PhyMask: An Adaptive Masking Paradigm for Efficient Self-Supervised Learning in IoT

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

Yatong Chen
Shanghai Jiao Tong University
Shanghai, Shanghai, China
chenyatong@sjtu.edu.cn

Jinyang Li
Ruijie Wang
Yizhuo Chen
Tianshi Wang
University of Illinois
Urbana-Champaign, USA
Urbana, Illinois, USA
jinyang7@illinois.edu

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

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

Abstract

This paper introduces PhyMask, an adaptive masking paradigm designed to enhance the efficiency and interpretability of Masked Autoencoders (MAEs) in analyzing IoT sensing signals. Different from all mainstream MAEs, which rely on random masking techniques, PhyMask employs an adaptive masking strategy that aligns with critical signal information. Its main contributions are threefold. First, PhyMask leverages the energy significance of frequency components to prioritize information-rich time-frequency regions, improving the reconstruction of original signals. Second, it includes a coherence-based masking component to identify and preserve essential temporal dynamics within the data. Finally, PhyMask integrates these components into an adaptive masking paradigm tailored to optimize the sensing context awareness within the masking configuration, focusing on the most informative parts of the data. This allows PhyMask to mask up to 96% of the input, reducing memory requirements by 14% and accelerating pre-training. Evaluations across two sensing applications, four datasets, and two real-world deployments demonstrate PhyMask's superior performance. PhyMask improves MAE accuracy by 7%, reduces pre-training data requirements by up to 75%, and enhances robustness to domain shifts and signal quality variations, making it of great value to robust and efficient intelligent IoT deployments.

CCS Concepts

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

Keywords

Multimodal sensing, Self-supervised learning, Internet of Things

ACM Reference Format:
Denizhan Kara, Tomoyoshi Kimura, Yatong Chen, Jinyang Li, Ruijie Wang, Yizhuo Chen, Tianshi Wang, Shengzhong Liu, and Tarek Abdelzaher. 2024. PhyMask: An Adaptive Masking Paradigm for Efficient Self-Supervised Learning in IoT. In *The 22nd ACM Conference on Embedded Networked Sensor Systems (SENSYS '24), November 4–7, 2024, Hangzhou, China.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3666025.3699325

SENSYS '24, November 4–7, 2024, Hangzhou, China

ACM ISBN 979-8-4007-0697-4/24/11
https://doi.org/10.1145/3666025.3699325

1 Introduction

Self-supervised learning has emerged as a powerful tool for leveraging the wealth of unlabeled data available in various domains, including IoT signal analysis. Self-supervised learning typically unfolds in a two-stage framework: pre-training on a vast, unlabeled dataset, followed by fine-tuning for a specific downstream task. This process has been shown to enhance downstream task performance, accelerate convergence, bolster robustness, and mitigate model overfitting [19, 22, 30]. In the realm of IoT sensing, the challenge of effectively encoding the intricacies of sensor data into useful representations has led to the exploration of masked autoencoders (MAEs) [21, 29, 31] and contrastive learning [8, 13, 37, 46, 59, 70] as primary self-supervised learning approaches.

In particular, MAEs have gained traction for their efficient self-supervised nature. MAE operates on inputs divided into sub-regions, or patches, of equivalent size. In the context of IoT signals, these patches typically refer to time-frequency segments of the input spectrogram [2, 24, 29, 45]. During pretraining, the MAE randomly masks a substantial portion (*e.g.,* 75%) of these patches and reconstructs them from the unmasked patches using an encoder-decoder architecture. Masking allows the MAE to encode only a small fraction of the original input, facilitating a straightforward and memory-efficient training regimen [3, 7, 18, 25]. However, the efficiency of MAEs heavily depends on the adopted masking strategy [34, 57, 61]. Traditional strategies like random patch [21], tube masking [61], and frame masking [7] explore various ways of selecting visible

time-frequency regions from the input signals [24, 45]. While random patch sampling can be somewhat effective, it intrinsically assumes homogeneous importance among data patches, which is seldom true in IoT sensing applications [29, 35]. Time-series signals vary significantly in their informational content, with many carrying redundant or minimal information. Random masking strategies often miss regions rich in spatiotemporal information, resulting in trivial reconstructions, a longer path to meaningful representation learning [3, 7, 32], and more training iterations compared to contrastive learning due to lack of focus on informative regions [4, 32]. A more discerning masking strategy is needed to prioritize information-rich regions, especially in the context of sensing signals where the region's importance can vary significantly due to the nature of physical phenomena being captured.

Moreover, a superior masking strategy needs to balance between masking "easy" and "hard" patches [4, 32]. Masking a large number of "hard" patches (*e.g.,* fundamental frequencies [56, 65, 66]) might challenge the model's ability to reconstruct them correctly from the remaining information, whereas masking too many "easy" patches (*e.g.,* low-energy harmonics, background noise) could make the self-supervised task too simple, hindering effective representation learning. This balance is crucial for achieving optimal performance, especially in sensing applications where the information density can vary greatly across different time-frequency regions [29, 35].

To address these challenges, we introduce PhyMask, a novel adaptive masking strategy tailored for MAEs that optimizes the pretraining process by aligning masking with critical sensor signal information. Central to our approach are three pivotal contributions:

First, PhyMask introduces an *energy-based masking component* to prioritize information-rich regions by evaluating their energy content. By identifying and sampling (*i.e.,* selecting as a visible patch) more from high-energy regions and less from low-energy backgrounds, PhyMask *guides mask sampling towards the most informative parts of the data, enhancing the learning of meaningful representations.*

Second, we introduce a *coherence-based masking* component to identify and preserve information that consistently manifests across various temporal sequences of data samples [29, 35, 49]. It leverages inter-sample coherent information to guide the masking process, complementing the energy-based masking component, which may mistake sudden noise blips or outliers for informative regions due to their high energy content. Moreover, the coherence-based masking component recognizes low signal-to-noise ratio (SNR) conditions, where the signal's energy is low but coherent information is still present. For instance, in moving target detection via radar [26, 50] or acoustic sensing [15, 65], coherent information may be present even if the signal energy is low due to the target's distance from the sensor. By focusing on regions with high coherence, even in low SNR conditions, and sampling fewer patches from regions with low coherence, PhyMask captures underlying patterns and changes in physical phenomena over time, guiding the mask sampling towards regions with significant spatiotemporal patterns.

Finally, PhyMask combines the above two masking components through a *Masking Adaptation Algorithm* to achieve dynamic masking configurations that effectively utilize both temporal coherence and frequency significance. Masking adaptation is achieved in two

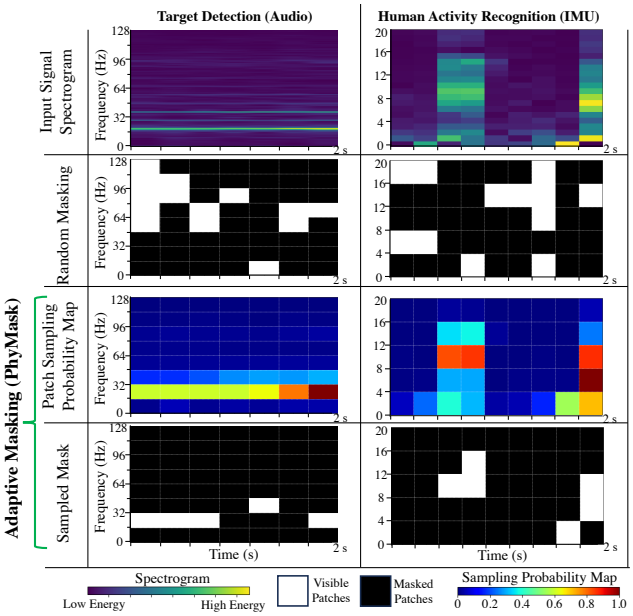

Figure 1: PhyMask vs. random masking at a 75% masking ratio. PhyMask adaptively samples more patches from regions with high spatiotemporal information and fewer patches from the background.

alternative methods: *(i) a weighted adaptation* that allows the user to control the weights of the coherence and energy significance for mask sampling, or *(ii) a self-adaptive paradigm* that decides the best adaptation through a Bayesian algorithm. This adaptation determines the weighting between the two masking components. Depending on the sensing application, Masking Adaptation allows the incorporation of prior knowledge (*e.g.,* consistent patterns in the data) into the mask sampling process or enables the model to automatically learn the best mask configuration through a self-adaptive paradigm when no prior knowledge is available. Hence, the masking adaptation guides PhyMask to effectively balance between easy and hard patches for the given sensing context, ensuring appropriate complexity for the self-supervised task and improved fine-tuning performance in downstream tasks.

We empirically show that PhyMask adaptively samples more patches from regions with high spatiotemporal information compared to random masking (Fig. 1). The left column shows an audio signal trace from a moving vehicle detection task, and the right column shows an IMU signal trace of a walking person for the Human Activity Recognition task. The audio signal from the vehicle has a highly coherent frequency pattern, while the IMU signal from the walking person has a more scattered pattern with less coherence. The second row shows random masking, which samples patches uniformly across both signals, ignoring unique signal characteristics and the importance of the patches. The third row shows the patch sampling probability calculated by PhyMask, which determines the probability of sampling (*i.e.,* selecting as visible) each time-frequency patch based on the signal's energy and coherence characteristics. The fourth row shows the actual patches sampled by PhyMask, which adaptively selects more patches from regions with high spatiotemporal information and fewer patches from the

background. The sampled patches are kept visible and used for encoding and reconstructing the masked patches, ensuring that *(i)* reconstruction is always performed from informative regions, and *(ii)* some informative regions are masked, balancing between easy and hard patches.

Our evaluation shows that PhyMask's adaptive mask allocation allows for high masking ratios (up to 96%) during MAE pre-training. This improves pre-training efficiency, reduces GPU memory requirements, and enhances downstream task accuracy. We evaluate PhyMask across two sensing applications and four datasets, demonstrating its effectiveness in various contexts. Additionally, we provide two real-world evaluations, showing that PhyMask outperforms traditional masking strategies in different environmental conditions and signal quality/distance variations. Specifically, PhyMask enhances MAE robustness to domain shifts with a 22% increase in accuracy and improves robustness to signal quality variations due to distance by an average of 21.3% in Moving Object Detection tasks [29, 37, 65] and 9% in indoor environments with shorter distances. Overall, PhyMask achieves an average accuracy improvement of up to 28% with 90% less labeled data, *emphasizing PhyMask's practical utility in dynamic, real-life scenarios, making it a valuable tool for representing information from dynamic sensing streams.*

2 Preliminaries

This section outlines the foundational concepts of masked autoencoders and the inspirations behind PhyMask's design.

2.1 Masked Autoencoders

MAEs have emerged as a mainstream self-supervised learning framework, particularly effective in domains with abundant unlabeled data, like vision [21] and natural language processing [5, 11, 64]. Unlike traditional contrastive learning paradigms for IoT signals, which rely heavily on domain-specific augmentations, MAEs provide an augmentation-free approach that significantly reduces dependence on labeled data across various sensing contexts.

An MAE adopts a two-phase training procedure: self-supervised pretraining followed by supervised task fine-tuning. During pretraining, the model learns to reconstruct masked portions of the input data, capturing essential semantic representations. The encoder processes the masked input, transforming it into low-dimensional embeddings, from which a decoder reconstructs the masked regions, minimizing reconstruction error. This stage leverages vast unlabeled data to learn versatile representations that can be easily fine-tuned for specific downstream tasks, offering two advantages: *(i)* fewer labels needed and *(ii)* faster training[29, 35].

2.2 Information Distribution in IoT Signals

To effectively handle the unique information distribution, it is crucial to consider the spatiotemporal characteristics inherent in IoT signals. Such data often exhibit complex patterns, where certain regions are significantly more informative than others [29, 35]. For example, in vibration data, the frequency domain can contain critical information about underlying physical phenomena, while the temporal coherence can reveal patterns and trends over time.

Mask sampling techniques are critical to the success of Masked Autoencoders (MAE) [3, 24, 32]. Previous studies have explored various sampling techniques, such as random "patch", "time", and "frequency" masking [24, 45]. While random patch sampling has been shown to work well in some cases, it assumes a uniform probability distribution over all input time-frequency patches, which is sub-optimal, contradicting the physical sensing nature. With these random masking strategies, visible patches may be sampled from redundant or low-information regions instead of high-information ones, leading to inaccurate reconstructions. This inhibits MAEs from learning meaningful representations and often requires more training iterations compared to contrastive learning methods. In contrast, our approach selects patches based on their spatiotemporal information. Unlike uniform random sampling, PhyMask's adaptive strategy ensures that the masking process focuses on the most informative regions of the data. This targeted approach not only enhances the quality of the learned representations but also improves the efficiency of the training process by reducing the number of training iterations before convergence.

The following sections detail the main insights that pave the foundation of PhyMask designs leveraging spatiotemporal IoT signal characteristics to optimize the masking process.

2.2.1 ***High Energy for Informativeness:*** The energy of frequency components in the spectral domain is a strong indicator of meaningful content in sensory data. Frequencies with higher energy typically reflect significant sensory information and meaningful patterns. By recognizing the importance of these high-energy frequency components, one can ensure that the most informative segments of the data are prioritized. For instance, in human activity recognition, the energy of specific frequency components can indicate the presence of critical physical actions, such as walking, running, or jumping [51, 54]. Highlighting these components allows for effective representation learning and accurate classification of different activities, leading to improved performance in activity recognition tasks.

2.2.2 ***Information Consistency for Robustness:*** Using energy alone to identify informative time-frequency patches makes the masking process prone to noise and outliers. For instance, a sudden spike in energy may not necessarily indicate an informative region but rather noise or an outlier. Moreover, focusing solely on energy may overlook consistent patterns that are crucial for understanding the underlying data dynamics. For example, in moving object classification, consistent (i.e., coherent) patterns over time, even if exhibiting lower energy compared to random noise in any one sample, can indicate the presence of specific objects or movements [65]. By identifying these consistent patterns (via coherence calculations), the essential dynamics of the data can be effectively captured, improving the robustness and quality of learned representations and enhancing the model's ability to classify moving objects accurately.

2.2.3 ***Adaptive Information Combination:*** While the high-energy frequency component and the inter-sample coherence component are effective in identifying informative regions, they both present natural limitations that depend in part on the inference task at hand. The high-energy frequency component could be sensitive to noise and outliers since it only considers sample energy. On the other hand, the inter-sample coherence component alone cannot distinguish between consistent patterns caused by target signals

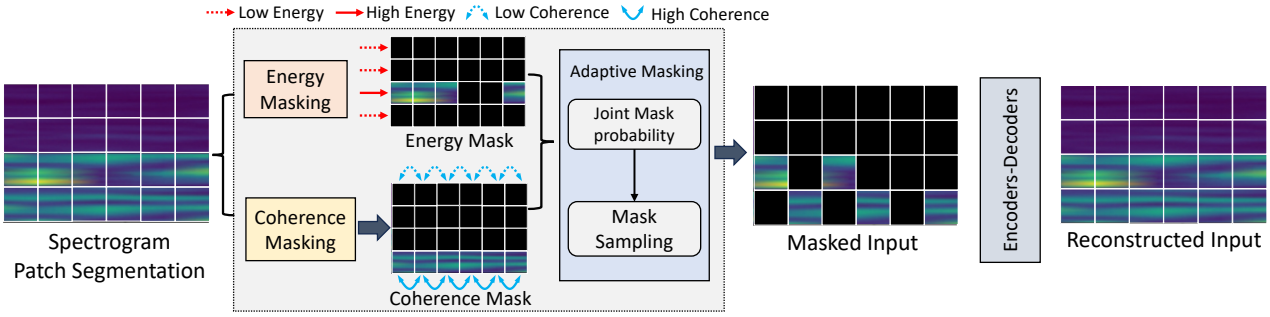

Figure 2: PhyMask overview with self-supervised pretraining workflow.

and those caused by signals irrelevant to the task, such as wind noise in target tracking. The inter-sample coherence component may falsely mark these regions as informative with high coherence.

It is therefore crucial to combine these two components in a way that leverages each of their strengths while mitigating their weaknesses. Some tasks have inherently consistent patterns, such as vehicle classification, while others may have more sporadic patterns, such as human activity recognition or anomaly detection, with less consistent but more sudden energy signatures. Hence, for instance, adding a higher coherence component in tasks with consistent patterns can help filter out noise and outliers, while adding a higher energy component in tasks with sporadic patterns can help capture sudden changes and critical information. By adaptively combining these two components, an approach can effectively leverage the strengths of each while mitigating their weaknesses to retain essential information while filtering out noise and outliers. We compare two approaches, one informed by the task at hand and one self-adaptive. The comparison helps us understand the trade-off between utilizing prior task knowledge versus relying on full automation in combining energy and coherence to inform masking policy.

3 PhyMask Framework

In this section, we first present an overview of PhyMask[1] and then describe its three novel components, motivated by the discussed properties of time-series signals.

3.1 Overview

PhyMask is integrated as the masking strategy of the MAE framework. Consider P modalities $M = \{M_1, M_2, \ldots, M_P\}$ and N unlabeled training samples $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$, where each $\mathbf{X}_i$ represents a multimodal input window, such that the window of modality M_j is denoted by $\mathbf{X}_{ij}$. Each window $\mathbf{X}_{ij}$ comprises a number of consecutive sensor data values of the indicated modality. Each multimodal window is called a sample. The objective is to obtain embeddings as sensor signal representations, utilizing modality-specific encoders for embedding generation. Input samples, $\mathbf{X}_{ij}$, are converted into spectrograms via the Short-Time Fourier Transform (STFT) for detailed time-frequency analysis. During pretraining, as shown in Figure 2, spectrograms are segmented into patches for linear projection embeddings. Positional embedding is excluded, as supported by [39], since the time-frequency representation inherently captures the temporal information.

Algorithm 1 PhyMask with Integrated Mask Components

Require: Time series data for each modality M_j: $X_j = \{X_{ij}(t) : \text{for } i = 1, 2, 3, \ldots, |D|\}$

for $V_i \in D$ **do**
  $X_i \leftarrow \text{PatchSegmentation}(V_i)$
  % Phase 1 - Mask sampling probability calculation
  $C_j \leftarrow \text{CoherenceMask}(X_i)$
  $E_j \leftarrow \text{EnergyMask}(X_i)$
  % Phase 2 - Adaptive Mask Combination
  $p \leftarrow \text{AdaptiveMask}(C_j, E_j)$
  $\mathcal{M}_{M_j} \leftarrow \text{BernoulliSample}(p)$ ▷ Mask sampling
  $X_v \leftarrow X_i \sim \mathcal{M}_{M_j}$ ▷ Visible patches after masking
  % Phase 3 - MAE reconstruction with PhyMask
  $F_v \leftarrow \text{Encoder}(X_v)$ ▷ Encode visible features
  $F_m \leftarrow \text{Mask}(X_i[\mathcal{M}_{M_j}])$ ▷ Masked patches
  $F' \leftarrow F_v \oplus F_m$ ▷ Concatenate masked patches
  $X' \leftarrow \text{Decoder}(F')$ ▷ Decode combined features
  $\mathcal{L}_R \leftarrow ||X'[\mathcal{M}_{M_j}] - X_i[\mathcal{M}_{M_j}]||_2$ ▷ Reconstruction loss
  backward
end for

The full PhyMask process is outlined in Algorithm 1. The process begins by segmenting the input data into patches, which are then used to calculate the *Energy Mask* and *Coherence Mask* for each modality (as shown in Figure 3). The *Energy Mask* is based on the normalized energy density of the frequency components, highlighting regions with significant energy. The *Coherence Mask* is calculated based on the average coherence between sample frequencies which measures the similarity and consistency of frequency components across different samples. Then, an Adaptive Mask algorithm combines the *Energy Mask* and the *Coherence Mask* to compute the final mask sampling probability, with a Bernoulli process to construct the final mask. Finally, we train the MAE with this mask through reconstruction and compute the mean squared error between the masked input and the reconstructed output for back-propagation to update the model parameters. This process is repeated for each modality in the dataset, allowing the model to learn the underlying structure of the data and generate embeddings that capture the essential features of the input data. Considering the two-dimensional nature of spectrograms for time-frequency

[1] https://github.com/denizhankara/PhyMask

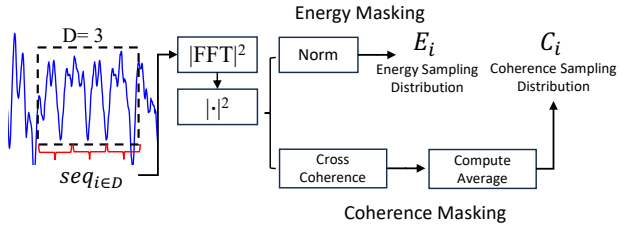

Figure 3: Two Masking schemes: Energy (left) and Coherence (right).

Algorithm 2 EnergyMask Pseudocode

Require: $X_j = \{X_{ij}(t) : i = 1, 2, 3, \ldots, |D|\}$ ▷ Time series data for modality M_j

% Phase 1 - Compute Energy Spectrum for each sample
for $X_{ij} \in D$ **do**
  $X_{ij}(f) \leftarrow \sum_{t=0}^{|T|-1} X_{ij}(t) \cdot e^{-\frac{2\pi i}{|T|} ft}$ ▷ FFT computation
  $P_{x_{ij}}(f) \leftarrow |X_{ij}(f)|^2$ ▷ Energy Spectrum
end for
% Phase 2 - Scale Energy Spectrum for Energy Distribution
for $X_{ij} \in D$ **do**
  $E_{ij} \leftarrow \frac{P_{x_{ij}}(f) - \min(P_{x_{ij}}(f))}{\max(P_{x_{ij}}(f)) - \min(P_{x_{ij}}(f))}$ ▷ $0 \le E_{ij} \le 1$
end for
return E_j ▷ Energy Mask Sampling Distribution

analysis, we found PhyMask's approach superior to both unstructured random masking and various structured masking strategies for pretraining while applying higher masking ratios.

3.2 Energy Masking

Sensor signals, particularly those from time-series data, exhibit strong frequency significance that random masking fails to leverage. This information is essential for capturing the dynamics of physical phenomena. Motivated by this, we propose *Energy Masking*, which ensures that spatially important regions are either preserved or masked based on their energy significance. Inspired by the relative energy of frequency components being a significant indicator of meaningful content in sensory data (see Section 2.2.1) [29, 35], Energy Masking assigns varying probabilities of sampling to different frequency regions based on their energy content. During the sampling process, spatially significant patches are more often selected to be visible for encoding. Moreover, the energy-based masking strategy effectively reduces the probability of low-energy regions, such as background noise or low-energy harmonics, being sampled. By filtering out noise and irrelevant information from the reconstruction process, energy masking allows the models to always have access to information-rich patches, facilitating a more accurate and efficient learning process.

Algorithm 2 outlines the EnergyMask process. After FFT, the energy spectrum of the frequency components is calculated, and the energy distribution is scaled to a range of $[0, 1]$. The scaled energy distribution is then used to create the *Energy Mask* sampling distribution for the modality M_j. The energy mask is then used to sample the patches of the input data for masking to select the most informative frequency components for learning.

Assigning the masking probability proportional to the energy metrics keeps the information distribution simple and consistent

Algorithm 3 CoherenceMask Pseudocode

Require: $X_j = \{X_{ij}(t) : i = 1, 2, 3, \ldots, |D|\}$ ▷ Time series data for modality M_j

% Phase 1 - Compute FFT and auto PSD for each sample
for $X_{ij} \in D$ **do** ▷ Sample i
  $X_{ij}(f) \leftarrow \sum_{t=0}^{|T|-1} X_{ij}(t) \cdot e^{-\frac{2\pi i}{|T|} ft}$ ▷ Compute FFT
  $P_{x_{ij}x_{ij}}(f) \leftarrow |X_{ij}(f)|^2$ ▷ PSD of $X_{ij}(f)$
end for
% Phase 2 - Compute average coherence for each sample
for $X_{ij} \in D$ **do** ▷ Sample i
  $S_{ij}(f) \leftarrow 0$ ▷ Initialize coherence sum for sample i
  for $X_{kj} \in D$ **do** ▷ Sample k
    $P_{x_{ij}x_{kj}}(f) \leftarrow X_{ij}(f) \cdot (X_{kj}(f))^*$ ▷ Cross PSD
    $C_{ijk}(f) \leftarrow \frac{|P_{x_{ij}x_{kj}}(f)|^2}{P_{x_{ij}x_{ij}}(f) \cdot P_{x_{kj}x_{kj}}(f)}$ ▷ Cross coherence
    $S_{ij}(f) \leftarrow S_{ij}(f) + C_{ijk}(f)$ ▷ Coherence sum
  end for
  $C_{ij}(f) \leftarrow \frac{S_{ij}(f)}{|D|-1}$ ▷ Sample i average coherence, $0 \le C_{ij}(f) \le 1$
end for
return C_j ▷ Coherence Mask Sampling Distribution

with the energy characteristics of the signal. By leveraging the higher energy in a signal patch being indicative more significant information content [17, 52], we normalize the energy values and use them directly in a Bernoulli process for masking. This approach prioritizes informative regions during training by probabilistically distributing them between masked and unmasked portions based on their energy content. For instance, harmonic frequency bands [1, 42] can be distributed in this way between masked and unmasked portions, allowing the model to reconstruct masked parts from unmasked ones and capture the harmonic structure of the signal. This method simplifies the masking strategy without additional complexity, effectively enhancing the self-supervised learning process by focusing on the most informative regions.

3.3 Coherence Masking

PhyMask's *Coherence Masking* component addresses shortcomings of Energy Mask (See Section 2.2.2) by leveraging temporal coherence to identify and preserve consistent patterns across temporal sequences. This approach helps to filter out inconsistent or noisy regions, allowing the model to focus on the most informative and stable parts of the data. These components are weighted together to generate the final mask sampling probabilities.

Algorithm 3 outlines the CoherenceMask process. The coherence between different samples is calculated via cross-power spectral density (CPSD). The average coherence is computed to create the *Coherence Mask* sampling distribution, assigning higher probabilities to regions with higher coherence, ensuring that consistent patterns are prioritized during training.

We assign the masking probability proportional to the coherence metrics because coherence is inherently a normalized measure ranging from 0 to 1, representing the consistency of signal sections across samples [44, 63]. Using coherence directly as the masking probability provides consistent probability distribution, ensuring that selected energetic sections are also informative while avoiding introducing additional complexity or computational overhead to the mask components. This approach allows us to adaptively combine

Algorithm 4 WeightedAdaptiveMask Pseudocode

Require: Coherence distribution C_j and Energy distribution E_j for modality M_j
Require: Validation set $\mathcal{V}$ for tuning weights W_C and W_E
1: *% Initialize weights*
2: Initialize weights W_C, W_E s.t. $0 \leq W_C, W_E \leq 1$
3: *% Tune weights using the validation set $\mathcal{V}$*
4: **for** each combination of W_C and W_E **do**
5: Apply mask using $p = \frac{W_C \cdot C_j + W_E \cdot E_j}{W_C + W_E}$
6: Train model and evaluate performance on $\mathcal{V}$
7: **end for**
8: Select W_C^* and W_E^* that maximize performance on $\mathcal{V}$
9: *% Generate final mask probabilities*
10: $p \leftarrow \frac{W_C^* \cdot C_j + W_E^* \cdot E_j}{W_C^* + W_E^*}$ ▷ Combine and normalize
11: **return** p ▷ Return mask sampling probability

and weight the energy and coherence components when constructing the final mask without concerns about differing units or scales. By keeping the probability calculations based on the same signal property, energy, we can effectively account for both energetic and consistent sections of the signal, enhancing the model's ability to focus on meaningful patterns during training while avoiding the need for more complex weighting schemes.

3.4 Adaptive Mask Construction

Using energy or coherence alone to guide the masking process in self-supervised learning for IoT signals presents several challenges. Energy-based masking effectively captures high-energy, informative segments but is prone to noise and outliers. On the other hand, coherence-based masking focuses on consistent patterns over time but may overlook high-energy, sporadic events that carry significant information. To address these limitations, we propose an Adaptive Mask Construction approach that combines the strengths of both energy and coherence masking. By leveraging both methods, we ensure that the most informative segments of the data, whether characterized by high energy or temporal consistency, are prioritized and more often kept visible for encoding during training.

This approach balances "easy" and "hard" patch masking by assigning masking probabilities based on the level of information content (hardness). The energy and coherence components select and distribute informative signal parts between masked and unmasked portions by assigning varying probabilities to these regions. By combining the probabilities from both components, the adaptive masking methods distribute informative regions between masked and unmasked sections, considering both the significance (energy) and consistency (coherence) of the information. This results in a balanced reconstruction task where the model must reconstruct informative patches using context from other informative (unmasked) patches. Consequently, the model learns to capture complex patterns and relationships within the data, improving its ability to generalize to downstream tasks. Additionally, adaptive masking results in a balanced distribution of informative components between masked and unmasked portions, effectively guiding the learned information toward meaningful content through its binomial probabilistic distribution algorithm.

Algorithm 5 BayesianAdaptiveMask Pseudocode

Require: Coherence distribution C_j and Energy distribution E_j for modality M_j
1: *% Initialize prior and likelihoods*
2: $P(\mathcal{M}) \leftarrow \mathcal{U}(0,1)$ ▷ Initialize prior mask sampling probability as random
3: $P(\mathcal{M}|C) \leftarrow C_j$ ▷ Sampling probability given coherence
4: $P(\mathcal{M}|E) \leftarrow E_j$ ▷ Sampling probability given energy
5: *% Calculate mask sampling posterior*
6: $P(\mathcal{M}|C,E) \leftarrow \frac{P(\mathcal{M}|C) \cdot P(\mathcal{M}|E)}{P(\mathcal{M})}$ ▷ Sampling Posterior
7: **return** $p = P(\mathcal{M}|C,E)$ ▷ Mask sampling distribution

We evaluate two algorithms for combining energy and coherence masking: *(i)* Weighted Adaptive Mask Generation (Algorithm 4) assigns weights to each component based on optimal validation performance, while *(ii)* Bayesian Adaptive Mask Generation (Algorithm 5) learns the importance of each component during the training process.

3.4.1 Weighted Adaptive Mask Generation. The Weighted Adaptive Mask algorithm calculates the final mask sampling probability by tuning the weights of the energy (W_E) and coherence (W_C) masking probabilities using a validation set. The algorithm iterates over various weight combinations to find the optimal weights that maximize performance on the validation set. The final mask sampling probability is then generated using these tuned weights, providing a flexible and adaptive masking strategy by integrating available prior information, such as inherent coherence in actions (*e.g.,* continuous moving object) or sporadic high-energy frequency components (*e.g.,* human activity recognition).

However, this approach still requires manual tuning of weights, which can be time-consuming and may not always yield optimal results. Additionally, the algorithm may be less flexible when prior information on the target task is not available, as it relies on the user to determine the importance of each component based on the data's characteristics.

3.4.2 Bayesian Adaptive Mask Generation. The limitations above motivate us to propose a Bayesian Adaptive Mask algorithm that estimates the masking probability using Bayesian posterior inference based on a random prior probability (as in Random Masking in MAE literature [21, 29, 45]) and the energy and coherence masking probabilities, rather than purely relying on prior distribution.

The prior mask sampling probability $P(\mathcal{M})$ is approximated by a uniform random distribution due to the lack of prior information about the spatiotemporal structure of the data. The algorithm then calculates the posterior mask sampling probability by combining the estimated energy $P(\mathcal{M}|E)$ and coherence $P(\mathcal{M}|C)$ masking probabilities as conditional distributions given the coherence and energy distributions, as shown in Algorithm 5. The combination (line 6 of Algorithm 5) follows the Bayesian formulation [28] to create a conditional probability distribution of masking regions. We assume that the energy and coherence metrics, i.e., $P(E, C|\mathcal{M}) = P(E|\mathcal{M})P(C|\mathcal{M})$, are conditionally independent given the masking distribution from the signal information. Intuitively, this is the case if energy and coherence components are independent sources of information. This assumption is reasonable because energy and coherence capture different aspects of the signal's information

Table 1: Finetune results on different datasets. We mark the best and second best results.

	PAMAP2		RWHAR		ACIDS		Parkland		Average	
Metric	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CMC	0.7571	0.7223	0.8211	0.8384	0.7836	0.6452	**0.9049**	**0.9023**	0.8167	0.7770
Cosmo	0.7910	0.7469	0.8529	0.7968	0.8776	0.7298	0.3228	0.3241	0.7111	0.6494
SimCLR	0.7346	0.6635	0.7830	0.7181	0.5658	0.4879	0.7535	0.7434	0.7092	0.6532
TS2Vec	0.5706	0.4942	0.6117	0.5002	0.6539	0.4913	0.7649	0.7632	0.6503	0.5622
TS-TCC	0.7871	0.7107	0.8684	0.8227	0.8758	0.7400	0.7709	0.7744	0.8256	0.7619
Vanilla MAE	0.7382	0.6999	0.8638	0.8700	0.8521	0.6908	0.7817	0.7793	0.8090	0.7600
LIMU-BERT	0.7847	0.7612	0.7946	0.7261	0.5023	0.3171	0.2157	0.1236	0.5743	0.4820
AudioMAE	0.7808	0.7478	0.8163	0.7437	0.7845	0.6120	0.7274	0.7249	0.7773	0.7071
CAVMAE	0.7995	0.6711	0.9113	**0.9153**	0.7995	0.6711	0.5432	0.5266	0.7663	0.6996
B-PhyMask	0.7940	0.7582	0.9038	0.8985	0.9347	**0.8272**	0.7970	0.7930	0.8574	0.8192
W-PhyMask	**0.8156**	**0.7719**	**0.9159**	0.9137	**0.9365**	0.8044	0.8929	0.8916	**0.8794**	**0.8325**

content (i.e., information significance and consistency, respectively) and can be considered independent sources of information.

The Bayesian approach offers a fully automated and self-adaptive masking strategy that can effectively learn the importance of each component during training and does not require manual tuning of the prior weights, allowing for a more flexible and adaptive masking strategy tailored to the specific characteristics of the data.

4 Evaluation

In this section, we describe our experimental setups and extensive evaluations to demonstrate PhyMask's performance and resiliency. We further ablate PhyMask to understand the contributions of its components. Lastly, we present two real-world case studies on vibration-based applications to illustrate PhyMask's deployment feasibility.

4.1 Experimental Setup

4.1.1 Datasets and Preprocessing: We evaluate PhyMask on four multimodal time-series datasets on human activity recognition and vehicle detection applications. **PAMAP2** [51] includes IMU data from accelerometer, gyroscope, and magnetometer sensors collected from 9 participants performing 18 different activities. **RealWorld-HAR (RWHAR)** [54] consists of accelerometer, gyroscope, magnetometer, and light sensor signals collected from 15 participants in 8 physical activities. **ACIDS** is a private dataset involving acoustic and seismic identification with 9 vehicle types across three terrains. **MOD** [37] is a vibration-based dataset containing seismic and acoustic signals describing nearby moving vehicles.

We preprocess the time-series data by segmenting it into fixed-size windows to create samples for training and evaluation. Within each window, we compute the Fourier transform to generate the frequency spectrum, resulting in spectrograms that serve as input to PhyMask. The lengths of the windows and samples are determined based on the specific characteristics of each dataset to effectively capture the relevant temporal dynamics. Since different modalities may have varying sampling rates, PhyMask handles each modality separately with dedicated feature encoders, accommodating differences in data resolution.

For training, we partition each dataset into training, validation, and test sets using an 8:1:1 ratio, ensuring a realistic split by leaving entire sessions out of the training set. The training set is further divided based on different labeling rates (e.g., 100%, 10%, 1%) to assess the model's performance under varying amounts of labeled data during fine-tuning. This approach allows us to evaluate PhyMask's effectiveness in scenarios with scarce labeled data, demonstrating its capability to learn meaningful representations from limited data.

4.1.2 Baselines: We compare PhyMask with three contrastive learning (CMC [59], Cosmo [46], SimCLR [8], two time-series (TS2Vec [70], TS-TCC [13]), and four reconstruction-based (Vanilla MAE [21], LIMU-BERT [68], AudioMAE[24], CAVMAE [18]) self-supervised learning frameworks.

4.1.3 Backbone models: We leverage Swin Transformers (SW-T) [39] as the backbone encoders to generate modality embeddings from their spectrogram inputs. SW-T is a state-of-the-art transformer model originally designed for image processing. We adapt it to process spectrogram inputs by partitioning the input spectrograms into non-overlapping time-frequency patches. For multimodal inputs, we use separate SW-T encoders for each modality to extract modality-specific features. These features are then fused using additional self-attention layers to capture cross-modal relationships. In reconstruction-based frameworks, the modality embeddings are used to reconstruct the unmasked input through decoders that mirror the encoder configuration. This approach enables the model to learn meaningful representations by predicting the masked portions of the input from the visible patches.

4.1.4 Training Details: In this section, we explain the hyperparameters and training strategies used for both PhyMask and the baseline models in our evaluations. During the pretraining phase, we apply the AdamW optimizer along with cosine learning rate schedulers for all models. The initial learning rate for each framework is adjusted according to its convergence behavior, typically set to 1e-4. We use a batch size of 128, with each batch consisting of randomly chosen samples from the unlabeled training data. A weight decay of 0.05 is applied as regularization during pretraining. For the contrastive learning frameworks, we apply eight time-domain augmentations (scaling, permutation, negation, time warp, magnitude warp, horizontal flip, jitter, and channel shuffle) and one frequency-domain augmentation (phase shift), which are commonly used in contrastive learning practices [27, 35–37, 55]. These augmentations are applied to the input data to create positive and negative pairs for contrastive learning. The models are trained on a workstation

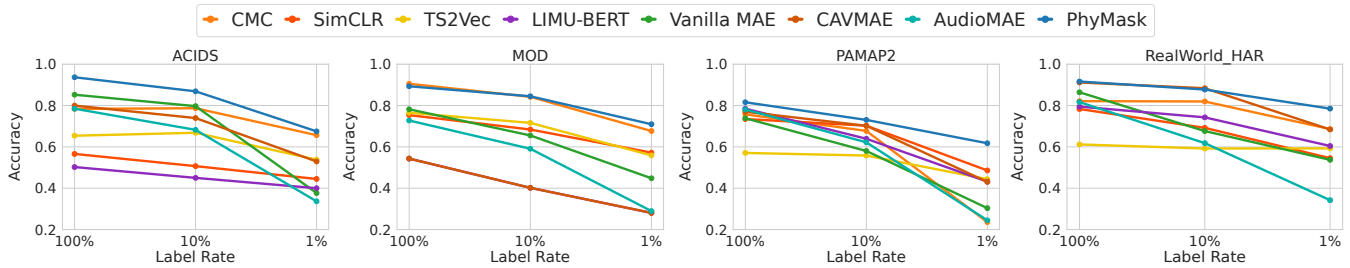

Figure 4: Accuracy comparison of PhyMask with different labeling rates.

with an AMD Threadripper PRO 3000WX processor (64 cores) and NVIDIA RTX 3090 GPUs, and the implementation is built on PyTorch 2.0.1.

In the fine-tuning phase, we adopt the Adam optimizer coupled with a step learning rate scheduler, where the learning rate decays by a factor of 0.2 at the end of every 50 epochs. Fine-tuning spans 200 epochs for each downstream task. The batch size remains at 128, and the weight decay is adjusted for each framework to balance training and validation performance. All models are fine-tuned using labeled data specific to each downstream task. We vary the amount of labeled data to assess performance under different labeling rates (e.g., 1%, 10%, 50%, 100%) for different tasks. Similarly, we train models with varying amounts of pretraining data to assess performance under different data ratios for pretraining efficiency.

4.2 Evaluation Results

4.2.1 Full Evaluations: Table 1 compares PhyMask's performance with Vanilla MAE and other self-supervised learning frameworks across different datasets. All evaluations use the same backbone model, SWIN-Transformers, with a linear layer added for classification tasks. Results show that PhyMask improves Vanilla MAE's performance across all datasets, averaging a 7% accuracy and 7.3% F1 score improvement. Moreover, PhyMask outperforms other self-supervised learning frameworks and MAE variants by at least 5.4% in accuracy and 5.6% in F1 score, demonstrating its utility in enhancing the MAE framework.

Both Weighted (W-PhyMask) and Bayesian (B-PhyMask) adaptation variants perform well across all datasets, with W-PhyMask achieving the highest average accuracy and F1 score due to its application-guided adaptability with flexible weighting. B-PhyMask shows a slight performance drop compared to W-PhyMask, likely due to its inability to capture optimal mask sampling combinations across varying dataset characteristics. Notably, B-PhyMask experiences a larger accuracy drop in the MOD task, possibly due to the dataset's unique features, such as the presence of multiple vehicle types and terrains, which complicates optimal mask sampling. However, B-PhyMask offers a fully automated mask generation process that is more straightforward and less computationally intensive than W-PhyMask, which requires additional hyperparameter tuning of weights to achieve optimal performance. Overall, the results demonstrate the effectiveness of PhyMask in enhancing the performance of Vanilla MAE and other self-supervised learning frameworks across various datasets.

4.2.2 Label Efficiency on Finetune: Here, we evaluate the performances of baselines and PhyMask with different labeling rates during the finetuning phase, ranging from 1% to 100%. Figure 4 presents the comparison results across all datasets. Higher labeling rates tend to yield improved accuracies across most models. PhyMask consistently outperforms the baseline models across all labeling rates and datasets. Importantly, the performance improvement of PhyMask over the baselines is more pronounced at lower labeling rates, particularly at the 1% label rate. This demonstrates PhyMask's ability to effectively learn from limited labeled data by focusing on the most informative content in physical sensing data. By leveraging its adaptive masking strategy, PhyMask can extract more meaningful representations from limited data, resulting in better performance even when labeled data is scarce. We observe that only CMC is competitive with PhyMask, leveraging a rich set of augmentations for effective learning from unlabeled data. However, CMC experiences significant performance degradation in HAR tasks and relies heavily on carefully designed augmentations for optimal performance [29, 62]. Overall, these results demonstrate that *PhyMask's superior performance, especially at low labeling rates, makes it a practical and efficient solution for real-world applications where labeled data is limited.*

4.2.3 Pretraining Data Efficiency: Here, we compare the pretraining efficiency of PhyMask in terms of unlabeled data requirements and convergence speed with Vanilla MAE. Figure 5 shows the performance of PhyMask with different amounts of pretraining data. PhyMask consistently outperforms Vanilla MAE across both datasets, demonstrating its efficiency in extracting the most informative features from unlabeled data. Vanilla MAE experiences a significant decline in performance as the amount of pretraining data decreases since it disregards the informative parts of the limited data. In contrast, PhyMask maintains high performance even when only 10% of the pretraining data is used.

Additionally, we analyze the convergence speed of Vanilla MAE and PhyMask during pretraining. Figure 6 shows the pretraining loss curves. Vanilla MAE exhibits a slower convergence rate, requiring more epochs to reach a lower loss value. Despite being trained 2.5 times longer in the ACIDS case, Vanilla MAE still does not converge. In contrast, PhyMask converges faster than Vanilla MAE, reaching a lower loss value in fewer epochs and converging quickly in around 100 epochs. These results demonstrate that

Table 2: Ablation study results for PhyMask across different datasets.

Datasets	ACIDS		MOD		RWHAR		PAMAP2	
Components	Acc	F1	Acc	F1	Acc	F1	Acc	F1
PhyMask	**0.9365**	**0.8044**	**0.8894**	**0.8880**	**0.9159**	**0.9137**	**0.8156**	**0.7719**
PhyMask-OnlyEnergy	0.8890	0.7546	0.7790	0.7735	0.8725	0.8790	0.7677	0.7377
PhyMask-OnlyCoherence	0.9005	0.7309	0.8227	0.8201	0.8436	0.8589	0.6589	0.6177
Vanilla (Random Mask)	0.8521	0.6908	0.7817	0.7793	0.8638	0.8700	0.7382	0.6999

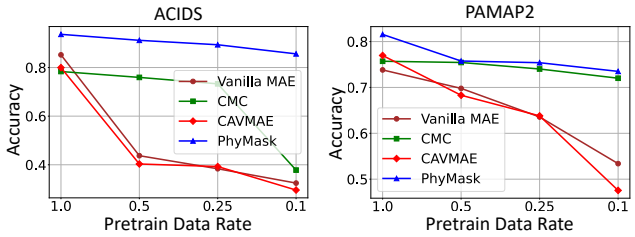

Figure 5: Performance analysis for different Pretrain Data Ratios on ACIDS (left) and PAMAP2 (right).

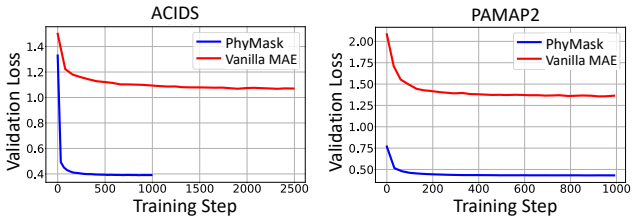

Figure 6: Pretrain loss for Vanilla MAE and PhyMask.

PhyMask masking is more efficient in leveraging spatiotemporal information within unlabeled data, enabling faster convergence and higher performance with less data.

4.2.4 Ablation Study: Table 2 presents an ablation study using all datasets to assess the contributions of the coherence and energy terms in PhyMask. We studied three PhyMask variants: *(i)* OnlyEnergy (without the Coherence mask), *(ii)* OnlyCoherence (without the Energy mask), and *(iii)* PhyMask (with both coherence and energy masks). Results show that the coherence mask component is more effective in vehicle and moving object classification tasks, where there are coherent and consistent temporal movement patterns. Such patterns are hard to capture by only considering the energy significance of short samples, as it overlooks consistently occurring signatures within time series sequences. In contrast, the energy mask component is more effective in HAR tasks, where there are more sporadic movement patterns with less consistency but more sudden energy signatures. The coherence term alone tries to capture coherent behavior where no consistent frequency patterns exist, degrading accuracy. This leads to the largest performance drop in the PAMAP2 case, which includes more daily activities with aperiodic patterns (*e.g.,* lying, sitting, ironing *etc.*) [51]. Overall, the results demonstrate *the necessity of both PhyMask components for optimal performance and adaptability across various tasks.*

4.2.5 Sensitivity Analysis: We evaluate the effect of varying weights on coherence and energy terms in PhyMask and observe their impact on accuracy, as shown in Figure 7. Consistent with the ablation study, higher coherence weights emphasize harmonic coherence

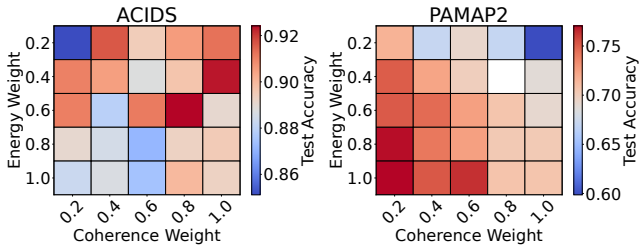

Figure 7: Sensitivity analysis for different masking weights on ACIDS (left) and PAMAP2 (right).

over individual sample energy, improving performance in tasks with consistent motion or behavior, such as vehicle movement. The ACIDS dataset shows notable improvement with increased coherence weights, validating the importance of capturing coherent spatiotemporal patterns. Conversely, higher energy weights enhance performance in tasks with sporadic movement patterns, such as activity recognition. The PAMAP2 dataset demonstrates significant performance gains with increased energy weights, underscoring the need to focus on informative regions within individual samples for effective activity recognition. Overall, the ablation and sensitivity analyses highlight *PhyMask's flexibility and adaptability in handling various tasks by adjusting the emphasis on signal coherence and energy characteristics.*

4.2.6 Mask Sampling Techniques: Table 3 compares PhyMask with other mask sampling techniques, including Tube [61] (masking individual frequencies), Random [21], and Time [7] (masking time intervals), each using a 75% mask ratio [16, 24, 29]. TimeMask shows competitive performance, especially in the HAR dataset, due to its ability to capture sporadic energy patterns crucial for HAR tasks. TubeMask, however, degrades performance across all datasets by missing temporal information and failing to capture coherent spatiotemporal patterns, essential for vibration-based tasks. RandomMask performs well overall, surpassing TimeMask in the ACIDS and MOD datasets and offering better average performance than TubeMask. By randomly masking both frequency and time intervals, RandomMask can capture coherent and energy patterns to some extent, leading to competitive performance. Finally, PhyMask outperforms RandomMask in all datasets by employing a smarter mask sampling strategy that adaptively combines both coherent and energy patterns, leading to superior performance across all tasks. Consistent performance improvements achieved by PhyMask suggest that its adaptive masking strategy effectively generalizes across different data and task complexities within the same domain. In cases where a dataset allows for multiple downstream tasks, PhyMask is likely to maintain similar performance trends compared to other masking methods, since it consistently outperforms them across different datasets within the same domain (see Table 1 and

Table 3: Results of different masking strategies across various datasets.

Datasets	Masking Ratio: 0.75									
Masking Strategy	Time		Tube		Time+Tube		Random		PhyMask	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
ACIDS	0.7722	0.5996	0.7009	0.5751	0.6886	0.5319	0.8521	0.6908	**0.9365**	**0.8044**
PAMAP2	0.8013	0.7695	0.5753	0.4776	0.7650	0.7293	0.7599	0.7120	**0.8156**	**0.7719**
MOD	0.3168	0.2846	0.4628	0.4513	0.4957	0.4897	0.7817	0.7793	**0.8894**	**0.8780**
RWHAR	0.8771	0.8800	0.7721	0.7235	0.8569	0.8500	0.8638	0.8700	**0.9159**	**0.9137**

Table 4: Mask sampling techniques on ACIDS dataset. PhyMask works better than other masking strategies and requires less memory.

Case	Mask Rate	Accuracy	F1	Memory
Random	75%	0.8521	0.6908	4.15 GB
Time	75%	0.7722	0.5996	4.14 GB
Frame	75%	0.7009	0.5751	4.14 GB
Random	87.5%	0.8176	0.6485	3.77 GB
PhyMask (adaptive)	**95%**	**0.9365**	**0.8044**	**3.55 GB (-14%)**

Table 5: Average mask rates (MR) of PhyMask

	Vanilla MAE (Random)		B-PhyMask		W-PhyMask	
Metric	MR	Accuracy	MR	Accuracy	MR	Accuracy
ACIDS	75%	0.8521	96%	0.9347	95%	0.9365
MOD	75%	0.7817	99%	0.7970	80%	0.8929
PAMAP2	75%	0.7847	98%	0.7940	85%	0.8156
RealWorld HAR	75%	0.8638	98%	0.9038	96%	0.9159

Table 3). This indicates that PhyMask's approach captures underlying informative patterns beneficial for various applications in a domain, demonstrating its robustness and versatility.

Next, we study the memory gain of PhyMask compared to other mask sampling techniques. Table 4 shows the memory consumption of different mask sampling techniques on the ACIDS dataset. We observe that increasing mask percentage to 87.5 % for random masking leads to more efficient memory usage, as the model can learn more from the masked data. However, this comes at the cost of decreased accuracy, as shown in Table 4. PhyMask, on the other hand, can effectively utilize a higher mask rate (95%) while maintaining high accuracy, resulting in a 14% memory reduction compared to RandomMask. This demonstrates *PhyMask's efficiency in memory usage and performance.*

Finally, we evaluate the average mask rates calculated by PhyMask across different datasets in Table 5. Bayesian PhyMask (B-PhyMask) achieves the highest mask rate across all datasets, demonstrating its adaptability to varying dataset characteristics. However, it degrades performance in MOD due to the dataset's unique features, such as multiple vehicle types and terrains, which complicate optimal mask sampling. The high masking rate destroys important information in the MOD dataset, leading to a 10% performance drop. This highlights a limitation of B-PhyMask in capturing optimal mask sampling combinations across varying dataset characteristics. In contrast, Weighted PhyMask (W-PhyMask) achieves optimal performance across all datasets with competitive mask rates, demonstrating its adaptability to different domains with flexible weighting possibilities. An 80% mask rate is optimal for the MOD dataset, balancing the need for information retention and destruction to achieve optimal performance. Overall, the results demonstrate the effectiveness of both PhyMask variants in achieving high mask rates while maintaining high accuracy across various datasets, leading to performance and computational efficiency.

Table 6: Compute Overhead Comparison.

Model	Size (MB)	Pretrain Overhead (s)	Infer. Time (s)
ViT	47.48	0.063	0.256
Vanilla MAE	49.06	0.112	0.230
PhyMask	49.06	0.142	0.238

4.2.7 Computational Overhead: Table 6 compares the compute overhead of PhyMask with Vanilla MAE and ViT [39] baselines in terms of parameter size, model size, and inference time. We deploy PhyMask to a Raspberry Pi 4 with 8 GB RAM and a 1.8 GHz Quad-core Cortex-A72 CPU. The inference time is the average duration to infer one sample (2 seconds) over 100 experiments. Results show PhyMask maintains the same model size and inference time as Vanilla MAE since PhyMask is a masking paradigm applied only during pretraining, not affecting the model size or inference time. PhyMask also achieves similar inference times to a supervised ViT classifier, demonstrating the feasibility of self-supervised pretraining in real-world applications. While PhyMask has a slightly higher pretraining overhead than Vanilla MAE due to additional masking computations, this increase is minimal and only affects the offline pretraining stage. Overall, the results demonstrate that *PhyMask is lightweight and efficient, making it suitable for deployment on edge devices with limited resources and real-time applications.*

4.3 Feasibility in Real-World Deployment

This section presents two real-life case studies to demonstrate the feasibility of PhyMask in real-world deployment scenarios. We evaluate the performance of PhyMask in two vibration-based applications: indoor human activity recognition and outdoor vehicle classification. We show that PhyMask can be effectively used in real-world applications, achieving high accuracy with minimal labeled data. Moreover, PhyMask is robust to noise and distance variations, making it suitable for various applications.

*4.3.1 Real-World Case Study 1: Human Activity Recognition.*We evaluate the performance of PhyMask for monitoring office events. Five sensor nodes are placed in the office, each equipped with a Raspberry Shake to measure seismic signals and a microphone array to record audio signals. These nodes effectively capture daily office activities and equipment usage from various directions, encompassing 13 floor-based and desktop activities, including walking, typing, shredder usage, fan usage, gaming, boiling water, moving chairs, wiping the table, clothes cleaner, robot master, refrigerator usage, washing machine usage, and grinder usage. There are 6 volunteers

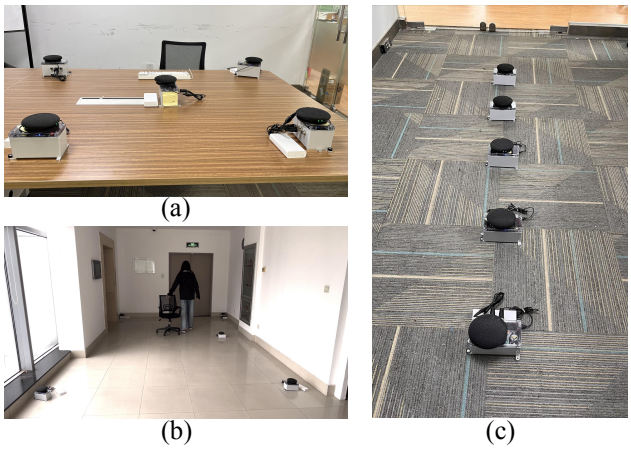

Figure 8: (a) and (b) are the scenarios of human activity recognition on the table and floor. (c) is the arrangement of different sensor distances.

in total (2 female and 4 male). Informed consent was obtained from all human participants for the experiments and data collection. The collecting scenarios consider different floor textures, such as carpet, wood, and tile; different table thicknesses (thin and thick tables); and different sensor distances, ranging from 5 cm to 250 cm. Figure 8 illustrates the experimental setup: (a) and (b) are for activity recognition, while (c) is used for sensor distance classification and activity recognition at different sensor distances. Data from the same sensing node can only appear in one set. The training sets include 4 nodes, and the testing sets include 1 node.

In Figure 9, we present the activity classification performance of PhyMask compared to other baselines across different fine-tuning label ratios (100%, 50%, 10%). PhyMask consistently outperforms other frameworks in the office deployment scenarios across all label ratios. At 100% and 50% label ratios, PhyMask achieves superior accuracy performance compared to the baselines. Even at a low label ratio of 10%, PhyMask maintains good performance with accuracy remaining above 80%, while other frameworks show more significant degradation. This slight performance decrease at very low label ratios is due to less consistent information being available from the limited labeled samples. However, PhyMask still offers robust performance in these cases and largely outperforms other frameworks in scenarios with limited data availability (see 1% label rates in Figure 4). This demonstrates PhyMask's effectiveness in real-world deployment scenarios where labeled data may be scarce. Moreover, comparing the Bayesian adaptive mask generation with weighted adaptive mask generation, we observe that the latter shows greater adaptability across all label ratios. This highlights its effectiveness in handling varying amounts of labeled data due to its flexibility in weighting the energy and coherence components.

We further assess PhyMask's capability in different downstream tasks, specifically classifying sensor distances in both floor and table scenarios as shown in Figure 10. In the floor scenario, each sensor is placed 50 centimeters apart, while in the table scenario, the distance is 30 centimeters. Activities are performed vertically above the sensor nodes. We observe that PhyMask variants are particularly robust in the floor scenario, where distances are larger and signals are more attenuated. Even in shorter distances on the

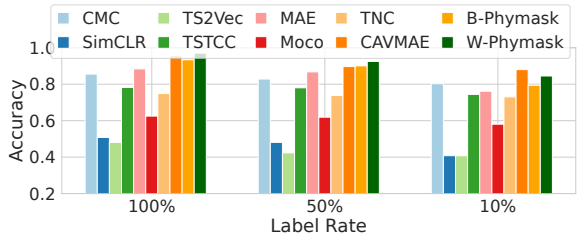

Figure 9: Inoffice Activity Classification performance across different label ratios.

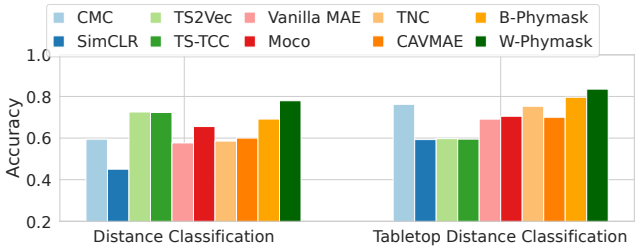

Figure 10: Comparison of classification performance in an office environment.

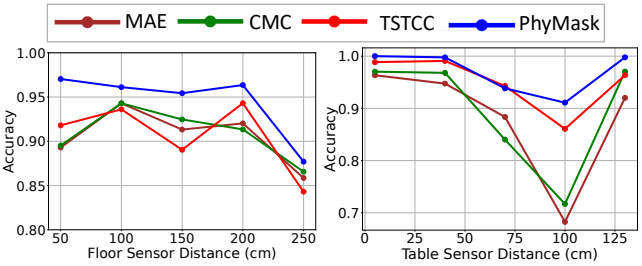

Figure 11: Detection accuracy versus different sensor deployment distances.

tabletop, where contrastive baselines perform well due to less attenuation, PhyMask still outperforms them, showing its robustness across various scenarios. Additionally, we examine PhyMask's sensitivity to different sensor distances on the table and floor separately in Figure 11. Floor evaluations show that PhyMask maintains over 90% accuracy within a sensor distance of 200 centimeters, consistently outperforming other baselines. Table evaluations further reveal that PhyMask's robustness becomes more evident as the distance increases, significantly surpassing other baselines. Overall, PhyMask *demonstrates strong resilience against environmental and signal quality variations due to distance, making it suitable for real-world IoT applications.*

4.3.2 Real-World Case Study 2: Vehicle Detection. We conduct a vehicle detection case study to evaluate PhyMask's performance and robustness against environmental and signal quality variations in real-world applications. Figure 12 shows the satellite view of the case study environment, a 200x50 meter outdoor field with gravel and sand surfaces. We deploy eight Raspberry Pi nodes with the same microphone and geophone setup as described earlier [33]. Four distinct vehicles, representing different brands and types with diverse acoustic and seismic signatures, are selected for this study. GPS information tracks each vehicle's real-time distance from the sensors. During each run, a single vehicle drives in cyclical patterns

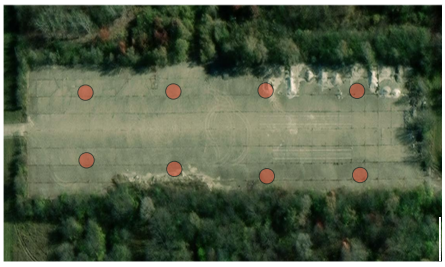

Figure 12: Vehicle detection deployment condition. Sensor nodes and object traces are also marked in colors.

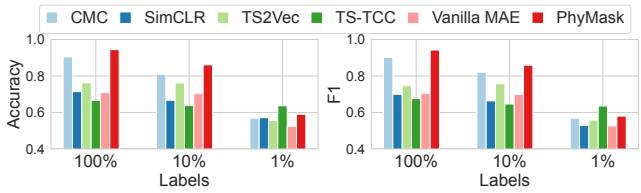

Figure 13: Vehice Detection Deployment Performance.

for an hour, varying speed, direction, and distance from the sensors. We fine-tune the models pretrained on the MOD dataset with the first 15 minutes of collected data and test with the remaining 45 minutes. Notably, the location and vehicles used differ from those in MOD, introducing domain shifts to simulate realistic conditions and challenges.

Figure 13 shows the performance of PhyMask compared to other methods across different fine-tuning label ratios (100%, 10%, and 1%). PhyMask offers significant performance benefits across various label availability scenarios. Notably, PhyMask maintains over 90% accuracy with all labels and 85% with only 10% of the labels, outperforming prior works. While PhyMask experiences a performance drop with 1% labels compared to TS-TCC due to the diminishing benefits of information accumulation from coherence masking with fewer labels, it still maintains competitive accuracy and tangible benefits over other baselines and Vanilla MAE.

Figure 14 illustrates the performance of PhyMask at various sensor distances. Each subplot represents the performance of the model fine-tuned at the specific sensor distance, with the x-axis indicating the maximum testing sensor distance. As the distance between the vehicle and the sensors increases, performance naturally declines due to the attenuation of vibration signals, causing strong signal quality variations. However, PhyMask shows less degradation in accuracy and consistently outperforms existing methods, especially when targets are more than 50 meters away from the sensors. Distance benefits are more pronounced when the model is fine-tuned with data collected from closer sensor distances. For example, baselines fine-tuned with data from 10 meters show a significant performance drop when tested with targets beyond 40 meters. In contrast, PhyMask maintains high accuracy across all distances, achieving the maximum performance gap with baseline methods when targets are farthest from the sensors. Overall, these results demonstrate *PhyMask's robustness to environmental and signal quality variations due to its ability to capture coherent and energy patterns even with low SNR signatures, making it suitable for real-world IoT applications.*

5 Related Work

Self-Supervised Representation Learning. Recent advancements in self-supervised learning (SSL) have primarily focused on contrastive learning (CL) and Masked Autoencoders (MAE). CL [6, 8, 9, 19, 36] aims to bring similar instances closer together in the representation space while pushing dissimilar instances apart. Besides treating augmented views as positive pairs (similar) in the unimodal context, CL has also been extensively applied to multimodal data, considering samples from different modalities as positive pairs [37, 47, 48, 59, 74]. Additionally, CL has been explored for time-series data, employing various temporal augmentations and contrasting based on temporal correspondences [13, 14, 60, 69, 70, 72].

On the other hand, Masked Autoencoders (MAE) [2, 21, 24, 58, 64, 67] focus on masking parts of the input data and learn to extract features that are most important for reconstruction. In contrast to CL, MAE does not heavily rely on data augmentations to generate different views. MAE has also been applied to time-series data, such as audio spectrograms [24, 45] and videos [20, 61]. Recently, CAV-MAE [18] introduced a contrastive audio-visual masked autoencoder that integrates contrastive learning with masked reconstruction for multimodal self-supervised learning. By combining the strengths of both CL and MAE, CAV-MAE is a promising benchmark for multimodal self-supervised learning. For physical sensing, CL has shown significant advancement using multimodal sensor data from wearable devices [10, 23, 46]. Others have also focused on the reconstruction of IMU signals for Human Activity Recognition [41, 68, 73].

Masking strategies in MAE. Different masking strategies have also been explored [4, 29, 71]. For instance, AdaMAE utilizes an adaptive masking strategy to enhance spatiotemporal learning by sampling visible tokens based on semantic context to improve performance on video action classification tasks [4]. Another approach, SemMAE, incorporates semantic information into the training process by proposing a Semantic-Guided Masking strategy, which guides the network to learn various information from intra-part patterns to inter-part relations, achieving state-of-the-art performance on various vision tasks [32]. Additionally, improving MAEs by learning where to mask, as explored in AutoMAE, shows that adaptive masking can lead to more efficient pre-training by focusing on patches with higher information density [7]. While these methods have been demonstrated to work well in vision-based tasks, they are not tailored to the unique characteristics of physical sensing data, such as seismic, acoustic or IMU signals. Such signals often contain time-series or frequency domain information, which is very different from image data [29]. For instance, AutoMAE relies on detecting informative regions (i.e., continuous foreground objects) within the image to use object-centric priors in mask sampling, helping the model focus on areas with objects. This approach of finding informative regions is not suitable for physical sensing data, where the information distribution in a sample is not continuous but uniquely defined by the frequency domain structure of the signal [29, 35] and distributed across sparse frequency bands such as fundamental frequency and signal harmonics [1, 42, 53]. Moreover, physical sensing data may include sparse noise and signal attenuations [38, 40, 43], and inherent temporal dependencies where data points are correlated across time [12, 29, 43]. These characteristics

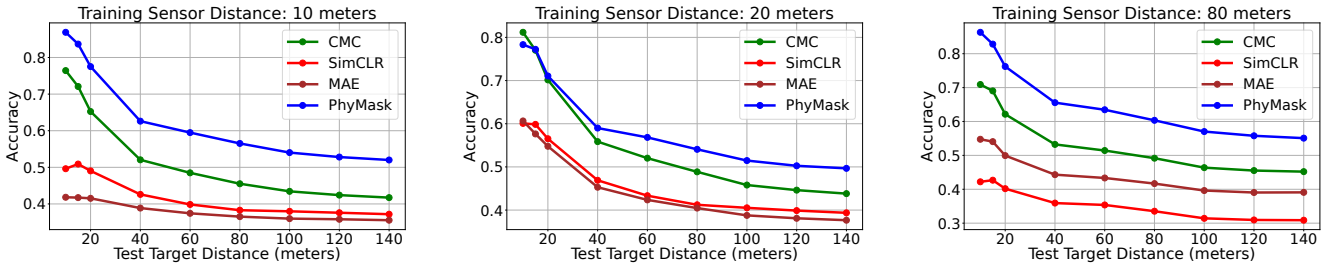

Figure 14: Detection accuracy vs. target distance for models trained with data collected from specific sensor distances. Each graph represents a training scenario where the model is trained only with data from vehicles within the indicated sensor distance and then tested across vehicle targets with various distances. Performance distribution at other distances shows a similar pattern and is therefore omitted due to space constraints.

disrupt the assumptions made by existing adaptive masking strategies, which are designed for vision-based tasks. PhyMask, on the other hand, is specifically designed to address these challenges by uniquely identifying sparsely and consistently informative regions through the energy and coherence metrics that are suitable for physical sensing data.

6 Discussion and Conclusions

We presented PhyMask, an adaptive masking strategy for Masked Autoencoders. Through physics-informed masking, PhyMask significantly enhanced the learning efficiency and the representation quality of time series sensing data. Our evaluations across diverse datasets and two real-world case studies demonstrated PhyMask's superior performance, deployment feasibility and maintained robustness during domain shifts in IoT sensing applications. PhyMask has a potential limitation that opens avenues for future extensions. When significant data variation occurs due to multiple environmental factors, noise, and target variations, the adaptive masking strategy may struggle to capture all variations, leading to suboptimal performance. Future research will explore more advanced adaptive masking strategies that can dynamically adjust the masking configuration to enhance robustness in complex scenarios.

Acknowledgments

Research reported in this paper was sponsored in part by NSF CNS 20-38817 and the Boeing Company. Yatong Chen and Shengzhong Liu are supported by China NSF grant No. 62472278, 62332014, and 62332013, as well as SJTU Kunpeng&Ascend Center of Excellence. 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 US government. The US government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation hereon.

References

[1] Jürgen Altmann, Sergey Linev, and Axel Weiß. 2002. Acoustic–seismic detection and classification of military vehicles—developing tools for disarmament and peace-keeping. Applied Acoustics 63, 10 (2002), 1085–1107.

[2] Alan Baade, Puyuan Peng, and David Harwath. 2022. Mae-ast: Masked autoencoding audio spectrogram transformer. arXiv preprint arXiv:2203.16691 (2022).

[3] W. G. C. Bandara, Naman Patel, A. Gholami, Mehdi Nikkhah, M. Agrawal, and Vishal M. Patel. 2022. AdaMAE: Adaptive Masking for Efficient Spatiotemporal Learning with Masked Autoencoders. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 14507–14517. https://doi.org/10.1109/CVPR52729.2023.01394

[4] Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M Patel. 2023. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 14507–14517.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision. 9650–9660.

[7] Haijia Chen, Wendong Zhang, Yunbo Wang, and Xiaokang Yang. 2023. Improving Masked Autoencoders by Learning Where to Mask. ArXiv (2023). https://consensus.app/papers/improving-masked-autoencoders-learning-where-mask-chen/7dce063d4078594db8b66ecb776481ea/?utm_source=chatgpt

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In International conference on machine learning. PMLR, 1597–1607.

[9] Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. In IEEE/CVF International Conference on Computer Vision (CVPR).

[10] Shohreh Deldari, Hao Xue, Aaqib Saeed, Daniel V Smith, and Flora D Salim. 2022. COCOA: Cross Modality Contrastive Learning for Sensor Data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6, 3 (2022), 1–28.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

[12] Zhaocai Dong, Kun Liu, Dongyu Han, Yuan Cao, and Yuanqing Xia. 2022. Reconstruction-based Multi-Scale Anomaly Detection for Cyber-Physical Systems. 2022 4th International Conference on Industrial Artificial Intelligence (IAI) (2022), 1–6. https://doi.org/10.1109/IAI55780.2022.9976844

[13] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2021. Time-Series Representation Learning via Temporal and Contextual Contrasting. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. 2352–2359.

[14] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2023. Self-supervised contrastive representation learning for semi-supervised time-series classification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[15] Naoko Evans. 2010. Automated vehicle detection and classification using acoustic and seismic signals. Ph. D. Dissertation. University of York.

[16] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. 2022. Masked autoencoders as spatiotemporal learners. Advances in neural information processing systems 35 (2022), 35946–35958.

[17] Ripul Ghosh, Aparna Akula, Satish Kumar, and HK Sardana. 2015. Time–frequency analysis based robust vehicle detection using seismic sensor. Journal of Sound and Vibration 346 (2015), 424–434.

[18] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. 2023. Contrastive Audio-Visual Masked Autoencoder. In The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=QPtMRyk5rb

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. Advances in Neural Information Processing Systems 33 (2020), 21271–21284.

[20] Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li. 2024. Siamese masked autoencoders. Advances in Neural Information Processing Systems 36 (2024).

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16000–16009.

[22] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. 2019. Using self-supervised learning can improve model robustness and uncertainty. Advances in neural information processing systems 32 (2019).

[23] Zhiqing Hong, Zelong Li, Shuxin Zhong, Wenjun Lyu, Haotian Wang, Yi Ding, Tian He, and Desheng Zhang. 2024. CrossHAR: Generalizing Cross-dataset Human Activity Recognition via Hierarchical Self-Supervised Pretraining. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 8, 2 (2024), 1–26.

[24] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2022. Masked autoencoders that listen. arXiv preprint arXiv:2207.06405 (2022).

[25] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2022. SemMAE: Semantic-Guided Masking for Learning Masked Autoencoders. Advances in Neural Information Processing Systems 35 (2022), 28708–28720.

[26] E. Hyun, Youngseok Jin, and Jong hun Lee. 2016. A Pedestrian Detection Scheme Using a Coherent Phase Difference Method Based on 2D Range-Doppler FMCW Radar. Sensors (Basel, Switzerland) 16 (2016). https://doi.org/10.3390/s16010124

[27] Brian Kenji Iwana and Seiichi Uchida. 2021. Time series data augmentation for neural networks by time warping with a discriminative teacher. In 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 3558–3565.

[28] James Joyce. 2003. Bayes' theorem. (2003).

[29] 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 on Web Conference 2024. 2795–2806.

[30] Tomoyoshi Kimura, Jinyang Li, Tianshi Wang, Denizhan Kara, Yizhuo Chen, Yigong Hu, Ruijie Wang, Maggie Wigness, Shengzhong Liu, Mani Srivastava, et al. 2024. On the Efficiency and Robustness of Vibration-based Foundation Models for IoT Sensing: A Case Study. arXiv preprint arXiv:2404.02461 (2024).

[31] Lingjing Kong, Martin Q. Ma, Guangyi Chen, Eric P. Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. 2023. Understanding Masked Autoencoders via Hierarchical Latent Variable Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 7918–7928.

[32] Gang Li, Heliang Zheng, Daqing Liu, Bing Su, and Changwen Zheng. 2022. SemMAE: Semantic-Guided Masking for Learning Masked Autoencoders. ArXiv (2022). https://consensus.app/papers/semmae-semanticguided-masking-learning-masked-li/db05155d3afd56958dd95cdf3d84e0d2/?utm_source=chatgpt

[33] Jinyang Li, Yizhuo Chen, Tomoyoshi Kimura, Tianshi Wang, Ruijie Wang, Denizhan Kara, Yigong Hu, Li Wu, Walid A. Hanafy, Abel Souza, Prashant Shenoy, Maggie Wigness, Joydeep Bhattacharyya, Jae Kim, Guijun Wang, Greg Kimberly, Josh Eckhardt, Denis Osipychev, and Tarek Abdelzaher. 2024. Acies-OS: A Content-Centric Platform for Edge AI Twinning and Orchestration. In 2024 33rd International Conference on Computer Communications and Networks (ICCCN). Big Island, HI, 1–1.

[34] Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang. 2022. Uniform Masking: Enabling MAE Pre-training for Pyramid-based Vision Transformers with Locality. ArXiv abs/2205.10063 (2022). https://consensus.app/papers/uniform-masking-enabling-pretraining-pyramidbased-li/c8f23e288bc05c088d6f575ec595d641/?utm_source=chatgpt

[35] Dongxin Liu. 2022. Self-supervised learning frameworks for IoT applications. Ph. D. Dissertation.

[36] Dongxin Liu, Tianshi Wang, Shengzhong Liu, Ruijie Wang, Shuochao Yao, and Tarek Abdelzaher. 2021. Contrastive self-supervised representation learning for sensing signals from the time-frequency perspective. In 2021 International Conference on Computer Communications and Networks (ICCCN). IEEE, 1–10.

[37] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher. 2024. FOCAL: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space. Advances in Neural Information Processing Systems 36 (2024).

[38] Wei Liu, S. Cao, Y. Chen, and S. Zu. 2016. An effective approach to attenuate random noise based on compressive sensing and curvelet transform. Journal of Geophysics and Engineering 13 (2016), 135–145. https://doi.org/10.1088/1742-2132/13/2/135

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision. 10012–10022.

[40] Zhaolun Liu, K. Lu, and X. Ge. 2018. Convolutional Sparse Coding for Noise Attenuation of Seismic Data. SEG 2018 Workshop: SEG Maximizing Asset Value Through Artificial Intelligence and Machine Learning, Beijing, China, 17–19 September 2018 (2018). https://doi.org/10.1190/AIML2018-02.1

[41] Shenghuan Miao, Ling Chen, and Rong Hu. 2024. Spatial-Temporal Masked Autoencoder for Multi-Device Wearable Human Activity Recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 7, 4 (2024), 1–25.

[42] Shailesh Mohine, Babankumar S Bansod, Pravendra Kumar, Rakesh Bhalla, and Anshul Basra. 2020. Single acoustic sensor-based time–frequency spectrum sensing approach for land vehicle detection. IEEE Sensors Journal 20, 13 (2020), 7275–7282.

[43] D. Moreau, B. Cazzolato, A. Zander, and C. Petersen. 2008. A Review of Virtual Sensing Algorithms for Active Noise Control. Algorithms 1 (2008), 69–99. https://doi.org/10.3390/a1020069

[44] N. Nakata, R. Snieder, T. Tsuji, K. Larner, and T. Matsuoka. 2011. Shear wave imaging from traffic noise using seismic interferometry by cross-coherence. Geophysics 76 (2011). https://doi.org/10.1190/geo2010-0188.1

[45] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. 2022. Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation. arXiv preprint arXiv:2204.12260 (2022).

[46] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. 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. 324–337. https://doi.org/10.1145/3495243.3560519

[47] Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Natasa Sladoje. 2020. CoMIR: Contrastive multimodal image representation for registration. Advances in neural information processing systems 33 (2020), 18433–18444.

[48] Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S Melo, Ana Paiva, and Danica Kragic. 2022. Geometric Multimodal Contrastive Representation Learning. In International Conference on Machine Learning. 17782–17800.

[49] Mashud Rana, Ashfaqur Rahman, and Daniel V. Smith. 2023. A Semi-supervised Approach for Activity Recognition from Indoor Trajectory Data. ArXiv abs/2301.03134 (2023). https://doi.org/10.48550/arXiv.2301.03134

[50] X. Rao, Xiangsheng Zhu, M. Gao, F. Kan, and T. Zhong. 2020. Weak moving target coherent integration detection based on compressed sensing. IET International Radar Conference (IET IRC 2020) 2020 (2020), 1750–1757. https://doi.org/10.1049/icp.2021.0517

[51] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In 2012 16th international symposium on wearable computers. IEEE, 108–109.

[52] Carlos Rivem-Moreno and Boris Escalante-Ramires. 1996. Seismic signal detection with time-frequency models. Proceedings of Third International Symposium on Time-Frequency and Time-Scale Analysis (TFTS-96) (1996), 345–348. https://doi.org/10.1109/TFSA.1996.547484

[53] Diego Seuret-Jiménez, Eduardo Trutié-Carrero, José Manuel Nieto-Jalil, Erick Daniel García-Aquino, Lorena Díaz-González, Laura Carballo-Sigler, Daily Quintana-Fuentes, and Luis Manuel Gaggero-Sager. 2023. Feature Extraction of a Non-Stationary Seismic–Acoustic Signal Using a High-Resolution Dyadic Spectrogram. Sensors 23, 13 (2023), 6051.

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

[55] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. 2020. Exploring contrastive learning in human activity recognition for healthcare. arXiv preprint arXiv:2011.11542 (2020).

[56] DW Thomas and Brian R Wilkins. 1972. The analysis of vehicle sounds for recognition. Pattern Recognition 4, 4 (1972), 379–389.
[57] Li Tian, Yuan Cheng, and Zhibin Li. 2022. Pseudo Random Masked AutoEncoder for Self-supervised Learning. In Proceedings of the 2022 6th International Conference on Video and Image Processing. https://doi.org/10.1145/3579109.3579133
[58] Yijun Tian, Kaiwen Dong, Chunhui Zhang, Chuxu Zhang, and N. Chawla. 2022. Heterogeneous Graph Masked Autoencoders. ArXiv (2022). https://consensus.app/papers/graph-masked-autoencoders-tian/08183c5489f4545183671d4adee15e98/?utm_source=chatgpt
[59] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. Springer, 776–794.
[60] Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. 2021. Unsupervised Representation Learning for Time Series with Temporal Neighborhood Coding. In International Conference on Learning Representations. https://openreview.net/forum?id=8qDwejCuCN
[61] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. arXiv preprint arXiv:2203.12602 (2022).
[62] Xiao Wang and Guo-Jun Qi. 2022. Contrastive learning with stronger augmentations. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).
[63] Roy E. White. 1984. Signal and noise estimation from seismic reflection data using spectral coherence methods. Proc. IEEE 72 (1984), 1340–1356. https://doi.org/10.1109/PROC.1984.13022
[64] Alex Wilf, Syeda Akter, Leena Mathur, Paul Liang, Sheryl Mathew, Mengrou Shou, Eric Nyberg, and Louis-Philippe Morency. 2023. Difference-Masking: Choosing What to Mask in Continued Pretraining. In Findings of the Association for Computational Linguistics: EMNLP 2023. 13222–13234.
[65] Hongwei Wu and Jerry M Mendel. 2007. Classification of battlefield ground vehicles using acoustic features and fuzzy logic rule-based classifiers. IEEE transactions on fuzzy systems 15, 1 (2007), 56–72.
[66] Yiwei Xia, Jun Ma, ChuYue Yu, XunHuan Ren, Boriskevich Anatoliy Antonovich, and Viktar Yurevich Tsviatkou. 2022. Recognition system of human activities based on time-frequency features of accelerometer data. In 2022 International Conference on Intelligent Systems and Computer Vision (ISCV). IEEE, 1–5.
[67] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: A simple framework for masked image modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9653–9663.
[68] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems. 220–233.
[69] Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. 2023. SimPer: Simple Self-Supervised Learning of Periodic Targets. In International Conference on Learning Representations. https://openreview.net/forum?id=EKpMeEV0hOo
[70] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. 2022. Ts2vec: Towards universal representation of time series. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. 8980–8987.
[71] Qi Zhang, Yifei Wang, and Yisen Wang. 2022. How mask matters: Towards theoretical understandings of masked autoencoders. Advances in Neural Information Processing Systems 35 (2022), 27127–27139.
[72] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. 2022. Self-Supervised Contrastive Pre-Training For Time Series via Time-Frequency Consistency. In Proceedings of Neural Information Processing Systems, NeurIPS.
[73] Yue Zhang, Zhizhang Hu, Uri Berger, and Shijia Pan. 2023. 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. 96–109. https://doi.org/10.1145/3583120.3586960
[74] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. 2021. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 1450–1459.