PROMOTING SALIENCY FROM DEPTH: DEEP UNSUPERVISED RGB-D SALIENCY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

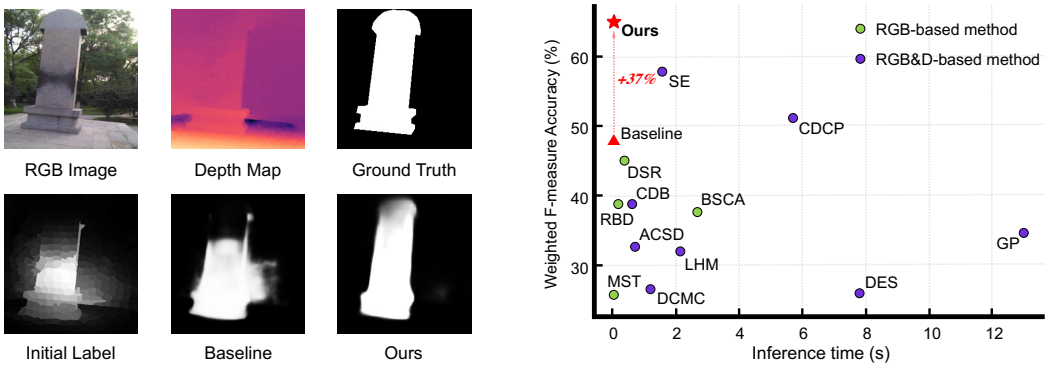
Growing interests in RGB-D salient object detection (RGB-D SOD) have been witnessed in recent years, owing partly to the popularity of depth sensors and the rapid progress of deep learning techniques. However, existing RGB-D SOD methods typically demand large quantity of training images being thoroughly annotated at pixel-level. The laborious and time-consuming manual annotation becomes a real bottleneck in various practical scenarios. On the other hand, current unsupervised RGB-D SOD methods still heavily rely on handcrafted feature representations. In this paper, we make the earliest effort to propose a *deep unsupervised RGB-D saliency detection* approach that does not need any manual pixel-level annotations in the training stage. It is made possible by two key ingredients in our training pipeline. First, a depth-disentangled saliency update (DSU) framework is designed to automatically produce pseudo-labels with iterative follow-up refinements, which provides more trustworthy supervision signals for training the saliency network. Second, an attentive training strategy is introduced to tackle the issue of noisy pseudo-labels, by properly re-weighting to highlight the more reliable pseudo-labels. Extensive experiments demonstrate the superior efficiency and effectiveness of our approach in addressing unsupervised RGB-D SOD. Moreover, our approach can be easily adapted to tackle the fully-supervised scenario, where empirical studies show that by incorporating our method, the performance of existing RGB-D SOD models can be significantly improved. *Source code will be made publicly available.*

1 INTRODUCTION

The recent development of RGB-D salient object detection, *i.e.*, RGB-D SOD, is especially fueled by the increasingly accessible 3D imaging sensors (Giancola et al., 2018) and their diverse applications, including image caption generation (Xu et al., 2015), image retrieval (Ko et al., 2004; Shao & Brady, 2006) and video analysis (Liu et al., 2008; Wang et al., 2018), to name a few. Provided the multi-modality input of an RGB image and its depth map, the task of RGB-D SOD is to effectively identify and segment the most distinctive objects in a scene.

The state-of-the-art RGB-D SOD approaches (Li et al., 2021; Ji et al., 2021; Chen et al., 2020b; Li et al., 2020b) typically entail an image-to-mask mapping pipeline that is based on the powerful deep learning paradigms of *e.g.*, VGG16 (Simonyan & Zisserman, 2015) or ResNet50 (He et al., 2016), which have proven to be effective, and have achieved excellent performance. On the other hand, these RGB-D SOD methods are fully supervised, thus demand a significant amount of pixel-level training annotations. This however becomes much less appealing in practical scenarios, owing to the laborious and time-consuming process in obtaining manual annotations. It is therefore natural and desirable to contemplating unsupervised alternatives. Unfortunately, existing unsupervised RGB-D SOD methods, such as global priors (Ren et al., 2015), center prior (Zhu et al., 2017b), and depth contrast prior (Ju et al., 2014), rely primarily on handcrafted feature representations. This is in stark contrast to the deep representations learned by their supervised SOD counterparts, which in effect imposes severe limitations on the feature representation power that may otherwise benefit greatly from the potentially abundant unlabeled RGB-D images.

These observations motivate us to explore a new problem of *deep unsupervised RGB-D saliency detection*: given an unlabeled set of RGB-D images, deep neural network is trained to predict saliency without any laborious human annotations in the training stage. A relatively straightforward idea is to exploit the outputs from traditional RGB-D method as pseudo-labels, which are internally employed



(a) Deep unsupervised RGB-D saliency detection (b) Accuracy and inference speed comparison

Figure 1: (a) An illustration of deep unsupervised RGB-D saliency detection. 'Initial label' is generated by traditional method. 'Baseline' shows the saliency map generated by saliency network trained with initial pseudo-labels. 'Ours' shows our final results. (b) Efficiency and effectiveness comparison over a wide range of unsupervised SOD methods on the NLPR benchmark. **Promoting Saliency From Depth:** our approach achieves a large-margin improvement over the baseline, by engaging depth information to improve pseudo-labels in the training process, without introducing additional computational cost during inference, shown in red arrow.

to train the saliency prediction network ('baseline'). Moreover, the input depth map may serve as a complementary source of information in refining the pseudo-labels, as it contains cues of spatial scene layout that may help in exposing the salient objects. Nevertheless, practical examination reveals two main issues: *(1) Inconsistency and large variations in raw depth maps*: as illustrated in Fig. 1 (a), similar depth values are often shared by a salient object and its surrounding, making it very difficult in extracting the salient regions from depth without explicit pixel-level supervision; *(2) Noises from unreliable pseudo-labels*: some unreliable pseudo-labels may inevitably bring false positive into training, resulting in severe damage in its prediction performance.

To address the above challenges, we propose a dedicated method that contains the following *two* key components. First, a depth-disentangled saliency update (DSU) framework is proposed to iteratively refine & update the pseudo-labels by engaging the depth knowledge. In it, a depth-disentangled network is to explicitly learn the discriminative saliency cues and non-salient background from raw depth map, denoted as saliency-guided depth D_{Sal} and non-saliency-guided depth D_{NonSal}, respectively. This is followed by a depth-disentangled label update (DLU) module that takes advantage of D_{Sal} to emphasize saliency response from pseudo-label, and utilizes D_{NonSal} to eliminate the background influence, gearing toward facilitating more trustworthy supervision signals in training the saliency network. We note that the DSU module is not performed at test time, so does not affect the inference speed. That is, in the inference stage, our trained model takes as input only the RGB image, without involving the depth map into computation. Second, an attentive training strategy is introduced to alleviate the issue of noisy pseudo-labels, by re-weighting the training samples in each training batch to focus on those more reliable pseudo-labels. As demonstrated in Fig. 1 (b), our approach works effectively and efficiently in practice, significantly outperforming current unsupervised SOD methods on the widely-used NLPR benchmark. Specifically, it improves over the baseline by 37%, a significant amount without incurring extra computation cost. Consequently, the test time execution of our approach is at 35 frame-per-second (FPS), the fastest among all RGB-D unsupervised methods, and on par with the most efficient RGB-based methods. In summary, our main contributions are as follows:

- To our knowledge, our work is the first in exploring deep representation to tackle the problem of unsupervised RGB-D saliency detection. This is enabled by two key components in the training process, namely the DSU strategy to produce & refine pseudo-labels, and the attentive training strategy to alleviate the influence of noisy pseudo-labels. This results in a light-weight architecture that engages only RGB data at test time (*i.e.,* w/o depth map), achieving a significant improvement without extra computation cost.
- Empirically, our approach outperforms state-of-the-art unsupervised methods on four public benchmarks. Moreover, it runs in real time at 35 FPS, much faster than existing unsupervised RGB-D SOD methods, and at least on par with the fastest RGB counterparts.

- Our approach could be adapted to work with fully-supervised scenario. As demonstrated in Sec. 4.4, augmented with our proposed DSU module, the performance of existing RGB-D SOD models can be significantly boosted. *Source code will be made publicly available.*

2 Related Work

Remarkable progresses have been made recently in salient object detection (Ma et al., 2021; Xu et al., 2021; Wei et al., 2020; Wu et al., 2019b; Tsai et al., 2018; Liu et al., 2018), where the performance tends to deteriorate in the presence of cluttered or low-contrast backgrounds. Moreover, promising results are shown by a variety of recent efforts (Chen et al., 2021a; Luo et al., 2020; Chen et al., 2020a; Cong et al., 2020; 2018; Desingh et al., 2013; Li et al., 2020a; Liao et al., 2020; Shigematsu et al., 2017) that integrate effective depth cues to tackle these issues. Those methods, however, typically demand extensive annotations, which are labor-intensive and time-consuming. This naturally leads to the consideration of unsupervised SODs that do not rely on such annotations.

Prior to the deep learning era (Wang et al., 2021; Zhao et al., 2020a), traditional RGB-D saliency methods are mainly based on the manually-crafted RGB features and depth cues to infer a saliency map. Due to their lack of reliance on manual human annotation, these traditional methods can be regarded as early manifestations of unsupervised SOD. Ju *et al*. (Ju et al., 2014) consider the incorporation of a prior induced from anisotropic center-surround operator, and an additional depth prior. Ren *et al*. (Ren et al., 2015) also introduce two priors: normalized depth prior and the global-context surface orientation prior. In contrary to direct utilization of the depth contrast priors, local background enclosure prior is explicitly developed by (Feng et al., 2016). Interested readers may refer to (Fan et al., 2020a; Peng et al., 2014; Zhou et al., 2021) for more in-depth surveys and analyses. However, by relying on manually-crafted priors, these methods tend to have inferior performance.

Meanwhile, considerable performance gain has been achieved by recent efforts in RGB-based unsupervised SOD (Zhang et al., 2017; 2018; Nguyen et al., 2019; Hsu et al., 2018a), which instead construct automated feature representations using deep learning. A typical strategy is to leverage the noisy output produced by traditional methods as pseudo-label (*i.e.*, supervisory signal) for training saliency prediction net. The pioneering work of Zhang *et al*. (Zhang et al., 2017) fuses the outputs of multiple unsupervised saliency models as guiding signals in CNN training. In (Zhang et al., 2018), competitive performance is achieved by fitting a noise modeling module to the noise distribution of pseudo-label. Instead of directly using pseudo-labels from handcrafted methods, Nguyen *et al*. (Nguyen et al., 2019) further refine pseudo-labels via a self-supervision iterative process. Besides, deep unsupervised learning has been considered by (Hsu et al., 2018b; Tsai et al., 2019) for the co-saliency task, where superior performance has been obtained by engaging a fusion-learning scheme (Hsu et al., 2018b), or utilizing dedicated loss terms (Tsai et al., 2019). These methods demonstrate that the incorporation of powerful deep neural network brings better feature representation than those unsupervised SOD counterparts based on handcrafted features.

In this work, a principled research investigation on *deep unsupervised RGB-D saliency detection* is presented for the first time. Different from existing deep RGB-based unsupervised SOD (Zhang et al., 2017; 2018) that use pseudo-labels from the handcrafted methods directly, our work aims to refine and update pseudo-labels to remove the undesired noise. And compared to the refinement method (Nguyen et al., 2019) using self-supervision technique, our approach instead adopts disentangled depth information to promote pseudo-labels.

3 Methodology

3.1 Overview

Fig. 2 presents an overview of our Depth-disentangled Saliency Update (DSU) framework. Overall our DSU strives to significantly improve the quality of pseudo-labels, that leads to more trustworthy supervision signals for training the saliency network. It consists of three key components. **First** is a saliency network responsible for saliency prediction, whose initial supervision signal is provided by traditional handcrafted method without using human annotations. **Second**, a depth network and a depth-disentangled network are designed to decompose depth cues into saliency-guided depth D_{Sal} and non-saliency-guided depth D_{NonSal}, to explicitly depict saliency cues in the spatial layout. **Third**, a depth-disentangled label update (DLU) module is devised to refine and update the pseudo-labels, by engaging the learned D_{Sal} and D_{NonSal}. The updated pseudo-labels could in turn provide more trustworthy supervisions for the saliency network. Moreover, an attentive training strategy (ATS) is incorporated when training the saliency network, tailored for noisy unsupervised learning

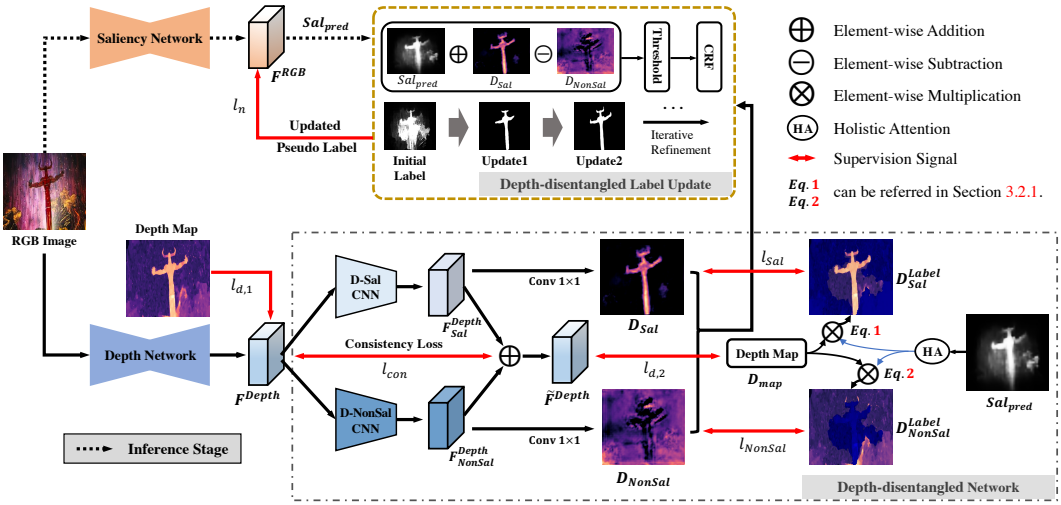


Figure 2: Overview of the proposed method. The saliency network is trained with the iteratively updated pseudo-labels. The depth network and depth-disentangled network are designed to decompose raw depth into saliency-guided depth D_{Sal} and non-saliency-guided depth D_{NonSal}, which are subsequently fed into the depth-disentangled label update (DLU) module to refine and update pseudo-labels. The inference stage involves only the black dashed portion.

by mitigating the ambiguities caused by noisy pseudo-labels. We note that the DSU and ATS are not performed at test time, so does not affect the inference speed. In other words, the inference stage involves only the black dashed portion of the proposed network architecture in Fig. 2, *i.e.*, only RGB images are used for predicting saliency, which enables very efficient detection.

3.2 Depth-disentangled Saliency Update

The scene geometry embedded in the depth map can serve as a complementary information source to refine the pseudo-labels. Nevertheless, as discussed in Sec.1, a direct adoption of the raw depth may not necessarily lead to good performance due to the large value variations and inherit noises in raw depth map. Without explicit pixel-wise supervision, the saliency model may unfortunately be confused by nearby background regions having the same depth but with different contexts; meanwhile, depth variations in salient or background regions may also result in false responses. These observations motivate us to design a depth-disentangled network to effectively capture discriminative saliency cues from depth, and a depth-disentangled label update (DLU) strategy to improve and refine pseudo-labels by engaging the disentangled depth knowledges.

3.2.1 Depth-disentangled Network

The depth-disentangled network aims at capturing valuable saliency as well as redundant non-salient cues from raw depth map. As presented in the bottom of Fig. 2, informative depth feature F^{Depth} are first extracted from the depth network under the supervision of raw depth map, using the mean square error (MSE) loss function, *i.e.*, $l_{d,1}$. The F^{Depth} is then decomposed into saliency-guided depth D_{Sal} and non-saliency-guided depth D_{NonSal} following two principles: 1) explicitly guiding the model to learn saliency-specific cues from depth; 2) ensuring the coherence between the disentangled and original depth features.

Specifically, in the bottom right of Fig. 2, we first construct the spatial supervision signals for the depth-disentangled network. Given the rough saliency prediction Sal_{pred} from the saliency network and the raw depth map D_{map}, the (non-)saliency-guided depth masks, *i.e.*, D^{Label}_{Sal} and D^{Label}_{NonSal}, can be obtained by multiplying Sal_{pred} (or $1 - Sal_{pred}$) and depth map D_{map} in a spatial attention manner. Since the predicted saliency may contain errors introduced from the inaccurate pseudo-labels, we employ a holistic attention (HA) operation (Wu et al., 2019a) to smooth the coverage area of the predicted saliency, so as to effectively perceive more saliency area from depth. Formally, the (non-)saliency-guided depth masks are generated by:

$$D^{Label}_{Sal} = \Psi_{\max}(\mathcal{F}_G(Sal_{pred}, k), Sal_{pred}) \otimes D_{map}, \tag{1}$$

$$D^{Label}_{NonSal} = \Psi_{\max}(\mathcal{F}_G(1 - Sal_{pred}, k), 1 - Sal_{pred}) \otimes D_{map}, \tag{2}$$

where $\mathcal{F}_G(\cdot, k)$ represents the HA operation, which is implemented using the convolution operation with Gaussian kernel k and zero bias; the size and standard deviation of the Gaussian kernel k are

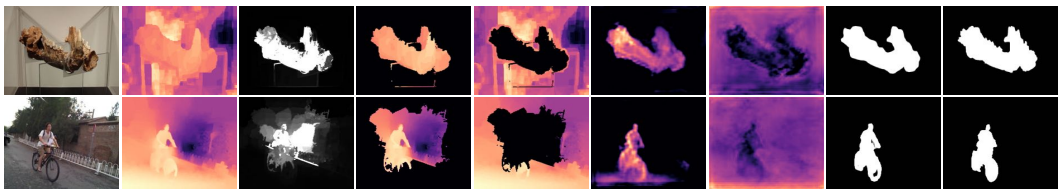


Figure 3: The internal inspections of the proposed DSU. It is observed that the updated label exhibits more reliable saliency signals than initial pseudo-label.

initialized with 32 and 4, respectively, which are then finetuned through the training procedure; $\Psi_{\max}(\cdot,\cdot)$ is a maximum function to preserve the higher values from the Gaussian filtered map and the original map; $\otimes$ denotes pixel-wise multiplication.

Building upon the guidance of D_{Sal}^{Label} and D_{NonSal}^{Label}, F^{Depth} is fed into D-Sal CNN and D-NonSal CNN to explicitly learn valuable saliency and redundant non-salient cues from depth map, generating D_{Sal} and D_{NonSal}, respectively. The loss functions here (*i.e.,* l_{Sal} and l_{NonSal}) are MSE loss. Detailed structures of D-Sal and D-NonSal CNNs are in the appendix. To further ensure the coherence between the disentangled and original depth features, a consistency loss, l_{con}, is employed as:

$$l_{con} = \frac{1}{H \times W \times C} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{C} \| F_{i,j,k}^{Depth}, \tilde{F}_{i,j,k}^{Depth} \|_2, \tag{3}$$

where $\tilde{F}^{Depth}$ is the sum of the disentangled F_{Sal}^{Depth} and F_{NonSal}^{Depth}, which denotes the regenerated depth feature; H, W and C are the height, width and channel of F^{Depth} and $\tilde{F}^{Depth}$; $\|\cdot\|_2$ represents Euclidean norm. Here, $\tilde{F}^{Depth}$ is also under the supervision of depth map, using MSE loss, *i.e.,* $l_{d,2}$.

Then, the overall training objective for the depth network and the depth-disentangled network is as:

$$\mathcal{L}_{depth} = \frac{1}{5N} \sum_{n=1}^{N} (l_{d,1}^{n} + l_{d,2}^{n} + l_{Sal}^{n} + l_{NonSal}^{n} + \lambda l_{con}^{n}), \tag{4}$$

where n denotes the n_{th} sample in a mini-batch with N training samples; λ is set to 0.2 in the experiments to balance the consistency loss l_{con} and other loss terms. As empirical evidences in Fig. 3 suggested, comparing to the raw depth map, D_{Sal} is better at capturing discriminative cues of the salient objects, while D_{NonSal} better depicts the complementary background context.

In the next step, the learned D_{Sal} and D_{NonSal} are fed into our Depth-disentangled Label Update, to obtain the improved pseudo-labels.

3.2.2 Depth-disentangled Label Update

To maintain more reliable supervision signals in training, a depth-disentangled label update (DLU) strategy is devised to iteratively refine & update pseudo-labels. Specifically, as shown in the upper stream of Fig. 2, using the obtained Sal_{pred}, D_{Sal} and D_{NonSal}, the DLU simultaneously highlights the salient regions in the coarse saliency prediction by the sum of Sal_{pred} and D_{Sal}, and suppresses the non-saliency negative responses by subtracting D_{NonSal} in a pixel-wise manner. This process can be formulated as:

$$\mathcal{S}_{temp} = Sal_{pred}^{i,j} + D_{Sal}^{i,j} - D_{NonSal}^{i,j} \Big|_{i \in [1,H]; j \in [1,W]}. \tag{5}$$

To avoid the value overflow of the obtained $\mathcal{S}_{temp}$ (*i.e.*, removing negative numbers and normalizing the results to the range of [0, 1]), a thresholding operation and a normalization process are performed as:

$$\mathcal{S}_{\mathcal{N}} = \frac{\mathcal{S}_n^{i,j} - min(\mathcal{S}_n)}{max(\mathcal{S}_n) - min(\mathcal{S}_n)}, \text{where } \mathcal{S}_n = \begin{cases} 0, & \text{if } \mathcal{S}_{temp}^{i,j} < 0 \\ \mathcal{S}_{temp}^{i,j}, & \text{others} \end{cases}, i \in [1,H]; j \in [1,W], \tag{6}$$

where $min(\cdot)$ and $max(\cdot)$ denote the minimum and maximum functions. Finally, a fully-connected conditional random field (CRF (Hou et al., 2017)) is applied to $\mathcal{S}_{\mathcal{N}}$, to generate the enhanced saliency map $\mathcal{S}_{map}$ as the updated pseudo-labels. The empirical result of the proposed DSU step is showcased in Fig. 3, with quantitative results presented in Table 4. These internal evidences suggest that the DSU leads to noticeable improvement after merely several iterations.

3.3 ATTENTIVE TRAINING STRATEGY

When training the saliency network using pseudo-labels, an attentive training strategy (ATS) is proposed to tailor for the deep unsupervised learning context, to reduce the influence of ambiguous pseudo-labels, and concentrate on the more reliable training examples. This strategy is inspired by the human learning process of understanding new knowledge, that is, from general to specific understanding cycle (Dixon, 1999; Peltier et al., 2005). The ATS alternates between two steps to re-weigh the training instances in a mini-batch.

To be specific, we first start by settling the related loss functions. For the n_{th} sample in a mini-batch with N training samples, we define the binary cross-entropy loss between the predicted saliency Sal^n_{pred} and the pseudo-label $\mathcal{S}^n_{map}$ as:

$$l_n = -(\mathcal{S}^n_{map} \cdot \log Sal^n_{pred} + (1 - \mathcal{S}^n_{map}) \cdot \log(1 - Sal^n_{pred})). \tag{7}$$

Then, the training objective for the saliency network in current mini-batch is defined as an attentive binary cross-entropy loss $\mathcal{L}_{sal}$, which can be represented as follows:

$$\mathcal{L}_{sal} = \frac{1}{\sum_{n=1}^{N} \alpha_n} \sum_{n=1}^{N} (\alpha_n \cdot l_n), \alpha_n = \begin{cases} 1, & \text{step one,} \\ \frac{\sum_{i \in N}^{i \neq n} e^{l_i}}{\sum_{i \in N} e^{l_i}}, & \text{step two,} \end{cases} \tag{8}$$

where α_n represents the weight of the n_{th} training sample at current training mini-batch.

The ATS starts from a uniform weight for each training sample in step one to learn general representations across a lot of training data. Step two decreases the importance of ambiguous training instances through the imposed attentive loss; the higher the loss value, the less weight an instance is to get.

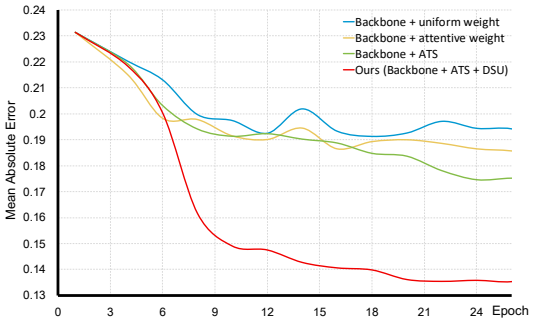


Figure 4: Analysis of the proposed attentive training strategy, when evaluated on NJUD testset. The 'backbone' refers to the saliency network trained with initial pseudo-labels.

In this paper, we define step one and two as a training round (2τ epochs). During each training round, the saliency loss $\mathcal{L}_{sal}$ and depth loss $\mathcal{L}_{depth}$ are optimized simultaneously to train their network parameters. The proposed DLU is taken at the end of each training round to update the pseudo-labels for the saliency network; meanwhile, D^{Label}_{Sal} and D^{Label}_{NonSal} in Eqs.1 and 2 are also updated using the improved Sal_{pred}. This allows the network to make efficient use of the updated pseudo labels. As suggested by Fig. 4 and Table 3, the use of our attentive training strategy leads to more significant error reduction compared with the backbone solely using uniform weight or attentive weight for each instance. Besides, when our DSU and the attentive training strategy are both incorporated to the backbone, better results are achieved.

4 EXPERIMENTS AND ANALYSES

4.1 DATASETS AND EVALUATION METRICS

Extensive experiments are conducted over four large-scale RGB-D SOD benchmarks. NJUD (Ju et al., 2014) in its latest version consists of 1,985 samples, that are collected from the Internet and 3D movies; NLPR (Peng et al., 2014) has 1,000 stereo images collected with Microsoft Kinect; STERE (Niu et al., 2012) contains 1,000 pairs of binocular images downloaded from the Internet; DUTLF-Depth (Piao et al., 2019) has 1,200 real scene images captured by a Lytro2 camera. We follow the setup of (Fan et al., 2020a) to construct the training set, which includes 1,485 samples from NJUD and 700 samples from NLPR, respectively. Data augmentation is also performed by randomly rotating, cropping and flipping the training images to avoid potential overfitting. The remaining images are reserved for testing.

Here, five widely-used evaluation metrics are adopted: E-measure (E_ξ) (Fan et al., 2018), weighed F-measure (F^w_β) (Margolin et al., 2014), F-measure (F_β) (Achanta et al., 2009), Mean Absolute Error (MAE or $\mathcal{M}$) (Borji et al., 2015), and inference time(s) or FPS (Frames Per Second).

Implementation details are presented in the Appx A.3.

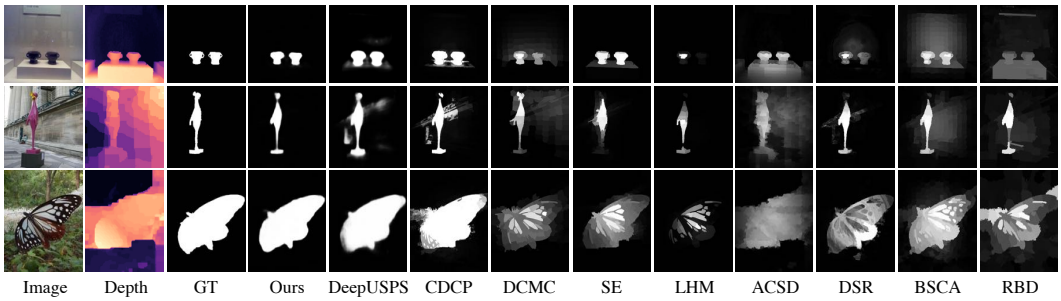


Figure 5: Qualitative comparison with unsupervised saliency detection methods. GT denotes ground-truth.

Table 1: Quantitative comparison with unsupervised saliency detection methods. 'Backbone' refers to the feature extraction network of DeepLabV3+ (Chen et al., 2018) adopted in our pipeline, *i.e.* the one without the two proposed key components. The RGB-based methods are specifically marked by †. UnSOD is shorthand for unsupervised SOD.

	*	Inference Time(s)↓	NJUD				NLPR				STERE				DUTLF-Depth			
			$E_\xi\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$
Handcrafted UnSOD	RBD† (Zhu et al., 2014)	0.189	.684	.387	.556	.256	.765	.388	.590	.211	.730	.443	.610	.223	.733	.447	.619	.222
	MST† (Tu et al., 2016)	0.030	.670	.291	.436	.281	.762	.257	.491	.199	.681	.312	.447	.269	.678	.254	.401	.279
	BSCA† (Qin et al., 2015)	2.665	.756	.446	.623	.216	.745	.376	.554	.178	.803	.497	.676	.179	.808	.479	.682	.181
	DSR† (Li et al., 2013)	0.376	.739	.436	.594	.196	.757	.451	.545	.120	.785	.486	.645	.165	.797	.478	.640	.164
	ACSD (Ju et al., 2014)	0.718	.790	.448	.696	.198	.751	.327	.547	.171	.793	.425	.661	.200	.250	.210	.188	.668
	DES (Cheng et al., 2014)	7.790	.421	.241	.165	.448	.735	.259	.583	.301	.673	.383	.592	.297	.733	.386	.668	.280
	LHM (Peng et al., 2014)	2.130	.722	.311	.625	.201	.772	.320	.520	.119	.772	.360	.703	.171	.767	.350	.659	.174
	GP (Ren et al., 2015)	12.98	.730	.323	.666	.204	.813	.347	.670	.144	.785	.371	.710	.182	-	-	-	-
	CDB (Liang et al., 2018)	0.600	.752	.408	.650	.200	.810	.388	.618	.108	.808	.436	.713	.166	-	-	-	-
	SE (Guo et al., 2016)	1.570	.780	.518	.715	.164	.853	.578	.701	.085	.825	.546	.747	.143	.730	.339	.474	.196
	DCMC (Cong et al., 2016)	1.210	.796	.506	.715	.167	.684	.265	.328	.196	.832	.529	.743	.148	.712	.290	.406	.243
	MB (Zhu et al., 2017a)	-	.643	.369	.492	.202	.814	.574	.637	.089	.693	.455	.572	.178	.691	.464	.577	.156
	CDCP (Zhu et al., 2017b)	5.720	.751	.522	.618	.181	.785	.512	.591	.114	.797	.596	.666	.149	.794	.530	.633	.159
Deep UnSOD	USD† (Zhang et al., 2018)	0.0180	.768	.565	.630	.163	.786	.536	.580	.119	.796	.572	.670	.146	.795	.545	.650	.157
	DeepUSPS† (Nguyen et al., 2019)	0.0292	.771	.576	.647	.159	.809	.622	.639	.088	.806	.632	.682	.124	.798	.573	.654	.149
	Backbone	0.0286	.759	.510	.627	.186	.760	.479	.570	.126	.794	.555	.666	.158	.798	.512	.644	.167
	Δ gains	-	↑5%	↑17%	↑15%	↓27%	↑16%	↑37%	↑31%	↓48%	↑8%	↑22%	↑16%	↓37%	↑7%	↑27%	↑18%	↓36%
	Ours	0.0286	**.797**	**.597**	**.719**	**.135**	**.879**	**.657**	**.745**	**.065**	**.857**	**.678**	**.774**	**.099**	**.854**	**.650**	**.763**	**.107**

4.2 Comparison with the State-of-the-arts

Our approach is compared with 15 unsupervised SOD methods, *i.e.*, without using any human annotations. Their results are either directly furnished by the authors of the respective papers, or generated by re-running their original implementations. In this paper, we make the first attempt to address deep-learning-based unsupervised RGB-D SOD. Since existing unsupervised RGB-D methods are all based on handcrafted feature representations, we additionally provide several RGB-based methods (*e.g.*, USD and DeepUSPS) for reference purpose. This gives more observational evidences for the related works. These RGB-based methods are specifically marked by † in Table 1.

Quantitative results are listed in Table 1, where our approach clearly outperforms the state-of-the-art unsupervised SOD methods in both RGB-D and RGB only scenarios. This is due to our DSU framework that leads to trustworthy supervision signals for saliency network. Furthermore, our network design leads to a light-weight architecture in the inference stage, shown as the black dashed portion in Fig. 2. This enables efficient & effective detection of salient objects and brings a large-margin improvement over the backbone network without introducing additional depth input and computational costs, as shown in Fig. 1 and Table 1. Qualitatively, saliency predictions of competing methods are exhibited in Fig. 5. These results consistently proves the superiority of our method.

4.3 Analysis of the Empirical Results

The focus here is on the evaluation of the contributions from each of the components, and the evaluation of the obtained pseudo-labels as intermediate results.

Effect of each component. In Table 2, we conduct ablation study to investigate the contribution of each component. To start with, we consider the backbone (a), where the saliency network is trained with initial pseudo-labels. As our proposed ATS and DSU are gradually incorporated, increased performance has been observed on both datasets. Here we first investigate the benefits of ATS by applying it to the backbone and obtaining (b). We observe increased F-measure scores of 3% and 5.8% on NJUD and NLPR benchmarks, respectively. This clearly shows that the proposed ATS can effectively improve the utilization of reliable pseudo-labels, by re-weighting ambiguous training data. We then investigate in detail all components in our DSU strategy. The addition of the entire DSU leads to (f), which significantly improves the F-measure metric by 11.3% and 23.5% on each

of the benchmarks, while reducing the MAE by 22.4% and 41.9%, respectively. This verifies the effectiveness of the DSU strategy to refine and update pseudo-labels. Moreover, as we gradually exclude the consistency loss l_{con} (row (e)) and HA operation (row (d)), degraded performances are observed on both datasets. For an extreme case where we remove the DLU and only maintain CRF to refine pseudo-labels, it is observed that much worse performance is achieved. These results consistently demonstrate that all components in the DSU strategy are beneficial for generating more accurate pseudo-labels.

Table 2: Ablation study of our deep unsupervised RGB-D SOD pipeline, using the F-measure and MAE metrics.

Index	Model Setups		NJUD		NLPR	
			$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
(a)	Backbone		0.627	0.186	0.570	0.126
(b)	(a) + attentive training strategy		0.646	0.174	0.603	0.112
(c)	DSU strategy	(b) + CRF	0.674	0.160	0.663	0.093
(d)		(b) + DSU (w/o l_{con}&HA)	0.703	0.141	0.716	0.074
(e)		(b) + DSU (w/o l_{con})	0.712	0.137	0.735	0.068
(f)		(b) + DSU (**Ours**)	0.719	0.135	0.745	0.065

We also display the visual evidence of the updated pseudo-labels obtained from the DSU strategy in Fig. 6. It is shown that the initial pseudo-labels unfortunately tend to miss important parts as well as fine-grained details. The application of CRF helps to filter away background noises, while salient parts could still be missing out. By adopting our DSU and attentive training strategy, the missing parts could be retrieved in the updated pseudo-labels, with the object silhouette also being refined. These numerical and visual results consistently verify the effectiveness of our pipeline in deep unsupervised RGB-D saliency detection.

Detailed analysis of the ATS. The influence of the attentive training strategy (ATS) is studied in detail here. The error reduction curves in Fig. 4 have shown that the use of ATS could lead to greater test error reduction compared with the backbone that uses uniform weight or attentive weight for each instance. When combining the ATS with the DSU strategy as shown in Table 3, the ATS also enables the network to learn more efficiently than using uniform weight or attentive weight alone, by comparing 'Ours' with 'Setting 1' and 'Setting 2'. In addition, we discuss the effect of different alternation intervals τ in ATS. As listed in Table 3, the larger or smaller interval leads to inferior performance due to the insufficient or excessive learning of saliency models.

Table 3: Analyzing attentive training strategy (ATS) with different settings. 'Setting 1' is backbone + uniform weight + DSU, and 'Setting 2' is backbone + attentive weight + DSU. The last four columns show backbone + ATS + DSU with different alternation interval τ.

*		Analysis of the ATS		**Ours**	Analysis of the interval τ		
		Setting 1	Setting 2	$\tau = 3$	$\tau = 1$	$\tau = 5$	$\tau = 10$
NJUD	$F_\beta \uparrow$	0.695	0.707	**0.719**	0.711	0.712	0.687
	$\mathcal{M} \downarrow$	0.148	0.142	**0.135**	0.143	0.138	0.147
NLPR	$F_\beta \uparrow$	0.706	0.737	**0.745**	0.742	0.739	0.698
	$\mathcal{M} \downarrow$	0.071	0.067	**0.065**	0.069	0.066	0.073

Table 4: Internal mean absolute errors, each is evaluated between current pseudo-labels and the corresponding true labels (only used for evaluation purpose) during the training process.

Pseudo-label Update	Initial	Update 1	Update 2	Update 3	Update 4
Mean absolute error	0.162	0.124	0.117	0.116	0.116

Table 5: Comparison of different pseudo-label generation variants. 'CRF' refers to fully-connected CRF. 'OTSU' represents the standard Otsu image thresholding method.

Label Accuray	$E_\xi \uparrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
Initial pseudo-label	0.760	0.526	0.614	0.162
Initial pseudo-label + CRF	0.763	0.578	0.634	0.144
Our DSU	**0.792**	**0.635**	**0.708**	**0.116**
Depth map	0.419	0.284	0.164	0.414
Depth map + OTSU	0.465	0.398	0.429	0.332

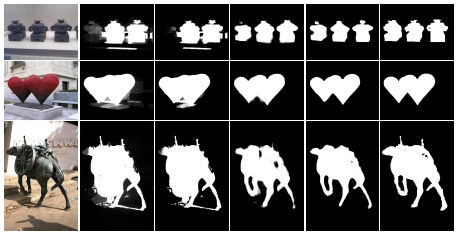


Figure 6: Visual examples of the intermediate pseudo-labels used in our approach. 'Initial' shows the initial pseudo-labels generated by traditional handcrafted method. '+CRF' refers to the pseudo-labels after applying fully-connected CRF. Update 1&2 represent the updated pseudo-labels produced in our pipeline over two training rounds. 'GT' means the ground truth, used for reference purpose only.

Analysis of pseudo-labels. We analyze the quality of pseudo-labels over the training process in Table 4, where the mean absolute error scores between the pseudo-labels at different update rounds and the ground-truth labels are reported. It is observed that the quality of pseudo-labels is significantly improved during the first two rounds, which then remains stable in the consecutive rounds. Fig. 6 also shows that the initial pseudo-label is effectively refined, where the updated pseudo-label is close to the true label. This provides more reliable guiding signals for training the saliency network.

Comparison of different pseudo-label variants. In Table 5, we investigate other possible pseudo-label generation variants, including, initial pseudo-label with CRF refinement, raw depth map, and raw depth map together with OTSU thresholding (Otsu, 1979). It is shown that, compared with the direct use of CRF, our proposed DSU is able to provide more reliable pseudo-labels, by disentangling depth to promote saliency. It is worth noting that a direct application of the raw depth map or together with an OTSU adaptive thresholding of the depth map, may nevertheless lead to awful results. We conjecture this is because of the large variations embedded in raw depth, and the fact that foreground objects may be affected by nearby background stuffs that are close in depth.

Is depth-disentanglement necessary? In our DSU, we disentangle the raw depth map into valuable salient parts as well as redundant non-salient areas to simultaneously highlight the salient regions and suppress the non-salient negative responses. To verify the necessity of depth disentanglement, we remove the D_{NonSal} branch in the depth-disentangled network and only use the D_{Sal} to update pseudo-labels in DLU. Empirical results in Table 6 reveal that removing D_{NonSal} (row (a)) leads to significant inferior performance when comparing with the original DSU (row (c)), which proves the superiority of our DSU design.

Table 6: Discussion on different pseudo-label updating settings.

Different Pseudo-label Updating Setups	NJUD		NLPR	
	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
(a) DSU w/o D_{NonSal}	0.698	0.143	0.715	0.072
(b) DSU using D_{Sal}^{Label}, D_{NonSal}^{Label}	0.691	0.154	0.707	0.079
(c) DSU using D_{Sal}, D_{NonSal} (**Ours**)	0.719	0.135	0.745	0.065

How about using raw depth labels in DSU? We also consider to directly utilize the raw disentangled depth labels, *i.e.,* D_{Sal}^{Label} and D_{NonSal}^{Label}, to update pseudo-labels in DSU. However, a direct adoption of the raw depth labels does not lead to good performance, as empirically exhibited in Table 6 (b). The DSU using raw labels performs worse compared to our original DSU design in row (c). This is partly due to the large value variations and intrinsic label noises in raw depth map as discussed before. On the other hand, benefiting from the capability of deep-learning-based networks to learn general representations from large-scale data collections, using D_{Sal} and D_{NonSal} in our DSU strategy is able to eliminate the potential biases in a single raw depth label. Visual evidences as exhibited in the 4th-7th columns of Fig. 3 also support our claim.

4.4 Application to Fully-supervised Setting

To show the generic applicability of our approach, a variant of our DSU is applied to several cutting-edge fully-supervised RGB-D models to improve their performance. This is made possible by re-defining the quantities $D_{Sal}^{Label} = \mathcal{S}_{GT} \otimes D_{map}$ and $D_{NonSal}^{Label} = (1 - \mathcal{S}_{GT}) \otimes D_{map}$, with $\mathcal{S}_{GT}$ being the ground-truth saliency. Then the saliency network (*i.e.,* existing RGB-D models) and the depth-disentangled network are trained by $\mathcal{S}_{GT}$ and the new D_{Sal}^{Label} and D_{NonSal}^{Label}, respectively. After training, the proposed DLU is engaged to obtain the final improved saliency. In Table 7, we report the original results of four RGB-D methods and the new results of incorporating our DSU strategy on two popular benchmarks. It is observed that our supervised variants have consistent performance improvement comparing to each of existing models. For example, the average MAE scores of four RGB-D methods on NJUD and NLPR benchmarks are reduced by 13.9% and 16.6%, respectively. We attribute the performance improvement to our DSU strategy that can exploit the learned D_{Sal} to facilitate the localization of salient object regions in a scene, as well as suppress the redundant background noises by subtracting D_{NonSal}. More experiments are presented in the Appx A.2.

Table 7: Applying our DSU to existing fully-supervised RGB-D SOD methods.

*	NJUD		NLPR	
	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
DMRA (Piao et al., 2019)	0.872	0.051	0.855	0.031
+ Our DSU	0.893	0.044	0.879	0.026
CMWN (Li et al., 2020c)	0.878	0.047	0.859	0.029
+ Our DSU	0.901	0.041	0.882	0.025
FRDT (Zhang et al., 2020e)	0.879	0.048	0.868	0.029
+ Our DSU	0.903	0.038	0.901	0.023
D3Net (Fan et al., 2020a)	0.863	0.047	0.857	0.030
+ Our DSU	0.891	0.043	0.888	0.025

5 Conclusion and Outlook

This paper tackles the problem of deep unsupervised RGB-D saliency detection, which is the first such effort to our knowledge. Our key insight is to internally engage the pseudo-labels, consolidated by two critical ingredients, namely depth-disentangled saliency update in iteratively fine-tuning the pseudo-labels & attentive training strategy in alleviating the issue of noisy pseudo-labels. Extensive empirical experiments demonstrate the superior performance and realtime efficiency of our approach. For future work, our plan is to explore the scenarios involving partial labels.

Ethics Statement. This work does not involve any negative societal and ethical impacts, and we believe the most proximate impacts of this work will be positive. In particular, class-agnostic unsupervised segmentation problems like RGB-D salient object detection problem present a major obstacle to the requirement of a significant amount of laborious pixel-level human annotations. The proposed Depth-disentangled Saliency Update framework can produce more reliable supervision signal for training saliency network, which exhibits an appealing prospect and tradeoff between model accuracy and human consumption. Furthermore, our method achieves fast inference speed and good generic applicability, which is useful and appealing for light-weight deployment in practice.

Reproducibility Statement. As for social responsibility, the project team member will responsible for the quality of the project, like no cheating, being honest and academic integrity. We also promise that our source code with easy-to-use instruction will be publicly available upon paper acceptance.

REFERENCES

Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604, 2009.

Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015.

John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

Chenglizhao Chen, Jipeng Wei, Chong Peng, Weizhong Zhang, and Hong Qin. Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion. *IEEE Transactions on Image Processing*, 29:4296–4307, 2020a.

Chenglizhao Chen, Jipeng Wei, Chong Peng, and Hong Qin. Depth-quality-aware salient object detection. *IEEE Transactions on Image Processing*, 30:2350–2363, 2021a.

Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for RGB-D salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3051–3060, 2018.

Hao Chen and Youfu Li. Three-stream attention-aware network for RGB-D salient object detection. *IEEE Transactions on Image Processing*, 28(6):2825–2835, 2019.

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. *Pattern Recognition*, 86:376–385, 2019.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, pp. 801–818, 2018.

Qian Chen, Ze Liu, Yi Zhang, Keren Fu, Qijun Zhao, and Hongwei Du. RGB-D salient object detection via 3D convolutional neural. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021b.

Shuhan Chen and Yun Fu. Progressively guided alternate refinement network for RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision*, 2020.

Zuyao Chen, Runmin Cong, Qianqian Xu, and Qingming Huang. DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection. *IEEE Transactions on Image Processing*, 2020b.

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*, pp. 23–27, 2014.

Runmin Cong, Jianjun Lei, Changqing Zhang, Qingming Huang, Xiaochun Cao, and Chunping Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. *IEEE Signal Processing Letters*, 23(6):819–823, 2016.

Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Nam Ling. HSCS: Hierarchical sparsity based co-saliency detection for RGBD images. *IEEE Transactions on Multimedia*, 21(7):1660–1671, 2018.

Runmin Cong, Jianjun Lei, Huazhu Fu, Junhui Hou, Qingming Huang, and Sam Kwong. Going from RGB to RGBD saliency: A depth-guided transformation model. *IEEE Transactions on Systems, Man, and Cybernetics*, 50(8):3627–3639, 2020.

Karthik Desingh, K Madhava Krishna, Deepu Rajan, and CV Jawahar. Depth really matters: Improving visual salient region detection with depth. In *British Machine Vision Conference*, 2013.

Nancy M Dixon. *The organizational learning cycle: How we can learn collectively*. Gower Publishing, Ltd., 1999.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 698–704, 2018.

Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020a.

Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In *Proceedings of the European Conference on Computer Vision*, 2020b.

David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. Local background enclosure for RGB-D salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2343–2350, 2016.

Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3052–3062, 2020.

Silvio Giancola, Matteo Valenti, and Remo Sala. *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies*. Springer, 2018.

Jingfan Guo, Tongwei Ren, and Jia Bei. Salient object detection for RGB-D image via saliency evolution. In *IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2016.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

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*, pp. 3203–3212, 2017.

Kuang-Jui Hsu, Yen-Yu Lin, Yung-Yu Chuang, et al. Co-attention cnns for unsupervised object co-segmentation. In *IJCAI*, pp. 748–756, 2018a.

Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, Xiaoning Qian, and Yung-Yu Chuang. Unsupervised cnn-based co-saliency detection with graphical optimization. In *European Conference on Computer Vision*, pp. 485–501, 2018b.

Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, and Li Cheng. Calibrated rgb-d salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9471–9481, 2021.

Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *IEEE International Conference on Image Processing*, pp. 1115–1119, 2014.

ByoungChul Ko, Soo Yeong Kwak, and Hyeran Byun. SVM-based salient region(s) extraction method for image retrieval. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 2, pp. 977–980, 2004.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Chongyi Li, Runmin Cong, Sam Kwong, Junhui Hou, Huazhu Fu, Guopu Zhu, Dingwen Zhang, and Qingming Huang. ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection. *IEEE Transactions on Cybernetics*, PP(99):1–13, 2020a.

Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy. RGB-D salient object detection with cross-modality modulation and selection. In *Proceedings of the European Conference on Computer Vision*, 2020b.

Gongyang Li, Zhi Liu, Linwei Ye, Yang Wang, and Haibin Ling. Cross-modal weighting network for RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision*, 2020c.

Gongyang Li, Zhi Liu, Minyu Chen, Zhen Bai, Weisi Lin, and Haibin Ling. Hierarchical alternate interaction network for RGB-D salient object detection. *IEEE Transactions on Image Processing*, 2021.

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*, pp. 2976–2983, 2013.

Fangfang Liang, Lijuan Duan, Wei Ma, Yuanhua Qiao, Zhi Cai, and Laiyun Qing. Stereoscopic saliency model using contrast and depth-guided-background prior. *Neurocomputing*, 275:2227–2238, 2018.

Guibiao Liao, Wei Gao, Qiuping Jiang, Ronggang Wang, and Ge Li. MMNet: Multi-stage and multi-scale fusion network for RGB-D salient object detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2436–2444, 2020.

Huiying Liu, Shuqiang Jiang, Qingming Huang, and Changsheng Xu. A generic virtual content insertion system based on visual attention analysis. In *Proceedings of the 16th ACM International Conference on Multimedia*, pp. 379–388, 2008.

Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3917–3926, 2019.

Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3089–3098, 2018.

Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for RGB-D saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13756–13765, 2020.

Ao Luo, Xin Li, Fan Yang, Zhicheng Jiao, Hong Cheng, and Siwei Lyu. Cascade graph neural networks for RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision*, 2020.

Mingcan Ma, Changqun Xia, and Jia Li. Pyramidal feature shrinking for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2311–2318, 2021.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2014.

Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. DeepUSPS: Deep robust unsupervised saliency prediction via self-supervision. In *Advances in Neural Information Processing Systems*, pp. 204–214, 2019.

Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 454–461, 2012.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision*, 2020.

James W Peltier, Amanda Hay, and William Drago. The reflective learning continuum: Reflecting on reflection. *Journal of marketing education*, 27(3):250–263, 2005.

Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. RGBD salient object detection: a benchmark and algorithms. In *Proceedings of the European Conference on Computer Vision*, pp. 92–109, 2014.

Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7254–7263, 2019.

Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9060–9069, 2020.

Yao Qin, Huchuan Lu, Yiqun Xu, and He Wang. Saliency detection via cellular automata. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 110–119, 2015.

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.

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*, pp. 25–32, 2015.

Ling Shao and Michael Brady. Specific object retrieval based on salient regions. *Pattern Recognition*, 39(10): 1932–1948, 2006.

Riku Shigematsu, David Feng, Shaodi You, and Nick Barnes. Learning RGB-D salient object detection using background enclosure, depth contrast, and top-down features. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 2749–2757, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi Li. Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1407–1417, 2021.

Chung-Chi Tsai, Weizhi Li, Kuang-Jui Hsu, Xiaoning Qian, and Yen-Yu Lin. Image co-saliency detection and co-segmentation via progressive joint optimization. *IEEE Transactions on Image Processing*, 28(1):56–71, 2018.

Chung-Chi Tsai, Kuang-Jui Hsu, Yen-Yu Lin, Xiaoning Qian, and Yung-Yu Chuang. Deep co-saliency detection via stacked autoencoder-enabled fusion and self-trained cnns. *IEEE Transactions on Multimedia*, 22(4): 1016–1031, 2019.

Wei-Chih Tu, Shengfeng He, Qingxiong Yang, and Shao-Yi Chien. Real-time salient object detection with a minimum spanning tree. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2334–2342, 2016.

Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4894–4903, 2018.

Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3907–3916, 2019a.

Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7264–7273, 2019b.

Binwei Xu, Haoran Liang, Ronghua Liang, and Peng Chen. Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3004–3012, 2021.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pp. 2048–2057, 2015.

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*, pp. 4048–4056, 2017.

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*, pp. 9029–9038, 2018.

Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8582–8591, 2020a.

Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12546–12555, 2020b.

Miao Zhang, Sun Xiao Fei, Jie Liu, Shuang Xu, Yongri Piao, and Huchuan Lu. Asymmetric two-stream architecture for accurate RGB-D saliency detection. In *Proceedings of the European Conference on Computer Vision*, 2020c.

Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for RGB-D saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3472–3481, 2020d.

Miao Zhang, Yu Zhang, Yongri Piao, Beiqi Hu, and Huchuan Lu. Feature reintegration over differential treatment: A top-down and adaptive fusion network for RGB-D salient object detection. In *ACM International Conference on Multimedia*, 2020e.

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 Conference on Computer Vision and Pattern Recognition*, pp. 3927–3936, 2019.

Jiawei Zhao, Yifan Zhao, Jia Li, and Xiaowu Chen. Is depth really necessary for salient object detection? In *ACM International Conference on Multimedia*, 2020a.

Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision*, 2020b.

Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. RGB-D salient object detection: A survey. *Computational visual media*, pp. 1–33, 2021.

Chunbiao Zhu, Ge Li, Xiaoqiang Guo, Wenmin Wang, and Ronggang Wang. A multilayer backpropagation saliency detection algorithm based on depth mining. In *International Conference on Computer Analysis of Images and Patterns*, pp. 14–23, 2017a.

Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. An innovative salient object detection using center-dark channel prior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1509–1515, 2017b.

Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. PDNet: Prior-model guided depth-enhanced network for salient object detection. In *IEEE International Conference on Multimedia and Expo*, pp. 199–204, 2019.

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*, pp. 2814–2821, 2014.

A APPENDIX

In this appendix, we first elaborate on the evaluation metrics used in this paper, in Appx A.1. Then, more details and experiments of our DSU are presented in Appx A.2, including the detailed structures of D-Sal and D-NonSal CNNs, more experiments on unsupervised and fully-supervised scenarios, and other pseudo-label initialization. These empirical results consistently demonstrate the effectiveness and scalability of our approach. In Appx A.3, we illustrate on the implementation details of our unsupervised pipeline. In Appx A.4, we discuss the potential limitations which can be addressed in the near future. Finally, in Appx A.5, we additionally conduct some interesting experiments to further verify the efficiency of our method.

A.1 EVALUATION METRICS

We adopt four widely-used metrics in SOD to evaluate the performance of saliency models, *i.e.*, E-measure, F-measure, weighted F-measure and MAE. *The lower the MAE, the better. For other metrics, the higher score is better.* Concretely, F-measure is an overall performance measurement and is computed by the weighted harmonic mean of the precision and recall:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{9}$$

where β^2 is set to 0.3 to emphasize the precision (Achanta et al., 2009). We use different fixed [0, 255] thresholds to compute the F-measure metric. Weighted F-measure (F_β^w) (Margolin et al., 2014) is its weighted measurement. MAE ($\mathcal{M}$) (Borji et al., 2015) represents the average absolute difference between the predicted saliency map and ground truth. It is used to calculate how similar a normalized saliency maps $\mathcal{S} \in [0,1]^{W \times H}$ is compared to the ground truth $\mathcal{G} \in [0,1]^{W \times H}$:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\mathcal{S}(x,y) - \mathcal{G}(x,y)|, \tag{10}$$

where W and H denote the width and height of $\mathcal{S}$, respectively. Enhanced-alignment measure (E_ξ) (Fan et al., 2018) is proposed based on cognitive vision studies to capture image-level statistics and their local pixel matching information, as in

$$E_\xi = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi_s(i,j), \tag{11}$$

where $\phi_s(\cdot)$ is the enhanced-alignment matrix (Fan et al., 2018), which reflects the correlation between $\mathcal{S}$ and $\mathcal{G}$ after subtracting their global means.

A.2 MORE ANALYSES ON THE PROPOSED DSU

Detailed structures. The detailed structures of the D-Sal CNN and D-NonSal CNN are illustrated in Fig. 7. The 'Conv 3×3' represents a convolutional layer with the kernel size of 3.

Visual results for unsupervised RGB-D SOD models. More visual comparisons with unsupervised RGB-D saliency models are shown in Fig. 8, where our unsupervised pipeline generates promising saliency prediction that is close to true label.

Comparison with current state-of-the-art RGB-D SOD methods under fully-supervised scenario. As discussed in Sec. 4.4 of the main text, our method can boost the performance of existing SOD models greatly, by applying DSU to existing methods. In this section, we further compare our fully-supervised variant of DSU with 25 existing state-of-the-art RGB-D saliency models. We adopt the same feature-extraction baseline as in BBSNet (Fan et al., 2020b) and DCF (Ji et al., 2021) to build the saliency network in the DSU framework. Numerical results in Table 8 and visual results in Fig. 9 consistently show the superiority and scalability of our method.

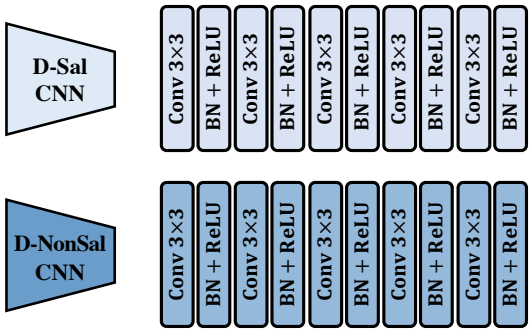


Figure 7: Detailed structure of D-Sal and D-NonSal CNNs in DSU. BN means batch normalization operation.

Pseudo-label Initialization. We investigate other initial pseudo-labels (*e.g.*, DCMC (Cong et al., 2016)), to further verify the effectiveness of our proposed method. The superior performance is also observed, by comparing backbone with ours: 0.167 *vs.* 0.146 on NJUD, 0.119 *vs.* 0.084 on NLPR, 0.142 *vs.* 0.119 on STERE and 0.146 *vs.* 0.128 on DUTLF-Depth, using MAE metric.

A.3 IMPLEMENTATION DETAILS

The proposed deep unsupervised pipeline is implemented with PyTorch and trained using a single Tesla P40 GPU. Both Saliency and Depth Networks employ the same feature extraction backbone (DeepLabv3+) (Chen et al., 2018), which is equipped with ResNet-50 (He et al., 2016), with initial parameters pre-trained in ImageNet (Krizhevsky et al., 2012). All training & testing images are uniformly resized to the size of 352×352. Throughout training, the learning rate is set to 1×10^{-4}, and the Adam optimizer is used with a mini-batch size of 10. Our approach is trained in an unsupervised manner, *i.e.*, without any human annotations, where initial pseudo-labels are the outputs of handcrafted RGB-D saliency method CDCP (Zhu et al., 2017b). As for the proposed attentive training strategy, its alternation interval is set to 3, amounting to $2\tau = 6$ epochs in a training round. During inference, our approach directly predicts saliency maps based on an RGB image, without access to any depth map.

A.4 LIMITATION AND DISCUSSION

As demonstrated in this paper, our *deep unsupervised RGB-D saliency detection* approach achieves an appealing trade-off between detection accuracy and human consumption. However, due to the lack of accurate pixel-level annotations, the model still fails to comprehensively detect the fine-grained details, *i.e.*, the edges of salient objects. To tackle this challenge, a doable solution is to introduce auxiliary edge constraint. For example, the edge detection loss can be employed to low-level features of the model, which forces the model to produce discriminative features highlighting object details (Zhang et al., 2020b; Liu et al., 2019). The edge maps can be generated by classical Canny operator (Canny, 1986) in an unsupervised manner. Hopefully this could encourage more inspirations and contributions to this community and further pave the way for its booming future.

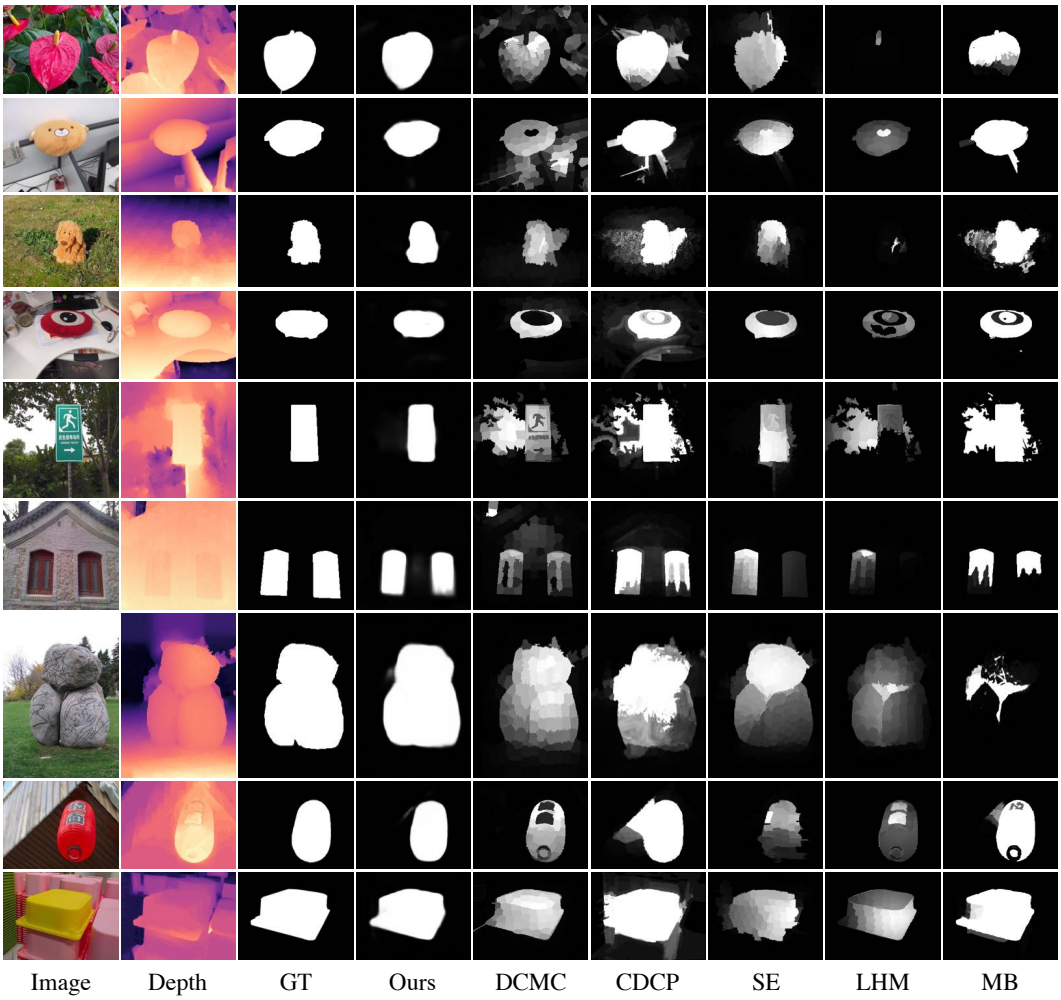


Figure 8: Qualitative comparison with unsupervised RGB-D SOD models. GT means ground-truth.

Table 8: Quantitative results of fully-supervised RGB-D saliency detection methods. The best results are highlighted in **boldface**. When evaluating the newly released DUTLF-Depth dataset, the specific setup used by (Piao et al., 2019) is adopted to make a fair comparison.

Method	NJUD				NLPR				STERE				DUTLF-Depth			
	$E_\xi \uparrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
CTMF (Han et al., 2017)	.864	.732	.788	.085	.869	.691	.723	.056	.841	.747	.771	.086	.884	.690	.792	.097
DF (Qu et al., 2017)	.818	.552	.744	.151	.838	.524	.682	.099	.691	.596	.742	.141	.842	.542	.748	.145
PCA (Chen & Li, 2018)	.896	.811	.844	.059	.916	.772	.794	.044	.887	.801	.826	.064	.858	.696	.760	.100
TANet (Chen & Li, 2019)	.893	.812	.844	.061	.916	.789	.795	.041	.893	.804	.835	.060	.866	.712	.779	.093
PDNet (Zhu et al., 2019)	.890	.798	.832	.062	.876	.659	.740	.064	.880	.799	.813	.071	.861	.650	.757	.112
MMCI (Chen et al., 2019)	.878	.749	.813	.079	.871	.688	.729	.059	.873	.757	.829	.068	.855	.636	.753	.113
CPFP (Zhao et al., 2019)	.895	.837	.850	.053	.924	.820	.822	.036	.912	.808	.830	.051	.814	.644	.736	.099
DMRA (Piao et al., 2019)	.908	.853	.872	.051	.942	.845	.855	.031	.923	.841	.876	.049	.927	.858	.883	.048
SSF (Zhang et al., 2020d)	.913	.871	.886	.043	.949	.874	.875	.026	.921	.850	.867	.046	.946	.894	.914	.034
A2dele (Piao et al., 2020)	.897	.851	.874	.051	.945	.867	.878	.028	.915	.855	.874	.044	.924	.864	.890	.043
JL-DCF (Fu et al., 2020)	-	-	-	-	.954	.882	.878	**.022**	.919	.857	.869	.040	.931	.863	.883	.043
S2MA (Liu et al., 2020)	-	-	-	-	.938	.852	.853	.030	.907	.825	.855	.051	.921	.861	.866	.044
UCNet (Zhang et al., 2020a)	-	-	-	-	.953	.878	.890	.025	.922	.867	.885	.039	.903	.821	.856	.056
FRDT (Zhang et al., 2020e)	.917	.862	.879	.048	.946	.863	.868	.029	.925	.858	.872	.042	.941	.878	.902	.039
D3Net (Fan et al., 2020a)	.913	.860	.863	.047	.943	.854	.857	.030	.920	.845	.855	.046	.847	.668	.756	.097
HDFNet (Pang et al., 2020)	.915	.879	.893	.038	.948	.869	.878	.027	.925	.863	.879	.040	.934	.865	.892	.040
CMWNet (Li et al., 2020c)	.910	.855	.878	.047	.940	.856	.859	.029	.917	.847	.869	.043	.916	.831	.866	.056
DANet (Zhao et al., 2020b)	-	-	-	-	.949	.858	.871	.028	.914	.830	.858	.047	.925	.847	.884	.047
PGAR (Chen & Fu, 2020)	.915	.871	.893	.042	.955	.881	.885	.024	.919	.856	.880	.041	.944	.889	.914	.035
ATSA (Zhang et al., 2020c)	.921	.883	.893	.040	.945	.867	.876	.028	.919	.866	.874	.040	.947	.901	.918	.032
BBSNet (Fan et al., 2020b)	.924	.884	.902	**.035**	.952	.879	.882	.023	.925	.858	.885	.041	.833	.663	.774	.120
HAINe (Li et al., 2021)	.921	.887	.898	.038	.956	.890	.891	.024	.925	.871	.883	.040	.938	.883	.906	.038
RD3D (Chen et al., 2021b)	.918	.889	.900	.037	.956	.894	.890	**.022**	.926	.877	.885	.038	.949	.908	.914	**.031**
DSA^{2}F (Sun et al., 2021)	.923	.889	.901	.039	.950	.889	.897	.024	**.927**	.877	**.895**	.038	.949	.908	.924	.032
DCF (Ji et al., 2021)	.924	.893	.902	**.035**	.956	.892	.891	**.022**	**.927**	.873	.885	.039	**.952**	.909	**.926**	**.031**
Ours (fully supervised)	**.926**	**.894**	**.909**	.036	**.957**	**.897**	**.907**	**.022**	**.927**	**.882**	**.895**	**.037**	.951	**.911**	.918	**.031**

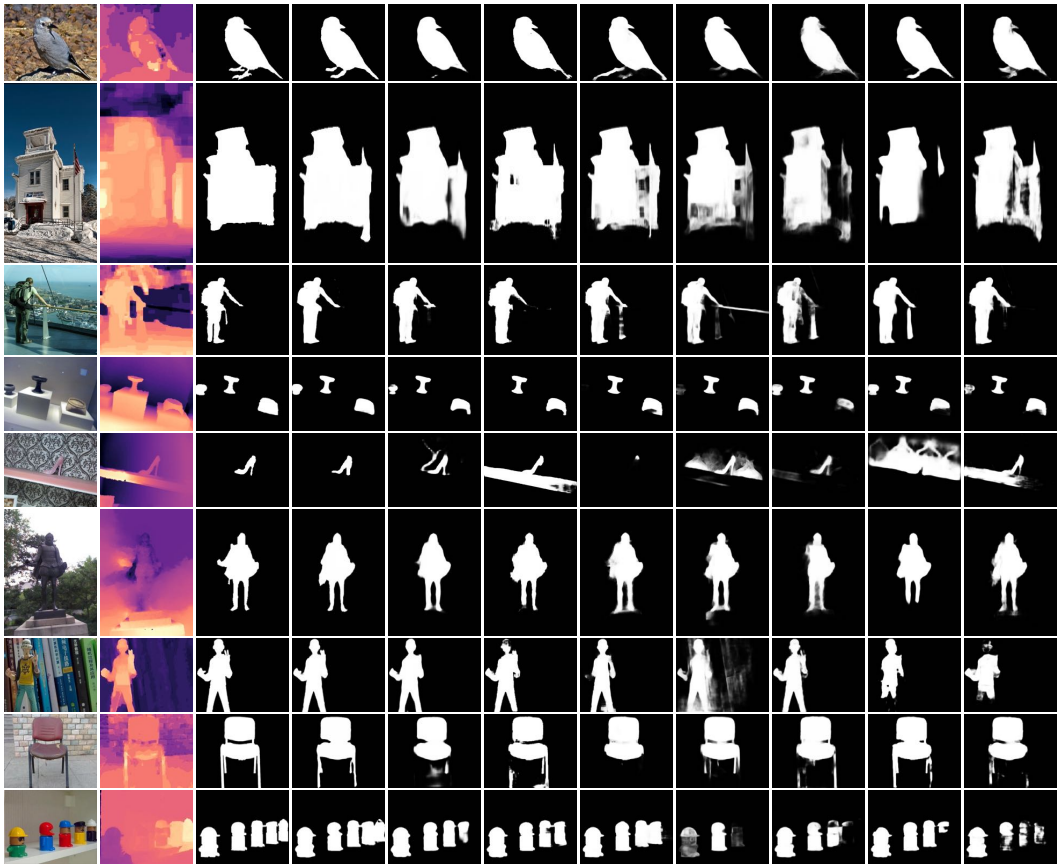


Figure 9: Qualitative comparison of fully-supervised RGB-D SOD methods. Obviously, our fully-supervised variant infers more appealing saliency maps compared to existing SOTA models, including DCF (Ji et al., 2021). DSA^{2}F (Sun et al., 2021), HAINet (Li et al., 2021), BBSNet (Fan et al., 2020b), S2MA (Liu et al., 2020), UCNet (Zhang et al., 2020a) and FRDT (Zhang et al., 2020e).

A.5 ADDITIONAL ANALYSIS

In this paper, our DSU design achieves appealing trade-off between model performance and complexity since our method has additional benefits of not introducing extra test-time cost and not relying on the depth map during the inference stage, as shown in Table 1 and Fig. 1. Meanwhile, our method can be easily adapted to integrate useful depth cues by engaging the proposed DSU and DLU. In Table 9, we conduct an interesting experiment, comparing the model performance, inference speed, FLOPs and parameter number, when using only RGB stream versus using RGB and depth simultaneously during inference. It is observed that introducing depth information can further improve the model performance, but at the same time, it will lead to the decrease of inference speed, and the increases of FLOPs and parameters. These results further demonstrate the effectiveness and efficiency of our DSU design.

Table 9: Quantitative results of our DSU when using only RGB stream versus using RGB and depth simultaneously during inference.

	Model Complexity			Model Performance (**NLPR**)			
	Speed↑	FLOPs↓	Parameters↓	E_ξ↑	F_β^w↑	F_β↑	$\mathcal{M}$↓
Our lightweight design (only using RGB)	35 FPS	17.87 G	47.85 MB	0.879	0.657	0.745	0.065
with additional depth (using RGB and depth)	21 FPS	45.16 G	96.53 MB	0.885	0.670	0.762	0.062