Learning with Proxy Supervision for End-To-End Visual Learning

Jiří Čermák[1*] Anelia Angelova[2]

***Abstract*— Learning with deep neural networks forms the state-of-the-art in many tasks such as image classification, image detection, speech recognition, text analysis. We here set out to gain understanding in learning in an 'end-to-end' manner for an autonomous vehicle, which refers to directly learning the decision which will result from the perception of the scene. For example, we consider learning a binary 'stop'/'go' decision, with respect to pedestrians, given the input image. In this work we propose to use additional information, referred to as 'proxy supervision', for improved learning and study its effects on the overall performance. We show that the proxy labels significantly improve the robustness of learning, while achieving as good, or better, accuracy than in the original task of binary classification.**

I. INTRODUCTION

Understanding complex urban scenes and producing intelligent behavior of autonomous or semi-autonomous vehicles is still an extremely challenging problem [4], [14], [12], [18], [10], [9], [8], [3]

Pedestrian detection, for example, has been a topic of research for more than 20 years [17], [7], [5], [6], [2]. Whereas accurate detection of all persons in the scene is very important, we here explore an alternative approach in which we wish to predict a direct end-to-end decision, of whether the vehicle needs to 'stop' (slow down) or 'go' (drive), given the currently observed situation ahead. That is, we train an end-to-end convolutional network detecting whether driving in the scene, seen in an image taken by the frontal camera mounted on the car, is safe or dangerous with respect to people in the scene. For simplicity, we here explore stop and go decisions with respect to pedestrians and cyclists (Figure 1).

Learning end-to-end is appealing as it avoids losses at intermediate steps, i.e. if we were to train several consecutive classifiers instead (e.g. train a pedestrian detector, followed by pose estimator, followed by a gaze detector in order to determine the pedestrian's intended behavior). However, it comes with its challenges, as in a direct end-to-end learning a very complex decision is in fact being learned. We propose to enact a 'proxy supervision' by additional labels which may be strongly or weakly related to the task at hand. We mainly focus on the effect of the additional proxy supervision, forcing the network to learn to detect pedestrian and cyclist bounding boxes, on the overall performance of the network. Figure 2 shows a schematic of our approach which uses proxy supervision. The proxy labels, which may come from manual labeling or other sources, are used as additional supervision during learning. As we show later in the paper, the proxy labels do not need to be present for all examples and have utility even as a very weak supervision. We also explore supervision to a seemingly non-related task, e.g. using bounding boxes for cars, for the purposes of making a stop/go decision with respect to persons. We also see that such supervision is beneficial, although of naturally less pronounced impact, which we attribute to their relations in scene understanding. Overall we find that using proxy supervision is very helpful for end-to-end learning and for stabilizing the learning process.

*This work was done while at Google Brain.

[1]Czech Technical University in Prague, Czech Republic jiri.cermak@agents.fel.cvut.cz

[2]Google Brain, Mountain View, USA anelia@google.com

'Stop' instance

'Go' instance

Fig. 1. End-to-end learning of a 'stop' or 'go' decision with respect to pedestrians or cyclists, given the input scene. This paper explores if directly learning the decision of 'stop' or 'go' is feasible and whether auxiliary proxy labels as supervision are beneficial to learning. Note that a 'stop' decision is a very conservative one.

II. PREVIOUS WORK

Understanding pedestrian behavior in urban environments is of high importance for autonomous driving systems [14], [2], [20].

Learning with auxiliary tasks has been embodied as training to additional losses, as was done in the Inception Network [19]. Recent work has showed that training to predict depth is beneficial [13] in the context of learning navigation with Reinforcement Learning. Our work is most aligned to these approaches.

Recent work in learning from visual inputs have gravitated towards more end-to-end systems, i.e. instead of separating the tasks into subproblems and solving them as separate learning tasks, to solve them as a holistic end-to-end problem. Such systems have shown initial promise [3], [16], [1].

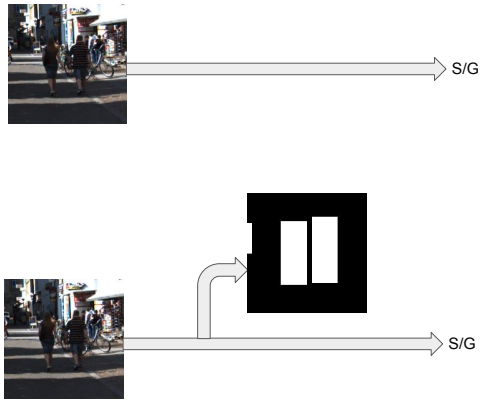

Fig. 2. Stop/go learning with proxy supervision. Top: direct learning of stop/go decision with respect to pedestrians. Bottom: Proposed stop/go learning with proxy labels. In this case proxy labels may come from manual annotation of pedestrian bounding box. Proxy labels do not need to be present for all examples.

We are not aware of another work that examines end-to-end system behavior with respect to behavior of pedestrians/cyclists in urban environments.

III. LEARNING WITH PROXY SUPERVISION

In end-to-end scenarios the main goal is to obtain direct decisions given the input information from an image or image stream. For the case of this study, we use as an input an image obtained from an onboard camera from the KITTI dataset [11]. The output is a stop/go decision, with respect to pedestrians/cyclists which takes a couple of variations. More specifically, we explore the case where the decision is made based on an automatic, or direct, criterion, and when it is judged by a human annotator.

We here apply a deep convolutional neural network [15] for learning and classification, due to the major success of such algorithms in many tasks, including object recognition and detection.

A. Joint learning with proxy labels

We explore, how end-to-end learning is affected by introducing side objectives, which we call proxy supervision. Figure 2 demonstrates the idea. The main feedforward network will learn to predict stop/go decisions per given visual input. In addition, a *proxy* label can be applied during learning as a side objective, e.g., to force the network to additionally learn to predict pedestrian and cyclist bounding boxes. Clearly, such an objective is correlated with the final decision to stop to yield to pedestrians, or drive because no pedestrians in dangerous positions are present.

B. Problem Formulation

More formally, as an input we use 192×192 3-channel images, denoted as x_i representing the situation in front of the car. In addition to x_i we are provided with proxy labels p_i in form of 192×192 single channel images containing the bounding boxes of pedestrians and cyclists present in x_i. Finally, we use a binary label l_i identifying the scene as either safe or dangerous.

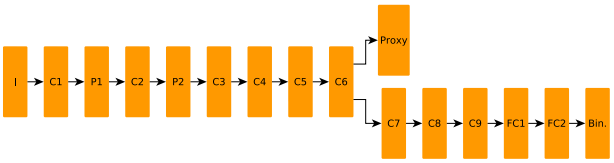

Fig. 3. Model architecture for learning with proxy labels. It is a feedforward convolutional network in which the proxy supervision is applied after a split by the end of the network. Thus all prior layers are shared. In the graph, 'C' refers to a convolutional layer, 'P' is a 2x2 pooling layer, whereas 'FC' refers to a fully-connected one.

TABLE I
MODEL ARCHITECTURE. PROXY SUPERVISION IS APPLIED AFTER C6.

Name	Convolutions	Filters	Stride
C1	5x5	64	2
C2	3x3	128	2
C3	3x3	128	1
C4	3x3	256	1
C5	3x3	256	1
C6	5x5	16	1
C7	3x3	128	1
C8	3x3	256	1
C9	3x3	256	1
FC1	- (FC)	256	-
FC2	- (FC)	1	-

The cost that is being optimized is as follows:

$$L(\theta) = \sum_i (L_{proxy}(x_i, p_i, \theta) + \alpha L_{binary}(x_i, l_i, \theta)) \quad (1)$$

where L_{proxy} and L_{binary} are the cross entropy losses for the variables corresponding to the proxy labels and the binary labels respectively. θ represents the parameters of the network and will be omitted in the following text.

$$L_{proxy}(x_i, p_i) = -\frac{1}{N}(\sum_{j,k} p_i^{j,k} log(s(f_p^{j,k}(x_i))) + \quad (2)$$

$$(1 - p_i^{j,k}) log(1 - s(f_p^{j,k}(x_i)))) \quad (3)$$

$$L_{binary}(x_i, l_i) = -\frac{1}{N}(l_i log(s(f_l(x_i))) + \quad (4)$$

$$(1 - l_i) log(1 - s(f_l(x_i)))) \quad (5)$$

where $s(x) = 1/(1+e^{-x})$ is the sigmoid function, and f_p, f_l are the feature predictions from the network that correspond to predicting the proxy labels and the binary labels, respectively. N is the number of examples, the index i spans examples, whereas j,k span pixels for each example (in order to apply proxy supervision which spans the full image). α is a coefficient that balances the two costs and we used a fixed value of 10^5 in all our experiments.

We use a feedforward convolutional neural network (Figure 3) for training. It consists of several convolutional layers,

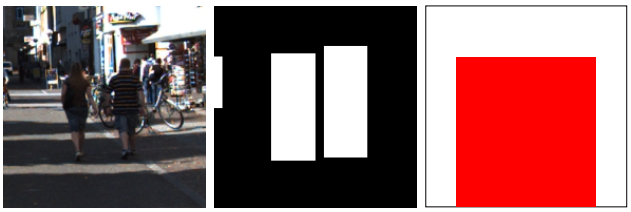

Fig. 4. Definition of the direct criterion. For each input patch (left) and provided person bounding boxes (middle), the criterion determines if any person bounding box overlaps with the central 'danger' zone (marked as red in the right image).

followed by a fully-connected ones (Table I). Dropout is applied after the first fully connected layer. Pooling and normalization layers are applied (denoted as P1, P2 in Figure 3). An alternative architecture with a skip connection is explored later in the paper.

IV. End-to-end decision criteria

Since the KITTI dataset does not contain stop/go labeling of images we provide an independent criteria for these labels. We explore two different stop/go criteria with respect to pedestrians and bicyclists. The first one is called a *direct* criterion since the criterion is generated automatically as a (nonlinear) function of the persons that are available in the input image (Section IV-A). The ground truth for pedestrians/cyclists are obtained from the KITTI benchmark. We sampled patches from KITTI images and labeled them automatically by this criterion. The second criterion (Section IV-B), concerns the full scene and is based on human judgment, as to whether a person is in dangerous position or not, in order to determine a stop/go decision.

A. Direct criterion

The direct criterion is derived as a nonlinear function of the proxy labels. It is intended for experimental purposes, as it is known that the proxy labels and the final decision are related by a nonlinear function. In particular, the criterion (see Figure 4) determines if any person bounding box overlaps with the central 'danger' zone (marked as red in the right image). Thus persons who are outside the danger zone, are either not directly in front of the car (e.g. to the side) or are too far (at the top of the box). This criterion is created so that it is a complex function of the provided bounding boxes (which will be used as proxy labels). Thus, we can test whether the networks can learn this function, with or without proxy labels.

Figure 5 visualizes example input images and their corresponding proxy labels. Note that since the criterion is built with respect to persons it is a 'go' decision for the bottom image, but would have been a 'stop' decision if it were with respect to cars.

B. Human decision based criterion

1) Manual labeling tool: To collect data with human judgment of whether a situation is safe for pedestrians, we created a manual labeling tool which allows a labeler to provide stop/go decisions and mark the pedestrians or cyclists in danger. Figure 6 visualizes the input and output from the human labeling tool. The input image contains all pedestrian and cyclist bounding boxes in the scene (shown in blue in the top of Figure 6). The labeler marks the ones that can be potentially in danger (green bounding boxes at the bottom), while taking into consideration the direction of their movement, whether they are on the sidewalk, intending to cross, etc.

Fig. 5. Examples of a stop (top) and go (bottom) instances and the corresponding proxy labels (right). The proxy label contains the full bounding box over pedestrians and bicyclists as a binary mask.

2) Data with human decision criterion: The decision, as mentioned above, is taken by a human and takes into consideration the potential intentions of the persons in the scene. This criterion is more complex than the direct one. Furthermore, while the proxy labels for this case are expected to be correlated with the final decision, the overall decision is not a direct function of all marked pedestrian boxes.

V. Experimental results

Here we present results of our experiments. First we introduce the dataset and then present results with the two criteria for proxy supervision: the direct criterion which is obtained automatically and the human labeled, which is obtained by subjective judgement of a human labeller.

In the experiments, we show the effect of additional side objective to the performance of the stop/go classification, when the side objective is to learn pedestrian and cyclist bounding boxes. Next, we explore the effect of learning car bounding boxes as the side objective. We further experiment with pre-training the neural network. Following are experiments where the network learns safe/dangerous classification based on labeling provided by a person and experiments with a skip-connection architecture.

A. Dataset

We use the KITTI dataset [11] containing 7481 1224×370 colored images taken by a frontal camera mounted on a car. This data set provides bounding boxes of cars, pedestrians, cyclists etc., present in the scene. The KITTI dataset contains complex urban scenes with variety of scenarios in which

Fig. 6. Stop/go manual labelling tool. Example image with provided bounding boxes. A labeler decides whether some of the persons (in the top row) intend to cross the path of the vehicle (assuming the vehicle is driving along the road, or straight otherwise). In that case the image is labeled as stop instance. All persons who are in dangerous positions, e.g. the ones in or close to the street, are labeled as green (the bottom row). Best viewed in color.

pedestrians may be crossing the street, leaving a parked car, or simply walking on the sidewalk without interference to the vehicle.

B. Direct Criterion

In all our experiments, we use 5 randomly generated crossvalidation sets with images randomly chosen from the KITTI dataset. We use 4:1 ratio of training to validation images with approximately 4:1 ratio of safe to dangerous labels. Additionally, the easily computable direct criterion allows us to generate richer training and validation data from the crossvalidation sets. We use random crops and rescaling to generate about 1500000 training and 350000 validation colored 192×192 images in every crossvalidation set.

We first examine if training with proxy supervision is useful and test if portion of proxy labels are helpful. In the Figure 7 we provide the average top1 error of stop/go classification with standard deviation obtained on the cross-validation sets as we reduce the percentage of images which have the proxy label available. When using no proxy labels in this experiments, it was very difficult to force the network to overcome the local optimum, where it simply classified everything as the majority go instance. When using the proxy labels, however, the robustness of the learning increased significantly. We were able to move from the local optimum without any additional hyperparameter tuning even when using only 10% of the proxy labels. Furthermore, when increasing the amount of proxy labels available during training, the average error consistently decreases. As mentioned, without proxy labels, we achieved a very high error ($\sim$22% which corresponds to one class overwhelming the other).

In Figure 8 we test the performance in the case in which 100% of proxy labels are used and only a portion of stop/go labels are included. More specifically, we report the average top1 error of stop/go classification with standard deviation obtained on the crossvalidation sets, but this time we reduce the percentage of images which have the stop/go label available. Here we further confirm the increase in the robustness of learning, since when we apply all the proxy labels the network is capable of overcoming the learning issues even when using only 10% of the stop/go labels. Clearly, with only 10% of the stop/go labels, and all proxy labels the error is 2x higher than when training the other way around, since the direct supervision is the most useful.

We further examine if pretraining by the same proxy labels would have effect. To that end we first train a network to predict the proxy masks (with the same architecture up until C6) which has to output the given pedestrian boxes.

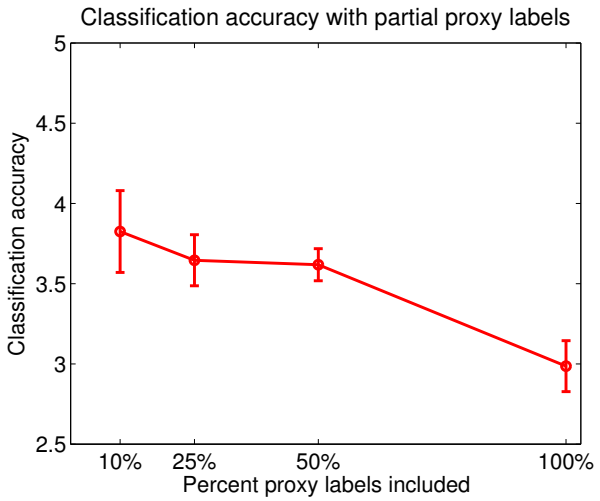

Fig. 7. Average error and standard deviation of the binary classification obtained by 5 fold crossvalidation using the direct criterion when only given amount of data is labeled by proxy label (with 100% stop/go labels).

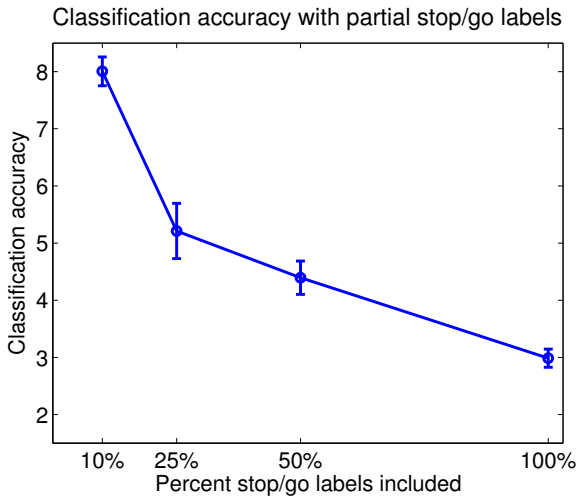

Fig. 8. Average error and standard deviation of the binary classification obtained by 5 fold crossvalidation using the direct criterion when only given amount of data is labeled by the stop/go label (with 100% proxy labels).

In Table II we show the average top1 error achieved after we pre-train the network prior to training. In the first row, we report the best average error achieved when using all the proxy labels followed by the result achieved by using no proxy labels at all. As we can see from the results the pretraining is very helpful (which is observed in other works and experiments). However, the proxy labels still improve the overall performance of the network while making the learning more robust.

We then wanted to understand if learning with seemingly non-related labels as proxy supervision is beneficial. In Table III we report the results of an experiment where we use the car bounding boxes as proxy supervision and compare to proxy supervision by the person bounding boxes. Note that car bounding boxes are unrelated to the stop/go decision problem with regards to persons, although it is related to the overall perception of the scene. From our experiments we clearly see that using car proxy labels is more useful than none at all. This is even in the case of pretraining, when the pretraining is done on bounding boxes with persons. And naturally, using person labels as proxy supervision is more beneficial than the car ones.

Overall, we find that using proxy supervision is beneficial, even in the presence of more adequate pre-training. We note that the error rate on the final stop/go decision is still relatively large (slightly below 3%), since the criterion is imperfect and may produce contradictory labels, e.g. if a person's bounding box is 1 pixel more in the dangerous zone, it will be labeled as a stop, otherwise a go decision.

C. Human Labeled Data

In this subsection we use the stop/go labels provided by human labelling, as described in Section IV-B. This dataset is more challenging as the manual decision is subjective and involves complex scene situations. Since a labeler evaluates the entire image, we use the full image as input to the prediction. These experiments therefore use only the 7481 images available in the KITTI benchmark, again partitioned to 5 crossvalidation sets as described above. The ratio of safe/dangerous classes provided in the human made labels is approximately 5:1. The only preprocessing used was the transformation of the original 1224×370 colored image to 192×192 colored image by shrinking it and thus reducing the quality somewhat.

TABLE II

AVERAGE ERROR AND STANDARD DEVIATION OF THE BINARY CLASSIFICATION OBTAINED BY 5 FOLD CROSSVALIDATION, WHEN PRETRAINING WITH KITTI BOUNDING BOXES.

Percent Proxy label	No Pretraining	Pretraining
100%	2.986 $\pm$ 0.159%	2.826 $\pm$ 0.189%
0%	22.0%	3.625 $\pm$ 0.215%

In Table IV we present the average error and standard deviation when learning the stop/go labels provided by a human labeler with no pretraining (first row) and with pretraining done in the same way as in the case of the direct criterion (second row). The error, as before, is averaged over the crossvalidated runs. The overall average error increased, since the data set is more challenging and is of smaller size. We again observed significant increase in the robustness of the learning when using proxy labels, both when pretraining and when used as an additional supervision.

D. Exploring skip connections

Here we try to reduce the error on the human labeled data by exploring additional network architectures. This is motivated by the fact that the baseline network (Figure 3) applies the logic for stop/go classification after the $C6$ convolution layer which is forced by the proxy cost to detect the bounding boxes of the pedestrians. We here present results for the skip architecture (Figure 9) which is conceptually similar to our main architecture, but has a skip connection. The skip link

TABLE III

PEDESTRIAN PROXY LABELS VS CAR PROXY LABELS USED AS SUPERVISION(EXPERIMENTS DONE WITH PRETRAINING).

Learning scenario	pedestrian proxy labels	car proxy labels	no proxy labels
Classification error	2.826 ± 0.189%	3.222 ± 0.250%	3.625 ± 0.215%

TABLE IV

AVERAGE ERROR ON HUMAN LABELED DATA (WITH AND WITHOUT PROXY SUPERVISION).

Proxy label	100%	0%
pedestrian	5.601 ± 0.667 %	6.520 ± 0.422 %
pedestrian, pretr.	5.492 ± 0.569 %	6.374 ± 0.383 %

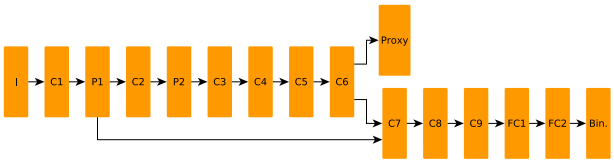

Fig. 9. The skip-connection architecture.

essentially forwards the information from the earlier layers to the part of the network handling the stop/go classification.

Comparison between the baseline and the architecture is presented in Table V. The skip architecture provides promising results, as it almost halves the average error achieved by the baseline feedforward architecture.

TABLE V

SKIP-CONNECTION NETWORK. COMPARISON OF ARCHITECTURES' PERFORMANCE. PEDESTRIAN PROXY LABELS, WITH PRETRAINING. HUMAN LABELED DATA.

Architecture \ Proxy label	100%	0%
Feedforward arch.	5.492 ± 0.569 %	6.374 ± 0.383 %
Skip arch.	2.885 ± 0.297 %	3.199 ± 0.327 %

VI. CONCLUSIONS AND FUTURE WORK

We present an approach for end-to-end learning based on additional, proxy supervision. We have found that the proxy labels are beneficial during the training process and can achieve better accuracy, even when pre-training is used. We also find that the proposed use of proxy-based learning helps in training stability. Proxy supervision which aims to detect other components of the scene, e.g. cars, are also helpful to a seemingly not related task for pedestrians. More advanced architectures, such as in our case the skip architecture may decrease the differences in accuracy.

This is a first attempt at improving an end-to-end learning scenario, and there are many future directions. Much more complex decisions need to be made in urban environments, since many more agents are present in the road, vehicles, traffic signs, etc. Other aspects of proxy supervision can be considered. Learning to make decisions in time is valuable to improve the performance, e.g. to estimate walking direction. Learning more complex decisions, rather than a binary one, is another future direction.

ACKNOWLEDGMENTS

Special thanks to Alex Krizhevsky and to Matthieu Devin, Samy Bengio, Oriol Vinyals and other members of the Brain team for their help and insights. We also thank the anonymous reviewers for their comments.

REFERENCES

[1] Deeptesla: http://selfdrivingcars.mit.edu/deeptesla/.

[2] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? *2nd Workshop on Road scene understanding and Autonomous driving, ECCV*, 2014.

[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *CoRR: 1604.07316*, 2016.

[4] E. Coelingh, A. Eidehall, and M. Bengtsson. Collision warning with full auto brake and pedestrian detection - a practical example of automatic emergency braking. *13th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2010.

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. *CVPR*, 2009.

[6] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 2012.

[7] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Trans. on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2009.

[8] A. Ess, B. Leibe, K. Schindler, , and L. van Gool. A mobile vision system for robust multi-person tracking. *CVPR*, 2008.

[9] D. Petrich et al. Map-based long term motion prediction for vehicles in traffic environments. *Intelligent Transportation Conference*, 2016.

[10] A. Geiger, M. Lauer, C. Stiller C. Wojek, and R. Urtasun. 3d traffic scene understanding from movable platforms. *Trans. PAMI*, 2014.

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013.

[12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. *CVPR*, 2012.

[13] M. Jaderberg, V. Mnih, W. Czarnecki, T. Schaul, J. Leibo, D. Silver, and K. Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *CoRR:1611.05397*, 2016.

[14] C. Keller, C. Hermes, and D. Gavrila. Will the pedestrian cross? probabilistic path prediction based on learned motion features. *DAGM*, 2011.

[15] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

[16] Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J. Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, Dharshan Kumaran, and Raia Hadsell. Learning to navigate in complex environments. *CoRR: 1611.03673*, 2016.

[17] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. *CVPR*, 1997.

[18] L. Schneider, M. Cordts, T. Rehfeld, D. Pfeiffer, M. Enzweiler, U. Franke, M. Pollefeys, and S. Roth. Semantic stixels: Depth is not enough. *Intelligent Vehicles Symposium*, 2016.

[19] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. 2015.

[20] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Understanding pedestrian behaviors from stationary crowd groups. *ECCV*, 2015.