Supplementary material for: Learning Where to Drive by Watching Others

Miguel A. Bautista, Patrick Fuchs, Björn Ommer

Heidelberg Collaboratory for Image Processing
IWR, Heidelberg University, Germany
firstname.lastname@iwr.uni-heidelberg.de

1 Ablation Studies

We have presented a self-supervised approach for the prediction of drivable areas in images. Our strategy makes use of large collections of unlabeled dashcam videos to teach a FCN which areas are drivable by watching others drive. We now analyze the impact of our contributions on the overall performance of our method by means of ablations. We evaluate several ablation methods: *(I)* Self-supervision via a fixed drivable area. Instead of obtaining self-supervision by tracking patches we fix a drivable area in front of the car bumper and collect self-supervision (e.g. Fig. 2(a) vs. 2(c)). *(II)* Single patch-based training of a binary CNN classifier. As opposed to a spatial-pyramid approach [6], we train a CNN for binary classification of drivable patches obtained via our tracking approach. *(III)* Training of a binary CNN classifier on a spatial-pyramid encoding of drivable patches [6]. *(IV)* Utilizing a FCN with dense up-stream convolutions for predicting pixel-wise labels of drivability obtained by self-supervision. *(V)* Our approach. In Tab. 1 we show different evaluation measures for all the ablation methods. We can see that the biggest performance improvement is obtained when comparing our approach with the fixed area self-supervision strategy, which does not track patches the other cars have driven over. In addition, we show that simple binary classification of drivable patches, even with spatial-pyramid encoding is not as successful as a FCN. Finally, using dilated convolutions gives us a broader context, which further improves results.

	No tracking (I)			Single patch (II)			Context-pyramid (III)			FCN dense upconv. (IV)			Ours (V) Sect. 3		
	UM	UMM	UU	UM	UMM	UU	UM	UMM	UU	UM	UMM	UU	UM	UMM	UU
MaxF	62.4	57.5	68.2	78.4	78.8	72.4	85.3	81.1	84.6	83.4	86.0	80.2	**90.9**	**87.5**	**88.2**
AP	45.7	47.4	55.0	84.7	85.9	79.0	91.2	91.2	91.8	83.5	87.7	81.8	**88.6**	**89.2**	**87.6**
PRE	65.8	72.5	71.9	77.4	82.0	70.1	83.5	83.4	83.8	78.9	83.0	78.7	**90.6**	**88.5**	**87.6**
REC	59.3	47.7	64.8	79.4	75.8	74.8	87.2	78.9	85.4	88.5	89.3	81.8	**91.3**	**86.6**	**88.7**
FPR	6.9	5.5	4.0	4.5	5.1	5.0	3.3	4.8	2.6	4.6	5.6	3.5	**1.8**	**3.4**	**2.0**
FNR	40.7	52.3	35.2	20.6	24.2	25.2	12.8	21.1	14.6	11.5	10.7	18.2	**8.7**	**13.4**	**11.3**

Table 1: Ablation experiments performed on KITTI [3].

2 Extended Quantitative Experimentation on KITTI

In addition to the experiments reported in the main submission, we also tested our approach on the KITTI [3] benchmark suite. In order to do so, we disregard the training labels provided by the benchmark and only use the 60 unlabeled video sequences provided with KITTI, utilizing just monocular color images. We then play the sequences backwards in time and generate the self-supervised labeling of drivable surfaces, gathering 42000 frames labeled with our self-supervision strategy.

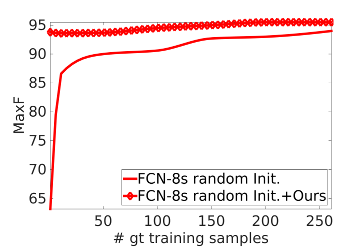
2.1 Zero-shot Learning

To assess the performance of our self-supervision method we tackle the problem of zero-shot learning of drivable areas on KITTI [3]. That is, methods are provided with 0 ground-truth labeled training images. We compare state-of-the-art fully convolutional architectures with and without our self-supervision method trained on the unlabeled sequences of KITTI. Tab. 1 summarizes the performance of two different architectures with and without our self-supervision method. We show results for our variant of FCN-8s [7] (with dilated upconvolutional layers), with and without Imagenet [2] pre-training. In addition, we also make use of the ResNet-101 model [4] pre-trained on Imagenet. In Tab. 1(a) we observe that our proposed approach for self-supervision drastically boosts the performance of zero-shot learning for all different architectures, with a performance improvement of at least 52%.

In addition to the zero-shot learning analysis we also show how our approach behaves when presented with few labeled samples, taking FCN-8s [7] as a particular instance. Fig. 1(b) shows how performance increases as a function of the number of labeled training samples. We see how our self-supervision training greatly amplifies the generalization capabilities of the network, consistently outperforming the same network without using our self-supervised pre-training

Model	UM	UMM	UU
FCN-8s Random Init.	25.5	36.8	22.4
FCN-8s Random Init. + Ours	90.7	85.8	87.0
FCN-8s Imagenet Init.	27.9	37.9	23.9
FCN-8s Imagenet Init + Ours	90.1	85.8	86.4
ResNet-101 Imagenet Init.	29.4	38.7	20.6
ResNet-101 Imagenet Init. + Ours	91.0	85.9	87.6

(a)



(b)

Fig. 1: (a) Zero-shot MaxF results for KITTI benchmark, where our model was trained on the unlabeled sequences of KITTI. (b) MaxF score as a function of the number of labeled ground-truth training samples. FCN-8s is trained from random weight initialization with and without our self-supervised pre-training.

Fig. 2: Sample score maps of drivable areas for zero-shot learning on KITTI.

Method	#gt labeled samples	UM	UMM	UU	ALL
MultiNet [10]	289	93.99	96.15	93.69	94.88
DDN [8]	289	93.65	94.17	91.76	93.43
Up-Conv-Poly [9]	289	92.20	95.52	92.65	93.83
FTP [5]	289	91.20	92.98	89.62	91.61
FCN-8s Random Init. + GT	289	89.50	92.81	84.50	89.83
FCN-8s Random Init. + Ours	0	87.39	86.14	84.96	85.74
Alvarez et. al [1]	1	73.69	86.21	72.25	79.02

Table 2: MaxF score for different method on the KITTI test server.

training. Finally, we show few score maps of drivable area yielded by our self-supervised approach on KITTI [3] in Fig. 2. Note that our method does not use any ground-truth labeled image during training.

To put our approach into context with state-of-the-art methods we report the results obtained by our self-supervised strategy on the test server of KITTI [3]. Since source code for top performing methods of KITTI is not available we take the widely used FCN-8s architecture as a study case. We then see that training FCN-8s using the KITTI ground-truth yields 5% worse performance that the top method. This situation is understandable since [10] is a more complex model than FCN-8s. To asses the quality of our approach we now train FCN-8s using self-supervision on the unlabeled KITTI video sequences, and compare it to FCN-8s trained on ground truth. We then see that the performance gap between using the KITTI training set and our self-supervised approach is 4%, despite using no labeled samples at all. In addition, we compare our approach with the one-shot method of Alvarez et. al [1] (which requires similar quantities of supervision as our approach) obtaining a performance improvement of 14% over it.

2.2 Transfer Learning

Conversely to Sect. 4.3 of the main submission in which we evaluate the potential of transferring a model trained on KITTI to Cityscapes, we now evaluate how a model trained on CityScapes transfers to KITTI.

The underlying rationale is that if a model is performing well on CityScapes it should also perform equivalently on KITTI. Therefore, we utilize the unlabeled sequences of KITTI for pre-training the FCNs using our self-supervised strategy, before using the CityScapes ground-truth labels to perform supervised learning. We evaluate transfer learning based on two separate network architectures, FCN-8s [7] and ResNet-101 [4]. In Tab. 3 we show the MaxF and IoU scores of the different models with and without our self-supervised pre-training. We can see that our self-supervised pre-training is extremely useful when transferring models between datasets, boosting performance by at least 10%. This performance improvement is due to the regularization properties of our self-supervision, which prevents the model from over-fitting to CityScapes-like scenarios, thus improving the capability to generalize to previously unseen scenarios.

3 Qualitative Results

In addition to the previous quantitative evaluation we also report qualitative results in the form of video sequences for different tasks.

3.1 Self-supervision

We now show how our training data is collected. We therefore take the unlabelled video sequences from KITTI and Cityscapes and apply our self-supervision strategy. To clearly illustrate our self-supervision approach we include few video sequences showing qualitative results in the folder `./self_supervision`. In these sequences, blue patches denote regions that have been driven over by the car equipped with the daschcam, while green patches are the ones driven over by other cars. Not drivable areas of the image are marked with red patches. Note how by playing videos back in time and tracking the patches that different cars have driven over, rich supervision can be obtained to learn which regions of an image are drivable.

3.2 Zero-shot learning

In addition, we also show how our FCN-8s performs when trained using our self-supervision strategy, without requiring tedious pixel-wise annotations of drivable areas. We collect few test sequences, which were not used for extracting

Model	MaxF	IoU
FCN-8s Random Init.	43.4	27.7
FCN-8s Random Init. + Ours	55.1	38.0
FCN-8s Imagenet Init.	50.1	33.4
FCN-8s Imagenet Init. + Ours	74.2	59.0
ResNet-101 Imagenet Init.	72.5	56.8
ResNet-101 Imagenet Init. + Ours	82.0	69.5

Table 3: Transfer Learning results from KITTI to Cityscapes benchmark.

self-supervision on neither KITTI or Cityscapes and let the network predict pixel-wise estimations of drivability. These zero-shot learning predictions can be found in `./zero_shot_pred`.

3.3 Difficult scenarios: Snow and Sand

Finally, we include two sample sequences where our classifier was trained to predict drivable areas on both on a road completely covered in snow and on a dessert trail. In order to do this we had two different instances of our network trained on several YouTube dashcam sequences with roads covered in snow, and in different sandy desert videos. We then show our results in two video sequences which were not used during the training process. Our goal is to illustrate that our method is not bounded to predict drivability on asphalt regions, but can learn a general notion of drivability when trained with suitable data.

References

1. Jose M Alvarez, Theo Gevers, Yann LeCun, and Antonio M Lopez. Road scene segmentation from a single image. In *Computer Vision–ECCV 2012*, pages 376–389. Springer, 2012.
2. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
3. Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.
4. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.
5. Ankit Laddha, Mehmet Kemal Kocamaz, Luis E. Navarro-Serment, and Martial Hebert. Map-supervised road detection. In *IEEE Intelligent Vehicles Symposium Proceedings*, 2016.
6. Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2169–2178. IEEE, 2006.
7. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.
8. Rahul Mohan. Deep deconvolutional networks for scene parsing, 2014.
9. Gabriel Oliveira, Wolfram Burgard, and Thomas Brox. Efficient deep methods for monocular road segmentation. 2016.
10. Marvin Teichmann, Michael Weber, J. Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. *CoRR*, abs/1612.07695, 2016.