Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing

Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, Jingdong Wang

www.xinggangw.info

Huazhong University of Science and Technology
Weakly-supervised visual learning (WSVL)

- Weakly-supervised visual learning is a new trend in CVPR

Search keyword “weakly supervised” and “weakly-supervised” in CVPR 17&18

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Weakly supervised</th>
<th>Weakly-supervised</th>
<th>In total</th>
</tr>
</thead>
<tbody>
<tr>
<td>cvpr17</td>
<td>14</td>
<td>5</td>
<td>19/783</td>
</tr>
<tr>
<td>cvpr18</td>
<td>19</td>
<td>10</td>
<td>29/979</td>
</tr>
</tbody>
</table>

Huazhong University of Science and Technology
The task of WSSS

WSSS overcomes the deficiency problem in semantic segmentation labelling.

Huazhong University of Science and Technology
The development of WSSS

\[(x_l, y_l) = \arg \max_{\forall l \in \mathcal{L}_I} \hat{p}_l(x, y)\]

MIL LOSS = \(-\frac{1}{|\mathcal{L}_I|} \sum_{l \in \mathcal{L}_I} \log \hat{p}_l(x_l, y_l)\)

MIL-FCN, Pathak et al, Arxiv 14, ICLRW 15

CAM, Zhou et al, CVPR 16

STC, Wei et al, TPAMI 15

Proposal classification, Qi et al, ECCV 16

Built-in FG/BG Model

Saleh et al, ECCV 16

Adversarial erasing, Wei et al, CVPR 17

Figures are from the original papers
The development of WSSS

1. Multi-instance learning
2. Saliency guided
3. Built-in network information
4. Adversarial learning
5. Seeding loss

Seeding loss, Kolesnikov et al, ECCV 17
Saliency guided labler, Oh et al, CVPR 17

Figures are from the original papers
The basic framework in our paper

Step 1: Foreground seeds from CAM

Step 2: Background seeds derived salient region detection [Jiang et al, CVPR13]

Figures are from the original papers
The basic framework in our paper

Step 3: FCN with seeding loss

Step 4: Retrain with FCN

Huazhong University of Science and Technology
A small trick: balanced seeding loss

Balance the weights between foreground and background

\[
\ell_{seed} = -\frac{1}{\sum_{c \in \mathcal{C}} |\mathcal{S}_c|} \sum_{c \in \mathcal{C}} \sum_{u \in \mathcal{S}_c} \log H_{u,c} \\
- \frac{1}{\sum_{c' \in \overline{\mathcal{C}}} |\mathcal{S}_{c'}|} \sum_{c' \in \overline{\mathcal{C}}} \sum_{u \in \mathcal{S}_{c'}} \log H_{u,c'}
\]
However, the seeds are sparse

In practice, to retain the precision of seeds, there are about 40% pixels have labels.
How to improve the quality and quantity of seeds

- Better “CAM” network
- Saliency guidance
- Adversarial erasing
- ...

- Online seeded region growing
Deep seeded region growing

Region growing criteria:

\[ P(H_{u,c}, \theta_c) = \begin{cases} 
\text{TRUE} & H_{u,c} \geq \theta_c \text{ and } c = \arg \max_{c'} H_{u,c'}, \\
\text{FALSE} & \text{otherwise.}
\end{cases} \]

1. **Directly use deep prob features**
2. **Cheap to compute**
3. **Online supervision updating**

Progressively check the neighborhood pixels

Huazhong University of Science and Technology
Deep seeded region growing
Algorithm 2 Deep Seeded Region Growing Training

1: **Input:** Training data \( D = \{(I_i, S_i)\}_{i=1}^N \).
2: **Initialize:** initialize \( M_0 \), \( t = 1 \).
3: **while** \( t \leq \text{max\_iter} \) **do**
4: Select a sample \( \{I_i, S_i\} \) from input data randomly;
5: \( H_i = M_{t-1}(I_i) \);
6: Perform \( G_i = DSRG(S_i, H_i) \) for seed expansion
7: Compute the loss \( \ell(G_i, H_i) \)
8: back propagate the error and update model from \( M_{t-1} \) to \( M_t \)
9: **end while**
10: **Output:** \( M \)
Experiments

- **Datasets**
  - PASCAL VOC 2012, 10582 train, 1449 val, 1456 test
  - COCO, 80k train, 40k val

- mIoU criterion

- Classification network: VGG-16
- Segmentation network: DeepLab-ASPP
## Main Results

### PASCAL VOC

<table>
<thead>
<tr>
<th>Method</th>
<th>Training</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCSM[2]</td>
<td>10k</td>
<td>44.1</td>
<td>45.1</td>
</tr>
<tr>
<td>BFBP[3]</td>
<td>10k</td>
<td>46.6</td>
<td>48.0</td>
</tr>
<tr>
<td>STC [4]</td>
<td>50k</td>
<td>49.8</td>
<td>51.2</td>
</tr>
<tr>
<td>SEC [5]</td>
<td>10k</td>
<td>50.7</td>
<td>51.7</td>
</tr>
<tr>
<td>AF-SS [6]</td>
<td>10k</td>
<td>52.6</td>
<td>52.7</td>
</tr>
<tr>
<td>Combining Cues [7]</td>
<td>10k</td>
<td>52.8</td>
<td>53.7</td>
</tr>
<tr>
<td>AE-PSL [8]</td>
<td>10k</td>
<td>55.0</td>
<td>55.7</td>
</tr>
<tr>
<td>DCSP [9]</td>
<td>10k</td>
<td>58.6</td>
<td>59.2</td>
</tr>
<tr>
<td>DSRG (VGG16)</td>
<td>10k</td>
<td>59.0</td>
<td>60.4</td>
</tr>
<tr>
<td>DSRG (Resnet101)</td>
<td>10k</td>
<td>61.4</td>
<td>63.2</td>
</tr>
</tbody>
</table>

### COCO

<table>
<thead>
<tr>
<th>Method</th>
<th>Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFBP[3]</td>
<td>20.4</td>
</tr>
<tr>
<td>SEC [5]</td>
<td>22.4</td>
</tr>
<tr>
<td>DSRG (Ours)</td>
<td>26.0</td>
</tr>
</tbody>
</table>
Ablation studies

The contributions of **Balanced seeding loss, DSRG & Retrain**

Table 2. Comparison of mIoU using different settings of our approach on VOC 2012 val set

<table>
<thead>
<tr>
<th>Method</th>
<th>bkg</th>
<th>plane</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>motor</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>LV</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>82.5</td>
<td>67.5</td>
<td>23.2</td>
<td>65.7</td>
<td>29.7</td>
<td>47.5</td>
<td>71.8</td>
<td>66.8</td>
<td>76.7</td>
<td>23.3</td>
<td>51.7</td>
<td>26.2</td>
<td>69.7</td>
<td>54.2</td>
<td>63.2</td>
<td>57.2</td>
<td>33.7</td>
<td>64.5</td>
<td>33.5</td>
<td>48.7</td>
<td>46.1</td>
<td>52.5</td>
</tr>
<tr>
<td>+BSL</td>
<td>82.4</td>
<td>71.9</td>
<td>29.1</td>
<td>67.7</td>
<td>32.4</td>
<td>49.8</td>
<td>75.5</td>
<td>67.9</td>
<td>74.7</td>
<td>22.8</td>
<td>54.9</td>
<td>26.6</td>
<td>64.3</td>
<td>55.7</td>
<td>64.7</td>
<td>56.0</td>
<td>35.0</td>
<td>67.7</td>
<td>32.7</td>
<td>50.2</td>
<td>45.8</td>
<td>53.6</td>
</tr>
<tr>
<td>+DSRG</td>
<td>86.6</td>
<td>70.5</td>
<td>28.8</td>
<td>70.6</td>
<td>34.7</td>
<td>55.7</td>
<td>74.9</td>
<td>70.1</td>
<td>80.2</td>
<td>24.1</td>
<td>63.6</td>
<td>24.8</td>
<td>76.6</td>
<td>64.1</td>
<td>64.9</td>
<td>72.3</td>
<td>38.5</td>
<td>68.7</td>
<td>35.8</td>
<td>51.8</td>
<td>51.9</td>
<td>57.6</td>
</tr>
<tr>
<td>+Retrain</td>
<td>87.5</td>
<td>73.1</td>
<td>28.4</td>
<td>75.4</td>
<td>39.5</td>
<td>54.5</td>
<td>78.2</td>
<td>71.3</td>
<td>80.6</td>
<td>25.0</td>
<td>63.3</td>
<td>25.4</td>
<td>77.8</td>
<td>65.4</td>
<td>65.2</td>
<td>72.8</td>
<td>41.2</td>
<td>74.3</td>
<td>34.1</td>
<td>52.1</td>
<td>53.0</td>
<td>59.0</td>
</tr>
</tbody>
</table>
Ablation studies

Image | w/o DSRG | +DSRG | Ground Truth

Huazhong University of Science and Technology
Ablation studies

The quality of the dynamic supervision (%) with respect to the epochs.

Performance on PASCAL val dataset for different $\theta$

Huazhong University of Science and Technology
Video demo
How to interpret DSRG

- A Neural network generates new label by itself.
- The inner structure of image/video helps, e.g., [Ahn & Kwak, CVPR 18].

From the perspective of SSL, pseudo label/supervision [Lee, ICMLw 13, Wang et al, MM 16] works.
Discussion

- Current limitations of WSSS
  - Hard to obtain precise boundaries
  - Does not work well in complex dataset, e.g., COCO & Kitti

- Let deep networks know what is an object, e.g., unsupervised learning from video.
- Weakly and semi-supervised (WASS) visual learning.
The paper is available at
http://www.xinggangw.info/pubs/cvpr18-dsrg.pdf

Codes will be available at
https://github.com/speedinghzl/DSRG
Thanks for your attention!