SUPPLEMENTARY MATERIAL OF **Self-Supervised Relative Depth Learning for Urban Scene Understanding**

Huaizu Jiang[1], Gustav Larsson[2], Michael Maire[2,3]
Greg Shakhnarovich[3], and Erik Learned-Miller[1]

[1] UMass Amherst {hzjiang, elm}@cs.umass.edu
[2] University of Chicago {larsson, mmaire}@cs.uchicago.edu
[3] TTI-Chicago gregory@ttic.edu

1 Evaluation on PASCAL VOC

We evaluate our pre-trained AlexNet on the PASCAL VOC 2007 classification and VOC 2012 segmentation datasets, following the protocols used in the self-supervised colorization approach [1]. Results can be found in Table 1. It can be seen that our approach achieves competitive results with state-of-the-art self-supervised methods on the classification task. Although on the dense pixel-wise labeling task (*i.e.*, semantic segmentation), our approach doesn't perform as well as colorization or context self-supervision, it performs significantly better than no pre-training (*i.e.*, weights are randomly initialized) on both tasks. Good classification and reasonable segmentation performance validates the effectiveness of our proposed approach.

Our self-supervised approach aims to induce better visual representations for urban scene understanding. It turns out that our approach also achieves competitive results on a generic visual recognition dataset (*i.e.*, PASCAL VOC).

2 Direct comparison with [7]

We train an AlexNet version of [7] using our CityDriving dataset.[4] Regarding camera intrinsic parameters, we assume the principal point is in the center of the image plane. We adopt focal lengths of CityScapes since the CityDriving dataset looks more similar to CityScapes. We then fine-tune the pre-trained network on three semantic segmentation datasets. As we can see in Table 2, our approach performs better on two out of three datasets.

Also, we note that the approach of [7] was not designed for self-supervised feature learning. No such experimental results were presented in [7] at all. Re-purposing [7] for self-supervised feature learning is novel.

[4] We replace the monocular depth estimation network with AlexNet and keep the camera pose network unchanged. We use the PyTorch implementation of https://github.com/ClementPinard/SfmLearner-Pytorch.

Table 1: PASCAL classification and segmentation results.

method	supervision	Classification (mAP)	Segmentation (mIoU)
supervised	ImageNet	**79.9**	**48.0**
none	-	53.3	19.8
tracking [2]	motion	58.7	-
moving [3]	ego-motion	54.2	-
context [4]	appearance	56.5	29.7
colorization [5,6]	color	**65.9**	**35.0**
Ours	relative depth	61.7	27.5

Table 2: Semantic segmentation results (mIoU) of AlexNet FCN32s using different self-supervised models. CS=CityScapes, K=KITTI, CV=CamVid.

method	supervision	CS	K	CV
Ours	rel. depth	**45.4**	**42.6**	53.4
Ours	abs. depth	44.2	39.8	51.8
SfM Learner [7]	reconstruction	43.9	42.5	**53.6**

3 Pre-training using ground-truth depth

We train an AlexNet using absolute depth ground-truth data provided in the KITTI dataset, captured using LiDAR. We then fine-tune it on the semantic segmentation benchmark dataset. Results are reported in Table 2. It can be seen that pre-training using relative depth consistently performs better on all three benchmark datasets.

Since ground-truth depth annotations are sparse and hard to capture, compared with our automatically recovered dense relative depth, we are only able to train the model using about 20K images, where we used 1.1M images for our relative depth. Although the self-supervised depth is noisy, the goal is not to achieve high performance on the proxy task. Rather, the eventual goal is to induce useful visual representations that can be transferred to downstream tasks.

References

1. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV. (2016)
2. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: ICCV. (2015)
3. Agrawal, P., Carreira, J., Malik, J.: Learning to see by moving. In: CVPR. (2015)
4. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.: Context encoders: Feature learning by inpainting. In: CVPR. (2016)
5. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: ECCV. (2016)

6. Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: CVPR. (2017)
7. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR. (2017)