2nd workshop on Unsupervised Learning for Automated Driving

Leveraging single and cross-modal unlabeled data for learning with limited labels

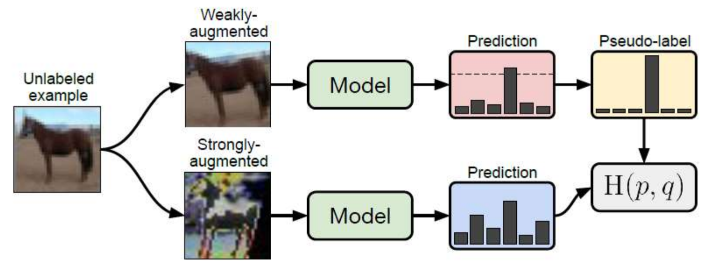
Zsolt Kira
Assistant Professor
School of Interactive Computing
Georgia Tech

Georgia Tech Machine Learning

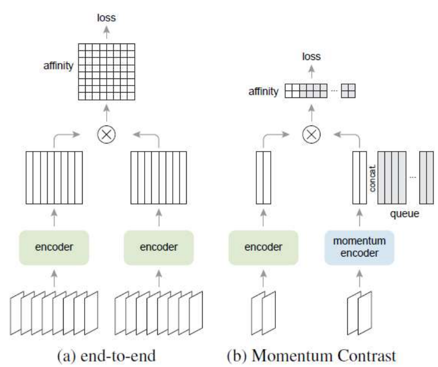
RIPL Robotics Perception and Learning

The Fast-Paced Landscape of Reducing Labels

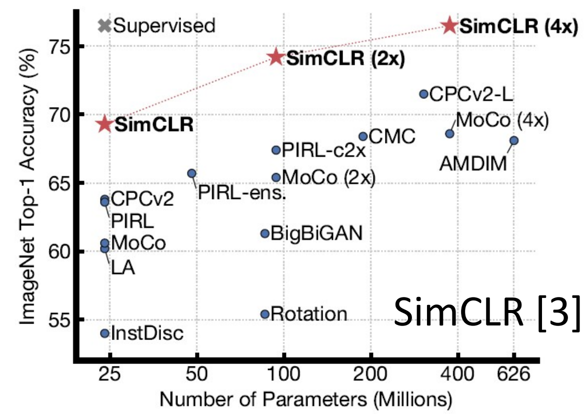
- The past two years have seen tremendous progress in:
  - Zero-shot learning
  - Few-shot learning
  - Semi-supervise learning
  - Self-supervised learning
  - Domain adaptation/generalization
  - Weakly supervised learning
  - Long-tailed datasets



FixMatch [1]



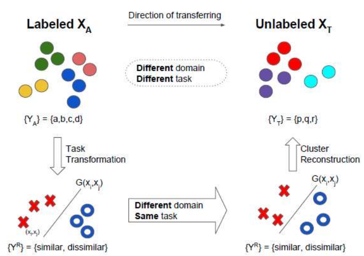
MoCo-v2 [2]



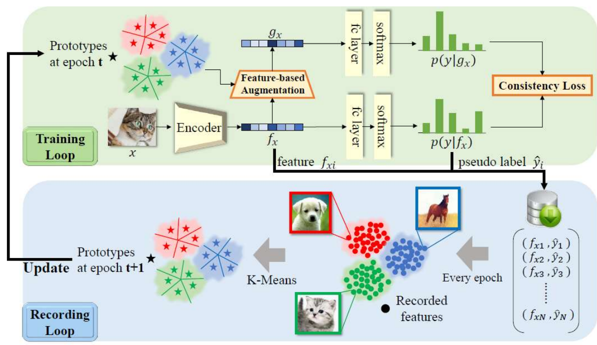
SimCLR [3]

[1] FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence, Sohn et al.
[2] Improved Baselines with Momentum Contrastive Learning, Chen et al.
[3] A Simple Framework for Contrastive Learning of Visual Representations, Chen et al.

Our Contributions

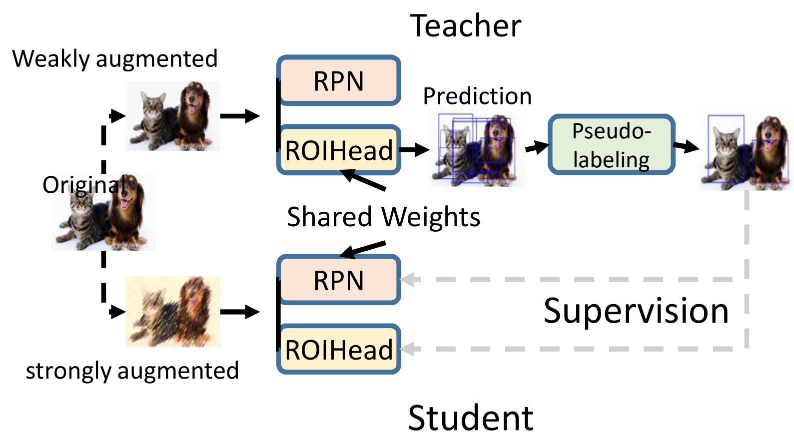


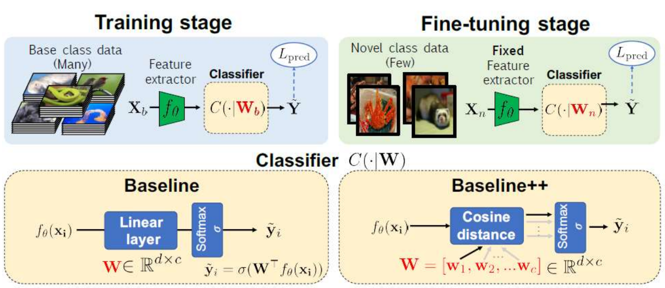
Pairwise Similarity for Cross-Task Object Discovery
[ICLR 2018, 2019]



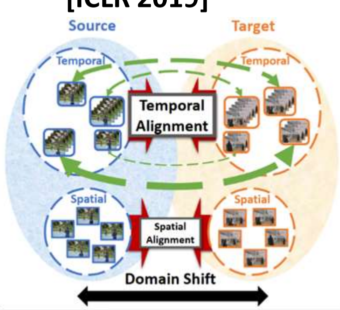
Complex Data Augmentation
Domain Generalization/SSL
[ECCV 2020]

Student-Teacher for Semi-Supervised Object Detection
[in submission, with FB]





Closer Look @ Few-Shot (w/ VT)
[ICLR 2019]



Video Domain adaptation
[CVPR 2019, ICCV 2019]

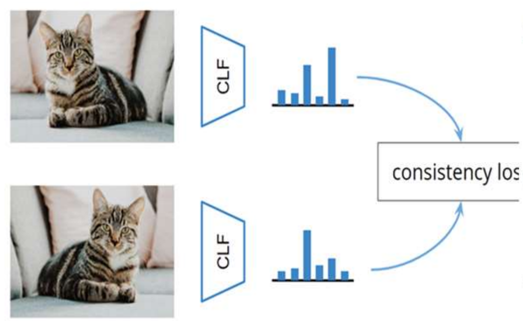
2D to 3D Inflation for semi-supervised learning
[https://arxiv.org/abs/2008.10592, with Argo AI]

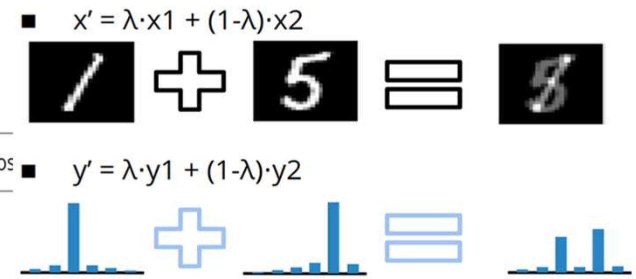
The Methods are Surprisingly Simple

- A handful of common techniques:
  - **Data augmentation**
  - **Pseudo-labeling** / distillation
  - Surrogate tasks / contrastive losses
  - Temperature scaling / Entropy maximization
  - Cosine/metric learning
  - **Prototypes**
  - **Graph neural networks**
  - Meta-learning

Setting: Semi-Supervised Learning

- Setting
  - Small amount of labeled data
  - Large amount of unlabeled data
- Example Datasets
  - SVHN
  - CIFAR-10
  - CIFAR-100
  - mini-ImageNet



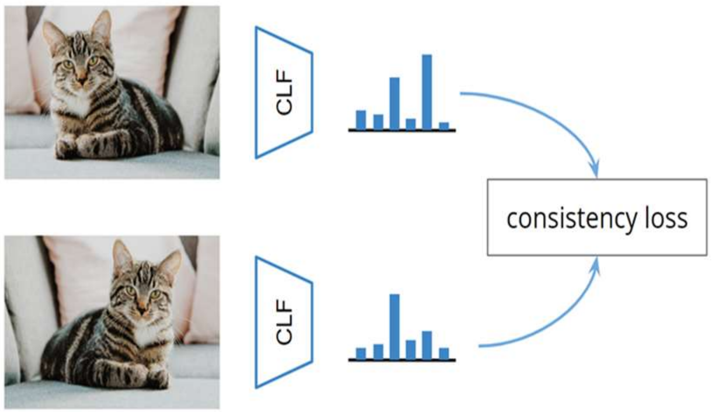


- Previous SoA method: MixMatch [1], with key contributions:
  - Consistency
  - Mixup
  - Mean teacher (for more reliable pseudo-labeling)

[1] MixMatch: A Holistic Approach to Semi-Supervised Learning

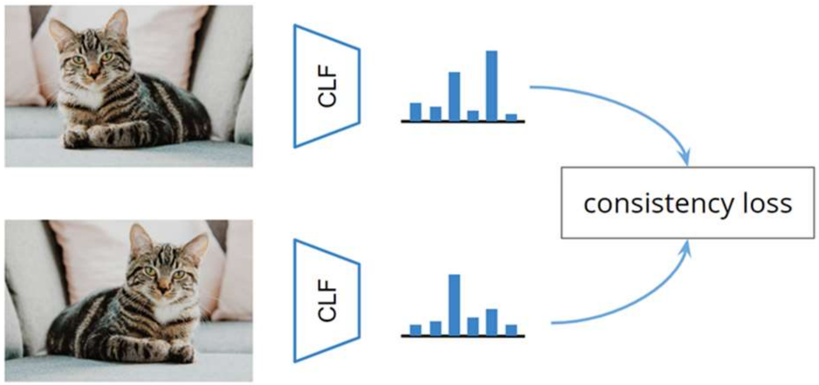
Data Augmentation

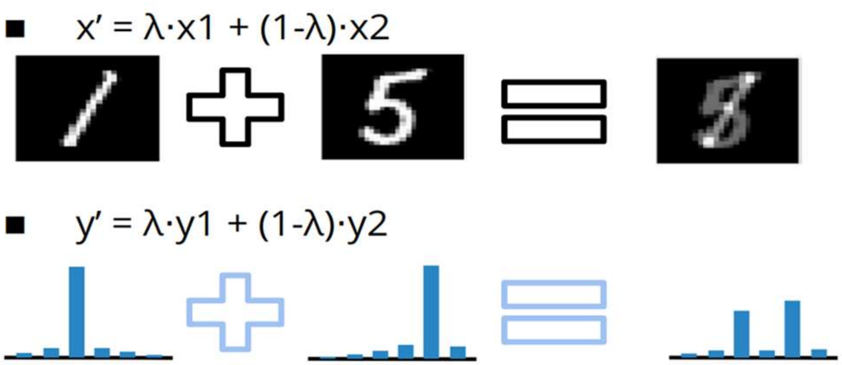
- Data augmentation key to many different areas, including:
  - Semi-supervised learning
  - Self-supervised learning
  - Showing up in few-shot learning, etc.



Limitations of Consistency-Based Method

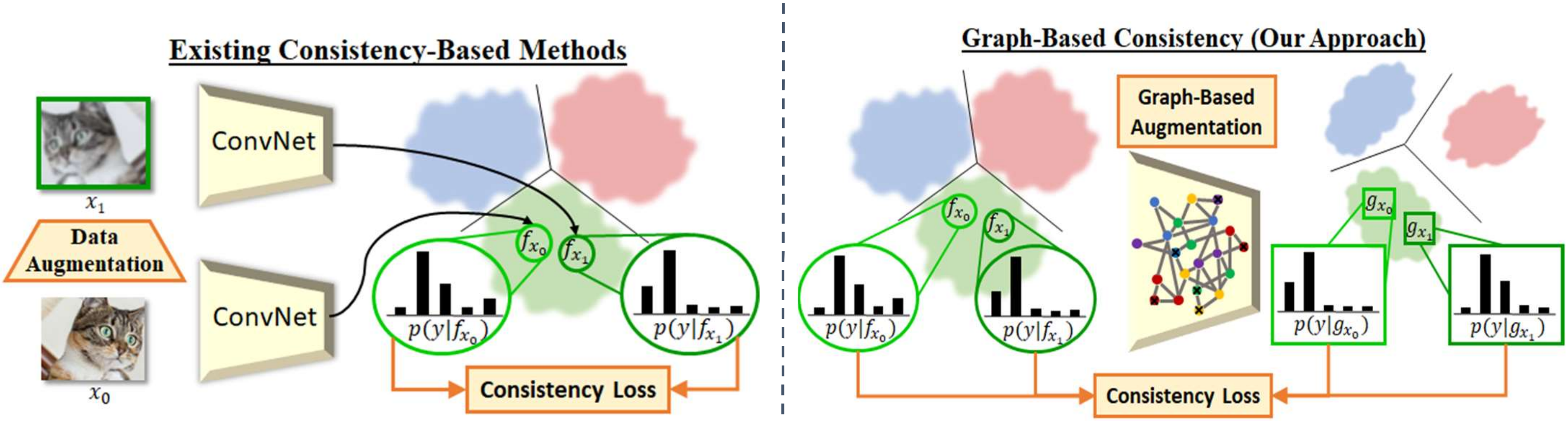
- Data augmentation only operates in **image space**
  - limits the possible transformations to textural or geometric within images.
- Data augmentation operates within a single instance
  - fails to transform data with the knowledge of other instances **(manifold structure)**
- Mixup method (sort of) addressed these issues.
  - However, the transformations are still in image space





FeatMatch: Proposed Graph-Based Consistency Method

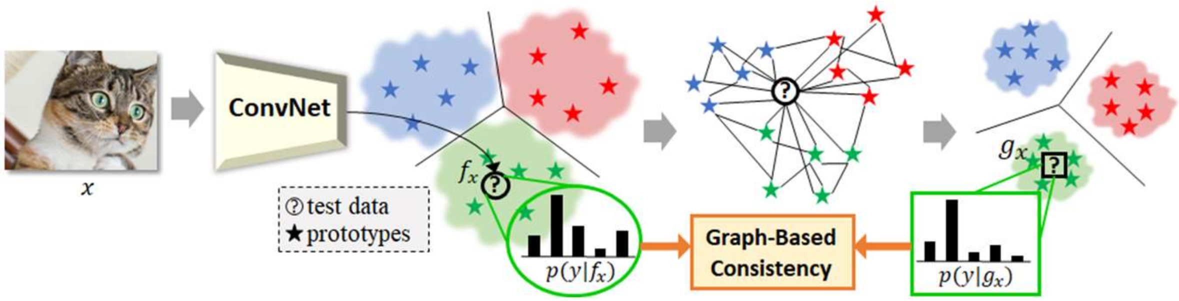
- Construct a graph neural network in the feature space for feature augmentation.
- Method is orthogonal to consistency-based methods (here combined with MixMatch, can also use FixMatch)



Kuo et al., FeatMatch: Feature-Based Augmentation for Semi-Supervised Learning, ECCV 2020

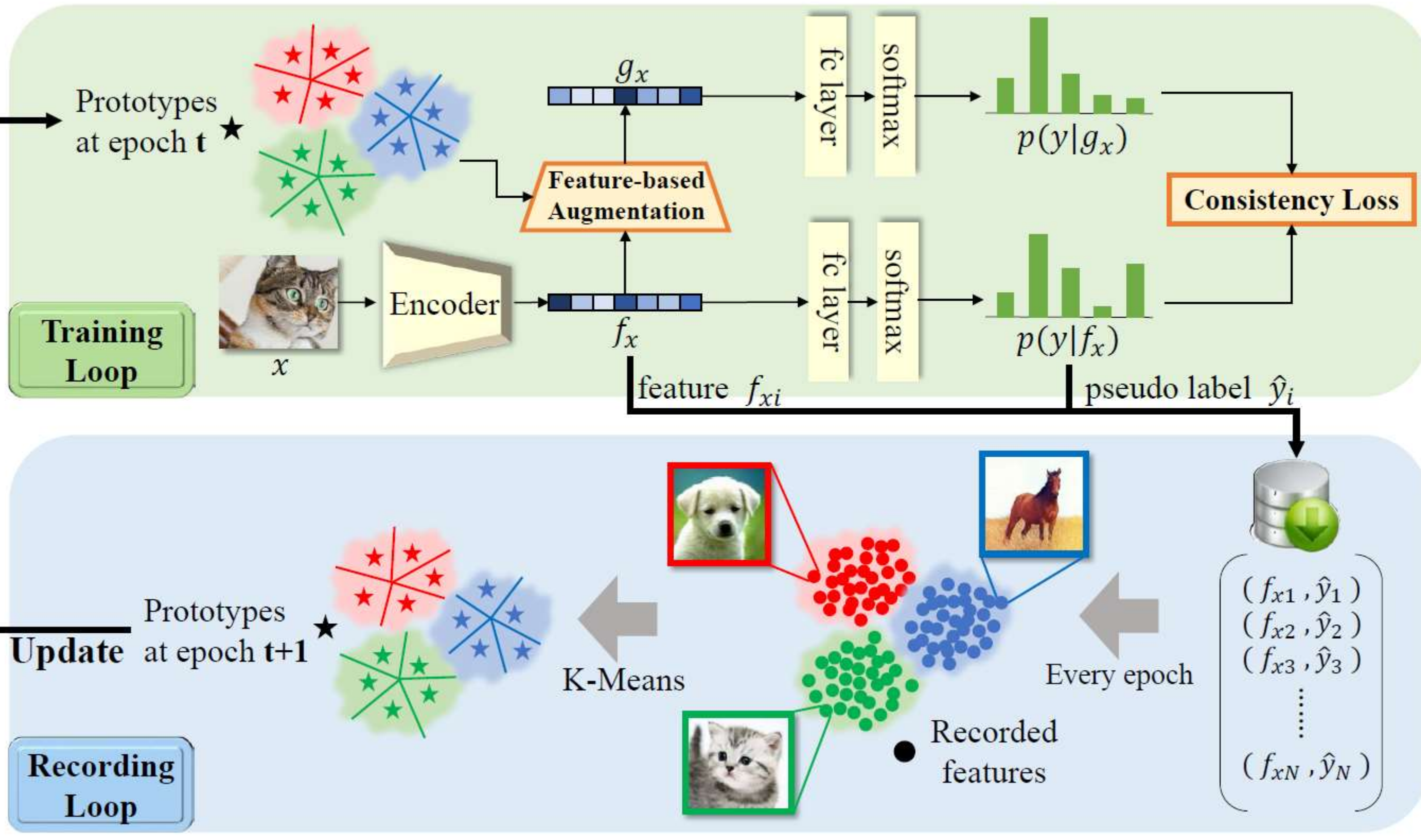
Graph Module

- The graph is constructed between image features and a set of prototypes
  - Computationally intractable to construct graph over the entire dataset
- Prototype extraction
  - K-mean centroids for each class every K iterations
  - Small set of labeled data: extract features on the fly
  - Large set of unlabeled data: record previous features ([2] has similar idea)



[2] Momentum Contrast for Unsupervised Visual Representation Learning (https://arxiv.org/abs/1911.05722)

Updating Prototypes



Comparison to Other Methods

	ReMixMatch [3]	MixMatch [4]	Mean Teacher [26]	ICT [30]	PLCB [1]	**FeatMatch (Ours)**
Feature-Based Augmentation	-	-	-	-	-	✔
Image-Based Augmentation	✔	✔	✔	✔	✔	✔
Temporal Ensembling	✔	✔	✔			-
Self-Supervised Loss	✔	-	-	-	-	-
Alignment of Class Distribution	✔	-	-	-	✔	-

Quantitative Results (Standard Datasets)

Method	Model (param.)	CIFAR-10 # Labeled samples 250	1,000	4,000	SVHN # Labeled samples 250	1,000	4,000
SSL with Memory [5]	CNN-13 (3M)	-	-	11.91± 0.22	8.83	4.21	-
Deep Co-Training [22]		-	-	8.35 ± 0.06	-	3.29 ± 0.03	-
Weight Averaging [2]		-	15.58 ± 0.12	9.05 ± 0.21	-	-	-
ICT [30]		-	15.48 ± 0.78	7.29 ± 0.02	4.78 ± 0.68	3.89 ± 0.04	-
Label Propagation [14]		-	16.93 ± 0.70	10.61 ± 0.28	-	-	-
SNTG [18]		-	18.41 ± 0.52	9.89 ± 0.34	4.29 ± 0.23	3.86 ± 0.27	-
PLCB [1]		-	6.85 ± 0.15	5.97 ± 0.15	-	-	-
Π-model [25]	WRN (1.5M)	53.02 ± 2.05	31.53 ± 0.98	17.41 ± 0.37	17.65 ± 0.27	8.60 ± 0.18	5.57 ± 0.14
PseudoLabel [17]		49.98 ± 1.17	30.91 ± 1.73	16.21 ± 0.11	21.16 ± 0.88	10.19 ± 0.41	5.71 ± 0.07
Mixup [13]		47.43 ± 0.92	25.72 ± 0.66	13.15 ± 0.20	39.97 ± 1.89	16.79 ± 0.63	7.96 ± 0.14
VAT [19]		36.03 ± 2.82	18.68 ± 0.40	11.05 ± 0.31	8.41 ± 1.01	5.98 ± 0.21	4.20 ± 0.15
Mean Teacher [26]		47.32 ± 4.71	17.32 ± 4.00	10.36 ± 0.25	6.45 ± 2.43	3.75 ± 0.10	3.39 ± 0.11
MixMatch [4]		11.08 ± 0.87	7.75 ± 0.32	6.24 ± 0.06	3.78 ± 0.26	3.27 ± 0.31	2.89 ± 0.06
ReMixMatch [3]		**6.27 ± 0.34**	**5.73 ± 0.16**	5.14 ± 0.04	**3.10 ± 0.50**	**2.83 ± 0.30**	**2.42 ± 0.09**
FeatMatch (Ours)		7.50 ± 0.64	**5.76 ± 0.07**	**4.91 ± 0.18**	**3.34 ± 0.19**	**3.10 ± 0.06**	2.62 ± 0.08

Quantitative Results (Larger Datasets)

- Our method is scalable to larger datasets and categories

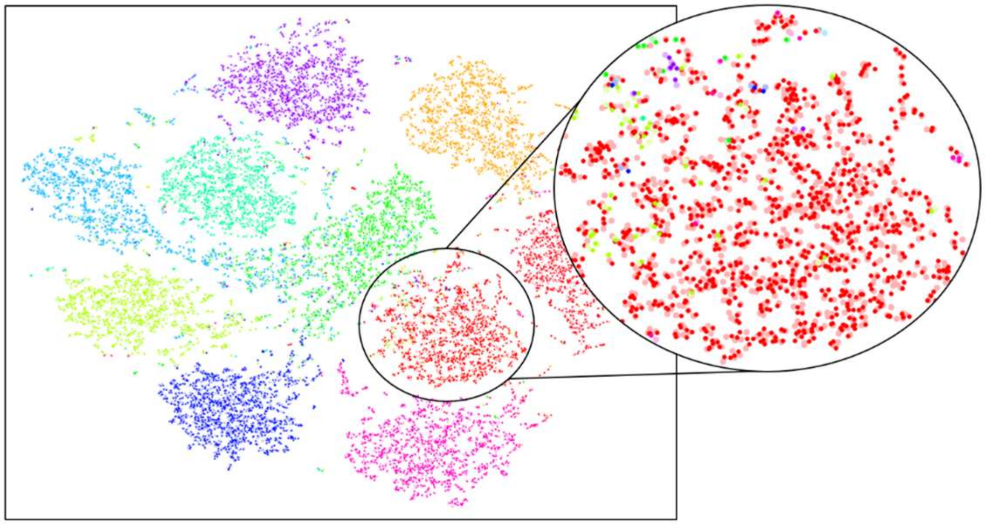
Method	CIFAR-100 # Labeled samples 4,000	CIFAR-100 # Labeled samples 10,000	mini-ImageNet # Labeled samples 4,000	mini-ImageNet # Labeled samples 10,000
Π-model [25]	-	39.19 ± 0.36	-	-
SNTG [18]	-	37.97 ± 0.29	-	-
SSL with Memory [5]	-	34.51 ± 0.61	-	-
Deep Co-Training [22]	-	34.63 ± 0.14	-	-
Weight Averaging [2]	-	33.62 ± 0.54	-	-
Mean Teacher [26]	45.36 ± 0.49	36.08 ± 0.51	72.51 ± 0.22	57.55 ± 1.11
Label Propagation [14]	43.73 ± 0.20	35.92 ± 0.47	70.29 ± 0.81	57.58 ± 1.47
PLCB [1]	37.55 ± 1.09	32.15 ± 0.50	56.49 ± 0.51	46.08 ± 0.11
FeatMatch (Ours)	$\mathbf{31.06 \pm 0.41}$	$\mathbf{26.83 \pm 0.04}$	$\mathbf{39.05 \pm 0.06}$	$\mathbf{34.79 \pm 0.22}$

Quantitative Results (DomainNet)

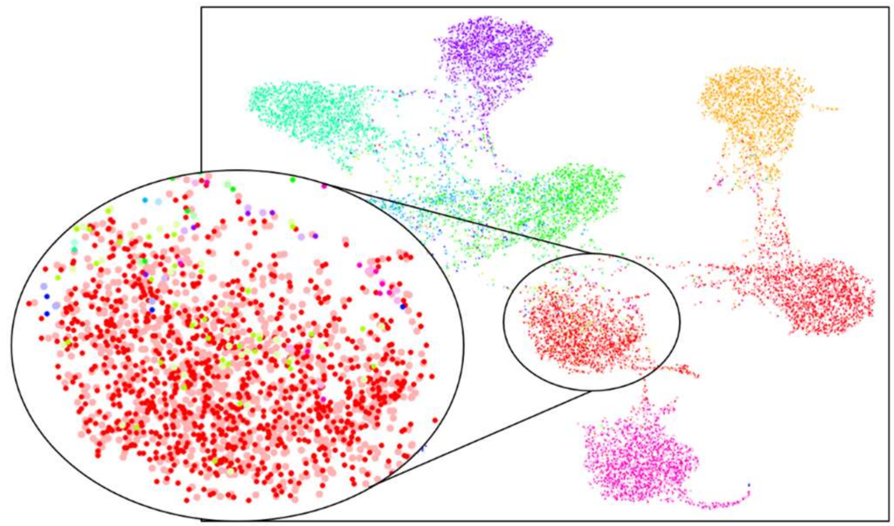
- Our method is more robust to out-of-distribution unlabeled data
  - Half of data from shifted domain for r = 50%

Method (5% labeled samples)	$r_u = 0\%$	$r_u = 25\%$	$r_u = 50\%$	$r_u = 75\%$
(Semi-supervised) Baseline	56.63 ± 0.17	62.44 ± 0.67 %	65.82 ± 0.07	70.50 ± 0.51
FeatMatch (Ours)	**40.66 ± 0.60**	**46.11 ± 1.15**	**54.01 ± 0.66**	**58.30 ± 0.93**
Supervised baseline (5% labeled samples, lower bound)		77.25 ± 0.52		
Supervised baseline (100% labeled samples, upper bound)		31.91 ± 0.15		

Qualitative Results



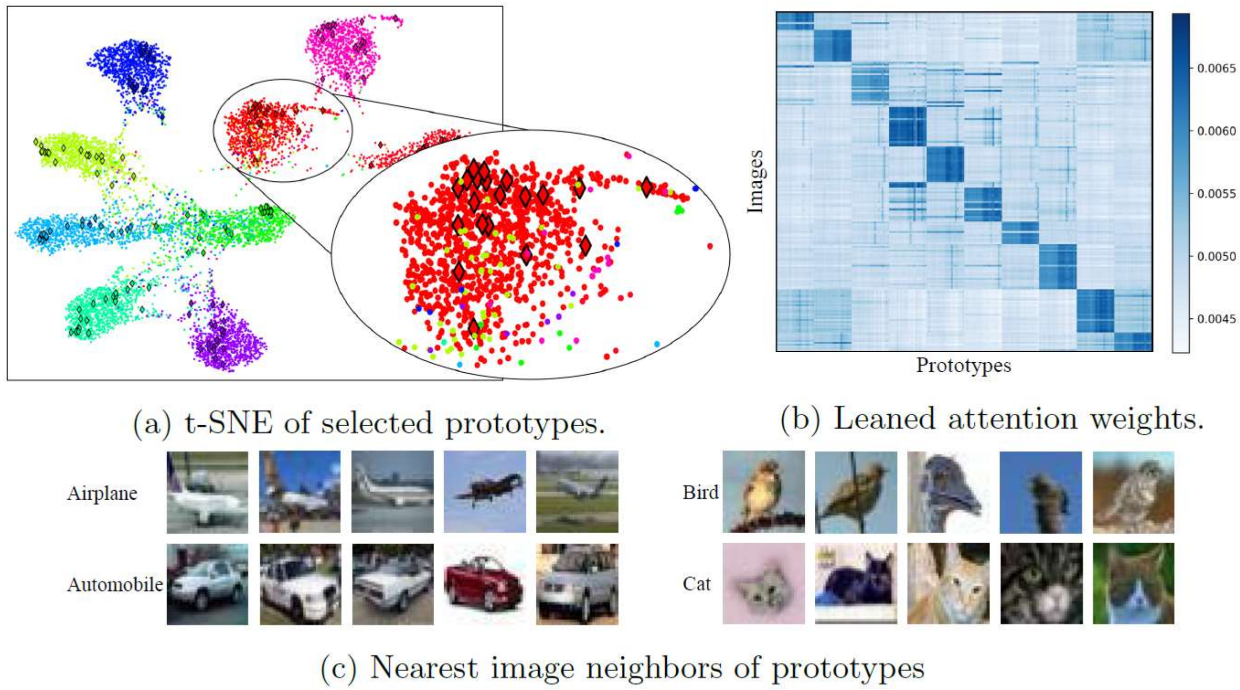
Data Augmentation



Graph-Based
Feature Augmentation

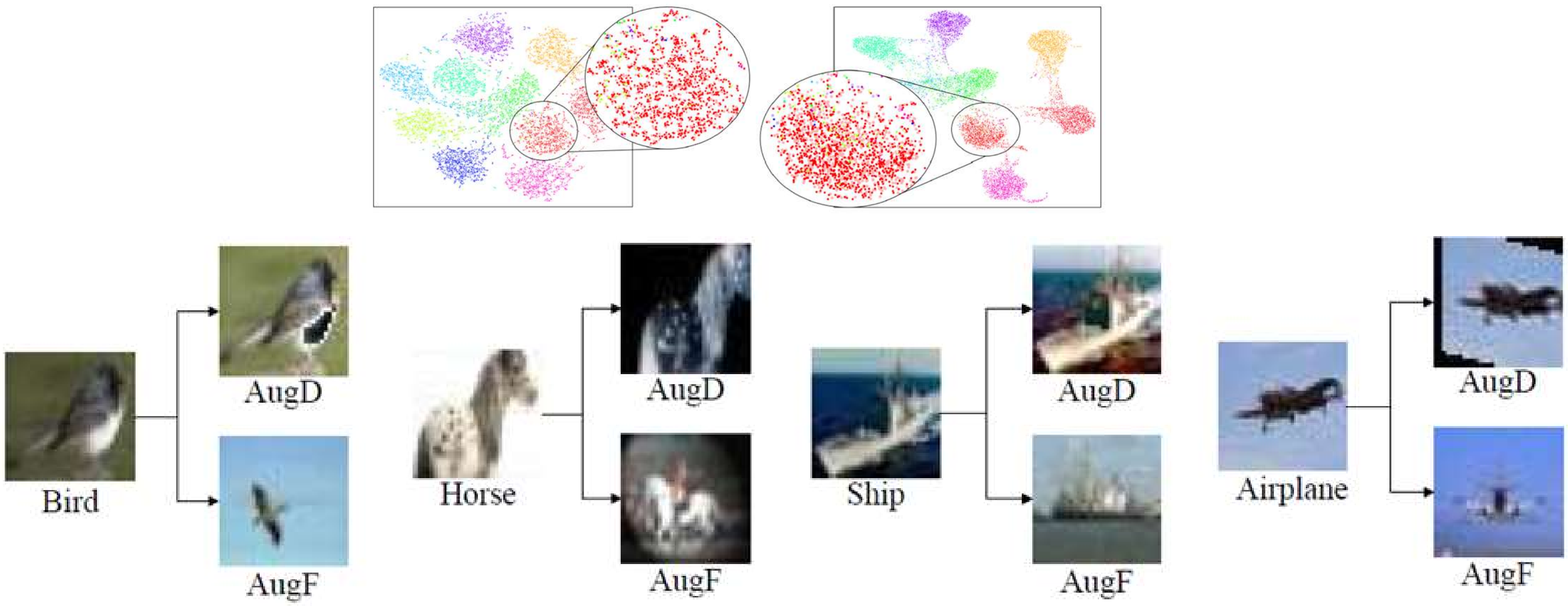
15

Prototypes and Similarity



- We can see variability in prototypes, and similarity function largely focuses on same-class prototypes

Augmentation Visualization



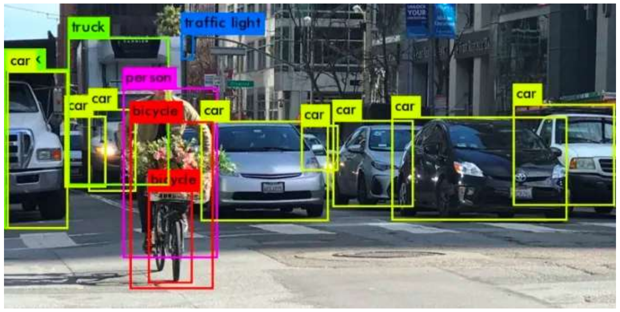
- Graph-based augmentation produces more variable, uniformly distributed augmentations

The Methods are Surprisingly Simple

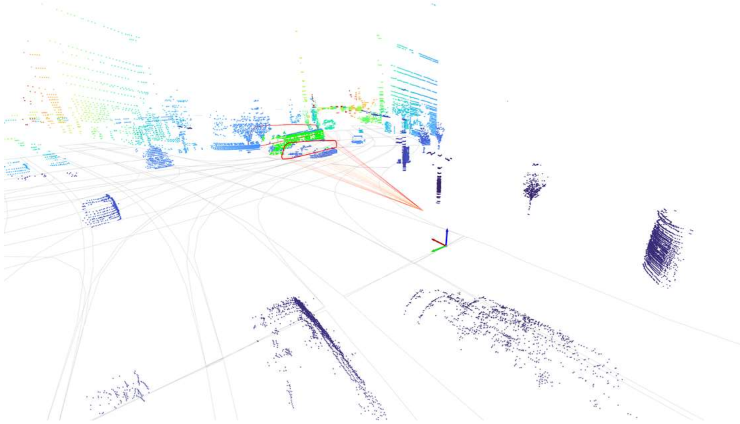
- A handful of common techniques:
  - **Data augmentation**
  - **Pseudo-labeling** / distillation
  - Surrogate tasks / contrastive losses
  - Temperature scaling / Entropy maximization
  - Cosine/metric learning
  - **Prototypes**
  - **Graph neural networks**
  - Meta-learning

- Autonomous vehicles:
  - **Object Detection?**
  - **Multi-modal?**

What about Object Detection?



What about 3D?



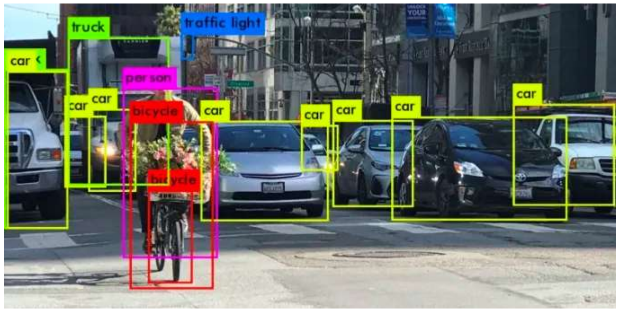
Motivation: Object Detection

- Few papers address 2-stage detector under semi-supervised setting.
    - ISD [1] and CSD [2] mainly focus on 1-stage detector
- 2-stage detectors generally have more accurate predictions.
    - Previous works focus on ROIheads training, and RPN net training is seldom explored.

Image Classification
(Single Object)

"Cat"

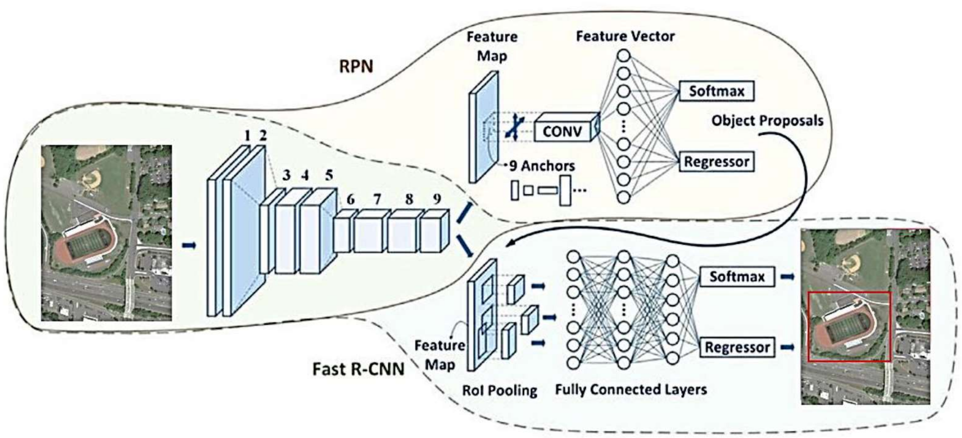
Object Detection
(Multiple objects + Bounding boxes)



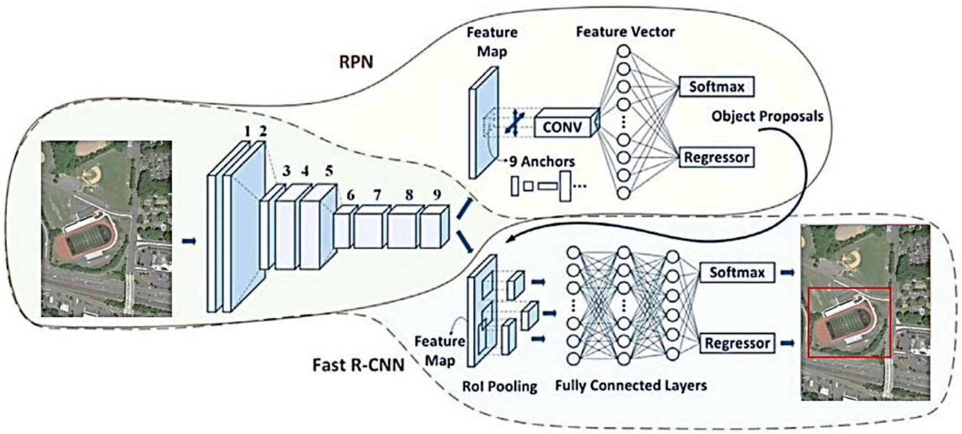
"Bike": (120, 201, 356, 347)
"Car": (345, 318, 945, 847)
"Truck": (420, 512, 601, 782)
"Traffic Light": (430, 60, 467, 123)

[1] "Consistency-based Semi-supervised Learning for Object detection", Jeong et al., NeurIPS 2019
[2] "Interpolation-based semi-supervised learning for object detection", Jeong et al., arXiv June 2020

1. Two-stage detector
   - 3 modules – Feature Backbone, Region Proposal Network, and ROIHead



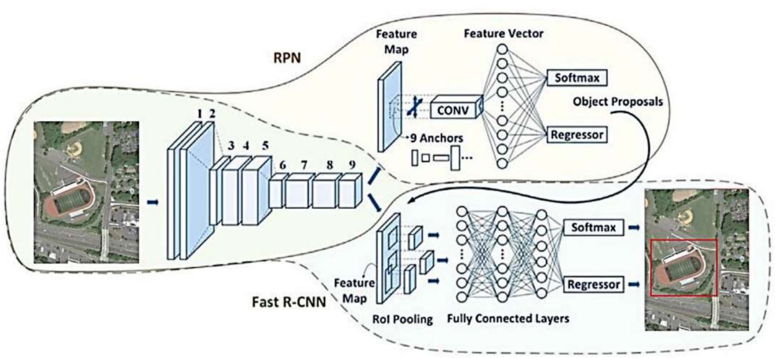
1. Two-stage detector
   - 4 predictions/losses
     - RPN - Foreground/Background Detection, Bounding Box Regression
     - ROI head – Patch Classification, Bounding Box Regression

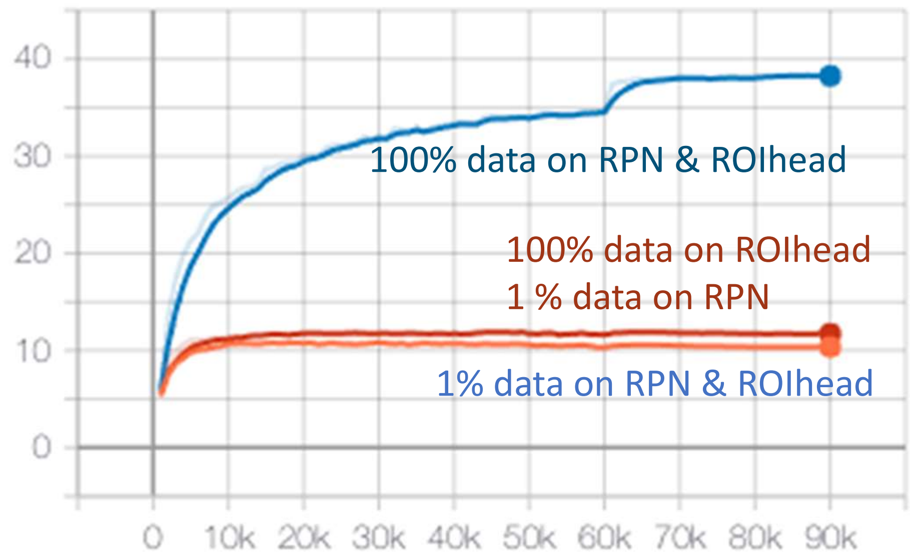


Observation 1: A good RPN is necessary!

Goal: Verify a good RPN is important; A good ROIhead requires a good RPN

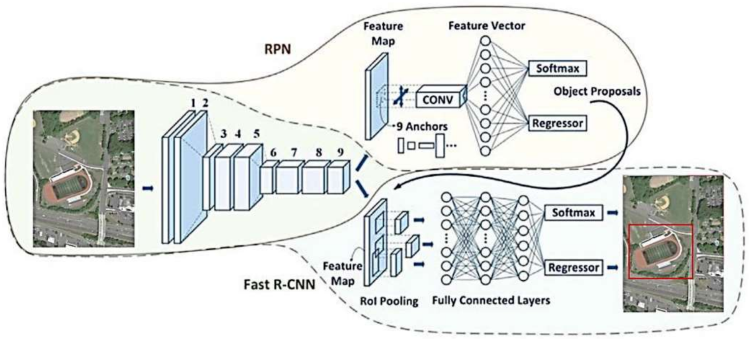
Train a model with
- RPN: 1% supervised data
- ROIHead: 100% supervised data



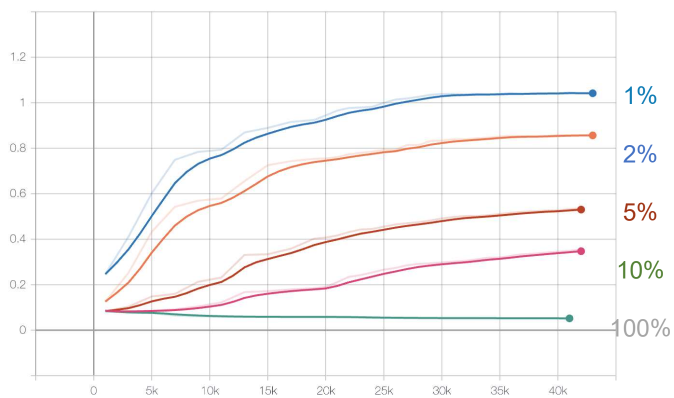


Observation 2: Overfitting!

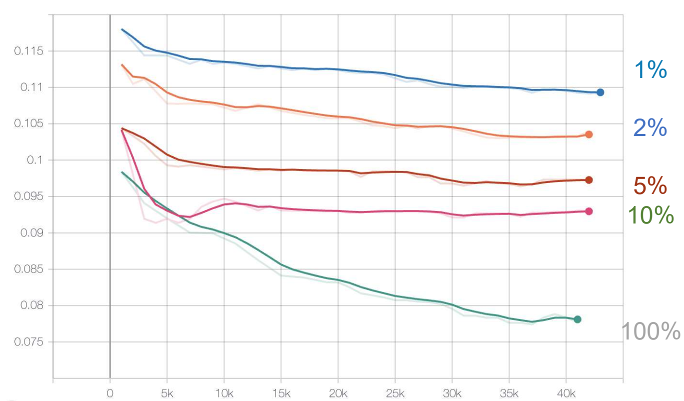
2. When the labels are insufficient …
    - Overfitting!
        - Foreground/Background Classification



Validation Fg-Bg Classification Loss
(RPN)



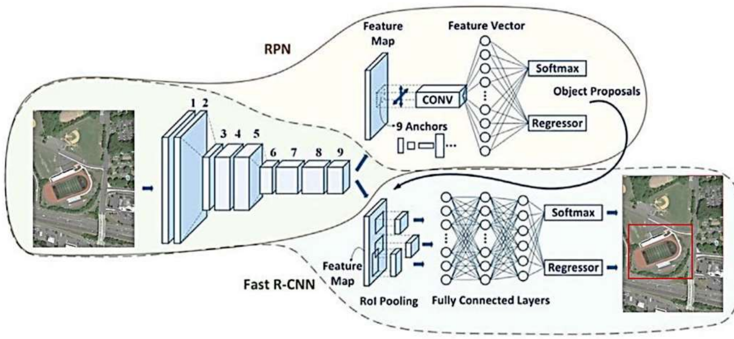
Validation Box Regression Loss
(RPN)



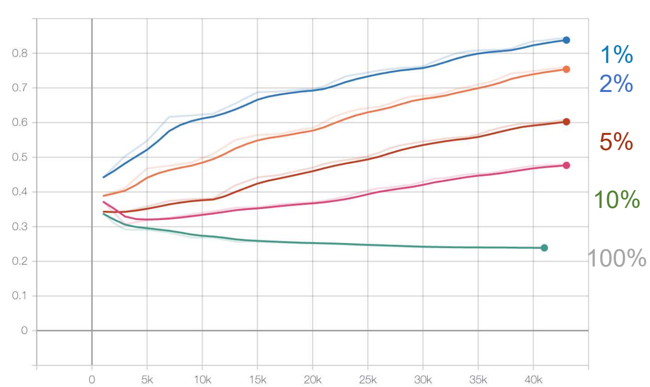
23

Observations

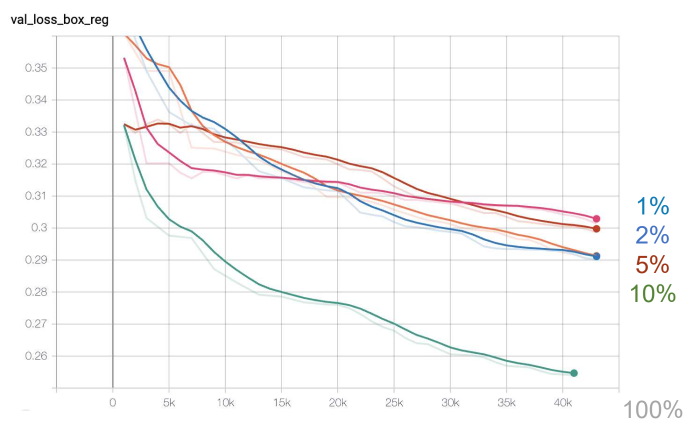
2. When the labels are insufficient …
    - Overfitting!
        - Foreground/Background Classification
        - Patch Classification



Validation Patch Classification Loss
(ROIHead)



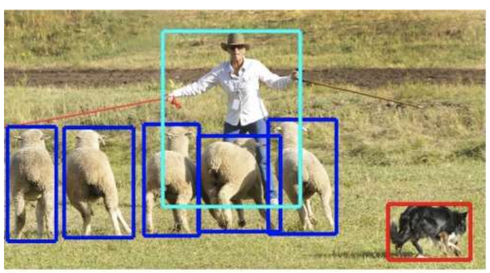
Validation Box Regression Loss
(ROIHead)



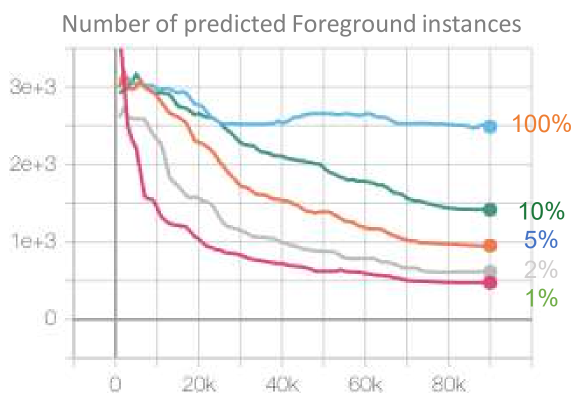
Why does overfitting occur?

1. Foreground-background imbalance
2. Foreground Classes imbalance

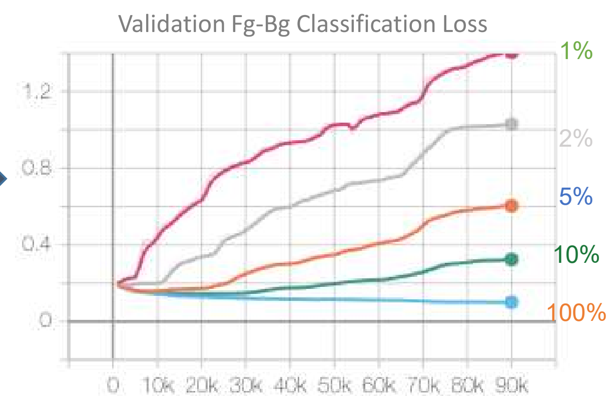
Foreground : Background = 1 : 3
(Ground-truth data)



Model biases to Background



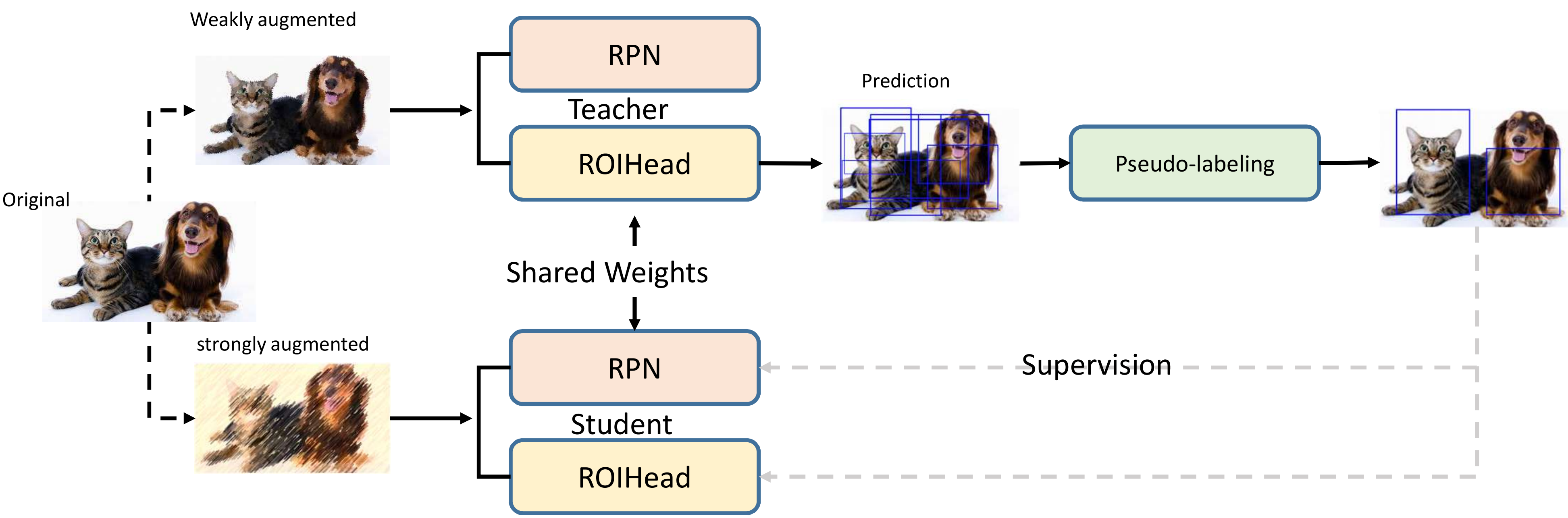
Overfitting for fg-bg prediction



Semi-supervised Object Detection

Problem 1: Labeled data is insufficient

Trial: Using the state-of-the-art semi-supervised classification method [4]?

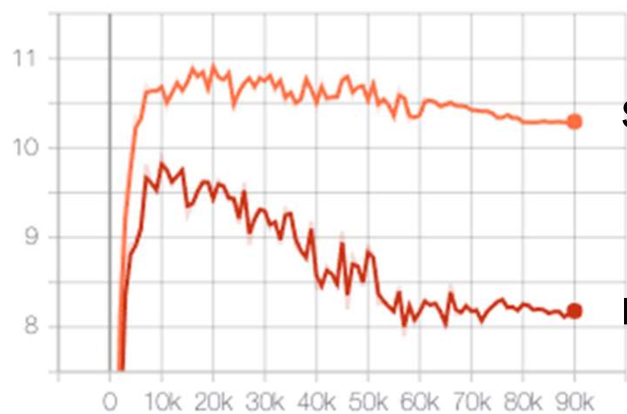


Semi-supervised Object Detection

Problem 1: Labeled data is insufficient

Trial: Using the state-of-the-art semi-supervised classification method [4]?

Average Precision



Supervised only Model (1% supervised data)

FixMatch [4] (1% supervised data)

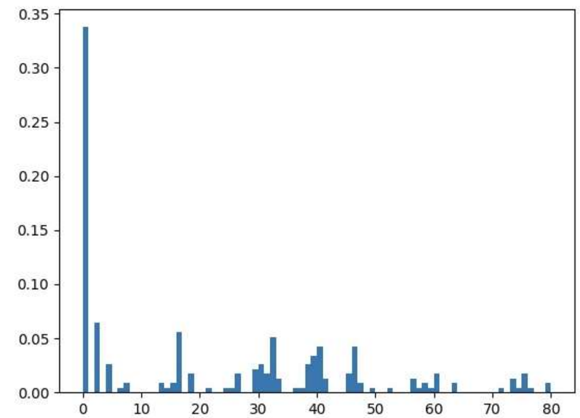
[4] "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence", Sohn et al. arXiv, Jan 2020

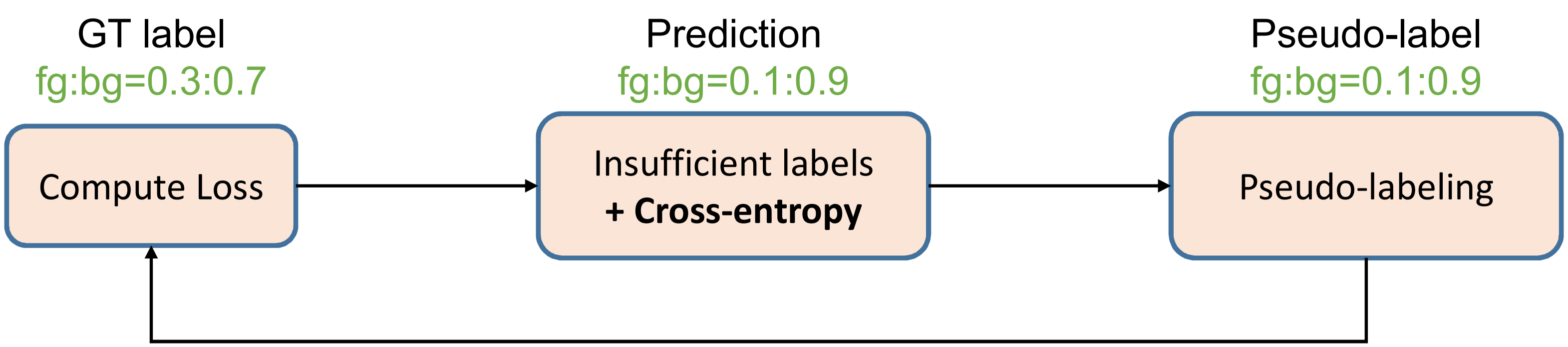
Semi-supervised Object Detection

Why SOTA semi-supervised classification cannot work?

Reason: Class imbalance!

Pseudo-labeling is a closed loop

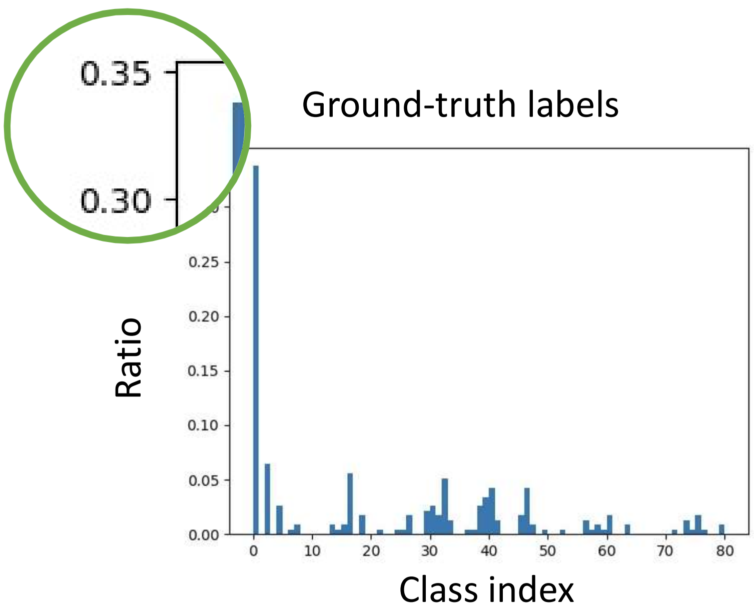




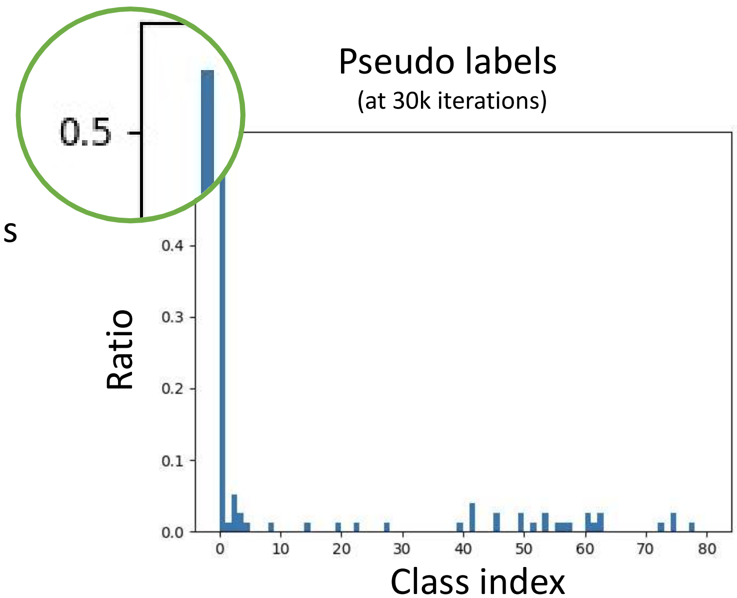
Semi-supervised Object Detection

Problem 2: Pseudo-label biases

- Cross-entropy → Model biases majority classes

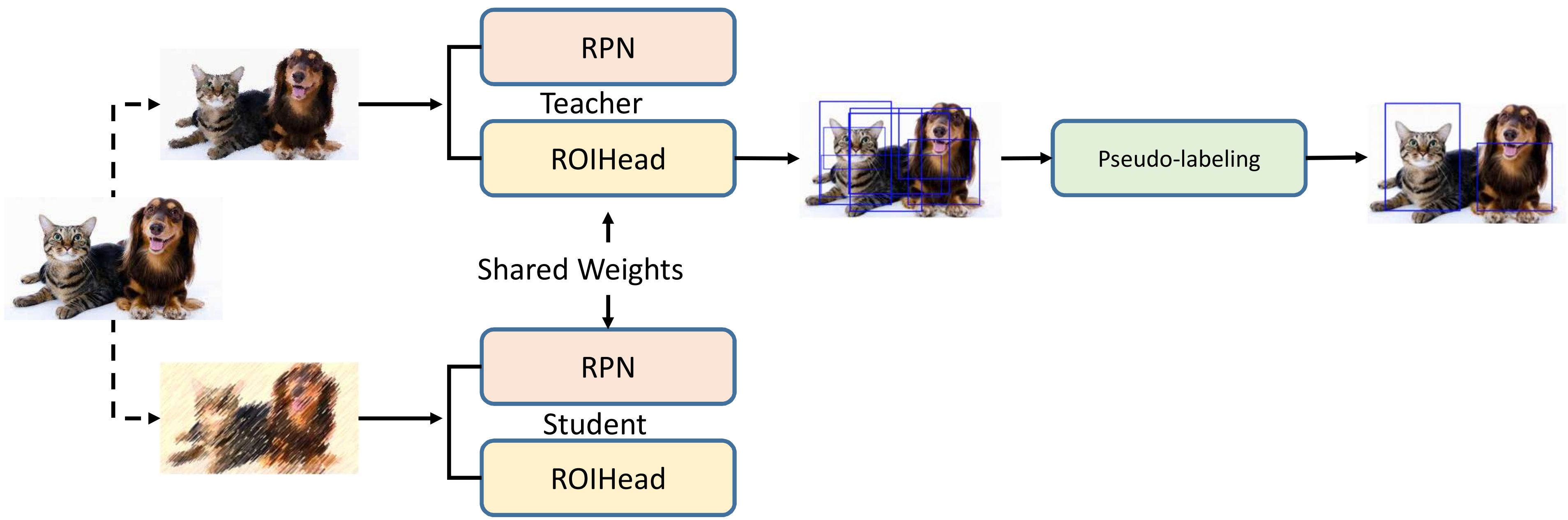


After training
for 30k iterations



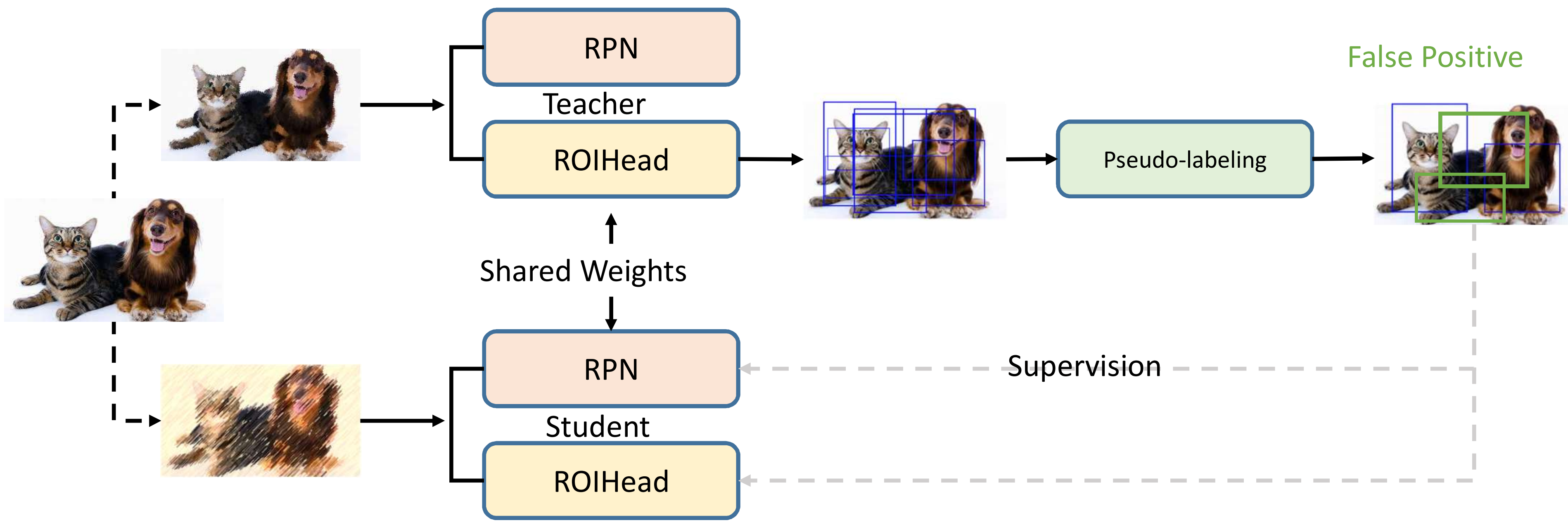
Semi-supervised Object Detection

Problem 3: Pseudo-label generation is not stable; False positive samples are detrimental



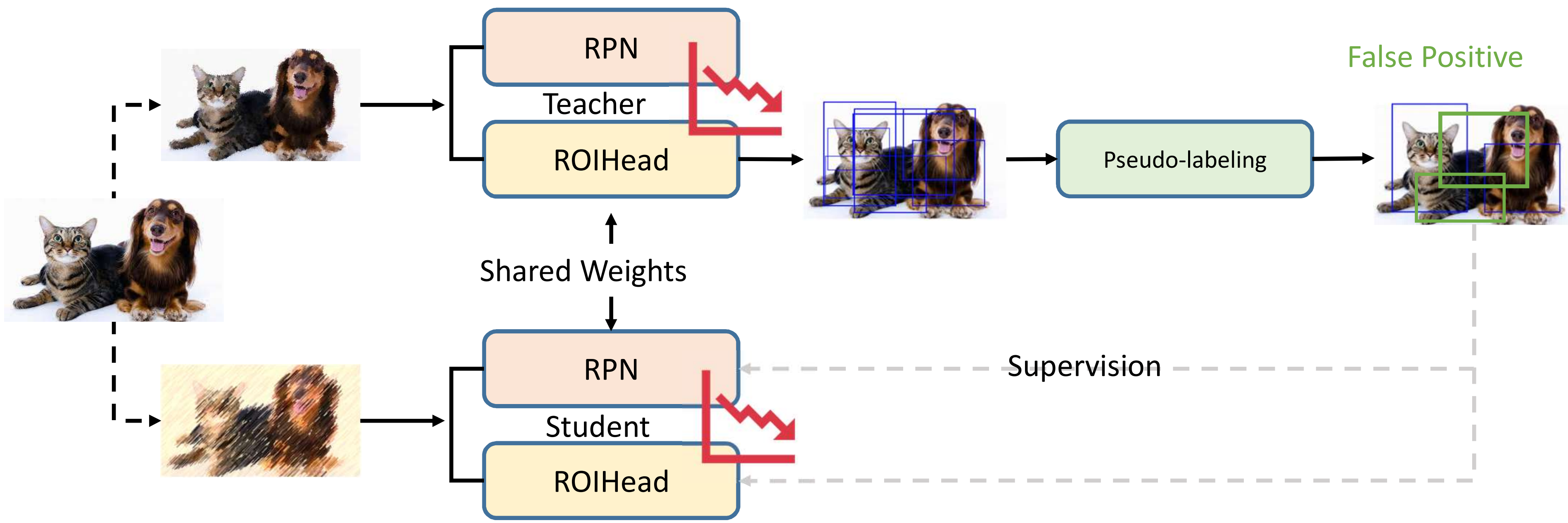
Semi-supervised Object Detection

Problem 3: Pseudo-label generation is not stable; False positive samples are detrimental

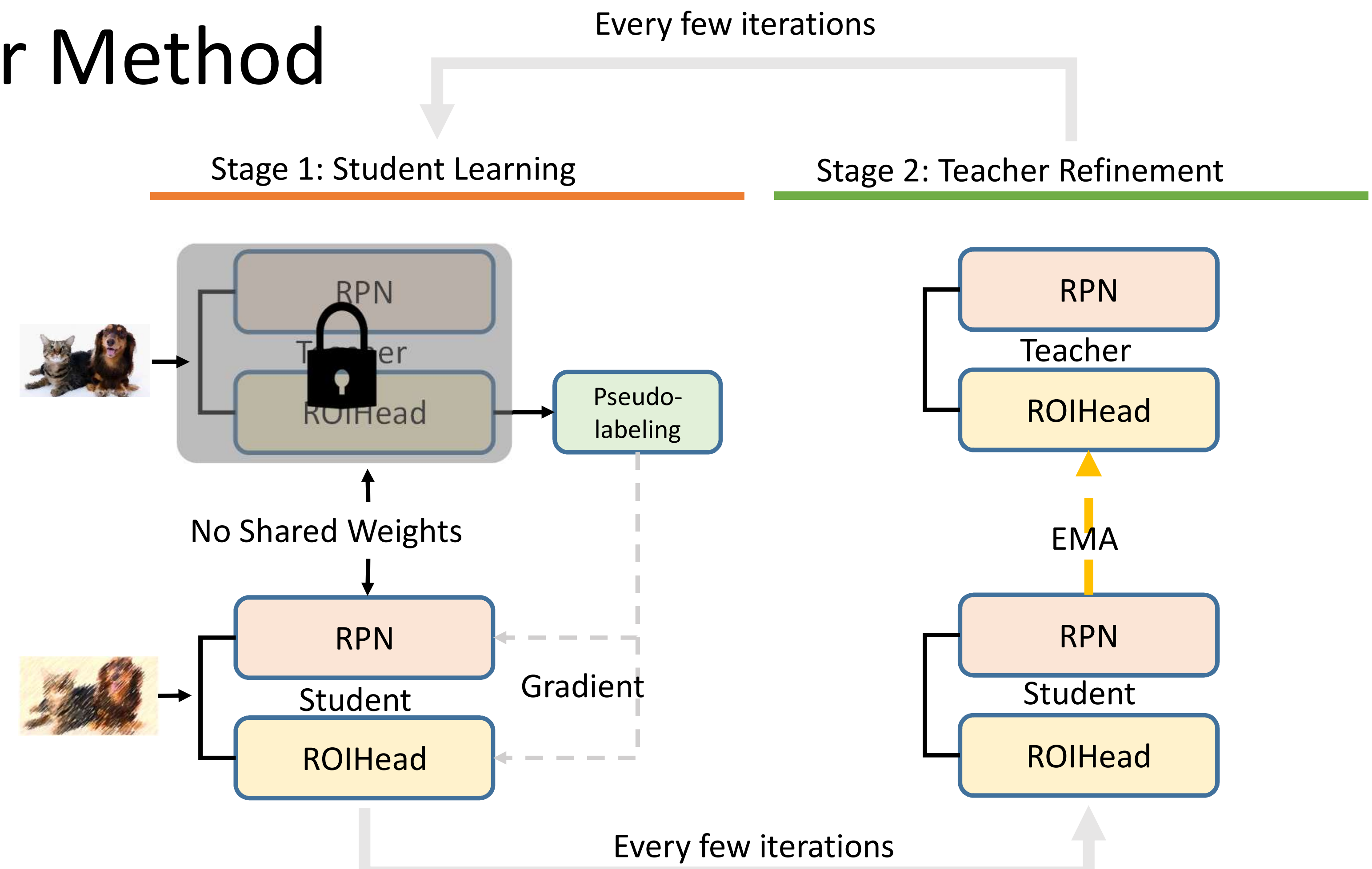


Semi-supervised Object Detection

Problem 3: Pseudo-label generation is not stable; False positive samples are detrimental



Our Method


Every few iterations
Stage 1: Student Learning
Stage 2: Teacher Refinement
RPN
Teacher
ROIHead
Pseudo-
labeling
No Shared Weights
RPN
Student
ROIHead
Gradient
RPN
Teacher
ROIHead
EMA
RPN
Student
ROIHead
Every few iterations

Experiments

	1%	2%	5%	10%
Supervised only	9.05	12.70	18.47	23.86

34

Experiments

	1%	2%	5%	10%
Supervised only	9.05	12.70	18.47	23.86
STAC [2] (SOTA from Google)	13.97 (+4.92)	18.25 (+5.55)	24.38 (+5.91)	28.64 (+4.78)

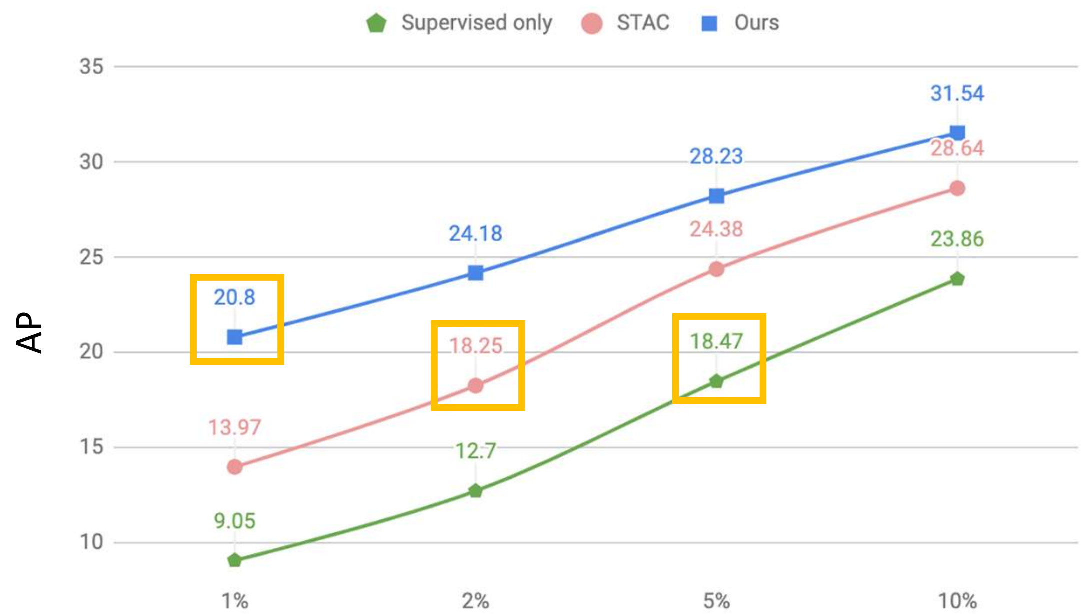
[2]"A Simple Semi-Supervised Learning Framework for Object Detection", Sohn et al., arXiv May 2020

Experiments

	1%	2%	5%	10%
Supervised only	9.05	12.70	18.47	23.86
STAC [2] (SOTA from Google)	13.97 (+4.92)	18.25 (+5.55)	24.38 (+5.91)	28.64 (+4.78)
Ours	20.80 (+11.75)	24.18 (+11.48)	28.23 (+9.76)	31.54 (+7.68)

[2]"A Simple Semi-Supervised Learning Framework for Object Detection", Sohn et al., arXiv May 2020

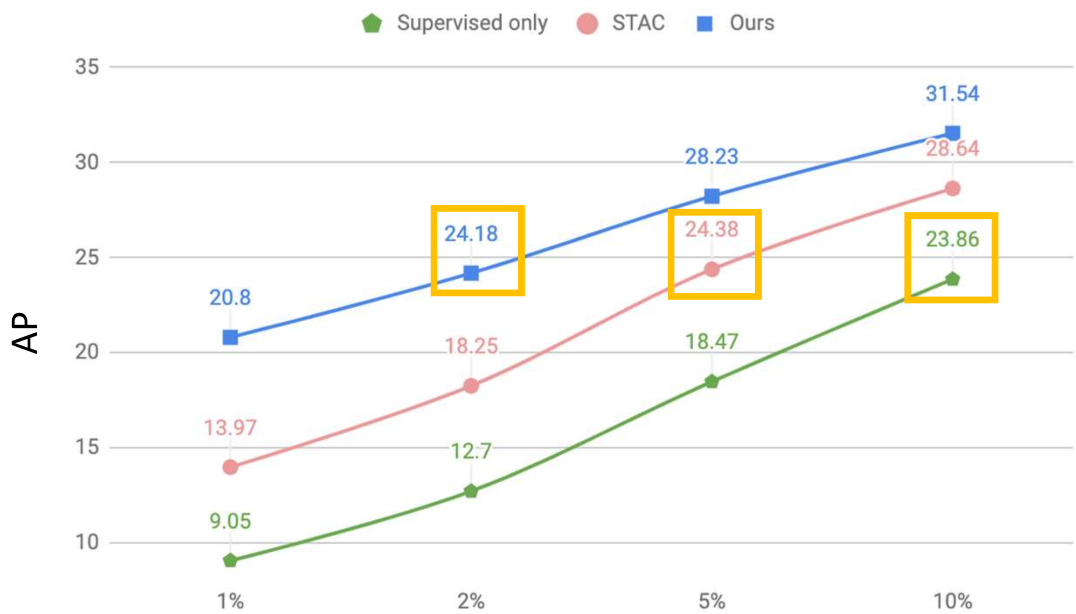
Experiments



	1%	2%	5%	10%
Supervised only	9.05	12.70	18.47	23.86
STAC [2] (SOTA from Google)	13.97 (+4.92)	18.25 (+5.55)	24.38 (+5.91)	28.64 (+4.78)
Ours	20.80 (+11.75)	24.18 (+11.48)	28.23 (+9.76)	31.54 (+7.68)

[2]"A Simple Semi-Supervised Learning Framework for Object Detection", Sohn et al., arXiv May 2020

Experiments

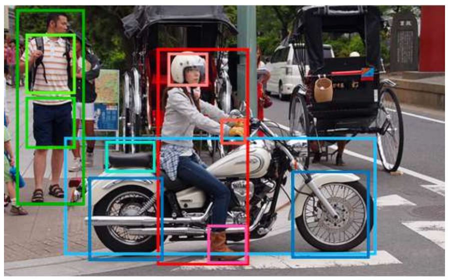


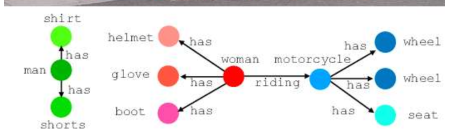
	1%	2%	5%	10%
Supervised only	9.05	12.70	18.47	23.86
STAC [2] (SOTA from Google)	13.97 (+4.92)	18.25 (+5.55)	24.38 (+5.91)	28.64 (+4.78)
Ours	20.80 (+11.75)	24.18 (+11.48)	28.23 (+9.76)	31.54 (+7.68)

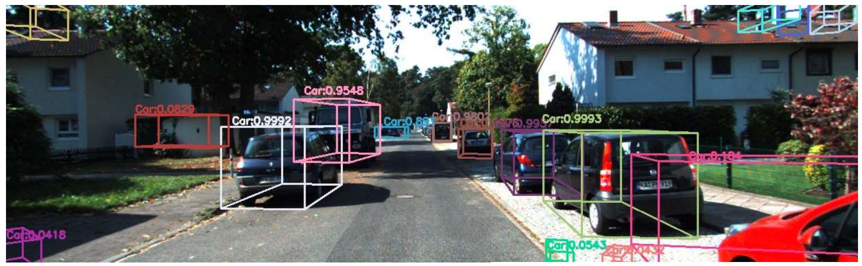
[2]"A Simple Semi-Supervised Learning Framework for Object Detection", Sohn et al., arXiv May 2020

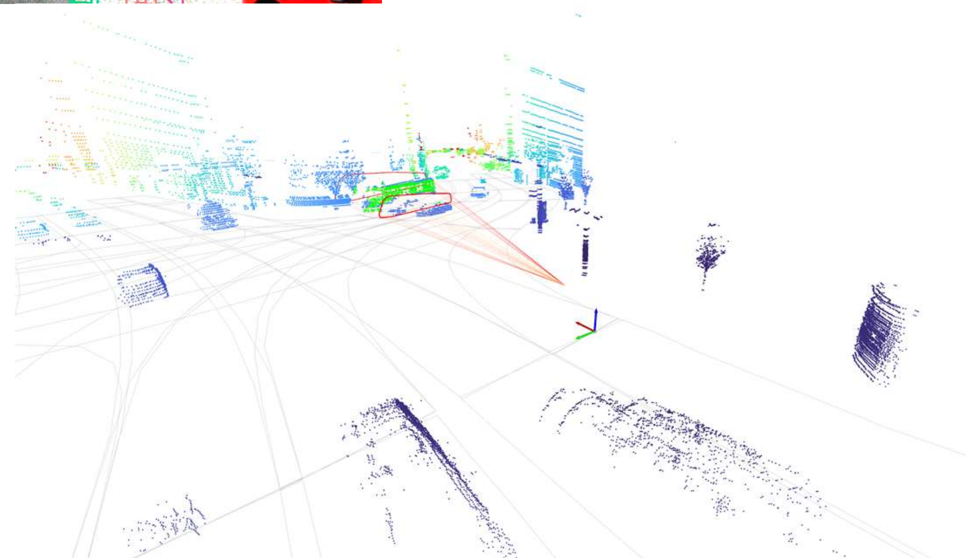
Summary

- Perform the state-of-art on semi-supervised object detection
- Different tasks based on object detection can benefit from this model









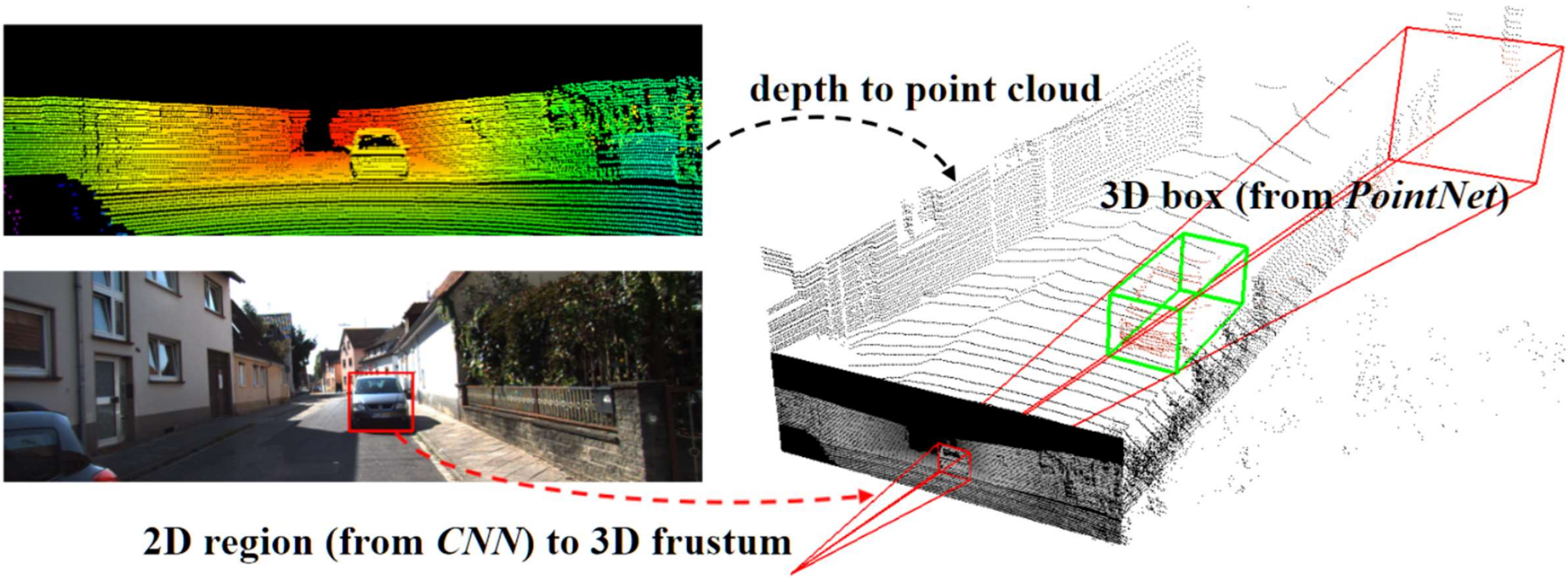
What about 3D?

3D Object Detection Goal:
Recover the *amodal* 3D spatial extent and heading of objects in a scene.

Can we “inflate” 2D instance segmentations into 3D cuboids?

Slide Credit: Ben Wilson & James Hays / Argo AI

Wilson et al., 3D for Free: Crossmodal Transfer Learning using HD Maps, https://arxiv.org/abs/2008.10592



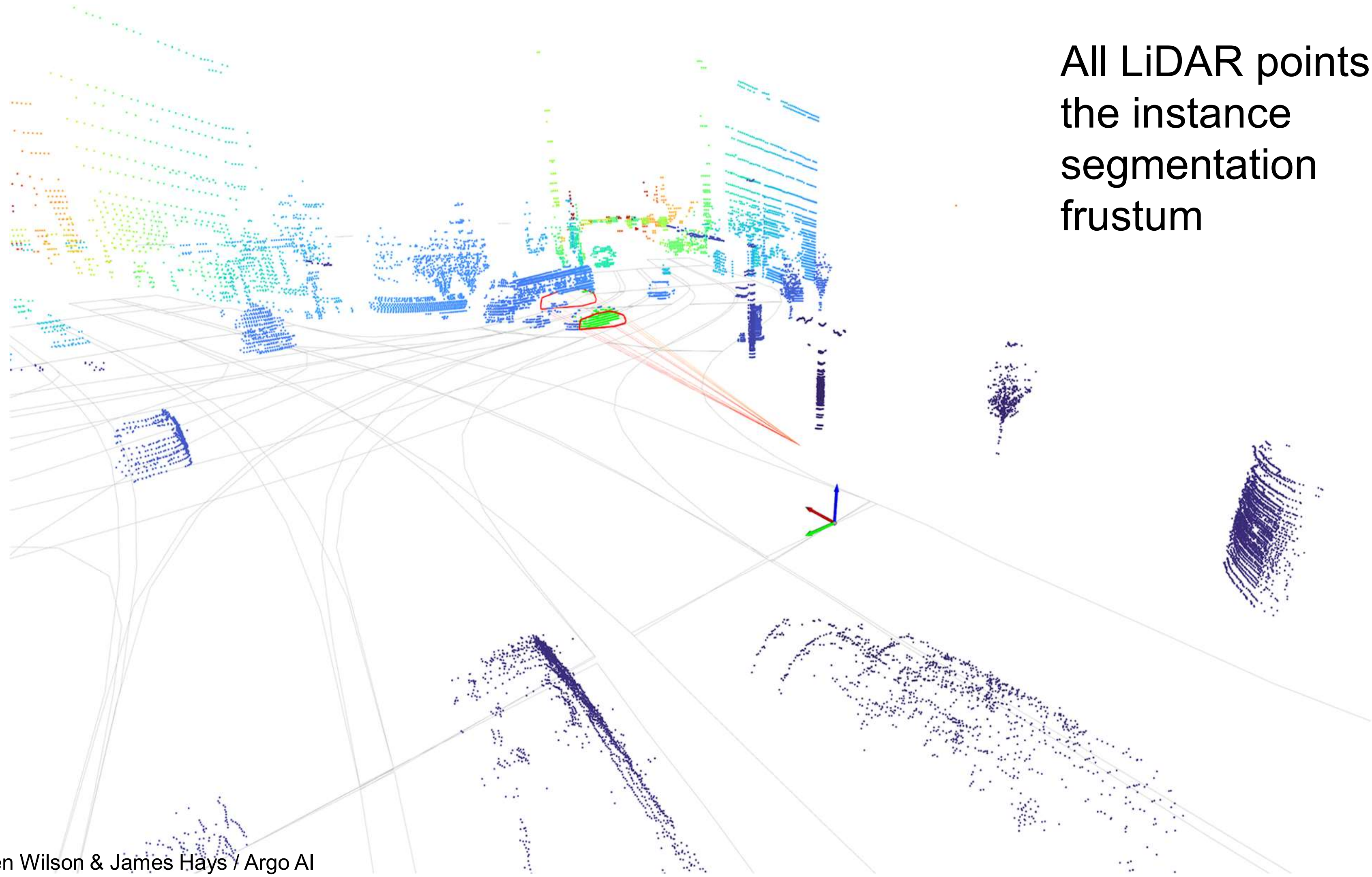
Frustum PointNets for 3D Object Detection from RGB-D Data.
Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, Leonidas J. Guibas. 2017.

Slide Credit: Ben Wilson & James Hays / Argo AI

Slide Credit: Ben Wilson & James Hays / Argo AI

Detections from: Youngwan Lee and Jong
Park. CenterMask: Real-Time Anchor-Free
Instance Segmentation. November 2019.

Slide Credit: Ben Wilson & James Hays / Argo AI

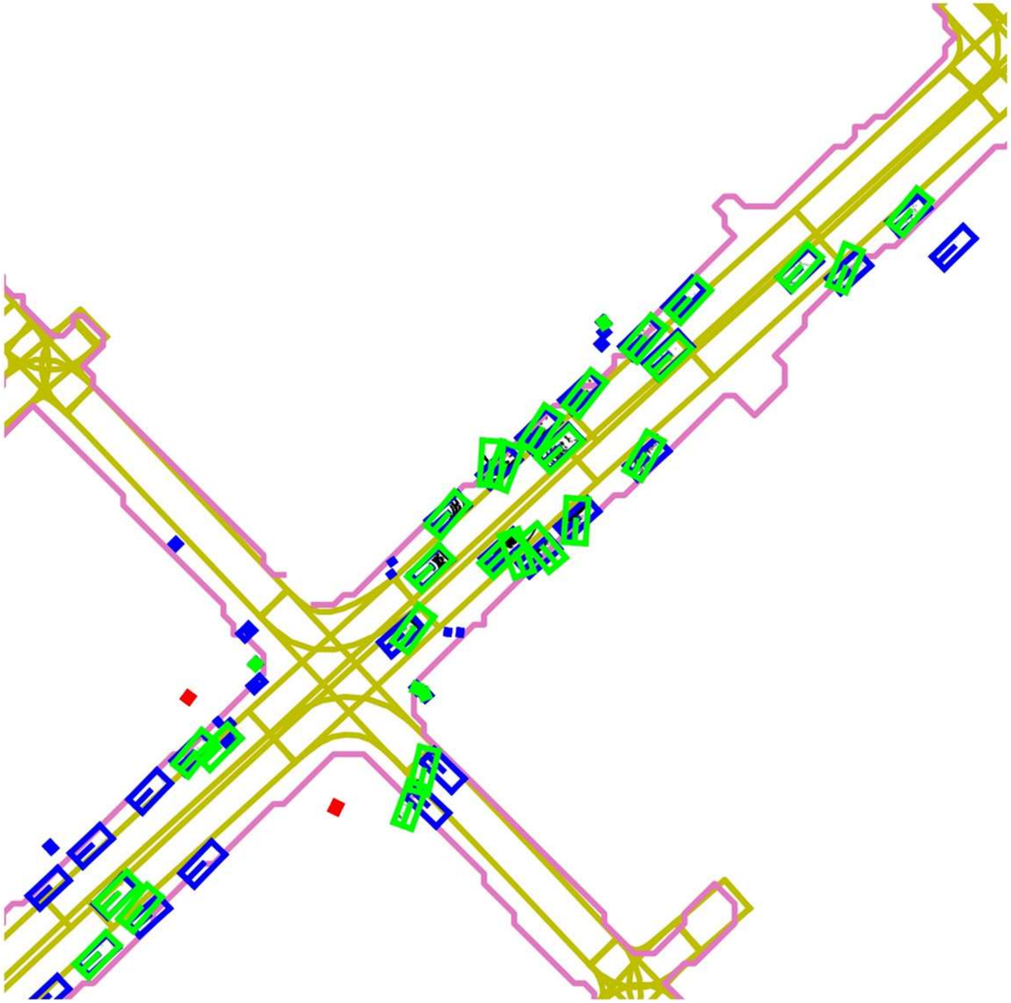


All LiDAR points in the instance segmentation frustum

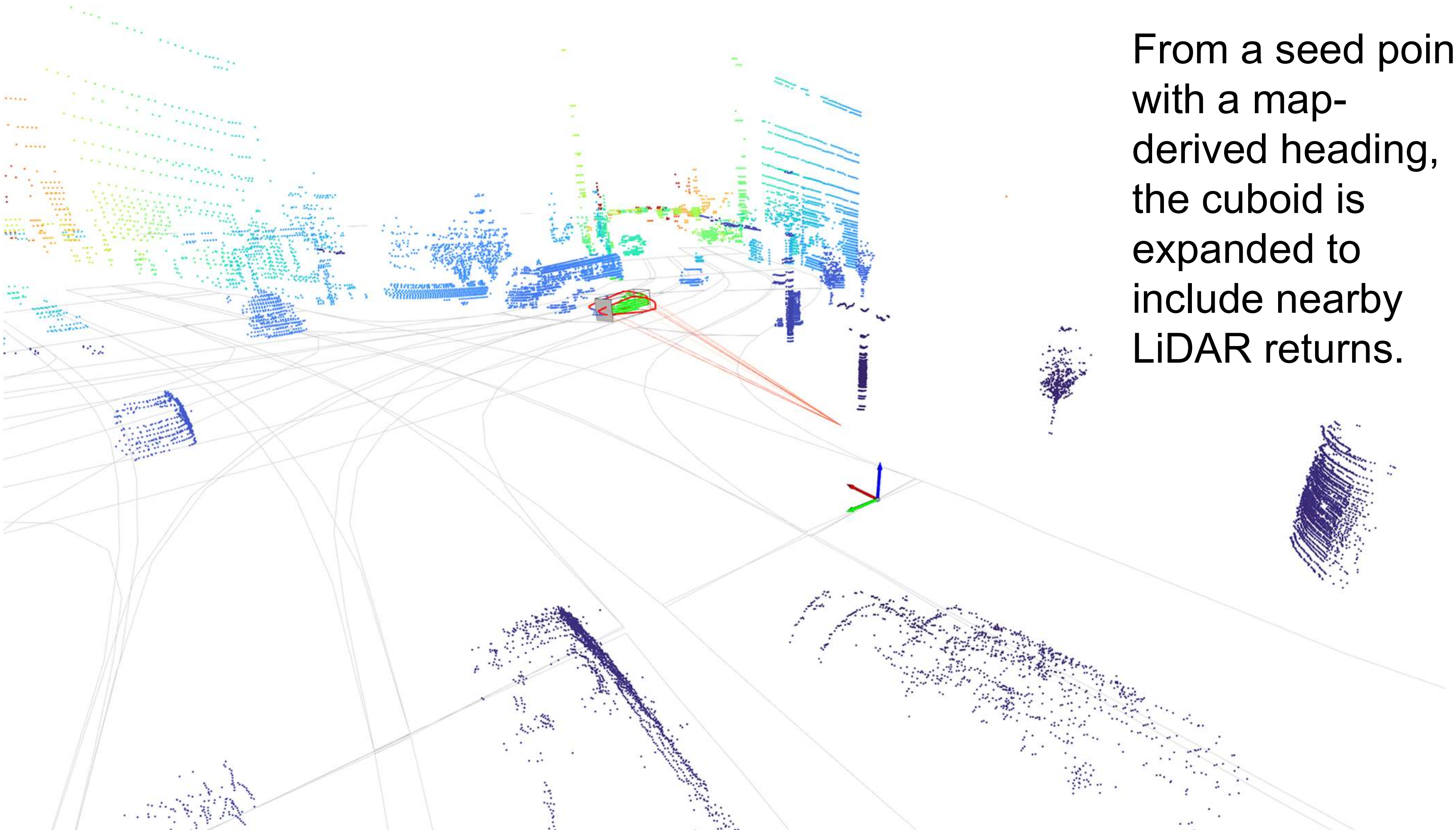
Slide Credit: Ben Wilson & James Hays / Argo AI

Using HD Maps (Centerlines)

- We use *centerlines* to improve orientation estimates

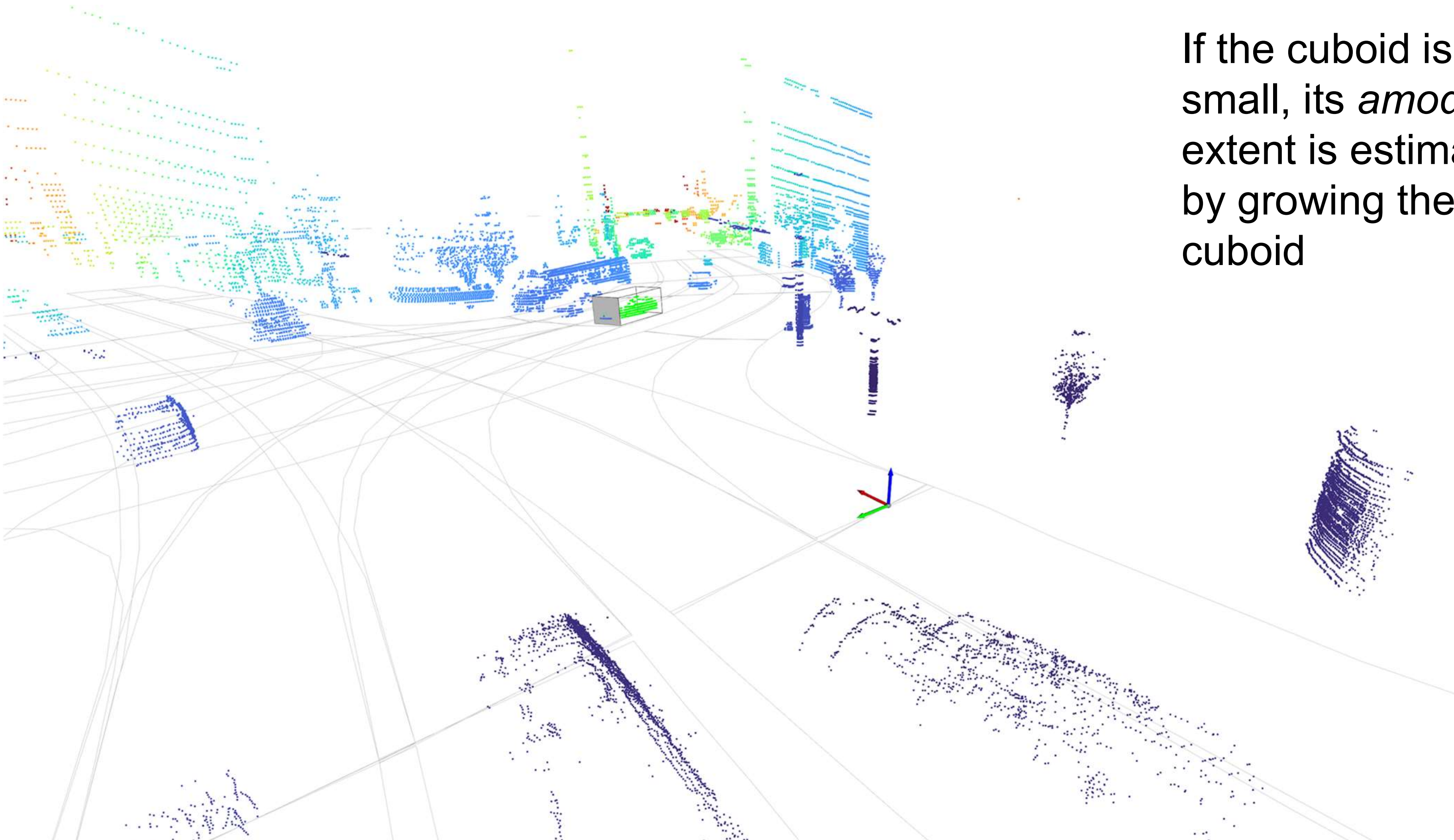


Slide Credit: Ben Wilson & James Hays / Argo AI



From a seed point, with a map-derived heading, the cuboid is expanded to include nearby LiDAR returns.

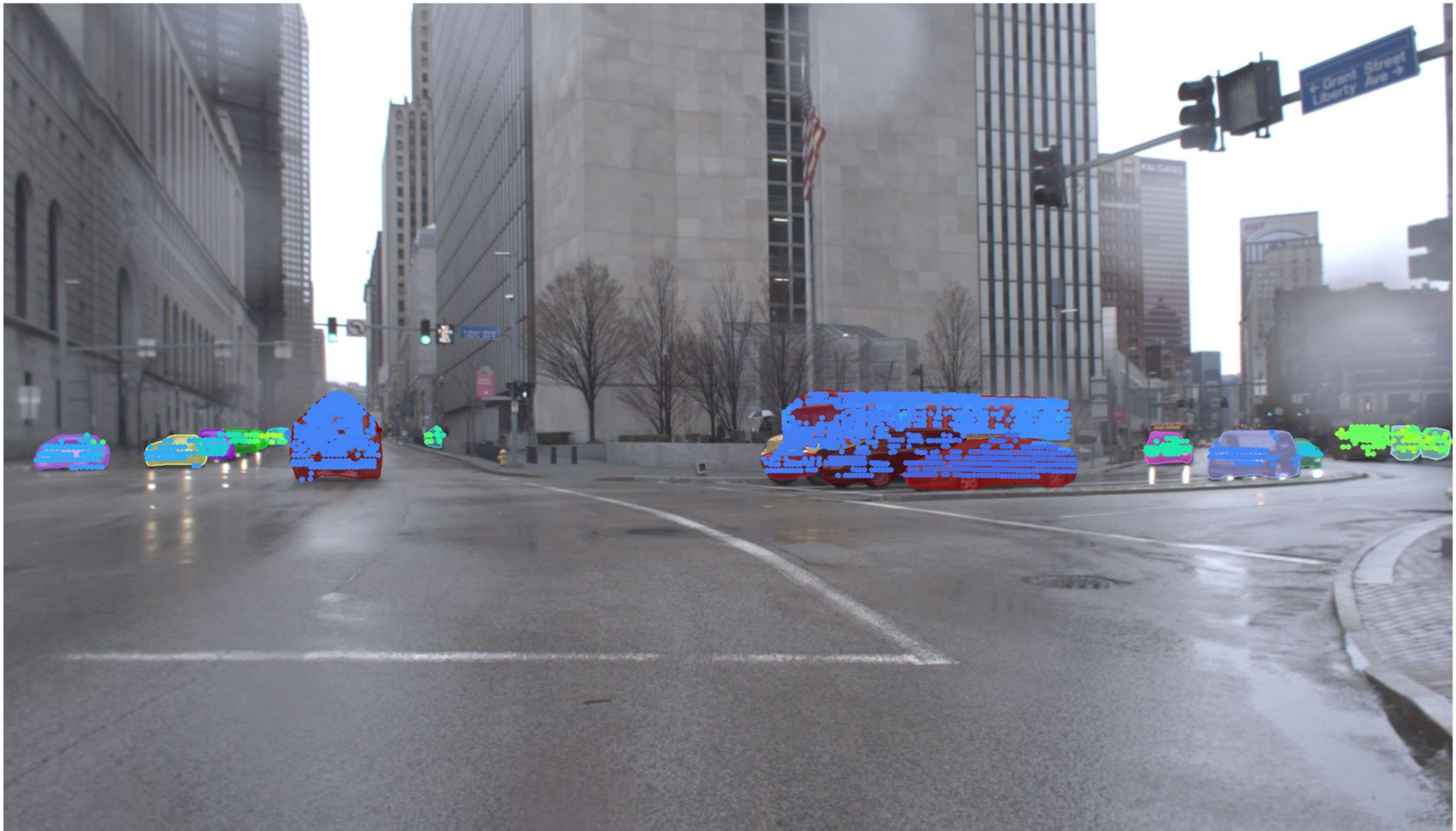
Slide Credit: Ben Wilson & James Hays / Argo AI



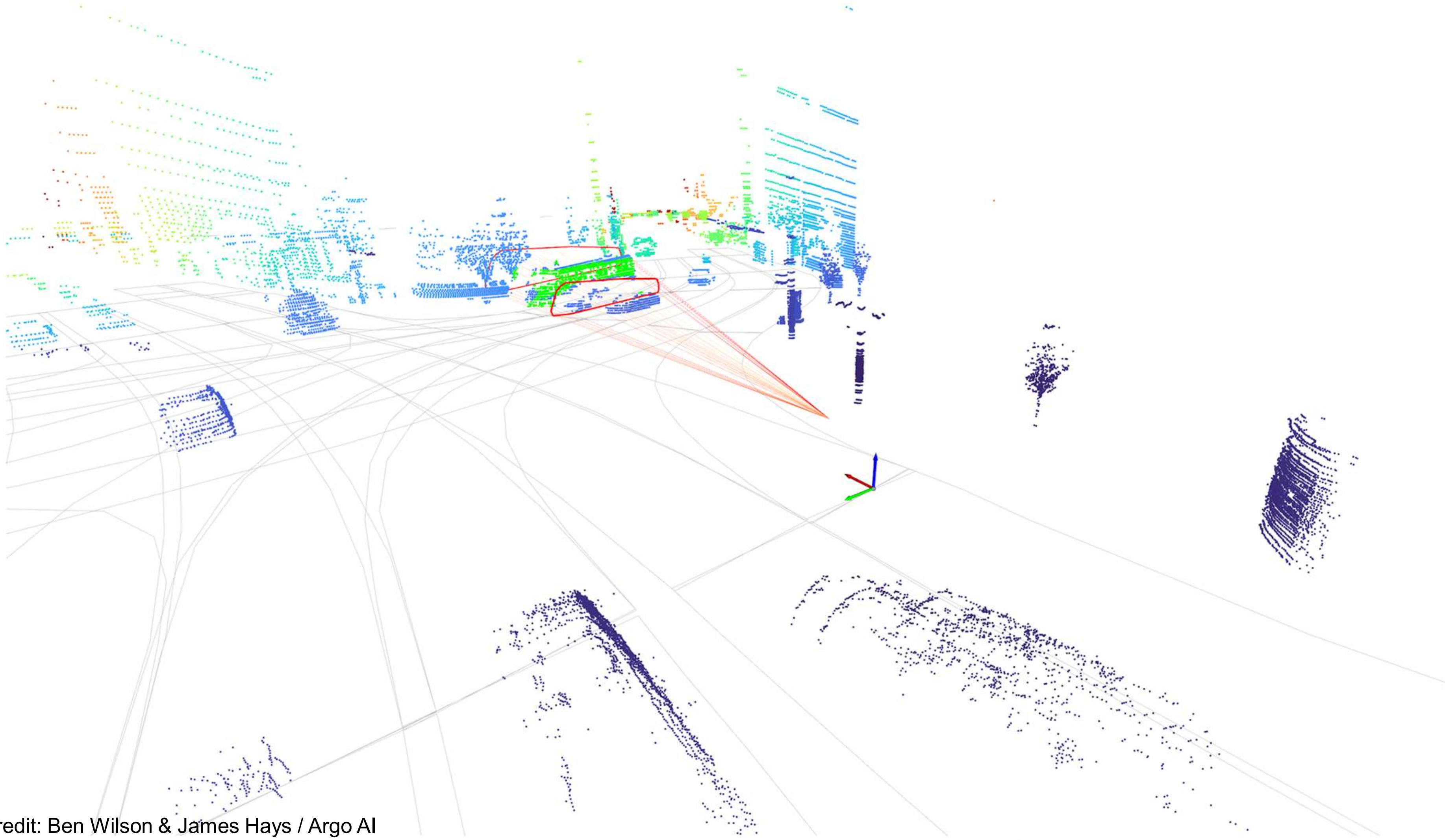
If the cuboid is too small, its *amodal* extent is estimated by growing the cuboid

Slide Credit: Ben Wilson & James Hays / Argo AI

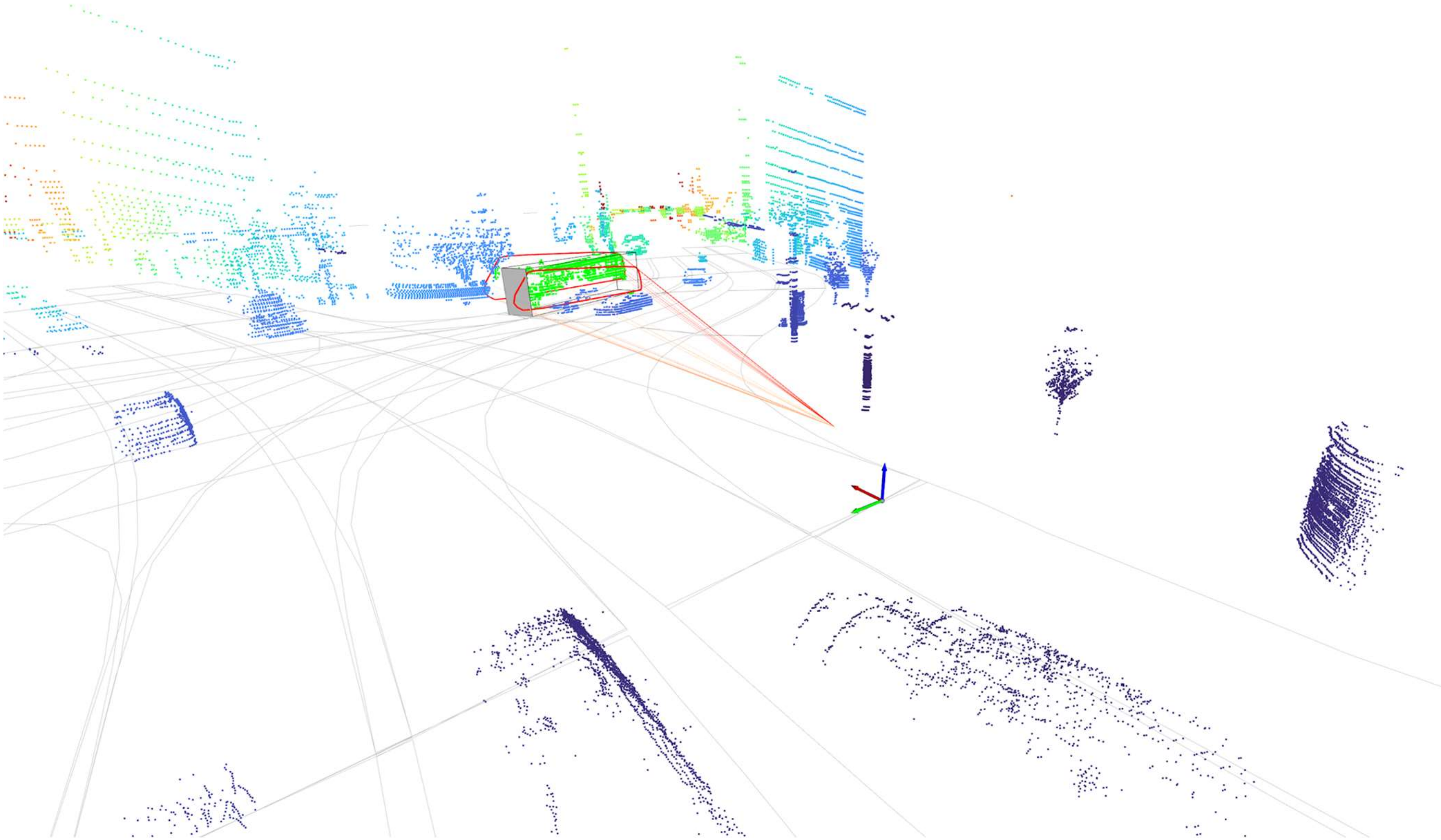
Slide Credit: Ben Wilson & James Hays / Argo AI



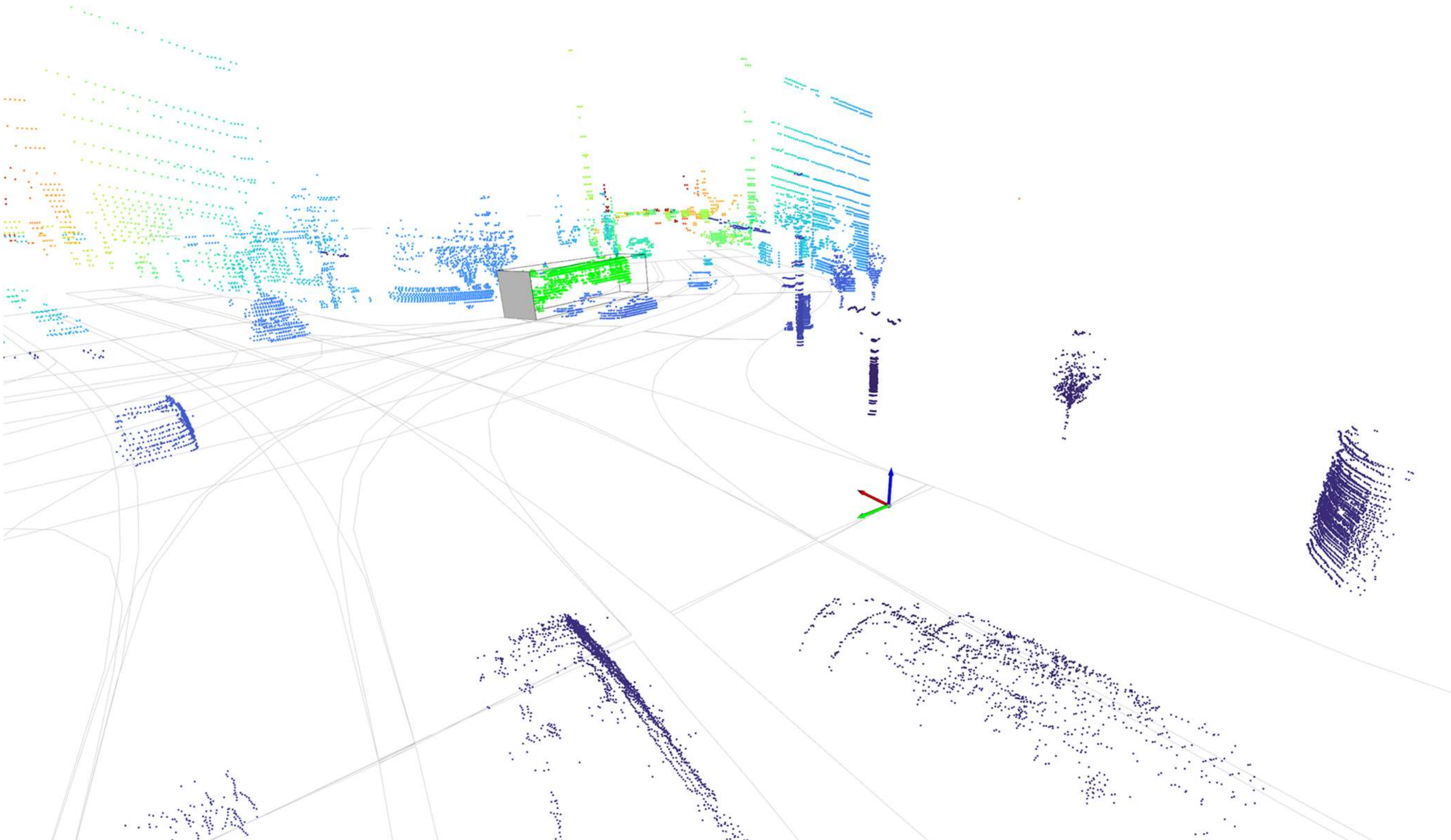
Slide Credit: Ben Wilson & James Hays / Argo AI



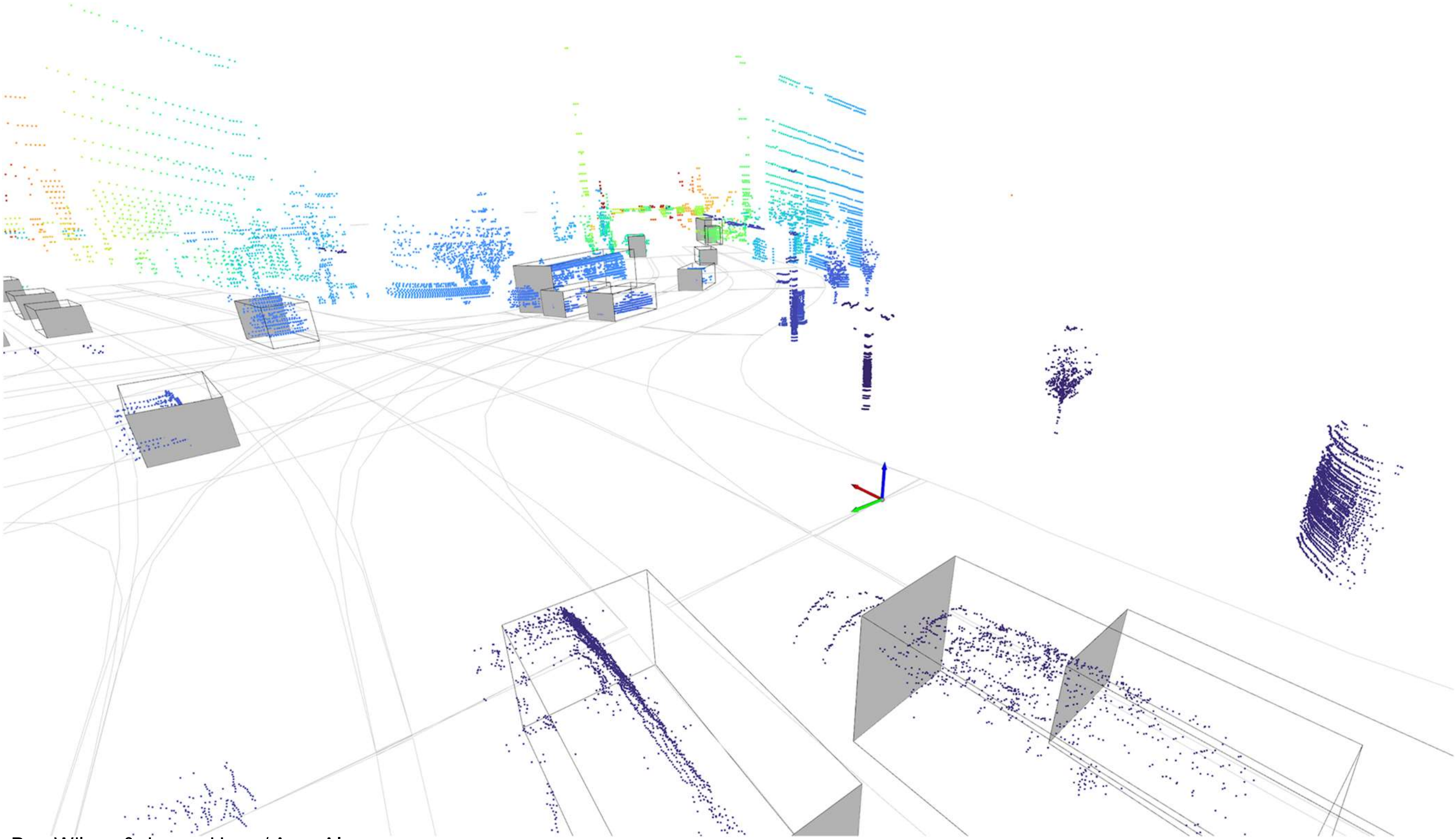
Slide Credit: Ben Wilson & James Hays / Argo AI



Slide Credit: Ben Wilson & James Hays / Argo AI

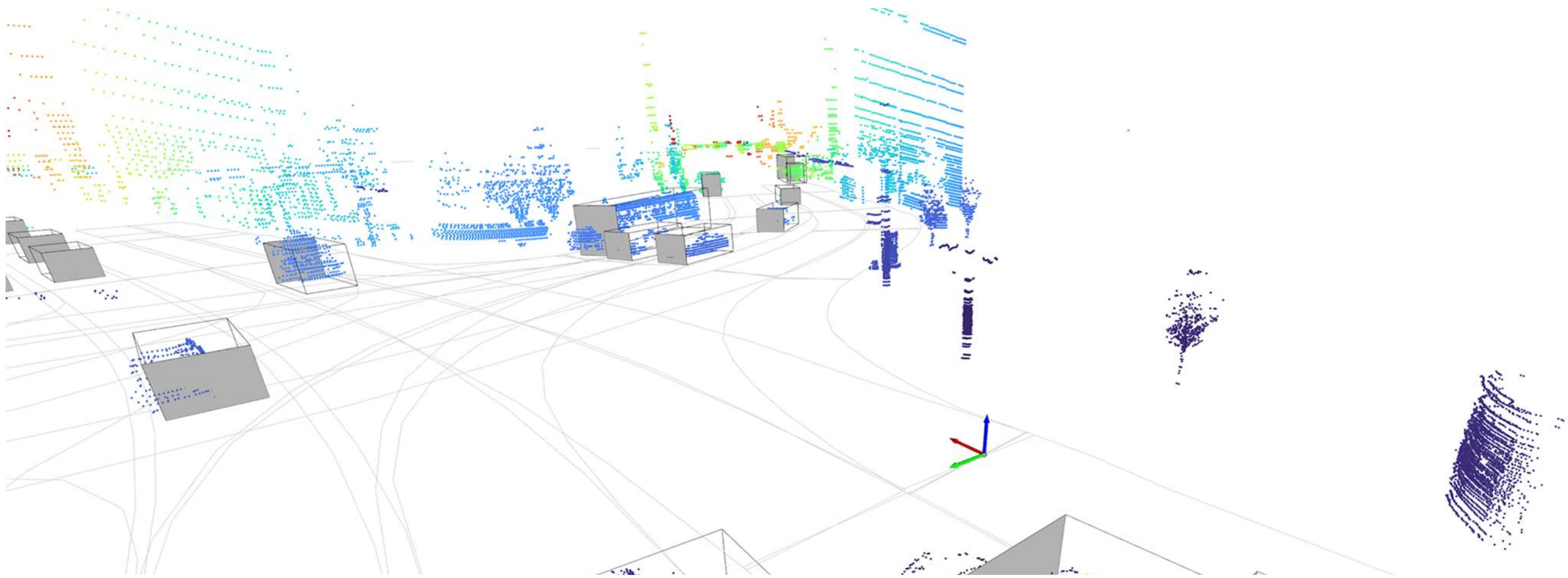


Slide Credit: Ben Wilson & James Hays / Argo AI



Slide Credit: Ben Wilson & James Hays / Argo AI

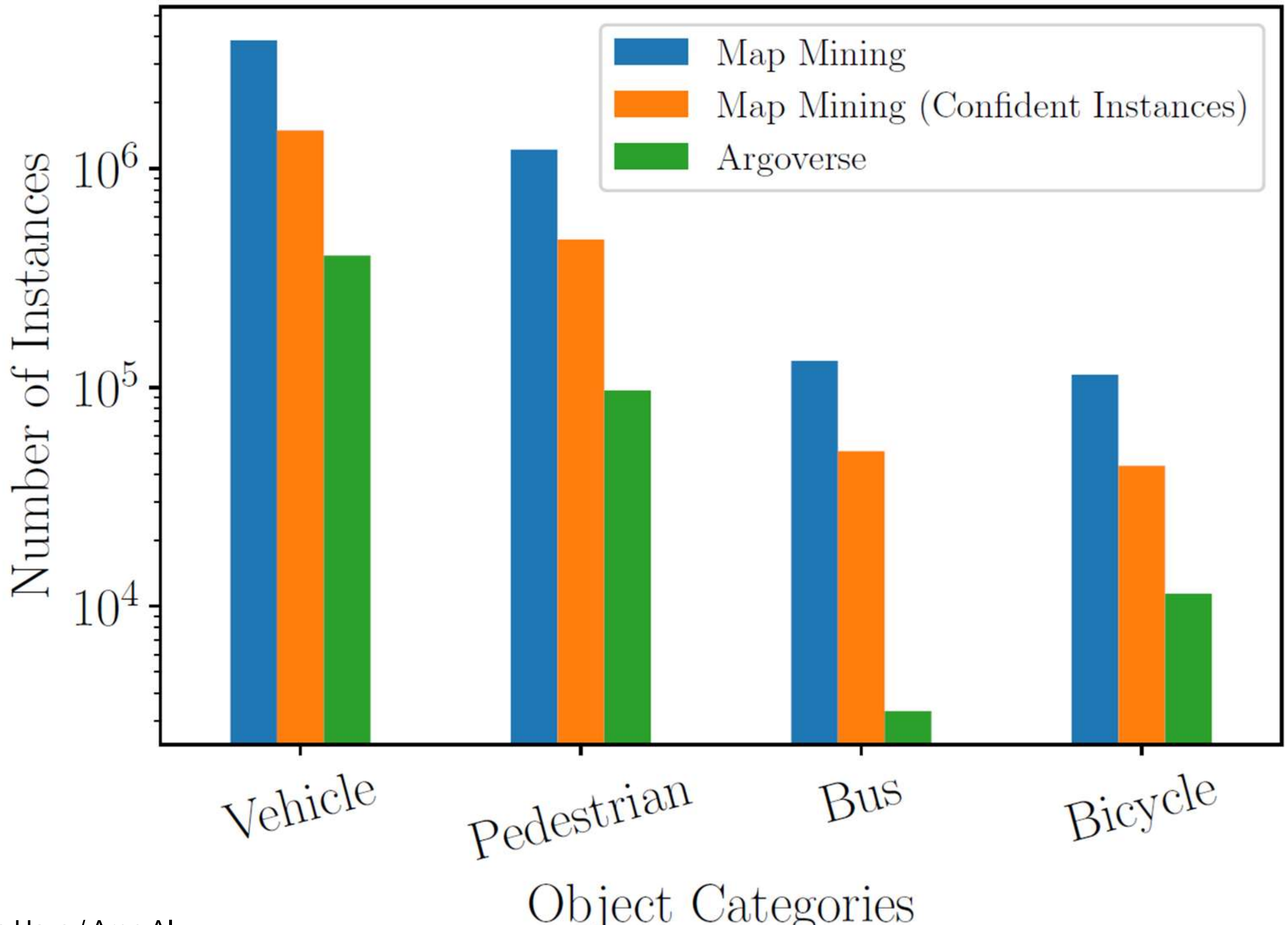
LiDAR baseline is Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced Grouping and Sampling for Point Cloud 3D Object Detection. arXiv, August 2019



Method	Human Annotation	↑ mAP	↓ mATE	↓ mASE	↓ mAOE	↑ CDS
Supervised Baseline	✓	23.76	0.39	**0.22**	0.86	18.48
Inflating with Rotating Calipers		23.62	0.55	0.32	1.68	15.18
Inflating with with Map Mined Data		27.66	0.44	0.30	**0.77**	19.72

Slide Credit: Ben Wilson & James Hays / Argo AI

If we *mine* 1,151 unlabeled vehicle logs...



Slide Credit: Ben Wilson & James Hays / Argo AI

Deep Learning is Robust to Massive Label Noise

David Rolnick [*1] Andreas Veit [*2] Serge Belongie [2] Nir Shavit [3]

Abstract

Deep neural networks trained on large supervised datasets have led to impressive results in image classification and other tasks. However, well-annotated datasets can be time-consuming and expensive to collect, lending increased interest to larger but noisy datasets that are more easily obtained. In this paper, we show that deep neural networks are capable of generalizing from training data for which true labels are massively outnumbered by incorrect labels. We demonstrate remarkably high test performance after training on corrupted data from MNIST, CIFAR, and ImageNet. For example, on MNIST we obtain test accuracy above 90 percent even after each clean training example has been diluted with 100 randomly-labeled examples. Such behavior holds across multiple patterns of label noise, even when erroneous labels are biased towards confusing classes. We show that training in this regime requires a significant but manageable increase in dataset size

Thus, annotation can be expensive and, for tasks requiring expert knowledge, may simply be unattainable at scale.

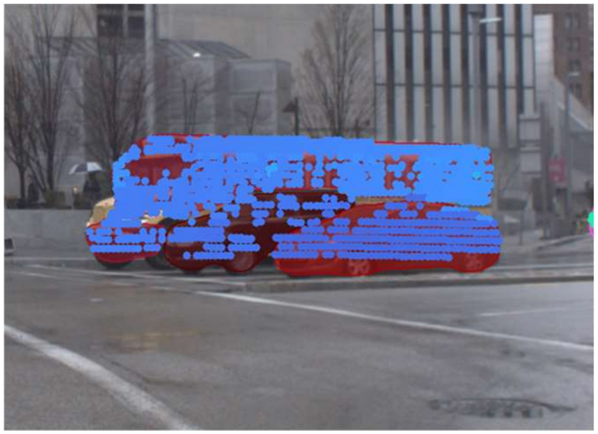
To address this limitation, other training paradigms have been investigated to alleviate the need for expensive annotations, such as unsupervised learning (Le, 2013), self-supervised learning (Pinto et al., 2016; Wang & Gupta, 2015) and learning from noisy annotations (Joulin et al., 2016; Natarajan et al., 2013; Veit et al., 2017). Very large datasets (e.g., Krasin et al. (2016); Thomee et al. (2016)) can often be obtained, for example from web sources, with partial or unreliable annotation. This can allow neural networks to be trained on a much wider variety of tasks or classes and with less manual effort. The good performance obtained from these large, noisy datasets indicates that deep learning approaches can tolerate modest amounts of noise in the training set.

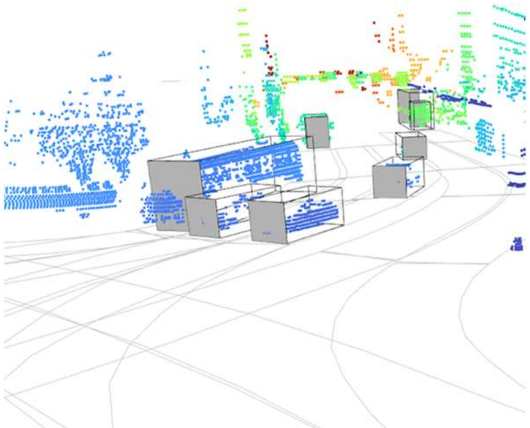
In this work, we study the behavior of deep neural networks under extremely low label reliability, only slightly above chance. The insights from our study can help guide future settings in which arbitrarily large amounts of data are easily

David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. **Deep Learning is Robust to Massive Label Noise**. arXiv, May 2017.

Learning from mined, inflated data

Method	Human Annotation	↑ mAP	↓ mATE	↓ mASE	↓ mAOE	↑ CDS
Supervised Baseline	✓	23.76	0.39	**0.22**	0.86	18.48
Inflating with Rotating Calipers		23.62	0.55	0.32	1.68	15.18
Inflating with with Map Mined Data		27.66	0.44	0.30	**0.77**	19.72
Training with Map Mined Data (ours)		**30.56**	0.39	0.28	1.01	**22.18**





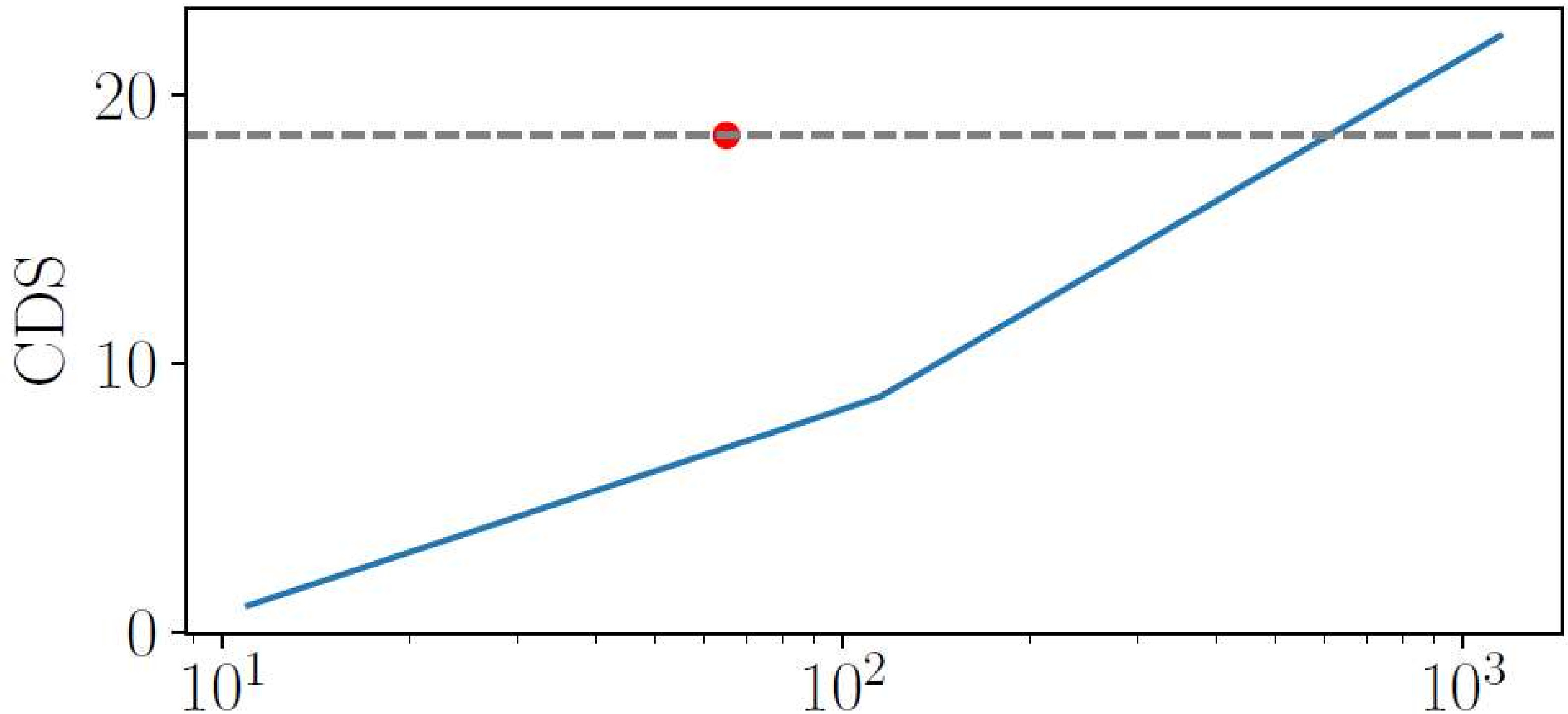
Slide Credit: Ben Wilson & James Hays / Argo AI

Learning from mined, inflated data

Method	Human Annotation	↑ mAP	↓ mATE	↓ mASE	↓ mAOE	↑ CDS
Supervised Baseline	✓	23.76	0.39	**0.22**	0.86	18.48
Inflating with Rotating Calipers		23.62	0.55	0.32	1.68	15.18
Inflating with with Map Mined Data		27.66	0.44	0.30	**0.77**	19.72
Training with Map Mined Data (ours)		**30.56**	0.39	0.28	1.01	**22.18**

Method	Vehicle	Bus	Motorcycle	Bicycle	Pedestrian	↑ mAP	↑ CDS
Supervised Baseline	**58.50**	**12.70**	2.70	2.10	**42.80**	23.76	18.48
Training with Map-Mined Data (ours)	52.40	12.10	**17.00**	**35.50**	35.80	**30.56**	**22.18**

Slide Credit: Ben Wilson & James Hays / Argo AI



CDS = mAP (mATE + mASE + mAOE)

Slide Credit: Ben Wilson & James Hays / Argo AI

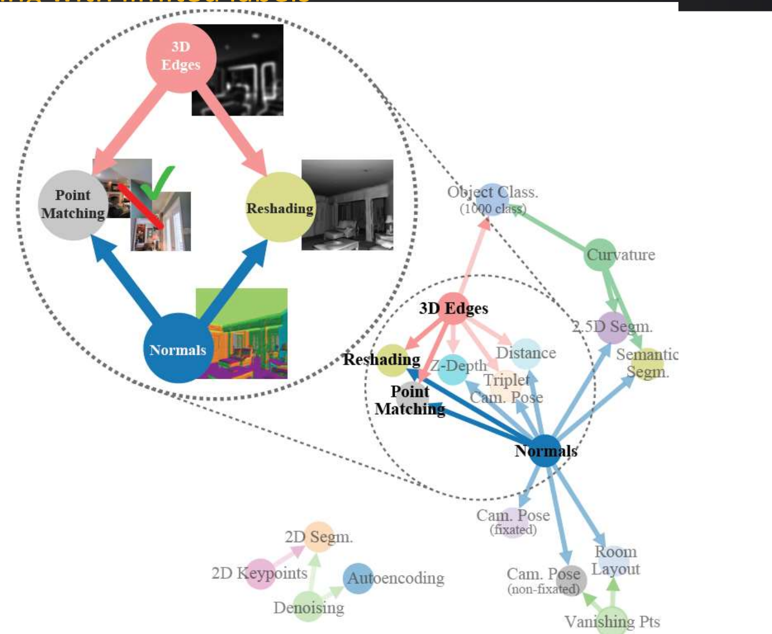
The Methods are Surprisingly Simple

- A handful of common techniques:
  - **Data augmentation**
  - **Pseudo-labeling** / distillation
  - Surrogate tasks / contrastive losses
  - Temperature scaling / Entropy maximization
  - Cosine/metric learning
  - **Prototypes**
  - **Graph neural networks**
  - Meta-learning

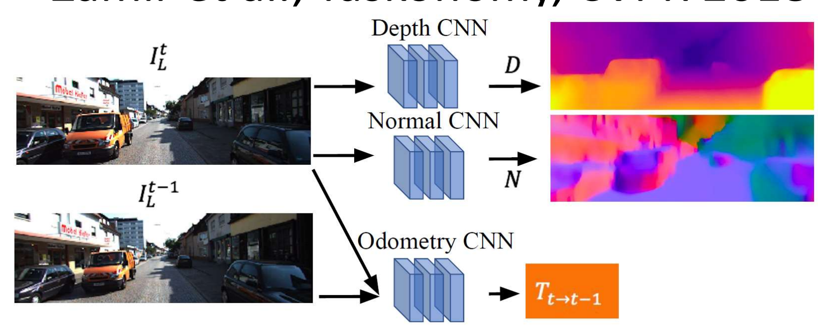
- Autonomous vehicles: **Multi-modal and multi-task!**

Conclusions

- Amazing gains have been made across learning with limited labels
- **Data augmentation** a crucial aspect; we develop methods for:
  - **Complex feature-space augmentation** using graphs and leveraging manifold structure
- Move **beyond image classification** for autonomous vehicles
  - Object Detection
  - 3D
  - Leverage large amount of unlabeled data though *many tasks*

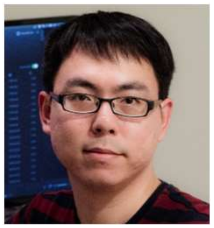


Zamir et al., Taskonomy, CVPR 2018



Zhan et al., Self-supervised Learning for Single View Depth and Surface Normal Estimation, CVPR 2018

Acknowledgement and Questions?



Chia-Wen Kuo

Ben Wilson

Chih-Yao Ma

Yen-Chang Hsu

Yen-Cheng Liu

RESEARCH AMERICA