Self-supervised approach for free space estimation

By: Anand Uday Gokhale, Sumanth R Hegde

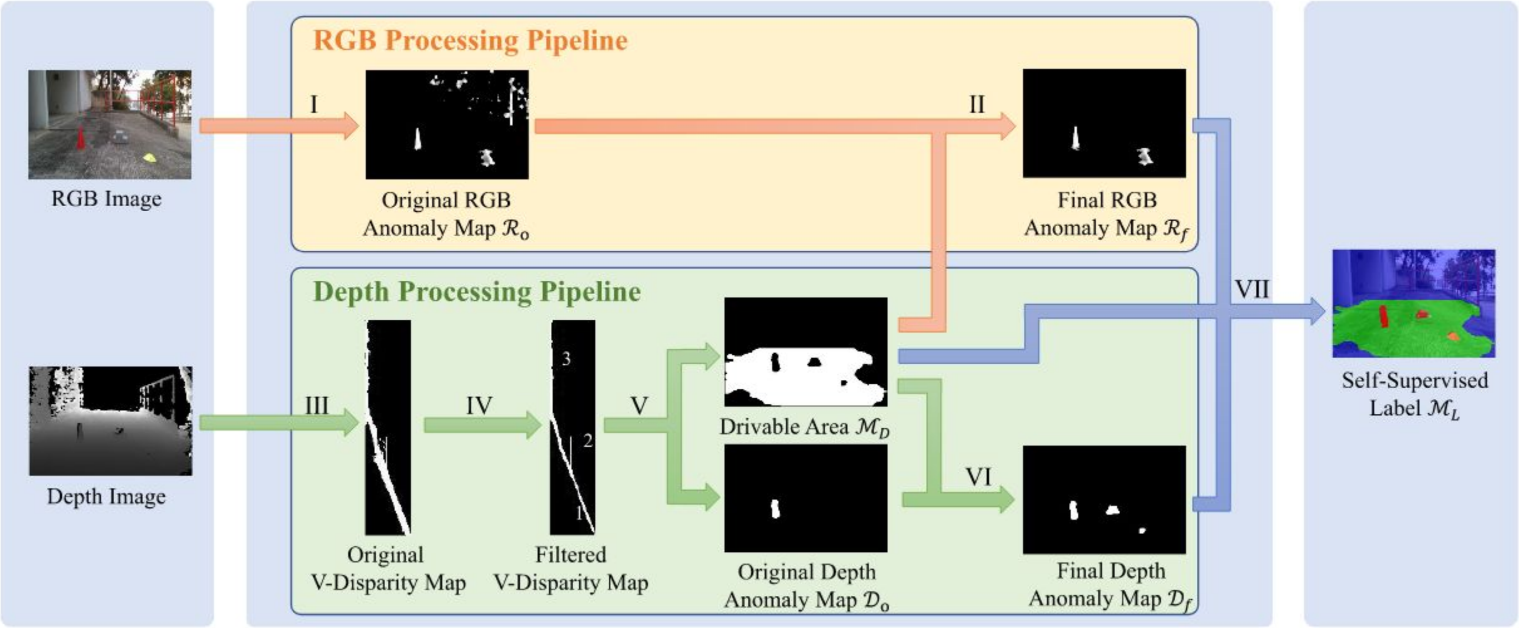
Overview of work done

- Analysed performance of existing SOTA road segmentation algorithms on the India Driving Dataset (IDD)
- Experimented with CRFs for generating weak labels from predictions of SegNet model
- Trained a self supervised model, using predictions from a recent SOTA algorithm as weak labels.
- Employed a multimodal fusion scheme to achieve better mean Intersection-over-Union results on IDD validation (val) set.
- Analysed the difference between Kitti and IDD

Previous work

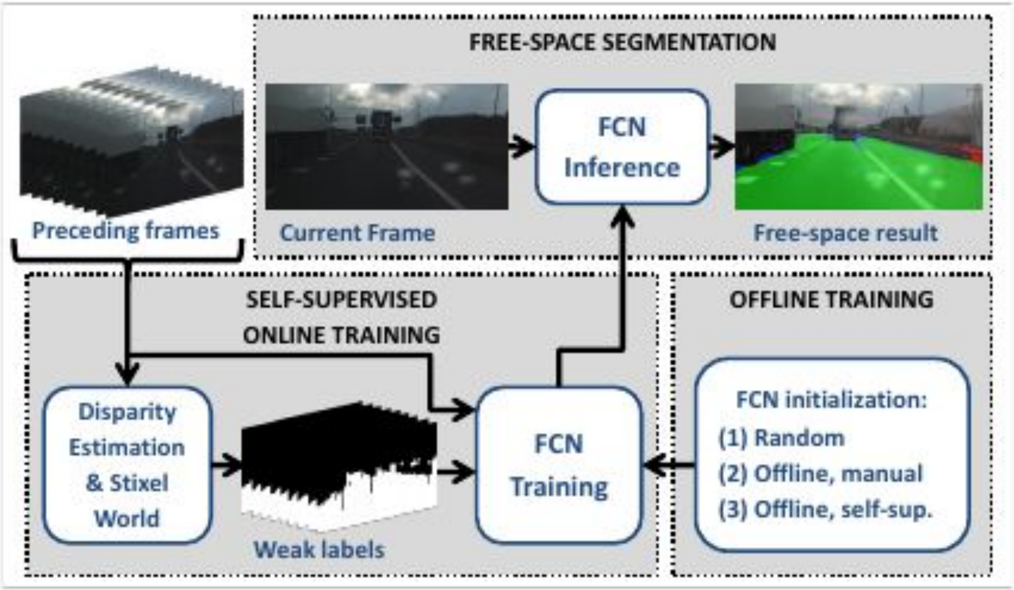
[1]

Using the V-disparity map



[2]

A Self supervised FCN-based approach



A CRF based approach

[3]

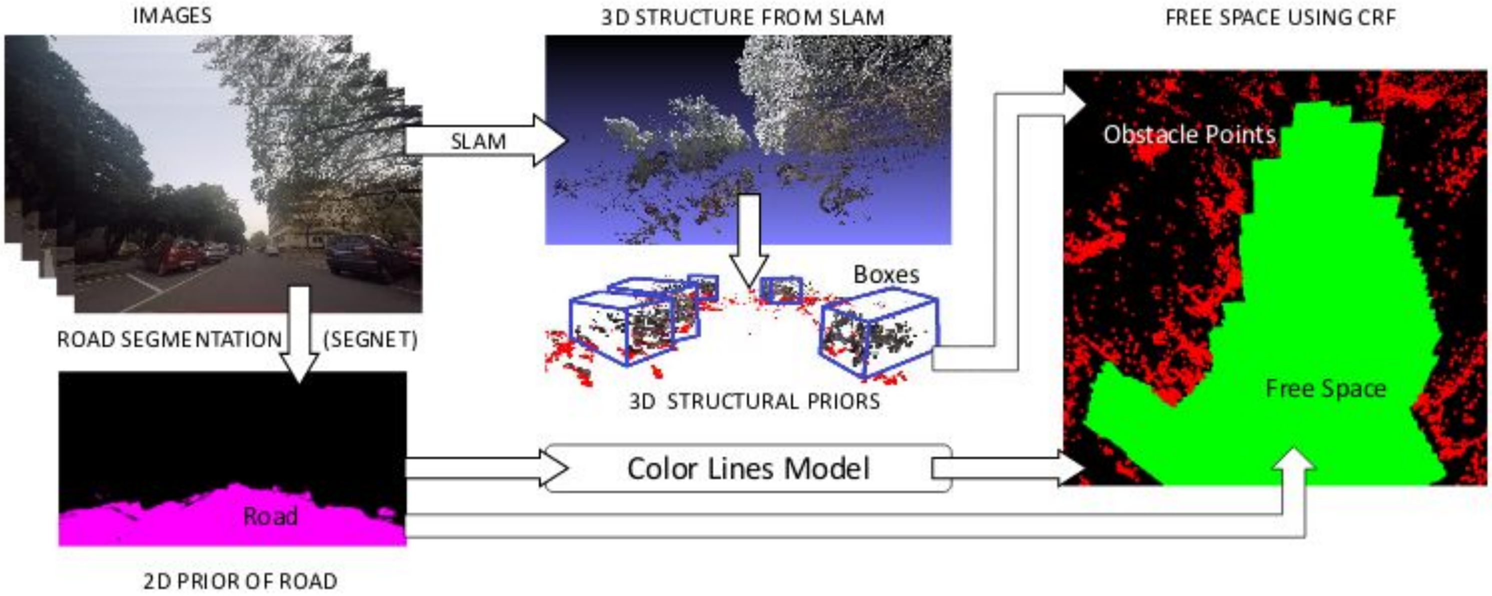
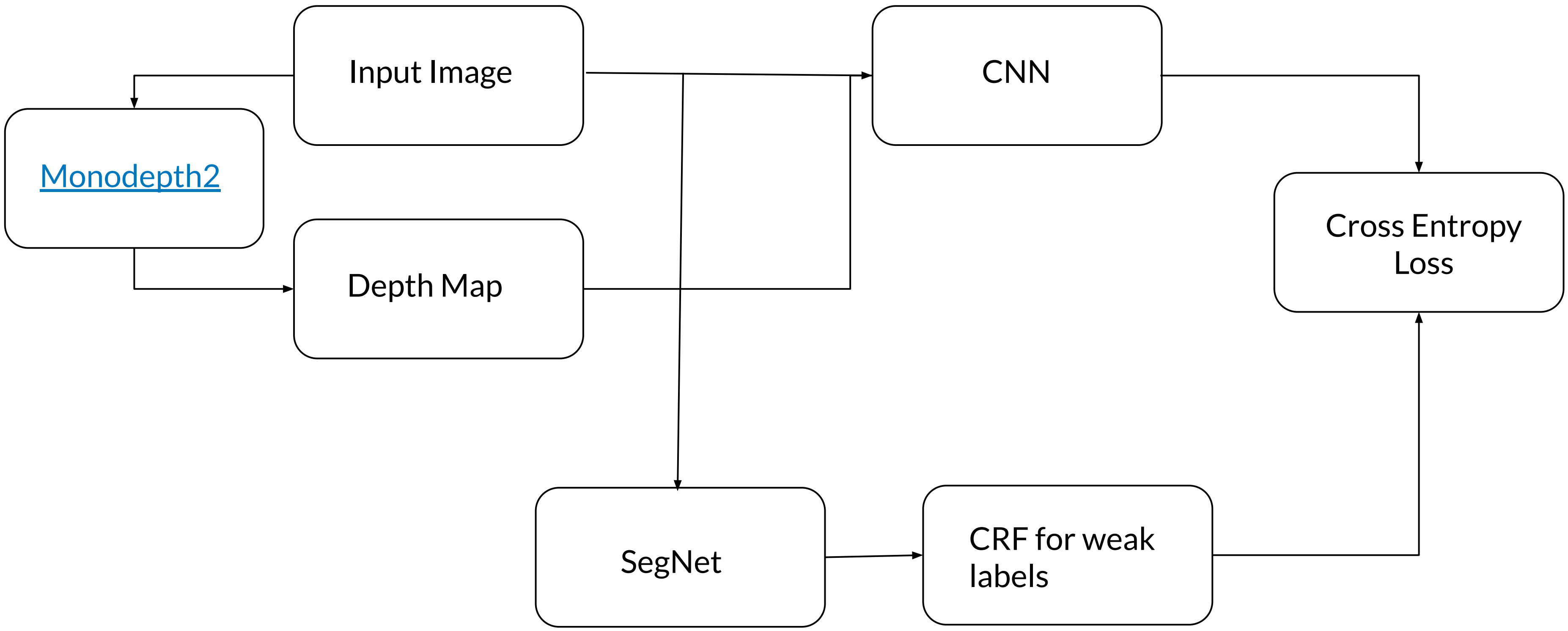
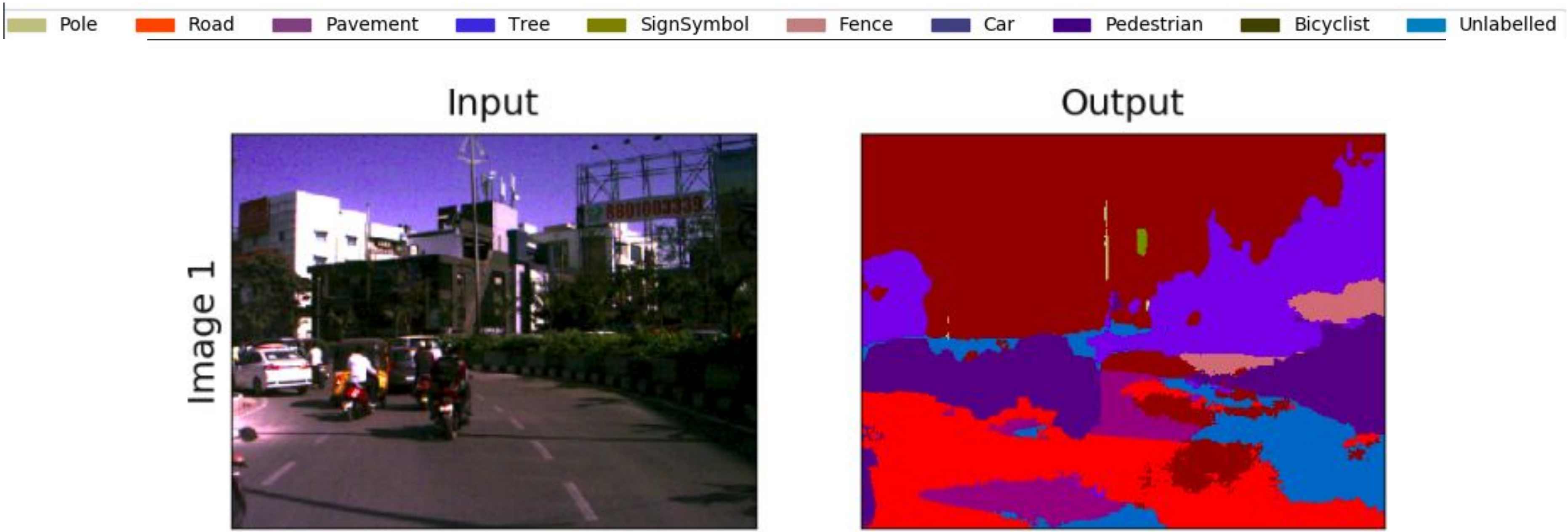


Figure 2. Overall framework of our method

Our Initial Approach





[SegNet](https://) predictions

CRF formulation

$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j),$$

Pydensecrf :

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \underbrace{\sum_{m=1}^{K} w^{(m)} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j)}_{k(\mathbf{f}_i, \mathbf{f}_j)},$$

Our CRF Formulation

- High confidence predictions from SegNet (for the road class) are used as part of our prior.
- SegNet mis-classifies certain road pixels as a different class with high probability.
- We also use predictions about classes (such as “sky”) that were never assigned to a true road pixel as part of our prior.
- Pairwise features include color, position and gradient of the depth map.
- However, without explicit information about obstacles, the algorithm performs poorly.

Challenges faced

- Needed a reliable way to obtain information about obstacles on the road.
- In the CRF formulation, the parameters were not learnable and thus hand-crafted.

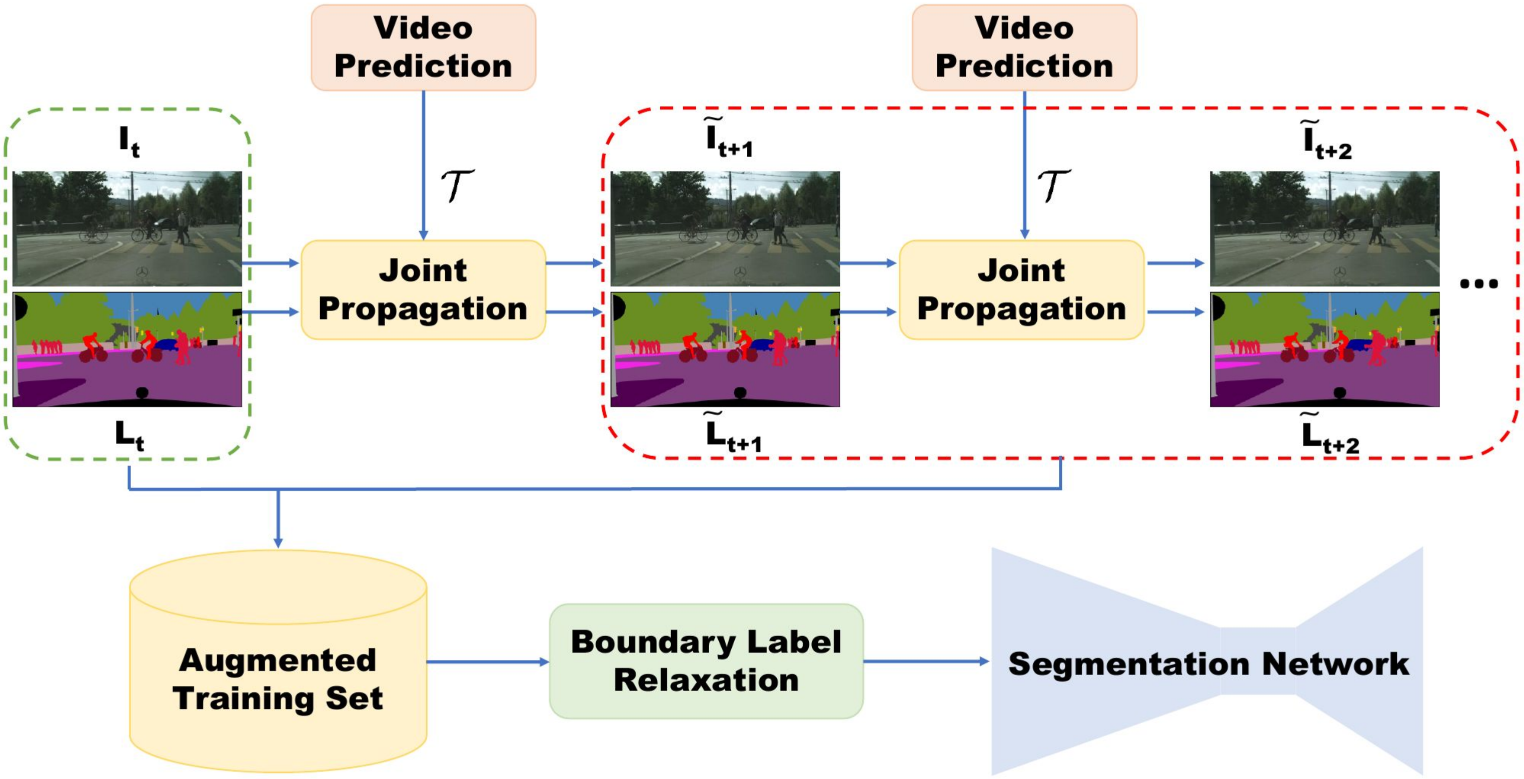
Result after the CRF : Input prior used was high confidence pixels of building, road and sky class

Experimentation with newer models

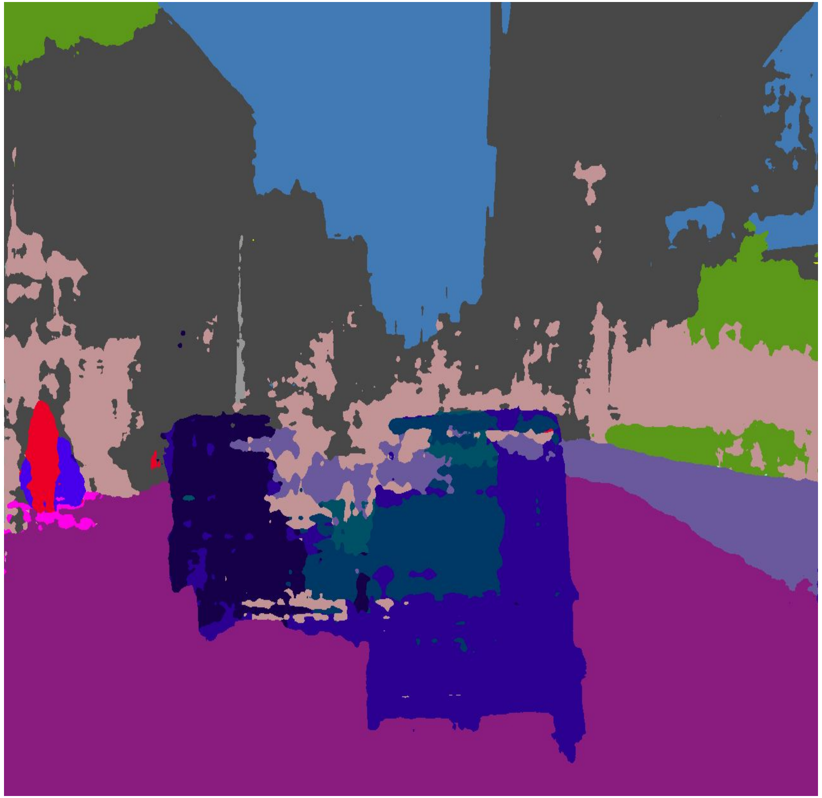
- A recent 2019 paper by NVIDIA achieves state of the art performance in semantic segmentation via label propagation and boundary relaxation.
- As an input pre-processing step, the algorithm performs a simple mean and std. deviation transform (hereon referred to as the “data transformation”)
- Road segmentation results obtained by this model were much more reliable

[4]

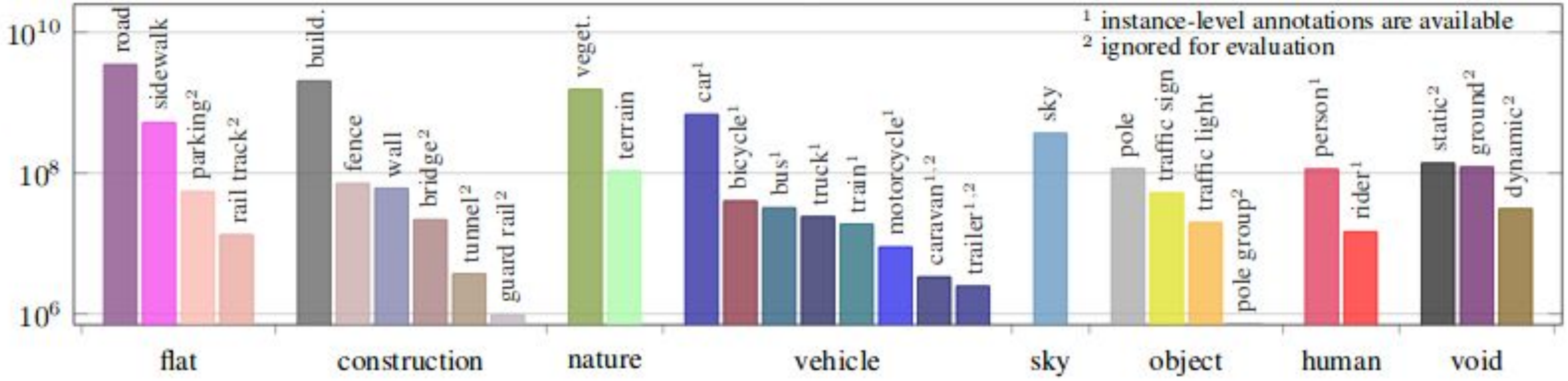
Semantic Segmentation via Video Propagation and Label Relaxation



Example Output



Label Legend for Cityscapes



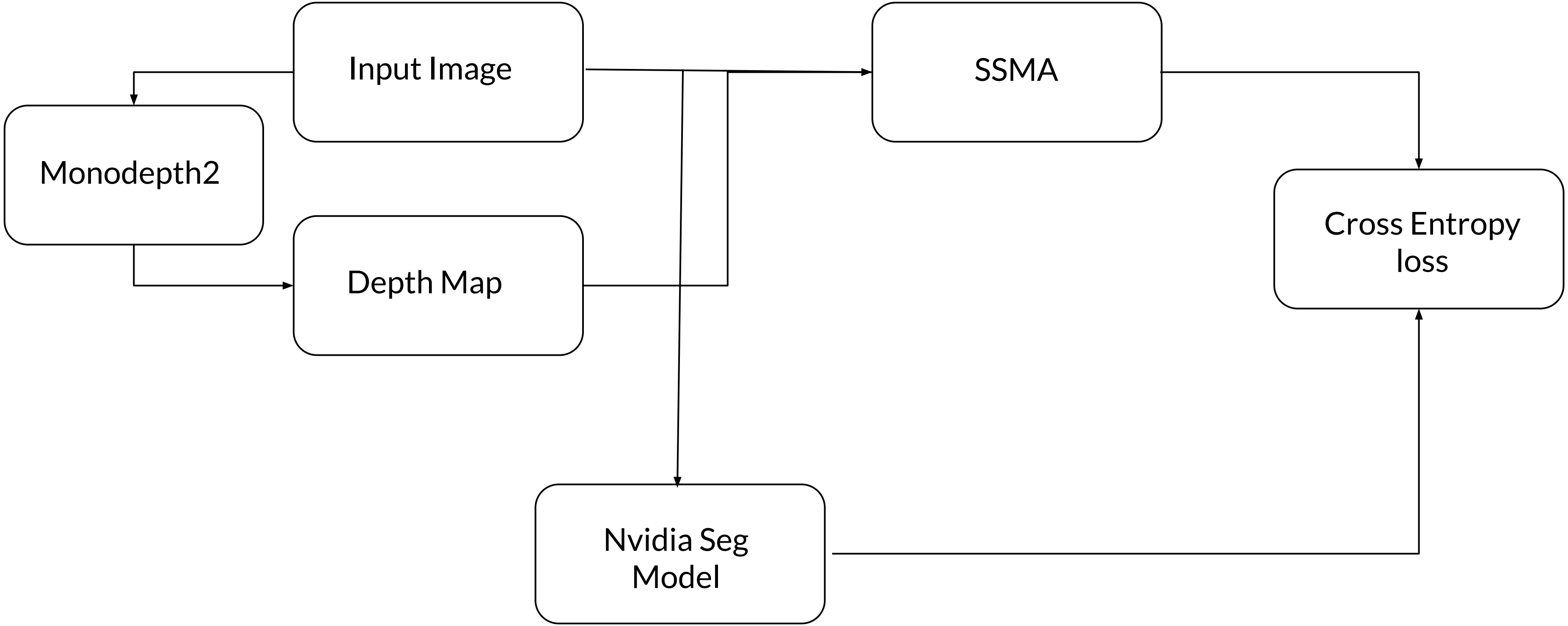
The Data transformation

An image from the India Driving Dataset

An Image from KITTI before and after the transformation

An Image from Cityscapes before and after the transformation

Our approach

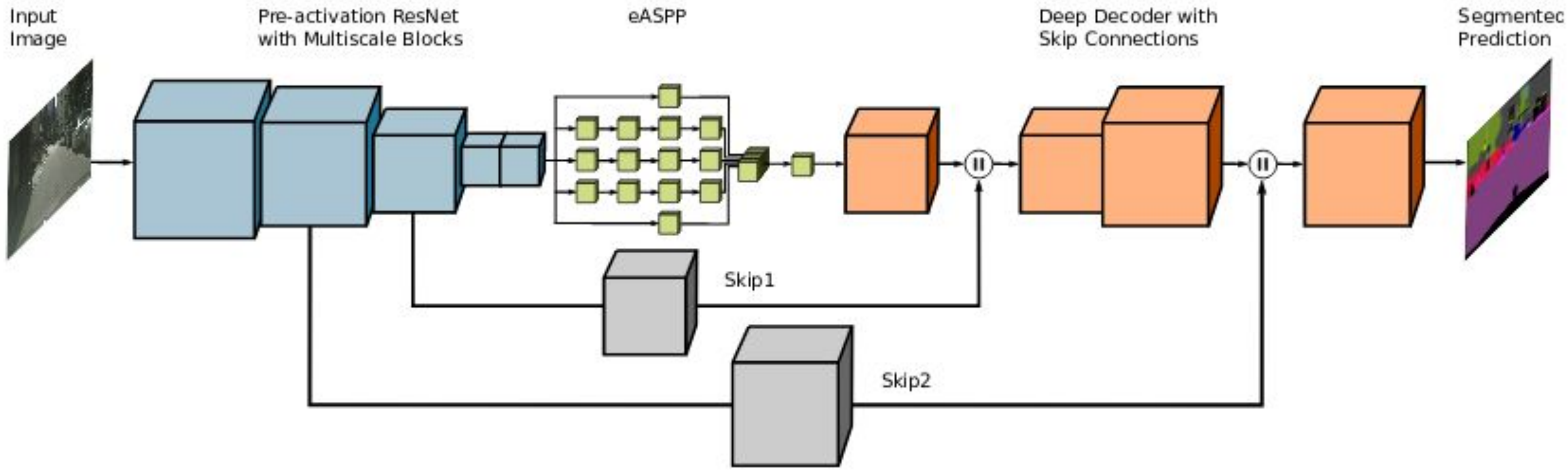


Self-Supervised Model Adaptation [5]

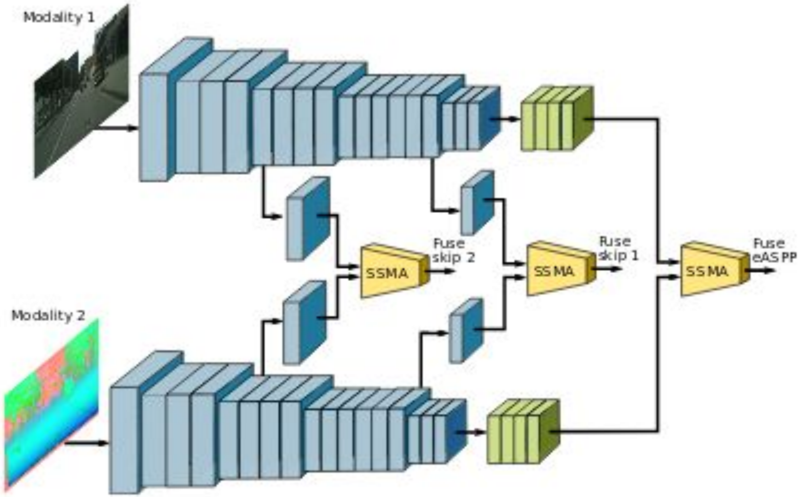
- Illumination and varying weather conditions can significantly affect many current segmentation methods that are trained solely on RGB data

- Complementary modalities can enable learning of richer representations that are resilient to such perturbations

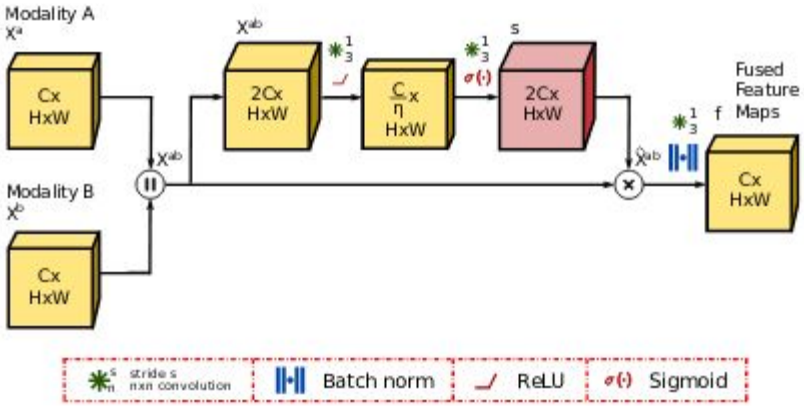
- SSMA is a novel multimodal fusion scheme that combines multimodal data using an attention scheme.



The Encoder Architecture : AdapNet++, an improvement over Adapnet



Combining information from different modalities (Encoder stage)



The Self Supervised Model Adaptation (SSMA) block

The SSMA model architecture

Training details

- Two arms of the SSMA architecture - individually trained; one with RGB input, one with depth map as input.
- Depth maps are prepared using Monodepth 2 model.
- For the combined architecture, we trained using the weights obtained above, for both arms, as initialization.
- Along with the standard cross entropy loss, we also used two auxillary loss functions, with outputs at intermediate stages of the architecture.

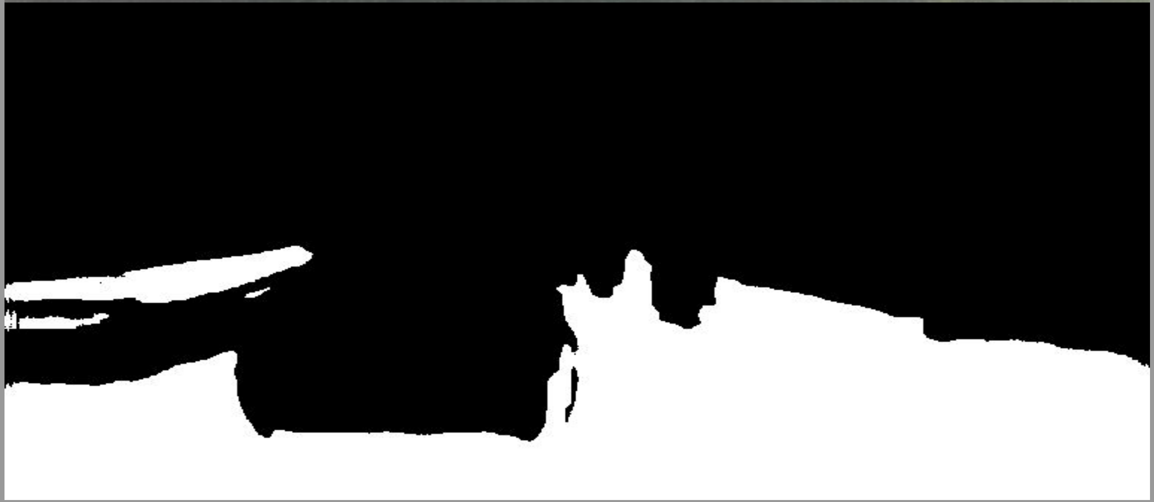
AdapNet++ training details

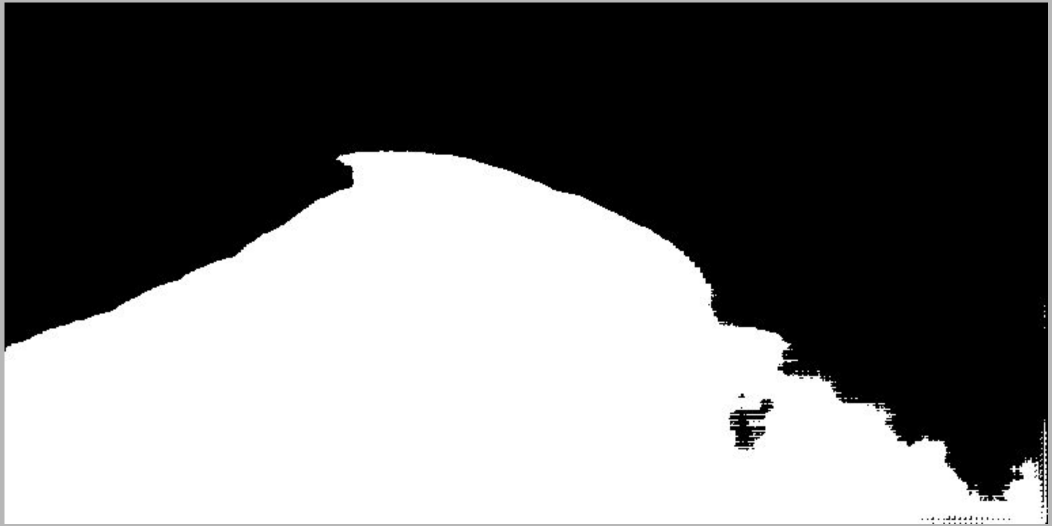
- RGB arm :
  - The train set was augmented using random flipping and cropping.
  - Model was trained on the (RGB, weak labels) pair for 20,000 iterations with a batch size of 8.
  - Initial learning rate of 0.001 with a polynomial decay.
  - Performance improved with the use of auxillary loss functions
- Depth arm :
  - Model was on the (Depth map, weak labels) pair for 10k iterations with other configurations being same as above.

SSMA training details

- Augmentation using flipping and cropping, increased train data from 6993 image triplets to 27000.
- Batch size: 16
- Number of steps: 20000
- Initial learning rate: 0.001
- Learning rate with polynomial decay has power: 0.9

Results





Comparative Results

Model	SegNet	Deeplabv3	Nvidia (w-t)*	Nvidia	**Ours**
mIoU on val set (981 img)	0.775 (std : **0.19)**	0.774	0.782	0.852	**0.858**
mIoU on train set** (6993)	-	-	-	0.822	**0.852**

** - Results quoted only for best performing models

* w-t : Without the "data transformation"

Failure cases


MOTOR CAB
TS08UA6906

Predictions by the Nvidia model :
captures fine object boundaries

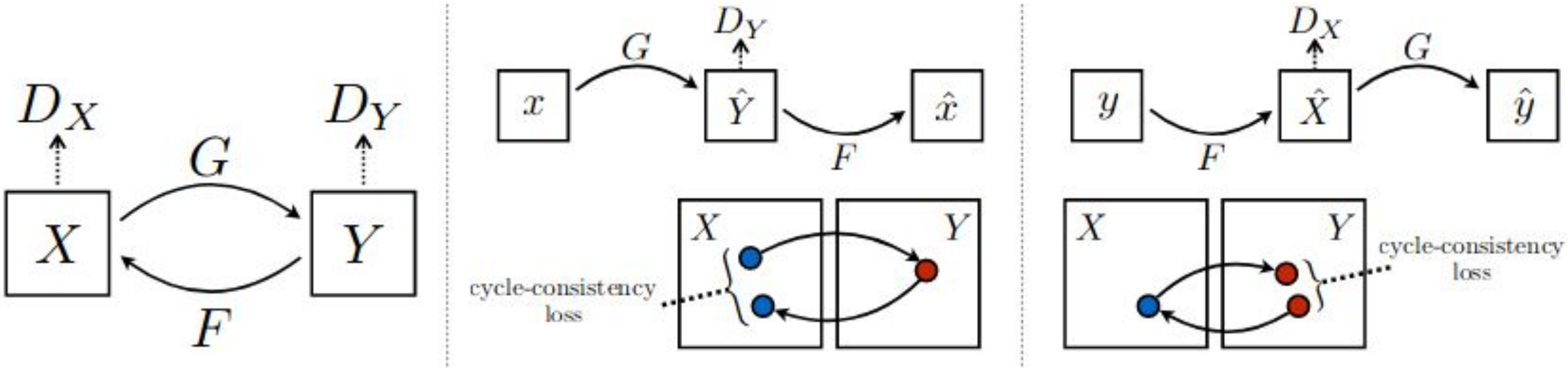
IDD vs KITTI/Cityscapes

- SegNet fails miserably on many images where there are large shadows, bad illumination,etc
- We wanted to further investigate what changes in an image of IDD made it more like an image in KITTI/Cityscapes.
- We observed that overall increment in brightness gave better results with Segnet, although unnaturally whitened the images.

CycleGAN : learning the transformation [6]

- We tried to capture the hidden style of an IDD image and a KITTI image
- Used CycleGAN to perform image-to-image translation task
- We note that the GAN attempts to remove the shadows, and sometimes increases colour saturation.

Model Architecture



$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y\sim p_{\text{data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))],$$

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y\sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1].$$

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) + \lambda\mathcal{L}_{\text{cyc}}(G, F),$$

Training Details

- Model was trained for 70 Epochs, over a dataset of 27000 images, obtained after augmenting the IDD train set, along with an equal number of KITTI images.
- Learning Rate : 0.0002, with Adam Optimizer
- Lambda = 10.
- We also experimented with using resize convolution instead of deconvolution in the generator, but the results showed no improvement.

Original
IoU of SegNet
predictions: 0.5390

IDD image in the
“style of Kitti”
IoU of SegNet
predictions: 0.6531

Original

IDD image in the “style of Kitti”

Unnatural artifacts

Original

IDD image in the “style of Kitti”

Future improvements

- Model performance can be improved using a semi supervised approach, where we incorporate about 100-200 true labels into our dataset. Also look at performance gains with a completely supervised approach.
- Multimodal information can be incorporated better using modified convolutional architectures, such as the recent Depth-aware CNN model.[7]
- A CRF algorithm can be used as a post-processing step for further improving results.
- Better modelling of transformations between datasets.

Conclusion

- Multimodal fusion is a promising approach for semantic segmentation.
- For the task of road segmentation, recent work demonstrates reasonable generalization across datasets.
- Self-supervised approaches, with multimodal data is a strong candidate for problems in low-data scenarios.
- Generative models can provide valuable insights into the difference between different datasets.

Thank You!

Questions?