A2dele: Adaptive and Attentive Depth Distiller for Efficient RGB-D Salient Object Detection

Yongri Piao[1*] Zhengkun Rong[1*] Miao Zhang[1,2†] Weisong Ren[1] Huchuan Lu[1,3]
[1]Dalian University of Technology, China
[2]Key Lab for Ubiquitous Network and Service Software of Liaoning Province,
Dalian University of Technology, China
[3]Pengcheng Lab
{yrpiao, miaozhang, lhchuan}@dlut.edu.cn, {rzk911113, beatlescoco}@mail.dlut.edu.cn

Abstract

Existing state-of-the-art RGB-D salient object detection methods explore RGB-D data relying on a two-stream architecture, in which an independent subnetwork is required to process depth data. This inevitably incurs extra computational costs and memory consumption, and using depth data during testing may hinder the practical applications of RGB-D saliency detection. To tackle these two dilemmas, we propose a depth distiller (A2dele) to explore the way of using network prediction and attention as two bridges to transfer the depth knowledge from the depth stream to the RGB stream. First, by adaptively minimizing the differences between predictions generated from the depth stream and RGB stream, we realize the desired control of pixel-wise depth knowledge transferred to the RGB stream. Second, to transfer the localization knowledge to RGB features, we encourage consistencies between the dilated prediction of the depth stream and the attention map from the RGB stream. As a result, we achieve a lightweight architecture without use of depth data at test time by embedding our A2dele. Our extensive experimental evaluation on five benchmarks demonstrate that our RGB stream achieves state-of-the-art performance, which tremendously minimizes the model size by 76% and runs 12 times faster, compared with the best performing method. Furthermore, our A2dele can be applied to existing RGB-D networks to significantly improve their efficiency while maintaining performance (boosts FPS by nearly twice for DMRA and 3 times for CPFP).

1. Introduction

The emergence of convolutional neural networks (CNNs), together with larger datasets [31, 17, 30, 29] have recently led to remarkable progress in RGB-D salient object detection. In RGB-D methods, the depth information provides a preponderance of discriminative power in location and spatial structure, which plays an important role in the task of saliency detection [2]. Many pioneering works [31, 3, 5, 4, 41, 43] have demonstrated its effectiveness, especially in challenging scenes.

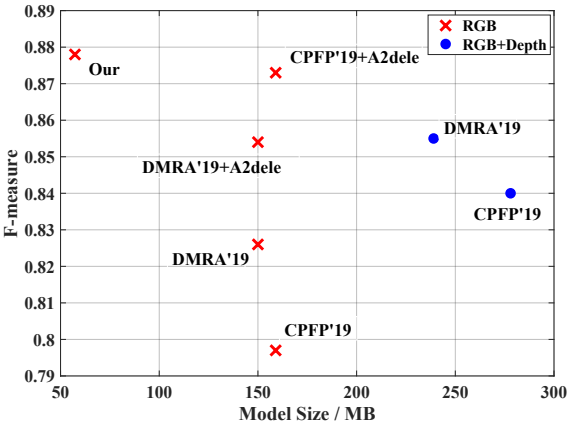

Figure 1. F-measure *vs.* Model Size on NLPR dataset [30]. By embedding our A2dele (CPFP'19 [41]+A2dele and DMRA'19 [31]+A2dele marked with ×), we achieve comparable accuracy compared to the original models (CPFP'19 and DMRA'19 marked with •) at a significantly smaller model size.

Learning discriminative representations for visual saliency, from two modalities, has been widely explored. For learning cross-model complementarity, RGB and depth data are often learnt separately in a two-stream architecture illustrated in Figure 2(a), where a multi-level fusion decoder is then appended to learn joint representations and cooperated predictions [31, 3, 5, 4]. On the other hand, approaches for learning enhanced RGB representations rely on exploring depth information by a tailor-maid subnetwork [41, 43], illustrated in Figure 2(b).

*Equal Contributions
†Corresponding Author

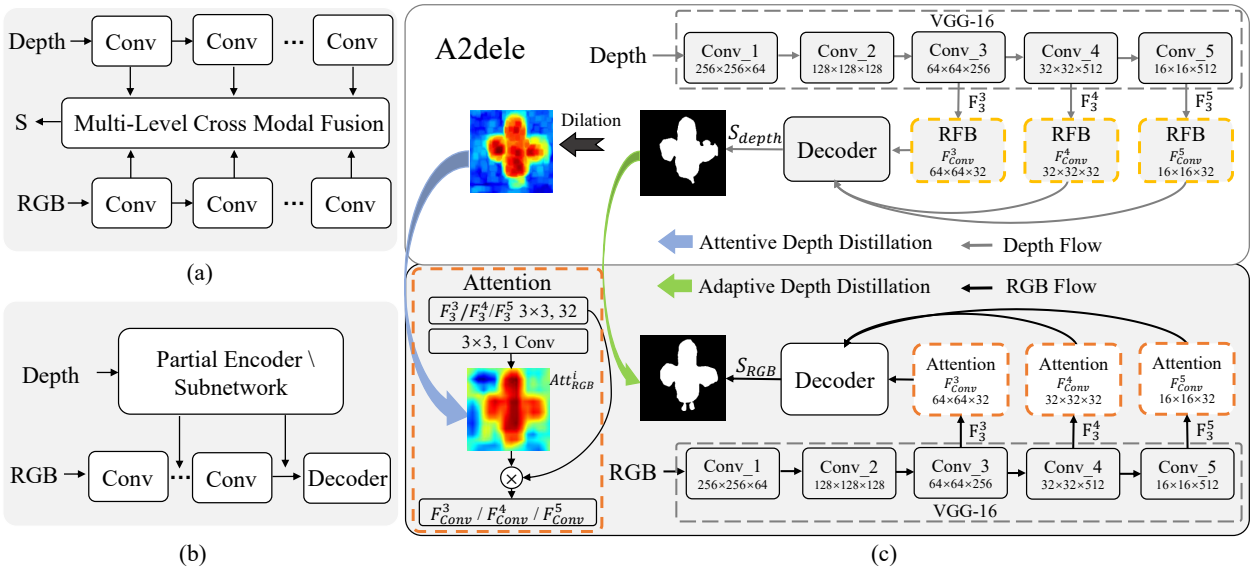

Figure 2. (a) Exploiting cross-modal complementarity by a two-stream architecture (e.g.[31, 3, 5, 4]). (b) Using depth information to enhance RGB features by a tailor-maid subnetwork (e.g.[41, 43]). (c) Our RGB stream embedded with the proposed depth distiller (A2dele). By embedding our A2dele, we achieve free use of depth stream during testing.

The strategy of leveraging RGB-D data and CNNs produces the impressive results, but it remains challenging in terms of two aspects. First, RGB-D approaches inevitably incur extra computational costs and memory consumption during inference of the two-stream model in which an independent encoder or subnetwork is required to process depth data, as shown in the F-measure vs. model size plot on the NLPR dataset [30] in Figure 1. We observe from the plot that the model size of the RGB-D networks is 1.5 larger than their RGB networks. Second, The use of depth information during testing may hinder the practical applications of RGB-D saliency detection. Despite the fact that the advent of consumer grade RGB-D cameras leaves open the possibility of opening a path towards a boarder application of 3D vision, depth sensors may pose a high risk to accurate saliency detection as they can be easily influenced by a number of factors, such as the temperature of the camera, background illumination, and distance and reflectivity of the observed objects. Considering these two challenges, our goal is to design a mechanism that learns from RGB-D data during training and is free of the use of depth data during testing, while maximizing performance.

To achieve this goal, we propose a depth distiller (A2dele), in which two bridges are adopted to connect RGB and depth modalities for transferring depth knowledge to the RGB stream as shown in Figure 2(c). First, we use the network prediction as a bridge for adaptively transferring the pixel-wise depth knowledge to the prediction of the RGB stream, namely an adaptive depth distillation scheme. More precisely, we selectively minimize the differences between predictions generated from the depth stream and RGB stream by an adaptive factor. This scheme realizes the desired control of pixel-wise depth knowledge transferred to RGB stream. Second, we use the network attention as an another bridge for transferring localization knowledge of salient objects to RGB features, namely an attentive depth distillation scheme. Specifically, we improve the prediction of depth stream via dilation operation to ensure the holistic coverage of salient objects, so that the dilated prediction can serve as reliable localization cues. By encouraging consistencies between the dilated prediction and attention map of the RGB stream, background area activations can be effectively suppressed in RGB features. Furthermore, our A2dele can facilitate other existing RGB-D approaches to achieve high efficiency while preserving accuracy. Figure 1 shows that the CPFP'19 [41]+A2dele and DMRA [31]+A2dele achieve comparable accuracy at a significantly smaller model size, compared to the original models.

Our core insight is that we embrace the challenges and move away from attempting to train and test a model both on paired RGB and depth images, and instead test the model over only the single RGB modality. Our approach is to design a depth distiller that uses the network prediction and attention as two bridges connecting RGB and depth modalities while being free of using depth maps during testing. In such way, our adaptive and attention distillation schemes ensure the reliable depth information being transferred by screening out the erroneous depth knowledge. The source code is released[1]. Concretely, we make following contributions:

- We propose a depth distiller (A2dele), which explores the way of using network prediction and attention

[1]https://github.com/OIPLab-DUT/CVPR2020-A2dele

as two bridges to transfer the depth knowledge from the depth stream to the RGB stream. As a result, a lightweight architecture, being free of the depth stream at test time, can be achieved by embedding our proposed A2dele at training time.

- Extensive experimental results on five benchmark datasets demonstrate that our RGB stream achieves state-of-the-art performance, which tremendously minimizes the model size by 76% and runs 12 times faster, compared with the best performing method.

- Our depth distiller (A2dele) can be applied to improve existing RGB-D approaches. Compared to the original models, the ones embedded by our A2dele achieve comparable performance while running much faster (FPS is boosted by nearly twice for DMRA [31] and 3 times for CPFP [41]) at a significantly smaller model size (model size is minimized by 37% for DMRA [31] and 43% for CPFP [41]).

2. Related Work

RGB-D Salient Object Detection. Early RGB-D saliency detection methods [30, 8, 17, 34] manually design handcrafted features and break the new ground. Recently, CNNs-based RGB-D approaches have yielded a qualitative leap in performance due to the powerful ability of CNNs in hierarchically extracting informative features. Zhu et.al [43] use an independent encoder network to make full use of depth cues and assist the RGB-stream network. Chen et.al [3] exploit the cross-model complement across all the levels by a complementarity-aware fusion module. Chen et.al [5] propose a multi-scale multi-path fusion network with cross-modal interactions to enable sufficient and efficient fusion. Chen et.al [4] introduce a cross-modal distillation stream to learn new discriminative multi-modal features in each level. Zhao et.al [41] propose to use the contrast-enhanced depth map as an attention map to suppress distractors in the RGB features. Piao et.al [31] propose a recurrent attention module based on ConvLSTM to progressively learn the internal semantic relation of the multi-modal features.

However, existing RGB-D approaches require an additional network to process depth data which incurs extra computational cost and memory consumption. Moreover, depth maps are easily influenced, which may pose a high risk to accurate saliency detection. These severely impede practical applications of RGB-D saliency detection. In contrast, by embedding our A2dele, we achieve free use of the depth stream at test time, while maximizing performance.

Distillation and Learning under Privileged Information. Our depth distiller is inspired by the generalized distillation [26] that combines distillation [14] and privileged information [36]. In distillation, knowledge is transferred from the teacher network to the student by minimizing the differences between the soft target from the teacher and the class probabilities from the student. Knowledge distillation has been exploited in many computer vision tasks, such as domain adaptation [10], object detection [21, 15], depth estimation [32] and semantic segmentation [13, 25]. In a similar spirit, our goal is to transfer knowledge from the depth stream to the RGB stream, being free use of depth stream during testing. The learning under privileged information provides a network with extra information which is only available in the training stage. Recent works [19, 37, 27] propose to use privileged depth information in semantic segmentation and action recognition. In our case, depth is the privileged information available for training, along with RGB data, but only RGB data is used at test time.

Different from the aforementioned distillation designs which indiscriminately transfer knowledge, we propose a tailor-made depth distiller (A2dele) to achieve the discriminative transfer of useful depth knowledge. It is well known that the unstable quality of depth map can impose negative effects on RGB-D salient object detection. Our A2dele can transfer useful depth information to the RGB stream and meanwhile suppressing erroneous ones.

3. Method

3.1. Overview

Existing methods for RGB-D salient object detection inevitably incur extra computational costs and memory due to requiring an independent subnetwork to process depth data, and the use of depth information during testing may hinder the practical applications of RGB-D saliency detection. To confront those challenges, we propose a depth distiller (A2dele) to improve RGB-D saliency detection taking a single RGB image as input at test time. An overview of the proposed framework is shown in Figure 2(c).

Depth, we train the depth stream to not only locate salient objects accurately but also transfer privileged knowledge for the RGB stream. The encoder in the depth stream is based on VGG16 [35], in which 5 convolutional blocks are maintained and the last pooling and fully-connected layers are discarded. Then we select the high-level features (F^3_{Conv}, F^4_{Conv} and F^5_{Conv}) to detect salient objects. Moreover, we boost the quality of depth features by applying a receptive field block (RFB) [24] in each level. The RFB can capture global contrast information which is suitable to the aim of depth stream. Finally, the decoder takes the depth features as input and make a final prediction. The detailed architecture of the decoder is shown in Figure 3.

RGB, we design an efficient RGB stream to effectively leverage both RGB information and depth knowledge transferred from the depth stream. The RGB stream has the same architecture with the depth stream. The only difference is

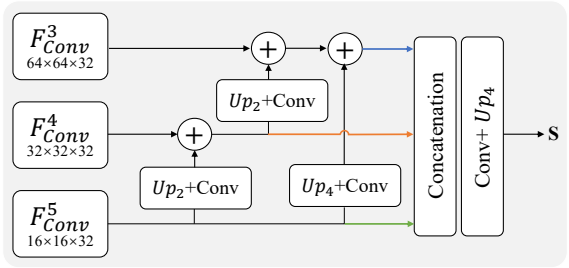

Figure 3. Detailed structure of the decoder in the depth stream or RGB stream.

that we replace the RFB with an attention module. The attention module is lightweight and consists of only one 3×3 convolutional layer. The training of the RGB stream is supervised by our proposed depth distiller (A2dele), which consists of an adaptive depth distillation scheme and an attentive depth distillation scheme (Details in Section 3.2).

3.2. The Proposed Depth Distiller (A2dele)

Inspired by distillation [14] and privileged information [36], we build two bridges connecting RGB and depth modalities via a depth distiller (A2dele) for transferring privileged depth knowledge to the RGB stream. The knowledge is defined as two parts: (1) The first part is designed to achieve the desired control of pixel-wise depth knowledge transferred to the prediction of RGB stream. (2) The second part is designed to transfer localization knowledge of salient objects to RGB features. Next, we elaborate on each distillation scheme in A2dele.

3.2.1 Adaptive Depth Distillation Scheme

In our proposed depth distiller, we use the network prediction as the first bridge across RGB and depth modalities for transferring pixel-wise depth knowledge to the prediction of the RGB stream. To this end, we train the RGB network by minimizing the loss between predictions produced from the depth stream and RGB stream. When we obtain an accurate prediction from the depth stream, this strategy will effectively help the RGB stream easily discriminate salient objects from background. On the contrary, if the prediction is not reliable due to the low-quality depth map, this strategy may introduce side effects in RGB prediction. Based on this observation, we propose an adaptive depth distillation scheme to ensure the desired depth knowledge transfer. More precisely, we design an adaptive factor λ to modulate the influence of the depth stream. The λ is defined as:

$$\lambda = exp(-\alpha L_{CE}(S_{depth}, Y)), \tag{1}$$

where Y represents the ground truth and the hyper-parameter α is set to 70 for keeping the λ ranging from 0 to 1. The λ is inversely related to the loss between the output of the depth stream and ground truth. This indicates that the RGB stream learns from the depth stream when the predictions of the depth stream are reliable; otherwise the RGB stream learns from the ground truth. Thus, the complete loss function is written as:

$$L_{Adap} = \lambda L_{KL}(S_{RGB} \| S_{depth}) + (1-\lambda) L_{CE}(S_{RGB}, Y), \tag{2}$$

where L_{KL} is the Kullback-Leibler divergence loss in which the temperature hyper-parameter T is set to 20 and L_{CE} is the cross entropy loss. Compared to directly enforcing the RGB stream to mimic the output from depth stream with a fixed weight, our proposed adaptive depth distillation scheme allows the RGB stream to selectively absorb the useful depth information from the depth stream.

3.2.2 Attentive Depth Distillation Scheme

Our attentive distillation scheme goes a further step: we choose the network attention as the second bridge for transferring localization knowledge to RGB features. This is achieved by encouraging consistency between the prediction of the depth stream and the attention map in the RGB stream. To minimize the inconsistency, the RGB stream must learn an attention map to approach to the prediction of the depth stream. As the attention map is improved in quality, the distractors of RGB features are suppressed gradually, inching the RGB stream toward accurate localization of salient objects. However, when the depth stream infers incomplete detection of salient objects, this strategy may lead to unsatisfactory segmentation results. To ensure the reliable localization knowledge, we enlarge the coverage area of the prediction from the depth stream to improve its effectiveness via dilation operation as observed in Figure 2(c). The Dilation is achieved by using the max-pooling operation and expressed as:

$$Dilation(S_{depth}) = Maxpool(S_{depth}, kernelsize = 11). \tag{3}$$

By covering more complete regions of salient objects, the dilated prediction of the depth stream can act as better localization cues and help boost the RGB features. In summary, the attentive depth distillation scheme can be defined as:

$$L_{Atten} = \sum_{i=1}^{N} L_{CE}(Att^i_{RGB}, Dilation(S_{depth})), \tag{4}$$

where Att^i_{RGB} represents the i^{th} attention map in the RGB stream. N means the total number of the levels and is set to 3. By minimizing the loss L_{Atten}, the response from outside the salient objects is suppressed, focusing the response on the salient regions.

3.3. Optimization

The training process of our method involves two stages as is presented in Algorithm 1. In stage 1, the depth stream

is supervised by the cross entropy loss L_{CE} with the ground truth Y. During the knowledge distillation process (stage 2), the parameters of the depth stream are kept frozen. The RGB stream is supervised by a combination of the adaptive depth distillation loss L_{Adap} in Eq.(2) and the attentive distillation loss L_{Atten} in Eq.(4). W_D and W_R are the parameters of the depth stream and RGB stream, respectively.

Algorithm 1: Training Process of Our Method

Stage 1 : Training the depth stream.
Input : Depth map.
$W_D = \text{argmin}_{W_D} L_{CE}(S_{depth}, Y)$
Stage 2 : Training the RGB stream.
Input : RGB.
$W_R = \text{argmin}_{W_R} (L_{Adap} + L_{Atten})$

4. Experiments

4.1. Benchmark Datasets

We conduct our experiments on five following widely-used RGB-D datasets. **DUT-RGBD** [31]: contains 1200 images captured by Lytro camera in real life scenes. **NJUD** [17]: includes 1985 stereo image pairs, in which the stereo images are collected from 3D movies, the Internet and photographs are taken by a Fuji W3 stereo camera. **NLPR** [30]: contains 1000 images captured by Kinect under different illumination conditions. **STEREO** [29]: includes 797 stereoscopic images gathered from the Internet. **RGBD135** [6]: includes 135 images captured by Kinect.

For comparison, we adopt the same training set as in [31], which contains 800 samples from the DUT-RGBD dataset, 1485 samples from NJUD and 700 samples from NLPR for training. The remaining images and other two datasets are used for testing to verify the generalization ability of saliency models. To avoid overfitting, we augment the training set by flipping, cropping and rotating.

4.2. Experimental Setup

Evaluation Metrics. We use generally-recognized F-measure (F_β) [1], weighted F-measure (F_β^w) [28] and Mean Absolute Error (MAE). These three evaluation metrics can provide comprehensive and reliable evaluation results and have been well explained in many literatures. We also 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. During the training phrase, we use the Adam optimization [18] algorithm to train our depth stream and RGB stream. The batch size is set as 10 and the initial learning rate is set to 1e-4. The maximum epoches of the depth stream and RGB stream are set to 100 and 50, respectively. All the training images are resized to 256×256.

4.3. Comparison with State-of-the-arts

We compare our RGB stream with 18 other state-of-the-arts methods including 9 RGB-D methods (remarked with $\star$): CTMF* [11], DF* [33], CDCP* [44], PCA* [3], PDNet* [43], MMCI* [5], TANet* [4], CPFP* [41], DMRA* [31]; and 9 RGB methods: DSS [16], Amulet [39], R^{3}Net [7], PiCANet [23], PAGRN [40], PoolNet [22], AFNet [9], CPD [38], EGNet [42]. We implement these models with authorized codes or directly evaluate results provided by authors. Note that CPD [38] and EGNet [42] have two settings (with VGG16 [35] and ResNet50 [12] backbone networks). For fair comparison, we show the results of CPD [38] and EGNet [42] using the same VGG16 backbone network as ours.

Quantitative Evaluation. Table 1 shows the quantitative comparison in terms of three evaluation metrics on five datasets. It can be seen that our proposed RGB stream can outperform both RGB methods and RGB-D methods across five datasets, except second-best weighted F-measure scores on NJUD and RGBD135. Especially, our RGB stream outperforms all other methods by a large margin on DUT-RGBD, NLPR and STEREO, where the images are comparably complicated. This indicates that our distiller can transfer the qualified depth knowledge to facilitate the RGB stream.

Qualitative Evaluation. In Figure 4, we show the qualitative comparison in some challenging cases: low intensity environment (1^{st} row), similar foreground and background (2^{nd} and 3^{rd} rows), transparent object (5^{th} row), small object (5^{th} and 6^{th} rows) and multiple objects (4^{th}, 5^{th} and 6^{th} rows). Compared to the RGB methods (last 4 columns), our method makes it easier to discriminate the salient objects from background and achieves more complete predictions. This indicates that our RGB stream is positively influenced by the depth knowledge transferred from the depth stream, leading to robust results. Moreover, compared to the RGB-D methods ($5^{th} - 8^{th}$ columns), our method also locates and segments salient objects more accurately. It further demonstrates the superiority of our proposed A2dele in transferring depth knowledge.

Complexity Evaluation. Moreover, we compare the model size and FPS (Frames Per Second) with other models for complexity evaluation as shown in Table 1. It can be observed that our RGB stream runs 12 times faster and minimizes the model size by 76% than the best performing method DMRA* [31]. Not only that, compared to the most efficient model CPD [38], we also achieve a large improvement on DUT-RGBD, NJUD and NLPR with half model size and nearly double FPS. Those results further verify that our A2dele enables a high-accuracy and low-cost RGB-D saliency detection model.

Table 1. Quantitative comparisons of F-measure (F_β) [1], weighted F-measure (F_β^w) [28] and Mean Absolute Error (MAE) scores on five RGB-D datasets. $\star$ represents RGB-D methods. - means no available results. (**red**: best, **blue**: second best, **green**: third best).

Methods	Years	FPS↑	Size↓	DUT-RGBD			NJUD			NLPR			STEREO			RGBD135		
				F_β^w ↑	F_β ↑	MAE↓	F_β^w ↑	F_β ↑	MAE↓	F_β^w ↑	F_β ↑	MAE↓	F_β^w ↑	F_β ↑	MAE↓	F_β^w ↑	F_β ↑	MAE↓
DSS	CVPR'17	23	447	.628	.732	.127	.678	.776	.108	.614	.755	.076	.718	.814	.087	.556	.697	.098
Amulet	ICCV'17	21	**133**	.762	.803	.083	.758	.798	.085	.716	.722	.062	.811	.842	.062	.701	.725	.070
R^{3}Net	IJCAI'18	22	225	.709	.781	.113	.736	.775	.092	.611	.649	.101	.752	.800	.084	.693	.728	.066
PiCANet	CVPR'18	5	197	.741	.826	.080	.768	.806	.071	.707	.761	.053	.792	.835	.062	.741	.797	.042
PAGRN	CVPR'18	-	-	.746	.836	.079	.746	.827	.081	.707	.795	.051	.774	.856	.067	.748	.834	.044
PoolNet	CVPR'19	32	279	**.836**	.871	**.049**	.816	.850	.057	.771	.791	.046	.849	**.877**	**.045**	.814	.852	**.031**
AFNet	CVPR'19	26	144	.817	.851	.064	**.832**	**.857**	**.056**	.796	.807	.043	**.850**	**.876**	**.046**	.816	.840	.034
CPD	CVPR'19	**66**	**112**	.835	**.872**	.055	.821	.853	.059	**.829**	**.840**	**.037**	**.851**	**.880**	**.046**	**.841**	**.860**	**.028**
EGNet	ICCV'19	21	412	.805	.866	.059	.808	.846	.060	.774	.800	.047	.835	**.876**	.049	.787	.831	.035
CTMF*	Tcyb'17	**50**	826	.690	.792	.097	.732	.788	.085	.691	.723	.056	.727	.786	.087	.694	.765	.055
DF*	TIP'17	-	-	.542	.748	.145	.552	.744	.151	.524	.682	.099	.576	.761	.142	.397	.566	.130
CDCP*	ICCV'17	-	-	.530	.633	.159	.522	.618	.181	.512	.591	.114	.595	.680	.149	.484	.583	.119
PCA*	CVPR'18	15	534	.696	.760	.100	.811	.844	.059	.772	.794	.044	.810	.845	.061	.718	.763	.049
PDNet*	ICME'19	-	-	.650	.757	.112	.798	.832	.062	.659	.740	.064	.799	.833	.064	.731	.800	.050
MMCI*	PR'19	19	930	.636	.753	.112	.749	.813	.079	.688	.729	.059	.747	.812	.080	.656	.750	.064
TANet*	TIP'19	-	-	.712	.779	.093	.812	.844	.061	.789	.795	.041	.811	.849	.059	.745	.782	.045
CPFP*	CVPR'19	7	278	.644	.736	.099	-	-	-	.820	.822	.036	-	-	-	.794	.819	.037
DMRA*	ICCV'19	10	239	**.858**	**.883**	**.048**	**.853**	**.872**	**.051**	**.845**	**.854**	**.031**	**.850**	.868	.047	**.849**	**.857**	**.029**
Our	-	**120**	**57.3**	**.870**	**.892**	**.042**	**.851**	**.874**	**.051**	**.867**	**.878**	**.028**	**.867**	**.884**	**.043**	**.845**	**.865**	**.028**

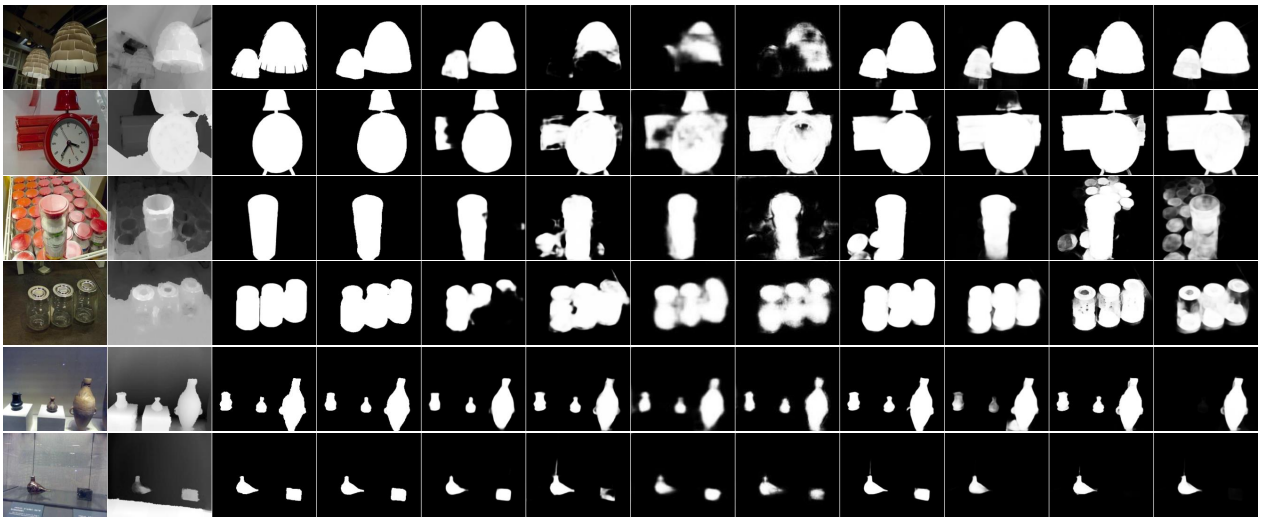

Figure 4. Visual comparison of our RGB stream with top-ranking CNNs-based methods in some challenging scenes.

4.4. Ablation Studies

Effect of Adaptive Depth Distillation Scheme. Our adaptive depth distillation scheme aims to transfer the desired pixel-wise depth knowledge to the prediction of RGB stream. We look in to the effect of enabling our adaptive depth distillation scheme as shown in Table 2. It is seen that our adaptive distillation largely improves the baseline RGB stream (leveraging RGB only) across four datasets. We also show the visual effects in Figure 5. It can be observed that our adaptive distillation scheme can help the RGB stream distinguish the salient objects from background by transferring the high-quality depth knowledge (1^{st} and 2^{nd} rows), and remove the negative effects caused by inaccurate depth map (3^{rd} row). Moreover, to make a

Table 2. The effect of different distillation schemes in our proposed A2dele. λ denotes the adaptive depth distillation scheme with fixed λ and L_{Adap} denotes our proposed adaptive factor. L_{Atten} represents attentive distillation scheme.

Model	DUT-RGBD			NJUD			NLPR			STEREO		
	$F_\beta^w \uparrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	MAE$\downarrow$
Depth	.829	.852	.054	.815	.835	.061	.811	.825	.043	.648	.702	.116
RGB	.836	.873	.052	.817	.848	.058	.834	.850	.036	.829	.860	.053
RGB+λ=0.3	.856	.883	.048	.841	.862	.053	.849	.863	.032	.850	.869	.048
RGB+λ=0.5	.858	.884	.048	.840	.863	.053	.854	.869	.031	.855	.875	.046
RGB+λ=0.7	.834	.863	.056	.823	.844	.058	.830	.843	.037	.832	.852	.054
RGB+L_{Adap}	.861	.886	.045	.845	.867	**.051**	.855	.870	.032	.858	.877	.046
RGB+L_{Adap}+L_{Atten}	**.870**	**.892**	**.042**	**.851**	**.874**	**.051**	**.867**	**.878**	**.028**	**.867**	**.884**	**.043**

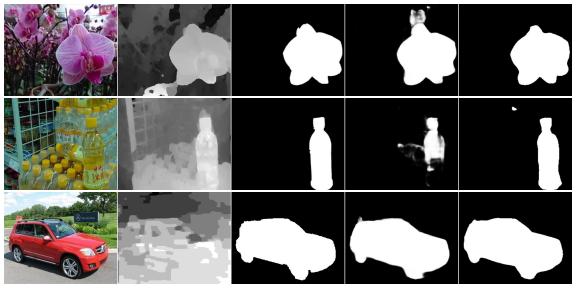

Figure 5. Visual analysis of the adaptive depth distillation scheme.

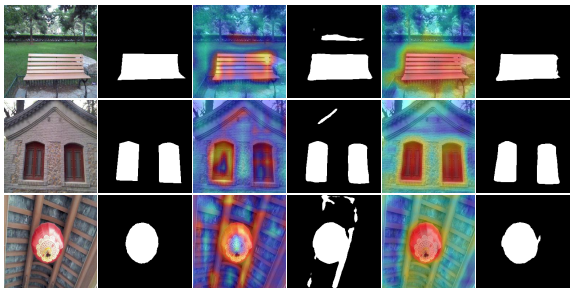

Figure 6. Visual analysis of the attentive depth distillation scheme. (a) and (b) denote the attention map and prediction generated from RGB+L_{Adap}, respectively. (c) and (d) represent the attention map and prediction generated from RGB+L_{Adap}+L_{Atten}, respectively.

deeper analysis about the core component in the adaptive depth distillation scheme – the adaptive factor λ, we add comparisons with fixed λ (0.3, 0.5, 0.7) in Table 2. It can be seen that our 'RGB+L_{Adap}' achieves the overall best results. Learning from the depth stream with fixed λ cannot maximize the benefits of depth stream. By contrast, our adaptive factor can tackle this dilemma by selectively transferring the depth knowledge to the RGB stream according to the performance of the depth stream.

Effect of Attentive Depth Distillation Scheme. Our attentive depth distillation scheme aims to transfer localization knowledge to RGB features. To prove the effect of the attentive depth distillation scheme, we visualize the attention map and saliency prediction in the absence of this scheme as shown in Figure 6. It is obvious that without our attentive depth distillation scheme, the attention map (Figure 6(a)) can not effectively filter the distractors of RGB features, introducing some background noise in the saliency prediction (Figure 6(b)). In contrast, the attention map generated by adding attentive depth distillation scheme (Figure 6(c)) can effectively suppress the distractions of background in RGB features and as a result, the prediction highlights the salient objects successfully (Figure 6(d)). These visual improvements are reasonable since the useful RGB features are emphasized and background area activations are suppressed by the proposed attentive depth distillation scheme. Also in Table 2, the improved performances across four datasets are achieved by adding our attentive depth distillation scheme.

4.5. Applying A2dele in Existing RGB-D Models

In this paper, we apply the proposed A2dele in two top-ranking RGB-D models (CPFP [41], DMRA [31]) to achieve improved efficiency, as well as comparable accuracy. CPFP uses a contrast-enhanced subnet to process depth data and DMRA adopts a VGG-19 to encode depth features. We first replace the original depth stream (the subnet in CPFP and VGG-19 in DMRA) with ours, and then impose the proposed two distillation schemes. Specifically, to apply our attentive depth distillation scheme, we add the same attention module in each level. The attention module is lightweight and nearly does not cause extra computation cost, referred to Table 3. And for DMRA, the depth vector in the depth-induced multi-scale weighting module is set to one. For fair comparisons, we adopt the same training sets and test sets with their original settings.

In Table 3, we show the quantitative comparison of the original models and the improved models (+A2dele). It can be observed that our A2dele largely improves the efficiency of original models. In detail, our A2dele boots the FPS of CPFP by 340% and the FPS of DMRA by 180%, and tremendously minimizes the model size of CPFP by 43% and the model size of DMRA by 37%. On the other hand, by applying our A2dele, we improve the performance of

Table 3. Quantitative Comparison of applying A2dele on top-ranking RGB-D models with original models. '-RGB' represents the RGB-D models without depth stream, and '+A2dele' represents embedding '-RGB' with our A2dele.

Methods	Size(M)↓	FPS↑	LFSD [20]		NJU2000 [17]		RGBD135 [6]		NLPR [30]	
			F_β ↑	MAE↓	F_β ↑	MAE↓	F_β ↑	MAE↓	F_β ↑	MAE↓
CPFP-RGB	**159.37**	**31**	.759	.123	.844	.057	.804	.040	.797	.041
CPFP	278	7	**.811**	**.088**	.850	**.053**	.815	**.037**	.840	.036
CPFP+A2dele	**159.42**	**31**	.806	.094	**.861**	**.053**	**.818**	.043	**.873**	**.033**
Methods	Size(M)↓	FPS↑	DUT-RGBD [31]		NJUD [17]		NLPR [30]		STEREO [29]	
			F_β ↑	MAE↓	F_β ↑	MAE↓	F_β ↑	MAE↓	F_β ↑	MAE↓
DMRA-RGB	**150.14**	**28**	.874	.054	.828	.061	.826	.036	.844	.054
DMRA	238.8	10	.883	.048	**.872**	**.051**	**.855**	**.031**	.868	.047
DMRA+A2dele	**150.15**	**28**	**.889**	**.040**	.867	**.051**	.854	.032	**.869**	**.046**

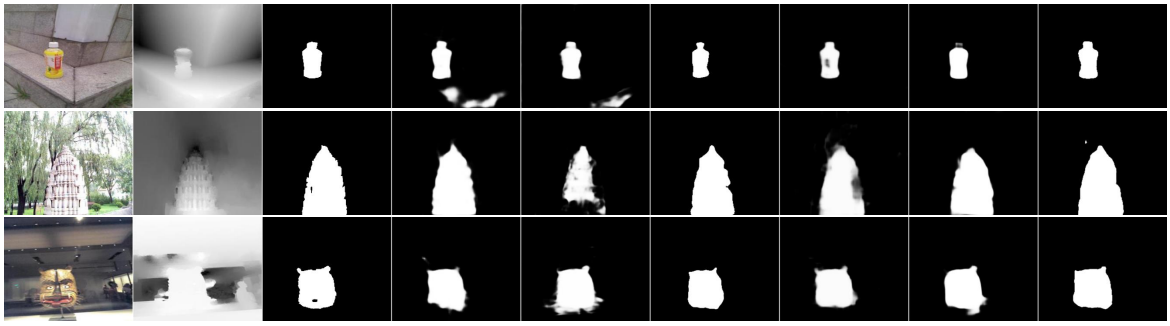

Figure 7. Visual Comparison of applying A2dele on top-ranking RGB-D models with original models on NLPR dataset. The meaning of indexes has been explained in the caption of Table 3.

CPFP-RGB and DMRA-RGB (without depth stream) by a dramatic margin across four datasets. These results further verify the generalization of our A2dele. Meanwhile, we also achieve comparable performance compared to the original models (CPFP and DMRA). Especially, the CPFP+A2dele achieves large improvements on NLPR and DMRA+A2dele improves the performance on DUT-RGBD by a large margin. Moreover, our A2dele leaves the depth data unused during testing, allowing the original model to be more applicable.

In Figure 7, we show some challenging cases in NLPR dataset: inaccurate depth map (1^{st} and 2^{nd} rows) or depth map with extremely low contrast between salient objects and non-salient regions (3^{rd} row). We can see that the CPFP+A2dele segments more uniform salient objects than the original model. Consistently, CPFP+A2dele achieves large improvements in F-measure score as shown in Table 3. This improvement is reasonable since CPFP does not consider the bad effects caused by low-quality depth map, but our A2dele can screen out the erroneous effects due to its discriminative ability of transferring useful depth knowledge. Meanwhile, the DMRA+A2dele also benefits from our A2dele and improves robustness in these challenging scenes.

5. Conclusion

In this paper, we propose a distiller (A2dele) within a two-stream framework that learns from RGB-D data and can be tested on RGB only, while maximizing performance. The proposed A2dele uses the network prediction as the first bridge to adaptively transfer the desired pixel-wise depth knowledge to the prediction of the RGB stream, while the network attention serves as the second bridge for transferring the localization knowledge of salient objects to RGB features. We conduct the experiments on five benchmark datasets and demonstrate that our method achieves state-of-the-arts performance and runs significantly faster at a much smaller model size than existing RGB-D and RGB methods. To prove the generalization of our A2dele, we apply it on top-ranking RGB-D networks. Extensive experiments show that our A2dele can improve the efficiency of RGB-D methods by a large margin, while maintaining performance.

Acknowledgements. This work was supported by the Science and Technology Innovation Foundation of Dalian (2019J12GX034), the National Natural Science Foundation of China (61976035, 61725202, U1903215, U1708263, 61829102, 91538201 and 61751212) , and the Fundamental Research Funds for the Central Universities (DUT19JC58).

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, number CONF, pages 1597–1604, 2009.

[2] Ali Borji, Ming Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Eprint Arxiv*, 16(7):3118, 2014.

[3] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *CVPR*, pages 3051–3060, 2018.

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

[5] Hao Chen, Youfu Li, and Dan Su. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. *PR*, 86:376–385, 2019.

[6] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *Proceedings of international conference on internet multimedia computing and service*, page 23. ACM, 2014.

[7] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. R3net: Recurrent residual refinement network for saliency detection. In *IJCAI*, pages 684–690. AAAI Press, 2018.

[8] David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. Local background enclosure for rgb-d salient object detection. In *CVPR*, pages 2343–2350, 2016.

[9] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *CVPR*, pages 1623–1632, 2019.

[10] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *CVPR*, pages 2827–2836, 2016.

[11] 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 Tcyb*, 48(11):3171–3183, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[13] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. In *CVPR*, pages 578–587, 2019.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[15] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *CVPR*, June 2016.

[16] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *CVPR*, pages 3203–3212, 2017.

[17] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *ICIP*, pages 1115–1119. IEEE, 2014.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2014.

[19] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. Spigan: Privileged adversarial learning from simulation. *ICLR*, 2019.

[20] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *CVPR*, pages 2806–2813, 2014.

[21] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *CVPR*, pages 6356–6364, 2017.

[22] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. *CVPR*, 2019.

[23] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, pages 3089–3098, 2018.

[24] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *ECCV*, pages 385–400, 2018.

[25] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *CVPR*, pages 2604–2613, 2019.

[26] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

[27] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *ECCV*, September 2018.

[28] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, pages 248–255, 2014.

[29] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, pages 454–461. IEEE, 2012.

[30] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *ECCV*, pages 92–109. Springer, 2014.

[31] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, October 2019.

[32] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *CVPR*, pages 9768–9777, 2019.

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

[34] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. Exploiting global priors for rgb-d saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 25–32, 2015.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[36] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *Journal of machine learning research*, 16(2023-2049):2, 2015.

[37] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. *arXiv preprint arXiv:1904.01886*, 2019.

[38] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019.

[39] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, pages 202–211, 2017.

[40] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, pages 714–722, 2018.

[41] 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 *CVPR*, 2019.

[42] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet:edge guidance network for salient object detection. In *ICCV*, Oct 2019.

[43] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. Pdnet: Prior-model guided depth-enhanced network for salient object detection. In *ICME*, pages 199–204. IEEE, 2019.

[44] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. An innovative salient object detection using center-dark channel prior. In *ICCV*, pages 1509–1515, 2017.