Uncertainty-Aware Active Domain Adaptive Salient Object Detection

Guanbin Li, *Member, IEEE*, Zhuohua Chen, Mingzhi Mao, Liang Lin, *Fellow, IEEE*, and Chaowei Fang, *Member, IEEE*

***Abstract*— Due to the advancement of deep learning, the performance of salient object detection (SOD) has been significantly improved. However, deep learning-based techniques require a sizable amount of pixel-wise annotations. To relieve the burden of data annotation, a variety of deep weakly-supervised and unsupervised SOD methods have been proposed, yet the performance gap between them and fully supervised methods remains significant. In this paper, we propose a novel, cost-efficient salient object detection framework, which can adapt models from synthetic data to real-world data with the help of a limited number of actively selected annotations. Specifically, we first construct a synthetic SOD dataset by copying and pasting foreground objects into pure background images. With the masks of foreground objects taken as the ground-truth saliency maps, this dataset can be used for training the SOD model initially. However, due to the large domain gap between synthetic images and real-world images, the performance of the initially trained model on the real-world images is deficient. To transfer the model from the synthetic dataset to the real-world datasets, we further design an uncertainty-aware active domain adaptive algorithm to generate labels for the real-world target images. The prediction variances against data augmentations are utilized to calculate the superpixel-level uncertainty values. For those superpixels with relatively low uncertainty, we directly generate pseudo labels according to the network predictions. Meanwhile, we select a few superpixels with high uncertainty scores and assign labels to them manually. This labeling strategy is capable of generating high-quality labels without incurring too much annotation cost. Experimental results on six benchmark SOD datasets demonstrate that our method outperforms the existing state-of-the-art weakly-supervised and unsupervised SOD methods and is even comparable to the fully supervised ones. Code will be released at: https://github.com/czh-3/UADA.**

***Index Terms*— Salient object detection, domain adaptation, active learning.**

Manuscript received 1 July 2023; revised 7 April 2024; accepted 31 May 2024. Date of publication 18 June 2024; date of current version 4 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62376206 and Grant 62322608 and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515010255. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaolin Hu. *(Corresponding author: Chaowei Fang.)*

Guanbin Li and Liang Lin are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: liguanbin@mail.sysu.edu.cn; linliang@ieee.org).

Zhuohua Chen and Mingzhi Mao are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: chenzhh79@mail2.sysu.edu.cn; mcsmmz@mail.sysu.edu.cn).

Chaowei Fang is with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: chaoweifang@outlook.com).

Digital Object Identifier 10.1109/TIP.2024.3413598

I. Introduction

SALIENT object detection (SOD) aims to accurately segment visually distinctive object regions within an image. Traditional methods heavily rely on the hand-crafted features [1], [2], [3], which often lack representational power and result in poor generalization performance. Recent advancements in deep learning have significantly improved the performance of SOD by leveraging large-scale pixel-wise labeled datasets [4], [5], [6]. However, the acquisition of such high-quality and large-scale annotations is time-consuming and costly [7], [8], thereby limiting the practical applicability of deep learning-based SOD methods.

To alleviate the burden of data annotation, there has been considerable research interest in weakly-supervised [9], [10], [11] and unsupervised [12], [13], [14] SOD algorithms. Weakly-supervised methods utilize less expensive annotation forms, such as image-level categories [9], captions [10], and scribbles [11] to train models. Unsupervised SOD methods aim to learn models without any manual annotations by leveraging noisy pseudo-labels [13], [15], [16] generated by traditional unsupervised SOD methods. However, the simple annotations used in weakly-supervised methods provide limited supervision, while the quality of pseudo-labels generated by traditional methods is compromised. As a result, a substantial performance gap still exists between weakly-supervised or unsupervised approaches and fully-supervised methods.

In this paper, we propose an approach to alleviate the data annotation burden in SOD by leveraging a synthetic saliency detection dataset. Specifically, we collect a large number of images with pure background content and insert foreground objects into these backgrounds, creating synthetic images with salient foreground objects. The masks of these inserted objects serve as ground-truth saliency maps. Two examples of synthetic images are illustrated in Fig. 1 (a). This synthetic dataset allows for training SOD models without incurring additional data annotation costs. However, training models solely on the synthetic dataset leads to suboptimal performance on real-world images due to the domain gap between synthetic and real-world domains.

To address this limitation, we introduce a novel method named Uncertainty-Aware Active Domain Adaptive (UADA) SOD to transfer the saliency detector trained on the synthetic dataset to real-world datasets. The main differences between our adaptation algorithm and existing weakly supervised SOD methods include two folds: 1) producing pseudo-labels along

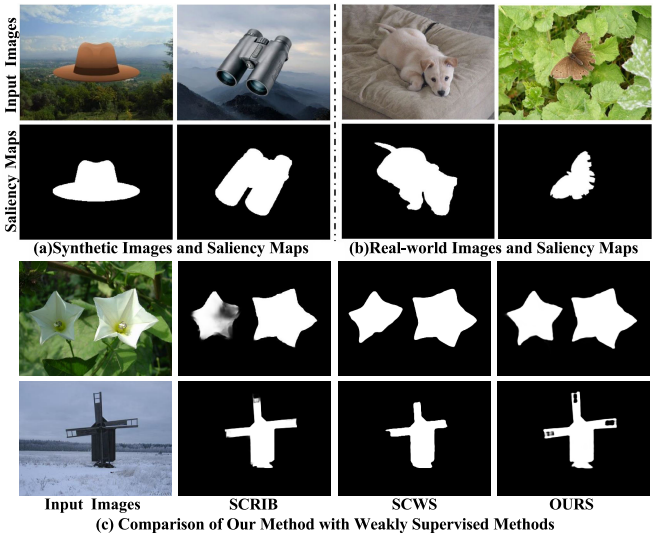


Fig. 1. The contrast between the foreground object and the background content is apparently distinct between synthetic images (a) and real-world images (b). In (c), We present two examples for comparing our method with existing weakly supervised methods SCRIB [11] and SCWS [17].

with their certainty levels from a saliency detector pre-trained with a synthetic dataset; 2) generating supervision information for target images by combining pseudo-labels and few manually collected labels. Pseudo-labels are obtained from the model's predictions and are widely used in unsupervised domain adaptation for exploring unlabeled images [18], [19]. However, these pseudo-labels may contain noise that hinders the optimization of network parameters. To evaluate the reliability of pseudo-labels, we propose a superpixel-level uncertainty estimation algorithm based on the inference variance among differently augmented images. This algorithm helps identify unreliable pseudo-labels with larger inference variances, which indicate a higher likelihood of noise. Conversely, unreliable pseudo-labels often correspond to more challenging samples, which are crucial for improving the robustness of the SOD model. Therefore, we design an active superpixel-level labeling system to create manual labels for these challenging samples without incurring significant annotation costs.

We train the salient object detection model using the synthetic dataset as the source domain and the DUTS dataset [20] as the target domain. During testing, we evaluate the model on six public benchmarks, including DUTS [20], ECSSD [21], DUT-O [22], HKU-IS [23], PASCAL-S [24], and SOD [25]. The experimental results indicate that our method achieves state-of-the-art performance. We provide two examples in Fig. 1 (c), which demonstrate that our method achieves more accurate saliency predictions than existing weakly supervised methods, SCRIB [11] and SCWS [17].

Main contributions of this paper are summarized as follows:

- We construct a synthetic SOD dataset and make the first attempt to learn the saliency detector from the synthetic dataset via domain adaptation and active labeling techniques, which varies from existing deep weakly supervised SOD algorithms based on coarse labels.
- We propose an active domain adaptive SOD algorithm that exploits reliable pseudo labels acquired via superpixel-level uncertainties and minima manual labels obtained by active labeling to adapt the saliency detector trained on the synthetic dataset to real-world scenarios.
- Our method outperforms existing weakly-supervised or unsupervised SOD methods and is comparable to fully supervised methods, as indicated by evaluation results on six benchmark datasets, namely DUTS, DUT-O, ECSSD, HKU-IS, PASCL-S, and SOD.

II. Related Work

A. Salient Object Detection

This subsection provides a brief introduction to literature of salient object detection (SOD) which is closely related to our paper. A comprehensive survey about this field can be referred to [26] which summarizes various types of SOD methods and offers extensive analysis about the robustness and generalization of those SOD methods.

Traditional SOD methods [1], [2], [3], [27], [28] rely on saliency priors and handcrafted features. However, recent advancements are made by leveraging deep convolutional neural networks (DCNNs), leading to significant improvements in SOD performance [29], [30], [31], [32], [33]. Li and Yu [34] set up a large-scale benchmark dataset and devise a DCNN model based on hierarchical features for implementing superpixel-wise saliency inference. Li et al. [7] propose a multi-scale framework which uses a contour prediction branch to achieve instance-level salient object extraction. Apart from a pixel-wise prediction branch, a superpixel-wise inference branch is incorporated to generate saliency maps with clearer boundaries in [35]. Wang et al. [6] propose to merge the multiple features recurrently from the bottom level to the top level. In [36], a complementary information selection module is devised to aggregate different levels of features selectively, and a recursive feature feedback mechanism is applied to eliminate the differences among different feature levels. Pang et al. [37] design inter-level and intra-level feature interaction modules to make full use of multi-level features. A few methods [38], [39] are dedicated to preventing information dilution during the top-down decoding process. Wei et al. [4] decouple the saliency map into a body map and a detail map to alleviate the imbalance in edge pixel distribution. Despite their effectiveness, DCNN-based approaches are data-hungry and heavily depend on a substantial number of manually annotated pixel-wise labels, posing significant challenges and costs.

To address the laborious data annotation process, researchers have proposed weakly supervised SOD methods that learn from less expensive forms of supervision, such as image-level categories [9], captions [10], scribbles [11], and sketches [40]. These approaches avoid the reliance on costly pixel-wise annotations by generating coarse pixel-wise annotations from weak supervision information. Yu et al. [17] propose a local coherence loss to propagate labels to unlabeled regions based on image features and pixel distance. Additionally, a series of methods [13], [15], [16], [41], [42] focus on learning saliency prediction models from automatically generated annotations, instead of relying on manual annotations. They rely on traditional unsupervised methods to generate pseudo-labels for training images. Due

to the lack of high-quality pixel-wise annotations, these weakly-supervised or unsupervised methods still have a large room for performance improvement.

In contrast to the aforementioned methodologies, we propose a novel perspective for SOD that involves learning from synthetic but clean labels. The fundamental idea is to construct a synthetic dataset consisting of images with pure background content. Foreground objects are then inserted into these background images, resulting in images with salient foreground objects. The masks of these inserted objects serve as ground-truth saliency maps, providing accurate and reliable labels for model training. With this synthetic dataset, we can effectively reduce the burden of data annotation while still achieving high-quality results.

B. Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) is a research field that focuses on transferring knowledge obtained from a labeled source domain to an unlabeled target domain. It has garnered significant attention in various computer vision tasks, including image classification [43], pose estimation [44], object detection [45], and semantic segmentation [46]. Among these tasks, semantic segmentation is the most closely related to salient object detection (SOD) due to their shared goal of pixel-level labeling.

Recent advances in UDA for semantic segmentation can be broadly categorized into two main approaches. The first approach aims to align the feature representations between the source and target domains. This can be achieved through techniques such as style transfer in the input space [47], feature space alignment [48], [49], or adversarial learning [50]. By reducing the distribution discrepancy between the domains, these methods facilitate knowledge transfer from the source to the target domain.

The second approach in UDA involves generating pseudo-labels for confident samples in the target domain [51], [52]. Practically, pseudo-labels are assigned to target domain samples based on their confidence scores generated by the current model. Retraining the model with the target domain data and pseudo-labels helps to eliminate the domain gap.

In the context of salient object detection, we present the first exploration of domain adaptation techniques. Specifically, we investigate domain adaptation for SOD by considering synthetic images as the source domain data. For real-world images, we can generate pseudo-labels from the model prediction and estimate superpixel-level uncertainty according to the prediction consistency across data augmentations. Then, superpixels with low uncertainty levels are utilized for fine-tuning the model, thus bridging the domain gap between synthetic and real-world images. This pioneering approach opens up new possibilities for improving the performance of SOD models in real-world scenarios through domain adaptation.

C. Active Learning

Active learning has emerged as a crucial area of research in SOD, with the objective of training models using a minimal number of annotations. The central challenge in active learning lies in the selection of samples that carry the highest importance and informative value during the training process. Various sample selection strategies have been proposed for image classification and semantic segmentation, including the uncertainty-based approach [53], [54], [55], diversity-based approach [56], [57], and expected model output change [58]. A comprehensive analysis and evaluation for different types of active learning methods is provided in [59]. This kind of techniques has been adopted for boosting domain adaptation algorithms for image classification [60], [61] and semantic segmentation [62]. To enhance the annotation quality for real-world images, we sample out those informative superpixels with high prediction uncertainty levels and allocate manual labels to them. With this superpixel selecting and labeling process, we can supply informative supervision information efficiently for further improving the model performance.

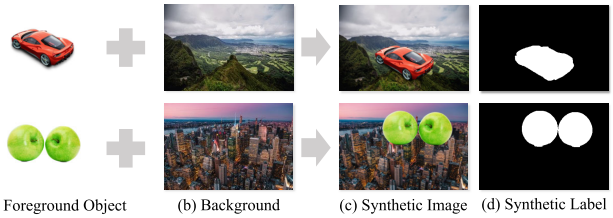


Fig. 2. Illustration of the generation procedure of SYNSOD dataset. We first collect foreground object images (a) with transparent background regions and cluttered images (b) with non-salient scenes. Then, the foreground images are pasted into the cluttered images, forming synthetic images (c). The masks of foreground objects are regarded as synthetic labels of saliency maps (d).

III. Synthetic Salient Object Detection Dataset

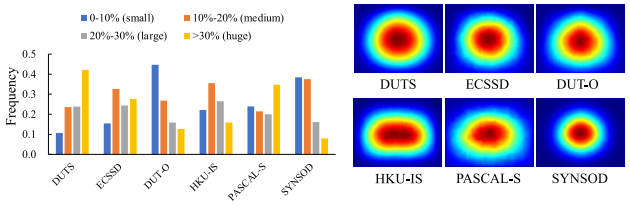
In this section, we introduce the construction process of our synthetic salient object detection (SYNSOD) dataset and provide a comprehensive illustration about its statistics.

A. Dataset Collection

To facilitate the training of salient object detection models, we propose an image synthesis approach that leverages the inherent characteristics of salient objects. Given that salient objects predominantly correspond to foreground objects within images, we synthesize training images by seamlessly integrating foreground objects into background images devoid of salient content.

To generate synthetic images with salient objects, we gather a substantial collection of foreground object images (refer to Fig.2 (a)) and background images (refer to Fig.2 (b)) from diverse non-copyrighted image sources. Employing a copy-paste strategy, we generate synthetic images by compositing foreground objects onto background images. The composition results can be observed from Fig. 2 (c).

The image synthesis process encompasses the following steps: 1) Randomly selecting a pair of a background image and a foreground object image; 2) Resizing the foreground object image using a scaling factor randomly drawn from the range of $[0.5, 1.1]$, followed by a random spatial shift; 3) Compositing the foreground object image and the background image using an alpha channel, where the transparency level of the background content is 100%. The pixel-wise



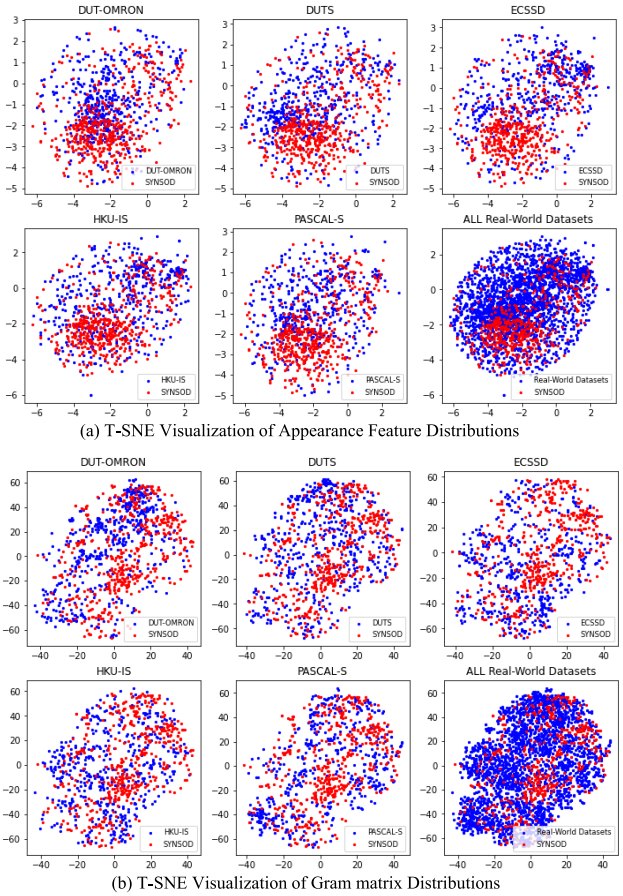
(a) Distribution of object size (b) Center bias

Fig. 3. Statistics of our SYNSOD dataset and existing real-world SOD datasets including DUTS [20], ECSSD [21], DUT-O [22], HKU-IS [23], and PASCAL-S [24]. The object size distribution and center bias have significant difference between our SYNSOD dataset and existing real-world SOD datasets.

salient object annotation is obtained by applying a threshold of 0.5 to the alpha channel. Following this image synthesis procedure, we construct a synthetic salient object detection dataset, referred to as SYNSOD, comprising 11,197 synthetic images accompanied by pixel-wise annotations.

B. Dataset Statistics

We visualize the object size and center bias statistics of our SYNSOD dataset and five existing SOD datasets [20], [21], [22], [23], [24] in Fig. 3. We categorize salient objects into four groups according to their spatial size evaluated with the ratio between the salient object area and full image area, i.e., small $(0-10\%)$, medium $(10-20\%)$, large $(20-30\%)$, and huge $(>30\%)$ objects. The frequency values of four groups are illustrated in Fig. 3 (a). The size of most salient objects in SYNSOD is less than 20%, and the average is 14.72%. For presenting the center bias of all datasets, we calculate the average of saliency maps and visualize it in Fig. 3 (b). The analysis of both object size and center bias statistics indicates the existence of a moderate domain gap between our SYNSOD dataset and real-world SOD datasets. To demonstrate the differences in appearance and style between our synthetic dataset and real-world datasets, we randomly select 500 images from each dataset and use VGG19 to extract their appearance features as well as gram matrices. Then, t-SNE is used to visualize these appearance features and gram matrices as shown in Fig. 4. It can be observed that there exists evident distribution gap in appearance features between our synthetic dataset and most real-world datasets, whereas the differences in gram matrix distributions are minor.



(a) T-SNE Visualization of Appearance Feature Distributions

(b) T-SNE Visualization of Gram matrix Distributions

Fig. 4. T-SNE visualization of appearance features and gram matrices distributions between synthetic dataset and real-world datasets. The first five images of each subplot show the differences between our synthetic dataset and single real-world datasets, while the last image shows the differences between the synthetic dataset and the union of five real-world datasets. The appearance feature distribution of our synthetic dataset differs evidently from that of real-world datasets, while the differences in gram matrices are relatively minor.

IV. Proposed Methodology

In this section, we demonstrate the technical details of our proposed method. The method in this paper is an extension of our conference paper [63], with new contributions as follows: 1) Aiming to improve the quality of pseudo-labels on target domain images, we propose a superpixel-level uncertainty estimation strategy based on the inference variance among differently augmented images. 2) To use pixels with unreliable pseudo-labels, we devise an active labeling system that assigns manual labels to superpixels with high uncertainty values.

A. Problem Formulation

To alleviate the burden of data annotation in salient object detection, this paper proposes to adapt models from synthetic source images to real-world target images. Only a limited number of superpixel-level labels are allowed for each target image. Let us denote the synthetic source image set as $\mathcal{X}^s = \{\mathbf{x}_i^s\}_{i=1}^{N^s}$, where $\mathbf{x}_i^s \in \mathbb{R}^{H\times W\times 3}$ represents an image from our SYNSOD dataset, and N^s is the number of synthetic source images. Additionally, $\mathcal{Y}^s = \{\mathbf{y}_i^s\}_{i=1}^{N^s}$ represents the set of ground-truth saliency maps corresponding to $\mathcal{X}^s$. The target image dataset is composed of N^t real-world images, denoted as $\mathcal{X}^t = \{\mathbf{x}_j^t\}_{j=1}^{N^t}$.

To tackle the aforementioned problem, we propose a novel uncertainty-aware active domain adaptive (UADA) algorithm, as illustrated in Fig. 5. The algorithm consists of K rounds: in the first round, the model is pre-trained with labeled source images, and in the remaining rounds, the model is fine-tuned with both source and target images. In each round except the first, pseudo-labels are generated for the target images by applying a threshold to the prediction confidences of the model obtained from the previous round. Pixel-level and superpixel-level uncertainties of the predicted saliency map are estimated by analyzing the output variance across different image augmentations. Subsequently, the training loss is re-weighted based on the pixel-level uncertainty values, with

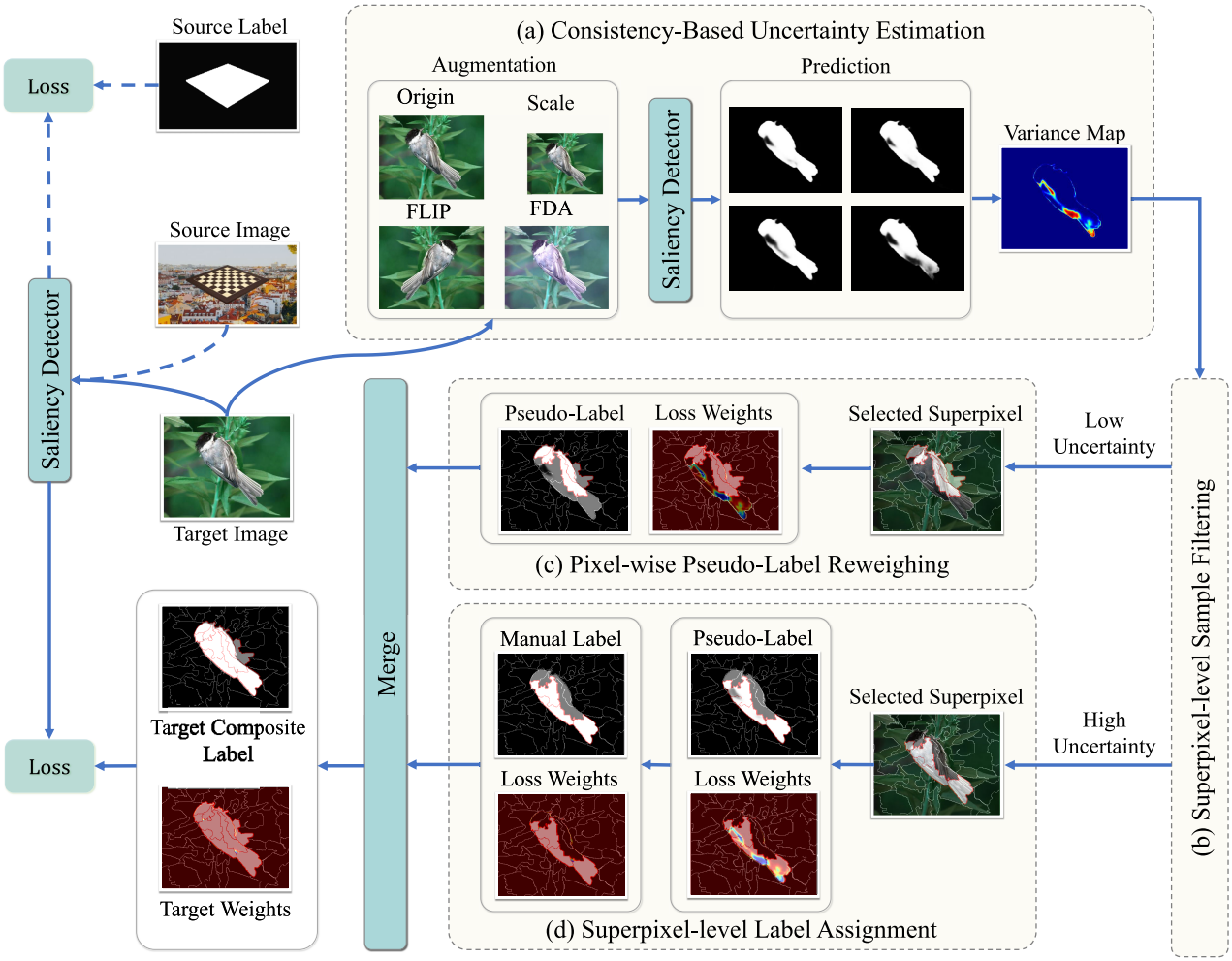


Fig. 5. The overall framework of the proposed active domain adaptive SOD method. It learns saliency prediction from source domain images with synthetic labels and target domain images with composite labels formed by combining reliable pseudo-labels and superpixel-level manual labels. Here, we first estimate the pixel-wise uncertainties based on the prediction variances across multiple augmented versions of the input image, including the original image and its augmented variants generated by scaling, flipping, and Fourier domain adaptation (FDA). Then, superpixel-level uncertainties are calculated by averaging pixel-wise uncertainties within every superpixel. Afterwards, we generate pseudo-labels along with loss weights for pixels inside super-pixels with relatively low uncertainties, while assigning manual labels to those pixels inside super-pixels with high uncertainties.

low uncertainty pixels receiving higher weights. Additionally, an active labeling strategy is devised by manually annotating superpixels with high uncertainty values.

The training process of each round can be formulated as the optimization of network parameters with the following objective function:

$$\mathcal{L} = \mathcal{L}(\mathcal{X}^s, \mathcal{Y}^s, \Omega^s) + \gamma \mathcal{L}(\mathcal{X}^t, \hat{\mathcal{Y}}^t, \Omega^t), \tag{1}$$

where $\hat{\mathcal{Y}}^t = \{\hat{\mathbf{y}}_j^t\}_{j=1}^{N^t}$ denotes the set of pseudo-labels for the target images, and γ is a constant. Ω^s and Ω^t represent the sets of re-weighting maps of $\mathcal{Y}^s$ and $\hat{\mathcal{Y}}^t$ respectively. Specifically, $\Omega^s = \{\boldsymbol{\omega}_i^s\}_{i=1}^{N_s}$ and $\Omega^t = \{\boldsymbol{\omega}_j^t\}_{j=1}^{N_t}$, where $\boldsymbol{\omega}_i^s$ and $\boldsymbol{\omega}_j^t$ denote the re-weighting maps for $\mathbf{y}_i^s$ and $\hat{\mathbf{y}}_j^t$, respectively. The function $\mathcal{L}(\cdot,\cdot,\cdot)$ accumulates the training loss on source or target domain images and is defined as,

$$\mathcal{L}(\mathcal{X}, \mathcal{Y}, \Omega) = \sum_{i=1}^{|\mathcal{X}|} \sum_{h=1}^{H} \sum_{w=1}^{W} \omega_i^{(h,w)} \ell(p_i^{(h,w)}, y_i^{(h,w)}), \tag{2}$$

where $\ell(\cdot,\cdot)$ represents the binary cross-entropy function. Here, $\mathbf{p}_i \in [0,1]^{H\times W}$ denotes the saliency probability map predicted from the image $\mathbf{x}_i$. Furthermore, $\omega_i^{(h,w)}$, $p_i^{(h,w)}$, and $y_i^{(h,w)}$ refer to the value at pixel (h,w) of $\boldsymbol{\omega}_i$, $\mathbf{p}_i$, and $\mathbf{y}_i$, respectively. The re-weighting maps of the source domain images are set to full one matrices, i.e., $\boldsymbol{\omega}_i^s = \mathbf{1}^{H\times W}$ for $i \in [1, 2, \cdots, N^s]$. In the first training round, γ is set to 0; otherwise, γ is set to 1.

In the subsequent sections, we provide detailed explanations of the pseudo-labeling process for real-world images.

B. Uncertainty-Aware Pixel-Level Pseudo-Labeling

To address the performance degradation caused by the domain gap between synthetic and real-world images, we propose an uncertainty-aware active domain adaptation strategy consisting of four steps.

1) Consistency-Based Uncertainty Estimation: We can assign pseudo-labels to target domain images based on the saliency maps predicted by the current model. However, due to the evident distribution differences between source and target

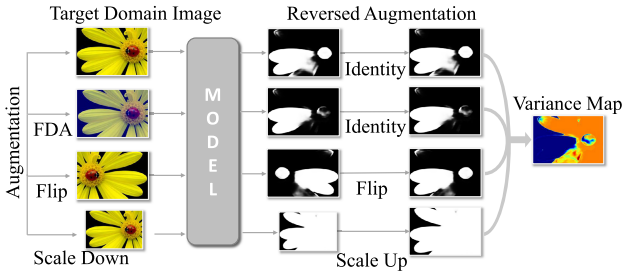


Fig. 6. Procedure for generating the pixel-wise uncertainty map.

domain images, these pseudo labels often contain noise. It is critical to estimate the reliability of pseudo labels and mitigate the impact of unreliable ones. To achieve this, we leverage the smoothness assumption [64], which states that a prediction is accurate if it remains stable under subtle variations in the input. Consequently, we propose estimating the uncertainty of the predicted saliency maps by analyzing variances across different image augmentations.

Practically, we apply reversible augmentation strategies, such as horizontal flipping, rescaling, and Fourier domain adaptation (FDA) [65], to each target domain image. For FDA, a random source image is selected to provide the low-frequency amplitudes. Denote the total number of augmentation strategies as U. For a given target image $\mathbf{x}_j^t$, we generate a new image $\mathbf{x}_{j,u}^t$ using the u-th augmentation strategy, denoted as $\mathbf{x}_{j,u}^t = \mathcal{A}_u(\mathbf{x}_j^t)$, where $\mathcal{A}_u(\cdot)$ represents the u-th image augmentation function. This augmented image $\mathbf{x}_{j,u}^t$ is then fed into the saliency detector, resulting in the saliency map $\mathbf{p}_{j,u}^t$. To ensure consistency, we apply reverse spatial transformations to the saliency map $\mathbf{p}_{j,u}^t$, producing $\tilde{\mathbf{p}}_{j,u}^t$. The reverse spatial transform operation is an identity mapping for augmentation strategies that do not introduce spatial distortions, such as FDA.

Finally, we evaluate the pixel-level uncertainty map $\mathbf{v}_j \in \mathcal{R}^{H\times W}$ using the following formulation,

$$\mathbf{v}_j = \frac{\sum_{u=0}^{U}(\tilde{\mathbf{p}}_{j,u}^t - \frac{1}{U+1}\sum_{u'=0}^{U}\tilde{\mathbf{p}}_{j,u'}^t)^2}{U+1}, \tag{3}$$

where $\tilde{\mathbf{p}}_{j,0}^t = \mathbf{p}_j^t$ represents the saliency map predicted from the original image $\mathbf{x}_j^t$, and the square operation is performed element-wise. This formulation measures the pixel-wise variance among saliency maps predicted from differently augmented images, allowing us to estimate the pseudo-label uncertainty for each target domain image.

2) Superpixel-Level Sample Filtering (SSF): To address the issue of pixel-wise uncertainty estimation, we adopt a superpixel-based approach to re-estimate the uncertainty of pseudo-labels. To implement this, we decompose the target domain image $\mathbf{x}_j^t$ into a set of superpixels $\mathcal{S}_j$ with the SEEDS algorithm [66]. Suppose the number of superpixels within $\mathbf{x}_j^t$ be M_j. We can denote $\mathcal{S}_j$ as $\mathcal{S}_j = \{S_j^{(m)}\}_{m=1}^{M_j}$, where $S_j^{(m)}$ represents the m-th superpixel and can also be described by a binary mask of size $H \times W$ denoted as $\mathbf{S}_j^{(m)}$. The uncertainty value $V_j^{(m)}$ of the superpixel $S_j^{(m)}$ is estimated by averaging the uncertainty values of its inner pixels,

$$V_j^{(m)} = \frac{\sum_{h=1}^{H}\sum_{w=1}^{W} v_j^{(h,w)} S_j^{(m,(h,w))}}{\sum_{h=1}^{H}\sum_{w=1}^{W} S_j^{(m,(h,w))}}, \tag{4}$$

where $v_j^{(h,w)}$ and $S_j^{(m,(h,w))}$ represents the value at the pixel position (h, w) of $\mathbf{v}_j$ and $\mathbf{S}_j^{(m)}$, respectively.

Since the model is relatively weak in the early training stage and is progressively improved during training, we assume that: 1) only pseudo-labels with low uncertainty need to be selected; 2) the number of pseudo-labels involved during training should be gradually increased with respect to the training iteration. During the k-th training round, we rank all superpixels based on their uncertainty values, and select the bottom $0.2\times(k-1)$ ratio of superpixels which have the lowest uncertainty values. These selected superpixels are denoted as $\mathcal{S}_{low}^{(k)}$ and are used to filter pixel-wise pseudo-labels as below,

$$\hat{y}_j^{t,(h,w)} = \begin{cases} p_j^{t,(h,w)}, & \text{if } \mathcal{I}_{sp}(\mathbf{x}_j^t, (h,w)) \in \mathcal{S}_{low}^{(k)}; \\ -1, & \text{otherwise.} \end{cases} \tag{5}$$

Here, $\mathcal{I}_{sp}(\mathbf{x}_j, (h, w))$ maps the pixel (h, w) in $\mathbf{x}_j$ to its corresponding superpixel identity. $\hat{y}_j^{t,(h,w)} = -1$ indicates that the label for pixel (h, w) remains unknown.

3) Pixel-Wise Pseudo-Label Reweighting (PPR): The pseudo-labels filtered by superpixel-level uncertainties as in Eq. (5) may still contain noises. To mitigate the negative effects of noisy pseudo-labels, we propose a pixel-wise pseudo-label reweighting strategy based on the uncertainty map $\mathbf{v}_j$ from Eq.(3). The reweighting map $\boldsymbol{\omega}_j^t$ is calculated as follows:

$$\omega_j^{t,(h,w)} = \begin{cases} e^{-\mu v_j^{(h,w)}}, & \text{if } \mathcal{I}_{sp}(\mathbf{x}_j^t, (h,w)) \in \mathcal{S}_{low}^{(k)}; \\ 0, & \text{otherwise;} \end{cases} \tag{6}$$

where μ is a hyper-parameter for controlling the attenuation degree of the weights.

C. Active Superpixel-Level Labeling

Solely training the model on superpixels with low uncertainty scores in target domain images may lead to suboptimal performance, as it disregards superpixels with high uncertainty scores. To strike a balance between annotation cost and label quality, we propose an active labeling strategy, termed Active Superpixel-level Labeling (ASL), to incorporate these high-uncertainty superpixels during training. At the beginning of every training round, we identify a small percentage of previously unlabeled superpixels with the highest uncertainty scores. Subsequently, we manually annotate each selected superpixel with a dominant label.

For each manually labeled superpixel in the target domain image $\mathbf{x}_j^t$, we update labels of its constituent pixels in $\hat{\mathbf{y}}_j^t$ with the assigned manual label. Moreover, we assign a weight value of 1 to the corresponding entries in the weight map $\boldsymbol{\omega}_j^t$. This labeling strategy enables us to explore high-uncertainty superpixels without incurring a heavy annotation burden, as only a limited number of superpixel-level labels are required.

Algorithm 1 Uncertainty-Aware Active Domain Adaptive Salient Object Detection Algorithm

Require:
Source domain images, $\mathcal{X}^s$; source domain labels, $\mathcal{Y}^s$; source domain weights, Ω^s; target domain images, $\mathcal{X}^t$; number of total training rounds, K; number of epochs during each finetuning round, E;

Initialize the saliency detector and warm up it on source domain dataset $\{\mathcal{X}^s, \mathcal{Y}^s\}$;
Decompose target domain images $\mathcal{X}^t$ into superpixels $\mathcal{S} = \{\mathcal{S}_j\}_{j=1}^{N^t}$;
for $k = 2$ to K **do**
  Augment target domain images and use the current model to predict saliency maps for them, resulting in $\{\tilde{\mathbf{p}}_{j,u}^t | u = 0, \cdots, U\}_{j=1}^{N^t}$;
  Calculate pixel-level uncertainty maps according to Eq. 3, resulting in $\{\mathbf{v}_j\}_{j=1}^{N^t}$;
  Estimate uncertainty scores of target domain images' superpixels by averaging the uncertainty scores of their inner pixels according to Eq. 4, resulting in $\{V_j^{(m)} | m = 1, \cdots, M_j;\ j = 1, \cdots, N^t\}$;
  Generate pseudo labels for target domain images according to Eq. 5, resulting in $\{\hat{\mathbf{y}}_j^t\}_{j=1}^{N^t}$;
  Assign pixel-wise weights to pseudo labels according to Eq. 6, resulting in $\Omega^t = \{\boldsymbol{\omega}_j^t\}_{j=1}^{N^t}$;
  Select superpixels with the highest uncertainty scores and annotate them manually;
  Update values of pixels inside the selected superpixels in $\hat{\mathbf{y}}_j^t$-s and $\boldsymbol{\omega}_j^t$-s with manual labels and 1, respectively;
  for $i = 1$ to E **do**
    Calculate loss on target domain images: $L^t = \mathcal{L}(\mathcal{X}^t, \hat{\mathcal{Y}}^t, \Omega^t)$;
    Select source domain images and calculate loss on them: $L^s = \mathcal{L}(\mathcal{X}^s, \mathcal{Y}^s, \Omega^s)$;
    Update parameters of the saliency detector by minimizing $L^s + L^t$.
  end for
end for

Additionally, superpixels, being composed of visually similar pixels and preserving object boundaries well, tend to yield higher-quality labels compared to pixel-wise pseudo-labels for these high-uncertainty superpixels.

In summary, the training procedure of our proposed method consists of two stages. Initially, the model is pre-trained using source domain images. Subsequently, the target domain images are annotated through a combination of pseudo labels and actively collected manual labels. The complete training procedure is outlined in Algorithm 1.

V. Experiments

A. Experimental Setup

1) Implementation Details: We adopt LDF [4] with ResNet50 [72] backbone as the saliency detector. During training, the proposed synthetic dataset SYNSOD is regarded as the source domain, and the training set of DUTS [20] is regarded as the target domain. We set the number of training rounds K to 4. The number of epochs is set to 48 for the first round, and 24 for the remaining three rounds. During the four rounds, 100%, 40%, 20%, and 10% of the source domain images are randomly selected in turn. In the k-th round ($k \in \{1, 2, 3, 4\}$), the proportion of selected target domain superpixels for pseudo-label learning is set to $0.2(k-1)$. From the second round to the fourth round, 5% additional superpixels are selected for active labeling per round. μ is set to 20. We adopt the SGD optimizer with momentum of 0.9 and weight decay of 5×10^{-4}. The learning rate is scheduled by the one cycle policy [73]. Specifically, in the k-th training round, we set the maximum learning rate to $0.0025 \times 0.9^{k-1}$ for the backbone and $0.05 \times 0.9^{k-1}$ for the detector. The batch size is set to 32 throughout the training process. During testing, images are resized to 352×352.

2) Datasets and Evaluation Metrics: To evaluate the performance of our proposed method, we test on six real-world SOD datasets including DUTS [20], DUT-OMRON [22], ECSSD [21], HKU-IS [23], PASCL-S [24], and SOD [25]. We apply six widely used evaluation metrics, including S-measure (S_m) [74], F-measure (F) [75], weighted F-measure (F_β^w) [76], E-measure (E) [77], and mean absolute error (MAE, M) [78]. Precision-recall (PR) curves are also provided to illustrate the robustness of SOD algorithms.

3) Competing Methods: We compare our method with existing state-of-the-art methods, including 10 fully supervised methods: R3 [67], DGRL [6], TSPOA [68], BAS [5], SCRN [69], AFNET [70], GCPA [38], GateNet [71], MINet [37], and LDF [4]; and 7 weakly-/un-supervised methods: ASMO [9], MNL [14], MWS [10], USPS [13], EDNL [12], SCRIB [11], and SCWS [17].

B. Comparison With State-of-the-Art

1) Quantitative Comparison: Table I and II provide a comprehensive evaluation of our proposed method and competing approaches on six datasets. The results demonstrate the superiority of our method over existing weakly-/unsupervised methods across all datasets. Our method achieves substantial performance gains when compared to the state-of-the-art weakly-supervised method SCWS. Specifically, we observe an average improvement of 3.53%, 2.71%, 2.96%, 2.25%, and 0.97% in terms of S_m, F_β^w, F, E, and M, respectively. These significant improvement verifies the effectiveness of our approach. Furthermore, our method demonstrates competitive performance even when compared to state-of-the-art fully-supervised SOD methods. Notably, on the HKU-IS dataset, our method outperforms all other approaches except for LDF. To visually illustrate the performance of SOD methods,

TABLE I
QUANTITATIVE COMPARISON OF SALIENT OBJECT DETECTION METHODS ON DUTS, DUT-OMRPN, AND ECSSD DATASETS

Method	DUTS					DUT-OMRON					ECSSD				
	$S_m \uparrow$	$F_\beta^w \uparrow$	$F \uparrow$	$E \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$F \uparrow$	$E \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$F \uparrow$	$E \uparrow$	$M \downarrow$
R3 [67]	0.836	0.713	0.798	0.882	0.066	0.818	0.679	0.759	0.862	0.071	0.903	0.860	0.9917	0.943	0.056
DGRL [6]	0.842	0.774	0.805	0.898	0.050	0.806	0.709	0.739	0.853	0.062	0.903	0.891	0.914	0.946	0.041
TSPOA [68]	0.860	0.767	0.828	0.907	0.049	0.818	0.697	0.750	0.858	0.061	0.907	0.876	0.919	0.942	0.046
BAS [5]	0.866	0.803	0.838	0.903	0.048	0.836	0.751	0.779	0.871	0.056	0.916	0.904	0.931	0.951	0.037
SCRN [69]	0.885	0.803	0.864	0.925	0.040	0.837	0.720	0.772	0.875	0.056	**0.927**	0.900	0.38	0.956	0.037
AFNET [70]	0.867	0.785	0.838	0.910	0.046	0.826	0.717	0.759	0.861	0.057	0.913	0.886	0.924	0.947	0.042
GCPA [38]	0.891	0.821	0.869	0.929	0.038	**0.839**	0.734	0.775	0.869	0.056	**0.927**	0.903	0.936	0.955	0.035
GateNet [71]	0.885	0.809	0.869	0.928	0.040	0.838	0.729	0.781	**0.876**	0.055	0.920	0.894	0.933	0.952	0.040
MINet [37]	0.884	0.825	0.865	0.927	0.037	0.833	0.738	0769	0.869	0.056	0.925	0.911	**0.938**	**0.957**	**0.033**
LDF [4]	**0.892**	**0.845**	**0.877**	**0.930**	**0.034**	**0.839**	**0.752**	**0.782**	0.869	**0.052**	0.924	**0.915**	**0.938**	0.954	0.034
ASMO [9]	0.697	0.488	0.647	0.789	0.116	0.752	0.559	0.684	0.816	0.101	0.802	0.702	0.811	0.864	0.110
MNL [14]	-	-	-	-	-	0.788	0.633	0.723	0.848	0.076	0.870	0.823	0.883	0.920	0.069
MWS [10]	0.759	0.586	0.720	0.833	0.091	0.756	0.527	0.677	0.816	0.109	0.828	0.716	0.859	0.909	0.096
USPS [13]	0.788	0.700	0.746	0.853	0.068	0.793	0.698	0.737	0.849	0.063	0.862	0.844	0.881	0.912	0.062
EDNL [12]	0.820	0.701	0.785	0.877	0.065	0.783	0.633	0.718	0.842	0.076	0.871	0.827	0.882	0.920	0.059
SCRIB [11]	0.803	0.709	0.755	0.873	0.062	0.785	0.669	0.708	0.841	0.068	0.865	0.835	0.871	0.920	0.059
SCWS [17]	0.840	0.792	0.823	0.907	0.049	0.812	0.731	0.756	0.868	0.060	0.882	0.875	0.902	0.932	0.049
UDASOD [63]	0.846	0.783	0.820	0.896	0.050	0.808	0.711	0.744	0.847	0.059	0.899	0.885	0.917	0.933	0.043
OURS	**0.881**	**0.824**	**0.864**	**0.927**	**0.039**	**0.836**	**0.741**	**0.780**	**0.873**	**0.054**	**0.920**	**0.902**	**0.932**	**0.953**	**0.038**

TABLE II
QUANTITATIVE COMPARISON OF SALIENT OBJECT DETECTION METHODS ON HKU-IS, PASCAL-S, AND SOD DATASETS

Method	HKU-IS					PASCAL-S					SOD				
	$S_m \uparrow$	$F_\beta^w \uparrow$	$F \uparrow$	$E \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$F \uparrow$	$E \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$F \uparrow$	$E \uparrow$	$M \downarrow$
R3 [67]	0.892	0.833	0.900	0.943	0.048	0.809	0.730	0.808	0.853	0.104	0.738	0.700	0.807	0.819	0.136
DGRL [6]	0.894	0.875	0.900	0.949	0.036	0.836	0.800	0.837	0.891	0.072	0.774	0.738	0.800	0.832	0.103
TSPOA [68]	0.902	0.862	0.909	0.95	0.038	0.841	0.779	0.834	0.886	0.078	0.775	0.718	0.808	0.844	0.115
BAS [5]	0.909	0.889	0.919	0.951	0.032	0.836	0.795	0.837	0.884	0.077	0.772	0.728	0.803	0.832	0.112
SCRN [69]	0.916	0.876	0.921	0.956	0.034	0.868	0.811	0.858	0.909	0.064	0.792	0.732	0.825	0.865	0.105
AFNET [70]	0.905	0.869	0.91	0.949	0.036	0.849	0.802	0.848	0.895	0.071	-	-	-	-	-
GCPA [38]	**0.920**	0.889	0.927	0.958	0.031	0.866	0.815	0.855	**0.908**	0.062	**0.808**	0.761	0.826	0.866	**0.088**
GateNet [71]	0.915	0.880	0.920	0.955	0.033	0.858	0.801	0.852	0.903	0.069	0.801	0.753	**0.837**	**0.870**	0.098
MINet [37]	0.919	0.897	0.926	**0.960**	**0.029**	0.856	0.814	0.852	0.903	0.064	0.805	**0.768**	0.836	**0.870**	0.092
LDF [4]	0.919	**0.904**	**0.929**	0.958	0.028	**0.862**	**0.826**	**0.859**	0.907	**0.061**	0.800	0.765	0.834	0.866	0.093
ASMO [9]	0.804	0.701	0.822	0.884	0.086	0.714	0.578	0.709	0.79	0.152	0.669	0.551	0.702	0.753	0.185
MNL [14]	0.884	0.834	0.892	0.942	0.047	0.824	0.743	0.824	0.879	0.093	0.736	0.663	0.793	0.836	0.144
MWS [10]	0.818	0.685	0.835	0.908	0.084	0.767	0.614	0.758	0.832	0.134	0.702	0.571	0.768	0.821	0.166
USPS [13]	0.876	0.857	0.886	0.936	0.041	0.773	0.715	0.768	0.83	0.108	0.713	0.659	0.748	0.768	0.143
EDNL [12]	0.884	0.838	0.892	0.942	0.046	0.819	0.739	0.819	0.875	0.095	0.739	0.669	0.791	0.832	0.142
SCRIB [11]	0.865	0.831	0.865	0.931	0.047	0.796	0.736	0.783	0.861	0.094	0.727	0.668	0.767	0.835	0.129
SCWS [17]	0.882	0.872	0.898	0.943	0.038	0.819	0.788	0.826	0.880	0.078	0.757	0.723	0.793	0.814	0.108
UDASOD [63]	0.897	0.879	0.912	0.950	0.035	0.822	0.773	0.813	0.880	0.080	0.788	0.750	0.812	0.846	**0.095**
OURS	**0.920**	**0.899**	**0.929**	**0.961**	**0.029**	**0.853**	**0.805**	**0.844**	**0.897**	**0.068**	**0.794**	**0.757**	**0.827**	**0.868**	0.097

Figure 7 showcases the precision-recall curves for the six datasets. It is evident that our method consistently surpasses other weakly-/unsupervised methods and achieves comparable performance to fully-supervised methods. These results validate the efficacy of our proposed method and its potential to address the challenges of weakly-/unsupervised SOD task.

2) Qualitative Comparison: Fig. 8 showcases the saliency predictions on 12 diverse images capturing a range of scenarios such as images with tiny objects (first row), complex object shapes (second to fourth rows), multiple instances (fifth and sixth rows), and highly camouflaged objects (seventh row). Our method consistently generates comprehensive and accurate saliency maps with well-defined boundaries, exhibiting remarkable performance across all images. Comparing the visualization results of our method with other weakly-/unsupervised methods, it is evident that our approach significantly outperforms them. Furthermore, our method's saliency predictions are on par with those achieved by fully supervised methods, further highlighting its effectiveness.

C. Ablation Study

To verify the effectiveness of our proposed uncertainty-aware active domain adaptive (UADA) algorithm, we conduct the ablation study on the main components and discuss the sensitivity to key hyper-parameters.

1) Effectiveness of Components in Pseudo-Label Learning: We conduct a series of experiments to evaluate the impact of critical components in our pseudo-label learning algorithm. The results are summarized in Table III. The 'Source Only' approach, which trains the saliency detector solely on synthetic source domain data, achieves performance comparable to some weakly-/unsupervised methods (as shown in Table I and II), such as ASMO [9] and MWS [10]. This finding demonstrates the feasibility of learning from synthetic data.

To assess the effectiveness of our pseudo-label learning algorithm, we train the saliency detector using unlabeled target domain images through vanilla pseudo-label learning (Vanilla PL) without employing superpixel-level sample filtering (SSF), pixel-wise pseudo-label reweighting (PPR), or active superpixel-level labeling (ASL). It is evident that

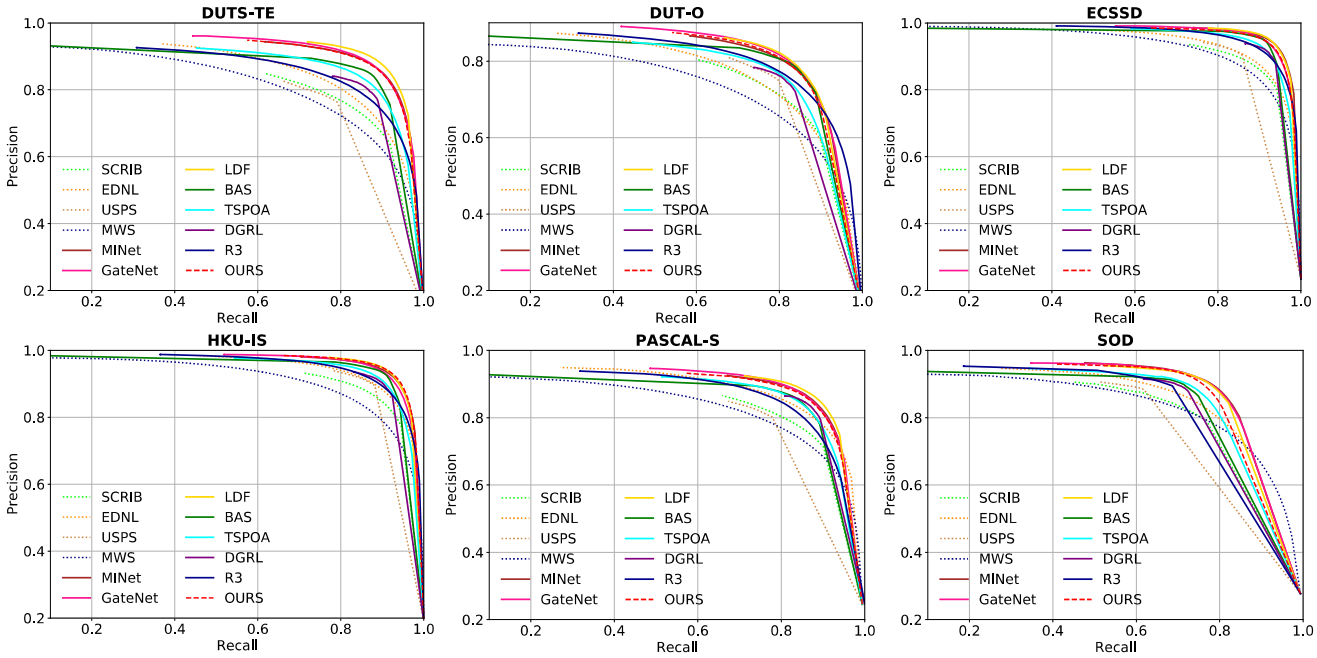


Fig. 7. The PR curves of different SOD methods on six datasets. The fully supervised methods are represented by solid lines, weakly-supervised/unsupervised methods are represented by dotted lines, and our method is represented by a red dashed line.

TABLE III

ABLATION STUDY ON KEY COMPONENTS OF PSEUDO-LABEL LEARNING

Method	DUTS		DUTS-OMRON		ECSSD		HKU-IS		PASCAL-S		SOD	
	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$
Source Only	0.670	0.068	0.572	0.083	0.812	0.068	0.803	0.056	0.705	0.106	0.640	0.134
Vanilla PL	0.711	0.065	0.627	0.079	0.818	0.069	0.819	0.053	0.710	0.106	0.669	0.132
UADA w/o SSF	0.803	0.043	0.722	0.060	0.899	0.038	0.894	0.030	0.799	0.070	0.748	0.099
UADA w/o PPR	0.809	0.042	0.725	0.058	0.894	0.040	0.889	0.031	0.798	0.071	0.751	0.097
UADA(Ours)	**0.824**	**0.039**	**0.741**	**0.054**	**0.902**	**0.038**	**0.899**	**0.029**	**0.805**	**0.068**	**0.757**	**0.097**

TABLE IV

ABLATION STUDY ON COMPONENTS OF ACTIVE LABELING

Method	DUTS		DUTS-OMRON		ECSSD		HKU-IS		PASCAL-S		SOD	
	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$
UADA-G	0.728	0.053	0.645	0.068	0.845	0.058	0.836	0.046	0.761	0.081	0.677	0.117
UADA-R	0.785	0.046	0.709	0.062	0.872	0.046	0.873	0.036	0.782	0.075	0.740	0.101
UADA(Ours)	0.824	0.039	0.741	0.054	0.902	0.038	0.899	0.029	0.805	0.068	0.757	0.097
UADA-F	**0.842**	**0.036**	**0.751**	**0.053**	**0.914**	**0.034**	**0.910**	**0.026**	**0.822**	**0.062**	**0.763**	**0.095**

training with vanilla pseudo-labels helps bridge the domain gap between synthetic and real-world data, leading to a significant improvement in saliency detection performance.

We also perform ablation experiments by removing SSF and PPR from our Uncertainty-aware Active Domain Adaptation (UADA) approach, resulting in 'UADA w/o SSF' and 'UADA w/o PPR', respectively. Specifically, for 'UADA w/o SSF', we conducted pseudo-label learning with all superpixels except those selected for active labeling. For 'UADA w/o PPR', we assign a weight of 1 to all pixels when calculating the training loss using pseudo-labels. Compared to the complete UADA approach, both 'UADA w/o SSF' and 'UADA w/o PPR' exhibit a slight drop in performance across the six datasets. This indicates that SSF effectively eliminates incorrect pseudo-labels, while PPR mitigates the negative impact of pixel-wise noisy pseudo-labels.

2) Effectiveness of Components in Active Superpixel-Level Labeling: To evaluate the effectiveness of components in our active superpixel-level labeling algorithm, we implemented four variants and analyzed their performance, as shown in Table IV.

The first variant, called UADA-G, decomposes target domain images into superpixels using uniform grids. However, this approach exhibits a significant degradation in performance compared to UADA, which utilizes the SEEDS algorithm [66] for superpixel generation. Specifically, the F_β^w metric decreases by over 0.1 on the DUTS and DUTS-OMRON datasets. This performance decline can be attributed to the fact that uniform grids cannot guarantee adherence to object boundaries, as depicted in Fig. 9. As a result, a large number of incorrectly labeled pixels are introduced, leading to a substantial deterioration in model performance. In contrast, superpixels generated by SEEDS effectively retain boundary information and generally consist of semantically coherent pixels, ensuring higher labeling accuracy during the manual annotation process.

The second variant, named UADA-R, randomly selects superpixels for manual label assignment. The results

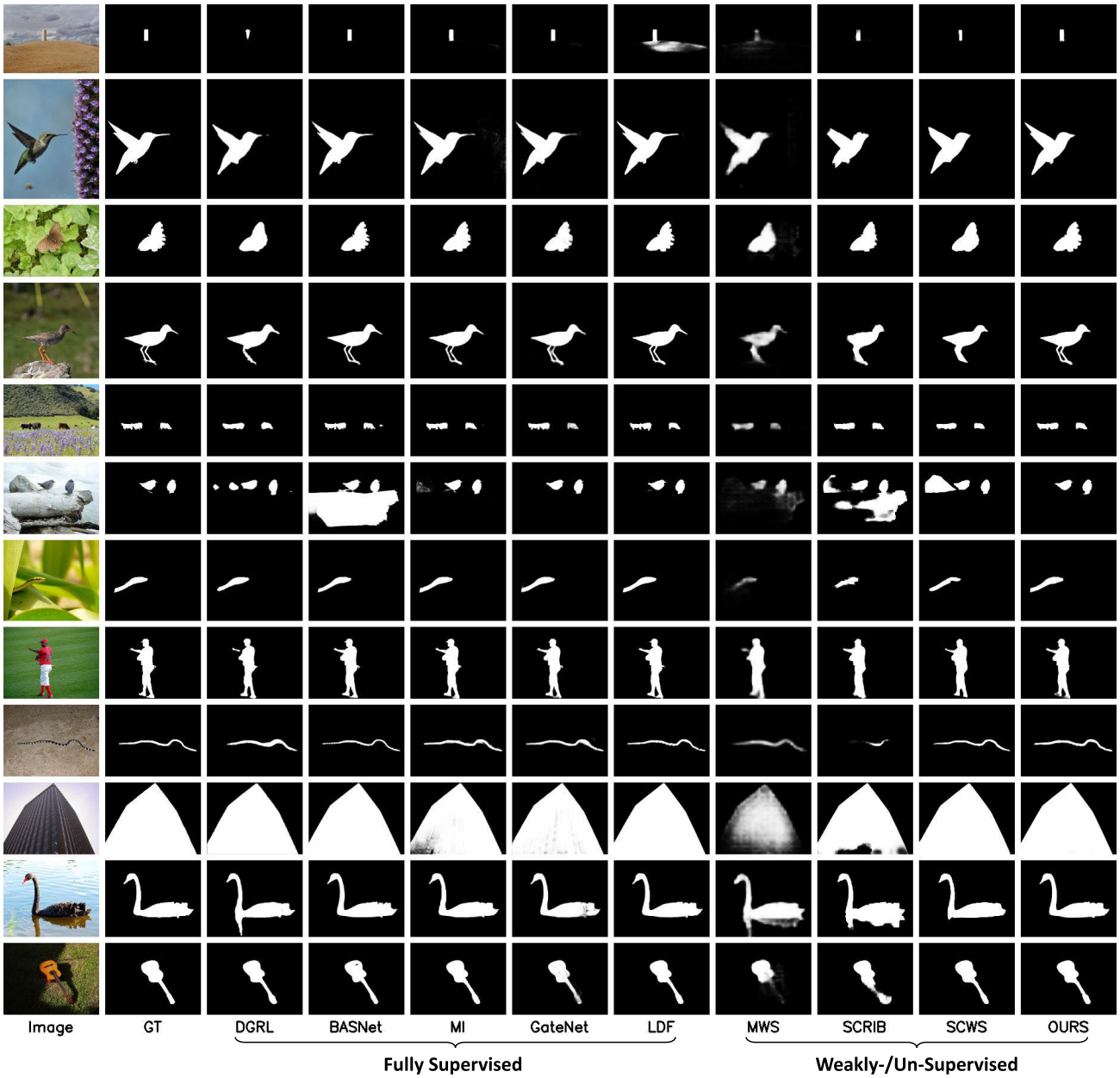


Fig. 8. Qualitative comparisons of our method against other fully supervised, weakly-/unsupervised. The performance of our method is better than that of weakly-/unsupervised methods and is comparable to that of fully supervised methods.

demonstrate that UADA-R achieves worse performance in terms of F_β^w and M on all six datasets. This finding confirms the effectiveness of selecting superpixels based on their uncertainty values, as opposed to a random selection strategy.

Additionally, we introduced another variant called UADA-F, where the selected superpixels are assigned pixel-wise precise labels instead of dominant labels. UADA-F can be considered an upper bound of our method, providing an estimation of the best achievable performance.

3) Sensitivity to Hyper-Parameter μ in Pixel-Wise Pseudo-Label Reweighing: In the PPR module, μ is a hyper-parameter for controlling the attenuation degree of the weights. Experiments are conducted to demonstrate the robustness of our method against the variance of μ. As shown in Table V, the performance drops evidently in the case where $\mu = 0$ (equivalent to removing PPR module, i.e. 'UADA w/o PPR' in Table III). Meanwhile, there only exists slight fluctuation when varying μ in $\{10, 20, 30\}$, which verifies the robustness of the proposed method to μ.

4) Sensitivity to the Number of Superpixel: Table VI presents an analysis of the sensitivity to the number of superpixels in a single image. The extreme case denoted by 'HW' in Table VI represents a scenario where each superpixel consists of only one pixel, which is equivalent to the UADA-F variant discussed in Table IV. The results indicate that as the number of superpixels increases, the performance of the method improves consistently. This improvement can be attributed to the finer division of superpixels, which reduces the error rate associated with assigning the dominant label. However, it is important to consider the tradeoff between performance and

TABLE V
SENSITIVITY TO HYPER-PARAMETER μ IN PIXEL-WISE PSEUDO-LABEL REWEIGHING

μ	DUTS		DUTS-OMRON		ECSSD		HKU-IS		PASCAL-S		SOD	
	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$
0	0.809	0.042	0.725	0.058	0.894	0.040	0.889	0.031	0.798	0.071	0.751	0.097
10	0.815	0.041	0.734	0.058	0.899	0.038	0.896	0.030	0.800	0.069	**0.774**	**0.090**
20	**0.824**	**0.039**	0.741	0.054	**0.902**	**0.038**	0.899	**0.029**	0.805	0.068	0.757	0.097
30	0.820	0.040	**0.745**	**0.053**	**0.902**	0.038	**0.901**	**0.029**	**0.811**	**0.066**	0.754	0.100

TABLE VI
SENSITIVITY TO THE NUMBER OF SUPERPIXELS IN A SINGLE IMAGE. HW REPRESENTS THE NUMBER OF PIXELS IN TRAINING IMAGES

Number of Superpixels	DUTS		DUTS-OMRON		ECSSD		HKU-IS		PASCAL-S		SOD		Error Rate
	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	
25	0.779	0.046	0.708	0.060	0.879	0.046	0.877	0.035	0.779	0.079	0.716	0.109	5.7%
50	0.806	0.042	0.719	0.057	0.887	0.043	0.882	0.034	0.789	0.072	0.732	0.104	4.9%
100	0.824	0.039	0.741	0.054	0.902	0.038	0.899	0.029	0.805	0.068	0.757	0.097	3.8%
HW	**0.842**	**0.036**	**0.751**	**0.053**	**0.914**	**0.034**	**0.910**	**0.026**	**0.822**	**0.062**	**0.763**	**0.095**	**0.0%**

TABLE VII
SENSITIVITY TO THE PORTION OF SUPERPIXELS FOR MANUAL LABEL ASSIGNMENT

Portion of Superpixels	DUTS		DUTS-OMRON		ECSSD		HKU-IS		PASCAL-S		SOD	
	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$
9%	0.814	0.042	0.732	0.056	0.899	0.040	0.891	0.033	0.802	0.069	0.732	0.106
15%(Ours)	0.824	0.039	0.741	**0.054**	0.902	**0.038**	0.899	0.029	0.805	**0.068**	0.757	**0.097**
21%	0.825	**0.039**	**0.744**	**0.054**	0.901	0.038	0.899	0.029	0.803	0.069	0.735	0.101
27%	**0.826**	**0.039**	0.743	0.055	**0.903**	0.038	**0.902**	**0.029**	**0.807**	0.068	**0.758**	0.097

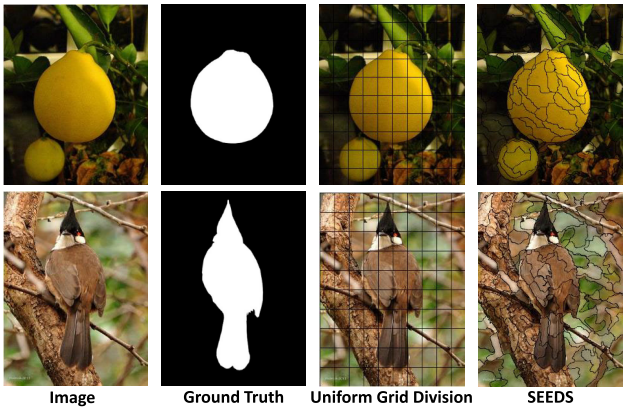


Fig. 9. Visualization of superpixel generated by uniform grid division and SEEDS [66]. Superpixels generated by SEEDS can better adhere to the object boundaries which cannot be ensured by uniform grid division.

annotation cost. In our experiments, we find that setting the number of superpixels to 100 strikes a good balance between achieving satisfactory performance and managing annotation costs.

5) Sensitivity to the Portion of Superpixels for Manual Label Assignment: In Table VII, we investigate the sensitivity of the model with regards to the proportion of superpixels selected for manual label assignment. Four experiments are conducted using proportions of 9%, 15%, 21%, and 27%, respectively, with increments of 3%, 5%, 7%, and 9% per round after the warm-up phase. Notably, a rapid improvement in performance is observed as the proportion increases from 9% to 15%. However, as the proportion continues to increase, the performance tends to reach a state of saturation.

6) Effectiveness of Iterative Training: We adopt an iterative training scheme consisting of four rounds. In Fig. 10, we illustrate the variation of F_β^w and M during the training process. As the training rounds progress, we consistently

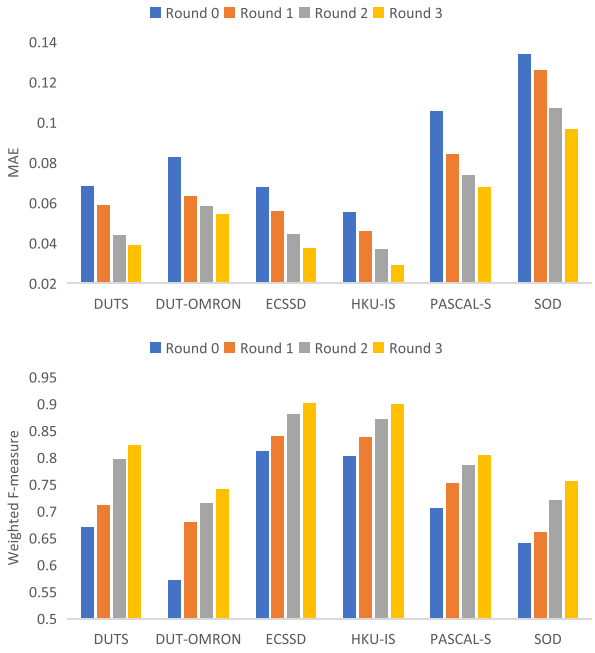


Fig. 10. Performance after each training round.

observe an increase in F_β^w and a decrease in M. This provides evidence of the effectiveness of our iterative training paradigm. Furthermore, Fig. 11 presents two examples where incorrectly predicted pixels are gradually reduced and the prediction accuracy is improved as the training rounds increase.

7) Choice of Data Augmentation Strategies: Table VIII showcases the performance achieved by employing diverse data augmentation strategies in the calculation of pixel-wise uncertainty values. The label 'None' denotes the vanilla pseudo-label learning approach without the application of superpixel-level sample filtering or pixel-wise pseudo-label reweighting. It is noteworthy that the adoption of a single

TABLE VIII
SENSITIVITY TO DIFFERENT KINDS OF DATA AUGMENTATION IN CONSISTENCY-BASED UNCERTAINTY ESTIMATION

Augmentation	DUTS		DUTS-OMRON		ECSSD		HKU-IS		PASCAL-S		SOD	
	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$M \downarrow$
None	0.711	0.065	0.627	0.079	0.818	0.069	0.819	0.053	0.710	0.106	0.669	0.132
Flip	0.806	0.043	0.725	0.058	0.889	0.043	0.885	0.033	0.794	0.072	0.733	0.105
Rescaling	0.812	0.041	0.730	0.059	0.898	0.038	0.898	0.029	0.803	0.069	0.764	0.093
FIA	0.805	0.043	0.721	0.058	0.888	0.042	0.886	0.033	0.794	0.072	0.724	0.107
All(Ours)	**0.824**	**0.039**	**0.741**	**0.054**	**0.902**	**0.038**	**0.899**	**0.029**	**0.805**	**0.068**	**0.757**	**0.097**

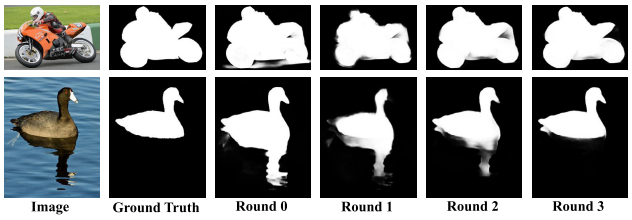


Fig. 11. Two examples of saliency prediction after each training round.

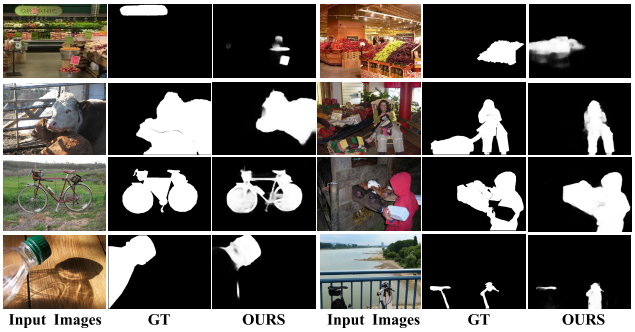


Fig. 12. Typical failure cases in which our method misses salient regions or regards non-salient regions as salient regions.

data augmentation technique, such as random flip, rescaling, or FIA, results in a significant enhancement in performance. This indicates the efficacy of individual augmentation strategies in improving uncertainty estimation. Furthermore, by integrating all three augmentation techniques, an additional performance boost is observed. This fusion of augmentations contributes to a more robust estimation of uncertainty, thereby yielding improved results.

VI. DISCUSSION

A. Failure Cases

In Fig. 12, we illustrate a series of failure cases encountered by our saliency detection algorithm. The first two rows highlight instances of false positives and negatives. Specifically, in the first row, we observe instances where non-salient objects are mistakenly identified as salient, and in the second row, only part of genuinely salient objects are detected. The primary reason behind these inaccuracies can be traced back to the synthetic nature of our training data, which typically features high contrast between foreground and background elements. This characteristic leads to a dataset that, while effective in simpler scenarios, struggles to replicate the complexity found in real-world environments. Such complexity often manifests as intricate background details or low contrast between the object and its surroundings, that are conditions under which our model's performance may diminish. To mitigate this, future iterations of our work will aim to incorporate more diverse and complex foreground and background images into the synthetic dataset, enhancing the model's ability to handle such challenging conditions.

The third and fourth rows demonstrate the algorithm's difficulties in accurately segmenting salient objects with intricate structures or high levels of transparency. These shortcomings are partly due to the simplicity of the foreground objects included in our synthetic dataset. Future work will explore the inclusion of more complex and varied foreground materials to better simulate the diversity of real-world scenarios. Additionally, the task of precisely segmenting objects with complex structures poses a significant challenge, necessitating the use of base models with advanced segmentation capabilities. Enhancing our dataset to include a wider range of object complexities and employing more sophisticated models for segmentation are crucial steps toward overcoming these limitations and improving the overall performance of our saliency detection algorithm.

B. Limitations of Our Approach

One limitation of our approach stems from the synthetic dataset construction process. Although we aim to replicate the complexity and diversity of real-world data by compositing foreground objects onto different background images, there is an inherent risk of creating scenarios that are unlikely or even impossible in the real world. Such combinations can introduce noisy training samples, which can potentially degrade the performance of the saliency detector when applied to real-world data, as the model may have learned to recognize patterns or scenarios that do not exist outside the synthetic environment. To address this limitation, future work could explore more sophisticated methods for generating synthetic datasets. Techniques such as adversarial training might be employed to ensure that synthetic images are indistinguishable from real images, thereby minimizing the domain gap.

The second limitation arises from the quality of the superpixels generated by our algorithm. Superpixels that contain both foreground and background elements can introduce noise. This issue is particularly problematic during the active labeling process, where superpixels of poor quality are likely to be selected for manual annotation. Even with human intervention, the presence of noise is inevitable and affects the overall performance of the model. Improving the quality of the superpixels is a straightforward approach to mitigating this issue. This improvement could be achieved by fine-tuning the parameters like the granularity of the superpixel algorithm or adopting more advanced superpixel segmentation methods that offer a better balance between boundary adherence and

computational efficiency. Moreover, during the manual annotation phase, human annotators can easily identify and correct severe inaccuracies introduced by poorly segmented superpixels. Implementing a semi-automated annotation process in which annotators are assisted by suggestions from the model could reduce the noise level and enhance the quality of training data.

C. Broader Implications

The implications of our research on SOD utilizing synthetic datasets and advanced domain adaptation techniques extend far beyond the immediate improvements in annotation efficiency and model performance. Here is a detailed exploration of the broader implications of our work.

One of the most immediate implications of our work is the substantial reduction in the need for manual annotation. This not only lowers the cost and time investment required for training SOD models but also makes it feasible to scale model training to datasets of virtually unlimited size. As a result, models can be trained on a wider variety of data, potentially leading to improvements in their generalizability and robustness.

Our approach addresses one of the critical challenges in utilizing synthetic data: the domain gap between synthetic and real-world images. By developing and implementing uncertainty-aware pixel-level and active superpixel-level labeling techniques, we provide a pathway for effectively leveraging synthetic data in training models that perform well on real-world tasks. This has significant implications for the entire field of machine learning, suggesting that with the right techniques, the gap between synthetic and real data can be bridged more efficiently than previously thought.

Finally, our work lays the groundwork for future research in several areas. It challenges researchers to further improve synthetic data generation techniques, to develop more sophisticated domain adaptation algorithms, and to explore new applications of these methods in other fields of artificial intelligence. Moreover, it highlights the importance of interdisciplinary collaboration, combining insights from computer vision, machine learning, and domain-specific knowledge to tackle complex problems.

VII. Conclusion

In this research, we propose a novel approach for salient object detection that addresses the challenge of limited data annotations by leveraging both synthetic and real-world datasets. Our approach introduces an uncertainty-aware pseudo-labeling algorithm, which includes superpixel-level sample filtering and pixel-wise pseudo-label reweighting, to effectively exploit the target domain data. By selectively excluding unreliable pseudo-labels and assigning different importance values to reliable ones, we enhance the robustness of the pseudo-labels. Additionally, we propose an active superpixel-level labeling algorithm that replaces unreliable pseudo-labels with manually annotated labels, thereby improving the overall quality of the labels without imposing significant annotation costs. We conduct extensive experiments on multiple benchmark datasets, including DUTS, DUT-O, ECSSD, HKU-IS, PASCL-S, and SOD, to evaluate the performance of our method. The results demonstrate the superiority of our approach compared to existing weakly-/unsupervised methods, as we achieve significant improvements in salient object detection accuracy. Moreover, our method remains competitive even when compared to state-of-the-art approaches. In conclusion, our proposed method makes a valuable contribution to the field of salient object detection by addressing the challenge of limited data annotations and achieving remarkable performance improvements.

References

[1] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 FPS," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1404–1412.

[2] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2814–2821.

[3] X. Li, H. Lu, L. Zhang, X. Ruan, and M. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2976–2983.

[4] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13025–13034.

[5] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.

[6] T. Wang et al., "Detect globally, refine locally: A novel approach to saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3127–3135.

[7] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2386–2395.

[8] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 186–202.

[9] G. Li, Y. Xie, and L. Lin, "Weakly supervised salient object detection using image labels," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–8.

[10] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu, "Multi-source weak supervision for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6067–6076.

[11] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12546–12555.

[12] J. Zhang, J. Xie, and N. Barnes, "Learning noise-aware encoder–decoder from noisy labels by alternating back-propagation for saliency detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 349–366.

[13] T. Nguyen et al., "DeepUSPS: Deep robust unsupervised saliency prediction via self-supervision," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[14] J. Zhang, T. Zhang, Y. Daf, M. Harandi, and R. Hartley, "Deep unsupervised saliency detection: A multiple noisy labeling perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9029–9038.

[15] J. Zhang, Y. Dai, T. Zhang, M. Harandi, N. Barnes, and R. Hartley, "Learning saliency from single noisy labelling: A robust model fitting perspective," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2866–2873, Aug. 2021.

[16] D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4068–4076.

[17] S. Yu, B. Zhang, J. Xiao, and E. G. Lim, "Structure-consistent weakly supervised salient object detection with local saliency coherence," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 3234–3242.

[18] J. Choi, M. Jeong, T. Kim, and C. Kim, "Pseudo-labeling curriculum for unsupervised domain adaptation," 2019, *arXiv:1908.00262*.

[19] L. Kong, B. Hu, X. Liu, J. Lu, J. You, and X. Liu, "Constraining pseudo-label in self-training unsupervised domain adaptation with energy-based model," *Int. J. Intell. Syst.*, vol. 37, no. 10, pp. 8092–8112, Oct. 2022.

[20] L. Wang et al., "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 136–145.

[21] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1155–1162.

[22] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[23] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5455–5463.

[24] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 280–287.

[25] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 49–56.

[26] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3239–3259, Jun. 2021.

[27] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2994–3002.

[28] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, and J. Han, "A self-paced multiple-instance learning framework for co-saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 594–602.

[29] W. Ji et al., "DMRA: Depth-induced multi-scale recurrent attention network for RGB-D saliency detection," *IEEE Trans. Image Process.*, vol. 31, pp. 2321–2336, 2022.

[30] C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, and J. Li, "Pyramid grafting network for one-stage high resolution saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11707–11716.

[31] Z. Wu, G. Allibert, F. Meriaudeau, C. Ma, and C. Demonceaux, "HiDAnet: RGB-D salient object detection via hierarchical depth awareness," *IEEE Trans. Image Process.*, vol. 32, pp. 2160–2173, 2023.

[32] X. Tian, J. Zhang, M. Xiang, and Y. Dai, "Modeling the distributional uncertainty for salient object detection models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19660–19670.

[33] Y. Wang, R. Wang, X. Fan, T. Wang, and X. He, "Pixels, regions, and objects: Multiple enhancement for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10031–10040.

[34] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.

[35] G. Li and Y. Yu, "Contrast-oriented deep neural networks for salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6038–6051, Dec. 2018.

[36] J. Wei, S. Wang, and Q. Huang, "F^3Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 12321–12328.

[37] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9413–9422.

[38] Z. Chen, Q. Xu, and R. Cong, "Global context-aware progressive aggregation network for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 10599–10606.

[39] R. Cong et al., "Densely nested top-down flows for salient object detection," *Sci. China Inf. Sci.*, vol. 65, no. 8, 2022, Art. no. 182103.

[40] A. K. Bhunia et al., "Sketch2Saliency: Learning to detect salient objects from human drawings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2733–2743.

[41] Y. Wang, W. Zhang, L. Wang, T. Liu, and H. Lu, "Multi-source uncertainty mining for deep unsupervised saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11727–11736.

[42] H. Zhou, B. Qiao, L. Yang, J. Lai, and X. Xie, "Texture-guided saliency distilling for unsupervised salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7257–7267.

[43] O. Sener, H. O. Song, A. Saxena, and S. Savarese, "Learning transferrable representations for unsupervised domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[44] J. Cao, H. Tang, H.-S. Fang, X. Shen, Y.-W. Tai, and C. Lu, "Cross-domain adaptation for animal pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9497–9506.

[45] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3339–3348.

[46] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y. F. Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2011–2020.

[47] J. Hoffman et al., "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.

[48] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2507–2516.

[49] I. Nejjar, Q. Wang, and O. Fink, "DARE-GRAM : Unsupervised domain adaptation regression by aligning inverse Gram matrices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11744–11754.

[50] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[51] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 289–305.

[52] Z. Zheng and Y. Yang, "Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1106–1120, Apr. 2021.

[53] W. H. Beluch, T. Genewein, A. Nurnberger, and J. M. Kohler, "The power of ensembles for active learning in image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9368–9377.

[54] H. Ranganathan, H. Venkateswara, S. Chakraborty, and S. Panchanathan, "Deep active learning for image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3934–3938.

[55] Y. Siddiqui, J. Valentin, and M. Nießner, "ViewAL: Active learning with viewpoint entropy for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9433–9443.

[56] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proc. Int. Conf. Mach. Learning(ICML)*, 2017, pp. 1183–1192.

[57] A. Parvaneh, E. Abbasnejad, D. Teney, R. Haffari, A. Van Den Hengel, and J. Q. Shi, "Active learning by feature mixing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12227–12236.

[58] A. Freytag, E. Rodner, and J. Denzler, "Selecting influential examples: Active learning with expected model output changes," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 562–577.

[59] P. Munjal, N. Hayat, M. Hayat, J. Sourati, and S. Khan, "Towards robust and reproducible active learning using neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 223–232.

[60] M. Xie et al., "Learning distinctive margin toward active domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 7993–8002.

[61] D. Huang, J. Li, W. Chen, J. Huang, Z. Chai, and G. Li, "Divide and adapt: Active domain adaptation via customized learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7651–7660.

[62] B. Xie, L. Yuan, S. Li, C. H. Liu, and X. Cheng, "Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8058–8068.

[63] P. Yan, Z. Wu, M. Liu, K. Zeng, L. Lin, and G. Li, "Unsupervised domain adaptive salient object detection through uncertainty-aware pseudo-label learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022, pp. 3000–3008.

[64] G. Bonaccorso, *Mastering Machine Learning Algorithms: Expert Techniques to Implement Popular Machine Learning Algorithms and Fine-Tune Your Models*. Birmingham, U.K.: Packt Publishing Ltd, 2018.
[65] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4085–4095.
[66] M. V. D. Bergh, X. Boix, G. Roig, B. D. Capitani, and L. V. Gool, "SEEDS: Superpixels extracted via energy-driven sampling," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2012, pp. 13–26.
[67] Z. Deng et al., "R^3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Palo Alto, CA, USA, Jul. 2018, pp. 684–690.
[68] Y. Liu, Q. Zhang, D. Zhang, and J. Han, "Employing deep part-object relationships for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1232–1241.
[69] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7264–7273.
[70] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1623–1632.
[71] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 35–51.
[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[73] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," *Proc. SPIE*, vol. 11006, pp. 369–386, May 2019.
[74] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.
[75] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.
[76] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.
[77] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," 2018, *arXiv:1805.10421*.
[78] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.

Guanbin Li (Member, IEEE) received the Ph.D. degree in computer science from The University of Hong Kong, Hong Kong, in 2016. He is currently a Full Professor with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. He has authored or coauthored more than 120 papers in top-tier academic journals and conferences. His research interests include computer vision, image processing, and deep learning. He was a recipient of the ICCV 2019 Best Paper Nomination Award. He serves as an Area Chair for CVPR 2024. He has been a reviewer for numerous academic journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *International Journal of Computer Vision*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CYBERNETICS, Conference on Computer Vision and Pattern Recognition, International Conference on Computer Vision, European Conference on Computer Vision, and Conference on Neural Information Processing Systems.

Zhuohua Chen received the B.Eng. degree from Sun Yat-sen University in 2022, where he is currently pursuing the master's degree with the School of Computer Science. His research interests include computer vision and deep learning.

Mingzhi Mao received the B.S. degree in computer science, the M.S. degree in computational mathematics, and the Ph.D. degree in computer software mathematics from Sun Yat-sen University, Guangzhou, China, in 1988, 1998, and 2008, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, Sun Yat-sen University. His research interests include intelligence algorithm, software engineering, and management information systems.

Liang Lin (Fellow, IEEE) received the Ph.D. degree in computer science and technology from Beijing Institute of Technology, Beijing, China, in 2008. He was the Executive Director and a Distinguished Scientist with SenseTime Group from 2016 to 2018, leading the research and development teams for cutting-edge technology transferring. He is currently a Full Professor of computer science with Sun Yat-sen University, Guangzhou, China. He has authored or coauthored more than 200 papers in leading academic journals and conferences, and his papers have been cited more than 20 000 times. He is a fellow of IET. He was a recipient of numerous awards and honors, including the Wu Wen-Jun Artificial Intelligence Award, the First Prize of China Society of Image and Graphics, the ICCV Best Paper Nomination in 2019, the Annual Best Paper Award by *Pattern Recognition* (Elsevier) in 2018, the Best Paper Dimond Award from IEEE ICME 2017, and the Google Faculty Award in 2012. His supervised Ph.D. students have received the ACM China Doctoral Dissertation Award, the CCF Best Doctoral Dissertation, and the CAAI Best Doctoral Dissertation. He was an Area Chair of numerous conferences, such as CVPR, ICCV, SIGKDD, and AAAI. He is an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS.

Chaowei Fang (Member, IEEE) received the Ph.D. degree from The University of Hong Kong, Hong Kong, in 2019. He is currently an Associate Professor with the School of Artificial Intelligence, Xidian University, Xi'an, China. He has contributed as an author or a coauthor to over 40 publications featured in prestigious journals and conferences. His research interests include low-level image processing, medical image analysis, and machine learning. He served as a Senior Program Committee Member for ECAI 2024.