Pixel Is All You Need: Adversarial Trajectory-Ensemble Active Learning for Salient Object Detection

Zhenyu Wu[1], Lin Wang[2], Wei Wang[3], Qing Xia[4], Chenglizhao Chen[5*], Aimin Hao[1,6], Shuo Li[7]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[2]School of Transportation Science and Engineering, Beihang University
[3]Harbin Institute of Technology
[4]SenseTime Group Limited
[5]China University of Petroleum
[6]Peng Cheng Laboratory
[7]Case Western Reserve University

Abstract

Although weakly-supervised techniques can reduce the labeling effort, it is unclear whether a saliency model trained with weakly-supervised data (e.g., point annotation) can achieve the equivalent performance of its fully-supervised version. This paper attempts to answer this unexplored question by proving a hypothesis: there is a point-labeled dataset where saliency models trained on it can achieve equivalent performance when trained on the densely annotated dataset. To prove this conjecture, we proposed a novel yet effective adversarial trajectory-ensemble active learning (ATAL). Our contributions are three-fold: 1) Our proposed adversarial attack triggering uncertainty can conquer the overconfidence of existing active learning methods and accurately locate these uncertain pixels. 2) Our proposed trajectory-ensemble uncertainty estimation method maintains the advantages of the ensemble networks while significantly reducing the computational cost. 3) Our proposed relationship-aware diversity sampling algorithm can conquer oversampling while boosting performance. Experimental results show that our ATAL can find such a point-labeled dataset, where a saliency model trained on it obtained 97%-99% performance of its fully-supervised version with only 10 annotated points per image.

Introduction

Recently, deep learning based SOD methods (Ji et al. 2021; Fan et al. 2021; Sun et al. 2021; Liu et al. 2021a; Tang et al. 2021; Gu et al. 2021; Zhou et al. 2021) have achieved great success with the help of well-annotated large-scale datasets. Unfortunately, it is a prohibitive cost to annotate a large amount of pixel-wise data.

Weakly-supervised salient object detection (WSOD) methods have been proposed to alleviate the dependency on pixel-wise data. Existing WSOD methods (Zeng et al. 2019; Li, Xie, and Lin 2018; Wang et al. 2017) utilize weak annotation, such as image-level label and scribble annotation, which is easier to collect than a densely annotated label. More recently, (Gao et al. 2022) proposed a point-supervised SOD framework with the help of edge informa-
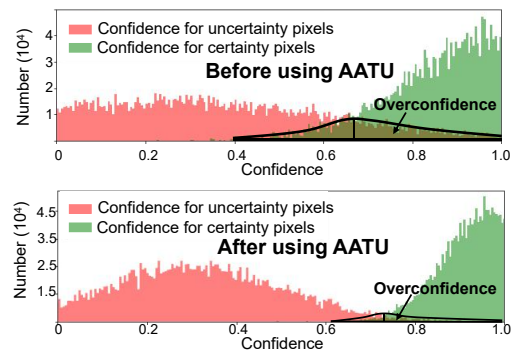

Figure 1: Our AATU can conquer the overconfidence issue of existing active learning methods and accurately identify these uncertain pixels.

tion. Despite the substantial progress, it is unclear whether a saliency model trained with weakly-supervised data (e.g., point-annotation) can achieve the equivalent performance of its fully-supervised version (i.e., pixel-wise annotation).

In this paper, we answer this unexplored question by proving a hypothesis: *there is a point labeled dataset where saliency models trained on it could achieve equivalent performance when trained on the densely annotated dataset.* This conjecture, if it is true, has rather promising practical significance–it suggests that the pixel-wise annotation is in fact unnecessary, freeing humans from the heavy burden of labeling pixel-wise data. While finding such a point-labeled dataset is non-trivial, there are two main challenges: **1)** In standard active learning (AL), uncertain pixels are defined as the low confidence pixels to the current model that has been trained over the present labeled set. However, our experiments revealed that existing AL methods are prone to produce "overconfident" predictions for uncertain pixels (see Fig. 1), resulting in inaccurate estimation. **2)** Despite deep ensemble networks (DEN) (Lakshminarayanan, Pritzel, and Blundell 2017; Franchi et al. 2020; Zaidi et al. 2021) is the most effective methodology for uncertainty estimation, it introduces prohibitive training cost, which is mainly incurred by training N networks repeatedly (see Fig. 2a).

$$o = \frac{1}{N} \sum_{i=1}^{N} \alpha f_i(x)$$

N networks, training N times

Input x

f_i f_{i+1} f_{i+2} **Spatial-ensemble**

(a) DEN

One network, training only once

Input x

Iteration t f_{θ_t} $f_{\theta_{t+1}}$ $f_{\theta_{t+2}}$

$$o = \frac{1}{T} \sum_{t=1}^{T} \alpha f_{\theta_t}(x)$$
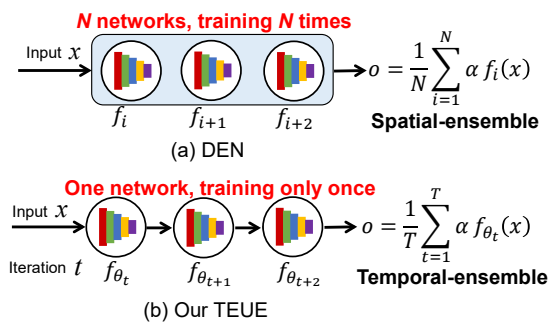
Temporal-ensemble

(b) Our TEUE

Figure 2: Unlike traditional DEN, which requires training N networks, our TEUE requires one network and training only once, compressing the computational overhead to $1/N$.

To address these challenges, we proposed a novel yet effective Adversarial Trajectory-ensemble Active Learning (ATAL), which can accurately identify these uncertain pixels and run cheaply as the vanilla training baseline. **First**, instead of focusing on developing a new regularizer to mitigate the overconfidence of existing softmax AL methods, we argue that continued research progress in uncertainty estimation requires new insights. In this paper, we take drastically different views, and propose a surprisingly effective technique, Adversarial Attack Triggering Uncertainty (AATU), to explicitly identify these uncertain pixels by injecting the adversarial perturbation into the input image. In contrast to previous methods, AATU allows us to comprehensively evaluate the uncertainty of each pixel and conquer the overconfidence of modern AL methods (see the 2nd row of Fig. 1). **Second**, by rethinking the standard deep ensemble networks and their variants, we proposed a temporal-ensemble model called Trajectory-Ensemble Uncertainty Estimation (TEUE) by aggregating network weights of the history model on the optimization path during the process of training. Unlike the previous spatial-ensemble techniques (see Fig. 2a), which require N networks and repeat training N times, our TEUE only requires one network and training only once (see Fig. 2b). Besides, our proposed Relationship-aware Diversity Sampling (RDS) can alleviate the oversampling issues and further boost the performance by considering the relationship among these sampling pixels.

Our ATAL has several desirable properties: **1) High Performance.** Our ATAL can achieve averagely $97\% - 99\%$ performance of its fully-supervised version with only ten annotated pixels per image (see Table. 2). In addition, our proposed method also outperforms existing WSOD models by a large margin. **2) Low Computational Cost.** Compared to traditional N-ensemble networks, our ATAL can compress the computational overhead to $1/N$. **3) Generalization.** Our ATAL can be easily integrated into existing SOD models as a plug-and-play module. Besides, a point labeled dataset selected by one SOD model can be used to train other SOD models well. In summary, our main contributions are:

- For the first time, we demonstrated that a saliency model trained with a point-labeled dataset could achieve equivalent performance trained on the pixel-wise dataset.

- Our AATU provides a new insight for uncertainty estimation and identifies these uncertain pixels by injecting the adversarial attacks into the input image, which can alleviate the overconfidence of modern AL methods and accurately locate these uncertain pixels.

- Our TEUE reduces the computational overheads stemming from repeatedly network training while maintaining the advantages brought by the ensemble networks.

- Our proposed RDS algorithm can alleviate the oversampling issues and further boost the performance by considering the relationship among these sampling pixels.

Related Works

Weakly-supervised SOD. With recent advances in weakly-supervised learning, a few existing works exploit the potential of training saliency model on image-level (Zeng et al. 2019; Li, Xie, and Lin 2018; Wang et al. 2017; Piao et al. 2021) and scribble-level (Yu et al. 2021; Zhang et al. 2020; Zhang, Xie, and Barnes 2020) to relax the dependency of manually annotated pixel-level masks. These approaches follow the same technical route, i.e., producing the initial saliency maps with image-level labels and then refining them via iterative training. Recently, point annotation was proposed in (Gao et al. 2022), but it requires extra data information (e.g., edge) to recover integral object structure. **Differences.** Distinct from all these works, our work attempt to demonstrate that a saliency model trained on a point-labeled dataset can achieve equivalent performance when trained on the pixel-wise dataset.

Deep Active Learning. Active learning is a set selection problem that aims to determine the most informative subset given a labeling budget. It has been successfully used as a method for reducing labeling costs. Recently, deep active learning has been attracting increasingly more attention in many areas, including classification (Beluch et al. 2018; Gal, Islam, and Ghahramani 2017; Krishnamurthy et al. 2017; Gao et al. 2020), object detection (Yoo and Kweon 2019; Yuan et al. 2021; Aghdam et al. 2019; Choi et al. 2021), semantic segmentation (Kim et al. 2021; Siddiqui, Valentin, and Nießner 2020; Dai et al. 2020; Yang et al. 2017). **Differences.** In contrast to previous methods, our ATAL can conquer the overconfidence of modern AL methods and accurately locate these uncertain pixels, significantly reducing the computational cost.

Methodology

The proposed ATAL framework (see Fig. 3) consists of four modules: **1)** a surprisingly effective adversarial attack for better identify uncertain pixels, **2)** a lightening trajectory-ensemble networks for uncertainty estimation, **3)** a novel relationship-aware diversity sampling strategy for pixel sampling, and **4)** an effective region label acquisition method.

Adversarial Attack Triggering Uncertainty

The proposed AATU explicitly identifies these uncertain pixels via injecting the adversarial perturbation into the input image, providing a new insight for uncertainty estimation. Concretely, given a saliency network f and an input x, the
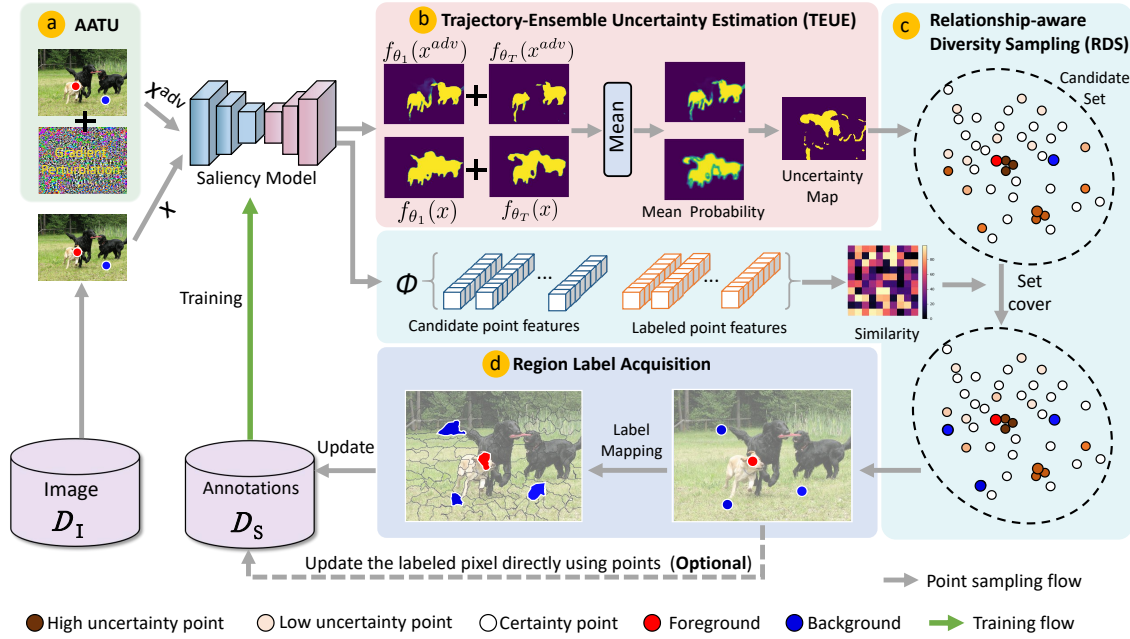
Figure 3: Overview of the ATAL. First, we train a saliency network on the initial labeled clean data x and employ the PGD to generate the adversarial image x^{adv}. Second, we use the proposed TEUE to calculate the uncertainty score for clean data x and adversarial data x^{adv}, and then obtain the candidate set. Third, we select a batch of uncertain pixels from the candidate set by using our RDS. Finally, we propagate the annotated pixel label to its corresponding superpixel block and update the current labeled set D_S for the next round of training. This process is repeated until reaching the largest budget or desired performance.

output $o = f(x)$, where $x \in \mathbb{R}^{H \times W \times 3}$ and $o \in \mathbb{R}^{H \times W}$. For a clean sample x, pixel $x(i, j)$ is called clean pixel; for the adversarial sample x^{adv}, which is obtained by injecting adversarial perturbation into x, pixel $x^{adv}(i, j)$ is called adversarial pixel. In practice, we employ the standard first-order adversarial attack, i.e., Projected Gradient Descent (PGD) (Kurakin, Goodfellow, and Bengio 2016), which iteratively updates the adversarial example under the l_∞-norm threat model by:

$$x^{adv_{t+1}} = \text{Clip}(x^{adv_t} + \alpha \cdot \text{sign}(\nabla_{x^{adv_t}}(\mathcal{L}(f(x^{adv_t}), y)))) \quad (1)$$

where x^{adv_t} is the adversarial sample after the t-th attack step, $\text{Clip}(\cdot)$ forces it ouput to reside in the range of $[x - \epsilon, x + \epsilon]$, ϵ is the perturbation range, $\text{sign}(\cdot)$ is the sign function, and α is the step size.

We further propose to divide these adversarial pixels into different categories according to their anti-perturbation ability. As shown in Fig. 4a, adversarial pixels can be classified into three categories: Safety Region (SR), Perturbation Insensitive Region (PIR), and Perturbation Sensitive Region (PSR). **1) SR:** clean pixels and their corresponding adversarial pixels are classified into the same category in the output space. These safety pixels are robust to adversarial attack and commonly stay far away from the decision boundary. **2) PIR:** clean pixels and their corresponding adversarial pixels are classified into different categories. These clean pixels stay far away from the decision boundary, while adversarial pixels stay near the decision boundary. If we enhance

the model's robustness, the PIR can be turned into SR. **3) PSR:** clean pixels and their corresponding adversarial pixels are classified into different categories, and these clean pixels stay near the decision boundary.

By injecting the PGD attack, we can easily identify these certain pixels that $x^{clean}(i, j)$ and $x^{adv}(i, j)$ have the same prediction. However, capturing these most uncertain pixels is non-trivial because they are entangled with these perturbation-insensitive pixels. We employ the DEN to distinguish these most uncertain pixels from PIR by enhancing the robustness of PIR and calculating the uncertainty score for each pixel. Despite its outstanding performance (see the 2nd row of Table 1), the DEN introduces prohibitive training costs, which will be addressed in the following section.

Advantages: We surprisingly find that these perturbation-sensitive pixels triggered by adversarial attacks correspond to these most uncertain pixels. Moreover, our AATU can alleviate the overconfidence of existing AL methods (see Fig. 1) and accurately locate these uncertain pixels.

Trajectory-Ensemble Uncertainty Estimation

Our TEUE mainly focuses on reducing the computational overheads stemming from repeating network training while attempting to retain the benefits brought by deep ensemble networks (DEN). We begin by revisiting the procedure of the traditional ensemble method for uncertainty estimation. **First**, a set of $\{f_1, f_2, ..., f_N\}$ train on the current labeled dataset $\{x_i, y_i\}_{i=1}^{M}$ with different initializations, forming N well-trained model with different predictions. **Second**, these
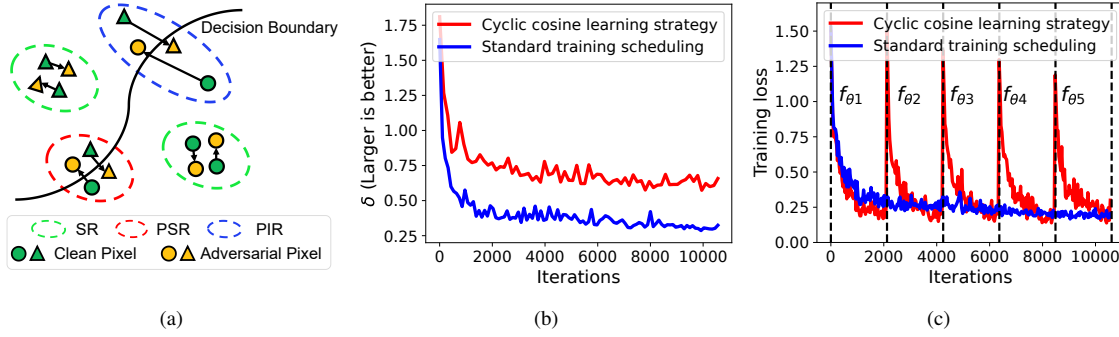
Figure 4: (a) The illustration of the SR, PIR, and PSR division. (b) Our CCLS can alleviate the homogenization phenomenon of trajectory-ensemble networks. (c) Our CCLS can force the model to converge to its local minimum just a few epochs, which ensure that our TEUE can work like deep ensemble networks.

	Training	DUT-OMRON		DUTS-TE	
	cost	$\max F_\beta \uparrow$	S-m$\uparrow$	$\max F_\beta \uparrow$	S-m$\uparrow$
Vanilla	M	.7254	.7367	.7852	.7768
DEN	MN	**.7716**	**.7930**	**.8383**	**.8325**
w/o CCLS	M	.7336	.7483	.7964	.7875
w/ CCLS	M	**.7708**	**.7921**	**.8374**	**.8326**

Table 1: Our DEN can achieve similar performance to DEN while reducing the computational cost to $1/N$.

clean images x and their adversarial counterparts x^{adv} are then fed into these networks to obtain the probability maps $p(x)$ and $p(x^{adv})$:

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x), \quad p(x^{adv}) = \frac{1}{N}\sum_{i=1}^{N} f_i(x^{adv}) \quad (2)$$

Finally, the obtained $p(x)$ and $p(x^{adv})$ can be used to calculate the uncertainty score map $u(x)$:

$$u(x) = \text{Diff}\left(\text{BvSB}\left(u\left(x\right)\right), \text{BvSB}\left(u\left(x^{adv}\right)\right)\right) \quad (3)$$

where $\text{BvSB}(\cdot)$ is the best-versus-second best margin (Joshi, Porikli, and Papanikolopoulos 2009), $\text{Diff}(x_1, x_2)$ denotes these pixels where x_1 are inconsistency with x_2.

Let's consider the training cost of one epoch. We denote the cost of a single forward and backward pass for one image as 1, and the dataset size as M. Then, the cost of vanilla training for one epoch is:

$$\text{Cost(vanilla)} = M \quad (4)$$

Similarly, the training cost of DEN is:

$$\text{Cost(DEN)} = M \times N \quad (5)$$

Obviously, DEN increases the training cost by a factor of N compared to the vanilla training baseline.

To reduce the computational cost, we first give a naive attempt to simplify DEN by ensembling the pre-recorded weights of model optimization trajectories, called trajectory-ensemble. Thus, we reformulate the Eq. 2 as:

$$p(x) = \frac{1}{T}\sum_{t=1}^{T} f_{\theta_t}(x), \quad p(x^{adv}) = \frac{1}{T}\sum_{t=1}^{T} f_{\theta_t}(x^{adv}) \quad (6)$$

where f_{θ_t} is the t-th snapshot of model optimization trajectory. As shown in the 3rd row of Table 1, this strategy severely degrades the original DEN's performance (i.e., 73.4% vs.77.2% on the DUT-OMRON dataset).

After carefully diagnosing, we observed that, in the trajectory-ensemble setting, the model f_{θ_t} obtained at relatively late stages loses the superiority of the ensemble due to the "homogenization phenomenon" as shown in Fig. 4b (blue line). We define the homogenization of a model by calculating the difference of these output of trajectory models over a period of iteration τ:

$$\delta = \frac{1}{|\tau|}\sum_{t}^{t+\tau} |f_{\theta_{t+1}}(x) - f_{\theta_t}(x)| \quad (7)$$

We introduce a novel cyclic cosine learning strategy (CCLS) to address the homogenization of trajectory-ensemble networks. Concretely, we start training the model with a very large learning rate and then lower it at a very fast pace, forcing the model to converge to its local minimum after just a few epochs. The optimization is then restarting at a larger learning rate, which perturbs the model and moves it away from the local minimum. We repeat this procedure several times to obtain multiple networks. Formally, our cyclic cosine learning rate is defined as:

$$\eta = \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\left(\pi\frac{\text{mod}\left(i-1, [I/L]\right)}{[I/L]}\right)\right) \quad (8)$$

where η_{max} and η_{min} are ranges for the learning rate, i is the iteration number, I is the total number of training iterations, and L is the number of training cyclic. Fig. 4c shows the difference between the standard training scheduling and our CCLS. After training L cycles, we obtained L model snapshots $\{f_1, ..., f_L\}$, each of which will be used in the final ensemble. As shown in Fig. 4b, our CCLS can relieve the homogenization phenomenon.

Advantages: As shown in the 4th row of Table 1, our TEUE not only largely reduces the training cost to $1/N$ but also maintains the advantages brought by the ensemble networks.
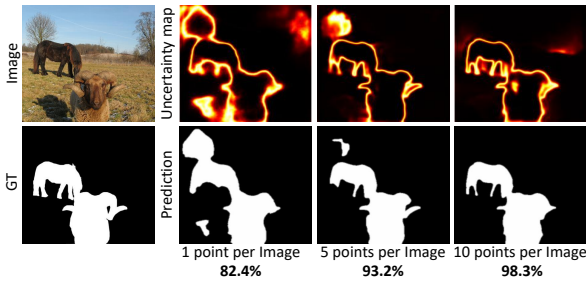
Figure 5: Our ATAL achieves 98.3% performance of its fully-supervised version with 10 labeled pixels per image.

Relationship-aware Diversity Sampling

Our proposed relationship-aware diversity sampling (RDS) aims to conquer the oversampling issue (selected points tend to be distributed in the most highlight region, see the uncertainty maps of Fig. 5) by considering the relationship among these labeled pixels. Given an uncertainty map $u(x)$, unlike previous methods simply selecting the top-k uncertainty points, we first sample the top $K\%$ pixels forming the candidate set D_C. We then select k points from the candidate set D_C, where the selected points should satisfy the following properties: **1)** covering the candidate set D_C; **2)** dissimilar with the current labeled set D_S.

To this end, we first need to find a set of points V that cover the candidate set D_C as possible, which can be viewed as a variant of the set cover problem (Feige 1998). Let $\sigma(p_u, p_v)$ denote the Euclidean distance between a pair of points $p_u, p_v \in D_C$. Our objective is to find a set $V \subseteq D_C$ to minimize the maximum distance of any point of D_C to its closest cluster center, which is an NP-hard problem.

$$\min_{V \subseteq D_C} \sum_{p_u \in D_C} \sum_{p_v, \in V} \sigma(p_u, p_v). \tag{9}$$

Instead of iterating through each pixel of the candidate set D_C, we introduce a greedy approximation algorithm (see Algorithm 1) to find pixels $p_u \in D_C$ that maximizes $\sum_{p_v \in V} \sigma(p_u, p_v)$.

After obtaining the cover set V, we then calculate the similarity of D_S for candidate point $p_x \in V$ as below:

$$\phi(p_x, D_S) = \frac{1}{|D_S|} \sum_{p_j \in D_S} \mathrm{Sim}(p_x, p_j), \tag{10}$$

where $\mathrm{Sim}(\cdot)$ is commonly used *cosine* similarity, $|D_S|$ denotes the number of labelled points. Finally, we rank all pixels in V using the $\phi(p_x, D_S)$ and select the top-k dissimilar points. In this way, the selected points not only have a high uncertainty score but also consider the relationship between the candidate points and the labeled points.

Advantages: Our proposed RDS can address the oversampling issues by considering the relationship of these labeled pixels, achieving high performance with a few labeled pixels (see Fig. 5).

Algorithm 1: Our greedy approximation algorithm.

Input: a candidate set D_C, an integer $m \in \mathbb{N}$
Output: $V \subseteq D_C$, $|V| = m$, with minimum max distance to points of D_C as eq. 9, $V \leftarrow p_u$, for $p_u \in D_C$ arbitrary
1: **while** $|V| < m$ **do**
2:     let $p_u \in D_C \setminus V$ be the element maximizing $\sum_{p_v \in V} \sigma(p_u, p_v)$
3:     update the V, $V \leftarrow p_u$
4: **end while**
5: **return** V

Region Label Acquisition

To exploit the structure and locality of the image, we argue that selecting regions instead of single pixels can achieve better performance. Unlike previous region-based AL methods, which select the closest rectangle, we propose to map the annotated pixel label to its superpixel block. In this work, we employ an off-the-shelf SEEDS algorithm (Van den Bergh et al. 2012) due to its good performance in ensuring class coherency within each superpixel. Please note that our proposed framework is universal, and any other superpixels algorithms can also be employed.

Experiments

Experimental Setup

Implementation Detail. Network. In this work, we do not focus on network architecture design, in our experiments, we adopt F3Net (Wei, Wang, and Huang 2020), MINet (Pang et al. 2020) and PFSN (Ma, Xia, and Li 2021) as our saliency model to validate the effectiveness of our ATAL algorithm. The detailed parameters setting of these saliency networks can be found in their paper. **Point Annotation.** For the point annotation collection, instead of using a real annotator, the point annotation process can be seamlessly combined with the DUTS-TR groundtruth. We can automatically classify the selected points into the correct category according to the positions in the groundtruth without any annotation efforts. **Loss Function.** In our ATAL algorithm, we treat each labeled pixel as an independent training sample. Thus, we can train these SOD models just like these fully-supervised models by minimizing the cross-entropy loss between prediction and annotated pixel in the same coordinate. We initially randomly select 2 points to label, and the max budget is set to 20. According to experimental results, when K=3, the saliency model obtains the best results. The number of training cyclic L is set to 5.

Datasets and Evaluation Metrics. We follow previous approaches to use the DUTS-TR (Wang et al. 2017) as the training set, and the only difference is that our model is supervised with point annotation. We evaluate our model on 5 datasets, including DUTS-OMRON (Yang et al. 2013), DUTS-TE (Wang et al. 2017), ECSSD (Yan et al. 2013), HKU-IS (Zhao et al. 2015) and PASCAL-S (Li et al. 2014). We adopt the Precision-Recall, F-measure, Mean Absolute Error (MAE), S-measure, and E-measure to evaluate our method.

		Fully-Supervised Models								Semi/Weakly-Supervised Models							
	Metric	DGRL	PAGR	BAS	CPD	SAMN	MINet	F3Net	PFSN	MWS	ENDS	WS^{3}A	MFNet	FCS	MINet$^*_{10}$	F3Net$^*_{10}$	PFSN$^*_{10}$
DUTS-O	maxF↑	.7742	.7709	.8053	.7966	.8026	.8098	.8133	**.8233**	.7176	.7581	.7532	.6874	.7170	.7895	.7908	**.8030**
	S-m↑	.8059	.7751	.8362	.8248	.8299	.8329	.8385	**.8425**	.7559	.7832	.7848	.7258	.7448	.8126	.8131	**.8252**
	MAE↓	.0618	.0709	.0565	.0560	.0652	.0555	**.0526**	.0545	.1086	.0759	.0684	.0982	.0656	.6130	.0604	**.0584**
	W-F↑	.7093	.6012	**.7520**	.7047	.6483	.7194	.7200	.7422	.4230	.5372	.6518	.5104	.6205	.7014	.7109	**.7133**
	E-m↑	.8425	.6036	**.8573**	.7874	.6882	.7975	.8031	.8134	.3364	.6161	.7854	.6005	.7257	.7761	**.8129**	.8027
DUTS-TE	maxF↑	.8287	.8545	.8591	.8654	.8360	.8835	.8905	**.8949**	.7686	.8173	.7889	.7632	.8296	.8632	.8624	**.8716**
	S-m↑	.8410	.8369	.8649	.8684	.8479	.8834	.8881	**.8916**	.7573	.8190	.8021	.7764	.8206	.8587	.8586	**.8741**
	MAE↓	.0500	.0562	.0480	.0438	.0582	.0375	**.0358**	.0359	.0920	.0657	.0628	.0798	.0459	.0468	.0454	**.0433**
	W-F↑	.7677	.6845	.7930	.7689	.6816	.7966	.8011	**.8198**	.5312	.6129	.6897	.5948	.7496	.7704	.7748	**.7996**
	E-m↑	.8856	.6130	**.8858**	.8377	.7124	.8518	.8640	.8737	.6103	.6954	.8214	.6836	.8616	.8324	.8548	**.8657**
ECSSD	maxF↑	.9224	.9268	.9424	.9392	.9279	.9445	.9453	**.9523**	.8778	.9002	.8880	.8727	.9108	.9236	.9290	**.9327**
	S-m↑	.9028	.8892	.9162	.9181	.9071	.9239	.9242	**.9298**	.8275	.8707	.8655	.8366	.8787	.8957	.8938	**.8993**
	MAE↓	.0407	.0609	.0370	.0371	.0501	.0334	.0333	**.0309**	.0963	.0676	.0590	.0841	.0471	.0478	.0461	**.0439**
	W-F↑	.8909	.8218	.9043	.8889	.8337	.8880	.9020	**.9142**	.6520	.7862	.8318	.7508	.8678	.8752	.8773	**.8807**
	E-m↑	.9373	.5582	**.9382**	.9020	.8010	.9012	.9150	.9199	.5554	.8232	.9033	.7987	.9212	.9124	.9233	**.9335**
HKU-IS	maxF↑	.9105	.9176	.9285	.9251	.9147	.9351	.9368	**.9428**	.8560	.9041	.8805	.8746	.8992	.9112	.9149	**.9245**
	S-m↑	.8945	.8873	.9090	.9055	.8983	.9160	.9173	**.9244**	.8182	.8838	.8649	.8523	.8718	.8905	.8931	**.8969**
	MAE↓	.0356	.0475	.0322	.0342	.0449	.0285	.0280	**.0259**	.0843	.0461	.0470	.0582	.0389	.0364	.0359	**.0346**
	W-F↑	.8752	.8049	8893	.8660	.8105	.9026	.8900	**.9029**	.6131	.7838	.8250	.7492	.8561	.8740	.8674	**.8761**
	E-m↑	**.9380**	.5065	.9356	.8879	.7857	.9059	.9120	.9192	.5075	.8164	.9051	.8143	.9295	.9376	**.9445**	.9398
PASCAL-S	maxF↑	.8808	.8691	.8757	.8841	.8894	.8894	.8948	**.8986**	.8140	.8706	.8374	.8230	.8742	.8762	.8896	**.8916**
	S-m↑	.8278	.7925	.8194	.8277	.8027	.8333	.8404	**.8431**	.7532	.8025	.7805	.7663	.8102	.8078	.8140	**.8234**
	MAE↓	.0823	.1149	.0924	.0890	.1130	.0828	.0799	**.0790**	.1509	.1144	.1106	.1310	**.0849**	.0923	.0915	.0879
	W-F↑	.7806	.7008	.7762	.7707	.7239	.7789	.7890	**.7981**	.6125	.7111	.7294	.6704	.7704	.7538	.7684	**.7788**
	E-m↑	.8380	.5921	.8340	.8273	.7653	.8347	.8430	**.8482**	.5456	.8028	.8308	.7533	.8560	.8401	.8591	**.8612**

Table 2: Our ATAL can achieve about 97% – 99% F-measure of its fully-supervised version with only 10 annotated pixels per image and outperform existing weakly-supervised methods. The "F3Net$^*_{10}$" denotes F3Net trained on point labeled dataset.

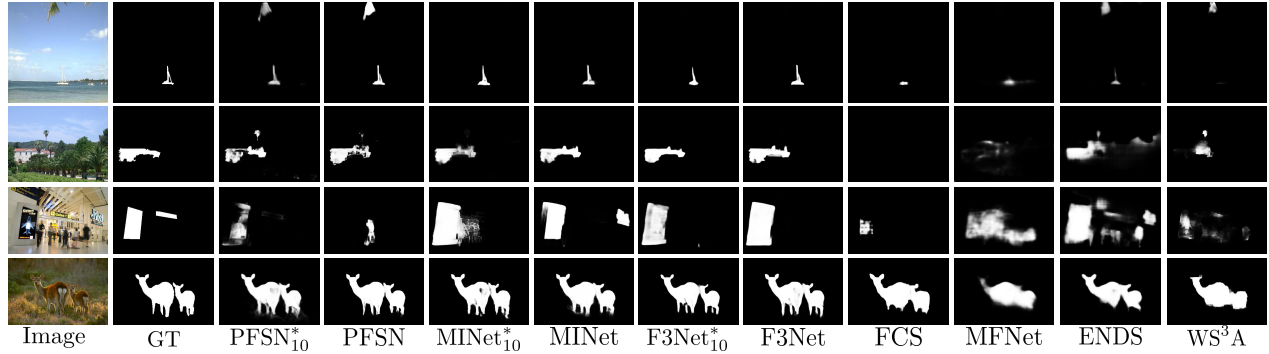| Image | GT | PFSN$^*_{10}$ | PFSN | MINet$^*_{10}$ | MINet | F3Net$^*_{10}$ | F3Net | FCS | MFNet | ENDS | WS^{3}A |

Figure 6: Our ATAL can identify an informative point-labeled dataset, where a saliency model trained on it can achieve the equivalent performance of its fully-supervised version.

Train \ Test	MINet	F3Net	PFSN
MINet	**97.4%**	96.4%	96.1%
F3Net	96.8%	**97.2%**	96.6%
PFSN	96.3%	96.5%	**97.5%**

Table 3: Cross-validation results demonstrate that our ATAL has good generability to different SOD models.

Comparisons with State-of-the-Art (SOTA)

We compare the our approach with 13 SOTA SOD models, including MWS (Zeng et al. 2019), EDNS (Zhang, Xie, and Barnes 2020), WS^{3}A (Zhang et al. 2020), MFNet (Piao et al. 2021), FCS (Zhang, Tian, and Han 2020), DGRL (Wang et al. 2018), PAGR (Zhang et al. 2018), BAS (Qin et al. 2019), CPD (Wu, Su, and Huang 2019), MINet (Pang et al. 2020), F3Net (Wei, Wang, and Huang 2020), PFSN (Ma, Xia, and Li 2021), and SAMN (Liu et al. 2021b).

Quantitative Comparison. To validate the effectiveness of our ATAL, we conduct experiments on MINet, F3Net, and PFSN. As shown in Table 2, our ATAL can achieve about 97% – 99% F-measure of its fully-supervised version with only ten annotated pixels per image and outperform existing weakly-supervised methods by a large margin. Besides, we conducted nine cross-validations on MINet, F3Net, and PFSN. As shown in Table 3, the point dataset selected by F3Net can be used to train MINet and PFSN with negligible performance degradations. Similar observations can also be found in Minet and PFSN models. These results demonstrate the effectiveness of our ATAL, and also indicate our ATAL has good generability to different SOD models.

Qualitative Comparison. As shown in Fig. 6, a saliency model trained on our ATAL selected point-labeled datasets
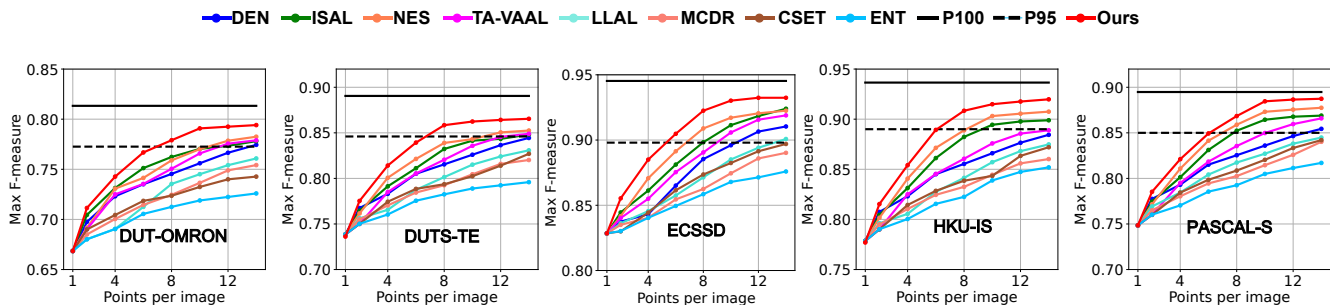
Figure 7: Our ATAL outperforms all other active learning methods. The horizontal solid line (p100) at the top represents model performance with the pixel-wise labeled training set. The dashed line (p95) represents 95% of that performance.

No.	DEN	AATU	TEUE	RDS	SP	DUT-OMRON S-m↑	DUTS-TE S-m↑
①	✔					.7682	.8126
②	✔	✔				.7930	.8325
③		✔	✔			.7921	.8326
④		✔	✔	✔		.8015	.8421
⑤		✔	✔	✔	✔	**.8131**	**.8586**

Table 4: Ablation study on each component demonstrates that all components are necessary for the proposed ATAL.

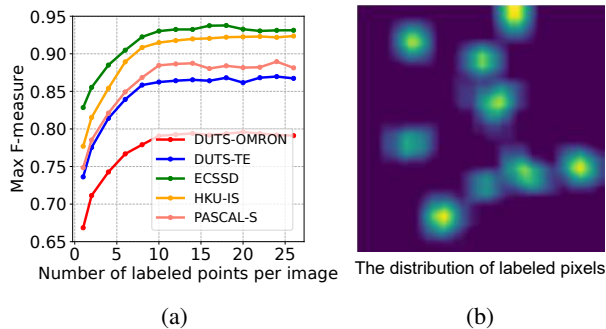(a)   (b)

Figure 8: (a) Ablation study on the number of annotated points per image; (b) Our RDS can conquer the oversampling issue, obtaining a set of diverse labeled points.

can achieve the equivalent performance of its fully-supervised version and outperforms existing weakly-supervised SOD in various challenging scenarios, small objects (the 1st row), occlusion scenes (the 2nd row), cluttered backgrounds (the 3rd row), and low contrast (the 4th row).

Comparisons Against Active Learning Methods. We compare our ATAL method with 8 commonly used AL methods, including **ISAL**(Liu et al. 2021c), **NES** (Zaidi et al. 2021), **TA-VAAL**) (Kim et al. 2021), **LLAL** (Yoo and Kweon 2019), **MCDR** (Mackowiak et al. 2018), **CSET** (Sener and Savarese 2018), **DEN** (Lakshminarayanan, Pritzel, and Blundell 2017), and **ENT** (Wang et al. 2016). As shown in Fig. 7, our ATAL achieves better performance than other AL methods with the same number of labeled data.

Ablation Study of Our Innovation

Number of Labeled Points Per Image. As shown in Fig. 8(a), with the increase of labeled points, the performance of F3Net improves rapidly before ten labeled points per image. After that, the performance improvements were only about 0.0044 in max-F from 10 to 20. In our experiments, we use the 10-point-labeled dataset, which is 2× faster than the 20-point-labeled dataset.

Effectiveness of AATU. We verify the effectiveness of AATU by combining our AATU with the model ①. As we can see, DEN+AATU can improve model ① performance by a considerable margin.

Effectiveness of TEUE. To verify the effectiveness of TEUE, we further replace the DEN with our TEUE. As we can see, model ③ can achieve similar performance to model ② while reducing the computational cost to $1/5$.

Effectiveness of RDS. To verify the effectiveness of RDS,

we replace the top-k sampling strategy in model ③ with our RDS (model ④). As we can see, our RDS can further improve the performance of model ③. In Fig. 8(b), we also show the distribution of labeled points on DUTS-TR, indicating the effect of our RDS.

Effectiveness of Superpixel (SP). Finally, we evaluate the effect using SP selection, forming our complete ATAL algorithm (model ⑤). Compared to model ④, superpixel selection achieves better performance than pixel selection. When using superpixel as a basic labeled unit, the total labeled pixels account for 2.2% of the DUTS-TR.

Conclusions

In this paper, we surprisingly find that there is a sparse point labeled dataset where saliency models trained on it can achieve equivalent performance when trained on the densely annotated dataset. As far as we know, this is the first work that gives empirical evidence to the existence of such a sparse point labeled dataset. Our results suggest that sparse point supervision has promising to replace the pixel-wise data in the feature. Besides, we also present an effective and efficient adversarial trajectory-ensemble active learning to identify such a sparse point labeled dataset.

Acknowledgments

References

Aghdam, H. H.; Gonzalez-Garcia, A.; Weijer, J. v. d.; and López, A. M. 2019. Active learning for deep detection neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 3672–3680.

Beluch, W. H.; Genewein, T.; Nürnberger, A.; and Köhler, J. M. 2018. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9368–9377.

Choi, J.; Elezi, I.; Lee, H.-J.; Farabet, C.; and Alvarez, J. M. 2021. Active Learning for Deep Object Detection via Probabilistic Modeling. In *Proceedings of the IEEE International Conference on Computer Vision*, 10264–10273.

Dai, C.; Wang, S.; Mo, Y.; Zhou, K.; Angelini, E.; Guo, Y.; and Bai, W. 2020. Suggestive Annotation of Brain Tumour Images with Gradient-Guided Sampling. In *International conference on medical image computing and computer-assisted intervention*, 156–165.

Fan, Q.; Fan, D.-P.; Fu, H.; Tang, C.-K.; Shao, L.; and Tai, Y.-W. 2021. Group Collaborative Learning for Co-Salient Object Detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 12288–12298.

Feige, U. 1998. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45(4): 634–652.

Franchi, G.; Bursuc, A.; Aldea, E.; Dubuisson, S.; and Bloch, I. 2020. TRADI: Tracking deep neural network weight distributions. In *European Conference on Computer Vision*, 105–121. Springer.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, 1183–1192.

Gao, M.; Zhang, Z.; Yu, G.; Arık, S. Ö.; Davis, L. S.; and Pfister, T. 2020. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, 510–526.

Gao, S.; Zhang, W.; Wang, Y.; Guo, Q.; Zhang, C.; He, Y.; and Zhang, W. 2022. Weakly-Supervised Salient Object Detection Using Point Supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Gu, Y.-C.; Gao, S.-H.; Cao, X.-S.; Du, P.; Lu, S.-P.; and Cheng, M.-M. 2021. iNAS: Integral NAS for Device-Aware Salient Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 4934–4944.

Ji, W.; Li, J.; Yu, S.; Zhang, M.; Piao, Y.; Yao, S.; Bi, Q.; Ma, K.; Zheng, Y.; Lu, H.; and Cheng, L. 2021. Calibrated RGB-D Salient Object Detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9471–9481.

Joshi, A. J.; Porikli, F.; and Papanikolopoulos, N. 2009. Multi-class active learning for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2372–2379.

Kim, K.; Park, D.; Kim, K. I.; and Chun, S. Y. 2021. Task-aware variational adversarial active learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8166–8175.

Krishnamurthy, A.; Agarwal, A.; Huang, T.-K.; Daumé III, H.; and Langford, J. 2017. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, 1915–1924.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30.

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*.

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, N.; Zhao, W.; Zhang, D.; Han, J.; and Shao, L. 2021a. Light Field Saliency Detection With Dual Local Graph Learning and Reciprocative Guidance. In *Proceedings of the IEEE International Conference on Computer Vision*, 4712–4721.

Liu, Y.; Zhang, X.-Y.; Bian, J.-W.; Zhang, L.; and Cheng, M.-M. 2021b. SAMNet: Stereoscopically Attentive Multi-Scale Network for Lightweight Salient Object Detection. *IEEE Transactions on Image Processing*, 30: 3804–3814.

Liu, Z.; Ding, H.; Zhong, H.; Li, W.; Dai, J.; and He, C. 2021c. Influence selection for active learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 9274–9283.

Ma, M.; Xia, C.; and Li, J. 2021. Pyramidal Feature Shrinking for Salient Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Mackowiak, R.; Lenz, P.; Ghori, O.; Diego, F.; Lange, O.; and Rother, C. 2018. Cereals-Cost-Effective REgion-based Active Learning for Semantic Segmentation. In *British Machine Vision Conference*.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020. Multi-Scale Interactive Network for Salient Object Detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9413–9422.

Piao, Y.; Wang, J.; Zhang, M.; and Lu, H. 2021. MFNet: Multi-filter Directive Network for Weakly Supervised Salient Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 4136–4145.

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 conference on computer vision and pattern recognition*, 7479–7489.

Sener, O.; and Savarese, S. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Siddiqui, Y.; Valentin, J.; and Nießner, M. 2020. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9433–9443.

Sun, P.; Zhang, W.; Wang, H.; Li, S.; and Li, X. 2021. Deep RGB-D Saliency Detection With Depth-Sensitive Attention and Automatic Multi-Modal Fusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1407–1417.

Tang, L.; Li, B.; Zhong, Y.; Ding, S.; and Song, M. 2021. Disentangled High Quality Salient Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 3580–3590.

Van den Bergh, M.; Boix, X.; Roig, G.; de Capitani, B.; and Van Gool, L. 2012. Seeds: Superpixels extracted via energy-driven sampling. In *European Conference on Computer Vision*, 13–26.

Wang, K.; Zhang, D.; Li, Y.; Zhang, R.; and Lin, L. 2016. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12): 2591–2600.

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^{3}Net: Fusion, Feedback and Focus for Salient Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12321–12328.

Wu, Z.; Su, L.; and Huang, Q. 2019. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3907–3916.

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, L.; Zhang, Y.; Chen, J.; Zhang, S.; and Chen, D. Z. 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, 399–407.

Yoo, D.; and Kweon, I. S. 2019. Learning loss for active learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 93–102.

Yu, S.; Zhang, B.; Xiao, J.; and Lim, E. G. 2021. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yuan, T.; Wan, F.; Fu, M.; Liu, J.; Xu, S.; Ji, X.; and Ye, Q. 2021. Multiple instance active learning for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5330–5339.

Zaidi, S.; Zela, A.; Elsken, T.; Holmes, C. C.; Hutter, F.; and Teh, Y. 2021. Neural ensemble search for uncertainty estimation and dataset shift. *Advances in Neural Information Processing Systems*, 34: 7898–7911.

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 conference on computer vision and pattern recognition*, 6074–6083.

Zhang, D.; Tian, H.; and Han, J. 2020. Few-cost salient object detection with adversarial-paced learning. In *Advances in Neural Information Processing Systems*.

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

Zhang, J.; Yu, X.; Li, A.; Song, P.; Liu, B.; and Dai, Y. 2020. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 12546–12555.

Zhang, X.; Wang, T.; Qi, J.; Lu, H.; and Wang, G. 2018. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 714–722.

Zhao, R.; Ouyang, W.; Li, H.; and Wang, X. 2015. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1265–1274.

Zhou, T.; Fu, H.; Chen, G.; Zhou, Y.; Fan, D.-P.; and Shao, L. 2021. Specificity-Preserving RGB-D Saliency Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 4681–4691.