Texture-guided Saliency Distilling for Unsupervised Salient Object Detection

Huajun Zhou[1], Bo Qiao[1], Lingxiao Yang[1], Jianhuang Lai[1,2,3], Xiaohua Xie[1,2,3*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Guangdong Province Key Laboratory of Information Security Technology, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China

Abstract

Deep Learning-based Unsupervised Salient Object Detection (USOD) mainly relies on the noisy saliency pseudo labels that have been generated from traditional handcraft methods or pre-trained networks. To cope with the noisy labels problem, a class of methods focus on only easy samples with reliable labels but ignore valuable knowledge in hard samples. In this paper, we propose a novel USOD method to mine rich and accurate saliency knowledge from both easy and hard samples. First, we propose a Confidence-aware Saliency Distilling (CSD) strategy that scores samples conditioned on samples' confidences, which guides the model to distill saliency knowledge from easy samples to hard samples progressively. Second, we propose a Boundary-aware Texture Matching (BTM) strategy to refine the boundaries of noisy labels by matching the textures around the predicted boundaries. Extensive experiments on RGB, RGB-D, RGB-T, and video SOD benchmarks prove that our method achieves state-of-the-art USOD performance. Code is available at `www.github.com/moothes/A2S-v2`.

1. Introduction

Unsupervised Salient Object Detection (USOD) methods aim to correctly localize and precisely segment salient objects simultaneously without using manual annotations. Compared to the supervised methods, USOD methods can easily adapt to more practical scenarios (*e.g.*, industrial or medical images) where a large number of labeled images may be very hard to collect. Moreover, USOD methods also can assist some related methods for other tasks, *e.g.*, object recognition [16,44] and object detection [11,35]. However, diverse objects, complex backgrounds, and other challeng-
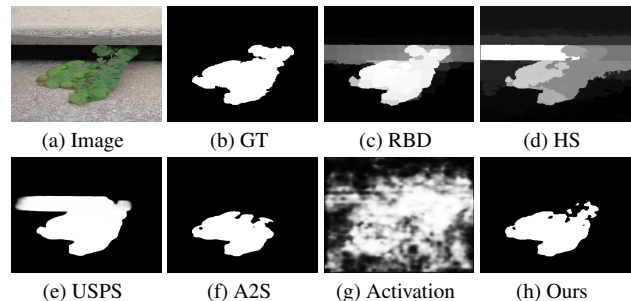
Figure 1. Saliency maps generated by USOD methods. Our result is generated based on the activation map of a deep network.

ing conditions bring severe challenges to USOD methods.

Most Deep Learning-based (DL-based) methods [21,29, 32, 55, 63, 66] base on the saliency cues extracted by traditional SOD methods (Fig. 1-c and 1-d). These handcrafted features related cues are employed as pseudo labels to train deep networks under certain constraints, *e.g.*, binary cross-entropy (BCE) loss. However, saliency cues by traditional methods usually shift away from target objects, especially in complex scenes. Moreover, conventional constraints, such as BCE loss, works well on fully-supervised SOD methods, but is suboptimal when fitting the noisy labels for unsupervised methods (Fig. 1-e). Recently, Zhou et al. [72] addressed the first issue by extracting saliency cues (Fig. 1-f) based on a unsupervisedly pre-trained network (Fig. 1-g) instead of using traditional methods. During training, they focus on learning reliable saliency knowledge from easy samples, but ignore latent knowledge in hard samples. The main reason is that hard samples may be wrongly-labeled and corrupt the fragile saliency knowledge learned in the early training phase. Therefore, to leverage hard samples, we argue that all samples should be employed in a meaningful order (i.e., from high reliable to low reliable), which is crucial for mining accurate knowledge from noisy labels. Trained by such a strategy, the network can mine valuable knowledge from hard examples without corrupting the knowledge learned from easy samples.

Deep networks can learn to localize salient regions from

noisy labels [72], but still struggle to find the precise boundaries of target objects. Generally, the appearance around saliency boundary has a similar texture as in saliency map. Therefore, matching the textures between different maps can serve as a guidance for producing reasonable saliency boundaries. We will demonstrate that above strategies are applicable to multimodal data besides RGB image, including depth map, thermal image, and optical flow.

Based on the above analysis, we propose a novel framework to tackle the Unsupervised Salient Object Detection (USOD) tasks. Specifically, two strategies are proposed to mine saliency knowledge from noisy saliency labels. First, we propose a Confidence-aware Saliency Distilling (CSD) scheme that scores samples with noisy labels conditioned on samples' confidences. Then, our CSD guides the network to learn saliency knowledge from easy samples to more complex ones progressively by employing an adaptive loss conditioned on the training progress. Second, we propose a Boundary-aware Texture Matching (BTM) strategy to refine the saliency boundaries of noisy labels by matching the textures around the predicted boundaries. During training, the predicted saliency boundaries are shifting toward surrounding edges in the appearance space of the whole image. Finally, guided by above two mechanisms, our method can produce high-quality pseudo labels to train generalized saliency detectors. Extensive experiments on RGB, RGB-D, RGB-T, and video SOD benchmarks prove that our method achieves state-of-the-art performance compared to existing USOD methods.

The main contributions of our novel USOD method are:

1. We propose a Confidence-aware Saliency Distilling (CSD) to mine rich and accurate saliency knowledge from noisy labels, which breaks through the limitation that existing methods cannot utilize hard samples.

2. We propose a Boundary-aware Texture Matching (BTM) to refine the boundary of the predicted saliency maps by matching textures in different spaces.

3. Extensive experiments on RGB, RGB-D, RGB-T and video SOD benchmarks prove that our method achieves state-of-the-art USOD performance.

2. Related Works

2.1. Supervised Salient Object Detection

Researchers have developed a large family of fully-supervised Salient Object Detection (SOD) algorithms [5,7, 15,19,30,39,54,58,67–69,71,73] in the past decades. Ronneberger et al. [41] proposed a U-shape structure that progressively upsamples and concatenates the smaller features to the larger ones. To ease the annotation burden, Zhang et al. [65] relabeled the DUTS-TR dataset [52] with scribbles

and leveraged an edge detection model for boundary localization. Yu et al. [62] proposed a local coherence loss to find precise boundary based on scribbles annotations.

Multimodal SOD tasks aim at using other modality data to improve the SOD performance, such as depth map, thermal image and optical flow. Recently, abundant methods [4,8,14,17,40,45–48,57,74] were proposed. Most of these methods employ a two-stream encoder-decoder structure to aggregate multi-level information in multimodal data. To reduce the annotating cost, Zhao et al. [70] proposed a video SOD dataset with scribbles annotations to indicate the location of salient objects.

Although the above methods have achieved significant performance, they require numerous human annotations for training, which are expensive to collect.

2.2. Unsupervised Salient Object Detection

Traditional SOD methods [22, 26, 60, 76] extracted saliency cues from images by modeling the correlation of hand-crafted features. Inspired by the centralization prior, Jiang et al. [22] considered the distances between boundary superpixels and non-boundary superpixels as saliency scores. Yan et al. [60] employed a tree-structure graphical model to compute the saliency results based on several over-segmented maps. The above methods fail to accurately localize salient objects because the global information in hand-crafted features are not representative enough.

Existing DL-based USOD methods can be divided into two pipelines based on the method used to extract saliency cues from images. First, most USOD methods [29, 32, 55, 63, 64, 66] focused on refining the coarse saliency cues extracted by several traditional SOD methods. For example, Zhang et al. [63] weighted these saliency cues by combining both intra-image and inter-image fusion streams. Zhang et al. [66] designed a noise modeling module to deal with noises in these saliency cues. Nguyen et al. [32] used these saliency cues as labels to train multiple deep networks. Ji et al. [21] refined the saliency maps extracted by traditional SOD methods to produce more accurate saliency predictions. Second, to prevent the localization errors caused by traditional SOD methods, Zhou et al. [72] proposed a novel framework that converts the activation maps of a pre-trained network to high-quality pseudo labels.

In this paper, we propose two novel mechanisms to tackle USOD tasks and explore the potential of improving the quality of pseudo labels by using multimodal data.

3. Our Approach

In our method, we propose two novel strategies to mine accurate saliency knowledge based on the noisy activation maps generated by a deep network Φ, as shown in Fig. 2.

Activation map generation. Following previous work [72], we generate an activation map Y for input image X
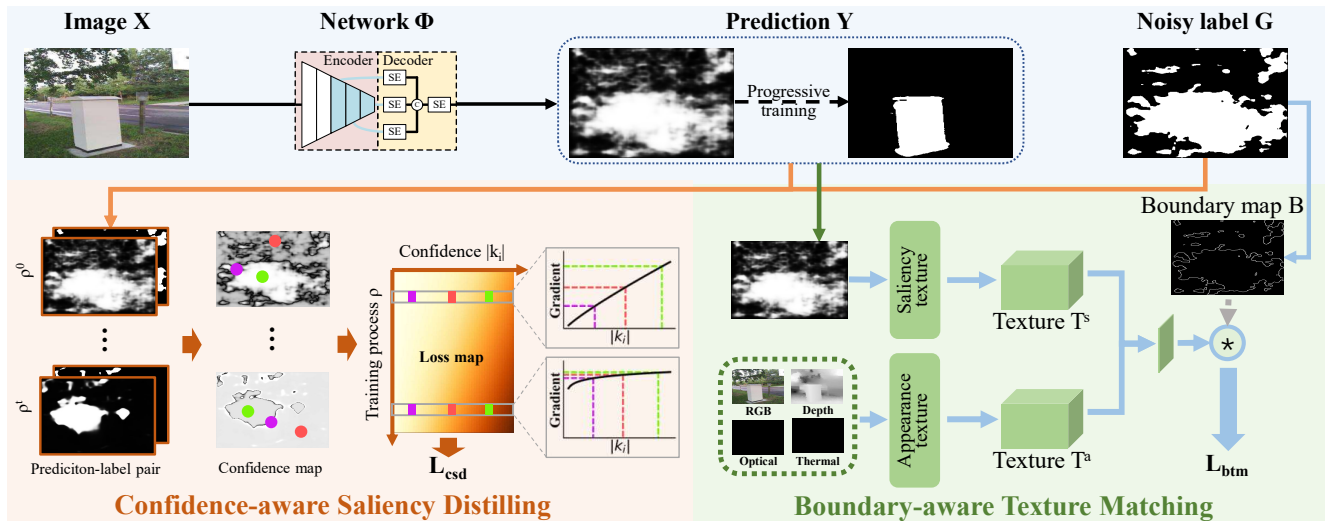
Figure 2. Overview of the proposed unsupervised framework. The network are guided by two mechanisms: (1) Condidence-aware Saliency Distilling mines rich and accurate saliency knowledge from noisy labels, and (2) Boundary-aware Texture Matching makes the boundary of the predicted saliency map more accurate. We pad some constant maps for the missing modalities.

using a deep network $\Phi(X) = Y$:

$$
\begin{aligned}
E_3, E_4, E_5 &= Encoder(X), \\
F_i &= SE(E_i), i \in \{3, 4, 5\}, \\
H &= SE(concat(F_3, F_4, F_5)), \\
Y &= inv(Sigmoid(sum(H - \bar{H}))),
\end{aligned} \tag{1}
$$

where $\bar{H}$ is the spatial mean of H. Specifically, we employ the ResNet-50 [18] pre-trained by MoCo-v2 [6] as our encoder, which is trained without extra manual annotations. In our decoder, four SE blocks [20] integrate multiple encoder features into H. After that, we set $H - \bar{H}$ to ensure that (1) noisy labels are adaptive to input images when using a fixed threshold and (2) coexisting of positive and negative samples. Next, features are summed over the channel dimension to produce one-channel activation map. Finally, the $Sigmoid$ function produces the final saliency scores, and the inv function identifies the regions with more corner pixels as background.

Training strategy. The activation maps produced by network Φ perceive some discriminative regions in input images [72], but still are low-quality because of widespread noises. To improve the quality, we employ three strategies to train the network Φ by:

$$
L_\Phi = \lambda_c(L_{csd} + \hat{L}_{csd}) + \lambda_b(L_{btm} + \hat{L}_{btm}) + \lambda_m L_{ms}, \tag{2}
$$

where all λ are hyperparameters. $\hat{L}$ is the loss for the predictions of a different scale. First, we propose a Confidence-aware Saliency Distilling (CSD) scheme to excavate valuable saliency knowledge from simple examples to more complex ones. Second, we propose a Boundary-aware Texture Matching (BTM) strategy to align the boundaries in

appearance and the predicted saliency maps. In addition, a multi-scale consistency loss L_{ms} ensures our method to produce consistent predictions for multi-scale inputs.

3.1. Confidence-aware Saliency Distilling

The primary challenge for USOD methods is how to localize salient regions. Existing DL-based USOD methods [32, 63, 72] are based on noisy saliency cues from traditional SOD methods [22, 26, 76, 77] or pre-trained networks [6, 18]. Following [72], we employ the activation maps produced by a pre-trained network as our saliency cues instead of the saliency predictions of traditional SOD methods. The main reasons are two-fold: (1) traditional methods require extra computation loads; (2) network can focus on mining saliency knowledge instead of fitting the inductive bias of traditional methods. We binarize the saliency cues as the initial labels using a fixed threshold of 0.5, while Eqn. 1 ensures that these labels remain adaptive to the input image.

To generate high-quality pseudo labels, USPS [32] uses these initial labels to train deep networks like other fully-supervised methods. However, the simple training strategy in fully-supervised methods is suboptimal for unsupervised methods. Specifically, we define pixels with saliency scores close to 0.5 as hard examples because of their low confidences. In the noisy labels, hard examples are likely to be wrongly labeled. Therefore, it is difficult to learn robust saliency knowledge from these hard samples using traditional loss functions (*e.g.*, BCE loss). To this end, A2S [72] learns reliable saliency knowledge from easy samples. However, for hard examples, the saliency knowledge hidden in noises is not fully explored. In summary, a customized strategy that organizes the samples in a more meaningful order is crucial for USOD methods.
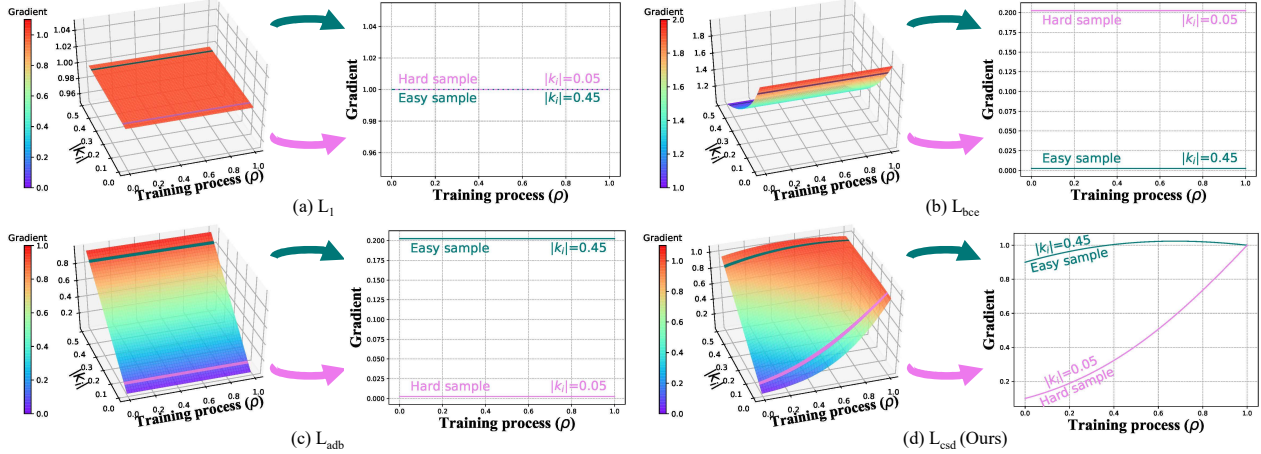
Figure 3. The gradient landscapes of different loss functions.

To address the above problem, we propose a Confidence-aware Saliency Distillation (CSD) scheme that scores samples with noisy labels conditional on their confidences and training progress. Concretely, easy samples contain reliable knowledge, and thus can assist our method to learn reliable saliency knowledge. On the contrary, the saliency knowledge in hard samples is hidden within noises, and may corrupt the fragile saliency patterns in the early stages of network training. Inspired by self-paced learning [2], we dynamically adjust the gradients for samples by introducing a factor ρ, which is linearly increasing from 0 to 1 as training proceeding. Our L_{csd} loss can be formulated as:

$$L_{csd} = -\frac{1}{N} \sum_i^N |\Phi(\mathbf{x}_i) - 0.5|^{(2^{1-\rho})}, \quad (3)$$

where N is the number of pixels, and $\Phi(\mathbf{x}_i)$ is the saliency prediction of pixel $\mathbf{x}_i$. Notice that noisy label G is omitted in Eqn. 3 because it is generated based on $\Phi(\mathbf{x}_i)$. Thus, we can calculate the confidence score $|k_i|$ of pixel $\mathbf{x}_i$ in a simpler way as: $|k_i| = |\Phi(\mathbf{x}_i) - 0.5|$, where the constant value 0.5 is related to $\bar{H}$ in Eqn. 1. The partial derivative of our L_{csd} over $\Phi(\mathbf{x}_i)$ is:

$$\frac{\partial L_{csd}}{\partial \Phi(\mathbf{x}_i)} = -sign(k_i)2^{(1-\rho)}|k_i|^{2^{(1-\rho)}-1}, \quad (4)$$

where $sign(k_i) \in \{-1, 1\}$ for negative and positive values, respectively.

For an intuitive comparison, we draw the gradient landscapes of different losses in Fig. 3, including L_1, L_{bce}, L_{adb} [72] and our L_{csd}. The gradient of other losses are consistent throughout the training process, while our L_{csd} varies with the training process. Specifically, in the beginning, our L_{csd} assigns low gradients to hard samples to learn reliable saliency knowledge from easy samples. As training proceeding, the gradients of hard samples are increasing to mine more valuable saliency knowledge.
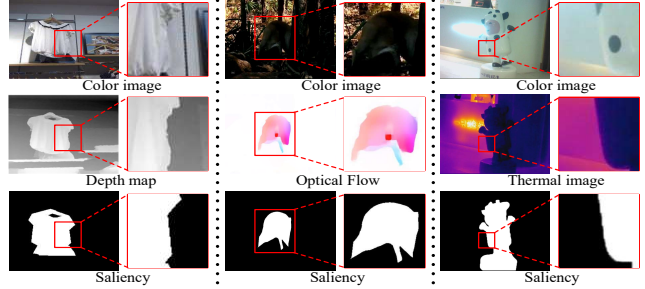

Figure 4. Appearance information in multimodal data.

3.2. Boundary-aware Texture Matching

In general, the appearance of saliency boundary has a similar texture as in saliency prediction. Therefore, matching the textures between different maps can guide our method to produce reasonable saliency scores. This strategy is also applicable to other modalities, such as depth map, thermal image, and optical flow, as shown in Fig. 4. By aggregating the rich appearance information in multimodal data, we can provide a more generalized guidance for our USOD method.

Based on the above analysis, we propose a Boundary-aware Texture Matching (BTM) strategy to align the saliency boundaries and image edges by matching their textures. First of all, similar to the LBP feature [33], we extract texture vectors from saliency predictions and appearance, respectively. For saliency predictions, the texture vector $T_i^s = [t_{i,1}^s, t_{i,2}^s, ..., t_{i,k^2}^s]$ for the i-th pixel is computed by $t_{i,j}^s = |\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_j)|, j \in K_i$, where K_i is $k \times k$ neighborhoods around pixel $\mathbf{x}_i$. For appearance information, it may be one of RGB image, optical flow, depth map or thermal image. To extract more distinctive features from multimodal data, we formulate the texture vector $T_i^a = [t_{i,1}^a, t_{i,2}^a, ..., t_{i,k^2}^a]$ as $t_{i,j}^a = exp(-\alpha \sum_m \|\mathbf{x}_i^{(m)} - \mathbf{x}_j^{(m)}\|^2), j \in K_i$, where α is a hyperparameter and $\mathbf{x}_i^{(m)}$
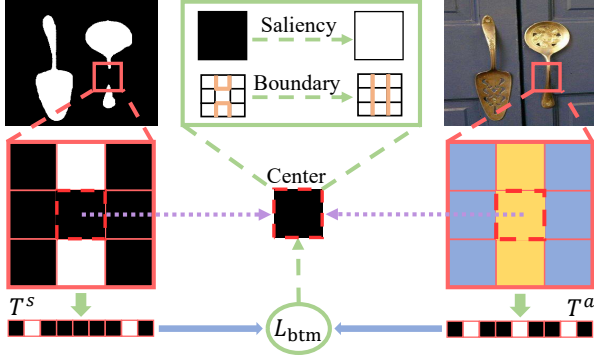
Figure 5. Graphical illustration of the proposed L_{btm}.

means the appearance data of modal m. It is noteworthy that a small difference between two pixels produces a small element in T^s, but a large element in T^a. Therefore, $T_i^s \cdot (T_i^a)^T$ is defined as the matching penalty for considering pixel $\mathbf{x}_i$ as saliency boundary, contrary to similarity. Finally, the formula of our L_{btm} is:

$$L_{btm} = \frac{\sum_i b_i T_i^s \cdot (T_i^a)^T}{\sum_i b_i}, \qquad (5)$$

where b_i is the binary boundary mask of saliency prediction.

An intuitive example is shown in Fig. 5. For the predicted boundary pixels, we find some nearby pixels that differ significantly in saliency scores. In general, the appearances between boundary and these pixels are also different because they locate within salient objects and backgrounds, respectively. If not the case, we expect the saliency score of this boundary pixel to be close to those nearby pixels, so that the boundary will shift in the opposite direction. After an iterative training process, our method can align the predicted saliency boundaries with image edges.

3.3. Multi-scale Consistency

Salient objects are consistent in multi-scale inputs. Thus, we resize input images to a reference scale and encourage our method to produce consistent predictions by:

$$L_{ms} = \sum_i |y_i - resize(\hat{y}_i)|^2, \qquad (6)$$

where y_i and $\hat{y}_i$ are saliency predictions of different scales.

3.4. Training a Generalized Detector

Following previous USOD methods [29,32,72], we generate pseudo labels $\ddot{Y}$ based on saliency predictions Y through $\ddot{Y} = CRF(Y)$ and use them to train an extra saliency detector with the IOU loss. For RGB SOD, we use the same detector as the previous method [72]. For multimodal SOD tasks, we employ the MIDD [47] network as our detector. To ensure a fair comparison with existing SOD methods, we train the extra detector only using task-specific data and corresponding pseudo labels.

4. Experiments

Implementation details. All experiments are implemented on a single GTX 1080 Ti GPU. The batch size is 8 and input images are resized to 320^2. The reference scale is randomly selected from $(192^2, 256^2, 384^2, 448^2)$. Only horizontal flipping is employed as our data augmentation strategy. We train our method for 20 epochs using the SGD optimizer with an initial learning rate of 0.1, which is decayed linearly. α, λ_c, λ_b, and λ_m are set to 200, 1, 0.05 and 1, respectively. We train the extra detectors for 10 epochs using the SGD optimizer with a learning rate of 0.005.

Datasets. In our experiments, we follow the prevalent settings of different SOD tasks. Specifically, for RGB SOD, we use the training subsets of MSRA-B [31] or DUTS [52] to train our method, respectively. ECSSD [42], PASCAL-S [27], HKU-IS [25], DUTS-TE [52], DUT-O [61], and the testing subset of MSRA-B are employed for evaluation. For RGB-D SOD, we choose 2185 samples from NLPR [36] and NJUD [23] as the training set. RGBD135 [9], SIP [13] and the testing subsets of NJUD and NLPR are employed for evaluation. For RGB-T SOD, 2500 images in VT5000 [49] are for training, while VT1000 [50], VT821 [51] and the rest 2500 images in VT5000 are for testing. For video SOD, we choose the training splits of DAVIS [37] and DAVSOD [14] to train our method. Moreover, we randomly select 5 frames of each video in DAVSOD to avoid overfitting. The testing splits of DAVIS, DAVSOD, SegV2 [24] and FBMS [3] are employed for evaluation.

Metrics. We adopt three criteria for evaluation, including ave-F_β, Mean Absolute Error ($\mathcal{M}$) and E-Measure (E_ξ) [12]. Specifically, $F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$, where β^2 is set to 0.3 [1]. The ave-F_β is the F_β scores by setting the threshold as two times mean values. $\mathcal{M}$ is the absolute error between predictions and ground truth. E_ξ measures the global statistics and local pixel matching information.

4.1. Results on RGB SOD

As shown in Tab. 1, we compare the proposed method with **fully-supervised** methods, MINet [34], LDF [56], and KRN [59], **weakly-supervised** methods, WSSA [65], MFNet [38], and SCW [62], and **unsupervised** methods, SBF [63], MNL [66], USPS [32], EDNS [64], A2S [72], DCFD [29], and SelfMask [43]. We list our results with different settings: (1) saliency results without extra detector and post-processing, denoted as "Ours$_{s1}$"; (2) our full method as "Ours"; (3) training our method using multimodal data, while only DUTS-TR or MSRA-B datasets with pseudo labels are employed to train the extra detector ("Ours$_{mm}$").

Either training on MSRA-B or DUTS-TR, the proposed method achieves significant improvements compared to existing USOD methods. Moreover, our unsupervised method

Table 1. Experiment results on SOD benchmarks. "Sup." indicates the supervised signals used to train SOD methods. "F", "W" and "U" mean fully-supervised, weakly-supervised and unsupervised, respectively. Best scores are in bold.

| Methods	Year	Sup.	ECSSD			MSRA-B			DUT-O			PASCAL-S			DUTS-TE			HKU-IS		
			$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$
Trained on MSRA-B																				
SBF [63]	2017	U	.812	.878	.087	.867	.929	.058	.611	.771	.106	.711	.795	.131	.627	.785	.105	.805	.895	.074
MNL [66]	2018	U	.874	.906	.069	.881	.932	.053	.683	.821	.076	.792	.846	**.091**	–	–	–	.874	.932	.047
USPS [32]	2019	U	.875	.903	.064	.896	.938	.042	.715	.839	.069	.770	.828	.107	.730	.840	.072	.880	.933	.043
DCFD [29]	2022	U	.880	.900	.064	.903	.938	.041	**.731**	.838	**.064**	.773	.830	.105	.744	.832	.068	.887	.926	.044
A2S [72]	2022	U	.888	.911	.064	.902	.941	.041	.719	.841	.069	.790	.838	.106	.750	.860	.065	.887	.937	.042
Ours	–	U	**.902**	**.923**	**.056**	**.912**	**.948**	**.036**	**.731**	**.851**	.065	**.803**	**.848**	.099	**.767**	**.871**	**.061**	**.891**	**.939**	**.041**
Trained on DUTS-TR																				
MINet [34]	2020	F	.924	**.953**	.033	.903	.948	.038	.756	.873	.055	.842	.899	.064	.828	.917	.037	.908	**.961**	.028
LDF [56]	2020	F	.930	.951	.034	.902	.944	.037	.773	.881	.052	**.853**	**.903**	.062	.855	.929	.034	.914	.960	.028
KRN [59]	2021	F	**.931**	.951	**.032**	**.911**	**.950**	**.036**	**.793**	**.893**	**.050**	.851	.894	.068	**.865**	**.934**	**.033**	**.920**	**.961**	**.027**
WSSA [65]	2020	W	.870	.917	.059	.869	.929	.049	.703	.845	.068	.785	.855	.096	.742	.869	.062	.860	.932	.047
MFNet [38]	2021	W	.844	.889	.084	.872	.923	.059	.621	.784	.098	.756	.824	.115	.693	.832	.079	.839	.919	.058
SCW [62]	2021	W	**.900**	**.931**	**.049**	**.898**	**.940**	**.040**	**.758**	**.862**	**.060**	**.827**	**.879**	**.080**	**.823**	**.890**	**.049**	**.896**	**.943**	**.038**
EDNS [64]	2020	U	.872	.906	.068	.880	.932	.051	.682	.821	.076	.801	.846	.097	.735	.847	.065	.874	.933	.046
SelfMask [43]	2022	U	.856	.920	.058	.844	.925	.050	.668	.815	.078	.774	.856	.087	.714	.848	.063	.819	.915	.053
DCFD [29]	2022	U	.888	.915	.059	.888	.930	.045	.710	.837	.070	.795	.860	.090	.764	.855	.064	.889	.935	.042
Ours$_{s1}$	–	U	.847	.912	.057	.849	.925	.054	.622	.773	.111	.763	.840	.093	.676	.814	.082	.819	.914	.053
Ours	–	U	**.916**	.938	.044	**.904**	.944	**.039**	**.745**	**.863**	.061	.830	.882	.074	**.810**	**.901**	**.047**	**.902**	**.947**	.037
Ours$_{mm}$	–	U	.913	**.942**	**.043**	.899	**.945**	**.039**	.737	.855	.064	**.831**	**.890**	**.073**	.803	.899	.048	.898	.946	**.037**
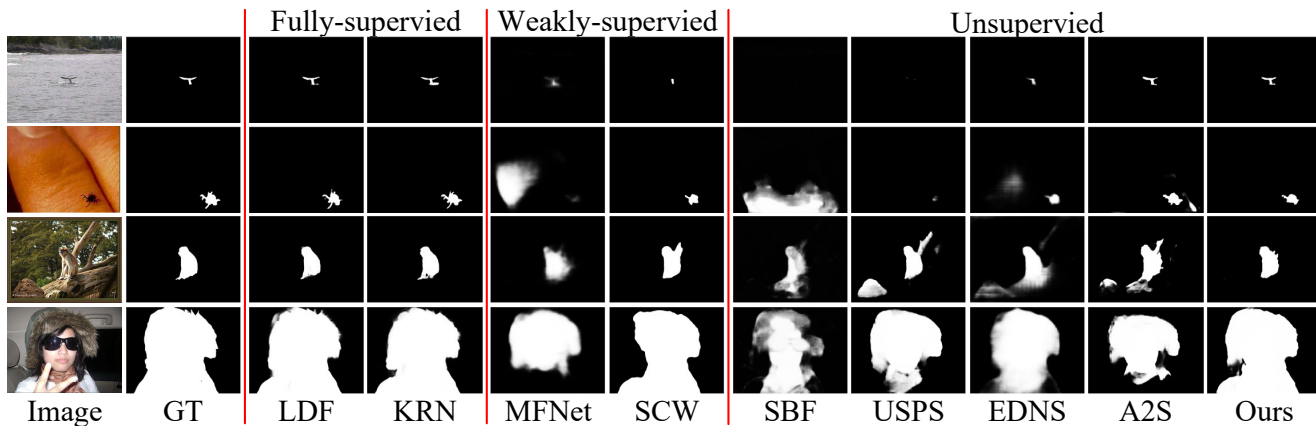
Figure 6. Visual comparison between state-of-the-art SOD methods and ours.

is competitive to recent weakly-supervised SCW [62] and fully-supervised methods KRN [59]. Furthermore, the results of "Ours$_{s1}$" prove that our method extracts precise saliency knowledge from training samples, but somehow weaken when deploying on unseen images. In addition, the similar results of "Ours$_{mm}$" and "Ours" imply that extra multimodal data may not bring significant improvements to RGB SOD task. As proved in [53], the large-scale DUTS-TR dataset is saturated for training saliency detectors, even in the unsupervised case.

A qualitative comparison is illustrated in Fig. 6. Overall, our saliency predictions are much more precise than other unsupervised methods. For example, in the first example, our method and two fully-supervised methods well segment target object with precise boundaries, while other methods fail to capture the tiny salient object.

4.2. Results on Multimodal SOD

We conduct more experiments on multimodal tasks, including RGB-D, RGB-T and video SOD. Noted that we pre-compute the optical flow for each video frame as an extra modality. To verify the improvements brought by multimodal data, we build two sets for training: (1) Task-specific data ("Ours"). For example, for RGB-T SOD task, we only

Table 2. Results on RGB-D SOD benchmarks. "F" and "U" mean fully-supervised and unsupervised, respectively.

Methods	Year	Sup.	RGBD135 $F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	NJUD $F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	NLPR $F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	SIP $F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$
DSA2F [45]	2021	F	.899	.958	.021	.901	.937	.039	.897	.953	.024	–	–	–
SPNet [74]	2021	F	**.927**	**.984**	**.013**	–	–	–	.903	.959	**.021**	**.893**	**.931**	**.043**
CCFE [28]	2022	F	.911	.964	.020	**.914**	**.953**	**.032**	**.907**	**.962**	**.021**	.889	.923	.047
DSU [21]	2022	U	.767	.895	.061	.719	.797	.135	.745	.879	.065	.619	.774	.156
Ours	–	U	.834	.945	.037	.787	.829	.109	.834	.924	.043	.716	.784	.124
Ours$_{mm}$	–	U	**.877**	**.946**	**.029**	**.862**	**.908**	**.060**	**.852**	**.931**	**.034**	**.873**	**.925**	**.051**

Table 3. Results on VSOD benchmarks. "F", "W" and "U" mean fully-supervised, weakly-supervised and unsupervised, respectively.

Methods	Year	Sup.	DAVSOD $F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	DAVIS $F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	SegV2 $F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	FBMS $F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$
PCSA [17]	2020	F	.556	.749	.077	.794	.922	.022	.789	.931	.019	.783	.868	.042
TENet [40]	2020	F	**.595**	**.773**	**.067**	**.821**	**.941**	**.017**	–	–	–	**.851**	**.915**	**.026**
STVS [4]	2021	F	.563	.764	.080	.812	.940	.022	**.835**	**.950**	**.016**	.821	.903	.042
WSVSOD [70]	2021	W	.492	.710	.103	.731	.900	.036	.711	.909	.031	.736	.840	.084
Ours	–	U	.534	.747	**.084**	.751	**.913**	.042	.751	.914	.033	.732	.794	.100
Ours$_{mm}$	–	U	**.547**	**.762**	.085	**.756**	.908	**.037**	**.808**	**.927**	**.021**	**.795**	**.876**	**.060**

Table 4. Results on RGB-T SOD benchmarks. "F" and "U" mean fully-supervised and unsupervised, respectively.

Methods	Year	Sup.	VT5000 $F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	VT1000 $F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	VT821 $F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$
MIED [46]	2020	F	.761	.880	.050	.853	.928	.030	.760	.877	.050
MIDD [47]	2021	F	.801	.899	.043	.882	.942	.027	.805	.898	.045
APNet [75]	2021	F	.821	.918	.035	.885	.951	.021	.818	.912	.034
CCFE [28]	2022	F	**.859**	**.937**	**.030**	**.906**	**.963**	**.018**	**.857**	**.934**	**.027**
Ours	–	U	**.810**	**.904**	**.046**	**.885**	.939	**.031**	**.805**	**.900**	**.043**
Ours$_{mm}$	–	U	.807	.903	.047	.881	**.939**	.032	**.805**	.899	.044

Table 5. Label quality on multimodal SOD datasets.

Methods	RGB $F_\beta\uparrow$	$\mathcal{M}\downarrow$	RGB-D $F_\beta\uparrow$	$\mathcal{M}\downarrow$	VSOD $F_\beta\uparrow$	$\mathcal{M}\downarrow$	RGB-T $F_\beta\uparrow$	$\mathcal{M}\downarrow$
Ours	.917	.038	.746	.110	.557	.099	.932	.029
Ours$_{rgb}$	.917	.040	.804	.068	.613	.089	.920	.030
Ours$_{mm}$	.918	.040	.826	.062	.637	.079	.923	.029

use 2500 RGB-thermal image pairs for training as prevalent RGB-T SOD methods [46,47,75]; (2) Multimodal data from the training sets of four SOD tasks ("Ours$_{mm}$").

Overall, our method reports state-of-the-art performance on all multimodal SOD tasks. Specifically, for the RGB-D SOD task in Tab. 2, our method surpasses the latest unsupervised method DSU [21]. For video and RGB-T SOD tasks in Tab. 3 and 4, to the best of our knowledge, our method is the first unsupervised method and is competitive to supervised methods. Furthermore, our method trained on multimodal data ("Ours$_{mm}$") produce high-quality pseudo labels for all multimodal SOD datasets simultaneously and has achieved performance improvements on most tasks.

4.3. Ablation Study

Label quality comparison. In Tab. 5, we exhibit the scores of pseudo labels on multimodal SOD datasets. Ex-

cept for the above two training sets, we employ an additional set that collects all RGB images from four tasks to train our method, denoted as "Ours$_{rgb}$". Overall, "Ours$_{mm}$" reports more generalized results on all datasets. Moreover, training on task-specific data ("Ours") slightly surpasses the performance of "Ours$_{mm}$" on the training sets of RGB and RGB-T tasks. Since our method is trained without ground truth, fitting a hybrid dataset may cause a slight performance drop on some subsets. In addition, the comparison between "Ours$_{rgb}$" and "Ours$_{mm}$" prove that multimodal data significantly improves the quality of the generated pseudo labels. For an intuitive comparison, we show some examples in Fig. 7. Trained with task-specific data, our method precisely localizes salient regions but fails to segment the whole objects. With more training data, the network perceives fine-grained concepts of target objects, resulting in more complete segmentations.

Impact of initial saliency cues. Our method is based on the activation maps of a pre-trained network [6, 18] instead of traditional methods [22,26,76] as other DL-based USOD methods [29,32,55,63,66]. We compare the performance of these two types of designs in Tab. 6. Overall, the activation map is worse before training, whereas better than traditional methods [22,26,76] after training. The activation maps are
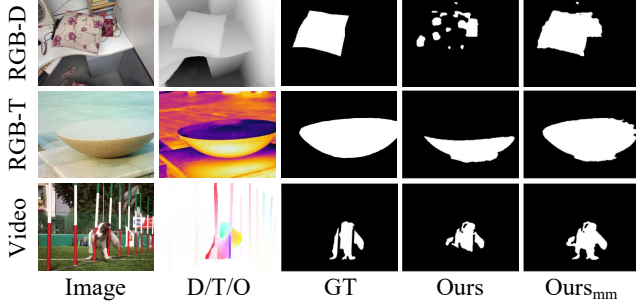
Figure 7. Examples of the generated pseudo labels. The "D/T/O" means depth map, thermal image or optical flow.

Table 6. Comparison between different saliency cues.

Saliency cues	Before training			After training		
	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
MC [22]	.810	.877	.145	.889	.914	.053
DSR [26]	.780	.867	.118	.886	.901	.059
RBD [76]	.803	.883	.109	.891	.912	.053
Ours	.451	.659	.353	.914	.940	.039

Table 7. Ablation studies on loss functions.

Tag	Loss	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
A1	L_{adb} [72] (Baseline)	.882	.915	.071
A2	L_{adb} [72] $+ L_{ms}$	.891	.921	.066
A3	$L_{csd} + L_{ms}$	.908	.937	.050
A4	$L_{csd} + L_{btm} + L_{ms}$ (Ours)	**.917**	**.945**	**.038**
B1	$L_{bce} + L_{btm} + L_{ms}$	.558	.540	.208
B2	$L_1 + L_{btm} + L_{ms}$	.895	.928	.044
B3	L_{adb} [72] $+ L_{btm} + L_{ms}$	.903	.934	.042
B4	L_{spl} [2] $+ L_{btm} + L_{ms}$	.910	.938	.041
B5	$L_{csd} + L_{btm} + L_{ms}$ (Ours)	**.917**	**.945**	**.038**
C1	$L_{csd} + L_{c1} + L_{ms}$	.735	.819	.120
C2	$L_{csd} + L_{lsc}$ [62] $+ L_{ms}$	.906	.923	.052
C3	$L_{csd} + L_{c2} + L_{ms}$	.907	.926	.051
C4	$L_{csd} + L_{btm} + L_{ms}$ (Ours)	**.917**	**.945**	**.038**

Table 8. Label quality on DUTS-TR dataset.

Pre-training	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
Supervised	.915	.942	.039
Unsupervised (Ours)	.917	.945	.038

dynamic during training, such that the network strengthens the learned saliency knowledge incrementally. On the contrary, using traditional methods, the network learns from fixed labels throughout the training process, and thus fits the biased knowledge in those traditional methods. In addition, traditional methods introduce extra computation loads, while the pre-trained network is indispensable for all DL-based USOD methods.

Effectiveness of loss functions. We conduct ablation studies on various combinations of loss functions in Tab. 7 and more details are introduced as follows.

In ablation study A, our method receives continuous improvements by appending new losses (L_{ms} and L_{btm}) or replacing the original L_{adb} to our L_{csd}. This experiment prove that these three losses can assist the network mine more detailed saliency knowledge from different perspectives. More importantly, these losses are supplementary to each other, such that their combination can collaborate to produce high-quality pseudo labels for SOD datasets.

In ablation study B, our L_{csd} loss can provide precise saliency localization information and thus achieves the best performance among all competitors. Specifically, L_{bce} and L_1 fail to mine detailed and reliable saliency knowledge from noisy saliency cues. L_{adb} [72] improves the robustness of the learned saliency knowledge by focusing on easy samples, and thus surpasses the L_1 loss. L_{spl} [2] also employs a dynamic schedule, but its binary weights filter out latent saliency knowledge in hard samples.

In ablation study C, we compare our L_{btm} with L_{lsc} [62] and two variants: (1) using L1 distance for the texture features in appearance space, denoted as L_{c1}; (2) removing the

boundary masks in L_{btm}, denoted as L_{c2}. In general, our L_{btm} outperforms other variants. Specifically, the L1 distance in L_{c1} causes the texture features to be not distinctive enough, resulting in ambiguous saliency boundaries. L_{lsc} focuses more on the adjacent pixels so that the boundaries are more susceptible to a limited region rather than a larger patch. Moreover, using L_{c2}, edges within objects or backgrounds may corrupt the learned saliency knowledge.

Supervised or unsupervised pre-training? Under the unsupervised setting, whether we can use a encoder with supervised pre-training is controversial. For the ImageNet dataset [10], the class labels of images usually indicate the category of the most salient object. It means that the supervised encoder receives extra saliency knowledge from these manual labels. Thus, for completely unsupervised SOD, we employ unsupervised MoCo-v2 to initialize our encoder. As listed in Tab. 8, the performance of our method when using supervised encoder is comparable.

5. Conclusion

In this paper, we propose an Unsupervised Salient Object Detection (USOD) method guided by two novel mechanisms. First, we propose a Confidence-aware Saliency Distilling (CSD) to learn saliency knowledge from easy samples to hard ones with noisy labels progressively. Second, we propose a Boundary-aware Texture Matching (BTM) to make the location of saliency prediction more accurate. As a result, the proposed method produces high-quality pseudo labels to train saliency detectors. Experiments on RGB, RGB-D, RGB-T, and video SOD benchmarks prove that our method outperforms existing USOD methods.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009. 5

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 4, 8

[3] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *European conference on computer vision*, pages 282–295. Springer, 2010. 5

[4] Chenglizhao Chen, Guotao Wang, Chong Peng, Yuming Fang, Dingwen Zhang, and Hong Qin. Exploring rich and efficient spatial temporal interactions for real-time video salient object detection. *IEEE Transactions on Image Processing*, 30:3995–4007, 2021. 2, 7

[5] Peijia Chen, Jianhuang Lai, Guangcong Wang, and Huajun Zhou. Confidence-guided adaptive gate and dual differential enhancement for video salient object detection. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 2

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3, 7

[7] Zixuan Chen, Huajun Zhou, Jianhuang Lai, Lingxiao Yang, and Xiaohua Xie. Contour-aware loss: Boundary-aware learning for salient object segmentation. *IEEE Transactions on Image Processing*, 30:431–443, 2020. 2

[8] Xiaolong Cheng, Xuan Zheng, Jialun Pei, He Tang, Zehua Lyu, and Chuanbo Chen. Depth-induced gap-reducing network for rgb-d salient object detection: An interaction, guidance and refinement approach. *IEEE Transactions on Multimedia*, 2022. 2

[9] 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*, pages 23–27, 2014. 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 8

[11] Wenhui Diao, Xian Sun, Xinwei Zheng, Fangzheng Dou, Hongqi Wang, and Kun Fu. Efficient saliency-based object detection in remote sensing images using deep belief networks. *IEEE Geoscience and Remote Sensing Letters*, 13(2):137–141, 2016. 1

[12] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 698–704, 2018. http://dpfan.net/e-measure/. 5

[13] 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.

[14] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8554–8564, 2019. 2, 5

[15] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1623–1632, 2019. 2

[16] Carola Figueroa Flores, Abel Gonzalez-Garcia, Joost van de Weijer, and Bogdan Raducanu. Saliency for fine-grained object recognition in domains with scarce training data. *Pattern Recognition*, 94:62–73, 2019. 1

[17] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10869–10876, 2020. 2, 7

[18] 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*, pages 770–778, 2016. 3, 7

[19] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017. 2

[20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[21] Wei Ji, Jingjing Li, Qi Bi, Chuan Guo, Jie Liu, and Li Cheng. Promoting saliency from depth: Deep unsupervised rgb-d saliency detection. *arXiv preprint arXiv:2205.07179*, 2022. 1, 2, 7

[22] Bowen Jiang, Lihe Zhang, Huchuan Lu, Chuan Yang, and Ming-Hsuan Yang. Saliency detection via absorbing markov chain. In *Proceedings of the IEEE international conference on computer vision*, pages 1665–1672, 2013. 2, 3, 7, 8

[23] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *2014 IEEE international conference on image processing (ICIP)*, pages 1115–1119. IEEE, 2014. 5

[24] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE international conference on computer vision*, pages 2192–2199, 2013. 5

[25] Guanbin Li and Yizhou Yu. Visual saliency based on multi-scale deep features. In *CVPR*, 2015. 5

[26] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE international conference on computer vision*, pages 2976–2983, 2013. 2, 3, 7, 8

IEEE Transactions on neural networks and learning systems, 32(5):2075–2089, 2020. 5

[27] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014. 5

[28] Guibiao Liao, Wei Gao, Ge Li, Junle Wang, and Sam Kwong. Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 7

[29] Xiangru Lin, Ziyi Wu, Guanqi Chen, Guanbin Li, and Yizhou Yu. A causal debiasing framework for unsupervised salient object detection. In *Thirty-sixth AAAI conference on artificial intelligence*, 2022. 1, 2, 5, 6, 7

[30] 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*, pages 3089–3098, 2018. 2

[31] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2010. 5

[32] Duc Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction with self-supervision. *arXiv preprint arXiv:1909.13055*, 2019. 1, 2, 3, 5, 6, 7

[33] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002. 4

[34] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9413–9422, 2020. 5, 6

[35] Prashant W Patil, Subrahmanyam Murala, Abhinav Dhall, and Sachin Chaudhary. Msednet: multi-scale deep saliency learning for moving object detection. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1670–1675. IEEE, 2018. 1

[36] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *European conference on computer vision*, pages 92–109. Springer, 2014. 5

[37] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 5

[38] Yongri Piao, Jian Wang, Miao Zhang, and Huchuan Lu. Mfnet: Multi-filter directive network for weakly supervised salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4136–4145, 2021. 5, 6

[39] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019. 2

[40] Sucheng Ren, Chu Han, Xin Yang, Guoqiang Han, and Shengfeng He. Tenet: Triple excitation network for video salient object detection. In *European Conference on Computer Vision*, pages 212–228. Springer, 2020. 2, 7

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[42] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2015. 5

[43] Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3971–3980, 2022. 5, 6

[44] Ali Shokoufandeh, Ivan Marsic, and Sven J Dickinson. View-based object recognition using saliency maps. *Image and Vision Computing*, 17(5-6):445–460, 1999. 1

[45] 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/CVF conference on computer vision and pattern recognition*, pages 1407–1417, 2021. 2, 7

[46] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. Multi-interactive encoder-decoder network for rgbt salient object detection. *arXiv e-prints*, pages arXiv–2005, 2020. 2, 7

[47] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. Multi-interactive dual-decoder for rgb-thermal salient object detection. *IEEE Transactions on Image Processing*, 30:5678–5691, 2021. 2, 5, 7

[48] Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu. Rgbt salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 2022. 2

[49] Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu. Rgbt salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 2022. 5

[50] Zhengzheng Tu, Tian Xia, Chenglong Li, Xiaoxiao Wang, Yan Ma, and Jin Tang. Rgb-t image saliency detection via collaborative graph learning. *IEEE Transactions on Multimedia*, 22(1):160–173, 2019. 5

[51] Guizhao Wang, Chenglong Li, Yunpeng Ma, Aihua Zheng, Jin Tang, and Bin Luo. Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In *Chinese Conference on Image and Graphics Technologies*, pages 359–369. Springer, 2018. 5

[52] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 2, 5

[53] 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. *arXiv preprint arXiv:1904.09146*, 2019. 6

[54] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1448–1457, 2019. 2

[55] Yifan Wang, Wenbo Zhang, Lijun Wang, Ting Liu, and Huchuan Lu. Multi-source uncertainty mining for deep unsupervised saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11727–11736, 2022. 1, 2, 7

[56] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 13025–13034, 2020. 5, 6

[57] Hongfa Wen, Chenggang Yan, Xiaofei Zhou, Runmin Cong, Yaoqi Sun, Bolun Zheng, Jiyong Zhang, Yongjun Bao, and Guiguang Ding. Dynamic selective network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:9179–9192, 2021. 2

[58] 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*, pages 3907–3916, 2019. 2

[59] 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*, pages 1–9, 2021. 5, 6

[60] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1155–1162, 2013. 2

[61] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 5

[62] Siyue Yu, Bingfeng Zhang, Jimin Xiao, and Eng Gee Lim. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Palo Alto, CA, USA, 2021. 2, 5, 6, 8

[63] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4048–4056, 2017. 1, 2, 3, 5, 6, 7

[64] Jing Zhang, Jianwen Xie, and Nick Barnes. Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection. In *European conference on computer vision*, pages 349–366. Springer, 2020. 2, 5, 6

[65] 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/CVF conference on computer vision and pattern recognition*, pages 12546–12555, 2020. 2, 5, 6

[66] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9029–9038, 2018. 1, 2, 5, 6, 7

[67] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 202–211, 2017. 2

[68] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, 2017. 2

[69] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 714–722, 2018. 2

[70] Wangbo Zhao, Jing Zhang, Long Li, Nick Barnes, Nian Liu, and Junwei Han. Weakly supervised video salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16826–16835, 2021. 2, 7

[71] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020. 2

[72] Huajun Zhou, Peijia Chen, Lingxiao Yang, Xiaohua Xie, and Jianhuang Lai. Activation to saliency: Forming high-quality labels for unsupervised salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1, 2, 3, 4, 5, 6, 8

[73] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9141–9150, 2020. 2

[74] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4681–4691, 2021. 2, 7

[75] Wujie Zhou, Yun Zhu, Jingsheng Lei, Jian Wan, and Lu Yu. Apnet: Adversarial learning assistance and perceived importance fusion network for all-day rgb-t salient object detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021. 7

[76] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2814–2821, 2014. 2, 3, 7, 8

[77] Wenbin Zou and Nikos Komodakis. Harf: Hierarchy-associated rich features for salient object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 406–414, 2015. 3