$\mathbf{R^3}$**Net:** Recurrent Residual Refinement Network for Saliency Detection

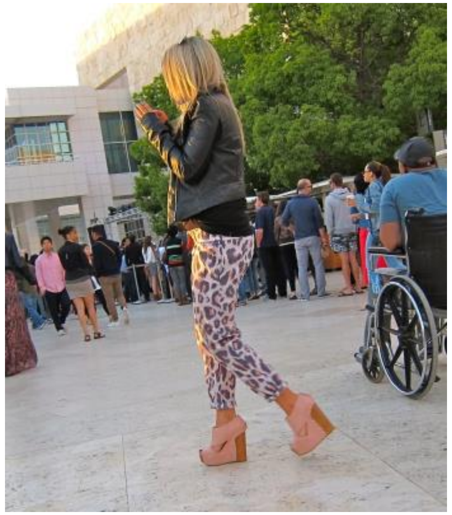
Zijun Deng[1,*], Xiaowei Hu[2,*], Lei Zhu[3,2], Xuemiao Xu[1], Jing Qin[3], Guoqiang Han[1], and Pheng–Ann Heng[2,4]

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

IJCAI ECAI 2018

Saliency Detection

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



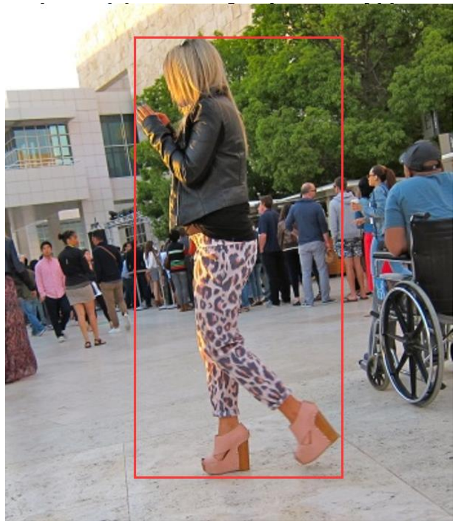
input image

saliency map

➢ 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)

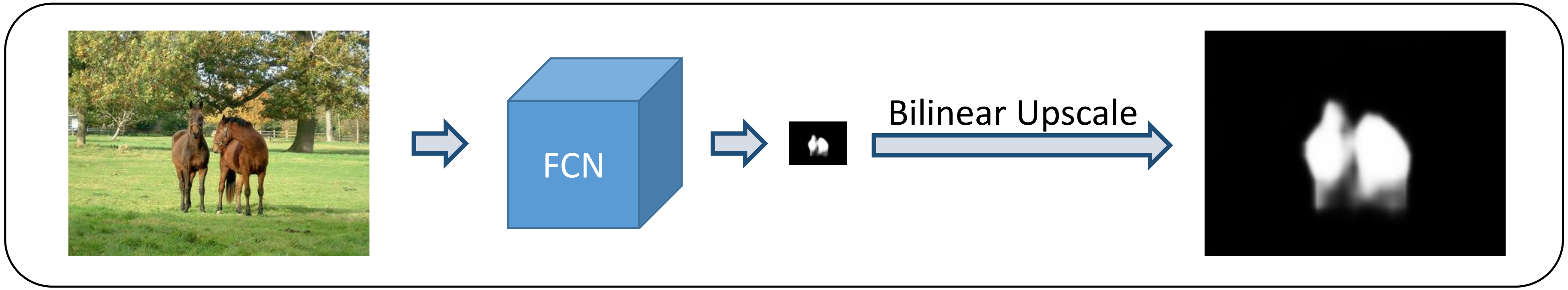


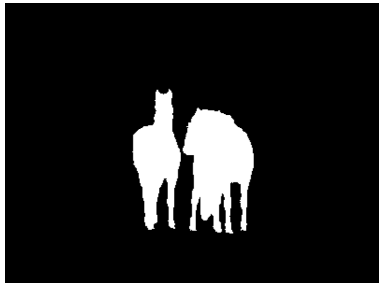
2. Segment the accurate regions of salient objects
   - Precise object localization

"Salient object detection: A survey”, Ali, Cheng, Hou, Jiang, Li. ArXiv 2014

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

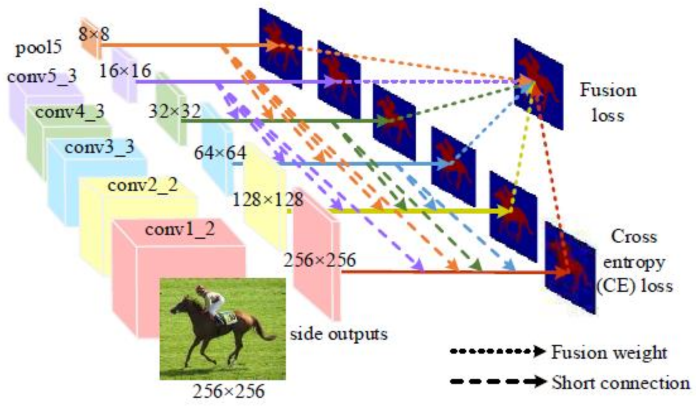




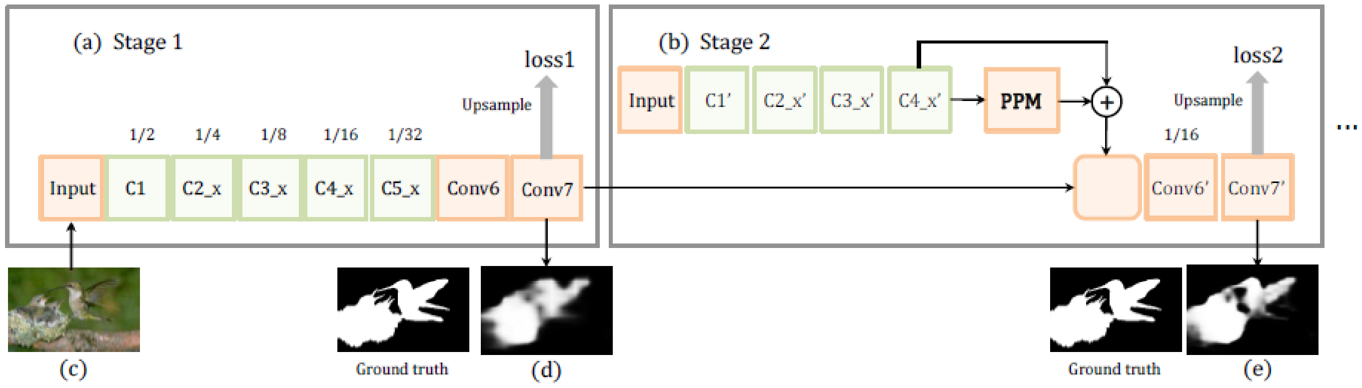
What we want indeed!

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]



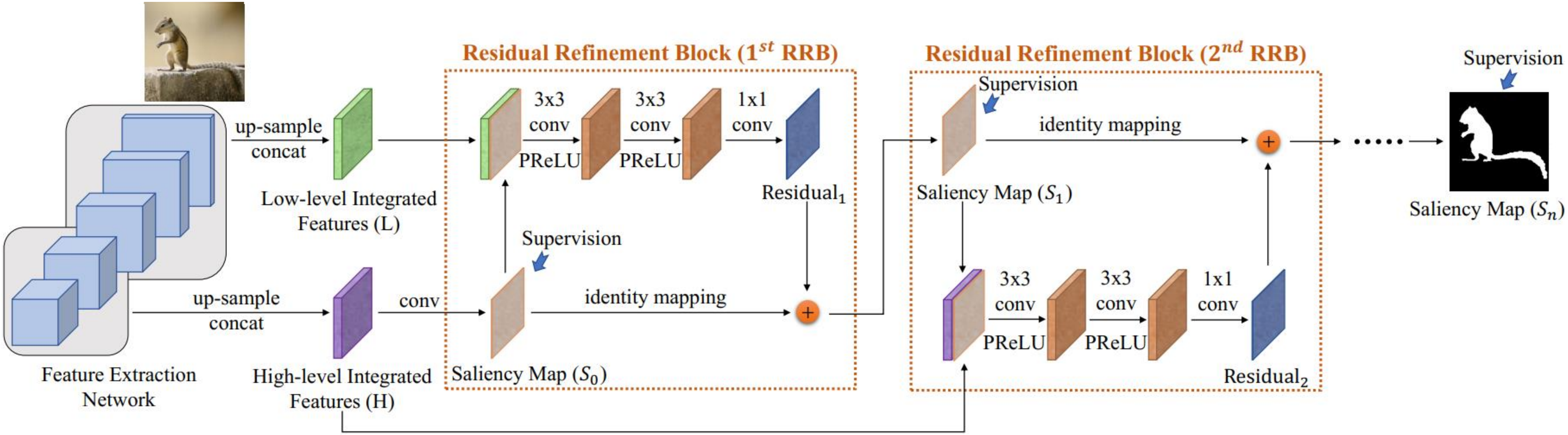
SRM [Wang et al., 2017]

Our Motivation

- ***Alternatively*** leverage the low-level detailed features and the high-level semantic features to do refinement

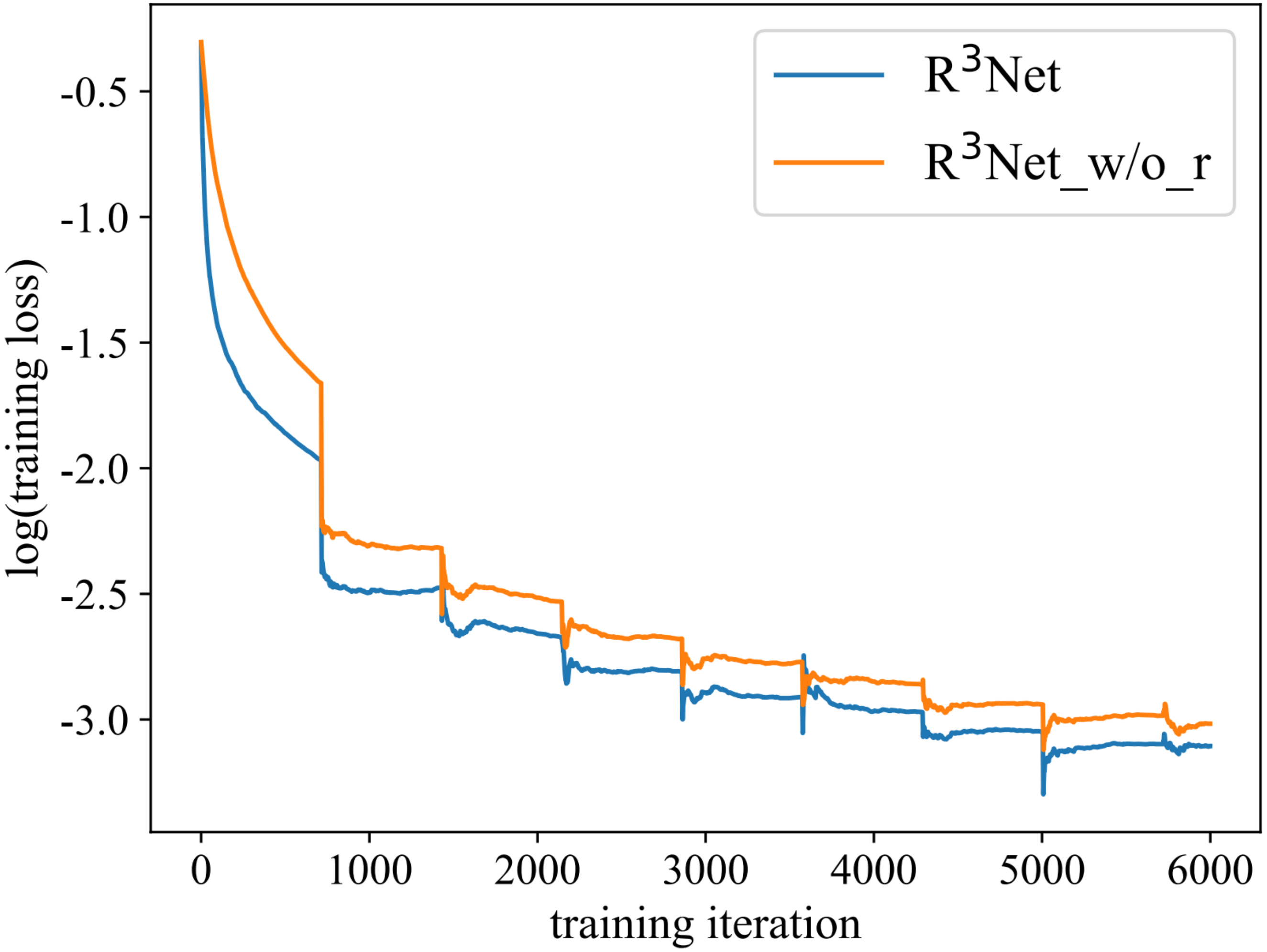
- Apply ***residual learning*** to saliency map refinement

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

Experimental Setting

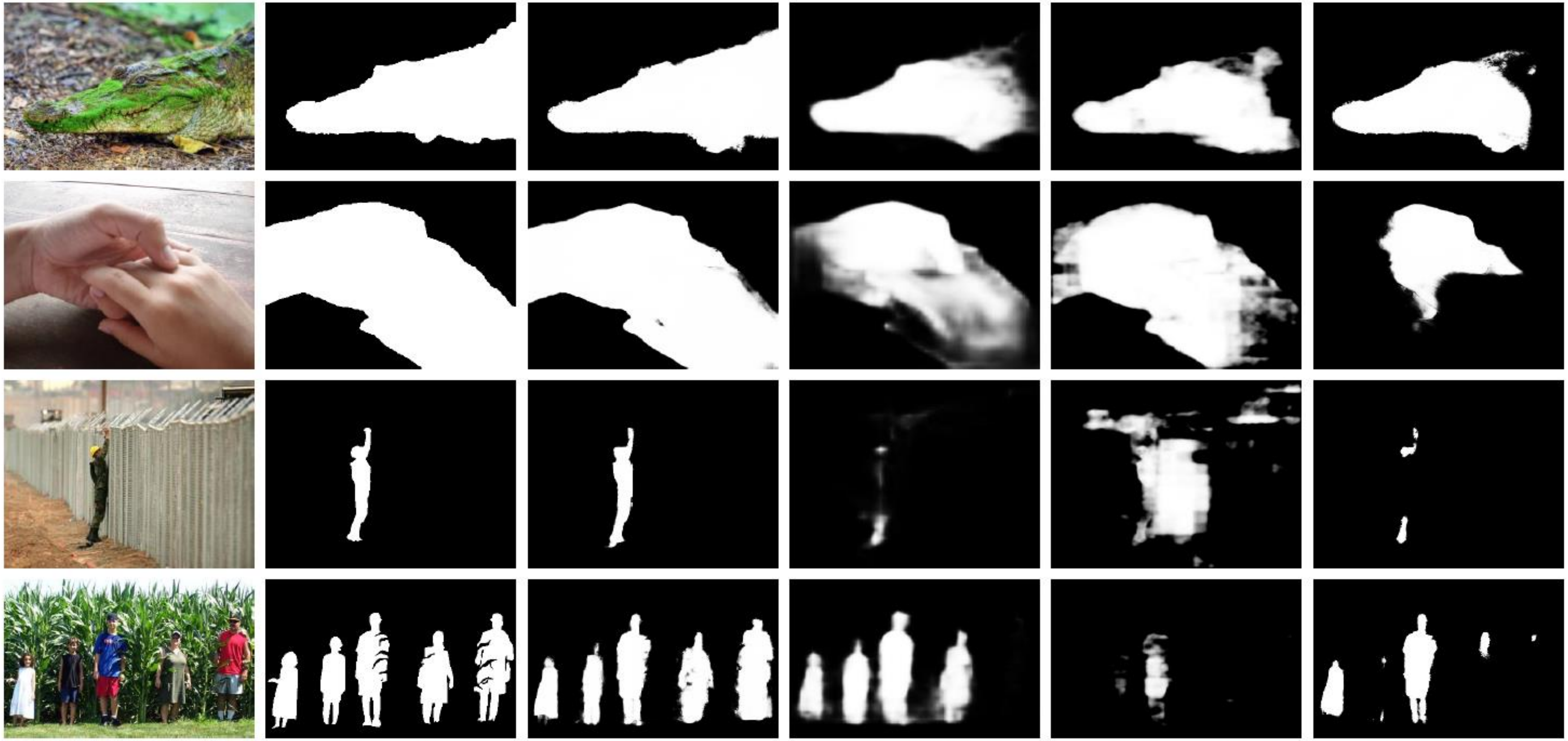
- Training
  - On the MSRA10K dataset (10K images)
  - ResNeXt101-32x4d as feature extraction network, pre-trained on ImageNet
  - Takes only 80 minutes on a single GPU

- 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

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 *et al.*, 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
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
SRM [Wang *et al.*, 2017]	0.917	0.056	0.906	0.046	0.844	0.087	0.843	0.126	0.769	0.069
R^3Net (ours)	0.935	0.040	0.916	0.036	0.845	0.100	0.847	0.124	0.805	0.063

Visual Comparison



(a) Inputs (b) GT (c) Ours (d) SRM (e) Amulet (f) DSS

Ablation Analysis

Method	ECSSD		HKU-IS		PASCAL-S		SOD		DUT-OMRON	
	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
R^3Net-0	0.918	0.049	0.900	0.044	0.831	0.101	0.816	0.128	0.769	0.079
R^3Net-1	0.926	0.044	0.910	0.038	0.841	0.100	0.833	0.125	0.783	0.073
R^3Net-2	0.931	0.043	0.911	0.038	0.844	0.104	0.836	0.127	0.787	0.073
R^3Net-3	0.934	0.041	0.915	**0.036**	0.847	0.100	0.841	0.123	0.794	0.066
R^3Net-4	0.932	0.042	0.912	0.038	0.843	0.102	0.841	0.125	0.782	0.073
R^3Net-5	0.933	0.042	0.913	0.037	0.845	0.100	0.841	0.122	0.791	0.069
R^3Net-6	**0.935**	**0.040**	**0.916**	**0.036**	0.845	0.100	**0.847**	0.124	**0.805**	**0.063**
R^3Net-7	0.934	**0.040**	0.914	**0.036**	**0.848**	**0.096**	0.842	**0.121**	0.804	**0.063**
R^3Net_w/o_r	0.931	0.042	0.910	0.039	0.839	0.103	0.839	**0.121**	0.782	0.077
R^3Net_w_s	0.933	0.041	0.914	0.037	0.841	0.102	0.842	0.122	0.794	0.070
R^3Net_LL	0.932	0.041	0.910	0.038	0.844	0.100	0.839	0.125	0.778	0.080
R^3Net_HH	0.926	0.046	0.902	0.042	0.836	0.101	0.819	0.128	0.786	0.071

- Performance increases in the first 6 iterations, and then becomes stable
- Total recurrent step: 6 (balancing the performance and time complexity)

Ablation Analysis

Method	ECSSD		HKU-IS		PASCAL-S		SOD		DUT-OMRON	
	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
R^3Net-0	0.918	0.049	0.900	0.044	0.831	0.101	0.816	0.128	0.769	0.079
R^3Net-1	0.926	0.044	0.910	0.038	0.841	0.100	0.833	0.125	0.783	0.073
R^3Net-2	0.931	0.043	0.911	0.038	0.844	0.104	0.836	0.127	0.787	0.073
R^3Net-3	0.934	0.041	0.915	**0.036**	0.847	0.100	0.841	0.123	0.794	0.066
R^3Net-4	0.932	0.042	0.912	0.038	0.843	0.102	0.841	0.125	0.782	0.073
R^3Net-5	0.933	0.042	0.913	0.037	0.845	0.100	0.841	0.122	0.791	0.069
R^3Net-6	**0.935**	**0.040**	**0.916**	**0.036**	0.845	0.100	**0.847**	0.124	**0.805**	**0.063**
R^3Net-7	0.934	**0.040**	0.914	**0.036**	**0.848**	**0.096**	0.842	**0.121**	0.804	**0.063**
R^3Net_w/o_r	0.931	0.042	0.910	0.039	0.839	0.103	0.839	**0.121**	0.782	0.077
R^3Net_w_s	0.933	0.041	0.914	0.037	0.841	0.102	0.842	0.122	0.794	0.070
R^3Net_LL	0.932	0.041	0.910	0.038	0.844	0.100	0.839	0.125	0.778	0.080
R^3Net_HH	0.926	0.046	0.902	0.042	0.836	0.101	0.819	0.128	0.786	0.071

- Model with residual refinement is better than that without residual refinement

Ablation Analysis

Method	ECSSD		HKU-IS		PASCAL-S		SOD		DUT-OMRON	
	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
R^3Net-0	0.918	0.049	0.900	0.044	0.831	0.101	0.816	0.128	0.769	0.079
R^3Net-1	0.926	0.044	0.910	0.038	0.841	0.100	0.833	0.125	0.783	0.073
R^3Net-2	0.931	0.043	0.911	0.038	0.844	0.104	0.836	0.127	0.787	0.073
R^3Net-3	0.934	0.041	0.915	**0.036**	0.847	0.100	0.841	0.123	0.794	0.066
R^3Net-4	0.932	0.042	0.912	0.038	0.843	0.102	0.841	0.125	0.782	0.073
R^3Net-5	0.933	0.042	0.913	0.037	0.845	0.100	0.841	0.122	0.791	0.069
R^3Net-6	**0.935**	**0.040**	**0.916**	**0.036**	0.845	0.100	**0.847**	0.124	**0.805**	**0.063**
R^3Net-7	0.934	**0.040**	0.914	**0.036**	**0.848**	**0.096**	0.842	**0.121**	0.804	**0.063**
R^3Net_w/o_r	0.931	0.042	0.910	0.039	0.839	0.103	0.839	**0.121**	0.782	0.077
R^3Net_w_s	0.933	0.041	0.914	0.037	0.841	0.102	0.842	0.122	0.794	0.070
R^3Net_LL	0.932	0.041	0.910	0.038	0.844	0.100	0.839	0.125	0.778	0.080
R^3Net_HH	0.926	0.046	0.902	0.042	0.836	0.101	0.819	0.128	0.786	0.071

- The result confirms the advantage of alternatively leveraging L and H.

Conclusion

- A recurrent residual refinement network (R^3Net) 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

Q & A

Thank you!

Code & Results:
github.com/zijundeng/R3Net