Playing for Benchmarks

Vladlen Koltun

Stephan Richter

Zeeshan Hayder

Visual Perception vs Computer Vision

Visual perception:

- Broad competence
- Predictive model that supports planning and action
- Integrates information over time
- Robust (e.g., time of day, weather)

Visual Perception vs Computer Vision

Visual perception:

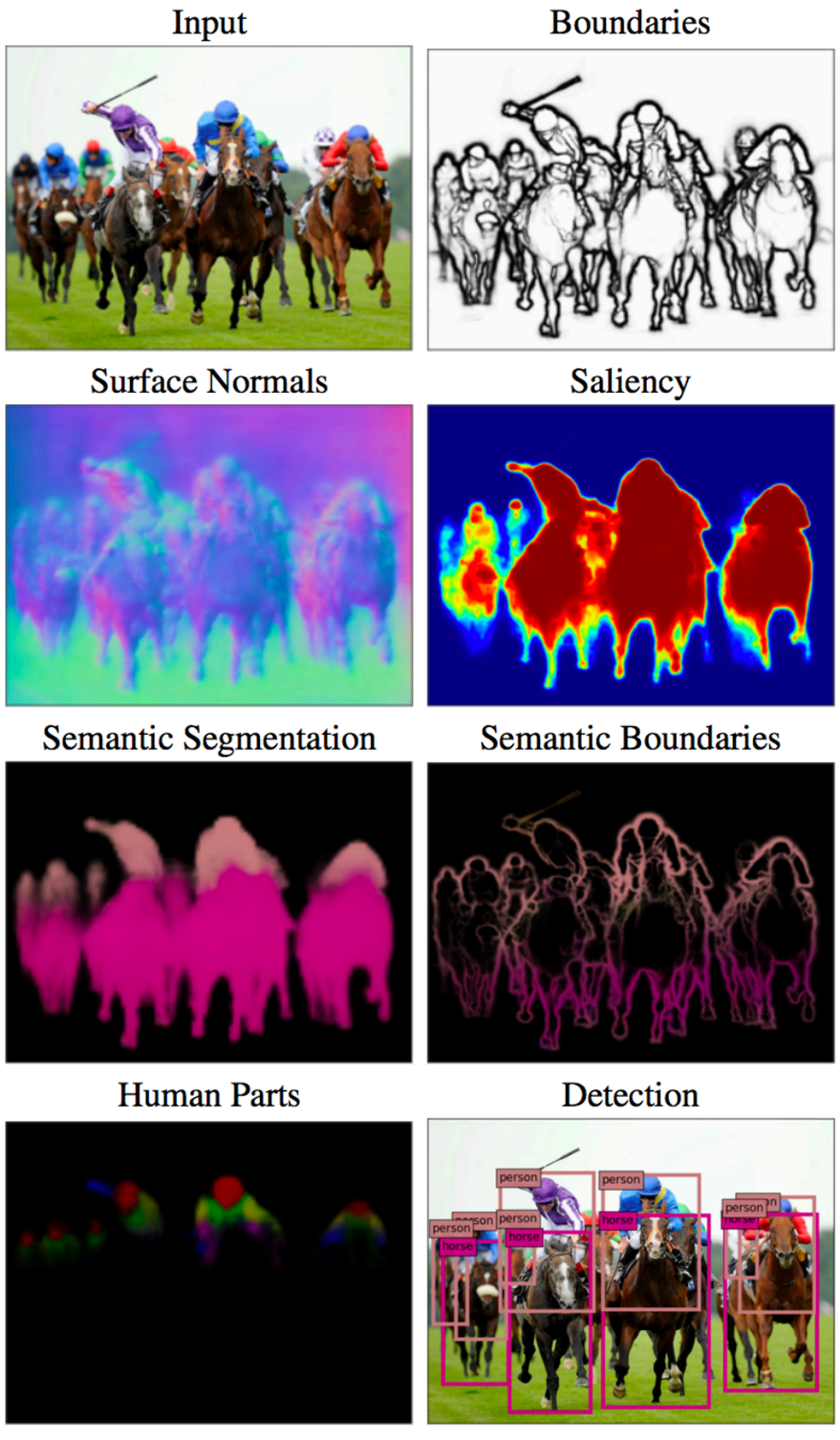
- Broad competence
- Predictive model that supports planning and action
- Integrates information over time
- Robust (e.g., time of day, weather)

Computer vision:

- Narrow, isolated modules: flow, semantic segmentation, etc.
- Mostly images
- Brittle

Broad Competence

- Malik et al., *The three R's of computer vision*, PRL 2016
- Kokkinos, *UberNet*, CVPR 2017
- Bilen and Vedaldi, *Integrated perception with recurrent multi-task neural networks*, NIPS 2016
- Misra, Shrivastava, Gupta, Hebert, *Cross-stitch networks for multi-task learning*, CVPR 2016



Kokkinos, CVPR 2017

Video and motion

- Galasso et al., *A unified video segmentation benchmark*, ICCV 2013
- Agrawal, Carreira, Malik, *Learning to see by moving*, ICCV 2015
- Jayaraman and Grauman, *Learning image representations tied to ego-motion*, ICCV 2015
- Pathak et al., *Learning features by watching objects move*, CVPR 2017

Robustness

Maddern et al., *The Oxford RobotCar dataset*, IJRR 2017

How can we accelerate progress?

Visual perception:

- Integrated process
- Broad competence
- Predictive model that supports planning and action
- Integrates information over time
- Robust (e.g., time of day, weather)

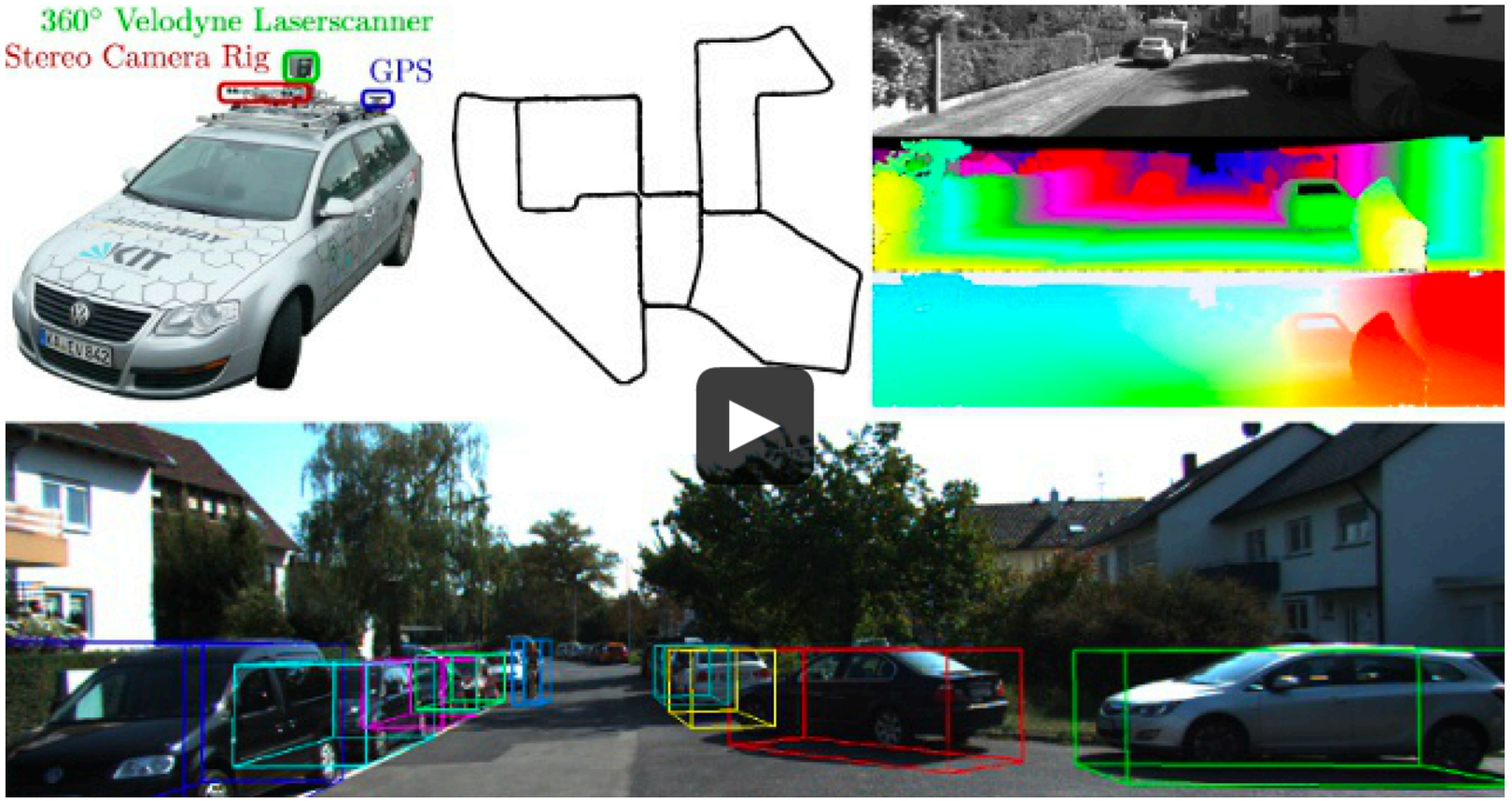
Computer vision:

- Narrow, isolated modules: flow, semantic segmentation, etc.
- Mostly images
- Brittle

Common task framework

- "What gets measured gets optimized."
- Donoho, *50 years of data science*, 2015. Based on reports by Marc Liberman.
- A methodological framework that counteracts susceptibility to "glamour and deceit".
- Public dataset, benchmark, error measures. Objective evaluation on a sequestered test set.

Existing benchmark suites



Geiger, Lenz, Urtasun, *The KITTI vision benchmark suite*, CVPR 2012

+ Both low-level and high-level tasks
+ Video
-- Limited accuracy (e.g., sparse LiDAR, fitted CAD models, limited annotation for a small number of frames and classes)
-- Single town in fair weather

Existing benchmark suites

Cordts et al., *The Cityscapes dataset*, CVPR 2016

+ 50 cities in Europe
+ Video
-- 5K frames are annotates in detail (90 min per frame annotation time)
-- Semantic segmentation and instance segmentation only
-- Daytime images in fair weather

Goals

- HD video, densely annotated
- Low-level and high-level tasks on the same data
- Ground-truth for all tasks on all frames
- Pixel-level segmentation and correspondences
- Diverse environmental conditions (e.g., rain, night)

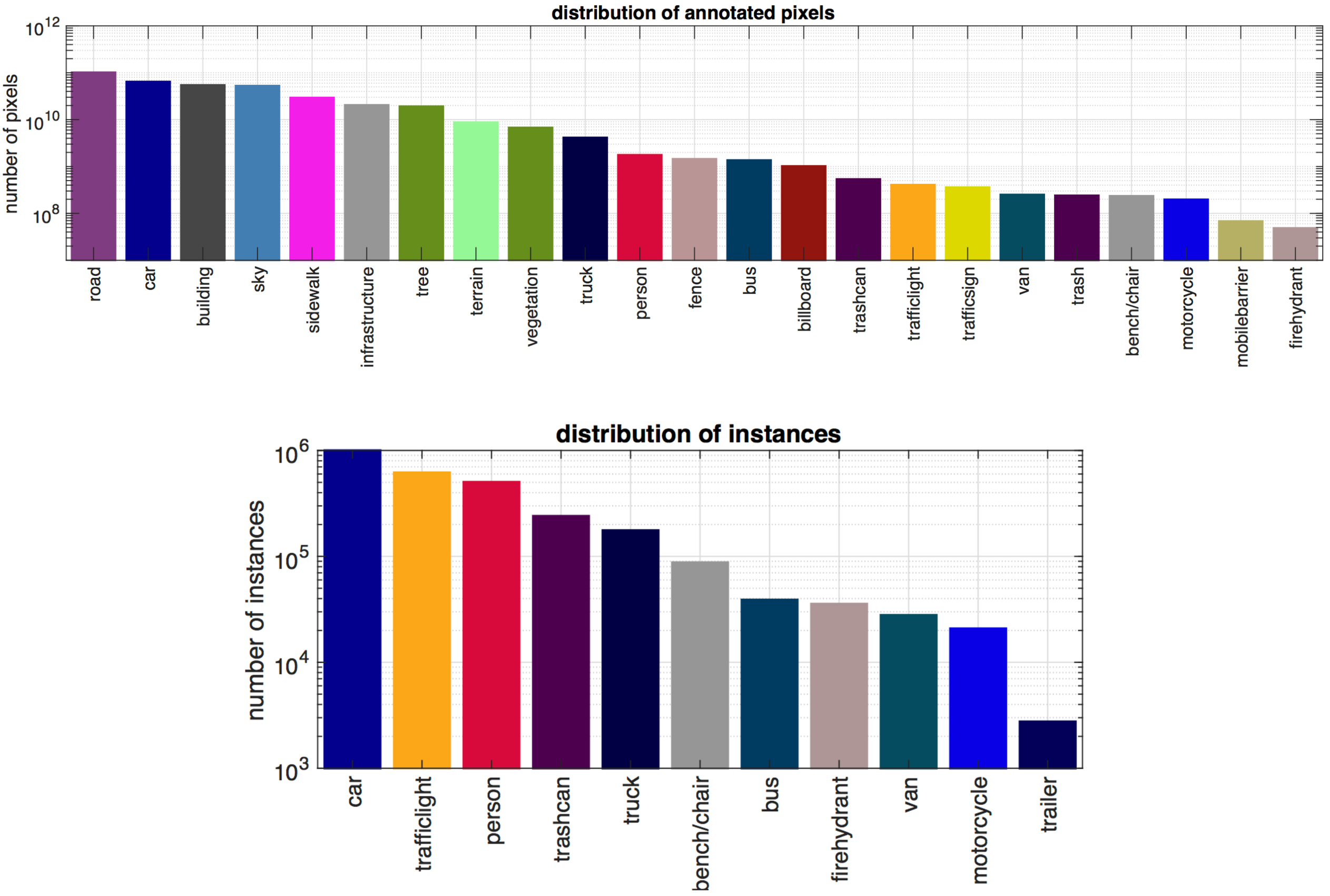
The VIPER benchmark suite

- 250K video frames (1080p, 15fps)
- Driving, riding, and walking 184 km
- Day, sunset, rain, snow, night
- Ground-truth for all tasks on all frames
- Pixel-level segmentation and correspondences
- Semantic segmentation, semantic instance segmentation and tracking, 3D scene layout (and tracking), optical flow, visual odometry, ...

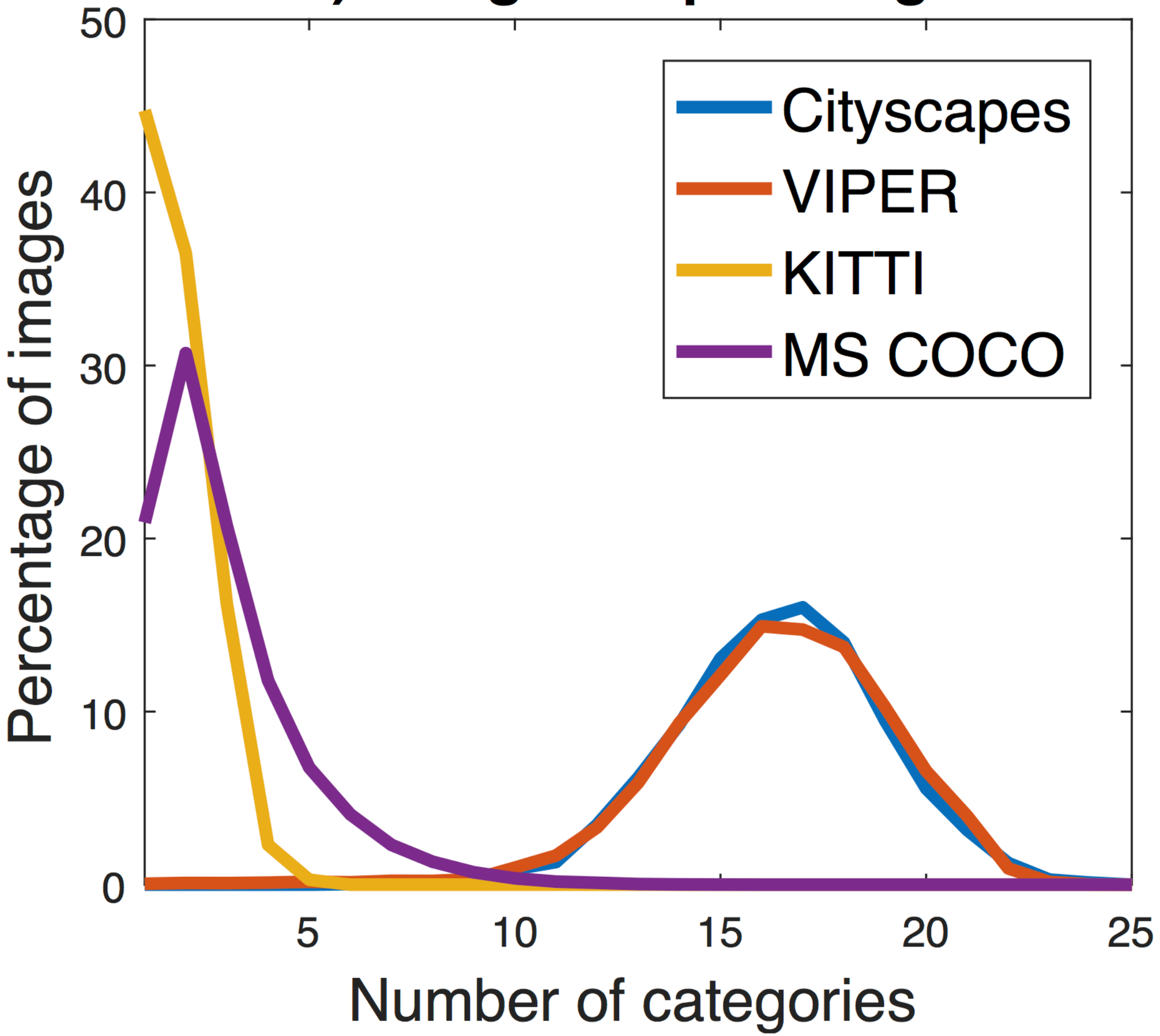
The VIPER benchmark suite

Why did we use a game?

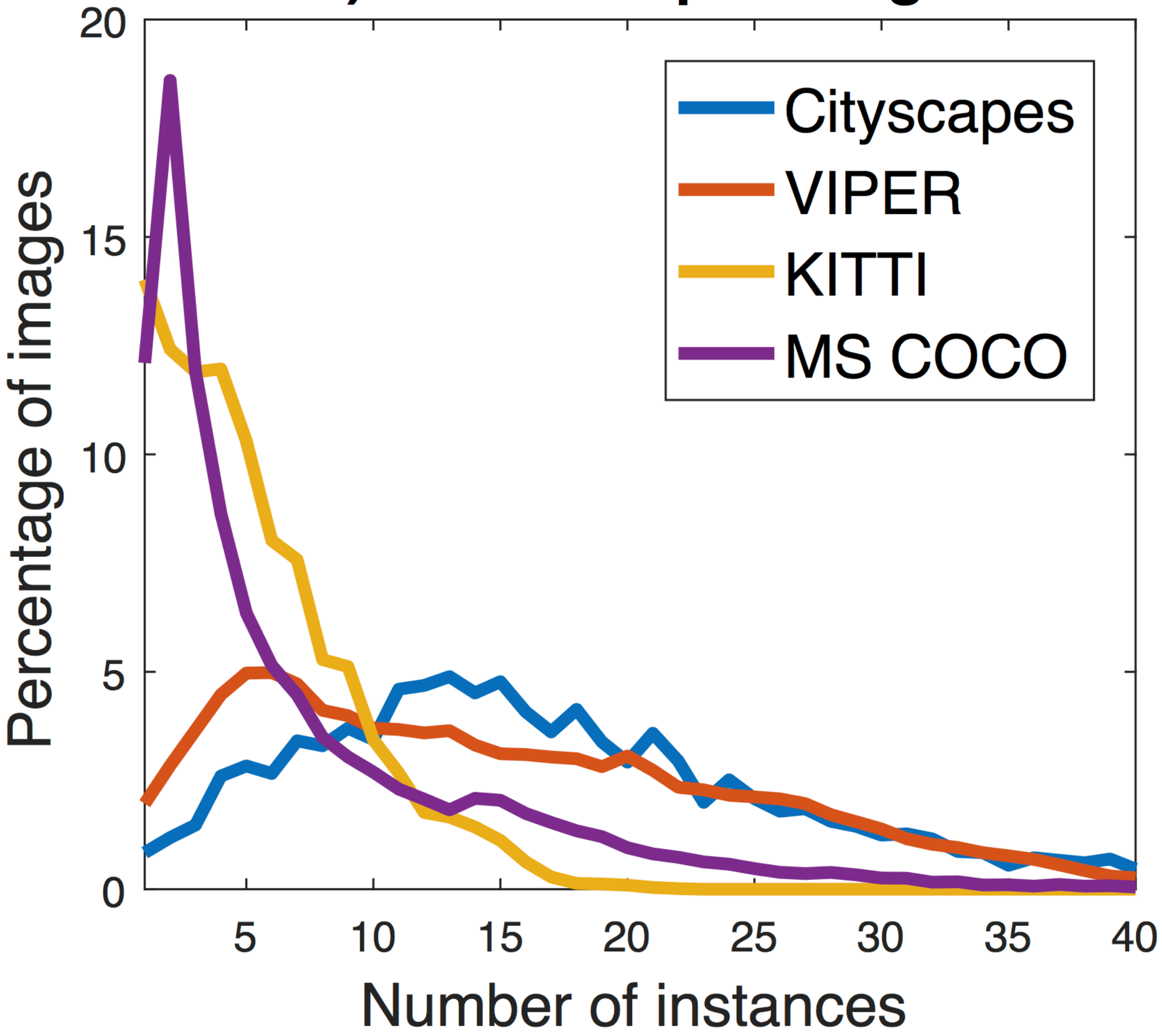
- Why not the physical world?
- Why not a custom virtual environment?


distribution of annotated pixels
number of pixels
10^{12}
10^{10}
10^{8}
road
car
building
sky
sidewalk
infrastructure
tree
terrain
vegetation
truck
person
fence
bus
billboard
trashcan
trafficlight
trafficsign
van
trash
bench/chair
motorcycle
mobilebarrier
firehydrant
distribution of instances
number of instances
10^{6}
10^{5}
10^{4}
10^{3}
car
trafficlight
person
trashcan
truck
bench/chair
bus
firehydrant
van
motorcycle
trailer

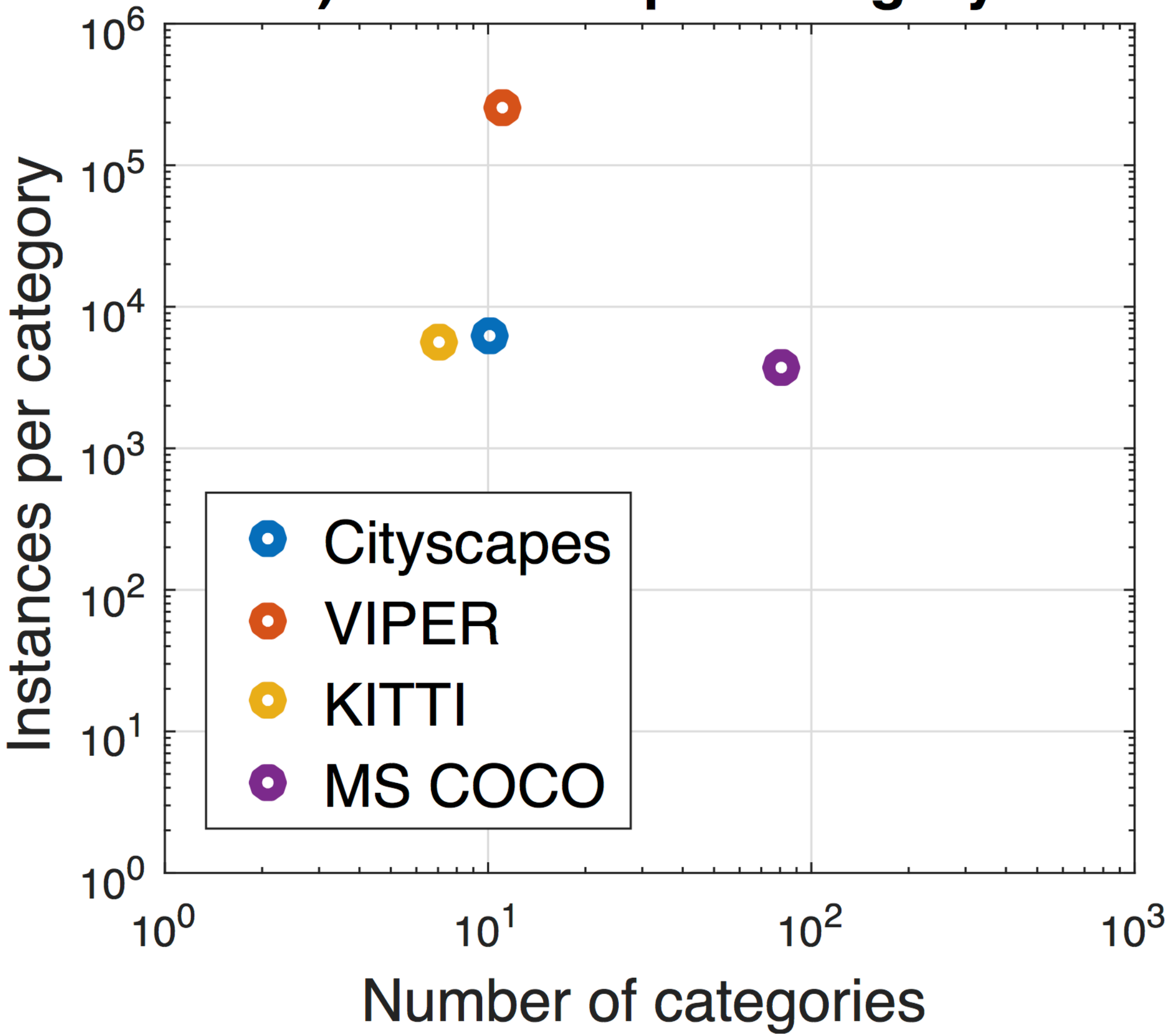
a) categories per image



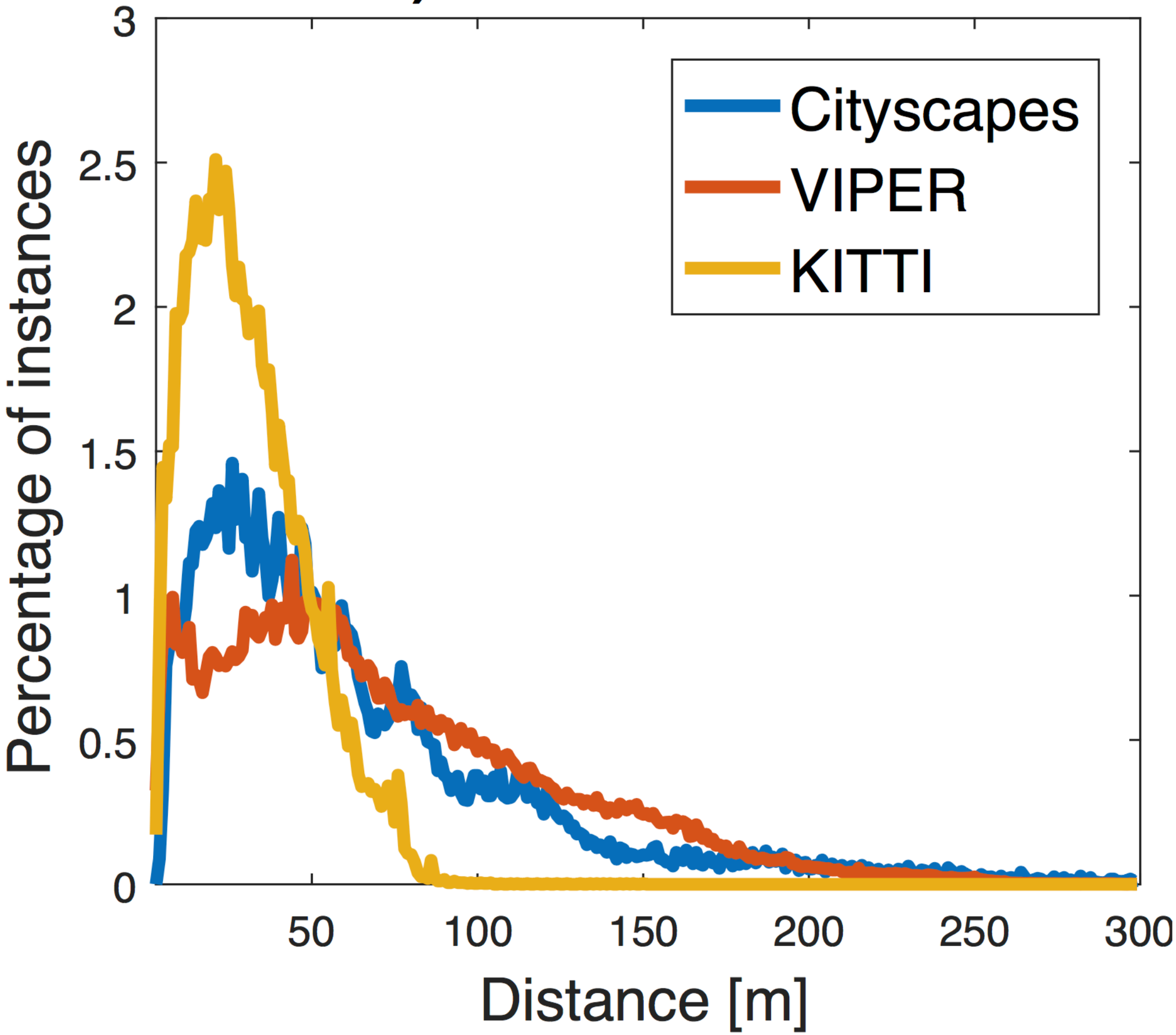
b) instances per image

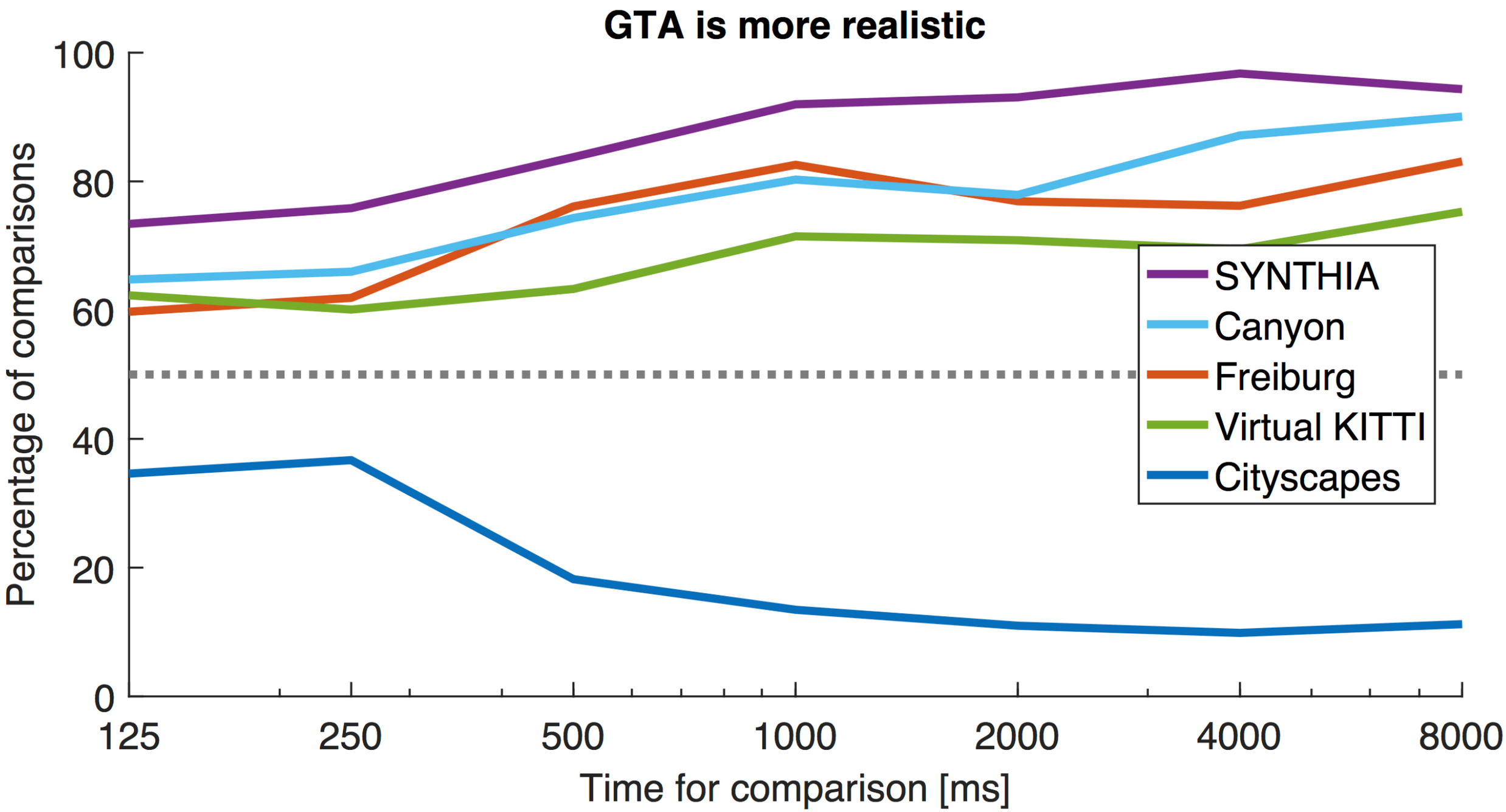


c) instances per category



d) vehicle distance




GTA is more realistic
100
80
60
40
20
0
Percentage of comparisons
125
250
500
1000
2000
4000
8000
Time for comparison [ms]
SYNTHIA
Canyon
Freiburg
Virtual KITTI
Cityscapes

Methodology

- Augment the game's shaders to render out resource IDs and other metadata
- Based on dynamic software updating. Operate directly on the bytecode.
- Instances and 3D layout: track and cluster transformation matrices.
- Dense correspondences: dense 3D coordinates in the object's coordinate frame. Trace through geometric transforms and vertex shaders.

Baselines and analysis

We evaluate several baselines on each task.

playing-for-benchmarks.org

Training set

Modality	Size	Links
Image (JPG)	27.1 Gb	Part 1, Part 2, Part 3, Part 4, Part 5, Part 6
Image (PNG)	-	Coming soon
Camera poses	5.3 Mb	Download
Semantic class labels	4.5 Gb	Download
Semantic instance labels	0.7 Gb	Download
Semantic instance labels (Cityscapes encoding)	0.9 Gb	Download
2D object bounding boxes	-	Coming soon
3D object bounding boxes	-	Coming soon
Optical flow	-	Coming soon
Semantic boundaries	-	Coming soon

Tasks

- Semantic segmentation
- Semantic instance segmentation
- 3D scene layout
- Optical flow
- Visual odometry
- Semantic boundaries
- Detection and tracking (2D and 3D)
- Prediction
- *Suggestions?*

Thank you