Dataset Culling: Towards Efficient Training of Distillation-based Domain Specific Models

K.Yoshioka[1][2], E. Lee[2], S. Wong[2], M. Horowitz[2]
(1) Toshiba
(2) Stanford University

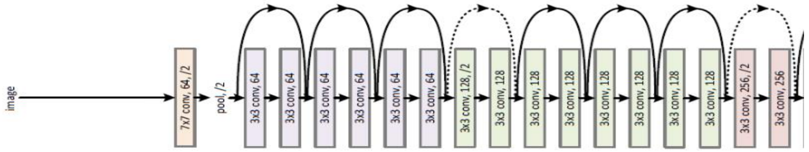
IEEE ICIP 2019 Sept. 25

Introduction

- Deep Learning based object detection has excellent accuracy.
  - e.g. Vision for security, infrastructure, transportation..
- Cost?
  - Requires many GPU-hours, difficult to scale.
  - Has accuracy-cost tradeoff.

101-layer Resnet:
Imagenet accuracy 78%

10-layer Resnet:
Imagenet accuracy 60%

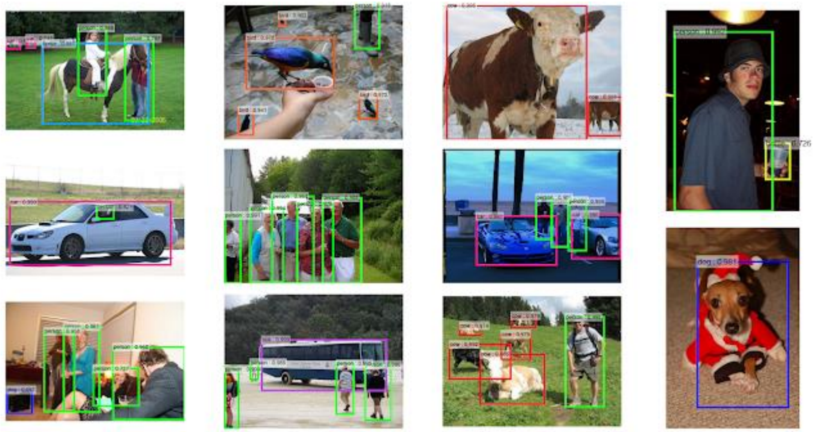


- How can we break this tradeoff?

Introduction: Domain Specific Models

- Training compact domain specific models (DSMs) [1,2]
    - DSMs: a specialized model for specific env.
      {conference room, your house, your office, etc.}
- Cuts down computation cost <u>5-20x</u>

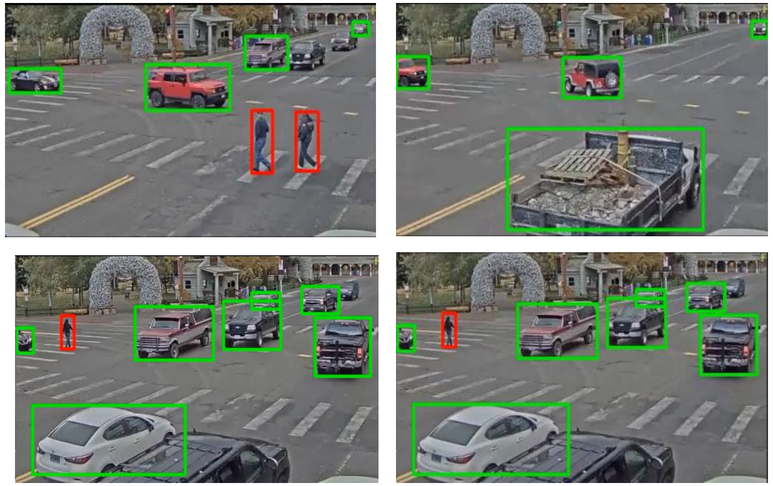
General dataset



Images from MS-COCO(http://cocodataset.org/)

⇔

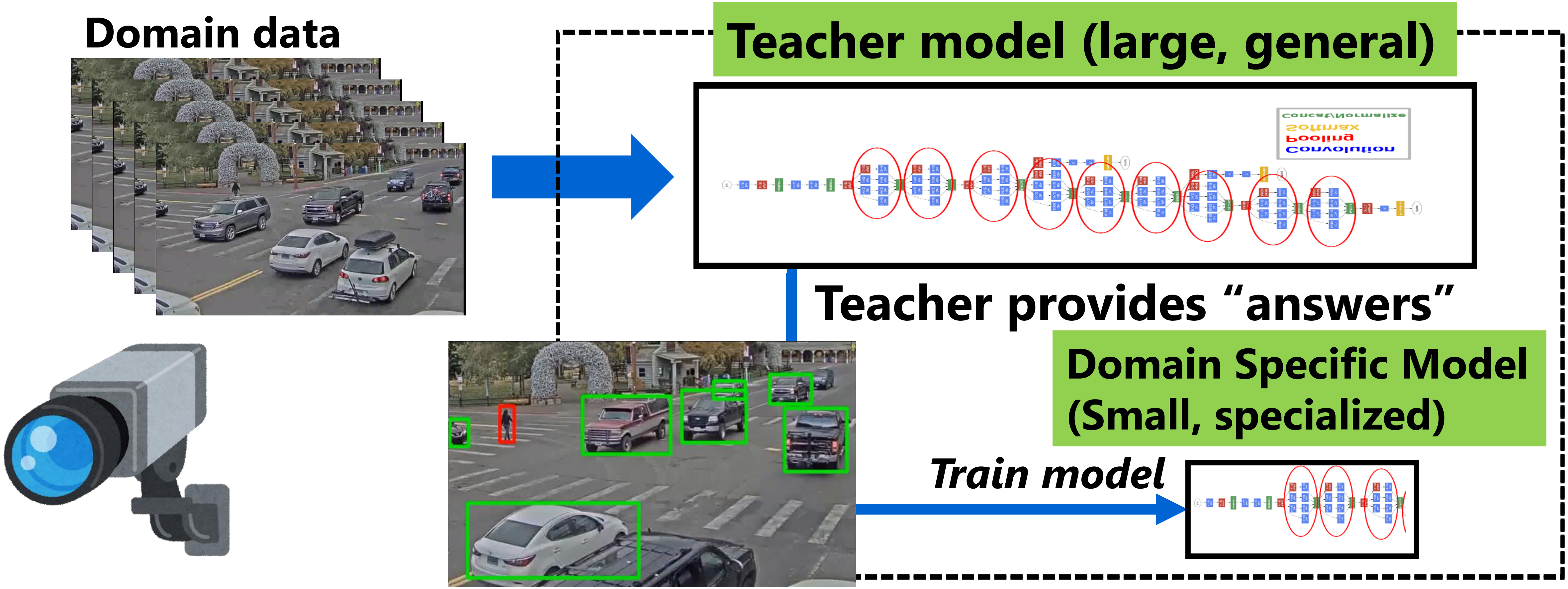
Surveillance cam. data



[1]D. Kang, “Noscope: optimizing neural network queries over video at scale,”
[2]R.Mullapudi “Online model distillation for efficient video inference,”

Introduction: What is Distillation?

- **Teacher model teaches the small student model to learn**
    - **Works without human interference**



Introduction: The Problem

- Can gather lots of training data easily..
  - A day's worth of surveillance data
  =86,400 images @ 1FPS

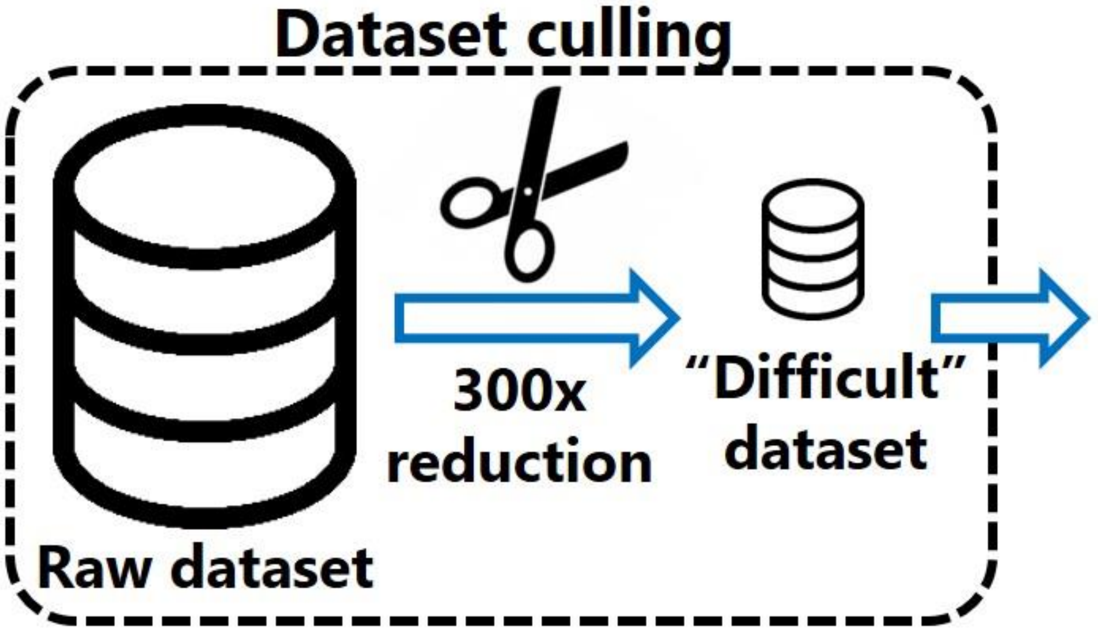
- Training 86,400 images require over 100 GPU-hours (Nvidia K80 on AWS) to train.
  - Unable to scale to deploying thousands of cameras

- Reducing the DSM training cost has not been explored.

Dataset Culling

Basic Idea of Dataset Culling

- Reduces the dataset size <u>300x</u>
    - Culls only <u>"Easy"</u> data; model accuracy is not harmed



Total training time:
104 → 2.2 GPU-hours

47x improvement ☺

What is good training data?

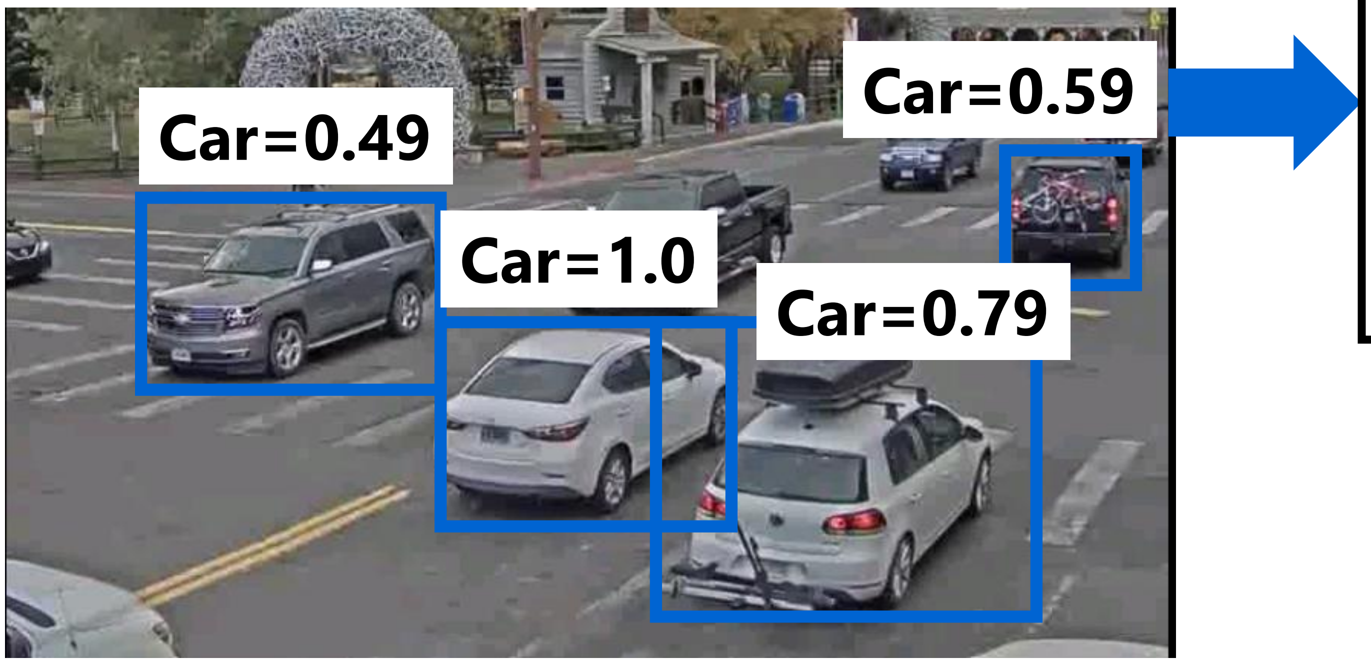
- <u>"Difficult" data</u> which the model makes a lot of mistakes.
    - No backprop is done if the model can perfectly predict.

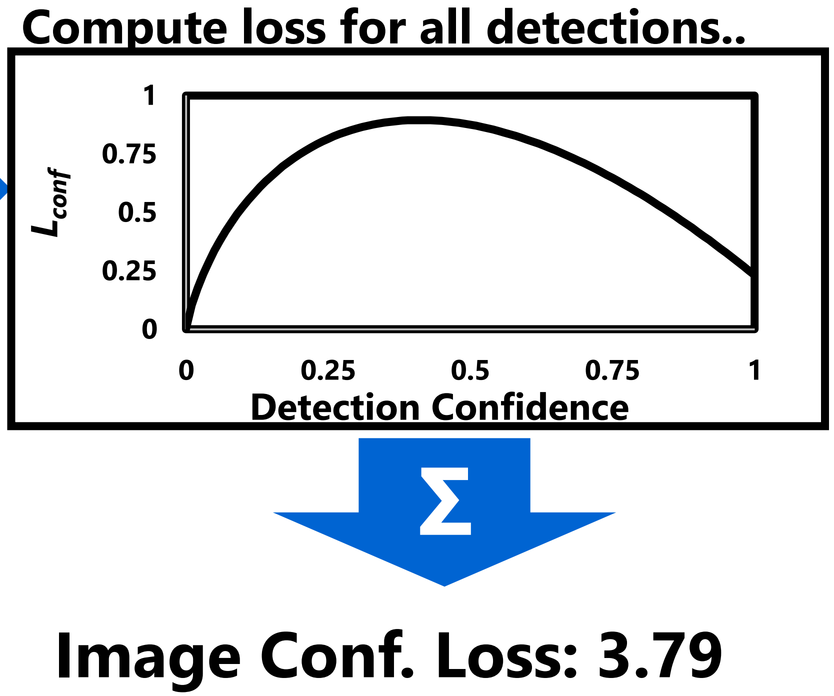
      → Does not contribute to training.
    - Comparing teacher-student predictions are costly..

- **Can we assess from student predictions only?**

How can we pick good training data?

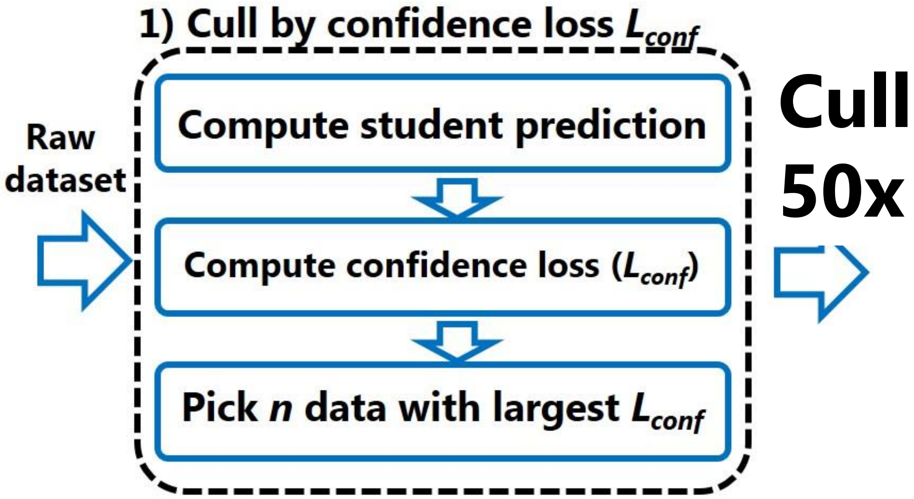
- **Quantify good data by proposed "confidence loss"**
  - **Assesses the difficulty of prediction from the output probability**





Dataset culling pipeline

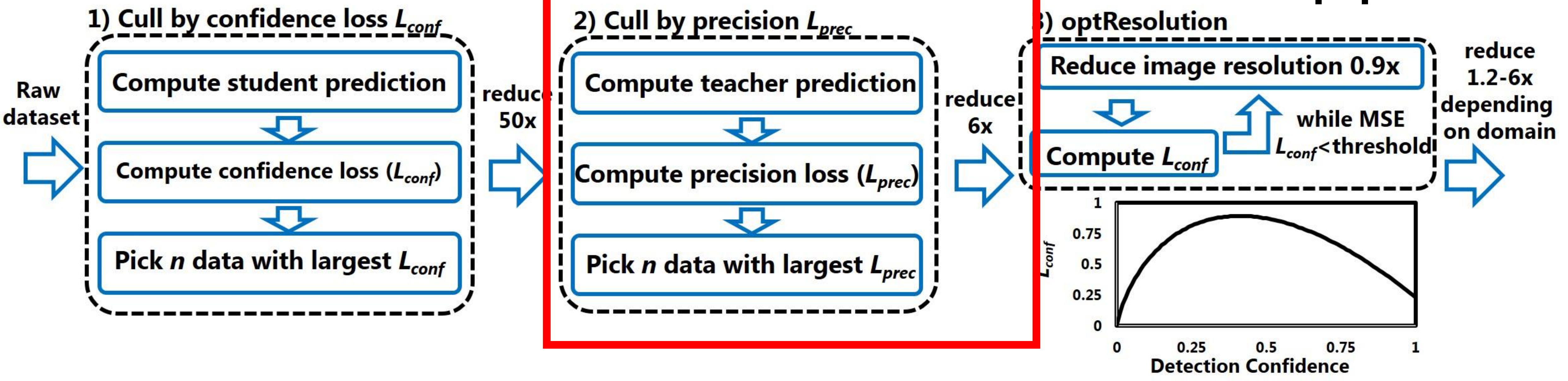
- First, cull dataset using <u>only the student model</u>
  - Culls out majority of the data first (50x).
  - Cheap; does not require costly teacher inference.



Dataset culling pipeline

- **Then, conduct a secondary culling using both <u>teacher-student predictions</u>.**
  - **Directly determine errors the student makes.**
  - **Data is culled up to <u>300x</u> by the pipeline.**

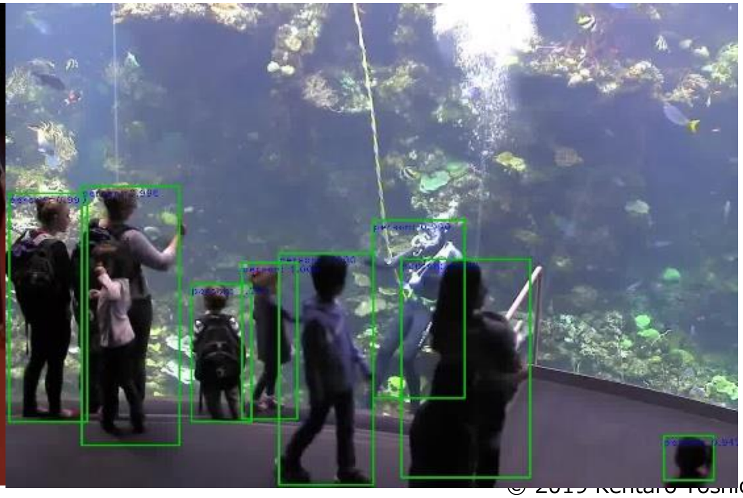
Details in paper



Experiments

Experiment setups

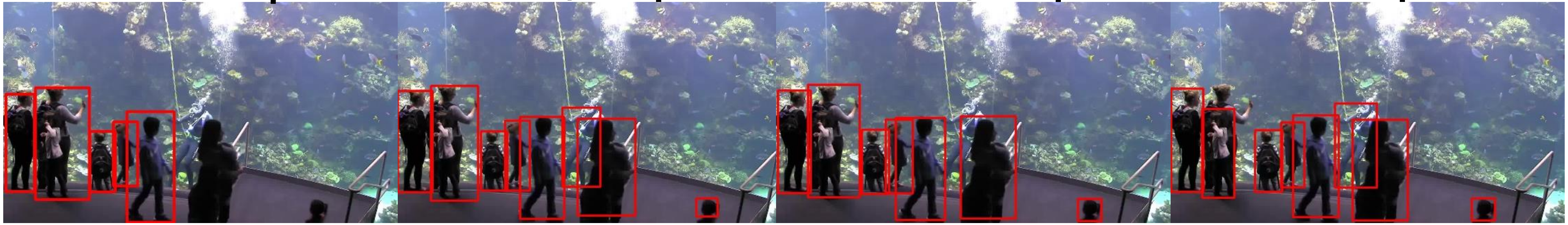
- **Models pretrained on MS-COCO:**
  - **Student: Resnet-18 based Faster-RCNN**
  - **Teacher: Resnet-101 based Faster-RCNN**
- **Dataset: 8 custom videos acquired from Youtube.**
  - **Train: first 24-hours**
  - **Validation: Subsequent 6-hours**
  - **Utilize teacher output as ground-truths**



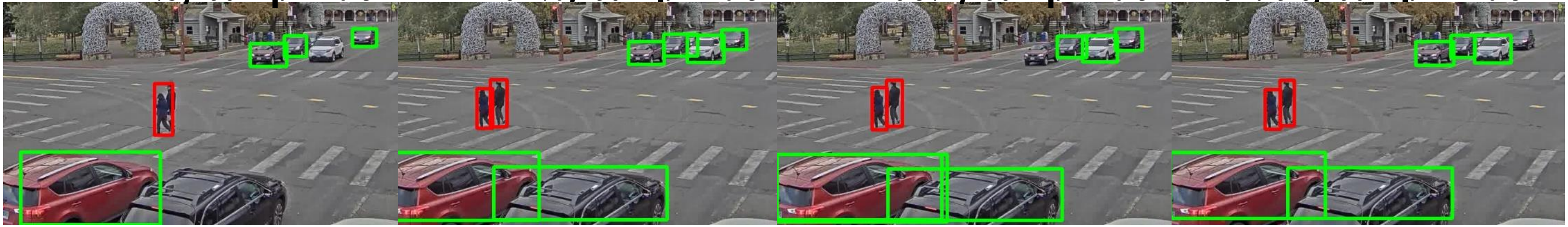
Qualitative results

RawStudent | *TrainStudent* | *TrainStudent+optResolution* | *Teacher*

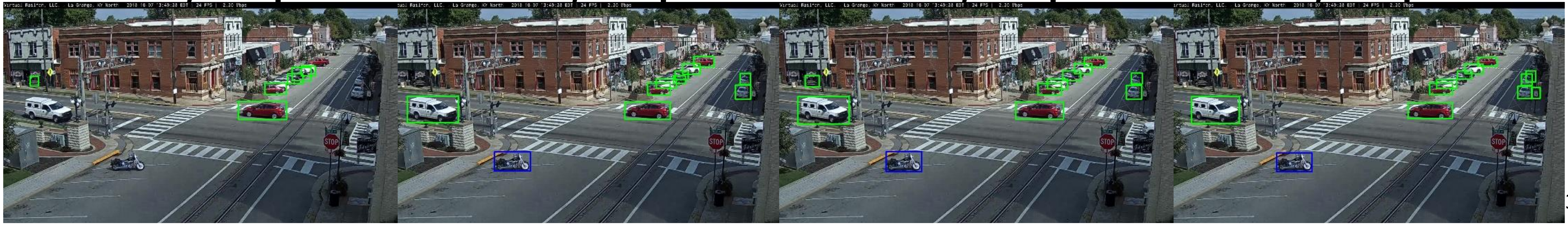
mAP=78.2, comp=28G | **mAP=90.2, comp=28G** | **mAP=89.6, comp=7G** | **Oracle, comp=128G**



mAP=71.3, comp=28G | **mAP=94.8, comp=28G** | **mAP=93.2, comp=18G** | **Oracle, comp=128G**



mAP=52.7, comp=28G | **mAP=81.6, comp=28G** | **mAP=80.7, comp=18G** | **Oracle, comp=128G**



Quantitative Results

- Can cull the dataset size to 300x, without accuracy drops or even with improvements.

Culled dataset size

	64	128	256	Full (86,400)	No Training
Mean Accuracy [mAP]	85.56 (-3.0%)	88.3 (-0.3%)	**89.3 (+0.8%)**	**88.5**	**58.6**
Total train time [hours]	1.9 (54x)	2.0 (50x)	**2.2 (47x)**	**104**	**-**
Student Prediction [hours]	1.54	1.54	1.54	0	-

Conclusions

- While DSMs can reduce the inference cost, training them can take many GPU-hours.

- We proposed Dataset Culling, which reduces the DSM training cost by 47x.
  - Only easy to predict data are culled to minimize the accuracy drop.
  - Evaluated on our long-duration dataset, we saw little to no accuracy penalty even with culling.

Codes and dataset available:
https://github.com/kentaroy47/DatasetCulling

Ablation study

- **Entropy implements the loss function for active learning.**
- **Using teacher-student comparisons achieve best accuracy (Precision)**
- **Our dataset culling pipeline with Confidence + Precision has the best tradeoff of accuracy and training time.**

Filtering strategy	Intermittent Samp.	Entropy[9]	Confidence	Precision	**Confidence + Precision**	Full dataset
mAP	0.731	0.866	0.911	0.954	**0.948**	0.958
GPU hours	0.15	1.7	1.7	8.0	**2.0**	104