2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)
Windsor Oceanico Hotel, Rio de Janeiro, Brazil, November 1-4, 2016

Pose-RCNN: Joint Object Detection and Pose Estimation Using 3D Object Proposals

Markus Braun[1,2], Qing Rao[1], Yikang Wang[2] and Fabian Flohr[1,2]

***Abstract*— This paper presents a novel approach for joint object detection and orientation estimation in a single deep convolutional neural network utilizing proposals calculated from 3D data. For orientation estimation, we extend a R-CNN like architecture by several carefully designed layers. Two new object proposal methods are introduced, to make use of stereo as well as lidar data. Our experiments on the KITTI dataset show that by combining proposals of both domains, high recall can be achieved while keeping the number of proposals low. Furthermore, our method for joint detection and orientation estimation outperforms state of the art approaches for cyclists on the easy test scenario of the KITTI test dataset.**

I. INTRODUCTION

Significant progress has been made over the last few years in video-based object detection. Especially in the domain of intelligent vehicle, this technique enabled market introduction of active safety systems which are able to brake automatically in dangerous traffic situations. The PRE-Safe brake system with pedestrian recognition available in the latest Mercedes-Benz C-, E-, and S-Class models is such an example. An autonomous vehicle needs to predict the movement of surrounding Vulnerable Road Users (VRUs) and cars far in advance, in order to be able to brake and/or employ evasive maneuvers in time. Due to the high maneuverability of such surrounding objects, any auxiliary context information that can reduce the uncertainty of the movement prediction should be utilized. Using pose information for example can help increase the prediction horizon up to 1 second without increasing the false alarm rate [1].

Today, deep learning methods are able to capture complex context information by using powerful, multi-layer visual representations. The visual representations are extracted from a set of object proposals estimated by preceding proposal methods. The recall rate of the proposal methods is crucial because it specifies an upper bound for the overall detection performance. The detection performance of standard regional convolutional neural network (R-CNN) on the KITTI benchmark [2] is limited due to the low recall performance of its proposal method, such as Selective Search [3]. Recent work [4] showed that the detection performance can be greatly improved by using more powerful 3D proposals from lidar data.

In this paper, we introduce Pose-RCNN, a combined approach for object detection and pose estimation based on a single R-CNN-like neural network. Pose estimation is carried out through an orientation regression network attached to an R-CNN architecture. The regression net is trained by using a carefully designed von Mises loss function [5] combined with a Biternion representation [6] of the orientation. Inspired by the good results achieved in [4], we present two different 3D proposal methods: One originates from the stixel world [7], the other uses lidar point clouds. The proposed Pose-RCNN is evaluated on the KITTI dataset [2]. We achieve competitive results in both detection and orientation estimation. Both introduced proposal methods achieve similar recall performance as the state of the art and significantly outperform methods that only make use of 2D image data. Fig. 2 shows an example of detected objects with their bounding box and orientation regression.

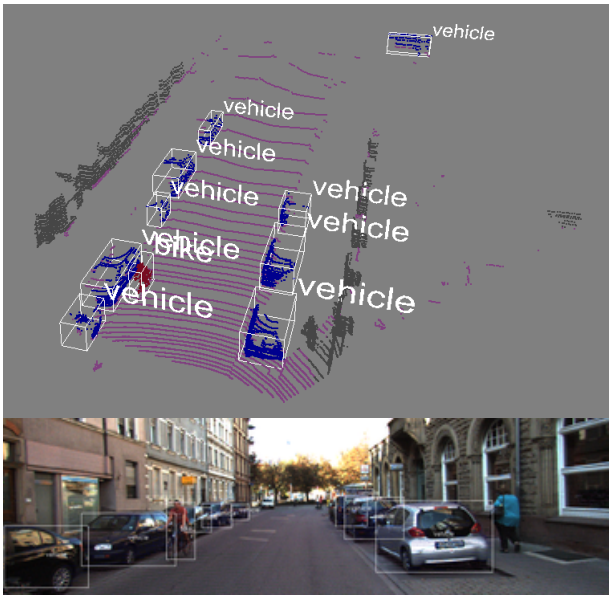

Fig. 1. Example of lidar bounding box proposal through clustering.

II. RELATED WORK

Candidate region of interests (ROIs) that are more likely to contain objects can be generated through bounding box proposal, also referred to as region proposal. Several well-known existing approaches include objectness [8], Multiscale Combinatorial Grouping (MCG) [9], Constrained Parameteric Min-Cut (CPMC) [10], Selective Search (SS) [3], etc. A comprehensive survey of color image based region proposal is given in [11]. Recent works [4][12][13] also showed performance improvements by taking advantage of 3D sensors including RGB-D camera and stereo camera.

[1]Daimler AG, Research and Development. Ulm, Germany.
[2]Informatics Institute, Faculty of Science, University of Amsterdam. Amsterdam, The Netherlands.

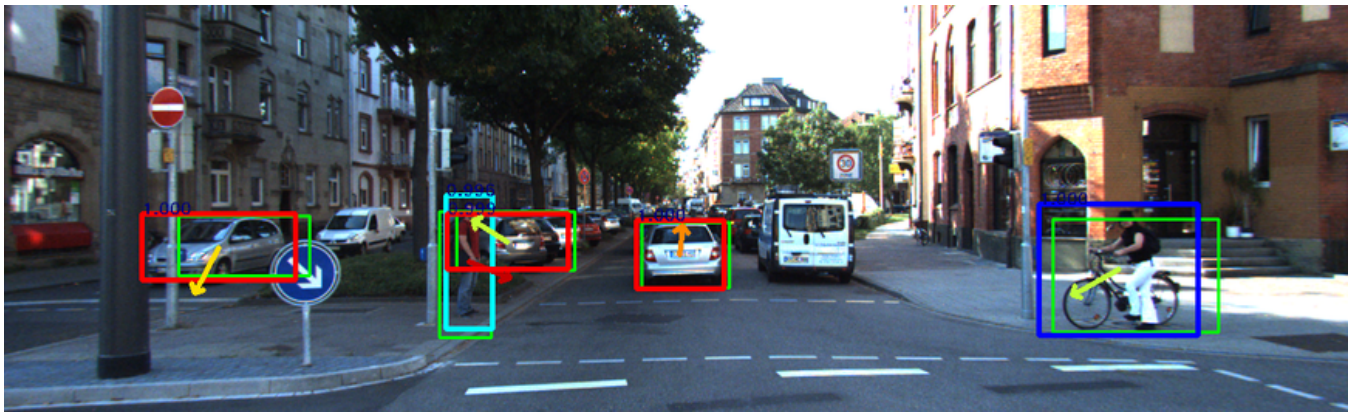

Fig. 2. Example results on the KITTI dataset using our novel Pose-RCNN framework. We show bounding box proposals (green), final detections (after bounding box regression) for car (red), pedestrian (cyan), and cyclist (blue). Results from the orientation regression are shown by the arrow inside the detected bounding boxes.

In the past, detectors in the computer vision area often based on HOG [14], DPM [15] or ACF [16] methods. Today, the domain of object detection and classification is without doubt dominated by Deep Learning based approaches. In 2014, Girshick et al. [17] combined region proposals with Convolutional Neural Networks (R-CNN), which achieved a break-through in detection performance on the Pascal VOC 2012 challenge [18]. Since then, a number of improvements including Fast R-CNN [19] and Faster R-CNN [20] were proposed attempting to make R-CNN realtime capable. In the context of autonomous driving, the current top performing algorithms in object detection on the KITTI vision benchmark [2] include e.g. Subcategory-aware CNN (SubCNN) [21] and the already mentioned 3D Object Proposals (3DOP) [4], which makes use of lidar-based proposals.

Recently, deep neural networks have also been used for estimating orientations of common objects in traffic scenarios [22][23][24]. In these works, orientation estimation was considered a multiclass classification problem. Beyer et al. [6] showed however that a correctly performed regression is a more nature way to address the problem of orientation estimation. They introduced Biternion Net which was capable of regressing fine-grained orientation angles. The Biternion representation is adapted in the Pose-RCNN approach proposed in this paper.

Our paper contributions are three-fold. First, we propose a novel deep CNN architecture called Pose-RCNN for joint object detection and pose estimation based on the well known R-CNN method [19]. We differentiate here to other methods by modeling orientation regression with a careful designed von Mises loss function based on a Biternion representation, while e.g. [4] applies a simple L1 regression. Whereas detection and orientation estimation was treated separately in most other works e.g. [25], we present a joint method for detection and orientation estimation by using one single CNN architecture. Second, we present two 3D proposal methods based on lidar and stixel information. Compared to [4], we introduce a new proposal method based on stixel data without using lidar information. We also show that a combination of lidar and stixel proposals can greatly enhance the recall performance. Third, we show that by using our best proposal method with our new Pose-RCNN architecture, we are competitive with state of the art approaches on the KITTI benchmark.

III. Proposed Approach

A. Lidar Proposal Generation

3D object proposals are generated in a straightforward way by clustering an unorganized lidar scan of the 3D environment into smaller clusters. A particular approach to cluster a traffic scene is to remove ground points and group the rest using the nearest-neighbor clustering technique, as shown in Fig. 1. Ground estimation is carried out through progressive morphological filter (PMF) [26], which distinguishes non-ground measurements such as buildings, vehicles, vegetations etc. from the ground plane. Subsequently, the non-ground lidar points are clustered by grouping nearest neighbors together using a kd-tree search structure [27]. In a last step, the 3D bounding box of each lidar cluster is projected onto the image plane in order to generate 2D object proposals. The 2D proposals are augmented again through spatial translation and scaling.

The recall rate of lidar proposals is highly affected by the parameter settings of the PMF and the nearest neighbor clustering. Here, we experiment two different parameter settings in our work: The first one *Li1* attempts to rigorously keep the false negative rate as low as possible, whereas the second set *Li2* allows more smaller object clusters in order to increase the recall rate. Table I shows the detailed parameter settings of *Li1* and *Li2*.

Additionally, ground estimation by *Li2* is only performed on lidar points within a short range, since the laser scan hardly reaches the ground above a certain distance. In other words, ground points are clustered together with wide range objects by *Li2*. This helps us catch more wide range objects and thus increase the recall rate. The range threshold is set to be 20 meters.

TABLE I

DETAILED PARAMETER SETTINGS OF LIDAR PROPOSAL.

Step	Parameter	*Li1*	*Li2*
Ground estimation	initial ground distance	0.2m	0.15m
	maximal ground distance	0.5m	0.15m
Euclidean clustering	cluster distance	0.3m	0.45m
	minimal number of points	50	10

B. Stereo Proposal Generation

Similarly to [22], we use the stixel representation of the world [28] to generate proposals (see Fig. 3). Stixels are calculated based on stereo data in a joint energy optimization that minimizes the variance of the depth within a stixel. Hence stixels are an efficient and sparse representation of objects having approximately vertical surfaces like vulnerable road users and cars. If the stixel calculation is supported by a ground plane estimation (i.e. in an automotive setup), the 3D position of the bottom of a stixel can be adjusted to match the ground plane. A priori knowledge of the size of possible objects is used to get proposals from the stixels. First, the stixels are filtered by their height in world coordinates that has to be within $[1.2m, 2.4m]$ and their distance that has to be less than $100m$. If an estimated ground plane is available, stixels that are more than $0.5m$ above the ground are also removed. In a second step, the width is adapted to match different aspect ratios. For each stixel, there will be one proposal per aspect ratio. In this work, we evaluate two different parameter settings *SP* and *SPLJ*. The applied aspect ratios for both are 0.5, 1, and 2. The width of the stixels in the energy minimization is fixed to seven pixels for *SP* and three pixels for *SPLJ*. By *SPLJ*, each proposal is augmented through four additional proposals sampled in the surrounding. Therefore, the position of the proposals is adapted by 10% of the witdh to the left and right and 10% of the height to the top and bottom.

C. Pose-RCNN

We extend Fast R-CNN [19] by attaching a small orientation regression network on top of the ROI pooling layer. Based on the last convolution layer of a VGG16 [29] architecture, ROI-pooling is done in the same way as in the original Fast R-CNN version. A softmax probability for classification and per-class bounding-box offsets are estimated from the pooled feature vectors at the end. Using the orientation regression network, we estimate an additional per-class orientation angle from each pooled feature vector. Figure 4 shows the architecture of the proposed Pose-RCNN. It is essential for an orientation regression network to have a carefully-designed loss function and a "mathematically convenient" representation of orientation. In the presented work, we use the von Mises [5] distribution for designing a loss for the orientation regression as done in [6]. The von Mises distribution is an analog of the normal distribution for the circular domain which avoids the problem of angular discontinuity, and it is everywhere differentiable and thus

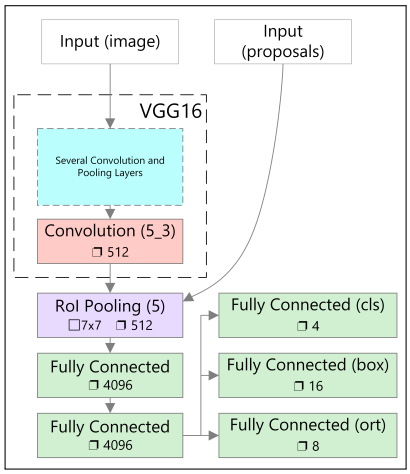

Fig. 4. Net architecture of the proposed Pose-RCNN. □ denotes the size of the ROI pooling layer, and ❒ shows the layer depth. Inputs to the Pose-RCNN comprise an image and a number of bounding box proposals. Feature vectors per bounding box proposal are extracted through the ROI pooling layer, and they are mapped by three fully connected layers in order to generate outputs of the network. These include: softmax probability for classification, per-class bounding-box regression offsets, and per-class orientation regression.

optimal for gradient-based optimization. It has the form

$$g(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}, \tag{1}$$

where θ is an angle, μ is the mean angle of the distribution, κ is the concentration parameter, and I_0 is the *modified Bessel function* of order 0. By inverting and scaling constants, we can derive the von Mises loss function which measures the probabilistic distance between a predicted angle θ and the target angle t as

$$L_{VM}(\theta|t; \kappa) = 1 - e^{\kappa(\cos(\theta - t) - 1)}, \tag{2}$$

with κ as a hyper parameter controlling the shape of the used von Mises distribution. To predict a periodic value using a linear operation, [6] introduces the Biternion representation $\mathbf{y} = (\cos\theta, \sin\theta)$. By combining the von Mises loss function with the Biternion representation and by using common trigonometric identities, the loss function of the orientation regression network and its gradient for back-propagation can be expressed by:

$$L_{VM}(\mathbf{y}|\mathbf{t}; \kappa) = 1 - e^{\kappa(\mathbf{y} \cdot \mathbf{t} - 1)}, \tag{3}$$

$$\frac{\partial L_{VM}}{\partial \mathbf{y}} = -e^{\kappa(\mathbf{y} \cdot \mathbf{t} - 1)} \kappa \mathbf{y}. \tag{4}$$

The derivation shown here is based on the assumption that the biternion representation takes valid values, which means that $\|\mathbf{y}\| = \cos^2\theta + \sin^2\theta = 1$. Since the orientation regression network cannot guarantee the vector length, normalization is required before loss computation. Therefore, we add a normalization layer ensuring the estimations to be always on the unit circle. Given a d-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, the forward pass is simply the normalization operation, which is

$$\mathbf{y} = \frac{\mathbf{x}}{\|\mathbf{x}\|} = \frac{\mathbf{x}}{\sqrt{\mathbf{x} \cdot \mathbf{x}}}. \tag{5}$$

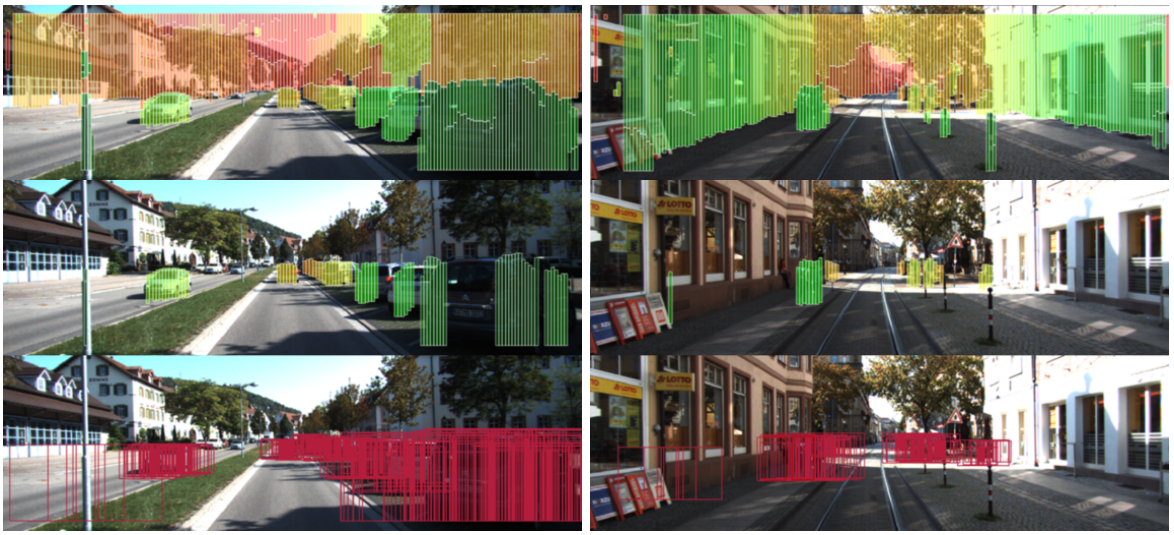

Fig. 3. Qualitative results of the stereo proposal generation. First the complete set of stixels (first row) is filtered by several constraints. The resulting proposals displayed in the last row have the height of the remaining stixels in the second row. Their width is adapted to match given aspect ratios.

For backward pass, we need to derive the partial derivative of the loss with respect to each dimension of $\mathbf{x}$

$$\frac{\partial L}{\partial x_j} = \sum_{i=1}^{d} \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial x_j}, \; j \in \{1, 2, \ldots, d\}, \tag{6}$$

while $\frac{\partial L}{\partial y_i}$ comes from the succeeding layer. Followed by basic derivation rules, we further get

$$\frac{\partial y_i}{\partial x_j} = \frac{\delta_{ij} - y_i y_j}{\sqrt{\mathbf{x} \cdot \mathbf{x}}}, \tag{7}$$

where $\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$ represents the Kronecker delta.

IV. Experiments and Evaluation

A. Experimental Setup

We evaluate our proposed approach for proposal generation, object detection, and orientation estimation on the KITTI object benchmark, which consists of 7481 training images and 7518 test images that are captured by the left RGB camera mounted on the recording vehicle. For each left color image in the object dataset, a corresponding right color image is available, which enables computing disparities and stixels from stereo pairs. Lidar point clouds and calibration information between the lidar and the cameras are also provided. A total of 80256 labeled objects in common traffic scenes including cars, pedestrians, and cyclists are available in the public training dataset.

We evaluate the different 3D proposal methods *SP*, *SPLJ*, *Li1*, *Li2* as well as certain combinations. Hereby *SP-Li1*, *SPLJ-Li1*, and *SPLJ-Li2* are just the union of the corresponding proposals without any filtering of duplicates.

The framework and parameter settings provided by [4] and our different proposal methods are used for the training of our Pose-RCNN model. 50% of the images of the training dataset serve as validation set.

B. Results

1) Proposal methods: We compare our proposal methods with several state of the art approaches – among others Selective Search (*SS*) [3] and *3DOP* [4]. Therefore, we plot the Recall as a function of the IoU threshold between 0.5 and 1.0 (see Fig. 6). Average recall (AR) defined in [11] is used as the metric for comparison. Note that the number of proposals is not the same for all methods. A fair comparison can be made between the results of *SP-Li1*, *3DOP* and *SS* for 500 proposals. *SP-Li1* achieves a higher AR than *3DOP* for the moderate and hard setting of the cyclist class and a lower AR for the other cases. *SS* and other state of the art methods like MCG [9] and BING [32] are outperformed (results for these are shown in [4]). The average recall of *SPLJ* is higher than for the corresponding setting *SP* while increasing the number of proposals (see Fig. 6). Note how the combination of stereo and lidar proposals further boosts the average recall score for *SPLJ-Li1* and *SPLJ-Li2*. The

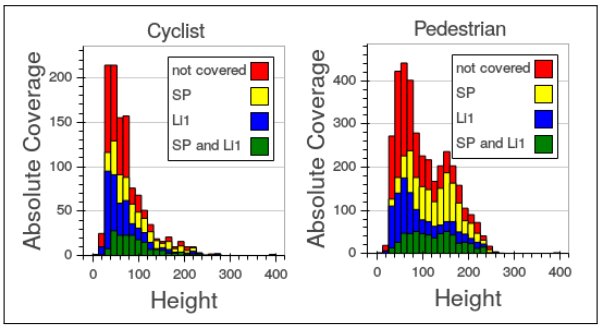

Fig. 5. Absolute coverage of the groundtruth samples of the pedestrian and cyclist classes (moderate difficulty). A sample counts as covered if there is a proposal with an IoU greater than 0.7. The samples are grouped by the method of the covering proposal and their height values. The absolute count of these groups is represented by the height of the bar segments.

combination of proposals of stereo and lidar data is further analyzed in Fig. 5. Although duplicates are not deleted when

TABLE II

AVERAGE PRECISION (IN %) ON THE TEST SET OF THE KITTI BENCHMARK

	Cars			Pedestrian			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
ACF [16]	55.89	54.74	42.98	44.49	39.81	37.21	-	-	-
R-CNN [30]	-	-	-	61.61	50.13	44.79	-	-	-
DPM-VOC+VP [31]	74.95	64.71	48.76	59.48	44.86	40.37	42.43	31.08	28.23
3DOP [4]	**93.04**	88.64	79.10	81.78	67.47	64.70	78.39	68.94	61.37
SubCNN [21]	90.81	**89.04**	**79.27**	**83.28**	**71.33**	**66.36**	79.48	**71.06**	**62.68**
Ours	88.43	75.80	66.57	77.53	63.40	57.49	**80.79**	68.79	60.40

TABLE III

AVERAGE ORIENTATION SIMILARITY (IN %) ON THE TEST SET OF THE KITTI BENCHMARK

	Cars			Pedestrian			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
DPM-VOC+VP [31]	72.28	61.84	46.54	53.55	39.83	35.73	30.52	23.17	21.58
3DOP [4]	**91.44**	86.10	76.52	72.94	59.80	57.03	70.13	58.68	52.35
SubCNN [21]	90.67	**88.62**	**78.68**	**78.45**	**66.28**	**61.36**	72.00	**63.65**	**56.32**
Ours	88.34	75.41	66.07	73.95	59.90	54.27	**75.49**	62.87	55.47

combining proposals, a lot of groundtruth samples are only covered by one or more proposals of exactly one method. A great amount of small objects is only covered by lidar proposals. That is due to missing stereo data in the far range.

2) Detection and orientation regression: Our Pose-RCNN model trained with proposals of the *SPLJ-Li2* setting achieves the best results on the validation set. For evaluation on the test dataset, we train an additional Pose-RCNN model with *SPLJ-Li2* proposals on the combined training and validation dataset. Table II and III show the respective test results. Especially for cyclists, our approach performs very well. We achieve the highest detection and orientation scores on the easy test scenario and competitive ones on the other scenarios.

C. Discussion

Average orientation similarity as defined in [2] is strongly correlated with the average precision. The average precision score is even the upper bound for the average orientation similarity score. The ratio between the orientation and detection score can not be higher than one. For cars, 3DOP [4], SubCNN [21], and our approach already achieve a ratio of nearly one. For cyclists and pedestrians, the ratio achieved by our approach is higher than by 3DOP and SubCNN, which evinces the potential of our proposed approach. The improvement of the detection performance for example by improving the average recall of the proposals could automatically boost the average orientation similarity score.

V. CONCLUDING REMARKS

We presented a novel approach called Pose-RCNN for joint object detection and pose estimation. The proposed approach exploits the power of deep learning to jointly perform object bounding box regression, classification, and orientation estimation. It is supported by a combination of 3D object proposals from stereo and lidar measurements. Quantitative evaluation results show that the proposal generation as well as the joint detection and orientation approach are competitive with or even outperform other state of the art approaches.

Future work will first focus on improving the proposal generation by reducing the number of proposals while keeping the average recall high. This will further improve the detection performance and will enable us to integrate the Pose-RCNN into an in-vehicle realtime framework. We believe that this novel technique then will serve as a basis on which various features for autonomous driving can be built.

REFERENCES

[1] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-Based Pedestrian Path Prediction," in *ECCV '14*, pp. 618–633.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *CVPR '12*, pp. 3354–3361.

[3] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective Search for Object Recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

[4] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3D Object Proposals for Accurate Object Class Detection," in *NIPS '15*, pp. 424–432.

[5] K. V. Mardia and P. E. Jupp, *Directional Statistics*. John Wiley & Sons, Inc., 2008.

[6] L. Beyer, A. Hermans, and B. Leibe, "Biternion Nets: Continous Head Pose Regression from Discrete Training Labels," in *GCPR '15*, pp. 157–168.

[7] D. Pfeiffer and U. Franke, "Towards a Global Optimal Multi-Layer Stixel Representation of Dense 3D Data," in *BMVC '11*, pp. 51.1–51.12.

[8] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the Objectness of Image Windows," *IEEE T-PAMI*, vol. 34, no. 11, pp. 2189–2202, 2012.

[9] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale Combinatorial Grouping," in *CVPR '14*, pp. 328–335.

[10] J. Carreira and C. Sminchisescu, "CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts," *IEEE T-PAMI*, vol. 34, no. 7, pp. 1312–1328, 2012.

[11] J. Hosang, R. Benenson, P. Dollar, and B. Schiele, "What Makes for Effective Detection Proposals?" *IEEE T-PAMI*, vol. 38, no. 4, pp. 814–830, 2016.

[12] D. Lin, S. Fidler, and R. Urtasun, "Holistic Scene Understanding for 3D Object Detection with RGBD Cameras," in *ICCV '13*, pp. 1417–1424.

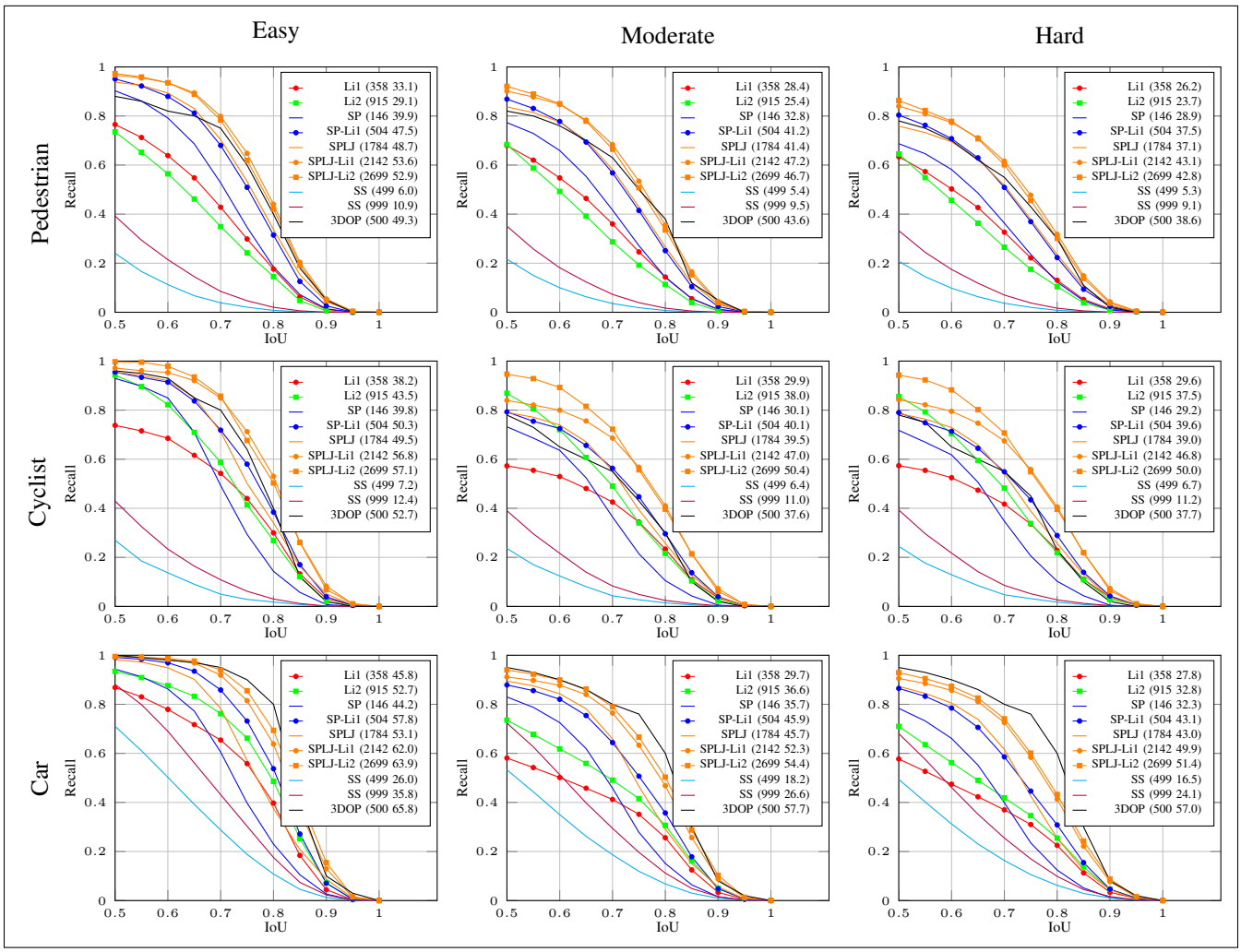

Fig. 6. Recall as a function of the IoU threshold for several proposal methods. The average number of proposals per frame (first number) and the average recall (in %) are displayed inside the brackets. The *3DOP* curve is sampled from the original paper [4].

[13] S. Gupta, R. B. Girshick, P. Arbeláez, and J. Malik, "Learning Rich Features from RGB-D Images for Object Detection and Segmentation," in *ECCV '14*, pp. 345–360.

[14] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *CVPR '05*, vol. 1, pp. 886–893.

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-based Models," *IEEE T-PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[16] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast Feature Pyramids for Object Detection," *IEEE T-PAMI*, vol. 36, no. 8, pp. 1532–1545, 2014.

[17] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *CVPR '14*, pp. 580–587.

[18] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *IJCV*, vol. 111, no. 1, pp. 98–136, 2015.

[19] R. B. Girshick, "Fast R-CNN," in *ICCV '15*, pp. 1440–1448.

[20] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *NIPS '15*, pp. 91–99.

[21] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware Convolutional Neural Networks for Object Detection," *arXiv*, 2016.

[22] X. Li, F. Flohr, Y. Yang, H. Xiong, M. Braun, S. Pan, K. Li, and D. M. Gavrila, "A New Benchmark for Vison-Based Cyclist Detection," in *IV '16*.

[23] H. Su, C. R. Qi, Y. Li, and L. Guibas, "Render for CNN: Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views," in *ICCV '15*, pp. 2686–2694.

[24] S. Tulsiani and J. Malik, "Viewpoints and Keypoints," in *CVPR '15*, pp. 1510–1519.

[25] F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij, and D. M. Gavrila, "A Probabilistic Framework for Joint Pedestrian Head and Body Orientation Estimation," *IEEE TITS*, vol. 16, no. 4, pp. 1872–1882, 2015.

[26] K. Zhang, S.-C. Chen, D. Whitman, M.-L. Shyu, J. Yan, and C. Zhang, "A Progressive Morphological Filter for Removing Nonground Measurements from Airborne LIDAR Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 4, pp. 872–882, 2003.

[27] R. Rusu, "Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments," Ph.D. dissertation, Computer Science department, Technische Universität München, Germany, 2009.

[28] M. Enzweiler, M. Hummel, D. Pfeiffer, and U. Franke, "Efficient Stixel-based Object Recognition," in *IV '12*, pp. 1066–1071.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[30] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a Deeper Look at Pedestrians," in *CVPR '15*, 2015, pp. 4073–4082.

[31] B. Pepik, M. Stark, P. Gehler, and B. Schiele, "Multi-view and 3D Deformable Part Models," *IEEE T-PAMI*, vol. 37, no. 11, pp. 2232–2245, 2015.

[32] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr, "BING: Binarized Normed Gradients for Objectness Estimation at 300fps," in *CVPR '14*, pp. 3286–3293.