MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Salient Object Detection for Complex Scenes

A thesis presented in partial fulfilment of the
requirements for the degree of

Doctor of Philosophy
in
Computer Science

School of Mathematical and Computational Sciences,
Massey University, Albany, Auckland,
New Zealand

Yi Wang

September 2024

Knowledge is the beginning of action,
and action is the completion of knowledge.

–Yangming Wang

Abstract

Salient Object Detection (SOD), a primary objective in computer vision, aims to locate and segment the region most visually striking within an image. In this thesis, we present three innovative methods based on deep learning to improve SOD performance in complex scenes.

Firstly, we introduce the Multiple Enhancement Network (MENet), inspired by boundary perception, gradual enhancement, frequency decomposition, and content integrity of the Human Visual System (HVS). We propose a flexible multi-scale feature enhancement module to aggregate and refine features and use iterative training to improve boundary and adaptive features in the dual-branch decoder of MENet. A multi-level hybrid loss guides the network in learning pixel-, region-, and object-level features. Evaluations of benchmark datasets show that MENet outperforms other SOD models, especially when the salient region has multiple objects with varied appearances or complex shapes.

Secondly, we propose TFGNet, an effective frequency-guided network for saliency detection based on Transformer. TFGNet has a parallel two-branch decoder, which leverages a pixel-wise decoder and a Transformer decoder to optimise high-spatial frequency boundary details and low-spatial frequency salient features. A novel loss is also designed to use frequency distribution similarity measurement to further improve performance. The experimental results indicate that TFGNet can accurately locate salient objects with more complete and precise boundaries on various complex backgrounds. This framework also rekindles awareness of the advantages of exploiting images' spatial frequency features in SOD.

Thirdly, we design a multi-source weakly supervised SOD (WSOD) framework that can effectively utilise pseudo-background (non-salient region) labels combined with scribble labels to obtain more accurate salient features. We first create a comprehensive salient pseudo-mask generator from multiple self-learning features. Also, we pioneer the exploration of generating salient pseudo-labels via point-prompted and box-prompted Segment-Anything Models (SAM). Then, a Transformer-based WSOD network named WBNet is proposed, which leverages pixel-decoder and transformer-decoder with auxiliary edge predictor with multi-source loss function to handle complex saliency detection tasks.

In summary, we contribute three novel approaches to address salient object detection in complex scenes. Each model achieves cutting-edge performance across prestigious datasets validated through comprehensive experiments.

Acknowledgements

First and foremost, I would like to express my gratitude to everyone who has supported and helped me during my doctoral studies.

I extend my heartfelt gratitude and utmost respect to my principal supervisor, Professor Ruili Wang. He epitomized the qualities of a scientist, guiding me with meticulous care and setting a shining example for me to follow. Under his mentorship, I gained a deep understanding of the rigor required in scientific research and the importance of perseverance. His personal guidance and exemplary conduct have been invaluable to me.

I am also indebted to my esteemed co-authors, Dr. Tianzhu Wang, Dr. Tao Jiang, and Prof. Xiangjian He, for their collaboration, expert feedback, and encouragement. Their contributions have added depth and richness to my research and broadened my scholarly horizons.

Lastly, I thank all the authors of the references cited in this work, whose ideas and methods have been a great source of inspiration and guidance for this study.

Publications

During my doctoral studies, the following research papers have been published in (or submitted to) international journals and conferences:

1. **Yi Wang**, Ruili Wang*, Xin Fan, Tianzhu Wang, and Xiangjian He, Pixels, Regions, and Objects: Multiple Enhancement for Salient Object Detection, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023:10031-10040.
   URL: http://doi.org/10.1109/CVPR52729.2023.00967 (CORE Rank A*)

2. **Yi Wang**, Ruili Wang*, Xiangjian He, Tianzhu Wang, WBNet: Weakly-Supervised Saliency Detection via Scribble and Pseudo-Background Priors, Pattern Recognition, 2024, 154, 11057.
   URL: https://doi.org/10.1016/j.patcog.2024.110579 (CORE Rank A)

3. **Yi Wang**, Ruili Wang*, Tao Jiang, Tianzhu Wang, Frequency-Guided Saliency Detection for Complex Scenes, Applied Soft Computing (Revision)

4. Jiang Tao, **Yi Wang***, Hou Feng, Ruili Wang, IENet: inheritance enhancement network for video salient object detection. Multimedia Tools and Applications, 2024, 83(28):1-20.
   URL: https://doi.org/10.1007/s11042-024-18408-4

5. Xiaofang Li, **Yi Wang***, Tianzu Wang, Ruili Wang, Spatial frequency enhanced salient object detection, Information Sciences, 2023, 647, 119460.
   URL: https://doi.org/10.1016/j.ins.2023.119460 (CORE Rank A)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter is an overview of the content and structure of the thesis. Section 1.1 delves into the background and previous studies on salient object detection. Section 1.2 discusses the motivation behind this research, highlighting the challenges existing approaches pose. The research objectives are delineated in Section 1.3. Lastly, Section 1.4 elucidates the organisation of this thesis.

1.1 Overview

Salient Object Detection (SOD) aims to locate and segment the most visually conspicuous region in an image by emulating the Human Visual System (HVS) [1, 2]. It is crucial to eliminate redundant information and improve computational efficiency in a variety of advanced computer vision applications, such as action recognition, image segmentation, image captioning, image editing, video compression, video summarization, and so on. SOD emphasises predicting overall salient objects/regions with exquisite boundaries, limited pre-processing, and low computational complexity.

The main challenge of learning salient features comes mainly from the different requirements for features in various parts of salient regions. In general, salient regions require globally consistent (invariant) features that can handle changes in appearance (such as texture, shape, colour, and size) in various scenes. In contrast, the salient region also needs features that can fully express the boundary details and has intense discrimination against backgrounds. Therefore, there is a contradiction in the feature learning of salient objects.

A U-Net-like encoder-decoder architecture [3] can handle these two contradictory requirements [1]. This architecture consists of an encoder and a decoder. The encoder uses convolution and down-sampling to determine the coarse location of salient objects and extract global semantic information. The decoder then refines the salient map by amalgamating features from diverse layers, generating a fine salient map. Skip connections, which link the corresponding layers in both paths, are crucial to capturing high-level semantic features and detailed spatial information. However, the U-Net framework is initially designed for semantic segmentation tasks (i.e., the pixel-level image classification task) [1, 3]. It employs various combinations of convolutions and pooling in the encoder; thus, the spatial resolution of feature maps decreases layer by layer. Although high-resolution features of the encoder are also introduced in the bottom-up decoding process, without proper guidance, erroneous global information may mislead the learning of boundary details [1].

To address these limitations and improve the accuracy of the SOD, researchers have explored various enhancement strategies. Some models have adopted a multi-stage enhancement approach to capture salient details across different scales, resulting in more comprehensive and accurate saliency maps [4–11]. Other methods have employed edge/contour supervision or designed multi-task learning to simultaneously detect edges and saliency [6, 10, 12, 13]. Transformer-based models have recently shown promise, surpassing CNN-based methods on standard benchmarks [14–17].

However, current SOD models still face various challenges when dealing with different types of complex scene, each of which presents unique difficulties for salient feature learning: (i) Small/Tiny Objects: Small objects are often difficult to detect because they only occupy a small portion of the image. The features that represent these objects can be easily overwhelmed by background noise or lost during down-sampling in the network; (ii) Large Objects: Due to their size, there is often a higher chance of encountering similar textures within different parts of the object or across the object and background. This can cause confusion in the model, leading to inaccurate saliency maps where the boundaries of the object are not well defined; (iii) Multiple Connected Objects: The boundaries between these objects may not be clear, leading to errors in detecting which parts of the image should be considered salient. (iv) Reflections: Reflections can provide misleading indicators in an image as they frequently duplicate the geometry, colour, and texture of an object. (v) Low Contrast: Subtle intensity and colour differences are hard to detect, causing the object to blend into the background,

resulting in poor saliency maps; (vi) Blurred Boundaries: Blurred boundaries from motion blur, depth of field, or soft edges make it difficult for SOD models to accurately delineate objects, causing errors in identifying the salient region. These challenges highlight the need for advanced techniques and strategies in SOD to better handle complex scenes and improve the accuracy of salient feature learning.

1.2 Motivation

To design a versatile SOD model that can handle various complex scene challenges robustly and accurately, we need to address common issues in saliency feature learning. In this thesis, we identify the following three key issues that need to be solved to achieve this goal:

- **Under-utilization of HVS mechanisms in SOD model design**. The HVS is complex and remarkable. It allows us to focus on salient objects in various environments. While SOD models based on deep learning have adopted some of HVS mechanisms, such as boundary sensitivity, content integrity, and iterative refinement. However, current SOD methods often fail to fully leverage these human visual and cognitive mechanisms to their advantage. This oversight reduces performance, particularly in complex scenes. Therefore, it is imperative to explore and harness human visual and cognitive mechanisms to enhance SOD methods effectively.
- **Interference from the mixture of frequency information of an image for salient object feature learning.** Accurately predicting boundaries in complex scenes presents a significant challenge. Many methods leverage edge information as an auxiliary tool or use a multi-task network to enhance boundary feature representation in such scenarios. However, these methods often incorporate a communication strategy between layers or branches, which can introduce errors propagating across layers or branches, leading to incorrect localization of salient objects and inaccurate boundary segmentation. Frequency decomposition, a technique that separates high-frequency components (e.g., edges and details) from low-frequency components, offers a potential solution. Therefore, a Transformer-based network architecture that effectively separates high and low-frequency saliency features while comprehensively expressing and enhanc-

ing both global and detailed salient features merits further exploration.

- **Bias and inconsistency in Weakly Supervised Salient Object Detection (WSOD) using sparse and multi-source pseudo Labels.** Sparse labels, such as scribbles, often provide only coarse annotations, lacking precise information about salient object locations and boundaries. Consequently, WSOD methods relying on these annotations may suffer from issues like over- or under-segmentation. To mitigate the challenges posed by limited supervision, single-source weakly supervised methods sometimes necessitate diverse pseudo annotations from different sources. However, these labels may introduce noise and uncertainty, complicating the label fusion process. Inconsistent labels can lead to conflicting information during training, potentially hindering the learning process. Addressing these challenges requires carefully considering and designing the label integration process and learning strategies.

1.3 Research Objectives

The objectives of this thesis are delineated as follows:

- **Objective 1** is to harness the full potential of HVS mechanisms to enhance saliency feature learning in complex scenes. This involves building a framework that effectively integrates holistic and continuous observation, boundary/contour and structural information sensitivity, and visual-spatial frequency decomposition to learn accurate and robust salient features.
- **Objective 2** is to minimize potential inaccuracies in localization and boundary prediction due to frequency mixing in the image. The objective is to create a concise Transformer-based network that leverages the image spatial frequency decomposition and synthesis mechanism, learns more precise and resilient frequency salient features, and subsequently merges these frequency features to produce the ultimate saliency features.
- **Objective 3** is to effectively harness and incorporate broader and more robust saliency cues from various pseudo annotations by developing an innovative Transformer-based weakly supervised SOD network. The aim is to include a pixel-level decoder, a Transformer decoder, and an auxiliary boundary predictor

into a network to predict saliency information with a comprehensive hybrid loss function utilizing scribble and pseudo-background labels.

1.4 Contributions

This thesis concentrates on two sub-problems of salient object detection: fully-supervised SOD (Chapter 2 and Chapter 3) and weakly supervised SOD (Chapter 4) in complex scenes. The contributions in each chapter are summarised as follows.

(i) Multiple enhancement network for accurate saliency detection.

- We propose to incorporate similarity measures at pixel, region, and object levels into the loss function to enhance prediction accuracy and robustness. Subsequently, we devise a hybrid loss function that integrates these measures at multiple levels.
- We create a multi-scale feature enhancement module (ME-Module) to emulate the holistic and progressive operations of the HVS mechanisms. The ME-Module can gradually generate global or detailed features by modifying the input order.
- We introduce a new Multiple Enhancement Network (MENet). MENet integrates multiple HVS mechanisms into the network structure, training strategy, and loss function. Specifically, MENet features a two-branch decoding process that progressively uses ME-Modules to refine boundary and adaptive features.

(ii) Frequency guidance framework for effective salient object detection.

- We present TFGNet, an efficient and concise Transformer-based Frequency Guidance Framework for SOD. TFGNet leverages the image spatial frequency decomposition and synthesis mechanism, learning frequency salient features in separate branches. Using a pixel-decoder and a Transformer-decoder, each branch learns comprehensive embeddings that lead to more accurate and robust predictions.
- We design a novel histogram-based loss function by using frequency distribution similarity measurement to improve TFGNet performance. To our knowledge, this is the first instance of the histogram-based being employed in a SOD model.

(iii) Effective weakly supervised SOD using multi-source pseudo annotations.

- We present a highly effective, weakly supervised SOD network named WBNet. A pixel-level decoder, a Transformer decoder, and an auxiliary boundary predictor are incorporated into this network to predict saliency information with a comprehensive hybrid loss function utilising scribble and pseudo-background labels.
- We design a self-learning feature-based pseudo-mask generator that utilises multi-source self-learning features, clustering techniques, and saliency-priors filtering strategies to produce comprehensive pseudo-masks that align consistently with scribble annotations.
- We employ the Segmentation Anything Module (SAM) to generate pseudo-masks for WSOD. We design two effective prompt generation methods: one relies on foreground scribble points, and the other leverages a bounding box derived from pseudo-labels obtained through self-learning methods.

1.5 Organization

The rest of this thesis is organised as follows.

Chapter 2 presents a multiple enhancement SOD network, which is based on our work published in the proceedings of *the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) 2023*, titled 'Pixels, Regions, and Objects: Multiple Enhancement for Salient Object Detection' [7].

Chapter 3 presents an effective frequency guidance transformer network for saliency detection and was submitted to the journal *Applied Soft Computing*, titled 'Frequency-Guided Saliency Detection for Complex Scenes'. (currently in revision).

Chapter 4 presents a new multi-source weakly supervised SOD framework that can effectively utilise pseudo-background (non-salient region) labels combined with scribble labels to obtain more accurate salient features, based on our work entitled titled 'Weakly-Supervised Salient Object Detection via Scribble and Pseudo-Background Priors' which is published in journal *Pattern Recognition*.

Chapter 5 summarises this thesis and offers insight into future research avenues.

Note that the reference list for each chapter is provided at the end of the respective chapters, and the Statements of Contributions are inserted at the beginning of each relevant chapter.

References

[1] 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 Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(6):3239–325, 2021.

[2] 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.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, page 234–241, 2015.

[4] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3902–3911, Long Beach, CA, USA, 2019.

[5] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13022–13031, Virtual, Online, USA, 2020.

[6] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3004–3012, 2021.

[7] Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[8] Lv Tang, Bo Li, Yijie Zhong, Shouhong Ding, and Mofei Song. Disentangled high quality salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3580–3590, 2021.

[9] Yuhuan Wu, Yun Liu, Le Zhang, Mingming Cheng, and Bo Ren. Edn: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing (TIP).*, 31:3125–3136, 2022.

[10] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, page 8779–8788, Seoul, Korea, Republic of, 2019.

[11] Yuming Fang, Haiyan Zhang, Jiebin Yan, Wenhui Jiang, and Yang Liu. Udnet: Uncertainty-aware deep network for salient object detection. *Pattern Recognition*, 134:109099, 2023.

[12] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7471–7481, Long Beach, CA, USA, 2019.

[13] Jiangjiang Liu, Qibin Hou, Mingming Cheng, Jiashi Feng, and Jiang Jianmin. A simple pooling-based design for realtime salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3917–3926, 2019.

[14] Mingchen Zhuge, Dengping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(3):3738–3772, 2023.

[15] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4722–4732, 2021.

[16] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. Learning generative vision transformer with energy-based latent space for saliency prediction. *Advances in Neural Information Processing Systems (ANIPS)*, 34:15448–15463, 2021.

[17] Yi Ke Yun and Weisi Lin. Selfreformer: Self-refined network with transformer for salient object detection. *arXiv preprint arXiv:2205.11283*, 2022.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Yi Wang
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 2

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, Xiangjian He. Pixels, Regions, and Objects: Multiple Enhance ment for Salient Object Detection, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR):10031-10040

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:
- The percentage of the manuscript/published work that was contributed by the candidate:
- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	wangyi 数字签名者：wangyi 日期：2024.04.02 11:47:30 +13'00'
Date:	02-4 月-2024
Primary Supervisor's Signature:	Prof Ruili Wang Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, c=NZ, o=Massey University , ou=School of Natural and Computational Sciences , email=ruili.wang@massey.ac.nz Date: 2024.04.02 17:19:03 +13'00'
Date:	2-4 月-2024

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 2

Multiple Enhancement Network for Saliency Detection

Salient Object Detection (SOD) aims to replicate the mechanism and cognition process of the Human Visual System (HVS) to identify and segment salient regions in an image. However, existing approaches are imperfect due to the under-utilisation of these mechanisms. To address this issue, we propose a novel approach, called the Multiple Enhancement Network (MENet), that integrates the boundary sensibility, content integrity, iterative refinement, and frequency decomposition mechanisms of the HVS. We first design a multilevel hybrid loss to guide the network in learning pixel-level, region-level, and object-level features. Next, we introduce a flexible multi-scale feature enhancement module that gradually aggregates and refines global or detailed features by altering the size order of the input feature sequence. We employ an iterative training strategy to enhance boundary and adaptive features in the dual-branch decoder of MENet. Extensive evaluations of six benchmark datasets demonstrate that MENet achieves state-of-the-art results. Note that the content presented in this chapter has been published in the CVPR 2023 conference.

2.1 Introduction

Salient Object Detection (SOD) aims to identify the regions most visually noticeable in an image that are consistent with the Human Visual System (HVS) and cognition mechanisms [1–3]. SOD can eliminate redundant information and improve computational performance for many high-level computer vision tasks, such as action recognition [4, 5], image segmentation [6, 7], image captioning [8], object tracking [9], and video summary [10]. Fully Convolutional Neural Networks (FCNs) [11] based SOD models have been particularly effective in improving SOD performance in recent years [1]. However, accurate segmentation of salient regions remains a challenging task for SOD, especially when the geometry and boundaries of these objects are complex or the scenes are chaotic or cluttered [3, 12].

An intuitive solution to address this problem is to explore the mechanisms of the Human Vision System (HVS) [13], some of which have been used in SOD models, as described below. (i) A human tends to improve recognition by alternating between viewing the whole and details in a complex scene, which has been used for various visual tasks [14–17]. (ii) HVS is sensitive to an object's boundary/contour and structure. Dual-branch feature refinement structures have been developed to incorporate extra-edge information to enhance salient feature learning [16, 18–23]. Some structural similarity measurements (e.g., Structural Similarity Index (SSIM) [24]) and regional similarity measurements (e.g., Intersection over Union (IoU) [25] and Dice [26]) are also adopted by SOD models [20, 22, 27–29] in loss functions. (iii) Human vision is holistic and continuous so that the HVS perceives an object as an organised whole [30], composed of meaningful and coherent parts with each other. ICON [28] proposes to improve the integrity from macro- and micro-level perspectives by improving the integrity information hidden in the feature channels. EDN [31] employs a powerful down-sampling technique to effectively learn a global view of the entire image. (iv) According to the human visual spatial frequency model [32], images can be decomposed into or synthesised into high- and low-spatial frequency parts. As a starting point in this work, we intend to use the mechanisms outlined above to further improve the SOD performance for complex scenes further.

In this chapter, we propose a Multiple Enhancement Network (MENet) that effectively integrates the above HVS mechanisms into a U-Net-like [33] 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 existing two-branch (or edge-aware) methods [16, 18, 19, 21, 22, 34, 35] that use one branch for the boundary and the other one for the entire object, such as EGNet [18] and AFNet [19]. In particular, there is no interaction between the intermediate features of the two branches of MENet, so it reduces the interference of inaccurate boundary local information with global features. Although LDF [20] 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 Multi-scale Feature Enhancement Module (ME-Module) as the core of each branch by leveraging the Atrous Spatial Pyramid Pooling (ASPP) [36] and global-local attention [37].

In addition, we introduce the HVS holistic and continuous mechanism to the design of loss functions. We present a multilevel hybrid loss that evaluates the pixel, region, and object level similarities between predicted and ground-truth (GT) saliency maps. For loss at the pixel level, we also use Binary Cross Entropy (BCE) [38] loss to ensure network accuracy and convergence. 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 [39] and IoU [25]. We are then inspired by SSIM and the S-measure [40], 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 [29], but it uses a simple combination of BCE, IoU, and SSIM for the whole saliency map without partitioning regions. The following is a summary of our contributions.

- We propose to leverage not only the pixel level, but also the region- and object-level similarity measures in loss to increase prediction accuracy and integrity and then design a multilevel hybrid loss to implement this proposal.
- We design a Multi-scale Feature Enhancement Module (ME-Module) to mimic HVS bottom-up and top-down refinement mechanisms. ME-Module can gradu-

ally 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 functions. Specifically, a two-branch decoder equipped with the ME-Module 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 that MENet outperforms the state-of-the-art methods by a large margin.

The rest of this chapter is structured as shown below: Section 2.2 briefly reviews related SOD approaches in this chapter. Section 2.3 explains the details of MENet. Section 2.4 demonstrates and discusses the proposed method through quantitative and qualitative experiments. Section 2.5 outlines the contributions of this chapter.

2.2 Related Work

In recent years, SOD approaches based on encoder-decoder and feature aggregation architecture have achieved high performance [1, 2, 41]. In the following, we briefly review models related to this chapter.

MLMSNet [42] takes advantage of the supervision of foreground boundary detection and edge detection. AFNet [19] develops an attentive feedback module to better explore the target structure. EGNet [18] uses complementary information about salient edges and objects to propose an edge guidance network. CPDNet [43] proposes a partial decoder to refine high-level features to generate precise saliency maps. BASNet [29] sequentially stacks two U-Nets[33] with different configurations. AADFNet [21] uses an attentional dense ASPP-based network to selectively use small and large dilated rate convolutions to obtain local and global saliency information. GateNet [44] designs a gated dual-branch structure to establish a cooperative relationship between features of different levels to increase network discriminability. U2Net [45] proposes a novel ReSidual U-block (RSU), which can obtain multi-resolution features in the intrastate without reducing the resolution of the feature map. MINet [46] proposes to enhance the feed-forward neural network by adopting

a refinement mechanism for multiple stages. LDF [20] designs a two-branch decoder to predict saliency maps using body complement and detailed information of the objects. SAC [17] implements a spatial attenuation context module to propagate and aggregate salient features through two rounds of recurrent translations. CANet [14] 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 [22] uses intermediate edge supervision in its coarse location module. ICON [28] introduces three different aggregations of features, enhancement of the integrity channel, and verification of the entire SOD. EDN [31] uses an extreme downsampling 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 salient objects' boundaries (or contours). Our work draws inspiration from the above-mentioned methods (e.g., two-branch decoders, ASPP, attention, and iterative refinement), but our model employs these methods uniquely.

2.3 Methodology

2.3.1 Framework Overview

The proposed MENet utilises the encoder-decoder structure, as shown in Fig. 2.1. The initial encoding part consists of a backbone (e.g., ResNet-50 [47]), 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 supervision ($\mathcal{L}_g^{(k)}$) and adaptive features (i.e., inner body features) $EaB{=}\{Eab_i\}_{i=1}^{N}$, respectively. Afterward, EgB and EaB concatenate 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 channels again and 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 $\mathcal{L}_s$.

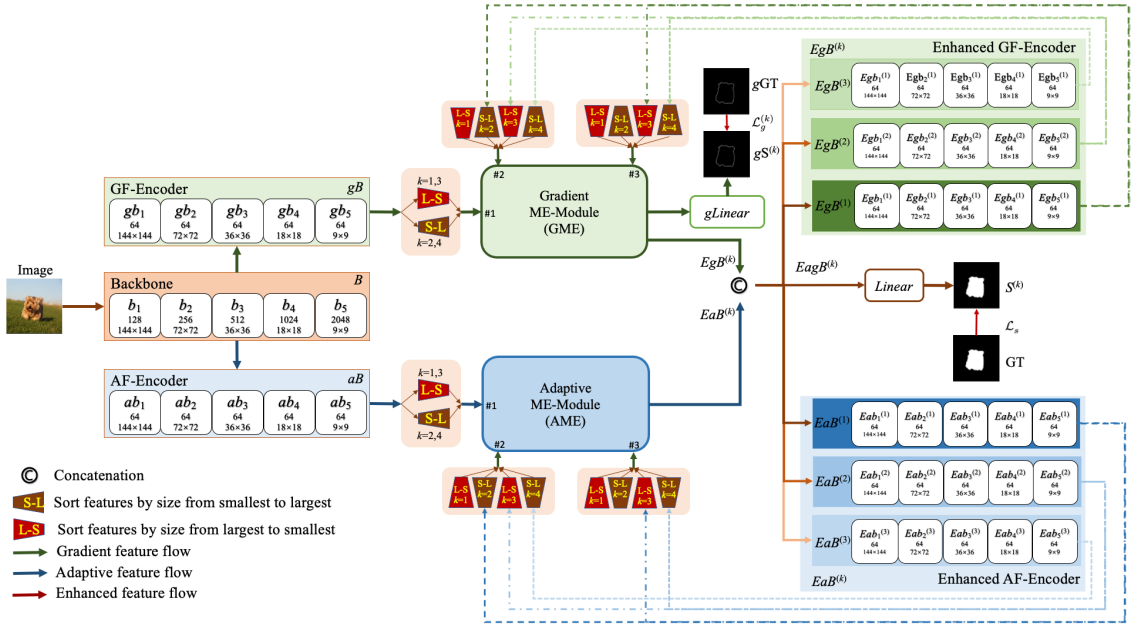


Figure 2.1: Illustration of the overall architecture and the pipeline of the MENet.

2.3.2 Multi-Scale Feature Enhancement

Figure 2.2 demonstrates the composition of the ME-Module. Using ASPP [36] and ME-Attention, the 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 focusses on diversifying visual fields, while the ME-Attention module emphasises the location of a salient object through global and detailed attention [48]. An 'interpolator performs up-sampling and down-sampling operations accordingly, to match the scale of features of the next stage. The ME-Module is versatile because it can output global or detailed features simply by adjusting the spatial order of multi-scale feature maps.

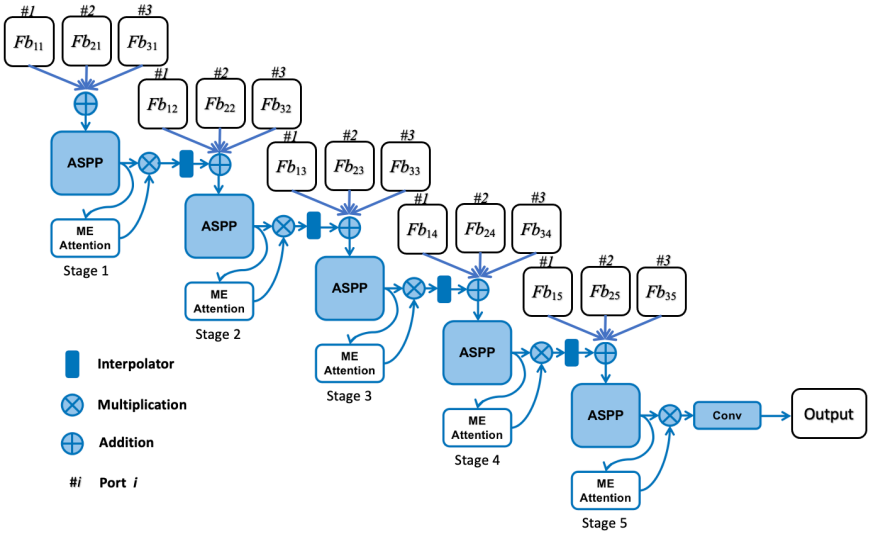


Figure 2.2: Illustration of the proposed ME-Module.

2.3.3 Iterative Enhancement

Because the ME module can be used as a global feature enhancer or a detailed feature enhancer, we can alternately change the input sequence of the feature maps and make them complement each other.

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 indicated as $EgB^{(k)}$ and $EaB^{(k)}$ ($k \in [1, 4]$), respectively. The entire enhancement process is summarised in Eqs.(2.1),(2.2),(2.3). The pipeline is illustrated in Fig. 2.1.

$$EgB^{(k)} = GME(\, V(gB),\, V(EgB^{(k-1)}),\, V(EgB^{(k-2)})), \tag{2.1}$$

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 $p_i = Null$.

$$EaB^{(k)} = AME(\, V(aB),\, V(EaB^{(k-1)}),\, V(EaB^{(k-2)})), \tag{2.2}$$

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

$$S^{(k)} = Linear(Concat(EgB^{(k)}, EaB^{(k)}), \tag{2.3}$$

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

Figure 2.3 shows two visual examples of the saliency map $S^{(k)}$ in each iteration. The odd-number iteration extracts global features, then refined by the even-number iterations. Thus, the global and detailed features complement and promote each other.

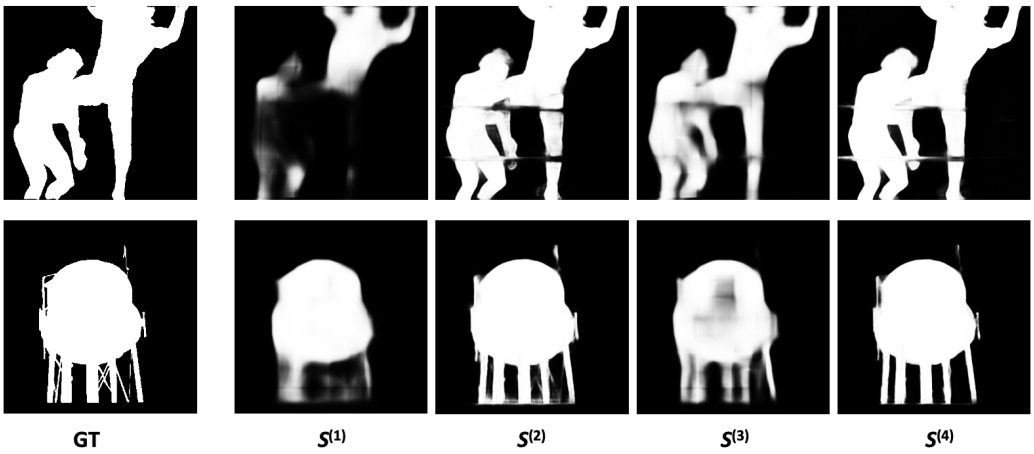


Figure 2.3: Examples of the saliency maps in each iteration.

2.3.4 Supervision Strategy

In the MENet, we set two supervisions in each enhancement round, as shown in Fig. 2.1. 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 in Eq. (2.4).

$$\mathcal{L} = \sum_{k=1}^{4} ({\alpha_1}^{(k)} \mathcal{L}_g^{(k)} + {\alpha_2}^{(k)} \mathcal{L}_s^{(k)}), \tag{2.4}$$

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 in Eq. (2.5)

$$\mathcal{L}_g = -\sum(G_g log S_g + (1 - G_g) log(1 - S_g)), \tag{2.5}$$

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 guide 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 in Eq. (2.6).

$$\mathcal{L}_s = \beta_1 \mathcal{L}_{s_{bce}} + \beta_2 \mathcal{L}_{s_{reg}} + \beta_3 \mathcal{L}_{s_{obj}}, \tag{2.6}$$

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 Eq. (2.7)

$$\mathcal{L}_{s_{bce}} = -\sum(G log S + (1 - G) log(1 - S)), \tag{2.7}$$

Region-Level Loss: We design $L_{s_{Reg}}$ based on a structural measure SSIM [24] and a regional measure IoU [25] of images. We divide S and G evenly into four equal sub-regions (i.e., S_i and G_i, $i \in [1, 4]$), which can be processed as a batch in implementation. Then $\mathcal{L}_{Reg}$ can be represented by Eq. (2.8).

$$\mathcal{L}_{s_{reg}} = 1 - \sum_{i=1}^{4} \omega_i (\theta_1 SSIM_i + \theta_2 IoU_i), \tag{2.8}$$

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 be formulated as a product of three components [24, 39]: the luminance comparison, the contrast comparison and the structure comparison. The $SSIM_i$ is defined in Eq. (2.9).

$$SSIM_i = \frac{2\mu_{S_i}\mu_{G_i} + c_1}{\mu_{S_i}^2 + \mu_{G_i}^2 + c_1} \cdot \frac{2\sigma_{S_i}\sigma_{G_i} + c_2}{{\sigma_{S_i}}^2 + {\sigma_{G_i}}^2 + c_2} \cdot \frac{\sigma_{S_i G_i} + c_3}{\sigma_{S_i}\sigma_{G_i} + c_3}, \tag{2.9}$$

where μ_{S_i} and μ_{G_i} are the means, σ_{S_i} and σ_{G_i} are the standard deviations, and $\sigma_{S_i G_i}$

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 [39].

Furthermore, we also measure the overlap of spatial regions between predictions and labels using the IoU in $L_{s_{Reg}}$, which is defined in Eq. (2.10).

$$IoU_i = \frac{\sum\sum(S_i G_i)}{\sum\sum(G_i + S_i - G_i S_i)}. \tag{2.10}$$

Object-Level Loss: Inspired by SSIM and the S measurement [40], we introduce the measurement of image similarity at the object level $\mathcal{L}_{s_{obj}}$. Since GT maps usually have sharp foreground-background contrast and uniform distribution, the predicted saliency maps should also have these properties [24, 40]. 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 an accurate foreground is the primary objective in training the whole network, we only consider the foregrounds (denoted S_o and G_o, respectively) of S and G and compute the foreground distribution. This difference from the S-measure considers both the distributions of the foregrounds and the backgrounds of S and G. Thus, $L_{s_{obj}}$ is defined in Eq. (2.11).

$$\mathcal{L}_{s_{obj}} = 1 - \frac{1}{(\frac{\mu_{S_o}^2 + \mu_{G_o}^2}{2\mu_{S_o}\mu_{G_o}} + \lambda\frac{\sigma_{S_o}}{\mu_{S_o}})}, \tag{2.11}$$

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. (2.11) can be simplified in Eq. (2.12).

$$\mathcal{L}_{s_{obj}} = 1 - \frac{2\mu_{S_o}}{\mu_{S_o}^2 + 1 + 2\lambda\sigma_{S_o}}. \tag{2.12}$$

2.4 Experiment and Discussion

2.4.1 Training and Testing Setting

We use ImageNet [49] to pre-train the backbone network and then use the DUTS-TR [50] to fine-tune the proposed MENet. Other MENet parameters are initialised randomly in a normal distribution. The inputs are scaled to $[352 \times 352]$, $[320 \times 320]$,

$[288 \times 288]$, $[256 \times 256]$, and $[224 \times 224]$ for data augmentation. We use the stochastic gradient descent (SGD) optimizer [51] and set the maximum learning rate of the backbone network to 0.00025 and the other parts to 0.0025. The momentum is set at 0.9, the weight decay is set at 0.0005, and the batch size is set at 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 test image scaled to $[352 \times 352]$ takes just 0.022 s (45 fps).

2.4.2 Testing Dataset

We assess all methods on the six SOD benchmark datasets listed below.

DUT-OMRON dataset [52]: comprises 5,168 images of diverse objects against complex backgrounds.

DUTS-TE dataset [50]: a subset of the DUTS dataset, comprises 5,019 images from the ImageNet DET test set and the SUN dataset, encompassing highly challenging scenarios.

HKU-IS dataset [53]: has 4,447 images that contain multiple scattered salient regions, and/or at least one prominent boundary region, and/or exhibiting background similarities.

ECSSD dataset [54]: comprises 1,000 semantically meaningful images with complicated structures and complex backgrounds.

PASCAL-S dataset [55] has 850 images. It is recognized for having less bias compared to many other salient datasets.

2.4.3 Evaluation Criteria

We extensively assess the proposed method using the following commonly employed metrics [56].

Mean Absolute Error (MAE) [57] evaluates the difference between the ground truth map (G) and the predicted map (S) at the average pixel level and can be

formulated in Eq. (2.13).

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{i=1}^{H} |G_{ij} - S_{ij}|, \tag{2.13}$$

where W and H denote input image dimensions. A smaller MAE value is better.

Enhanced Alignment Measure (E_m) [58] computes the pixel-level matching and image-level statistics for the foregrounds of the predicted salient map in Eq. (2.14).

$$E_m = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{i=1}^{H} \phi_s(i,j), \tag{2.14}$$

where $\phi(\ ,\ )$ is an enhanced alignment matrix. We report the average E_m (denoted mE_m) in the experiments.

Structure Measure (S_m) [40] measures how similar the predicted map is to the GT map Eq. (2.15).

$$S_m = \alpha \times S_o + (1 - \alpha) \times S_r. \tag{2.15}$$

Here, S_o and S_r denote region-specific and object-specific structural similarities, respectively, with α usually set to 0.5. S_m as a structure-wise measure complements pixel-wise errors.

F-Measure (F_β) [59] is beneficial for unbalanced datasets and can be mathematically represented in Eq. (2.16).

$$F_\beta = \frac{(1 + \beta^2) Precision \times Recall}{\beta^2 Precision + Recall}. \tag{2.16}$$

We report the maximum F_β (denoted F_β^{max}) and the mean F_β (denoted mF) in the experiments.

We also use Precision Recall (PR) curves and Fm curves to assess the performance of the classification model. The surveys [1, 56] provide detailed descriptions.

Table 2.1: Quantitative comparison on the DUT-OMRON and DTUS-TE datasets.

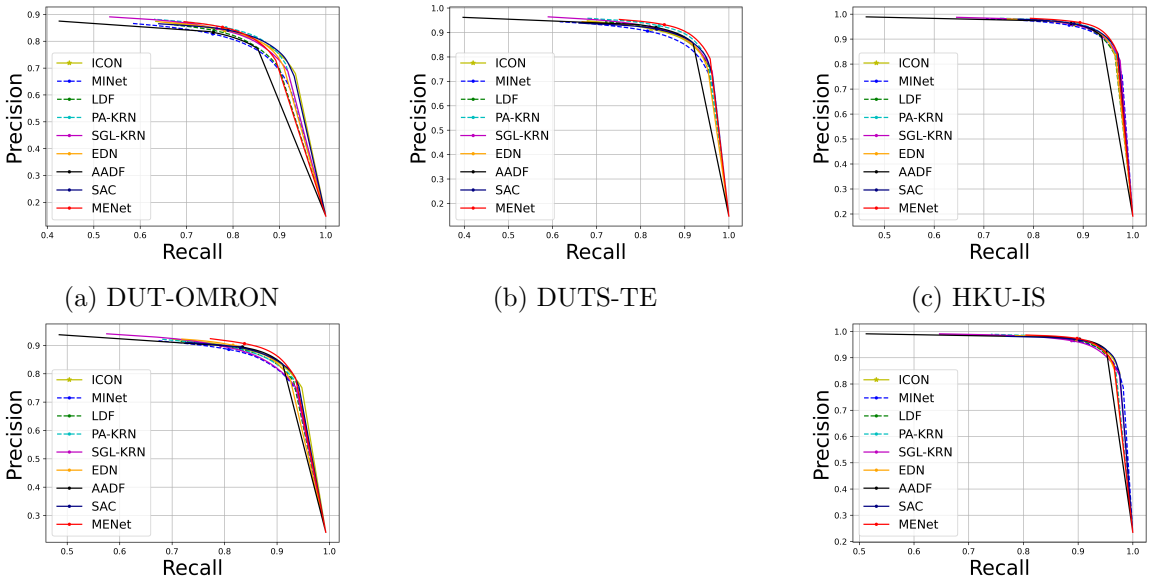
			DUT-OMRON [52] (5,168 images)					**DUTS-TE** [50] (5,019 images)				
Year	**Method**	**Backbone**	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$
2019	MLMSNet [42]	VGG-16	0.0635	0.7740	0.7455	0.8387	0.8093	0.0484	0.8511	0.8137	0.8631	0.8618
2019	AFNet [19]	VGG-16	0.0573	0.7972	0.7766	0.8595	0.8263	0.0453	0.8623	0.8340	0.8929	0.8672
2019	EGNet [18]	VGG-16	0.0564	0.8087	0.7855	0.8642	0.8357	0.0431	0.8764	0.8472	0.8927	0.8786
2019	EGNet [18]	ResNet-50	0.0528	0.8155	0.7942	0.8738	0.8412	0.0386	0.8880	0.8597	0.9040	0.8873
2019	CPD [43]	VGG-16	0.0567	0.7935	0.7800	0.8685	0.8178	0.0425	0.8638	0.8458	0.9038	0.8669
2019	CPD [43]	ResNet-50	0.0560	0.7966	0.7807	0.8726	0.8248	0.0429	0.8649	0.8431	0.9009	0.8691
2019	BASNet [29]	ResNet-34	0.0565	0.8053	0.7906	0.8691	0.8362	0.0472	0.8589	0.8416	0.8790	0.8660
2020	AADFNet [21]	ResNet-50	0.0488	0.8143	0.8050	0.8744	0.8389	0.0314	0.8993	0.8911	0.9225	0.8914
2020	GateNet [44]	VGG-16	0.0613	0.7940	0.7691	0.8534	0.8209	0.0448	0.8695	0.8388	0.8856	0.8705
2020	GateNet [44]	ResNet-50	0.0552	0.8180	0.7914	0.8682	0.8382	0.0399	0.8870	0.8552	0.9004	0.8852
2020	GateNet [44]	ResNet-101	0.0547	0.8210	0.7944	0.8736	0.8449	0.0380	0.8919	0.8615	0.9075	0.8910
2020	U2Net [45]	RSU	0.0544	0.8226	0.8023	0.8716	0.8467	0.0443	0.8719	0.8479	0.8840	0.8738
2020	MINet [46]	VGG-16	0.0572	0.7936	0.7755	0.8644	0.8218	0.0395	0.8761	0.8550	0.9070	0.8753
2020	MINet [46]	ResNet-50	0.0559	0.8098	0.7911	0.8734	0.8329	0.0373	0.8833	0.8597	0.9132	0.8842
2020	LDF [20]	ResNet-50	0.0517	0.8199	0.8015	0.8814	0.8392	0.0336	0.8968	0.8779	0.9232	0.8924
2021	SAC [17]	ResNet-101	0.0523	0.8287	0.8092	0.8833	0.8487	0.0339	0.8944	0.8732	0.9208	0.8957
2021	CANet [14]	CNN	0.0581	0.8101	0.7796	0.8593	0.8356	0.0437	0.8755	0.8382	0.8896	0.8781
2021	SGL-KRN [22]	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 [22]	ResNet-50	0.0496	0.8101	0.7956	0.8880	**0.8533**	0.0328	0.8945	0.8761	0.9353	0.9005
2022	ICON [28]	ResNet-50	0.0569	0.8254	0.8013	0.8791	0.8445	0.0370	0.8917	0.8665	0.9142	0.8889
2022	EDN [31]	VGG-16	0.0565	0.7818	0.7686	0.8628	0.8376	0.0410	0.8636	0.8457	0.9118	0.8829
2022	EDN [31]	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**

Table 2.2: Quantitative comparison on the HKU-IS and PASCAL-S datasets.

			HKU-IS [53] (4,447 images)					**PASCAL-S** [55] (850 images)				
Year	**Method**	**Backbone**	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$
2019	MLMSNet [42]	VGG-16	0.0387	0.9207	0.8891	0.9379	0.9066	0.0736	0.8552	0.8254	0.8447	0.8443
2019	AFNet [19]	VGG-16	0.0358	0.9226	0.8998	0.9475	0.9055	0.0700	0.8629	0.8409	0.8851	0.8494
2019	EGNet [18]	VGG-16	0.0345	0.9273	0.9050	0.9503	0.9100	0.0776	0.8585	0.8371	0.8714	0.8475
2019	EGNet [18]	Resnet-50	0.0309	0.9352	0.9122	0.9564	0.9180	0.0740	0.8653	0.8437	0.8772	0.8521
2019	CPD [43]	VGG-16	0.0333	0.9239	0.9075	0.9501	0.9042	0.0721	0.8612	0.8441	0.8837	0.8446
2019	CPD [43]	ResNet-50	0.0342	0.9250	0.9047	0.9503	0.9056	0.0706	0.8595	0.8414	0.8873	0.8484
2019	BASNet [29]	ResNet-34	0.0322	0.9284	0.9113	0.9458	0.9090	0.0758	0.8539	0.8344	0.8527	0.8380
2020	AADFNet [21]	ResNet-50	0.0255	0.9415	**0.9339**	0.9592	0.9190	0.0550	0.8797	0.8677	0.9051	0.8658
2020	GateNet [44]	VGG-16	0.0361	0.9287	0.9036	0.9470	0.9100	0.0684	0.8696	0.8439	0.8692	0.8574
2020	GateNet [44]	ResNet-50	0.0337	0.9335	0.9097	0.9534	0.9154	0.0680	0.8690	0.8459	0.8842	0.8580
2020	GateNet [44]	ResNet-101	0.0320	0.9375	0.9136	0.9567	0.9195	0.0668	0.8702	0.8468	0.8924	0.8622
2020	U2Net [45]	RSU	0.0312	0.9352	0.9133	0.9484	0.9161	0.0740	0.8592	0.8386	0.8500	0.8444
2020	MINet [46]	VGG-16	0.0316	0.9302	0.9133	0.9540	0.9119	0.0645	0.8650	0.8450	0.8961	0.8544
2020	MINet [46]	ResNet-50	0.0292	0.9349	0.9166	0.9600	0.9189	0.0643	0.8665	0.8461	0.8981	0.8563
2020	LDF [20]	ResNet-50	0.0275	0.9394	0.9224	0.9597	0.9196	0.0596	0.8741	0.8577	0.9048	0.8630
2021	SAC [17]	ResNet-101	0.0257	0.9416	0.9260	0.9636	0.9253	0.0622	0.8772	0.8585	0.9022	0.8656
2021	CANet [14]	CNN	0.0371	0.9297	0.8977	0.9455	0.9100	0.0728	0.8662	0.8392	0.8790	0.8552
2021	SGL-KRN [22]	ResNet-50	0.0280	0.9301	0.9154	0.9539	0.9206	0.0678	0.8502	0.8373	0.8941	0.8556
2021	PA-KRN [22]	ResNet-50	0.0271	0.9349	0.9198	0.9561	0.9235	0.0665	0.8530	0.8388	0.8964	0.8578
2022	ICON [28]	ResNet-50	0.0289	0.9395	0.9196	0.9585	0.9202	0.0644	0.8757	0.8514	0.8931	0.8611
2022	EDN [31]	VGG-16	0.0286	0.9286	0.9141	0.9504	0.9208	0.0650	0.8555	0.8406	0.8955	0.8605
2022	EDN [31]	ResNet-50	0.0264	0.9325	0.9196	0.9548	0.9241	0.0617	0.8600	0.8489	0.9015	0.8646
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**

Table 2.3: Quantitative comparison on the ECSSD and SOD datasets.

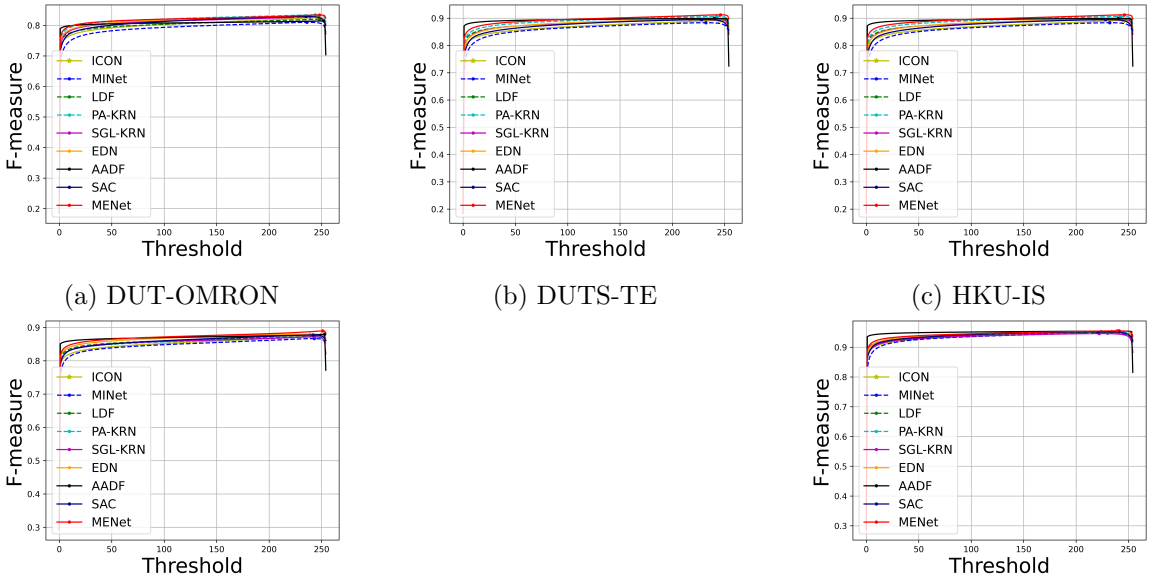
			ECSSD [54] (1,000 images)					**SOD** [60] (300 images)				
Year	**Method**	**Backbone**	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$	$MAE\downarrow$	$MaxF\uparrow$	$mF\uparrow$	$mE_m\uparrow$	$S_m\uparrow$
2019	MLMSNet [42]	VGG-16	0.0446	0.9284	0.9007	0.9161	0.9112	0.1060	0.8517	0.8291	0.8019	0.7898
2019	AFNet [19]	VGG-16	0.0418	0.9350	0.9157	0.9414	0.9135	-	-	-	-	-
2019	EGNet [18]	VGG-16	0.0405	0.9434	0.9232	0.9408	0.9193	0.1100	0.8589	0.8426	0.8209	0.7882
2019	EGNet [18]	Resnet-50	0.0374	0.9474	0.9288	0.9469	0.9246	0.0969	0.8778	0.8610	0.8422	0.8067
2019	CPD [43]	VGG-16	0.0402	0.9360	0.9233	0.9433	0.9103	0.1125	0.8480	0.8365	0.8124	0.7715
2019	CPD [43]	ResNet-50	0.0371	0.9393	0.9244	0.9494	0.9182	0.1097	0.8568	0.8376	0.8174	0.7711
2019	BASNet [29]	ResNet-34	0.0370	0.9425	0.9274	0.9210	0.9163	0.1124	0.8487	0.8368	0.7793	0.7721
2020	AADFNet [21]	ResNet-50	**0.0280**	0.9543	**0.9478**	0.9529	0.9299	0.0903	0.8677	0.8579	0.8051	0.7929
2020	GateNet [44]	VGG-16	0.0418	0.9413	0.9191	0.9314	0.9169	-	-	-	-	-
2020	GateNet [44]	ResNet-50	0.0408	0.9454	0.9250	0.9431	0.9198	-	-	-	-	-
2020	GateNet [44]	ResNet-101	0.0357	0.9508	0.9301	0.9501	0.9302	-	-	-	-	-
2020	U2Net [45]	RSU	0.0330	0.9510	0.9325	0.9251	0.9276	0.1061	0.8588	0.8428	0.7993	0.7891
2020	MINet [46]	VGG-16	0.0370	0.9435	0.9296	0.9475	0.9192	-	-	-	-	-
2020	MINet [46]	ResNet-50	0.0342	0.9475	0.9309	0.9532	0.9250	-	-	-	-	-
2020	LDF [20]	ResNet-50	0.0335	0.9501	0.9379	0.9509	0.9245	-	-	-	-	-
2021	SAC [17]	ResNet-101	0.0309	0.9512	0.9376	**0.9586**	**0.9312**	0.0934	**0.8804**	0.8695	0.8482	0.8087
2021	CANet [14]	CNN	0.0441	0.9378	0.9103	0.9362	0.9154	0.0992	0.8650	0.8406	0.8331	0.8007
2021	SGL-KRN [22]	ResNet-50	0.0360	0.9368	0.9241	0.9462	0.9231	-	-	-	-	-
2021	PA-KRN [22]	ResNet-50	0.0323	0.9425	0.9301	0.9503	0.9278	-	-	-	-	-
2022	ICON [28]	ResNet-50	0.0318	0.9503	0.9336	0.9543	0.9290	**0.0841**	0.8790	**0.8711**	**0.8516**	**0.8238**
2022	EDN [31]	VGG-16	0.0336	0.9408	0.9285	0.9508	0.9283	-	-	-	-	-
2022	EDN [31]	ResNet-50	0.0320	0.9410	0.9304	0.9508	0.9267	-	-	-	-	-
2023	**MENet(Ours)**	ResNet-50	0.0307	**0.9549**	0.9422	0.9544	0.9279	0.0874	0.8780	0.8684	0.8381	0.8089



(a) DUT-OMRON (b) DUTS-TE (c) HKU-IS

(d) PASCAL-S (e) ECSSD

Figure 2.4: PR-curves comparison.



(a) DUT-OMRON (b) DUTS-TE (c) HKU-IS

(d) PASCAL-S (e) ECSSD

Figure 2.5: Fm-curves comparison.

2.4.4 Quantitative and Qualitative Comparison

We compare the proposed MENet with 16 recent SOD models, including MLM-SNet [42], AFNet [19], EGNet [18], CPD [43], BASNet [29], AADFNet [21], GateNet [44], U2Net [45], MINet [20], SAC [17], CANet [14], SGL-KRN [22], PA-KRN [22], EDN [31], and ICON [28]. 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 [61], ResNet-50 [47], and ResNet-101 [47]), because some models do not provide codes or predictive results for all the backbones, as shown in Tables 2.1- 2.3. The symbols '↓'/'↑' indicate the lower/higher the evaluation metric, the better the model is. The symbol '-' means that the model is not available. The top three results are depicted in red, green, and blue. Overall, the proposed MENet has superior performance. 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 [28] and SAC [17] on the SOD dataset, and AADFNet [21] 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. ICON only has better performance on the SOD dataset.

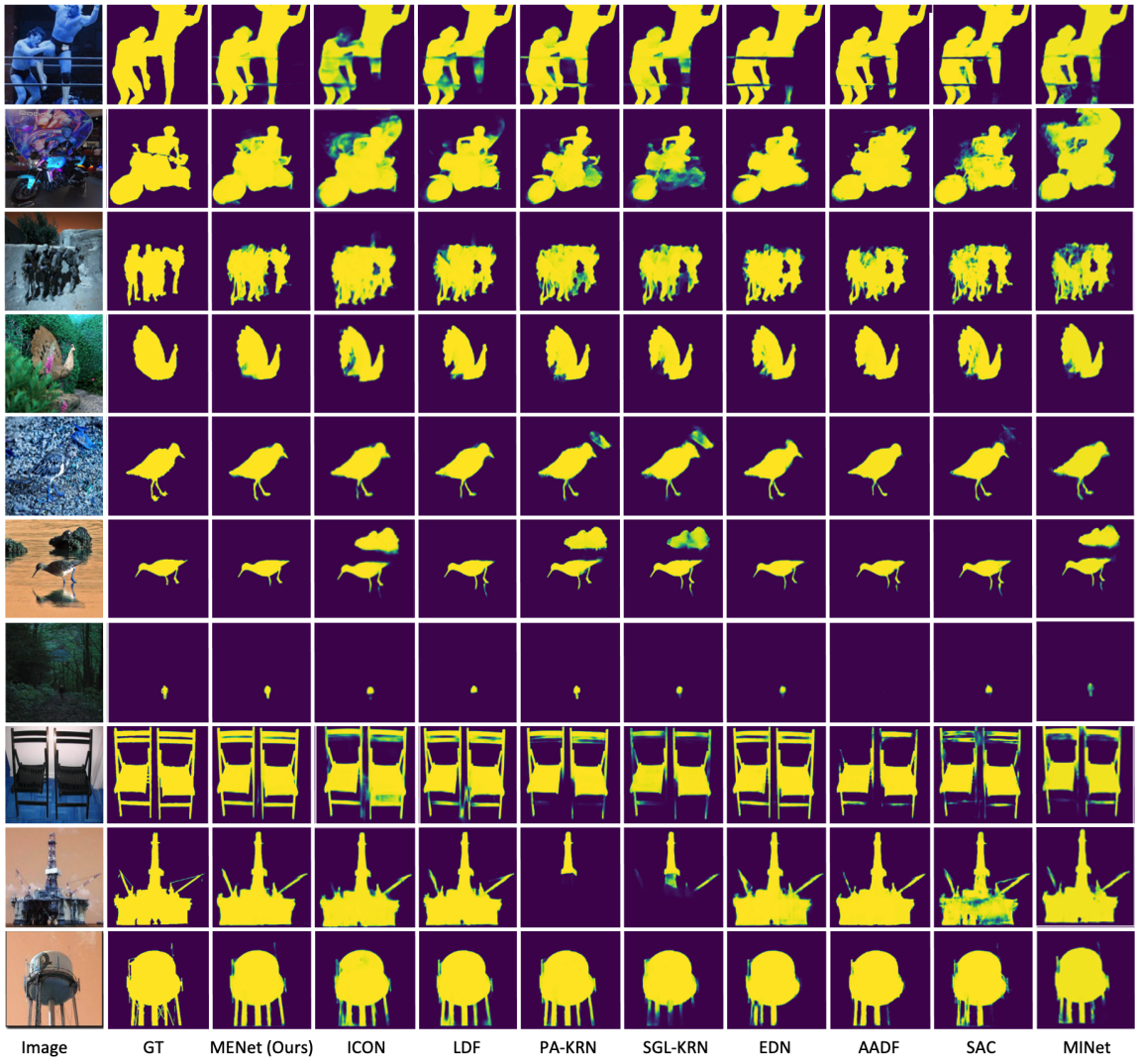


Figure 2.6: Qualitative performance comparison for complex scenes.

PR and F-Measure curves are shown in Fig. 2.4 and Fig. 2.5, 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. 2.6. The prediction results are given in probability. The determined value '1' is marked in yellow, the other probability values are indicated in shades of green, and the value '0' is black. As we can see, the proposed MENet achieves the most accurate overall detection results for

Table 2.4: Comparison of MENet with different iterative enhancement times. The best results are shown in bold red.

Dataset	ITimes	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$mF_\beta \uparrow$	$mE_m \uparrow$	$S_m \uparrow$
DUT-OMRON [52]	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 [50]	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 [53]	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 [55]	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 [54]	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 [60]	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**

salient regions. The boundaries of the targets are also more precise and complete.

2.4.5 Ablation Study

We investigated the different settings for improving and supervising MENet, using Resnet-50 as the backbone.

Number of iterative enhancements: Table 2.4 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 2.5 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 DUT-OMRON, DUTS-TE, HKU-IS, and PASCAL-S datasets. The accuracy is further improved, especially when gradient supervision ($\mathcal{L}_g$) is added. 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.

Table 2.5: Ablation tests for loss settings. The best results are shown in bold red.

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 [52]	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 [50]	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 [53]	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 [55]	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 [54]	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 [60]	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

Sub-region numbers of region-level loss: Table 2.4 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.

2.5 Conclusion

This chapter proposes a Multiple Enhancement Eetwork (MENet) to improve the performance of SOD for complex scenes by fully utilising the Human Visual System (HVS) and cognition mechanisms. Two multi-scale feature enhancement modules are the core of MENet, and they gradually propagate and fuse boundary and global features, respectively. We also propose a novel multilevel hybrid loss that measures similarities at the pixel, region, and object levels. Comprehensive experiments on

six challenging databases demonstrate that MENet achieves the new state-of-the-art for SOD. As shown in the studies in this chapter, SOD in blurred and low-contrast natural scenes is still a valuable but challenging topic that warrants more research efforts.

References

[1] 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 Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(6):3239–3259, 2021.

[2] 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.

[3] Dengping Fan, Mingming Cheng, Jiangjiang Liu, Shanghua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 186–202, 2018.

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

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

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

[7] Tao Chen, Yazhou Yao, Lei Zhang, Qiong Wang, Guo-Sen Xie, and Fumin Shen. Saliency guided inter-and intra-class relation constraints for weakly supervised semantic segmentation. *IEEE Transaction Multimedia (TMM)*, 25:1727–1737, 2022.

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

[9] Filiz Gurkan, Llukman Cerkezi, Ozgun Cirakman, and Bilge Gunsel. Tdiot:

Target-driven inference for deep video object tracking. *IEEE Transactions on Image Processing (TIP)*, 30:7938–7951, 2021.

[10] Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing (TIP)*, 30:948–962, 2021.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[12] Dengping Fan, Jing Zhang, Gang Xu, Mingming Cheng, and Ling Shao. Salient objects in clutter. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(2):2344–2366, 2022.

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

[14] Qinghua Ren, Shijian Lu, Jinxia Zhang, and Renjie Hu. Salient object detection by fusing local and global contexts. *IEEE Transaction Multimedia (TMM)*, 23:1442–1453, 2021.

[15] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4722, 2022.

[16] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 355–370, 2018.

[17] Xiaowei Hu, Chi Wing Fu, Lei Zhu, Tianyu Wang, and Pheng Ann Heng. Sacnet: Spatial attenuation context for salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 31(3):1079–1090, 2021.

[18] 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 *International Conference on Computer Vision (ICCV)*, pages 8779–8788, 2019.

[19] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1623–1632, 2019.

[20] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *Proceedings of IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 13022–13031, 2020.

[21] 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 Transactions on Circuits and Systems for Video Technology (TCSVT)*, 30(10):3358–3371, 2020.

[22] 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 Conference on Artificial Intelligence*, volume 35, pages 3004–3012, 2021.

[23] Jiangjiang Liu, Qibin Hou, Mingming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3917–3926, 2019.

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

[25] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on Visual Computing*, pages 234–244. Springer, 2016.

[26] 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 *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) Brainlesion Workshops*, pages 64–76. Springer, 2017.

[27] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 12321–12328, 2020.

[28] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(3):3738–3752, 2022.

[29] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR), pages 7471–7481, 2019.

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

[31] Yuhuan Wu, Yun Liu, Le Zhang, Mingming Cheng, and Bo Ren. Edn: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing (TIP)*, 31:3125–3136, 2022.

[32] 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.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, page 234–241, 2015.

[34] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9141–9150, 2020.

[35] Yuanfang Zhang, Jiangbin Zheng, Wenjing Jia, Wenfeng Huang, Long Li, Nian Liu, Fei Li, and Xiangjian He. Deep rgb-d saliency detection without depth. *IEEE Transaction Multimedia (TMM)*, 24:755–767, 2021.

[36] 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 Transaction Multimedia (TMM)*, 25:1991–2006, 2022.

[37] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WCACV)*, pages 3560–3569, 2021.

[38] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005.

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

[40] Mingming Cheng and Dengping Fan. Structure-measure: A new way to evaluate foreground maps. *International Journal of Computer Vision (IJCV)*, 129(9):2622–2638, 2021.

[41] Yuxin Mao, Jing Zhang, Zhexiong Wan, Yuchao Dai, Aixuan Li, Yunqiu Lv,

Xinyu Tian, Dengping Fan, and Nick Barnes. Generative transformer for accurate and reliable salient object detection. *arXiv e-prints*, pages arXiv–2104, 2021.

[42] 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 *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8142–8151, 2019.

[43] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3902–3911, 2019.

[44] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 23–28, 2020.

[45] 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. *Pattern Recognition*, 106:107404, 2020.

[46] Lihe Zhang, Jie Wu, Tiantian Wang, Ali Borji, Guohua Wei, and Huchuan Lu. A multistage refinement network for salient object detection. *IEEE Transactions on Image Processing (TIP)*, 29:3534–3545, 2020.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

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

[49] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems Conference (ANIPS)*, volume 25, pages 1097–1105, Red Hook, NY, USA, 2012.

[50] 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 *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3796–3805, 2017.

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

[52] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3166–3173, 2013.

[53] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5455–5463, 2015.

[54] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1162, 2013.

[55] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 280–287, 2014.

[56] Huajun Zhou, Yang Lin, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Benchmarking deep models on salient object detection. *Pattern Recognition*, 145:109951, 2024.

[57] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–740, 2012.

[58] Dengping Fan, Cheng Gong, Yang Cao, Bo Ren, Mingming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, page 698–704, 2018.

[59] Ran Margolin, Lihi Zelnik Manor, and Ayellet Tal. How to evaluate foreground maps. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, Columbus, OH, USA, 2014.

[60] Vida Movahedi and James H. Elder. Design and perceptual validation of performance measures for salient object segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 49–56, 2010.

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

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Yi Wang
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 3

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

● The manuscript is currently under review for publication – please indicate:

- The name of the journal:

Applied Soft Computing

- The percentage of the manuscript/published work that was contributed by the candidate: 80.00

- Describe the contribution that the candidate has made to the manuscript/published work:

For the manuscript, the contribution that the candidate has made include conceptualization,investigation,methodology,validation and draft writing.

○ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	wangyi 数字签名者：wangyi 日期：2024.04.02 11:47:30 +13'00'
Date:	02-4 月-2024
Primary Supervisor's Signature:	Prof Ruili Wang Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, c=NZ, o=Massey University , ou=School of Natural and Computational Sciences , email=ruili.wang@massey.ac.nz Date: 2024.04.02 17:18:32 +13'00'
Date:	2-4 月-2024

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 3

Frequency-Guided Saliency Detection

Saliency detection based on deep learning models has a contradictory issue. It needs to learn both global context and fine detail simultaneously. In this chapter, we propose an effective frequency-guided network called TFGNet based on the Transformer to address this issue. TFGNet features a two-branch feature aggregation structure, enabling simultaneous learning of high-frequency boundary details and low-frequency inner regions of salient objects. A lightweight but powerful multi-scale frequency feature enhancement module is designed to optimise frequency feature learning in each branch. In addition, a global intensity distribution similarity measure is designed to further increase performance. In comprehensive evaluations with 24 state-of-the-art methods on five widely used SOD datasets, TFGNet demonstrated superior performance in complex scenes.

3.1 Introduction

The main difficulty of SOD lies in balancing two contrasting goals: capturing global contexts and preserving local details [1]. Capturing the global context becomes even more critical when the salient region comprises multiple objects with substantial differences in appearance or complex geometric shapes. Global contexts are essential for understanding the overall structure of the salient objects, which requires globally

consistent (invariant) features (such as semantic features). Boundary details are vital for accurate segmentation, but they usually have imbalance distribution or bad consistency for small/tiny objects or multi-connected objects.

A U-Net-like encoder-decoder architecture [2] has been used to resolve the above contradiction issue of salient feature learning. It has been the mainstay of SOD models [3–10]. This structure adopts a top-down encoder and a down-top decoder to locate and refine details. Using convolution and down-sampling, the top-down path determines the approximate location of salient objects and extracts their global semantic information; in the reverse path, the rough salient map is refined by integrating features from different layers, and the fine salient map is generated at the highest level. Skip connections are specially implemented between the corresponding layers to retain global and detailed information. Based on this architecture, SOD models have introduced various refinement strategies to further enhance performance. Some models adopt a multi-scale feature enhancement strategy to capture salient details at different scales, obtaining more comprehensive and accurate saliency maps [5–7, 9, 11–14]. Some works explicitly use auxiliary boundary information to improve overall segmentation quality [3, 7, 13, 15]. However, these methods tend to adopt some communication strategy between layers or branches [3, 6, 9, 13, 16]. Although such strategies may complement global and local feature information, they can also introduce errors that propagate across layers or branches, causing salient objects to be incorrectly localised and boundary segmentation to be inaccurate. Recently, Transformer-based models [10, 17–19] have shown promising results in benchmark datasets and outperformed CNN-based methods. However, accurately detecting boundaries and locations in complex natural and social scenes remains a challenge, as demonstrated in Fig. 3.1, where there are multiple connected objects, blurred boundaries, low contrast, and reflections.

One promising approach to address this issue is to re-examine the characteristics of image frequency [9, 16]. An image contains varying levels of information on different scales. Different objects or textures in an image can exhibit various spatial frequencies, with some areas containing fine details (high-spatial frequency (HSF) components) and others having smoother or less gradual changes (low-spatial frequency (LSF) components). So, we can decompose an image into the HSF and LSF parts and then learn and enhance its detailed and global salient features, respectively. Furthermore, when combined with multi-scale information, the salient feature of the

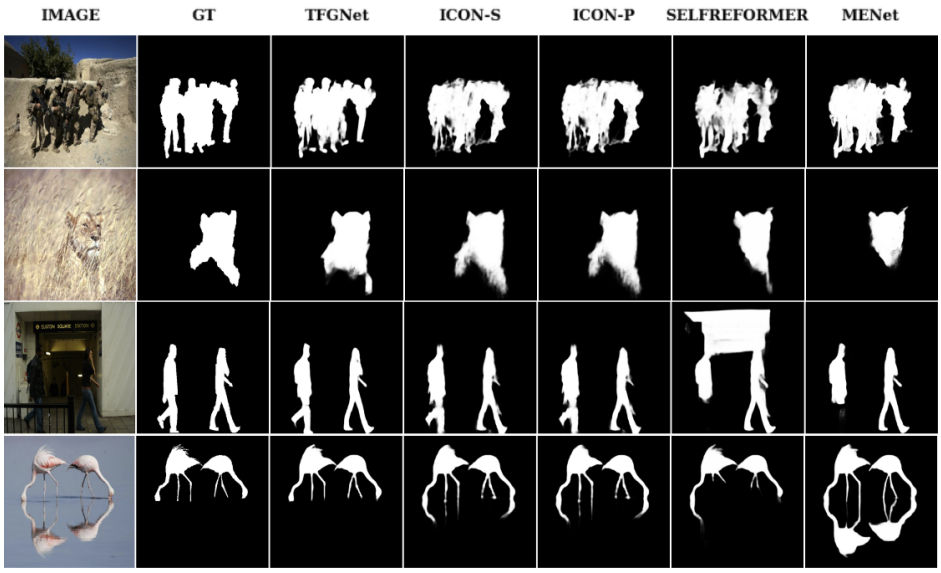


Figure 3.1: Visual examples of challenging real-world scenes.

boundary can be more fully expressed and improved [9, 16].

This chapter presents a concise and compelling Transformer-based Frequency-Guided Network (TFGNet). First, TFGNet's network structure strictly adheres to an image's high- and low-frequency decomposition and synthesis mechanisms by employing a parallel two-branch decoder that refines boundary details and inner regions of salient objects. The HSF and LSF features then construct a full-band saliency prediction. The HSF and final full-band predictions are supervised, while low-frequency features are adaptively learnt. The difference from existing edge-aware/two-branch decoding structures lies in the following points: (i) It learns high-frequency (boundary details) and low-frequency (internal body regions, rather than the entire area as in other two-branch-based models [7, 8, 13]) features separately. This can reduce the impact of inaccurate edge information on global feature representation learning. (ii) No information exchange between intermediate high- and low-frequency features in TFGNet. This design minimises the potential inaccuracies in positioning and boundary predictions resulting from erroneous high-frequency and low-frequency information communication. MENet [9] and SFENet [16] adopt a similar boundary- and inner-region-separated feature learning strategy. However, SFENet uses a two-stage process, while MENet uses four stages to enhance the features. In each stage, the HSF and LSF information from the previous stage is fused, so global and detailed information is exchanged to complement each other multiple times. In contrast, TFGNet only

needs a one-stage decoding process with comprehensive decoders and loss functions designed for the network.

We adopt a pixel-wise decoder combined with a Transformer (TF) decoder, inspired by Maskformer [20], to optimise the features by combining per-pixel embedding and mask-level embedding in each branch. Unlike Maskformer, TFGNet's pixel-wise decoder uses frequency feature enhancement (FFE), gradually enhancing, upsampling, and aggregating multi-scale features. FFE adopts a global-local attention mechanism [21] to handle objects at different scales and capture essential contextual information. The TF-decoder learns mask-level discriminative features. Combining FFE with TF-decoder ensures a comprehensive understanding of salient regions and improves the local representation of this Transformer-based network.

Loss functions are also essential for training a model. Binary Cross Entropy (BCE) [22] is widely employed in SOD. However, BCE loss does not account for spatial dependencies, so it may not be sufficient to predict structural integrity. To overcome this limitation, several SOD models [3, 6, 7, 9, 10, 12] have incorporated structural measures such as intersection-over-union (IoU) loss [23]. The intensity values of pixels can be interpreted as signals, and if two images are highly similar, their global signal intensity distributions should also be identical. Histograms provide a visual representation of the frequency distribution of these intensity values.

We introduce histogram dissimilarity measurement as a loss to enhance the overall accuracy of saliency predictions. A histogram of an image shows the global frequency distributions of the image. Measurement of the histogram of predicted maps with GT maps can provide a global view of the pixel distribution from frequency perception. Then, by combining the histogram-based loss with the BCE and IoU losses, we create a hybrid loss that ensures a comprehensive and effective training process for TFGNet. As far as we are aware, this is the first instance of a histogram-based loss being utilised in a SOD model.

This chapter presents the following key contributions:

- We present TFGNet, an efficient and concise Transformer-based Spatial-Frequency Guidance Framework for SOD. TFGNet leverages the image spatial frequency decomposition and synthesis mechanism, learning HSF and LSF salient features in separate branches. Using a pixel-decoder and Transformer-decoder,

each branch learns comprehensive embedding that lead to more accurate and robust predictions.

- A novel hybrid loss is introduced by combining frequency distribution (i.e., histogram) similarity measurement with a BCE loss and an IoU loss to improve TFGNet performance further. As far as we are aware, this is the first instance of utilizing the histogram-based loss in a SOD model.
- A comprehensive evaluation of TFGNet with 24 recent SOD methods on five widely used SOD datasets and an underwater SOD dataset. The results demonstrate that TFGNet can accurately local salient objects with more complete and precise boundaries on various complex backgrounds.

The rest of this chapter is structured as shown below: Section 3.2 briefly reviews related SOD approaches in this chapter. Section 3.3 explains the details of TFGNet. Section 3.4 demonstrates and discusses the proposed method through quantitative and qualitative experiments. Section 3.5 outlines the contributions of this chapter.

3.2 Related Work

Recently, various SOD approaches have been successively proposed [1, 24]. The following brief overview provides a summary of recent strategies in SOD research.

Convolutional Neural Network (CNN)-based SOD methods demonstrate promising results across various benchmark datasets [1, 24]. Most of these models adopt a U-Net [2]-like framework. However, semantic global features in top layers are not enough to accurately localise salient objects, so it is possible to guide detail learning incorrectly. Hence, the design of the decoder is crucial for generating accurate salient maps. Some methods employ a multitasking architecture to predict boundaries and salient regions simultaneously. For example, EGNet [13] explicitly obtains salient edge features to guide multi-resolution salient feature learning. LDF [6] divides a GT map into a body map and a detailed map, allowing collaborative supervision of body parts and details of saliency regions. UDNet [14] locates pixels within and surrounding the contours of the region using internal contour uncertainty maps, whole saliency maps, and external contour uncertainty maps. DCN [15] uses a multitasking network to simultaneously predict salient maps, edges, and skeleton maps. Then, cross-task aggregation and cross-layer aggregation modules are used to integrate multi-level and

multitasking features for the final results. PA-KRN [7] performs intermediate edge supervision on its five feature layers in its exemplary segmentation module to ensure that the boundaries provided by the encoding process are clear. AADFNet [8] selectively uses small and large dilated rate convolutions to obtain local and global relevant information. SFENet [16] uses a spatial frequency enhancement (SFE) module to refine saliency features by extracting and exchanging frequency information among multiple in-stage and cross-stage feature maps. MENet [9] integrates multiple human visual systems (HVS) operations into the network structure and the loss function. Specifically, MENet features a two-branch decoding process that progressively refines the boundary and adaptive features.

Although these methods utilise additional edge information and / or edge supervision to improve detection accuracy, accurate boundary/contour prediction in complex scenes is still tricky. There are fewer boundary data than inner region data, which results in sub-optimal results when using edges directly as supervision [6]. In addition, boundary labels contain less information. They are easily disturbed by similar textures within large objects [1], especially when the difference between the background and the foreground is limited or the object has blurred boundaries. Several methods apply a certain communication design between levels of different branches that complement each other. Error information can also be exchanged at the same time.

Transformers show promise for computer vision tasks like SOD due to their effective modelling of long-range dependencies using self-attention [25, 26]. Several transformer-based SOD models [10, 17–19] have shown promising results in benchmark data sets and outperformed CNN-based methods. Visual Saliency Transformer (VST) [17] introduced a token-based multitask structure. The decoder detects salient regions and boundaries simultaneously using task-related tokens. EBMG [18] uses the vision transformer that generates latent variables to detect salient objects based on an informative energy-driven prior. Using global and local context generated by the Context Refinement Module, the Self-Refined Transformer (SelfReformer) [19] can guide and correct itself, and the predictions are reshaped to ground truth using Pixel Shuffle from Super-Resolution (SR). ICON-P/ICON-S [10] (based on the PVT [27] and Swin [28] backbones, respectively) incorporate three critical components to achieve complete SOD: aggregation of diverse features, enhancement of the integrity channel, and verification of the whole.

Given images, the image's high spatial frequencies often represent edges and finer details. In contrast, low-spatial frequencies represent overall structures, and we can adopt the frequency decomposition technique of images in the network design to learn these two features separately, reducing the interference between them. Improving local representation in the transformer network is the key to improving accuracy. Therefore, we designed a pixel-decoder combined with a Transformer decoder in each frequency branch for TFGNet to improve salient representation accuracy.

3.3 Methodology

3.3.1 Framework

The proposed frequency-guided network (TFGNet) is shown in Fig. 3.2. The backbone network produces multistage multi-scale features denoted as $B_i \in \mathcal{R}^{H_i \times W_i \times C_i}$. Since widely used backbones such as Swin [28], all produce four-stage multi-scale features, we let i range from 1 to 4. The notation H_i, W_i, and C_i represent the block height, width, and channel number in the order given.

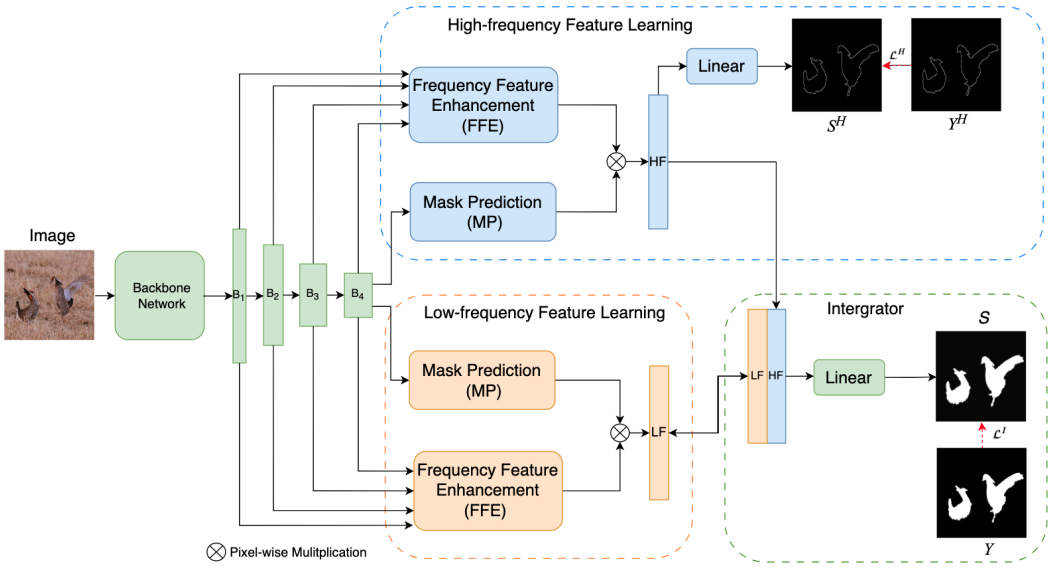


Figure 3.2: Illustration of the overall architecture of the proposed TFGNet.

The decoding process is divided into a high-spatial-frequency (HSF) feature learning process and a low-spatial-frequency (LSF) feature learning process. Each stream consists of a spatial frequency-feature enhancement (FFE) module, also known as a pixel

decoder, and a mask prediction (MP) module, which utilises a Transformer decoder. In the FFE module, the network gradually propagates and aggregates multi-scale spatial frequency features extracted from the backbone network. This process helps compute per-pixel embedding, which captures fine-grained details and local information. The MP module operates at the mask level using the TF-decoder. It takes the input image or feature maps and divides them into segments, which are called masks. Each mask is then classified into the salient or non-salient category through the Transformer decoder layer, producing per-mask embedding. This TF-decoder focusses on capturing global information and high-level features at the mask level. The optimised frequency feature embedding is computed by performing a product operation between the two embeddings. This fusion of per-pixel and per-mask embedding ensures a comprehensive understanding of the salient object regions and achieves more accurate and robust salient features.

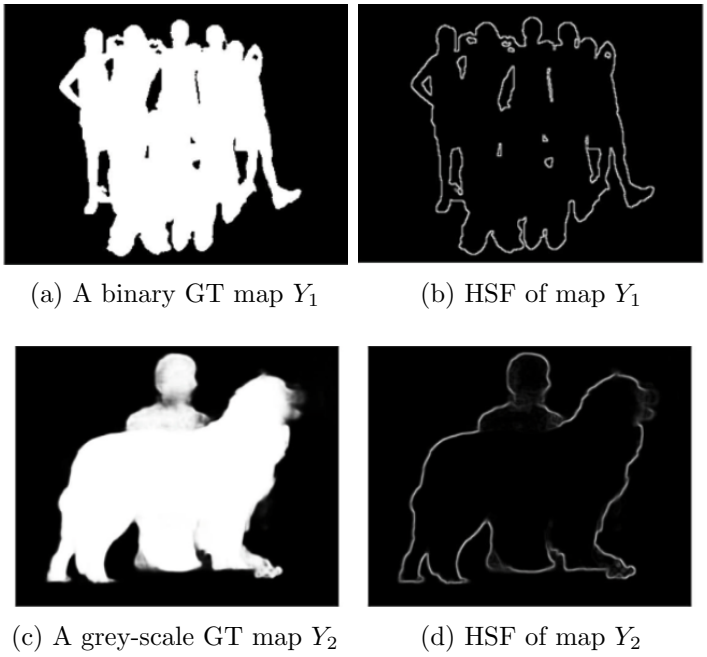
In the last part of TFGNet, the enhanced frequency salient features are concatenated into an integrator, which outputs a full band salient map, denoted S. During training, S is supervised by the proposed hybrid loss function.

The following subsections offer a detailed description of the functionalities of each module.

3.3.2 GT Map Decomposition

An image can be decomposed into multiple spatial frequencies. Since SOD is a binary classification task, a ground truth (GT) map ($Y = \{y_i \mid y_i \in [0,1]\}_{i=1}^{H \cdot W}$) is typically binary or greyscale (as illustrated in Fig. 3.3), with distinct boundary and uniformly distributed inner region. This allows us to decompose a GT map into the HSF component (Y^H) and the LSF component (Y^L) as supervisions for the learning of the salient features of HSF and LSF, respectively.

HSF information in a GT map exhibits the object's details, edges, and other fine-grained features. The LSF part can be computed by $Y^L = Y - Y^H$, which is the inner region of the salient object. Since a GT map is a simple binary or grey-scale image with no background or foreground noise, gradient-based edge detectors [29], such as Prewitt, Roberts, and Sobel, are usually used to extract Y^H by convolving the image with different size kernels. We use the Sobel detector in our implementation, as shown in Fig. 3.3.



(a) A binary GT map Y_1 (b) HSF of map Y_1

(c) A grey-scale GT map Y_2 (d) HSF of map Y_2

Figure 3.3: Two types of GT maps and the HSF components of the GT maps computed by the Sobel edge detector.

In contrast to some edge-aware methods, such as EGNet and MENet, which utilise edge detectors to extract salient edges from GT maps as prior information, our approach takes a different perspective based on the frequency decomposition of the GT maps. This ensures theoretical consistency with the network structure.

3.3.3 Salient Features Learning

We combine a pixel-level decoder and a Transformer decoder to optimise salient features.

Pixel Decoder: It is a multi-scale frequency feature enhancement (FFE) module, as shown in Fig. 3.4. Enhancing features begins with the feature B_4. Then, the three remaining feature blocks (i.e., B_3, B_2, and B_1) are aggregated sequentially, proceeding from small to large resolutions. We introduce a Global-Local Attention (GL-Attention for short in Fig. 3.4 module to the FFE to tackle the challenge of recognizing and detecting objects in extreme scale variations. GL-Atten is based on multi-scale channel attention (MS-CAM) [21]. By incorporating GL-Attention into the FFE, the network becomes more adept at handling objects at different scales and capturing essential contextual information, leading to accurate pixel-wise classification.

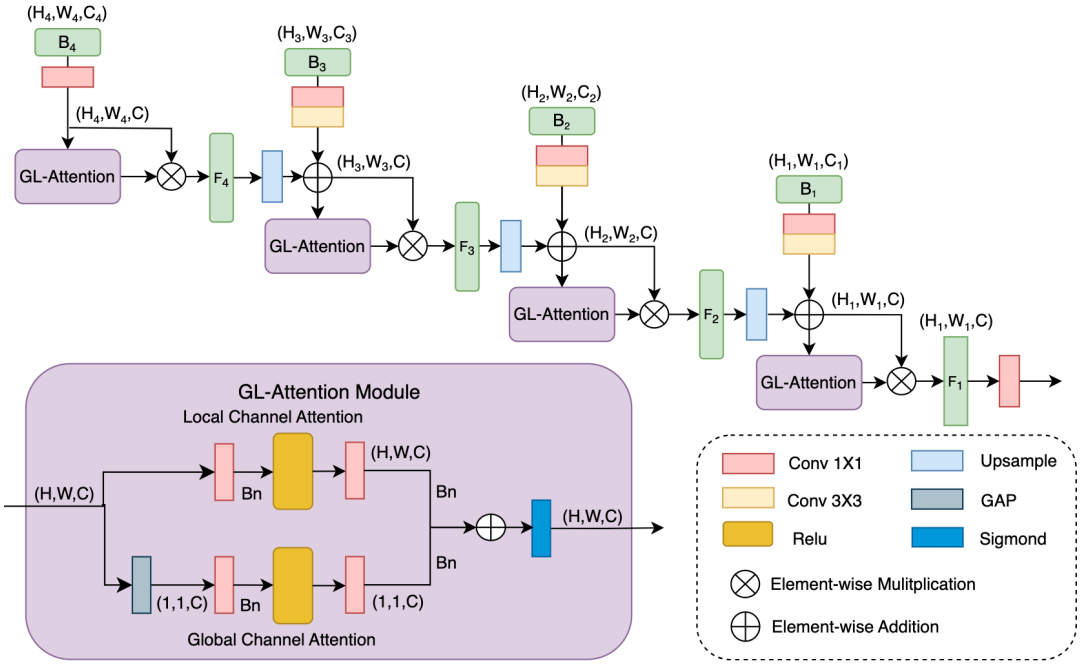


Figure 3.4: FFE module structure.

Starting from the feature block B_4, a $[1 \times 1]$ convolutional layer (denoted as $Conv_1(\,)$) compresses the number of channels to C. In the implementation, we let $C = 256$. Next, we apply $[1 \times 1]$ followed by a $[3 \times 3]$ convolutional layers (denoted $Conv_3(\,)$) to the feature blocks B_i to make these feature blocks the same channel value C, ensuring consistency in channel dimensions.

Then, a feature block B_i is refined by a global channel attention process ($gAtten(\,)$) and a local channel attention process ($lAtten(\,)$), synchronously, in the following Eq. (3.1):

$$\begin{aligned} &glAtten(B_i) = Sig(gAtten(B_i) \oplus lAtten(B_i)) \\ &= Sig(Conv_1(ReLU(Bn(Conv_1(GAP(B_i)))))) \oplus Conv_1(ReLU(Bn(Conv_1(B_i))))), \end{aligned} \tag{3.1}$$

where $ReLU(\,)$ is the Rectified Linear Unit, $GAP(\,)$ is the global average pooling, $Sig(\,)$ is the Sigmond Activation function, $Bn(\,)$ is the Batch Normalization, and $\oplus$ is the pixel-wise addition.

let $\hat{B}_i = UP(F_{i+1}) \oplus Conv_3(Conv_1(B_i))(i = 3, 2, 1)$, where $Up(\,)$ is the up-sampling operation, then the FFE refine process can be concluded by the following four steps

in Eq. (3.2)-(3.5).

$$F_4 = glAtten(Conv_1(B_4)) \otimes Conv_1(B_4), \tag{3.2}$$

$$F_3 = glAtten(\hat{B}_3) \otimes \hat{B}_3, \tag{3.3}$$

$$F_2 = glAtten(\hat{B}_2) \otimes \hat{B}_2, \tag{3.4}$$

$$F_1 = glAtten(\hat{B}_1) \otimes \hat{B}_1. \tag{3.5}$$

Here, the symbol $\otimes$ represents the multiplication of elements in a given direction.

Transformer Decoder: It is a mask prediction (MP) module based on a Transformer (TF). The MP divides the input into segments, also known as masks, and classifies them into salient and non-salient categories to produce segment embeddings by a TF-decoder layer. These segment embeddings are then processed by a Multi-Layer Perception (MLP) with two hidden layers to convert into N mask embeddings. Segments with the same category label are merged, yielding the final mask embeddings.

Next, the optimised salient features $HF \in \mathcal{R}^{H_1 \times W_1 \times 256}$ and $LF \in \mathcal{R}^{H_1 \times W_1 \times 256}$, where $[H_1 \times W_1]$ are the size of B_1, are obtained through a product operation between the feature embeddings of the pixel-decoder and the TF-decoder. This Pixel-Decoder and TF-decoder optimisation process enables TFGNet to learn and efficiently generate accurate saliency predictions for complex scenes.

Learning Strategies: We use different loss settings and learning strategies for HSF and LSF feature learning branches. Considering the complexity of the network architecture in the HSF feature learning branch, we only apply the BCE loss on the *HF*. Regarding the LSF feature learning branch, both the pixel-decoder and the TF-decoder employ an adaptive learning strategy. Consequently, the LF is not supervised. This strategy has been experimentally proven optimal, helping the network refine high- and low-frequency features more effectively.

Integration: Since the output of HSF and LSF feature learning branches (i.e., *HF* and LF) share the exact resolution and channel number, we concatenate them. Next, a $[3 \times 3]$ convolutional layer is applied to this concatenated map, obtaining the full-band saliency map. Then a $[1 \times 1]$ convolutional layer and an up-sampling operation are used to generate the final saliency map $S \in \mathcal{R}^{H \times W \times 1}$. S is supervised by the

proposed hybrid loss in Section 3.3.4.

3.3.4 Hybrid Loss

There are two supervisions within the network. One ($\mathcal{L}^H$) is for the predicted HSF feature. The other ($\mathcal{L}^I$) is applied to the final predicted full-band map (S). No supervision is applied to the LSF feature learning branch. So the training loss $\mathcal{L}$ can be represented in Eq. (3.6)

$$\mathcal{L} = \alpha_1 \mathcal{L}^H + \alpha_2 \mathcal{L}^I, \tag{3.6}$$

We set $\alpha_1 = \alpha_2 = 0.5$ in the implementation.

We use the BCE loss [22] in L^H which is defined in Eq. (3.7)

$$\mathcal{L}^H = -\sum [Y^H log S^H + (1 - Y^H) log(1 - S^H)], \tag{3.7}$$

where S^H and Y^H are the HSF predicted and GT maps, respectively.

We propose to use a hybrid loss for L^I, defined in Eq. (3.8)

$$\mathcal{L}^I = \beta_1 L^I_{bce} + \beta_2 L^I_{iou} + \beta_3 L^I_{hist}, \tag{3.8}$$

where $\mathcal{L}^I_{bce}$, $\mathcal{L}^I_{iou}$, and $\mathcal{L}^I_{hist}$ denote the BCE, the IoU [23], and the proposed histogram-based losses, respectively. Empirically, we set $\beta_1 = \beta_2 = \beta_3 = 1$.

$\mathcal{L}^I_{bce}$ is defined in Eq. (3.9).

$$\mathcal{L}^I_{bce} = -\sum [Y log S + (1 - Y) log(1 - S)], \tag{3.9}$$

where Y is the GT map.

The IoU loss is defined in Eq. (3.10).

$$\mathcal{L}^I_{iou} = 1 - \frac{\sum [Y * S]}{\sum [Y + S - Y * S]}. \tag{3.10}$$

IoU loss measures global structural similarity by evaluating the overlap between predicted and ground truth saliency regions.

Here, we suggest using a histogram to evaluate the similarity of the global pixel-intensity frequency distributions of Y and S. The histogram is the first-order data statistic that represents the global signal intensity distribution. The loss based on the histogram is represented in Eq. (3.11).

$$\mathcal{L}_{hist}^{I} = \sum_{1}^{N_{bin}} |H_s(i) - H_y(i)|/N_{bin}, \tag{3.11}$$

where $H_s(\ )$ and $H_y(\ )$ are the histograms of S and Y, and N_{bin} is total bin number in the histogram.

The loss based on histograms lacks spatial information. It remains zero when Y and S have different shapes but identical histograms. Incorporating a structure-ware loss function like IoU loss can complement such limitations. Therefore, combining these losses can provide a comprehensive training process for the TFGNet model.

In Section 3.4.5, we explore how these losses impact overall network performance.

3.4 Experiment and Discussion

3.4.1 Training and Testing Setting

We use ImageNet [30] to pre-train and DUTS-TR [31] to fine-tune the proposed TFGNet. Data augmentation includes horizontal flips and random crops. Five popular pixel-wise annotated SOD benchmark datasets are used for evaluation: DUTS-TE [31], DUT-OMRON [32], HKU-IS [33], Pascal-S [34], and ECSSD [35]. In training, TFGNet uses SwinV2 [36] as the backbone network. The maximum learning rate of the backbone is 0.0001, and 0.001 for other parts. The momentum is 0.9, and the weight decay is 0.001. The 'poly' learning rate strategy is also adopted. The maximum iteration number is defaulted to 99. The batch size is 12. All evaluations are performed on a server with A100 (40G) and AMD EPYC 7763 64-Core Processor (1T).

3.4.2 Evaluation Criteria

We use Mean Absolute Error (MAE) [37], maximum F-measure (denoted by F_{β}^{max}), mean F-measure (mF) [38], mean Enhanced-alignment Measure (mE_m) [39], and

S-measure (S_m) [40] to evaluate SOD models. The Precision Recall (PR) and F-measure curves are plotted to demonstrate overall performance. Further details on these metrics can be found in Chapter 2.

3.4.3 Quantitative and Qualitative Comparison

We report three TFGNet configurations using three SwinV2 backbones, referred to as TFGNet-B256, TFGNet-B384, and TFGNet-L384, to show the effectiveness of TFGNet under different settings. Here, 'B' represents the SwinV2-Base model, 'L' means the SwinV2-Large model, and '256' and '384' indicate the input image size.

The comparison models include: (i) CNN-based models: CPD [5], EGNet [13], BASNet [3], AADFNet [8], GateNet [4], MINet [41], LDF [6], U2Net [42], PA-KRN [7], SGL-KRN [7], HQSOD [11], DCN [15], SAC [43], EDN [12], TSNet [44], CFNet [45], ICON-R [10], UDNet [14] and MENet [9], and (ii) Transformer-backbone based models: VST [17], EBMG [18], SelfReformer [19], ICON-P [10], and ICON-S [10].

3.4.3.1 Quantitative evaluation

Tables 3.1- 3.3 illustrate quantitative comparison results. Models without special annotations in the CNN category use ResNet50 as their backbone. The symbols '↓' / '↑' indicate that a lower/higher value of the evaluation metric corresponds to better model performance. The symbol '-' means that the value is not available. Red, green, and blue denote the best three results. Fig. 3.5 illustrates the average values of five metrics in the five data sets. For a more comprehensive comparison, we rank all models across backbones.

The comparison demonstrates that TFGNet-L384 outperforms other models on various metrics and datasets. TFGNet-L384 reduces *MAE* scores by an impressive 12.77%. TFGNet-L384 also has better F-measure and PR curves than other approaches, as depicted in Fig. 3.6 and Fig. 3.7. TFGNet-L384 is stable in all five datasets. Furthermore, both TFGNet-B256 and TFGNet-B384 outperform other methods overall in the five data sets, reinforcing the statistical evidence of the superiority of TFGNet.

As a large model, TFGNet-L384 has more parameters than the other models. However, TFGNet-L384 needs the commendable speed of 20 frames-per-second (FPS) to

Table 3.1: Quantitative comparison on the DUT-OMRON dataset.

			OMRON [32] (5168 images)				
Year	**Method**	**Params(M)**	$MAE\downarrow$	$F_\beta^{max}\uparrow$	$mF\uparrow$	$E_m\uparrow$	$S_m\uparrow$
	ResNet						
2019	CPD [5]	47.85	0.0560	0.7966	0.7807	0.8726	0.8248
2019	EGNet [13]	111.65	0.0528	0.8155	0.7942	0.8738	0.8412
2020	BASNet(R34) [3]	87.06	0.0565	0.8053	0.7906	0.8691	0.8362
2020	AADFNet [8]	26.46	0.0488	0.8143	0.8050	0.8744	0.8389
2020	GateNet(R101) [4]	130.02	0.0547	0.8210	0.7944	0.8736	0.8449
2020	MINet [41]	162.37	0.0559	0.8098	0.7911	0.8734	0.8329
2020	LDF [6]	25.15	0.0517	0.8199	0.8015	0.8814	0.8392
2020	U2Net(RSU) [42]	87.06	0.0544	0.8226	0.8023	0.8716	0.8467
2021	PA-KRN [7]	68.69	0.0496	0.8101	0.7956	0.8880	0.8533
2021	SGL-KRN [7]	72.37	0.0492	0.7961	0.7830	0.8783	0.8464
2021	HQSOD [11]	-	0.0509	0.8160	0.8091	0.8805	0.8404
2021	DCN [15]	37.95	0.0511	0.8230	0.8056	0.8853	0.8455
2021	SAC(R101) [43]	-	0.0523	0.8287	0.8092	0.8833	0.8487
2022	EDN [12]	42.84	0.0494	0.7992	0.7880	0.8774	0.8495
2022	TSNet(R50+Vgg16) [44]	-	0.0510	0.8312	0.8083	0.8871	0.8503
2022	CFNet(R101) [45]	62.96	0.0500	0.8030	-	0.8700	0.8410
2023	ICON-R [10]	33.09	0.0569	0.8254	0.8013	0.8791	0.8445
2023	UDNet [14]	27.81	-	-	-	-	-
2023	MENet [9]	-	0.0450	0.8337	0.8178	0.8911	0.8496
	Transformer						
2021	VST [17]	44.63	0.0582	0.8245	0.7967	0.8718	0.8503
2021	EBMG [18]	118.96	0.0505	0.8386	0.8179	0.8951	0.8584
2022	SelfReformer [19]	90.70	0.0433	0.8367	0.8189	0.8928	0.8608
2023	ICON-P [10]	65.68	0.0468	0.8519	0.8228	0.8951	0.8654
2023	ICON-S [10]	94.30	0.0426	0.8546	0.8350	0.9073	0.8693
2024	**TFGNetB-256 (Ours)**	80.74	0.0438	0.8557	0.8388	0.9074	0.8717
2024	**TFGNetB-384 (Ours)**	80.74	0.0441	0.8605	0.8462	0.9087	0.8758
2024	**TFGNetL-384 (Ours)**	162.86	**0.0402**	**0.8614**	**0.8488**	**0.9118**	**0.8799**

Table 3.2: Quantitative comparison on the DUTS-TE and ECSSD datasets.

		DUTS-TE [31] (5019 images)					**ECSSD** [35] (1000 images)				
Year	**Method**	$MAE\downarrow$	$F_\beta^{max}\uparrow$	$mF\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$MAE\downarrow$	$F_\beta^{max}\uparrow$	$mF\uparrow$	$E_m\uparrow$	$S_m\uparrow$
	ResNet										
2019	CPD [5]	0.0429	0.8649	0.8431	0.9009	0.8691	0.0371	0.9393	0.9244	0.9494	0.9182
2019	EGNet [13]	0.0386	0.8880	0.8597	0.9040	0.8873	0.0374	0.9474	0.9288	0.9469	0.9246
2020	BASNet(R34) [3]	0.0472	0.8589	0.8416	0.8790	0.8660	0.0370	0.9425	0.9274	0.9210	0.9163
2020	AADFNet [8]	0.0314	0.8993	0.8911	0.9225	0.8914	0.0280	0.9543	0.9478	0.9529	0.9299
2020	GateNet(R101) [4]	0.0380	0.8919	0.8615	0.9075	0.8910	0.0357	0.9508	0.9301	0.9501	0.9302
2020	MINet [41]	0.0373	0.8833	0.8597	0.9132	0.8842	0.0342	0.9475	0.9309	0.9532	0.9250
2020	LDF [6]	0.0336	0.8968	0.8779	0.9232	0.8924	0.0335	0.9501	0.9379	0.9509	0.9245
2020	U2Net(RSU) [42]	0.0443	0.8719	0.8479	0.8840	0.8738	0.0330	0.9510	0.9325	0.9251	0.9276
2021	PA-KRN [7]	0.0328	0.8945	0.8761	0.9353	0.9005	0.0323	0.9425	0.9301	0.9503	0.9278
2021	SGL-KRN [7]	0.0337	0.8833	0.8649	0.9311	0.8929	0.0360	0.9368	0.9241	0.9462	0.9231
2021	HQSOD [11]	0.0326	0.8932	0.8867	0.9271	0.8920	0.0294	0.9520	0.9456	0.9600	0.9276
2021	DCN [15]	0.0348	0.8935	0.8742	0.9180	0.8917	0.0315	0.9524	0.9396	0.9575	0.9282
2021	SAC(R101) [43]	0.0339	0.8944	0.8732	0.9208	0.8957	0.0309	0.9512	0.9376	0.9586	0.9312
2022	EDN [12]	0.0351	0.8784	0.8634	0.9250	0.8924	0.0320	0.9410	0.9304	0.9508	0.9267
2022	TSNet(R50+Vgg16) [44]	0.0319	0.9025	0.8775	0.9249	0.8995	0.0297	0.9532	0.9391	0.9561	0.9322
2022	CFNet(R101) [45]	0.0330	0.8840	-	0.9310	0.8950	0.0300	0.9430	-	0.9540	0.9300
2023	ICON-R [10]	0.0370	0.8917	0.8665	0.9142	0.8889	0.0318	0.9503	0.9336	0.9543	0.9290
2023	UDNet [14]	0.0320	0.9060	-	0.9180	0.9020	0.0310	0.9550	-	0.9310	0.9320
2023	MENet [9]	0.0281	0.9123	0.8930	0.9368	0.9049	0.0307	0.9549	0.9422	0.9544	0.9279
	Transformer										
2021	VST [17]	0.0372	0.8898	0.8579	0.9153	0.8963	0.0337	0.9507	0.9258	0.9571	0.9323
2021	EBMG [18]	0.0288	0.9091	0.8863	0.9331	0.9088	0.0232	0.9591	0.9452	0.9632	0.9416
2022	SelfReformer [19]	0.0266	0.9155	0.8921	0.9210	0.9111	0.0273	0.9577	0.9414	0.9361	0.9356
2023	ICON-P [10]	0.0255	0.9218	0.8932	0.9386	0.9173	0.0240	0.9594	0.9432	0.9624	0.9401
2023	ICON-S [10]	0.0242	0.9196	0.8998	0.9470	0.9171	0.0235	0.9608	0.9458	0.9669	0.9414
2024	**TFGNetB-256 (Ours)**	0.0228	0.9242	0.9062	0.9500	0.9212	0.0214	0.9633	0.9496	0.9714	0.9452
2024	**TFGNetB-384 (Ours)**	0.0208	0.9340	0.9198	0.9545	0.9312	0.0198	0.9673	0.9559	0.9728	0.9493
2024	**TFGNetL-384 (Ours)**	**0.0193**	**0.9375**	**0.9238**	**0.9566**	**0.9340**	**0.0186**	**0.9679**	**0.9568**	**0.9741**	**0.9510**

Table 3.3: Quantitative comparison on the HKU-IS and PASCAL-S datasets.

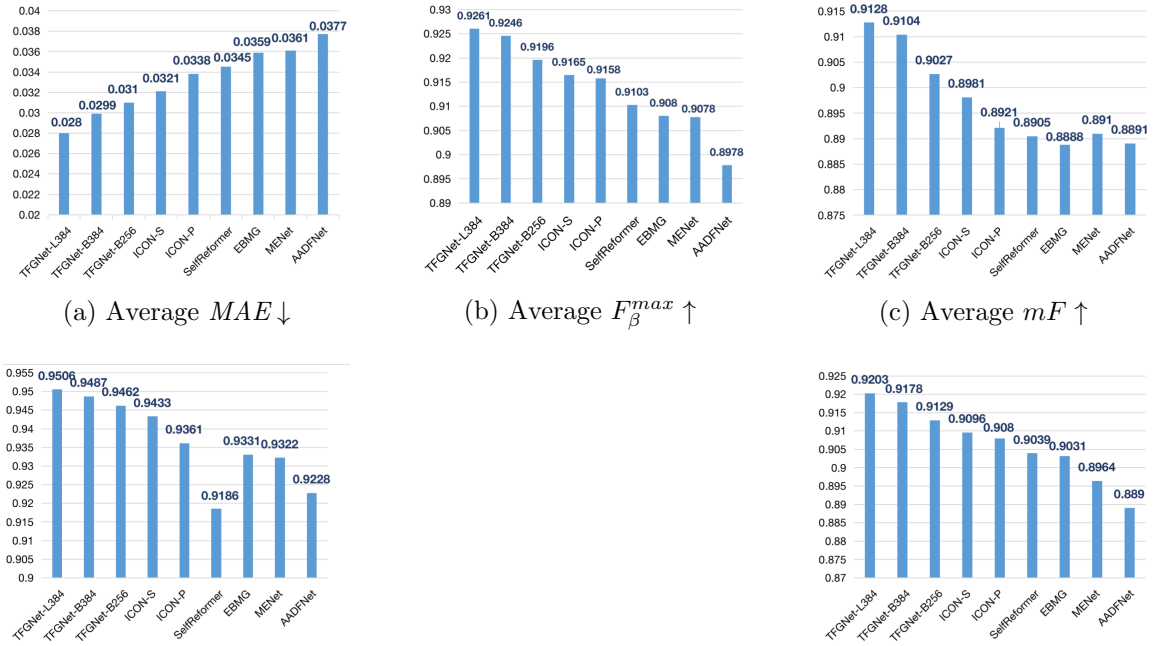
		HKU-IS [33] (4447 images)					**PASCAL-S** [34] (850 images)				
Year	**Method**	$MAE\downarrow$	$F_\beta^{max}\uparrow$	$mF\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$MAE\downarrow$	$F_\beta^{max}\uparrow$	$mF\uparrow$	$E_m\uparrow$	$S_m\uparrow$
	ResNet										
2019	CPD [5]	0.0342	0.9250	0.9047	0.9503	0.9056	0.0706	0.8595	0.8414	0.8873	0.8484
2019	EGNet [13]	0.0309	0.9352	0.9122	0.9564	0.9180	0.0740	0.8653	0.8437	0.8772	0.8521
2019	BASNet [3]	0.0322	0.9284	0.9113	0.9458	0.9090	0.0758	0.8539	0.8344	0.8527	0.8380
2020	AADFNet [8]	0.0255	0.9415	0.9339	0.9592	0.9190	0.0550	0.8797	0.8677	0.9051	0.8658
2020	GateNet(R101) [4]	0.0320	0.9375	0.9136	0.9567	0.9195	0.0668	0.8702	0.8468	0.8924	0.8622
2020	MINet [41]	0.0292	0.9349	0.9166	0.9600	0.9189	0.0643	0.8665	0.8461	0.8981	0.8563
2020	LDF [6]	0.0275	0.9394	0.9224	0.9597	0.9196	0.0596	0.8741	0.8577	0.9048	0.8630
2020	U2Net [42]	0.0312	0.9352	0.9133	0.9484	0.9161	0.0740	0.8592	0.8386	0.8500	0.8444
2021	PA-KRN [7]	0.0271	0.9349	0.9198	0.9561	0.9230	0.0665	0.8530	0.8388	0.8964	0.8578
2021	SGL-KRN [7]	0.0280	0.9301	0.9154	0.9539	0.9206	0.0678	0.8502	0.8373	0.8941	0.8556
2021	HQSOD [11]	0.0252	0.9428	0.9351	0.9639	0.9235	0.0597	0.8798	0.8698	0.9074	0.8603
2021	DCN [15]	0.0268	0.9394	0.9226	0.9624	0.9217	0.0618	0.8723	0.8543	0.9017	0.8612
2021	SAC(R101) [43]	0.0257	0.9416	0.9260	0.9636	0.9253	0.0622	0.8772	0.8585	0.9022	0.8656
2022	EDN [12]	0.0264	0.9325	0.9196	0.9548	0.9241	0.0617	0.8600	0.8489	0.9015	0.8646
2022	TSNet [4]	0.0266	0.9417	0.9220	0.9622	0.9223	0.0573	0.8800	0.8599	0.9064	0.8684
2022	CFNet(R101) [45]	0.0260	0.9380	-	0.9590	0.9270	0.0590	0.8680	-	0.9090	0.8700
2023	ICON-R [10]	0.0289	0.9395	0.9196	0.9585	0.9202	0.0644	0.8757	0.8514	0.8931	0.8611
2023	UDNet [14]	0.0270	0.9430	-	0.9560	0.9240	0.0590	0.8860	-	0.8650	0.8620
2023	MENet [9]	0.0234	0.9483	0.9319	0.9657	0.9274	0.0535	0.8896	0.8701	0.9132	0.8721
	Transformer										
2021	VST [17]	0.0297	0.9424	0.9129	0.9597	0.9283	0.0620	0.8755	0.8457	0.9024	0.8716
2021	EBMG [18]	0.0229	0.9466	0.9288	0.9673	0.9304	0.0542	0.8866	0.8659	0.9070	0.8765
2022	SelfReformer [19]	0.0241	0.9474	0.9265	0.9606	0.9310	0.0510	0.8943	0.8736	0.8825	0.8809
2023	ICON-P [10]	0.0216	0.9521	0.9325	0.9698	0.9353	0.0510	0.8927	0.8690	0.9145	0.8819
2023	ICON-S [10]	0.0216	0.9512	0.9331	0.9717	0.9355	0.0484	0.8961	0.8767	0.9237	0.8849
2024	**TFGNet-B256(Ours)**	0.0200	0.9548	0.9382	0.9738	0.9387	0.0468	0.9000	0.8806	0.9284	0.8874
2024	**TFGNet-B384(Ours)**	0.0179	0.9596	0.9456	0.9769	0.9441	0.0471	0.9018	0.8843	0.9304	0.8887
2024	**TFGNet-L384(Ours)**	**0.0176**	**0.9603**	**0.9472**	**0.9774**	**0.9449**	**0.0442**	**0.9038**	**0.8874**	**0.9332**	**0.8919**

infer an [384 × 384] image. TFGNet-B384, with fewer parameters, demonstrates a faster inference speed of 28 FPS for an image of [384 × 384]. On the other hand, TFGNet-B256 achieves an acceptable speed of 24 FPS for an image of [256 × 256]. These efficient inference speeds indicate that TFGNet is well suited for real-time SOD applications.

3.4.3.2 Qualitative comparison

We selected 15 challenging scenes for comparison, as shown in Fig. 3.8. TFGNet has superior comprehensive performance (zoom in for more details). The predicted results are given as a probability.

TFGNet achieves the most precise detection results overall, especially about high integrity (e.g., *Rows* 1, 8, 13, 14), correct localisation (e.g., *Rows* 2-5, 7, 9), low-contrast backgrounds (e.g., *Rows* 10, 11, 13) and precise boundaries (e.g., *Rows* 2, 10, 13, 15). In contrast, other models, such as ICON-S, ICON-P, and SelfReformer, have less integrity and accurate boundary predictions.



(a) Average $MAE \downarrow$ (b) Average $F_\beta^{max} \uparrow$ (c) Average $mF \uparrow$

(d) Average $E_m \uparrow$ (e) Average $S_m \uparrow$

Figure 3.5: Average values of the five metrics for five test datasets.

3.4.4 Limitation

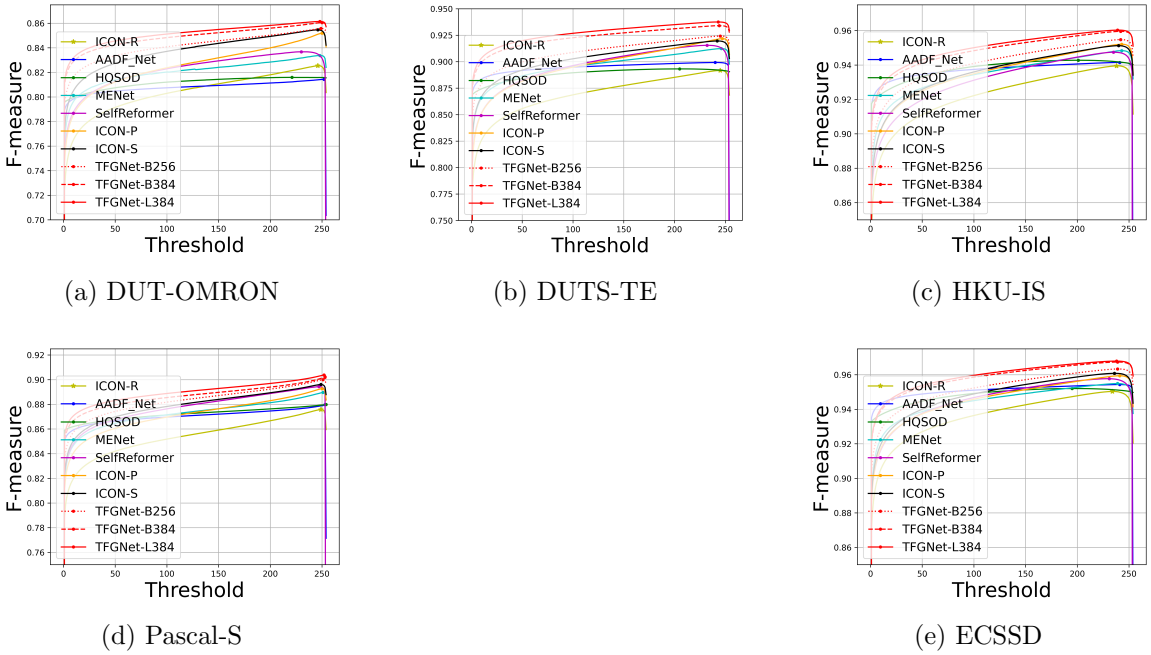
Figure 3.9 presents some representative failure cases. In these instances, TFGNet misclassified other foreground objects as salient objects. This result is mainly attributed to the ambiguity and subjective bias in the annotations present in some complex scenes [1, 44]. Such cases only occupy a small part, which causes the model to struggle with accurately detecting these scenarios. It should be noted that these challenging cases are not unique to TFGNet; other methods, such as ICON, SelfReformer, VST, and EBMG, also face difficulties in handling similar ambiguous scenes.

3.4.5 Ablation Study

3.4.5.1 Loss function

This experiment aims to verify the impact of losses ($\mathcal{L}^H$, $\mathcal{L}^L$, and $\mathcal{L}^I$) on the performance of TFGNet.

We divide the test cases into four categories: (i) Both branches are unsupervised (*Case* 1 and *Case* 2); (ii) Both branches are supervised (*Case* 3 and *Case* 4); (iii) Only the high-frequency branch is under supervision (*Case* 5 and *Case* 6). Each



(a) DUT-OMRON (b) DUTS-TE (c) HKU-IS

(d) Pascal-S (e) ECSSD

Figure 3.6: Fm-curves comparison.

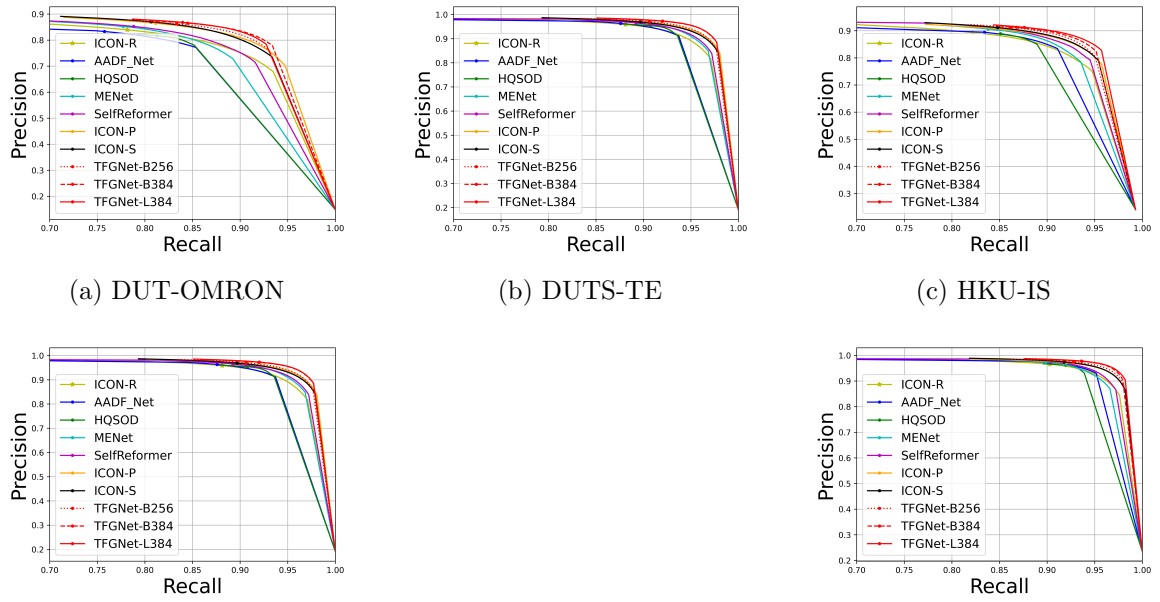
category also tests the effect of introducing the proposed histogram-based loss ($\mathcal{L}^I_{hist}$) in the integrator, as shown in Table 3.4 (red and green highlight the two best results).

Unsupervised cases for both branches (*Case* 1 and *Case* 2) yield inferior performance compared to other cases. However, when both branches are fully supervised (*Case* 3 and *Case* 4), the network may become over-supervised. It is worth noting that when the low-frequency branch is not explicitly supervised, it benefits from the implicit integrator's full-band supervision. This is due to the complementary nature of high and low frequencies. Furthermore, it has been shown that setting supervision only in the high-frequency branch (*Case* 7 and *Case* 8) leads to superior results.

In particular, regardless of the loss setting for the branches, incorporating the proposed histogram-based loss ($\mathcal{L}^I_{hist}$) in the integrator loss ($\mathcal{L}^I$) outperforms the cases without $\mathcal{L}^I_{hist}$ in most scenarios. Consequently, *Case* 6 with $\mathcal{L}^I_{hist}$ is selected as the optimal TFGNet configuration.

3.4.5.2 Network configuration

The TFGNet structure in the previous section consists of two streams: an HSF feature learning branch and an LSF feature learning branch. In addition to this, we also



(a) DUT-OMRON (b) DUTS-TE (c) HKU-IS

(d) Pascal-S (e) ECSSD

Figure 3.7: PR-curves comparison.

Table 3.4: Ablation study for loss functions.

Dataset	No.	$\mathcal{L}^H$	$\mathcal{L}^L$	$\mathcal{L}^I$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$mF \uparrow$	$mE_m \uparrow$	$S_m \uparrow$
OMRON [32]	1			$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0444	0.8592	0.8424	0.9079	0.8744
	2			$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	0.0431	0.8605	0.8443	0.9094	0.8757
	3	✓	✓	$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0432	0.8610	0.8452	0.9107	0.8760
	4	✓	✓	$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	0.0424	0.8627	0.8464	0.9099	0.8765
	5	✓		$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	**0.0398**	**0.8638**	**0.8494**	**0.9129**	0.8778
	6	✓		$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	0.0402	0.8614	0.8488	0.9118	**0.8799**
DUTS-TE [31]	1			$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0227	0.9296	0.9124	0.9511	0.9260
	2			$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	0.0224	0.9299	0.9135	0.9517	0.9267
	3	✓	✓	$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0221	0.9309	0.9145	0.9520	0.9270
	4	✓	✓	$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	0.0205	0.9341	0.9175	0.9542	0.9304
	5	✓		$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0198	0.9357	0.9197	0.9556	0.9311
	6	✓		$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	**0.0193**	**0.9375**	**0.9238**	**0.9566**	**0.9340**
HKU-IS [33]	1			$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0184	0.9580	0.9426	0.9758	0.9427
	2			$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	0.0182	0.9582	0.9435	0.9761	0.9431
	3	✓	✓	$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0183	0.9582	0.9441	0.9758	0.9429
	4	✓	✓	$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	0.0181	0.9590	0.9449	0.9765	0.9436
	5	✓		$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0183	0.9590	0.9449	0.9760	0.9428
	6	✓		$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	**0.0176**	**0.9603**	**0.9472**	**0.9774**	**0.9449**
PASCAL-S [34]	1			$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0470	0.8997	0.8800	0.9272	0.8883
	2			$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	0.0463	0.9009	0.8812	0.9281	0.8888
	3	✓	✓	$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0466	0.9017	0.8833	0.9275	0.8888
	4	✓	✓	$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	0.0458	0.9006	0.8821	0.9298	0.8897
	5	✓		$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0451	0.9025	0.8838	0.9309	0.8898
	6	✓		$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	**0.0442**	**0.9038**	**0.8874**	**0.9332**	**0.8919**
ECSSD [35]	1			$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0213	0.9647	0.9519	0.9697	0.9465
	2			$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	0.0210	0.9649	0.9535	0.9700	0.9472
	3	✓	✓	$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0213	0.9650	0.9532	0.9694	0.9467
	4	✓	✓	$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	0.0205	0.9664	0.9548	0.9717	0.9488
	5	✓		$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}$	0.0202	0.9673	0.9558	0.9719	0.9491
	6	✓		$\mathcal{L}^I_{bce}+\mathcal{L}^I_{iou}+\mathcal{L}^I_{hist}$	**0.0186**	**0.9679**	**0.9568**	**0.9741**	**0.9510**

developed a one-stream model, as shown in Fig. 3.10, to compare the effectiveness of the frequency decomposition strategy for the network structure. The one-stream setting only uses the hybrid loss for the final prediction.

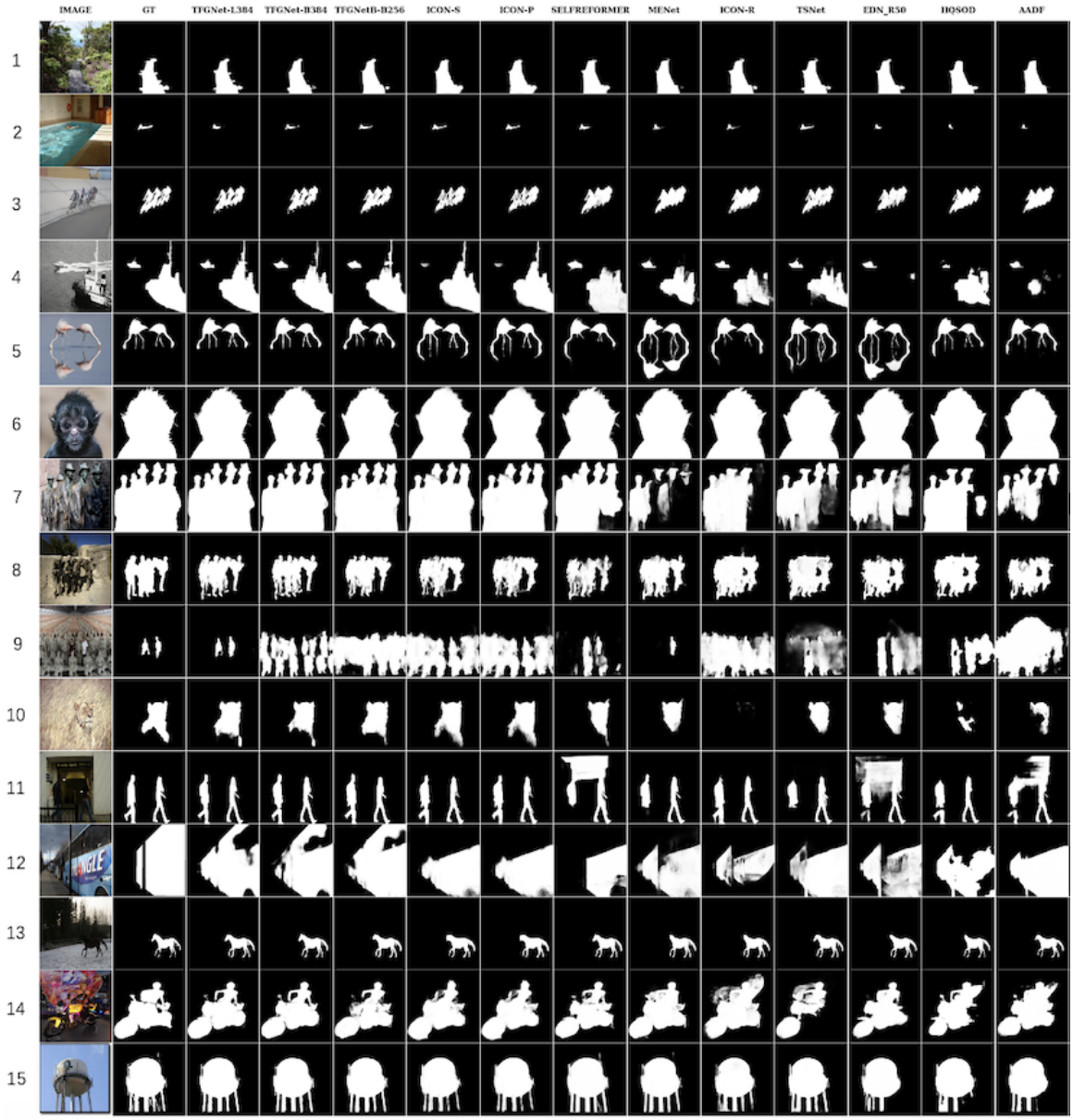


Figure 3.8: Qualitative performance comparison for complex scenes.

The experimental results are listed in Table 3.5. Observations reveal that the two-stream TFGNet setting outperforms the one-stream setting on all five datasets. Specifically, the two-stream setting achieves reductions of 12.99%, 5.85%, 1.68%, 2.64%, and 6.10% of *MAE* compared to the one-stream setting for five datasets, respectively. This demonstrates the effectiveness of the frequency decomposition strategy in the network structure, allowing the model to better capture and use the salient features of HSF and LSF, leading to more accurate and robust saliency predictions.

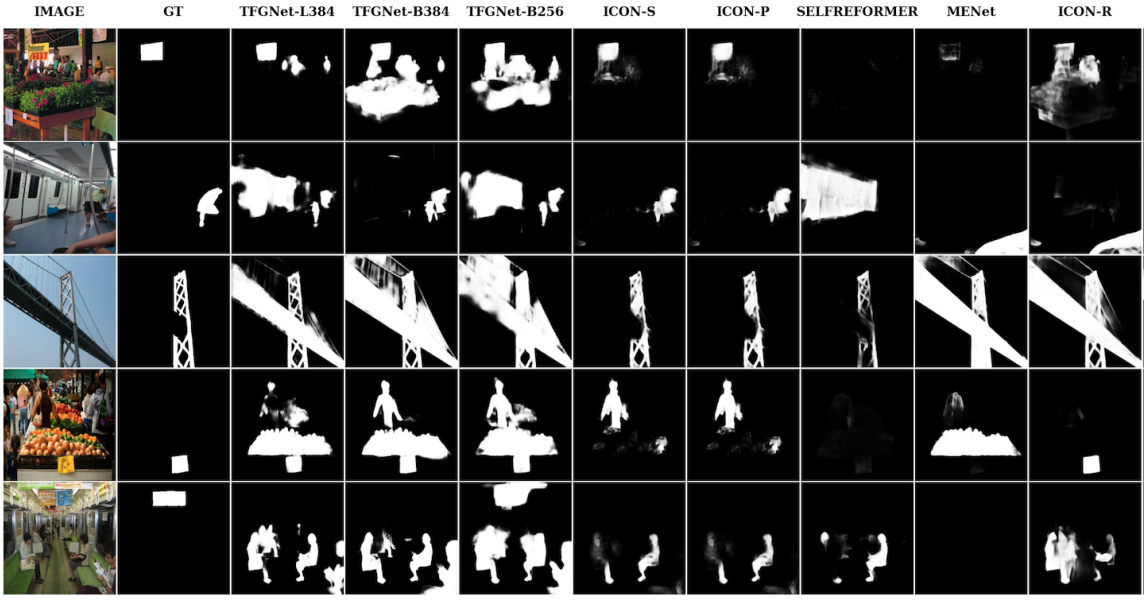


Figure 3.9: Failure cases.

Table 3.5: Configuration comparison.

Dataset	No.	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$mF \uparrow$	$mE_m \uparrow$	$S_m \uparrow$
DUT-OMRON [32]	one	0.0462	0.8602	0.8438	0.9067	0.8760
	two	**0.0402**	**0.8614**	**0.8488**	**0.9118**	**0.8799**
DUTS-TE [31]	1	0.0205	0.9348	0.9191	0.9551	0.9321
	2	**0.0193**	**0.9375**	**0.9238**	**0.9566**	**0.9340**
HKU-IS [33]	1	0.0179	0.9596	0.9445	0.9771	0.9444
	2	**0.0176**	**0.9603**	**0.9472**	**0.9774**	**0.9449**
PASCAL-S [34]	1	0.0454	0.9035	0.8824	0.9318	**0.8922**
	2	**0.0442**	**0.9038**	**0.8874**	**0.9332**	0.8919
ECSSD [35]	1	0.0196	0.9658	0.9527	0.9723	0.9491
	2	**0.0186**	**0.9679**	**0.9568**	**0.9741**	**0.9510**

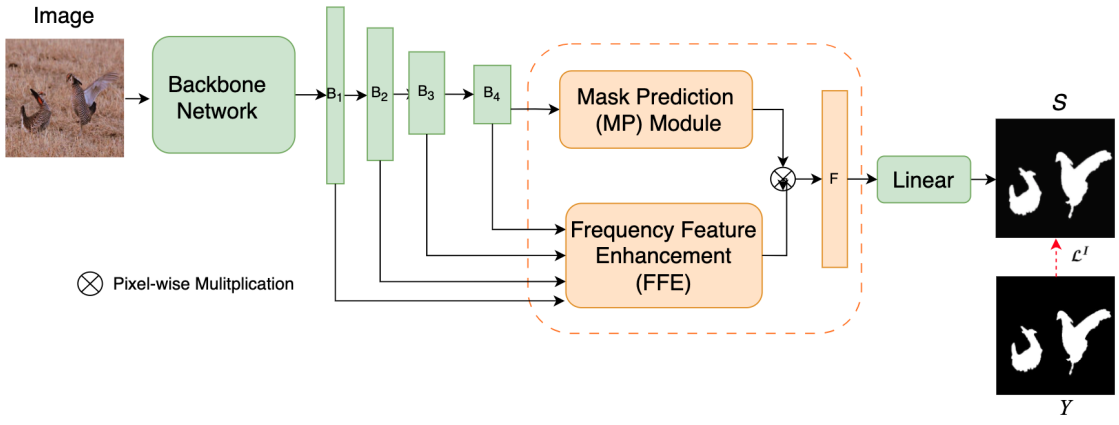


Figure 3.10: One-stream TFGNet structure.

3.5 Conclusion

We introduce TFGNet, an effective frequency-guided network for SOD based on the Transformer. TFGNet has a parallel two-branch decoder, which refines high-frequency boundary details and low-frequency inner regions of salient objects under the guidance of the decomposed frequency supervisions. TFGNet solves the problem of directly predicting the entire salience map for complex scenes. This framework also rekindles awareness of the advantages of multi-scale spatial frequency features in saliency detection.

We design a pixel decoder combined with a Transformer (TF) decoder in each frequency branch for TFGNet to improve the salient representation accuracy. The pixel decoder employs a frequency feature enhancement (FFE) module to diversify salient information of multi-scale feature maps to obtain more comprehensive and robust salient features. The TF-decoder generates mask-level discriminative features and per-mask embedding. By combining FFE with the TF decoder, we can improve local representation in the transformer network.

Furthermore, we propose an improved hybrid loss function that combines histogram dissimilarity measurement with BCE and IoU losses, leading to enhanced optimisation during training.

References

[1] 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 Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(6):3239–325, 2021.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, page 234–241, 2015.

[3] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7471–7481, Long Beach, CA, USA, 2019.

[4] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 23–28, Glasgow, UK, 2020.

[5] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3902–3911, Long Beach, CA, USA, 2019.

[6] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13022–13031, Virtual, Online, USA, 2020.

[7] 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 Conference on Artificial Intelligence*, volume 35, pages 3004–3012, 2021.

[8] Lei Zhu, Jiaxing Chen, Xiaowei Hu, Chiwing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Aggregating attentional dilated features for salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 30(10):3358–3371, 2020.

[9] Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[10] Mingchen Zhuge, Dengping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(3):3738–3772, 2023.

[11] Lv Tang, Bo Li, Yijie Zhong, Shouhong Ding, and Mofei Song. Disentangled high quality salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3580–3590, 2021.

[12] Yu-Huan Wu, Yun Liu, Le Zhang, Mingming Cheng, and Bo Ren. Edn: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing (TIP)*, 31:3125–3136, 2022.

[13] Jiaxing Zhao, Jiangjiang Liu, Dengping Fan, Yang Cao, Jufeng Yang, and Mingming Cheng. Egnet: Edge guidance network for salient object detection. In

Proceedings of the IEEE International Conference on Computer Vision (ICCV), page 8779–8788, Seoul, Korea, Republic of, 2019.

[14] Jiebin Yan Wenhui Jiang Yang Liu Yuming Fang, Haiyan Zhang. Udnet: Uncertainty-aware deep network for salient object detection. *Pattern Recognition*, 134:109099, 2023.

[15] Zhe Wu, Li Su, and Qingming Huang. Decomposition and completion network for salient object detection. *IEEE Transactions on Image Processing (TIP)*, 30:6226–6239, 2021.

[16] Xiaofang Li, Yi Wang, Tianzhu Wang, and Ruili Wang. Spatial frequency enhanced salient object detection. *Information Science*, 647:119460, 2023.

[17] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4722–4732, 2021.

[18] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. Learning generative vision transformer with energy-based latent space for saliency prediction. *Advances in Neural Information Processing Systems (ANIPS)*, 34:15448–15463, 2021.

[19] Yi Ke Yun and Weisi Lin. Selfreformer: Self-refined network with transformer for salient object detection. *arXiv preprint arXiv:2205.11283*, 2022.

[20] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems (ANIPS)*, volume 34, pages 17864–17875, 2021.

[21] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WCACV)*, pages 3560–3569, 2021.

[22] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005.

[23] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *Proceedings of the International Symposium on Visual Computing (ISVC)*, pages 234–244. Springer, 2016.

[24] 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.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need.

Advances in Neural Information Processing Systems (ANIPS), 30, 2017.

[26] Huajun Zhou, Yang Lin, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Benchmarking deep models on salient object detection. *Pattern Recognition*, 145:109951, 2024.

[27] Wenhai Wang, Enze Xie, Xiang Li, Dengping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[29] Haldo Spontón and Juan Cardelino. A review of classic edge detectors. *Image Process. Online*, 5:90–123, 2015.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems (ANIPS)*, volume 25, page 1097–1105, Red Hook, NY, USA, 2012.

[31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3796–3805, Honolulu, HI, USA, 2017.

[32] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Minghsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3166–3173, Portland, OR, USA, 2013.

[33] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5455–5463, Boston, MA, USA, 2015.

[34] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 280–287, Columbus, OH, USA, 2014.

[35] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1162, Portland, OR, USA, 2013.

[36] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022.

[37] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–740, Providence, RI, USA, 2012.

[38] Ran Margolin, Lihi Zelnik Manor, and Ayellet Tal. How to evaluate foreground maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, Columbus, OH, USA, 2014.

[39] Dengping Fan, Cheng Gong, Yang Cao, Bo Ren, Mingming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, page 698–704, Stockholm, Sweden, 2018.

[40] Mingming Cheng and Dengping Fan. Structure-measure: A new way to evaluate foreground maps. *International Journal of Computer Vision (IJCV)*, 129(9):2622–2638, 2021.

[41] Lihe Zhang, Jie Wu, Tiantian Wang, Ali Borji, Guohua Wei, and Huchuan Lu. A multistage refinement network for salient object detection. *IEEE Transactions on Image Processing (TIP)*, 29:3534–3545, 2020.

[42] 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. *Pattern Recognition*, 106:107404, 2020.

[43] Xiaowei Hu, Chi Wing Fu, Lei Zhu, Tianyu Wang, and Pheng Ann Heng. Sacnet: Spatial attenuation context for salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 31(3):1079–1090, 2021.

[44] Zhenyu Wu, Shuai Li, Chenglizhao Chen, Aimin Hao, and Hong Qin. Deeper look at image salient object detection: Bi-stream network with a small training dataset. *IEEE Transactions on Multimedia (TMM)*, 24:73–86, 2022.

[45] Xinyu Yan, Meijun Sun, Yahong Han, Zheng Wang, and Qi Tian. Effective full-scale detection for salient object based on condensing-and-filtering network. *Pattern Recognition*, 131:108904, 2022.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Yi Wang
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 4

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Yi Wang, Ruili Wang, Xiangjian He, Tianzhu Wang, WBNet: Weakly-Supervised Saliency Detection via Scribble and Pseudo-Background Priors, Pattern Recognition, 2024: 110579.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:
- The percentage of the manuscript/published work that was contributed by the candidate:
- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	wangyi 数字签名者：wangyi 日期：2024.04.02 11:47:30 +13'00'
Date:	02-4 月-2024
Primary Supervisor's Signature:	Prof Ruili Wang Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, c=NZ, o=Massey University , ou=School of Natural and Computational Sciences , email=ruili.wang@massey.ac.nz Date: 2024.04.02 17:17:56 +13'00'
Date:	2-4 月-2024

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 4

Multi-Source Weakly-Supervised Saliency Detection

Weakly supervised salient object detection (WSOD) methods endeavor to boost sparse labels to get more salient cues in various ways. Among them, an effective approach is using pseudo labels from multiple unsupervised self-learning methods, but inaccurate and inconsistent pseudo labels could ultimately lead to detection performance degradation. To tackle this problem, we develop a new multi-source WSOD framework, WBNet, that can effectively utilise pseudo-background (non-salient region) labels combined with scribble labels to obtain more accurate salient features. We first design a comprehensive salient pseudo-mask generator from multiple self-learning features. Then, we pioneer the exploration of generating salient pseudo-labels via point-prompted and box-prompted Segment-Anything Models (SAM). Then, WBNet leverages a pixel-level Feature Aggregation Module (FAM), a mask-level Transformer-decoder (TFD), and an auxiliary Boundary Prediction Module (EPM) with a hybrid loss function to handle complex saliency detection tasks.

4.1 Introduction

Salient object detection (SOD) based deep learning requires pixel-wise dense ground-truth (GT) labels. However, obtaining such labels is expensive, time-consuming, or infeasible in some practical scenarios [1, 2]. To alleviate this burden, weakly

supervised SOD methods based on sparse labeling [1, 3] have been explored, which can achieve competitive performance with limited annotated data.

Sparse labelling refers to a small subset of pixels marked as salient or non-salient regions in an image. Scribbles, bounding boxes, points, categories, and captions are the sparse labels used for WSOD. Considering scribbles provide accurate salient foreground (salient region) and background (non-salient region) information but have similar annotation costs as other sparse labels [4], we focus on scribble-based WSOD in this chapter.

Scribble labels are comprised of two distinct strokes or lines: one is inside the salient region, and the other is inside the non-salient region. The inherent absence of salient information, such as boundaries and structural details, poses a significant challenge when training a high-performance WSOD model using scribble labels. To address this limitation of scribble labels, a practical approach is using pixel-wise pseudo labels as supplementary cues for scribbles [4, 5]. Some prior studies have employed Class Activation Maps (CAMs) [6] to synthesise pseudo labels from image-level category labels [7, 8]. Other methods leverage the rich appearance information available in RGB images to refine CAMs [9]. However, these pseudo labels can often appear fuzzy and imprecise, as illustrated in Fig. 4.1. In this figure, we also display the saliency pseudo masks generated by the proposed S-PMG module solely using a single self-learning model (denoted by S-PMG (MoCoV2), S-PMG (DINO), or S-PMG (SwAV)) and the combination of these three self-learning methods (denoted by S-PMG (Multiple Models)), as well as the Segment-Anything Model (SAM) [10] (with either points or a box as the prompt). For scribble labels, we colour the foreground (salient region) mask red, the background (non-salient region) mask green, and the black area indicates unlabelled regions. For other masks, the grey value represents the probability of being the foreground and the black value represents the background. As we can see, pseudo-labels generated from different sources (or models) of an image exhibit diversity.

More recently, unsupervised self-learning models have made significant progress (e.g., DINO [11], SwAV [12], and MoCov2 [13]) for detection and segmentation tasks. The features learnt from these models can be used to generate pseudo-labels for SOD [14]. In addition, pseudo labels generated from self-learning features of various models exhibit variations, as depicted in Fig. 4.1. To mitigate the biases associated with a single

model, it can be useful to take advantage of this diversity of pseudo labels [7]. However, it is essential to note that these pseudo-labels sometimes exhibit inconsistencies. For example, certain areas may be identified as salient by one type of pseudo-label but non-salient according to another. Similarly, there may be contradictory issues between pseudo labels and precise sparse labels, which can potentially impact the model's overall performance when used together.

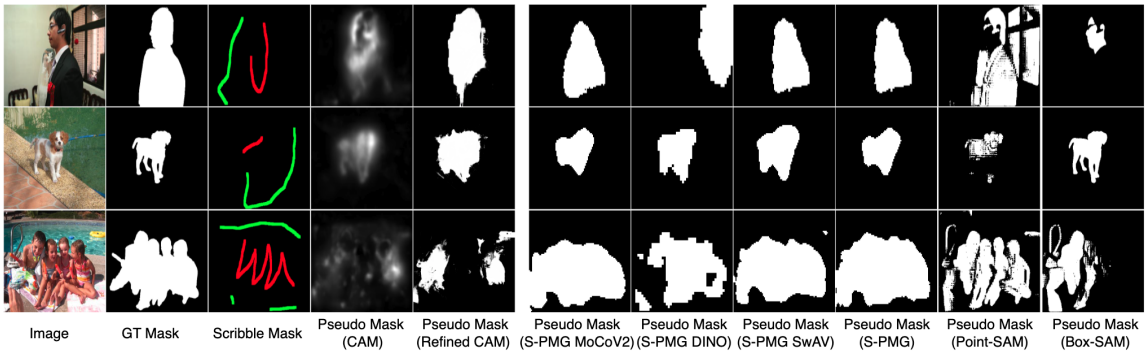


Figure 4.1: Illustration of various annotations in the saliency detection task.

This challenge motivates us to effectively investigate how to leverage more comprehensive salient information via multi-source pseudo-labels and how to make pseudo-labels consistent with scribble labels to boost performance. Specifically, we propose a self-learning feature-based pseudo mask generator (S-PMG). This generator employs clustering techniques, such as spectral clustering [15] and k-Means, to derive candidate masks from various self-learning features from multiple models. Subsequently, we design saliency filtering and selection strategies incorporating various salient constraints to yield more comprehensive pseudo masks compatible with scribble labels.

We also observe that pseudo-background labels are more functional and robust for completing scribble labels than pseudo-full labels. This discrepancy arises because, in most scenarios, the background region typically occupies a larger area than the foreground, sometimes encompassing the foreground. This expansive background area poses a challenge when trying to manually encompass all relevant background features using lines or strokes alone in the annotation process. Consequently, scribble background labels may not be effective in facilitating feature learning. As evidenced in Fig. 4.2 (where the foreground masks are highlighted red and the background masks are highlighted green for clarity), there are instances where essential elements, such as the dark forest in *Row 1* and the waves in *Row 2*, are not included in the scribble

labels. To address this limitation, we propose augmenting scribble-background labels with pseudo-background labels. Pseudo-background labels are advantageous as they encompass a broader spectrum of background content, thereby embedding more comprehensive background features that help distinguish non-salient areas more effectively.

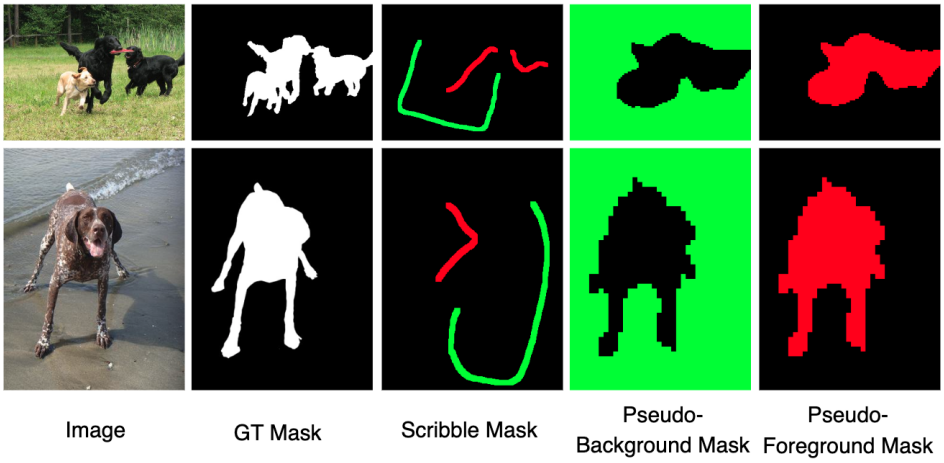


Figure 4.2: Illustration of the proposed pseudo-background enhancing scribble masks.

In addition, we explore the generation of pseudo-labels leveraging the Segment-Anything Model (SAM) [10]. This technique employs salient prompts effectively extracted from the bounding boxes of pseudo-foreground masks derived from self-learning methods and taking points as prompts from foreground scribble annotations, as shown in Fig. 4.1. While this utilisation may not strictly adhere to the conventional weakly supervised SOD concept, it is valuable to explore methods that harness pre-trained large models. We consider this exploration an initial foray into this domain, recognizing the potential for future advancements.

Overall, we propose an innovative Transformer-based weakly supervised SOD network, named WBNet, which can effectively harness pseudo labels and scribble labels in improving performance. Inspired by MaskFormer [16], which was initially designed for semantic / instance segmentation, the main salient feature learning stream of our framework consists of a Transformer-decoder (TFD) and a Pixel-Level Feature Aggregation Module (FAM). TFD captures global contextual information, while FAM propagates and aggregates features across multiple scales. These two modules enable the network to efficiently grasp intricate specifics and a broader universal perspective of saliency regions. In addition to the core salient feature learning, we incorporate a

Boundary Prediction Module (EPM) to recover structural information and boundary details within salient features. Then, we design a comprehensive hybrid loss function to evaluate the prediction with multi-source GT maps.

The contributions of this chapter are concluded as follows.

- We propose a highly effective weakly supervised salient object detection network named WBNet. A pixel-level feature accumulation module (FAM), a mask-level Transformer-decoder (TFD), and an auxiliary Boundary Prediction Module (EPM) are incorporated into this network to predict saliency information with a comprehensive hybrid loss function utilising scribble and pseudo-background labels.
- We design a self-learning feature-based Pseudo-Mask Generator (S-PMG) that utilizes multi-source self-learning features, clustering techniques, and saliency-priors filtering strategies to produce comprehensive pseudo-masks that align consistently with scribble annotations.
- For the first time, we employ the Segmentation Anything Module (SAM) to generate pseudo-masks for WSOD. We design two effective prompt generation methods: one relies on foreground scribble points, and the other leverages a bounding box derived from pseudo-labels obtained through self-learning.
- We extensively evaluate WBNet on five widely used benchmark datasets with recent SOD and WSOD methods. The results demonstrate that WBNet significantly surpasses other WSOD models and could be compared to some SOD methods in complex scenarios.

The following content of this chapter is: Section 4.2 briefly reviews recent WSOD methods. Section 4.3 details the proposed WBNet. Section 4.4 gives quantitative and qualitative experiments. Section 4.5 is the conclusion of the chapter.

4.2 Related Work

Weakly supervised SOD (WSOD) methods have become a promising way to address the challenges associated with obtaining pixel-level dense annotations in SOD. The following brief overviews analyse and summarise recent strategies in this field.

4.2.1 Sparse Annotation

Early works concentrated on identifying influential sparse labels, such as image classification labels, bounding boxes, scribbles, and captions for WSOD. The researchers delved deeply into the strengths and limitations of each type of sparse label.

In 2017, Wang et al. [17] introduced the first WSOD model using image-level labels. A foreground inference network is first trained based on FCNs [18]. The network is then fine-tuned with the results in the first stage as GT maps using an iterative Conditional Random Field (CRF) [19]. In 2020, WSSA [3] first utilized scribble annotations in WSOD through an auxiliary boundary prediction task and a gated structure-aware loss. In 2021, SCWSSOD [4] aggregated multi-level features with a loss of consistency of the saliency structure, ensuring consistent saliency maps. PSOD [20] extended DUTS [17] with point labels (PDUTS). A Transformer-based model is used to generate the initial predicted maps. Then, Non-Salient Suppression (NSS) optimised erroneous saliency maps in the second training round. In 2023, Zhao et al. [5] used a Cluster-based Scribble Inference (CSI) and Pooling-based Scribble Inference (PSI) to boost scribble labels, and these boosted labels assisted their WSOD model in achieving better performance.

Single-source sparse labels, such as scribbles, often provide only coarse-level annotations, lacking precise information about the exact position and boundaries of salient objects. Consequently, weakly supervised methods relying on these annotations may suffer from issues like over-segmentation or under-segmentation of salient regions. To mitigate the challenges posed by limited supervision, single-source weakly supervised methods sometimes necessitate additional components like a label-boosting network or an auxiliary edge prediction network. Although these components improve model performance, they make the training process more complex and resource-intensive.

4.2.2 Multi-Source Annotation

Recent advances focus on combining multiple label sources, leveraging self-supervised learning, and exploring novel loss functions to improve robustness and accuracy.

In 2021, MFNet [7] synthesised pixel-wise and superpixel-wise pseudo-labels from CAMs based on an image-level classification network. MFNet also uses multiple directive filters to get more accurate predictions from a few noisy pseudo-annotations.

MWS [21] used category labels and captions to produce pseudo-pixel-level labels in one stage and then used image pairs synthesised from Web images to train an SOD network using attention, shift loss, and consistency loss. In 2022, NSAL [8] was developed to guide SOD with pseudo labels obtained from the classification network and a noise-robust discriminator network. Hybrid-SOD [22] incorporated pseudo labels from unsupervised methods and 10% real labels and iteratively trained a coarse label refinement network (R-Net) and a SOD network (S-Net). Li et al. [23] used limited-labelled datasets and unlabelled datasets as training datasets to train a classification network (MFRN) to get boosted labels. In the second phase, these boosted labels supervise a salient region prediction network (SORN) with an edge enhancement branch.

While multi-source label-based methods offer advantages over single-source label-based approaches, they also have weaknesses and limitations. On the one hand, integrating diverse and potentially conflicting annotations from different sources may introduce noise and uncertainty, making the label fusion process more intricate. In addition, ensuring consistency between weak labels from various sources can be difficult, as each source may have biases and inaccuracies. Inconsistent labels can lead to conflicting information during training, potentially hampering the model's learning process. Addressing these weaknesses requires careful consideration and design of the label integration process and the learning strategies.

To tackle the above-mentioned challenges, we present an approach to enhance prediction accuracy by combining multi-source pseudo-background labels with scribble labels. This method has three new features: (i) We harness pseudo-saliency cues generated by multiple self-supervised SOD models within saliency constraints, ensuring their comprehensiveness and alignment with scribble labels. We propose only to use pseudo-background masks to mitigate the influence of inaccuracies in pseudo-foreground labels on precise sparse labels; (ii) We introduce background saliency cues from large-scale pre-trained model SAM into the framework, which marks the inaugural application of a large-scale model in WSOD; (iii) Our primary saliency detection network comprises a pixel-level decoder, a Transformer decoder, and an edge prediction module. This combination obtains details at the local pixel level, global contextual information, and structural and boundary features. Moreover, we designed a comprehensive hybrid loss function for WBNet, encompassing a scribble loss, a disparity smoothness loss, a pseudo-background loss, a local saliency consistency loss,

and an edge loss, to assess predictive performance from various perspectives. Experimental results demonstrate that WBNet improves the saliency detection performance for WSOD tasks.

4.3 Methodology

The schematic diagram of WBNet, as depicted in Fig. 4.3, comprises two main parts: a pseudo mask generation module and a saliency prediction network. In the following, we introduce these two components.

4.3.1 Pseudo-Mask Generation

This section explains how three types of pseudo masks are generated through the feature learning of self-learning models, box-prompted SAM, and point-prompted SAM. This also explains why the backgrounds of these pseudo labels are used to supplement scribble labels.

4.3.1.1 Self-learning pseudo-mask generator (S-PMG)

Many unsupervised self-learning SOD models have emerged to mitigate the need for extensive human annotation [1]. However, the absence of ground-truth (GT) labels results in a notable disparity in features generated by different self-supervised learning networks. Drawing inspiration from unsupervised approaches like Self-mask [14], we leverage three self-learning models (i.e., DINO [11], MoCoV2 [13], and SwAV [12]) to amalgamate the strengths of each.

We first employ several self-learning models to generate image features. Subsequently, we utilize clustering algorithms, such as spectral clustering [15] and k-Means, to produce multiple candidate masks for each self-learning model. Figure **??** illustrates the process of generating 9 candidate masks (27 in total from three models) for each model—DINO, MoCoV2, and SwAV—using different numbers of clusters (i.e., $k = 2, 3, 4$). Following this, all candidate masks are fed into a Filtering and Selection Module (FSM) to output the mask that best conforms to the SOD annotation criteria. The filtering procedure unfolds as follows.

Step 1: Saliency Filtering. Following saliency principles, salient regions are typically situated nearer the center of an image and rarely extend beyond its bound-

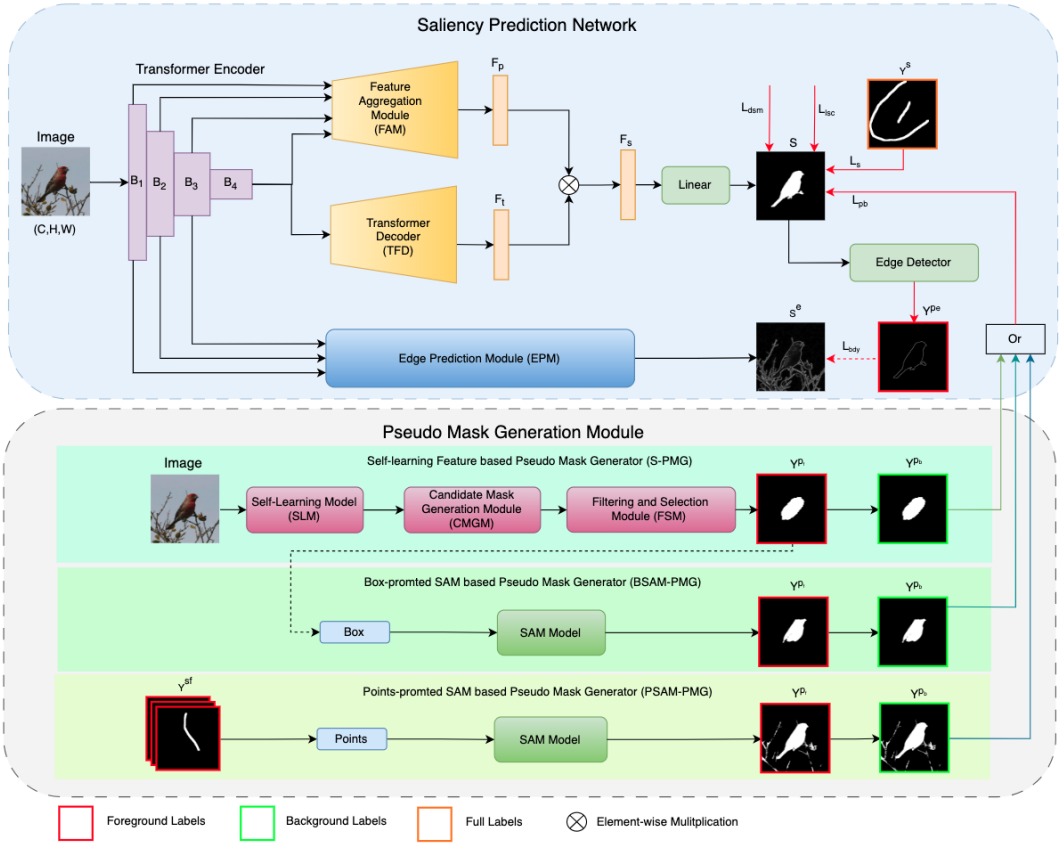


Figure 4.3: Schematic diagram of WBNet, comprising a module for generating pseudo masks and a network for saliency prediction.

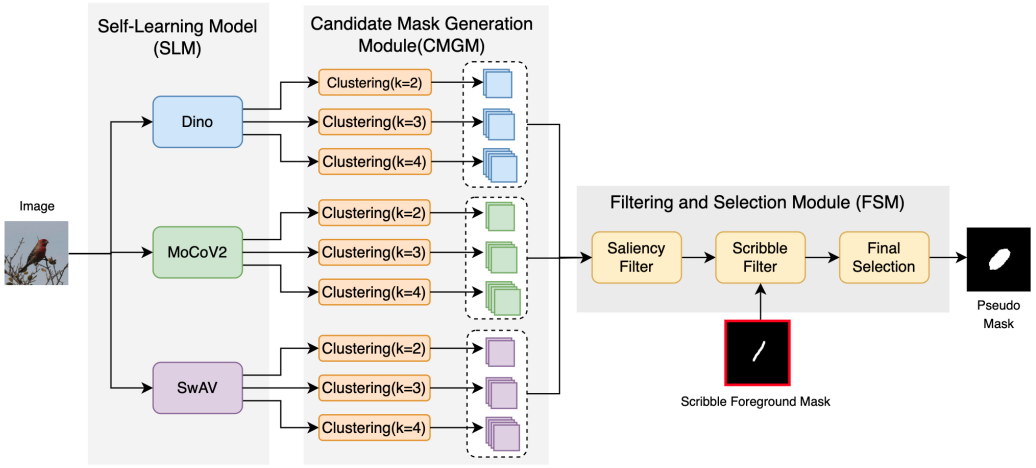


Figure 4.4: Illustration of the Self-learning Pseudo-mask Generator (S-PMG).

aries [1]. We select masks with the shortest average distance from their constituent pixels to the image's center. Additionally, we discard candidate masks that intersect with the image's width and/or length boundaries.

Step2: Scribble-Label based Filtering. We utilize the foreground (salient region) of scribble labels as the criterion for the saliency filtering process, expecting that the predicted pseudo masks would cover the foreground scribble mask. In this regard, candidate masks that do not completely encompass the foreground scribble mask are excluded from consideration. If none of the candidate masks meet this criterion, all proceed to the next selection process. This approach distinguishes our filtering procedure from Self-Mask, which operates without prior scribble labels and relies solely on the inherent saliency characteristics of the images in its pseudo-label filtering strategy.

Step 3: Final Selection. Among the remaining candidate masks, we adopt a selection criterion similar to the approach used in Self-mask [14] to remove masks that are excessively elongated and deviate significantly from the centre. To be more specific, the mask that shows the most excellent average pairwise similarity, identified as *IS*, is selected via the Intersection over Union (IoU) [24] operation, computed in Eq. (4.1).

$$IS = \arg \max_{i \in [1,\ldots,n]} \{ \frac{{M_i}^T \bigcap M_i}{{M_i}^T \bigcup M_i} \}, \tag{4.1}$$

where M_i represents the candidate mask, it can also be regarded as a matrix. ${M_i}^T$ denotes the transpose of M_i. We randomly select one when multiple masks share identical *IS* scores.

From the above description, the structure of S-PMG adopts a modular design, allowing the incorporation of any number of self-learning models, clustering methods, and filtering strategies, making it easily scalable.

4.3.1.2 SAM-based pseudo-label generation

We utilise SAM [10], the Segmentation Anything Module, to generate pseudo labels. SAM is a prompt-based segmentation model that works with different prompts, such as points, rectangular boxes, masks, or texts.

Point-prompted SAM-based pseudo-mask generator (PSAM-PMG): We

utilize foreground scribble annotations to act as point-prompts to generate pseudo-labels. SAM can simultaneously receive these points to create a segmentation result. We refer to the WBNet with this pseudo-label type as **WBNet-PSAM**.

Box-prompted SAM-based pseudo-mask generator (BSAM-PMG): For the box-prompted generator, we avoid using weakly supervised GT (e.g., scribbles) to directly create boxes of salient regions, as they may not cover the entire object accurately. Instead, we use the full-pseudo mask output from the Self-learning Pseudo-Mask generator (S-PMG) to compute the box prompts. We denote WBNet with this pseudo-label type as **WBNet-BSAM**.

4.3.2 Full Pseudo-labels versus Background Pseudo-labels

We provide further discussion on pseudo labels below, starting with additional examples generated by the S-PMG, Box-SAM, and Point-SAM modules as illustrated in Fig. 4.5, where the foreground and background labels are highlighted red and green, respectively, for clear.

A discrepancy exists between the pseudo masks (S-PMG masks, Box-SAM masks, and Point-SAM masks) and the pixel-level ground-truth (GT) masks when predicting the foreground region, i.e., the saliency region. Specifically, the S-PMG mask's foreground regions contain more background pixels around boundaries or lack foreground pixels in certain parts. While Point-SAM and Box-SAM masks provide more detailed information than S-PMG masks around boundaries, Point-SAM foreground masks exhibit a textured pattern of discrete point collections and include extra background pixels in their foreground regions. However, Box-SAM masks omit some foreground pixels. Consequently, utilising such inaccurate foreground pseudo labels to augment accurate scribble foreground labels may introduce errors.

From the examples in Fig. 4.5, we also observe that the background prediction is generally more accurate than the foreground prediction by the S-PMG, Box-SAM and Point-SAM modules. While the primary goal of SOD is the accurate foreground, SOD is a binary classification problem that involves predicting both foreground (salient) and background (non-salient) regions. Therefore, the richness and precision of background information play a crucial role in effectively distinguishing non-salient areas and, consequently, improving the prediction of foreground (salient) regions. Considering the inaccuracies in predicting foreground pseudo-labels, we propose only extend-

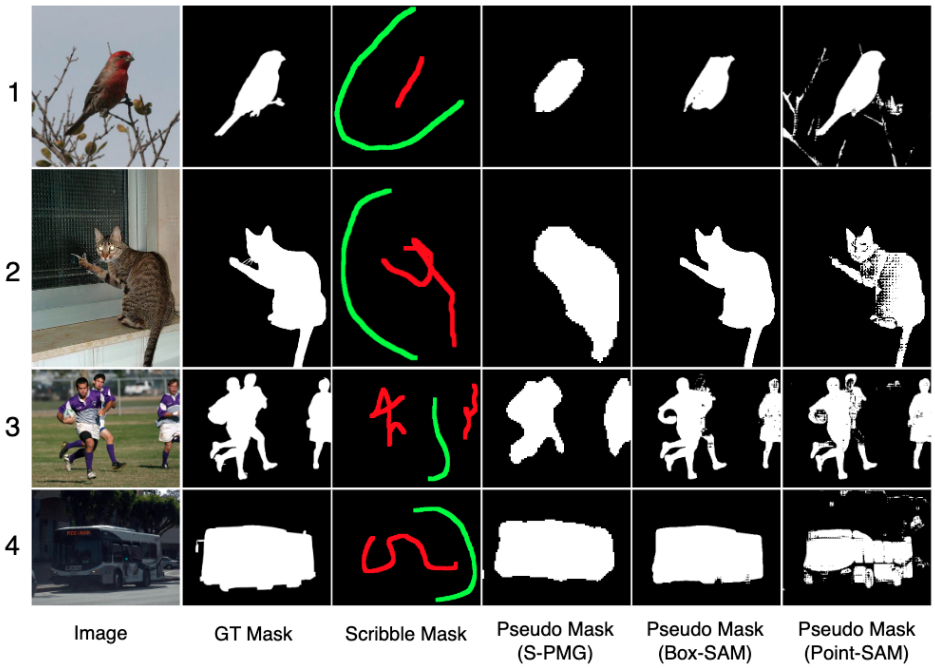


Figure 4.5: Comparison of pseudo labels generated from S-PMG, Box-SAM, and Point-SAM modules.

ing scribble background labels with pseudo-background masks. In Section 4.4.4.2, we will perform numerical experiments to demonstrate that pseudo-background labels significantly enhance prediction accuracy compared to pseudo-full labels.

4.3.3 Saliency Prediction Network

As Fig. 4.3 shows, WBNet has a Transformer Encoder, a pixel-level Feature Aggregation Module (FAM), a Transformer Decoder (TFD), an edge prediction module (EPM) and supervision. Next, we delve into each component in detail.

4.3.3.1 Encoder

Given an image $I \in \mathcal{R}^{C\times H\times W}$, multi-scale feature blocks with increasing channels (C) and decreasing sizes ($H \times W$) are first generated. Considering Transformer backbones (e.g., SwinV2 [25]), produce four-stage of feature blocks; we denote them by $B_i \in \mathcal{R}^{C_i\times H_i\times W_i}(i = 1, ..., 4)$ in our Transformer Encoder.

4.3.3.2 Decoder

Inspired by MaskFormer [16], originally designed for semantic / instance segmentation tasks, the primary salient feature learning pathway in our network consists of a pixel-level Feature Aggregation Module (FAM) and a Transformer Decoder (TFD). In contrast to MaskFormer, we additionally incorporate an Edge Prediction Module (EPM) to enhance boundary precision in saliency prediction.

Transformer Decoder (TFD): generate per-segment embedding (denoted by F_t) by standard from input image features and positional queries similar to Maskformer [16]. The transformer creates class predictions based on global information collected from all image features, alleviating the need for the per-pixel module for heavy context aggregation.

Feature Aggregation Module (FAM): In addition to the upsampling and gradually fused operation of the pixel-level decoder in the Maskformer, we incorporate multi-scale channel attention (MS-CAM) [26] similar to MENet [27] at each intermediate stage of aggregation to enhance salient features' ability to handle various size objects, as depicted in Fig. 4.6. Subsequently, F_p has the exact resolution as the first feature block (i.e., B_1) with C_p channels (C_p is set to 256 in experiments). Then, the salient feature F_s is obtained by matrix multiplication between F_p and F_t, followed by Sigmoid activation. We then apply a $[1 \times 1]$ convolutional layer and an up-sampling layer, generating the final prediction $S \in \mathcal{R}^{1\times H\times W}$.

Edge Prediction Module (EPM): Edge prediction significantly impacts WSOD because it contributes to recovering structural information and enhancing boundary details. Many models [3, 20] use edge prediction as an auxiliary task to assist salient feature learning. Our edge prediction module uses B_1, B_2, and B_3 feature blocks as inputs, as illustrated in Fig. 4.7. We first map each feature block into a channel of C_e (Empirically, C_e is set to be 32) by convolutions and the ReLU activation. We then upsample them into $[H \times W]$ scale and concatenate them to 96 channels. Next, a Residual Channel Attention Block (RCAB) [28] is used to suppress the non-edge information, and a classifier is used to finally produce the edge map $S^e \in \mathcal{R}^{1\times H\times W}$.

Since ground truth boundary labels are not available, we adopt an edge detector to generate pseudo boundary labels, denoted as Y^{pe}, derived from the final prediction of the saliency map S. Subsequently, Binary Cross-Entropy (BCE) loss [29] is adopted

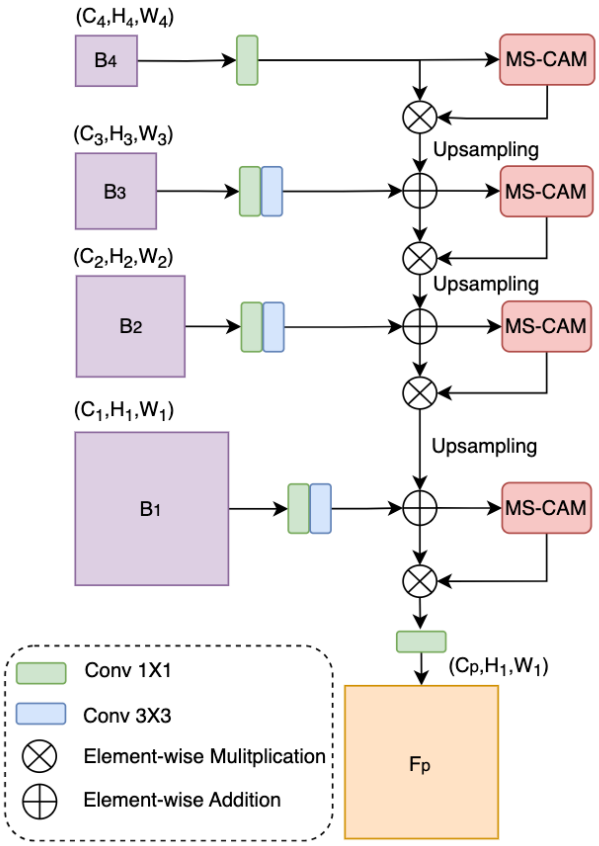


Figure 4.6: Illustration of the Feature Aggregation Module (FAM).

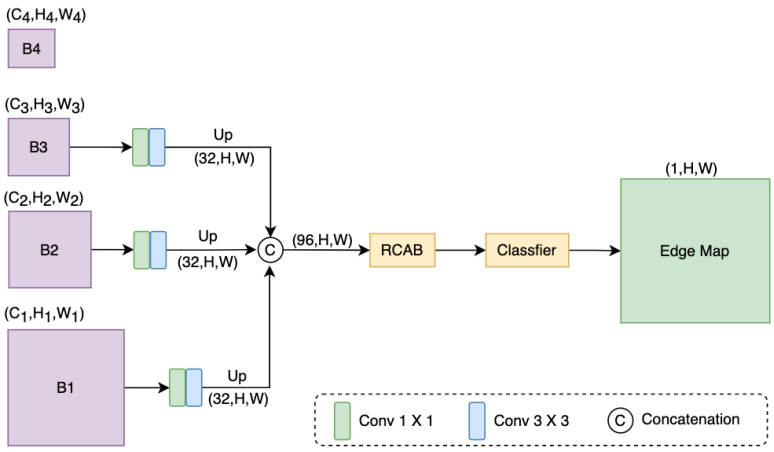


Figure 4.7: Illustration of the Boundary Prediction Module (BPM).

to compute $\mathcal{L}_{bdy}$ by Eq. (4.2).

$$\mathcal{L}_{bdy} = -\sum Y^{pe} log S^{e} + (1 - Y^{pe}) log(1 - S^{e}), \tag{4.2}$$

where S^e is the predicted boundary map. In particular, we separate $\mathcal{L}_{bdy}$ from Y^{pe}.

4.3.3.3 Training objective

WBNet uses a hybrid loss ($\mathcal{L}$) in training, as depicted in Fig. 4.3. $\mathcal{L}$ is composed of a scribble loss $\mathcal{L}_s$, a smoothness loss of the disparity $\mathcal{L}_{dsm}$, a pseudo-background loss $\mathcal{L}_{pb}$, and a local saliency consistency $\mathcal{L}_{lsc}$ for the saliency map S, as well as an edge loss L_{bdy} for the EPM branch, as defined in Eq. (4.3).

$$\mathcal{L} = \alpha_1 \mathcal{L}_s + \alpha_2 \mathcal{L}_{dsm} + \alpha_3 \mathcal{L}_{pb} + \alpha_4 \mathcal{L}_{lsc} + \alpha_5 L_{bdy}, \tag{4.3}$$

Empirically, we set $\alpha_1 = 2(H \times W)/(N_{sf} + N_{sb})$, $\alpha_2 = 0.3$, $\alpha_3 = 0.05$, $\alpha_4 = 1$, and $\alpha_5 = 1$. Here, N_{sf} is the pixel number of the foreground of a scribble mask, while N_{bf} is the pixel number of all background labels for scribble masks.

As L_{bdy} has been introduced in Eq. (4.2), the following provides details for the other four loss functions.

Scribble Loss $\mathcal{L}_s$: Partial cross-entropy loss is adopted for computing $\mathcal{L}_s$ for S and the scribble GT maps, which is in Eq. (4.4).

$$\mathcal{L}_s = -\sum_{i \in PS} [Y_i^s log S_i + (1 - Y^s) log(1 - S_i)], \tag{4.4}$$

where PS represents scribble labels and Y^s denotes the scribble GT maps.

Disparity Smoothness Loss $\mathcal{L}_{dsm}$: We use an edge-aware disparity smoothness penalty to let the salient region be similar to the values in its neighbour with the closest appearance. The disparity smoothness loss is defined by Eq. (4.5).

$$\mathcal{L}_{dsm} = \frac{1}{N} \sum [|\partial_x S| e^{-\|\partial_x I_g\|} + |\partial_y S| e^{-\|\partial_y I_g\|}], \tag{4.5}$$

where I_g is the greyscale version of the input image I and N denotes the total pixel count in S. The symbols ∂_x and ∂_y represent the partial derivative with respect to x

and y, respectively.

Pseudo-background Loss $\mathcal{L}_{pb}$: this loss consists of partial cross-entropy loss [30] for S and background pseudo-masks Y^{pb}. $\mathcal{L}_{pb}$ is defined by Eq. (4.6).

$$\mathcal{L}_{pb} = -\sum_{i \in PBG} [Y_i^{pb} log S_i + (1 - Y_i^{pb}) log(1 - S_i)], \tag{4.6}$$

where PBG is the pixel set of the pseudo background masks.

Local Saliency Coherence Loss $\mathcal{L}_{lsc}$: To obtain better precision and enforce boundary pixels that have consistent saliency scores, we follow SCWSSOD [4] and use local saliency coherence loss for S. Here, we take the original images as GT maps. The input and output are resized to a quarter of the original size to make it more efficient. Therefore, $\mathcal{L}_{lsc}$ is defined by Eq. (4.7).

$$\mathcal{L}_{lsc} = \sum_{p_i} \sum_{p_j \in K_i} F(p_i, p_j) D(p_i, p_j), \tag{4.7}$$

where K_i represents a $[k \times k]$ kernel around pixel i; $D(i, j) = |S_i - S_j|$ is the salient difference between pixels p_i and p_j computed by L_1 distance; S_i and S_j are salient scores for p_i and p_j, respectively; $F(p_i, p_j)$ is a pixel position filter using Gaussian kernels and its definition can be found in [4].

These loss functions work together to improve the overall comprehensiveness and quality of the predicted outcome: $\mathcal{L}_s$ uses scribble annotations to guide the propagation of scribble pixels into the foreground regions, while $\mathcal{L}_{pb}$ enforces background similarity, suppressing foreground expansion; $\mathcal{L}_{dsm}$ ensures local smoothness; $\mathcal{L}_{lsc}$ guarantees coherence in saliency scores among neighbouring pixels, and $\mathcal{L}_{bdy}$ enhances boundary details. Section 4.4.4 will give an experimental evaluation to assess the impact and effectiveness of these loss settings.

4.4 Experiment and Discussion

4.4.1 Training and Testing Setting

We use S-DUTS [3] to train and DUTS-TE [17], DUT-OMRON [31], HKU-IS [32], Pascal-S [33], and ECSSD [34] to test the models. We employ SwinV2-Base [25] as the

backbone. The maximum learning rate is 0.0001 for the backbone and 0.001 for the other parts. The momentum is 0.9, and the weight decay is 0.0001. Additionally, we employed the 'poly' learning rate strategy. The input images are scaled to $[384 \times 384]$. The training batch size is 10, and the training times are 99 epochs. Evaluations are conducted on a server with an A100 (40G) GPU and an AMD EPYC 7763 64-Core Processor (1T).

Evaluation Criteria: We use Mean Absolute Error (MAE) [35], maximum F-measure (denoted by F_β^{max}), mean F-measure (mF) [36], mean Enhanced-alignment Measure (mE_m) [37], and the measure of S (S_m) [38] to evaluate the SOD models. Precision-recall (PR) and Fm curves are plotted to demonstrate overall performance. Further details on these metrics can be found in Chapter 2.

4.4.2 Quantitative and Qualitative Comparison

In this section, we perform a comprehensive comparative analysis between our proposed models and the state-of-the-art SOD models. There are three categories: (i) SOD approaches (i.e., VST [39], EBMG [40], SelfReformer [41], ICON-P [42], and ICON-S [42]);(ii) WSOD models (MWS [43], WSSA [3], SCWSSOD [4], MFNet-D169 [7], MFNet [7], MFRN-SRPN [23], NSAL [8], PSOD [20], and CSI-PSI [5]), and (iii) a not strictly WSOD model: HybridSOD [44]. Here, we clarify: The pseudo-labels used by the WBNet and WBNet-K models, generated by unsupervised self-learning models, fall within the weakly supervised category. In contrast, WBNet-PSAM and WBNet-BSAM utilise pseudo-labels generated by the SAM model, which is supervised and pre-trained. HybridSOD uses 10% full pixel labels in the training dataset. Hence, these three are categorised as not strictly weakly supervised methods. We obtain the saliency maps of these models from their authors or deployment codes for fair comparison.

4.4.2.1 Quantitative performance evaluation

Table 4.1 and Table 4.2 show quantitative comparison results for five data sets. The supervision types ('Sup.') are: 'Full' (pixel-wise full labels), 'Cla' (image-level classification labels), 'UPse' (pseudo labels from unsupervised methods), 'Scr' (scribble labels), 'Point' (Point labels), 'Cap' (caption labels), 'CPse' (pseudo labels generated by CAM), 'BUPse' (background pseudo labels from unsupervised methods), and

'BPseSAM' (background pseudo labels from SAM). The symbol '-' means that the method does not provide predicted saliency maps to compute specific metrics; other metric values are from their papers. The three top results for the fully supervised SOD model and the WSOD are coloured red, green, and blue, respectively.

In particular, among the WSOD models, the proposed WBNet and WBNet-K distinguish themselves by presenting significantly superior results to the single-source and multi-source WSOD methods and even close to some supervised models, that is, EBMG [40] and VST [39]. In particular, WBNet-K, employing K-means in the pseudo-label computing module (S-PMG), excels on DUTS-TE and PASCAL datasets. At the same time, WBNet, using spectral clustering in S-PMG, is superior on the DUT-OMRON, HKU-IS, and ECSSD datasets. The observed variation in the results can be attributed to the various distribution features of these datasets. Consequently, the selection of the clustering method in S-PMG can impact the quality of pseudo-labels differently depending on the dataset, thereby influencing the overall accuracy of the algorithm. However, both WBNet and WBNet-K have a considerable advantage over other models. Primarily, in terms of *MAE*, 20.29% and 20.32% improvements in DUT-OMRON; 16.33% and 28.20% improvements in DUTS-TE, 8.61% and 8.07% improvements in HKU-IS and 10.05% and 18.44% improvements in ECSSD.

In the comparison of WBNet-PSAM, WBNet-BSAM, and HybridSOD (which uses 10% full pixel labels in the training dataset), it is apparent that these three methods, although not strictly WSOD, offer unique insights into label utilisation strategies. As the tables indicate, WBNet-PSAM outperforms WBNet-BSAM on DUTS-TE, HKU-IS, and ECSSD. Furthermore, WBNet-PSAM and WBNet-BSAM are not better than WBNet and WBNet-K overall.

Comparisons of the Fm and PR curves are depicted in Fig. 4.8 and Fig. 4.9. WBNet and WBNet-K's F-measure curves outperform others overall on five datasets. In most cases, WBNet is slightly superior to WBNet-K, consistent with the quantitative comparison in Table 4.1 and Table 4.2. Despite PSOD being somewhat better than WBNet and WBNet-K after a certain threshold, WBNet and WBNet-K are stable across all five datasets. Based on these observations, WBNet and WBNet-K have more advantages for WSOD tasks.

Table 4.1: Quantitative comparison on the DUT-OMRON and DUTS-TE datasets.

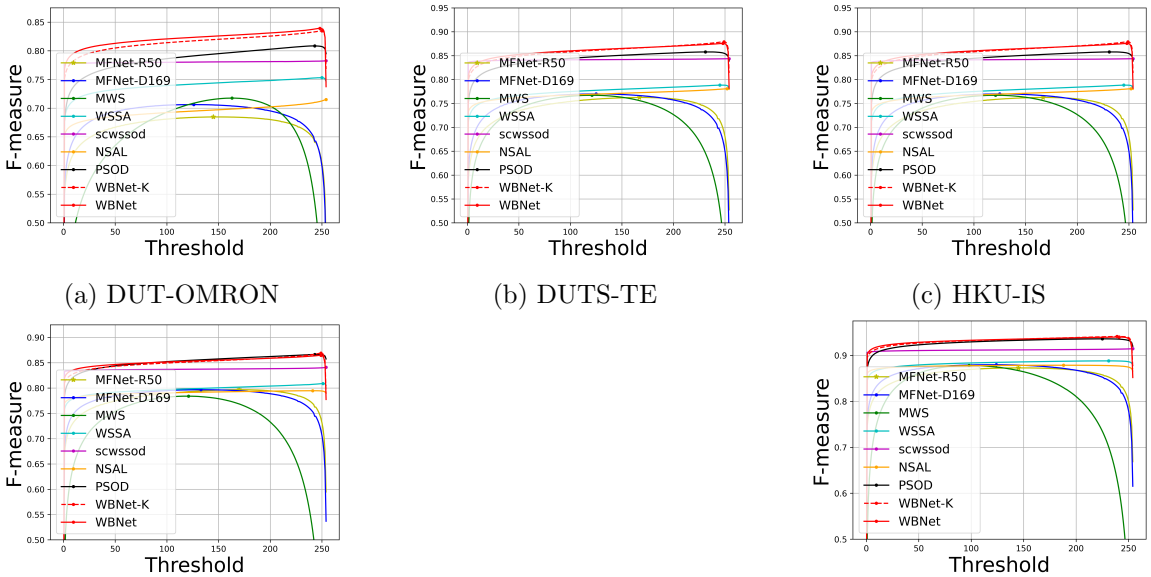
			DUT-OMRON [31] (5168 images)					**DUTS-TE** [17] (5019 images)				
Year	**Model**	**Sup.**	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$mF_\beta \uparrow$	$mE_m \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$mF_\beta \uparrow$	$mE_m \uparrow$	$S_m \uparrow$
	SOD											
2021	VST [39]	Full	0.0582	0.8245	0.7967	0.8718	0.8503	0.0372	0.8898	0.8579	0.9153	0.8963
2021	EBMG [40]	Full	0.0505	0.8386	0.8179	0.8951	0.8584	0.0288	0.9091	0.8863	0.9331	0.9088
2023	ICON-P [42]	Full	0.0468	0.8519	0.8228	0.8951	0.8654	0.0255	**0.9218**	0.8932	0.9386	**0.9173**
2023	ICON-S [42]	Full	**0.0426**	**0.8546**	**0.8350**	**0.9073**	**0.8693**	**0.0242**	0.9196	**0.8998**	**0.9470**	0.9171
2024	SelfReformer [41]	Full	0.0433	0.8367	0.8189	0.8928	0.8608	0.0266	0.9155	0.8921	0.9210	0.9111
	WSOD											
2019	MWS [21]	Cla+Cap	0.1077	0.7175	0.6443	0.7642	0.7559	0.0913	0.7671	0.7108	0.8142	0.7590
2020	WSSA [3]	Scr	0.0684	0.7532	0.7373	0.8448	0.7849	0.0621	0.7883	0.7723	0.8641	0.8037
2021	SCWSSOD [4]	Scr	0.0602	0.7827	0.7779	0.8699	0.8120	0.0488	0.8437	0.8389	0.8967	0.8407
2021	MFNet-D169 [7]	CPse	0.0867	0.7062	0.6845	0.8037	0.7419	0.0761	0.7699	0.7455	0.8373	0.7750
2021	MFNet [7]	CPse	0.0982	0.6847	0.6650	0.7844	0.7259	0.0787	0.7625	0.7375	0.8303	0.7781
2022	MFRN-SRPN [23]	UPse+Cla	0.0880	-	0.6790	-	0.7570	0.0800	-	0.7240	-	0.7670
2022	PSOD [20]	Point	0.0642	0.8086	0.7836	0.8648	0.8245	0.0447	0.8578	0.8404	0.8988	0.8536
2023	NSAL [8]	CPse	0.0884	0.7150	0.6918	0.8025	0.7450	0.0728	0.7808	0.7660	0.8431	0.7817
2024	CSI-PSI [5]	Scr	0.0601	-	0.7800	0.8650	-	0.0500	-	0.8330	0.9000	-
2024	**WBNet-K**	Scr+BUPse	0.0486	0.8347	0.8119	0.8917	0.8523	**0.0359**	**0.8789**	0.8562	**0.9084**	**0.8774**
2024	**WBNet**	Scr+BUPse	**0.0479**	**0.8392**	**0.8187**	**0.8943**	**0.8550**	0.0374	0.8756	**0.8575**	0.9083	0.8764
	Not strictly WSOD											
2022	HybridSOD [22]	10%Full+UPse	-	-	-	-	-	0.0500	0.8030	-	-	0.8370
2024	**WBNet-BSAM**	Scr+BPseSAM	**0.0504**	**0.8374**	**0.8085**	0.8813	**0.8527**	0.0391	0.8710	0.8459	0.8951	**0.8738**
2024	**WBNet-PSAM**	Scr+BPseSAM	0.0524	0.8240	0.8070	**0.8856**	0.8461	**0.0386**	**0.8730**	**0.8589**	**0.9042**	0.8731

Table 4.2: Quantitative comparison on the HKU-IS and PASCAL-S datasets.

			HKU-IS [32] (4447 images)					**PASCAL-S** [33] (850 images)				
Year	**Method**	**Sup.**	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$mF_\beta \uparrow$	$mE_m \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$mF_\beta \uparrow$	$mE_m \uparrow$	$S_m \uparrow$
	SOD											
2021	VST [39]	Full	0.0297	0.9424	0.9129	0.9597	0.9283	0.0620	0.8755	0.8457	0.9024	0.8716
2021	EBMG [40]	Full	0.0229	0.9466	0.9288	0.9673	0.9304	0.0542	0.8866	0.8659	0.9070	0.8765
2023	ICON-P [42]	Full	**0.0216**	**0.9521**	0.9325	0.9698	0.9353	0.0510	0.8927	0.8690	0.9145	0.8819
2023	ICON-S [42]	Full	**0.0216**	0.9512	**0.9331**	**0.9717**	**0.9355**	**0.0484**	**0.8961**	**0.8767**	**0.9237**	**0.8849**
2024	SelfReformer [41]	Full	0.0241	0.9474	0.9265	0.9606	0.9310	0.0510	0.8943	0.8736	0.8825	0.8809
	WSOD											
2019	MWS [21]	Cla+Cap	0.0858	0.8560	0.7750	0.8957	0.8183	0.1342	0.7839	0.7136	0.7911	0.7675
2020	WSSA [3]	Scr	0.0470	0.8806	0.8708	0.9322	0.8651	0.0924	0.8088	0.7954	0.8568	0.7975
2021	SCWSSOD [4]	Scr	0.0375	0.9086	0.9031	0.9428	0.8823	0.0775	0.8411	0.8350	0.8806	0.8200
2021	MFNet-D169 [7]	CPse	0.0585	0.8767	0.8533	0.9222	0.8466	0.1149	0.7967	0.7785	0.8206	0.7695
2021	MFNet-R50 [7]	CPse	0.0582	0.8747	0.8504	0.9187	0.8525	0.1118	0.7968	0.7770	0.8236	0.7817
2022	MFRN-SRPN [23]	UPse+Cla	0.0560	-	0.8470	-	0.8480	0.1090	-	0.7730	-	0.7790
2022	PSOD [20]	Point	0.0322	0.9235	0.9134	0.9581	0.9022	0.0647	0.8663	0.8499	**0.8957**	**0.8529**
2023	NSAL [8]	CPse	0.0511	0.8825	0.8759	0.9231	0.8540	0.1103	0.7947	0.7885	0.8260	0.7671
2024	CSI-PSI [5]	Scr	0.0360	-	0.9080	0.9440	-	0.1200	-	**0.8650**	0.8300	-
2024	**WBNet-K**	Scr+BUPse	0.0296	0.9309	0.9177	0.9571	0.9125	**0.0638**	**0.8691**	0.8483	0.8714	0.8513
2024	**WBNet**	Scr+BUPse	**0.0291**	**0.9310**	**0.9203**	**0.9585**	**0.9137**	0.0658	0.8646	0.8499	0.8723	0.8508
	Not strictly WSOD											
2022	Hybrid-SOD [22]	10%Full+UPse	0.0380	0.8920	-	-	0.8870	0.0760	0.8270	-	-	0.8280
2024	**WBNet-BSAM**	Scr+BPseSAM	0.0295	0.9299	0.9157	0.9546	**0.9153**	0.0671	0.8614	0.8419	0.8572	**0.8495**
2024	**WBNet-PSAM**	Scr+BPseSAM	**0.0291**	**0.9316**	**0.9220**	**0.9582**	0.9129	**0.0665**	**0.8624**	**0.8467**	**0.8685**	0.8481

Table 4.3: Quantitative comparison on the ECSSD dataset.

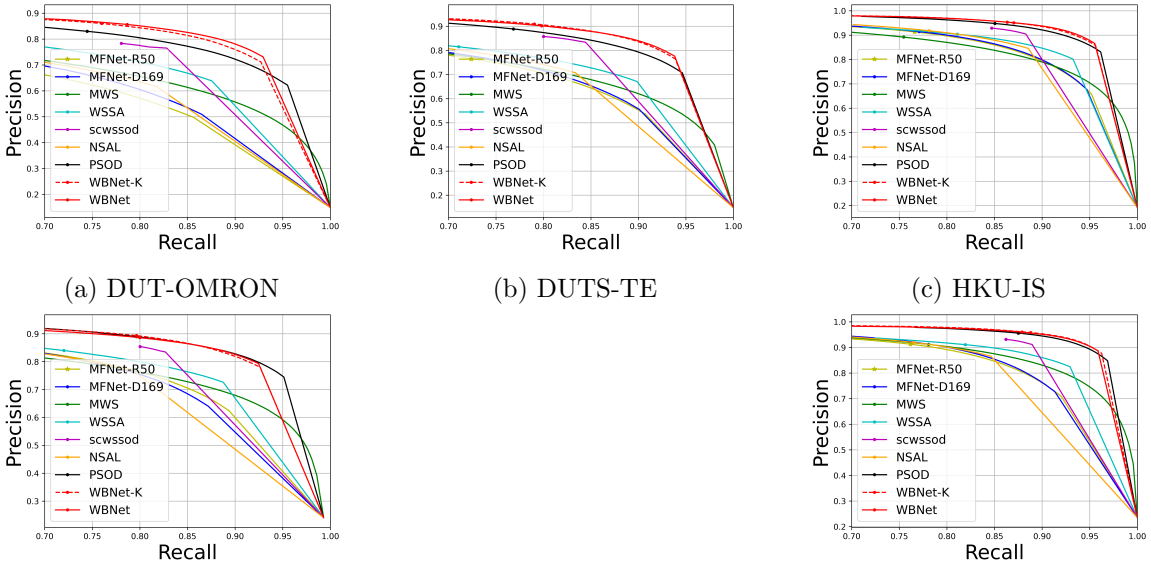
			ECSSD [34] (1000 images)				
Year	**Method**	**Sup.**	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$mF_\beta \uparrow$	$mE_m \uparrow$	$S_m \uparrow$
	SOD						
2021	VST [39]	Full	0.0337	0.9507	0.9258	0.9571	0.9323
2021	EBMG [40]	Full	**0.0232**	0.9591	0.9452	0.9632	**0.9416**
2023	ICON-P [42]	Full	0.0240	0.9594	0.9432	0.9624	0.9401
2023	ICON-S [42]	Full	0.0235	**0.9608**	**0.9458**	**0.9669**	0.9414
2024	SelfReformer [41]	Full	0.0273	0.9577	0.9414	0.9361	0.9356
	WSOD						
2019	MWS [21]	Cla+Cap	0.0985	0.8779	0.8049	0.8849	0.8278
2020	WSSA [3]	Scr	0.0590	0.8880	0.8803	0.9172	0.8656
2021	SCWSSOD [4]	Scr	0.0489	0.9145	0.9091	0.9313	0.8820
2021	MFNet-D169 [7]	CPse	0.0843	0.8796	0.8600	0.8890	0.8347
2021	MFNet-R50 [7]	CPse	0.0841	0.8727	0.8542	0.8894	0.8368
2022	MFRN-SRPN [23]	UPse+Cla	0.6600	-	0.8720	-	0.0858
2022	PSOD [20]	Point	0.0358	0.9359	0.9255	0.9526	0.9137
2023	NSAL [8]	CPse	0.0777	0.8785	0.8742	0.8893	0.8338
2024	CSI-PSI [5]	Scr	0.0480	-	0.9140	0.9320	-
2024	**WBNet-K**	Scr+BUPse	**0.0296**	0.9309	0.9177	**0.9571**	0.9125
2024	**WBNet**	Scr+BUPse	0.0322	**0.9398**	**0.9298**	0.9377	**0.9189**
	Not strictly WSOD						
2022	Hybrid-SOD [22]	10%Full+UPse	0.0510	0.8990	-	-	0.8860
2024	**WBNet-BSAM**	Scr+BPseSAM	0.0318	0.9398	0.9270	0.9355	0.9211
2024	**WBNet-PSAM**	Scr+BPseSAM	**0.0309**	**0.9412**	**0.9323**	**0.9356**	**0.9203**



(a) DUT-OMRON (b) DUTS-TE (c) HKU-IS

(d) PASCAL-S (e) ECSSD

Figure 4.8: Fm-curves comparison.



(a) DUT-OMRON (b) DUTS-TE (c) HKU-IS

(d) PASCAL-S (e) ECSSD

Figure 4.9: PR-curves comparison.

4.4.2.2 Qualitative performance comparison

We selected 13 challenging scenes from the test datasets for comparison of saliency detection, as shown in Fig. 4.10. With high integrity and more precise boundaries, the proposed WBNet and WBNet-K achieve the most accurate overall detection results. For example, WBNet demonstrates precise localisation of salient objects, avoiding missed detection in some models. In scenarios involving occlusion, WBNet can estimate the obscured portions. In scenes featuring camouflaged objects, WBNet identifies crucial entities, as seen in *Rows* 6 and 7. The boundary delineation is also notably accurate, as evident in the depiction of the monkey's fur in *Row* 5. WBNet performs exceptionally well at predicting geometric objects, particularly those with sharp and elongated features, as shown in *Rows* 12 and 13.

4.4.3 Limitation

From Tables 4.1- 4.3, it is evident that both WBNet and WBNet-K, as weakly supervised models, exhibit specific gaps in precision compared to supervised SOD, particularly in terms of *MAE* scores. This difference is also noticeable in Fig. 4.10 that the results of WBNet deviate from the ground-truth images, especially in cases involving geometrically complex objects in *Rows* 2, 3, 12, 13.

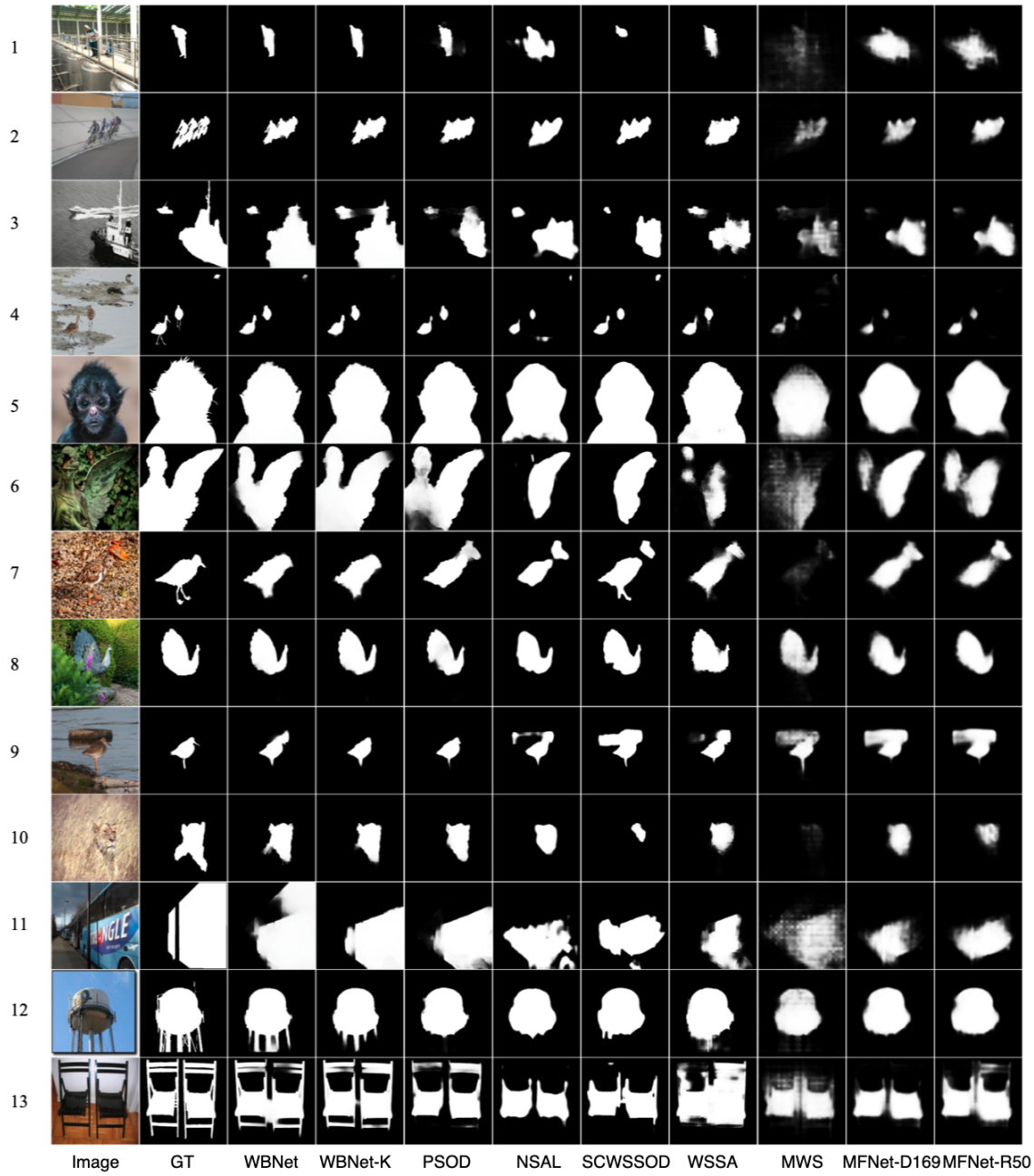


Figure 4.10: Qualitative performance comparison.

Moreover, there are some notable failure cases, as shown in Fig. 4.11. In these instances, all WSOD models, including our WBNet and WBNet-K, exhibit the common challenge of misclassifying specific nonsalient foreground objects as salient. This misclassification arises due to the inherent challenges posed by complex scenes characterised by ambiguity and subjective bias in annotations [1]. It is imperative to emphasize that these challenging instances constitute only a tiny fraction of the data. However, despite being a minority, they limit the model's ability to handle complex scenes, which indicates areas for improvement in future WSOD models.

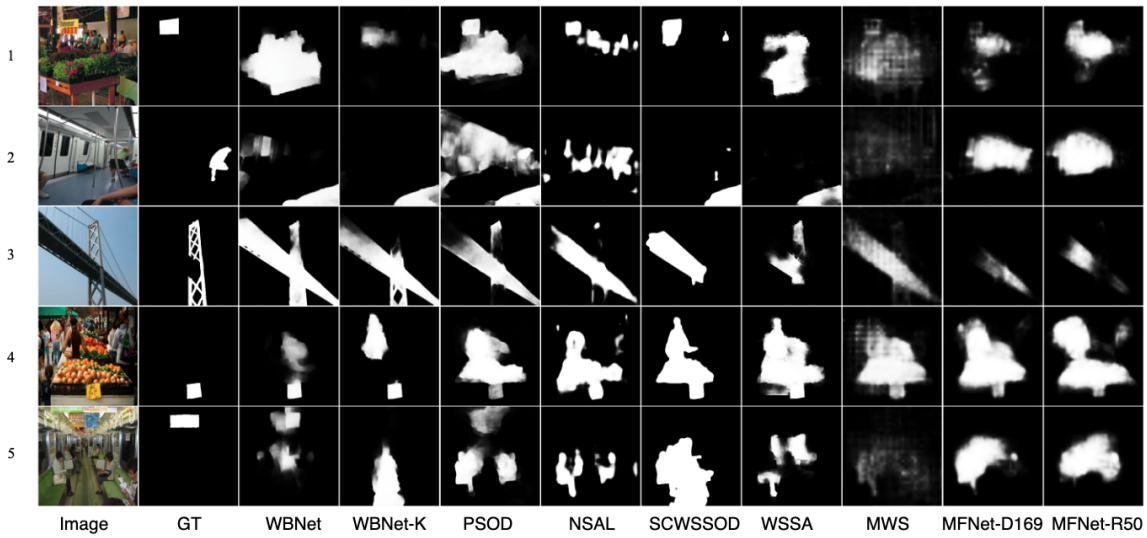


Figure 4.11: Examples of some failure cases.

4.4.4 Ablation Study

A comprehensive demonstration of the effectiveness of different combinations of loss of WBNet is studied.

4.4.4.1 S-PMG configuration

We examined the impact of using different combinations of self-learning models in S-PMG, with the results presented in Table 4.4 and Table 4.5. The configuration composed of three self-learning models (in *Row* 5) serves as the baseline for WBNet. The baseline utilises the 27 candidate masks from three self-learning models as input for the Filtering and Selection Module (FSM) of the S-PMG. However, in *Rows* 1-4, only 9 candidate masks from a single self-learning model are used as input for the FSM.

Interestingly, when using a single self-learning model, DINO [11] achieved superior results, followed by SwAV [12]. SwAV [12] exhibits superior performance in HKU-IS and ECSSD. When combining DINO, MoCoV2 [13], and SwAV in S-PMG, the results surpass those attained by employing any single self-learning model, confirming that pseudo-equivalents of multiple sources can improve performance. It should be noted that, despite DINO-V2 being the advanced version of DINO, it does not outperform DINO on these five **SOD** datasets, either used in isolation or mixture with other self-learning models for WBNet.

Figure 4.12 illustrates a qualitative comparison of pseudo-masks generated by using DION, MoCoV2, SwAV, and the combination of them from the S-PMG module (corresponding to the configurations *No.* 1, 3, 4, and 5 in Table 4.4 and Table 4.5, respectively). It is apparent that the masks created using each of the three self-learning methods have their own set of advantages and disadvantages in various situations. The masks' effectiveness is compromised when significant errors occur, such as in *Rows* 4 and 8 for S-PMG DINO, *Row* 7 for S-PMG MoCoV2, and *Row* 9 for S-PMG SwAV in Fig. 4.12. Therefore, we feed all candidate masks generated by the three self-learning models into the Filtering Module to identify the optimal one that adheres most closely to the saliency rules. This approach enables us to leverage the strengths of each self-learning method while mitigating their weaknesses, resulting in more stable pseudo-masks overall.

4.4.4.2 Pseudo-full label or pseudo-background label

We conducted a comparison between using pseudo-full labels (including both foreground and background masks) and using only pseudo-background labels in WBNet, as presented in *Rows* 5-8 in Table 4.4 and Table 4.5.

Utilising pseudo-background labels leads to significantly improved performance compared to full pseudo labels. This result can be credited to the characteristics of scribble labels, which tend to be more accurate at representing the foreground but may lack comprehensive foreground features. When combined with inaccurate pseudo-foreground labels, there is a risk of introducing erroneous foreground information into the pseudo-labels, potentially leading to imprecise feature learning. In contrast, pseudo-background labels offer broader coverage of background features, compensating for any inaccuracies in the scribble background label. Importantly, they introduce

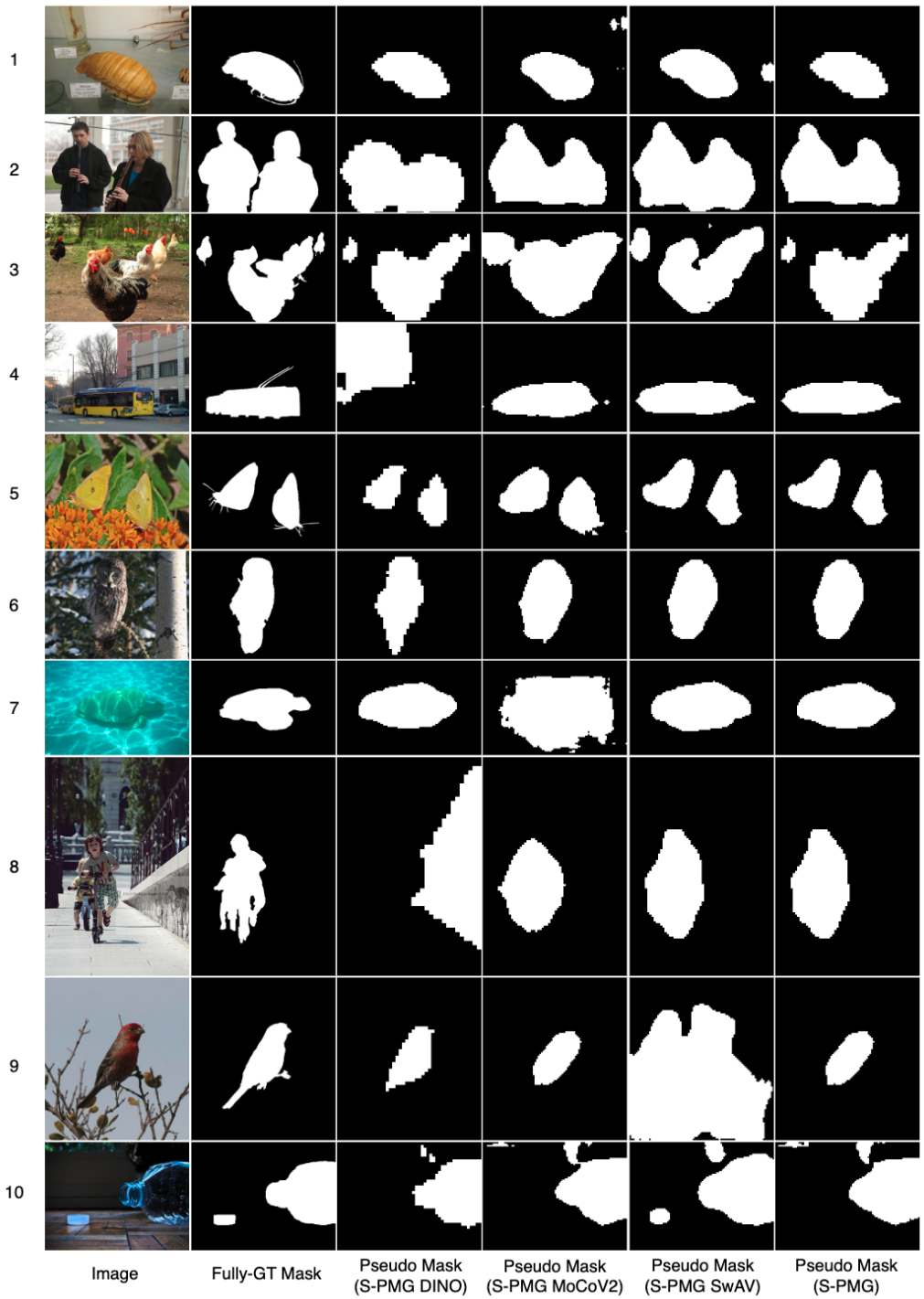


Figure 4.12: Examples of saliency pseudo masks generated by difference configuration of S-PMS module.

Table 4.4: Configuration comparison of the S-PMG module on the DUT-OMRON and DUTS-TE datasets.

			DUT-OMRON [31]			DUTS-TE [17]		
No.	**Label**	**Configuration**	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$
1	Bg	DINO [11]	0.0483	0.8350	0.8526	0.0375	0.8737	0.8750
2	Bg	DINO-V2 [45]	0.0625	0.8204	0.8367	0.0441	0.8642	0.8665
3	Bg	MoCoV2 [13]	0.0548	0.8140	0.8375	0.0428	0.8549	0.8617
4	Bg	SwAV [12]	0.0568	0.8226	0.8389	0.0429	0.8656	0.8654
5	Bg	**DINO,MoCoV2,SwAV**	**0.0479**	**0.8392**	**0.8550**	**0.0374**	**0.8756**	**0.8764**
6	Bg	DINO-V2,MoCoV2,SwAV	0.0545	0.8326	0.8481	0.0393	0.8740	0.8735
7	Full	DINO,MoCoV2,SwAV	0.0689	0.8292	0.8354	0.0508	0.8756	0.8626
8	Full	DINO-V2,MoCoV2,SwAV	0.0650	0.8364	0.8381	0.0493	0.8723	0.8616

Table 4.5: Configuration comparison of the S-PMG module on the HKU-IS, PASCAL-S, and ECSSD datasets.

			HKU-IS [32]			PASCAL-S [33]			ECSSD [34]		
No.	**Label**	**Configuration**	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$
1	Bg	DINO [11]	0.0310	0.9247	0.9103	0.0759	0.8524	0.8388	0.0353	0.9354	0.9149
2	Bg	DINO-V2 [45]	0.0319	0.9247	0.9093	0.0721	0.8585	0.8442	0.0345	0.9344	0.9168
3	Bg	MoCoV2 [13]	0.0355	0.9118	0.8994	0.0689	0.8549	0.8460	0.0400	0.9246	0.9057
4	Bg	SwAV [12]	0.0307	0.9274	9089	0.0689	0.8588	0.8441	0.0343	0.9366	0.9159
5	Bg	**DINO,MoCoV2,SwAV**	0.0291	**0.9310**	0.9137	**0.0658**	0.8646	**0.8508**	0.0322	0.9398	0.9189
6	Bg	DINO-V2,MoCoV2,SwAV	**0.0287**	0.9308	**0.9142**	0.0683	0.8606	0.8474	**0.0310**	**0.9403**	**0.9206**
7	Full	DINO,MoCoV2,SwAV	0.0361	0.9295	0.9101	0.0754	0.8637	0.8434	0.0372	0.9363	0.9189
8	Full	DINO-V2,MoCoV2,SwAV	0.0365	0.9269	0.9071	0.0739	**0.8652**	0.8432	0.0373	0.9364	0.9167

less interference with foreground information, making them a more suitable complement to scribble labels.

4.4.4.3 Network configuration

This experiment aims to elucidate the effects of three key modules of WBNet, including the Feature Aggregation Module (FAM), Transformer Decoder (TFD), and Edge Prediction Module (EPM). The full-module configuration, representing WBNet's default setting, serves as a baseline for comparison in Table 4.6 and Table 4.7. According to the tables, combining all three modules yields the most significant performance enhancement, highlighting their synergistic effects on boosting WBNet's overall performance. Using FAM alone also demonstrates comparable performance in HKU-IS, PASCAL, and ECSSD datasets. TFD significantly improves performance; although using it alone is not superior to using FAM independently, combining FAM and TFD significantly improves performance compared to utilising either of these two modules individually. Similarly, EPM proves to be effective, showing substantial improvements in various evaluation metrics when combined with FAM or TFD. Even the combination of FAM and EPM achieves the third-best results across all configurations.

Table 4.6: Configuration comparison of the WBNet on the DUT-OMRON and DUTS-TE datasets.

				DUT-OMRON [31]			**DUTS-TE** [17]		
No.	**FAM**	**TFD**	**EPM**	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$
1	√	–	–	0.0572	0.8291	0.8452	0.0415	0.8670	0.8685
2	–	√	–	0.0616	0.7904	0.8252	0.0496	0.8239	0.8397
3	√	–	√	0.0539	0.8316	0.8483	0.0405	0.8696	0.8706
4	√	√	–	0.0529	0.8340	0.8487	0.0385	0.8730	0.8746
5	–	√	√	0.0578	0.7975	0.8312	0.0478	0.8291	0.8425
6	√	√	√	**0.0479**	**0.8392**	**0.8550**	**0.0374**	**0.8756**	**0.8764**

Table 4.7: Configuration comparison of the WBNet on the HKU-IS, PASCAL-S, and ECSSD datasets.

				HKU-IS [32]			**PASCAL-S** [33]			**ECSSD** [34]		
No.	**FAM**	**TFD**	**EPM**	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$
1	√	–	–	0.0299	0.9269	0.9109	0.0675	0.8580	0.8459	**0.0321**	0.9368	**0.9191**
2	–	√	–	0.0397	0.8959	0.8830	0.0770	0.8347	0.8257	0.0416	0.9137	0.8951
3	√	–	√	0.0294	0.9278	0.9124	0.0699	0.8555	0.8453	0.0322	0.9373	0.9180
4	√	√	–	**0.0291**	0.9307	0.9134	0.0663	**0.8651**	0.8498	**0.0321**	0.9391	0.9185
5	–	√	√	0.0399	0.8964	0.8824	0.0786	0.8366	0.8239	0.0426	0.9148	0.8928
6	√	√	√	**0.0291**	**0.9310**	**0.9137**	**0.0658**	0.8646	**0.8508**	0.0322	**0.9398**	0.9189

Table 4.8: Loss comparison on the DUT-OMRON and DUTS-TE datasets.

						DUT-OMRON [31]			**DUTS-TE** [17]		
No.	L_s	L_{pb}	L_{bdy}	L_{dsm}	L_{lsc}	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$
1	√	-	√	√	√	0.0532	0.8277	0.8451	0.0389	0.8682	0.8700
2	√	√	-	√	√	0.0529	0.8340	0.8487	0.0385	0.8730	0.8746
3	√	√	√	-	√	0.0515	0.8370	0.8521	0.0376	**0.8773**	**0.8786**
4	√	√	√	√	-	0.0539	0.8243	0.8356	0.0408	0.8612	0.8608
5	√	√	√	√	√	**0.0479**	**0.8392**	**0.8550**	**0.0374**	0.8756	0.8764

Table 4.9: Loss comparison on the HKU-IS, PASCAL-S, and ECSSD datasets.

						HKU-IS [32]			**PASCAL-S** [33]			**ECSSD** [34]		
No.	L_s	L_{pb}	L_{bdy}	L_{dsm}	L_{lsc}	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$S_m \uparrow$
1	√	-	√	√	√	**0.0287**	0.9272	0.9111	0.0690	0.8558	0.8430	0.0308	0.9365	0.9185
2	√	√	-	√	√	0.0291	0.9307	0.9134	0.0663	0.8651	0.8498	0.0321	0.9391	0.9185
3	√	√	√	-	√	0.0288	0.9306	0.9156	0.0675	0.8598	0.8498	**0.0306**	**0.9422**	**0.9226**
4	√	√	√	√	-	0.0357	0.9283	0.9093	**0.0639**	**0.9144**	**0.8995**	0.0357	0.9283	0.9093
5	√	√	√	√	√	0.0291	**0.9310**	**0.9203**	0.0658	0.8646	0.8508	0.0322	0.9398	0.9189

4.4.4.4 Loss configuration

This experiment aims to assess the influence of individual loss. We established four different comparison scenarios by excluding one type of loss from each setting. The full-loss configuration, WBNet's default setting, serves as the baseline for comparison. This comparative analysis is presented in Table 4.8 and Table 4.9 (the best two results are, respectively, represented by red and green). We observe that excluding specific loss components has different impacts across datasets. For example, L_{pb} helps reduce MAE on DUT-OMRON, DUTS-TE, and ECSSD, L_{bdy} helps reduce MAE on DUT-OMRON but has less effect on the others, L_{lsc} helps reduce MAE on DUT-OMRON and HKU-IS significantly but hampers MAE on PASCAL-S, and L_{dsm} contributes to S_m more than other metrics on five datasets. However, full-loss configurations are versatile enough to address varied databases effectively despite these variations.

4.5 Conclusion

WSOD methods aim to extract more salient information from limited annotations, often employing pseudo labels generated by unsupervised self-learning techniques. However, the accuracy and consistency of these pseudo labels can hinder detection performance.

To address this challenge, we explore the generation and utilisation of multi-source pseudo-labels in Weakly Supervised Object Detection (WSOD). We introduce an innovative multi-source weakly supervised SOD framework that leverages pseudo-background (non-salient) labels alongside scribble labels to enhance salient feature extraction accuracy. We first develop a comprehensive salient pseudo-mask generator, utilising information from various self-learning features. We also pioneered the generation of salient pseudo labels through a point-prompted or box-prompted Segment-Anything Model (SAM), which, while not strictly conforming to conventional WSOD paradigms, marks a promising step in this direction.

Furthermore, we develop a Transformer-based WSOD network (WBNet) based on scribble labels and pseudo-background labels. WBNet incorporates a pixel decoder, a transformer decoder, and an auxiliary edge prediction module with a multi-source hybrid loss function. Comprehensive evaluations, including comparisons with state-

of-the-art weakly supervised SOD methods on five widely recognised datasets, demonstrated that WBNet achieved a substantial improvement in performance.

References

[1] 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 Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(6):3239–325, 2021.

[2] Huajun Zhou, Yang Lin, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Benchmarking deep models on salient object detection. *Pattern Recognition*, 145:109951, 2024.

[3] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12546–12555, 2020.

[4] Siyue Yu, Bingfeng Zhang, Jimin Xiao, and Eng Gee Lim. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3234–3242, 2021.

[5] Shuo Zhao, Peng Cui, Jing Shen, and Haibo Liu. Local saliency consistency-based label inference for weakly supervised salient object detection using scribble annotations. *CAAI Transactions on Intelligence Technology*, 9(1):239–249, 2024.

[6] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[7] Yongri Piao, Jian Wang, Miao Zhang, and Huchuan Lu. Mfnet: Multi-filter directive network for weakly supervised salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4136–4145, 2021.

[8] Yongri Piao, Wei Wu, Miao Zhang, Yongyao Jiang, and Huchuan Lu. Noise-sensitive adversarial learning for weakly supervised salient object detection. *IIEEE Transactions on Multimedia (TMM)*, 25:2888–2897, 2023.

[9] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from

image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4253–4262, 2020.

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

[12] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems (ANIPS)*, 33:9912–9924, 2020.

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[14] Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3971–3980, 2022.

[15] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(8):888–905, 2000.

[16] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems (ANIPS)*, volume 34, pages 17864–17875. Curran Associates, Inc., 2021.

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3796–3805, Honolulu, HI, USA, 2017.

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, Boston, MA, USA, 2015.

[19] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):834–848, 2017.

[20] Shuyong Gao, Wei Zhang, Yan Wang, Qianyu Guo, Chenglong Zhang, Yangji He, and Wenqiang Zhang. Weakly-supervised salient object detection using point supervison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 670–678, 2022.

[21] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, and Yizhou Yu. Multi-source weak supervision for saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6074–6083, 2019.

[22] Runmin Cong, Qi Qin, Chen Zhang, Qiuping Jiang, Shiqi Wang, Yao Zhao, and Sam Kwong. A weakly supervised learning framework for salient object detection via hybrid labels. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 33(2):534–548, 2022.

[23] Xuan Li, Yuhang Xu, Lei Ma, Zhi Yang, Zhenghua Huang, Hanyu Hong, and Jinwen Tian. Multi-source weakly supervised salient object detection via boosting weak-annotation source and constraining object structure. *Digital Signal Processing*, 126:103461, 2022.

[24] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *Proceedings of the International Symposium on Visual Computing (ISVC)*, pages 234–244. Springer, 2016.

[25] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022.

[26] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WCACV)*, pages 3560–3569, 2021.

[27] Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10031–10040, 2023.

[28] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu.

Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[29] Pieter T De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134:19–67, 2005.

[30] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1818–1827, 2018.

[31] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3166–3173, Portland, OR, USA, 2013.

[32] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5455–5463, Boston, MA, USA, 2015.

[33] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 280–287, Columbus, OH, USA, 2014.

[34] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1162, Portland, OR, USA, 2013.

[35] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–740, Providence, RI, USA, 2012.

[36] Ran Margolin, Lihi Zelnik Manor, and Ayellet Tal. How to evaluate foreground maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, Columbus, OH, USA, 2014.

[37] Dengping Fan, Cheng Gong, Yang Cao, Bo Ren, Mingming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, page 698–704, Stockholm, Sweden, 2018.

[38] Mingming Cheng and Dengping Fan. Structure-measure: A new way to evaluate foreground maps. *International Journal of Computer Vision (IJCV)*, 129(9):2622–2638, 2021.

[39] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4722–4732, 2021.

[40] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. Learning generative vision transformer with energy-based latent space for saliency prediction. *Advances in Neural Information Processing Systems (ANIPS)*, 34:15448–15463, 2021.

[41] Yike Yun and Weisi Lin. Towards a complete and detail-preserved salient object detection. *IIEEE Transactions on Multimedia (TMM)*, 26:4667–4680, 2024.

[42] Mingchen Zhuge, Dengping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(3):3738–3772, 2023.

[43] Hongshuang Zhang, Yu Zeng, Huchuan Lu, Lihe Zhang, Jianhua Li, and Jinqing Qi. Learning to detect salient object with multi-source weak supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(7):3577–3589, 2021.

[44] Runmin Cong, Qi Qin, Chen Zhang, Qiuping Jiang, Shiqi Wang, Yao Zhao, and Sam Kwong. A weakly supervised learning framework for salient object detection via hybrid labels. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 33(2):534–548, 2022.

[45] Yao Wei and Shunping Ji. Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022.

Chapter 5

Summary

This chapter provides the summary for this thesis, which has made contributions to various aspects of salient object detection. Section 5.1 reviews the new methods proposed and Section 5.2 discusses potential future research directions, and Section 5.3 presents potential applications and social impacts of SOD techniques.

5.1 Research Summary

In this thesis, we have conducted an in-depth study on improving salient object detection (SOD) in complex scenes from multiple perspectives. Our investigations include methods like multi-enhancement and frequency decomposition to enhance salient feature learning in fully-supervised SOD, as well as employing multi-source pseudo annotations to enhance weakly-supervised SOD (WSOD). The following is a summary of the methods and contributions we have made.

Chapter 2 proposes a multi-enhancement network (MENet) to harness the full potential of Human Visual System (HVS) mechanisms to enhance saliency feature learning in complex scenes. The content involves constructing a framework that effectively integrates holistic and continuous observation, boundary/contour and structural information sensitivity, and the visual-spatial frequency model to learn accurate and robust salient features.

Chapter 3 proposes TFGNet, an effective frequency-guided network for SOD based

on the Transformer. TFGNet has a parallel two-branch decoder, which refines high-frequency boundary details and low-frequency inner regions of salient objects gradually under the guidance of the decomposed frequency supervisions. TFGNet solves the problem of directly predicting the entire salience map for complex scenes. This framework also rekindles awareness of the advantages of exploiting images' multi-scale spatial frequency features in saliency detection.

Chapter 4 introduces an innovative multi-source weakly supervised SOD (WSOD) framework that leverages pseudo-background (non-salient) labels alongside scribble labels to enhance the extraction accuracy of salient features. A comprehensive salient pseudo-mask generator is first developed, utilising information from diverse self-learning features. Furthermore, a Transformer-based WSOD network (WBNet) is proposed based on scribble and pseudo-background labels. Also, this chapter pioneers the generation of salient pseudo-labels through a point-prompted or box-prompted Segment-Anything Model (SAM), which, while not strictly conforming to conventional WSOD paradigms, marks a promising step in this direction.

In conclusion, this thesis introduces three novel SOD methods. These approaches have demonstrated a notable increase in accuracy for salient and weakly supervised object detection. Contributions have been recognised through publications at prestigious conferences and journals, such as CVPR2023 and Pattern Recognition.

5.2 Future Research Direction

We suggest two specific directions for potential research.

5.2.1 Cross-Domain Multi-Task Learning based SOD

Given that salient segmentation, semantic segmentation, and instance segmentation are all related tasks. It is wise to use a Multi-Task Learning (MTL) framework [1, 2] to accomplish these segmentation tasks simultaneously, reducing data and computational requirements during training and inference. However, current multitask training sets are limited, which presents challenges in designing and training multitask learning models. Specifically, the following aspects need to be addressed: how to effectively integrate datasets from different domains and annotation types to improve feature learning and model generalisation; how to handle label differences between

different tasks to ensure that the model can simultaneously adapt to the requirements of multiple tasks; and how to address domain differences and distribution imbalances between datasets to avoid performance degradation of the model during testing.

Considering these issues, it is hoped that an effective method can be found to use multiple source datasets for feature learning, thereby constructing a more comprehensive and robust complex scene perception framework with which SOD can be improved.

5.2.2 Personalized and/or Task Orientated SOD/WSOD

Another exciting direction is to make saliency detection more personalised, tailored to individual differences in how people perceive and process visual information. Human perception is influenced by a wide range of factors, including age, gender, cultural background, education, socioeconomic status, and even health conditions. These elements can significantly alter the way we focus on and interpret scenes. Incorporating such personal factors into SOD models would greatly enhance their applicability across various contexts. The relatively promising research avenue lies in task-orientated saliency detection. The relevance of visual information changes depending on the context and the task at hand. SOD models that can adapt to different tasks or environments, dynamically prioritising what is important while ignoring irrelevant details, would be a significant advancement.

These tasks require models to go beyond simply detecting salient objects; they need to understand the context and intent behind what is important to different users or tasks. Integrating more semantic information into existing datasets is crucial to advance personalised and task-orientated SOD models.

5.3 Potential Application and Societal Impact

Saliency Object Detection (SOD) offers transformative benefits in a wide range of applications by prioritising and highlighting the most relevant visual information. Its impact spans various fields, including autonomous driving, robotics, medical imaging, augmented reality, video surveillance, and more.

In autonomous vehicles, SOD is essential to effectively detect and prioritise key ob-

jects instantaneously. It is not just about detecting pedestrians or obstacles, but also about understanding road signs, lane markings, and other critical visual cues in complex and dynamic environments. Autonomous vehicles can make better decisions, improving both safety and navigation. In robotics, especially for tasks that involve object manipulation or navigation, saliency detection helps the robot focus on relevant objects and actions, improving task efficiency and safety. In medical imaging, saliency detection is essential for accurately identifying and delineating tumours or other abnormalities in various organs. Precise boundary detection is particularly critical during surgeries, where accurate tumour localisation is necessary to remove it while preserving healthy tissue. In this context, saliency detection ensures that critical features, such as tumour edges, are clearly identified, supporting more effective diagnosis and treatment planning. SOD can improve user experiences in AR and VR by dynamically highlighting important objects or features in a scene, making interactions more intuitive and immersive. For individuals with visual impairments, SOD can enhance the usability of assistive devices by highlighting important objects and obstacles in their environment, aiding navigation and interaction.

In conclusion, saliency detection is a powerful tool that can drive innovation and effectiveness in various practical applications, underscoring its importance in the advancement of technology and improving daily life.

References

[1] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.

[2] Apoorv Khattar, Srinidhi Hegde, and Ramya Hebbalaguppe. Cross-domain multi-task learning for object detection and saliency estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3639–3648, 2021.