Webly Supervised Semantic Segmentation

Bin Jin¹, Maria V. Ortiz-Segovia², Sabine Süsstrunk³
¹IC, EPFL ²Océ Print Logic Technologies

**Motivation:**
The pixel-wise annotation of images to obtain accurate semantic segmentation ground-truth is both expensive and time-consuming. By exploiting the vast collection of labeled web images with rich context, we can bypass this tedious task. Our webly supervised segmentation outperforms the state-of-the-art weakly supervised segmentation methods by a significant margin.

**Web images:**
From these websites, we collect three sets of web images as training data:

- W: a white background set, built by querying the text-based image search engine, e.g., Google or Microsoft Bing, with the query "class" on white background.
- C: a common background set, built by retrieving images from image sharing websites, e.g., Flickr or Imgur1, with common background keywords.
- R: a realistic images set, constructed by crawling image sharing websites with the given class name or using existing datasets.

**References:**

**Methods:**
Images in (W) are first segmented with a saliency algorithm[1] combined with dense CRF[2]. We then train a semantic segmentation network in a three-stage pipeline:

(a) Stage 1: we train a Shallow Neural Network (SNN) for each class to output class-specific segmentation masks, using the hypercolumn features from a pre-trained network.

(b) Stage 2: we iteratively refine the SNNs based on the realistic images in {R}. A Conditional Random Field (CRF) is applied during each iteration.

(c) Stage 3: we assemble all SNNs into one deep convolution neural network (DCNN) by training the DCNN end-to-end with the segmentation masks generated by the SNNs.

**Results:**
Dataset: For training, we collect 6807 white background images for (W), 1491 images for (C) and 10,582 images for (R). We evaluate the performance on the PASCAL VOC 2012 segmentation benchmark with the Intersection over Union (IoU) metric.

<table>
<thead>
<tr>
<th>Method</th>
<th>CCCN³</th>
<th>DCSM⁴</th>
<th>BFBP⁵</th>
<th>STC⁶</th>
<th>SEC⁷</th>
<th>OURS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation set</td>
<td>35.3</td>
<td>44.1</td>
<td>46.6</td>
<td>49.8</td>
<td>50.7</td>
<td>53.4</td>
</tr>
<tr>
<td>Test set</td>
<td>35.6</td>
<td>45.1</td>
<td>48.0</td>
<td>51.2</td>
<td>51.7</td>
<td>55.3</td>
</tr>
</tbody>
</table>

IoU: Our method produces excellent results on both the validation and test set, outperforming all state-of-the-art weakly-supervised semantic segmentation methods.