Soft Contrastive Learning for Irregular Multivariate Time Series

Junghoon Lim [1] **Seunghan Lee** [1] **Taeyoung Park** [1]

Abstract

Irregular multivariate time series (IMTS) are characterized by irregular observation times, resulting in 1) misaligned time points across features (i.e., *misalignment*) and 2) inconsistent intervals between observations (i.e., *inconsistency*). However, existing time series methods often overlook these irregularities, leading to suboptimal performance, or depend on large labeled datasets. To this end, we introduce **SITS**, a simple yet effective *soft contrastive learning* strategy tailored for IMTS, where pairs are constructed from a *single* instance that shares the *same* irregularities, rather than from different instances with varying irregularities. Specifically, different views of a single instance are generated with varying masking ratios, where higher masking ratios correspond to smaller soft label values. Furthermore, we propose **SeqTAND**, a model architecture that handles misalignment and inconsistency in a *sequential* manner, which is shown to be more effective than addressing them in parallel. Experimental results demonstrate that SITS outperforms state-of-the-art methods in both classification and interpolation tasks.

1. Introduction

Irregular multivariate time series (IMTS) are widely observed in various fields such as healthcare (Zhang et al., 2022a), industry (Liu et al., 2021), and climatology (Cao et al., 2023). Nonetheless, handling IMTS is challenging due to irregularities, which can be categorized into: 1) *misalignment*, where features are observed at varying time points, and 2) *inconsistency*, characterized by irregular intervals between observations, as shown in Figure 1. However, existing models for regular TS analysis typically assume aligned time points and consistent intervals, which may lead to performance degradation (Chowdhury et al., 2023).

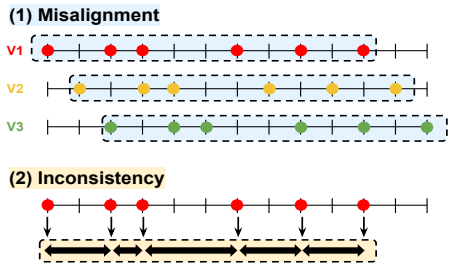

Figure 1: Irregularities in IMTS.

To address these issues, recent IMTS studies have mostly relied on supervised learning (SL), which is constrained by its need for large labeled datasets. In contrast, self-supervised learning (SSL), despite its success in regular time series (Lee et al., 2024; Dong et al., 2024; Nie et al., 2022), remains underexplored in IMTS. Although PrimeNet (Chowdhury et al., 2023) adopts an SSL approach, its reliance on hard contrastive learning (CL) with subseries consistency (Franceschi et al., 2019) suffers from level-shift sensitivity and a limited ability to preserve correlations (Yue et al., 2022; Lee et al., 2024).

To this end, we propose **S**oft Contrastive Learning for **I**rregular Multivariate **T**ime **S**eries (*SITS*), a simple yet effective *soft contrastive learning* strategy tailored for IMTS, where pairs are formed from a *single* instance that shares the same irregularities, rather than from different instances with varying irregularities. Specifically, different views are obtained by masking a single instance with varying masking ratios, where a pair containing an instance with a higher masking ratio is assigned a smaller soft label value.

Furthermore, we introduce **Seq**uential m**TAND** (*SeqTAND*) to effectively handle the irregularities of IMTS by addressing misalignment and inconsistency sequentially. This approach proves to be more effective than the parallel approach.

We conduct experiments on real-world datasets to demonstrate the effectiveness of the proposed method, achieving state-of-the-art (SOTA) performance across various tasks. Our main contributions are summarized as follows:

*Equal contribution [1]Department of Statistics and Data Science, Yonsei University. Correspondence to: Junghoon Lim <meaningfull9502@yonsei.ac.kr>, Seunghan Lee <seunghan9613@yonsei.ac.kr>, Taeyoung Park <tpark@yonsei.ac.kr>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025.

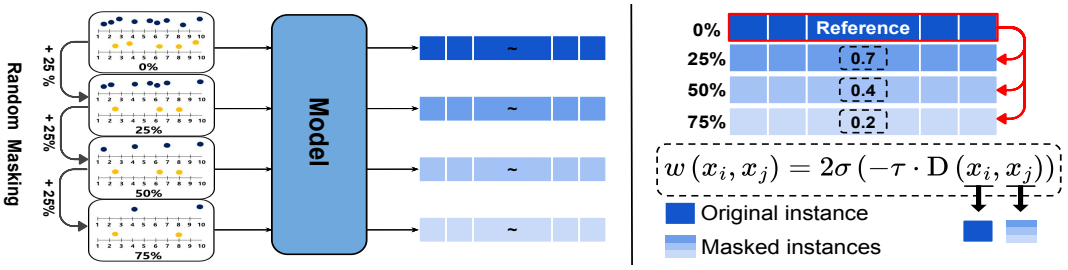

Figure 2: **Overall framework of SITS.** SITS constructs paired views from a single instance by applying random masking, thereby sharing the same irregularities. In contrast to conventional hard CL, which assigns hard labels (i.e., binary labels) to sample pairs, SITS utilizes assigns soft labels based on the similarity between augmented instances.

- We propose SITS, a simple yet effective soft contrastive learning strategy for IMTS, where pairs are formed from a single instance to ensure that they share the same irregularities while applying varying masking ratios. To the best of our knowledge, SITS is the first method to apply soft contrastive learning to IMTS.
- We introduce SeqTAND, a model that addresses misalignment and inconsistency sequentially, demonstrating greater effectiveness than the method that handles these irregularities in parallel.
- We conduct extensive experiments, where our method demonstrates SOTA performance across various datasets in both classification and interpolation tasks.

2. Preliminaries

Notation. Let $D = (T, X, M)$ represent a single IMTS instance, where $T \in \mathbb{R}^S$, $X \in \mathbb{R}^{S \times K}$, and $M \in \mathbb{R}^{S \times K}$. Here, T denotes the union of time points across all features, with S representing the total number of timestamps. X represents sequence values corresponding to time points T across K features, and M denotes an observation mask indicating the presence (1) or absence (0) of data at each time point.

3. Soft Contrastive Learning

In this section, we introduce SITS, a soft CL strategy for IMTS where soft assignments based on the similarity between augmented instances that share the irregularities of the original instance. The overall framework of SITS is illustrated in Figure 2. As IMTS exhibits unique irregularities in each instance, contrasting different instances with different irregularities may pose a challenge. To address this issue, we propose soft CL that contrasts *augmented instances from the same original instance* that share the same irregularities, rather than contrasting different original instances.

Data Augmentation. To generate diverse augmented views while ensuring consistent irregularities are shared across all instances, we progressively apply random masking at different levels in a cumulative manner. Each step builds on the previously masked data. An adjusted ratio r (Equation 1) regulates this process, ensuring that masking is applied based on a predefined ratio relative to the original instance.

$$\mathrm{Adj}(r) = \begin{cases} \mathrm{ratio}(i), & \text{if } i = 0, \\ \dfrac{\mathrm{ratio}(i) - \mathrm{ratio}(i-1)}{1 - \mathrm{ratio}(i-1)}, & \text{if } i \neq 0. \end{cases} \tag{1}$$

Soft Label. Soft labels are assigned based on the similarity between augmented instances and the original instance. We compute a soft assignment for a pair of data indices (x_i, x_j), which is used in the contrastive loss, as:

$$w(x_i, x_j) = 2\sigma\left(-\tau \cdot \mathrm{D}(x_i, x_j)\right). \tag{2}$$

where $\mathrm{D}(\cdot)$ is an arbitrary similarity metric, and τ is a hyperparameter controlling the sharpness of the soft labels.

Metric. Among various choices for D, we use KL-divergence (Kullback & Leibler, 1951) to measure the similarity between instances in terms of information loss, as random masking partially removes information from the original instance. Notably, unlike other metrics, KL-divergence is computed on M. Through experiments with various choices for D, as shown in Table E.4, we demonstrate that KL-divergence is the most suitable metric.

Loss. Specifically, the soft assignments are calculated based on the similarity between $x^{(0)}$ and $x^{(k)}$, where $x^{(0)}$ is the original instance and $x^{(k)}$ is the k-th augmented instance of $x^{(0)}$. The soft contrastive loss for x_i can be written as:

$$\ell^{(i,t)} = -\sum_{k \in \Omega} w(x_i^{(0)}, x_i^{(k)}) \log p(r_i^{(k)}), \tag{3}$$

$$p(r_i^{(k)}) = \frac{\exp(r_i^{(0)} \circ r_i^{(k)})}{\sum_{s \in \Omega} \exp(r_i^{(0)} \circ r_i^{(s)})}. \tag{4}$$

where $r_i^{(k)}$ is the representation of $x_i^{(k)}$, and Ω denotes the indices of augmented instances.

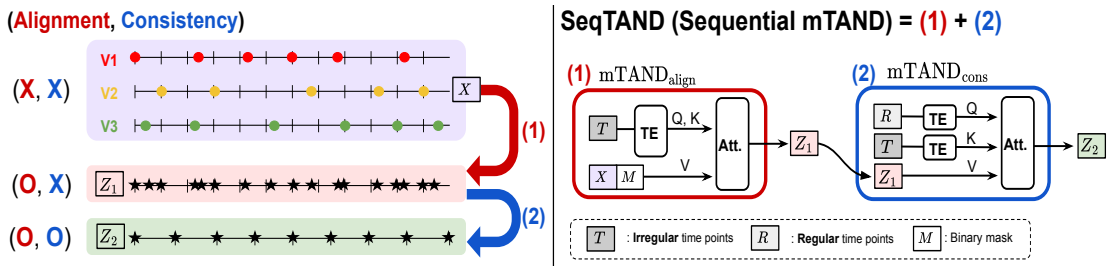

Figure 3: **Architecture of SeqTAND.** SeqTAND sequentially addresses misalignment and inconsistency in IMTS. $\text{mTAND}_{\text{align}}$ handles the *misalignment* using the irregular (observed) time points as reference points, while $\text{mTAND}_{\text{const}}$ handles the *inconsistency* using the regular time points as reference points.

4. Model Architecture

The proposed model is based on the Time Embedding (Shukla & Marlin, 2021), SeqTAND, and the Transformer Encoder (Vaswani et al., 2017), as shown in Figure 3.

Time Embedding (TE). Time Embedding $\phi_h(t)$ maps a continuous time point t into a d_r-dimensional representation, with the i-th dimension defined as:

$$\phi_h(t)[i] = \begin{cases} \omega_{0,h} \cdot t + \alpha_{0,h}, & \text{if } i = 0, \\ \sin(\omega_{i,h} \cdot t + \alpha_{i,h}) & \text{if } 0 < i < d_r. \end{cases} \tag{5}$$

Here, $w_{i,h}$ and $\alpha_{i,h}$ are learnable parameters that represent the frequency and phase of the sine function, respectively.

Sequential mTAND (SeqTAND). We introduce SeqTAND, which utilizes mTAND (Shukla & Marlin, 2021) sequentially, addressing the misalignment before handling the inconsistency.

To address the misalignment in IMTS, $\text{mTAND}_{\text{align}}$ uses the observed time points T as reference points instead of regular time points R, which can be expressed as:

$$\text{mTAND}_{\text{align}}(t = T, v = X, m = M) = (M \odot A_D)X, \tag{6}$$

$$A_D = \text{Softmax}\left(\frac{Q_D K_D}{d_r}\right), \tag{7}$$

$$Q_D = K_D = \phi_h(T). \tag{8}$$

where Q_D and K_D are the vectors obtained by feeding the observed time points T into $\phi_h(\cdot)$. Then, attention scores A_D, representing the similarity between the time points of Q_D and K_D, are calculated and element-wise multiplied by M to generate an irregular time representation Z_1.

After addressing the misalignment, $\text{mTAND}_{\text{cons}}$ handles the inconsistency by using the regular time points R as reference points, generating a regular time representation Z_2 as follows:

$$\text{mTAND}_{\text{cons}}(t = R, v = Z_1, m = \mathbf{1}) = A_I Z_1, \tag{9}$$

$$A_I = \text{Softmax}\left(\frac{Q_I K_I}{d_r}\right), \tag{10}$$

$$Q_I = \phi_h(R),\ K_I = \phi_h(T). \tag{11}$$

All procedures are the same as in $\text{mTAND}_{\text{align}}$, except that Q_I is obtained using R instead of T and Z_1, derived from $\text{mTAND}_{\text{align}}$, is used in place of X and M.

Transformer Encoder. To capture relationships between features, attention is performed on Z_2, the output of $\text{mTAND}_{\text{cons}}$, which serves as the query, key, and value vector. Also we use Time Embedding for positional encoding to further leverage the known exact time points for each vector. Like a typical Transformer Encoder, residual and feed-forward layers are applied to the output of the attention.

5. Experiment

Interpolation. We conduct experiments on the PhysioNet and Human Activity datasets. For the interpolation task, we simulate a scenario where [10%, 30%, 50%, 70%, 90%] of the observed time points are randomly missing and predict the values. As shown in Table 1, SITS achieves outstanding performance in these experiments.

Classification. For the classification task, we use three datasets: PhysioNet, MIMIC-III, and Human Activity. The results in Table 2 demonstrate that SITS consistently outperforms the baseline models across all datasets.

Comparison of Performance under Various Settings. Table 3 compares performance under various settings on the classification task across three datasets. Since human activity involves classifying all originally observed time points, sequential approaches are not applicable. In the supervised

Methods		SSL models				IMTS models						SSL+IMTS models			
		TNC		TST		P-LSTM		mTAND		t-PatchGNN		PrimeNet		SITS (Ours)	
Dataset	Ratio (%)	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PhysioNet	10	**0.049**	**0.070**	0.109	0.169	0.056	0.087	0.083	0.156	0.074	0.114	0.068	0.129	0.051	0.093
	30	0.137	**0.147**	0.260	0.255	0.147	0.170	0.153	0.196	0.129	0.166	0.136	0.183	**0.121**	0.155
	50	0.224	**0.210**	0.408	0.337	0.247	0.243	0.226	0.239	0.217	0.241	0.203	0.232	**0.187**	0.212
	70	0.317	0.273	0.555	0.411	0.322	0.283	0.315	0.284	0.270	0.295	0.285	0.283	**0.268**	**0.263**
	90	0.505	0.385	0.703	0.494	0.502	0.390	0.482	0.363	0.441	0.368	0.434	0.360	**0.406**	**0.341**
Human Activity	10	0.058	0.023	0.399	0.083	0.078	0.032	0.052	0.031	0.052	0.022	0.048	0.023	**0.045**	**0.020**
	30	0.183	0.053	1.154	0.168	0.215	0.065	0.158	0.058	0.161	0.054	0.147	0.050	**0.143**	**0.048**
	50	0.3d40	0.086	1.905	0.252	0.373	0.097	0.268	0.085	0.259	0.087	**0.248**	0.077	0.251	**0.076**
	70	0.585	0.131	2.656	0.333	0.608	0.138	0.389	0.111	0.397	0.113	0.376	0.106	**0.370**	**0.104**
	90	1.213	0.213	3.406	0.414	1.270	0.223	0.640	0.159	0.606	0.156	0.597	0.151	**0.579**	**0.148**
Avg.		0.361	0.159	1.156	0.292	0.382	0.173	0.277	0.168	0.261	0.162	0.254	0.159	**0.242**	**0.146**

Table 1: **Performance of the interpolation task under various missing ratios.** Best results are highlighted in **red**, and second-best in blue.

Dataset	AUC		Acc.
	PhysioNet	MIMIC-III	Human Activity
SSL models			
TNC	$77.5_{\pm 0.003}$	$83.3_{\pm 0.003}$	$87.2_{\pm 0.002}$
TS2Vec	$78.1_{\pm 0.005}$	$82.6_{\pm 0.005}$	$89.8_{\pm 0.002}$
TST	$77.5_{\pm 0.004}$	$79.8_{\pm 0.006}$	$66.1_{\pm 0.001}$
IMTS models			
P-LSTM	$77.6_{\pm 0.008}$	$83.8_{\pm 0.001}$	$85.5_{\pm 0.007}$
RNN-VAE	$57.7_{\pm 0.004}$	$51.8_{\pm 0.002}$	$34.3_{\pm 0.009}$
ODE-RNN	$80.8_{\pm 0.002}$	$84.5_{\pm 0.004}$	$88.5_{\pm 0.004}$
L-ODE	$81.2_{\pm 0.002}$	$84.3_{\pm 0.006}$	$87.0_{\pm 0.002}$
mTAND	$80.9_{\pm 0.003}$	$83.3_{\pm 0.005}$	$90.0_{\pm 0.001}$
t-PatchGNN	$65.7_{\pm 0.002}$	$82.9_{\pm 0.003}$	$72.0_{\pm 0.001}$
SSL+IMTS models			
PrimeNet	$82.2_{\pm 0.003}$	$84.3_{\pm 0.001}$	$88.9_{\pm 0.002}$
SITS (Ours)	**84.0**$_{\pm 0.001}$	**85.2**$_{\pm 0.002}$	**90.9**$_{\pm 0.001}$

Table 2: **Performance of the classification task.** The best results for each dataset are highlighted in **red**, and the second-best in blue.

Sequential	SITS	PhysioNet	MIMIC-III	Human Activity
		81.7	83.4	88.7
✓		82.8	83.8	-
	✓	83.2	84.8	**90.9**
✓	✓	**84.0**	**85.2**	-

Table 3: Comparison of performance across various settings.

learning (SL) setting, the model with sequential mTAND consistently outperforms the parallel approach. While SITS alone improves performance in SSL, the combination of sequential mTAND and SITS achieves the highest performance.

Effectiveness of SeqTAND. Table 4 evaluates the effectiveness of SeqTAND on classification(AUC) and reconstruction(MSE, MAE) tasks using the PhysioNet dataset. We compare three models: (1) mTAND, which applies Cons once; (2) A model with the same number of parameters as SeqTAND that applies Cons twice; (3) SeqTAND, which applies Aligns and Cons sequentially. The number of parameters is averaged across tasks. In reconstruction, the hidden dimension is set to 32, smaller than the number of variables (41). Learning and inference times are measured per epoch in seconds. SeqTAND achieves the best performance among the models. This suggests that SeqTAND effectively captures IMTS information (Bank et al., 2023), emphasizing that the improvement comes from the sequential approach.

Approach	# Parameters	AUC	MSE	MAE	Learning	Inference
Cons	221k	81.7	0.041	0.098	4.332	0.707
Cons → Cons	247k	82.0	0.044	0.105	4.662	0.736
Align → Cons	247k	**82.8**	**0.016**	**0.052**	5.249	0.785

Table 4: Effectiveness of SeqTAND.

CL		PhysioNet	MIMIC-III	Human Activity
Hard	Soft			
		82.8	83.8	88.7
✓		82.7	84.2	89.8
	✓	**84.0**	**85.2**	**90.9**

Table 5: Effectiveness of soft CL.

Effectiveness of soft CL. Table 5 demonstrates the effectiveness of the proposed soft CL on the classification task across three datasets. In the hard setting, only augmented versions were treated as positives with a label of 1. The results show that the soft approach consistently outperforms the hard approach.

6. Conclusion

In this paper, we propose SITS, a simple yet effective soft contrastive learning strategy for IMTS, where pairs are formed from a single instance to ensure that they share the same irregularities. Soft labels are assigned based on similarity, with higher masking ratios leading to smaller soft label values. Extensive experiments validate the effectiveness of our method. We hope this work highlights the potential of CL strategies for IMTS and inspires further research into SSL approaches.

References

Bank, D., Koenigstein, N., and Giryes, R. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pp. 353–374, 2023.

Cao, D., Enouen, J., Wang, Y., Song, X., Meng, C., Niu, H., and Liu, Y. Estimating treatment effects from irregular time series observations with hidden confounders. In *AAAI*, 2023.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607. PMLR, 2020.

Chen, Y., Ren, K., Wang, Y., Fang, Y., Sun, W., and Li, D. Contiformer: Continuous-time transformer for irregular time series modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=YJDz4F2AZu.

Chowdhury, R. R., Zhang, X., Shang, J., Gupta, R. K., and Hong, D. Tarnet: Task-aware reconstruction for time-series transformer. In *KDD*, pp. 212–220, 2022.

Chowdhury, R. R., Li, J., Zhang, X., Hong, D., Gupta, R. K., and Shang, J. Primenet: Pre-training for irregular multivariate time series. In *AAAI*, 2023.

Dong, J., Wu, H., Zhang, H., Zhang, L., Wang, J., and Long, M. Simmtm: A simple pre-training framework for masked time-series modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwoh, C.-K., Li, X., and Guan, C. Time-series representation learning via temporal and contextual contrasting. In *IJCAI*, 2021.

Franceschi, J.-Y., Dieuleveut, A., and Jaggi, M. Unsupervised scalable representation learning for multivariate time series. In *NeurIPS*, 2019.

Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, 2021.

Horn, M., Moor, M., Bock, C., Rieck, B., and Borgwardt, K. Set functions for time series. In *ICML*, pp. 4353–4363. PMLR, 2020.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Kaluža, B., Mirchevska, V., Dovgan, E., Luštrek, M., and Gams, M. An agent-based approach to care in independent living. In *Ambient Intelligence: First International Joint Conference, AmI 2010, Malaga, Spain, November 10-12, 2010. Proceedings 1*, pp. 177–186. Springer, 2010.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

Lee, S., Park, T., and Lee, K. Soft contrastive learning for time series. In *ICLR*, 2024.

Liu, Z., Wang, H., Zhang, Y., and Wang, X. Deep masking generative network for irregularly sampled multivariate time series. In *2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pp. 296–300. IEEE, 2021.

Logeswaran, L. and Lee, H. An efficient framework for learning sentence representations. In *ICLR*, 2018.

Neil, D., Pfeiffer, M., and Liu, S.-C. Phased lstm: Accelerating recurrent network training for long or event-based sequences. In *NeurIPS*, 2016.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. *ICLR*, 2022.

Rubanova, Y., Chen, R. T., and Duvenaud, D. K. Latent ordinary differential equations for irregularly-sampled time series. In *NeurIPS*, 2019.

Shukla, S. N. and Marlin, B. M. Interpolation-prediction networks for irregularly sampled time series. In *ICLR*, 2019.

Shukla, S. N. and Marlin, B. M. Multi-time attention networks for irregularly sampled time series. In *ICLR*, 2021.

Silva, I., Moody, G., Scott, D. J., Celi, L. A., and Mark, R. G. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 computing in cardiology*, pp. 245–248. IEEE, 2012.

Tan, Q., Ye, M., Yang, B., Liu, S., Ma, A. J., Yip, T. C.-F., Wong, G. L.-H., and Yuen, P. Data-gru: Dual-attention time-aware gated recurrent unit for irregular multivariate time series. In *AAAI*, 2020.

Tonekaboni, S., Eytan, D., and Goldenberg, A. Unsupervised representation learning for time series with temporal neighborhood coding. In *ICLR*, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *NeurIPS*, volume 30, 2017.

Wang, X., Zhang, R., Shen, C., Kong, T., and Li, L. Dense contrastive learning for self-supervised visual pretraining. In *CVPR*, pp. 3024–3033, 2021.

Yalavarthi, V. K., Madhusudhanan, K., Scholz, R., Ahmed, N., Burchert, J., Jawed, S., Born, S., and Schmidt-Thieme, L. Grafiti: Graphs for forecasting irregularly sampled time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16255–16263, 2024.

Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., and Xu, B. Ts2vec: Towards universal representation of time series. In *AAAI*, 2022.

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., and Eickhoff, C. A transformer-based framework for multivariate time series representation learning. In *KDD*, pp. 2114–2124, 2021.

Zhang, W., Yin, C., Liu, H., Zhou, X., and Xiong, H. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *NeurIPS*, 2024.

Zhang, X., Zeman, M., Tsiligkaridis, T., and Zitnik, M. Graph-guided network for irregularly sampled multivariate time series. In *ICLR*, 2022a.

Zhang, X., Zhao, Z., Tsiligkaridis, T., and Zitnik, M. Self-supervised contrastive pre-training for time series via time-frequency consistency. In *NeurIPS*, 2022b.

Zhang, X., Li, S., Chen, Z., Yan, X., and Petzold, L. R. Improving medical predictions by irregular multimodal electronic health records modeling. In *ICML*, pp. 41300–41313. PMLR, 2023.

A. Related Work

A.1. Self-supervised Learning for Time Series

Self-supervised learning (SSL) has shown success in computer vision (Chen et al., 2020; Wang et al., 2021) and natural language processing (Logeswaran & Lee, 2018; Gao et al., 2021), leading to its application in time series (TS). Contrastive learning methods such as T-loss (Franceschi et al., 2019) and TNC (Tonekaboni et al., 2021), define positive and negative pairs based on temporal characteristics. TS-TCC (Eldele et al., 2021) employs a temporal contrastive loss, while TS2Vec (Yue et al., 2022) applies hierarchical contrastive learning for robust contextual representations. TF-C (Zhang et al., 2022b) captures time-frequency relationships, and SoftCLT (Lee et al., 2024) enhances both instance-wise and temporal contrastive learning with soft assignments. Masked modeling approaches, such as TST (Zerveas et al., 2021) and TARNet (Chowdhury et al., 2022), reconstruct masked segments to learn informative representations. While these methods can be adapted to IMTS via binning (Shukla & Marlin, 2019), they often fail to capture critical irregularities, leading to performance degradation (Chowdhury et al., 2023). This highlights the need for self-supervised models tailored to IMTS.

A.2. Methods for Irregular Time Series

Recently, fully and semi-supervised methods have been extensively researched in the context of IMTS and have proven effective in handling irregularities, resulting in strong performance. GRU-D (Che et al., 2018) integrates timestamps into GRU to manage missing values. DATA-GRU (Tan et al., 2020) employs dual-attention mechanisms, while P-LSTM (Neil et al., 2016) updates memory selectively through a time gate. ODE-RNN (Rubanova et al., 2019) and Latent-ODE (Rubanova et al., 2019) leverage neural ODEs for continuous modeling. ContiFormer (Chen et al., 2023) integrates neural ODEs with Transformers for continuous-time modeling. GraFITi (Yalavarthi et al., 2024) applies graph neural networks (GNNs) to forecast irregular series by predicting edge weights. t-PatchGNN (Zhang et al., 2024) transforms univariate series into temporal patches and utilizes time-adaptive GNNs to capture inter-series relationships. IP-Nets (Shukla & Marlin, 2019) employs semi-parametric interpolation, while SeFT (Horn et al., 2020) uses a set function-based classification approach. mTAND (Shukla & Marlin, 2021) employs an attention mechanism to embed continuous time values and generate a fixed-length representation. UTDE (Zhang et al., 2023) combines embeddings from mTAND and imputed TS with learnable gates. PrimeNet (Chowdhury et al., 2023) introduces pre-training strategies, including hard contrastive learning (CL) and masked modeling. However, it faces limitations in capturing inter-series correlations and addressing level shift issues in subseries consistency.

B. Preliminaries

mTAND (Shukla & Marlin, 2021). mTAND produces a *fixed-length* representation of a TS with a *variable* number of observations across variables. Specifically, it processes a query time point t and a set of observed time points and their corresponding values (i.e., (T, X)) as keys and values. By utilizing predefined regular time points R as reference points, mTAND aligns continuous time points to consistent intervals. The process of mTAND can be expressed as:

$$\text{mTAND}(t = R, v = X, m = M) = Z, \tag{12}$$

where Z is a fixed-length representation of predefined regular time points R.

C. Dataset

- **PhysioNet Challenge** (Silva et al., 2012) dataset comprises multivariate time series data derived from intensive care unit (ICU) records, featuring 41 physiological variables. Each record contains 48 hours of measurements collected after ICU admission. The main goal of this dataset is to predict in-hospital mortality, which is framed as a binary classification task.
- **MIMIC-III** (Johnson et al., 2016) dataset consists of multivariate time series data with sparse and irregular sampling of physiological signals, collected from the Beth Israel Deaconess Medical Center between 2001 and 2012. It includes 17 key clinical variables. The irregular sampling and missing data pose significant challenges, and the dataset is mainly utilized to predict clinical outcomes, such as in-hospital mortality.
- **Human Activity** (Kaluža et al., 2010) dataset includes 3D positional data captured from sensors placed on the waist,

chest, and ankles of five individuals performing various activities, such as walking, sitting, lying, and standing. It is designed for activity recognition and classification, offering detailed positional information to facilitate the analysis of human movements.

D. More on Experiment Section

D.1. Baselines

- **Self-Supervised TS Methods**
  - **TNC** (Tonekaboni et al., 2021) sets temporal neighborhoods based on local signal smoothness.
  - **TS2Vec** (Yue et al., 2022) employs hierarchical CL to learn robust contextual representations.
  - **TST** (Zerveas et al., 2021) pre-trains a Transformer-based model using fixed-length MM.

- **Irregular TS Methods**
  - **P-LSTM** (Neil et al., 2016) introduces a time gate controlled by a parametrized oscillation.
  - **RNN-VAE** comprises an RNN encoder and a decoder within a variational autoencoder model.
  - **ODE-RNN** (Rubanova et al., 2019) employs neural ODEs for hidden state dynamics, updating them with an RNN using new observations.
  - **L-ODE** (Rubanova et al., 2019) refers to the Latent ODE, which uses an ODE-RNN as the encoder and a neural ODE as the decoder.
  - **mTAND** (Shukla & Marlin, 2021) uses a multi-time attention module to produce a fixed-length time representation.
  - **t-patchGNN** (Zhang et al., 2024) uses transformable patches and a time-adaptive GNN for forecasting.
  - **PrimeNet** (Chowdhury et al., 2023) designs time-sensitive CL and constant-time MM.

D.2. Experimental Protocols

During pre-training, we determine the hyperparameters τ and $n(\Omega)$ by performing a grid search to select the values from $(1, 2, 3, 4, 5)$ and $(2, 3, 4, 5)$ that yield the best performance. In the fine-tuning stage, we update the parameters for both the task-specific layers and SITS. For the classification task, we use cross-entropy loss, while for the interpolation task, we use mean squared error (MSE) as the loss function. Early stopping is applied based on the validation dataset, with patience set to 20 epochs for pre-training, 50 epochs for classification tasks, and 100 epochs for interpolation tasks. The learning rates are set to 0.001 and 0.0005, and the batch size is fixed at 128. The time embedding dimension and hidden vector dimension are both set to 128, and the number of attention heads is set to 1. These configurations follow those used in PrimeNet (Chowdhury et al., 2023). For baselines that do not report experimental settings on the target dataset, we apply the same experimental settings as ours; otherwise, we follow the experimental settings reported in the original papers. All experiments are repeated three times with different random seeds for model initialization.

D.3. Task and evaluation metrics

We demonstrate the effectiveness of the proposed SITS on two downstream tasks: interpolation and classification tasks. In interpolation task, we condition on the observed data points to predict the values for the missing points and assess interpolation performance using MSE(10^{-2}) and MAE(10^{-1}) for PhysioNet and MSE(10^{-1}) and MAE for Human Activity. In classification task, we assess classification performance using the Area Under the ROC Curve (AUC), while for the Human Activity dataset, which involves multi-class prediction, we use accuracy (Acc.).

E. Analysis

Robustness to the number of augmentations. Table E.1 compares the effect of the number of augmented instances on the classification task using three datasets, where the masking ratio is determined by n-quantiles. The results show that performance remains consistent across different numbers of augmentations. Based on these findings, we select $n(\Omega) = 3, 3, 4$ for PhysioNet, Human Activity, and MIMIC-III, which yield the best performance.

$n(\Omega)$	PhysioNet	MIMIC-III	Human Activity
2	83.6	84.8	90.6
3	**84.0**	84.9	**90.9**
4	83.5	**85.2**	90.7
5	83.5	84.7	90.4

Table E.1: Performance across different numbers of augmented instances.

τ	PhysioNet	MIMIC-III	Human Activity
1.0	82.9	84.5	**90.9**
2.0	83.4	84.5	90.4
3.0	83.5	**85.2**	90.2
4.0	**84.0**	84.8	90.1
5.0	83.7	84.4	90.3

Table E.2: Performance across different values of the hyperparameter τ.

Cumul.	Physionet	MIMIC-III	Human Activity
	83.4	83.2	90.5
✓	**84.0**	**85.2**	**90.9**

Table E.3: Effect of cumulative random masking.

Metric	PhysioNet	MIMIC-II	Human Activity
DTW	83.1	84.2	90.5
Rank	82.6	84.4	89.7
Ratio	83.6	84.6	89.9
KL-divergence	**84.0**	**85.2**	**90.9**

Table E.4: Metrics for soft CL.

Robustness to the hyperparameter τ. In soft CL, τ is introduced to control the sharpness of soft labels. Table E.2 illustrates the robustness of performance to different values of τ on the classification task across three datasets. Based on the experimental results, we select $\tau = 4, 3, 1$ for PhysioNet, MIMIC-III, and Human Activity.

Effect of cumulative random masking. Table E.3 demonstrates the effectiveness of the proposed cumulative random masking on the classification task using three datasets. The results show that the cumulative approach consistently outperforms the non-cumulative approach.

Metrics for soft CL. Table E.4 evaluates several metrics: dynamic time warping (DTW), Rank, Ratio, and KL-divergence, on the classification task across three datasets. The Rank is assigned in ascending order, starting with the smallest masking ratio, and Ratio corresponds to the masking ratio. The results demonstrate that although the degree of performance improvement varies across metrics, most of them are effective. Among these, KL-divergence achieves the best performance and is thus selected.