Multivariate Time Series Anomaly Detection and Interpretation using Hierarchical Inter-Metric and Temporal Embedding

Zhihan Li, Youjian Zhao, Jiaqi Han, Ya Su, Rui Jiao, Xidao Wen, and Dan Pei. 2021.
In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’21)
2025년 2월 18일

이규원

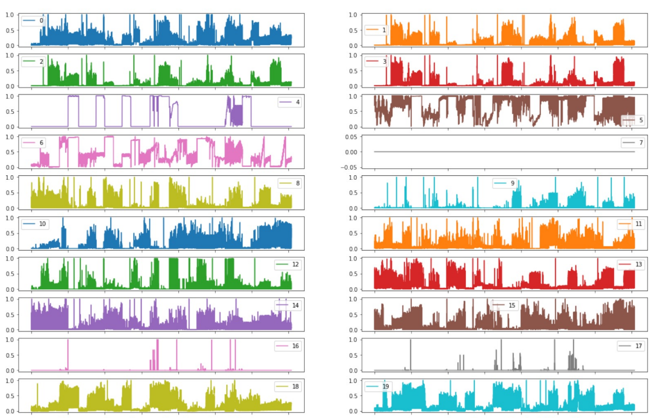
Department of Computer Science and Engineering
Chung-Ang University

Index

- Backgrounds
- Motivation
- Model Architecture
- Experimental Results
- Conclusion

Backgrounds

- **Multivariate Time Series**
  - Involves more than one variable over time
  - High dimensional data



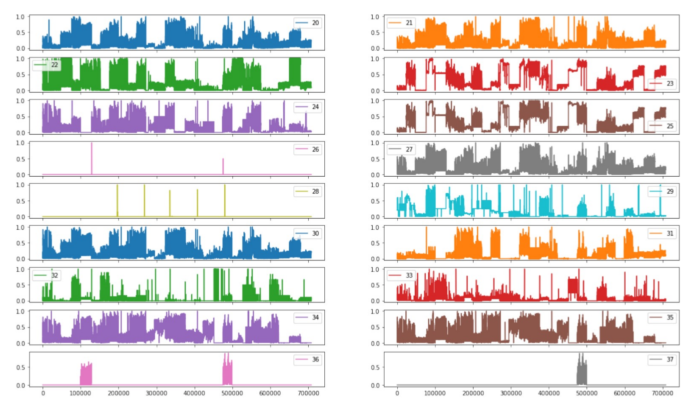
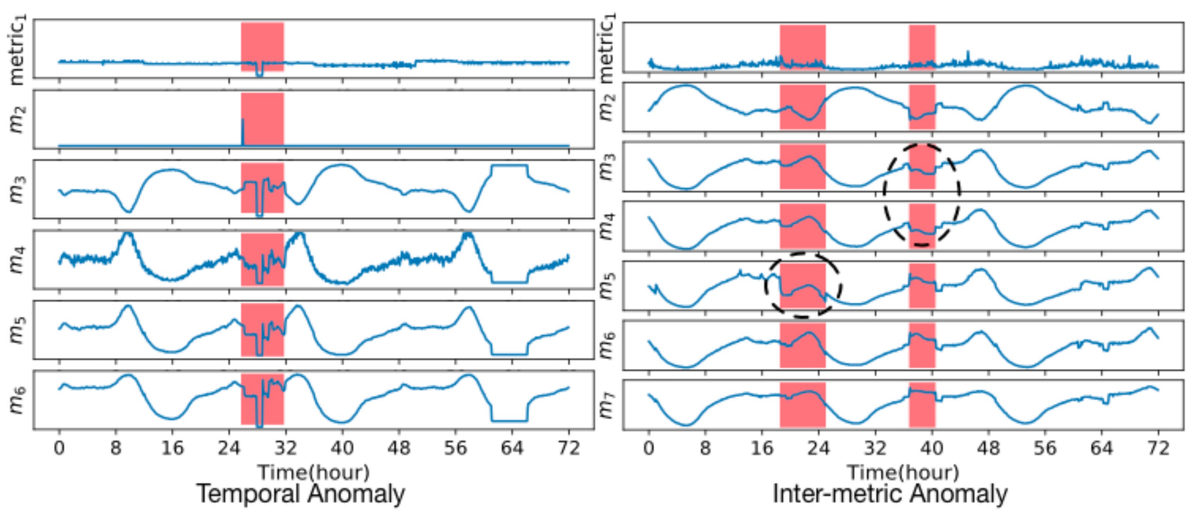


Figure 6: Waveform of SMD dataset.

Xu, Zhijian, Ailing Zeng, and Qiang Xu. "FITS: Modeling time series with $10 k $ parameters." arXiv preprint arXiv:2307.03756 (2023).

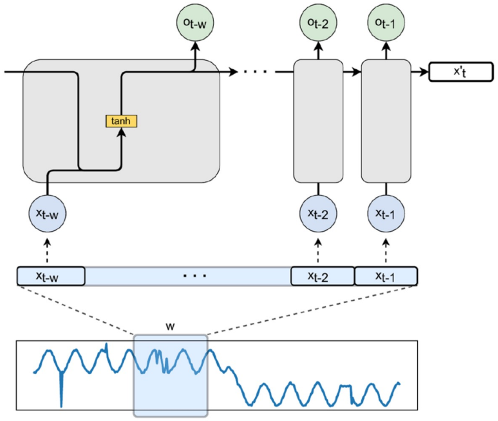
Backgrounds

- **Time Series Anomaly Detection**
  - Identifying patterns or events in time-ordered data that deviate significantly from expected behavior or norms
- **Time Series Anomaly**
  - Temporal: Significantly deviate from their historical patterns
  - Inter-metric: The relationships among different metrics violate their expected correlations

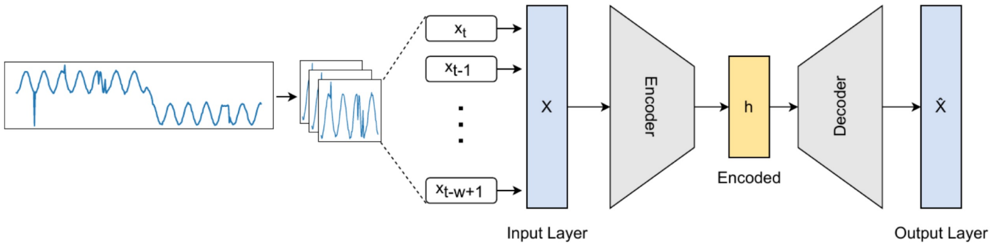


Backgrounds

- **Forecasting-based Model**
  - To predict a future point or subsequence based on a point or a recent window
- **Reconstruction-based Model**
  - To reconstruct normal data well while failing to reconstruct anomalous data



Forecasting

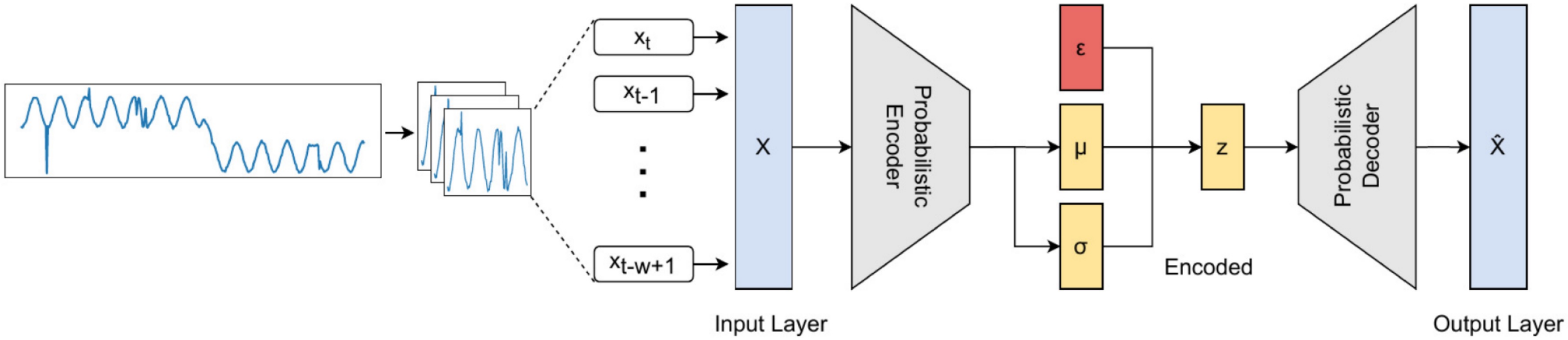


Reconstruction

Zamanzadeh Darban, Zahra, et al. "Deep learning for time series anomaly detection: A survey." ACM Computing Surveys 57.1 (2024): 1-42.

Backgrounds

- **Stochasticity**
  - Uncertainty (factors we cannot provide as input) affects time-series data, causing variations
- **Variational Auto Encoder**
  - Decoder: Finds the probabilistic generative model $p_\theta(x|z)$ that makes data most probable from latent space z
  - Encoder: Uses $q_\phi(z|x)$ to approximate $p_\theta(z|x)$ due to its intractability
  - Estimate parameters of data distributions at each timestamp (e.g., μ, σ of a Gaussian)



(b) Variational Auto-Encoder

Zamanzadeh Darban, Zahra, et al. "Deep learning for time series anomaly detection: A survey." ACM Computing Surveys 57.1 (2024): 1-42.

Backgrounds

- **Variational Auto Encoder – ELBO**

$$p_\theta(x) = \int p_\theta(x|z)\,\underbrace{p_\theta(z)}_{\text{prior of } z}\,\boxed{dz}$$

모든 z에 대해 적분할수 없다.

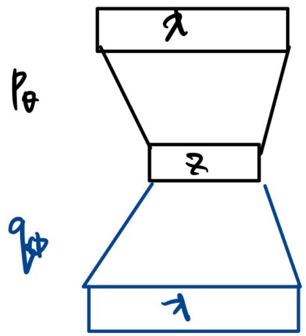
구해보자.

$$p_\theta(z|x) = \frac{p_\theta(x|z)\,p_\theta(z)}{\boxed{p_\theta(x)}}$$

못구했다.

Backgrounds

- **Variational Auto Encoder – ELBO**



$$q_\phi(z|x) \simeq p_\theta(z|x)$$

$\mathbb{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log(p_\theta(x^{(i)}))\right]$를 최대화해보자.

$$p_\theta(x^{(i)}|z) = \frac{p_\theta(z|x^{(i)})\,p_\theta(x^{(i)})}{p_\theta(z)}$$

$$p_\theta(x^{(i)}) = \frac{p_\theta(x^{(i)}|z)\,p_\theta(z)}{p_\theta(z|x^{(i)})}$$

$$\mathbb{E}_z\left[\log\frac{p_\theta(x^{(i)}|z)\,p_\theta(z)}{p_\theta(z|x^{(i)})}\right]$$

$$= \mathbb{E}_z\left[\log\frac{p_\theta(x^{(i)}|z)\,p_\theta(z)}{p_\theta(z|x^{(i)})}\cdot\frac{q_\phi(z|x^{(i)})}{q_\phi(z|x^{(i)})}\right]$$

$$= \mathbb{E}_z\left[\log p_\theta(x^{(i)}|z)\right] - \mathbb{E}_z\left[\log\frac{q_\phi(z|x^{(i)})}{p_\theta(z)}\right] + \mathbb{E}_z\left[\log\frac{q_\phi(z|x^{(i)})}{p_\theta(z|x^{(i)})}\right]$$

lower bound

$$= \mathbb{E}_z\left[\log p_\theta(x^{(i)}|z)\right] - D_{KL}\left(q_\phi(z|x^{(i)})\,\|\,p_\theta(z)\right) + D_{KL}\left(q_\phi(z|x^{(i)})\,\|\,\boxed{p_\theta(z|x^{(i)})}\right)$$

못구한다.

reconstruction 잘 돼나.

p_θ는 Gaussian, Bernoulli decoder

샘플링함수가 prior를 따르도록하라.

→ 왜? 오버피팅 방지

$D_{KL} \geq 0$

$$\mathbb{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log\frac{q_\phi(z|x^{(i)})}{p_\theta(z)}\right]$$

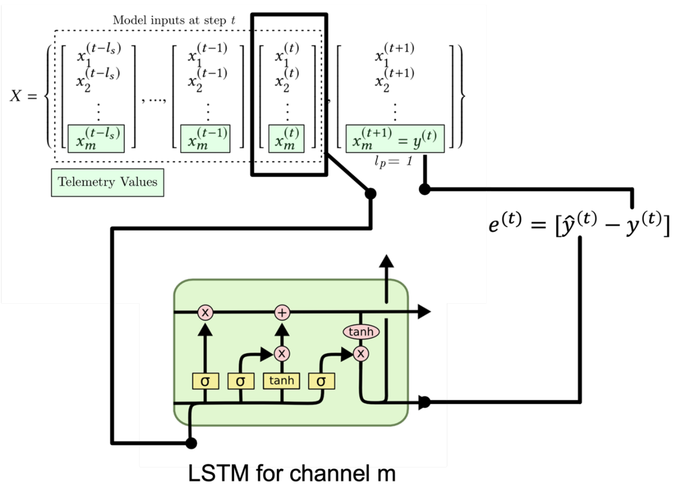
$$= \int_z q_\phi(z|x^{(i)})\log\frac{q_\phi(z|x^{(i)})}{p_\theta(z)}\,dz$$

$$KL(P\|Q)$$

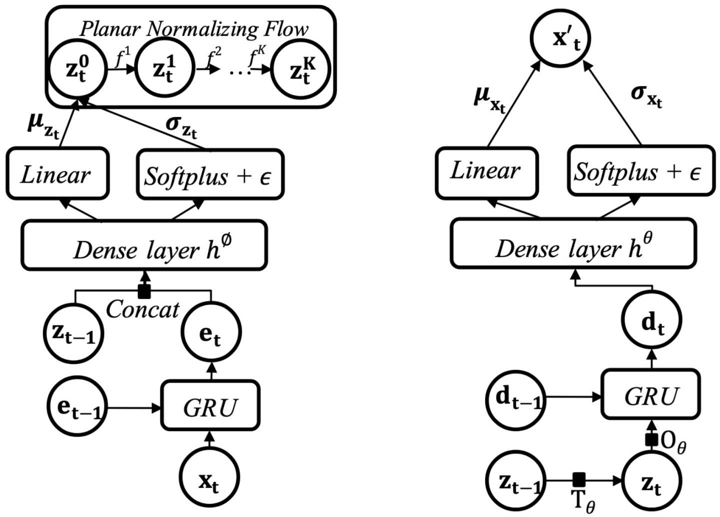
$$= \sum_x P(x)\log\frac{P(x)}{Q(x)}$$

Motivation

- **Previous Methods**
  - LSTM-NDT
  - MSCRED
  - LSTM-VAE, OmniAnomaly
  - MADGAN



LSTM-NDT

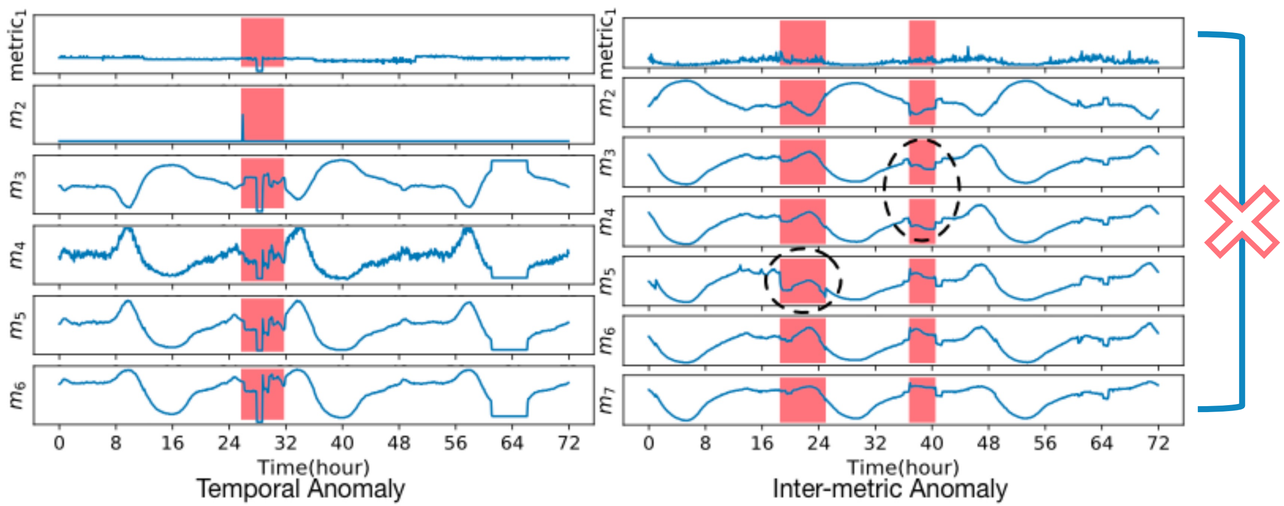


OmniAnomaly

Hundman, Kyle, et al. "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding." Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018.
Su, Ya, et al. "Robust anomaly detection for multivariate time series through stochastic recurrent neural network." Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019.

Motivation

- **Limitation**
  - Modeled either **temporal dependency** or **intermetric dependency**
  - MSCRED Captures intermetric **correlation** rather than learning explicit intermetric embeddings

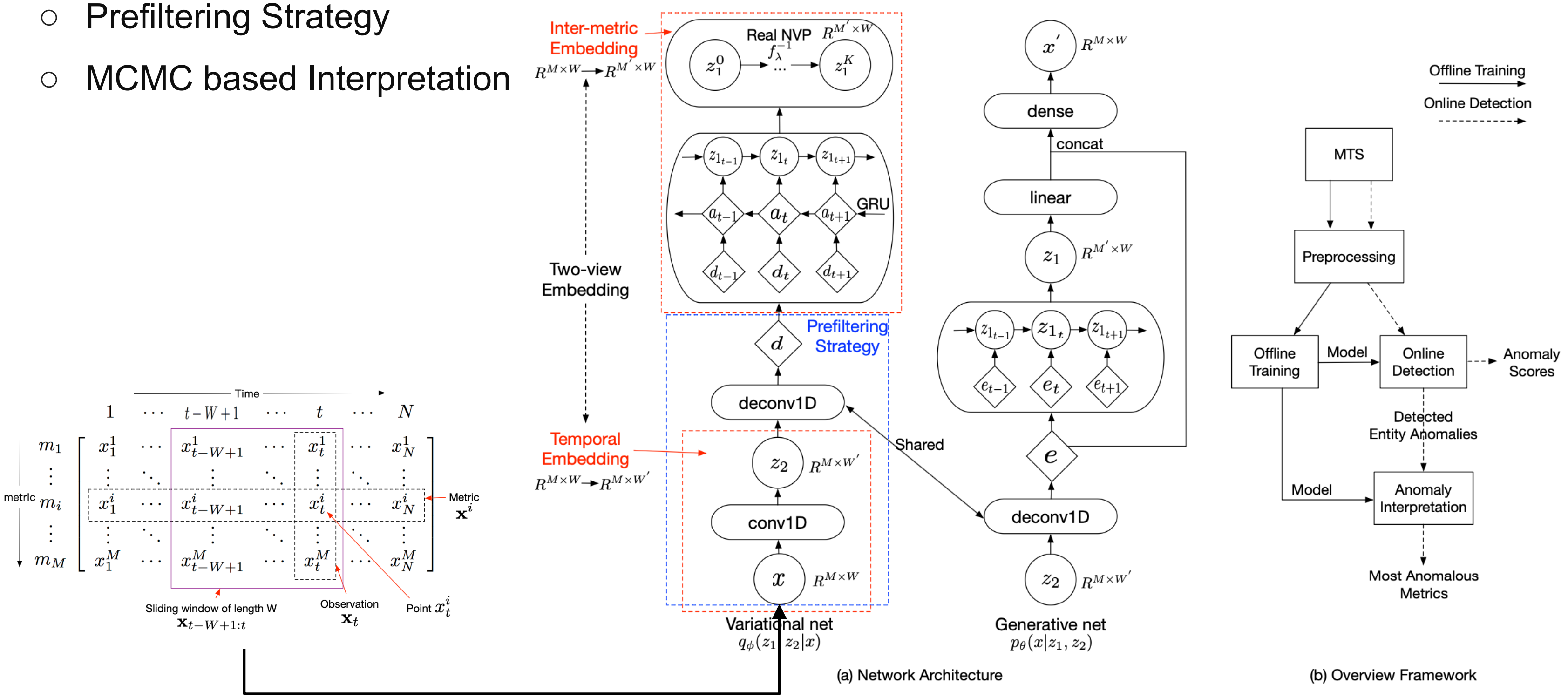


Motivation

- **Purpose**
  - Explicitly learn the low-dimensional intermetric and temporal representations
  - Robustness to noise and anomalies in training data
  - Improving interpretability of anomaly detection
- **Idea**
  - Model based on HVAE and Two-view Embedding
  - Prefiltering Strategy
  - MCMC based Interpretation

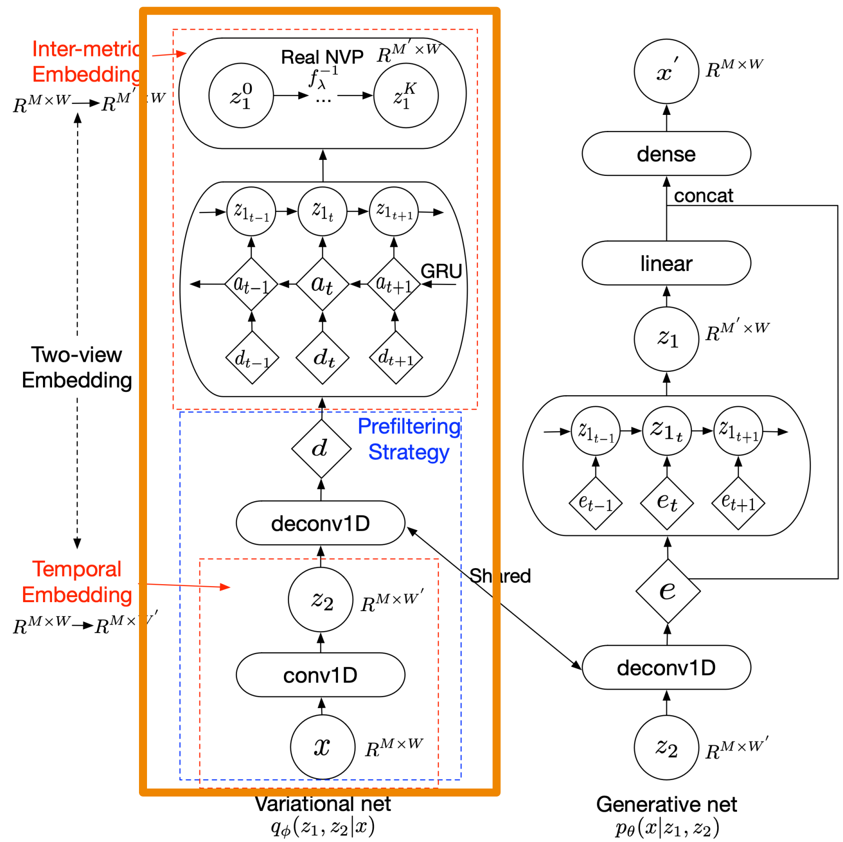
Model Architecture

- **Idea**
  - Model based on HVAE and Two-view Embedding
  - Prefiltering Strategy
  - MCMC based Interpretation



Model Architecture

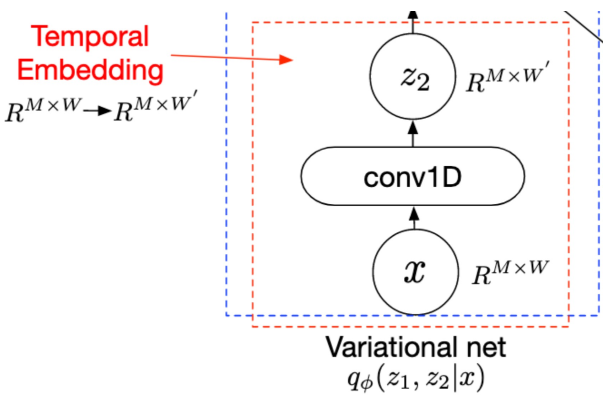
- **HVAE**
  - Temporal → Intermetric
  - Rather than learning independently

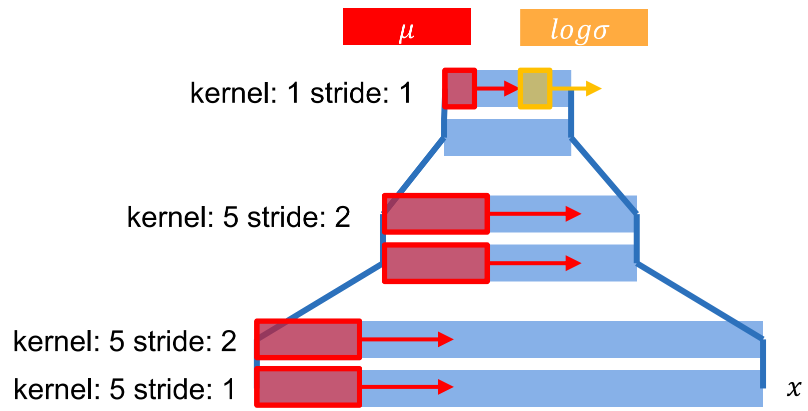


(a) Network Architecture

Model Architecture

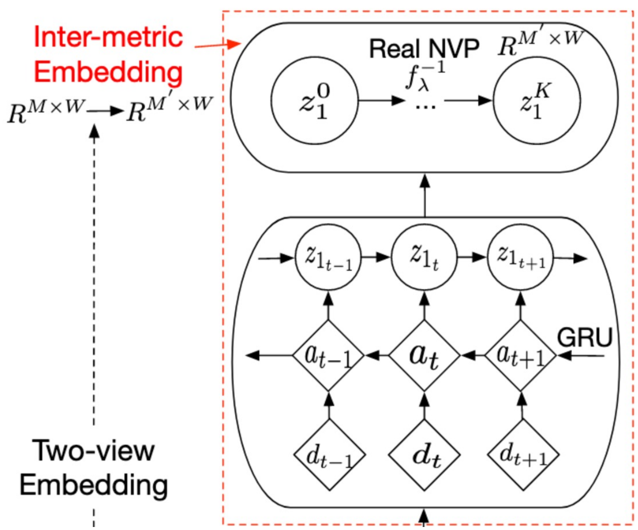
- **HVAE – Temporal Embedding**
  - Captures the temporal dependency using Conv1D layers along the time dimension

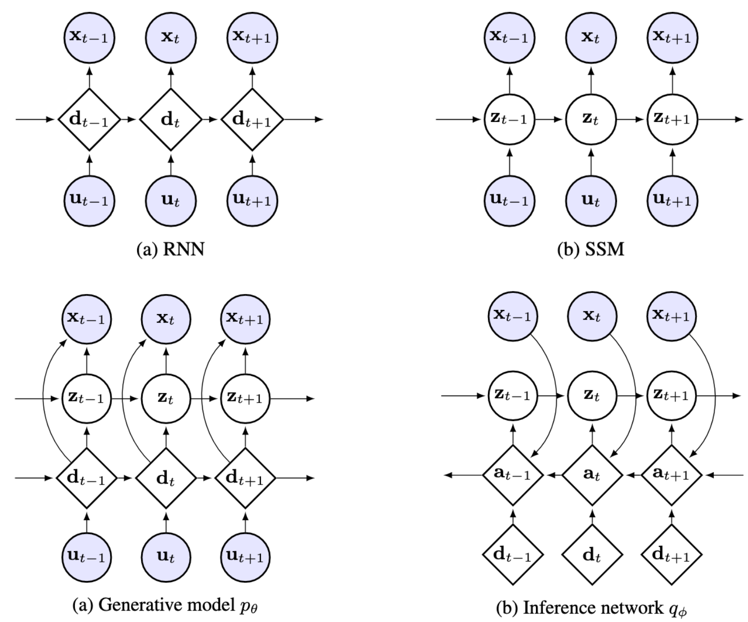




Model Architecture

- **HVAE – Intermetric Embedding**
  - Represents the intermetric dependency using an SRNN-like architecture
  - Real NVP learns an invertible, stable, mapping between a $q_{\emptyset}(z_1^k)$and a latent distribution $q_{\emptyset}(z_1^0)$ (typically a Gaussian)



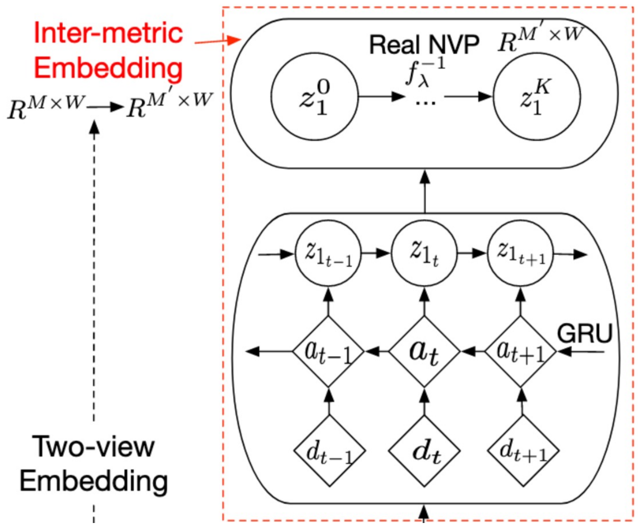


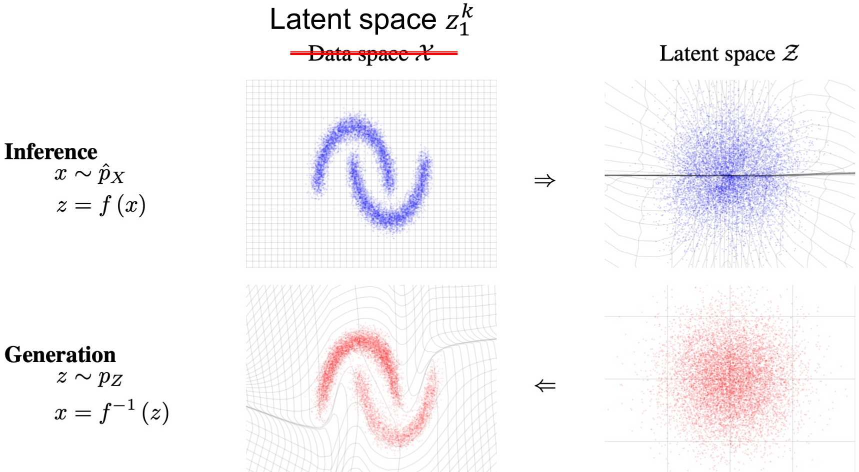
Stochastic RNN

Fraccaro, Marco, et al. "Sequential neural models with stochastic layers." Advances in neural information processing systems 29 (2016).

Model Architecture

- **HVAE – Intermetric Embedding**
  - Represents the intermetric dependency using an SRNN-like architecture
  - Real NVP learns an invertible, stable, mapping between a $q_{\emptyset}(z_1^k)$and a latent distribution $q_{\emptyset}(z_1^0)$ (typically a Gaussian)





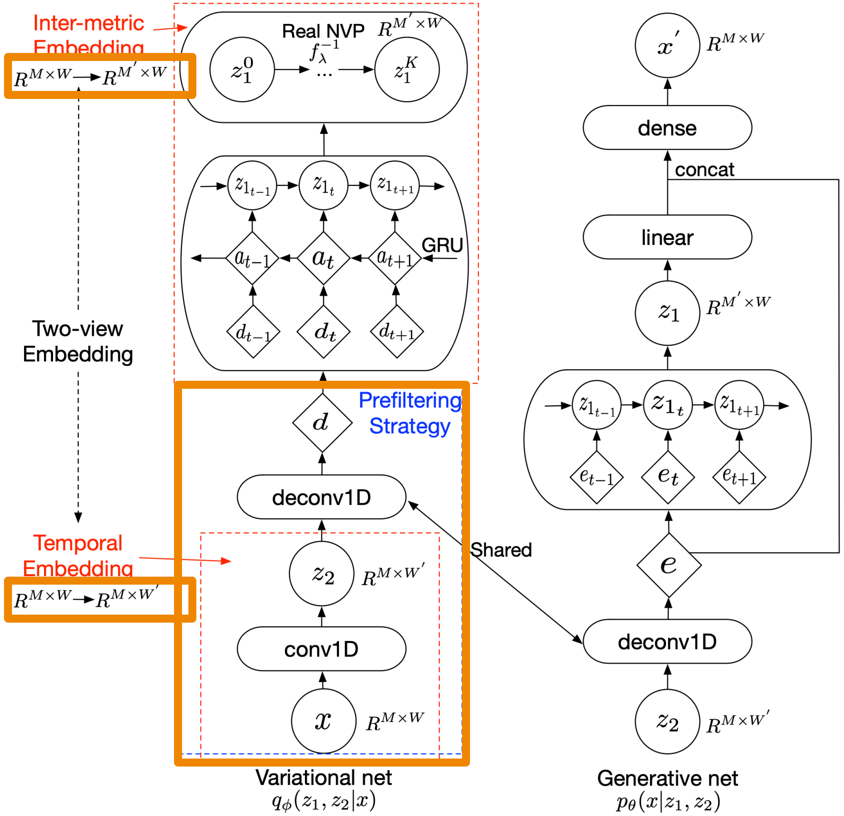
Normalizing Flow

Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP." *International Conference on Learning Representations*. 2022.

Model Architecture

- **Two-view Embedding**
  - Derives intermetric embeddings using the reconstructed d $R^{M\times W}$
  - To preserve time consistency

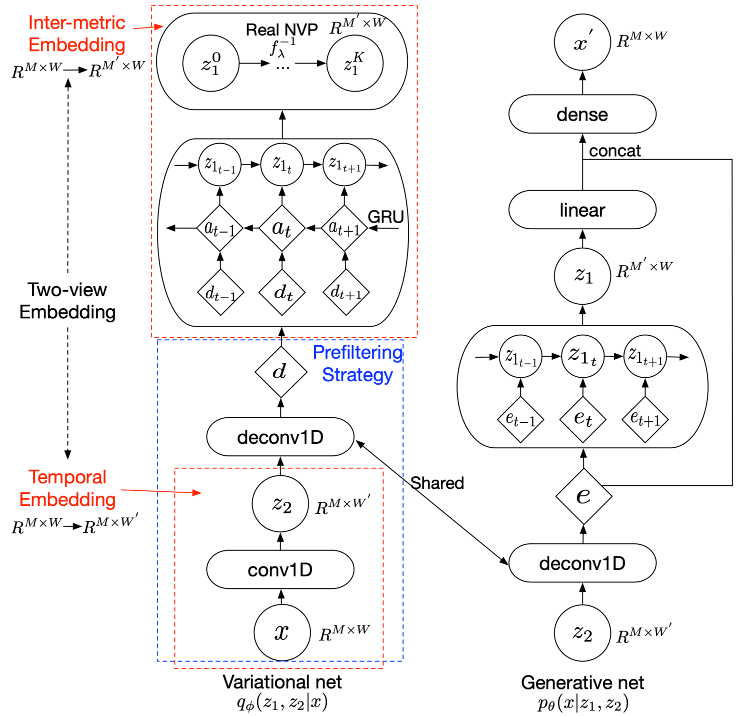
- **Prefiltering Strategy**
  - Filtering noise and anomalies in real world training data
  - z_1 derived from reconstructed d
  - d is pretrained for initial reconstruction



(a) Network Architecture

Model Architecture

- **Generative net**
  - p_θ uses z_1, z_2 sampled from q_ϕ to reconstruct
  - $p_\theta(z_1)$ without backward GRU and Real NVP
  - $p_\theta(z_2)$ is Normal distribution
  - $p_\theta(z_1), p_\theta(z_2)$ are used in ELBO calculation, specifically in the regulaization term



(a) Network Architecture

$$\mathcal{L}(\mathbf{x}, \theta, \phi) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z_1}, \mathbf{z_2}, \mathbf{d}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z_1}, \mathbf{z_2}, \mathbf{e})]}_{\text{Reconstruction term}} - \underbrace{D_{KL}(q_\phi(\mathbf{z_1}, \mathbf{z_2}, \mathbf{d}|\mathbf{x})||p_\theta(\mathbf{z_1}, \mathbf{z_2}, \mathbf{e}))}_{\text{Regularization term}}$$

Model Architecture

- **Interpretation (MCMC imputation)**
  - Masking anomalous data points (metric, time)
  - Infer the normal values of severe anomalies
  - $\tilde{x}$: Anomalous parts have been replaced with inferred normal values
  - Perform $\mathbb{E}_{q_\phi(z_1,z_2|\tilde{x})}[\log p_\theta(x|z_1,z_2)]$
  - Detect both severe and subtle anomalies

Algorithm 1: *InterFusion* Anomaly Interpretation

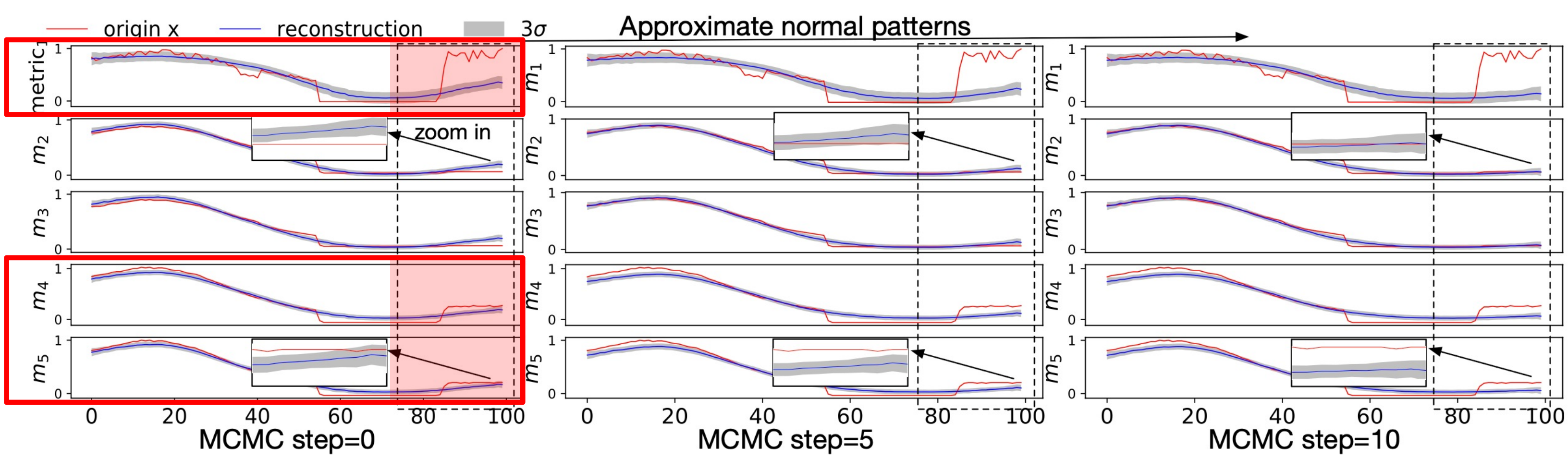
Input: input sequence $\mathbf{x} \in R^{M\times W}$, original reconstruction probability $\mathbf{r}^0$, normal baseline b, window length W, number of metrics M, small constant ratio β_{init}, β_{inc}
Output: revised anomaly score $AS \sim R^{M\times W}$ for interpretation

```
n_p ← number of points (x_{m,t}) where r^0_{m,t} < b/(M*W);
n_init ← β_init n_p, n_inc ← β_inc n_p, n ← n_init, r^a = Σ_{m,t} r^0_{m,t};
while not (r^a ≥ b or n > n_p) do
    x_m ← top n points in x that have the lowest r^0_{m,t};
    x_o ← other points in x but not in x_m;
    Denote x' = x = (x_o, x_m);
    for s ← 1 to S do        // MCMC imputation for S times
        sample (z_1, z_2) from q_φ(z_1, z_2|x_o, x_m);
        reconstruct (x'_o, x'_m) from p_θ(x_o, x_m|z_1, z_2);
        update x' ← (x_o, x'_m);
    end
    /* Approximate the true reconstruction prob of the
       input window using revised x'                        */
    r^a = (M*W)/(M*W−n) E_{q_φ(z_1,z_2|x')}[Σ_{x_i∈x_o} log p_θ(x_i|z_1, z_2)];
    add r^a to rlist, n ← n + n_inc;
end
x̃ ← x' that achieves the highest r^a in rlist;
r^f = E_{q_φ(z_1,z_2|x̃)}[log p_θ(x|z_1, z_2)], AS = −r^f;
```

$n_p \leftarrow$ number of points $(\mathbf{x}_{m,t})$ where $\mathbf{r}^0_{m,t} < \frac{b}{M*W}$;
$n_{init} \leftarrow \beta_{init} n_p$, $n_{inc} \leftarrow \beta_{inc} n_p$, $n \leftarrow n_{init}$, $r^a = \sum_{m,t} \mathbf{r}^0_{m,t}$;
$r^a = \frac{M*W}{M*W-n}\mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x}')}[\sum_{\mathbf{x}_i\in\mathbf{x}_o}\log p_\theta(\mathbf{x}_i|\mathbf{z}_1,\mathbf{z}_2)]$;
$\mathbf{r}^f = \mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\tilde{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_1,\mathbf{z}_2)]$, $AS = -\mathbf{r}^f$;

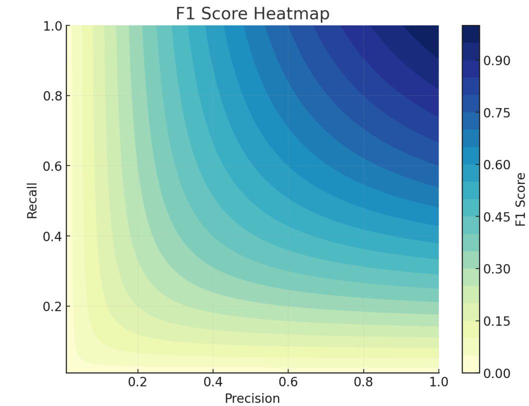
Model Architecture

- **Interpretation (MCMC imputation)**
  - Masking anomalous data points (metric, time)
  - Infer the normal values of severe anomalies
  - $\tilde{x}$ represents the input where anomalous parts have been replaced with inferred normal values
  - Perform $\mathbb{E}_{q_\phi(z_1, z_2|\tilde{x})}[\log p_\theta(x|z_1, z_2)]$
  - Detect both severe and subtle anomalies



Experimental Results

- **Evaluation Metric**
  - Point Adjust Precision/Recall
  - Interpretation Score(Proposed)
- **Thresholding Method**
  - Best F1 score



$$IPS = \sum_{a=1}^{A} w_a \frac{|G_{\Phi_a} \cap I_{\Phi_a}|}{|G_{\Phi_a}|}, \quad w_a = \frac{N_{\phi_a}}{\sum_{a=1}^{A} N_{\phi_a}}$$

Interpretation Score

$\phi_a : An\ anomaly\ segement$

$G_{\phi_a} : Ground - truth\ anomalous\ metric\ set\ for\ segment\ \phi_a$

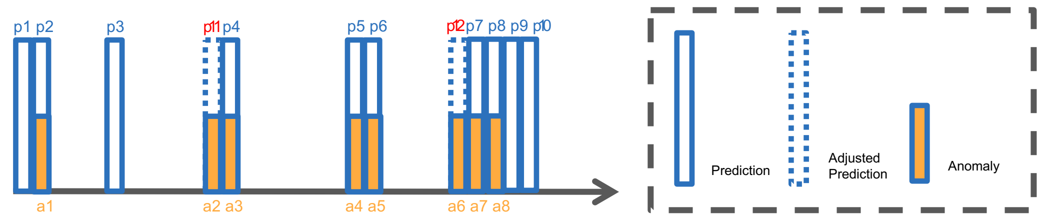
$I_{\phi_a} : The\ set\ of\ top\ k\ anomalous\ metrics\ in\ the\ segment\ \phi_a$

$k = |G_{\phi_a}|, \qquad w = length\ weight$

PA Precision : $\frac{Correct\ Predictions}{Total\ Predictions}$

$(Predictions = Predictions + Adjusted\ Predictions)$

PA Recall : $\frac{Detected\ Anomalies}{Total\ Anomalies}$



PA Precision = 0.75, PA Recall = 1
F1 Score = 0.857

Experimental Results-F1 Score

Table 1: Average best-F1 for *InterFusion* and baselines.

Methods	SWaT	WADI	SMD	ASD	Avg.
LSTM-NDT	0.8133	0.5067	0.7687	0.4061	0.6237
MSCRED	0.8346	0.5469	0.8252	0.5948	0.7004
MAD-GAN	0.8431	0.7085	0.8966	0.6325	0.7702
OmniAnomaly	0.7344	0.7927	0.9628	0.8344	0.8311
DSANet	0.8924	0.8739	0.9630	0.8740	0.9008
USAD	0.8227	0.4275	0.9024	0.7987	0.7378
VAEpro	0.8369	0.8200	0.8693	0.8522	0.8446
InterFusion	**0.9280**	**0.9103**	**0.9817**	**0.9531**	**0.9433**

Experimental Results-Interpretation Score

Table 2: Interpretation IPS for *InterFusion* and baselines.

Methods	SMD	ASD	Avg.
LSTM-NDT	0.5751	0.8619	0.7185
MSCRED	0.6421	0.7652	0.7037
OmniAnomaly	0.8008	0.8029	0.8019
DSANet	0.6713	0.8123	0.7418
VAEpro	0.5681	0.8236	0.6959
VAEpro*	0.7433	0.8916	0.8175
InterFusion-nI	0.7752	0.8881	0.8317
InterFusion	**0.8340**	**0.9107**	**0.8724**

No MCMC imputation
Original reconstruction probability

Experimental Results-Ablation Study

Model Variant	Description
TimeVAE	Only Temporal
m-SRNN	Only Intermetric
IF-p	No HVAE
IF-s	No Two-view Embeddings
IF-x	No Prefiltering Strategy
PureAE	AutoEncoder
IF-AERNN	AutoEncoder + RNN

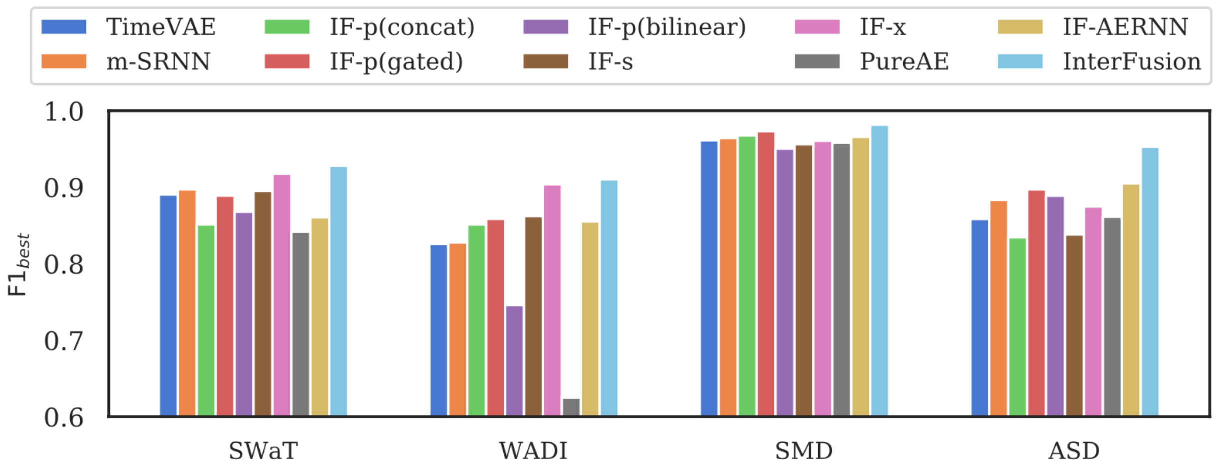


Figure 4: Average anomaly detection best-F1 for *InterFusion* and its variants. 'IF' denotes *InterFusion* for short.

Conclusion

- **Capturing both temporal and intermetric dependencies**
  - HVAE, Two-view Embedding
- **Filtering noise and anomalies in training data**
  - Prefiltering Strategy
- **Enhancing interpretability**
  - MCMC-based Interpretation

Thank you