Referring Image Segmentation via Recurrent Refinement Networks Supplementary Material

Ruiyu Li[†], Kaican Li[†], Yi-Chun Kuo[†], Michelle Shu[‡],
Xiaojuan Qi[†], Xiaoyong Shen[§], Jiaya Jia[†,§]

[†]The Chinese University of Hong Kong, [‡]Johns Hopkins University, [§]YouTu Lab, Tencent

{ryli,kcli5,yckuo5,xjqi,leojia}@cse.cuhk.edu.hk, mshu1@jhu.edu, goodshenxy@gmail.com

Table 1, 2, 3, 4 present the full experimental results on ReferIt [3], UNC [6], UNC+ [6], G-Ref [5] datasets respectively. More visualization and segmentation masks are shown in Fig. 1-8.

References

[1] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 1

[2] R. Hu, M. Rohrbach, S. Venugopalan, and T. Darrell. Utilizing large scale vision and text datasets for image segmentation from referring expressions. *arXiv preprint arXiv:1608.08305*, 2016. 3

[3] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1

[4] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, 2017. 1, 2, 3

[5] J. Mao, H. Jonathan, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. *arXiv preprint arXiv:1511.02283*, 2015. 1

[6] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1

Model	Set	*prec*@0.5	*prec*@0.6	*prec*@0.7	*prec*@0.8	*prec*@0.9	ovreall IoU
LSTM-CNN [1]	test	34.02	26.71	19.32	11.63	3.92	48.03
DeepLab+RMI [4]	test	44.33	36.13	27.20	16.99	6.43	57.34
DeepLab+RMI+DCRF [4]	test	46.08	38.90	30.77	20.62	8.54	58.73
RRN (with plain structure)	test	50.41	42.81	34.39	23.97	11.39	60.66
RRN (with plain structure, DCRF)	test	51.13	44.25	36.16	25.51	11.48	61.11
RRN (with vanilla RNN)	test	51.19	43.41	34.59	24.13	11.59	60.86
RRN (with vanilla RNN, DCRF)	test	52.01	44.78	36.58	25.57	11.65	61.29
RRN (with LSTM)	test	55.72	47.78	38.49	26.72	12.53	63.12
RRN (with LSTM, DCRF)	test	**56.71**	**49.22**	**40.36**	**28.39**	**12.68**	**63.63**

Table 1. Experimental results on ReferIt dataset.

Model	Set	*prec*@0.5	*prec*@0.6	*prec*@0.7	*prec*@0.8	*prec*@0.9	ovreall IoU
DeepLab+RMI [4]	val	41.27	29.71	18.41	7.37	0.76	44.33
DeepLab+RMI+DCRF [4]	val	42.99	33.24	22.75	12.11	2.23	45.18
RRN (with plain structure)	val	51.84	42.36	30.64	17.39	3.84	49.74
RRN (with plain structure, DCRF)	val	53.07	44.27	34.45	22.08	6.19	50.56
RRN (with vanilla RNN)	val	49.85	40.22	29.49	16.62	4.05	48.86
RRN (with vanilla RNN, DCRF)	val	51.46	42.77	33.19	20.91	6.24	49.51
RRN (with LSTM)	val	60.19	50.19	38.32	23.87	5.66	54.26
RRN (with LSTM, DCRF)	val	**61.66**	**52.50**	**42.40**	**28.13**	**8.51**	**55.33**
DeepLab+RMI [4]	testA	40.68	30.14	18.99	8.03	0.88	44.74
DeepLab+RMI+DCRF [4]	testA	42.99	33.59	23.69	12.94	2.44	45.69
RRN (with plain structure)	testA	53.46	43.49	31.89	18.37	3.82	51.31
RRN (with plain structure, DCRF)	testA	54.30	45.84	34.93	23.09	6.63	52.12
RRN (with vanilla RNN)	testA	51.42	41.70	31.04	17.89	3.75	49.79
RRN (with vanilla RNN, DCRF)	testA	52.06	43.43	34.40	22.61	6.36	50.41
RRN (with LSTM)	testA	63.00	52.93	40.99	24.47	5.50	56.21
RRN (with LSTM, DCRF)	testA	**64.13**	**54.66**	**44.37**	**29.15**	**8.08**	**57.26**
DeepLab+RMI [4]	testB	42.75	30.40	18.19	7.83	0.86	44.63
DeepLab+RMI+DCRF [4]	testB	44.99	34.21	22.69	11.84	2.65	45.57
RRN (with plain structure)	testB	50.74	40.37	29.38	17.29	4.95	49.49
RRN (with plain structure, DCRF)	testB	51.91	42.47	33.50	21.37	8.15	50.34
RRN (with vanilla RNN)	testB	48.81	39.54	29.28	18.07	5.32	48.68
RRN (with vanilla RNN, DCRF)	testB	50.17	41.57	32.62	22.02	8.07	49.46
RRN (with LSTM)	testB	57.51	47.71	36.51	22.87	6.91	52.71
RRN (with LSTM, DCRF)	testB	**59.35**	**50.32**	**39.82**	**27.30**	**10.05**	**53.95**

Table 2. Experimental results on UNC dataset.

Model	Set	*prec*@0.5	*prec*@0.6	*prec*@0.7	*prec*@0.8	*prec*@0.9	ovreall IoU
DeepLab+RMI [4]	val	18.39	11.50	5.86	1.85	0.20	29.91
DeepLab+RMI+DCRF [4]	val	20.52	14.02	8.46	3.77	0.62	29.86
RRN (with plain structure)	val	21.82	14.83	8.78	4.11	0.61	32.73
RRN (with plain structure, DCRF)	val	23.22	16.59	10.83	5.78	1.12	32.50
RRN (with vanilla RNN)	val	22.53	15.22	8.82	3.98	0.48	32.84
RRN (with vanilla RNN, DCRF)	val	23.77	17.22	11.08	5.71	0.90	32.61
RRN (with LSTM)	val	35.45	25.93	16.60	8.11	1.19	39.23
RRN (with LSTM, DCRF)	val	**37.32**	**28.96**	**20.31**	**11.33**	**2.66**	**39.75**
DeepLab+RMI [4]	testA	18.76	11.67	6.08	1.78	0.26	30.37
DeepLab+RMI+DCRF [4]	testA	21.22	14.43	8.99	3.91	0.49	30.48
RRN (with plain structure)	testA	25.10	17.46	10.57	4.86	0.86	34.61
RRN (with plain structure, DCRF)	testA	26.21	19.66	13.20	7.28	1.43	34.50
RRN (with vanilla RNN)	testA	26.27	18.37	11.23	4.92	0.77	34.63
RRN (with vanilla RNN, DCRF)	testA	28.08	20.40	13.90	7.72	1.68	34.47
RRN (with LSTM)	testA	39.71	29.11	19.04	9.24	1.34	41.68
RRN (with LSTM, DCRF)	testA	**40.80**	**31.66**	**22.74**	**12.78**	**2.78**	**42.15**
DeepLab+RMI [4]	testB	19.08	12.11	6.44	2.70	0.31	29.43
DeepLab+RMI+DCRF [4]	testB	20.78	14.56	8.80	4.58	0.80	29.50
RRN (with plain structure)	testB	18.86	12.31	7.65	3.80	0.84	29.86
RRN (with plain structure, DCRF)	testB	19.88	14.17	9.39	5.17	1.60	29.61
RRN (with vanilla RNN)	testB	18.98	12.66	7.61	3.58	0.76	29.96
RRN (with vanilla RNN, DCRF)	testB	20.31	14.58	9.29	5.11	1.23	29.78
RRN (with LSTM)	testB	30.19	21.64	14.03	7.57	1.43	35.63
RRN (with LSTM, DCRF)	testB	**32.42**	**24.69**	**17.10**	**9.92**	**2.78**	**36.11**

Table 3. Experimental results on UNC+ dataset.

Model	Set	*prec*@0.5	*prec*@0.6	*prec*@0.7	*prec*@0.8	*prec*@0.9	ovreall IoU
LSTM-CNN [2]	val	15.25	8.37	3.75	1.29	0.06	28.14
DeepLab+RMI [4]	val	26.19	18.46	10.68	4.28	0.73	34.40
DeepLab+RMI+DCRF [4]	val	27.77	21.06	13.92	6.83	1.43	34.52
RRN (with plain structure)	val	30.47	22.92	15.87	8.80	2.12	34.43
RRN (with plain structure, DCRF)	val	31.30	24.56	17.85	10.89	3.26	34.40
RRN (with vanilla RNN)	val	28.42	21.48	14.69	7.75	1.70	33.92
RRN (with vanilla RNN, DCRF)	val	28.98	22.77	16.63	9.68	2.66	33.66
RRN (with LSTM)	val	35.01	27.65	19.89	10.93	2.38	36.32
RRN (with LSTM, DCRF)	val	**36.00**	**29.77**	**22.78**	**14.06**	**3.74**	**36.45**

Table 4. Experimental results on G-Ref dataset.

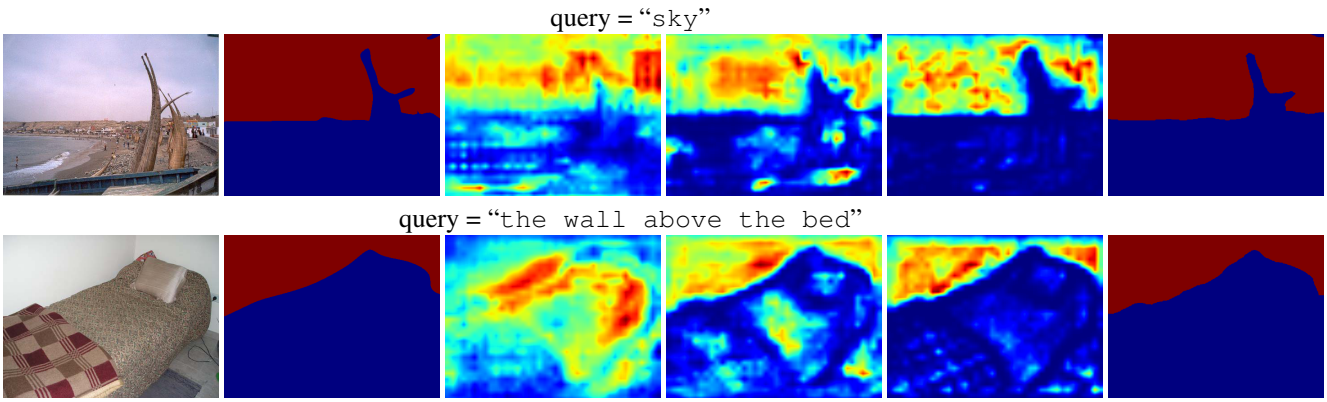

Figure 1. Visualization of convolutional LSTM on ReferIt dataset. From left to right are input images, ground truth masks, the strongest activated channel of hidden states after combining C_5, C_4,C_3 features, and the predicted mask.

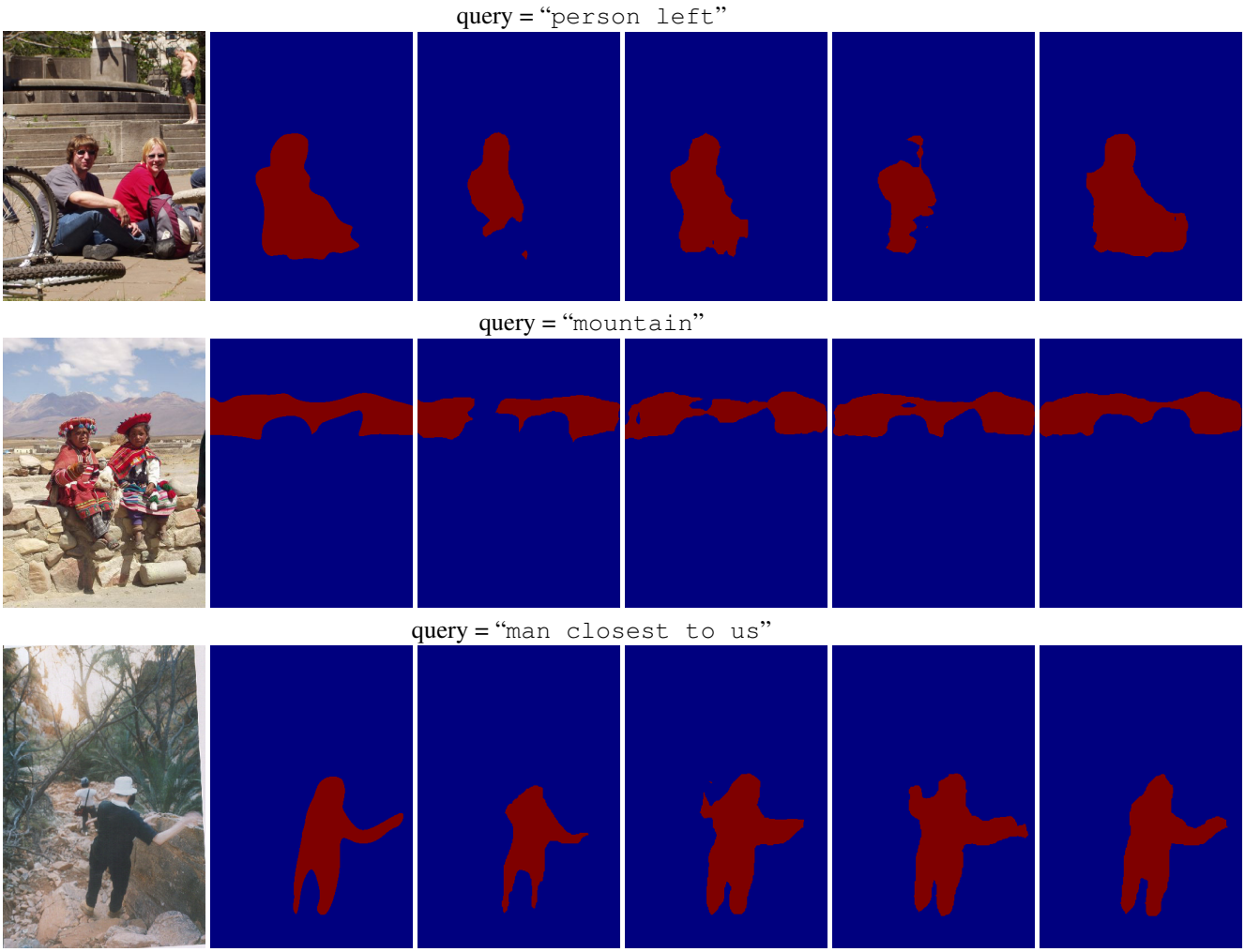

Figure 2. Segmentation results on ReferIt dataset. From left to right are input images, ground truth masks, results from baseline, plain structure, RNN, and LSTM respectively.

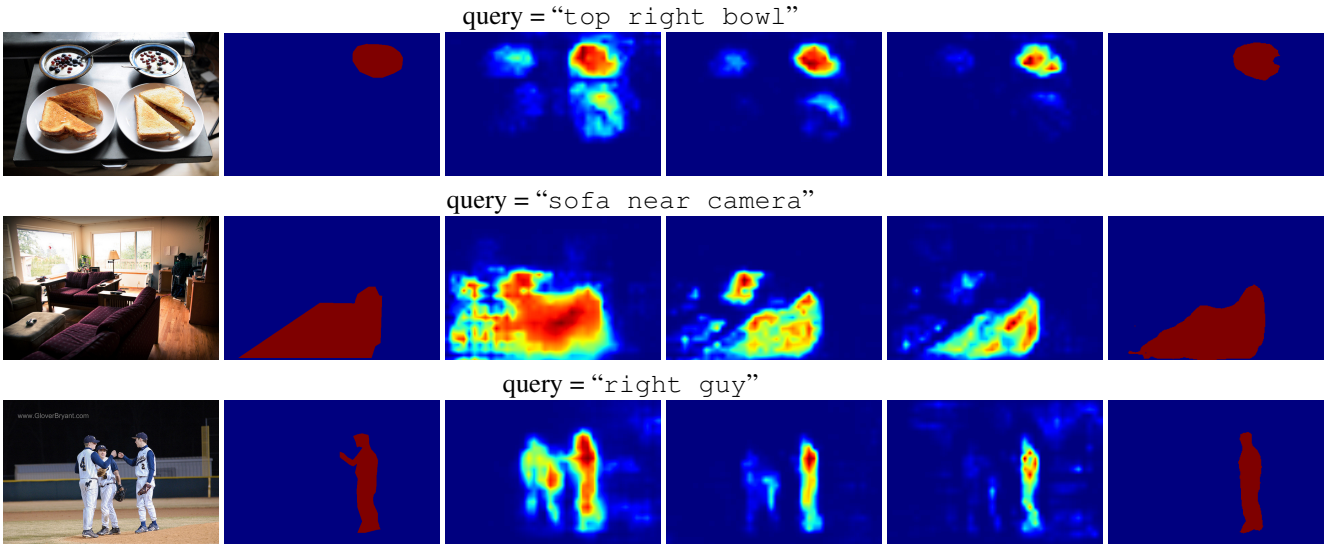

Figure 3. Visualization of convolutional LSTM on UNC dataset. From left to right are input images, ground truth masks, the strongest activated channel of hidden states after combining C_5, C_4,C_3 features, and the predicted mask.

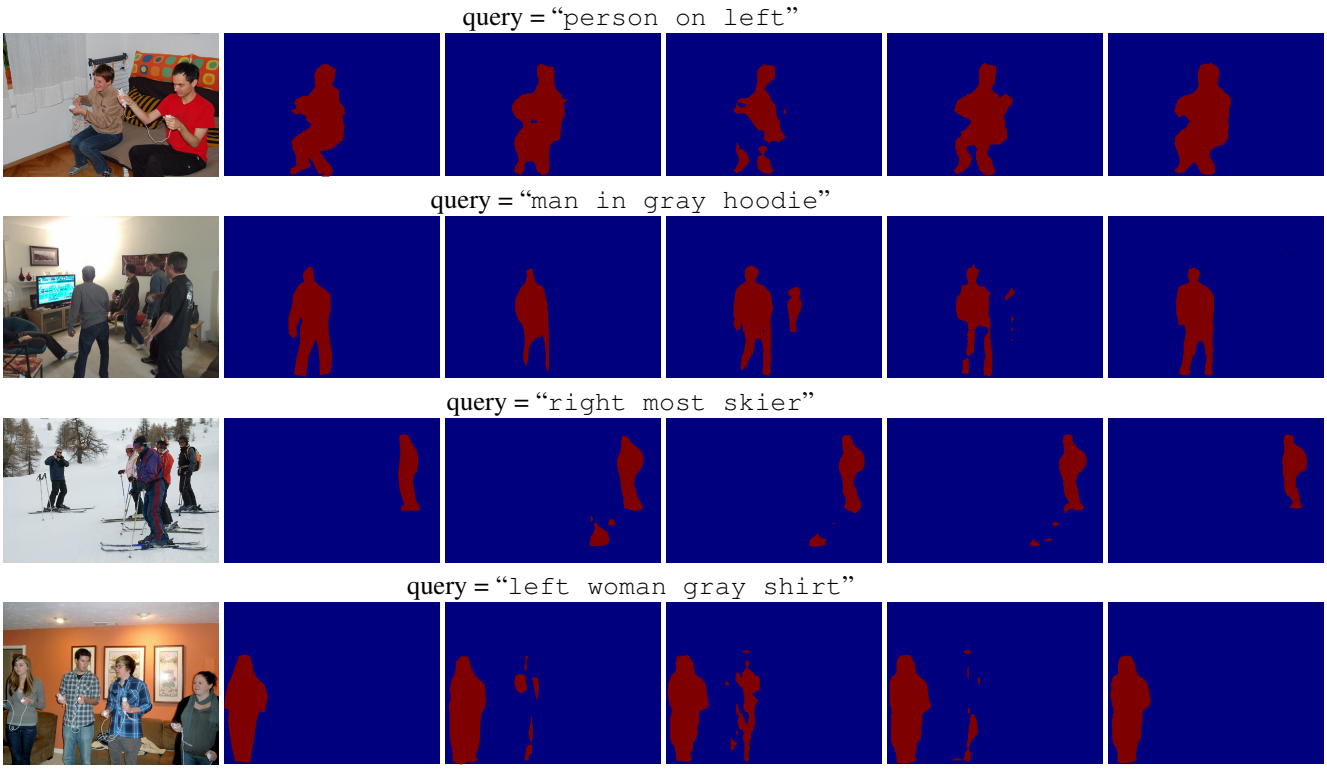

Figure 4. Segmentation results on UNC dataset. From left to right are input images, ground truth masks, results from baseline, plain structure, RNN, and LSTM respectively.

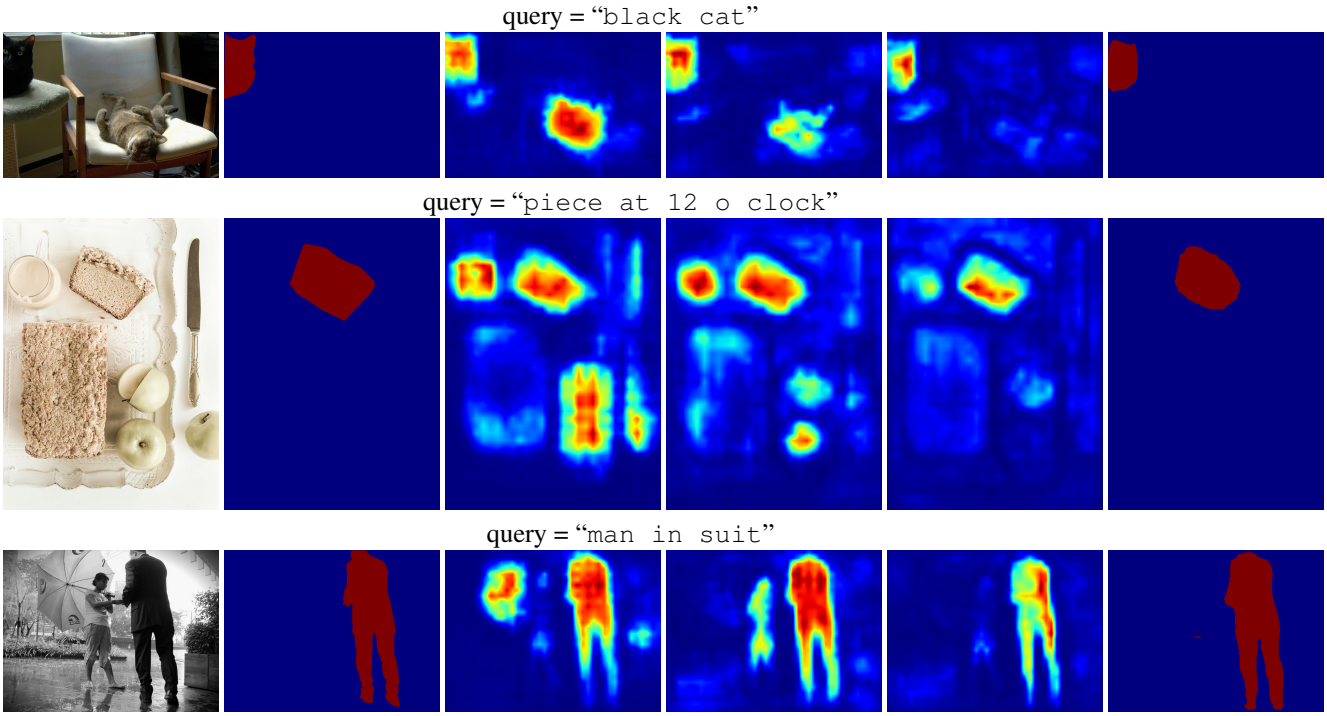

Figure 5. Visualization of convolutional LSTM on UNC+ dataset. From left to right are input images, ground truth masks, the strongest activated channel of hidden states after combining C_5, C_4,C_3 features, and the predicted mask.

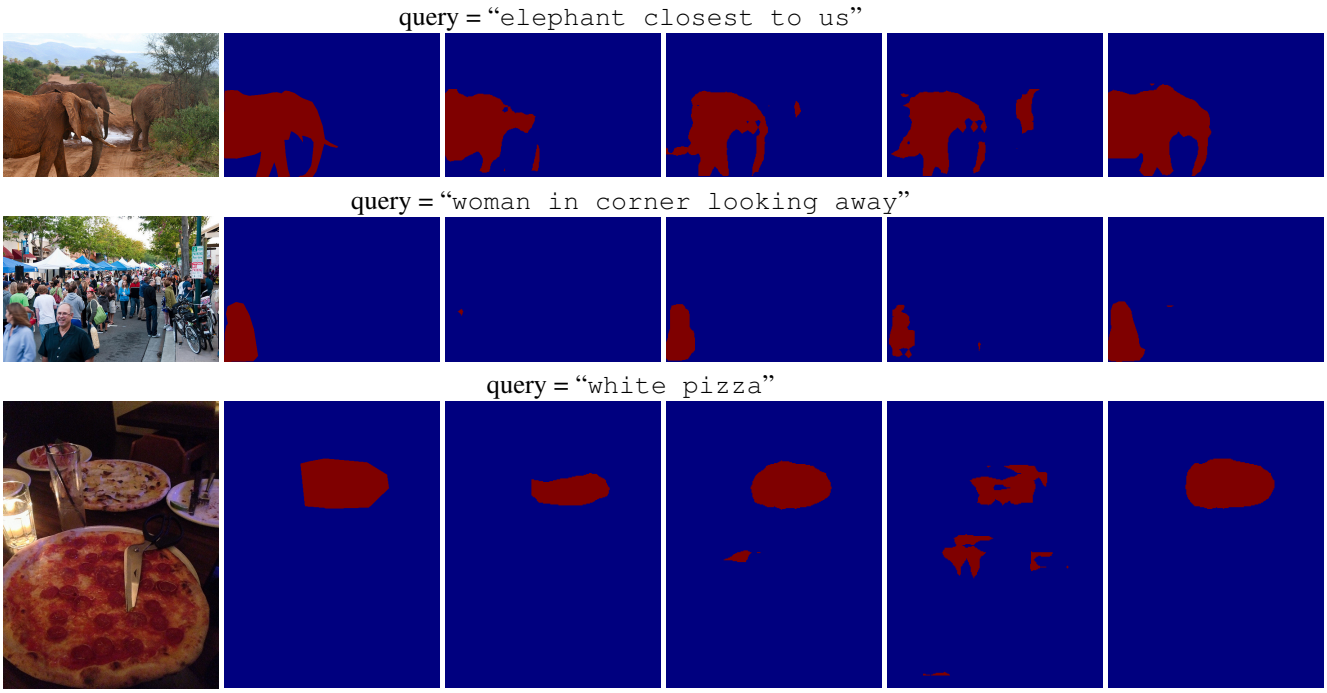

Figure 6. Segmentation results on UNC+ dataset. From left to right are input images, ground truth masks, results from baseline, plain structure, RNN, and LSTM respectively.

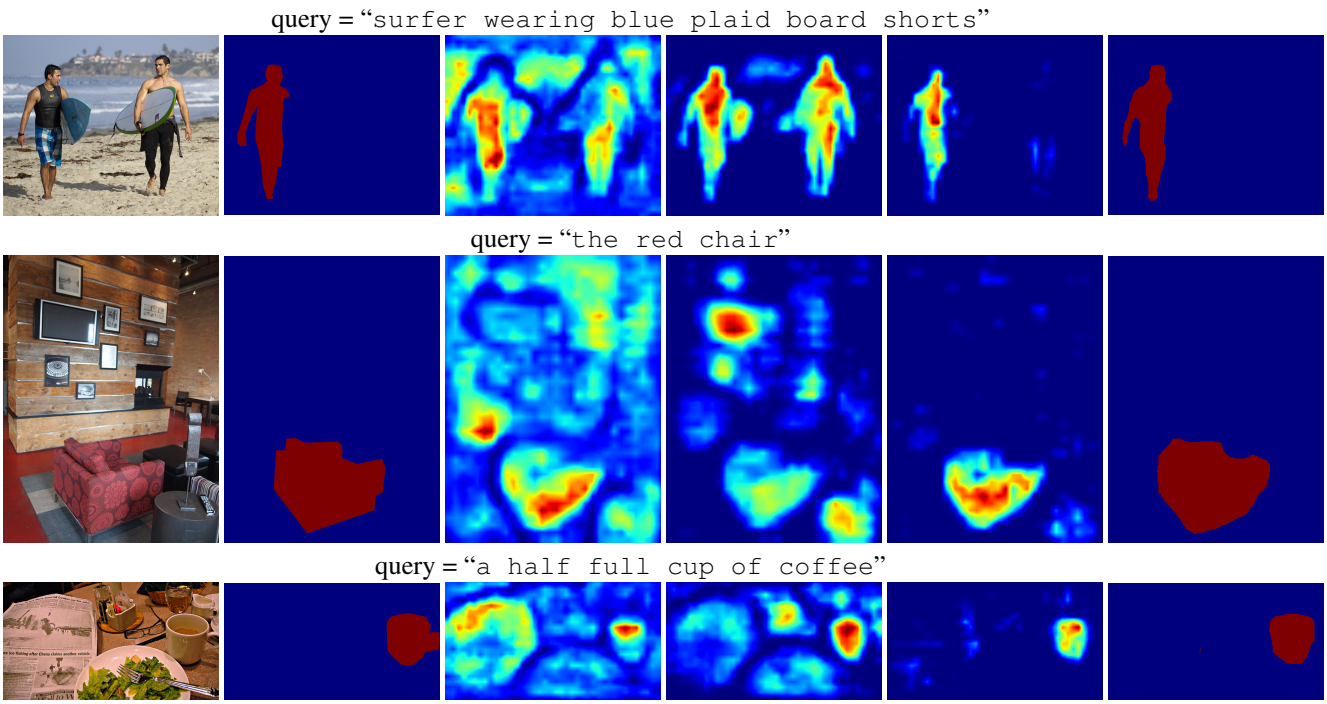

Figure 7. Visualization of convolutional LSTM on G-Ref dataset. From left to right are input images, ground truth masks, the strongest activated channel of hidden states after combining C_5, C_4,C_3 features, and the predicted mask.

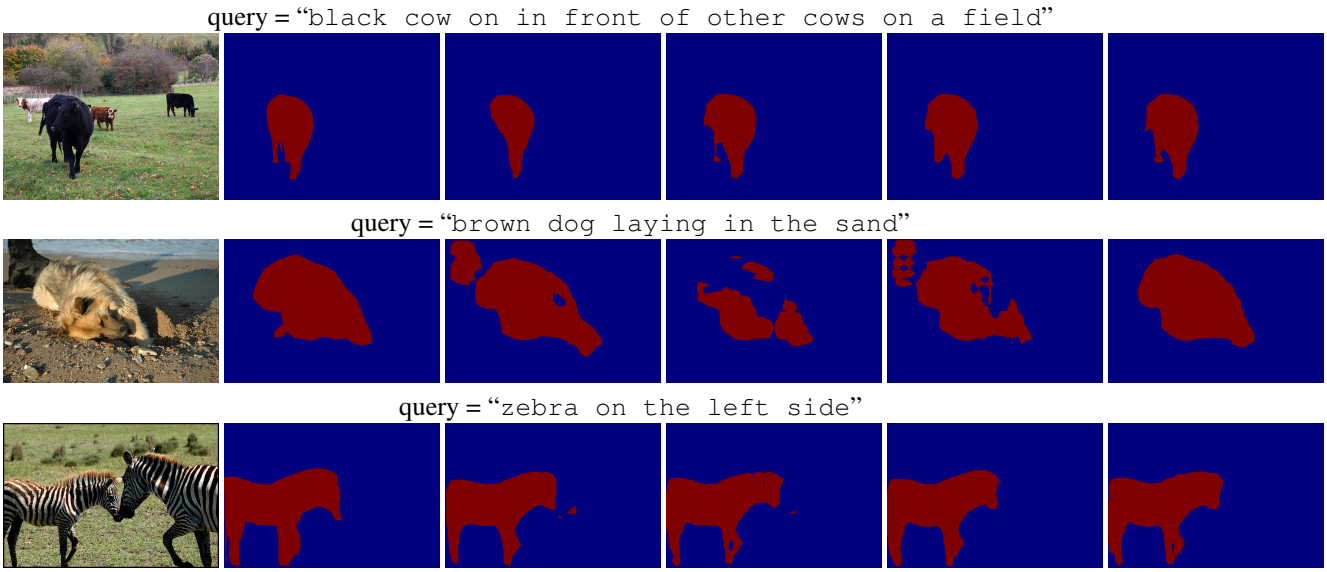

Figure 8. Segmentation results on G-Ref dataset. From left to right are input images, ground truth masks, results from baseline, plain structure, RNN, and LSTM respectively.