Understanding the Effect of Bias in Deep Anomaly Detection

Ziyu Ye, Yuxin Chen, Haitao Zheng

THE UNIVERSITY OF CHICAGO

What is Anomaly Detection (AD)?

What is Anomaly Detection (AD)?

- AD is critical in many real-world applications.

  e.g., intrusion detection, fraud detection, adversarial attacks, medical diagnosis, time series analysis, system monitoring, …

What is Anomaly Detection (AD)?

- AD is critical in many real-world applications.

  e.g., intrusion detection, fraud detection, adversarial attacks, medical diagnosis, time series analysis, system monitoring, …

- **Basic assumption**: limited knowledge for anomalies.

What is Anomaly Detection (AD)?

- AD is critical in many real-world applications.

  e.g., intrusion detection, fraud detection, adversarial attacks, medical diagnosis, time series analysis, system monitoring, …

- **Basic assumption**: limited knowledge for anomalies.
  - *Few* samples of anomalies in training;

What is Anomaly Detection (AD)?

- AD is critical in many real-world applications.

  e.g., intrusion detection, fraud detection, adversarial attacks, medical diagnosis, time series analysis, system monitoring, …

- **Basic assumption**: limited knowledge for anomalies.
  - *Few* samples of anomalies in training;
  - *Unexpectable* target distribution of anomalies.

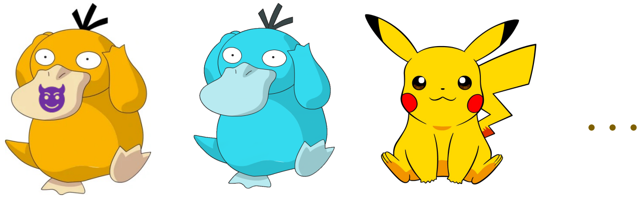
What is Anomaly Detection (AD)?

(An illustrative example of Pokémon.)

Normal

Abnormal

Known types in source distribution



Unknown types in target distribution

- AD is critical in many real-world applications.

  e.g., intrusion detection, fraud detection, adversarial attacks, medical diagnosis, time series analysis, system monitoring, …

- **Basic assumption**: limited knowledge for anomalies.
  - *Few* samples of anomalies in training;
  - *Unexpectable* target distribution of anomalies.

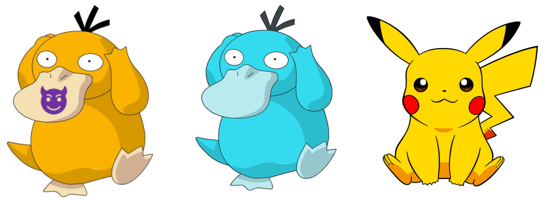
What is Anomaly Detection (AD)?

(An illustrative example of Pokémon.)

Normal

Abnormal

Known types in source distribution



. . .

Unknown types in target distribution

- AD is critical in many real-world applications.

  e.g., intrusion detection, fraud detection, adversarial attacks, medical diagnosis, time series analysis, system monitoring, …

- **Basic assumption**: limited knowledge for anomalies.
  - *Few* samples of anomalies in training;
  - *Unexpectable* target distribution of anomalies.

➜ Challenging to get a *representative anomaly set*.

Existing Approaches for AD

Existing Approaches for AD

Low High
Anomaly Score

Normal Data
Abnormal Data

Semi-Supervised (AD)

[AC15], [ZP17], [Ruf+18], [Zon+18], [Goy+20], ...

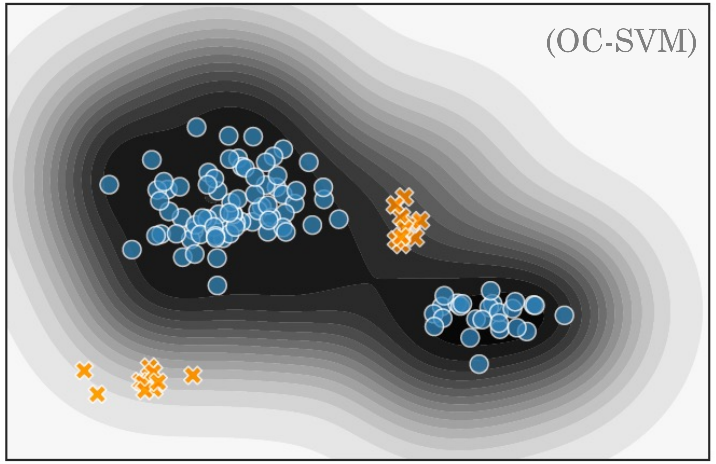


Image Source: Ruff, Lukas, et al. "Deep semi-supervised anomaly detection." In *proc. of ICLR*, 2020.

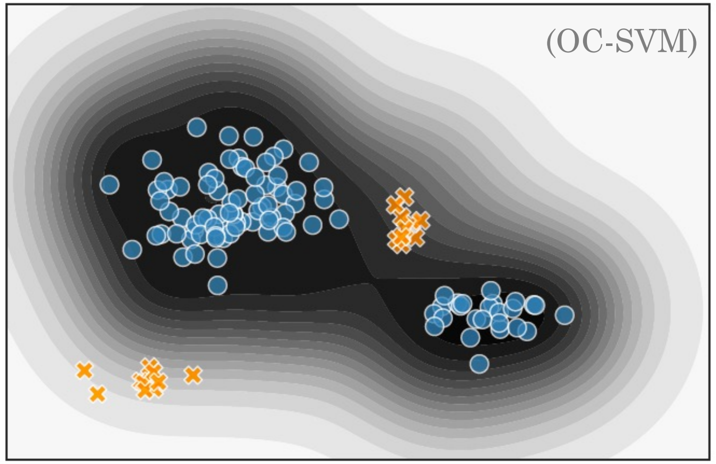
Existing Approaches for AD

Low *High*
Anomaly Score

Normal Data
Abnormal Data

Semi-Supervised (AD)

[AC15], [ZP17], [Ruf+18], [Zon+18], [Goy+20], ...



Pro A compact enclosing of the normal.

Image Source: Ruff, Lukas, et al. "Deep semi-supervised anomaly detection." In *proc. of ICLR*, 2020.

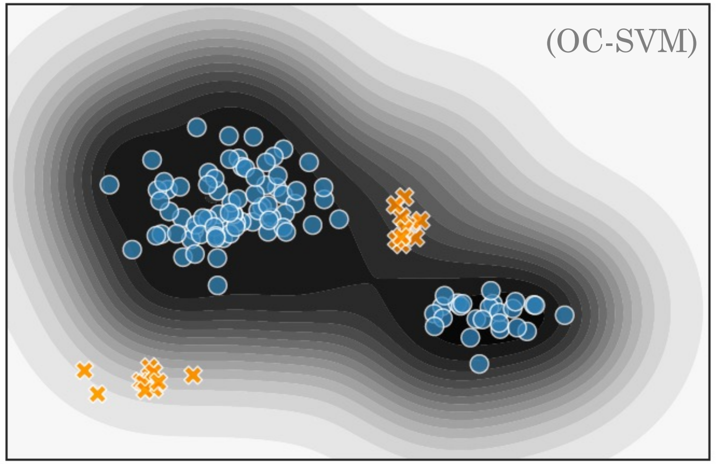
Existing Approaches for AD

Low *High*
Anomaly Score

Normal Data
Abnormal Data

Semi-Supervised (AD)

[AC15], [ZP17], [Ruf+18], [Zon+18], [Goy+20], ...



Pro A compact enclosing of the normal.

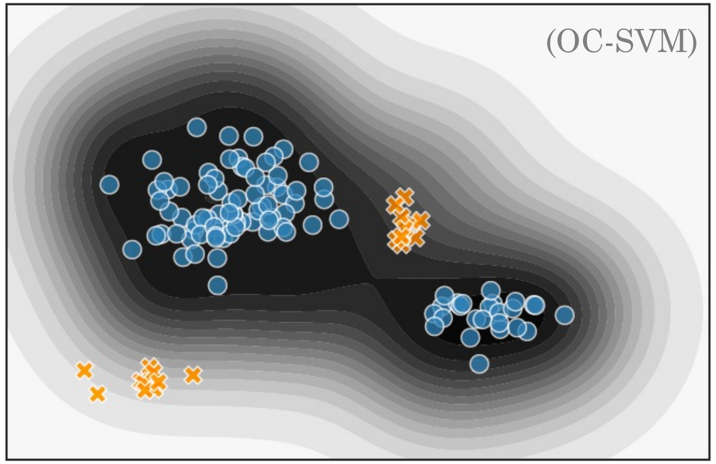
Con Unable to identify *hard anomalies*.
→ Underfitting bias.

Image Source: Ruff, Lukas, et al. "Deep semi-supervised anomaly detection." In *proc. of ICLR*, 2020.

Existing Approaches for AD

Semi-Supervised (AD)

[AC15], [ZP17], [Ruf+18], [Zon+18], [Goy+20], ...



Pro A compact enclosing of the normal.

Con Unable to identify *hard anomalies*.
→ Underfitting bias.

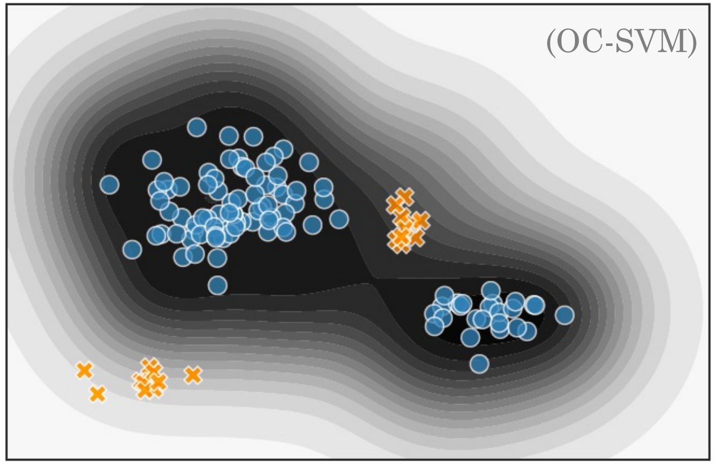
Can we make use of *additional labeled anomalies*?

Image Source: Ruff, Lukas, et al. "Deep semi-supervised anomaly detection." In *proc. of ICLR*, 2020.

Existing Approaches for AD

Semi-Supervised (AD)

[AC15], [ZP17], [Ruf+18], [Goy+20], ...



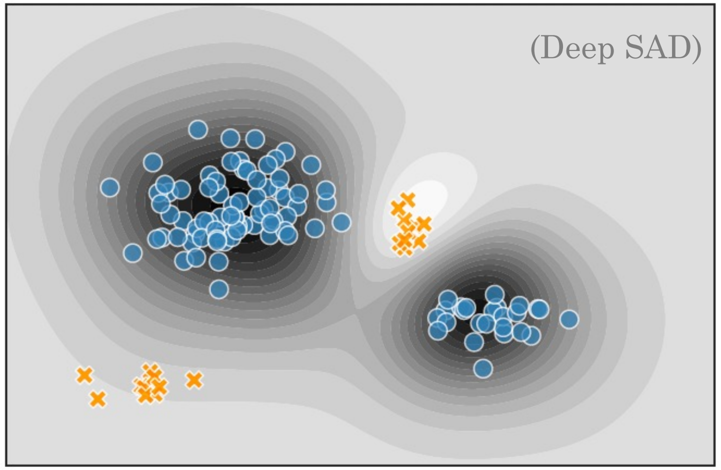
Pro A compact enclosing of the normal.

Con Unable to use *additional labels*.
→ Underfitting bias.

Train with *additional labeled anomalies* →

Supervised (AD)

[Pan+19], [Yam+19], [Hen+19], [Ruf+20], [Goy+20], ...



Pro A compact enclosing of the normal.
+
Discriminating on *known* anomalies.

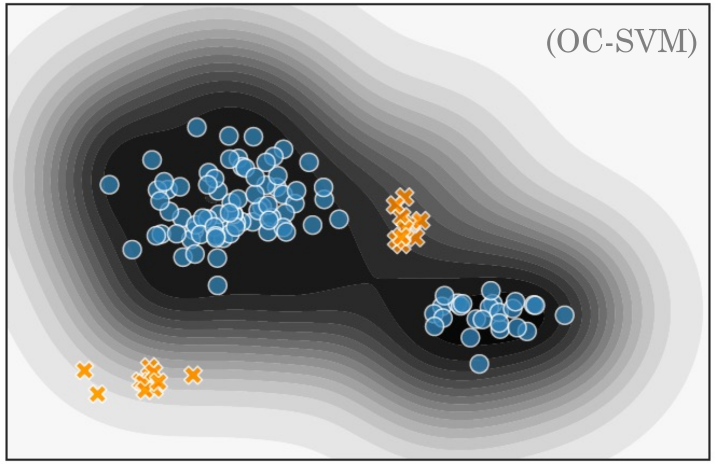
Image Source: Ruff, Lukas, et al. "Deep semi-supervised anomaly detection." In *proc. of ICLR*, 2020.

Existing Approaches for AD

Low — *High*
Anomaly Score
Normal Data
Abnormal Data

Semi-Supervised (AD)

[AC15], [ZP17], [Ruf+18], [Goy+20], ...



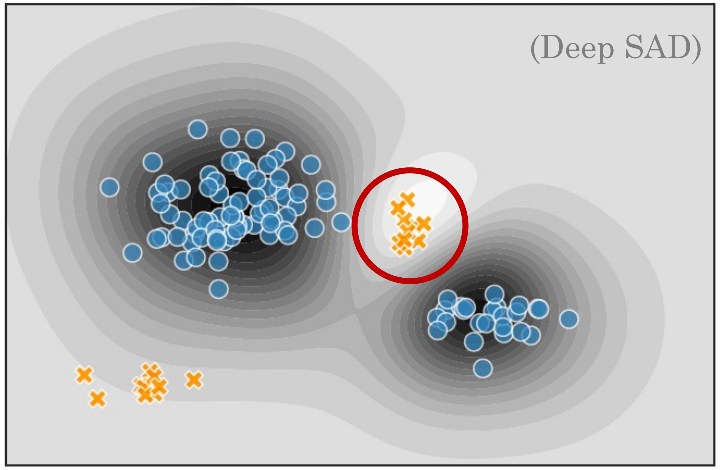
Pro A compact enclosing of the normal.

Con Unable to use *additional labels*.
→ Underfitting bias.

Train with *additional labeled anomalies* →

Supervised (AD)

[Pan+19], [Yam+19], [Hen+19], [Ruf+20], [Goy+20], ...



Pro A compact enclosing of the normal.
+
Discriminating on *known* anomalies.

Image Source: Ruff, Lukas, et al. "Deep semi-supervised anomaly detection." In *proc. of ICLR*, 2020.

Motivation

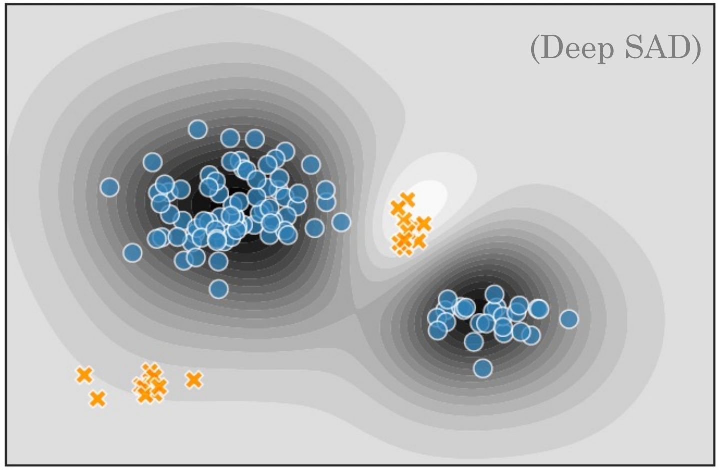
Low High
Anomaly Score

Normal Data
Abnormal Data

Will *additional labels* do *harm*?

Supervised (AD)

[Pan+19], [Yam+19], [Hen+19], [Ruf+20], [Goy+20], ...



Pro A compact enclosing of the normal.

+

Discriminating on *known* anomalies.

Image Source: Ruff, Lukas, et al. "Deep semi-supervised anomaly detection." In *proc. of ICLR*, 2020.

Motivation

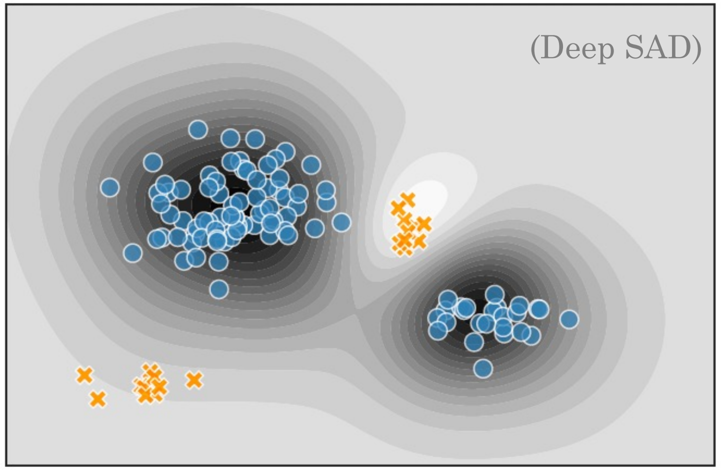
Low High
Anomaly Score
Normal Data
Abnormal Data

Will *additional labels* do *harm*?

Can **unseen anomalies** suffer from **bias***?

Supervised (AD)

[Pan+19], [Yam+19], [Hen+19], [Ruf+20], [Goy+20], ...



Pro A compact enclosing of the normal.
+
Discriminating on *known* anomalies.

* Note that such bias is novel compared to the aforementioned in literature (c.f. Section 2 of our paper).
Image Source: Ruff, Lukas, et al. "Deep semi-supervised anomaly detection." In *proc. of ICLR*, 2020.

A Counter-Intuitive Example

Training with *additional labeled anomalies* can bring *disastrous harmful bias*.

A Counter-Intuitive Example

Training with *additional labeled anomalies* can bring *disastrous harmful bias*.

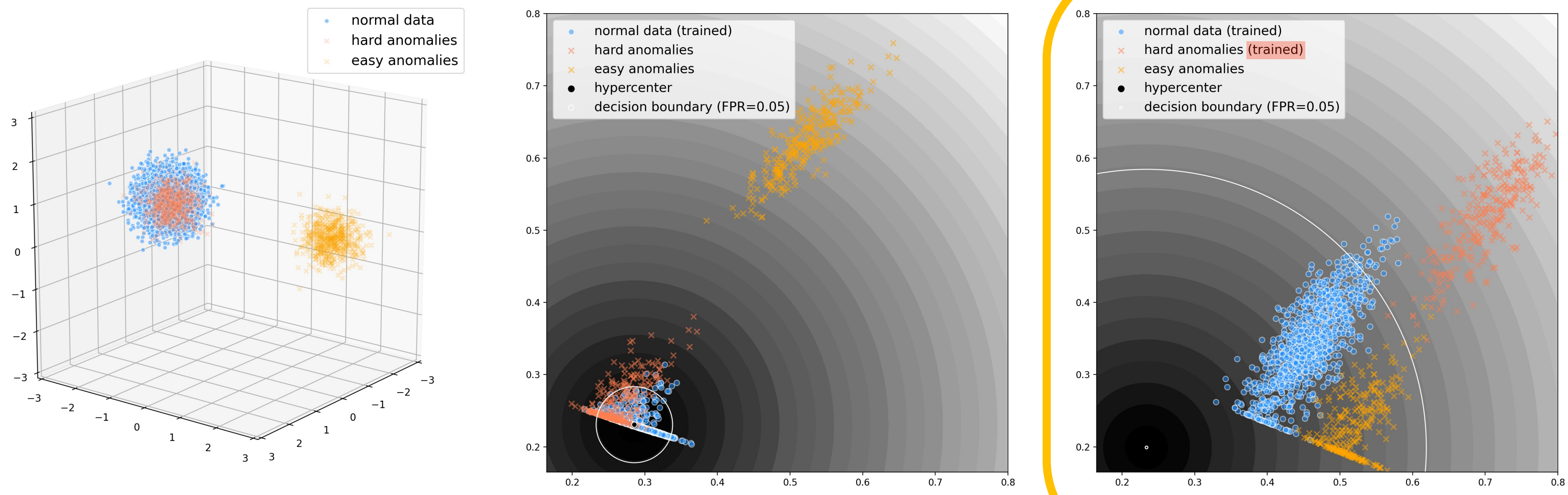


Fig 1. Original 3D Space

Fig 2. 2D Latent Space (*Semi-Supervised* AD)

Fig 3. 2D Latent Space (*Supervised* AD)

Source code available on github.com/ZIYU-DEEP/Understanding-Bias-in-Deep-Anomaly-Detection-PyTorch/blob/main/network/gaussian3d_net.py

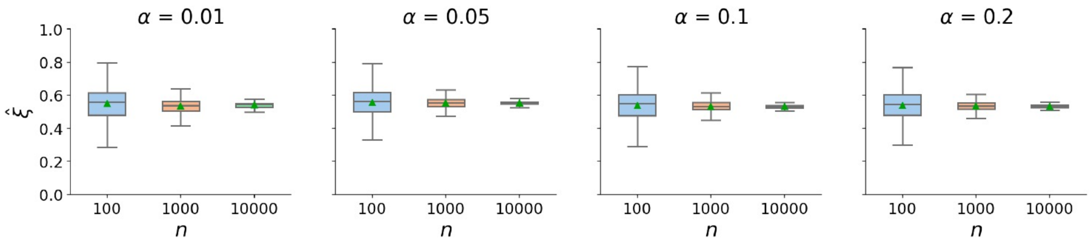
Our Contributions

1 *Define* Bias: A formal ERM Framework

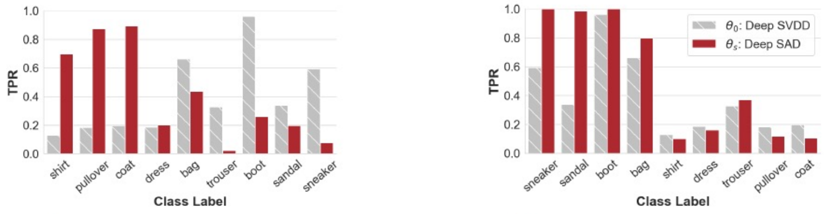
$$\text{bias}(\hat{s}_\theta, \hat{\tau}_\theta) := \arg\max_{(s_\theta, \tau_\theta):\theta \in \Theta} \text{TPR}(s_\theta, \tau_\theta) - \text{TPR}(\hat{s}_\theta, \hat{\tau}_\theta)$$

2 *Estimate Bias*: The First PAC Analysis

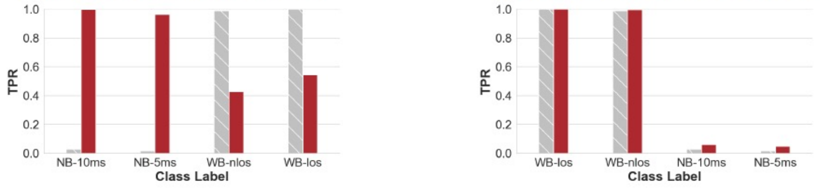
$$n \geq \frac{8}{\epsilon^2} \cdot \left(\log \frac{2}{1-\sqrt{1-\delta}} \cdot \left(\frac{2-\alpha}{\alpha}\right)^2 + \log \frac{2}{\delta} \cdot \frac{1}{1-\alpha} \left(\left(\frac{\ell_a}{\ell_0^-}\right)^2 + \left(\frac{\ell_a'}{\ell_0'}\right)^2 \right) \right)$$



3 *Characterize* Bias: Empirical Experiments



(a) Fashion-MNIST: Scenario 1 (b) Fashion-MNIST: Scenario 2



(c) Spectrum Misuse: Scenario 1 (d) Spectrum Misuse: Scenario 2

[Clarification] Bias in *AD* $\neq$ Bias in *Supervised Learning*

[Clarification] Bias in *AD* $\neq$ Bias in *Supervised Learning*

- Problem formulation is different.

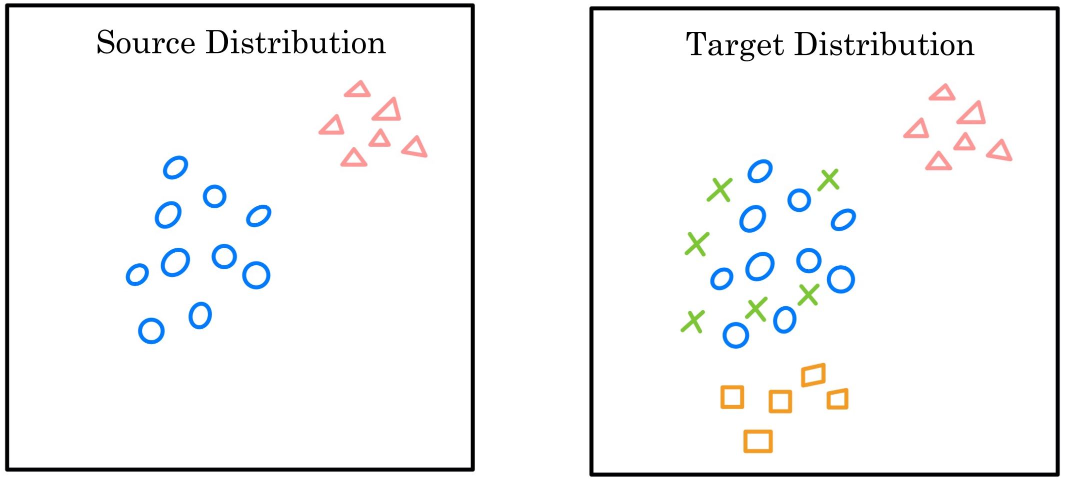


Fig 1. Data distribution of AD problem. The blue represent the normal data, and other different colors represent different *subtypes* of anomalies.

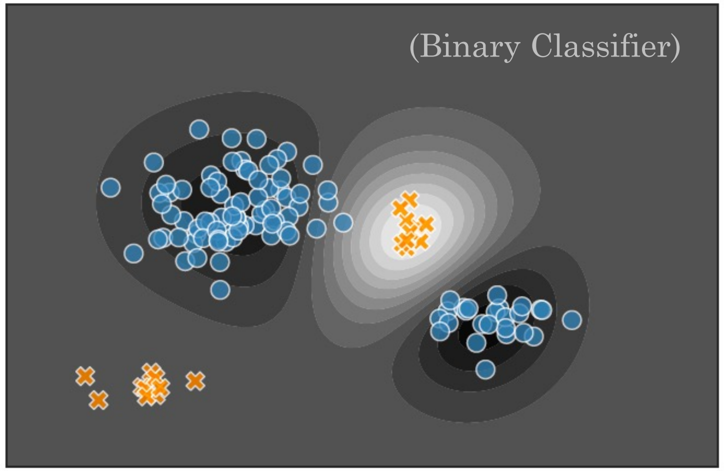
Task Type	Distribution Shift	Known Target Distribution	Known Target Label Set
Imbalanced Classification [Johnson and Khoshgoftaar, 2019]	No	N/A	N/A
Closed Set Domain Adaptation [Saenko *et al.*, 2010]	Yes	Yes	Yes
Open Set Domain Adaptation [Panareda Busto and Gall, 2017]	Yes	Yes	No
Anomaly Detection [Chalapathy and Chawla, 2019]	Yes	No	No

Table 1: Comparison of anomaly detection tasks with other relevant classification tasks.

[Clarification] Bias in *AD* ≠ Bias in *Supervised Learning*

- Training mechanism is different.

Supervised (Classifier)

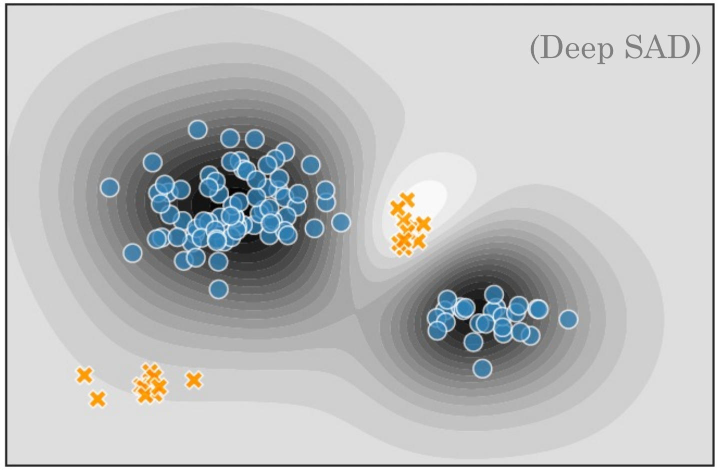


Pro Discriminating on *known* anomalies.

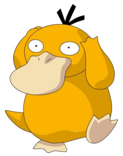
Con Overfitting to *known anomalies*.
➜ Overfitting bias.

Supervised (AD)

[Pan+19], [Yam+19], [Hen+19], [Ruf+20], [Goy+20], ...

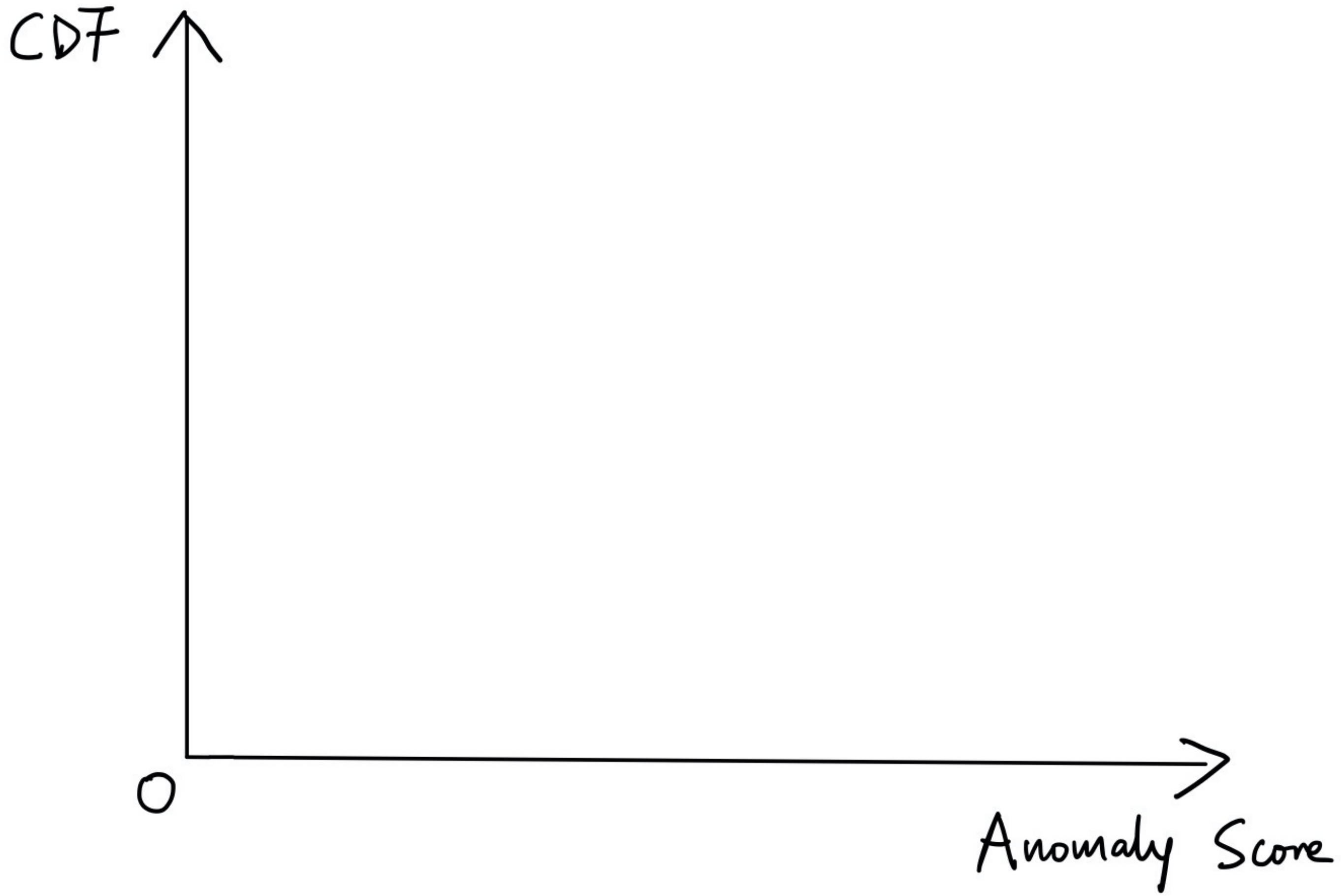


Pro A compact enclosing of the normal.
+
Discriminating on *known* anomalies.

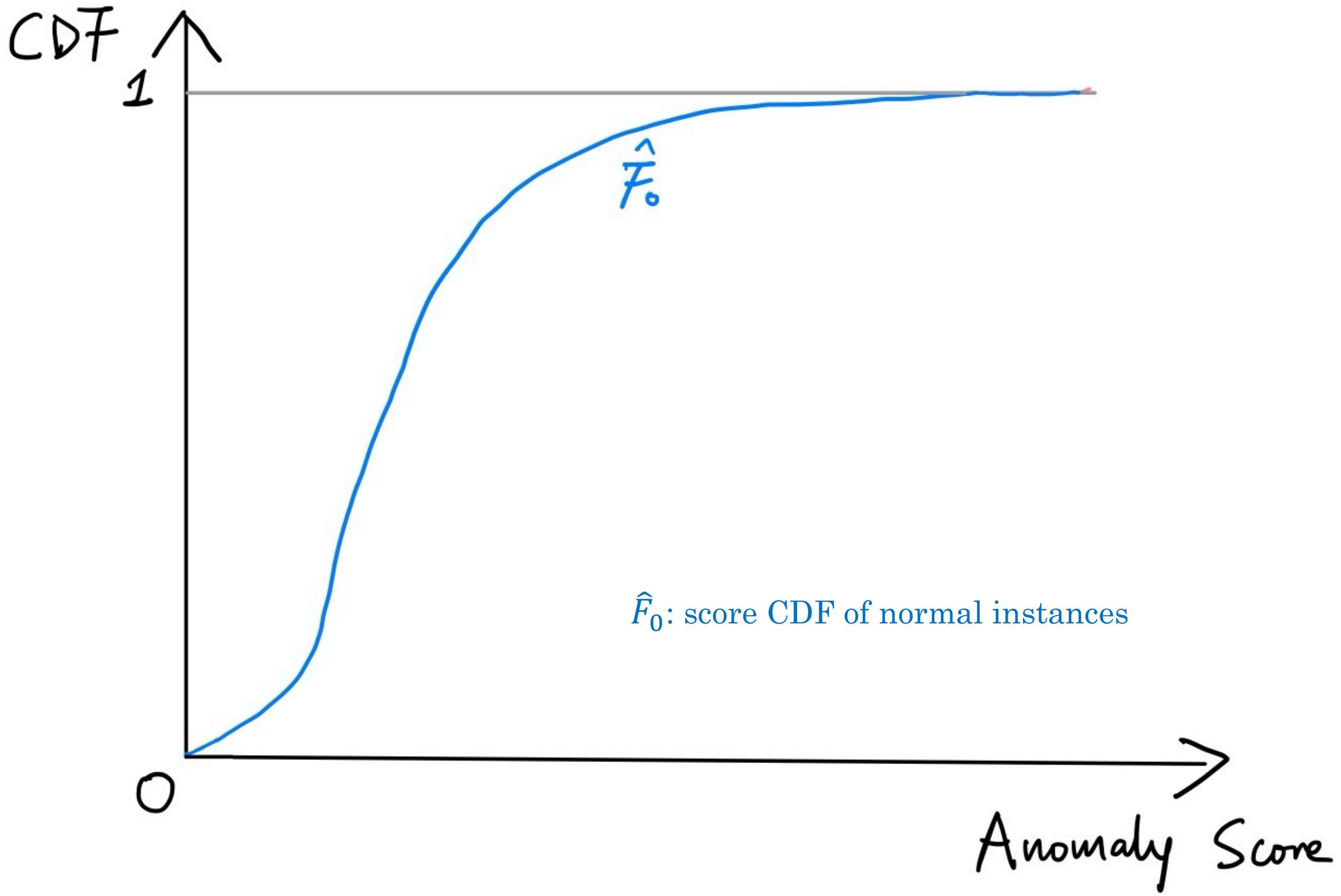


How to *Define* Bias in AD?

A General AD Framework

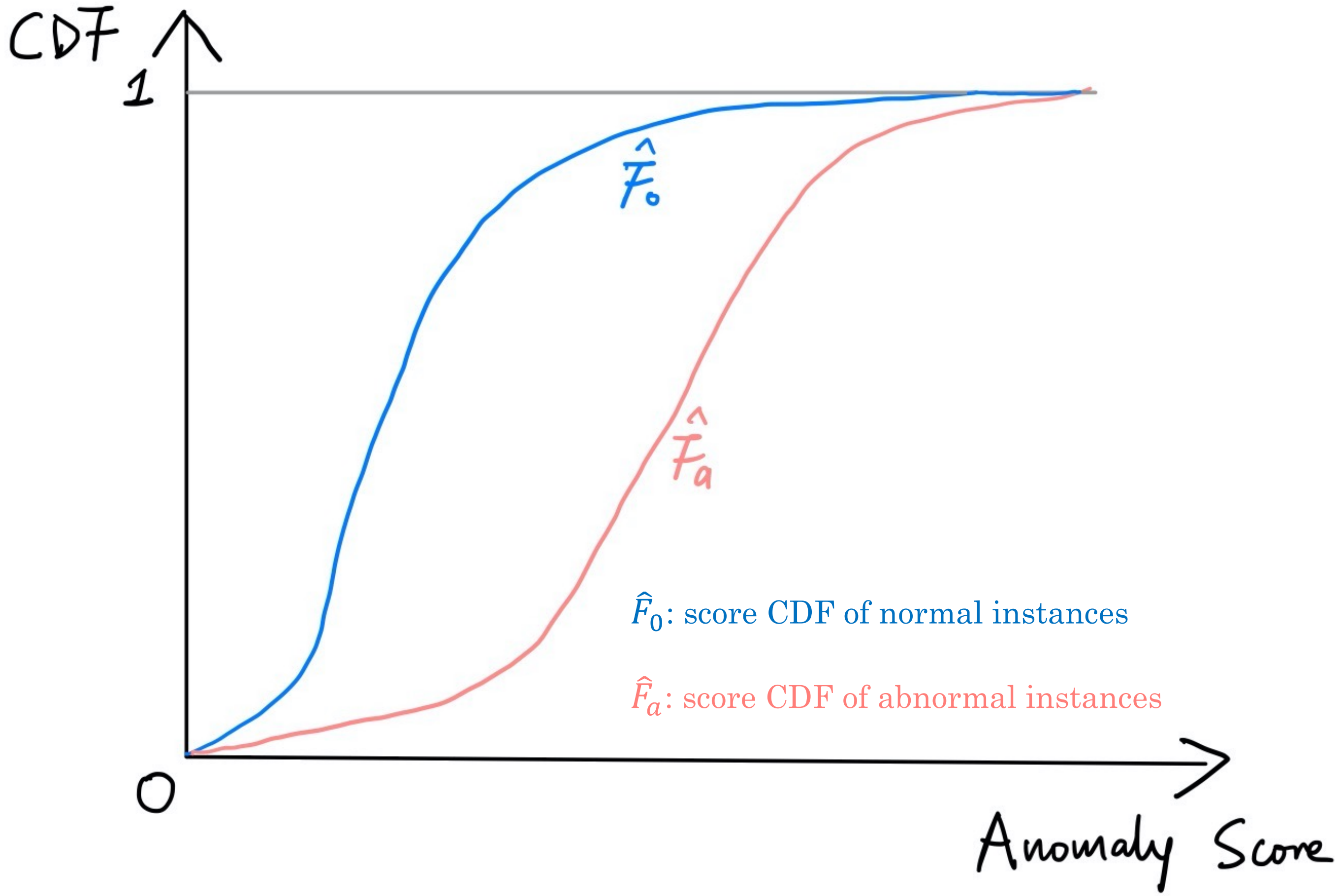


A General AD Framework

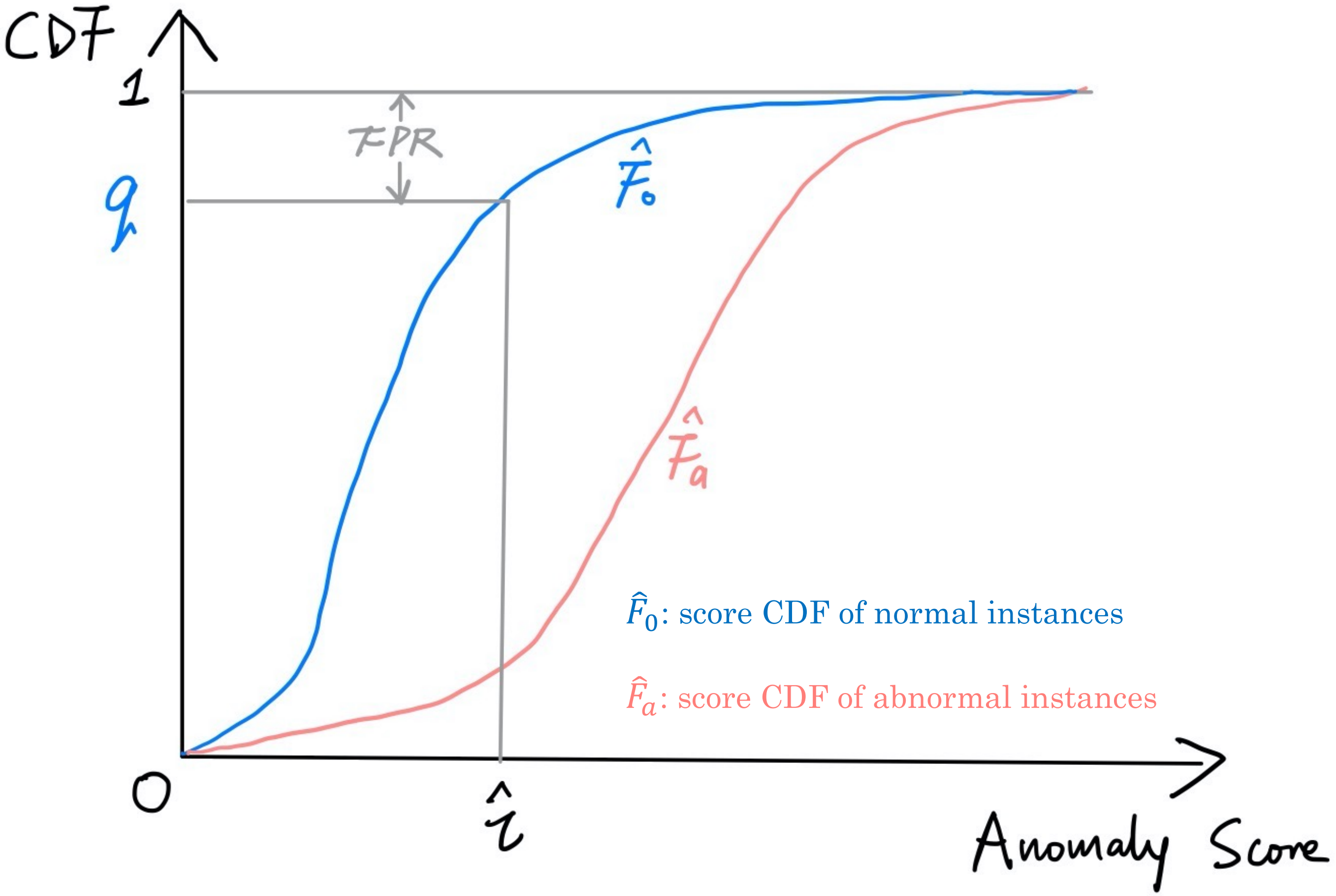


$\hat{F}_0$: score CDF of normal instances

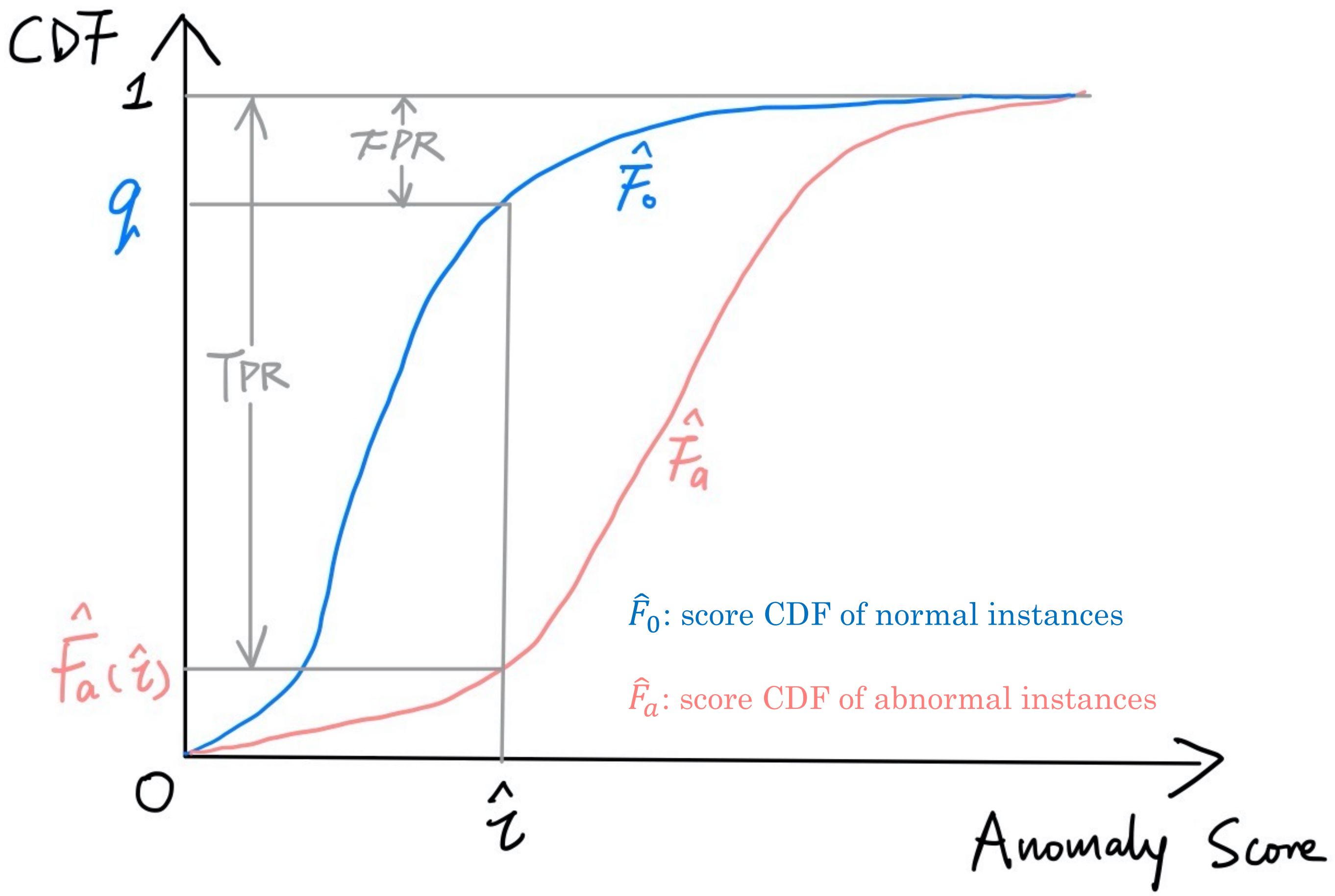
A General AD Framework



A General AD Framework



A General AD Framework



This formulation generalizes most practical AD frameworks, e.g, [Liu+18], [Li+19], [AsS+20], etc.

ERM-Style Scoring Bias

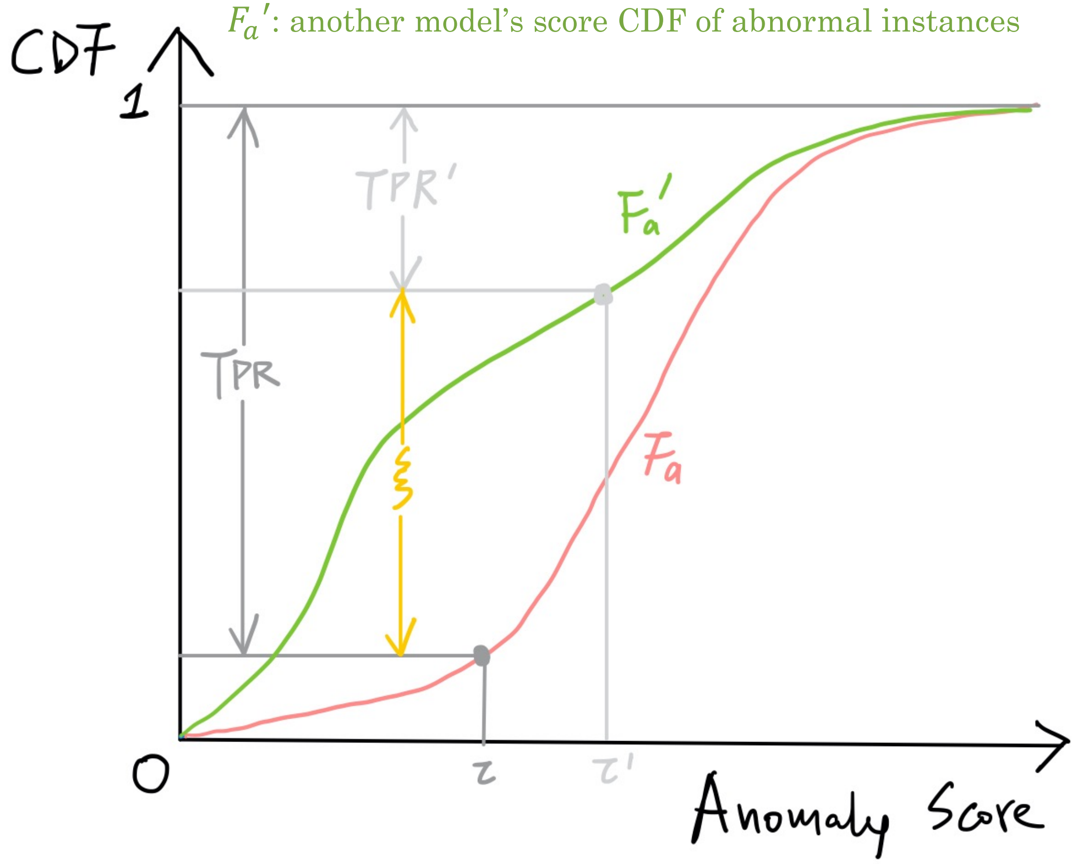
1 Scoring Bias

$$\text{bias}(\hat{s}_\theta, \hat{\tau}_\theta) := \underset{(s_\theta, \tau_\theta):\theta\in\Theta}{\arg\max}\ \text{TPR}(s_\theta, \tau_\theta) - \text{TPR}(\hat{s}_\theta, \hat{\tau}_\theta)$$

ERM-Style Scoring Bias

F_a: one model's score CDF of abnormal instances

$F_a{}'$: another model's score CDF of abnormal instances



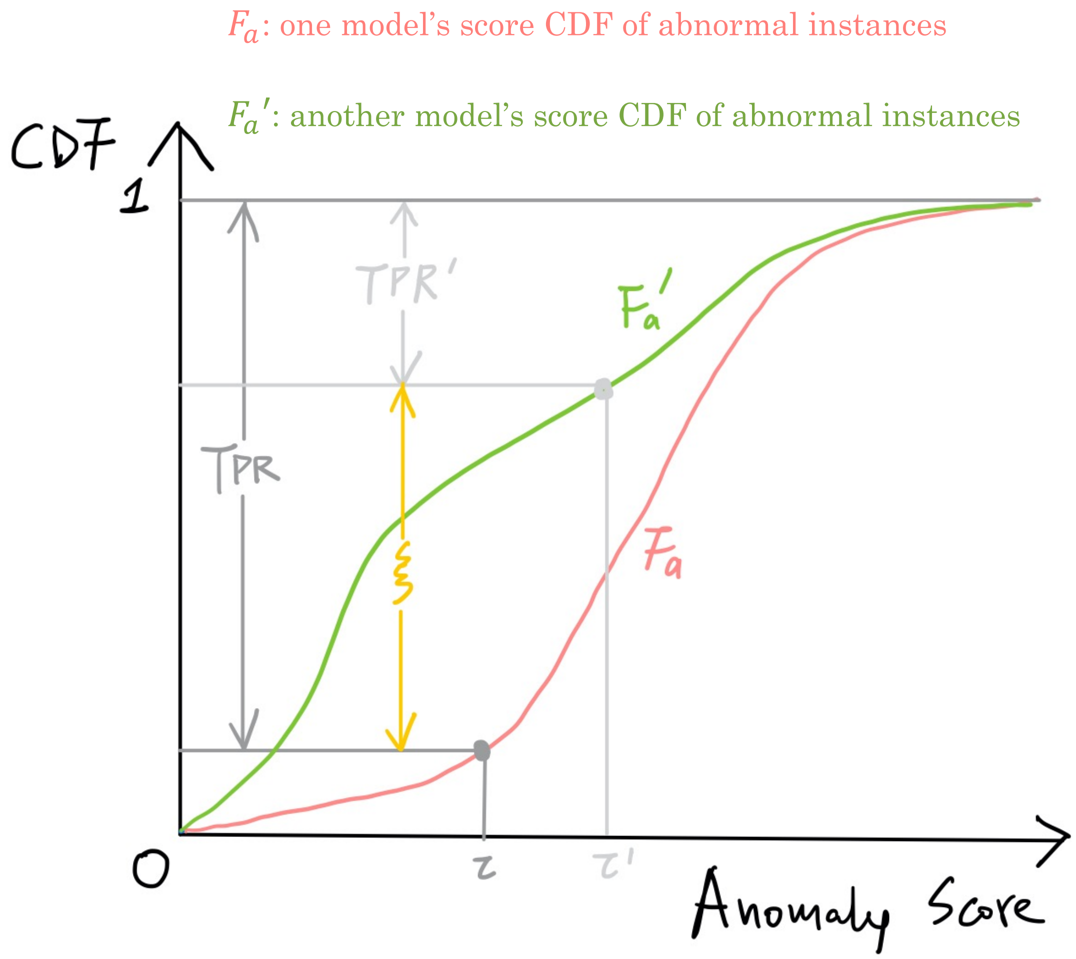
1 **Scoring Bias**

$$\mathrm{bias}(\hat{s}_\theta, \hat{\tau}_\theta) := \underset{(s_\theta, \tau_\theta):\theta\in\Theta}{\arg\max}\ \mathrm{TPR}(s_\theta, \tau_\theta) - \mathrm{TPR}(\hat{s}_\theta, \hat{\tau}_\theta)$$

2 ***Relative* Scoring Bias**

$$\begin{aligned}\xi(s, s') &:= \mathrm{bias}(s, \tau) - \mathrm{bias}(s', \tau') \\ &= \mathrm{TPR}(s', \tau') - \mathrm{TPR}(s, \tau)\end{aligned}$$

ERM-Style Scoring Bias



1 Scoring Bias

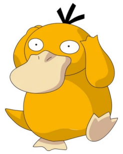
$$\text{bias}(\hat{s}_\theta, \hat{\tau}_\theta) := \underset{(s_\theta, \tau_\theta):\theta \in \Theta}{\arg\max} \ \text{TPR}(s_\theta, \tau_\theta) - \text{TPR}(\hat{s}_\theta, \hat{\tau}_\theta)$$

2 *Relative* Scoring Bias

$$\begin{aligned} \xi(s, s') &:= \text{bias}(s, \tau) - \text{bias}(s', \tau') \\ &= \text{TPR}(s', \tau') - \text{TPR}(s, \tau) \end{aligned}$$

3 *Empirical Relative* Scoring Bias

$$\hat{\xi}(s, s') := \widehat{\text{TPR}}(s', \tau') - \widehat{\text{TPR}}(s, \tau)$$



How to *Estimate* Bias in AD?

Finite Sample Guarantee

- **Goal**: a theoretical guarantee on model performance in terms of bias.

  e.g., how can we almost surely say that additional labeled data helps, or hurts?

Finite Sample Guarantee

- **Goal**: a theoretical guarantee on model performance in terms of bias.

  e.g., how can we almost surely say that additional labeled data helps, or hurts?

Proposition 1. *Given two scoring functions s, s' and a target FPR q, the relative scoring bias is* $\xi(s, s') = F_a(F_0^{-1}(q)) - F'_a(F_0'^{-1}(q))$.

Finite Sample Guarantee

- **Goal**: a theoretical guarantee on model performance in terms of bias.

  e.g., how can we almost surely say that additional labeled data helps, or hurts?

Proposition 1. *Given two scoring functions s, s' and a target FPR q, the relative scoring bias is* $\xi(s, s') = F_a(F_0^{-1}(q)) - F'_a(F_0'^{-1}(q))$.

Theorem 3. *Assume that $F_a, F'_a, F_0^-, F_0'^-$ are Lipschitz continuous with Lipschitz constant $\ell_a, \ell'_a, \ell_0^-, \ell_0'^-$, respectively. Let α be the fraction of abnormal data from the mixture distribution. Then, w.p. at least $1 - \delta$, with*

$$n = \mathcal{O}\left(\frac{1}{\alpha^2 \epsilon^2} \log \frac{1}{\delta}\right)$$

the empirical relative scoring bias satisfies $|\hat{\xi} - \xi| \leq \epsilon$.

---

* The exact form and detailed proof of Theorem 3 can be found in our full paper.

Convergence of Scoring Bias: *Empirical Results*

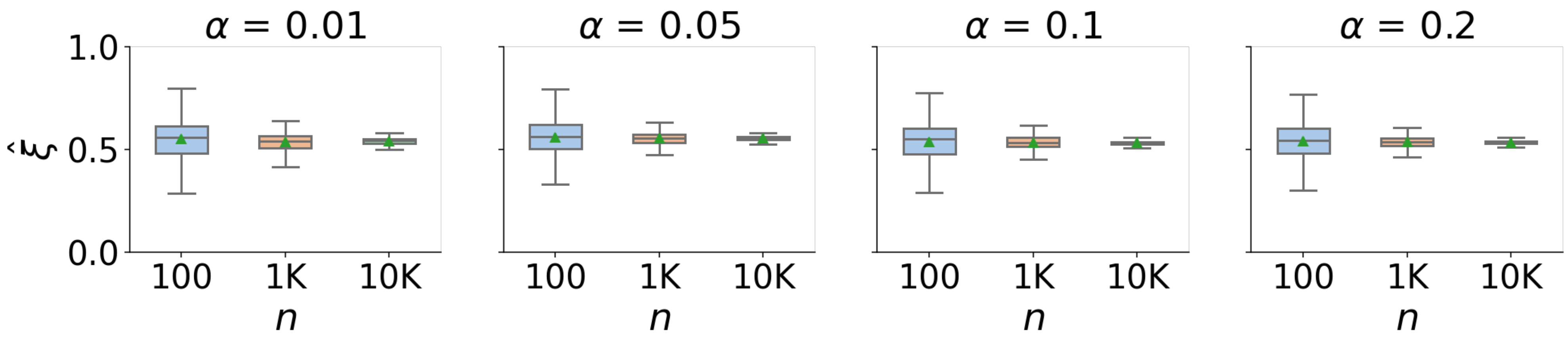
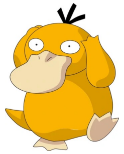


Fig 1. $\hat{\xi}$ is the scoring bias of Deep SVDD relative to Deep SAD.

- The estimation error ϵ decreases at the rate of $\frac{1}{\sqrt{n}}$.
- The sample complexity n grows as $\mathcal{O}\left(\frac{1}{\alpha^2\epsilon^2}\log\frac{1}{\delta}\right)$.



How does Bias *Impact* AD?

Recall on Our Observations...

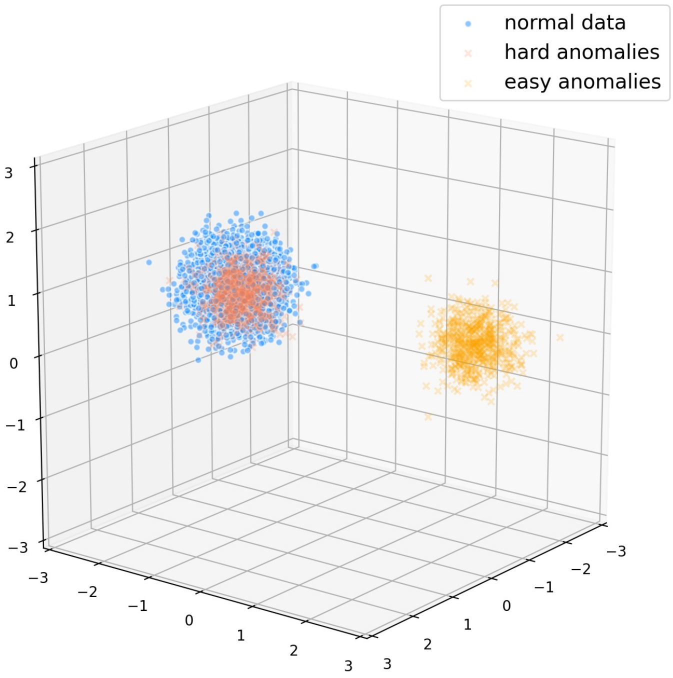


Fig 1. Original 3D Space

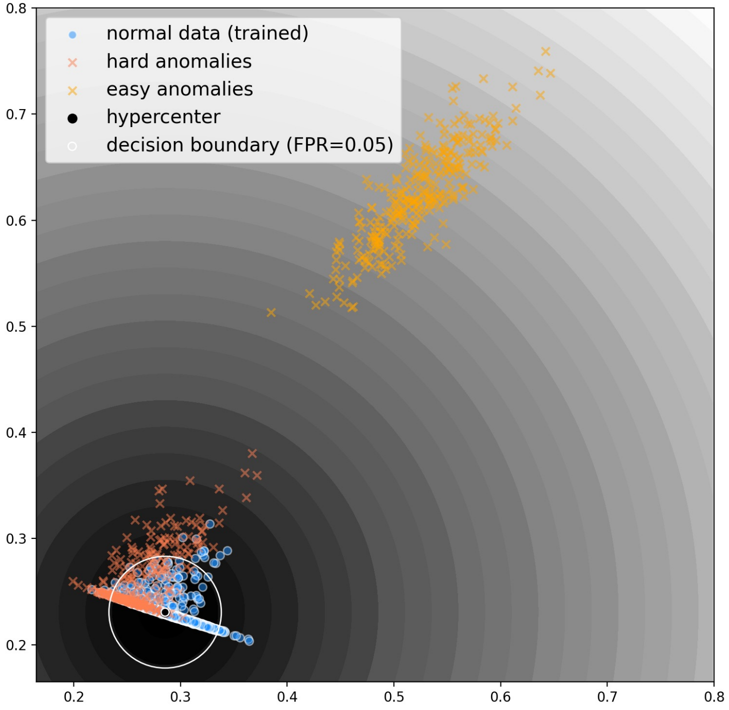


Fig 2. 2D Latent Space (*Semi-Supervised* AD)

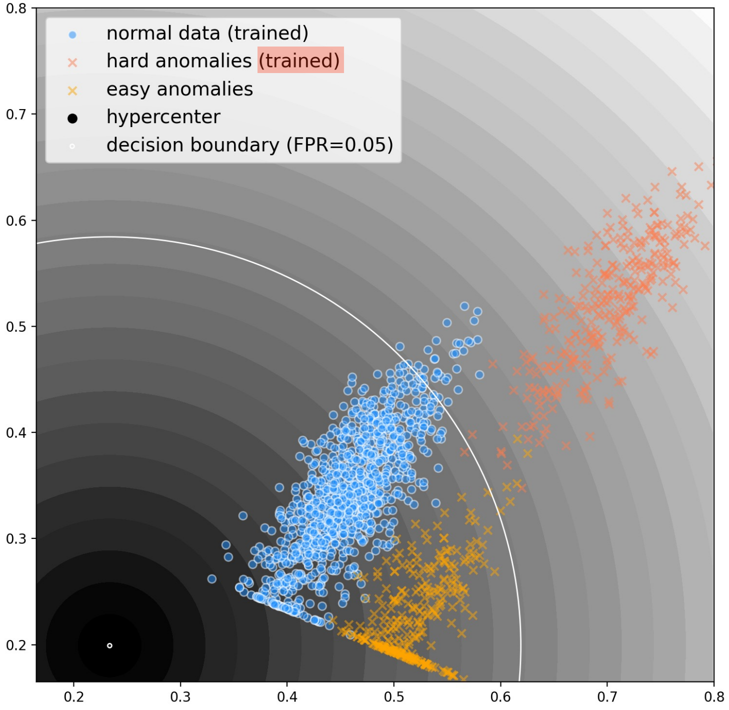


Fig 3. 2D Latent Space (*Supervised* AD)

Recall on Our Observations...

Training with different distributions affects normal enclosing *unevenly*.

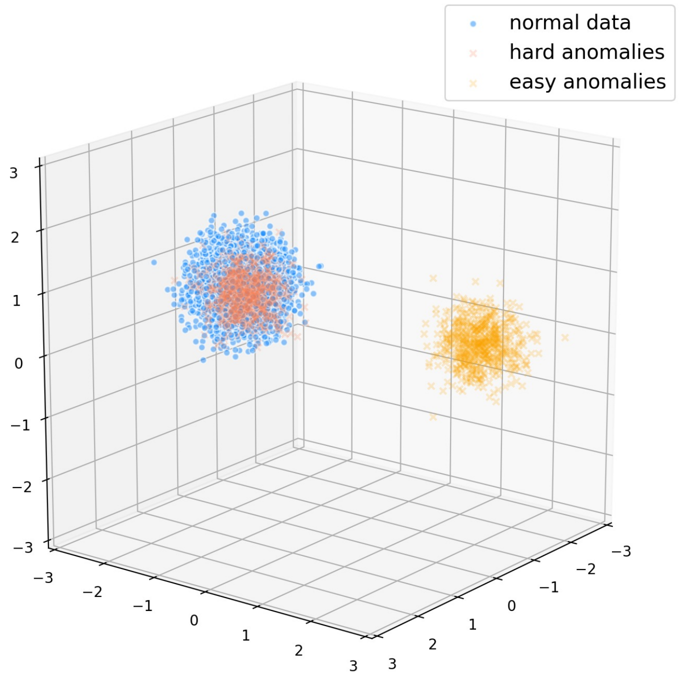


Fig 1. Original 3D Space

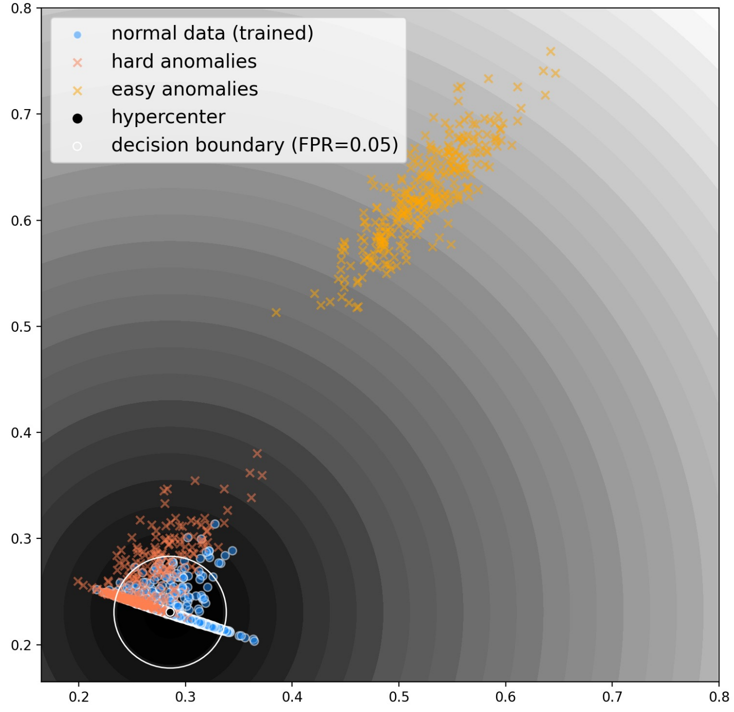


Fig 2. 2D Latent Space (*Semi-Supervised* AD)

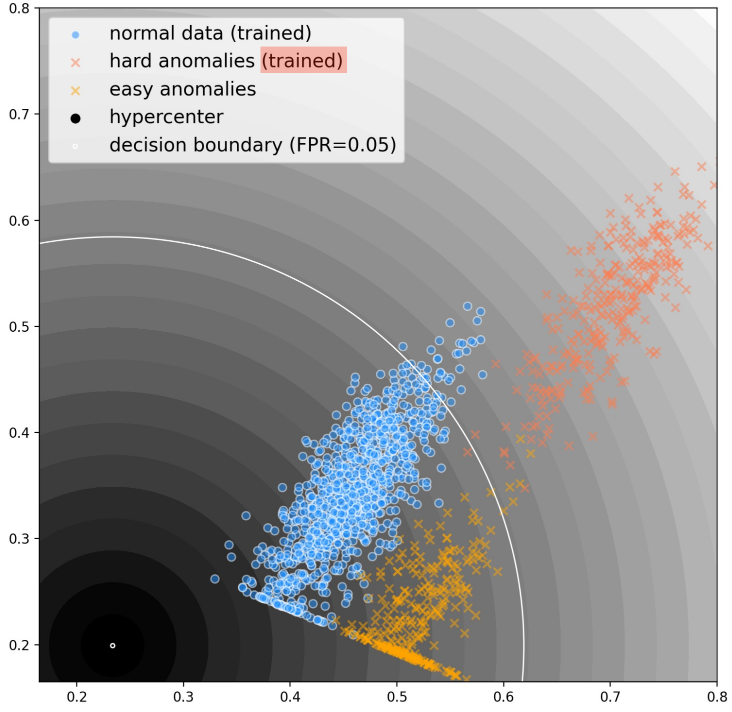
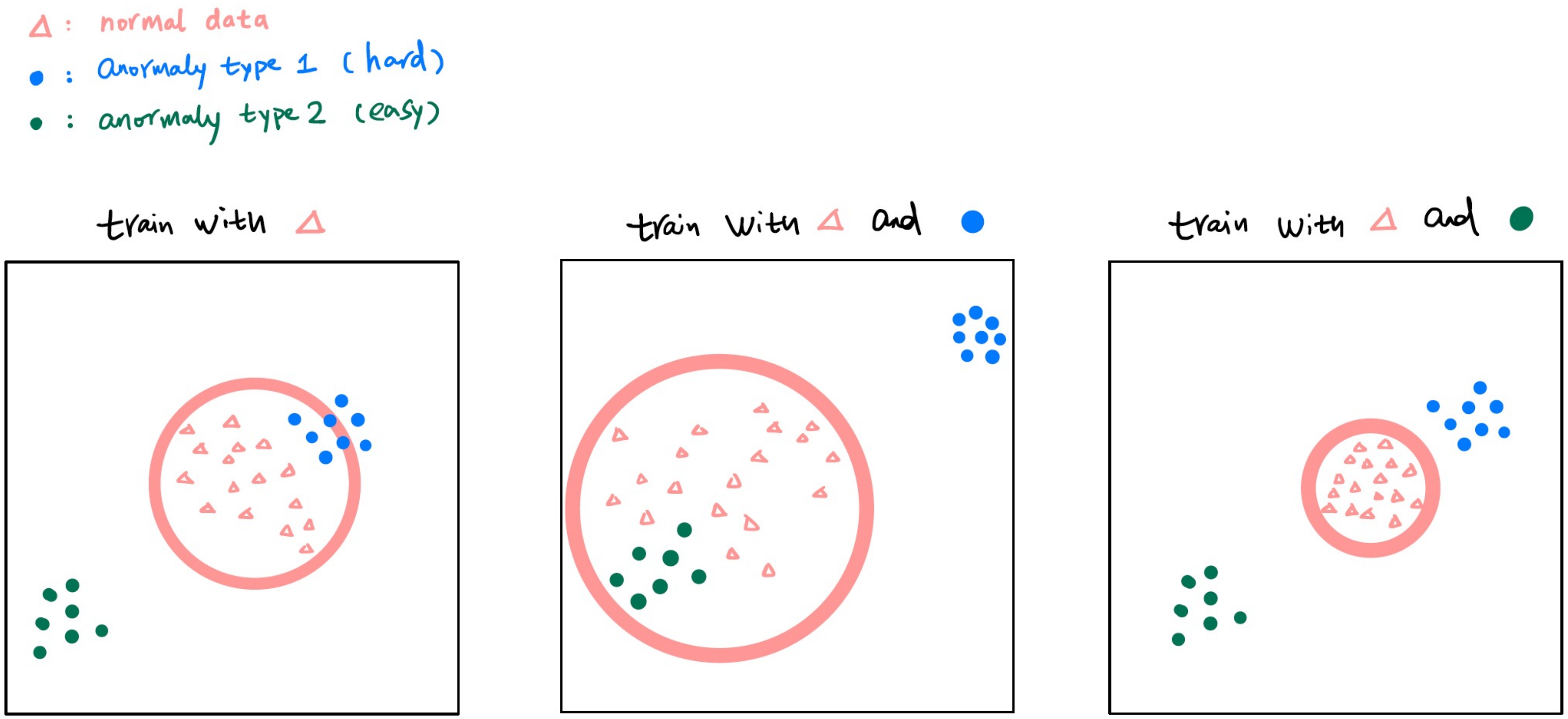


Fig 3. 2D Latent Space (*Supervised* AD)

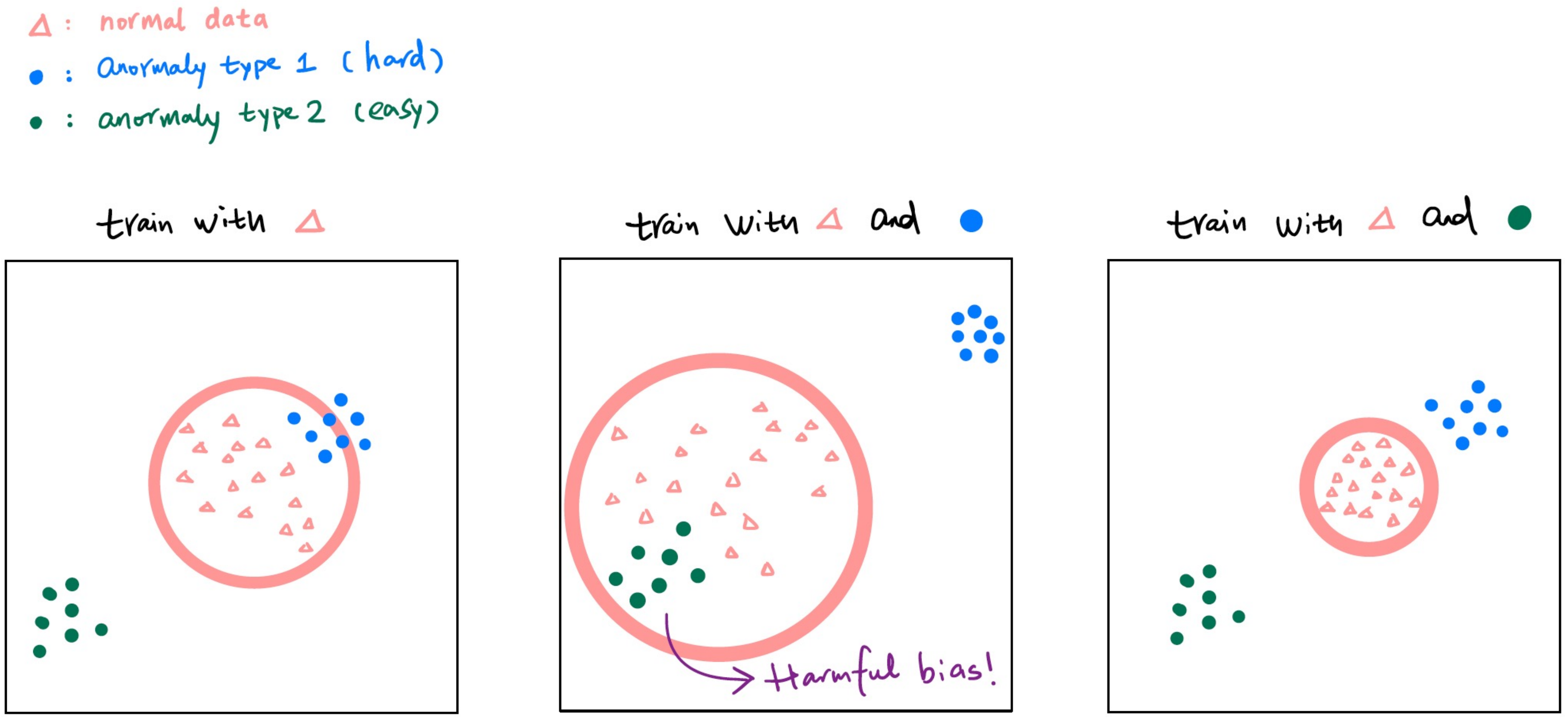
Our Hypothesis

Training with different distributions affects normal enclosing *unevenly*.



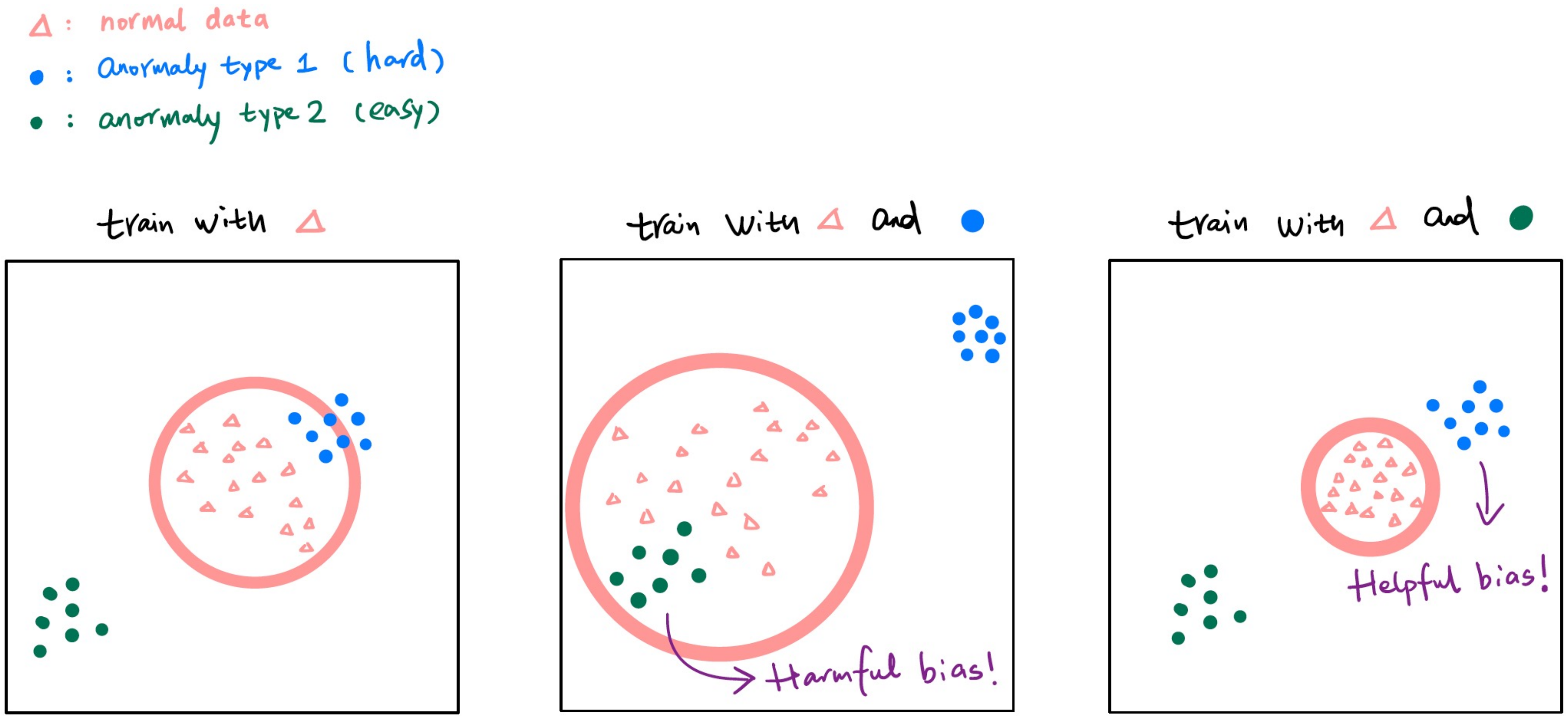
Our Hypothesis

Training with different distributions affects normal enclosing *unevenly*.



Our Hypothesis

Training with different distributions affects normal enclosing *unevenly*.

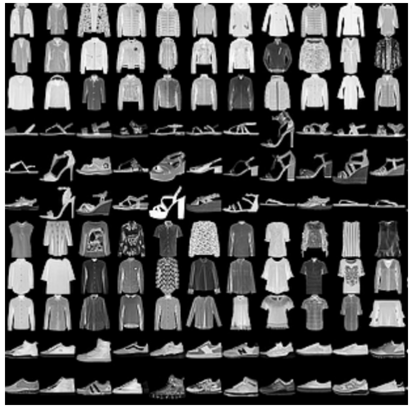


Experiment Setup

Models

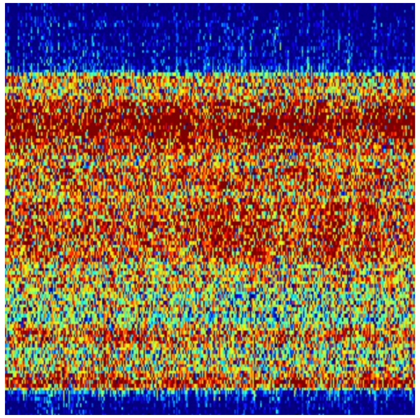
Type	Semi-supervised (trained on normal data)	Supervised (trained on normal & some abnormal data)
Hypersphere-based	Deep SVDD [Ruff *et al.*, 2018]	Deep SAD [Ruff *et al.*, 2020b], Hypersphere Classifier (HSC) [Ruff *et al.*, 2020a]
Reconstruction-based	Autoencoder (AE) [Zhou and Paffenroth, 2017]	Supervised AE (SAE)[7], Autoencoding Binary Classifier (ABC) [Yamanaka *et al.*, 2019]

Datasets



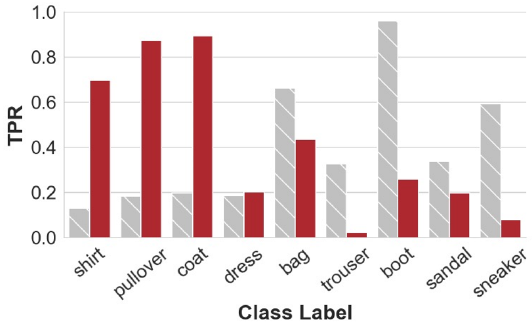
Fashion-MNIST

Landsat Satellite



Spectrum Misuse

Scenario 1: Training w/ the *Hard* Anomalies

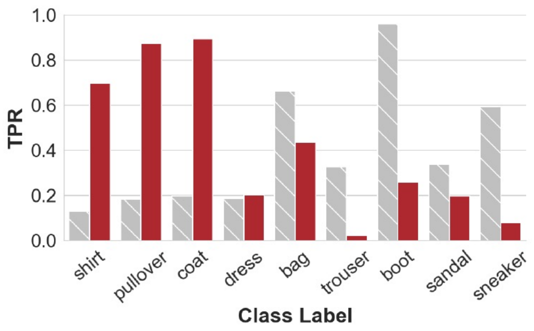


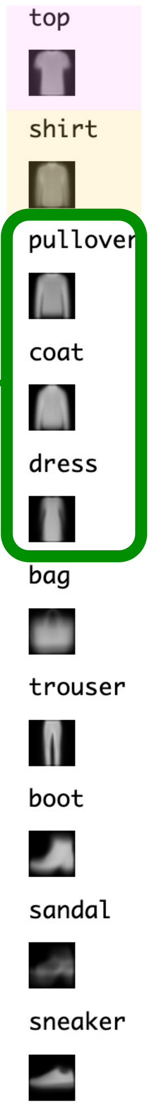
training normal = `top`, training abnormal = `shirt`

Test data	Deep SVDD	Deep SAD	HSC	AE	SAE	ABC	L^2 to shirt
shirt	0.09 ± 0.01	0.71 ± 0.01 ↑	0.70 ± 0.01 ↑	0.12 ± 0.01	0.72 ± 0.01 ↑	0.72 ± 0.01 ↑	0
pullover	0.13 ± 0.02	0.90 ± 0.01 ↑	0.89 ± 0.01 ↑	0.19 ± 0.02	0.84 ± 0.02 ↑	0.85 ± 0.01 ↑	0.01
coat	0.14 ± 0.03	0.92 ± 0.02 ↑	0.92 ± 0.01 ↑	0.15 ± 0.02	0.92 ± 0.02 ↑	0.92 ± 0.01 ↑	0.01
dress	0.17 ± 0.03	0.24 ± 0.03 ↑	0.24 ± 0.03 ↑	0.11 ± 0.01	0.20 ± 0.03 ↑	0.21 ± 0.03 ↑	0.04
bag	0.49 ± 0.07	0.38 ± 0.08 ↓	0.36 ± 0.07 ↓	0.70 ± 0.03	0.52 ± 0.09 ↓	0.53 ± 0.07 ↓	0.04
trouser	0.32 ± 0.10	0.07 ± 0.04 ↓	0.06 ± 0.03 ↓	0.59 ± 0.04	0.07 ± 0.04 ↓	0.16 ± 0.07 ↓	0.06
boot	0.92 ± 0.03	0.29 ± 0.15 ↓	0.27 ± 0.16 ↓	0.98 ± 0.02	0.90 ± 0.09 ↓	0.90 ± 0.08 ↓	0.08
sandal	0.30 ± 0.04	0.26 ± 0.08 ↓	0.26 ± 0.12 ↓	0.82 ± 0.02	0.46 ± 0.10 ↓	0.56 ± 0.09 ↓	0.09
sneaker	0.55 ± 0.09	0.12 ± 0.10 ↓	0.14 ± 0.12 ↓	0.74 ± 0.09	0.47 ± 0.19 ↓	0.46 ± 0.18 ↓	0.10

Table 3: The model TPR under scenario 1, Fashion-MNIST. The normal class `top` is similar to the abnormal training class `shirt`. Their L^2 distance = 0.02.

Scenario 1: Training w/ the *Hard* Anomalies





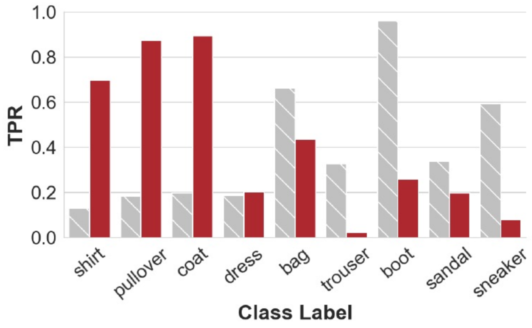
Positive bias!

training normal = `top`, training abnormal = `shirt`

Test data	Deep SVDD	Deep SAD	HSC	AE	SAE	ABC	L^2 to shirt
shirt	0.09 ± 0.01	$0.71 \pm 0.01 \uparrow$	$0.70 \pm 0.01 \uparrow$	0.12 ± 0.01	$0.72 \pm 0.01 \uparrow$	$0.72 \pm 0.01 \uparrow$	0
pullover	0.13 ± 0.02	$0.90 \pm 0.01 \uparrow$	$0.89 \pm 0.01 \uparrow$	0.19 ± 0.02	$0.84 \pm 0.02 \uparrow$	$0.85 \pm 0.01 \uparrow$	0.01
coat	0.14 ± 0.03	$0.92 \pm 0.02 \uparrow$	$0.92 \pm 0.01 \uparrow$	0.15 ± 0.02	$0.92 \pm 0.02 \uparrow$	$0.92 \pm 0.01 \uparrow$	0.01
dress	0.17 ± 0.03	$0.24 \pm 0.03 \uparrow$	$0.24 \pm 0.03 \uparrow$	0.11 ± 0.01	$0.20 \pm 0.03 \uparrow$	$0.21 \pm 0.03 \uparrow$	0.04
bag	0.49 ± 0.07	$0.38 \pm 0.08 \downarrow$	$0.36 \pm 0.07 \downarrow$	0.70 ± 0.03	$0.52 \pm 0.09 \downarrow$	$0.53 \pm 0.07 \downarrow$	0.04
trouser	0.32 ± 0.10	$0.07 \pm 0.04 \downarrow$	$0.06 \pm 0.03 \downarrow$	0.59 ± 0.04	$0.07 \pm 0.04 \downarrow$	$0.16 \pm 0.07 \downarrow$	0.06
boot	0.92 ± 0.03	$0.29 \pm 0.15 \downarrow$	$0.27 \pm 0.16 \downarrow$	0.98 ± 0.02	$0.90 \pm 0.09 \downarrow$	$0.90 \pm 0.08 \downarrow$	0.08
sandal	0.30 ± 0.04	$0.26 \pm 0.08 \downarrow$	$0.26 \pm 0.12 \downarrow$	0.82 ± 0.02	$0.46 \pm 0.10 \downarrow$	$0.56 \pm 0.09 \downarrow$	0.09
sneaker	0.55 ± 0.09	$0.12 \pm 0.10 \downarrow$	$0.14 \pm 0.12 \downarrow$	0.74 ± 0.09	$0.47 \pm 0.19 \downarrow$	$0.46 \pm 0.18 \downarrow$	0.10

Table 3: The model TPR under scenario 1, Fashion-MNIST. The normal class `top` is similar to the abnormal training class `shirt`. Their L^2 distance = 0.02.

Scenario 1: Training w/ the *Hard* Anomalies

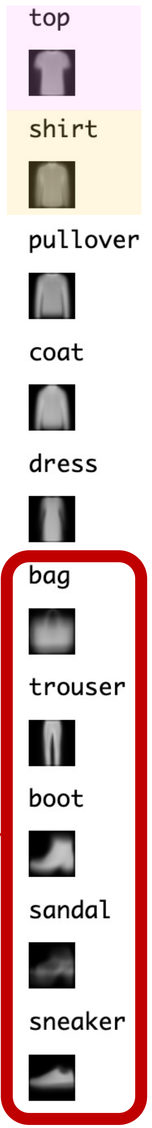


training normal = `top`, training abnormal = `shirt`

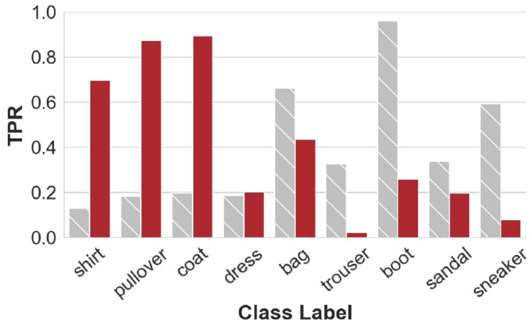
Test data	Deep SVDD	Deep SAD	HSC	AE	SAE	ABC	L^2 to shirt
shirt	0.09 ± 0.01	0.71 ± 0.01 ↑	0.70 ± 0.01 ↑	0.12 ± 0.01	0.72 ± 0.01 ↑	0.72 ± 0.01 ↑	0
pullover	0.13 ± 0.02	0.90 ± 0.01 ↑	0.89 ± 0.01 ↑	0.19 ± 0.02	0.84 ± 0.02 ↑	0.85 ± 0.01 ↑	0.01
coat	0.14 ± 0.03	0.92 ± 0.02 ↑	0.92 ± 0.01 ↑	0.15 ± 0.02	0.92 ± 0.02 ↑	0.92 ± 0.01 ↑	0.01
dress	0.17 ± 0.03	0.24 ± 0.03 ↑	0.24 ± 0.03 ↑	0.11 ± 0.01	0.20 ± 0.03 ↑	0.21 ± 0.03 ↑	0.04
bag	0.49 ± 0.07	0.38 ± 0.08 ↓	0.36 ± 0.07 ↓	0.70 ± 0.03	0.52 ± 0.09 ↓	0.53 ± 0.07 ↓	0.04
trouser	0.32 ± 0.10	0.07 ± 0.04 ↓	0.06 ± 0.03 ↓	0.59 ± 0.04	0.07 ± 0.04 ↓	0.16 ± 0.07 ↓	0.06
boot	0.92 ± 0.03	0.29 ± 0.15 ↓	0.27 ± 0.16 ↓	0.98 ± 0.02	0.90 ± 0.09 ↓	0.90 ± 0.08 ↓	0.08
sandal	0.30 ± 0.04	0.26 ± 0.08 ↓	0.26 ± 0.12 ↓	0.82 ± 0.02	0.46 ± 0.10 ↓	0.56 ± 0.09 ↓	0.09
sneaker	0.55 ± 0.09	0.12 ± 0.10 ↓	0.14 ± 0.12 ↓	0.74 ± 0.09	0.47 ± 0.19 ↓	0.46 ± 0.18 ↓	0.10

Table 3: The model TPR under scenario 1, Fashion-MNIST. The normal class `top` is similar to the abnormal training class `shirt`. Their L^2 distance = 0.02.

Negative bias!



Scenario 1: Training w/ the *Hard* Anomalies



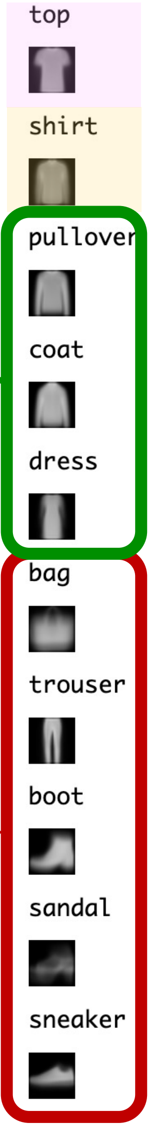
training normal = top, training abnormal = shirt

Test data	Deep SVDD	Deep SAD	HSC	AE	SAE	ABC	L^2 to shirt
shirt	0.09 ± 0.01	0.71 ± 0.01 ↑	0.70 ± 0.01 ↑	0.12 ± 0.01	0.72 ± 0.01 ↑	0.72 ± 0.01 ↑	0
pullover	0.13 ± 0.02	0.90 ± 0.01 ↑	0.89 ± 0.01 ↑	0.19 ± 0.02	0.84 ± 0.02 ↑	0.85 ± 0.01 ↑	0.01
coat	0.14 ± 0.03	0.92 ± 0.02 ↑	0.92 ± 0.01 ↑	0.15 ± 0.02	0.92 ± 0.02 ↑	0.92 ± 0.01 ↑	0.01
dress	0.17 ± 0.03	0.24 ± 0.03 ↑	0.24 ± 0.03 ↑	0.11 ± 0.01	0.20 ± 0.03 ↑	0.21 ± 0.03 ↑	0.04
bag	0.49 ± 0.07	0.38 ± 0.08 ↓	0.36 ± 0.07 ↓	0.70 ± 0.03	0.52 ± 0.09 ↓	0.53 ± 0.07 ↓	0.04
trouser	0.32 ± 0.10	0.07 ± 0.04 ↓	0.06 ± 0.03 ↓	0.59 ± 0.04	0.07 ± 0.04 ↓	0.16 ± 0.07 ↓	0.06
boot	0.92 ± 0.03	0.29 ± 0.15 ↓	0.27 ± 0.16 ↓	0.98 ± 0.02	0.90 ± 0.09 ↓	0.90 ± 0.08 ↓	0.08
sandal	0.30 ± 0.04	0.26 ± 0.08 ↓	0.26 ± 0.12 ↓	0.82 ± 0.02	0.46 ± 0.10 ↓	0.56 ± 0.09 ↓	0.09
sneaker	0.55 ± 0.09	0.12 ± 0.10 ↓	0.14 ± 0.12 ↓	0.74 ± 0.09	0.47 ± 0.19 ↓	0.46 ± 0.18 ↓	0.10

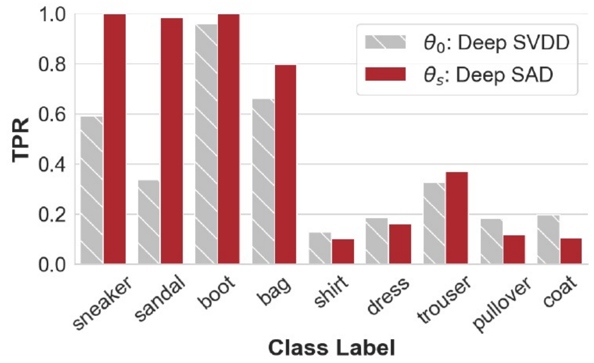
Table 3: The model TPR under scenario 1, Fashion-MNIST. The normal class top is similar to the abnormal training class shirt. Their L^2 distance = 0.02.

Positive bias!

Negative bias!



Scenario 2: Training w/ the *Easy* Anomalies

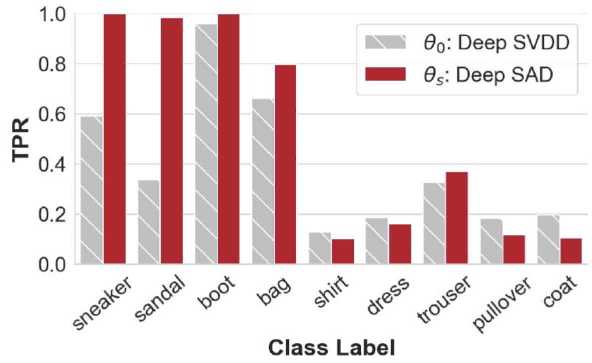


training normal = `top`, training abnormal = `sneaker`

Test data	Deep SVDD	Deep SAD	HSC	AE	SAE	ABC	L^2 to sneaker
`sneaker`	0.55 ± 0.09	$1.00 \pm 0.00 \uparrow$	$1.00 \pm 0.00 \uparrow$	0.74 ± 0.09	$1.00 \pm 0.00 \uparrow$	$1.00 \pm 0.00 \uparrow$	0
`sandal`	0.30 ± 0.04	$0.99 \pm 0.01 \uparrow$	$0.98 \pm 0.02 \uparrow$	0.82 ± 0.02	$1.00 \pm 0.00 \uparrow$	$1.00 \pm 0.00 \uparrow$	0.02
`boot`	0.92 ± 0.03	$1.00 \pm 0.00 \uparrow$	$0.97 \pm 0.02 \uparrow$	0.98 ± 0.02	$1.00 \pm 0.00 \uparrow$	$1.00 \pm 0.00 \uparrow$	0.07
`bag`	0.49 ± 0.07	$0.80 \pm 0.05 \uparrow$	$0.81 \pm 0.11 \uparrow$	0.70 ± 0.03	$0.84 \pm 0.03 \uparrow$	$0.82 \pm 0.03 \uparrow$	0.07
`shirt`	0.09 ± 0.01	$0.11 \pm 0.02 \uparrow$	$0.12 \pm 0.01 \uparrow$	0.12 ± 0.01	$0.13 \pm 0.01 \uparrow$	$0.15 \pm 0.01 \uparrow$	0.10
`trouser`	0.32 ± 0.09	0.31 ± 0.10	$0.11 \pm 0.12 \downarrow$	0.58 ± 0.04	0.58 ± 0.03	0.58 ± 0.05	0.12
`dress`	0.16 ± 0.03	0.16 ± 0.04	$0.11 \pm 0.01 \downarrow$	0.11 ± 0.01	0.11 ± 0.01	0.12 ± 0.01	0.13
`pullover`	0.13 ± 0.02	0.13 ± 0.03	0.14 ± 0.05	0.19 ± 0.02	0.21 ± 0.03	0.19 ± 0.02	0.13
`coat`	0.14 ± 0.03	0.13 ± 0.03	0.13 ± 0.06	0.15 ± 0.02	0.16 ± 0.02	0.15 ± 0.02	0.14

Table 6: The model TPR under scenario 2, Fashion-MNIST. The normal class `top` is dissimilar to the abnormal training class `sneaker`, and the L^2 distance between the two is 0.13.

Scenario 2: Training w/ the *Easy* Anomalies

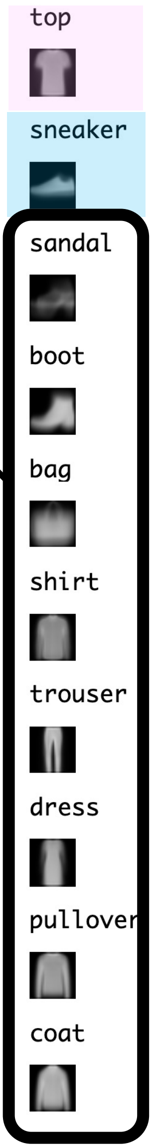


training normal = `top`, training abnormal = `sneaker`

Test data	Deep SVDD	Deep SAD	HSC	AE	SAE	ABC	L^2 to sneaker
sneaker	0.55 ± 0.09	1.00 ± 0.00 ↑	1.00 ± 0.00 ↑	0.74 ± 0.09	1.00 ± 0.00 ↑	1.00 ± 0.00 ↑	0
sandal	0.30 ± 0.04	0.99 ± 0.01 ↑	0.98 ± 0.02 ↑	0.82 ± 0.02	1.00 ± 0.00 ↑	1.00 ± 0.00 ↑	0.02
boot	0.92 ± 0.03	1.00 ± 0.00 ↑	0.97 ± 0.02 ↑	0.98 ± 0.02	1.00 ± 0.00 ↑	1.00 ± 0.00 ↑	0.07
bag	0.49 ± 0.07	0.80 ± 0.05 ↑	0.81 ± 0.11 ↑	0.70 ± 0.03	0.84 ± 0.03 ↑	0.82 ± 0.03 ↑	0.07
shirt	0.09 ± 0.01	0.11 ± 0.02 ↑	0.12 ± 0.01 ↑	0.12 ± 0.01	0.13 ± 0.01 ↑	0.15 ± 0.01 ↑	0.10
trouser	0.32 ± 0.09	0.31 ± 0.10	0.11 ± 0.12 ↓	0.58 ± 0.04	0.58 ± 0.03	0.58 ± 0.05	0.12
dress	0.16 ± 0.03	0.16 ± 0.04	0.11 ± 0.01 ↓	0.11 ± 0.01	0.11 ± 0.01	0.12 ± 0.01	0.13
pullover	0.13 ± 0.02	0.13 ± 0.03	0.14 ± 0.05	0.19 ± 0.02	0.21 ± 0.03	0.19 ± 0.02	0.13
coat	0.14 ± 0.03	0.13 ± 0.03	0.13 ± 0.06	0.15 ± 0.02	0.16 ± 0.02	0.15 ± 0.02	0.14

Table 6: The model TPR under scenario 2, Fashion-MNIST. The normal class `top` is dissimilar to the abnormal training class `sneaker`, and the L^2 distance between the two is 0.13.

Mostly harmless bias!



Scenario 3: Mixed Training

training normal = top, training abnormal = 50% shirt and 50% sneaker

Test data	Deep SVDD	Deep SAD	HSC	AE	SAE	ABC	L^2 to shirt	L^2 to sneaker
shirt	0.09 ± 0.01	0.69 ± 0.01 ↑	0.69 ± 0.02 ↑	0.12 ± 0.01	0.67 ± 0.01 ↑	0.66 ± 0.01 ↑	0	0.10
sneaker	0.55 ± 0.09	1.00 ± 0.00 ↑	1.00 ± 0.00 ↑	0.74 ± 0.09	1.00 ± 0.00 ↑	1.00 ± 0.00 ↑	0.10	0
pullover	0.13 ± 0.02	0.90 ± 0.01 ↑	0.90 ± 0.01 ↑	0.19 ± 0.02	0.82 ± 0.02 ↑	0.83 ± 0.02 ↑	0.01	0.13
coat	0.14 ± 0.03	0.91 ± 0.02 ↑	0.90 ± 0.01 ↑	0.15 ± 0.02	0.86 ± 0.02 ↑	0.87 ± 0.02 ↑	0.01	0.14
dress	0.17 ± 0.03	0.23 ± 0.04 ↑	0.24 ± 0.04 ↑	0.11 ± 0.01	0.19 ± 0.03 ↑	0.18 ± 0.02 ↑	0.04	0.13
bag	0.49 ± 0.07	0.63 ± 0.06 ↑	0.62 ± 0.07 ↑	0.70 ± 0.03	0.76 ± 0.05 ↑	0.78 ± 0.03 ↑	0.04	0.07
trouser	0.32 ± 0.10	0.05 ± 0.04 ↓	0.04 ± 0.02 ↓	0.59 ± 0.04	0.22 ± 0.08 ↓	0.34 ± 0.06 ↓	0.06	0.12
boot	0.92 ± 0.03	0.95 ± 0.03	0.95 ± 0.03	0.98 ± 0.02	1.00 ± 0.00 ↑	1.00 ± 0.00 ↑	0.08	0.07
sandal	0.30 ± 0.04	0.92 ± 0.04 ↑	0.92 ± 0.04 ↑	0.82 ± 0.02	0.96 ± 0.01 ↑	0.97 ± 0.01 ↑	0.09	0.02

Table 9: The model TPR under configuration 1 of weighted mixture training on Fashion-MNIST.

Takeaways and Future Directions

- Additional labeled data in AD poses a *hidden threat* for model practitioners.

- **Potential debiasing strategies**:
  - Data-based strategy
    - Using *active learning* and to get representative anomaly labels on the fly.
    - Leveraging *synthetic samples*;
  - Model-based strategy
    - *Robust* model design (e.g., *ensembles*).

References

[SHS05] I. Steinwart, D. Hush, and C. Scovel, “A classification framework for anomaly detection,” Journal of Machine Learning Research, vol. 6, no. Feb, pp. 211–232, 2005.

[Gör+13] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld, “Toward supervised anomaly detection,” Journal of Artificial Intelligence Research, vol. 46, pp. 235–262, 2013.

[AC15] J. An and S. Cho. “Variational autoencoder based anomaly detection using reconstruction probability.” Special Lecture on IE 2, no. 1 (2015): 1-18.

[ZP17] C. Zhou and R. Paffenroth. “Anomaly detection with robust deep autoencoders”. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017.

[Ruf+18] L. Ruff, R. Vandermeulen, N. Gornitz, L. Deecke, S. Siddiqui, A. Binder, E. Muller, and M. Kloft. “Deep one-class classification.” In Proc. of ICML, 2018.

[Goy+20] S. Goyal, A. Raghunathan, M. Jain, H. Simhadri, and P. Jain. “Drocc: Deep robust one classification.” In Proc. of ICML, 2020.

[Pan+19] G. Pang, C. Shen, and A. Hengel. “Deep anomaly detection with deviation networks.” In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 353–362, 2019.

[Yam+19] Y. Yamanaka, T. Iwata, H. Takahashi, M. Yamada, and S. Kanai. “Autoencoding binary classifiers for supervised anomaly detection.” arXiv preprint arXiv:1903.10709, 2019.

[Hen+19] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. “Using self-supervised learning can improve model robustness and uncertainty.” In Advances in Neural Information Processing Systems, pages 15663–15674, 2019.

[Ruf+20] L. Ruff, R. Vandermeulen, N. Gornitz, A. Binder, E. Muller, K. Muller, and M. Kloft. “Deep semi-supervised anomaly detection.” In Proc. of ICLR, 2020.

[AsS+20] B. AsSadhan, R. AlShaalan, D. Diab, A. Alzoghaiby, S. Alshebeili, J. Al-Muhtadi, H. Bin-Abbas, F. Abd El-Samie. “A robust anomaly detection method using a constant false alarm rate approach.” Multimedia Tools and Applications. 2020 Jan 22:1-24.

[Li+19] Z. Li, Z. Xiao, B. Wang, B. Zhao, and H. Zheng. “Scaling deep learning models for spectrum anomaly detection.” In Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing, page 291–300, 2019.

[Liu+18] S. Liu, R. Garrepalli, T. Dietterich, A. Fern, and D. Hendrycks. “Open category detection with PAC guarantees.” In Proc. of ICML, 2018.