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

Nian Liu[1] Junwei Han[1] Ming-Hsuan Yang[2,3]
[1]Northwestern Polytechincal University [2]University of California, Merced [3]Google 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 4^{th} and the 5^{th} 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 5^{th} to the 1^{st} Conv blocks in the decoding modules $\mathcal{D}^5$ to $\mathcal{D}^1$. We adopt global PiCANets in $\mathcal{D}^5$ and $\mathcal{D}^4$, and local PiCANets in the last three modules, respectively. In each decoding module $\mathcal{D}^i$, we use the final Conv feature map of the i^{th} Conv block in the ResNet50 encoder (*e.g.* res4f and res3d) as the incorporated encoder feature map $\boldsymbol{En}^i$ 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 $\mathcal{D}^5, \mathcal{D}^4, \cdots, \mathcal{D}^1$ 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. "+75G432LP" means using **G**lobal PiCANets in $\mathcal{D}^7$ and $\mathcal{D}^5$, and **L**ocal **P**iCANets in $\mathcal{D}^4$, $\mathcal{D}^3$, $\mathcal{D}^2$. Other settings can be inferred similarly. **Blue** indicates the best performance.

Settings	DUT-O [11]			DUTS-TE [8]		
	F_β	F_β^ω	MAE	F_β	F_β^ω	MAE
U-Net [7]	0.761	0.651	0.073	0.819	0.715	0.060
+7GP	0.772	0.660	0.071	0.826	0.722	0.058
+75GP	0.778	0.662	0.071	0.834	0.724	0.057
+75G4LP	0.785	0.678	0.069	0.840	0.736	0.056
+75G43LP	0.791	0.682	**0.068**	0.848	0.740	0.055
+75G432LP	**0.794**	**0.691**	**0.068**	**0.851**	**0.748**	**0.054**

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 $\mathcal{D}^7, \mathcal{D}^5, \mathcal{D}^4, \cdots, \mathcal{D}^1$ 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 $\mathcal{D}^7$ and $\mathcal{D}^5$ can attend to foreground objects for background pixels and backgrounds for foreground pixels. The local attention learned in $\mathcal{D}^4$, $\mathcal{D}^3$, and $\mathcal{D}^2$ 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

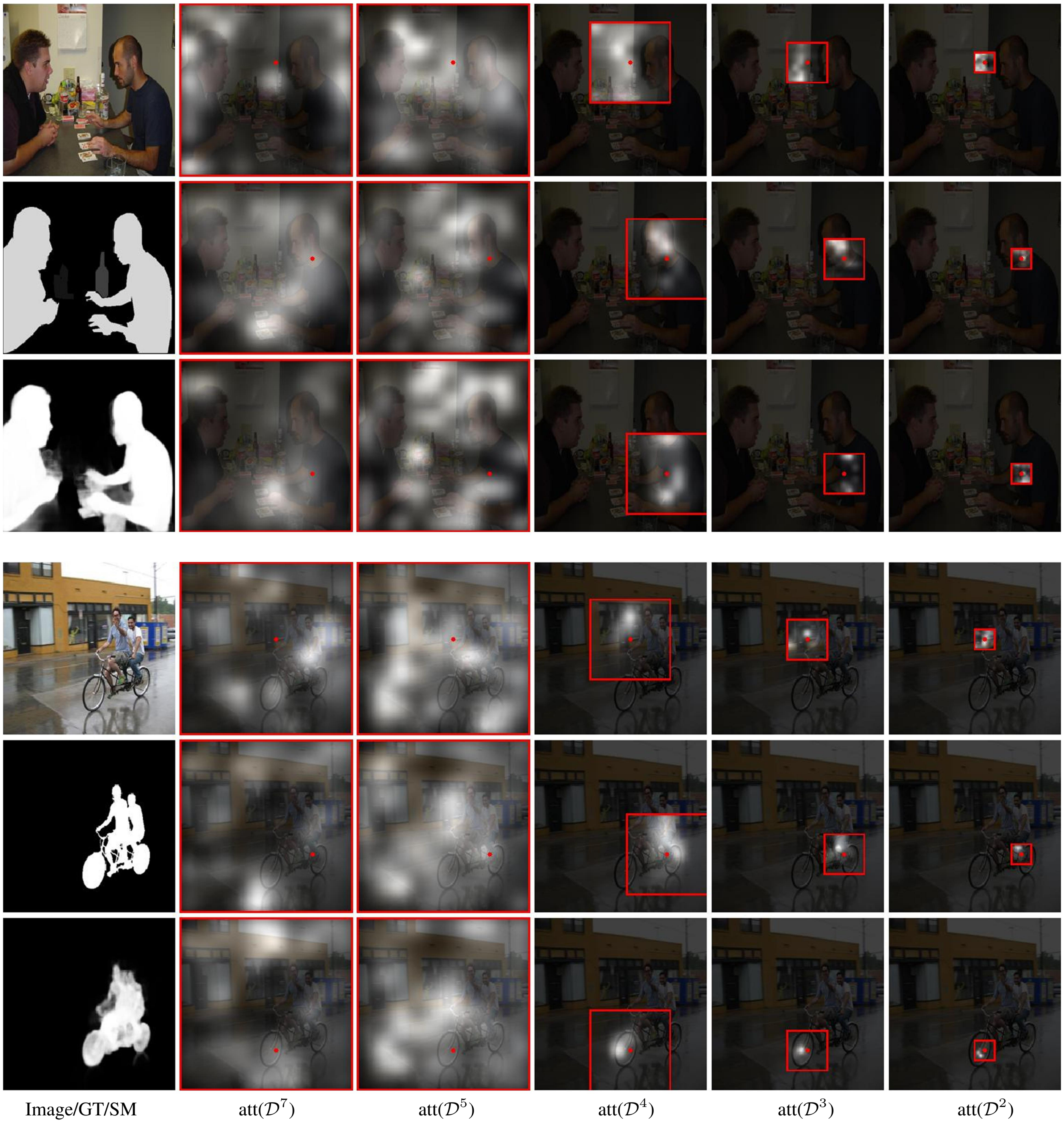

Figure 1. Illustration of the learned attention maps of the proposed PiCANets. The first column shows two images and their corresponding ground truth masks and the predicted saliency maps of our model while the last five columns show the attention maps in five attended decoding modules, respectively. For each image, we give three example pixels (denoted as red dots. The first row shows a background pixel and the bottom two rows show two foreground pixels). The attended context regions are marked by red rectangles.

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

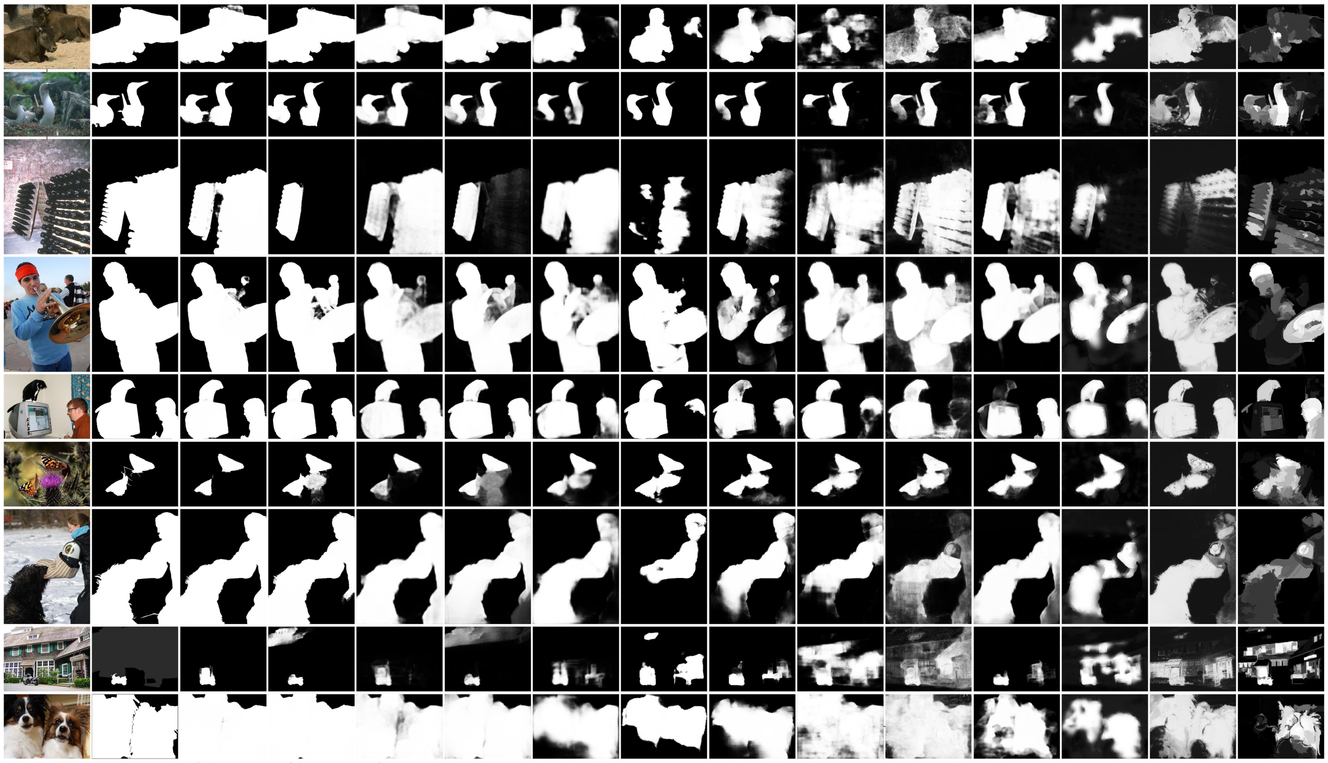

Figure 2. Qualitative comparison. (GT: ground truth)

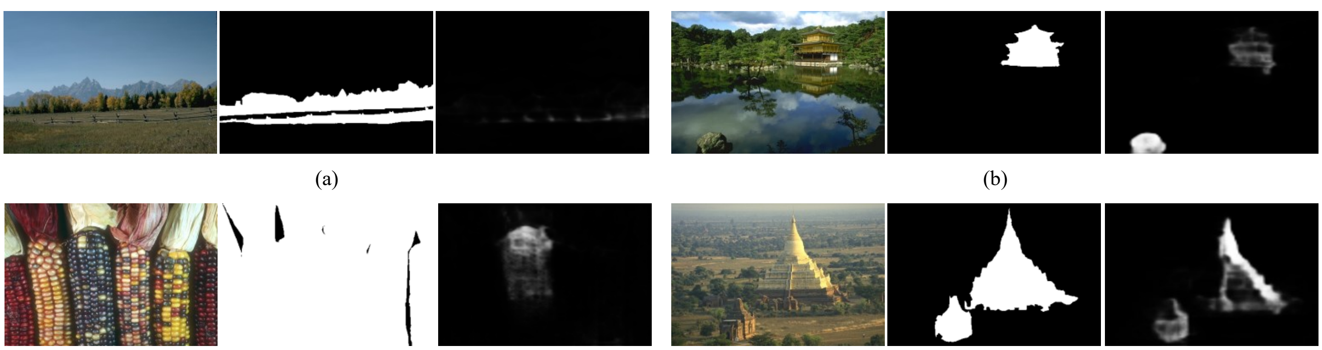

Figure 3. Failure cases. The three images in each set are the input image, the ground truth, and our result, respectively.

(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

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[2] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. In *CVPR*, 2017. 1, 3

[3] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *CVPR*, 2015. 3

[4] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *CVPR*, 2016. 3

[5] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016. 3

[6] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin. Non-local deep features for salient object detection. In *CVPR*, 2017. 3

[7] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MIC-*

CAI, 2015. 1

[8] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 1

[9] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, 2016. 3

[10] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, 2017. 3

[11] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 1

[12] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017. 3

[13] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, 2017. 3