signal-to-noise ratio (PSNR) metric [17] for comparing image reconstruction quality [38], we define the overall training objective of FreqMAE (in dB) as:

$$WPSNR = 10 \cdot \log \left(\frac{E^{\lambda}}{WMSE} \right), \tag{6}$$

where λ is the hyperparameter, ranging from 0 to 1, that controls the scale of the energy component. We utilize the negative of WPSNR as the pretraining loss for FreqMAE. Since MSE fundamentally represents the "mean residual energy", both the logarithm in the numerator and the denominator are in the same unit.

The WPSNR objective guides pretraining to prioritize high-fidelity reconstruction of high-energy (low WMSE) samples. In summary, the WPSNR enables the model to emphasize essential frequency components within a sample while comparatively assessing the semantic importance of different samples for efficient representation.

Table 1: Dataset Summary

Dataset	# Classes	Modalities[1]	# Samples	Application
MOD	7	MP, S	39,609	VC
ACIDS	9	MP, S	27,597	VC
RealWorld-HAR	8	A, G, M, L	12,887	HAR
PAMAP2	18	A, G, M	9,611	HAR

4 EVALUATION

Here, we present our experimental setups and extensive evaluations to demonstrate FreqMAE's performance, resiliency, and feasibility.

4.1 Experimental Setup

4.1.1 Datasets and Preprocessing. We evaluate FreqMAE using four datasets from prior work [8, 33, 46, 54] across two application domains, *(i)* Vehicle Classification (VC) and *(ii)* Human Activity Recognition (HAR). Datasets feature diverse sensors, classes, and environments (see Table 1): **(1) MOD** is a self-collected dataset with microphone arrays (acoustic) and geophones (seismic) for classifying six vehicle types and human walking. **(2) ACIDS** is from the US Army Research Lab, focusing on acoustic and seismic identification with 9 vehicle types across three terrains. **(3) RealWorld-HAR** is a public dataset with accelerometer, gyroscope, magnetometer, and light sensors for detecting eight physical activities collected from 15 participants. **(4) PAMAP2** is another public dataset capturing 18 physical human activities via accelerometer, gyroscope, and magnetometers. More dataset details are given in Appendix A.

In preprocessing, we divide time-series data into evenly sized windows and apply the Fourier transform to each to generate spectrograms, with lengths based on data characteristics. FreqMAE processes these spectrograms for feature representation. Note that FreqMAE *can handle different sampling rates among modalities* since they have separate feature encoders. During training, datasets are split into training, validation, and test sets in an 8:1:1 ratio, leaving sessions out for a realistic split. Training data is further split into different ratios of available labels (100%, 10%, 1%)—the **label ratio**—to *show FreqMAE's effectiveness under label scarcity.* For additional preprocessing and training details, see Appendix B.

4.1.2 Baselines. We compare FreqMAE with 10 baselines: a supervised benchmark, five self-supervised frameworks (MAE [16], SimCLR [7], CAV-MAE [14], AudioMAE [19], LIMU-BERT [53]), two modality-matching contrastive baselines (CMC [42], Cosmo [30]), and two state-of-the-art (SOTA) contrastive time series frameworks (TS-TCC [11], TS2Vec [55]). Detailed introductions of baselines are in Appendix C. We append a linear classification layer for downstream tasks during fine-tuning. For the contrastive settings, we keep the backbone encoders the same as FreqMAE for a fair comparison. A set of eight time-domain augmentations, and a frequency domain augmentation is used from common practices [20, 25, 41] for contrastive baselines (details in Appendix B). Note that contrastive frameworks' performance depends on the used augmentations, while *FreqMAE eliminates dependency on used augmentations and is generalizable* (analysis at Section 4.2.1).

[1]MP=microphone, S=seismic, A=accelerometer, G=gyroscope, L=light, M=magnetometer.

4.2 Evaluation Results

4.2.1 Overall Performances. Table 2 compares the performance of FreqMAE with other baselines using a 100% label ratio. All evaluations use fixed encoders and a linear layer on top of pretrained sample features for a fair assessment of representational quality. The results show FreqMAE surpasses all baselines by at least 6.6 % and 8 % in average accuracy and F1, affirming its effectiveness. While supervised training slightly outperforms FreqMAE on the PAMAP2 task with full labels, we suspect this is due to PAMAP2 including human activities with shorter bandwidth (similar to RealWorld-HAR), therefore self-supervised representations being less detailed to outperform supervised training with full labels. Moreover, supervised training suffers from label shortage and degrades significantly with fewer labels (see Section 4.2.2). Thus, FreqMAE's overall superior performance indicates the high quality of its extracted features. The primary competitors of FreqMAE, TS-TCC and CMC frameworks, are heavily dependent on augmentation design and often underperform with fewer augmentations [49]. Figure 9 demonstrates their performance drop when using only six or three out of nine random augmentations. Further evaluations of FreqMAE on downstream tasks and representation quality are in Appendix D.

4.2.2 Varying Labeling Ratio. In this experiment, we evaluate the performances of baselines and FreqMAE with different labeling rates, varying from 1% to 100%. Figure 8 presents the comparison results with all datasets. Higher labeling rates tend to yield improved accuracies across most models. However, FreqMAE consistently outperforms the baseline models in all scenarios. Notably, there are consistent performance gaps between FreqMAE and other models toward lower labeling rates. We note that only TS-TCC consistently competes with FreqMAE. This is because TS-TCC efficiently leverages the temporally correlated nature of sensing signals through temporal contrasting views. However, TS-TCC also relies on a rich set of augmentations and experiences performance degradation with fewer augmentations, as shown in Figure 9. This suggests that FreqMAE *effectively learns general representations from unlabeled data, and thus a linear classifier is enough to achieve higher accuracy.*

4.2.3 Ablation Study. Table 3 presents an ablation study using ACIDS for VC and PAMAP2 for HAR tasks to assess the contribution of each design component. We studied four FreqMAE variants: **w/o Weighted Loss** using standard MSE for reconstruction (Equation 2), **w/o Energy Scaling** applying only WMSE loss without energy scaling (Equation 3), **w/o TS-T** employing Swin Transformer instead of TS-Transformer, and **w/o Fusion** without shared fusion and doing separate modality reconstruction during training.

First, the contribution of all components is evident in both tasks. Comparatively, the fusion component and weighted loss scheme are more helpful in improving task performance, which shows learning relations across modalities can reveal underlying patterns in the frequency domain. Such patterns might be hard to capture without considering modality relations, as different sensor modalities often provide complementary information [31]. Second, the focus of the weighted loss objective on prioritizing informative content within and across samples offers extra self-supervision for pretraining. Finally, the absence of TS-T configuration has a larger impact on the PAMAP2 task than on ACIDS. We suspect this difference is due to

Table 2: Finetune results with 100 % labels. We mark the best and second best values.

	ACIDS		MOD		PAMAP2		RealWorld-HAR		Average	
Metric	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Supervised	0.9137	0.7770	0.8948	0.8931	**0.8612**	**0.8384**	0.9313	0.9278	0.9002	0.8591
CMC	0.7813	0.6216	0.9049	0.9023	0.7571	0.7223	0.8211	0.8384	0.8161	0.7712
Cosmo	0.8776	0.7298	0.3228	0.3241	0.7910	0.7469	0.8529	0.7968	0.7111	0.6494
SimCLR	0.5658	0.4879	0.7535	0.7434	0.7346	0.6635	0.7830	0.7181	0.7092	0.6532
TS2Vec	0.6539	0.4913	0.7649	0.7632	0.5706	0.4942	0.6117	0.5002	0.6503	0.5622
TS-TCC	0.9046	0.7651	0.7709	0.7744	0.7871	0.7107	0.8684	0.8227	0.8328	0.7682
Vanilla MAE	0.8872	0.7604	0.9015	0.8460	0.7382	0.6999	0.8638	0.8700	0.8477	0.7941
LIMU-BERT	0.5023	0.3171	0.2157	0.1236	0.7847	0.7612	0.7946	0.7261	0.5743	0.4820
CAV-MAE	0.7995	0.6711	0.5184	0.4941	0.7697	0.7351	0.9215	0.9267	0.7523	0.7068
AudioMAE	0.7845	0.6120	0.7274	0.7249	0.7808	0.7478	0.8163	0.7437	0.7773	0.7071
FreqMAE	**0.9365**	**0.7919**	**0.9524**	**0.9514**	0.8420	0.8205	**0.9250**	**0.9327**	**0.9140**	**0.8741**

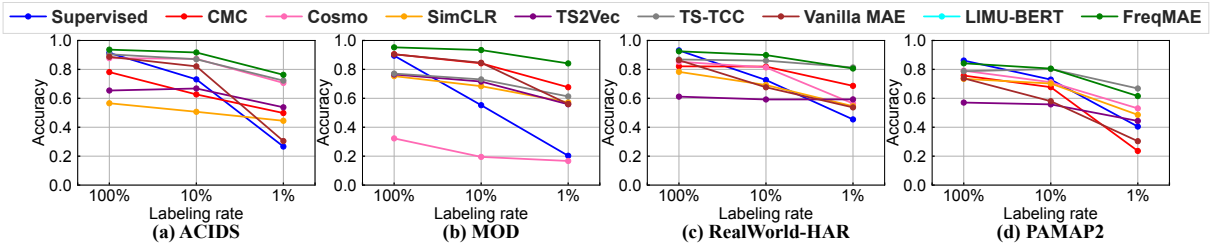


Figure 8: Accuracy comparison of FreqMAE with different labeling rates.

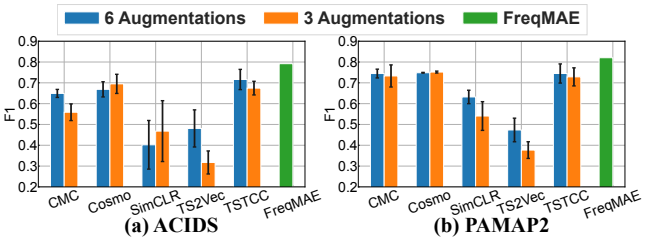


Figure 9: Sensitivity to Data Augmentations.

Table 3: Ablation Study on FreqMAE components.

Dataset	ACIDS		PAMAP2	
Metric	Acc	F1	Acc	F1
w/o Weighted Loss	0.9068	0.7674	0.8249	0.8046
w/o Energy Scaling	0.9265	0.7642	0.8222	0.8013
w/o TS-T	0.9324	0.7876	0.8238	0.7991
w/o Fusion	0.9183	0.7636	0.8186	0.7905
FreqMAE	**0.9365**	**0.7919**	**0.8420**	**0.8205**

the audio and seismic data from the moving vehicles having sparser frequency content with larger temporal correlation (*i.e.,* more stable movement) than HAR tasks. Therefore, the contribution of localized attention and temporal interaction is relatively more limited.

4.3 Feasibility in Real-World Deployment

4.3.1 Computation Overhead. Table 4 compares FreqMAE with baselines in terms of parameters, model size, and inference time. By running FreqMAE on a single-board Raspberry Pi 3 with 1 GB RAM and a 1.2 GHz quad-core CPU, we evaluate memory and inference time on deployment. The inference time is the execution time for inferring one sample (2-seconds length), averaged over 1000 experiments. Results show that although FreqMAE incurs slightly more inference time than DeepSense [54], a state-of-the-art supervised model for performance comparisons [23, 53], the overhead is comparable and affordable for the considered COTS devices. Moreover, the localized attention mechanism significantly reduces the computational overhead compared to Vanilla MAE, which utilizes a global attention mechanism. Finally, although FreqMAE has comparable size to the ViT, *FreqMAE's local attention mechanism significantly reduces the computational overhead and inference time while improving performance* in sensory data. Hence, *FreqMAE incurs 37% less overhead than its counterparts and allows real-time inference.*

Table 4: Compute Overhead Comparison.

Model	Parameters (M)	Size (MB)	Infer. Time (s)
DeepSense	0.563	2.193	0.491
ViT	2.821	10.850	1.503
Vanilla MAE	2.821	10.849	1.538
FreqMAE	3.036	11.693	0.972

Table 5: MOD variations for robustness experiments.

Variations	Sensor Locations	Vehicle Types	Terrain	# Labels
MOD-A	✓	✗	✗	3229
MOD-B	✗	✓	✗	6748
MOD-C	✗	✗	✓	1163

4.3.2 Robustness Test. Figure 10 illustrates our field testbed deployment across three distinct parking lot environments: MOD-A, B, and C. We placed FreqMAE sensor nodes with acoustic and seismic sensors strategically. The pretrained model from the MOD (see Table 1) is utilized for each classification, including variations listed in Table 5. MOD-A aligns closely with the original data, differing only in sensor placement. MOD-B has a similar terrain to MOD-A but uses different vehicles, while MOD-C is set on a concrete building rooftop, introducing distinct acoustic and seismic behaviors.

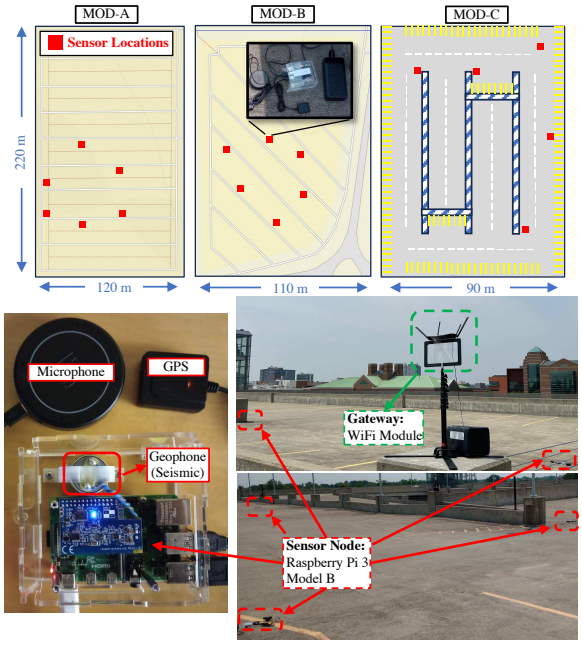


Figure 10: Robustness experiments were conducted in three environments with different variations.

Table 6 presents the robustness evaluations, *demonstrating FreqMAE's impressive resilience to environmental variations across deployments.* In MOD-A, changes to sensor locations are less challenging for models, as they mostly influence measurement intensity without significantly altering frequency signatures. For MOD-B, all frameworks struggle with vehicles absent during pretraining due to differing acoustic and seismic signatures with vehicle types. Yet, FreqMAE's performance excels, showcasing its ability to generalize and classify even unseen targets. Finally, in MOD-C, seismic alterations arise due to the concrete environment. However, FreqMAE effectively harnesses insights from physics-based pretraining and the fusion of complementary stable acoustic information, proving adept at distinguishing features even with domain shifts.

Contrastive baselines TS-TCC and CMC, though competitive in standard benchmarks (refer to Table 2 and Figure 8), underperform in changing environments. This drop can be attributed to the nature of contrastive frameworks. While they excel at extracting patterns through similarities among various sample "views", they lack the robustness provided by guidance based on generalized physical features, thereby affecting adaptability in dissimilar environments.

5 RELATED WORK

Self-Supervised Multi-Modal Representation Learning. Self-supervised learning advancements in language and vision have leveraged contrastive methods, relying on tailored spatial augmentations [7, 15], and generative approaches such as MAE [16]. While frameworks like CMC [42] and GMC [35] handle multimodal data, they overlook frequency aspects in time series. Unimodal time series have seen contrastive adaptations [11, 43, 55–57], and multimodal sensing has been addressed by Cosmo [30] and Cocoa [9], but without fine-tuning for modality-specific characteristics. Parallel to contrastive learning, Masked Image Modeling has shown

Table 6: Robustness against deployment variations.

	MOD-A		MOD-B		MOD-C	
Metric	Acc	F1	Acc	F1	Acc	F1
CMC	0.7415	0.7390	0.5760	0.4983	0.6412	0.5691
Cosmo	0.4205	0.3059	0.5816	0.5214	0.5496	0.2376
SimCLR	0.6733	0.6685	0.5377	0.3922	0.6107	0.3730
TS2Vec	0.6563	0.6439	0.5260	0.3521	0.5725	0.4487
TS-TCC	0.6051	0.5910	0.5012	0.1720	0.5802	0.4099
Vanilla MAE	0.8580	0.8602	0.6626	0.6347	0.6794	0.6326
LIMU-BERT	0.5000	0.1667	0.4233	0.1983	0.5649	0.2407
CAV-MAE	0.4801	0.4431	0.50309	0.21076	0.5419	0.3409
AudioMAE	0.5113	0.4981	0.4839	0.3475	0.4961	0.4571
FreqMAE	**0.8750**	**0.8766**	**0.6885**	**0.6622**	**0.7710**	**0.7340**

equivalent performance in vision [4, 16, 52]. Vision-language multimodal modeling has been widely explored [12], and LIMU-BERT [53] specifically targets generative modeling for IMU data. Differing from these, FreqMAE uniquely integrates multimodal features with a masked fusion approach and a physical domain-weighted objective, improving multi-modal sensor data representation learning.

Masked Spectrogram Learning. MAE, prevalent in vision-based self-supervised learning, is now being applied to Masked Spectrogram Learning [13]. While AudioMAE [19] and MSM-MAE [29] tackle single-modality audio spectrograms, and CAV-MAE [14] blends modality matching with MAE for image and audio, none address the unique characteristics of physical sensory data we motivate. Contrarily, *FreqMAE integrates physical insights in a multimodal approach for enhanced time series representation learning.*

6 DISCUSSION AND CONCLUSIONS

The paper introduced an IoT-centric masked autoencoding framework, informed by physics-based insights for sensor signals, to effectively capture crucial semantics for intelligent sensing tasks. Experimental evaluations showed that FreqMAE surpasses current state-of-the-art baselines across different tasks and reduces the need for data labeling, maintaining robustness during domain shifts. A potential limitation of FreqMAE may arise when a significant portion of the unlabeled pretraining data is noisy, potentially affecting the energy supervision from the weighted loss. In such scenarios, adjusting the energy contribution in the training objective to emphasize the reconstruction of important frequency content, typically less noisy, can be beneficial. In future work, we aim to explore training objectives more resilient to such noisy data.

ACKNOWLEDGMENTS

Research reported in this paper was sponsored in part by the Army Research Laboratory under Cooperative Agreement W911NF-17-20196, NSF CNS 20-38817, DARPA award HR001121C0165, DARPA award HR00112290105, DoD Basic Research Office award HQ00342110002, and the Boeing Company. Shengzhong Liu is also supported by the National Natural Science Foundation of China (Grant No. BE0300076, BC0301315, BC0301340). The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the CCDC Army Research Laboratory, or the US government. The US government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] E. K. Antonsson and R. W. Mann. The frequency content of gait. *Journal of biomechanics*, 18(1):39–47, 1985.

[2] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] A. Bansal, N. Aggarwal, D. Vij, and A. Sharma. An off the shelf cnn features based approach for vehicle classification using acoustics. In *Proceedings of the International Conference on ISMAC in Computational Vision and Bio-Engineering 2018 (ISMAC-CVB)*, pages 1163–1170. Springer, 2019.

[4] H. Bao, L. Dong, S. Piao, and F. Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*.

[5] T. Boroushaki, I. Perper, M. Nachin, A. Rodriguez, and F. Adib. Rfusion: Robotic grasping via rf-visual sensing and learning. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, SenSys '21, page 192–205, New York, NY, USA, 2021. Association for Computing Machinery.

[6] Y. Cao, A. Dhekne, and M. Ammar. Itracku: Tracking a pen-like instrument via uwb-imu fusion. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '21, page 453–466, New York, NY, USA, 2021. Association for Computing Machinery.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] A. D. Cobb, B. A. Jalaian, N. D. Bastian, and S. Russell. Robust decision-making in the internet of battlefield things using bayesian neural networks. In *2021 Winter Simulation Conference (WSC)*, pages 1–12. IEEE, 2021.

[9] S. Deldari, H. Xue, A. Saeed, D. V. Smith, and F. D. Salim. Cocoa: Cross modality contrastive learning for sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–28, 2022.

[10] I. Dirgová Luptáková, M. Kubovčík, and J. Pospíchal. Wearable sensor-based human activity recognition with transformer model. *Sensors*, 22(5):1911, 2022.

[11] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan. Time-series representation learning via temporal and contextual contrasting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2352–2359, 2021.

[12] X. Geng, H. Liu, L. Lee, D. Schuurmans, S. Levine, and P. Abbeel. Multimodal masked autoencoders learn transferable representations. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*.

[13] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10699–10709, 2022.

[14] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2023.

[15] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[17] A. Hore and D. Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[18] Z. Hu, Y. Zhang, T. Yu, and S. Pan. Vma: Domain variance-and modality-aware model transfer for fine-grained occupant activity recognition. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 259–270. IEEE, 2022.

[19] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer. Masked autoencoders that listen. *arXiv preprint arXiv:2207.06405*, 2022.

[20] B. K. Iwana and S. Uchida. Time series data augmentation for neural networks by time warping with a discriminative teacher. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3558–3565. IEEE, 2021.

[21] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

[22] S. Li, R. R. Chowdhury, J. Shang, R. K. Gupta, and D. Hong. Units: Short-time fourier inspired neural networks for sensory time series classification. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, SenSys '21, page 234–247, New York, NY, USA, 2021. Association for Computing Machinery.

[23] S. Li, R. R. Chowdhury, J. Shang, R. K. Gupta, and D. Hong. Units: Short-time fourier inspired neural networks for sensory time series classification. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 234–247, 2021.

[24] D. Liu. *Self-supervised learning frameworks for IoT applications*. PhD thesis, 2022.

[25] D. Liu, T. Wang, S. Liu, R. Wang, S. Yao, and T. Abdelzaher. Contrastive self-supervised representation learning for sensing signals from the time-frequency perspective. In *2021 International Conference on Computer Communications and Networks (ICCCN)*, pages 1–10. IEEE, 2021.

[26] T. Liu, M. Gao, F. Lin, C. Wang, Z. Ba, J. Han, W. Xu, and K. Ren. Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 97–110, 2021.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[28] Z. Liu and Y. Shao. Retromae: Pre-training retrieval-oriented transformers via masked auto-encoder. *arXiv preprint arXiv:2205.12035*, 2022.

[29] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino. Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation. *arXiv preprint arXiv:2204.12260*, 2022.

[30] X. Ouyang, X. Shuai, J. Zhou, I. W. Shi, Z. Xie, G. Xing, and J. Huang. Cosmo: Contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, MobiCom, page 324–337, 2022.

[31] S. Pan, M. Berges, J. Rodakowski, P. Zhang, and H. Y. Noh. Fine-grained recognition of activities of daily living through structural vibration and electrical sensing. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 149–158, 2019.

[32] L. C. Parra and C. Spence. Separation of non-stationary natural signals. *Independent component analysis: principles and practice*, pages 135–157, 2001.

[33] N. H. Parrish, A. J. Llorens, and A. E. Driskell. An agent-ensemble for thresholded multi-target classification. *Applied Sciences*, 10(4):1376, 2020.

[34] R. Pintelon and J. Schoukens. *System identification: a frequency domain approach*. John Wiley & Sons, 2012.

[35] P. Poklukar, M. Vasco, H. Yin, F. S. Melo, A. Paiva, and D. Kragic. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*, pages 17782–17800, 2022.

[36] A. Reiss and D. Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pages 108–109. IEEE, 2012.

[37] C. A. Ronao and S.-B. Cho. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert systems with applications*, 59:235–244, 2016.

[38] U. Sara, M. Akter, and M. S. Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019.

[39] M. Schaekermann, E. Law, K. Larson, and A. Lim. Expert disagreement in sequential labeling: A case study on adjudication in medical time series analysis. In *SAD/CrowdBias@ HCOMP*, pages 55–66, 2018.

[40] T. Sztyler and H. Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–9. IEEE, 2016.

[41] C. I. Tang, I. Perez-Pozuelo, D. Spathis, and C. Mascolo. Exploring contrastive learning in human activity recognition for healthcare. *arXiv preprint arXiv:2011.11542*, 2020.

[42] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[43] S. Tonekaboni, D. Eytan, and A. Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations*, 2021.

[44] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[46] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu. Deep learning models for real-time human activity recognition with smartphones. *Mobile Networks and Applications*, 25:743–755, 2020.

[47] L. Wang, P. Luc, A. Recasens, J.-B. Alayrac, and A. v. d. Oord. Multimodal self-supervised learning of general audio representations. *arXiv preprint arXiv:2104.12807*, 2021.

[48] T. Wang, D. Kara, J. Li, S. Liu, T. Abdelzaher, and B. Jalaian. The methodological pitfall of dataset-driven research on deep learning: An iot example. In *MILCOM 2022-2022 IEEE Military Communications Conference (MILCOM)*, pages 1082–1087. IEEE, 2022.

[49] X. Wang and G.-J. Qi. Contrastive learning with stronger augmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[50] Y. Wei, H. Hu, Z. Xie, Z. Zhang, Y. Cao, J. Bao, D. Chen, and B. Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022.

[51] H. Wu and J. M. Mendel. Classification of battlefield ground vehicles using acoustic features and fuzzy logic rule-based classifiers. *IEEE transactions on fuzzy*

systems, 15(1):56–72, 2007.
[52] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.
[53] H. Xu, P. Zhou, R. Tan, M. Li, and G. Shen. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 220–233, 2021.
[54] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th international conference on world wide web*, pages 351–360, 2017.
[55] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8980–8987, 2022.
[56] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. In *Proceedings of Neural Information Processing Systems, NeurIPS*, 2022.
[57] Y. Zhang, Z. Hu, U. Berger, and S. Pan. Cma: Cross-modal association between wearable and structural vibration signal segments for indoor occupant sensing. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*, IPSN '23, page 96–109, New York, NY, USA, 2023. Association for Computing Machinery.

A DATASETS

FreqMAE was tested on four datasets from Vehicle Classification (VC) and Human Activity Recognition (HAR) applications, covering various sensors, classes and environments, as shown in Table 1.

Moving Object Detection (MOD). This dataset, collected with a RaspberryShake 1D and microphone array at two sites, records vibrations from passing vehicles, including seven object types such as humans, at various speeds and distances. Seismic data was sampled at 100 Hz and acoustic at 16000 Hz.

Acoustic-seismic identification Data Set (ACIDS) [8]. Created by the US Army Research Lab, ACIDS includes over 270 runs of nine ground vehicle types across three environmental conditions, digitized at 1025 Hz. In VC, we added speed and distance classification tasks to evaluate FreqMAE's domain shift adaptability, with speeds (5, 10, 15, 20 mph) and distances (close, mid-range, far).

RealWorld-HAR [40]. This dataset records eight human activities using accelerometers, gyroscopes, magnetometers, and light sensors at 100 Hz from 15 participants' waists.

PAMAP2 [36]. It includes 18 physical activities from nine individuals, monitored using IMUs on the wrist, capturing data from a 3-axis accelerometer, gyroscope, and magnetometer at 100 Hz. This study focused solely on wrist data.

B PREPROCESSING AND TRAINING STRATEGIES

Here, we give preprocessing and training strategies at Section 4.1.

B.1 Preprocessing

During preprocessing, time-series data is segmented into equal-sized windows and further split into overlapping or non-overlapping intervals for creating spectrograms with Fourier transform. FreqMAE accommodates varying sampling rates of each modality with different encoders, using spectrograms for feature representation.

Datasets are split into training, validation, and test sets in an 8:1:1 ratio, with a realistic session-based division. Training data is varied in label availability (100%, 10%, 1%) for finetuning's **label ratio**. Self-supervised pretraining uses unlabeled data, and finetuning uses a linear classifier trained on the labeled training set.

B.2 Data Augmentations

This section details the augmentation methods from Section 4.1 used for contrastive baselines, chosen based on established practices to improve training. *Note that, unlike contrastive frameworks, FreqMAE does not require crafted augmentations as a self-automated framework capable of generalizing across various IoT task domains.*

Time-domain augmentations prior to spectrogram conversion:
- **Scaling:** Multiplies signals by Gaussian random values for scaling.
- **Permutation:** Randomly rearranges intervals within samples.
- **Jitter.** We introduce random Gaussian noise into the signals.
- **Negation:** Applies a -1 multiplier to signal values.
- **TimeWarp:** Uses a smooth curve to warp signal time locations.
- **MagnitudeWarp:** Modifies magnitudes with a cubic spline curve.
- **Horizontal Flip:** Flips the time series along the time axis.
- **Channel Shuffle:** Shuffles channels in multivariate data, such asthree-axis accelerometer input (X, Y, Z dimensions).

Frequency-domain augmentations after spectrogram conversion:
- **Phase Shift:** Applies a random phase shift in the range of $-\pi$ to π to the complex frequency spectrum's phase values.

B.3 Training Strategies

This section describes the hyperparameters and training methods for the models, as detailed in Section 4 and Table 7. Configurations remain mostly uniform across models. We utilize the AdamW optimizer with cosine scheduling, adjusting the initial learning rate for each model, a batch size of 128, and fine-tuning the temperature parameter for peak performance, as indicated in Table 8. We apply a 0.05 weight decay for regularization. For fine-tuning, we switch to the Adam optimizer with a step scheduler, lowering the learning rate by 0.2 every epoch across 200 epochs with 50 periods each, and adjusting weight decay to optimize training outcomes.

C BASELINES

Here, we provide baselines' introductions described in Section 4.1.
- **Supervised.** We train the entire model (*i.e.,* the encoder and linear classifier) in a supervised way with all of the available labels.
- **SimCLR [7].** Uses contrastive learning and NT-Xent loss to facilitate similarity between augmented views of the same sample.
- **CMC [42].** Creates embeddings by matching same-sample representations across modalities (positive pairs) and distancing different samples (negative pairs), leveraging multimodal data to enhance modality agreement using random batches and augmentations.
- **TS2VEC [55].** Enhances time series representation through contrastive tasks across window sizes, identifying same-sample augmentations and contexts as positive, and different samples or sequences as negative, supporting temporal and instance learning.
- **Cosmo [30].** Creates multimodal time-series representation through contrastive fusion, mapping modal embeddings to a hypersphere. It treats similar features as positive and dissimilar as negative pairs.
- **MAE [16].** MAE uses a self-supervised auto-encoding approach with Transformers, masking significant input parts and focusing on unmasked segments. It encodes modalities separately, then integrates embeddings to minimize reconstruction errors, using modality encoders and linear layers for inference.
- **LIMU-BERT [53].** Designed for unlabeled IMU data, LIMU-BERT adapts BERT's self-supervised learning to sensor data, capturing temporal patterns with custom adjustments for IMU specifics.
- **CAV-MAE [14].** CAV-MAE merges MAE's approach with contrastive learning for audio-visual data, using multi-stream processing for input reconstruction, distinct encoders for each modality, and a combined encoder to enhance cross-modal learning.

Table 7: TS-Transformer Configurations.

Dataset	MOD	ACIDS	RealWorld-HAR	PAMAP2
Dropout Ratio	0.2	0.2	0.2	0.2
Patch Size	aud: [1, 40], sei: [1, 1]	[1, 8]	[1, 2]	[1, 2]
Temporal Window Size	[1, 9]	[1, 8]	[1, 9]	[1, 8]
Mod Feature Block Num	[2, 2, 4]	[2, 2, 4]	[2, 2, 2]	[2, 2, 2]
Mod Feature Block Channels	[64, 128, 256]	[64, 128, 256]	[32, 64, 128]	[32, 64, 128]
Mod Fusion Channel	256	256	128	128
Mod Fusion Head Num	4	4	4	4
Mod Fusion Block	2	2	2	2
FC Dim	512	512	256	128
Temporal Shift	1	1	1	1

Table 8: Training configurations. (We use LR for Learning Rate)

Dataset	MOD	ACIDS	RealWorld-HAR	PAMAP2
Temperature	0.07	0.2	0.07	0.07
Lambda	0.1	0.3	1.0	0.3
Gamma	0.5	1.0	4.0	1.0
Pretrain Optimizer	AdamW	AdamW	AdamW	AdamW
Pretrain Max LR	Default: 1e − 5	Default: 1e − 4	Default: 1e − 4	Default: 1e − 4
Pretrain Epochs	6000	3000	1000	1000
Finetune Start LR	0.0001	0.0003	0.0005	0.001

• **AudioMAE [19].** AudioMAE, building on MAE [16], uses a Transformer with global and local attention for audio representation, setting a baseline for TS-T design evaluations. It transforms audio into spectrogram patches, masking some for efficient encoding.

D ADDITIONAL EVALUATIONS

D.1 Additional Downstream Tasks.

We assess pretrained models on distance and speed classification tasks using the MOD dataset. Results in Figure 13 reveal consistent outperformance by contrastive frameworks (SimCLR, CMC, TS-TCC) over other self-supervised methods (MAE, LIMU-BERT). FreqMAE's integration of modality, temporal characteristics, and physical insights enables superior adaptation on both tasks.

D.1.1 Representation Visualization. We employ the t-SNE algorithm [44] to visualize the fused embeddings of FreqMAE to show representation quality. Figure 11 illustrates FreqMAE embeddings, showing well-separated clusters in ACIDS and RealWorld-HAR datasets, indicating effective capture of underlying data structure In MOD and PAMAP2, cohesive clusters are observed, albeit with more overlap, due to a more challenging dataset structure.

D.2 Effect of Masking Strategies.

Figure 14 presents FreqMAE's performance with varying masking rates (60% to 90%) and strategies, comparing random unstructured masking to three structured variants: *(i)* Time masking for vertical spectrogram patches, *(ii)* Frequency masking for horizontal patches, and *(iii)* Time+Frequency masking, applied with equal probability.

Masking Rate. Similar to MAE in vision, a pretraining masking ratio of 70%-80% is ideal for learning spectrogram features, utilizing the redundancy in continuous signals (see Figure 2). Vehicle classification is more impacted by the masking ratio than HAR tasks, due to audio and seismic data's wider, complex frequency range. Very high masking ratios (e.g., 90%) decrease performance, underscoring the importance of a balanced self-supervised challenge for IoT data.

Masking Scheme. Unstructured (random) masking outperforms structured methods in self-supervised pretraining by using nearby contexts to estimate missing spectrogram sections. Frequency masking reduces performance by removing harmonic bands, while time masking effectively captures temporal correlations by reconstructing missing temporal content from related elements. Combining time and frequency masking approaches the effectiveness of unstructured masking through extrapolation from adjacent content.

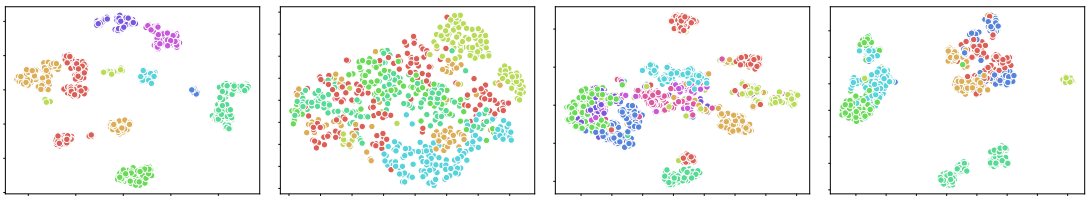
D.3 Fusion Hyperparameter (γ) Analysis.

Figure 12-(a, b) shows the effect of the information scaling hyperparameter (γ) on combining shared and private feature embeddings, as detailed in Section 3.3. This was tested with different settings across two datasets (ACIDS and PAMAP2) for VC and HAR tasks. A higher γ value emphasizes shared modal features, while a lower one highlights individual modality information. The aim is to find the best fusion approach for various tasks with FreqMAE.

Figure 12-(a) shows that VC tasks on the ACIDS dataset perform better with smaller fusion weights, due to the difficulty in reconstructing one modality from another in its audio-seismic combination and the imbalance in spectral content between modalities. Conversely, Figure 12-(b) demonstrates that HAR tasks on the PAMAP2 dataset, which involves multiple IMUs, benefit from larger fusion weights, enhancing classification due to the richer cross-modality fusion. *Such versatility enables FreqMAE to be applied broadly across various sensing tasks, providing an efficient and generalizable time series data representation framework for practitioners.*

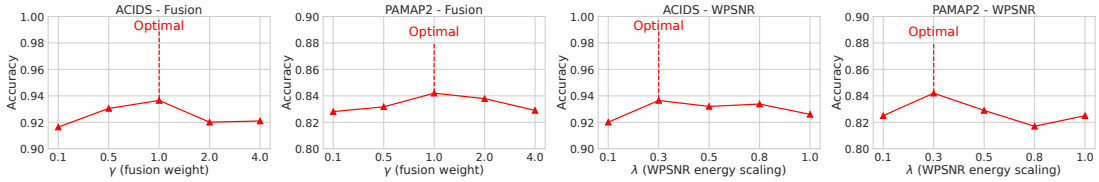
D.4 WPSNR Hyperparameter (λ) Analysis.

Figure 12-(c, d) shows the effect of energy contribution (λ) on training outcomes, with tests on ACIDS and PAMAP2 datasets for VC and HAR tasks, as discussed in Section 3.4. Increasing λ prioritizes high-energy samples, enhancing detection across tasks. Too low a λ decreases performance by failing to distinguish between high



(a) ACIDS (b) MOD (c) PAMAP2 (d) RealWorld-HAR

Figure 11: t-SNE visualization of FreqMAE embeddings. Different colors represent different ground truth labels.



(a) Fusion Analysis: ACIDS (b) Fusion Analysis: PAMAP2 (c) WPSNR Analysis: ACIDS (d) WPSNR Analysis: PAMAP2

Figure 12: Fusion (γ) and WPSNR energy contribution (γ) hyperparameter anaylsis.

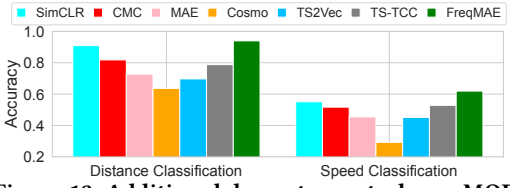
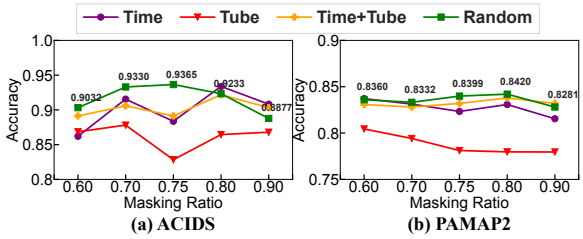


Figure 13: Additional downstream tasks on MOD.



(a) ACIDS **(b) PAMAP2**

Figure 14: Effect of masking strategy on performance.

signal-to-noise ratio (SNR) samples and those with negligible information (*e.g.,* background data or no observed content). For HAR tasks, a higher λ significantly improves performance, leveraging the energy content in IMU sensor readings to detect human activities effectively, as confirmed by optimal configurations in Table 8. This suggests that adjusting the energy contribution within the loss function can optimize model learning, especially for tasks where energy content is a critical indicator.

For ACIDS, overly high λ values disproportionately prioritize energy in learned representations, detrimental compared to HAR tasks where IMU sensors, placed directly on the body, exhibit less activity-related energy variation [40]. Conversely, in VC tasks, audio and seismic modalities face rapid energy fluctuations due to external deployment on moving vehicles, suggesting that high λ values might neglect low-energy yet informative signals. Hence, employing the WPSNR objective with energy supervision refines model focus towards high-quality representations, enhancing adaptability across sensor types and modality characteristics, *positioning FreqMAE as a versatile framework for diverse sensing applications.*

D.5 Effect of Positional Encodings.

This section evaluates the role of positional encoding in masked representation learning. Following Swin-Transformers [27], we add one-dimensional absolute positional embeddings (APE) to the patches, organizing patch inputs into a sequence by channel, then time, for various input lengths. These embeddings are combined with the inputs before entering the backbone network.

Table 9: Effect of positional encoding on FreqMAE

Setting	ACIDS Acc/F1	PAMAP2 Acc/F1	RealWorld-HAR Acc/F1	MOD Acc/F1
With	0.9265/0.7596	0.8312/0.8120	0.8783/0.8916	0.9377/0.9356
Without	0.9365/0.7919	0.8420/0.8205	0.9250/0.9327	0.9524/0.9514

Table 9 shows positional encoding's effect on frameworks, comparing TS-Transformer's classification with and without embeddings. Echoing [27], positional embeddings don't clearly enhance and may even reduce accuracy in sensing tasks, likely due to spectrogram non-stationarity. Spectrogram harmonic sequences display temporal shifts, and using positional information could cause overfitting to these changing sequences, which clashes with the TS-Transformer's Temporal Shift approach for temporal decoding.

TS-Transformer distinctively uses local attention for frequency details and Temporal Shifting for dynamic harmonics, focusing on inter-frequency links and adapting to physical data's non-stationarity without overfitting to exact positional frequency details.