– APPENDIX –
TFEC: MULTIVARIATE TIME-SERIES CLUSTERING VIA TEMPORAL-FREQUENCY ENHANCED CONTRASTIVE LEARNING

Zexi Tan[a], Tao Xie[a], Haoyi Xiao[a], Baoyao Yang[a], Yuzhu Ji[a], An Zeng[a], Xiang Zhang[a], Yiqun Zhang[a,b,]*

[a]Guangdong University of Technology, [b]Hong Kong Baptist University

1. PSEUDO CODE FOR TFEC

This section describes Algorithm 1 for TFEC in detail.

Algorithm 1 TFEC: Temporal-Frequency Enhanced Contrastive Learning Framework

Require: Multivariate time-series dataset $\mathcal{X}$, number of clusters k, hyperparameters α, β.
Ensure: Cluster assignments $\mathbf{Y}$, representation $\mathbf{R}$.

Temporal-Frequency Co-Enhancement:
for each $\mathbf{x}_i \in \mathcal{X}$ **do**
  Identify proximate neighbors $\mathcal{N}_i$ via density wave detection
  Apply aligned cropping to obtain segments x_i^L, $x_{(i,p)}^L$
  Compute FFT: $\mathbf{q}_i^L = \mathcal{F}(x_i^L)$, $\mathbf{q}_{(i,p)}^L = \mathcal{F}(x_{(i,p)}^L)$
  Mix frequencies: $\mathbf{F} = \mathbf{q}_i^L + \sum_p \delta_p \cdot \mathbf{q}_{\langle i,p \rangle}^L$
  Synthesize enhanced sample via inverse FFT
end for
Form Enhanced MTS Embedding (EME)
Dual-Path Learning:
repeat
  PGCL Path:
  Extract encodings $\mathbf{r}, \mathbf{r}'$ from dual views
  Fuse views: $\mathbf{R} = (\mathbf{r} + \mathbf{r}')/2$
  Initialize clusters via K-means on $\mathbf{R}$
  Compute confidence scores CONF_i (Eq. 3)
  Construct contrastive pairs $\mathcal{P}$, $\mathcal{N}$ using high-confidence samples
  Update embeddings via $\mathcal{L}_{\text{con}}$ (Eq. 4)
  READ Path:
  Reconstruct masked EME via autoencoder
  Compute reconstruction loss $\mathcal{L}_{\text{recon}}$
  Update model parameters by minimizing $\mathcal{L}_{\text{total}} = \beta\mathcal{L}_{\text{con}} + (1-\beta)\mathcal{L}_{\text{recon}}$
until convergence

2. DETAILED EXPERIMENTAL SETTINGS

This section mainly shows detailed experiment settings and other extended experiments to validate the proposed method.

2.1. Dataset Description

To rigorously evaluate the efficacy and generalizability of the proposed TFEC framework, this paper conducts extensive experiments on six real-world multivariate time-series (MTS) datasets from the UEA archive [1]. These datasets span diverse domains and exhibit variations in sequence length (T), dimensionality (F), sample size (N), and number of classes, as summarized in Table 1 of full paper. Below we provide a concise description of each dataset and its associated clustering challenges.

AtrialFibrillation: Comprises ECG recordings ($F = 2$, $T = 640$) for detecting arrhythmia types. The small sample size ($N = 15$) and subtle inter-class differences pose significant challenges for robust representation learning.

ERing: Contains motion sensor data ($F = 4$, $T = 65$) from hand gestures across six classes. The short sequence length and fine-grained motion patterns require precise temporal feature extraction.

RacketSports: Comprises accelerometer data ($F = 6$, $T = 30$) from four badminton shot types. The ultra-short sequences and high intra-class variability demand effective discrimination of rapid motion patterns.

Libras: Consists of movement trajectories ($F = 2$, $T = 45$) representing 15 Brazilian sign language gestures. The high number of classes and low-dimensional spatiotemporal signals test model capacity for fine-grained clustering.

StandWalkJump: Includes sensor recordings ($F = 4$, $T = 2500$) of three human activities. The extreme sequence length and temporal scale variations require modeling long-range dependencies while maintaining discriminability.

NATOPS: Contains motion capture data ($F = 24$, $T = 51$) from six aircraft handling gestures. The high dimensionality and complex multi-class structure challenge both representation learning and cluster separation.

The selected benchmark datasets enable comprehensive evaluation of TFEC across varying temporal lengths, dimensionalities, and clustering complexities, underscoring its practicality for real-world MTS analysis.

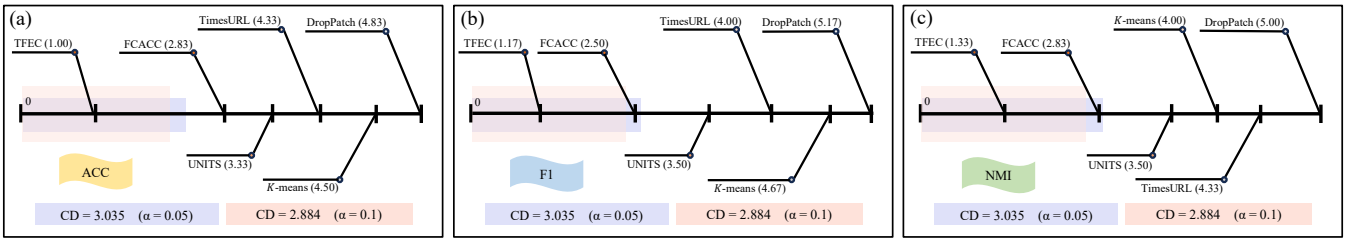
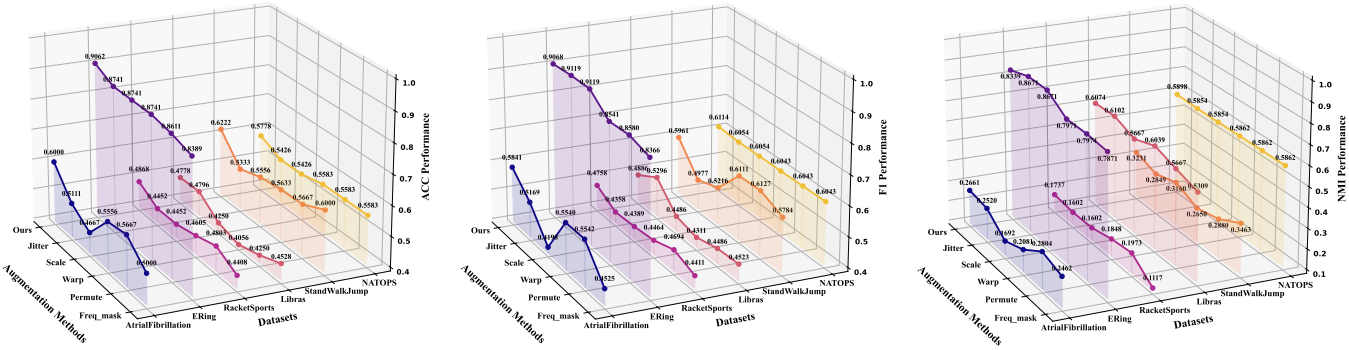
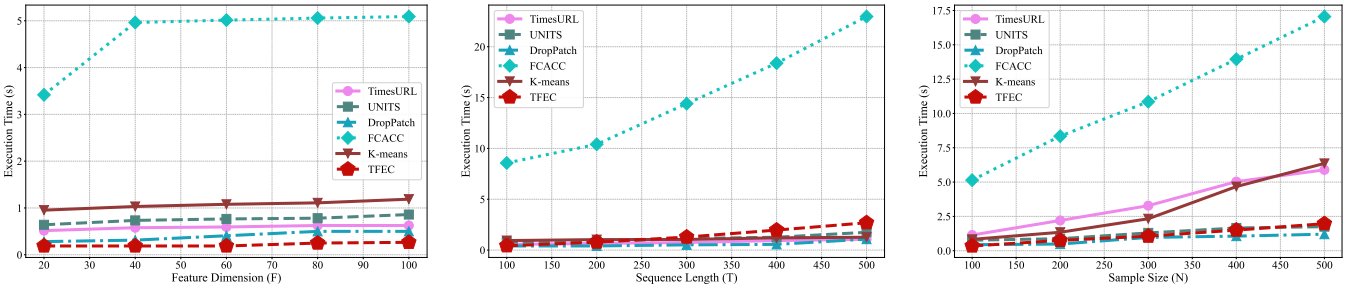


Fig. 1: BD test (a), (b) and (c) based on the average ranks of ACC, F1 and NMI in the Fig.2 of the full paper. Methods ranked outside the CD intervals are believed to perform significantly differently from TFEC.



(a) ACC Performance (b) F1 Performance (c) NMI Performance

Fig. 2: Comparison the proposed temporal-frequency enhancement against five common augmentation strategies: jitter, scale, warp, permute, and frequency mask (freq_mask).



(a) Execution time w.r.t. F (b) Execution time w.r.t. T (c) Execution time w.r.t. N

Fig. 3: Comparison of execution time with respect to different dimensions

2.2. Evaluation Metrics

To quantitatively assess the performance of TFEC framework and facilitate a fair comparison with state-of-the-art baselines, we employ three established metrics that evaluate clustering quality from complementary perspectives: alignment with ground truth labels, purity of the discovered clusters, and the overall harmonic mean of precision and recall.

Clustering Accuracy (ACC). ACC directly measures the alignment between the predicted cluster assignments and the ground truth labels. It is defined as the maximum accuracy achieved by finding the optimal one-to-one mapping between clusters and labels, which resolves the inherent permutation invariance in clustering. Formally, given the true labels $\mathbf{Y} = \{y_1, \ldots, y_N\}$ and the predicted cluster assignments $\mathbf{C} = \{c_1, \ldots, c_N\}$, ACC is computed as:

$$\text{ACC} = \max_m \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i = m(c_i)), \tag{1}$$

where m is the mapping function from clusters to labels, and $\mathbb{I}(\cdot)$ is the indicator function. ACC ranges from 0 to 1, with 1 indicating a perfect match.

F1-Score (F1). The F1-Score [2] provides a balanced measure of clustering performance by computing the harmonic mean of precision and recall. For each cluster, preci-

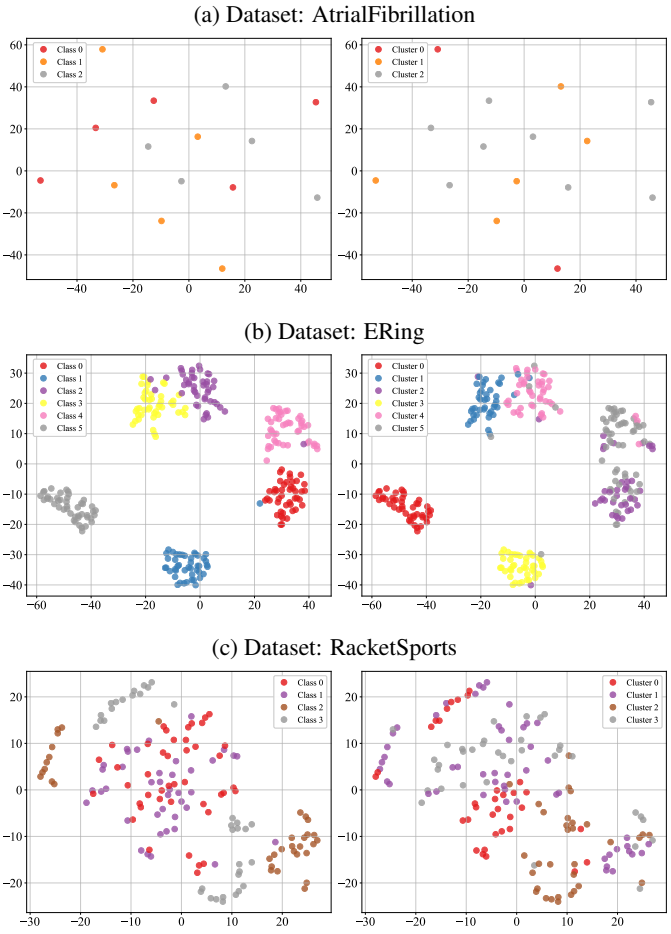


Fig. 4: Visualization of clustering results on AtrialFibrillation, ERing and RacketSports datasets. Left side of each subfigure shows the true label distribution, and right side shows the clustering label distribution.

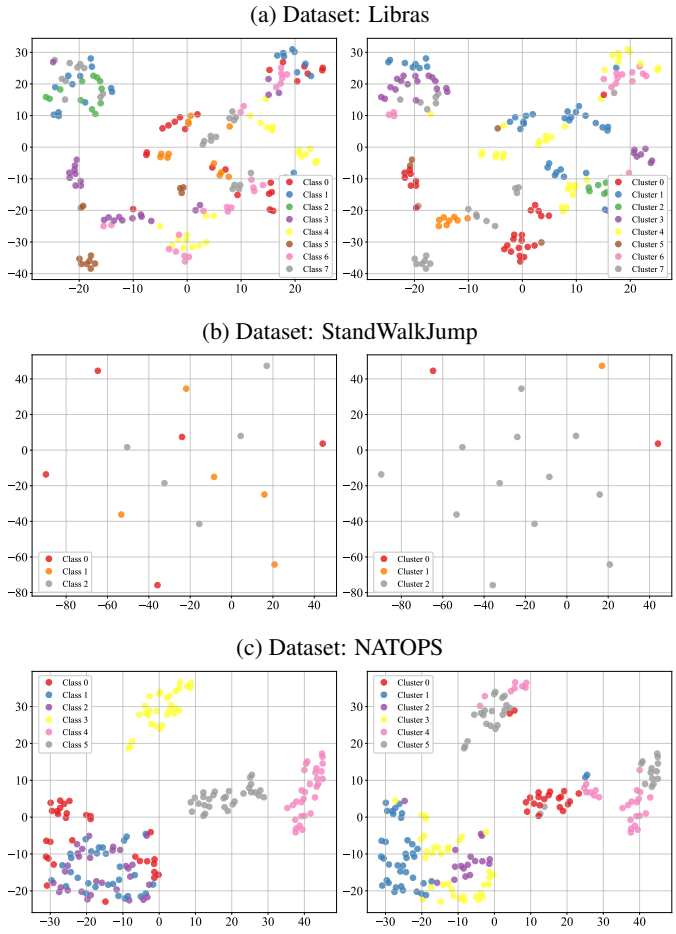


Fig. 5: Visualization of clustering results on Libras, StandWalkJump and NATOPS datasets. Left side of each subfigure shows the true label distribution, and right side shows the clustering label distribution.

sion and recall are calculated with respect to its best-matching true class. The macro-averaged F1-Score is then computed across all clusters:

$$\text{F1} = \frac{1}{K}\sum_{k=1}^{K}\text{F1}_k, \tag{2}$$

where F1_k is the F1-Score for the k-th cluster. This metric balances the trade-off between cluster compactness and completeness, with a value of 1 indicating perfect clustering.

Normalized Mutual Information (NMI). NMI assesses the quality of the discovered clusters by quantifying the statistical information shared between $\mathbf{C}$ and $\mathbf{Y}$. It normalizes the Mutual Information (MI) by the average entropy of the two distributions, making it invariant to the number of clusters. It is defined as:

$$\text{NMI}(\mathbf{Y}, \mathbf{C}) = \frac{2 \cdot I(\mathbf{Y}; \mathbf{C})}{H(\mathbf{Y}) + H(\mathbf{C})}, \tag{3}$$

where $I(\mathbf{Y}; \mathbf{C})$ is the mutual information, and $H(\cdot)$ denotes entropy. NMI yields a score between 0 and 1, with higher values indicating better cluster purity.

2.3. Significance Test

This paper also conducted BD tests to the results of clustering performance in Fig.2 of the full paper. And the test result based on CD interval is visualized in Fig. 1 in the Appendix. The lengths of the CD, when comparing six approaches on six real-world benchmark datasets, are 3.035 and 2.884 with $\alpha = 0.05$ and 0.1, respectively. It can be seen that TFEC significantly outperforms most of its counterparts, including four state-of-the-art methods, which indicates the competitiveness of TFEC in multivariate time-series clustering.

2.4. Comparison with Other Augmentation Mechanism

As shown in Fig. 2, this paper compares the proposed temporal-frequency enhancement against five common augmentation strategies: jitter, scale, warp, permute, and frequency mask (freq_mask). Results across all six datasets demonstrate the consistent superiority of TFEC. It achieves the highest or competitive performance across all metrics, with notable advantages on challenging datasets such as *AtrialFibrillation* and *NATOPS*. While certain augmentations (e.g., warp on *RacketSports*) occasionally perform well, they exhibit instability and performance degradation on more complex data. These findings confirm that TFEC's enhancement mechanism effectively preserves temporal structure and enriches discriminative features, outperforming heuristic augmentations that often introduce unrealistic distortions or break inherent periodicities.

2.5. Efficiency Evaluation

As shown in Fig. 3. This paper evaluates the computational efficiency of TFEC against five baselines under three scaling scenarios. When fixing $N = 50, T = 2$ and increasing feature dimension F from 20 to 100, TFEC maintains low runtime (0.1875s to 0.2656s), significantly outperforming FCACC. With fixed $N = 50, F = 2$ and increasing sequence length T from 100 to 500, TFEC scales linearly (0.4219s to 2.6875s), comparable to TimesURL and UNITS. Under fixed $T = 2, F = 2$ with sample size N increasing from 100 to 500, TFEC demonstrates sub-linear growth (0.3438s to 1.9688s), substantially more efficient than FCACC and TimesURL. Overall, TFEC achieves competitive scalability across all dimensions, offering practical efficiency for multivariate time-series clustering.

2.6. Visualization

To qualitatively evaluate the clustering performance of TFEC, we visualize the learned embeddings of all six datasets using t-SNE [3]. Fig. 4 and 5 present a comparative visualization between the ground truth labels (left) and the cluster assignments produced by TFEC (right).

The results show that TFEC effectively preserves the underlying manifold structure of the data. On datasets with clear separability such as *ERing*, the clusters are well-formed and distinct. For more challenging datasets like *Libras* (15 classes) and *NATOPS*, TFEC still maintains high intra-cluster cohesion and inter-cluster separation despite the increased complexity. Visualizations on *AtrialFibrillation* and *StandWalkJump* further confirm the model's ability to capture subtle temporal variations and model long-range dependencies, respectively. These visual outcomes align with quantitative metrics, demonstrating that TFEC learns discriminative, cluster-friendly representations that faithfully reflect the intrinsic structure of multivariate time-series data.

3. REFERENCES

[1] A. J. Bagnall et al., "The UEA multivariate time series classification archive," arXiv preprint arXiv:1811.00075, 2018.

[2] N. Chinchor, "MUC-4 evaluation metrics," in *MUC*, USA, 1992, pp. 22—-29.

[3] L. v. d. Maaten et al., "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.