

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

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

1 Evaluation with Fine-tuning

Figures 1-4 compare the behavior of our convolutional networks in the training phase. We present the evolution of *objective value*, *pixel classification accuracy*, *class average accuracy*, and *mean intersection-over-union* for CamVid, Cityscapes, CamVid+, and Cityscapes+ datasets. Pre-training on synthetic data consistently improves the initialization and the final solution, and in most cases also outperforms pre-training on real-world data.

Figure 5 compares the per-class accuracy of each training strategy on the test set of CamVid and the validation set of Cityscapes. Using synthetic data yields a consistent improvement over the baseline. On CamVid, pre-training on real data leads to a better model than pre-training on synthetic data, but the mixed approach has the best accuracy. On Cityscapes, however, pre-training on synthetic data has a higher average accuracy than pre-training on real-world data. Figure 6 shows the per-class accuracy on the Cityscapes+ dataset. Similar to the previous experiments using synthetic data results in more improvement than using real-world data. Combining synthetic and real data gives the highest performance boost in these experiments.

2 Cross-dataset Evaluation

In the cross-dataset setting, we train one network on each dataset and evaluate the accuracy of each network on the other datasets. The purpose of this experiment is to measure and compare the generalization power of the networks that are trained on synthetic or real data only. Figure 7 shows the per-class accuracy for evaluation on the Camvid+ dataset. The `Baseline` network is directly trained on the target dataset, while the `Real` network is trained on the alternative real dataset, and the `Synthetic` network is trained on synthetic data only. Without domain adaptation, both of the `Real` and `Synthetic` networks have a

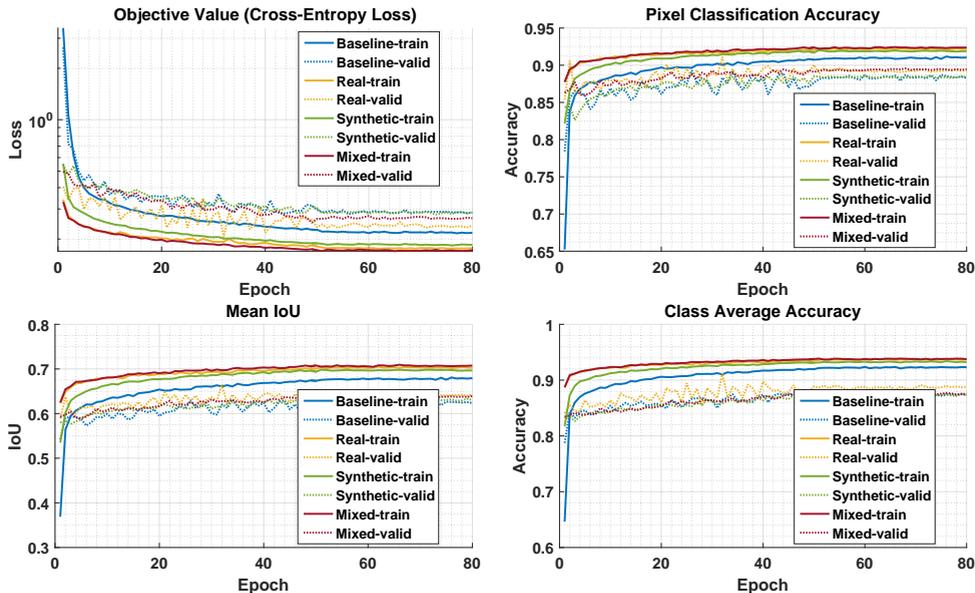


Figure 1: The influence of various pre-training approaches on the CamVid dataset. The solid lines are the evaluation results on the training set, the dashed lines are the results on the validation set.

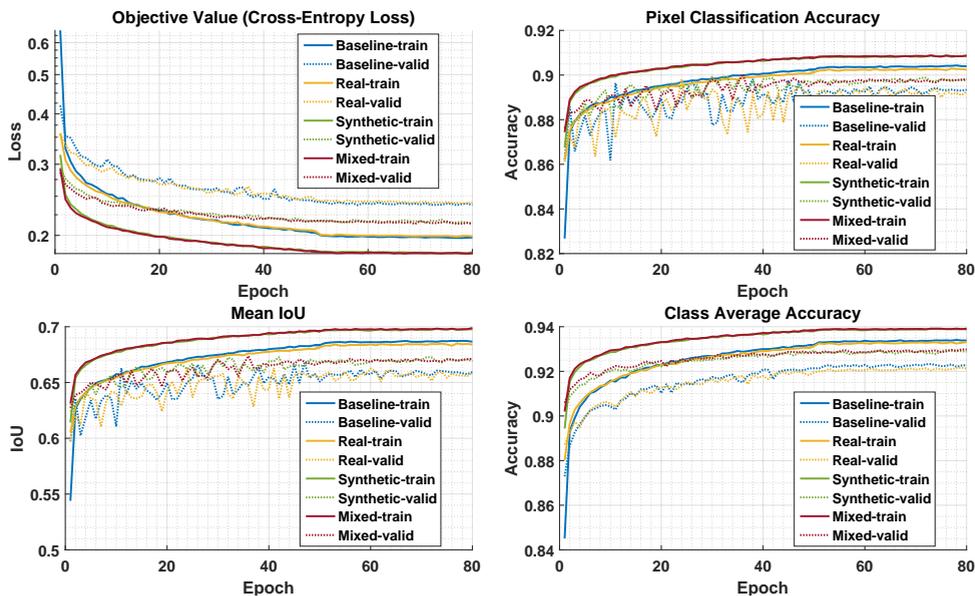


Figure 2: The influence of various pre-training approaches on the Cityscapes dataset. The solid lines are the evaluation results on the training set, the dashed lines are the results on the validation set.

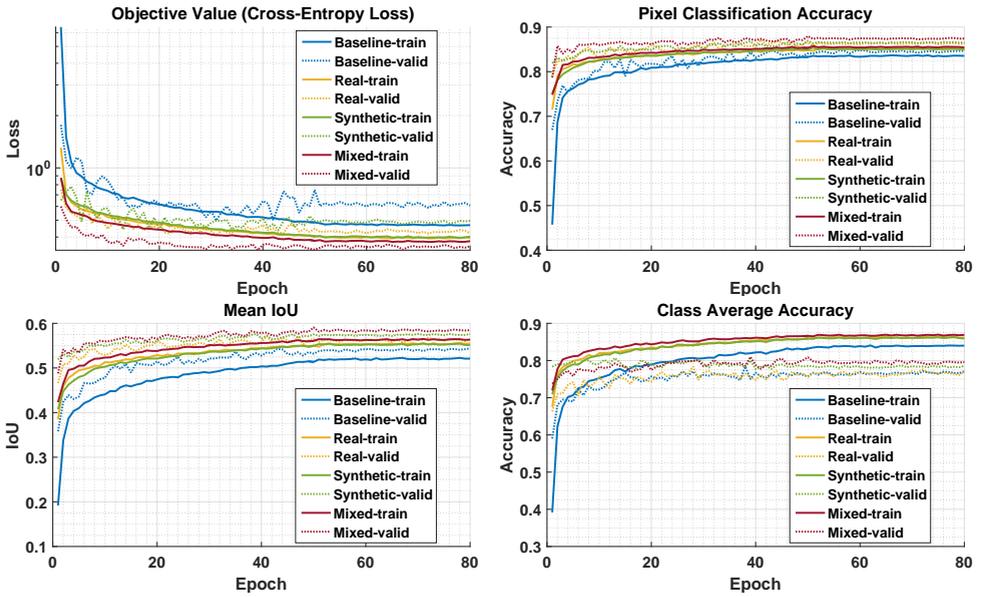


Figure 3: The influence of various pre-training approaches on the CamVid+ dataset. The solid lines are the evaluation results on the training set, the dashed lines are the results on the validation set.

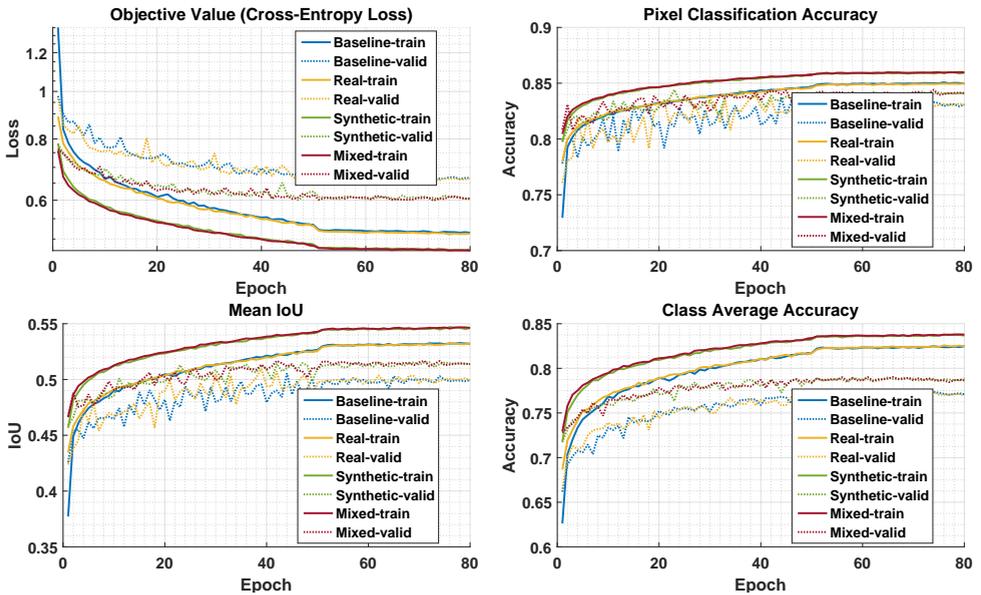


Figure 4: The influence of various pre-training approaches on the Cityscapes+ dataset. The solid lines are the evaluation results on the training set, the dashed lines are the results on the validation set.

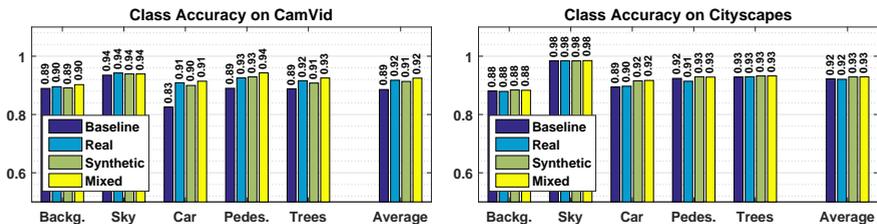


Figure 5: The per-class accuracy on CamVid (left) and Cityscapes (right).

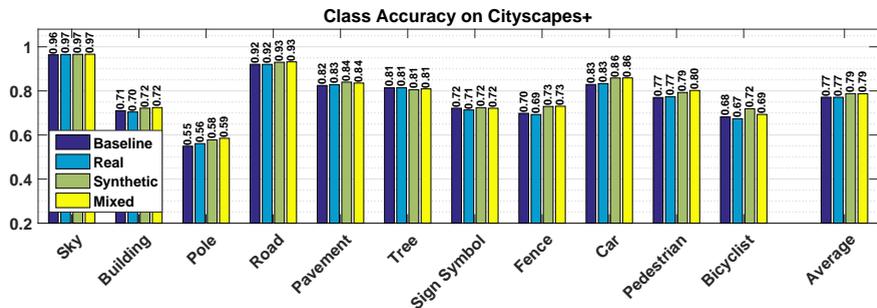


Figure 6: The per-class accuracy on Cityscapes+.

lower accuracy than the Baseline. The network that is trained on real data has a better accuracy than the network that is trained on synthetic data only. Even though the Synthetic network is only trained on synthetic data, it outperforms the real network on ‘Building’, ‘Pole’, and ‘Fence’. While the Synthetic network does not exceed the accuracy of the Real network on average, the small gap indicates that the network with synthetic data is relying on relevant features and is not merely overfitting to the game specific textures.

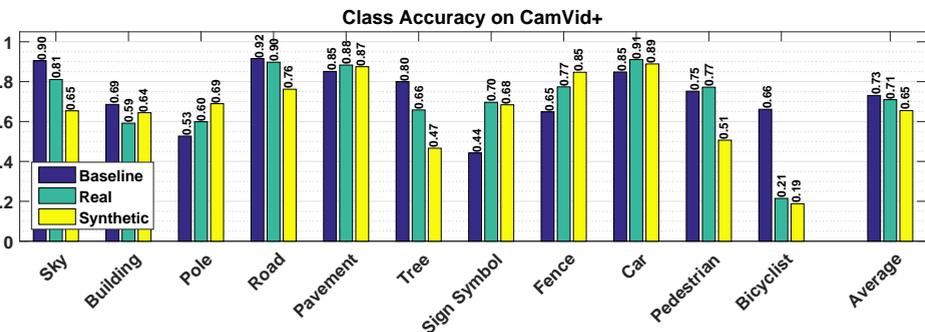


Figure 7: Cross-dataset evaluation. The per-class accuracy on the test set of the CamVid+ dataset.

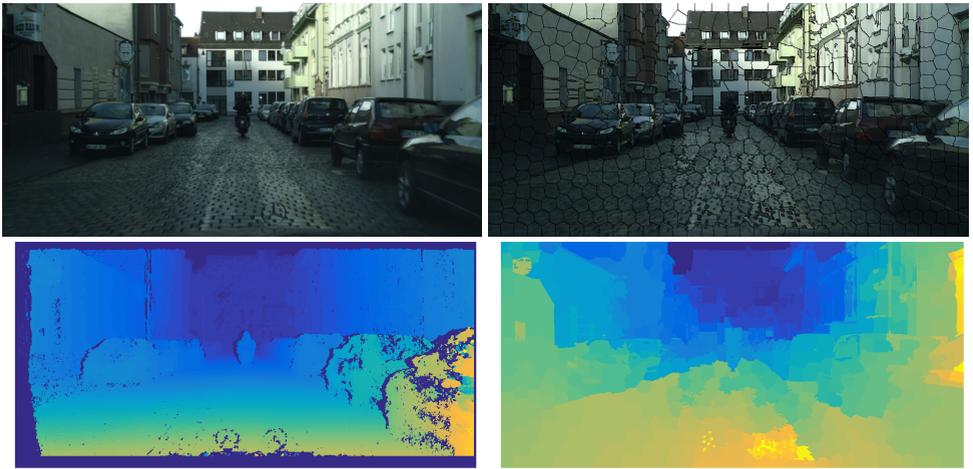


Figure 8: (top left) A sample image from the Cityscapes [1] dataset, (top right) decomposition of the RGB image to SLIC superpixels [2], (bottom left) the groundtruth disparity map, (bottom right) the globalized depth output of the method presented by Zoran *et al.* [3].

3 Depth Estimation from RGB

Zoran *et al.* [3] present a depth estimation method that only relies on the ordinal relationships between a set of image patch pairs. The image is first decomposed into SLIC [2] superpixels. A deep convolutional network classifies the ordinal relationship between the adjacent superpixels by generating a local relationship label $\{<, =, >\}$ with the corresponding probabilities. A quadratic program is then constructed to generate a total ordering (ranking) over the superpixels which will represent the depth. Note that this method does not rely on the depth measurement unit. Hence, it is not directly comparable to the prior work that directly regress to the depth value. We use this method because the depth information that is collected in the video game is not directly comparable to the real-world depth metrics. The ordinal relationships, however, can be consistently inferred from the extracted depth information. In the main paper, we demonstrated how using the synthetic RGB images can improve the patch classifier of Zoran *et al.* [3]. Figure 8 shows the groundtruth depth and the predicted depth image of a sample input from the Cityscapes dataset.

References

- [1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 2012.
- [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] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T. Freeman. Learning ordinal relationships for mid-level vision. In *ICCV*, 2015.