FCNs for Free-Space Detection with Self-Supervised Online Training

Willem P. Sanberg, Gijs Dubbelman and Peter H.N. de With

Abstract— **Recently, vision-based Advanced Driver Assist Systems have gained broad interest. In this work, we investigate free-space detection, for which we propose to employ a Fully Convolutional Network (FCN). We show that this FCN can be trained in a *self-supervised* manner and achieve similar results compared to training on manually annotated data, thereby reducing the need for large manually annotated training sets. To this end, our self-supervised training relies on a stereo-vision disparity system, to automatically generate (weak) training labels for the color-based FCN. Additionally, our self-supervised training facilitates *online* training of the FCN instead of offline. Consequently, given that the applied FCN is relatively small, the free-space analysis becomes highly adaptive to any traffic scene that the vehicle encounters. We have validated our algorithm using publicly available data and on a new challenging benchmark dataset. Experiments show that the online training boosts performance with 5% over offline training, both for $F_{\max}$ and AP.**

I. INTRODUCTION

In recent years, much research has been dedicated to developing vision-based Advanced Driver Assist Systems (ADAS). These systems help drivers in controlling their vehicle by, for instance, warning against lane departure, hazardous obstacles in the vehicle path or a too short distance to the preceding vehicle. As these systems evolve with more advanced technology and higher robustness, they are expected to increase traffic safety and comfort. A key component of ADAS is free-space detection, which provides information about the surrounding drivable space. In this work, we employ a Fully Convolutional Network (FCN) for this task and explore *online* training in a *self-supervised* fashion, to increase the robustness of the free-space detection system.

Neural nets with deep learning are becoming increasingly successful and popular for image analysis. In the field of Intelligent Vehicles, many of the recent state-of-the-art algorithms rely on neural nets, mostly on Convolutional Neural Nets (CNNs). They excel in a wide variety of ADAS applications, such as stereo disparity estimation, object detection for cars and pedestrians and road estimation, as can be seen in the corresponding KITTI evaluation tables[1].

In literature, training a neural net typically requires many data samples for proper convergence of the large amount of parameters and proper generalization of the classifier. Different strategies are adopted throughout the field to handle this. For image recognition and object detection problems in natural environments, a common method is to start with a net that is trained on a large and generic dataset and adapt it to the task at hand [1][2].

Scene labeling, in contrast to scene or object recognition, requires a per-pixel classification. Recently, Fully Convolutional Networks (FCNs), have been employed for this task [3]. An FCN is a Convolutional Neural Network without any fully connected layers (they can be replaced by their convolutional counterpart). This adaptation transforms the net into a deep filter that preserves spatial information, since it only consists of filtering layers that are invariant to translation. FCNs have several attractive properties for scene parsing. For example, FCNs have no constraints on the size of their input data and execute inference efficiently in a single pass per image, thereby avoiding analyzing sliding windows, or, for instance, regions [4].

Even though weakly- or unsupervised training methods of CNNs are improving, they are currently still outperformed by fully supervised methods [2][5]. Together with the fact that creating large amounts of pixel-accurate training labels is inherently much work, we propose a middle-way in this paper: self-supervised training. If training labels can be generated automatically, the amount of supervised training data available becomes practically unlimited. However, this leads to a paradox, since it requires an algorithm that can generate the labeling, which is exactly the issue that needs to be solved. Therefore, we propose to rely on an algorithm based on traditional (non-deep learning) computer vision methods. This algorithm needs not to be perfect but at least sufficiently good to generate weak training labels. The goal is then that the FCN, trained with these weak labels, outperforms the traditional algorithm.

For next-generation ADAS, stereo cameras and multi-view cameras are an increasingly used sensor configuration. Stereo cameras provide insight into the geometry of the scene by means of the stereo disparity signal, which is valuable information for free-space detection. A state-of-the-art algorithm to distinguish free space and obstacles is the Disparity Stixel World [6]. We will use this algorithm to generate free-space masks and exploit these as weak training labels, and we will rely on the generalization power of FCNs to deal with the errors in the weak labeling. In essence, we use a stixel-based disparity vision system to train a pixel-accurate free-space segmentation system, based on an FCN, and refer to this as self-supervised training.

As a further contribution, our proposed self-supervised training is enhanced by combining it with the aforementioned strategies of task-specific fine-tuning of neural nets. Since traffic scenes come in a wide variety (urban versus rural,

Willem Sanberg, Gijs Dubbelman and Peter de With are with the Department of Electrical Engineering, Video Coding and Architectures Research Group, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands `w.p.sanberg@tue.nl`

[1] `http://www.cvlibs.net/datasets/kitti/index.php`.

highway versus city-center), with varying imaging conditions (good or bad weather, day or night), ADAS have to be both flexible and robust. A potential strategy is to train many different classifiers and to select the one that is most relevant at the moment (e.g., based on time and geographical location), or train a complex single classifier to handle all cases. In contrast, we show in this paper that it is feasible to fine-tune a relatively simple, single classifier in an online fashion. This is obtained by using the same self-supervised strategy as for offline learning, namely, based on generally correct segmentation by the disparity Stixel World. This results in automatically improved robustness of the free-space detection, as the algorithm is adapted while driving.

Considering the overall approach, our work is also related to [7], where automatically generated labels are exploited to train a CNN for road detection, which is applied as a sliding-window classifier. They also have an online component, which analyzes a small rectangular area at the bottom of the image (assumed road) and calculates a color transform to boost the uniformity of road appearance. The results of offline and online classifications are combined with Bayesian fusion. Our proposed work differs in several key points. Firstly, we do not need to assume that the bottom part of a image is road in the online training step, which is often an invalid assumption in stop-and-go traffic, since we exploit the stereo disparity as an additional signal. Secondly, their offline and online method is a hybrid combination of supervised and hand-crafted features, whereas our method can be trained and tuned in a fully end-to-end fashion, using a single FCN, while avoiding an additional fusion step. Thirdly, we do not require a sliding window in our inference step, since we use an FCN and not a CNN.

A comprehensive version of this extended abstract is available online[2].

II. METHOD

A. Fully Convolutional Network (FCN)

The color-based segmentation algorithm used as a basis of our work is an FCN [3]. For our experimentation, we have relied on the CN24 framework as described in [8]. Provided that the context (road detection) and data (images captured from within a vehicle [9]) are comparable to our research, we adopt their network architecture and their recommendations about the optimal training strategy. The network consists of several convolutional, max pooling and non-linear layers: Conv ($7 \times 7 \times 12$); MaxP (2×2); ReLU; Conv ($5 \times 5 \times 6$); ReLU; Full ($48\times$); ReLU; Full ($192\times$) + spatial prior; ReLU; Full ($1\times$) + tanh. The fully connected layers are interpreted and executed as convolutional layers by the CN24 library.

Note that our current work is not meant to offer an exhaustive test on optimizing the network architecture or hyper parameters. Our results may be improved by investigating that more properly, but the focus in this paper is to show the feasibility of self-supervised training and the additional benefits of our proposed online tuning in the context of free-space segmentation.

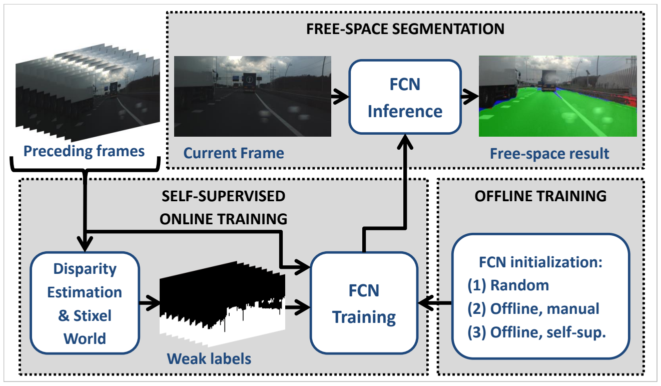

Fig. 1. Schematic overview of our free-space detection method with online self-supervised training.

B. Self-Supervised Training

Self-supervised training requires an algorithm that generates (weak) training labels. The training label generation algorithm is chosen to be an independent algorithm, which exploits an additional signal modality, namely stereo disparity. This disparity-based algorithm estimates masks of the drivable surface. As these masks are not perfect, we say they represent *weak* training labels. The reason to use a different modality (disparity) as basis for the weak training labeling than the modality (color) that is analyzed by the FCN, is to increase the chance that the trained algorithm can correct unavoidable errors in the weak labels, instead of stepping into the same pitfalls in difficult situations.

Stereo disparity is an attractive modality, since it is computationally inexpensive and yet provides relevant information in the context of free-space detection. We propose to analyze the disparity signal with the disparity Stixel World algorithm. This is a probabilistic framework that segments traffic scenes into vertically stacked, rectangular patches that are labeled as either ground or obstacle. These regions subsequently serve as the weak labels for the corresponding color image in our self-supervised training process.

The challenge is that the generated weak labels will contain errors, as is visible in the second column of Figure 2, potentially hampering the training process. We rely on the generalization power of the FCN training process to deal with these inconsistencies in the labeling, which we can validate by comparing the results of our self-supervised training with the results of training on manually annotated frames.

C. Online Training

For online training, we adopt the training strategies as introduced in [10]. In that work, the stereo-disparity signal is analyzed for several frames, and the resulting segmentation labels are exploited to construct a color model of ground and obstacle regions. The color model is exploited to segment a new frame in the sequence with their color Stixel World algorithm. Additional experiments over different color spaces

[2] https://arxiv.org/abs/1604.02316

showed that no single space is optimal for all frames [11]. In other words, their color representation can be potentially improved by adapting it better to the imaging circumstances. Building further upon that observation, we propose to apply end-to-end learning in this work to exploit an FCN training algorithm for finding the representation of the image that is most relevant in the current situation.

A schematic overview of our experimental framework for free-space detection is shown in Figure 1. We train an FCN from scratch (with random initialization), or start with one of the offline trained models and tune the entire model with online data. By comparing these online strategies with results from solely offline training, we show the importance and added value of adapting the classifier online to the changing environment. If this adaptation can be realized in a reliable and realistic way, our free-space detection system improves without putting extra effort and computational power into training and executing a larger, more advanced FCN. By limiting the complexity of our system, real-time execution in a driving car becomes feasible in the near future.

III. DATA AND EXPERIMENTS

We utilize two publicly available datasets as the training set for our offline training of the FCN (188 frames with manual annotation and its 10 unlabeled preceding frames)[10]. For our test set, we employ newly annotated data[3] that was captured in a similar configuration. It consists of 265 hand-annotated frames (and 10 unlabeled preceding frames) of urban and highway traffic scenes, both under good and adverse imaging conditions. There is a large variety in scenes, covering crowded city centers, small streets, large road crossings, road-repair sites, highways, etc.

A. Experiment 1: Supervised versus Self-Supervised Training

To validate the feasibility of our self-supervised training method, we compare three FCNs that have an equal architecture but are trained with different data. The first model is trained offline on manually annotated labels, as a reference result for offline, supervised training. The second model is trained offline on the same frames but now using automatically generated weak labels instead of the manual version. This model serves as a demonstration of offline, self-supervised training. Thirdly, we train a model in a self-supervised fashion on *all* available frames in the dataset, including frames for which no manual labels are provided. This experiment tests the added value of training on additional data in our framework, which is realized efficiently because of the initial choice of fully self-supervised training.

B. Experiment 2: Offline versus Online Training

We perform three key experiments to test the benefits of online training for our FCN-based free-space detection and compare this to the offline experiments of Section III-A. Similar to Experiment 1, we train on different data while the architecture of our FCN is kept identical. First, we train an FCN from scratch (with random initialization) on the weakly labeled preceding frames of each test frame. Additionally, we validate the benefits of online tuning. To this end, we initialize the net of each training sequence with one of the offline trained models (trained on either manual or self-supervised labels). Note that the labels for the online training itself are always self-supervised, since the preceding frames of each sequence are not manually annotated.

Furthermore, we perform an experiment to show the power and benefit of *'over-tuning'* for our framework. To this end, we test the online trained FCNs on test frames of different sequences than of the ones for which they were trained. By doing so, we can investigate the extent to which the online trained FCNs are tuned to their specific sequence. If the FCNs are over-tuned, we expect them to perform well if the training sequence and the test frame are aligned, but simultaneously expect them to perform poorly when they are misaligned. To validate this, we conduct three different misalignment experiments: shift one training sequence ahead or one back, and randomly permutate all training sequences. Note that our data sequences are ordered in time, therefore, there can still be correlation between training sequences and test frames when shifting back or forth a single training sequence. We reduce this correlation as much as possible by randomly permutating all training sequences.

IV. RESULTS AND CONCLUSIONS

Figure 2 shows qualitative results of our experiments. In general, the offline-trained FCN detects less false obstacles than the Stixel World baseline. However, it misses the trailer and part of the lamppost and it detects false obstacles on some shadows. In contrast, our online training outperforms both the Stixel World baseline and the offline methods. It segments the scene with raindrops on the car windscreen robustly and it classifies the trailer and shadows correctly.

We adopt the quantitative pixel metrics as employed for the KITTI dataset: $F_{\max}$ (an indication of the optimal performance) and the Average Precision AP, which captures the Precision score over the full range of Recall [9].

The trends of our quantitative results over the number of training iterations are shown in Figure 3. The training converges after 5,000 to 10,000 iterations. For offline learning, the results of supervised (manual labels) and self-supervised (disparity-based labels) are nearly identical. This confirms

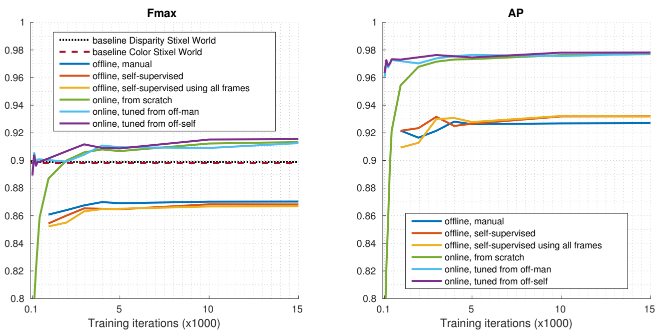

Fig. 3. Fmax and AP convergence over training iterations.

[3] available at `http://www.willemsanberg.net`.

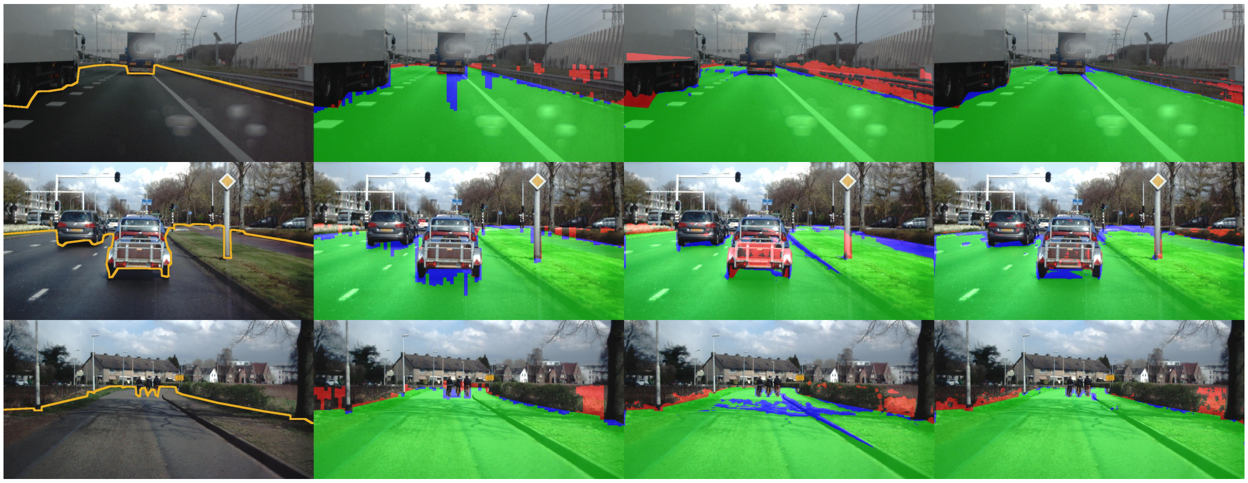

Fig. 2. Qualitative results on the frames shown in the leftmost column with hand-annotated freespace boundary. In the next columns, from left to right: Disparity Stixel World baseline; our result with offline training on manual labels; our result with online tuning. Colors indicate the freespace detections true positives (green), false negatives (blue) and false positives (red). Best viewed in color.

the feasibility of self-supervised learning, as relying on weak labels does not hamper the performance of our system. Self-supervised training on more data did not lead to a clear improvement of our results, as illustrated by the graph. This may show that our network is too small to exploit the additional data, or that the correlation within the new samples is too high to be informative. Regarding our online training strategies, Figure 3 shows that these outperform the offline training by 5%, both for $F_{\max}$ and AP.

An important conclusion of the experiments is that the contribution of online-tuned training is most significant in the speed of convergence, and less relevant for the final result after convergence. Specifically, the tuned models outperform the other methods already after 100 iterations of training (which takes less than half a second on a GeForce GTX970 graphics card), whereas models trained from scratch need at least 500 iterations to match the offline FCN and more than 2000 to exceed the Stixel World algorithms.

The results of the misalignment of the training sequences and the test frames with the online-trained FCNs are provided in Table I. It is clear that the misalignment has a negative impact on the performance of the online training approach, as was expected. The scores drop even below that of the models that are trained offline, also for the FCNs that were initialized with offline pre-trained nets. As the online FCNs outperform all other methods when their training sequence and test frame are aligned, this validates our claim that the online training is giving the system flexibility to adapt to new circumstances, and that over-tuning can be exploited beneficially in the context of free-space detection for ADAS. In conclusion, we exploit the fact that our adaptive strategy is not required to generalize to a large amount of traffic scenes with a single detector. Hence, the detector can -and should-be 'over-tuned' on currently relevant data. In turn, this allows for a small FCN whose training converges fast enough to facilitate real-time deployment in the near future.

TABLE I

RESULTS OF ONLINE TRAINING ON DIFFERENT TRAINING SEQUENCES.

	offline	trained online (scratch)			tuned online (off-self)		
	(man.)	normal	+1/-1	random	normal	+1/-1	rand.
$F_{\max}$	0.87	0.91	0.83	0.79	0.92	0.83	0.80
AP	0.93	0.98	0.91	0.84	0.98	0.92	0.86

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25 (NIPS)*. Curran Associates, Inc., 2012, pp. 1097–1105.

[2] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in *Advances in Neural Information Processing Systems 27 (NIPS)*. Curran Associates, Inc., 2014, pp. 766–774.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Computer Vision and Pattern Recognition (CVPR) (to appear)*, vol. abs/1411.4038, Nov. 2015.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[6] D.-I. D. Pfeiffer, "The stixel world," Ph.D. dissertation, Humboldt-Universität zu Berlin, 2012.

[7] J. M. Alvarez, T. Gevers, Y. LeCun, and A. M. Lopez, "Road Scene Segmentation from a Single Image," in *Eur. Conf. on Computer Vision (ECCV)*, 2012, pp. 376–389.

[8] C.-A. Brust, S. Sickert, M. Simon, E. Rodner, and J. Denzler, "Convolutional patch networks with spatial prior for road detection and urban scene understanding," *Int. Conf. on Computer Vision Theory and Applications (VISAPP)(to appear)*, vol. abs/1502.06344, 2015.

[9] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *Int. Conf. on Intelligent Transportation Systems (ITSC)*. IEEE, 2013.

[10] W. P. Sanberg, G. Dubbelman, and P. H. de With, "Color-based free-space segmentation using online disparity-supervised learning," in *Int. Conf. on Intelligent Transportation Systems (ITSC)*. IEEE, September 2015, pp. 906–912.

[11] ——, "Free-space detection using online disparity-supervised color modeling," in *7th IROS Workshop on Planning, Perception and Navigation for Intelligent Vehicles (IROS-PPNIV)*, Sep. 2015, pp. 105–110.