Unsupervised Domain Adaptive Salient Object Detection Through Uncertainty-Aware Pseudo-Label Learning

Pengxiang Yan[1,2*], Ziyi Wu[1*], Mengmeng Liu[1], Kun Zeng[1], Liang Lin[1], Guanbin Li[1,3†]

[1]Sun Yat-sen University [2]ByteDance Inc. [3]Shenzhen Research Institute of Big Data
yanpx@live.com, wuzy39@mail2.sysu.edu.cn, liumm97@outlook.com,
zengkun2@mail.sysu.edu.cn, linliang@ieee.org, liguanbin@mail.sysu.edu.cn

Abstract

Recent advances in deep learning significantly boost the performance of salient object detection (SOD) at the expense of labeling larger-scale per-pixel annotations. To relieve the burden of labor-intensive labeling, deep unsupervised SOD methods have been proposed to exploit noisy labels generated by handcrafted saliency methods. However, it is still difficult to learn accurate saliency details from rough noisy labels. In this paper, we propose to learn saliency from synthetic but clean labels, which naturally has higher pixel-labeling quality without the effort of manual annotations. Specifically, we first construct a novel synthetic SOD dataset by a simple copy-paste strategy. Considering the large appearance differences between the synthetic and real-world scenarios, directly training with synthetic data will lead to performance degradation on real-world scenarios. To mitigate this problem, we propose a novel unsupervised domain adaptive SOD method to adapt between these two domains by uncertainty-aware self-training. Experimental results show that our proposed method outperforms the existing state-of-the-art deep unsupervised SOD methods on several benchmark datasets, and is even comparable to fully-supervised ones.

Introduction

Salient object detection (SOD) aims to accurately locate and segment out the most visually distinctive object region in a scene. In recent years, the development of deep convolutional neural networks (DCNN) significantly boosts the performance of salient object detection (Wei et al. 2020; Qin et al. 2019; Wang et al. 2018) and has taken place of conventional hand-crafted feature-based algorithms (Zhang et al. 2015; Zhu et al. 2014; Li et al. 2013) to become the dominant methods in salient object detection. However, such promising performance comes at a cost of a large number of pixel-wise annotated images to train the DCNN-based models. Moreover, to ensure the quality and consistency of labeling, it generally requires multiple human annotators to annotate fine pixel-level masks for the same image (Li et al. 2017; Fan et al. 2018). The time-consuming and laborious labeling work limits the amount of training data and thus hampers the further development of DCNN-based SOD methods.

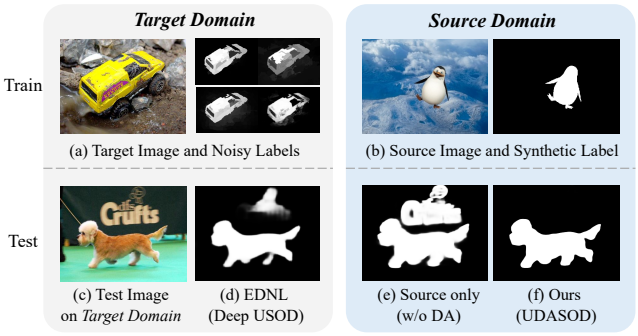


Figure 1: Deep unsupervised salient object detection (USOD) achieved by two training settings. Existing deep USOD algorithms are mainly trained on **(a)** real-world images (target domain) with noisy labels generated by traditional USOD methods. While we propose to exploit **(b)** the synthetic saliency data (source domain) for training. However, due to the discrepancy between two domains, the saliency detector trained only on synthetic data (**(e)** Source only) without domain adaption (DA) usually fails to performs well on **(c)** real images. To solve this problem, we propose **(f)** an unsupervised domain adaptive SOD (UDASOD) method, which can generate more accurate saliency predictions than **(d)** the best-performing deep USOD method EDNL (Zhang, Xie, and Barnes 2020).

To alleviate the burden of pixel-wise labeling but take full advantage of the end-to-end training advantages of DCNN, weakly-supervised (Zhang et al. 2020b; Zeng et al. 2019; Li, Xie, and Lin 2018) and deep unsupervised (Zhang, Xie, and Barnes 2020; Nguyen et al. 2019; Zhang et al. 2018) SOD algorithms have been proposed. Weakly-supervised SOD algorithms mainly focus on learning saliency inference from simple but clean manual annotations, such as image classes (Li, Xie, and Lin 2018), image captions (Zeng et al. 2019), and scribbles (Zhang et al. 2020b). While deep unsupervised SOD methods aim to learn saliency detection without resorting to any manual annotations. Existing deep unsupervised SOD methods mainly focus on learning from the dense noisy labels generated by single (Zhang et al. 2020a) or multiple (Zhang, Xie, and Barnes 2020; Nguyen et al. 2019; Zhang et al. 2018; Zhang, Han, and Zhang 2017) tra-

*The first two authors have equal contribution. †Corresponding author is Guanbin Li (Email: liguanbin@mail.sysu.edu.cn).

ditional unsupervised SOD methods (as shown in Fig. 1 (a)), which can be achieved through noise modeling (Zhang, Xie, and Barnes 2020; Zhang et al. 2018) or pseudo-label self-training (Zhang et al. 2020a; Nguyen et al. 2019; Zhang, Han, and Zhang 2017). However, traditional unsupervised methods that rely on manual features and specific saliency priors are arduous to deal with the complex situation of low foreground/background contrast. The generated pseudo-labels are rich in noise and are almost impossible to be repaired in iterative training based on pseudo-labels, especially for the boundary of the salient objects.

Instead of struggling with the generated noisy labels of real images, in this paper, we propose that learning saliency from the synthetic but clean labels (Fig. 1 (b)) would be yet another feasible solution. There are massive object images with transparent backgrounds as well as pure background images without salient objects that can be easily collected from the design resources or photography websites on the Internet. Since the salient objects of a scene are usually the foreground objects, we construct a new large-scale synthetic salient object detection (SYNSOD) dataset with clean labels by simply copying foreground objects and pasting them on the background images. The SYNSOD dataset can be applied to existing fully-supervised SOD methods to relieve the burden of manual annotations. However, as shown in Fig. 1, due to the presence of large appearance differences between real images (*target domain*) and synthetic images (*source domain*) known as "domain gap", the model directly trained on SYNSOD (Fig. 1 (e)) fails to performs well on the real-world dataset such as DUTS (Wang et al. 2017).

To resolve the above issues, we propose a novel unsupervised domain adaptive salient object detection (UDASOD) algorithm to adapt the DCNN-based saliency detector trained on the synthetic dataset to the real-world SOD datasets. The proposed UDASOD algorithm is an iterative method that exploits an uncertainty-aware pseudo-learning (UPL) strategy to achieve adaption between two domains. Specifically, in each round of iteration, UDASOD leverages the source images with synthetic labels and the target images with weighted pseudo-labels to jointly train the saliency detector. After the training of each round, UPL dynamically updates the training set and pseudo-labels of the target domain through three major steps, including pixel-wise uncertainty estimation, image-level sample selection and pixel-wise pseudo-label reweighting. The main contributions of this paper can be summarized as follows:

- To our knowledge, we are the first attempt to achieve SOD by exploiting unsupervised domain adaption from synthetic data, which varies from existing deep unsupervised SOD algorithms targeted at noisy labels.
- We construct a synthetic SOD dataset and further propose UDASOD that exploits uncertainty-aware pseudo-label learning to adapt the saliency detector trained on the synthetic dataset to real-world scenarios.
- Experimental results show that our proposed domain adaptive SOD method outperforms all existing state-of-the-art deep unsupervised SOD methods and is comparable to the fully-supervised ones.

Related Work

Salient Object Detection

Conventional SOD is mainly achieved by different saliency priors or handcrafted features (Zhang et al. 2015; Zhu et al. 2014; Li et al. 2013). Recent advances in DCNNs significantly boost the performance of SOD at a cost of numerous pixel-wise annotations (Chen et al. 2020; Wei, Wang, and Huang 2020; Pang et al. 2020; Liu et al. 2019; Wang et al. 2018). To mitigate the labeling costs, weakly supervised SOD is proposed to learn saliency under weak supervision such as image classes (Li, Xie, and Lin 2018), captions (Zeng et al. 2019), and scribbles (Zhang et al. 2020b). Deep unsupervised SOD is further proposed to learn saliency without resorting to any manual annotations. Existing deep unsupervised methods mainly rely on learning from noise labels generated by conventional SOD methods, which can be achieved through noise modeling (Zhang, Xie, and Barnes 2020; Zhang et al. 2018) or pseudo-label self-training (Zhang et al. 2020a; Nguyen et al. 2019). In this paper, we propose to solve SOD from a novel perspective, *i.e.*, learning from synthetic but clean labels.

Unsupervised Domain Adaption

Unsupervised domain adaptation (UDA) aims to transfer the knowledge learned from the label-rich source domain to an unlabeled target domain. It is widely studied on various vision tasks such as image classification (Sener et al. 2016), object detection (Chen et al. 2018), semantic segmentation (Chen et al. 2017), *etc*. Among these tasks, semantic segmentation shares most characteristics with SOD. The primary approach of UDA for semantic segmentation is to minimize the discrepancy between two domain distributions through adversarial learning (Hoffman et al. 2018; Luo et al. 2019; Tsai et al. 2018). There are some self-training-based UDA methods (Zou et al. 2018, 2019) that assign pseudo-labels to confident target samples and directly use pseudo-labels as target domain supervision to reduce domain mismatch. To the best of our knowledge, we are the first attempt to relieve the burden of large-scale manual annotations by leveraging UDA on salient object detection.

Pseudo-Label Learning

Pseudo-label learning, which is initially explored in semi-supervised learning scenario (Lee et al. 2013), has recently attracted wide attention due to its simplicity and effectiveness. The goal of pseudo-label learning is to fully exploit the unlabeled data by generating and updating pseudo-labels for unlabeled samples with a model trained on labeled data. Thus, it can be applied to benefit various tasks such as semi-supervised learning (Lee et al. 2013; Yan et al. 2019), domain adaption (Zheng and Yang 2021; Li et al. 2020), and noisy label learning (Zhang et al. 2020a; Tanaka et al. 2018). There are also some SOD methods (Li, Xie, and Lin 2018; Nguyen et al. 2019) that exploit the pseudo-label learning technique. Different from them, our proposed method exploits an uncertainty-aware pseudo-learning strategy that treats each pseudo-label differently and is also free of time-consuming post-processing like fully-connected CRF.

(a) Foreground Object (b) Background (c) Synthetic Image (d) Synthetic Label

Figure 2: Examples of the dataset construction of SYNSOD. Each foreground object is matched with a unique background to generate a synthetic sample through a simple copy-paste strategy. The pixel-level synthetic label can be obtained from the alpha channel of the foreground image.

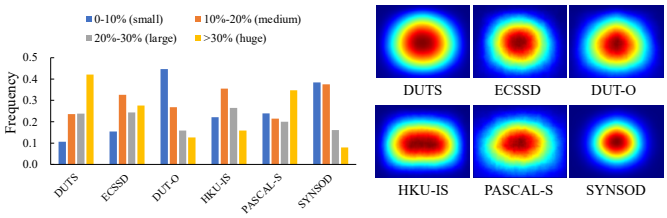
Proposed Dataset

In this section, we detail the proposed SYNSOD dataset from the following aspects.

Image Collection. As salient objects are usually the foreground objects of a scene, we can intuitively obtain a synthetic image with salient objects by pasting the foreground objects on a background image. Thus, to construct a novel synthetic SOD dataset, we first collect a large number of object images with transparent backgrounds (RGBA color) from several websites with non-copyrighted design resources, each of which contains single or multiple objects of diverse appearances and categories. Next, we collect background photos from multiple non-copyrighted photography websites, which contains various non-salient scenes such as forest, grass, sky, ocean, *etc*. The collection process is executed through a designed spider program and images with low resolution will be automatically removed.

Data Generation. Given the premise of the collected foreground and background images, we can easily generate the synthetic SOD dataset by a simple copy-paste strategy. As shown in Fig. 2, we match each foreground object image with a unique non-salient background image. Then, we randomly scale the object image with a ratio ranging from 0.5 to 1.1. Next, we set an object center in the background image to cover its surrounding pixels with the non-transparent object pixels, resulting in the synthetic image for SOD. The pixel-level synthetic label can be easily obtained by binarizing the corresponding alpha channel of the foreground object pixels in a synthetic image. In this way, we construct a large-scale synthetic SOD dataset (SYNSOD), containing 11,197 synthetic images and corresponding pixel-level labels.

Dataset Statistics. As shown in Fig. 3, we present the following dataset statistics on our proposed SYNSOD dataset and five public benchmark SOD datasets (Wang et al. 2017; Yan et al. 2013; Yang et al. 2013; Li and Yu 2015; Li et al. 2014). **1) Object size.** As shown in Fig. 3 (a), the ratio of salient object size in SYNSOD ranges from 0.39% to 86.96% (avg.: 14.72%), yielding a border range. **2) Center bias.** To reveal the degree of center bias, we compute the average saliency maps over all images of each dataset. As shown in Fig. 3 (b), SYNSOD is center-biased and the degree of center-bias is slightly stronger than others, which shows strong domain gaps with other real-world datasets.



(a) Distribution of object size (b) Center bias

Figure 3: Statistics of our proposed SYNSOD dataset including distribution of salient object size and center bias.

Methodology

Problem Formulation

To achieve SOD without resorting to manual annotations or noisy labels, we propose to learn saliency from the synthetic but clean labels through a novel unsupervised domain adaptive salient object detection (UDASOD) framework. As shown in Fig. 4, UDASOD is formulated as an iterative training paradigm, which can leverage existing deep learning-based saliency detectors to learn saliency prediction from synthetic source data and unsupervisedly adapt it to the real target scenarios. To fully exploit unlabeled target images, UDASOD is jointly trained with the pseudo-labels of target images and the synthetic labels of source images.

To formulate UDASOD, we start with the synthetic training set denoted as the source domain $\mathcal{D}_{src} = \{(I_s, y_s)\}_{s=1}^{S^i}$, where I_s is a synthetic RGB color image of size $H \times W$, $y_s \in \{0,1\}^{H \times W}$ is the corresponding binary saliency map, and S^i is the number of source images in round i. The proposed UDASOD framework will unsupervisedly adapt the saliency detector from the synthetic dataset to the real SOD dataset denoted as target domain $\mathcal{D}_{trg} = \{(I_t, \hat{y}_t)\}_{t=1}^{T^i}$, where I_t is a real RGB color image, $\hat{y}_t \in [0,1]^{H \times W}$ is the corresponding pseudo-label, and T^i is the number of pseudo-labels in round i. Thus, the training process of round i can be formulated as optimization of the network parameters θ of the saliency detector as follows:

$$\theta^i = \arg\min_{\theta} \mathcal{L}(\theta, i), \tag{1}$$

where the loss function $\mathcal{L}(\theta, i)$ under the joint supervision of source $\mathcal{D}_{src}$ and target $\mathcal{D}_{trg}$ domains is defined as:

$$\mathcal{L}(\theta|i) = \mathcal{L}_{src}(\theta|I_s^i, Y_s^i) + \mathcal{L}_{trg}(\theta|I_t^i, \hat{Y}_t^i). \tag{2}$$

Here, Y_s and $\hat{Y}_t$ denote the set of synthetic source labels and the set of pseudo target labels, respectively. $\mathcal{L}_{src}$ and $\mathcal{L}_{trg}$ refer to the specific loss calculation of source and target samples, which will be detailed in the following.

However, since the pseudo-labels of target domain are generated by the saliency detector initially trained on the source domain, the pseudo-labels inevitably contain incorrect pixel-level prediction due to the significant distribution gap between the two domains. To avoid error accumulation in the iterative training process, we propose that the samples of the target domain need to be carefully selected to

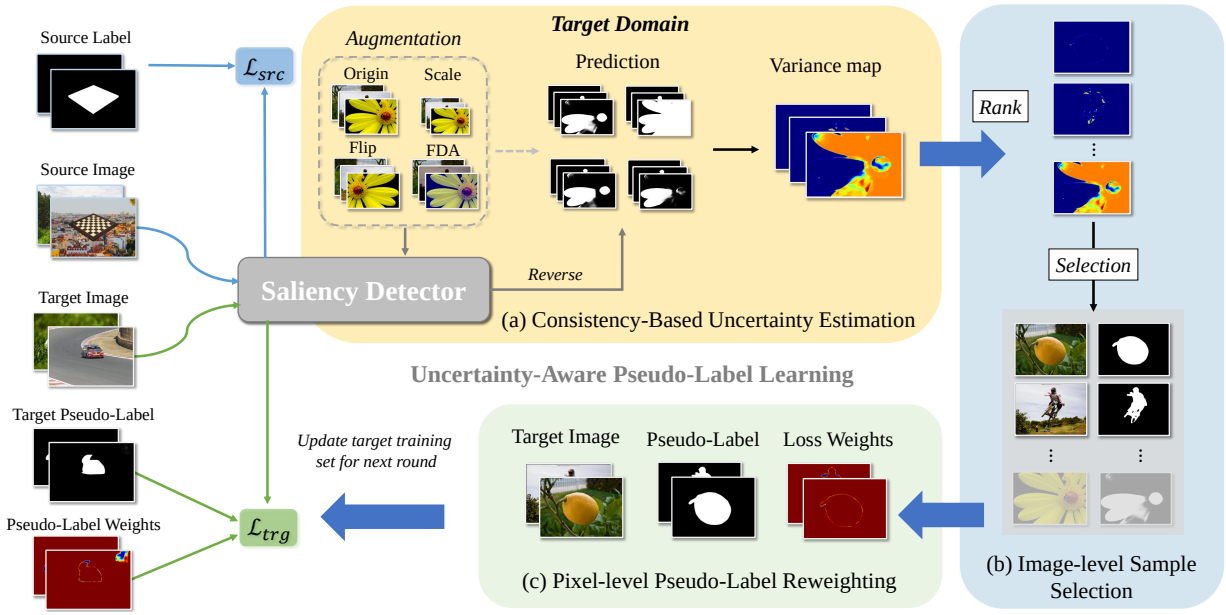


Figure 4: The overall framework of our proposed unsupervised domain adaptive salient object detection method. It iteratively learns saliency from the synthetic labels (source domain) and pseudo-labels of real images (target domain). The pseudo-labels will be dynamically updated after each training round through an uncertainty-aware pseudo-label learning strategy that contains three major steps, *i.e.*, (a) consistency-based uncertainty estimation, (b) image-level sample selection, and (c) pixel-wise pseudo-label reweighting. We use three kinds of data augmentations for consistency-based uncertainty estimation, including 1) horizontal flipping (Flip), 2) rescale input image to 224×224 (Scale), and 3) randomly swap image style with other target images via FDA (Yang and Soatto 2020).

participate in the training, and each pixel of the selected sample should be adaptively assigned different weights. Therefore, the loss function for each predicted saliency map $p \in [0,1]^{H\times W}$ is formulated with a weight matrix $\omega \in (0,1]^{H\times W}$ as follows:

$$\mathcal{L}(y,p,\omega) = \sum_{h=1}^{H}\sum_{w=1}^{W} \omega^{(h,w)} \ell(y^{(h,w)}, p^{(h,w)}), \tag{3}$$

where $\ell(.)$ denotes the binary cross-entropy loss for each pixel and $y \in [0,1]^{H\times W}$ denotes the dense label of p. Then, the loss function for the source and target samples can be formulated as:

$$\mathcal{L}_{src}(\theta|X_s^i, Y_s^i) = \sum_{s=1}^{S_i} \mathcal{L}(y_s, p_\theta(I_s), \omega_s), \tag{4}$$

$$\mathcal{L}_{trg}(\theta|X_t^i, Y_t^i) = \sum_{t=1}^{T_i} \mathcal{L}(\hat{y}_t, p_\theta(I_t), \omega_t), \tag{5}$$

where $p_\theta(I)$ denotes the prediction of saliency detector with parameters θ for input image I. In practice, we only assign different pixel-wise weights to pseudo-labels while setting $\omega_s = \mathbb{1} \in \mathbb{R}^{H\times W}$ in source domain. At the end of each round, the target training set with pseudo-labels will be dynamically updated and assigned with pixel-level weights based on our proposed uncertainty-aware pseudo-label learning strategy.

Uncertainty-Aware Pseudo-Label Learning

Instead of equally using all the pseudo-labels, we propose to select target pseudo-labels and assign pixels with different weights through an uncertainty-aware pseudo-label learning strategy (UPL) that contains the following three major steps.

1) Consistency-Based Uncertainty Estimation. To update the target pseudo-labels, we first perform consistency-based uncertainty estimation. Specifically, as shown in Fig. 4, given a saliency detector with fixed parameters $\tilde{\theta}$, we feed each real target image I_t into the saliency detector to obtain its pseudo-label $\hat{y}_t = p_{\tilde{\theta}}(I_t)$. To model the uncertainty of a target pseudo-label, we consider the following two aspects. First, the saliency detector will be robust to different small noises on target samples of high-confidence / low-uncertainty. Second, as it is recognized data augmentation can be regarded as a noise injection method (Xie et al. 2020), we model the uncertainty by evaluating the consistency of the saliency predictions of the target image I_t under multiple data augmentations. The salient prediction under the data augmentation $\{\alpha_j(.)\}_{j=1}^N$ can be formulated as:

$$\tilde{y}_t^j = \alpha_j^{-1}(p_{\tilde{\theta}}(\alpha_j(I_t))). \tag{6}$$

Here, we only adopt the data augmentation $\alpha(.)$ that can be reversed and $\alpha^{-1}(.)$ will be applied for each saliency prediction $\tilde{y}_t^j$ to transform it back to the same condition (*e.g.*, direction, scale) as the pseudo-label $\hat{y}_t$. Inspired by (Zheng and Yang 2021), we leverage variance to evaluate the consistency of the pseudo-label and other saliency predictions of

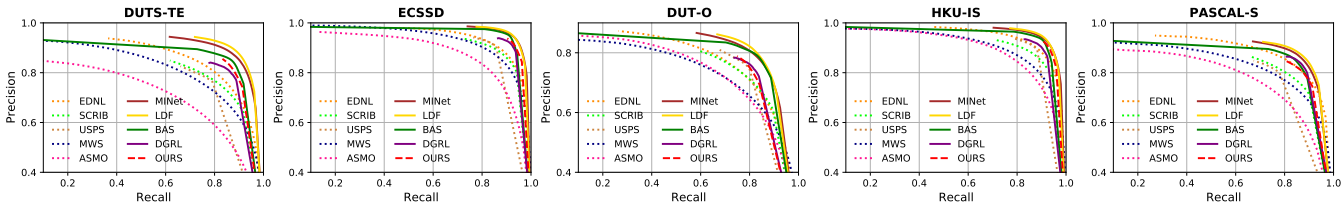


Figure 5: Quantitative comparison with state-of-the-art SOD methods in terms of Precision-Recall curves.

different variants of data augmentations. For simplification, we let $\tilde{y}_t^1 = \hat{y}_t$. The variance map v_t of the sample I_t can be formulated as:

$$Var(I_t, \tilde{\theta}) = \mathbb{E}[(\tilde{y}_t^j - \frac{1}{N}(\sum_{j=1}^{N} \tilde{y}_t^j))^2], \tag{7}$$

where $\mathbb{E}(.)$ denotes the mathematical expectation. The dense variance map $v_t \in \mathbb{R}^{H \times W}$ can be used to represent the pixel-level uncertainty of the target pseudo-label $\hat{y}_t$.

2) Image-level Sample Selection (ISS). Since the saliency detector is generally weak in the early training stage and is gradually improved during iterative training, we propose that 1) only the pseudo-labels of low uncertainty should be selected and 2) the number of pseudo-label should slowly increase with the increase of training rounds. As shown in Fig. 4, the variance maps can reflect the pixel-level uncertainty of the target pseudo-labels, where red and blue indicate high and low uncertainty, respectively. Thus, to rank the target sample by their uncertainty, we introduce the image-level uncertainty score U based on the mean value of variance (Eq. (7)). The uncertainty score of the target image I_t can be formulated as:

$$U(I_t, \tilde{\theta}) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} Var(I_t, \tilde{\theta})^{(h,w)}. \tag{8}$$

We rank all the target domain samples according to the uncertain score and select a certain proportion of target samples with low uncertainty for each round. The proportion will increase with the improvement of the saliency detector. Note that here we also empirically discard those pseudo-labels composed of nearly all salient or non-salient pixels.

3) Pixel-wise Pseudo-Label Reweighting (PPR). Although the selected target pseudo-labels generally reflect a low-uncertainty level, there still exists high uncertainty regions such as object boundaries as shown in their variance maps. Therefore, we suggest that each pixel of the pseudo-labels should be treated differently during the training process and further propose a pixel-wise pseudo-label reweighting strategy Ω based on the variance maps Var. The pixel-wise weight matrix $w_t \in (0, 1]^{H \times W}$ mentioned in Eq. (5) can be replaced by $\Omega(I_t, \tilde{\theta})$ that is computed as:

$$\Omega(I_t, \tilde{\theta}) = \exp(-k\, Var(I_t, \tilde{\theta})), \tag{9}$$

where $k \in \mathbb{R}^+$ indicates the descent degree of the soft weights. We set $k = 20$ in our experiments.

Experiments

Experimental Setup

Implementation Details. We adopt ResNet-50-based (He et al. 2016) LDF (Wei et al. 2020) as our saliency detector. During training, we adopt SYNSOD (11,197 images) as the source domain and the training set of DUTS (Wang et al. 2017) (10,533 images) as the target domain. We set the total number of training rounds to six. The proportion of the selected source and target domain samples are set to $\{1.0, 0.5, 0.25, 0.125, 0.0625, 0.03125\}$ and $\{0.0, 0.1, 0.2, 0.4, 0.6, 0.6\}$ respectively in the six rounds. The source samples are randomly selected while the target samples are selected via the proposed image-level sample selection strategy ISS. We use an SGD optimizer and adopt the linear one cycle learning rate policy (Smith and Topin 2019) to schedule each training round. The whole training process takes about 20 hours with a batch size of 32 on a workstation with a NVDIA GTX 1080 GPU. During testing, each image is resized to 352×352, and fed into the network for saliency prediction without any post-processing. More implementation details are provided in the supplemental materials.

Datasets and Evaluation Metrics. To evaluate the performance of our method, we conduct testing on six real-world benchmark SOD datasets including DUTS-TE (Wang et al. 2017) (5,017 images), ECSSD (Yan et al. 2013) (1,000 images), DUT-O (Yang et al. 2013) (5,168 images), HKU-IS (Li and Yu 2015) (4,447 images), PASCAL-S (Li et al. 2014) (850 images), SOD (Movahedi and Elder 2010) (300 images). We adopt four widely used evaluation metrics, *i.e.*, precision-recall (PR) curve, mean absolute error (MAE, $\mathcal{M}$) (Perazzi et al. 2012), weighted F-measure (F_β^w) (Margolin, Zelnik-Manor, and Tal 2014), and S-measure (S_m) (Fan et al. 2017).

Comparison with State-of-the-Art

Quantitative Comparison. In Table 1, we compare our method with eight fully supervised deep saliency prediction methods: R3Net, DGRL, Capsal, TSPOA, BASNet, MINet, GateNet, LDF, two handcrafted unsupervised methods : MB+, RBD, and five deep weakly-/un-supervised methods: ASMO, MWS, SCRIB, USPS, EDNL. For a fair comparison, we evaluate all the saliency maps provided by the authors with the same evaluation code. As shown in the table, our method consistently outperforms exsiting weakly-supervised and unsupervised SOD methods by a large margin over all six datasets. Specifically our method achieves an average gain of 3.65%, 5.56% 1.61% w.r.t S_m, F_β^w

Method	Sup.	DUTS-TE			ECSSD			DUT-O			HKU-IS			PASCAL-S			SOD		
		$S_m \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$
R3Net (Deng et al. 2018)	F&D	.836	.713	.066	.903	.860	.056	.818	.679	.071	.892	.833	.048	.809	.730	.104	.738	.700	.136
DGRL (Wang et al. 2018)	F&D	.842	.774	.050	.903	.891	.041	.806	.709	.062	.894	.875	.036	.836	.800	.072	.774	.738	.103
Capsal (Zhang et al. 2019)	F&D	.819	.689	.063	.828	.775	.073	.677	.489	.099	.852	.780	.058	.838	.790	.073	.684	.573	.152
TSPOA (Liu et al. 2019)	F&D	.860	.767	.049	.907	.876	.046	.818	.697	.061	.902	.862	.038	.841	.779	.078	.775	.718	.115
BASNet (Wang et al. 2018)	F&D	.866	.803	.048	.916	.904	.037	.836	.751	.056	.909	.889	.032	.836	.795	.077	.772	.728	.112
MINet (Pang et al. 2020)	F&D	.884	.825	.037	**.925**	.911	**.033**	.833	.738	.056	.919	.897	.029	.856	.814	.064	**.805**	**.768**	**.092**
GateNet (Zhao et al. 2020)	F&D	.885	.809	.040	.920	.894	.040	.838	.729	.055	.915	.880	.033	.858	.801	.069	.801	.753	.098
LDF (Wei et al. 2020)	F&D	**.892**	**.845**	**.034**	.924	**.915**	.034	**.839**	**.752**	**.052**	**.919**	**.904**	**.028**	**.862**	**.826**	**.061**	.800	.765	.093
MB+ (Zhang et al. 2015)	U&H	.595	.307	.149	.595	.389	.199	.612	.331	.143	.609	.383	.166	.528	.296	.224	.490	.280	.255
RBD (Zhu et al. 2014)	U&H	.567	.278	.305	.667	.423	.271	.572	.288	.310	.648	.385	.271	.621	.389	.297	.612	.398	.305
ASMO (Li, Xie, and Lin 2018)	W&D	.697	.488	.116	.802	.702	.110	.752	.559	.101	.804	.701	.086	.714	.578	.152	.669	.551	.185
MWS (Zeng et al. 2019)	W&D	.759	.586	.091	.828	.716	.096	.756	.527	.109	.818	.685	.084	.767	.614	.134	.702	.571	.166
USPS (Nguyen et al. 2019)	U&D	.788	.700	.068	.862	.844	.062	.793	.698	.063	.876	.857	.041	.773	.715	.108	.713	.659	.143
EDNL (Zhang, Xie, and Barnes 2020)	U&D	.820	.701	.065	.871	.827	.068	.783	.633	.076	.884	.838	.046	.819	.739	.095	.739	.669	.142
SCRIB (Zhang et al. 2020b)	W&D	.803	.709	.062	.865	.835	.059	.785	.669	.068	.865	.831	.047	.796	.736	.094	.727	.668	.129
Ours	U&D	**.846**	**.783**	**.050**	**.899**	**.885**	**.043**	**.808**	**.711**	**.059**	**.897**	**.879**	**.035**	**.822**	**.773**	**.080**	**.788**	**.750**	**.095**

Table 1: Quantitative comparison with state-of-the-art SOD methods on six datasets in terms of S-measure $S_m \uparrow$, weighted F-measure $F_\beta^w \uparrow$, and MAE $\mathcal{M} \downarrow$. $\uparrow$ and $\downarrow$ indicate larger and smaller is better, respectively. The best performance of fully-supervised and weak-/un-supervised methods is marked in **bold**, respectively. 'Sup.' denotes supervision type. 'F&D' means fully-supervised and deep learning-based methods. 'U&H' means unsupervised and handcrafted methods. 'W&D' refers to weakly-supervised and deep learning-based methods. 'U&D' means unsupervised and deep learning-based methods.

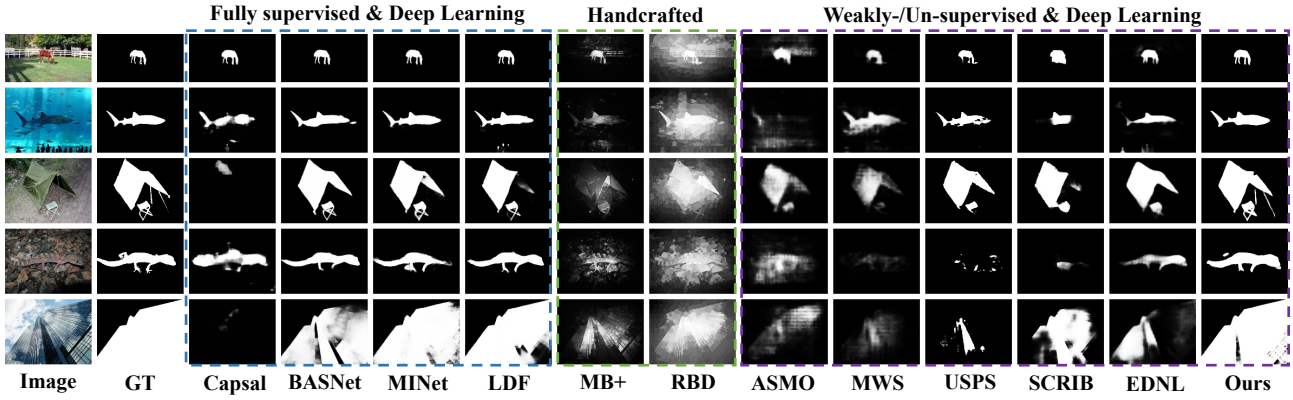


Figure 6: Visual comparisons of different types of SOD methods, where each row displays an input image. Our proposed method (Ours) consistently generates saliency maps close to the ground truth (GT).

and $\mathcal{M}$ compared with previous state-of-the-art weakly-supervised method SCRIB (Zhang et al. 2020b) on six datasets. As for previous state-of-the-art deep unsupervised method EDNL (Zhang, Xie, and Barnes 2020), our approach obtains an average gain of 2.4%, 6,23%, 2.15% w.r.t S_m, F_β^w and $\mathcal{M}$ over six datasets. Moreover, the performance of our proposed UDASOD method is comparable to state-of-the-art fully-supervised SOD methods, and even better than several of them, such as R3Net (Deng et al. 2018), DGRL (Wang et al. 2018), TSPOA (Liu et al. 2019). Fig. 5 presents the precision-recall curves of different SOD methods on five datasets, where weakly-/un-supervised methods are represented by dotted lines. From the figure, we can observe that our method overall lies above other weakly-/un-supervised methods and is even comparable to some fully supervised methods.

Qualitative Comparison. Fig. 6 presents several representative visual examples of predicted saliency maps. These examples reflect various scenarios, including small object (1st row), object with a complex background (2nd row), object with thread-like boundary (3rd row), low contrast between salient object and image background (4th row), and object with a border-connected region (5th row). It can be seen that our proposed method produces accurate and complete saliency maps with sharp boundaries and coherent details, which consistently outperforms the weakly-/un-supervised models and even some fully supervised models.

Ablation Study

Effectiveness of UDASOD. To demonstrate the effectiveness of our proposed unsupervised domain adaptive salient object detection (UDASOD) through the uncertainty-aware pseudo-label learning (UPL) strategy, we conduct the ablation study from the following aspects and report the performance of different variants in Table 2.

1) Synthetic Data. The saliency detector trained with only synthetic source data (Source only) achieves comparable performance to other unsupervised models (as shown in Table 1), indicating the feasibility of learning salient object detection from the proposed synthetic dataset SYNSOD.

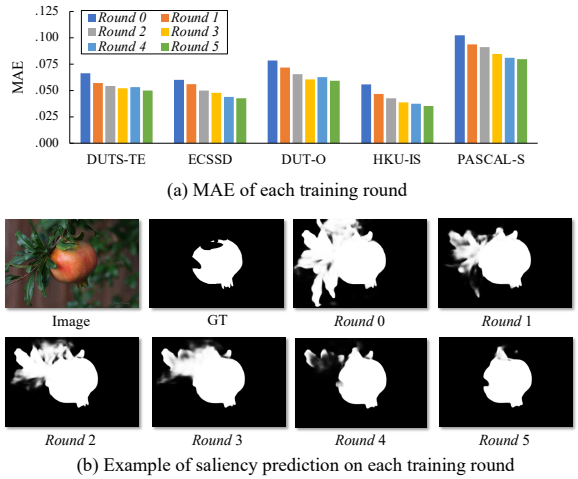
2) Unsupervised Domain Adaption. Introducing the unlabeled real target data through vanilla pseudo-label learning (Vanilla PL) strategy can improve the performance of source only model, which demonstrates that a simple unsu-

Method	Training		UPL		DUTS-TE (Wang et al. 2017)			ECSSD (Yan et al. 2013)			DUT-O (Yang et al. 2013)			HKU-IS (Li and Yu 2015)			PASCAL-S (Li et al. 2014)		
	Source	Target	ISS	PPR	$S_m \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$
Source only	√				.802	.695	.066	.873	.836	.060	.752	.608	.079	.863	.814	.056	.795	.725	.103
Vanilla PL	√	√			.818	.724	.067	.877	.845	.057	.769	.643	.084	.875	.836	.049	.800	.732	.100
UPL w/o ISS	√	√		√	.823	.741	.063	.883	.861	.052	.778	.666	.077	.880	.850	.044	.805	.749	.092
UPL w/o PPR	√	√	√		.842	.777	.052	.894	.878	.046	.803	.704	.064	.894	.875	.037	.822	.774	.082
UPL (Ours)	√	√	√	√	**.846**	**.783**	**.050**	**.899**	**.885**	**.043**	**.808**	**.711**	**.059**	**.897**	**.879**	**.035**	**.822**	**.774**	**.080**

Table 2: Ablation study on five benchmark datasets using S-measure $S_m \uparrow$, weighted F-measure $F_\beta^w \uparrow$, and MAE $\mathcal{M} \downarrow$.

Augmentation	DUTS-TE		ECSSD		DUT-O		HKU-IS	
	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$
Scale	.765	.055	.873	.049	.664	.072	.869	.039
FDA	.774	.053	.872	.047	.700	.067	.874	.038
Flip	.777	.053	.879	.045	.697	.066	.875	.038
Flip+Scale+FDA	**.783**	**.050**	**.885**	**.043**	**.711**	**.059**	**.879**	**.035**

Table 3: Sensitivity to different kinds of data augmentation in consistency-based uncertainty estimation.



(a) MAE of each training round

(b) Example of saliency prediction on each training round

Figure 7: Quantitative and visual performance of our proposed method on each training round.

pervised domain adaption through pseudo-label learning can help to mitigate the domain gap between the synthetic and real domains. While our proposed method (UPL) can further boost the performance of vanilla PL by a large margin by exploiting image-level sample selection (ISS) and pixel-level pseudo-label reweighting (PPR).

3) Uncertainty-Aware Pseudo-Label Learning. To further verify the effectiveness of each component in the proposed UPL. We conduct the ablation study by removing PPR and ISS from UPL, respectively, *i.e.*, UPL w/o PPR and ISS w/o ISS in Table 2. Compared to UPL, the performance of UPL w/o PPR slightly drops on five datasets, which indicates that the selected low uncertainty pseudo-labels still contain some misclassified pixels and the PPR module can alleviate the noise of pseudo labels by adjusting the weights of pixels. UPL w/o ISS is iteractively trained with all the target pseudo-labels without image-level selection, resulting in a severe performance degradation compared to UPL. Theoretically, image-level selection can be approximated as a special case of pixel-level reweighting. However, in practice, using only pixel-level reweighting (UPL w/o ISS) performs worse than image-level selection (UPL w/o PPR). We conjecture that without image-level selection, the pseudo-labels of those high-uncertainty samples naturally have lots of misclassified pixels that will be suppressed by the pixel-wise reweighting. As suggested by(Shin et al. 2020), this will lead to sparse pseudo-labels and inevitably increase the difficulty of network convergence. Whereas, the ISS and PPR modules are complementary to each other and can further boost the performance of our proposed method.

Sensitivity to Data Augmentation. Our proposed method leverages multiple data augmentations as noise injection methods to estimate the uncertainty of pseudo-labels. To demonstrate that our method is applicable to different data augmentations, we report the performance using the augmentations mentioned in Sec. . As shown in Table 3, our proposed method is not limited to a single kind of data augmentation. When applying only one data augmentation (*i.e.*, Flip, Scale, FDA) the proposed uncertainty-aware pseudo-label learning (UPL) strategy can still work and outperform the vanilla pseudo-label learning strategy (Vanilla PL in Table 2) by a large margin, which indicates the robustness of our proposed UPL. Moreover, when combining different data augmentations (Flip+Scale+FDA), the performance of UPL can be further improved as the combination leads to more stable uncertainty measurement.

Sensitivity to Training Rounds. Our proposed method adopts an iterative training paradigm that contains multiple rounds. To show the performance of each training round more intuitively, we present the MAE results and predicted saliency maps in Fig. 7. As shown in Fig. 7 (a), MAE is consistently improved with the increase of training rounds over all datasets. Moreover, as shown in Fig. 7 (b), the non-salient pixels of the predicted saliency map are gradually suppressed and lead to a more accurate result.

Conclusion

In this paper, we propose to tackle deep unsupervised salient object detection from a novel perspective, *i.e.*, learning from synthetic but clean labels. To achieve this goal, we construct a new synthetic salient object detection dataset and introduce a novel unsupervised domain adaptive salient object detection framework to learn and adapt from the synthetic dataset. Specifically, the proposed algorithm exploiting an uncertainty-aware pseudo-label learning strategy to mitigate the domain gap between the synthetic source domain and the real target domain. Extensive experiments on multiple benchmark datasets demonstrate the effectiveness and robustness of our proposed method, which makes it superior to all state-of-the-art deep unsupervised methods and even comparable to fully-supervised methods.

Acknowledgements

This work was supported in part by the Chinese Key-Area Research and Development Program of Guangdong Province (2020B0101350001), in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No.2020B1515020048, in part by the National Natural Science Foundation of China under Grant No.61976250, No.U1811463 and No.61906049, and in part by the Guangdong Provincial Key Laboratory of Big Data Computing, the Chinese University of Hong Kong, Shenzhen.

References

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, 3339–3348.

Chen, Y.-H.; Chen, W.-Y.; Chen, Y.-T.; Tsai, B.-C.; Frank Wang, Y.-C.; and Sun, M. 2017. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the IEEE International Conference on Computer Vision*, 1992–2001.

Chen, Z.; Xu, Q.; Cong, R.; and Huang, Q. 2020. Global context-aware progressive aggregation network for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10599–10606.

Deng, Z.; Hu, X.; Zhu, L.; Xu, X.; Qin, J.; Han, G.; and Heng, P.-A. 2018. R3net: Recurrent residual refinement network for saliency detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 684–690.

Fan, D.-P.; Cheng, M.-M.; Liu, J.-J.; Gao, S.-H.; Hou, Q.; and Borji, A. 2018. Salient objects in clutter: Bringing salient object detection to the foreground. In *Proceedings of the European Conference on Computer Vision*, 186–202.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, 4548–4557.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, 1989–1998.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3.

Li, G.; Kang, G.; Liu, W.; Wei, Y.; and Yang, Y. 2020. Content-consistent matching for domain adaptive semantic segmentation. In *European Conference on Computer Vision*, 440–456.

Li, G.; Xie, Y.; and Lin, L. 2018. Weakly supervised salient object detection using image labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Li, G.; Xie, Y.; Lin, L.; and Yu, Y. 2017. Instance-level salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2386–2395.

Li, G.; and Yu, Y. 2015. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5455–5463.

Li, X.; Lu, H.; Zhang, L.; Ruan, X.; and Yang, M.-H. 2013. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2976–2983.

Li, Y.; Hou, X.; Koch, C.; Rehg, J. M.; and Yuille, A. L. 2014. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 280–287.

Liu, Y.; Zhang, Q.; Zhang, D.; and Han, J. 2019. Employing deep part-object relationships for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1232–1241.

Luo, Y.; Liu, P.; Guan, T.; Yu, J.; and Yang, Y. 2019. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6778–6787.

Margolin, R.; Zelnik-Manor, L.; and Tal, A. 2014. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Movahedi, V.; and Elder, J. H. 2010. Design and perceptual validation of performance measures for salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 49–56.

Nguyen, D. T.; Dax, M.; Mummadi, C. K.; Ngo, T.-P.-N.; Nguyen, T. H. P.; Lou, Z.; and Brox, T. 2019. DeepUSPS: Deep Robust Unsupervised Saliency Prediction With Self-Supervision. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9413–9422.

Perazzi, F.; Krähenbühl, P.; Pritch, Y.; and Hornung, A. 2012. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 733–740.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7479–7489.

Sener, O.; Song, H. O.; Saxena, A.; and Savarese, S. 2016. Learning transferrable representations for unsupervised domain adaptation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2118–2126.

Shin, I.; Woo, S.; Pan, F.; and Kweon, I. S. 2020. Two-phase Pseudo Label Densification for Self-training based Domain Adaptation. In *European Conference on Computer Vision*, 532–548.

Smith, L. N.; and Topin, N. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, 1100612. International Society for Optics and Photonics.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5552–5560.

Tsai, Y.-H.; Hung, W.-C.; Schulter, S.; Sohn, K.; Yang, M.-H.; and Chandraker, M. 2018. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, 7472–7481.

Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; and Ruan, X. 2017. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 136–145.

Wang, T.; Zhang, L.; Wang, S.; Lu, H.; Yang, G.; Ruan, X.; and Borji, A. 2018. Detect globally, refine locally: A novel approach to saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3127–3135.

Wei, J.; Wang, S.; and Huang, Q. 2020. F^3Net: Fusion, Feedback and Focus for Salient Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12321–12328.

Wei, J.; Wang, S.; Wu, Z.; Su, C.; Huang, Q.; and Tian, Q. 2020. Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13025–13034.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised Data Augmentation for Consistency Training. In *Advances in Neural Information Processing Systems*, volume 33, 6256–6268.

Yan, P.; Li, G.; Xie, Y.; Li, Z.; Wang, C.; Chen, T.; and Lin, L. 2019. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7284–7293.

Yan, Q.; Xu, L.; Shi, J.; and Jia, J. 2013. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1155–1162.

Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2013. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3166–3173.

Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4085–4095.

Zeng, Y.; Zhuge, Y.; Lu, H.; Zhang, L.; Qian, M.; and Yu, Y. 2019. Multi-source weak supervision for saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6074–6083.

Zhang, D.; Han, J.; and Zhang, Y. 2017. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE International Conference on Computer Vision*, 4048–4056.

Zhang, J.; Dai, Y.; Zhang, T.; Harandi, M. T.; Barnes, N.; and Hartley, R. 2020a. Learning Saliency from Single Noisy Labelling: A Robust Model Fitting Perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J.; Sclaroff, S.; Lin, Z.; Shen, X.; Price, B.; and Mech, R. 2015. Minimum barrier salient object detection at 80 fps. In *Proceedings of the IEEE International Conference on Computer Vision*, 1404–1412.

Zhang, J.; Xie, J.; and Barnes, N. 2020. Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection. In *Proceedings of the European Conference on Computer Vision*, 349–366.

Zhang, J.; Yu, X.; Li, A.; Song, P.; Liu, B.; and Dai, Y. 2020b. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12546–12555.

Zhang, J.; Zhang, T.; Dai, Y.; Harandi, M.; and Hartley, R. 2018. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9029–9038.

Zhang, L.; Zhang, J.; Lin, Z.; Lu, H.; and He, Y. 2019. Capsal: Leveraging captioning to boost semantics for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6024–6033.

Zhao, X.; Pang, Y.; Zhang, L.; Lu, H.; and Zhang, L. 2020. Suppress and balance: A simple gated network for salient object detection. In *Proceedings of European Conference on Computer Vision*, 35–51.

Zheng, Z.; and Yang, Y. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 1–15.

Zhu, W.; Liang, S.; Wei, Y.; and Sun, J. 2014. Saliency optimization from robust background detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2814–2821.

Zou, Y.; Yu, Z.; Kumar, B.; and Wang, J. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision*, 289–305.

Zou, Y.; Yu, Z.; Liu, X.; Kumar, B.; and Wang, J. 2019. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5982–5991.