Using A Priori Knowledge to Improve Scene Understanding in Autonomous Navigation

Brigit Schroeder[1], Alexandre Alahi[2]
[1]Department of Computer Engineering, University of California, Santa Cruz [2] Visual Intelligence for Transportation Lab, EPFL

1 Motivation

Autonomous vehicles have been predicted to make significant changes to the field of transportation in the next decades. In order to support the growth of this technology, the development of scene understanding algorithms which can robustly identify the objects in the vehicle's viewpoint is crucial for improving both navigation and safety. Fine-grained semantic labeling of static and dynamic objects in a scene, such as roadways, sidewalks, sign posts, cars, pedestrians and cyclists (see Figure 1), provides valuable information as to what is in the vehicle's near and far field [6]. This information can be fed to systems such as collision avoidance and trajectory prediction, depending upon the level of autonomy of the vehicle, to further improve both vehicle and pedestrian safety.

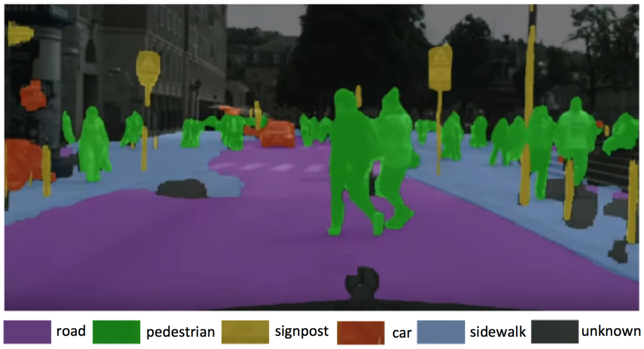

Figure 1: Scene Understanding: An example of a semantically labeled (or segmented) image which can be used to improve navigation and safety [5].

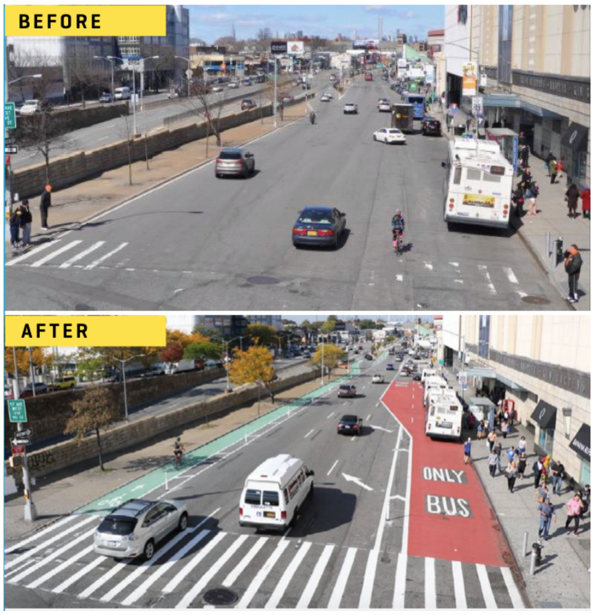

Figure 2: Spatial Priors: The "before" and "after" images of an intersection after improvements have been made. The before image provides a strong structural prior for labeling objects in the new ("after") image.

An autonomous vehicle is typically outfitted with several sensor modalities which can be used for mapping of environment [8] through which it drives (e.g Google self-driving cars continuously map the campus and streets of Mountain View, CA, near the headquarters) [3]. A priori knowledge of a given area can be derived from maps created during previous traversal through an intersection, for example. These scene priors, in the form of imagery and semantic labeling, can be incorporated into scene understanding algorithms to improve the semantic segmentation of the scene in its current state. As seen in Figure 2, the "before" image provides a strong spatial prior: the scene need not have the exact appearance to be useful as the fundamental layout of the road, sidewalk and buildings represent a strong structural prior. Time series data in the form of video from a moving vehicle can also provide a rich temporal prior. Earlier frames captured within a time window of the current scene often share a high degree of visual coherence (especially for objects in the distance) which can be leveraged in scene understanding algorithms. This idea is illustrated in Figure 3, where the "temporal prior" is captured from a car driving down a highway several seconds before the "current image".

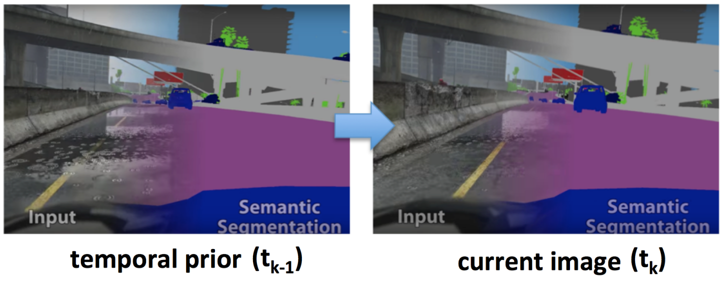

Figure 3: Temporal Priors: Earlier frames captured within a time window of the current scene often share a high degree of visual similarity and structure.

Modeling a prior is a challenging task. Some objects, such as cars and pedestrians, are mobile and are not in the same location between frames (or timesteps). Also, the appearance of the scene can shift quite a bit, depending upon the speed of the vehicle. For these reasons, it can be difficult to know which semantic labels should be selectively propagated from the semantic segmentation prior to accurately represent the current scene. There are naive approaches for transferring this information, such as estimating the motion shift between frames, something more useful for static priors. However, we want to use machine learning to learn from driving data how to more accurately transfer information from the prior. This will help avoid problems like spurious labeling in the final segmentation.

This research seeks improve an autonomous vehicle's situational awareness of its environment by using a machine learning-based approach with spatial and temporal priors to improve scene understanding algorithms.

2 Methodology

In the past several years, deep learning has come to the forefront of machine learning, computer vision and robotics research, proving to be highly effective by generating state-of-the-art results in several areas such as object detection and classification [7]. This work has recently been extended to the field of autonomous driving with success. In the past year, NVIDIA created a deep learning-based system, known as PilotNet, which is able to predict steering angles and identify relevant object on the road by using images an input [2].

More specifically, deep learning uses feed-forward, multi-layered convolutional neural networks (CNN) which can take various types of image input and produce output such as bounding box detections, scene depth estimation or pixel-level classification of objects in a scene. The latter is called semantic segmentation where each pixel in an image is labeled with one of a set of pre-defined classes and is used for our scene understanding task.

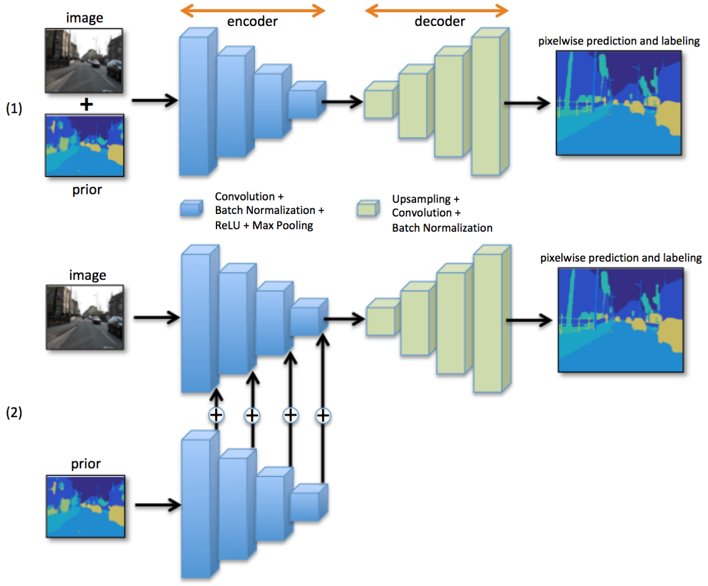

Figure 4: Model Architecture: Two variants of a convolutional neural network with an encoder-decoder architecture which takes an image and semantic segmentation as a prior (either spatial or temporal) to produce a semantic segmentation of the input scene.

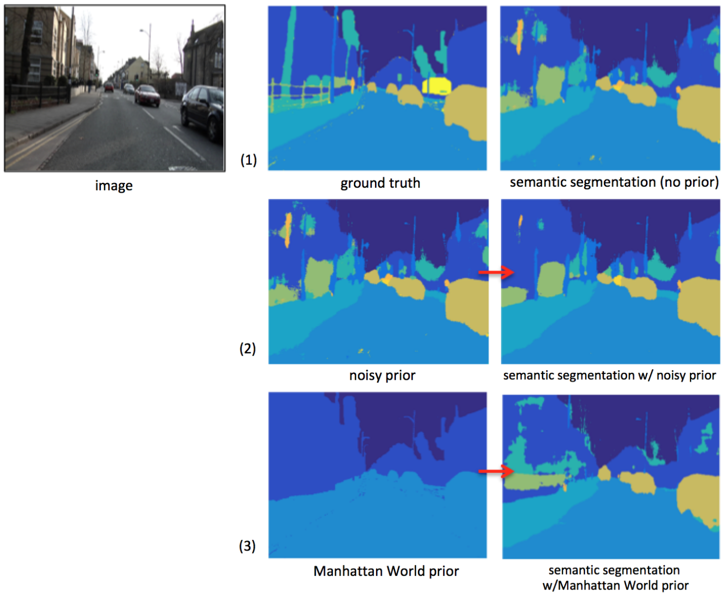

Figure 5: Semantic Segmentation: Qualitative comparison of semantically labeled images for networks which use and do not use priors.

Each network has an architecture that is specifically designed to the task at hand. In the case of our models, we use a fully-convolutional encoder-decoder architecture for semantic labeling, motivated by [1]. The strength of this type of model is that it is able to map the network's input ("encode") to a completely different representation ("decode"). The encoder projects raw pixel values into a lower dimensional representation by progressively convolving the input with a set of learned feature kernels and similarly the decoder progressively deconvolves (or upsamples) this representation into a semantically labeled image.

Network Architecture	Global	Class	Inter Over Union
(1) No Prior: All Classes	80.85	62.41	46.03
(2) With Noisy Prior: All Classes	82.33	61.89	47.57
(3) With Manhattan World Prior: All Classes	84.46	51.84	45.00
(1) No Prior: Dynamic Classes	78.34	78.84	34.62
(2) Noisy Prior: Dynamic Classes	76.30	80.57	37.54
(3) Manhattan World Prior: Dynamic Classes	51.62	85.13	26.14
(1) No Prior: Static Classes	80.98	58.63	50.32
(2) Noisy Prior: Static Classes	82.65	58.14	51.32
(3) Manhattan World Prior: Static Classes	86.18	54.33	52.07

Table 1: Multi-Class Comparison: Pixel-wise labeling accuracy for semantic segmentation networks which use and do not use priors. Results for all classes, dynamic classes (car, person, bicycle) and static classes (road, sidewalk, signpost, fence, etc.). Dynamic classes comprise 2% of test dataset annotations. Network architectures correspond to cases illustrated in Figure 4.

We have trained several variants of the encoder-decoder architecture for scene understanding which incorporate prior knowledge at different levels in the network (see Figure 3). We define a scene prior as a semantically segmented image of a given location which has been captured in a spatially or temporally similar way. Each model is trained using the CamVid road scene dataset [4] which contains several driving sequences with object class semantic labels, collected at various times of the day. The challenge for each model is to be able to label eleven classes such as road, sidewalk, building, car, pedestrians, bicycle, sign, poles, sidewalk, etc.

3 Results and Discussion

The preliminary results of our experiments indicate that using priors, when available, increases the overall accuracy of the semantic labeling in the test set. The performance of the scene segmentation can be measured using three standard metrics: global accuracy, class accuracy and intersection-over-union. Global accuracy is the overall mean per-pixel labeling accuracy and class accuracy is the mean class-wise accuracy. Intersection-over-union is the average of the intersection of the prediction and ground truth regions over the union of them.

Table 1 shows a quantitative comparison of the network architecture variants, where architecture (1) (no prior) is the baseline . Two different types of priors were used during model training, a "noisy" prior which was previous semantic segmentation of the scene (as described in Section 1) and a Manhattan World prior (see Figure 5, 3rd row) which simplifies class labels into approximately three planar surfaces (horizontal: sidewalks, roads, vertical: buildings, poles, trees, pedestrians, overhead: sky). Both priors were obtained from semantic segmentation models trained solely for generating priors. The global accuracy increased both for models (2) and (3) of per-pixel labeling increased by 1.5 to 4.4% and the intersection-over-union for model (2) increased by approximately 1.5%. The improvement in (3) indicates that just by collapsing object classes into orientation-similar labels, a deep learning network is able to learn well from a simplified structural prior. One reason for this may be that noisy spurious labels have been removed from the prior, decreasing the amount of error propagated.

An example where a spatial prior is applied can be seen in Figure 5, where a comparison models with no priors, noisy and Manhattan World priors is shown. The semantic segmentation with no prior (1) contains a lot of noisy artifacts of mislabeled classes. Even when a noisy prior (with some mislabeling of pixels) of the scene is used (2), the effect is that the final segmentation is much cleaner and smoother around object boundaries. The prior acts as an extra weighting in the feature maps that are produced by convolutional kernels at each level of the network. The final segmentation for the Manhattan World prior (3) lacks much of the mislabeled pixels propagated in (2) which can be attributed to using a simpler, cleaner prior.

4 References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[2] Mariusz Bojarski, Philip Yeres, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Lawrence D. Jackel, and Urs Muller. Explaining how a deep neural network trained with end-to-end learning steers a car. *CoRR*, abs/1704.07911, 2017. URL `http://arxiv.org/abs/1704.07911`.

[3] Neil E Boudette. Building a road map for the self-driving car. *The New York Times*, Mar 2017. URL `https://www.nytimes.com/2017/03/02/automobiles/wheels/self-driving-cars-gps-maps.html`.

[4] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, xx(x):xx–xx, 2008.

[5] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun. 3d traffic scene understanding from movable platforms. *Pattern Analysis and Machine Intelligence (PAMI)*, 2014.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`.

[8] Guowei Wan, Xiaolong Yang, Renlan Cai, Hao Li, Hao Wang, and Shiyu Song. Robust and precise vehicle localization based on multi-sensor fusion in diverse city scenes. *CoRR*, abs/1711.05805, 2017. URL `http://arxiv.org/abs/1711.05805`.