Learning from Pixel-Level Noisy Label : A New Perspective for Light Field Saliency Detection (Supplementary)

We provide more implementation detail and experimental results in this supplementary material. In detail, we provide more ConvLSTM module details in Section A, we present more ablation studies on DUT-LF [10] dataset to thoroughly analyze our proposed modules in Section B, we report elicitation details of cross-scene noise penalty loss in Section C, and we present additional qualitative and quantitative comparisons in Section D.

A. More Architecture Details

To further illustrate the proposed network, we show more details of the ConvLSTM module [12]. The features extracted from focal slices and all-focus central view image in m-th layers (denoted as g_m) are fed into a ConvLSTM structure in our architecture to gradually refine the abundant information for accurately identifying the salient objects. The procedure is defined as:

$$\begin{aligned}
i_t &= \sigma(w_{xi} * g_m + w_{hi} * H_{t-1} + w_{ci} \otimes C_{t-1} + b_i)\\
f_t &= \sigma(w_{xf} * g_m + w_{hf} * H_{t-1} + w_{cf} \otimes C_{t-1} + b_f)\\
C_t &= f_t \otimes C_{t-1} + i_t \otimes \tanh(w_{xc} * g_m + w_{hc} * H_{t-1} + b_c)\\
o_t &= \sigma(w_{xo} * g_m + w_{ho} * H_{t-1} + w_{co} \otimes C_t + b_o)\\
H_t &= o_t \otimes \tanh(C_t)
\end{aligned} \tag{1}$$

where $\otimes$ denotes pixel-wise multiplication and $\sigma(\cdot)$ is softmax function. Memory cell C_t stores previous information. All $*$, w_* and b_* represent convolution operator and convolution parameters to be learned. The memory cell C_t, the gates i_t, f_t, o_t and hidden state H_t are 3D tensors.

As shown in Fig.2 of our main paper, the weighted focal slices features for m layers are regarded as a sequence of inputs corresponding to consecutive time steps, feeding into ConvLSTM modules to gradually refine their spatial information (Fig. 1(a)). Then, the updated features F'_m and R'_m are further input to ConvLSTM modules to summarize information (Fig. 1(b) and Fig. 1(c)).

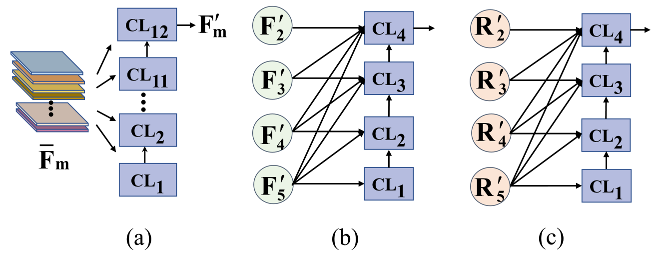

Figure 1: ConvLSTM modules (denoted as CL_m) used in our framework to process the weighted focal slices $\bar{F}_m$ (a), the focal slices features F'_m (b) and all-focus central view image features R'_m (c).

B. Ablation Study

To explore the optimal hyperparameters in the proposed cross-scene noise penalty loss and evaluate influence of different supervision information, we conduct additional ablation studies reported as following.

B.1. Hyperparameters of loss function

In this part, we present parameters details of cross-scene noise penalty loss $\mathcal{L}_t$. Based on m_l pairs of cross-scene samples, $\mathcal{L}_t$ for pixel (u, v) is defined as:

$$\mathcal{L}_t(s_i^{(u,v)}, \hat{y}_i^{(u,v)}) = \mathcal{L}(s_i^{(u,v)}, \hat{y}_i^{(u,v)}) - \frac{\alpha}{m_l - 1} \sum_{n,n'=2}^{m_l} (l(s_{i_n}^{(u,v)}, \hat{y}_{i_{n'}}^{(u,v)})), \tag{2}$$

The second term is linearly combined with two hyperparameters α and m_l. For α, we define a dynamic hyperparameter recursive process as $\alpha_{t+1} = \alpha_t + c/m_t$, where c is the maximal value of α and m_t denotes the maximal number of training iterations for each sample ($m_t = 30$ in our experiments). We report the affect of various value of c on saliency detection performance in Table 1. It can be seen that performance achieves optimal when $c = 0.30$.

c	0	0.10	0.15	0.20	0.25	0.30	0.35	0.40
$F\uparrow$	0.751	0.766	0.789	0.794	0.811	**0.813**	0.802	0.793
$M\downarrow$	0.168	0.144	0.112	0.114	0.102	**0.091**	0.052	0.026

Table 1: Experimental results of our model trained with different settings of the hyperparameter c (the maximal value of α) in Eq.(1) on saliency detection performance while keeping other settings unchanged.

In addition, we conduct experiments on our framework keeping other settings unchanged with different m_l values from 2 to 5. As shown in Table 2, we can see that more samples can consistently boost the performance. However, the increasing number of samples will consume a lot of computing resources. We set $m_l = 4$ to achieve the balance of performance and training time in our experiment.

m_l	2	3	4	5
$F\uparrow$	0.797	0.801	0.813	0.815
$M\downarrow$	0.096	0.090	0.091	0.090

Table 2: Model performances with regard to different number of correlation samples in Eq.(1) while keeping other settings unchanged.

Settings	Metrics	DUT-LF	HUFT	LFSD
BaselineDSR	$F\uparrow$	0.641	0.485	0.594
	$M\downarrow$	0172	0.253	0.191
DSR	$F\uparrow$	0.812	0.640	0.812
	$M\downarrow$	0.104	0.121	0.094
BaselineGT	$F\uparrow$	0.842	0.762	0.835
	$M\downarrow$	0.012	0.084	0.066
GT	$F\uparrow$	0.851	0.760	0.844
	$M\downarrow$	0.007	0.010	0.047

Table 3: Results on the generalization capability of our proposed method.

B.2. Learn from different noisy label generator

In the main paper, we use conventional unsupervised method RBD [20] to generate noisy labels. We further conduct an experiment with noisy labels generated by conventional method DSR [4]. We build a concise baseline that only contains separate features extraction branches for focal slices and all-focus central view image in the main paper. In this section, we treat the saliency maps generated by DSR method as supervision, and train the model using the setting of baseline. The results are reported in Table 3, we can find a huge gap between DSR and BaselineDSR, indicating the generalization capability of our method. Our method can handle noisy pixels even the noise generation mechanism is different. It is reasonable since that we correlate noisy pixels across whole dataset. In our experiment, we control the actual portion of noisy pixels by using different noisy labelling, instead adding noise to clean data.

B.3. Learn from ground truth labels

The ground truth can be treated as a special case of noisy label. We directly conduct experiments on ground truth using the setting of baseline. Then, we train our proposed model using ground truth (denoted as GT in Table 3). The performance improvement indicates the necessary of noise handling even training model on the ground truth.

C. Elicitation details of loss function

We define pixel (u, v) in noisy labels and ground truth as $\hat{Y}^{(u,v)}$ and $Y^{(u,v)}$, then the error rate can be denoted as:

$$\begin{aligned} e_{+1} &= p(\hat{Y}^{(u,v)} = -1 \mid Y^{(u,v)} = +1) \\ e_{-1} &= p(\hat{Y}^{(u,v)} = +1 \mid Y^{(u,v)} = -1) \end{aligned} \tag{3}$$

Similar to e_{+1} and e_{-1}, we define the error rates for noisy labels and the initial noisy saliency maps generated by our method:

$$\begin{aligned} p(\hat{y}^{(u,v)} = -1 \mid y^{(u,v)} = +1) = e_{+1}, \\ p(\hat{y}^{(u,v)} = +1 \mid y^{(u,v)} = -1) = e_{-1} \end{aligned} \tag{4}$$

$$\begin{aligned} p(s^{(u,v)} = -1 \mid y^{(u,v)} = +1) = e^*_{+1}, \\ p(s^{(u,v)} = +1 \mid y^{(u,v)} = -1) = e^*_{-1} \end{aligned} \tag{5}$$

where $\hat{y}^{(u,v)}$ and $y^{(u,v)}$ represent pixel (u, v) in noisy label and ground truth respectively.

Consider a binary segmentation case (salient object and background): $p(y^{(u,v)} = -1) = 0.4, p(y^{(u,v)} = +1) = 0.6$, the noise in the labels are $e_{-1} = 0.3, e_{+1} = 0.4$ and $e^*_{-1} = 0.2, e^*_{+1} = 0.3$.

Firstly, we compute the marginals of $s^{(u,v)}$ and $\hat{y}^{(u,v)}$:

$$\begin{aligned} &p(s^{(u,v)} = -1) \\ =&p(s^{(u,v)} = -1 \mid y^{(u,v)} = -1)p(y^{(u,v)} = -1) \\ +&p(s^{(u,v)} = -1 \mid y^{(u,v)} = +1)p(y^{(u,v)} = +1) \\ =&(1 - e^*_{-1}) \cdot 0.4 + e^*_{+1} \cdot 0.6 = 0.5, \end{aligned} \tag{6}$$

and easily

$$p(s^{(u,v)} = +1) = 1 - p(s^{(u,v)} = -1) = 0.5 \tag{7}$$

for noisy labels:

$$\begin{aligned} &p(\hat{y}^{(u,v)} = -1) \\ =&p(\hat{y}^{(u,v)} = -1 \mid y^{(u,v)} = -1)p(y^{(u,v)} = -1) \\ +&p(\hat{y}^{(u,v)} = -1 \mid y^{(u,v)} = +1)p(y^{(u,v)} = +1) \\ =&(1 - e_{-1}) \cdot 0.4 + e_{+1} \cdot 0.6 = 0.52 \end{aligned} \tag{8}$$

and

$$p(\hat{y}^{(u,v)} = +1) = 1 - p(\hat{y}^{(u,v)} = -1) = 0.48 \tag{9}$$

for the joint distribution,

$$\begin{aligned} &p(s^{(u,v)} = -1, y^{(u,v)} = -1) \\ =&p(s^{(u,v)} = -1, y^{(u,v)} = -1 \mid y^{(u,v)} = -1)p(y^{(u,v)} = -1) \\ +&p(s^{(u,v)} = -1, y^{(u,v)} = -1 \mid y^{(u,v)} = +1)p(y^{(u,v)} = +1) \\ =&(1 - e^*_{-1})(1 - e_{-1}) \cdot 0.4 + e^*_{+1} \cdot e_{+1} \cdot 0.6 = 0.296 \\ &p(s^{(u,v)} = -1, \hat{y}^{(u,v)}) = +1) \\ =&p(s^{(u,v)} = -1) - p(s^{(u,v)} = -1, y^{(u,v)} = -1) \\ =&0.264 \end{aligned} \tag{10}$$

further,

$$\begin{aligned} &p(s^{(u,v)} = +1, \hat{y}^{(u,v)} = -1) \\ =&p(\hat{y}^{(u,v)} = -1) - p(s^{(u,v)} = -1, y^{(u,v)} = -1) \\ =&0.224 \\ &p(s^{(u,v)} = +1, y^{(u,v)} = +1) \\ =&p(s^{(u,v)} = +1) - p(s^{(u,v)} = +1, y^{(u,v)} = -1) \\ =&0.216 \end{aligned} \tag{11}$$

Dataset	Metrics	Fully Supervised Models										Conventional Model		Noisy label Model			Ours
		RGB				RGB-D				Light field		RGB	Light field	RGB			
		C2S [5]	DSS [3]	DHS [6]	UCF [17]	CPFP [18]	DF [8]	PDNet [19]	CTMF [2]	DLLF [11]	Mo-LF [16]	DSR [4]	DILF [14]	SBF [13]	DUSPS [7]	MNL [15]	
DUT-LF	F ↑	0.791	0.728	0.801	0.769	0.730	0.733	0.763	0.790	0.868	0.843	0.645	0.641	0.583	0.736	0.716	0.813
	M ↓	0.084	0.128	0.090	0.107	0.101	0.151	0.111	0.100	0.070	0.052	0.164	0.168	0.135	0.062	0.086	0.091
HFUT	F ↑	0.618	0.606	0.542	0.596	0.594	0.531	0.608	0.620	0.863	0.627	0.518	0.529	–	0.705	–	0.652
	M ↓	0.112	0.138	0.129	0.144	0.096	0.156	0.112	0.103	0.093	0.095	0.153	0.148	–	0.087	–	0.108
LFSD	F ↑	0.749	0.644	0.761	0.748	0.524	0.750	0.780	0.791	–	0.819	0.631	0.728	–	0.795	–	0.804
	M ↓	0.113	0.190	0.133	0.169	0.186	0.162	0.116	0.119	–	0.089	0.208	0.168	–	0.105	–	0.111

Table 4: Additional quantitative comparisons between our method and competing methods on three light field datasets. ↑ & ↓ denote larger and smaller is better respectively.

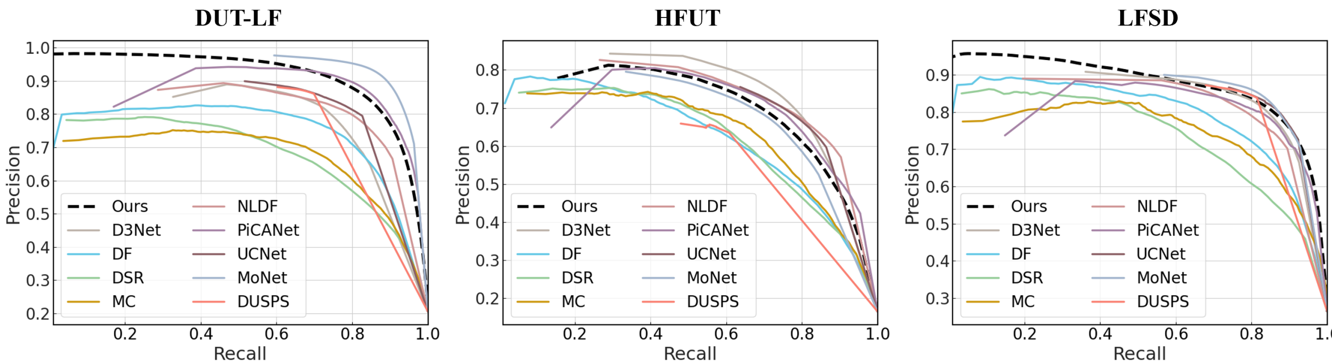

Figure 2: The PR curves of our method and other methods on three light field datasets, including fully supervised RGB, RGBD and light field models, conventional models and unsupervised RGB methods.

With above, the entries in $\Delta_{a,b}$ can be computed easily, for instance

$$\begin{aligned}\Delta_{1,1} =& p(s^{(u,v)} = -1, y^{(u,v)} = -1) \\ & - p(s^{(u,v)} = -1) \cdot p(y^{(u,v)} = -1) \\ =& 0.296 - 0.5 \cdot 0.52 = 0.036\end{aligned} \tag{12}$$

Then we have

$$\begin{bmatrix} 0.036 & -0.036 \\ -0.036 & 0.036 \end{bmatrix} \Rightarrow \mathrm{Sgn}(\Delta) \\ = \Omega(s^{(u,v)}, \hat{y}^{(u,v)}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{13}$$

The above implies that for $\Delta_{a,b}, \forall a, b = \{1, 2\}$, the marginal correlation is positive while for off-diagonal entries, they are negatively correlated.

CA [9] [1] requires each pixel in the predicted salient map to perform multiple tasks: compute the correlation with its corresponding noisy label and exploit the correlation between predictions of other scenes and unpaired noisy labels as the penalty to current scene. Ultimately the scoring function for each task, is defined as follows:

$$S(s_i^{(u,v)}, \hat{y}_i^{(u,v)}) = \Omega(s_i^{(u,v)}, \hat{y}_i^{(u,v)}) - \Omega(s_{i_1}^{(u,v)}, \hat{y}_{i_2}^{(u,v)}) \tag{14}$$

Next, we have the cross scene penalty loss defined as Eq.(2). The terms in cross scene penalty loss is simulating the marginal correlation probability in Δ.

D. Experiment Results

D.1. Quantitative Results

We extend the quantitative studies as a supplement to the main paper. Firstly, we present results of additional 10 methods in Table 4, including 4 fully supervised RGB methods (C2S [5], DSS [3], DHS [6], UCF [17]), 4 supervised RGB-D methods (CPFP [18], DP [8], PDNet [19], CTMF [2]) and 2 conventional unsupervised methods (DSR [4], DILF [14]). Results of competing methods are generated by authorized codes or directly provided by authors. Our model consistently achieves higher scores on all datasets across two evaluation metrics. Secondly, we compare our method with the state-of-the-art methods on three benchmark datasets and the PR curves are shown in Figure 2. Compared to the state-of-the-art fully supervised RGB and RGB-D methods, our method achieves significant advantages with a relatively small training set DUT-LF. It can be seen we still achieves competitive performance when compared with a number of fully supervised light field methods.

D.2. Qualitative Comparison

To further prove the superior of our method, we visualize results for our method and others. As shown in Figure 3, our results have a significant improvement for challenging scenes compared with fully supervised RGB, RGB-D methods and unsupervised RGB method. We still show a competitive performance compared with fully supervised light field method. With our proposed method and the abundant cues of light field data, our model has better noise invariant

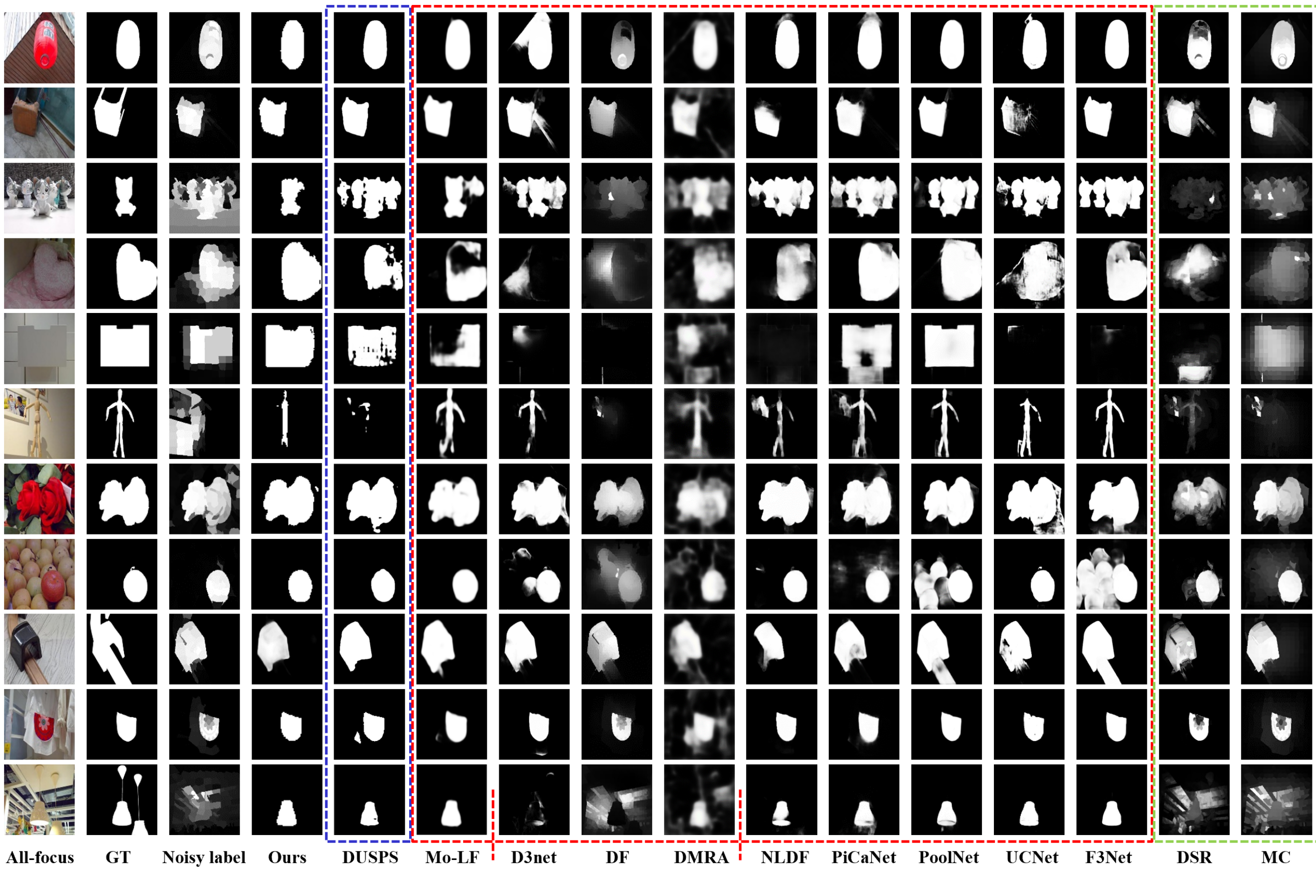

Figure 3: Additional qualitative comparisons between our method and others on DUT-LF [10]. The saliency maps in the blue box are predicted from noisy labels supervised RGB methods, the saliency maps in the red box are predicted from fully supervised light field, RGB-D and RGB saliency method respectively and the saliency maps in the green box are predicted from conventional models.

capability for salient object detection learning from pixel-level noisy labels.

References

[1] Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, pages 319–330, 2013. 3

[2] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. *IEEE transactions on cybernetics*, 48(11):3171–3183, 2017. 3

[3] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3203–3212, 2017. 3

[4] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE international conference on computer vision*, pages 2976–2983, 2013. 2, 3

[5] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 355–370, 2018. 3

[6] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 678–686, 2016. 3

[7] Duc Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: deep robust unsupervised saliency prediction with self-supervision. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 204–214, 2019. 3

[8] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5):2274–2285, 2017. 3

[9] Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C Parkes. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 179–196, 2016. 3

[10] Tiantian Wang, Yongri Piao, Xiao Li, Lihe Zhang, and Huchuan Lu. Deep learning for light field saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8838–8848, 2019. 1, 4

[11] Tiantian Wang, Yongri Piao, Xiao Li, Lihe Zhang, and Huchuan Lu. Deep learning for light field saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8838–8848, 2019. 3

[12] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015. 1

[13] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion:

Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4048–4056, 2017. 3
[14] Jun Zhang, Meng Wang, Jun Gao, Yi Wang, Xudong Zhang, and Xindong Wu. Saliency detection with a deeper investigation of light field. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015. 3
[15] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9029–9038, 2018. 3
[16] Miao Zhang, Jingjing Li, Wei Ji, Yongri Piao, and Huchuan Lu. Memory-oriented decoder for light field salient object detection. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 898–908, 2019. 3
[17] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *Proceedings of the IEEE International Conference on computer vision*, pages 212–221, 2017. 3
[18] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for rgbd salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3927–3936, 2019. 3
[19] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. Pdnet: Prior-model guided depth-enhanced network for salient object detection. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 199–204. IEEE, 2019. 3
[20] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2814–2821, 2014. 2