# Using A Priori Knowledge to Improve Scene Understanding

Brigit Schroeder
University of California, Santa Cruz
Intel AI Lab
brschroe@ucsc.edu

Alexandre Alahi
Visual Intelligence for Transportation Lab, EPFL
alexandre.alahi@epfl.ch

## Abstract

*Semantic segmentation algorithms that can robustly segment objects across multiple camera viewpoints are crucial for assuring navigation and safety in emerging applications such as autonomous driving. Existing algorithms treat each image in isolation, but autonomous vehicles often revisit the same locations. We propose leveraging this a priori knowledge to improve semantic segmentation of images from sequential driving datasets. We examine several methods to fuse these temporal scene priors, and introduce a prior fusion network that is able to learn how to transfer this information. Our model improves the accuracy of dynamic object classes from 69.1% to 73.3%, and static classes from 88.2% to 89.1%.*

## 1. Introduction

An autonomous vehicle is typically outfitted with several sensor modalities which can be used for mapping the environment [1] through which it drives (e.g Google self-driving cars continuously map the campus and streets of Mountain View, CA) [2]. *A priori* knowledge of a given area can be derived from data collected created during previous traversal through an intersection. These scene priors, in the form of temporal video frames, can be incorporated into scene understanding algorithms to improve the semantic segmentation of the scene. Earlier frames captured from a moving vehicle, 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. As seen in Figure 1, the image on the left provides a strong prior spatially: 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. Both recorded and live video (e.g. data previous to the current scene) provide a rich temporal prior and are a source of often unleveraged data that can enhance scene understanding.

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 time steps). The appearance of the scene can shift slightly, depending on the speed of the objects. Therefore, it can be difficult to discern which semantic labels to propagate from the prior to accurately inform the current scene. Naive approaches for selection, such as estimating the motion shift between frames, are more useful for static scenes. Here we use a learned module to determine from raw driving data how to propagate information from the prior.


Figure 1. **Scene Prior.** The image on the left is a temporal prior (one second previous) to the image on the right (representing the current scene).

There are existing video-based approaches for semantic segmentation. Some are limited to only segmenting a limited number of objects per scene[3] or require the use of optical flow networks [4] into the overall architecture, thereby increasing the complexity and processing time of the network. We show in our preliminary approach that we are able to do full scene segmentation with multiple frames using a low complexity model. This method could potentially be adapted to other semantic segmentation frameworks.

## 2. Methodology

### 2.1. Prior Fusion Network Architecture

We use a fully-convolutional encoder-decoder architecture for semantic labeling, motivated by SegNet [5]. These models feature a bottleneck stage, where the input image in projected to a lower dimensional representation. We hypothesize that this bottleneck representation could serve as the location for incorporating prior knowledge before the

decoder network expands the representation. We define a scene prior as an image of a given location which has been captured at an earlier time step (such as frames preceding the current frame). For our experiments, we use a prior that was captured one second earlier. Early experiments showed that using a frame too far in the past to be more detrimental than helpful as the differences in the scenes (both structurally and visually) were too high.

We tested three architectures:

- **Baseline.** Our baseline is a fully convolutional network with eight layers. The encoder has 64-128-256-512 features, and the decoder has 512-256-128-64 features. This baseline had no access to the temporal prior.
- **Embedding Prior.** In this approach, both the prior $x_0$ and the image $x_1$ are passed through an encoder (Figure 2, top). To fuse the representations, we use a weighted sum with a $tanh$ activation function (module $A$ in Figure 2), an idea borrowed from recurrent neural networks [6].
- **Decoder Prior.** This model is similar to the embedding prior, except the prior is applied to the decoder, and the features fused at each level of the decoder.

Importantly, the weights for the encoder-decoder are shared between the prior network and the image network, so the only difference in parameter count between the three models above are small contributions from the $A$ modules. In early experiments, we also tested a naive approach of concatenating the bottleneck representations, but the model performed poorly, so we exclude this model.

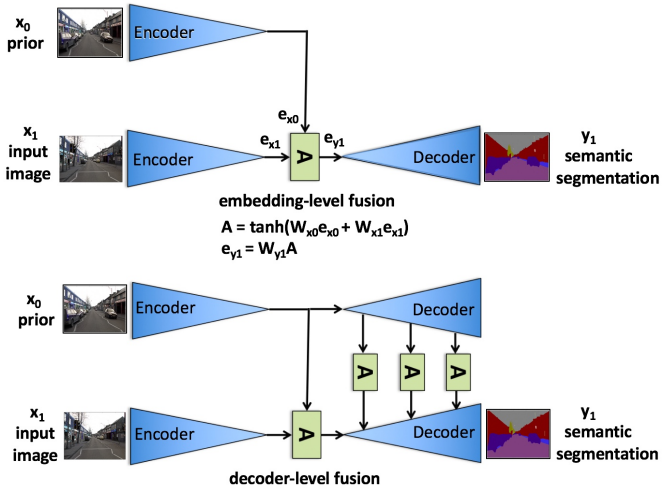


Figure 2. **Network Architectures.** A compact encoder-decoder network architecture is used with the addition embedding-level (top network architecture) and decoder-level (bottom network architecture) prior fusion. Prior features are combined using *tanh* activation functions, similar to those in recurrent networks.

| | Global | Class | IoU |
|---|---|---|---|
| **(1) Baseline (no prior)** | 87.3 | 62.8 | 54.7 |
| **(2) Embedding Prior** | 88.0 | 60.5 | 53.7 |
| **(3) Decoder Prior** | **88.4** | **63.1** | **55.5** |

Table 1. **Semantic Segmentation Evaluation.** Performance of three semantic segmentation network variants on the CamVid test set, evaluated using global pixel accuracy, class accuracy and intersection over union. For model descriptions, see Section 2.

| | Static Objs | Dynamic Objs |
|---|---|---|
| **(1) Baseline (no prior)** | 88.2 | 69.1 |
| **(2) Embedding Prior** | +0.7% | **+4.2%** |
| **(3) Decoder Prior** | **+0.9%** | +3.8% |

Table 2. **Global Accuracy of Static and Dynamic Classes.** Comparison of classes which are divided into static objects (e.g. buildings, roads, trees, etc.) and dynamic objects (e.g. pedestrians, car, bicycles, etc.). The addition of a prior increases the accuracy of both types of objects.

### 2.2. Dataset

Each model is trained using the CamVid road scene dataset [7] which contains several driving sequences with object class semantic labels, collected at various times of the day. There are 367 train images and 233 test images. Due to the small size of the dataset, models were initially trained with 227 x 227 random image crops from the full 360 x 480 image as in [8], and then final models were fine-tuned from these models using the full-sized images.

## 3. Experiments and Results

We measured performance of scene segmentation using three standard metrics [9]: global accuracy, class accuracy and intersection-over-union (IoU). 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. As shown in Table 1, models that incorporate priors (Decoder Prior and Embedding Prior) outperform the baseline across all three metrics in many cases.

Prior fusion improves upon all metrics when done at the decoder level over the baseline. The global accuracy of per-pixel labeling increased both for fusion models (2) and (3) for per-pixel labeling, more than 1% in the best performing model. Embedding prior fusion contributes to an increase in global accuracy but a decrease in class accuracy and IoU (60.4% versus 62.81%), suggesting that only fusing at the bottleneck layer, which has rich features but poor spatial resolution, only benefits specific classes. In contrast, when priors are fused at different feature resolutions throughout

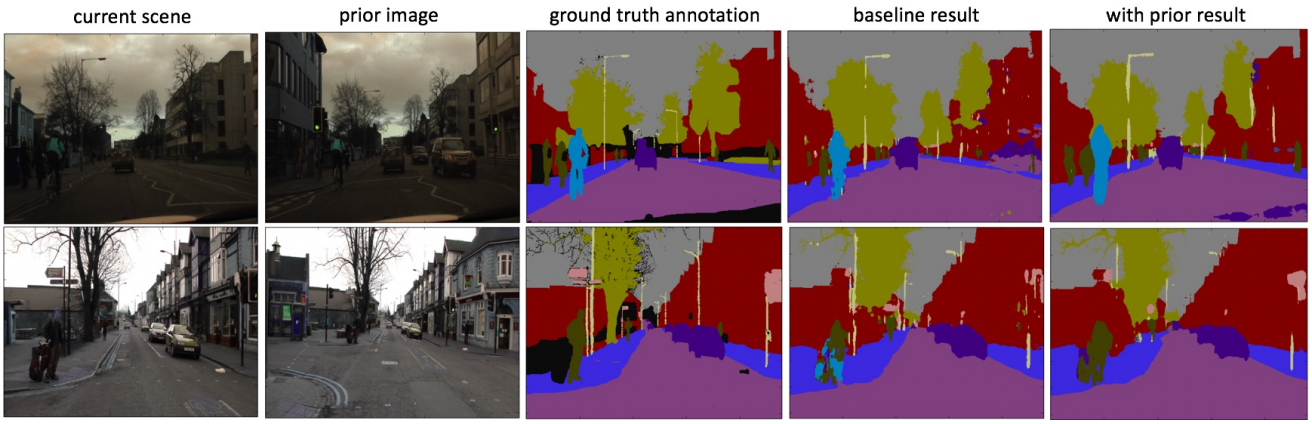


Figure 3. **Semantic Segmentation with Priors.** Qualitative comparison of semantically labeled images for networks which use and do not use priors. Results from the baseline model are shown in the fourth column and Decoder Prior (our best-performing model) in the last column.

the decoder, both fine-grained and coarser feature classes see a gain in class accuracy and IoU.

The performance improvement from incorporating priors is significantly enhanced when we examine dynamic versus static objects. We divided the CamVid object classes into static objects (e.g. buildings, roads, light posts, signs, trees, etc.) and dynamic objects (e.g. pedestrians, cyclists, cars, etc.). The global accuracy for both for is reported in Table 2. Importantly, we observed that the priors had a significant impact on the semantic segmentation of dynamic objects (73.3% versus 69.1%), which tend to be of smaller size and lower frequency than the static objects.

Overall, priors decrease the spurious semantic labeling of pixels, which can be seen in Figure 3. Note that the prior model (fifth column) reduces a lot of noise in the pixel labeling and improves the labelling for fine-grained feature classes such as pedestrian and street sign, when compared to the baseline model.

## 4. Discussion and Future Work

We have demonstrated that the addition of prior knowledge to a deep convolutional network can increase the performance of semantic segmentation, particularly for dynamic objects. We introduce a method using a learned fusion module to incorporate prior information, and demonstrate that fusing at multiple feature resolutions improves performance. This general technique could be applied to other models beyond the encoder-decoder model architecture.

In future work, we plan to examine how multiple priors across time steps could benefit semantic segmentation, as well as applying perceptual loss approaches derived from the prior image. Other representations of the scene, such as scene graphs which encode semantic relationships between objects in an image, could also be provided as prior knowledge for the network to exploit.

## References

[1] Guowei Wan, Xiaolong Yang, Renlan Cai, Hao Li, Yao Zhou, Hao Wang, and Shiyu Song. Robust and precise vehicle localization based on multi-sensor fusion in diverse city scenes. In *2018 IEEE International Conference on Robotics and Automation, ICRA*. 1

[2] Neil E Boudette. Building a road map for the self-driving car. *The New York Times*, Mar 2017. 1

[3] Sepehr Valipour, Mennatullah Siam, Martin Jägersand, and Nilanjan Ray. Recurrent fully convolutional networks for video segmentation. In *2017 IEEE Winter Conference on Applications of Computer Vision, WACV*, pages 29–36, 2017. 1

[4] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *2018 Computer Vision and Pattern Recognition*. 1

[5] 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. 1

[6] Jeff Donahue et al. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans. Pattern Anal. Mach. Intell.* 2

[7] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 2008. 2

[8] Simon Jégou et al. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *2017 IEEE CVPR Workshops*. 2

[9] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and José García Rodríguez. A review on deep learning techniques applied to semantic segmentation. *CoRR*, abs/1704.06857, 2017. 2