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

Abstract. The most prominent approach for autonomous cars to learn what areas of a scene are drivable is to utilize tedious human supervision in the form of pixel-wise image labeling for training deep semantic segmentation algorithms. However, the underlying CNNs require vast amounts of this training information, rendering the expensive pixel-wise labeling of images a bottleneck. Thus, we propose a self-supervised approach that is able to utilize the myriad of easily available dashcam videos from YouTube or from autonomous vehicles to perform fully automatic training by simply watching others drive. We play training videos backwards in time and track patches that cars have driven over together with their spatio-temporal interrelations, which are a rich source of context information. Collecting large numbers of these local regions enables fully automatic self-supervision for training a CNN. The proposed method has the potential to extend and complement the popular supervised CNN learning of drivable pixels by using a rich, presently untapped source of unlabeled training data.

1 Introduction

The amount of video being recorded by dashboard cameras is increasing exponentially, thus becoming a potentially valuable source of training data for driver assistance and autonomous driving systems. However, the prevalent approach for many of these systems has been supervised learning of Convolutional Neural Networks (CNN) for semantic segmentation of traffic scenes [23,22,7]. A core sub-task of many of these systems is the detection of drivable surfaces (i.e. road detection to avoid lane departure or to plan driving trajectory), where tedious pixel-wise supervision of drivable areas needs to be collected in order to train a supervised model. While this supervision can improve the model performance on particular evaluation sub-sets [14,9], the statistics captured by these datasets may not be transferrable to other scenarios, e.g. a model trained on data collected in Germany [9] cannot be expected to perform equally in a UK based traffic scenario. Alternatively, scaling the labeling effort to the wide variety of traffic scenarios present all around the world is a futile undertaking.

Therefore, what we need is not simply more labelled data, as human annotators would always present a bottleneck to scale up learning. What we strive for is to enable learning algorithms to use the virtually unlimited number of dashcam videos available (YouTube, etc.), which are presently inaccessible for the current

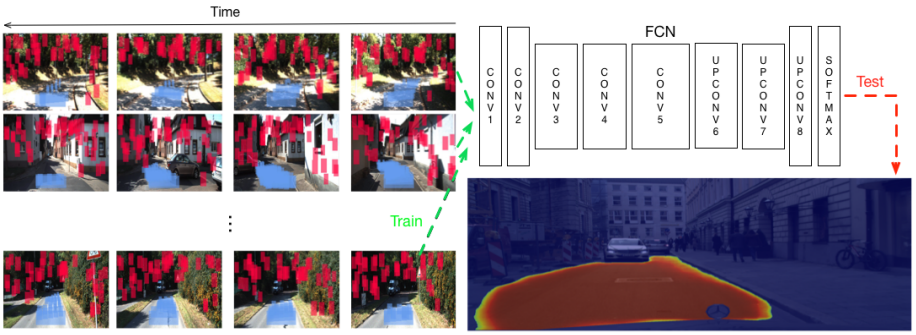


Fig. 1: Pipeline of the proposed approach for self-supervised learning of drivable surface. To obtain self-supervision, dashcam sequences are played back in time and patches that cars have driven over are marked as drivable and tracked. Using this self-supervision we propose to train an FCN that is able to effectively predict drivability of pixels without using any ground-truth labeled samples for training.

supervised learning methods since they are lacking labels altogether. For instance, the popular KITTI [14] and Cityscapes [9] datasets contain 290 and 25000 labeled training frames, respectively, compared to the potentially unlimited amount of unlabeled dashcam footage, which can be easily collected for all possible scenarios (i.e. different continents, countries, cities, weather conditions, etc.). Therefore, assuming that a supervised algorithm trained on datasets like [14,9] can be deployed in diverse real traffic scenarios is, at the very least, unrealistic. In addition, regardless of the amount of labeling effort and cost invested, the volume of unlabeled data will always be magnitudes larger.

A clear example are autonomous driving corporations like Tesla, Waymo, Uber, etc. where the competitive advantage is held by the corporation with more data collected. As an example, Tesla claims to have 1.3B miles of collected data, which to improve the autonomous capability of their cars has to be evaluated by experts and annotators. We hypothesize that a much more favorable situation would be to let an algorithm do the heavy-lifting and utilize the virtually infinite unlabeled dashcam footage collected every day. Moreover, the algorithm should not only learn from its own mistakes as in classical boosting but additionally from watching other cars drive. Moreover, large amounts of training data have become even more critical for the presently thriving deeper CNNs [31,18]. In recent years, there has thus been an increasing general interest in unsupervised training of CNNs on surrogate tasks [33,12] in order to exploit large amounts of unlabeled data that are infeasible to label. These unsupervised models can be then directly used as predictors [4,3] or further fine-tuned with only few labeled samples in order to compensate for the shift in both the task and data distribution [32].

The goal is then to learn a model for predicting drivable areas in an unsupervised manner, where we are only provided with unlabeled video sequences of a car interacting with other traffic partakers. Our hypothesis is the following: *can*

a CNN learn what areas are drivable by simply watching others? The motivation has its roots in Experiential Learning inspired by Kurt Lewin and colleagues [21]. Here, the key paradigm is to learn through reflection on a human performing a particular task [5]. Human beings can learn by reflecting on experience [26] collected by watching other humans performing a particular task. For example, when walking over a frozen lake, a person will find a safe path to walk, by watching other persons on the ice and following their path. Furthermore, when analyzing how human beings learn to drive, we observe that a driving instructor is present only for a very limited initial period of time of only a few hours. However, after driving lessons we continuously improve our driving skills by watching others drive and reflecting on that experience. In addition, a human driver may have acquired their skills in particular country with small variation in conditions, but when travelling to another region can quickly learn to adapt by watching other traffic partakers and learning by reflection. In this sense, while the detection of drivable areas in a supervised manner is successfully tackled by current machine learning approaches, the adaptation and improvement that human drivers are capable by experiential reflection on other drivers remains an unexplored problem.

Motivated by this observation we propose to extract self-supervision of drivable surfaces from large video collections recorded only using inexpensive monocular dashboard cameras and no other more sophisticated and expensive sensor modalities such as RADAR [24], LIDAR [35], stereo cameras [8], etc., since they are far less likely to find on, for example, YouTube. To accomplish this task, we play the training videos backwards in time and track patches so as to find regions that cars have driven over (including the car that the camera is in). Similarly, we obtain patches that are unlikely to be drivable. All gathered patches are then used for discriminative training of a Fully Convolutional Network (FCN), following up on recent advances for self-supervised learning of CNNs [33,12,1,10,6]. Obviously, we only play backwards the unlabeled training videos for extracting self-supervision. During testing, the FCN predicts drivable areas in an image without using any extra information and by only computing a single forward pass. A visual example of the proposed pipeline is shown in Figure 1.

We evaluate our approach for unsupervised learning of drivable areas on the widely used KITTI [14] and Cityscapes [9] datasets, using only unlabeled video sequences provided with each dataset. In addition, we also gathered a collection of dashcam videos from YouTube and used them to train our model. The proposed approach shows how meaningful representations can be extracted from large volumes of unlabeled traffic footage. The goal is obviously not to completely replace supervised CNN training, but to open up the potential of adding a rich, presently untapped source of unlabeled training data.

2 Related Work

In this section, we review drivable surface detection methods both for the supervised and unsupervised settings. However, for a more extended survey of road detection methods, the readers may refer to a recent survey work [19].

A lot of attention has been paid to supervised classification models for drivable surface detection. Guo et al. [16] formulated the road detection problem in a maximum a posteriori (MAP) framework, and used a Markov random field to solve this problem. In latter work, they also incorporated semantic information and applied a graphical model to infer the road boundary [17]. Furthermore, following recent results of CNNs, Alvarez et al. [2] used a CNN to extract the appearance feature of a patch and classify a road scene into the sky, the vertical regions and the drivable regions. Furthermore, with the advent of CNNs approaches that trained FCNs on large labeled datasets have obtained successful results [23,22,7], at the cost of requiring tedious pixel-wise labeling. To circumvent the high cost of labeling images at the pixel level, virtual datasets have recently gained a lot of attention [29,28,15], while such datasets provide inexpensive labeling they fail to encode the variability of different traffic scenarios and lack the degree of realism provided by videos recorded in the physical world.

In the avenue of unsupervised learning recent works have exploited spatial and temporal context to obtain supervisory signal for learning feature representations using CNNs, obtaining very satisfying results. In this sense, Wang et. al [33], showed that video can be used as supervisory signal by tracking objects over time, obtaining comparable results to supervised methods on Imagenet dataset [11]. In addition, Agrawal et. al [1] showed that ego-motion is a useful source of intrinsic supervision for visual feature learning in mobile agents. However, this approach minimized the error between the ego-motion information (i.e. camera transformation) obtained from its motor system and ego-motion predicted using its visual inputs only. This is not directly applicable to our case since the ego-motion information from the car motor system is not available.

3 Unsupervised Learning of Drivable Surfaces

In this section we describe our approach for unsupervised learning of drivable areas. Our goal is to learn an FCN for prediction of drivable surfaces in a completely unsupervised manner. However, standard training of FCNs requires huge amounts of labeled training data, which is infeasible to collect for the many different driving scenarios in which these systems are deployed. In order to circumvent this problem, we propose to generate self-supervision by experiential reflection on unlabeled video sequences.

3.1 Self-supervision by Experiential Reflection

Video sequences recorded by dashboard cameras contain unparalleled numbers of traffic scenes in which a car drives around while interacting with other road users,

which cannot be utilized by supervised learning methods. A simple approach to exploit this tremendous source of untapped traffic information would be to assume that a small fixed area in front of the car is drivable, generating self-supervision as the car moves during the video sequence. However, this generated self-supervision will only model the statistics of a small fixed area in front of the car bumper, neglecting the rich information provided by the scene and other traffic partakers. In addition, such approach will fail to model drivable areas far from the car, left and right turns or changes of drivable areas due to environmental causes (i.e. changes in lighting, shadows, road marks, etc.). A visual example of such self-supervision is shown in Fig. 2(a).

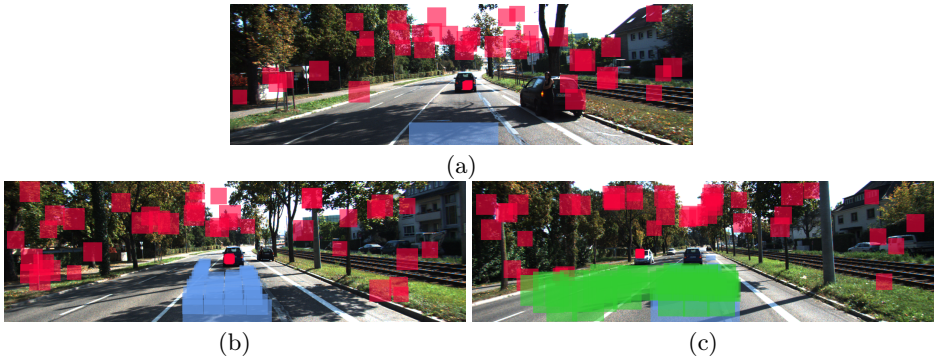
Interestingly, if such video sequences are played backwards in time, a human observer can easily point what areas have been driven over by different traffic partakers, and thus, can reflect on the experience (i.e. self-supervision) accumulated by watching such sequences to learn what makes an area drivable. Following this observation, we propose to automatically obtain self-supervision by rewinding sequences back in time and keeping track of surfaces, i.e., image patches, that different road users have driven over, reflecting on the experience obtained by watching others to collect supervision.

Given a training video sequence $\mathcal{S} = \{\mathbf{I}_1, \ldots, \mathbf{I}_t\}$, we obtain self-supervision by tracking patches that cars have driven over while playing the sequence backwards in time. To initialize the drivable patches to track, we can assume that a small area in front of a moving car bumper is a drivable area composed of several patches $\mathbf{P}_t^i \in \mathbb{R}^{h \times w \times 3},\ \forall i \in \{1, \ldots, p\}$. For the point of view of the car recording the sequence this is a trivial task, since the bumper position is fixed. To extend this assumption to the rest of cars in the sequence we simply obtain object proposals using a car detector [27] and assume that the area directly below the detection is a drivable surface. When rewinding the sequence we track these patches using optical flow [25]. Flow is computed densely between pairs of consecutive frames $(\mathbf{I}_{t-1}, \mathbf{I}_t)$ and used to estimate a projective transformation with RANSAC [13]. This transformation compensates for the ego motion and is used to establish correspondences between a patch $\mathbf{P}_{t-1}^i$ and its successor $\mathbf{P}_t^i$, therefore being able to track patches that different cars have driven over (cf. supplementary material for sample video sequences).

To take account of tracking errors (e.g. patch drift) and of the movement of other road users, we compute the similarity between a patch $\mathbf{P}_{t-1}^i$ and its projected successor $\mathbf{P}_t^i$ to eliminate unreliable correspondences: since the projective transformation is estimated using consecutive frames, we can assume that the color histogram of a patch $\mathbf{P}_{t-1}^i$ and its projected successor $\mathbf{P}_t^i$ is highly similar. Let $\mathbf{h}(\mathbf{P}_t^i)$ denote the normalized color histogram of patch $\mathbf{P}_t^i$. Then, given the pairwise similarities between two consecutive patches:

$$s(\mathbf{P}_{t-1}^i, \mathbf{P}_t^i) = \mathbf{h}(\mathbf{P}_{t-1}^i)^\top \mathbf{h}(\mathbf{P}_t^i), \tag{1}$$

We compute the distribution of distances and truncate it at the pivot point, thus effectively eliminating unreliable projected drivable patches (i.e. false positive patches which were projected to a highly dissimilar region) by stopping their



(a)

(b) (c)

Fig. 2: (a) Self-supervision obtained by fixing a drivable area in front of the car. (b) Self-supervision obtained using only the car that is recording the video sequence, where blue and red patches correspond to drivable and non-drivable surfaces, respectively. (c) Self-supervision computed using different traffic partakers, where blue patches correspond to drivable surfaces of the egocentric point of view, green patches denote self-supervision obtained from other cars and red patches correspond to non-drivable surfaces.

tracks, while reliable tracks are computed until the similarity of two consecutive patches lies below the pivot point. Drivable patches are tracked for 63 frames on average with a maximum of 191 frames.

To obtain negatives, i.e., non-drivable patches, we randomly sample patches from image regions which were not driven over by any car and whose tracks consisted for more than ten frames (to avoid drifting patches or those occluded by other objects). In addition, we place negative patches inside the car bounding boxes computed in the previous step to include other cars in the negative training set. A visual example is shown in Fig. 2(b) where red and blue patches denote non-drivable and drivable surfaces, respectively. Furthermore, Fig. 2(c) also shows in green the self-supervision obtained by tracking patches that other road users drove over.

3.2 Learning Drivable Surfaces

CNNs are presently among the most powerful classification frameworks for Computer Vision [31]. Since our self-supervision strategy generates patches of drivable and non-drivable areas on an image, a natural choice would be to train a CNN to classify these image patches [20]. However, since our ultimate objective is to predict the drivability of individual pixels, such approach has two shortcomings: *(i)* estimating the drivability of a pixel using only local patch information is a hard task due to the lack of context information. *(ii)* testing is computationally prohibitive since all patches in an image need to be evaluated.

A more efficient approach is to cast the problem as pixel-wise labeling for each training image using the available self-supervision. All pixels contained in

a drivable training patch are considered positive and all pixels contained in a non-drivable patch are negative. Then, a FCN architecture [23,22,7] is trained for predicting a pixel-wise label for each training image, where pixels which are not labeled as positive nor negative during the self-supervision step are not used during training. FCN architectures naturally incorporate context which is encoded by means of convolution and pooling operations, while being extremely efficient during testing due to their weight sharing scheme. However, a common problem of FCNs is to model long-range contextual interactions between pixels. In our particular case these long-range interactions encode extremely useful context that helps modelling what makes up a drivable area (e.g. drivable pixels on the correct side of a curb, road marks, etc.). Therefore, to include these long-range interactions we use dilated convolutions [34] through all our up-stream convolutional layers.

Once our FCN is trained, we compute a confidence map for each training image. However, thresholding these confidence maps to obtain predictions would not take into account the labeling of neighbor pixels. Therefore, we use GrabCut [30] to obtain a discrete labeling that aggregates local context at the pixel-wise label level, using this new self-supervised ground truth we further fine-tune our FCN and repeat this process for three iterations. Finally, we need discrete output prediction of drivability. Thus, provided the accuracy of the probability heat-maps yielded by our iterative approach, we simply threshold these confidence maps. Computing a confidence map for an image only requires a forward-pass, making it very effective for real-time deployment[1].

4 Experimental Results

In this section, we present both quantitative and qualitative results obtained by the proposed unsupervised learning approach on the main benchmarks for pixel-wise drivable surface detection, Cityscapes [9]. Our hypothesis is two fold: (i) Our self-supervised approach should perform well in zero-shot learning scenarios, where no labeled samples are provided for training. (ii) Self-supervised training should serve as a regularizer that helps to generalize when transferring between similar datasets. Extensive experimental results can be found in the supplementary material. All models, datasets and generated labels are publicly available at[2].

4.1 Datasets

Cityscapes Cityscapes [9] is the biggest dataset available for semantic segmentation of traffic scenes. Cityscapes contains 25000 fully labeled images at pixel-level. For evaluation purposes we only utilize the *road* category. Cityscapes [9] does not provide the video sequences of their vehicles driving around different cities, and only allowed us to use three sequences from a single city containing 30000 frames in total to utilize our self-supervised method.

[1] Our approach runs at 15 FPS on a NVIDIA Titan X GPU.

[2] `https://hcicloud.iwr.uni-heidelberg.de/index.php/s/tutGQ2J3XoUyqkU`

YouTube Dashcam Dataset To further evaluate the generality of our approach, we additionally acquired 13 dashcam sequences totalling 100000 frames that were recorded in the wild and uploaded to YouTube. These videos were all recorded in different cities from Germany and the United States showing varying weather conditions and seasons, therefore spanning a large variety of scenarios. Furthermore, we also collected two more sequences with difficult conditions: a dusty desert trail and a road covered in snow and mud, where each sequence contains 10000 frames.

4.2 Zero-shot Learning

To assess the performance of our self-supervision method we tackle the problem of zero-shot learning of drivable areas on CityScapes [9]. 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 CityScapes (cf. Sect. 4.1). 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 [23] (with dilated upconvolutional layers), with and without Imagenet [11] pre-training. In addition, we also make use of the ResNet-101 model [18] pre-trained on Imagenet. In Tab. 1 we observe that our proposed approach for self-supervision drastically boosts the performance of zero-shot learning for all different architectures, where our self-supervision computed on the unlabeled sequences of Cityscapes is extremely helpful for both randomly initialized and pre-trained models. Note that a randomly initialized model with our self-supervision strategy is able to attain equivalent performance than the same model pre-trained on Imagenet [11]. Thus, being able to circumvent the use of the 1.2M of Imagenet labeled samples. To the best of our knowledge, this is the first time that a self-supervised method that performs equivalently in the absence of the widely adopted Imagenet pre-training strategy.

Model	MaxF	IoU
FCN-8s Random Init.	49.5	32.9
FCN-8s Random Init. + Ours	82.6	70.4
FCN-8s Imgnet. Init.	51.3	34.7
FCN-8s Imgnet. Init. + Ours	81.9	69.4
ResNet-101 Imgnet. Init.	52.5	34.9
ResNet-101 Imgnet. Init. + Ours	79.6	66.1

Table 1: Zero-shot results for Cityscapes benchmark.

Finally, we show few score maps of drivable area yielded by our self-supervised approach Cityscapes [9] in Fig. 3. Note that our method does not use any ground-truth labeled image during training.

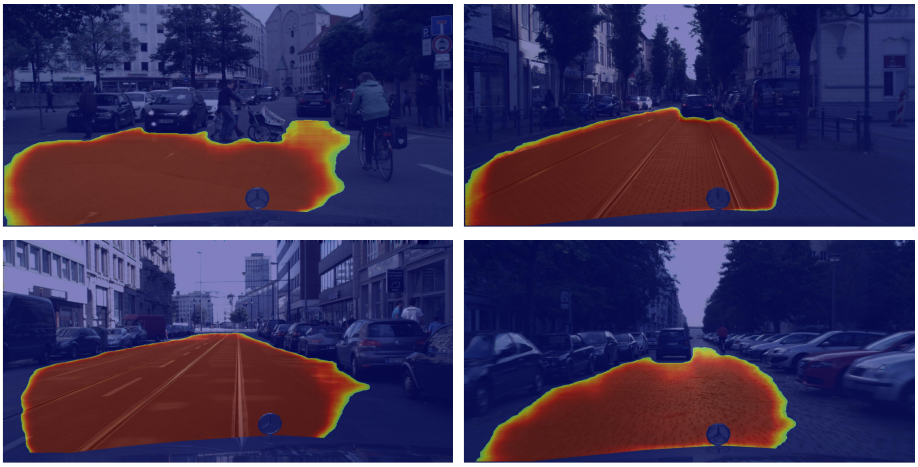


Fig. 3: Sample score maps for zero-shot learning on Cityscapes.

4.3 Transfer Learning

We now evaluate the ability of our approach to boost performance on a transfer learning task where a model is trained on one dataset and then transferred to another for evaluation. We employ the two most prominent datasets for estimating drivable areas, KITTI [14] and Cityscapes [9]. The underlying rationale is that if a model is performing well on KITTI it should also perform equivalently on Cityscapes. Therefore, we utilize the unlabeled sequences of Cityscapes (cf. Sect. 4.1) for pre-training the FCNs using our self-supervised strategy, before using the KITTI ground-truth labels to perform supervised learning. After training this model is then evaluated on Cityscapes. We evaluate transfer learning based on two separate network architectures, FCN-8s [23] and ResNet-101 [18]. In Tab. 2 we show the MaxF and IoU scores on Cityscapes of the different models transferred from KITTI with and without our self-supervised pre-training (first two columns, denoted with KITTI-TF). 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 KITTI-like scenarios, thus improving the capability to generalize to previously unseen scenarios.

4.4 Self-supervision from the Wild: YouTube Dashcam dataset

After providing results using the most prominent datasets for predicting drivable areas, we now study the problem of collecting self-supervision from the wild. We therefore utilize the YouTube Dashcam dataset described in Sect. 4.1, collecting self-supervision for 100000 frames of video data from different locations under different environmental conditions. For clarity, we follow the evaluation protocol

for Cityscapes and evaluate the suitability of models trained with self-supervision from the wild for zero-shot learning. Tab. 2 reports the result of our self-supervised YouTube training (last two columns, denoted with YT-SS) where we show results for FCN-8s [23], with and without Imagenet [11] pre-training. In addition, we also make use of the ResNet-101 model [18] pre-trained on Imagenet. We can therefore observe how the self-supervision that we collected from video sequences from the wild is of tremendous power, boosting performance by 18% at least.

Model	MaxF/KITTI-TF	IoU/KITTI-TF	MaxF/YT-SS	IoU/YT-SS
FCN-8s Rand.	43.4	27.7	49.5	32.9
FCN-8s Rand. + Ours	55.1	38.0	70.6	54.5
FCN-8s Imgnet.	50.1	33.4	49.5	32.9
FCN-8s Imgnet Init. + Ours	74.2	59.0		
ResNet-101 Imgnet.	72.5	56.8	49.5	32.9
ResNet-101 Imgnet. + Ours	82.0	69.5	75.8	61

Table 2: Transfer learning results on Cityscapes when training using self-supervision from KITTI and from Youtube Dashcam dataset.

5 Conclusions

In this paper we have presented a self-supervised approach for learning drivable regions from unlabeled dashcam videos. Based on experiential learning we aim to learn about drivable areas by watching others drive. Our simple, yet effective method makes large amounts of unlabeled training videos usable for training standard FCNs for pixel-wise predictions. Playing unlabeled dashcam sequences backwards in time and tracking patches the other cars have driven over allows us to gather large amounts of self-supervision which can be successfully leveraged by an FCN. For comparison and reproducibility, we train and evaluate on both KITTI and Cityscapes datasets obtaining competitive results for zero-shot learning tasks, where no ground-truth labeled image samples are provided for training. In addition, we introduce a novel dataset of dashcam sequences from YouTube where we collected self-supervision from 100000 frames and show that we can obtain valuable self-supervision from such an unconstrained source of video. The results obtained by the proposed approach show that powerful pixel-wise predictors of drivability can be learnt from large unlabeled video collections and demonstrate the potential of adding this rich, previously inaccessible resource to supervised FCN learning.

References

1. Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 37–45, 2015.

2. 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.
3. Miguel A Bautista, Artsiom Sanakoyeu, and Björn Ommer. Deep unsupervised similarity learning using partially ordered sets. *In Proceedings of IEEE Computer Vision and Pattern Recognition*, 2017.
4. Miguel A Bautista, Artsiom Sanakoyeu, Ekaterina Tikhoncheva, and Bjorn Ommer. Cliquecnn: Deep unsupervised exemplar learning. In *Advances In Neural Information Processing Systems*, pages 3846–3854, 2016.
5. Evelyn M Boyd and Ann W Fales. Reflective learning key to learning from experience. *Journal of Humanistic Psychology*, 23(2):99–117, 1983.
6. Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating accidents in dashcam videos. In *Asian Conference on Computer Vision*, pages 136–153. Springer, 2016.
7. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.
8. Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2015.
9. 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.
10. Hendrik Dahlkamp, Adrian Kaehler, David Stavens, Sebastian Thrun, and Gary R Bradski. Self-supervised monocular road detection in desert terrain. In *Robotics: science and systems*, volume 38. Philadelphia, 2006.
11. 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.
12. Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.
13. Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
14. 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.
15. Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. *CoRR*, abs/1605.06457, 2016.
16. Chunzhao Guo, Seiichi Mita, and David McAllester. Mrf-based road detection with unsupervised learning for autonomous driving in changing environments. In *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pages 361–368. IEEE, 2010.
17. Chunzhao Guo, Takayuki Yamabe, and Seiichi Mita. Robust road boundary estimation for intelligent vehicles in challenging scenarios based on a semantic graph. In *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pages 37–44. IEEE, 2012.

18. 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.
19. Aharon Bar Hillel, Ronen Lerner, Dan Levi, and Guy Raz. Recent progress in road and lane detection: a survey. *Machine Vision and Applications*, 25(3):727–745, 2014.
20. Junqi Jin, Kun Fu, and Changshui Zhang. Traffic sign recognition with hinge loss trained convolutional neural networks. *Intelligent Transportation Systems, IEEE Transactions on*, 15(5):1991–2000, 2014.
21. David A Kolb. *Experiential learning: Experience as the source of learning and development*. FT press, 2014.
22. Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.
23. 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.
24. Meng Lu, Kees Wevers, and Rob Van Der Heijden. Technical feasibility of advanced driver assistance systems (adas) for road traffic safety. *Transportation Planning and Technology*, 28(3):167–187, 2005.
25. Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. *IJCAI*, 81:674–679, 1981.
26. Andrew N Meltzoff and Rechele Brooks. Self-experience as a mechanism for learning about others: a training study in social cognition. *Developmental psychology*, 44(5):1257, 2008.
27. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
28. Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV 2016 – Proceedings of the 14th European Conference on Computer Vision – Volume Part II*, pages 102–118, 2016.
29. Germán Ros, Laura Sellart, Joanna Materzynska, David Vázquez, and Antonio M. Lopez. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3234–3243, 2016.
30. Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.
31. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
32. X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015.
33. Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.
34. Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
35. Wende Zhang. Lidar-based road and road-edge detection. In *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pages 845–848. IEEE, 2010.