Time-Frequency Augmented Multi-level Contrastive Clustering for Time Series

Congyu Wang[1], **Mingjing Du**[1*], **Xiang Jiang**[1]

[1]School of Artificial Intelligence and Computer Science, Jiangsu Normal University
wangcongyu@jsnu.edu.cn, dumj@jsnu.edu.cn, xjiang@jsnu.edu.cn

Abstract

Current unsupervised time series clustering methods often struggle to fully exploit the inherent characteristics of time series data and commonly adopt a two-stage training strategy that separates feature learning from the clustering process. To address these limitations, this paper proposes a novel deep clustering framework, **T**ime-**F**requency augmented **M**ulti-level **C**ontrastive **C**lustering (TFMCC). TFMCC employs a multi-scale time-frequency augmentation strategy, where each training iteration stochastically selects time and frequency scales to generate diverse augmented views, enhancing the model's ability to learn robust and generalizable representations. In addition, a multi-level contrastive learning mechanism is introduced to jointly capture temporal dependencies, inter-sample similarities, and cluster structures. By jointly optimizing these components, TFMCC enables the learning of temporally-aware and clustering-friendly representations. Experimental results on 40 benchmark datasets demonstrate that TFMCC outperforms six existing methods in clustering accuracy.

Code — https://github.com/Du-Team/TFMCC

Introduction

Time series clustering aims to discover inherent patterns and structural relationships within sequential data and plays a vital role in fields such as finance, meteorology, and healthcare (Yang et al. 2023; Cini, Mandic, and Alippi 2024). Recently, contrastive learning–based clustering methods have shown strong potential for this task by constructing positive and negative sample pairs to automatically extract informative representations in an unsupervised manner (Zhong, Huang, and Wang 2023). Despite promising progress, existing methods still face key limitations.

Firstly, traditional data augmentation approaches, such as flipping or permutation (Eldele et al. 2021), can disrupt temporal dependencies (Liu and Chen 2024), while time-series-specific augmentations typically operate only in the time domain (Yue et al. 2022) and at a single scale (Peng et al. 2024), limiting their ability to capture rich patterns (e.g., trend and seasonality) and multi-scale dependencies (e.g., short-, medium-, and long-term). In addition, many contrastive frameworks prioritize instance-level similarity while overlooking temporal relationships that are critical for time series representation learning. Moreover, the widespread use of two-stage training approaches separates representation learning from clustering, hindering joint optimization and increasing tuning complexity (Ma et al. 2020).

To tackle these issues, we propose **T**ime-**F**requency augmented **M**ulti-level **C**ontrastive **C**lustering (TFMCC), an end-to-end deep clustering framework composed of three modules: data augmentation, representation learning, and clustering. In the augmentation module, we introduce a multi-scale augmentation strategy in both time and frequency domains. Across training iterations, context-base cropping with stochastic temporal scales captures multi-granular patterns while wavelet transforms with varying decomposition levels enrich spectral features, collectively generating diverse views that enhance both robustness and generalization. For representation learning, a multi-level contrastive mechanism combines temporal-, instance-, and cluster-level objectives. Lower layers employ temporal-level and instance-level contrastive losses to capture temporal dependencies and instance relationships, while higher layers utilize a cluster-level contrastive loss to optimize the membership matrix for clustering refinement. In summary, the main contributions of this paper are as follows:

- We develop TFMCC, a unified end-to-end deep clustering framework for time series. This framework integrates representation learning and the clustering task into a single optimization process, significantly improving clustering performance on time series data.
- We introduce a multi-scale time-frequency augmentation strategy that dynamically applies context-based cropping with varying temporal scales and wavelet-based transformations with different frequency levels at each training iteration, generating different augmented views to capture diverse patterns and enhance generalization.
- We propose a multi-level contrastive learning mechanism that integrates temporal-, instance-, and cluster-level objectives to preserve temporal dependencies, capture inter-instance similarities, and uncover clustering structures, thereby producing more discriminative and clustering-friendly representations.

*Corresponding Author

Related Work

In this section, we review deep time series clustering methods and contrastive learning algorithms that are closely related to this study.

Deep Time Series Clustering

In recent years, numerous deep learning-based time series clustering methods have been proposed (Tiano, Bonifati, and Ng 2021). These methods automatically extract latent features via neural networks and perform clustering in the latent space, greatly enhancing the expressiveness and adaptability of clustering models (Li and Liu 2021). Existing methods can be broadly categorized into separated and joint optimization strategies (Xu et al. 2022; Alqahtani et al. 2021).

Separated optimization typically adopts a "feature-then-cluster" paradigm, where time series are first encoded by deep models, and clustering is subsequently performed using traditional algorithms such as k-means (McQueen 1967). For example, Chen et al. (2022) employ RNNs for representation learning before clustering in a shared latent space, while T-LSTM (Baytas et al. 2017) captures both short- and long-term dependencies in medical time series to support patient subtyping. However, these methods decouple representation learning from clustering, preventing mutual refinement between the two stages and often resulting in suboptimal performance due to increased tuning complexity (Ma et al. 2020).

To address these limitations, joint optimization strategies have gained traction by integrating feature learning and clustering into a unified training process. DTC (Madiraju 2018) incorporates clustering objectives into an autoencoder framework, while STCN (Ma et al. 2020) adopts a self-supervised framework that iteratively refines pseudo-labels and model parameters. TCGAN (Huang and Deng 2023) leverages adversarial training to improve clustering through latent distribution learning, and DTCC (Zhong, Huang, and Wang 2023) introduces contrastive learning to enhance both representation quality and clustering accuracy in a self-supervised manner.

Unlike the above methods, our model achieves end-to-end joint clustering by incorporating temporal-, instance-, and cluster-level contrastive objectives into a unified framework. The entire process is optimized through backpropagation, achieving superior clustering performance.

Contrastive Learning

Contrastive learning, which aims to maximize the similarity between positive pairs and minimize it between negative pairs in the feature space, has gained significant attention in time series analysis (Wang et al. 2023; Zhong et al. 2021; Sun et al. 2025). Franceschi et al. (2019) introduce subsequence-based contrastive learning to bring representations closer to their sampled segments. TS2Vec (Yue et al. 2022) further improves robustness by encouraging each timestamp to reconstruct itself under varying contexts, enhancing context consistency. More recently, contrastive learning has been applied to deep clustering to enhance representation quality and clustering performance. For instance, CC (Li et al. 2021) proposes a single-stage framework combining instance- and cluster-level contrastive learning, while SACC (Deng et al. 2023) extends dual-view augmentation to multiple views with varying strengths to improve clustering outcomes.

In time series clustering, contrastive learning has also been explored. CDCC (Peng et al. 2024) incorporates both time- and frequency-domain information within a contrastive clustering framework to boost clustering accuracy.

Compared with these methods, our model performs multi-scale augmentation in both time and frequency domains without relying on predefined scales, allowing better adaptation to diverse temporal patterns. Moreover, rather than focusing on a single representation level, we introduce multi-level contrastive constraints to capture rich structural information across different layers, leading to more informative and clustering-friendly representations.

Method

This section will provide a detailed introduction to the time-frequency augmented multi-level contrastive clustering method proposed in this paper. The method primarily consists of two core components: the multi-scale time-frequency augmentation module and the multi-level contrastive learning module, both of which are jointly optimized in a unified end-to-end deep architecture. The overall architecture is illustrated in Figure 1.

Multi-Scale Time-Frequency Augmentation

Time series data inherently contains rich temporal dynamics (Tonekaboni, Eytan, and Goldenberg 2021) and frequency characteristics (Zhang et al. 2022). Specifically, temporal features can manifest as patterns at varying lengths, while frequency components reflect periodicity across different bands, both of which exhibit multi-scale characteristics (Wang et al. 2024; Hu et al. 2025). However, existing contrastive learning methods often employ fixed-scale augmentations or overlook the interplay between time and frequency dimensions, thus failing to fully capture the structural diversity embedded in time series. To address this issue, we propose a multi-scale time-frequency augmentation strategy.

As illustrated in Figure 2, the augmentation pipeline consists of two branches at a single training iteration. In the first branch, a time series $x \in \mathbb{R}^T$ undergoes cropping with randomized window sizes and contextual expansion, producing a time-domain view. In the second branch, the cropped sequence is further transformed using Discrete Wavelet Transform (DWT)-based frequency perturbation through random decomposition level selection and coefficient scaling, reconstructed through Inverse DWT (IDWT) to yield a frequency-perturbed view. The augmentation dynamically generates diverse augmented views across training iterations, allowing the model to learn diverse patterns in time series data.

Multi-Scale Time Domain Augmentation In the time domain, TFMCC adopts context-based random cropping to generate augmented samples. A window length is first randomly selected from a predefined multi-scale range to en-

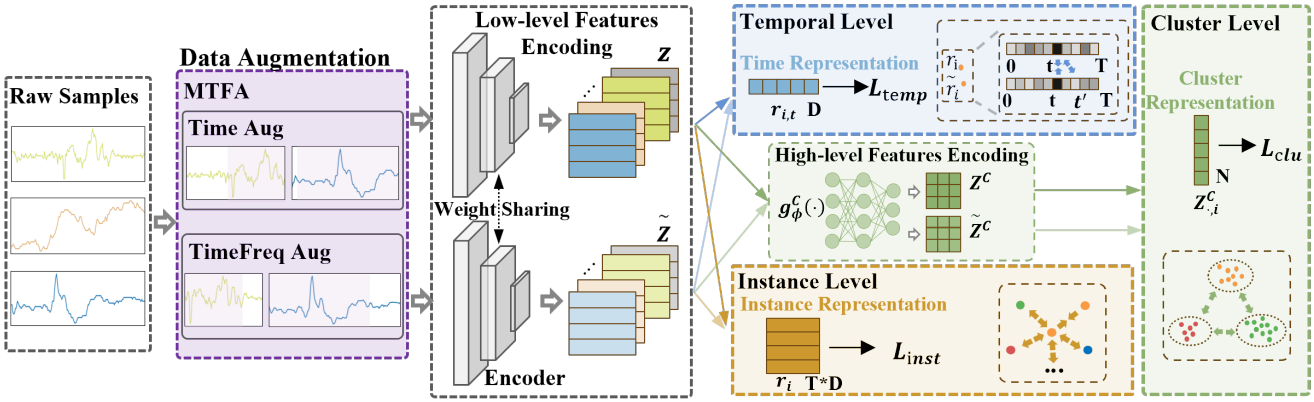

Figure 1: Overall Structure of TFMCC Framework.

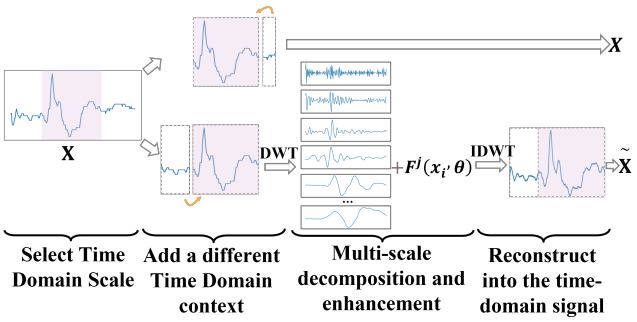

Figure 2: Data Augmentation Process.

sure temporal diversity. Based on this length, a subsequence is cropped from the original time series. To preserve temporal dependencies and introduce contextual variation, we follow a context consistency strategy (Yue et al. 2022), where adjacent segments before and after the crop are randomly sampled and concatenated. This dual randomness in scale and context generates multi-scale augmented views while retaining meaningful temporal structure, improving the model's ability to learn diverse temporal patterns.

Multi-Scale Frequency Domain Augmentation In the frequency domain, TFMCC leverages the Discrete Wavelet Transform (DWT) (Mallat 2002) to decompose time series into multi-resolution components, enabling the model to access frequency features at different scales (Murad, Aktukmak, and Yilmaz 2025; Dai et al. 2024). In our implementation, we adopt Daubechies wavelets (db4). The wavelet transform employs an iterative decomposition method using high-pass and low-pass filters to extract wavelet coefficients. The high-pass filter extracts detail information (i.e., detail coefficients X_{D_i}), while the low-pass filter extracts low-frequency information (i.e., approximation coefficients X_{A_i}). At each decomposition level, the approximation coefficients from the previous level are further split into new approximation and detail coefficients. Specifically, a time series X is decomposed into one approximation and multiple detail coefficient sequences via:

$$[X_{A_m}, X_{D_m}, ..., X_{D_1}] = \text{Decomp}(X, \psi, m), \quad (1)$$

where ψ is the wavelet type, and m is a randomly specified level of DWT decomposition. To reduce redundancy, only the final-level approximation coefficient X_{A_m} is retained.

To generate frequency-domain augmentations, we ran select a level $j \in \{1, ..., m\}$ and apply a transformation $F^j(x_i, \theta)$, with random scaling factor $\theta \sim \mathcal{U}(0.2, 1.8)$, to perturb the detail coefficients:

$$y_{i,D_j} = F^j(x_i, \theta), \quad (2)$$

$F^j(x_i, \theta)$ scales the detail coefficients at the selected level j with a random factor to produce y. This selective enhancement introduces controlled frequency perturbations at different scales without distorting the overall structure. The augmented detail coefficients $\{Y_{D_j}\}$, together with X_{A_m}, are then used to reconstruct the time-domain signal via IDWT:

$$Y = \text{Recon}_\psi(X_{A_m}, \{Y_{D_k}\}_{k=1}^m) \quad (3)$$

This stochastic multi-scale perturbation preserves temporal structure while enhancing frequency-aware representation learning, improving model generalization across varying periodic patterns.

Multi-Level Contrastive Learning

To effectively capture the rich characteristics of time series data while simultaneously optimizing for clustering performance, we propose a unified optimization framework that jointly learns feature representations and cluster assignments through a multi-level contrastive learning approach. To avoid conflicts that may arise from simultaneously optimizing heterogeneous objectives within the same feature space(Xu et al. 2022), we distribute three complementary contrastive losses across different levels of the network. Specifically, temporal-level and instance-level contrastive losses are applied to lower-level features to preserve temporal consistency and capture inter-instance distinctions, while the cluster-level contrastive loss is applied to higher-level features to uncover underlying clustering structures. This

integrated optimization strategy enables the model to learn representations that are temporally coherent, discriminative across instances, and clustering-friendly, thereby bridging the gap between representation learning and downstream clustering objectives.

Temporal-Level Contrastive Loss The temporal-level contrastive learning component focuses on modeling the internal temporal structure of a sequence by performing contrastive learning at the timestamp level (Yue et al. 2022). This mechanism views a time series as a composition of individual time points and encourages consistency between corresponding timestamps across different augmented views of the same sequence. By enforcing alignment at the time-step level, the model is guided to capture long-term dependencies and periodic patterns. Formally, the temporal-level contrastive loss for the t-th time step of the i-th sample is defined as:

$$\ell_{\text{temp}}^{(i,t)} = -\log \frac{\exp\left(\frac{r_{i,t}\cdot\tilde{r}_{i,t}}{\tau^t}\right)}{\sum_{t'\in\Omega}\left(\exp\left(\frac{r_{i,t}\cdot\tilde{r}_{i,t'}}{\tau^t}\right)+\mathbb{I}_{[t\neq t']}\exp\left(\frac{r_{i,t}\cdot r_{i,t'}}{\tau^t}\right)\right)} \tag{4}$$

where $\mathbb{I}$ is the indicator function, and Ω denotes the overlapping time indices between two views. $r_{i,t}$ and $\tilde{r}_{i,t}$ represent the feature representations of the i-th sample at timestamp t obtained from two different augmentations, and τ^t is the temporal-level temperature coefficient.

By aggregating over all samples and timestamps, the overall temporal-level contrastive loss is expressed as:

$$L_{\text{temp}} = \frac{1}{N\cdot T}\sum_i\sum_t \ell_{\text{temp}}^{(i,t)} \tag{5}$$

This formulation allows the model to maintain timestamp-level consistency across augmentations, reinforcing its ability to model temporal dependencies.

Instance-Level Contrastive Loss The instance-level contrastive learning component operates at the sample level to enhance discriminative representation learning across different time series instances. The objective is to pull closer the representations of different augmented versions of the same time series instance and push apart those of different instances (Li et al. 2021). Formally, the instance-level contrastive loss at timestamp t for the i-th sample is given by:

$$\ell_{\text{inst}}^{(i,t)} = -\log \frac{\exp\left(\frac{r_{i,t}\cdot\tilde{r}_{i,t}}{\tau^I}\right)}{\sum_{j=1}^{N}\left(\exp\left(\frac{r_{i,t}\cdot\tilde{r}_{i,t}}{\tau^I}\right)+\mathbb{I}_{[i\neq j]}\exp\left(\frac{r_{i,t}\cdot r_{j,t}}{\tau^I}\right)\right)} \tag{6}$$

where N is the number of samples in the dataset, $\mathbb{I}$ denotes the indicator function, and τ^I is the instance-level temperature coefficient.

The total instance-level contrastive loss is computed by aggregating over all timestamps and instances as:

$$L_{\text{inst}} = \frac{1}{N\cdot T}\sum_i\sum_t \ell_{\text{inst}}^{(i,t)} \tag{7}$$

The instance-level contrastive loss operates in synergy with the temporal-level objective within the lower feature space, where their complementary roles jointly capture rich temporal patterns.

Cluster-Level Contrastive Loss To further capture high-level semantic structures and promote clustering-friendly representation learning, we incorporate a cluster-level contrastive loss that operates on the distribution of samples over clusters. This loss is designed to increase similarity between representations belonging to the same cluster while enforcing separation across different clusters, thereby enhancing intra-cluster compactness and inter-cluster separability. Positive and negative pairs are constructed at the cluster level by comparing the cluster assignment distributions obtained from different augmented views of the data.

Specifically, the encoder output is first passed through a dedicated cluster-level contrastive head, denoted as $g_\phi^C(\cdot)$, which consists of a two-layer MLP followed by a softmax activation. Given a pooled feature vector r_i, the soft label representation is computed as:

$$\begin{aligned} z_i^C = g_\phi^C(r_i), \quad \tilde{z}_i^C = g_\phi^C(\tilde{r}_i), \\ g_\phi^C(h) = \text{softmax}(\text{MLP}^2(h)), \quad h \in \{r_i, \tilde{r}_i\} \end{aligned} \tag{8}$$

The soft label $z_i^C \in \mathbb{R}^k$ represents the distribution over k clusters for the i-th sample. Following the "labels as representations" principle, the i-th column of the soft assignment matrix $Z^C \in \mathbb{R}^{n\times k}$ corresponds to the representation of the i-th cluster, with $Z_{(\cdot,i)}^C$ and $\tilde{Z}_{(\cdot,i)}^C$ denoting the cluster-wise representations from two augmented views. Cosine similarity is used to quantify the alignment between representations:

$$s(z,\tilde{z}) = \frac{z^\top \tilde{z}}{\|z\|\cdot\|\tilde{z}\|} \tag{9}$$

The contrastive loss for the i-th cluster is defined as:

$$\mathcal{L}_{Z_{(\cdot,i)}^C} = -\log \frac{\exp(s(Z_{(\cdot,i)}^C, \tilde{Z}_{(\cdot,i)}^C)/\tau^C)}{\sum_{\substack{j=1\\ j\neq i}}^{k}\exp(s(Z_{(\cdot,i)}^C, \tilde{Z}_{(\cdot,j)}^C)/\tau^C)+\exp(s(Z_{(\cdot,i)}^C, Z_{(\cdot,j)}^C)/\tau^C)} \tag{10}$$

where τ^C is the cluster-level temperature coefficient. To prevent degenerate solutions where most samples collapse into a single cluster, we introduce a cross-entropy-based regularization:

$$\mathcal{L}_{\text{ce}} = -\sum_{i=1}^{k}\left(P_i^C \log P_i^C + \tilde{P}_i^C \log \tilde{P}_i^C\right) \tag{11}$$

where $P_i^C = \frac{1}{n}\sum_{j=1}^n Z_{(j,i)}^C$ and $\tilde{P}_i^C = \frac{1}{n}\sum_{j=1}^n \tilde{Z}_{(j,i)}^C$.

This term ensures a more balanced cluster assignment. By aggregating over all cluster-level losses, the total cluster-level contrastive loss is defined as:

$$\mathcal{L}_{\text{clu}} = \frac{1}{2k}\sum_{i=1}^{k}\left(\mathcal{L}_{Z_{(\cdot,i)}^C} + \mathcal{L}_{\tilde{Z}_{(\cdot,i)}^C}\right) + \mathcal{L}_{\text{ce}} \tag{12}$$

Finally, we unify all three contrastive losses at the temporal, instance, and cluster levels into the overall training objective.

$$\mathcal{L} = \mathcal{L}_{\text{temp}} + \mathcal{L}_{\text{inst}} + \mathcal{L}_{\text{clu}} \tag{13}$$

This multi-level contrastive learning framework enables the model to progressively capture temporal dynamics, instance-level distinctions, and cluster-level semantics, which collectively improve the quality of learned representations and the effectiveness of downstream clustering tasks.

Dataset	NMI							RI						
	STCN	TCGAN	CDCC	TimesURL	TS2Vec	R-Clust	ours	STCN	TCGAN	CDCC	TimesURL	TS2Vec	R-Clust	ours
AGWX	0.152	0.273	0	0.263	0.3	0.28	**0.343**	0.773	0.724	0.099	0.819	0.762	0.816	**0.847**
AGWY	0.188	0.221	0	0.25	0.126	0.189	**0.281**	0.768	0.787	0.099	0.831	0.606	0.811	**0.84**
AGWZ	0.133	0.205	0	0.257	0.173	0.228	**0.43**	0.805	0.803	0.099	0.824	0.754	0.82	**0.867**
Beef	0.244	0.29	0.006	0.234	0.293	0.247	**0.296**	0.668	0.639	0.297	0.671	**0.726**	0.67	0.703
CBF	0.393	0.983	0.983	0.793	0.844	0.95	**1**	0.728	0.996	0.996	0.92	0.922	0.984	**1**
Car	0.269	0.245	0.26	0.167	**0.526**	0.503	0.443	0.705	0.64	0.71	0.659	**0.795**	0.721	0.756
CricketX	0.174	0.343	0.324	0.244	0.052	0.337	**0.419**	0.824	0.849	0.873	0.75	0.519	0.869	**0.883**
CricketY	0.216	0.373	**0.438**	0.176	0.164	0.389	0.421	0.824	0.862	**0.888**	0.767	0.664	0.874	0.885
CricketZ	0.168	0.337	0.28	0.209	0.059	0.331	**0.366**	0.822	0.852	0.868	0.715	0.52	0.864	**0.877**
DLD	0.285	**0.396**	0	0.164	0.348	0.392	0.378	0.761	0.727	0.138	0.728	0.704	0.782	**0.789**
ECGFiveDays	0.019	0.002	0.526	0.071	0.24	0.02	**0.627**	0.513	0.5	0.817	0.548	0.648	0.513	**0.862**
FaceAll	0.313	0.462	0.624	0.345	0.382	0.661	**0.843**	0.871	0.893	0.925	0.797	0.77	0.916	**0.959**
FaceFour	0.328	0.618	0.39	0.357	**1**	0.573	0.774	0.739	0.81	0.758	0.699	**1**	0.79	0.901
FacesUCR	0.358	0.708	0.626	0.268	0.378	0.648	**0.864**	0.873	0.935	0.926	0.676	0.718	0.919	**0.963**
FiftyWords	0.444	**0.719**	0.603	0.278	0.343	0.665	0.704	0.905	**0.959**	0.951	0.852	0.749	0.956	0.958
FordB	0.003	**0.291**	0.011	0.05	0.071	0.035	0.253	0.502	0.604	0.508	0.532	0.548	0.522	**0.609**
GMAD3	0.294	0.429	0	0.369	0.307	**0.537**	0.43	0.883	0.879	0.036	0.92	0.905	0.933	**0.935**
GPZ1	0.21	0.382	0	0.05	0.459	0.452	**0.779**	0.764	0.779	0.164	0.653	0.803	0.792	**0.918**
GPZ2	0.165	0.357	0	0.055	0.305	0.455	**0.733**	0.755	0.782	0.164	0.708	0.73	0.792	**0.892**
GPOVY	0.001	0.344	**1**	0.344	**1**	**1**	**1**	0.499	0.619	**1**	0.619	**1**	**1**	**1**
Lightning7	0.302	0.488	0.475	0.403	0.449	**0.53**	0.509	0.77	0.771	0.828	0.793	0.81	0.816	**0.831**
Meat	0.441	0.521	0.014	0.586	0.576	**0.708**	0.644	0.704	0.739	0.558	0.789	0.781	**0.861**	0.76
MP	0.251	0.571	0	0.551	0.668	0.621	**0.669**	0.705	0.809	0.1	0.869	**0.909**	0.896	0.906
MPOAG	0.394	0.397	0.391	0.368	0.391	0.4	**0.423**	0.732	0.732	**0.735**	0.729	0.734	0.733	0.725
MoteStrain	0.237	0.006	0.496	0.011	0.615	0.453	**0.758**	0.656	0.506	0.798	0.509	0.846	0.768	**0.909**
OSULeaf	0.201	0.282	0.464	0.111	0.289	0.457	**0.576**	0.744	0.731	0.814	0.686	0.711	0.81	**0.853**
PGWZ	0.548	**0.719**	0	0.432	0.705	0.602	0.659	0.866	**0.914**	0.091	0.843	0.9	0.876	0.905
PPTW	0.543	0.567	0.512	0.569	0.517	0.553	**0.591**	0.802	**0.863**	0.792	0.801	0.787	0.78	0.787
SGWZ	0.617	0.694	0	0.642	0.855	0.719	**0.916**	0.869	0.854	0.091	0.885	0.956	0.914	**0.978**
ShapeletSim	0	0.015	0.32	0.033	**1**	**1**	**1**	0.498	0.507	0.703	0.52	**1**	**1**	**1**
ShapesAll	0.485	0.714	0.635	0.387	0.469	0.756	**0.793**	0.905	0.964	0.968	0.809	0.848	0.982	**0.983**
SS	0.316	0.369	0.394	0.234	0.537	0.304	**1**	0.694	0.717	0.667	0.65	0.74	0.663	**1**
SARS1	0.672	0.573	0.761	0.291	0.747	0.634	**0.859**	0.876	0.823	0.905	0.684	0.92	0.871	**0.953**
SARS2	0.24	0.258	0.478	0.116	0.369	0.415	**0.541**	0.651	0.668	0.755	0.585	0.681	0.74	**0.765**
SwedishLeaf	0.439	0.742	0.615	0.554	0.626	**0.76**	0.749	0.844	0.916	0.923	0.913	0.913	0.942	**0.947**
Symbols	0.72	0.84	0.82	0.724	0.92	0.952	**0.965**	0.884	0.917	0.934	0.883	0.975	0.987	**0.991**
SC	0.582	0.808	0.792	0.805	0.791	0.806	**0.995**	0.83	0.872	0.922	0.927	0.908	0.897	**0.999**
TS2	0.037	0.026	0.258	0.003	0.284	0.205	**0.31**	0.513	0.504	0.602	0.497	0.602	**0.694**	0.592
TwoPatterns	0.116	0.005	0.018	0.006	0.208	0.32	**0.39**	0.655	0.622	0.631	0.575	0.686	**0.725**	0.724
WS	0.282	0.545	0.464	0.204	0.168	0.472	**0.554**	0.849	0.906	0.9	0.766	0.675	0.899	**0.909**
AVG NMI/RI	0.295	0.428	0.349	0.299	0.464	0.514	**0.632**	0.751	0.774	0.626	0.735	0.781	0.830	**0.874**
Best	0	4	2	0	4	6	**29**	0	3	3	0	6	5	**28**
AVG Rank	5.7	3.675	4.7	5.5	3.85	2.93	**1.35**	5.15	4.425	4.175	5.175	3.925	3.15	**1.6**

Table 1: Overall performance comparison. AVG NMI/RI indicates average of NMI or RI over all datasets.

Clustering

TFMCC unifies feature learning and clustering within a single framework, enabling the model to learn representations that are inherently aligned with the clustering objective. In this framework, the cluster-level representation Z^C serves as the foundation for determining the final cluster assignment Y. Specifically, the cluster label for each sample is obtained by identifying the cluster with the highest predicted probability in the soft assignment vector generated by the cluster-level contrastive head. Formally, the cluster assignment can be expressed as:

$$Y = \arg\max \left(g_\phi^C(\text{Encoder}(X)) \right) \tag{14}$$

where $g_\phi^C(\cdot)$ denotes the cluster-level contrastive head, and $\text{Encoder}(\cdot)$ represents the time series encoder. This formulation allows the learned representations to be directly mapped to clustering outcomes, ensuring consistency between the representation space and the clustering structure.

Experiments

Experimental Setup

Dataset and Evaluation Metrics To validate the effectiveness of the proposed model, experiments are conducted on 40 datasets from the UCR[1] (Dau et al. 2019). The cluster-

[1]AGWX: AllGestureWiimoteX, AGWY: AllGestureWiimoteY, AGWZ: AllGestureWiimoteZ, DLD: DodgerLoopDay, GMAD3:

ing performance is evaluated using two widely used metrics: Normalized Mutual Information(NMI) (Strehl and Ghosh 2002) and Rand Index(RI) (Rand 1971).

Baseline Methods To evaluate the performance of TFMCC, this paper selects six representative methods for comparative experiments, including one self-supervised model (STCN (Ma et al. 2020)) and five unsupervised representation learning models (TCGAN(Huang and Deng 2023), CDCC (Peng et al. 2024), R-Clust (Jorge and Rubén 2024), TS2Vec (Yue et al. 2022), TimesURL (Liu and Chen 2024)). Among them, TS2Vec and TimesURL adopt separated optimization strategies, with k-means used for clustering.

Model Architecture and Experiment Details For the comparative methods, we use open-source code implementations and follow the parameter configurations and experimental analyses specified in their respective papers. In the TFMCC model, the temperature coefficients at the time level, instance level, and cluster level are set as $\tau^t = 0.2$, $\tau^I = 0.2$, and $\tau^C = 0.5$. To optimize the network training process, we select the Adam optimizer (Yao et al. 2021). The experiments are conducted on a computer equipped with an AMD 9800X3D processor, a 5090D (32GB) graphics card, and 64 GB of memory.

The encoder consists of an input projection layer and an expanded CNN module. For each input x_i, the input projection layer is a fully connected layer that maps each time step's observation $x_{i,t}$ to a high-dimensional space, forming a vector $z_{i,t}$. The expanded CNN module includes ten dilated convolution blocks, with each block containing a residual block and a one-dimensional convolution layer. Each block includes two one-dimensional convolution layers with dilation parameters (the dilation parameter for the l-th block is 2^l), providing a larger receptive field (Bai, Kolter, and Koltun 2018).

Overall Performance Comparing

In Table 1, we compare the proposed TFMCC method with STCN, TCGAN, TS2Vec, TimesURL, CDCC, and R-Clust. Bold text indicates that the method ranks first in the corresponding dataset, while underlined text indicates that the method ranks second in the corresponding dataset. TFMCC achieves the highest NMI in 29 out of 40 datasets and the best RI in 28 datasets. It also achieves the highest average NMI (0.632), the highest average RI (0.874), and the highest average rank, with values of 1.45 (NMI) and 1.675 (RI). Notably, TFMCC achieves an NMI and RI of 1 in the CBF, GPOVY, ShapeletSim, and SS.

In the comparison of various deep clustering methods, TFMCC performs excellently on most datasets. Compared to methods that separate representation and clustering (such as TS2Vec), TFMCC further improves clustering performance by jointly optimizing representation learning and the clustering objective. Compared to traditional methods that jointly optimize representation and clustering (such as STCN, TCGAN, etc.), TFMCC achieves higher-quality clustering through multi-scale augmentation in both time and frequency domains. Although CDCC combines both time-domain and frequency-domain information, TFMCC utilizes multi-level contrastive learning to leverage the multi-level information in time series data, resulting in stronger clustering performance.

GestureMidAirD3, GPZ1: GesturePebbleZ1, GPZ2: GesturePebbleZ2, GPOVY: GunPointOldVersusYoung, MP: MelbournePedestrian, MPOAG: MiddlePhalanxOutlineAgeGroup, PGWZ: PickupGestureWiimoteZ, PPTW: ProximalPhalanxTW, SGWZ: ShakeGestureWiimoteZ, SS: SmoothSubspace, SARS1: SonyAIBORobotSurface1, SARS2: SonyAIBORobotSurface2, SC: SyntheticControl, TS2: ToeSegmentation2, WS: WordSynonyms

	Avg.NMI
TFMCC	**0.632**
w/o TLCL	0.591(-4.2%)
w/o ILCL	0.545(-8.7%)
w/o CLCL	0.584(-4.8%)
w/o MTDA	0.565(-6.8%)
w/o MFDA	0.566(-6.6%)
w/o MTFA	0.545(-8.7%)

Table 2: Ablation results on 40 UCR datasets.

Ablation Study

To validate the effectiveness of the components proposed in TFMCC, Table 2 shows a comparison between the full TFMCC and its six variants across 40 UCR datasets. Specifically: (1) w/o TLCL removes the Temporal-Level Contrastive Loss, (2) w/o ILCL removes the Instance-Level Contrastive Loss, (3) w/o CLCL removes the Cluster-Level Contrastive Loss, performs k-means instead, (4) w/o MTDA replaces the Multi-scale Time Domain Augmentation with subseries consistency (Franceschi, Dieuleveut, and Jaggi 2019), (5) w/o MFDA removes the Multi-scale Frequency Domain Augmentation, and (6) w/o MTFA replaces Multi-scale Temporal-Frequency Augmentation with subseries consistency.

The ablation studies demonstrate that TFMCC's performance relies on the synergistic operation of all components. Removing the temporal-level contrastive loss reduces NMI to 0.591 (4.2% decrease), confirming its importance in capturing temporal dynamics. The instance-level contrastive loss proves most critical, with its removal causing the largest NMI drop to 0.545 (8.7% decrease), as it directly optimizes inter-sample relationships. Similarly, eliminating the cluster-level contrastive loss lowers NMI to 0.584 (4.8% decrease), underscoring its role in cluster structure learning.

The multi-scale augmentation strategies also show substantial impact: time-domain augmentation removal decreases NMI by 6.6%, while frequency-domain removal causes a 6.8% reduction. When both augmentations are disabled (MTFA), performance drops by 8.7%, revealing their complementary nature in capturing multi-scale time-frequency features.

These results collectively validate that each module contributes uniquely to TFMCC's effectiveness.

To further assess the effectiveness of multi-scale diversity in the time and frequency domains, we introduce two controlled variants: w/ FTDA&MFDA and w/ FFDA&MTDA.

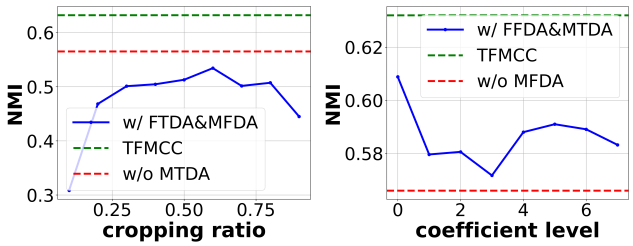

Figure 3: Effectiveness of Multi-Scale Data Augmentation

In w/ FTDA&MFDA, we fixed the cropping ratio for time domain augmentation throughout training, while still applying multi-scale frequency domain augmentation. w/ FFDA&MTDA applies a fixed-level DWT coefficient frequency augmentation across training iterations, with multi-scale time-domain augmentation still operating.

Figure 3 presents the average NMI results on the 40 UCR datasets. In the left subfigure, the green dashed line denotes the full method, the red dashed line represents w/o MTDA, and the blue curve shows the results of w/ FTDA&MFDA across different fixed cropping ratios. The performance of w/ FTDA&MFDA consistently falls below both the full model and w/o MTDA, indicating that rigid single-scale temporal cropping not only limits the ability to capture diverse temporal structures but also performs worse than strategies leveraging subseries consistency.

In the right subfigure, the NMI performance of w/ FFDA&MTDA under different fixed-level DWT coefficients is shown. The blue curve lies above w/o MFDA (red dashed) but below the full method (green dashed), demonstrating that even fixed-scale frequency-domain augmentation brings measurable gains, while full multi-scale augmentation further improves performance by better capturing frequency-specific characteristics.

Overall, these results highlight the critical role of multi-scale augmentation in both domains. By introducing dynamic diversity across time and frequency resolutions, the model benefits from enhanced generalization capability, thereby achieving superior clustering performance.

Parameter Analysis

To assess the impact of temperature coefficients on model performance, we analyze the key parameters of TFMCC, namely the time-level (τ^t), instance-level (τ^I), and cluster-level (τ^C) temperature coefficients, using the FaceAll, FacesUCR, SmoothSubspace, and SyntheticControl datasets. As illustrated in Figure 4, although some datasets exhibit fluctuations under specific temperature coefficient settings, the clustering ability of the model remains generally stable, with no significant impact on clustering performance.

Convergence Analysis

We analyze the clustering quality convergence of TFMCC on the FaceAll, FacesUCR, SmoothSubspace, and SyntheticControl datasets, as shown in Figure 5. The results indicate that as the number of epochs increases, the model's clustering performance steadily improves until it reaches convergence. These findings suggest that the proposed clustering model exhibits ideal convergence behavior.

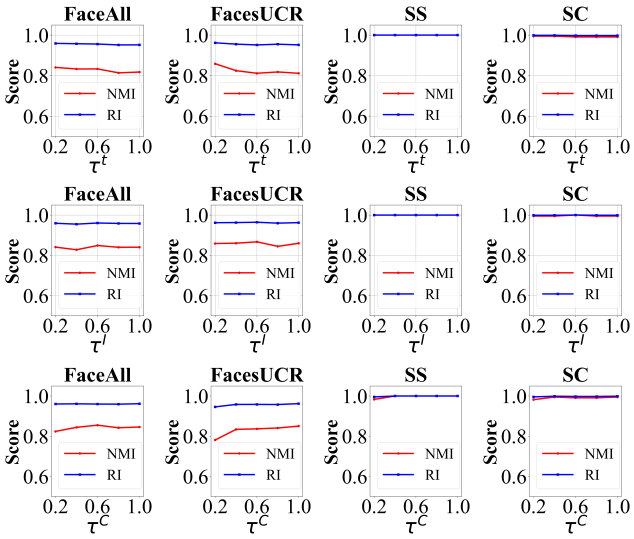

Figure 4: The impact of temperature coefficients

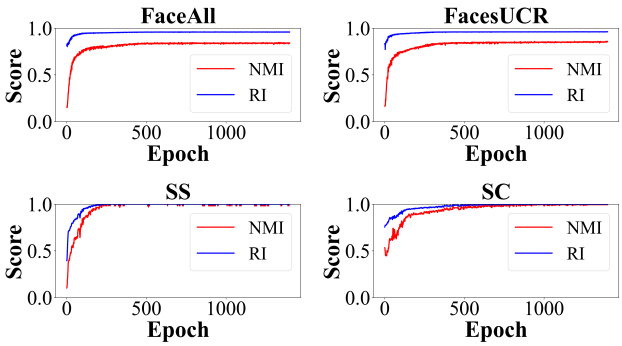

Figure 5: The convergence of clustering performance.

Conclusion

This paper presents an innovative time series clustering framework (TFMCC), which significantly enhances the model's representation ability and clustering performance through multi-scale time-frequency augmentation and multi-level contrastive learning. Experiments on 40 benchmark datasets show that TFMCC outperforms existing methods in overall performance. Ablation experiments further validate the effectiveness of the key modules. Future work will focus on exploring its application in multimodal data.

Acknowledgments

This work is supported by the Qinglan Project of Jiangsu Province of China, the National Natural Science Foundation of China (No. 62006104), Postgraduate Research & Practice Innovation Program of Jiangsu Normal University (No. 2024XKT2583).

References

Alqahtani, A.; Ali, M.; Xie, X.; and Jones, M. W. 2021. Deep time-series clustering: A review. *Electronics*, 10(23): 3001.

Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Baytas, I. M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A. K.; and Zhou, J. 2017. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 65–74.

Chen, I. Y.; Krishnan, R. G.; and Sontag, D. 2022. Clustering interval-censored time-series for disease phenotyping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 6211–6221.

Cini, A.; Mandic, D. P.; and Alippi, C. 2024. Graph-based Time Series Clustering for End-to-End Hierarchical Forecasting. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Dai, T.; Wu, B.; Liu, P.; Li, N.; Yuerong, X.; Xia, S.-T.; and Zhu, Z. 2024. DDN: Dual-domain dynamic normalization for non-stationary time series forecasting. *Advances in Neural Information Processing Systems*, 37: 108490–108517.

Dau, H. A.; Bagnall, A.; Kamgar, K.; Yeh, C.-C. M.; Zhu, Y.; Gharghabi, S.; Ratanamahatana, C. A.; and Keogh, E. 2019. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6): 1293–1305.

Deng, X.; Huang, D.; Chen, D.-H.; Wang, C.-D.; and Lai, J.-H. 2023. Strongly augmented contrastive clustering. *Pattern Recognition*, 139: 109470.

Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C. K.; Li, X.; and Guan, C. 2021. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*.

Franceschi, J.-Y.; Dieuleveut, A.; and Jaggi, M. 2019. Unsupervised scalable representation learning for multivariate time series. *Advances in Neural Information Processing systems*, 32.

Hu, Y.; Liu, P.; Zhu, P.; Cheng, D.; and Dai, T. 2025. Adaptive multi-scale decomposition framework for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17359–17367.

Huang, F.; and Deng, Y. 2023. TCGAN: Convolutional Generative Adversarial Network for time series classification and clustering. *Neural Networks*, 165: 868–883.

Jorge, M.-B.; and Rubén, C. 2024. Time series clustering with random convolutional kernels. *Data Mining and Knowledge Discovery*, 1–27.

Li, H.; and Liu, Z. 2021. Multivariate time series clustering based on complex network. *Pattern Recognition*, 115: 107919.

Li, Y.; Hu, P.; Liu, Z.; Peng, D.; Zhou, J. T.; and Peng, X. 2021. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8547–8555.

Liu, J.; and Chen, S. 2024. Timesurl: Self-supervised contrastive learning for universal time series representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13918–13926.

Ma, Q.; Li, S.; Zhuang, W.; Wang, J.; and Zeng, D. 2020. Self-supervised time series clustering with model-based dynamics. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9): 3942–3955.

Madiraju, N. S. 2018. *Deep temporal clustering: Fully unsupervised learning of time-domain features*. Master's thesis, Arizona State University.

Mallat, S. G. 2002. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7): 674–693.

McQueen, J. B. 1967. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, 281–297.

Murad, M. M. N.; Aktukmak, M.; and Yilmaz, Y. 2025. Wpmixer: Efficient multi-resolution mixing for long-term time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19581–19588.

Peng, F.; Luo, J.; Lu, X.; Wang, S.; and Li, F. 2024. Cross-Domain Contrastive Learning for Time Series Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8921–8929.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336): 846–850.

Strehl, A.; and Ghosh, J. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec): 583–617.

Sun, X.; Jin, J.; Wang, H.; Sun, X.; Wang, X.; and Zhu, J. 2025. Riding the Wave: Multi-Scale Spatial-Temporal Graph Learning for Highway Traffic Flow Prediction Under Overload Scenarios. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 3317–3325.

Tiano, D.; Bonifati, A.; and Ng, R. 2021. FeatTS: Feature-based time series clustering. In *Proceedings of the 2021 International Conference on Management of Data*, 2784–2788.

Tonekaboni, S.; Eytan, D.; and Goldenberg, A. 2021. Unsupervised Representation Learning for Time Series with Temporal Neighborhood Coding. In *International Conference on Learning Representations*.

Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J. Y.; and ZHOU, J. 2024. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. In *International Conference on Learning Representations (ICLR)*.

Wang, Y.; Han, Y.; Wang, H.; and Zhang, X. 2023. Contrast everything: A hierarchical contrastive framework for medical time-series. *Advances in Neural Information Processing Systems*, 36: 55694–55717.

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16051–16060.

Yang, Z.; Li, H.; Tuo, X.; Li, L.; and Wen, J. 2023. Unsupervised clustering of microseismic signals using a contrastive learning model. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–12.

Yao, Z.; Gholami, A.; Shen, S.; Mustafa, M.; Keutzer, K.; and Mahoney, M. 2021. Adahessian: An adaptive second order optimizer for machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10665–10673.

Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; and Xu, B. 2022. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8980–8987.

Zhang, X.; Zhao, Z.; Tsiligkaridis, T.; and Zitnik, M. 2022. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35: 3988–4003.

Zhong, H.; Wu, J.; Chen, C.; Huang, J.; Deng, M.; Nie, L.; Lin, Z.; and Hua, X.-S. 2021. Graph contrastive clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9224–9233.

Zhong, Y.; Huang, D.; and Wang, C.-D. 2023. Deep temporal contrastive clustering. *Neural Processing Letters*, 55(6): 7869–7885.