Efficient Biomedical Time-series Contrastive Learning: Revisiting Taylor Expansion in Loss Optimization

Yunho Jeong[1†], Chanyu Moon[1†], Jinmo Kim[2], Ji-Woong Choi[123*]
[1]*Dept. of Artificial Intelligence, DGIST, Daegu, Republic of Korea*
[2]*Dept. of Electrical Engineering and Computer Science, DGIST, Daegu, Republic of Korea*
[3]*Brain Engineering Convergence Research Center, DGIST, Daegu, 42988, Republic of Korea*
{wjddbsgh4152, anscksdb0127, jmkim, jwchoi}@dgist.ac.kr

***Abstract*—Contrastive learning has become a cornerstone for self-supervised time-series representation learning, yet its practical application is often hindered by the quadratic computational and memory complexity of the standard Information Noise-Contrastive Estimation (InfoNCE) loss function. This bottleneck severely limits the use of large batch sizes and complex hierarchical models, such as TS2Vec, which are crucial for learning high-quality representations. To address this fundamental limitation, we leverage a first-order Taylor expansion to approximate the InfoNCE loss. This approach circumvents the explicit computation of the full similarity matrix, effectively reducing the complexity to a manageable linear scale. We integrated our Taylor-approximated loss into the TS2Vec framework and evaluated its performance on the Human Activity Recognition (HAR) benchmark, from University of California, Irvine (UCI) Machine Learning Repository. Our experiments demonstrate a substantial improvement in efficiency—achieving up to a 7.3x speedup and an 8.5x reduction in peak memory usage—while maintaining classification accuracy and discriminative power highly comparable to the original, resource-intensive model. By mitigating the prohibitive costs of large-batch training, our work enables the deployment of powerful time-series models in resource-constrained settings, paving the way for broader applications in fields like biomedical signal processing and federated learning.**

***Index Terms*—Time-Series Analysis, Contrastive Learning, Representation Learning, Loss Function Optimization, Taylor Expansion, Computational Efficiency**

INTRODUCTION

Contrastive learning has emerged as a dominant paradigm in self-supervised representation learning, fundamentally reshaping the acquisition of meaningful data representations from vast unlabeled datasets [1], [2]. This approach obviates the need for costly and labor-intensive manual annotation by training models to distinguish between similar and dissimilar data points. The core principle involves generating semantically related "positive pairs" and unrelated "negative pairs" through data augmentations [3], [4]. In the learned embedding space, the representations of positive pairs are encouraged to be proximal, while those of negative pairs are repelled. Central to this process is a contrastive loss function, with Information Noise-Contrastive Estimation (InfoNCE) being one of the most effective and widely adopted variants [2], [5]. By maximizing the mutual information between latent representations of positive pairs, InfoNCE enables the model to learn a highly structured embedding space that captures intricate data semantics. The resulting representations have demonstrated remarkable transferability, achieving state-of-the-art performance on various downstream tasks, often rivaling or surpassing fully supervised methods, particularly in data-scarce domains [3], [6].

Despite its empirical success, the standard implementation of InfoNCE-based contrastive learning is hampered by a significant computational bottleneck, namely its quadratic computational and memory complexity. To compute the loss for a mini-batch of size B, it is necessary to construct a $B \times B$ similarity matrix containing scores for all possible pairs, resulting in memory and computational costs that scale quadratically ($O(B^2)$) with the batch size. This quadratic scaling poses a critical barrier to scalability, as the performance of contrastive learning often correlates positively with the use of larger batch sizes, which provide a greater diversity of negative examples [3], [7]. This challenge is exacerbated in advanced hierarchical architectures such as TS2Vec, a universal framework for time series representation learning [8], where applying contrastive loss across multiple semantic levels compounds the memory burden.

While distributed training strategies that partition data and computation across multiple accelerators have been proposed to mitigate this issue [9], they are not universally applicable. Such methods are often infeasible in settings where privacy or hardware constraints mandate on-device computation, a common requirement in biomedical applications and federated learning [10], [11].

This motivates the need for a more fundamental solution. Inspiration can be drawn from recent work in the imaging domain, which has highlighted that contrastive learning performance is highly dependent on the penalty strength, or "hardness," applied to negative samples [16]. To investigate this relationship, their study utilized a simplified, non-exponential contrastive loss, enabling a direct comparison of

†These authors contributed equally to this work.
* Corresponding author: Ji-Woong Choi (jwchoi@dgist.ac.kr)

performance across different 'hardness' levels. But we point out that simplified, non-exponential contrastive loss has further advantages. By leveraging Taylor expansion to contrastive loss, computational burden can be addressed.

While previous studies have focused on the theoretical behavior of the InfoNCE loss, its practical application is often hindered by significant computational and memory costs. In this paper, we address this computational challenge by leveraging a first-order Taylor approximation of the InfoNCE loss, which effectively linearizes its complexity. We apply this computationally efficient loss to memory-intensive hierarchical time series models, such as TS2Vec, and conduct a comparative analysis of its time and space complexity and downstream task performance against the standard InfoNCE. Our experiments demonstrate that this approach yields substantial improvements in memory efficiency, allowing for training on a single resource-constrained device, and maintains the high quality of the learned representations, showing significant promise for biomedical time series analysis.

I. Related work

A. Time series contrastive learning

The success of contrastive learning in computer vision has inspired its application to time-series representation learning. Various methods have been proposed to adapt this paradigm to the unique characteristics of temporal data. Early approaches adapted triplet loss frameworks for time-series data to learn discriminative representations [12]. More recent works have developed sophisticated augmentation strategies and contrastive objectives. For instance, TS-TCC learns robust representations by performing temporal and contextual contrasting simultaneously [13], while CoST focuses on disentangling seasonal and trend components through a specialized contrastive objective [14].

Among these, TS2Vec [8] has emerged as a particularly powerful and universal framework, achieving state-of-the-art performance across various benchmarks. However, a common thread among these advanced models, including TS2Vec, is their reliance on a contrastive loss that requires pairwise similarity computations. As established in the Introduction, this shared mechanism leads to the foundational $O(B^2)$ complexity bottleneck that limits their scalability. Our work directly addresses this fundamental issue, proposing an efficient approximation that can benefit not only TS2Vec but also the broader landscape of time-series contrastive learning models.

B. TS2Vec

As our work directly optimizes the loss function used in TS2Vec [8], we use it as the baseline framework to validate Taylor-approximated method. TS2Vec is designed to learn universal representations for arbitrary time series from different domains without requiring domain-specific augmentations or model modifications.

1) Encoder: TS2Vec first generates two augmented views by randomly cropping overlapping subseries from an input time series. These subseries are then fed into a shared encoder network. The encoder consists of an input projection layer, a timestamp masking mechanism, and a stack of dilated convolution blocks to capture temporal dependencies across various receptive fields.

2) Positive Pairs: Positive pairs are defined based on the principle of contextual consistency. For two augmented subseries, only the representations corresponding to the same original timestamp are considered a positive pair. All other pairs, whether they are from different timestamps within the same subseries or from different timestamps between the two subseries, are treated as negative pairs.

3) Hierarchical Contrastive Loss: A key feature of TS2Vec is its hierarchical contrasting mechanism. The contrastive loss is computed not only at the final output of the encoder but also at intermediate layers. This forces the model to learn multi-scale contextual information, capturing both fine-grained and coarse-grained patterns within the time series. However, this hierarchical approach exacerbates the memory burden, as it requires computing multiple large similarity matrices.

II. Method

A. InfoNCE loss

Let $r, z \in \mathbb{R}^{B\times C}$ be two matrices of embeddings, where B is the batch size, C is the embedding dimension and $\mathbb{R}$ is a set of real number. The i-th embedding vector from each matrix is denoted by $r_i, z_i \in \mathbb{R}^C$.

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(\text{sim}(r_i, z_i))}{\sum_j^B \exp(\text{sim}(r_i, z_j))} \tag{1}$$

Typically, the similarity function $\text{sim}(x, y)$ is the dot product $x \cdot y$, assuming the vectors are L2-normalized. Thus, the conventional InfoNCE loss is expressed as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(r_i \cdot z_i)}{\sum_j^B \exp(r_i \cdot z_j)} \tag{2}$$

B. Loss Function Optimization by Taylor Expansion

However, a direct computation of the InfoNCE loss requires constructing a similarity matrix. Storing this matrix requires $O(B^2)$ space, which introduces a significant computational and memory bottleneck, especially for large batch sizes. To overcome this limitation, we leveraged a Taylor expansion of the InfoNCE loss to circumvent the need for an explicit similarity matrix.

The computational challenge arises from the denominator term in Eq. (2). The loss can be decomposed as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{i}^{B}(r_i \cdot z_i) + \frac{1}{B}\sum_{i}^{B}\log\{\sum_{j}^{B}\exp(r_i \cdot z_j)\} \tag{3}$$

We approximate the log-sum-exp term in Eq. (3) using a first-order Taylor expansion.

1) First-Order Taylor Expansion: We define a function $f(t)$ as:

$$f(t) = \log \sum_{j}^{B} \exp(st), \quad \text{where } s, t \in \mathbb{R} \tag{4}$$

The first derivative of $f(t)$ with respect to t is:

$$\frac{df(t)}{dt} = \frac{\sum_j^B s \exp(st)}{\sum_j^B \exp(st)} \tag{5}$$

The first-order Taylor expansion of $f(t)$ around the center point $t = 0$ is approximated as $f(t) \simeq f(0) + f'(0)t$. This gives us:

$$f(t) \simeq \log(B) + \left(\frac{\sum_j^B s}{B} \right) t \tag{6}$$

By setting $t = 1$ and substituting s with the dot product similarity $r_i \cdot z_j$, we obtain the approximation for the log-sum-exp term. Substituting this back into Eq. (3), the final approximated InfoNCE loss can be written as:

$$\mathcal{L}_{\text{ap}} = -\frac{1}{B} \sum_{i}^{B} (r_i \cdot z_i) + \frac{1}{B} \sum_{i}^{B} \left\{ \log(B) + \left(\frac{\sum_j^B r_i \cdot z_j}{B} \right) \right\} \tag{7}$$

$$= -\frac{1}{B} \sum_{i}^{B} (r_i \cdot z_i) + \frac{1}{B} \sum_{i}^{B} \log(B) + \frac{1}{B^2} \sum_{i}^{B} \sum_{j}^{B} (r_i \cdot z_j) \tag{8}$$

$$= -\frac{1}{B} \sum_{i}^{B} (r_i \cdot z_i) + \log(B) + \left(\frac{1}{B} \sum_{i}^{B} r_i \right) \cdot \left(\frac{1}{B} \sum_{j}^{B} z_j \right) \tag{9}$$

This final expression allows for the calculation of the approximate loss without explicitly constructing and storing the similarity matrix.

III. Result

A. Temporal/Spatial Efficiency

To rigorously evaluate the computational efficacy of Taylor-approximated Hierarchical Contrastive Loss, we conducted a series of benchmarks against the original formulation. Our primary experiment was designed to simulate common use cases by systematically varying the batch size (B) from 8 to 1024, while the sequence length (T) and feature dimension (C) were held constant at 256. Performance was assessed using two main metrics: execution time and maximum memory usage. The results of this benchmark are presented in Fig. 1.

The empirical results of the comprehensive benchmark, as illustrated in Fig. 1A and 1B, reveal a significant performance disparity between the two methods. The original loss formulation exhibits a super-linear growth trend in both execution time and memory consumption as the batch size increases. This trend becomes particularly acute beyond the X-Large (B=128) configuration, underscoring a critical scalability bottleneck. In contrast, our Taylor-approximated method demonstrates a substantially more favorable performance profile, maintaining significantly lower execution times and memory footprints across all tested configurations. The slight deviation from perfect linearity in the Taylor method's execution time (Fig. 1A) is attributed to the measurement encompassing the entire input-to-output process, whereas the core loss computation itself scales linearly. The near-linear growth in resource consumption for our method highlights its superior scalability and efficiency.

To quantify the relative advantages of our approach, we analyzed the speedup factor and memory reduction factor, presented in Fig. 1C and Fig. 1D, respectively. The speedup conferred by our method is directly correlated with the batch size. While the gains are modest for smaller configurations (ranging from 1.1x to 1.2x), they become increasingly pronounced at larger scales, culminating in a 7.3x speedup for the XXX-Large+ (B=1024) configuration (Fig. 1C). This non-linear improvement demonstrates the profound algorithmic advantage of our method in computationally intensive settings. The benefits are even more significant with respect to memory efficiency (Fig. 1D). Taylor approximated method achieves a minimum 2.0x memory reduction even at the smallest Tiny (B=8) configuration. This efficiency gain steadily amplifies with the batch size, reaching an 8.5x reduction at the largest scale.

These experimental findings empirically validate that our Taylor-approximated loss effectively mitigates the critical computational complexity and memory bottlenecks inherent to the standard contrastive loss formulation. Crucially, the amplification of these efficiency gains with increasing batch size confirms that our method enables the practical and feasible training of large-scale models, even within resource-constrained environments such as a single GPU.

B. Accuracy

This section aims to verify that the computational efficiency gains of Taylor-expanded methodology do not compromise classification performance. To this end, we conduct an evaluation on the UCI HAR dataset, a standard benchmark in time series classification [15]. This dataset comprises time-series data from smartphone accelerometers and gyroscopes, collected from 30 subjects performing six distinct activities: Walking, Upstairs, Downstairs, Sitting, Standing, and Laying.

Table I summarizes the quantitative benchmark results on the UCI HAR dataset. This significant gain in efficiency is achieved at the cost of a marginal decrease in classification performance. Key metrics such as Accuracy and AUPRC show a slight reduction of approximately 1-2%. However, we argue that this minor trade-off is practically acceptable, especially considering the substantial benefits in computational efficiency that enable the deployment of large-scale models in resource-constrained environments.

To further assess discrimination performance of the model, we analyzed the Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) scores, shown in Fig. 2C and Fig. 2D. The ROC curve illustrates a model's ability to distinguish between classes, with performance improving

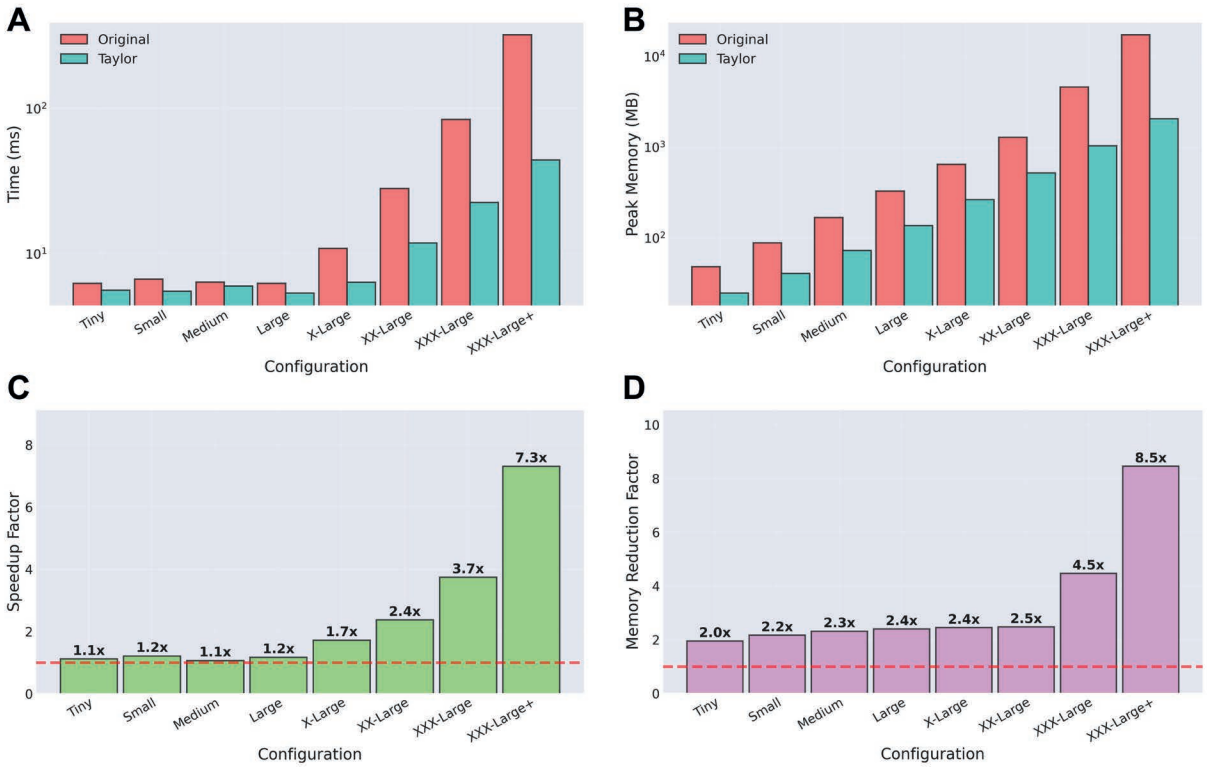

Fig. 1. Comparison of execution time and memory consumption across different batch sizes. (A) Execution time comparison, (B) Peak memory usage comparison, (C) Speedup factor, and (D) Memory reduction factor.

TABLE I
QUANTITATIVE BENCHMARK RESULTS ON THE UCI HAR DATASET

	Classification Performance		Computational Efficiency		
Method	**Accuracy**	**AUPRC**	**Time (s)**	**GPU Peak (MB)**	**Throughput (items/sec)**
Original	0.9213	0.9485	60.1	20576.0	122.4
Taylor	0.9087	0.9291	31.9	8214.0	230.6

as the curve approaches the top-left corner. For both models, the ROC curves for all six classes are tightly clustered in the top-left corner, with corresponding AUC scores approaching 1.0. This signifies excellent class separability. Critically, there are no discernible differences between the curves of the original model (Fig. 2C) and our Taylor-approximated model (Fig. 2D). This provides compelling evidence that the discriminative power of the learned representations is fully maintained.

The evaluation on the UCI HAR dataset demonstrates that Taylor-approximated method achieves substantial computational efficiency while maintaining classification accuracy and discriminative power that are highly comparable to the original model. This validates our approach as a practical and effective alternative, capable of delivering performance comparable to that of more resource-intensive models.

IV. CONCLUSION

In this paper, we address a critical computational bottleneck in contrastive learning for time-series analysis: the quadratic ($O(B^2)$) complexity inherent in the standard InfoNCE loss function. To overcome this limitation, we leverage a first-order Taylor approximation to estimate the loss, thereby circumventing the explicit computation and storage of the full similarity matrix. This approach successfully reduces computational and memory complexity to a manageable linear ($O(B)$) scale with respect to the batch size.

Our empirical evaluations demonstrate the efficacy and efficiency of Taylor-approximated contrastive loss which achieved a substantial speedup of up to 7.3x and a memory reduction of up to 8.5x compared to the original hierarchical contrastive loss in TS2Vec, confirming its superior scalability. Critically, this significant gain in computational efficiency was realized with only a marginal trade-off in performance. On the UCI HAR classification benchmark, it maintained discriminative power and accuracy highly comparable to the original, resource-intensive formulation.

By mitigating the prohibitive costs associated with large-batch training, our work shows the effective application of

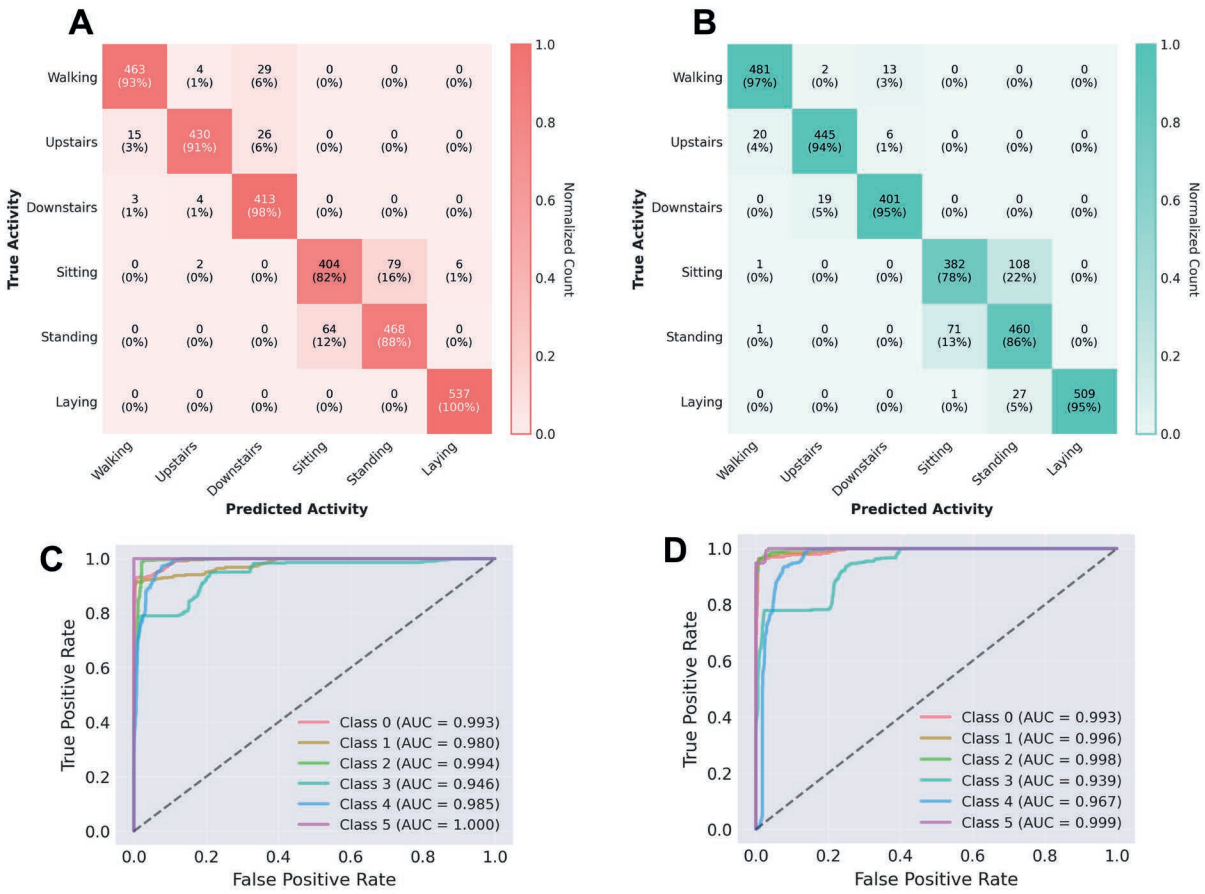

Fig. 2. Classification performance on the UCI HAR dataset. (A) Confusion matrix for the original model. (B) Confusion matrix for the Taylor-approximated model. (C) ROC curves for the original model. (D) ROC curves for the Taylor-approximated model.

powerful contrastive learning models in resource-constrained environments, such as on a single GPU. This advancement holds considerable promise for democratizing access to state-of-the-art time-series representation learning and broadens its applicability in domains where on-device processing is crucial for data privacy and real-time responsiveness, including biomedical and wearable sensor data analysis.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00415347, RS-2024-00428887 and RS-2024-00442085).

References

[1] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 9729–9738.

[2] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 1597–1607.

[4] J.-B. Grill *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," in Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 21271–21284.

[5] K. Sohn, "Improved deep metric learning with a multi-class n-pair loss objective," in Adv. Neural Inf. Process. Syst., vol. 29, 2016, pp. 1857–1865.

[6] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 15535–15545.

[7] J. Robinson, K. C. M. Chuang, S. Sra, and S. Jegelka, "When does contrastive learning preserve semantics?," *arXiv preprint arXiv:2106.01189*, 2021.

[8] Z. Yue *et al.*, "TS2Vec: Towards universal representation of time series," in Proc. AAAI Conf. Artif. Intell., vol. 36, no. 8, 2022, pp. 9180–9187.

[9] Z. Cheng *et al.*, 'Breaking the Memory Barrier of Contrastive Loss via Tile-Based Strategy', in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 10036–10045.

[10] K. Bonawitz *et al.*, "Towards federated learning at scale: System design," in Proc. Mach. Learn. Syst., vol. 1, 2019, pp. 374–388.

[11] A. G. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," Nat. Mach. Intell., vol. 2, no. 6, pp. 305–311, 2020.

[12] J.-Y. Franceschi, A. Dieuleveut, and M. Jaggi, "Unsupervised scalable representation learning for multivariate time series," in Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 4650–4661.

[13] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, and X. Li, "Time-series representation learning via temporal and contextual contrasting," in Proc. Int. Joint Conf. Artif. Intell. (IJCAI), 2021, pp. 2374–2381.

[14] G. Woo, C. Liu, D. Sahoo, A. Kumar, and S. Hoi, "CoST: Contrastive learning of disentangled seasonal-trend representations for time series forecasting," in Proc. Int. Conf. Learn. Represent. (ICLR), 2022.

[15] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in

Proc. Eur. Symp. Artif. Neural Netw. Comput. Intell. Mach. Learn. (ESANN), 2013, pp. 437–442.
[16] F. Wang and H. Liu, “Understanding the behaviour of contrastive loss,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 260–269.