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

Alireza Shafaei
http://cs.ubc.ca/~shafaei
James J. Little
http://cs.ubc.ca/~little
Mark Schmidt
http://cs.ubc.ca/~schmidtm

Department of Computer Science
University of British Columbia
Vancouver, Canada

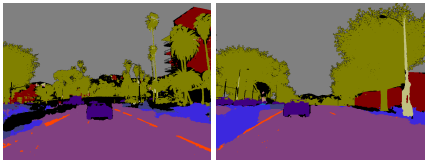

Figure 1: Densely-labeled samples from the synthetic dataset.

Are video games realistic enough to be used in the training of computer vision models to address practical, real-world problems? We explore this idea and deliver a proof of concept by experimenting with the synthetic RGB images that we sample from a video game. We collect over 60,000 synthetic samples with similar conditions to the real-world CamVid [1] and Cityscapes [2] datasets. We provide several experiments to demonstrate that the synthetically generated RGB images can be used to improve the performance of deep neural networks on both image segmentation and depth estimation.

Dataset. We capture the synthetic dataset by sampling the game every second while an autonomous driver is wandering in the city. Each sample contains the RGB image, semantic segmentation, depth image, and the surface normals. See Fig. 1.

Experiments. We use the FCN8 [3] architecture for the dense image classification task, and for the depth estimation experiments we use the approach of Zoran *et al*. [4].

Results. We show that in a cross-dataset setting, the CNNs that we obtain from synthetic data have a similar test error as the networks that we train on real-world data (Fig. 2). Furthermore, the synthetically generated RGB images can provide similar or better results compared to the real-world datasets if a simple domain adaptation technique is applied (Tab. 1). We also show that pre-training on synthetic data results in a better initialization and final local minima in the optimization. For the depth estimation task, we present similar improvements.

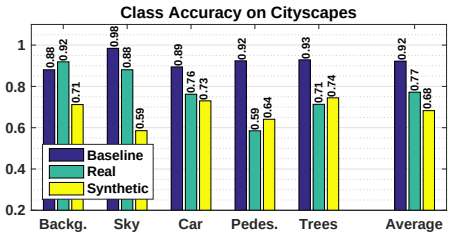

Figure 2: The cross-dataset per-class accuracy. The baseline is trained on the target dataset, the real is trained on the CamVid dataset, and the synthetic is trained on synthetic data only.

	Cityscapes		
Model	Pixel Acc.	Class Acc.	IoU
Baseline	83%	77%	50%
Real	83%	77%	50%
Synthetic	**84%**	**79%**	51%
Mixed	**84%**	**79%**	**52%**

Table 1: Evaluation of different pre-training strategies.

Conclusion. Our results suggest that video games with photorealistic environments are potentially useful for a variety of computer vision tasks as they can offer an alternative way to compile large realistic datasets for training and evaluation.

[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.