Published as a conference paper at ICLR 2022

ANOMALY TRANSFORMER: TIME SERIES ANOMALY DETECTION WITH ASSOCIATION DISCREPANCY

Jiehui Xu,* Haixu Wu,* Jianmin Wang, Mingsheng Long (✉)
School of Software, BNRist, Tsinghua University, China
{xjh20,whx20}@mails.tsinghua.edu.cn, {jimwang,mingsheng}@tsinghua.edu.cn


Jiehui Xu*


Haixu Wu*


Jianmin Wang

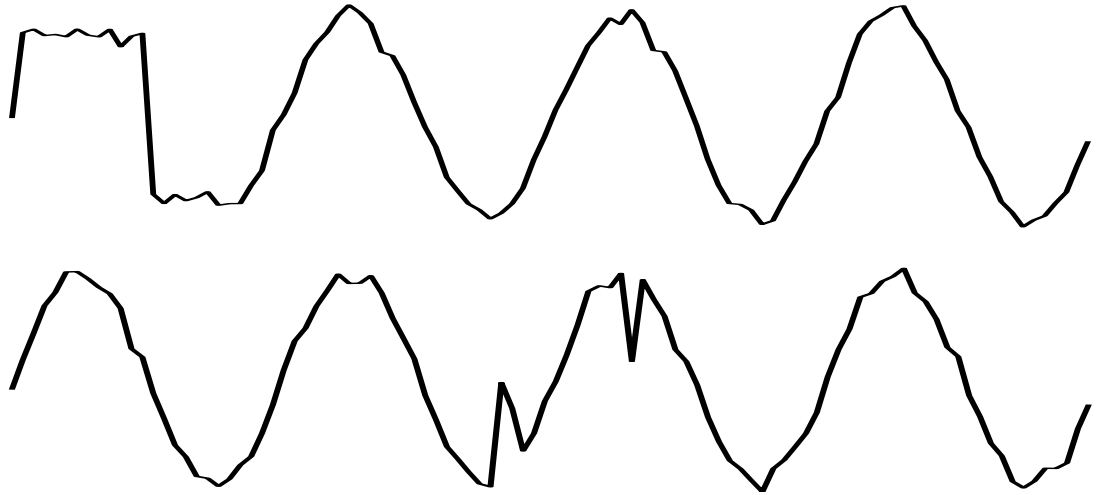

Mingsheng Long

Time Series Anomaly Detection

Service Monitoring

Space & earth Exploration

Water Treatment



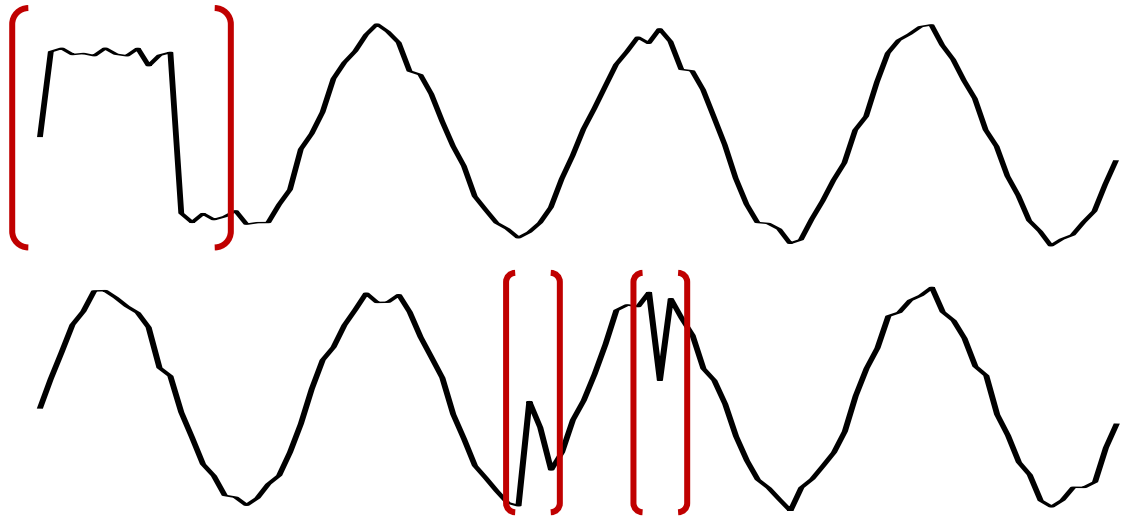
Real-world systems always work continuously and generate successive measurements.

Time Series Anomaly Detection

Service Monitoring

Space & earth Exploration

Water Treatment



Discovering the malfunctions to ensure security and avoid financial loss

Detecting the abnormal time points from time series

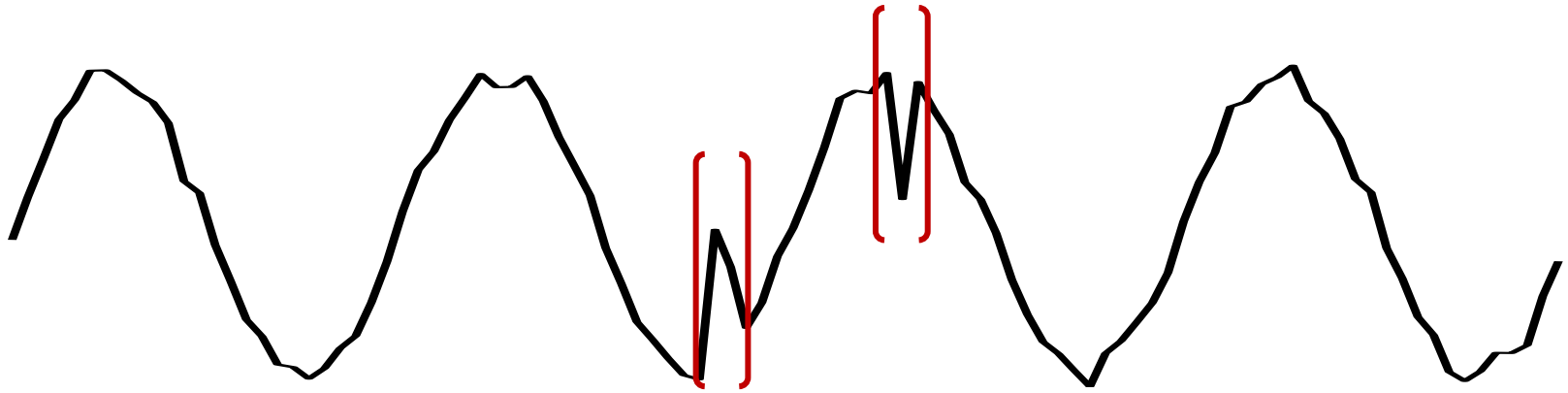
Unsupervised Time Series Anomaly Detection

Anomalies are usually **rare** and **hidden by vast normal time points**, making labeling hard and expensive.

Unsupervised tasks

Time series ⇨ **Informative** representation for each time point ⇨ **Distinguishable** anomaly criterion

Related Work



(1) Classic methods
(e.g. LOF、OC-SVM、SVDD)

- Do not consider the temporal information in time series.
- Are hard to generalize to unseen real scenarios.

(2) Recurrent networks self-supervised by reconstruction and autoregressive tasks

- Point-wise representation is **less informative** and can be dominated by normal points.
- Reconstruction or predictive error is point by point without comprehensive description.

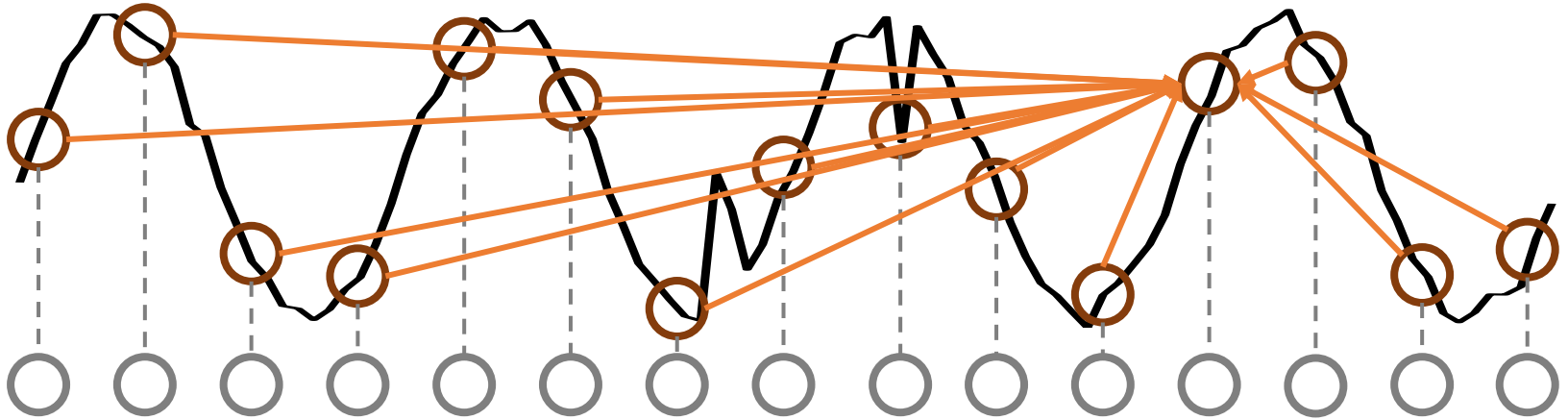
Related Work

(3) Explicit association learning

(e.g. vector autoregression, state space models)

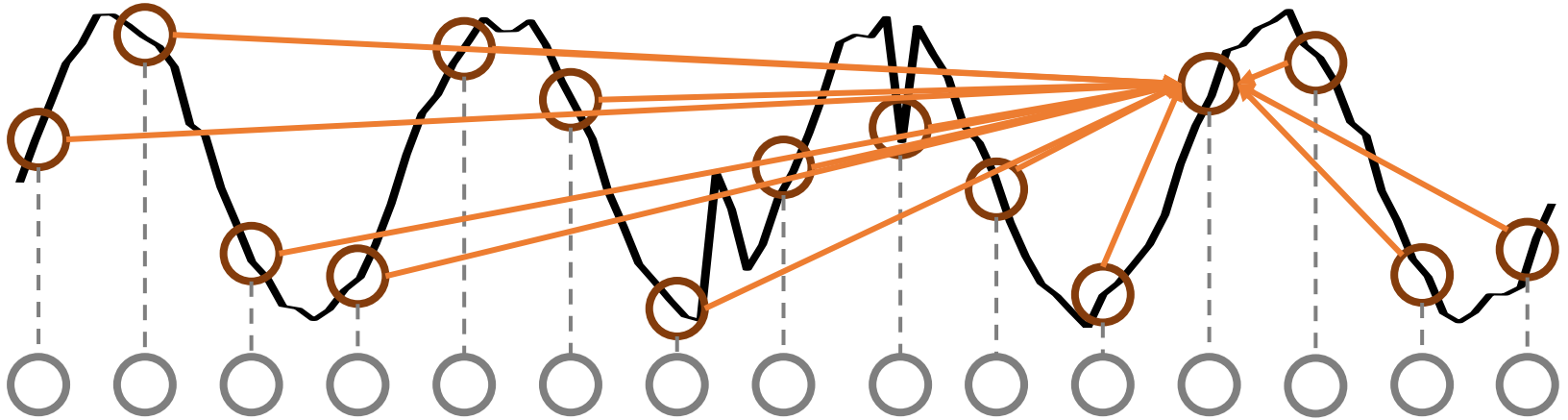
- GNN for multivariate time series -> limited to single time point and insufficient for complex temporal patterns.
- Subsequence-based similarity calculation -> cannot capture fine-grained associations for each time point.

Temporal Association



Temporal Association: a distribution of association weights
to all the time points along the temporal dimension.

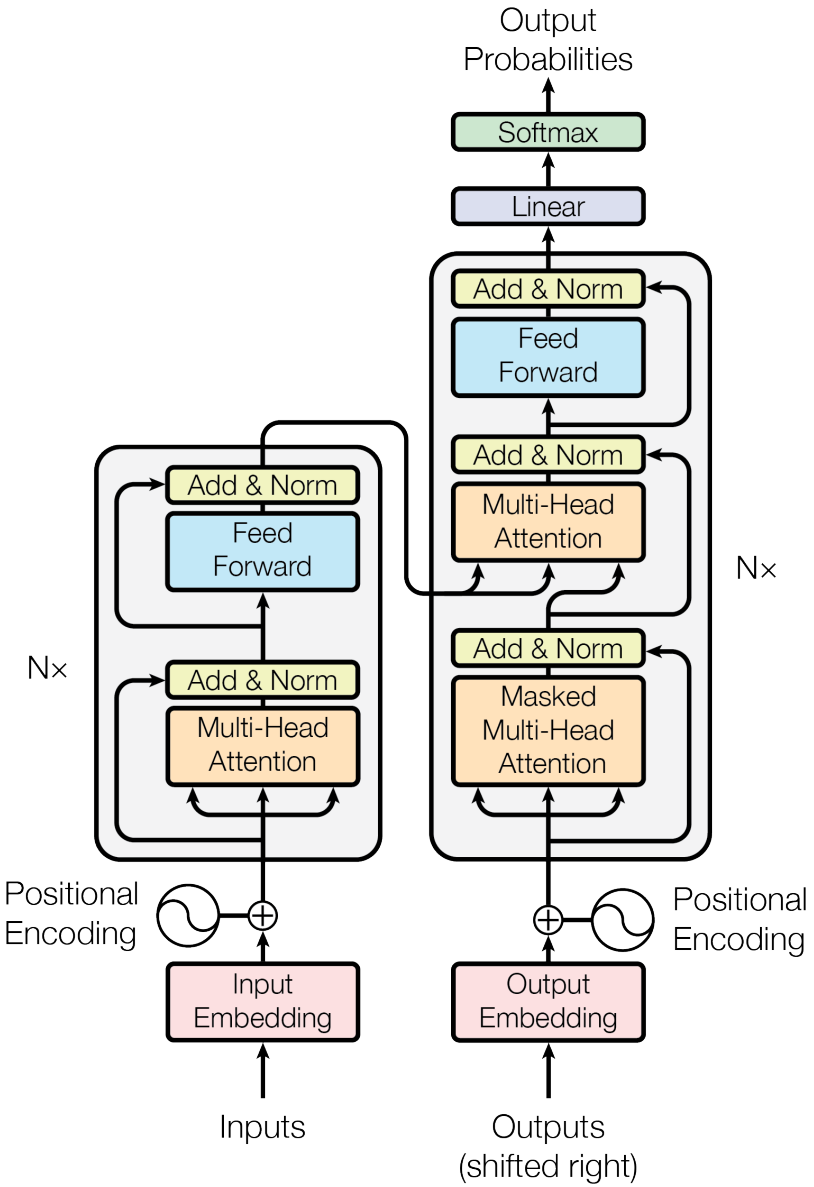
Temporal Association

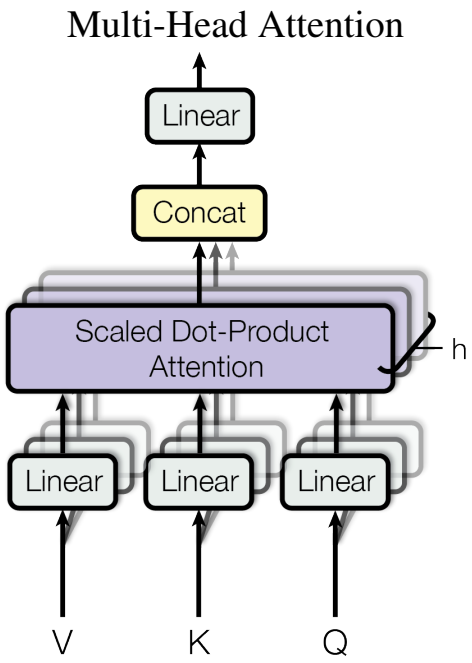


Temporal Association: a distribution of association weights
to all the time points along the temporal dimension.

More **Informative** for the temporal context,
indicating temporal patterns, such as the period or trend of time series.

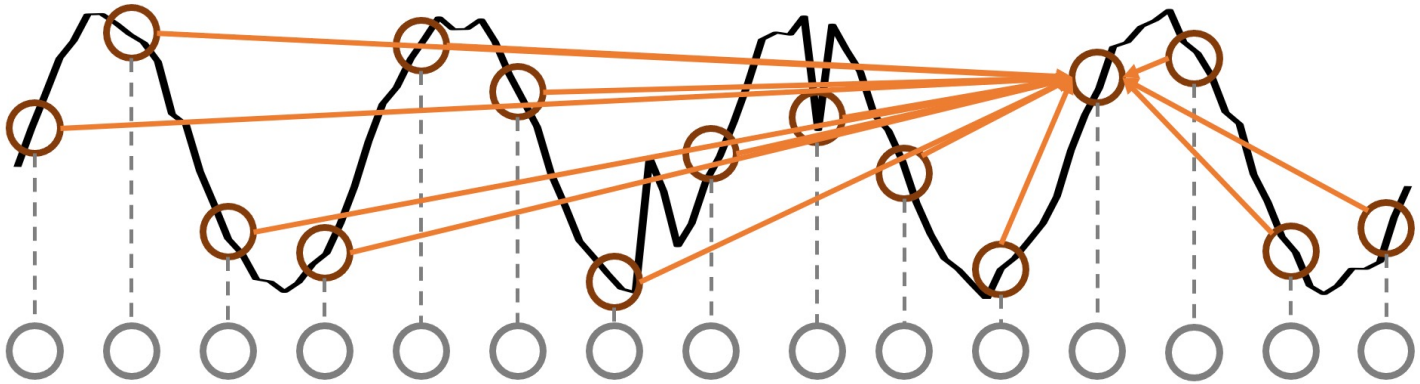
Transformer for Series-Association



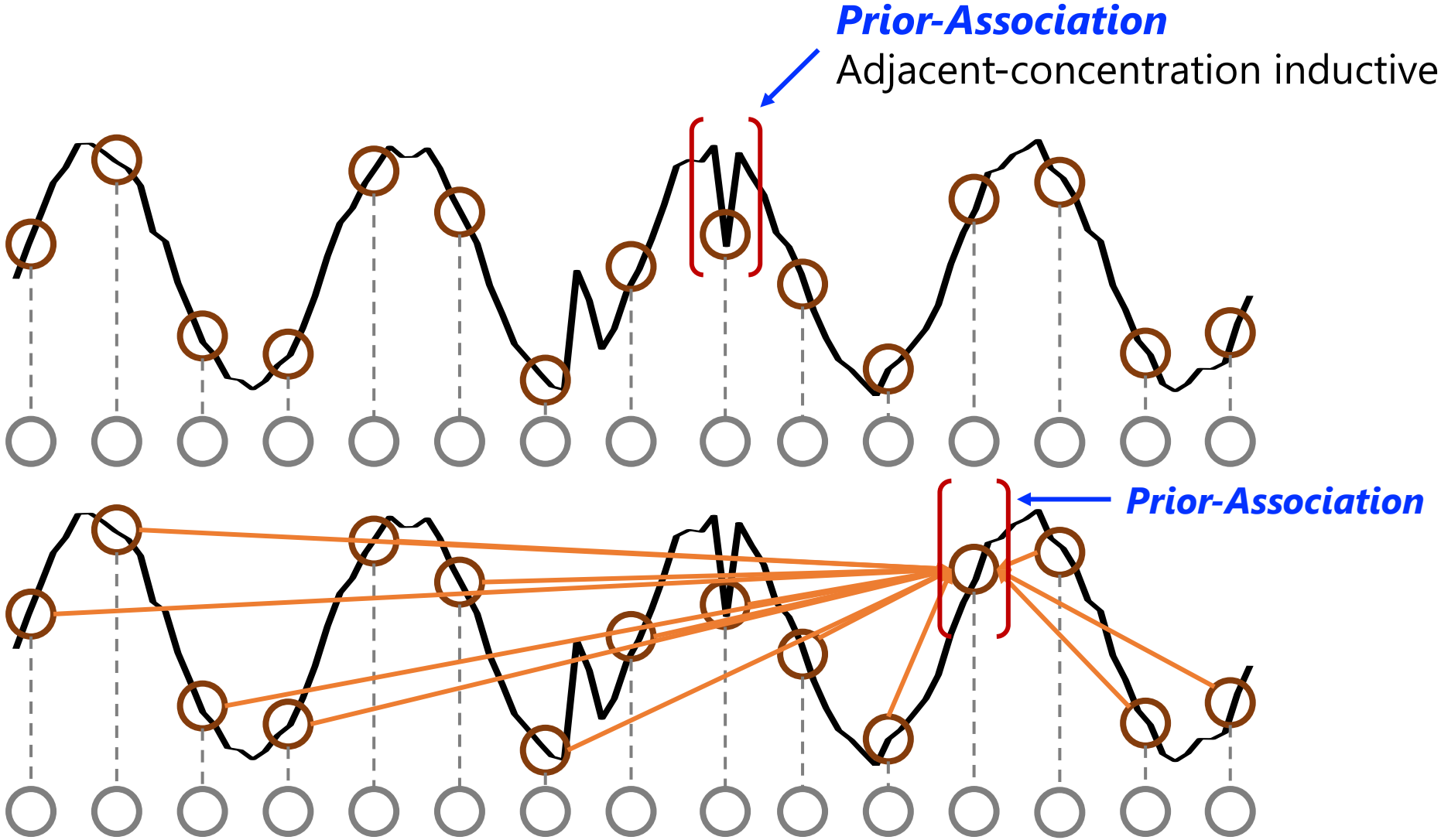


Series-Association

learned from raw time series

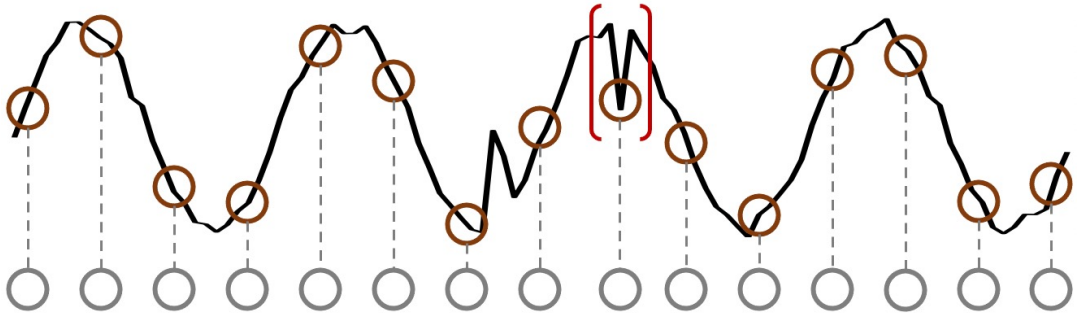


Adjacent-concentration for Prior-Association



Association Discrepancy

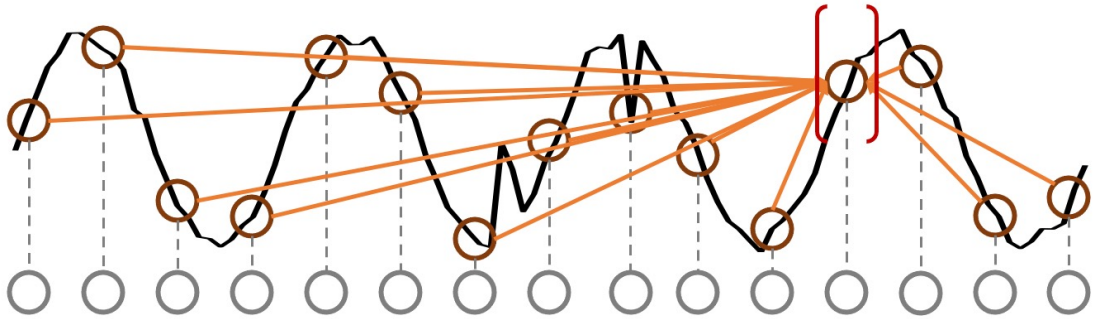
Abnormal
Time Point



Prior-Association

Series-Association

Normal
Time Point

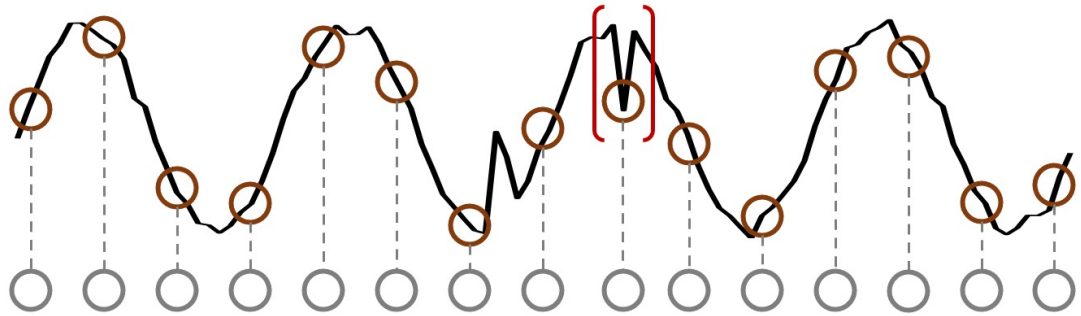


Prior-Association

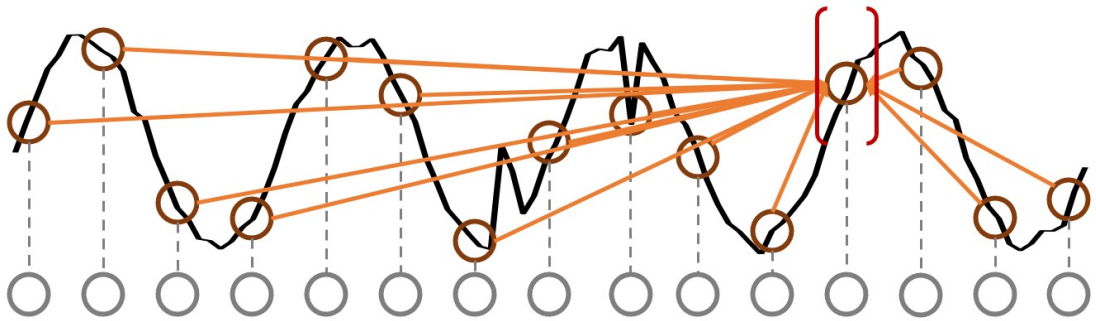
Series-Association

Association Discrepancy

Abnormal
Time Point



Normal
Time Point



Prior-Association

Series-Association

Small
Association
Discrepancy

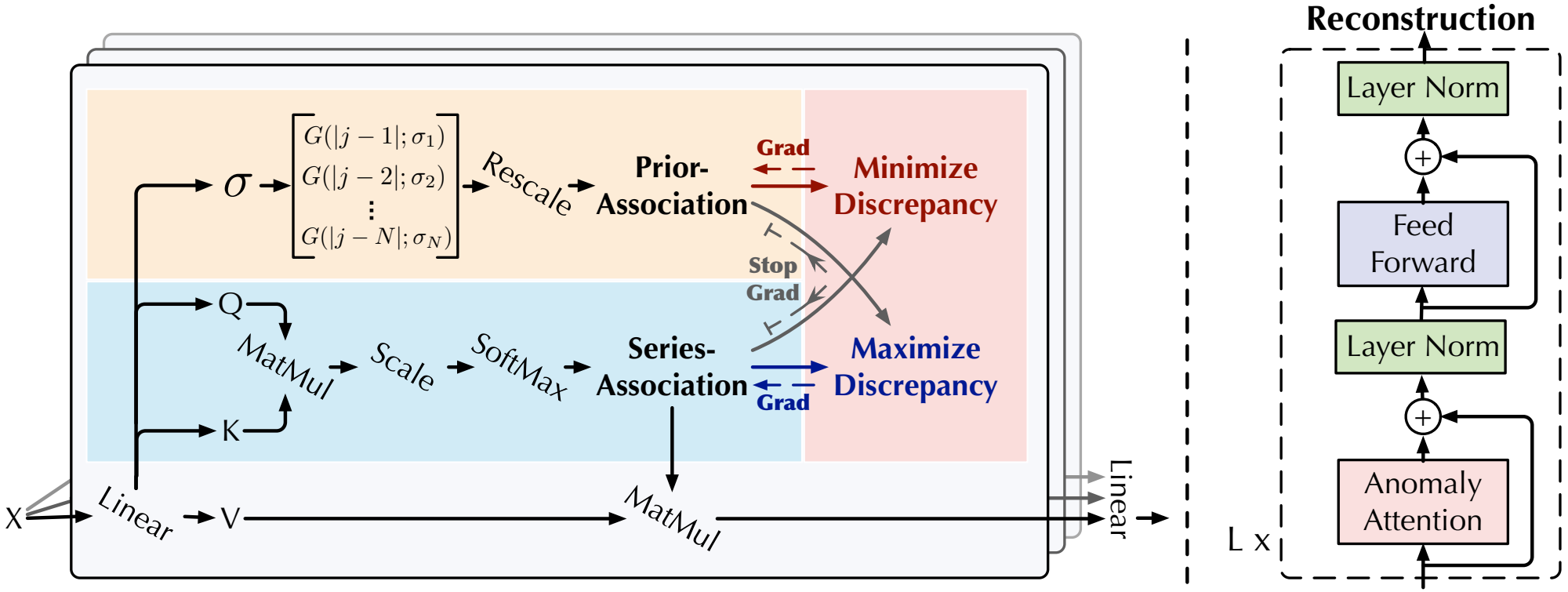
Inherent distinguishable criterion

Prior-Association

Series-Association

Large
Association
Discrepancy

Anomaly Transformer

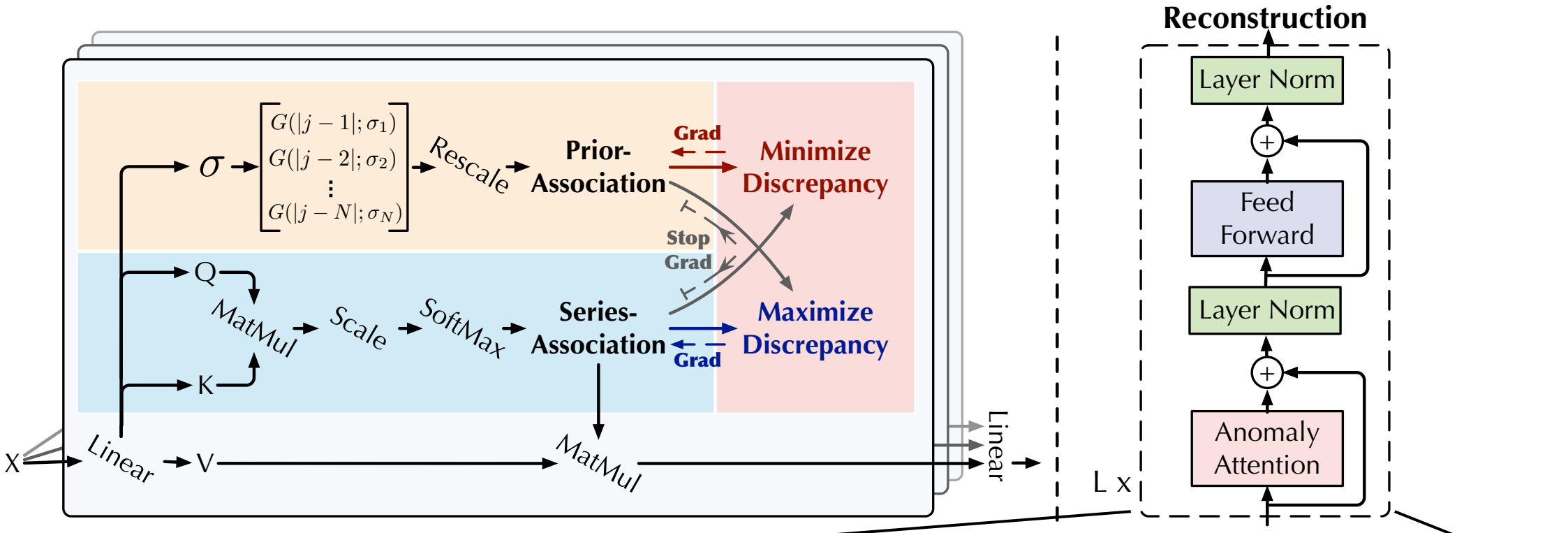


(1) Architecture: Anomaly Transformer with Anomaly-Attention

(2) Training Strategy: Minimax Association Learning

(3) Criterion: Association-based Anomaly Criterion

Overall Architecture

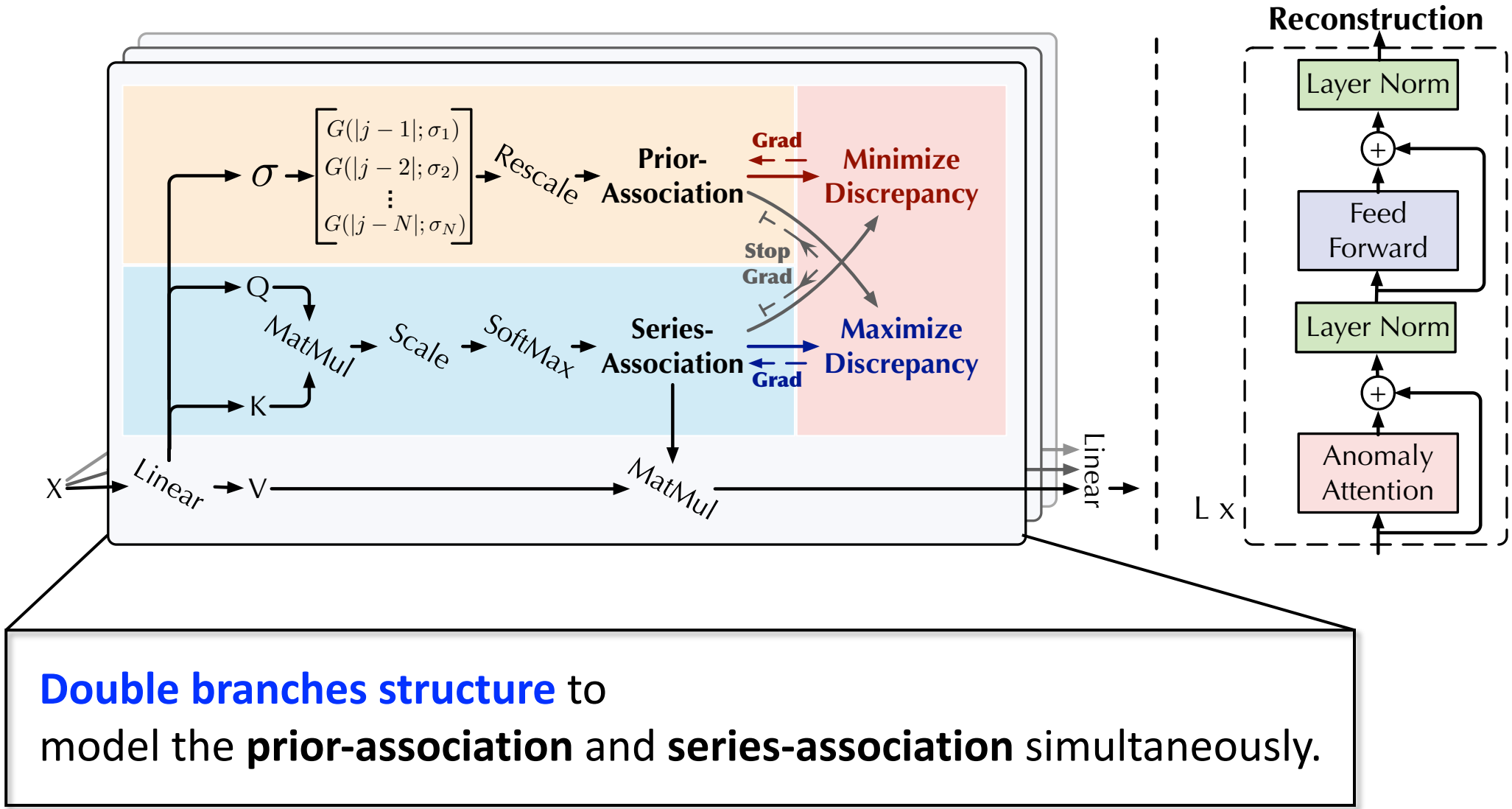


$$\mathcal{Z}^l = \text{Layer-Norm}\Big(\text{Anomaly-Attention}(\mathcal{X}^{l-1}) + \mathcal{X}^{l-1}\Big)$$

$$\mathcal{X}^l = \text{Layer-Norm}\Big(\text{Feed-Forward}(\mathcal{Z}^l) + \mathcal{Z}^l\Big),$$

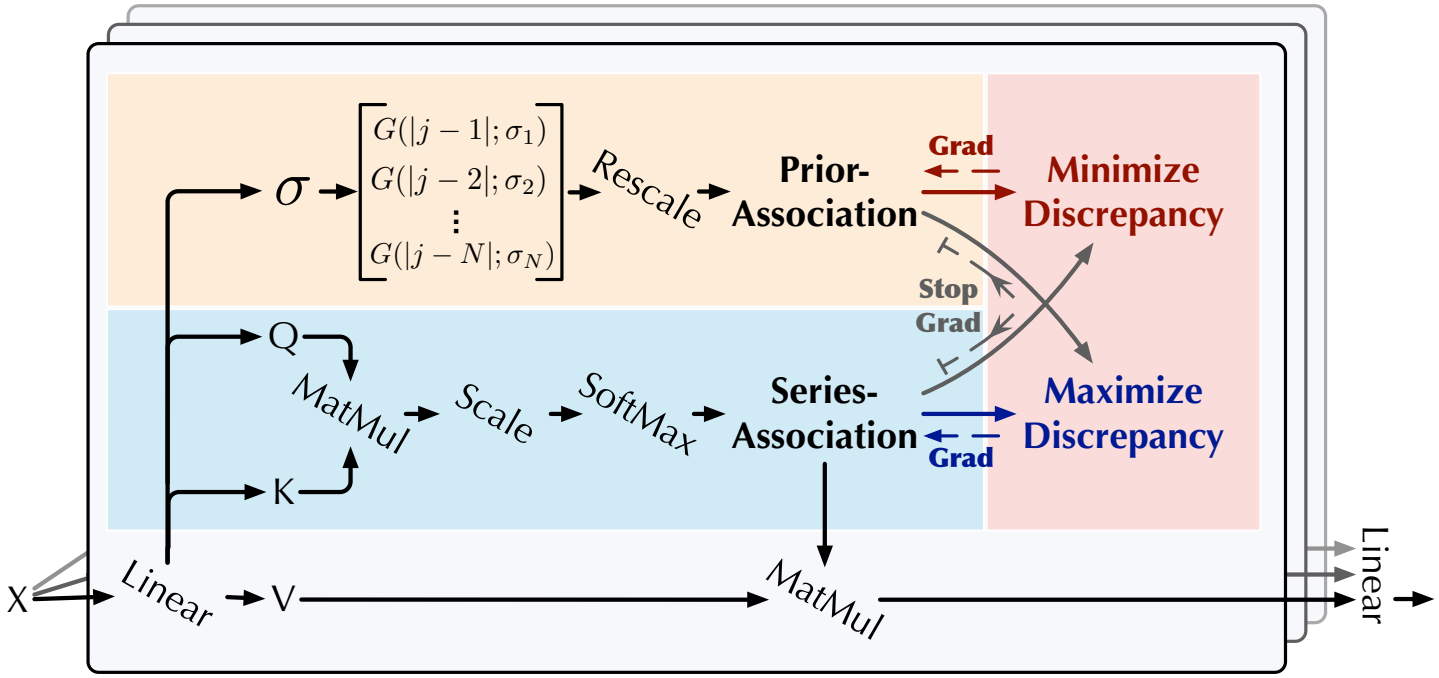
Learning underlying associations from deep **multi-level** features.

Overall Architecture



Double branches structure to
model the **prior-association** and **series-association** simultaneously.

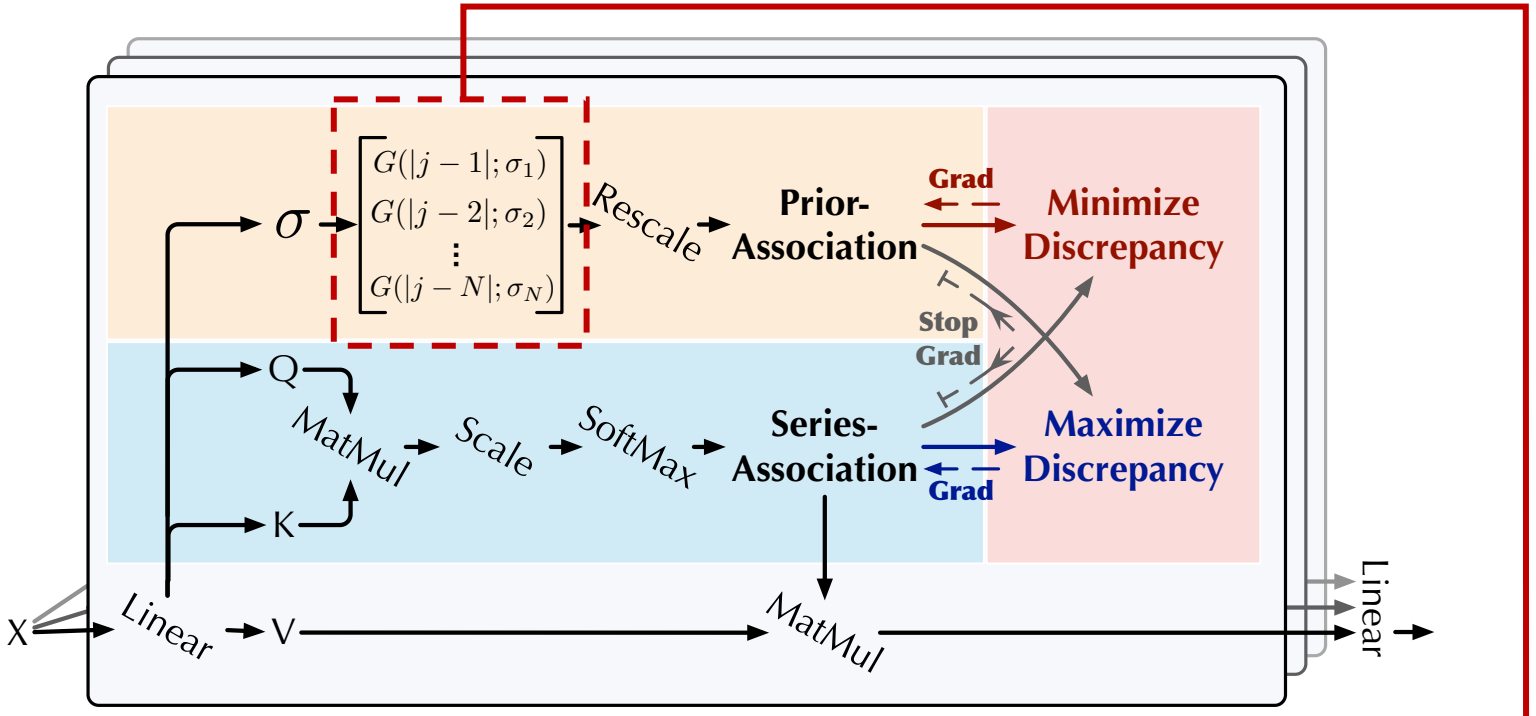
Anomaly-Attention



$$\text{Prior-Association: } \mathcal{P}^l = \text{Rescale}\left(\left[\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)\right]_{i,j\in\{1,\cdots,N\}}\right)$$

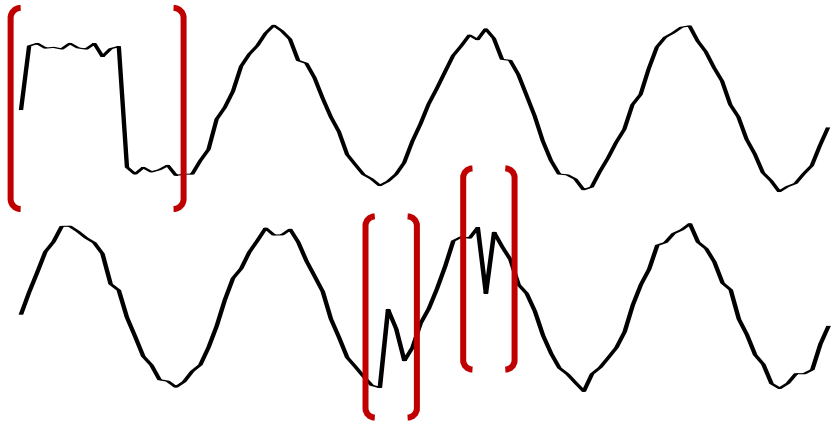
$$\text{Series-Association: } \mathcal{S}^l = \text{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^{\mathrm{T}}}{\sqrt{d_{\text{model}}}}\right)$$

Anomaly-Attention



learnable Gaussian kernel

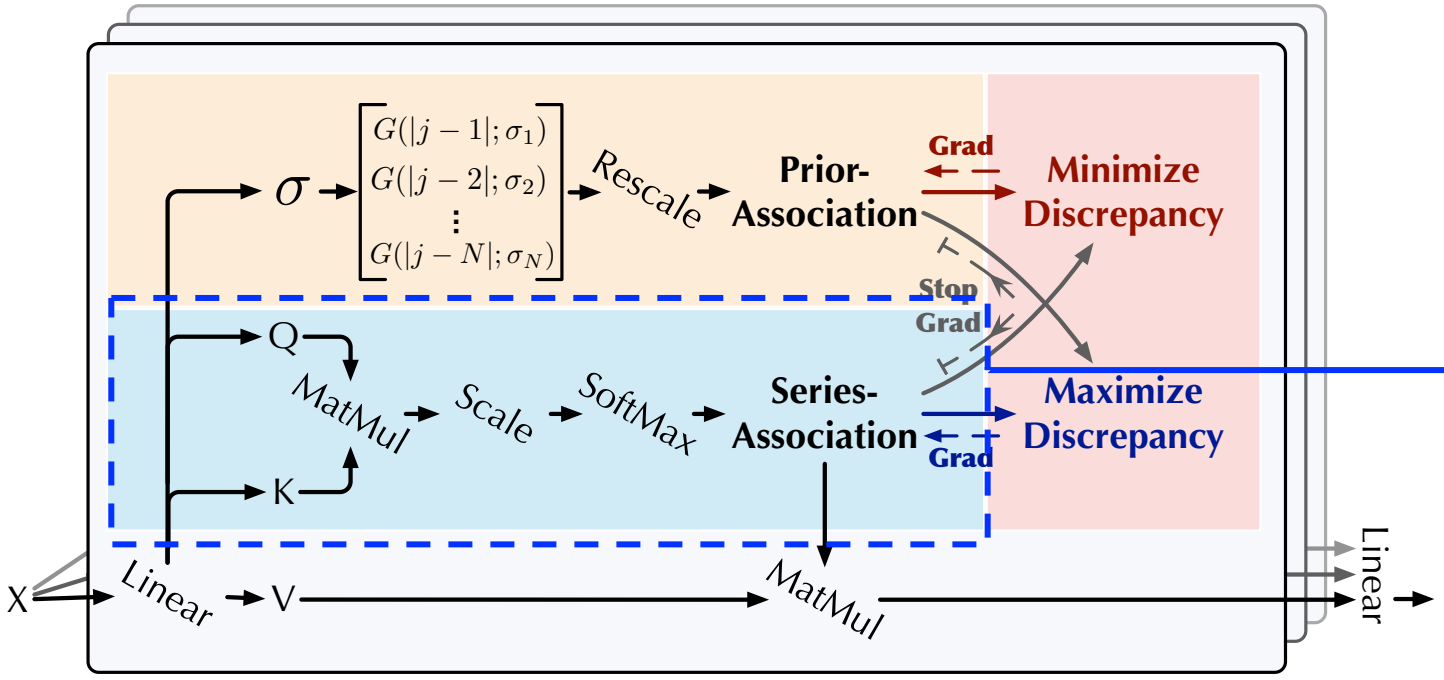
making prior-associations adapt to the **various time series patterns**



$$\text{Prior-Association: } \mathcal{P}^l = \text{Rescale}\left(\left[\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)\right]_{i,j\in\{1,\cdots,N\}}\right)$$

$$\text{Series-Association: } \mathcal{S}^l = \text{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^{\mathrm{T}}}{\sqrt{d_{\text{model}}}}\right)$$

Anomaly-Attention



Self-Attention

Find the most effective associations from raw series

$$\text{Prior-Association: } \mathcal{P}^l = \text{Rescale}\left(\left[\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)\right]_{i,j\in\{1,\cdots,N\}}\right)$$

$$\text{Series-Association: } \mathcal{S}^l = \text{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^{\mathrm{T}}}{\sqrt{d_{\text{model}}}}\right)$$

Association Discrepancy

$$\text{Prior-Association: } \mathcal{P}^l = \text{Rescale}\left(\left[\frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{|j-i|^2}{2\sigma_i^2} \right) \right]_{i,j \in \{1,\cdots,N\}} \right)$$

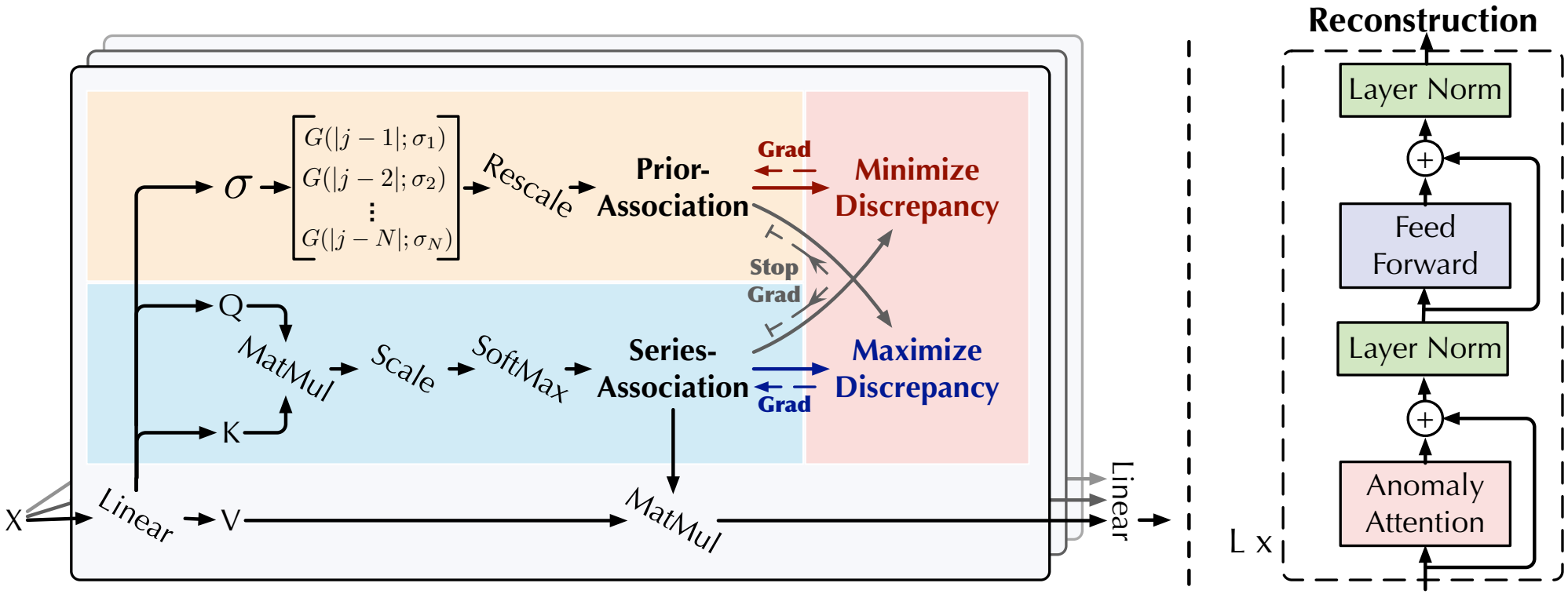
$$\text{Series-Association: } \mathcal{S}^l = \text{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^{\mathrm{T}}}{\sqrt{d_{\text{model}}}} \right)$$

$$\text{AssDis}(\mathcal{P}, \mathcal{S}; \mathcal{X}) = \left[\frac{1}{L} \sum_{l=1}^{L} \left(\text{KL}(\mathcal{P}_{i,:}^l \| \mathcal{S}_{i,:}^l) + \text{KL}(\mathcal{S}_{i,:}^l \| \mathcal{P}_{i,:}^l) \right) \right]_{i=1,\cdots,N}$$

Symmetrized KL divergence between **multi-level** prior- and series- associations

(The adjacent-concentration property of series-association)

Training Strategy (Vanilla Version)



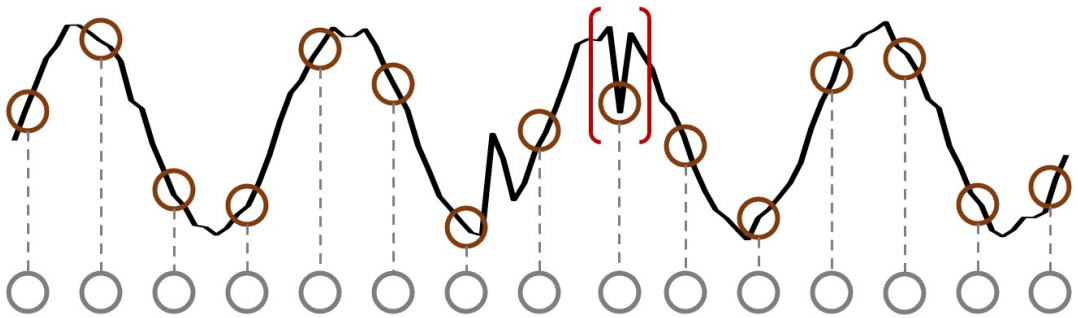
$$\mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}}, \mathcal{P}, \mathcal{S}, \lambda; \mathcal{X}) = \|\mathcal{X} - \widehat{\mathcal{X}}\|_{\text{F}}^2 - \lambda \times \|\text{AssDis}(\mathcal{P}, \mathcal{S}; \mathcal{X})\|_1$$

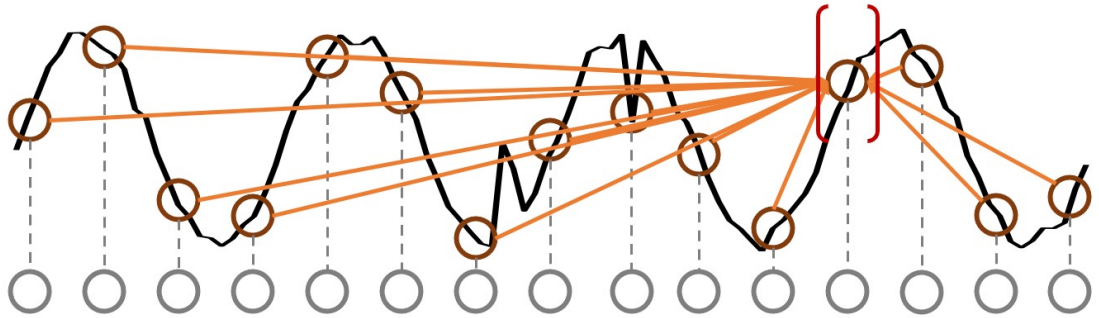
Representation Learning

Training Strategy (Vanilla Version)

$$\mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}}, \mathcal{P}, \mathcal{S}, \lambda; \mathcal{X}) = \|\mathcal{X} - \widehat{\mathcal{X}}\|_{\text{F}}^2 - \lambda \times \|\text{AssDis}(\mathcal{P}, \mathcal{S}; \mathcal{X})\|_1$$

Enlarge the association discrepancy → Paying less attention to adjacent area





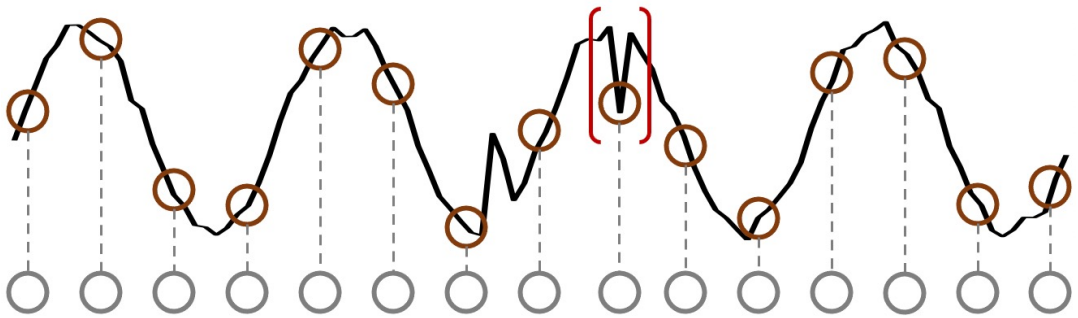
Abnormal time points have the adjacent-concentrate inductive bias → Making the reconstruction of abnormal time points harder

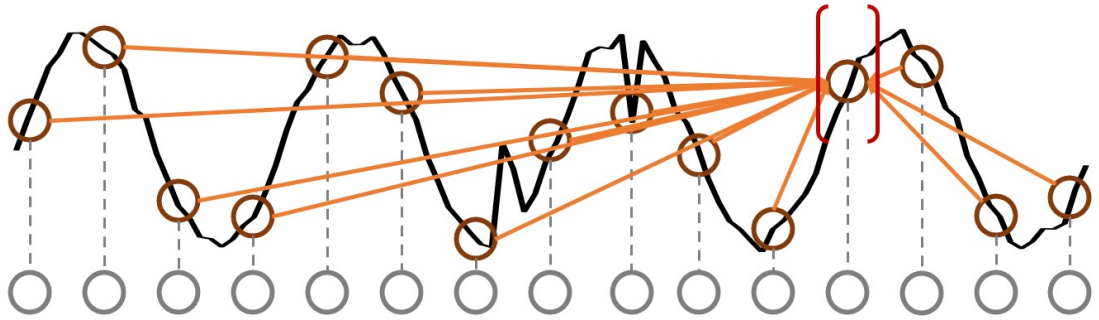
Association discrepancy: *the adjacent-concentration property of series-association*

Training Strategy (Vanilla Version)

$$\mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}}, \mathcal{P}, \mathcal{S}, \lambda; \mathcal{X}) = \|\mathcal{X} - \widehat{\mathcal{X}}\|_{\text{F}}^2 - \lambda \times \|\text{AssDis}(\mathcal{P}, \mathcal{S}; \mathcal{X})\|_1$$

Enlarge the association discrepancy → Paying less attention to adjacent area





Abnormal time points have the adjacent-concentrate inductive bias → Making the reconstruction of abnormal time points harder

Amplify the difference between normal and abnormal points

Association discrepancy: *the adjacent-concentration property of series-association*

Training Strategy (Vanilla Version)

$$\mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}}, \mathcal{P}, \mathcal{S}, \lambda; \mathcal{X}) = \|\mathcal{X} - \widehat{\mathcal{X}}\|_{\text{F}}^2 - \lambda \times \|\text{AssDis}(\mathcal{P}, \mathcal{S}; \mathcal{X})\|_1$$

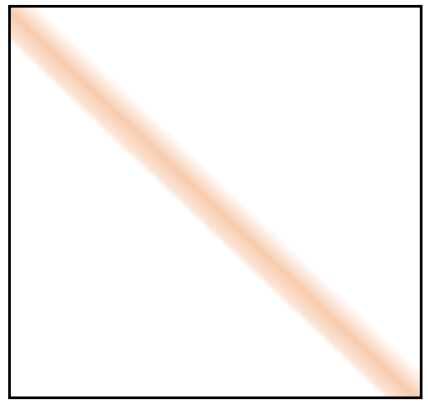
May fail for optimization ☹

Directly maximizing the association discrepancy will
extremely reduce the scale parameter of the Gaussian kernel

Series-Association

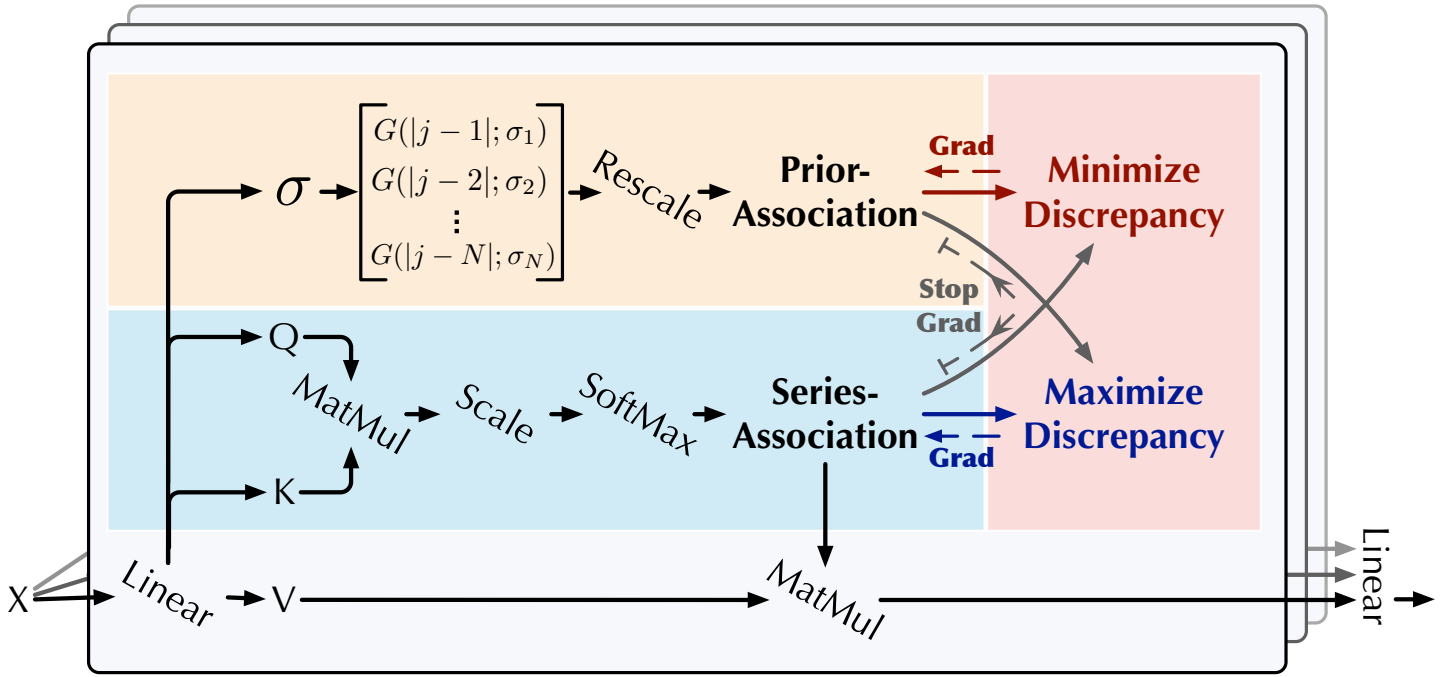
Prior-Association (well-optimized)

Prior-Association (degenerated $\sigma \to 0$)



Association discrepancy: *the adjacent-concentration property of series-association*

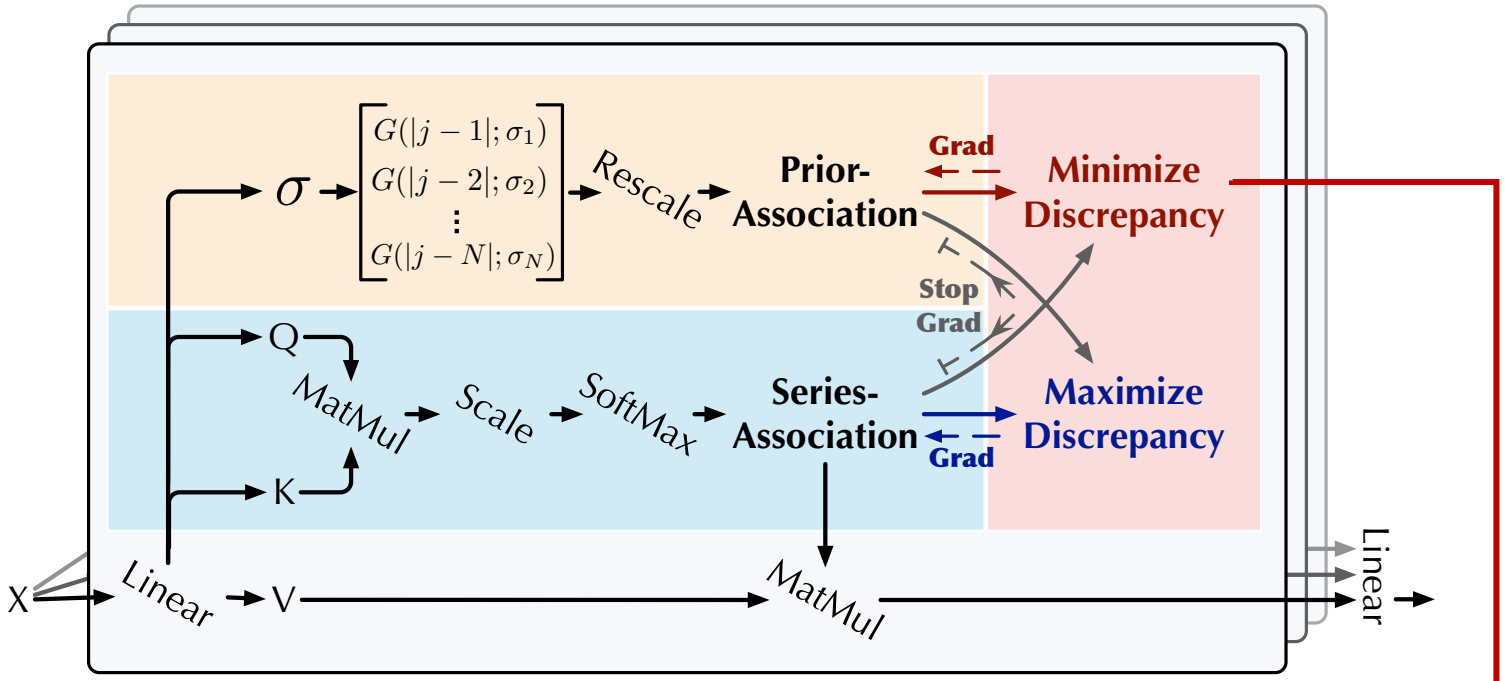
Minimax Association Learning



Minimize Phase: $\mathcal{L}_{\text{Total}}(\mathcal{X}, \mathcal{P}, \mathcal{S}_{\text{detach}}, -\lambda; \mathcal{X})$

Maximize Phase: $\mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}}, \mathcal{P}_{\text{detach}}, \mathcal{S}, \lambda; \mathcal{X}),$

Minimax Association Learning

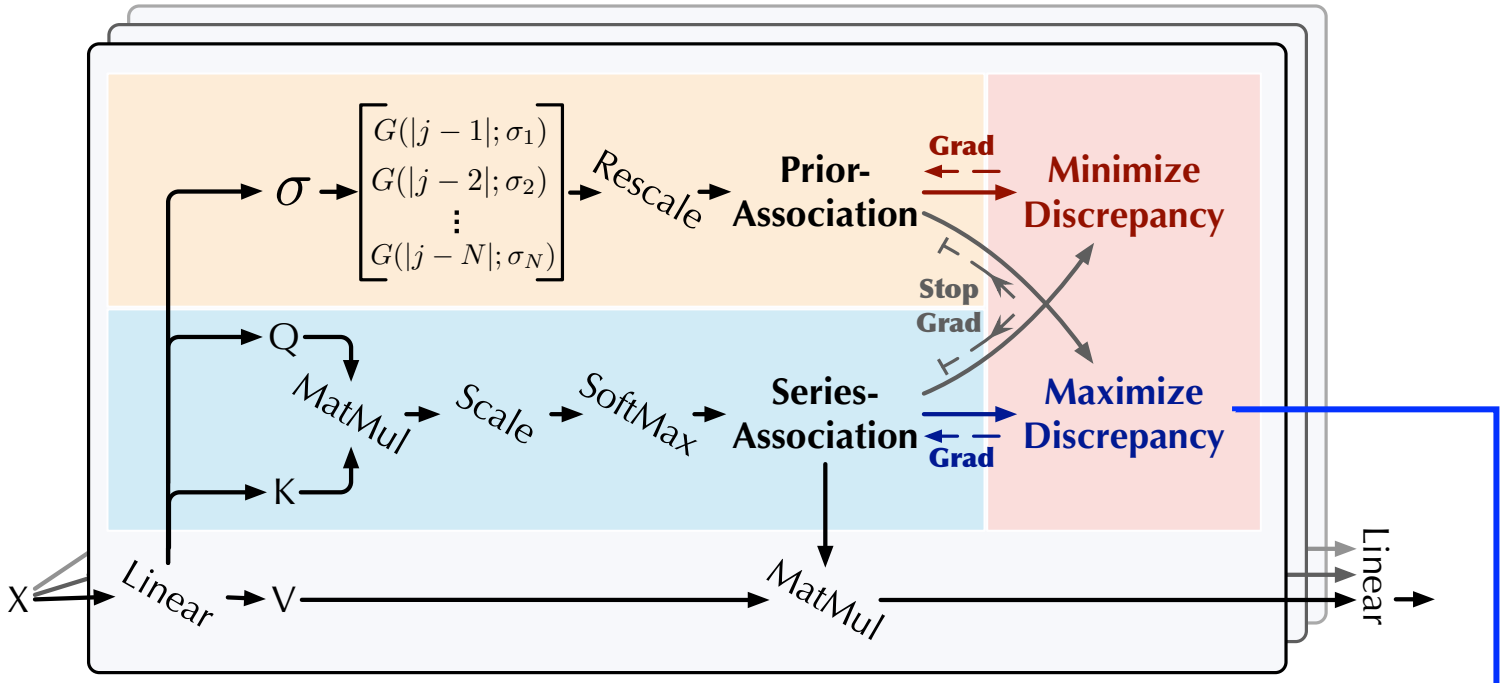


Learning prior-association $\mathcal{P}$ to avoid degeneration

Minimize Phase: $\mathcal{L}_{\text{Total}}(\mathcal{X}, \mathcal{P}, \mathcal{S}_{\text{detach}}, -\lambda; \mathcal{X})$

Maximize Phase: $\mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}}, \mathcal{P}_{\text{detach}}, \mathcal{S}, \lambda; \mathcal{X}),$

Minimax Association Learning

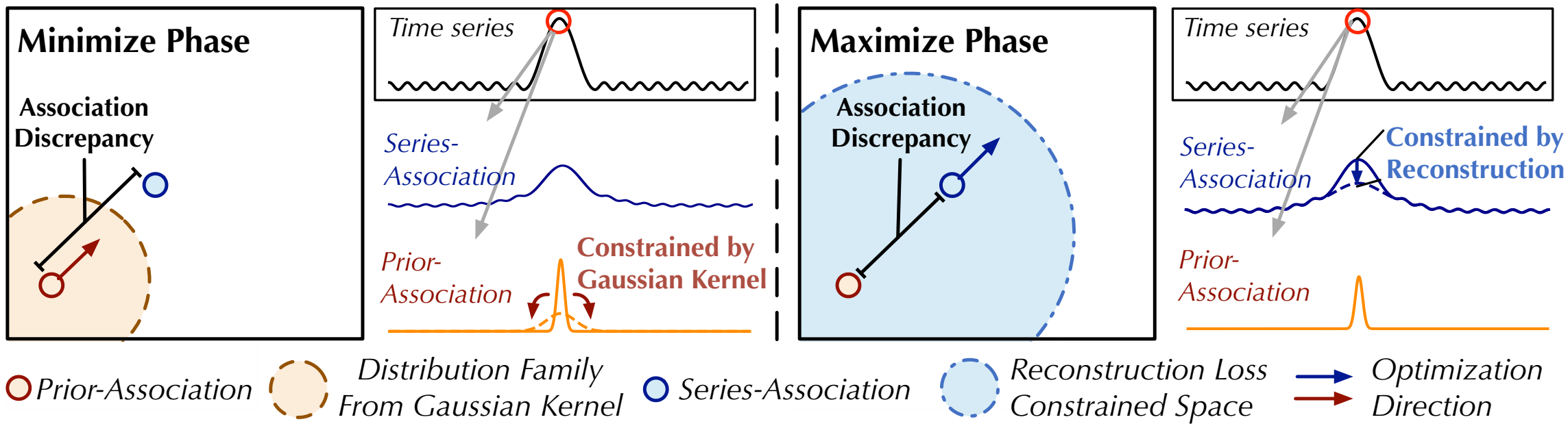


Amplify the difference between normal and abnormal points

Minimize Phase: $\mathcal{L}_{\text{Total}}(\mathcal{X}, \mathcal{P}, \mathcal{S}_{\text{detach}}, -\lambda; \mathcal{X})$

Maximize Phase: $\mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}}, \mathcal{P}_{\text{detach}}, \mathcal{S}, \lambda; \mathcal{X}),$

Minimax Association Learning



(1) Obtain a better estimation of association discrepancy

(2) Amplifying the normal-abnormal distinguishability

Association-based Anomaly Criterion

$$\mathrm{AnomalyScore}(\mathcal{X}) = \underbrace{\mathrm{Softmax}\Big(-\mathrm{AssDis}(\mathcal{P},\mathcal{S};\mathcal{X})\Big)}_{\text{Normalized Association Discrepancy}} \odot \underbrace{\Big[\|\mathcal{X}_{i,:}-\widehat{\mathcal{X}}_{i,:}\|_2^2\Big]_{i=1,\cdots,N}}_{\text{Reconstruction Error}}$$

Item for Anomalies	(1) Good Reconstruction	(2) Bad Reconstruction
Normalized Association Discrepancy	**Larger**	Unknown
Reconstruction Error	Smaller	**Larger**

Collaborate with each other to improve detection performance.

Points with anomaly scores larger than threshold will be detected as Anomalies.

Experiments

Table 13: Details of benchmarks. AR represents the truth abnormal proportion of the whole dataset.

Benchmarks	Applications	Dimension	Window	#Training	#Validation	#Test (labeled)	AR (Truth)
SMD	Server	38	100	566,724	141,681	708,420	0.042
PSM	Server	25	100	105,984	26,497	87,841	0.278
MSL	Space	55	100	46,653	11,664	73,729	0.105
SMAP	Space	25	100	108,146	27,037	427,617	0.128
SWaT	Water	51	100	396,000	99,000	449,919	0.121
NeurIPS-TS	Various Anomalies	1	100	20,000	10,000	20,000	0.018

Six benchmarks for three practical applications

Main Results (SOTA over 18 baselines)

Table 1: Quantitative results for Anomaly Transformer (*Ours*) in five real-world datasets. The *P*, *R* and *F1* represent the precision, recall and F1-score (as %) respectively. F1-score is the harmonic mean of precision and recall. For these three metrics, a higher value indicates a better performance.

Dataset	SMD			MSL			SMAP			SWaT			PSM		
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Classic methods / Density-based: OCSVM	44.34	76.72	56.19	59.78	86.87	70.82	53.85	59.07	56.34	45.39	49.22	47.23	62.75	80.89	70.67
Density-based: IsolationForest	42.31	73.29	53.64	53.94	86.54	66.45	52.39	59.07	55.53	49.29	44.95	47.02	76.09	92.45	83.48
Density-based: LOF	56.34	39.86	46.68	47.72	85.25	61.18	58.93	56.33	57.60	72.15	65.43	68.62	57.89	90.49	70.61
Density-based: Deep-SVDD	78.54	79.67	79.10	91.92	76.63	83.58	89.93	56.02	69.04	80.42	84.45	82.39	95.41	86.49	90.73
Density-based: DAGMM	67.30	49.89	57.30	89.60	63.93	74.62	86.45	56.73	68.51	89.92	57.84	70.40	93.49	70.03	80.08
Density-based: MMPCACD	71.20	79.28	75.02	81.42	61.31	69.95	88.61	75.84	81.73	82.52	68.29	74.73	76.26	78.35	77.29
Autoregression-based: VAR	78.35	70.26	74.08	74.68	81.42	77.90	81.38	53.88	64.83	81.59	60.29	69.34	90.71	83.82	87.13
Autoregression-based: LSTM	78.55	85.28	81.78	85.45	82.50	83.95	89.41	78.13	83.39	86.15	83.27	84.69	76.93	89.64	82.80
Autoregression-based: CL-MPPCA	82.36	76.07	79.09	73.71	88.54	80.44	86.13	63.16	72.88	76.78	81.50	79.07	56.02	99.93	71.80
Autoregression-based: ITAD	86.22	73.71	79.48	69.44	84.09	76.07	82.42	66.89	73.85	63.13	52.08	57.08	72.80	64.02	68.13
Reconstruction-based: LSTM-VAE	75.76	90.08	82.30	85.49	79.94	82.62	92.20	67.75	78.10	76.00	89.50	82.20	73.62	89.92	80.96
Reconstruction-based: BeatGAN	72.90	84.09	78.10	89.75	85.42	87.53	92.38	55.85	69.61	64.01	87.46	73.92	90.30	93.84	92.04
Reconstruction-based: OmniAnomaly	83.68	86.82	85.22	89.02	86.37	87.67	92.49	81.99	86.92	81.42	84.30	82.83	88.39	74.46	80.83
Reconstruction-based: InterFusion	87.02	85.43	86.22	81.28	92.70	86.62	89.77	88.52	89.14	80.59	85.58	83.01	83.61	83.45	83.52
Clustering-based: THOC	79.76	90.95	84.99	88.45	90.97	89.69	92.06	89.34	90.68	83.94	86.36	85.13	88.14	90.99	89.54
Ours	89.40	95.45	**92.33**	92.09	95.15	**93.59**	94.13	99.40	**96.69**	91.55	96.73	**94.07**	96.91	98.90	**97.89**

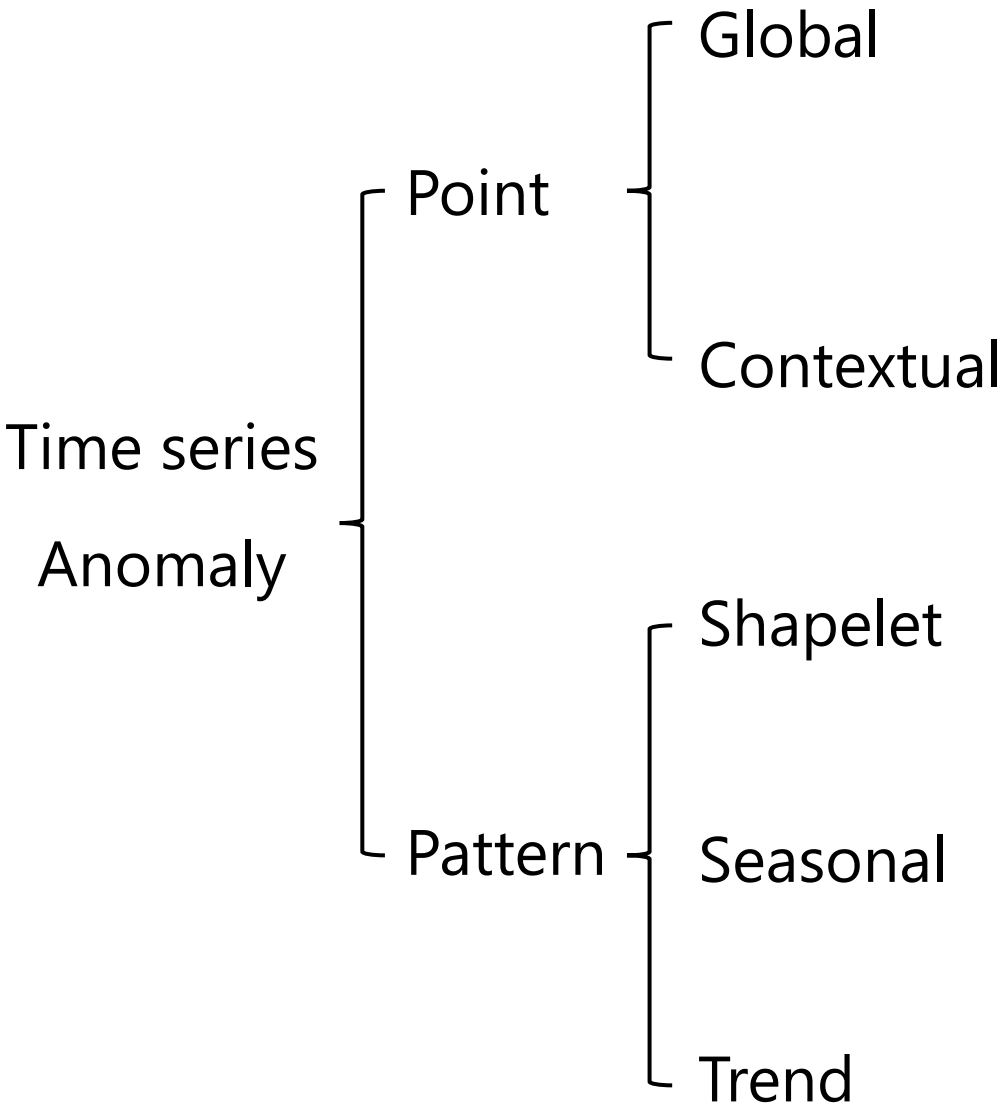
Main Results (SOTA over 18 baselines)

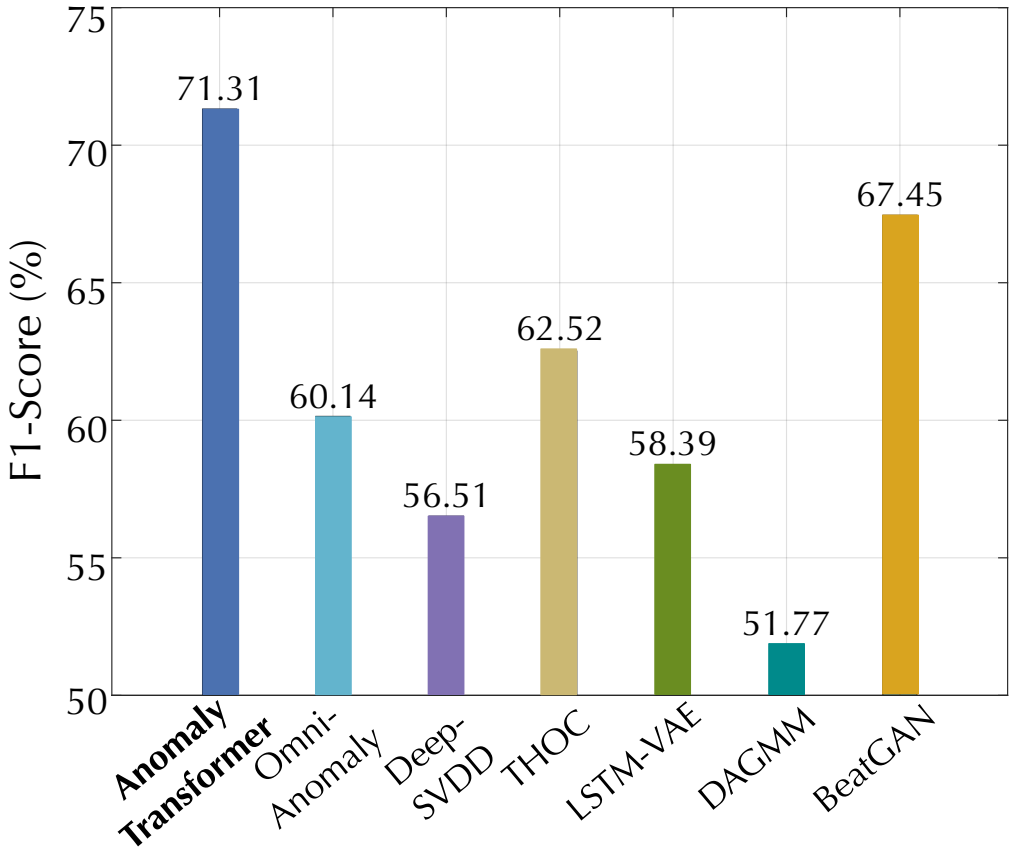
Table 1: Quantitative results for Anomaly Transformer (*Ours*) in five real-world datasets. The *P*, *R* and *F1* represent the precision, recall and F1-score (as %) respectively. F1-score is the harmonic mean of precision and recall. For these three metrics, a higher value indicates a better performance.

Dataset	SMD			MSL			SMAP			SWaT			PSM		
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
OCSVM	44.34	76.72	56.19	59.78	86.87	70.82	53.85	59.07	56.34	45.39	49.22	47.23	62.75	80.89	70.67
IsolationForest	42.31	73.29	53.64	53.94	86.54	66.45	52.39	59.07	55.53	49.29	44.95	47.02	76.09	92.45	83.48
LOF	56.34	39.86	46.68	47.72	85.25	61.18	58.93	56.33	57.60	72.15	65.43	68.62	57.89	90.49	70.61
Deep-SVDD	78.54	79.67	79.10	91.92	76.63	83.58	89.93	56.02	69.04	80.42	84.45	82.39	95.41	86.49	90.73
DAGMM	67.30	49.89	57.30	89.60	63.93	74.62	86.45	56.73	68.51	89.92	57.84	70.40	93.49	70.03	80.08
MMPCACD	71.20	79.28	75.02	81.42	61.31	69.95	88.61	75.84	81.73	82.52	68.29	74.73	76.26	78.35	77.29
VAR	78.35	70.26	74.08	74.68	81.42	77.90	81.38	53.88	64.83	81.59	60.29	69.34	90.71	83.82	87.13
LSTM	78.55	85.28	81.78	85.45	82.50	83.95	89.41	78.13	83.39	86.15	83.27	84.69	76.93	89.64	82.80
CL-MPPCA	82.36	76.07	79.09	73.71	88.54	80.44	86.13	63.16	72.88	76.78	81.50	79.07	56.02	99.93	71.80
ITAD	86.22	73.71	79.48	69.44	84.09	76.07	82.42	66.89	73.85	63.13	52.08	57.08	72.80	64.02	68.13
LSTM-VAE	75.76	90.08	82.30	85.49	79.94	82.62	92.20	67.75	78.10	76.00	89.50	82.20	73.62	89.92	80.96
BeatGAN	72.90	84.09	78.10	89.75	85.42	87.53	92.38	55.85	69.61	64.01	87.46	73.92	90.30	93.84	92.04
OmniAnomaly	83.68	86.82	85.22	89.02	86.37	87.67	92.49	81.99	86.92	81.42	84.30	82.83	88.39	74.46	80.83
InterFusion	87.02	85.43	86.22	81.28	92.70	86.62	89.77	88.52	89.14	80.59	85.58	83.01	83.61	83.45	83.52
THOC	79.76	90.95	84.99	88.45	90.97	89.69	92.06	89.34	90.68	83.94	86.36	85.13	88.14	90.99	89.54
Ours	89.40	95.45	**92.33**	92.09	95.15	**93.59**	94.13	99.40	**96.69**	91.55	96.73	**94.07**	96.91	98.90	**97.89**

Previous SOTA

NeurIPS-TS benchmark





Achieve SOTA on various anomalies.

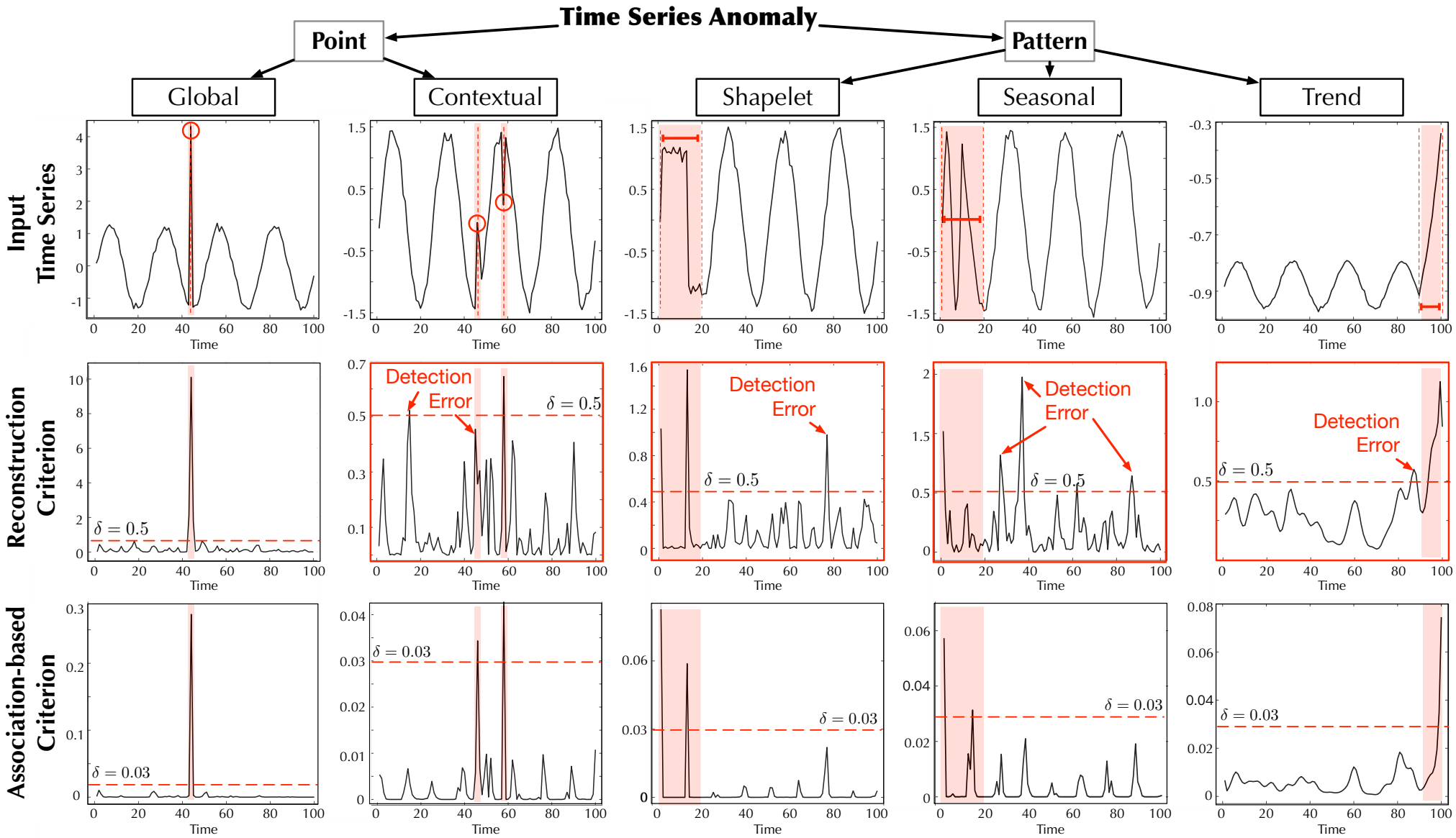
Ablation study

Table 2: Ablation results (F1-score) in anomaly criterion, prior-association and optimization strategy. *Recon*, *AssDis* and *Assoc* mean the pure reconstruction performance, pure association discrepancy and our proposed association-based criterion respectively. *Fix* is to fix *Learnable* scale parameter σ of prior-association as 1.0. *Max* and *Minimax* ref to the strategies for association discrepancy in the maximization (Equation 4) and minimax (Equation 5) way respectively.

Architecture	Anomaly Criterion	Prior-Association	Optimization Strategy	SMD	MSL	SMAP	SWaT	PSM	Avg F1 (as %)
Transformer	Recon	×	×	79.72	76.64	73.74	74.56	78.43	76.62
Anomaly Transformer	Recon	Learnable	Minmax	71.35	78.61	69.12	81.53	80.40	76.20
	AssDis	Learnable	Minmax	87.57	90.50	90.98	93.21	95.47	91.55
	Assoc	Fix	Max	83.95	82.17	70.65	79.46	79.04	79.05
	Assoc	Learnable	Max	88.88	85.20	87.84	81.65	93.83	87.48
*final	Assoc	Learnable	Minmax	**92.33**	**93.59**	**96.90**	**94.07**	**97.89**	**94.96**

(1) Anomaly Criterion **18.76%↑** (2) Prior-association **8.43%↑** (3) optimization strategy **7.84%↑**

Visualization of Anomaly Criterion



Visualization of Prior-Association

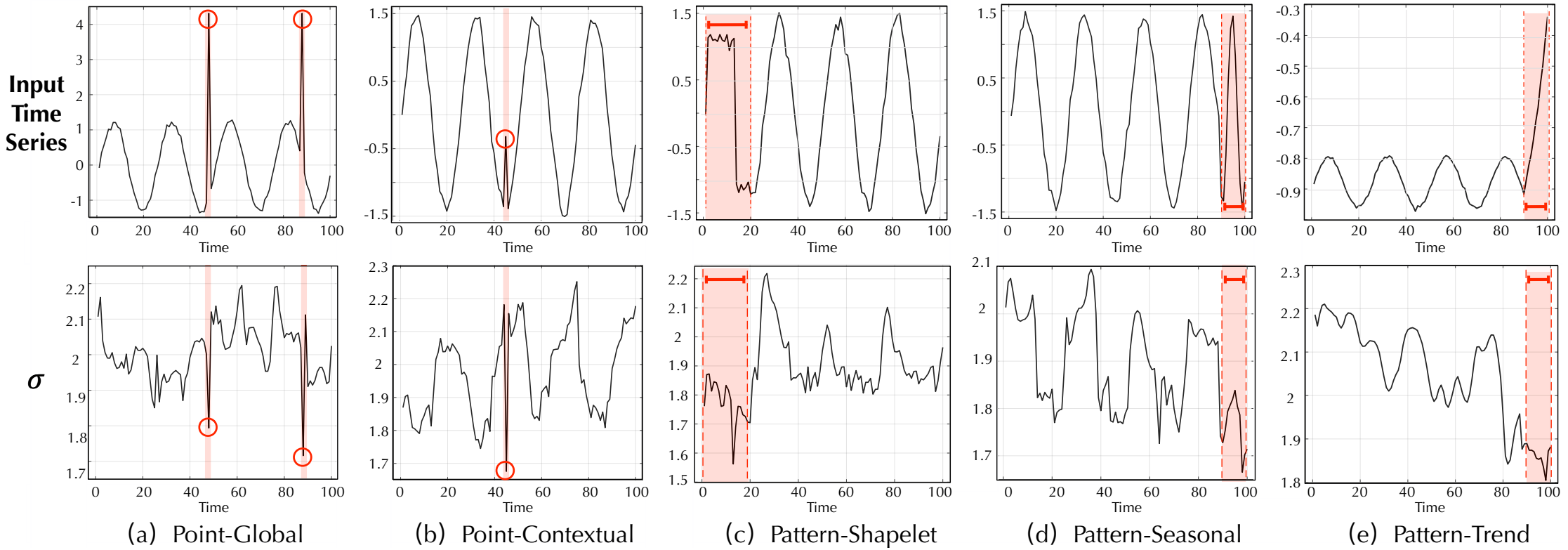


Figure 6: Learned scale parameter σ for different types of anomalies (highlight in red).

Prior-association can adapt to various data patterns of time series.

Abnormal time points show the adjacent-concentration property.

Statistics of Optimization Strategy

Table 3: The statistical results of adjacent association weights for *Abnormal* and *Normal* time points respectively. *Recon*, *Max* and *Minimax* represent the association learning process that is supervised by reconstruction loss, direct maximization and minimax strategy respectively. A higher contrast value ($\frac{\text{Abnormal}}{\text{Normal}}$) indicates a stronger distinguishability between normal and abnormal time points.

Dataset	SMD			MSL			SMAP			SWaT			PSM		
Optimization	Recon	Max	Ours	Recon	Max	Ours	Recon	Max	Ours	Recon	Max	Ours	Recon	Max	Ours
Abnormal (%)	1.08	0.95	0.86	1.01	0.65	0.35	1.29	1.18	0.70	1.27	0.89	0.37	1.02	0.56	0.29
Normal (%)	0.94	0.75	0.36	1.00	0.59	0.22	1.23	1.09	0.49	1.18	0.78	0.21	0.99	0.54	0.11
Contrast ($\frac{\text{Abnormal}}{\text{Normal}}$)	1.15	1.27	**2.39**	1.01	1.10	**1.59**	1.05	1.08	**1.43**	1.08	1.14	**1.76**	1.03	1.04	**2.64**

Directly maximizing association discrepancy will cause degeneration.

Minimax association learning will amplifying the normal-abnormal distinguishability.

Thank You!

whx20@mails.tsinghua.edu.cn