Play and Learn: Using Video Games to Train Computer Vision Models

Alireza Shafaei, James J. Little, Mark Schmidt
University of British Columbia

BMVC 2016

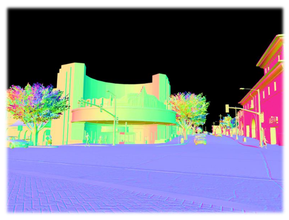
Video Games vs. Reality

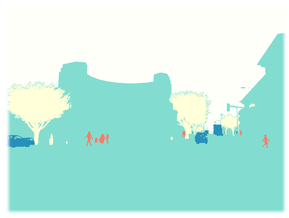
Video game

Google street view

Are existing video games visually realistic enough to improve computer vision models in practice?

Why Video Games?





- **Free Groundtruth Annotation**: image segmentation, depth maps, surface normals, shadows, precise localization, optical flow, etc.
- **Controllable Environment**: variations of seasons, times of the day, climate settings, points of view, interactions, etc.
- **Automation and Scalability.**

Method

- **Dense Image Classification**. Measure the performance of FCN8s [3] in two approaches: (i) fine-tuning on a real-world dataset with various pre-training strategies, and (ii) cross-dataset evaluation.
- **Depth Estimation.** Measure the improvement in image patch ordering task under the method of Zoran *et al*. [4].

Summary of Results

- A convolutional network trained on synthetic data achieves a similar test error to a network that is trained on real-world data for dense image classification on a new dataset.
- The video game dataset can deliver similar or better results compared to the real-world datasets if a simple domain adaptation technique is applied.
- Pre-training on synthetic data results in better initialization and final local minima in the optimization of convolutional networks.
- Video games can offer an alternative way to compile large datasets for direct training or augmenting real-world datasets.

Selected References

[1] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 2009.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[4] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T. Freeman. Learning ordinal relationships for mid-level vision. In *ICCV*, 2015.

Datasets

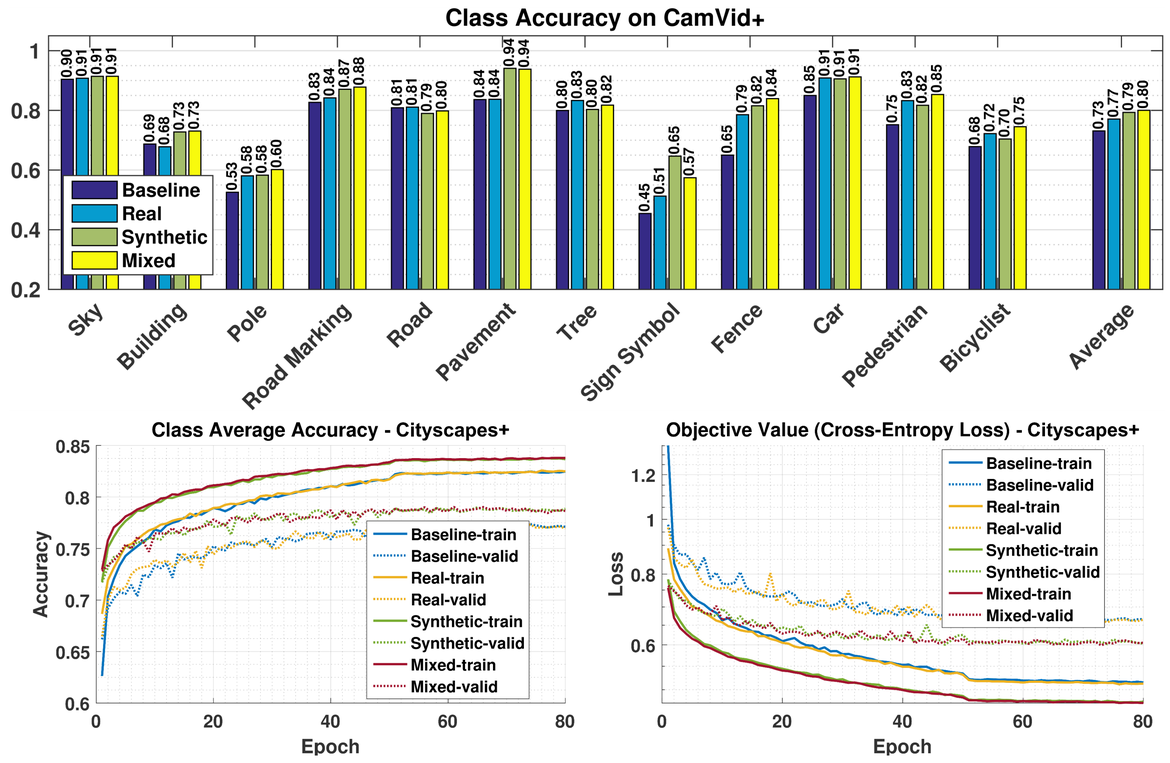
Synthetic

CamVid [1]

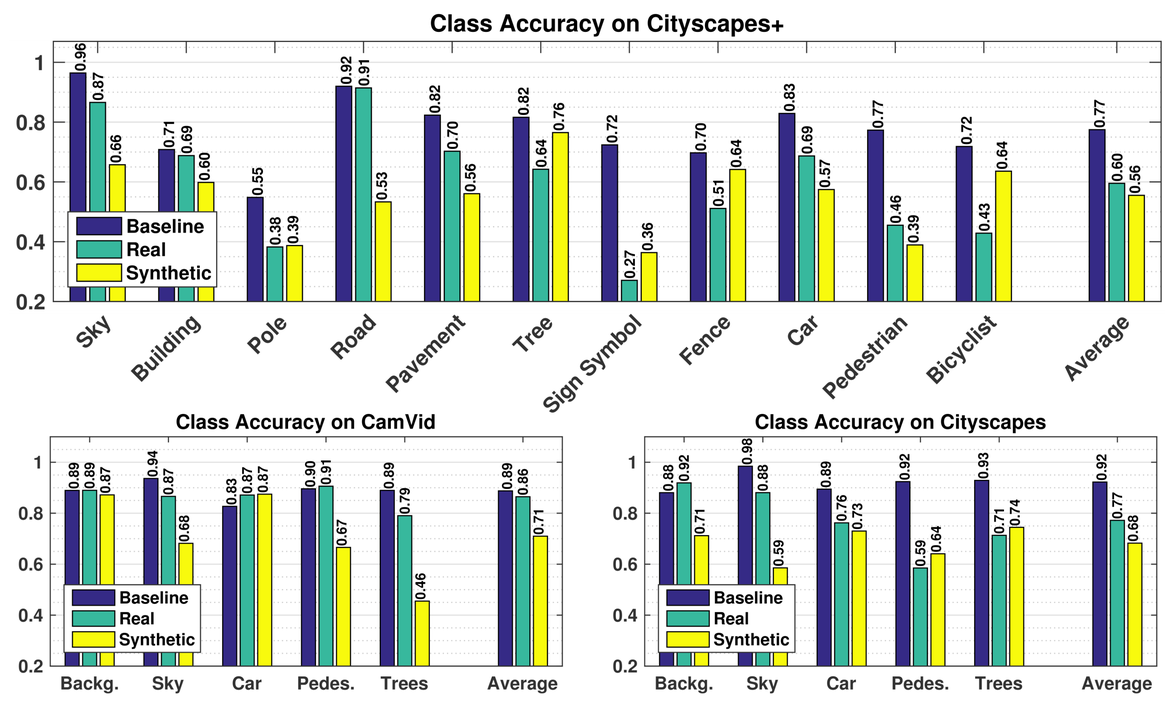
Cityscapes [2]

- **Synthetic.** A camera is mounted on a car, and an autonomous driver wanders around the city while a separate process captures data. We collect over 60,000 samples with annotation.
- **CamVid and Cityscapes.** A 5-class annotation of the data.
- **CamVid+ and Cityscapes+.** A 12-class annotation of the data.

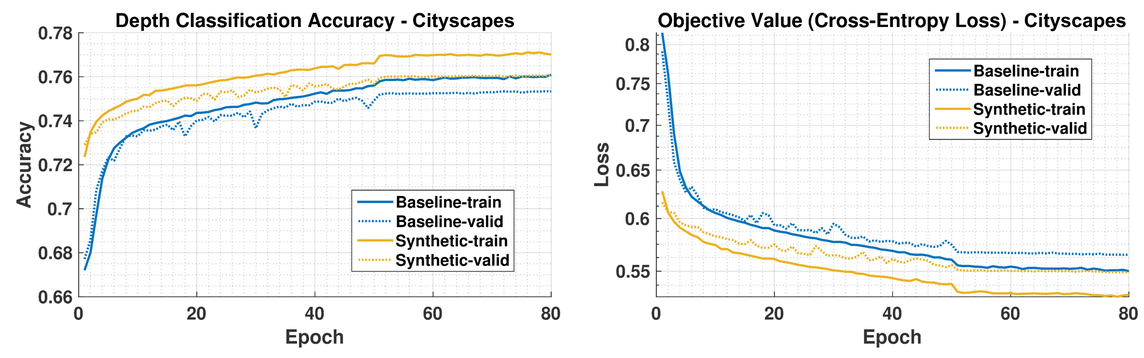
Fine-tuned Dense Image Classification



Cross-dataset Dense Image Classification



Depth Estimation from RGB



NSERC CRSNG