

# Semantic Image Segmentation Guided by Scene Geometry

Sotirios Papadopoulos, Dr. Ioannis Mademlis, Prof.  
Ioannis Pitas

Aristotle University of Thessaloniki

[papadops@csd.auth.gr](mailto:papadops@csd.auth.gr)

[www.aiia.csd.auth.gr](http://www.aiia.csd.auth.gr)

# Introduction

- Semantic image segmentation networks work RGB features via RGB input images.
- Knowing the geometry of the depicted scene can provide useful info in complex areas (shadow-y spots, similar texture/color in adjacent semantically different objects etc.) for semantic segmentation to benefit from.
- *Trivial solution:* RGB+Depthmap as input to network.
- *Drawback:* need for RGBD datasets (difficult and costly to acquire).

# Introduction

- *Popular solution:* Multitask network for simultaneous estimation of depth maps and semantic segmentation maps.
- *Drawback:* Difficult to train (especially when the depth branch is trained with self-supervision), high computational complexity.
- *Proposed solution:*
  - Pretrain a separate depth estimation network (self-supervision),
  - Train an off-the-shelf semantic segmentation network to get semantic maps.
  - During training, force the output segmentation maps to share similar structure to the depth maps of the pretrained depth estimation network.

# Semantic Image Segmentation

- CNNs for Semantic image segmentation typically uses a cascade of an **encoding** and a **decoding subnetwork**.
- The final output of the decoder is a **semantic image map**, having:
  - **same spatial resolution** as the input and
  - as **many channels** as the object class number.
- **Per-pixel** image classification is performed.



# Semantic Image Segmentation

- Input image:  $\mathbf{I} = \{I_{ij}\}_{\substack{1 \leq i \leq N_1 \\ 1 \leq j \leq N_2}}, N_1, N_2 \in \mathbb{N}$ .
- Target:  $\mathbf{S} = \{S_{ij}\}_{\substack{1 \leq i \leq N_1 \\ 1 \leq j \leq N_2}}$  (semantic segmentation map)
- $S_{ij} \in \mathcal{C}$ : label of  $I_{ij}$ ,  $\mathcal{C}$  is the set of supported semantic class labels.
- Network output:  $\hat{\mathbf{S}} \in \mathbb{R}_1^{N_1 \times N_2 \times |\mathcal{C}|}$ , probabilities for each class for each pixel.

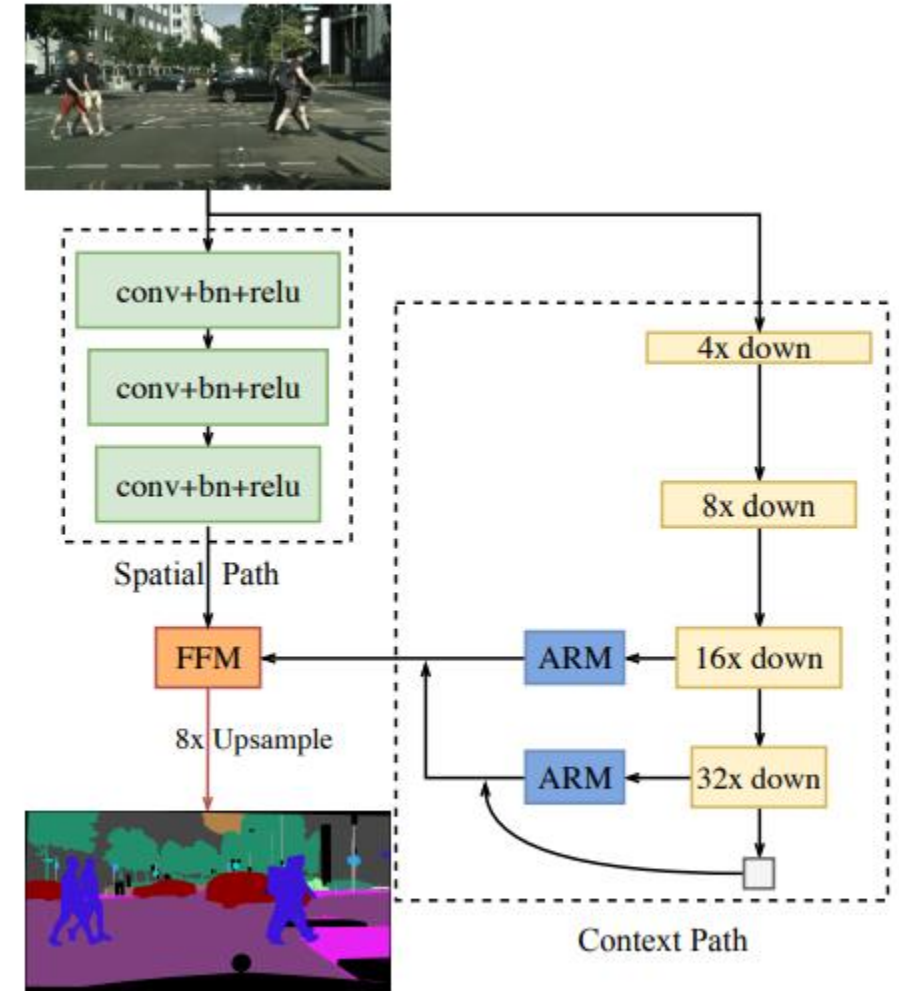
# Semantic Image Segmentation

Baseline network: BiSeNet [1]

- accurate real-time semantic segmentation
- two separate network branches:
  - **Spatial path**: a shallow branch to preserve spatial details, and
  - **Context path**: a deep lightweight feature extractor for high level context.

The two branches are later concatenated and fed to a shallow CNN module for the final prediction.

[1] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation.", In Proceedings of the European Conference on Computer Vision (ECCV), 2018.



(a) Network Architecture

# Self-supervised Depth estimation



Depth estimation from monocular image without supervision.

- Video:  $\mathcal{I} = \{\mathbf{I}_0, \dots, \mathbf{I}_t, \mathbf{I}_{t+1}, \dots\}$ .
- Depth map  $\mathbf{D}_t$  corresponding to  $\mathbf{I}_t$  is estimated with the help of  $\mathbf{I}_{t+1}$  (no ground truth depth map).
- Camera intrinsics matrix:  $\mathbf{K}$ .

# Self-supervised Depth estimation



Training:

- Estimate relative camera pose  $\mathbf{T}_{t \rightarrow t+1}$  between consecutive video frames  $\mathbf{I}_t$  and  $\mathbf{I}_{t+1}$  via a dedicated CNN.
- Find coordinates of the projection of  $\mathbf{p}_t \in \mathbf{I}_t$  on the plane of  $\mathbf{I}_{t+1}$ :

$$\mathbf{p}_{t+1} \approx \mathbf{K}\mathbf{T}_{t \rightarrow t+1}\mathbf{D}(\mathbf{p}_t)\mathbf{K}^{-1}\mathbf{p}_t.$$

- Transform  $\mathbf{I}_{t+1}$  to form an approximation  $\mathbf{I}'_t$  of  $\mathbf{I}_t$  via differentiable bilinear interpolation.



# Self-supervised Depth estimation



Training:

- Minimize photometric cost function:

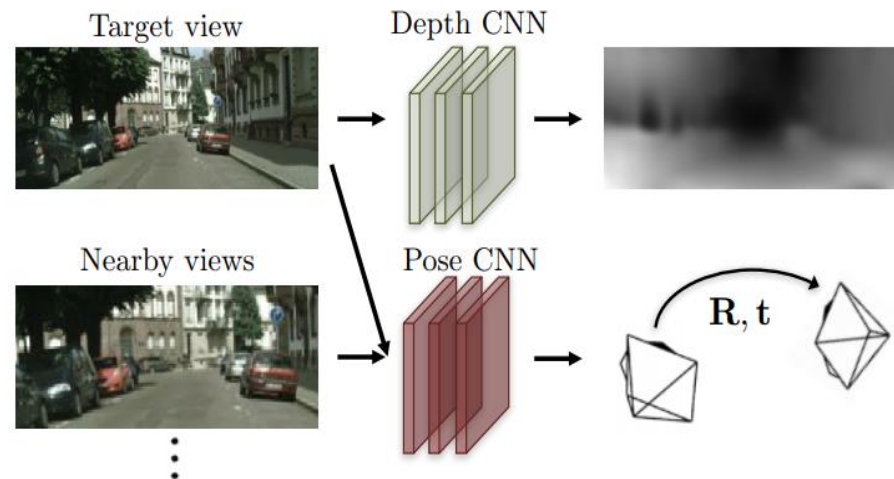
$$L_{photo} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{p} \in \mathcal{V}} \|\mathbf{I}_t(\mathbf{p}) - \mathbf{I}'_t(\mathbf{p})\|_1,$$

- Where  $\mathcal{V}$  is the set of pixels that fell exactly onto  $\mathbf{I}_{t+1}$  after the projection.

# Disparity/Depth map Estimation with NNs



(a) Training: unlabeled video clips.



(b) Testing: single-view depth and multi-view pose estimation.

Depth and pose estimation DNNs.

# Proposed Method

- Intuitive observation: semantic objects tend to stand out in depth maps → co-occurrence of image gradients in the two tasks.
- Idea: to enhance semantic segmentation accuracy, force semantic edges to be absent in areas where there are not any depth edges.
- Depth branch is used only for training, can be totally omitted during testing.

# Proposed Method

- Per-class consistency loss:

$$L_p = \sum_{c=1}^c \text{mean} \left( \left\{ \left| \frac{dS}{dx}(i, j, c) \right| \cdot e^{-\left| \frac{dD}{dx}(i, j) \right|} \right\}_{\substack{1 \leq i \leq N_1 \\ 1 \leq j \leq N_2}} \right) +$$

$$\text{mean} \left( \left\{ \left| \frac{dS}{dy}(i, j, c) \right| \cdot e^{-\left| \frac{dD}{dy}(i, j) \right|} \right\}_{\substack{1 \leq i \leq N_1 \\ 1 \leq j \leq N_2}} \right)$$

# Proposed Method

- Holistic consistency loss:

$$L_h = \text{mean} \left( \left\{ |S'_x(i, j)| \cdot e^{-\left| \frac{dD}{dx}(i, j) \right|} \right\}_{\substack{1 \leq i \leq N_1 \\ 1 \leq j \leq N_2}} \right) +$$

$$\text{mean} \left( \left\{ |S'_y(i, j)| \cdot e^{-\left| \frac{dD}{dx}(i, j) \right|} \right\}_{\substack{1 \leq i \leq N_1 \\ 1 \leq j \leq N_2}} \right)$$

- where  $S'_k = \left\{ \max \left( \left| \frac{dS}{dk}(i, j) \right| \right) \right\}_{\substack{1 \leq i \leq N_1 \\ 1 \leq j \leq N_2}}$ .

# Proposed Method

## Pros:

- No depth ground truth data required,
- No runtime overhead during inference,
- Does not require any architectural modifications to the semantic segmentation CNN.

## Con:

- An appropriate training dataset is difficult to find. Requires:
  - RGB images with semantic segmentation ground truth,
  - Images must be consecutive video frames,
  - Camera intrinsics matrix must be provided.

# Evaluation

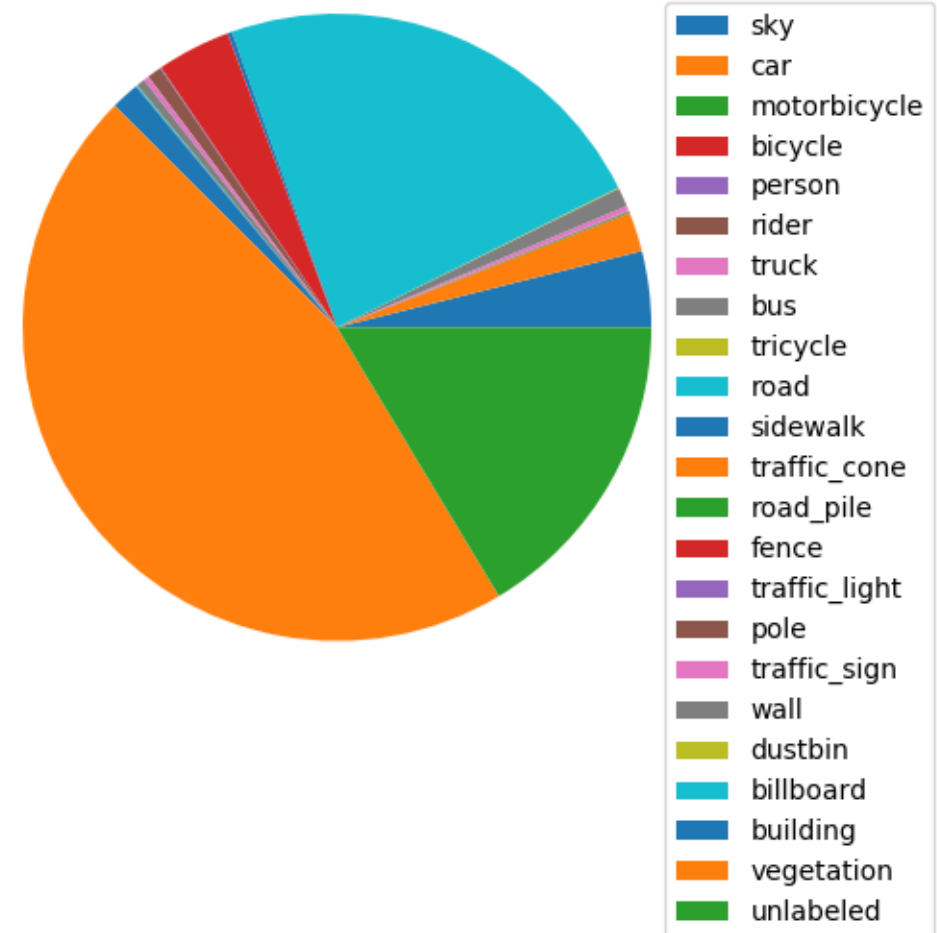
## Dataset:

- Apolloscape dataset: stereo RGB images of size  $3384 \times 2710$ .
- Video shot from a moving vehicle while driving on city streets.
- Contains semantic segmentation ground truth for every video frame.
- Camera intrinsics matrix is provided.



# Evaluation

Low mean IoU values are expected:  
 Some classes are too dominant.  
 A lot of poorly represented classes.

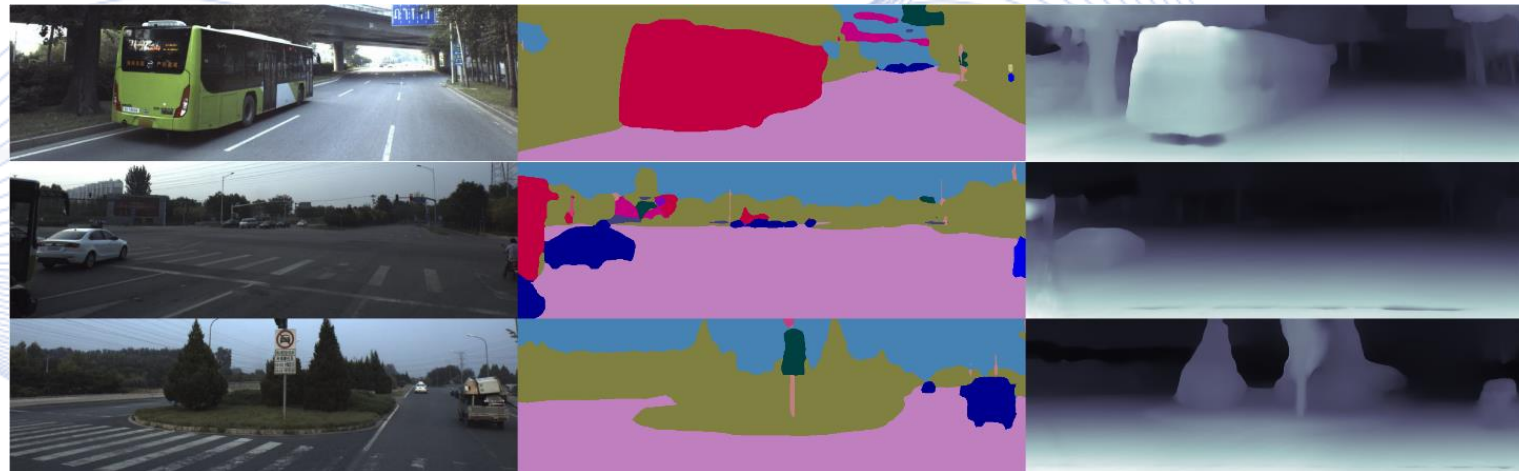




# Evaluation

## Specifications:

- Rule out problematic areas: cut upper part of each image (lots of sky pixels) as well as the lower one (filming car bonnet)
- Image rescaling to  $832 \times 256$ .
- Backbone network: ResNet-50.



# Evaluation

Method	Mean IoU	Inference runtime (msec)
Baseline (no depth) [1]	39.557%	<b>6.2</b>
[2] (multitask)	34.318%	6.4
Baseline + [3] (multitask)	37.683%	8.3
Baseline + [4] regularizer (pretrained)	39.610%	<b>6.2</b>
Baseline + [4] regularizer (multitask)	38.153%	9
Baseline + $L_h$ (pretrained, proposed)	<b>40.597%</b>	<b>6.2</b>

- [1] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018.
- [2] M. Klingner, A. Bar, and T. Fingscheidt, "Improved noise and attack robustness for semantic segmentation by using multi-task training with self-supervised depth estimation," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020.
- [3] J. Novosel, P. Viswanath, and B. Arsenali, "Boosting semantic segmentation with multi-task self-supervised learning for autonomous driving applications," in Proceedings of Advances in Neural Information Processing Systems (NIPS), 2019.
- [4] P.Y. Chen, A. H Liu, Y.C. Liu, and Y.C.F Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

# Q & A

**Thank you very much for your attention!**

**Contact: Prof. I. Pitas  
pitas@csd.auth.gr**