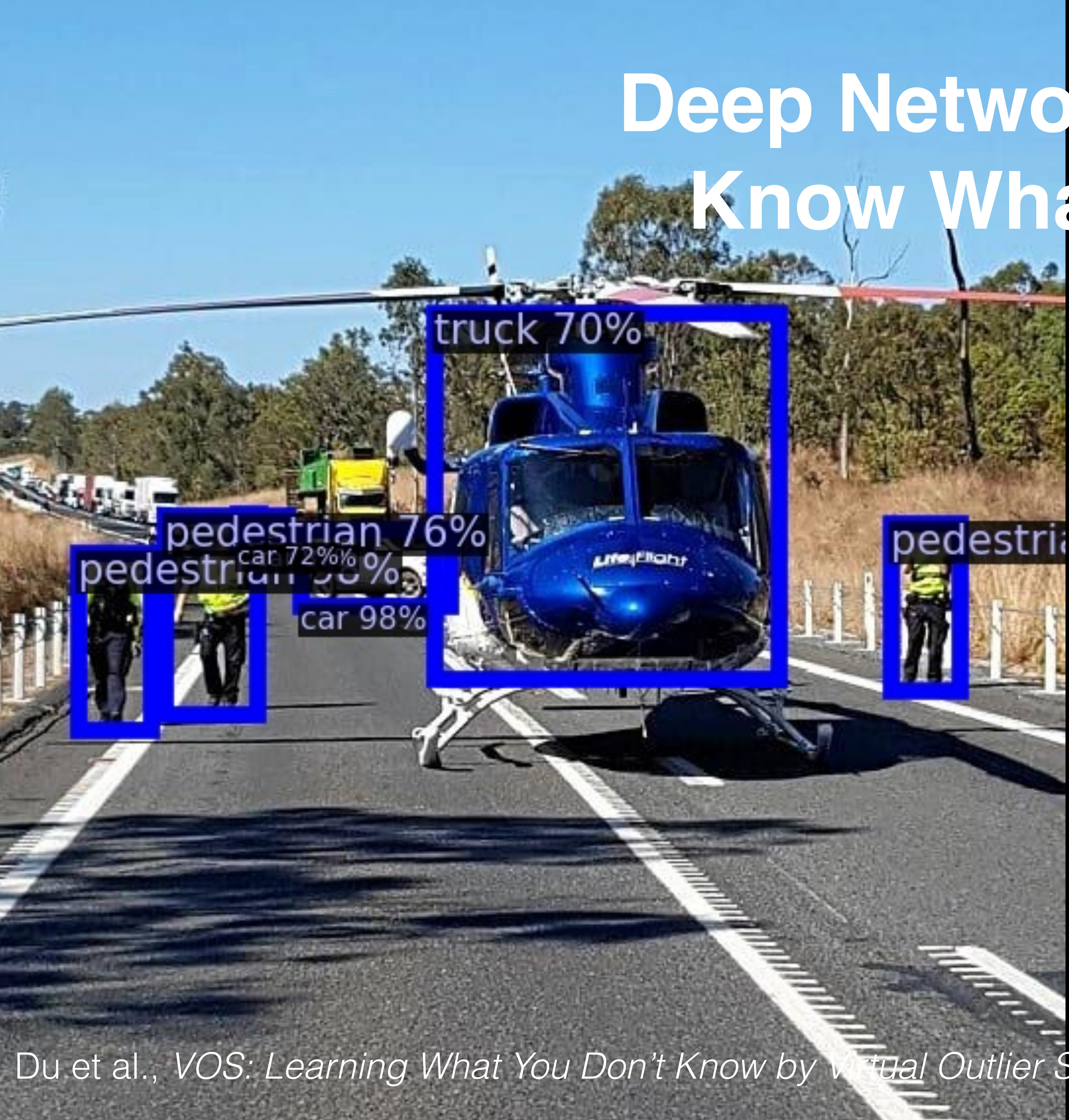
A Tutorial on Out-of-Distribution Detection

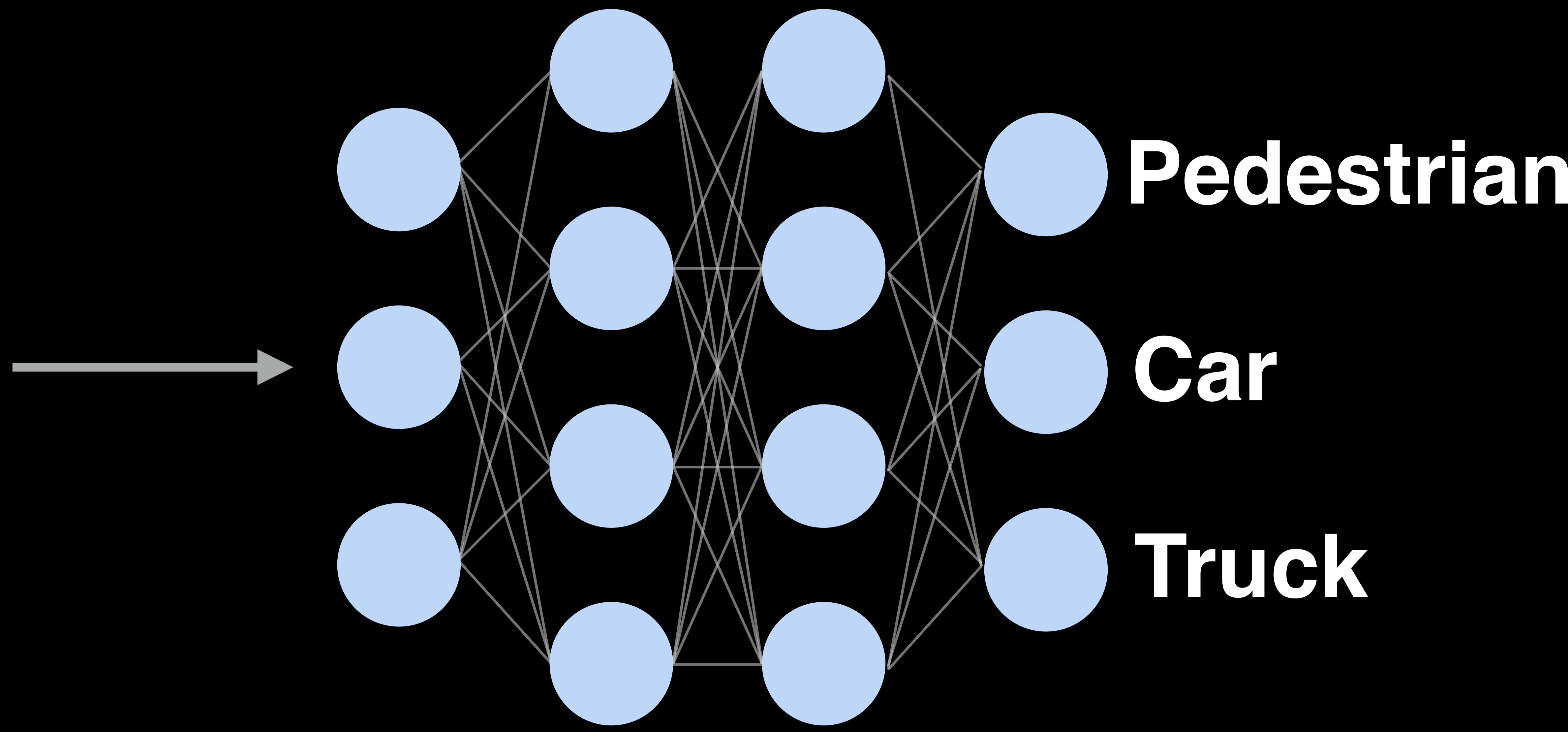
Sharon Yixuan Li
Department of Computer Sciences
University of Wisconsin-Madison

@ICCV 2023

Video source: Berkeley Deep Drive-100k

Deep Networks Do Not Necessarily Know What They Don't Know…





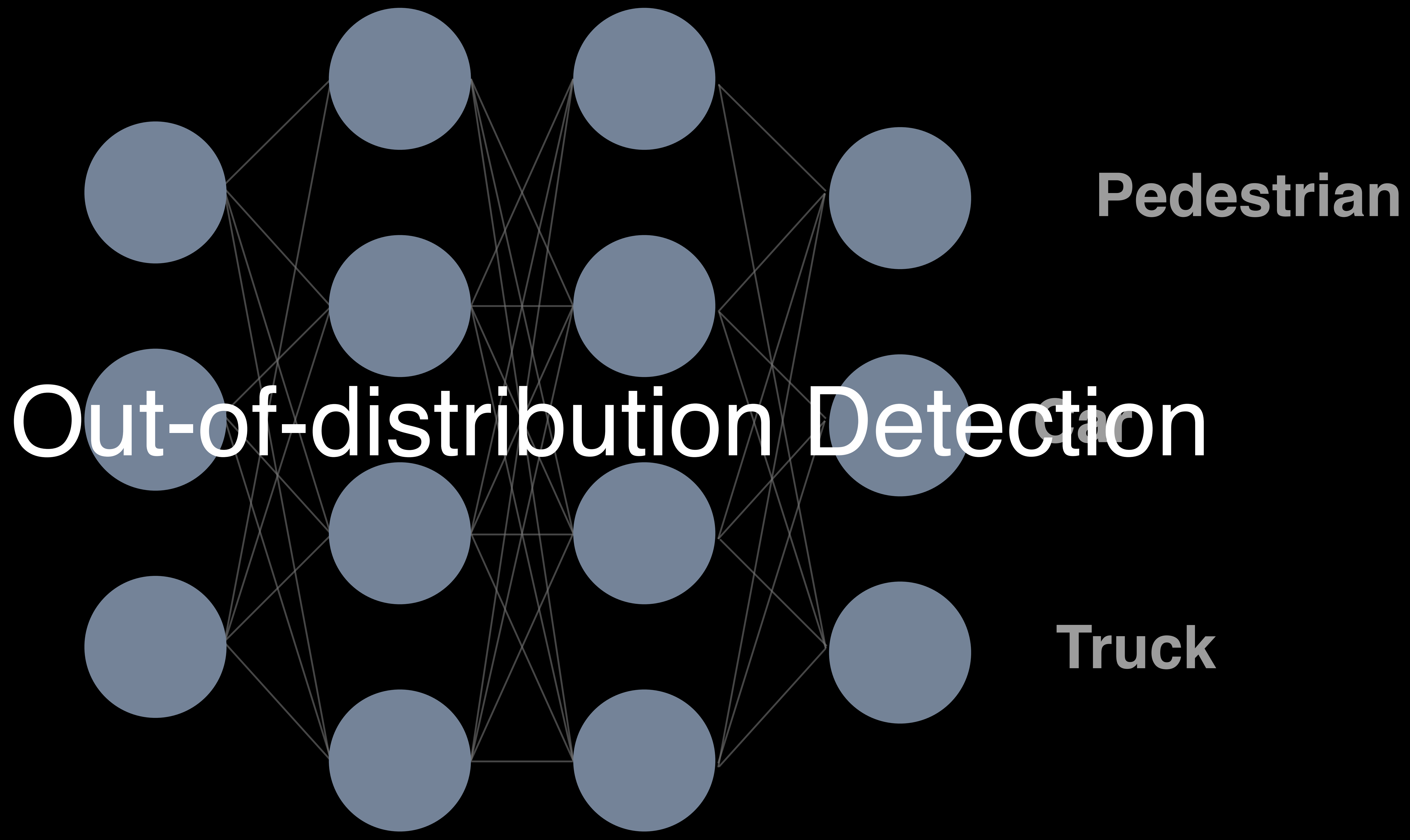
Model trained on BDD dataset produces overconfident predictions for unknown object "helicopter"

Du et al., *VOS: Learning What You Don't Know by Virtual Outlier Synthesis*, ICLR, 2022

A Tesla vehicle using ‘Smart Summon’ appears to crash into a $3.5 million private jet

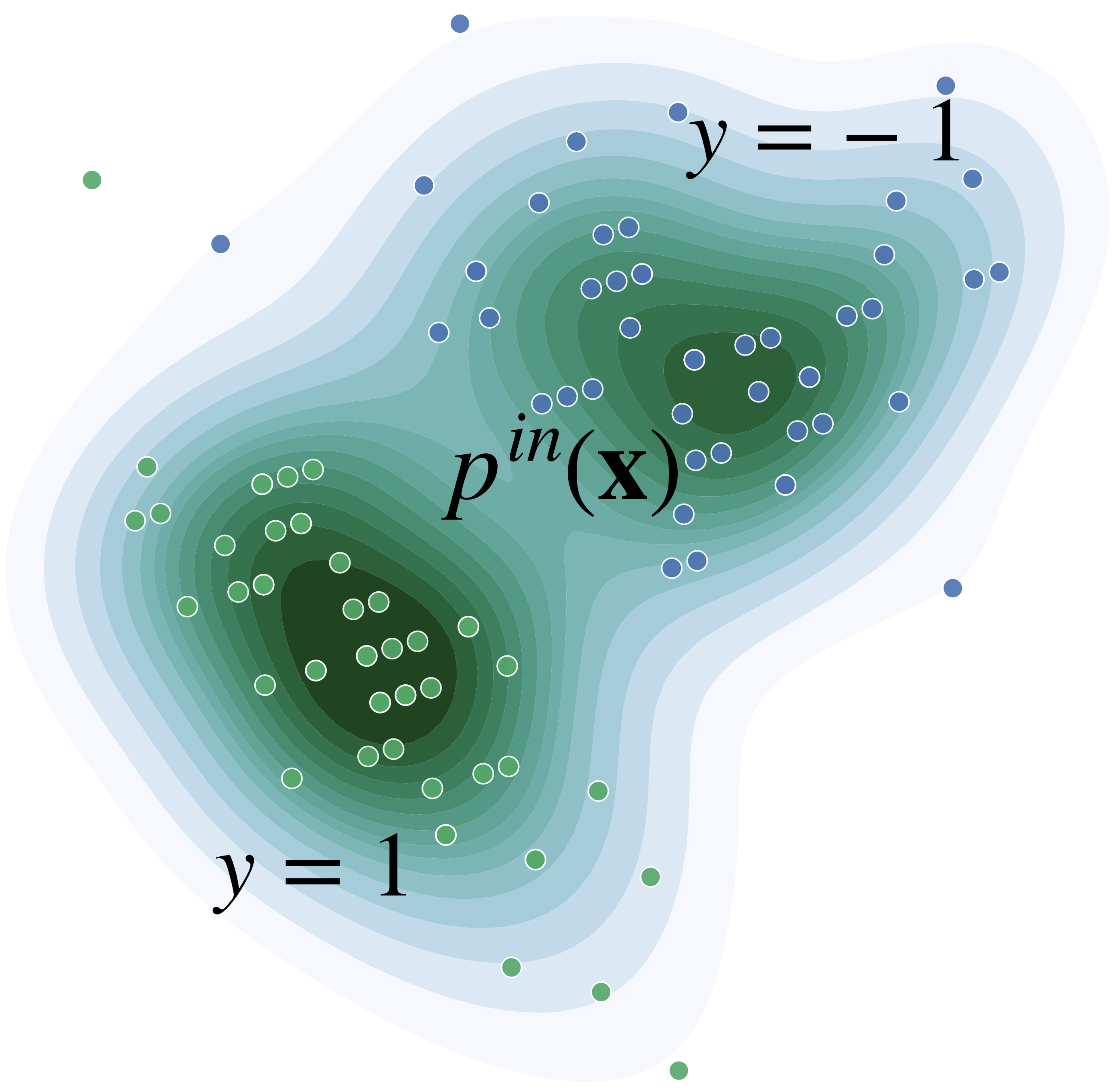
More money, more problems

By Andrew J. Hawkins | @andyjayhawk | Apr 22, 2022, 3:03pm EDT


Out-of-distribution Detection
Pedestrian
Car
Truck

Out-of-distribution Detection: A Simple View

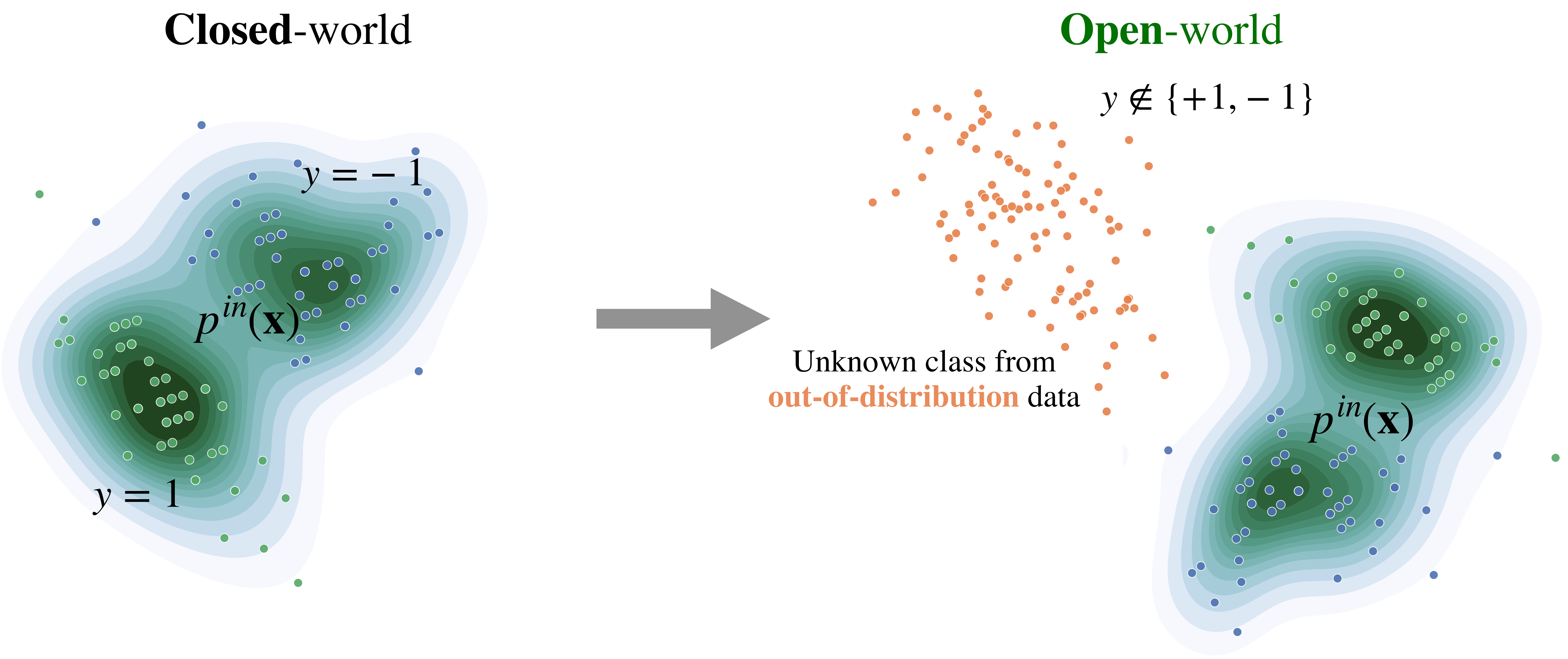
Closed-world



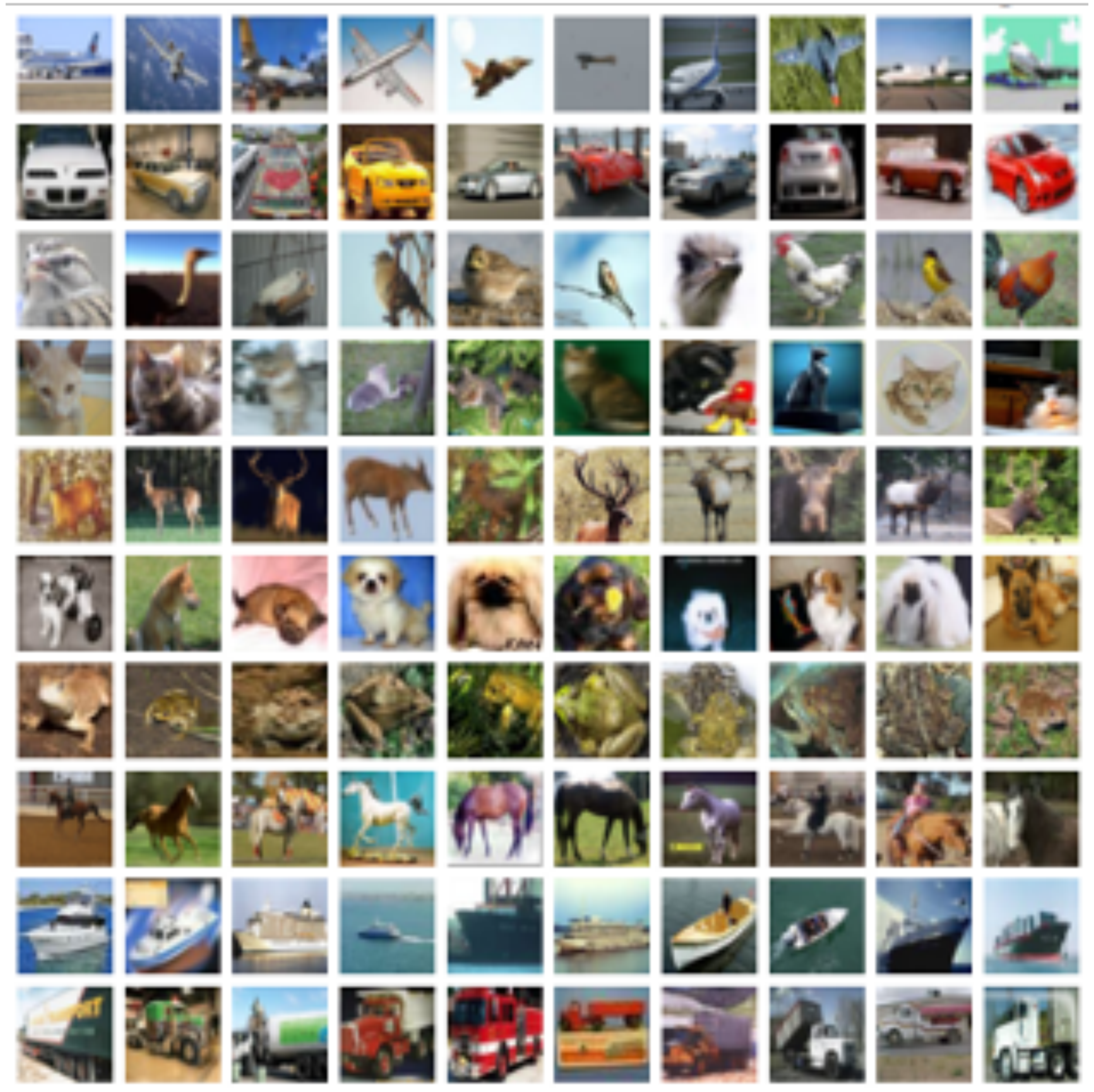
Input space: $\mathcal{X} = \mathbb{R}^d$

Label space: $\mathcal{Y} = \{1, -1\}$

Out-of-distribution Detection: A Simple View



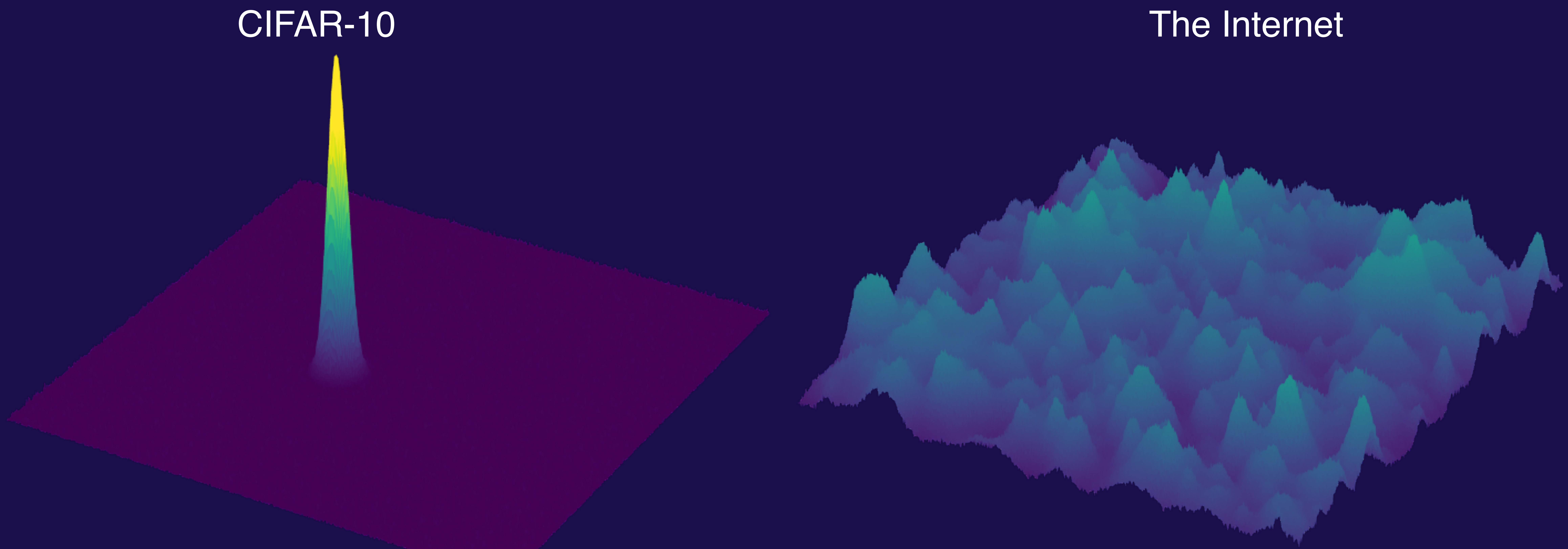
Out-of-distribution Detection



CIFAR-10 (in-distribution)

SVHN (OOD)

Out-of-distribution Detection

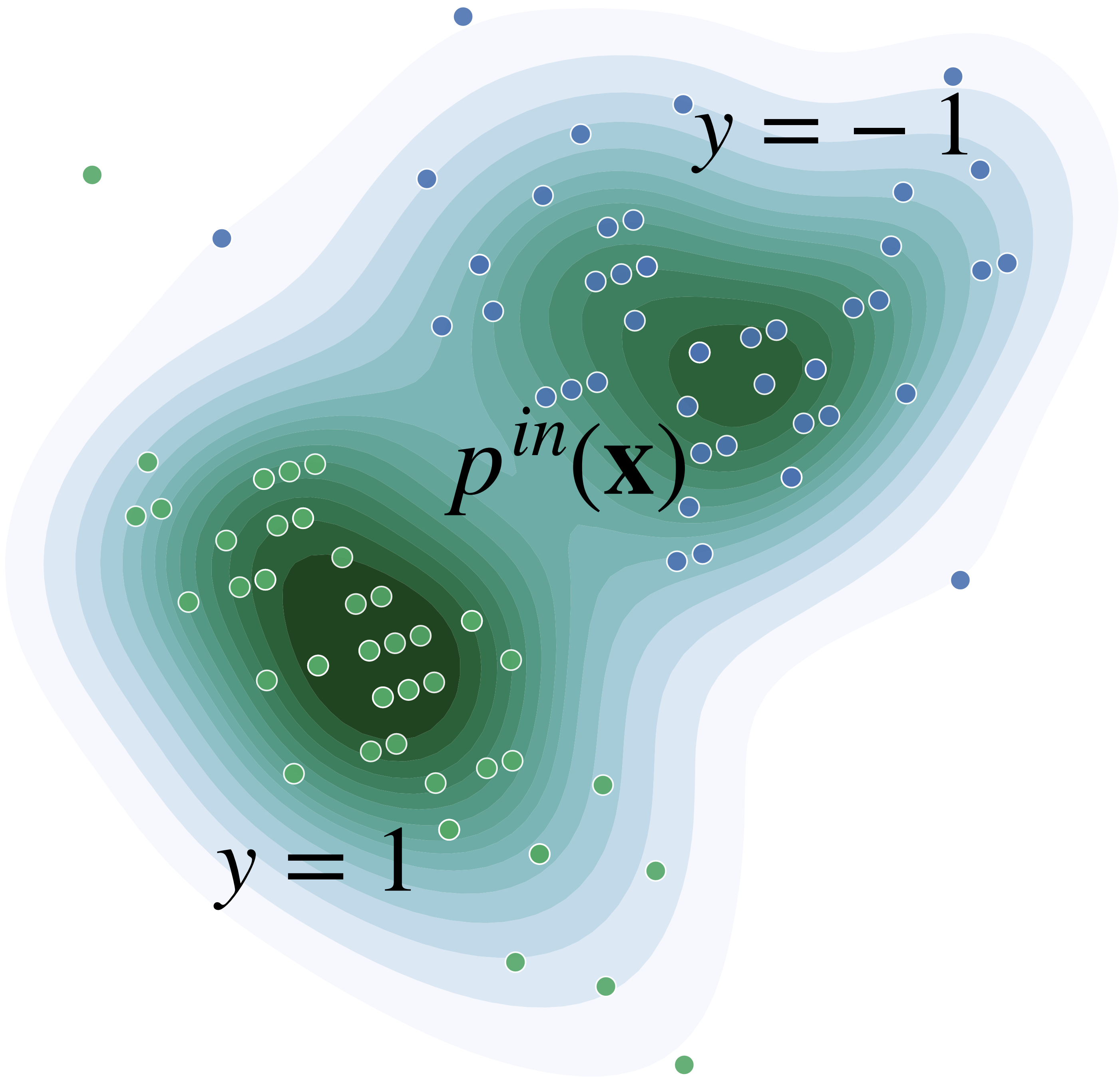


Slide from OpenAI

Out-of-distribution detection is a hard problem. Why?

Challenges

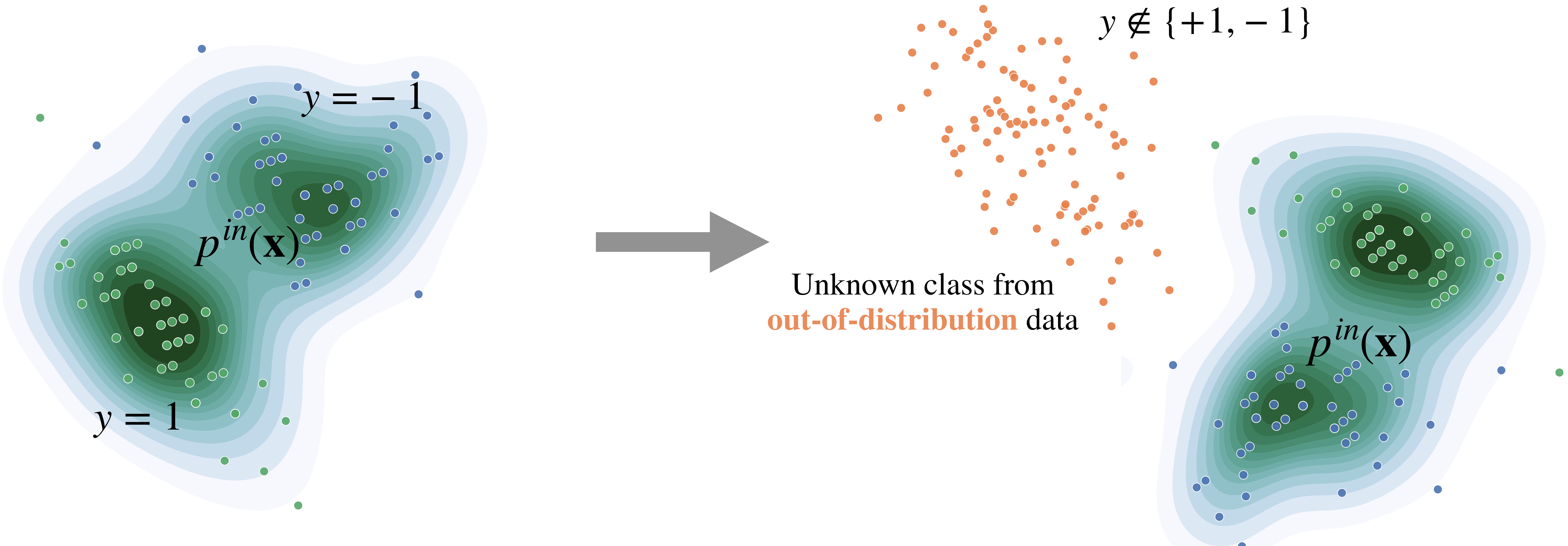
- Lack of supervision from unknowns during training
  (model is trained only on the green and blue dots, using empirical risk minimization)



Challenges

- Lack of supervision from unknowns during training
  (model is trained only on the green and blue dots, using empirical risk minimization)

- Huge space of unknowns in the high-dimensional space
  (hard to anticipate orange dots in advance)

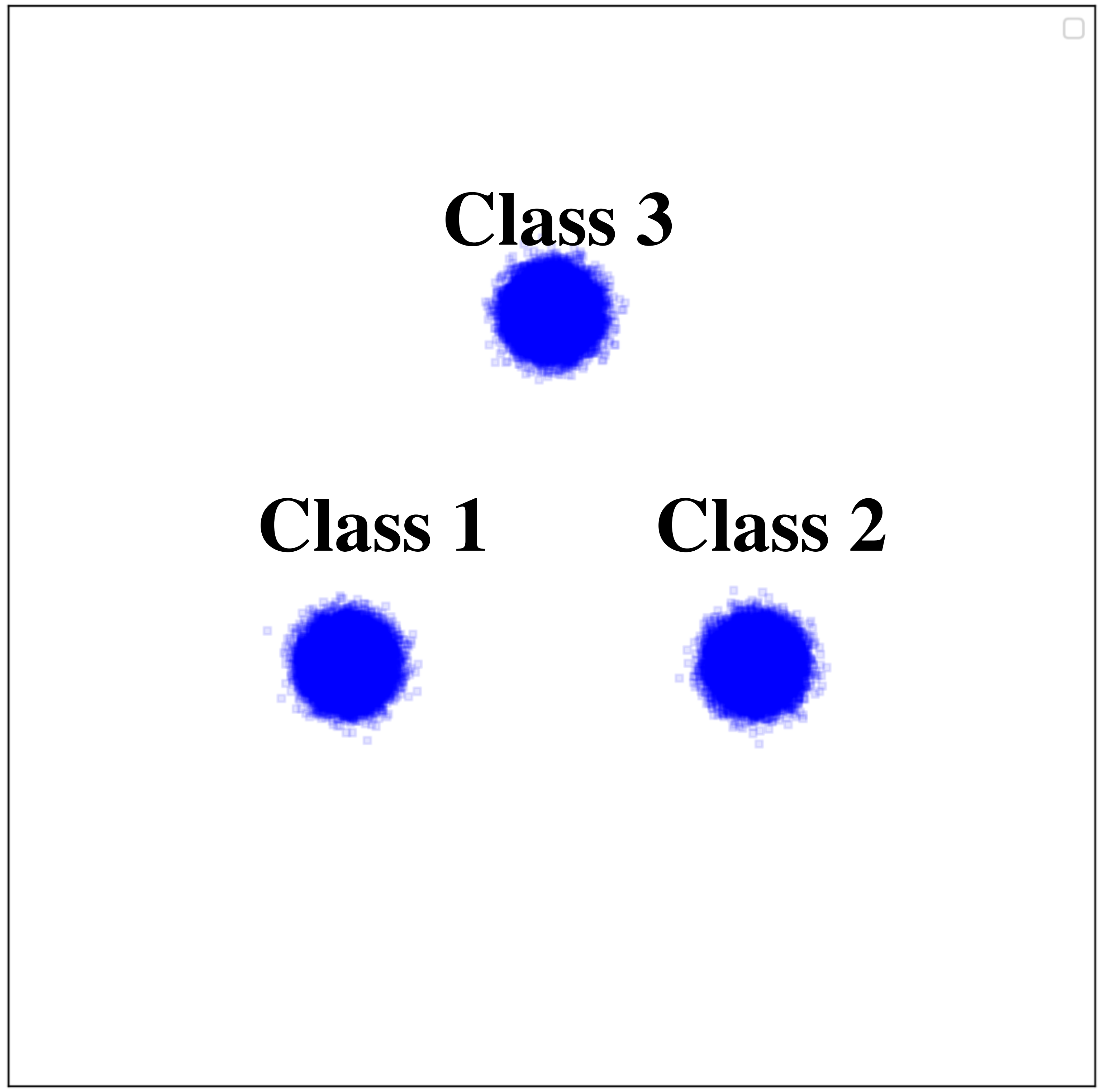


Challenges

- High-capacity neural networks exacerbate over-confident predictions
  (ill-fated decision boundary which cannot distinguish ID vs. OOD)

Challenges

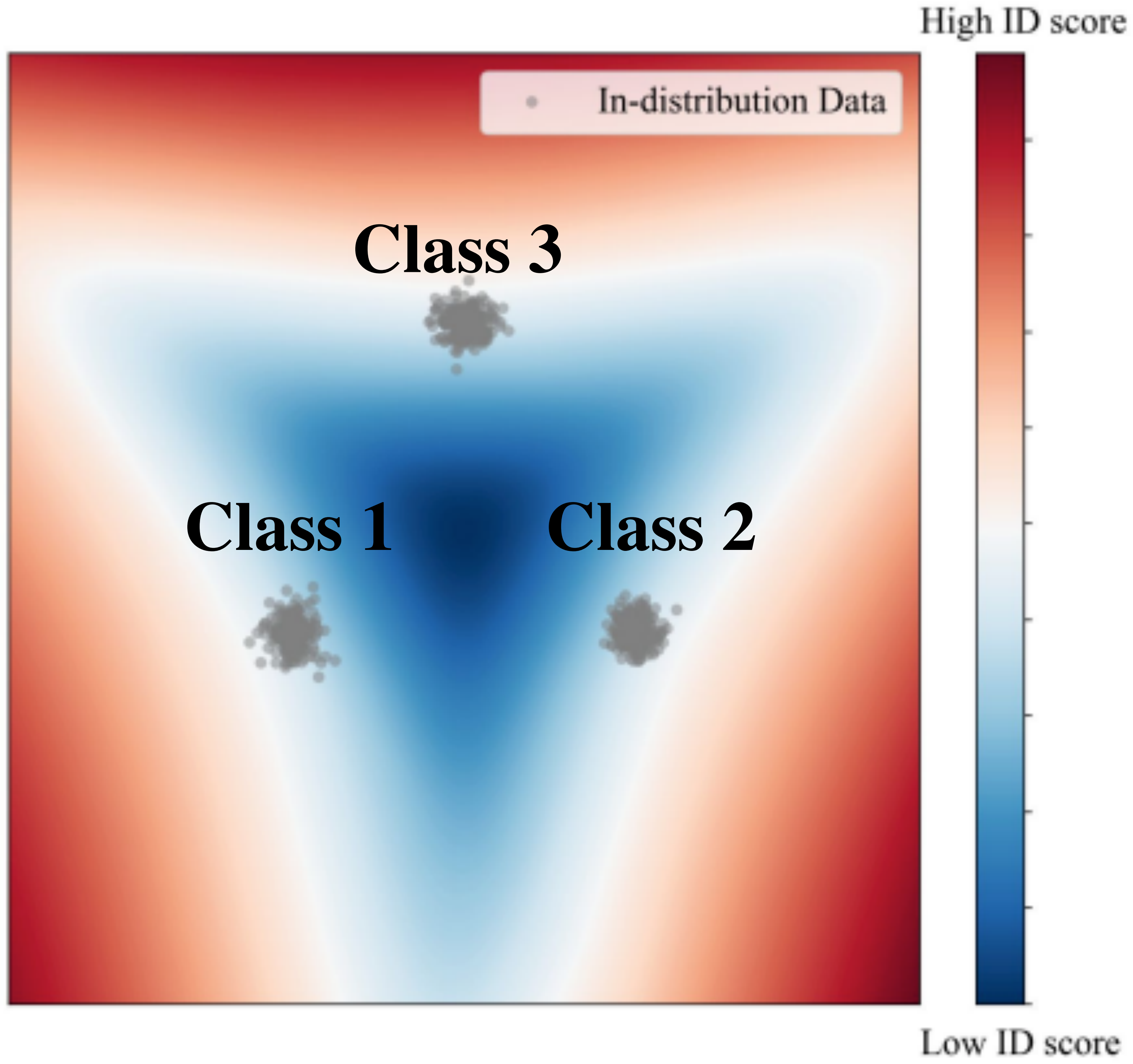
- High-capacity neural networks exacerbate over-confident predictions **(ill-fated decision boundary which cannot distinguish ID vs. OOD)**



In-distribution: mixture of 3 Gaussians

Challenges

- High-capacity neural networks exacerbate over-confident predictions **(ill-fated decision boundary which cannot distinguish ID vs. OOD)**



Decision boundary learned by a simple MLP
(Overconfident in red regions)

Challenges

- Real-world images are composed of numerous objects and components.

(Need finer-grained understanding of OOD at the **object-level,** not just image-level)

Thriving literature on OOD detection

arXiv > cs > arXiv:2110.11334

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 21 Oct 2021 (v1), last revised 3 Aug 2022 (this version, v2)]

Generalized Out-of-Distribution Detection: A Survey

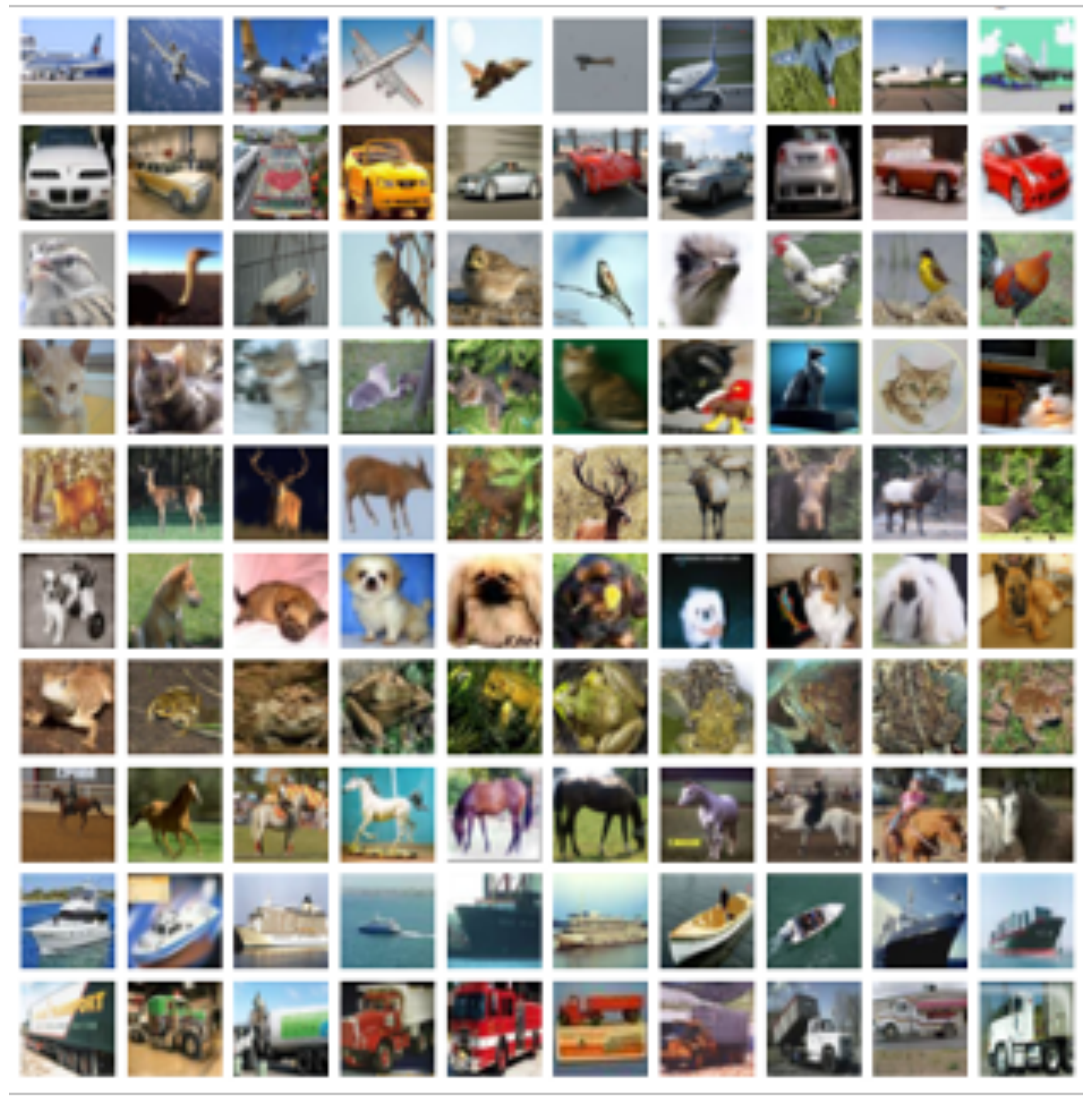
Jingkang Yang, Kaiyang Zhou, Yixuan Li, Ziwei Liu

§ 5 Out-of-Distribution Detection	§ 5.1 Classification	§ 5.1.1.a: Post-hoc Detection	[55], [187], [188], [189], [190], [191], [192]
		§ 5.1.1.b: Conf. Enhancement	[58], [192], [193], [194], [195], [196], [197], [198], [199], [200], [201], [202], [203], [204], [205], [206], [207], [208], [209]
		§ 5.1.1.c: Outlier Exposure (OE)	[52], [210], [211], [212], [213], [214], [215], [216], [217], [218], [219]
		§ 5.1.2: OOD Data Generation	[220], [221], [222], [223]
		§ 5.1.3: Gradient-based Methods	[188], [191]
		§ 5.1.4: Bayesian Models	[224], [225], [226], [227], [228], [229]
		§ 5.1.5: Large-scale OOD Detection	[168], [171], [230], [231]
	§ 5.2: Density-based Methods		[87], [88], [89], [90], [92], [121], [207], [232], [233], [234], [235], [236], [237], [238], [239], [240]
	§ 5.3: Distance-based Methods		[207], [241], [242], [243], [244], [245], [246]

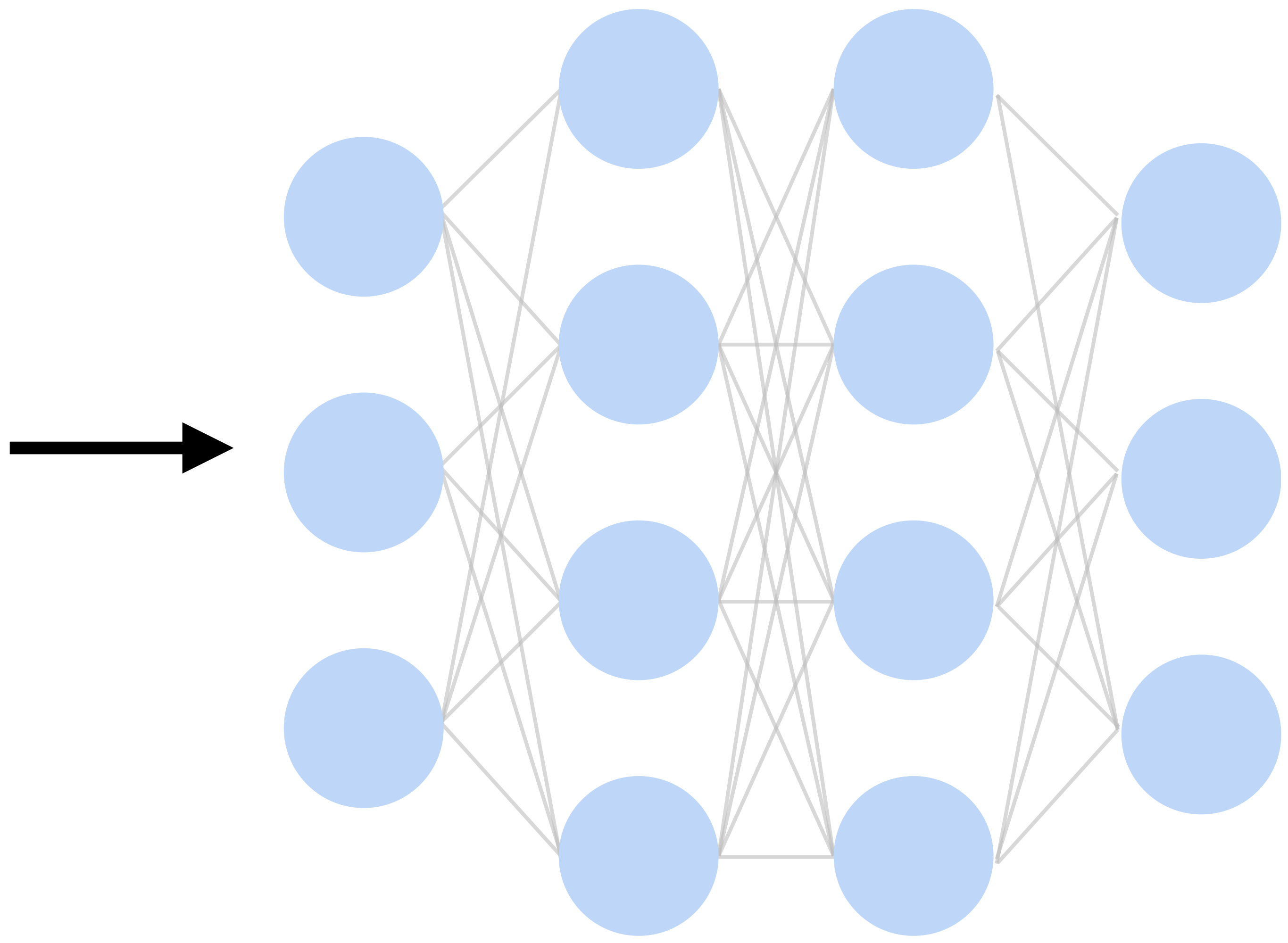
Tutorial Outline

- **Inference-time OOD detection**
  - Output-based methods
  - Distance-based methods
- **Training-time regularization for OOD detection**
  - Safety-aware learning objective
  - Synthesizing virtual outliers
  - Leveraging wild unlabeled data

Inference-Time Out-of-distribution Detection
Method Overview



$f(x;\theta)$



Empirical risk minimization:

$$R_{\text{closed}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i)$$

$$f^* = \operatorname{argmin}_{f \in \mathcal{F}} R_{\text{closed}}(f)$$

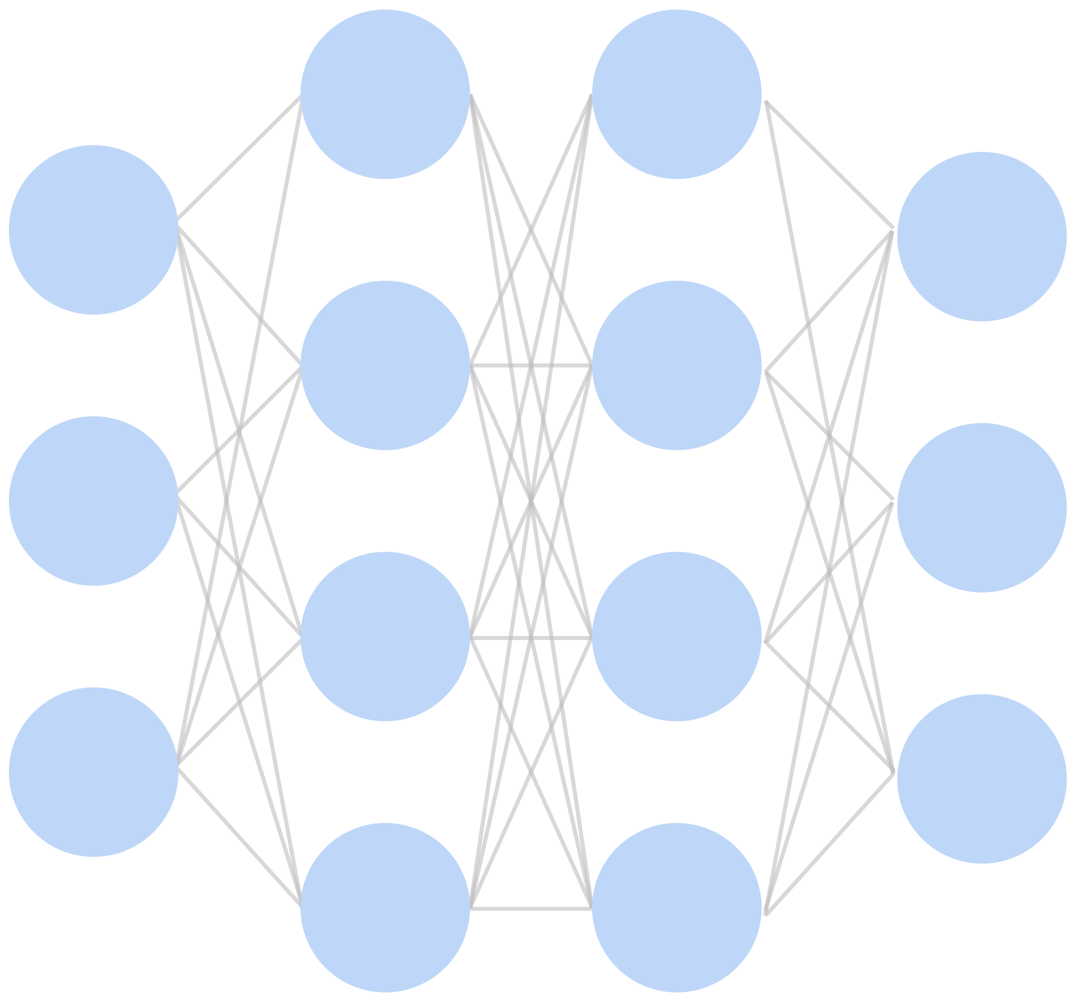
Trained on in-distribution data
(e.g., CIFAR-10), freeze parameters

V. Vapnik. *Principles of risk minimization for learning theory. NIPS 1991*

Out-of-distribution Detection Method Overview

$f(x;\theta)$

Test input x $\longrightarrow$



$\longrightarrow$

$$G_\lambda(\mathbf{x}; f) = \begin{cases} \text{in} & S(\mathbf{x}; f) \geq \lambda \\ \text{out} & S(\mathbf{x}; f) < \lambda \end{cases}$$

S: Scoring function

Trained on in-distribution data
(e.g., CIFAR-10), freeze parameters

How to define OOD scoring function?

Published as a conference paper at ICLR 2017

A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks

Dan Hendrycks*
University of California, Berkeley
hendrycks@berkeley.edu

Published as a conference paper at ICLR 2018

Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks

Shiyu Liang
Coordinated Science Lab, Department of ECE
University of Illinois at Urbana-Champaign
sliang26@illinois.edu

Yixuan Li
University of Wisconsin-Madison*
sharonli@cs.wisc.edu

R. Srikant
Coordinated Science Lab, Department of ECE
University of Illinois at Urbana-Champaign
rsrikant@illinois.edu

A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks

Kimin Lee[1], **Kibok Lee**[2], **Honglak Lee**[3,2], **Jinwoo Shin**[1,4]
[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]University of Michigan
[3]Google Brain
[4]AItrics

Out-of-Distribution Detection with Deep Nearest Neighbors

Yiyou Sun [1] **Yifei Ming** [1] **Xiaojin Zhu** [1] **Yixuan Li** [1]

Abstract

Out-of-distribution (OOD) detection is a critical task for deploying machine learning models in the open world. Distance-based methods have demonstrated promise, where testing samples are detected as OOD if they are relatively far away

A rich line of OOD detection algorithms has been developed recently, among which distance-based methods demonstrated promise (Lee et al., 2018; Tack et al., 2020; Sehwag et al., 2021). Distance-based methods leverage feature embeddings extracted from a model, and operate under the assumption that the test OOD samples are relatively far

Energy-based Out-of-distribution Detection

Weitang Liu
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093, USA
wel022@ucsd.edu

Xiaoyun Wang
Department of Computer Science
University of California, Davis
Davis, CA 95616, USA
xiywang@ucdavis.edu

John D. Owens
Department of Electrical and Computer Engineering
University of California, Davis
Davis, CA 95616, USA
jowens@ece.ucdavis.edu

Yixuan Li
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53703, USA
sharonli@cs.wisc.edu

Tutorial Outline

- **Inference-time OOD detection**
  - Output-based methods
  - Distance-based methods
- **Training-time regularization for OOD detection**
  - Safety-aware learning objective
  - Synthesizing virtual outliers
  - Leveraging wild unlabeled data

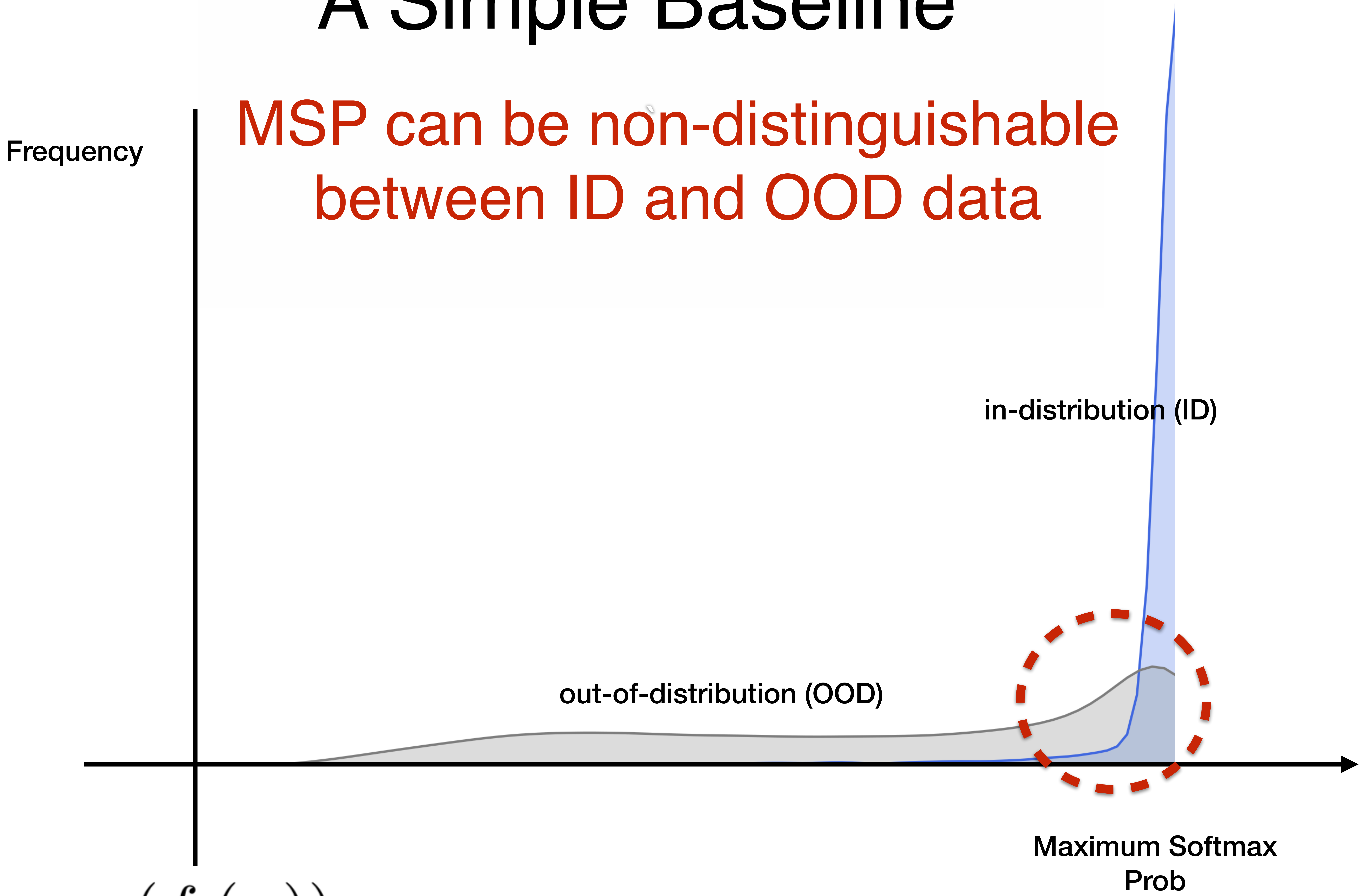
A Simple Baseline

$$S(\mathbf{x};f) = \max_i \frac{\exp\left(f_i(\boldsymbol{x})\right)}{\sum_{j=1}^{N} \exp\left(f_j(\boldsymbol{x})\right)},$$

Hendycks et al., *A Baseline for Detecting Misclassified and Out-of-distribution Samples in Neural Networks. ICLR 2017*

A Simple Baseline

MSP can be non-distinguishable between ID and OOD data

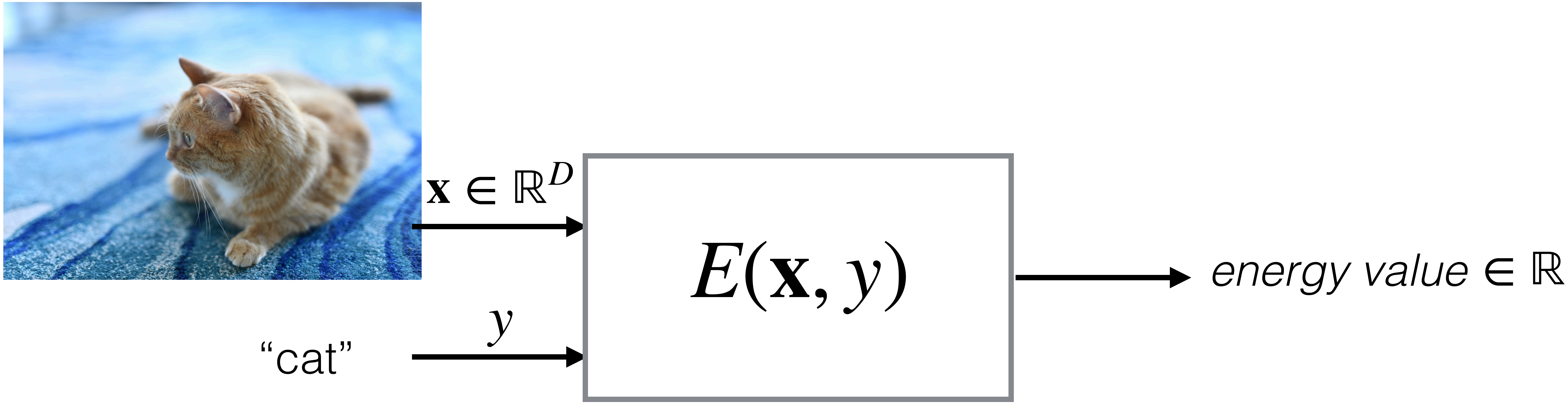


Energy-based Out-of-distribution Detection

$$g(\mathbf{x}; \tau, f) = \begin{cases} 0 & \text{if } -E(\mathbf{x}; f) \leq \tau, \\ 1 & \text{if } -E(\mathbf{x}; f) > \tau, \end{cases}$$

Liu et al., *Energy-based Out-of-distribution Detection*, NeurIPS 2020

Energy-based Model



Energy can be turned into **probability** through Gibbs distribution:

$$p(y \mid \mathbf{x}) = \frac{e^{-E(\mathbf{x},y)/T}}{\int_{y'} e^{-E(\mathbf{x},y')/T}} = \frac{e^{-E(\mathbf{x},y)/T}}{e^{-E(\mathbf{x})/T}}$$

Energy-based Model

Energy can be turned into probability through Gibbs distribution:

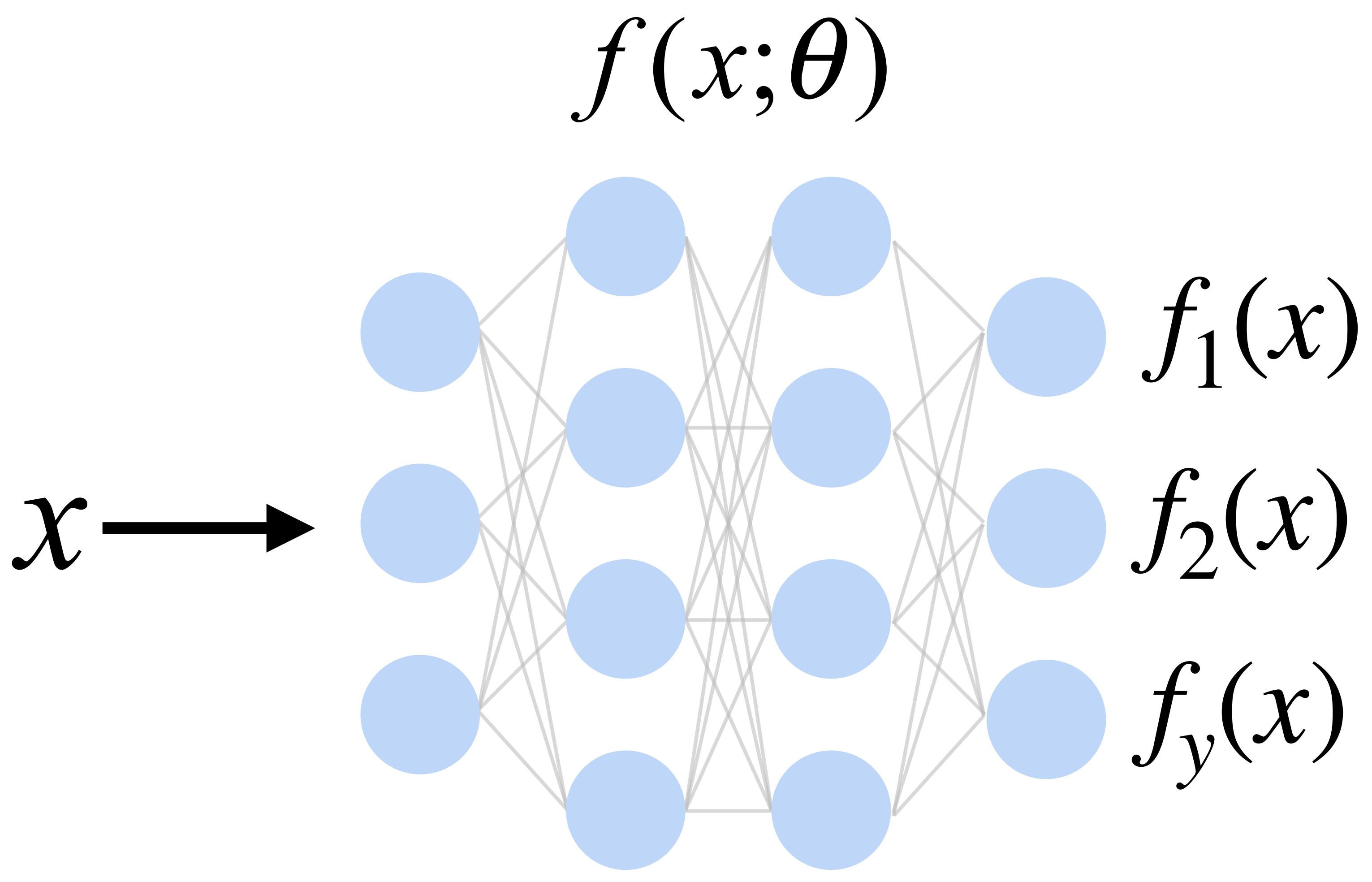
$$p(y \mid \mathbf{x}) = \frac{e^{-E(\mathbf{x},y)/T}}{\int_{y'} e^{-E(\mathbf{x},y')/T}} = \frac{e^{-E(\mathbf{x},y)/T}}{e^{-E(\mathbf{x})/T}}$$

Partition function

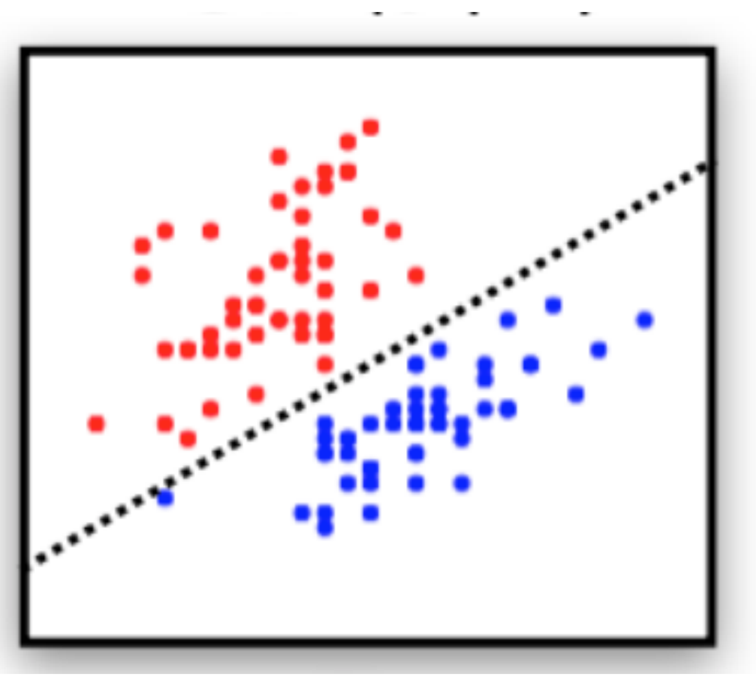
Free energy can be expressed as the negative of the log partition function:

$$E(\mathbf{x}) = -T \cdot \log \int_{y'} e^{-E(\mathbf{x},y')/T}$$

Energy-based Interpretation of Classification Model



softmax

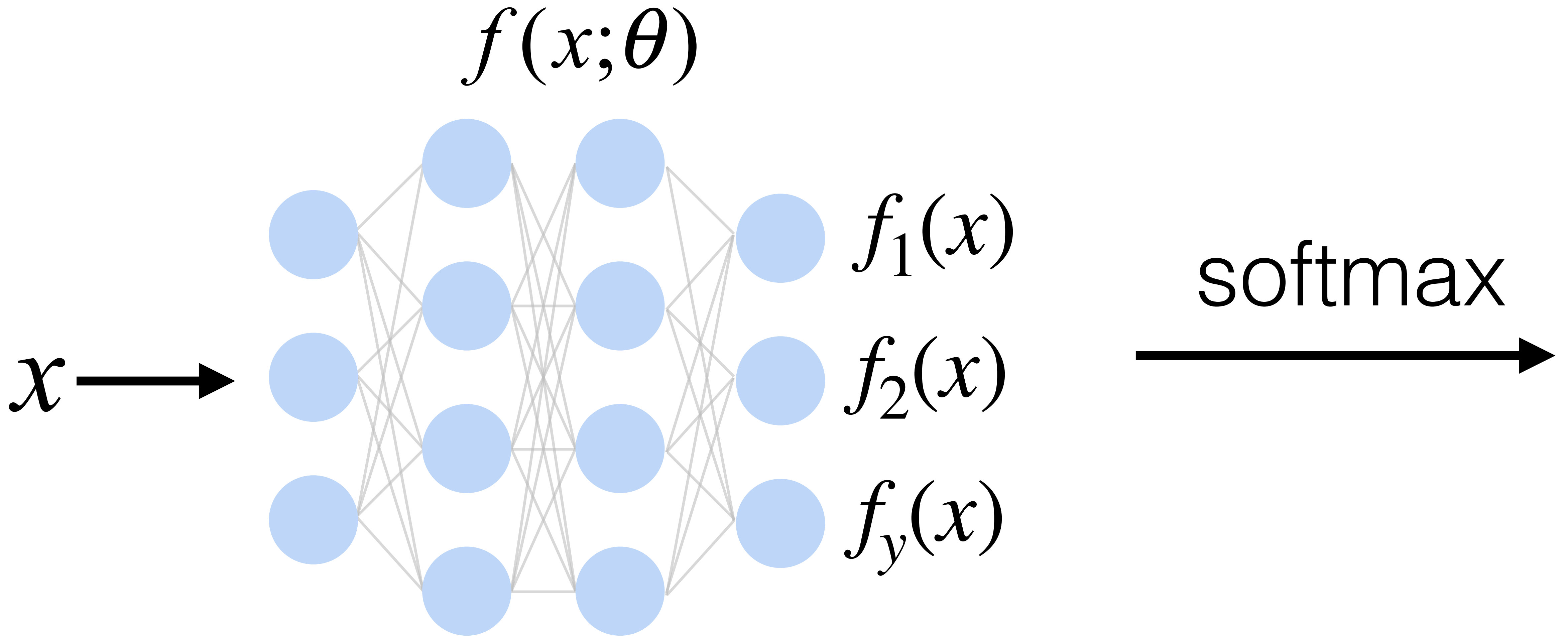


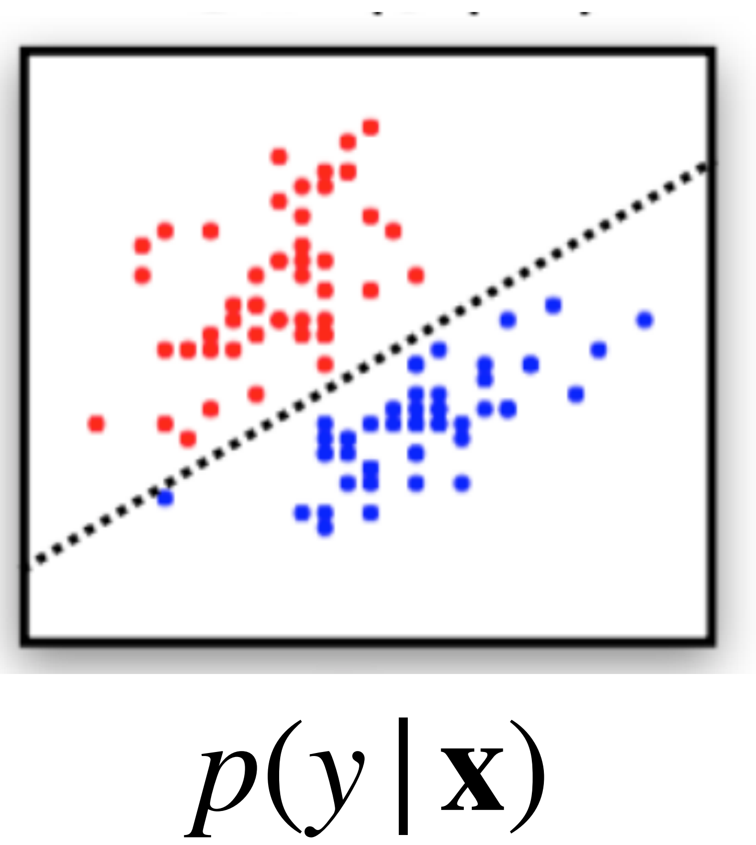
$p(y \mid \mathbf{x})$

$$E(\mathbf{x}, y) = -f_y(\mathbf{x})$$

$$p(y \mid \mathbf{x}) = \frac{e^{f_y(\mathbf{x})/T}}{\sum_i^K e^{f_i(\mathbf{x})/T}} \quad \Longleftrightarrow \quad p(y \mid \mathbf{x}) = \frac{e^{-E(\mathbf{x},y)/T}}{\int_{y'} e^{-E(\mathbf{x},y')/T}}$$

Energy-based Interpretation of Classification Model





Free energy can be expressed as the negative of the **LogSumExp**:

$$E(\mathbf{x}) = -T \cdot \log \int_{y'} e^{-E(\mathbf{x},y')/T}$$

input

Neural nets

Energy-based Out-of-distribution Detection

$$g(\mathbf{x};\tau,f)=\begin{cases}0 & \text{if } -E(\mathbf{x};f)\le\tau,\\ 1 & \text{if } -E(\mathbf{x};f)>\tau,\end{cases}$$

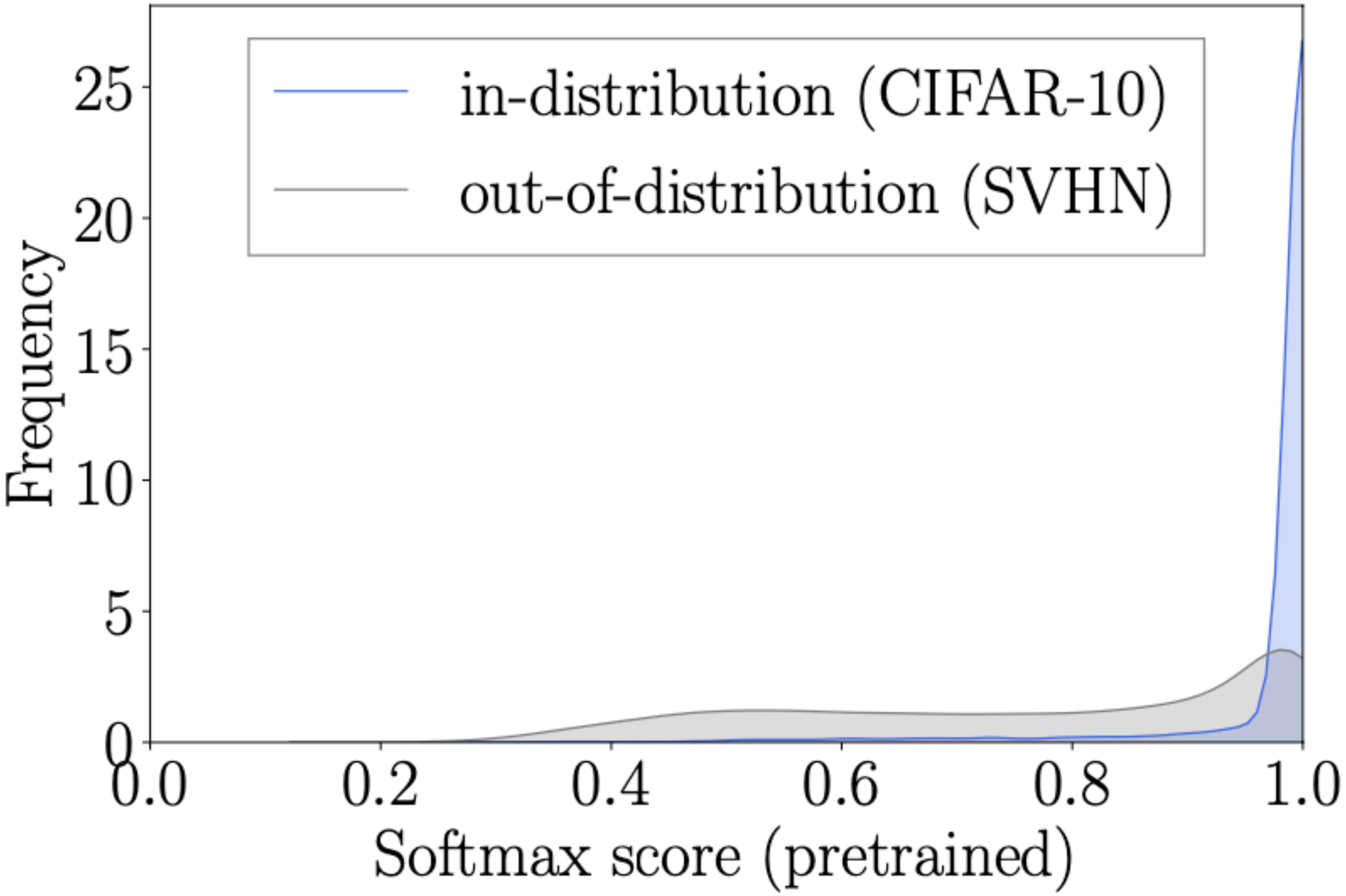
Softmax vs. energy scores

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})} = \frac{e^{f_y(\mathbf{x};\theta)}}{\sum_{k=1}^{K} e^{f_k(\mathbf{x};\theta)}}$$

$$E(\mathbf{x};\theta) := -\log \sum_{k=1}^{K} e^{f_k(\mathbf{x};\theta)}$$

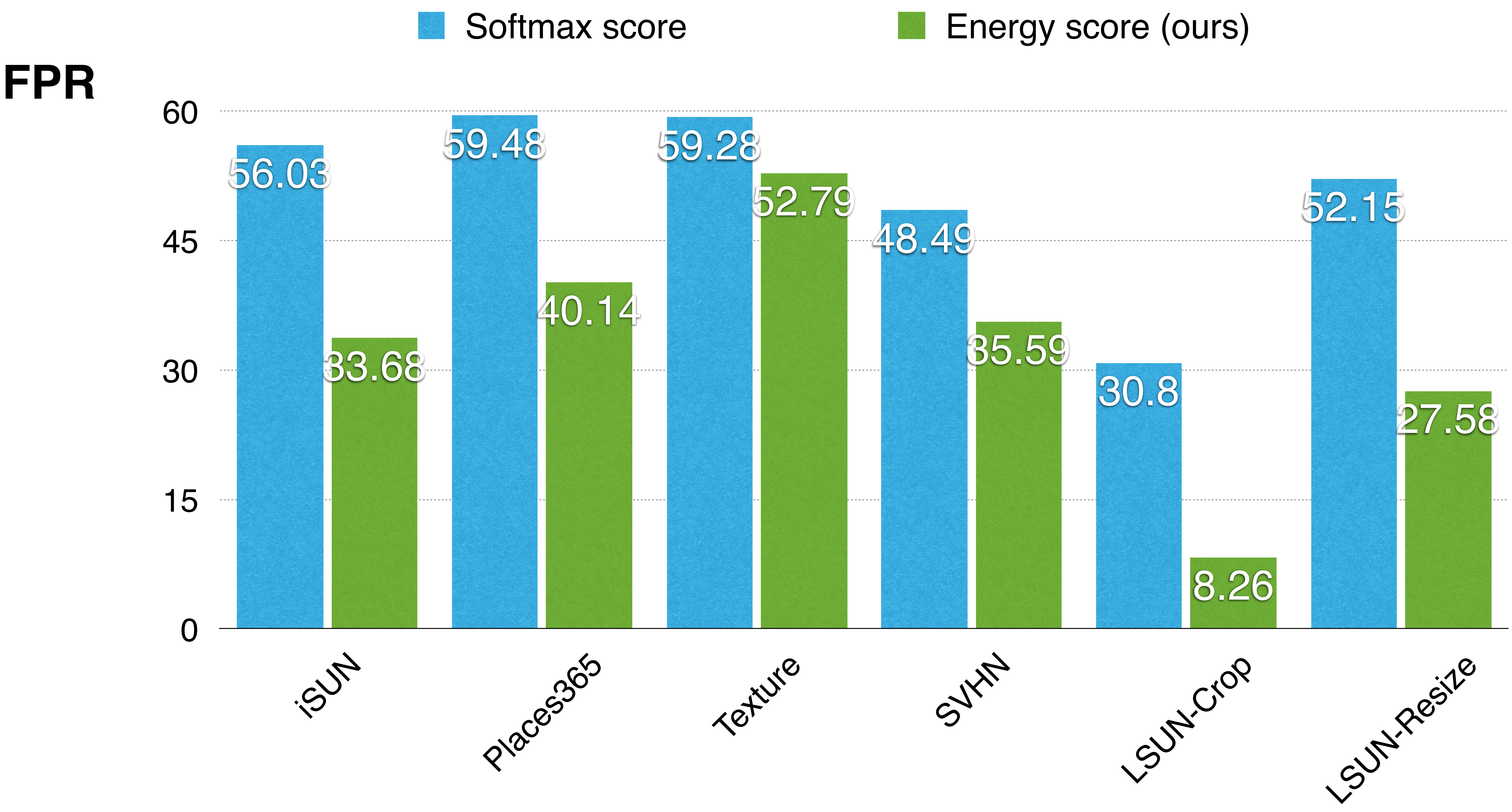
Energy function has an inherent connection to the log likelihood (but not identical — will come back to this).

Softmax vs. energy scores



(a) **FPR95: 48.87**

More Results

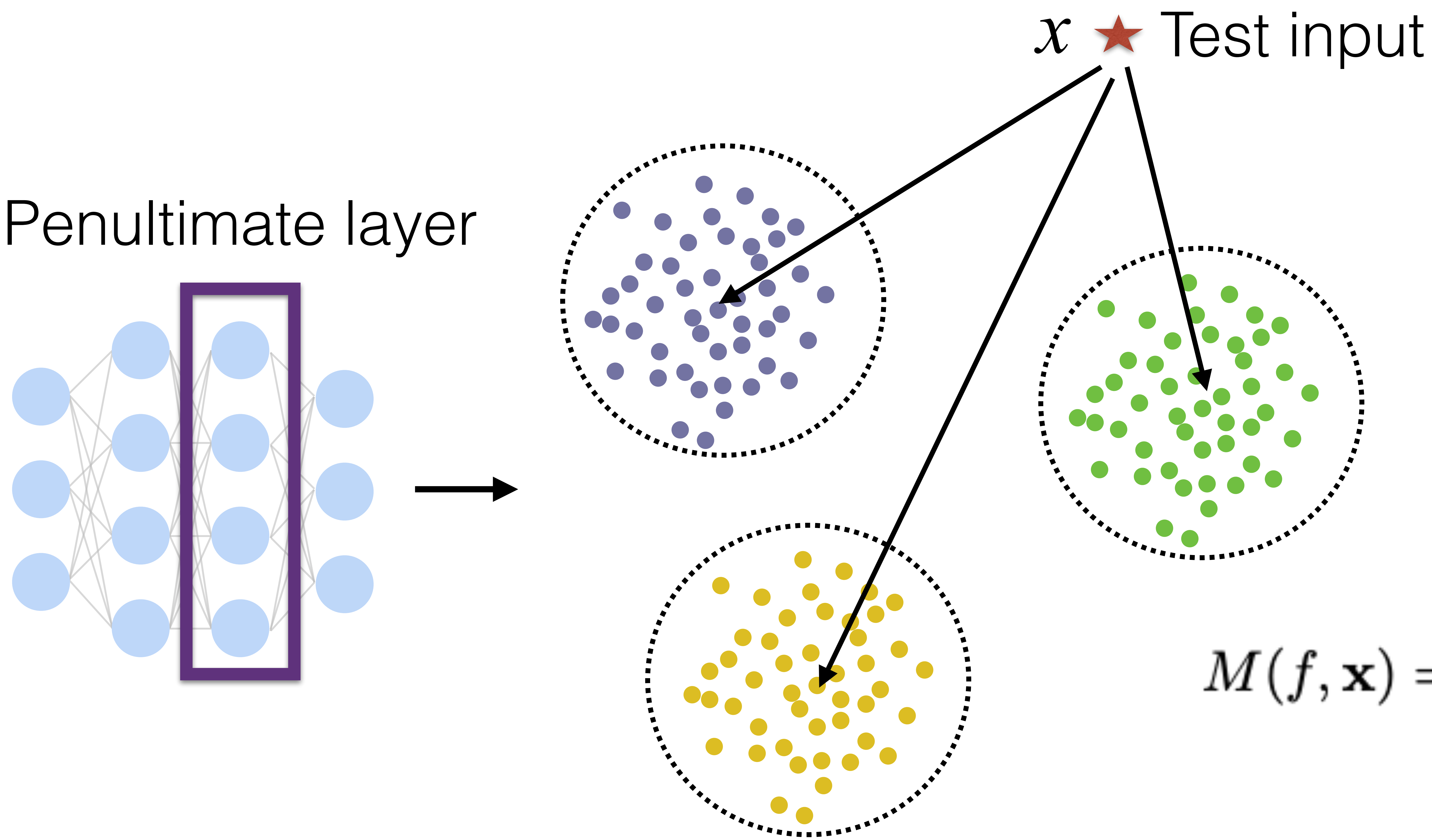


Tutorial Outline

- **Inference-time OOD detection**
  - Output-based methods
  - Distance-based methods
- **Training-time regularization for OOD detection**
  - Safety-aware learning objective
  - Synthesizing virtual outliers
  - Leveraging wild unlabeled data

Mahalanobis distance (parametric)

Idea: Model the feature space as a mixture of multivariate Gaussian distribution, one for each class. Use distance to the closest centroid as a proxy for OOD measure.



$$M(f, \mathbf{x}) = \max_i -(\mathbf{x} - \boldsymbol{\mu}_i)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$$

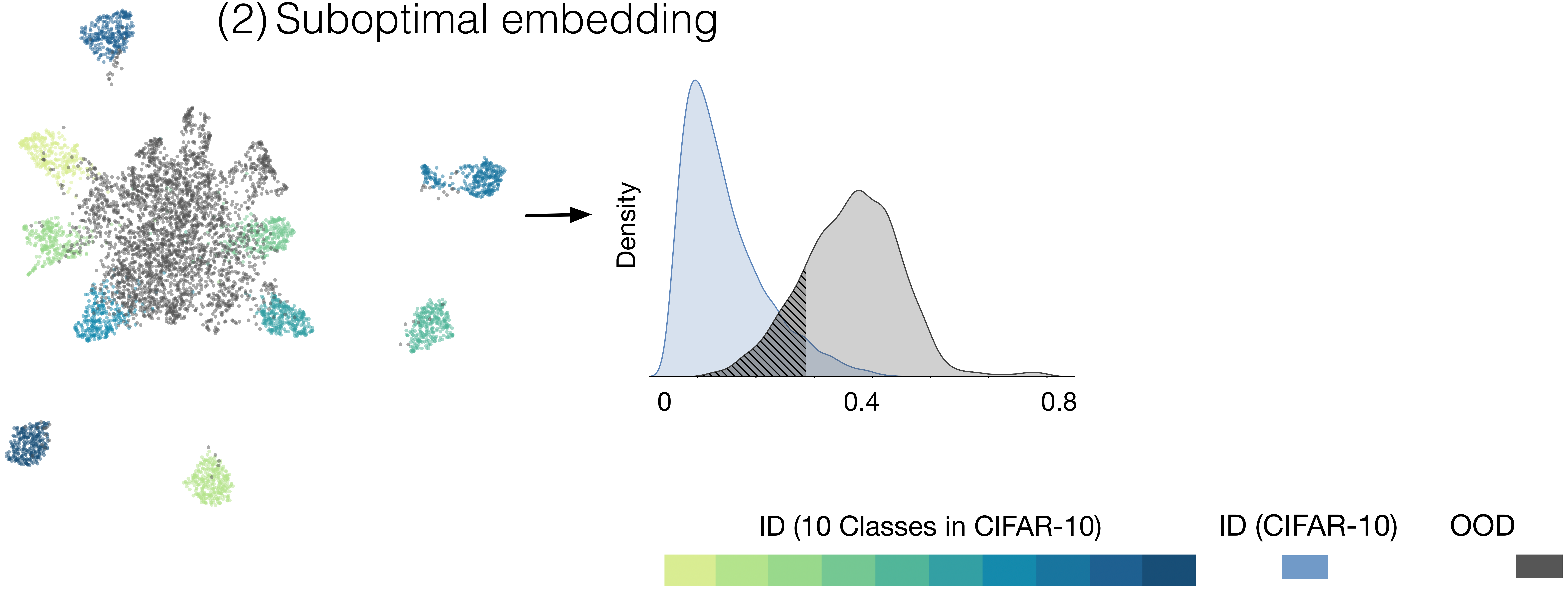
Lee et al., *A simple unified framework for detecting out-of-distribution samples and adversarial attacks. NeurIPS 2018*

Mahalanobis distance (parametric)

Limitations:

(1) strong distributional assumption (features may not necessarily be Gaussian-distributed)

(2) Suboptimal embedding

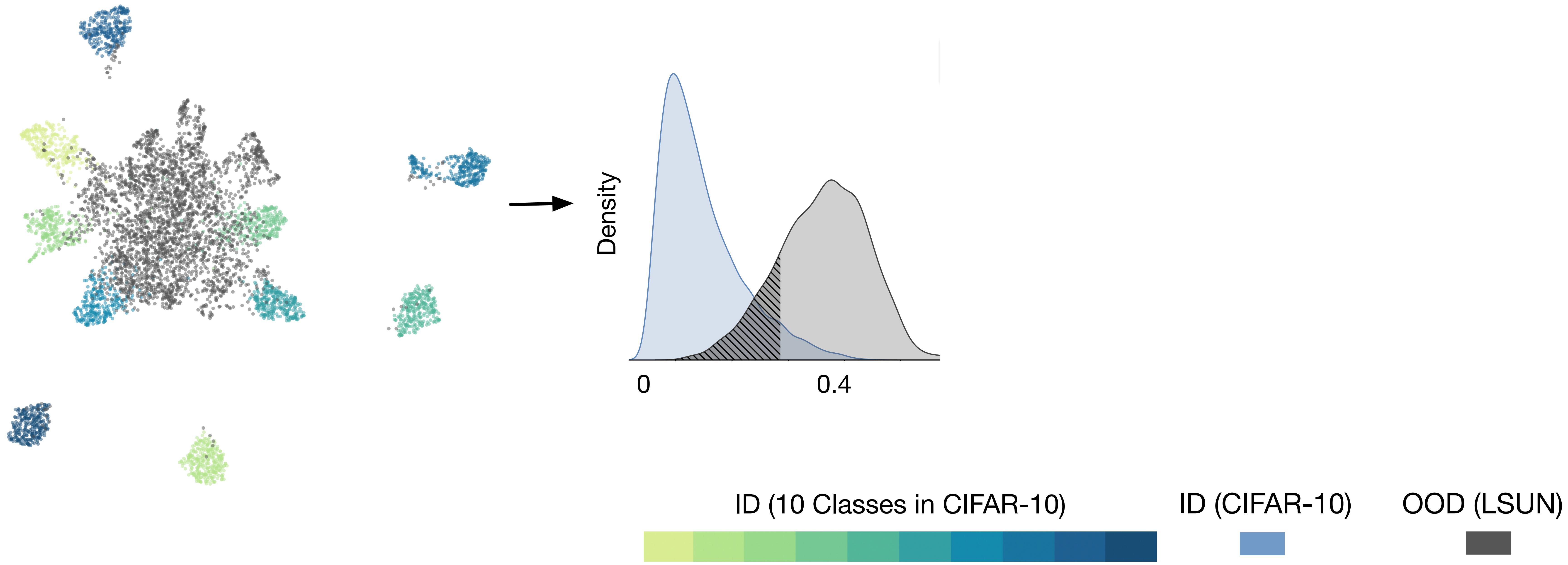


Nearest Neighbor Distance (non-parametric)

Limitations of Maha distance:

(1) strong distributional assumption

(2) Suboptimal embedding

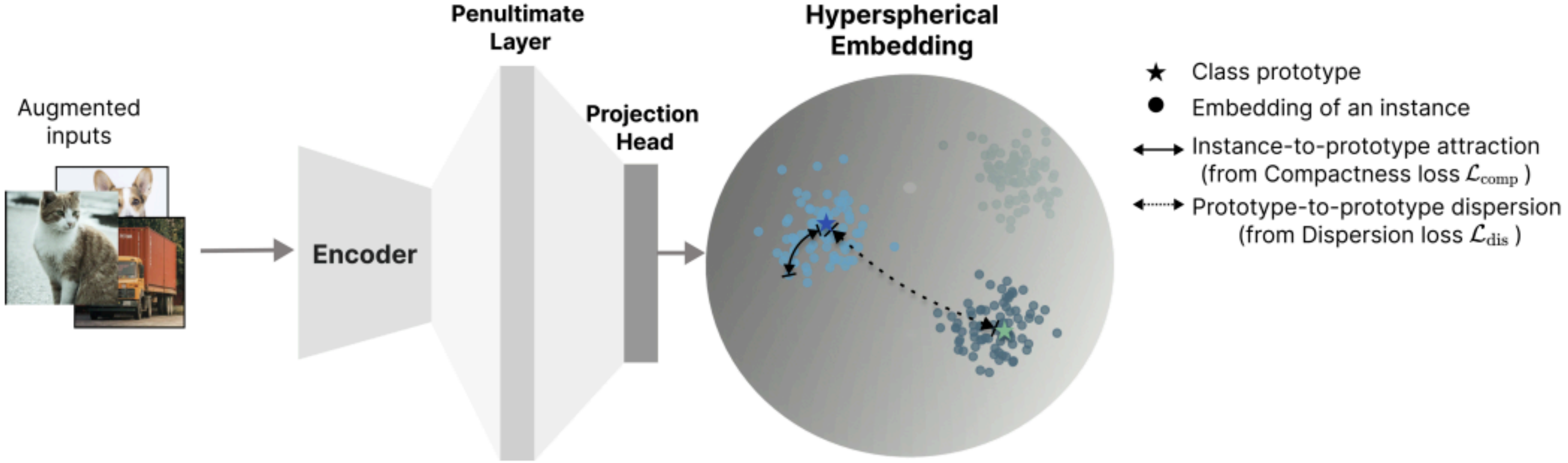


Sun et al., *Out-of-distribution Detection with Deep Nearest Neighbors*, ICML 2022

CIDER

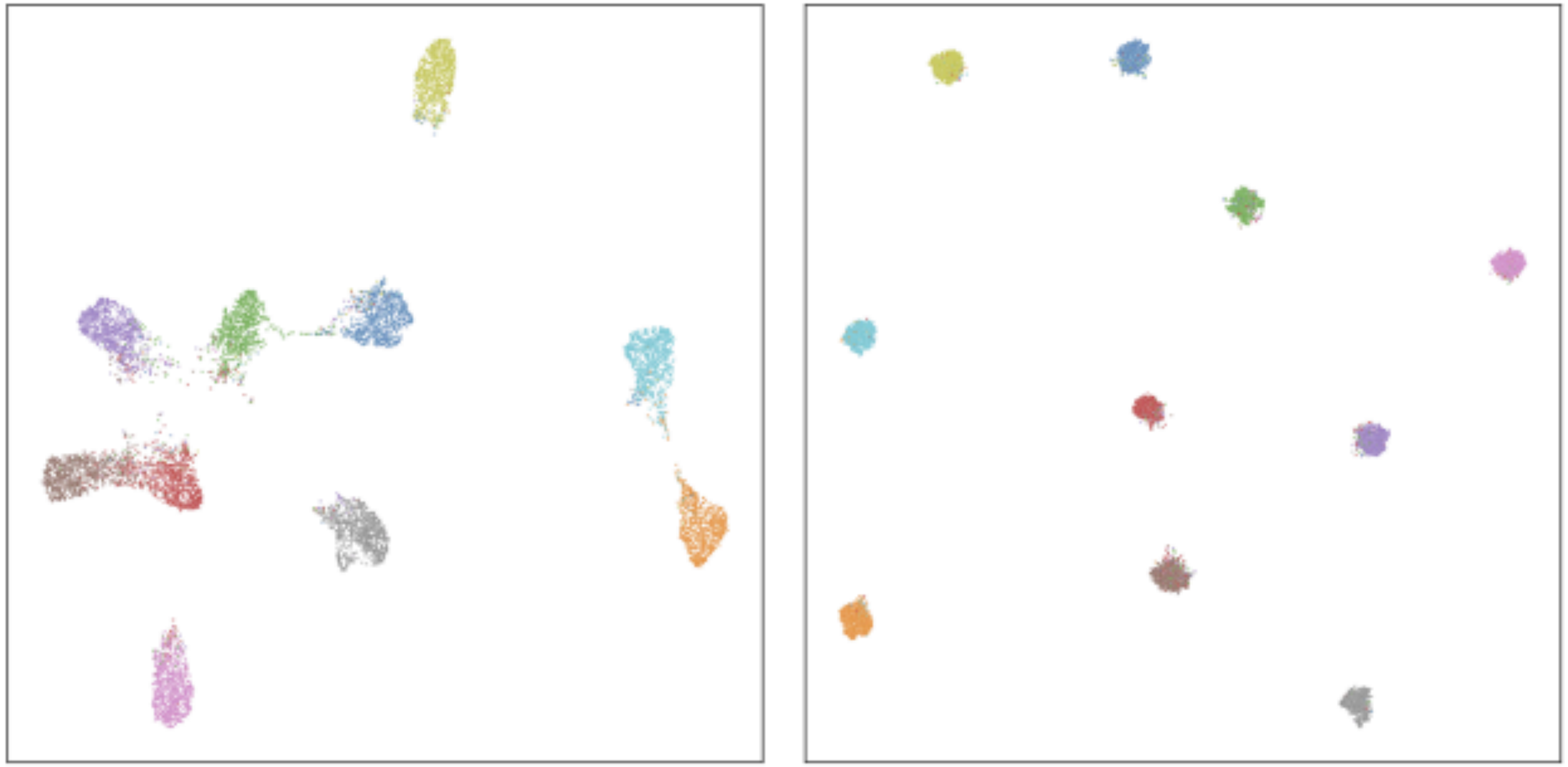
Learning optimal hyper-spherical embeddings for OOD detection

$$p_d(\mathbf{z}; \boldsymbol{\mu}_c, \kappa) = Z_d(\kappa) \exp\left(\kappa \boldsymbol{\mu}_c^\top \mathbf{z}\right),$$

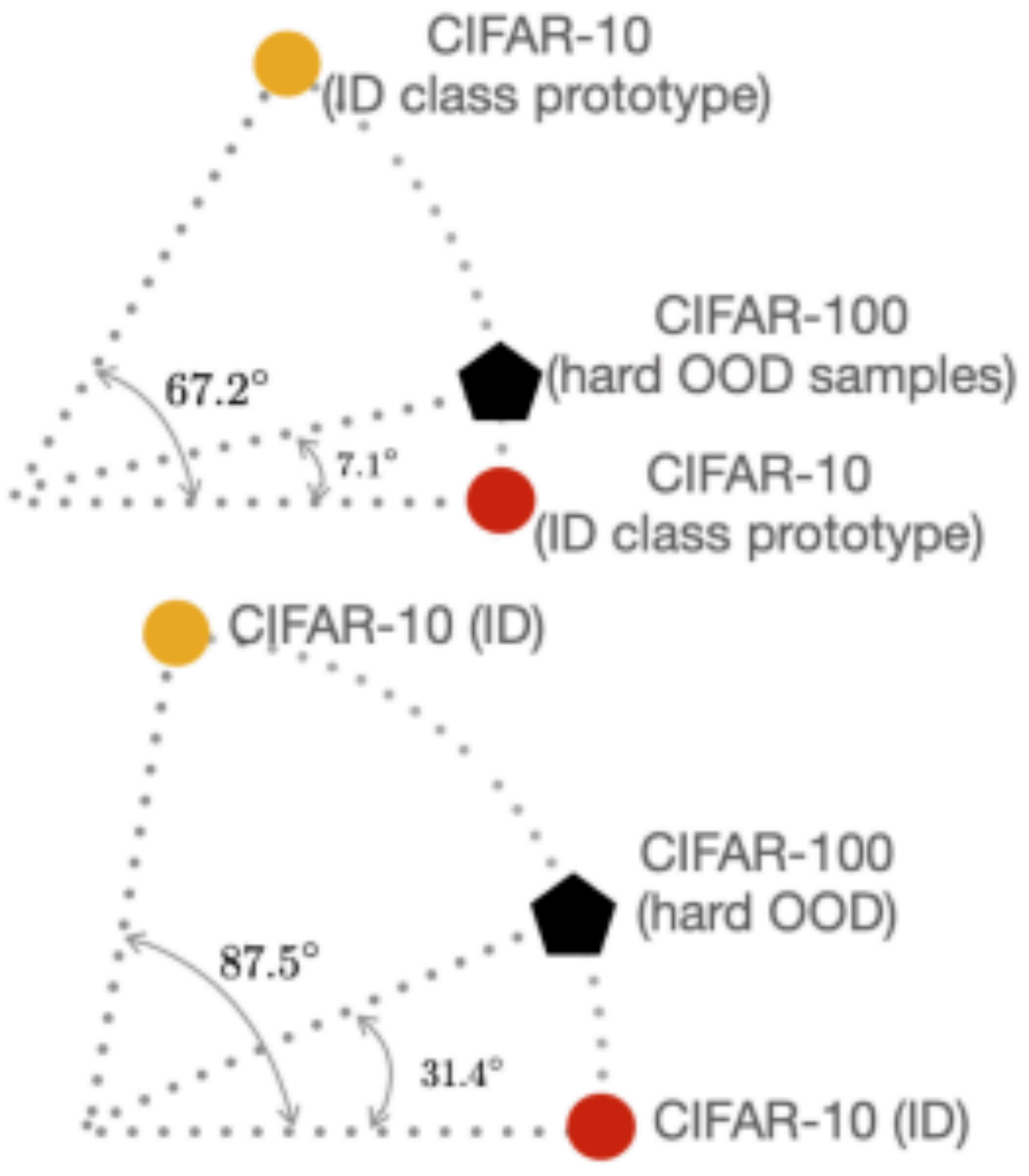


Ming et al., *How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? ICLR* 2023

CIDER



(a) ID embeddings of `CE` (left) and `CIDER` (right)



(b) ID & OOD of `CE` (top) and `CIDER` (down)

Method	**OOD Dataset** SVHN	Places365	LSUN	iSUN	Texture	**AVG FPR95**
SupCon+KNN (KNN+)	39.23	80.74	48.99	74.99	57.15	60.22
CIDER+KNN	**23.09**	**79.63**	**16.16**	**71.68**	**43.87**	**46.89**

Ming et al., *How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? ICLR* 2023

Scoring function is only part of the solution…

Mitigating OOD Risk Requires Rethinking

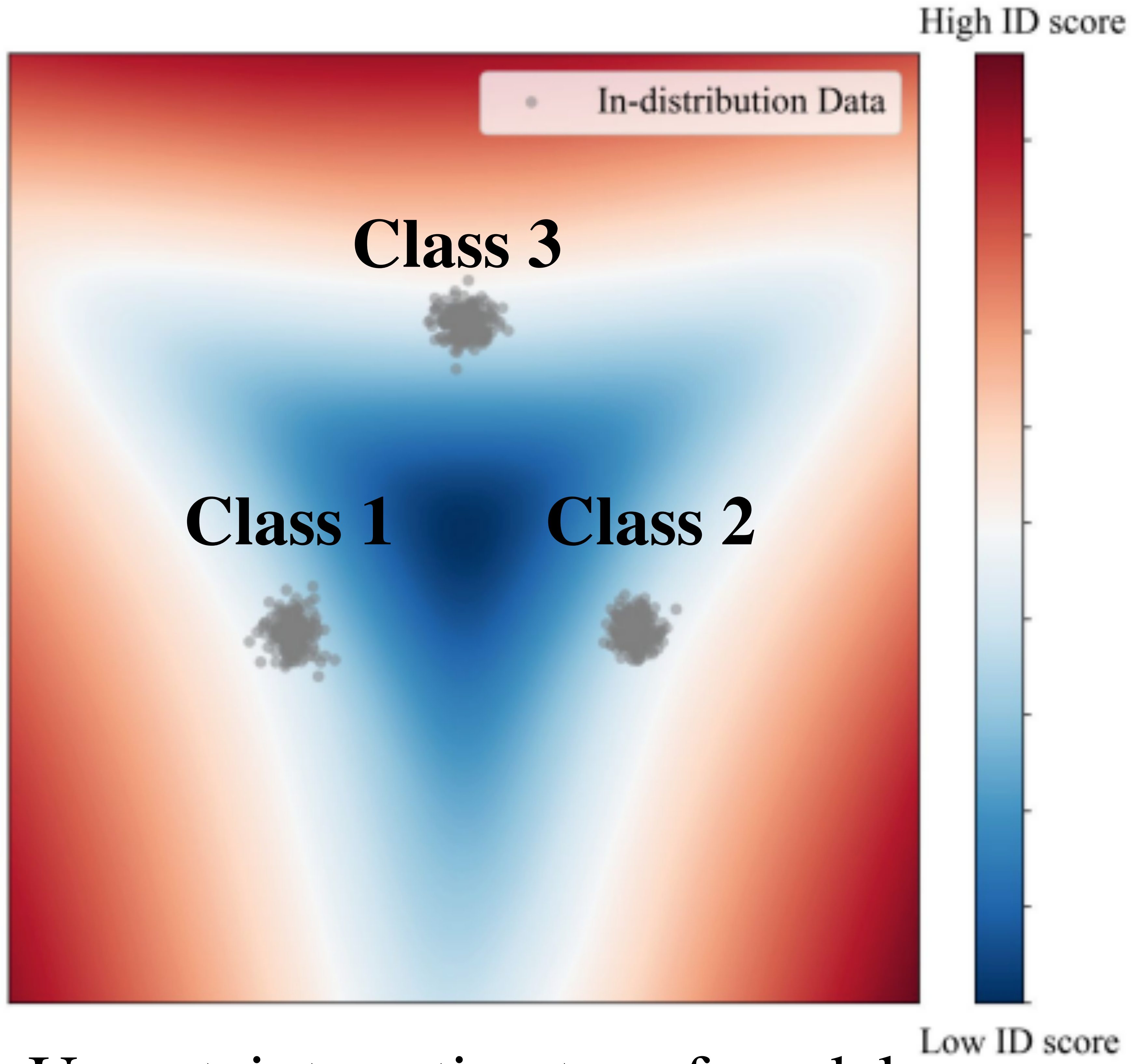
Learning Algorithm Design

Tutorial Outline

- **Inference-time OOD detection**
  - Output-based methods
  - Distance-based methods
- **Training-time regularization for OOD detection**
  - Safety-aware learning objective
  - Synthesizing virtual outliers
  - Leveraging wild unlabeled data

Insufficiency of ERM

- Existing learning algorithms are primarily driven by optimizing accuracy ***only*** on the ID data, but do not account for uncertainty from outside ID data.



Uncertainty estimates of model trained using standard CE loss **(not ideal)**

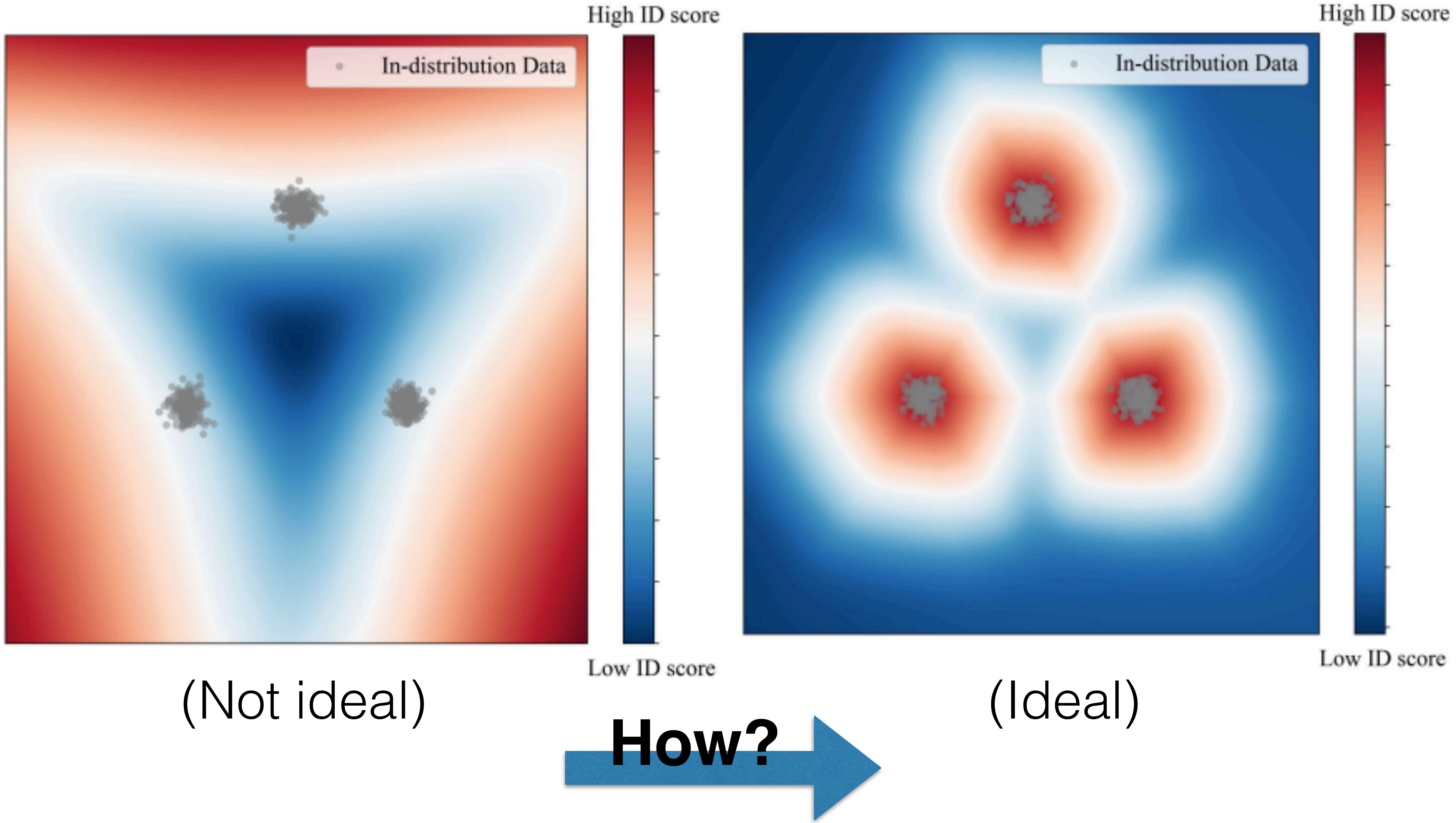
Empirical risk minimization:

$$R_{\text{closed}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i)$$

$$f^* = \operatorname{argmin}_{f\in\mathcal{F}} R_{\text{closed}}(f)$$

Going beyond ERM

- We need training-time regularization that ***explicitly*** accounts for uncertainty outside ID data.



(Not ideal)

(Ideal)

Safety-aware learning objective

Dual objectives in learning (ID classification and OOD detection):

$$\text{argmin} \left[\underbrace{R_{\text{closed}}(f)}_{\text{Classification error on ID}} + \alpha \cdot \underbrace{R_{\text{open}}(g)}_{\text{Error of OOD detector}} \right]$$

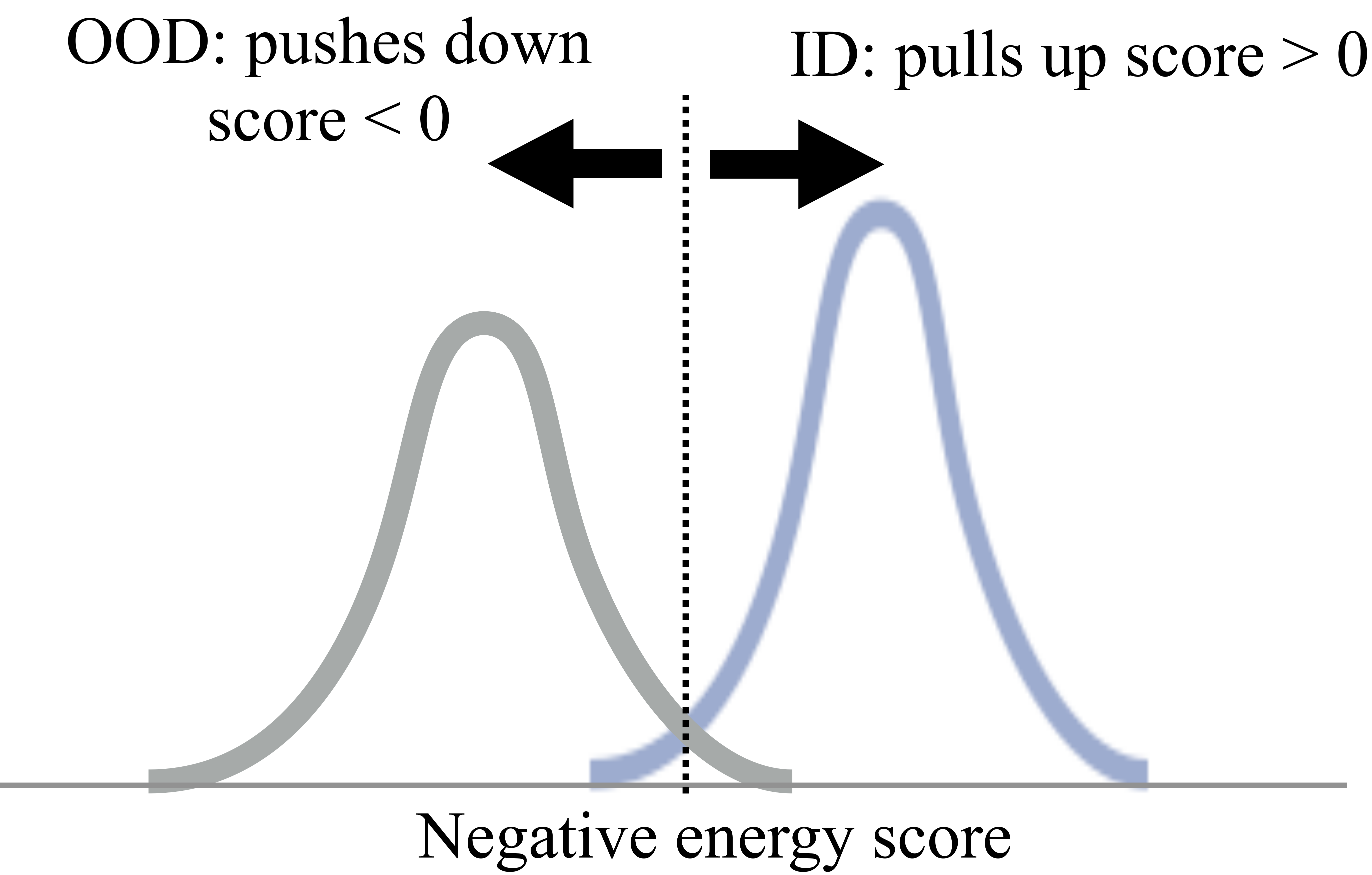
Safety-aware learning objective

Dual objectives in learning (ID classification and OOD detection):

$$\operatorname{argmin} \left[\underbrace{R_{\text{closed}}(f)}_{\text{Classification error on ID}} + \alpha \cdot \underbrace{R_{\text{open}}(g)}_{\text{Error of OOD detector}} \right]$$

$$\operatorname{argmin}\left[\underbrace{R_{\text{closed}}(f)}_{\text{Classification error on ID}} + \alpha \cdot \underbrace{R_{\text{open}}(g)}_{\text{Error of OOD detector}}\right]$$

$$R_{\text{open}}(g)$$



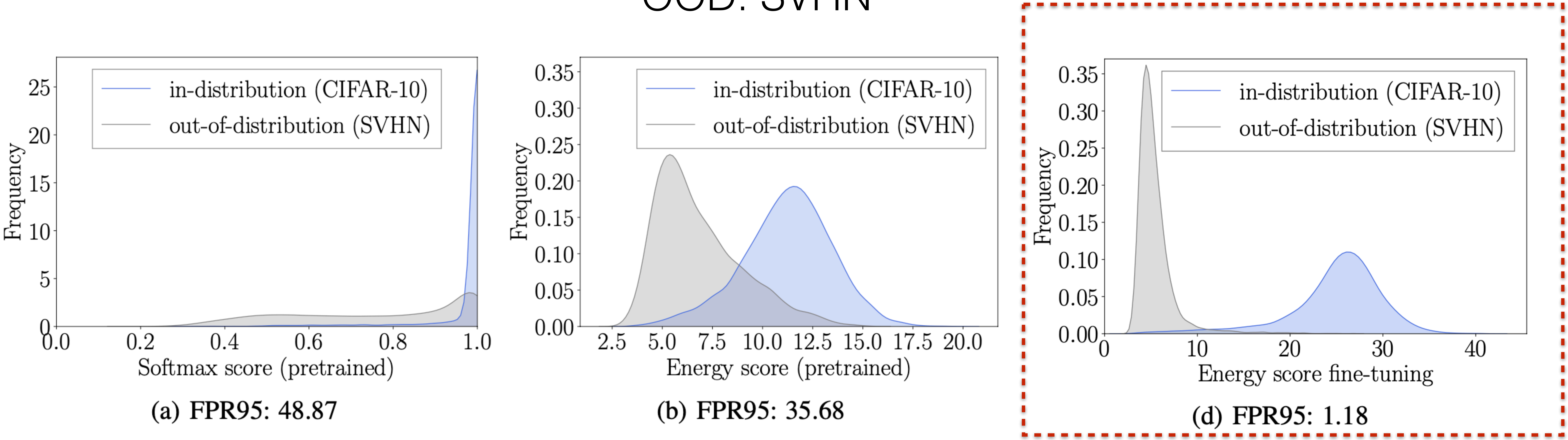
Recall that:

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{\boxed{p(\mathbf{x})}} = \frac{e^{f_y(\mathbf{x};\theta)}}{\boxed{\sum_{k=1}^{K} e^{f_k(\mathbf{x};\theta)}}}$$

$$E(\mathbf{x};\theta) := -\log \sum_{k=1}^{K} e^{f_k(\mathbf{x};\theta)}$$

Training-time Regularization Improves ID/OOD Separability

ID: CIFAR-10
OOD: SVHN



Caveat: requires auxiliary outlier training data, which can be difficult to obtain

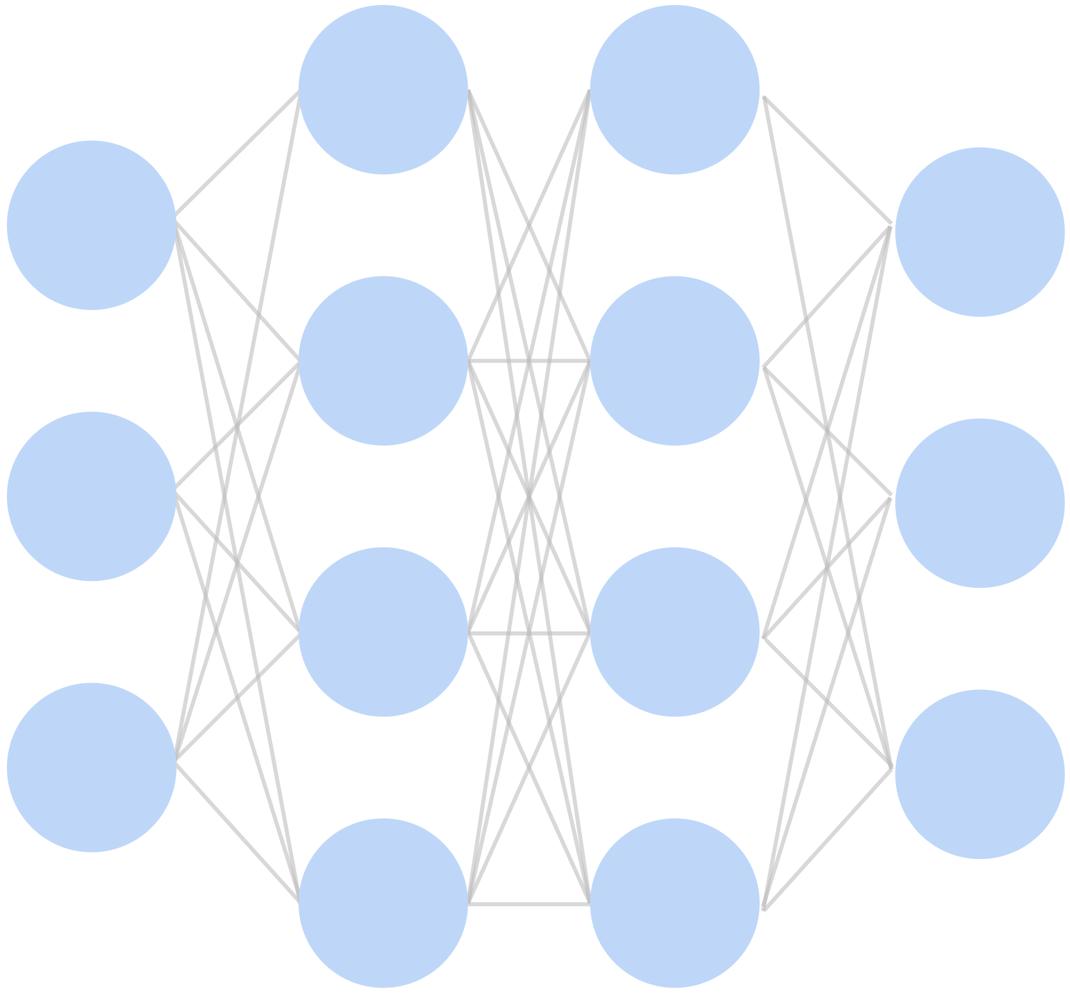
How to obtain auxiliary OOD training data, for free?

Tutorial Outline

- **Inference-time OOD detection**
  - Output-based methods
  - Distance-based methods
- **Training-time regularization for OOD detection**
  - Safety-aware learning objective
  - Synthesizing virtual outliers
  - Leveraging wild unlabeled data

Virtual Outlier Synthesis

Sample low-likelihood data points in the **feature space** for model regularization

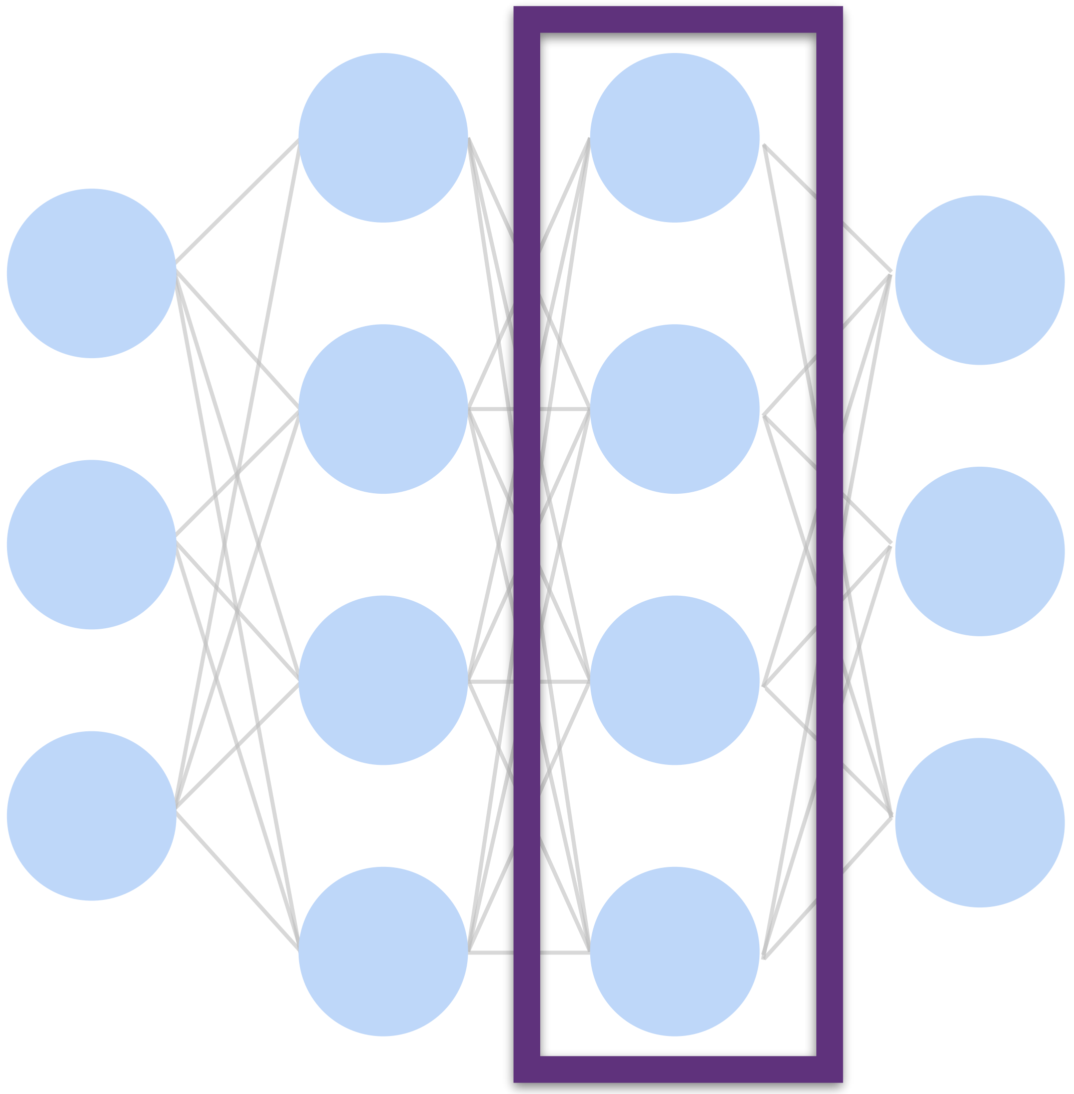


Du et al., *VOS: Learning What You Don't Know by Virtual Outlier Synthesis*, ICLR, 2022

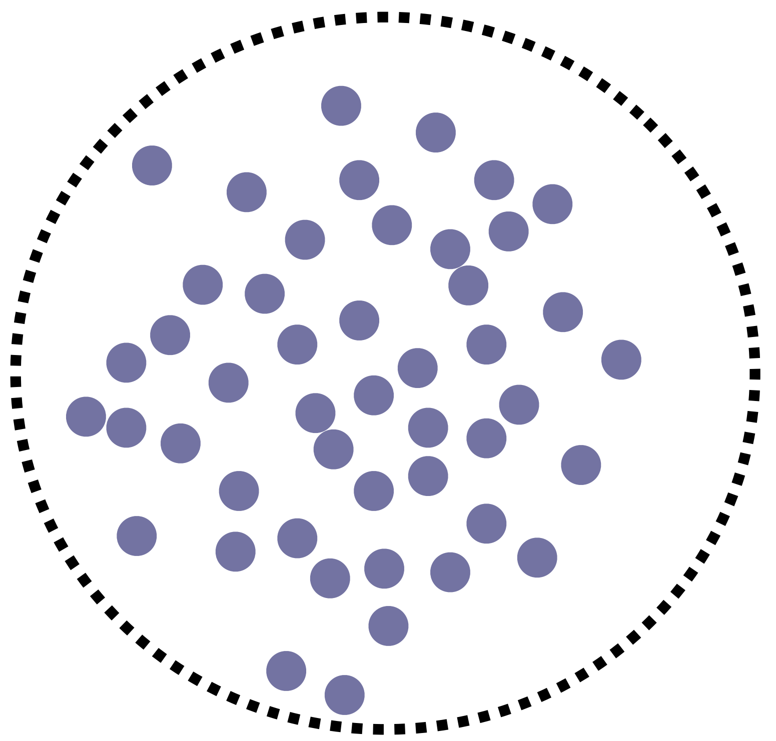
Virtual Outlier Synthesis

Sample low-likelihood data points in the **feature space** for model regularization

Modeling feature representation as class-conditional Gaussian distribution



$h(x; \theta)$

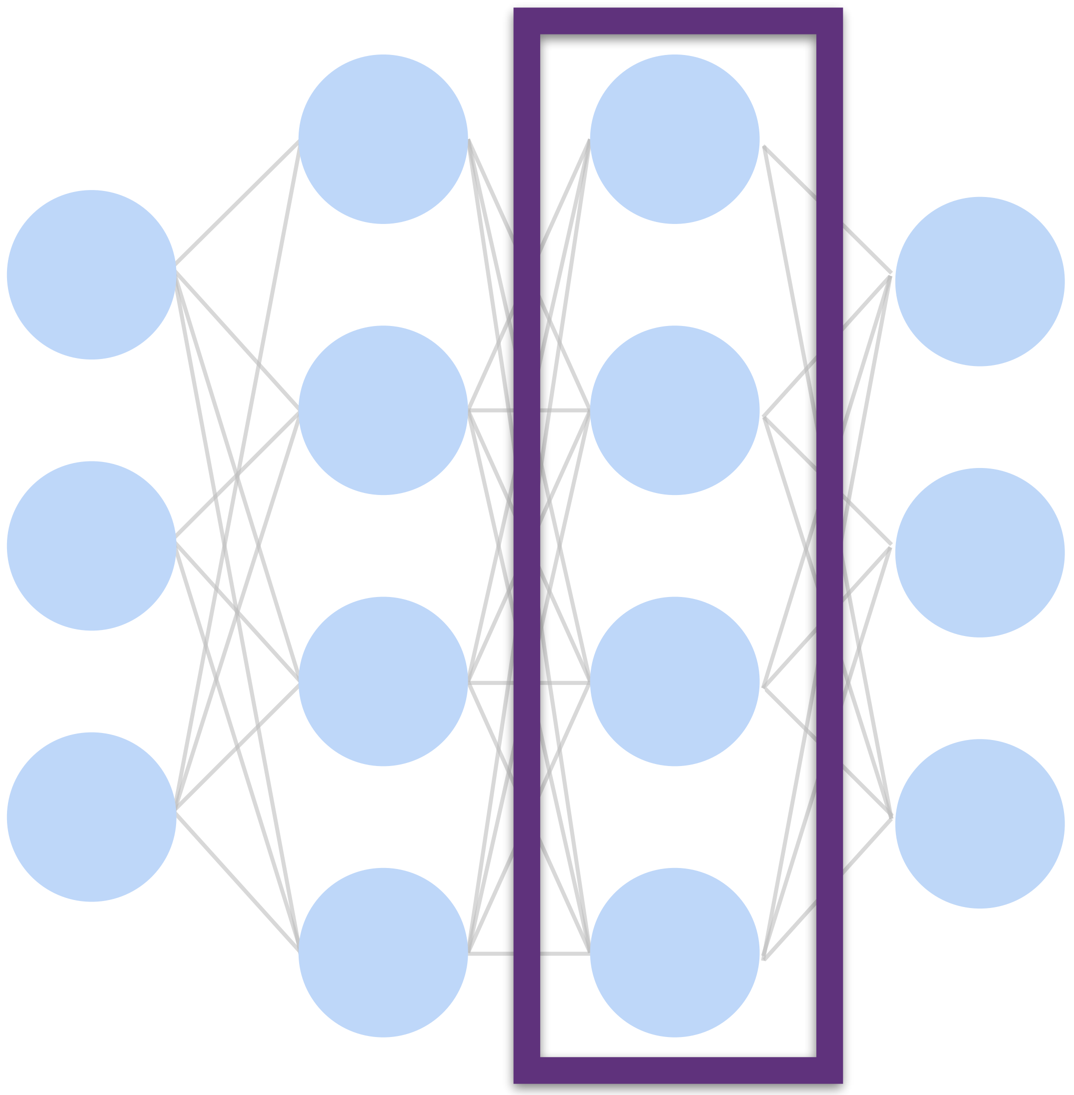


$$p_\theta(h(\mathbf{x}, \mathbf{b}) | y = k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

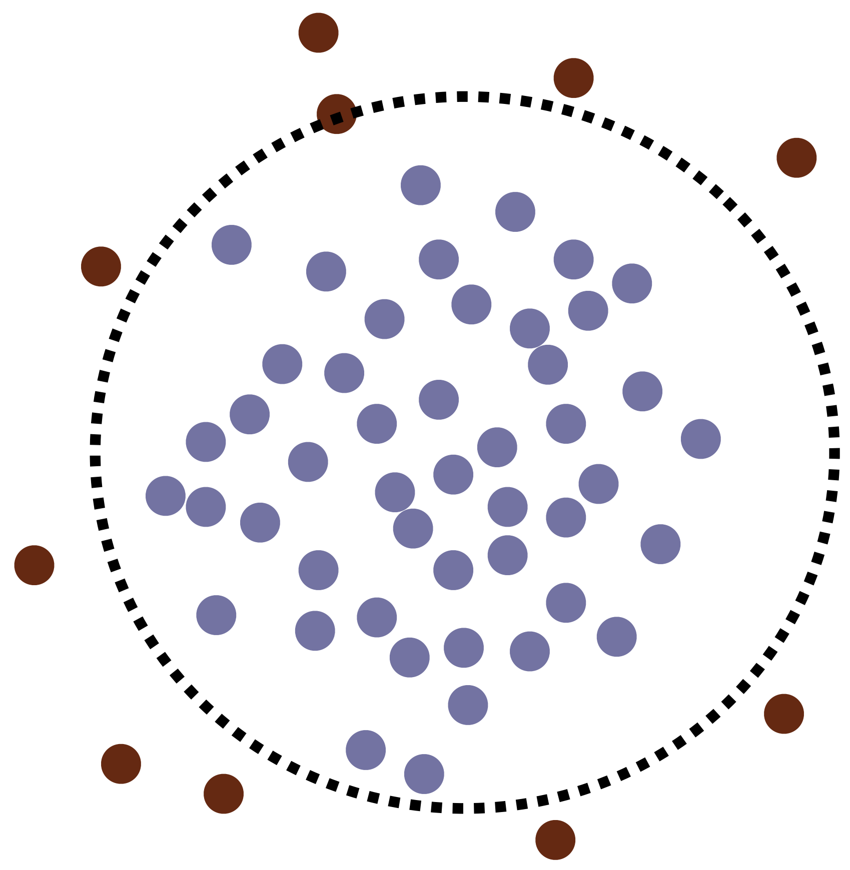
Du et al., *VOS: Learning What You Don't Know by Virtual Outlier Synthesis,* ICLR, 2022

Virtual Outlier Synthesis

Sample low-likelihood data points in the **feature space** for model regularization



Sample virtual outliers from the class-conditional Gaussian distribution

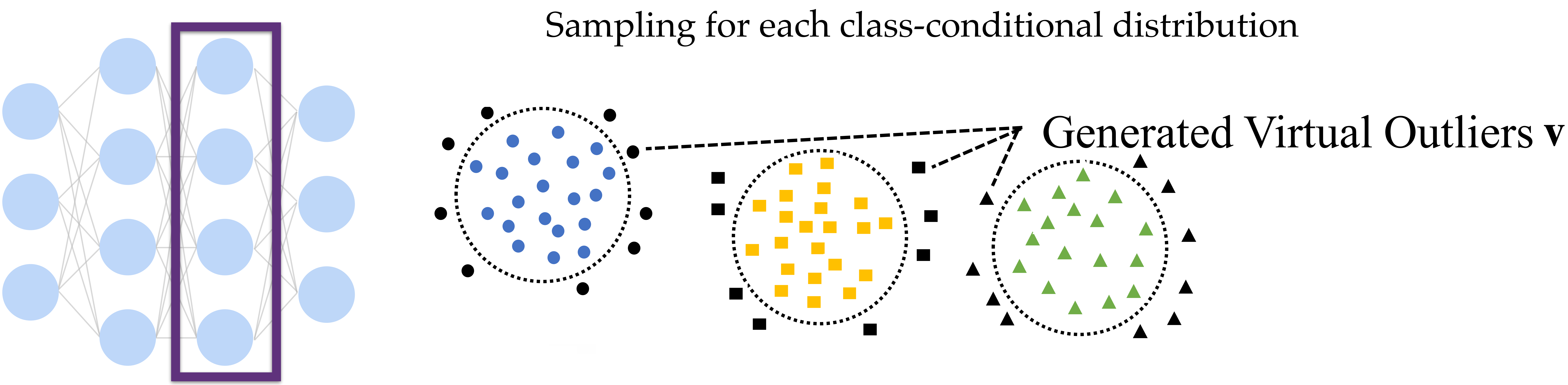


$$\mathcal{V}_k = \{\mathbf{v}_k | \frac{1}{(2\pi)^{m/2}|\widehat{\boldsymbol{\Sigma}}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{v}_k - \widehat{\boldsymbol{\mu}}_k)^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{v}_k - \widehat{\boldsymbol{\mu}}_k)\right) < \epsilon\}$$

Du et al., *VOS: Learning What You Don't Know by Virtual Outlier Synthesis,* ICLR, 2022

Virtual Outlier Synthesis

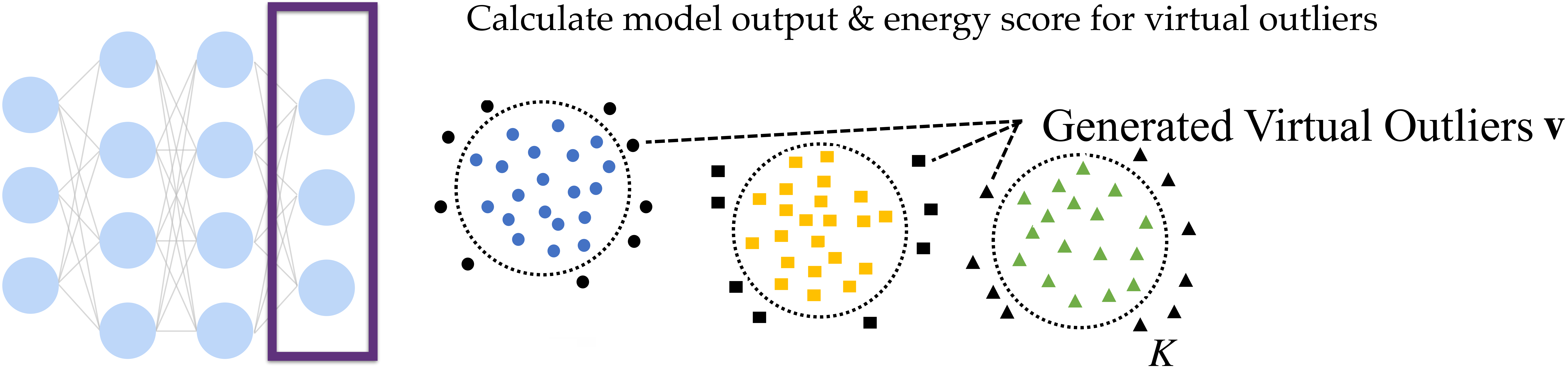
Sample low-likelihood data points in the **feature space** for model regularization



Du et al., *VOS: Learning What You Don't Know by Virtual Outlier Synthesis*, ICLR, 2022

Virtual Outlier Synthesis

Sample low-likelihood data points in the **feature space** for model regularization

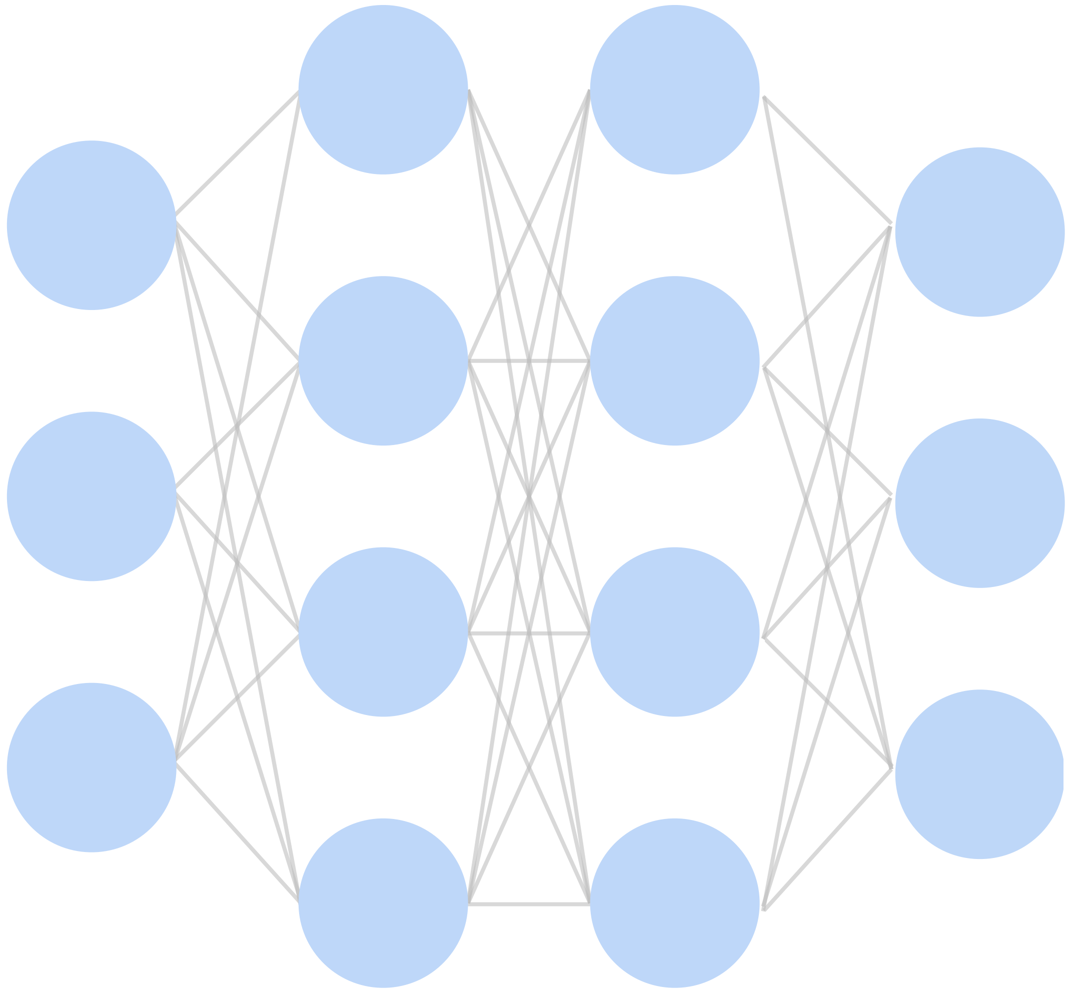


$$f(\mathbf{v};\theta) = W_{\text{cls}}^{\top}\mathbf{v}, \longrightarrow E(\mathbf{v};\theta) = -\log \sum_{k=1}^{K} e^{f_k(\mathbf{v};\theta)}$$

Du et al., *VOS: Learning What You Don't Know by Virtual Outlier Synthesis*, ICLR, 2022

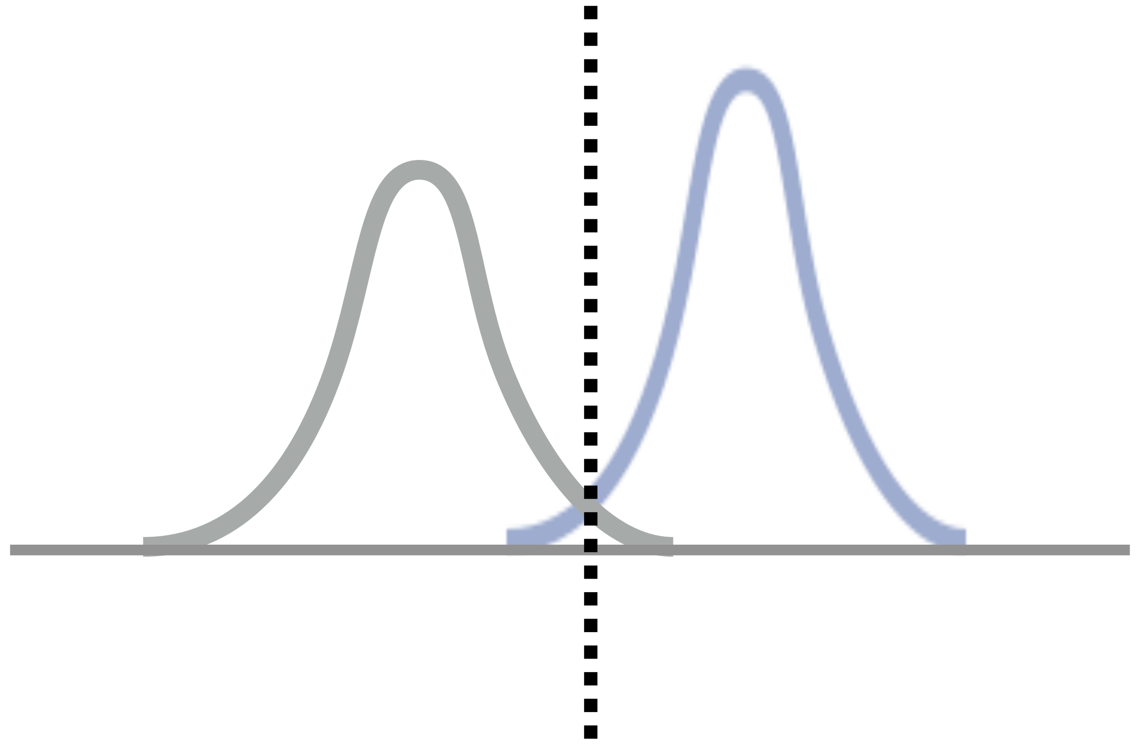
Learning Objective with Virtual Outliers

Our learning framework **jointly** optimizes for both: (1) accurate classification of samples from ID, and (2) reliable detection of data from outside ID.



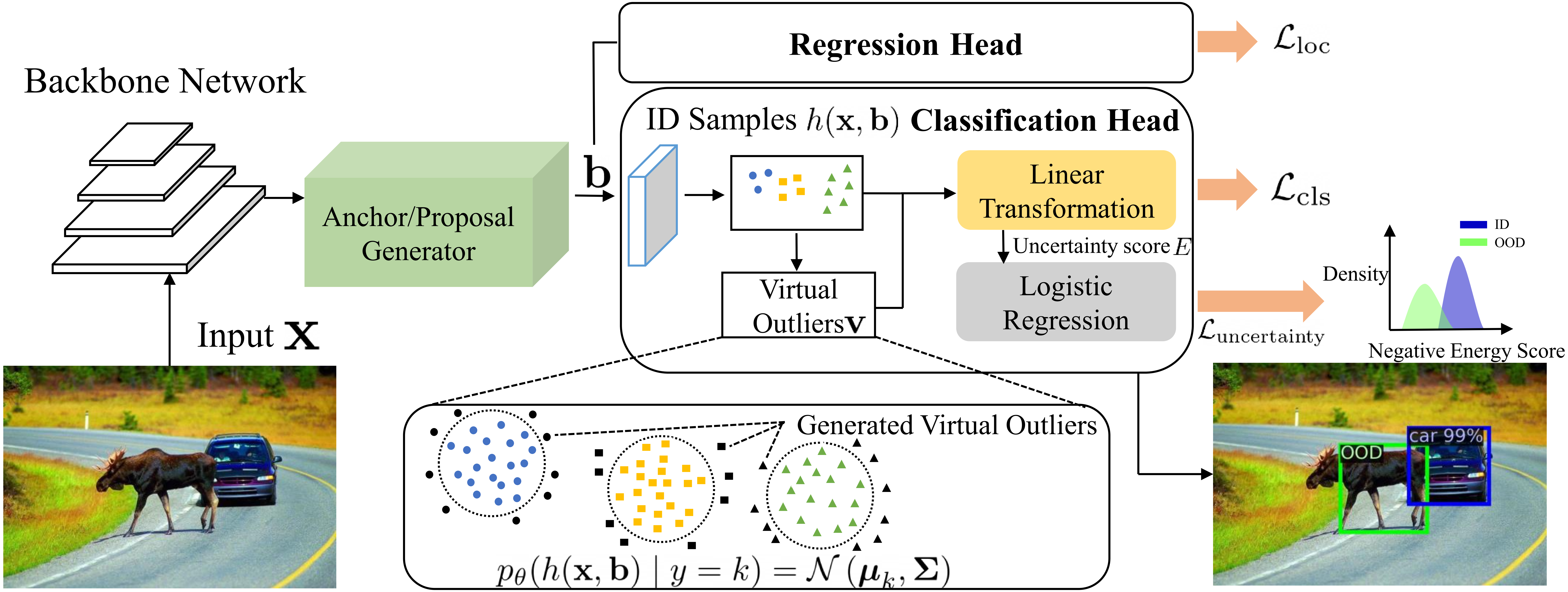
$$\text{argmin} \left[\underbrace{R_{\text{closed}}(f)}_{\text{Classification error on ID}} + \alpha \cdot \underbrace{R_{\text{open}}(g)}_{\text{Error of OOD detector}} \right]$$

$$R_{\text{open}}(g) = \mathbb{E}_{\mathbf{v}\sim\mathcal{V}} \ \mathbb{1}\{E(\mathbf{v};\theta) > 0\} + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \ \mathbb{1}\{E(\mathbf{x};\theta) \leq 0\}$$



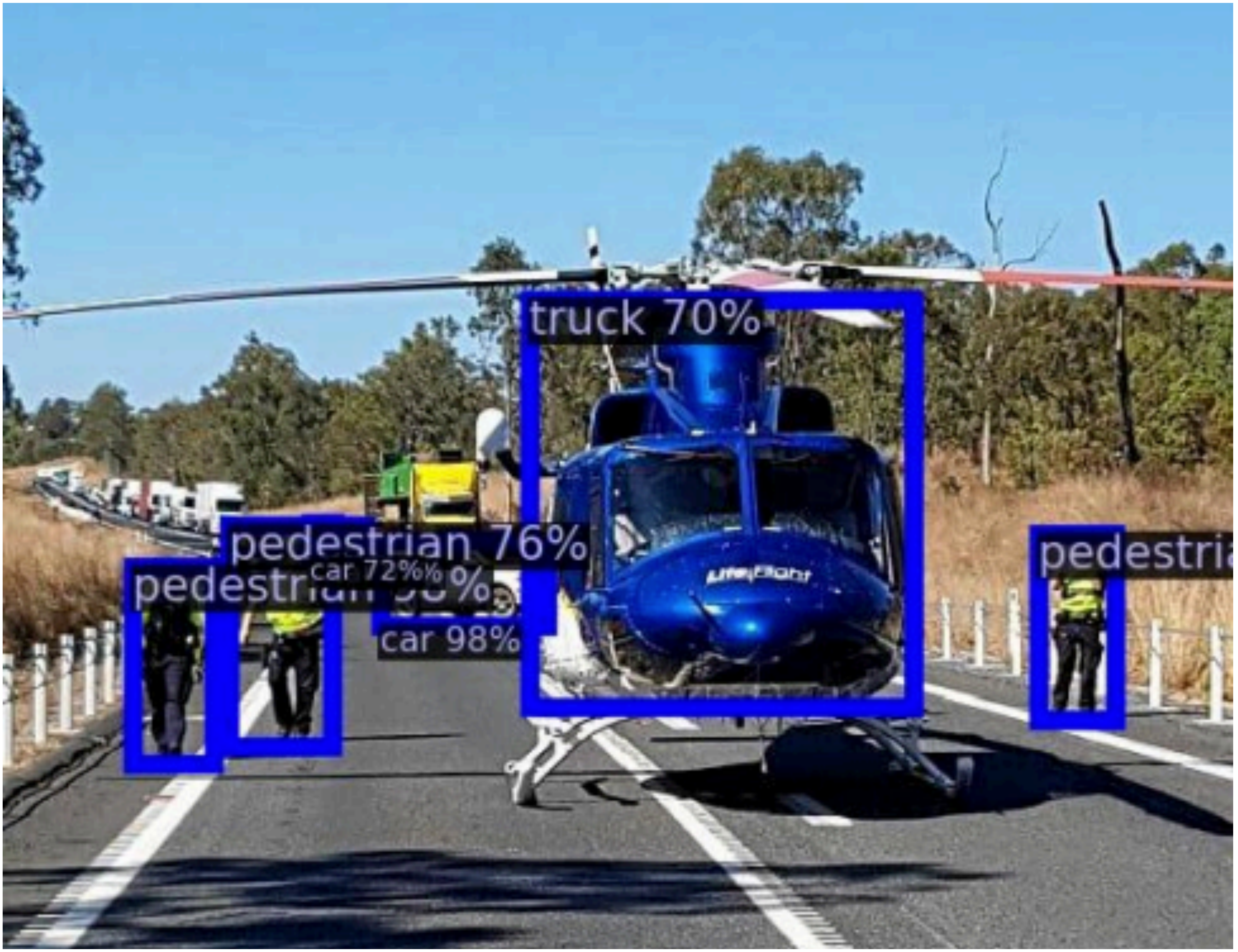
Du et al., *VOS: Learning What You Don't Know by Virtual Outlier Synthesis,* ICLR, 2022

Virtual Outlier Synthesis for Object Detection



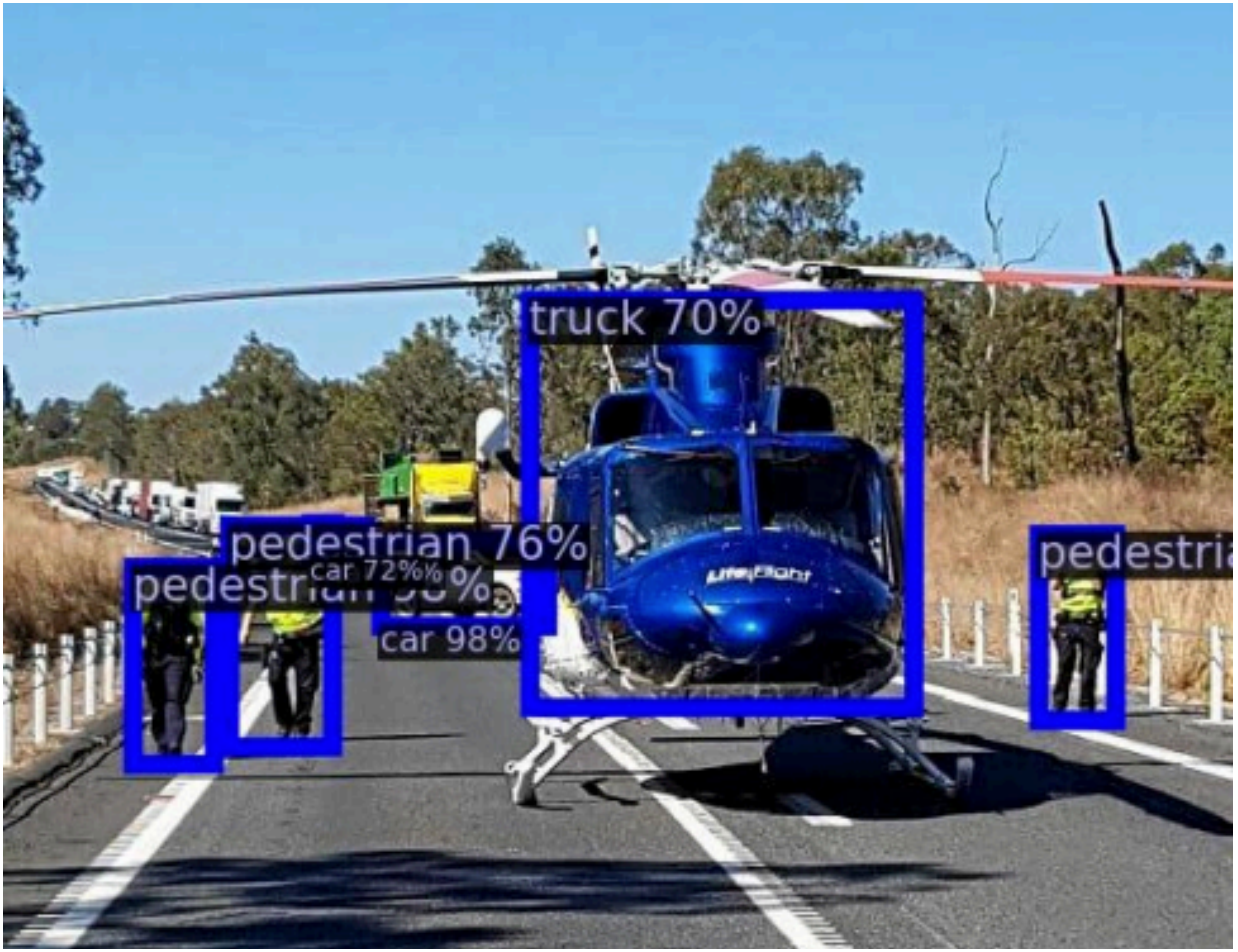
VOS is a general learning framework that is suitable for both object detection and image classification tasks.

Results

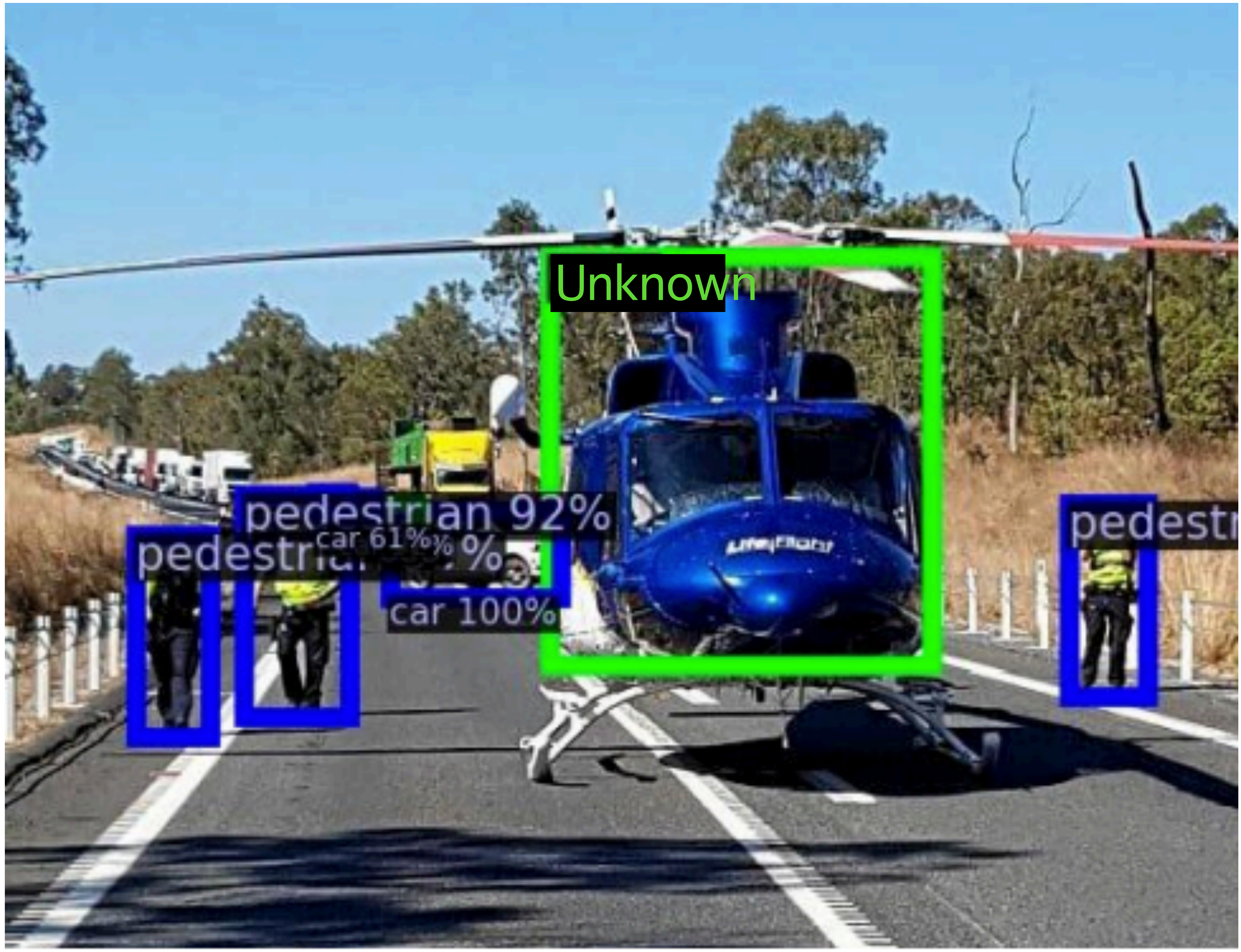


Without VOS

Results



Without VOS



With VOS

Non-Parametric Outlier Synthesis

NON-PARAMETRIC OUTLIER SYNTHESIS

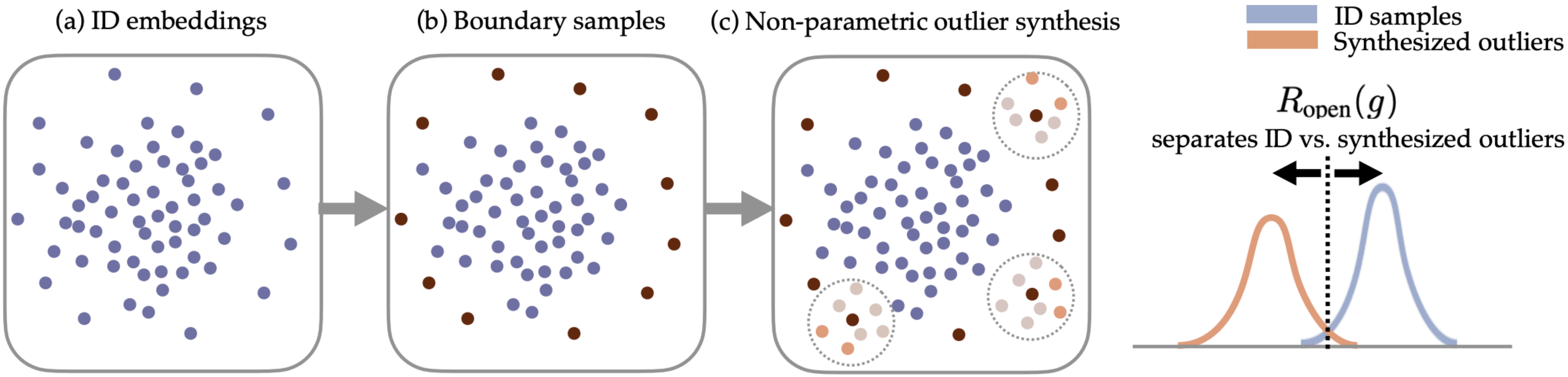
Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) detection is indispensable for safely deploying machine learning models in the wild. One of the key challenges is that models lack supervision signals from unknown data, and as a result, can produce overconfident predictions on OOD data. Recent work on outlier synthesis modeled the feature space as parametric Gaussian distribution, a strong and restrictive assumption that might not hold in reality. In this paper, we propose a novel framework, *non-parametric outlier synthesis* (NPOS), which generates artificial OOD training data and facilitates learning a reliable decision boundary between ID and OOD data. Importantly, our proposed synthesis approach does not make any distributional assumption on the ID embeddings, thereby offering strong flexibility and generality. We show that our synthesis approach can be mathematically interpreted as a rejection sampling framework. Extensive experiments show that NPOS can achieve superior OOD detection performance, outperforming the competitive rivals by a significant margin.

Tao et al., *Non-Parametric Outlier Synthesis,* ICLR, 2023

Non-Parametric Outlier Synthesis



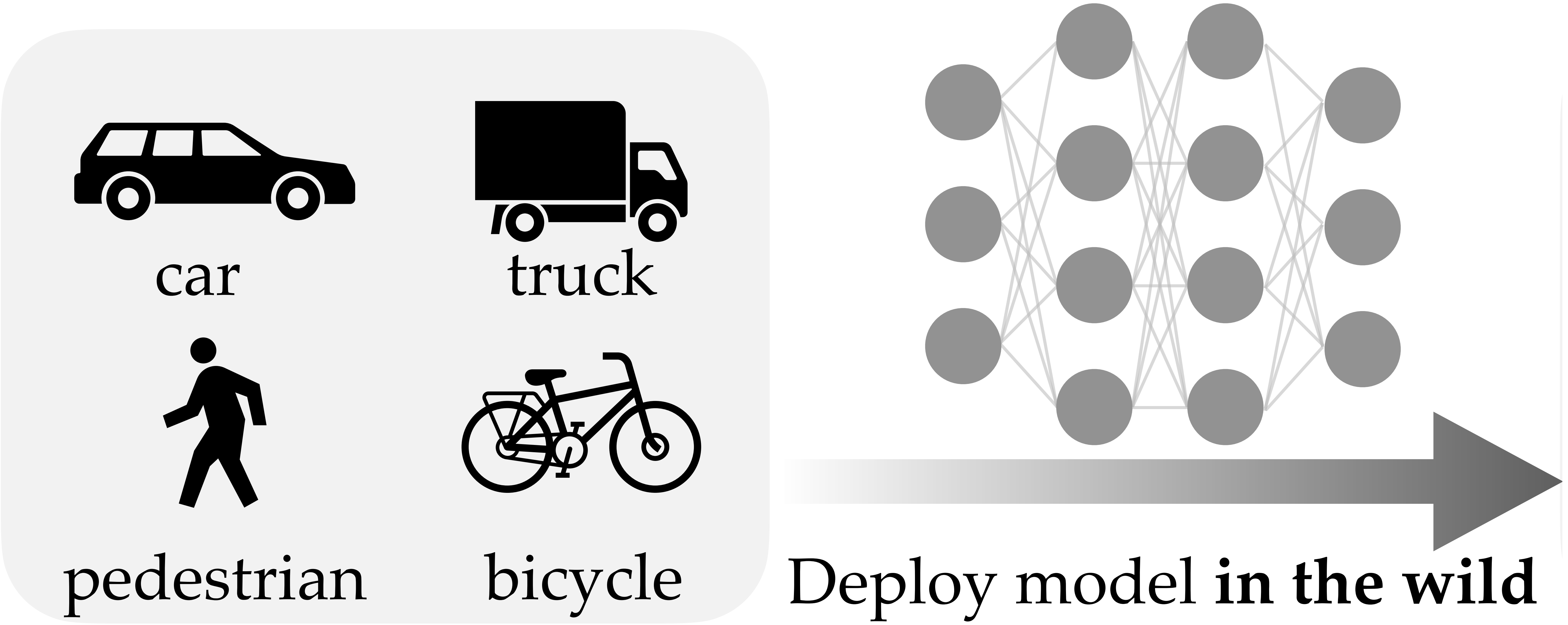
Sampling virtual outliers **without making distributional assumption** about feature embedding. Strong generality and flexibility.

How to obtain **natural** outlier training data, for free?

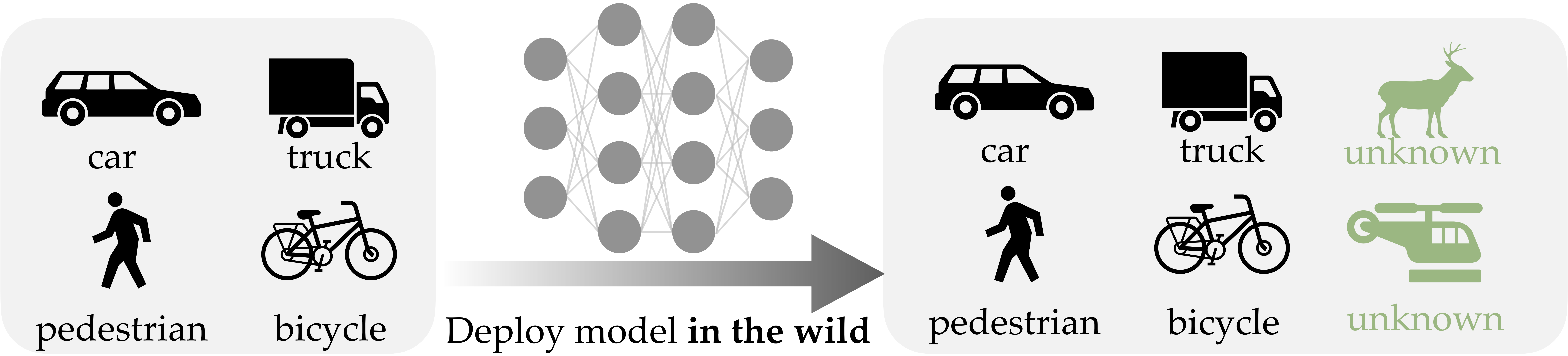
Tutorial Outline

- **Inference-time OOD detection**
  - Output-based methods
  - Distance-based methods
- **Training-time regularization for OOD detection**
  - Safety-aware learning objective
  - Synthesizing virtual outliers
  - Leveraging wild unlabeled data

Leveraging Wild Unlabeled Data for OOD Detection

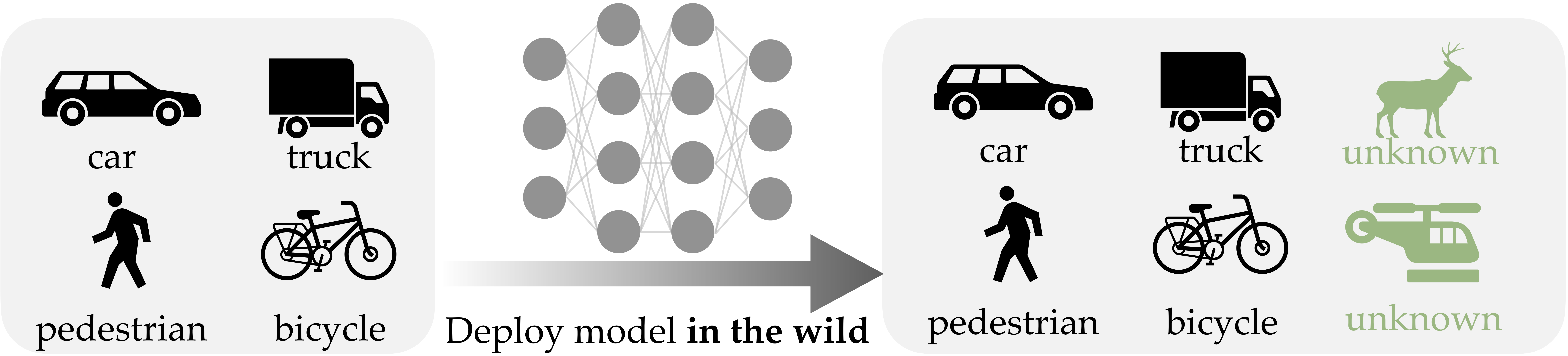


Leveraging Wild Unlabeled Data for OOD Detection



Advantages: (1) data is available in abundance, (2) does not require any human annotation, and (3) is often a much better match to the true test time distribution than data collected offline.

Leveraging Wild Unlabeled Data for OOD Detection



Challenges: Wild data is not pure, and consists of both ID data and OOD data

$$\mathbb{P}_{\text{wild}} := (1-\pi)\mathbb{P}_{\text{in}} + \pi\mathbb{P}_{\text{out}}$$

Training OOD Detectors in their Natural Habitats

Julian Katz-Samuels [*1] **Julia Nakhleh** [*2] **Robert Nowak** [3] **Yixuan Li** [2]

Abstract

Out-of-distribution (OOD) detection is important for machine learning models deployed in the wild. Recent methods use auxiliary outlier data to regularize the model for improved OOD detection.

Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild

Xuefeng Du[1], Xin Wang[2], Gabriel Gozum[1], and Yixuan Li[1]
[1]University of Wisconsin-Madison, [2]Microsoft Research
{xfdu,sharonli}@cs.wisc.edu, wanxin@microsoft.com, ggozum@wisc.edu

Abstract

Building reliable object detectors that can detect out-of-distribution (OOD) objects is critical yet underexplored. One of the key challenges is that models lack supervision signals from unknown data, producing overconfident predictions on OOD objects. We propose a new unknown-aware object detection framework through Spatial-Temporal Unknown Distillation (STUD), which dis-

(a) Overconfident Predictions (b) Unknown objects in videos

Figure 1. (a) Vanilla object detectors can predict OOD objects (*e.g.*, deer) as an ID class (*e.g.*, pedestrian) with high confidence.

[1] Katz-Samuels et al,, *Training OOD Detectors in their Natural Habitats*, ICML 2022
[2] Du et al., *Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild,* CVPR 2022
[3] Bai et al., Feed Two Birds with One Scone: Exploiting Wild Data for Both OOD Generalization and Detection, ICML 2023

Summary

- **Inference-time OOD detection**
  - Output-based methods
  - Distance-based methods
- **Training-time regularization for OOD detection**
  - Safety-aware learning objective
  - Synthesizing virtual outliers
  - Leveraging wild unlabeled data

Thank you!

sharonli@cs.wisc.edu