Responsible CV:
How do models fail and what can we do about it?

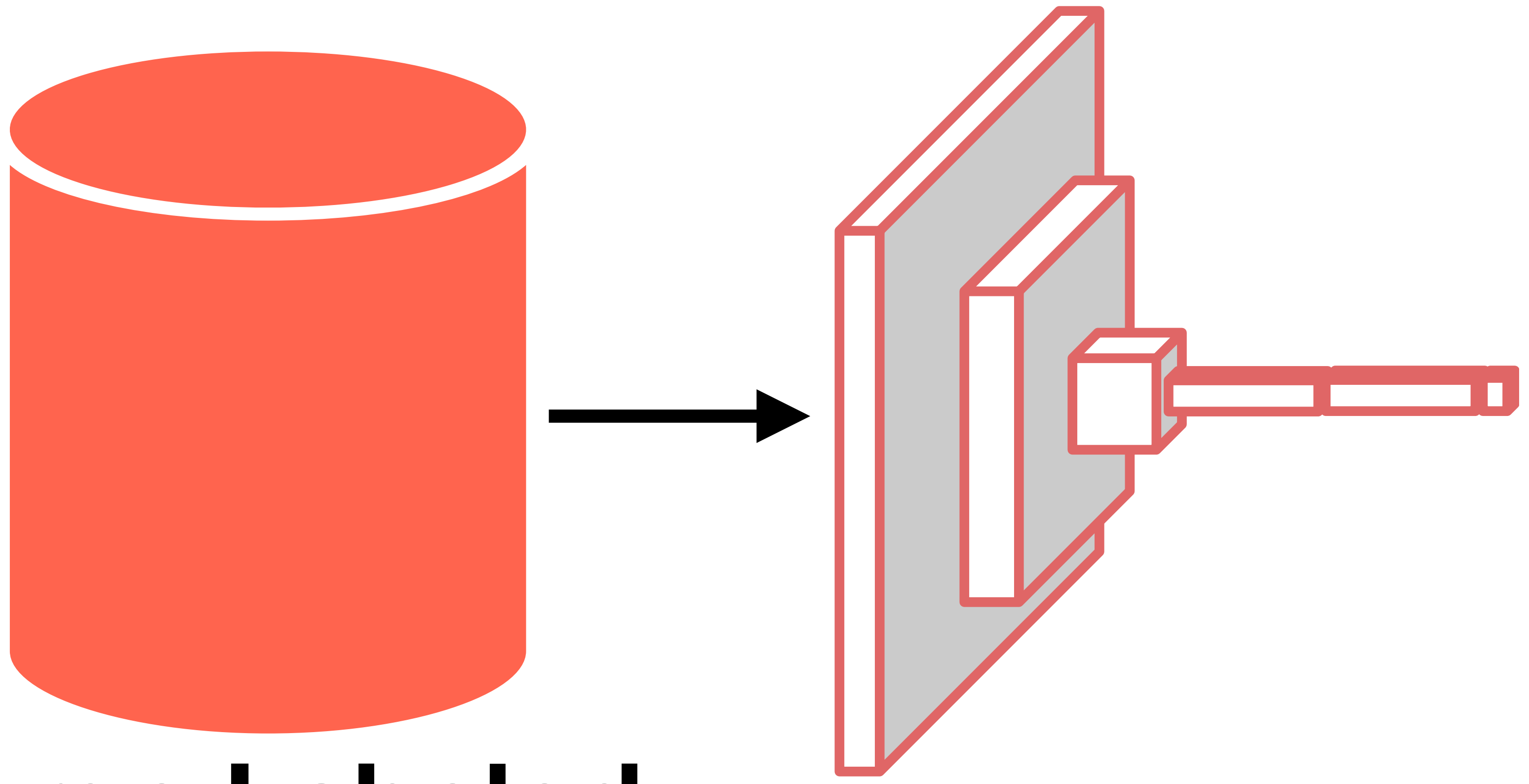
Judy Hoffman and Viraj Prabhu

Human-centered AI Tutorial @CVPR

June 20, 2022

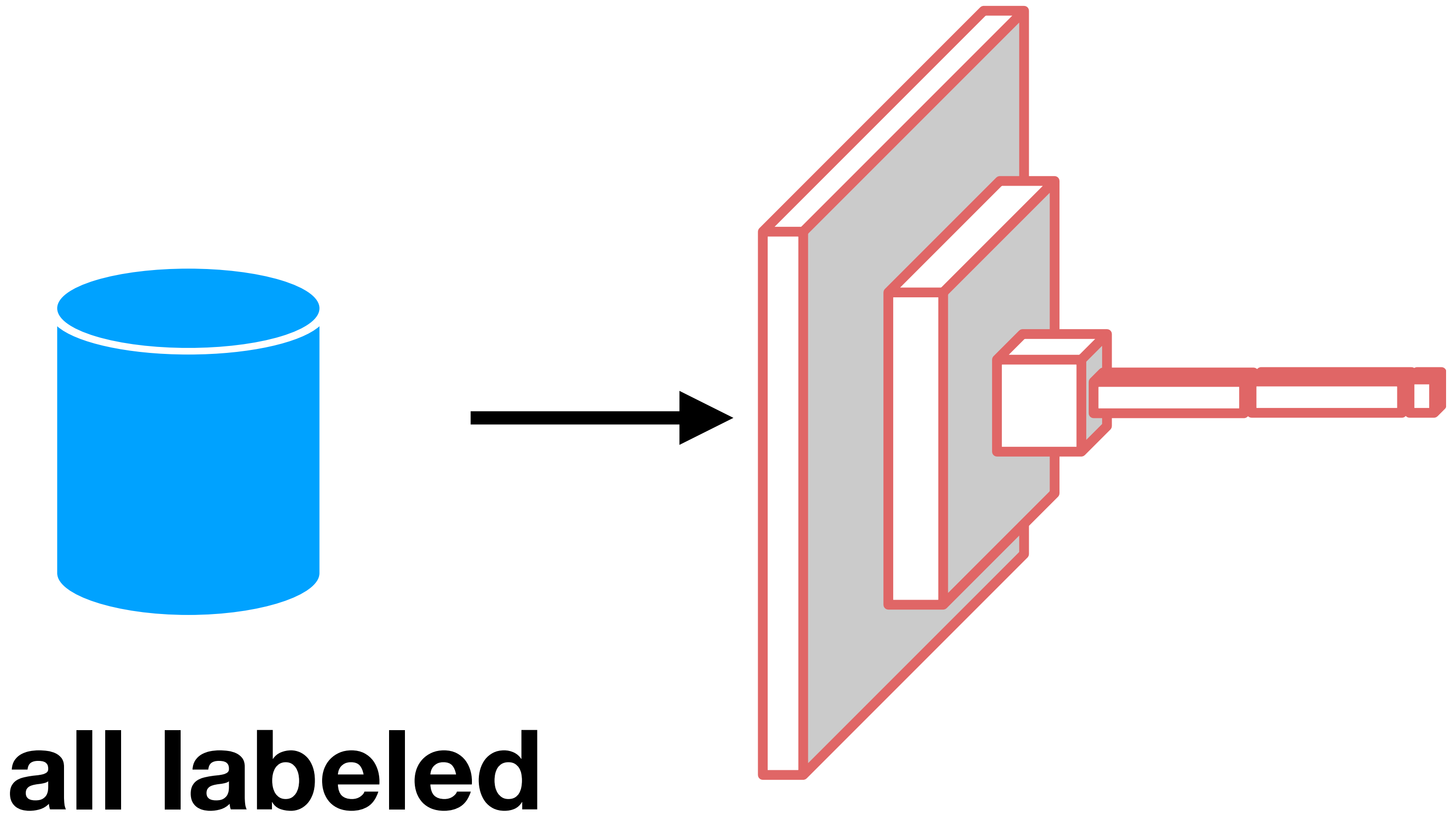
Practical Transfer Learning

Initial Training



Large Labeled source data

Finetuning



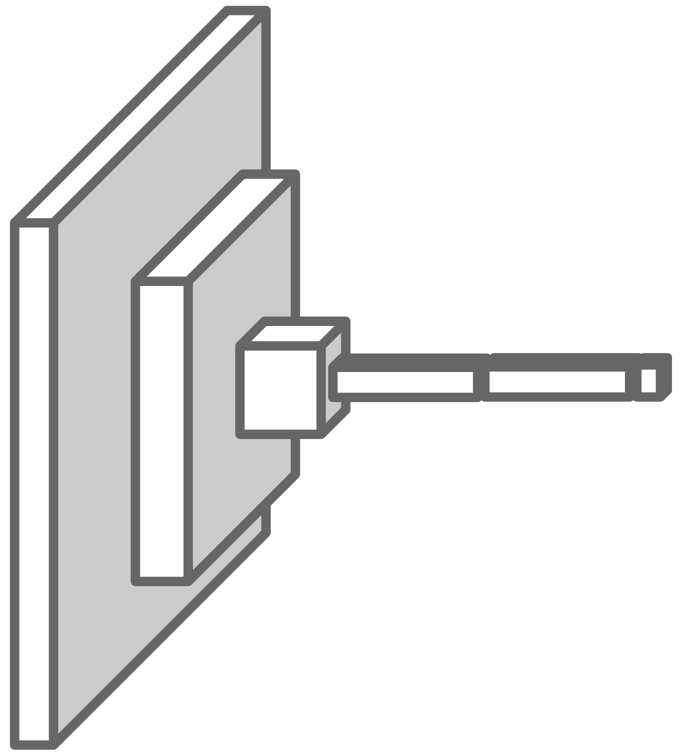
Small labeled target data

Frequently select model that performs best on ImageNet

Standard Visual Recognition Pipeline

1. Collect Data

2. Annotate Data



3. Train Model

Visual Recognition Benchmarks

Classification

PASCAL

COCO
Common Objects in Context

CITYSCAPES
DATASET

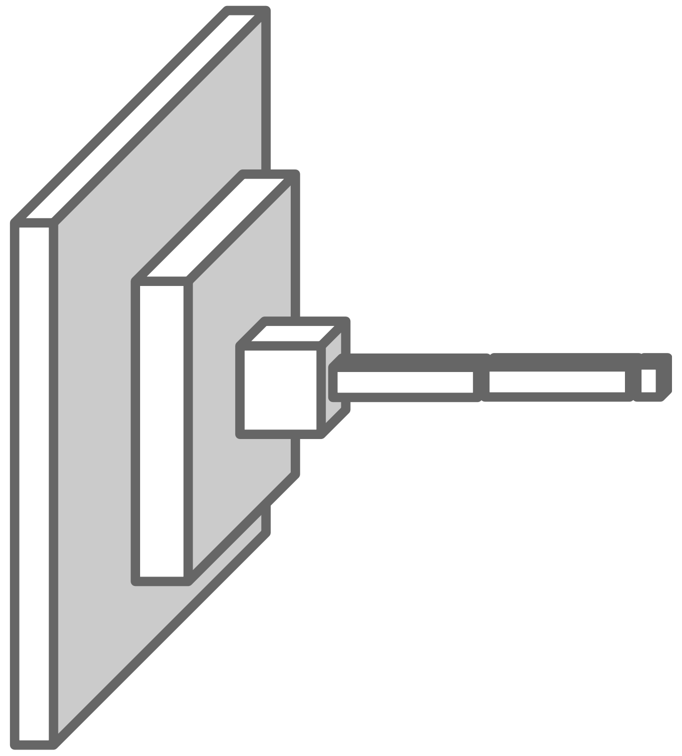
LVIS

Detection / Segmentation

Standard Visual Recognition Pipeline

1. Collect Data

2. Annotate Data



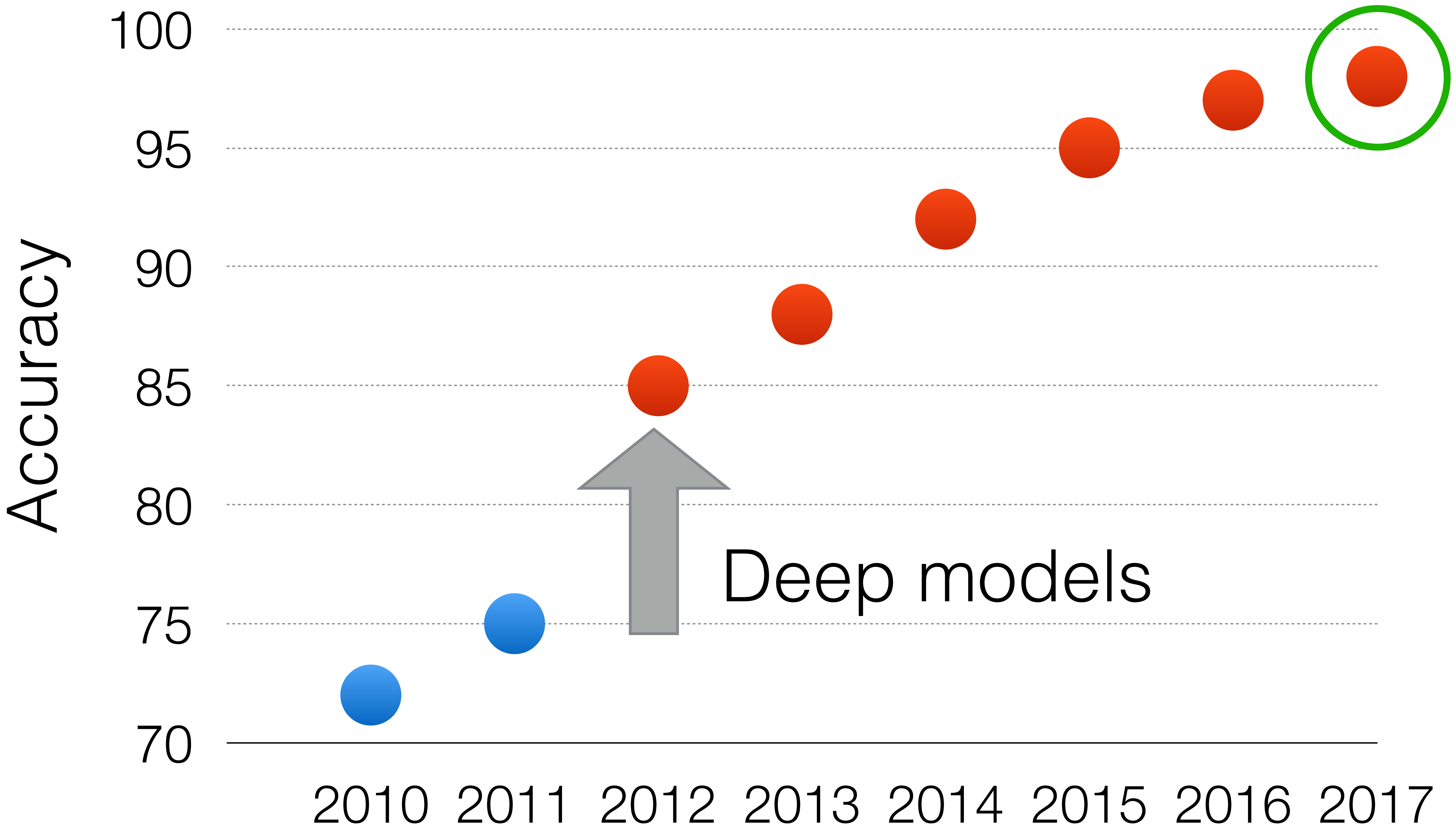
3. Train Model

4. Validate Model

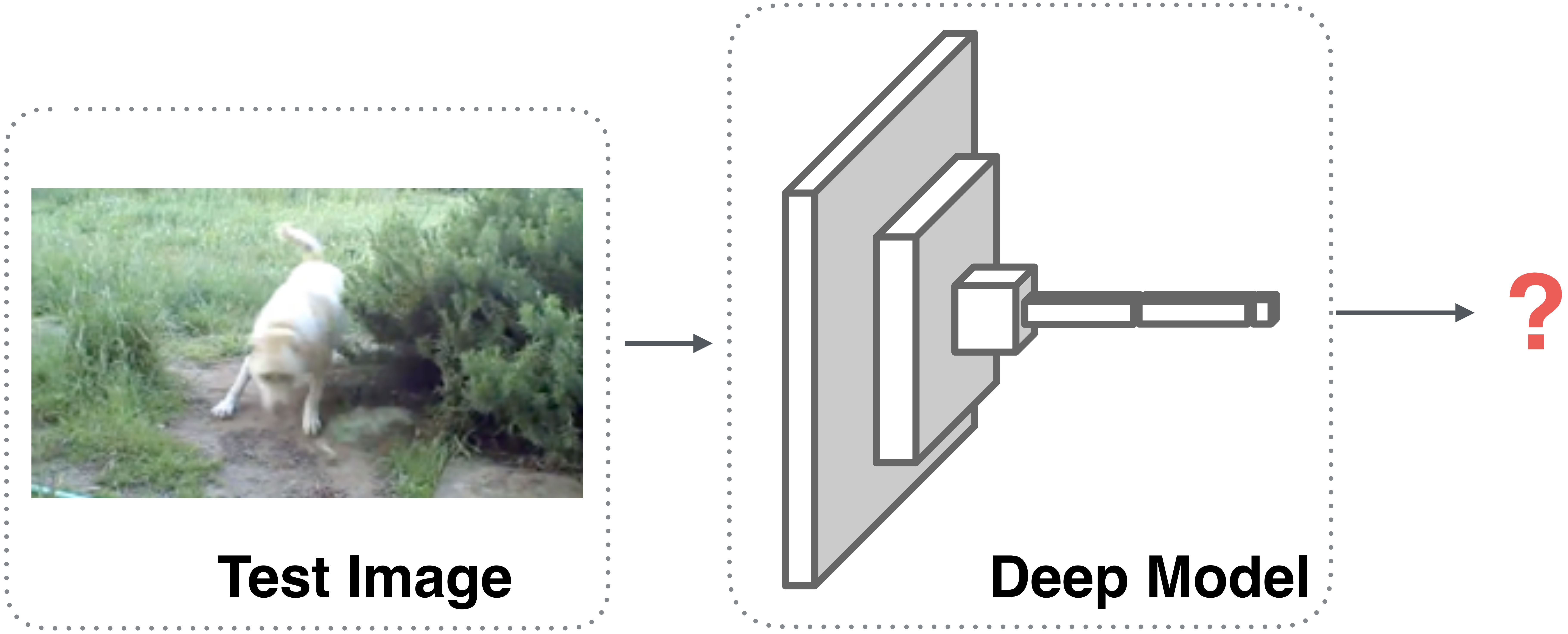
Benchmark Performance

Millions of Images

Challenge to recognize 1000 categories



Dataset Bias



Dataset Bias

Dataset Bias

IMAGENET

Dataset Bias

Low resolution

Motion Blur

Pose Variety

The world has high natural variation

Large Potential for Change
Different: Weather, City, Car

Car
Road
Sidewalk
Person
Sky
Vegetation
Street Sign
Building

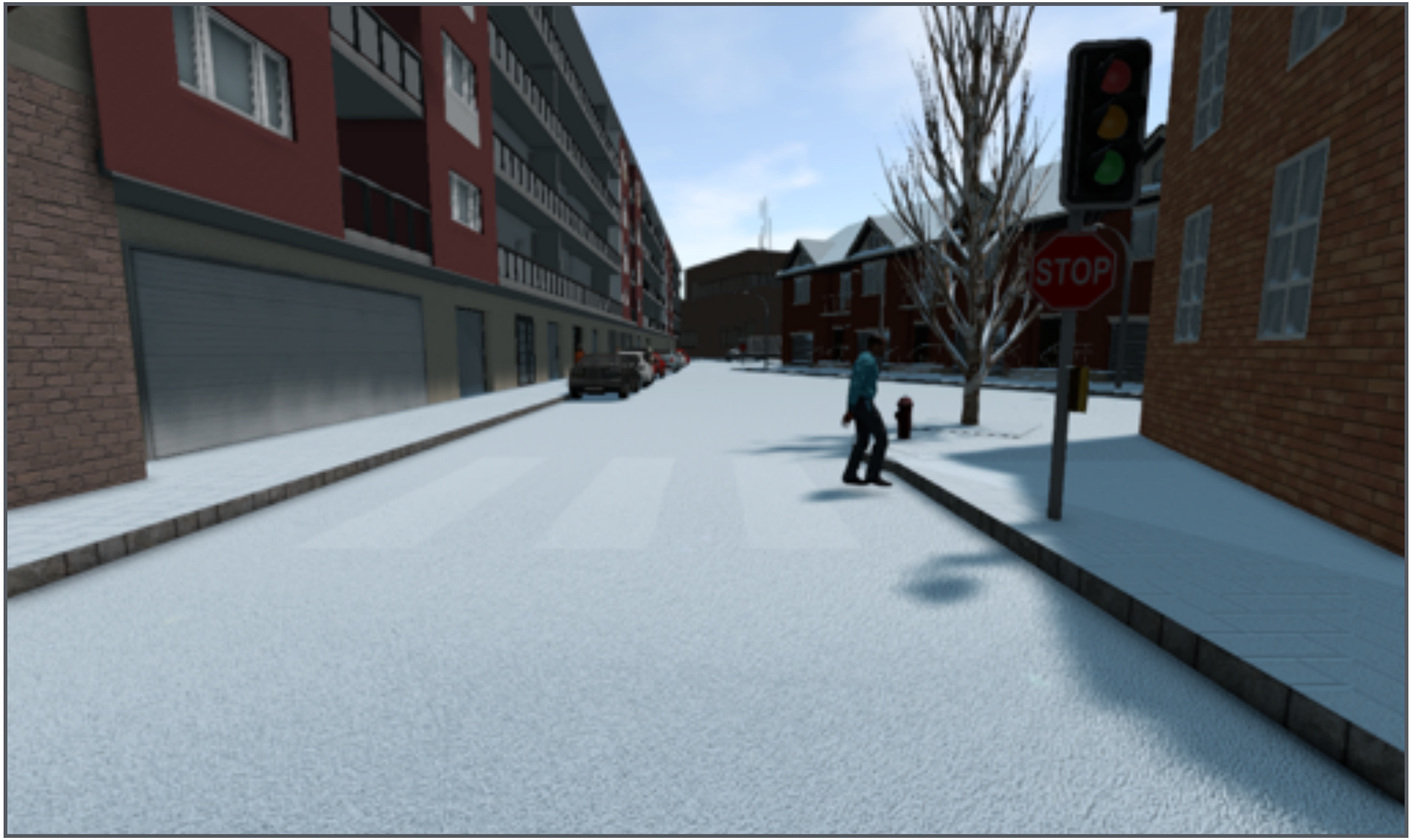
Expensive
($10-12 per image)

Train in Sunny Weather

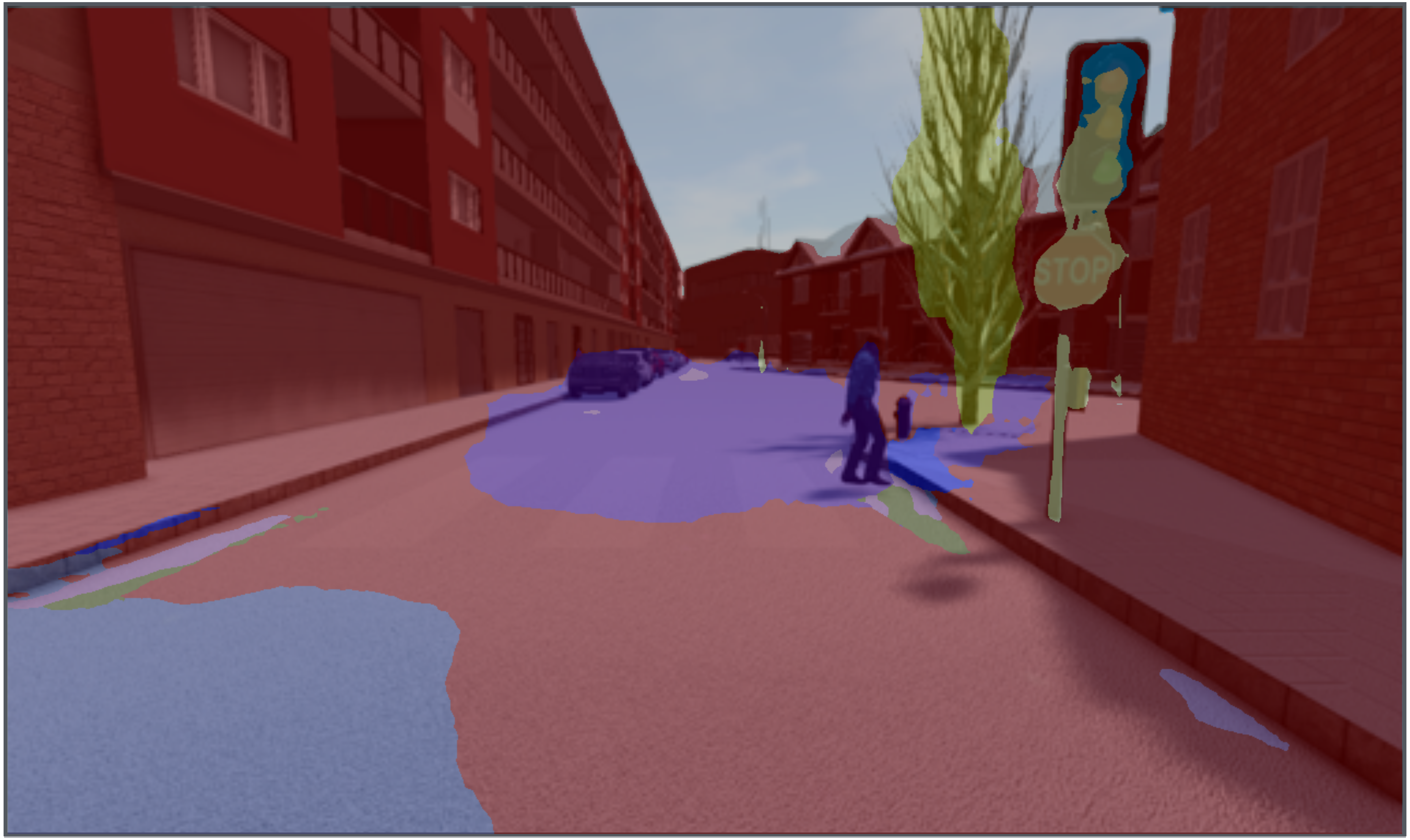
Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Efros, Darrell, ICML 2018.

Robust to Weather Changes?

- Car
- Road
- Sidewalk
- Person
- Sky
- Vegetation
- Street Sign
- Building
- Traffic Light



Robust to Weather Changes?



- Car
- Road
- Sidewalk
- Person
- Sky
- Vegetation
- Street Sign
- Building
- Traffic Light

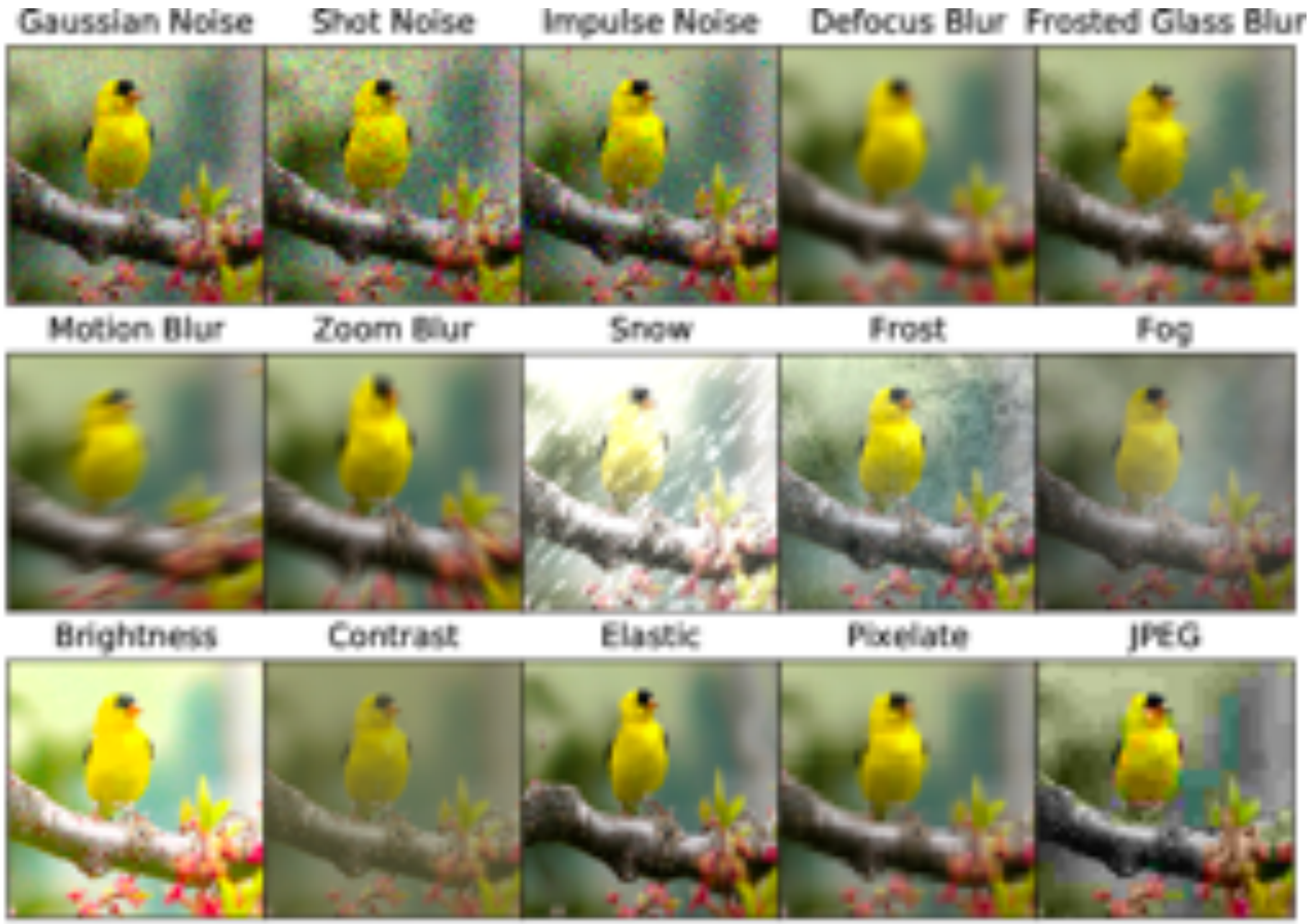
Impact of Input Corruptions on Recognition

CiFAR-10, ResNet-18

Clean Acc = 94.2

↓

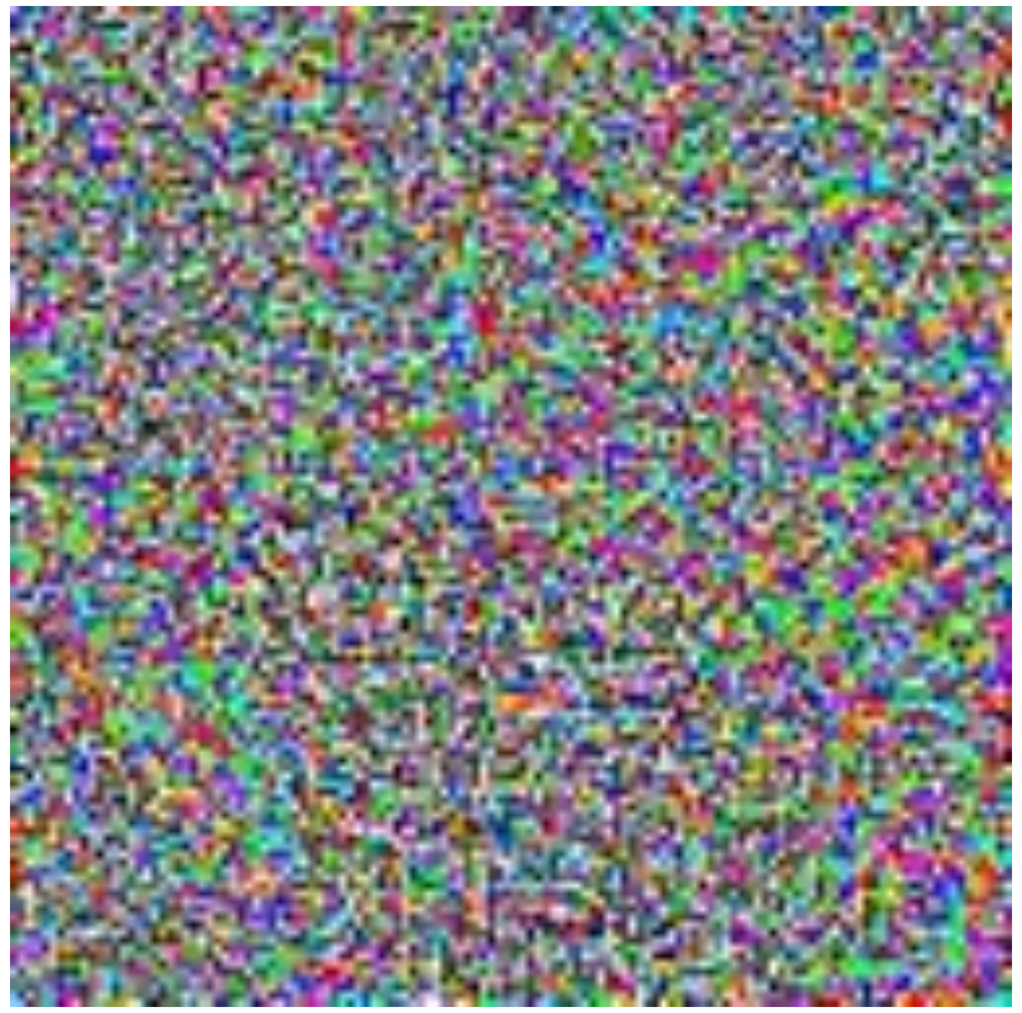
Corrupt Acc = 72.7



Hendrycks and Dietterich, ICLR, 2019.

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]

Benchmark Challenge Adversarial

The Art of Robustness:
Devil and Angel in Adversarial Machine Learning

Workshop at IEEE Conference on Computer Vision and Pattern Recognition 2022

RobustNav
Towards Benchmarking Robustness in Embodied Navigation

ICCV 2021

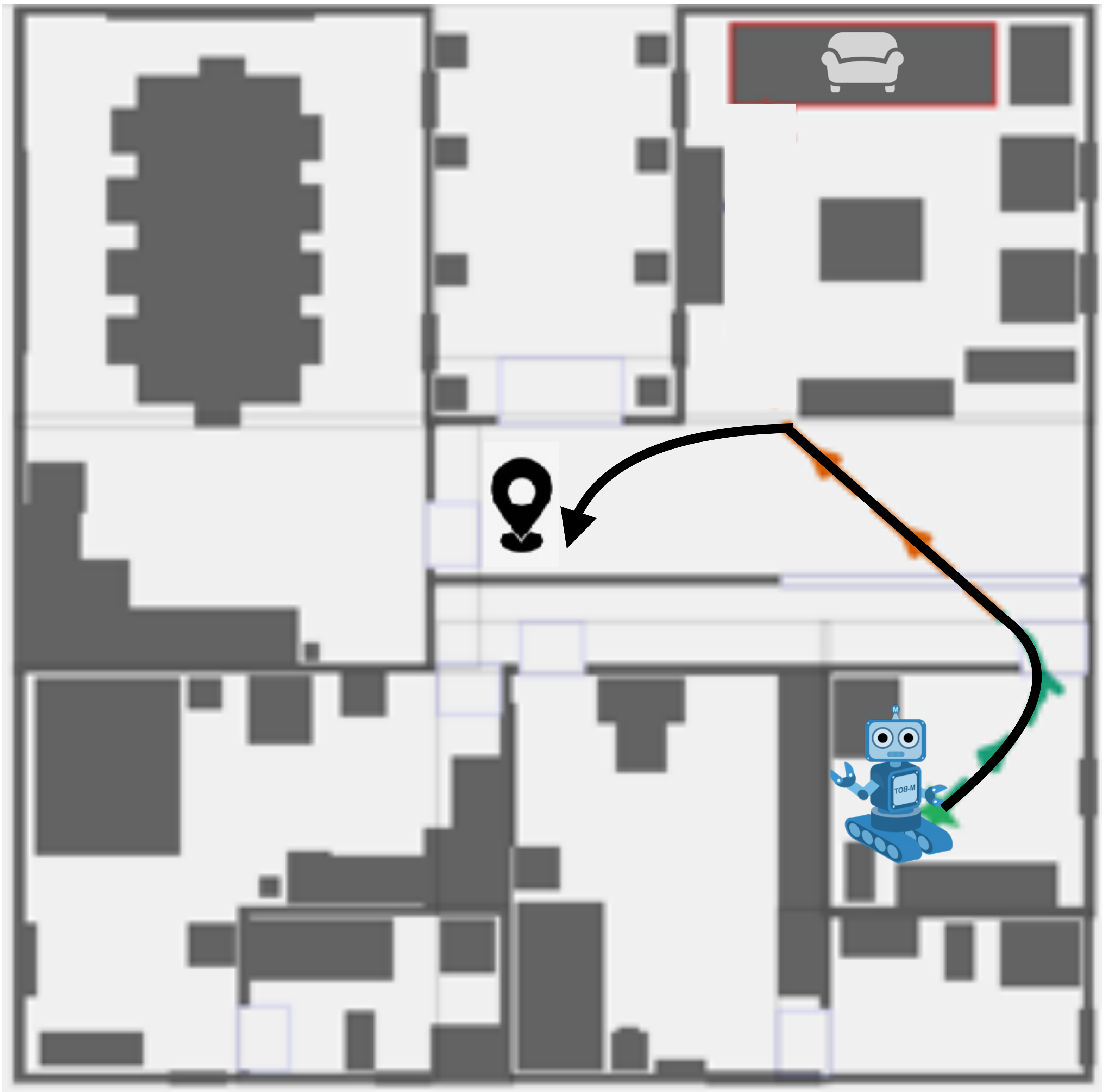



Prithvijit Chattopadhyay[1]   Judy Hoffman[1]   Roozbeh Mottaghi[2]   Ani Kembhavi[2]

[1] Georgia Tech   [2] PRIOR

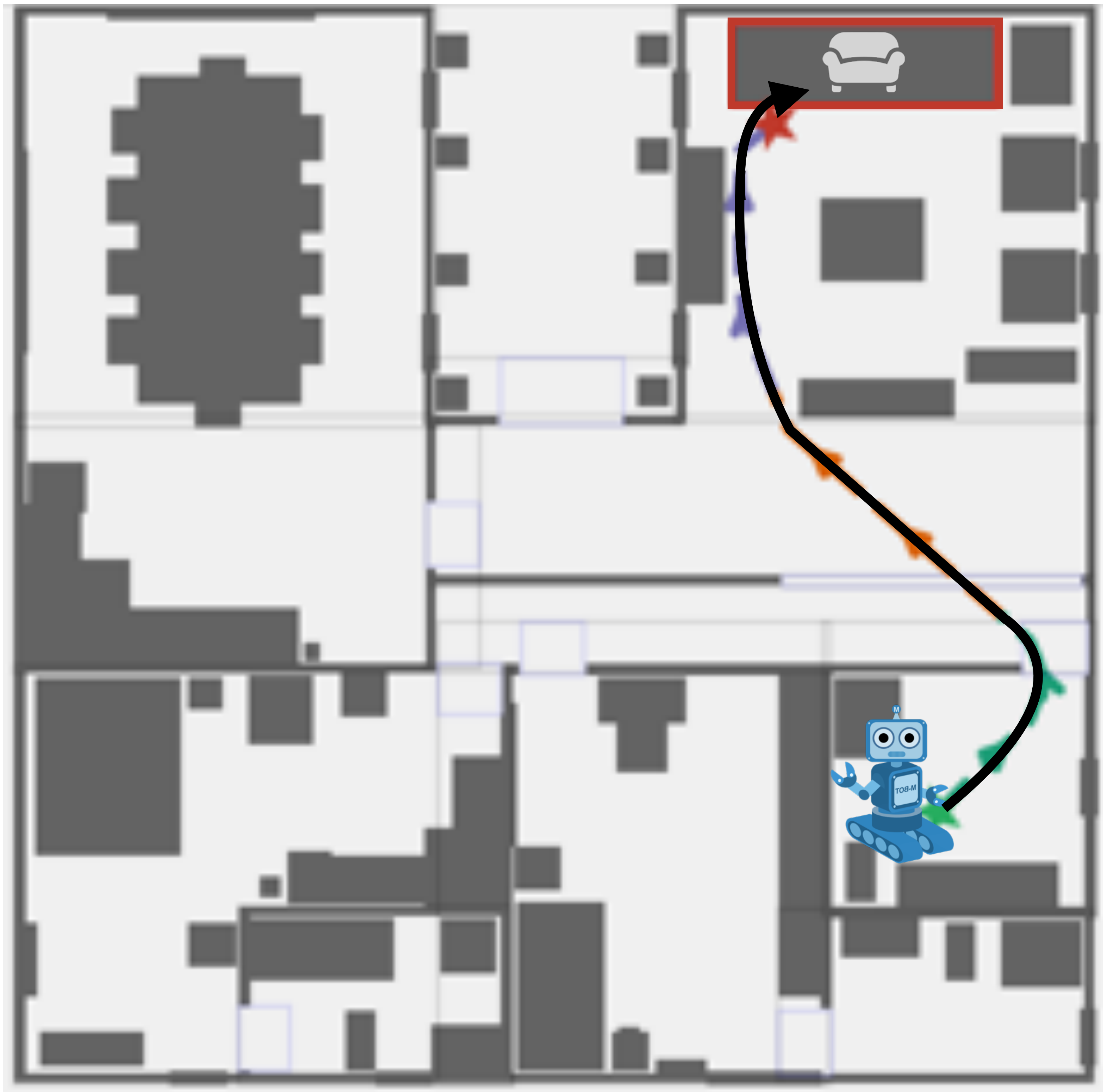
Visual Navigation (RGB+Depth)

PointNav
GPS + Compass enabled navigation



Task: Go to (r, θ) location

ObjectNav
Semantic, Target-driven Navigation



Task: Go to a "sofa"

Figure Credits: Abhishek Das

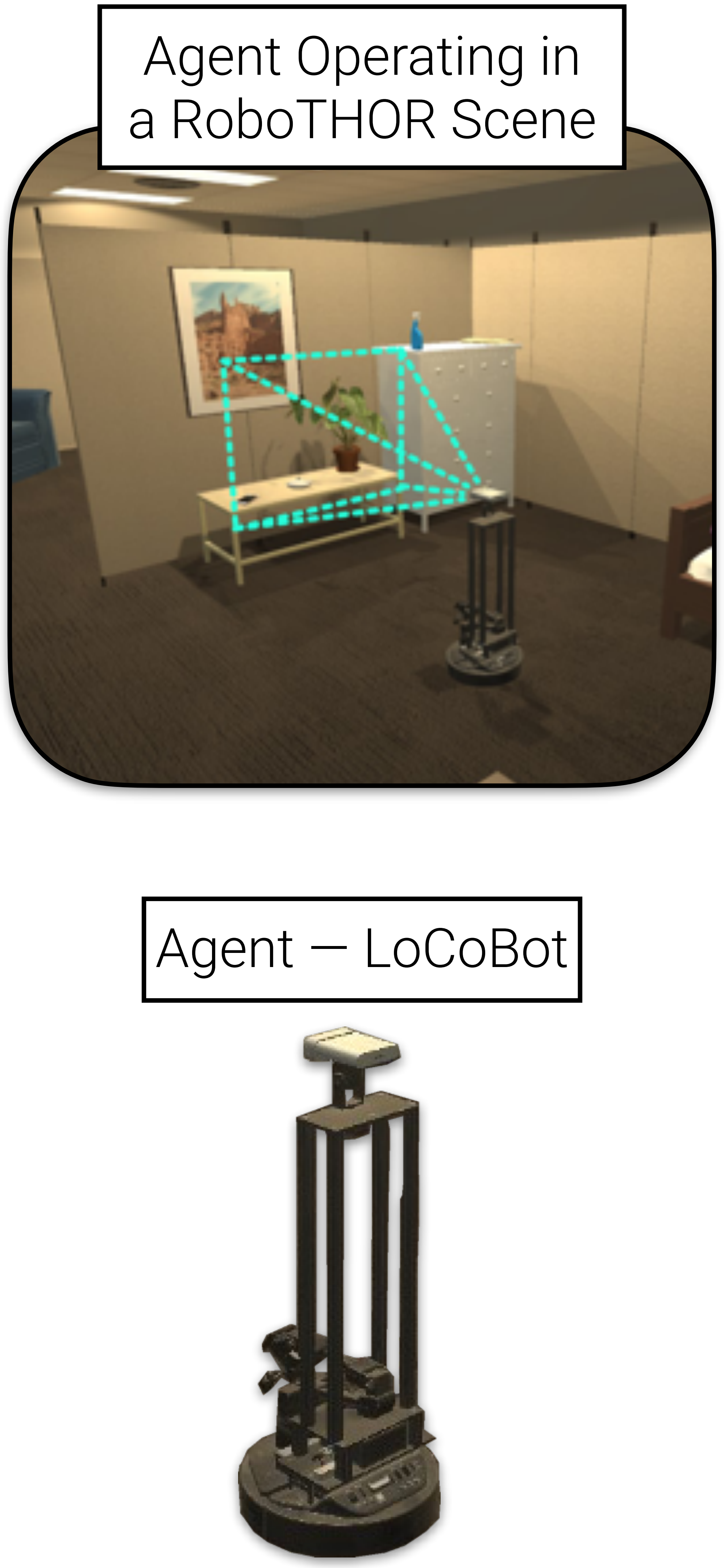
Visual Navigation

Agents don’t have access to a “map”, and must navigate based solely on sensory inputs

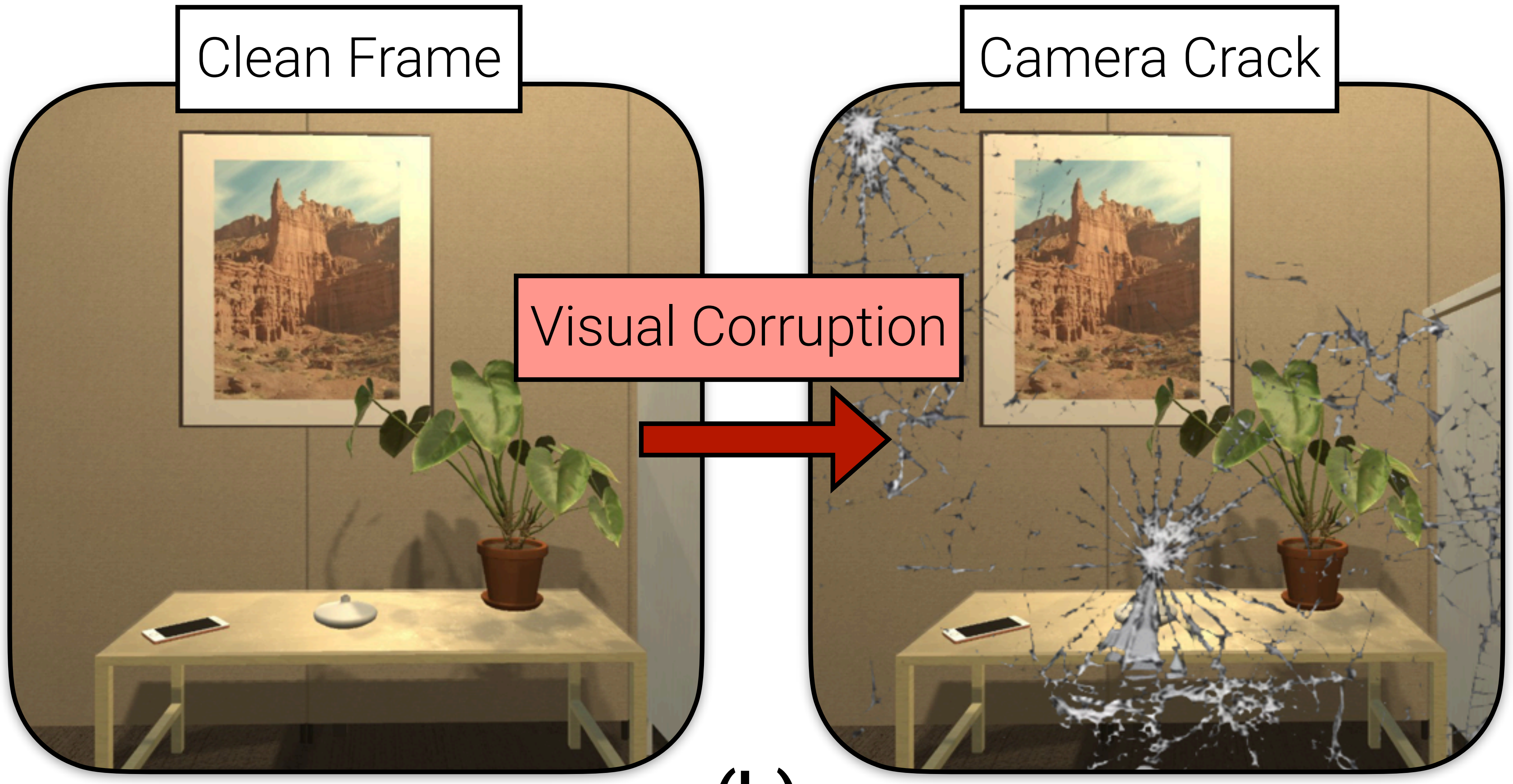
RobustNav

7 visual corruptions
at **5** levels of severity

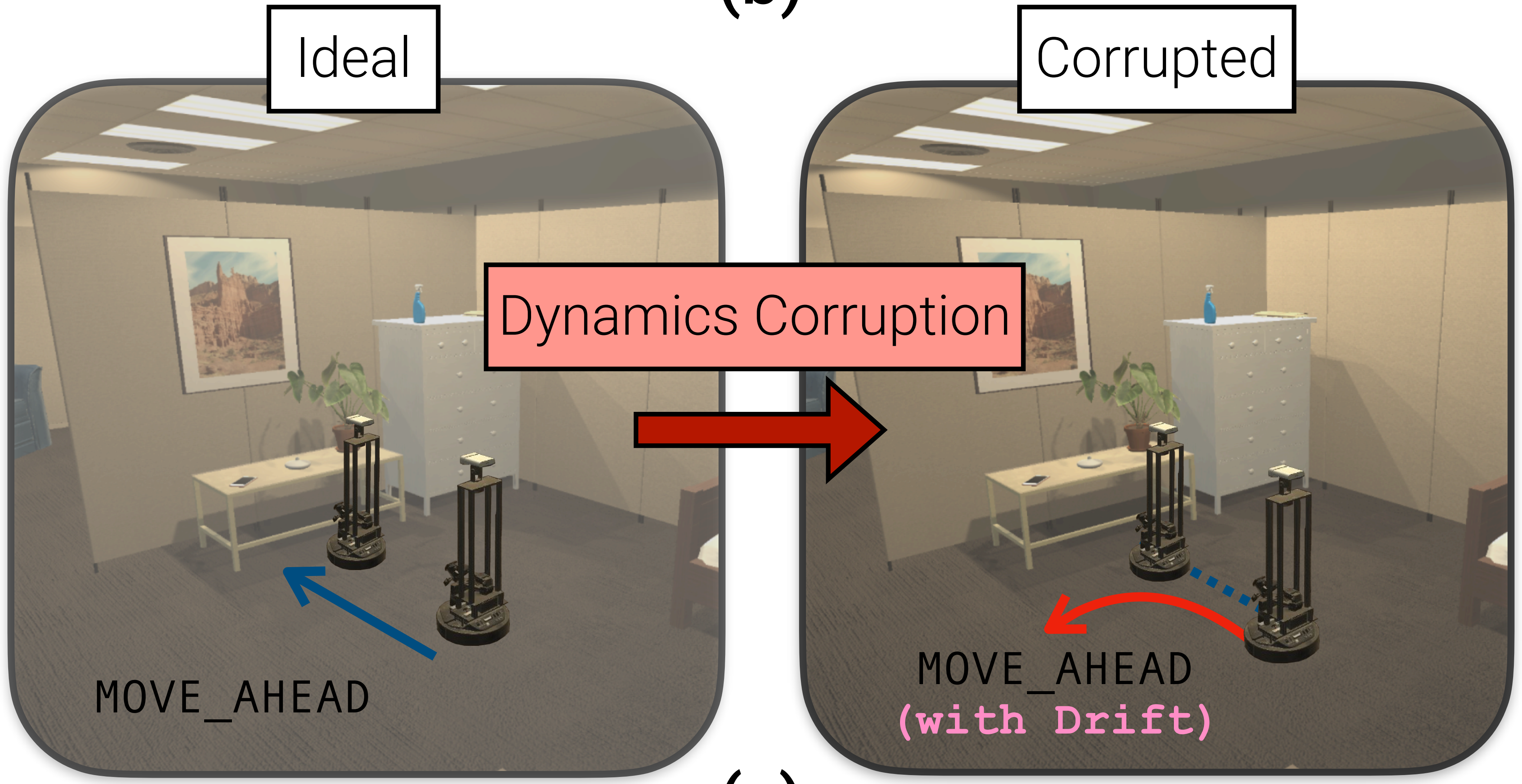
4 dynamics corruptions
Corruptions can be due to sensor or environment variations



(a)



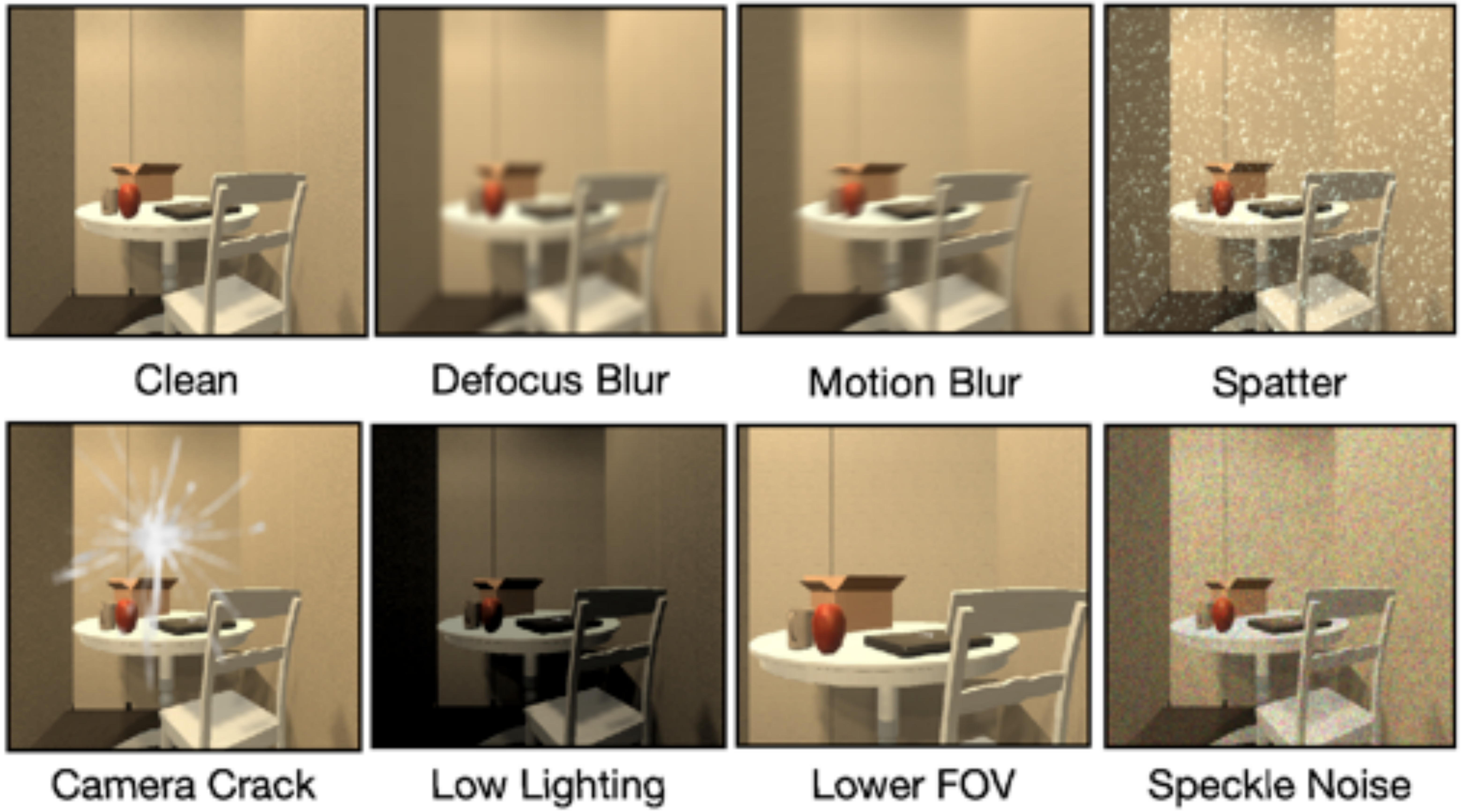
(b)



(c)

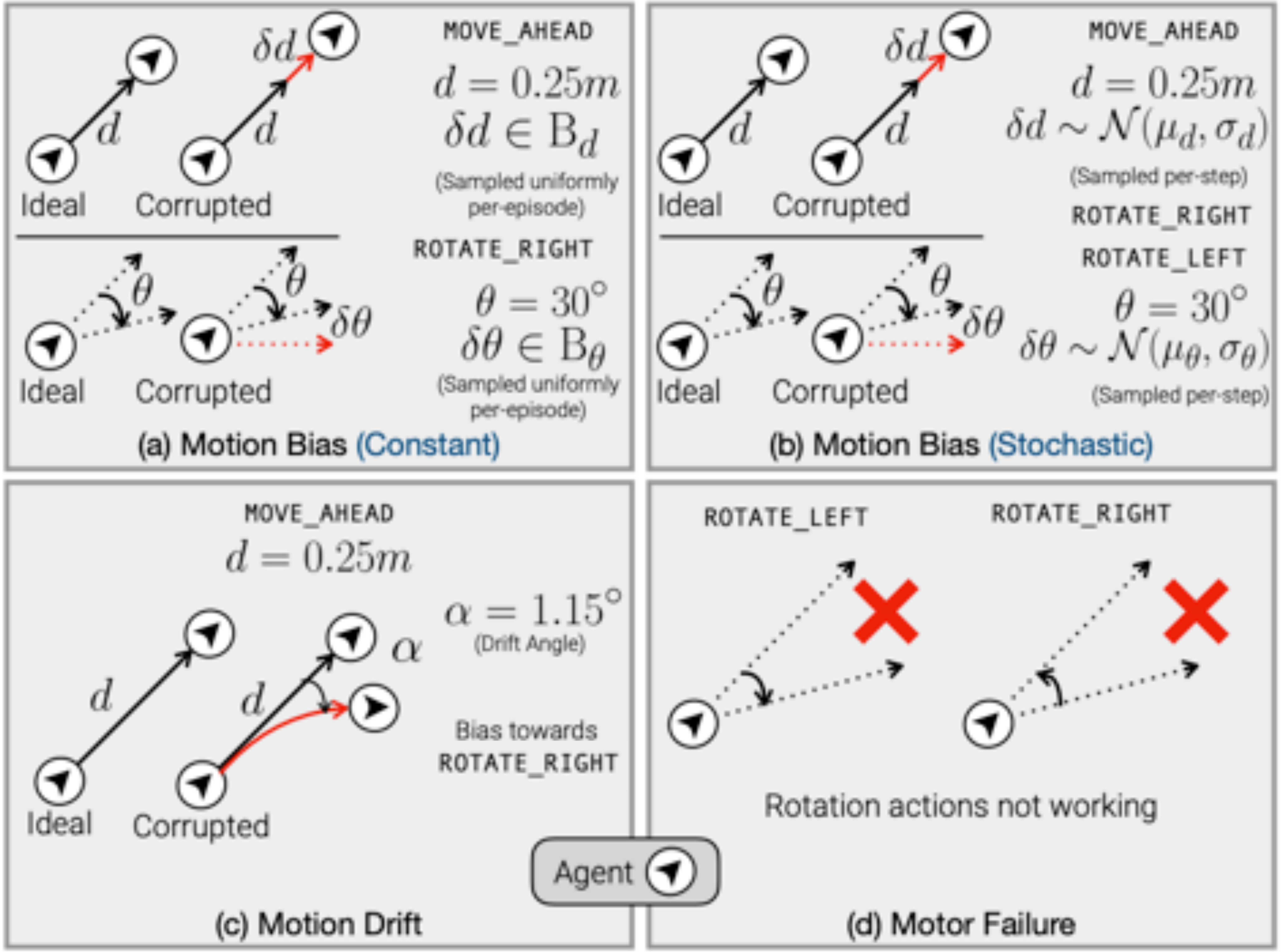
Chattopadhyay, Hoffman, Mottaghi, Kembhavi. ICCV 2021

RobustNav Visual Corruptions



Chattopadhyay, Hoffman, Mottaghi, Kembhavi. ICCV 2021

RobustNav Dynamics Corruptions

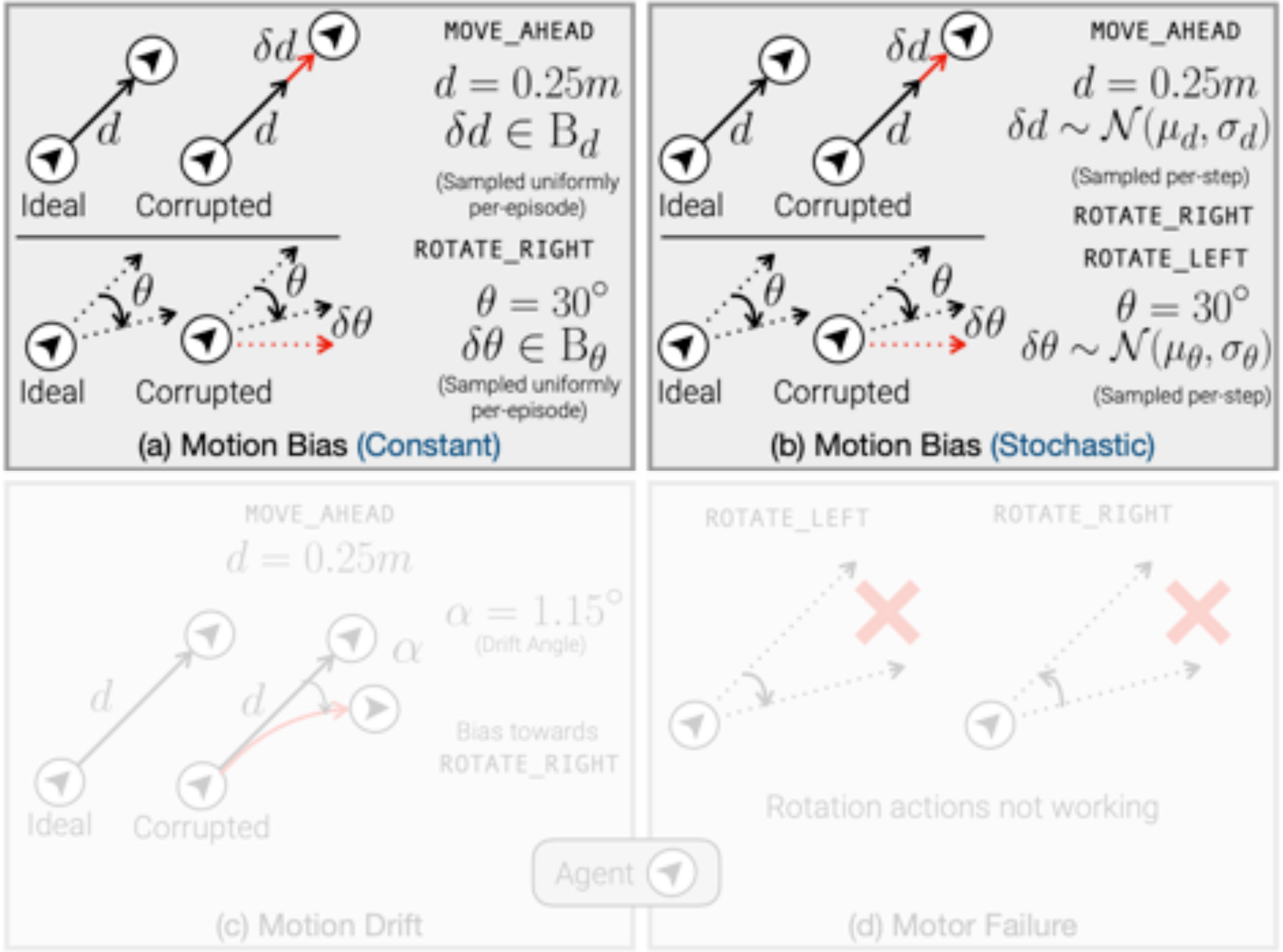


Chattopadhyay, Hoffman, Mottaghi, Kembhavi. ICCV 2021

RobustNav Dynamics Corruptions

Due to Environment

Scene-level friction



High and low friction zones

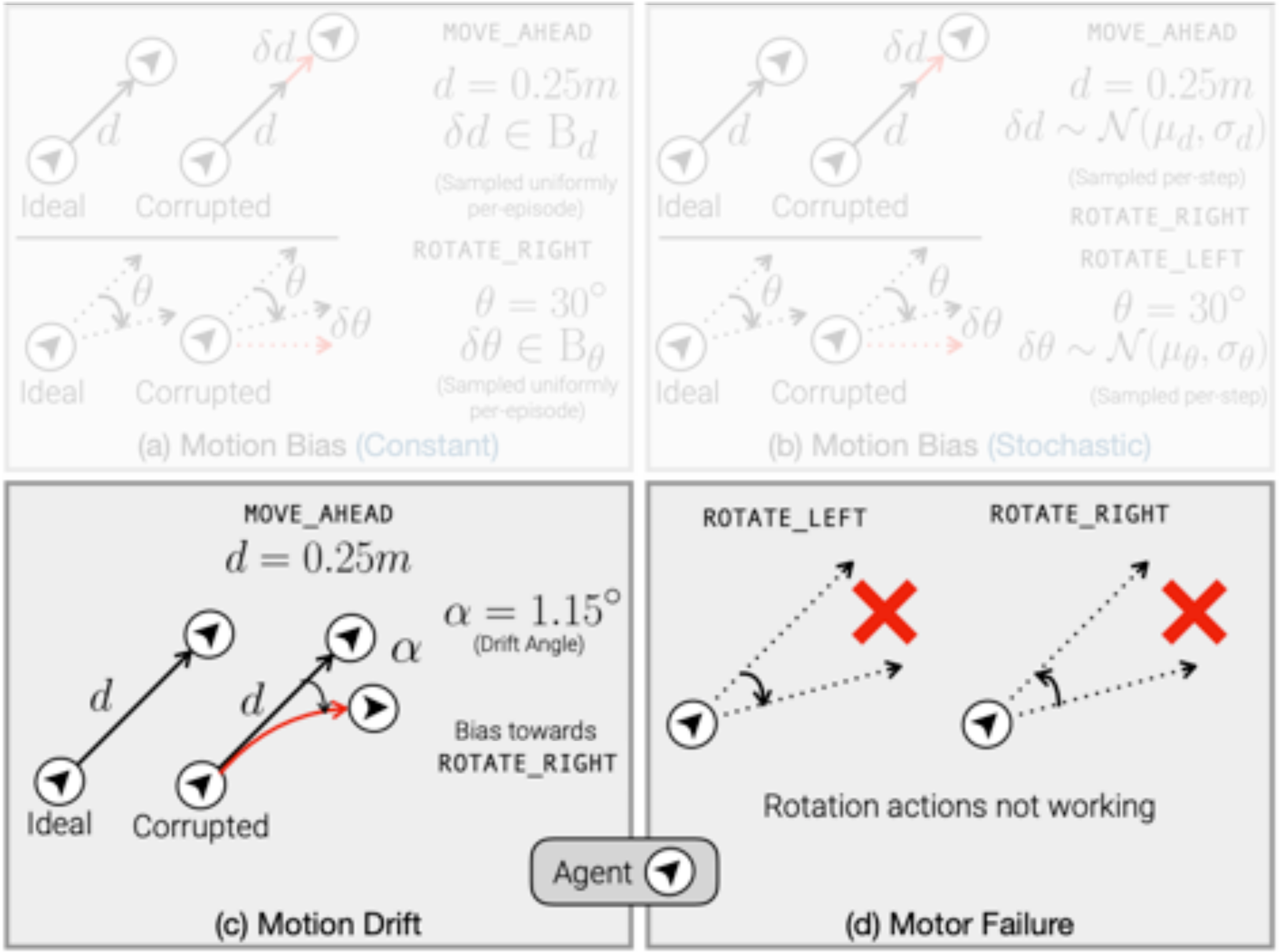
Chattopadhyay, Hoffman, Mottaghi, Kembhavi. ICCV 2021

RobustNav Dynamics Corruptions

Due to Environment

Scene-level friction

High and low friction zones



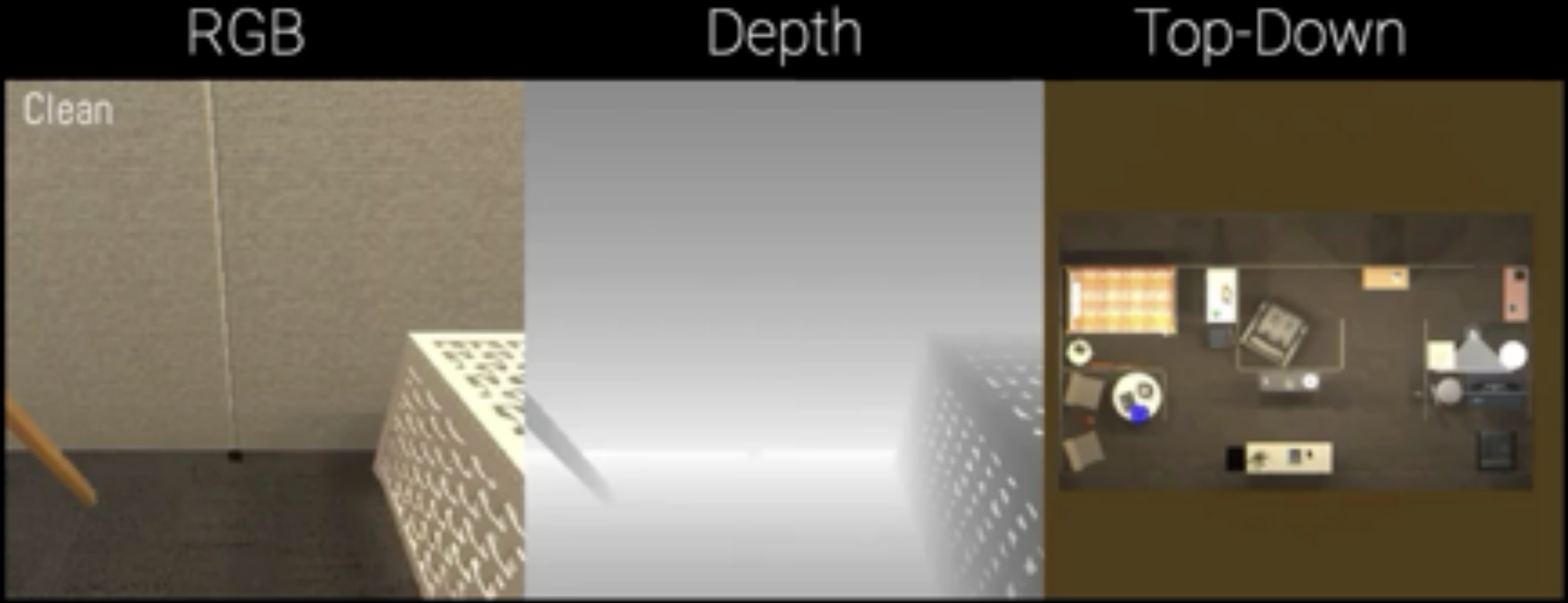
Due to Faulty Movements

Malfunctioning components

Chattopadhyay, Hoffman, Mottaghi, Kembhavi. ICCV 2021

ObjectNav RGBD — Target Object in "Blue"

Clean Conditions
(Success = True)



Chattopadhyay, Hoffman, Mottaghi, Kembhavi. ICCV 2021

Synthetic to Real Pixel Adaptation

Train

GTA (synthetic)

Test

CityScapes (Germany)

Hoffman et.al. ICML 2018

Domain Adaptation: Train on Source Test on Target

Domain Adaptation: Train on Source Test on Target

Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Domain Adaptation: Train on Source Test on Target

Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Domain Adaptation: Train on Source Test on Target

Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Domain Adaptation: Train on Source Test on Target

Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Domain Adaptation: Train on Source Test on Target

Source Domain $\sim P_S(X_S, Y_S)$
lots of **labeled** data

$\neq$

Target Domain $\sim P_T(X_T, Y_T)$
unlabeled or limited labels

Domain Adaptation: Train on Source Test on Target

$\neq$

Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Target Domain $\sim P_T(X_T, Y_T)$

unlabeled or limited labels

Domain Adaptation: Train on Source Test on Target

$\neq$

Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Target Domain $\sim P_T(X_T, Y_T)$

unlabeled or limited labels

Domain Adaptation: Train on Source Test on Target

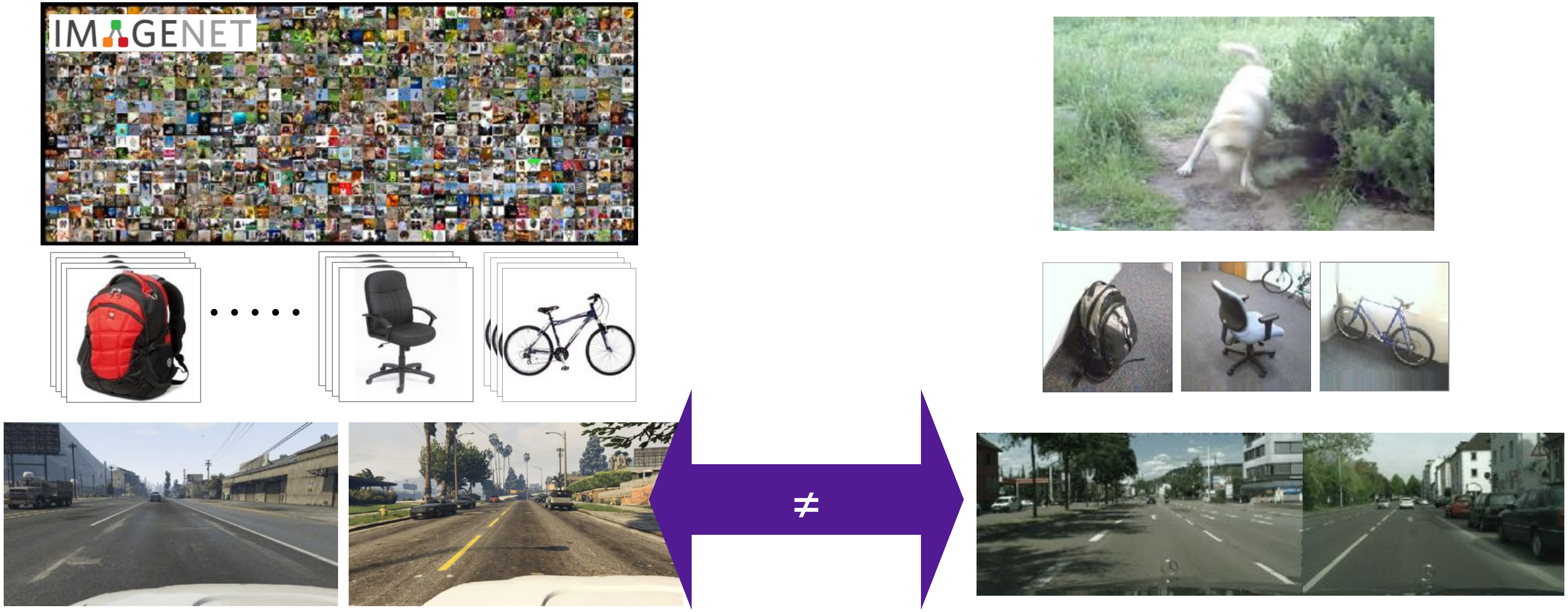
Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Target Domain $\sim P_T(X_T, Y_T)$

unlabeled or limited labels

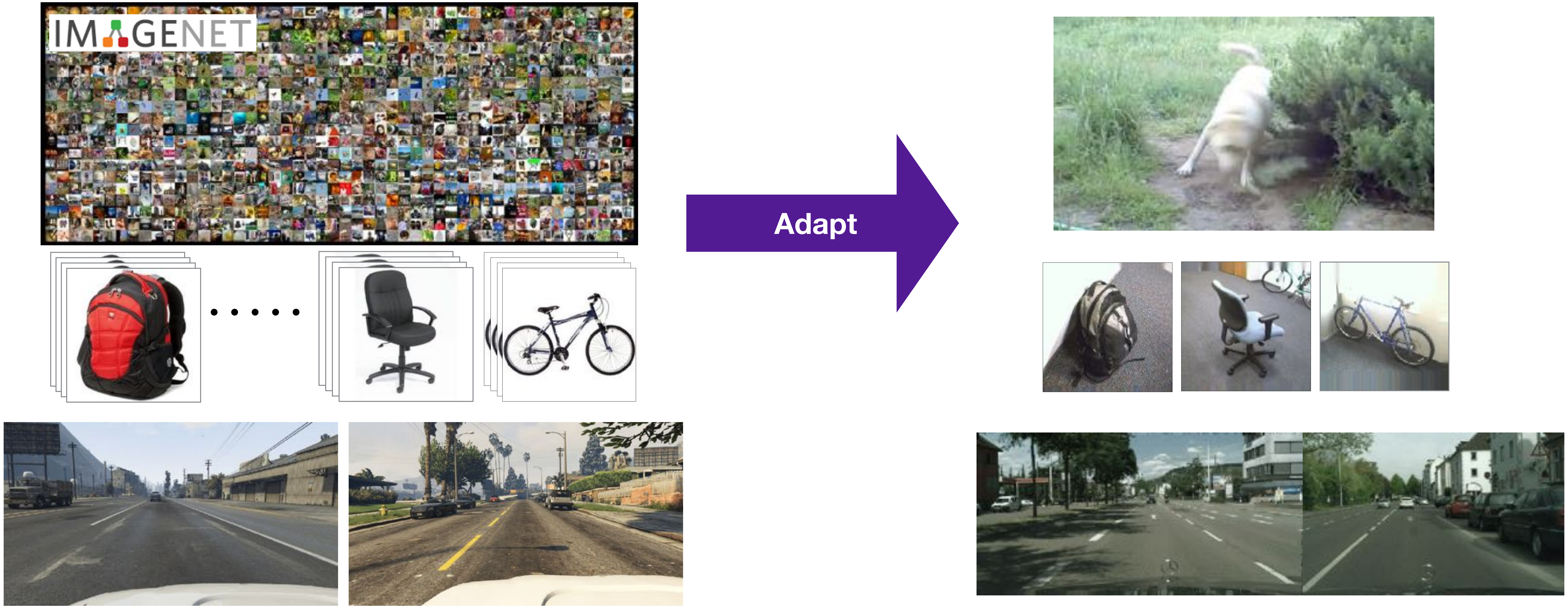
Domain Adaptation: Train on Source Test on Target



Source Domain $\sim P_S(X_S, Y_S)$
lots of **labeled** data

Target Domain $\sim P_T(X_T, Y_T)$
unlabeled or limited labels

Domain Adaptation: Train on Source Test on Target



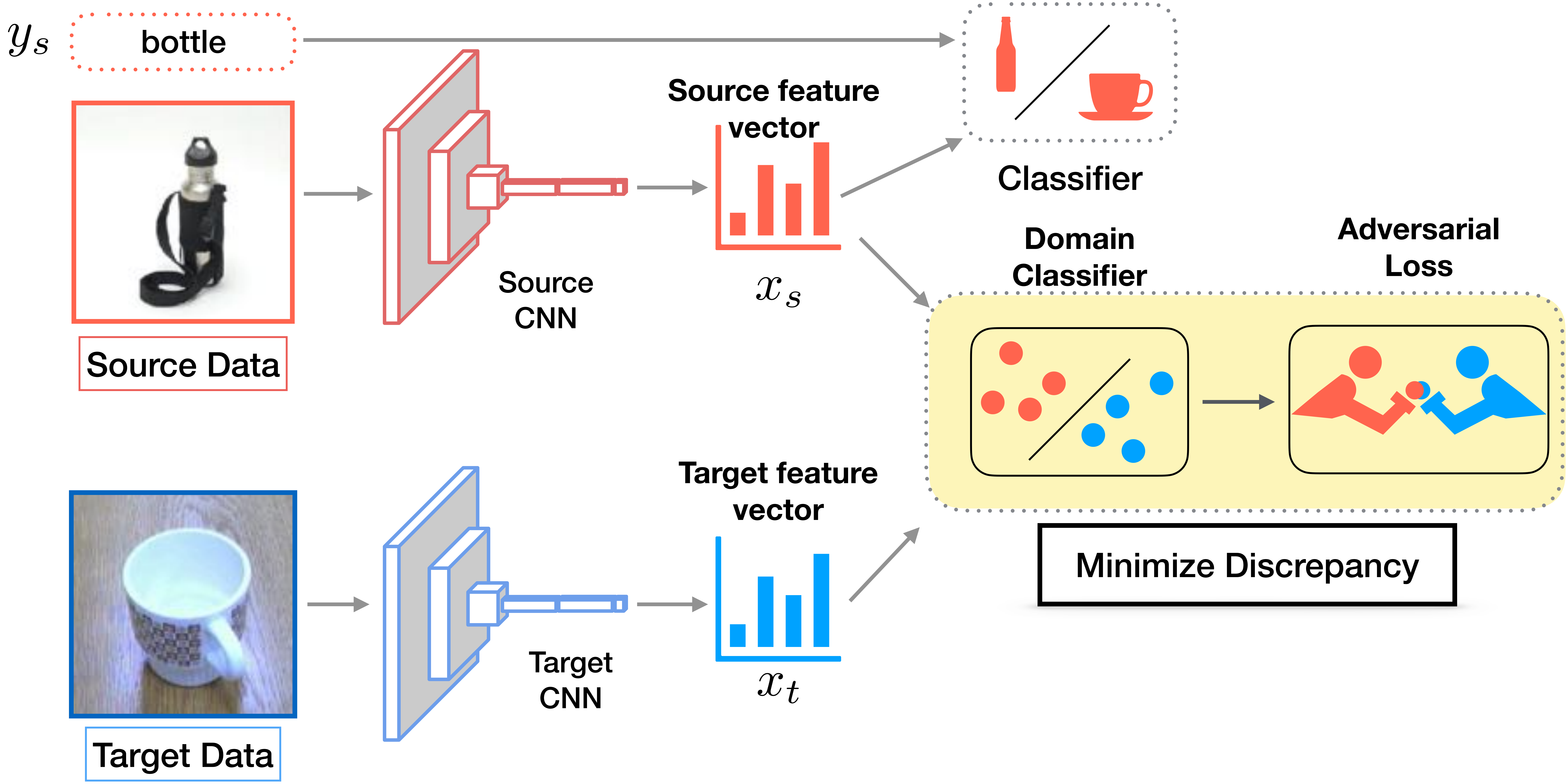
Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Target Domain $\sim P_T(X_T, Y_T)$

unlabeled or limited labels

Domain Adversarial Adaptation

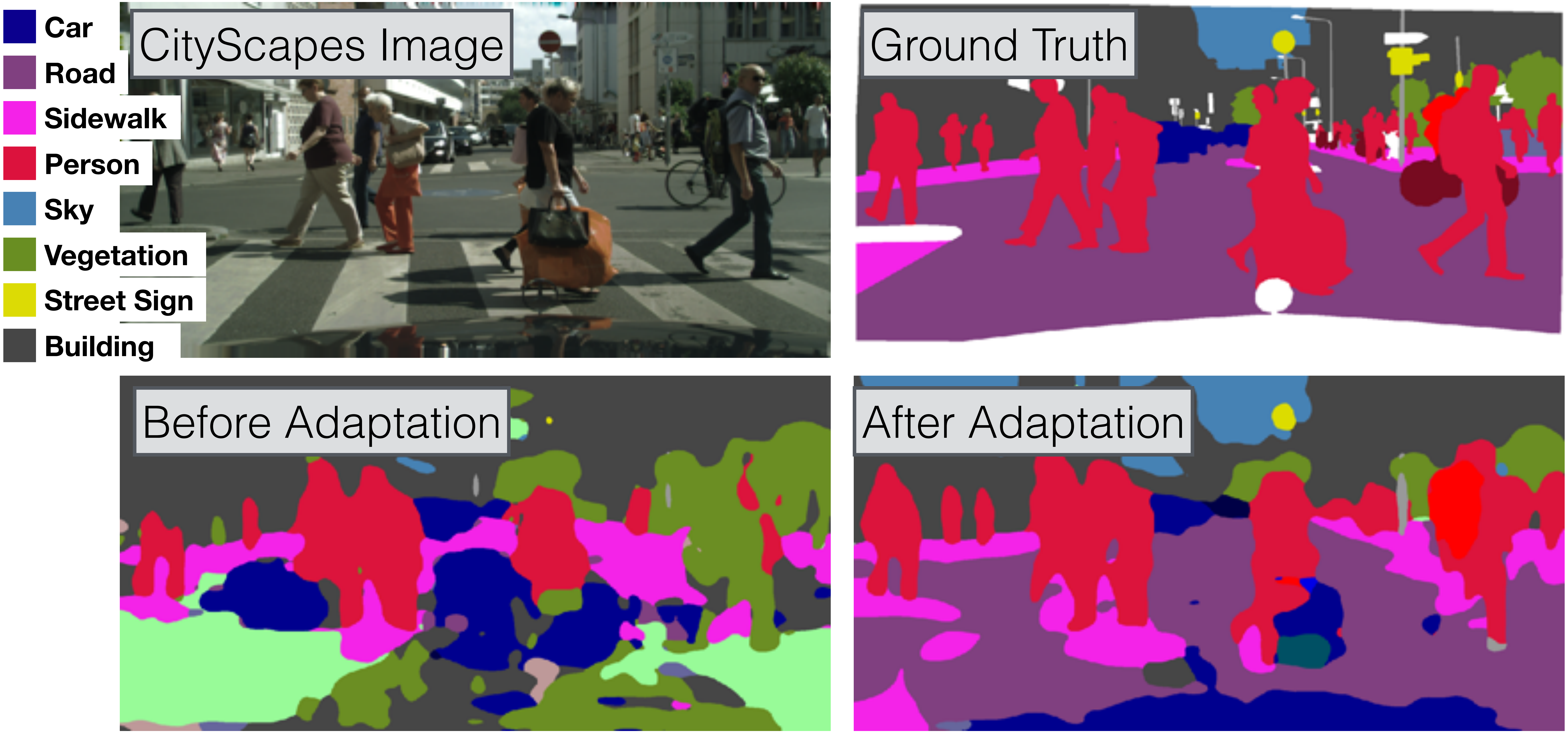


Ganin ICML 2015, Long ICML 2015, Tzeng et al ICCV 2015, Tzeng et al CVPR 2017, Hoffman et al ICML 2018

Synthetic to Real Pixel Adaptation

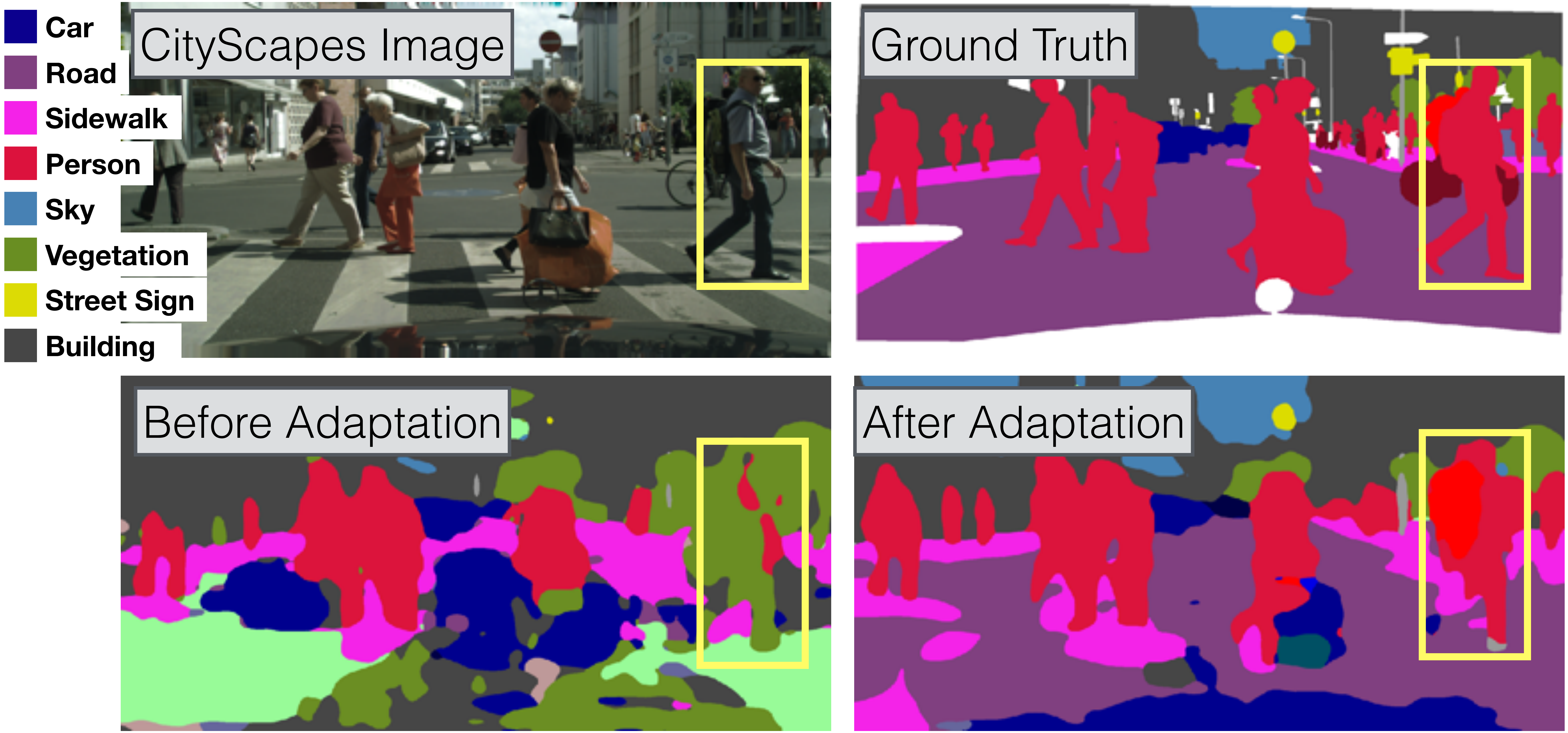
Hoffman et.al. ICML 2018

CyCADA Results: CityScapes Evaluation



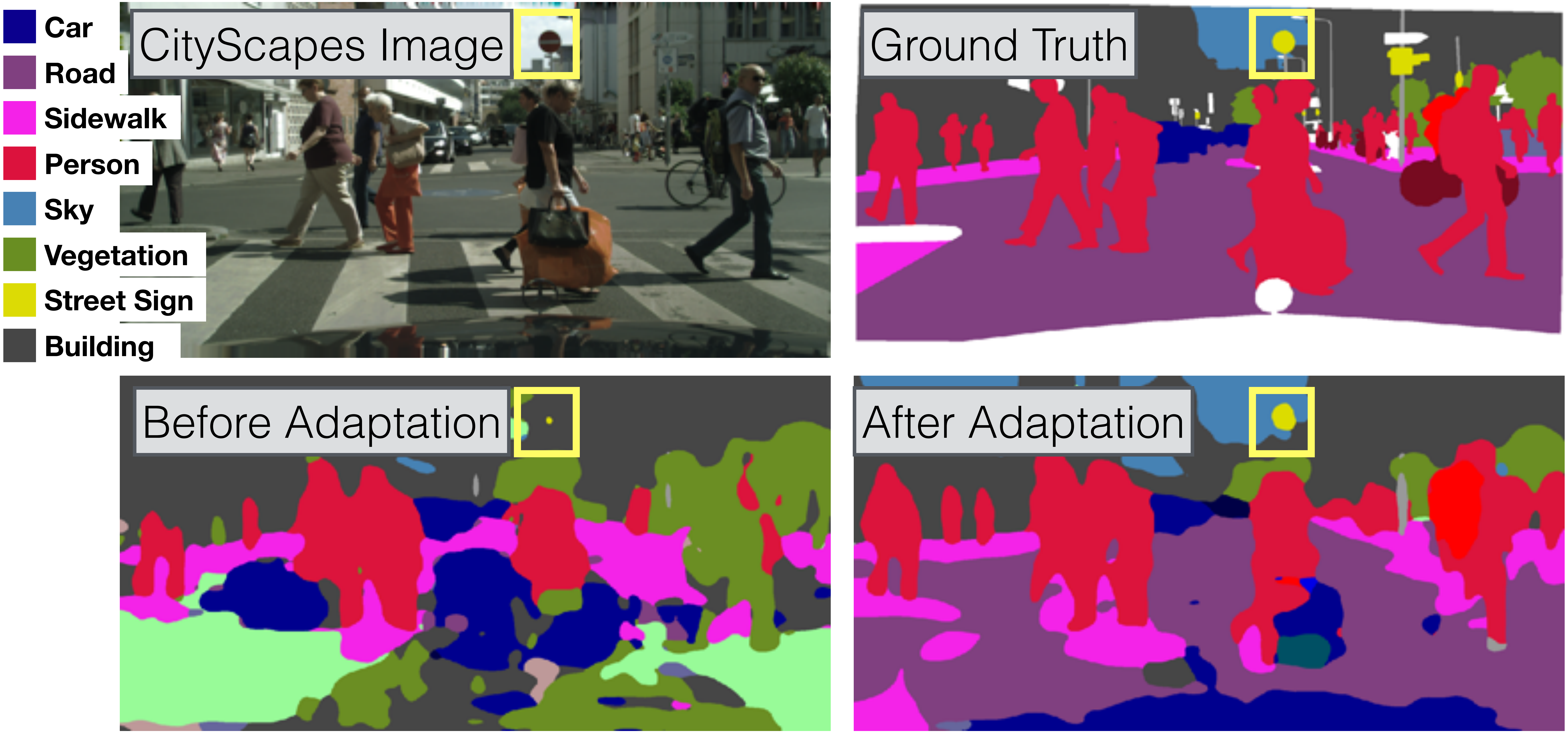
Hoffman et.al. ICML 2018

CyCADA Results: CityScapes Evaluation



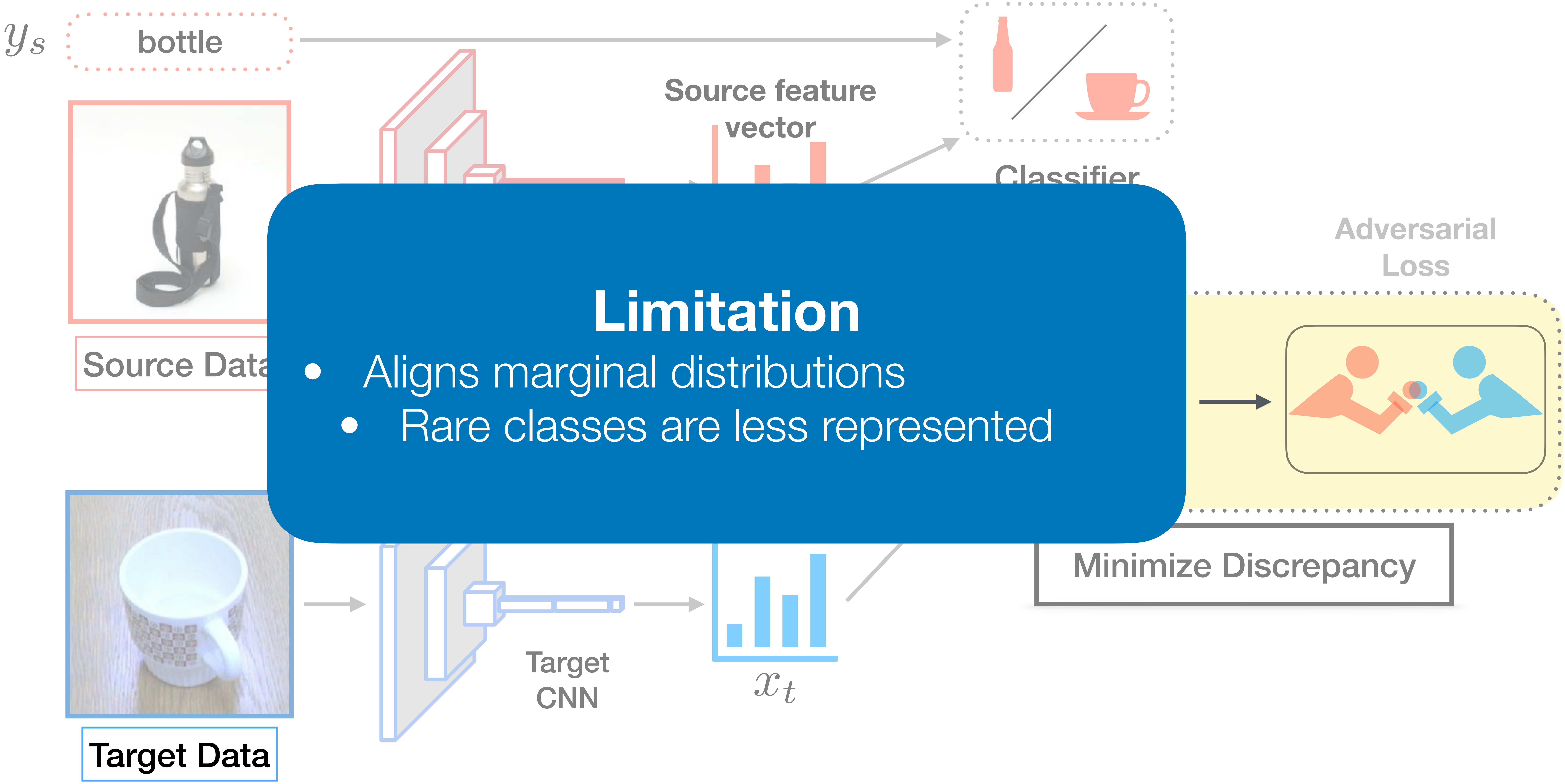
Hoffman et.al. ICML 2018

CyCADA Results: CityScapes Evaluation



Hoffman et.al. ICML 2018

Domain Adversarial Adaptation



Limitation

- Aligns marginal distributions
  - Rare classes are less represented

Tzeng et al ICCV 2015, Tzeng et al CVPR 2017, Hoffman et al ICML 2018

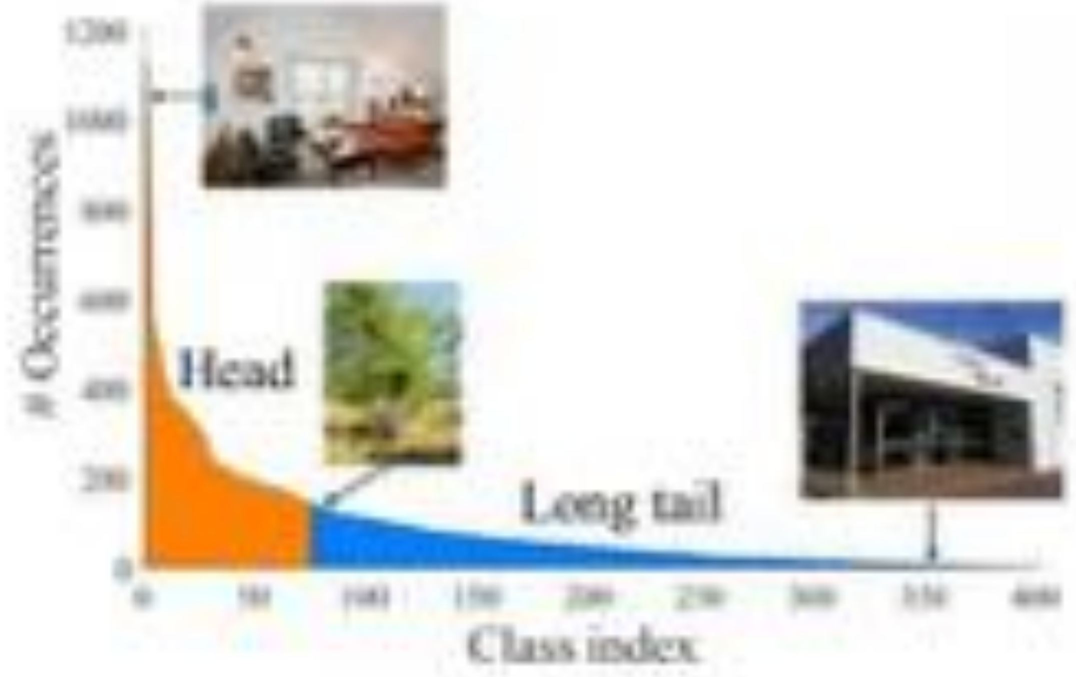
Adapting to Imbalanced Data

Source data may be curated to be balanced

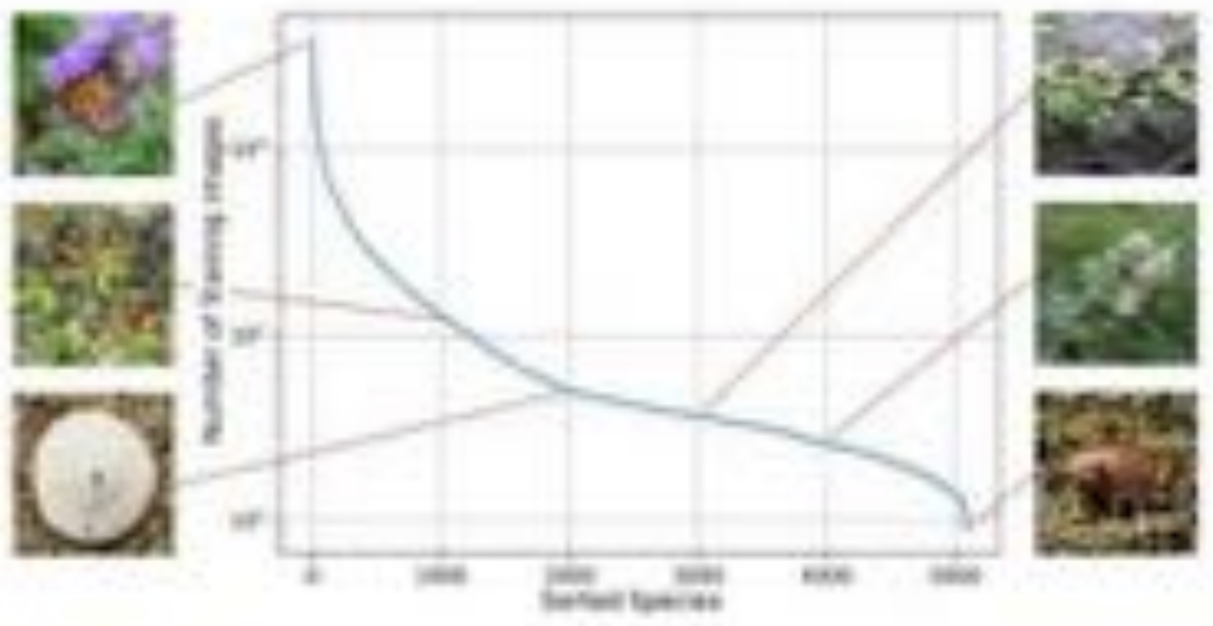
We have no control over target datasets!

Goal: Adapt under both data and label distribution shift

Faces [Zhang et al. 2017]



Places [Wang et al. 2017]



Species [Van Horn et al. 2019]

Actions [Zhang et al. 2019]

Adapting to Imbalanced Data

- **Challenge:** Existing DA methods (eg. domain adversarial) struggle in this setting!
  - Implicitly assume[1,2] similar label distributions

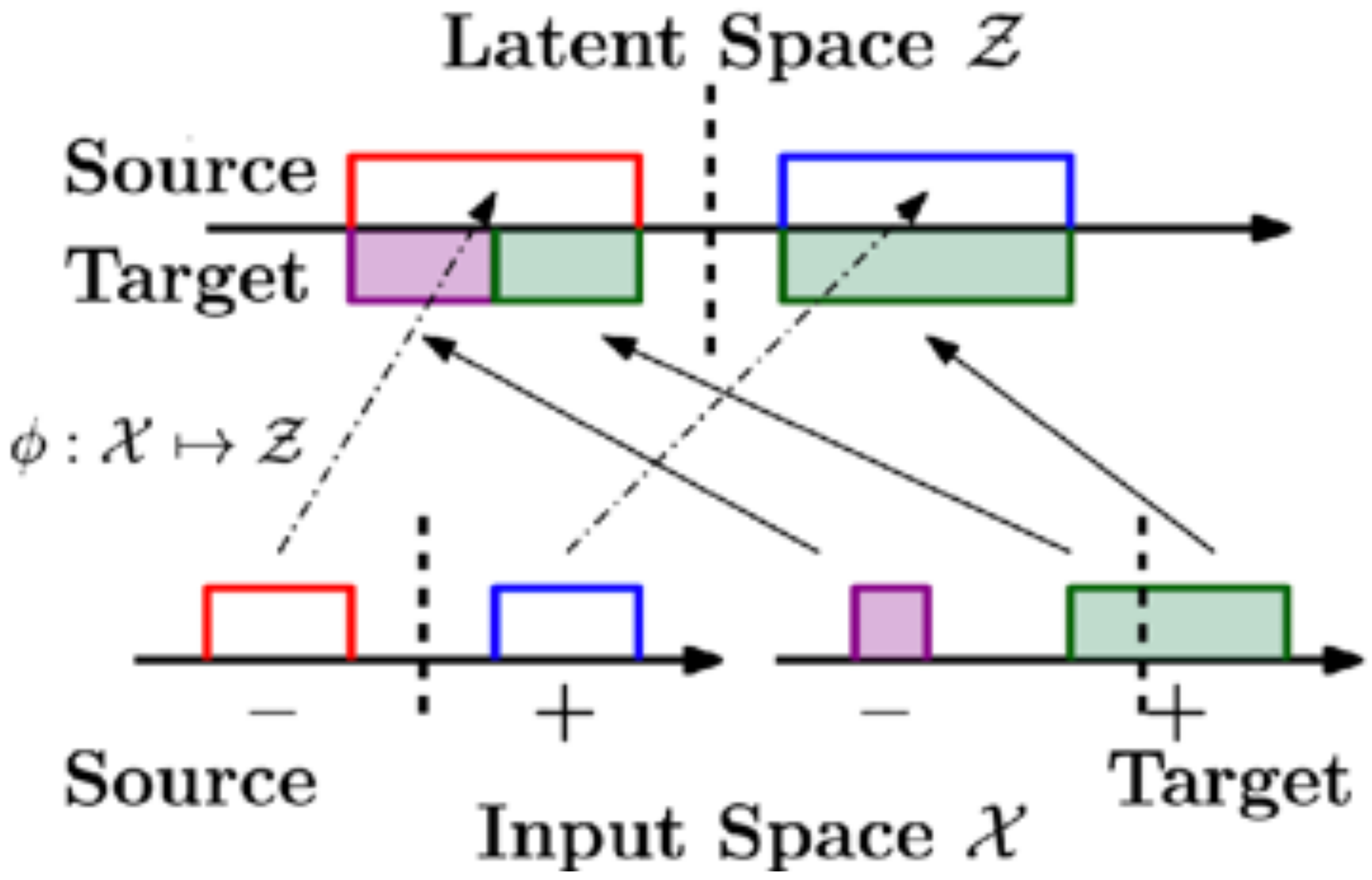


Figure credit: Wu *et al.*, ICML 2019

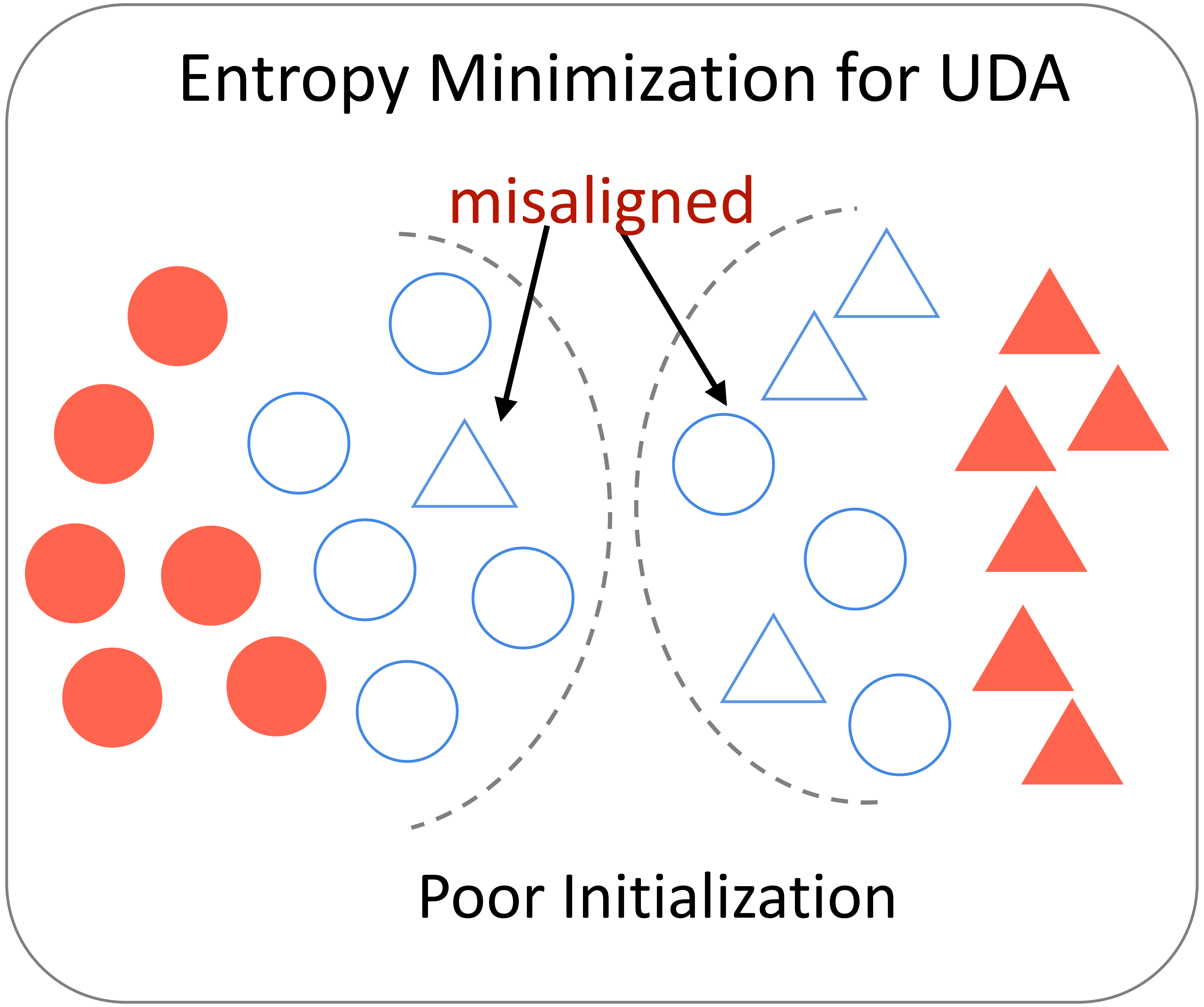
- We turn to simpler DA approaches based on **self-training**[3,4]
  - **Algorithm:** Training on model predictions on unlabeled target
  - No requirement of similar source/target label distributions

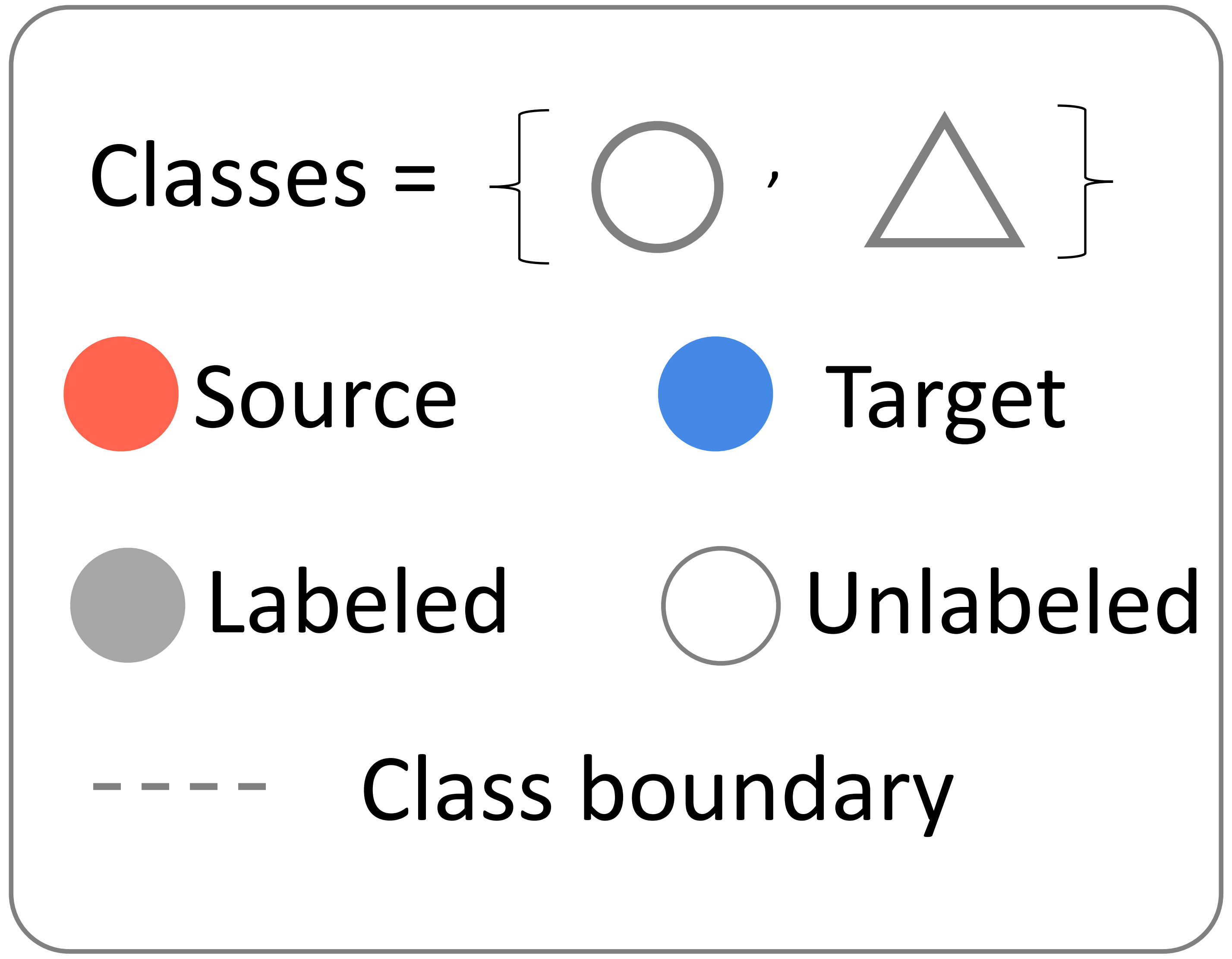
1. Wu *et al.*, ICML 2019. 2. Li *et al.*, arXiv 2020. 3. Grandvalet *et al.*, NeurIPS 2004. 4. Tan *et al.*, ECCVW 2020

Adaptation with Self-Training

- **Domain Shift**: Target data is misaligned

- Entropy minimization can reinforce errors

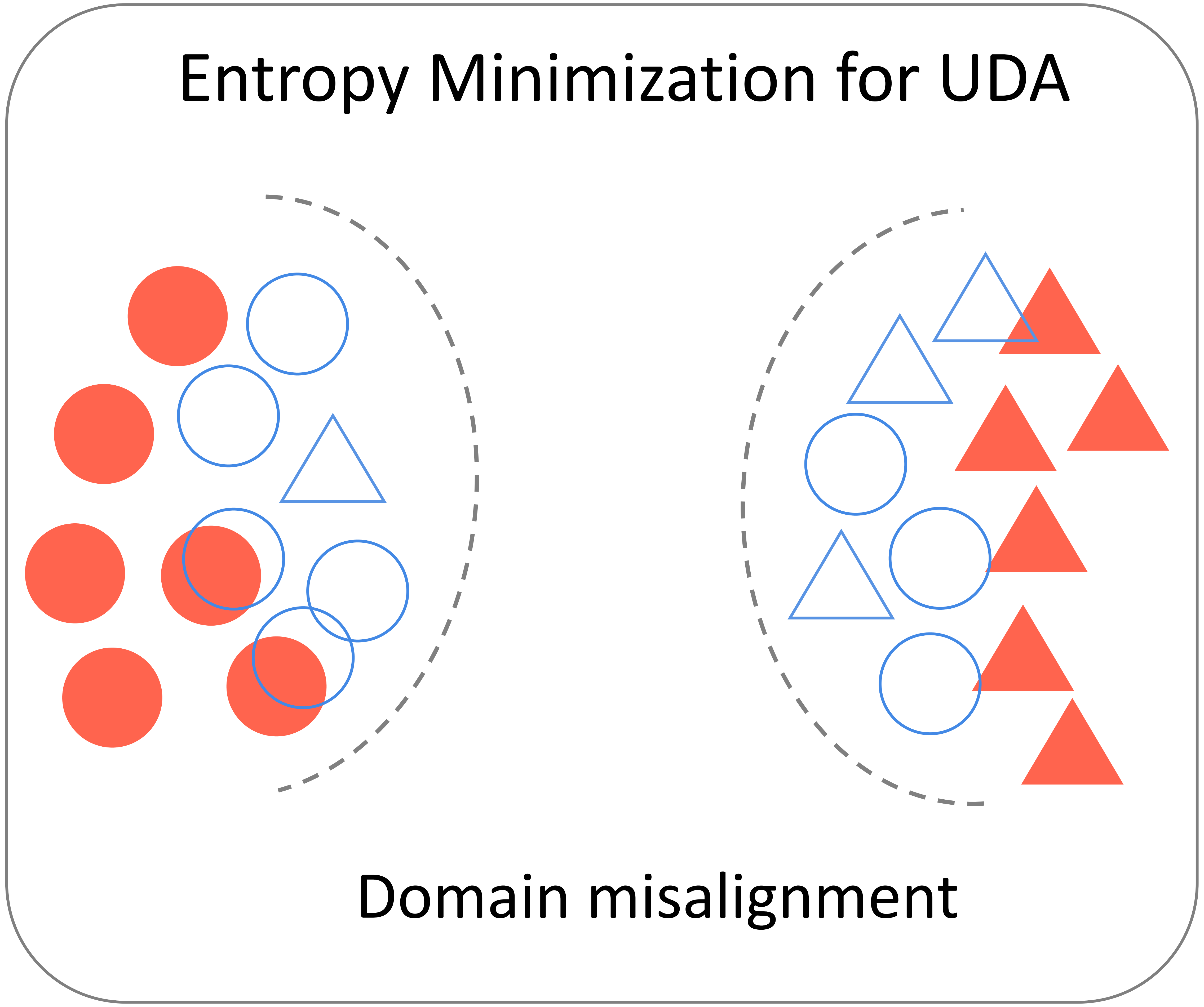


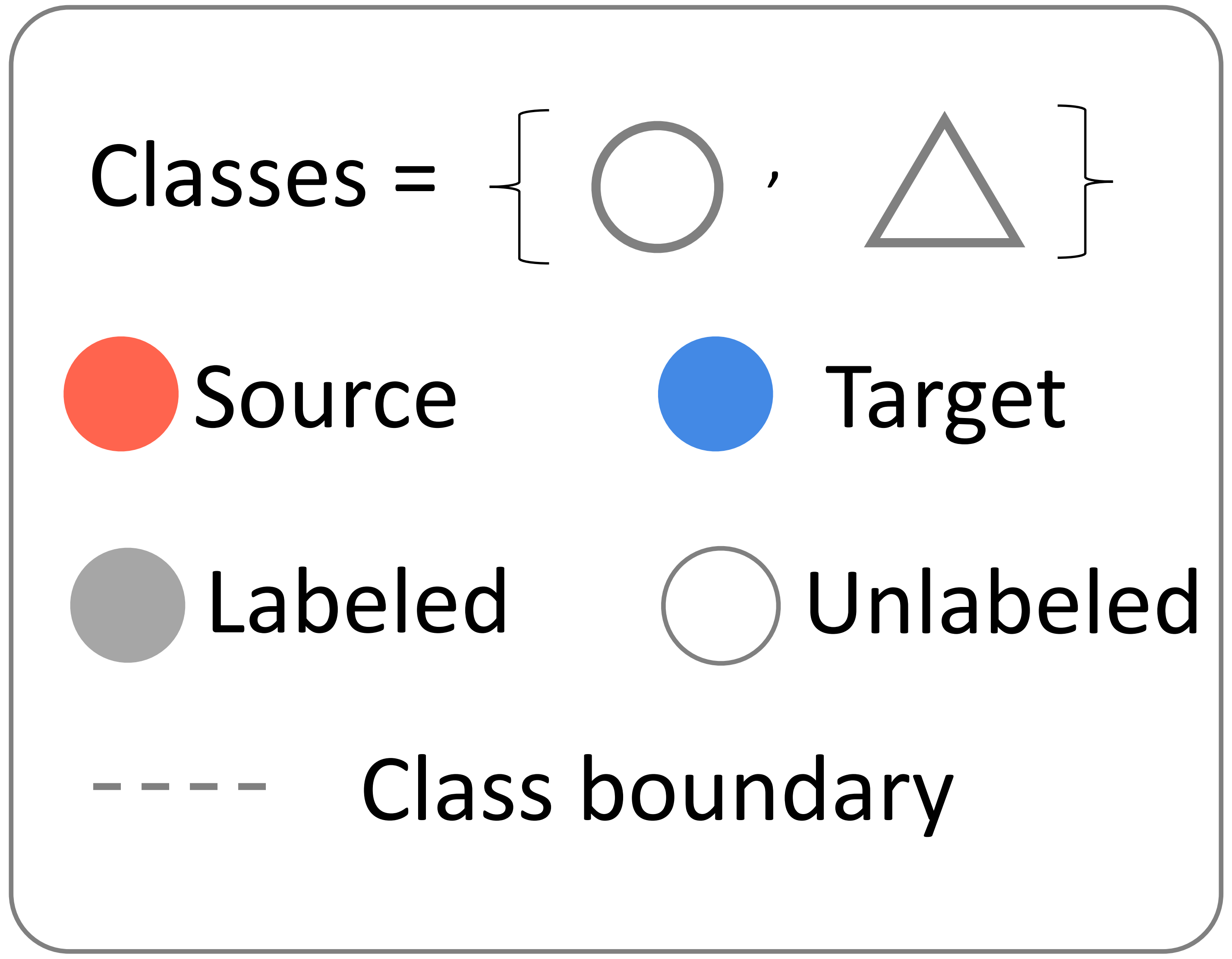


$$\mathcal{L}_{CEM} = \mathbb{E}_{\mathbf{x}_\mathcal{T}\sim\mathcal{P}_\mathcal{T}}[\mathcal{H}_\Theta(y \mid \mathbf{x}_\mathcal{T})]$$
$$= \mathbb{E}_{\mathbf{x}_\mathcal{T}\sim\mathcal{P}_\mathcal{T}}\left[\sum_{c=1}^{C} -p_\Theta(y=c \mid \mathbf{x}_\mathcal{T}) \log p_\Theta(y=c \mid \mathbf{x}_\mathcal{T})\right]$$

Adaptation with Self-Training

- **Domain Shift**: Target data is misaligned
- Entropy minimization can reinforce errors

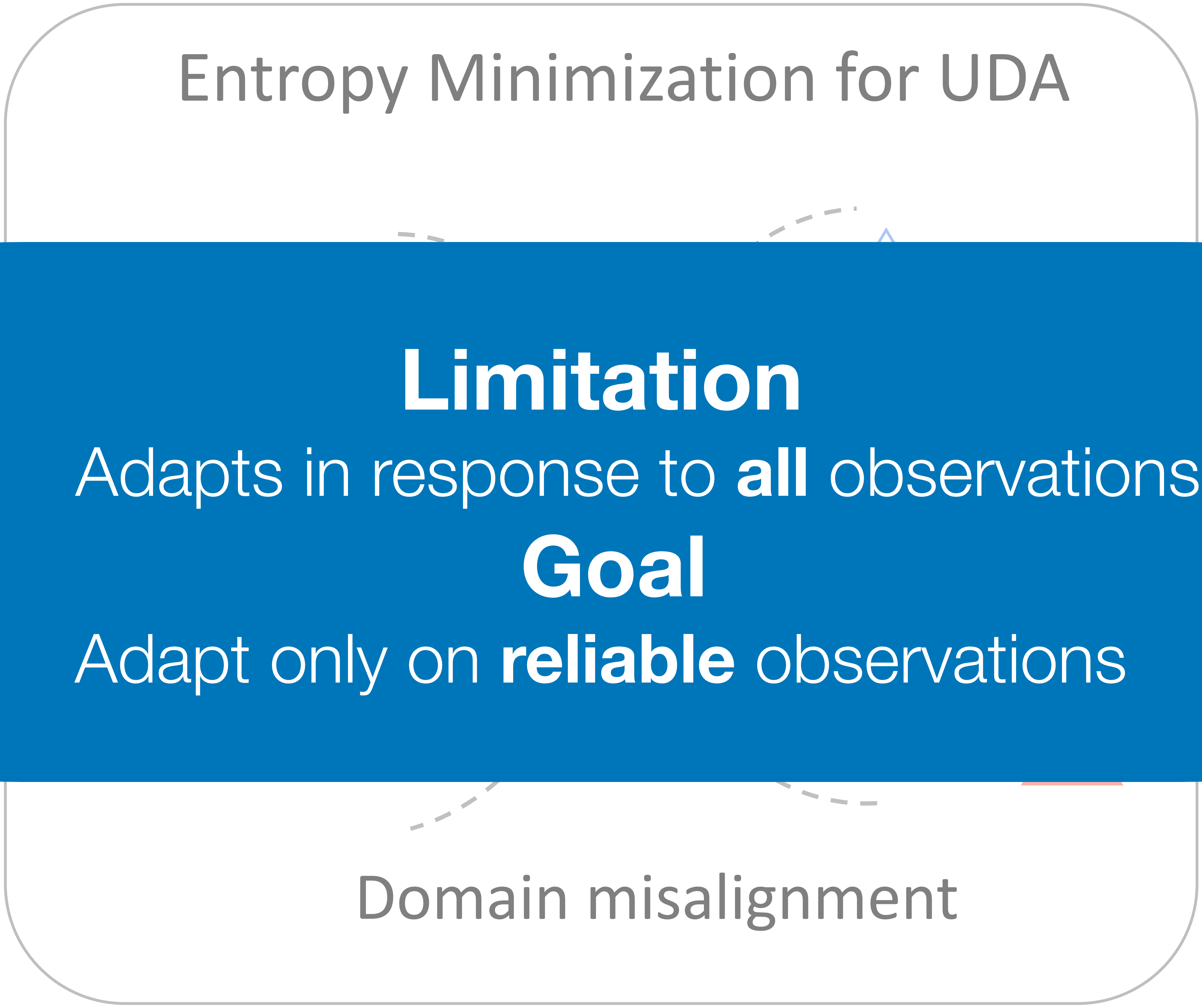


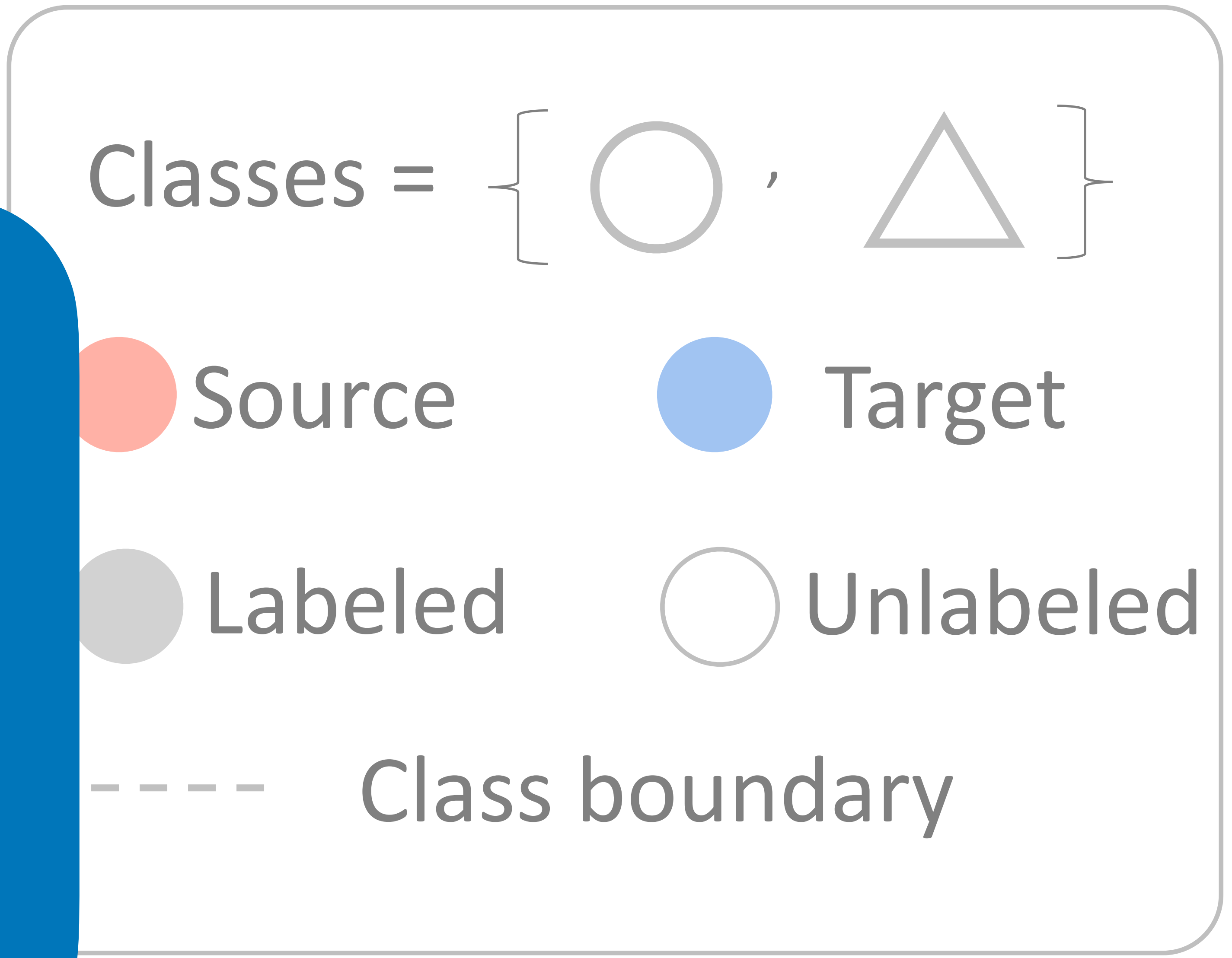


$$\mathcal{L}_{CEM} = \mathbb{E}_{\mathbf{x}_\mathcal{T} \sim \mathcal{P}_\mathcal{T}}[\mathcal{H}_\Theta(y \mid \mathbf{x}_\mathcal{T})]$$
$$= \mathbb{E}_{\mathbf{x}_\mathcal{T} \sim \mathcal{P}_\mathcal{T}}\left[\sum_{c=1}^{C} -p_\Theta(y = c \mid \mathbf{x}_\mathcal{T}) \log p_\Theta(y = c \mid \mathbf{x}_\mathcal{T})\right]$$

Adaptation with Self-Training

- **Domain Shift**: Target data is misaligned
- Entropy minimization ca reinforce errors





Limitation

- Adapts in response to **all** observations

Goal

- Adapt only on **reliable** observations

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

SENTRY
Selective Entropy Optimization via Committee Consistency for Unsupervised Domain Adaptation

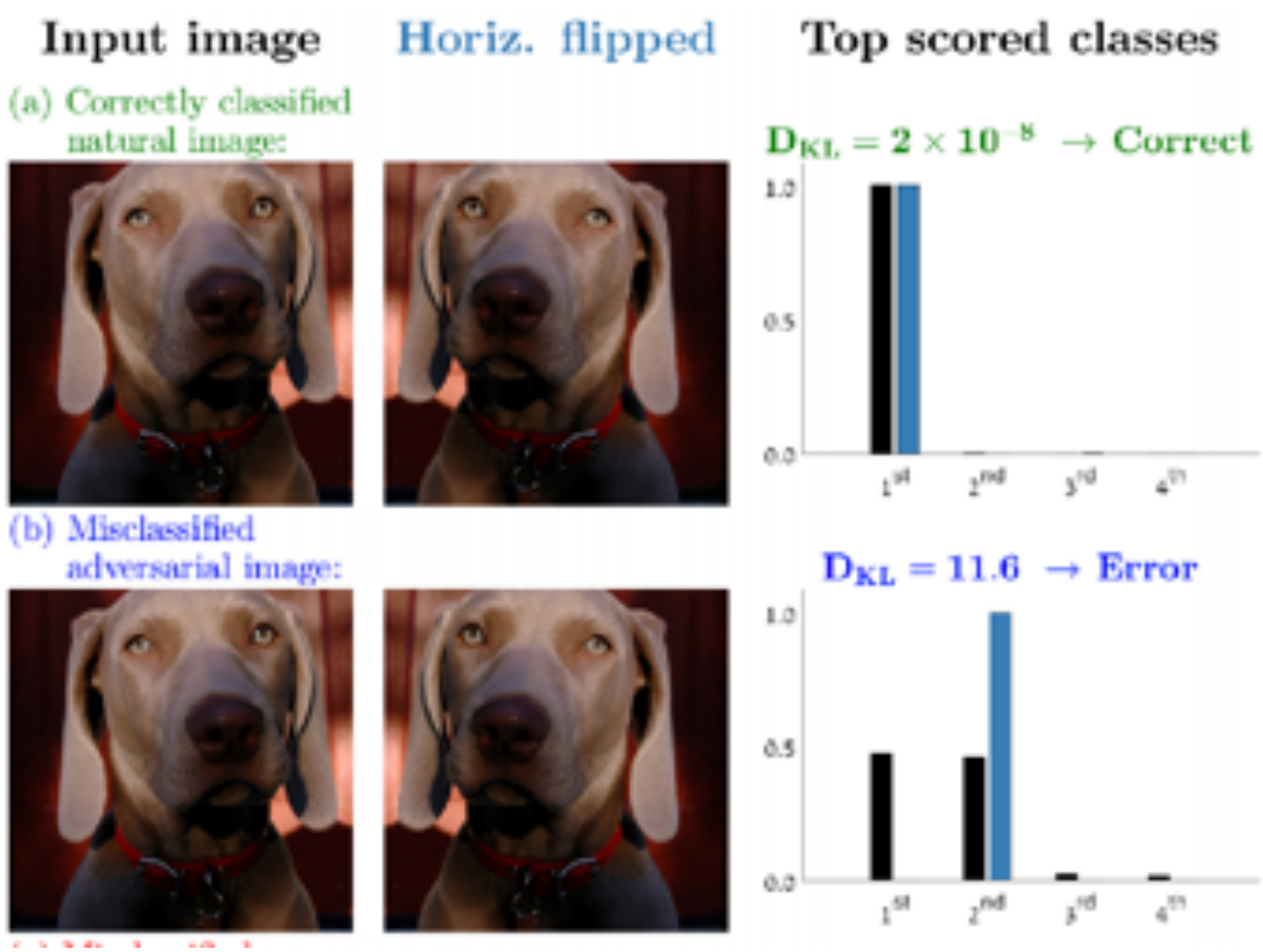



Viraj Prabhu Shivam Khare Deeksha Karthik Judy Hoffman

Georgia Tech

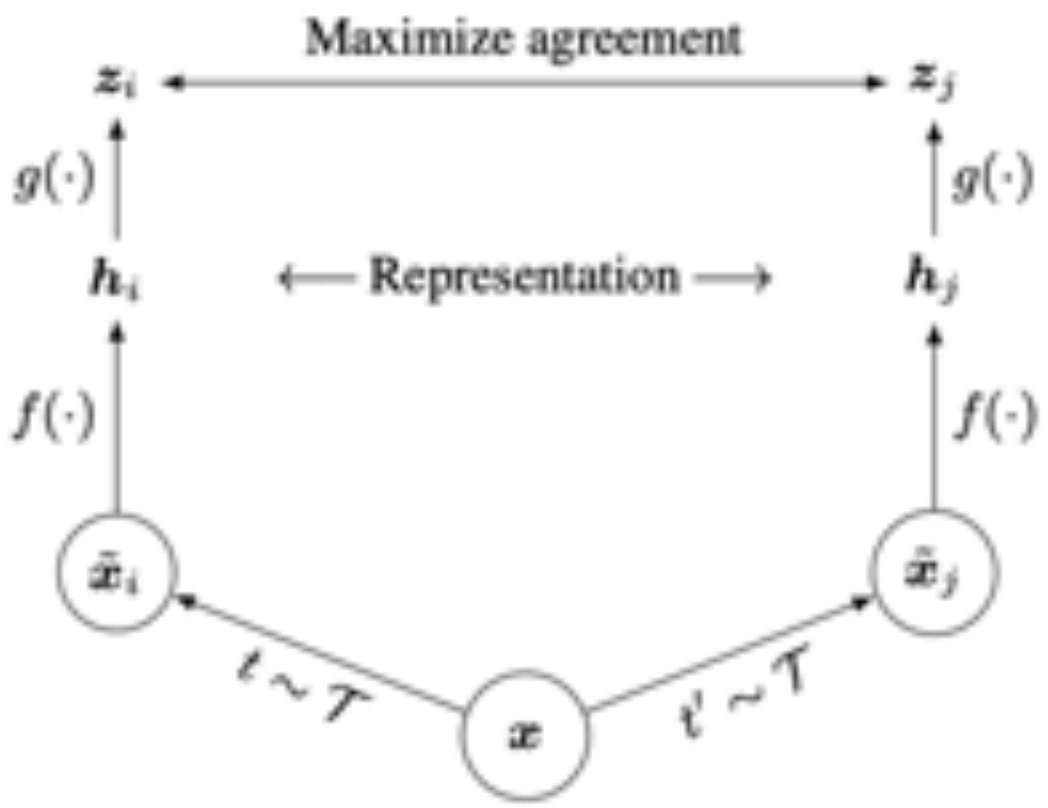
ICCV 2021

Prior Work: Predictive Consistency across Aug

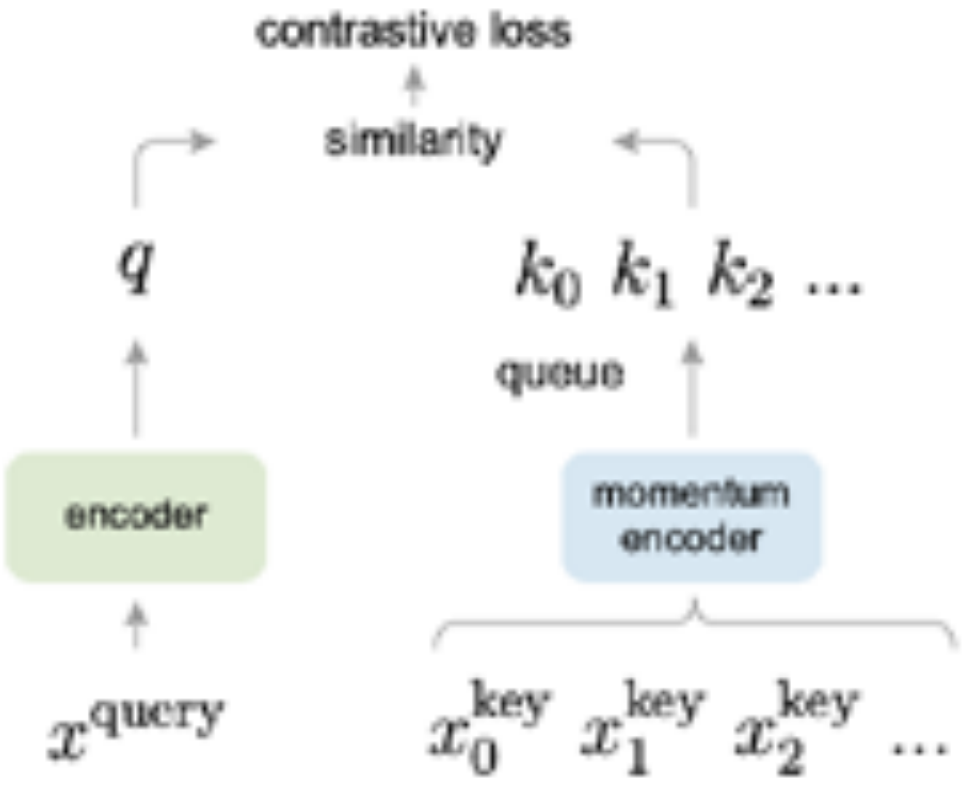


Natural and Adversarial Error Detection using Invariance to Image Transformations. Irani *et al.*, arXiv 2019

Detecting Errors



SimCLR, Chen et al.
ICML 2020



MoCo, He et al.
CVPR 2020

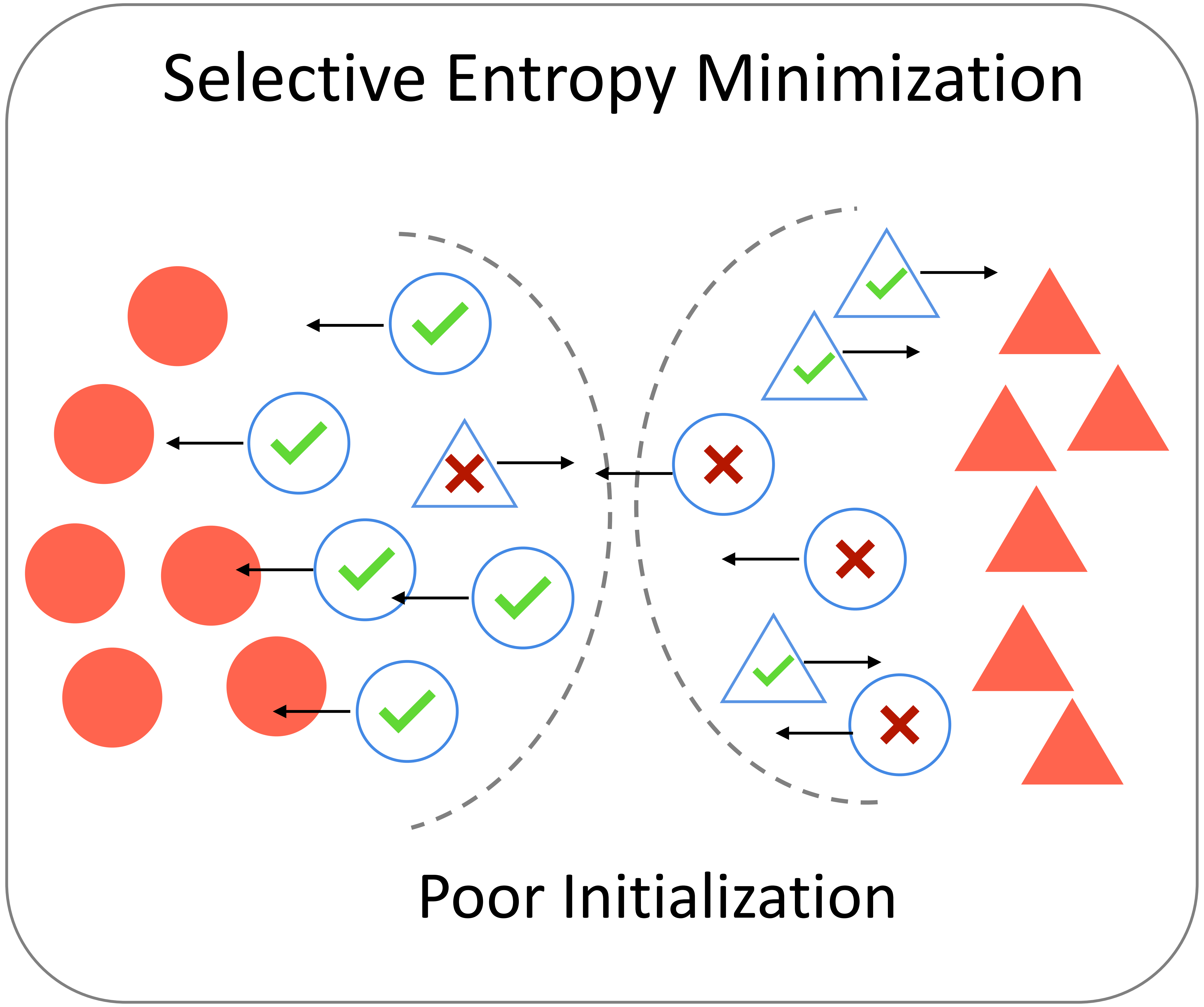
Learned Invariance
(Contrastive Learning)

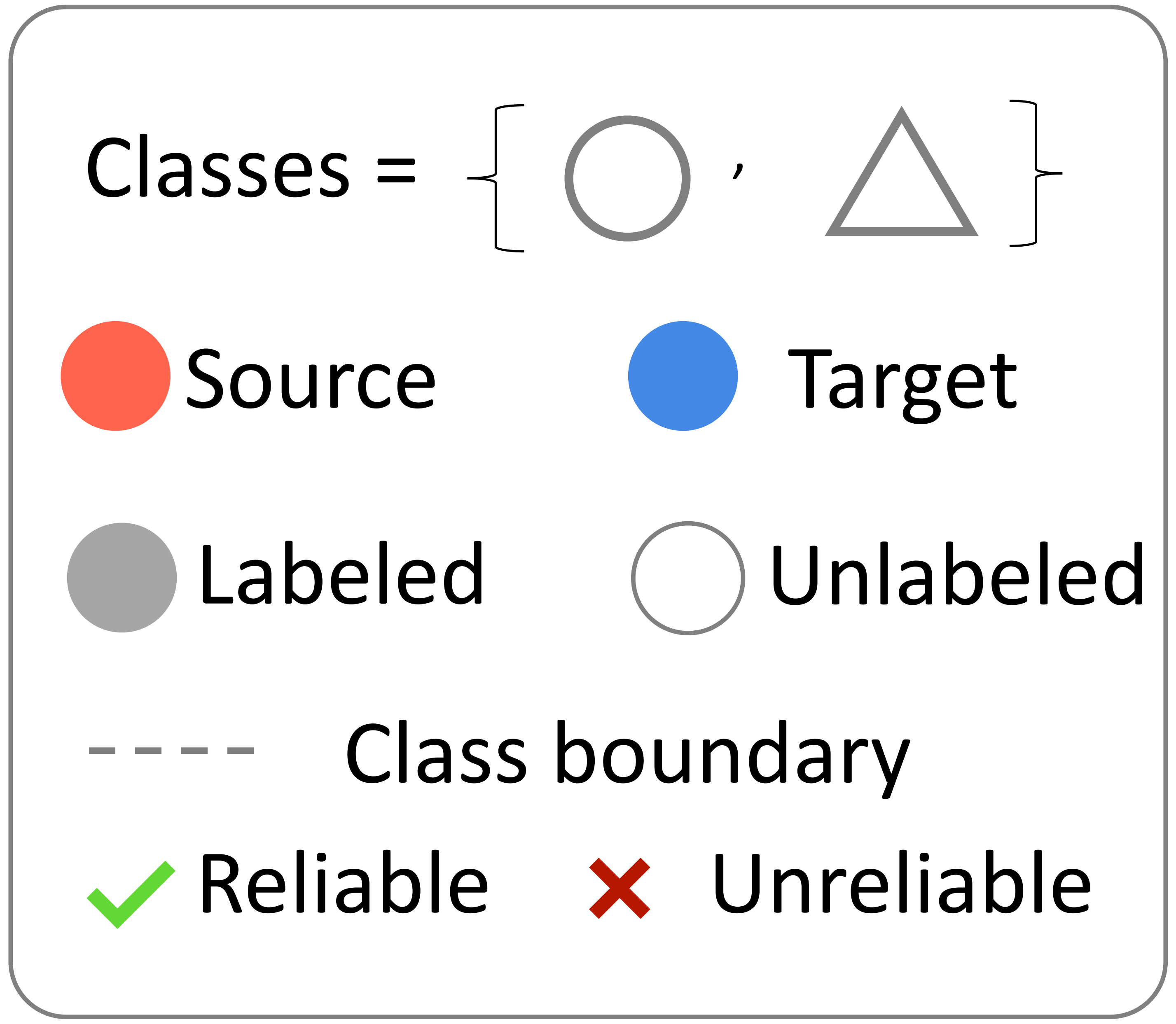
SENTRY: Selective Entropy Optimization

Key Idea

Identify reliable target instances via ~~model confidence~~ Predictive consistency[1,2,3]

Increase confidence on consistent instances





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

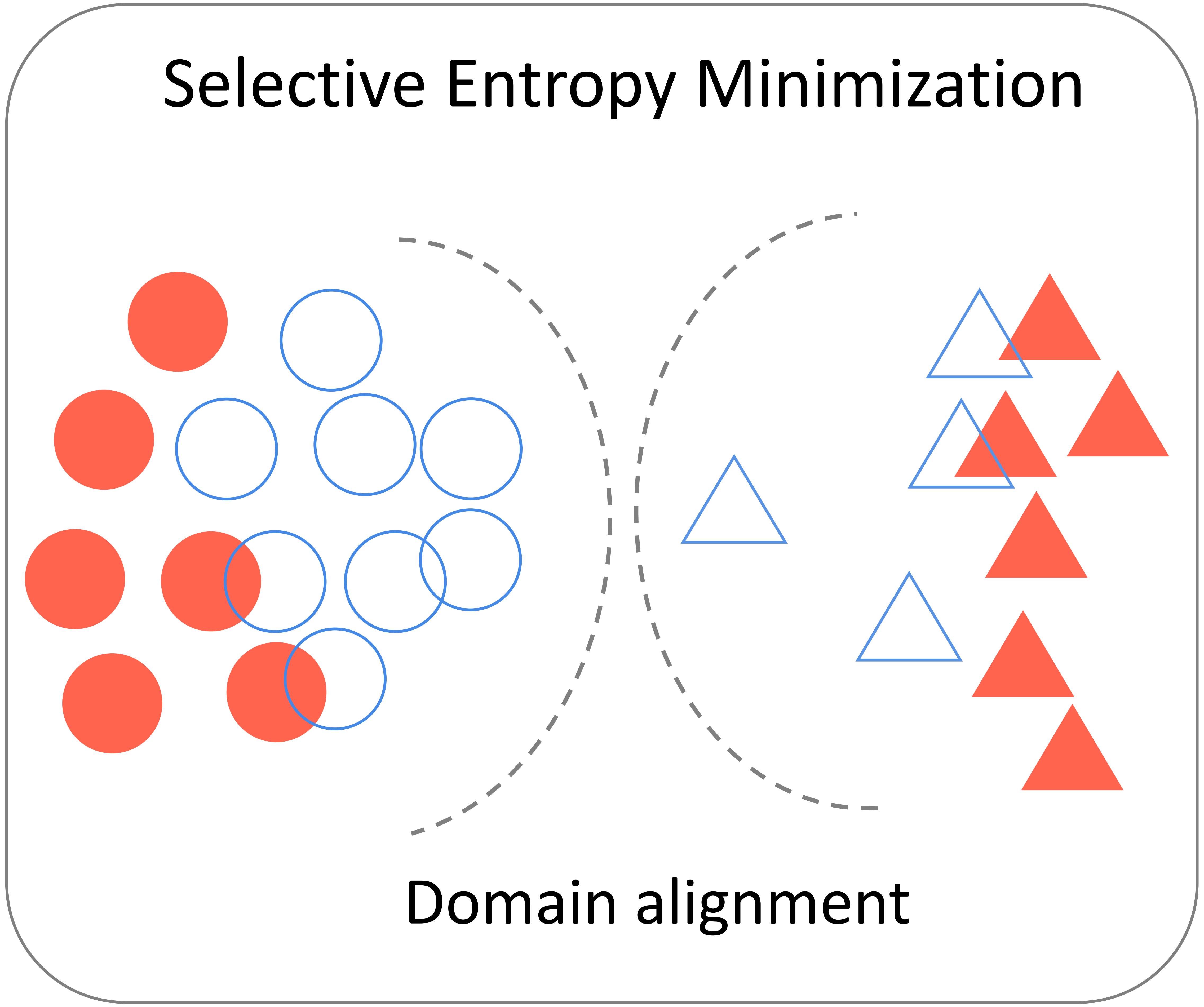
Prabhu, Khare, Karthik, Hoffman. ICCV 2021

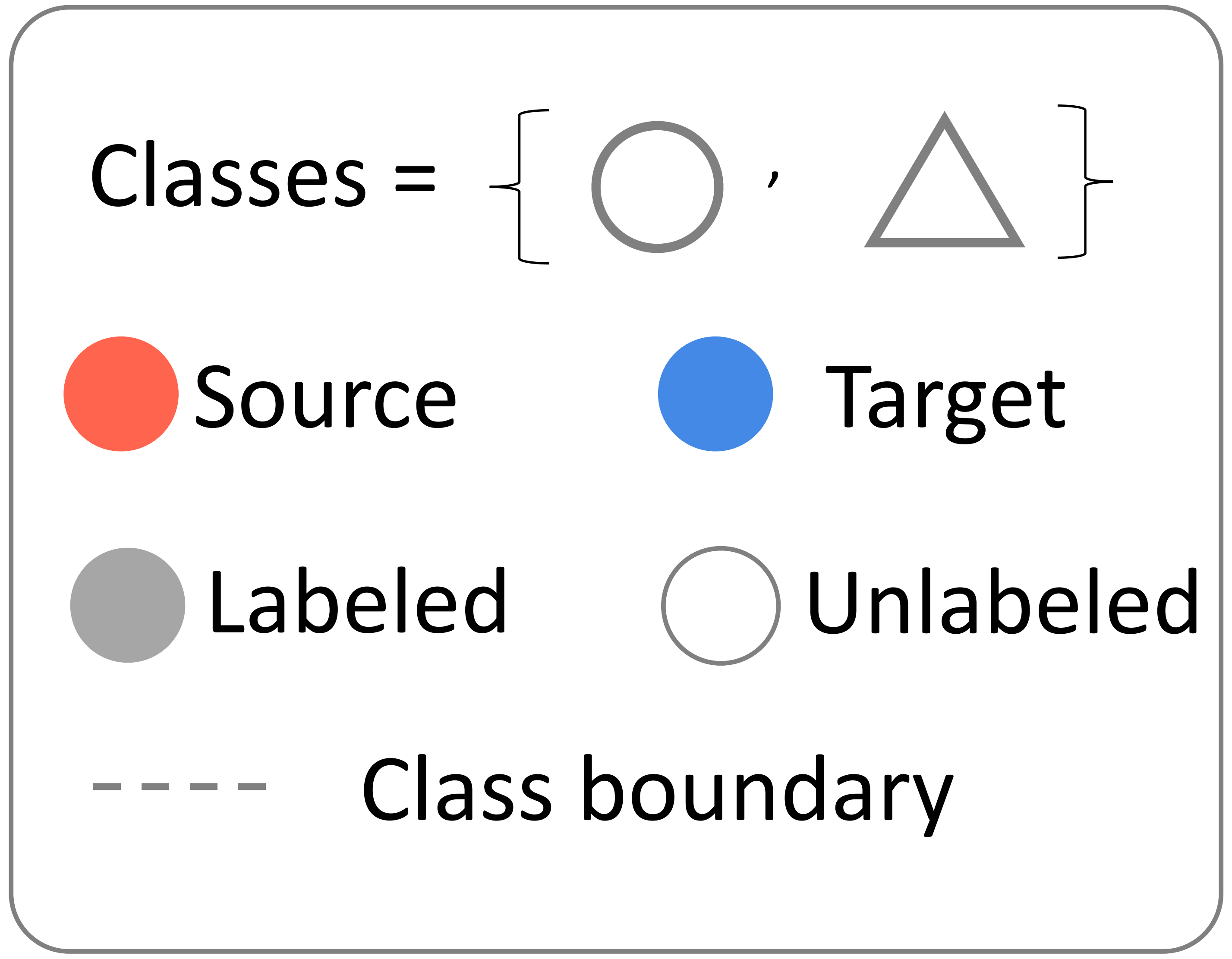
SENTRY: Selective Entropy Optimization

Key Idea

Identify reliable target instances via ~~model confidence~~ Predictive consistency[1,2,3]

Increase confidence on consistent instances



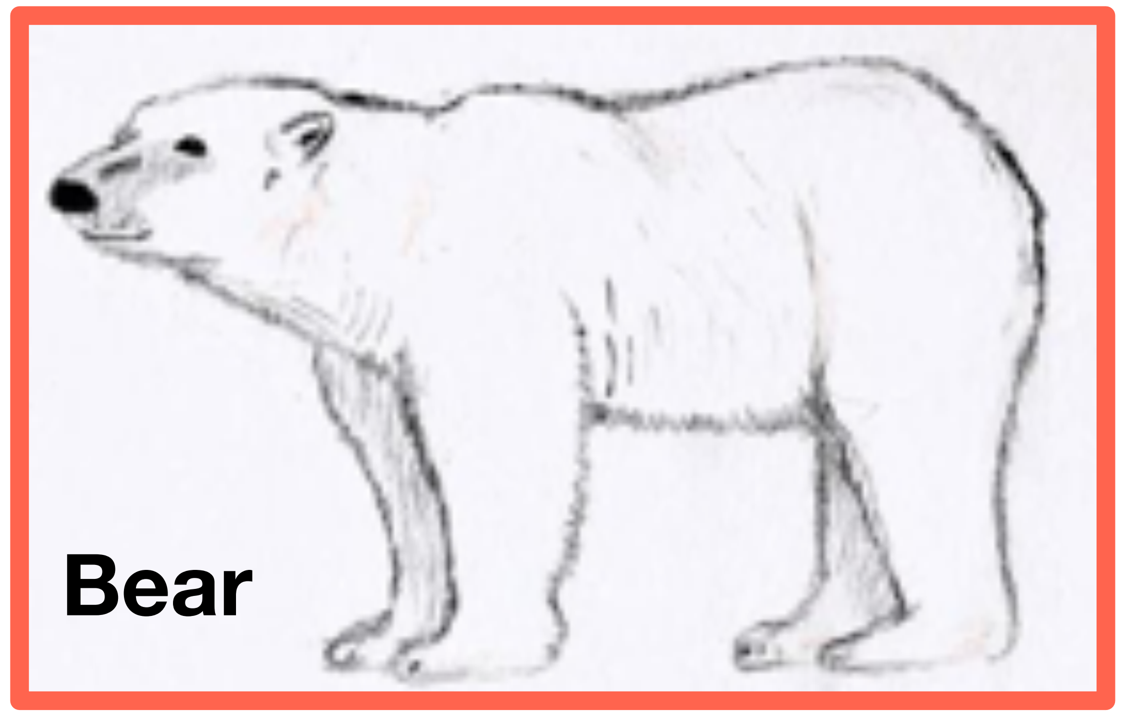


1. Bahat *et al.*, arXiv 2019.
2. Chen *et al.*, ICML 2020.
3. Sohn *et al.*, NeurIPS 2020.

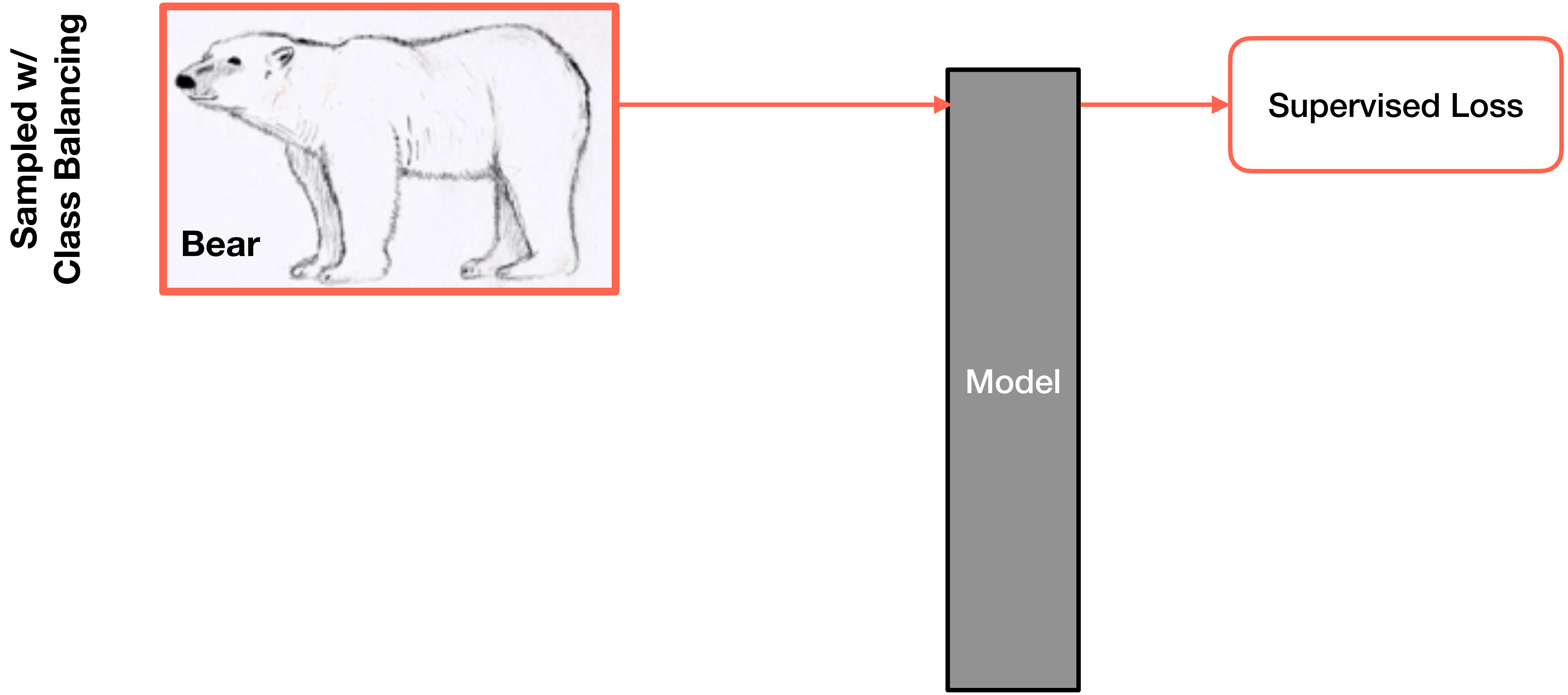
Prabhu, Khare, Karthik, Hoffman. ICCV 2021

SENTRY: Selective Entropy Optimization via Committee Consistency

Sampled w/
Class Balancing

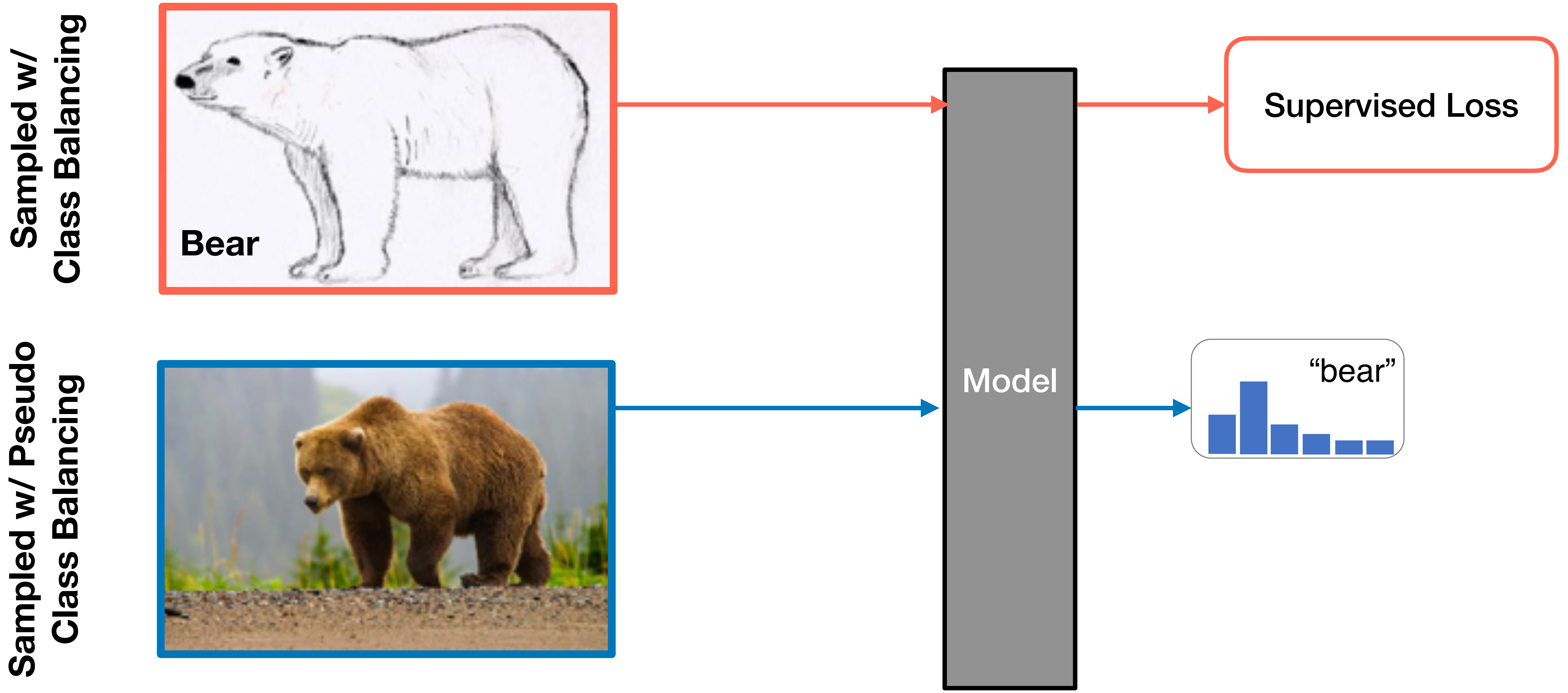


SENTRY: Selective Entropy Optimization via Committee Consistency



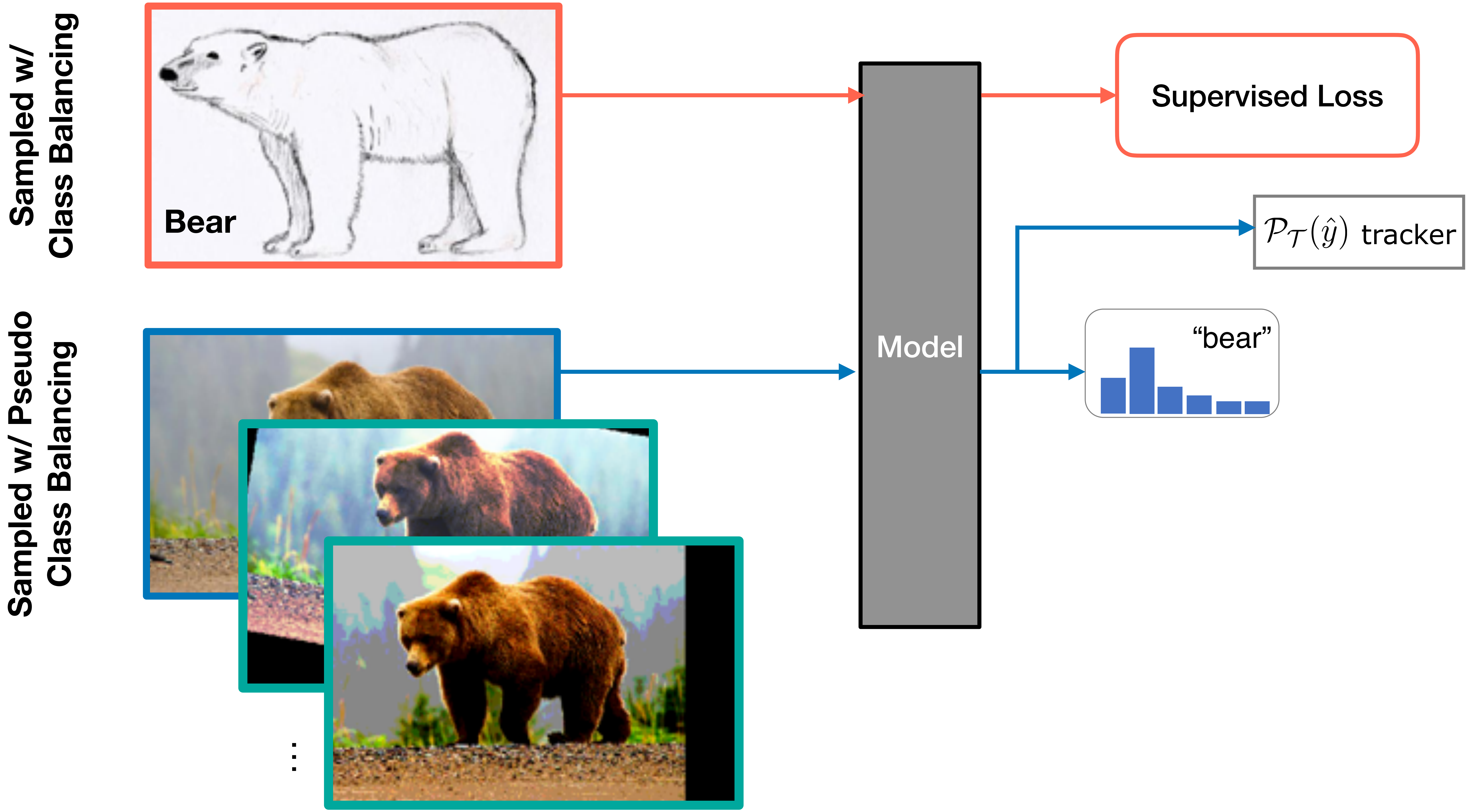
Prabhu, Khare, Karthik, Hoffman. ICCV 2021

SENTRY: Selective Entropy Optimization via Committee Consistency



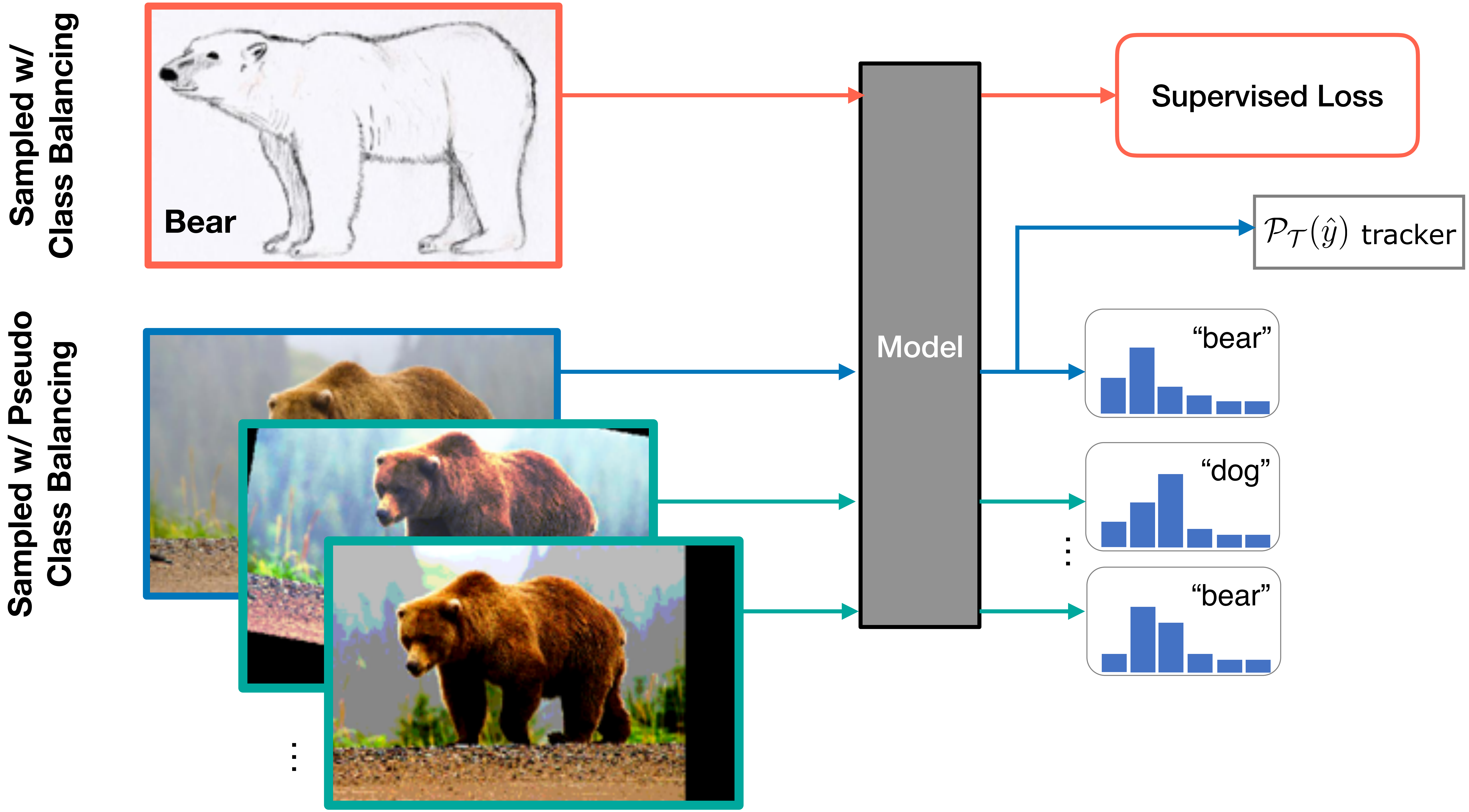
Prabhu, Khare, Karthik, Hoffman. ICCV 2021

SENTRY: Selective Entropy Optimization via Committee Consistency



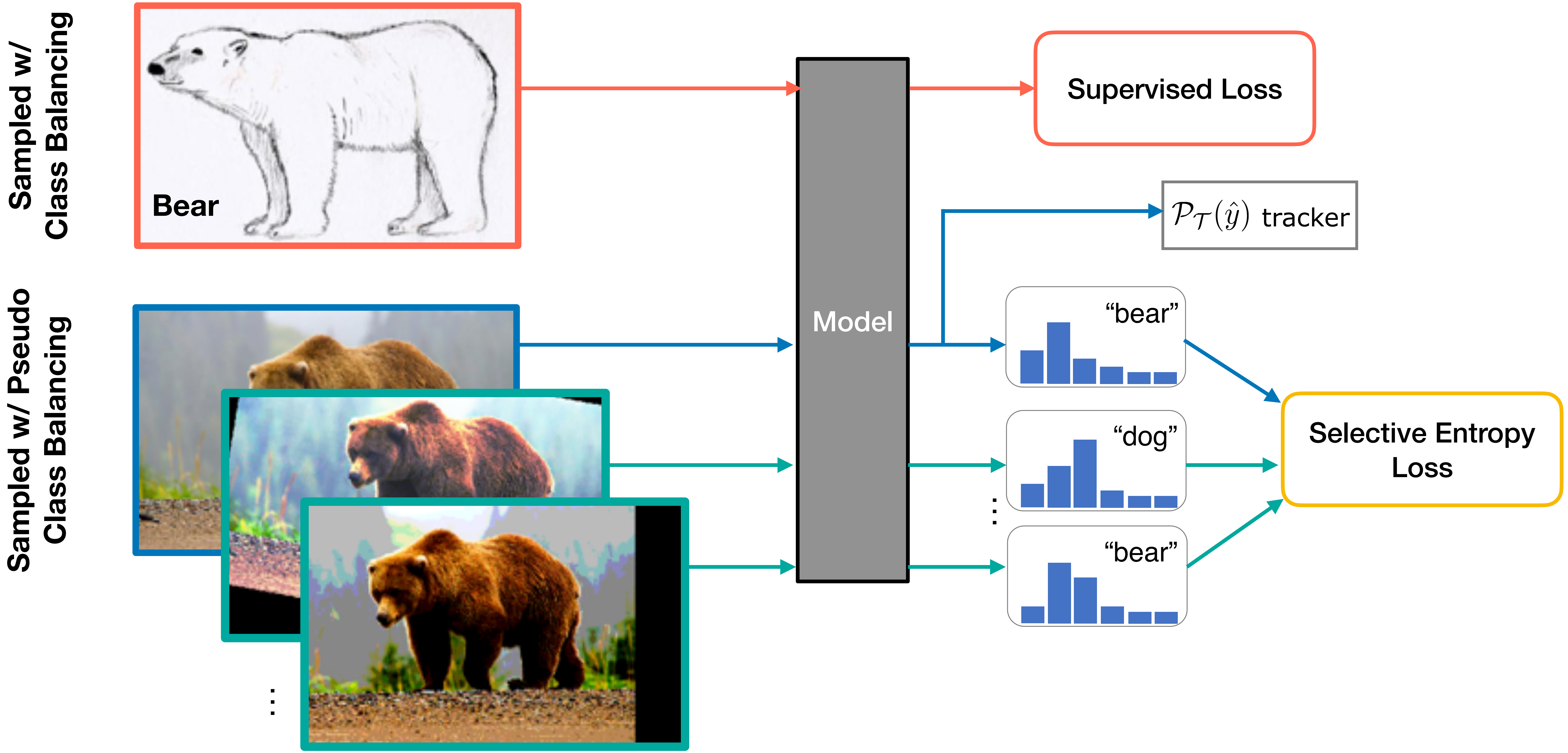
Prabhu, Khare, Karthik, Hoffman. ICCV 2021

SENTRY: Selective Entropy Optimization via Committee Consistency



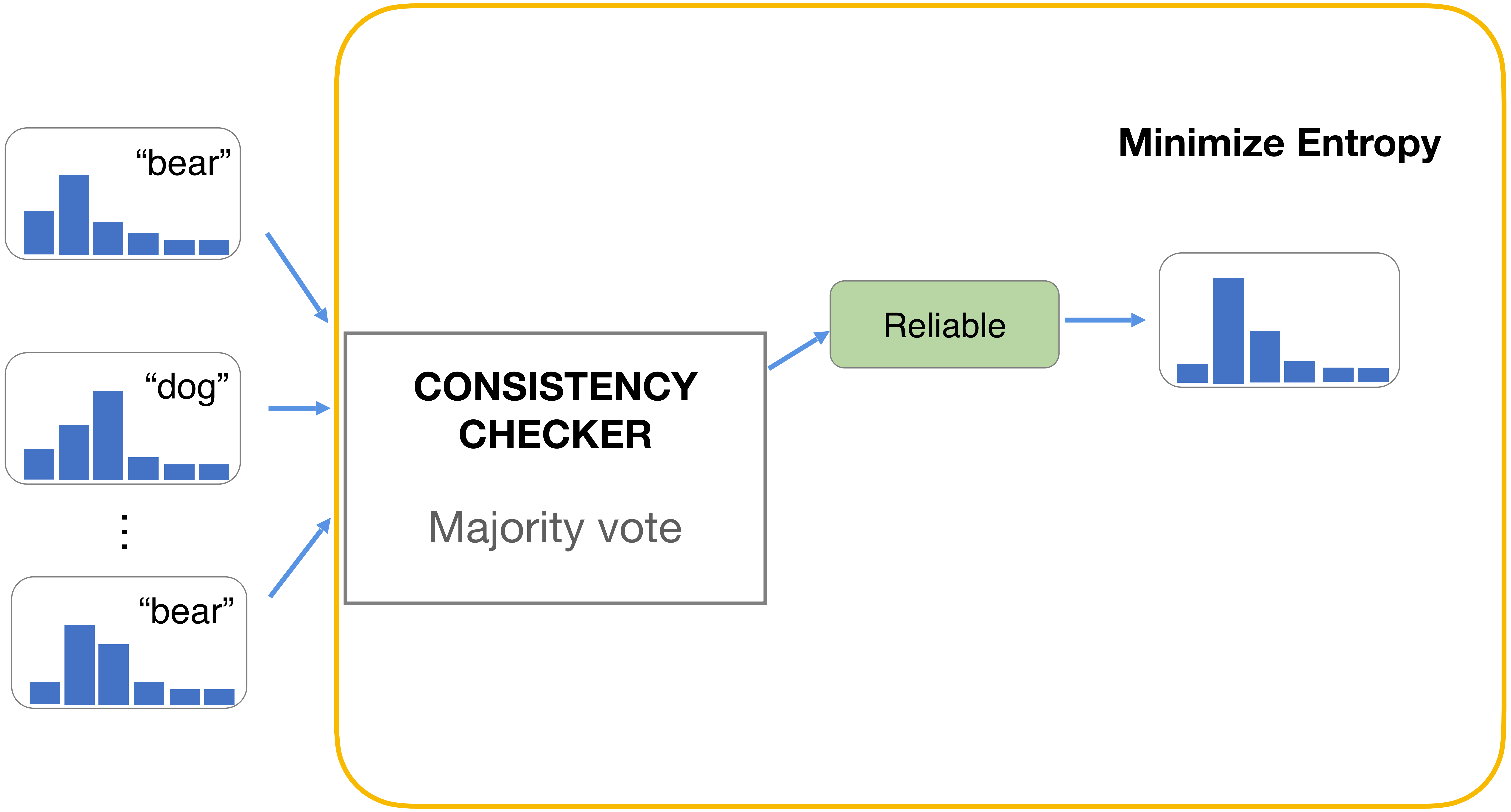
Prabhu, Khare, Karthik, Hoffman. ICCV 2021

SENTRY: Selective Entropy Optimization via Committee Consistency



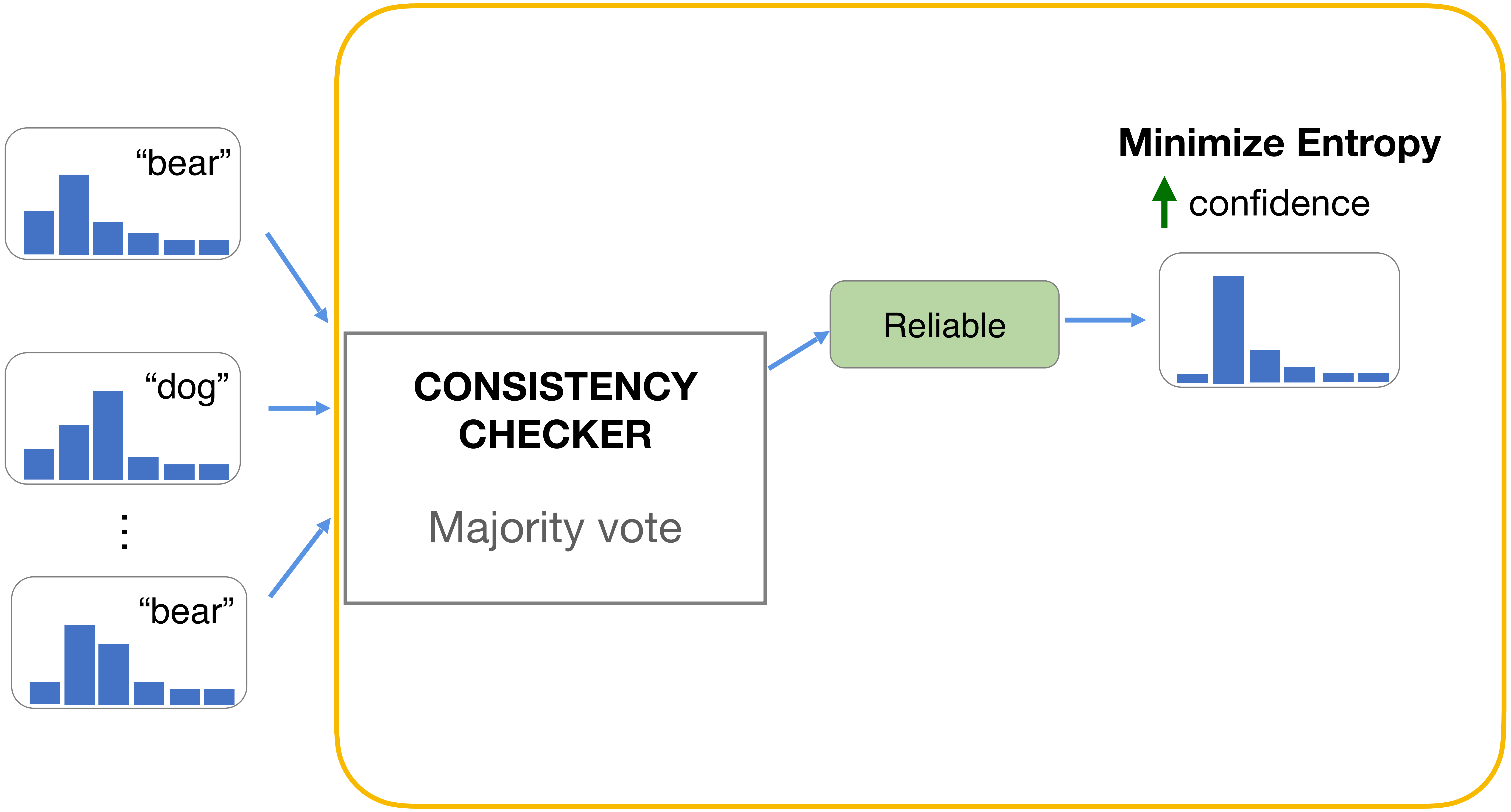
Prabhu, Khare, Karthik, Hoffman. ICCV 2021

Selective Entropy Loss



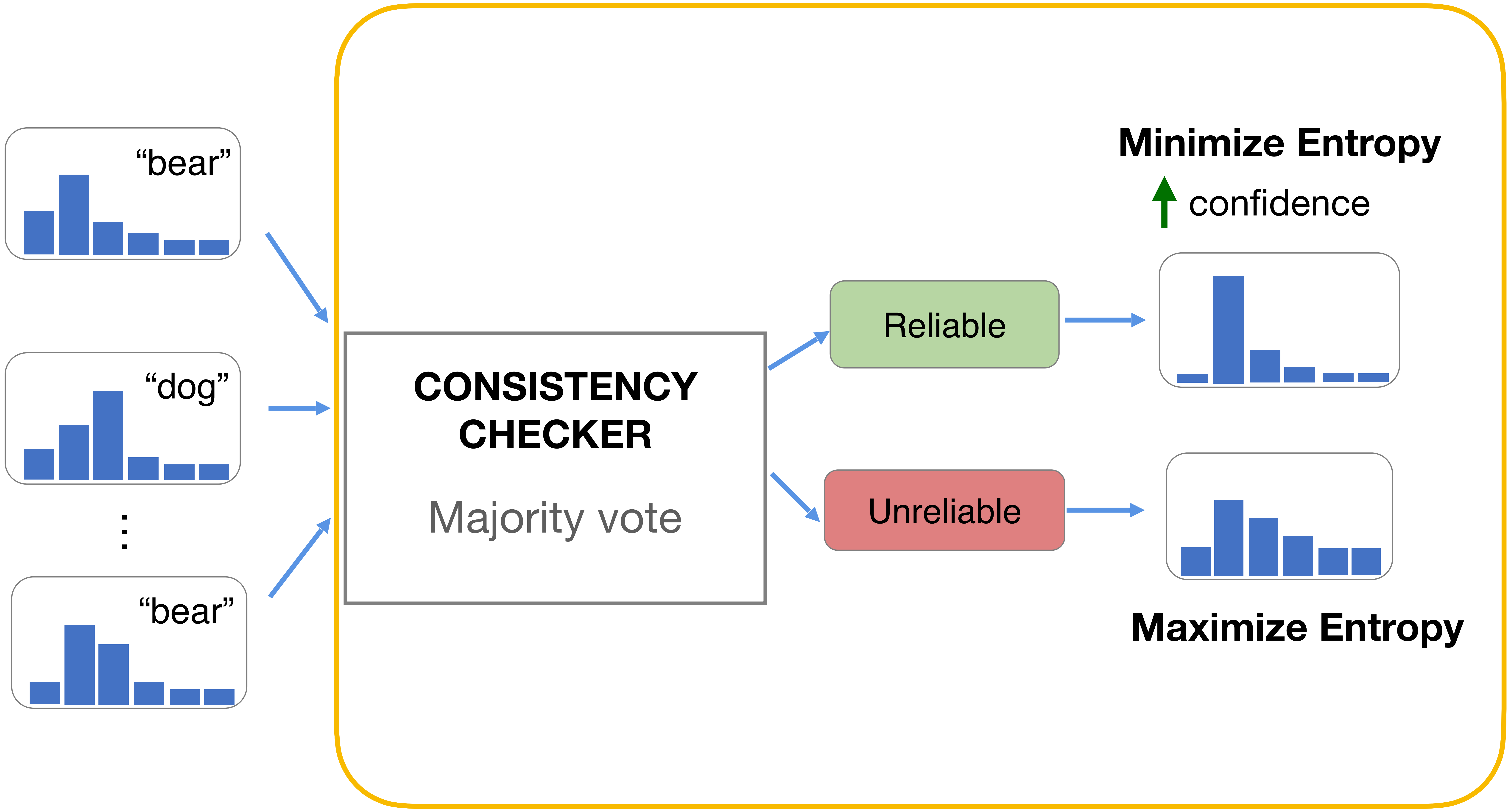
Prabhu, Khare, Karthik, Hoffman. ICCV 2021

Selective Entropy Loss



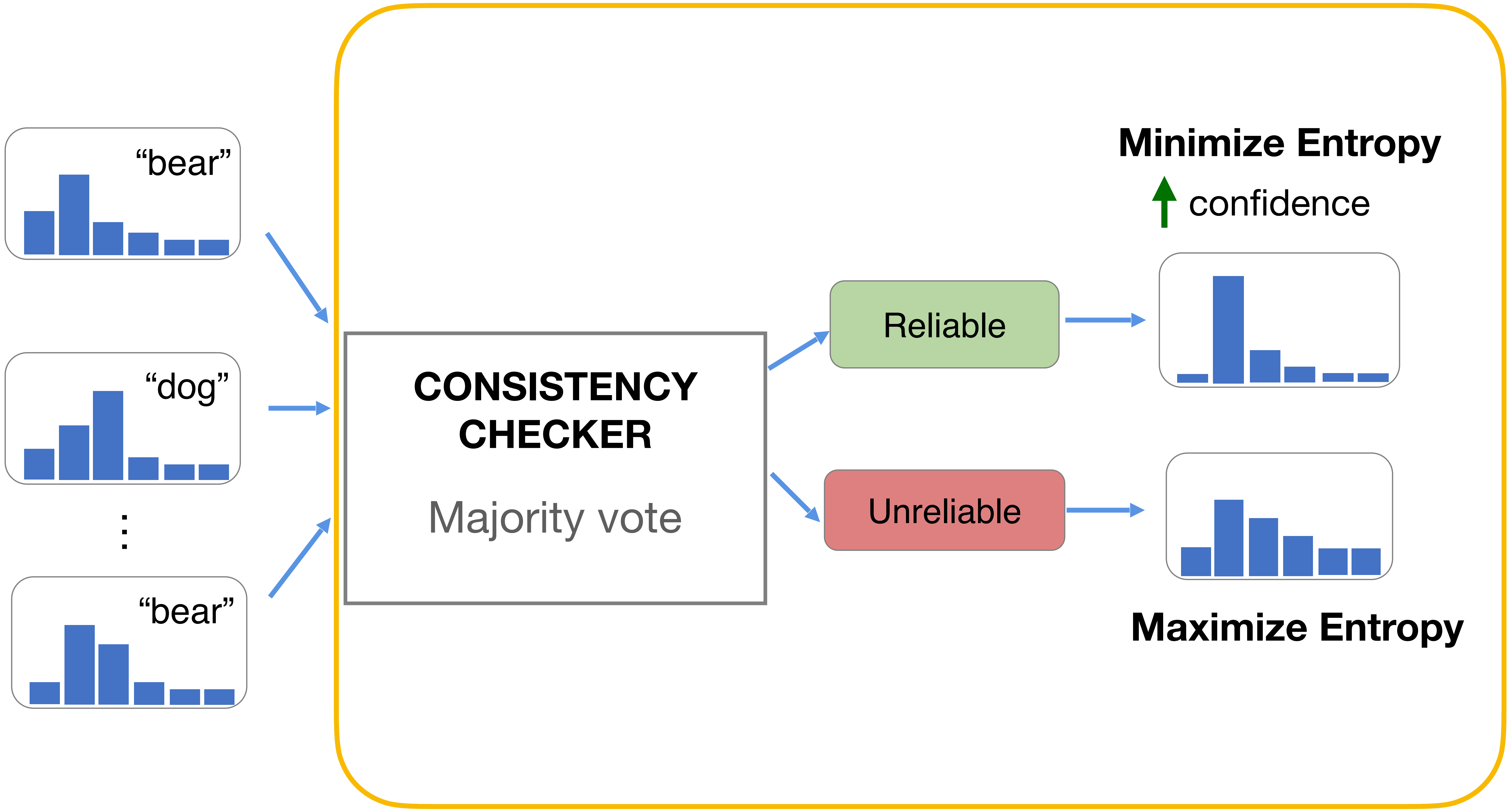
Prabhu, Khare, Karthik, Hoffman. ICCV 2021

Selective Entropy Loss



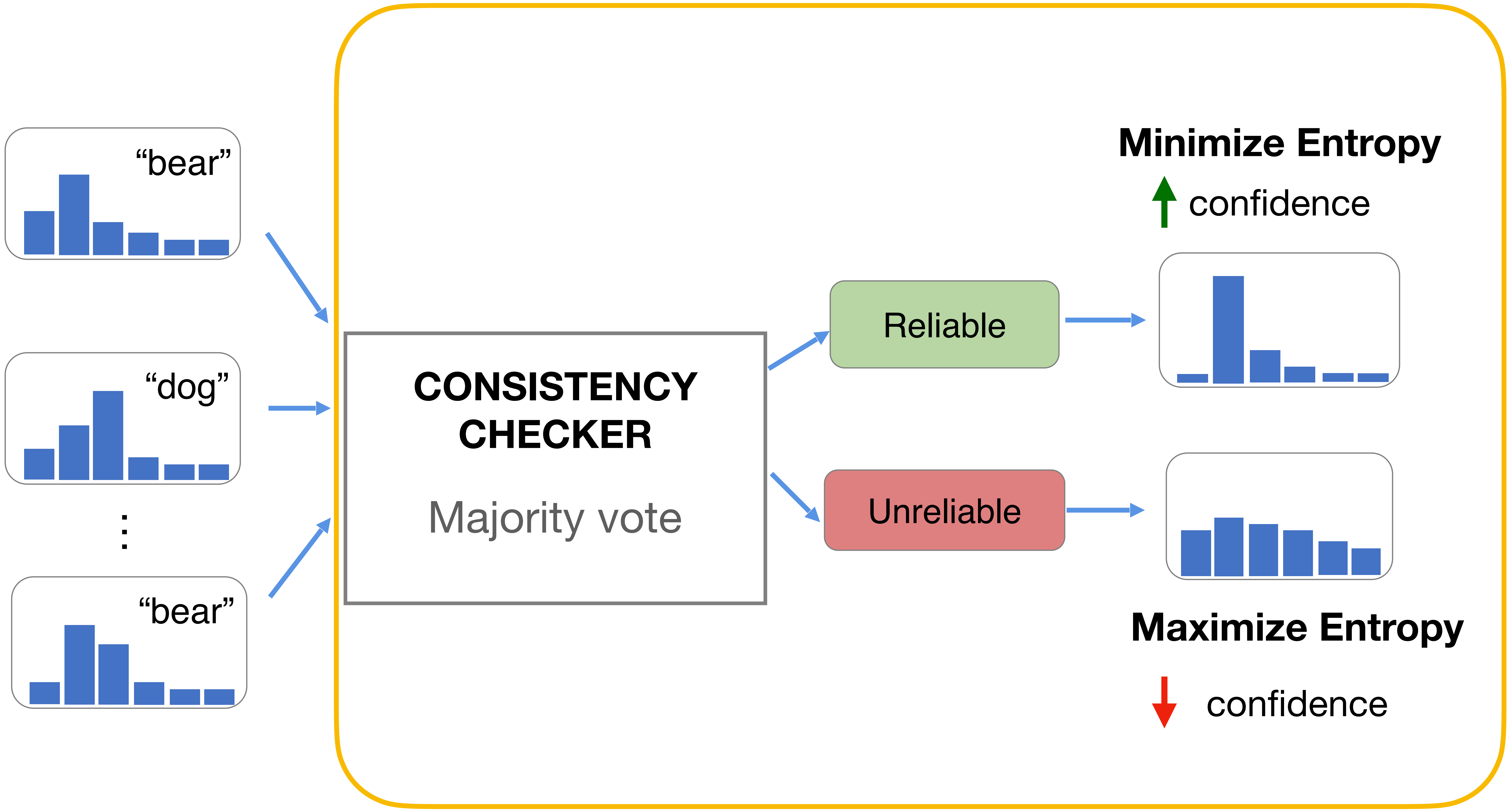
Prabhu, Khare, Karthik, Hoffman. ICCV 2021

Selective Entropy Loss



Prabhu, Khare, Karthik, Hoffman. ICCV 2021

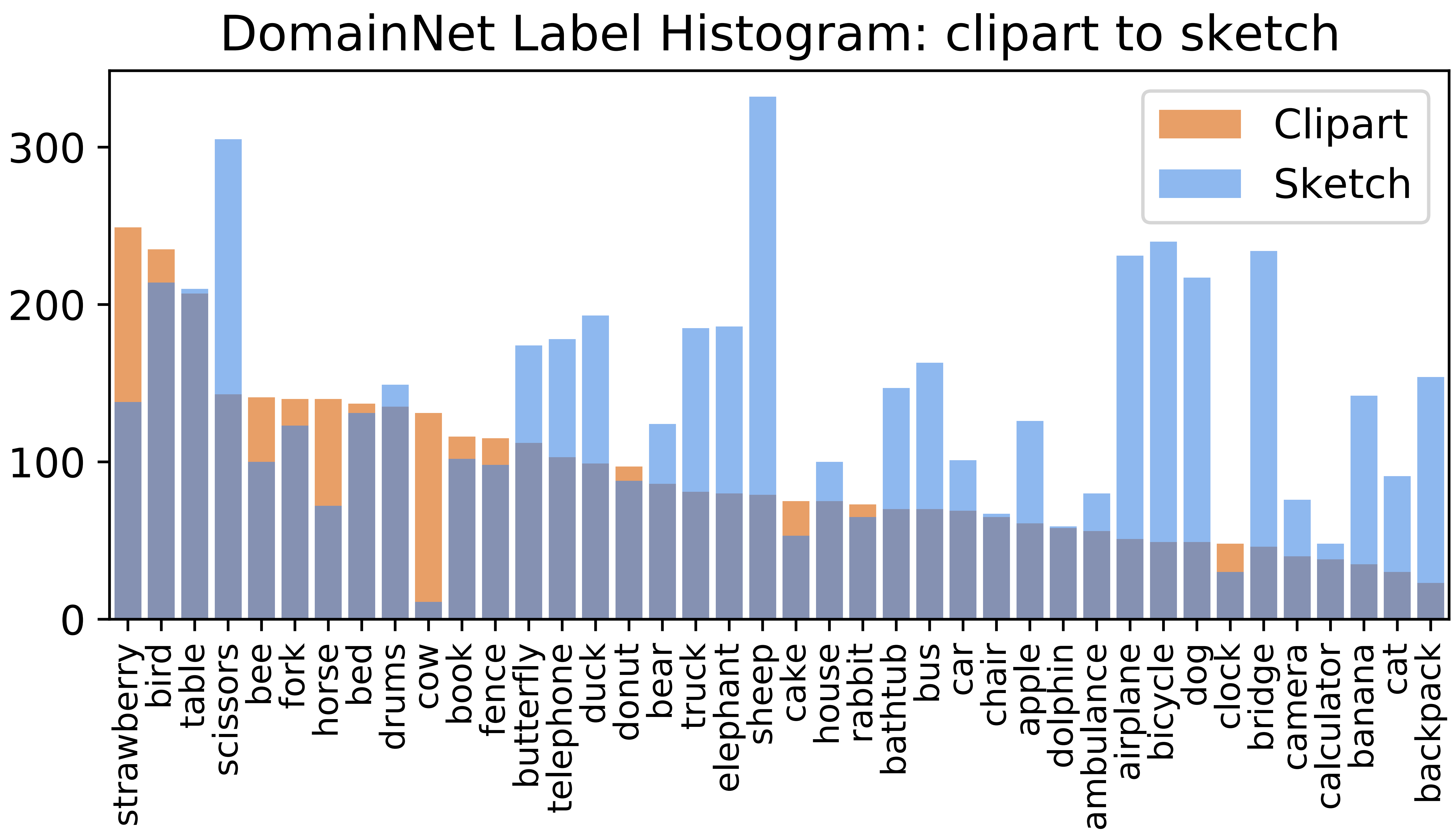
Selective Entropy Loss



Prabhu, Khare, Karthik, Hoffman. ICCV 2021

SENTRY Results: Image Classification

Natural label shifts



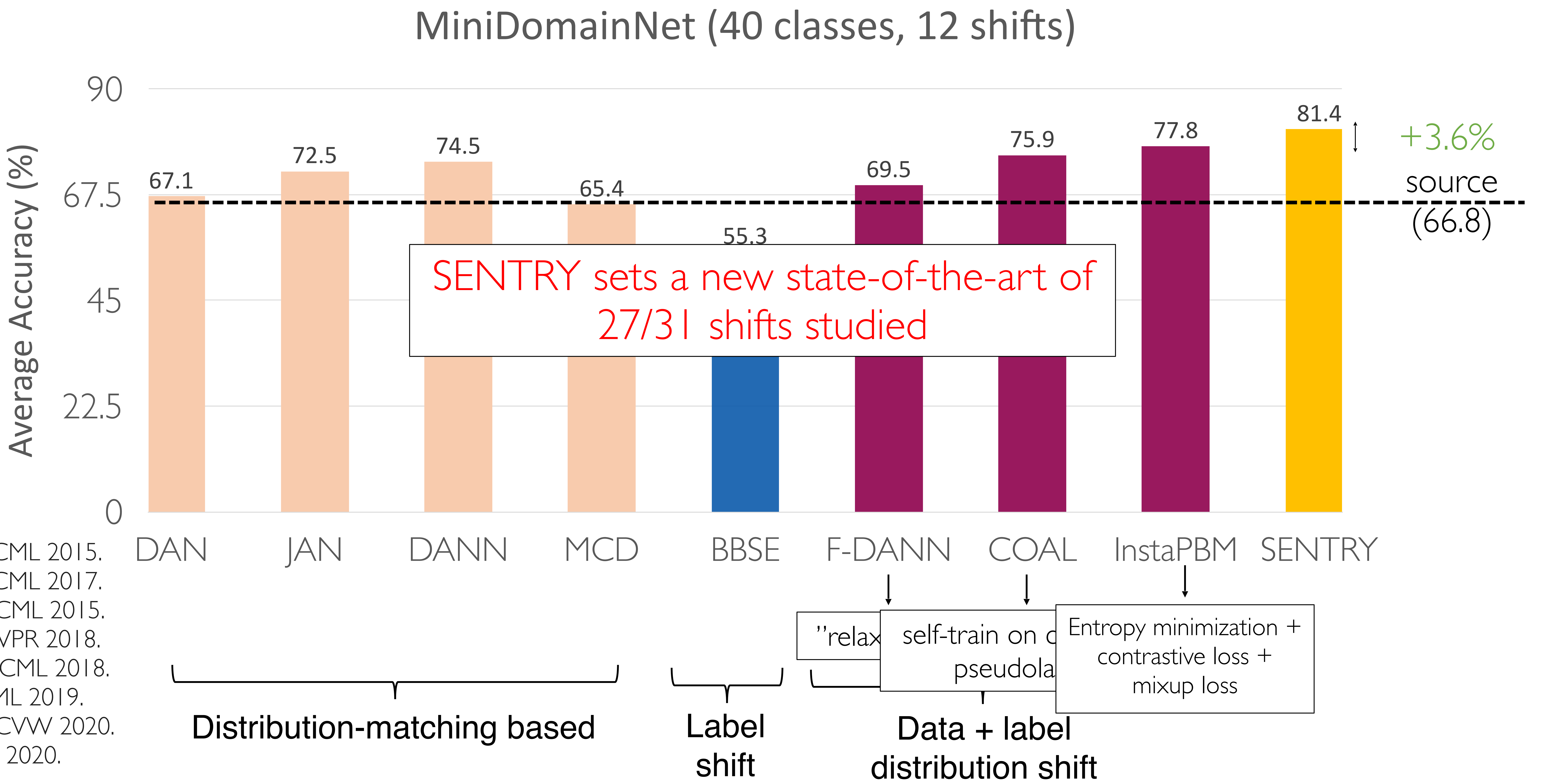
MiniDomainNet[1,2]

1. Peng et al., ICCV 2019.

2. Tan et al., ECCVW 2020.

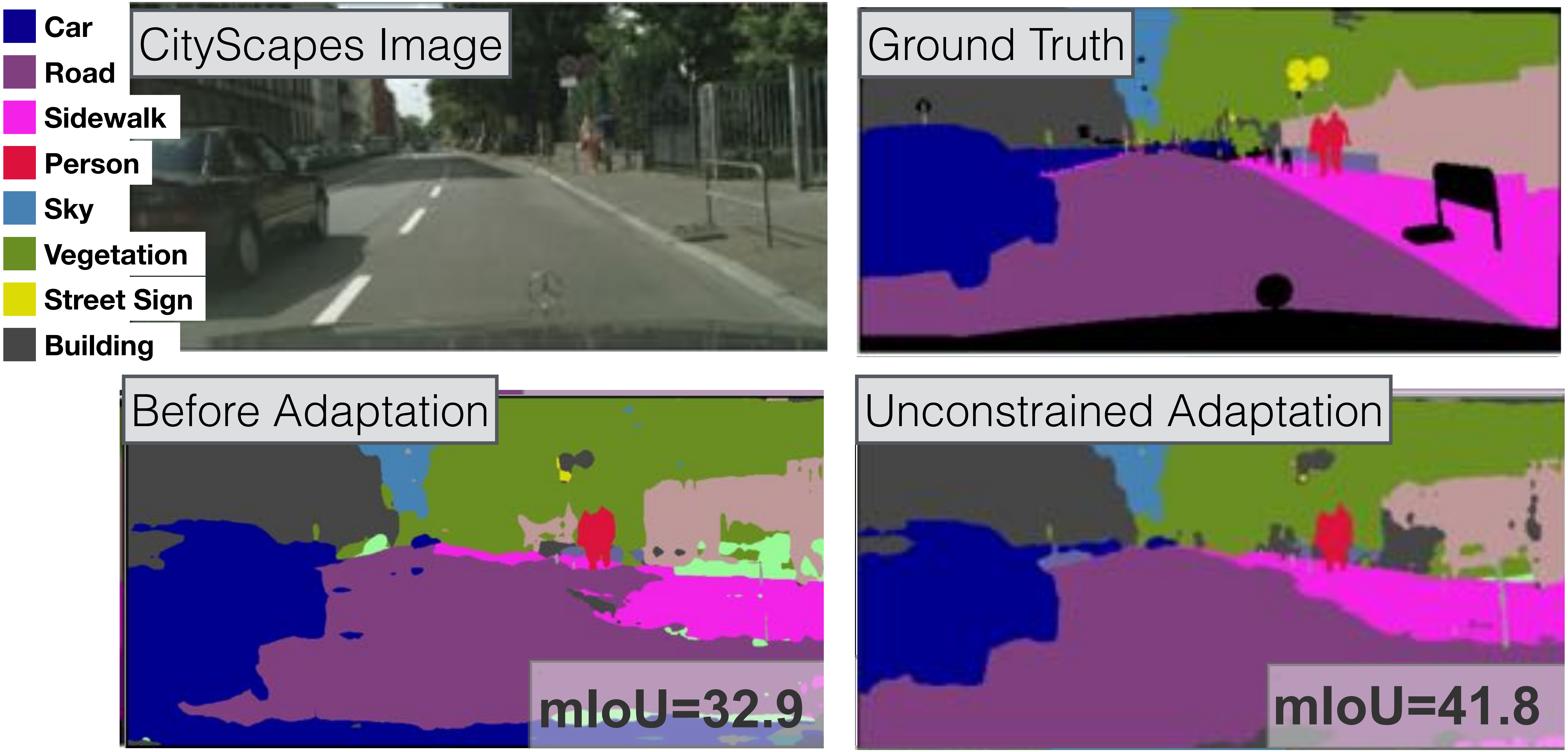
Prabhu, Khare, Karthik, Hoffman. ICCV 2021

SENTRY Results: MiniDomainNet



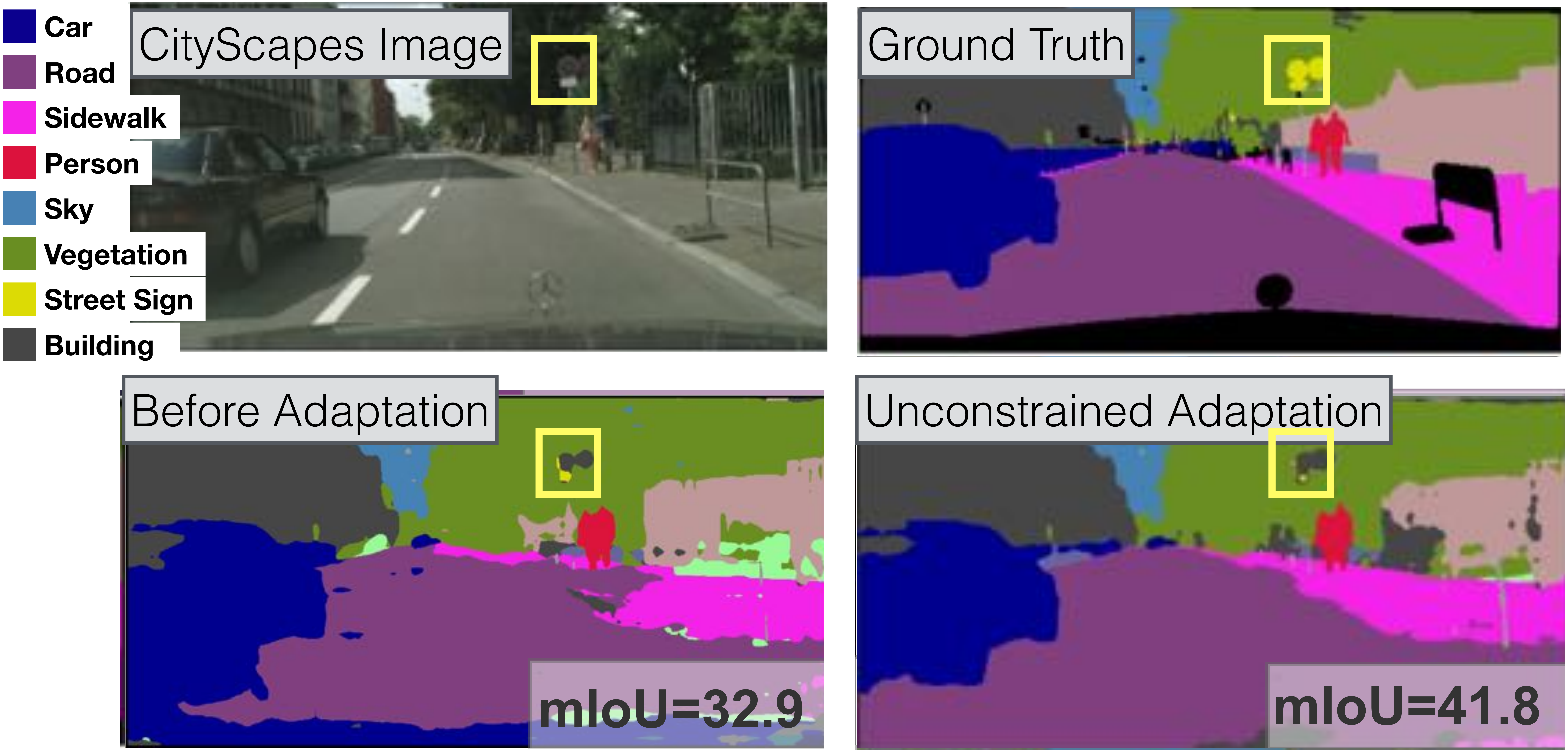
1. Long et al., ICML 2015.
2. Long et al., ICML 2017.
3. Ganin et al., ICML 2015.
4. Saito et al., CVPR 2018.
5. Lipton et al., ICML 2018.
6. Wu et al., ICML 2019.
7. Tan et al., ECCVW 2020.
8. Li et al., arXiv 2020.

Extension to Semantic Segmentation



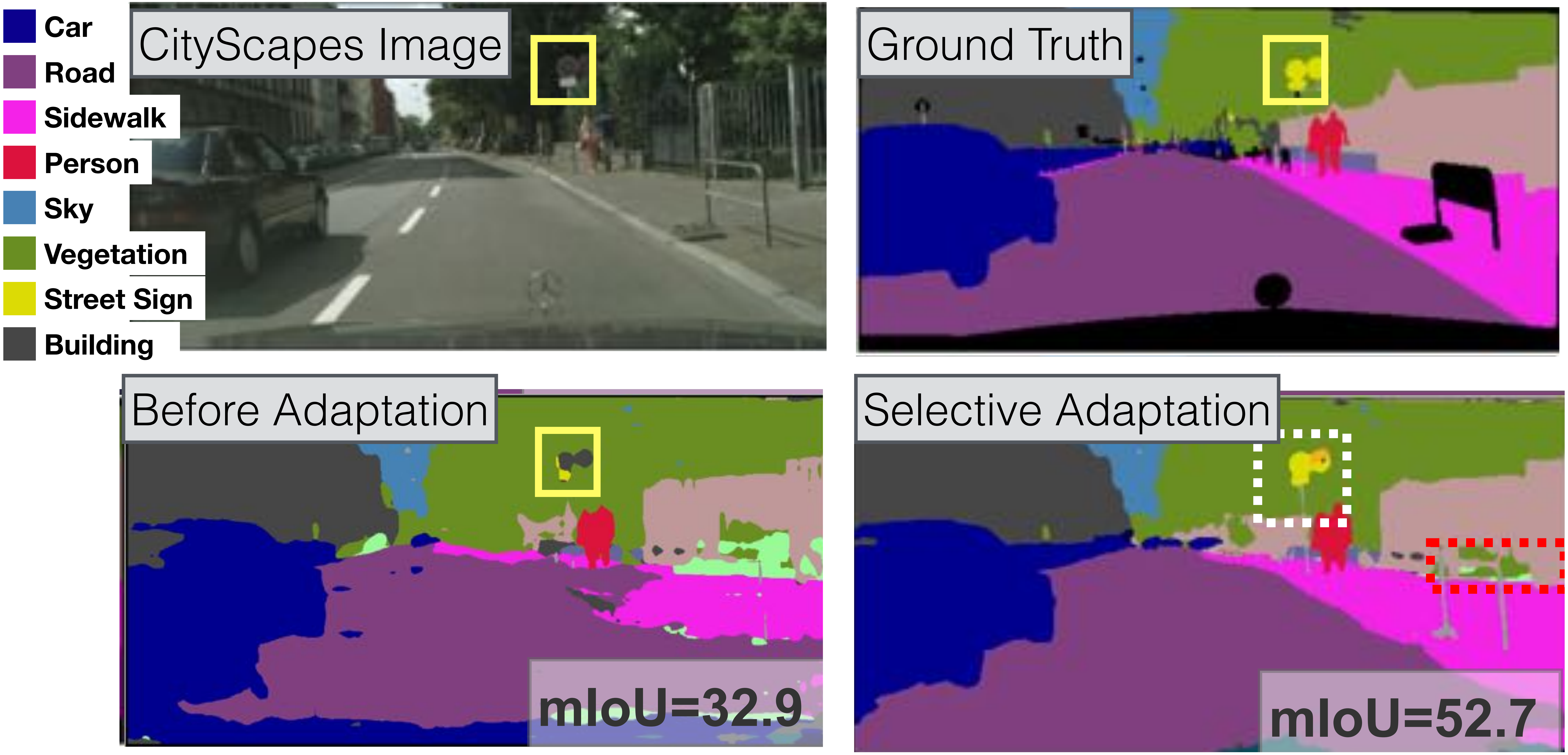
AUGCO: Augmentation Consistency-guided Self-training for Source-free Domain Adaptive Semantic Segmentation, Prabhu*, Khare*, Karthik, Hoffman. arXiv 2021

Extension to Semantic Segmentation



AUGCO: Augmentation Consistency-guided Self-training for Source-free Domain Adaptive Semantic Segmentation, Prabhu*, Khare*, Karthik, Hoffman. arXiv 2021

Extension to Semantic Segmentation



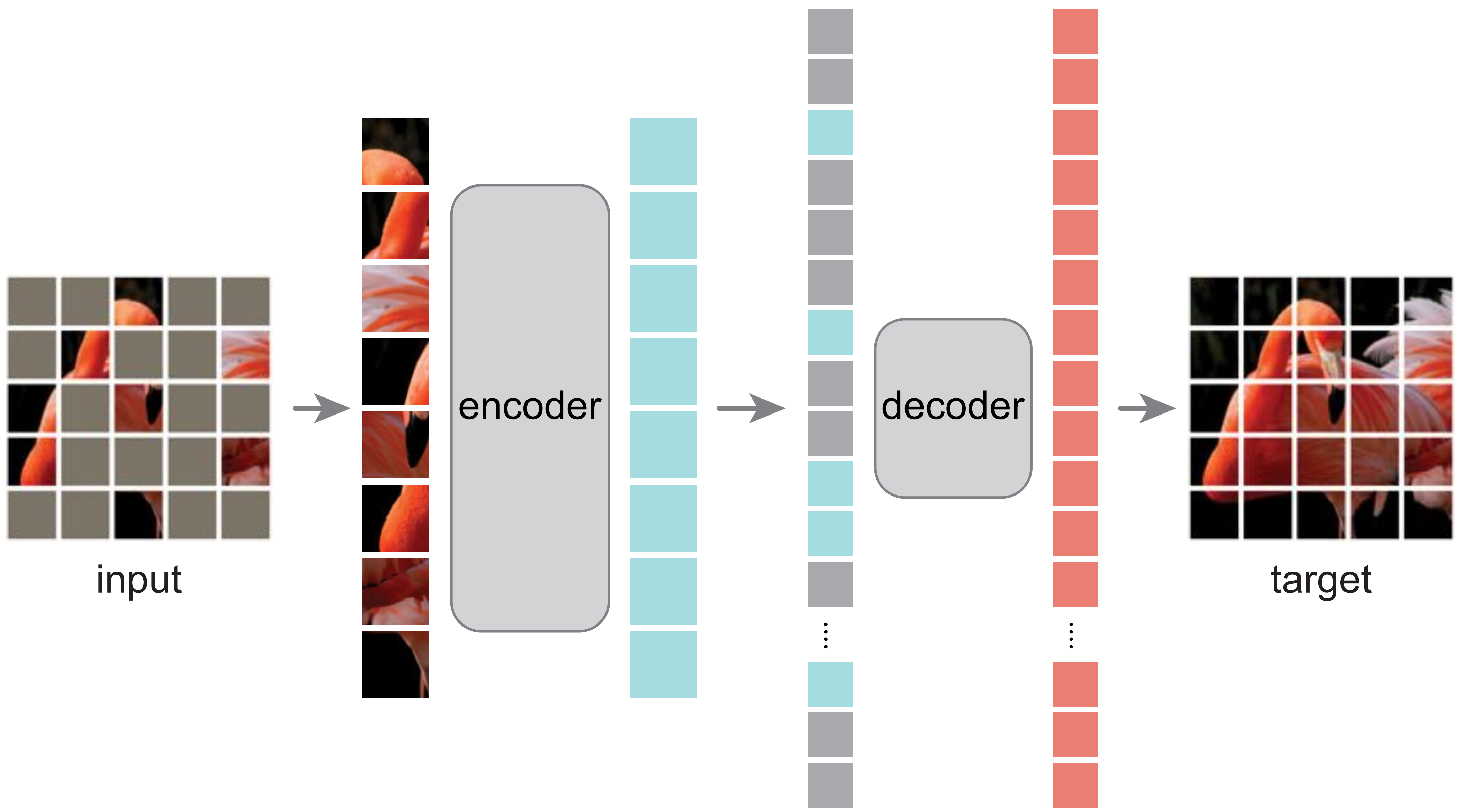
AUGCO: Augmentation Consistency-guided Self-training for Source-free Domain Adaptive Semantic Segmentation, Prabhu*, Khare*, Karthik, Hoffman. arXiv 2021

Consistency via attention-conditioned masking

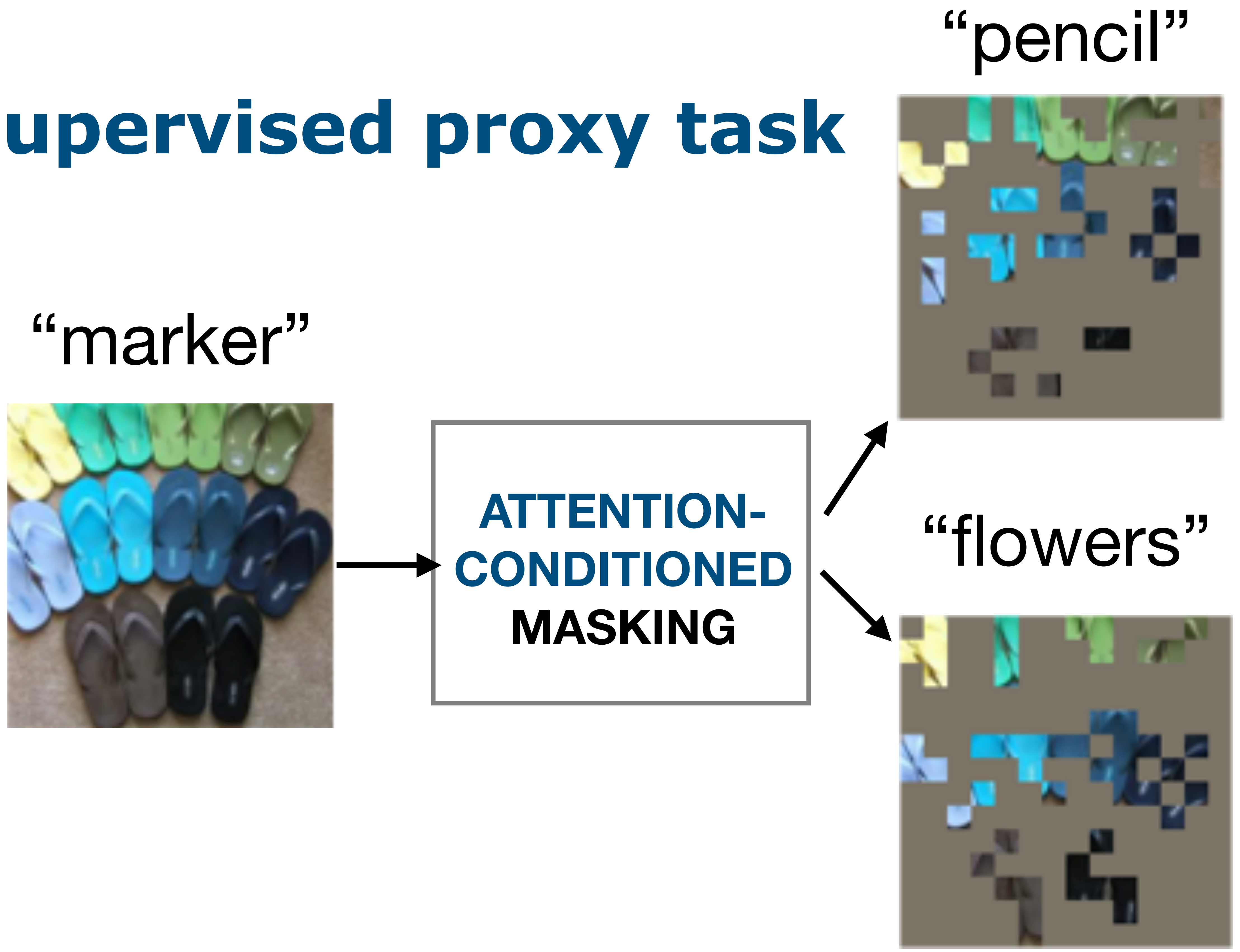
CNN ⇨ Vision Transformer (ViT)   Supervised ⇨ Self-Supervised Init

Key Idea
Measure predictive consistency under:

Random augmentations ⇨ **Self-supervised proxy task**



Masked Autoencoders, He *et al.*, CVPR 2022



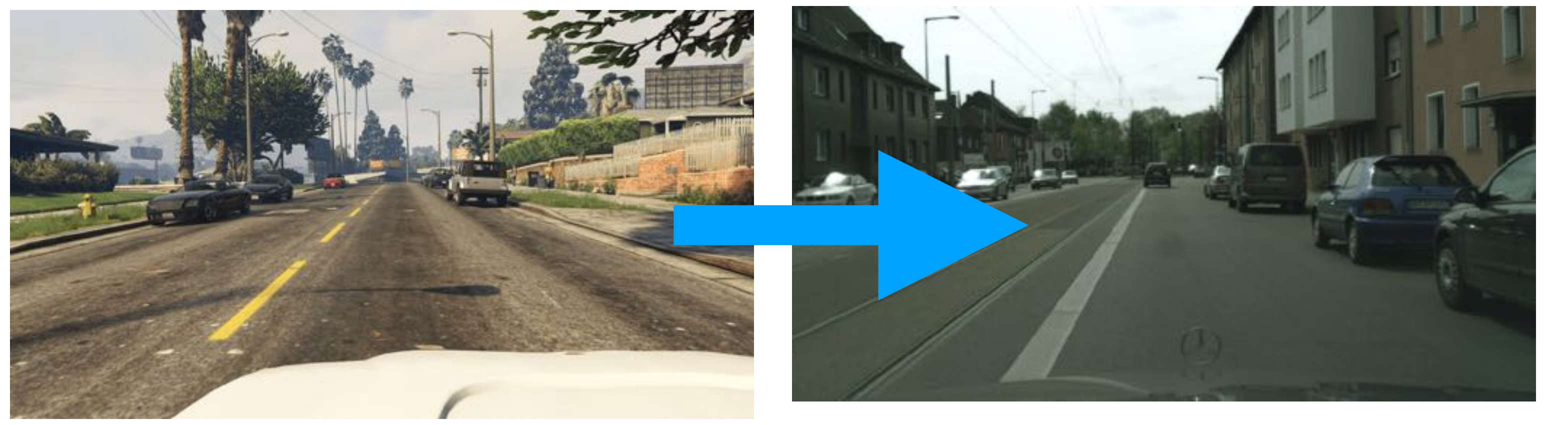
Adapting Self-Supervised Vision Transformers by Probing Attention-Conditioned Masking Consistency, Prabhu*, Yenamandra*, Singh, Hoffman. arXiv 2022

Performance Degradation from Bias

Curation

Weather

Sensor

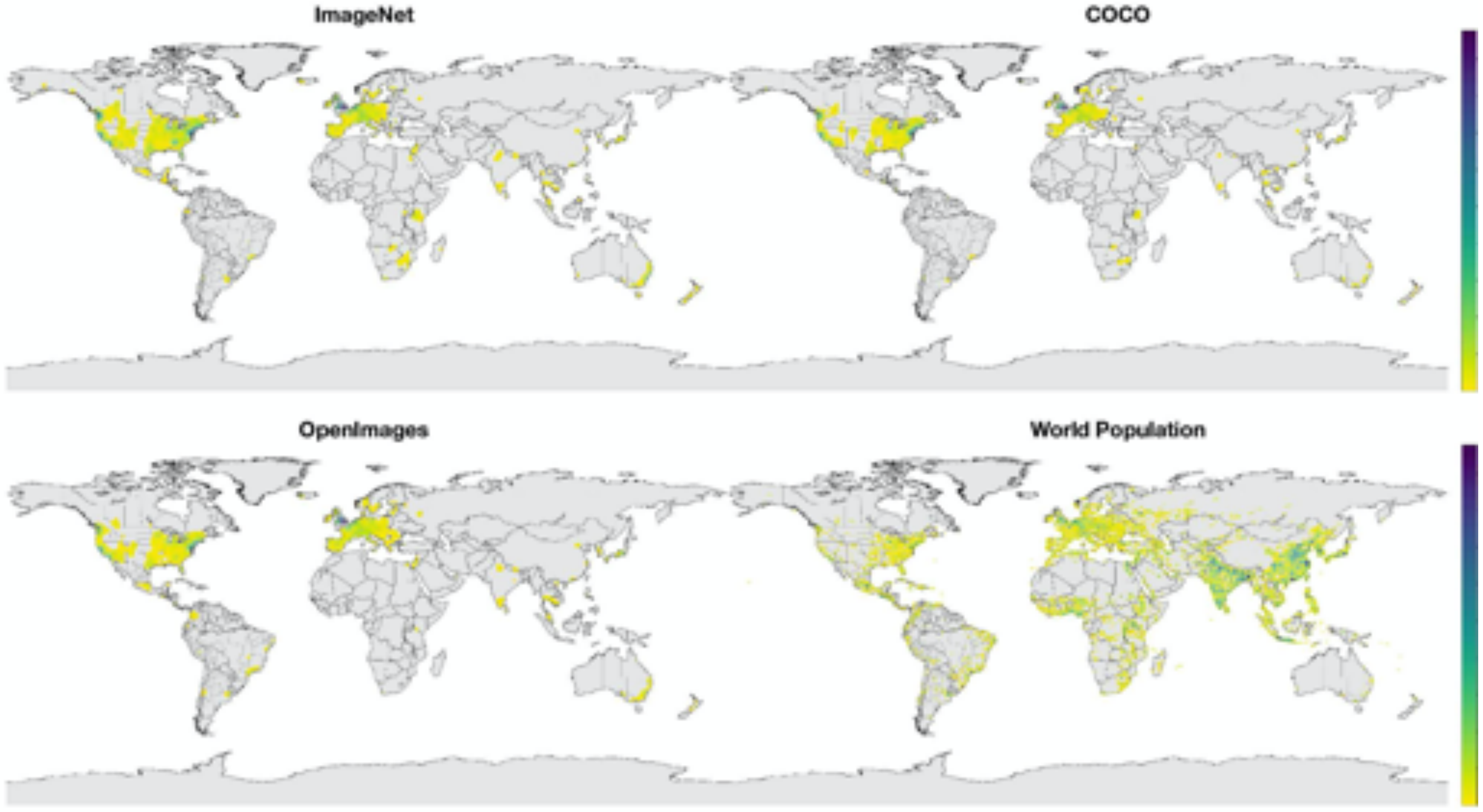


Demographic

Systems can underperform for certain subpopulations

Often caused by underrepresentation

Geographic Bias



DeVries et al. CVPRW 2019.

Does object recognition work for everyone?

Ground truth: Soap **Nepal, 288 $/month**

Azure: food, cheese, bread, cake, sandwich
Clarifai: food, wood, cooking, delicious, healthy
Google: food, dish, cuisine, comfort food, spam
Amazon: food, confectionary, sweets, burger
Watson: food, food product, turmeric, seasoning
Tencent: food, dish, matter, fast food, nutriment

Ground truth: Soap **UK, 1890 $/month**

Azure: toilet, design, art, sink
Clarifai: people, faucet, healthcare, lavatory, wash closet
Google: product, liquid, water, fluid, bathroom accessory
Amazon: sink, indoors, bottle, sink faucet
Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser
Tencent: lotion, toiletry, soap dispenser, dispenser, after shave

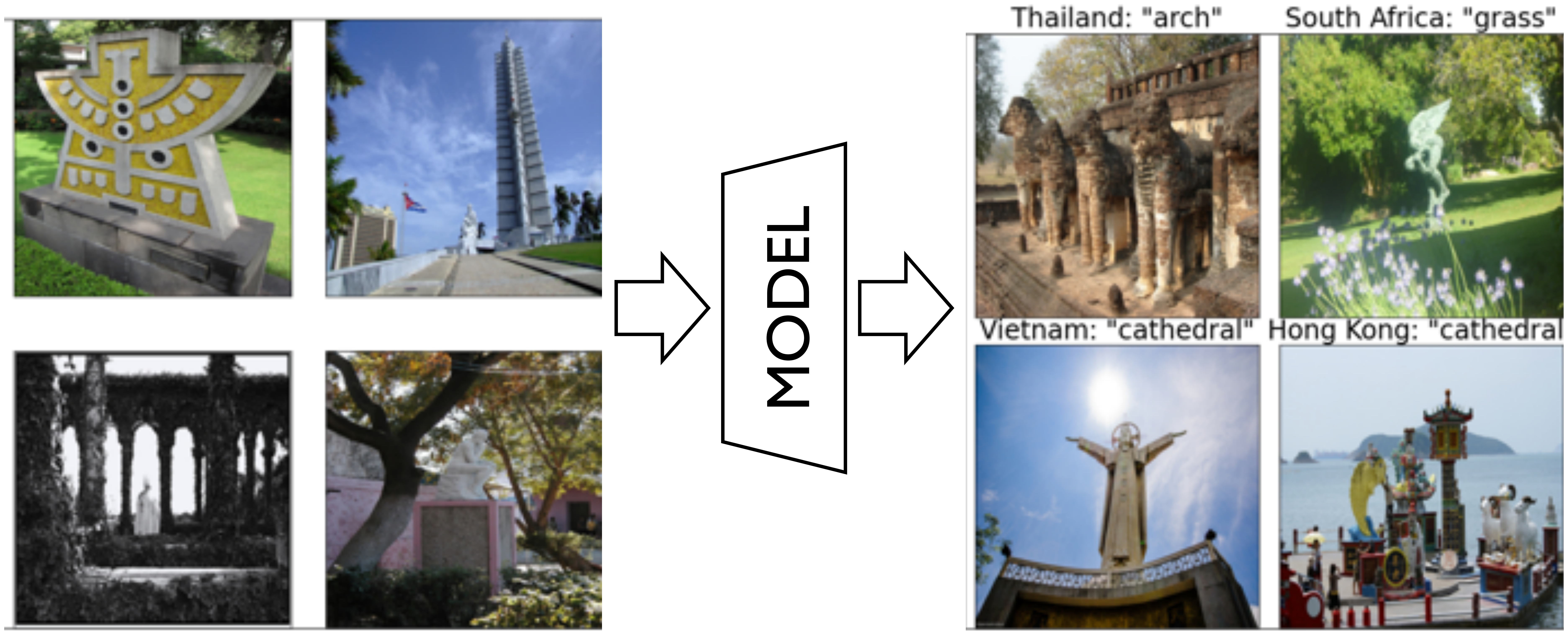
DeVries et al. CVPRW 2019.

Can domain adaptation make obj rec work for everyone?

Train (North America)

label = “statue”

Test (Rest of the world)



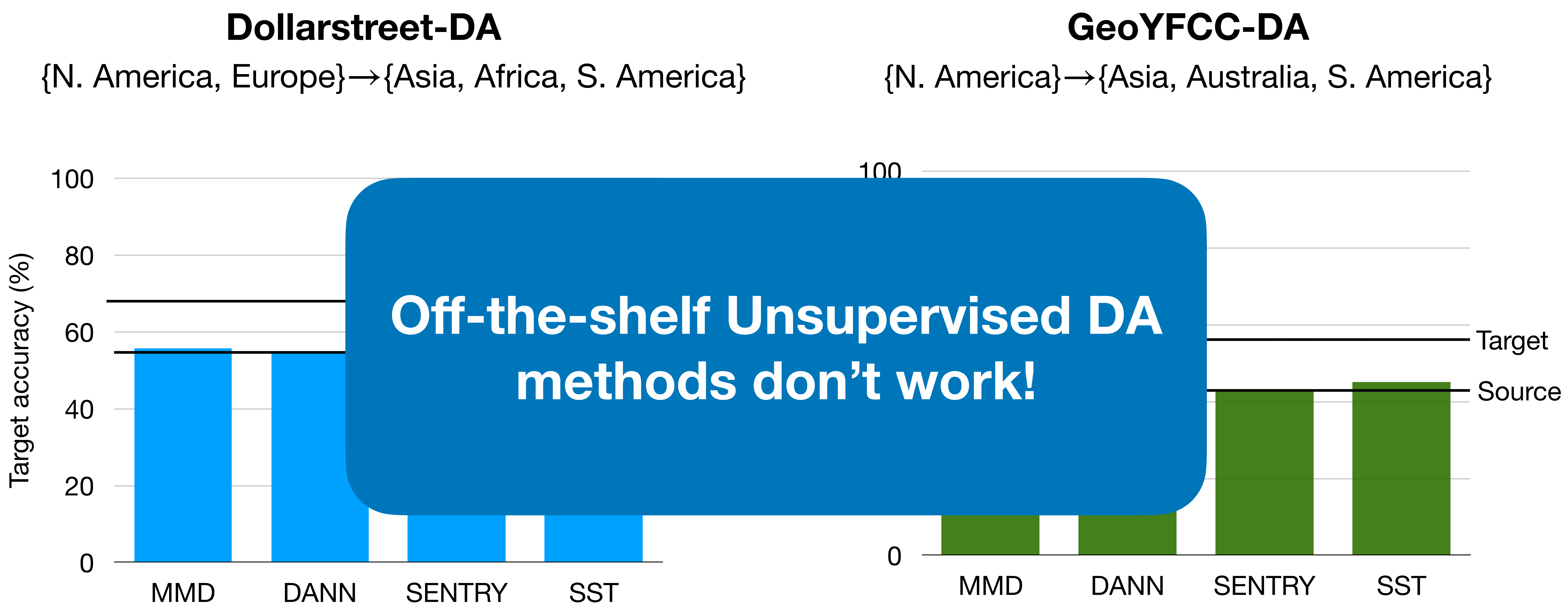
Prabhu, Selvaraju, Hoffman, Naik. CVPR L3D Workshop, 2022

Geographically diverse data

Prabhu, Selvaraju, Hoffman, Naik. CVPR L3D Workshop, 2022

1. https://www.gapminder.org/dollar-street 2. Dubey *et al.*, CVPR 2021.

Results



1. Long *et al.*, ICML 2015
2. Ganin *et al.*, ICML 2015
3. Prabhu *et al.*, ICCV 2021

Prabhu, Selvaraju, Hoffman, Naik. CVPR L3D Workshop, 2022

Additional challenges in GeoDA

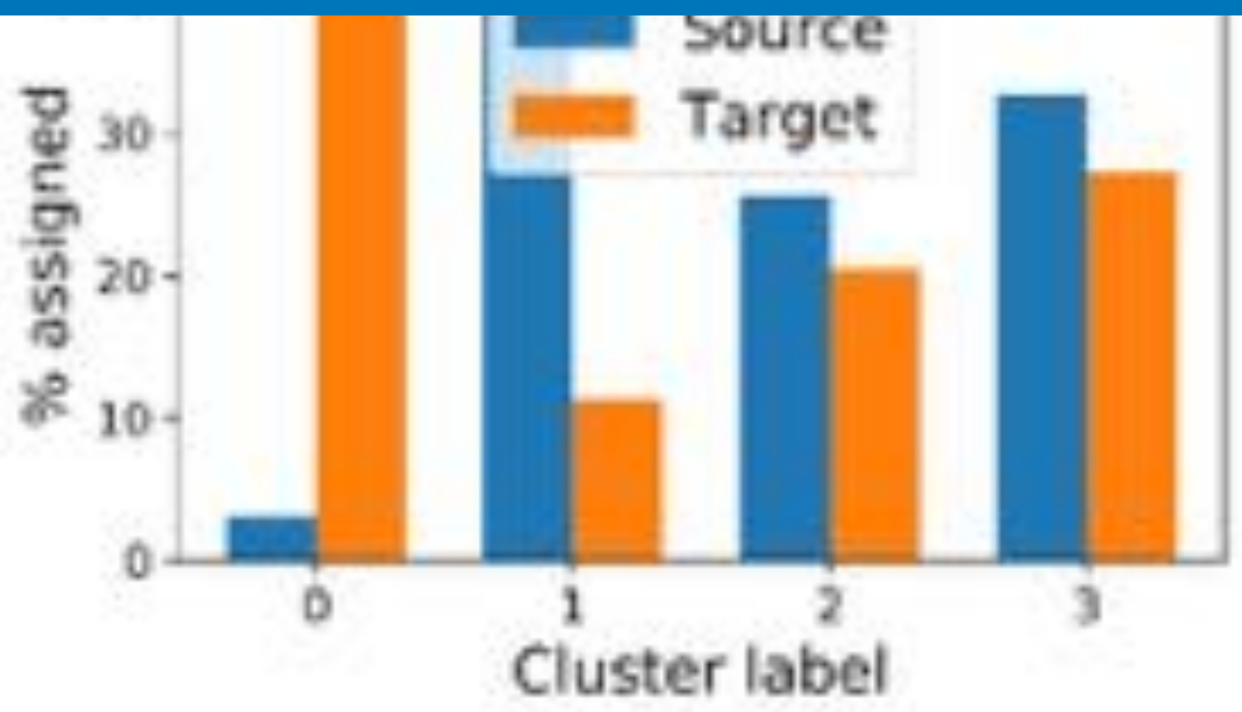
Context Shift

$P_S(c(\mathbf{x})\,|\,y) \neq P_T($

Specialized solutions are needed for Geo DA!

Subpopulation Shift

$P_S(\mathbf{x}\,|\,y) \neq P_T(\mathbf{x}\,|\,y)$



Prabhu, Selvaraju, Hoffman, Naik. CVPR L3D Workshop, 2022

Summary: Responsible Vision

Reliability Goal: Perform vision tasks as expected at deployment time.

Benchmarks

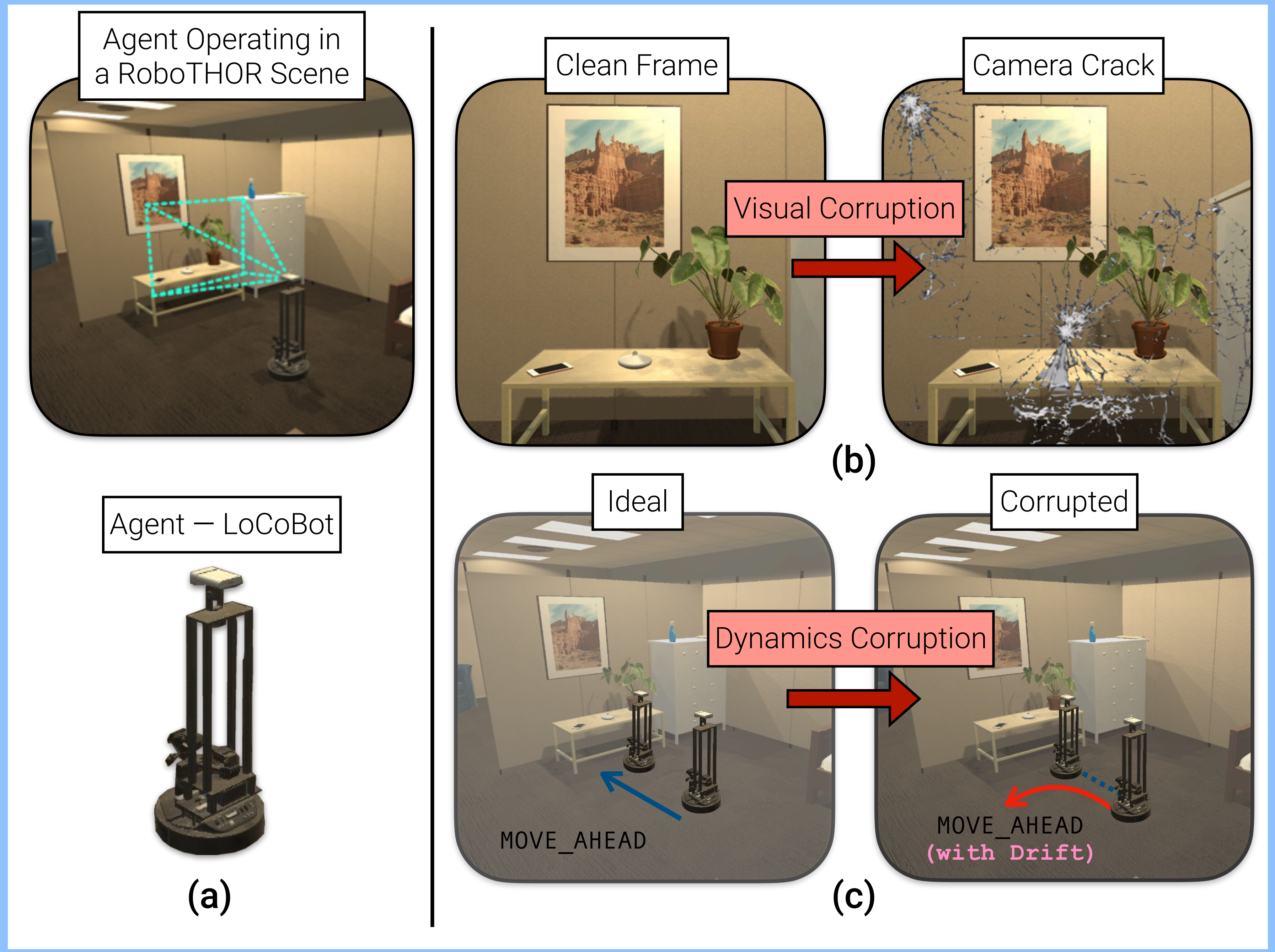
Need benchmarks to define expectations

Resilience

Withstand or adapt to a diverse set of visual conditions

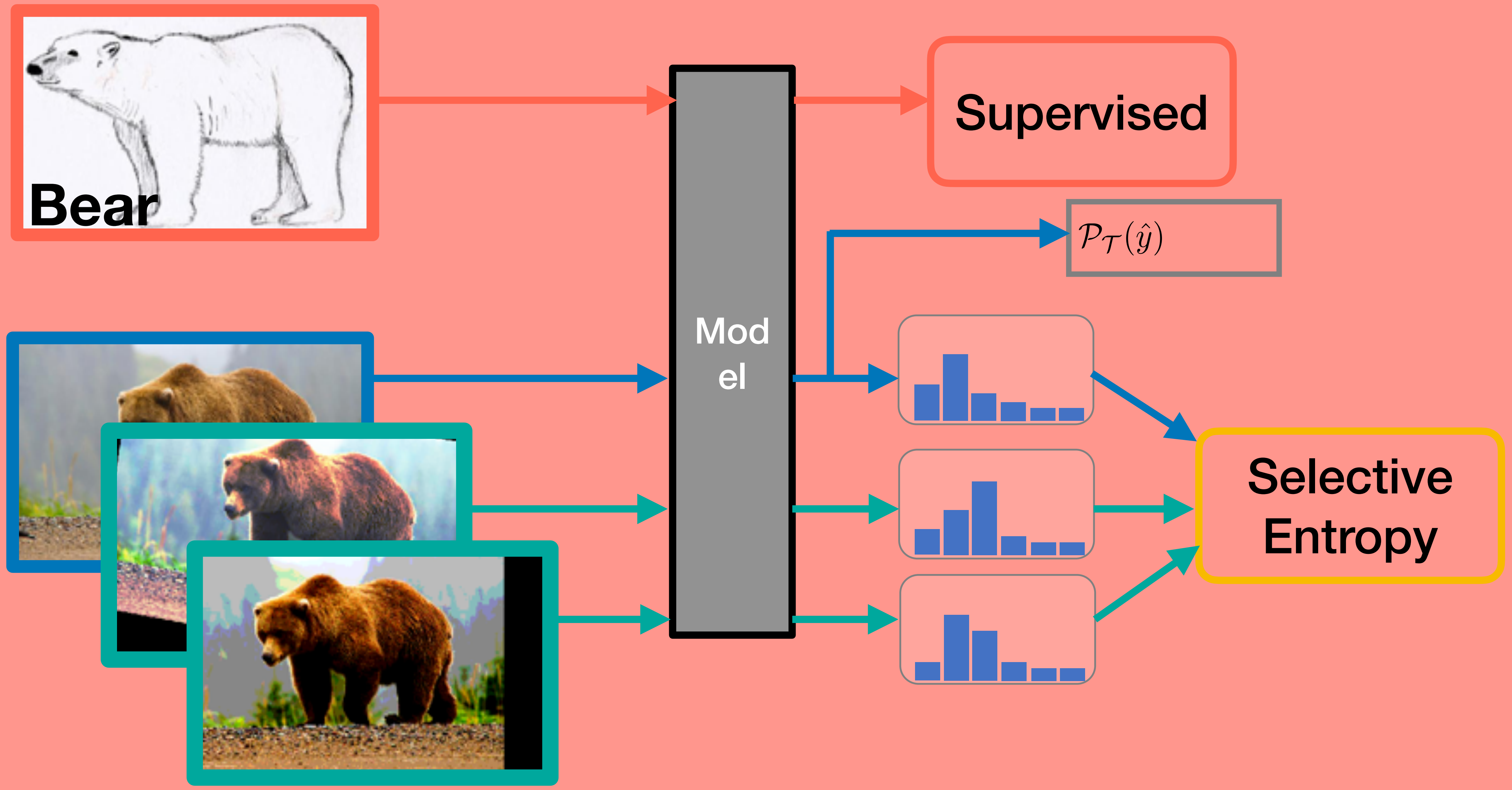
Summary: Responsible Vision

Benchmarks for Analysis



RobustNav for Embodied Nav study
Chattopadhyay et al, ICCV 2021

Domain Adaptation



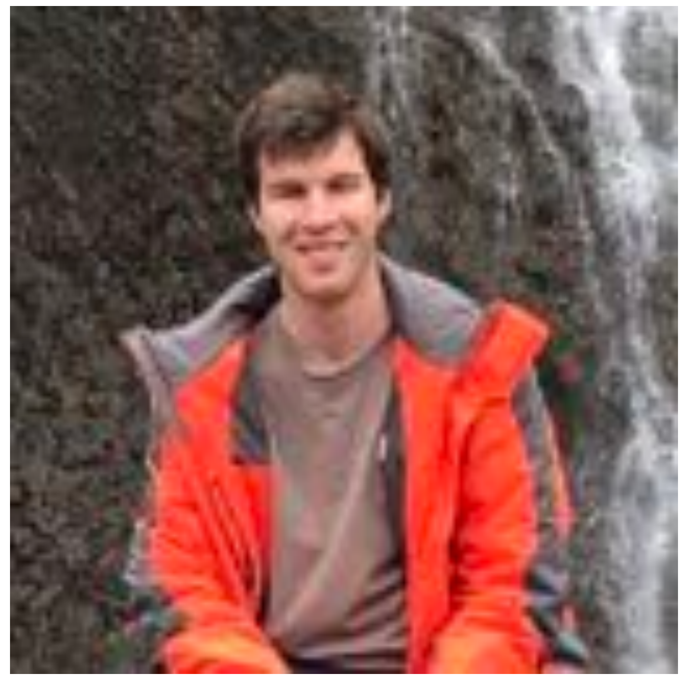
SENTRY: Selective Updates
Prabhu et al, ICCV 2021

Thank you

Sean Foley

Daniel Bolya

Sruthi Sudhakar



George Stoica

Aayushi Agarwal

Kartik Sarangmath

Prithvijit Chattopadhyay

Viraj Prabhu

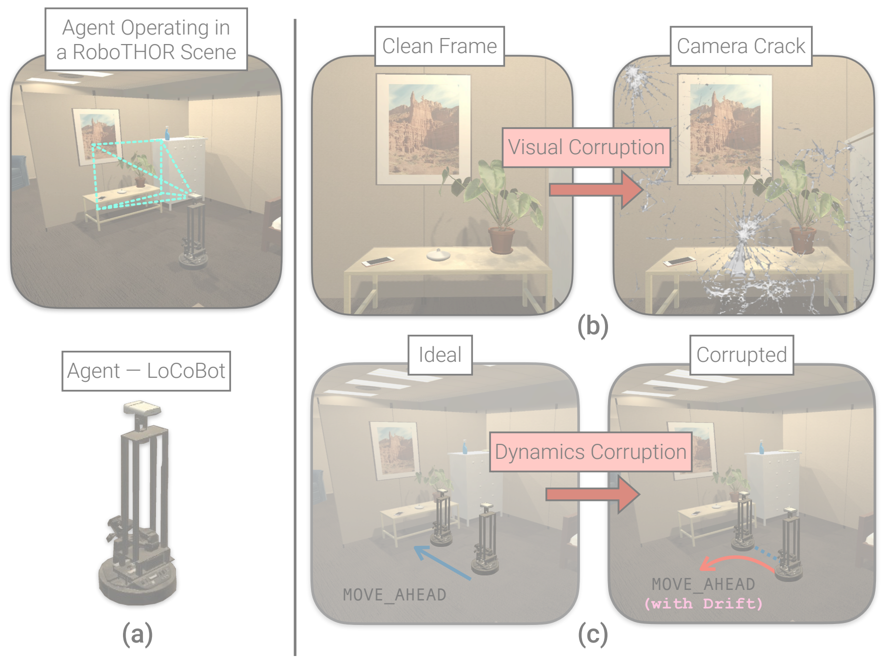
Shivam Khare

Deeksha Karthik

Bhavika Devnani

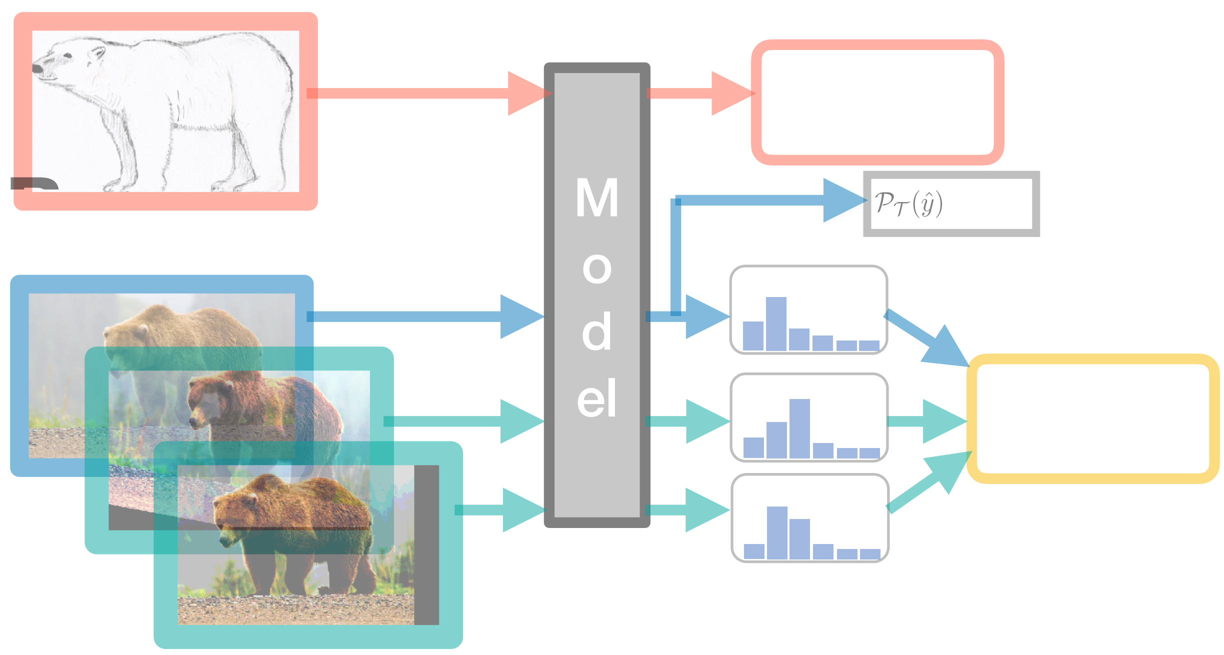
Deepanshi Deepanshi

Summary: Responsible Vision



Benchmarks for Analysis

RobustNav for Embodied Nav study
Chattopadhyay et al, ICCV 2021



Domain Adaptation

SENTRY: Selective Updates
Prabhu et al, ICCV 2021

Thank you!
Questions?
{judy,virajp}@gatech