1. Introduction

- Task & General Approaches:
  - Semantic Segmentation
  - Intra-object Variations
  - Inter-object Contexts

- Motivation:
  - Combine ConvNets and MRF into a unified framework:
    - End-to-end Training
    - Rich Pairwise Relationship

- Existing Works:
  - DenseCRF [CVPR 2011]
  - FCN [CVPR 2015]
  - DeepLab [CVPR 2015]
  - CRFasRNN [ICCV 2015]
  - DPN

- Our Idea:
  - High-order MRF as One-pass CNN:
    \[ E(y) = \sum_{i \in V} \Phi(y_i) + \sum_{i,j \in E} \Psi(y_i, y_j) \]

2. Approach

- Unary Term
  \[ \Phi(y_i) = -\ln p_i \]
  \( p_i \) indicates the probability of the presence of label \( u \) at pixel \( i \)

- Pairwise Term
  \[ \Psi(y_i, y_j) = \sum_{k=1}^{K} \lambda_k \mu_k(i, u, j, v) \sum_{z \in E} d(j, z) p_z \]

- Mean Field Solver
  \[ q_i^u \propto \exp \left( -\Phi_i^u - \sum_{k=1}^{K} \lambda_k \mu_k(i, u, j, v) \sum_{z \in E} d(j, z) q_z^v \right) \]
  (each \( q_i^u \) is initialized by the corresponding \( p_i^u \))

3. Network Architecture

Deep Parsing Network (DPN) : \( 512 \times 512 \times 3 \) input image; \( 512 \times 512 \times 21 \) output label maps

- Convolution
- Max Pooling
- Deconvolution
- Local Convolution


4. Effectiveness of DPN

- Label-Label Space
- Spatial Label Space
- Pairwise Terms Comparisons
- End-to-End Learning

5. Overall Performance

<table>
<thead>
<tr>
<th>Approach</th>
<th>area</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td>76.8</td>
<td>34.2</td>
<td>68.9</td>
<td>49.4</td>
<td>60.3</td>
<td>75.3</td>
<td>74.7</td>
<td>77.6</td>
<td>74.4</td>
<td>62.5</td>
<td>46.8</td>
<td>71.8</td>
<td>63.9</td>
<td>76.5</td>
<td>73.9</td>
<td>45.2</td>
<td>72.4</td>
<td>87.4</td>
<td>70.9</td>
<td>55.1</td>
</tr>
<tr>
<td>DeepLab</td>
<td>89.2</td>
<td>46.7</td>
<td>88.5</td>
<td>63.5</td>
<td>68.4</td>
<td>87</td>
<td>81.2</td>
<td>86.3</td>
<td>32.6</td>
<td>80.7</td>
<td>62.4</td>
<td>81</td>
<td>81.3</td>
<td>84.3</td>
<td>82.1</td>
<td>56.2</td>
<td>84.6</td>
<td>58.3</td>
<td>76.2</td>
<td>67.2</td>
</tr>
<tr>
<td>RNN</td>
<td>90.4</td>
<td>55.3</td>
<td>88.7</td>
<td>68.4</td>
<td>69.8</td>
<td>88.3</td>
<td>82.4</td>
<td>85.1</td>
<td>32.6</td>
<td>78.5</td>
<td>64.4</td>
<td>79.6</td>
<td>81.9</td>
<td>86.4</td>
<td>81.8</td>
<td>58.6</td>
<td>82.4</td>
<td>53.5</td>
<td>77.4</td>
<td>70.1</td>
</tr>
<tr>
<td>BoxSupv</td>
<td>89.8</td>
<td>38</td>
<td>89.2</td>
<td>68.9</td>
<td>68</td>
<td>89.6</td>
<td>81</td>
<td>87.7</td>
<td>34.4</td>
<td>83.6</td>
<td>67.1</td>
<td>81.5</td>
<td>83.7</td>
<td>85.2</td>
<td>83.5</td>
<td>58.6</td>
<td>84.9</td>
<td>55.8</td>
<td>81.2</td>
<td>70.7</td>
</tr>
<tr>
<td>DPN</td>
<td>87.7</td>
<td>59.4</td>
<td>78</td>
<td>64.9</td>
<td>70.3</td>
<td>89.3</td>
<td>83.5</td>
<td>86.1</td>
<td>31.7</td>
<td>79.9</td>
<td>62.6</td>
<td>81.9</td>
<td>80</td>
<td>83.5</td>
<td>82.3</td>
<td>60.5</td>
<td>83.2</td>
<td>53.4</td>
<td>77.9</td>
<td>65</td>
</tr>
<tr>
<td>DPN*</td>
<td>89</td>
<td>61.6</td>
<td>87.7</td>
<td>66.8</td>
<td>74.7</td>
<td>91.2</td>
<td>84.3</td>
<td>87.6</td>
<td>36.5</td>
<td>86.3</td>
<td>66.1</td>
<td>84.4</td>
<td>87.8</td>
<td>85.6</td>
<td>85.4</td>
<td>63.6</td>
<td>87.3</td>
<td>61.3</td>
<td>79.4</td>
<td>66.4</td>
</tr>
</tbody>
</table>

Per-class results on VOC12 test. The approaches pre-trained on COCO are marked with *.

6.2 Visual Quality Comparisons

Visual quality comparison of different semantic image segmentation methods:
- (a) input image
- (b) ground truth
- (c) FCN
- (d) DeepLab
- (e) DPN

7. Conclusion

- DPN employs one-pass CNN to model high-order MRF
- High performance by approximating one iteration of MF
- DPN incorporates various types of pairwise terms
- Rich contextual information
- DPN contains only conventional operations of CNN
- Easier to be parallelized and speeded up in GPU