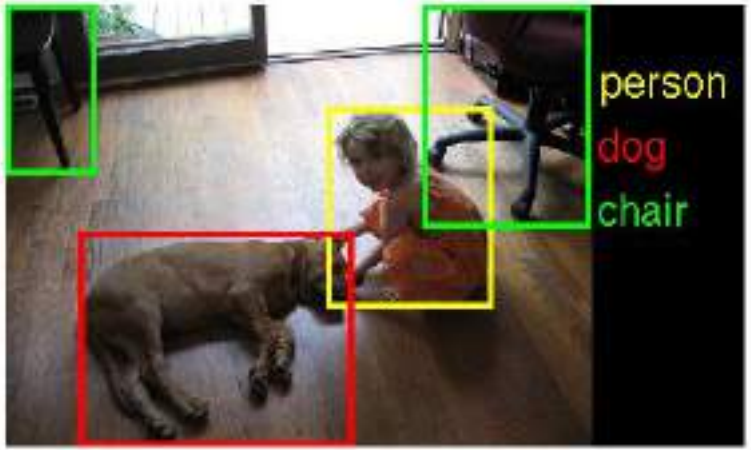
# Computer Vision for Embedded Systems

Yung-Hsiang Lu
Purdue University
yunglu@purdue.edu

PURDUE UNIVERSITY

# Bottleneck in Supervised Learning

- Labeling data is time-consuming and expensive
- Computer vision (and many other machine learning tasks) is not perfect and needs "teachers" to provide correct answers.
- Labeling requires human effort, slow and expensive
- Acquiring "rare" events is difficult (or impossible)



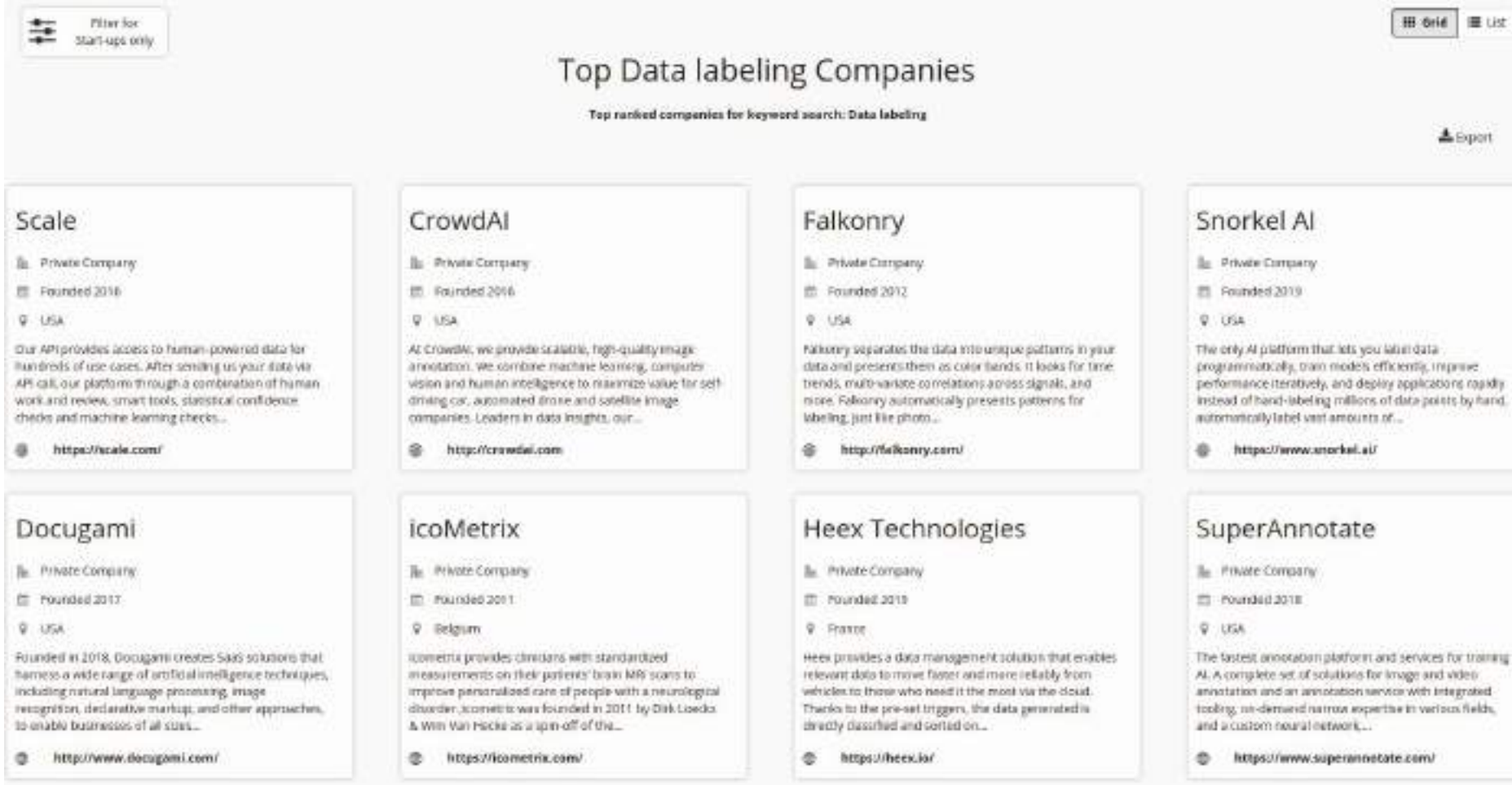

ImageNet

DAVIS: Densely Annotated VIdeo Segmentation

DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort (CVPR 2021)

"Labeling a complex scene with 50 objects can take anywhere between 30 to 90 minutes"
(for semantic segmentation)

30 frames / second x 60 second/minute x 90 minutes = 162,000




https://www.ventureradar.com/keyword/Data%20labeling


## Medical Diagnostics

Artificial intelligence is transforming healthcare by allowing practitioners to use big data to identify and treat diseases.


## Retail Automation

As customer expectations evolve, companies are turning to AI to make the retail experience more convenient and customized.


## Precision Agriculture

Technology holds great promise for solving the many challenges and inefficiencies in the production and distribution of food.

https://info.cloudfactory.com/



https://www.v7labs.com/pricing

# Crowdsourcing Annotations for Visual Object Detection (Conference on Artificial Intelligence 2012)

## Desired outcomes: objects' bounding boxes in images

- Quality: tight bounding boxes
- Coverage: every object is labeled (for positive and negative examples)

## Challenge: how to know the labels are correct and high quality?

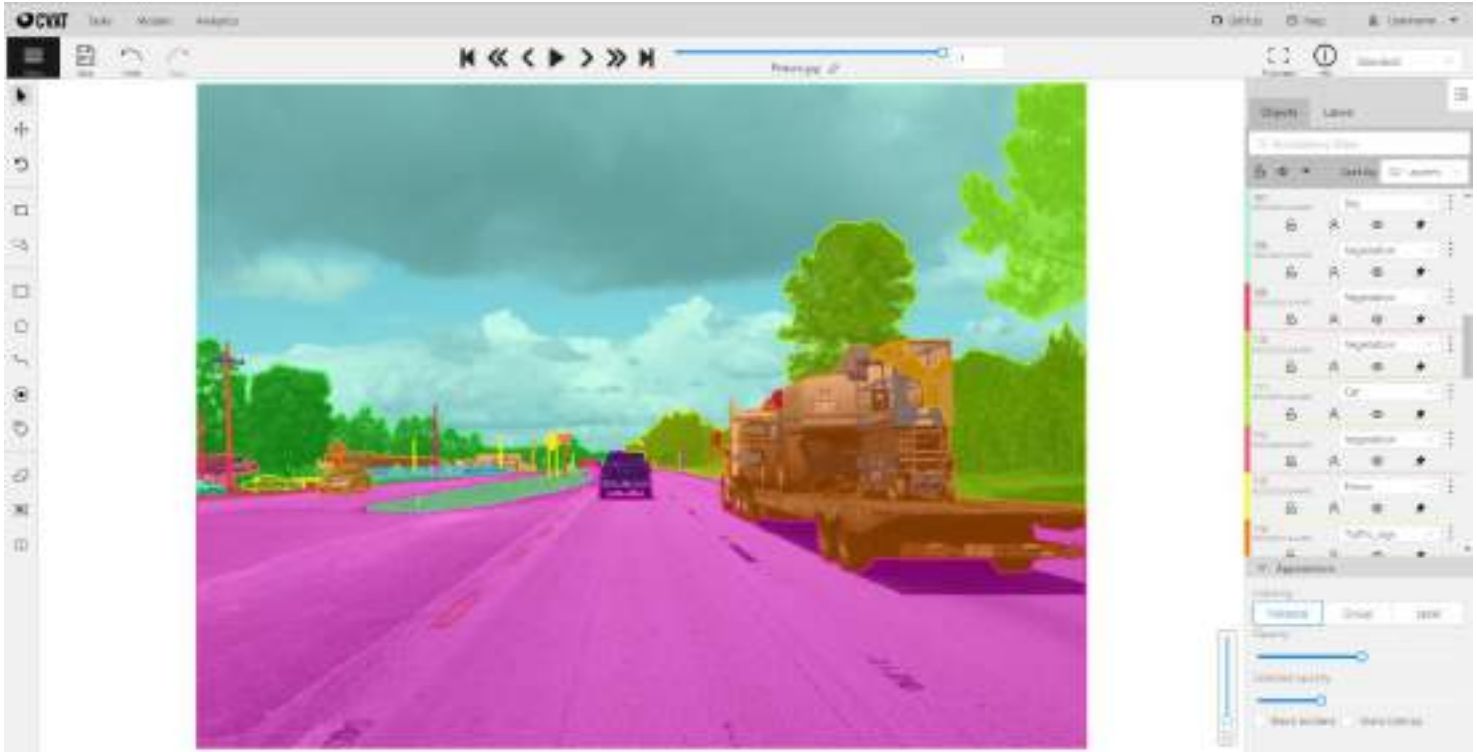
- How to obtain trustable results from crowds?
- What is the right incentive to the crowds?
- "chicken-egg" problem:
  - no labels and no truth
  - no truth and cannot verify
  - cannot verify and no trusted label

"bottle" category

# Crowdsourcing for Labeling

# Labeling Tools

## Computer Vision Annotation Tool (CVAT)



Yung-Hsiang Lu, Purdue University

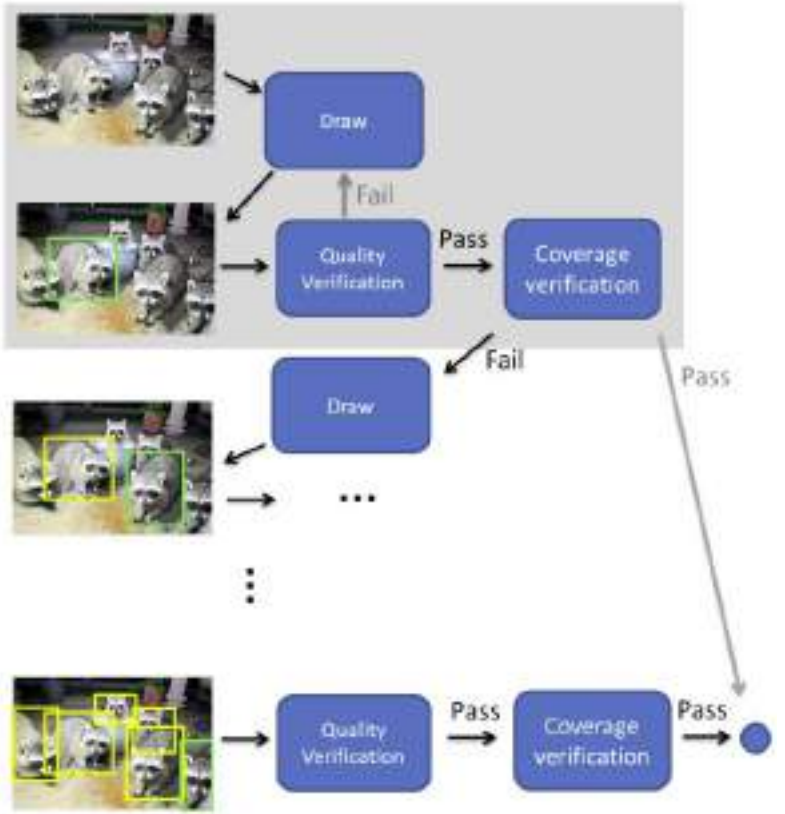

Objects
Labels
Stop Creation
Merge Shapes
Group Shapes
Open Menu
Filter:
Reset
Propagation
Relations
Frame

# Draw-Verification Procedure

30 second/image x 1M image x 3 people
= 25,000 hours
= 1,042 days

If person x 2 hours/day
⇒ 520 days

Each person draws only one bounding box



# *Procedure for Crowdsourced Labeling*

**Drawing Task:**

1. Include all visible parts and draw as tightly as possible.
2. If there are multiple instances, include only ONE (any one ).
3. Draw on a new instance if an instance has a bounding box.
4. If every instance already has a bounding box, check the check box.

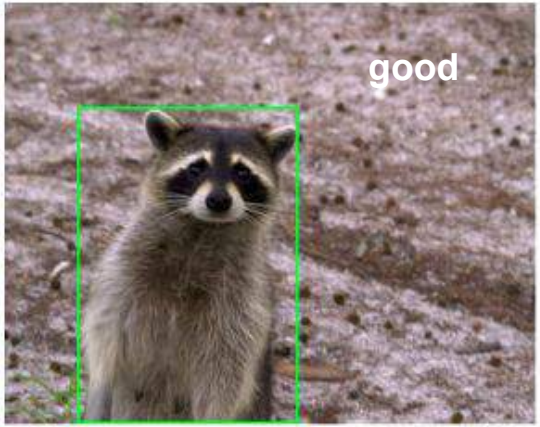
**Verification Task:**

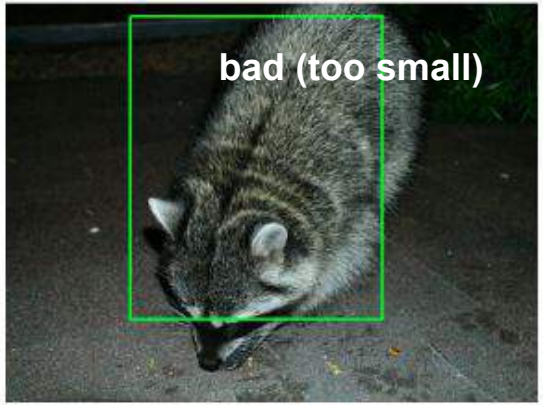
1. A bounding box must include an instance of the required object.
2. A bounding box must include all visible parts and be as tight as possible.
3. If there are multiple instances, a good bounding box must include only ONE ( any one )

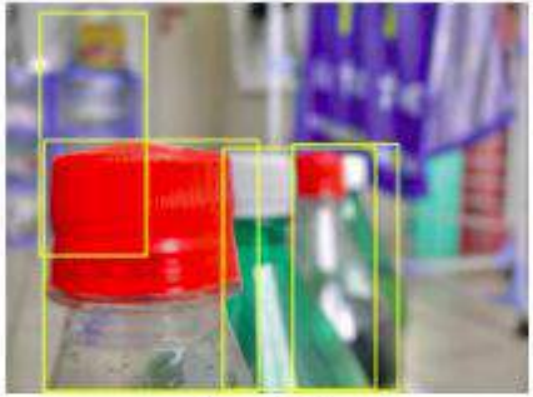
Before given a real labeling task, a participant must pass a test.

# Experiments

- 20,000 categories
- 14 million images
- 97.9% correct bounding boxes
- common errors: bounding boxes too small
- 88 second per bounding box



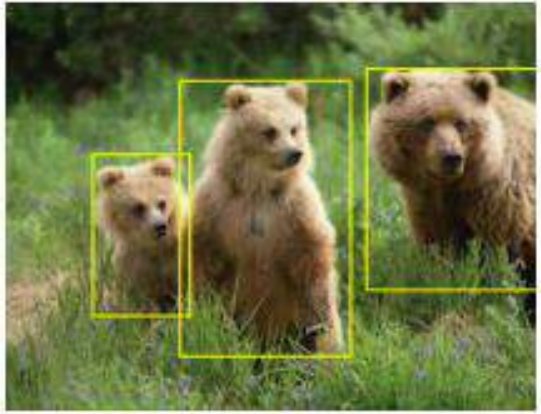





bottle

bed

bear



bear

balloon

bird

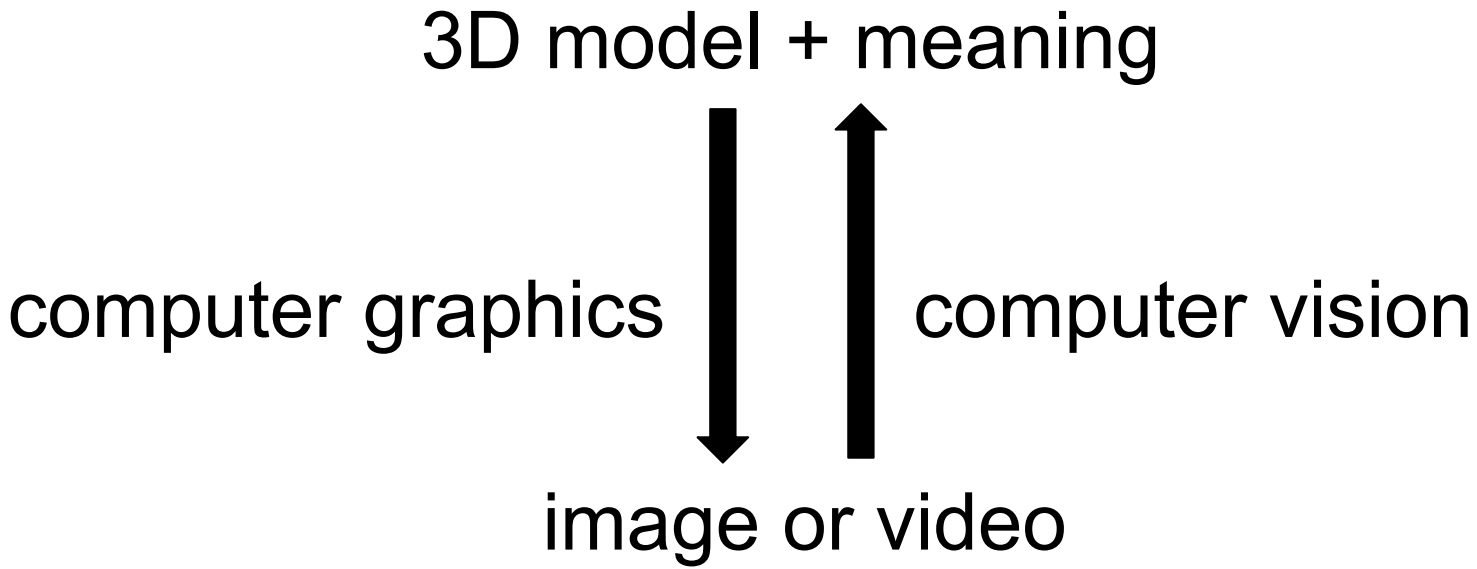

JAZZ #1

# *Research Environments for Machine Learning*

- social networks + photo hosting services ⇒ easy to collect data
- web-based human interaction ⇒ crowdsourcing
- crowdsourcing framework ⇒ distribute work
- "micro payment" ⇒ monetary incentive
- select the right tasks ⇒ "ordinary" people can contribute
- well-defined problems ⇒ crowd can participate

# Synthesized Data + Labels

# *Vision and Graphics are Reverse Problems*




Computer Graphics: also called CG, animation, special effects … in movies

# Computer Graphics and Vision

**Graphics**

**Vision**

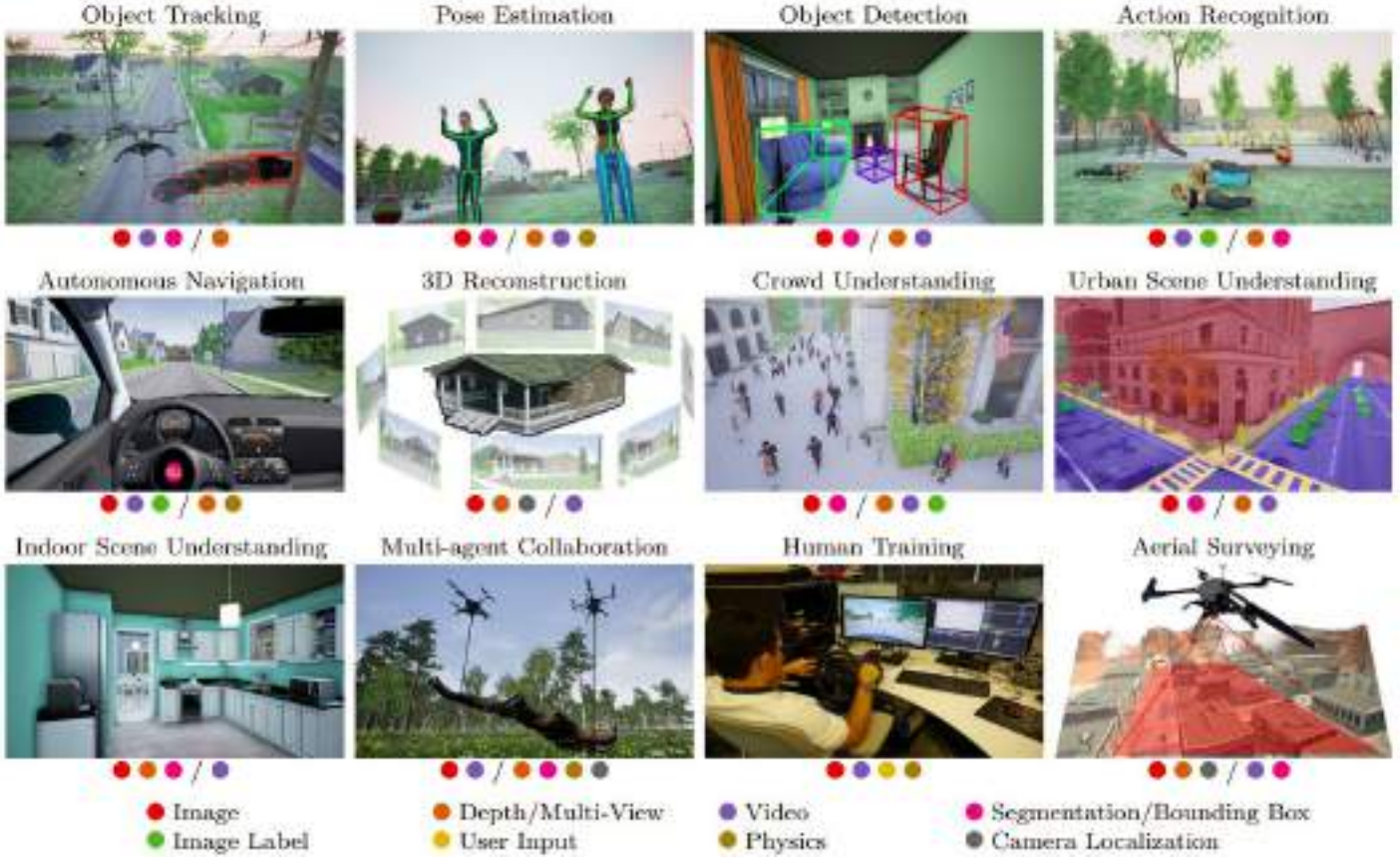
| **Meaning** | **Visual Data** |
|---|---|
| Objects: Airplane, Human, Vehicle, Tree | |
| Season, Time, environment: Autumn, football game | |
| Actions: Walking, Running, Flying | |
| Relationship of objects: Above, Fighting | |

https://brianmmurray.wordpress.com/2013/02/28/feeling-anxious-spend-time-with-nature/
https://www.jconline.com/picture-gallery/sports/2020/10/31/zander-horvath-look-purdue-football-running-back/6103553002/
https://www.rogerebert.com/reviews/ip-man-4-the-finale-movie-review-2019

# *Why to synthesize data?*

- rare events (disasters, accidents, endangered species)
- dangerous environments (fire, pedestrians jumping into traffic)
- seasonal delays (evaluate vision's responses to winter)
- augmented reality
- flexible viewing angles
- scale up to interactions of multiple objects
- repeatable evaluation
- additional information: depth, speed, weight and volume

# *Sim4CV: A Photo-Realistic Simulator for Computer Vision Applications* *International Journal of Computer Vision (2018)*



# *Factors to consider in synthesized data*

- Photo-realistic or not
- Physics (inertia, gravity, turbulence, mass, size, elasticity)
- Weather (wind, rain, fog, sun, shadow)
- Time
- Power and energy
- Surface properties (e.g., reflection)
- Human behavior
- Interaction with physical components

# Applications

# *Photorealistic Image Synthesis for Object Instance Detection* *IEEE International Conference on Image Processing 2019*

Physics-based modeling:

- scattering
- refraction and reflection
- diffuse
- usually very slow

Photorealistic images synthesized by the proposed approach

Real images from LineMod [4] and Rutgers APC [2] datasets

Scene 1

Scene 2

Scene 3

Scene 4

Scene 5

Scene 6

# Virtual Worlds as Proxy for Multi-Object Tracking Analysis

Adrien Gaidon, Qiao Wang, Yohann Cabon, Eleonora Vig

CVPR 2016

Is this real or virtual?

Is this real or virtual?

Segmentation and Depth

# *Challenge in Data Labeling (for Video)*

- volume of data (30 frames per second, multiple objects)
- movement
- occlusion
- diversity

Meanwhile, synthesizing data also has many challenges:

- photorealism
- time-consuming to generate high-quality data
- Playing video games need human players

# *Generate Virtual World*

1. acquire real-world data as a starting point for calibration
2. clone this real-world data into a virtual world
3. generate synthetic sequences with different weather conditions
4. create ground truth annotations
5. evaluate the “usefulness” of the synthetic data

**Figure 2:** Frames from 5 real KITTI videos (left, sequences 1, 2, 6, 18, 20 from top to bottom) and rendered virtual clones (right).

**Figure 3:** Simulated conditions. From top left to bottom right: clone, camera rotated to the right by 15°, to the left by 15°, "morning" and "sunset" times of day, overcast weather, fog, and rain.

**Figure 5:** Predicted tracks on matching frames of two original videos (top) and their synthetic clones (bottom) for both DP-MCF (left) and MDP (right). Note the visual similarity of both the scenes and the tracks. Most differences are on occluded, small, or truncated objects.

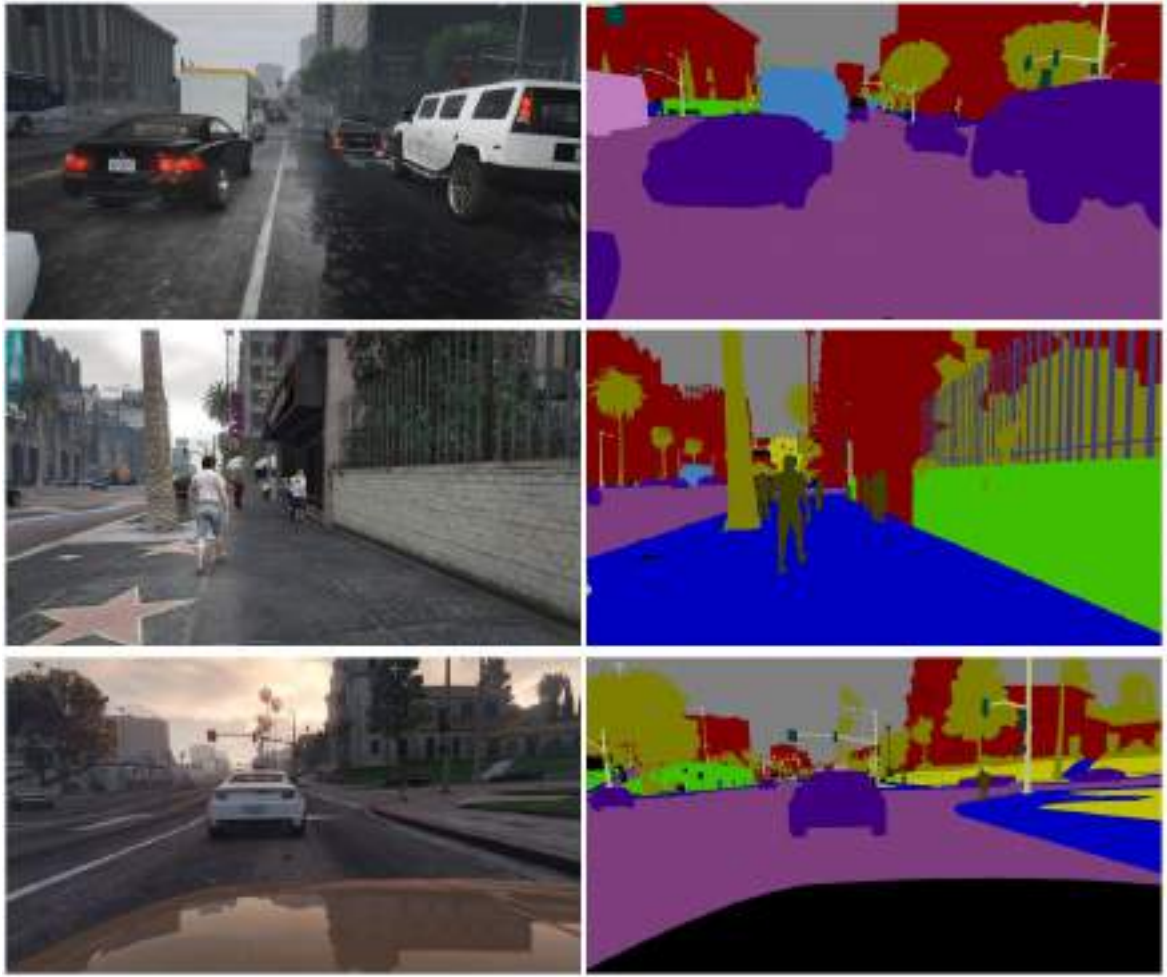
# Playing for Data: Ground Truth from Computer Games

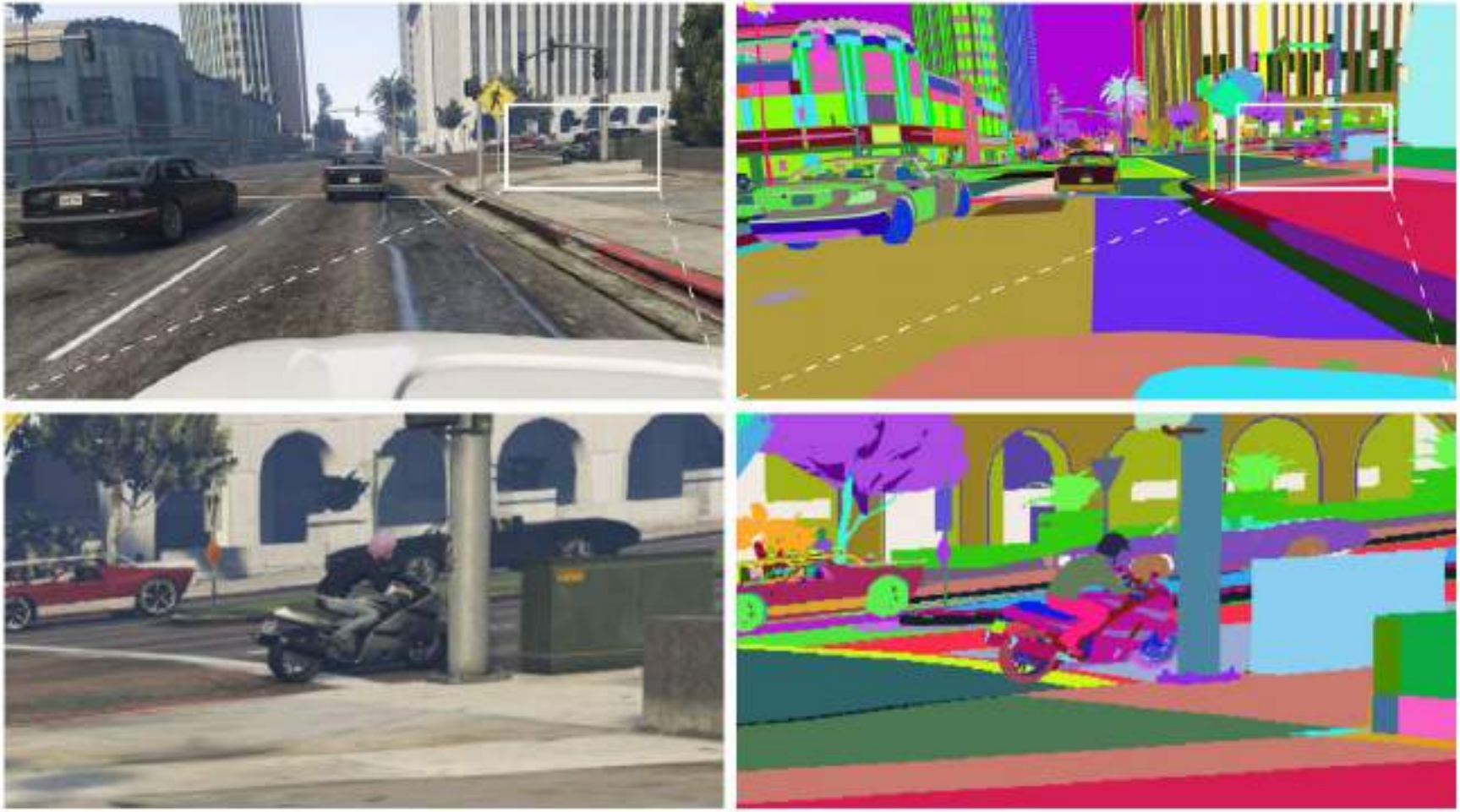
Stephan R. Richter, Vibhav Vineet, Stefan Roth, Vladlen Koltun

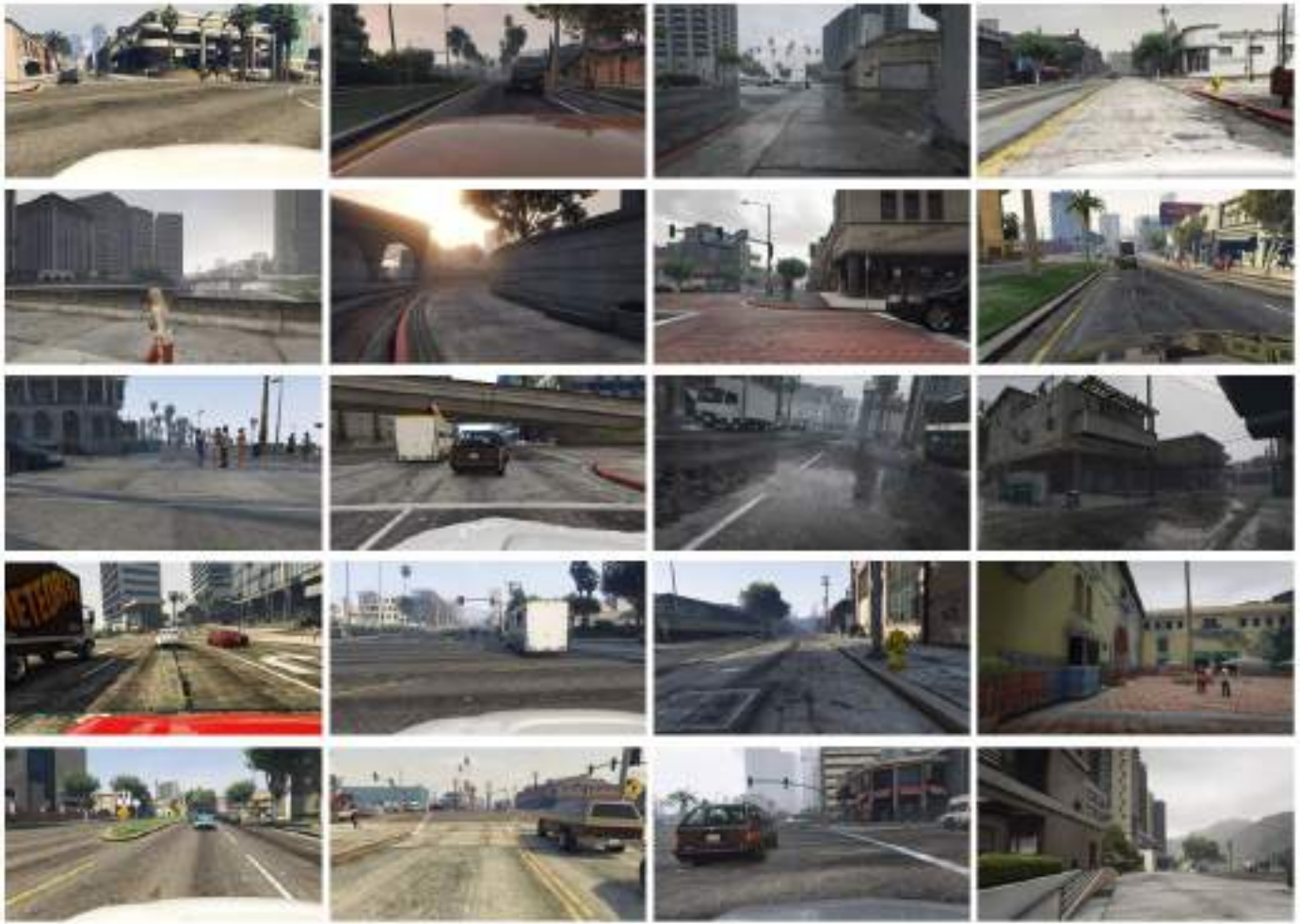
ECCV 2016

# *Capture Graphics Commands in Game*

- open-source games lack details like commercial games
- source code of commercial game engine unavailable
- create pixel-wise semantic labels
- record and reproduce rendering commands at operating systems
- method: intercept communication between software and hardware
  - identify relevant function calls
  - identify hardware resources
  - format for annotation
- 25,000 images, 49 hours for labeling, about 7 seconds / image







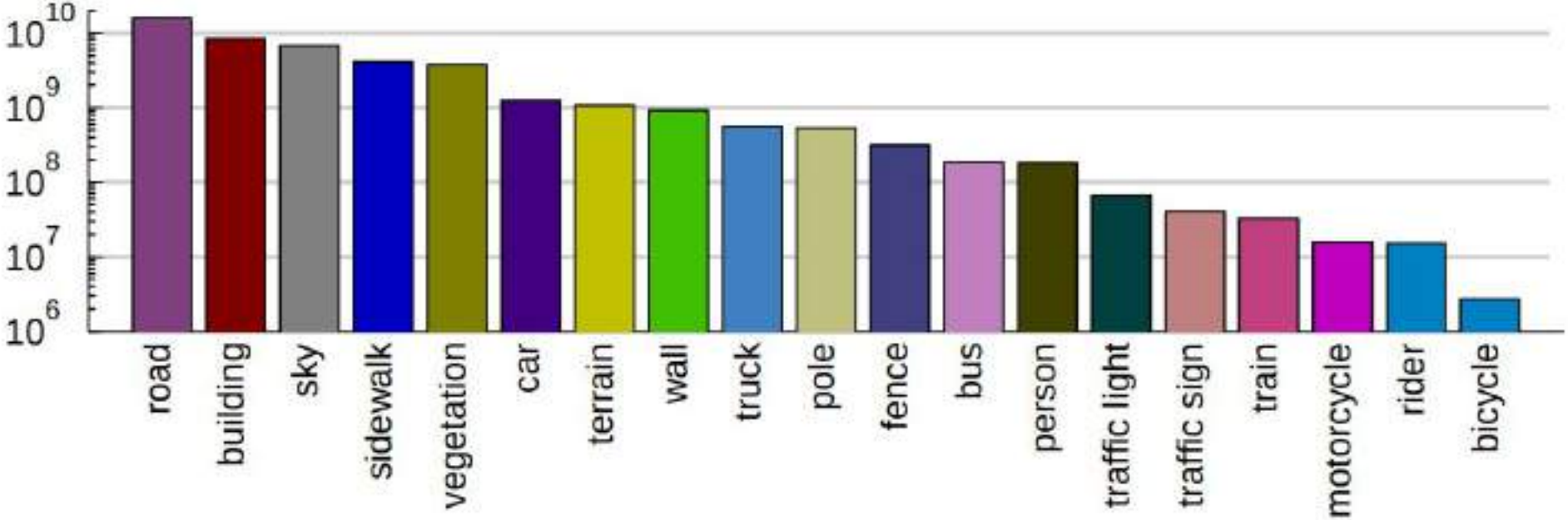



**Fig. 4.** Number of annotated pixels per class in our dataset. Note the logarithmic scale.

| | #pixels [$10^9$] | annotation density [%] | annotation time [sec/image] | annotation speed [pixels/sec] |
|---|---|---|---|---|
| GTA5 | 50.15 | 98.3 | 7 | 279,540 |
| Cityscapes (fine) [11] | 9.43 | 97.1 | 5400 | 349 |
| Cityscapes (coarse) [11] | 26.0 | 67.5 | 420 | 3095 |
| CamVid [8] | 0.62 | 96.2 | 3,600 | 246 |
| KITTI [39] | 0.07 | 98.4 | N/A | N/A |