R³Net: Recurrent Residual Refinement Network for Saliency Detection

Zijun Deng¹,*, Xiaowei Hu²,*, Lei Zhu³,², Xuemiao Xu¹, Jing Qin³, Guoqiang Han¹, and Pheng-Ann Heng²,⁴

¹South China University of Technology
²The Chinese University of Hong Kong
³The Hong Kong Polytechnic University
⁴Shenzhen Institutes of Advanced Technology
Saliency Detection

- Goal: highlight the most *visually distinctive* objects in an image

- Applications: weakly-supervised object detection, visual tracking, etc.
Saliency Detection: A Two-stage View

1. Detect the salient objects
   • Global perception of saliency
   • (where the salient objects are)

2. Segment the accurate regions of salient objects
   • Precise object localization

Recent Work: FCNs

- Extract discriminative saliency features while keeping spatial information
  - Process the two stages simultaneously

- Deep high-level features are better for detection than hand-crafted features

- High-level features are unfriendly to segmentation due to its low resolution

What we want indeed!
Recent Work: FCNs

- [Hou et al.] exploited complementary information of multi-level features
  - Conduct prediction at one stage, making results still unsatisfactory

- [Wang et al.] presented a stage-wise refinement network
  - Low-level features tend to introduce non-salient regions
  - Do not preserve the previous saliency maps in multi-stage refinement

DSS [Hou et al., 2017]  
SRM [Wang et al., 2017]
Our Motivation

- \textit{Alternatively} leverage the low-level detailed features and the high-level semantic features to do refinement

- Apply \textit{residual learning} to saliency map refinement
Our Model

- Residual Refinement Block for multiple-stage refinement
- Alternatively leverage low-level features and high-level features
- Deep supervision for initial prediction and each refinement stage
Residual Refinement

- Ease the optimization task with a faster convergence at early stages
- Reduce the training error over directly learning the underlying saliency mapping
Training
- On the MSRA10K dataset (10K images)
- ResNeXt101-32x4d as feature extraction network, pre-trained on ImageNet
- Takes only 80 minutes on a single GPU

Testing
- On five benchmark datasets: ECSSD (1K images), HKU-IS (~4K images), PASCAL-S (0.8K images), SOD (0.3K images), DUT-OMRON (~6K images)
- Apply CRF (fully connected conditional random field) to enhance the saliency maps
## Comparison with State-of-the-arts

<table>
<thead>
<tr>
<th>Method</th>
<th>ECSSD $F_\beta$</th>
<th>ECSSD MAE</th>
<th>HKU-IS $F_\beta$</th>
<th>HKU-IS MAE</th>
<th>PASCAL-S $F_\beta$</th>
<th>PASCAL-S MAE</th>
<th>SOD $F_\beta$</th>
<th>SOD MAE</th>
<th>DUT-OMRON $F_\beta$</th>
<th>DUT-OMRON MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR [Yang et al., 2013]</td>
<td>0.736</td>
<td>0.189</td>
<td>0.715</td>
<td>0.174</td>
<td>0.666</td>
<td>0.223</td>
<td>0.619</td>
<td>0.273</td>
<td>0.610</td>
<td>0.187</td>
</tr>
<tr>
<td>wCtr* [Zhu et al., 2014]</td>
<td>0.716</td>
<td>0.171</td>
<td>0.726</td>
<td>0.141</td>
<td>0.659</td>
<td>0.201</td>
<td>0.632</td>
<td>0.245</td>
<td>0.630</td>
<td>0.144</td>
</tr>
<tr>
<td>BSCA [Qin et al., 2015]</td>
<td>0.758</td>
<td>0.183</td>
<td>0.723</td>
<td>0.174</td>
<td>0.666</td>
<td>0.224</td>
<td>0.634</td>
<td>0.266</td>
<td>0.616</td>
<td>0.191</td>
</tr>
<tr>
<td>MC [Zhao et al., 2015]</td>
<td>0.822</td>
<td>0.106</td>
<td>0.798</td>
<td>0.102</td>
<td>0.740</td>
<td>0.145</td>
<td>0.688</td>
<td>0.197</td>
<td>0.703</td>
<td>0.088</td>
</tr>
<tr>
<td>LEGS [Wang et al., 2015]</td>
<td>0.827</td>
<td>0.118</td>
<td>0.770</td>
<td>0.118</td>
<td>0.756</td>
<td>0.157</td>
<td>0.707</td>
<td>0.215</td>
<td>0.669</td>
<td>0.133</td>
</tr>
<tr>
<td>MDF [Li and Yu, 2015]</td>
<td>0.831</td>
<td>0.108</td>
<td>0.860</td>
<td>0.129</td>
<td>0.759</td>
<td>0.142</td>
<td>0.785</td>
<td>0.155</td>
<td>0.694</td>
<td>0.092</td>
</tr>
<tr>
<td>ELD [Lee et al., 2016]</td>
<td>0.867</td>
<td>0.080</td>
<td>0.844</td>
<td>0.071</td>
<td>0.771</td>
<td>0.121</td>
<td>0.760</td>
<td>0.154</td>
<td>0.719</td>
<td>0.091</td>
</tr>
<tr>
<td>DS [Li et al., 2016]</td>
<td>0.882</td>
<td>0.123</td>
<td>-</td>
<td>-</td>
<td>0.758</td>
<td>0.162</td>
<td>0.781</td>
<td>0.150</td>
<td>0.745</td>
<td>0.120</td>
</tr>
<tr>
<td>RFCN [Wang et al., 2016]</td>
<td>0.898</td>
<td>0.097</td>
<td>0.895</td>
<td>0.079</td>
<td>0.827</td>
<td>0.118</td>
<td>0.805</td>
<td>0.161</td>
<td>0.747</td>
<td>0.095</td>
</tr>
<tr>
<td>DCL [Li and Yu, 2016]</td>
<td>0.898</td>
<td>0.071</td>
<td>0.904</td>
<td>0.049</td>
<td>0.822</td>
<td>0.108</td>
<td>0.832</td>
<td>0.126</td>
<td>0.757</td>
<td>0.080</td>
</tr>
<tr>
<td>DHSNet [Liu and Han, 2016]</td>
<td>0.907</td>
<td>0.059</td>
<td>0.892</td>
<td>0.052</td>
<td>0.827</td>
<td>0.096</td>
<td>0.823</td>
<td>0.127</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NLDF [Luo et al., 2017]</td>
<td>0.905</td>
<td>0.063</td>
<td>0.902</td>
<td>0.048</td>
<td>0.831</td>
<td>0.099</td>
<td>0.810</td>
<td>0.143</td>
<td>0.753</td>
<td>0.080</td>
</tr>
<tr>
<td>UCF [Zhang et al., 2017b]</td>
<td>0.910</td>
<td>0.078</td>
<td>0.886</td>
<td>0.073</td>
<td>0.821</td>
<td>0.120</td>
<td>0.800</td>
<td>0.164</td>
<td>0.735</td>
<td>0.131</td>
</tr>
<tr>
<td>DSS [Hou et al., 2017]</td>
<td>0.916</td>
<td>0.053</td>
<td>0.911</td>
<td>0.040</td>
<td>0.829</td>
<td>0.102</td>
<td>0.842</td>
<td>0.118</td>
<td>0.771</td>
<td>0.066</td>
</tr>
<tr>
<td>Amulet [Zhang et al., 2017a]</td>
<td>0.913</td>
<td>0.059</td>
<td>0.887</td>
<td>0.053</td>
<td>0.828</td>
<td>0.095</td>
<td>0.801</td>
<td>0.146</td>
<td>0.737</td>
<td>0.083</td>
</tr>
<tr>
<td>SRM [Wang et al., 2017]</td>
<td>0.917</td>
<td>0.056</td>
<td>0.906</td>
<td>0.046</td>
<td>0.844</td>
<td>0.087</td>
<td>0.843</td>
<td>0.126</td>
<td>0.769</td>
<td>0.069</td>
</tr>
<tr>
<td><strong>R$^2$Net (ours)</strong></td>
<td><strong>0.935</strong></td>
<td><strong>0.040</strong></td>
<td><strong>0.916</strong></td>
<td><strong>0.036</strong></td>
<td><strong>0.845</strong></td>
<td><strong>0.100</strong></td>
<td><strong>0.847</strong></td>
<td><strong>0.124</strong></td>
<td><strong>0.805</strong></td>
<td><strong>0.063</strong></td>
</tr>
</tbody>
</table>
Visual Comparison
Ablation Analysis

- Performance increases in the first 6 iterations, and then becomes stable
- Total recurrent step: 6 (balancing the performance and time complexity)
Model with residual refinement is better than that without residual refinement
The result confirms the advantage of alternatively leveraging L and H.
A recurrent residual refinement network (R³Net) to progressively refine the saliency maps by building a sequence of RRBs to alternatively use the low-level features and high-level features.

Achieve the best performance on all the five benchmark datasets.

Code & Results:

github.com/zijundeng/R3Net
Thank you!

Code & Results: github.com/zijundeng/R3Net