

# On the Efficiency and Robustness of Vibration-based Foundation Models for IoT Sensing: A Case Study

Tomoyoshi Kimura<sup>†</sup>, Jinyang Li<sup>†</sup>, Tianshi Wang<sup>†</sup>, Denizhan Kara<sup>†</sup>, Yizhuo Chen<sup>†</sup>, Yigong Hu<sup>†</sup>, Ruijie Wang<sup>†</sup>, Maggie Wigness<sup>\*</sup>, Shengzhong Liu<sup>‡</sup>, Mani Srivastava<sup>§</sup>, Suhas Diggavi<sup>§</sup>, Tarek Abdelzaher<sup>†</sup>

<sup>†</sup>University of Illinois at Urbana-Champaign, <sup>\*</sup>DEVCOM Army Research Laboratory

<sup>‡</sup>Shanghai Jiao Tong University, <sup>§</sup>University of California, Los Angeles

{tkimura4, jinyang7, tianshi3, kara4, yizhuoc, yigongh2, ruijiew2}@illinois.edu,

{maggie.b.wigness}.civ@army.mil, shengzhong@sjtu.edu.cn,

{mbs, suhas}@ee.ucla.edu, zaher@illinois.edu

**Abstract**—This paper demonstrates the potential of vibration-based Foundation Models (FMs), pre-trained with *unlabeled sensing data*, to improve the robustness of run-time inference in (a class of) IoT applications. A case study is presented featuring a vehicle classification application using acoustic and seismic sensing. The work is motivated by the success of foundation models in the areas of natural language processing and computer vision, leading to generalizations of the FM concept to other domains as well, where significant amounts of unlabeled data exist that can be used for self-supervised pre-training. One such domain is *IoT applications*. Foundation models for selected sensing modalities in the IoT domain can be pre-trained in an environment-agnostic fashion using available *unlabeled* sensor data and then fine-tuned to the deployment at hand using a small amount of labeled data. The paper shows that the pre-training/fine-tuning approach improves the robustness of downstream inference and facilitates adaptation to different environmental conditions. More specifically, we present a case study in a real-world setting to evaluate a simple (vibration-based) FM-like model, called FOCAL, demonstrating its superior robustness and adaptation, compared to conventional supervised deep neural networks (DNNs). We also demonstrate its superior convergence over supervised solutions. Our findings highlight the advantages of vibration-based FMs (and FM-inspired self-supervised models in general) in terms of inference robustness, runtime efficiency, and model adaptation (via fine-tuning) in resource-limited IoT settings.

**Index Terms**—Foundation Model, Internet of Things

## I. INTRODUCTION

The paper presents a real-world case study of a target classification application, based on seismic and acoustic sensing, that demonstrates how a self-supervised neural network model pre-trained with *unlabeled sensor data* (using pre-training techniques common to foundation models [1]) can significantly improve run-time inference robustness and adaptation. Modalities, such as acoustic or seismic sensing, are particularly sensitive to environmental factors. Even in the same application domain, such as target tracking, a target (e.g., some vehicle on a road) may generate different sensory signatures depending on a variety of factors, such as the type of terrain (paved road, gravel, sand, ...), background noise

(rain, wind, construction, traffic, ...), and other natural and/or human disturbances. Training an inference task (e.g., a target classifier) to handle all such contingencies is a daunting undertaking. Inspired by pre-training solutions used for foundation models, can one pre-train a general target-independent and environment-independent model once, based on large amounts of unlabeled data (henceforth called a *foundation model*), then fine-tune it in a very light-weight fashion to each deployment environment and set of targets of interest?

Early *supervised* solutions for intelligent IoT applications are label-hungry due to the large sizes of modern deep neural networks (DNNs) that call for commensurately large volumes of (labeled) input training data. In the absence of sufficient amounts of labeled data, supervised neural-network training techniques suffer from overfitting, thereby dramatically reducing the robustness of run-time inference [2]. In contrast, by obviating the need for labeled data in pre-training (and requiring only small amounts of labeled data for fine-tuning), foundation models developed for intelligent IoT applications can improve inference robustness and adaptation to domain shifts and environmental noise.

Unlike supervised training techniques that directly teach a neural network *how to perform a particular inference task*, a foundation model is the output of (pre-)training that aims to teach the neural network *a better internal representation of domain-specific data*. By empirically learning statistical properties and patterns found in large domain-specific datasets, such an internal representation encodes (empirical approximations of) higher-level semantics or “knowledge” of the domain. Clearly, the degree to which such outcomes can be elicited depends on the amount of data used. Three important features thus characterize the pre-training of foundation models. First, it is *self-supervised*; no labeled data are needed. Second, it is *task-agnostic*; it does not know the downstream inference task(s) and, as such, can in principle support several different tasks, deployments, or environments. Finally, it generally uses a large amount of (unlabeled) data. For the sake of a proof of concept, we sacrifice the last property a bit in this study.

The feasibility of pre-training in the absence of labeled data and without knowing the exact downstream task(s) makes the approach attractive to IoT applications, especially from a robustness perspective. Interestingly, despite the use of a smaller (and thus more manageable) amount of pre-training data in this paper, the robustness advantages of the resulting model are still possible to illustrate. We show that the pre-trained model can be fine-tuned with only a minimal amount of labeled data for a specific downstream deployment, allowing more robust classification than baseline (supervised) approaches.

Another challenge for IoT sensing is the computational limitations in IoT devices. Rapid advances in computational resources have led to increasingly large DNNs [3]. However, many IoT devices remain limited by their resource constraints [4]. These devices, from simple sensors to complex wearables, often lack the necessary processing power, memory, and energy efficiency to support the training and operation of large-scale DNNs in real-time. This discrepancy poses significant challenges for deploying and training advanced DNNs in IoT applications [5], introducing bottlenecks to the model performance on IoT devices. We show that the pre-trained model we use is capable of execution on a Raspberry-Pi class of devices in real-time with a higher fine-tuning convergence rate while offering more robust performance than its supervised counterparts.

The rest of the paper is organized as follows. We cover a brief background on foundation model pre-training and the specific model used in this paper in Section II. We describe our case study and experimental set-up in Section III. Section IV presents the evaluation results, followed by discussion in Section V. Section VI covers related work. Section VII concludes the paper.

## II. SELF-SUPERVISED MODEL PRE-TRAINING

While many techniques were proposed recently for self-supervised pre-training of foundation models, two are particularly widespread: learning to reconstruct masked [6], [7] (or distorted [8]) inputs and contrastive learning [9], [10], [11], [12]. They differ in the way they train the model useful concepts from the domain, without the need for labeled data. Specifically, masking/distortion removes/distorts parts of the input, and then rewards the model for the correct reconstruction of these parts. Clearly, a model that learns correct reconstruction must have encapsulated some knowledge about the target domain. Contrastive learning teaches the model what “similarity” means in the target domain (by contrasting similar and dissimilar sample pairs), such that similar inputs are grouped closer together in a latent space. To do so without labels, it often relies on semantics-invariant input transformations that convert individual input samples to “similar” ones (without necessarily knowing what the sample labels or semantics are). An example of such transformations in vision is image resizing. An example in time-series data is adding simulated noise. The result of rewarding the model for putting similar samples closer together in the latent space is a well-organized learned latent representation, where proximity implies semantic similarity.

In this paper, for pre-training, we use a contrastive learning framework, called FOCAL [1], recently proposed for (pre-training in) intelligent multimodal sensing applications. FOCAL pre-trains *an encoder* to extract a structured latent representation of the input multimodal sensing data. This latent representation separates shared and private subspaces. The shared subspace contains common information shared across the different sensing modalities. The private subspaces hold additional modality-exclusive information. An orthogonality constraint is applied among the private subspaces, as well as between each private subspace and the shared subspace to enforce information independence among these subspaces. A pre-trained encoder is fine-tuned by appending a single linear layer whose weights are adapted to the downstream use scenario (using a small amount of task-specific labeled data).

We utilize FOCAL to train two popular DNN encoders (DeepSense [13] and SWIN-Transformer [14]) on a multi-modal Moving Object Detection [1] (MOD) dataset that consists of acoustic and seismic signals. Then, we perform a two-day experiment in a real-world neighborhood as a case study to examine the performance of FOCAL against supervised counterparts. The experimental setting and results are described below.

## III. AN EXPERIMENTAL STUDY

Our experiment was conducted at an outdoor research facility located on (repurposed) state park grounds. Sensors were deployed and vehicles were driven past the sensors over a period spanning two days. On the first day, the environment surrounding the experiment was controlled and disturbance-free. The second day featured significant interference (as described in the next section).

FOCAL was *pre-trained* on a previously published dataset [1] collected from acoustic and seismic sensors, deployed in different urban and rural environments that varied in terrain (paved, gravel, dirt, rooftop, etc) and environmental conditions (quiet, windy, etc), recording the passage of a variety of target types, mostly focusing on civilian automobiles, bikes, and humans. The pre-training data did *not* include any from the deployment reported in this paper. To experiment with the robustness of the pre-trained model, we *fine-tune* it on part of the data collected in the new deployment and *test* the fine-tuned model’s performance under the same or different deployment conditions. A comparison is carried out with supervised approaches.

We show that FOCAL is more robust to domain changes than other methods. Additionally, we demonstrate FOCAL’s superiority in terms of label efficiency by comparing performance under different amounts of labels used for fine-tuning. Finally, we present our findings on the computational efficiency of FOCAL and explore the potential applications for real-time IoT systems.

### A. The Experimental Setup

The experiment was performed using four deployed multimodal sensor nodes. Figure 1 shows a satellite view of

TABLE I

TRAINING CONFIGURATIONS: BELOW, WE DETAIL THE TRAINING PARAMETERS INCLUDING THE BATCH SIZE (NUMBER OF SAMPLES PER BATCH), THE OPTIMIZER FOR UPDATING MODEL PARAMETERS, THE INITIAL LEARNING RATE (LR), AND THE LR SCHEDULER FOR DYNAMIC LR ADJUSTMENTS, ALONGSIDE ITS LR DECAY RATE. THE TABLE ALSO LISTS THE TOTAL TRAINING EPOCHS AND THE DATA AUGMENTATIONS APPLIED.

Stage	Batch Size	Optimizer	Initial LR	LR Scheduler	LR Decay	Epochs	Augmentations
Supervised	128	AdamW [15]	1e-4	Cosine [16]	0.2	500	Mixup, Phase Shift
Pretrain	256	AdamW [15]	0.0001	Cosine [16]	0.05	6000	Permutation, Negation, Time Warp, Horizontal Flip, Magnitude Warp, Scaling, Phase Shift
fine-tune	256	Adam [17]	1e-3	Cosine [16]	0.2	200	Mixup, Phase Shift

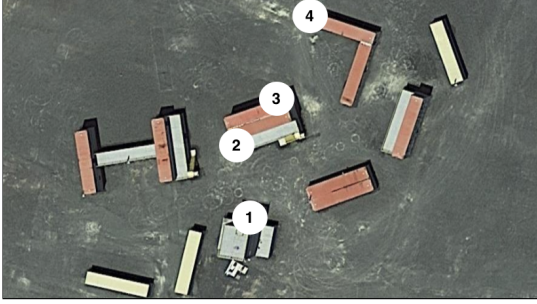


Fig. 1. The satellite view of the case study neighborhood with labeled nodes.

the facility and the four locations where we set up the sensor nodes. Nodes 1 & 4 utilized the RaspberryShake<sup>1</sup> 4D, while Nodes 2 & 3 utilized the RaspberryShake 1D. Each node featured a geophone and a microphone array, collecting seismic and acoustic vibration signals from nearby objects. In each run, a specific target navigated the neighborhood, passing the sensors in some arbitrary order within a short time window. Four distinct target types were used: (i) a Polaris<sup>2</sup> off-road vehicle, (ii) a Warthog<sup>3</sup> all-terrain unmanned ground robot, (iii) a Husky unmanned outdoor field robot<sup>4</sup>, and (iv) a standard civilian automobile. We collected 23 runs in total, each lasting approximately 10 minutes.

Although the sensors and targets were identical on both days, the sensor data distributions of the two days varied substantially due to different on-site events. On the first day, we conducted a controlled test run with only our operators around. On the second day, 5-6 research groups simultaneously worked on multiple experiments, which led to increased interference. Individuals walked and talked near our sensors, introducing human-related (acoustic and seismic) noise. Loud motor-powered generators were used by some teams creating additional acoustic and seismic interference. Strong wind on the second day further added to environmental disturbances. Thus, we partition the data collected by day. We refer to data collected on the first day as the *Control Set* and data collected on the second day as the *Noisy Set*.

## B. Datasets

We also consider the MOD dataset released in [1]. This dataset contains multi-modal seismic and acoustic signals

describing nearby moving objects. Therefore, we have three sets of data subject to varying distribution shifts — *MOD*, *Control Set*, and *Noisy Set*. We follow the same setup as [1] to process these three sets with a 0.2-second overlapping ratio between 2 seconds samples of 8000Hz acoustic 100Hz seismic data. We then partition MOD into a set of unlabeled data used to pre-train the FM and a set of labeled data for supervised training and fine-tuning. The distribution of the MOD dataset is significantly different compared to the other two sets due to distinct locations, sensor placements, and moving targets. The Control Set and the Noisy Set, though have similar targets, have different distributions as well due to runtime conditions.

## C. Training Pipelines

1) *Training Frameworks*: We choose FOCAL [1] as our self-supervised training framework. We use FOCAL with two different backbone encoders:

- **DeepSense** [13] is a DNN classifier designed for time-series sensory inputs. It applies convolution layers on modality spectrograms to extract general features and then utilizes recurrent layers (stacked GRU) to further extract global temporal relationships.
- **SWIN-Transformer (SW-T)** [14] is a variant of Vision Transformer (ViT)[18], proposing to extract a hierarchical representation through downsampling and shifting window operations. Similar to ViT, it partitions the sample into patches. What makes SW-T different from ViT is that SW-T groups different patches into non-overlapping windows and computes self-attention within each window to minimize computation costs. These windows are further shifted to take advantage of the cross-window connection.

2) *Pretraining*: We pretrain FOCAL with the unlabeled set from the MOD dataset. We randomly apply different time and frequency augmentations in the time domain and frequency domain. We use STFT to convert each modality sample into the frequency domain and then extract the modality embedding. Training configurations used during pre-training are presented in Table I.

3) *Supervised Training/fine-tuning*: In the fine-tuning stage, we use labeled samples to perform supervised fine-tuning on the pretrained model. We freeze the pretrained model and add a linear layer for target classification (from the concatenated modality embeddings). We would like to note that only the linear layer is trained at the fine-tuning stage. During fine-tuning, We apply mixup [19] augmentation in the time domain and phase shift augmentation in the frequency domain. We

<sup>1</sup><https://raspberrysake.org/>

<sup>2</sup><https://www.polaris.com/>

<sup>3</sup><https://clearpathrobotics.com/warthog-unmanned-ground-vehicle-robot/>

<sup>4</sup><https://clearpathrobotics.com/husky-unmanned-ground-vehicle-robot/>

also separately train supervised DNNs for the two backbone encoders as the benchmarks. The supervised model contains an additional fusion layer to fuse the modality embeddings for classification. Training configurations for fine-tuning and supervised benchmark can be found in Table I. We also use a supervised model initially trained on the MOD dataset and later fine-tuned on its final classification layer, mirroring FOCAL’s fine-tuning approach. We call it the supervised-fine-tuned baseline.

#### IV. EVALUATION RESULTS

Below, we examine FOCAL performance after fine-tuning with some target domain data then compare the computational efficiency of the supervised and the Foundation models.

##### A. Model Retraining/fine-tuning

We divide the Control Set into training, validation, and testing data with a ratio of 8:1:1. We train and fine-tune the models using different amounts of labeled samples from the training data of the Control Set (100%, 50%, 10%, 1%) and then evaluate their performance on the testing data from the same set. Table II summarizes the performance of the retrained DNNs on the Control Set, under different label ratios. When the amount of labeled data used is high (100% or 50%), the supervised approaches work well. In fact, they slightly outperform FOCAL (that tunes its last layer only). However, as the amount of labeled data decreases (10% and 1%), the supervised approaches degrade substantially, whereas FOCAL suffers a much lower penalty in performance, suggesting a higher label efficiency. Note also that the supervised-fine-tuned benchmark is dominated by FOCAL, offering no advantage across the board. The gap between the supervised-fine-tuned benchmark and FOCAL underscores FM’s ability to encode prior knowledge much more robustly than the supervised models. As such, it can be more easily adapted to various IoT deployment conditions with a minimal amount of labels.

Next, we use the models trained on the Control Set and evaluate them on the Noisy Set. While the Control Set and the Noisy Set contain identical target objects collected on the same set of sensors, the dynamic nature of the IoT System can still affect the collected data distribution. We examine such domain shift effect in Table III. As we lower the label ratio, supervised models experience significant degradation, whereas FOCAL remains relatively stable. As before, FOCAL dominates the supervised-fine-tuned approach, suggesting that its self-supervised pre-training has better knowledge transfer.

##### B. Training Efficiency

In this section, we compare the training efficiency of the supervised models and the fine-tuning efficiency of FOCAL (which we refer to as “training” efficiency as well, for the sake of brevity, below). We define the training efficiency as the convergence speed or the number of training epochs needed for convergence. As shown in Table II, both the supervised model and FOCAL perform well after training on the Control Set. We compare their convergence speed by observing the training

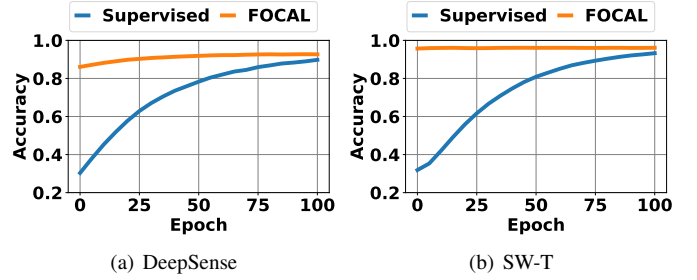


Fig. 2. Accuracy curves of Supervised Training and FOCAL fine-tuning.

accuracy curves in Figure 2 during the first 100 epochs. On both backbone encoders, FOCAL (fine-tuning) converges much faster with near-optimal performance achieved in the first few epochs, compared to training the supervised model. This shows that the pre-trained representation is useful for the downstream task and can easily transfer knowledge to achieve high performance in a short time. On the other hand, since the supervised models are trained from scratch, they begin at a lower accuracy and with more parameters to train (FOCAL only updates the linear classification layer during fine-tuning, as opposed to the supervised benchmark that trains all its parameters). Thus, the supervised algorithm approaches FOCAL performance only towards the end of the 100 epochs. We do not consider the supervised-fine-tune benchmarks since they are dominated by the others. While not shown, FOCAL also requires much less memory to fine-tune its single last layer, compared to retraining an entire supervised model from scratch.

#### V. DISCUSSION

The experimental study reported in this paper suggests that the task-agnostic nature of pre-training of self-supervised models endows them with greater robustness, making them ideally suited for IoT application deployment across various environments with only limited fine-tuning needed to achieve high-quality inference. Unlike traditional supervised models, these pre-trained models exploit large amounts of unlabeled data, leading to enhanced resilience against domain shifts. Although this paper only leverages a small-scale unlabeled dataset for pre-training, the pre-trained models already exhibit characteristics of foundation models with great robustness in different domains. This is particularly useful in IoT sensing scenarios where different sensor deployments (even within the same application) may be subjected to vastly different conditions. For example, target tracking using acoustic/vibration sensing will see significant distributional shifts across urban areas, rural roads, freeways, gravel parking lots, etc, as well as across different weather conditions (wind, snow, rain), and different target types. Pre-training the foundation model with a larger scale dataset and larger backbone encoders could also potentially improve its downstream robustness, and we leave that to future work.

The high label efficiency of pre-trained models further facilitates their rapid deployment to a wide array of downstream tasks, where label scarcity is a critical challenge. These models

TABLE II  
FINE-TUNE RESULTS ON THE CONTROL SET

Label Ratio		100%		50%		10%		1%	
Encoder	Framework	Acc	F1	Acc	F1	Acc	F1	Acc	F1
DeepSense	Supervised	<b>0.9684</b>	<b>0.9637</b>	<b>0.9425</b>	<b>0.9328</b>	0.8078	0.7714	0.5247	0.5019
	Supervised-fine-tune	0.7933	0.7578	0.7762	0.7379	0.7383	0.6892	0.5974	0.5392
	FOCAL	0.9330	0.9293	0.9204	0.9154	<b>0.8976</b>	<b>0.8893</b>	<b>0.8078</b>	<b>0.7876</b>
SW-T	Supervised	<b>0.9842</b>	<b>0.9840</b>	<b>0.9608</b>	<b>0.9589</b>	0.7434	0.7107	0.3660	0.2802
	Supervised-fine-tune	0.6372	0.5829	0.6327	0.5778	0.6056	0.5592	0.5607	0.5037
	FOCAL	0.9526	0.9473	0.9558	0.9524	<b>0.9425</b>	<b>0.9372</b>	<b>0.8312</b>	<b>0.8176</b>

TABLE III  
TEST RESULTS ON THE NOISY SET

Label Ratio		100%		50%		10%		1%	
Encoder	Framework	Acc	F1	Acc	F1	Acc	F1	Acc	F1
DeepSense	Supervised	<b>0.6769</b>	<b>0.6843</b>	<b>0.6603</b>	<b>0.6639</b>	0.5805	0.5764	0.4688	0.4919
	Supervised-fine-tune	0.5766	0.5735	0.5689	0.5650	0.5539	0.5458	0.4358	0.4041
	FOCAL	0.6558	0.6640	0.6515	0.6601	<b>0.6578</b>	<b>0.6634</b>	<b>0.6101</b>	<b>0.6153</b>
SW-T	Supervised	0.5454	0.5397	0.5126	0.5040	0.4180	0.3962	0.2838	0.2157
	Supervised-fine-tune	0.4179	0.3968	0.4149	0.3944	0.4072	0.3883	0.3862	0.3527
	FOCAL	<b>0.6641</b>	<b>0.6788</b>	<b>0.6742</b>	<b>0.6819</b>	<b>0.6924</b>	<b>0.7050</b>	<b>0.5549</b>	<b>0.5508</b>

exhibit exceptional adaptability to varying physical environments, which makes them suitable to meet the demands of real-time CPS. Merely training a single linear layer in FOCAL, for fine-tuning, can easily reach optimal performance within a few epochs. This efficiency not only enhances the practicality of FMs in dynamic settings but also opens opportunities for on-device training, making it feasible to train FMs on resource-constrained IoT devices.

## VI. RELATED WORK

Deep Learning has catalyzed significant advances in inference from IoT sensing data [20], with DNNs becoming integral to a wide range of IoT applications [21], [22]. However, domain-specific challenges still lead to many limitations in building robust DNNs for IoT sensing. Deployed DNNs must handle unpredictable interference in the field that greatly alters the statistical distribution of collected sensor data. The altered distribution, or *domain shift*, can significantly degrade DNN performance, leading to inaccurate results and potentially severe consequences.

More recently, Foundation Models (FMs) [23] have gained increasing popularity, most notably in language [6], [24] and vision [7], [25]. The techniques were then generalized to other areas where domain-specific FMs emerged, such as security [26], [27], networking [28], [29], and meteorology [30].

Contrastive Learning (CL) [1], [31], [32], [33], [34], [35] has been a popular form of SSL to extract a robust embedding space during pre-training. The main idea is to pull similar samples closer while pushing other samples further apart in the embedding space. Unimodal CL frameworks like [36], [31] apply random augmentations to learn transformation invariant information. [37], [33], [32] are multi-modal CL frameworks that enforce cross-modal consistency. CL for time series has also been extensively studied in [34], [35].

Improving resilience against domain shifts has been widely studied in recent years. [38], [39] investigate improving the efficiency of unsupervised domain adaptation for IoT applications. However, these works primarily consider classifiers trained in a supervised manner. Others have also worked on Federated Learning-based domain generalization [40], [41]. Numerous works analyze SSL for domain generalization [42], [43], but less has been explored for IoT applications.

## VII. CONCLUSIONS

In this paper, we examined an FM-based approach, specifically FOCAL, against conventional supervised models in the context of IoT sensing. Through our real-world case study, we have demonstrated how Foundation Models require minimal domain-specific tuning while allowing robust real-time inference. Our results highlight promising opportunities for Foundation Models in the IoT landscape. Our future work will focus on developing more scalable Foundation Models for generalized IoT systems.

## VIII. ACKNOWLEDGEMENTS

Research reported in this paper was sponsored in part by DEVCOM ARL under Cooperative Agreement W911NF-17-2-0196 (ARL IoBT CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

## REFERENCES

- [1] S. Liu, T. Kimura, D. Liu, R. Wang, J. Li, S. Diggavi, M. Srivastava, and T. Abdelzaher, "Focal: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space," in *Advances in Neural Information Processing Systems*, 2023.

- [2] 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)*. IEEE, 2022, pp. 1082–1087.
- [3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [4] B. Chatterjee, N. Cao, A. Raychowdhury, and S. Sen, "Context-aware intelligence in resource-constrained iot nodes: Opportunities and challenges," *IEEE Design & Test*, vol. 36, no. 2, pp. 7–40, 2019.
- [5] S. Yao, Y. Zhao, H. Shao, S. Liu, D. Liu, L. Su, and T. Abdelzaher, "Fastdeepiot: Towards understanding and optimizing neural network execution time on mobile and embedded devices," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018, pp. 278–291.
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [7] 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*, 2022, pp. 16 000–16 009.
- [8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.
- [9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [10] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," *Advances in neural information processing systems*, vol. 33, pp. 5812–5823, 2020.
- [11] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka, "De-biased contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 8765–8775, 2020.
- [12] 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)*. IEEE, 2021, pp. 1–10.
- [13] 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 *International Conference on World Wide Web (WWW)*, 2017.
- [14] 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 *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2021.
- [15] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.
- [16] —, "Sgdr: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2016.
- [17] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.
- [18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.
- [19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.
- [20] M. Srivatsa, T. Abdelzaher, and T. He, Eds., *Artificial Intelligence for Edge Computing*. Springer, 2023.
- [21] B. Salehi, G. Reus-Muns, D. Roy, Z. Wang, T. Jian, J. Dy, S. Ioannidis, and K. Chowdhury, "Deep learning on multimodal sensor data at the wireless edge for vehicular network," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7639–7655, 2022.
- [22] V. Radu, C. Tong, S. Bhattacharya, N. D. Lane, C. Mascolo, M. K. Marina, and F. Kawsar, "Multimodal deep learning for activity and context recognition," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 1, no. 4, pp. 1–27, 2018.
- [23] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.
- [24] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.
- [25] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DinoV2: Learning robust visual features without supervision," 2023.
- [26] J. G. Almaraz-Rivera, J. A. Cantoral-Ceballos, and J. F. Botero, "Enhancing iot network security: Unveiling the power of self-supervised learning against ddos attacks," *Sensors*, vol. 23, no. 21, p. 8701, 2023.
- [27] Z. Zhang, S. Bu, Y. Zhang, and Z. Han, "Market-level integrated detection against cyber attacks in real-time market operations by self-supervised learning," *IEEE Transactions on Smart Grid*, 2024.
- [28] S. Zhang, O. T. Ajayi, and Y. Cheng, "A self-supervised learning approach for accelerating wireless network optimization," *IEEE Transactions on Vehicular Technology*, 2023.
- [29] M. S. Towhid and N. Shahriar, "Encrypted network traffic classification using self-supervised learning," in *2022 IEEE 8th International Conference on Network Softwarization (NetSoft)*. IEEE, 2022, pp. 366–374.
- [30] G. Mai, W. Huang, J. Sun, S. Song, D. Mishra, N. Liu, S. Gao, T. Liu, G. Cong, Y. Hu *et al.*, "On the opportunities and challenges of foundation models for geospatial artificial intelligence," *arXiv preprint arXiv:2304.06798*, 2023.
- [31] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2021.
- [32] 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 *International Conference on Mobile Computing And Networking (MobiCom)*, 2022.
- [33] 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 (ICML)*, 2022.
- [34] S. Tonekaboni, D. Eytan, and A. Goldenberg, "Unsupervised representation learning for time series with temporal neighborhood coding," in *International Conference on Learning Representations (ICLR)*, 2021.
- [35] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu, "Ts2vec: Towards universal representation of time series," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.
- [36] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020.
- [37] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European Conference on Computer Vision (ECCV)*, 2020.
- [38] J. Li, M. Jing, H. Su, K. Lu, L. Zhu, and H. T. Shen, "Faster domain adaptation networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5770–5783, 2021.
- [39] Y. Zhao, D. Saxena, and J. Cao, "Memory-efficient domain incremental learning for internet of things," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 1175–1181.
- [40] L. Zhang, X. Lei, Y. Shi, H. Huang, and C. Chen, "Federated learning for iot devices with domain generalization," *IEEE Internet of Things Journal*, 2023.
- [41] Y. Huang, M. Du, H. Zheng, and X. Feng, "Incremental unsupervised adversarial domain adaptation for federated learning in iot networks," in *2022 18th International Conference on Mobility, Sensing and Networking (MSN)*. IEEE, 2022, pp. 186–190.
- [42] I. Achituve, H. Maron, and G. Chechik, "Self-supervised learning for domain adaptation on point clouds," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 123–133.
- [43] J. Xu, L. Xiao, and A. M. López, "Self-supervised domain adaptation for computer vision tasks," *IEEE Access*, vol. 7, pp. 156 694–156 706, 2019.