Pixels, Regions, and Objects: Multiple Enhancement for Salient Object Detection

Yi Wang
DUT-RU International School of
Information Science and Engineering
Dalian University of Technology, China
dlutwangyi@dlut.edu.cn

Ruili Wang [*]
School of Mathematical and
Computational Sciences
Massey University, New Zealand
ruili.wang@massey.ac.nz

Xin Fan
DUT-RU International School
of Information Science and Engineering
Dalian University of Technology, China
xin.fan@dlut.edu.cn

Tianzhu Wang
School of Mathematical and
Computational Sciences
Massey University, New Zealand
wangtz@126.com

Xiangjian He [*]
School of Computer Science
University of Nottingham Ningbo China, Ningbo, China
sean.he@nottingham.edu.cn

Abstract

Salient object detection (SOD) aims to mimic the human visual system (HVS) and cognition mechanisms to identify and segment salient objects. However, due to the complexity of these mechanisms, current methods are not perfect. Accuracy and robustness need to be further improved, particularly in complex scenes with multiple objects and background clutter. To address this issue, we propose a novel approach called Multiple Enhancement Network (MENet) that adopts the boundary sensibility, content integrity, iterative refinement, and frequency decomposition mechanisms of HVS. A multi-level hybrid loss is firstly designed to guide the network to learn pixel-level, region-level, and object-level features. A flexible multiscale feature enhancement module (ME-Module) is then designed to gradually aggregate and refine global or detailed features by changing the size order of the input feature sequence. An iterative training strategy is used to enhance boundary features and adaptive features in the dual-branch decoder of MENet. Comprehensive evaluations on six challenging benchmark datasets show that MENet achieves state-of-the-art results. Both the codes and results are publicly available at https://github.com/yiwangtz/MENet.

1. Introduction

Salient object detection (SOD) aims to identify the most visually conspicuous regions in an image that are consistent with the human visual system (HVS) and cognition mechanisms [9,13,40]. SOD can eliminate redundant information and improve computational performance for many high-level computer vision tasks, such as action recognition [4, 60], image segmentation [3, 33], image captioning [52], object tracking [14], and video summary [58]. Fully convolutional networks (FCNs) [25] based SOD models have been particularly effective at improving SOD performance in recent years [40]. However, accurately segmenting complex object boundaries remains a challenging task for SOD. This is especially true when the geometry and/or boundaries of these objects are complex, or when scenes are chaotic or cluttered [9, 10], as shown in Fig. 1.

An intuitive solution for addressing this problem is to explore the mechanisms of the human vision system (HVS) [55] and some of which have been used to improve SOD models, as described below. (i) A human tends to enhance recognition by alternating between viewing the entire object and the details of complex scenes, which has been utilized for various visual tasks [16, 17, 22, 36]. (ii) HVS is sensitive to both boundary/contour and structural information, so dual-branch feature refinement structures have been developed to incorporate extra-edge information to enhance salient feature learning [11, 22, 24, 43, 47, 53, 57]. Some

[*] The corresponding authors

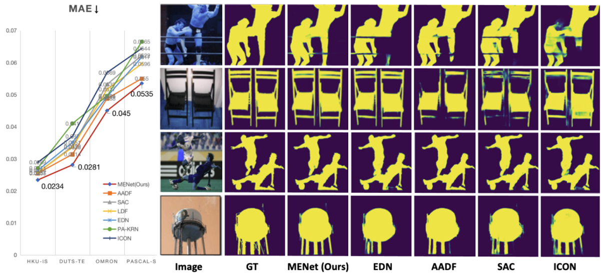


Figure 1. Illustration of *MAE* (left part) and some visual results (right part) for the proposed MENet with some recent state-of-the-art SOD methods: EDN [45], AADFNet [57], SAC [16], and ICON [59]. Please refer to Sec.5 for detailed experimental settings. The MENet model achieves the lowest *MAE* score with the most precise and complete boundaries.

structural similarity measurements (e.g., Structural Similarity Index (SSIM) [41]) and regional similarity measurements (e.g., Intersection over Union (IoU) [35] and Dice [12]) are also adopted by SOD models [32, 42, 43, 47, 59] in the loss functions. (iii) Human vision is indeed holistic and continuous so that it perceives objects and scenes as organized wholes [18], which are composed of parts that are meaningful and coherent in relation to each other. ICON [59] proposes to improve the integrity from both macro- and micro-level perspectives by enhancing integrity information hidden in channels of features. EDN [45] employs a powerful down-sampling technique to learn a global view of the whole image effectively. (iv) According to the human visual spatial frequency model [30], an image can be decomposed into or synthesized by high-spatial frequency and low-spatial frequency parts. As a starting point in this work, we intend to use the mechanisms outlined above to further improve SOD performance for complex scenes.

In this work, we propose a multi-enhancement network (MENet) that effectively integrates the above HVS mechanisms in a U-Net-like [37] encoder-decoder framework to produce more accurate SOD for complex scenes. Foremost, MENet employs the image frequency decomposition idea to design a two-stream feature learning decoder for boundaries (high frequencies) and inner body regions (low frequencies). This setting is different from the existing two-branch (or edge-aware) methods [11, 22, 47, 51, 53, 56, 57] that use one branch for the boundary and the other one for the entire object, such as EGNet [53] and AFNet [11]. Particularly, there is no interaction between the intermediate features of the two branches of MENet, so it reduces the interference of inaccurate boundary information with global features. This is because boundary features need to be highly discriminative against the background, while global features need consistency and robustness. Although LDF [43] also learns internal regional features in one branch, its detailed map and body map cannot be computed accurately and efficiently for geometrically complex objects.

Then, we propose an iterative training strategy to progressively enhance features by alternately aggregating high- and low-level features to mimic HVS bottom-up and top-down refinement mechanisms. To produce high- and low-level features flexibly, we design a multiscale feature enhancement module (ME-Module) as the core of each branch by leveraging atrous spatial pyramid pooling (ASPP) [34] and global-local attention [6].

In addition, we introduce the HVS holistic and continuous mechanism to loss function design. We present a multi-level hybrid loss, which evaluates the pixel-, region-, and object-level similarities between predicted saliency maps and ground-truth (GT) saliency maps. For pixel-level loss, we also use Binary Cross Entropy (BCE) [7] loss to ensure network accuracy and convergence speed. As for region-level loss, we divide a saliency map into four sub-regions of equal size and then calculate the sum of weighted regional similarities through SSIM and IoU. Then, inspired by SSIM and S-measure [5], an object-level loss is designed by the contrast and distribution statistics of the foreground between the GT map and the predicted map. A similar hybrid loss is reported in BASNet [32], but it uses a simple combination of BCE, IoU, and SSIM for the whole saliency map without partitioning regions. Following is a summary of our contributions.

- We propose to leverage not only pixel-level but also region-level and object-level similarity measures in loss to increase prediction accuracy and integrity, and then design a multi-level hybrid loss to implement this proposal.

- We design a multiscale feature enhancement module (ME-Module) to mimic HVS bottom-up and top-down refinement mechanisms. ME-Module can gradually propagate and produce comprehensive global or detailed features by changing the size and order of the input features.

- We propose a novel Multiple Enhancement Network (MENet) for dealing with SOD in complex scenes by integrating multiple HVS mechanisms into the network structure and loss function. Specifically, a two-branch decoder equipped with ME-Modules is designed to incrementally refine the boundary and adaptive features by an iterative training strategy and the proposed multilevel hybrid loss.

The results of quantitative and qualitative experiments on six datasets demonstrate MENet outperforms the state-of-the-art methods by a large margin, as shown in Fig. 1.

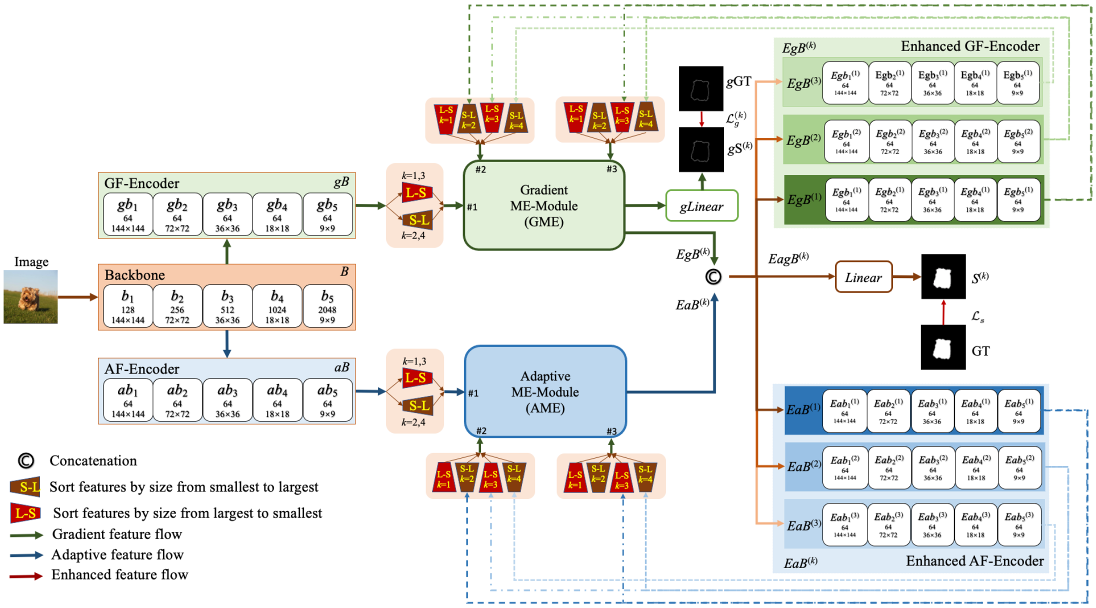


Figure 2. Illustration of the overall architecture and the pipeline of the MENet.

2. Related Work

In recent years, salient object detection (SOD) approaches based on encoder-decoder and feature aggregation architecture have achieved high performance [13, 26, 40]. Below, we briefly review related models to this work.

MLMSNet [44] exploits the supervision of foreground boundary detection and edge detection. AFNet [11] develops an attentive feedback module to better explore target structure. EGNet [53] uses complementary information about salient edges and salient objects to propose an edge guidance network. CPDNet [46] proposes a partial decoder to refine high-level features to generate precise saliency maps. BASNet [32] has a fine prediction model via sequentially stacking two U-Nets [37] with different configurations. AADFNet [57] is designed to use an attentional dense ASPP-based network to selectively use small dilated rate convolutions and big dilated rate convolutions to get local and global saliency information. GateNet [54] designs a gated dual-branch structure to establish a cooperative relationship between features of different levels to increase network discriminability. U2Net [31] puts forward a novel ReSidual U-block (RSU), which can obtain intrastate multi-resolution features without reducing feature map resolution. MINet [50] proposes to enhance the feed-forward neural network by adopting a refinement mechanism for multiple stages. LDF [43] designs a two-branch decoder to predict saliency maps by utilizing the complement of body and detail information of objects. SAC [16] implements a spatial attenuation context module to propagate and aggregate salient features through two rounds of recurrent translations. CANet [36] presents a context-aware attention module, which detects salient regions by simultaneously establishing links between each pixel and its surrounding global and local contexts. KRN [47] uses intermediate edge supervision in its coarse locating module. ICON [59] introduces three diverse feature aggregations, integrity channel enhancement, and part-whole verification to SOD. EDN [45] uses an extreme down-sampling method to effectively learn global features and Scale-Correlated Pyramid Convolution in the decoder to recover local details. When a scene has low contrast and is blurred, it is still challenging to discern the fine boundaries (or contours) of salient objects. In our work, we draw inspiration from the above-mentioned methods (e.g., two-branch decoders, ASPP, attention, and iterative refinement), but our model employs these methods uniquely.

3. Methodology

3.1. Framework Overview

The proposed MENet utilizes the encoder-decoder structure, as shown in Fig. 2. The initial encoding part consists of a backbone (e.g., ResNet-50 [15]), a gradient feature encoder (GF-Encoder), and an adaptive feature encoder (AF-Encoder). Specifically, a 3-channel image of size $[W \times H]$ is fed directly into the backbone network, and then N (e.g., $N = 5$ in Resnet-50) multistage feature maps (denoted by B=$\{b_i\}_{i=1}^{N}$) are extracted. Then, B is squeezed into 64-channel features gB=$\{gb_i\}_{i=1}^{N}$ and aB=$\{ab_i\}_{i=1}^{N}$ by the GF-Encoder and the AF-Encoder, respectively. Then, gB and aB are passed to a gradient ME-Module (GME) and an adaptive ME-Module (AME) to learn gradient features (i.e., boundary features) EgB=$\{Egb_i\}_{i=1}^{N}$ under the supervision ($\mathcal{L}_g^{(k)}$) and adaptive features (i.e., inner body features) EaB=$\{Eab_i\}_{i=1}^{N}$, respectively. Afterward, EgB and EaB are concatenated as an enhanced feature block $EagB$ which is then fed to an Enhanced GF-Encoder and an Enhanced AF-Encoder to be squeezed into 64-channel again then put into the GME and AME together with gB and aB for the next iteration, respectively. A four-iteration enhancing procedure alternates propagating and aggregating high- and low-level EgB and high- and low-level EaB, in parallel, with the proposed multilevel hybrid loss $\mathcal{L}_s$.

3.2. Multiscale Feature Enhancement

Figure 3 demonstrates the ME-module composition. By utilizing ASPP [34] and ME-Attention, ME-Module gradually propagates and fuses features in five stages. Each stage first receives and makes a pixel-wise addition for a specific size of the sub-features Fb_{ij} of the input feature block from three input ports {#1, #2, #3} and the output of the previous stage, where i is the port index and j is the corresponding sub-feature number to the five stages in the three input feature blocks. After that, ASPP focuses on diversifying visual fields, while the ME-Attention module emphasizes salient object location through global and detailed attention [21]. An 'Interpolator' performs up-sampling and down-sampling operations, accordingly, to match the scale of features of the next stage. ME-Module is versatile in that it can output global or detailed features simply by adjusting the spatial order of multiscale feature maps.

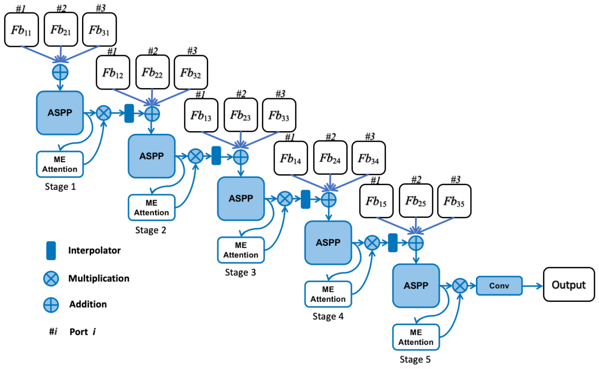


Figure 3. Illustration of the proposed ME-Module.

3.3. Iterative Enhancement

Because the ME-Module can be used as either a global feature enhancer or a detailed feature enhancer, we can alternately change the input sequence of the feature maps and make them complement one another.

Given gradient and adaptive backbone feature blocks gB and aB, our first step is to sort their sub-features from small to large, respectively. Afterward, we iteratively enhance their global and local features by changing their input order through four rounds of iteration, respectively. The features after each round of enhancement are denoted as $EgB^{(k)}$ and $EaB^{(k)}$ ($k \in [1, 4]$), respectively. The entire enhancement process is summarized in Eqs. 1,2,3. The pipeline is illustrated in Fig. 2.

$$EgB^{(k)} = GME(V(gB), V(EgB^{(k-1)}), V(EgB^{(k-2)})), \quad (1)$$

where the $GME(p_1, p_2, p_3)$ represents the Graduate ME-Module and $V(\ )$ is used to get the reverse of the feature list. If $k \leq 0$, we let $p_i = Null$.

$$EaB^{(k)} = AME(V(aB), V(EaB^{(k-1)}), V(EaB^{(k-2)})), \quad (2)$$

where $AME(q_1, q_2, q_3)$ represents the Adaptive ME-Module. If $k \leq 0$, we let $q_i = Null$.

$$S^{(k)} = Linear(Concat(EgB^{(k)}, EaB^{(k)}), \quad (3)$$

where $Linear(\ )$ is the linear layer and $Concat(\ ,\ )$ is the concatenation operation.

Figure 4 shows two visual examples of the saliency map $S^{(k)}$ in each iteration. The odd number iteration extracts global features, which are then refined by the even number iterations. Thus, the global and detailed features complement each other and promote each other.

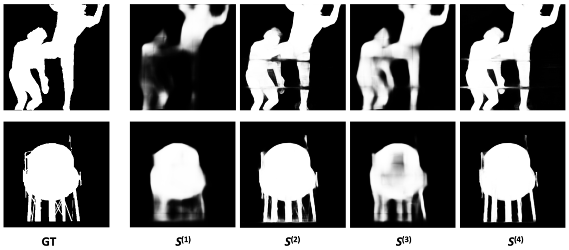


Figure 4. Examples of the saliency map ($S^{(k)}$) in each iteration.

4. Supervision Strategy

In MENet, we set two supervisions in each enhancement round, as shown in Fig. 2. One is gradient supervision in the GME module with a BCE loss (denoted by $\mathcal{L}_g^{(k)}$). The other supervision is for the overall saliency maps ($S^{(k)}$) with the proposed multilevel hybrid loss (denoted by $\mathcal{L}_s^{(k)}$). We do not assign any supervision for the Adaptive ME-Module, so the overall training loss $\mathcal{L}$ of MENet is defined by

$$\mathcal{L} = \sum_{k=1}^{4} ({\alpha_1}^{(k)} \mathcal{L}_g^{(k)} + {\alpha_2}^{(k)} \mathcal{L}_s^{(k)}), \quad (4)$$

where we set $\alpha_1^{(k)} = \alpha_2^{(k)} = 0.5$ in the implementation.

$\mathcal{L}_g$ in each round is a pixel-wise BCE loss defined by

$$\mathcal{L}_g = -\sum(G_g log S_g + (1-G_g)log(1-S_g)), \quad (5)$$

where $G_g \in [0,1]$ is the gradient map of the GT map $G \in [0,1]$, and $S_g \in [0,1]$ is the gradient map of the predicted saliency map $S \in [0,1]$. G_g is the high-frequency part of the GT map, which can direct the learning of boundary features.

The multilevel hybrid loss $\mathcal{L}_s$ in each round, composed of a pixel-level loss ($\mathcal{L}_{s_{bce}}$), a region-level loss ($\mathcal{L}_{s_{reg}}$), and an object-level loss ($\mathcal{L}_{s_{obj}}$), is defined by

$$\mathcal{L}_s = \beta_1 \mathcal{L}_{s_{bce}} + \beta_2 \mathcal{L}_{s_{reg}} + \beta_3 \mathcal{L}_{s_{obj}}, \quad (6)$$

where we set $\beta_1 = \beta_2 = 0.4$ and $\beta_3 = 0.2$ in the implementation.

Pixel-level Loss: $\mathcal{L}_{s_{bce}}$ is a BCE loss and can be computed by

$$\mathcal{L}_{s_{bce}} = -\sum(G log S + (1-G)log(1-S)), \quad (7)$$

Region-level Loss: We design $L_{s_{Reg}}$ based on a structural measure SSIM [41] and a regional measure IoU [35] of images. Particularly, we divide S and G evenly into four equal sub-regions (i.e., S_i and G_i, $i \in [1,4]$), respectively, which can be processed as a batch in implementation. Then $\mathcal{L}_{Reg}$ can be represented by

$$\mathcal{L}_{s_{Reg}} = 1 - \sum_{i=1}^{4} \omega_i(\theta_1 \mathit{SSIM}_i + \theta_2 \mathit{IoU}_i), \quad (8)$$

where $\theta_1 = \theta_2 = 0.5$, and ω_i is the ratio of the predicted foreground (i.e., the salient regions) to the corresponding GT foreground in each region.

The SSIM_i between a pair of regions S_i and G_i can simply be formulated as a product of three components [1,41]: the luminance comparison, the contrast comparison, and the structure comparison by

$$\mathit{SSIM}_i = \frac{2\mu_{S_i}\mu_{G_i} + c_1}{\mu_{S_i}^2 + \mu_{G_i}^2 + c_1} \cdot \frac{2\sigma_{S_i}\sigma_{G_i} + c_2}{{\sigma_{S_i}}^2 + {\sigma_{G_i}}^2 + c_2} \cdot \frac{\sigma_{S_iG_i} + c_3}{\sigma_{S_i}\sigma_{G_i} + c_3}, \quad (9)$$

where μ_{S_i} and μ_{G_i} are the means, σ_{S_i} and σ_{G_i} are the standard deviations, and $\sigma_{S_iG_i}$ is the covariance of the predicted regional saliency map S_i and regional GT map G_i, respectively, and c_1, c_2, and c_3 are small quantities introduced for numerical stab [1].

Additionally, we also measure the overlap of spatial regions between predictions and labels using the IoU in $L_{s_{Reg}}$, which is defined as

$$IoU_i = \frac{\sum\sum(S_iG_i)}{\sum\sum(G_i + S_i - G_iS_i)}. \quad (10)$$

Object-level Loss: Inspired by SSIM and S-measure [5], we introduce object-level image similarity measurement to the object-level loss $L_{s_{obj}}$. Since GT maps usually have sharp foreground-background contrast and a uniform distribution, the predicted saliency maps should also have these properties [5] [41]. We use the luminance component of SSIM and the coefficient of variation (i.e., the ratio of mean to deviation) to model these two properties, respectively. Since accurate foreground is the primary objective in training the whole network, we only consider the foregrounds (denoted by S_o and G_o, respectively) of S and G and compute the foreground distribution. This is the difference from the S-measure that considers both the distributions of the foregrounds and the backgrounds of S and G. Thus, $L_{s_{obj}}$ is defined by

$$L_{s_{obj}} = 1 - \frac{1}{(\frac{\mu_{S_o}^2 + \mu_{G_o}^2}{2\mu_{S_o}\mu_{G_o}} + \lambda\frac{\sigma_{S_o}}{\mu_{S_o}})}, \quad (11)$$

where μ_{S_o} and μ_{G_o} denote the means of S_o and G_o, respectively, σ_{S_o} is the standard deviation of S_o, and λ is the weight. Since μ_{G_o} is exactly 1 in practice, Eq.11 can be simplified by

$$L_{s_{obj}} = 1 - \frac{2\mu_{S_o}}{\mu_{S_o}^2 + 1 + 2\lambda\sigma_{S_o}}. \quad (12)$$

5. Experiments

5.1. Experimental Settings

Training and Testing Strategy: We use ImageNet [19] to pre-train the backbone network and then use the DUTS-TR [39] to fine-turn the proposed MENet. Other MENet parameters are initialized randomly in a normal distribution. Inputs are scaled to [352 × 352], [320 × 320], [288 × 288], [256 × 256], and [224 × 224] for data augmentation. We utilize the stochastic gradient descent (SGD) optimizer [2] and set the maximum learning rate of the backbone network to 0.00025 and other parts to 0.0025. The momentum is set to 0.9, the weight decay is set to 0.0005, and the batch size is set to 24. The 'poly' learning rate strategy is also adopted. Our network is built on PyTorch 1.12 on a computer server with AMD EPYC 7742 (2.25GHz) and an NVIDIA A100 GPU (with 40 GB of memory). The network is trained for 99 epochs. Inference for a testing image scaled to [352 × 352] takes just 0.022s (45 fps).

Testing Datasets: We evaluate all methods on the following six popular SOD benchmark datasets. OMRON [49] has 5,168 images of salient objects with relatively complex backgrounds. DUTS-TE is part of DUTS [39] designed for testing, including 5,019 images. HKU-IS [20] has 4,447 images, primarily containing multiple scattered salient objects. PASCAL-S [23] is a subset of PASCAL-VOC 2010, containing 850 images, and has less bias than most saliency

			OMRON [49] (5,168 images)					DUTS-TE [39] (5,019 images)					SOD [28] (300 images)				
Year	Method	Backbone	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$
2019	MLMSNet [44]	VGG-16	0.0635	0.7740	0.7455	0.8387	0.8093	0.0484	0.8511	0.8137	0.8631	0.8618	0.1060	0.8517	0.8291	0.8019	0.7898
2019	AFNet [11]	VGG-16	0.0573	0.7972	0.7766	0.8595	0.8263	0.0453	0.8623	0.8340	0.8929	0.8672	-	-	-	-	-
2019	EGNet [53]	VGG-16	0.0564	0.8087	0.7855	0.8642	0.8357	0.0431	0.8764	0.8472	0.8927	0.8786	0.1100	0.8589	0.8426	0.8209	0.7882
2019	EGNet [53]	ResNet-50	0.0528	0.8155	0.7942	0.8738	0.8412	0.0386	0.8880	0.8597	0.9040	0.8873	0.0969	0.8778	0.8610	0.8422	0.8067
2019	CPD [46]	VGG-16	0.0567	0.7935	0.7800	0.8685	0.8178	0.0425	0.8638	0.8458	0.9038	0.8669	0.1125	0.8480	0.8365	0.8124	0.7715
2019	CPD [46]	ResNet-50	0.0560	0.7966	0.7807	0.8726	0.8248	0.0429	0.8649	0.8431	0.9009	0.8691	0.1097	0.8568	0.8376	0.8174	0.7711
2019	BASNet [32]	ResNet-34	0.0565	0.8053	0.7906	0.8691	0.8362	0.0472	0.8589	0.8416	0.8790	0.8660	0.1124	0.8487	0.8368	0.7793	0.7721
2020	AADFNet [57]	ResNet-50	0.0488	0.8143	0.8050	0.8744	0.8389	0.0314	0.8993	0.8911	0.9225	0.8914	0.0903	0.8677	0.8579	0.8051	0.7929
2020	GateNet [54]	VGG-16	0.0613	0.7940	0.7691	0.8534	0.8209	0.0448	0.8695	0.8388	0.8856	0.8705	-	-	-	-	-
2020	GateNet [54]	ResNet-50	0.0552	0.8180	0.7914	0.8682	0.8382	0.0399	0.8870	0.8552	0.9004	0.8852	-	-	-	-	-
2020	GateNet [54]	ResNet-101	0.0547	0.8210	0.7944	0.8736	0.8449	0.0380	0.8919	0.8615	0.9075	0.8910	-	-	-	-	-
2020	U2Net [31]	RSU	0.0544	0.8226	0.8023	0.8716	0.8467	0.0443	0.8719	0.8479	0.8840	0.8738	0.1061	0.8588	0.8428	0.7993	0.7891
2020	MINet [50]	VGG-16	0.0572	0.7936	0.7755	0.8644	0.8218	0.0395	0.8761	0.8550	0.9070	0.8753	-	-	-	-	-
2020	MINet [50]	ResNet-50	0.0559	0.8098	0.7911	0.8734	0.8329	0.0373	0.8833	0.8597	0.9132	0.8842	-	-	-	-	-
2020	LDF [43]	ResNet-50	0.0517	0.8199	0.8015	0.8814	0.8392	0.0336	0.8968	0.8779	0.9232	0.8924	-	-	-	-	-
2021	SAC [16]	ResNet-101	0.0523	0.8287	0.8092	0.8833	0.8487	0.0339	0.8944	0.8732	0.9208	0.8957	0.0934	**0.8804**	0.8695	0.8482	0.8087
2021	CANet [36]	CNN	0.0581	0.8101	0.7796	0.8593	0.8356	0.0437	0.8755	0.8382	0.8896	0.8781	0.0992	0.8650	0.8406	0.8331	0.8007
2021	SGL-KRN [47]	ResNet-50	0.0492	0.7961	0.7830	0.8783	0.8464	0.0337	0.8833	0.8649	0.9311	0.8929	-	-	-	-	-
2021	PA-KRN [47]	ResNet-50	0.0496	0.8101	0.7956	0.8880	0.8533	0.0328	0.8945	0.8761	0.9353	0.9005	-	-	-	-	-
2022	ICON [59]	ResNet-50	0.0569	0.8254	0.8013	0.8791	0.8445	0.0370	0.8917	0.8665	0.9142	0.8889	**0.0841**	0.8790	**0.8711**	**0.8516**	**0.8238**
2022	EDN [45]	VGG-16	0.0565	0.7818	0.7686	0.8628	0.8376	0.0410	0.8636	0.8457	0.9118	0.8829	-	-	-	-	-
2022	EDN [45]	ResNet-50	0.0494	0.7992	0.7880	0.8774	0.8495	0.0351	0.8784	0.8634	0.9250	0.8924	-	-	-	-	-
2023	**MENet(Ours)**	ResNet-50	**0.0450**	**0.8337**	**0.8178**	**0.8911**	**0.8496**	**0.0281**	**0.9123**	**0.8930**	**0.9368**	**0.9049**	0.0874	0.8780	0.8684	0.8381	0.8089

Table 1. Quantitative comparisons on OMRON [49], DTUS-TE [39], and SOD [28] datasets. The best results are marked in bold. The symbols '↓ / ↑' indicate the lower/higher the evaluation metric, the better the model is. The symbol '-' means that the model is not available. Overall, the proposed MENet has superior performance.

			HKU-IS [20] (4,447 images)					PASCAL-S [23] (850 images)					ECSSD [48] (1,000 images)				
Year	Method	Backbone	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$
2019	MLMSNet [44]	VGG-16	0.0387	0.9207	0.8891	0.9379	0.9066	0.0736	0.8552	0.8254	0.8447	0.8443	0.0446	0.9284	0.9007	0.9161	0.9112
2019	AFNet [11]	VGG-16	0.0358	0.9226	0.8998	0.9475	0.9055	0.0700	0.8629	0.8409	0.8851	0.8494	0.0418	0.9350	0.9157	0.9414	0.9135
2019	EGNet [53]	VGG-16	0.0345	0.9273	0.9050	0.9503	0.9100	0.0776	0.8585	0.8371	0.8714	0.8475	0.0405	0.9434	0.9232	0.9408	0.9193
2019	EGNet [53]	Resnet-50	0.0309	0.9352	0.9122	0.9564	0.9180	0.0740	0.8653	0.8437	0.8772	0.8521	0.0374	0.9474	0.9288	0.9469	0.9246
2019	CPD [46]	VGG-16	0.0333	0.9239	0.9075	0.9501	0.9042	0.0721	0.8612	0.8441	0.8837	0.8446	0.0402	0.9360	0.9233	0.9433	0.9103
2019	CPD [46]	ResNet-50	0.0342	0.9250	0.9047	0.9503	0.9056	0.0706	0.8595	0.8414	0.8873	0.8484	0.0371	0.9393	0.9244	0.9494	0.9182
2019	BASNet [32]	ResNet-34	0.0322	0.9284	0.9113	0.9458	0.9090	0.0758	0.8539	0.8344	0.8527	0.8380	0.0370	0.9425	0.9274	0.9210	0.9163
2020	AADFNet [57]	ResNet-50	0.0255	0.9415	**0.9339**	0.9592	0.9190	0.0550	0.8797	0.8677	0.9051	0.8658	**0.0280**	0.9543	**0.9478**	0.9529	0.9299
2020	GateNet [54]	VGG-16	0.0361	0.9287	0.9036	0.9470	0.9100	0.0684	0.8696	0.8439	0.8692	0.8574	0.0418	0.9413	0.9191	0.9314	0.9169
2020	GateNet [54]	ResNet-50	0.0337	0.9335	0.9097	0.9534	0.9154	0.0680	0.8690	0.8459	0.8842	0.8580	0.0408	0.9454	0.9250	0.9431	0.9198
2020	GateNet [54]	ResNet-101	0.0320	0.9375	0.9136	0.9567	0.9195	0.0668	0.8702	0.8468	0.8924	0.8622	0.0357	0.9508	0.9301	0.9501	0.9302
2020	U2Net [31]	RSU	0.0312	0.9352	0.9133	0.9484	0.9161	0.0740	0.8592	0.8386	0.8500	0.8444	0.0330	0.9510	0.9325	0.9251	0.9276
2020	MINet [50]	VGG-16	0.0316	0.9302	0.9133	0.9540	0.9119	0.0645	0.8650	0.8450	0.8961	0.8544	0.0370	0.9435	0.9296	0.9475	0.9192
2020	MINet [50]	ResNet-50	0.0292	0.9349	0.9166	0.9600	0.9189	0.0643	0.8665	0.8461	0.8981	0.8563	0.0342	0.9475	0.9309	0.9532	0.9250
2020	LDF [43]	ResNet-50	0.0275	0.9394	0.9224	0.9597	0.9196	0.0596	0.8741	0.8577	0.9048	0.8630	0.0335	0.9501	0.9379	0.9509	0.9245
2021	SAC [16]	ResNet-101	0.0257	0.9416	0.9260	0.9636	0.9253	0.0622	0.8772	0.8585	0.9022	0.8656	0.0309	0.9512	0.9376	**0.9586**	**0.9312**
2021	CANet [36]	CNN	0.0371	0.9297	0.8977	0.9455	0.9100	0.0728	0.8662	0.8392	0.8790	0.8552	0.0441	0.9378	0.9103	0.9362	0.9154
2021	SGL-KRN [47]	ResNet-50	0.0280	0.9301	0.9154	0.9539	0.9206	0.0678	0.8502	0.8373	0.8941	0.8556	0.0360	0.9368	0.9241	0.9462	0.9231
2021	PA-KRN [47]	ResNet-50	0.0271	0.9349	0.9198	0.9561	0.9235	0.0665	0.8530	0.8388	0.8964	0.8578	0.0323	0.9425	0.9301	0.9503	0.9278
2022	ICON [59]	ResNet-50	0.0289	0.9395	0.9196	0.9585	0.9202	0.0644	0.8757	0.8514	0.8931	0.8611	0.0318	0.9503	0.9336	0.9543	0.9290
2022	EDN [45]	VGG-16	0.0286	0.9286	0.9141	0.9504	0.9208	0.0650	0.8555	0.8406	0.8955	0.8605	0.0336	0.9408	0.9285	0.9508	0.9283
2022	EDN [45]	ResNet-50	0.0264	0.9325	0.9196	0.9548	0.9241	0.0617	0.8600	0.8489	0.9015	0.8646	0.0320	0.9410	0.9304	0.9508	0.9267
2023	**MENet(Ours)**	ResNet-50	**0.0234**	**0.9483**	0.9319	**0.9657**	**0.9274**	**0.0535**	**0.8896**	**0.8701**	**0.9132**	**0.8721**	0.0307	**0.9549**	0.9422	0.9544	0.9279

Table 2. Quantitative comparisons on HKU-IS [20], PASCAL-S [23], and ECSSD [48] datasets. The best results are bolded. The symbols '↓ / ↑' indicate the lower/higher the evaluation metric, the better the model is. Overall, the proposed MENet has superior performance.

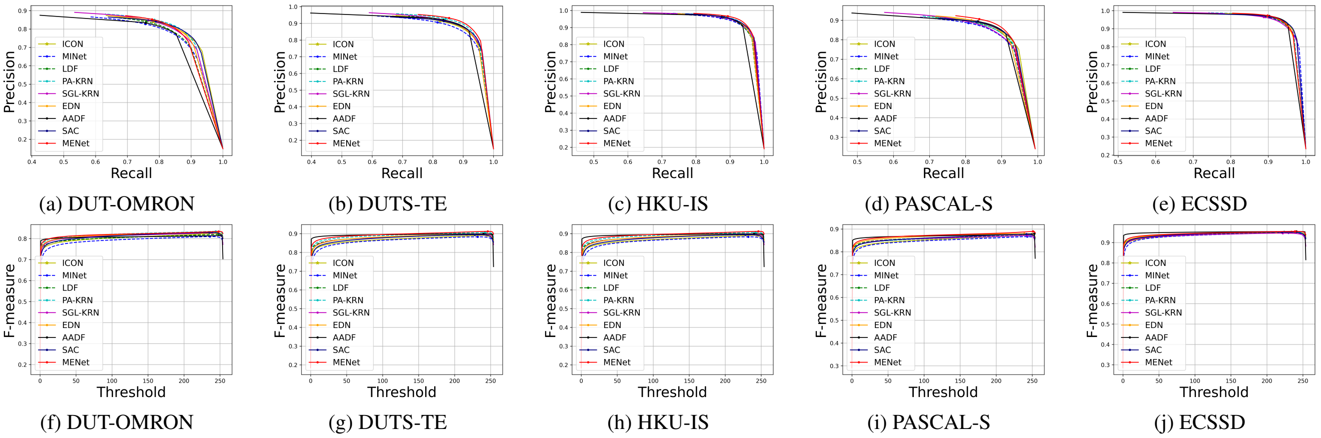
datasets. ECSSD [48] has 1,000 images with semantically meaningful but complex structures, and the backgrounds also have complexity. SOD [28] consists of 300 challenging images of seven subjects.

Evaluation Criteria: We adopt the Mean Absolute Error (*MAE*) [29], the max F-measure (*MaxF*) and the mean F-measure (*mF*) [27], the mean Enhanced-alignment Measure (mE_m) [8], and the S-measure (S_m) [5] to assess SOD models. We plot Precision-Recall (PR) curves and F-measure curves to show overall performance.

5.2. Comparison with the state-of-the-arts

We compare the proposed MENet with sixteen recent state-of-the-art models, including MLMSNet [44], AFNet [11], EGNet [53], CPD [46], BASNet [32], AADFNet [57], GateNet [54], U2Net [31], MINet [50], LDF [43], SAC [16], CANet [36], SGL-KRN [47], PA-KRN [47], EDN [45], and ICON [59]. For a fair comparison, the saliency maps are either provided by the authors or generated by the officially released pre-trained models.

Quantitative performance comparison: We rank the methods across the backbone networks (e.g., VGG-16 [38], ResNet-50 [15], and ResNet-101 [15]), because some models do not provide codes or predictive results for all the backbones, as shown in Tables. 1 and 2. Obviously, the proposed MENet (with ResNet50) outperforms other methods by a large margin. MENet performs well on the metrics mE_m and S_m. Although MENet is inferior to the ICON [59]



(a) DUT-OMRON (b) DUTS-TE (c) HKU-IS (d) PASCAL-S (e) ECSSD

(f) DUT-OMRON (g) DUTS-TE (h) HKU-IS (i) PASCAL-S (j) ECSSD

Figure 5. PR-curves (Row 1) and Fm-curves (Row 2) for SOD methods: ICON [59], MINet [50], LDF [43], PA-KRN [47], SGL-KRN [47], EDN [45], AADFNet [57], SAC [16], and MENet. The proposed MENet has the overall best performance on five datasets.

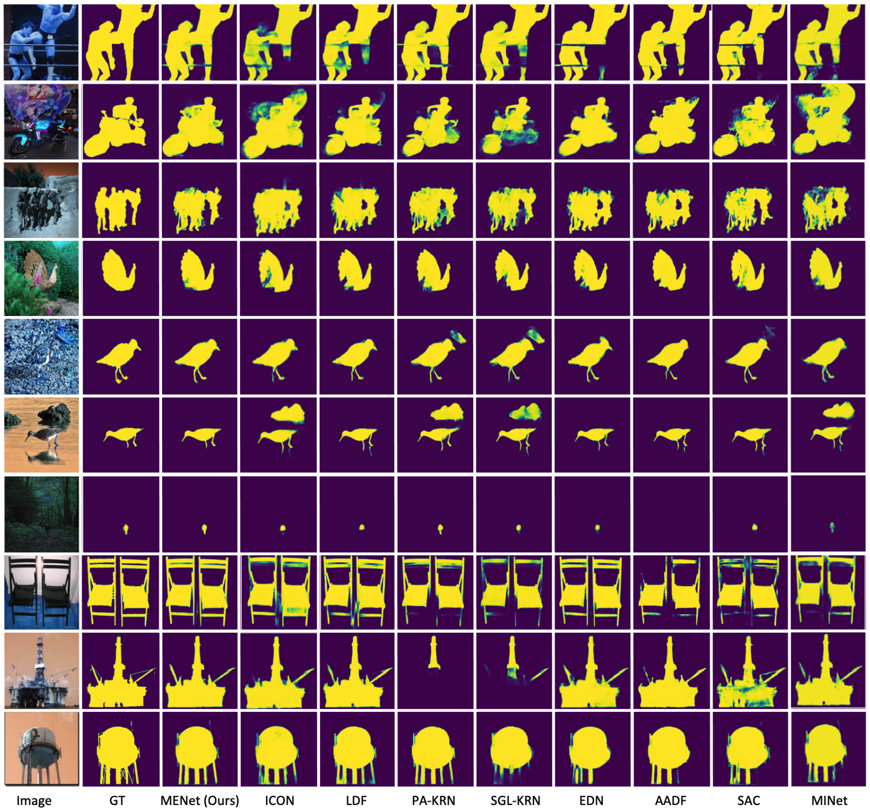


Figure 6. Quantitative comparisons with recent state-of-the-art methods: ICON [59], LDF [43], PA-KRN [47], SGL-KRN [47], EDN [45], AADFNet [57], SAC [16], and MINet [50]. The proposed MENet produces more complete and clear boundaries with fewer background noises for various complex scenes.

and SAC [16] on the SOD dataset, and AADFNet [57] on the ECSSD dataset, its performances are generally close to the leading ones. MENet uses Resnet-50 as the backbone network, while SAC uses ResNet-101 as the backbone network. ICON only has better performance on the SOD dataset. F-Measure and PR curves are shown in Fig. 5, respectively. MENet performs best overall on PR and F-Measure curves. All these statistical results reveal the superiority of MENet.

Qualitative performance comparison: We select a few challenging scenes for comparison, including small objects, reflections, large objects, multiple complex objects, and low-contrast environments, as shown in Fig. 6. The prediction results are given in probability. The determined value '1' is marked in yellow, the other probability values are indicated in different shades of green, and the value '0' is black. As we can see, the proposed MENet achieves the most accurate overall detection results for salient regions. The boundaries of the targets are also more precise and complete.

5.3. Ablation Study

We investigate the different settings for enhancement and supervision of MENet that uses Resnet-50 as the backbone.

Number of iterative enhancements: Table 3 shows that the four-round enhancement setting achieves the best results on all databases. According to our settings, the first and third rounds learn global features, and the second and fourth rounds learn detailed features, so the result of *Round 2* is better than that of *Round 3*. But the result of *Round 4* is much better than that of *Round 2*, demonstrating the effectiveness of our iterative training strategies.

Loss combinations: Table 4 shows that with the introduction of region-level loss ($\mathcal{L}_{s_{reg}}$) and object-level loss ($\mathcal{L}_{s_{obj}}$), the performance of all metrics is gradually increased on OMRON, DUTS-TE, HKU-IS, and PASCAL-S datasets. Especially, when gradient supervision ($\mathcal{L}_g$) is added, the overall accuracy is further improved. For the ECSSD and SOD datasets, although the metrics do not conform to this rule, overall, the combination we use in MENet is still close to the best ones.

Sub-region numbers of region-level loss: Table 4 shows that the four-sub-region setting improves performance significantly than the one-region setting, by reducing *MAE* scores by 4.66%, 4.75%, 8.24%, 7.12%, 5.83%, and 7.42% on six datasets, respectively. Therefore, partitioning regions is necessary.

6. Conclusion

This paper proposes a method to improve the performance of SOD for complex scenes by fully leveraging the human visual system (HVS) and cognition mechanisms. The method consists of a novel multilevel hybrid loss that measures pixel-, region- and object-level similarities, and a multi-enhancement network (MENet) that integrates multiple mechanisms of the HVS. Two multiscale feature enhancement modules are the core of MENet, and they gradually propagate and fuse boundary and global features, respectively. Comprehensive experiments on six challenging databases demonstrate that MENet achieves the new state-of-the-art for SOD. As demonstrated in all the studies in this paper, SOD in blurred and low-contrast real scenes is still a valuable but challenging topic that warrants more research efforts.

Dataset	ITimes	$MAE\downarrow$	$F_\beta^{max}\uparrow$	$mF_\beta\uparrow$	$mE_m\uparrow$	$S_m\uparrow$
OMRON [49]	1	0.0617	0.7867	0.7657	0.8701	0.7827
	2	0.0483	0.8304	0.8067	0.8854	0.8409
	3	0.0504	0.8066	0.7956	0.8755	0.8325
	4	**0.0450**	**0.8337**	**0.8178**	**0.8911**	**0.8496**
DUTS-TE [39]	1	0.0496	0.8560	0.8386	0.9025	0.8363
	2	0.0310	0.9071	0.8762	0.9313	0.8969
	3	0.0389	0.8816	0.8690	0.9144	0.8783
	4	**0.0281**	**0.9123**	**0.8930**	**0.9368**	**0.9049**
HKU-IS [20]	1	0.0504	0.9028	0.8864	0.9344	0.8649
	2	0.0266	0.9453	0.9184	0.9621	0.9205
	3	0.0383	0.9231	0.9102	0.9497	0.9024
	4	**0.0234**	**0.9483**	**0.9319**	**0.9657**	**0.9274**
PASCAL-S [23]	1	0.0868	0.8525	0.8373	0.8611	0.8166
	2	0.0576	0.8856	0.8567	0.9059	0.8652
	3	0.0685	0.8726	0.8605	0.8728	0.8587
	4	**0.0535**	**0.8896**	**0.8701**	**0.9131**	**0.8721**
ECSSD [48]	1	0.0673	0.9163	0.8988	0.9065	0.8617
	2	0.0344	0.9511	0.9298	0.9484	0.9213
	3	0.0503	0.9336	0.9208	0.9205	0.9016
	4	**0.0307**	**0.9549**	**0.9422**	**0.9544**	**0.9279**
SOD [28]	1	0.1449	0.8388	0.7809	0.7497	0.7028
	2	0.0895	0.8700	0.8640	0.8352	0.8063
	3	0.1147	0.8548	0.8277	0.7795	0.7666
	4	**0.0874**	**0.8780**	**0.8684**	**0.8381**	**0.8089**

Table 3. Comparison of MENet with different iterative enhancement times (denoted by **ITimes**). As shown by the bolded results, MENet's optimal enhancement is four iterations.

Dataset	No.	$\mathcal{L}_g$	$\mathcal{L}_{s_{bce}}$	$\mathcal{L}_{s_{reg}}$	$\mathcal{L}_{s_{obj}}$	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$
OMRON [49]	1		✓			0.0518	0.8141	0.7933	0.8684	0.8407
	2		✓	✓4		0.0498	0.8075	0.7892	0.8632	0.8388
	3		✓	✓4	✓	0.0492	0.8281	0.8093	0.8840	0.8437
	4	✓	✓			0.0486	0.8221	0.8046	0.8762	0.8362
	5	✓	✓	✓4		0.0470	0.8190	0.8017	0.8762	0.8451
	6	✓	✓	✓4	✓	**0.0450**	**0.8337**	**0.8178**	**0.8911**	**0.8496**
	7	✓	✓	✓1	✓	0.0472	0.8156	0.8027	0.8713	0.8339
DUTS-TE [39]	1		✓			0.0308	0.9030	0.8816	0.9288	0.8991
	2		✓	✓4		0.0305	0.8978	0.8755	0.9247	0.8945
	3		✓	✓4	✓	0.0295	0.9097	0.8871	0.9339	0.9007
	4	✓	✓			0.0301	0.9049	0.8797	0.9285	0.8935
	5	✓	✓	✓4		0.0295	0.9053	0.8856	0.9319	0.8993
	6	✓	✓	✓4	✓	**0.0281**	**0.9123**	**0.8930**	**0.9368**	**0.9049**
	7	✓	✓	✓1	✓	0.0295	0.9060	0.8887	0.9302	0.8980
HKU-IS [20]	1		✓			0.0237	0.9453	0.9269	0.9639	0.9266
	2		✓	✓4		0.0259	0.9394	0.9209	0.9584	0.9199
	3		✓	✓4	✓	0.0252	0.9450	0.9269	0.9621	0.9220
	4	✓	✓			0.0283	0.9411	0.9206	0.9555	0.9140
	5	✓	✓	✓4		0.0250	0.9438	0.9271	0.9605	0.9226
	6	✓	✓	✓4	✓	**0.0234**	**0.9483**	**0.9319**	**0.9657**	**0.9274**
	7	✓	✓	✓1	✓	0.0255	0.9434	0.9247	0.9622	0.9223
PASCAL-S [23]	1		✓			0.0572	0.8794	0.8608	0.9026	0.8663
	2		✓	✓4		0.0552	0.8836	0.8639	0.9054	0.8681
	3		✓	✓4	✓	0.0565	0.8839	0.8653	0.9100	0.8670
	4	✓	✓			0.0606	0.8831	0.8619	0.8985	0.8587
	5	✓	✓	✓4		0.0557	0.8865	0.8674	0.9055	0.8652
	6	✓	✓	✓4	✓	**0.0535**	**0.8896**	**0.8701**	**0.9132**	**0.8721**
	7	✓	✓	✓1	✓	0.0576	0.8845	0.8652	0.9024	0.8678
ECSSD [48]	1		✓			0.0308	0.9524	0.9374	0.9542	0.9252
	2		✓	✓4		0.0315	0.9494	0.9343	0.9526	0.9243
	3		✓	✓4	✓	**0.0297**	0.9536	0.9384	0.9554	**0.9290**
	4	✓	✓			0.0344	0.9477	0.9310	0.9484	0.9193
	5	✓	✓	✓4		0.0305	0.9537	0.9393	0.9544	0.9267
	6	✓	✓	✓4	✓	0.0307	**0.9549**	**0.9422**	0.9545	0.9279
	7	✓	✓	✓1	✓	0.0326	0.9514	0.9247	**0.9622**	0.9223
SOD [28]	1		✓			0.0910	0.8772	**0.8707**	0.8307	0.8024
	2		✓	✓4		0.0910	0.8595	0.8543	0.8137	0.7987
	3		✓	✓4	✓	**0.0841**	0.8648	0.8595	0.8250	**0.8089**
	4	✓	✓			0.0947	0.8725	0.8650	0.8150	0.7949
	5	✓	✓	✓4		0.0886	0.8667	0.8608	0.8133	0.8019
	6	✓	✓	✓4	✓	0.0874	**0.8780**	0.8684	**0.8381**	**0.8089**
	7	✓	✓	✓1	✓	0.0944	0.8729	0.8675	0.8171	0.7994

Table 4. Ablation tests for loss settings. Symbol ✓ means the loss is calculated. For $\mathcal{L}_{s_{reg}}$, '✓1' and '✓4' represent one region and four sub-regions, respectively. The best results marked in bold demonstrate that the No.6 combination of losses is optimal for the six datasets.

Acknowledgements. This work is partially supported by the National Natural Science Foundation of China (61976037 and U22B2052) and the Yongjiang Technology Innovation Project (2022A-097-G).

References

[1] Illya Bakurov, Marco Buzzelli, Raimondo Schettini, Mauro Castelli, and Leonardo Vanneschi. Structural similarity index (ssim) revisited: A data-driven approach. *Expert Syst. Appl.*, 189:116087, 2022. 5

[2] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012. 5

[3] Tao Chen, Yazhou Yao, Lei Zhang, Qiong Wang, Guosen Xie, and Fumin Shen. Saliency guided inter-and intra-class relation constraints for weakly supervised semantic segmentation. *IEEE TMM*, 2022. 1

[4] Zhe Chen, Ruili Wang, Zhen Zhang, Huibin Wang, and Lizhong Xu. Background–foreground interaction for moving object detection in dynamic scenes. *Inf. Sci.*, 483:65–81, 2019. 1

[5] Ming-Ming Cheng and Deng-Ping Fan. Structure-measure: A new way to evaluate foreground maps. *IJCV*, 129(9):2622–2638, 2021. 2, 5, 6

[6] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *WACV*, pages 3560–3569, 2021. 2

[7] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Ann. Oper. Res.*, 134(1):19–67, 2005. 2

[8] D.P. Fan, C. Gong, Y. Cao, B. Ren, M.M. Cheng, and A. Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, page 698–704, 2018. 6

[9] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, pages 186–202, 2018. 1

[10] Deng-Ping Fan, Jing Zhang, Gang Xu, Ming-Ming Cheng, and Ling Shao. Salient objects in clutter. *IEEE TPAMI*, 45(2):2344–2366, 2022. 1

[11] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *CVPR*, pages 1623–1632, 2019. 1, 2, 3, 6

[12] Lucas Fidon, Wenqi Li, Luis C Garcia-Peraza-Herrera, Jinendra Ekanayake, Neil Kitchen, Sébastien Ourselin, and Tom Vercauteren. Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In *International MICCAI brainlesion workshop*, pages 64–76. Springer, 2017. 2

[13] Ashish Kumar Gupta, Ayan Seal, Mukesh Prasad, and Pritee Khanna. Salient object detection techniques in computer vision—a survey. *Entropy*, 22(1174):1–49, 2020. 1, 3

[14] Filiz Gurkan, Llukman Cerkezi, Ozgun Cirakman, and Bilge Gunsel. Tdiot: Target-driven inference for deep video object tracking. *IEEE TIP*, 30:7938–7951, 2021. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 6

[16] Xiaowei Hu, Chi Wing Fu, Lei Zhu, Tianyu Wang, and Pheng Ann Heng. Sac-net: Spatial attenuation context for salient object detection. *IEEE TCSVT*, 31(3):1079–1090, 2021. 1, 2, 3, 6, 7

[17] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *CVPR*, pages 4713–4722, 2022. 1

[18] Kurt Koffka. *Principles of Gestalt psychology*. Routledge, 2013. 2

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Adv. neural inf. proces. syst.*, volume 25, pages 1097–1105, Red Hook, NY, USA, 2012. 5

[20] Guanbin Li and Yizhou Yu. Visual saliency based on multi-scale deep features. In *CVPR*, pages 5455–5463, 2015. 5, 6, 8

[21] Linghui Li, Sheng Tang, Yongdong Zhang, Lixi Deng, and Qi Tian. Gla: Global–local attention for image description. *IEEE TMM*, 20(3):726–737, 2017. 4

[22] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *ECCV*, pages 355–370, 2018. 1, 2

[23] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014. 5, 6, 8

[24] J. J. Liu, Q. Hou, M. M. Cheng, J. Feng, and J. Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, page 3917–3926, 2019. 1

[25] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1

[26] Yuxin Mao, Jing Zhang, Zhexiong Wan, Yuchao Dai, Aixuan Li, Yunqiu Lv, Xinyu Tian, Deng-Ping Fan, and Nick Barnes. Generative transformer for accurate and reliable salient object detection. *arXiv e-prints*, pages arXiv–2104, 2021. 3

[27] Ran Margolin, Lihi Zelnik Manor, and Ayellet Tal. How to evaluate foreground maps. In *CVPR*, pages 248–255, Columbus, OH, USA, 2014. 6

[28] Vida Movahedi and James H. Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPRW*, pages 49–56, 2010. 6, 8

[29] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012. 6

[30] S Plainis and IJ Murray. Neurophysiological interpretation of human visual reaction times: effect of contrast, spatial frequency and luminance. *Neuropsychologia*, 38(12):1555–1564, 2000. 2

[31] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *PR*, 106:107404, 2020. 3, 6

[32] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7471–7481, 2019. 2, 3, 6

[33] Shuang Qiu, Yao Zhao, Jianbo Jiao, Yunchao Wei, and Shikui Wei. Referring image segmentation by generative adversarial learning. *IEEE TMM*, 22(5):1333–1344, 2020. 1

[34] Yu Qiu, Yun Liu, Yanan Chen, Jianwen Zhang, Jinchao Zhu, and Jing Xu. A2sppnet: Attentive atrous spatial pyramid pooling network for salient object detection. *IEEE TMM*, pages 1–1, 2022. 2, 4

[35] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on Vis. Comput.*, pages 234–244. Springer, 2016. 2, 5

[36] Qinghua Ren, Shijian Lu, Jinxia Zhang, and Renjie Hu. Salient object detection by fusing local and global contexts. *IEEE TMM*, 23:1442–1453, 2021. 1, 3, 6

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In *MICCAI*, page 234–241, 2015. 2, 3

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[39] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 3796–3805, 2017. 5, 6, 8

[40] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE TPAMI*, 44(6):3239–3259, 2021. 1, 3

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 2, 5

[42] Jun Wei, Shuhui Wang, and Qingming Huang. F^{3}net: fusion, feedback and focus for salient object detection. In *AAAI*, volume 34, pages 12321–12328, 2020. 2

[43] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *CVPR*, pages 13022–13031, 2020. 1, 2, 3, 6, 7

[44] Runmin Wu, Mengyang Feng, Wenlong Guan, Dong Wang, Huchuan Lu, and Errui Ding. A mutual learning method for salient object detection with intertwined multi-supervision. In *CVPR*, pages 8142–8151, 2019. 3, 6

[45] Yu-Huan Wu, Yun Liu, Le Zhang, Ming-Ming Cheng, and Bo Ren. Edn: Salient object detection via extremely-downsampled network. *IEEE TIP*, 31:3125–3136, 2022. 2, 3, 6, 7

[46] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3902–3911, 2019. 3, 6

[47] Binwei Xu, Haoran Liang, Ronghua Liang, and Peng Chen. Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In *AAAI*, volume 35, pages 3004–3012, 2021. 1, 2, 3, 6, 7

[48] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013. 6, 8

[49] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013. 5, 6, 8

[50] Lihe Zhang, Jie Wu, Tiantian Wang, Ali Borji, Guohua Wei, and Huchuan Lu. A multistage refinement network for salient object detection. *IEEE TIP*, 29:3534–3545, 2020. 3, 6, 7

[51] Yuan-fang Zhang, Jiangbin Zheng, Wenjing Jia, Wenfeng Huang, Long Li, Nian Liu, Fei Li, and Xiangjian He. Deep rgb-d saliency detection without depth. *IEEE TMM*, 24:755–767, 2021. 2

[52] Zongjian Zhang, Qiang Wu, Yang Wang, and Fang Chen. Exploring pairwise relationships adaptively from linguistic context in image captioning. *IEEE TMM*, 24:3101–3113, 2022. 1

[53] Jia Xing Zhao, Jiang Jiang Liu, Deng Ping Fan, Yang Cao, Jufeng Yang, and Ming Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, pages 8779–8788, 2019. 1, 2, 3, 6

[54] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, pages 23–28, 2020. 3, 6

[55] Li Zhaoping. *Understanding vision: theory, models, and data*. OUP Oxford, 2014. 1

[56] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *CVPR*, pages 9141–9150, 2020. 2

[57] Lei Zhu, Jiaxing Chen, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Aggregating attentional dilated features for salient object detection. *IEEE TCSVT*, 30(10):3358–3371, 2020. 1, 2, 3, 6, 7

[58] Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE TIP*, 30:948–962, 2021. 1

[59] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE TPAMI*, 2022. 2, 3, 6, 7

[60] Ming Zong, Ruili Wang, Xiubo Chen, Zhe Chen, and Yuanhao Gong. Motion saliency based multi-stream multiplier resnets for action recognition. *Image Vis Comput.*, 107:104108, 2021. 1