Efficient Multivariate Time Series Anomaly Detection Through Transfer Learning for Large-Scale Web Services

Yongqian Sun, **Minghan Liang**, Zeyu Che, Dongwen Li, Tinghua Zheng, Shenglin Zhang* , Pengtian Zhu, Yuzhi Zhang, Dan Pei

ICWS 2023

Outline

Background

Approach

Evaluation

Conclusion

Outline

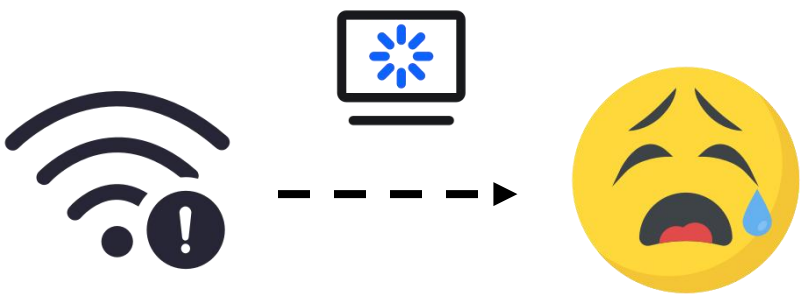
Background

Approach

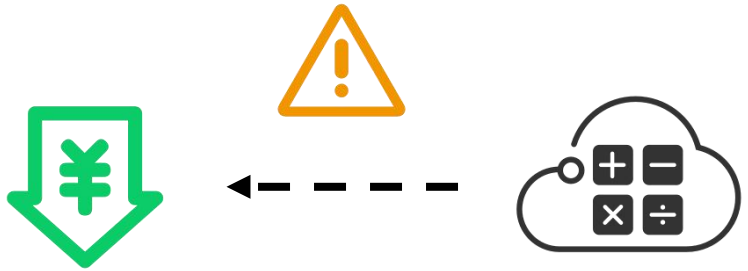
Evaluation

Conclusion

The Reliability of Web Services is Important



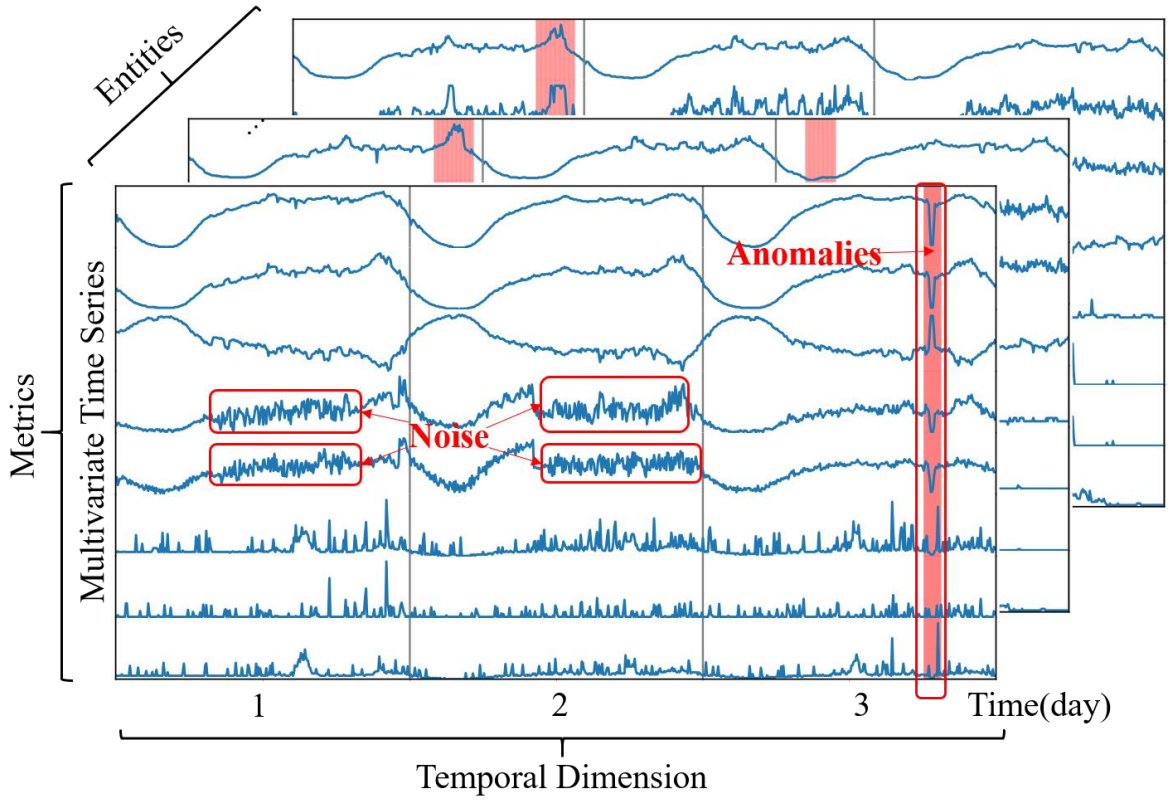
Poor user experience



A drop in revenue

MTS Anomaly Detection

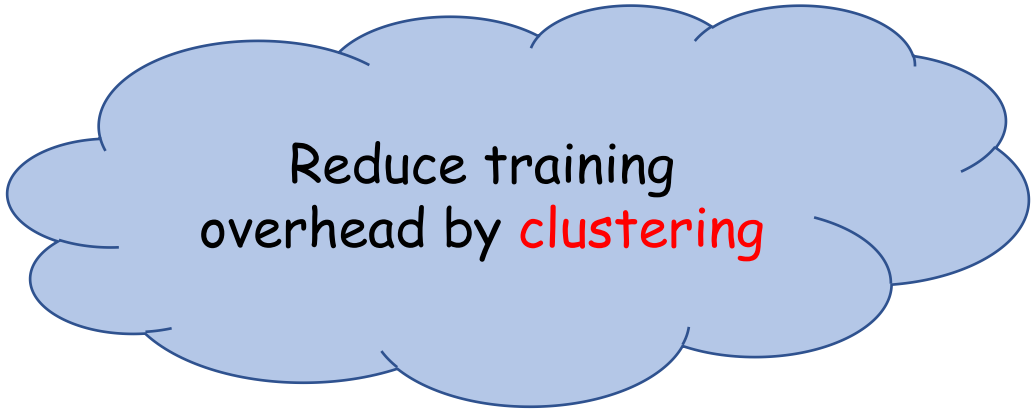
- The collected multiple metrics of each entity forms a multivariate time series (MTS)
- Determine whether the entity has anomaly behavior that deviates from the normal pattern
- Anomaly detection is critical to the quality of service (QoS) management



Motivations (1/3)

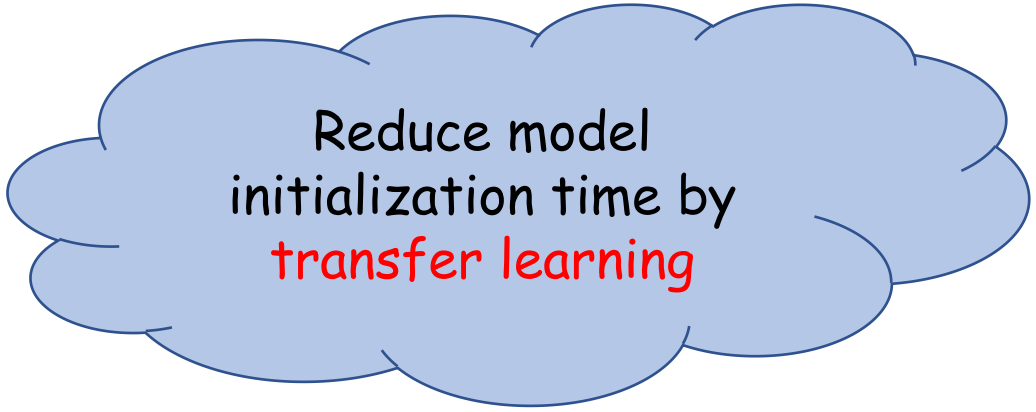
- Huge number of entities in large-scale Web services.
- Training an MTS anomaly detection model for **each entity** is resource-consuming.

Model	Time(1M Entities)
OmniAnomaly(KDD '19)	1.57years
InterFusion(KDD '21)	1.41years
DAGMM(ICLR '18)	6.09months
USAD(KDD '20)	5.72weeks
GDN(AAAI '21)	2.19weeks
TranAD(VLDB '22)	4.89weeks



Motivations (2/3)

- Frequent changes in web services lead to changes in the pattern of MTS.
- MTS anomaly detection model based on deep learning needs a lot of training data to achieve satisfactory detection performance.

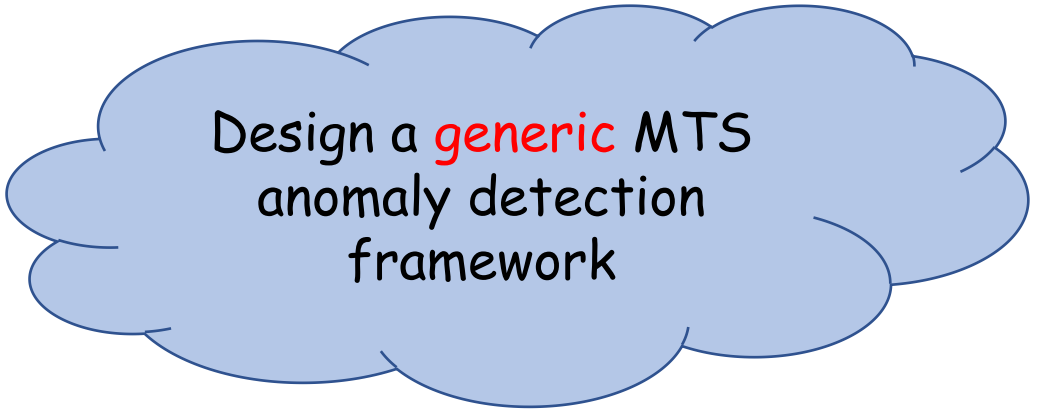


Model initialization time refers to the length of time that a model can train the required training data well

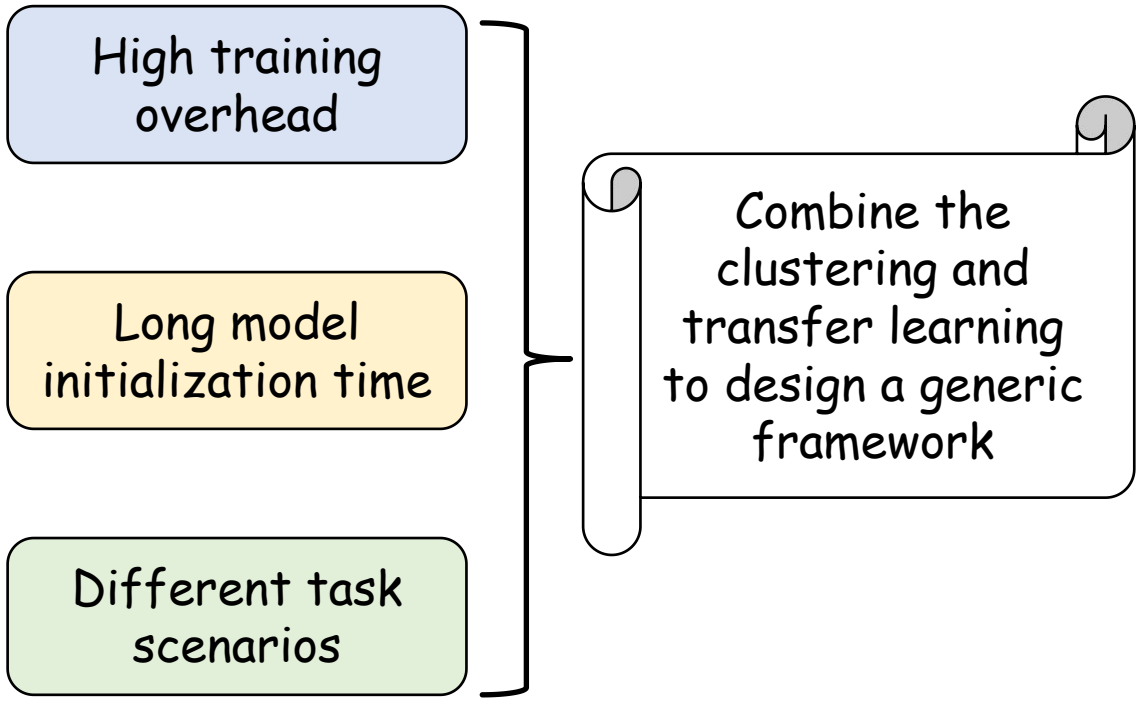
Motivations (3/3)

- Different MTS anomaly detection models are suitable for different task scenarios.

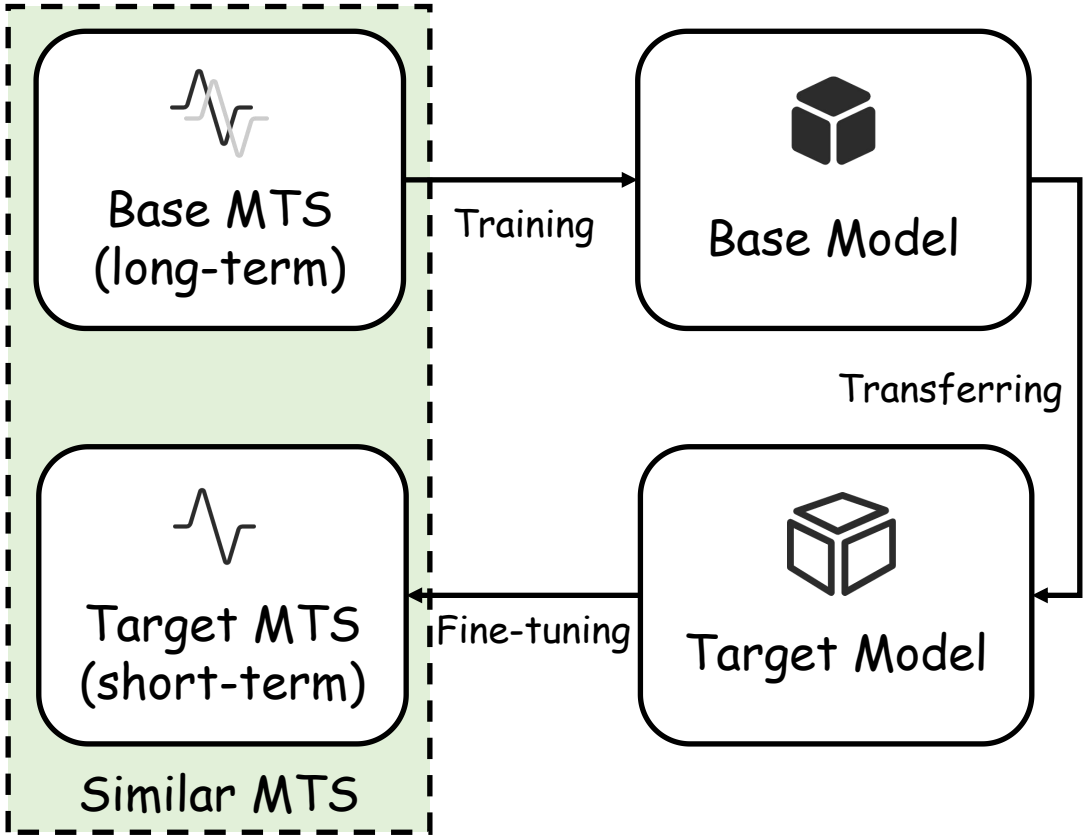
Model	Focus
OmniAnomaly(KDD '19)	Temporal dependence and stochasticity
InterFusion(KDD '21)	Inter-metric and temporal embeddings
DAGMM(ICLR '18)	Decoupling problem
USAD(KDD '20)	Stable and faster
GDN(AAAI '21)	Correlations among metrics
TranAD(VLDB '22)	Multi-modal feature extraction



Intuition



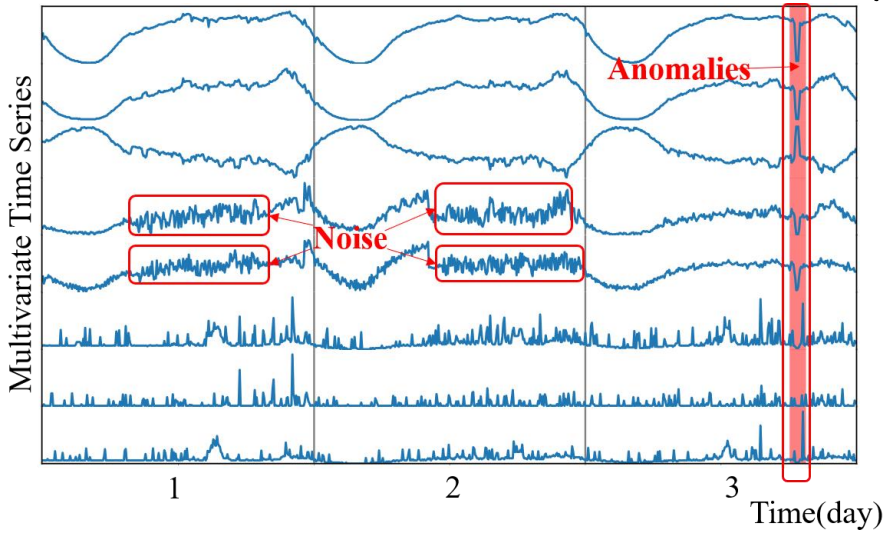
Motivation and Intuition



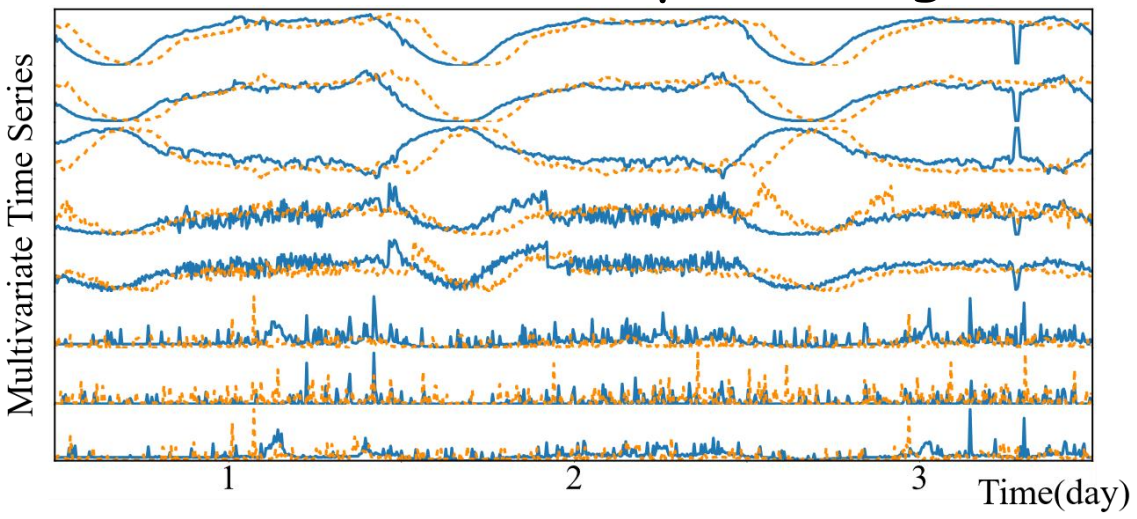
Transfer Learning

Challenges (1/2)

- **High diversity of MTS**
  - MTS can be generated by various entities with diverse patterns.
  - MTS contains irregular noises, anomalies, and extreme values.
  - MTS may have similar shapes but with phase shifts.
  - **Solution**: cluster MTS by baseline extraction and phase alignment



Anomalies and noise of MTS



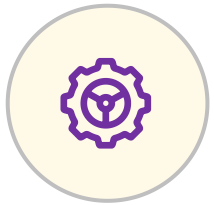
Phase shifts of MTS

Challenges (2/2)

- **Selection of transfer strategy**
  - **Various strategies** for transferring parameters from the base model to the target model.
  - **Various distances** between the base and target MTS making the optimal transfer strategy of each target MTS different.
  - The optimal transfer strategies for **different models** are diverse.
  - **Solution**: adaptive transfer strategy

Outline

Background

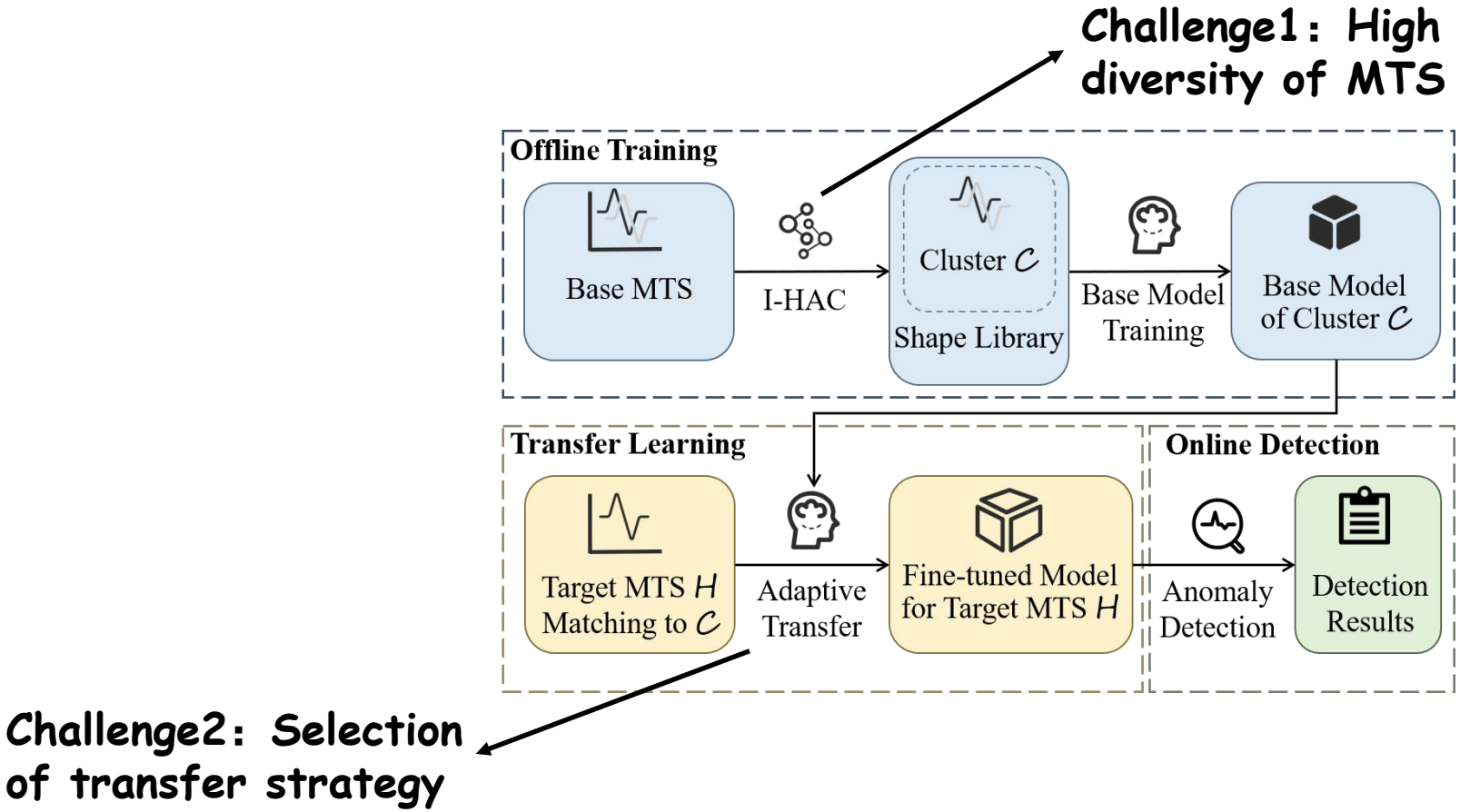


Approach

Evaluation

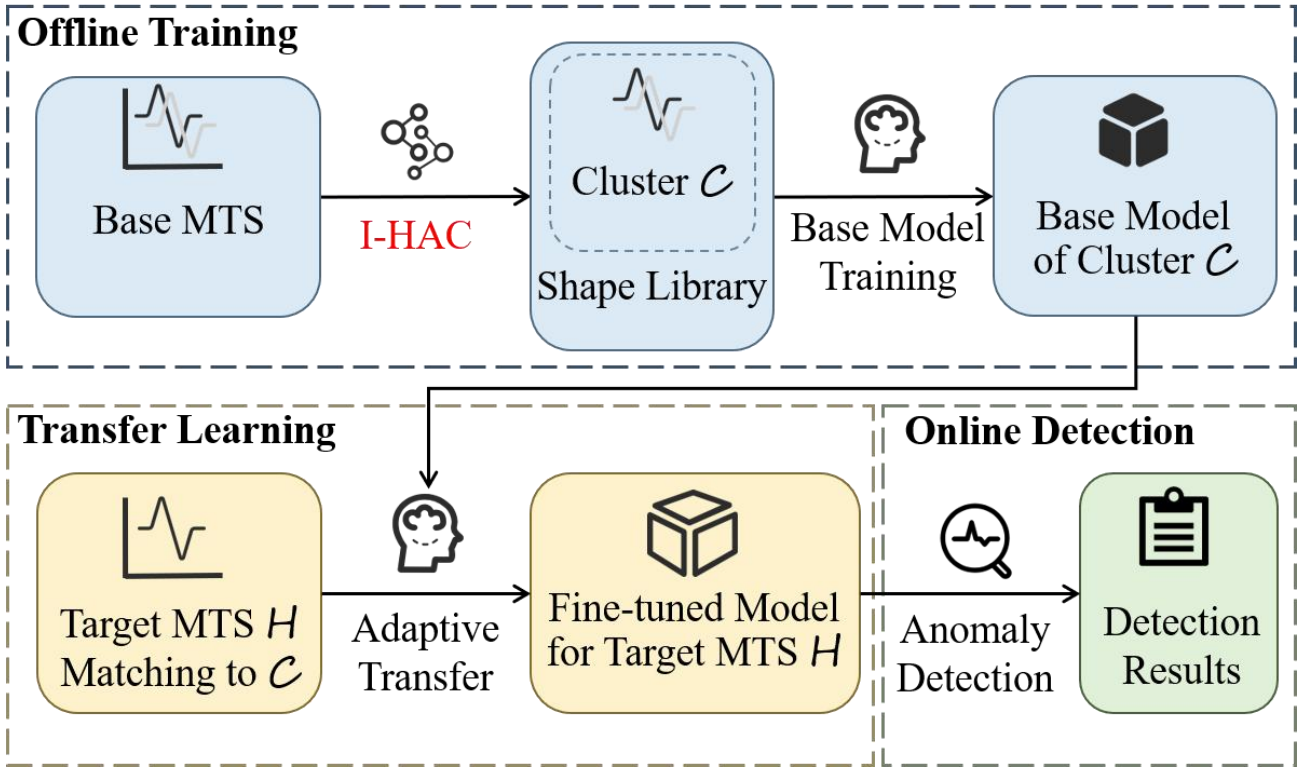
Conclusion

OmniTransfer Overview



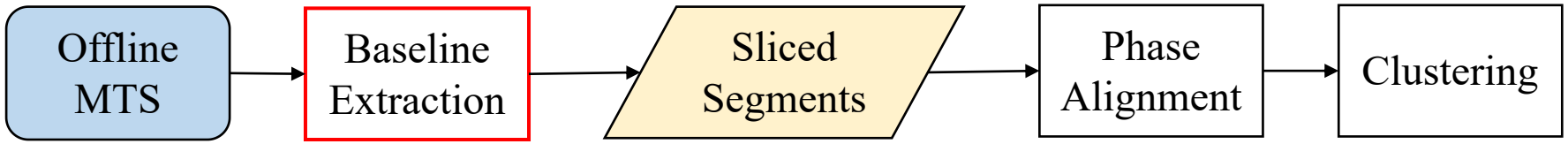
Framework of OmniTransfer

Improved Hierarchical Agglomerative Clustering



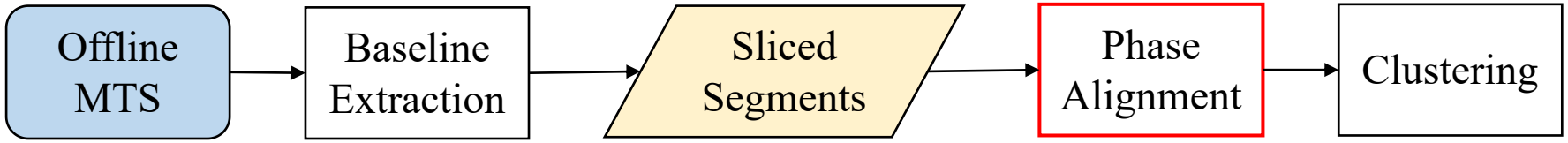
I-HAC in OmniTransfer

I-HAC(1/3)



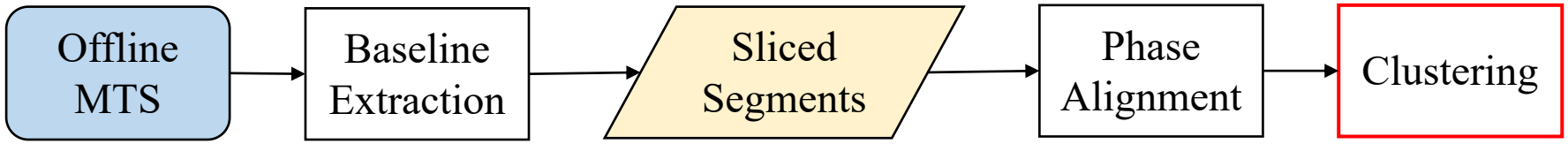
- Remove the top 5% data that deviates from the mean value
- Use linear interpolation to fill the vacancies
- Apply the moving average

I-HAC(2/3)



- Get the pivot **PVT** of the entire offline segments $PVT=\arg\min_{A\in D}\sum_{B\in D} Euc(A,B)$
- Wrap around the $MTS^{M\times n}$ by possible phase shifts $S\in[-n+1,n-1]$
- Normalized Cross-Correlation (NCC) reach the maximum value when s is close to the real phase shift
  - $NCC(A,B,s)=\sum_{j=1}^{M}\frac{\sum_{i=1}^{n}A(s)_i^j B(s)_i^j}{\left\|A(s)^j\right\|_2\left\|B(s)^j\right\|_2}$
  - $S^*=\arg\max_{s\in[-n+1,n-1]} NCC(PVT, MTS, s)$

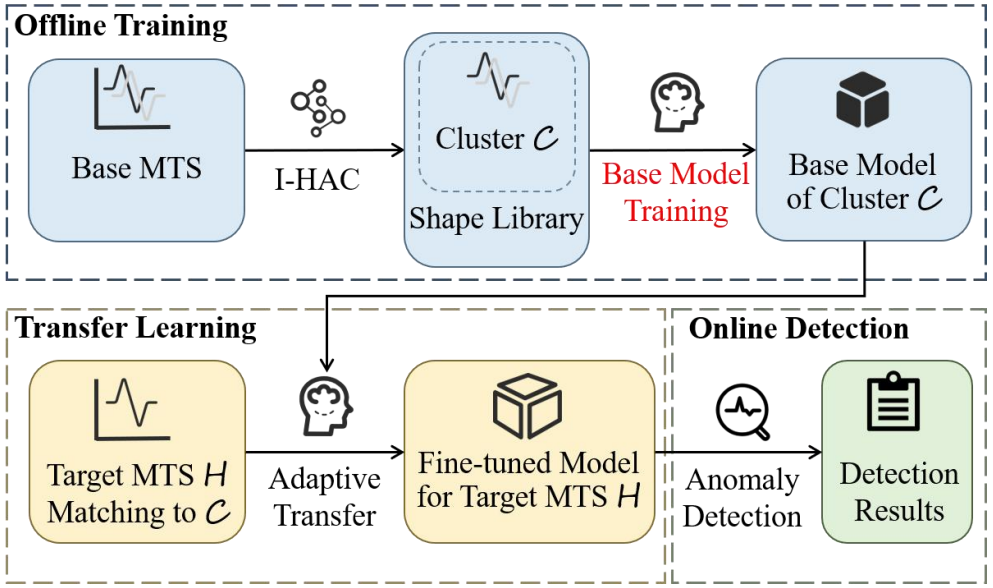
I-HAC(3/3)



- HAC with average linkage
  - Robust to extreme value
  - Each data has the same effect on the distance measure

Base Model Training

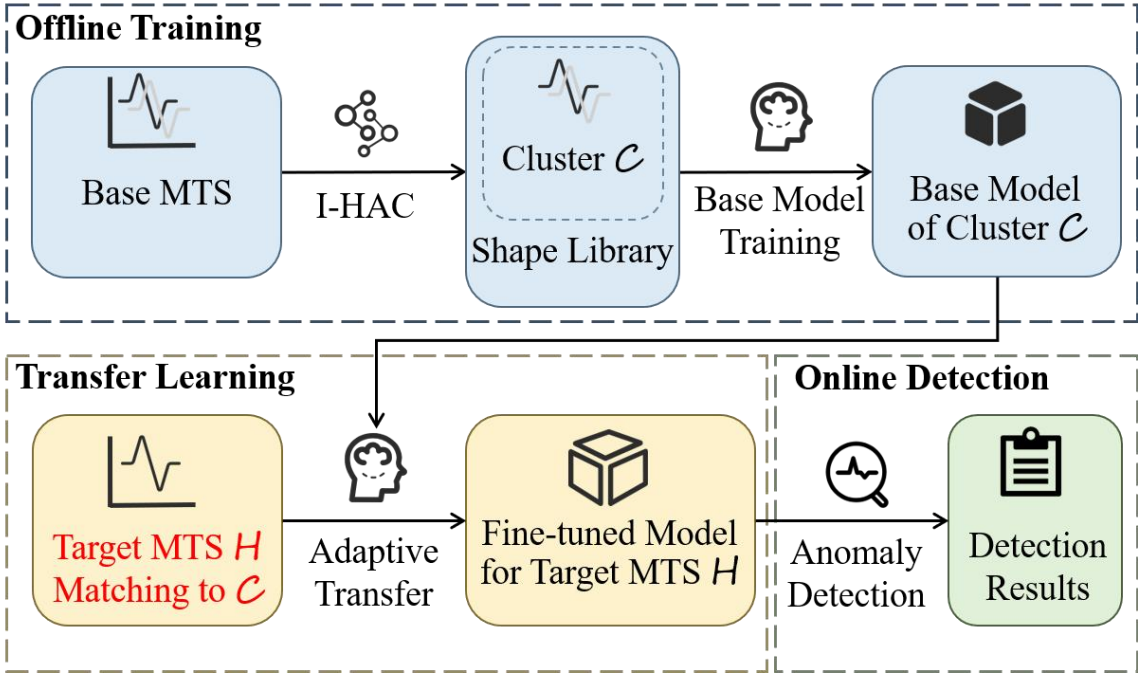
- VAE-based algorithms
  - $L_1 = E_{q_\phi(z|x)}\left[\log p_{\theta(x|z)}\right] - D_{KL}[q_\phi(z|x)||p_\theta(z)]$
- AE-based and prediction-based algorithms
  - $L_2 = MSE(target, output)$



Base model training in OmniTransfer

Transfer Learning (1/3)

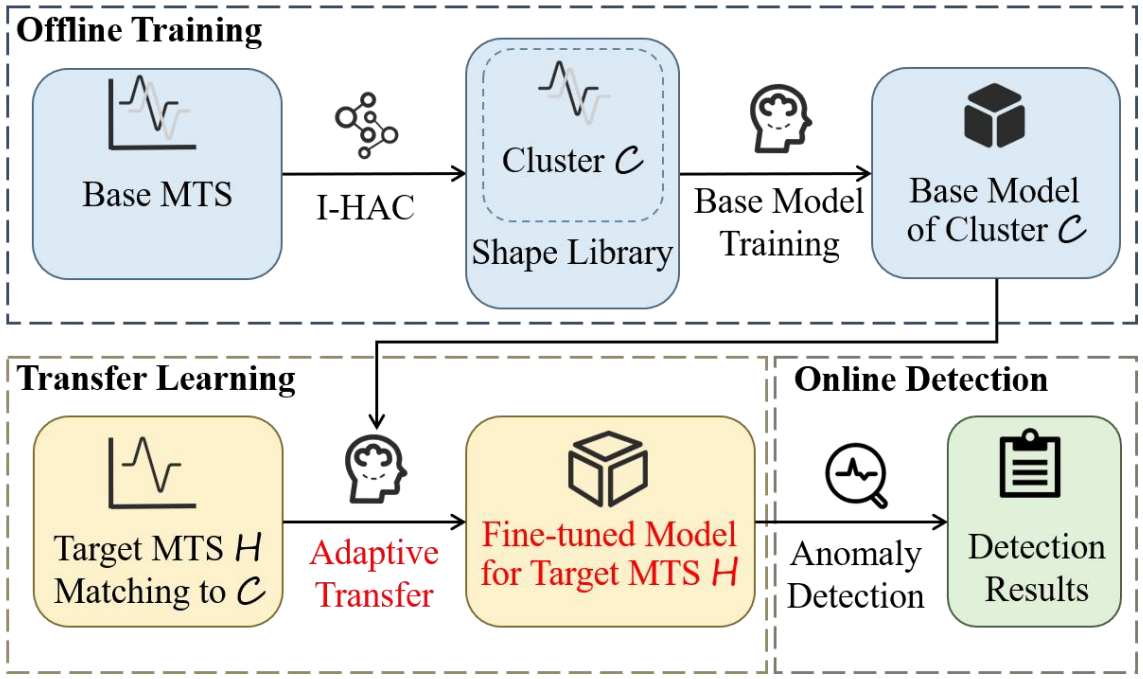
- *Transfer preparations*
  - *Target MTS **H** undergoes baseline extraction and phase alignment and get **H'***
  - *Calculate the distance between **H'** and each cluster centroid*
  - *Select the closest one and corresponding base model*
  - *Use H to fine-tune the base model*



Transfer preparations in OmniTransfer

Transfer Learning (2/3)

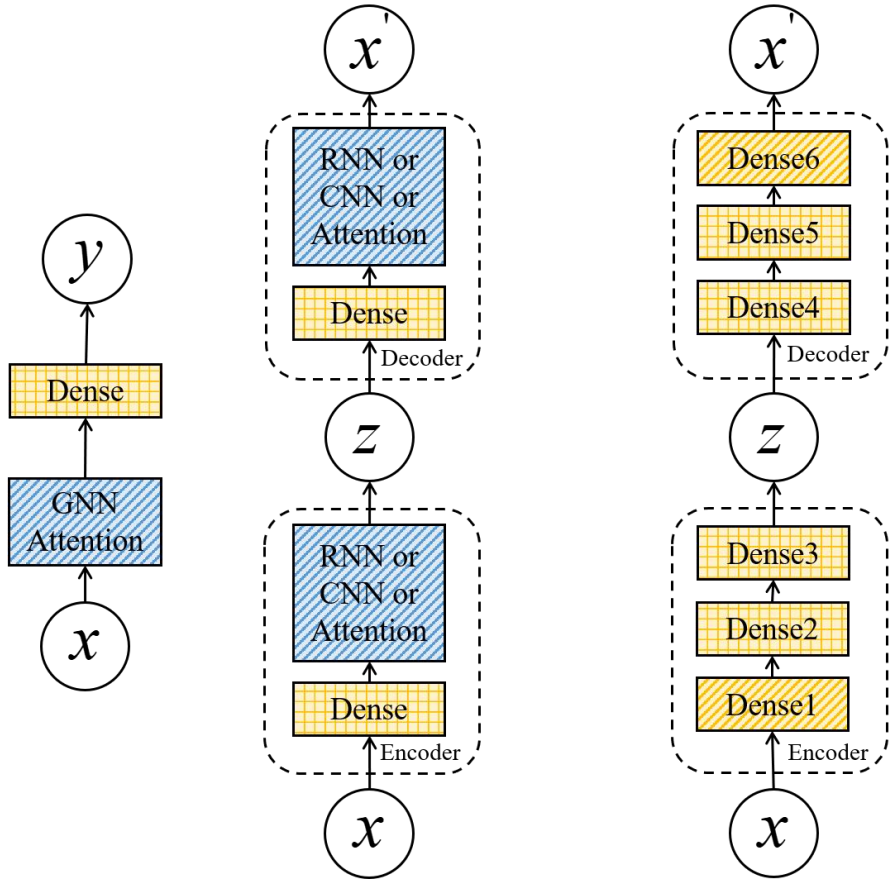
- *Adaptive transfer strategy*
  - *Use Euclidean distance to determine the degree of similarity between H′ and cluster centroid*
  - *Operators choose a threshold α empirically*
    - $\mathrm{Euc}\left(H', \mathrm{centroid}\right) \leq \alpha$, use full parameter transfer strategy
    - $\mathrm{Euc}\left(H', \mathrm{centroid}\right) > \alpha$, use partial parameter transfer strategy



Adaptive transfer in OmniTransfer

Transfer Learning (3/3)

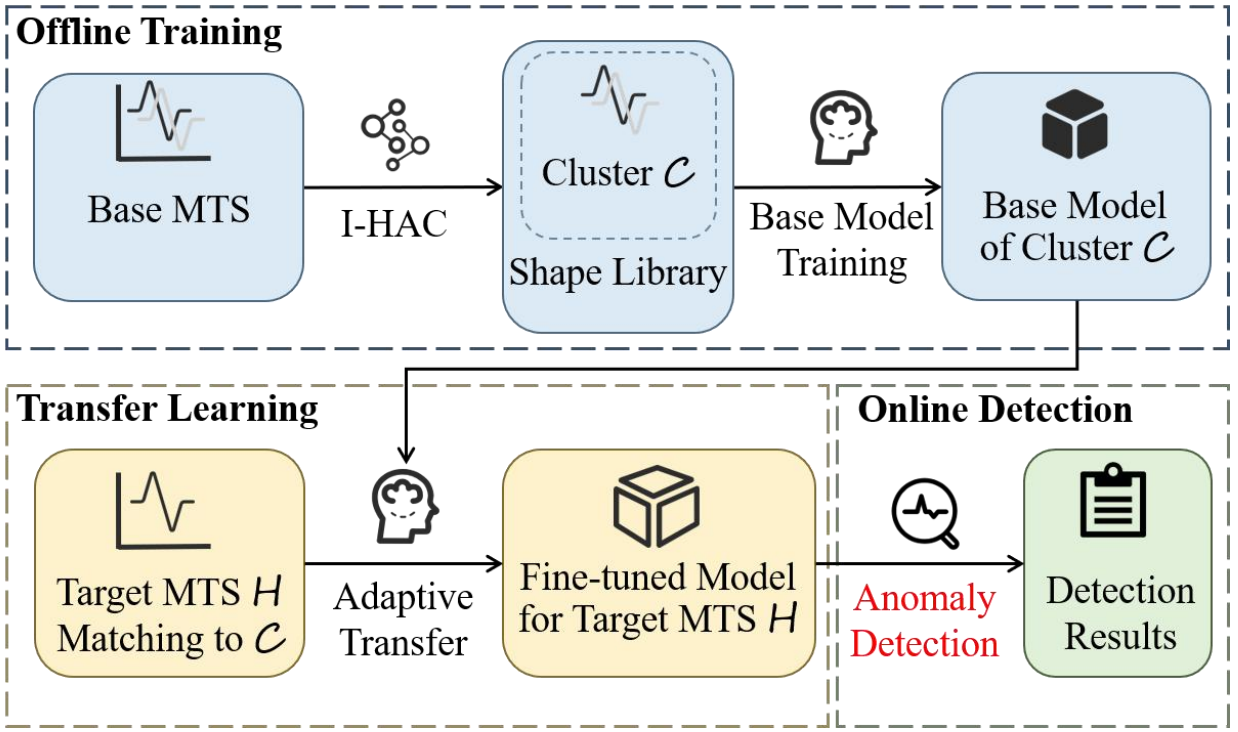
- *Transfer layer selection*
  - *Models consists of specialized layers*
    - *Specialized layers (RNN, CNN, GNN and attention): generic features*
    - *Other layers: specific tasks*
  - *Models consists of some fully connected layers*
    - *Outer layers: extensive tasks and generic features*
    - *Inner layers: task-specific features*



Architecture of MTS anomaly detection models

Online Detection

- VAE-based algorithms
  - $AS_1 = E_{q_\phi(z|x)}\left[\log p_{\theta(x|z)}\right]$
- AE-based and prediction-based algorithms
  - $AS_2 = MSE(target, output)$



Online detection in OmniTransfer

Outline

Background

Approach

Evaluation

Conclusion

Dataset & Evaluation Metric

- **Dataset:**
  - **Entities 400**
  - **Dimension of each entity 19KPIs x 2016 time points (frequency 5min, 7days)**
  - **Training 5th day, Testing last 2 days**
- **Evaluation Metric:**
  - **Micro-average F1**
  - **Model training time**
  - **Model Initialization time**

Research Questions

- RQ1. How does the effectiveness and effificiency of *OmniTransfer* compare to baseline methods?
- RQ2. How much initialization time can *OmniTransfer* reduce compared to non-transfer learning methods?
- RQ3. How much do the key techniques contribute to its overall performance?

Research Questions

- RQ1. How does the effectiveness and effificiency of *OmniTransfer* compare to baseline methods?

- RQ2. How much initialization time can *OmniTransfer* reduce compared to non-transfer learning methods?

- RQ3. How much do the key techniques contribute to its overall performance?

RQ1. OmniTransfer vs. baseline models

Model	OmniTransfer		OmniCluster (WWW '22)		One model/entity	
	F1	Time(s)	F1	Time(s)	F1	Time(s)
OmniAnomaly	**0.8865**	1212.99	0.5169	540.67	0.7000	9888.25
InterFusion	**0.8666**	1585.63	0.5830	566.56	0.4769	8884.94
DAGMM	**0.8375**	244.48	0.7104	137.37	0.8245	2947.47
USAD	**0.8222**	80.16	0.7468	109.04	0.7875	691.77
GDN	**0.8026**	54.55	0.6806	42.81	0.7405	265.27
TranAD	**0.8995**	114.53	0.7797	102.10	0.8538	591.67

Model	F1	Time(s)
JumpStarter (USENIX '21)	0.4211	4786.67
CTF (INFOCOM'21)	0.8661	4965.61

- **OmniCluster**: a model-agnostic clustering framework
- **One model/entity**: trains a model for each entity
- **JumpStarter**: uses the Compressed Sensing to reduce the model initialization time
- **CTF**: a framework for RNN+VAE models

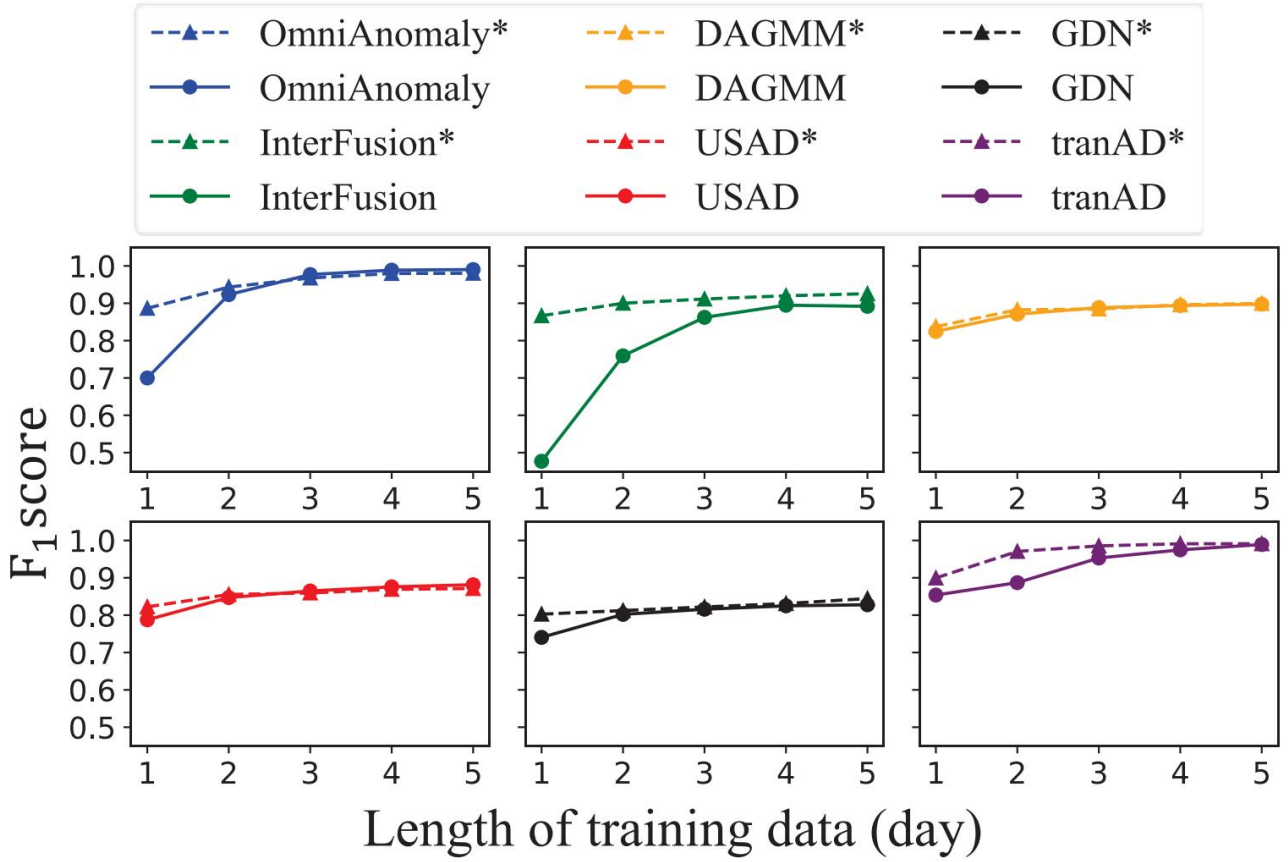
Research Questions

- RQ1. How does the effectiveness and effificiency of *OmniTransfer* compare to baseline methods?

- RQ2. How much initialization time can *OmniTransfer* reduce compared to non-transfer learning methods?

- RQ3. How much do the key techniques contribute to its overall performance?

RQ2. Effect on reducing model initialization time



*: OmniTransfer ;w/o *: one model/entity
OmniTransfer(2 days) ≈ one model/entity(5 days)

Research Questions

- RQ1. How does the effectiveness and effificiency of *OmniTransfer* compare to baseline methods?
- RQ2. How much initialization time can *OmniTransfer* reduce compared to non-transfer learning methods?
- RQ3. How much do the key techniques contribute to its overall performance?

RQ3. Ablation experiment

- Key technologies: clustering, phase alignment, transfer learning
- C1: Only one base model is used for transfer learning, and the data used to train the base model are randomly selected.
- C2: Do not align the phase shift.
- C3: The base model is directly used for anomaly detection of all MTS in the cluster.

Model	OmniTransfer	C1	C2	C3
OmniAnomaly	**0.8865**	0.6925	0.7979	0.7242
InterFusion	**0.8666**	0.6560	0.7668	0.7319
DAGMM	**0.8375**	0.7966	0.8071	0.7804
USAD	**0.8222**	0.7754	0.7928	0.8008
GDN	**0.8026**	0.7702	0.7647	0.7643
TranAD	**0.8995**	0.8805	0.8884	0.8436

Outline

Background

Approach

Evaluation

Conclusion

Conclusion

- The first general MTS anomaly detection framework using clustering and transfer learning techniques.
- Propose an adaptive transfer strategy. It can automatically select either full parameter transfer or partial parameter transfer strategy.
- Reduce the initialization time by 59.72% and the training overhead by 85.01% on average while maintaining high accuracy in detecting anomalies.

Thank you!

Q & A

minghanliang@nankai.edu.cn

ICWS 2023

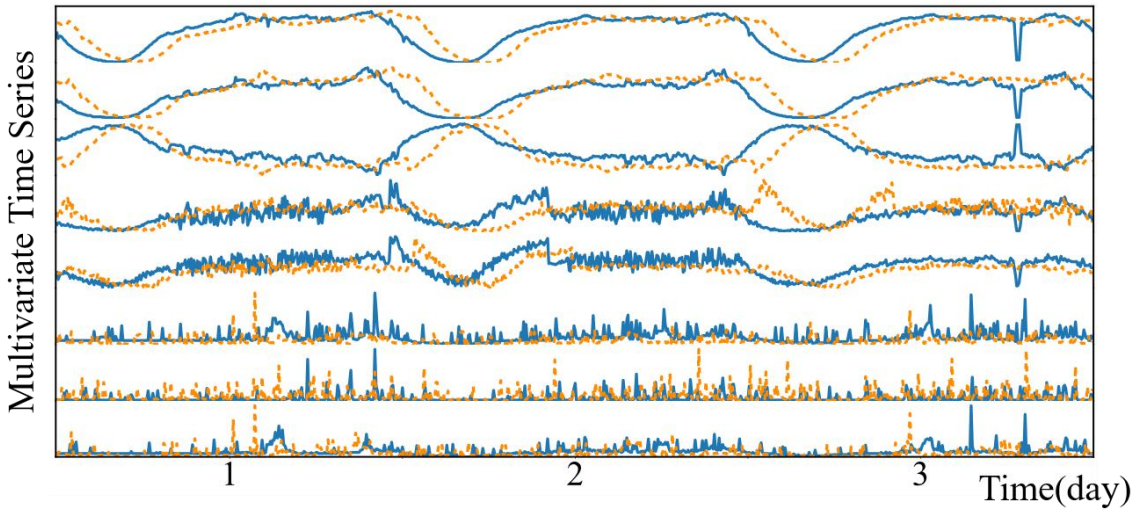
南開大學
Nankai University

Improvements

- Reduce the initialization time by 59.72% and the training overhead by 85.01% on average while maintaining high accuracy in detecting anomalies.
- Increase the universality of the anomaly detection framework.
- Use an adaptive transfer strategy to select optimal transfer strategy.

Phase shifts reason

- This can happen when large-scale Web services use different servers to serve users across a wide geographical area, resulting in similar MTS patterns with a time delay.
- The diversity can affect the distance calculation of MTS and lead to poor clustering performance.



Adaptive transfer strategy threshold

- The threshold of the adaptive Transfer strategy is based on the experience of the operator
- In view of the length of the paper, we did not consider the impact of this threshold on the adaptive Transfer strategy in detail. This is a good research issue in the future

Dataset size

- Please note that we only choose 400 entities from millions for evaluation since the labeling work is time-consuming
- We believe our framework can be applied to datasets with millions of entities