# Unified Unsupervised Salient Object Detection via Knowledge Transfer

**Yao Yuan**, **Wutao Liu**, **Pan Gao**[*], **Qun Dai** and **Jie Qin**[*]
Nanjing University of Aeronautics and Astronautics
{ayews233, wutaoliu, pan.gao, daiqun}@nuaa.edu.cn, qinjiebuaa@gmail.com

## Abstract

Recently, unsupervised salient object detection (USOD) has gained increasing attention due to its annotation-free nature. However, current methods mainly focus on specific tasks such as RGB and RGB-D, neglecting the potential for task migration. In this paper, we propose a unified USOD framework for generic USOD tasks. Firstly, we propose a Progressive Curriculum Learning-based Saliency Distilling (PCL-SD) mechanism to extract saliency cues from a pre-trained deep network. This mechanism starts with easy samples and progressively moves towards harder ones, to avoid initial interference caused by hard samples. Afterwards, the obtained saliency cues are utilized to train a saliency detector, and we employ a Self-rectify Pseudo-label Refinement (SPR) mechanism to improve the quality of pseudo-labels. Finally, an adapter-tuning method is devised to transfer the acquired saliency knowledge, leveraging shared knowledge to attain superior transferring performance on the target tasks. Extensive experiments on five representative SOD tasks confirm the effectiveness and feasibility of our proposed method. Code and supplement materials are available at https://github.com/I2-Multimedia-Lab/A2S-v3.

## 1 Introduction

Salient object detection (SOD) aims to identify the most visually significant objects in images. Supervised SOD methods have achieved excellent results, but due to their heavy reliance on pixel-level annotations for salient objects, unsupervised SOD (USOD) has been gaining increasing attention. USOD not only eliminates the need for annotated data but also exhibits strong generalization performance when applied to other tasks [Niu *et al.*, 2021; Wu *et al.*, 2021].

Traditional SOD methods rely heavily on hand-crafted features, such as color and contrast, for saliency extraction. Although these methods prove effective for simple scenes, they encounter difficulties in complex scenes due to the absence of high-level semantic information. Existing deep learning-based USOD methods [Nguyen *et al.*, 2019; Zhang *et al.*, 2018] mostly utilize the predictions generated by traditional SOD methods as saliency cues and incorporate semantic information to generate refined saliency predictions. Recently, based on the observation that CNNs pre-trained on large-scale data usually produce high activations on some primary objects, A2S [Zhou *et al.*, 2023a] have developed a method to distill saliency from the activation maps of deep networks and generate high-quality pseudo-labels. However, we found that during the initial training phase, the presence of hard samples in complex scenes or along object boundaries results in the accumulation of irreparable errors.

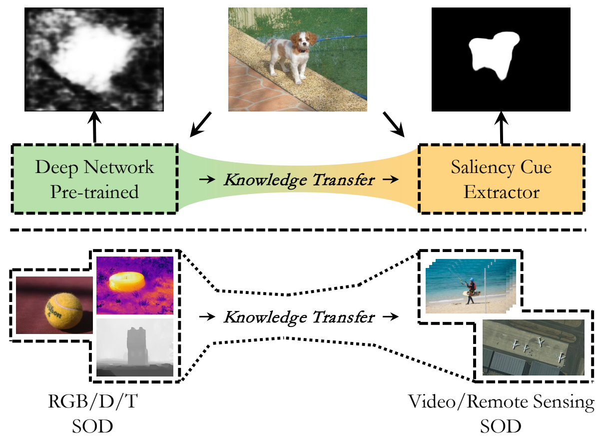


Figure 1: The proposed framework includes two types of knowledge transfer: (1) From pre-trained deep network to saliency cue extractor; (2) From Natural Still Image (NSI) SOD to non-NSI SOD.

Unsupervised SOD is generally considered to exhibit strong generalization and transferability due to its annotation-free nature. However, prevailing USOD methodologies predominantly focus on the Natural Still Image (NSI) domain, exemplified by RGB, RGB-D, and RGB-T. Consequently, USOD on non-NSI domain, encompassing video SOD and Remote Sensing Image (RSI) SOD, remains largely unexplored, presenting a notable research gap in the field. We believe that different SOD tasks share common knowledge, and exploiting this shared knowledge can benefit transfer performance. On the other hand, compared to NSI SOD, the available datasets for video SOD or RSI SOD are relatively small

[*]Corresponding authors

and burdensome to obtain. As a result, training models from scratch on these tasks to obtain satisfactory performance is currently deemed impractical. Therefore, we advocate for the investigation of a more universally applicable unsupervised saliency knowledge transfer method.

To address the aforementioned issues, we design a unified framework for generic unsupervised SOD tasks. Firstly, we propose the Progressive Curriculum Learning-based Saliency Distilling (PCL-SD) mechanism to guide the extraction of saliency cues. At the early stages of training, we only extract preliminary saliency cues from easy samples. As the training progresses, we progressively incorporate hard samples to mine deeper saliency knowledge. The employment of PCL-SD effectively mitigates the initial accumulation of errors, leading to a more stable and robust training process. Next, we utilize the obtained saliency cues to train a saliency detector and design a Self-rectify Pseudo-label Refinement (SPR) mechanism to improve the quality of pseudo-labels. On one hand, the proposed SPR employs the saliency predictions of the model during training to rectify incorrect predictions within the pseudo-labels. On the other hand, it incorporates the prior knowledge of the input image to prevent the model from becoming complacent. The SPR mechanism demonstrates a strong capability in self-supervised learning, resulting in improved pseudo-label quality. Finally, we devise an adapter-tuning method to transfer the acquired saliency knowledge to non-NSI SOD tasks, such as video SOD and RSI SOD. Specifically, we selectively fine-tune the deep features, ensuring effective adaptation of the model to the target task while mitigating the risk of model degradation.

Our main contributions can be summarized as follows:

- We propose the Progressive Curriculum Learning-based Saliency Distilling (PCL-SD) mechanism to extract saliency cues from easy samples to hard ones.
- We design the Self-rectify Pseudo-label Refinement (SPR) mechanism to gradually improve the quality of pseudo-labels during the training process.
- We devise an adapter-tuning method to transfer saliency knowledge from NSI SOD to non-NSI SOD tasks, achieving impressive transfer performance.

Note that we are the first to consider knowledge transfer from NSI domain to non-NSI domain, and develop a unified framework for generic USOD tasks. Experiments on RGB, RGB-D, RGB-T, video SOD and RSI SOD benchmarks confirm the state-of-the-art USOD performance of our method.

## 2 Related Works

### 2.1 Unsupervised Salient Object Detection

Traditional SOD methods rely on hand-crafted features to extract saliency cues. For instance, [Perazzi *et al.*, 2012] estimates saliency by evaluating the contrast in uniqueness and spatial distribution within the image. [Jiang *et al.*, 2011] employ a combination of bottom-up salient stimuli and object-level shape prior to segment salient objects. Although these approaches perform well in simple scenes, they face challenges in handling complex scenes due to the lack of high-level semantic information.

Existing deep learning-based methods for USOD typically involve two stages. In the first stage, pseudo-labels are obtained, while in the second stage, a network is trained using these pseudo-labels. For instance, [Zhang *et al.*, 2017] fuses multiple noisy saliency cues to generate supervisory signals for training the deep salient object detector. In [Nguyen *et al.*, 2019], a set of refinement networks are initially trained to enhance the quality of these saliency cues, and the refined pseudo-labels are subsequently utilized to train a saliency detector. A more recent approach, A2S [Zhou *et al.*, 2023a], proposes a method to distill saliency from the activation maps of deep networks, achieving high-quality pseudo-labels.

### 2.2 Knowledge Transfer in SOD

Knowledge transfer involves applying models or features trained in one task or domain to another related task or domain. A typical example is fine-tuning a deep network that was pre-trained on large-scale data for a specific target task. However, the exploration of knowledge transfer across different SOD tasks remains insufficient. Among the limited studies, [Fu *et al.*, 2022] addresses the RGB-D SOD task as a few-shot learning problem and enhances performance by incorporating knowledge from RGB SOD. [Zhou *et al.*, 2023b] employs data from multiple SOD tasks to train a generalized saliency detector. Nevertheless, when extending to generic SOD tasks, the inherent gap between various SOD tasks can impede effective model training. Consequently, it becomes crucial to devise a knowledge transfer approach that is rooted in shared knowledge.

## 3 Proposed Method

Figure 2 illustrates the proposed two-stage framework. In stage 1, we train a saliency cue extractor (SCE) to transfer saliency knowledge from a pre-trained deep network. The proposed Progressive Curriculum Learning-based Saliency Distilling is employed to mitigate the initial accumulation of errors in training and ensure the stability and robustness of the training process. In stage 2, we utilize the obtained saliency cues as initial pseudo-labels to train a saliency detector (SD). CRF [Krähenbühl and Koltun, 2011] is adopted to enhance the initial pseudo-labels, and we employ the proposed Self-rectify Pseudo-label Refinement mechanism to improve pseudo-labels quality during the training process gradually.

Initially, we train our base model on Natural Still Image (NSI) SOD and subsequently transfer the model to non-NSI SOD tasks. Throughout the training of the base model, we combine all the NSI data for training. However, during the transfer process, we only employ task-specific data for training. For example, when migrating to video SOD, we solely utilize video frames and optical flow as input. The transfer process also follows a two-stage training approach, while we applied the proposed fine-tuning method to optimize the SCE instead of training it from scratch. Besides, ResNet-50 [He *et al.*, 2016] pre-trained by MoCo-v2 [Chen *et al.*, 2020], A2S [Zhou *et al.*, 2023a] and MIDD [Tu *et al.*, 2021] are employed as the pre-trained deep network, SCE, and SD, respectively. A more detailed description and explanation can be found in supplementary materials.

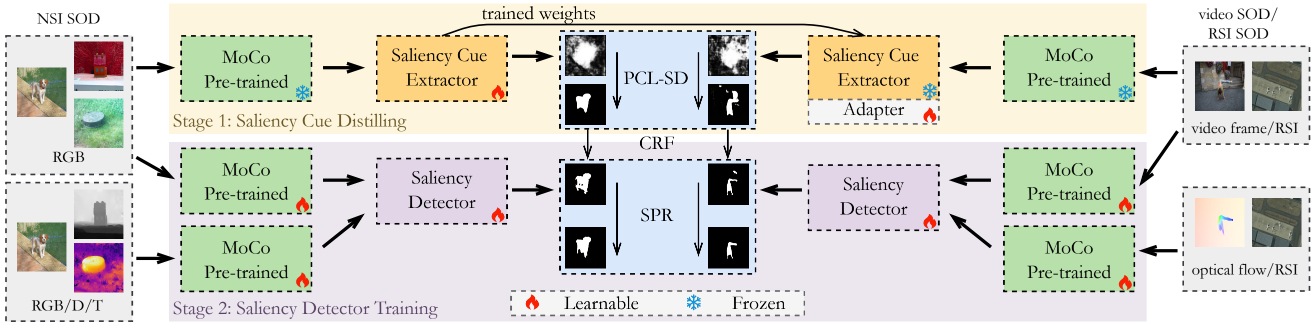


Figure 2: Overview of the proposed method. The left side represents the training process on NSI SOD, while the right side shows the training process of transferring to non-NSI SOD tasks.

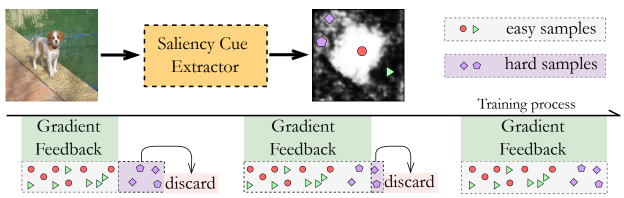


Figure 3: Illustration of the proposed PCL-SD. Hard samples are progressively incorporated as the training progresses.

### 3.1 Progressive Curriculum Learning-based Saliency Distilling

The problem of obtaining saliency cues or extracting salient information from scratch has always been a challenge for unsupervised salient object detection methods. Earlier deep learning-based methods [Zhang *et al.*, 2017] relied on noisy saliency cues generated by traditional SOD methods, while approaches like A2S [Zhou *et al.*, 2023a] employ the activation maps produced by a pre-trained network as saliency cues. This method effectively extracts the saliency information embedded in the pre-trained network. However, at the early stages of training, hard samples in complex scenes may corrupt the fragile saliency patterns in the network, leading to irreparable accumulation errors and the risk of pattern collapse. To address this issue, we introduce the concept of curriculum learning into saliency distilling and propose Progressive Curriculum Learning-based Saliency Distilling (PCL-SD). As can be seen in Figure 3, the proposed PCL-SD rigidly excludes hard samples at the early stages of training and gradually incorporates them as training progresses. As a result, the model progressively extracts saliency knowledge from easy to hard samples, and the entire training process becomes more robust and stable.

The process of saliency distilling can be formulated as:

$$\mathcal{L}_{sal} = 0.5 - \frac{1}{N}\sum_{i}^{N} \|S(i) - 0.5| \tag{1}$$

Here, $N$ represents the number of pixels, and $S(i)$ denotes pixel $i$ in the saliency prediction $S$ output by the saliency cue extractor (SCE). To be intuitively described, $\mathcal{L}_{sal}$ pulls the predicted values of each pixel towards either 0 or 1. However, during the early stages of training, $\mathcal{L}_{sal}$ may pull hard samples with values close to 0.5 in the wrong direction, which we refer to as the problem of error accumulation. The proposed PCL-SD strategy focuses on two essential aspects: (1) how to define hard samples, and (2) how to gradually incorporate them. Firstly, the determination of a pixel in the saliency prediction $S$ as a hard sample is based on its prediction value. Specifically, a pixel $S(i)$ is classified as a hard sample if

$$|S(i) - 0.5| < p. \tag{2}$$

Here, $p$ is the threshold for dividing hard samples, with a larger $p$ indicating more hard samples are divided. Secondly, the value of $p$ is initially set as 0.2 and progressively decreased during the training process until all samples are included. This decrease is governed by the formula:

$$p = Max(0, 0.2 - 0.6 \times E_c/E_t), \tag{3}$$

where $E_c$ and $E_t$ denote current epoch and total epoch, respectively. Finally, we define PCL-SD as:

$$\begin{aligned} M(i) &= \begin{cases} 0 & if\ |S(i) - 0.5| < p, \\ 1 & otherwise, \end{cases} \\ \mathcal{L}_{pcl-sd} &= 0.5 - \frac{1}{N}\sum_{i}^{N} |M(i) \odot S(i) - 0.5| \end{aligned} \tag{4}$$

where $\odot$ denotes the Hadamard product for matrices.

### 3.2 Self-rectify Pseudo-label Refinement

Obtaining high-quality pseudo-labels is crucial for training a saliency detector (SD). As shown in Figure 4, the saliency prediction $S$ output by SD can partially rectify errors within the pseudo-labels. We define saliency prediction as posterior rectification: $R_{\text{post}} = S$. However, while this posterior rectification can rectify errors in initial pseudo-labels, it also introduces the risk of the model becoming overly confident and stagnant. To overcome this, we introduce prior information from the input image to optimize saliency prediction, in order to avoid the model falling into a self-complacent trap.

Previous approaches primarily rely on CRF for prior rectification, which entails significant computational costs. Inspired by [Ru *et al.*, 2022], we employ a real-time pixel refiner to provide efficient prior rectification based on the input image. To start, let $I$ and $P$ represent the input image and position information, while $\sigma_f$ and $\sigma_p$ denote the standard deviation of feature values and position differences, respectively.

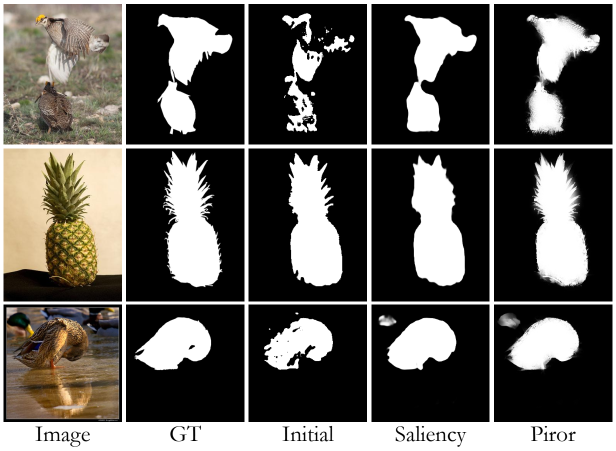


Figure 4: The comparison between initial pseudo-label, saliency prediction, and prior rectification.

The parameters $\omega_1$ and $\omega_2$ control the smoothness. We define the feature distance $d_f^{i,j}$ and position distance $d_p^{i,j}$ between pixels as follows:

$$d_f^{i,j} = -\frac{\|I(i) - I(j)\|}{\omega_1 \sigma_f}, d_p^{ij} = -\frac{\|P(i) - P(j)\|}{\omega_2 \sigma_p} \qquad (5)$$

Then, the refiner $R(\cdot)$ is then defined as:

$$R(I) = \sum_{j \in \mathcal{N}(i)} \left( \frac{exp(d_f^{ij})}{\sum\limits_{k \in \mathcal{N}(i)} exp(d_f^{ik})} + \omega_3 \frac{exp(d_p^{ij})}{\sum\limits_{k \in \mathcal{N}(i)} exp(d_p^{ik})} \right) \qquad (6)$$

Here, $\mathcal{N}(\cdot)$ represents the set of neighboring pixels in an 8-way manner. Finally, the prior rectification can be defined as:

$$R_{\text{pri}} = R(I) \odot S \qquad (7)$$

where $\odot$ denotes the Hadamard product for matrices. At last, the refined pseudo-label is defined as:

$$G_{\text{ref}} = \lambda_1 R_{\text{pri}} + \lambda_2 R_{\text{post}} + \lambda_3 G_{\text{pre}} \qquad (8)$$

Here, $G_{\text{ref}}$ refers to the pseudo-labels after refinement, and $G_{\text{pre}}$ refers to the previous pseudo-labels. The introduction of $G_{\text{ref}}$ aims to improve the stability of the refinement process. $\lambda_1$, $\lambda_2$, $\lambda_3$ are empirically assigned as 0.2, 0.6, and 0.2, respectively. As shown in Figure 4, prior rectification has effectively compensated for the considerable loss of local details. The proposed SPR mechanism combines prior and posterior rectification, gradually improving the quality of pseudo labels during the training process, demonstrating strong self-supervised performance.

### 3.3 Knowledge Transfer to Non-NSI SOD Tasks

We investigate the transferability of the proposed method to video SOD and Remote Sensing Image (RSI) SOD. Figure 5 illustrates the varying degrees of relevance between different SOD tasks. The tasks within NSI SOD benefit from a greater amount of shared knowledge, allowing for the joint training of multiple tasks to achieve a better generalization performance. However, as we broaden our focus to generic SOD

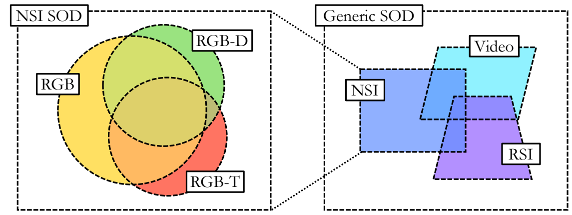


Figure 5: The relevance between different SOD tasks. The overlaps can be seen as shared common knowledge.

tasks, the inherent gap between tasks becomes the primary influencing factor. Joint training becomes more challenging and poses risks of model degradation. More discussions on this topic can be found in supplementary materials.

We posit that identifying an appropriate fine-tuning method can effectively address the issue of model degradation. Inspired by recent studies on Adapter-tuning [Houlsby *et al.*, 2019], we design a simple but effective fine-tuning method for knowledge transfer from NSI SOD to non-NSI SOD tasks. Specifically, for end-to-end tasks in SOD, the prevailing methods and models employ the U-net [Ronneberger *et al.*, 2015] structure and utilize multi-scale feature aggregation to achieve accurate saliency predictions. We contend that shallow features primarily contribute to local details and possess a degree of cross-task generality, while deep features play a pivotal role in salient object localization and exhibit task-specific characteristics. Hence, we suppose that fine-tuning solely the network layers or modules responsible for deep feature handling allows the model to adapt to the target task while circumventing degradation. Technically, we define the deep feature handling process in the model as

$$\hat{F} = T(F) \qquad (9)$$

Here, $F$ represents the deep features extracted by the backbone, $\hat{F}$ denotes the processed features, and $T$ signifies the network layer or module performing the processing. In specific end-to-end SOD models, $T$ can comprise a convolutional layer that modifies the number of feature channels or a network module that enhances the features. Our adapter-tuning approach can be defined as:

$$\hat{F} = T(F) + T_a(F) \qquad (10)$$

In this equation, $T_a$ refers to the adapter, which possesses a structure consistent with $T$. Following the processing, $T_a$ is connected to the original network through a residual connection. During fine-tuning, we exclusively optimize the weights of $T_a$ while keeping the remaining weights of the model frozen. The detailed description can be found in supplementary materials. It is worth mentioning that this fine-tuning method is universal for any kind of SOD method or task.

### 3.4 Supervision Strategy

We initially train the saliency cue extractor (SCE) in the first stage, followed by training the saliency detector (SD) in the second stage. In the training of the first stage, we also incorporate Boundary-aware Texture Matching (BTM) [Zhou *et al.*, 2023b] to introduce extra structural cues, and is defined as $\mathcal{L}_{btm}$. Moreover, a structural consistency loss is employed to achieve transformation-invariant predictions, and is formulated as:

$$\mathcal{L}_{sc} = \sum_{i}^{N} \|S(i) - \hat{S}(i)\|. \tag{11}$$

Here, $\hat{S}$ denotes saliency prediction after transformation. To ensure training stability, only random scaling is adopted. The total loss for training SCE can be defined as:

$$\mathcal{L}_{sce} = \mathcal{L}_{pcl-sd} + \gamma\mathcal{L}_{btm} + \mathcal{L}_{sc}. \tag{12}$$

$\gamma$ is empirically assigned as 0.05.

We train the saliency detector (SD) with IoU loss, which is defined as:

$$\mathcal{L}_{IoU} = 1 - \frac{\sum_{i=1}^{N}(S(i)G(i))}{\sum_{i=1}^{N}(S(i) + G(i) - S(i)G(i))}, \tag{13}$$

$G$ refers to the pseudo-labels, and the total loss for training SD can be defined as:

$$\mathcal{L}_{sd} = \mathcal{L}_{IoU} + \mathcal{L}_{sc}. \tag{14}$$

# 4 Experiments

## 4.1 Implementation Details

### Training Settings

The batch size is set to 8 and input images are resized to 320×320. Horizontal flipping is employed as our data augmentation. We train the saliency cue extractor for 20 epochs using the SGD optimizer with an initial learning rate of 0.1, which is decayed linearly. We train the saliency detector for 10 epochs using the SGD optimizer with a learning rate of 0.005. All experiments were implemented on a single RTX 3090 GPU.

### Datasets

We follow the prevalent settings of SOD and relevant tasks. Here are some details about the datasets we used. **(1) RGB SOD:** We use the training subsets of DUTS [Wang *et al.*, 2017] to train our method. ECSSD [Yan *et al.*, 2013], PASCALS [Li *et al.*, 2014], HKU-IS [Li and Yu, 2015], DUTS-TE [Wang *et al.*, 2017] and DUT-O [Yang *et al.*, 2013] are employed for evaluation. **(2) RGB-D SOD:** We choose 2185 samples from the training subsets of NLPR [Peng *et al.*, 2014] and NJUD [Ju *et al.*, 2014] as the training set. RGBD135 [Cheng *et al.*, 2014], SIP [Fan *et al.*, 2020] and the testing subsets of NJUD and NLPR are employed for evaluation. **(3) RGB-T SOD:** 2500 images in VT5000 [Tu *et al.*, 2022a] are for training, while VT1000 [Tu *et al.*, 2019], VT821 [Wang *et al.*, 2018] and the rest 2500 images in VT5000 are for testing. **(4) Video SOD:** We choose the training splits of DAVIS [Perazzi *et al.*, 2016] and DAVSOD [Fan *et al.*, 2019] to train our method. SegV2 [Li *et al.*, 2013], FBMS [Brox and Malik, 2010] and the testing splits of DAVIS and DAVSOD are employed for evaluation. **(5) Remote Sensing Image SOD:** We choose the training splits of ORSSD [Li *et al.*, 2019] and EORSSD [Zhang *et al.*, 2020b] to train our method. The testing splits of ORSSD and EORSSD are employed for evaluation.

### Metrics

We employ three metrics to evaluate our model and the existing state-of-the-art methods, including Mean Absolute Error $M$, average F-measure ($F_\beta$) [Achanta *et al.*, 2009] and E-measure ($E_\xi$) [Fan *et al.*, 2018]. Specifically, ($M$) measures the average pixel-wise difference between the prediction $P$ and the ground truth $G$, and is calculated as $M = \frac{1}{N}\sum_{i=1}^{N}|P(i) - G(i)|$. $F_\beta$ considers both precision and recall values of the prediction map, and can be computed as $F_\beta = \frac{(1+\beta^2)\times Precision \times Recall}{\beta^2 \times Precision + Recall}$, with $\beta^2$ set to 0.3. $E_\xi$ takes into account the local pixel values along with the image-level mean value, and is defined as $E_\xi = \frac{1}{N}\sum_{i=1}^{N}\phi_\xi(i,j)$, where $\phi_\xi$ represents the enhanced alignment matrix.

## 4.2 Comparisons With State-of-the-Art

We report the performance of our method on five representative SOD tasks, and more qualitative results are provided in supplementary materials.

### Results on RGB SOD

Table 1 presents a quantitative comparison between the proposed method and recent fully-supervised, weakly-supervised, and unsupervised methods. The fully-supervised methods include MINet [Pang *et al.*, 2020] and VST [Liu *et al.*, 2021], the weakly-supervised methods contain WSSA [Zhang *et al.*, 2020a] and MFNet [Piao *et al.*, 2021], and the unsupervised methods comprise SBF [Zhang *et al.*, 2017], TSD [Zhou *et al.*, 2023b] and STC [Song *et al.*, 2023]. Our results are presented under different settings: (1) Training our method using task-specified data, denoted as "Ours$_{t.s.}$", for a fair comparison; (2) Training our method using NSI data, including RGB, RGB-D, and RGB-T datasets, referred to as "Ours".

The results presented in Table 1 clearly indicate that the proposed method outperforms existing USOD methods, leading to significant improvements in performance. Additionally, our unsupervised approach demonstrates competitive performance in comparison to recent weakly-supervised and fully-supervised methods. Notably, our method, referred to as "Ours", exhibits a slight superiority over "Ours$_{t.s.}$". We suppose that this advantage stems from the utilization of a more extensive training dataset, which enhances the model's generalization ability and leads to improved performance when applied to unseen images.

A qualitative comparison is presented in Figure 6. As can be seen, our method has achieved more accurate and complete saliency prediction. Moreover, our approach exhibits remarkable performance in dealing with multiple objects (row 2).

### Results on RGB-D and RGB-T SOD

Table 2 and 3 present a comparison between the proposed method and recent methods on RGB-D and RGB-T benchmarks, respectively. For a fair comparison, we also train our method using task-specified data, denoted as "Ours$_{t.s.}$". VST [Liu *et al.*, 2021], CCFE[Liao *et al.*, 2022], DSU [Ji *et al.*, 2022], TSD, MIDD [Tu *et al.*, 2021] and SRS [Liu *et al.*, 2023] are employed for comparison. Our method has achieved state-of-the-art performance on both RGB-D and

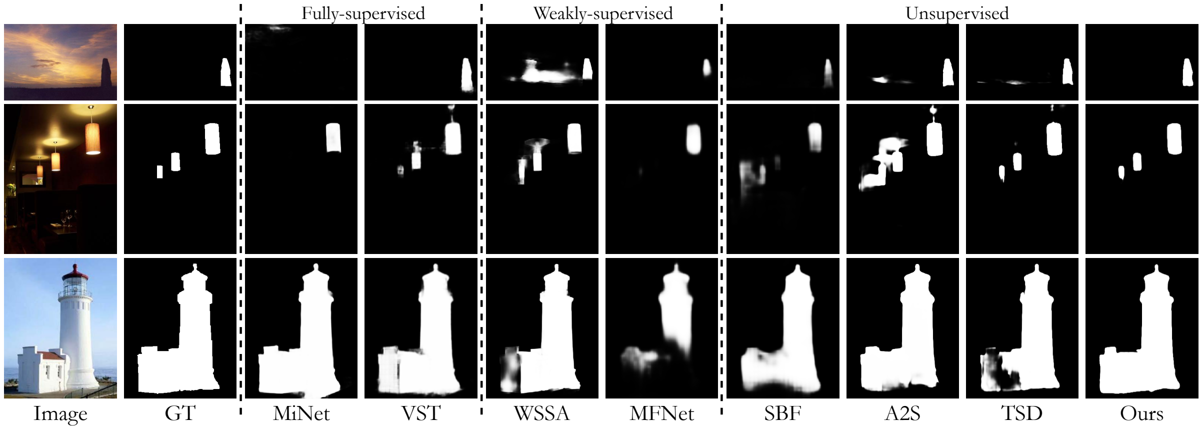


Figure 6: Visual comparison between the proposed method and other state-of-the-art SOD methods on RGB SOD datasets.

| dataset | | | DUT-O | | | DUTS-TE | | | ECSSD | | | HKU-IS | | | PASCAL-S | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Year | Sup. | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ |
| MINet | 2020 | F | **.055** | **.756** | **.873** | **.037** | **.828** | **.917** | **.033** | **.924** | .953 | **.028** | **.908** | **.961** | .064 | **.842** | .899 |
| VST | 2021 | F | .058 | **.756** | .872 | **.037** | .818 | .916 | **.033** | .92 | **.957** | .029 | .9 | .96 | **.061** | .829 | **.902** |
| WSSA | 2020 | W | **.068** | **.703** | **.845** | **.062** | **.742** | **.869** | **.047** | **.860** | **.932** | **.059** | **.870** | **.917** | **.096** | **.785** | **.855** |
| MFNet | 2021 | W | .098 | .621 | .784 | .079 | .693 | .832 | .058 | .839 | .919 | .084 | .844 | .889 | .115 | .756 | .824 |
| TSD | 2023 | U | **.061** | .745 | .863 | .047 | .810 | .901 | .044 | .916 | .938 | .037 | .902 | .947 | .074 | .830 | .882 |
| STC | 2023 | U | .068 | .753 | .852 | .052 | .809 | .891 | .050 | .903 | .935 | .041 | .891 | .942 | .076 | .827 | .881 |
| Ours$_{t.s.}$ | - | U | .063 | .749 | .864 | **.046** | .814 | **.906** | **.038** | .922 | .95 | .034 | .905 | .953 | **.068** | .841 | .898 |
| Ours | - | U | .062 | **.759** | **.868** | .047 | **.816** | **.906** | **.038** | **.923** | **.951** | **.033** | **.908** | **.954** | .069 | **.844** | **.899** |

Table 1: Quantitative comparison on RGB SOD benchmarks. "Sup." indicates the supervised signals used to train SOD methods. "F", "W" and "U" mean fully-supervised, weakly-supervised and unsupervised, respectively. The best results are shown in **bold**.

| dataset | | | RGBD-135 | | | NJUD | | | NLPR | | | SIP | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Year | Sup. | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ |
| VST | 2021 | F | **.017** | **.917** | **.979** | .034 | .899 | .943 | .023 | .886 | .956 | **.04** | **.895** | **.941** |
| CCFE | 2022 | F | .020 | .911 | .964 | **.032** | **.914** | **.953** | **.021** | **.907** | **.962** | .047 | .889 | .923 |
| DSU | 2022 | U | .061 | .767 | .895 | .135 | .719 | .797 | .065 | .745 | .879 | .156 | .619 | .774 |
| TSD | 2023 | U | .029 | 877 | **.946** | .060 | .862 | .908 | .034 | .852 | .931 | .051 | .873 | .925 |
| Ours$_{t.s.}$ | - | U | .027 | .882 | .945 | .053 | .862 | .915 | .034 | .853 | .935 | .042 | .876 | **.935** |
| Ours | - | U | **.025** | **.888** | .94 | **.049** | **.876** | **.923** | **.028** | **.871** | **.945** | **.04** | **.879** | .931 |

Table 2: Quantitative comparison on RGB-D SOD benchmarks.

| dataset | | | VT5000 | | | VT1000 | | | VT821 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Year | Sup. | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ |
| CCFE | 2022 | F | **.030** | **.859** | **.937** | **.018** | **.906** | **.963** | **.027** | **.857** | **.934** |
| SRS | 2023 | W | .042 | .817 | .905 | .027 | .899 | .95 | .036 | .84 | .909 |
| TSD | 2023 | U | .047 | .807 | .903 | .032 | .881 | .939 | .044 | .805 | .899 |
| Ours$_{t.s.}$ | - | U | .041 | .809 | .907 | .024 | .886 | .948 | .057 | .789 | .883 |
| Ours | - | U | **.038** | **.843** | **.924** | **.023** | **.904** | **.956** | **.041** | **.846** | **.918** |

Table 3: Quantitative comparison on RGB-T SOD benchmarks.

| dataset | | | DAVSOD | | | DAVIS | | | SegV2 | | | FBMS | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Year | Sup. | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ | $M \downarrow$ | $F_\beta \uparrow$ | $E_\xi \uparrow$ |
| STVS | 2021 | F | .080 | .563 | .764 | .022 | .812 | .940 | .016 | .835 | .950 | .042 | .821 | .903 |
| WSVSOD | 2021 | W | .103 | .492 | .710 | .036 | .731 | .900 | .031 | .711 | .909 | .084 | .736 | .840 |
| TSD | 2023 | U | .085 | .547 | .762 | .037 | .756 | .908 | .021 | .808 | .927 | .060 | .795 | .876 |
| Ours | - | U | .092 | .572 | .754 | .041 | .764 | .897 | **.018** | **.842** | .92 | .052 | **.822** | .891 |
| Ours$_f$ | - | U | **.084** | **.576** | **.764** | **.030** | **.793** | **.917** | .019 | .83 | **.936** | **.051** | .82 | **.896** |

Table 4: Quantitative comparison on video SOD benchmarks.

| dataset | | | ORSSD | | | EORSSD | | |
|---|---|---|---|---|---|---|---|---|
| Method | Year | Sup. | $M\downarrow$ | $F_\beta\uparrow$ | $E_\xi\uparrow$ | $M\downarrow$ | $F_\beta\uparrow$ | $E_\xi\uparrow$ |
| LVNet | 2019 | F | .021 | .751 | .92 | .015 | .628 | .845 |
| MJRB | 2022 | F | **.016** | **.802** | **.933** | **.010** | **.707** | **.890** |
| Ours | - | U | .057 | .669 | .83 | **.053** | .545 | .755 |
| Ours$_f$ | - | U | **.053** | **.726** | **.874** | .064 | **.625** | **.808** |

Table 5: Quantitative comparison on RSI SOD benchmarks.

| Method | RGB | | RGB-D | | RGB-T | | video | | RSI | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $M\downarrow$ | $F_\beta\uparrow$ | $M\downarrow$ | $F_\beta\uparrow$ | $M\downarrow$ | $F_\beta\uparrow$ | $M\downarrow$ | $F_\beta\uparrow$ | $M\downarrow$ | $F_\beta\uparrow$ |
| Ours$_{t.s.}$ | **.033** | **.928** | .052 | .854 | **.019** | .949 | - | - | - | - |
| Ours | **.033** | .927 | **.047** | **.87** | .020 | **.953** | **.068** | .696 | .074 | .634 |
| Ours$_f$ | - | - | - | - | - | - | .070 | **.698** | **.051** | **.743** |

Table 6: Evaluation on Pseudo-label Quality.

RGB-T SOD. Moreover, our proposed approach, referred to as "Ours", exhibits a substantial performance improvement compared to "Ours$_{t.s.}$". We attribute this improvement to the limited size of the training datasets for these specific tasks. In contrast, "Ours" was trained on a diverse range of datasets encompassing RGB, RGB-D, and RGB-T SOD, effectively utilizing shared common knowledge across different SOD tasks.

**Results on Video SOD and RSI SOD**

Table 4, 5 present a comparison between the proposed method and recent methods on video SOD and RSI SOD benchmarks, respectively. STVS [Chen *et al.*, 2021], WSVSOD [Zhao *et al.*, 2021], LVNet [Li *et al.*, 2019], MJRB [Tu *et al.*, 2022b] and TSD are employed for comparison. We consider video SOD and RSI SOD as two types of target transfer tasks. Thus, in the table, "Ours" represents zero-shot transfer results, while "Ours$_f$" refers to the outcomes obtained by fine-tuning the transferred model on the target task using our proposed knowledge transfer approach. Note that the transfer for video SOD and RSI SOD is conducted separately. Our model exhibits excellent adaptability to the target task following fine-tuning, and exhibits remarkable performance.

## 4.3 Ablation Study

**Evaluation on Pseudo-label Quality**

We assessed the quality of the pseudo-labels generated by models trained on different datasets. As previously mentioned, "Ours$_{t.s.}$" denotes the model trained using task-specific data, whereas "Ours$_f$" refers to the model transferred to the target task. The results are presented in Table 6. In comparison to "Ours$_{t.s.}$", "Ours" exhibits slightly inferior performance on the RGB training set, but displays a notable improvement on the RGB-D and RGB-T training sets, which possess a comparatively limited amount of training data. This indicates that a larger training dataset yields superior model

| Refine Settings | | | RGB | | RGB-D | | RGB-T | |
|---|---|---|---|---|---|---|---|---|
| $G_{\text{res}}$ | $R_{\text{pri}}$ | $R_{\text{post}}$ | $M\downarrow$ | $F_\beta\uparrow$ | $M\downarrow$ | $F_\beta\uparrow$ | $M\downarrow$ | $F_\beta\uparrow$ |
| ✓ | ✗ | ✗ | .04 | .918 | .064 | .825 | .03 | .923 |
| ✓ | ✗ | ✓ | .034 | .925 | .048 | .868 | .022 | .951 |
| ✓ | ✓ | ✓ | **.033** | **.927** | **.047** | **.87** | **.020** | **.953** |

Table 7: Evaluation on Self-rectify Pseudo-label Refinement.

| Loss Settings | RGB | | DUTS-TE | | NLPR | |
|---|---|---|---|---|---|---|
| | $M\downarrow$ | $F_\beta\uparrow$ | $M\downarrow$ | $F_\beta\uparrow$ | $M\downarrow$ | $F_\beta\uparrow$ |
| w/o PCL-SD | **.044** | .895 | .077 | .7 | .05 | .757 |
| w/ PCL-SD | **.044** | **.896** | **.074** | **.713** | **.047** | **.77** |
| $\mathcal{L}_{bce}$ | .034 | .924 | .050 | .784 | .033 | .84 |
| $\mathcal{L}_{iou}$ | .034 | .926 | .049 | .806 | .029 | .866 |
| $\mathcal{L}_{iou}$+$\mathcal{L}_{bce}$ | **.033** | **.928** | .049 | .799 | .032 | .851 |
| $\mathcal{L}_{iou}$+$\mathcal{L}_{ms}$ | **.033** | .927 | **.047** | **.816** | **.028** | **.871** |

Table 8: Evaluation on Supervision Strategy. "RGB" denotes the training set of RGB SOD.

performance and enhanced generalization ability. Furthermore, "Ours$_f$" shows a slight improvement in video SOD, whereas it exhibits a substantial enhancement in RSI SOD. This indicates that video SOD and NSI SOD share more common knowledge, while RSI SOD requires greater fine-tuning and adaptation. More analysis on the adaptation to target tasks is presented in supplementary materials.

**Evaluation on SPR**

We evaluated the influence of various rectifications on the pseudo-labels, and the results are presented in Table 7. The posterior rectification $R_{\text{post}}$ effectively corrects the erroneous predictions present in pseudo-labels, while the prior rectification $R_{\text{pri}}$ adequately compensates for the lack of local details in pseudo-labels. As a result, the proposed SPR gradually enhances the quality of pseudo-labels, thereby improving the model's performance.

**Evaluation on Supervision Strategy**

We evaluated the effectiveness of the proposed supervision strategy, as shown in Table 8. We treat all samples as easy samples to examine the effectiveness of PCL-SD. Upon applying PCL-SD, the model exhibits a slight improvement on the training set. Nonetheless, an impressive enhancement in performance can be observed on the test set. This improvement substantiates the model's heightened generalization capability. Additionally, we explored the training of the saliency detector using different loss functions. The results indicate that the commonly employed binary cross-entropy (bce) in supervised SOD did not lead to effective performance enhancement. We hypothesize that this ineffectiveness may be attributed to the errors and interference stemming from incorrect predictions in pseudo-labels. In contrast, the self-supervised loss $\mathcal{L}_{ms}$ delivered a noteworthy improvement.

# 5 Conclusion

In this paper, we propose a two-stage unified unsupervised SOD framework for generic SOD tasks, with knowledge transfer as the foundation. Specifically, we introduce two innovative mechanisms: Progressive Curriculum Learning-based Saliency Distilling (PCL-SD) and Self-rectify Pseudo-label Refinement (SPR), which aim to extract saliency cues and optimize pseudo-labels. Additionally, we present a universal fine-tuning method to transfer the acquired saliency knowledge to generic SOD tasks. Extensive experiments on five representative SOD tasks validate the effectiveness and feasibility of our proposed method.

## Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No. 62272227 & No. 62276129), and the Natural Science Foundation of Jiangsu Province (No. BK20220890).

## References

[Achanta *et al.*, 2009] Radhakrishna Achanta, Sheila S. Hemami, Francisco J. Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.

[Brox and Malik, 2010] 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.

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

[Chen *et al.*, 2021] 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.

[Cheng *et al.*, 2014] 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.

[Fan *et al.*, 2018] 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 Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 698–704. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[Fan *et al.*, 2019] 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.

[Fan *et al.*, 2020] 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*, 32(5):2075–2089, 2020.

[Fu *et al.*, 2022] Keren Fu, Jing He, and Xiao Yang. Few-shot learning-based rgb-d salient object detection: A case study. *Neurocomputing*, 512:142–152, 2022.

[He *et al.*, 2016] 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.

[Houlsby *et al.*, 2019] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[Ji *et al.*, 2022] 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.

[Jiang *et al.*, 2011] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Tie Liu, Nanning Zheng, and Shipeng Li. Automatic salient object segmentation based on context and shape prior. In *British Machine Vision Conference*, volume 6, page 9, 2011.

[Ju *et al.*, 2014] 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.

[Krähenbühl and Koltun, 2011] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.

[Li and Yu, 2015] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5455–5463, 2015.

[Li *et al.*, 2013] 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.

[Li *et al.*, 2014] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014.

[Li *et al.*, 2019] Chongyi Li, Runmin Cong, Junhui Hou, Sanyi Zhang, Yue Qian, and Sam Kwong. Nested network with two-stream pyramid for salient object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):9156–9166, 2019.

[Liao *et al.*, 2022] 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*, 32(11):7646–7661, 2022.

[Liu *et al.*, 2021] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4722–4732, October 2021.

[Liu *et al.*, 2023] Zhengyi Liu, Xiaoshen Huang, Guanghui Zhang, Xianyong Fang, Linbo Wang, and Bin Tang. Scribble-supervised rgb-t salient object detection. *arXiv preprint arXiv:2303.09733*, 2023.

[Nguyen *et al.*, 2019] 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. *Advances in Neural Information Processing Systems*, 32, 2019.

[Niu *et al.*, 2021] Menghui Niu, Kechen Song, Liming Huang, Qi Wang, Yunhui Yan, and Qinggang Meng. Unsupervised saliency detection of rail surface defects using stereoscopic images. *IEEE Transactions on Industrial Informatics*, 17(3):2271–2281, 2021.

[Pang *et al.*, 2020] 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 (CVPR)*, June 2020.

[Peng *et al.*, 2014] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 92–109. Springer, 2014.

[Perazzi *et al.*, 2012] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE conference on computer vision and pattern recognition*, pages 733–740. IEEE, 2012.

[Perazzi *et al.*, 2016] 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.

[Piao *et al.*, 2021] 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 (ICCV)*, pages 4136–4145, October 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[Ru *et al.*, 2022] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16855, 2022.

[Song *et al.*, 2023] Yicheng Song, Shuyong Gao, Haozhe Xing, Yiting Cheng, Yan Wang, and Wenqiang Zhang. Towards end-to-end unsupervised saliency detection with self-supervised top-down context. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5532–5541, 2023.

[Tu *et al.*, 2019] 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.

[Tu *et al.*, 2021] 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.

[Tu *et al.*, 2022a] 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.

[Tu *et al.*, 2022b] Zhengzheng Tu, Chao Wang, Chenglong Li, Minghao Fan, Haifeng Zhao, and Bin Luo. Orsi salient object detection via multiscale joint region and boundary model. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[Wang *et al.*, 2017] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Wang *et al.*, 2018] 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 *Image and Graphics Technologies and Applications: 13th Conference on Image and Graphics Technologies and Applications, IGTA 2018, Beijing, China, April 8–10, 2018, Revised Selected Papers 13*, pages 359–369. Springer, 2018.

[Wu *et al.*, 2021] Jhih-Ciang Wu, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Learning unsupervised metaformer for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4369–4378, October 2021.

[Yan *et al.*, 2013] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[Yang *et al.*, 2013] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3166–3173, 2013.

[Zhang *et al.*, 2017] 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.

[Zhang *et al.*, 2018] 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.

[Zhang *et al.*, 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/CVF conference on computer vision and pattern recognition*, pages 12546–12555, 2020.

[Zhang *et al.*, 2020b] Qijian Zhang, Runmin Cong, Chongyi Li, Ming-Ming Cheng, Yuming Fang, Xiaochun Cao, Yao Zhao, and Sam Kwong. Dense attention fluid network for salient object detection in optical remote sensing images. *IEEE Transactions on Image Processing*, 30:1305–1317, 2020.

[Zhao *et al.*, 2021] 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.

[Zhou *et al.*, 2023a] 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*, 33(2):743–755, 2023.

[Zhou *et al.*, 2023b] Huajun Zhou, Bo Qiao, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Texture-guided saliency distilling for unsupervised salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7257–7267, 2023.