Semantic Segmentation

Prepared for CSC2541: Visual Perception for Autonomous Driving
Stefania Raimondo
March 15, 2015
Semantic segmentation

- Pixel-level classification

(Badrinarayanan, Kendall, & Cipolla, 2015)

(Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)
Previously...

- Hand-engineered features, various classifiers

- Deep Convolutional Neural Nets
  - Success at other *high-level* vision tasks (abstract representations)

- DCNN hurdles for low-level tasks:
  - Signal down-sampling ---+ reduced signal resolution
  - Spatial invariance ---+ limits spatial accuracy

- In general – hard to train
#1 “DeepLab” (2014)

_Semantic image segmentation with deep convolutional nets and fully connected crfs_ – Chen et al., 2014

Idea

Overcome the two hurdles of DCNNs using the “atrous” algorithm (downsampling issue) and CRFs (spatial insensitivity)

Do not rely on front-end segmentation systems
DeepLab

• Deeper (more max-pooling)...
  • ... increased invariance and large receptive fields
  • ... loss of spatial accuracy

• Previous solutions:
  • Segmentation – 2 stage approaches
  • Harness information from multiple layers

• DeepLab Alternative: CRF
DeepLab DCNN

- Modify ImageNet pre-trained VGG-16 (Simonyan & Zisserman, 2014)
  - fully convolutional
  - dense features
  - Upsampling by bilinear interpolation

(image modified from Badrinarayanan, Kendall, & Cipolla, 2015)
DeepLab CRF

Image/G.T.  DCNN output  CRF Iteration 1  CRF Iteration 2  CRF Iteration 10

Score  Belief

(Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)
DeepLab CRF

\[
E(x) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j)
\]

From DCNN label probabilities

Gaussian, pairwise

\[
w_1 \exp \left( -\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2} \right) + w_2 \exp \left( -\frac{||p_i - p_j||^2}{2\sigma_\gamma^2} \right)
\]

Differences in position and intensity

Just position

(equations from Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)
DeepLab Variations

- DeepLab (no CRF)
- DeepLab-CRF
- DeepLab-MSc (CRF)
- DeepLab-7x7 (CRF)
- DeepLab-4x4 (CRF)
- DeepLab-LargeFOV (CRF/MSc)

Multi-scale Prediction

Input image

5*128 = 640 channels

Soft-max

(image modified from Badrinarayanan, Kendall, & Cipolla, 2015)
DeepLab Results

• PASCAL VOC 2012
  • 20 classes + background
  • ~1.5k images for testing/training/validation
  • ~10.5k extra training annotations
  • Performance: IOU averaged across classes

• Most results/experiments provided on ‘val’ set
<table>
<thead>
<tr>
<th>Method</th>
<th>mean IOU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepLab</td>
<td>59.80</td>
</tr>
<tr>
<td>DeepLab-CRF</td>
<td>63.74</td>
</tr>
<tr>
<td>DeepLab-MSc</td>
<td>61.30</td>
</tr>
<tr>
<td>DeepLab-MSc-CRF</td>
<td>65.21</td>
</tr>
<tr>
<td>DeepLab-7x7</td>
<td>64.38</td>
</tr>
<tr>
<td>DeepLab-CRF-7x7</td>
<td>67.64</td>
</tr>
<tr>
<td>DeepLab-LargeFOV</td>
<td>62.25</td>
</tr>
<tr>
<td>DeepLab-CRF-LargeFOV</td>
<td>67.64</td>
</tr>
<tr>
<td>DeepLab-MSc-LargeFOV</td>
<td>64.21</td>
</tr>
<tr>
<td>DeepLab-MSc-CRF-LargeFOV</td>
<td>68.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>mean IOU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSRA-CFM</td>
<td>61.8</td>
</tr>
<tr>
<td>FCN-8s</td>
<td>62.2</td>
</tr>
<tr>
<td>TTI-Zoomout-16</td>
<td>64.4</td>
</tr>
<tr>
<td>DeepLab-CRF</td>
<td>66.4</td>
</tr>
<tr>
<td>DeepLab-MSc-CRF</td>
<td>67.1</td>
</tr>
<tr>
<td>DeepLab-CRF-7x7</td>
<td>70.3</td>
</tr>
<tr>
<td>DeepLab-CRF-LargeFOV</td>
<td>70.3</td>
</tr>
<tr>
<td>DeepLab-MSc-CRF-LargeFOV</td>
<td>71.6</td>
</tr>
</tbody>
</table>

(Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)
DeepLab Results

(Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)
DeepLabResults

(Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)
(a) FCN-8s vs. DeepLab-CRF
(b) TTI-Zoomout-16 vs. DeepLab-CRF

(Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)
DeepLab Summary + Future Work...

• State-of-the-art on Pascal Segmentation
• Fast – 8fps DCNN, 0.5s CRF
• Step away from relying on segmentation
  • ...but still requires post-processing of NN output
  • ...but still not trained end-to-end

• Future work:
  • End-to-end training of CNN + CRF
  • Apply to video, depth maps...
  • Training with weakly supervised annotations
#2 SegNet (2015)

*SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation – Badrinarayanan et al. 2015*

Idea:
Performance boosting support algorithms should not be used to hide deficiencies in core network performance

Should train end-to-end
Focus on decoding architecture
SegNet Contributions

• Efficient architecture (memory + computation time)
  • Upsampling reusing max-pooling indices
• Reasonable results without performance boosting addition
• Comparison to FCN
SegNet Architecture

(Badrinarayanan, Kendall, & Cipolla, 2015)
SegNet Evaluation

• Pascal VOC 2012
  • Lots of background – favors methods using weakly labelled data
  • Same objects with different backgrounds
• CamVid – for variants
  • 11 classes, day and dusk, ~300 testing/training images
• SUN RGB-D
  • 37 indoor scene classes, ~5000 training/testing images
SegNet Evaluation

• Measures
  • Global accuracy (G)- % pixels correctly classified
  • Class average accuracy (C) – mean accuracy over all classes
  • Mean intersection over union (I/U)
    • Penalizes false positives, not optimized for
SegNet Decoders

(Badrinarayanan, Kendall, & Cipolla, 2015)
SegNet Decoder Evaluation

• SegNet-Basic
  • 4 enc. + 4 dec., all max-pooling, no RELU, 7x7 kernel

• SegNet-Basic-SingleChannelDecoder
  • decoder only convolve their corresponding upsampled feature map

• FCN Basic
  • fully convolutional decoding technique

• FCN Basic-NoAddition
  • skips the addition step (space)

• Bilinear-Interpolation – no learning for upsampling
## SegNet Decoder Evaluation

<table>
<thead>
<tr>
<th>Variant</th>
<th>Params (M)</th>
<th>Encoder storage (MB)</th>
<th>Infer time (ms)</th>
<th>Median frequency balancing</th>
<th>Natural frequency balancing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Test G C I/U</td>
<td>Train G C I/U</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed upsampling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilinear-Interpolation</td>
<td>0.625</td>
<td>0</td>
<td>24.2</td>
<td>77.9 61.1 43.3 89.1 90.2 82.7</td>
<td>82.7 52.5 43.8 93.5 74.1 59.9</td>
</tr>
<tr>
<td>Upsampling using max-pooling indices</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SegNet-Basic</td>
<td>1.425</td>
<td>1x</td>
<td>52.6</td>
<td>82.7 62.0 47.7 94.7 96.2 92.7</td>
<td>84.0 54.6 46.3 96.1 83.9 73.3</td>
</tr>
<tr>
<td>SegNet-Basic-EncoderAddition</td>
<td>1.425</td>
<td>64x</td>
<td>53.0</td>
<td>83.4 <strong>63.6</strong> 48.5 94.3 95.8 92.0</td>
<td><strong>84.2</strong> 56.5 <strong>47.7</strong> 95.3 80.9 68.9</td>
</tr>
<tr>
<td>SegNet-Basic-SingleChannelDecoder</td>
<td>0.625</td>
<td>1x</td>
<td>33.1</td>
<td>81.2 60.7 46.1 93.2 94.8 90.3</td>
<td>83.5 53.9 45.2 92.6 68.4 52.8</td>
</tr>
<tr>
<td>Learning to upsample (bilinear initialisation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FCN-Basic</td>
<td>0.65</td>
<td>11x</td>
<td>24.2</td>
<td>81.7 62.4 47.3 92.8 93.6 88.1</td>
<td>83.9 55.6 45.0 92.0 66.8 50.7</td>
</tr>
<tr>
<td>FCN-Basic-NoAddition</td>
<td>0.65</td>
<td>n/a</td>
<td>23.8</td>
<td>80.5 58.6 44.1 92.5 93.0 87.2</td>
<td>82.3 53.9 44.2 93.1 72.8 57.6</td>
</tr>
<tr>
<td>FCN-Basic-NoDimReduction</td>
<td>1.625</td>
<td>64x</td>
<td>44.8</td>
<td><strong>84.1</strong> 63.4 <strong>50.1</strong> <strong>95.1</strong> <strong>96.5</strong> <strong>93.2</strong></td>
<td>83.5 <strong>57.3</strong> 47.0 97.2 <strong>91.7</strong> <strong>84.8</strong></td>
</tr>
<tr>
<td>FCN-Basic-NoAddition-NoDimReduction</td>
<td>1.625</td>
<td>0</td>
<td>43.9</td>
<td>80.5 61.6 45.9 92.5 94.6 89.9</td>
<td>83.7 54.8 45.5 95.0 80.2 67.8</td>
</tr>
</tbody>
</table>

(Badrinarayanan, Kendall, & Cipolla, 2015)
SegNet Evaluation

• SUN RGB-D Results – complex indoor scenes
• CamVid - out-door road scenes
• Pascal VOC 2012 – few classes, varying backgrounds

• Demo
  • http://mi.eng.cam.ac.uk/projects/segnet/#demo
SegNet CamVid Results

SegNet-Basic with only local contrast normalized RGB as input (median freq. balancing)

SegNet with only local contrast normalized RGB as input (pre-trained encoder, median freq. balancing)

SegNet with only local contrast normalized RGB as input (pretrained encoder, median freq. balancing + large training set)

(Badrinarayanan, Kendall, & Cipolla, 2015)
## SegNet CamVid Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Building</th>
<th>Tree</th>
<th>Sky</th>
<th>Car</th>
<th>Sign-Symbol</th>
<th>Road</th>
<th>Pedestrian</th>
<th>Fence</th>
<th>Column-Pole</th>
<th>Side-walk</th>
<th>Bicyclist</th>
<th>Class avg</th>
<th>Global avg</th>
<th>Mean IU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SfM+Appearance [26]</td>
<td>46.2</td>
<td>61.9</td>
<td>89.7</td>
<td>68.6</td>
<td>42.9</td>
<td>89.5</td>
<td>53.6</td>
<td>46.6</td>
<td>0.7</td>
<td>60.5</td>
<td>22.5</td>
<td>53.0</td>
<td>69.1</td>
<td>n/a</td>
</tr>
<tr>
<td>Boosting [27]</td>
<td>61.9</td>
<td>67.3</td>
<td>91.1</td>
<td>71.1</td>
<td>58.5</td>
<td>92.9</td>
<td>49.5</td>
<td>37.6</td>
<td>25.8</td>
<td>77.8</td>
<td>24.7</td>
<td>59.8</td>
<td>76.4</td>
<td>n/a</td>
</tr>
<tr>
<td>Dense Depth Maps [30]</td>
<td>85.3</td>
<td>57.3</td>
<td>95.4</td>
<td>69.2</td>
<td>46.5</td>
<td><strong>98.5</strong></td>
<td>23.8</td>
<td>44.3</td>
<td>22.0</td>
<td>38.1</td>
<td>28.7</td>
<td>55.4</td>
<td>82.1</td>
<td>n/a</td>
</tr>
<tr>
<td>Structured Random Forests [29]</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>51.4</td>
</tr>
<tr>
<td>Neural Decision Forests [60]</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>56.1</td>
</tr>
<tr>
<td>Local Label Descriptors [61]</td>
<td>80.7</td>
<td>61.5</td>
<td>88.8</td>
<td>16.4</td>
<td>n/a</td>
<td>98.0</td>
<td>1.09</td>
<td>0.05</td>
<td>4.13</td>
<td>12.4</td>
<td>0.07</td>
<td>36.3</td>
<td>73.6</td>
<td>n/a</td>
</tr>
<tr>
<td>Super Parsing [31]</td>
<td>87.0</td>
<td>67.1</td>
<td>96.9</td>
<td>62.7</td>
<td>30.1</td>
<td>95.9</td>
<td>14.7</td>
<td>17.9</td>
<td>1.7</td>
<td>70.0</td>
<td>19.4</td>
<td>51.2</td>
<td>83.3</td>
<td>n/a</td>
</tr>
<tr>
<td>SegNet-Basic</td>
<td>81.3</td>
<td>72.0</td>
<td>93.0</td>
<td>81.3</td>
<td>14.8</td>
<td>93.3</td>
<td>62.4</td>
<td>31.5</td>
<td>36.3</td>
<td>73.7</td>
<td>42.6</td>
<td>62.0</td>
<td>82.7</td>
<td>47.7</td>
</tr>
<tr>
<td>SegNet-Basic (layer-wise training [12])</td>
<td>75.0</td>
<td>84.6</td>
<td>91.2</td>
<td>82.7</td>
<td>36.9</td>
<td>95.3</td>
<td>55.0</td>
<td>37.5</td>
<td>44.8</td>
<td>74.1</td>
<td>16.0</td>
<td>62.9</td>
<td>84.3</td>
<td>n/a</td>
</tr>
<tr>
<td>SegNet</td>
<td><strong>88.8</strong></td>
<td>87.3</td>
<td>92.4</td>
<td>82.1</td>
<td>20.5</td>
<td>97.2</td>
<td>57.1</td>
<td>49.3</td>
<td>27.5</td>
<td>84.4</td>
<td>30.7</td>
<td>65.2</td>
<td><strong>88.5</strong></td>
<td>55.6</td>
</tr>
<tr>
<td>SegNet (3.5K dataset training)</td>
<td>73.9</td>
<td><strong>90.6</strong></td>
<td>90.1</td>
<td><strong>86.4</strong></td>
<td><strong>69.8</strong></td>
<td>94.5</td>
<td><strong>86.8</strong></td>
<td><strong>67.9</strong></td>
<td><strong>74.0</strong></td>
<td><strong>94.7</strong></td>
<td><strong>52.9</strong></td>
<td><strong>80.1</strong></td>
<td><strong>86.7</strong></td>
<td><strong>60.4</strong></td>
</tr>
</tbody>
</table>

(Badrinarayanan, Kendall, & Cipolla, 2015)
SegNet SUN RGB-D Results

(Badrinarayanan, Kendall, & Cipolla, 2015)
## SegNet Pascal Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Encoder size (M)</th>
<th>Decoder size (M)</th>
<th>Total size (M)</th>
<th>Class avg. acc.</th>
<th>Inference 500 × 500 pixels</th>
<th>Inference 224 × 224 pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepLab [14] (validation set)</td>
<td>n/a</td>
<td>n/a</td>
<td>&lt; 134.5</td>
<td>58</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>FCN-8 [2] (multi-stage training)</td>
<td>134</td>
<td>0.5</td>
<td>134.5</td>
<td>62.2</td>
<td>210ms</td>
<td>n/a</td>
</tr>
<tr>
<td>Hypercolumns [43] (object proposals)</td>
<td>n/a</td>
<td>n/a</td>
<td>&gt; 134.5</td>
<td>62.6</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>DeconvNet [9] (object proposals)</td>
<td>138.35</td>
<td>138.35</td>
<td>276.7</td>
<td>69.6</td>
<td>n/a</td>
<td>92ms (× 50)</td>
</tr>
<tr>
<td>CRF-RNN [10] (multi-stage training)</td>
<td>n/a</td>
<td>n/a</td>
<td>&gt; 134.5</td>
<td>69.6</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>SegNet</td>
<td><strong>14.725</strong></td>
<td><strong>14.725</strong></td>
<td><strong>29.45</strong></td>
<td>59.1</td>
<td><strong>94ms</strong></td>
<td><strong>28ms</strong></td>
</tr>
</tbody>
</table>

(Badrinarayanan, Kendall, & Cipolla, 2015)
SegNet Summary + Future Work

• Reasonable results w/out support methods
• Comparison of FCN decoding method
• Some failures:
  • Lacks smoothness on large objects
  • Cannot handle clutter

• Future work:
  • Estimate labelling uncertainty
  • Real-time application
  • Dropout during training and testing
#3 Joint Seg + 3D Reconstruction

*Joint Semantic Segmentation and 3D Reconstruction from Monocular Video – Kundu et al. 2014*

**Idea**

Structural and semantic information is necessary for some applications and can benefit from each other
Method

2D image segmentation

3D labelled voxel representation

SLAM: trajectory + point cloud

CRF

(images from Kundu, Li, Dellaert, Li, & Rehg, 2014)
Contributions

• Method for simultaneous 3D structure and semantics
  • ...but not the first to use 3D features to improve 2D segmentation
  • ...and not the first to use 2D segmentation to improve 3D depth estimation

• Benefits of this approach
  • Temporally consistent
  • Monocular
  • Does not require dense depth maps
  • Efficient for real-time applications

• First 3D reconstruction of monocular Camvid
CRF Model

\[ m_i \in \mathcal{M} \leftarrow \text{voxel } i\text{'s semantic label} \quad \mathcal{L}_\mathcal{M} = \{\text{Free, Road, Car, ...}\} \]

\[ \mathcal{D} = \{z_{1:P}^r, z_{1:Q}^s, g_{1:T}\} \leftarrow \text{input data/measurements} \]

- with-depth
  - pixel + pose
  - 2D label
  - depth
- semantic-only
  - pixel + pose
  - 2D label
- camera trajectory per image
CRF Model

\[ \mathcal{M}^* = \arg \max_{\mathcal{M}} P(\mathcal{M} | \mathcal{D}) \]  

\[ P(\mathcal{M} | \mathcal{D}) = P(\mathcal{M}) \prod_{p=1}^{P} P(z_p^w | \mathbf{m}_p, g_p) \prod_{q=1}^{Q} P(z_q^s | \mathbf{m}_q, g_q) \]

Map is independent of camera trajectory

Do not assume voxels are independent

Measurements are independent given the map

\[ P(\mathcal{M} | \mathcal{D}) = \frac{1}{Z(\mathcal{D})} \prod_{i} \psi^i_u(m_i) \prod_{i,j \in \mathcal{N}} \psi_p(m_i, m_j) \prod_{R \in \mathcal{R}} \psi_h(\mathbf{m}_R) \]
Unary potentials

$$\psi_u^i(m_i) = [\psi_{\text{MISS}}(m_i)]^{N_M} \prod_{l \in \mathcal{L} \setminus \text{Sky}} [\psi_{\text{HIT}}^l(m_i)]^{N_{H_l}}$$

(Kundu, Li, Dellaert, Li, & Rehg, 2014)
Updating Potentials

1) With-depth measurement
2) Semantic-only measurements

$$\psi_u^i(m_i) = \left[\psi_{\text{MISS}}(m_i)\right]^{N_M} \prod_{l \in \mathcal{L}_T \setminus \text{Sky}} \left[\psi_{\text{HIT}}^l(m_i)\right]^{N_{HI}}$$

(image modified from Kundu, Li, Dellaert, Li, & Rehg, 2014)
Updating Potentials (Semantic-only)

• Depth statistics (per grid cell)
• For low-depth-uncertainty categories:
  • Same as with-depth, add unary factors
• For high-depth-uncertainty categories:
  • Add a higher order factor
  • Joins voxels along the ray bwtwn min-max depth

(image modified from Kundu, Li, Dellaert, Li, & Rehg, 2014)
Updating Potentials

• Pairwise potentials
  • Neighbors in each direction are treated differently
    • Ex. road more likely in the horizontal direction
  • Lower cost for free neighbor
Implementation details…

• Octree data structure
  • Unused voxels are uninitialized
  • Minimal storage/computation
  • Pairwise/higher order are static across all voxels
  • Only store factor values, not the measurements

• Clamping
  • High probability (0.98) voxels are treated like evidence
  • 3D support for clamped voxels
    • Extra hit unaries for neighbors
    • Including free-space boundaries

• Improving SLAM
  • Reject matches if they lie on different semantic categories
  • Bundle adjustment (minimize re-projection errors)
Results

• Camvid, Leuven, KITTI
  • Fast forward-moving datasets

• Video

• Qualitative 3D reconstruction results

• Quantitative 2D segmentation results
  • Label accuracy
  • Temporal consistency (entropy)
Results

- 2D segmentation results...

<table>
<thead>
<tr>
<th>CAMVID seq05VD</th>
<th>Building</th>
<th>Road</th>
<th>Car</th>
<th>Sidewalk</th>
<th>Sky</th>
<th>Tree</th>
<th>Fence</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>NA</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>[20]</td>
<td>0.114</td>
<td>98.52</td>
<td>0.04</td>
<td>95.99</td>
<td>0.231</td>
<td>89.41</td>
<td>0.177</td>
<td>96.53</td>
</tr>
<tr>
<td>[24]</td>
<td>0.114</td>
<td>94.78</td>
<td>0.016</td>
<td>98.85</td>
<td>0.106</td>
<td>99.69</td>
<td>0.184</td>
<td>94.11</td>
</tr>
<tr>
<td>[31]</td>
<td>0.025</td>
<td>95.01</td>
<td>0.004</td>
<td>98.97</td>
<td>0.046</td>
<td>99.87</td>
<td>0.062</td>
<td>73.17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LEUVEN</th>
<th>Building</th>
<th>Road</th>
<th>Car</th>
<th>Sidewalk</th>
<th>Sky</th>
<th>Bike</th>
<th>Pedestrian</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>NA</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>[19]</td>
<td>0.046</td>
<td>95.84</td>
<td>0.116</td>
<td>98.75</td>
<td>0.150</td>
<td>91.42</td>
<td>0.429</td>
<td>74.89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>KITTI seq05</th>
<th>Building</th>
<th>Road</th>
<th>Car</th>
<th>Sidewalk</th>
<th>Sky</th>
<th>Fence</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>NA</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>[20]</td>
<td>0.165</td>
<td>97.47</td>
<td>0.113</td>
<td>87.85</td>
<td>0.203</td>
<td>98.14</td>
<td>0.158</td>
</tr>
</tbody>
</table>

(Kundu, Li, Dellaert, Li, & Rehg, 2014)
3D+Seg Summary

• CRF based method
• Dense reconstruction
• Temporally consistent
• “Tractable for large outdoor environments”

• Future work:
  • Real-time application
  • Incorporating multi-camera information (already done)
Overall Summary

• Semantic Segmentation – pixel-level
  • Need dense output
  • Need to preserve spatial details

• CRF (and other methods) can boost accuracy of NN models
• DCNN models may be able to stand on their own
  • If given sufficient training data and proper architecture
• Use 3D information to augment 2D segmentation
