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

Zexi Tan[a], Tao Xie[a], Haoyi Xiao[a], Baoyao Yang[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 $\mathbf{Y}$, embeddings $\mathbf{R}$.

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 Dataset (EMD)
Dual-Path Optimization:
repeat
  PGCL Path:
  Extract embeddings $\mathbf{r}, \mathbf{r}'$ from dual EMD views
  Fuse representations: $\mathbf{R} = (\mathbf{r} + \mathbf{r}')/2$
  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 EMD 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 UCR 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.

Beef: Comprises spectrographic MTS data from beef muscle across five quality classes. With $T = 470$ and $F = 5$, the dataset captures subtle biochemical variations, presenting challenges in distinguishing fine-grained classes under temporal variability and sensor noise.

Coffee: Contains chemical sensor recordings ($F = 5$, $T = 286$) from roasting coffee beans of two botanical origins (Arabica and Robusta). The clustering task is challenging due to high inter-class similarity and non-linear roasting profiles.

Adiac: Consists of shape outline sequences ($F = 5$, $T = 176$) extracted from animal images. The task involves clustering into 37 animal classes, requiring discriminative feature extraction from low-dimensional spatiotemporal signals.

ArrowHead: Includes outlines of archaeological arrowheads ($F = 5$, $T = 251$). The goal is to cluster artifacts into three morphological categories, testing sensitivity to subtle temporal-shape variations.

BME: Comprises motion sensor data ($F = 5$, $T = 128$) from body exercise activities. The small sample size ($N = 150$) and short sequences pose risks of overfitting, demanding robust representation learning.

Car: Contains sensor recordings ($F = 5$, $T = 577$) from vehicle operations. The long sequences and four-class structure require modeling prolonged temporal dependencies and distinguishing operational modes.

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

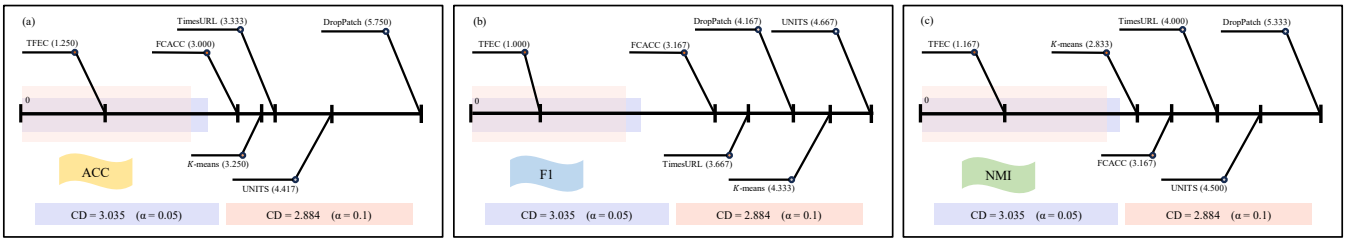
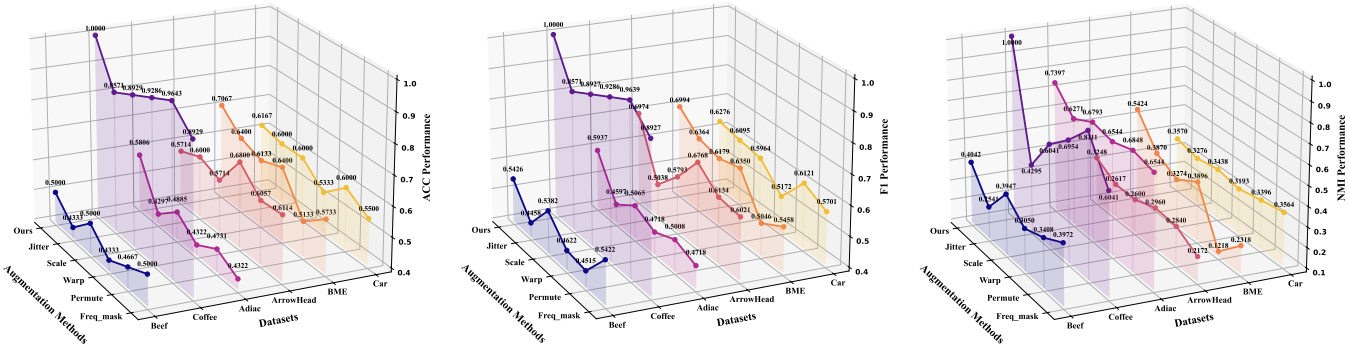
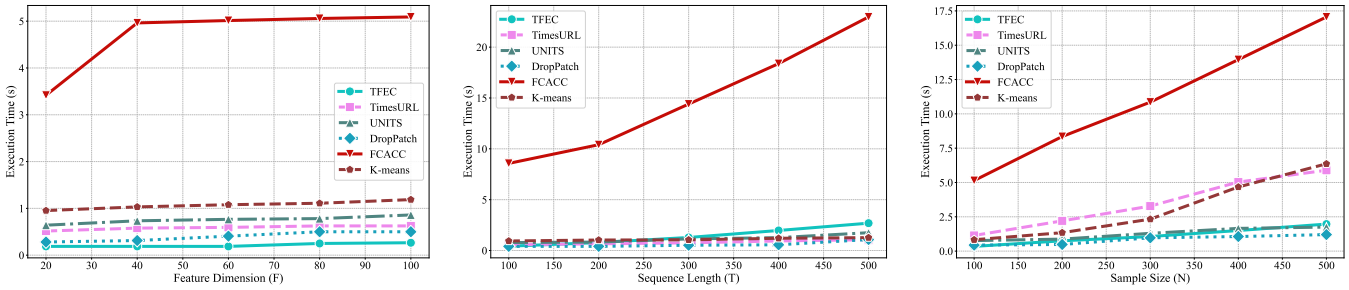


Fig. 1: BD test (a), (b) and (c) based on the average ranks of ACC, F1 and NMI in the Fig.2 of the full paper. Methods ranked outside the CD intervals are believed to perform significantly differently from TFEC.



(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-

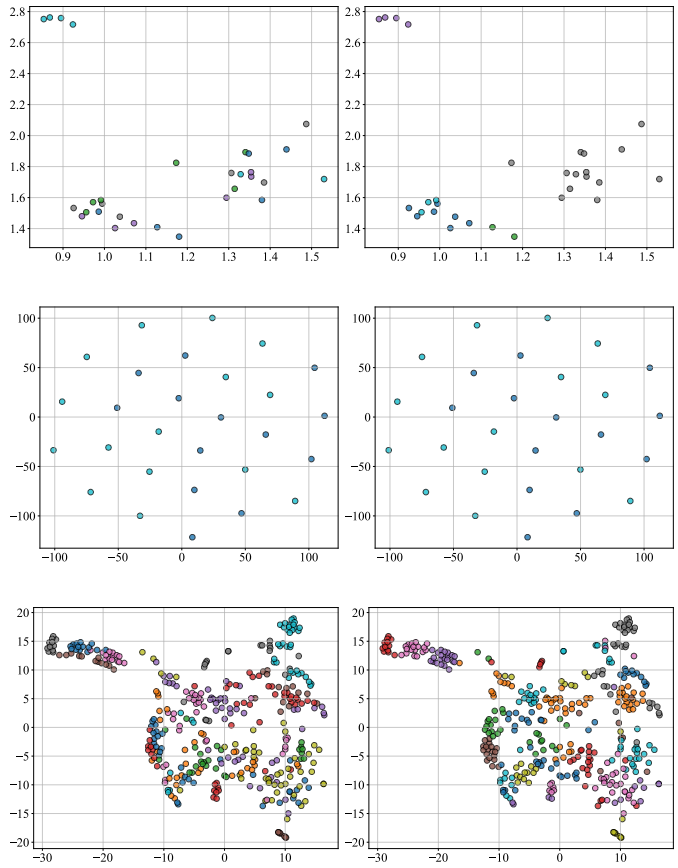


Fig. 4: Visualization of clustering results on Beef, Coffee and Adiac datasets. Left side of each subfigure shows the true label distribution, and right side shows the clustering label distribution.

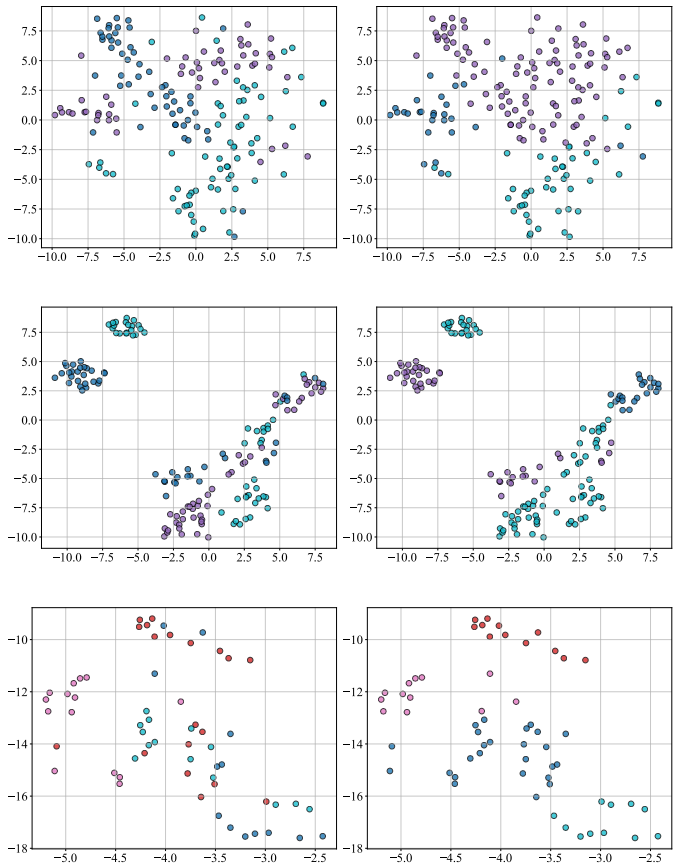


Fig. 5: Visualization of clustering results on ArrowHead, BME and Car datasets. Left side of each subfigure shows the true label distribution, and right side shows the clustering label distribution.

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 the result shown in Fig. 2, this paper further compares the proposed temporal-frequency enhancement against five

common augmentation strategies: jitter, scale, warp, permute, and frequency mask (freq_mask). Experimental results across all six datasets demonstrate the consistent superiority of our method. TFEC achieves the highest or competitive performance on all metrics, particularly excelling on complex datasets such as *Adiac* and *BME*. While some augmentations (e.g., warp on *ArrowHead*) occasionally perform well, they exhibit significant instability and degradation on more challenging data. These results 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 Test

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 demonstrate that TFEC effectively preserves the underlying manifold structure of the data. On datasets with clear separability such as *Coffee*, the clusters are nearly perfectly formed and separated. For more complex datasets like *Adiac* (37 classes) and *BME*, TFEC still maintains high intra-cluster cohesion and inter-cluster separation, despite the increased difficulty. The visualizations on *ArrowHead* and *Car* further confirm the model's capability to discern subtle temporal-shape variations and model long-range dependencies, respectively. These visual results align with the quantitative metrics, providing strong evidence that TFEC learns discriminative and cluster-friendly representations which faithfully reflect the intrinsic categorical structure of multivariate time-series data.

3. REFERENCES

[1] H. A. Dau et al., "The ucr time series archive," 2019.

[2] N. Chinchor, "Muc-4 evaluation metrics," in *Proceedings of the Conference on Message Understanding (MUC)*, USA, 1992, pp. 22—-29.

[3] L. v. d. Maaten, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.