Exploit and Replace: An Asymmetrical Two-Stream Architecture for Versatile Light Field Saliency Detection

Yongri Piao,[1] **Zhengkun Rong,**[1] **Miao Zhang,**[2,3*] **Huchuan Lu**[1,4]
[1]School of Information and Communication Engineering, Dalian University of Technology, China
[2]International School of Information and Software Engineering, Dalian University of Technology, China
[3]Key Lab for Ubiquitous Network and Service Software of Liaoning Province, Dalian University of Technology, China
[4]Peng Cheng Laboratory
{yrpiao, miaozhang, lhchuan}@dlut.edu.cn, rzk911113@mail.dlut.edu.cn

Abstract

Light field saliency detection is becoming of increasing interest in recent years due to the significant improvements in challenging scenes by using abundant light field cues. However, high dimension of light field data poses computation-intensive and memory-intensive challenges, and light field data access is far less ubiquitous as RGB data. These may severely impede practical applications of light field saliency detection. In this paper, we introduce an asymmetrical two-stream architecture inspired by knowledge distillation to confront these challenges. **First**, we design a teacher network to learn to exploit focal slices for higher requirements on desktop computers and meanwhile transfer comprehensive focusness knowledge to the student network. Our teacher network is achieved relying on two tailor-made modules, namely multi-focusness recruiting module (MFRM) and multi-focusness screening module (MFSM), respectively. **Second**, we propose two distillation schemes to train a student network towards memory and computation efficiency while ensuring the performance. The proposed distillation schemes ensure better absorption of focusness knowledge and enable the student to replace the focal slices with a single RGB image in an user-friendly way. We conduct the experiments on three benchmark datasets and demonstrate that our teacher network achieves state-of-the-arts performance and student network (ResNet18) achieves Top-1 accuracies on HFUT-LFSD dataset and Top-4 on DUT-LFSD, which tremendously minimizes the model size by 56% and boosts the Frame Per Second (FPS) by 159%, compared with the best performing method.

Introduction

Human attentional mechanism (HAM) allows us to focus on interesting regions and filter out irrelevant ones. This cognitive ability helps us quickly understand visual scenes out of an overwhelming amount of information. Over the past decades, many works devote to imitating HAM. This task, namely saliency detection, is essential for progress in image understanding and has shown great potential in various computer vision and image processing tasks, such as image segmentation (Li et al. 2014b), visual tracking (Hong et al. 2015; Smeulders et al. 2013), object recognition (Ren et al. 2015; Dai et al. 2016) and robot navigation (Craye, Filliat, and Goudou 2016).

The existing saliency detection methods can be roughly divided into three categories based on the 2D (RGB), 3D (RGB-D) and 4D (light field) input images. Different from 2D and 3D saliency detection methods, 4D methods exploit the light field data. The light field provides multi-view images of the scene through an array of lenslets and produces a stack of focal slices, containing abundant spatial parallax information as well as accurate depth information. Moreover, the stack of focal slices cater to human visual perception and are observed in sequence with a combination of eye movements and shifts in visual attention. Such abundant 4D data provides abundant saliency cues for saliency detection in challenging scenes such as similar foreground and background, small salient objects and complex background (Li et al. 2014a; Li, Sun, and Yu 2015; Piao et al. 2019).

However, light field saliency detection has significant drawbacks. **1)** Light field methods are both computation-intensive and memory-intensive as high dimensional data are employed, e.g., the model size of the first deep-learning based light field saliency detection network is more than 119 MB and FPS is only 2 on a single 1080Ti GPU card (Piao et al. 2019). **2)** Taking light field data as input is seemingly inconvenient because data taken by light field cameras is far less ubiquitous as RGB data taken by traditional digital cameras. In consequence, it is reasonably essential to design a versatile, efficient and user-friendly mechanism to address those issues while ensuring the performance.

In this paper, we propose a novel learning strategy leveraging the concept of knowledge distillation (Hinton, Vinyals, and Dean 2015) and learning under an asymmetrical two-stream network architecture, to confront those challenges (see Figure 1). **First**, we consider the Focal stream, operating on focal slices, as a teacher network. Our goal is to design a deep network not only to achieve superior performance for higher requirements on desktop computers but also to transfer comprehensive knowledge for the student network. Given two phrases—recruiting and screening that the eyes process all information in our visual field (Recruit-

*Contact Author

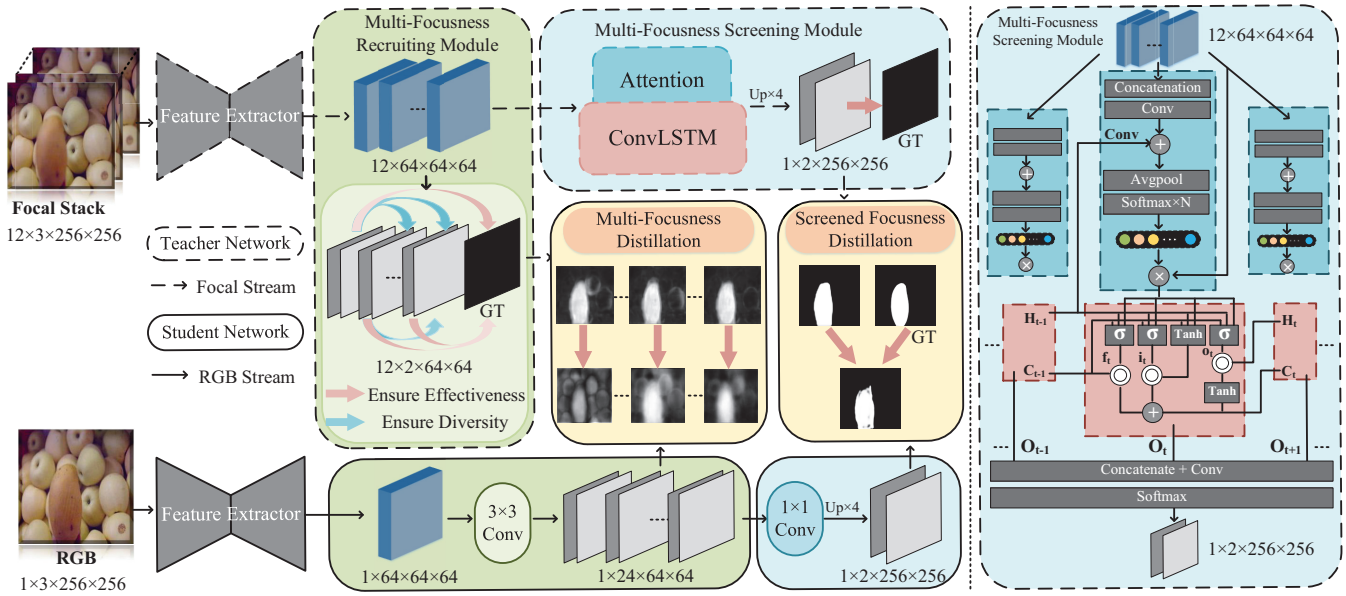

Figure 1: The whole pipeline.

ing is the act of gathering visual resources. Screening is the act of using these resources to select aspects of visual information), we propose two modules to meet the needs for the teacher network, namely *multi-focusness recruiting module (MFRM)* and *multi-focusness screening module (MFSM)*. The *MFRM* is designed to recruit rich saliency features from every single focal slice for ensuring both effectiveness and diversity, while the *MFSM* is designed to screen useful features by scanning various locations and emphasizing the most relevant ones. In the meanwhile, the *MFRM* and *MFSM* enable duo-transfer of knowledge tailored for the students network. **Second**, we aim to learn a lightweight and user-friendly student network, which highlights the challenges. Directly transferring the output of the teacher to the student overlooks the inherent differences between a RGB image and focal slices. We ably propose the *multi-focusness distillation* by encouraging multi-focusness consistencies between the Focal stream and RGB stream, and *screened focusness distillation* schemes by learning complementarity between the screened focusness knowledge from the Focal stream and appearance information from the RGB stream. The proposed distillation schemes ensure better absorption of focusness knowledge and enable the student to replace the focal slices with a single RGB image. **Last but not the least**, we demonstrate the effectiveness of the proposed framework on three light field datasets: DUT-LFSD (Zhang et al. 2019), HFUT-LFSD (Zhang et al. 2015), LFSD (Li et al. 2014a). Our teacher network achieves state-of-the-art results on three datasets, and student network (Resnet18) achieves Top-1 accuracies (MAE) on HFUT-LFSD dataset and Top-4 on DUT-LFSD. The student minimizes the model size by 56% and boosts the Frame Per Second (FPS) by 159%, compared with the best performing method. An important observation should be noted: our model achieves superior generalization on the training set (1100 samples) one order of magnitude smaller than the RGB training set (10553 samples). The source code can be found at **https://github.com/OIPLab-DUT/AAAI2020-Exploit-and-Replace-Light-Field-Saliency/**.

Related Work

Saliency Detection. Early 2D saliency detection methods (Tu et al. 2016; Qin et al. 2015; Li et al. 2013) focus on exploiting low-level hand-crafted features, such as color, region contrast, etc. With the development of convolutional neural networks, many new 2D methods based on CNNs are proposed. (Zhang et al. 2018) propose a multi-path recurrent network embedded with spatial and channel-wise attention mechanisms. (Liu et al. 2019) produce detail enriched saliency maps by two pooling-based modules which can progressively exploit the high-level features. (Wu, Su, and Huang 2019) come up with a cascaded partial decoder which can improve both efficiency and accuracy of the existing multi-level feature aggregation networks. (Zhao et al. 2019b) propose an edge-guided FCN to preserve good boundaries of salient objects by embedding edge prior knowledge into multi-level features.

In 3D saliency detection, depth is utilized to exploit geometric information for saliency detection. (Chen and Li 2018) propose a complementarity-aware fusion module which can learn complementary information from the paired modality. (Chen, Li, and Su 2019) utilize cross-modal interactions to encourage complements across both high-level and low-level features. (Chen and Li 2019) propose a three-stream network with an attention-aware mechanism which can adaptively select complementary features. (Zhao et al. 2019a) propose to enhance the depth map by contrast prior and use the enhanced depth information as an attention map

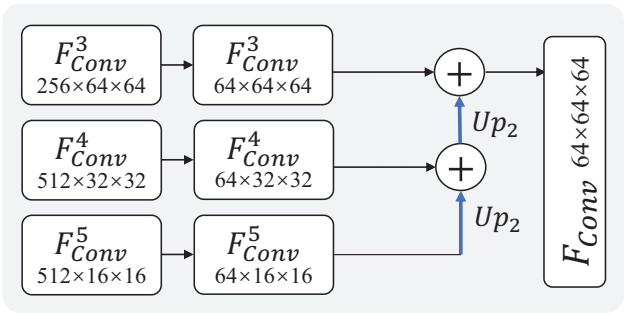

Figure 2: Detailed structure of the feature coder in the focal stream or RGB stream.

to work with RGB features for high-quality predictions.

Recently, a small number of works use light field information for saliency detection. (Li et al. 2014a) introduce the first light field saliency detection dataset. (Zhang et al. 2015) develop the background priors encoded by light field focusness to enhance the saliency and reduce the background distraction. (Li, Sun, and Yu 2015) propose a weighted sparse coding framework which can process the heterogenous type of input data effectively. (Zhang et al. 2017a) integrate multi light field cues extracted from all-in-focus image, depth map, focal slices and multi-view images based on a random-search-based strategy. (Piao et al. 2019) propose the first CNN-based network for processing multi-view images. These works suggest that light field information plays an important role in saliency detection.

However, higher dimension of data poses computational challenges. This severely impedes practical applications of light field saliency detection. In contrast, we propose an asymmetrical two-stream network in which teacher network exploits focal slices for higher requirement on desktop computers, while student network takes a single RGB image as input to achieve computational efficiency on mobile devices.

Knowledge Distillation. Knowledge distillation (Hinton, Vinyals, and Dean 2015) is a deep network compression method in which a small network (student) is trained to mimic a pre-trained, larger model (teacher). The knowledge distillation scheme has been verified valid in many computer vision tasks, such as object detection (Li, Jin, and Yan 2017), pedestrian re-identification (Chen, Wang, and Zhang 2018) and semantic segmentation (He et al. 2019). (Li, Jin, and Yan 2017) propose to train very efficient CNNs-based detectors by mimicking convolutional features without the need of ImageNet pre-training. (Chen, Wang, and Zhang 2018) introduce cross sample similarities for model compression and acceleration. (He et al. 2019) propose a knowledge adaptation scheme in which the reinterpreted knowledge is easy to be learned for student network, and an affinity distillation module to help the student network capture long-term dependencies.

Unlike directly applying knowledge distillation to light field saliency detection, we propose two tailored distillation schemes to transfer focusness knowledge to the RGB stream, which provide the Focal stream with an effective alternative.

Method

In this paper, in order to develop a versatile, efficient and user-friendly architecture for light field saliency detection, we introduce an asymmetrical two-stream architecture based on knowledge distillation (see Figure 1). The Focal stream, served as the teacher network, aims to learn to exploit focal slices for higher requirement on desktop computers. On the other hand, the student network takes a single RGB image as input and aims to learn to replace focal slices for computational efficiency on mobile devices. The feature extractor in the teacher network is based on VGG19 (Simonyan and Zisserman 2015), while the student feature extractor is based on ResNet18 (He et al. 2016). We select the high-level convolutional features (F^3_{Conv}, F^4_{Conv} and F^5_{Conv}) to detect salient objects. The detailed structure of the feature extractor is shown in Figure 2. In the following, we discuss the reasoning Focal stream (Learning to Exploit Focal Stack), and RGB stream (Learning to Replace Focal Stack) in detail.

Learning to Exploit Focal Stack

We propose two tailored modules in the teacher network to learn to exploit light field data, which are multi-focusness recruiting module (MFRM) and multi-focusness screening module (MFSM). The MFRM focuses on explicitly recruiting saliency information from each focal slice, and the MFSM aims to select useful features and suppress the unnecessary ones. The MFRM and MFSM are used to enable our teacher network with more accurate prediction for higher requirements on desktop computers, as well as duotransfer of rich focusness knowledge to the student. We visualize the effect of MFRM and MFSM, shown in Figure 3 and Figure 4, respectively. Detailed discussions about MFRM and MFSM are provided in ablation studies. Next, we elaborate each component in the teacher network.

Multi-Focusness Recruiting Module (MFRM). Inspired by the recruiting phrase in the process of visual attention, we aim to gather rich saliency features by processing every single focal slice. A simple method is to supervise the raw features with ground truth for avoiding the distraction of non-salient objects and ensuring the effectiveness of each focusness features. However, this strategy could reduce diversity and complementarity between multi-focusness features. To this end, we propose a multi-focusness recruiting module (MFRM) which encourages the raw multi-focusness features containing both adequate effectiveness and diversity to achieve optimal results. The MFRM is shown in Figure 1. We first connect a convolutional layer to convert each focusness feature from 64 channels to 2 channels. Then each focusness feature is supervised by following loss function:

$$L_R(f_k) = L_{CE}(f_k, Y) - \lambda \sum_{i=1, i\neq k}^{N} L_{MSE}(f_k, f_i), \quad (1)$$

where f_k is the k^{th} focusness feature, Y is the ground truth and N is the total number of focal slices. L_{CE} and L_{MSE} represent cross-entropy and mean squared error loss functions, respectively. The first item encourages effectiveness,

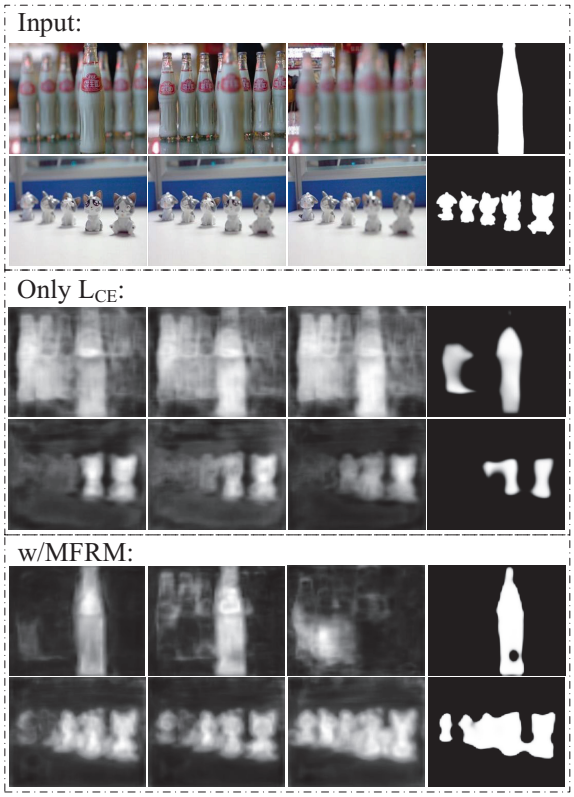

Figure 3: Visual comparisons in MFRM. The 1^{st} to 3^{rd} columns show the focal slices and corresponding multi-focusness features. The 4^{th} column shows the ground truth and saliency maps.

the second item enhances diversity, and the non-negative weight λ which is set to 10 expresses the trade-off between these two items.

Multi-Focusness Screening Module (MFSM). Inspired by the the screening phrase in the process of visual attention, which concerns with selectivity, we aim to efficiently select useful saliency information from multi-focusness features. To do this, we propose a multi-focusness screening module (MFSM) to resemble the screening phrase of how human select information of interest from visual resources by assigning different weights to different focal slices regarding the salient objects. The MFSM consists of a ConvLSTM model with an attention mechanism as shown in Figure 1. The attention module aims to emphasize the useful features and suppress the unnecessary ones to produce a screened features. The ConvLSTM module aims to summarize the spatial information from the screened features of historical steps and current step for accurately identifying the salient objects. As the time step increases, the MFSM can highlight the salient regions and block the non-salient ones gradually (see Figure 4). Detailed operations are expressed as follows.

In each time step t, the multi-focusness features $f = \{f_1, f_2, \cdots, f_N\}$ first go through a feature-wise attention

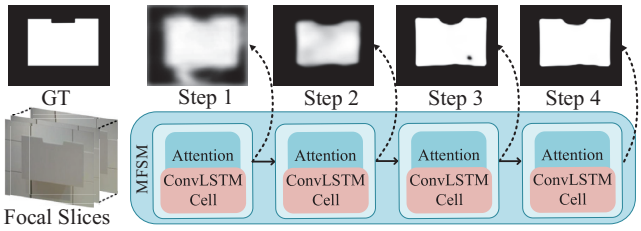

Figure 4: Visual comparisons of different steps in MFSM.

module and this procedure can be defined as:

$$\tilde{F}_t = \sum_{i=1}^{N} f \odot (\Phi(AvgPool(Cat\,[f_1; f_2; \cdots; f_N] * W_f + H_{t-1} * W_h))), \tag{2}$$

where Φ represents softmax function. $\odot$ and $*$ mean feature-wise multiplication and convolution operation, respectively. H_{t-1} represents the hidden state of the ConvLSTM cell in the $(t-1)^{th}$ step. The W_f and W_h are the parameters of the convolutional kernels. Then the combined features $\tilde{F}_t$ are fed into a ConvLSTM cell. The internal operations in ConvLSTM can be shown in Figure 1. After several time steps, we concatenate the O gates to summarize the saliency information from screened features, and make a final prediction. This procedure can be defined as:

$$S_{tea} = \Phi\left(W_s * Cat[O_1; \cdots, O_t; \cdots; O_T]\right), \tag{3}$$

where S_{tea} is the saliency map predicted from the teacher network and T represents the total time steps and is set to 4. W_s denotes the convolution parameter.

Learning to Replace Focal Stack

The most existing methods based on knowledge distillation take as same input for the teacher and student networks. Considering heavy focal computation and convenient access to focal slices, we aim to design a lightweight network in an user-friendly way by taking a single ubiquitous RGB image as input to replace focal slices. However, directly transferring the output from the teacher to the student overlooks the inherent differences between two types of data. Therefore, we propose two tailored distillation schemes to replace focal slices with a single RGB image by transferring the focusness knowledge. The focusness knowledge is defined as two parts: (1) The first part is designed to mimic multi-focusness features only using a single RGB image. (2) The second part is designed to learn complementary information from appearance and screened focusness knowledge. We show the visual effect of the proposed two distillation schemes in Figure 5, and give an in-depth discussion in ablation studies. Detailed description for each distillation scheme is given below.

Multi-Focusness Distillation Scheme (MFD). Unlike directly enforcing the student to mimic the output from teacher, the student is first trained to hallucinate multi-focusness features from the Focal stream by our proposed multi-focusness distillation scheme. This is mainly driven by the consideration of the inherent differences of input data

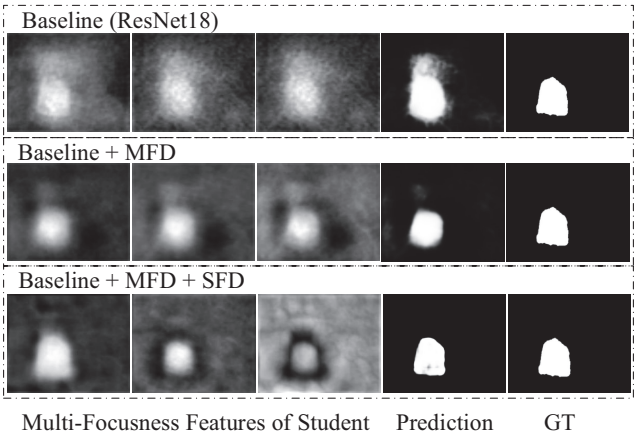

Figure 5: Visual results of enabling and disabling different components of our distillation system.

for the teacher and student. Moreover, multi-focusness features can be produced from a single RGB input without explicit focal computation. This leads to significantly faster inference. In detail, we reduce the Kullback-Leibler divergence loss between features of the penultimate layer in RGB stream, with features generated from the MFRM in the Focal stream:

$$L_{MFD}=\frac{1}{N}\sum_{i=1}^{N} KL\left(f_i^s \,\|\, f_i^t\right), \tag{4}$$

where f_i^t represents the feature map of the i^{th} focal slice produced from the teacher network and f_i^s represents the i^{th} feature map produced from the student network.

Screened Focusness Distillation Scheme (SFD). Our second distillation scheme goes a step further: we align the class probability of each pixel produced from the student network and teacher network, as well as the probability of each pixel between the output of the student network and the ground truth. We refer this distillation scheme as screened focusness distillation. This scheme allows the student network to learn complementary information from appearance and screened focusness information for accurate prediction. To enhance this process with screened focusness and appearance information, we train the student network by backpropagating a linear combination of KL and cross entropy losses through the entire network. The loss function is given as follows:

$$L_{SFD}=KL\left(S_{stu} \,\|\, S_{tea}\right)+\alpha L_{CE}(S_{stu}, Y), \tag{5}$$

where S_{tea} and S_{stu} represent the saliency map predicted from the teacher and student networks, respectively. The hyperparameter α is set to 1.

Training Process

As presented in Algorithm 1, the teacher network is supervised by two losses: the cross entropy loss L_{CE} with the ground truth and the recruiting loss $L_R(f_k)$ in Eq.(1). During the knowledge distillation process, the teacher is pretrained and the parameters are kept frozen. The student is supervised by a combination of the multi-focusness distillation loss L_{MFD} in Eq.(4) and the screened focusness distillation loss L_{SFD} in Eq.(5). W_T and W_S are parameters for the teacher and student, respectively.

Algorithm 1: Traning Process of Our Method

Stage 1 : Training the teacher network.
Inputs : Focal slices.
$W_T = \mathrm{argmin}_{W_T}\left(L_{CE}\left(S_{tea}, Y\right)+\sum_{k=1}^{N} L_R\left(f_k\right)\right)$
Stage 2 : Training the student network.
Inputs : Single RGB.
$W_S = \mathrm{argmin}_{W_S}\left(L_{MFD}+L_{SFD}\right)$

Experiment

Experimental Setup

Benchmark Datasets. To evaluate the performance of our framework, we conduct experiments on three widely-used light field benchmark datasets. DUT-LFSD (Zhang et al. 2019) is the largest dataset which contains 1462 light field images. HFUT-LFSD (Zhang et al. 2015) and LFSD (Li et al. 2014a) datasets include 255 and 100 samples, respectively. Each light field consists of an all-in-focus image, 12 focal slices focused at different depths and a corresponding manually labeled ground truth.

For comparison, we randomly select 1000 samples from DUT-LFSD dataset and 100 samples from HFUT-LFSD dataset as the training set. The remaining samples and the LFSD dataset are used for testing. To avoid overfitting, we augment the training set by flipping, cropping and rotating.

Evaluation Metrics. We use newly-proposed S-measure (Fan et al. 2017) and E-measure (Fan et al. 2018), as well as generally-recognized weighted F-measure (Margolin, Zelnik-Manor, and Tal 2014), F-measure (Achanta et al. 2009) and Mean Absolute Error (MAE) as evaluation metrics for comparing the performance of models. The four evaluation metrics can provide comprehensive and reliable evaluation results and have been well explained in many literatures. Also we adopt model size and Frames Per Second (FPS) to evaluate the complexity of each method.

Implementation Details. We implement our method based on the Pytorch toolbox with one GTX 1080Ti GPU. We train both the teacher network and student network using the SGD optimization algorithm in which the momentum, weight decay and learning rate are set to 0.9, 0.0005, 1e-10, respectively. The hyperparameter temperature T is set to 20 in all distillation loss functions. The minibatch size is 1 and maximum iterations of teacher and student network are set to 500000 and 300000, respectively.

Ablation Studies

Effect of MFRM. To prove the effect of the MFRM in terms of the recruiting ability, we conduct visual comparisons (see Figure 3) for the multi-focusness features generated with simple supervision (noted as Only L_{CE}) and our

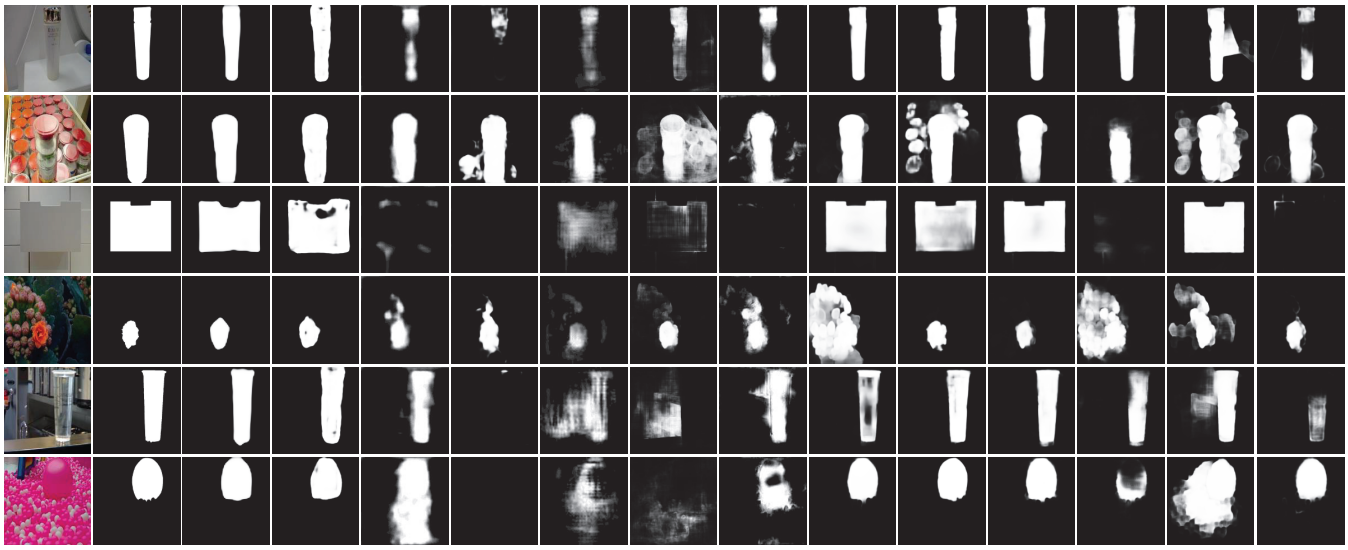

Figure 6: Visual comparisons of our method with top-ranking CNNs-based methods in some challenging scenes.

Table 1: Quantitative comparisons of E-measure, S-measure, weighted F-measure, F-measure and MAE scores on three light field datasets. $*$ represents conventional methods. - means no available results. (**boldface**: best, *italic*: second best, underline: third best, *underline*: fourth best).

Type	Methods	Years	DUT-LFSD					HFUT-LFSD					LFSD				
			$E_s\uparrow$	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	MAE$\downarrow$	$E_s\uparrow$	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	MAE$\downarrow$	$E_s\uparrow$	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	MAE$\downarrow$
4D	Teacher	-	**.943**	**.899**	**.880**	**.889**	**.040**	**.831**	*.777*	**.682**	**.705**	**.082**	**.889**	*.838*	**.809**	**.842**	**.080**
2D	Student	-	*.916*	*.848*	*.816*	*.838*	*.061*	**.824**	.736	.621	*.651*	**.085**	.820	.726	.672	.721	.137
4D	DLFS	IJCAI'19	.891	.841	.763	.801	.076	.783	.741	.590	.615	.098	.806	.737	.657	.715	.147
	LFS*	TPAMI'17	.728	.563	.288	.484	.240	.686	.579	.264	.430	.205	.771	.680	.479	.740	.208
	MCA*	TOOM'17	-	-	-	-	-	-	-	-	-	-	.841	.749	-	.815	.150
	WSC*	CVPR'15	-	-	-	-	-	-	-	-	-	-	.794	.706	.642	.706	.156
	DILF*	IJCAI'15	.805	.705	.494	.641	.168	.736	.695	.458	.555	.131	.810	.755	.604	.728	.168
3D	CPFP	CVPR'19	.808	.741	.634	.730	.101	.768	.701	.536	.594	.096	.669	.599	.465	.524	.186
	TANet	TIP'19	.861	.803	.702	.771	.096	.789	.744	.587	.638	.096	.849	.803	.727	.804	.112
	MMCI	PR'19	.853	.785	.629	.750	.116	.787	.741	.540	.645	.104	.848	.799	.685	.796	.128
	PCA	CVPR'18	.857	.800	.694	.762	.100	.782	.748	.598	.644	.095	.846	.807	.733	.801	.112
	PDNet	ICME'18	.864	.803	.655	.763	.111	.786	.770	.592	.629	.105	.849	.786	.728	.780	.116
	CTMF	Tcyb'17	.881	.823	.682	.790	.100	.784	.752	.544	.620	.103	.856	.801	.710	.791	.119
	DF	TIP'17	.838	.716	.523	.733	.151	.742	.670	.408	.562	.138	.816	.751	.607	.756	.162
2D	EGNet	ICCV'19	.914	.886	.829	.870	.053	.794	.772	.634	.672	.094	.776	.784	.717	.762	.118
	CPD	CVPR'19	.923	.890	.850	.887	.050	.810	.764	.652	.689	.097	.865	.846	.796	.841	.083
	PoolNet	CVPR'19	.919	.889	.832	.868	.051	.802	.776	.652	.683	.092	.786	.800	.717	.769	.118
	PiCANet	CVPR'18	.892	.829	.736	.821	.083	.726	.781	.556	.618	.115	.780	.729	.621	.671	.158
	PAGRN	CVPR'18	.878	.822	.733	.828	.084	.773	.717	.551	.635	.114	.805	.727	.642	.725	.147
	C2S	ECCV'18	.874	.844	.764	.791	.084	.786	.763	.630	.650	.111	.820	.806	.737	.749	.113
	R^3Net	IJCAI'18	.833	.819	.708	.783	.113	.728	.727	.566	.625	.151	.838	.789	.717	.781	.128
	Amulet	ICCV'17	.882	.847	.764	.805	.083	.760	.767	.616	.636	.110	.821	.773	.707	.757	.135
	UCF	ICCV'17	.850	.837	.708	.769	.107	.764	.754	.572	.623	.130	.776	.762	.655	.710	.169
	SRM	ICCV'17	.899	.848	.773	.832	.072	.801	.762	.623	.672	.096	.863	.826	.760	.827	.099
	NLDF	CVPR'17	.862	.786	.695	.778	.103	.807	.729	.590	.636	.091	.810	.745	.675	.748	.138
	DSS	CVPR'17	.827	.764	.624	.728	.128	.778	.715	.511	.626	.133	.749	.677	.570	.644	.190

Table 2: Quantitative results of the ablation analysis for our teacher network.

Module	Model	DUT-LFSD		HFUT-LFSD	
		$E_s \uparrow$	MAE$\downarrow$	$E_s \uparrow$	MAE$\downarrow$
MFRM	Only L_{CE}	.920	.066	.818	.090
	w/MFRM	**.943**	**.040**	**.831**	**.082**
MFSM	Step 1	.827	.218	.755	.271
	Step 2	.897	.091	.808	.100
	Step 3	.941	.041	.824	.084
	Ours (step 4)	**.943**	**.040**	**.831**	**.082**

proposed MFRM (noted as w/MFRM). We can observe that the multi-focusness features generated by simple supervision are almost the same and could lead to sub-optimal results, such as false positives (row 3) or incomplete detection of salient objects (row 4). In contrast, the MFRM encourages adequate diversity between features of different focal slices to achieve optimal results (row 5 and row 6). Numerically, our MFRM reduces the MAE performances by nearly 39% and 9% on two datasets as shown in Table 2.

Effect of MFSM. To give the evidence for the screening ability of the MFSM, we visualize the saliency maps in different time steps as shown in Figure 4. We can observe that the attention module contributes more on locating salient object accurately in step 1 and 2, while the ConvLSTM contributes more on refining the details of salient object in step 3 and 4. Also in Table 2, accumulative improvements are achieved as the time step increases. These improvements are resonable since the useful features are emphasized by attention module and spatial details are refined gradually by ConvLSTM.

Effect of MFD and SFD. To demonstrate the effectiveness of our distillation schemes, we analyze the performance in the absence of MFD and SFD. The experiments are conduct on three student networks, which are ResNet18, VGG16 and MobileNetV2 as shown in Table 3. It can be seen that as we add the MFD and SFD, the performance of student network achieves accumulative improvements by a large margin. Specifically, the MAE of ResNet18 gains 48% improvements on DUT-LFSD, and also 40% improvements on HFUT-LFSD.

The visual effects are shown in Figure 5. Compared to the baseline ResNet18 (row 1), the baseline with MFD can produce more effective multi-focusness features, such as features with accurate location of salient object (row 2, column 1), thus results in more accurate prediction (row 2, column 4). This validates the importance of transferring multi-focusness information from the Focal steam to the RGB stream. As we add the SFD, we observe a more boundary detailed prediction (row 3, column 4) and features with sharp boundary of salient object (row 3, column 3). This visual improvement is mainly due to the SFD which allows the student network to learn complementarity from appearance and screened focusness information.

Table 3: Ablation analysis of the proposed distillation schemes on different student networks.

Model	Size(M)	FPS	DUT-LFSD		HFUT-LFSD	
			$E_s \uparrow$	MAE$\downarrow$	$E_s \uparrow$	MAE$\downarrow$
ResNet18	49.1	171	.839	.119	.713	.142
+MFD	49.1	171	.869	.093	.737	.120
+MFD+SFD	49.1	171	**.916**	**.061**	**.824**	**.085**
VGG16	61.9	148	.879	.086	.765	.106
+MFD	61.9	148	.882	.074	.775	.093
+MFD+SFD	61.9	148	**.907**	**.067**	**.836**	**.079**
MobileNetV2Plus	27.5	65	.748	.250	.682	.239
+MFD	27.5	65	.764	.235	.692	.213
+MFD+SFD	27.5	65	**.811**	**.128**	**.726**	**.142**

Comparison with State-of-the-arts

We compare our method with 24 other state-of-the-arts ones including both deep-learning-based methods and conventional methods (remarked with $*$). There are **five** 4D light field methods: DLFS (Piao et al. 2019), LFS* (Li et al. 2014a), MCA* (Zhang et al. 2017a), WSC* (Li, Sun, and Yu 2015), DILF* (Zhang et al. 2015); **seven** 3D RGBD methods: CPFP (Zhao et al. 2019a), TANet (Chen and Li 2019), MMCI (Chen, Li, and Su 2019), PCA (Chen and Li 2018), PDNet (Zhu et al. 2018), CTMF (Han et al. 2017), DF (Qu et al. 2017); and **twelve** top-ranking RGB methods: EGNet (Zhao et al. 2019b), CPD (Wu, Su, and Huang 2019), PoolNet (Liu et al. 2019), PiCANet (Liu, Han, and Yang 2018), PAGRN (Zhang et al. 2018), C2S (Li et al. 2018), R^{3}Net (Deng et al. 2018), Amulet (Zhang et al. 2017b), UCF (Zhang et al. 2017c), SRM (Wang et al. 2017), NLDF (Luo et al. 2017), DSS (Hou et al. 2017). For fair comparison, the results from competing methods are generated by authorized codes or directly provided by authors.

Quantitative Evaluation. As shown in Table 1, the proposed teacher network can largely outperform other models across all the datasets in terms of five evaluation metrics, except second-best S-measure scores on HFUT-LFSD and LFSD datasets. It should be noted that our significant advantages are achieved on the training set (1100 samples) an order of magnitude smaller than the large RGB training set (10553 samples). Not only that, the proposed focusness knowledge distillation schemes can be seen as a good replacement for the Focal stream. This observation is further supported by the considerably good results of the student network (ResNet18), such as Top-1 accuracies (MAE) on HFUT-LFSD dataset and Top-4 on DUT-LFSD.

Qualitative Evaluation. Figure 6 provides some challenging samples of results comparing our method with other state-of-the-art methods. It can be seen that both the teacher and student network can achieve more complete and accurate prediction, when foreground and background are similar as shown in the 1^{st}, 2^{nd} and 3^{rd} rows, when salient object is small or transparent as shown in the 4^{th} and 5^{th}rows, when background is complex as shown in the 6^{th} row. It is worth noted that our student network can be positively influenced

Table 4: Complexity comparisons. The meaning of notation has been explained in Table 3.

Type	Methods	Years	Size(M)↓	FPS↑	DUT	HFUT	LFSD
					MAE↓	MAE↓	MAE↓
	Teacher	-	**92.5**	14	**.040**	**.082**	**.080**
	Student	-	**49.1**	**171**	*.061*	**.085**	.137
4D	DLFS	IJCAI'19	119	2	.076	.098	.147
3D	CPFP	CVPR'19	278	7	.101	.096	.186
	MMCI	PR'19	929.7	19	.116	.104	.128
	PCA	CVPR'18	533.6	15	.100	.095	.112
	PDNet	ICME'18	192	19	.111	.105	.116
	CTMF	Tcyb'17	825.8	50	.100	.103	.119
2D	EGNet	ICCV'19	412	21	.053	.094	.118
	CPD	CVPR'19	112	66	.050	.097	.083
	PoolNet	CVPR'19	278.5	32	.051	.092	.118
	PiCANet	CVPR'18	197.2	5	.083	.115	.158
	Amulet	ICCV'17	132.6	21	.083	.110	.135
	UCF	ICCV'17	117.9	23	.107	.130	.169
	SRM	ICCV'17	213.1	37	.072	.096	.099
	NLDF	CVPR'17	425.9	20	.103	.091	.138
	DSS	CVPR'17	447.3	23	.128	.133	.190

by the focusness knowledge transferred from teacher which leads to robust results in challenging scenes even with a single RGB input.

Complexity Evaluation. In Table 4, we compare the average execution time and model size with 15 representative models. We can see that the teacher network outperforms all other methods, and the student network achieves Top-4 accuracies (MAE) on DUT-LFSD and Top-1 accuracies on HFUT-LFSD. It is noted that the model size of our student network (ResNet18) is only 49.1 MB and FPS reaches up to 171. Compared to the best performing method CPD, our student network tremedously minimizes the model size by 56% and boosts the FPS by 159%.

Conclusion

In this paper, we develop a novel asymmetrical two-stream network architecture, which consists of Focal stream and RGB stream, to achieve versatility for both desktop computers and mobile devices. We consider the Focal stream as a teacher network, to learn to exploit focal slices and produce focusness knowledge tailored for student network. Our proposed MFRM and MFSM recruit and screen useful saliency information effectively and enable the teacher network to achieve superior performance. On the other hand, we train the student network, using single RGB input, to learn to replace focal slices relying on two tailor-made distillation schemes. The proposed distillation schemes allow the student to produce more effective multi-focusness features and learn complementarity between appearance and screened focusness information for accurate saliency prediction. Our extensive evaluation shows that the proposed asymmetrical network architecture can be applied on both PC and mobile terminals successfully.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (61976035), the Science and Technology Innovation Foundation of Dalian (2019J12GX034), and the Fundamental Research Funds for the Central Universities (DUT19JC58).

References

Achanta, R.; Hemami, S.; Estrada, F.; and Süsstrunk, S. 2009. Frequency-tuned salient region detection. In *CVPR*, number CONF, 1597–1604.

Chen, H., and Li, Y. 2018. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *CVPR*, 3051–3060.

Chen, H., and Li, Y. 2019. Three-stream attention-aware network for rgb-d salient object detection. *TIP* 28(6):2825–2835.

Chen, H.; Li, Y.; and Su, D. 2019. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. *Pattern Recognition* 86:376–385.

Chen, Y.; Wang, N.; and Zhang, Z. 2018. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In *AAAI*.

Craye, C.; Filliat, D.; and Goudou, J.-F. 2016. Environment exploration for object-based visual saliency learning.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. *International Conference on Neural Information Processing Systems (NIPS)* 379–387.

Deng, Z.; Hu, X.; Zhu, L.; Xu, X.; Qin, J.; Han, G.; and Heng, P.-A. 2018. R^3net: Recurrent residual refinement network for saliency detection. In *IJCAI*, 684–690.

Fan, D.; Cheng, M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 4558–4567.

Fan, D.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.; and Borji, A. 2018. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, 698–704.

Han, J.; Chen, H.; Liu, N.; Yan, C.; and Li, X. 2017. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. *IEEE Tcyb* PP:1–13.

He, K.; Zhang, X.; Ren, S.; and Jian, S. 2016. Deep residual learning for image recognition. In *CVPR*.

He, T.; Shen, C.; Tian, Z.; Gong, D.; Sun, C.; and Yan, Y. 2019. Knowledge adaptation for efficient semantic segmentation. In *CVPR*, 578–587.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hong, S.; You, T.; Kwak, S.; and Han, B. 2015. Online tracking by learning discriminative saliency map with convolutional neural network. *CVPR*.

Hou, Q.; Cheng, M.-M.; Hu, X.; Borji, A.; Tu, Z.; and Torr, P. H. 2017. Deeply supervised salient object detection with short connections. In *CVPR*, 5300–5309.

Li, X.; Lu, H.; Zhang, L.; Ruan, X.; and Yang, M.-H. 2013. Saliency detection via dense and sparse reconstruction. In *ICCV*, 2976–2983.

Li, N.; Ye, J.; Ji, Y.; Ling, H.; and Yu, J. 2014a. Saliency detection on light field. In *CVPR*, 2806–2813.

Li, Y.; Hou, X.; Koch, C.; Rehg, J. M.; and Yuille, A. L. 2014b. The secrets of salient object segmentation. 280–287.

Li, X.; Yang, F.; Cheng, H.; Liu, W.; and Shen, D. 2018. Contour knowledge transfer for salient object detection. In *ECCV*, 355–370.

Li, Q.; Jin, S.; and Yan, J. 2017. Mimicking very efficient network for object detection. In *CVPR*, 6356–6364.

Li, N.; Sun, B.; and Yu, J. 2015. A weighted sparse coding framework for saliency detection. In *CVPR*, 5216–5223.

Liu, J.-J.; Hou, Q.; Cheng, M.-M.; Feng, J.; and Jiang, J. 2019. A simple pooling-based design for real-time salient object detection. *CVPR*.

Liu, N.; Han, J.; and Yang, M.-H. 2018. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 3089–3098.

Luo, Z.; Mishra, A.; Achkar, A.; Eichel, J.; Li, S.; and Jodoin, P.-M. 2017. Non-local deep features for salient object detection. In *CVPR*, 6609–6617.

Margolin, R.; Zelnik-Manor, L.; and Tal, A. 2014. How to evaluate foreground maps? In *CVPR*, 248–255.

Piao, Y.; Rong, Z.; Zhang, M.; Li, X.; and Lu, H. 2019. Deep light-field-driven saliency detection from a single view. In *IJCAI*.

Qin, Y.; Lu, H.; Xu, Y.; and Wang, H. 2015. Saliency detection via cellular automata. In *CVPR*, 110–119.

Qu, L.; He, S.; Zhang, J.; Tian, J.; Tang, Y.; and Yang, Q. 2017. Rgbd salient object detection via deep fusion. *IEEE TIP* 26(5):2274–2285.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(6):1137–1149.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR*.

Smeulders, A.; Chu, D.; Cucchiara, R.; Calderara, S.; Dehghan, A.; and Shah, M. 2013. Visual tracking: An experimental survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tu, W.-C.; He, S.; Yang, Q.; and Chien, S.-Y. 2016. Real-time salient object detection with a minimum spanning tree. In *CVPR*, 2334–2342.

Wang, T.; Borji, A.; Zhang, L.; Zhang, P.; and Lu, H. 2017. A stagewise refinement model for detecting salient objects in images. In *ICCV*, 4019–4028.

Wu, Z.; Su, L.; and Huang, Q. 2019. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 3907–3916.

Zhang, J.; Wang, M.; Gao, J.; Wang, Y.; Zhang, X.; and Wu, X. 2015. Saliency detection with a deeper investigation of light field. In *IJCAI*, 2212–2218.

Zhang, J.; Wang, M.; Lin, L.; Yang, X.; Gao, J.; and Rui, Y. 2017a. Saliency detection on light field: A multi-cue approach. *ACM TOOM* 13(3):32.

Zhang, P.; Wang, D.; Lu, H.; Wang, H.; and Ruan, X. 2017b. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 202–211.

Zhang, P.; Wang, D.; Lu, H.; Wang, H.; and Yin, B. 2017c. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, 212–221.

Zhang, X.; Wang, T.; Qi, J.; Lu, H.; and Wang, G. 2018. Progressive attention guided recurrent network for salient object detection. In *CVPR*, 714–722.

Zhang, M.; Li, J.; Ji, W.; Piao, Y.; and Lu, H. 2019. Memory-oriented decoder for light field salient object detection. In *NeurIPS*.

Zhao, J.-X.; Cao, Y.; Fan, D.-P.; Cheng, M.-M.; Li, X.-Y.; and Zhang, L. 2019a. Contrast prior and fluid pyramid integration for rgbd salient object detection. In *CVPR*.

Zhao, J.-X.; Liu, J.-J.; Fan, D.-P.; Cao, Y.; Yang, J.; and Cheng, M.-M. 2019b. Egnet:edge guidance network for salient object detection. In *ICCV*.

Zhu, C.; Cai, X.; Huang, K.; Li, T. H.; and Li, G. 2018. Pdnet: Prior-model guided depth-enhanced network for salient object detection. *ICME* 199–204.