Towards a Generalized Framework for Anomaly Detection Encompassing Different Risk Levels and Supervision Settings

Rulan Wei[1], Zewei He[1], Martin Pavlovski[2], Fang Zhou[1]*

[1]School of Data Science and Engineering, East China Normal University, China
[2]Temple University, Philadelphia, PA, USA

* Corresponding author

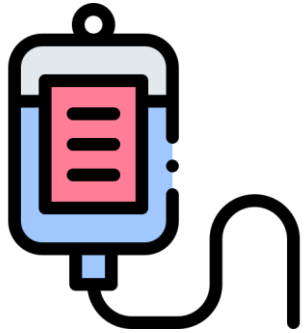
Background and Motivation

- **Anomaly detection (AD):** Identifying data objects significantly deviating from the majority of the data.
- **Applications:**

Fraud detection in finance [3, 28]

Risks management in banking [10]

Safe-guarding against network intrusions in cybersecurity [38]



Disease detection in healthcare [22]

[3] Cao, B. et al. (2018). Collective fraud detection capturing inter-transaction dependency. In *KDD 2017 Workshop on Anomaly Detection in Finance. PMLR*, 66–75.

[10] Lee, M. C. et al. (2020). Autoaudit: Mining accounting and time-evolving graphs. In *2020 IEEE International Conference on Big Data (Big Data). IEEE*, 950–956.

[22] Ruff, L. et al. (2021). A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5), 756-795.

[28] Tao, J. et al. (2019) Mvan: Multi-view attention networks for real money trading detection in online games. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Minin*g. 2536–2546.

[38] Zhang, S. et al. (2015). A novel anomaly detection approach for mitigating web-based attacks against clouds. In *2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing* (pp. 289-294). IEEE.

Background and Motivation

Existing approaches:

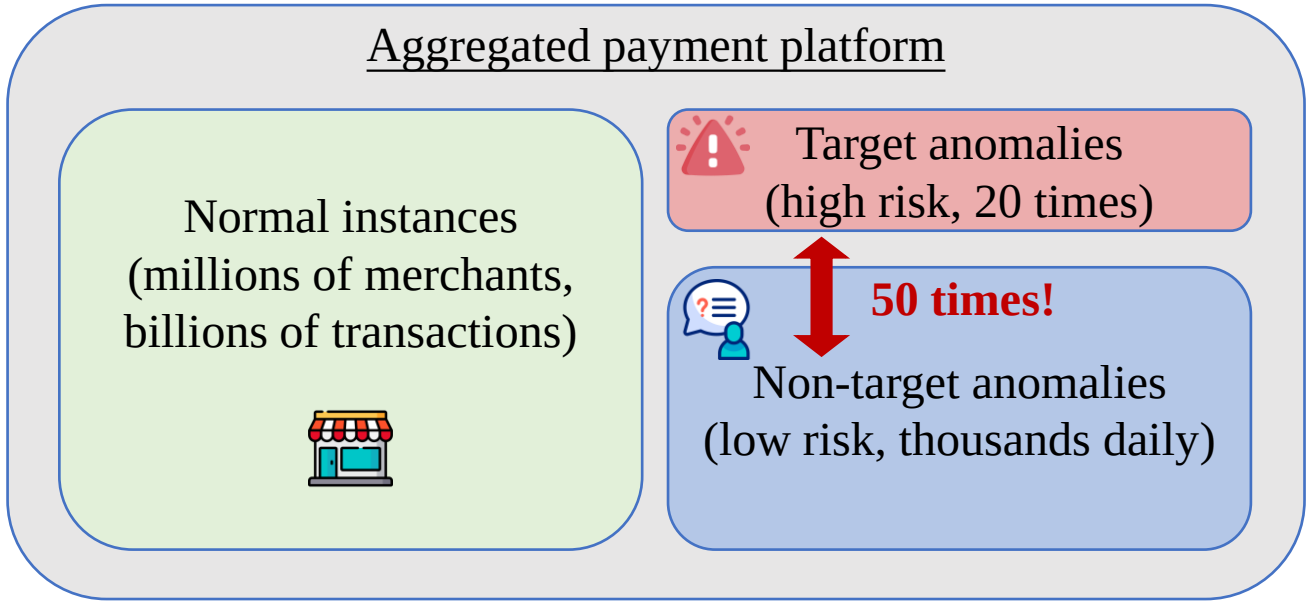
- Unsupervised AD methods: OC-SVM, Isolation forest, Deep SVDD, etc.
- Semi-supervised AD methods: Deep SAD, DevNet, the Kernel-based method, PIA-WAL, etc.

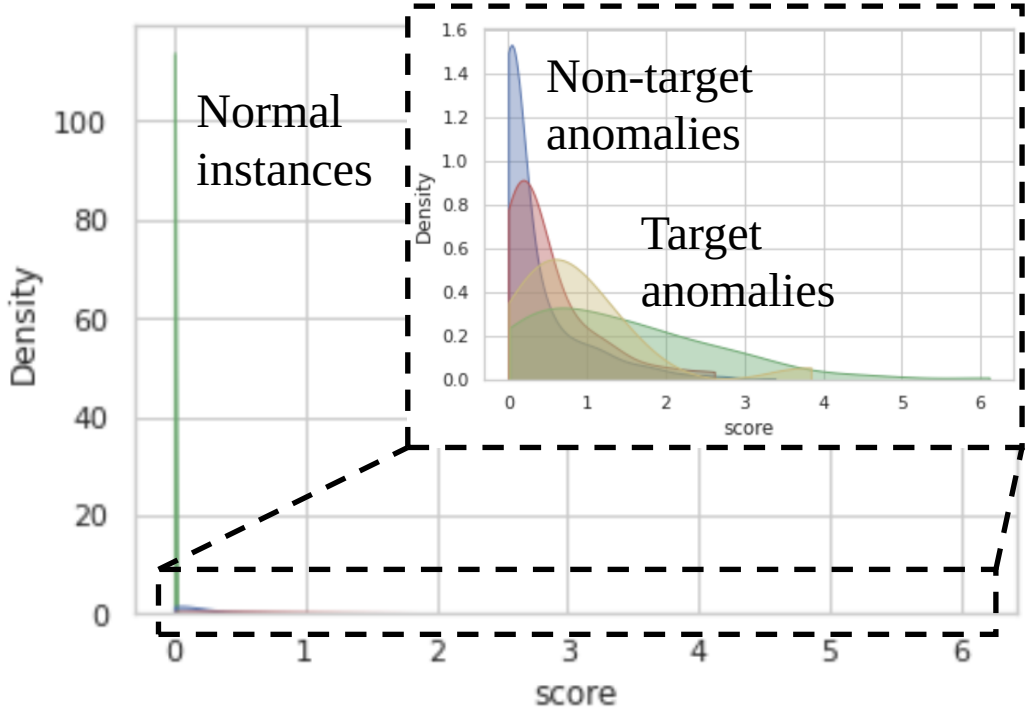
Limitations:

- Existing approaches aim to identify anomalies **uniformly** while **ignoring any priority** among anomalies
- However, in **real-world scenarios**, anomalies often **exhibit distinct levels of priority**

Example:

- Risk control scenario of an aggregated payment platform





Background and Motivation

Motivation:

- Given the **efficiency constraints** of risk management and **scarcity and high cost of human resources**, the **precise identification of high-risk anomalies** becomes imperative, while adopting a more **permissive stance toward low-risk behaviors** is deemed a preferable strategy.

Two straightforward solutions failed:

- Semi-supervised AD method (separate all types of anomalies from normal instances) → supervised binary classifier (differentiate target from non-target anomalies)
- Supervised three-class classification approach (distinguish between normal instances, non-target and target anomalies)

Challenges:

- **Labels for all types of non-target anomalies are challenging to obtain** as they are not of primary interest and hence are rarely labeled. In some scenarios, we only have access to labeled target anomalies of interest.
  ⇨ Maximizing the labeled data utilization is of paramount importance in such scenarios.

- **Class overlap occurs among target anomalies, non-target anomalies, and marginal normal instances.**
  ⇨ Effectively differentiating them to minimize the FPR is key to precise identification of target anomalies.

Related Work

<table>
<tr><th>Categories</th><th colspan="2">Models</th><th>Shortcomings</th></tr>
<tr><td rowspan="2">Unsupervised methods</td><td>classical methods</td><td>iForest (TKDD 2012)
OC-SVM (Neural computation 2001)</td><td rowspan="2">Suffer from limited inclusion of prior knowledge
⇨ leading to high FPR</td></tr>
<tr><td>deep learning-based methods</td><td>DeepSVDD (ICML 2018
DeepiForest (TKDE 2023)
ECOD (TKDE 2022)</td></tr>
<tr><td rowspan="3">Semi-supervised methods</td><td>leverage labeled anomalies to enhance performance</td><td>REPEN (SIGKDD 2018)
Deep SAD (ICLR 2020)
DevNet (SIGKDD 2019)</td><td rowspan="3">• Treat all anomalies equally
• Identify all types of anomalies into a single class.
⇨ May result in high FPR when only a specific subset of nomalies requires accurate identification.</td></tr>
<tr><td>mitigate the negative influence of noisy instances</td><td>Elite (SIGKDD 2021)
Kernel-based Method (IJCAI 2022)
ADMoE (AAAI 2023)</td></tr>
<tr><td>emphasis the marginal instances</td><td>PIA-WAL (DASFAA 2022)</td></tr>
<tr><td>Weakly-supervised methods</td><td>detect unseen/unknown anomalies</td><td>AABiGAN (IJCAI 2022)
DPLAN (SIGKDD 2023)</td><td></td></tr>
</table>

Problem Statement

Generalized Anomaly Detection:

- Unlabeled dataset: $\mathcal{D}^u = \{x_1^u, ..., x_{n_0}^u\} \in \mathcal{X}$
- Anomalies (M classes): decreasing priority → $1, 2, \dots, k, k+1, \dots, M$
  - $1, 2, \dots, k$: target anomalies
  - $k+1, \dots, M$: non-target anomalies

Problem	Goal	Data accessibility
1. $k = M$: conventional AD problem	y = +1 ➡ anomaly of any classes y = -1 ➡ normal instance	Labels of $\mathcal{D}^M$ are available.
2. $k < M$: partial AD problem	y = +1 ➡ anomaly belongs to first k classes y = -1 ➡ normal instance or anomalies not among first k classes	Labels of $\mathcal{D}^k \subseteq \mathcal{D}^M$ are available.
		✔ *2.1* fully-supervised partial AD: additional $M - k$ anomaly classes
		✘ *2.2* semi-supervised partial AD: additional $M - k$ anomaly classes

* Note that here the supervision refers exclusively to the availability of labels for the anomalies, not the normal instances.

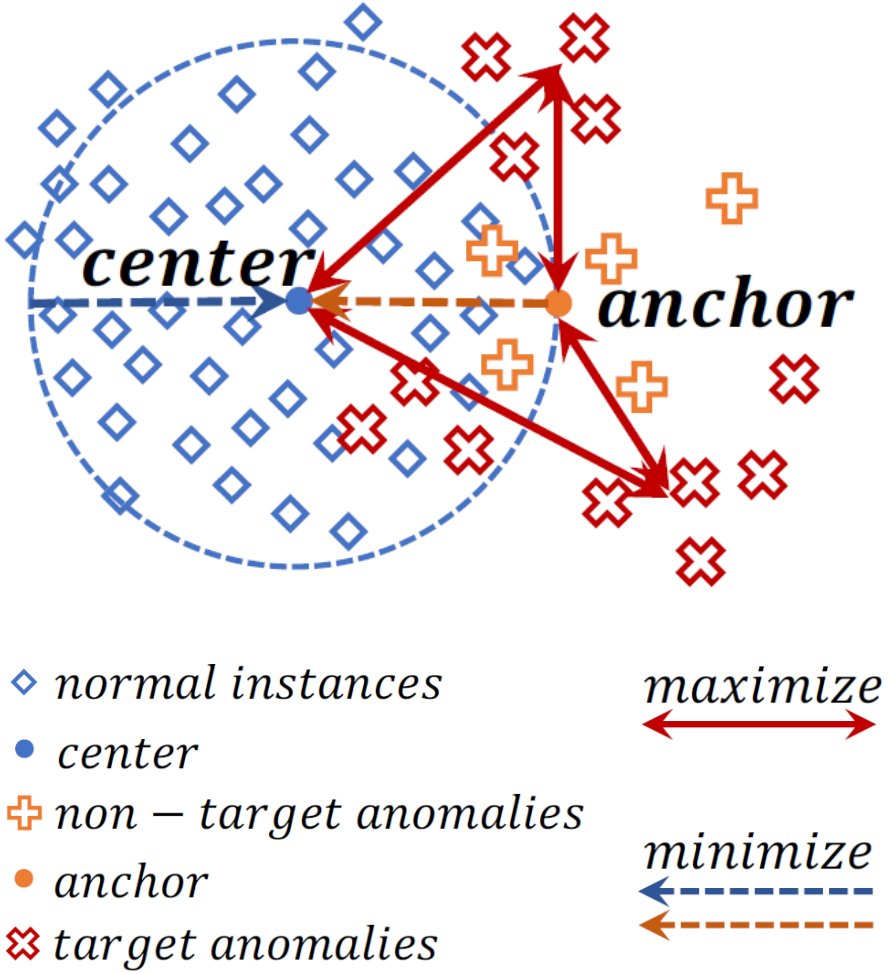
Proposed Method

The Dual-Center Mechanism

- Designed to maximize the distinction between normal and (target) anomalies, and moreover, between (target) anomalies and marginal instances (or non-target anomalies).

$$\boldsymbol{c} = \frac{\sum_{i=1}^{|\mathcal{D}^u|} \Phi(\boldsymbol{x}_i; \mathcal{W})}{|\mathcal{D}^u|}, \boldsymbol{x}_i \in \mathcal{D}^u$$

$$\boldsymbol{a} = \frac{\sum_{i=1}^{|\mathcal{D}^a|} \Phi(\boldsymbol{x}_i; \mathcal{W})}{|\mathcal{D}^a|}, \quad \boldsymbol{x}_i \in \mathcal{D}^a.$$



- The auxiliary data D^a consists of either labeled non-target anomalies or unlabeled marginal instances.

- $$L_{compact} = \sum_{i=1}^{|\mathcal{D}^u|} \|\Phi(\boldsymbol{x}_i; \mathcal{W}) - \boldsymbol{c}\|^2 + \eta \sum_{j=1}^{|\mathcal{D}^a|} \|\Phi(\boldsymbol{x}_j; \mathcal{W}) - \boldsymbol{c}\|^2$$

Proposed Method

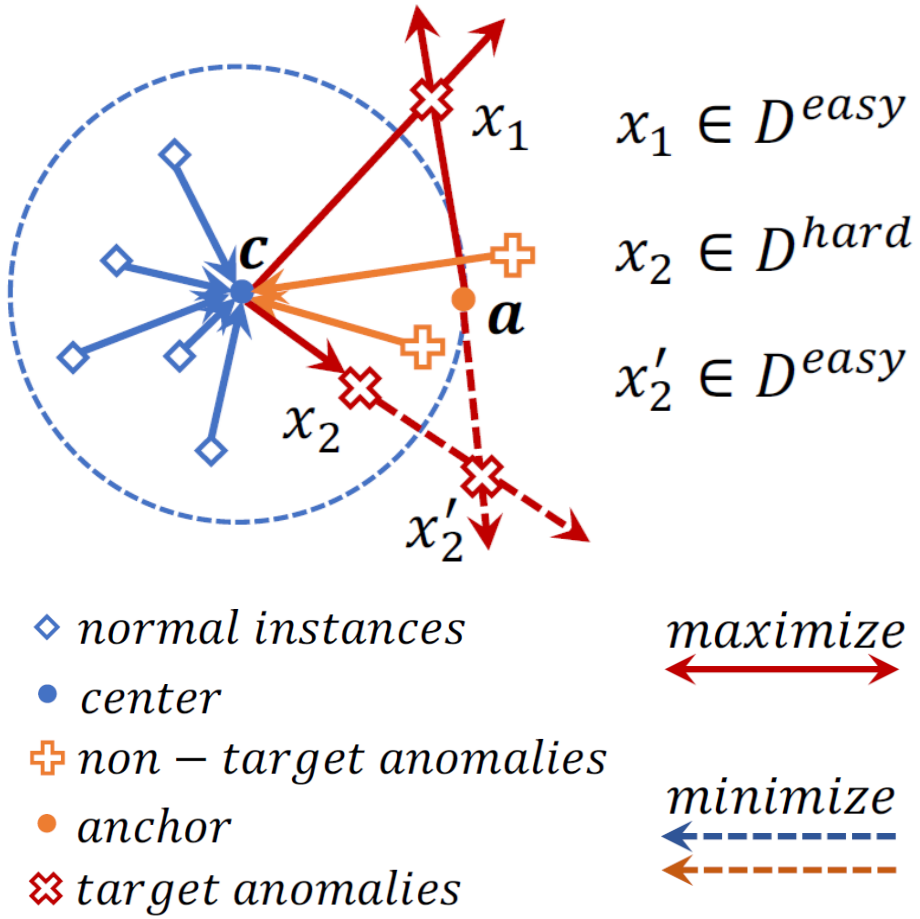
- The set of labeled target anomalies D^k is partitioned into easy-to-identify set D^{easy} and hard-to-identify set D^{hard} based on their respective positions in the latent space.

$$\mathcal{D}^{easy} = \{\boldsymbol{x} | dist(\Phi(\boldsymbol{x};\mathcal{W}), \boldsymbol{c}) \geq dist(\boldsymbol{a}, \boldsymbol{c}), \boldsymbol{x} \in \mathcal{D}^k\},$$

$$\mathcal{D}^{hard} = \{\boldsymbol{x} | dist(\Phi(\boldsymbol{x};\mathcal{W}), \boldsymbol{c}) < dist(\boldsymbol{a}, \boldsymbol{c}), \boldsymbol{x} \in \mathcal{D}^k\},$$

$$\mathcal{D}^{easy} \cup \mathcal{D}^{hard} = \mathcal{D}^t, \quad \mathcal{D}^{easy} \cap \mathcal{D}^{hard} = \emptyset,$$

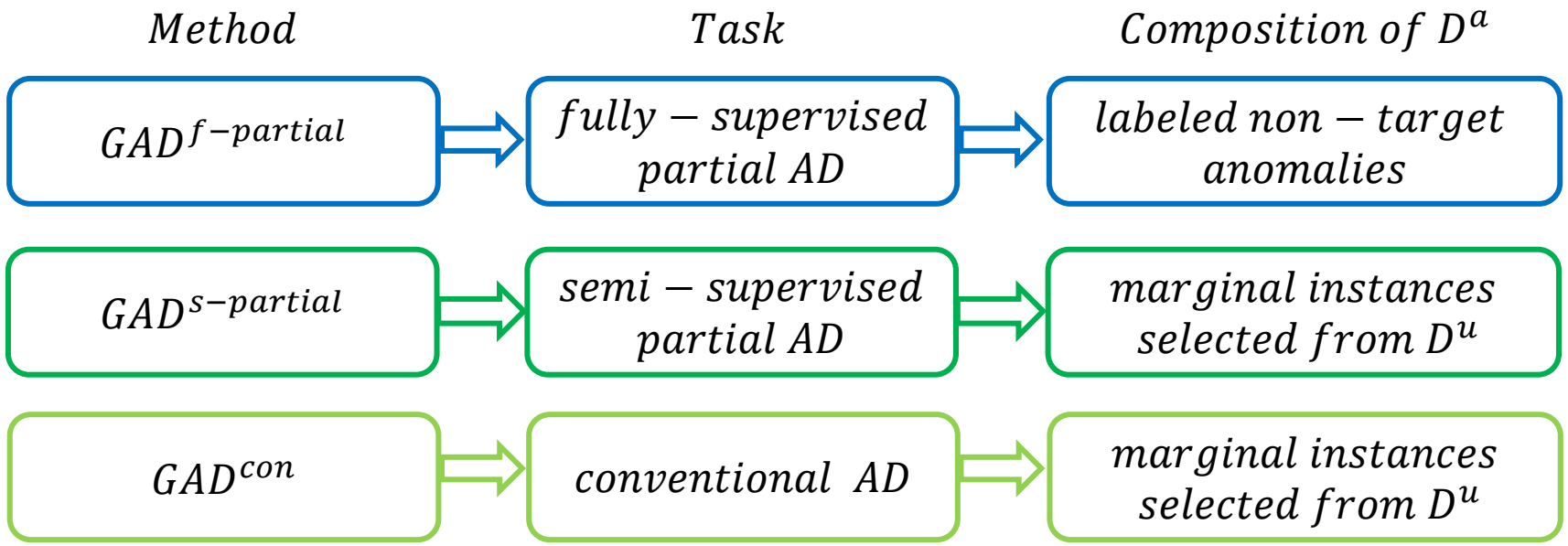
$$L_{target} = \sum_{k=1}^{|\mathcal{D}^{easy}|} \left(\frac{1}{\|\Phi(\boldsymbol{x}_k;\mathcal{W}) - \boldsymbol{c}\|^2} + \frac{1}{\|\Phi(\boldsymbol{x}_k;\mathcal{W}) - \boldsymbol{a}\|^2} \right) + \sum_{l=1}^{|\mathcal{D}^{hard}|} \frac{1}{\|\Phi(\boldsymbol{x}_l;\mathcal{W}) - \boldsymbol{c}\|^2}.$$



- **Overall objective function**: $\min_{\mathcal{W}} \frac{L_{compact} + L_{target}}{|\mathcal{D}^u| + |\mathcal{D}^a| + |\mathcal{D}^k|}$
- **Anomaly score:** $s(\boldsymbol{x}) = \|\Phi(\boldsymbol{x};\mathcal{W}) - \boldsymbol{c}\|^2$

Proposed Method

Three GAD variants and corresponding tasks



- $\boldsymbol{GAD^{f-partial}}$: D^a consists of labeled non-target anomalies.
- $\boldsymbol{GAD^{s-partial}}$: D^a is composed of marginal instances selected from D^u
- $\boldsymbol{GAD^{con}}$: D^a is composed of marginal instances selected from D^u

Experimental Setting

Datasets:

dataset		training set			validation set			testing set		
dataset name	d	unlabeled ($\mathcal{D}^u$)	target ($\mathcal{D}^k$)	non-target ($\mathcal{D}^a$)	normal	target	non-target	normal	target	non-target
UNSW_NB15	196	57,318	300(3)	400(4)	18,600	1,666(3)	2,335(4)	18,600	1,666(3)	2,335(4)
FMNIST[1], FMNIST[2]	28×28	5,100	100(1)	100(1)	1,000	100(1)	100(1)	1,000	100(1)	100(1)
FMNIST[3]	28×28	5,100	100(1)	0	1,000	100(1)	0	1,000	100(1)	0
SQB	182	134,299*	205(3)	205(5)	33,575*	41(3)	41(5)	148,323*	129(3)	463(5)

The number of distinct categories present in a dataset is surrounded with "()".

* Since normal instances are not available in the SQB dataset, we consider the unlabeled instances as normal for validation and testing.

- Default contamination ratio: 2%
- Labeled data proportion: 0.31%-4.05%

Competing methods:

- **Unsupervised** methods: OC-SVM (Neural Computation 2001), Isolation forest (ICDM 2008), Deep SVDD (PMLR 2018)
- **Semi-supervised** methods: Deep SAD (ICLR 2020), DevNet (KDD 2019), the Kernel-based method (IJCAI 2022), PIA-WAL (DASFAA 2022)
- **Fully-supervised** methods: Deep SAD + RF (2 classes), RF (3 classes)

Experiments

Effectiveness on Real-world Datasets

Model	use of labeled non-targ.	AUPRC					AUROC				
		Partial AD				Conventional AD	Partial AD				Conventional AD
		UNSW_NB15	SQB	FMNIST[1]	FMNIST[2]	FMNIST[3]	UNSW_NB15	SQB	FMNIST[1]	FMNIST[2]	FMNIST[3]
DeepSVDD	×	47.7±2.76	0.37±0.18	21.71±1.02	18.7±1.93	22.61±3.08	93.3±0.51	66.25±15.46	78.17±1.8	61.49±2.2	62.59±3.33
iForest	×	36.24±7.49	1.63±0.37	25.5±2.53	10.67±0.51	15.79±0.98	83.97±1.7	90.92±0.61	86.39±1.61	58.48±1.02	63.77±0.98
OC-SVM	×	30.93±0.0	1.03±0.0	14.69±0.0	11.14±0.0	15.63±0.0	88.82±0.0	84.98±0.0	74.42±0.0	59.11±0.0	63.27±0.0
DeepSAD	×	72.24±1.04	23.0±0.98	94.78±1.32	64.58±4.68	70.68±4.03	96.35±0.11	97.57±0.48	98.29±0.75	89.0±1.69	90.46±2.12
DevNet	×	65.71±1.42	14.89±0.89	94.38±1.52	44.08±6.45	57.87±3.96	94.95±0.3	97.36±0.84	98.56±0.5	81.98±2.26	85.94±1.95
Kernel-Based	×	68.58±5.39	2.53±0.52	88.97±2.47	44.83±8.5	44.69±4.61	94.57±0.34	85.11±0.61	97.12±0.72	75.06±5.26	75.0±3.1
PIA-WAL	×	72.2±2.08	18.58±1.02	64.31±10.8	19.46±8.57	35.12±7.42	95.65±0.14	96.14±1.0	88.85±3.2	68.32±8.69	75.58±2.94
GAD$^{s-partial}$	×	**74.74±1.1**	**30.23±1.12**	**97.1±0.3**	**72.6±1.04**	-	**97.3±0.07**	**98.74±0.72**	**99.31±0.2**	**92.12±0.62**	-
RF (3 classes)	✓	78.33±0.18	19.21±1.02	91.43±0.51	52.28±1.77	-	97.52±0.05	97.15±0.52	98.58±0.11	86.14±0.19	-
Deep SAD+RF (2 classes)	✓	54.31±3.10	1.73±0.85	76.82±6.64	14.74±2.73	-	13.25±0.64	0.27±0.01	15.51±2.50	11.84±1.41	-
GAD$^{f-partial}$	✓	**79.13±0.21**	**33.84±2.9**	**96.57±0.35**	**76.09±1.0**	-	**97.62±0.05**	**98.77±0.78**	**99.31±0.15**	**92.71±0.4**	-
RF (2 classes)	-	-	-	-	-	43.38±2.11	-	-	-	-	84.23±0.69
GAD^{con}	-	-	-	-	-	**78.29±1.3**	-	-	-	-	**93.24±0.57**

Observations:

- ***GAD$^{f-partial}$*** yields improvements of **0.8%-61.35%** over **fully-supervised baselines**.
- ***GAD$^{s-partial}$*** attains lifts ranging from **2.5% to 53.14%** relative to the **semi-supervised baselines**.
- ***GAD^{con}*** surpasses **all baselines** with lifts ranging from **7.61% to 62.66%**.

Experiments

Effect of Overlap Degree on AD Performance

(on different class combinations of FMNIST)

Left panel: UMAP embeddings of the FMNIST datasets.

- N: Normal
- T: Target
- 1-4: (non-target) anomaly selections

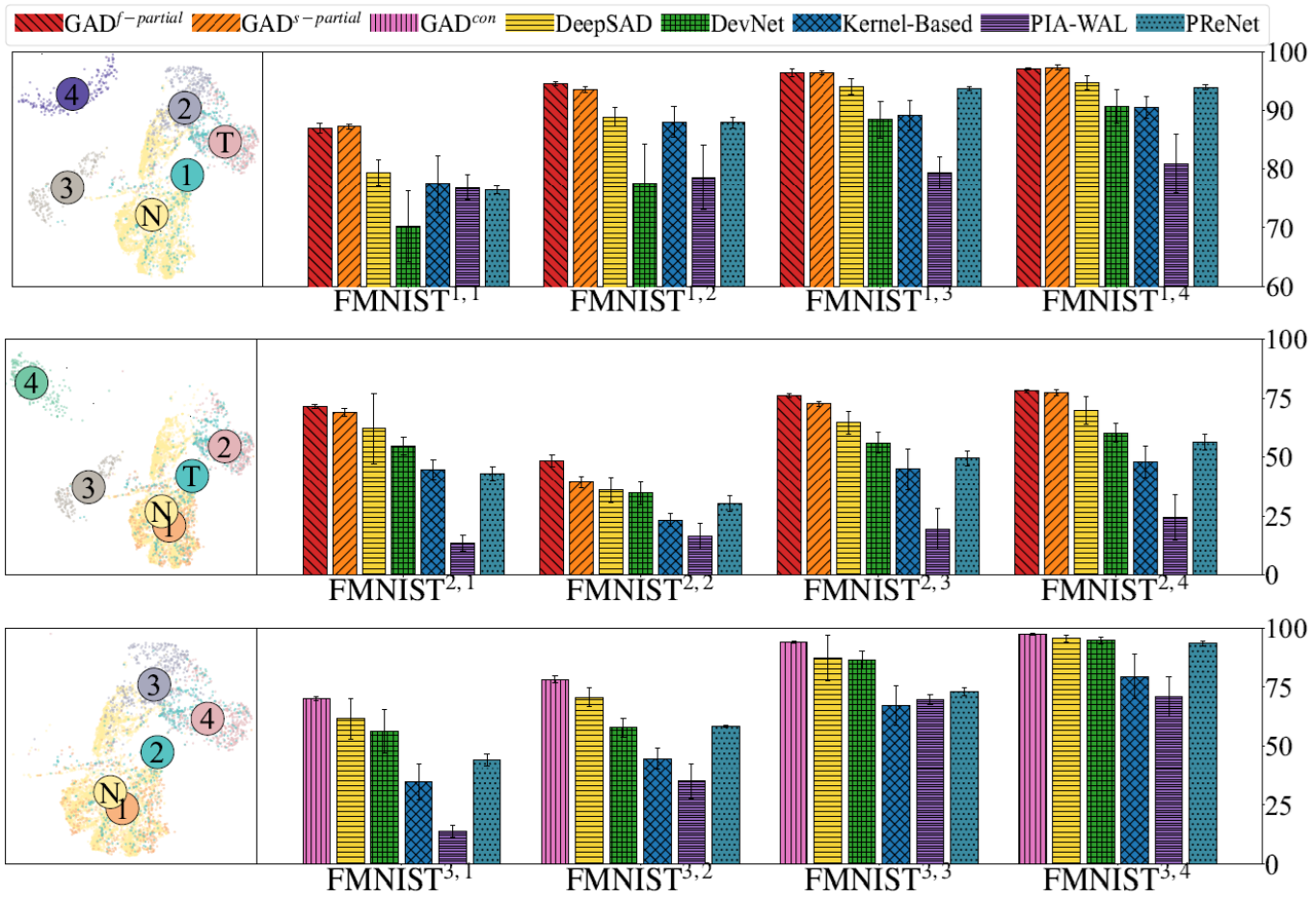
Right panel: AUPRC ± standard deviation

Partial AD scenarios:

- FMNIST¹ and FMNIST²: sorted by the ascending inter-class distance between non-target and target anomalies.

Conventional AD scenarios

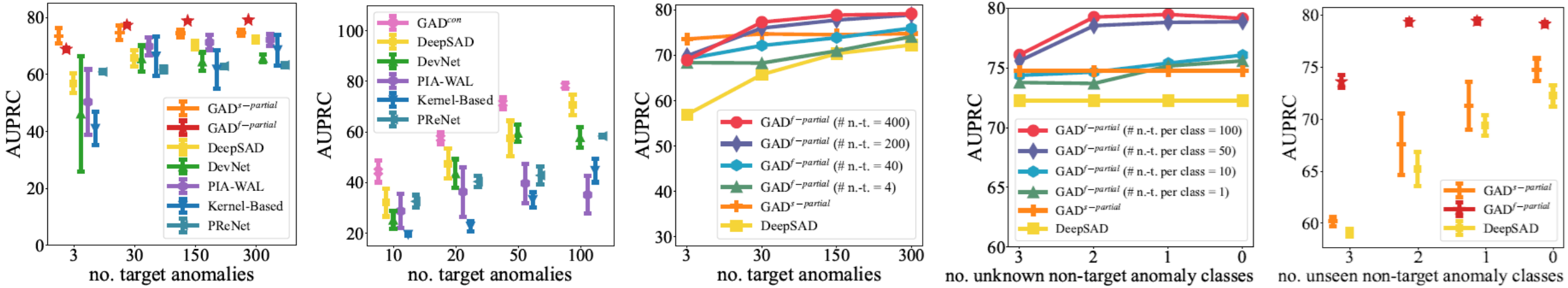
- FMNIST³ is sorted by the anomalies inter-class distance between anomaly and normal classes.



⇨ ***Observation:*** With **no assumption for the relative positions** of normal instances, target and non-target anomalies, GAD consistently and effectively identifies target anomalies.

Experiments

Effectiveness under Different Quantities of Labeled Anomalies



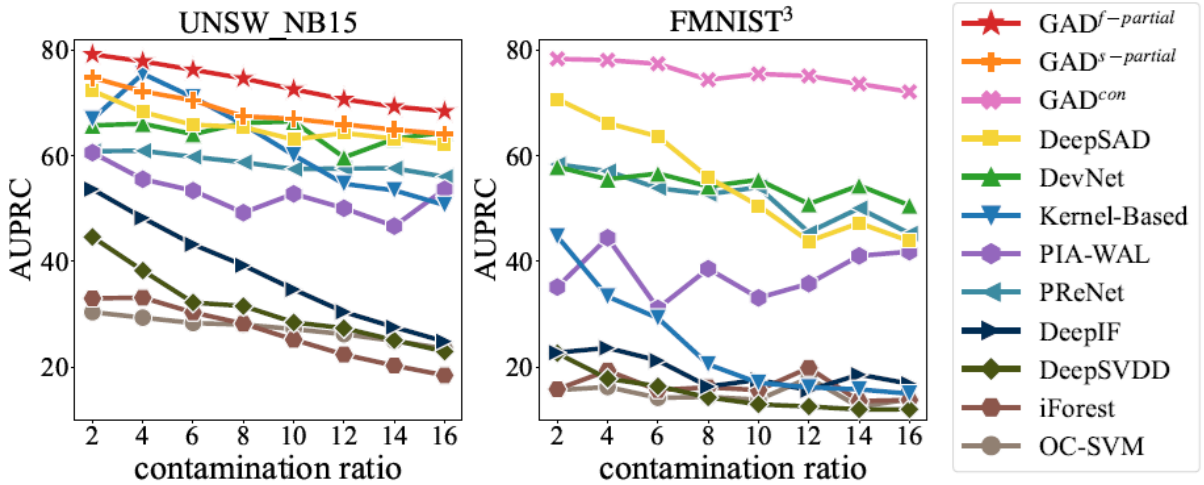
(a) partial AD (b) conventional AD (c) partial AD (d) partial AD (e) partial AD

Observations:

- All GAD variants significantly outperform the semi-supervised baselines across all levels of labeled target anomalies (a, b).
- Even without knowledge of all non-target anomaly classes, $\boldsymbol{GAD^{f-partial}}$ exhibits significant lifts over the baseline (c, d).
- $\boldsymbol{GAD^{f-partial}}$ demonstrates strong robust performance across unseen non-target anomaly classes (e).
- All GAD variants showcases **a substantial improvement in data utilization efficiency**.

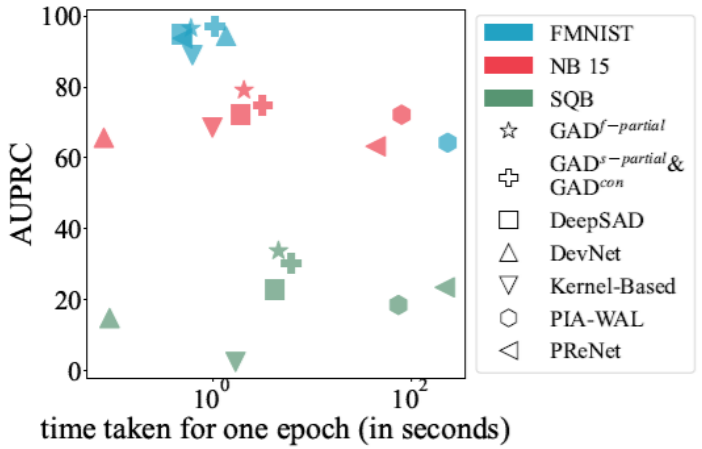
Experiments

Robustness under Anomaly Contamination



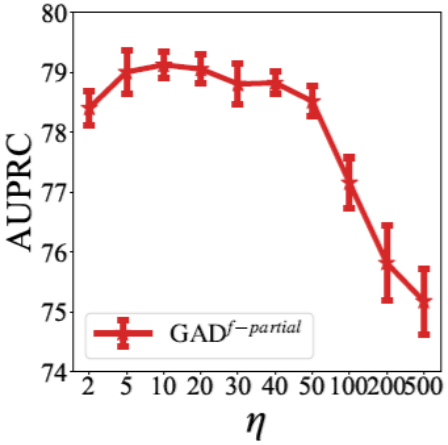
(a)

Time Efficiency



(b)

Sensitivity Analysis



(c)

Observations:

- All GAD variants maintain a stable and superior AUPRC, consistently outperforming all baselines across all contamination ratios.
- GAD achieves significant detection performance improvement with acceptable extra complexity.
- GAD is not sensitive to the hyperparameter η.

Conclusion

- This work emphasizes the concept of **priority among anomalies** and precisely identifies anomalies of primary interest to meet real-world requirements.

- We address a generalized anomaly detection problem which covers a **broader and more practical range** of real-world scenarios.

  ⇨ Proposed an ‘umbrella’ (all-encompassing) framework GAD that addresses different AD scenarios.

- GAD and its variants demonstrate **notable performance** as well as **high utilization of labeled data**.

  ⇨ Significantly reduce the FPR caused by class overlaps, insufficient and incomplete labeled data, and positions of non-target anomalies.

- Code link: https://github.com/ZhouF-ECNU/GAD