PLAY AND LEARN: USING VIDEO GAMES TO TRAIN COMPUTER VISION MODELS

Alireza Shafaei, James J. Little, and Mark Schmidt

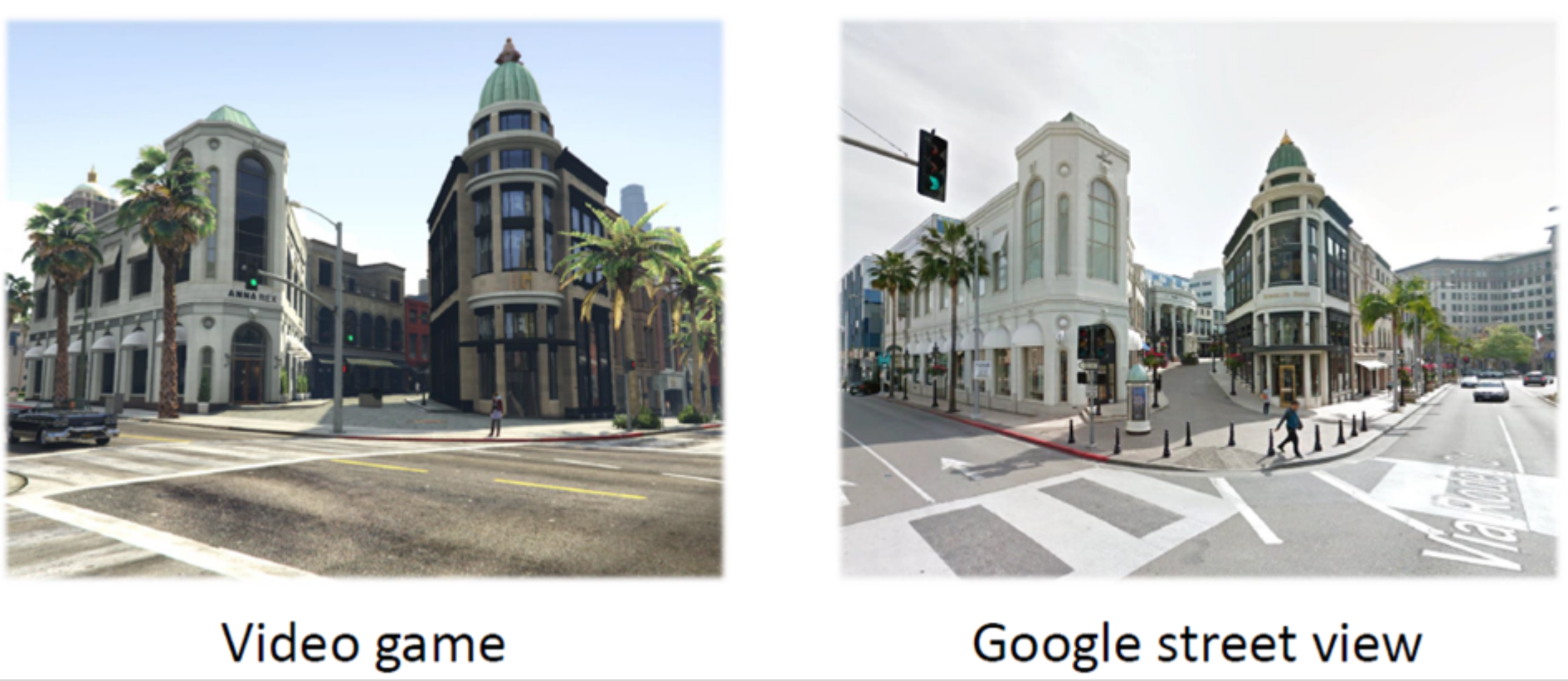
DECEMBER 2016 – CIFAR LEARNING IN MACHINES AND BRAINS

Single-Label vs. Dense-Labels

- CNNs are now the dominant models in computer vision.
  - Key ingredient: large labeled datasets (ImageNet: >13 million images).
- What about **dense labels** as in image segmentation?
  - Much harder to get large labeled datasets (Cityscapes: 5k images).

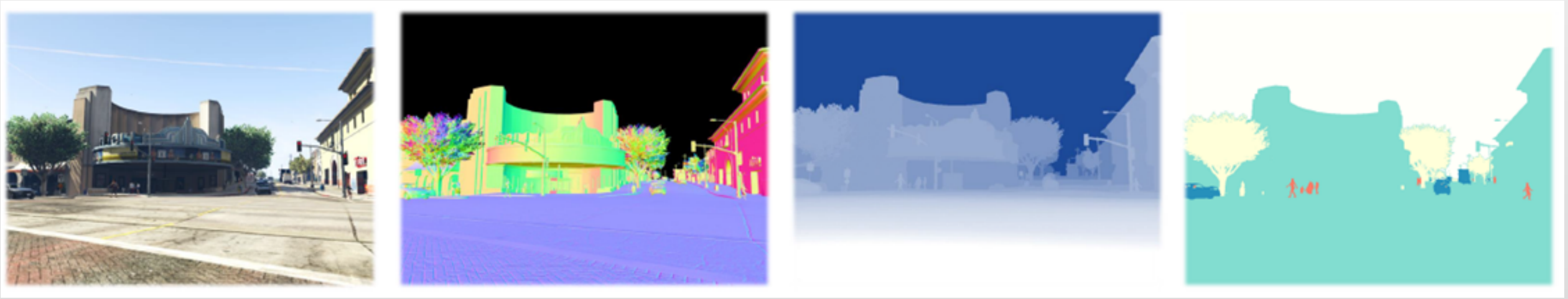
Simulations vs. Reality

- Classic solution: simulate more labeled examples.
- This work: can **video games** augment real data?



Why Video Games?

- **Free and scaleable groundtruth** annotation for many tasks:
  - Segmentation, depth, surface normal, shadows, optical flow, and so on.
- **Active control of environment**:
  - Viewpoints, time of day, weather, dangerous situations, and so on.

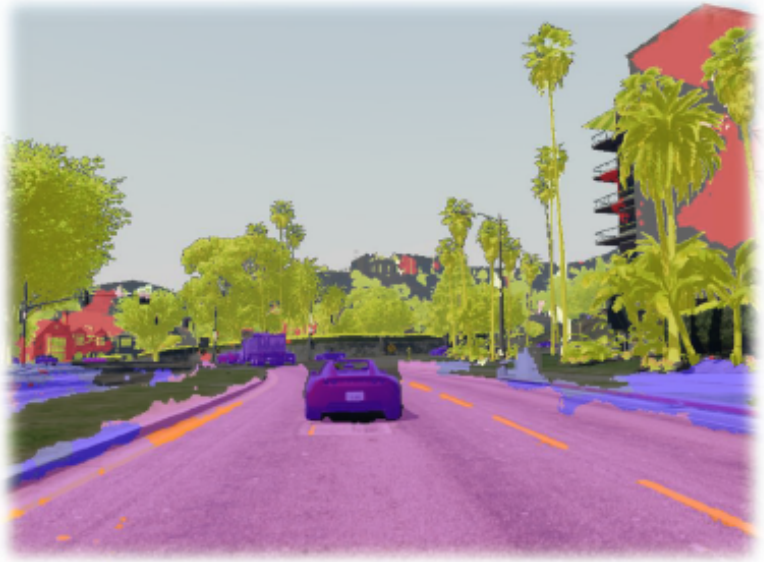


Method

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].

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-tuning performance (Segmentation)

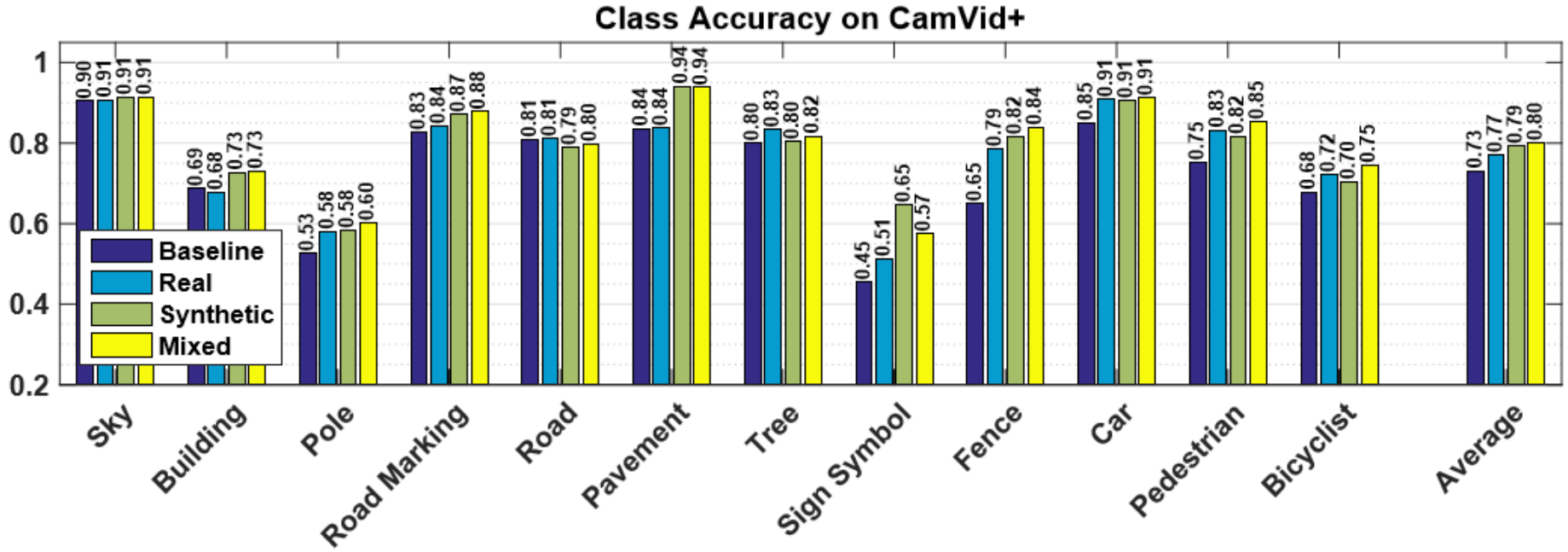
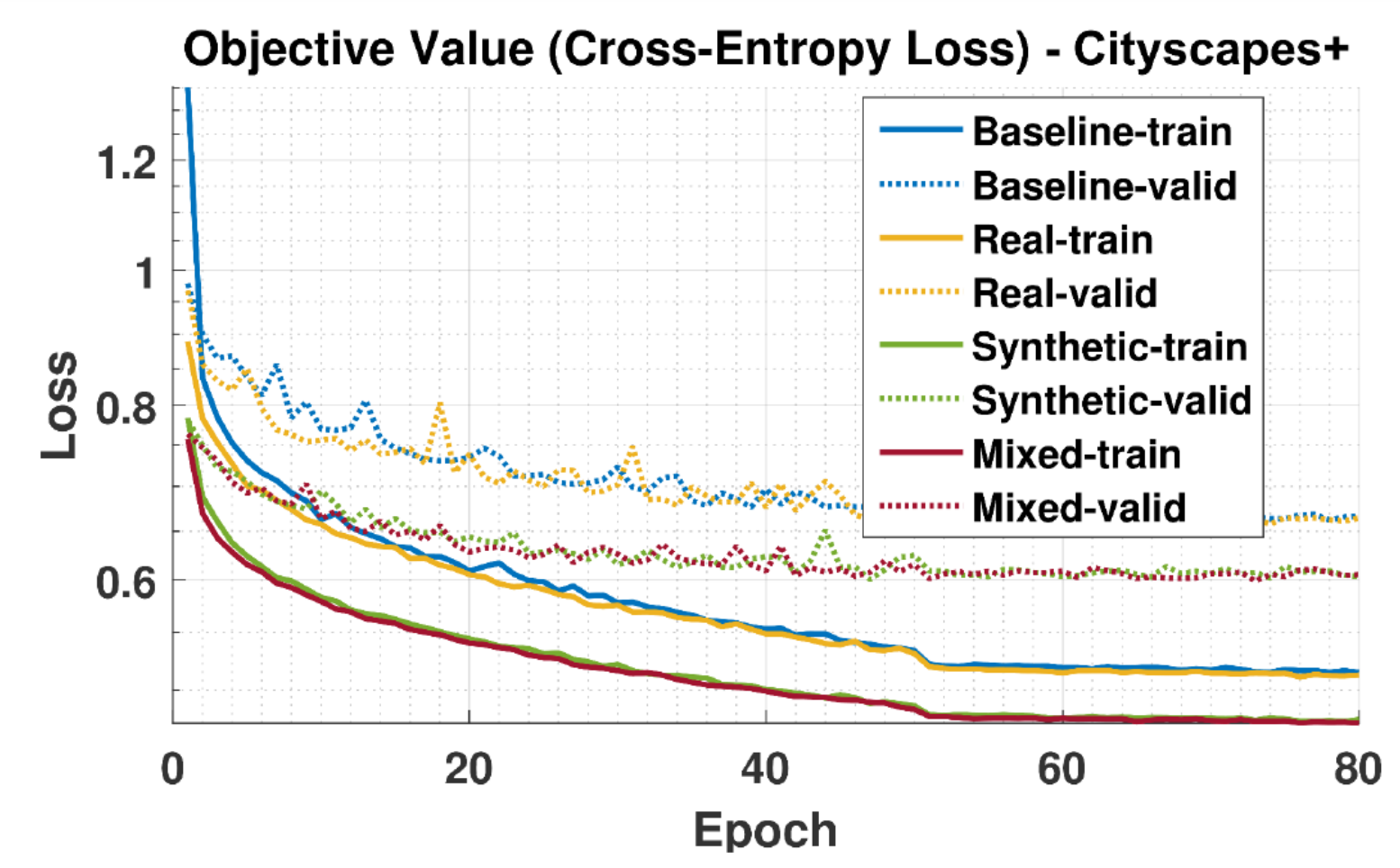
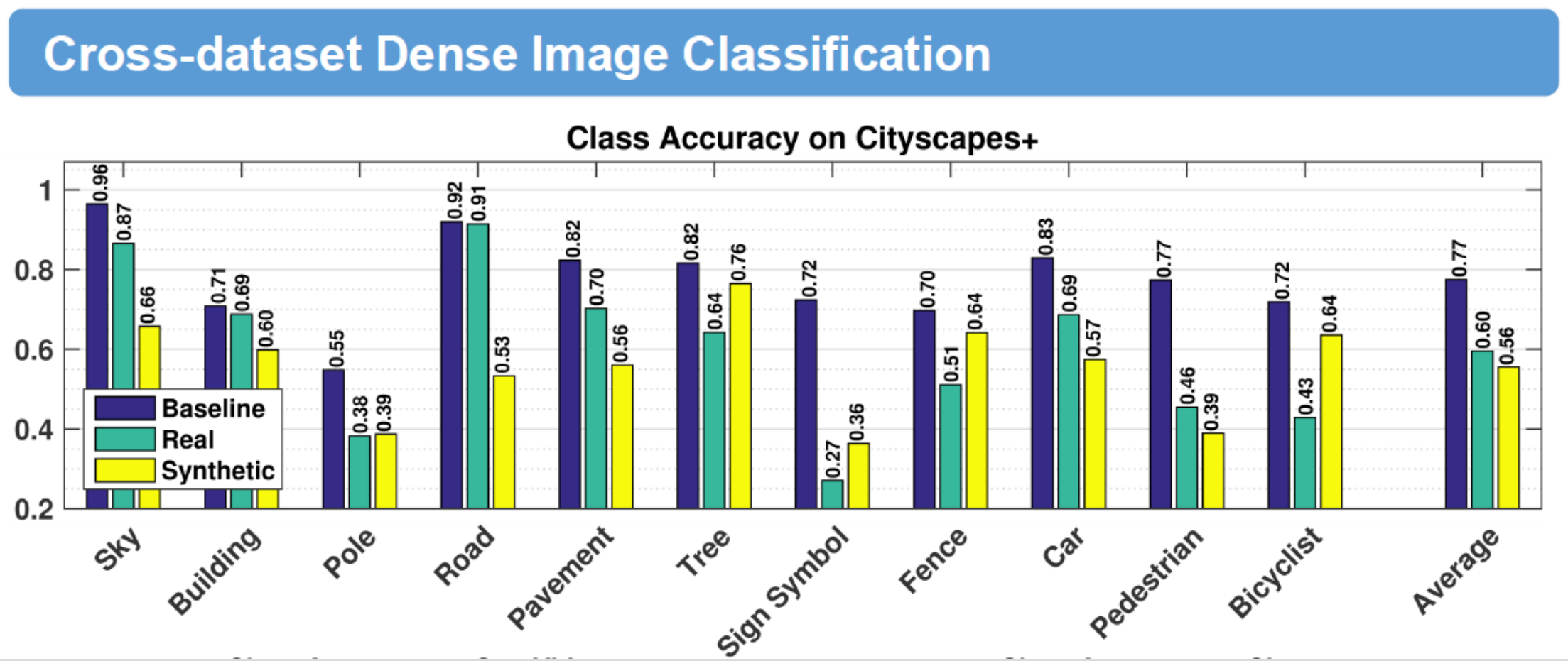


Figure 4: The per-class accuracy on the test set of `CamVid+` dataset. The `baseline` is trained on the target dataset, the `real` is pre-trained on the real alternative dataset, and the `synthetic` is pre-trained on `VG+`. The `Mixed` approach is pre-trained on both synthetic and the real alternative dataset. Pre-training the baseline with the synthetic `VG+` improves the average accuracy by 6%, while pre-training with the real-world `Cityscapes+` improves the average by 4%.

Fine-tuning performance (Segmentation)

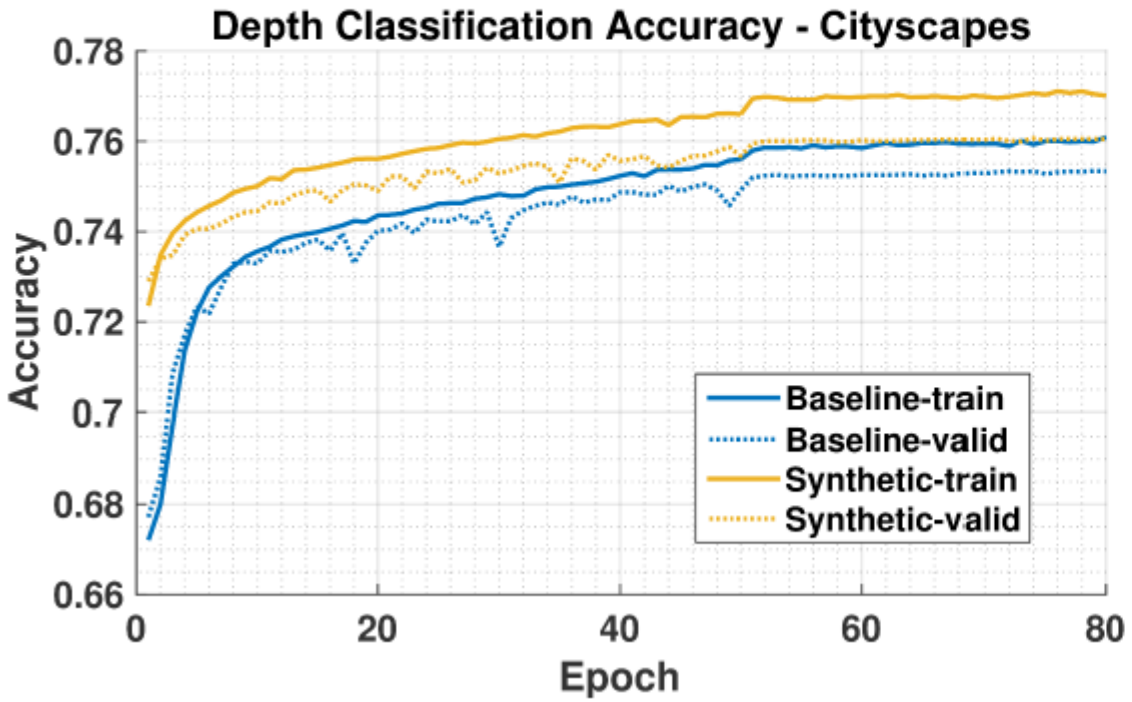


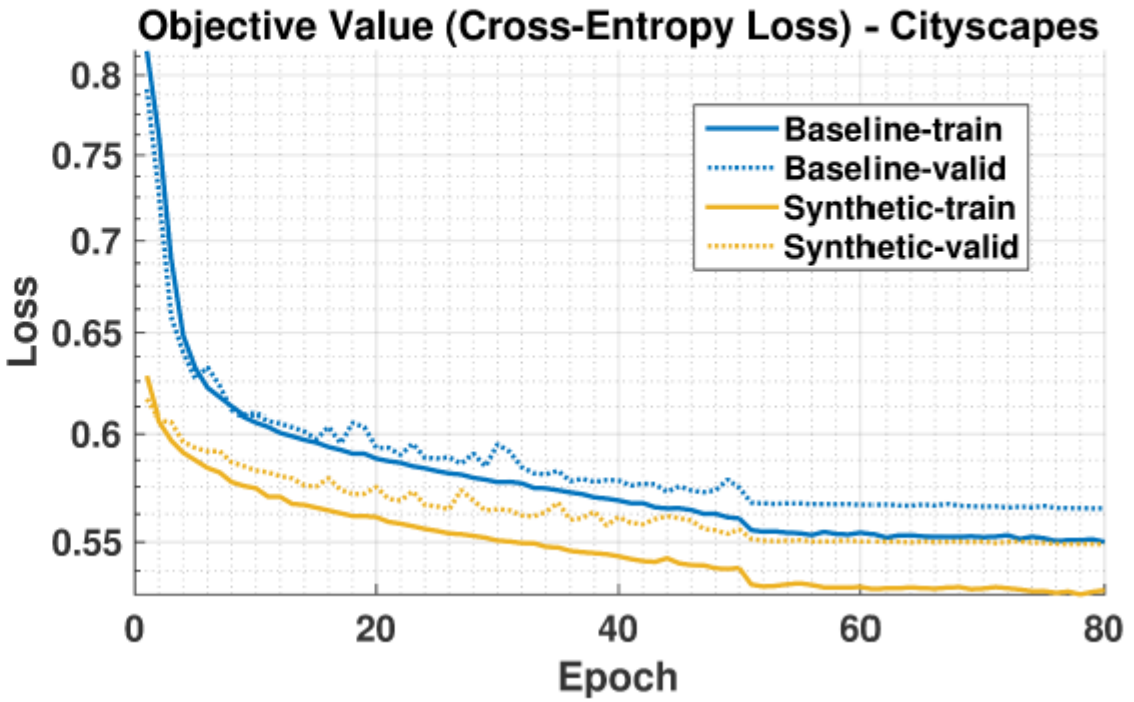
Performance without fine-tuning.



Fine-Tuning Performance (Depth)

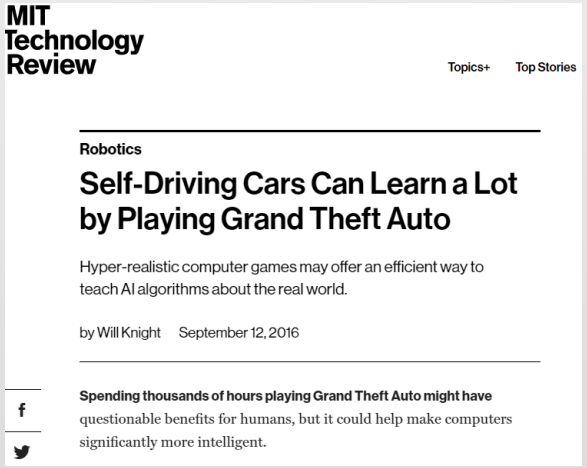
Depth Estimation from RGB

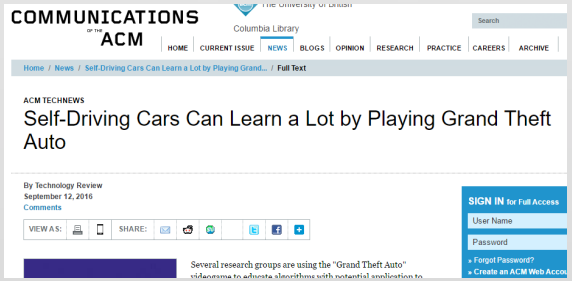


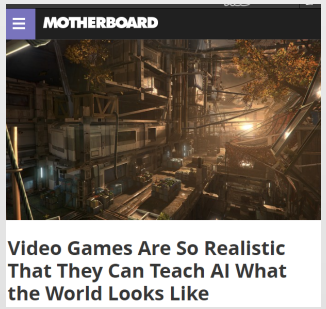


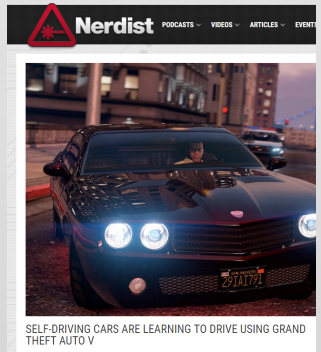
Take-Home Message

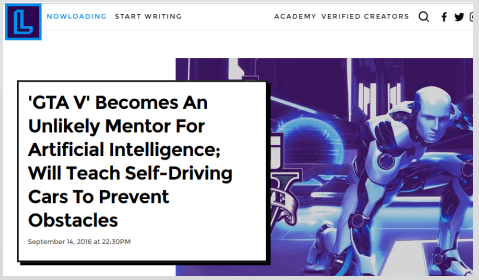
- Training on video game synthetic data achieved:
  - Similar test accuracy to training on another real dataset.
  - Better test accuracy when using real data to fine-tune.
- Video games may offer a way to compile large labeled datasets.

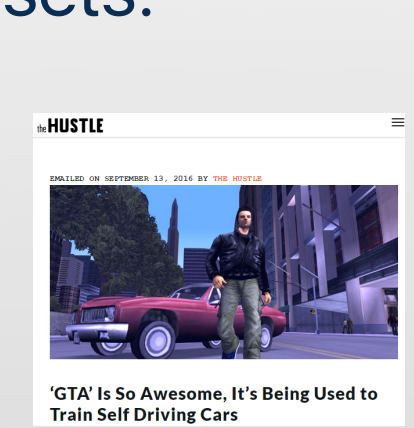


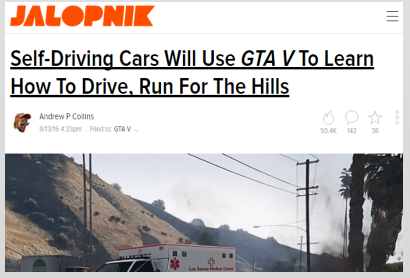













UBC CS is (also) hiring in all ranks/areas.