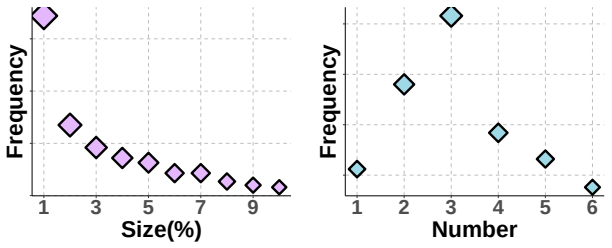
Size-Invariance Matters: Rethinking Metrics and Losses for Imbalanced Multi-object Salient Object Detection

Feiran Li [1 2] **Qianqian Xu** [3] **Shilong Bao** [1 2] **Zhiyong Yang** [4]
Runmin Cong [5 6 7] **Xiaochun Cao** [8] **Qingming Huang** [4 3 9]

Abstract

This paper explores the size-invariance of evaluation metrics in Salient Object Detection (SOD), especially when multiple targets of diverse sizes co-exist in the same image. We observe that current metrics are size-sensitive, where larger objects are focused, and smaller ones tend to be ignored. We argue that the evaluation should be size-invariant because bias based on size is unjustified without additional semantic information. In pursuit of this, we propose a generic approach that evaluates each salient object separately and then combines the results, effectively alleviating the imbalance. We further develop an optimization framework tailored to this goal, achieving considerable improvements in detecting objects of different sizes. Theoretically, we provide evidence supporting the validity of our new metrics and present the generalization analysis of SOD. Extensive experiments demonstrate the effectiveness of our method. The code is available at https://github.com/Ferry-Li/SI-SOD.



(a) Average size of objects (b) Average number of objects

Figure 1. Statistics on dataset MSOD. Fig. 1(a) illustrates the widely existing small salient objects, with *Size(%)* as the proportion of the size of an object over the whole image.Fig. 1(b) reveals that practical SOD scenarios usually involve multiple salient objects.

1. Introduction

Salient object detection (SOD), also known as salient object segmentation, aims at highlighting visually salient regions in images (Wang et al., 2022). To achieve this, a SOD model typically processes an RGB image to generate a binary mask, marking each pixel as either salient (1) or not (0). Recently, SOD has witnessed great progress in various applications (Mahadevan & Vasconcelos, 2009; Ren et al., 2014; Tang et al., 2017; Li et al., 2019; Zhang et al., 2020a; Jiang et al., 2023; Gui et al., 2024).

The progress of SOD primarily depends on two factors. One is the development of sophisticated models (say deep neural networks), which effectively disentangle diverse feature patterns for accurate SOD detection. Notable methods include (Wu et al., 2022; Luo et al., 2017; Wang et al., 2023; Ma et al., 2021; Zhang et al., 2021a). The other is the evaluation and selection of the best models for practical applications. Generally, a well-performed SOD model should simultaneously embrace a high True Positive Rate (TPR) and a low False Positive Rate (FPR) (Borji et al., 2019). To this end, various metrics (typically MAE and F-score) have been widely considered for evaluation and optimization (Chen et al., 2021; Sun et al., 2022).

In this paper, we argue that current evaluation metrics are

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China [2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China [3]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China [4]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China [5]Institute of Information Science, Beijing Jiaotong University, Beijing, China [6]School of Control Science and Engineering, Shandong University, Jinan, China [7]Key Laboratory of Machine Intelligence and System Control, Ministry of Education, Jinan, China [8]School of Cyber Science and Tech., Shenzhen Campus, Sun Yat-sen University [9]Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China. Correspondence to: Qianqian Xu <xuqianqian@ict.ac.cn>, Qingming Huang <qmhuang@ucas.ac.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

(a) Image (b) Label (c) EDN (d) EDN+Ours

Figure 2. (c) is the result of backbone EDN (Wu et al., 2022), and (d) is the prediction optimized by our approach. (c) detects fewer salient objects, yet enjoys lower MAE than (d). However, SI-MAE can correctly distinguish two detections.

size-sensitive, which is not a proper choice for SOD tasks when the sizes of objects in a given image are highly imbalanced. As demonstrated in Fig. 1, SOD tasks typically involve multiple salient objects with diverse sizes. In this sense, prediction errors would be dominated by those larger objects, leading models to overlook small salient objects. Taking MAE as an example, Fig. 2(c) could merely detect the larger salient object but miss the smaller one on the right, while Fig. 2(d) could successfully capture all salient objects. However, Fig. 2(d) induces a worse MAE than Fig. 2(c), which is **counter-intuitive** to our visual perceptions. Large objects dominate size-sensitive metrics, consequently leading to practical performance degradation because there are many cases where small objects are critical for downstream tasks. For example, in a street view, traffic lights are usually of small size, but they play a significant role in autonomous driving tasks.

To address the issues above, we are interested in the following problem:

Can we develop an effective size-invariant criterion for imbalanced multi-object SOD?

The answer is affirmative in this paper. To begin with, we present a novel unified framework to understand why popular SOD metrics are size-sensitive. Specifically, given an image, we show that common criterion can be reformulated as a weighted (denoted by $\mathbb{P}_{x_i}$) sum of multiple independent parts, with each weighted term $\mathbb{P}_{x_i}$ being **highly related to the size** of the corresponding part. This creates an inductive bias toward objects of different sizes.

Motivated by this, we thus propose a simple yet effective paradigm for size-invariant SOD evaluation. The key idea is to modify the size-related term $\mathbb{P}_{x_i}$ into a size-invariant constant, ensuring equal treatment for each salient object regardless of size. Meanwhile, we introduce a generic Size-Invariant SOD (SI-SOD) optimization loss to pursue our size-invariant goal practically.

To show the effectiveness of our proposed paradigm, we then investigate the generalization performance of the SI-SOD algorithm. To the best of our knowledge, such a problem **remains barely explored** in the SOD community. As a result, we find that for composite losses (defined in Sec. 3.1), the size-invariant loss function leads to a sharper bound than its size-sensitive counterparts.

Finally, extensive experiments over a range of benchmark datasets speak to the efficacy of our proposed method.

2. Related Work

In recent years, SOD achieved considerable progress with elaborate frameworks and well-designed losses. We give a brief overview of SOD methods here and a detailed description of evaluation metrics in App. A.

Architecture-focused methods usually adopt convolutional networks as basic modules since their great success. For example, UCF (Zhang et al., 2017b) introduced a reformulated dropout after specific convolutional layers to learn deep uncertain convolutional features. DCL (Li & Yu, 2016) adopted a multi-stream framework, with the pixel-level fully convolutional stream to improve pixel-level accuracy. A common way to extract multi-level features is to design a bottom-up/top-down architecture, which resembles the U-Net (Ronneberger et al., 2015). PiCANet (Liu et al., 2018a) proposed a pixel-wise contextual attention network to selectively attend to informative context locations for each pixel and embed global and local networks into a U-Net architecture. RDCPN (Wu et al., 2021) introduced a novel multi-level ROIAlign-based decoder to adaptively aggregate multi-level features for better mask predictions. Similar structures are also utilized in recent works, including EDN (Wu et al., 2022), ICON (Zhuge et al., 2022), Bi-Directional (Zhang et al., 2018), CANet (Ren et al., 2021), etc. (Piao et al., 2019) designed a refinement block to fully extract and fuse multi-scale features, successfully achieving excellent performance on most datasets. Based on this, (Ji et al., 2022) further exploited a cascaded hierarchical feature fusion strategy to promote efficient information interaction of multi-level contextual features and efficiently improve contextual represent ability.

Multi-source-based methods have recently become popular. Specifically, both PoolNet (Liu et al., 2019) and MENet (Wang et al., 2023) conducted joint supervision of salient objects and object boundaries at each side-output. (Ji et al., 2023) used thermal infrared images as extra input to deal with rainy, overexposure, or low-light occasions, and achieved effective results. Depth information is also widely used in SOD, which is usually named as RGB-D SOD. For instance, (Ji et al., 2020; Zhang et al., 2023; Li et al., 2023a;b) introduced depth map to SOD and significantly improved the detection performance. Furthermore, (Zhang et al., 2019; 2020b) utilized light field data as an auxiliary for SOD and achieved state-of-the-art performance

at that time. Some extensive works such as (Zhang et al., 2021b; Ji et al., 2023; Li et al., 2023a) successfully deal with video SOD tasks exploiting the inter-frame information.

There are also previous works analyzing the evaluation in SOD. (Bylinskii et al., 2019) provided a comprehensive analysis of eight different evaluation metrics and their properties. (Borji et al., 2013a) performed a comparison of dozens of methods on many datasets to explore the consistency between the model ranking and practical performance. However, little attention has been paid to occasions where multiple salient objects co-exist, which is quiet common in the real world.

3. A Novel Size-invariant Evaluation Protocol

In this section, we begin by discussing why the commonly used metrics, such as Mean Absolute Error (MAE) and F-score, are not suitable for evaluating on imbalanced multi-object occasions. We then introduce methods to improve these metrics, aiming for a size-invariant SOD evaluation.

3.1. Revisiting Current SOD Evaluation Metrics

We start our analysis from standard functions, which could be divided into two groups: *separable* and *composite* functions, expressed as follows:

Definition 3.1 (*Separable Function*)**.** Given a predictor f, a function g applied to f is separable if the following equation formally holds:

$$g(f(X), Y) = \sum_{i=1}^{n} w_{X_i} \cdot g(f(X_i), Y_i), \tag{1}$$

with

$$\bigcup_{i=1}^{n} X_i = X, \bigcap_{i=1}^{n} X_i = \varnothing, \tag{2}$$

where X is the input and Y is the ground truth; $(X_1, X_2, \cdots, X_n)$ are n non-intersect parts of X; and w_{X_i} is an X_i-related weight for the term $g(f(X_i), Y_i)$.

According to the definition above, we realize that the point-wise evaluation metrics in the SOD community are separable (say Mean Absolute Error (MAE) (Perazzi et al., 2012) and Mean Square Error (MSE)).

Definition 3.2 (*Composite Function*)**.** *Composite Functions* are a series of compositions of separable functions Eq. (1), denoted by

$$G(f(X), Y) = (g_1 \circ g_2 \circ \cdots \circ g_T)(f(X), Y),$$

where T is the number of compositions.

According to the definition above, complicated evaluation metrics such as F-score (Achanta et al., 2009), IOU (Girshick et al., 2014) and AUC (Borji et al., 2013b) are composite. In what follows, we will discuss each of them respectively. For simplicity, we abbreviate $g(f(X), Y)$ and $G(f(X), Y)$ as $g(f)$ and $G(f)$ for a clear presentation.

Current separable metrics are NOT size-invariant. In SOD, the model $f : \mathbb{R}^S \to \mathbb{R}^S$ takes an image X with label Y as input, aiming to make a binary classification for each pixel, where S is the size of the image and $Y = \{0, 1\}^S$ is the pixel-level ground-truth. In light of this, the image could be naturally divided into N_c parts based on the location of salient objects.

Therefore, given a certain separable SOD metric g, let $g(f_i) := g(f(X_i), Y_i)$, we can rewrite it as Eq. (1) does:

$$g(f) = \sum_{i=1}^{N_c} \mathbb{P}_{X_i} \cdot g(f_i), \tag{3}$$

where $\mathbb{P}_{X_i}$ is the **size-sensitive** weight for the i-th part of X, which brings about inductive bias in evaluation.

Taking MAE as an example, the following equation holds:

$$\begin{aligned} \mathsf{MAE}(f) &= \sum_{i=1}^{N_c} \frac{\|f(X_i) - Y_i\|_{1,1}}{S} \\ &= \sum_{i=1}^{N_c} \frac{S_i}{S} \cdot \frac{\|f(X_i) - Y_i\|_{1,1}}{S_i} \\ &= \sum_{i=1}^{N_c} \frac{S_i}{S} \cdot \mathsf{MAE}(f_i) \\ &= \sum_{i=1}^{N_c} \mathbb{P}_{X_i} \cdot \mathsf{MAE}(f_i). \end{aligned} \tag{4}$$

Here we have $\mathbb{P}_{X_i} = S_i/S$, where S_i represents the size of the i-th part. It is explicitly that the current MAE metric for SOD is size-sensitive, where **larger objects would be paid more attention**. Similar results can be drawn for other point-wise metrics in the SOD community.

Current composite metrics are NOT size-invariant. Similarly, we formally rewrite the composite metric $G(f)$ as follows:

$$G(f) = \frac{\sum_{i=1}^{N_c} \mathbb{P}_{X_i}(a_1 g_1(f_i) + \cdots + a_T g_T(f_i))}{\sum_{i=1}^{N_c} \mathbb{P}_{X_i}(b_1 g_1(f_i) + \cdots + b_T g_T(f_i))}, \tag{5}$$

where again $g_t(f_i) := g_t(f(X_i), Y_i), t \in [T]$ is a certain separable metric value over the i-th part X_i; a_i and b_i represent coefficients for different composite functions, and here $\mathbb{P}_{X_i}$ is also a **size-sensitive** weight for each separable part of X.

Specifically, in terms of the widely used F-score (Achanta et al., 2009), we have:

$$\begin{aligned}\mathsf{F}(f) &= \frac{2\sum_{i=1}^{N_c}\mathsf{TP}(f_i)}{\sum_{i=1}^{N_c}[2\mathsf{TP}(f_i)+\mathsf{FP}(f_i)+\mathsf{FN}(f_i)]} \\ &= \frac{2\sum_{i=1}^{N_c}\frac{S_i}{S}\cdot\frac{\mathsf{TP}(f_i)}{S_i}}{\sum_{i=1}^{N_c}[\frac{S_i}{S}\cdot(2\frac{\mathsf{TP}(f_i)}{S_i}+\frac{\mathsf{FP}(f_i)}{S_i}+\frac{\mathsf{FN}(f_i)}{S_i})]} \\ &= \frac{2\sum_{i=1}^{N_c}\mathbb{P}_{X_i}\cdot\mathsf{TPR}(f_i)}{\sum_{i=1}^{N_c}\mathbb{P}_{X_i}\cdot(2\mathsf{TPR}(f_i)+\mathsf{FPR}(f_i)+\mathsf{FNR}(f_i))},\end{aligned} \tag{6}$$

where $\mathsf{TP}(f_i)$, $\mathsf{FP}(f_i)$, $\mathsf{FN}(f_i)$ represent the number of **T**rue **P**ositives, **F**alse **P**ositives and **F**alse **N**egatives within X_i, and $\mathsf{TPR}(f_i), \mathsf{FPR}(f_i), \mathsf{FNR}(f_i)$ represent the corresponding **T**rue **P**ositive **R**ate, **F**alse **P**ositive **R**ate, and **F**alse **N**egative **R**ate, which are all separable functions mentioned above. In this case, we still have $\mathbb{P}_{X_i} = S_i/S$, which is sensitive to the size of salient objects. Similar conclusion also applies to metrics like AUC, with analysis in App. C.

Why size-invariance MATTERS? We have realized that the existing widely adopted metrics would inevitably introduce biased weights for objects of different sizes. With this imbalance, smaller objects are suppressed by larger ones, and therefore are easily ignored in both evaluation and prediction. Unfortunately, as shown in Fig. 1, practical SOD tasks usually involve multiple objects of various sizes, including small yet critical ones. For example, Fig. 2(c) totally overlooks a small object, but enjoys a similar MAE compared to Fig. 2(d), which contradicts our visual perceptions. To rectify this, we introduce the principles of size-invariant evaluation in the next section.

3.2. Principles of Size-Invariant Evaluation

Based on the discussions above, the fundamental limitation of the current evaluation lies in the size-sensitive $\mathbb{P}_{X_i}$. Therefore, a principal way to achieve size-invariant evaluation is to eliminate the effect of the weighting term $\mathbb{P}_{X_i}$. In this paper, we propose a simple yet effective size-invariant protocol:

$$g_{\mathsf{SI}}(f) = \frac{1}{N_c}\sum_{i=1}^{N_c} 1\cdot g(f_i), \tag{7}$$

$$G_{\mathsf{SI}}(f) = \frac{1}{N_c}\sum_{i=1}^{N_c} 1\cdot\frac{(a_1g_1(f_i)+\cdots+a_Tg_T(f_i))}{(b_1g_1(f_i)+\cdots+b_Tg_T(f_i))}, \tag{8}$$

where $\mathbb{P}_{X_i}$ is replaced by a constant 1. The size-sensitive weight is directly eliminated, and we naturally arrive at size-invariance.

In what follows, we will adopt widely used metrics, i.e., MAE, F-score and AUC, to instantiate our size-invariant principles. Note that our proposed strategy could also be applied to other metrics as mentioned in Sec. 4.

3.2.1. Size-Invariant MAE

According to Eq. (7), SI-MAE is expressed as follows:

$$\mathsf{SI\text{-}MAE}(f) = \frac{1}{N_c}\sum_{i=1}^{N_c} 1\cdot\mathsf{MAE}(f_i). \tag{9}$$

Here our primary focus is on dividing the image into N_c parts. Motivated by the success of object detection (Ren et al., 2015) (Redmon et al., 2016), we segment salient objects into a series of foreground frames by their minimum bounding boxes, and pixels that do not form part of any bounding box are treated as the background.

Ideally, assume that there are K salient objects in an image and let $C_k = \{(a_j, b_j)\}_{j=1}^{M_k}, k \in [K]$ be the coordinate set for the object k, then the minimum bounding box for object k could be determined clockwise by the following vertex coordinates:

$$\begin{aligned}X_k^{fore} =&\{(a_k^{min}, b_k^{max}), (a_k^{max}, b_k^{max}),\\ &(a_k^{max}, b_k^{min}), (a_k^{min}, b_k^{min})\},\end{aligned} \tag{10}$$

where $a_k^{min}, a_k^{max}, b_k^{min}, b_k^{max}$ are the minimum and maximum coordinates in C_k, respectively.

Correspondingly, the background frame is defined as follows:

$$X_{K+1}^{back} = X \setminus F, \tag{11}$$

where

$$F = \bigcup_{k=1}^{K} X_k^{fore} \tag{12}$$

is the collection of all minimum bounding boxes for salient objects.

However, since there is no instance-level label to distinguish different objects in most practical datasets, we instead regard each connected component composed of salient objects in the saliency map as an independent proxy C_k. Some examples of partitions are presented in Fig. 3, where an image will be divided into $N_c = K + 1$ parts, including K foreground frames and a background frame. Please refer to implementation details in Sec. 5.1 for more details of the connected component.

The bounding boxes are similar to those widely applied in the area of object detection (Xiao & Marlet, 2020) (Ding et al., 2022). However, object detection makes bounding box regression to match the predicted boxes as close to the ground-truth boxes as possible, while our approach generates the bounding boxes from the ground-truth binary masks and exploits them as auxiliary tools to calculate the loss and metric results around each salient object.

(a) Single-object scenario (b) Multi-object scenario

Figure 3. Examples of partitions. In Fig. 3(a), there is a foreground frame ① and a background frame ②. In Fig. 3(b), there are five foreground frames from ① to ⑤, and a background frame ⑥.

In this way, the goal of SI-MAE becomes

$$\text{SI-MAE}(f) = \frac{1}{K+\alpha}\left[\sum_{k=1}^{K}\text{MAE}(f_k^{fore}) + \alpha\text{MAE}(f_{K+1}^{back})\right], \tag{13}$$

where a parameter α, determined by the ratio of the size of the background and the sum of all foreground frames, namely $\alpha = \frac{S_{K+1}^{back}}{\sum_{k=1}^{K} S_k^{fore}}$, is further introduced to balance the model attention adaptively. By doing so, the predictor could not only pay equal consideration to salient objects of various sizes, but also impose an appropriate penalty for misclassifications in the background. This plays an important role in reducing the false positives as illustrated in Sec. 5.3.3.

In the following, we make a brief discussion between MAE and our proposed SI-MAE, with proof in App. D.1.

Proposition 3.3 (Informal)**.** *Given two different predictors f_A and f_B, the following two possible cases suggest that* SI-MAE *is more effective than* MAE *during evaluation.*

Case 1: *Assume that there is a single salient object (i.e., $K = 1$), with two different results from predictors f_A and f_B. In this case, there is no imbalance from different sizes of objects, and therefore* SI-MAE *is equivalent to* MAE*.*

Case 2: *Suppose there are two salient objects ($K = 2$) where f_A and f_B detect the same amount of salient pixels in an image X. Meanwhile, assume that f_A only predicts C_2 perfectly while f_B could somewhat recognize C_1 and C_2 partially. In this case, f_B should still be better than f_A since f_A totally fails on C_1. Unfortunately, if $S_1^{fore} < S_2^{fore}$, $\text{MAE}(f_A) = \text{MAE}(f_B)$ holds but we have $\text{SI-MAE}(f_A) > \text{SI-MAE}(f_B)$.*

Remark. Fig. 2 provides a toy example for **Case 2**. Discussions above support that MAE is sensitive to the size of objects concerning multiple object cases, yet SI-MAE can serve our expectations better. We also extend our analysis to the case with $K \geq 3$, which consistently suggests the efficacy of SI-MAE. The detailed discussion is attached to App. D.1.

3.2.2. Size-Invariant Composite Metrics

Here we instantiate our size-invariant principle with common composite metrics, including F-score and AUC.

As to the composite metric F-score, we define SI-F as follows:

$$\text{SI-F}(f) = \frac{1}{K}\sum_{i=1}^{K}\text{F}(f_k^{fore}), \tag{14}$$

where $X_1^{fore}, \cdots, X_K^{fore}$ denote foreground frames. Similar to SI-MAE, we give a proposition in App. B to support that in multiple object cases, SI-F can serve our expectations better.

As to another common composite metric AUC, we similarly define SI-AUC as follows:

$$\text{SI-AUC}(f) = \frac{1}{K}\sum_{i=1}^{K}\text{AUC}(f_k^{fore}), \tag{15}$$

where $X_1^{fore}, \cdots, X_K^{fore}$ denote foreground frames. The analysis of SI-AUC is deferred to App. C due to space limitations.

4. How to Practically Pursue Size-Invariance?

In previous sections, we outlined how to achieve size-invariant evaluation for SOD. Now this section explores how to directly optimize these size-invariant metrics to promote practical SOD performance.

4.1. A Generic Size-Invariant Optimization Goal

Motivated by the principles of the size-invariant evaluation, our optimization goal is expressed as follows:

$$\mathcal{L}_{\text{SI}}(f) = \sum_{k=1}^{K}\ell(f_k^{fore}) + \alpha\ell(f_{K+1}^{back}), \tag{16}$$

where $\ell(\cdot)$ could be any popular loss in the SOD community (such as BCE or IOU). For simplicity, we let $\ell(f) := \ell(f(X), Y)$ and $\ell(f_i) := \ell(f(X_i), Y_i)$. Similar to Eq. (13), if $\ell(\cdot)$ is separable, we set $\alpha = \frac{S_{K+1}^{back}}{\sum_{k=1}^{K} S_k^{fore}}$; for composite losses like DiceLoss (Milletari et al., 2016) and IOU Loss (Yu et al., 2016), we set $\alpha = 0$ because the TPR is always 0 in the background. Specifically in App. F.4, we describe detailed implementations of Size-Invariant Optimization for different backbones discussed in Sec. 5.

As discussed in Sec. 3.2, Eq. (16) ensures that the model treats all objects equally regardless of size, thus improving

the detection of smaller objects. We give the following proposition to illustrate the mechanism of SI-SOD, with proof in App. E.1.

Proposition 4.1 (Mechanism of SI-SOD)**.** *Given a separable loss function $\ell(\cdot)$ and its corresponding size-invariant loss $\mathcal{L}_{\mathsf{SI}}(\cdot)$, then for a certain scenario:*

1. *when $S_i < \frac{S}{K+\alpha}$, we have $w_{\mathcal{L}_{\mathsf{SI}}}(x_i) > w_\ell(x_i)$,*
2. *when $S_i < S_j$, we have $w_{\mathcal{L}_{\mathsf{SI}}}(x_i) > w_{\mathcal{L}_{\mathsf{SI}}}(x_j)$.*

*where $w_{\mathcal{L}_{\mathsf{SI}}}(x_i)$ is the **weight** of pixel-level loss in X_i with $\mathcal{L}_{\mathsf{SI}}$, and $w_\ell(x_i)$ is the **weight** of pixel-level loss in X_i with the original loss ℓ.*

Remark. Compared to standard losses such as BCE, SI-SOD adaptively adjusts the weight of pixel-level loss to ensure equal treatment on different objects.

Item 1 illustrates that smaller objects, which fall below a certain size threshold, will produce more loss. *Item* 2 describes that SI-SOD increases the weight for pixels in smaller salient objects, finally alleviating size-sensitivity.

4.2. Generalization Bound

In this section, we theoretically demonstrate that SI-SOD can generalize to common SOD tasks, despite several challenges.

First, SOD is considered as structured prediction (Ciliberto et al., 2020; Li et al., 2021), where couplings between output substructures make it difficult to directly apply Rademacher Complexity-based techniques in theoretical analysis. The standard result to bound the empirical Rademacher complexity (Michel Ledoux, 1991) holds when the prediction functions are real-valued. To overcome this, we adopt the vector contraction inequality (Maurer, 2016) to extend it from real-valued analysis to vector-valued ones, and consequently reach a **sharper** result with Lipschitz properties (Foster & Rakhlin, 2019).

Another challenge lies in the diversity of losses, which hinders exploring the generalization properties within a coordinated framework. Therefore, by studying from the view of separable and composite functions respectively, we obtain Lipschitz properties (Dembczyński et al., 2017) for both categories and ultimately achieve a **unified** conclusion.

We present our conclusions here, and the proof is deferred to App. E.2.

Theorem 4.2 (**Generalization Bound for SI-SOD**)**.** *Assume $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}^K\}$, where $K = H \times W$ is the pixel count in an image, $g^{(i)}$ is the risk over i-th sample, and is L-Lipschitz with respect to the l_∞ norm, (i.e. $\|g(x) - g(\tilde{x})\|_\infty \le L \cdot \|x - y\|_\infty$). When there are N i.i.d. samples, there exists a constant $C > 0$ for any $\epsilon > 0$, the following generalization bound holds with probability at least $1 - \delta$:*

$$\begin{aligned}&\sup_{f\in\mathcal{F}}(\mathbb{E}[g(f)] - \hat{\mathbb{E}}[g(f)])\\ &\le C \cdot \frac{L\sqrt{K}}{N} \cdot \max_i \mathfrak{R}_N(\mathcal{F}|_i) \cdot \log^{\frac{3}{2}+\epsilon}\left(\frac{N}{\max_i \mathfrak{R}_N(\mathcal{F}|_i)}\right)\\ &\quad + 3\sqrt{\frac{\log\frac{2}{\delta}}{2N}},\end{aligned} \tag{17}$$

where again $g(f(X,Y)) := g(f)$, $\mathbb{E}[g(f)]$ and $\hat{\mathbb{E}}[g(f)]$ represent the expected risk and empirical risk. $\mathfrak{R}_N(\mathcal{F}; x_{1:N}) = \max_{x_{1:N}\in\mathcal{X}} \mathfrak{R}(\mathcal{F}; x_{1:N})$ denotes the worst-case Rademacher complexity, and we let $\mathfrak{R}_N(\mathcal{F}|i)$ denote its restriction to output coordinate i. Specifically,

Case 1: *For separable loss functions $\ell(\cdot)$, if it is μ-Lipschitz, we have $L = \mu$.*

Case 2: *For composite loss functions, when $\ell(\cdot)$ is DiceLoss (Milletari et al., 2016), we have $L = \frac{4}{\rho}$, where $\rho = \min \frac{S_l^{1,i}}{S_l^i}$, which represents the minimum proportion of the salient object in the l-th frame within the i-th sample.*

Remark. We reach a bound of $\mathcal{O}(\frac{\sqrt{K}\log N}{N})$, which indicates reliable generalization with a large training set.

Specifically, for case 2, the original composite loss $\ell(\cdot)$, still taking DiceLoss as an example, will result in a $\rho' = \frac{S^{1,i}}{S^i}$, which represents the proportion of salient pixels in an image. It is obvious that $\rho > \rho'$ because SI-SOD enlarges the proportion by reducing the denominator from the whole image to a bounding box, and finally leads to a **smaller** L and **a sharper bound**.

5. Experiments

In this section, we describe some details of the experiments and present our results. **Due to space limitations, please refer to App. F for an extended version.**

5.1. Experimental Setups

Datasets. Eight datasets, DUTS (Wang et al., 2017), ECSSD (Yan et al., 2013), DUT-OMRON (Yang et al., 2013), HKU-IS (Li & Yu, 2015), MSOD (Deng et al., 2023), PASCAL-S (Yan et al., 2013), SOD (Movahedi & Elder, 2010) and XPIE (Xia et al., 2017), are included in the experiment. Following common practice, we train our network on

Table 1. Quantitative comparisons on MSOD and DUTS-TE. The better results are shown with **bold**, and darker color indicates superior results. Metrics with ↑ mean higher value represents better performance, while ↓ mean lower value represents better performance.

Dataset	Methods	MAE ↓	SI-MAE ↓	AUC ↑	SI-AUC ↑	F_m^β ↑	SI-F_m^β ↑	F_{max}^β ↑	SI-F_{max}^β ↑	E_m ↑
MSOD	PoolNet	0.0752	0.1196	0.9375	0.9563	0.6645	0.6397	0.7755	0.8402	0.7529
	+ ***Ours***	**0.0635**	**0.0924**	**0.9553**	**0.9721**	**0.7314**	**0.7467**	**0.8200**	**0.8867**	**0.8286**
	LDF	0.0508	0.0946	0.8719	0.9246	0.7589	0.6691	0.8144	0.7575	0.8241
	+ ***Ours***	**0.0506**	**0.0893**	**0.9530**	**0.9441**	**0.7796**	**0.7573**	**0.8415**	**0.8879**	**0.8726**
	ICON	0.0545	0.0945	0.8973	0.8909	0.7687	0.7029	0.8178	0.7789	0.8487
	+ ***Ours***	**0.0535**	**0.0830**	**0.9537**	**0.9514**	**0.7691**	**0.7738**	**0.8373**	**0.8665**	**0.8742**
	GateNet	**0.0442**	0.0808	0.9331	0.9244	0.8005	0.7581	0.8510	0.8434	0.8776
	+ ***Ours***	0.0444	**0.0734**	**0.9456**	**0.9436**	**0.8157**	**0.8083**	**0.8570**	**0.8724**	**0.8972**
	EDN	0.0467	0.0788	0.9196	0.9188	0.7925	0.7635	0.8410	0.8321	0.8712
	+ ***Ours***	**0.0453**	**0.0724**	**0.9401**	**0.9387**	**0.8057**	**0.7990**	**0.8555**	**0.8619**	**0.8936**
DUTS-TE	PoolNet	0.0656	0.0609	0.9607	0.9716	0.7200	0.7569	0.8245	0.8715	0.8103
	+ ***Ours***	**0.0621**	**0.0562**	**0.9706**	**0.9824**	**0.7479**	**0.8172**	**0.8438**	**0.9029**	**0.8478**
	LDF	**0.0419**	**0.0410**	0.9337	0.9680	**0.8203**	0.8201	0.8735	0.8802	0.8821
	+ ***Ours***	0.0440	0.0422	**0.9690**	**0.9756**	0.8076	**0.8388**	**0.8736**	**0.9117**	**0.8895**
	ICON	0.0461	0.0454	0.9469	0.9424	**0.8131**	0.8270	**0.8648**	0.8815	0.8858
	+ ***Ours***	**0.0454**	**0.0435**	**0.9640**	**0.9706**	0.8031	**0.8395**	0.8629	**0.8958**	**0.8921**
	GateNet	**0.0383**	0.0380	0.9629	0.9619	**0.8292**	0.8519	**0.8835**	0.9041	**0.9053**
	+ ***Ours***	0.0399	**0.0375**	**0.9663**	**0.9692**	0.8185	**0.8687**	0.8743	**0.9116**	0.9038
	EDN	**0.0389**	0.0388	0.9600	0.9611	**0.8288**	0.8565	0.8752	0.9017	0.9033
	+ ***Ours***	0.0392	**0.0381**	**0.9658**	**0.9687**	0.8260	**0.8672**	**0.8765**	**0.9119**	**0.9072**

the DUTS training set (DUTS-TR) and test it on the DUTS test set (DUTS-TE) and the other seven datasets. Detailed introductions on these datasets are deferred to App. F.1.

Competitors. To demonstrate the effectiveness of size-invariant loss, we integrate it into five state-of-the-art backbones: EDN (Wu et al., 2022), ICON (Zhuge et al., 2022), GateNet (Zhao et al., 2020), LDF (Wei et al., 2020), PoolNet (Liu et al., 2019). EDN, ICON, and LDF utilize DiceLoss or IOULoss to handle the potential imbalanced distribution, and ICON specifically focuses on the macro-integrity, which are summarized at App. F.2 with details. Specifically, we modify the original loss functions into their corresponding size-invariant versions following Eq. (16), and re-train the network with the same setting. Correspondingly, we also compare the time cost of our method with original optimization frameworks, which is deferred at App. G.6.

Evaluation Metrics. Apart from our proposed metrics SI-MAE, SI-F and SI-AUC, we also include common metrics such as Mean Absolute Error(MAE), max F-measure(F_{max}^β), mean F-measure(F_m^β) and AUC.

Another widely used metric is E_m introduced by (Fan et al., 2018), which is a newly proposed metric considering both global and local information. Definitions and calculations of all metrics are deferred to App. A.

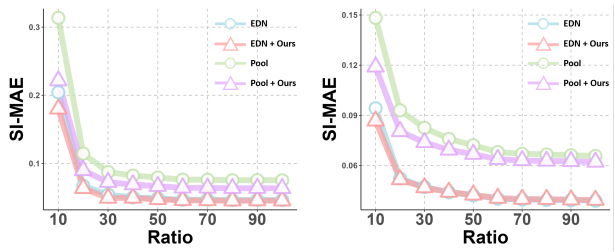
Implementation Details. We carried out the experiments on a single GeForce RTX 3090. To ensure fairness, both the original and modified backbones are trained under identical settings. All images are resized into 384×384 for training and testing, and the ResNet50 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) is loaded. Specific settings and optimization details for each backbone are deferred at App. F.3 and App. F.4.

We preprocess the dataset with the package *skimage*, which identifies connected components with the ground-truth mask. Then we obtain the minimum bounding box following Eq. (10). Note that all procedures can also be done during training without any preprocessing.

5.2. Overall Performance

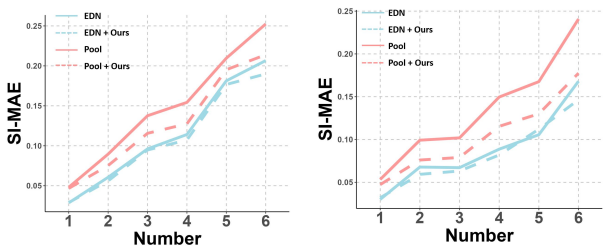
As mentioned above, we re-train the backbones with our size-invariant loss for a fair comparison. Tab. 1 shows the results on MSOD and DUTS-TE. The result is shown in a pair of backbones before and after applying our size-invariant loss, with the superior result highlighted in **bold**.

Since samples in MSOD contain multiple salient objects, it naturally arises that small objects can be overlooked due to the imbalance. Therefore, all backbones with our loss achieve considerable improvements on nearly all metrics, even including the original MAE. Averagely, our method outperforms other frameworks by around 0.012, 0.038, 0.070, 0.065, 0.038 on SI-MAE, SI-AUC, SI-F_m^β, SI-F_{max}^β and E_m, respectively.



(a) MSOD (b) DUTS

Figure 4. SI-MAE performance on objects with different sizes on two representative datasets, with EDN and PoolNet as backbones.



(a) MSOD (b) DUTS

Figure 5. SI-MAE performance with different object numbers on two representative datasets, with EDN and PoolNet as backbones.

Tab. 1 also shows the performance on DUTS-TE. Our method achieves better results on almost all size-invariant metrics and E_m, and stays competitive in terms of original MAE and F-score. This justifies that the size-invariant loss achieves similar performance on single-object scenarios, suggesting the superior generalization ability of our loss. Averagely, our method outperforms other frameworks by around 0.001, 0.012, 0.024, 0.020, 0.011 on SI-MAE, SI-AUC, SI-F_m, SI-F^{β}_{max}, E_m on DUTS-TE. Results on other datasets are deferred to Tab. 3.

Fig. 6 shows the qualitative comparison on different backbones. While the original backbones may fail to detect all the salient objects in some hard samples, our method can significantly improve the detection on multi-object occasions. For example, in the 1st image, EDN only finds the largest sailboat on the right but fails to detect two smaller targets on the left, while ours additionally detects two small sailboats. In the 4th image, EDN detects fewer false positive pixels and ICON detects one more salient object with our loss. In the 3rd, 5th, and 6th images, all backbones detect more salient objects at the right part after using our loss. More qualitative comparisons are deferred to App. G.2.

5.3. Fine-grained Analysis

5.3.1. Performance with Respect to Sizes

We conduct size-relevant analysis on five datasets. As it is the size of salient objects that our method focuses on, we divide all salient objects into ten groups according to their proportion to the entire image, ranging from [0%, 10%], [10%,20%], and finally up to [90%, 100%]. We evaluate the performance within each group, and here we only take foreground frames into account to concentrate on the detection performance of salient objects with different sizes.

From Fig. 4, we observe that all backbones perform well on larger objects but show remarkable improvements on smaller objects when using our method. This aligns with our objective to enhance the detection of smaller objects. Specifically, for objects with size in [0%, 10%] of the image, our method outperforms the previous backbone, say EDN, by around 0.024 on SI-MAE on the MSOD dataset. As seen in Fig. 1(a), small-size salient objects usually account for the majority, therefore such improvement firmly speaks to our progress. This is not reflected by size-sensitive metrics like MAE, but can be directly revealed by our proposed SI-MAE. Performance analysis with respect to the object size on other backbones and datasets is deferred to App. G.3.

5.3.2. Performance with Respect to Object Numbers

We also conduct number-relevant analysis on five datasets to evaluate the performance on single-object and multi-object scenarios, as shown in Fig. 5. With the number of salient objects increasing, the SOD tasks are getting imbalanced, where some objects are more likely to be ignored. Therefore, we divide all samples into several groups according to the number of salient objects in the image.

Generally, our method shows substantial improvements in multi-object scenarios and remains competitive in single-object cases, which again justifies the generalization and universality of our method. Specifically, for samples with greater than or equal to two salient objects, EDN gains an improvement by around 0.007 on SI-MAE on the MSOD dataset after employing our size-invariant loss. Performance analysis with respect to the object numbers on other backbones and datasets is deferred to App. G.4.

5.3.3. Ablation Studies

To investigate how the parameter α works, we conduct ablation studies on α to verify its effectiveness. Here we set α among $0, 1, \frac{S^{back}}{S^{fore}}$. $\alpha = 0$ indicates that we do not consider the background frame and pay all attention to foreground frames, while $\alpha = 1$ means that we consider the background frame equally as other foreground frames, and $\alpha = \frac{S^{back}}{S^{fore}}$ is exactly our method. Fig. 7 illustrates the ablations on dataset MSOD and DUTS with the backbone EDN. $\alpha = 0$ induces an extreme result with a high score within foreground frames and a low score in the background frame because it solely

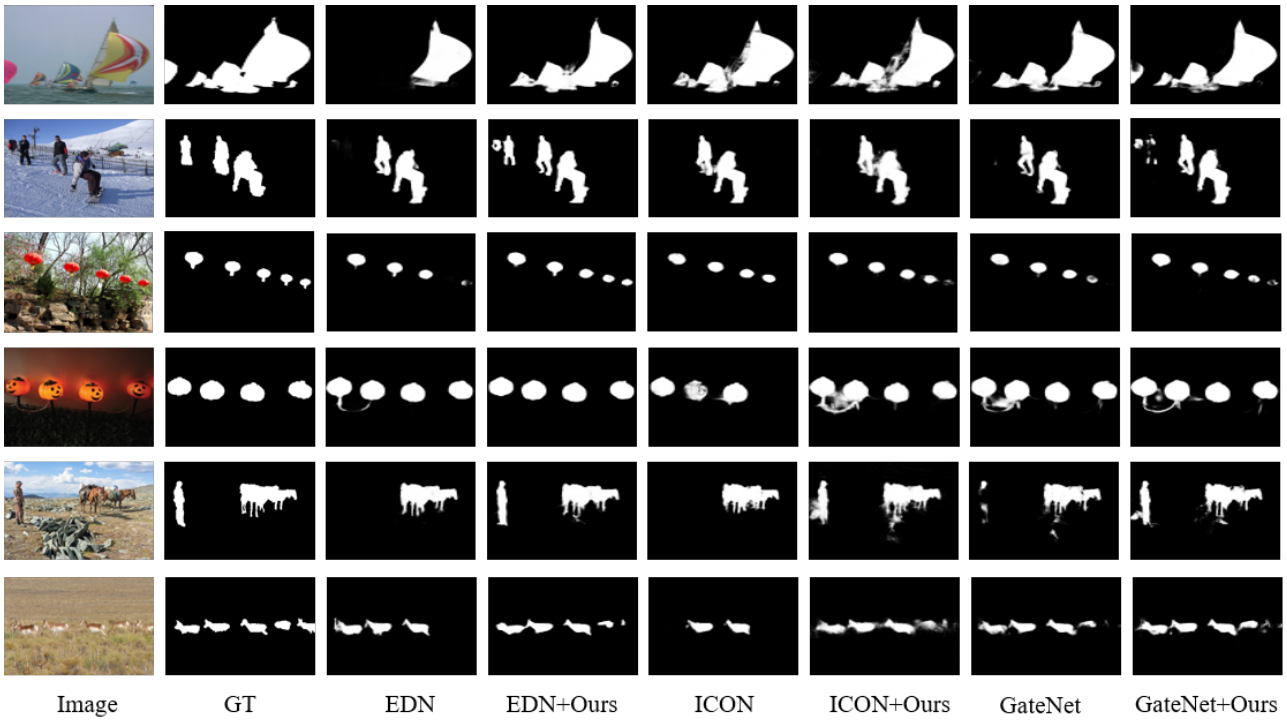
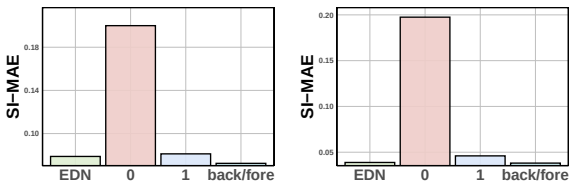


Figure 6. Qualitative comparison on different backbones.



(a) MSOD (b) DUTS

Figure 7. SI-MAE performance on two representative datasets with different value of α. EDN, 0, 1, and back/fore represent the original backbone EDN, $\alpha = 0$, $\alpha = 1$ and $\alpha = \frac{S^{back}}{S^{fore}}$, respectively.

focuses on foreground detection. $\alpha = 1$ alleviates the phenomenon, and surpasses the original framework on some metrics, but still predicts too many false positives, due to the slight penalty on the error within the background frame. Experiments on other datasets also speak to the efficacy of the α. More detailed results are deferred to App. G.5.

6. Conclusion

In this paper, we explore the size-invariance in SOD tasks. When multiple objects of various sizes co-exist, we observe that current evaluation metrics are size-sensitive, where larger objects are focused and smaller objects are likely overlooked. To rectify this, we introduce a generic approach to achieve size-invariance. Specifically, we propose SI-MAE and SI-F, which evaluate each salient object separately before merging their results. We further design an optimization framework directly toward this goal, which can adaptively balance the weights to ensure equal treatment on different objects. Theoretically, we provide evidence to support our proposed metrics and present the generalization analysis for our SI-SOD optimization loss. Comprehensive experiments consistently demonstrate the efficacy of our method.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102000, in part by National Natural Science Foundation of China: 62236008, U21B2038, U23B2051, U2001202, 61931008, 62122075, 61976202, 62206264 and 92370102, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDB0680000, in part by the Innovation Funding of ICT, CAS under Grant No.E000000, in part by the Taishan Scholar Project of Shandong Province under Grant tsqn202306079.

Impact Statement

We propose a general SOD method to deal with the potential bias toward small objects. For fairness-sensitive scenarios, it might be helpful to improve fairness for minority groups.

References

Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. Frequency-tuned salient region detection. In *CVPR*, Jun 2009.

Borji, A., Sihite, D. N., and Itti, L. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE TIP*, 22(1):55–69, 2013a.

Borji, A., Tavakoli, H. R., Sihite, D. N., and Itti, L. Analysis of scores, datasets, and models in visual saliency prediction. In *ICCV*, pp. 921–928, 2013b.

Borji, A., Cheng, M.-M., Hou, Q., Jiang, H., and Li, J. Salient object detection: A survey. *Computational Visual Media*, pp. 117–150, 2019.

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., and Durand, F. What do different evaluation metrics tell us about saliency models? *IEEE TPAMI*, 41(3):740–757, 2019.

Chen, H., Li, Y., Deng, Y., and Lin, G. Cnn-based rgb-d salient object detection: Learn, select, and fuse. *IJCV*, 129(7):2076–2096, 2021.

Ciliberto, C., Rosasco, L., and Rudi, A. A general framework for consistent structured prediction with implicit loss embeddings. *JMLR*, 21(1):3852–3918, 2020.

Dembczyński, K., Kotłowski, W., Koyejo, O., and Natarajan, N. Consistency analysis for binary classification revisited. In *ICML*, volume 70, pp. 961–969. PMLR, 06–11 Aug 2017.

Deng, B., French, A. P., and Pound, M. P. Addressing multiple salient object detection via dual-space long-range dependencies. *CVIU*, 235:103776, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Ding, J., Xue, N., Xia, G.-S., Bai, X., Yang, W., Yang, M. Y., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE TPAMI*, 44:7778–7796, 2022.

Fan, D.-P., Gong, C., Cao, Y., Ren, B., Cheng, M.-M., and Borji, A. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pp. 698–704, 7 2018.

Foster, D. J. and Rakhlin, A. ℓ_∞ vector contraction for rademacher complexity, 2019.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pp. 580–587, 2014.

Gui, S., Song, S., Qin, R., and Tang, Y. Remote sensing object detection in the deep learning era—a review. *Remote Sensing*, 16(2), 2024.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Ji, W., Li, J., Zhang, M., Piao, Y., and Lu, H. Accurate rgb-d salient object detection via collaborative learning. In *ECCV*, pp. 52–69, 2020.

Ji, W., Yan, G., Li, J., Piao, Y., Yao, S., Zhang, M., Cheng, L., and Lu, H. Dmra: Depth-induced multi-scale recurrent attention network for rgb-d saliency detection. *IEEE TIP*, 31:2321–2336, 2022.

Ji, W., Li, J., Bian, C., Zhou, Z., Zhao, J., Yuille, A. L., and Cheng, L. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *CVPR*, pp. 1094–1104, 2023.

Jia, S. and Bruce, N. D. B. Richer and deeper supervision network for salient object detection. *ArXiv*, 2019.

Jiang, Y., Hua, C., Feng, Y., and Gao, Y. Hierarchical set-to-set representation for 3-d cross-modal retrieval. *IEEE TNNLS*, pp. 1–13, 2023.

Li, G. and Yu, Y. Deep contrast learning for salient object detection. In *CVPR*, pp. 478–487, 2016.

Li, G. and Yu, Z. Visual saliency based on multiscale deep features. In *CVPR*, pp. 5455–5463, June 2015.

Li, J., Ji, W., Bi, Q., Yan, C., Zhang, M., Piao, Y., Lu, H., et al. Joint semantic mining for weakly supervised rgb-d salient object detection. *NeurIPS*, 34:11945–11959, 2021.

Li, J., Ji, W., Wang, S., Li, W., and Cheng, L. Dvsod: Rgb-d video salient object detection. In *NeurIPS*, pp. 8774–8787, 2023a.

Li, J., Ji, W., Zhang, M., Piao, Y., Lu, H., and Cheng, L. Delving into calibrated depth for accurate rgb-d salient object detection. *IJCV*, 131(4):855–876, 2023b.

Li, Z., Tang, J., and Mei, T. Deep collaborative embedding for social image understanding. *IEEE TPAMI*, 41(9): 2070–2083, 2019.

Liu, J.-J., Hou, Q., Cheng, M.-M., Feng, J., and Jiang, J. A simple pooling-based design for real-time salient object detection. In *CVPR*, 2019.

Liu, N., Han, J., and Yang, M.-H. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018a.

Liu, N., Han, J., and Yang, M.-H. Picanet: Pixel-wise contextual attention learning for accurate saliency detection. *IEEE TIP*, Dec 2018b.

Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., and Jodoin, P.-M. Non-local deep features for salient object detection. In *CVPR*, 2017.

Ma, M., Xia, C., and Li, J. Pyramidal feature shrinking for salient object detection. In *AAAI*, volume 35, pp. 2311–2318, 2021.

Mahadevan, V. and Vasconcelos, N. Saliency-based discriminant tracking. In *CVPR*, 2009.

Margolin, R., Zelnik-Manor, L., and Tal, A. How to evaluate foreground maps. In *CVPR*, Jun 2014.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, Nov 2002.

Mason, S. J. and Graham, N. E. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, 128(584):2145–2166, 2002.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory*, pp. 3–17, 2016.

Michel Ledoux, M. T. *Probability in Banach Spaces*. Springer-Verlag, New York, 1991.

Milletari, F., Navab, N., and Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, pp. 565–571, 2016.

Movahedi, V. and Elder, J. H. Design and perceptual validation of performance measures for salient object segmentation. In *CVPR workshop*, pp. 49–56, 2010.

Perazzi, F., Krahenbuhl, P., Pritch, Y., and Hornung, A. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, Jun 2012.

Piao, Y., Ji, W., Li, J., Zhang, M., and Lu, H. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pp. 7254–7263, 2019.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *CVPR*, pp. 779–788, 2016.

Ren, Q., Lu, S., Zhang, J., and Hu, R. Salient object detection by fusing local and global contexts. *IEEE TMM*, 23: 1442–1453, 2021.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015.

Ren, Z., Gao, S., Chia, L.-T., and Tsang, I. W.-H. Region-based saliency detection and its application in object recognition. *IEEE TCSVT*, pp. 769–779, 2014.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, Jan 2015.

Sun, P., Zhang, W., Li, S., Guo, Y., Song, C., and Li, X. Learnable depth-sensitive attention for deep rgb-d saliency detection with multi-modal fusion architecture search. *IJCV*, 130(11):2822–2841, 2022.

Tang, J., Shu, X., Qi, G.-J., Li, Z., Wang, M., Yan, S., and Jain, R. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE TPAMI*, 39(8):1662–1674, 2017.

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., and Ruan, X. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017.

Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., and Yang, R. Salient object detection in the deep learning era: An in-depth survey. *IEEE TPAMI*, pp. 3239–3259, 2022.

Wang, Y., Wang, R., Fan, X., Wang, T., and He, X. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *CVPR*, pp. 10031–10040, 2023.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.

Wei, J., Wang, S., Wu, Z., Su, C., Huang, Q., and Tian, Q. Label decoupling framework for salient object detection. In *CVPR*, June 2020.

Wu, Y.-H., Liu, Y., Zhang, L., Gao, W., and Cheng, M.-M. Regularized densely-connected pyramid network for salient instance segmentation. *IEEE TIP*, 30:3897–3907, 2021.

Wu, Y.-H., Liu, Y., Zhang, L., Cheng, M.-M., and Ren, B. Edn: Salient object detection via extremely-downsampled network. *IEEE TIP*, pp. 3125–3136, 2022.

Xia, C., Li, J., Chen, X., Zheng, A., and Zhang, Y. What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors. In *CVPR*, pp. 4399–4407, 2017.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pp. 3485–3492, 2010.

Xiao, Y. and Marlet, R. Few-shot object detection and viewpoint estimation for objects in the wild. In *ECCV*, volume PP of *3*, pp. 192–210, 2020.

Yan, Q., Xu, L., Shi, J., and Jia, J. Hierarchical saliency detection. In *CVPR*, 2013.

Yang, C., Zhang, L., Lu, H., Ruan, s., and Yang, M.-H. Saliency detection via graph-based manifold ranking. In *CVPR*, pp. 3166–3173. IEEE, 2013.

Yang, Z., Xu, Q., Bao, S., He, Y., Cao, X., and Huang, Q. Optimizing two-way partial auc with an end-to-end framework. *IEEE TPAMI*, 45(8):10228–10246, 2023.

Yu, J., Jiang, Y., Wang, Z., Cao, Z., and Huang, T. Unitbox: An advanced object detection network. In *ACM MM*, pp. 516–520, 2016.

Zhang, D., Zhang, H., Tang, J., Hua, X.-S., and Sun, Q. Causal intervention for weakly-supervised semantic segmentation. In *NeurIPS*, volume 33, pp. 655–666, 2020a.

Zhang, J., Xie, J., Barnes, N., and Li, P. Learning generative vision transformer with energy-based latent space for saliency prediction. *NeurIPS*, 34:15448–15463, 2021a.

Zhang, L., Dai, J., Lu, H., He, Y., and Wang, G. A bi-directional message passing model for salient object detection. In *CVPR*, Jun 2018.

Zhang, M., Li, J., Wei, J., Piao, Y., and Lu, H. Memory-oriented decoder for light field salient object detection. *NeurIPS*, pp. 896–906, 2019.

Zhang, M., Ji, W., Piao, Y., Li, J., Zhang, Y., Xu, S., and Lu, H. Lfnet: Light field fusion network for salient object detection. *IEEE TIP*, 29:6276–6287, 2020b.

Zhang, M., Liu, J., Wang, Y., Piao, Y., Yao, S., Ji, W., Li, J., Lu, H., and Luo, Z. Dynamic context-sensitive filtering network for video salient object detection. In *ICCV*, pp. 1553–1563, 2021b.

Zhang, M., Yao, S., Hu, B., Piao, Y., and Ji, W. C2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection. *IEEE TMM*, 25:5142–5154, 2023.

Zhang, P., Wang, D., Lu, H., Wang, H., and Ruan, X. Amulet: Aggregating multi-level convolutional features for salient object detection. *arXiv*, 2017a.

Zhang, P., Wang, D., Lu, H., Wang, H., and Yin, B. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, pp. 212–221, 2017b.

Zhao, X., Pang, Y., Zhang, L., Lu, H., and Zhang, L. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, 2020.

Zhuge, M., Fan, D.-P., Liu, N., Zhang, D., Xu, D., and Shao, L. Salient object detection via integrity learning. *IEEE TPAMI*, 2022.

Contents

A. Evaluation Metrics for SOD

Different from usual classification tasks where we calculate accuracy on the image level, SOD requires evaluation pixel by pixel. Other pixel-level tasks like semantic segmentation adopt mIOU, which utilizes the mean IOU over all classes as the metric. However, in SOD all salient objects are labeled with 1 without further class labels. Therefore, there is no representative metric, and the following are commonly utilized:

MAE **(Perazzi et al., 2012).** It measures the average absolute error pixel-wise. The prediction is normalized to $[0, 1]$ when calculating the errors from the ground truth. It is defined as:

$$\mathsf{MAE} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |f_N(X)_{(i,j)} - Y_{(i,j)}|, \tag{18}$$

where $f_N(X)$ and Y are the normalized prediction map and the saliency map, respectively.

F**-score (Achanta et al., 2009).** It is designed to deal with imbalanced distribution and comprehensively considers both precision and recall. The original F-score is defined as follows:

$$\mathsf{F} = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{19}$$

where

$$Precision = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FP}}, \ Recall = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FN}}, \tag{20}$$

where $\mathsf{TP}, \mathsf{TN}, \mathsf{FP}, \mathsf{FN}$ are **T**rue **P**ositive, **T**rue **N**egative, **F**alse **P**ositive and **F**alse **N**egative. A set of thresholds is applied to generate the binary result when calculating the metrics above.

According to empirical settings (Wu et al., 2022), (Liu et al., 2018b), (Zhang et al., 2017a), (Liu et al., 2019), we adopt F^β as previous works do (Margolin et al., 2014):

$$\mathsf{F}_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{21}$$

and set $\beta^2 = 0.3$ to emphasize the importance of precision, following (Wu et al., 2022), (Liu et al., 2018b), (Zhang et al., 2017a), (Liu et al., 2019), etc.

AUC**. (Borji et al., 2013b)** As SOD is essentially a binary classification task, it is natural that AUC is suitable for this problem. AUC considers both TPR and FPR, and is insensitive to data distribution (Yang et al., 2023). Geometrically, it can be calculated as follows:

$$\mathsf{AUC} = \int_{x=0}^{1} \mathsf{TPR}(\mathsf{FPR}^{-1}(x))dx. \tag{22}$$

It is equivalent to the Wilcoxon test of ranks (Mason & Graham, 2002), and an unbiased estimator of AUC can be expressed as:

$$\mathsf{AUC} = \mathbb{E}_{P_+, P_-}[\ell_{0,1}(f(x^+) - f(x^-))], \tag{23}$$

where $\ell_{0,1}(\cdot)$ denotes the 0-1 loss.

E_m**. (Fan et al., 2018)** It considers the match of global and local similarities simultaneously. It is specially designed for binary map evaluation and has been widely used in recent years. E_m is defined as follows:

$$\mathsf{E}_m = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} \phi_{FM}(x, y), \tag{24}$$

where w and h are the height and width of the image, and

$$\phi_{FM} = f(\xi_{FM}), \tag{25}$$

where $f(\cdot)$ is a convex function. Here we set $f(x) = \frac{1}{2}(1 + x)^2$ as (Fan et al., 2018) suggested. ξ is computed as:

$$\xi_{FM} = \frac{2\phi_{GT} \circ \phi_{GT}}{\phi_{GT} \circ \phi_{GT} + \phi_{FM} \circ \phi_{FM}}, \tag{26}$$

where $\circ$ represents Hadamard production, and $\phi_I = I - \mu_I \cdot \mathbb{A}$, with I as the input, μ_I as the global mean value, and $\mathbb{A}$ an all-ones matrix.

Toward the objectives above, most SOD methods are trained with two types of loss functions: pixel-level loss and region-level loss. The former focuses on pixel-level accuracy, and the latter aims at promoting regional performance.

Pixel-level loss includes binary cross-entropy (BCE), mean square error (MSE), etc. BCE is the most widely used loss function in SOD because it is essentially a binary classification task for each pixel. It is also reasonable to regard it as a regression task with MSE considering that there are few pixels labeled between 0 and 1. Specifically, GateNet (Zhao et al., 2020) and RDSN (Jia & Bruce, 2019) employ MSE as the loss function, while most of the other methods utilize BCE. Specifically, they are defined as follows:

$$\begin{aligned} \mathsf{BCE} &= \frac{1}{N}\sum_{i=1}^{N}[-(p_i \log(g_i) + (1-p_i)\log(1-g_i))], \\ \mathsf{MSE} &= \frac{1}{N}\sum_{i=1}^{N}(p_i - g_i)^2, \end{aligned} \tag{27}$$

where p_i and g_i is the prediction and ground-truth for i-th pixel.

Region-level loss can vary throughout different methods. Some widely used loss functions include DiceLoss (Milletari et al., 2016), and IOULoss (Yu et al., 2016). Both these losses consider the performance in a region, instead of focusing on certain pixels, which can therefore improve the performance from a higher level. DiceLoss is defined as follows:

$$\mathsf{DiceLoss} = 1 - \frac{2 \cdot \sum_i^N p_i \cdot g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}, \tag{28}$$

where the sums run over the N pixels, and p_i, g_i represent the prediction and ground truth, respectively.

IOULoss is computed as follows, which is slightly different from (Yu et al., 2016):

$$\mathsf{IOULoss} = 1 - \frac{\sum_i^N (p_i \cdot g_i)}{\sum_i^N (p_i + d_i) - \sum_i^N (p_i \cdot g_i)}, \tag{29}$$

where p_i and g_i represent the prediction and ground truth.

There are also other region-level losses, such as SSIM (Wang et al., 2004). Generally, they focus on regional detection performance and are therefore robust against imbalanced distribution.

B. SI-F Metric

Similar to SI-MAE, SI-F is expressed as follows according to Eq. (7):

$$\begin{aligned} &\mathsf{SI\text{-}F}(f) \\ &= \frac{1}{N_c}\sum_{i=1}^{N_c} 1 \cdot \frac{2 \cdot \mathsf{TPR}(f_i)}{2\mathsf{TPR}(f_i) + \mathsf{FPR}(f_i) + \mathsf{FNR}(f_i)} \\ &= \frac{1}{N_c}\sum_{i=1}^{N_c} \mathsf{F}(f_i), \end{aligned} \tag{30}$$

where $\mathbb{P}_{X_i}$ is replaced by 1. The definition of object frames is the same as that in SI-MAE. It is worth noting that we do not consider the background frame here and just leave it to SI-MAE because there are no salient pixels and the true positive rate is always 0.

Based on the discussions above, we now define the new metric SI-F as follows:

$$\mathsf{SI\text{-}F}(f) = \frac{1}{K}\sum_{i=1}^{K} \mathsf{F}(f_k^{fore}), \tag{31}$$

where $X_1^{fore}, \cdots, X_k^{fore}$ denote foreground frames.

Also, we give the following proposition to demonstrate the effectiveness of SI-F when there are multiple salient objects, with proof deferred to App. D.2:

Proposition B.1 (Informal)**.** *Given two different predictors f_A and f_B, the following case suggest that* SI-F *is more effective than* F *during evaluation.*

Suppose there are two salient objects ($K = 2$) where f_A and f_B detect the same amount of salient pixels for an image X. Meanwhile, assume that f_A only predicts C_2 perfectly while f_B could somewhat recognize C_1 and C_2 partially. In this case, f_B should still be better than f_A since the latter totally fails on C_1. Unfortunately, $\mathsf{F}(f_A) = \mathsf{F}(f_B)$ holds but we have $\mathsf{SI\text{-}F}(f_A) < \mathsf{SI\text{-}F}(f_B)$.

Remark. Fig. 2 provides a toy example. Discussions above support that F-score is sensitive to the sizes of objects in multi-object cases, yet SI-F can serve our expectations better.

C. SI-AUC Metric

Normally, AUC is defined as

$$\mathsf{AUC} = \int_0^1 \mathsf{TPR}_f(\mathsf{FPR}^{-1}{}_f(t))dt, \tag{32}$$

where f is the predictor, t is the probability threshold, TPR, FPR are the true positive rate and false positive rate, respectively. In the equation above, the integral makes it hard to analyze. Therefore, to further investigate the potential issue for AUC, we adopt another form:

$$\mathsf{AUC}(f) = \sum_{i=1}^{n^+}\sum_{j=1}^{n^-} \frac{\mathbb{I}(f(X_i^+) > f(X_j^-))}{n^+n^-}, \tag{33}$$

where $\mathbb{I}(x) = 1$ when x is True, X_i^+, X_i^- are sampled from salient and non-salient pixels.

Following Eq.(5) in the main text, we let

$$g(f_k) = \sum_{i=1}^{n_k^+}\sum_{j=1}^{n_k^-} \frac{\mathbb{I}(f(X_{k,i}^+) > f(X_{k,j}^-))}{n_k^+n_k^-}, \tag{34}$$

where n_k^+, n_k^- are the number of salient and non-salient pixels within the k-th part, $X_{k,i}^+$, $X_{k,j}^-$ are respectively sampled from salient and non-salient pixels within the k-th part, and $g(f_k)$ is actually the AUC value within the k-th part. Then following the definition of composite functions, we have:

$$\mathsf{AUC}(f) = \sum_{k=1}^{K} g(f_k)\frac{n_k^+n_k^-}{n^+n^-} = \sum_{k=1}^{K} g(f_k)\frac{S_k'}{S'}, \tag{35}$$

where K is the number of foreground frames, $S' = n^+n^-$ and $S_k' = n_k^+n_k^-$, which is also a size-related term. However, it also depends on the distribution of salient pixels, and is therefore more size-robust than MAE, F-score to some extent. To thoroughly eliminate the influence of size, we can define SI-AUC, similar to SI-F:

$$\mathsf{SI\text{-}AUC}(f) = \frac{1}{K}\sum_{k=1}^{K} \mathsf{AUC}(f_k^{fore}). \tag{36}$$

D. Proof for Propositions of Size-Invariant Metrics

D.1. Proof and extension for Proposition 3.3

Restate of Proposition 3.3. Given two different predictors f_A and f_B, the following two possible cases suggest that SI-MAE is more effective than MAE during evaluation.

Case 1: Assume that there is a single salient object (i.e., $K = 1$), with two different results from predictors f_A and f_B. In this case, there is no imbalance from different sizes of objects, and therefore SI-MAE is equivalent to MAE.

Case 2: Suppose there are two salient objects ($K = 2$) where f_A and f_B detect the same amount of salient pixels in an image X. Meanwhile, assume that f_A only predicts C_2 perfectly while f_B could somewhat recognize C_1 and C_2 partially. In this case, f_B should still be better than f_A since the latter totally fails on C_1. Unfortunately, if $S_1^{fore} < S_2^{fore}$, $\mathsf{MAE}(f_A) = \mathsf{MAE}(f_B)$ holds but we have $\mathsf{SI\text{-}MAE}(f_A) > \mathsf{SI\text{-}MAE}(f_B)$.

Proof. For **Case 1**, as there is only one salient object, we can easily divide the image into X_1^{fore} and X_2^{back}, and the weight $\alpha = S_2^{back}/S_1^{fore} = (S - S_1^{fore})/S_1^{fore}$. Therefore, we have the following:

$$\begin{aligned}
\mathsf{MAE}(f) &= \frac{\|f(X_1^{fore}) - Y_1^{fore}\|_{1,1} + \|f(X_2^{back}) - Y_2^{back}\|_{1,1}}{S}, \\
\mathsf{SI\text{-}MAE}(f) &= \frac{1}{1+\alpha} \cdot \left[\frac{\|f(X_1^{fore}) - Y_1^{fore}\|_{1,1}}{S_1^{fore}} + \alpha \cdot \frac{\|f(X_2^{back}) - Y_2^{back}\|_{1,1}}{S_2^{back}} \right] \\
&= \frac{1}{1 + \frac{S - S_1^{fore}}{S_1^{fore}}} \cdot \left[\frac{\|f(X_1^{fore}) - Y_1^{fore}\|_{1,1}}{S_1^{fore}} + \frac{S_2^{back}}{S_1^{fore}} \cdot \frac{\|f(X_2^{back}) - Y_2^{back}\|_{1,1}}{S_2^{back}} \right] \\
&= \frac{S_1^{fore}}{S} \cdot \left[\frac{\|f(X_1^{fore}) - Y_1^{fore}\|_{1,1} + \|f(X_2^{back}) - Y_2^{back}\|_{1,1}}{S_1^{fore}} \right] \\
&= \frac{\|f(X_1^{fore}) - Y_1^{fore}\|_{1,1} + \|f(X_2^{back}) - Y_2^{back}\|_{1,1}}{S} \\
&= \mathsf{MAE}(f).
\end{aligned} \tag{37}$$

For **Case 2**, we first suppose $\rho \in [0, 1]$ of C_2 is correctly detected by f_B. For convenience, we consider errors just in foreground frames. According to the proposition, f_A and f_B detect the same amount of salient pixels, and therefore it is clear that $\mathsf{MAE}(f_A) = \mathsf{MAE}(f_B)$. As there are two salient objects, we have $\alpha = S_3^{back}/(S_1^{fore} + S_2^{fore})$, and thus SI-MAE is calculated as follows:

$$\begin{aligned}
\mathsf{SI\text{-}MAE}(f_A) &= \frac{1}{2+\alpha} \cdot \frac{|C_1|}{S_1}, \\
\mathsf{SI\text{-}MAE}(f_B) &= \frac{1}{2+\alpha} \cdot \left[\frac{|C_2|(1-\rho)}{S_2} + \frac{|C_1| - |C_2|(1-\rho)}{S_1} \right],
\end{aligned} \tag{38}$$

then when $S_1^{fore} < S_2^{fore}$, we have:

$$\begin{aligned}
\mathsf{SI\text{-}MAE}(f_B) - \mathsf{SI\text{-}MAE}(f_A) &= \frac{1}{2+\alpha} \cdot \left[\frac{|C_2|(1-\rho)}{S_2} + \frac{|C_1| - |C_2|(1-\rho)}{S_1} - \frac{|C_1|}{S_1} \right] \\
&= \frac{1}{2+\alpha} \cdot \left[\frac{|C_2|(1-\rho)}{S_2} - \frac{|C_2|(1-\rho)}{S_1} \right] \\
&= \frac{|C_2|(1-\rho)}{2+\alpha} \cdot \left(\frac{1}{S_2} - \frac{1}{S_1} \right) \\
&< 0.
\end{aligned} \tag{39}$$

This completes the proof.

□

Extension of Proposition 3.3. Still, suppose f_A, f_B detect the same amount of salient pixels in an image X. We denote K as the number of salient objects in image X, with $S_1 \le S_2 \le \cdots \le S_K$. Meanwhile, assume f_A only predicts $C_{m+1}, \cdots, C_K$ perfectly while f_B could recognize $C_1, \ldots, C_K$ partially. In this case, f_B should still be better than f_A since the latter totally fails on $C_1, \cdots, C_m$. Unfortunately when S_{m+1} is sufficiently large, which means that larger objects dominate smaller ones, $\mathsf{MAE}(f_A) = \mathsf{MAE}(f_B)$ holds while $\mathsf{SI\text{-}MAE}(f_A) > \mathsf{SI\text{-}MAE}(f_B)$.

Proof. We first suppose $[\rho_1, \rho_2, \cdots, \rho_K], \rho_i \in [0,1]$ of $C_1, C_2, \cdots, C_K$ are correctly detected by f_B. For convenience, we consider errors just in foreground frames. According to the proposition above, f_A and f_B detect the same amount of salient pixels, and therefore $\mathsf{MAE}(f_A) = \mathsf{MAE}(f_B)$. Furthermore, we have the equation below:

$$\sum_{i=1}^{K} \rho_i |C_i| = \sum_{i=m+1}^{K} |C_i|. \tag{40}$$

As there are K salient objects, we have $\alpha = \frac{S_{K+1}^{back}}{\sum_{i=1}^{K} S_i^{fore}}$, and thus $\mathsf{SI\text{-}MAE}$ is calculated as follows:

$$\begin{aligned} \mathsf{SI\text{-}MAE}(f_A) &= \frac{1}{K+\alpha} \sum_{i=1}^{m} \frac{|C_i|}{S_i}, \\ \mathsf{SI\text{-}MAE}(f_B) &= \frac{1}{K+\alpha} \sum_{i=1}^{K} \frac{(1-\rho_i)|C_i|}{S_i}. \end{aligned} \tag{41}$$

For clear representation, we denote $t = (t_1, \cdots, t_K)$ where $t_i = \rho_i |C_i|$ and $S = (S_1, \cdots, S_K)$.

Therefore, when $S_{m+1} > \frac{\langle t, \mathbf{1} \rangle}{\langle t, \frac{1}{S} \rangle}$, we have:

$$\begin{aligned} \mathsf{SI\text{-}MAE}(f_B) - \mathsf{SI\text{-}MAE}(f_A) &= \frac{1}{K+\alpha} \left[\sum_{i=1}^{K} \frac{(1-\rho_i)|C_i|}{S_i} - \sum_{i=1}^{m} \frac{|C_i|}{S_i} \right] \\ &= \frac{1}{K+\alpha} \left[\sum_{i=m+1}^{K} \frac{|C_i|}{S_i} - \sum_{i=1}^{K} \frac{\rho_i |C_i|}{S_i} \right] \\ &\leq \frac{1}{K+\alpha} \left[\frac{\sum_{i=m+1}^{K} |C_i|}{S_{m+1}} - \sum_{i=1}^{K} \frac{\rho_i |C_i|}{S_i} \right] \\ &= \frac{1}{(K+\alpha)} \left[\frac{\sum_{i=1}^{K} \rho_i |C_i|}{S_{m+1}} - \sum_{i=1}^{K} \frac{\rho_i |C_i|}{S_i} \right] \\ &= \frac{1}{(K+\alpha)} \left[\frac{< t, 1 >}{S_{m+1}} - < t, \frac{1}{S} > \right] \\ &< 0. \end{aligned} \tag{42}$$

□

D.2. Proof for Proposition B.1

Restate of Proposition B.1. Given two different predictors f_A and f_B, the following case suggest that $\mathsf{SI\text{-}F}$ is more effective than F-score during evaluation.

Suppose there are two salient objects ($K = 2$) where f_A and f_B detect the same amount of salient pixels for an image X. Meanwhile, assume that f_A only predicts C_2 perfectly while f_B could somewhat recognize C_1 and C_2 partially. In this case, f_B should still be better than f_A since the latter totally fails on C_1. Unfortunately, $\mathsf{F}(f_A) = \mathsf{F}(f_B)$ holds but we have $\mathsf{SI\text{-}F}(f_A) < \mathsf{SI\text{-}F}(f_B)$.

Proof. We first suppose $\rho \in [0,1]$ of C_2 is correctly detected by f_B. For convenience, we consider errors just in foreground frames. According to the proposition, f_A and f_B detect the same amount of salient pixels, and therefore it is clear that $\mathsf{MAE}(f_A) = \mathsf{MAE}(f_B)$. As there are two salient objects, we have $\alpha = S_3^{back} / (S_1^{fore} + S_2^{fore})$, and thus $\mathsf{SI\text{-}F}$ is calculated as follows:

$$\begin{aligned} \mathsf{SI\text{-}F}(f_A) &= \frac{1}{2} \cdot (1+0) = \frac{1}{2}, \\ \mathsf{SI\text{-}F}(f_B) &= \frac{1}{2} \cdot \left[\frac{2\rho}{1+\rho} + \frac{2(1-\rho)|C_2|}{|C_1| + (1-\rho)|C_2|} \right], \end{aligned} \tag{43}$$

Therefore, we have $\mathsf{SI\text{-}F}(f_B) > \mathsf{SI\text{-}F}(f_A)$ because

$$\begin{aligned}\mathsf{SI\text{-}F}(f_B) - \mathsf{SI\text{-}F}(f_A) &= \frac{1}{2}\cdot\left[\frac{2\rho}{1+\rho} + \frac{2(1-\rho)|C_2|}{|C_1|+(1-\rho)|C_2|}\right] - \frac{1}{2} \\ &= \frac{1}{2}\cdot\left[\frac{1-\rho}{1+\rho} + \frac{2(1-\rho)|C_2|}{|C_1|+(1-\rho)|C_2|}\right] \\ &= \frac{1-\rho}{2}\cdot\left[\frac{1}{1+\rho} + \frac{2|C_2|}{|C_1|+(1-\rho)|C_2|}\right] \\ &> 0.\end{aligned} \tag{44}$$

This completes the proof. □

E. Proof for Properties of SI-SOD Loss

E.1. Proof for Proposition 4.1

Restate of Proposition 4.1 Given a separable loss function $\ell(\cdot)$ and its corresponding size-invariant loss $\mathcal{L}_{\mathsf{SI}}(\cdot)$, then:

1. when $S_i < \frac{S}{K+\alpha}$, we have $w_{\mathcal{L}_{\mathsf{SI}}}(x_i) > w_\ell(x_i)$,
2. when $S_i < S_j$, we have $w_{\mathcal{L}_{\mathsf{SI}}}(x_i) > w_{\mathcal{L}_{\mathsf{SI}}}(x_j)$.

where $w_{\mathcal{L}_{\mathsf{SI}}}(x_i)$ is the **weight** of pixel-level loss in X_i with $\mathcal{L}_{\mathsf{SI}}$, and $w_\ell(x_i)$ is the **weight** of pixel-level loss in X_i with the original loss ℓ.

Proof. Common separable loss functions include Cross Entropy(CE), Mean Absolute Error(MAE), Mean Square Error(MSE), etc.

For Item 1, we have the following as mentioned in Sec. 3.1:

$$\ell(f) = \frac{1}{S}\sum_{i=1}^{S}\ell(f(x_i), y_i), \tag{45}$$

where the weight of pixel-level loss $w_\ell(x_i) = 1/S$.

When using the size-invariant loss $\mathcal{L}_{\mathsf{SI}}$, we have:

$$\begin{aligned}\mathcal{L}_{\mathsf{SI}}(f) &= \frac{1}{K+\alpha}\left[\sum_{i=1}^{K}\ell(f_i) + \alpha\ell(f_{K+1})\right] \\ &= \frac{1}{K+\alpha}\left[\sum_{i=1}^{K}\frac{\sum_{j\in X_i}\ell(f(x_j), y_j)}{S_i} + \alpha\frac{\sum_{j\in X_{K+1}}\ell(f(x_j), y_j)}{S_{K+1}}\right],\end{aligned} \tag{46}$$

where the weight of foreground pixel-level loss $w_{\mathcal{L}_{\mathsf{SI}}}(x_i) = \frac{1}{(K+\alpha)S_i}$. Therefore, when $S_i < \frac{S}{K+\alpha}$, we have $w_{\mathcal{L}_{\mathsf{SI}}}(x_i) > w_\ell(x_i)$.

For Item 2, it is obvious that $w_{\mathcal{L}_{\mathsf{SI}}}(x_i)$ is in proportion to $1/S_i$. Therefore, when $S_i < S_j$, we have $w_{\mathcal{L}_{\mathsf{SI}}}(x_i) > w_{\mathcal{L}_{\mathsf{SI}}}(x_j)$. □

E.2. Proof for Theorem 4.2

E.2.1. Proof for Technical Lemmas

In this subsection, we present key lemmas that are important for the proof of Theorem 4.2.

Lemma E.1. *The empirical Rademacher complexity of function g with respect to the predictor f is defined as:*

$$\hat{\mathfrak{R}}_{\mathcal{F}}(g) = \mathbb{E}_\sigma[\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i g(f^{(i)})]. \tag{47}$$

where $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}^K\}$ is a family of predictors, and N refers to the size of the dataset, and σ_is are independent uniform random variables taking values in $\{-1, +1\}$. The random variables σ_i are called Rademacher variables.

Lemma E.2. *Let $\mathbb{E}[g]$ and $\hat{\mathbb{E}}[g]$ represent the expected risk and empirical risk, and $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}^K\}$. Then with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m, the generalization bound holds:*

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}[g(f)] - \hat{\mathbb{E}}[g(f)] \right) \leq 2\hat{\mathfrak{R}}_{\mathcal{F}}(g) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{48}$$

Lemma E.3. *(Foster & Rakhlin, 2019) Assume $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}^K\}$, and $(\sigma_1, \cdots, \sigma_n)$ is a sequence of i.i.d Rademacher random variables. When $\phi_1, \cdots, \phi_n$ are L-Lipschitz with respect to the ℓ_∞ norm, there exists a constant $C > 0$ for any $\delta > 0$, such that if $|\phi_t(f(x))| \vee \|f(x)\|_\infty \leq \beta$, the following holds:*

$$\mathfrak{R}(\phi \circ \mathcal{F}; x_{1:n}) \leq C \cdot L\sqrt{K} \cdot \max_i \mathfrak{R}_n(\mathcal{F}|_9) \cdot \log^{\frac{3}{2}+\delta} \left(\frac{\beta n}{\max_i \mathfrak{R}_n(\mathcal{F}|_i)} \right). \tag{49}$$

where

$$\mathfrak{R}(\phi \circ \mathcal{F}; x_{1:n}) = \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t), \tag{50}$$

where $\mathcal{F}$ is a class of functions $f : \mathcal{X} \to \mathbb{R}$, and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)$ is a sequence of $i.i.d.$ Rademacher random variables. Here we set $\beta = 1$ because it is obvious that $\|f(x)\|_\infty \leq 1$ and $\phi_t(x) \leq 1$. Therefore,

$$\mathfrak{R}(\phi \circ \mathcal{F}; x_{1:n}) \leq C \cdot L\sqrt{K} \cdot \max_i \mathfrak{R}_n(\mathcal{F}|_9) \cdot \log^{\frac{3}{2}+\delta} \left(\frac{n}{\max_i \mathfrak{R}_n(\mathcal{F}|_i)} \right). \tag{51}$$

Lemma E.4. *When $g(\cdot)$ is Lipschitz continuous, the following holds:*

$$\|g(x) - g(\tilde{x})\|_\infty \leq \sup \|\nabla_x g\|_p \cdot \|x - \tilde{x}\|_q, \tag{52}$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

Proof.

$$\begin{aligned} |g(x) - g(\tilde{x})| = & \left| \int_0^1 \langle \nabla g(\tau x + (1-\tau)\tilde{x}), x - \tilde{x} \rangle \, d\tau \right| \\ \leq & \sup_{x \in \mathcal{X}} \left[\|\nabla g\|_p \right] \cdot \left\| x - \tilde{x} \right\|_q \end{aligned} \tag{53}$$

□

Specifically, when $p = 1$ and $q = \infty$, we have

$$\|g(x) - g(\tilde{x})\|_\infty \leq \sup \|\nabla_x g\|_1 \cdot \|x - \tilde{x}\|_\infty. \tag{54}$$

Lemma E.5. *Common composite functions are p-Lipschitz, as (Dembczyński et al., 2017) stated:*

Definition E.6 (p-Lipschitzness)**.** $\Phi(u, v, p)$ is said to be p-Lipschitz if:

$$|\Phi(u, v, p) - \Phi(u', v', p')| \leq U_p|u - u'| + V_p|v - v'| + P_p|p - p'|, \tag{55}$$

where $\Phi(u(h), v(h), p)$ is defined as:

$$u(h) = \text{TP}(h), \qquad v(h) = \mathbb{P}(h = 1), \text{and} \quad p = \mathbb{P}(y = 1). \tag{56}$$

Any metric being a function of the confusion matrix can be parameterized in this way. According to (Dembczyński et al., 2017), Accuracy, AM, F-score, Jaccard, G-Mean, and AUC are all p-Lipschitz.

E.2.2. PROOF FOR THE GENERALIZATION BOUND

Restate of Theorem 4.2 Assume $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}^K\}$, where $K = H \times W$ is the number of pixels in an image, $g^{(i)}$ is the risk over i-th sample, and is L-Lipschitz with respect to the l_∞ norm, (i.e. $\|g(x) - g(\tilde{x})\|_\infty \leq L \cdot \|x - y\|_\infty$). When there are N $i.i.d.$ samples, there exists a constant $C > 0$ for any $\epsilon > 0$, the following generalization bound holds with probability at least $1 - \delta$:

$$\begin{aligned} &\sup_{f \in \mathcal{F}} (\mathbb{E}[g(f)] - \hat{\mathbb{E}}[g(f)]) \\ &\leq C \cdot \frac{L\sqrt{K}}{N} \cdot \max_i \mathfrak{R}_N(\mathcal{F}|_i) \cdot \log^{\frac{3}{2}+\epsilon} \left(\frac{N}{\max_i \mathfrak{R}_N(\mathcal{F}|_i)} \right) \\ &\quad + 3\sqrt{\frac{\log \frac{2}{\delta}}{2N}}, \end{aligned} \tag{57}$$

where again $g(f(X, Y)) := g(f)$, $\mathbb{E}[g(f)]$ and $\hat{\mathbb{E}}[g(f)]$ represent the expected risk and empirical risk, and $\mathfrak{R}_N(\mathcal{F}|_i) = \max_{x_{1:N} \in \mathcal{X}} \mathfrak{R}(\mathcal{F}; x_{1:N})$ denotes the worst-case Rademacher complexity. Specifically,

Case 1: For separable loss functions $\ell(\cdot)$, if it is μ-Lipschitz, we have $L = \mu$.

Case 2: For composite loss functions, when $\ell(\cdot)$ is DiceLoss (Milletari et al., 2016), we have $L = \frac{4}{\rho}$, where $\rho = \min \frac{S_l^{1,i}}{S_l^i}$, which represents the minimum proportion of the salient object in the l-th frame within the i-th sample.

Proof. In this subsection, we give the proof combining lemmas above.

Firstly, the empirical risk over the dataset is:

$$\hat{\mathbb{E}}[g(f)] = \frac{1}{N} \sum_{i=1}^{N} g^{(i)}(f(X), Y), \tag{58}$$

where X, Y are the prediction and ground truth, and

$$g^{(i)}(f(X), Y) = \frac{1}{|N_l^i|} \sum_{l \in [N_c^i]} \ell(f^{(i)}(X_{N_l}), Y_{N_l}^{(i)}), \tag{59}$$

where N_l stands for l-th foreground frame, and therefore $\ell(\cdot)$ is a matrix element function.

Combing Lem. E.1, Lem. E.2 and Lem. E.3, with probability at least $1 - \delta$, we have:

$$\begin{aligned} \sup_{f \in \mathcal{F}} (\mathbb{E}[g(f)] - \hat{\mathbb{E}}[g(f)]) &\leq 2\hat{\mathfrak{R}}_{\mathcal{F}}(g) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2N}} \\ &= 2\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i g(f^{(i)}) \right] + 3\sqrt{\frac{\log \frac{2}{\delta}}{2N}} \\ &\leq C \cdot \frac{L\sqrt{K}}{N} \cdot \max_i \mathfrak{R}_n(\mathcal{F}|_i) \cdot \log^{\frac{3}{2}+\epsilon} \left(\frac{n}{\max_i \mathfrak{R}_n(\mathcal{F}|_i)} \right) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2N}}. \end{aligned} \tag{60}$$

For ***Case 1*** separable loss functions $\ell(\cdot)$, we have the following equation according to definition Eq. (1):

$$g^{(i)}(f) = \frac{1}{N_c^{(i)}} \sum_{l \in [N_c^i]} \frac{1}{|N_l^i|} \sum_{(j,k) \in N_l^i} \ell(f_{j,k}^{(i)}, Y_{j,k}^{(i)}). \tag{61}$$

Assume $\ell(\cdot)$ is μ-Lipschitz, then for Lipschitz continuous of $g^{(i)}$, we have:

$$\begin{aligned} g^{(i)}(f) - g^{(i)}(\tilde{f}) &= \frac{1}{N_c^{(i)}} \sum_{l\in[N_c^i]} \frac{1}{|N_l^i|} \sum_{(j,k)\in N_l^i} \left(\ell(f_{j,k}^{(i)}, Y_{j,k}^{(i)}) - \ell(\tilde{f}_{j,k}^{(i)}, Y_{j,k}^{(i)})\right) \\ &\leq \frac{1}{N_c^{(i)}} \sum_{l\in[N_c^i]} \frac{1}{|N_l^i|} \sum_{(j,k)\in N_l^i} \max_{(j,k)} \left|\left(\ell(f_{j,k}^{(i)}, Y_{j,k}^{(i)}) - \ell(\tilde{f}_{j,k}^{(i)}, Y_{j,k}^{(i)})\right)\right| \\ &\leq \frac{1}{N_c^{(i)}} \sum_{l\in[N_c^i]} \frac{1}{|N_l^i|} \sum_{(j,k)\in N_l^i} \mu \max_{(j,k)} \left|f_{j,k}^{(i)} - \tilde{f}_{j,k}^{(i)}\right| \\ &= \mu \|f^{(i)} - \tilde{f}^{(i)}\|_\infty, \end{aligned} \tag{62}$$

where we use $f_{j,k}^{(i)}$ to denote $f^{(i)}(X_{(j,k)\in N_l^i})$ as abbreviation. We can always bound $\ell(f_{j,k}^{(i)}, Y_{j,k}^{(i)}) - \ell(\tilde{f}_{j,k}^{(i)}, Y_{j,k}^{(i)})$ with the maximum element $\max_{(j,k)} |(\ell(f_{j,k}^{(i)}, Y_{j,k}^{(i)}) - \ell(\tilde{f}_{j,k}^{(i)}, Y_{j,k}^{(i)}))|$ because there are finite pixels in an image. Therefore, for separable loss $g^{(i)}$, we let $L = \mu$, and complete the proof.

For ***Case 2*** composite loss, when $\ell(\cdot)$ is DiceLoss (Milletari et al., 2016), we have the following equation:

$$g^{(i)}(f) = \frac{1}{N_c^i} \sum_{l=1}^{N_c^i} \left[1 - \frac{2\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)} \cdot f_{j,k}^{(i)}}{\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)} + \sum_{(j,k)\in N_l^i} f_{j,k}^{(i)}}\right]. \tag{63}$$

Considering that formally DiceLoss $= 1 - \mathsf{F}$, combining Lem. E.4 and Lem. E.5, we turn to solve $\sup \|\nabla_x g\|_1 \cdot \|$ instead of directly pursuing the Lipschitz constant with respect to ℓ_∞ norm. Therefore, we can find the Lipschitz continuous of $g^{(i)}$:

$$\begin{aligned} \left\|\frac{\partial g^{(i)}}{\partial f_{j,k}^{(i)}}\right\|_1 &= 2 \cdot \left| \frac{Y_{j,k}^{(i)} \cdot \left(\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)} + \sum_{(j,k)\in N_l^i} f_{j,k}^{(i)}\right) - \sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)} \cdot f_{j,k}^{(i)}}{\left(\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)} + \sum_{(j,k)\in N_l^i} f_{j,k}^{(i)}\right)^2} \right| \\ &\leq 2 \cdot \left(\left| \frac{Y_{j,k}^{(i)}}{\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)} + \sum_{(j,k)\in N_l^i} f_{j,k}^{(i)}} \right| + \left| \frac{\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)} \cdot f_{j,k}^{(i)}}{(\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)} + \sum_{(j,k)\in N_l^i} f_{j,k}^{(i)})^2} \right| \right) \\ &\leq 2 \cdot \left(\frac{1}{\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)} + \sum_{(j,k)\in N_l^i} f_{j,k}^{(i)}} + \frac{\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)}}{\left(\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)} + \sum_{(j,k)\in N_l^i} f_{j,k}^{(i)}\right)^2} \right) \\ &\leq 2 \cdot \left(\frac{1}{\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)}} + \frac{\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)}}{\left(\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)}\right)^2} \right) \\ &= \frac{4}{\sum_{(j,k)\in N_l^i} Y_{j,k}^{(i)}} \\ &= \frac{4}{S_l^{1,i}}, \end{aligned} \tag{64}$$

where $S_l^{1,i}$ stands for the area of the object in i-th frame, and S_l^i stands for the area of the i-th frame. Therefore, for a frame

in the image,

$$\begin{aligned}\|\nabla g^{(i)}\|_1 &= \frac{1}{N_c^i}\sum_{l=1}^{N_c^i}\sum_{(j,k)\in N_l^i}|\nabla g_{j,k}^{(i)}| \\ &= \frac{1}{N_c^i}\sum_{l=1}^{N_c^i}\left\|\frac{\partial g^{(i)}}{\partial f_{j,k}^{(i)}}\right\|_1 \cdot S_l^i \\ &\le \frac{1}{N_c^i}\sum_{l=1}^{N_c^i} 4\cdot\frac{S_l^i}{S_l^{1,i}} \\ &\le \frac{4}{\rho},\end{aligned} \tag{65}$$

where $0 < \rho \le \frac{S_l^{1,i}}{S_l^i}$, which depicts the threshold, how much proportion the object occupies in the corresponding frame. Therefore, taking DiceLoss as an example of composite loss, we let $L = \frac{4}{\rho}$, and complete the proof. □

F. Additional Experiment Settings

In this section, we make a supplementation to Sec. 5.

F.1. Datasets

Here we give more detailed introductions to the five datasets used in the experiments, as shown in Tab. 2.

Dataset	Scale	Characteristics
DUTS (Wang et al., 2017)	10,553 + 5,019	Training set (10,553), as well as test set (5,019), is provided.
ECSSD (Yan et al., 2013)	1,000	Semantically meaningful but structurally complex contents are included.
DUT-OMRON (Yang et al., 2013)	5,168	It is characterized by complex background and diverse contents.
HKU-IS (Li & Yu, 2015)	4,447	Far more multiple disconnected objects are included.
MSOD (Deng et al., 2023)	300	It consists of the most challenging multi-object scenarios with 1342 objects in total.
PASCAL-S (Yan et al., 2013)	850	Images are from PASCAL VOC 2010 validation set with multiple salient objects.
SOD (Movahedi & Elder, 2010)	300	Many images have more than one salient object that is similar to the background.
XPIE (Xia et al., 2017)	10,000	It covers many complex scenes with different numbers, sizes and positions of salient objects.

Table 2. Statistics on Datasets.

DUTS is a widely used large-scale dataset, consisting of the training set DUTS-TR including 10,553 images, and the test set DUTS-TE including 5,019 images. All the images are sampled from the ImageNet DET training and test set (Deng et al., 2009), and some test images are also collected from the SUN data set (Xiao et al., 2010). It is common practice that SOD models are trained on DUTS-TR and tested on other datasets.

ECSSD, namely **E**xtended **C**omplex **S**cene **S**aliency **D**ataset, consists of 1,000 images with complex scenes, presenting textures and structures. One of the characteristics is that this dataset includes many semantically meaningful but structurally complex images. The images are acquired from the internet and 5 helpers are asked to produce the ground truth masks individually.

DUT-OMRON is composed of 5,168 images with complex backgrounds and diverse content. Images in this dataset have one or more salient objects and a relatively complex background.

HKU-IS has 4,447 images with relatively more multi-object scenarios. Particularly, around 50% images in this dataset have multiple disconnected salient objects, far beyond the three datasets above.

MSOD contains the most challenging multi-object scenes across the common SOD datasets. It comprises 300 test images with 1342 total objects. The dataset comprises a variety of object classes and a varied number of these objects across the image.

PASCAL-S is a dataset for salient object detection consisting of a set of 850 images from PASCAL VOC 2010 validation set with multiple salient objects on the scenes.

SOD consists of 300 images, constructed from (Martin et al., 2002). Many images have more than one salient object that is similar to the background or touches image boundaries.

XPIE contains 10,000 images with pixel-wise masks of salient objects. It covers many complex scenes with different numbers, sizes and positions of salient objects.

F.2. Competitors

Here we give a more detailed summary of the backbones mentioned in the experiments.

PoolNet (Liu et al., 2019) is a widely used baseline. They design various pooling-based modules for the first time to assist in improving the performance of SOD. Specifically, the model consists of two primary modules based on the feature pyramid networks, namely a global guidance module (GGM), and a feature aggregation module (FAM). GGM is an individual module, where high-level semantic information can be transmitted to all pyramid layers, and FAM aims at capturing local context information at different scales and then combining them with different weights.

LDF (Wei et al., 2020) contains a level decoupling procedure and a feature interaction network. It decomposes the saliency label into the body map and detail map to supervise the model. The feature interaction network is introduced to make full use of the complementary information between branches. Both branches iteratively exchange information to produce more precise saliency maps.

GateNet (Zhao et al., 2020) proposes a gated network to adaptively control the amount of information flowing into the decoder. The multilevel gate units help to balance the contribution of each encoder and suppress the information in non-salient regions. An ASPP module is exploited to capture richer context information. With a dual-branch architecture, it forms a residual structure, which complements each other to generate better results.

ICON (Zhuge et al., 2022) proposes micro-integrity and macro-integrity, aiming to focus on whole-part relevance within a single salient object and to identify all salient objects within the given image scene. It is composed of three parts: diverse feature aggregation, integrity channel enhancement, and part-whole verification.

EDN (Wu et al., 2022) directly utilizes an extreme down-sampling technique to capture effective high-level features for SOD and achieved competitive performance with high computational efficiency. The proposed Extremely Downsampled Block is to learn a global view of the whole image. It only introduces a tiny computational overhead but achieves competitive performance with a fast inference speed.

F.3. Implementation Details

The experiment platform is Ubuntu 18.04.5 LTS with Intel(R) Xeon(R) Gold 6246R CPU @ 3.40GHz. We implement our method with Python 3.8.13 and torch 2.0.1. For specific backbones, we report the settings as follows:

- For EDN, following (Wu et al., 2022), Adam optimizer with parameters $\beta_1 = 0.9$, and $\beta_2 = 0.99$ is adopted, with weight decay 10^{-4} and batch size 36. The initial learning rate is set to 5×10^{-5} with a poly learning rate strategy and the training lasts for 100 epochs in total.

- For ICON, we use the SGD optimizer, with the initial learning rate as 10^{-2}, the weight decay as 10^{-4}, and the momentum as 0.9. The batch size is set to 36, and the training lasts 100 epochs.

- For GateNet (Zhao et al., 2020), we fully follow (Zhao et al., 2020) to utilize the SGD optimizer, with the initial learning rate as 10^{-3}, momentum as 0.9 and the weight decay as 5×10^{-4}. The batch size is set to 12, and the network is iterated within 10^5 times.

- For LDF, as it is a two-stage framework, we integrate our method into the second stage, as most previous works do. Specifically, we use the SGD optimizer with the initial learning rate as 5×10^{-2}, momentum as 0.9 and the weight decay as 5×10^{-4}. The batch size is set to 32 and the training in the second stage lasts 40 epochs.

- For PoolNet, we fully follow (Liu et al., 2019) to utilize the Adam optimizer with the initial learning rate as 5×10^{-5}, and the weight decay as 5×10^{-4}. The batch size to 1, and the network is iterated every 10 gradients are accumulated. The training lasts 24 epochs.

F.4. Optimization Details for Different Backbones

Following Eq. (16), here we give a detailed description of the implementation of optimization for different backbones.

- For EDN, the original loss function is
$$\mathcal{L} = \mathsf{BCE}(f) + \mathsf{DiceLoss}(f), \tag{66}$$
and we modify it into
$$\mathcal{L} = \sum_{k=1}^{K}[\mathsf{BCE}(f_k^{fore}) + \mathsf{DiceLoss}(f_k^{fore})] + \alpha\mathsf{BCE}(f_{K+1}^{back}). \tag{67}$$
Specifically, DiceLoss is defined as:
$$\mathsf{DiceLoss}(\cdot) = \frac{2\cdot\sum_i^N p_i\cdot g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}, \tag{68}$$
where the sums run over the N pixels, and p_i, g_i represent the prediction and ground truth, respectively.

- For ICON, the original loss function is
$$\mathcal{L} = \mathsf{BCE}(f) + \mathsf{IOULoss}(f), \tag{69}$$
and we modify it into
$$\mathcal{L} = \sum_{k=1}^{K}[\mathsf{BCE}(f_k^{fore}) + \mathsf{IOULoss}(f_k^{fore})] + \alpha\mathsf{BCE}(f_{K+1}^{back}), \tag{70}$$
where IOULoss is defined as
$$\mathsf{IOULoss}(\cdot) = 1 - \frac{\sum_i^N (p_i\cdot g_i)}{\sum_i^N (p_i + d_i) - \sum_i^N (p_i\cdot g_i)}, \tag{71}$$
where p_i and g_i represent the prediction and ground truth.

- For GateNet, the original loss function is
$$\mathcal{L} = \ell(f), \tag{72}$$
where $\ell(\cdot)$ is anyone among the binary cross entropy, mean square error, and L1 loss. We modify it into
$$\mathcal{L} = \sum_{k=1}^{K}\ell(f_k^{fore}) + \alpha\ell(f_{K+1}^{back}). \tag{73}$$

- For LDF, the original loss function is the same as that in ICON. Therefore, we adopt the same modification.
- For PoolNet, the original loss function is the same as that in GateNet. Therefore, we adopt the same modification.

G. Additional Experiment Analysis

G.1. Quantitative comparisons

Here we display the quantitative results on other datasets.

Table 3. Quantitative comparisons on ECSSD, DUT-OMRON, and HKU-IS. The better results are shown with **bold**, and darker color indicates superior results. Metrics with ↑ mean higher value represents better performance, while ↓ mean lower value represents better performance.

Dataset	Methods	MAE ↓	SI-MAE ↓	AUC ↑	SI-AUC ↑	F_m^β ↑	SI-F_m^β ↑	F_{max}^β ↑	SI-F_{max}^β ↑	E_m ↑
ECSSD	PoolNet	0.0632	0.0467	0.9785	0.9817	0.8453	0.8630	0.9205	0.9309	0.8813
	+ Ours	**0.0575**	**0.0421**	**0.9839**	**0.9893**	**0.8588**	**0.8907**	**0.9287**	**0.9477**	**0.8989**
	LDF	**0.0450**	**0.0336**	0.9618	0.9774	**0.8949**	0.8984	0.9366	0.9421	0.9153
	+ Ours	0.0476	0.0353	**0.9799**	**0.9769**	0.8920	**0.9072**	**0.9404**	**0.9522**	**0.9236**
	ICON	**0.0395**	**0.0300**	0.9677	0.9746	**0.9036**	0.9084	**0.9407**	0.9452	0.9279
	+ Ours	0.0436	0.0328	**0.9756**	**0.9774**	0.8905	**0.9096**	0.9338	**0.9481**	**0.9281**
	GateNet	0.0378	0.0288	0.9773	0.9773	**0.9098**	0.9199	**0.9463**	**0.9520**	0.9358
	+ Ours	**0.0362**	**0.0274**	**0.9784**	**0.9853**	0.9075	**0.9253**	0.9423	0.9489	**0.9387**
	EDN	0.0363	0.0271	**0.9767**	**0.9754**	**0.9089**	0.9147	**0.9531**	**0.9560**	0.9338
	+ Ours	**0.0358**	**0.0269**	0.9762	0.9778	0.9084	**0.9216**	0.9456	0.9543	**0.9375**
DUT-OMRON	PoolNet	0.0727	0.0625	0.9403	0.9595	0.6757	0.7463	0.7919	0.8738	0.7807
	+ Ours	**0.0699**	**0.0585**	**0.9457**	**0.9687**	**0.6979**	**0.7847**	**0.8093**	**0.8927**	**0.8123**
	LDF	**0.0540**	**0.0464**	0.8834	0.9452	0.7154	0.7416	0.7918	0.8413	0.8080
	+ Ours	0.0584	0.0499	**0.9327**	**0.9535**	**0.7156**	**0.7813**	**0.7982**	**0.8944**	**0.8299**
	ICON	**0.0601**	**0.0525**	0.9160	0.9296	**0.7450**	0.7940	**0.8092**	0.8584	**0.8456**
	+ Ours	0.0646	0.0554	**0.9358**	**0.9423**	0.7098	**0.8108**	0.7874	**0.8798**	0.8342
	GateNet	**0.0548**	**0.0475**	0.9246	0.9246	**0.7403**	0.8031	**0.8116**	0.8704	0.8464
	+ Ours	0.0580	0.0501	**0.9329**	**0.9293**	0.7363	**0.8300**	0.8049	**0.8852**	**0.8510**
	EDN	**0.0551**	0.0484	0.9292	0.9407	**0.7529**	0.8224	**0.8117**	0.8798	s
	+ Ours	0.0557	**0.0483**	**0.9382**	**0.9507**	0.7544	**0.8381**	0.8163	**0.8912**	**0.8594**
HKU-IS	PoolNet	0.0526	0.0537	0.9804	0.9842	0.8294	0.8316	0.9066	0.9199	0.8816
	+ Ours	**0.0464**	**0.0440**	**0.9847**	**0.9891**	**0.8501**	**0.8770**	**0.9188**	**0.9395**	**0.9081**
	LDF	**0.0333**	0.0355	0.9575	0.9544	**0.8868**	0.8759	0.9263	0.9217	0.9234
	+ Ours	0.0346	**0.0344**	**0.9810**	**0.9839**	0.8815	**0.8910**	**0.9306**	**0.9437**	**0.9320**
	ICON	**0.0346**	0.0374	0.9616	0.9641	**0.8854**	0.8722	0.9232	0.9151	0.9277
	+ Ours	0.0357	**0.0361**	**0.9815**	**0.9787**	0.8788	**0.8898**	**0.9260**	**0.9369**	**0.9315**
	GateNet	0.0326	0.0338	0.9762	0.9830	0.8961	0.8993	**0.9346**	0.9385	0.9394
	+ Ours	**0.0292**	**0.0293**	**0.9785**	0.9779	**0.8995**	**0.9122**	0.9345	**0.9441**	**0.9443**
	EDN	**0.0279**	0.0294	0.9750	0.9738	**0.9004**	0.9017	**0.9417**	0.9364	0.9429
	+ Ours	0.0287	**0.0289**	**0.9776**	**0.9780**	0.8986	**0.9072**	0.9375	**0.9443**	**0.9442**

Table 4. Quantitative comparisons on SOD, PASCAL-S, and XPIE. The better results are shown with **bold**, and darker color indicates superior results. Metrics with ↑ mean higher value represents better performance, while ↓ mean lower value represents better performance.

Dataset	Methods	MAE ↓	SI-MAE ↓	AUC ↑	SI-AUC ↑	F_m^β ↑	SI-F_m^β ↑	F_{max}^β ↑	SI-F_{max}^β ↑	E_m ↑
SOD	PoolNet	0.1353	0.1219	0.9000	0.9039	0.6974	0.6685	0.8356	0.8370	0.7208
	+Ours	**0.1235**	**0.1097**	**0.9143**	**0.9251**	**0.7365**	**0.7245**	**0.8437**	**0.8657**	**0.7654**
	LDF	**0.0940**	**0.0884**	0.8934	0.8839	**0.7991**	**0.7584**	**0.8706**	0.8391	**0.8091**
	+Ours	0.1093	0.1009	**0.9058**	**0.9102**	0.7755	0.7404	0.8593	**0.8684**	0.7858
	ICON	0.1058	0.0995	0.8785	0.8661	0.7931	0.7461	**0.8567**	0.8182	0.7984
	+Ours	**0.0987**	**0.0920**	**0.9083**	**0.9088**	**0.7992**	**0.7759**	0.8557	**0.8513**	**0.8220**
	GateNet	0.0987	0.0949	0.8834	0.8727	0.7928	0.7514	0.8602	0.8274	0.7949
	+Ours	**0.0965**	**0.0910**	**0.8935**	**0.8880**	**0.8094**	**0.7792**	**0.8626**	**0.8483**	**0.8190**
	EDN	0.1093	0.1009	0.8823	0.8716	0.7795	0.7313	**0.8723**	0.8379	0.7873
	+Ours	**0.0982**	**0.0922**	**0.8892**	**0.8818**	**0.8110**	**0.7728**	0.8677	**0.8382**	**0.8207**
PASCAL-S	PoolNet	0.0944	0.0716	0.9462	0.9612	0.7556	0.8495	0.7932	0.8915	0.8165
	+Ours	**0.0919**	**0.0690**	**0.9535**	**0.9722**	**0.7651**	**0.8520**	**0.8346**	**0.9103**	**0.8359**
	LDF	**0.0662**	**0.0502**	0.9437	0.9536	**0.8117**	0.8504	**0.8759**	0.9106	**0.8742**
	+Ours	0.0705	0.0532	**0.9492**	**0.9626**	0.8075	**0.8509**	0.8710	**0.9150**	0.8716
	ICON	**0.0735**	**0.0565**	0.9308	0.9391	**0.8142**	0.8406	**0.8642**	0.8908	**0.8682**
	+Ours	0.0791	0.0599	**0.9456**	**0.9603**	0.7946	**0.8521**	0.8554	**0.9028**	0.8641
	GateNet	**0.0622**	**0.0473**	0.9474	0.9554	**0.8298**	0.8707	**0.8794**	0.9121	**0.8882**
	+Ours	0.0665	0.0504	**0.9478**	**0.9582**	0.8217	0.8750	0.8724	**0.9138**	0.8839
	EDN	0.0649	0.0494	0.9419	0.9506	0.8207	0.8431	**0.8841**	0.9086	0.8750
	+Ours	**0.0644**	**0.0491**	**0.9456**	**0.9551**	**0.8260**	**0.8684**	0.8757	**0.9114**	**0.8859**
XPIE	PoolNet	0.0622	0.0505	0.9667	0.9771	0.7904	0.8242	0.8710	0.9103	0.8494
	+Ours	**0.0599**	**0.0476**	**0.9733**	**0.9857**	**0.8042**	**0.8662**	**0.8786**	**0.9308**	**0.8738**
	LDF	**0.0428**	0.0347	0.9641	0.9700	**0.8520**	**0.8844**	**0.9015**	0.9324	**0.9054**
	+Ours	0.0458	**0.0372**	**0.9701**	**0.9785**	0.8467	0.8824	0.8965	**0.9360**	0.9013
	ICON	**0.0459**	**0.0381**	0.9468	0.9503	**0.8514**	0.8699	**0.8903**	0.9090	**0.8994**
	+Ours	0.0498	0.0405	**0.9662**	**0.9740**	0.8359	**0.8774**	0.8847	**0.9204**	0.8988
	GateNet	**0.0414**	**0.0339**	0.9615	0.9656	**0.8614**	0.8924	**0.9024**	0.9296	0.9107
	+Ours	0.0429	0.0347	**0.9649**	**0.9713**	0.8563	**0.9046**	0.8969	**0.9356**	**0.9125**
	EDN	**0.0409**	0.0337	0.9598	0.9642	0.8584	0.8793	**0.9044**	0.9256	0.9043
	+Ours	0.0416	**0.0337**	**0.9640**	**0.9707**	**0.8590**	**0.8986**	0.8959	**0.9317**	**0.9132**

G.2. Qualitative comparisons

Here we present some visualization examples on the effect of our size-invariant loss.

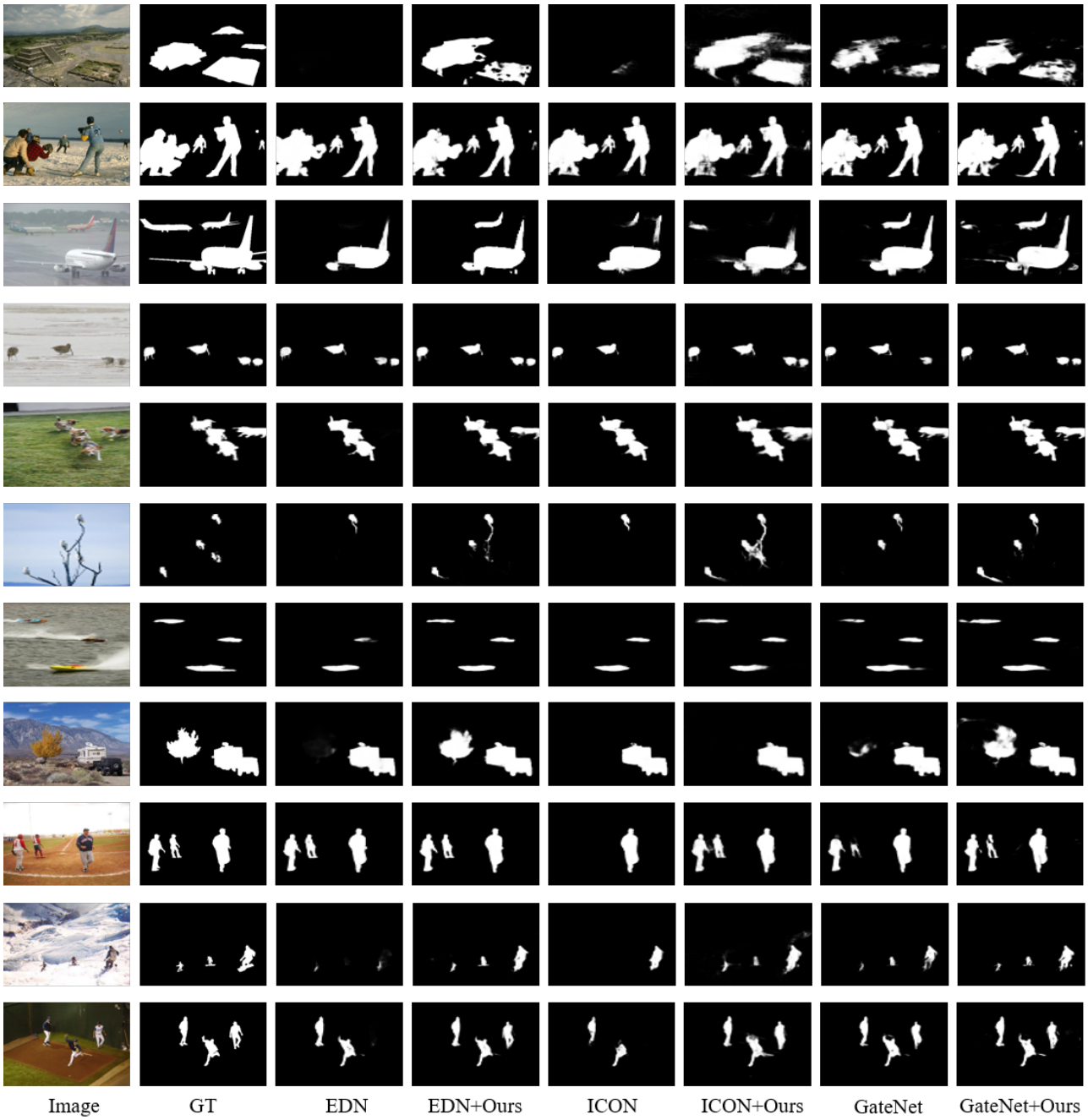
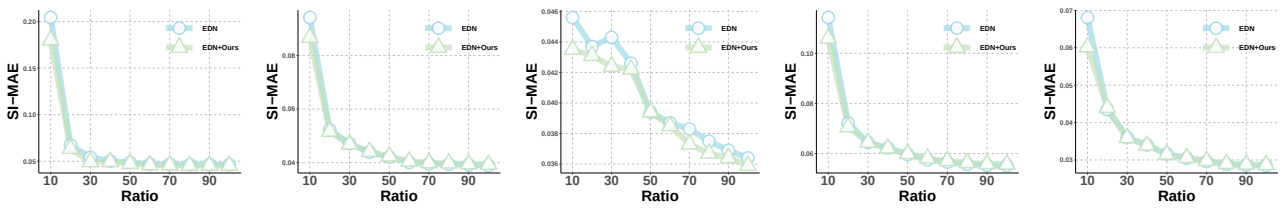


Figure 8. Qualitative comparison on different backbones.

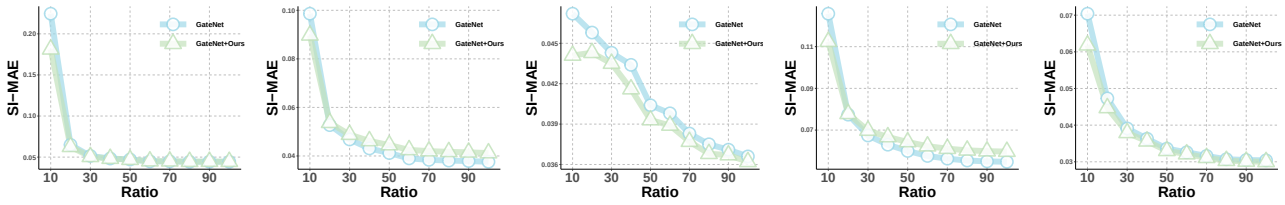
G.3. Performance with Respect to Sizes

Here we expand the size-relevant fine-grained analysis to other backbones and benchmarks.



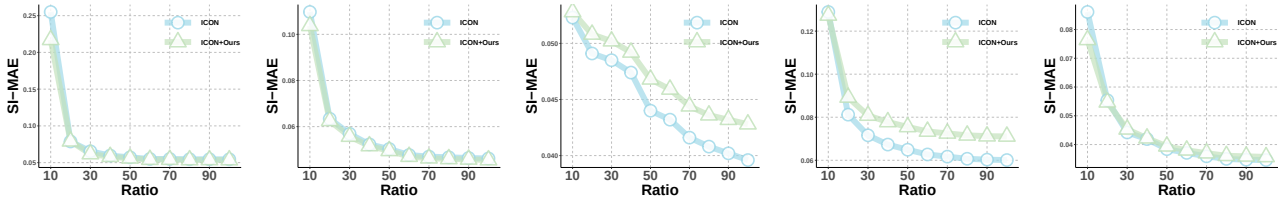
(a) MSOD (b) DUTS (c) ECSSD (d) DUT-OMRON (e) HKU-IS

Figure 9. SI-MAE performance on objects with different sizes on five datasets, with EDN as the backbone.



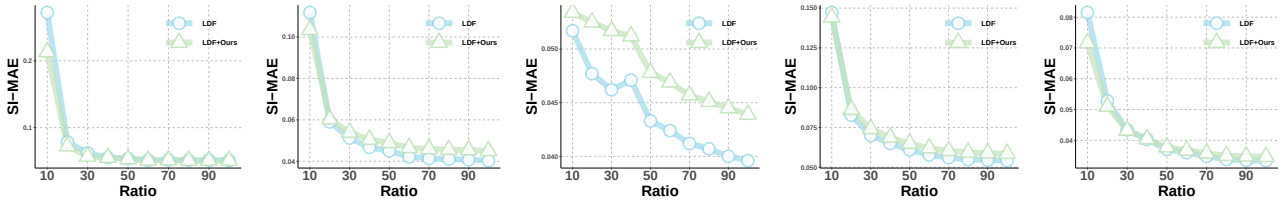
(a) MSOD (b) DUTS (c) ECSSD (d) DUT-OMRON (e) HKU-IS

Figure 10. SI-MAE performance on objects with different sizes on five datasets, with GateNet as the backbone.



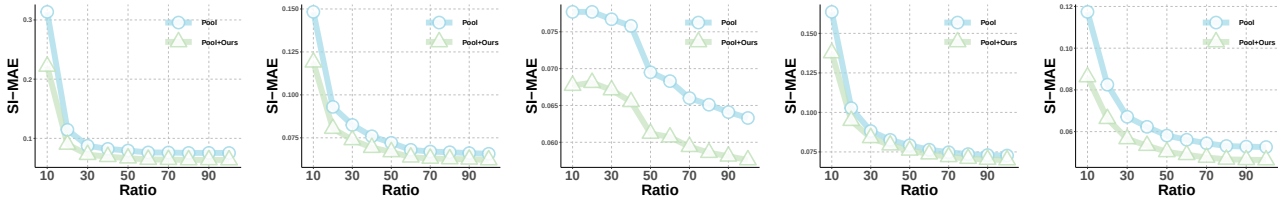
(a) MSOD (b) DUTS (c) ECSSD (d) DUT-OMRON (e) HKU-IS

Figure 11. SI-MAE performance on objects with different sizes on five datasets, with ICON as the backbone.



(a) MSOD (b) DUTS (c) ECSSD (d) DUT-OMRON (e) HKU-IS

Figure 12. SI-MAE performance on objects with different sizes on five datasets, with LDF as the backbone.

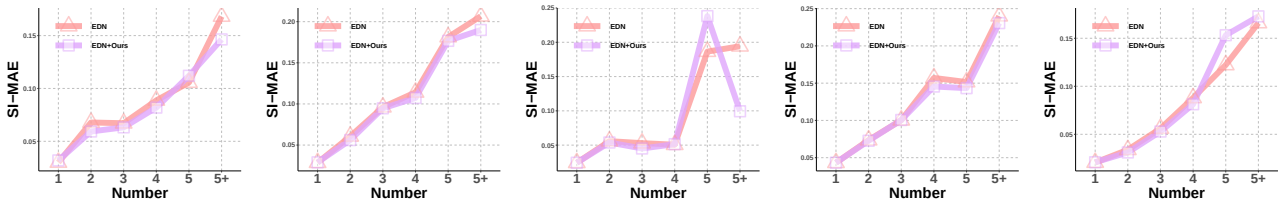


(a) MSOD (b) DUTS (c) ECSSD (d) DUT-OMRON (e) HKU-IS

Figure 13. SI-MAE performance on objects with different sizes on five datasets, with PoolNet as the backbone.

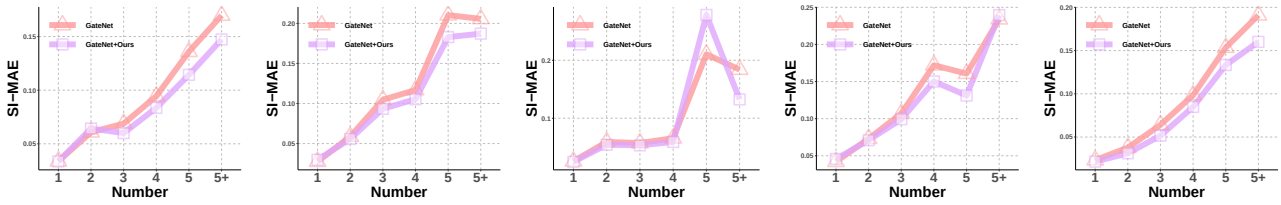
G.4. Performance with Respect to Object Numbers

Here we expand the number-relevant fine-grained analysis to other backbones and benchmarks.



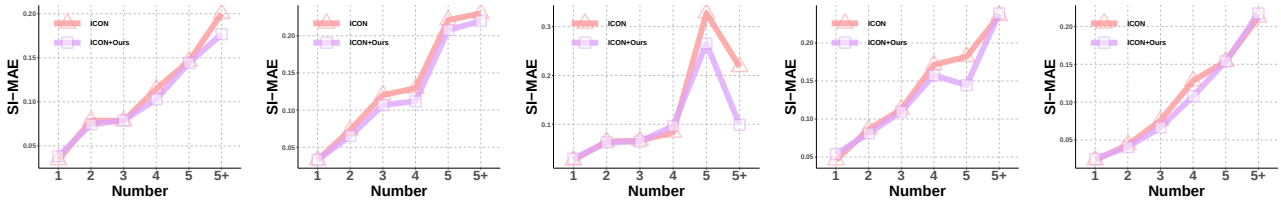
(a) MSOD (b) DUTS (c) ECSSD (d) DUT-OMRON (e) HKU-IS

Figure 14. SI-MAE performance on objects with different object numbers on five datasets, with EDN as the backbone.



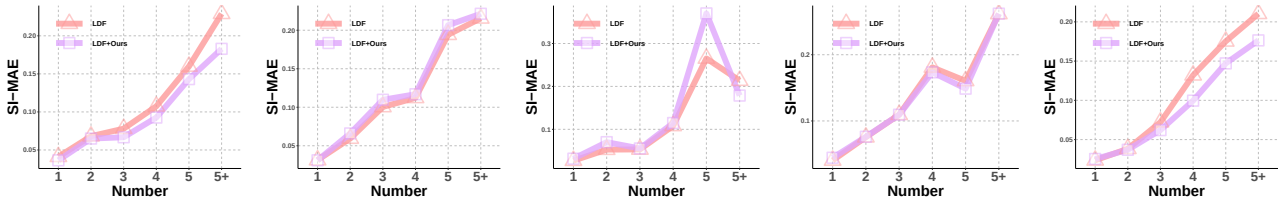
(a) MSOD (b) DUTS (c) ECSSD (d) DUT-OMRON (e) HKU-IS

Figure 15. SI-MAE performance on objects with different object numbers on five datasets, with GateNet as the backbone.



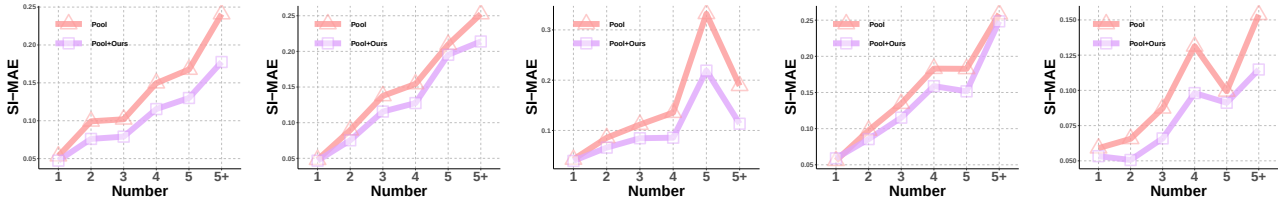
(a) MSOD (b) DUTS (c) ECSSD (d) DUT-OMRON (e) HKU-IS

Figure 16. SI-MAE performance on objects with different object numbers on five datasets, with ICON as the backbone.



(a) MSOD (b) DUTS (c) ECSSD (d) DUT-OMRON (e) HKU-IS

Figure 17. SI-MAE performance on objects with different object numbers on five datasets, with LDF as the backbone.



(a) MSOD (b) DUTS (c) ECSSD (d) DUT-OMRON (e) HKU-IS

Figure 18. SI-MAE performance on objects with different object numbers on five datasets, with PoolNet as the backbone.

G.5. Ablation Studies

Here we display the detailed results of the ablation studies in Sec. 5.3.3.

Table 5. Ablation on the parameter α on MSOD(300 images). The best results are marked in bold.

Methods	α	MAE ↓	SI-MAE ↓	AUC ↑	F_m^β ↑	SI-F_m^β ↑	F_{max} ↑	SI-F_{max} ↑	E_m ↑
EDN (A)	ResNet50	0.0467	0.0788	0.9196	0.7925	0.7635	0.8410	0.8321	0.8712
+ ***Ours*** (B)	0	0.2340	0.2000	0.9217	0.4790	**0.8547**	0.7695	**0.9136**	0.6199
+ ***Ours*** (C)	1	0.0544	0.0812	0.9360	0.7893	0.7850	0.8502	0.8912	0.8825
+ ***Ours*** (D)	$\frac{S^{fore}}{S^{back}}$	**0.0453**	**0.0724**	**0.9401**	**0.8057**	0.7990	**0.8555**	0.8619	**0.8936**

Table 6. Ablation on the parameter α on DUTS(5,019 images). The best results are marked in bold.

Methods	α	MAE ↓	SI-MAE ↓	AUC ↑	F_m^β ↑	SI-F_m^β ↑	F_{max} ↑	SI-F_{max} ↑	E_m ↑
EDN (A)	ResNet50	**0.0389**	0.0388	0.9600	**0.8288**	0.8565	0.8752	0.9017	0.9033
+ ***Ours*** (B)	0	0.2318	0.1975	0.9354	0.4621	**0.8730**	0.7705	**0.9235**	0.6069
+ ***Ours*** (C)	1	0.0489	0.0460	0.9634	0.8146	0.8585	0.8807	0.9182	0.8954
+ ***Ours*** (D)	$\frac{S^{fore}}{S^{back}}$	0.0392	**0.0381**	**0.9658**	0.8260	0.8672	**0.8765**	0.9119	**0.9072**

Table 7. Ablation on the parameter α on ECSSD(1,000 images). The best results are marked in bold.

Methods	α	MAE ↓	SI-MAE ↓	AUC ↑	F_m^β ↑	SI-F_m^β ↑	F_{max} ↑	SI-F_{max} ↑	E_m ↑
EDN (A)	ResNet50	0.0363	0.0271	**0.9767**	**0.9089**	0.9147	**0.9531**	0.9560	0.9338
+ ***Ours*** (B)	0	0.1656	0.1282	0.9633	0.6557	**0.9236**	0.9043	**0.9587**	0.7431
+ ***Ours*** (C)	1	0.0454	0.0340	0.9740	0.8986	0.9164	0.9457	0.9556	0.9282
+ ***Ours*** (D)	$\frac{S^{fore}}{S^{back}}$	**0.0358**	**0.0269**	0.9762	0.9084	0.9216	0.9456	0.9543	**0.9375**

Table 8. Ablation on the parameter α on DUT-OMRON(5,168 images). The best results are marked in bold.

Methods	α	MAE ↓	SI-MAE ↓	AUC ↑	F_m^β ↑	SI-F_m^β ↑	F_{max} ↑	SI-F_{max} ↑	E_m ↑
EDN (A)	ResNet50	**0.0514**	0.0484	0.9292	0.7529	0.8224	0.8117	0.8798	0.8514
+ ***Ours*** (B)	0	0.2693	0.2305	0.9098	0.4231	**0.8708**	0.7104	**0.9231**	0.564
+ ***Ours*** (C)	1	0.0642	0.0555	0.9284	0.7442	0.8239	**0.8190**	0.9048	0.8508
+ ***Ours*** (D)	$\frac{S^{fore}}{S^{back}}$	0.0557	**0.0483**	**0.9382**	**0.7544**	0.8381	0.8163	0.8912	**0.8594**

Table 9. Ablation on the parameter α on HKU-IS(4,447 images). The best results are marked in bold.

Methods	α	MAE ↓	SI-MAE ↓	AUC ↑	F_m^β ↑	SI-F_m^β ↑	F_{max} ↑	SI-F_{max} ↑	E_m ↑
EDN (A)	ResNet50	**0.0279**	0.0294	0.9750	**0.9004**	0.9017	**0.9417**	0.9364	0.9429
+ ***Ours*** (B)	0	0.1822	0.1431	0.9613	0.6027	**0.9132**	0.8921	**0.9541**	0.7100
+ ***Ours*** (C)	1	0.0383	0.0364	0.9760	0.888	0.9007	0.9377	0.9478	0.9347
+ ***Ours*** (D)	$\frac{S^{fore}}{S^{back}}$	0.0287	**0.0289**	**0.9776**	0.8986	0.9072	0.9375	0.9443	**0.9442**

G.6. Time Cost Comparison

Size-invariant optimization generally modifies the computation process of the original loss functions without bringing in too much computational burden. According to Eq. (16):

$$\mathcal{L}_{\mathsf{SI}}(f) = \sum_{k=1}^{K} \ell(f_k^{fore}) + \alpha \ell(f_{K+1}^{back}). \tag{74}$$

If the time cost of the original loss functions is $\mathcal{O}(t)$, then the theoretical time cost for size-invariant optimization will almost be $\mathcal{O}((K+1)\bar{t})$, where $\mathcal{O}(\bar{t}$ is the average time cost of processing one frame. Particularly, as it is common practice to train the SOD model on dataset DUTS-TR, we report the average $\bar{K}$ over the dataset: $\bar{K} \approx 1.21$. Also, we report the practical training time of different backbones when applying our method.

Table 10. Practical training cost per epoch. The results are displayed as mean $\pm$ std., with 'seconds' as the unit.

Backbone	Original optimization	SI optimization
EDN	340.0 $\pm$ 3.3s	543.8$\pm$0.7s
PoolNet	523.5$\pm$1.1s	690.2$\pm$1.5s
ICON	162.0$\pm$0.5s	340.1$\pm$0.1s
GateNet	561.2$\pm$0.8s	1270.2$\pm$35.3s
LDF	109.2$\pm$0.6s	244.5$\pm$1.4s

It is noteworthy that the time cost of calculating the connected components of the image is not included in the training process. All the calculations can be completed during the data pre-process. The pre-process mainly consists of two stages:

(a) Calculating the connected components of the image.

(b) Generating the weight mask according to the bounding boxes for components.

Here we display the practical pre-process time on some representative datasets, which shows that the bounding boxes and connected components can be obtained with acceptable efficiency.

Table 11. Practical pre-process time for each dataset.

Dataset	Stage(a)	Stage(b)	Total
DUTS-TE(5,019)	474.0s	188.2s	658.2s
ECSSD(1,000)	91.8s	40.5s	132.3s
DUT-OMRON(5,168)	470.7s	220.1s	690.8s
HKU-IS(4,447)	508.8s	190.1s	698.9s
XPIE(10,000)	1432.2s	316.5s	1748.7s