Recurrently Aggregating Deep Features for Salient Object Detection

Xiaowei Hu[1], Lei Zhu[2], Jing Qin[2], Chi-Wing Fu[1,3], and Pheng-Ann Heng[1,3]

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[2] Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong
[3] Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

Introduction

- **Task:**
  Salient object detection aims to identify the most visually distinctive objects in an input image.

- **Related Works:**
  - Hand-crafted visual features (e.g, color, texture, and contrast) with heuristic priors: ineffective to capture the high-level semantic knowledge.
  - Traditional fully convolutional neural network (FCN) based methods: neglect many fine details.
  - FCN-based methods with multi-level integrated features (MLIF): tend to contain many non-salient objects and simultaneously lose some parts (details) of salient objects (as shown in Figure 1).

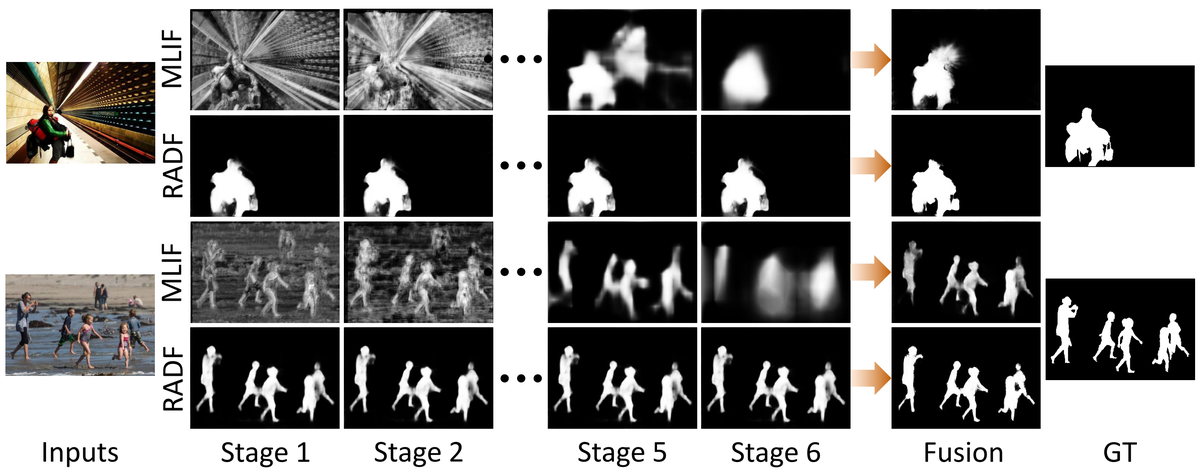

Figure 1. Visual comparisons of predicted saliency maps from different features. The 1st and 3rd rows in the middle show the predicted saliency maps using the multi-level integrated features (MLIF) and features at each layer, while the 2nd and 4th rows in the middle use our proposed method (RADF). Stages 1 to 6 show the saliency maps predicted from the shallow layers to deep layers.

- **Contributions of Recurrently Aggregating Deep Features (RADF):**
  - Effectively leverage the complementary information encoded in the deep features generated in different layers.
  - A novel scheme to aggregate the multi-level deep features to features of each layer in a recurrent manner. It leads to more distinguishing features containing both semantic and detailed information of salient objects.
  - Set a new state-of-the-art performance on salient object detection by comparing with 17 algorithms on 5 benchmarks.

Method

- **Architecture of RADF:**

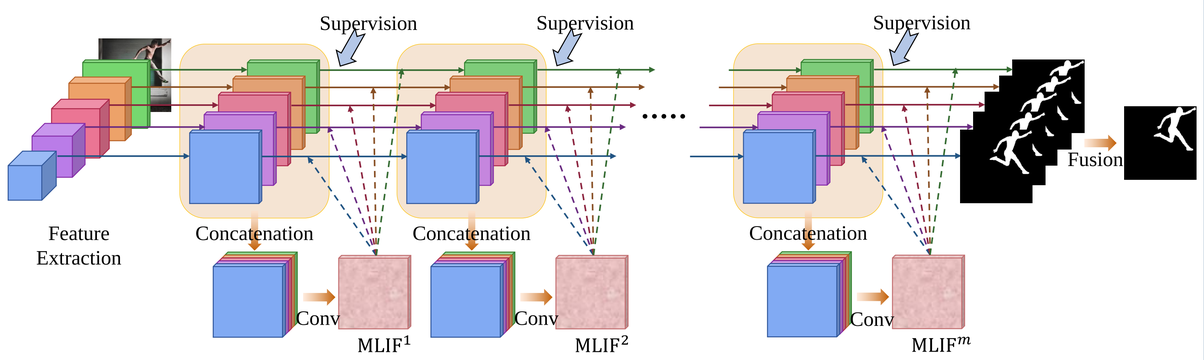

Figure 2. Schematic illustration of RADF. We extract features in different scales over the CNN layers from the input image. The feature maps with different scales are up-sampled to the size of input image and concatenated together as the multi-level integrated features (MLIF). The MLIF is added to the features of each layer and merged by a convolutional operation. This step is performed *m* iterations to alternatively refine MLIF and the features at each layer. Moreover, the deep supervision mechanism is imposed at each step. Finally, output score maps at the last step are merged together to generate the fusion score map.

- **Discussion**

1) Recurrently aggregate the MLIF to each individual layer.
   * Suppress non-saliency regions on the feature maps of shallow layers.
   * Enhance saliency boundary details on the feature maps of deep layers.
2) Refined features in individual layers are integrated together as refined MLIF.

Experiments

- **Comparison with the State-of-the-arts:**

Table 1. The top three results are highlighted in red, green, and blue, respectively. A better performance has a larger F-measure value and a smaller MAE value.

Method	ECSSD		HKU-IS		PASCAL-S		SOD		DUT-OMRON	
	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
MR (Yang et al. 2013)	0.736	0.189	0.715	0.174	0.666	0.223	0.619	0.273	0.610	0.187
wCtr* (Zhu et al. 2014)	0.716	0.171	0.726	0.141	0.659	0.201	0.632	0.245	0.630	0.144
BSCA (Qin et al. 2015)	0.758	0.183	0.723	0.174	0.666	0.224	0.634	0.266	0.616	0.191
MC (Zhao et al. 2015)	0.822	0.106	0.798	0.102	0.740	0.145	0.688	0.197	0.703	0.088
LEGS (Wang et al. 2015)	0.827	0.118	0.770	0.118	0.756	0.157	0.707	0.215	0.669	0.133
MDF (Li and Yu 2015)	0.831	0.108	0.860	0.129	0.759	0.142	0.785	0.155	0.694	0.092
ELD (Lee, Tai, and Kim 2016)	0.867	0.080	0.844	0.071	0.771	0.121	0.760	0.154	0.719	0.091
DS (Li et al. 2016)	0.882	0.123	-	-	0.758	0.162	0.781	0.150	0.745	0.120
FPN (Lin et al. 2016)	0.895	0.062	0.896	0.044	0.793	0.114	0.808	0.126	0.730	0.084
DeepLab (Chen et al. 2016)	0.904	0.053	0.890	0.041	0.812	0.108	0.810	0.128	0.765	0.068
RFCN (Wang et al. 2016)	0.898	0.097	0.895	0.079	0.827	0.118	0.805	0.161	0.747	0.095
DCL (Li and Yu 2016)	0.898	0.071	0.904	0.049	0.822	0.108	0.832	0.126	0.757	0.080
DHSNet (Liu and Han 2016)	0.907	0.059	0.892	0.052	0.827	0.096	0.823	0.127	-	-
NLDF (Luo et al. 2017)	0.905	0.063	0.902	0.048	0.831	0.099	0.810	0.143	0.753	0.080
UCF (Zhang et al. 2017b)	0.910	0.078	0.886	0.073	0.821	0.120	0.800	0.164	0.735	0.131
DSS (Hou et al. 2017)	0.916	0.053	0.911	0.040	0.829	0.102	0.842	0.118	0.771	0.066
Amulet (Zhang et al. 2017a)	0.913	0.059	0.887	0.053	0.828	0.095	0.801	0.146	0.737	0.083
RADF (ours)	0.924	0.049	0.914	0.039	0.832	0.102	0.835	0.125	0.789	0.060

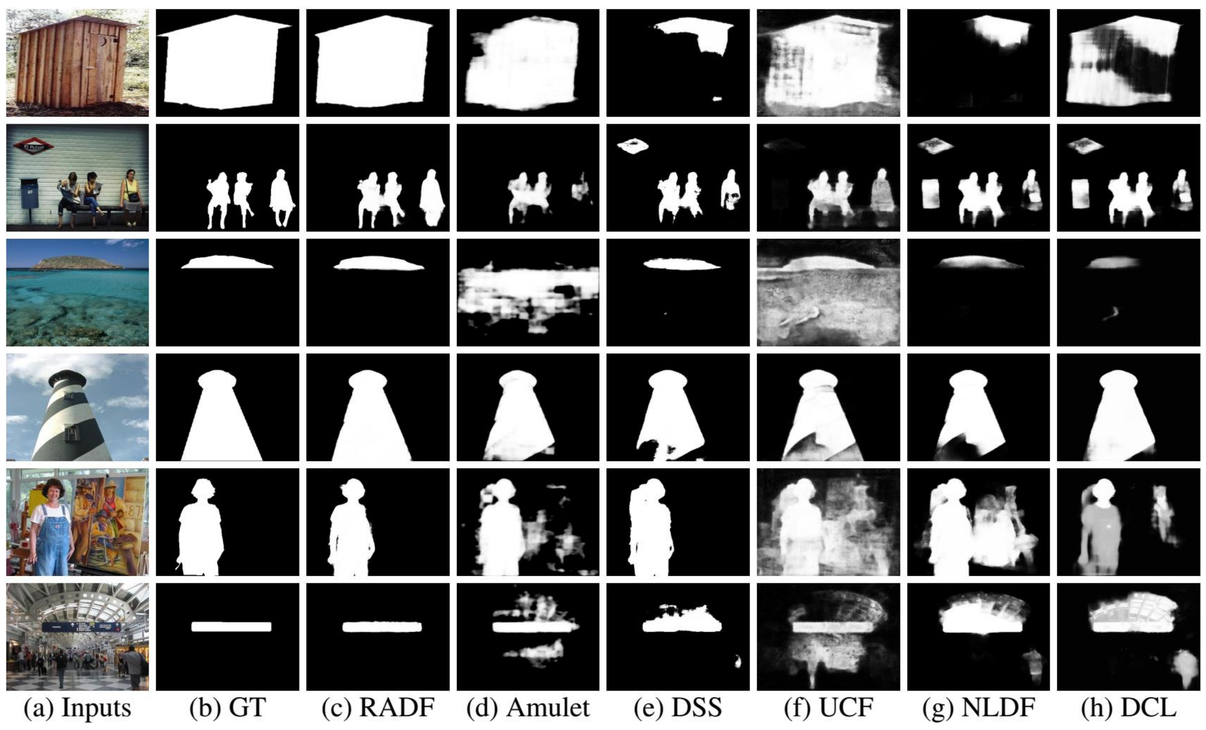

Figure 3. Visual comparison of saliency maps. Note that "GT" stands for "Ground truths". Apparently, our method (RADF) can produce more accurate saliency maps than others. More comparisons can be found at the paper's website (QR code is on the top of poster).

- **Ablation Analysis:**

Table 2. The F-measure and MAE of different settings on five saliency detection datasets.

Method	ECSSD		HKU-IS		PASCAL-S		SOD		DUT-OMRON	
	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
RADF-D	0.886	0.068	0.865	0.060	0.781	0.126	0.788	0.148	0.729	0.086
RADF-i	0.913	0.054	0.908	0.041	0.826	0.105	0.827	0.129	0.763	0.072
RADF-m	0.917	0.053	0.907	0.046	0.828	0.107	0.829	0.131	0.772	0.066
RADF1	0.923	0.050	0.913	0.041	0.828	0.106	0.831	0.129	0.787	0.063
RADF2	**0.924**	**0.049**	**0.914**	**0.039**	**0.832**	**0.102**	**0.835**	**0.125**	0.789	**0.060**
RADF2-s	0.923	**0.049**	0.911	0.043	0.830	0.105	0.829	0.131	**0.793**	0.062

1) "RADF-D" uses DenseNet-161 to extract features (Others use VGG-16).
2) "RADF-I" predicts the saliency maps just based on the features of each layer.
3) "RADF-m" uses the MLIF to predict the saliency maps directly.
4) "RADF-1/2" denotes we set different number of steps to aggregate the deep features.
5) "RADF2-s" represents the weights are shared in the two recurrent steps.

Conclusion

We present a novel FCN with recurrently aggregated deep features for salient object detection by employing multi-level features to progressively refine the features of each layer. During the recurrent aggregation procedure, non-salient noises in low-level features are gradually reduced and the saliency details in high-level features are continuously enhanced. The proposed feature aggregation scheme is general enough and has great potential to be used in other applications such as object detection and semantic segmentation.

e-mail: xwhu@cse.cuhk.edu.hk