A APPENDIX

A.1 VISUALIZATION

To make assessments about the clusterability of learned representations in the encoding space, we visualize the feature distribution by using t-SNE (Van der Maaten & Hinton, 2008). It is noted that if the information of the latent state is properly learned and encoded by the model, the representations from the same underlying state should cluster together. Figure 6 shows the comparisons about representations distribution of different models. It demonstrates that the representations learned by proposed BTSF from the same hidden state are better than the other approaches. The visualization results further prove the superior representation ability of our model. In Addition, we have evaluated on the all univariate time series datasets: the UCR archive. The corresponding critical difference diagram is shown in Figure 7. The BTSF significantly outperforms the other approaches with an average rank of almost 1.3.

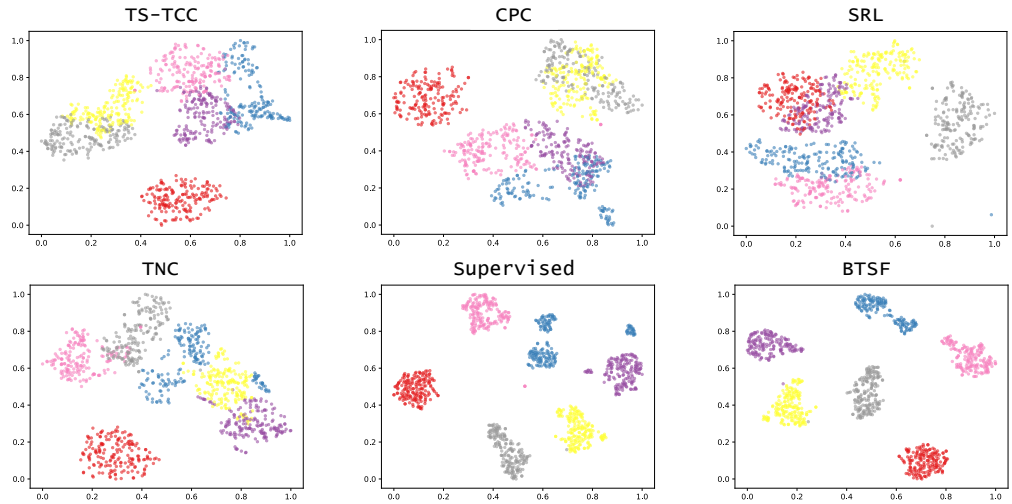

Figure 6: T-SNE visualization of signal representations for HAR dataset.

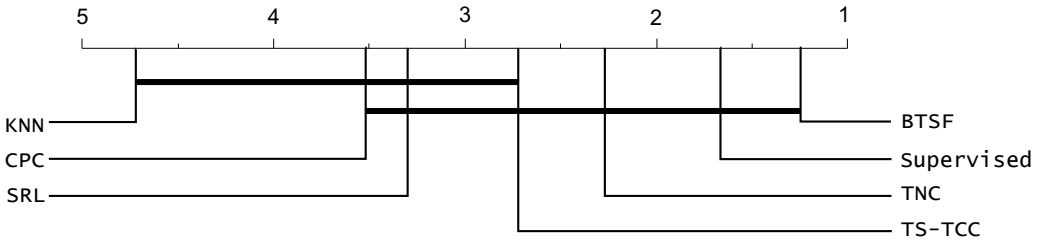

Figure 7: Critical difference diagram showing pairwise statistical difference comparison of BTSF and previous methods on the UCR archive.

A.2 EFFECTIVENESS

To prove the efficiency of our devised bilinear fusion, we provide the deduction of gradient flow from the loss function. Since the overall architecture is a directed acyclic graph, the parameters can be trained by back-propagating the gradients of the contrastive loss. The bilinear form simplifies the gradient computations. Let $\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}}$ be the gradient of $\mathcal{L}$ with respect to $\boldsymbol{f}$, then for Eq.(8) by chain rule

Table 6: Ablation experiments of BTSF.

Accuracy	Temporal	Spectral	Sum/Concat	Bilinear	Iterative Bilinear
Slicing	88.3	86.7	88.7	90.7	91.5
Dropout	89.4	88.4	89.8	92.4	**94.6**
Layer-Wise Dropout	89.8	89.1	90.4	93.1	**95.4**

of gradients we can get:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{F}_t} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{f}} \boldsymbol{W}_t + 2\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}} \boldsymbol{W} \boldsymbol{F}_s, \quad \frac{\partial \mathcal{L}}{\partial \boldsymbol{F}_s} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{f}} \boldsymbol{W}_s + 2\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}} \boldsymbol{W} \boldsymbol{F}_t \tag{10}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_t} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{f}} \boldsymbol{F}_t, \quad \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_s} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{f}} \boldsymbol{F}_s, \quad \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{f}} \boldsymbol{F}_t {\boldsymbol{F}_s}^T \tag{11}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_t} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{f}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{F}_t} \boldsymbol{W}_t + 2\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{F}_t} \boldsymbol{W} \boldsymbol{F}_s, \quad \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_s} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{f}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{F}_s} \boldsymbol{W}_s + 2\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{F}_s} \boldsymbol{W} \boldsymbol{F}_t \tag{12}$$

From the Eq.(10) and Eq.(12), , we conclude that the gradient update of parameters $\boldsymbol{\theta}_t$ in temporal feature $\boldsymbol{F}_t$ is closely related to the spectral feature since $\boldsymbol{F}_s$ is treated as a weighted coefficient straightly multiplying the gradient, and vice versa. Additionally, we can know that interaction matrix $\boldsymbol{W}$ has a strong connection with cross-domain affinities $\boldsymbol{F}_t {\boldsymbol{F}_s}^T$ from the Eq.(11) which leads to a better combination of temporal and spectral features. In hence, it is proved that our BTSF adequately explores and utilizes the underlying spectral and temporal information of time series.

A.3 More ablation studies

To quantify the promotion of each module in BTSF, we make a specific ablation study where all experiments are conducted on HAR dataset and results are in Table 6. We use TNC as a baseline which applies time slicing as augmentation with accuracy of 88.3%. We could find that our instance-level augmentation (dropout) is better than segment-level augmentation (slicing) and layer-wise dropout (adding dropout in internal layers) has a promotion by 1.5% compared with slicing. However, we do not apply layer-wise dropout in aforementioned experiments for fair comparisons otherwise our BTSF will have better performance. Besides, incorporating spectral feature with temporal feature by using summation or concatenation will also improve the results, which illustrates the necessity of cross-domain interaction. The accuracy is obviously promoted by 2%$\sim$3% when involving temporal and spectral information with bilinear fusion, and iterative operation will further improve the performance by enhancing and refining the temporal-spectral interaction. In conclusion, instance-level augmentation (dropout) and iterative bilinear fusion are two main modules of BTSF which largely improve the generalization ability of unsupervised learned representations with accuracy of 94.6%, an improvement of 6.3% to baseline.

Studies of hyperparameters In the proposed BTSF, there are some hyperparameters needed to be carefully set, the dropout rate, temperature number τ and the loops number of iterative bilinear fusion. Table 7 illustrates that when the rate is set to 0.1, BTSF acquires the best performance since setting too high value would lose the original properties of time series and setting too low value would bring about representation collapse. Table 8 demonstrates that when τ is set to 0.05 , BTSF has the best performance. It is reasonable that proper value of τ would promote the optimization of training process and make representations more discriminative with the adjustment. We also run the experiments of loops number of iterative bilinear fusion and the results are depicted in Figure 8. From the results, we conclude that our iterative bilinear fusion is effective and its performance converges after just three loops.

A.4 Datasets descriptions and more experiments

In all experiments, we use Pytorch 1.8.1 (Paszke et al., 2017) and train all the models on a GeForce RTX 2080 Ti GPU with CUDA 10.2. We apply an Adam optimizer (Kingma & Ba, 2017) with

Table 7: Ablation experiments of dropout rate

dropout rate	p=0.01	p=0.05	p=0.1	p=0.15	p=0.2	p=0.3
HAR	90.29	92.78	**94.63**	93.36	91.21	88.07
Sleep-EDF	82.76	85.34	**87.45**	86.01	83.44	80.92
ECG Waveform	93.13	96.56	**98.12**	97.28	95.63	92.05

Table 8: Ablation experiments on temperature number τ.

τ	0.001	0.01	0.05	0.1	1
HAR	90.04	92.91	**94.63**	93.04	91.85
Sleep-EDF	82.69	84.82	**87.45**	85.11	83.28
ECG Waveform	93.06	95.74	**98.12**	96.47	94.88

a learning rate of 3e-4, weight decay of 1e-4 and batch size is set to 256. In this part, we would introduce all the datasets used in our experiments which involve three kinds of downstream tasks, time series classification, forecasting and anomaly detection. The definitions of downstream tasks are detailed in the following:

- **Time Series Classification**: Given the univariate time series $\{x_1, x_2, \ldots, x_T\}$ or multivariate time series $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D\}$ as input, time series classification is to classify the input consisting of real-valued observations to a certain class.
- **Time Series Forecasting**: Given the past univariate observations $\{x_{t-T_1+1}, \ldots, x_t\}$ or multivariate ones $\{\boldsymbol{x}_{t-T_1+1}, \ldots, \boldsymbol{x}_t\}$ as input, time series forecasting aims to predict the future data points $\{x_{t+1}, x_{t+2}, \ldots, x_{t+T_2}\}$ or $\{\boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+2}, \ldots, \boldsymbol{x}_{t+T_2}\}$ based on the input.
- **Time Series Anomaly Detection**: Given the univariate time series $\{x_1, x_2, \ldots, x_T\}$ or multivariate time series $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D\}$ as input, time series anomaly detection is to find out which point ($\hat{x_i}$ or $\hat{\boldsymbol{x}_i}$) or subsequence ($\{\hat{x_1}, \hat{x_2}, \ldots, \hat{x_T}\}$ or $\{\hat{\boldsymbol{x}_1}, \hat{\boldsymbol{x}_2}, \ldots, \hat{\boldsymbol{x}_T}\}$) of the input behaves unusually when compared either to the other values in the time series (global outlier) or to its neighboring points (local outlier).

Data Preprocessing Following Franceschi et al. (2019); Zhou et al. (2021), for univariate time series classification task, we normalize datasets using z-score so that the set of observations for each dataset has zero mean and unit variance. For multivariate time series classification task, each variable is normalized independently using z-score. For forecasting tasks, all reported metrics are calculated based on the normalized time series.

A.4.1 CLASSIFICATION

In the time series classification task, we choose six popular datasets which are widely used in previous works. These six datasets are Human Activity Recognition (HAR) (Anguita et al., 2013), Sleep Stage Classification (Sleep-EDF) (Goldberger et al., 2000), Epilepsy Seizure Prediction (Andrzejak et al., 2001), ECG Waveform (Moody, 1983),UCR (Dau et al., 2019) and UEA (Bagnall et al., 2018). The detailed introduction to these datasets are as follows:

Human Activity Recognition HAR dataset contains 30 individual subjects which provide six activities for each subject. These six activities are walking, walking upstairs, downstairs, standing, sitting, and lying down. The data of HAR is collected by sensors with a sampling rate of 50 HZ and the collected signals record the continuous activity of every subject.

Sleep Stage Classification The dataset is designed for EEG signal classification task where each signal belongs to one of five categories: Wake (W), Non-rapid eye movement (N1, N2, N3) and Rapid Eye Movement (REM). And the Sleep-EDF dataset collects the PSG for the whole night, and we just used a single EEG channel, following previous works (Eldele et al., 2021a).

Epilepsy Seizure Prediction The Epileptic Seizure Prediction dataset contains EEG signals which are collected from 500 subjects. The brain activity for each subject was recorded for 23.6 seconds.

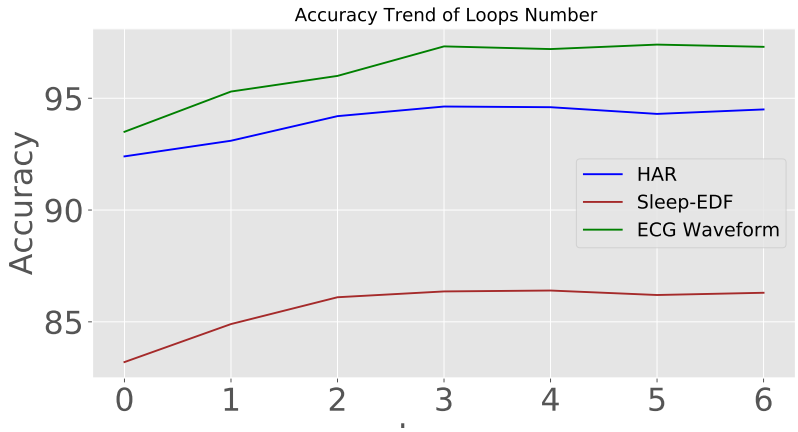

Figure 8: Accuracy trend of changing loops number on HAR, Sleep-EDF and ECG Waveform datasets.

Table 9: More comparisons of classification results about BTSF and previous work, results of TST (Zerveas et al., 2021), Rocket (Dempster et al., 2020) and Supervised (Zerveas et al., 2021) are quoted from TST for fair comparisons.

Methods	TST	Rocket	Supervised	**BTSF**
EthanolConcentration	32.6	45.2	33.7	**49.4**
FaceDetection	68.9	64.7	68.1	**73.0**
Handwriting	35.9	58.8	30.5	**62.3**
Heartbeat	77.6	75.6	77.6	**84.7**
JapaneseVowels	99.7	96.2	99.4	**99.8**
InsectWingBeat	68.7	-	68.4	**78.3**
PEMS-SF	89.6	75.1	91.9	**95.7**
SelfRegulationSCP1	92.2	90.8	92.5	**96.5**
SelfRegulationSCP2	60.4	53.3	58.9	**64.9**
SpokenArabicDigits	99.8	71.2	99.3	**99.8**
UWaveGestureLibrary	91.3	94.4	90.3	**97.1**
Avg Accuracy	74.8	72.5	74.2	**82.0**
Avg Rank	1.7	2.3	1.7	**1.2**

Additionally, the original classes of the dataset are five, and we preprocess the dataset for classification task like Eldele et al. (2021b).

ECG Waveform The ECG Waveform is a real-world clinical dataset, it includes 25 long-term Electrocardiogram (ECG) recordings (10 hours in duration) of human subjects with atrial fibrillation. Besides, it contains two ECG signals with a sampling rate of 250HZ.

UCR and UEA The UCR and UEA are widely used public datasets for time series analysis. The UCR archive consists of univariate datasets while UEA archive contains multivariate datasets, which cover multiple scenes in real world.

Table 9 shows the comparison results between BTSF with recent works following their evaluation protocols. The results show that BTSF significantly outperforms them in a large margin. Table 10 shows the classification results of Epileptic Seizure Prediction datasets. From the illustrated results, we conclude that our BTSF gets the best performance and exceeds other methods by a large margin in univariate and multivariate time series classification tasks.

Table 10: More comparisons of classification results of ESP dataset.

Methods	Epilepsy Seizure Prediction	
	Accuracy	AUPRC
Supervised	96.32±0.38	0.97±0.65
KNN	87.96±1.32	0.89±1.04
SRL	94.65±0.97	0.95±0.86
CPC	96.61±0.43	0.97±0.69
TS-TCC	97.23±0.10	0.98±0.21
TNC	96.15±0.33	0.96±0.45
BTSF	**99.01±0.12**	**0.99±0.06**

A.4.2 FORECASTING

In Section 4, we conduct experiments on four datasets about time series forecasting, including two collected real-world datasets for long sequence time-series forecasting (LSTF) problem and one public benchmark dataset as in Zhou et al. (2021). The detailed introduction to these datasets are as follows:

Electricity Transformer Temperature (ETT) The ETT is a crucial indicator in the electric power long-term deployment. The 2-year data was collected from two separated counties in China, which was first used to investigate the granularity on the LSTF problem with each data point containing the target value "oil temperature" and six power load features. ETTh1 , ETTh2 and ETTm1 represent for 1-hour-level and 15-minute-level respectively.

Weather This dataset contains local climatological data for about 1,600 U.S. places, 4 years from 2010 to 2013, where data points are collected every 1 hour with each data point consisting of the target value "wet bulb" and 11 climate features.

We run the forecasting tasks about prediction length of 48 and 1440 on ETT dataset and visualize the forecasting results of BTSF, TNC and supervised models. From Figure 9 and 10, we could find that our BTSF achieves the best forecasting results under both short-term and long-term settings since it adequately leverages the global context and utilize temporal-spectral relations which are helpful in producing more accurate predictive representations. The complete comparisons of forecasting results in Table 11 further prove the superiority of BTSF.

A.4.3 ANOMALY DETECTION

In Section 4, we conduct extensive experiments about time series anomaly detection on five widely used datasets, which are all public available. The detailed introduction to these datasets are illustrated as follows:

Secure Water Treatment (SWaT) The SWaT dataset is a scaled down version of a real-world industrial water treatment plant producing filtered water (Goh et al., 2016). The collected dataset (Mathur & Tippenhauer, 2016) consists of 11 days of continuous operation: 7 days collected under normal operations and 4 days collected with attack scenarios.

Water Distribution (WADI) This dataset is collected from an extension of the SWaT tesbed. It consists of 16 days of continuous operation: 14 days were collected under normal operation and 2 days with attack scenarios.

Server Machine Dataset (SMD) This dataset is a 5-week-long dataset from a large internet company which was collected and made publicly available (Su et al., 2019). It contains data from 28 server machines with each one monitored by m=33 metrics. SMD is divided into two subsets of equal size: the first half is the training set and the second half is the testing set.

Figure 9: Visualizing forecasting results of length 48 on ETT dataset.

Soil Moisture Active Passive (SMAP) and Mars Science Laboratory (MSL) SMAP and MSL are two real-world public datasets, expert-labeled datasets from NASA (Hundman et al., 2018). They contain respectively the data of 55/27 entities each monitored by m = 25/55 metrics.

The complete comparisons of all metrics (P, R and F1) in anomaly detection are illustrated in Table 12. Our BTSF outperforms other methods imcluding supervised method in a large margin. It demonstrates BTSF is more sensitive to the outliers in time series.

Table 11: Comparisons of multivariate forecasting Results.

Datasets	Length	Supervised		SRL		CPC		TS-TCC		TNC		**BTSF**	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	24	0.577	0.549	0.698	0.661	0.687	0.634	0.653	0.610	0.632	0.596	**0.541**	**0.519**
	48	0.685	0.625	0.758	0.711	0.779	0.768	0.720	0.693	0.705	0.688	**0.613**	**0.524**
	168	0.931	0.752	1.341	1.178	1.282	1.083	1.129	1.044	1.097	0.993	**0.640**	**0.532**
	336	1.128	0.873	1.578	1.276	1.641	1.201	1.492	1.076	1.454	0.919	**0.864**	**0.689**
	720	1.215	0.896	1.892	1.566	1.803	1.761	1.603	1.206	1.604	1.118	**0.993**	**0.712**
ETTh2	24	0.720	0.665	1.034	0.901	0.981	0.869	0.883	0.747	0.830	0.756	**0.359**	**0.432**
	48	1.451	1.001	1.854	1.542	1.732	1.440	1.701	1.378	1.689	1.311	**0.544**	**0.527**
	168	3.389	1.515	5.062	2.167	4.591	3.126	3.956	2.301	3.792	2.029	**1.669**	**0.875**
	336	2.723	1.340	4.921	3.012	4.772	3.581	3.992	2.852	3.516	2.812	**1.954**	**1.093**
	720	3.467	1.473	5.301	3.207	5.191	2.781	4.732	2.345	4.501	2.410	**2.566**	**1.276**
ETTm1	24	0.323	0.369	0.561	0.603	0.540	0.513	0.473	0.490	0.429	0.455	**0.302**	**0.342**
	48	0.494	0.503	0.701	0.697	0.727	0.706	0.671	0.665	0.623	0.602	**0.395**	**0.387**
	96	0.678	0.614	0.901	0.836	0.851	0.793	0.803	0.724	0.749	0.731	**0.438**	**0.399**
	288	1.056	0.786	2.471	1.927	2.066	1.634	1.958	1.429	1.791	1.356	**0.675**	**0.429**
	672	1.192	0.926	2.042	1.803	1.962	1.797	1.838	1.601	1.822	1.692	**0.721**	**0.643**
Weather	24	0.335	0.381	0.688	0.701	0.647	0.652	0.572	0.603	0.484	0.513	**0.324**	**0.369**
	48	0.395	0.459	0.751	0.883	0.720	0.761	0.647	0.691	0.608	0.626	**0.366**	**0.427**
	168	0.608	0.567	1.204	1.032	1.351	1.067	1.117	0.962	1.081	0.970	**0.543**	**0.477**
	336	0.702	0.620	2.164	1.982	2.019	1.832	1.783	1.370	1.654	1.290	**0.568**	**0.487**
	720	0.831	0.731	2.281	1.994	2.109	1.861	1.850	1.566	1.401	1.193	**0.601**	**0.522**

Table 12: Comparisons of multivariate anomaly detection.

Datasets	Metric	Supervised	SRL	CPC	TS-TCC	TNC	**BTSF**
SAaT	P	0.996	0.784	0.791	0.823	0.816	**0.997**
	R	0.842	0.603	0.644	0.712	0.726	**0.873**
	F1	0.901	0.710	0.738	0.775	0.799	**0.944**
WADI	P	0.720	0.459	0.473	0.522	0.561	**0.763**
	R	0.761	0.478	0.492	0.525	0.574	**0.801**
	F1	0.649	0.340	0.382	0.427	0.440	**0.685**
SMD	P	0.984	0.751	0.783	0.802	0.834	**0.993**
	R	0.963	0.790	0.774	0.811	0.806	**0.985**
	F1	0.958	0.768	0.732	0.794	0.817	**0.972**
SMAP	P	0.791	0.562	0.597	0.639	0.641	**0.881**
	R	0.985	0.755	0.781	0.812	0.826	**0.994**
	F1	0.842	0.598	0.620	0.679	0.693	**0.906**
MSL	P	0.937	0.728	0.778	0.825	0.819	**0.968**
	R	0.980	0.702	0.749	0.793	0.815	**0.993**
	F1	0.945	0.788	0.813	0.795	0.833	**0.984**

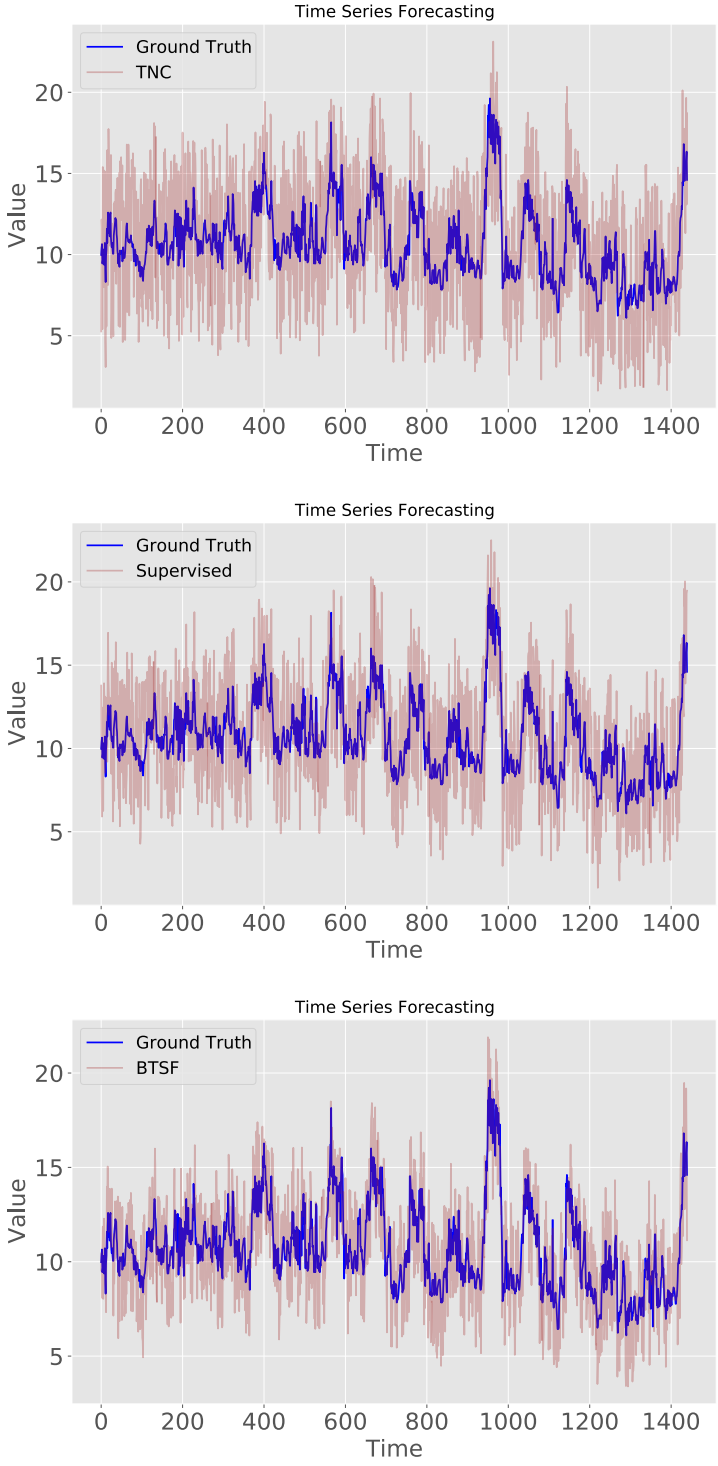

Figure 10: Visualizing long-term forecasting results of length 1440 on ETT dataset.