清華大學
Tsinghua University

清華大學电子工程系
Department of Electronic Engineering, Tsinghua University

面向多源异构数据的时间序列挖掘关键技术研究

答辩人： 张振威

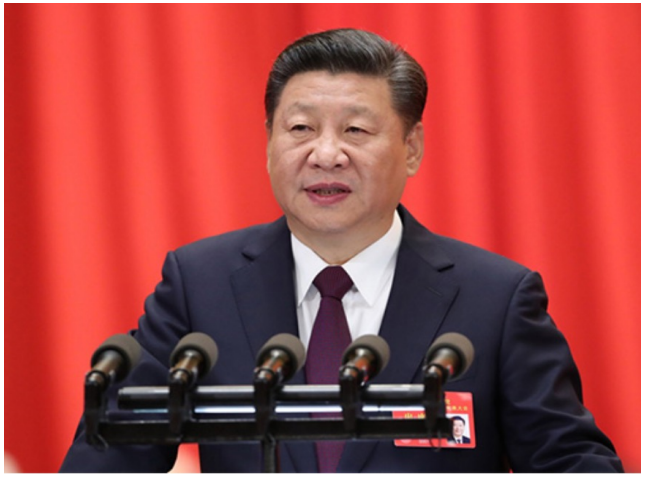
导师： 谷源涛 教授

2023年11月2日

目录

一、研究背景

服务国家重大战略需求



“当今时代，数字技术、数字经济是世界科技革命和产业变革的先机，是新一轮国际竞争重点领域，我们一定要抓住先机、抢占未来发展制高点。”

强化数据“多样性”处理。提升**数值**、**文本**、图形图像、**音频视频**等**多类型数据**的多样化处理能力。促进**多维度异构数据关联**，创新数据融合模式，提升多模态数据的综合处理水平，通过数据的完整性提升认知的全面性。

《“十四五”大数据产业发展规划》
2021年11月 · 工业和信息化部

坚持以习近平……，促进**工业数据**汇聚共享、深化数据融合创新、提升数据治理能力、加强数据安全管理，着力打造资源富集、应用繁荣、产业进步、治理有序的**工业大数据生态体系** ……支持企业建设数据汇聚平台，实现**多源异构数据**的融合和汇聚。

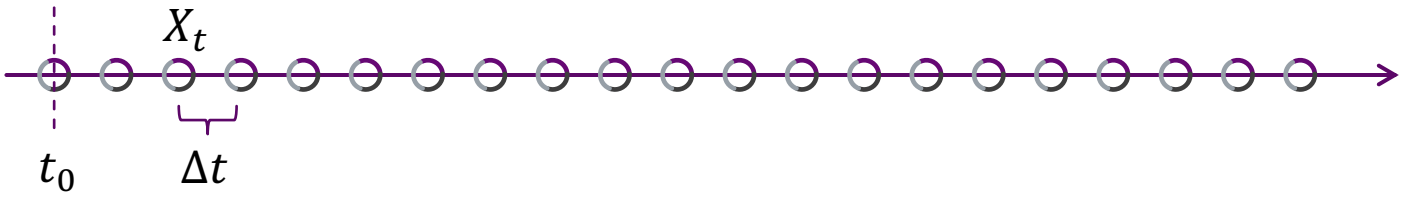
《工业和信息化部关于工业大数据发展的指导意见》
2020年4月 · 工业和信息化部

多源异构时间序列数据在中国数字技术和产业变革中有重要战略价值

什么是时间序列?

- 狭义上：时间序列是用时间排序的一组随机变量，通常的时间间隔为一恒定值

$$\{X_t : t = t_0 + n\Delta t,\ n \in \mathbb{Z}\}$$



- 广义上：时间序列是一组按照时间发生先后顺序进行排列的"观测值"序列。

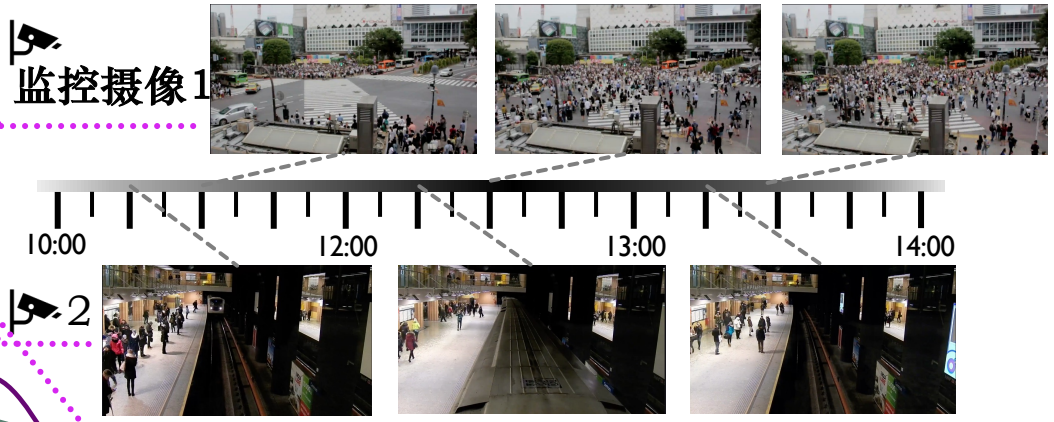
$$\{O_t : t \in T\}$$

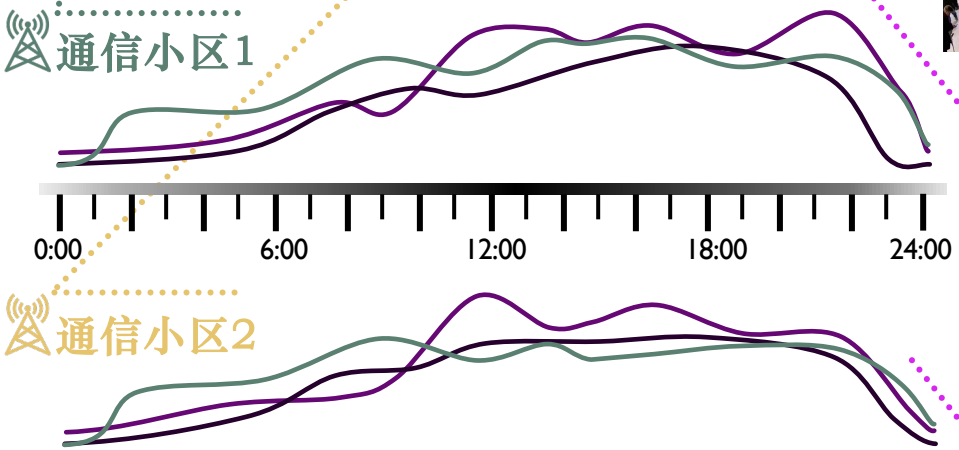
  - 时间间隔不一定相等
  - 每个时间点上的"观测值"不一定要是同质的或符合同一统计分布
  - “观测值”也不必都是数值型数据，可以是**多维向量、文本、图结构**等等

多源异构时间序列数据的挑战

研究挑战

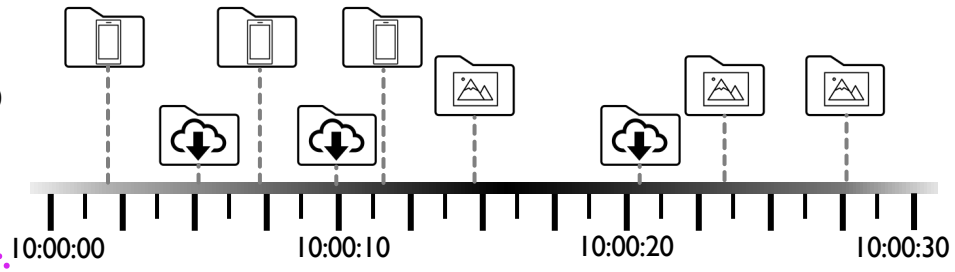
1. 时序模式复杂叠加
2. 多元变量关系不明
3. 数据形态结构各异
4. 数据语义弱，模式跨度大





电信KPIs：流量、用户数、资源占用率

网络流量



地铁闸机、安检人流、天气指数……

2.1 时间序列预测任务概述

2.2 时间序列异常检测任务概述

时间序列挖掘关键任务

2.3 时间序列分析的经典方法

2.4 时间序列分析的深度学习方法

时间序列挖掘常用方法

二、研究现状

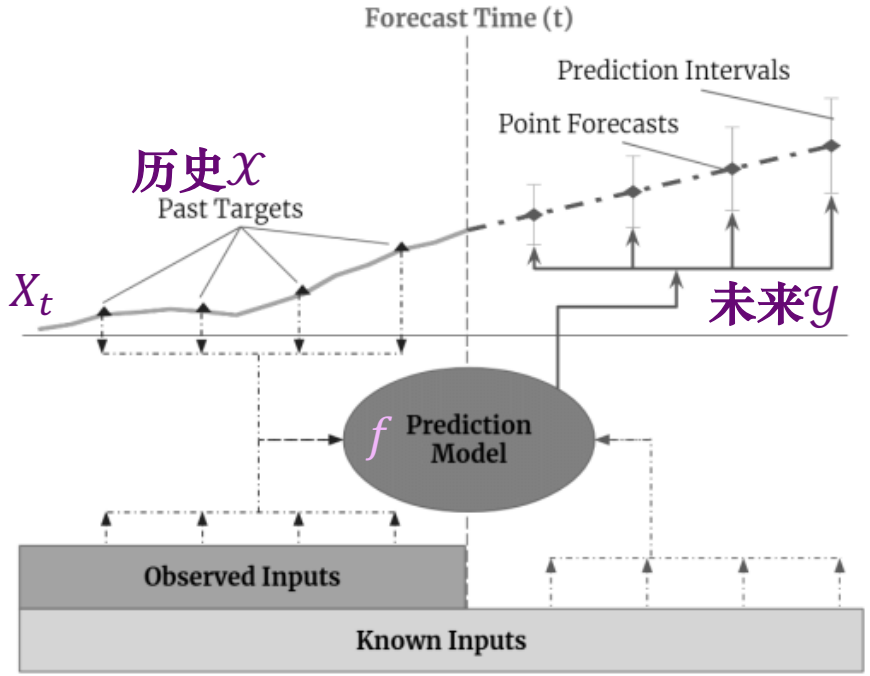
时间序列预测任务

- **多元时间序列预测（Multivariate Time Series Forecasting）目标是利用历史数据以及可选的已知信息来预测未来的一个或多个时间点。**

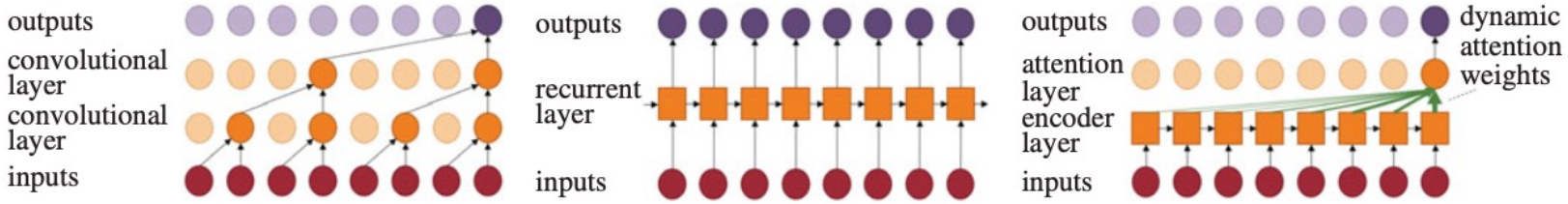
$$\mathcal{X} = [X_t, X_{t+1}, \cdots, X_{t+L-1}]$$

$$\mathcal{Y} = [X_{t+L}, \cdots, X_{t+L+T-1}]$$

$$f(\cdot): \mathcal{X} \rightarrow \mathcal{Y}$$



图：多元时间序列预测的一般框架



CNN model *[CVPR'17]*　　RNN model *[IJF'17]*　　attention-based model *[AAAI'21]*

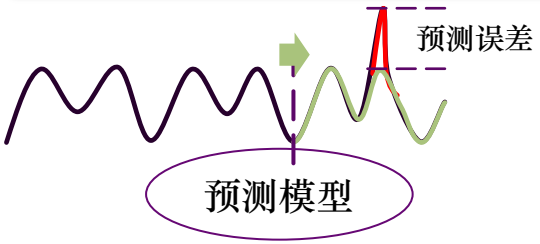
时间序列异常检测任务

- 时间序列异常：异常点对于正常样本分布的偏离现象
- 方法：对于异常分数的建模

$$\mathcal{X} = [X_t, X_{t+1}, \dots, X_{t+L-1}]$$
$$\mathcal{Y} = [Y_t, Y_{t+1}, \dots, Y_{t+L-1}], Y_i \in \{0,1\}$$

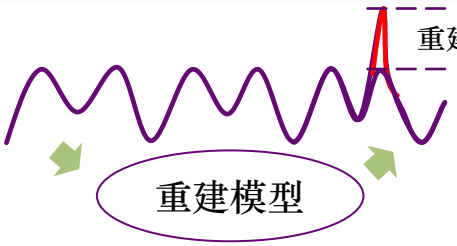
对于时序序列计算出异常分数S_i，以衡量其相对于正常样本的偏离程度，如果某处异常分数高于阈值($S_i \geq \epsilon$)，则认为样本在该点处出现异常，即$Y_i = 1$

① 基于预测的方法



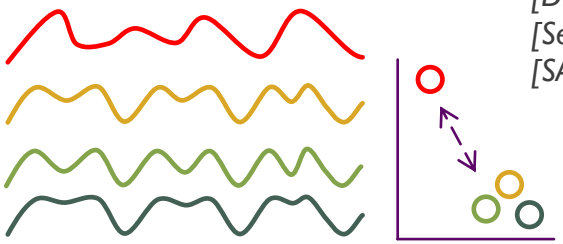
[ARIMA@1990]
[LSTM@KDD'18]
[HTM@Neurocomputing'18]

② 基于重建的方法



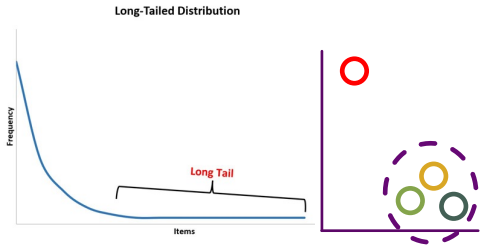
[DAGMM@ICLR'18]
[LSTM-VAE@IEEE RAL'18]
[OmniAnomaly@KDD'19]

③ 基于编码/距离的方法



[DBStream@TKDE'16]
[Series2Graph@PVLDB'20]
[SAND@PVLDB'21]

④ 基于分布的方法

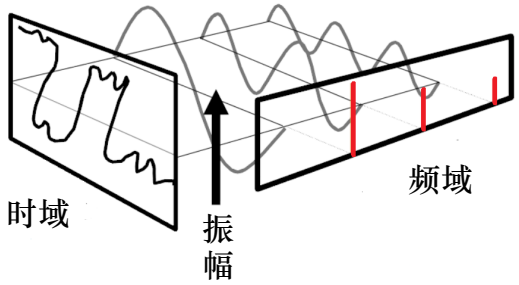


[COPOD@ICDM'20]
[DSPOT@KDD'17]
[DWT-MLEAD@ITISE'17]

时间序列分析的传统方法

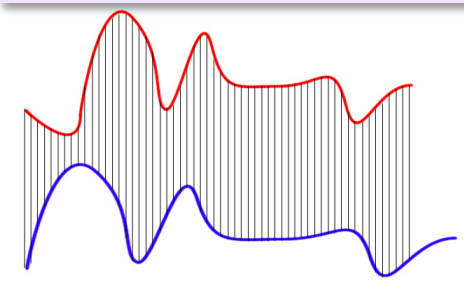
时频域分析方法

[STFT @IEEE TASSP'1984]
[Hilbert-Huang Transform]

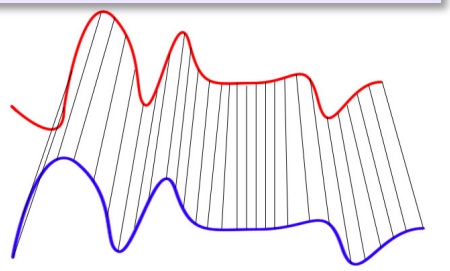


基于距离的方法

[DTW @KDD'1994]
[ERP @VLDB'2004]



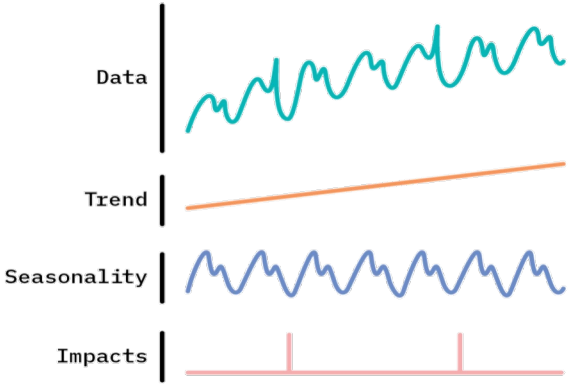
Euclidean Matching



Dynamic Time Warping Matching

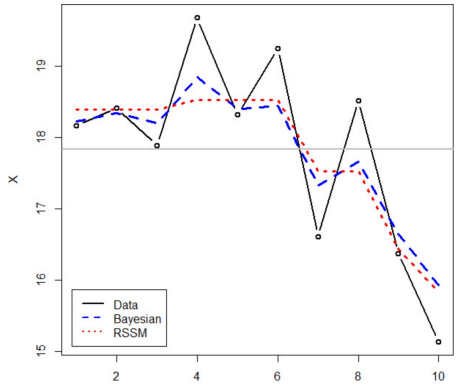
基于分解的方法

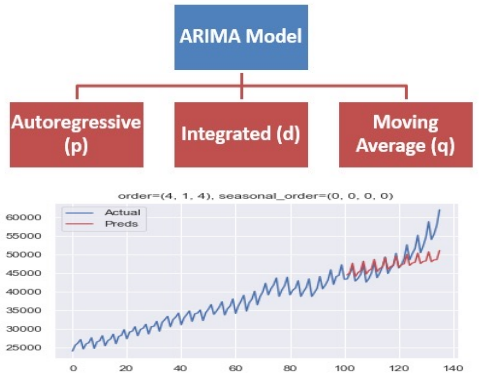
[X-11-ARIMA @1931]
[STL @JOS'1990]



基于回归的方法

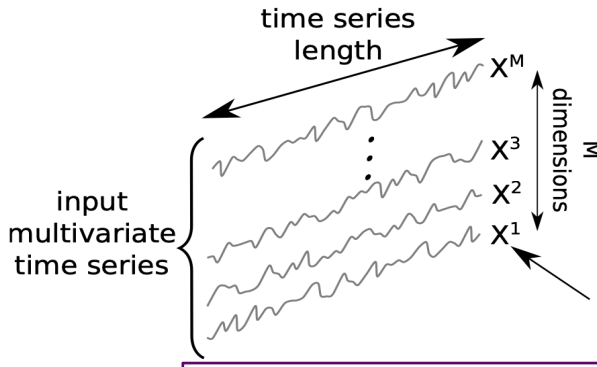
[ARIMA@Oper. Res. Q.'1975]
[LASSO @JRSSA'1996]





经典统计方法普遍依赖于线性分析模型，拟合能力有限；
且参数选择复杂，在应用过程中需要大量的人工判断

时间序列分析的深度学习方法



任务相关的神经网络 → 任务输出

- 建模**变量间**的关系
  - 降维表示
    - AE *[IJCAI'19]*
    - VAE *[TNNLS'20]*
  - 矩阵/张量
    - TF-Net *[KDD'20]*
  - 图结构
    - GNN *[KDD'20]*
- ？
  - **挑战：多元变量关系不明**
- 建模**变量内**时序模式
  - RNN *[IJF'17]*
  - CNN *[CVPR'17]*
  - Attention *[AAAI'21] [ICML'22]*
  - Hybrid *[SIGIR'18]*

如何设计网络，使其既能关注变量之间的信息
也能关注变量内部的时序模式?

3.1 基于数据驱动分解的时间序列预测与异常检测框架

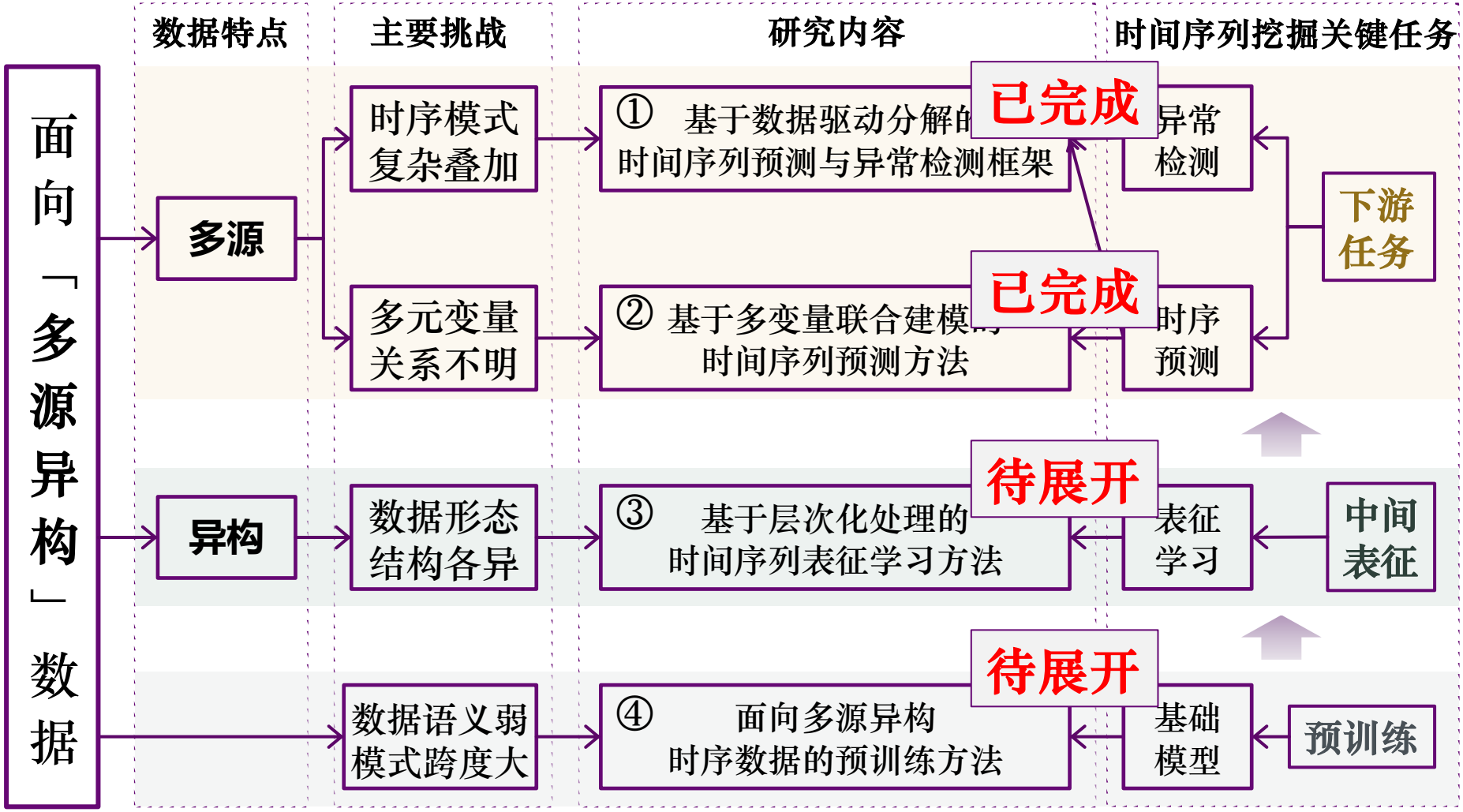
3.2 基于多变量联合建模的时间序列预测方法

3.3 基于层次化处理的时间序列表征学习方法

3.4 面向多源异构时序数据的基础模型研究

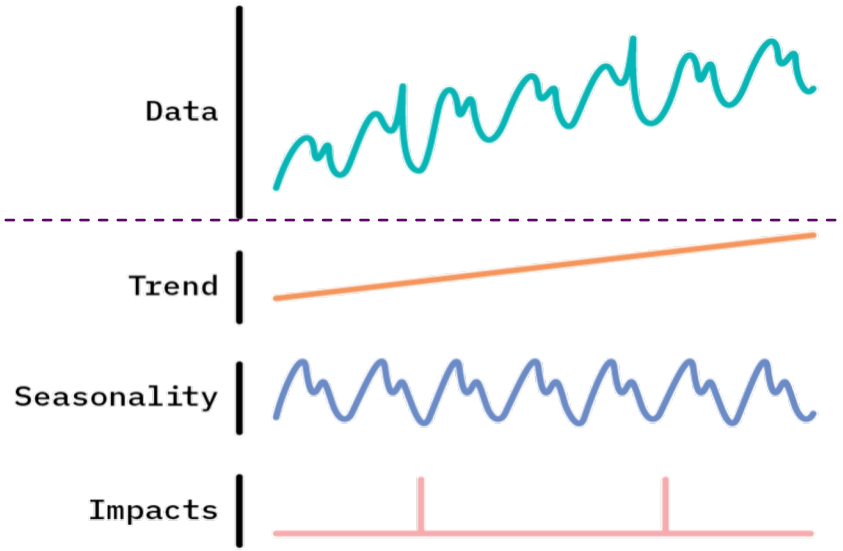
三、研究内容

研究课题设置与逻辑关系



研究内容一、基于数据驱动分解的时间序列预测与异常检测框架

- 研究问题：如何准确地识别与处理多源时间序列中的叠加模式?
- 主要挑战：时序模式复杂叠加 $x_t = \tau_t + s_t + r_t$
  - 趋势复杂性：如何识别并拟合非线性、多尺度的趋势
    - $\tau_t = f(t) + g(t^2) + \cdots$
  - 多周期性：时间序列可能包含多个不同的周期性模式
    - $s_t = s_{daily}(t) + s_{weekly}(t) + \cdots$
  - 噪声与异常值：如何有效的处理噪声和多种异常
    - $r_t = \epsilon_t + anomalies$



- 解决思路：时间序列分解（Time-series Decomposition）

小波变换（Wavelet）	经验模态分解（EMD）	季节趋势项分解（STL）
[Neurocomputing'2002] *[KDD'2018]*	*[Proc. R. Soc. A: Math.'1998]* *[Neurocomputiong'2019]*	*[JOS'1990]* *[AAAI'2019]*

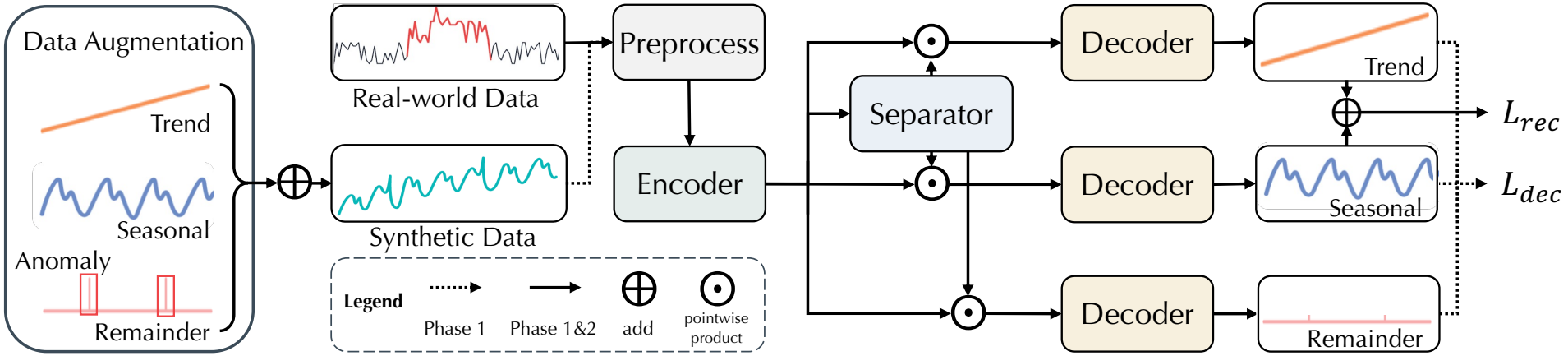
现有方法：时间序列分解

	小波变换（Wavelet）	经验模态分解（EMD）	季节趋势项分解（STL）
思路	使用基于不同尺度和位置的小波函数对时间序列进行多尺度分析	EMD 是一种自适应的方式，它将非平稳的数据分解成一组称为内部模式函数的基函数	使用局部回归（Loess）对时间序列进行季节性和趋势成分的分解
方法	1. 选择一个合适的母小波 2. 使用连续或离散的小波变换来分解时间序列	1. 通过局部极值找到所有最大和最小点 2. 使用样条插值计算这些点的包络 3. 从原数据中减去包络的平均值 4. 重复以上步骤直至满足 IMF 的条件	1. 使用局部回归平滑器（Loess）分离出趋势成分 τ_t 2. 从原始数据中减去趋势成分，获取季节性-剩余成分 $s_t + r_t$ 3. 再次使用局部回归平滑器对季节性-剩余成分进行季节性分解
优点	在不同的尺度上分析数据，对信号的局部特性有很好的捕捉能力	自适应，适用于非线性和非平稳的数据	灵活性高，能够适应不规则和非线性的数据
缺点	**需要事先选择合适的母小波，这一选择需要依赖具体问题**	**可能产生模式混叠，计算成本高。难以适用于大规模数据**	**计算成本较高，可能需要大量的调参。难以适用于大规模数据**

现有的时序分解方法大多需要人工干涉并手动选择参数
大规模的复杂模式时间序列处理需要数据驱动的分解！

创新思路：数据驱动的时序分解网络

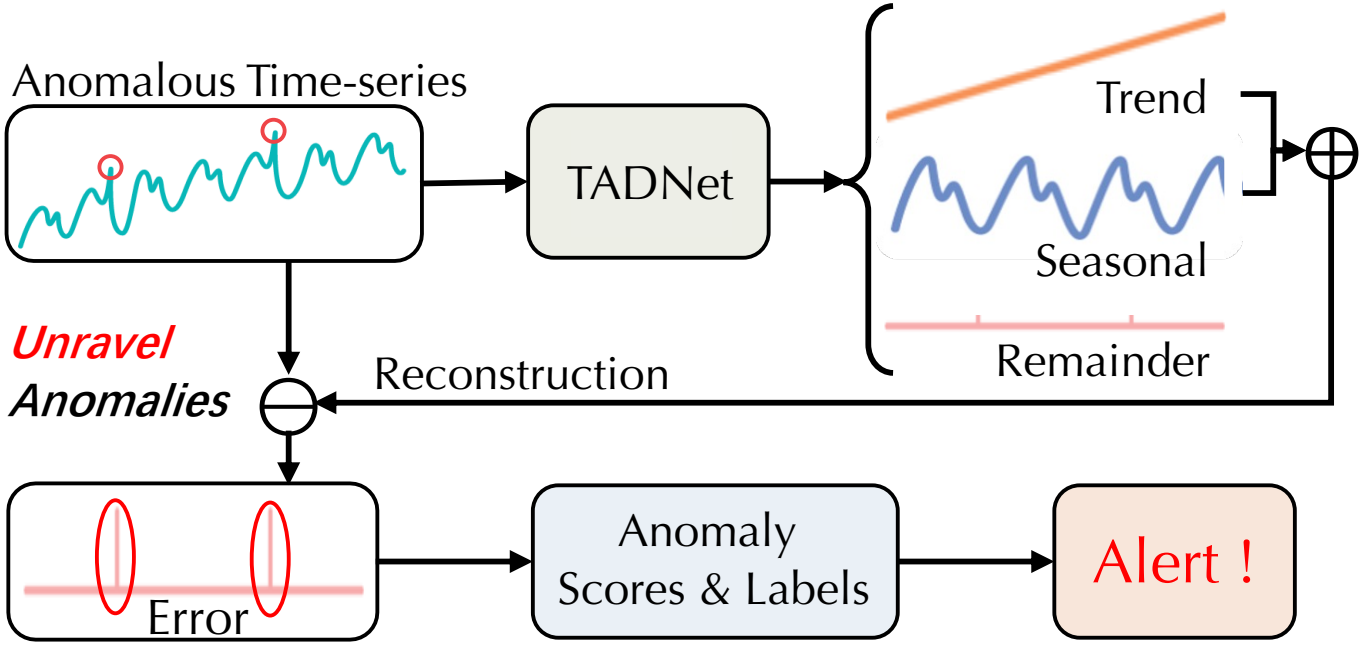
- 提出**TADNet**，采用数据驱动的方式，提供了灵活的端到端时间序列分解
  - 为了解决实际数据缺少分解后的监督信号问题，**使用数据增强的方法生成分解任务的合成数据集**
  - 为了适应真实场景，选用了**预训练-微调的范式**，在合成数据集上进行预训练，并在真实数据上对网络进行微调
  - 选择TasNet语音分离网络作为模型分解的骨干部分，直接对时域信号进行端到端的时间域分解学习



TADNet网络整体流程示意图

创新方法：时序分解帮助异常检测

- 时间序列分解可以帮助**揭示异常**：分解后更精细地理解数据的内在结构和动态变化
  1. 趋势分析：长期观察数据的增长或减少，突增或突减现象
  2. 季节性分析：在固定周期或特定时间段内，数据有与常规不符的波动
  3. 残差分析：数据中不规则或随机的波动，高或低于正常水平的残差



时序分解帮助异常检测示意图

创新方法：对异常进行更清晰分类

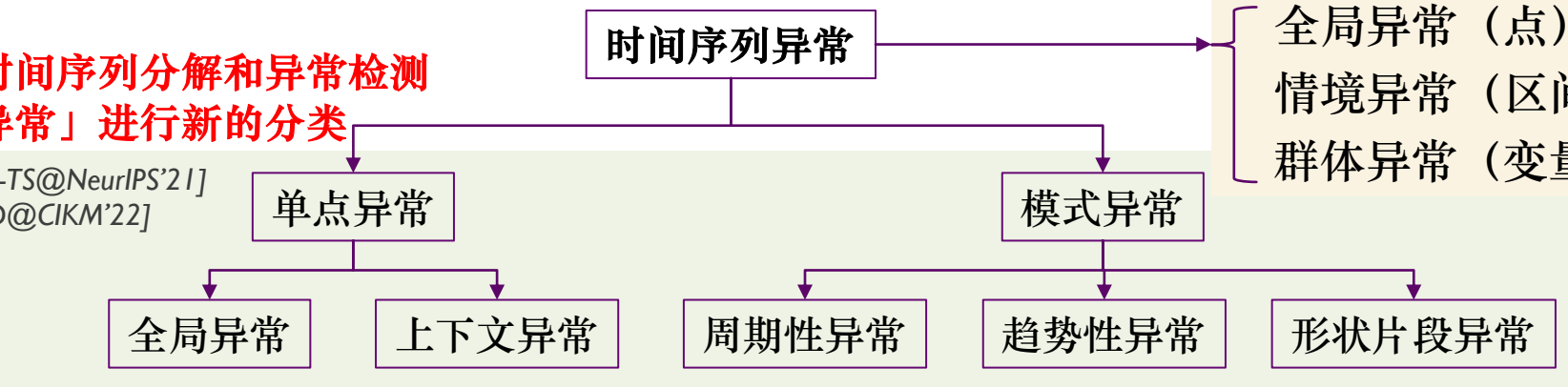
[CompEval@PVLDB'22]

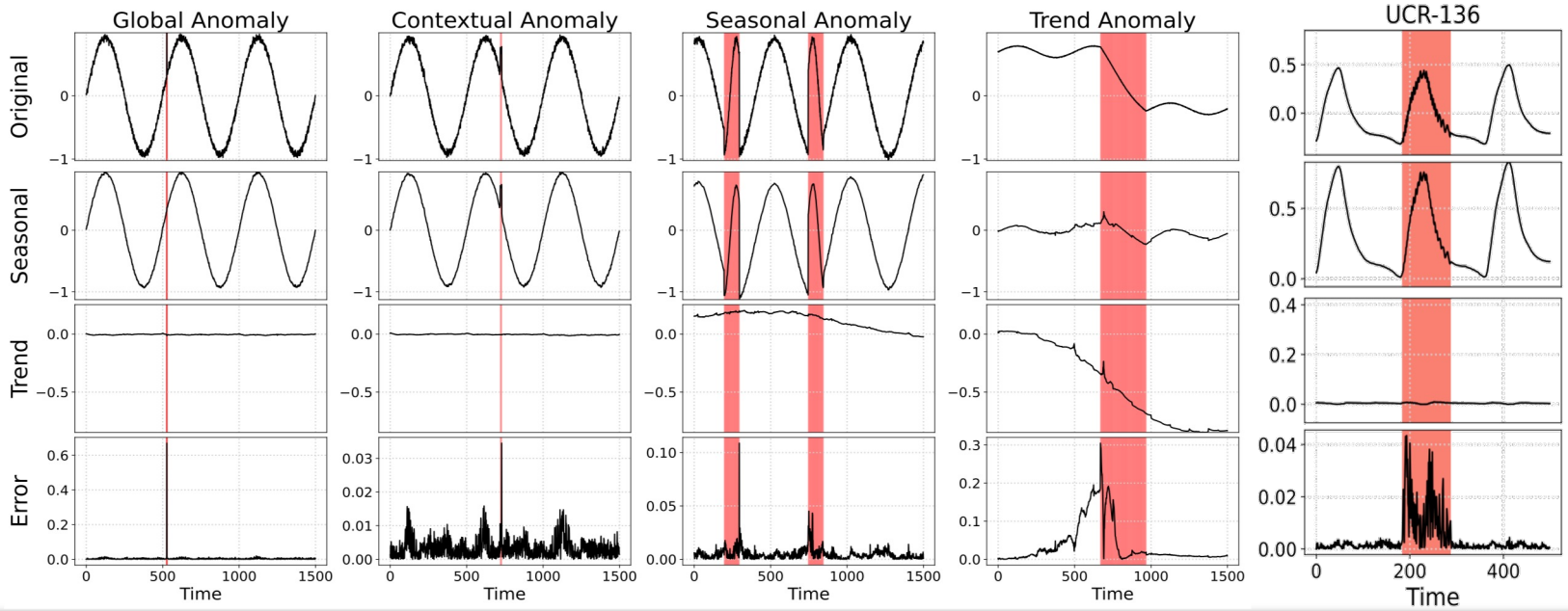
常用，但语义混淆

- 全局异常（点）
- 情境异常（区间）
- 群体异常（变量）

结合时间序列分解和异常检测
对「异常」进行新的分类

[NIPS-TS@NeurIPS'21]
[TFAD@CIKM'22]



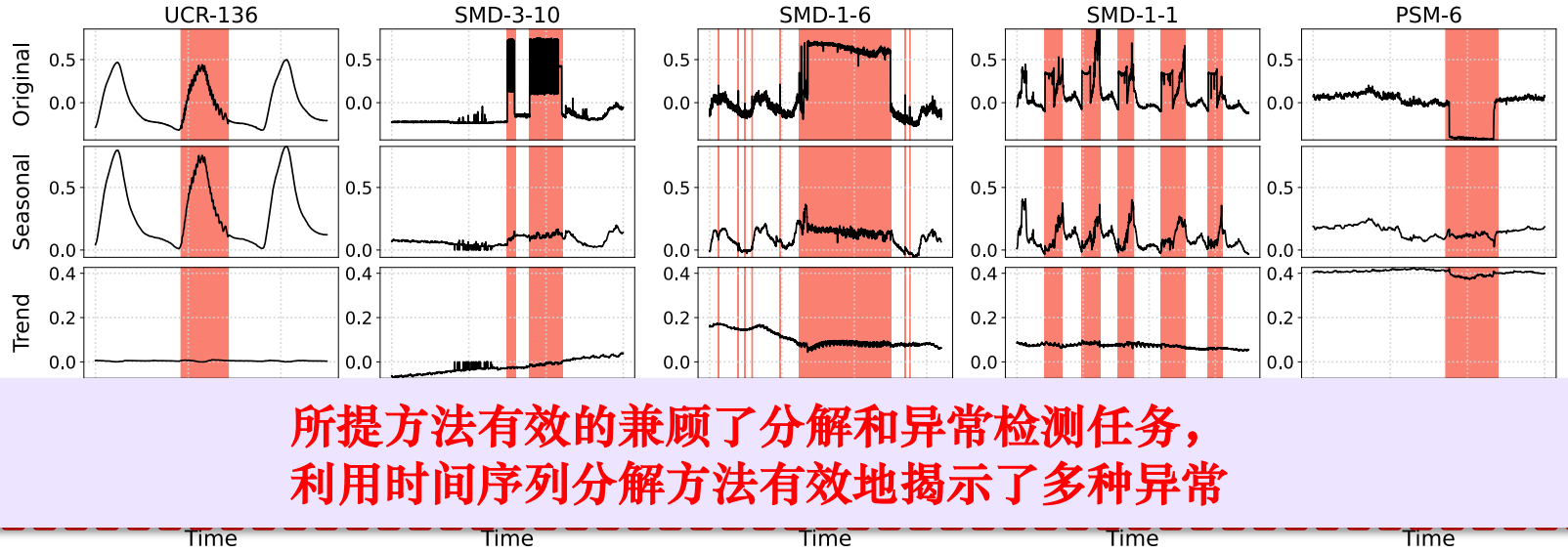


时间序列异常检测实验结果与分析

F1分数最高相对提升10%

Table 1. Quantitative results for TADNet across five real-world datasets use metrics *P*, *R*, and *F1* for precision, recall, and F1-score (%). Higher values indicate better performance. Best and second-best results are in bold and underlined, respectively. Dataset are followed by brackets, where *u* indicates univariate and *m* multivariate.

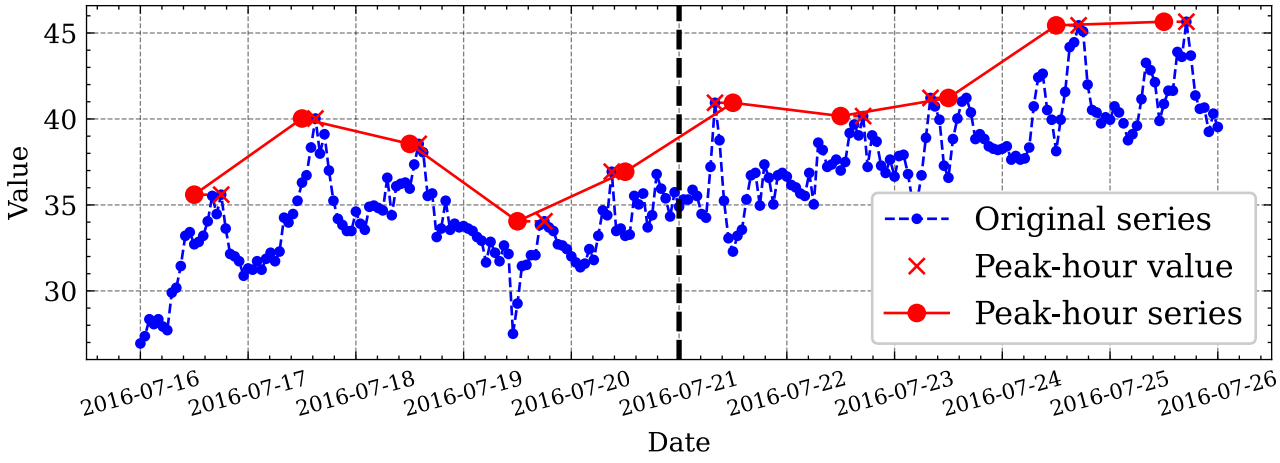
Dataset		UCR (*u*)			SMD (*m*)			SWaT (*m*)			PSM (*m*)			WADI (*m*)		
Metric		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
[NeurComput.'01]	OCSVM	41.14	94.00	57.23	44.34	76.72	56.19	45.39	49.22	47.23	62.75	80.89	70.67	61.89	62.31	62.10
[IJCAI'19]	BeatGAN	45.20	88.42	59.82	72.90	84.09	78.10	64.01	87.46	73.92	90.30	93.84	92.04	65.13	38.32	48.25
[SIGKDD'19]	OmniAnomaly	64.21	86.93	73.86	83.34	94.49	88.57	86.33	76.94	81.36	91.61	71.36	80.23	31.58	65.41	42.60
[SIGKDD'21]	InterFusion	60.74	95.20	74.16	87.02	85.43	86.22	80.59	85.58	83.01	83.61	83.45	83.52	80.26	30.38	44.08
[ICLR'21]	AnomalyTran	72.80	99.60	84.12	89.40	95.45	92.33	91.55	96.73	**94.07**	96.91	98.90	97.89	80.30	79.23	79.76
[PVLDB'22]	TranAD	94.07	100.00	96.94	88.03	89.42	88.72	97.60	69.97	81.51	96.44	87.37	91.68	35.29	82.96	49.51
[IEEEBigData'22]	DecompTran	71.58	96.83	82.31	89.32	93.94	91.57	95.17	80.30	87.10	97.65	87.21	92.14	79.40	81.01	80.20
TADNet(Ours)		97.51	100.00	**98.74**	94.81	91.93	**93.35**	92.15	88.35	90.21	98.12	99.21	**98.66**	94.03	82.96	**88.15**

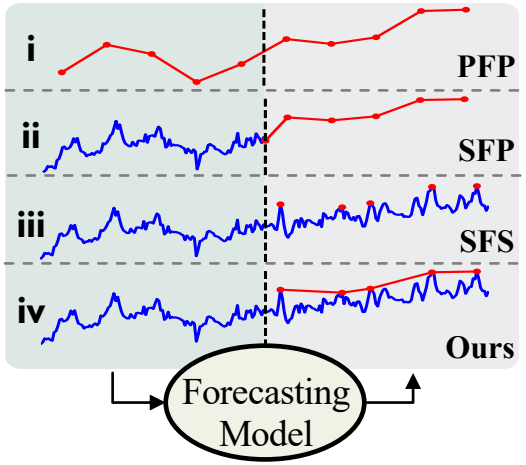


所提方法有效的兼顾了分解和异常检测任务，
利用时间序列分解方法有效地揭示了多种异常

研究内容一、基于数据驱动分解的时间序列预测与异常检测框架

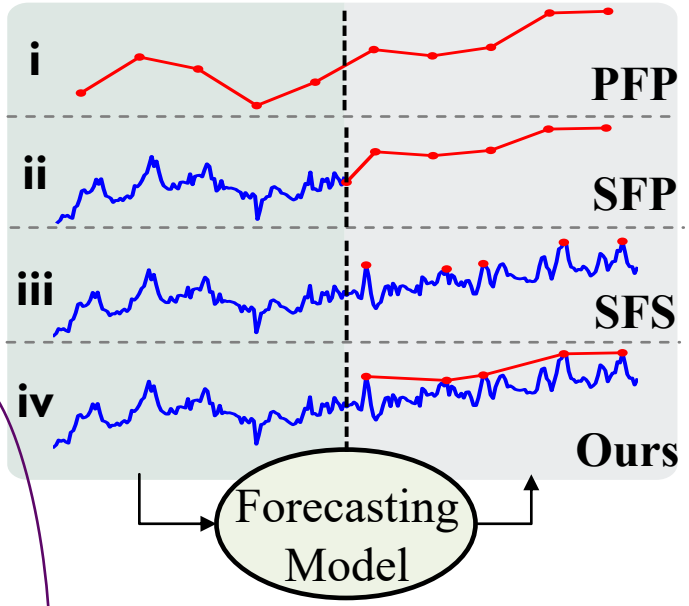
- 周期信息挖掘在多个行业中起到至关重要的作用。在某些场景下，最关心的并不总是完整的周期信息，而是**周期性变化中的最大值**。
  - **电信领域**: 工程师会根据周期性的最大流量来校准基站容量，以优化通信质量，或是提前关断部分基站帮助节能
  - **能源领域**: 每天的峰值用电量会影响原材料的供应和电力生产，某些公用事业公司也会根据峰值需求来进行计费
  - **交通领域**: 理解早晚高峰的交通模式对有效的城市规划至关重要。
- **率先提出「峰值时间序列预测任务」(Peak-hour series forecasting)。**

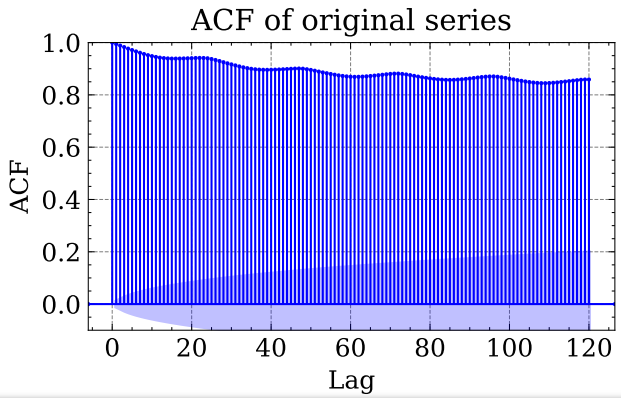


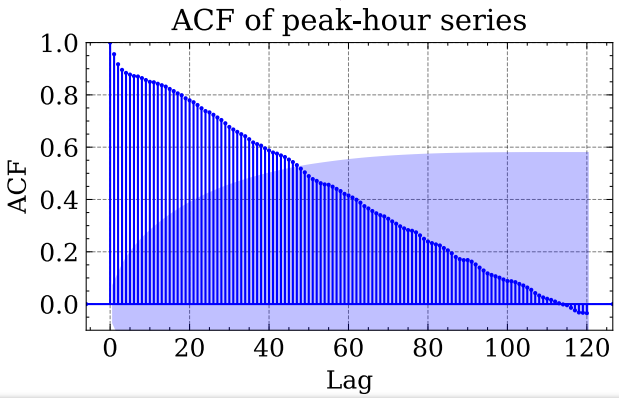


现有方法：峰值时间序列预测任务的三种范式

范式	方法描述	优点	缺点
PFP	仅依赖历史峰值序列对未来的峰值序列预测	简单，直接针对峰值数据	由于峰值序列的**自相关函数（ACF）**值很低，难以预测
SFP	使用完整序列进行峰值预测	简单，能够利用更多上下文信息	面临与PFP相似的挑战，忽视了峰值和原始系列之间的特殊关系
SFS	用完整历史序列预测完整未来，并手动提取每日最大值	结合了全序列和高峰时段信息	损失函数最小化完整序列的平均误差，损害了预测极端值的能力







创新方法：峰值时间序列预测框架

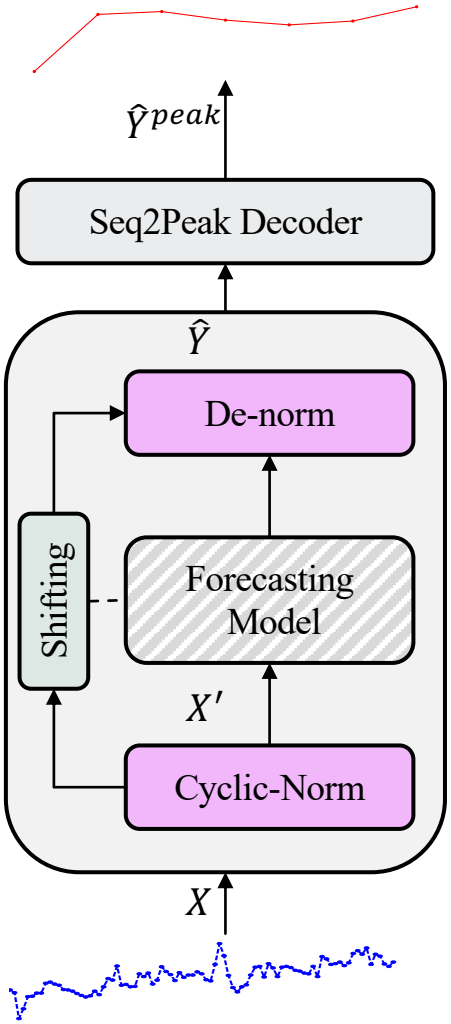
1. Cyclic Normalization（周期规范化）
   - 对原始输入时间序列进行周期性的标准化
   - 非平稳偏移(Shifting, $\mathcal{T}$)模块以模拟数据分布的时间变化
   - 生成更适合PHSF任务的特征

$$\mathrm{X}^{(i)} = \{x_i, x_{i+T}, x_{i+2T}, \cdots\}$$
$$\{X^{(i)}, \mu_i, \sigma_i\} = \mathrm{Norm}(X^{(i)})$$
$$(M', \Sigma') = \mathcal{T}(M, \Sigma)$$

2.Seq2Peak解码器
   - 最大池化层（核大小=周期）提取出峰值序列
   - 采用混合损失函数（hybrid loss function）
   - 同时优化原始时间序列和对应的峰值序列

$$l_{hy} = \alpha l_{seq} + (1-\alpha) l_{peak}$$

原始时间序列的损失　　峰值序列的损失

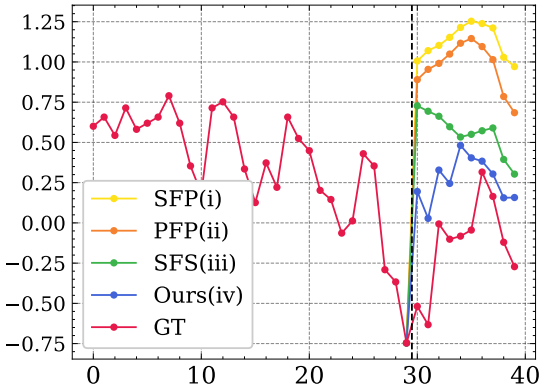


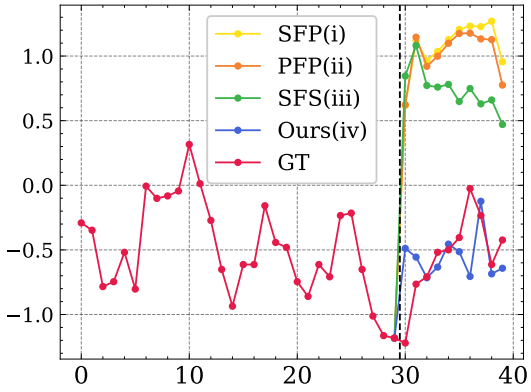
峰值时序预测实验结果与分析

公开数据

Table 1: Performance promotion by applying the proposed framework to four TSF models.

Dataset		[NeurIPS'17] Transformer		+Seq2Peak		[AAAI'21] Informer		+Seq2Peak		[NeurIPS'21] Autoformer		+Seq2Peak		[AAAI'23] DLinear		+Seq2Peak	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	5	2.585	1.403	**0.299**	**0.390**	2.120	1.232	**0.341**	**0.410**	0.486	0.519	**0.334**	**0.413**	0.306	0.378	**0.240**	**0.338**
	10	2.547	1.376	**0.330**	**0.406**	2.502	1.265	**0.367**	**0.420**	0.588	0.549	**0.365**	**0.426**	0.353	0.414	**0.263**	**0.358**
	15	2.668	1.342	**0.353**	**0.418**	2.635	1.358	**0.413**	**0.451**	0.617	0.561	**0.338**	**0.413**	0.393	0.442	**0.279**	**0.368**
	30	2.684	1.446	**0.439**	**0.466**	2.779	1.418	**0.503**	**0.507**	0.747	0.633	**0.398**	**0.440**	0.513	0.517	**0.335**	**0.399**
	Avg	2.621	1.392	**0.355**	**0.420**	2.509	1.318	**0.406**	**0.447**	0.610	0.566	**0.359**	**0.423**	0.391	0.438	**0.279**	**0.366**
ETTh2	5	1.847	1.094	**0.375**	**0.448**	2.223	1.278	**0.592**	**0.540**	0.448	0.499	**0.400**	**0.458**	**0.294**	**0.382**	0.301	0.384
	10	1.932	1.111	**0.391**	**0.444**	1.849	1.158	**0.540**	**0.522**	0.479	0.513	**0.429**	**0.471**	0.376	0.428	**0.360**	**0.423**
	15	1.873	1.103	**0.398**	**0.452**	1.886	1.190	**0.557**	**0.533**	0.517	0.532	**0.456**	**0.487**	0.426	0.459	**0.373**	**0.433**
	30	1.946	1.186	**0.437**	**0.474**	2.138	1.290	**0.646**	**0.566**	0.623	0.582	**0.591**	**0.547**	0.664	0.582	**0.435**	**0.476**
	Avg	1.900	1.124	**0.400**	**0.455**	2.024	1.229	**0.584**	**0.540**	0.517	0.532	**0.469**	**0.491**	0.440	0.463	**0.367**	**0.429**
Electricity	5	0.457	0.469	**0.273**	**0.345**	0.893	0.681	**0.307**	**0.366**	0.330	0.395	**0.299**	**0.361**	0.252	0.326	**0.229**	**0.305**
	10	0.551	0.513	**0.293**	**0.358**	1.030	0.739	**0.327**	**0.380**	0.356	0.410	**0.343**	**0.388**	0.286	0.351	**0.263**	**0.330**
	15	0.524	0.510	**0.326**	**0.380**	1.450	0.917	**0.335**	**0.383**	0.403	0.438	**0.354**	**0.399**	0.312	0.371	**0.289**	**0.350**
	30	0.557	0.517	**0.351**	**0.398**	1.530	0.959	**0.379**	**0.409**	0.459	0.466	**0.399**	**0.429**	0.373	0.415	**0.352**	**0.394**
	Avg	0.522	0.502	**0.311**	**0.370**	1.226	0.824	**0.337**	**0.385**	0.387	0.427	**0.349**	**0.394**	0.306	0.366	**0.283**	**0.345**
Traffic	5	4.030	1.007	**2.101**	**0.814**	7.148	1.729	**2.388**	**0.951**	3.801	1.050	**2.253**	**0.877**	2.303	0.905	**2.050**	**0.836**
	10	4.045	0.991	**2.024**	**0.796**	7.458	1.796	**2.513**	**0.986**	3.889	1.076	**2.254**	**0.876**	2.377	0.918	**2.110**	**0.844**
	15	4.068	0.993	**2.072**	**0.808**	8.138	1.933	**2.684**	**1.034**	3.932	1.068	**2.458**	**0.915**	2.447	0.931	**2.145**	**0.848**
	30	4.270	1.037	**2.206**	**0.834**	8.881	2.071	**3.139**	**1.172**	3.908	1.085	**2.513**	**0.920**	2.641	0.969	**2.256**	**0.868**
	Avg	4.103	1.007	**2.101**	**0.813**	7.906	1.882	**2.681**	**1.036**	3.883	1.070	**2.370**	**0.897**	2.442	0.931	**2.140**	**0.849**





平均相对提升37%

峰值时序预测实验结果与分析

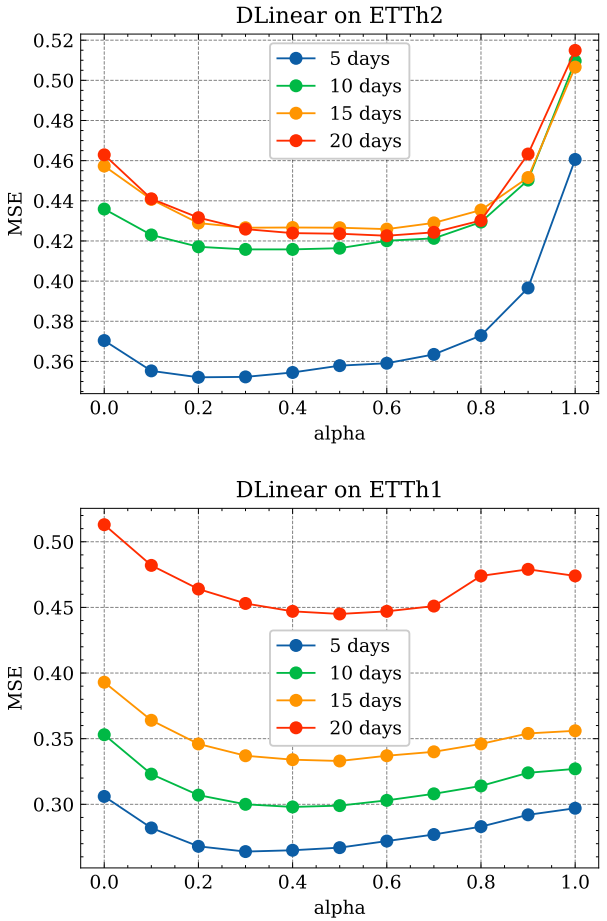
Table 2: Model architecture ablations.

Horizons	5		10		15	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE
Transformer	2.285	1.403	2.547	1.376	2.668	1.342
+ Decoder	2.138	1.156	1.988	1.194	0.935	0.728
+ CyclicNorm	0.423	0.459	0.421	0.467	0.439	0.480
+ Seq2Peak	**0.299**	**0.390**	**0.330**	**0.406**	**0.353**	**0.418**
DLinear	0.306	0.378	0.353	0.414	0.393	0.442
+ Decoder	0.267	0.370	0.299	0.393	0.333	0.416
+ CyclicNorm	0.265	0.354	0.294	0.379	0.313	0.392
+ Seq2Peak	**0.240**	**0.338**	**0.263**	**0.358**	**0.279**	**0.368**

The term "**+ Seq2Peak**" indicates the addition of the complete framework, which includes both the CyclicNorm and Decoder components.

消融实验证明我们模型设计的有效性

右图展示不同α的预测精度，说明混合损失的设计是必要的



研究内容一：小结

- ✓ 提出了基于数据驱动分解的异常检测框架，显式地给出分解结果，并帮助揭示各种异常，提升异常检测精度
- ✓ 率先提出「峰值时间序列预测任务」，并提出一种周期性挖掘框架解决此任务

论文成果：

- **Zhenwei Zhang**, Xin wang, Jingyuan Xie, Heling Zhang and Yuantao Gu. 2023. "Unlocking the Potential of Deep Learning in Peak-Hour Series Forecasting." In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23). Association for Computing Machinery, New York, NY, USA, 4415–4419. https://doi.org/10.1145/3583780.3615159（CCF-B推荐会议，接受率27%）

专利成果：

- 徐灏,**张振威**,闫思成,汪昕,谷源涛.人群聚集预测方法、装置及系统. 华为技术有限公司,清华大学.（发明专利，清华第一作者，已受理，申请号：202311291079.3）

在投论文：

- **Zhenwei Zhang**, Ruiqi Wang, Yuantao Gu. "Unravel Anomalies: An End-to-end Seasonal-Trend Decomposition Approach for Time Series Anomaly Detection." ICASSP, 2024.（电子系顶级会议，在投）

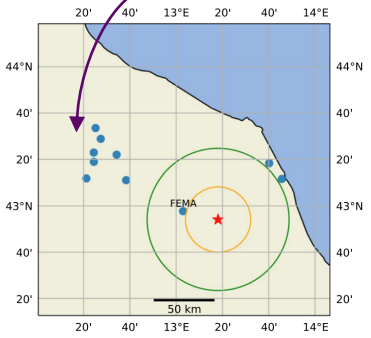
研究内容二、基于多变量联合建模的时间序列预测方法

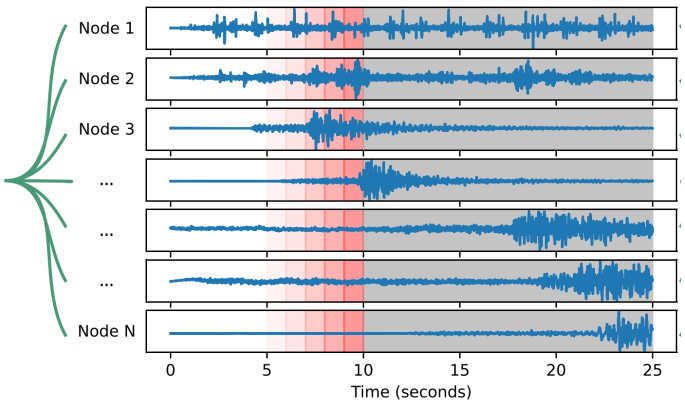
- 研究问题：如何有效地建模与利用多源时间序列中变量间与变量内的关系？
- 主要挑战：多元变量关系不明

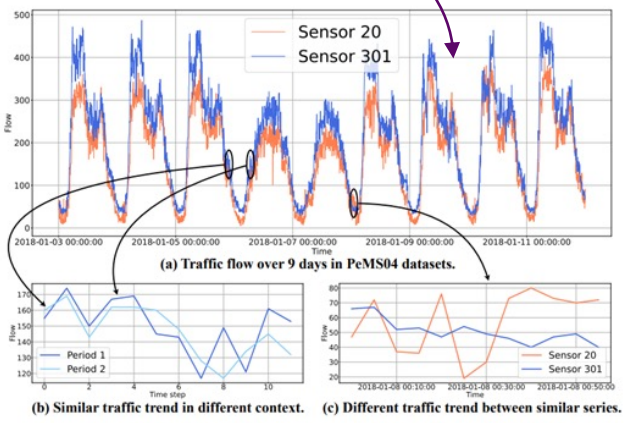
建模变量间的关系 inter-series

两者之间如何关联

建模变量内时序模式 intra-series







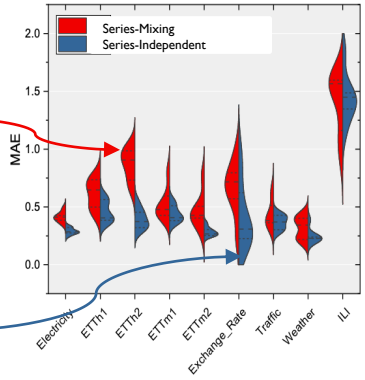
地震波随着与震中距离的增加而呈现时延特性

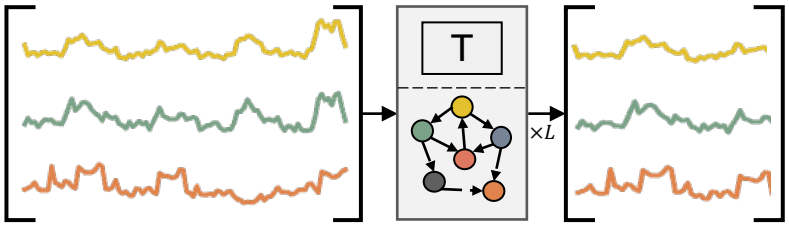
不同时刻具有相似的模式

相同时刻具有不同的模式

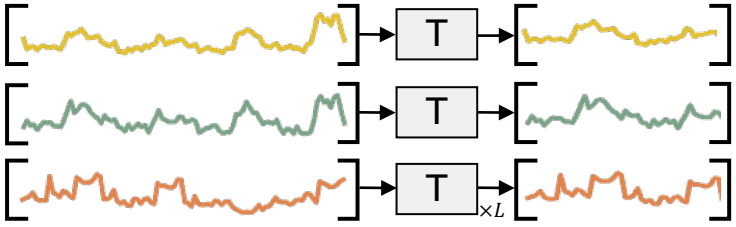
现有多元时间序列预测框架

	范式	方法描述	优点	缺点
[Stationary @NIPS'22] [Autoformer @NIPS'21] [Informer @AAAI'21]	变量混淆	将所有变量在每一个时间点通过线性变换混合成一个单一的向量	• 捕获时间依赖性强 • 结构简单	• 易受冗余信息影响 • 混淆个变量间的关系
[PatchTST @ICLR'23] [Dlinear @AAAI'23]	变量独立	每个变量都单独处理，不考虑它们之间的依赖关系	• 鲁棒性强，能够适应数据分布的变化 • 预测性能普遍优于SM框架	• 忽视了变量间的关系 • 在特定形式数据上表现很差





SM (b) Series-mixing Framework



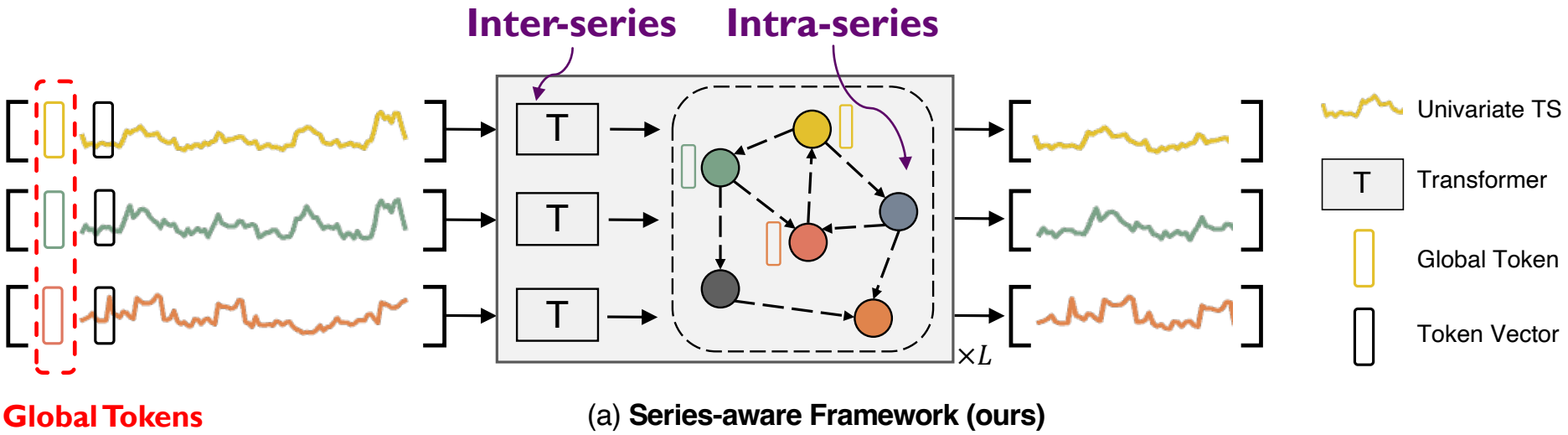
SI (c) Series-independent Framework

大部分情况下变量独立的预测效果优于变量混淆框架，但是特定情况下也会失效

亟需一个新的联合建模预测框架！

创新思路：全新多元关系联合建模预测框架

- **变量内建模**：引入全局令牌（Global Token）以捕获时间模式的多样性
- **变量间建模**：使用稀疏连接图结构（GNN）在全局令牌间传递信息，以减少冗余信息的影响
- 结合前述两个框架的特长：在SI框架的基础上增加Global token，使得即可以保留SI的优秀预测性能，同时有效建模了多变量间的依赖关系



(a) **Series-aware Framework (ours)**

所提方法：如何建模序列间的依赖关系？

多元时间序列

有先天图结构？

否

额外信息？

否

利用数据学习关系图

实验1

E随机初始化，学习时序数据关系

$\mathbf{M}_1 = \mathrm{act}_1(\mathbf{E}\Theta_1)$;

$\mathbf{M}_2 = \mathrm{act}_1(\mathbf{E}\Theta_2)$;

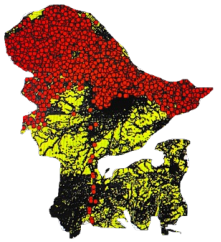
$\mathbf{A}' = Relu(\mathbf{M}_1\mathbf{M}_2^T - \mathbf{M}_2\mathbf{M}_1^T)$

非对称关系

是

实验2（华为合作，落地应用）

- 空间信息
- 文本标签
- 层次结构
- ……



否

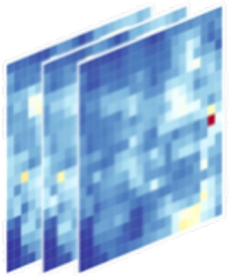
是

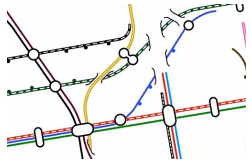
直接使用？

是

实验3

- 网格数据
- 社交网络
- 交通网络
- ……





实验1：公开数据集（无天然图结构）

Table 2: Long-term forecasting task. Bold/underline indicates the best/second. Blue background marks the models explicitly utilizing inter-series dependencies; green marks series-independent neural models; yellow marks series-mixing transformer-based models. All the results are averaged from 4 different prediction lengths, that is {24, 36, 48, 60} for ILI and {96, 192, 336, 720} for the others. See Table 8 in Appendix A for the full results.

Models	SageFormer (Ours)		Crossformer [ICLR'23]		MTGNN [KDD'20]		LSTnet [SIGIR'18]		PatchTST [ICLR'23]		DLinear [AAAI'23]		Stationary [NIPS'22]		Autoformer [NIPS'21]		Informer [AAAI'21]		Transformer [NIPS'17]	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Traffic	**0.436**	**0.285**	0.570	0.312	0.592	0.317	0.736	0.450	0.471	0.298	0.625	0.383	0.624	0.340	0.628	0.379	0.854	0.416	0.661	0.363
Electricity	**0.175**	**0.273**	0.314	0.366	0.333	0.378	0.440	0.494	0.200	0.288	0.212	0.300	0.193	0.296	0.227	0.338	0.311	0.397	0.272	0.367
Weather	**0.249**	**0.275**	0.256	0.305	0.290	0.348	0.768	0.672	0.256	0.279	0.265	0.317	0.288	0.314	0.338	0.382	0.634	0.548	0.611	0.557
ETTm1	**0.387**	**0.399**	0.509	0.507	0.566	0.537	1.947	1.206	0.389	0.401	0.403	0.407	0.481	0.456	0.588	0.517	0.961	0.734	0.936	0.728
ETTm2	**0.277**	**0.322**	1.433	0.747	1.287	0.751	2.639	1.280	0.280	0.328	0.350	0.401	0.306	0.347	0.327	0.371	1.410	0.810	1.478	0.873
ETTh1	**0.431**	**0.433**	0.615	0.563	0.679	0.605	2.113	1.237	0.443	0.443	0.456	0.452	0.570	0.537	0.496	0.487	1.040	0.795	0.919	0.759
ETTh2	**0.374**	**0.403**	2.170	1.175	2.618	1.308	4.382	2.008	0.381	0.404	0.559	0.515	0.526	0.516	0.450	0.459	4.431	1.729	4.492	1.691
Exchange	**0.354**	**0.400**	0.756	0.645	0.786	0.674	1.681	0.197	**0.354**	**0.400**	**0.354**	0.414	0.461	0.454	0.613	0.539	1.550	0.998	1.386	0.898
ILI	2.113	**0.877**	3.417	1.214	4.861	1.507	5.300	1.657	2.065	0.882	2.616	1.090	**2.077**	0.914	3.006	1.161	5.137	1.544	4.784	1.471

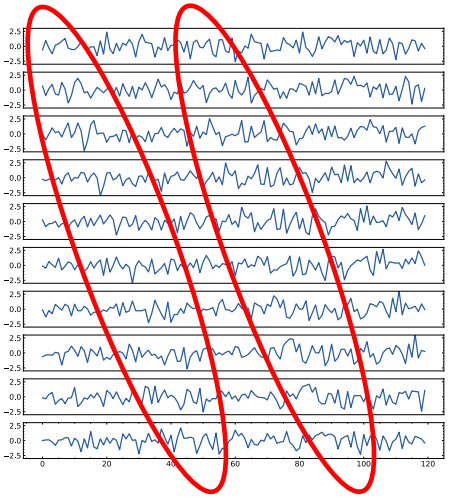
提升7.4%

提升9.3%

九个广泛使用的公开数据集

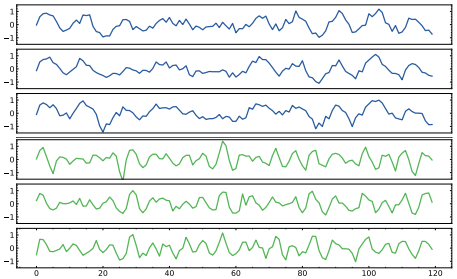
参考算法：　显式建模关系　变量独立　变量混淆

实验1：合成数据集



Directed Cycle Graph Dataset.

$$x_{i,t} \sim \mathcal{N}\left(\beta x_{(i-1 \bmod N),t-10}; \sigma^2\right)$$



Low-Rank Dataset.

$$x_i = \sum_{m=1}^{M} B_{i,m} \sin\left(2\pi\omega_{\lfloor i/K \rfloor,m} t\right) + \epsilon_{i,t}$$

- *Directed Cycle Graph Dataset*：循环依赖的多变量数据，为了检验模型是否可以学习到数据中真实的依赖关系
- *Low-Rank Dataset*：低秩数据集，为检验模型在高维变量情况下捕获信息以及对抗噪声的能力

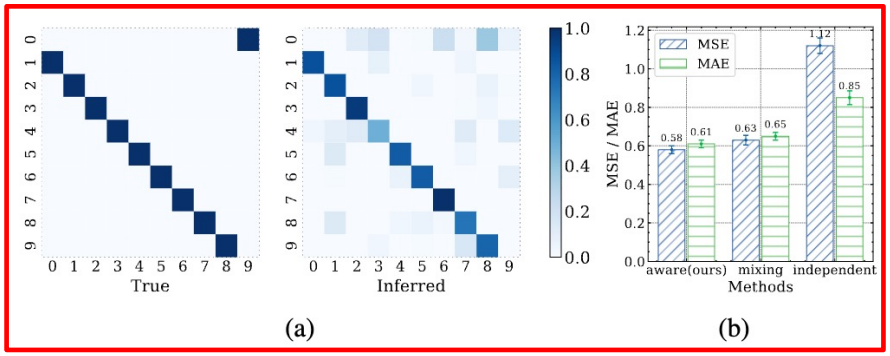


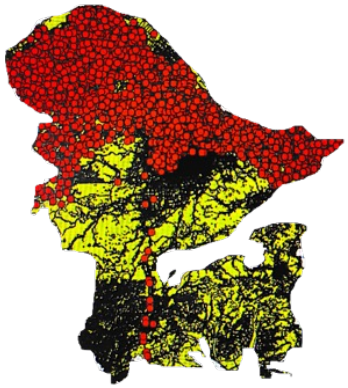
Figure 4: Evaluation on synthetic datasets. (a) The left side displays the heat map of the actual adjacency matrix, while the right side presents the inferred adjacency matrix by SageFormer, illustrating the effectiveness of our proposed method in learning the inherent graph structure.; (b) Prediction results of three different methods on the Directed Cycle Graph dataset; (c) Prediction MAE results for low-rank datasets with varying numbers of series (N). We selected the Nonstationary Transformer for the series-mixing method, and for the series-independent method, we chose PatchTST as a representative.

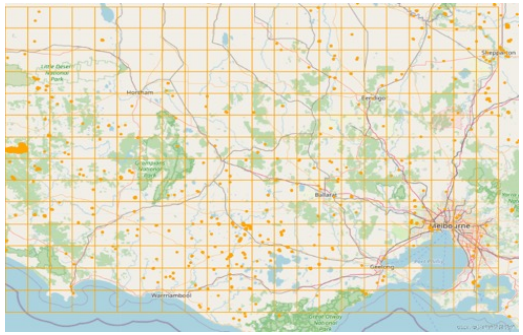
实验2：地理辅助信息构建图

如何联合电信时序流量和基站空间特征，构建统一的时空语义表征？

静态特征：

- **POI分类计数**：
  - 出入口、公司企业、购物、交通设施、教育培训、金融、酒店…
- **地面覆盖类别占比**：
  - **建筑**：超高建筑、高层建筑、中等建筑区、一般城区、工业区、郊区
  - **陆地**：城市开阔地、乡村开阔地
  - **植被**：湿地、普通绿地、灌木林、林地、公园绿地、城市树木
  - **水域**：海洋、内陆水域

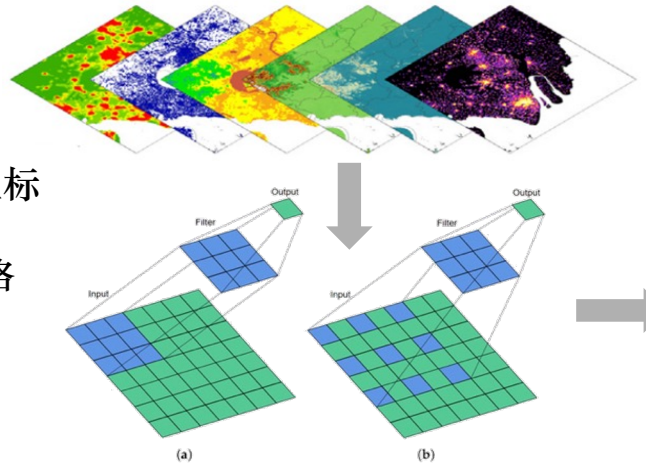




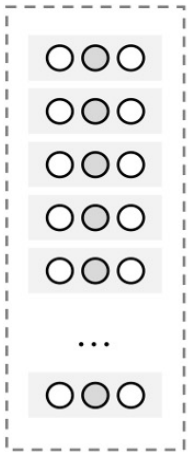
地图数据栅格化

转换UTM坐标
提取小区
周围的栅格

不同种类的栅格数据



CNN网络提取空间特征



小区邻域时空语义

实际应用：华为基站KPIs真实数据

宁波和石家庄的实际基站数据

Methods	24(1day)								
	用户数 Users			下行流量TrafficDL(1e3)			资源利用 PrbRatioDL		
	MSE	MAE	R2	MSE	MAE	R2	MSE	MAE	R2
Naive	213.4852	7.557	-0.836	2.5348	0.9507	-1.7979	176.0801	8.0461	-3.7229
HI	30.614	2.5624	0.9011	0.7901	0.5224	0.5897	40.0131	3.7831	0.658
ARIMA	231.6123	8.2522	-0.8093	2.7307	1.0399	-1.6782	180.738	8.9236	-3.4471
XGBoost	206.7967	7.4296	-0.7967	2.4292	0.9285	-1.7104	169.1741	7.8712	-3.5919
DLinear	25.7267	2.393	0.9108	0.5534	0.4327	0.724	29.1787	3.2099	0.7544
SciNet	36.233	3.0215	0.8326	0.6338	0. 4937	0.6244	34.7726	3.7111	0.6106
MTGNN	27.5354	2.3489	0.9102	0.5381	0.4197	0.7392	28.8914	3.1144	0.7698
Proposed	21.3096	2.1322	0.9277	0.5036	0.4088	0.7516	25.819	2.9708	0.7877
IMPROVE	**17.17%**	**9.23%**	**1.86%**	**6.41%**	**2.60%**	**1.68%**	**10.63%**	**4.61%**	**2.33%**

小时级别

平均提升：**6.28%**

Methods	30(days)								
	Users			TrafficDL(1e3)			PrbRatioDL		
	MSE	MAE	R2	MSE	MAE	R2	MSE	MAE	R2
Naive	133.779	6.7663	0.6664	2.8363	1.0367	0.5933	137.2541	7.5195	0.6306
HI	196.6731	6.7468	0.5094	3.6062	1.1495	0.4819	179.3554	8.4229	0.5167
ARIMA	147.7822	6.4139	0.6049	3.31338	1.1536	0.5811	153.2203	8.3663	0.5872
XGBoost	130.4876	6.0233	0.6745	2.7671	1.0185	0.6033	133.5232	7.3842	0.6407
SciNet	167.1014	7.1703	0.5812	3.1891	1.1437	0.5448	171.9799	8.6319	0.5374
MTGNN	152.0835	6.6977	0.6186	3.7079	1.2192	0.4706	155.5192	8.0487	0.5818
Proposed	116.5232	5.716	0.7089	2.4642	0.9643	0.6489	127.5225	7.2488	0.6577
IMPROVE	**10.70%**	**5.10%**	**5.10%**	**10.95%**	**5.32%**	**7.56%**	**4.49%**	**1.83%**	**2.65%**

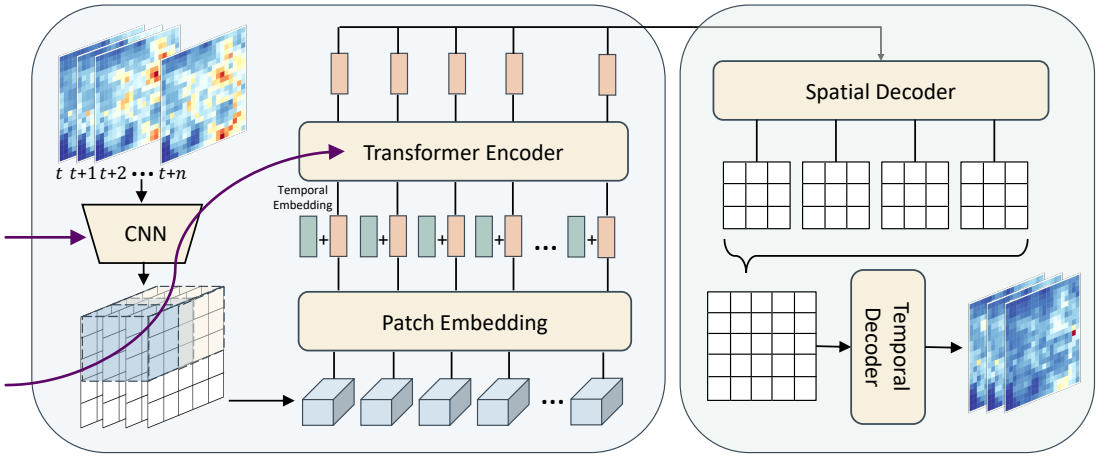
天级别

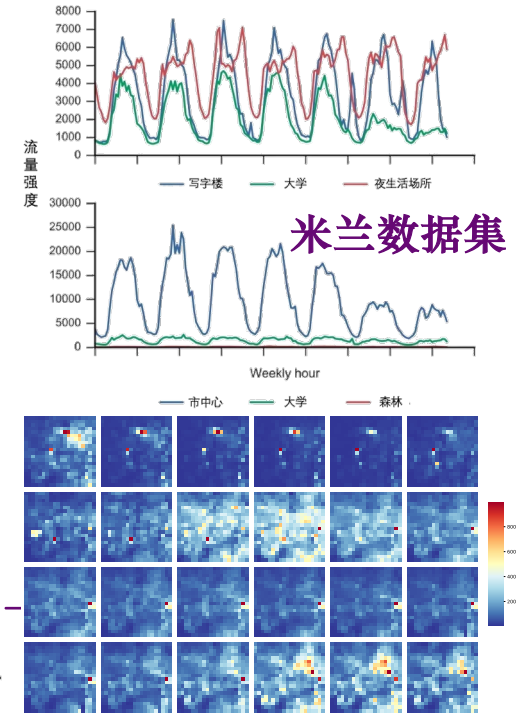
平均提升：**5.97%**

实验3：时空网格数据（特殊图结构）

短距离变量关系

长距离变量关系





米兰数据集

TABLE I
CELLULAR TRAFFIC PREDICTION RESULTS ON MILAN DATASET (THREE CASES).

	Dataset	SMS			Call			Internet		
	Metrics	RMSE↓	MAE↓	R2↑	RMSE↓	MAE↓	R2↑	RMSE↓	MAE↓	R2↑
Temporal	HI	78.0939	35.0401	0.7842	57.1934	31.1083	0.8832	223.5931	120.7588	0.9203
	ARIMA	71.8223	40.1875	0.7531	47.7327	24.6042	0.8226	240.3088	150.029	0.8902
	LSTM	72.8344	39.6033	0.7832	44.1100	22.3111	0.8399	231.7825	141.3343	0.9004
	DLinear	74.4177	35.6869	0.7804	46.5896	26.5153	0.9183	224.0433	130.6027	0.9400
Spatiotemporal	ST-DenseNet	60.3758	31.3021	0.8191	43.9073	19.6701	0.8437	196.3721	125.0611	0.9290
	STCNet	54.1664	30.3221	0.8590	34.3346	17.9901	0.9102	167.3321	93.8873	0.9500
	StTran	62.5428	33.4029	0.8198	39.8231	16.4982	0.8959	169.7015	91.3930	0.9489
	MVSTGN	49.0515	24.9796	**0.8856**	30.9443	14.6816	0.9310	165.0445	88.6983	0.9550
	ST2T(ours)	**40.0087**	**23.9413**	0.8594	**21.5861**	**14.4490**	**0.9352**	**138.649**	**79.658**	**0.9562**

研究内容二：小结

- ✓ 提出了一个结合变量关系和时序模式建模的多元时序预测框架，并在多个数据集上优于参考算法
- ✓ 针对多种数据（空间，POI等）设计处理网络，作为辅助变量提升时序预测性能

论文成果：

- **Zhenwei Zhang**, Leon Yan, Yuantao Gu. “ST2T: A Spatio-Temporal Transformer for Cellular Traffic Prediction in Digital Twin Systems.” 2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT), Qingdao, China, 2023, pp. 1112-1117, doi: 10.1109/ICEICT57916.2023.10245385.（EI会议）
- **Zhenwei Zhang**, Xin Wang and Yuantao Gu. “SageFormer: Series-Aware Graph-Enhanced Transformers for Multivariate Time Series Forecasting.” KDD-MILETS, 2023.（CCF- A推荐会议Workshop）

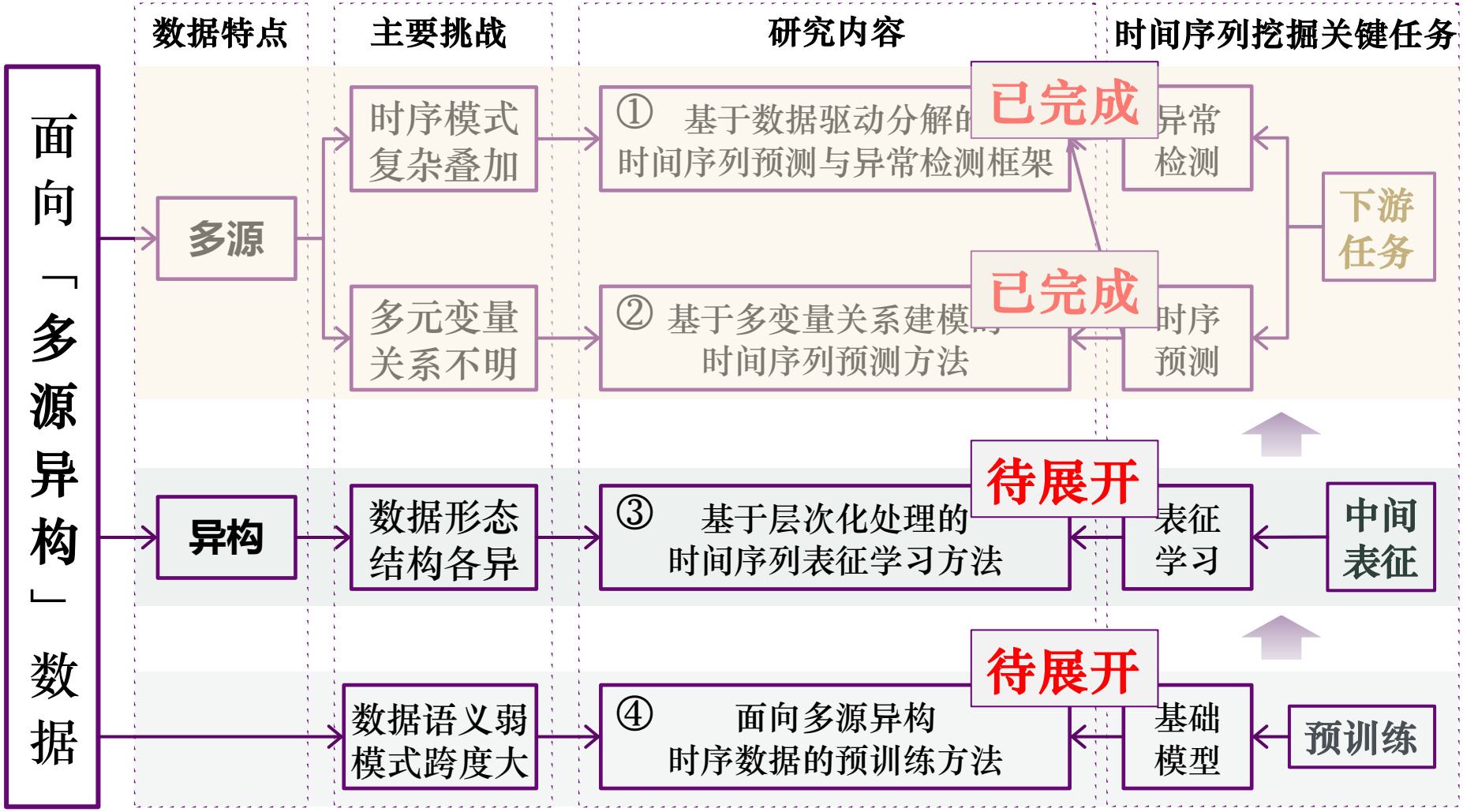
专利成果：

- 徐灏,**张振威**,陈艺茗,丁冉,谷源涛. 一种话务需求预测方法以及系统. 华为技术有限公司,清华大学.（发明专利，清华第一作者，已受理，申请号202311055232.2）

在投论文：

- **Zhenwei Zhang**, Linghang Meng, Yuantao Gu. “Series-Aware Framework for Long-term Multivariate Time Series Forecasting.”（IOTJ在投）

回顾：研究课题设置与逻辑关系



研究内容三、基于层次化处理的时间序列表征学习方法

研究问题：如何针对多源异构流式数据进行统一表征？

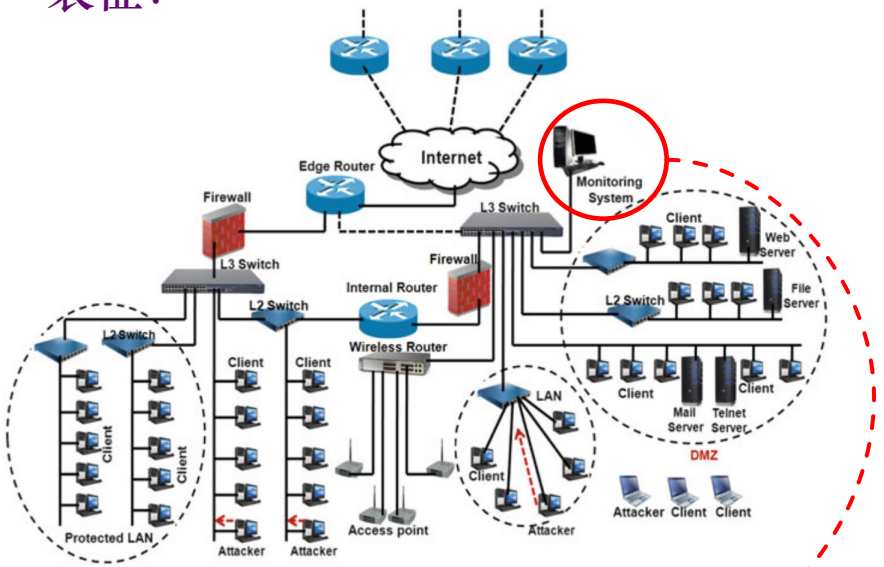
```
d4 c3 b2 a1 02 00 04 00
00 00 00 00 00 00 00 00
00 00 04 00 01 00 00 00
00 45 d4 5e 18 8e 0c 00
42 00 00 00 42 00 00 00
00 1e ec 26 d2 ac 26 02
06 49 6b 31 08 00 45 02
00 34 30 8c 40 00 72 06
81 7f 2e 69 63 a3 c0 a8
04 02 cf 3a 00 50 8d a5
ee 7b 00 00 00 00 80 c2
20 00 ac 29 00 00 02 04
05 78 01 03 03 08 01 01
04 02 00 45 d4 5e 2c 77
0d 00 36 00 00 00 36 00
00 00 00 1e ec 26 d2 ac
```

24 byte PCAP Header
Link-Layer Type = Ethernet (0x00000001)

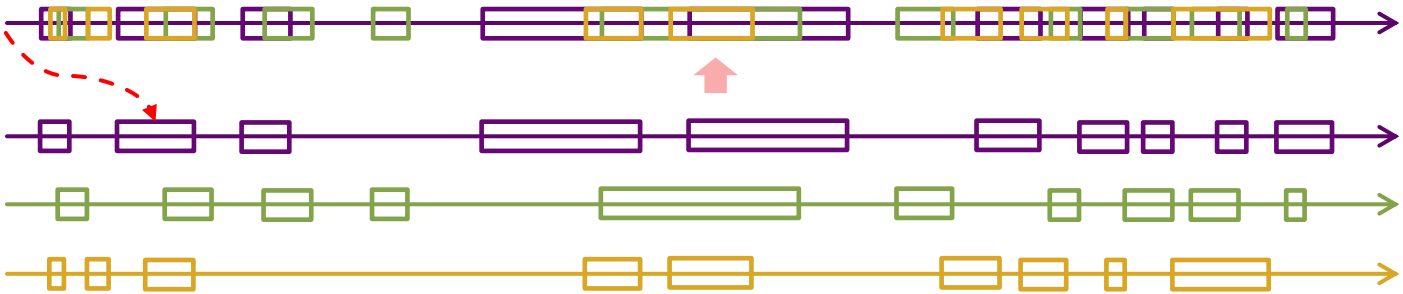
16 byte Packet Header
Timestamp = 1 June 2020
Packet length = 66 bytes (0x00000042)

66 bytes of Packet Data
Destination MAC = 00:1e:ec:26:d2:ac
Source MAC = 26:02:06:49:6b:31
Source IP = 46.105.99.163
Destination IP = 192.168.4.2

16 byte Packet Header
Packet length = 54 bytes (0x00000036)



观测值：



NetFlow 1

NetFlow 2

NetFlow 3

Packet：大小不同，到达间隔不同，协议格式不同，发送接收设备不同

工作基础：应用层协议推断

已实现七种业务中关键协议（共26种网络协议）的指纹推断

业务类型	完成的关键协议	官方文档
VOIP	SIP,RTCP	RFC3261，RFC3263，RFC4961
多媒体流	RTSP，RTP	RFC4567，RFC3550，RFC7604
批量数据传输	FTP_CONTROL, FTP_DATA,	RFC0114，RFC0141，RFC0959，RFC2228，RFC7151
交互式网络应用	XDMCP,VMware, Telnet, RDP,VNC, PCAnyWhere, SSH, TeamViewer	RFC0854，RFC1151，RFC8182，RFC4250~4256，RFC7869
邮件服务	POP3, SMTP,IMAP	RFC1939，RFC3501，RFC5321
WWW	HTTP，HTTP_Connect, HTTP_Proxy, AJP	RFC1945，RFC2068，RFC7230
数据库应用	PostgreSQL, MySQL, MsSQL-TDS，Redis, DRDA	各个数据库应用的官方文档

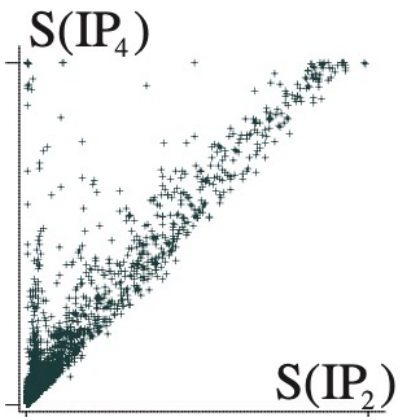
工作基础：主机时序行为画像

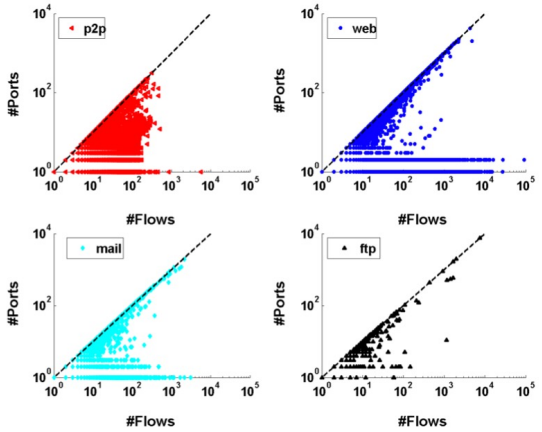
定义：在给定的时间间隔内对主机的一系列时序行为进行抽象描述

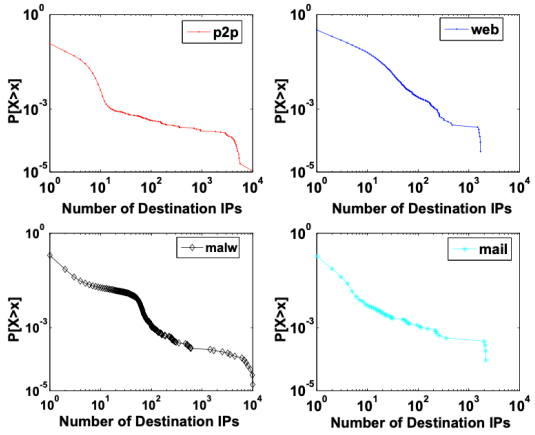
- **连接层次**
  - 通联IP个数
  - 通联IP集合各字节熵之比
- **功能层次**
  - 活跃源/目的端口个数
  - 系统、用户、动态端口
- **流量层次**
  - 网络流包长度
  - 网络流包个数
- **应用层次**
  - 网络各类业务

主机画像

- 主机ID
- 主机操作系统
- 主机IP地址
- 主机活跃的端口
- **主机业务分布**
- ……





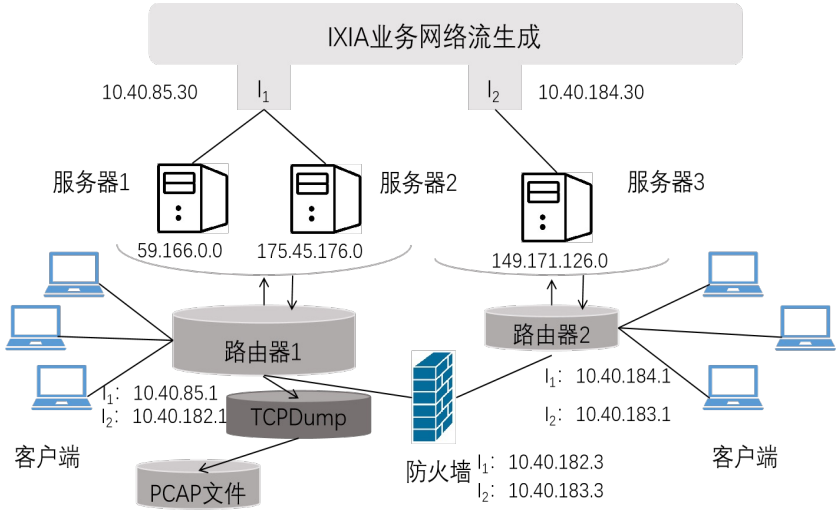


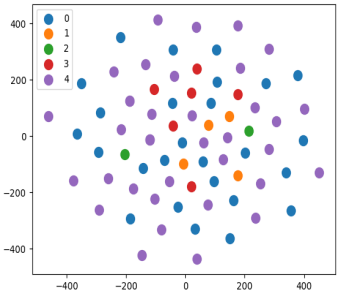
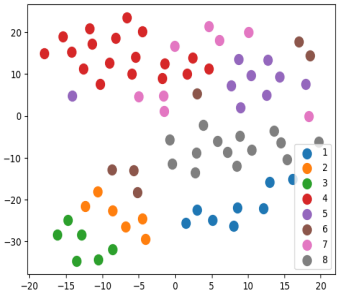
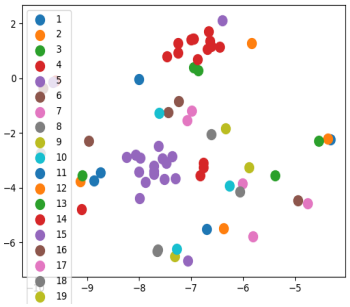
工作基础：初步实验验证

公开数据集：The UNSW-NB15 Dataset

- 大小：1.86GB
- 主机个数：39个
- 记录条数：351万

根据主机的时序行为模式和行为特征描述，聚类获取主机群类型和主机角色的推断

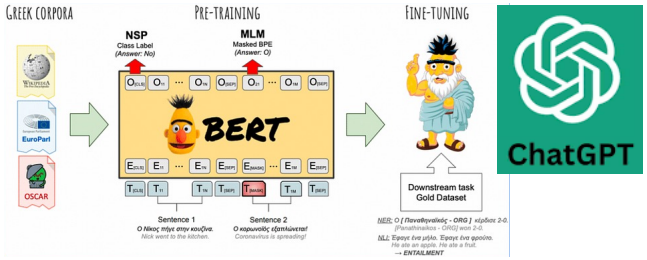


	BCH算法	PCH算法	所提方法
Silhouette Score 轮廓系数（类内外间距）	-0.9734	0.1792	**0.1943**
tSNE			

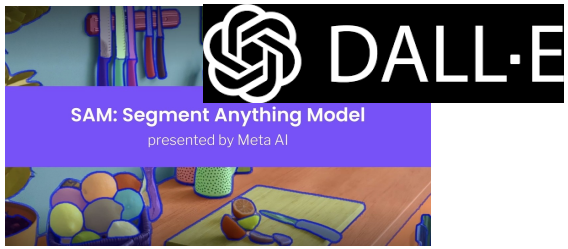
研究内容四、面向多源异构时间序列数据的预训练方法研究

研究问题：如何构建多源异构时间序列挖掘的预训练模型?

自然语言处理



计算机视觉

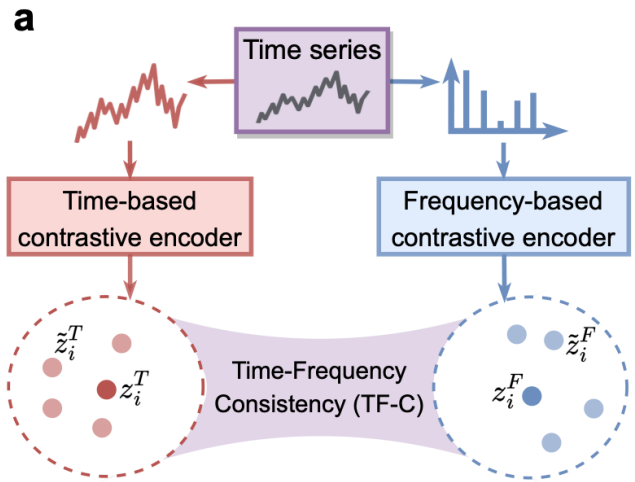


时间序列

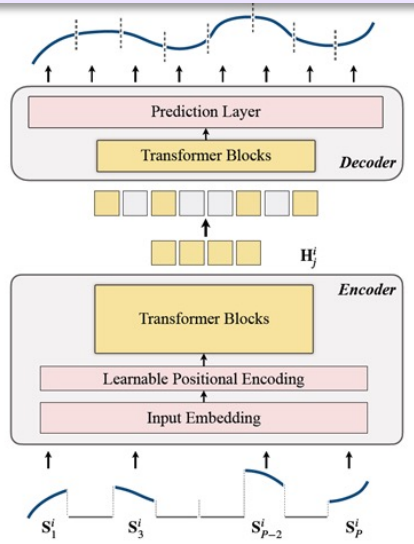
?

现有工作

[NeruIPS'22] 基于对比学习的预训练



[SIGKDD'22] 基于掩码重建的预训练



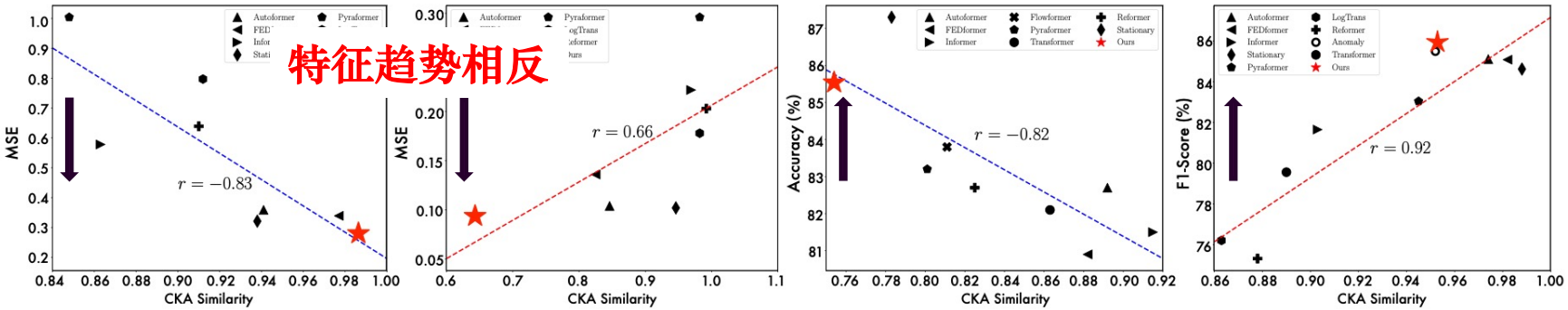
挑战：时序数据语义弱 模式跨度大

?

预训练无统一范式!

工作基础：时间序列预训练任务研究

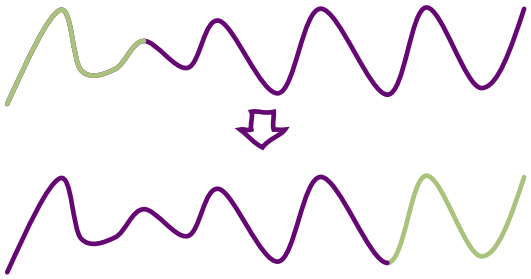
目前大部分的时间序列预训练任务都是基于MASK的补全任务，但是任务特征不通用



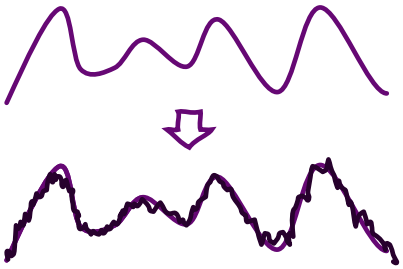
图源自[TimesNet@ICLR'23]

我们提出了三个辅助预测的新预训练任务，补充时间序列的预训练范式

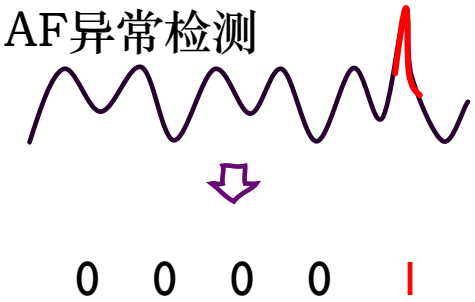
① IF反向



② CF平滑



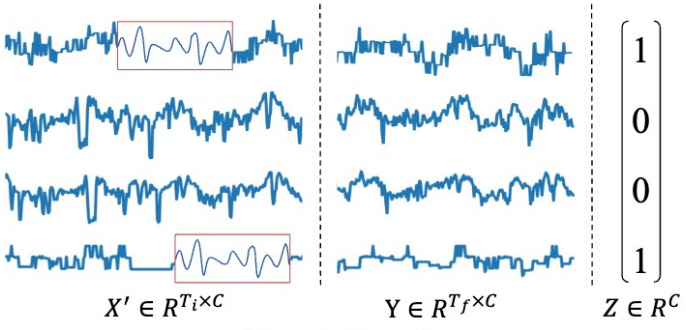
③ AF异常检测



研究内容四、面向多源异构时序数据的预训练方法

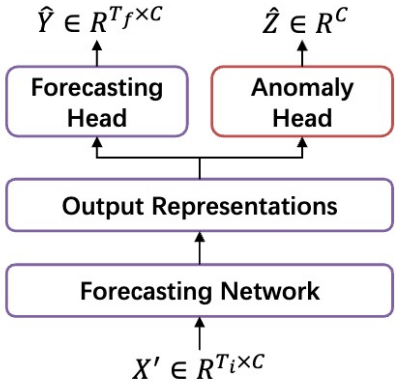
工作基础：时间序列预训练任务实验

③ AF异常检测预训练任务（续）



(a) Anomaly Generation

增强时序语义信息



(b) Anomaly Forecasting Backbone

Fig. 2. (a) gives an overview of data with anomaly series in the Anomaly Forecasting task. (b) is a brief schematic of the backbone of the Anomaly Forecasting task, where the purple framework is the same forecasting backbone as the downstream task.

公开数据集实验结果

FULL RESULTS FOR FORECASTING TASK. WE COMPARE MODELS PRE-TRAINED BY OUR PROPOSED METHODS WITH BASELINES UNDER DIFFERENT PREDICTION LENGTHS. THE INPUT SEQUENCE LENGTH IS SET TO 336 FOR THE ETTM1 AND ETTM2 DATASETS, AND 512 FOR THE OTHERS. *Avg.* IS AVERAGED FROM ALL FOUR PREDICTION LENGTHS.

Models		**Composite**		IF		CF		AF		SMAE [13]		PatchTST [13]	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	**0.142**	**0.191**	0.144	0.194	0.146	0.197	0.145	0.196	0.144	0.193	0.150	0.198
	192	**0.188**	**0.236**	0.189	0.237	0.191	0.240	0.195	0.244	0.190	**0.236**	0.197	0.241
	336	**0.243**	**0.277**	**0.243**	0.279	0.244	0.281	0.245	0.280	0.244	0.280	0.250	0.283
	720	**0.311**	**0.329**	0.313	0.334	0.313	0.332	0.318	0.334	0.320	0.335	0.325	0.337
	Avg.	**0.221**	**0.258**	0.222	0.261	0.224	0.263	0.226	0.264	0.225	0.261	0.231	0.265
ETTh1	96	**0.366**	**0.392**	**0.366**	0.393	0.373	0.398	0.371	0.397	0.366	0.397	0.367	0.394
	192	**0.398**	**0.417**	0.411	0.420	0.415	0.423	0.413	0.429	0.431	0.443	0.414	0.422
	336	**0.411**	**0.428**	0.418	0.430	0.425	0.444	0.414	0.430	0.450	0.456	0.432	0.436
	720	**0.432**	**0.459**	0.434	**0.459**	0.450	0.470	0.464	0.473	0.472	0.484	0.457	0.470
	Avg.	**0.403**	**0.424**	0.407	0.426	0.416	0.434	0.414	0.432	0.430	0.445	0.420	0.431
ETTh2	96	0.281	0.337	0.283	0.338	0.282	0.341	0.285	0.341	0.284	0.343	**0.275**	**0.336**
	192	0.350	**0.378**	**0.349**	0.383	0.363	0.393	0.361	0.390	0.355	0.387	0.359	0.392
	336	**0.325**	**0.377**	0.335	0.386	0.333	0.389	0.337	0.386	0.379	0.411	0.351	0.394
	720	**0.380**	0.423	0.390	0.429	0.385	0.427	0.382	0.424	0.400	0.435	**0.380**	**0.422**
	Avg	**0.334**	**0.378**	0.339	0.384	0.341	0.388	0.341	0.385	0.355	0.394	0.341	0.386
ETTm1	96	**0.285**	**0.340**	0.291	**0.340**	0.293	0.345	0.295	0.343	0.289	0.344	0.292	0.343
	192	**0.322**	**0.367**	0.332	0.371	0.363	0.392	0.339	0.374	0.323	0.368	0.331	0.369
	336	0.356	0.392	0.365	0.392	0.368	0.394	0.361	0.394	**0.353**	**0.387**	0.365	0.392
	720	0.415	0.425	0.413	0.422	0.418	0.424	0.407	0.418	**0.398**	**0.416**	0.421	0.425
	Avg	0.344	0.381	0.350	0.381	0.361	0.389	0.350	0.382	**0.341**	**0.379**	0.352	0.382
ETTm2	96	0.167	0.257	0.166	**0.252**	0.165	0.253	0.176	0.263	0.166	0.256	**0.164**	0.254
	192	0.227	0.295	0.221	0.293	**0.220**	**0.291**	0.235	0.304	0.221	0.295	0.221	0.292
	336	0.285	0.334	0.276	0.328	**0.275**	**0.327**	0.290	0.339	0.278	0.333	0.278	0.329
	720	**0.362**	0.383	0.367	0.384	0.364	**0.382**	0.381	0.395	0.365	0.388	0.367	0.385
	Avg	0.260	0.317	0.258	0.314	**0.256**	**0.313**	0.270	0.325	0.258	0.318	0.257	0.315
1st Count		25		6		5		0		5		5	

研究内容三和四、小结

- ✓ 实现了异构时序网络流量的解析框架，多种关键协议的指纹推断，并进行主机时序行为画像和聚类分析
- ✓ 提出了三个有效辅助预测的创新预训练任务，补充时间序列的预训练范式

论文成果：

- Leon Yan*, **Zhenwei Zhang***, Xin Wang, Yu Zhang, Yuantao Gu. "Pre-training Approaches for Multivariate Time Series Forecasting Frameworks." IEEE 6th International Conference on Electronic Information and Communication Technology, 2023. （EI会议）
- 蒙治伸,**张振威**,朱思义,谷源涛.基于网络流量特征分析的主机群分析研究.计算机仿真, 2022.（北大核心）

专利成果：

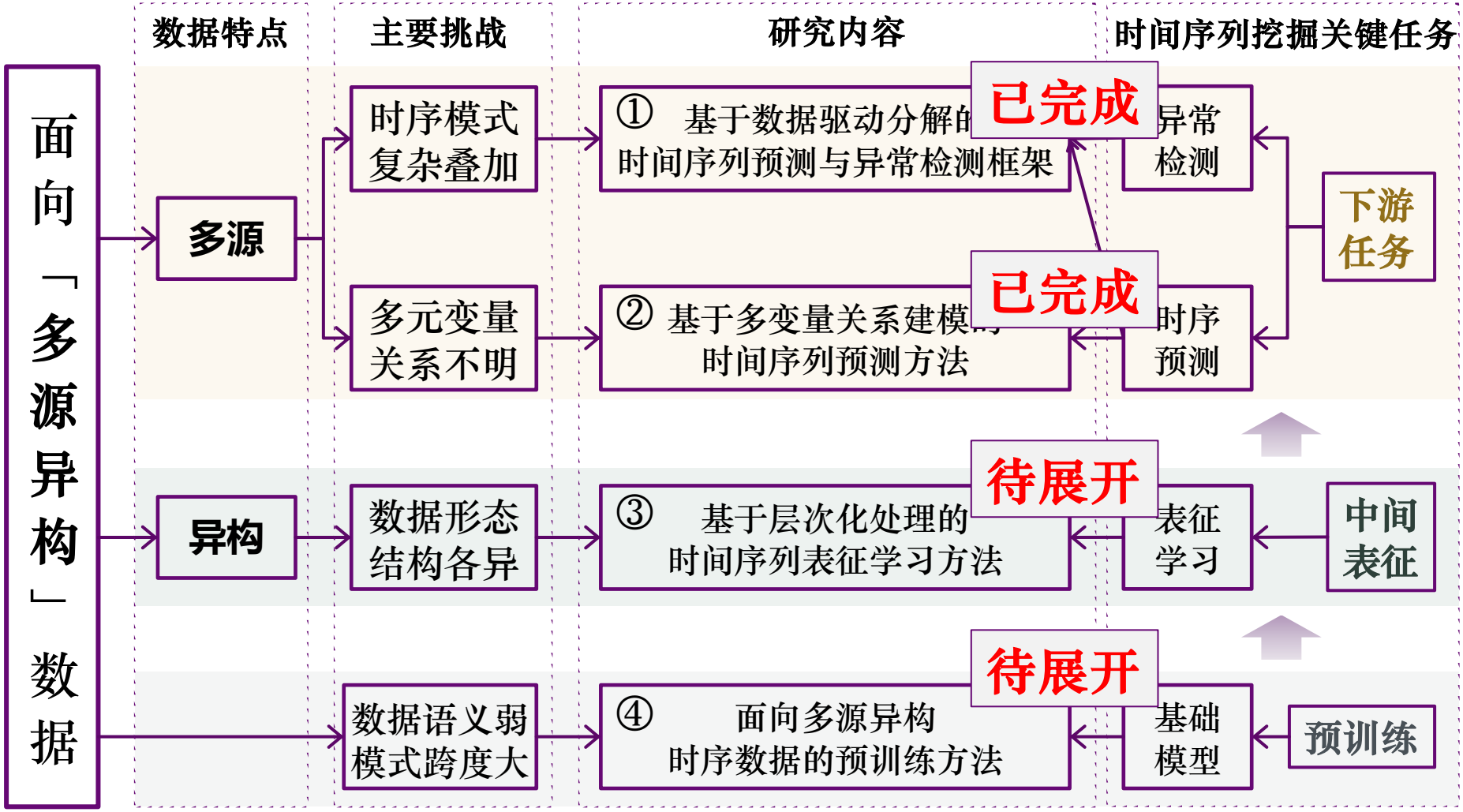
- 谷源涛,蒙治伸,**张振威**. 主机聚类方法和装置. 清华大学.（发明专利，学生第二作者，已授权，授权号：CN 113452714 B）

软件成果：

- 「网络流量特征分析软件」作为项目成果已交付给XXX，在多个地区部署应用。

四、未来研究计划

已完成和带展开的工作计划



研究内容三：基于层次化处理的时间序列表征学习方法

（待展开）

- 目前的时序异构数据大多都是针对特定数据进行处理和表征，尚无通用方法
- 拟以网络流量数据为出发点，设计多源异构时序数据的通用表征方法
- 利用层次化表征方法，对复杂异构数据进行分层表示，得到多层次的表征向量，便于后续多任务处理

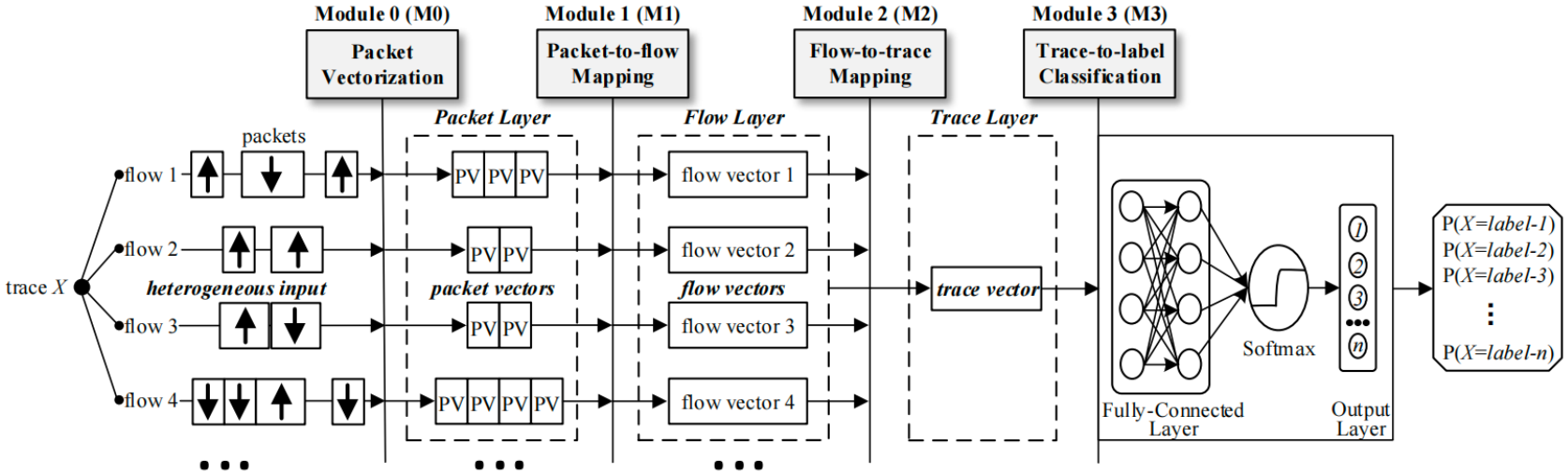


Figure 1: The input-agnostic hierarchical deep learning framework supporting heterogeneous input for traffic fingerprinting.

图源自[Trace@USENIX-Security'23]

研究内容四：面向多源异构时序数据的预训练方法

（待展开）

- 目前针对时间序列预训练方法尚无统一的框架，性能相比于直接训练并无本质提升
- NLP和CV领域虽有大量完善工作，但数据结构不同，时间序列语义更弱，大量工作难以直接迁移至本领域
- 拟针对多源异构时序数据，借鉴NLP领域的经验，探索更适用于时间序列数据的预训练任务与目标，建立时间序列基础模型

参考文献：

1. Yeh, Chin-Chia Michael, et al. "Toward a Foundation Model for Time Series Data." *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2023.
2. Garza, Azul, and Max Mergenthaler-Canseco. "TimeGPT-1." *arXiv preprint arXiv:2310.03589* (2023).
3. Zhang, Xiang, et al. "Self-supervised contrastive pre-training for time series via time-frequency consistency." *Advances in Neural Information Processing Systems* 35 (2022): 3988-4003.
4. Shao, Zezhi, et al. "Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting." *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022.
5. Sagheer, Alaa, and Mostafa Kotb. "Unsupervised pre-training of a deep LSTM-based stacked autoencoder for multivariate time series forecasting problems." *Scientific reports* 9.1 (2019): 19038.

时间进度安排

- 2023.10-2024.04
  - 完成「基于层次化处理的时间序列表征学习方法」
  - 整理研究结果并撰写学术论文
- 2024.04-2025.01
  - 完成「面向多源异构时序数据的预训练方法」
  - 整理研究结果并撰写学术论文
- 2025.02-2025.07
  - 资料整理，撰写学术论文
  - 撰写博士论文，准备答辩

五、已取得及预期成果

学术成果

◆ 已录用，第一作者：

1. **Zhenwei Zhang**, Xin wang, Jingyuan Xie, Heling Zhang and Yuantao Gu. “Unlocking the Potential of Deep Learning in Peak-Hour Series Forecasting.” CIKM, 2023. （CCF-B推荐会议，接受率27.4%）
2. **Zhenwei Zhang**, Leon Yan, Yuantao Gu. “ST2T: A Spatio-Temporal Transformer for Cellular Traffic Prediction in Digital Twin Systems.” IEEE 6th International Conference on Electronic Information and Communication Technology, 2023. （EI会议）
3. **Zhenwei Zhang**, Xin Wang and Yuantao Gu. “SageFormer: Series-Aware Graph-Enhanced Transformers for Multivariate Time Series Forecasting.” KDD-MILETS, 2023. （CCF- A推荐会议Workshop）
4. **Zhenwei Zhang**, Xin wang, Jingyuan Xie, Heling Zhang and Yuantao Gu. “Bridge the Performance Gap in Peak-hour Series Forecasting: The Seq2Peak Framework.” KDD-MILETS, 2023. （CCF- A推荐会议Workshop）

◆ 在投，第一作者：

1. **Zhenwei Zhang**, Ruiqi Wang, Yuantao Gu. “Unravel Anomalies: An End-to-end Seasonal-Trend Decomposition Approach for Time Series Anomaly Detection.” ICASSP, 2024. （电子系顶级会议，在投）
2. **Zhenwei Zhang**, Linghang Meng, Yuantao Gu. “Series-Aware Framework for Long-term Multivariate Time Series Forecasting.” （IOTJ在投）

◆ 已录用，非第一作者：

1. Leon Yan*, **Zhenwei Zhang***, Xin Wang, Yu Zhang, Yuantao Gu. “Pre-training Approaches for Multivariate Time Series Forecasting Frameworks.” IEEE 6th International Conference on Electronic Information and Communication Technology, 2023. （EI会议，共同一作）
2. 蒙治伸,**张振威**,朱思义,谷源涛.基于网络流量特征分析的主机群分析研究.计算机仿真, 2022.（北大核心）

其他成果

◆专利成果：

1. 谷源涛,蒙治伸,**张振威**. 主机聚类方法和装置. 清华大学.（发明专利，学生第二作者，已授权，授权号：CN 113452714 B）
2. 徐灏,**张振威**,陈艺茗,丁冉,谷源涛. 一种话务需求预测方法以及系统. 华为技术有限公司,清华大学.（发明专利，清华第一作者，已受理，申请号：202311055232.2）
3. 徐灏,**张振威**,闫思成,汪昕,谷源涛.人群聚集预测方法、装置及系统. 华为技术有限公司,清华大学.（发明专利，清华第一作者，已受理，申请号：202311291079.3）

◆软件成果：

1. 「网络流量特征分析软件」作为XXX项目成果已交付给XXXX，在多个地区部署应用。
2. 「基站流量分解，预测与生成」已在华为公司宁波、石家庄、深圳等多地部署应用。

主要参考文献

- [1]Wang J., Wang Z., Li J., and Wu J., "Multilevel Wavelet Decomposition Network for Interpretable Time Series Analysis," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, in KDD '18. New York, NY, USA: Association for Computing Machinery, Jul. 2018, pp. 2437–2446. doi: 10.1145/3219819.3220060.
- [2]K. Choi, J. Yi, C. Park, and S. Yoon, "Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines," *IEEE Access*, vol. 9, pp. 120043–120065, 2021, doi: 10.1109/ACCESS.2021.3107975.
- [3]Wang Y., Du X., Lu Z., Duan Q., and Wu J., "Improved LSTM-Based Time-Series Anomaly Detection in Rail Transit Operation Environments," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 12, pp. 9027–9036, Dec. 2022, doi: 10.1109/TII.2022.3164087.
- [4]K. Choi, J. Yi, C. Park, and S. Yoon, "Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines," *IEEE Access*, vol. 9, pp. 120043–120065, 2021, doi: 10.1109/ACCESS.2021.3107975.
- [5]Wen Q., Gao J., Song X., Sun L., Xu H., and Zhu S., "RobustSTL: A Robust Seasonal-Trend Decomposition Algorithm for Long Time Series," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, Art. no. 01, Jul. 2019, doi: 10.1609/aaai.v33i01.33015409.
- [6]Chen L. and Ng R., "On the marriage of Lp-norms and edit distance," in *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*, in VLDB '04. Toronto, Canada: VLDB Endowment, Aug. 2004, pp. 792–803.
- [7]Berndt D. J. and Clifford J., "Using dynamic time warping to find patterns in time series," in *Proceedings of the 3rd international conference on knowledge discovery and data mining*, 1994, pp. 359–370. Accessed: Oct. 17, 2023. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/3000850.3000887
- [10]Zhang X., Zhao Z., Tsiligkaridis T., and Zitnik M., "Self-Supervised Contrastive Pre-Training For Time Series via Time-Frequency Consistency." arXiv, Oct. 15, 2022. Accessed: Sep. 11, 2023. [Online]. Available: http://arxiv.org/abs/2206.08496
- [11]Tay Y. *et al.*, "UL2: UNIFYING LANGUAGE LEARNING PARADIGMS," 2023.
- [12]Yamaguchi A., Chrysostomou G., Margatina K., and Aletras N., "Frustratingly Simple Pretraining Alternatives to Masked Language Modeling." arXiv, Sep. 04, 2021. doi: 10.48550/arXiv.2109.01819.
- [13]"Time-series forecasting with deep learning: a survey | Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences." Accessed: Sep. 10, 2023. [Online].
- [14]Masini R. P., Medeiros M. C., and Mendes E. F., "Machine Learning Advances for Time Series Forecasting." arXiv, Apr. 09, 2021. Accessed: Sep. 10, 2023. [Online]. Available: http://arxiv.org/abs/2012.12802
- [15]Fawaz H. I., Forestier G., Weber J., Idoumghar L., and Muller P.-A., "Deep learning for time series classification: a review," *Data Min Knowl Disc*, vol. 33, no. 4, pp. 917–963, Jul. 2019, doi: 10.1007/s10618-019-00619-1.
- [16]Qu J. *et al.*, "An {Input-Agnostic} Hierarchical Deep Learning Framework for Traffic Fingerprinting," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 589–606.
- [17]M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, *Network Traffic Anomaly Detection and Prevention*. in Computer Communications and Networks. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-65188-0.
- [20]A. Woicik, M. Zhang, J. Chan, J. Ma, and S. Wang, "Extrapolating heterogeneous time-series gene expression data using Sagittarius," *Nat Mach Intell*, vol. 5, no. 7, Art. no. 7, Jul. 2023, doi: 10.1038/s42256-023-00679-5.

主要参考文献

- [21]L. Brinkmeyer, R. R. Drumond, J. Burchert, and L. Schmidt-Thieme, “Few-Shot Forecasting of Time-Series with Heterogeneous Channels.” arXiv, Aug. 18, 2022. Accessed: Sep. 09, 2023. [Online]. Available: http://arxiv.org/abs/2204.03456
- [22]Zhou T., Niu P., Wang X., Sun L., and Jin R., “One Fits All:Power General Time Series Analysis by Pretrained LM.” arXiv, May 25, 2023. Accessed: Sep. 07, 2023. [Online]. Available: http://arxiv.org/abs/2302.11939
- [23]Ma Q. *et al.*, “A Survey on Time-Series Pre-Trained Models.” arXiv, May 18, 2023. Accessed: Sep. 07, 2023. [Online]. Available: http://arxiv.org/abs/2305.10716
- [24]Z. Hajirahimi and M. Khashei, “Hybrid structures in time series modeling and forecasting: A review,” *Engineering Applications of Artificial Intelligence*, vol. 86, pp. 83–106, Nov. 2019, doi: 10.1016/j.engappai.2019.08.018.
- [25]李爱华, 续维佳, and 石勇, “基于‘物理—事理—人理’的多源异构大数据融合探究,” 中国科学院院刊, vol. 38, no. 8, pp. 1225–1233, 2023, doi: 10.16418/j.issn.1000-3045.20220921003.
- [26]余辉, 梁镇涛, and 鄢宇晨, “多来源多模态数据融合与集成研究进展,” 情报理论与实践, vol. 43, no. 11, pp. 169–178, 2020, doi: 10.16353/j.cnki.1000-7490.2020.11.027.
- [27]S. Schmidl, P. Wenig, and T. Papenbrock, “Anomaly detection in time series: a comprehensive evaluation,” *Proc. VLDB Endow.*, vol. 15, no. 9, pp. 1779–1797, May 2022, doi: 10.14778/3538598.3538602.
- [28]朱家明, “基于时间序列特征驱动分解的多尺度组合预测模型及其应用,” 博士, 安徽大学, 2019. Accessed: Aug. 29, 2023. [Online]. [29]张小波, “基于数据特征驱动分解的季节性时间序列预测模型研究及应用,” 博士, 东北财经大学, 2020. doi: 10.27006/d.cnki.gdbcu.2020.000028.
- [30]汤兴恒, “基于成分分解的时间序列预测方法研究,” 硕士, 山东财经大学, 2023. doi: 10.27274/d.cnki.gsdjc.2023.000770.
- [31]石巍巍, “大规模多源时间序列预处理与隐藏空间映射分析研究,” 博士, 上海交通大学, 2018. doi: 10.27307/d.cnki.gsjtu.2018.000447.
- [32]胡宇鹏, “时间序列数据挖掘中的特征表示与分类方法的研究,” 博士, 山东大学, 2018. Accessed: Aug. 29, 2023. [Online].
- [33张奥千, “时间序列数据清洗方法研究,” 博士, 清华大学, 2018. doi: 10.27266/d.cnki.gqhau.2018.000702.
- [34]张琪, “时间序列流数据异常检测问题的研究,” 博士, 山东大学, 2019. [Online].
- [35]钱爱玲, “复杂结构的时间序列数据挖掘与预测方法研究,” 博士, 华中科技大学, 2011. [Online].
- [36]周琨, “网络流量模型及异常检测技术研究,” 博士, 电子科技大学, 2021. [Online]. [37]展鹏, “基于时间序列挖掘的异常检测关键技术研究,” 博士, 山东大学, 2020. doi: 10.27272/d.cnki.gshdu.2020.005871.
- [38]田楚杰, “神经网络在时间序列与时空序列流量预测中的应用与研究,” 博士, 北京邮电大学, 2021. doi: 10.26969/d.cnki.gbydu.2021.000109.
- [39]Pasini K., “Forecast and anomaly detection on time series with dynamic context: Application to the mining of transit ridership data”.
- [40Chandola V., “A THESIS SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF THE UNIVERSITY OF MINNESOTA”.
- [41]孙友强, “时间序列数据挖掘中的维数约简与预测方法研究,” 博士, 中国科学技术大学, 2014. Accessed: Aug. 27, 2023. [Online].
- [42]刘海洋, “复杂环境下时间序列预测方法研究,” 博士, 北京交通大学, 2019. Accessed: Aug. 27, 2023. [Online].

清華大學
Tsinghua University

请各位老师专家批评指正！

答辩人：　张振威

导师：　谷源涛 教授

2023年11月2日