Detecting Reliable Instances for Learning

Judy Hoffman

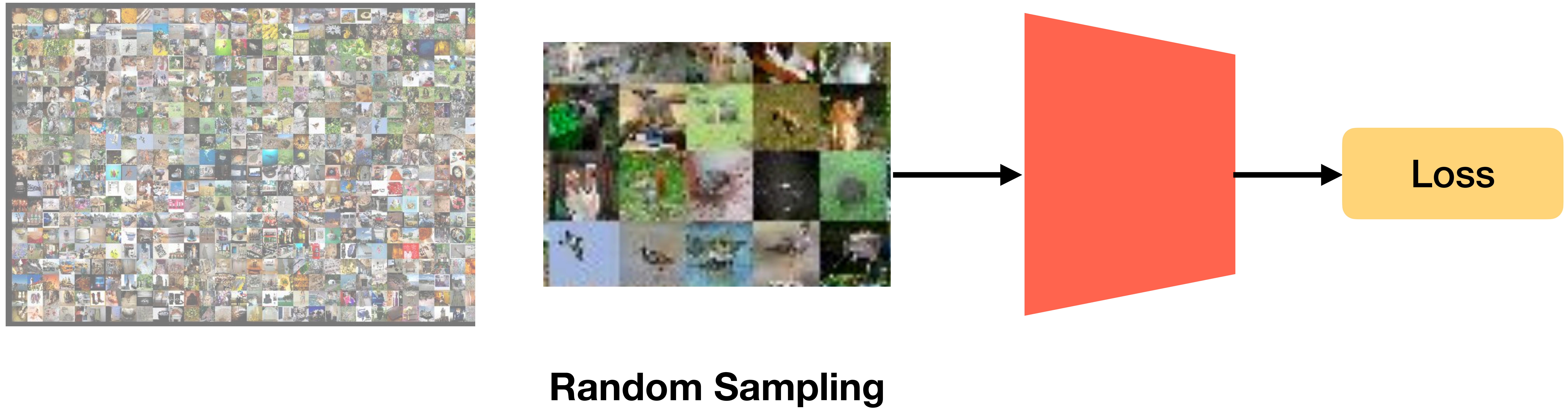
Adversarial Machine Learning in Computer Vision
CVPR Workshop 2021

Georgia Institute of Technology

ML@GT

Standard Supervised Learning

Standard Supervised Learning



Potential Data Pitfalls

Incorrect Label	Adversarial Manipulations	Variable Difficulty
		
Label: “Cat”	**Label:** “Dog”	**Label:** “Dog”
Prediction: “Cat”	**Prediction:** “Cat”	**Prediction:** “Cat”

Potential Data Pitfalls

Incorrect Label

Label: “Cat”

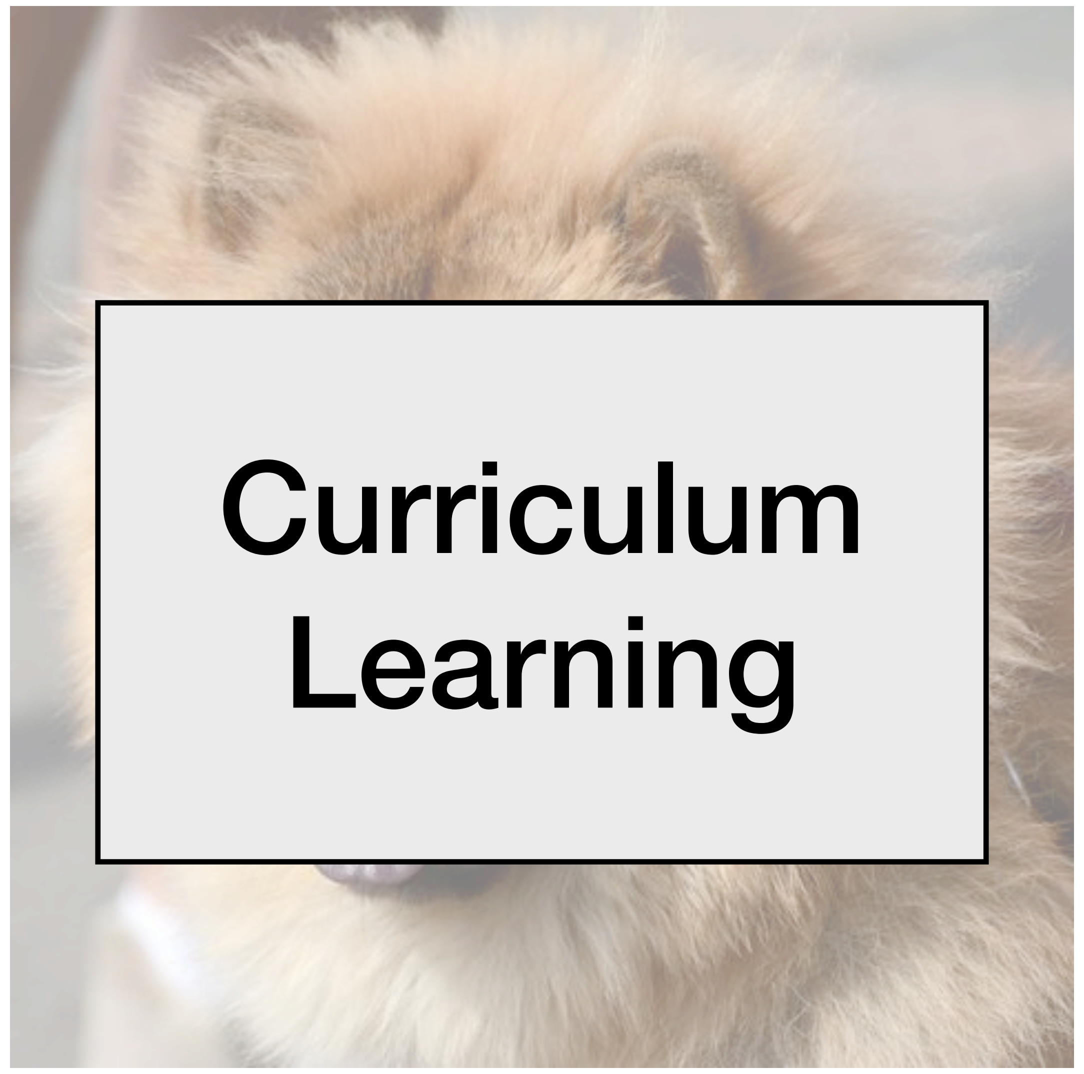
Prediction: “Cat”

Adversarial Manipulations

Label: “Dog”

Prediction: “Cat”

Variable Difficulty



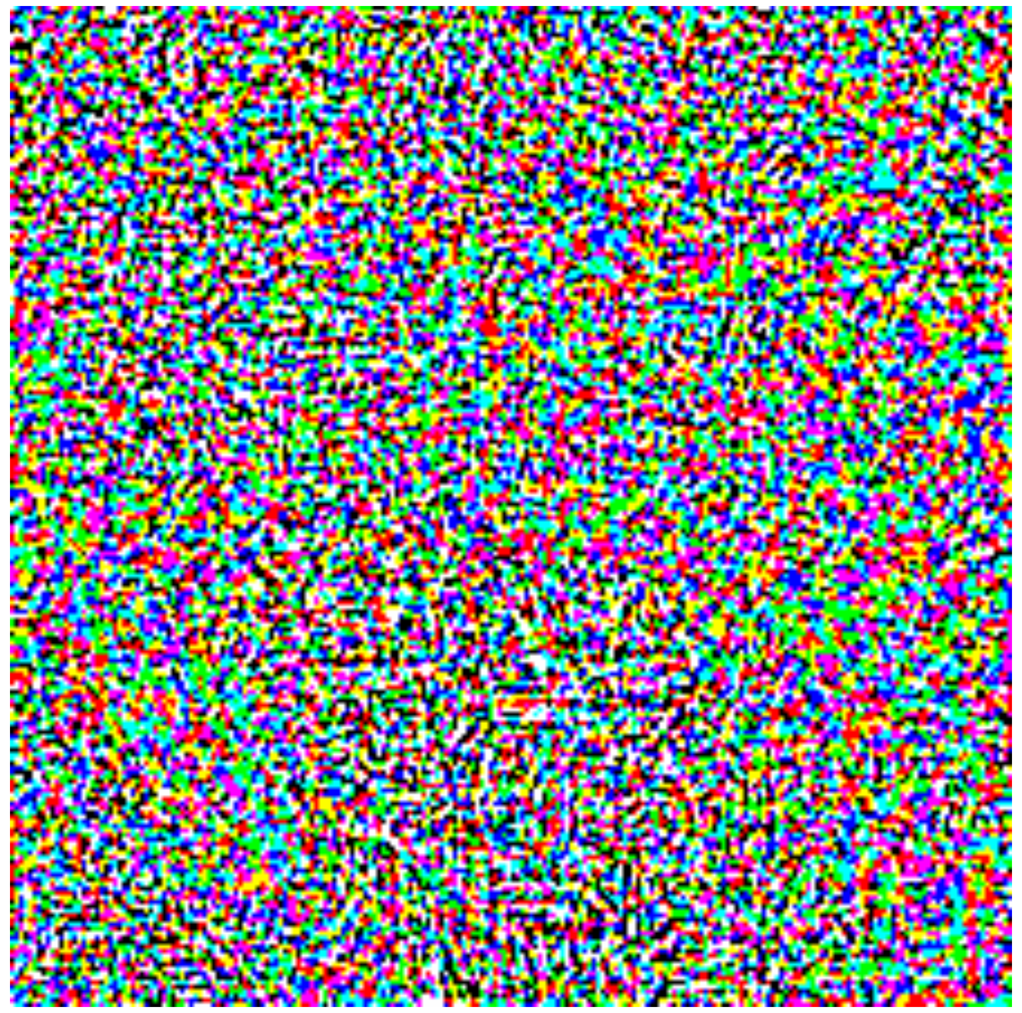
Label: “Dog”

Prediction: “Cat”

Enforcing Reliability

Adversarial Examples

$+ .007 \times$



$=$

$\boldsymbol{x}$
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

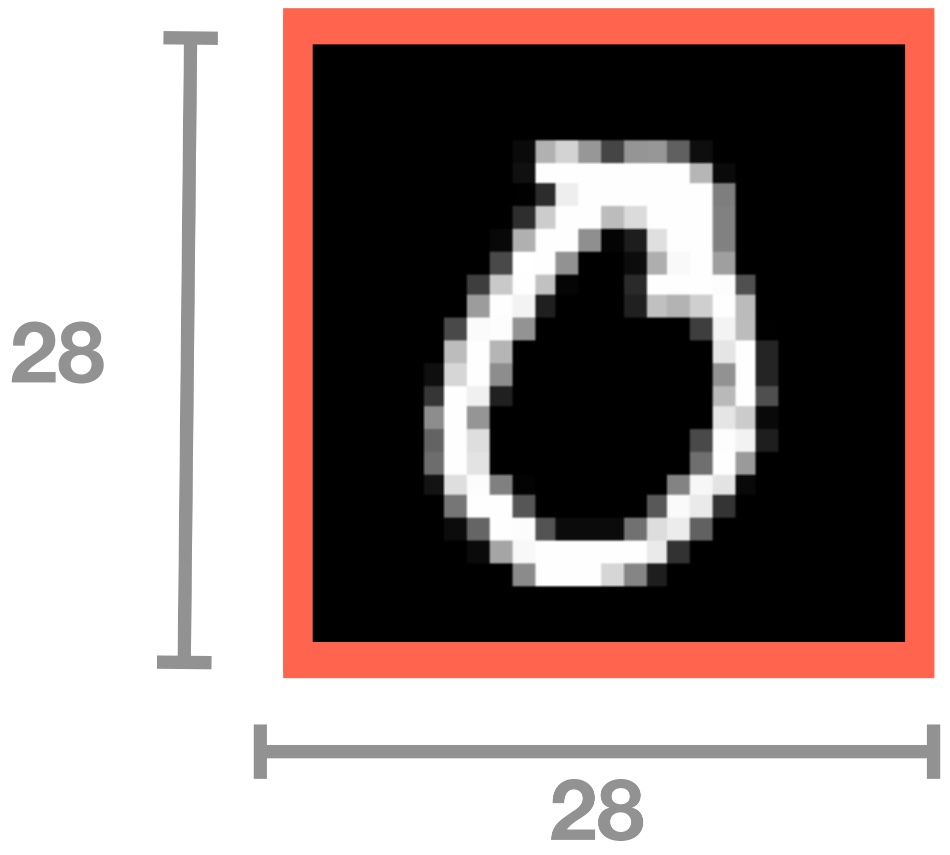
$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

[Goodfellow et al. ICLR 2015]

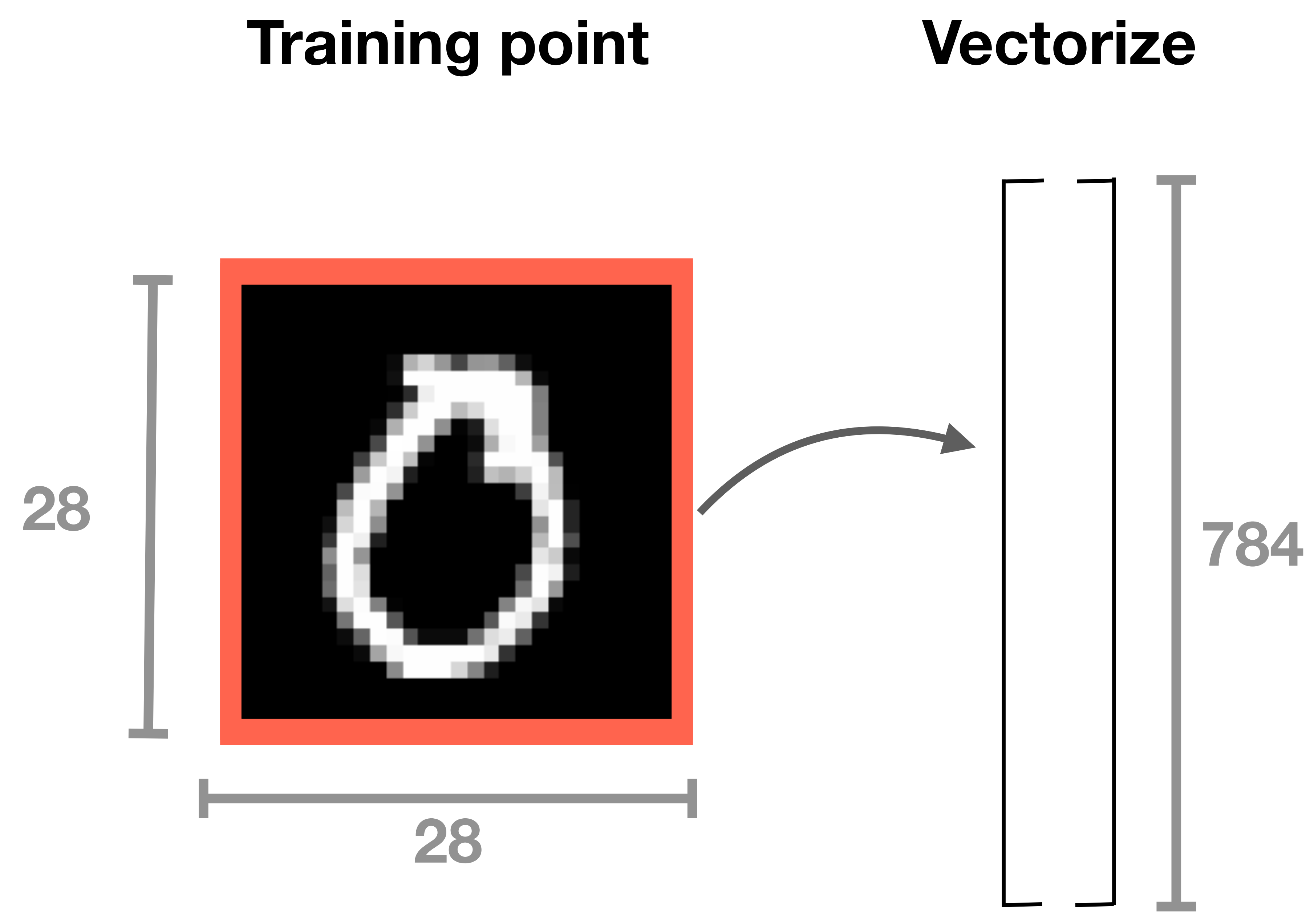
Visualize Perturbation Space

Visualize Perturbation Space

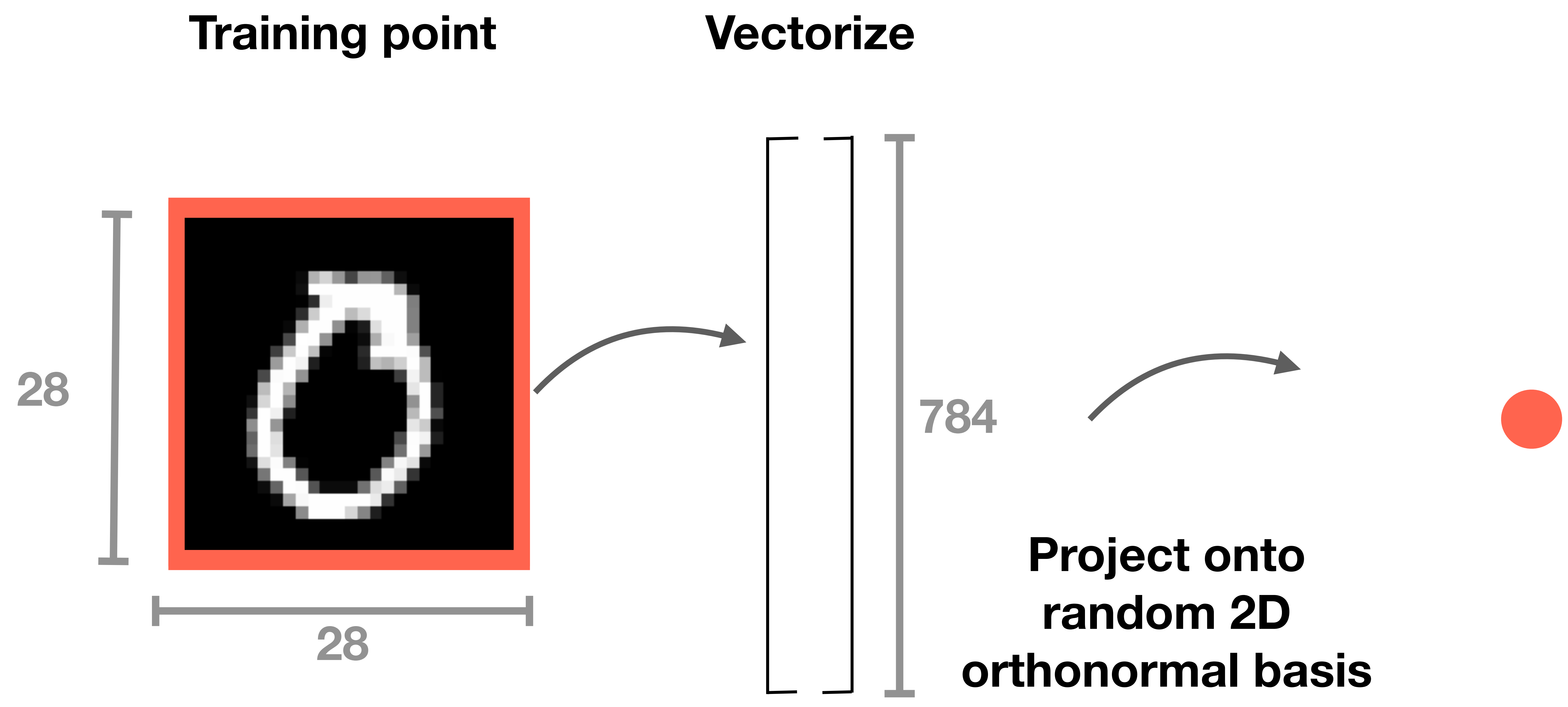
Training point



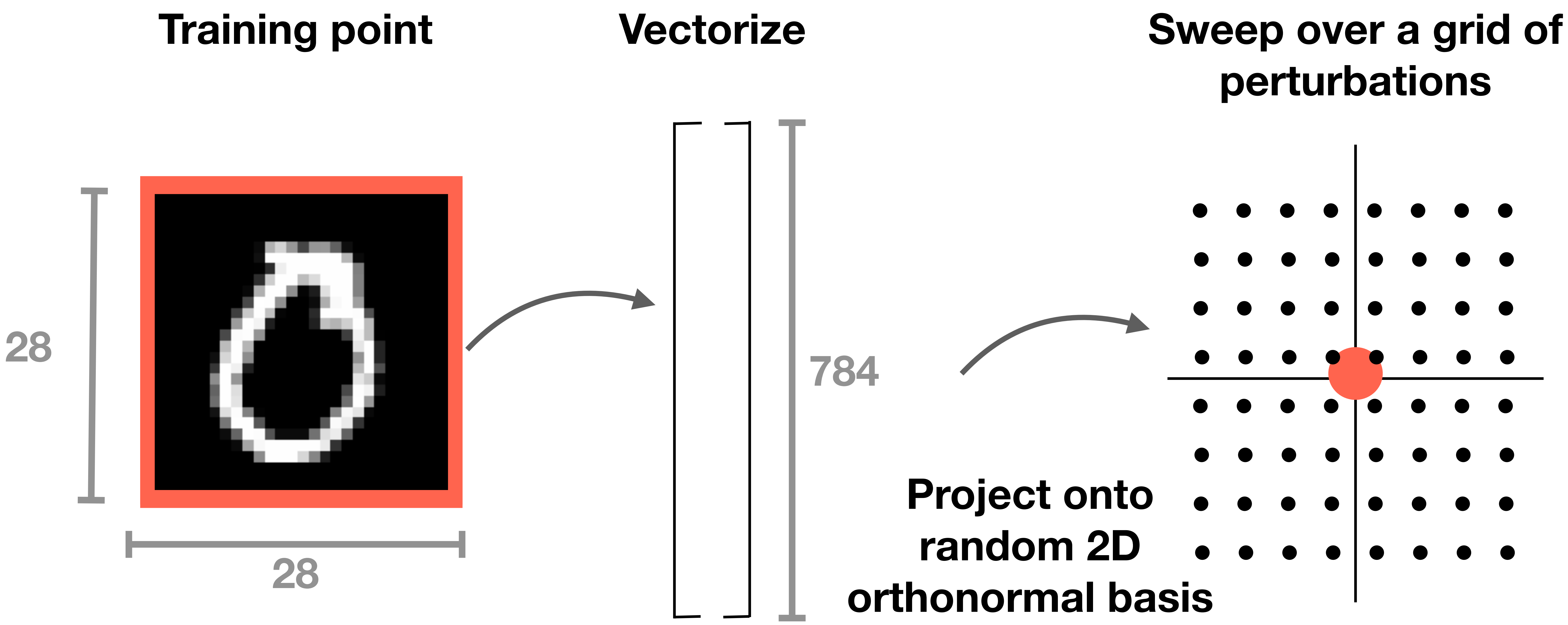
Visualize Perturbation Space



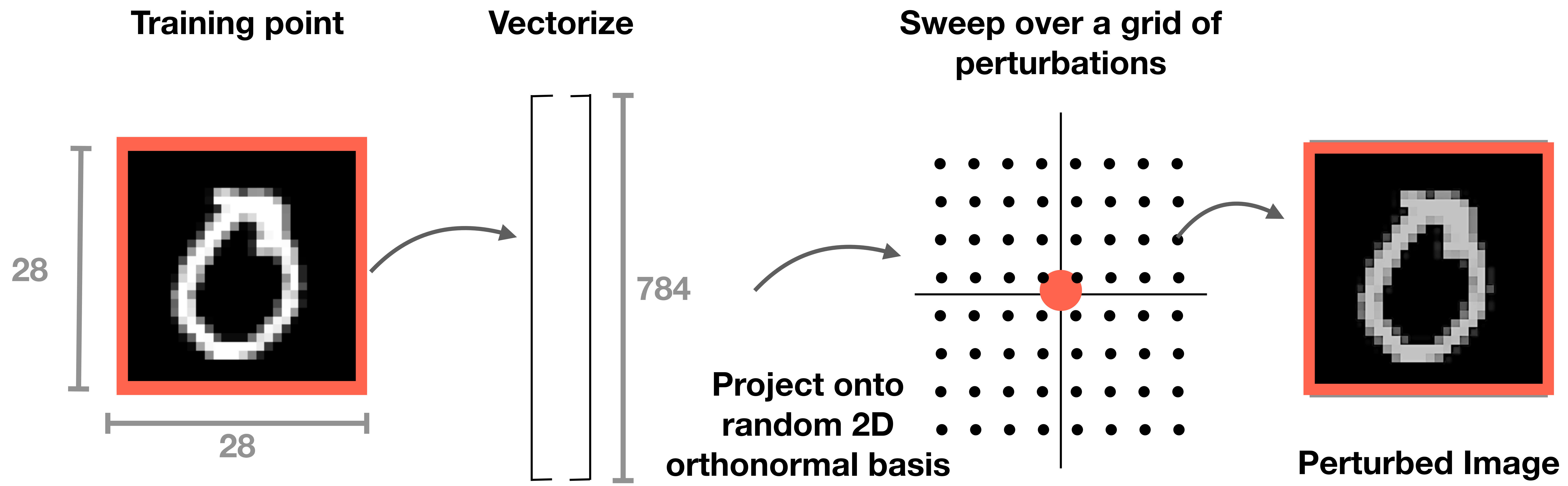
Visualize Perturbation Space



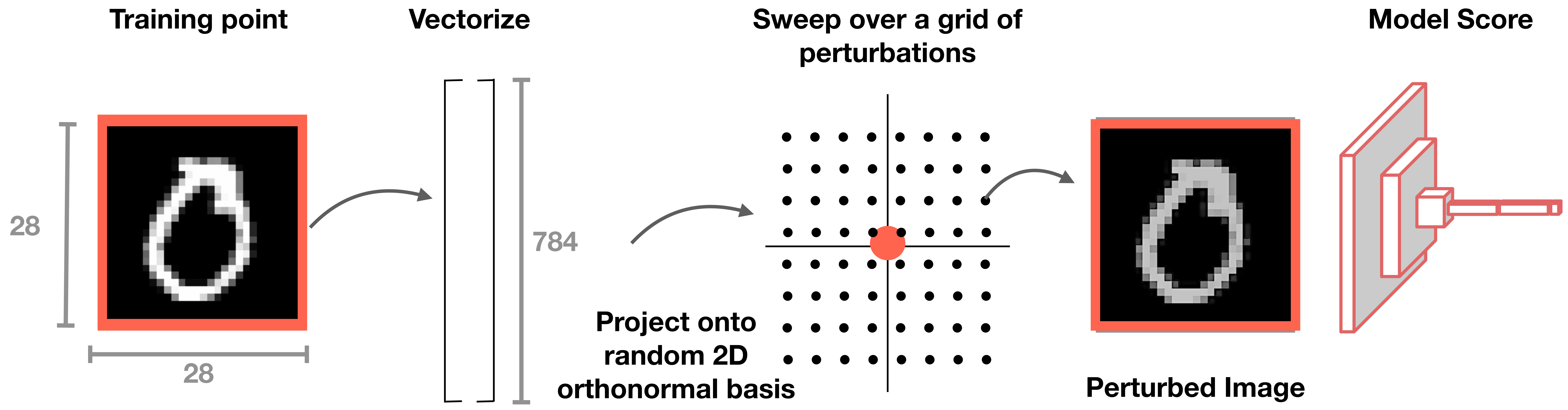
Visualize Perturbation Space



Visualize Perturbation Space



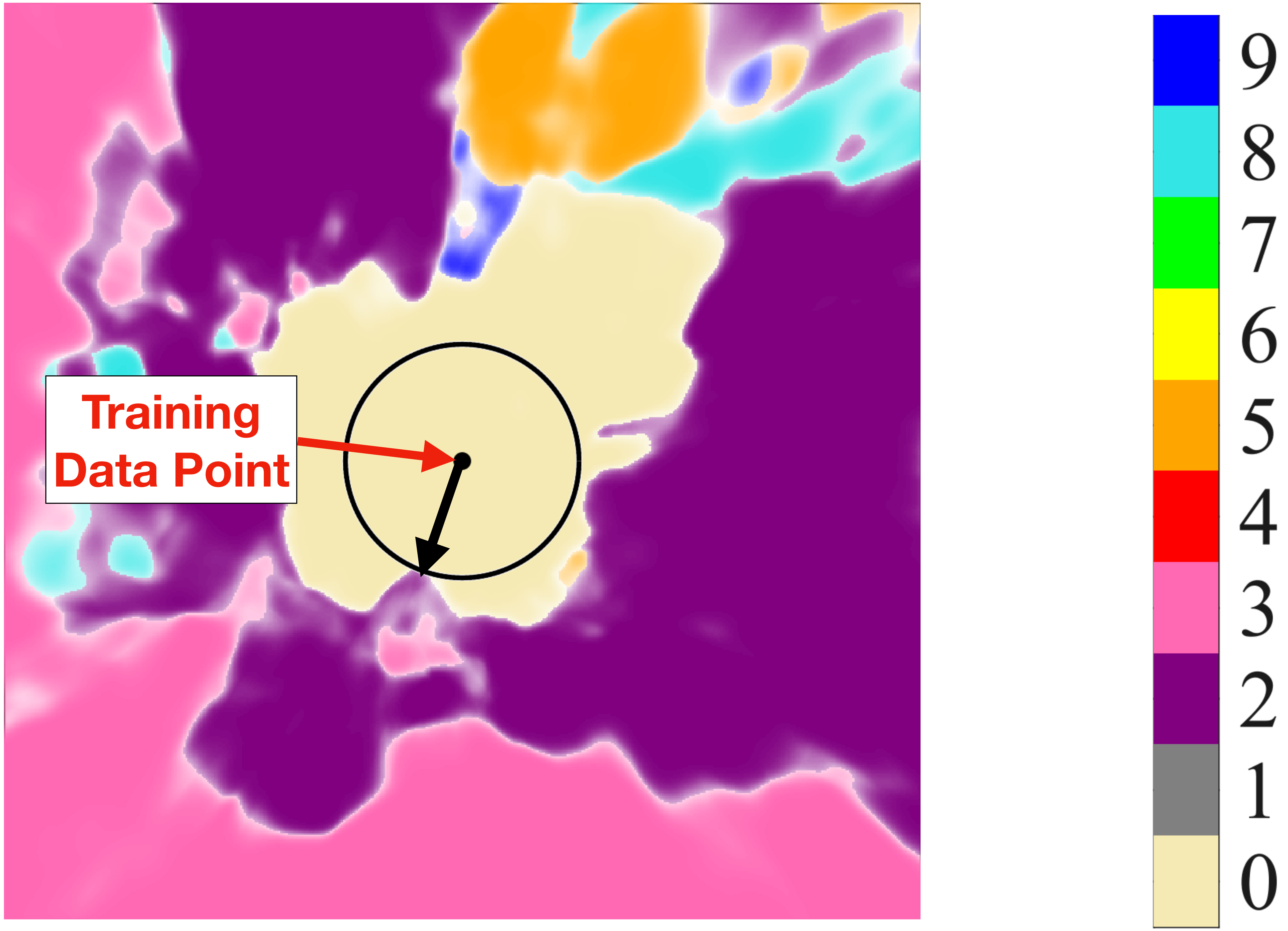
Visualize Perturbation Space



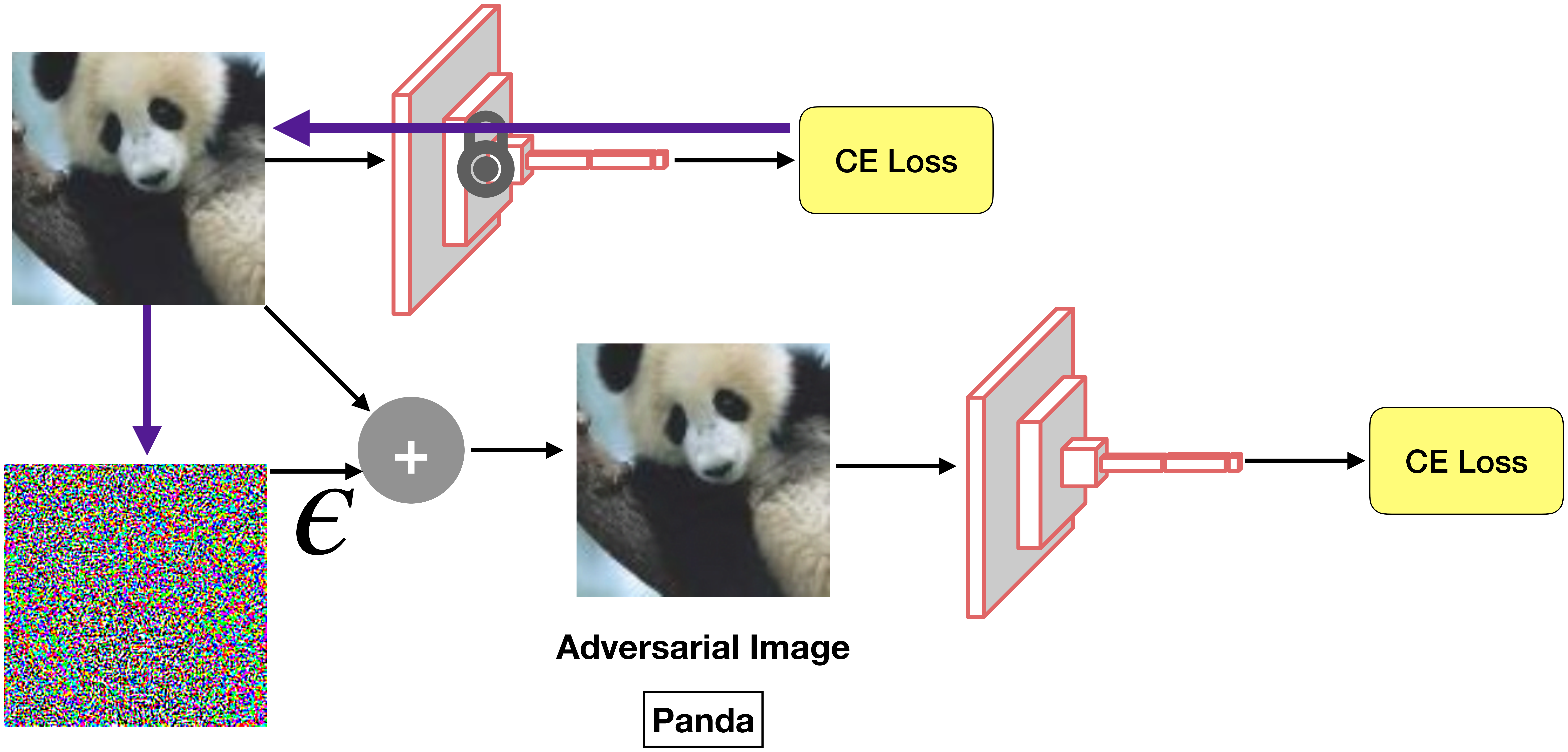
MNIST LeNet Decisions Around Training Point

Non-smooth Decision Boundary

Small perturbations lead to new outputs

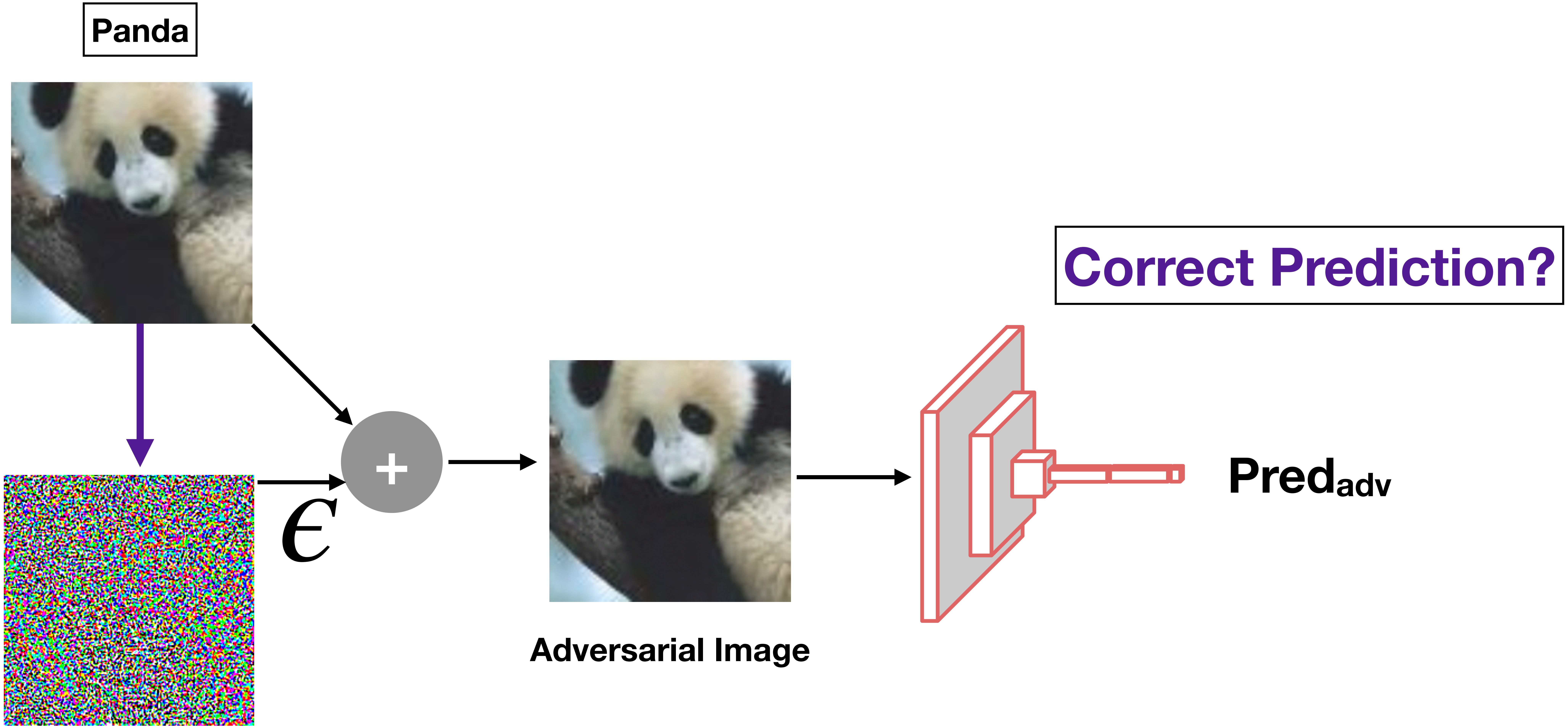


Adversarial Training

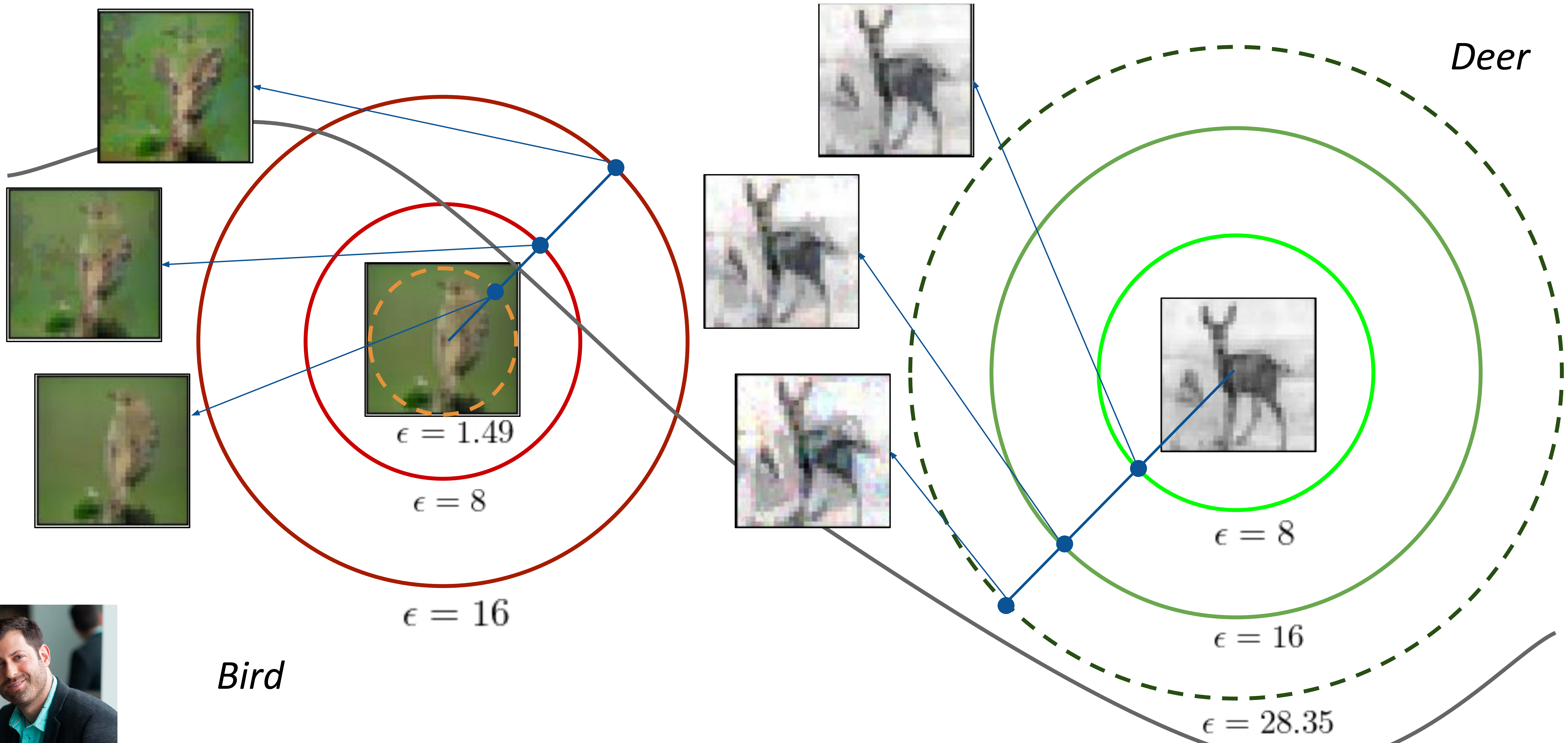


[Madry et. al. ICLR 18]

Adversarial Stability

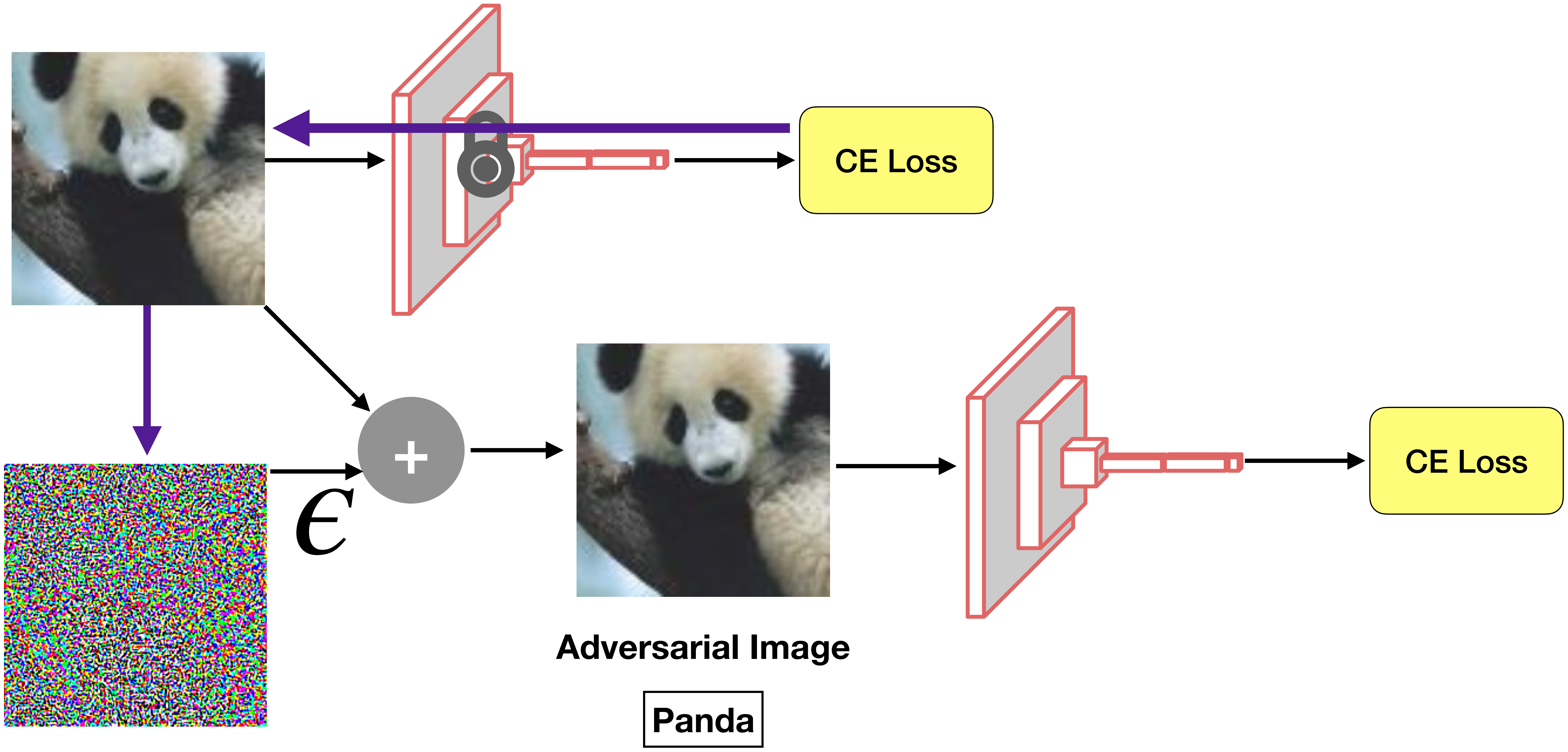


Instance Adaptive Adversarial Training



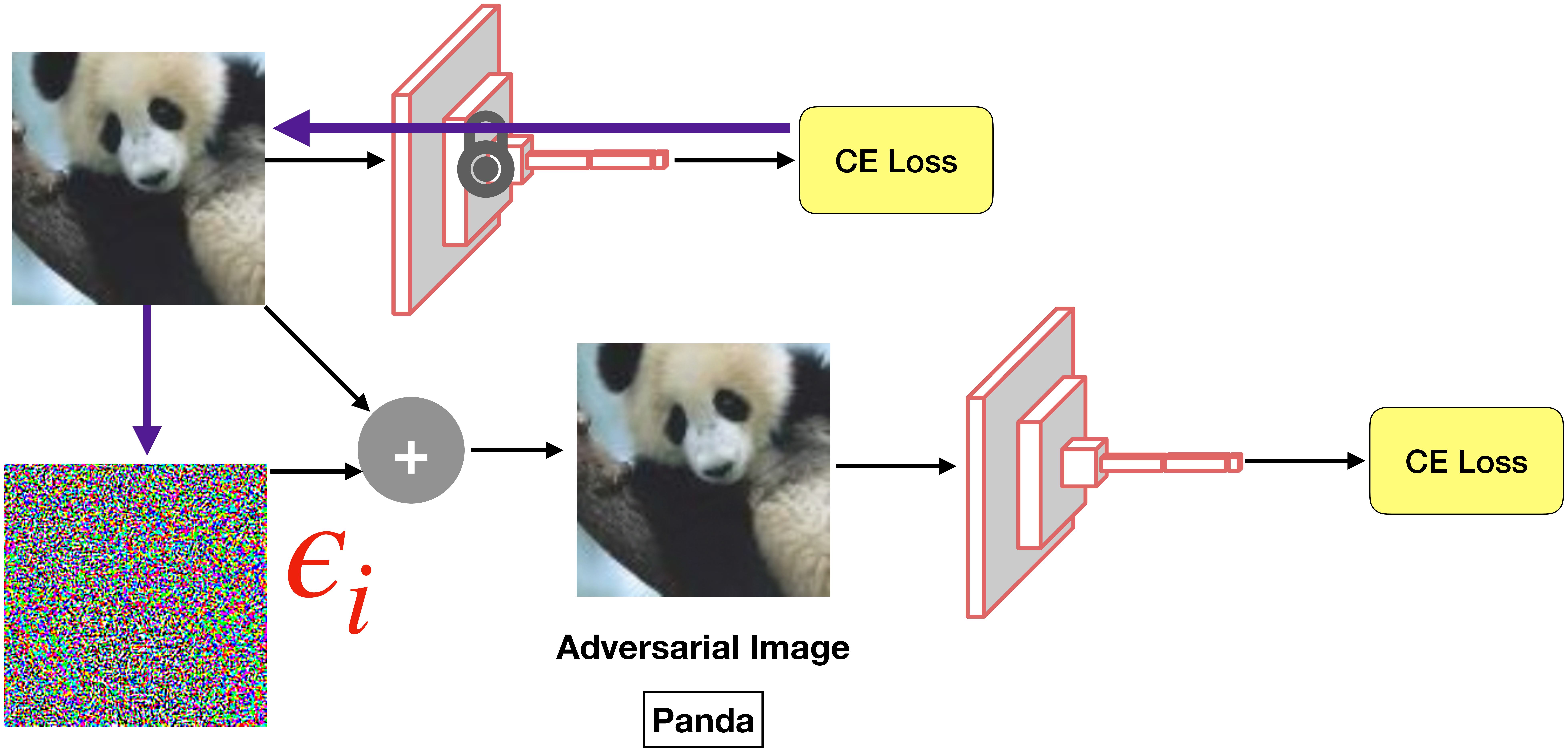
Balaji, Goldstein, Hoffman. arXiv 2020.

Adversarial Training



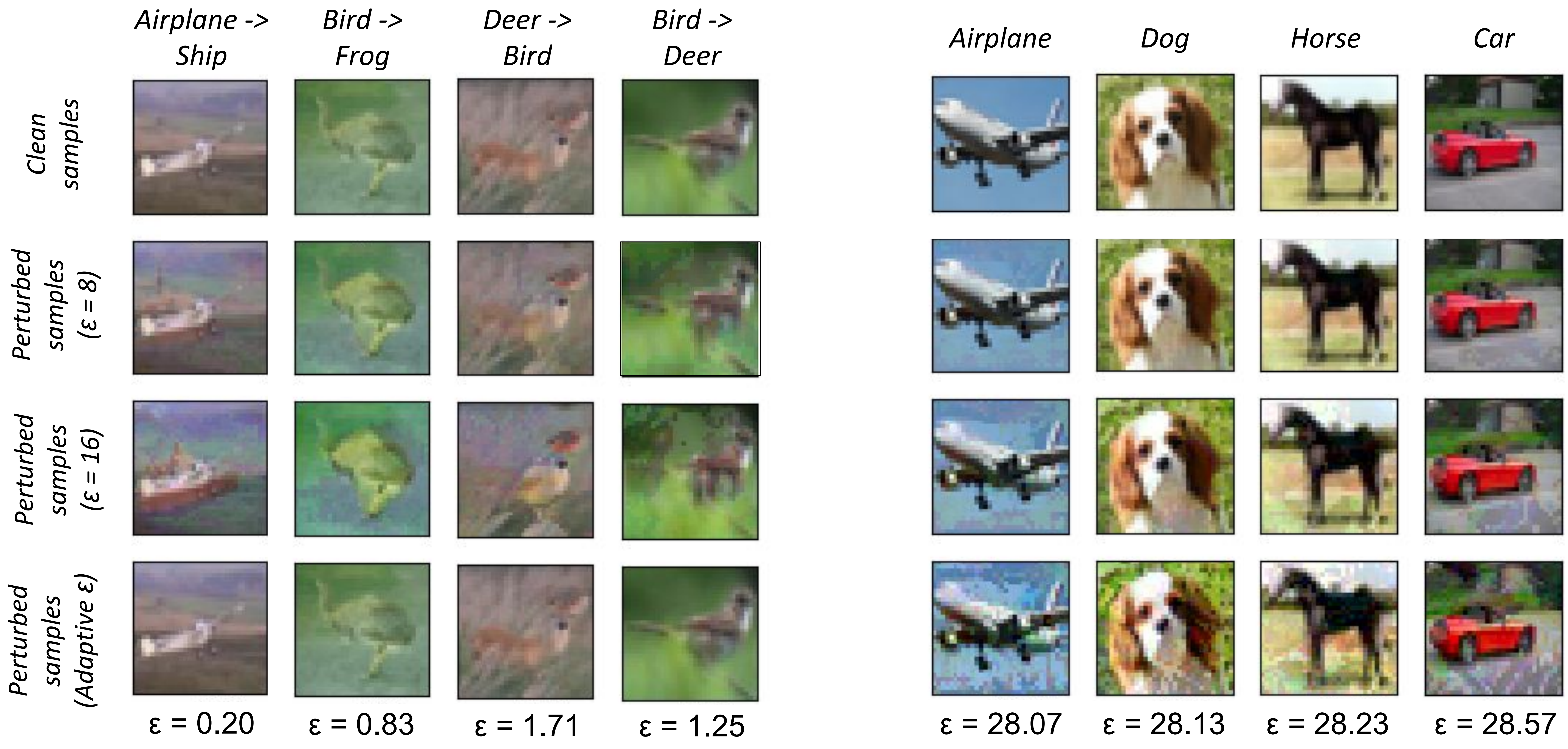
[Madry et. al. ICLR 18]

Instance Adaptive Adversarial Training



Balaji, Goldstein, Hoffman. arXiv 2020.

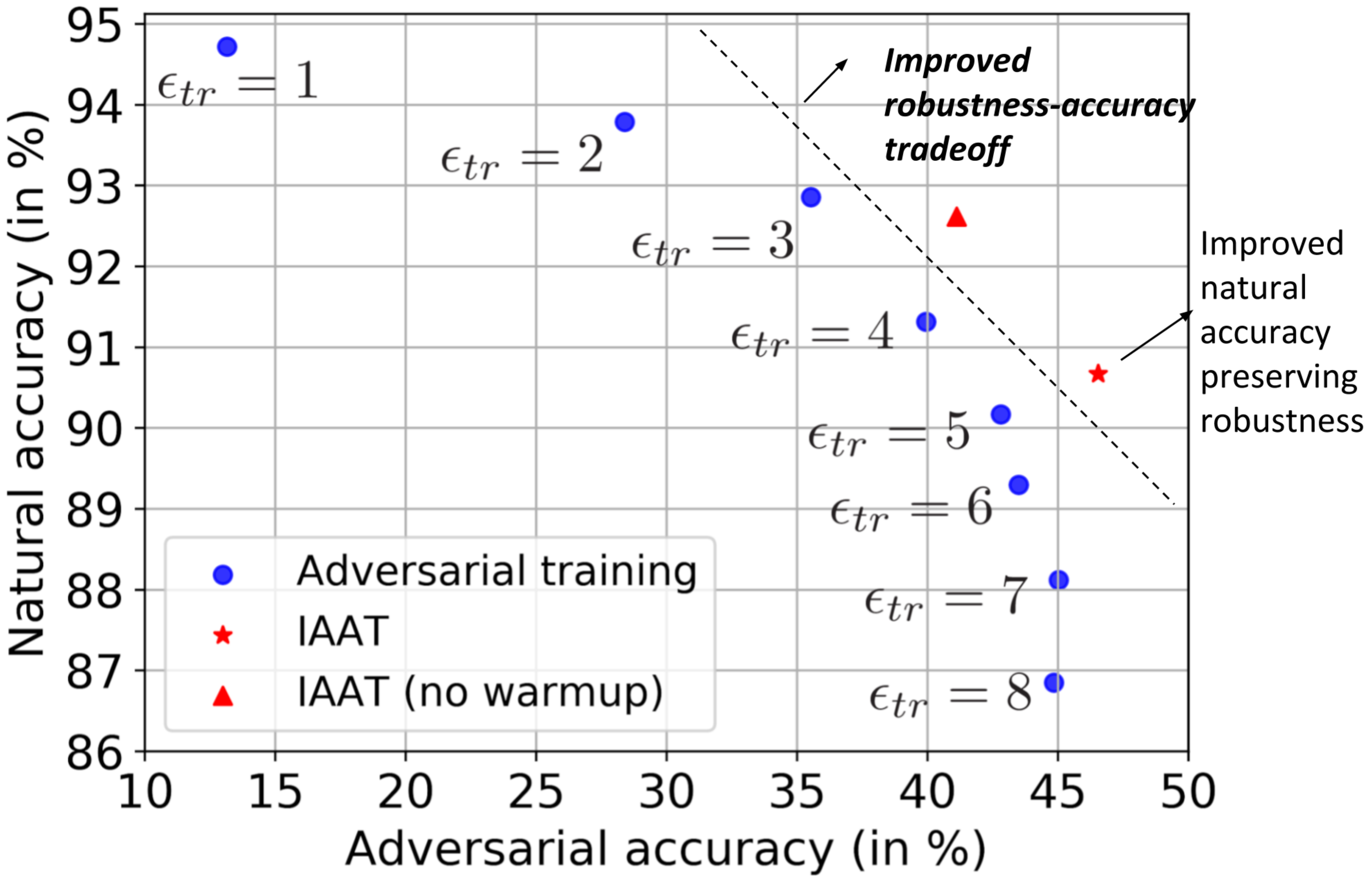
Instance Adaptive Adversarial Training



(a) Samples from bottom 1% ϵ (b) Samples from top 1% ϵ

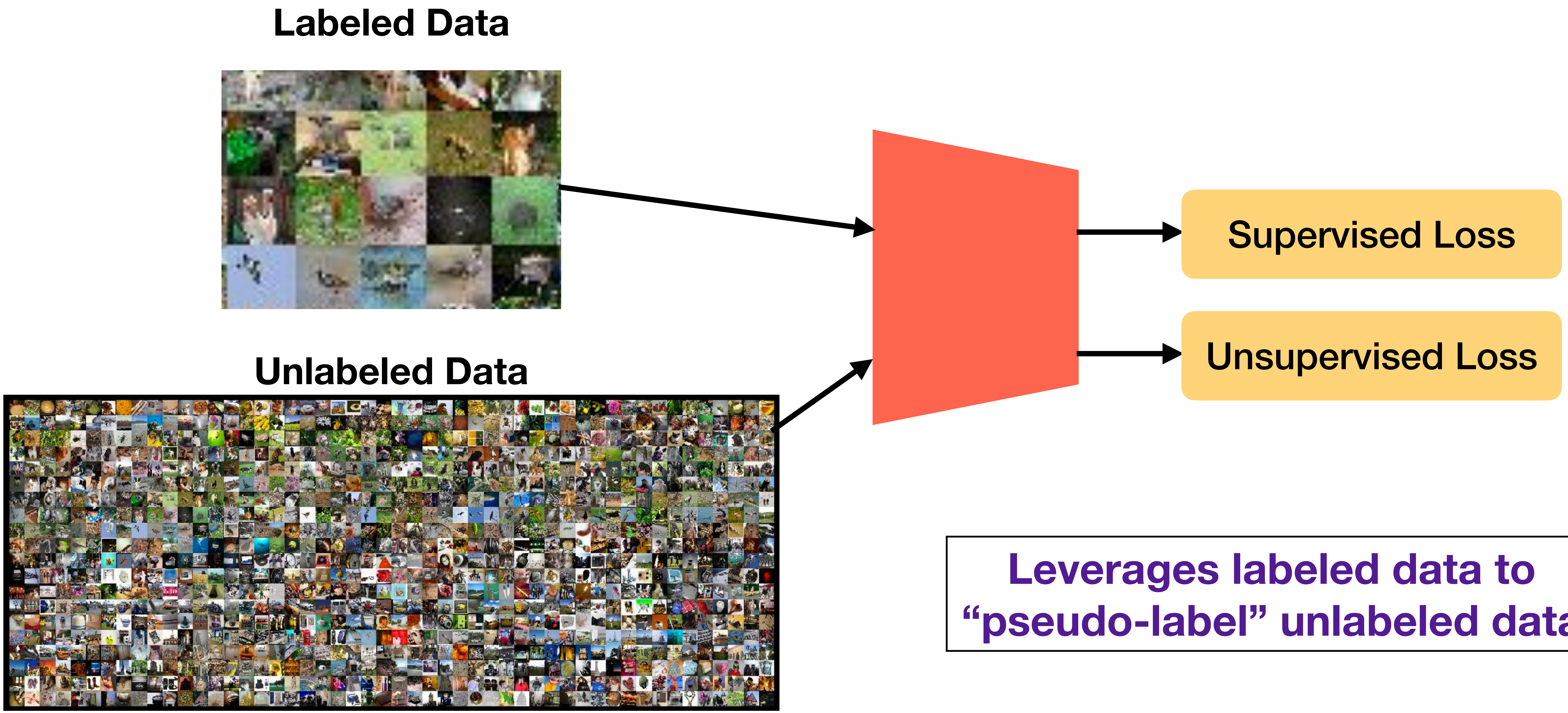
Balaji, Goldstein, Hoffman. arXiv 2020.

Adaptive Adversarial Training: CIFAR-10



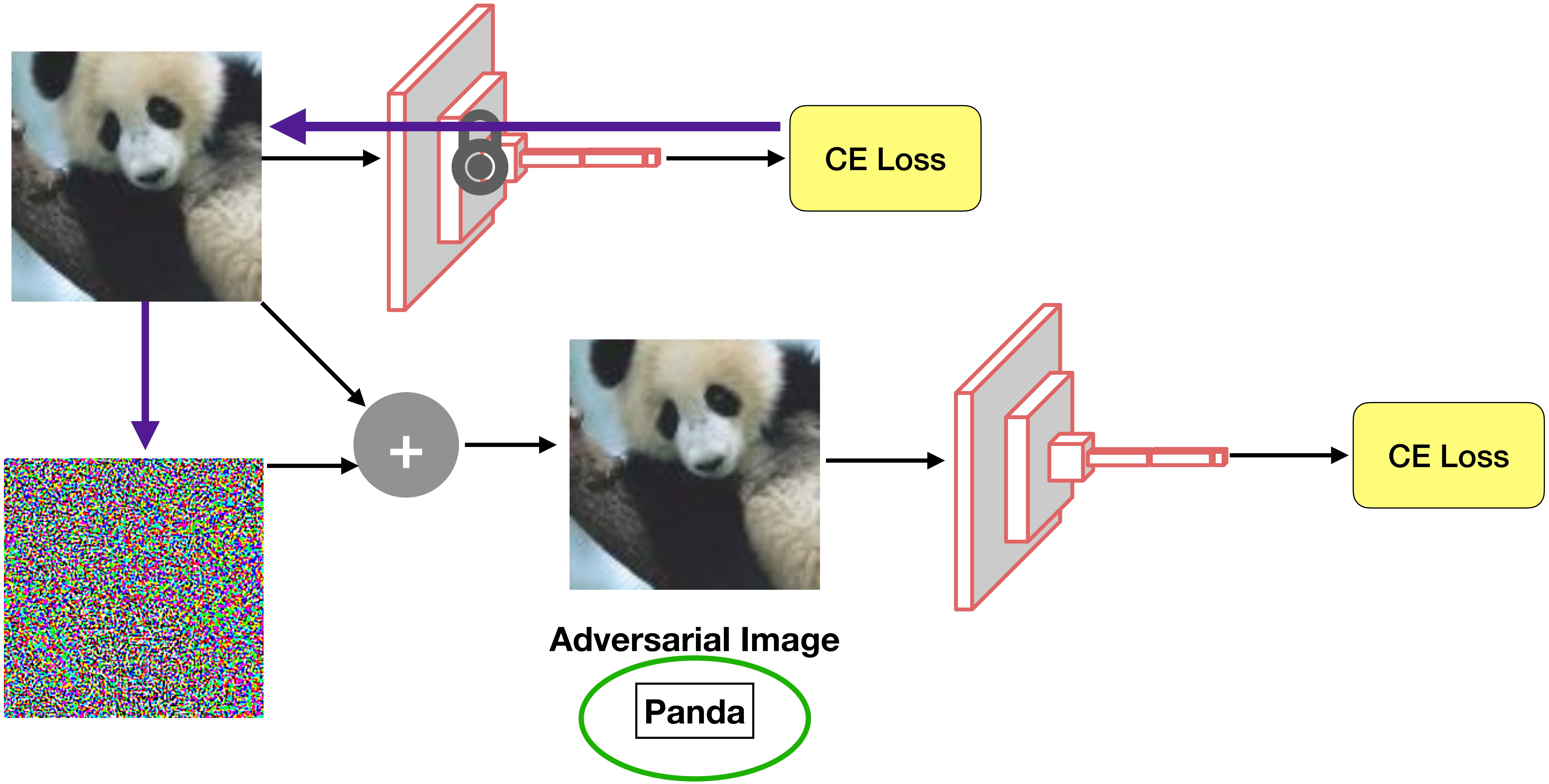
Balaji, Goldstein, Hoffman. arXiv 2020.

Semi-supervised Learning



Leverages labeled data to "pseudo-label" unlabeled data

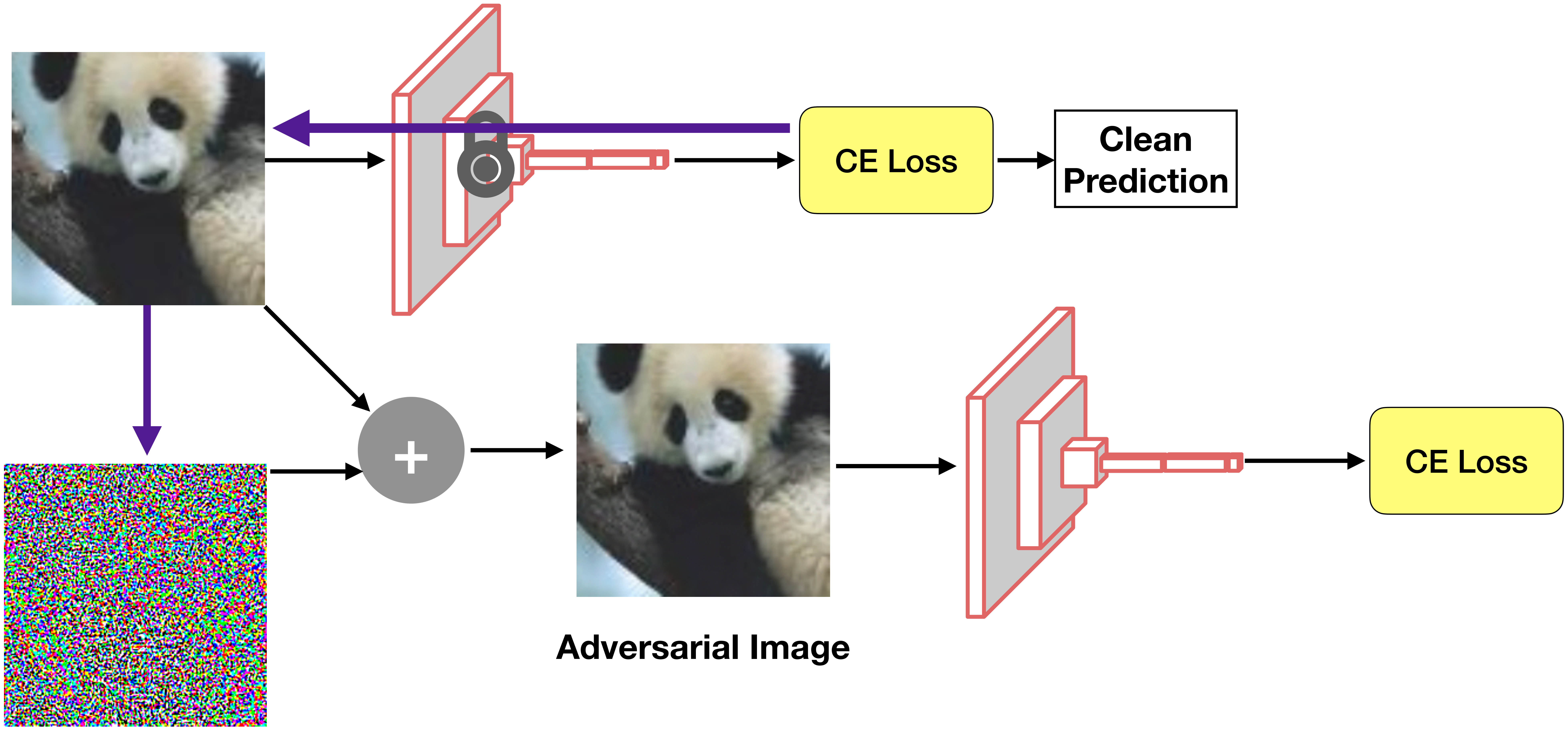
Adversarial Training



What if the label is unknown?

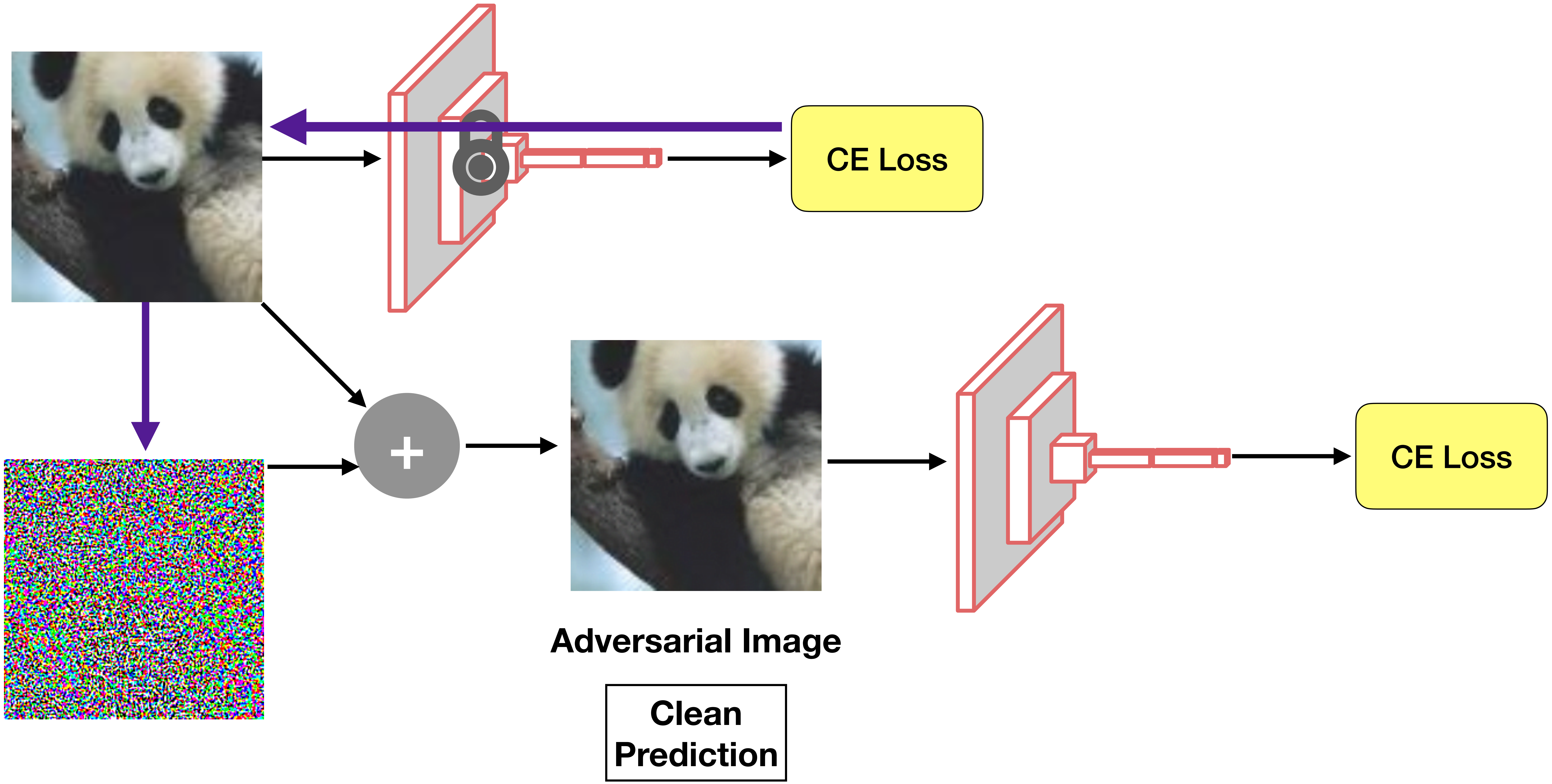
[Madry et. al. ICLR 18]

Virtual Adversarial Training



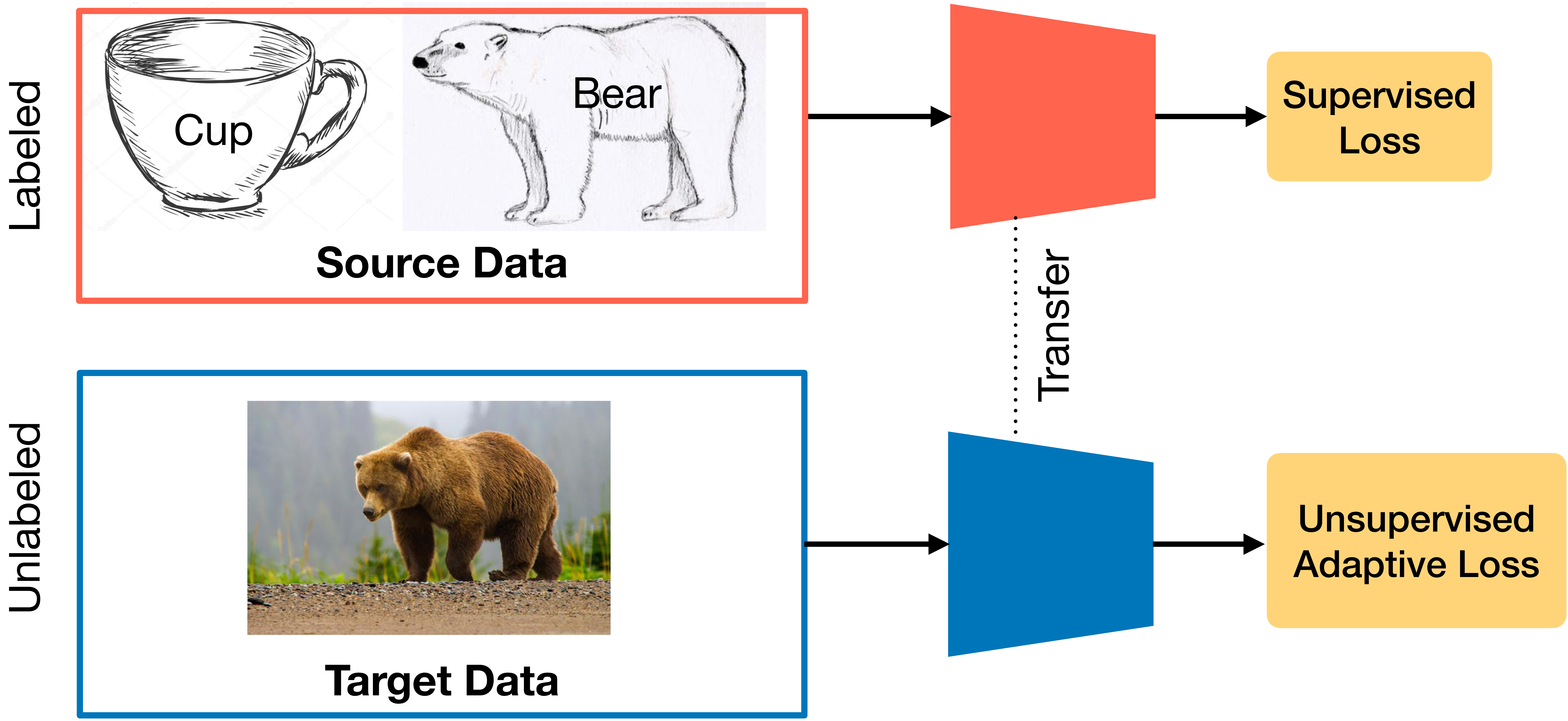
[Miyato et. al. ICLR 2016]

Virtual Adversarial Training

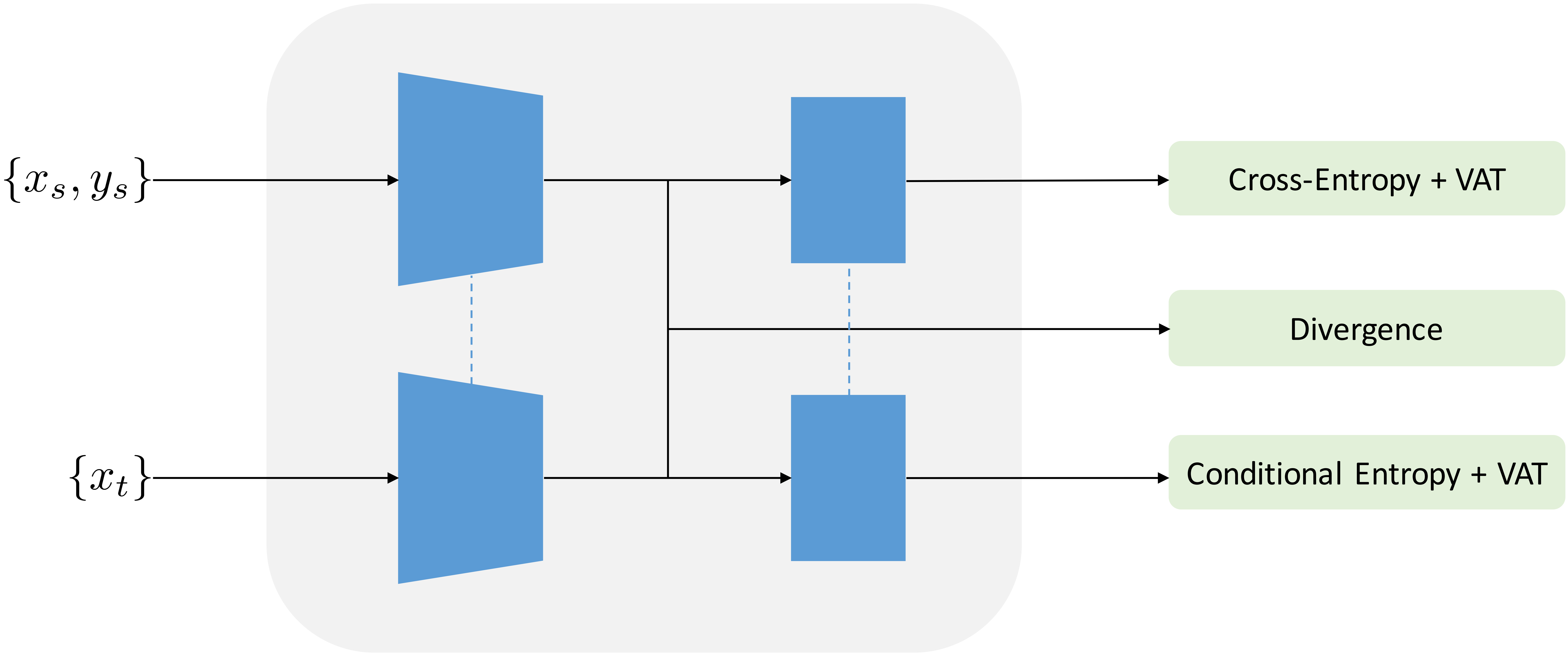


[Miyato et. al. ICLR 2016]

Unsupervised Domain Adaptation



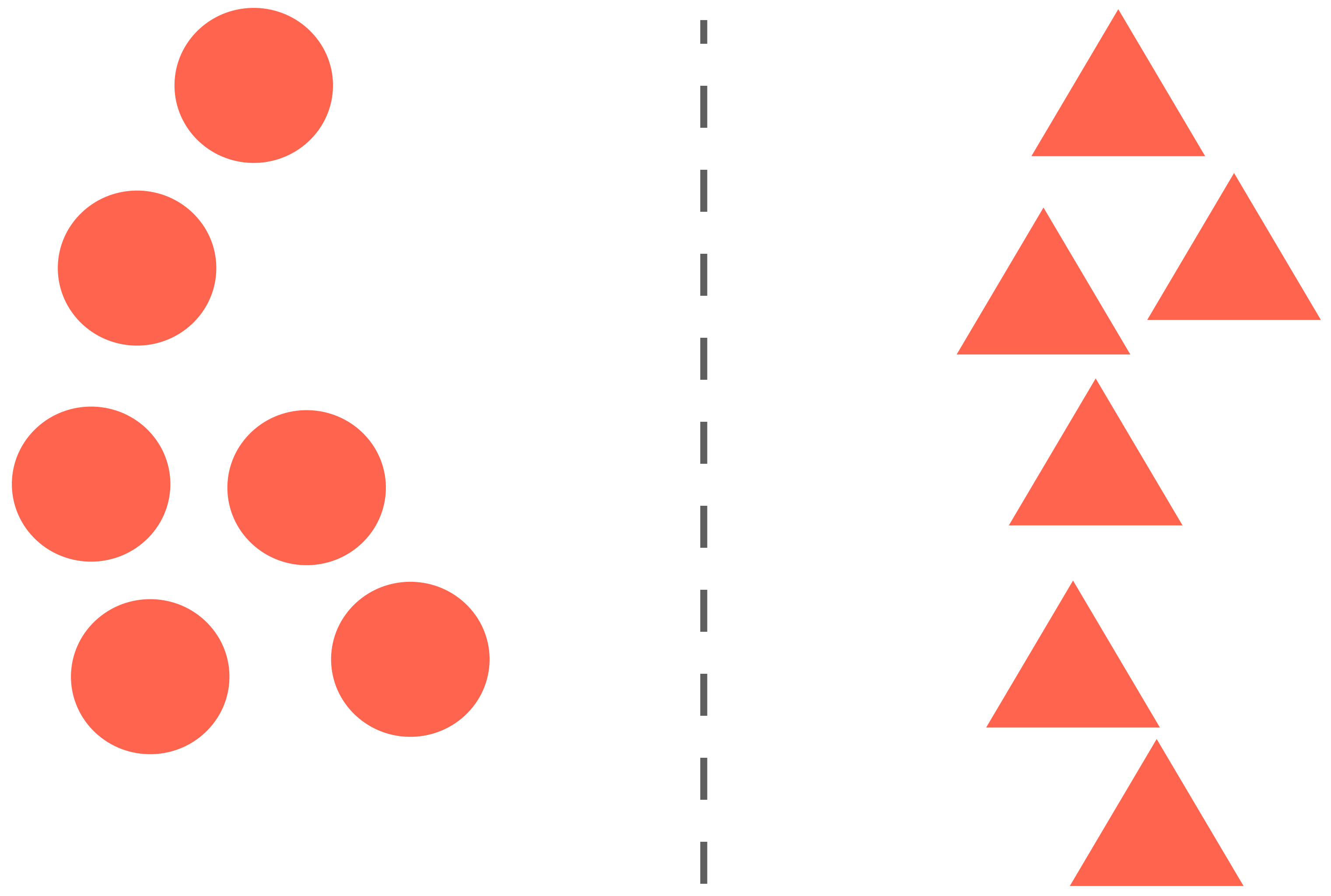
DIRT-T / VADA



Shu et. al. ICLR 2018

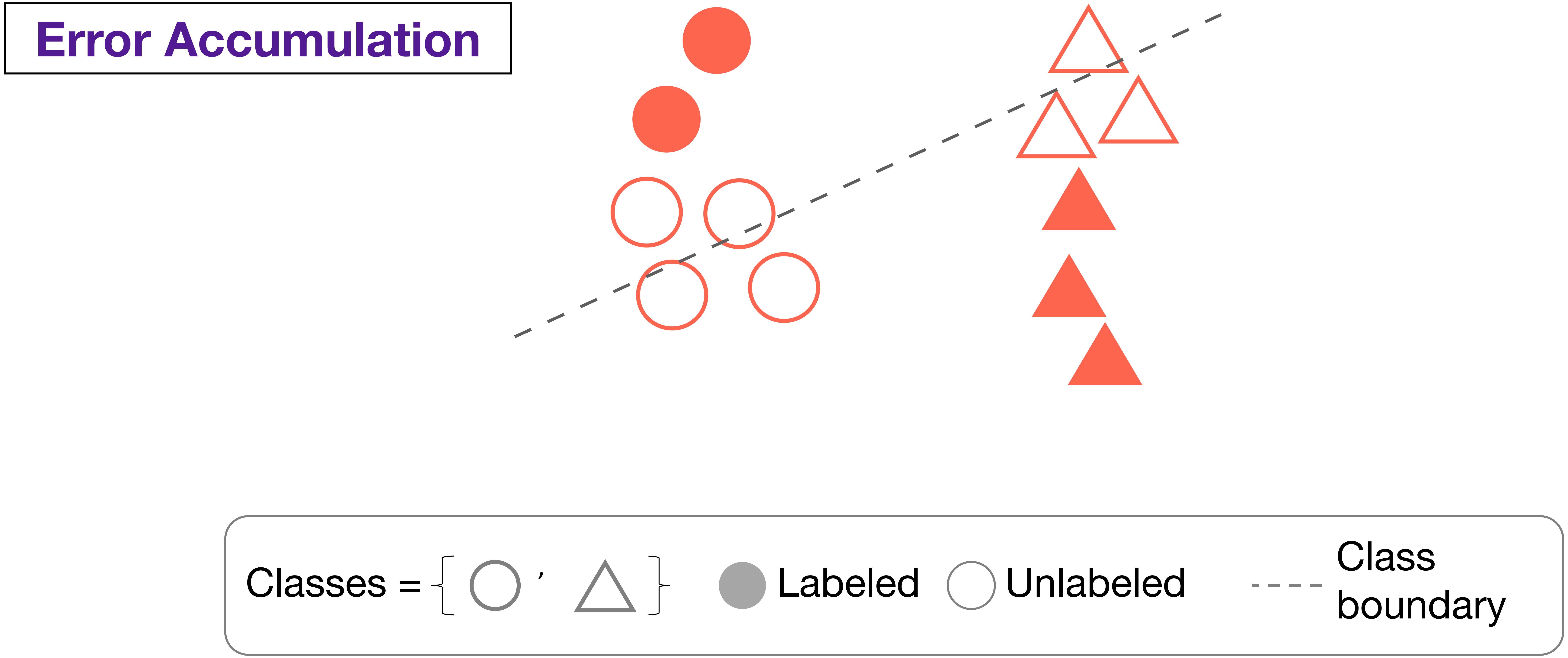
Example: Entropy Minimization

Supervised Decision Boundary



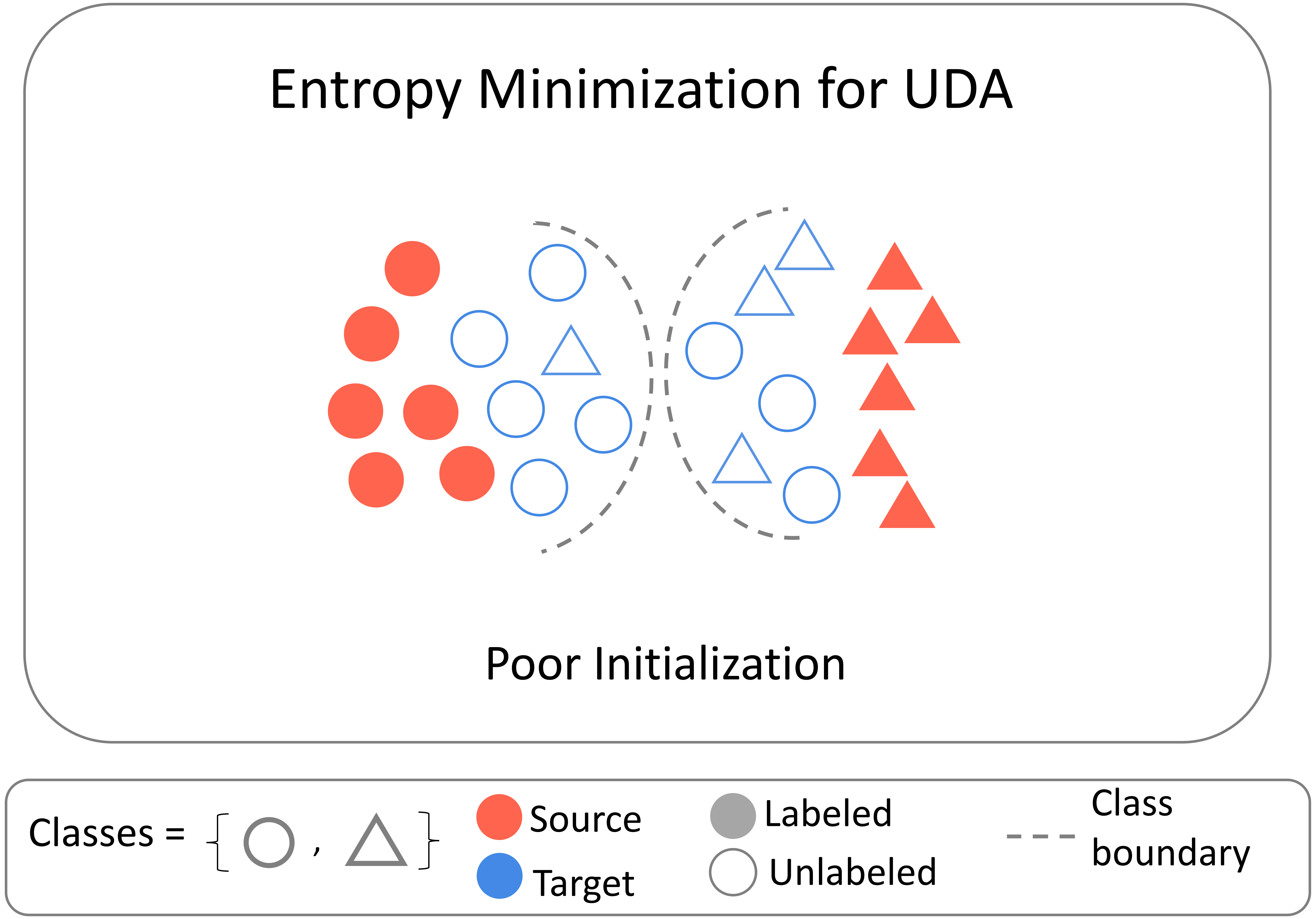
Classes = { ○ , △ }   ● Labeled   ○ Unlabeled   - - - - Class boundary

Example: Entropy Minimization



Self-Training with Unreliable Instances

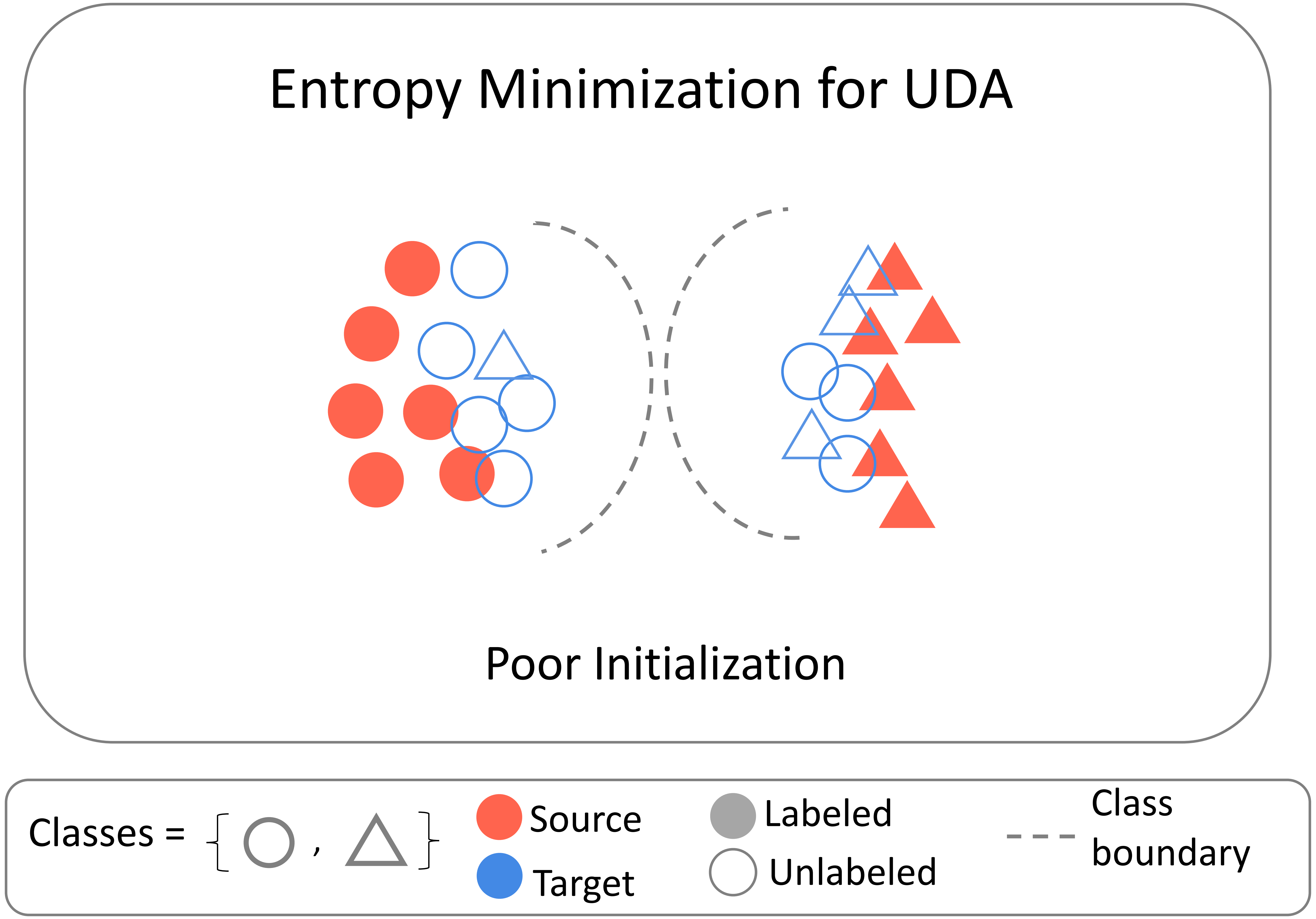
- Under a domain shift, some target categories may be misaligned
- Entropy minimization on such instances would increase model confidence, reinforcing errors



$$\mathcal{L}_{CEM} = \mathbb{E}_{\mathbf{x}_\mathcal{T} \sim \mathcal{P}_\mathcal{T}}[\mathcal{H}_\Theta(y \mid \mathbf{x}_\mathcal{T})]$$
$$= \mathbb{E}_{\mathbf{x}_\mathcal{T} \sim \mathcal{P}_\mathcal{T}}\left[\sum_{c=1}^{C} -p_\Theta(y=c \mid \mathbf{x}_\mathcal{T}) \log p_\Theta(y=c \mid \mathbf{x}_\mathcal{T})\right]$$

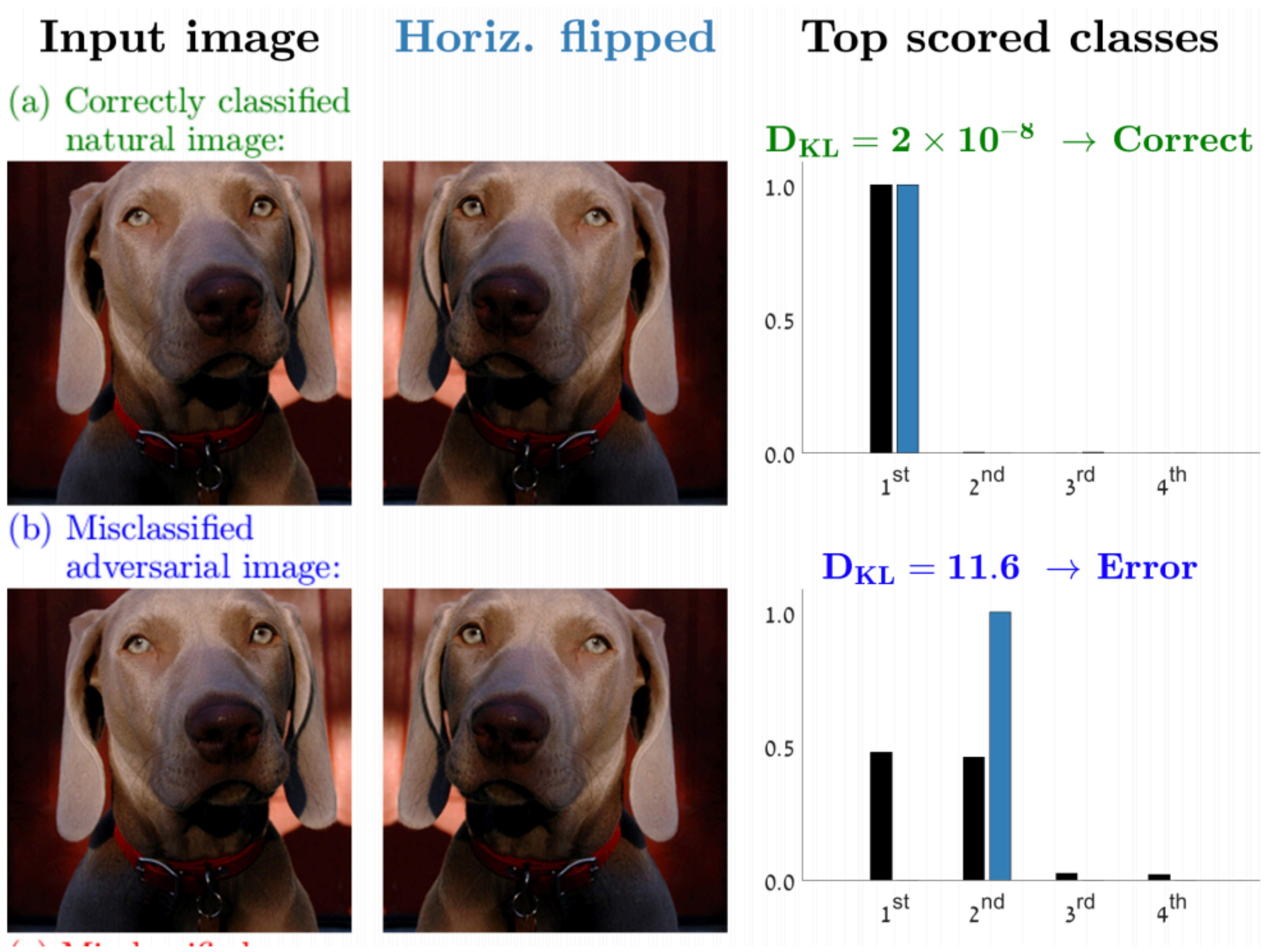
Self-Training with Unreliable Instances

- Under a domain shift, some target categories may be misaligned
- Entropy minimization on such instances would increase model confidence, reinforcing errors

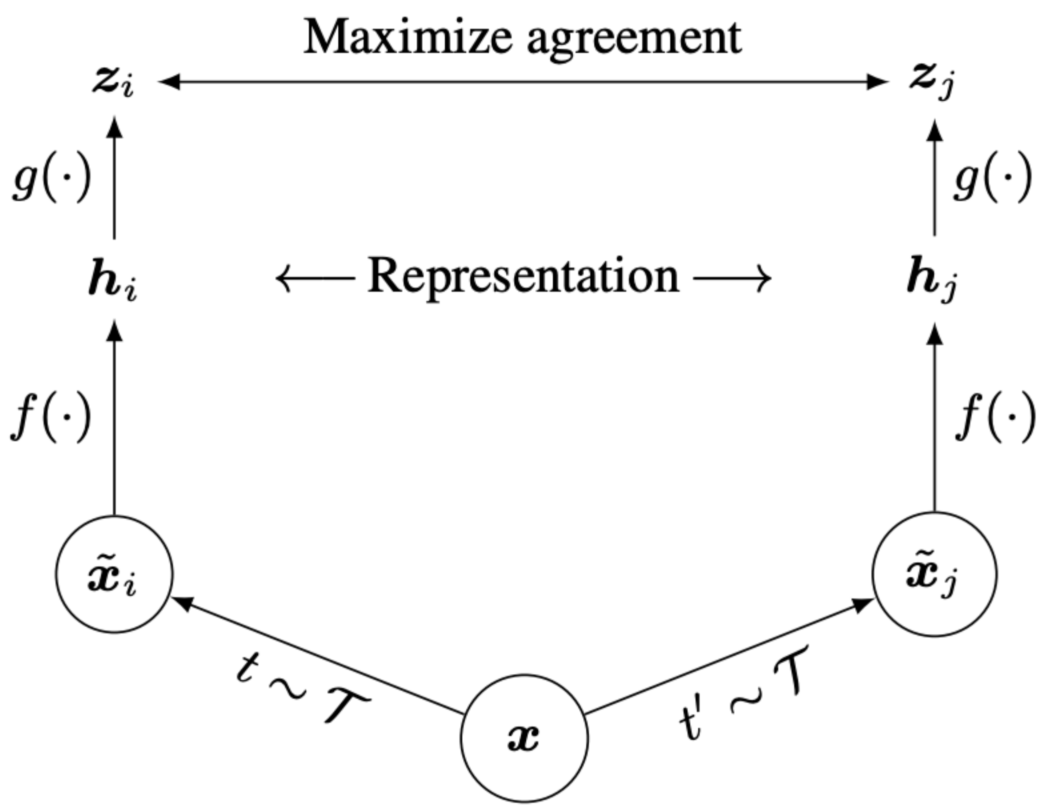


$$\mathcal{L}_{CEM} = \mathbb{E}_{\mathbf{x}_\mathcal{T} \sim \mathcal{P}_\mathcal{T}}[\mathcal{H}_\Theta(y \mid \mathbf{x}_\mathcal{T})]$$
$$= \mathbb{E}_{\mathbf{x}_\mathcal{T} \sim \mathcal{P}_\mathcal{T}}\left[\sum_{c=1}^{C} -p_\Theta(y=c \mid \mathbf{x}_\mathcal{T}) \log p_\Theta(y=c \mid \mathbf{x}_\mathcal{T})\right]$$

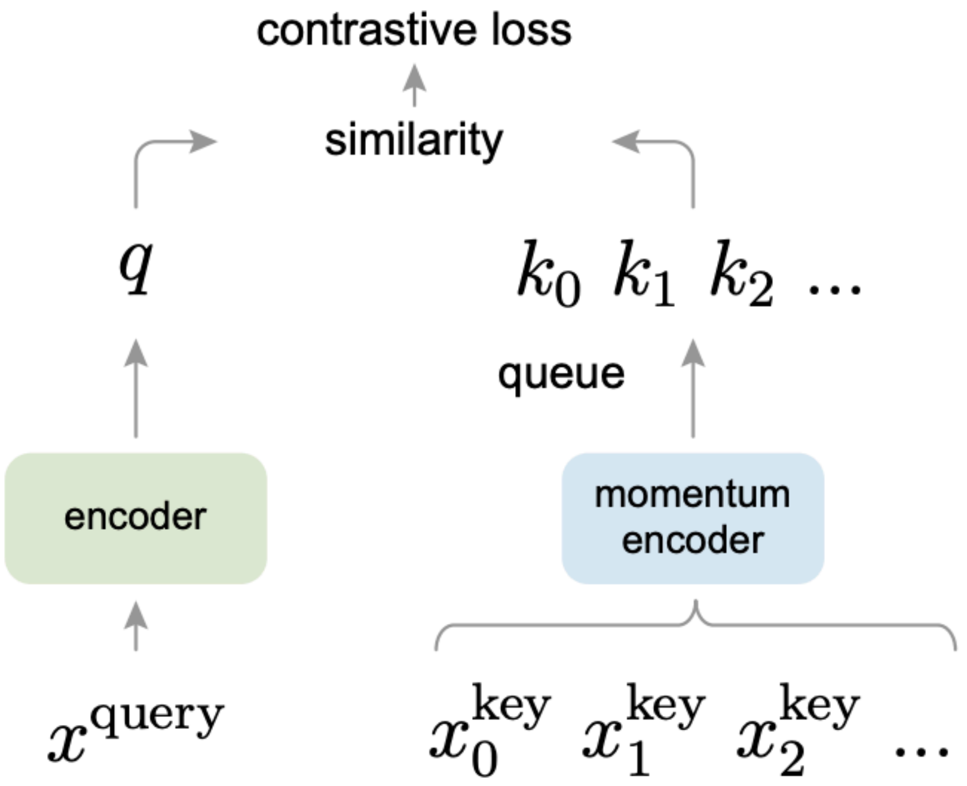
Prior Work: Measure Image Aug Differences



Natural and Adversarial Error Detection using Invariance to Image Transformations. Bahat, Irani, Shakhnarovich, arXiv 2019



SimCLR, Chen et al.
ICML 2020



MoCo, He et al.
CVPR 2020

Detecting Errors

Learned Invariance
(Contrastive Learning)

SENTRY

Selective Entropy Optimization via Committee Consistency for Unsupervised Domain Adaptation

Viraj Prabhu

Shivam Khare

Deeksha Karthik

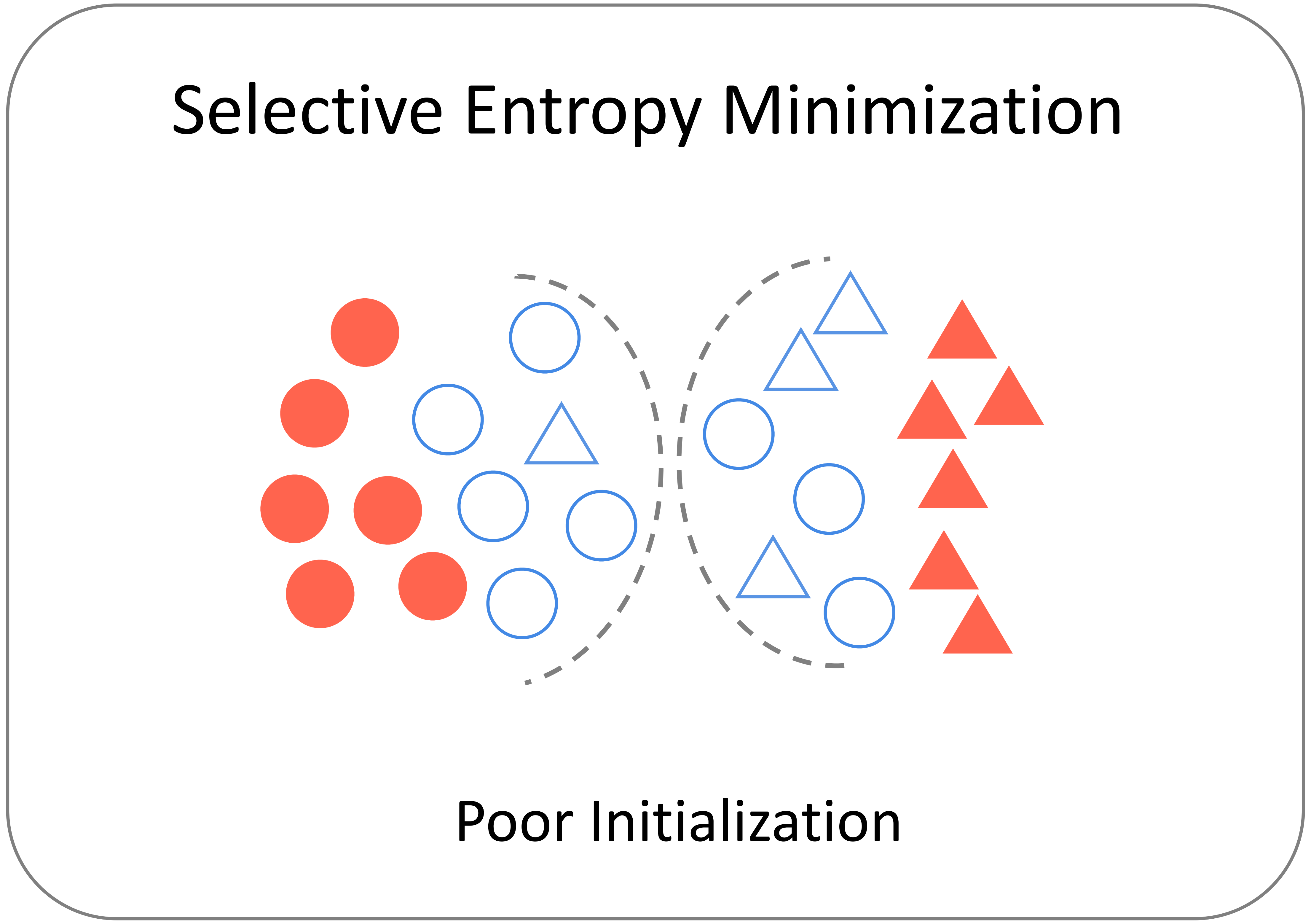
Judy Hoffman

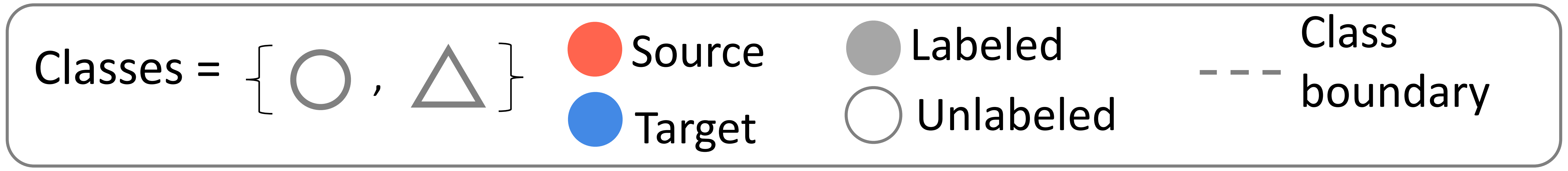
Georgia Tech

SENTRY: Selective Entropy Optimization via Committee Consistency

Key Idea

1. Identify reliable instances using predictive consistency[1,2,3]
   - Model confidence is known to be uncalibrated under distribution shift [Ovadia NeurIPS 2019]
2. Increase model confidence on highly consistent target instances, reduce on inconsistent





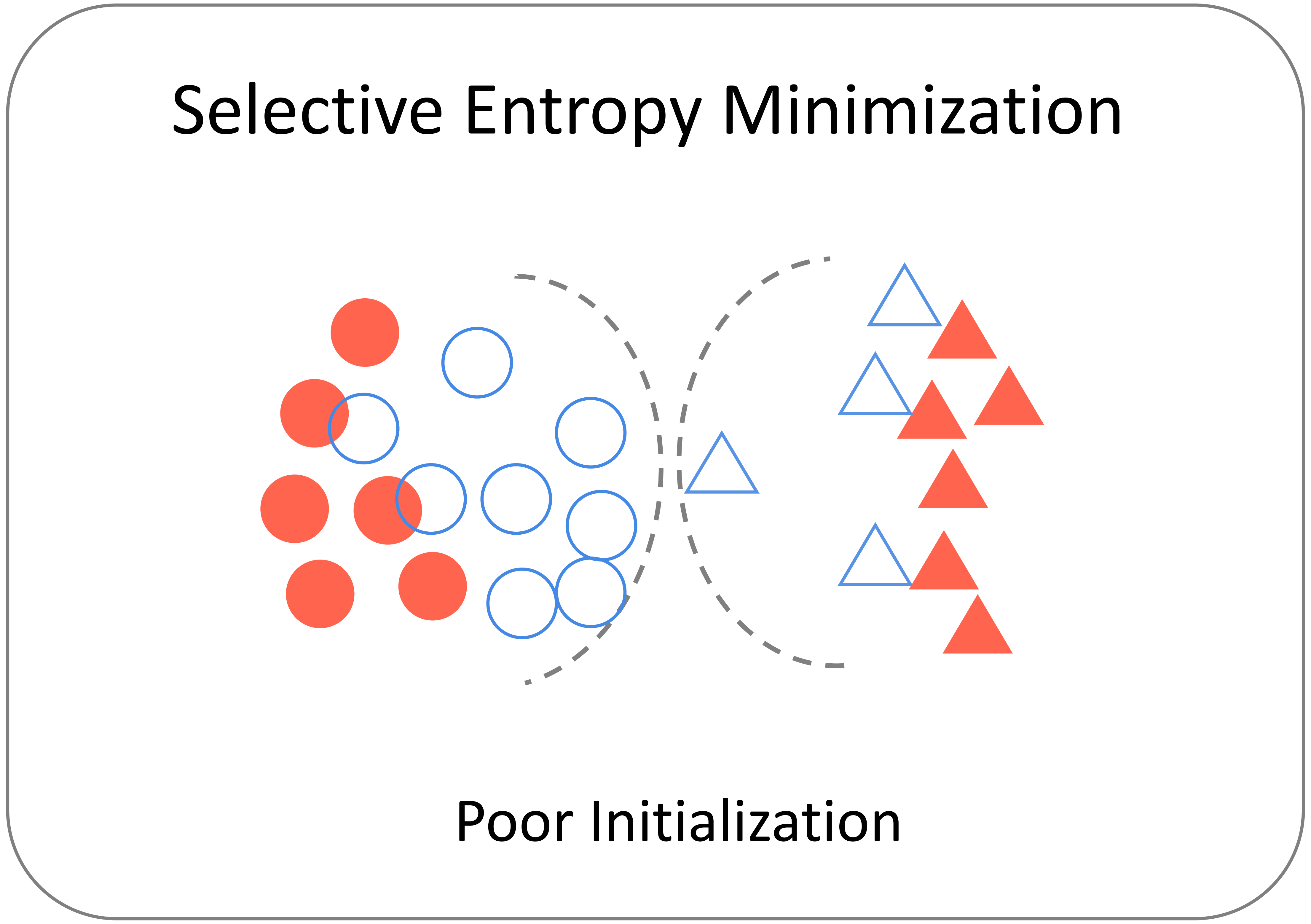
1. Bahat et al. arXiv 2019.
2. Chen et al. ICML 2020.
3. Sohn et al., NeurIPS 2020.

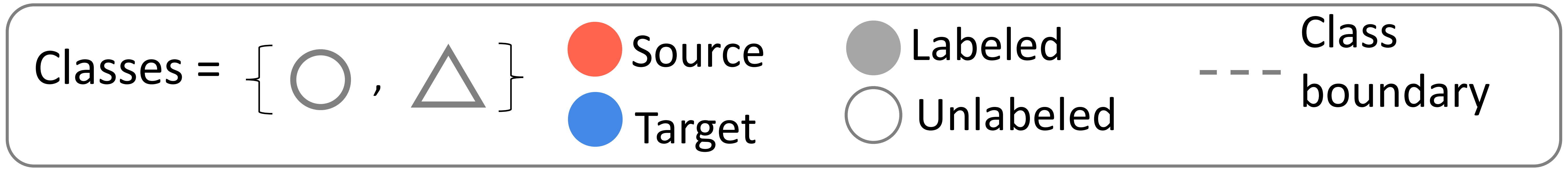
Prabhu, Khare, Karthik, Hoffman. arXiv:2012.11460, 2021

SENTRY: Selective Entropy Optimization via Committee Consistency

Key Idea

1. Identify reliable instances using predictive consistency[1,2,3]
   - Model confidence is known to be uncalibrated under distribution shift [Ovadia NeurIPS 2019]
2. Increase model confidence on highly consistent target instances, reduce on inconsistent



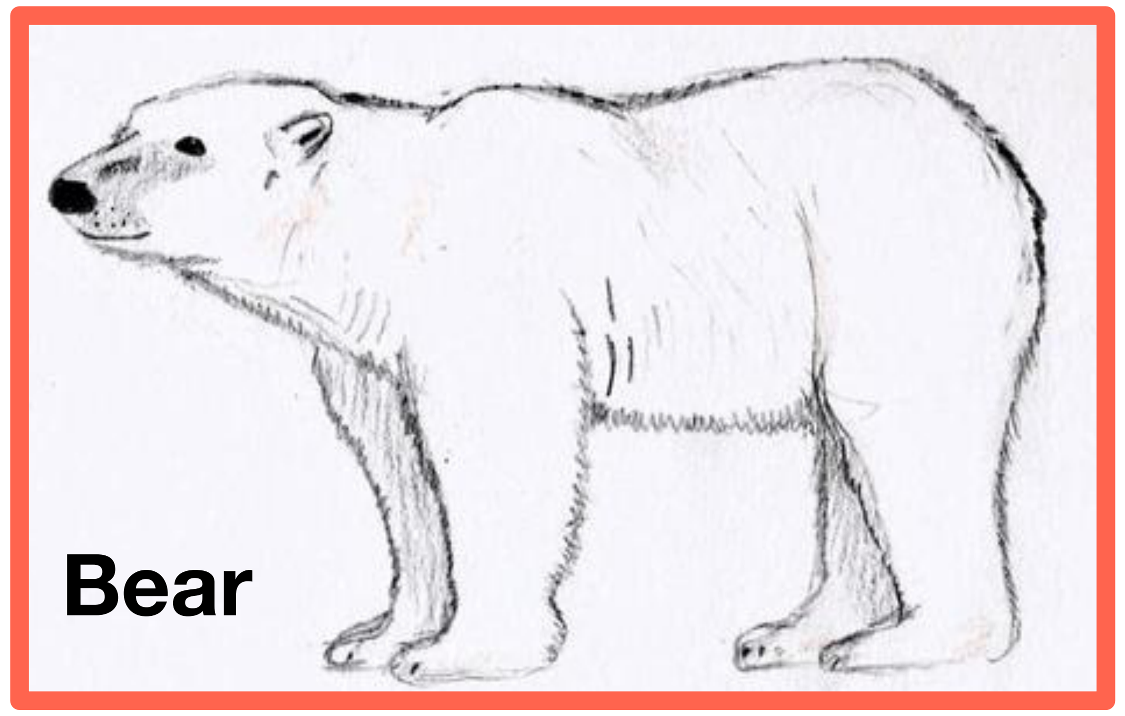


1. Bahat et al. arXiv 2019.
2. Chen et al. ICML 2020.
3. Sohn et al., NeurIPS 2020.

Prabhu, Khare, Karthik, Hoffman. arXiv:2012.11460, 2021

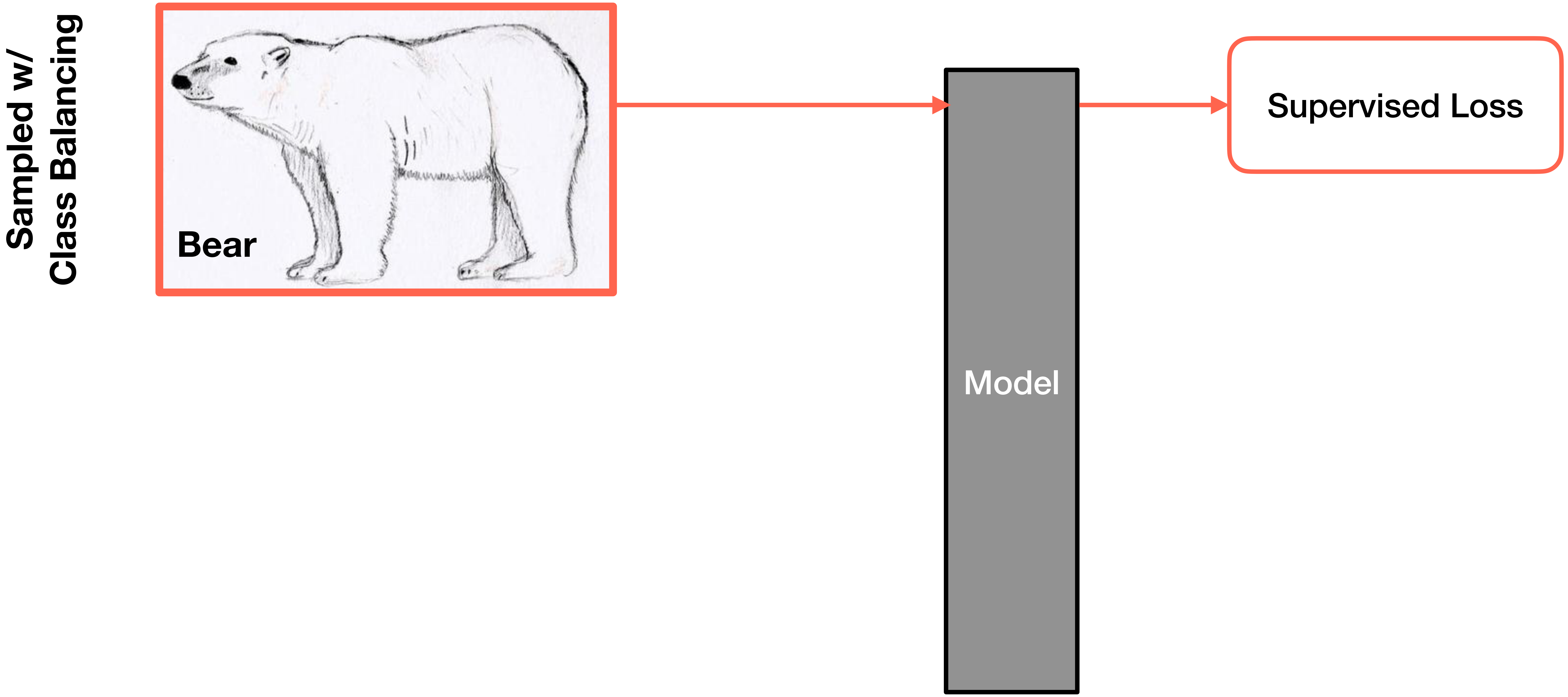
SENTRY: Selective Entropy Optimization via Committee Consistency

Sampled w/ Class Balancing



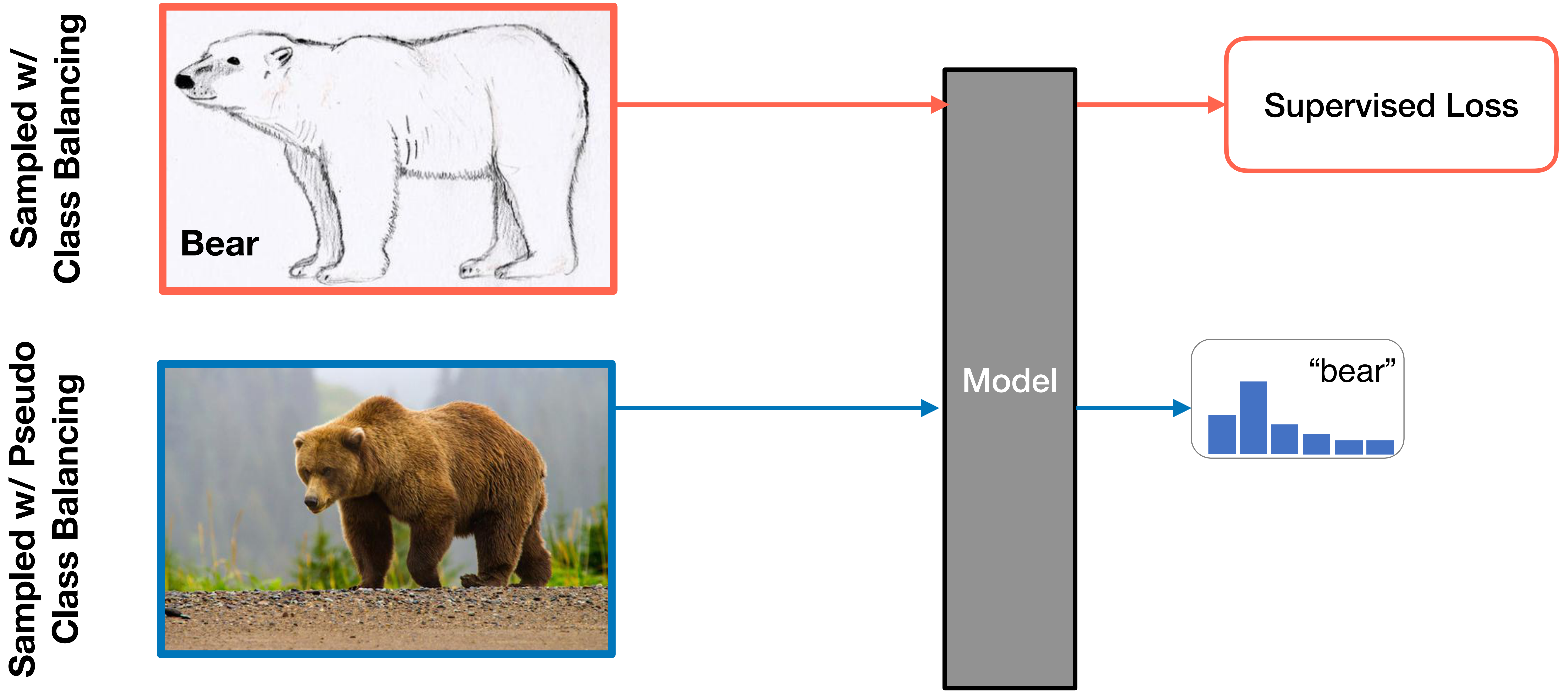
Prabhu, Khare, Karthik, Hoffman. arXiv, 2021

SENTRY: Selective Entropy Optimization via Committee Consistency



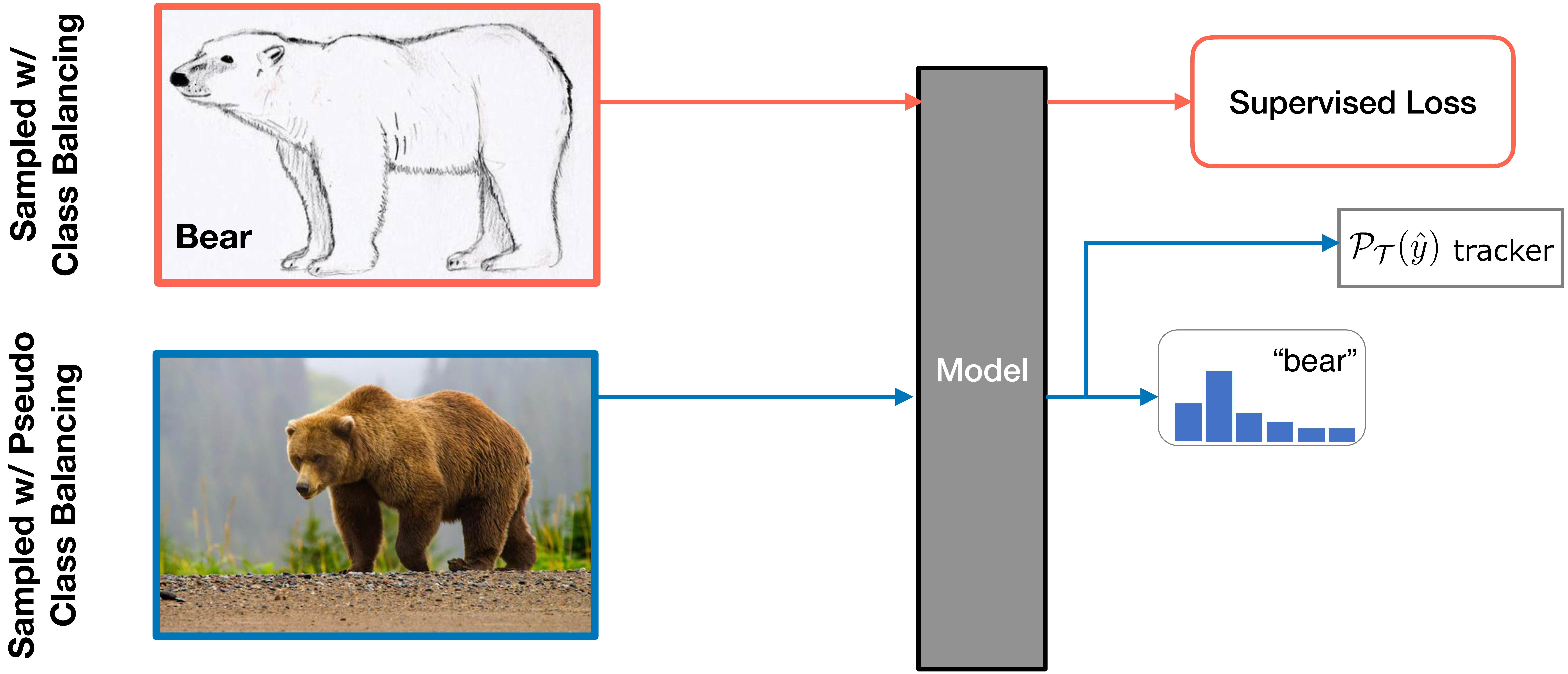
Prabhu, Khare, Karthik, Hoffman. arXiv, 2021

SENTRY: Selective Entropy Optimization via Committee Consistency



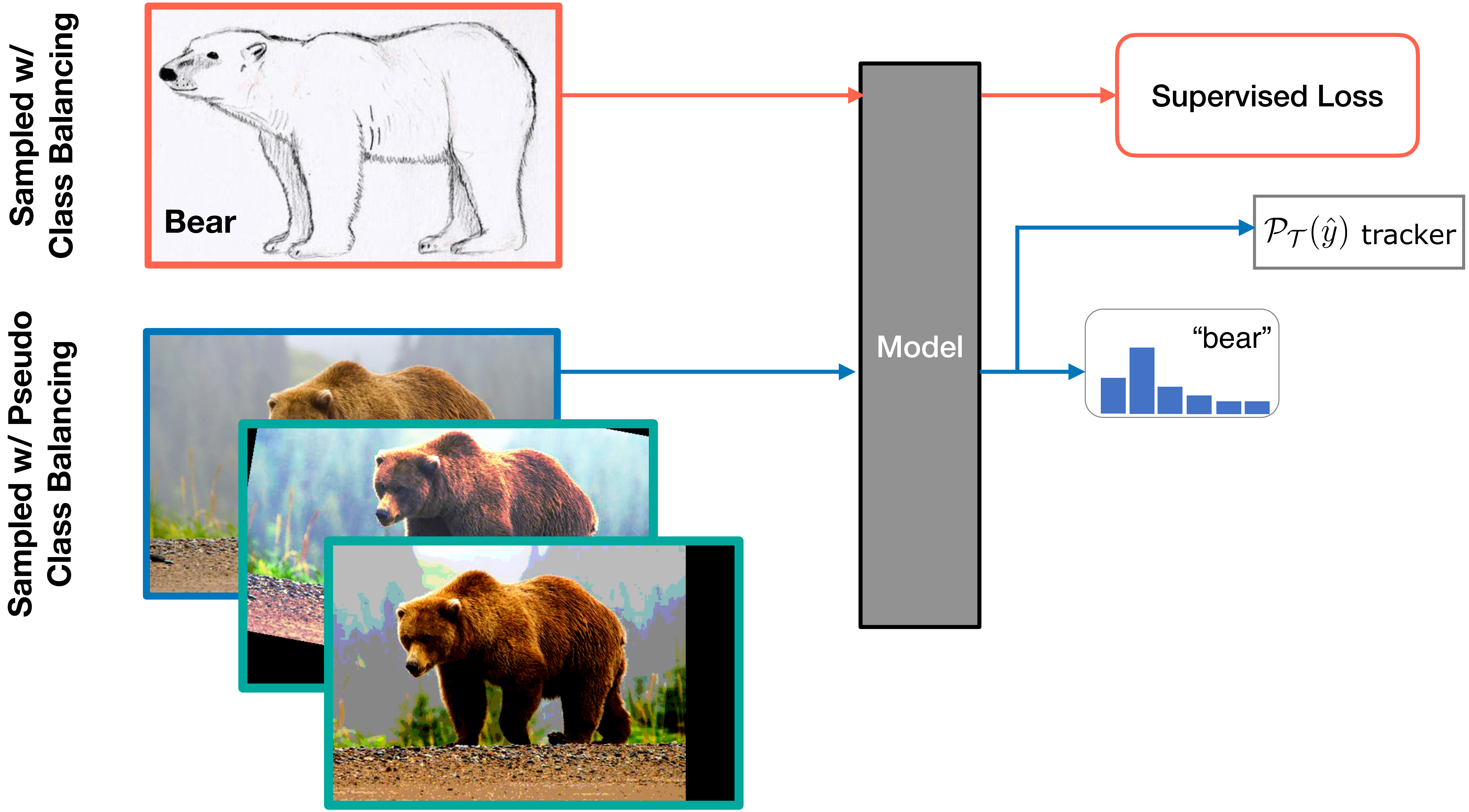
Prabhu, Khare, Karthik, Hoffman. arXiv, 2021

SENTRY: Selective Entropy Optimization via Committee Consistency



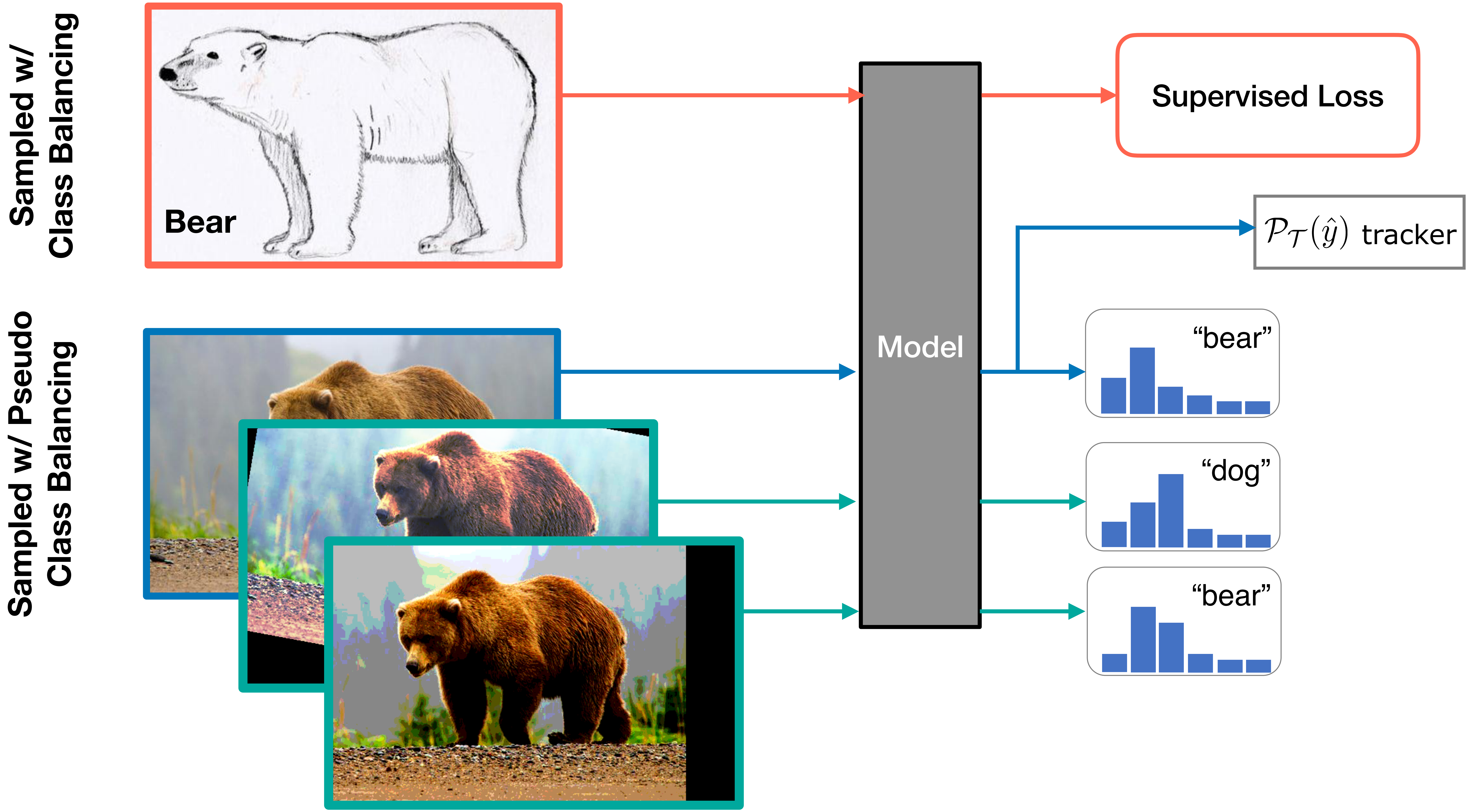
Prabhu, Khare, Karthik, Hoffman. arXiv, 2021

SENTRY: Selective Entropy Optimization via Committee Consistency



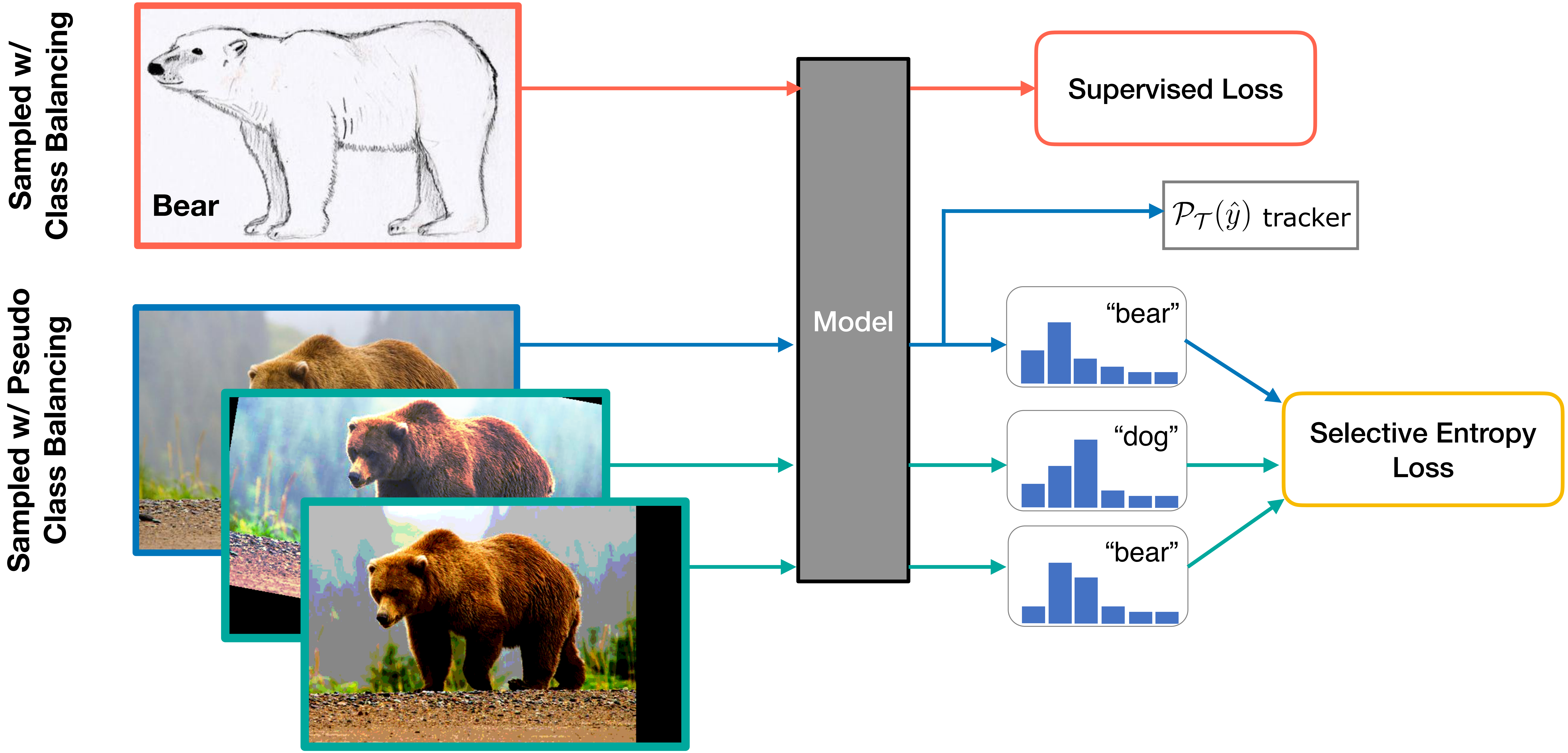
Prabhu, Khare, Karthik, Hoffman. arXiv, 2021

SENTRY: Selective Entropy Optimization via Committee Consistency



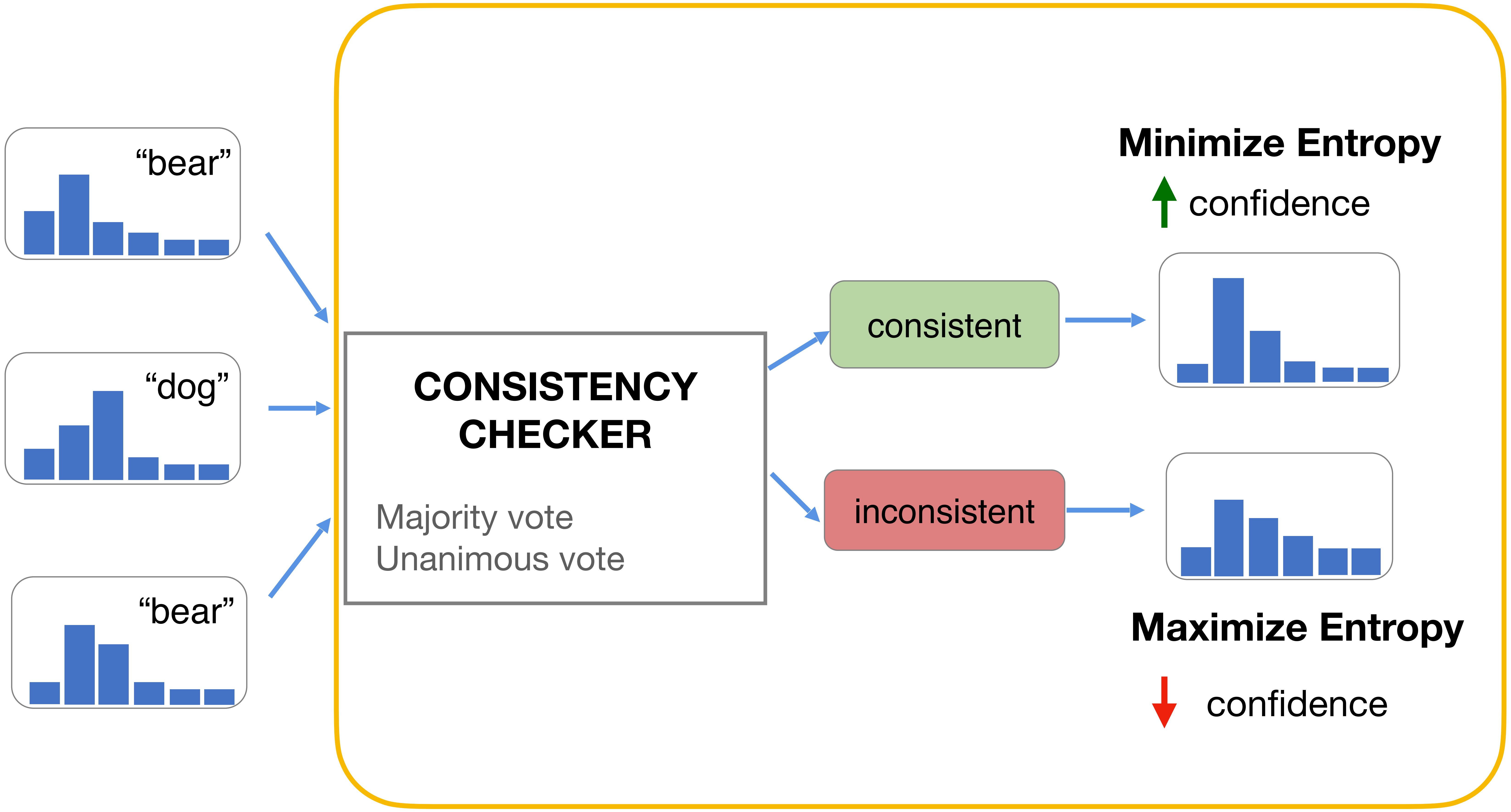
Prabhu, Khare, Karthik, Hoffman. arXiv, 2021

SENTRY: Selective Entropy Optimization via Committee Consistency



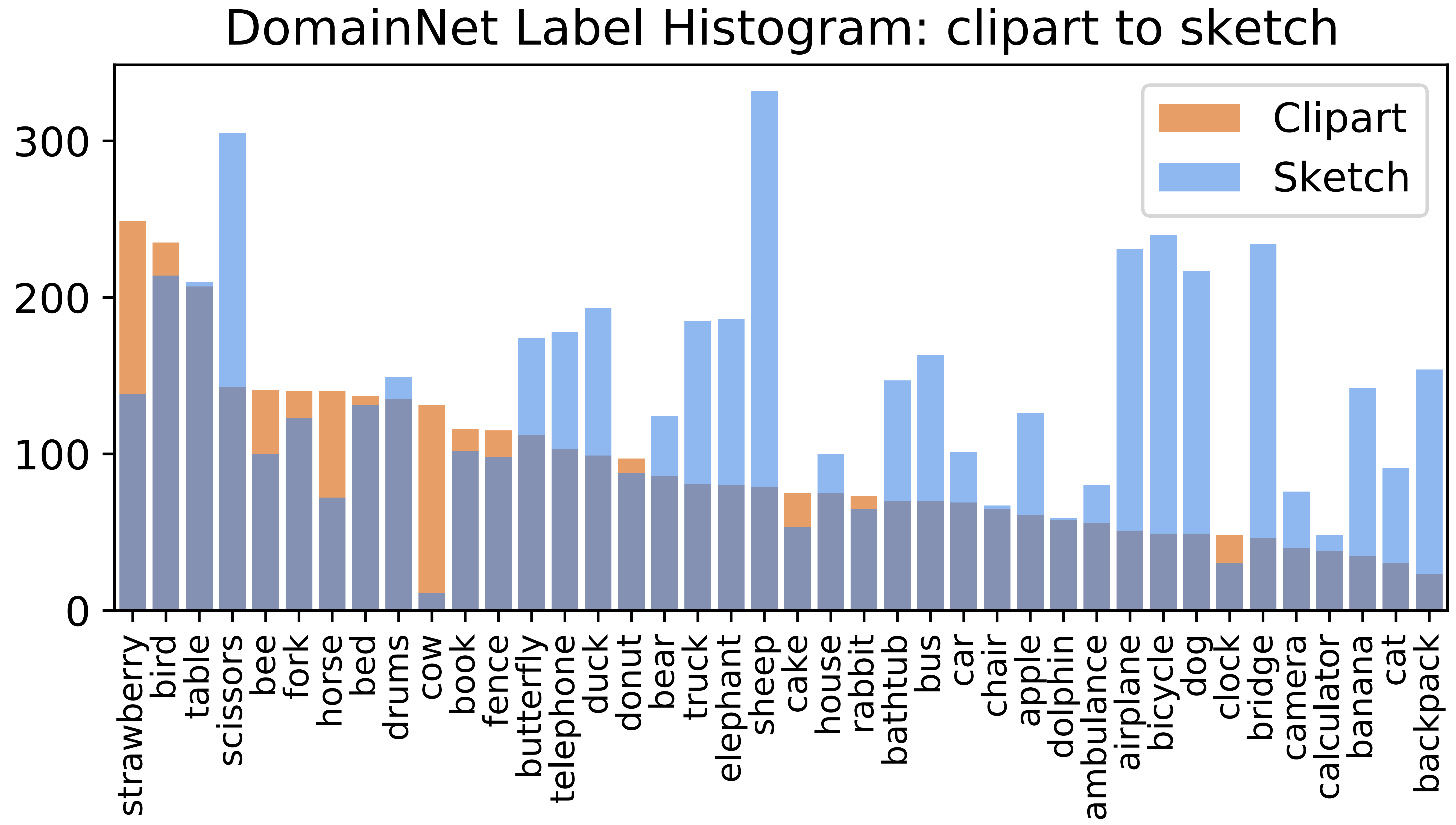
Prabhu, Khare, Karthik, Hoffman. arXiv, 2021

Selective Entropy Loss



Prabhu, Khare, Karthik, Hoffman. arXiv, 2021

SENTRY Results: Image Classification



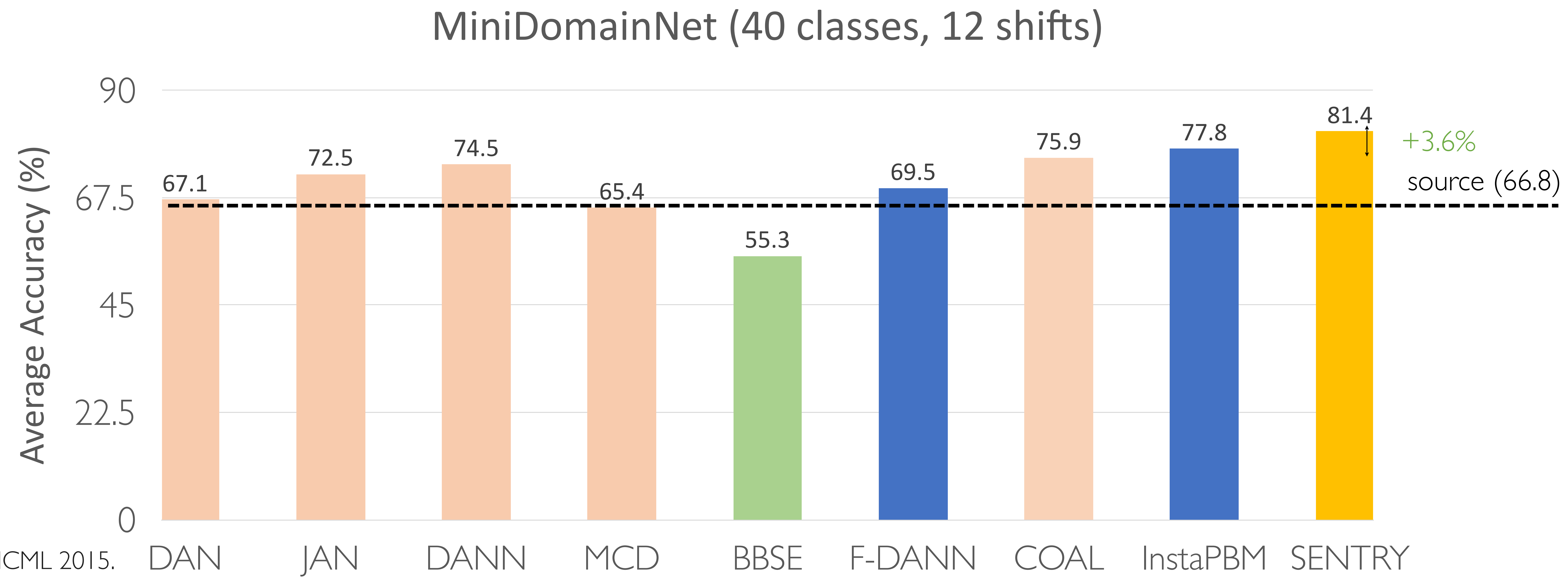
MiniDomainNet[1,2]

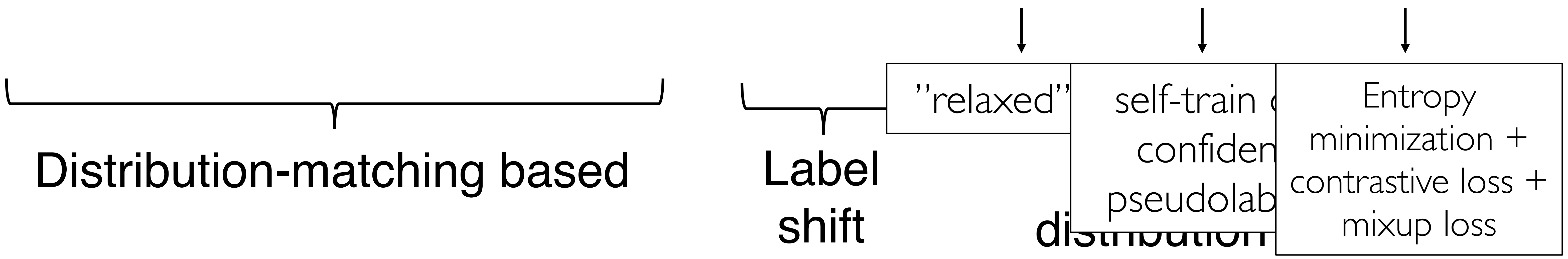
1. Peng et al., ICCV 2019.

2. Tan et al., ECCVW 2020.

Prabhu, Khare, Karthik, Hoffman 2021

SENTRY Results: MiniDomainNet

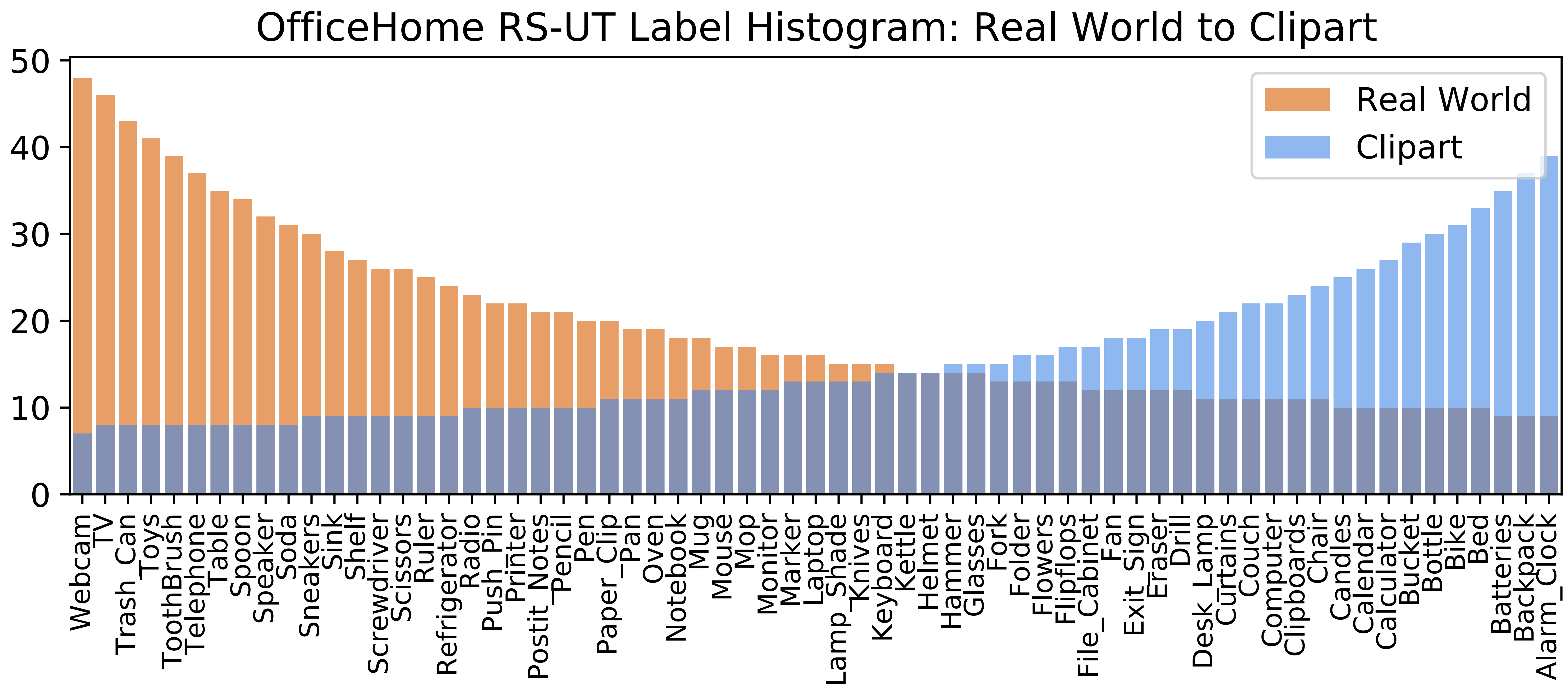




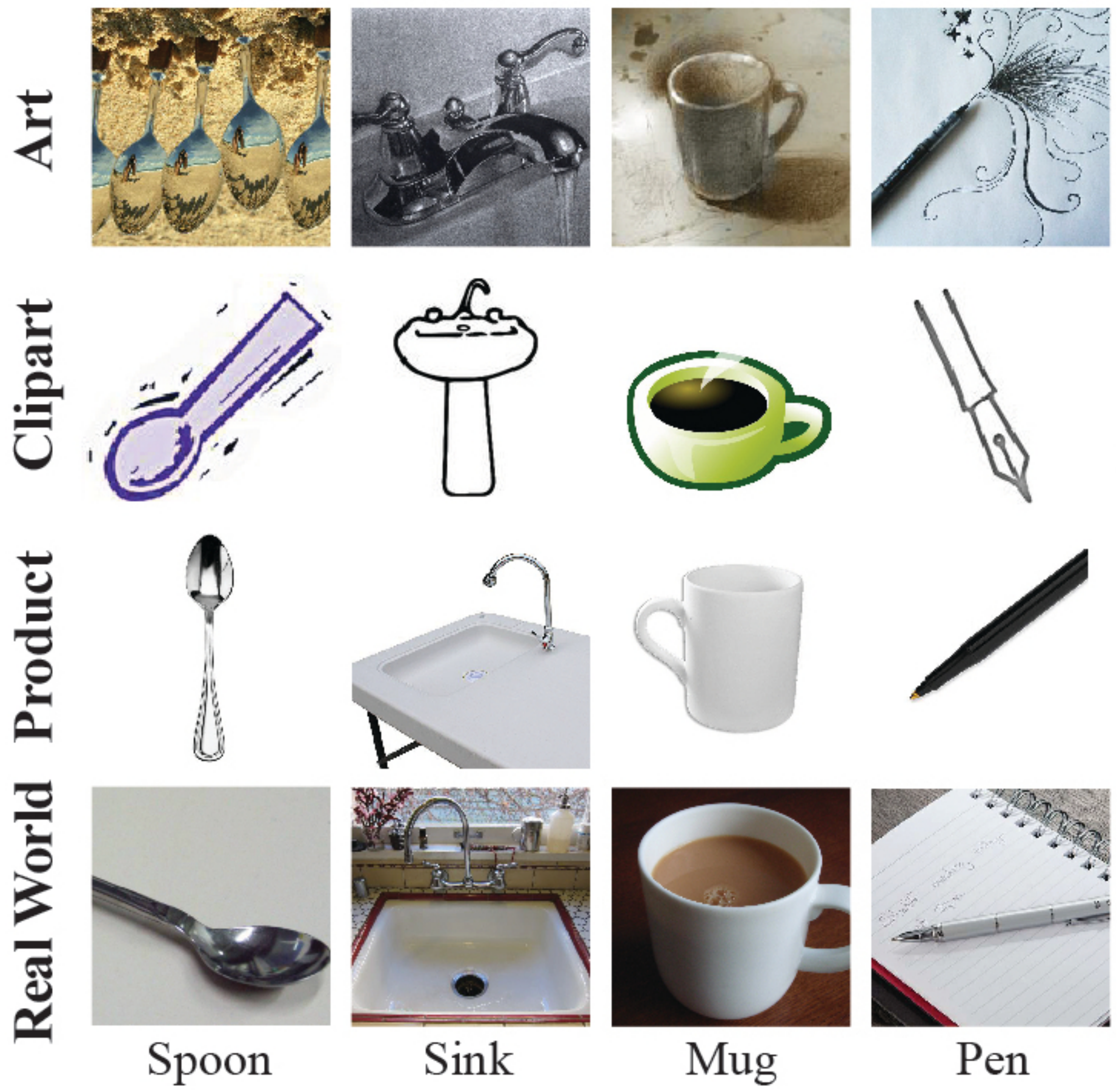
1. Long et al., ICML 2015.
2. Long et al., ICML 2017.
3. Ganin et al., ICML 2015.
4. Saito et al., CVPR 2018.
5. Lipton et al., ICML 2018.
6. Wu et al., ICML 2019.
7. Tan et al., ECCVW 2020.
8. Li et al., arXiv 2020.

SENTRY Results: Office Home

Custom label shifts[2]

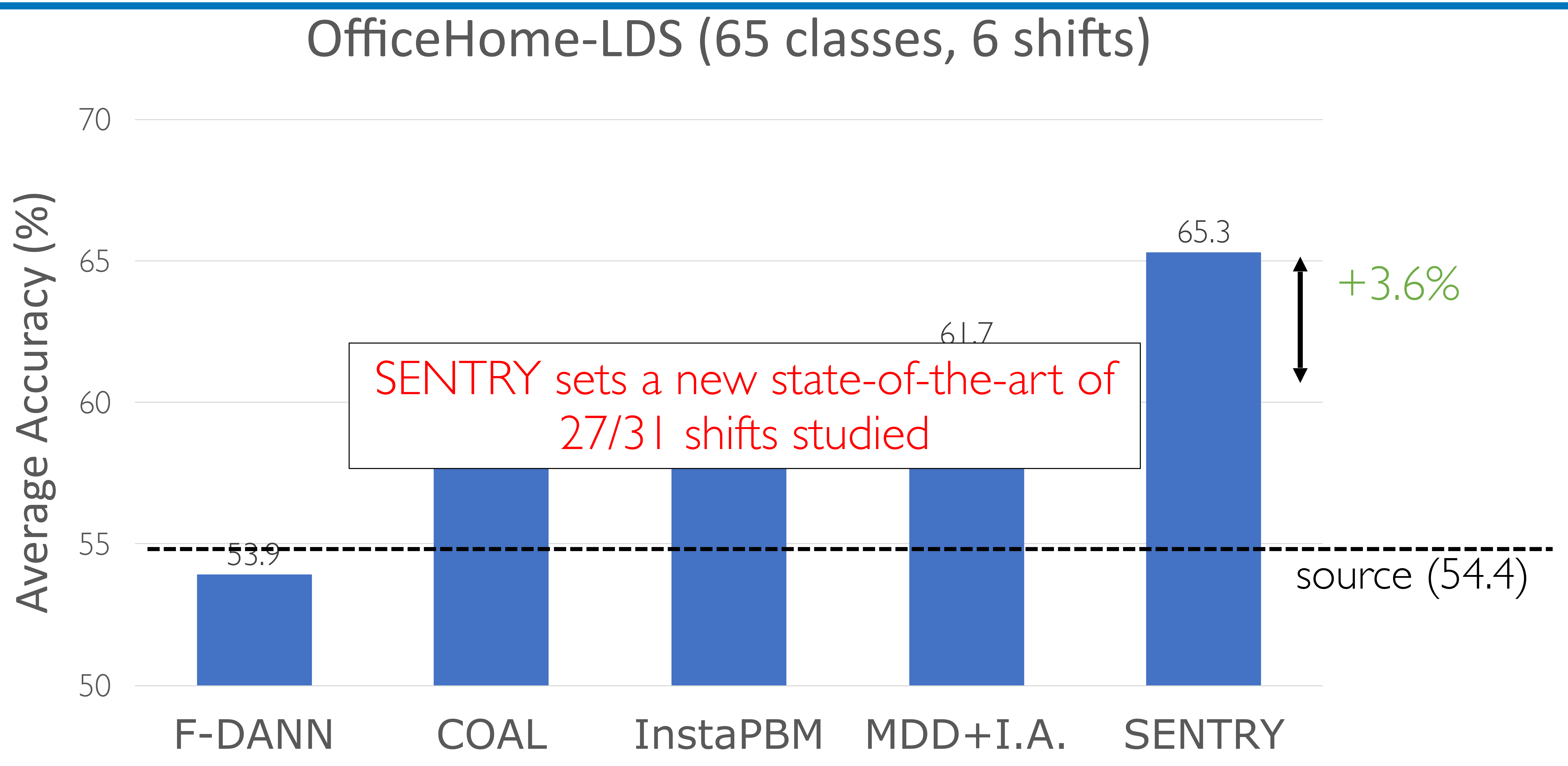


Office Home[1]



1. Venkateswara et al., CVPR 2017. 2. Tan et al., ECCVW 2020.

SENTRY Results: Office Home



Wu et al., ICML 2019.
Jiang et al., ICML 2020.

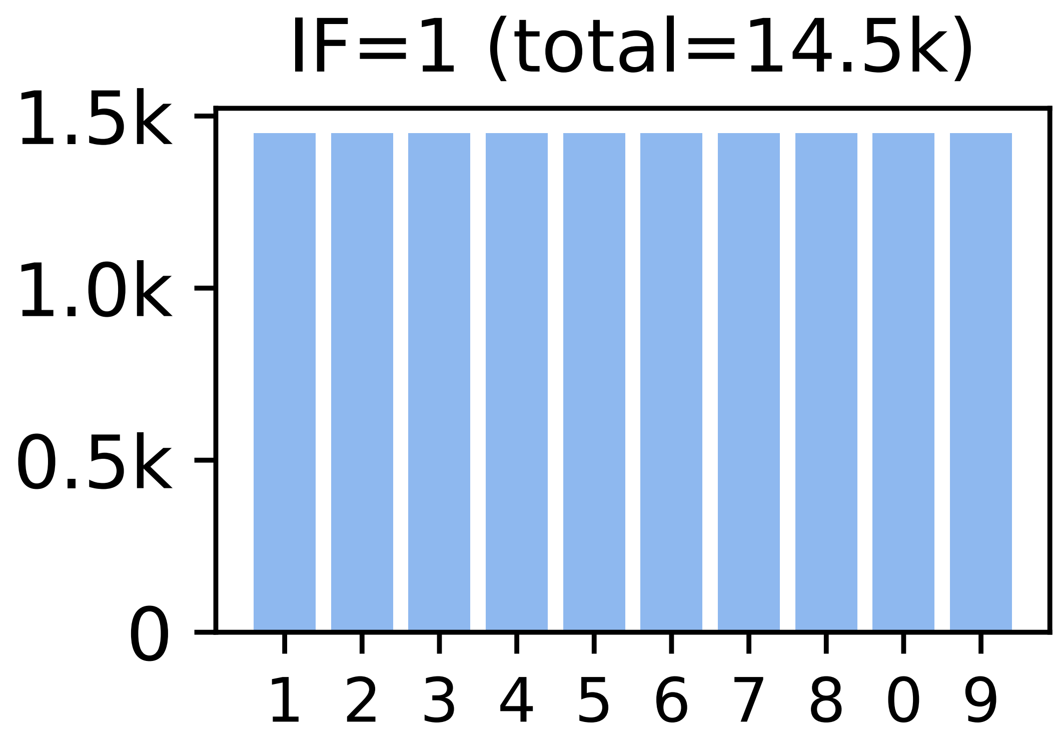
Tan et al., ECCVW 2020.
Long et al., NeurIPS 2018.

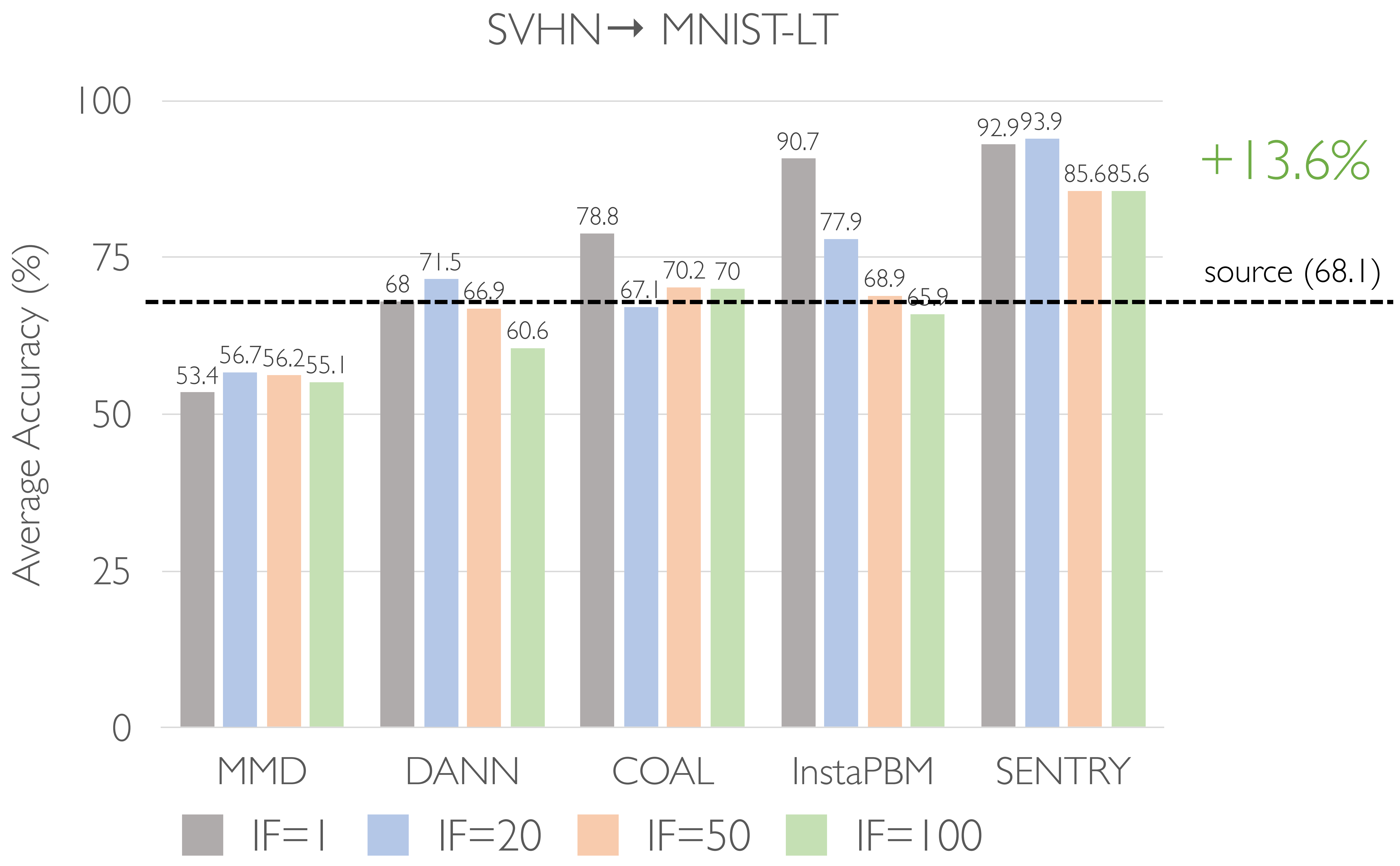
Li et al., arXiv 2020.
Zhang et al., ICML 2019

Prabhu, Khare, Karthik, Hoffman. arXiv, 2021

Results: Controlling Target Distribution

MNIST-LT label histograms

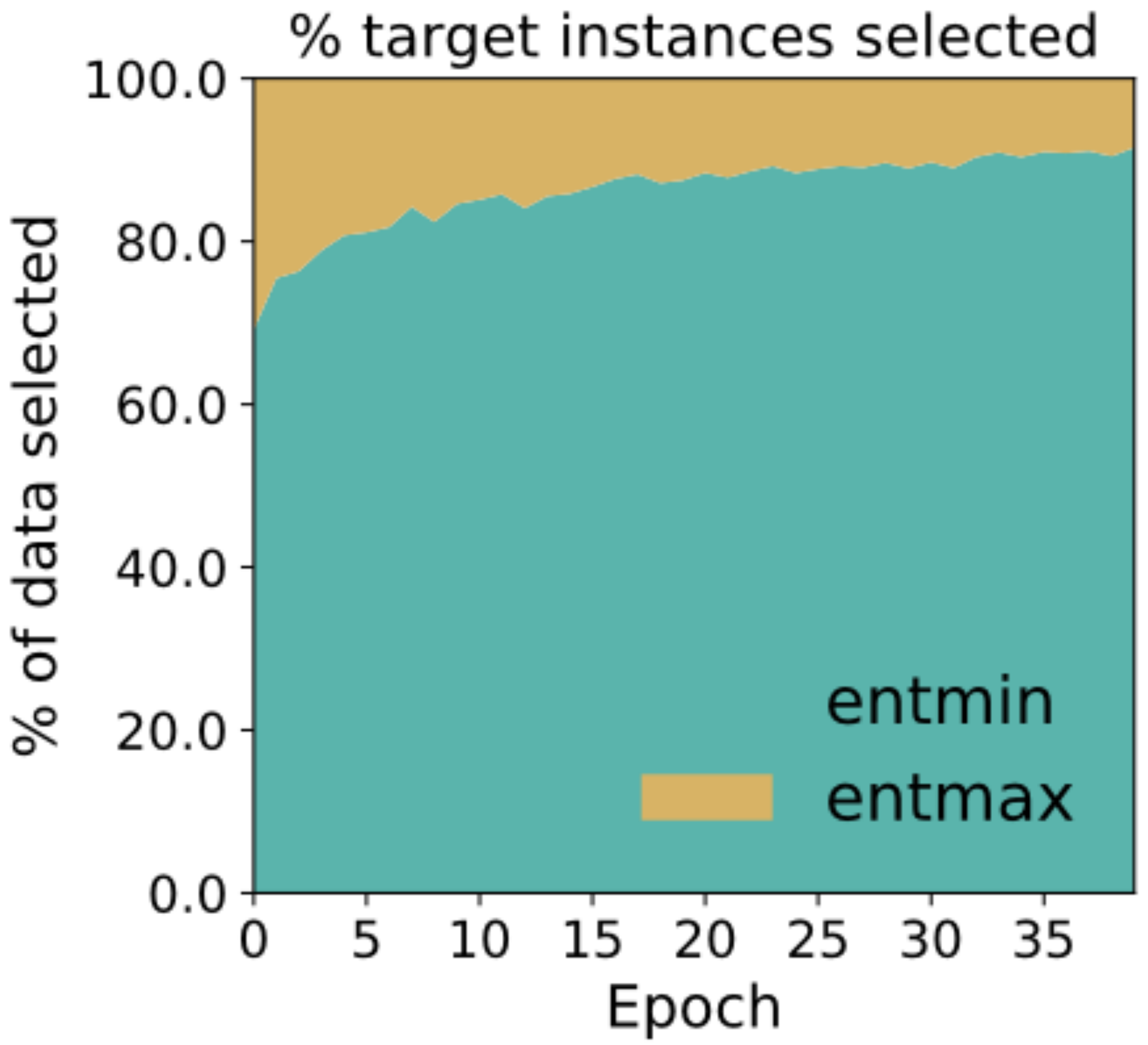




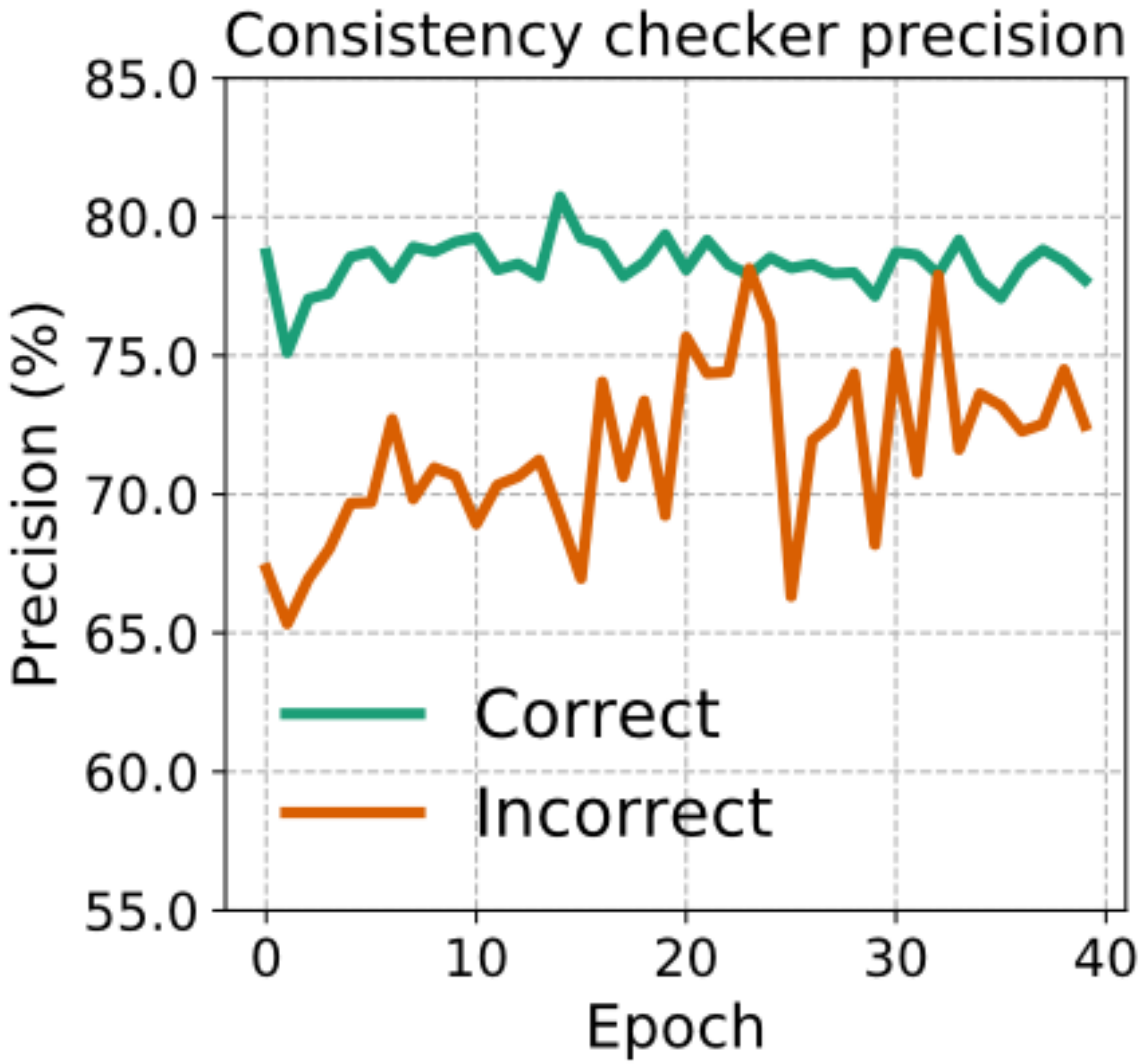
Long et al., ICML 2015. Ganin et al., ICML 2015. Tan et al., ECCVW 2020. Li et al., arXiv 2020.

Ablating SENTRY: Selection

Increasing % of data is aligned over time



Predictive consistency is a good reliability measure



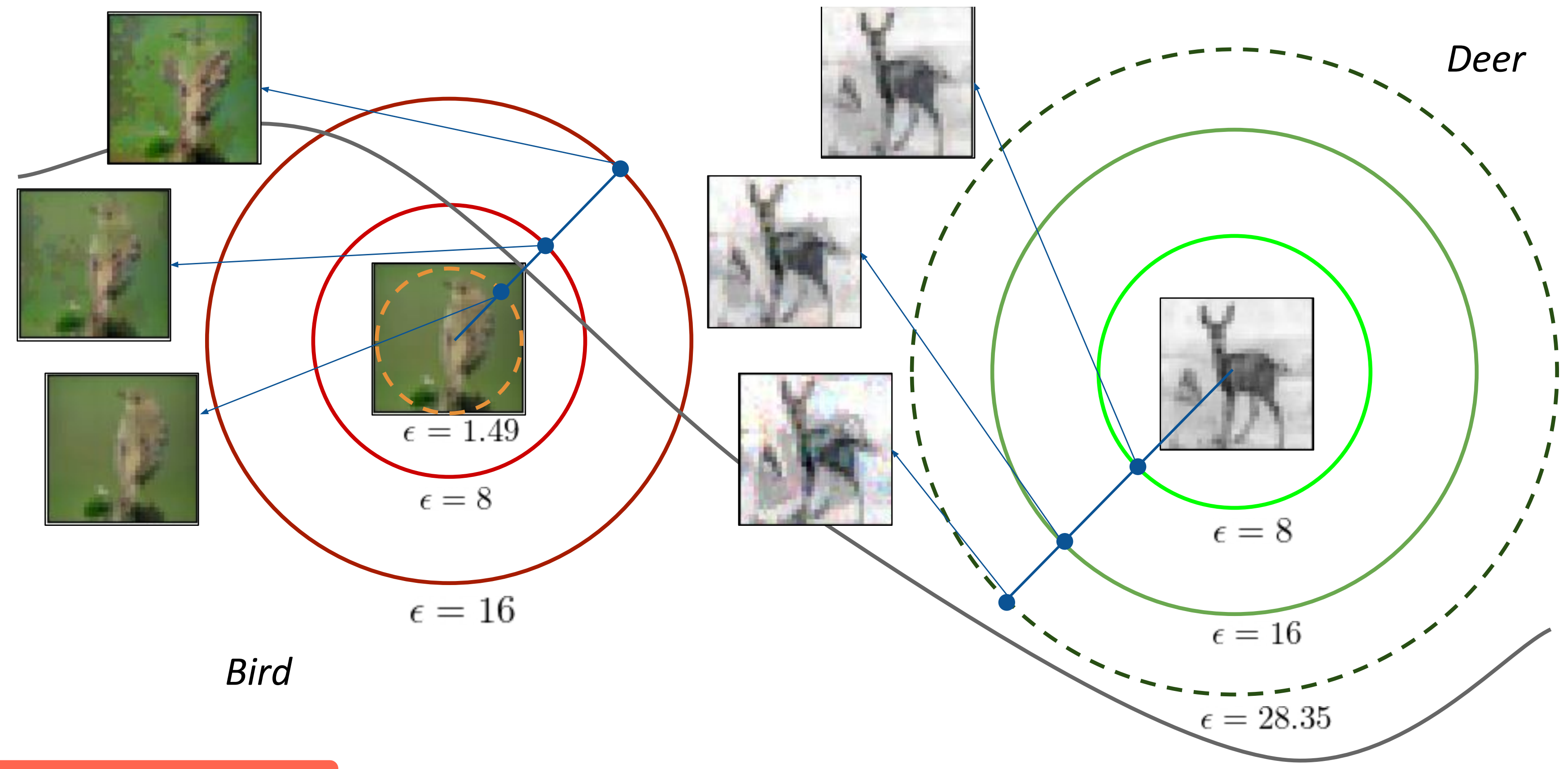
Prabhu, Khare, Karthik, Hoffman. arXiv:2012.11460, 2021

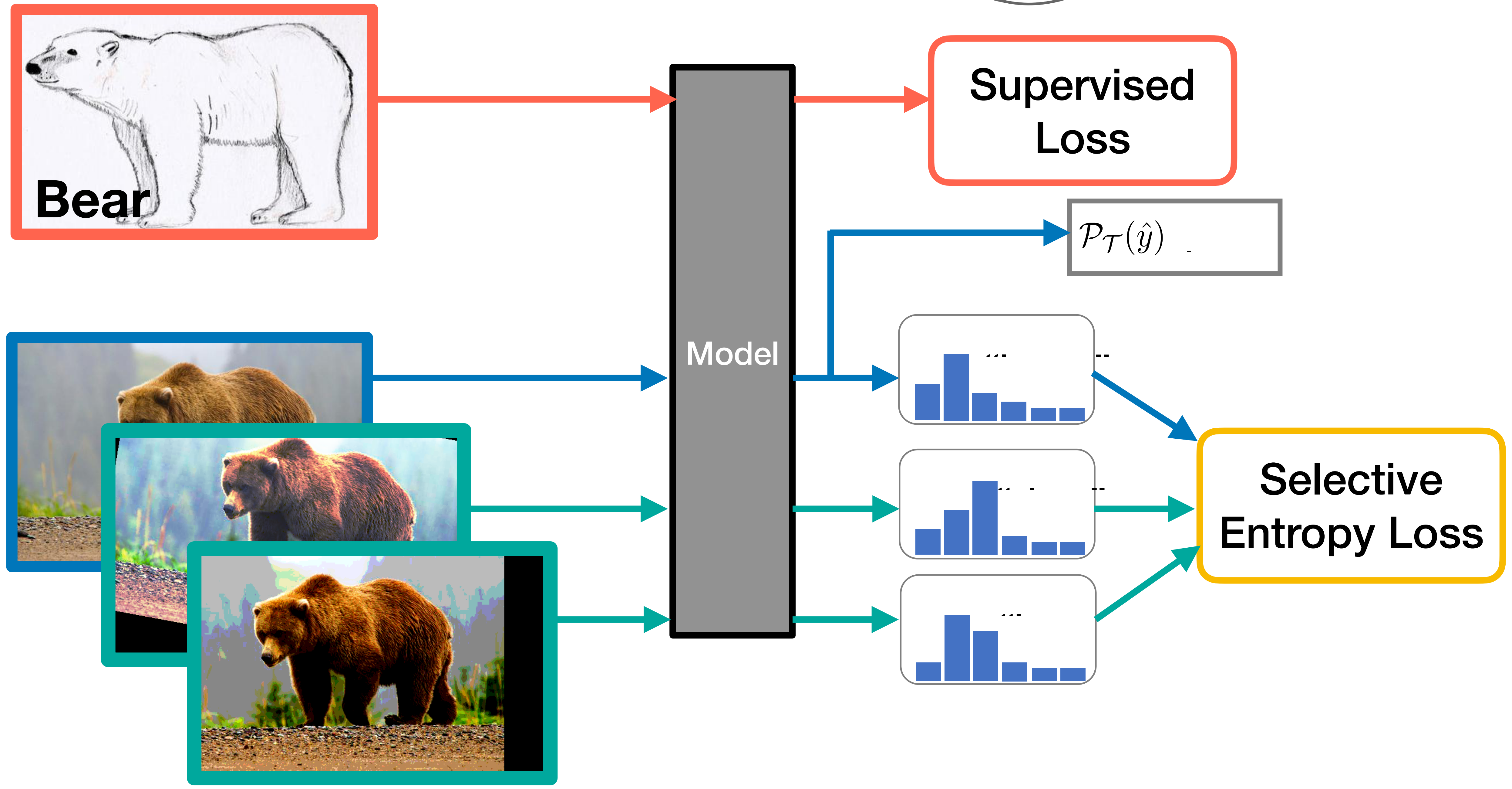
Summary

Key Idea: Decide when to learn

Learning can be derailed by

- Unreliable labels
  - Label noise
  - Manipulated
  - Model misalignment
- Unreliable samples
  - Inherently ambiguous
  - Different from prior data
  - Manipulated





Thank you

Sean Foley

Daniel Bolya

Sruthi Sudhakar

Arvind Krishnakumar

Rohit Mittapali

Kartik Sarangmath

Prithvijit Chattopadhyay

Viraj Prabhu

Shivam Khare

Deeksha Karthik

Bhavika Devnani

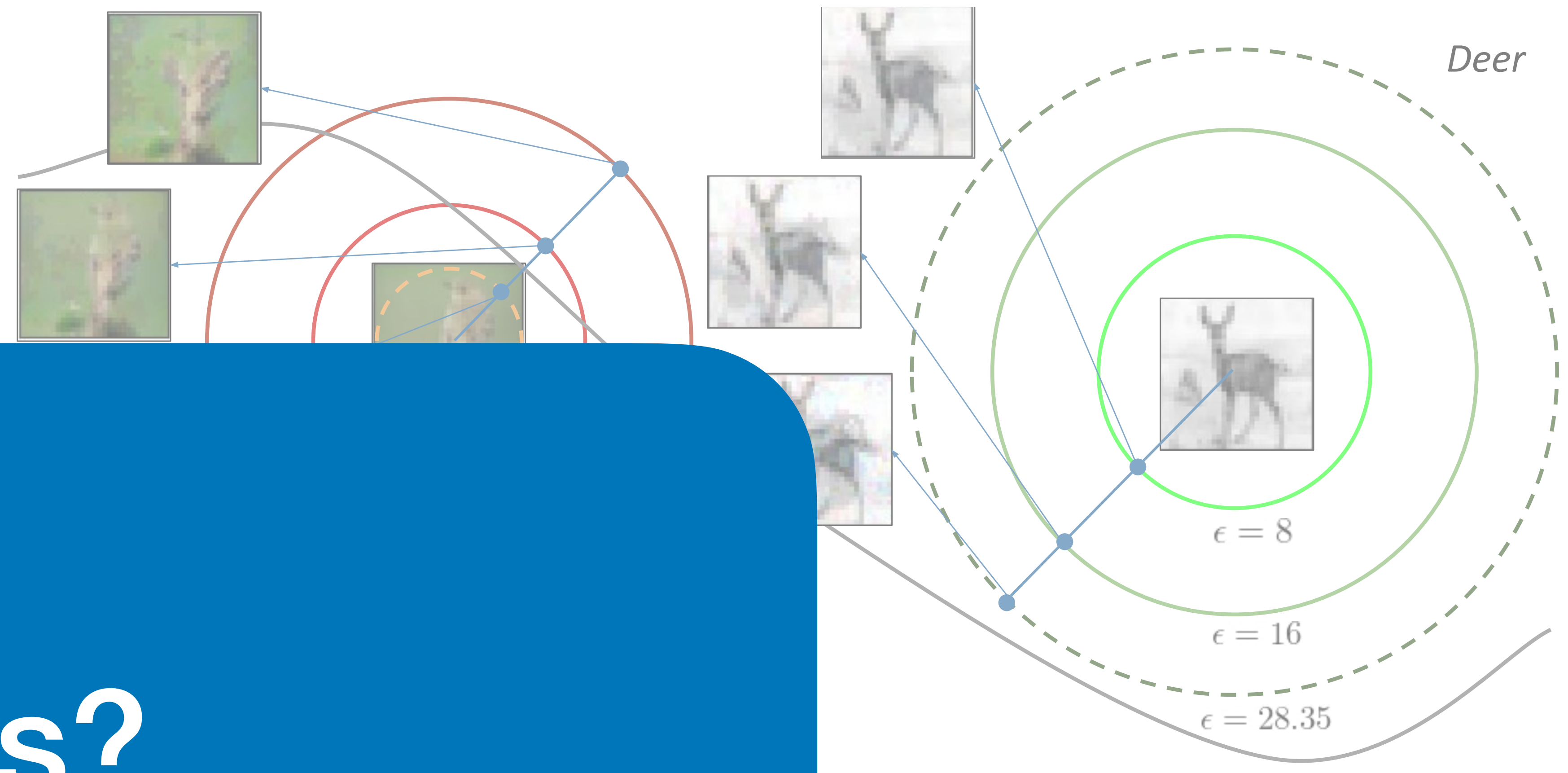
Luis Bermudez

Summary

Key Idea: Decide when to learn

Learning can be d

- Unreliable lab
  - Label nois
  - Manipulat
  - Model mi
- Unreliable sa
  - Inherently ambiguous
  - Different from prior data
  - Manipulated



Questions?

judy@gatech.edu