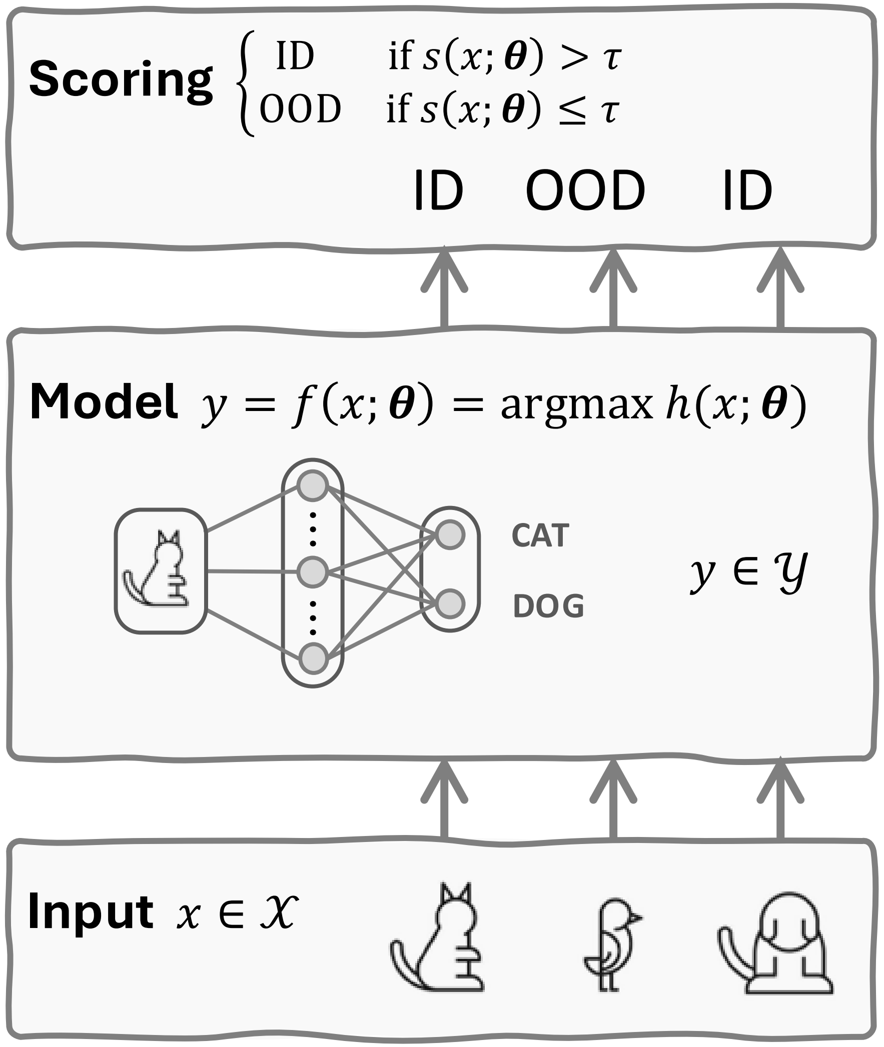
# On the Insights and Strategies for OOD Detection Learning

Dr. Qizhou WANG
RIKEN AIP

https://qizhouwang.github.io/homepage

# Post-hoc OOD Detection: Review

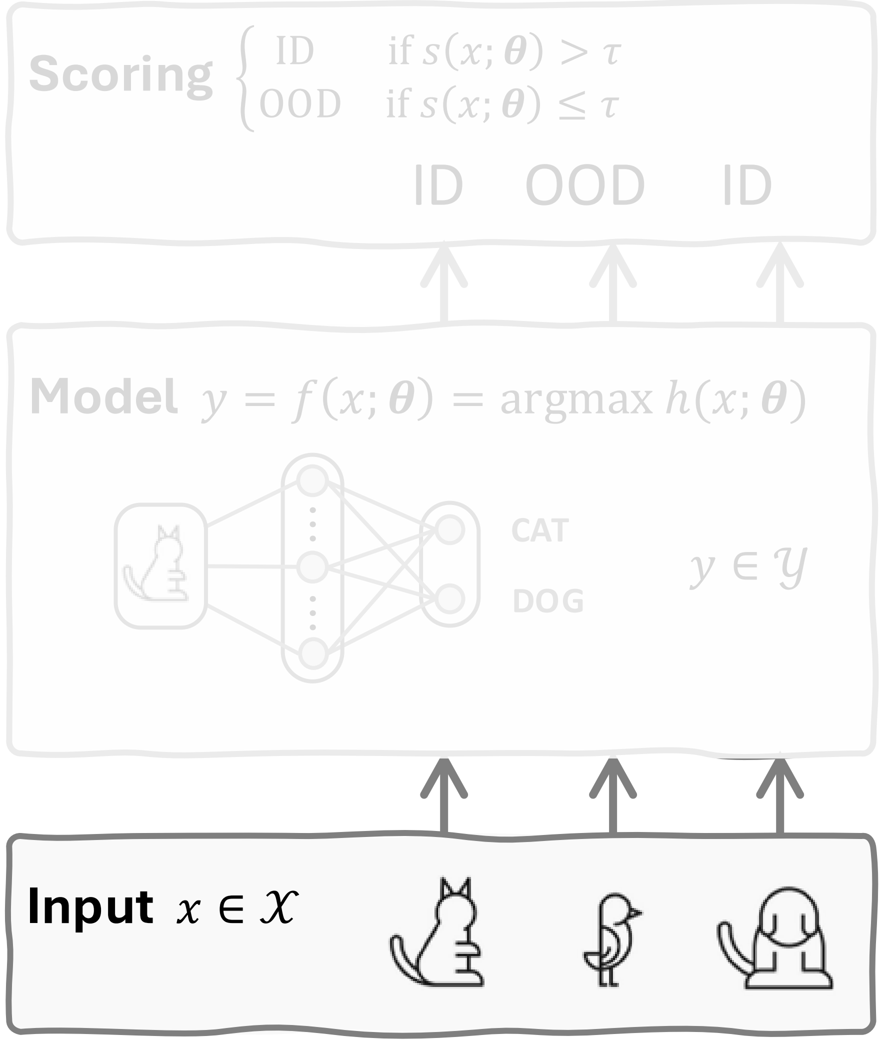



**Semantic Shift**: semantic relationship between inputs and labels changes.



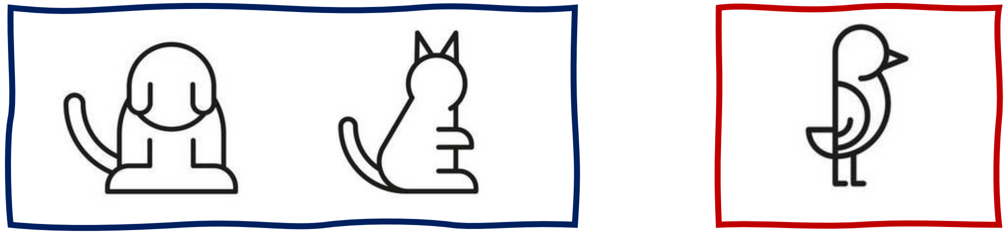

**Covariate Shift**: feature distribution changes while labels stay within the label space.

# Post-hoc OOD Detection: Review



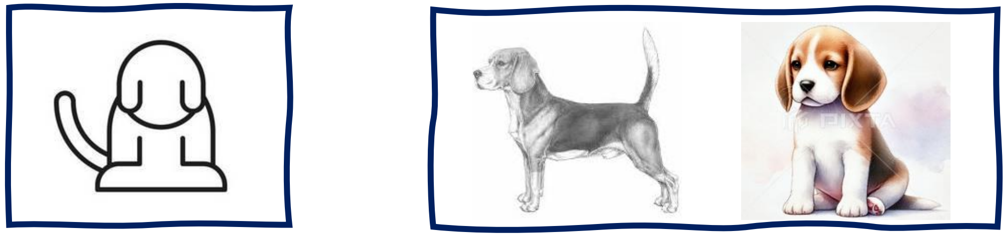

Interests of OOD detection

**Semantic Shift**: semantic relationship between inputs and labels changes.



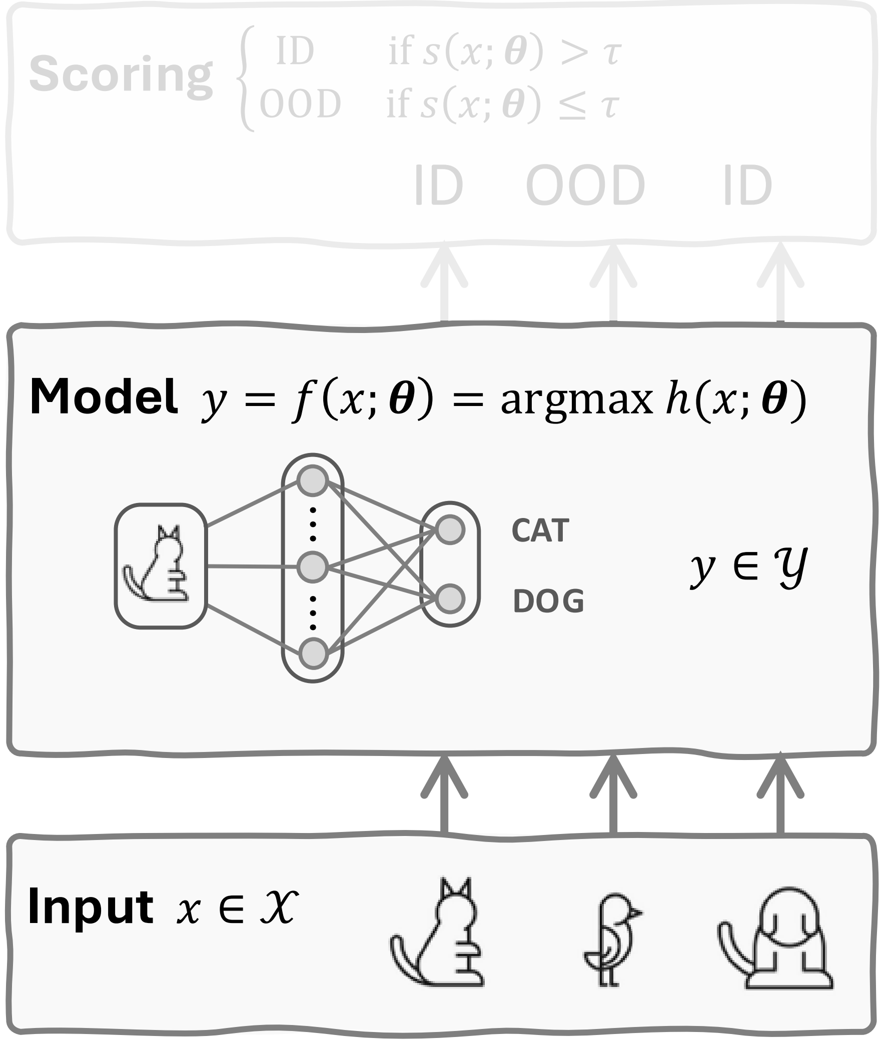
ID OOD

**Covariate Shift**: feature distribution changes while labels stay within the label space.

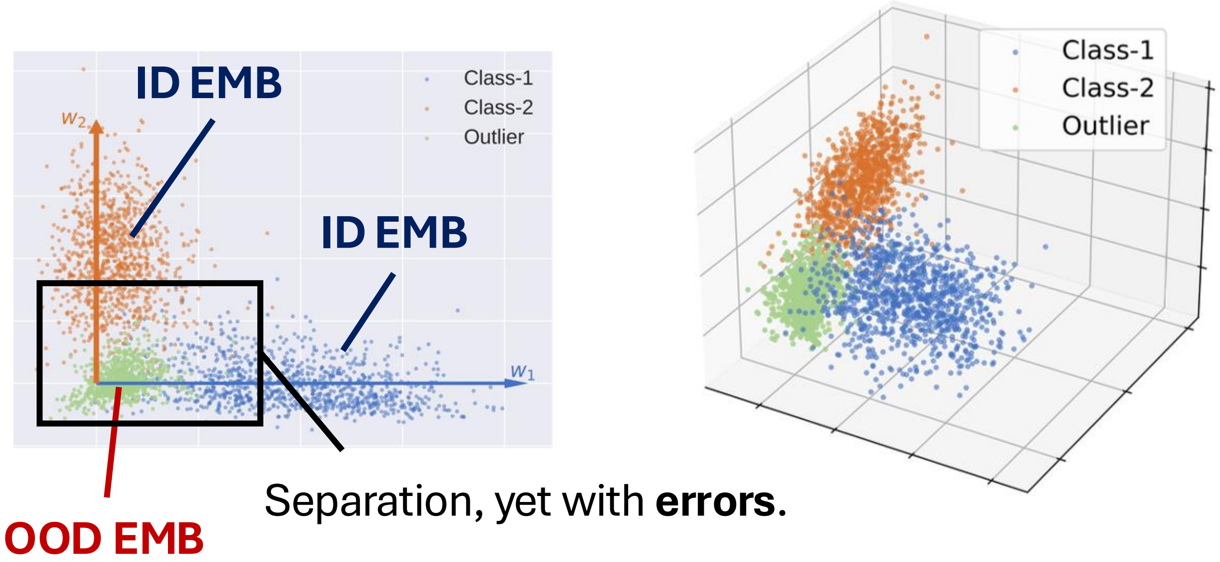


ID OOD

# Post-hoc OOD Detection: Review



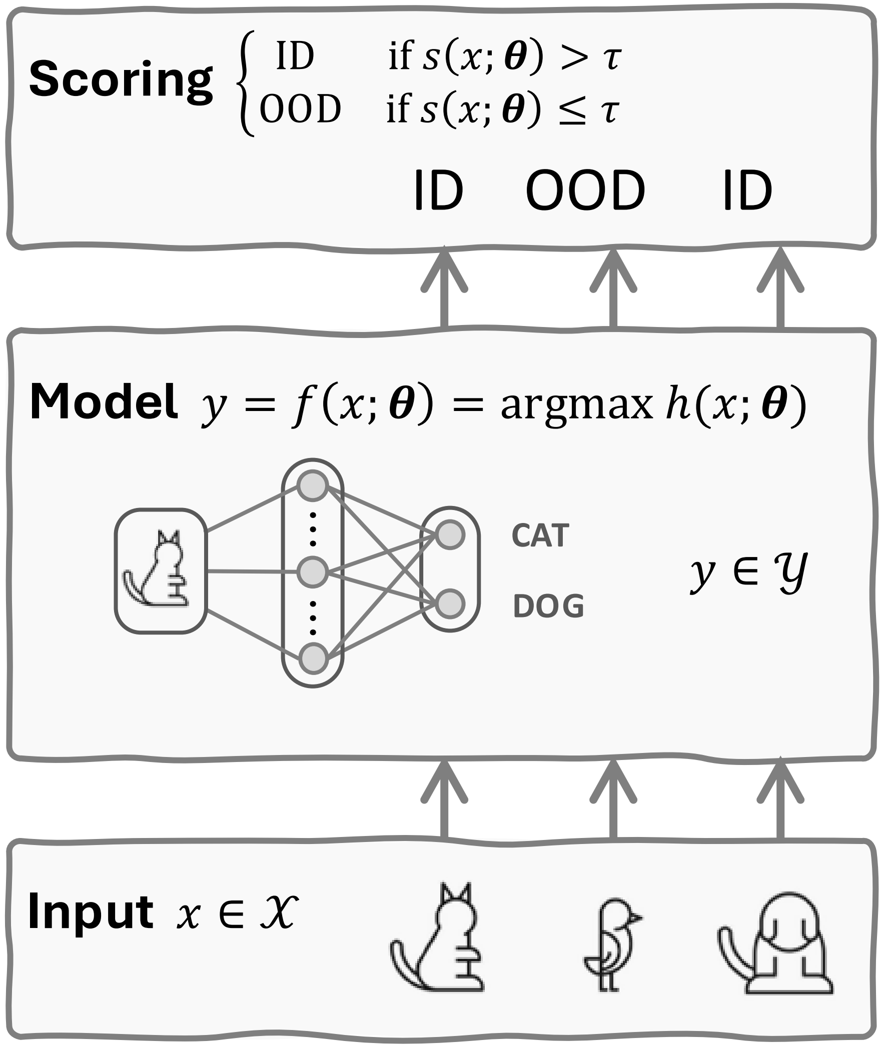

Pre-trained models can **separate ID and OOD data in embedding space** to some extent.




*Figures of dimensional-reduced embeddings [a].*

[a] Wu et al. Pursuing Feature Separation based on Neural Collapse for Out-of-Distribution Detection. In ICLR, 2025.

# Post-hoc OOD Detection: Review




**Model responses** can be used to craft OOD scoring functions.

maximal softmax prediction

❖ **Output Level**, MSP [a]

$$s_{\text{MSP}}(x;\boldsymbol{\theta}) = \max_{\text{k}} \text{softmax}_{\text{k}}\, h(x;\boldsymbol{\theta})$$

k-th nearest neighbor

❖ **Embedding Level**, KNN [b]

$$s_{\text{KNN}}(x;\boldsymbol{\theta}) = \left\| h(x;\boldsymbol{\theta}) - z_{(k)} \right\|_2$$

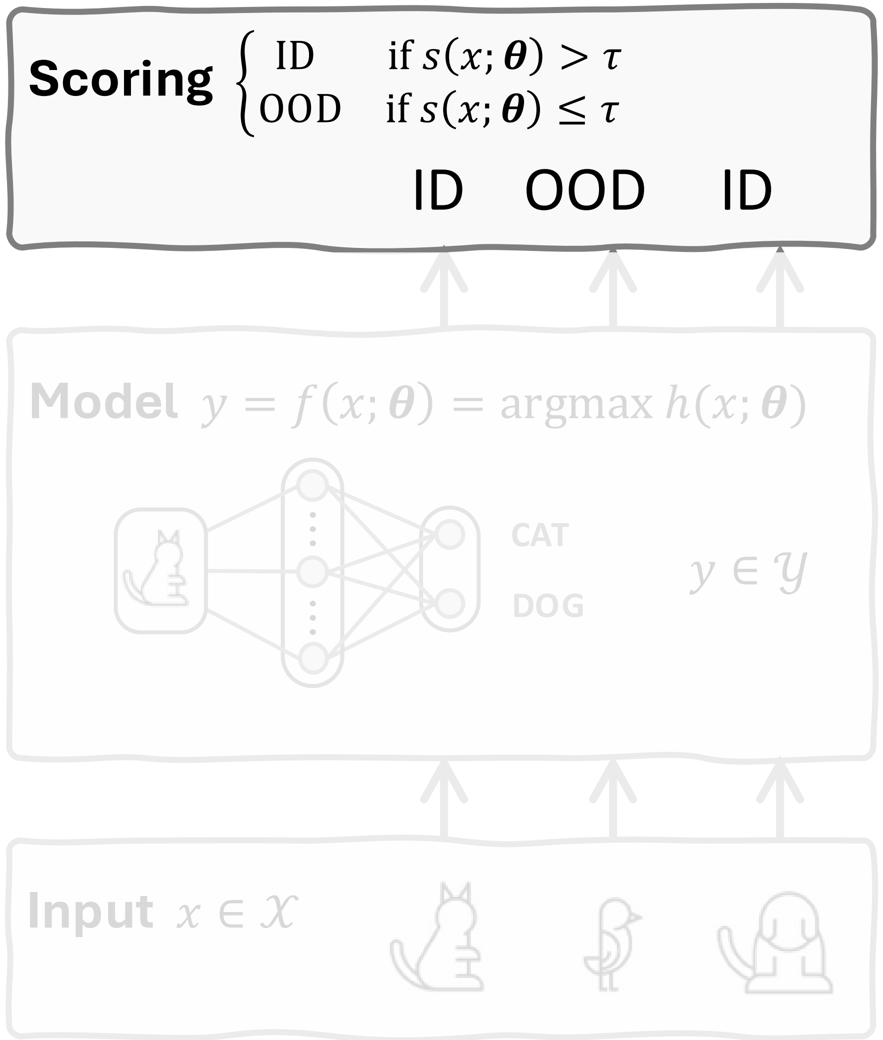
uniform distribution

❖ **Gradient Level**, GradNorm [c]

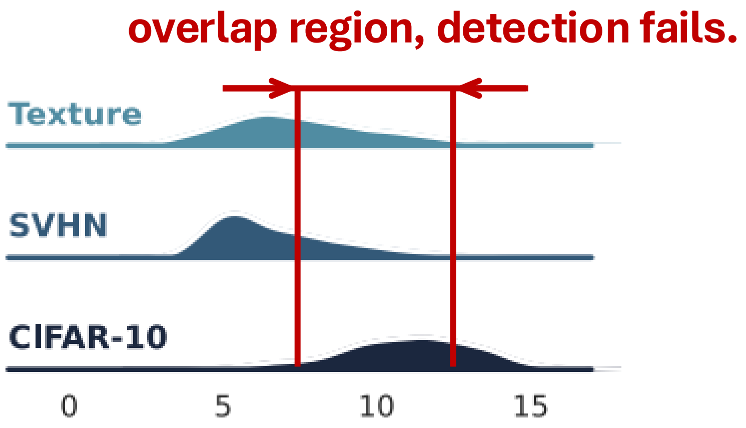
$$s_{\text{GN}}(x;\boldsymbol{\theta}) = \|\nabla_{\boldsymbol{\theta}} \text{KL}(u\|\text{softmax}(h(x;\boldsymbol{\theta})))\|_2$$


[a] Hendrycks et al. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In ICLR, 2017.
[b] Sun et al. Out-of-Distribution Detection with Deep Nearest Neighbors. In ICML, 2022.
[c] Huang et al. On the Importance of Gradients for Detecting Distributional Shifts in the Wild. In NeurIPS, 2021.

# Post-hoc OOD Detection: Challenges




Post-hoc OOD detection often **makes mistakes**, failing to discern many ID and OOD patterns.




*Figures of ID & OOD scoring distributions [a].*

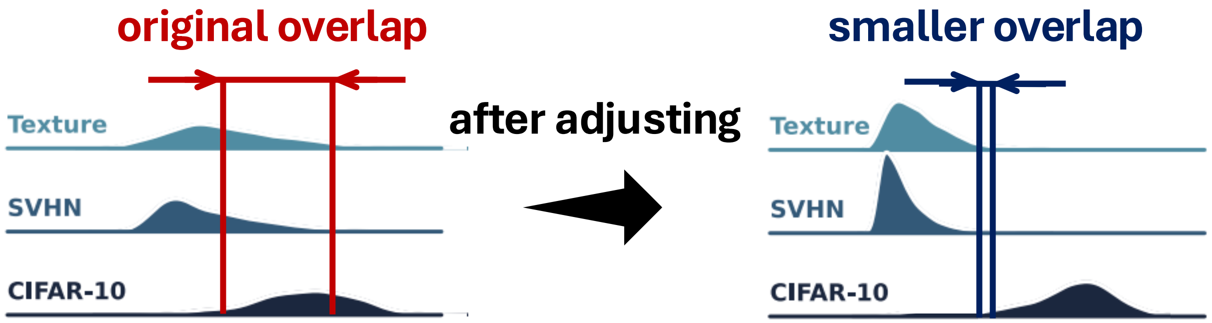
**Explanations:** For conventional-trained models, 1) their **representations** are not good enough, b) their **calibration** is inherently poor, and c) they cannot fully **classify** ID and OOD patterns.

[a] Wang et al. Watermarking for Out-of-distribution Detection. In NeurIPS, 2022.

# OOD Detection Learning

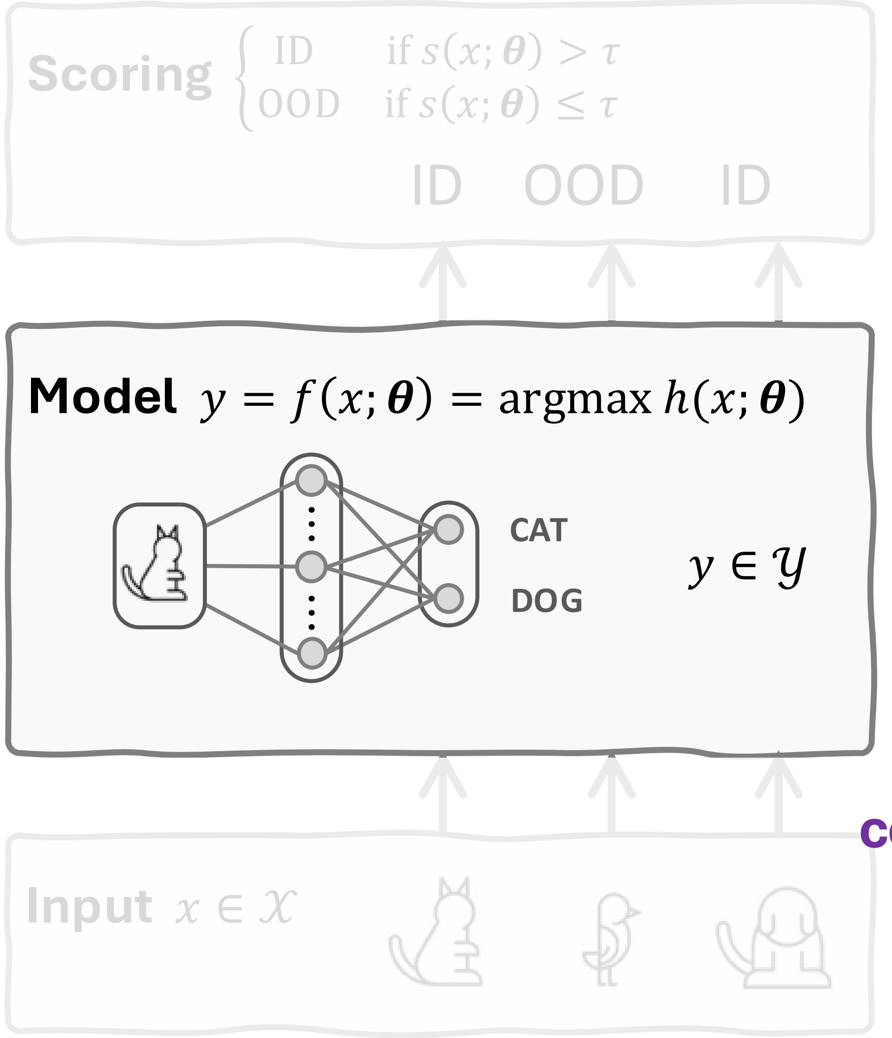



**Adjust** the system to improve OOD detection.

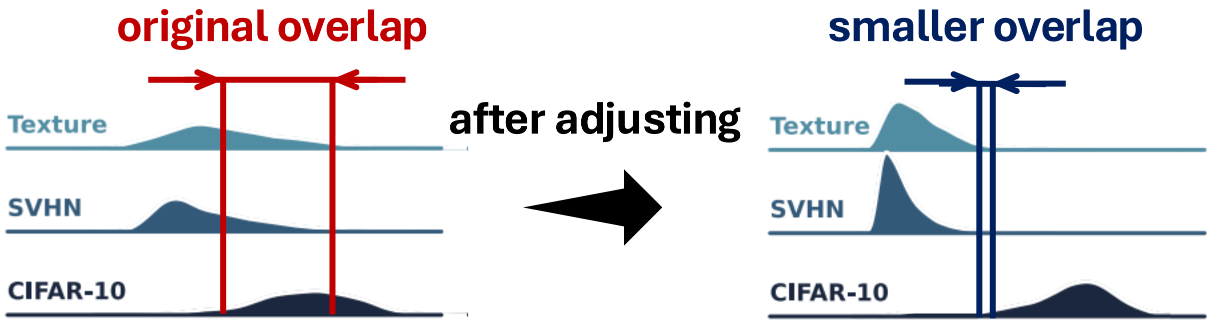



❖ **Model Level**, most works $s(x;\boldsymbol{\theta}) \rightarrow s(x;\boldsymbol{\theta}_{\text{adj}})$

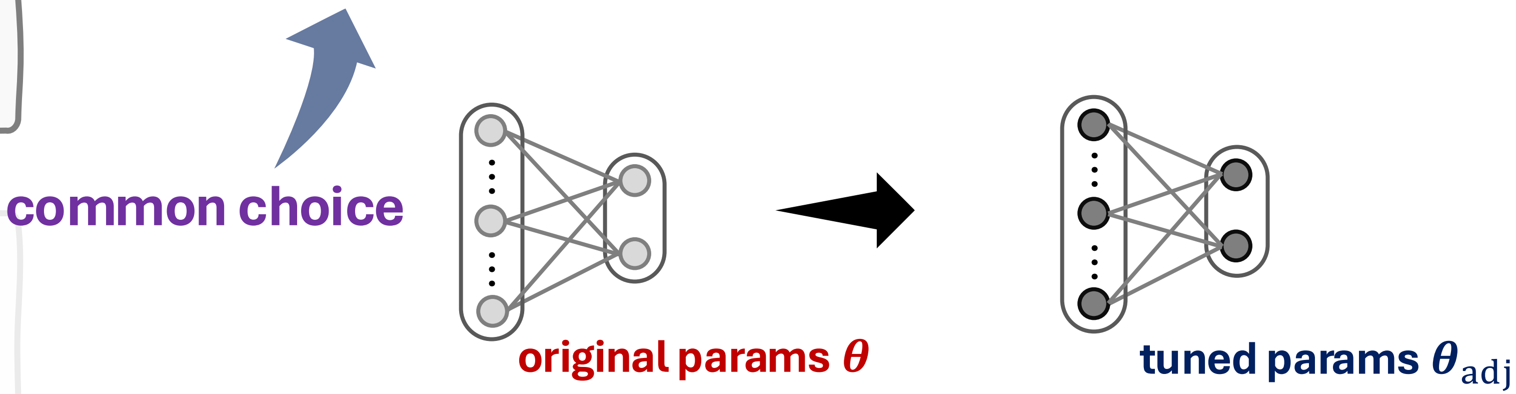
# OOD Detection Learning: Where to Adjust?



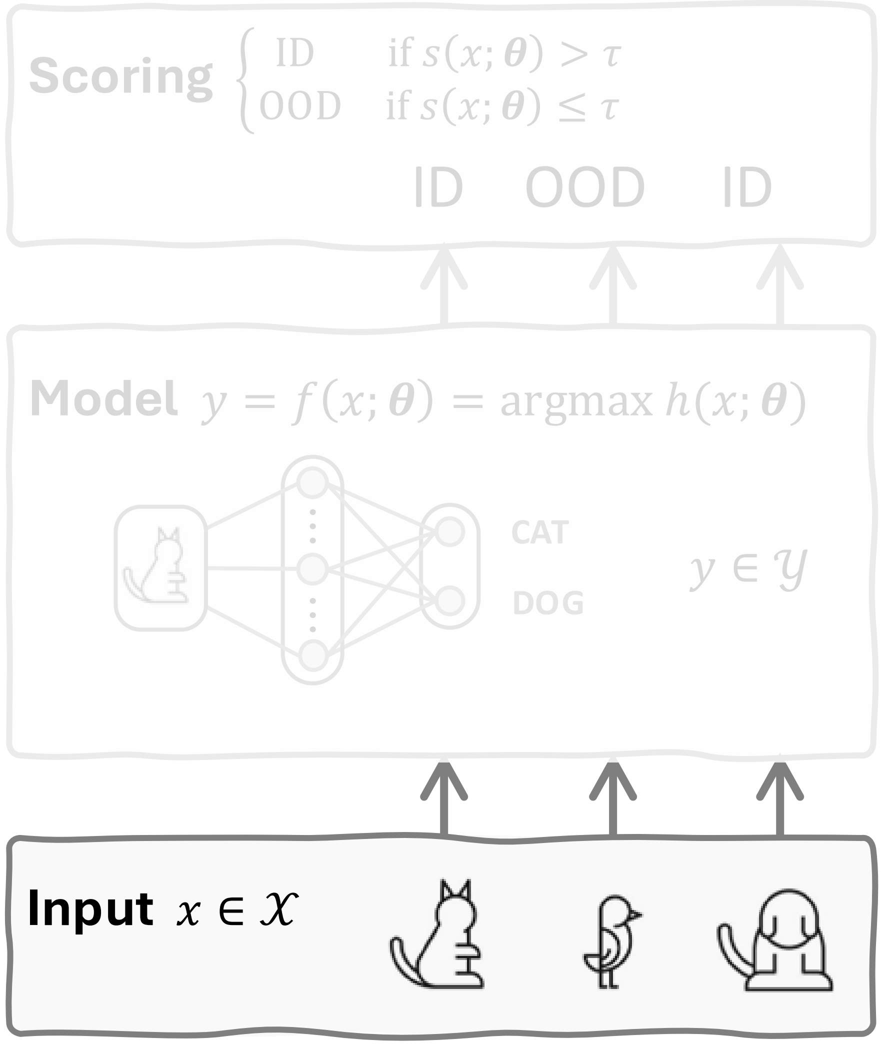

**Adjust** the system to improve OOD detection.



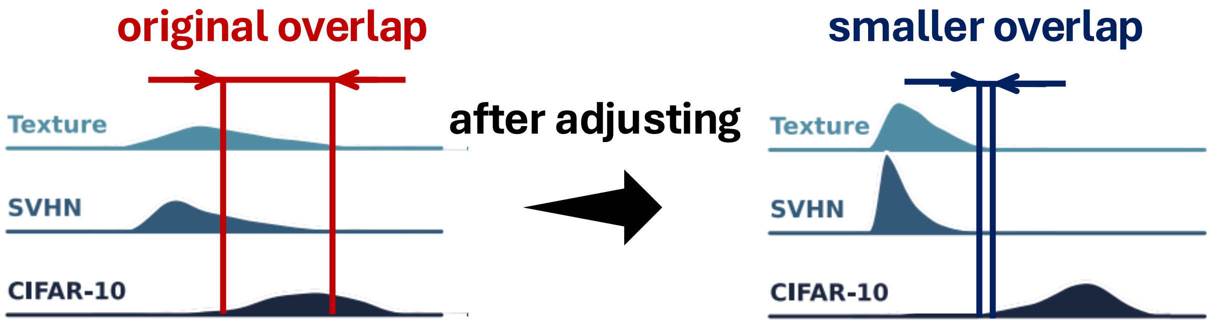

❖ **Model Level**, most works $s(x;\boldsymbol{\theta}) \rightarrow s(x;\boldsymbol{\theta}_{\text{adj}})$



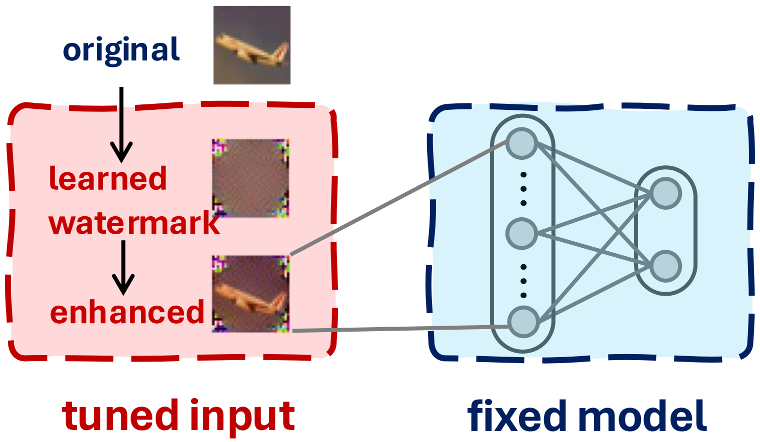
# OOD Detection Learning: Where to Adjust?




**Adjust** the system to improve OOD detection.



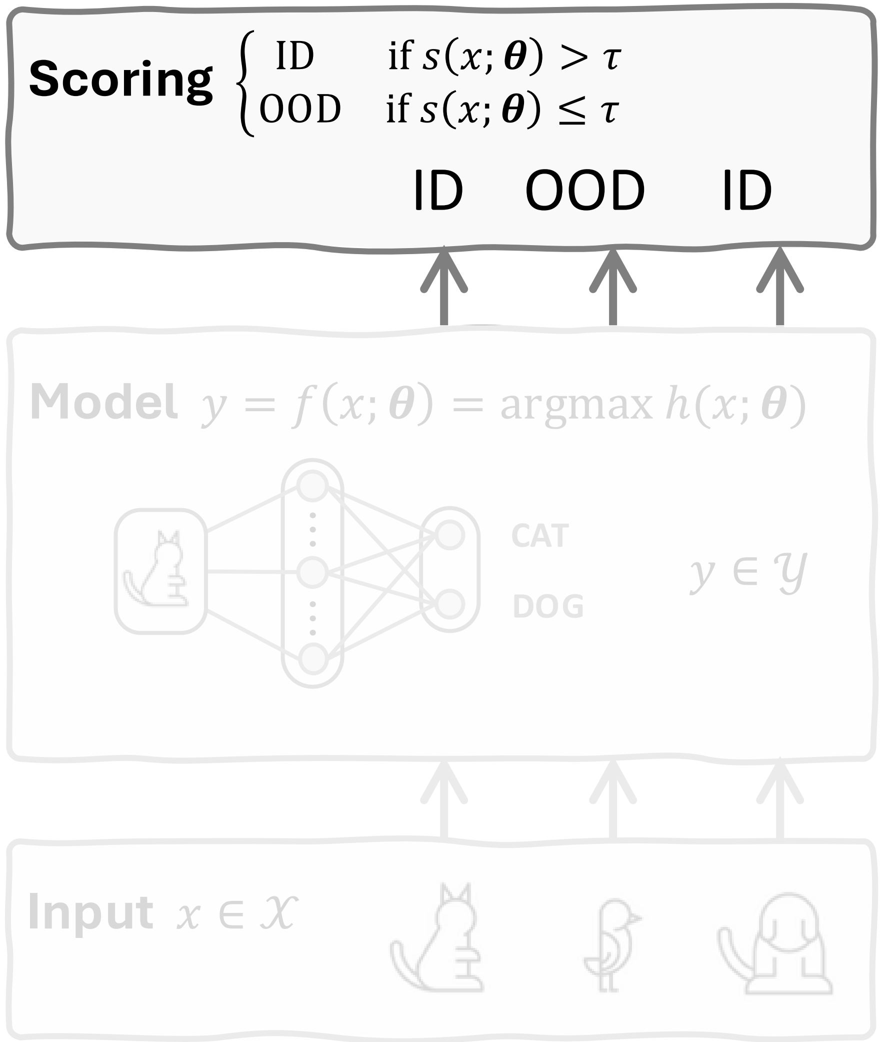

❖ **Input Level**, WM [a] $s(x;\boldsymbol{\theta}) \rightarrow s(x + w;\boldsymbol{\theta})$



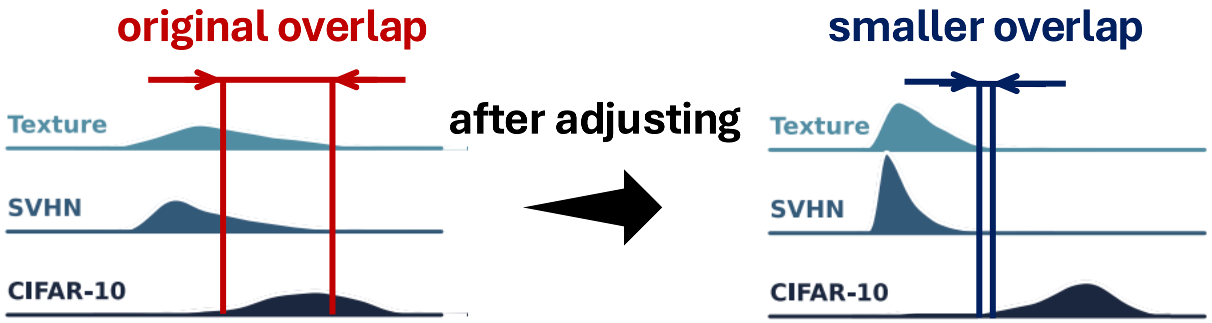

- *Watermark is static, **tuned** to enhance OOD detection.*
- *The pre-trained model remains **fixed**.*

[a] Wang et al. Watermarking for Out-of-distribution Detection. In NeurIPS, 2022.

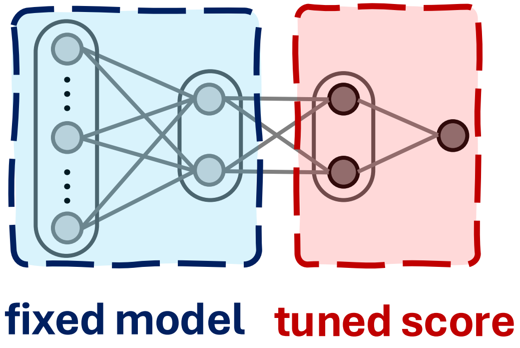
# OOD Detection Learning: Where to Adjust?




**Adjust** the system to improve OOD detection.




❖ **Score Level**, VOS [a] $s(x;\boldsymbol{\theta}) \rightarrow \boldsymbol{s}(h(x;\boldsymbol{\theta}),\boldsymbol{w})$



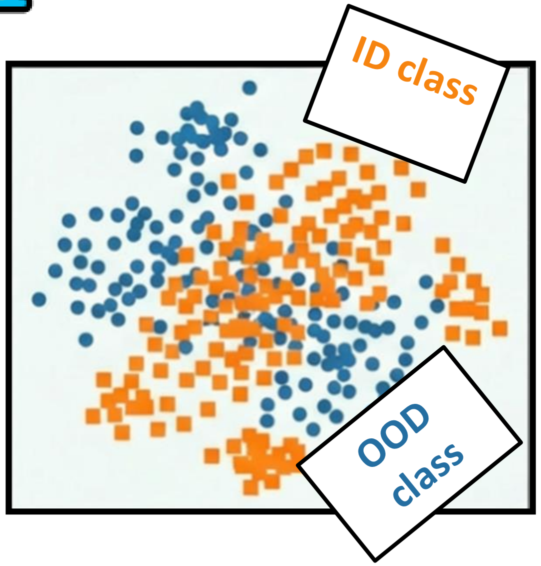

- ❖ *The scoring function introduces exact params to be* ***tuned****.*
- ❖ *The pre-trained model remains* ***fixed****.*

[a] Du et al. Learning What You Don't Know by Virtual Outlier Synthesis. In ICLR, 2022.

# OOD Detection Learning: How to Adjust?

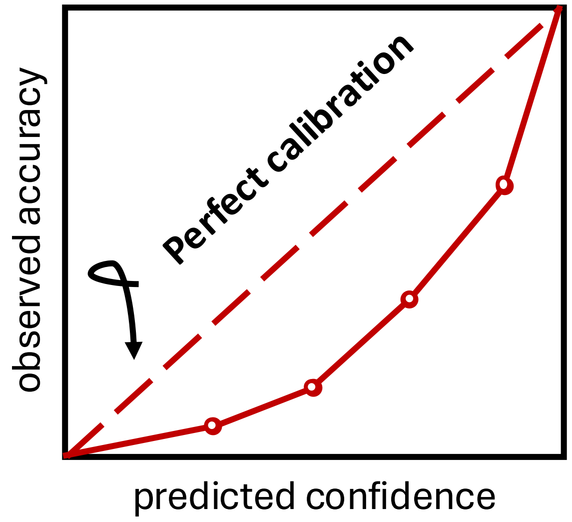
Let's recall **why does post-hoc OOD detection often fail**: For conventional-trained models, they have

## Poor Representation



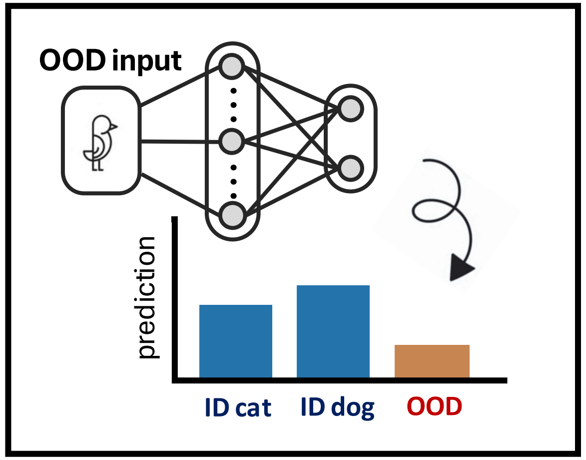

Data with **different semantics** may **not be perfectly separated** in the embedding space.

## Poor Calibration




**High model confidence** does not correspond to **high model accuracy**.

## Poor Classification

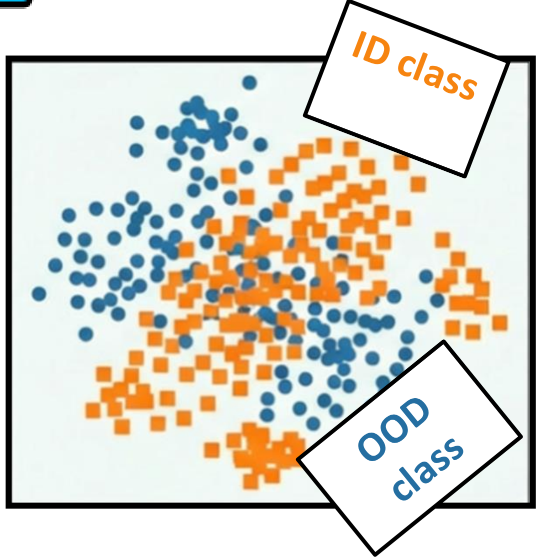



The model **predicts the wrong class**, despite the true class of being either ID or OOD.

# OOD Detection Learning: How to Adjust?

Let's recall why does post-hoc OOD detection often fail: For conventional-trained models, they have

## Poor Representation

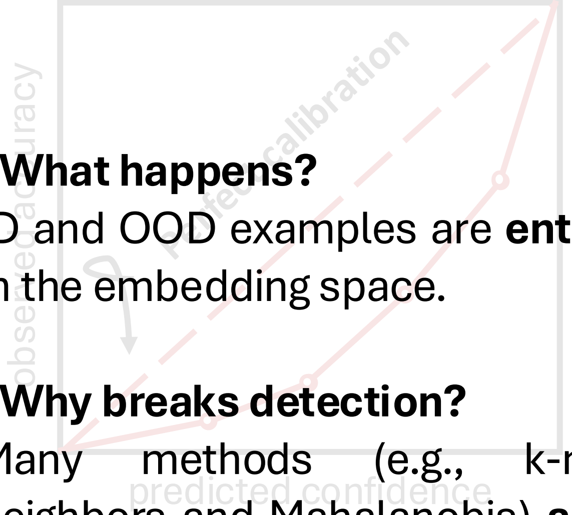



Data with **different semantics** may **not be perfectly separated** in the embedding space.

❖ **What happens?**
ID and OOD examples are **entangled** in the embedding space.

❖ **Why breaks detection?**
Many methods (e.g., k-nearest neighbors and Mahalanobis) **assume that OOD lies away from ID.**

## Poor Calibration




not correspond to **high model accuracy**.

## Poor Classification




The model **predicts the wrong class**, despite the true class of being either ID or OOD.

# OOD Detection Learning: How to Adjust?

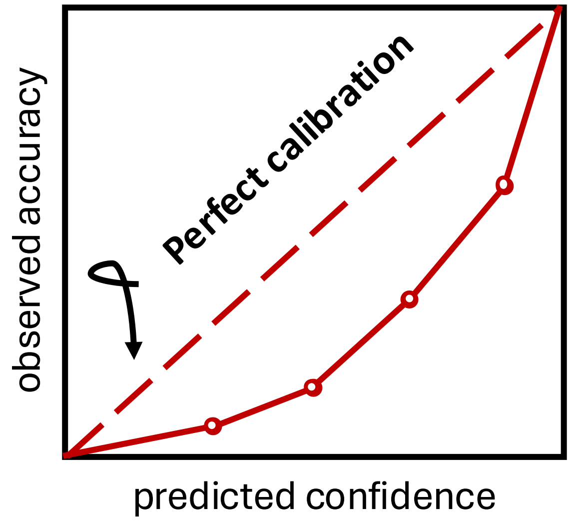
Let's recall why does post-hoc OOD detection often fail: For conventional-trained models, they have

### 😭 Poor Representation



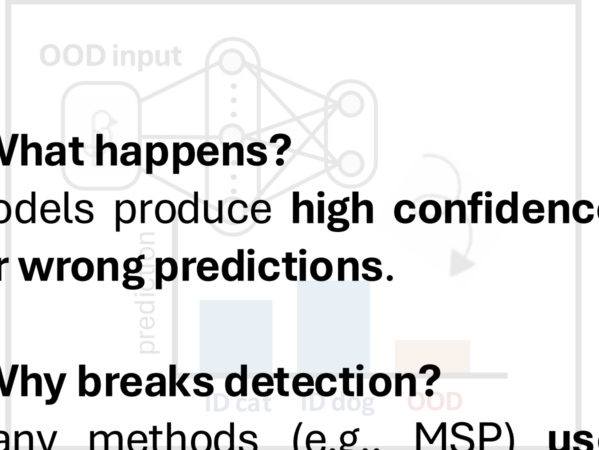

Data with **different semantics** may **not be perfectly separated** in the embedding space.

### 😭 Poor Calibration




**High model confidence** does not correspond to **high model accuracy**.

### 😭 Poor Classification




❖ **What happens?**
Models produce **high confidence for wrong predictions**.

❖ **Why breaks detection?**
Many methods (e.g., MSP) **use confidence-like scores.**

The model **predicts the wrong class**, despite the true class of being either ID or OOD.

# OOD Detection Learning: How to Adjust?

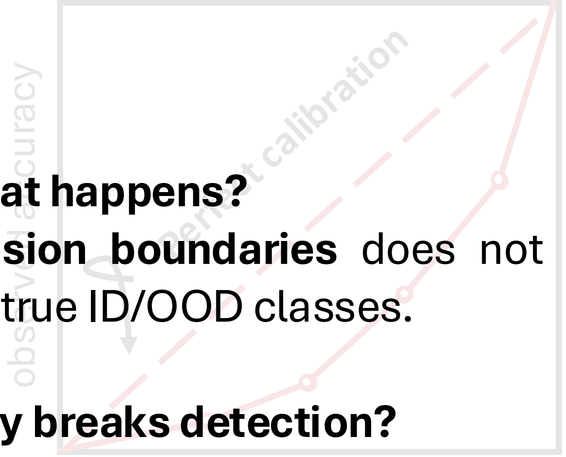
Let's recall why does post-hoc OOD detection often fail: For conventional-trained models, they have

## Poor Representation



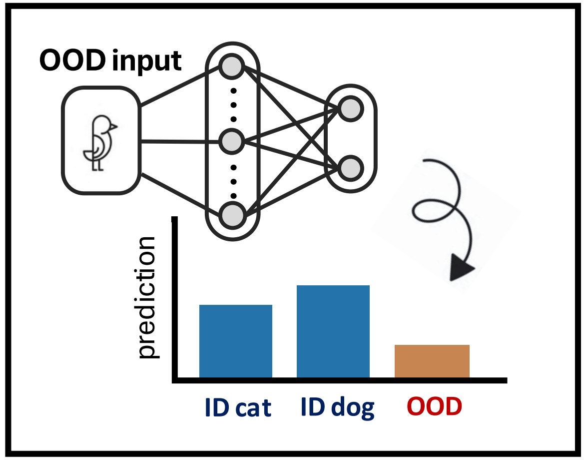

Data with **different semantics** may **not be perfectly separated** in the embedding space.

## Poor Calibration




**High model confidence** does not correspond to **high model accuracy**.

## Poor Classification



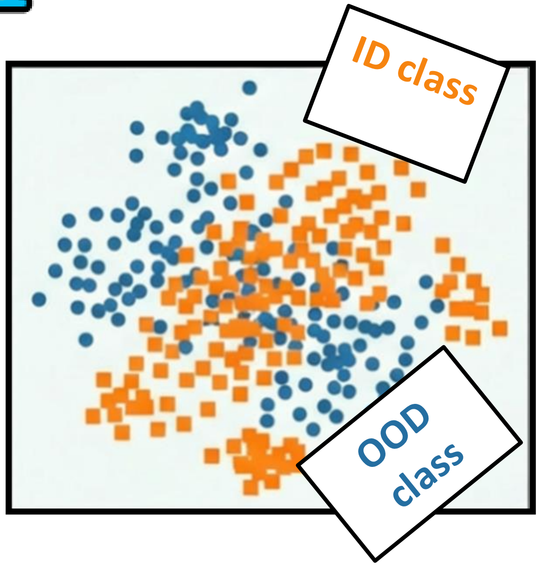

The model **predicts the wrong class**, despite the true class of being either ID or OOD.

- **What happens?**
  **Decision boundaries** does not align with true ID/OOD classes.
- **Why breaks detection?**
  Taking as an **extra classification task**, this classifier is not accurate.

# OOD Detection Learning: How to Adjust?

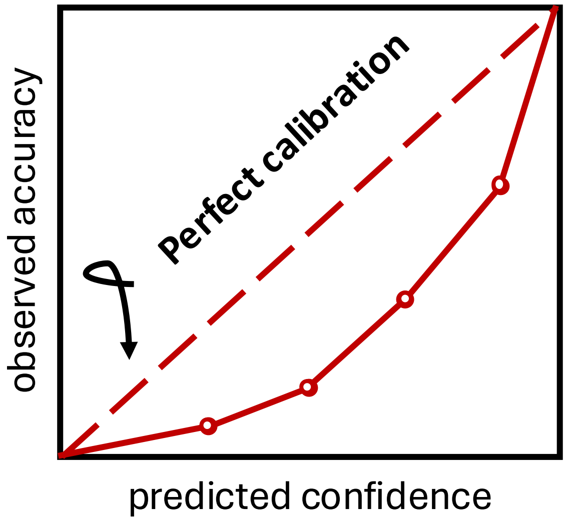
Let's recall why does post-hoc OOD detection often fail: For conventional-trained models, they have

### Poor Representation



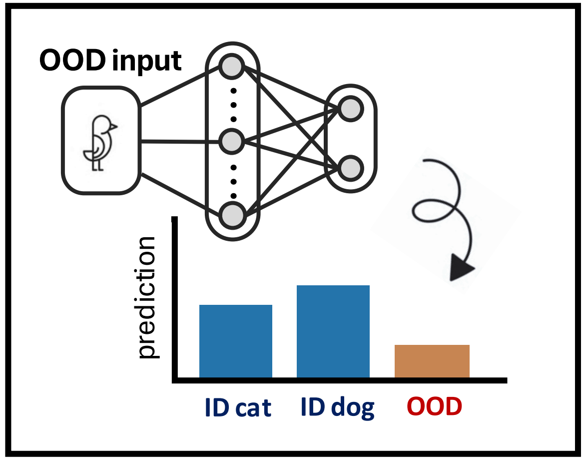

Data with **different semantics** may **not be perfectly separated** in the embedding space.

### Poor Calibration




**High model confidence** does not correspond to **high model accuracy**.

### Poor Classification

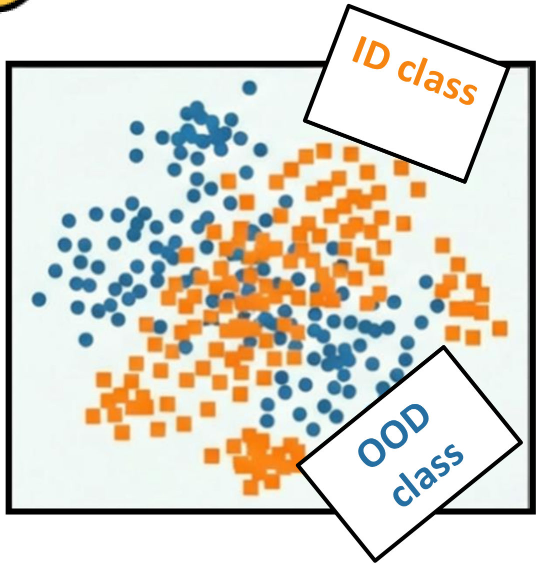



The model **predicts the wrong class**, despite the true class of being either ID or OOD.

# OOD Detection Learning: How to Adjust?

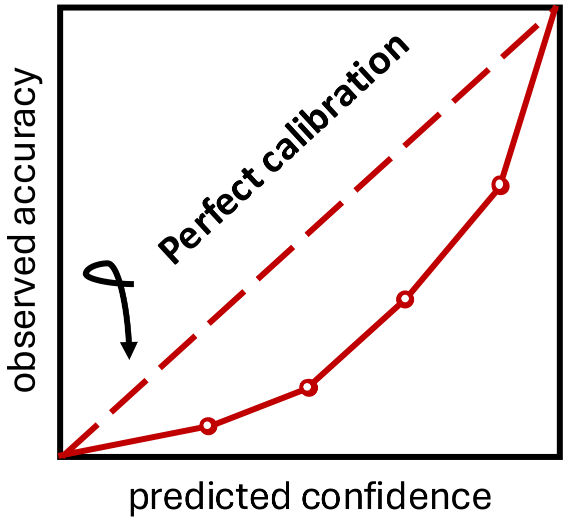
To **address the drawbacks** of post-hoc OOD detection, OOD detection learning can

### Improve Representation



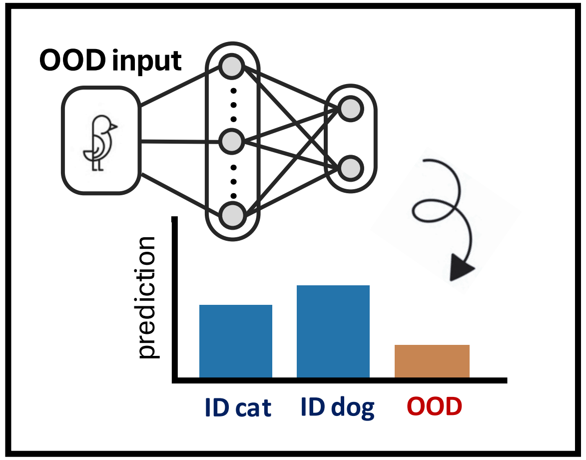

- **contrastive learning,**
- **reconstruction learning,**
- **pre-training,** et al.

### Improve Calibration




- **calibration techniques,**
- **density regularization,**
- **distribution modelling,** et al.

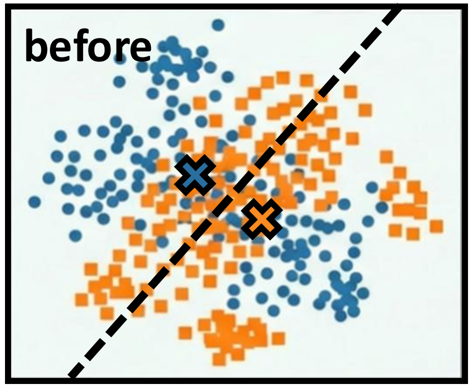
### Improve Classification




- **outlier exposure,**
- **data augmentation,**
- **sample selection**, et al.

# Representation: Overview

**Conventional-trained Classifiers** → **Representation-based OOD Learning**




ID/OOD boundary

**A simple detection strategy**: Nearest distances to K-means clustering.

$$s_{\text{KM}}(x;\boldsymbol{\theta}) = \left\| h(x;\boldsymbol{\theta}) - \mu_{(x)} \right\|_2$$

nearest centroid (ID/OOD)

- mainly learn patterns among training classes as a **shortcut to learn classification.**
- make **no assumptions about the form of OOD data** will take or the type of downstream task.




**Representations with different semantics are better separated.**

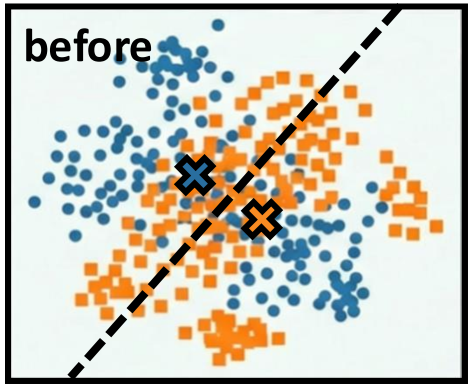
- low intra-class variance
- high inter-class variance

**Pretext tasks**

- **Contrastive Learning**: CSI, SSD
- **Reconstruction Learning**: MOOD
- **Pre-training**: CLIP

# Representation: Overview

**Conventional-trained Classifiers** → **Representation-based OOD Learning**

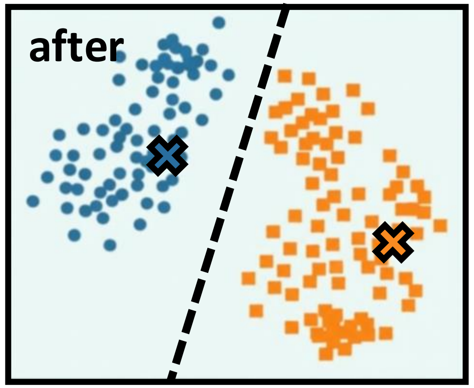



**ID/OOD boundary**

**A simple detection strategy**: Nearest distances to K-means clustering.

$$s_{\mathrm{KM}}(x;\boldsymbol{\theta}) = \left\| h(x;\boldsymbol{\theta}) - \mu_{(x)} \right\|_2$$

nearest centroid (ID/OOD)




**Representations with different semantics are better separated.**

- low intra-class variance
- high inter-class variance

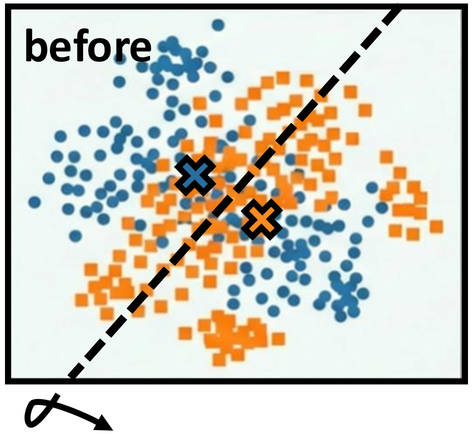
**Pretext tasks**

- **Contrastive Learning**: CSI, SSD
- **Reconstruction Learning**: MOOD
- **Pre-training**: CLIP

- mainly learn patterns among training classes as a **shortcut to learn classification.**
- make **no assumptions about the form of OOD data** will take or the type of downstream task.

# Representation: Overview

## Conventional-trained Classifiers → Representation-based OOD Learning

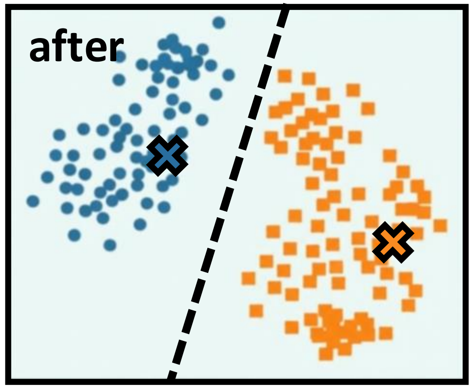



**A simple detection strategy**: Nearest distances to K-means clustering.

$$s_{\mathrm{KM}}(x;\boldsymbol{\theta}) = \left\| h(x;\boldsymbol{\theta}) - \mu_{(x)} \right\|_2$$

nearest centroid (ID/OOD)

- ❖ mainly learn patterns among training classes as a **shortcut to learn classification.**
- ❖ make **no assumptions about the form of OOD data** will take or the type of downstream task.




**Representations with different semantics are better separated.**

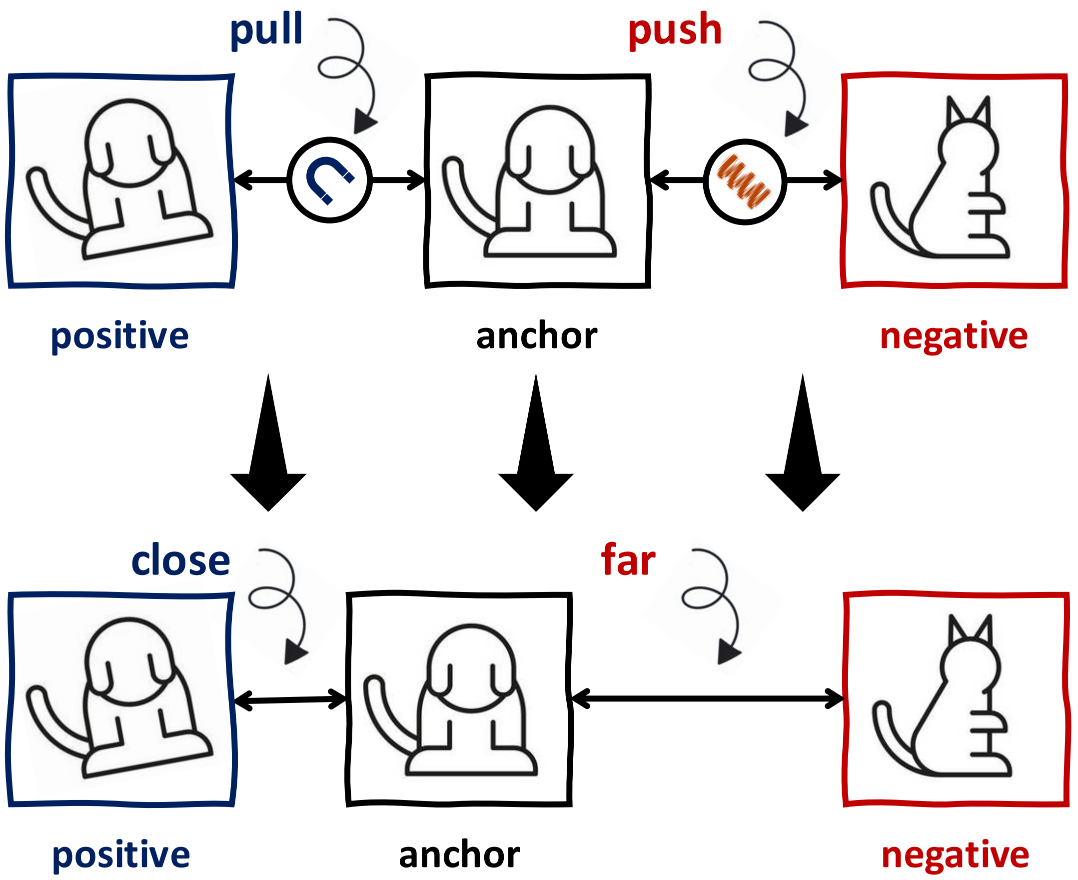
- ❖ low intra-class variance
- ❖ high inter-class variance

**Pretext tasks**

- ❖ **Contrastive Learning**: CSI, SSD
- ❖ **Reconstruction Learning**: MOOD
- ❖ **Pre-training**: CLIP

# Representation: Contrastive Learning

Contrastive learning **improves the semantic structure of the embedding space** by **pulling** semantically similar samples together and **pushing** dissimilar ones apart.




**Contrastive Learning**

positive anchor

$$\mathcal{L}_{\text{con}}(x, \{x_+\}, \{x_-\}) = -\frac{1}{|\{x_+\}|} \log \frac{\sum_{x' \in \boxed{\{x_+\}}} \exp\{\text{sim}(\boxed{z(x), z}(x'))\}}{\sum_{x' \in \{x_+\} \boxed{\cup \{x_-\}}} \exp\left(\text{sim}(z(x), z(x'))\right)}$$

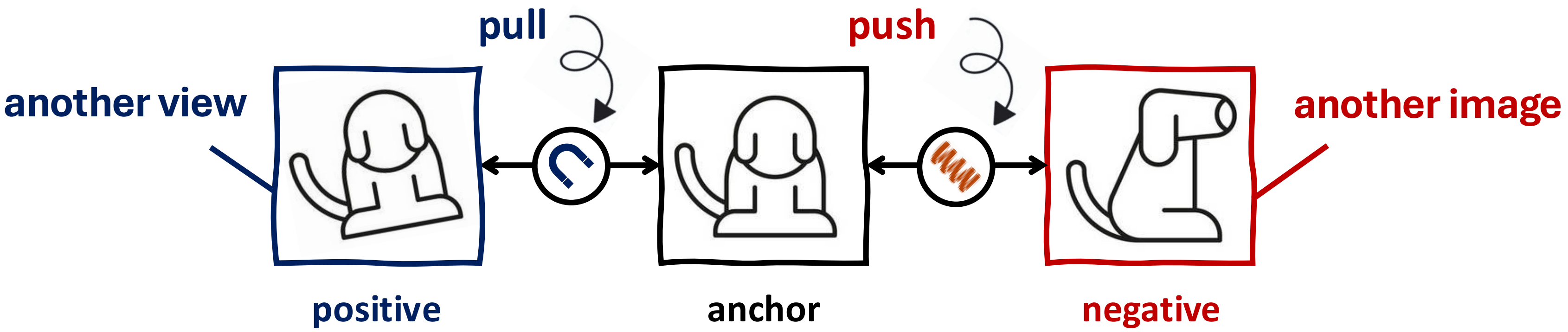
negative

***Increasing similarity*** *between the anchor and* ***positive samples****, while* ***decreasing similarity*** *to* ***negative samples****.*

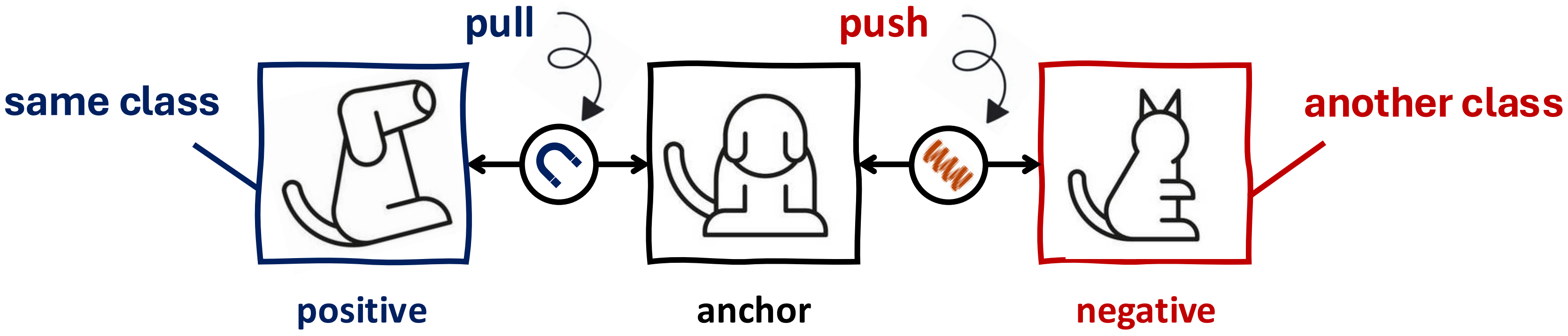
# Representation: Contrastive Learning | SSD

**SSD [a] Training Objective.**

❖ **SSD**, discriminating between **individuals (label-agnostic)**.




❖ **SSD+**, discriminating between **classes (label-aware)**.




[a] Sehwag et al. SSD: A Unified Framework for Self-supervised Outlier Detection. In ICLR, 2021.

# Representation: Contrastive Learning | SSD

**SSD OOD Detector.**

❖ **Mahalanobis**, cluster-conditioned detection **(OOD-agnostic)**.

k-th ID cluster centroid

$$s_{\mathrm{mah}}(x;\boldsymbol{\theta}) = \min_k (z(x)-\mu_k)^\top \boxed{\Sigma_k^{-1}} (z(x)-\boxed{\mu_k})$$

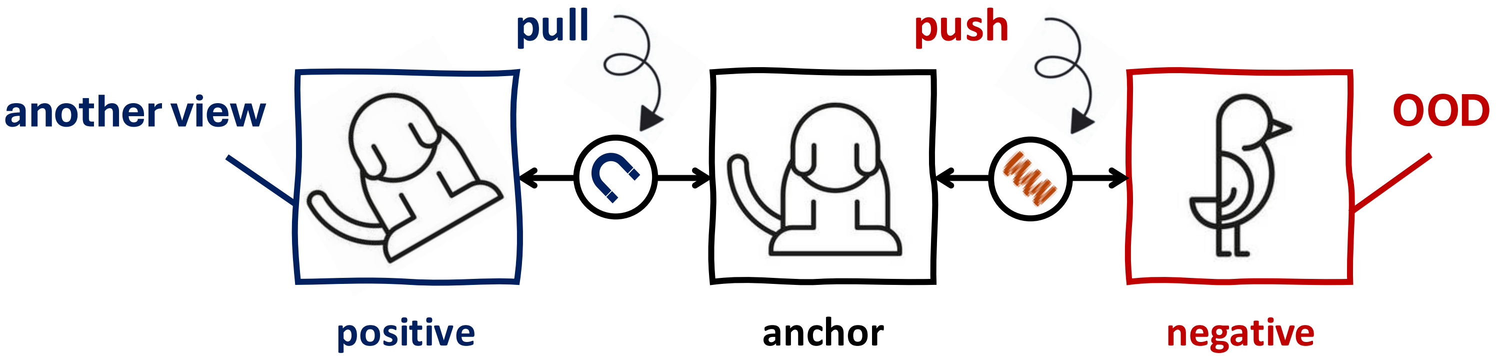
k-th ID cluster covariance, decorrelating features

❖ **Mahalanobis+**, a small OOD set is available before inference **(few-shot OOD)**.

$$s_{\mathrm{mah+}}(x;\boldsymbol{\theta}) = \underbrace{(z(x)-\mu_{id})^\top \Sigma_{id}^{-1}(z(x)-\mu_{id})}_{\textbf{distance to overall ID centroid}} - \underbrace{(z(x)-\mu_{ood})^\top \Sigma_{ood}^{-1}(z(x)-\mu_{ood})}_{\textbf{distance to overall OOD centroid}}$$

# Representation: Contrastive Learning | CSI

**CSI [a]** further studies effective **negative transformation** to discriminate OOD samples.




**What if no OOD data are available?** Those transformations previously found **ineffective as positive transformations** can instead be used as negative transformations.

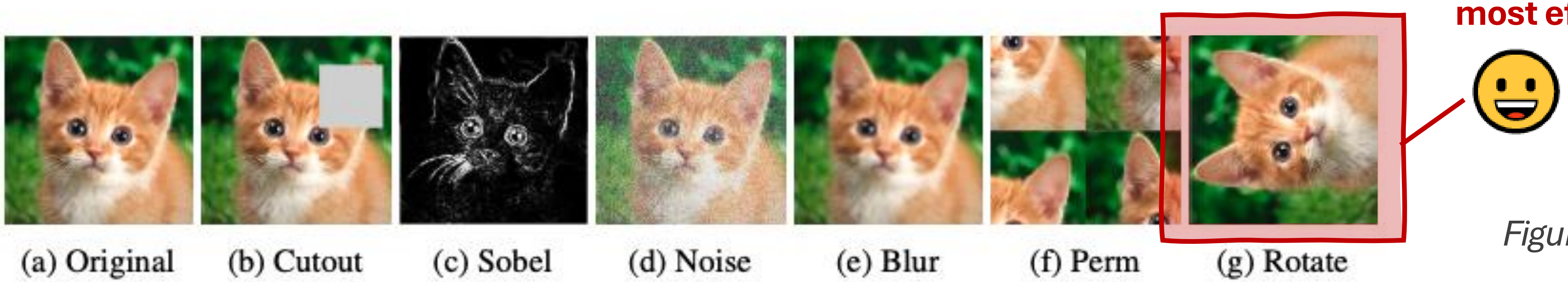
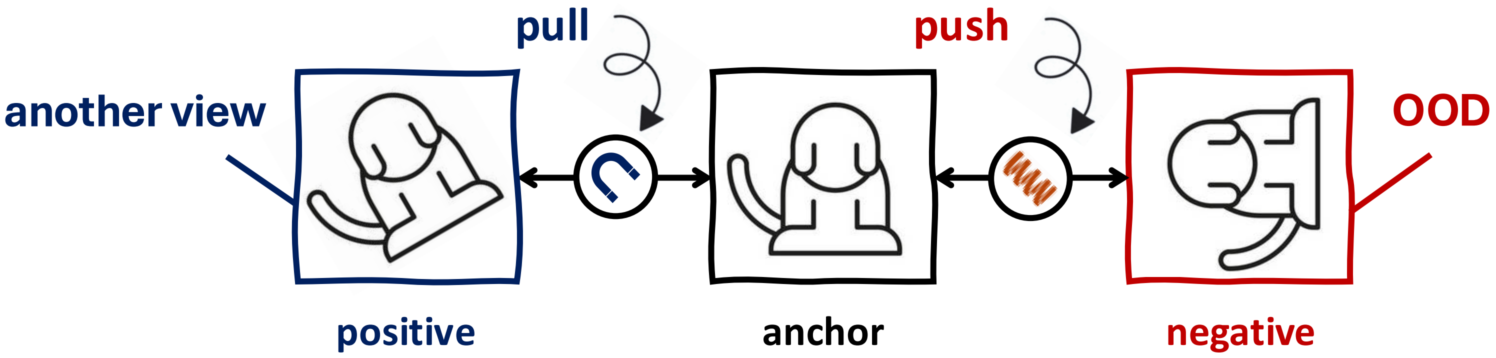



*Figure from [a].*

[a] Tack et al. CSI: Novelty Detection via Contrastive learning on Distributionally Shifted Instances. In NeurIPS, 2022.

# Representation: Contrastive Learning | CSI

**CSI [a]** further studies effective **negative transformation** to discriminate OOD samples.




**What if no OOD data are available?** Those transformations previously found **ineffective as positive transformations** can instead be used as negative transformations.

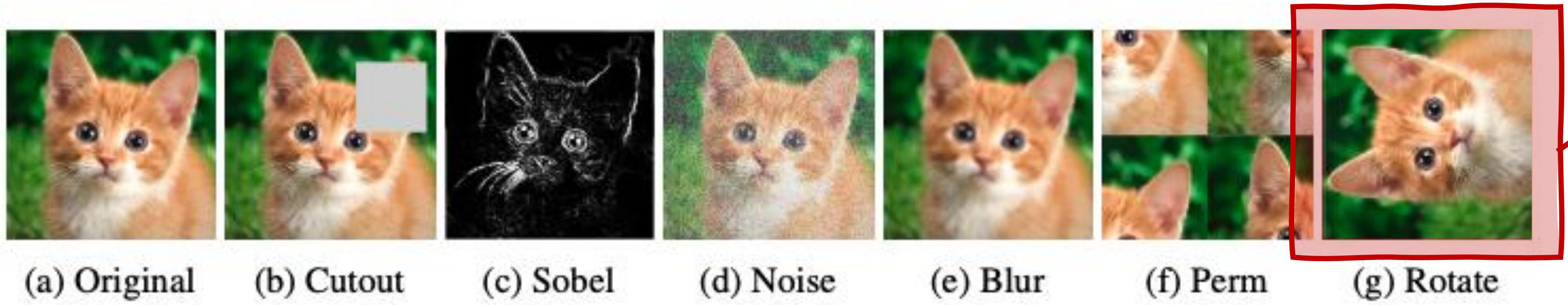
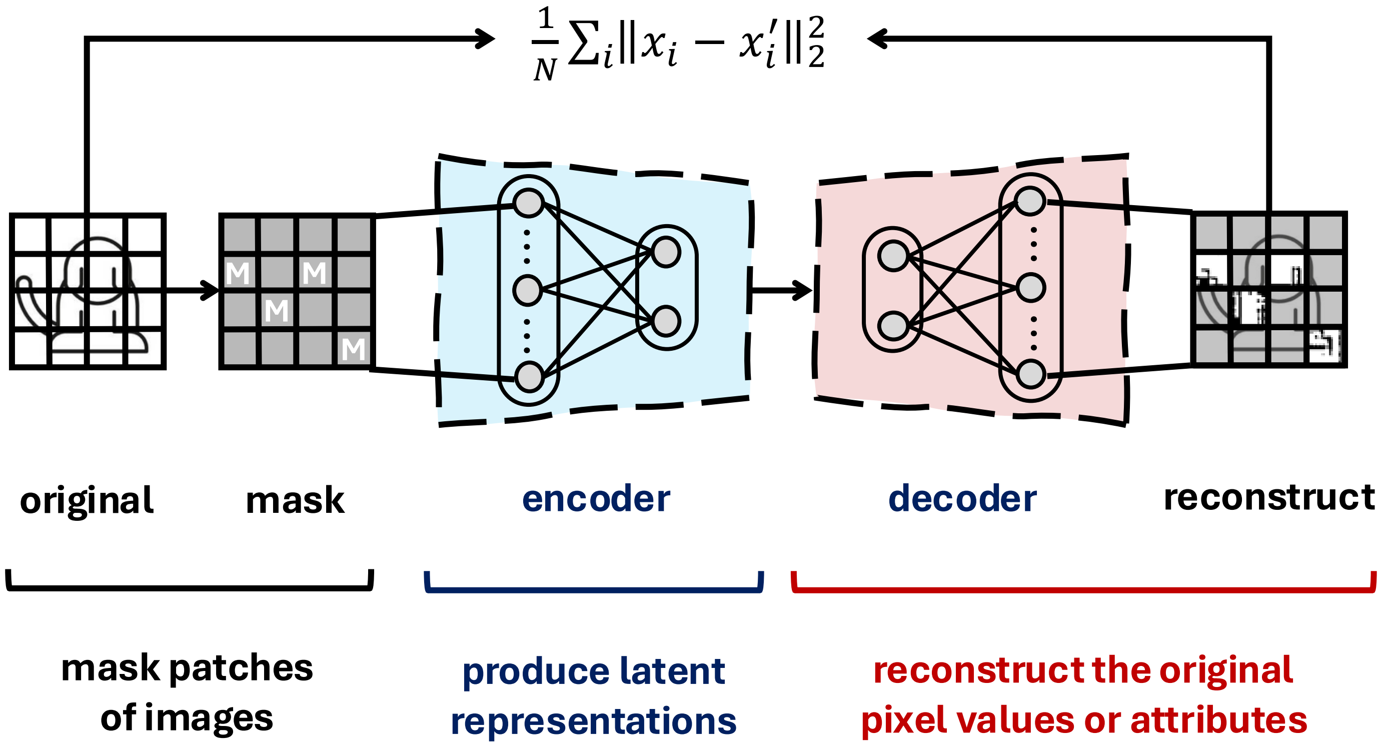



*Figure from [a].*

[a] Tack et al. CSI: Novelty Detection via Contrastive learning on Distributionally Shifted Instances. In NeurIPS, 2022.

# Representation: Reconstruction Learning | MOOD

**MOOD [a]** forces the models to **encode finer-grained information needed to rebuild input**, which is far beyond semantics as in contrastive learning.



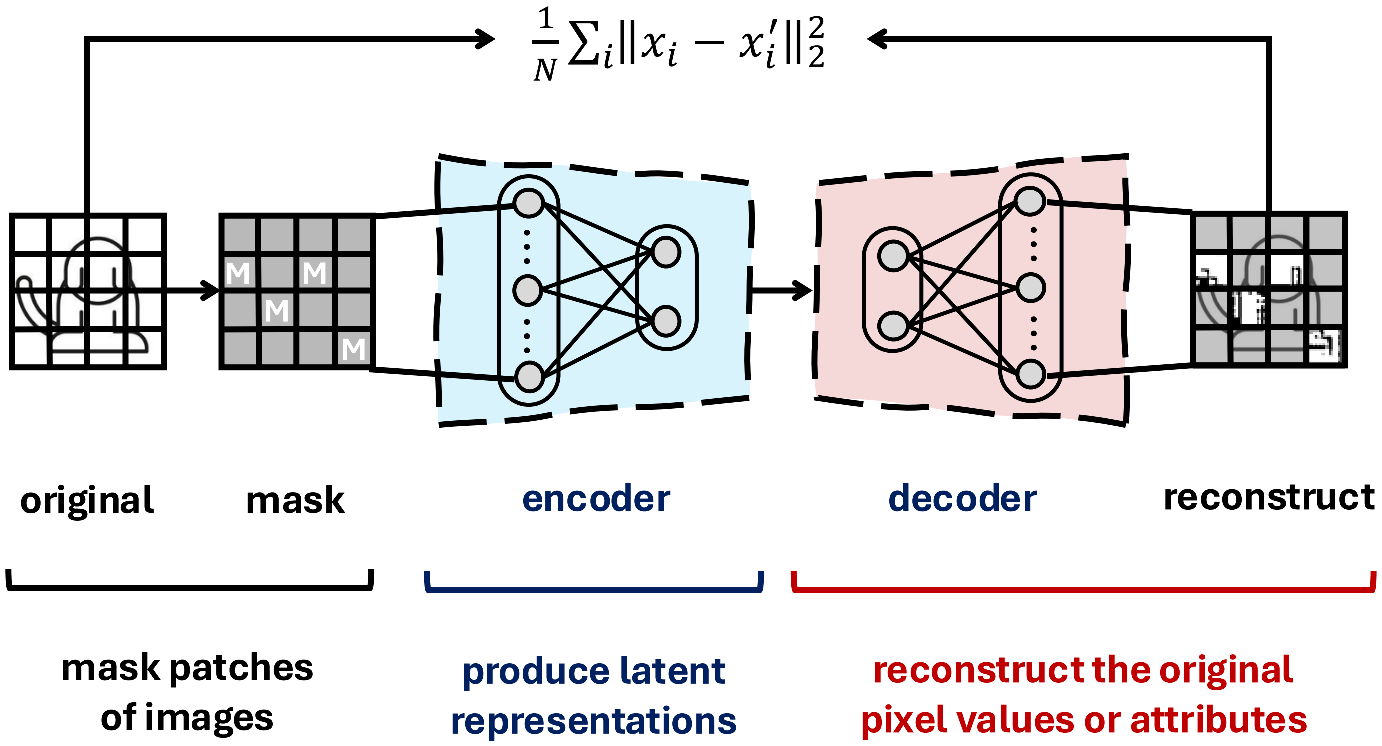


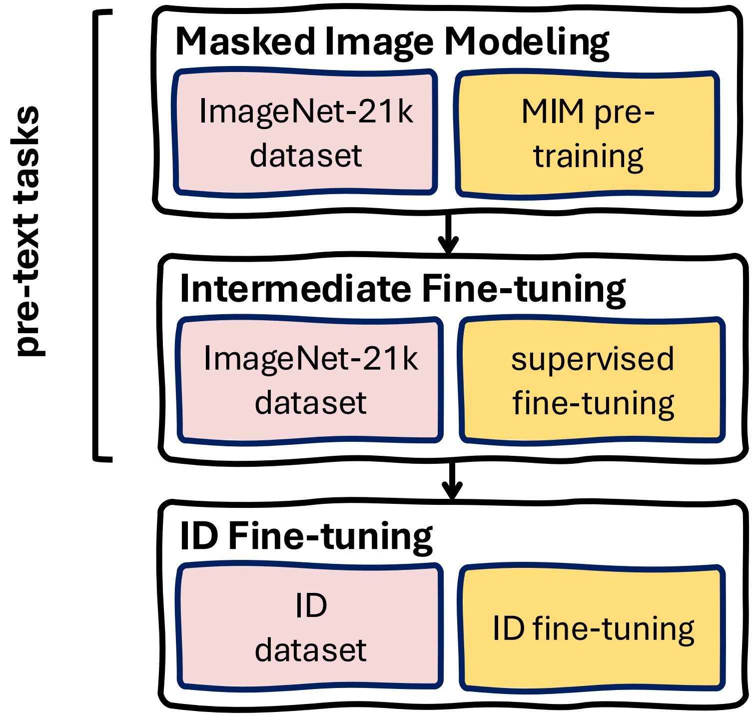


[a] Li et al. Rethinking Out-of-distribution (OOD) Detection: Masking Image Modeling is All You Need. In CVPR, 2023.

# Representation: Reconstruction Learning | MOOD

**MOOD [a]** forces the models to **encode finer-grained information needed to rebuild input**, which is far beyond semantics as in contrastive learning.

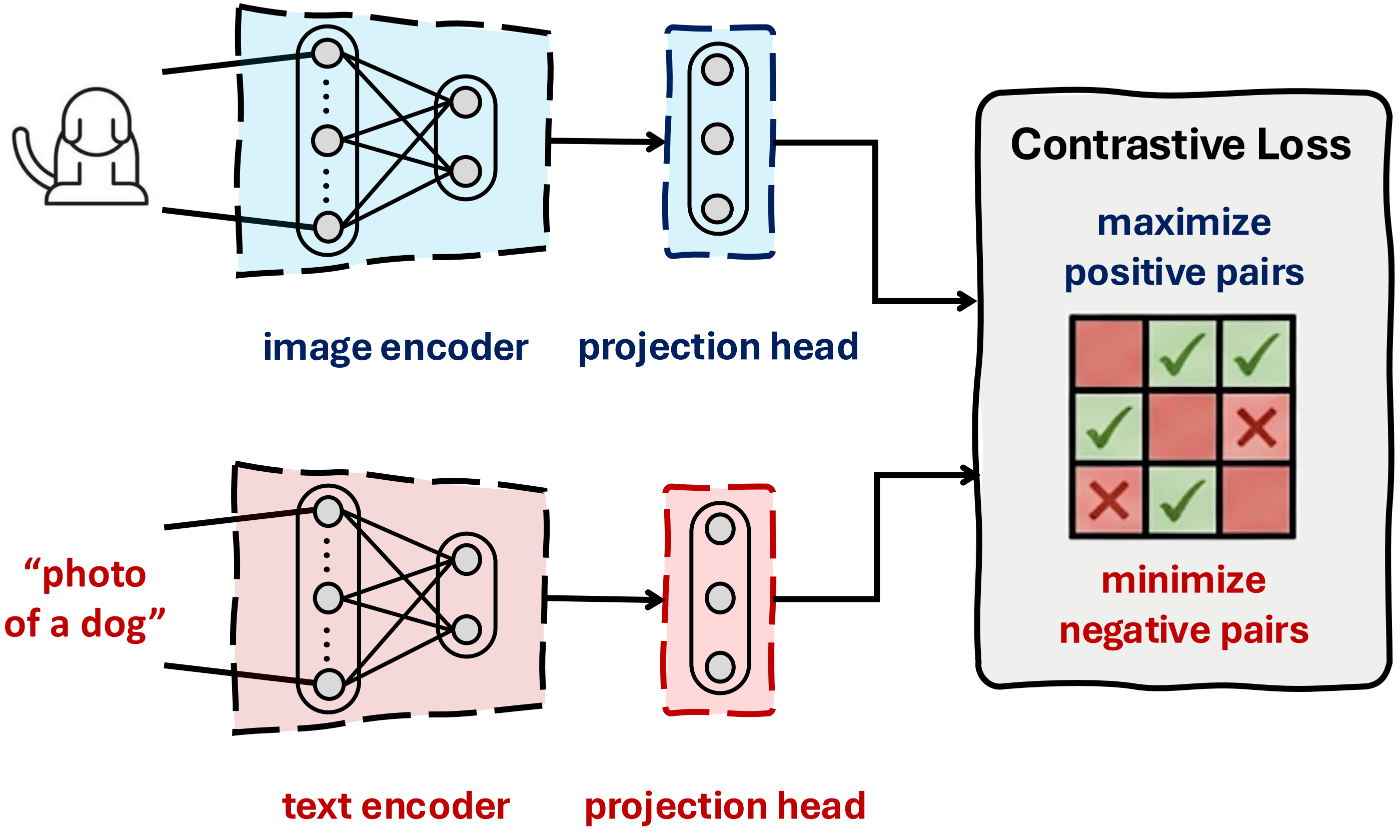



[a] Li et al. Rethinking Out-of-distribution (OOD) Detection: Masking Image Modeling is All You Need. In CVPR, 2023.

# Representation: Pre-training | CLIP

**Cross-modal alignment** reduces reliance on **the shortcut to learn classification**, encouraging more general, semantically meaningful features.





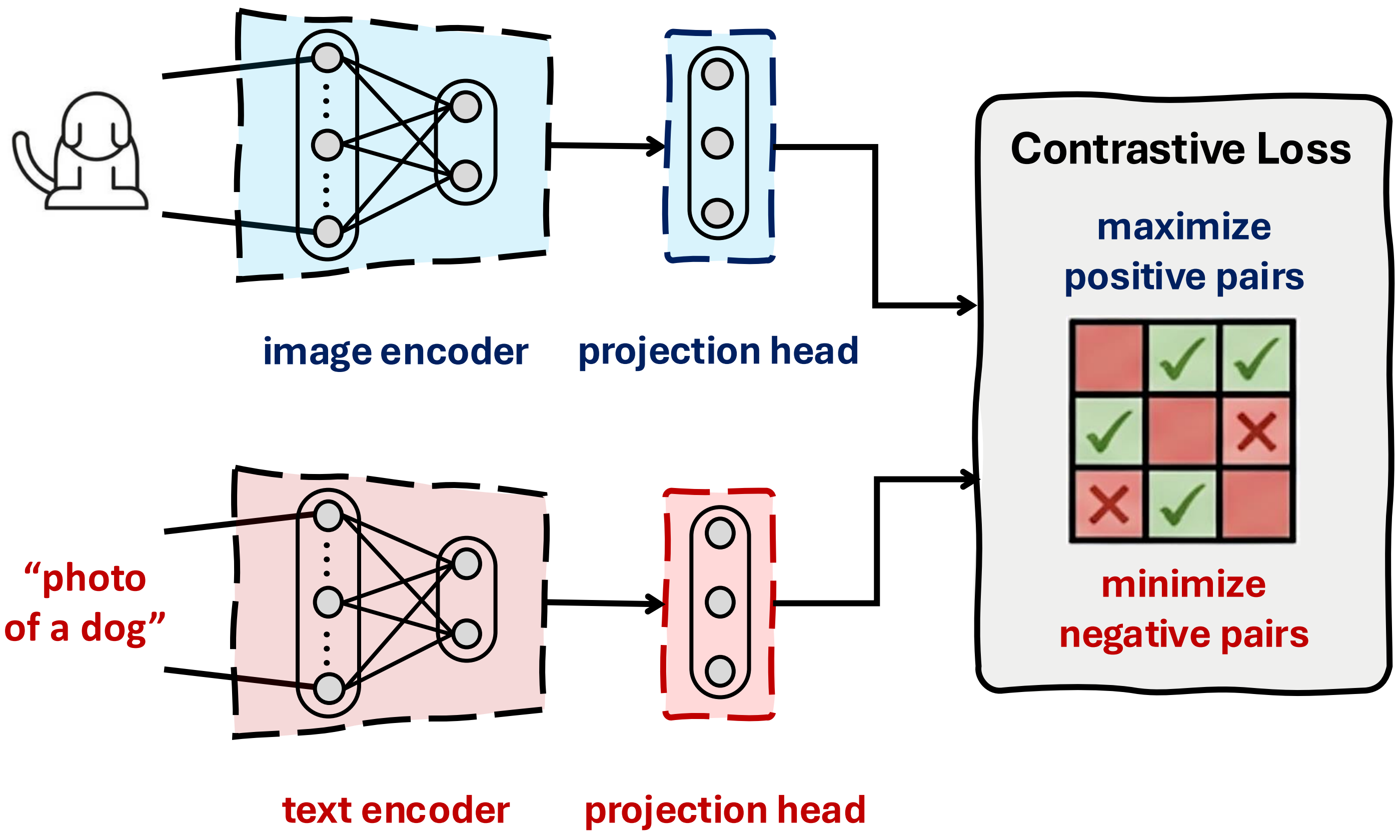


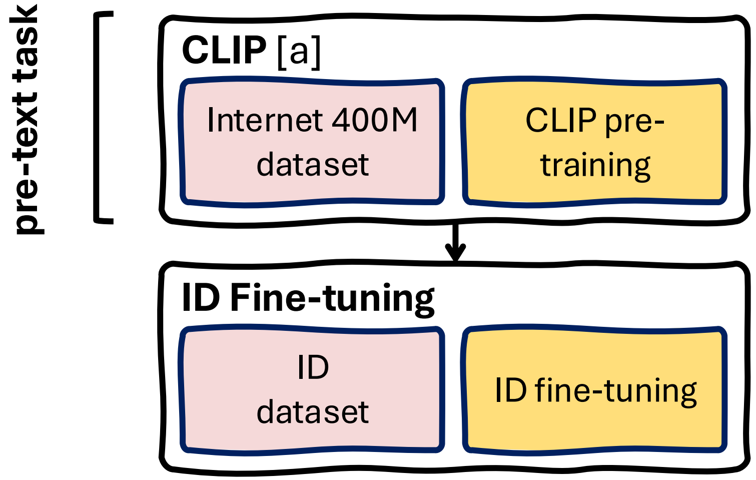
*Aligns **images and text representations** by **maximizing similarity of correct pairs**, meanwhile **minimizing similarity of incorrect pairs**.*

[a] Fort et al. Exploring the Limits of Out-of-Distribution Detection. In NeurIPS, 2021.

# Representation: Pre-training | CLIP

**Cross-modal alignment** reduces reliance on **the shortcut to learn classification**, encouraging more general, semantically meaningful features.



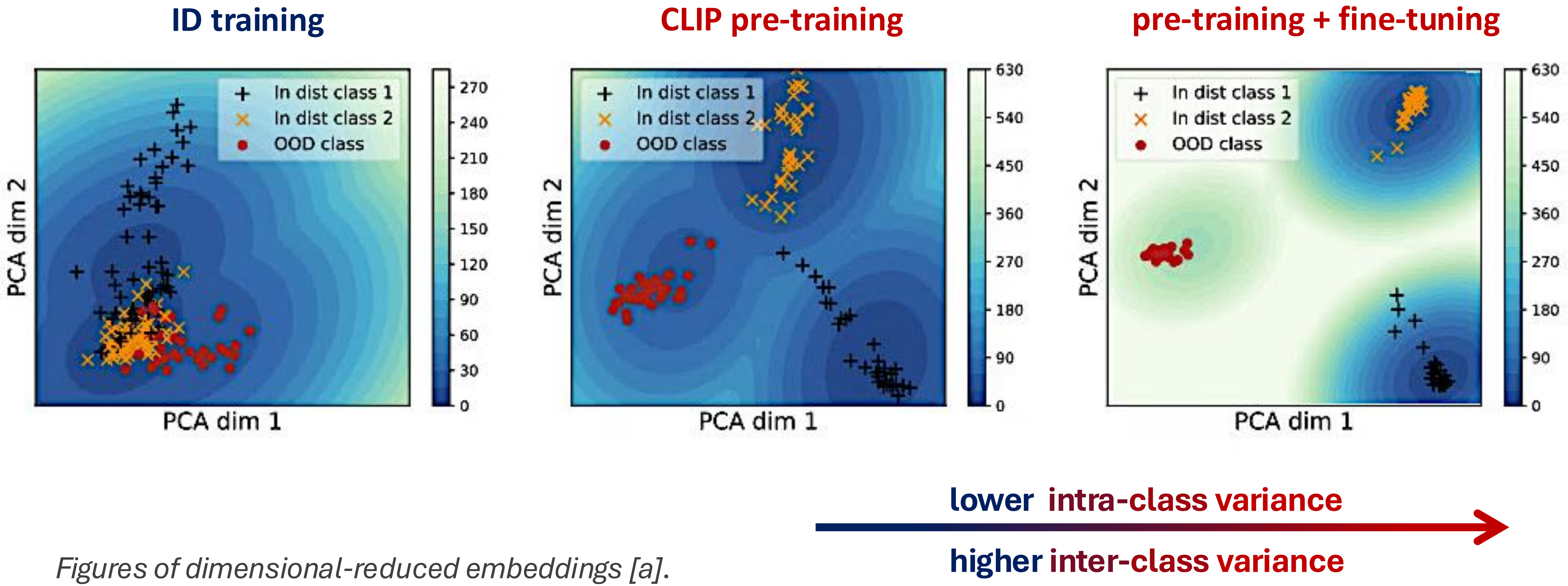




*Aligns **images and text representations** by **maximizing similarity of correct pairs**, meanwhile **minimizing similarity of incorrect pairs**.*

[a] Fort et al. Exploring the Limits of Out-of-Distribution Detection. In NeurIPS, 2021.

# Representation: Pre-training | CLIP

**Cross-modal alignment** reduces reliance on **the shortcut to learn classification**, encouraging more general, semantically meaningful features.




*Figures of dimensional-reduced embeddings [a].*

[a] Fort et al. Exploring the Limits of Out-of-Distribution Detection. In NeurIPS, 2021.

# Calibration: Overview

Let us review what we have learned from textbook [a].

## ❖ Maximum Likelihood

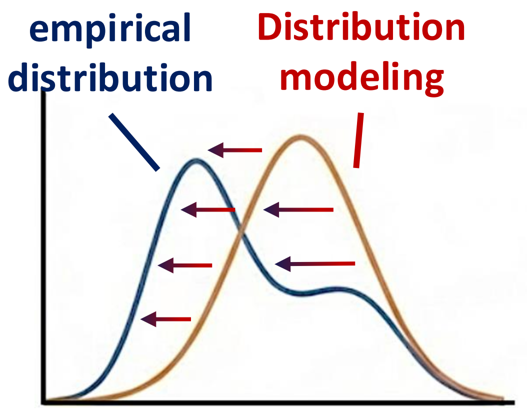
MLE minimizes KL between the **distribution model** and **empirical data distribution.**

**optimal solution**

$$p^* = \underbrace{\underset{p\in\mathcal{P}}{\operatorname{argmax}} \sum_i \log p(x_i)} = \underbrace{\underset{p\in\mathcal{P}}{\operatorname{argmin}} \mathrm{KL}\left(\widehat{\mathcal{D}}||p\right)}$$

**Condition 1.** proper distribution family

**Condition 2.** enough data



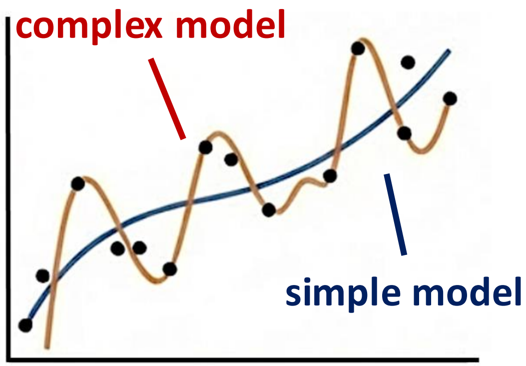

## ❖ Model Complexity

**Higher complexity increases estimation errors** for a fixed dataset size.

**estimation error**

$$\left\| \underset{x\sim\mathcal{D}}{\mathbb{E}}[\Phi(x)] - \underset{x\sim\widehat{\mathcal{D}}}{\mathbb{E}}[\Phi(x)] \right\|_\infty \le \underbrace{2\mathfrak{R}_m(\mathcal{H}) + r\sqrt{\frac{\log 2/\delta}{2m}}}$$

**Condition 3.** proper model complexity

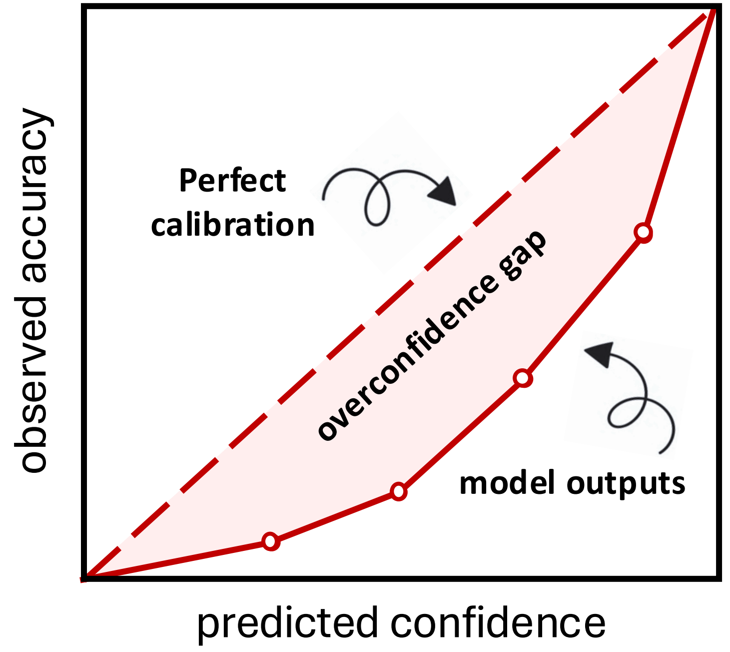



[a] Mohri et al. Foundations of Machine Learning. MIT press, 2018.

# Calibration: Overview

**Conditions 1-3** may not be fully satisfied in practice, leading to calibration failures.

### The overconfidence problem



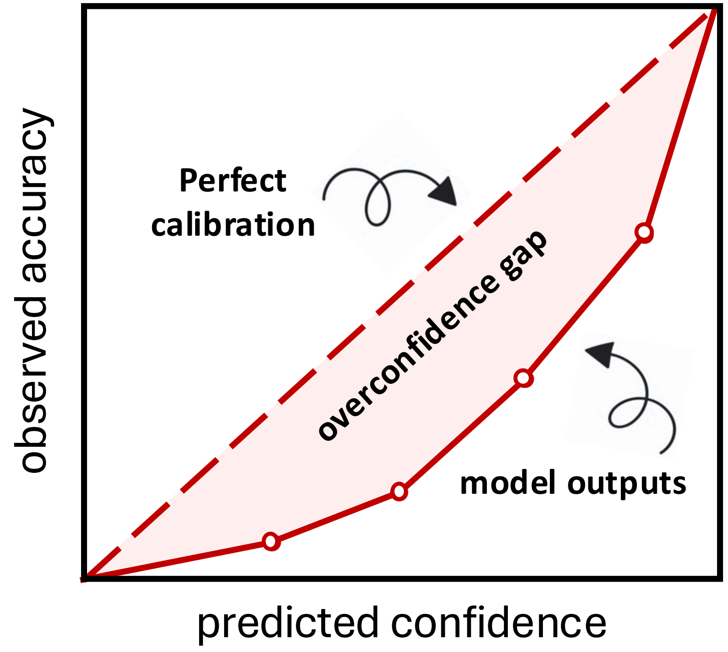

Overconfidence makes the model outputs **appear more reliable than they really are.**

### Consequences

# Calibration: Overview

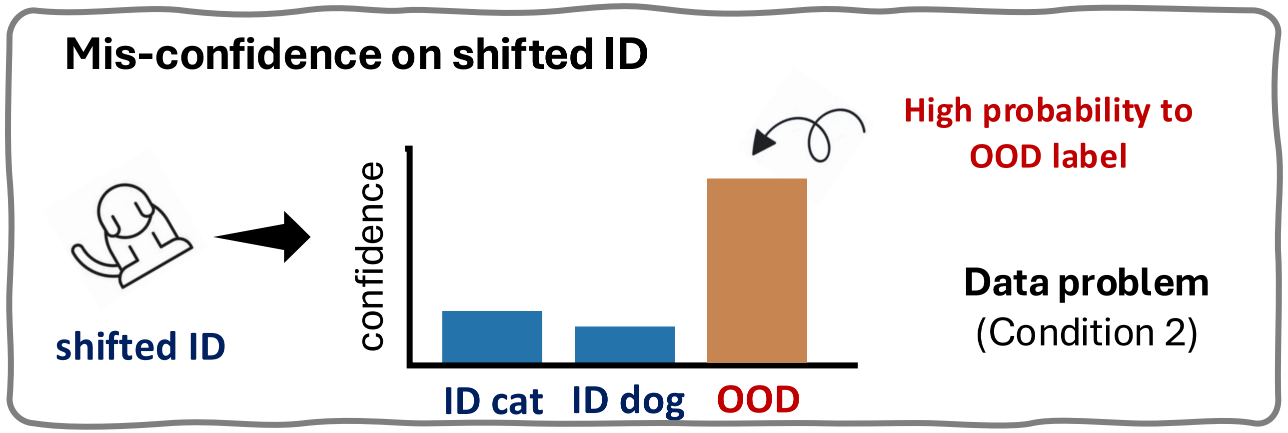
**Conditions 1-3** may not be fully satisfied in practice, leading to calibration failures.

### The overconfidence problem




Overconfidence makes the model outputs **appear more reliable than they really are.**

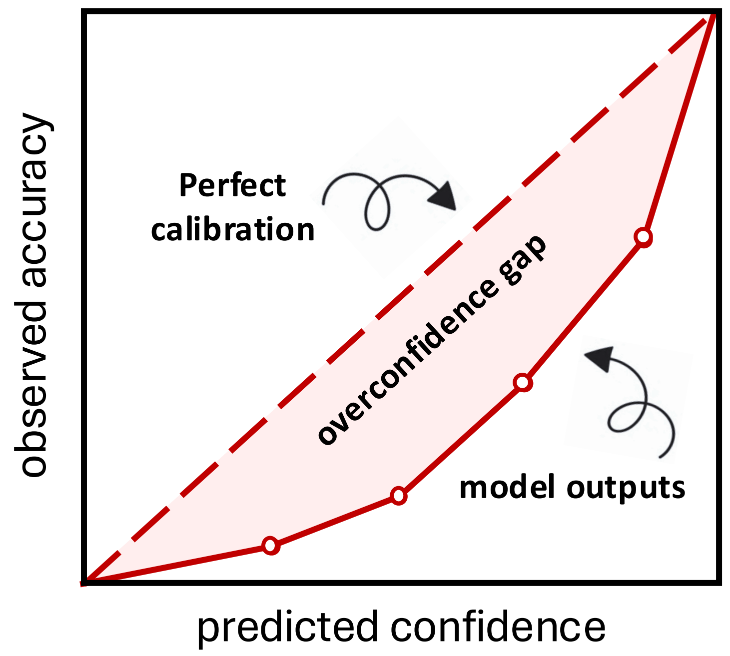
### Consequences



# Calibration: Overview

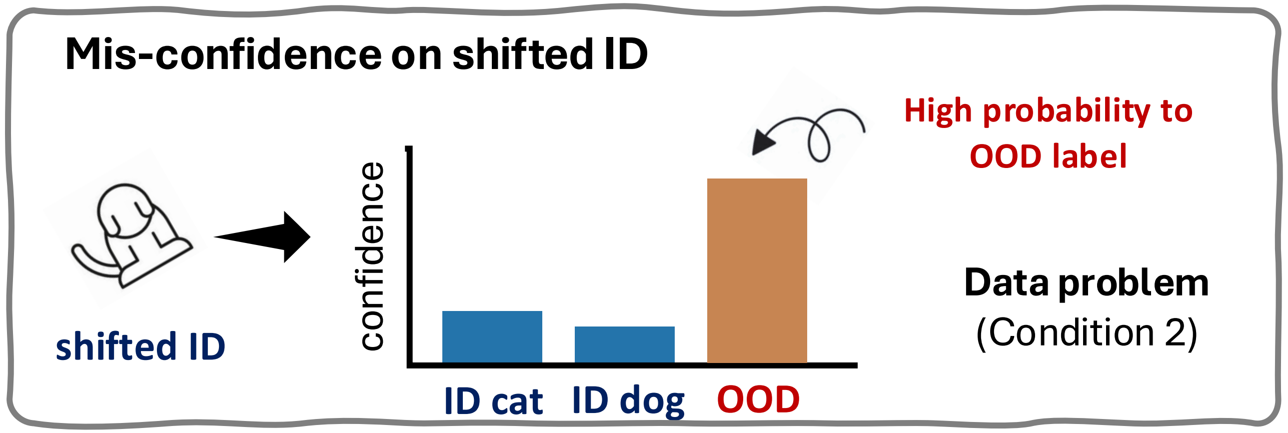
**Conditions 1-3** may not be fully satisfied in practice, leading to calibration failures.

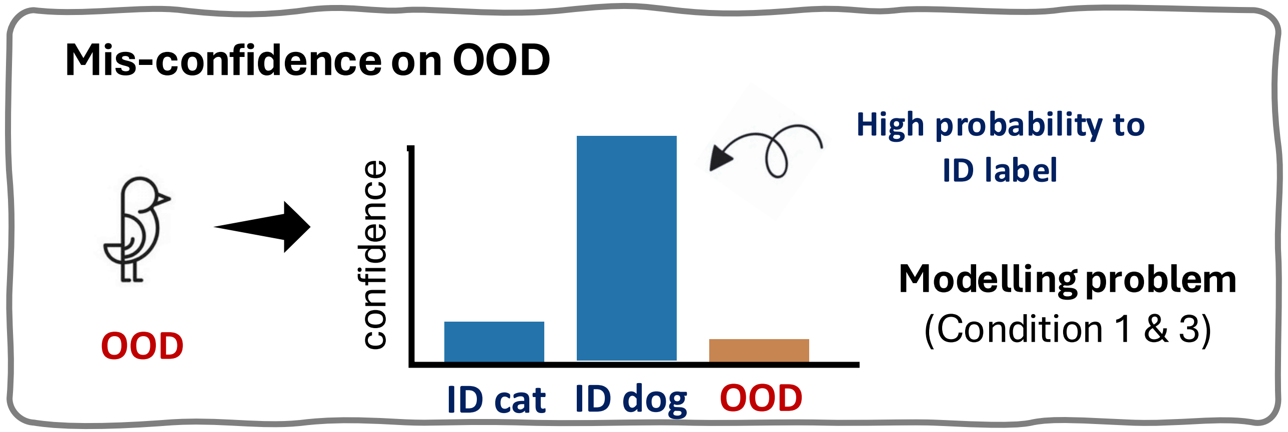
### The mis-confidence problem




Overconfidence makes the model outputs **appear more reliable than they really are.**

### Consequences



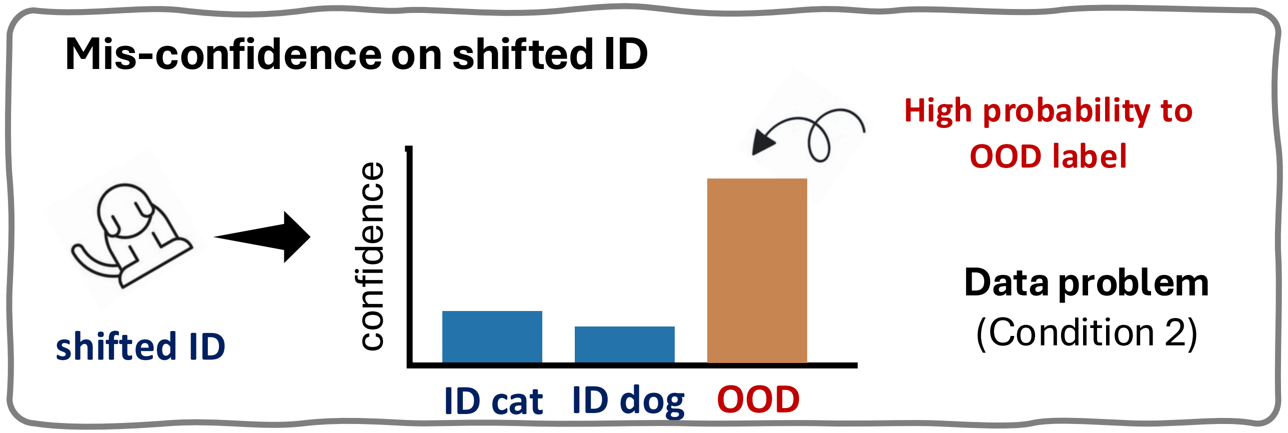



# Calibration: Overview

**Conditions 1-3** may not be fully satisfied in practice, leading to calibration failures.

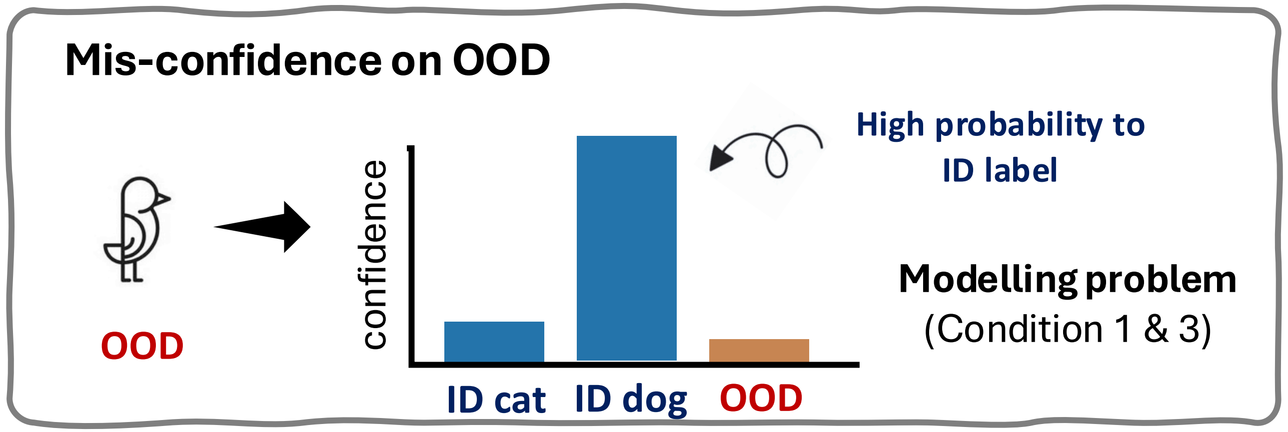
**The overconfidence problem**

## Consequences




- ❖ **Data-centric Solutions.**

  **conventional calibration methods,** such as label smoothing and mixup augmentations.




- ❖ **Model-centric Solutions.**

  various **regularization strategies.**

- ❖ **Distribution-centric Solutions.**

  modelling **beyond softmax.**

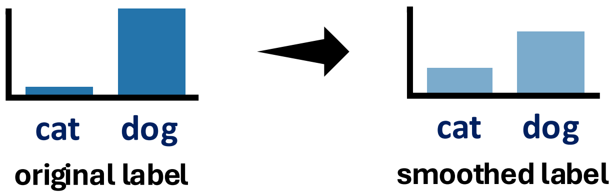
Overconfidence makes the model outputs appear more reliable than they really are.

# Calibration: Data-centric Solutions

**Conventional calibration strategies** have been empirically shown to enhance OOD detection.

### ❖ Change Labels

label smoothing [a]

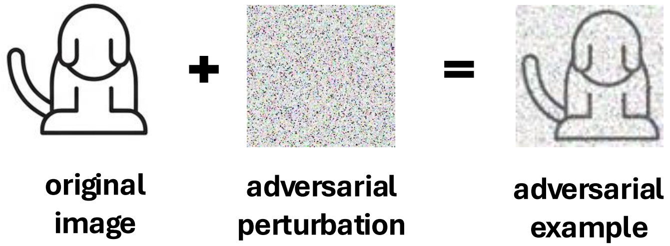



$$\tilde{y} = \overbrace{(1-\alpha)\underbrace{y}_{\text{original label}} + \underbrace{\frac{\alpha}{C}}_{\text{uniform}}}^{\text{interpolation}}$$

replaces one-hot labels with **slightly softened targets.**

### ❖ Change Inputs

adversarial examples [b]

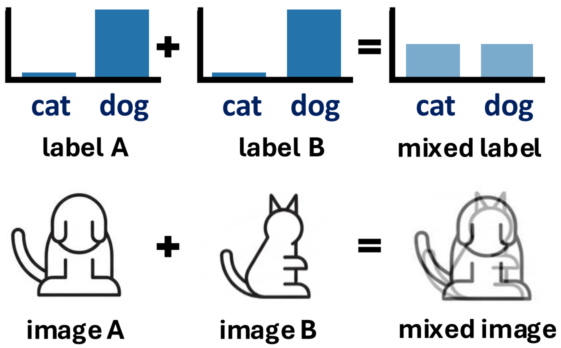



$$\underbrace{\max_{\delta\in\Delta}}_{\text{fail model}} \mathcal{L}(x+\underbrace{\delta}_{\text{perturb input}}, y, \theta)$$

perturbs inputs **imperceptibly** to fool the models and then train the model.

### ❖ Change Labels and Inputs

mixup [c]




$$\tilde{x} = \lambda x_{\mathrm{A}} + (1-\lambda)x_{\mathrm{B}}$$
$$\tilde{y} = \lambda y_{\mathrm{A}} + (1-\lambda)y_{\mathrm{B}}$$

creates virtual training examples that **interpolates** between data pairs.

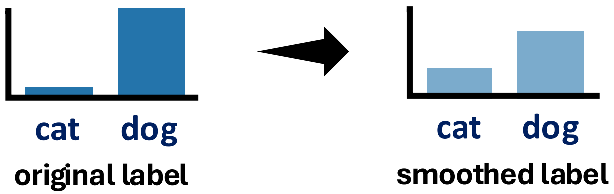

[a] Li et al. Rethinking Out-of-distribution (OOD) Detection: Masked Image Modeling is All You Need. In CVPR, 2023.
[b] Botschen et al. Out-of-Distribution Detection with Adversarial Outlier Exposure. In CVPR, 2025.
[c] Zhang et al. *mixup*: Beyond Empirical Risk Minimization. In ICLR, 2018.

# Calibration: Data-centric Solutions

**Conventional calibration strategies** have been empirically shown to enhance OOD detection.

### ❖ Change Labels

label smoothing [a]




$$\tilde{y} = \overbrace{(1-\alpha)\underbrace{y}_{\text{original label}} + \underbrace{\frac{\alpha}{c}}_{\text{uniform}}}^{\text{interpolation}}$$

replaces one-hot labels with **slightly softened targets.**

**discourages extreme 0/1 probabilities** and thus reduces overconfidence and improves calibrations.

### ❖ Change Inputs

adversarial examples [b]




perturb input

perturbs inputs **imperceptibly** to fool the models and then train the model.

### ❖ Change Labels and Inputs

mixup [c]




$$\tilde{x} = \lambda x_A + (1-\lambda)x_B$$
$$\tilde{y} = \lambda y_A + (1-\lambda)y_B$$

creates virtual training examples that **interpolates** between data pairs.


[a] Li et al. Rethinking Out-of-distribution (OOD) Detection: Masked Image Modeling is All You Need. In CVPR, 2023.
[b] Botschen et al. Out-of-Distribution Detection with Adversarial Outlier Exposure. In CVPR, 2025.
[c] Zhang et al. *mixup*: Beyond Empirical Risk Minimization. In ICLR, 2018.

# Calibration: Data-centric Solutions

**Conventional calibration strategies** have been empirically shown to enhance OOD detection.

## ❖ Change Labels

label smoothing [a]



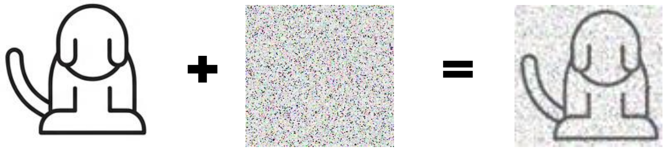

interpolation

$$\tilde{y} = \underbrace{(1-\alpha)y}_{\text{original label}} + \underbrace{\frac{\alpha}{C}}_{\text{uniform}}$$

replaces one-hot labels with **slightly softened targets.**

## ❖ Change Inputs

adversarial examples [b]




**original image** **adversarial perturbation** **adversarial example**

$$\underbrace{\max_{\delta \in \Delta}}_{\text{fail model}} \mathcal{L}(x + \delta, y, \theta)$$

**fail model** **perturb input**

perturbs inputs **imperceptibly** to fool the models and then train the model.

encourages the model to **spread probability mass more cautiously** around decision boundaries.

## ❖ Change Labels and Inputs

mixup [c]




creates virtual training examples that **interpolates** between data pairs.


[a] Li et al. Rethinking Out-of-distribution (OOD) Detection: Masked Image Modeling is All You Need. In CVPR, 2023.
[b] Botschen et al. Out-of-Distribution Detection with Adversarial Outlier Exposure. In CVPR, 2025.
[c] Zhang et al. *mixup*: Beyond Empirical Risk Minimization. In ICLR, 2018.

# Calibration: Data-centric Solutions

**Conventional calibration strategies** have been empirically shown to enhance OOD detection.

### ❖ Change Labels

label smoothing [a]




interpolation

$$\tilde{y} = (1-\alpha)y + \frac{\alpha}{c}$$

original label uniform

replaces one-hot labels with **slightly softened targets.**

### ❖ Change Inputs

adversarial examples [b]

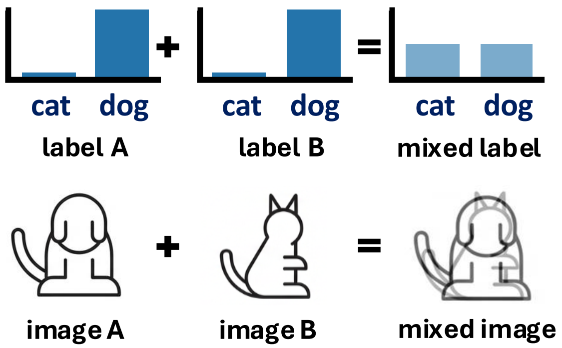



**interpolates empirical distribution** into a continuous one via a convex combinations of data and labels.

perturbs inputs **imperceptibly** to fool the models and then train the model.

### ❖ Change Labels and Inputs

mixup [c]




$$\tilde{x} = \lambda x_{\mathrm{A}} + (1-\lambda)x_{\mathrm{B}}$$
$$\tilde{y} = \lambda y_{\mathrm{A}} + (1-\lambda)y_{\mathrm{B}}$$

creates virtual training examples that **interpolates** between data pairs.


[a] Li et al. Rethinking Out-of-distribution (OOD) Detection: Masked Image Modeling is All You Need. In CVPR, 2023.
[b] Botschen et al. Out-of-Distribution Detection with Adversarial Outlier Exposure. In CVPR, 2025.
[c] Zhang et al. *mixup*: Beyond Empirical Risk Minimization. In ICLR, 2018.

# Calibration: Model-centric Solutions

**Comparison between MSP and Free Energy**

softmax prediction

$$P(Y=y|X) = \frac{\exp\{h_y(X)\}}{\sum_{y'} \exp\{h_{y'}(X)\}}$$

$\propto P(Y,X)$

$\propto P(X)$, free energy

Free energy, which models $P(X)$, is more reliable for OOD detection than maximum softmax prediction. **Why?**

---

**A Bayesian View [a].**

Considering the following two learning goals, **which one is more suitable for OOD detection?**

$$\underbrace{\mathbb{P}(y \neq \boxed{f(x)}, r(x)=0)}_{\text{Risk of accepting data that are misclassified}} + \underbrace{c_0 \mathbb{P}(\boxed{r(x)} = 1)}_{\text{Risk of rejecting data with cost } c_0}$$

classifier

rejector

$$\underbrace{\alpha \boxed{\mathbb{P}_{\text{in}}}(r(x)=1)}_{\text{Risk of rejecting ID}} + \underbrace{\beta \boxed{\mathbb{P}_{\text{ood}}}(r(x)=0)}_{\text{Risk of accepting OOD}}$$

ID distribution

OOD distribution

[a] Narasimhan et al. Learning to Reject Meets OOD Detection: Are All Abstentions Created Equal? Arxiv, 2023.

# Calibration: Model-centric Solutions

## Comparison between MSP and Free Energy

softmax prediction $\quad P(Y=y|X) = \dfrac{\exp\{h_y(X)\}}{\sum_{y'} \exp\{h_{y'}(X)\}}$

$\exp\{h_y(X)\} \propto P(Y,X)$

$\sum_{y'} \exp\{h_{y'}(X)\} \propto P(X)$, free energy

Free energy, which models $P(X)$, is more reliable for OOD detection than maximum softmax prediction. **Why?**

---

## A Bayesian View [a].

Considering the following two learning objectives, **which one is more suitable for OOD detection?**

$$\underbrace{\mathbb{P}(y \neq \boxed{f(x)}, r(x) = 0)}_{\text{Risk of accepting data that are misclassified}} + \underbrace{c_0 \mathbb{P}(\boxed{r(x)} = 1)}_{\text{Risk of rejecting data with cost } c_0}$$

**classifier** **rejector**

$$\underbrace{\boxed{\alpha \mathbb{P}_{\text{in}}}(r(x) = 1)}_{\text{Risk of rejecting ID}} + \underbrace{\boxed{\beta \mathbb{P}_{\text{ood}}}(r(x) = 0)}_{\text{Risk of accepting OOD}}$$

**ID distribution** **OOD distribution**

[a] Narasimhan et al. Learning to Reject Meets OOD Detection: Are All Abstentions Created Equal? Arxiv, 2023.

# Calibration: Model-centric Solutions

**Comparison between MSP and Free Energy**

$$P(Y = y|X) = \frac{\exp\{h_y(X)\}}{\sum_{y'} \exp\{h_{y'}(X)\}}$$

softmax prediction

$\propto P(Y, X)$

$\propto P(X)$, free energy

Free energy, which models $P(X)$, is more reliable for OOD detection than maximum softmax prediction. **Why?**

---

## A Bayesian View [a].

Considering the following two learning objectives, **which one is more suitable for OOD detection?**

$$\mathbb{P}(y \neq \boxed{f(x)}, r(x) = 0) + c_0 \mathbb{P}(\boxed{r(x)} = 1)$$

**classifier** — Risk of accepting data that are misclassified

**rejector** — Risk of rejecting data with cost $c_0$

$$\alpha \mathbb{P}_{\text{in}}(r(x) = 1) + \beta \mathbb{P}_{\text{ood}}(r(x) = 0)$$

**ID distribution** — Risk of rejecting ID

**OOD distribution** — Risk of accepting OOD

[a] Narasimhan et al. Learning to Reject Meets OOD Detection: Are All Abstentions Created Equal? Arxiv, 2023.

# Calibration: Model-centric Solutions

**Comparison between MSP and Free Energy**

$$\underbrace{P(Y=y|X)}_{\text{softmax prediction}} = \frac{\overbrace{\exp\{h_y(X)\}}^{\propto P(Y,X)}}{\underbrace{\sum_{y'}\exp\{h_{y'}(X)\}}_{\propto P(X)\text{, free energy}}}$$

Free energy, which models $P(X)$, is more reliable for OOD detection than maximum softmax prediction. **Why?**

## A Bayesian View [a].

Considering the following two learning objectives, **which one is more suitable for OOD detection?**

$$\underbrace{\mathbb{P}(y \neq \boxed{f(x)}, r(x) = 0)}_{\text{Risk of accepting data that are misclassified}} + \underbrace{c_0\mathbb{P}(\boxed{r(x)} = 1)}_{\text{Risk of rejecting data with cost } c_0}$$

classifier

rejector

$$\underbrace{\alpha\boxed{\mathbb{P}_{\text{in}}}(r(x) = 1)}_{\text{Risk of rejecting ID}} + \underbrace{\beta\boxed{\mathbb{P}_{\text{ood}}}(r(x) = 0)}_{\text{Risk of accepting OOD}}$$

**ID distribution**

**OOD distribution**

[a] Narasimhan et al. Learning to Reject Meets OOD Detection: Are All Abstentions Created Equal? Arxiv, 2023.

# Calibration: Model-centric Solutions

**Comparison between MSP and Free Energy**

$$P(Y = y|X) = \frac{\exp\{h_y(X)\}}{\sum_{y'} \exp\{h_{y'}(X)\}}$$

softmax prediction

$\propto P(Y, X)$

$\propto P(X)$, free energy

Free energy, which models $P(X)$, is more reliable for OOD detection than maximum softmax prediction. **Why?**

---

## A Bayesian View [a].

Considering the following two learning objectives, **which one is more suitable for OOD detection?**

$$\mathbb{P}(y \neq f(x), r(x) = 0) + c_0 \mathbb{P}(r(x) = 1)$$

used for **abstention-aware classification.**

MSP

$$r^*(x) = [\![ \max_y \mathbb{P}(y|x) < 1 - c_0 ]\!]$$

$$\alpha \mathbb{P}_{\text{in}}(r(x) = 0) + \beta \mathbb{P}_{\text{ood}}(r(x) = 1)$$

used for **OOD detection.**

Free Energy

$$r^*(x) = \left[\!\!\left[ \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{ood}}(x)} < \frac{\beta}{\alpha} \right]\!\!\right] \approx \left[\!\!\left[ \frac{\mathbb{P}_{\text{in}}(x)}{\tau} < \frac{\beta}{\alpha} \right]\!\!\right]$$

[a] Narasimhan et al. Learning to Reject Meets OOD Detection: Are All Abstentions Created Equal? Arxiv, 2023.

# Calibration: Model-centric Solutions

**Comparison between MSP and Free Energy**

$$P(Y=y|X) = \frac{\exp\{h_y(X)\}}{\sum_{y'} \exp\{h_{y'}(X)\}}$$

softmax prediction

$\exp\{h_y(X)\}$: $\propto P(Y,X)$

$\sum_{y'} \exp\{h_{y'}(X)\}$: $\propto P(X)$, free energy

Free energy, which models $P(X)$, is more reliable for OOD detection than maximum softmax prediction. **Why?**

---

**A Bayesian View [a].**

Considering the following two learning objectives, **which one is more suitable for OOD detection?**

😭 $\mathbb{P}(y \neq f(x), r(x)=0) + c_0 \mathbb{P}(r(x)=1)$

used for **abstention-aware classification.**

$r^*(x) = [\![ \underbrace{\max_y \mathbb{P}(y|x)}_{\text{MSP}} < 1 - c_0 ]\!]$

😀 $\alpha \mathbb{P}_{\text{in}}(r(x)=0) + \beta \mathbb{P}_{\text{ood}}(r(x)=1)$

used for **OOD detection.**

$r^*(x) = \left[\!\left[ \frac{\mathbb{P}_{\text{in}}(x)}{\mathbb{P}_{\text{ood}}(x)} < \frac{\beta}{\alpha} \right]\!\right] \approx \left[\!\left[ \underbrace{\frac{\mathbb{P}_{\text{in}}(x)}{\tau}}_{\text{Free Energy}} < \frac{\beta}{\alpha} \right]\!\right]$

[a] Narasimhan et al. Learning to Reject Meets OOD Detection: Are All Abstentions Created Equal? Arxiv, 2023.

# Calibration: Model-centric Solutions

**Comparison between MSP and Free Energy**

softmax prediction

$$P(Y=y|X)=\frac{\overbrace{\exp\{h_y(X)\}}^{\propto P(Y,X)}}{\underbrace{\sum_{y'}\exp\{h_{y'}(X)\}}_{\propto P(X)\text{, free energy}}}$$

Free energy, which models $P(X)$, is more reliable for OOD detection than maximum softmax prediction. However, $P(X)$ is also not well-calibrated. So, **how to calibrate** $P(X)$**?**

---

**Calibrating** $P(Y)$ **(low dimension)** offers a sufficient condition of **calibrating** $P(X)$ **(high dimension).**

**Mathematical**

$P(Y=y)=\frac{1}{Z}\int\exp\{h_y(x)\}dvx,$

where $Z=\sum_y\int\exp\{h_y(x)\}dvx$

**Monte Carlo**



**Empirical**

$\hat{P}(Y=y)=n_y/N$

**Count**



**Hypothesis Test**

$H_0: P(Y=y)=\hat{P}(Y=y)$

versus

$H_1: P(Y=y)\neq\hat{P}(Y=y)$

[a] Huang et al. Density-driven Regularization for Out-of-distribution Detection. In NeurIPS, 2022.

# Calibration: Model-centric Solutions | DCR [a]

**Comparison between MSP and Free Energy**

$$\text{softmax prediction} \quad P(Y=y|X) = \frac{\overbrace{\exp\{h_y(X)\}}^{\propto P(Y,X)}}{\underbrace{\sum_{y'} \exp\{h_{y'}(X)\}}_{\propto P(X)\text{, free energy}}}$$

Free energy, which models $P(X)$, is more reliable for OOD detection than maximum softmax prediction. However, $P(X)$ is also not well-calibrated. So, **how to calibrate $P(X)$?**

---

**Calibrating $P(Y)$ (low dimension)** offers a sufficient condition of **calibrating $P(X)$ (high dimension).**

**Mathematical**

$P(Y=y) = \frac{1}{Z}\int \exp\{h_y(x)\}dvx,$

where $Z = \sum_y \int \exp\{h_y(x)\}dvx$

Monte Carlo



**Hypothesis Test**

$H_0: P(Y=y) = \hat{P}(Y=y)$

versus

$H_1: P(Y=y) \neq \hat{P}(Y=y)$

**Empirical**

$\hat{P}(Y=y) = n_y/N$

Count



[a] Huang et al. Density-driven Regularization for Out-of-distribution Detection. In NeurIPS, 2022.

# Calibration: Model-centric Solutions | DCR [a]

### Comparison between MSP and Free Energy

softmax prediction

$$P(Y=y|X) = \frac{\exp\{h_y(X)\}}{\sum_{y'} \exp\{h_{y'}(X)\}}$$

$\propto P(Y, X)$

$\propto P(X)$, free energy

Free energy, which models $P(X)$, is more reliable for OOD detection than maximum softmax prediction. However, $P(X)$ is also not well-calibrated. So, **how to calibrate $P(X)$?**

---

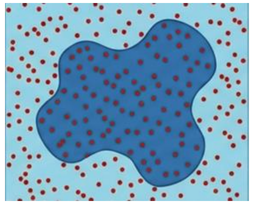
**Calibrating $P(Y)$ (low dimension)** offers a sufficient condition of **calibrating $P(X)$ (high dimension).**

**Mathematical**

$P(Y=y) = \frac{1}{Z}\int \exp\{h_y(x)\}dvx,$

where $Z = \sum_y \int \exp\{h_y(x)\}dvx$

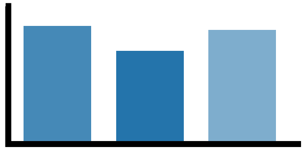
→ **Monte Carlo**



**Empirical**

$\hat{P}(Y=y) = n_y/N$

→ **Count**



**Hypothesis Test**

$H_0: P(Y=y) = \hat{P}(Y=y)$

versus

$H_1: P(Y=y) \neq \hat{P}(Y=y)$

[a] Huang et al. Density-driven Regularization for Out-of-distribution Detection. In NeurIPS, 2022.

# Calibration: Model-centric Solutions | DCR [a]

**Comparison between MSP and Free Energy**

softmax prediction
$$P(Y=y|X) = \frac{\exp\{h_y(X)\}}{\sum_{y'} \exp\{h_{y'}(X)\}}$$
$\propto P(Y,X)$

$\propto P(X)$, free energy

Free energy, which models $P(X)$, is more reliable for OOD detection than maximum softmax prediction. However, $P(X)$ is also not well-calibrated. So, **how to calibrate $P(X)$?**

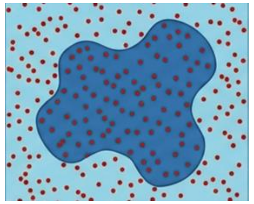
---

**Calibrating $P(Y)$ (low dimension)** offers a sufficient condition of **calibrating $P(X)$ (high dimension).**

**Mathematical**

$$P(Y=y) = \frac{1}{Z}\int \exp\{h_y(x)\}dvx,$$
where $Z = \sum_y \int \exp\{h_y(x)\}dvx$

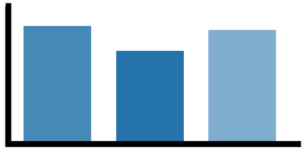
**Monte Carlo**



**Empirical**

$$\hat{P}(Y=y) = n_y/N$$

**Count**



**Hypothesis Test**

$$H_0: P(Y=y) = \hat{P}(Y=y)$$
versus
$$H_1: P(Y=y) \neq \hat{P}(Y=y)$$

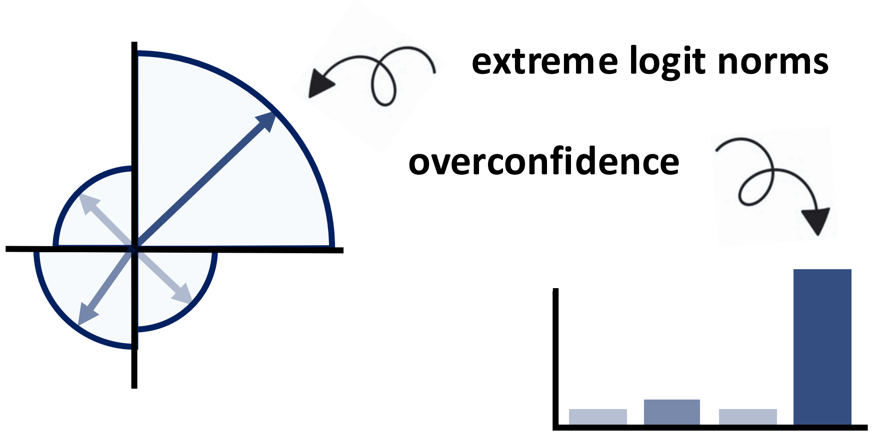
[a] Huang et al. Density-driven Regularization for Out-of-distribution Detection. In NeurIPS, 2022.

# Calibration: Distribution-centric Solutions | LogitNorm

We seek **distribution modelling beyond softmax** that are more proper for OOD detection.

### Why is softmax not sufficient?

- **Scaling logits** increases confidence and decreases the risk.
- **Training pushes near-zero risk**, driving large confidence and inflating logits.




### Logit Norm [a]

- **normalizes logit vector** to a constant norm during training, following

$$-\log \frac{\exp\{f_y/(\tau\|f_y\|)\}}{\sum_i \exp\{f_i/(\tau\|f_i\|)\}}$$

normalization



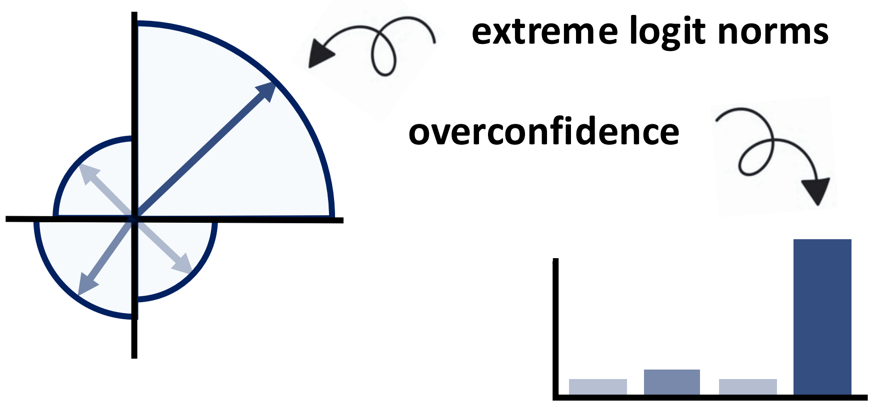

[a] Wei et al. Mitigating Neural Network Overconfidence with Logit Normalization. In ICML, 2022.

# Calibration: Distribution-centric Solutions | LogitNorm

We seek **distribution modelling beyond softmax** that are more proper for OOD detection.

### Why is softmax not sufficient?

- **Scaling logits** increases confidence and decreases the risk.
- **Training pushes near-zero risk**, driving large confidence and inflating logits.

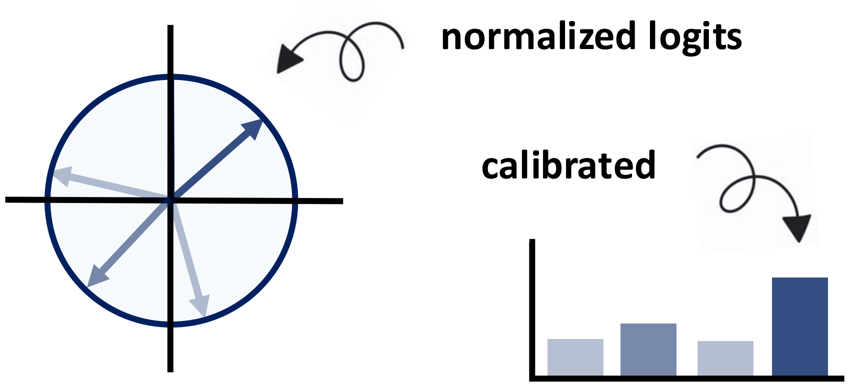



### Logit Norm [a]

- **normalizes logit vector** to a constant norm during training, following

$$-\log \frac{\exp\{f_y/(\tau\|f_y\|)\}}{\sum_i \exp\{f_i/(\tau\|f_i\|)\}}$$

normalization



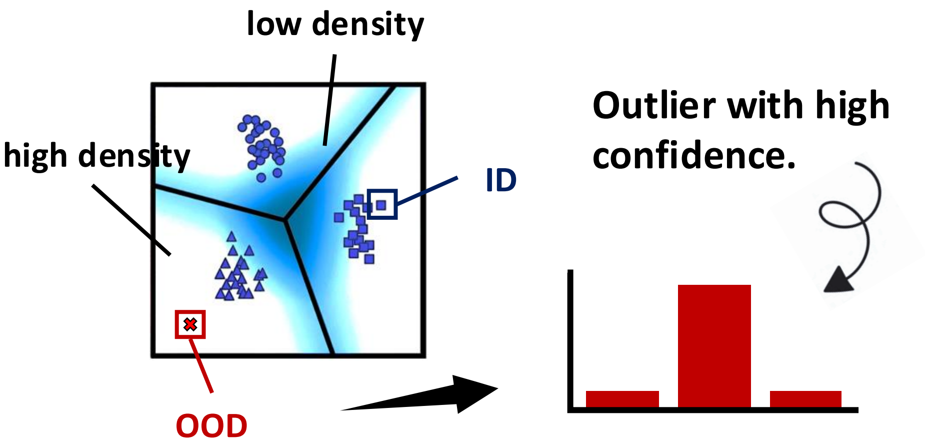

[a] Wei et al. Mitigating Neural Network Overconfidence with Logit Normalization. In ICML, 2022.

# Calibration: Distribution-centric Solutions | SIREN

We seek **distribution modelling beyond softmax** that are more proper for OOD detection.

### Why is softmax not sufficient?

- **The separation is piecewise linear** in feature space.
- Regions that **contain no training points can still be assigned confidently.**




### SIREN [a]

- The density depends on **similarity to prototypes** and vMF likelihood, creating a **curved, cluster-shaped** regions.

$$P(Y = c|r) \propto \exp\{\kappa_c \mu_c^\top r\}$$

concentration prototype



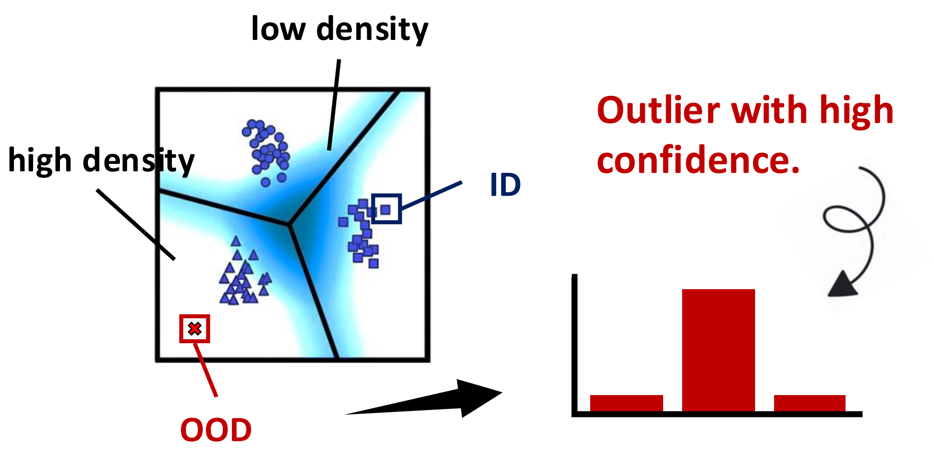

[a] Du et al. SIREN: Shaping Representations for Detecting Out-of-Distribution Objects. In NeurIPS, 2022.

# Calibration: Distribution-centric Solutions | SIREN

We seek **distribution modelling beyond softmax** that are more proper for OOD detection.

### Why is softmax not sufficient?

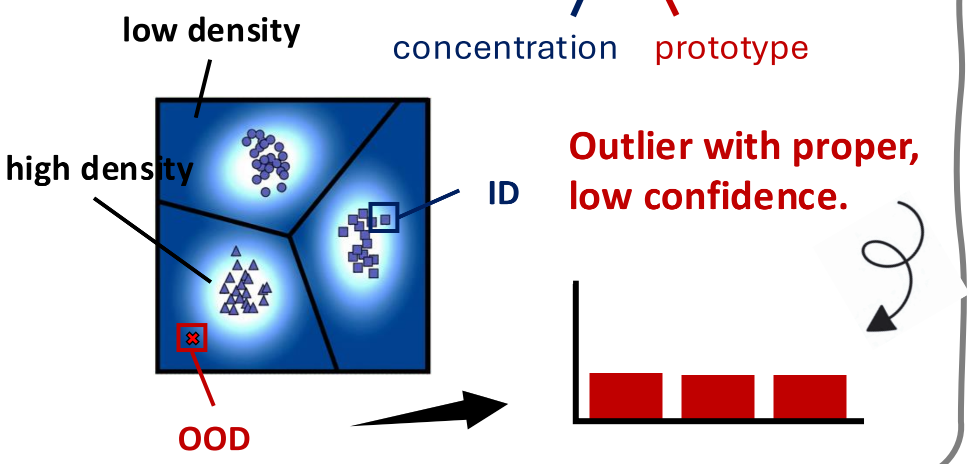
- **The separation is piecewise linear** in feature space.
- Regions that **contain no training points can still be assigned confidently**.




### SIREN [a]

- The density depends on **similarity to prototypes** and vMF likelihood, creating a **curved, cluster-shaped** regions.

$$P(Y = c|r) \propto \exp\{\underbrace{\kappa_c}_{\text{concentration}} \underbrace{\mu_c^\top}_{\text{prototype}} r\}$$

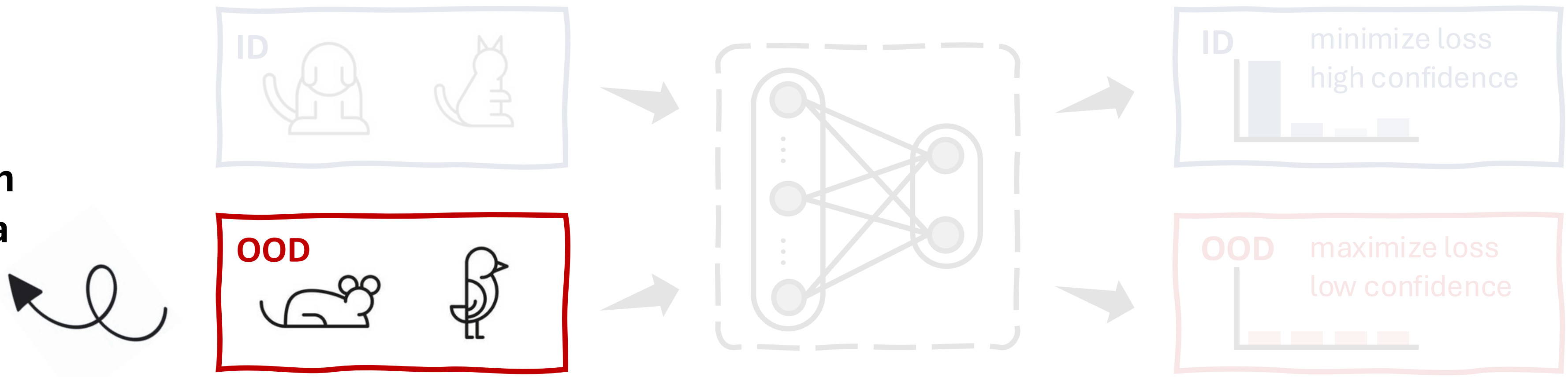


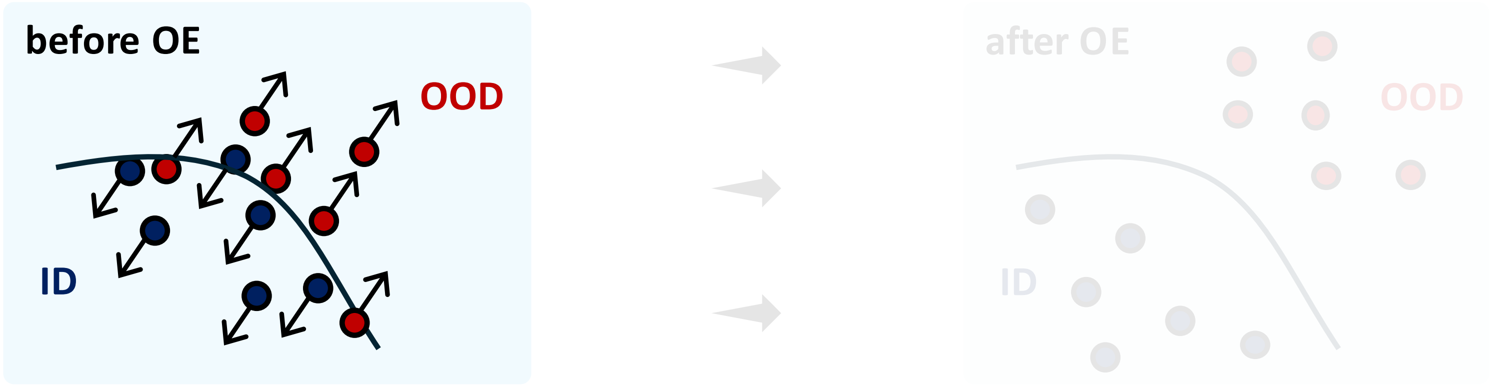

[a] Du et al. SIREN: Shaping Representations for Detecting Out-of-Distribution Objects. In NeurIPS, 2022.

# Outlier Exposure: Overview

Outlier exposure [a] takes OOD detection as **an additional binary classification task**, enabling models to directly learn to distinguish ID from OOD patterns.

make the model **learn from extra OOD data**



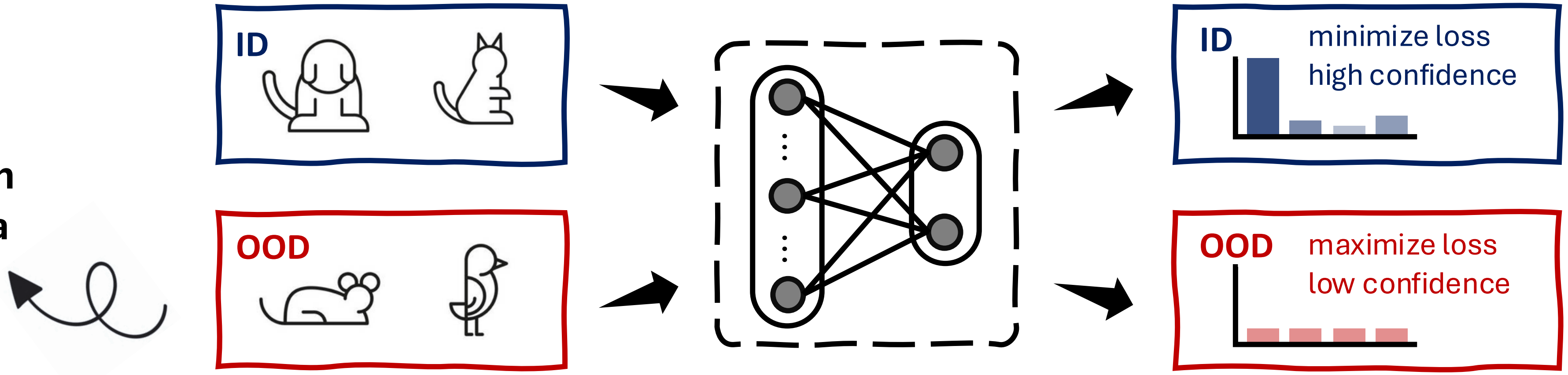

Decision boundary (measured by OOD scoring) **better separate ID and OOD.**

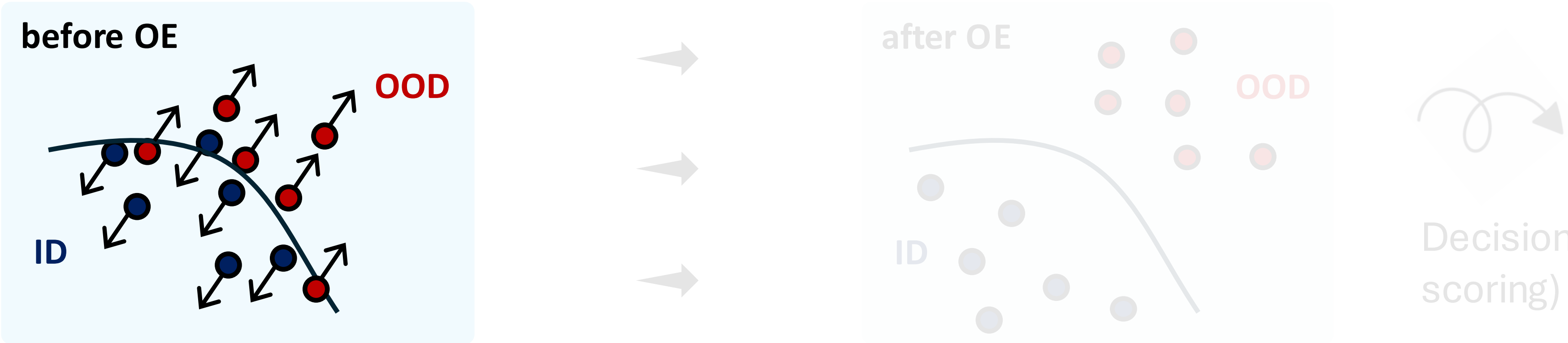
[a] Hendrycks et al. Deep Anomaly Detection with Outlier Exposure. In ICLR, 2019.

# Outlier Exposure: Overview

Outlier exposure [a] takes OOD detection as **an additional binary classification task**, enabling models to directly learn to distinguish ID from OOD patterns.

make the model **learn from extra OOD data**



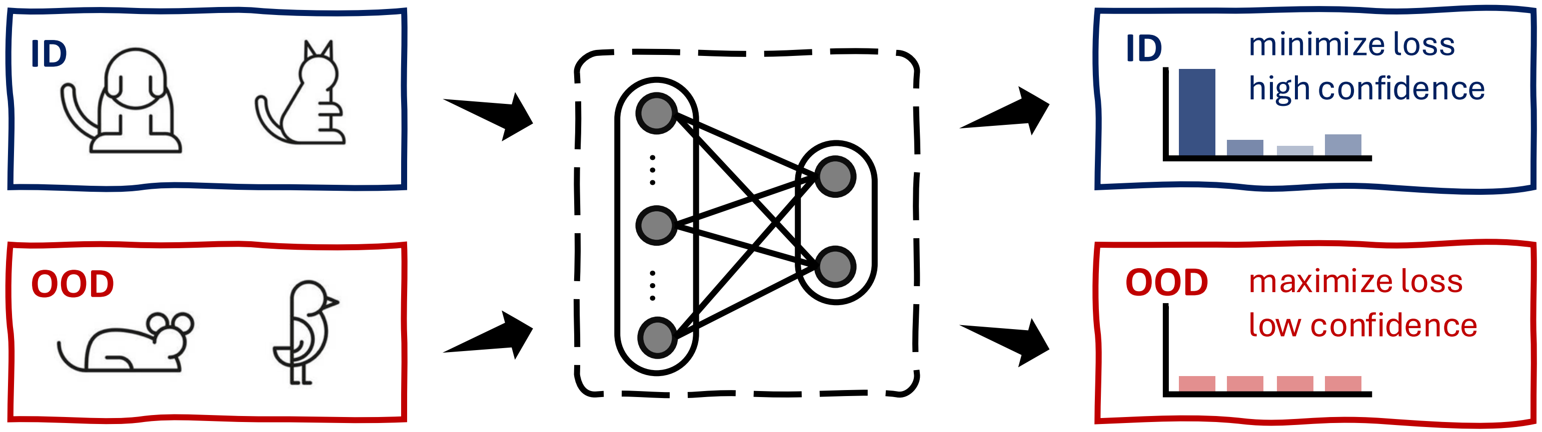

Decision boundary (measured by OOD scoring) **better separate ID and OOD.**

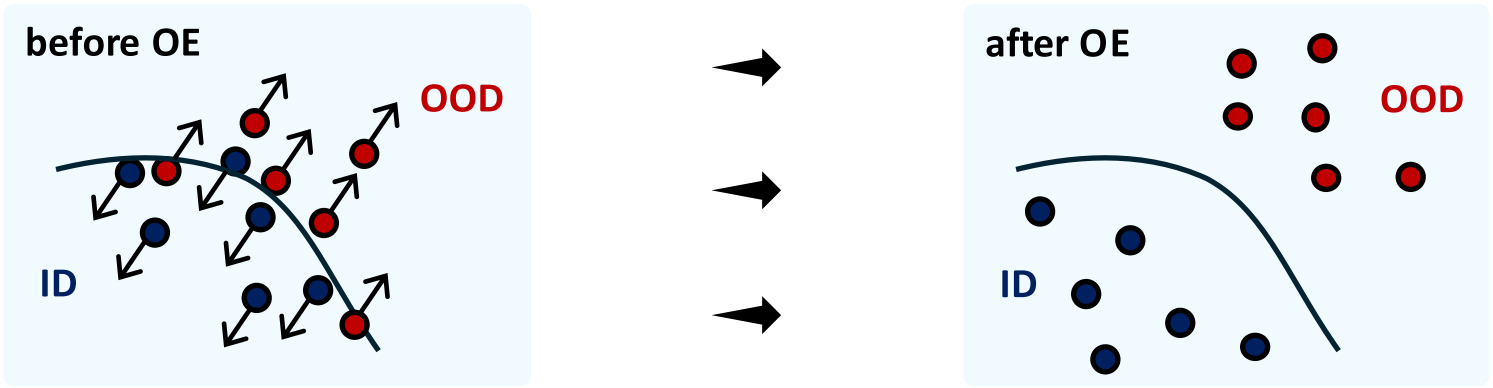
[a] Hendrycks et al. Deep Anomaly Detection with Outlier Exposure. In ICLR, 2019.

# Outlier Exposure: Overview

Outlier exposure [a] takes OOD detection as **an additional binary classification task**, enabling models to directly learn to distinguish ID from OOD patterns.

make the model **learn from extra OOD data**




Decision boundary (measured by OOD scoring) **better separate ID and OOD.**

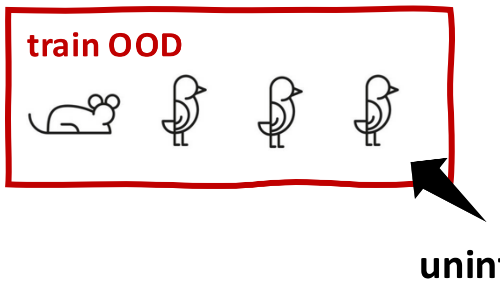
[a] Hendrycks et al. Deep Anomaly Detection with Outlier Exposure. In ICLR, 2019.

# Outlier Exposure: Challenges

Training and test OOD can differ, and **such distribution gap degrades OOD performance** [a].

❖ **Case 1.**

**Redundant Data**




Training data include test data,
but many samples are uninformative.

How to address Case 1?

**Data sampling / reweighting** to remove those uninformative samples.

❖ **Case 2.**

**Scarce Data**




Training data contain no test data.

How to address Case 2?

**Data augmentation** to enlarge the OOD coverage during training.

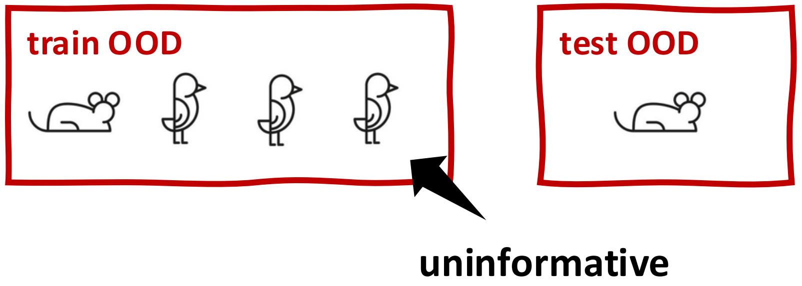
[a] Wang et al. Learning to Augment Distributions for Out-of-distribution Detection. In NeurIPS, 2023.

# Outlier Exposure: Challenges

Training and test OOD can differ, and **such distribution gap degrades OOD performance** [a].

❖ **Case 1.**

**Redundant Data**



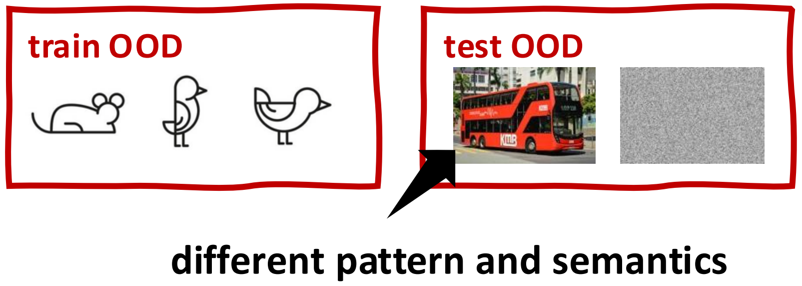

Training data include test data,
**but many samples are uninformative.**

How to address Case 1?

**Data sampling / reweighting** to remove those uninformative samples.

❖ **Case 2.**

**Scarce Data**




**Training data contain no test data.**

How to address Case 2?

**Data augmentation** to enlarge the OOD coverage during training.

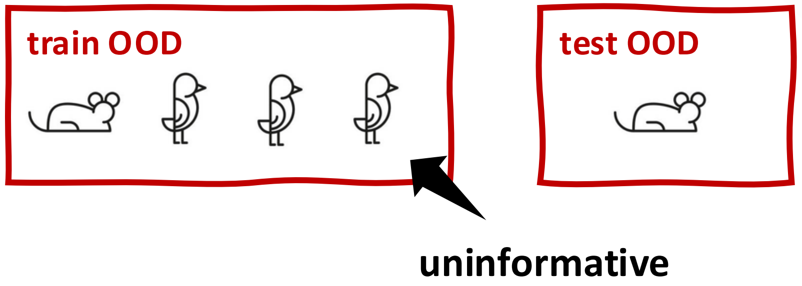
[a] Wang et al. Learning to Augment Distributions for Out-of-distribution Detection. In NeurIPS, 2023.

# Outlier Exposure: Improvements

**Data sampling learns from fewer examples** and **data augmentation from more**.

❖ **Case 1.**

**Redundant Data**



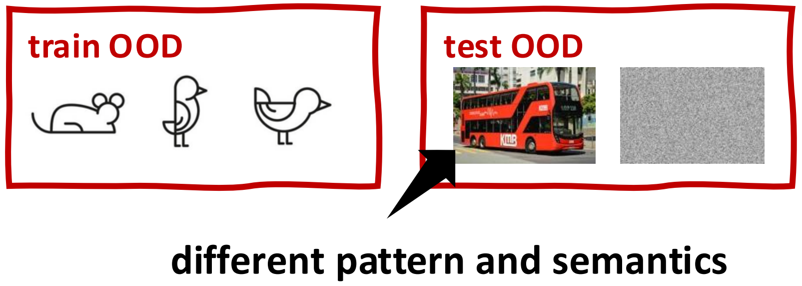

Training data include test data,
but many samples are uninformative.

**How to address Case 1?**

**Data sampling / reweighting** to remove those uninformative samples.

❖ **Case 2.**

**Scarce Data**




Training data contain no test data.

**How to address Case 2?**

**Data augmentation** to enlarge the OOD coverage during training.

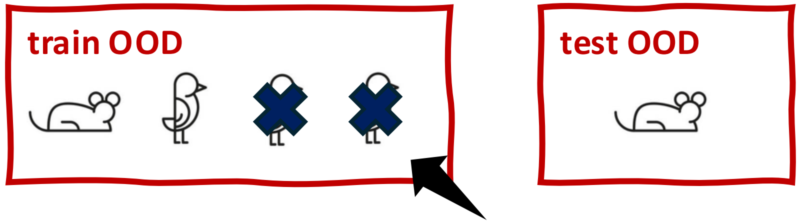
[a] Wang et al. Learning to Augment Distributions for Out-of-distribution Detection. In NeurIPS, 2023.

# Outlier Exposure: Improvements

**Data sampling learns from fewer examples** and **data augmentation from more**.

❖ **Case 1.**

**Redundant Data → Data Sampling**




**learn from less data**

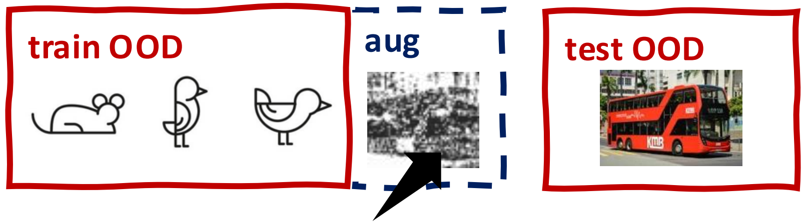
$$\mathbb{E}_{x\sim\mathcal{D}_{ood}}[\boxed{w_x}\ell_{ood}(x;\boldsymbol{\theta})]$$

**reweighting**

learn from resampled data

❖ **Case 2.**

**Scarce Data → Data Augmentation**




**learn from extra data**

$$\mathbb{E}_{x\sim\mathcal{D}_{ood}}[\ell_{ood}(x';\boldsymbol{\theta})],\ x' = \boxed{\Phi(x)}$$

**augmentation**

learn from extra data

[a] Wang et al. Learning to Augment Distributions for Out-of-distribution Detection. In NeurIPS, 2023.

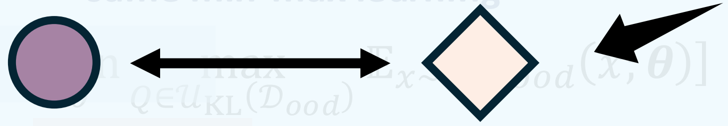
# Outlier Exposure: Improvements

Data sampling and augmentation **both improve distribution robustness** [a].

❖ Case 1.

**original distribution gap**, $\min_{\theta} \mathbb{E}_{x\sim\mathcal{D}}[\ell(x;\boldsymbol{\theta})]$

**The distribution gap degrade model performance.**




Redundant Data → Data Sampling

same min-max learning

different distribution set

$$w_x = \frac{\exp\{\ell_{ood}(x;\theta)\ /\eta\}}{\int \exp\{\ell_{ood}(x;\theta)\ /\eta\}dx}$$

up-weight OOD with larger loss

❖ Case 2.

**Scarce Data → Data Augmentation**

same min-max learning

$$\min_{\theta} \max_{Q\in\mathcal{U}_{\mathrm{W}}(\mathcal{D}_{ood})} \mathbb{E}_{x\sim Q}[\ell_{ood}(x;\boldsymbol{\theta})]$$

$$\mathcal{U}_{\mathrm{W}}(\mathcal{D}_{ood}) := \{Q: \mathrm{Wass}(Q,\mathcal{D}_{ood}) \le \rho\}$$

different distribution set

$$\min_{\theta} \mathbb{E}_{x\sim\mathcal{D}_{ood}}[\ell_{ood}(x';\theta)]$$

$$x' = \underbrace{\sup_{x'}\{\ell_{ood}(x';\theta) - \gamma\|x',x\|_2\}}_{\text{augment OOD with larger loss}}$$

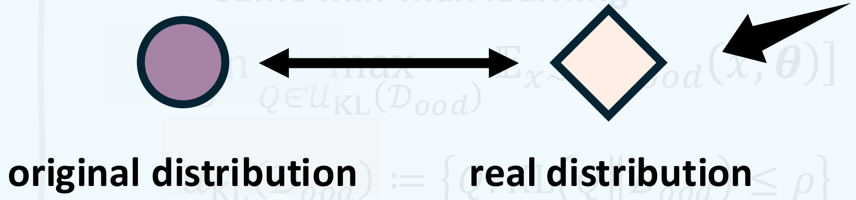
[a] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements

Data sampling and augmentation **both improve distribution robustness** [a].

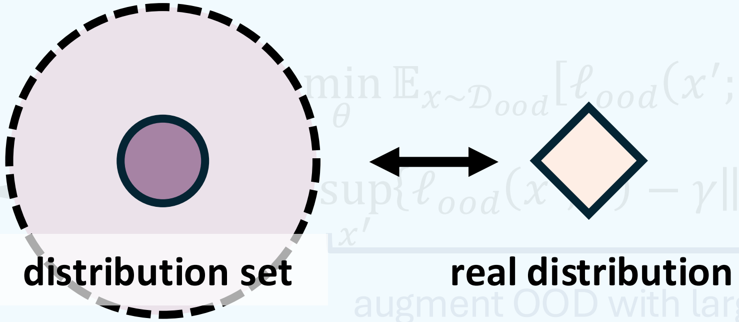
**original distribution gap**, $\min_{\theta} \mathbb{E}_{x\sim\mathcal{D}}[\ell(x;\boldsymbol{\theta})]$

**The distribution gap degrade model performance.**




**augmented distribution gap**, $\min_{\theta} \boxed{\max_{Q\in\mathcal{U}(\mathcal{D})}} \mathbb{E}_{x\sim Q}[\ell(x;\boldsymbol{\theta})]$

**The model performs well for all distribution within the set.**




[a] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements

Data sampling and augmentation **both improve distribution robustness** [a].

❖ **Case 1.**

**Redundant Data → Data Sampling**

**same min-max learning**

$$\min_{\theta} \max_{Q \in \mathcal{U}_{\mathrm{KL}}(\mathcal{D}_{ood})} \mathbb{E}_{x\sim Q}[\ell_{ood}(x;\boldsymbol{\theta})]$$

$$\mathcal{U}_{\mathrm{KL}}(\mathcal{D}_{ood}) \coloneqq \{Q: \mathrm{KL}(Q \| \mathcal{D}_{ood}) \le \rho\}$$

**different distribution set**

$$\min_{\theta} \mathbb{E}_{x\sim\mathcal{D}_{ood}}[w_x \ell_{ood}(x;\theta)]$$

$$\underbrace{w_x = \frac{\exp\{\ell_{ood}(x;\theta)\ /\eta\}}{\int \exp\{\ell_{ood}(x;\theta)\ /\eta\} dx}}_{\text{up-weight OOD with larger loss}}$$

❖ **Case 2.**

**Scarce Data → Data Augmentation**

**same min-max learning**

$$\min_{\theta} \max_{Q \in \mathcal{U}_{\mathrm{W}}(\mathcal{D}_{ood})} \mathbb{E}_{x\sim Q}[\ell_{ood}(x;\boldsymbol{\theta})]$$

$$\mathcal{U}_{\mathrm{W}}(\mathcal{D}_{ood}) \coloneqq \{Q: \mathrm{Wass}(Q, \mathcal{D}_{ood}) \le \rho\}$$

**different distribution set**

$$\min_{\theta} \mathbb{E}_{x\sim\mathcal{D}_{ood}}[\ell_{ood}(x';\theta)]$$

$$\underbrace{x' = \sup_{x'}\{\ell_{ood}(x';\theta) - \gamma\|x', x\|_2\}}_{\text{augment OOD with larger loss}}$$

[a] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements

Data sampling and augmentation **both improve distribution robustness** [a].

❖ **Case 1.**

**Redundant Data → Data Sampling**

same min-max learning

$$\min_{\theta} \max_{Q \in \mathcal{U}_{\mathrm{KL}}(\mathcal{D}_{ood})} \mathbb{E}_{x \sim Q}[\ell_{ood}(x; \boldsymbol{\theta})]$$

$$\mathcal{U}_{\mathrm{KL}}(\mathcal{D}_{ood}) \coloneqq \{Q: \mathrm{KL}(Q \| \mathcal{D}_{ood}) \le \rho\}$$

different distribution set

→

$$\min_{\theta} \mathbb{E}_{x \sim \mathcal{D}_{ood}}[w_x \ell_{ood}(x; \boldsymbol{\theta})]$$

$$\underbrace{w_x = \frac{\exp\{\ell_{ood}(x;\boldsymbol{\theta})\ /\eta\}}{\int \exp\{\ell_{ood}(x;\boldsymbol{\theta})\ /\eta\} dx}}_{\text{up-weight OOD with larger loss}}$$

**up-weight** OOD with larger loss

❖ **Case 2.**

**Scarce Data → Data Augmentation**

same min-max learning

$$\min_{\theta} \max_{Q \in \mathcal{U}_{\mathrm{W}}(\mathcal{D}_{ood})} \mathbb{E}_{x \sim Q}[\ell_{ood}(x; \boldsymbol{\theta})]$$

$$\mathcal{U}_{\mathrm{W}}(\mathcal{D}_{ood}) \coloneqq \{Q: \mathrm{Wass}(Q, \mathcal{D}_{ood}) \le \rho\}$$

different distribution set

→

$$\min_{\theta} \mathbb{E}_{x \sim \mathcal{D}_{ood}}[\ell_{ood}(x'; \boldsymbol{\theta})]$$

$$x' = \sup_{x'}\{\ell_{ood}(x'; \boldsymbol{\theta}) - \gamma \|x', x\|_2\}$$

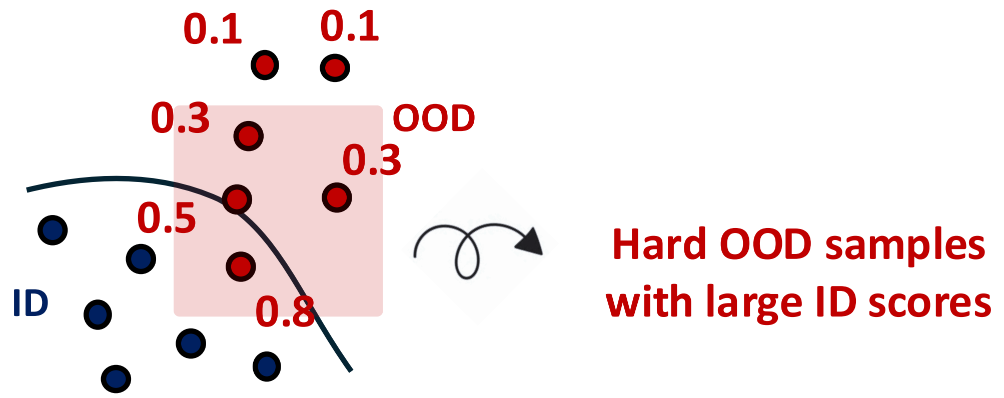
**augment** OOD with larger loss

[a] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements | Sampling

## ATOM [a]

- **Hard OOD data** should be sampled more during training.
- **Large ID scores on OOD samples** indicates hard OOD data.




## POEM [b]

- OOD data **near the ID/OOD boundary** should be sampled more.
- Bayesian classifiers capture parameter uncertainty, **improving exploration.**





[a] Chen et al. ATOM: Robustifying Out-of-distribution Detection Using Outlier Mining. In ECML PKDD, 2021.
[b] Ming et al. POEM: Out-of-distribution Detection with Posterior Sampling. In ICML, 2022.

# Outlier Exposure: Improvements | Sampling

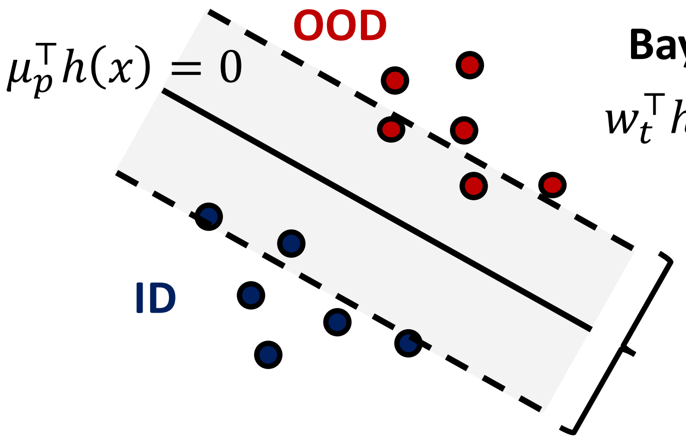
## ATOM [a]

- ❖ **Hard OOD data** should be sampled more during training.
- ❖ **Large ID scores on OOD samples** indicates hard OOD data.




## POEM [b]

- ❖ OOD data **near the ID/OOD boundary** should be sampled more.
- ❖ Bayesian classifiers capture parameter uncertainty, **improving exploration.**




**Bayesian ID/OOD classifier**

$w_t^\top h(x) = 0, w_t \sim \mathcal{N}(\mu_p, \Sigma_p)$

**High Variance in $w_t$, measured by $\Sigma_\mathrm{p}$.**

[a] Chen et al. ATOM: Robustifying Out-of-distribution Detection Using Outlier Mining. In ECML PKDD, 2021.
[b] Ming et al. POEM: Out-of-distribution Detection with Posterior Sampling. In ICML, 2022.

# Outlier Exposure: Improvements | Sampling

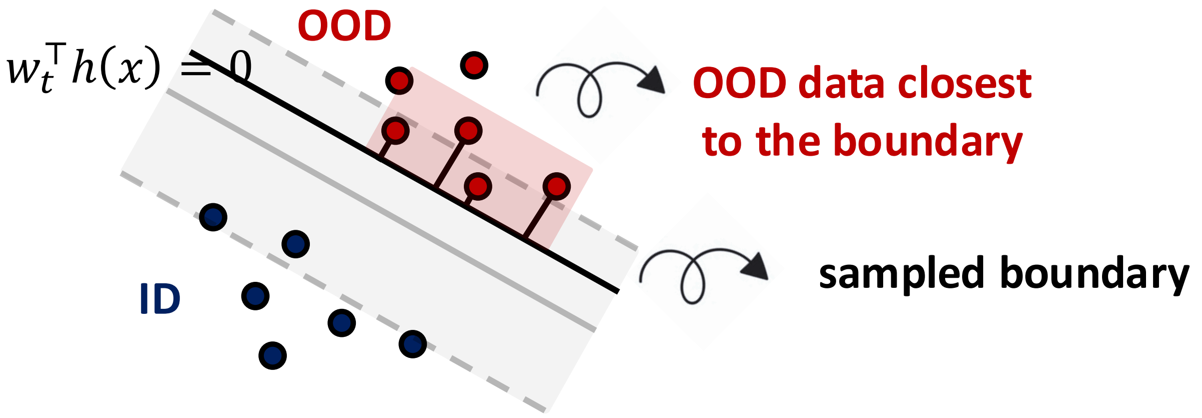
## **ATOM** [a]

- ❖ **Hard OOD data** should be sampled more during training.
- ❖ **Large ID scores on OOD samples** indicates hard OOD data.




## **POEM** [b]

- ❖ OOD data **near the ID/OOD boundary** should be sampled more**.**
- ❖ Bayesian classifiers capture parameter uncertainty, **improving exploration.**




[a] Chen et al. ATOM: Robustifying Out-of-distribution Detection Using Outlier Mining. In ECML PKDD, 2021.
[b] Ming et al. POEM: Out-of-distribution Detection with Posterior Sampling. In ICML, 2022.

# Outlier Exposure: Improvements | Sampling

**ATOM** [a]

❖ **Hard OOD data** should be sampled more during training.

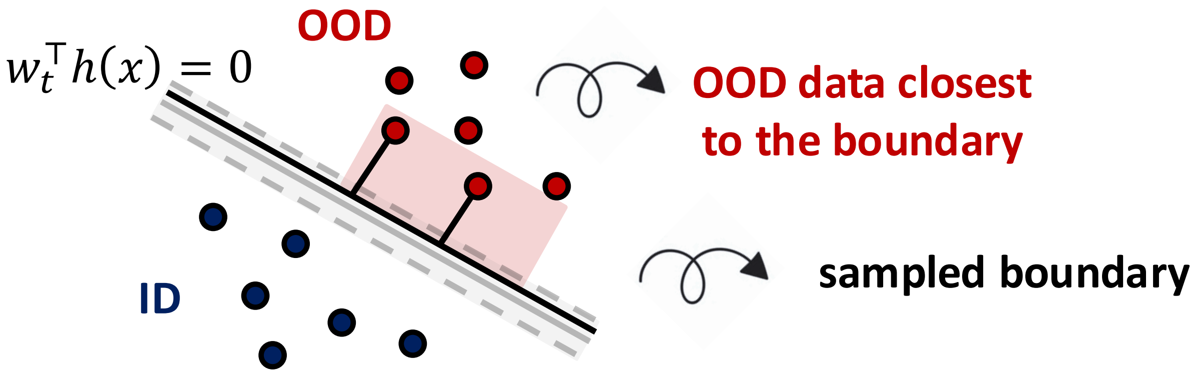
❖ **Large ID scores on OOD samples** indicates hard OOD data.




**POEM** [b]

❖ OOD data **near the ID/OOD boundary** should be sampled more.

❖ Bayesian classifiers capture parameter uncertainty, **improving exploration.**





[a] Chen et al. ATOM: Robustifying Out-of-distribution Detection Using Outlier Mining. In ECML PKDD, 2021.
[b] Ming et al. POEM: Out-of-distribution Detection with Posterior Sampling. In ICML, 2022.

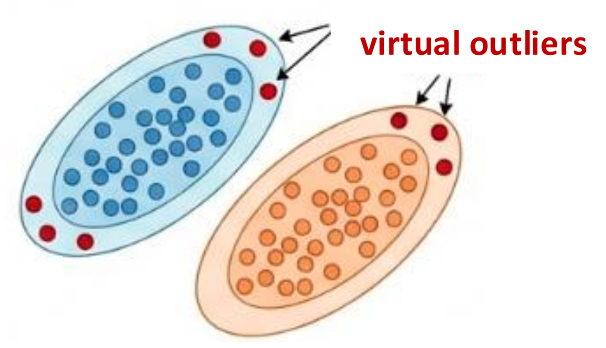
# Outlier Exposure: Improvements | Augmentation

Augmentation can be conducted in either **embedding space** or input space.

### ❖ Synthesis

VOS [a]

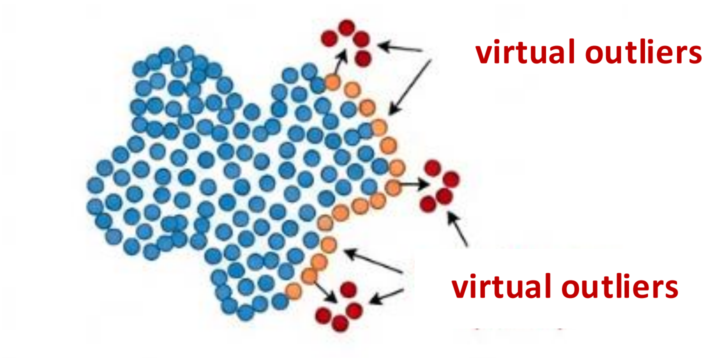
class-conditional **Gaussian modelling**




- ❖ estimate $\mu$ and $\Sigma$ per class.
- ❖ sample $v \sim \mathcal{N}(\mu_c, \Sigma)$.
- ❖ keep only low likelihood samples with $p(v) < \epsilon$.

NPOS [b]

**Non-parametric** K-NN modelling




- ❖ Identify boundary ID with high k-NN.
- ❖ sample locally around boundary $v \sim \mathcal{N}\left(h_{bdy}, \sigma^2 I\right)$.
- ❖ keep candidates with large k-NN.

### ❖ Augmentation

DAL [c]

worst-case **OOD augmentation**




- ❖ sample original OOD data.
- ❖ gradient ascent to increase their respective loss values.
- ❖ augmented OOD data for training.


[a] Du et al. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. In ICLR, 2022.
[b] Tao et al. Non-parametric Outlier Synthesis. In ICLR, 2023.
[c] Wang et al. Learning to Augment Distributions for Out-of-distribution Detection. In NeurIPS, 2023.

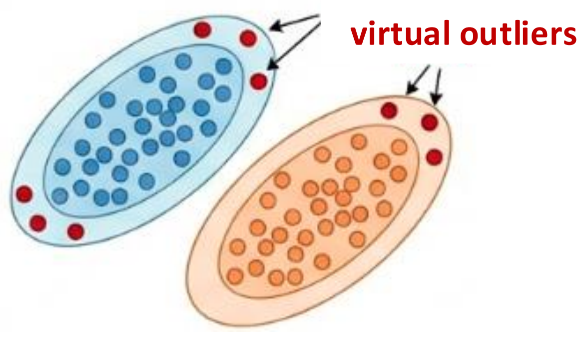
# Outlier Exposure: Improvements | Augmentation

Augmentation can be conducted in either **embedding space** or input space.

## ❖ Synthesis

VOS [a]

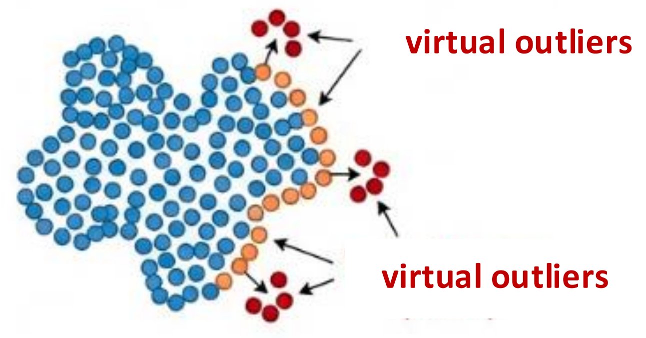
class-conditional **Gaussian modelling**




- estimate $\mu$ and $\Sigma$ per class.
- sample $v \sim \mathcal{N}(\mu_c, \Sigma)$.
- keep only low likelihood samples with $p(v) < \epsilon$.

NPOS [b]

**Non-parametric** K-NN modelling

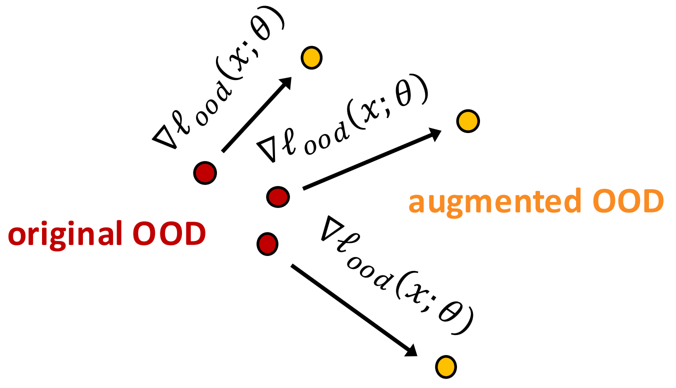



- Identify boundary ID with high k-NN.
- sample locally around boundary $v \sim \mathcal{N}\left(h_{bdy}, \sigma^2 I\right)$.
- keep candidates with large k-NN.

## ❖ Augmentation

DAL [c]

worst-case **OOD augmentation**



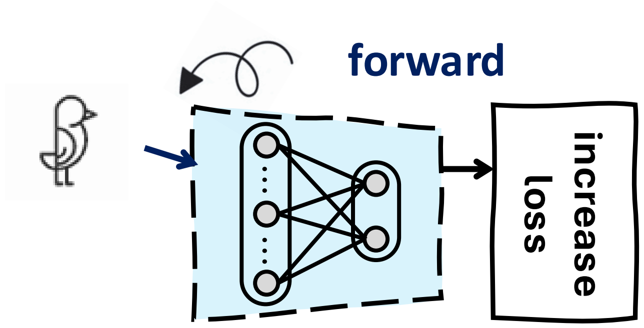

- sample original OOD data.
- gradient ascent to increase their respective loss values.
- augmented OOD data for training.


[a] Du et al. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. In ICLR, 2022.
[b] Tao et al. Non-parametric Outlier Synthesis. In ICLR, 2023.
[c] Wang et al. Learning to Augment Distributions for Out-of-distribution Detection. In NeurIPS, 2023.

# Outlier Exposure: Improvements | Augmentation

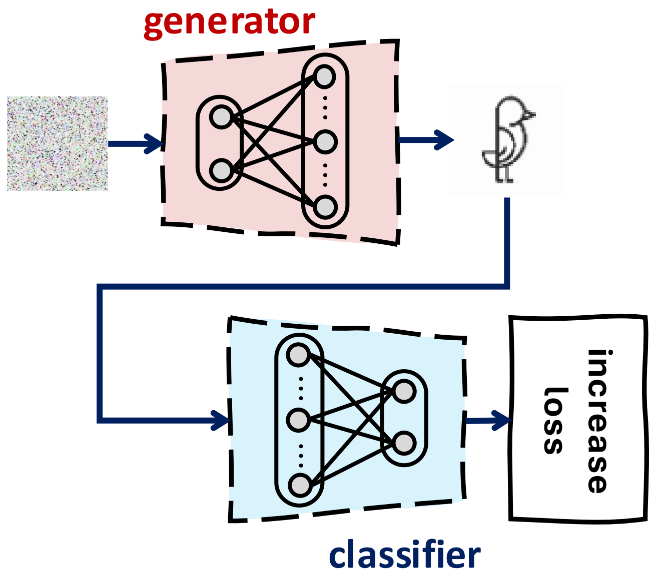
Augmentation can be conducted in either embedding space or **input space**.

### ❖ Adversarial Training [a]



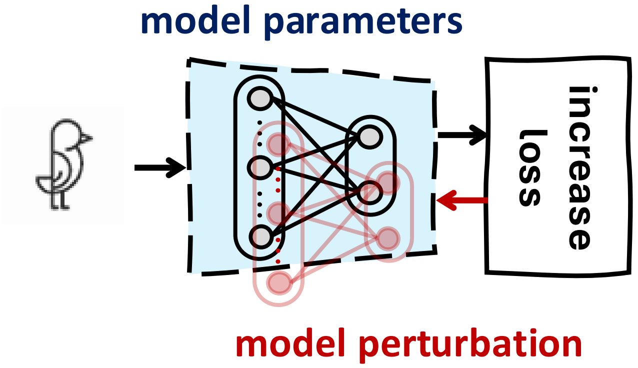

**perturb inputs** to maximize OOD loss values, thus creating worst-case examples.

### ❖ Generative Model [b]




**perturb latent embedding** to maximize OOD loss values, thus creating worst-case examples.

### ❖ Implicit Synthesis [c]




$$\underbrace{[W(I+\alpha A)]^\top x}_{\text{model perturb}} = W^\top \underbrace{[x+\alpha A^\top x]}_{\text{feature perturb}}$$

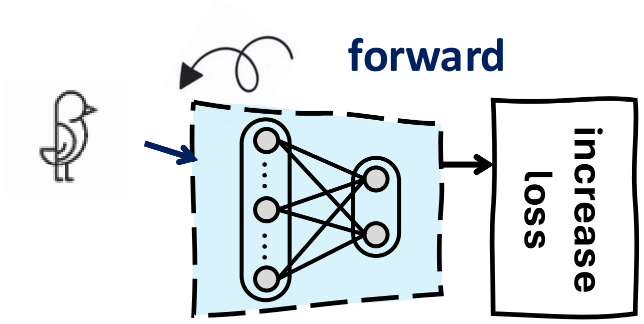
**perturb parameters** that maximize OOD loss can implicitly create worst-case examples.

[a] Zhu et al. Diversified Outlier Exposure for Out-of-distribution Detection via Informative Extrapolation. In NeurIPS, 2023.
[b] Lee et al. Training Confidence-calibrated Classifiers for Detecting Out-of-distribution Samples. In ICLR, 2018.
[c] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements | Augmentation

Augmentation can be conducted in either embedding space or **input space**.

❖ **Adversarial Training** [a]




**perturb inputs** to maximize OOD loss values, thus creating worst-case examples.

❖ **Generative Model** [b]




**perturb latent embedding** to maximize OOD loss values, thus creating worst-case examples.

❖ **Implicit Synthesis** [c]




$$\underbrace{[W(I+\alpha A)]^\top x}_{\text{model perturb}} = \underbrace{W^\top[x+\alpha A^\top x]}_{\text{feature perturb}}$$

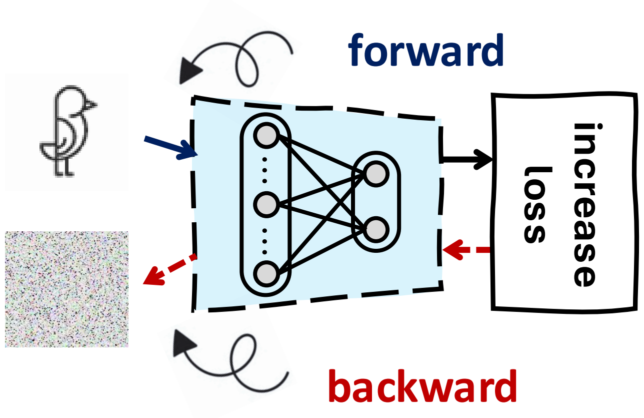
**perturb parameters** that maximize OOD loss can implicitly create worst-case examples.

[a] Zhu et al. Diversified Outlier Exposure for Out-of-distribution Detection via Informative Extrapolation. In NeurIPS, 2023.
[b] Lee et al. Training Confidence-calibrated Classifiers for Detecting Out-of-distribution Samples. In ICLR, 2018.
[c] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements | Augmentation

Augmentation can be conducted in either embedding space or **input space**.

### ❖ Adversarial Training [a]




**perturb inputs** to maximize OOD loss values, thus creating worst-case examples.

### ❖ Generative Model [b]




**perturb latent embedding** to maximize OOD loss values, thus creating worst-case examples.

### ❖ Implicit Synthesis [c]




$$\underbrace{[W(I+\alpha A)]^\top x}_{\text{model perturb}} = \underbrace{W^\top[x+\alpha A^\top x]}_{\text{feature perturb}}$$

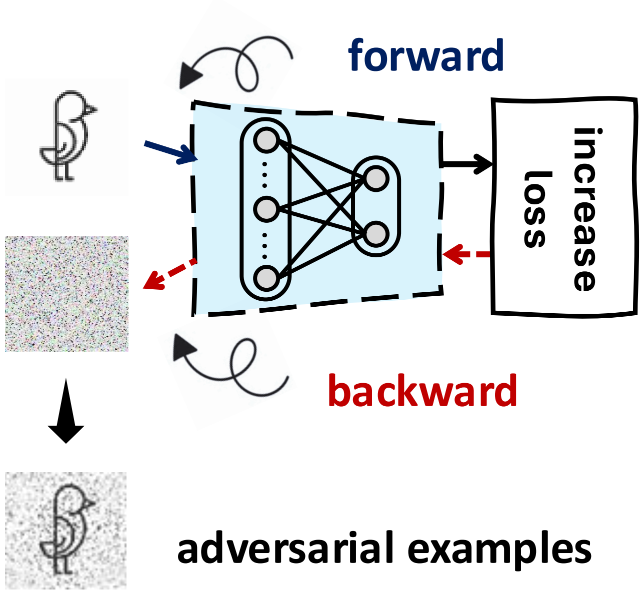
**perturb parameters** that maximize OOD loss can implicitly create worst-case examples.


[a] Zhu et al. Diversified Outlier Exposure for Out-of-distribution Detection via Informative Extrapolation. In NeurIPS, 2023.
[b] Lee et al. Training Confidence-calibrated Classifiers for Detecting Out-of-distribution Samples. In ICLR, 2018.
[c] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements | Augmentation

Augmentation can be conducted in either embedding space or **input space**.

### ❖ Adversarial Training [a]




**perturb inputs** to maximize OOD loss values, thus creating worst-case examples.

### ❖ Generative Model [b]




**perturb latent embedding** to maximize OOD loss values, thus creating worst-case examples.

### ❖ Implicit Synthesis [c]



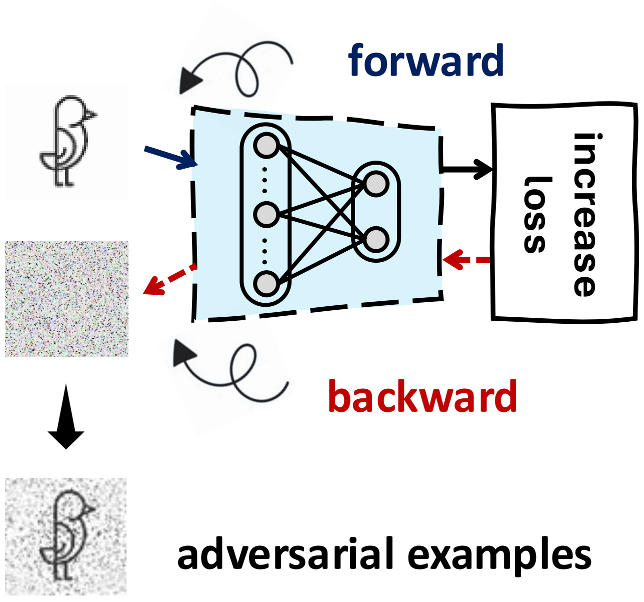

**perturb parameters** that maximize OOD loss can implicitly create worst-case examples.


[a] Zhu et al. Diversified Outlier Exposure for Out-of-distribution Detection via Informative Extrapolation. In NeurIPS, 2023.
[b] Lee et al. Training Confidence-calibrated Classifiers for Detecting Out-of-distribution Samples. In ICLR, 2018.
[c] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements | Augmentation

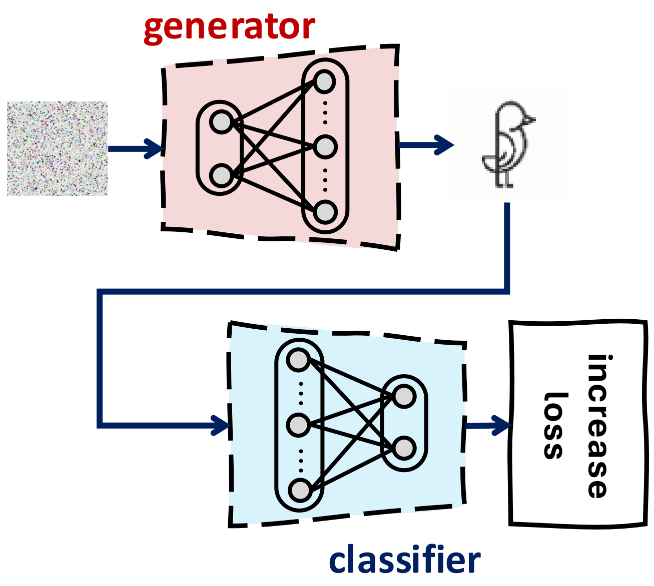
Augmentation can be conducted in either embedding space or **input space**.

### ❖ **Adversarial Training** [a]




**perturb inputs** to maximize OOD loss values, thus creating worst-case examples.

### ❖ **Generative Model** [b]




**perturb latent embedding** to maximize OOD loss values, thus creating worst-case examples.

### ❖ **Implicit Synthesis** [c]



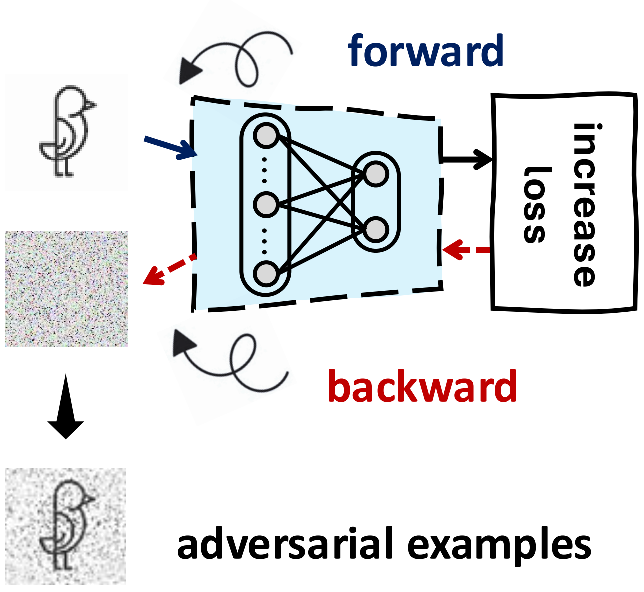

**perturb parameters** that maximize OOD loss can implicitly create worst-case examples.


[a] Zhu et al. Diversified Outlier Exposure for Out-of-distribution Detection via Informative Extrapolation. In NeurIPS, 2023.
[b] Lee et al. Training Confidence-calibrated Classifiers for Detecting Out-of-distribution Samples. In ICLR, 2018.
[c] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements | Augmentation

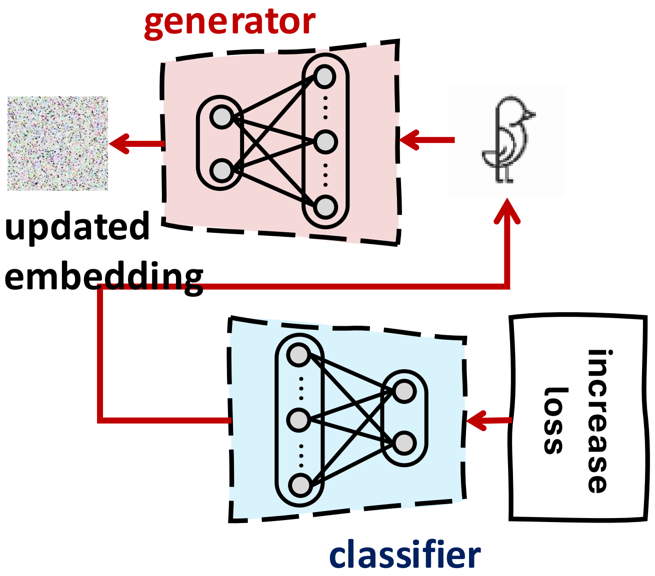
Augmentation can be conducted in either embedding space or **input space**.

### ❖ Adversarial Training [a]




**perturb inputs** to maximize OOD loss values, thus creating worst-case examples.

### ❖ Generative Model [b]




**perturb latent embedding** to maximize OOD loss values, thus creating worst-case examples.

### ❖ Implicit Synthesis [c]



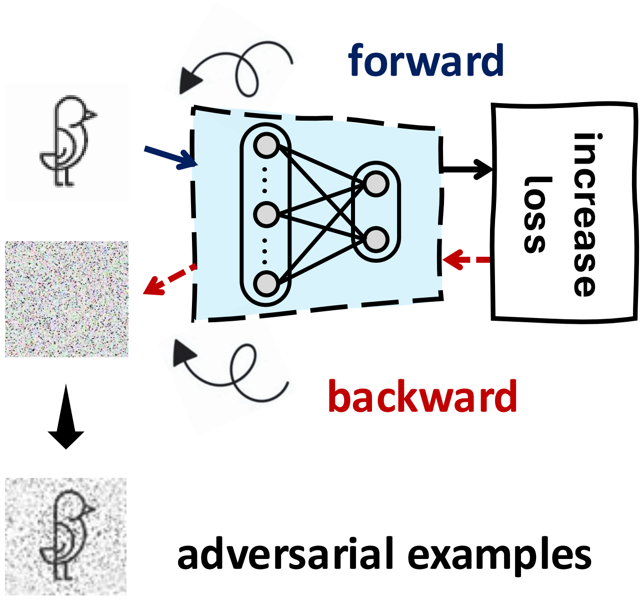

**perturb parameters** that maximize OOD loss can implicitly create worst-case examples.

[a] Zhu et al. Diversified Outlier Exposure for Out-of-distribution Detection via Informative Extrapolation. In NeurIPS, 2023.
[b] Lee et al. Training Confidence-calibrated Classifiers for Detecting Out-of-distribution Samples. In ICLR, 2018.
[c] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements | Augmentation

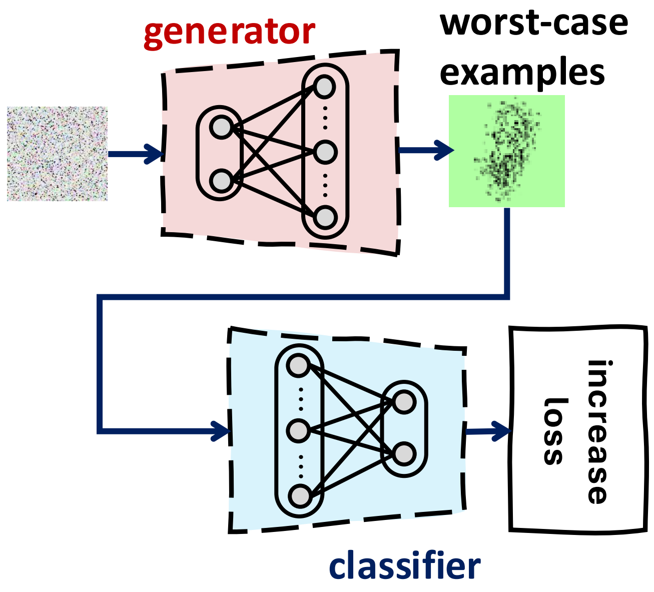
Augmentation can be conducted in either embedding space or **input space**.

### ❖ Adversarial Training [a]




**perturb inputs** to maximize OOD loss values, thus creating worst-case examples.

### ❖ Generative Model [b]




**perturb latent embedding** to maximize OOD loss values, thus creating worst-case examples.

### ❖ Implicit Synthesis [c]




$$\underbrace{[W(I+\alpha A)]^\top x}_{\text{model perturb}} = \underbrace{W^\top[x+\alpha A^\top x]}_{\text{feature perturb}}$$

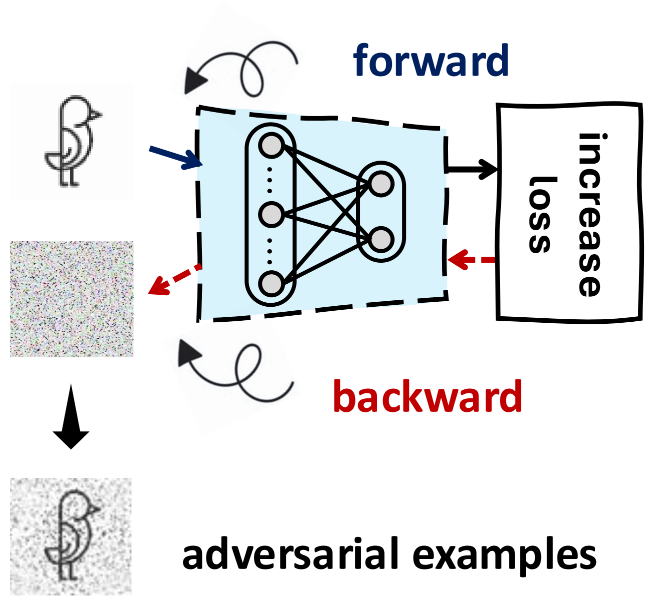
**perturb parameters** that maximize OOD loss can implicitly create worst-case examples.


[a] Zhu et al. Diversified Outlier Exposure for Out-of-distribution Detection via Informative Extrapolation. In NeurIPS, 2023.
[b] Lee et al. Training Confidence-calibrated Classifiers for Detecting Out-of-distribution Samples. In ICLR, 2018.
[c] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements | Augmentation

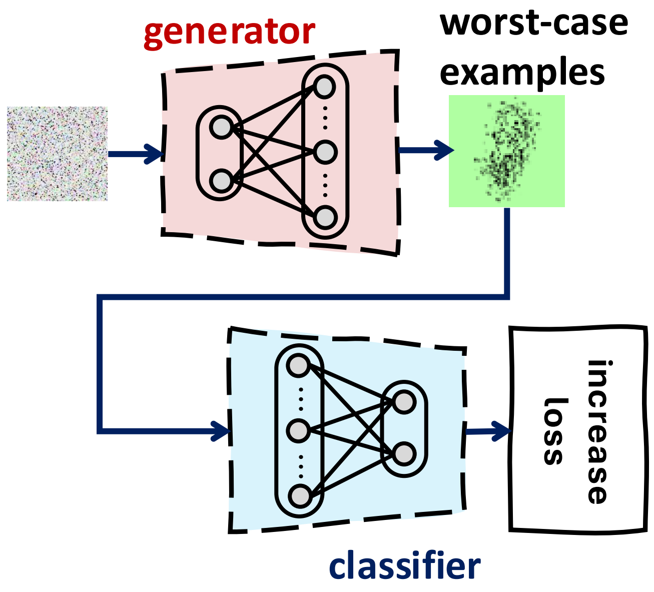
Augmentation can be conducted in either embedding space or **input space**.

❖ **Adversarial Training** [a]



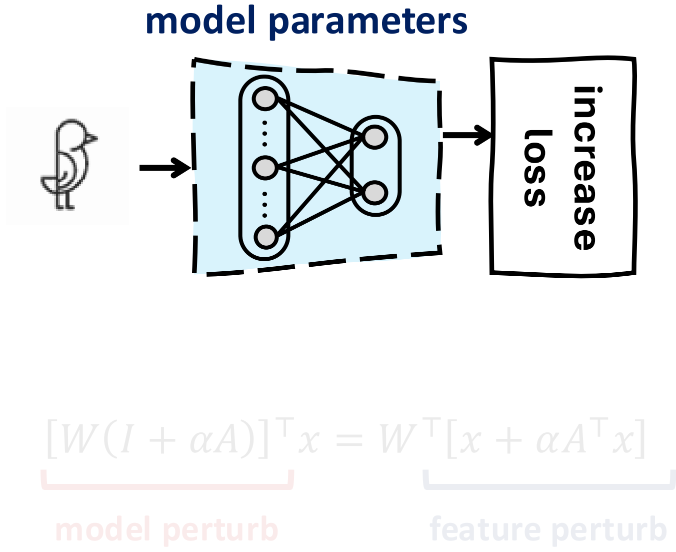

**perturb inputs** to maximize OOD loss values, thus creating worst-case examples.

❖ **Generative Model** [b]




**perturb latent embedding** to maximize OOD loss values, thus creating worst-case examples.

❖ **Implicit Synthesis** [c]



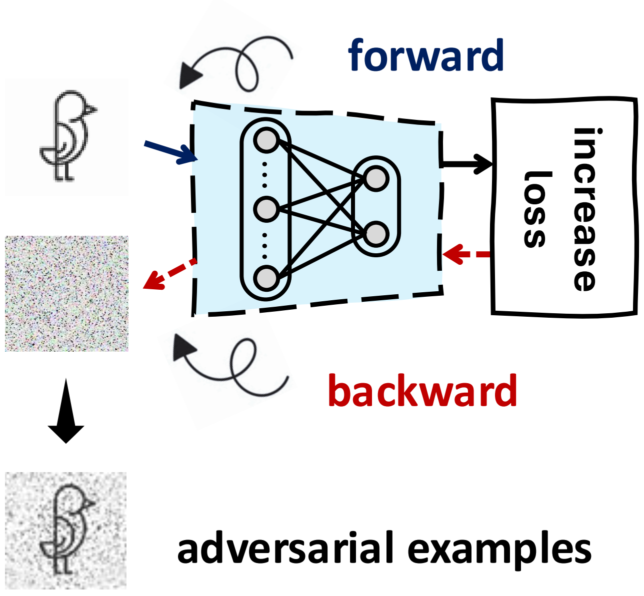

**perturb parameters** that maximize OOD loss can implicitly create worst-case examples.


[a] Zhu et al. Diversified Outlier Exposure for Out-of-distribution Detection via Informative Extrapolation. In NeurIPS, 2023.
[b] Lee et al. Training Confidence-calibrated Classifiers for Detecting Out-of-distribution Samples. In ICLR, 2018.
[c] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements | Augmentation

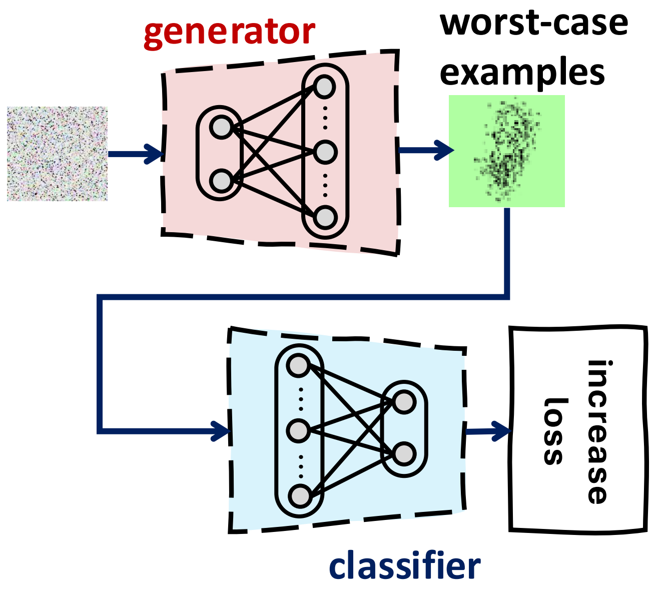
Augmentation can be conducted in either embedding space or **input space**.

❖ **Adversarial Training** [a]



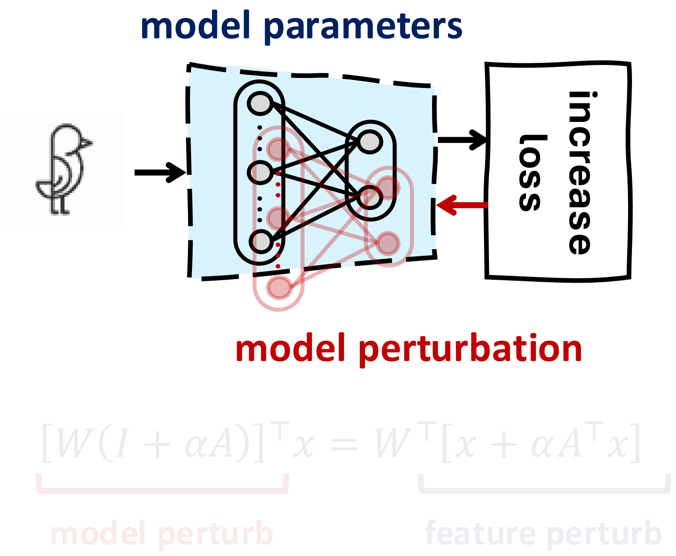

**perturb inputs** to maximize OOD loss values, thus creating worst-case examples.

❖ **Generative Model** [b]




**perturb latent embedding** to maximize OOD loss values, thus creating worst-case examples.

❖ **Implicit Synthesis** [c]



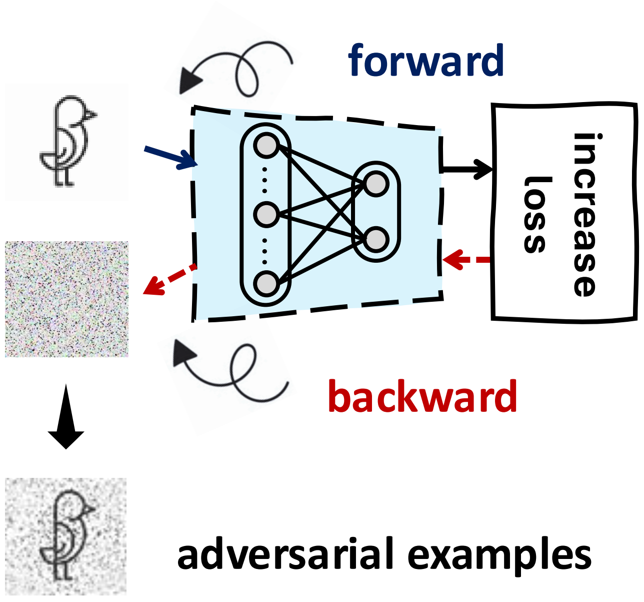

**perturb parameters** that maximize OOD loss can implicitly create worst-case examples.

[a] Zhu et al. Diversified Outlier Exposure for Out-of-distribution Detection via Informative Extrapolation. In NeurIPS, 2023.
[b] Lee et al. Training Confidence-calibrated Classifiers for Detecting Out-of-distribution Samples. In ICLR, 2018.
[c] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements | Augmentation

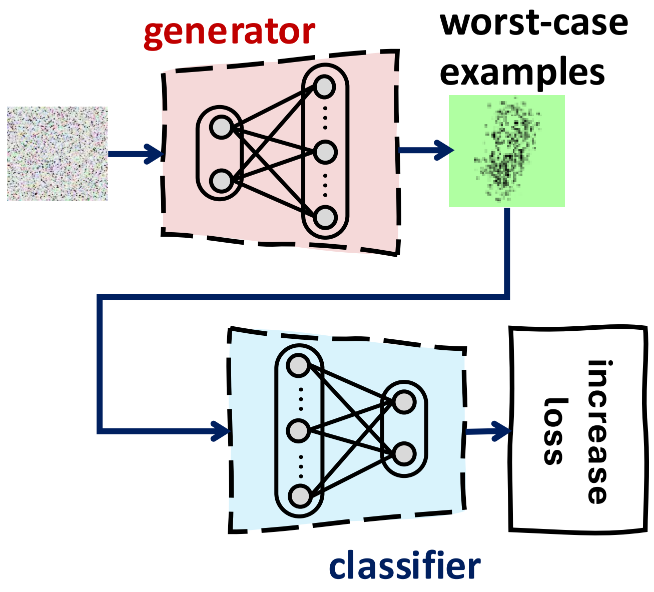
Augmentation can be conducted in either embedding space or **input space**.

### ❖ Adversarial Training [a]



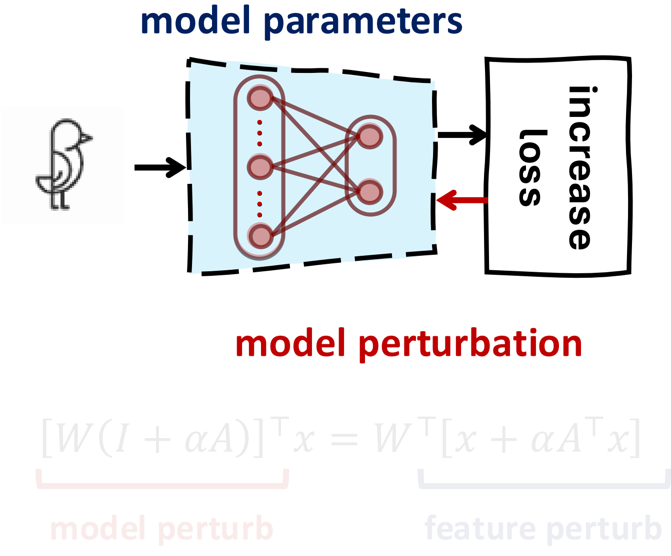

**perturb inputs** to maximize OOD loss values, thus creating worst-case examples.

### ❖ Generative Model [b]




**perturb latent embedding** to maximize OOD loss values, thus creating worst-case examples.

### ❖ Implicit Synthesis [c]



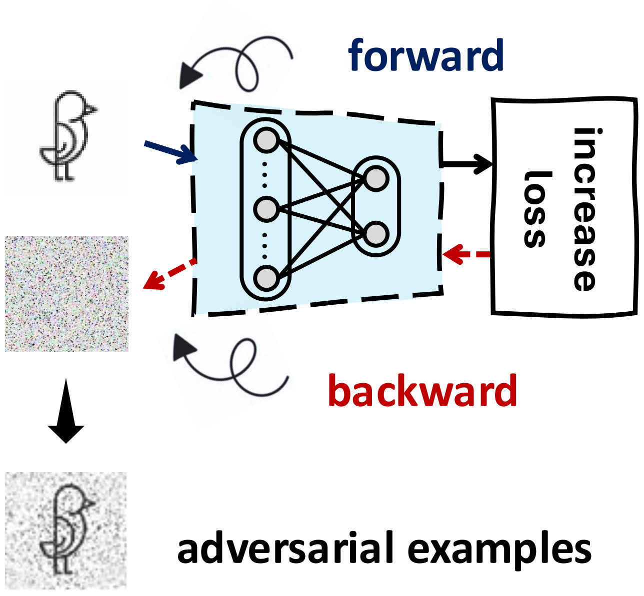

**perturb parameters** that maximize OOD loss can implicitly create worst-case examples.

[a] Zhu et al. Diversified Outlier Exposure for Out-of-distribution Detection via Informative Extrapolation. In NeurIPS, 2023.
[b] Lee et al. Training Confidence-calibrated Classifiers for Detecting Out-of-distribution Samples. In ICLR, 2018.
[c] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Outlier Exposure: Improvements | Augmentation

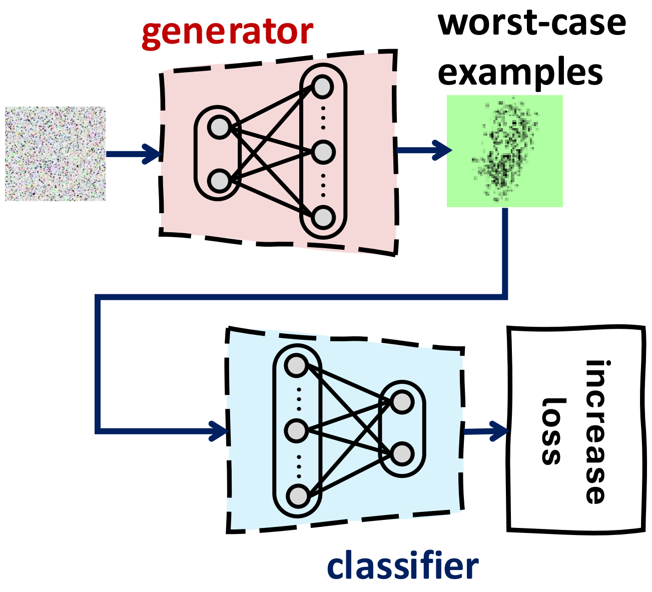
Augmentation can be conducted in either embedding space or **input space**.

❖ **Adversarial Training** [a]



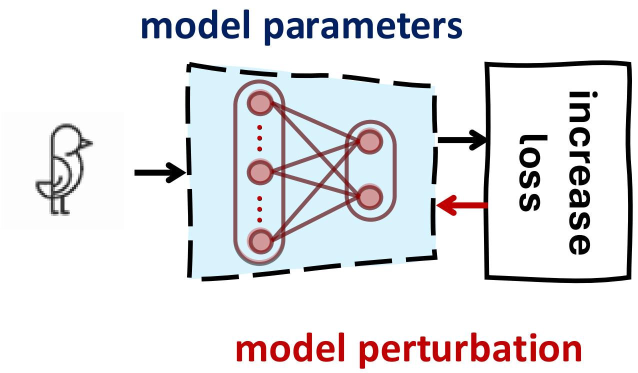

**perturb inputs** to maximize OOD loss values, thus creating worst-case examples.

❖ **Generative Model** [b]




**perturb latent embedding** to maximize OOD loss values, thus creating worst-case examples.

❖ **Implicit Synthesis** [c]



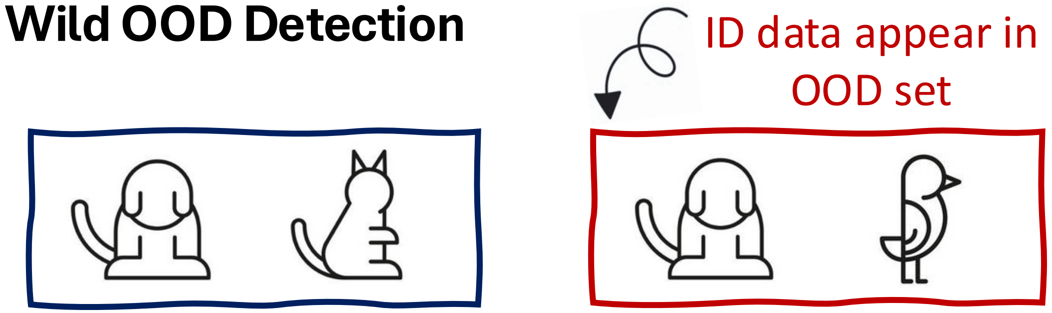

$$\underbrace{[W(I+\alpha A)]^\top}_{\text{model perturb}} x = W^\top \underbrace{[x+\alpha A^\top x]}_{\text{feature perturb}}$$

**perturb parameters** that maximize OOD loss can implicitly create worst-case examples.

[a] Zhu et al. Diversified Outlier Exposure for Out-of-distribution Detection via Informative Extrapolation. In NeurIPS, 2023.
[b] Lee et al. Training Confidence-calibrated Classifiers for Detecting Out-of-distribution Samples. In ICLR, 2018.
[c] Wang et al. Out-of-distribution Detection with Implicit Outlier Transformation. In ICLR, 2023.

# Challenges and Future Directions

❖ **Wild OOD Detection**




**Wrongly taking ID data as OOD for OE training** can degrade OOD detection performance.

# Challenges and Future Directions

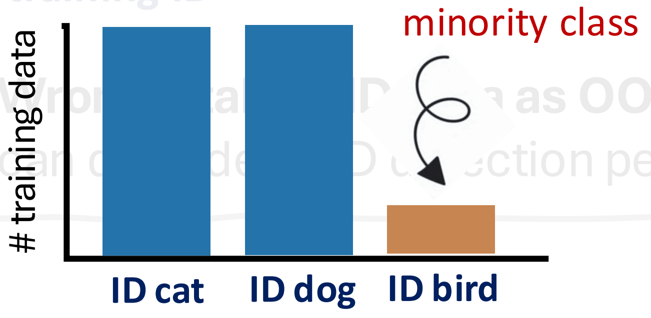
Wild OOD Detection

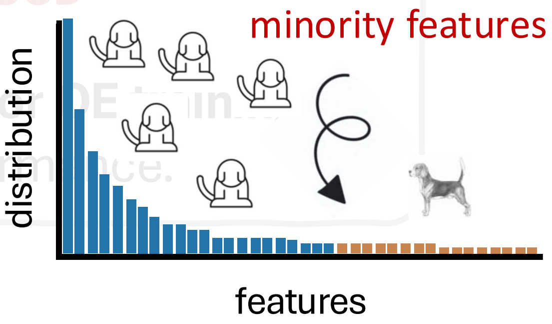
ID data appear in OOD set

training ID

training OOD

## ❖ Imbalanced and Long-tailed OOD Learning




**Minority classes or features** are easier to be confused with OOD data.

# Challenges and Future Directions

❖ Wild OOD Detection

ID data appear in OOD set

❖ Imbalanced and Long-tailed OOD Learning

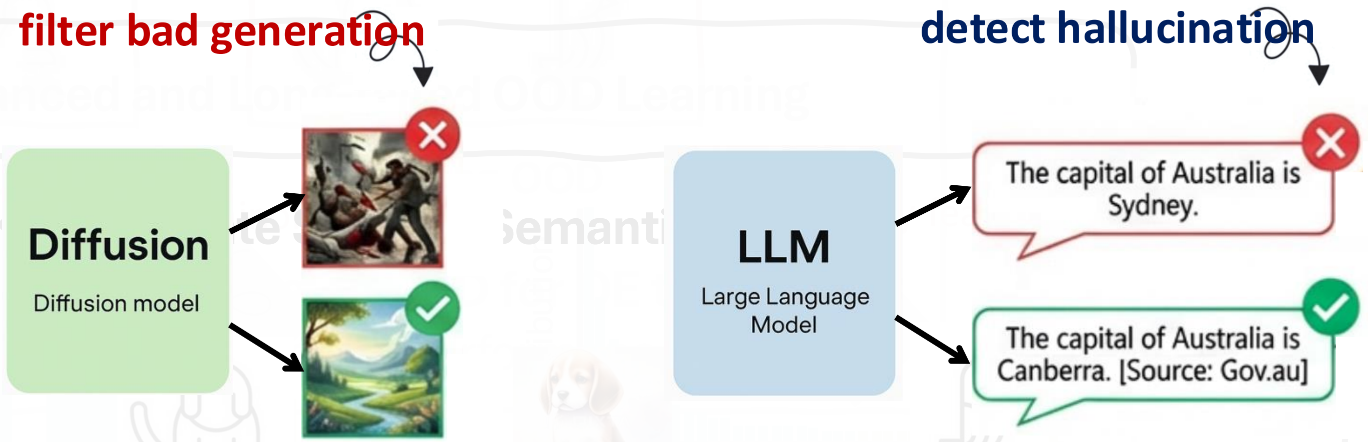
Minority classes or features are easier to be confused with OOD data.

❖ **Covariate Shift vs. Semantic Shift**


Models typically **fail to discern data with covariate shift and semantic shift.**

# Challenges and Future Directions

❖ **Large-scale Generative Models**




OOD detection can be used for **quality / safety control in large-scale generative models.**

# Thank you for listening!

Find my slides from my homepage:
https://qizhouwang.github.io/homepage