PiCANet: Learning Pixel-wise Contextual Attention for Saliency Detection
Supplemental Material

Nian Liu¹  Junwei Han¹  Ming-Hsuan Yang²,³
¹Northwestern Polytechnical University  ²University of California, Merced  ³Google Cloud
{liunian228, junweihan2010}@gmail.com  mhyang@ucmerced.edu

In this supplementary material, we include more implementation details, more ablation analyses, and more experimental results.

1. Additional Implementation Details

1.1. Using the ResNet50 Backbone

When using the ResNet50 network [1] as the backbone, we modify the 4th and the 5th residual blocks to have strides of 1 and dilations of 2 and 4, respectively, thus making the encoder have a stride of 8. Then we progressively fuse the feature maps from the 5th to the 1st Conv blocks in the decoding modules D7 to D1. We adopt global PiCANets in D7 and D5, and local PiCANets in the last three modules, respectively. In each decoding module Di, we use the final Conv feature map of the ith Conv block in the ResNet50 encoder (e.g. res4f and res3d) as the incorporated encoder feature map Eni and do not adopt the BN and the ReLU layers on it as shown in Figure 3(b) since the ResNet50 network has already used BN layers after each Conv layer. The final generated saliency map is of size 112 × 112 since the conv1 layer has a stride of 2.

As the same as when using the VGG-16 backbone, we empirically set the loss weights in D5, D4, · · · , D1 as 0.5, 0.5, 0.8, 0.8, and 1, respectively. The minibatch size of our ResNet50 based network is set to 8 due to the GPU memory limitation. The other hyperparameters are set as the same as the ones used in the VGG-16 based network. The testing time for one image is 0.236s.

1.2. Using the CRF Post-processing

When we adopt the CRF post-processing method, we use the same parameters and the same code used by [2]. It additionally costs another 0.09s for each image.

Table 1. Effectiveness of progressively embedding PiCANets. “+7GP” means using Global PiCANets in D7 and D5, and Local PiCANets in D4, D3, D2. Other settings can be inferred similarly. Blue indicates the best performance.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fβ</td>
<td>Fω</td>
</tr>
<tr>
<td>U-Net [7]</td>
<td>0.761</td>
<td>0.651</td>
</tr>
<tr>
<td>+7GP</td>
<td>0.772</td>
<td>0.660</td>
</tr>
<tr>
<td>+75GP</td>
<td>0.778</td>
<td>0.662</td>
</tr>
<tr>
<td>+7G4LP</td>
<td>0.785</td>
<td>0.678</td>
</tr>
<tr>
<td>+7G43LP</td>
<td>0.791</td>
<td>0.682</td>
</tr>
<tr>
<td>+7G432LP</td>
<td>0.794</td>
<td>0.691</td>
</tr>
</tbody>
</table>

2. Experiments

2.1. Effectiveness of Progressively Embedding PiCANets

Here we report a more detailed ablation study of progressively embedding PiCANets in each decoding module. As shown in Table 1, progressively embedding global and local PiCANets in D7, D5, D4, · · · , D1 can consistently improve the saliency detection performance, thus demonstrating the effectiveness of our proposed PiCANets and the saliency detection model.

2.2. More Visualization of the Learned Attention Maps

We illustrate more learned attention maps in Figure 1 for the five attended decoding modules. Figure 1 shows that the global attention learned in D7 and D5 can attend to foreground objects for background pixels and backgrounds for foreground pixels. The local attention learned in D4, D3, and D2 can attend to regions with similar semantics with the referred pixel.

2.3. More Visual Comparison Between Our Model and Stata-of-the-Art Methods

We also show more qualitative results in Figure 2. It shows that compared with other state-of-the-art methods, our model can highlight salient objects more accurately and
uniformly under various challenging scenarios even without using post-processing techniques.

2.4. Failure Cases

We show some failure cases of our PiCANet-R model in Figure 3. Basically, our model usually fails when the image has no obvious foreground objects, as shown in (a) and
(b). (c) shows that when the foreground object is extremely large, our model is also easy to fail. While these two situations are also challenging to other traditional and deep learning based saliency models, indicating that we still have much room to improve current models. (d) shows that the non-uniform illumination on the object may also mislead our model.

References


