Deep Learning-based Segmentation for Complex Scene Understanding

by

Wei Ji

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Engineering

Department of Electrical and Computer Engineering

University of Alberta

Abstract

Deep learning-based segmentation plays a crucial role in computer and robot vision. Traditional approaches have predominantly relied on RGB (*i.e.*, color) imagery, given its widespread availability and usage. However, the innate issues with color imagery, such as cluttered backgrounds and poor lighting, have significantly influenced the performance of existing segmentation methods under complex visual scenes. This thesis is an attempt seeking to advance the capabilities of deep learning-based segmentation for complex scene understanding by investigating additional imaging modalities.

This thesis navigates two innovative avenues: 1) incorporating depth data to comprehensively understand the 3D spatial layout of scenes, and 2) using thermal infrared imagery to enhance vision under adverse lighting conditions. In the first avenue, we concentrate on RGB-depth segmentation and propose three novel strategies to improve segmentation efficacy by optimizing three key aspects of deep learning models, namely, network input, network architecture, and network supervision. These strategies involve calibrating the inherent bias in depth inputs for better scene layout depiction, developing advanced network architectures for improved multimodal information fusion and contextual comprehension, and harnessing depth map geometry for facilitating unsupervised RGB-D segmentation, thus reducing reliance on extensive human annotations. In the second avenue, we delve into RGB-Thermal (multispectral) segmentation, which is a relatively less-explored territory. We introduce the SemanticRT dataset, an extensive and large-scale resource for segmenting images

under varied illumination conditions, and an innovative explicit complement modeling (ECM) framework to enhance modality-specific cue utilization and cross-modal feature fusion. Additionally, we pioneer the RGB-Thermal segmentation in the video domain, by presenting the first multispectral video semantic segmentation benchmark dataset - MVSeg, and developing an efficient MVNet baseline framework to jointly learn semantic representations from multispectral and temporal contexts.

Extensive evaluations across ten segmentation datasets demonstrate that our proposed methodologies significantly outperform existing state-of-the-art solutions in handling challenging scenarios, heralding advancements in deep learning-based segmentation. This thesis also discusses the benefits and limitations of a recent foundational model - Segment Anything Model, and outlines some compelling issues and future research avenues within the field. Importantly, we advocate for open access, making our source codes, models and datasets publicly available to foster reproducibility and encourage collaborative research efforts.

Preface

This thesis is an original work by Wei Ji.

Part I contains work on RGB-Depth salient object detection from:

- **Chapter 3**: [1] **W. Ji**, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu, L. Cheng. "Calibrated RGB-D salient object detection". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

  I am responsible for proposing the method, designing and implementing the code, conducting the experiments, and writing the draft. J. Li, S. Yu, M. Zhang Y. Piao and H. Lu provide constructive feedback for this work, help me prepare the draft. J. Li, S. Yao, and Q. Bi help me implement the method and pre-process the experimental data. K. Ma and Y. Zheng, provide sufficient GPU resources to support this project, and give several suggestions to improve this work. L. Cheng provides lots of valuable feedback, and revises the draft.

- **Chapter 4**: [2] **W. Ji**, G. Yan, J. Li, Y. Piao, S. Yao, M. Zhang, L. Cheng, H. Lu. "DMRA: Depth-induced multi-scale recurrent attention network for RGB-D saliency detection". IEEE Transactions on Image Processing (IEEE TIP), 2022.

  I am responsible for proposing the method, designing and implementing the code, conducting the experiments, and writing the draft. G. Yan and Y. Piao help me propose the method's technical design. The experimental results and analysis are carried out with the assistance of G. Yan, J. Li, S. Yao. M. Zhang and H. Lu provide valuable insights and help me prepare the draft. L. Cheng provides valuable comments for this work and also revises the paper.

- **Chapter 5**: [3] **W. Ji**, J. Li, Q. Bi, C. Guo, J. Liu, L. Cheng. "Promoting saliency from depth: Deep unsupervised RGB-D saliency detection". International Conference on Learning Representations (ICLR), 2022.

  I am responsible for proposing the method, designing and implementing the code, conducting the experiments, and writing the draft, with the assistance of J. Li. Q. Bi implements some state-of-the-art methods and carries out some experimental analyses. C. Guo and J. Liu help me pre-process the experimental data and give some useful feedback in improving the draft. L. Cheng gives insightful suggestions for this work and also revises the paper for several rounds.

Part II contains work on RGB-Thermal semantic segmentation from:

- **Chapter 6**: [4] **W. Ji**, J. Li, C. Bian, Z. Zhang, L. Cheng. "SemanticRT: A large-scale dataset and method for robust semantic segmentation in multispectral images". In Proceedings of the 31st ACM International Conference on Multimedia (ACM MM), 2023.

  I am responsible for proposing the method, designing and implementing the code, conducting the experiments, and writing the draft. I also lead the dataset collection and annotation with the assistance of J. Li, C. Bian and Z. Zhang. J. Li helps me pre-process the experimental data, provides lots of valuable feedback, and prepares the paper draft. L. Cheng gives insightful suggestions for this work and also revises the paper for several rounds.

- **Chapter 7**: [5] **W. Ji**, J. Li, C. Bian, Z. Zhou, J. Zhao, A. Yuille, L. Cheng. "Multispectral video semantic segmentation: a benchmark dataset and baseline". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

  I am responsible for proposing the method, designing and implementing the code, conducting the experiments, and writing the draft. J. Li helps me propose the method's technical design. The experimental results and analysis are carried out with the assistance of C. Bian and J. Zhao. The

dataset annotation is with the assistance of J. Li, C. Bian, Z. Zhou, and J. Zhao. A. Yuille provides valuable insights and helps me prepare the draft. L. Cheng gives insightful suggestions for this work and also revises the paper.

Chapter 8 contains discussion on recent segment anything model from:

- **Chapter 8**: [6] **W. Ji**, J. Li, Q. Bi, T. Liu, W. Li, L. Cheng. "Segment anything is not always perfect: An investigation of sam on different real-world applications". CVPR Workshop on Vision-based Industrial InspectiON (**Best Paper**, CVPR VISION Workshop), 2023.

  I am responsible for investigating the literature, conducting the experiments, and writing the draft. J. Li and Q. Bi give insightful suggestions and prepare the paper draft. T. Liu and W. Li help me pre-process experimental data and carry out part of the experimental results. L. Cheng provides some key insights for this work, and provides several suggestions to improve this work.

Acknowledgements

Throughout my doctoral studies, I feel incredibly fortunate to pursue a Ph.D. degree at the University of Alberta, under the guidance of Prof. Li Cheng. This phase of my academic pursuit holds profound significance in my personal and professional development. Numerous individuals have played pivotal roles in enriching this journey, making it stimulating, challenging, and enjoyable. Their support and assistance have been indispensable in shaping this thesis, and I am deeply thankful to each and every one of them.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Prof. Li Cheng, for his unwavering support, guidance, and mentorship. His reverence for and rigorous approach to research have profoundly influenced my research mindset. His openness has exposed me to fresh perspectives on a multitude of research topics. His passion for and commitment to research will continue to motivate me to advance and aspire to become an exceptional researcher. I am honored to have had the opportunity to learn from him.

Second, I extend my sincere thanks to the people who have guided me on my academic journey, including Shuang Yu, Cheng Bian, Wenbo Li, Zongwei Zhou, Alan Yuille, Zhicheng Zhang, and Yefeng Zheng. Their expertise, wisdom, and mentorship have played a pivotal role in my growth as a researcher and a professional. Without the help and support from them, I would not have been able to overcome the difficulties I encountered along the way and achieve my current accomplishments.

I am exceptionally fortunate to collaborate with remarkable individuals at the Vision and Learning Lab, including Shihao Zou, Chuan Guo, Ji Yang, Jingjing Li, Youdong Ma, Yuxuan Mu, Yilin Wang, Siyuan Li, Yande Li,

Muhammad Gohar Javed, Taivanbat Badamdorj, Size Wang, Xinxin Zuo, and Sen Wang. Their invaluable assistance and camaraderie have been truly appreciated. I am glad to work with them and am proud of the work we have done together.

Lastly, I am eternally grateful to my family for their unwavering love, support, and sacrifices. To my parents and grandparents, thank you for instilling in me the values of hard work and perseverance, and for always believing in me. To my wife, Goddess, your love, patience, and encouragement have been my cornerstone, and I could not have completed this journey without you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

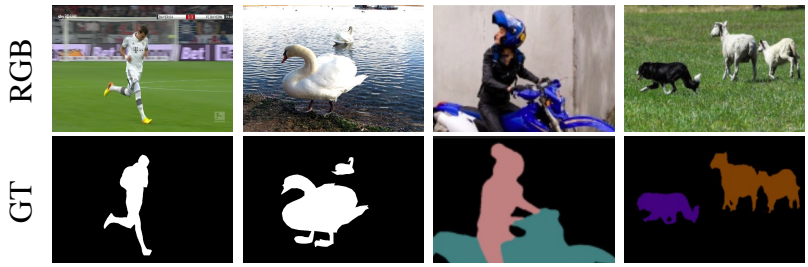
Imagine you are looking at a family photo. You can easily recognize family members, spot the family pet curled up in its favorite spot, and identify everyday items like the sofa or a lamp in the background. This task, effortless for us, presents a significant challenge to a computer.

Visual scene segmentation [7], a burgeoning field within computer vision, seeks to bridge this gap. It aims to decompose visual inputs into their constituent segments or regions, mirroring the human visual system's nuanced ability to scene understanding. This cutting-edge process leverages sophisticated algorithms and models capable of identifying, outlining, and categorizing elements within an image or video, thereby equipping machines with the ability to "understand" visual scenes. This capability has profound implications across a spectrum of real-world scenarios: from empowering autonomous vehicles to understand complex street scenes [8], [9], to enabling robots to navigate and interact within their spaces, from revolutionizing smart farming systems by pinpointing pest infestations to prevent crop reduction, to transforming medical imaging with precise identification of tissues or anomalies [10].

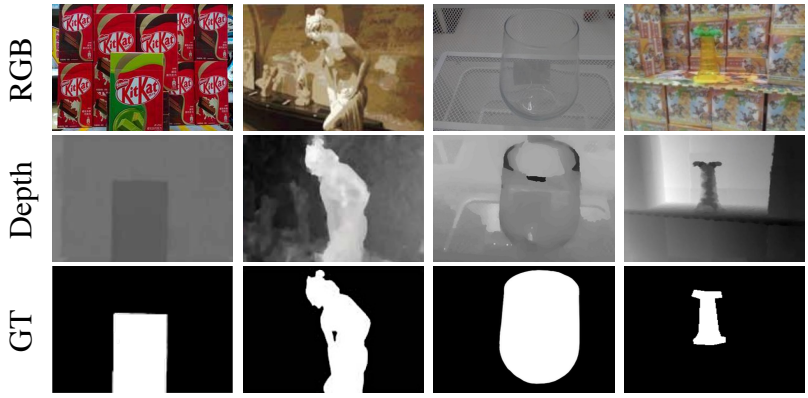
The increasing involvement of research institutions and companies in developing segmentation algorithms [11]–[20] has led to a wealth of literature on the subject. Earlier methodologies are mainly based on manually-crafted human priors, such as local contrasts [11], [12] to assess pixel or region rarity based on color and intensity variations within local surroundings, and back-

ground priors [13] which incorporate assumptions regarding the boundary and connectivity of the background to infer target objects. With the improvement of computer hardware performance and the popularity of big data, deep learning methods began to emerge in the field of computer vision. A turning point came with the emergence of AlexNet [14], which clinched victory in the ImageNet classification competition, heralding the dawn of the deep learning era. Harnessing the innate capability of deep learning to extract hierarchical representations from vast visual datasets, there has been a notable surge in the development of new segmentation approaches. Fully convolutional network (FCN [21]) stands as a milestone, catalyzing significant advancements in deep learning-based image segmentation. FCN removes fully-connected layers to accommodate arbitrarily-sized images and incorporates skip connections, allowing feature maps from deeper layers (providing semantic information) to be combined with those from shallower layers (providing appearance information). Subsequently, numerous ingenious techniques have been developed to enhance segmentation accuracy, including the deeplab series [15]–[17] which employ atrous convolution to preserve fine-grained details, [18], [19], leveraging deconvolution operations to reconstruct high-resolution regions from low-resolution counterparts, and [2], [20], harnessing convolutional/pooling kernels with varying sizes to capture diverse receptive fields within images, thereby effectively enriching contextual information.

Nevertheless, these researches focus on segmenting visual scenes solely based on appearance cues from RGB images or videos, which only perform well in easy and well-structured scenes. When adapting them to complex scenarios such as cluttered background, similar foreground and background, transparent objects, and poor lighting conditions, there remains a formidable challenge. As displayed in Fig. 1.1 (b) & (c), based solely on the appearance and textural cues from the visible RGB image, it is hard, even for human eyes, to discern the target objects in the complex environments. **Therefore, in this thesis, *we seek to advance the capabilities of deep learning-based segmentation for complex scene understanding by investigating additional imaging modalities.***



(a) Common examples



(b) Cluttered, similar background, transparent objects

(c) Adverse lighting conditions

Figure 1.1: Examples of visual scene segmentation in real-world environments. (a) Common examples. (b) Complex scenarios such as similar foreground and background, cluttered surroundings, and transparent objects. (c) Adverse lighting conditions such as over-exposure, low-light and darkness. The 'GTs' are presented with both binary masks in write/black format and semantic masks in color format to denote different semantic categories.

1.2 Motivations and Challenges

This thesis navigates two innovative avenues: 1) incorporating *depth data* to comprehensively understand the 3D spatial layout of scenes in RGB-Depth salient object detection[1], and 2) using *thermal infrared imagery* to enhance vision under adverse lighting conditions in RGB-Thermal semantic segmentation[2]. In what follows, we elaborate on the motivations and objectives of main components of this thesis.

1.2.1 RGB-Depth Salient Object Detection

In computer vision, the depth map can be seen as a digital simulation of human depth perception, which measures the distance from the camera to objects in a scene on a per-pixel basis. This depth information, imbued with rich 3D spatial structure and scene layouts, is crucial for enabling automated systems or machines to understand and interact with their surroundings effectively. This motivates our pursuit of RGB-Depth salient object detection (SOD) techniques, harmonizing 2D RGB and 3D depth information seamlessly to tackle the complexities inherent in diverse visual environments shown in Fig. 1.1 (b).

With the increasing prevalence of 3D imaging sensors in depth cameras, such as Kinect and Intel RealSense, as well as in mobile devices like iPhone 13 Pro, Huawei Mate 50, and Samsung Galaxy S21, the incorporation of depth information in addition to the conventional RGB image as input in RGB-D SOD is gaining increasing research interest. While prior efforts [22]–[25] have made commendable progresses, this emerging line of research has been considerably hindered by several common limitations and challenges:

- ***Noise and ambiguity that prevail in raw depth inputs.*** In essence, the actual value of depth in segmentation lies in its capability of discerning the object silhouette from background. Nevertheless, depth

[1]**Salient object detection** is a class-agnostic segmentation problem, which aims to segment out the most visually distinctive objects from background.

[2]**Semantic segmentation** is a class-aware segmentation problem, which aims to partition the scenes into segments according to predefined semantic categories.

maps are occasionally of low quality [26], [27] and thus may contain a lot of noise and misleading information, which results in the performance bottleneck of RGB-D SOD models to certain extent. Even with correct depth, the foreground object differs only slightly from the surrounding background in the depth maps. This may be hampered by the limitation of depth sensors and scene configurations such as occlusion [28], reflection [29], [30] and viewing distance [31].

Recently, there have been several emerging research works shedding light on the influence of unreliable depth and trying to address it. Zhao *et al*. [32] adopted a contrast prior loss to enhance the color difference between foreground and background of depth data. Similarly, Zhang *et al*. [33] proposed a semantic guided depth correction subnetwork to produce enhanced depth cues under the assumption that edges of depth map should be aligned with edges of the RGB image. Fan *et al*. [34] designed a three-stream feature learning network, and performed a depth depurator unit to filter low-quality depth maps during the test phase. Furthermore, Chen *et al*. [35] leveraged the retrieval of a small set of similar images from external datasets to acquire additional enhanced depth information, and employed a selective fusion way to extract handcrafted saliency clues from the enhanced depth, original depth and RGB image for more accurate segmentation.

In this thesis, we systematically address the depth-related side effects, and propose a depth calibration strategy to tackle the noise issue. Different from existing approaches, our work aims to directly calibrate the raw depth, and the calibrated depth provides more reliable complementary information for RGB-D SOD, which significantly boosts the performance. Meanwhile, when directly applying the calibrated depth to existing models, noticeable performance gain is also observed.

- ***Less effective network architectures.*** Prior studies [22]–[25], [36] have designed plausible network architectures aimed at enhancing the interaction between complementary cross-modal data. Commonly, these

designs adopt a dual-stream fusion approach, where RGB and depth information are processed independently in separate streams, followed by the integration of shared layers at either an initial or a later stage to facilitate the learning of combined representations and collaborative decision-making. Nonetheless, our analysis indicates that such architectures fall short of expectations, yielding suboptimal results.

When designing RGB-D SOD models, there are three points should to be considered. 1) Multiple objects in a scene have large variations in both depth and scale. Exploring the relationship between depth cues and objects with different scales can further provide vital guidance cues for obtaining informative feature representation. However, to our best knowledge, this relevance has never been researched in previous RGB-D SOD works. 2) Studies show that people perceive visual information using an Internal Generative Mechanism (IGM) [37], [38]. In the IGM, environment captured by human is not a straight translation of the ocular input, but a result of a series of active inferences of brains, especially in complex scenes. However, the benefits of IGM for comprehensively understanding a scene have never been explored in previous works. Particularly, the fused feature is directly used for prediction while the internal semantic relation in the fused feature is ignored. 3) Deep features in the hierarchical feature representations can provide discriminative semantic cues while the shallow features also contain affluent local details for accurately segmenting target objects. Designing an efficient multi-level feature fusion strategy is essential for the segmentation task.

To this end, we propose a depth-induced multi-scale recurrent attention network for RGB-D SOD, named as DMRA, which has three key components: 1) we design a depth-induced multi-scale weighting (DMSW) module, where the relationship between depth information and objects with different scales is explored for the first time in RGB-D SOD task. Ablation analysis shows that utilizing this relevance can improve detection accuracy and facilitate the integration of RGB and depth data; 2) we

design a novel recurrent attention module (RAM) inspired by the IGM of human brain. It can iteratively generate more accurate saliency results in a coarse-to-fine manner by comprehensively learning the internal semantic relation of the fused feature. When inferring the current result, the RAM retrieves the previous memory to aid current decision, thereby progressively optimizing local details with memory-oriented scene understanding; and 3) we devise a bottom-up cascaded hierarchical feature fusion strategy (CHFF) with a channel-specific contextual interaction block (CCIB) to progressively integrate multi-level cross-modality features. Such efficient feature interaction enables to obtain more reliable predictions.

- ***Non-sustainable annotation efforts in supervised segmentation.*** Modern RGB-D SOD has greatly benefited from deep learning advances. However, the development is impeded by a significant barrier: deep learning is data hungry by nature, demanding large-scale, high-quality annotated datasets, especially for the segmentation task which requires pixel-level annotations. This however becomes much less appealing in practical scenarios, owing to the laborious and time-consuming process in obtaining manual annotations. It is therefore natural and desirable to consider unsupervised alternatives.

  Unfortunately, existing unsupervised RGB-D SOD methods, such as global priors [39], center prior [40], and depth contrast prior [41], rely primarily on handcrafted feature representations. This is in stark contrast to the deep representations learned by their supervised counterparts, which in effect imposes severe limitations on the feature representation power that may otherwise benefit greatly from the potentially abundant unlabeled RGB-D images.

  These observations motivate us to explore a new problem of deep unsupervised RGB-D saliency detection: given an unlabeled set of RGB-D images, deep neural network is trained to predict saliency without any laborious human annotations in the training stage.

1.2.2 RGB-Thermal Semantic Segmentation

As showcased by the exemplary RGB images in Fig. 1.1 (c), it could be exceedingly challenging even for human eye to discern the pedestrians and vehicles in the scenarios of low-light night or when facing a strong coming headlight. On the contrary, thermal infrared imaging proves invaluable under these circumstances, serving as a complementary source of information to conventional RGB images by capturing the infrared radiation emitted by objects warmer than absolute zero [42].

This has naturally led to a growing interest in RGB-Thermal Semantic Segmentation, also known as, Multispectral Semantic Segmentation (MSS), where a pair of RGB and thermal (RGB-T) images is used as an input, to address the limitations of traditional RGB models in adverse lighting conditions [42]. This line of research has seen a range of real-world applications, from autonomous safe driving [9], night patrol [43], and fire rescue [44], to object tracking [45]. However, despite its importance, this area of study is still in the early stages of development, largely due to the scarcity of large-scale datasets. In this thesis, we aim to support the advancement of RGB-Thermal Semantic Segmentation in both the image and video domains:

- ***Multispectral image semantic segmentation.*** The progression of benchmark datasets has played essential roles underpinning the development of MSS methods. The pioneering benchmark, MFNet [42], offers 1,569 RGB-T images, accompanied with pixel-wise annotations that support the training and evaluation of MSS models. Another dataset, PST900 [46], contains 894 pairs of images captured in underground tunnels and caves, which could serve to validate the generalization capabilities of MSS models.

  Though these RGB-T datasets have contributed to MSS advancements, they present certain limitations. One primary obstacle is the absence of large-scale benchmarks. Contrasting to the RGB-based semantic segmentation datasets such as Cityscapes [8] and PASCAL-Context [47] that contain 5,000-10,103 finely annotated images, for the MSS commu-

nity, existing benchmark datasets are considerably smaller – the largest benchmark containing only 1,569 images. This has imposed a severe limit toward developing better MSS models. Additionally, existing datasets typically lack diversity in scene contents & categories, and feature low image resolutions. This may impede practical development in the MSS field. To tackle these challenges, we curate a large-scale SemanticRT dataset, comprising 11,371 RGB and thermal infrared image pairs, accompanied with high-quality, pixel-wise annotations over 13 categories. It also covers a diverse range of scenarios (*e.g.*, road, park, campus, street) in both daytime and nighttime settings. The majority (over 95%) of these RGB-T image pairs are of high-resolution (1280×1024). Furthermore, we incorporate in this new dataset several distinct attribute-based test subsets, configured according to often-used image properties, such as daytime, nighttime, multi-object, multi-class, and low-contrast. This is expected to be used for a comprehensive robustness evaluation of existing and new MSS algorithms.

With access to these rich multispectral cues, existing MSS methods have developed plausible solutions to unify the two types of information, by concatenating or summing multimodal features from separate encoders [42], [48], [49], direct incorporation of thermal images as an additional input channel [46], or weighted attention fusions [50], [51]. However, the results of these *implicit* fusion strategies are still unsatisfactory, since they often indiscriminately aggregate two modal cues extracted from individual feature extractor, which could bring an overemphasis on shared high-intensity information, and eventually dilute the useful modality-specific cues, finally weaken their discriminative power in scene representations. To address this problem, an explicit complement modeling (ECM) scheme is developed to better exploit the complementary characteristics of both the RGB and thermal modalities.

- ***Multispectral video semantic segmentation.*** Existing RGB-T segmentation methods are based on single images. However, the lack of

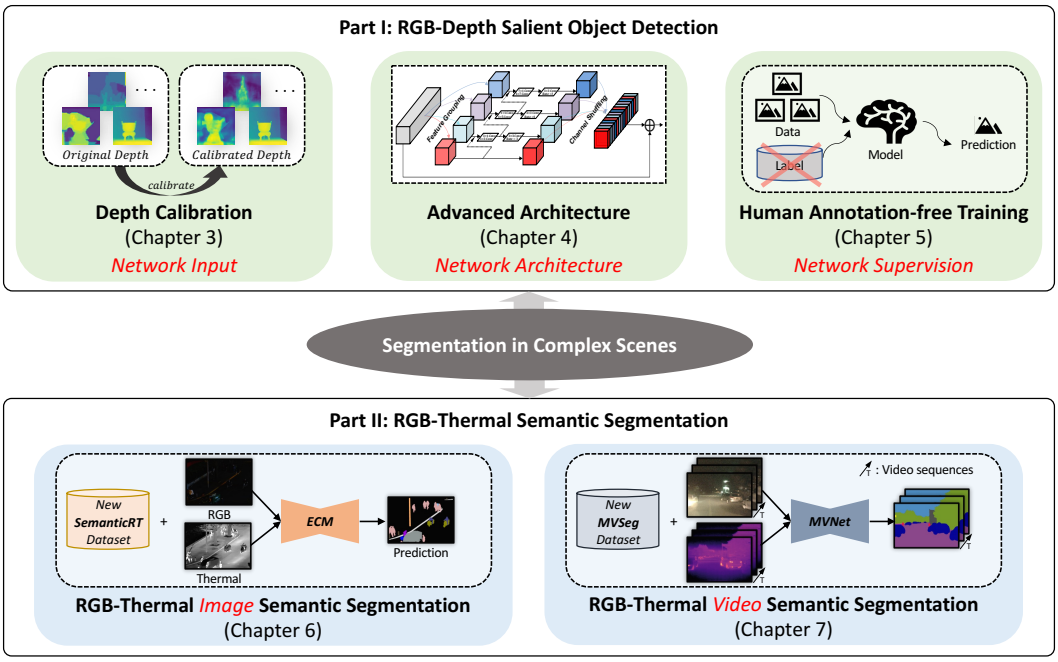


Figure 1.2: Thesis Overview.

mechanism to account for the temporal contexts may limit their performance when working with video inputs containing dynamic scenes, which are omnipresent in our daily lives. This leads us to explore in this thesis a relatively new task of Multispectral Video Semantic Segmentation, or in short MVSS, with a specific focus on RGB-T video inputs. The RGB frames and thermal frames can provide rich and often complementary information for identifying moving foreground objects and static background scenes in low-light night or facing strong headlights. To our knowledge, this is the first work to address such multispectral video semantic segmentation problem.

An in-house MVSeg dataset is thus curated, consisting of 738 calibrated RGB and thermal videos, accompanied by 3,545 fine-grained pixel-level semantic annotations of 26 categories. Our dataset contains a wide range of challenging urban scenes in both daytime and nighttime. Moreover, we propose an effective MVSS baseline, dubbed MVNet, which is to our knowledge the first model to jointly learn semantic representations from multispectral and temporal contexts.

1.3 Summary of Contributions

In this thesis, we present a comprehensive suite of methodologies and resources tailored for enhancing visual scene segmentation capabilities. As outlined in Fig. 1.2, our contributions are organized into two parts: 1) the integration of depth data in RGB-Depth Salient Object Detection, and 2) the employment of thermal infrared imagery in RGB-Thermal Semantic Segmentation.

The contributions are summarized as follows:

- **Part I: RGB-Depth Salient Object Detection**

  In this part, we propose three novel strategies to improve segmentation efficacy. We optimize three key aspects of deep learning models, namely, network input, network architecture, and network supervision:

  - *Calibrating depth input* (Chapter 3): We devise a novel depth calibration strategy that is capable of effectively calibrating/correcting the latent bias in the original depth images. In doing so, we design a depth discriminator to distinguish depth maps with bad quality (negative cases) from the good quality ones (positive cases), and devise a depth estimator to estimate good quality depth maps from RGB data. We then replace the original depth map with the weighted summation between the raw depth map and the estimated depth, based on weights determined by a reliability probability predicted by the discriminator. The calibrated depth has been proved to effectively improve the model performance. It can also serve as a preprocessing step that is directly applicable to existing RGB-D salient object detection methods to boost the performance.
  - *Advanced network architecture* (Chapter 4): We propose a novel depth-induced multi-scale recurrent attention network for RGB-D saliency detection, named as DMRA. It achieves dramatic performance especially in complex scenarios. Specifically, we combine depth cues with abundant spatial information with multi-scale contextual features for accurately locating salient objects. We also

devise an effective recurrent attention module inspired by Internal Generative Mechanism of human brain to generate more accurate saliency results via comprehensively learning the internal semantic relation of the fused feature and progressively optimizing local details with memory-oriented scene understanding. Finally, a cascaded hierarchical feature fusion strategy is established to promote efficient information interaction of multi-level contextual features and further improve the contextual representability of model. Extensive empirical experiments demonstrate that our method can accurately identify salient objects.

- *Human annotation-free training* (Chapter 5): We tackle a new task of deep unsupervised RGB-D saliency detection, which requires no manual pixel-level annotation during training. Our key insight is to internally engage and refine the pseudo-labels. It is realized by two key ingredients: 1) a depth-disentangled saliency update (DSU) framework is designed to automatically produce pseudo-labels with iterative follow-up refinements, which provides more trustworthy supervision signals for training the saliency network; 2) an attentive training strategy is introduced to tackle the issue of noisy pseudo-labels, by properly re-weighting to highlight the more reliable pseudo-labels. Extensive experiments demonstrate the superior efficiency and effectiveness of our approach in tackling the challenging unsupervised RGB-D SOD problem.

- **Part II: RGB-Thermal (Multispectral) Semantic Segmentation**

  In this part, we contribute to the advancement of RGB-Thermal Semantic Segmentation in both the image and video domains:

  - *Multispectral image semantic segmentation (MSS)* (Chapter 6): We present SemanticRT, a new large-scale multispectral semantic segmentation dataset that covers diverse scenarios and varying illumination conditions. Composed of high-quality RGB-T pairs with

pixel-wise annotations and attribute-based testing subsets, SemanticRT is expected to facilitate model development and comparison in the MSS field. Additionally, we propose an explicit complement modeling (ECM) framework that explicitly captures modality-specific useful cues and incorporates them into robust cross-modal feature fusion and encoding process. Extensive empirical results demonstrate the effectiveness of our approach.

– *Multispectral video semantic segmentation (MVSS)* (Chapter 7): we present a preliminary investigation on the new task of semantic segmentation of multispectral video inputs. Specifically, we have provided a new challenging and finely annotated MVSeg dataset, developed a simple but efficient baseline framework (*i.e.*, MVNet), conducted comprehensive benchmark experiments, and highlighted several potential challenges and future directions. The above contributions provide an opportunity for the community to design new algorithms for robust MVSS.

1.4 Organization of Thesis

This thesis is structured to provide a clear and comprehensive exploration of the research conducted. Below is an overview of the content and organization:

Chapter 2: Related Work

This chapter offers an extensive review of existing research, covering areas including RGB-based/RGB-D/unsupervised RGB-D salient object detection, RGB-based/RGB-Thermal semantic segmentation, as well as representative model architectures.

Part I: RGB-Depth Salient Object Detection (Chapters 3 - 5)

- Chapter 3 introduces a novel depth calibration strategy to correct latent noise inherited in raw depth maps, laying the groundwork for improved detection accuracy.

- Chapter 4 presents the Depth-induced Multi-scale Recurrent Network (DMRA), a new network architecture designed for enhanced RGB-D salient object detection.
- Chapter 5 explores a new task of Deep Unsupervised RGB-D Salient Object Detection, showcasing our method that forgoes the need for manual annotations.

Part II: RGB-Thermal Semantic Segmentation (Chapters 6 & 7)

- Chapter 6 unveils the SemanticRT dataset and the Explicit Complements Modeling (ECM) framework, both pivotal for advancing multispectral image semantic segmentation.
- Chapter 7 discusses multispectral video semantic segmentation (MVSS), introducing the MVSeg dataset and the baseline MVNet network, highlighting our contributions to this new task.

Chapter 8: Conclusion, Discussion and Future Work

The final chapter reflects a conclusion of the proposed methods. It also also discusses the benefits and limitations of a recent foundational model - Segment Anything Model, and outlines some compelling issues and future research avenues within the field.

Chapter 2

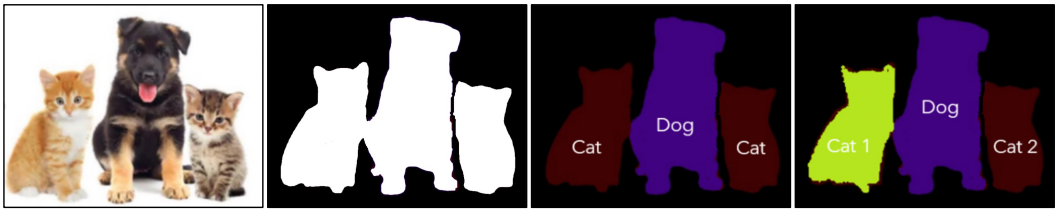
Related Work

The purpose of this chapter is to establish the foundational background for this thesis, encompassing several crucial components. We begin with a review of prior research in related fields in Sec. 2.1. Following that, in Sec. 2.2, we introduce several widely-used segmentation networks spanning various architectures and data sources.

2.1 Related Research Topics

Visual scene segmentation is a fundamental task in computer vision, with the objective of partitioning a scene into meaningful regions or objects. This field encompasses various subtasks, including salient/foreground object segmentation, semantic segmentation, instance segmentation, and *etc.*

As displayed in Fig. 2.1, salient/foreground object segmentation [52], also known as salient object detection, highlights specific objects or regions deemed important within an image. It typically exhibits a class-agnostic binary segmentation task, where the goal is to distinguish salient objects from the background. Semantic segmentation [21], on the other hand, assigns a predefined semantic label to every pixel in an image, providing a detailed understanding of the scene's content and structure. Unlike semantic segmentation, instance segmentation [53] can differentiate between multiple objects of the same class and provide separate masks for each instance. In this thesis, we primarily focus on salient object detection and semantic segmentation in complex scenarios.



Original Image | Salient Object Detection | Semantic Segmentation | Instance Segmentation

Figure 2.1: Illustrations of fundamental visual scene segmentation subtasks, including salient object detection, semantic segmentation, and instance segmentation.

2.1.1 RGB-based Salient Object Detection

Salient object detection (SOD) [11], [13], [39], [41], [54]–[58] aims to segment out the most visually striking or attention-grabbing objects or regions from the background. It is characterized by its class-agnostic nature, meaning it doesn't require prior knowledge of specific object classes. As a crucial branch of the visual scene segmentation field, SOD can be regarded as a binary image segmentation task, where each pixel is categorized as either foreground or background. This simplifies and/or changes the representation of an image into something that is more meaningful and easier to analyze.

At present, there are many different algorithms and approaches for SOD, and they can generally be grouped into two main categories: (1) Rule-based approaches: These methods rely on predefined rules or heuristics to identify and locate salient objects. These can include thresholding, appearance contrast [59], edge constrain [60] or background modeling [13]. Typically, Roberto *et al*. [59] develop a computational method to infer visual saliency in images, which is based on the assumption that salient objects should have local characteristics that are different than the rest of the scene, being edges, color or shape. Liu *et al*. [60] integrate multiscale contrast, center-surround histogram, and color spatial distribution, to describe a salient object locally, regionally, and globally. Subsequently, the boundary and connectivity priors are introduced in[13] to model the properties of background to obtain saliency. (2) Deep learning-based approaches: They [10], [16], [54], [55] usually adopt convolutional neural networks (CNNs) [14] to learn the powerful hierarchical features

from a lot of data and generate predictions. In SOD field, Wang *et al*. [61] use a CNN to predict saliency score for each pixel in local context first, then they refine the saliency score for each object proposal over the global view. Li and Yu [62] produce the saliency score for each superpixel by using multiscale CNN features. Similarly, Zhao *et al*. [63] predict the saliency score for each superpixel by incorporating local context and global context simultaneously in a multi-context CNN. To further learn enough global structures, DHSNet [64] adopt the whole image as the computational unit and propagates the global context information to local contexts hierarchically and progressively, being able to perceive global properties and avoid the distraction of local interferences from the beginning. Meanwhile, to obtain the fine edge details, in [57], the boundaries of salient objects are explicitly modeled, aiming to leverage the salient edge features to help the salient object features locate objects.

2.1.2 RGB-Depth Salient Object Detection

RGB-D SOD [34], [39]–[41], [56], [65]–[69] aims to identify interested target objects by taking advantage of complementary RGB-D data, especially in complex scenarios. This is attributed to the fact that depth cues can provide affluent spatial structure and 3D layout information, making it easy to find target regions from cluttered background.

Existing RGB-D SOD methods can be generally classified into two categories: (1) manually designing hand-crafted features; (2) automatically extracting features with CNNs. For hand-crafted methods, Peng *et al.* [56] utilize a multi-stage model combining RGB-produced saliency with new depth-induced saliency for SOD. Zhu *et al.* [40] propose to utilize a center saliency prior and a dark channel prior for extracting RGB-D complementary information. Ren *et al.* [39] exploit the normalized depth prior and the global-context prior for further predicting saliency. Those methods, mainly relying on human-designed priors and lacking of high-level semantic representations, are limited to the expression ability of handcrafted features and are difficult to be adapted for understanding global context. Recently, the emergence of CNNs (Convolutional Neural Networks [70]) have significantly pushed the performance of

low-level computer vision tasks for its powerful ability in automatically extracting hierarchical context features. This naturally leads to the effective integration of cross-modal features in both RGB and depth views. [71] utilizes hand-crafted features to train a CNN-based model and achieves significant improvements over traditional methods. Fan *et al.* [34] design a three-stream feature learning network, and perform a depth depurator unit to filter unreliable depth information. Minhyeok et al. [72] introduce a prototype sampling network designed to selectively sample prototypes representing salient objects in both RGB and depth perspectives. Subsequently, they employ a reliance selection module to assess the efficacy of individual RGB and depth feature maps, dynamically adjusting their weighting based on their reliability. Wu *et al.* [65] introduce a granularity-based attention scheme to strengthen the discriminatory power of RGB and depth features separately. This effectively promotes sufficient feature interactions.

2.1.3 Unsupervised RGB-Depth Salient Object Detection

Remarkable progresses have been made recently in RGB-D salient object detection (SOD)/image segmentation [35], [73]–[76] that integrate effective depth cues to tackle cluttered background issues. Those RGB-D SOD methods, however, typically demand extensive annotations, which are labor-intensive and time-consuming. This naturally leads to the consideration of unsupervised counterparts that do not rely on such human annotations.

Prior to the deep learning era [52], [77], traditional RGB-D methods are mainly based on the manually-crafted RGB features and depth cues to infer a saliency/segmentation map. Due to their lack of reliance on manual human annotation, these traditional methods can be regarded as early manifestations of unsupervised SOD. Ju *et al.* [41] and Feng *et al.* [12] present depth-aid saliency methods based on anisotropic center-surround difference prior or local background enclosure prior. Lang *et al.* [78] utilize Gaussian mixture models to model the distribution of depth-induced saliency. In contrary to direct utilization of the depth contrast priors, local background enclosure prior is

explicitly developed by [12]. However, by relying on manually-crafted priors, these methods tend to have inferior performance. Meanwhile, considerable performance gain has been achieved by recent efforts in RGB-based unsupervised SOD [79]–[82], which instead construct automated feature representations using deep learning. A typical strategy is to leverage the noisy output produced by traditional methods as pseudo-label (*i.e.*, supervisory signal) for training saliency prediction net. The pioneering work of Zhang *et al.* [79] fuses the outputs of multiple unsupervised saliency models as guiding signals in CNN training. In [80], competitive performance is achieved by fitting a noise modeling module to the noise distribution of pseudo-label. Instead of directly using pseudo-labels from handcrafted methods, Nguyen *et al.* [81] further refine pseudo-labels via a self-supervision iterative process. Besides, deep unsupervised learning has been considered by [83], [84] for the co-saliency task, where superior performance has been obtained by engaging a fusion-learning scheme [83], or utilizing dedicated loss terms [84]. These methods demonstrate that the incorporation of powerful deep neural network brings better feature representation than those unsupervised SOD counterparts based on handcrafted features.

2.1.4 RGB-based Semantic Segmentation

Semantic segmentation is to divide an image into meaningful segments or regions, where each pixel is assigned to a specific class label. In last years, this field has experienced significant growth, primarily due to the accessibility of large-scale datasets (*e.g.*, Cityscapes [8] and PASCAL-Context [47]), rapid advancements in convolutional networks (*e.g.*, VGG [70] and ResNet [85]), the evolution of segmentation models (*e.g.*, FCN [21], U-Net [10], and DeepLab series [15], [86]), and a wide range of practical applications (*e.g.*, autonomous driving [20], [87]–[89], scene understanding [1], [2], [90]–[92], video processing [5], [93]–[95] and medical diagnosis [96]–[98]).

A milestone in the field is the development of FCN [21], which introduces fully convolutional networks for per-pixel representation learning. Following this, numerous methods [19], [20], [99]–[112] have been proposed to aggre-

gate contextual information to augment network's feature representation. For example, [15], [20], [108] focus on improving the network's receptive field by rethinking the role of pyramid operations, and [110], [111] design novel attention mechanisms to suppress unnecessary spatial distractions. Recently, vision transformers have gained popularity in semantic segmentation [113]–[115] due to their ability to capture global context [116], leading to impressive performance improvements. Xie *et al.* [115] present SegFormer that comprises a hierarchically structured Transformer encoder which outputs multiscale features. It does not need positional encoding, thereby avoiding the interpolation of positional codes which leads to decreased performance when the testing resolution differs from training. [114] design a progressive shrinking pyramid and a spatial-reduction attention based on transformer backbone, enabling a convolution-free object detection pipeline. This makes the network adaptable for learning multi-scale and high-resolution features. Interested readers can refer to review articles of [7], [117], [118] for more details. Despite considerable progress has been made, when facing with challenging conditions such as darkness or dim lighting, these models are prone to erroneous prediction due to the limited appearance cues on RGB input alone.

2.1.5 RGB-Thermal Semantic Segmentation

To deal with the severe performance degradation of these RGB-based methods under poor lighting conditions, the incorporation of thermal imaging (*i.e.*, RGB-T/multispectral semantic segmentation) has been recently investigated [42], [46], [49]–[51], [119].

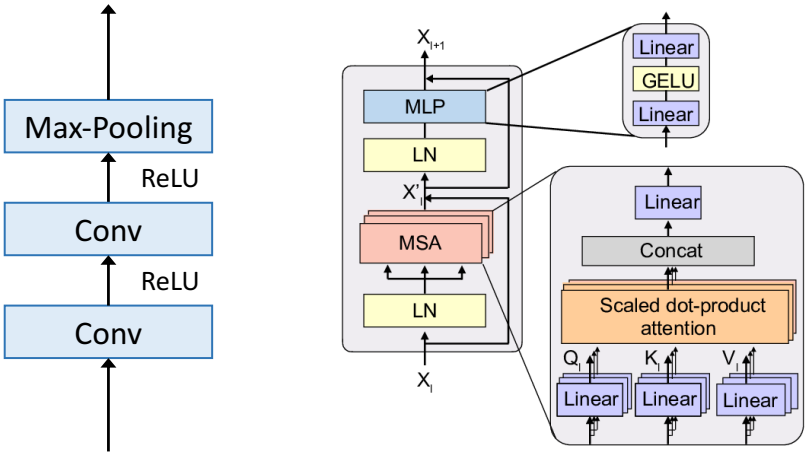
The seminar work of MFNet [42] introduces a two-stream structure with mini-inception blocks to extract complementary features, as well as contributes the first RGB-T semantic segmentation benchmark. [48] extracts thermal feature maps from an extra encoder and fuses them into the RGB encoder through element-wise summation. In the decoding stage, two types of upsampling inception blocks extract features and restore their resolution. At the same time, a FuseSeg [49] is developed based on the dense connection of [120] and a two stage fusion strategy is applied in the encoder and decoder. To verify the gen-

eralization capability of learned models, Shivakumar *et al.* [46] introduce the PST900 dataset and design a dual-branch CNN that integrates the segmentation result from independent RGB branch with original thermal input. Deng *et al.* [119] conceive a two-stage feature-enhanced attention network, aiming to excavate and use multi-level features from both the channel and spatial views. To obtain fine-grained predictions, EGFNet [51] designs its method to embed prior edge maps into the boundary features toward enhancing object-level details. Similarly, a multilabel supervision is designed in [121] to optimize the network in terms of semantic, binary, and boundary characteristics. In [122], the authors use a hybrid fusion module to integrate the complementary information across modalities while considering the propagation of fusion cues by incorporating previously fused features. A grid-like context-aware module is further designed to capture rich contextual information, thereby achieving the optimal performance. The work of Zhang *et al.* [50] sheds light on the influence of modality differences by proposing a bridging-then-fusing strategy, where two bi-directional image-to-image translation networks are used to reduce the differences from two modalities. To enhance segmentation accuracy further, an optimization strategy incorporating progressive deep supervision loss [123] is proposed. This strategy directly supervises both upper and lower layers of the RGB-T decoder, guiding it to achieve precise segmentation in a coarse-to-fine manner.

2.2 Related Model Architectures

2.2.1 Convolutional Neural Network

Convolutional Neural Networks (CNNs) [14] are widely recognized and highly effective architectures in the realm of deep learning, particularly for computer vision tasks. As shown in Fig. 2.2 (a), CNNs typically comprise three primary layers: 1) convolutional layers, which utilize weighted kernels or filters to extract features from input data; 2) nonlinear layers, applying activation functions to feature maps to model complex, nonlinear relationships; 3) pooling layers, which reduce spatial resolution by summarizing local neighbor-



(a) Conv Block (b) Transformer Block

Figure 2.2: Basic components in convolutional neuronal network and transformer. (a) Convolution, ReLU Nonlinear Unit and Max Pooling in VGG [70]. (b) Basic block in Transformer [124]. Figures adapted from [124].

hoods within feature maps using statistical measures like mean or max pooling. CNNs offer computational efficiency through weight sharing, significantly reducing the number of parameters compared to fully-connected networks. Well-known CNN architectures include VGG [70], and ResNet [85]. Following that, Fully Convolutional Networks (FCNs), introduced by Long *et al.* [21], marked a significant advancement in deep learning-based image segmentation. FCNs consist solely of convolutional layers, enabling them to produce segmentation maps with the same dimensions as the input image. To accommodate arbitrarily-sized images, FCNs modify existing CNNs (*e.g.*, VGG [70]) by removing fully-connected layers, resulting in spatial segmentation maps rather than classification scores. They incorporate skip connections, allowing feature maps from deeper layers (providing semantic information) to be combined with those from shallower layers (providing appearance information). This fusion enhances the model's ability to produce accurate and detailed segmentations. Motivated by the success of FCNs, various efficient structures like DeepLab [16], SegNet [18], and PSPNet [20] have emerged, aiming to further improve segmentation accuracy.

2.2.2 Vision Transformer

In recent years, researchers have been inspired by the success of Transformer architectures [124] in the field of Natural Language Processing (NLP) and have begun exploring attention mechanisms as replacements or enhancements for traditional convolutional layers. By utilizing multi-head attention mechanisms shown in Fig. 2.2 (b), Transformers excel at modeling long-range dependancies of data, which addresses limitations associated with the locality property of FCN-based methods [18], [21]. However, implementing attention in convolutional architectures can be computationally demanding, particularly for image. The computation cost of self-attention grows quadratically with image size, as each pixel attends to every other pixel. To tackle this challenge, Dosovitskiy *et al.* [125] propose dividing images into a sequence of patches, treating them as tokens similar to NLP. This shift from pixel-wise to patch-wise attention significantly reduces computational complexity while maintaining competitive performance. The resulting architecture has not only outperformed state-of-the-art FCN-based methods [15], [16], [20] but has also paved the way for subsequent research efforts. Notable examples include the Swin Transformer [113], which has consistently delivered impressive results. Building upon these advancements, more powerful fundamental segmentation architectures have emerged, such as Segformer [115] and PVT [114], taking segmentation performance to new heights.

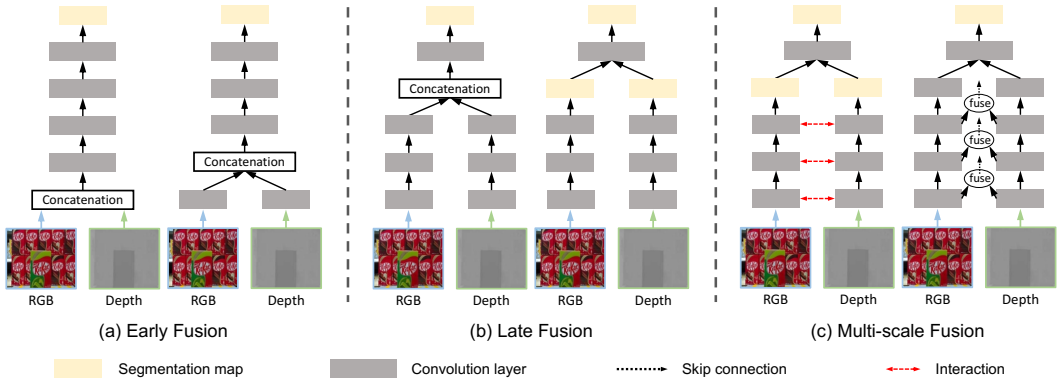


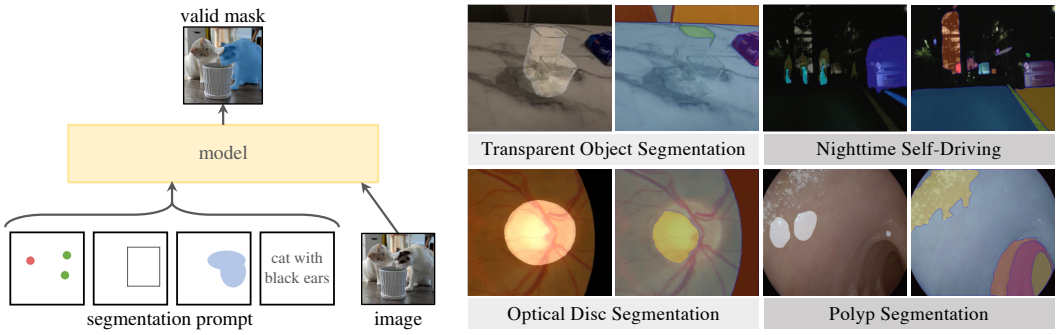
Figure 2.3: Fusion strategies in two-stream fusion networks: (a) early fusion, (b) late fusion, and (c) multi-scale fusion. Figures adapted from [34].

2.2.3 Two-Stream Fusion Network

With the popularity of depth and thermal infrared sensors, two-stream fusion networks [126] have garnered increased attention in the field of computer vision. These networks aim to effectively combine RGB images with complementary modalities like depth maps or thermal data to address challenges posed by complex backgrounds and varying lighting conditions. Various methods [2], [22], [42] have been developed to enhance the performance of such fusion networks. In the context of two-stream fusion networks, the fusion of RGB images with complementary modalities stands as a pivotal aspect.

These fusion strategies [2], [22], [25], [127]–[129] can be broadly classified into three distinct approaches, as depicted in Fig. 2.3. a) Firstly, early fusion entails the direct integration of RGB images and depth maps, forming a unified four-channel input [39], commonly as input fusion. An alternative early fusion technique involves processing RGB and depth images separately through dedicated networks. Subsequently, their low-level representations are combined before feeding into a subsequent network for prediction [71], termed early feature fusion. b) Late fusion methods can be categorized into two primary families. In the first approach, two parallel network streams are employed to acquire high-level features for RGB and depth data. These features are concatenated before generating the final saliency prediction [23], known as late feature fusion. In the second approach, two parallel network streams independently produce segmentation maps for RGB images and depth cues. These maps are then aggregated to obtain the final prediction [127], *i.e.*, late result fusion. c) To effectively capture the correlations between RGB images and depth maps, some methods employ a multi-scale fusion strategy [1], [2], [34], [130]. These models fall into two categories: The first category focuses on learning cross-modal interactions and integrating them into a feature learning network. For instance, Chen *et al.* [131] develop a multi-scale, multi-path fusion network with a cross-modal interaction module. This approach introduces cross-modal interactions at multiple layers, enhancing depth stream learning and exploring complementarity between low-level and high-level representa-

tions. The second category [132] involves fusing features from RGB images and depth maps at different layers and then integrating them into a decoder network, often using skip connections, to produce the final segmentation map.



(a) SAM (b) Results of SAM

Figure 2.4: Segment Anything Model (SAM [133]): (a) A global view of SAM; (b) Results of SAM on various real-world applications, where we adopt Everything mode to obtain SAM segmentations (right). The ground truth is masked with image for reference purpose (left). Figures adapted from [6], [133].

2.2.4 Segment Anything Model

Taking advantage of the accessibility of web-scale datasets and ample computing resources, there has been a notable increase of interest in foundational models. In 2023, Meta AI Research unveiled a promptable Segment Anything Model (SAM [133]). As illustrated in Fig. 2.4 (a), SAM showcases the ability to segment any object within images or videos without requiring additional training, a capability often termed as zero-shot transfer in the vision community. These prompts can involve various forms, including foreground/background points, rough boxes or masks, freeform text, or any other cues indicating the target of segmentation within an image. As indicated by [133], SAM's functionality is underpinned by a foundational vision model, such as Transformer, trained on an extensive SA-1B dataset comprising over 11 million images and one billion masks.

The advent of the promptable SAM has revolutionized segmentation models, owing partly to its unprecedentedly large segmentation dataset. It is of great practical interest to investigate how well SAM can be generalized to

challenging conditions such as transparent objects, low-light night or darkness environments. Several studies [6], [134], [135] have conducted experiments to exam SAM's performance across a diverse range of real-world segmentation applications. As observed in Fig. 2.4 (b), SAM encounters difficulties in detecting whole target objects in transparent or low-light conditions due to the limited information provided by a single modality. Meanwhile, SAM's performance in medical applications remains suboptimal due to its lack of medical-specific knowledge. Consequently, addressing challenging environments in specific applications remains an open problem.

Part I

RGB-Depth Salient Object Detection

In computer vision, the depth map serves as a digital simulation of human depth perception. It measures the distance from the camera to objects in a scene on a per-pixel basis. This depth information, enriched with rich 3D spatial structure and scene layouts, is essential for empowering automated systems or machines to interpret and interact with their surroundings effectively. Therefore, we focus on developing techniques for RGB-Depth salient object detection (SOD) in this section, aiming to integrate 2D RGB and 3D depth data flawlessly to navigate the complexities inherent in complex visual environments. This task has proven indispensable across various real-world applications, such as video conferencing [136] and image manipulation [137].

Within this part, we confront three main challenges associated with RGB-Depth salient object detection. First, in Chapter 3, we tackle the prevalent issue of noise and ambiguity in raw depth inputs by designing a novel depth calibration strategy. Next, in Chapter 4, we propose a depth-induced multi-scale recurrent attention network, known as DMRA, which enhances the fusion of multimodal information and contextual understanding. Lastly, in Chapter 5, we explore a new problem of deep unsupervised RGB-D saliency detection, aiming to minimize the dependency on extensive human labeling.

This part is based on the following publications:

- **Chapter 3**: [1] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu, L. Cheng. "Calibrated RGB-D salient object detection". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

- **Chapter 4**: [2] W. Ji, G. Yan, J. Li, Y. Piao, S. Yao, M. Zhang, L. Cheng, H. Lu. "DMRA: Depth-induced multi-scale recurrent attention network for RGB-D saliency detection". IEEE Transactions on Image Processing (IEEE TIP), 2022.

- **Chapter 5**: [3] W. Ji, J. Li, Q. Bi, C. Guo, J. Liu, L. Cheng. "Promoting saliency from depth: Deep unsupervised RGB-D saliency detection". International Conference on Learning Representations (ICLR), 2022.

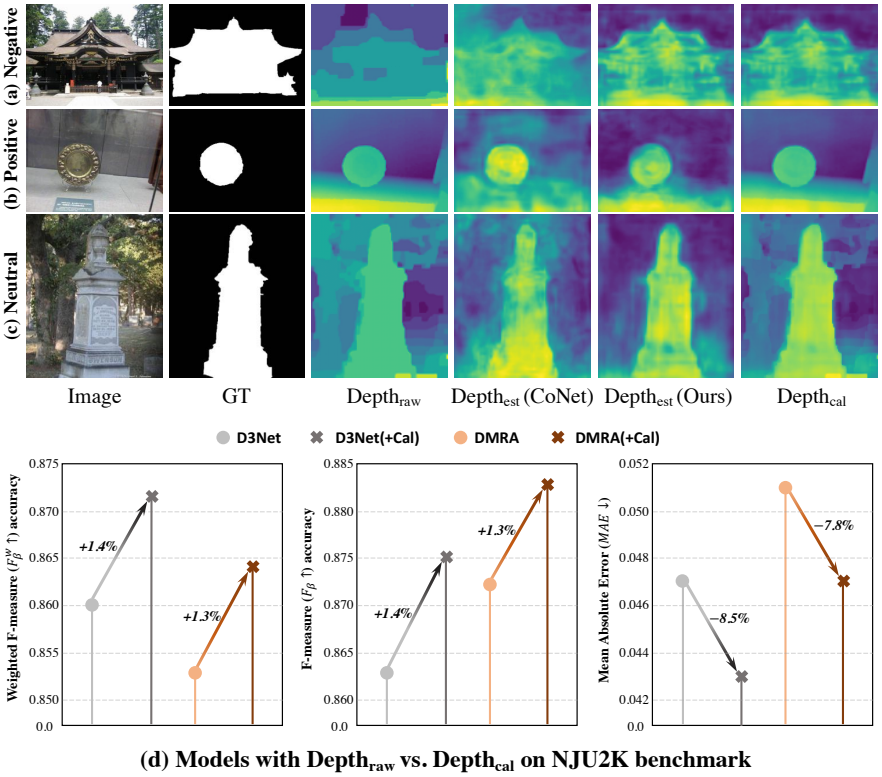
Chapter 3

Depth Calibration Strategy

3.1 Introduction

Salient object detection (SOD) excels at pinpointing and extracting regions of interest within a scene, characterized by its class-agnostic nature. This capability allows for the handling of various and arbitrary objects, making it indispensable for augmented reality (AR) applications such as object highlighting and target extraction. Due to its inherent characteristics, SOD also serves as a crucial tool for general object segmentation, often as a binary (foreground/background) segmentation task. It can be applied to a variety of downstream applications including visual tracking [138], object ranking [139], [140] and image retrieval [141], *etc*. To tackle the innate challenges in addressing difficult scenes with low texture contrast or in the presence of cluttered backgrounds, depth information has been incorporated as a complementary input source. The growing interests in the development of RGB-D SOD methods [74], [127], [142] are especially boosted by the rapid progress and flourish of varied 3D imaging sensors [143], ranging from the traditional stereo imaging that produces disparity maps, to the more recent structured lighting [27], [144], time-of-flight, light field [90], [107], [145] and LIDAR cameras that directly generate depth images. As showcased by the recent cross-modality fusion schemes [66], [131], [146], adding depth-map on top of RGB image as an extra input leads to superior performance in localizing salient objects on challenging scenes.

In essence, the actual value of depth in SOD lies in its capability of dis-



(d) Models with $\text{Depth}_{\text{raw}}$ vs. $\text{Depth}_{\text{cal}}$ on NJU2K benchmark

Figure 3.1: Top: Examples of different depth qualities; GT denotes the ground-truth saliency map; $\text{Depth}_{\text{raw}}$ denotes the original depth map; $\text{Depth}_{\text{est}}$ in the 4^{th} and 5^{th} columns are the estimated depth produced by CoNet [147] and our DCF, respectively; $\text{Depth}_{\text{cal}}$ of the last column is generated by our proposed depth calibration strategy. Bottom: Accuracy of two representative RGB-D SOD models (D3Net [34] and DMRA [91]) trained with original and calibrated depth ('+Cal'), respectively.

cerning the object silhouette from background. Nevertheless, practical examination as presented in Fig. 3.1 implies two main issues that hinder the full exploitation of depth map: 1) The depth maps are often exceedingly noisy at the object boundaries, as shown in Fig. 3.1(a), which may be hampered by the limitation of depth sensors and scene configurations such as occlusion [28], reflection [29], [30] and viewing distance [31]; 2) Even with correct depth, as exampled by Fig. 3.1(c), the foreground object often differs only slightly from the surrounding background in the depth maps. This severely limits the potential performance gain of incorporating depth maps compared to using RGB image as the sole input.

To tackle the above two challenges, a two-step depth calibration & fusion

(DCF) pipeline is developed: step one involves calibrating the depth image and correcting the latent bias in the original depth maps; step two introduces an effective cross reference module to fuse the feature representations from RGB and calibrated depth streams. Meanwhile, our depth calibration module can serve as a preprocessing step that is directly applicable to existing RGB-D SOD methods. By introducing the depth calibration module to the existing RGB-D based SOD methods, the MAE metric of D3Net [34] and DMRA [91] are decreased by 8.5% and 7.8%, respectively, when being evaluated on the widely-used NJU2K benchmark. Comprehensive experiments on public benchmarks are carried out to validate the effectiveness and generation applicability of the proposed methods.

3.2 Proposed Method

3.2.1 Method Overview

Fig. 3.2 provides an overview of the proposed DCF framework[1]. Based on a two-stream feature extraction network, it contains two core components: depth calibration and fusion strategies. As presented in Fig. 3.2, a depth calibration (DC) strategy is proposed to correct potential noise caused by unreliable raw depth maps and obtain the calibrated depth I_{depth} (or $Depth_{cal}$)[2]. As for the examples shown in Fig. 3.2, the calibrated depth can manifest the scene layout and identify foreground regions better than the original depth. Now, given the calibrated RGB-D paired data, RGB image I_{RGB} and the calibrated depth I_{depth} are fed into a two-stream feature extraction network to generate hierarchical features. For each stream, an encoder-decoder net [54] is adopted as the backbone. This is followed by a fusion strategy: cross reference modules (CRMs) are designed to integrate the valuable cues from both RGB features and depth features into the cross-modal fused features; this leads to three decoding branches that deal with RGB, depth and fused hierarchical features, respectively. Those features are separately processed and the corresponding

[1]Source code is publicly available at `https://github.com/jiwei0921/DCF`.

[2]In this thesis, symbols may have different meanings across chapters due to their context-specific use. Please interpret these symbols within the context of each respective chapter.

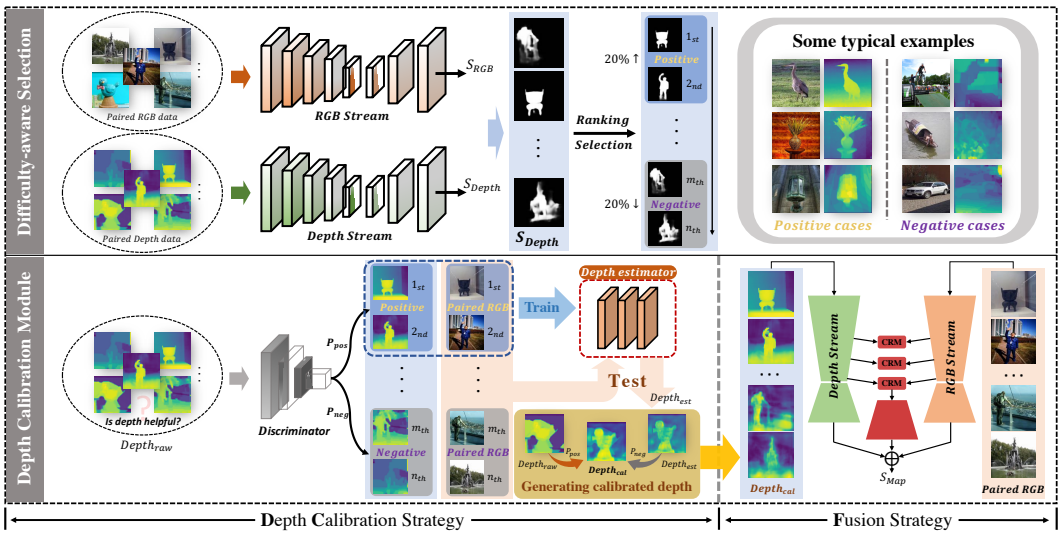


Figure 3.2: An overview of the proposed Depth Calibration and Fusion (DCF) network.

outputs are summed up to obtain the final saliency map S_{Map}.

3.2.2 Depth Calibration

Effective spatial information from depth map plays an essential role in assisting the localization of salient regions on challenging scenes such as cluttered backgrounds and low-contrast situations. However, unreliable raw depth and potential depth acquisition errors resulted by viewing distance, occlusion or reflection, will impede the model from extracting accurate information from the depth maps.

In order to tackle the performance bottleneck resulted by noisy depth maps, we attempt to calibrate the raw depth to better express the scene layout. There are two key issues that need to be addressed: 1) How can the model learn to distinguish depth maps with bad quality (negative cases) from the good quality ones (positive cases)? 2) How to produce the calibrated/refined depth maps that can both preserve helpful cues from good quality depth maps and correct unreliable information from the bad quality depth maps? Hence, we design the Depth Calibration (DC) strategy, which is the core component of our DCF, as shown in Fig. 3.2. Two sequential stages are involved to select the representative samples, and generate the calibrated depth maps.

Difficulty-aware Selection Strategy. A difficulty-aware selection strategy is proposed to solve the first key problem. As shown in Fig. 3.2, it aims to select the most typical negative and positive samples in the training database. These samples are then used to train a discriminator/classifier in predicting the quality of the depth maps, reflecting the reliabilities of depth maps.

We first pre-train two baseline models with the same architecture for RGB data and depth data individually as input under the supervision of saliency ground-truths, denoted as $\psi^{RGB}(\cdot)$, $\psi^{Depth}(\cdot)$, respectively. Then, a selection scheme is designed to measure whether a depth map is able to provide reliable information based on the saliences predicted by the two baseline models. Specifically, according to saliency results generated by the RGB stream and depth stream, we first compute the intersection over union (IoU) metric between the predicted saliency and the ground-truth saliency for the two streams, denoted as IoU_{depth} and IoU_{RGB}, respectively, for each training sample. Then, the IoU_{depth} scores for all the training samples will be sequentially sorted from large to small. Based on the ranking orders, the samples ranked top 20% of all the training samples will be regarded as typical positive set $\mathcal{P}_{set}$ (*i.e.*, the quality of depth map is acceptable) and the bottom 20% will be regarded as typical negative set $\mathcal{N}_{set}$ (*i.e.*, the quality of depth map is bad and unacceptable). In addition, when $IoU_{depth} > IoU_{RGB}$, these samples will be regarded as positive samples as well, which indicates that raw depth data provides richer global cues to identify foreground regions than RGB input. Some typical examples of both positive cases and negative cases are shown in the upper right corner of Fig. 3.2.

Depth Calibration Module. Based on the selected representative positive and negative samples, a ResNet-18 [85] based binary discriminator/classifier is trained to evaluate the reliability of the depth map. Here, the selected positive set and negative set are used for the training of the discriminator, $\{Depth_{raw}, 1\} \in \mathcal{P}_{set}$ and $\{Depth_{raw}, 0\} \in \mathcal{N}_{set}$. Our trained discriminator thus is capable of predicting a reliability score P_{pos}, indicating the probabilities of the depth map being positive or negative, respectively. The higher P_{pos} is, the better quality the original depth maps have.

In addition, a depth estimator is established, which contains several convolutional blocks using the same architecture as that of [147]. The depth estimator is trained with the RGB image and the good quality depth data pairs from the positive set, *i.e.*, $\{I_{RGB}, Depth_{raw}\} \in \mathcal{P}_{set}$, so as to mitigate the inherent noise resulted by inaccurate raw depth data. In the depth calibration module, instead of directly using the raw depth map which might be unreliable, we replace the original depth map with the weighted summation between the raw depth map and the estimated depth, and the weight is determined by the reliability probability P_{pos} predicted by the discriminator. Thereby, we obtain the calibrated depth map $Depth_{cal}$, as in:

$$Depth_{cal} = Depth_{raw} * P_{pos} + Depth_{est} * (1 - P_{pos}), \tag{3.1}$$

where $Depth_{est}$ and $Depth_{raw}$ represent the estimated depth from depth estimator and raw depth map, respectively. For better understanding, we visualize the intermediate results of the depth calibration procedure in Fig. 3.5. For the negative cases with bad quality depth, as seen in the 4^{th}-6^{th} rows in Fig. 3.5, $Depth_{cal}$ provides more reliable 3D layout information than $Depth_{raw}$. In terms of low-contrast depth data (as seen in the 4^{th} rows), our $Depth_{cal}$ can better manifest the complete scene structure compared to the original depth.

3.2.3 Cross Reference Module

After the depth calibration procedure, the calibrated depth map $Depth_{cal}$ together with the RGB image are fed to a two-stream feature extraction network to generate hierarchical features, denoted as $\{F_i^{Depth}\}_{i=3}^{5}$ and $\{F_i^{RGB}\}_{i=3}^{5}$, respectively. Note that we preserve the last three convolution blocks with plentiful semantic features and drop the first two convolutional blocks with high resolution to balance the computational cost. Generally, features extracted from the RGB channel contain rich semantic information and textural information; meanwhile, features from the depth channel contain more discriminative scene layout cues, which are complementary to that of the RGB features. In order to integrate the cross-modality information, our fusing strategy named Cross Reference Module (CRM), is designed and illustrated in Fig. 3.3.

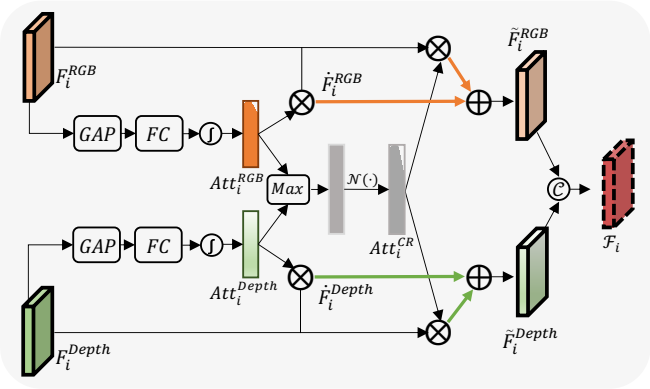


Figure 3.3: The architecture of the proposed CRM.

The proposed CRM aims to mine and combine the most discriminative channels (*i.e.*, feature detectors [110]) among depth and RGB features, and generate more informative features. More specifically, given two input features F_i^{RGB} and F_i^{Depth} produced by the i^{th} convolutional block of the RGB stream and depth stream, respectively, we first employ a global average pooling (GAP) to obtain the global statistics in the RGB and depth views. Then, the two feature vectors are separately fed into a fully connected layer (FC) and a softmax activation function $\delta(\cdot)$ to obtain the channel attention vectors Att_i^{RGB} and Att_i^{Depth}, reflecting the importance of the RGB features and depth features, respectively. The attention vectors are then applied on the input feature in a channel-wise multiplication manner. In this way, the CRM will explicitly focus on important features and suppress the unnecessary ones for scene understanding. This procedure can be defined as:

$$Att_i = \delta(\mathcal{W}_i * AvgPooling(F_i) + b_i), \tag{3.2}$$

where $\mathcal{W}_i$ and b_i represent the parameters of the FC layer for the i^{th} features, and $AvgPooling(\cdot)$ denotes the global average pooling operation. Then, the channel enhancing feature $\dot{F}_i = Att_i \otimes F_i$ is generated, where $\otimes$ denotes the channel-wise multiplication.

In addition, the attention vectors Att_i^{RGB} and Att_i^{Depth} are aggregated by the maximum function to preserve the useful feature channels from both the RGB stream and depth stream, which are then fed to the normalization operation $\mathcal{N}(\cdot)$ to normalize the output to the range from 0 to 1. And thus we

obtain the cross-referenced channel attention vector Att_i^{CR}. This procedure can be defined as:

$$Att_i^{CR} = \mathcal{N}(Max(Att_i^{RGB}, Att_i^{Depth})). \tag{3.3}$$

Based on the fusion channel attention vector Att_i^{CR}, the enhanced features $\tilde{F}_i^{RGB}$ and $\tilde{F}_i^{Depth}$ can be obtained by summing the $\dot{F}_i{}^{RGB}$ and $\dot{F}_i{}^{Depth}$ with the Att_i^{CR} enhanced features. The enhanced features from the RGB branch and depth branch are further concatenated and fed to the 1×1 convolutional layer to generate the cross-modal fused feature $\mathcal{F}_i$. The procedure can be described as:

$$\tilde{F}_i = \dot{F}_i + Att_i^{CR} \otimes F_i, \tag{3.4}$$

$$\mathcal{F}_i = Conv_{1\times1}(Concat(\tilde{F}_i{}^{RGB}, \tilde{F}_i{}^{Depth})). \tag{3.5}$$

Furthermore, a triplet loss is utilized to enhance the obtained cross-modal fused feature $\mathcal{F}_i$, so as to encourage the fused feature to be closer of the foreground, meanwhile enlarging the distance between the foreground feature and the background feature. We use $\mathcal{F}_i$ as the anchor features. Features corresponding to the saliency region are set as the positive, and features of the background region are set as the negative, as in:

$$\mathcal{F}_i^{pos} = \mathcal{F}_i \otimes \mathcal{S}, \tag{3.6}$$

$$\mathcal{F}_i^{neg} = \mathcal{F}_i \otimes (1 - \mathcal{S}), \tag{3.7}$$

where $\mathcal{S}$ represents the ground-truth saliency map.

The triplet loss $\mathcal{L}_{triplet}$ then can be calculated as:

$$\mathcal{L}_{triplet} = Max(d(\mathcal{F}_i, \mathcal{F}_i^{pos}) - d(\mathcal{F}_i, \mathcal{F}_i^{neg}) + m, 0), \tag{3.8}$$

where $d(\cdot)$ indicates the Euclidean distance; m denotes the margin parameter and is set as 1.0 following [148].

After the proposed CRM, we can obtain the cross-modal fused feature $\{\mathcal{F}_i\}_{i=3}^{5}$, which, together with the original features extracted from the RGB

stream $\{F_i^{RGB}\}_{i=3}^{5}$ and depth stream $\{F_i^{Depth}\}_{i=3}^{5}$, are further fed to three separate decoders supervised by $\mathcal{S}$. Finally, the predictions from three decoders are summed to generate the final saliency map S_{Map}.

The optimization objective $\mathcal{L}_{total}$ of the proposed method can be described as:

$$\mathcal{L}_{total} = \mathcal{L}_{RGB} + \mathcal{L}_{Depth} + \mathcal{L}_{fuse} + \frac{\alpha}{N}\sum_{i=3}^{5}\mathcal{L}_{triplet}^{i}, \tag{3.9}$$

where $\mathcal{L}_{RGB}$, $\mathcal{L}_{Depth}$ and $\mathcal{L}_{fuse}$ denote the binary cross entropy loss between the prediction of each decoder and the ground-truth saliency. $N = 3$ indicates the number of convolutional blocks involved in the triplet loss. In this paper, the hyper-parameter α is set as 0.2 empirically.

3.3 Experiments

3.3.1 Datasets

We evaluate the effectiveness of segmentation models on several public RGB-D SOD datasets. NJUD [41]: contains 1985 images in its latest version, which are collected from the Internet, 3D movies and photographs taken by a Fuji W3 stereo camera. NLPR [56]: includes 1000 images captured by Kinect. DUTLF-Depth [91]: includes 1200 paired RGB-D data captured by commercial Lytro2 camera in real life scenes. For simplicity, it is occasionally referred to as DUT-D. We follow the setup of [147], [149] to construct the training set, which consists of 1485 samples from NJUD and 700 samples from NLPR and 800 samples from DUTLF-Depth. STERE [150]: contains 1000 stereoscopic images downloaded from the Internet, which is a selected version of official STERE1000 [150]. SIP [34]: provides 929 RGB-D pairs that depicts salient people in varied scenarios. DES [151]: contains 135 images captured by Kinect. The remaining images in these datasets are all for evaluation to verify the generalization ability of saliency models. Data augmentation is also performed by randomly rotating, cropping and flipping the training images to avoid potential overfitting.

3.3.2 Evaluation Metrics

For comprehensively evaluating various saliency methods, we adopt six evaluation metrics including precision-recall (PR) curve, F-measure (F_β) [152] as well as its weighted measurement (F_β^w) [153], mean absolute error (MAE) [154] and recently proposed S-measure (S_λ) [155] and E-measure (E_γ) [156]. Concretely, the predicted saliency maps are binarized using a series of thresholds and several pairs of precision and recall are computed to plot the PR curve. F-measure is an overall performance measurement and is computed by the weighted harmonic mean of the precision and recall:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{3.10}$$

where β^2 is set to 0.3 as suggested in [152] to weight precision more than recall. MAE represents the average absolute difference between the saliency map and ground truth. It is used to calculate how similar a normalized saliency maps $\mathcal{S} \in [0,1]^{W \times H}$ is compared to the ground truth $\mathcal{G} \in \{0,1\}^{W \times H}$:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\mathcal{S}(x,y) - \mathcal{G}(x,y)|, \tag{3.11}$$

where W and H denote the width and height of $\mathcal{S}$, respectively. Structural measure (S-measure) evaluates the structural similarity between the predicted saliency maps and the binary ground truths. S-measure (denoted as S_λ) contains two terms, S_o and S_r, referring to object-aware and region-aware structural similarities, respectively:

$$S_\lambda = \lambda * S_o + (1-\lambda) * S_r \tag{3.12}$$

where λ is the balance parameter and is set to 0.5 as in [155]. Enhanced-alignment measure (E_s) considers the global means of the image and local pixel matching simultaneously.

$$E_s = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi_s(i,j), \tag{3.13}$$

where $\phi_s(\cdot)$ is the enhanced alignment matrix, which reflects the correlation between $\mathcal{S}$ and $\mathcal{G}$ after subtracting their global means, respectively. The lower the MAE, the better. For other metrics, the higher score is better.

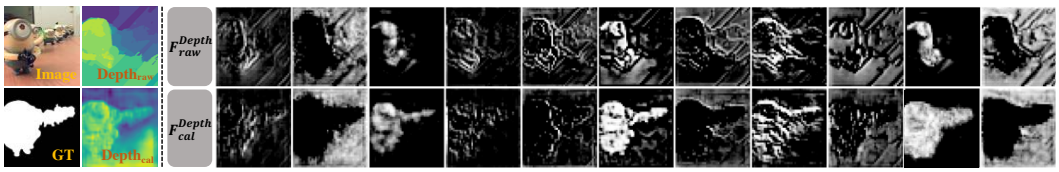


Figure 3.4: Visualization of feature representation maps in the proposed cross reference module (CRM), where F_{raw}^{Depth} and F_{cal}^{Depth} denote extracted features from backbone with raw depth and calibrated depth as input, respectively. It is observed that the calibrated depth feature maps capture richer structural information than feature maps from raw depth.

The performance of the estimated depth is evaluated with Root Mean Square Error ($RMSE$), absolute relative error ($AbsRel$), squared relative error ($SqRel$) and depth accuracy at various thresholds 1.25, 1.25^2 and 1.25^3, as suggested by [157].

3.3.3 Ablation Studies

To verify the effectiveness of the proposed modules, ablation studies are performed over each component of the DCF framework to investigate their performance gains.

RGB Stream vs. Depth Stream. Table 3.1 (a) and (b) compare the saliency prediction performance of the baseline models using RGB data as input (RGB stream) and the original depth data as input (depth stream), respectively. The RGB stream achieves better performance than that of the depth stream using original depth maps, indicating that the RGB input contains more semantic and texture information than that of the depth input. In addition, for the SIP dataset with high-quality depth maps, the performance of the depth stream is closer to that of the RGB stream, compared to other datasets with lower-quality depth maps. This again verifies the assumption that reliable depth cues can help the model to identify the salient regions better.

Effect of depth calibration strategy. To evaluate the effectiveness of the depth calibration strategy, we first compare the baseline model performance with the original depth as input (depth stream) versus that of using the cal-

Table 3.1: Quantitative comparison with different ablation settings.

Index.	Model.	**NJU2K** [41]				**NLPR** [56]				**STERE1000** [150]				**SIP** [34]			
		E_ξ	F_β^w	F_β	MAE	E_ξ	F_β^w	F_β	MAE	E_ξ	F_β^w	F_β	MAE	E_ξ	F_β^w	F_β	MAE
(a)	RGB Stream	.905	.866	.869	.046	.942	.860	.855	.028	.916	.856	.863	.047	.908	.813	.839	.063
(b)	Depth Stream	.885	.800	.831	.068	.915	.794	.800	.044	.823	.609	.695	.122	.903	.802	.845	.068
(c)	Calibrated Depth Stream	.896	.824	.840	.059	.925	.819	.821	.039	.873	.742	.778	.083	.906	.804	.852	.067
(d)	(a)+(c)+Direct fusion	.910	.867	.878	.043	.945	.862	.859	.026	.919	.863	.867	.044	.913	.822	.859	.060
(e)	(a)+(c)+CRM (w/o $\mathcal{L}_{triplet}$)	.919	.882	.890	.038	.954	.887	.885	.023	.921	.866	.877	.042	.919	.845	.869	.052
(f)	(a)+(c)+CRM (**Ours**)	**.924**	**.893**	**.902**	**.035**	**.957**	**.892**	**.891**	**.021**	**.927**	**.873**	**.885**	**.039**	**.920**	**.848**	**.875**	**.051**

ibrated depth (calibrated depth stream). As listed in Table 3.1 (b) and (c), the calibrated depth reduces the MAE metric by averagely 14.51% on four datasets. A relatively smaller performance gain is achieved on the SIP dataset compared with the rest datasets, which is reasonable since high-quality SIP has already provided reliable depth cues in the original depth maps. For better understanding, Fig. 3.1 visualizes several representative examples of the original depth map, the estimated depth as well as the final calibrated depth map. Meanwhile, as shown in Fig. 3.4, features map F_{cal}^{Depth} extracted from calibrated depth can capture scene layout information better than F_{raw}^{Depth} from raw depth (see 1^{st} vs. 2^{nd} rows).

Table 3.2: Quantitative comparison with state-of-the-art method CoNet [147] on the accuracy of the estimated depth, evaluating on two high-quality RGB-D datasets SIP [34] and DES [151]. ↑ and ↓ represent high and low scores are better, respectively.

*		$RMSE\downarrow$	$AbsRel\downarrow$	$SqRel\downarrow$	$\delta<1.25\uparrow$	$\delta<1.25^2\uparrow$	$\delta<1.25^3\uparrow$
SIP [34]	CoNet	0.4350	0.1507	0.0947	0.6713	0.9060	0.9846
	Ours	0.4289	0.1482	0.0907	0.6866	0.9168	0.9867
DES [151]	CoNet	0.6426	0.2586	0.2023	0.4364	0.7446	0.9317
	Ours	0.4794	0.1978	0.1192	0.5569	0.8764	0.9851

We have also evaluated quality of the estimated depth generated by the depth estimator on two datasets with high-quality depth maps, including the SIP dataset and the DES dataset. We follow the standard protocol of [157] to evaluate the quality of estimated depth. As listed in Table 3.2, our depth estimator achieves more accurate depth estimation, compared with CoNet [147]. Also note that our depth estimator is trained by only 20% of the training set, meanwhile CoNet was trained by 100% of the same training set, which also demonstrates the effectiveness of our difficulty-aware selection strategy.

Effect of fusion strategy. For the cross-modality fusion module to integrate the RGB and depth features, a straightforward solution is to use concatena-

tion followed by convolution operations to fuse the complementary features from RGB and depth (direct fusion). In Table 3.1, by comparing (d) and (f), we can see that the proposed CRM can better fuse the complementary information from RGB and depth features, compared with direct feature fusion. Meanwhile, compared to (f) in Table 3.1, *i.e.*, the final framework, when excluding the triplet loss from the framework, performance drop is observed on all the experimental datasets, indicating the effectiveness of the triplet loss in enhancing feature representations. In summary, quantitative and qualitative analysis showed that our DCF framework can effectively capture reliable depth information and integrate complementary cross-modal features.

Table 3.3: Quantitative comparison on five representative large-scale benchmark datasets. The best two results are shown in red and blue, respectively. * means non-deep-learning methods.

Pub.	Method	DUTLF-Depth [91]				NJUD [41]				NLPR [56]				STERE1000 [150]				SIP [34]			
		E_ξ	F_β^w	F_β	MAE	E_ξ	F_β^w	F_β	MAE	E_ξ	F_β^w	F_β	MAE	E_ξ	F_β^w	F_β	MAE	E_ξ	F_β^w	F_β	MAE
ICIMCS14	DES* [151]	.733	.386	.668	.280	.421	.241	.165	.448	.735	.259	.583	.301	.579	.281	.594	.295	.742	.352	.646	.300
SPL16	DCMC* [158]	.712	.290	.406	.243	.796	.506	.715	.167	.684	.265	.328	.196	.655	.551	.742	.148	.787	.426	.646	.186
ECCV14	LHM* [56]	.767	.350	.659	.174	.722	.311	.625	.201	.772	.320	.520	.119	.484	.379	.703	.172	.722	.286	.593	.182
CAIP17	MB* [159]	.691	.464	.577	.156	.643	.369	.492	.202	.814	.574	.637	.089	.693	.455	.572	.178	.715	.474	.573	.163
TCyb17	CTMF [23]	.884	.690	.792	.097	.864	.732	.788	.085	.869	.691	.723	.056	.841	.747	.771	.086	.824	.551	.684	.139
TIP17	DF [71]	.842	.542	.748	.145	.818	.552	.744	.151	.838	.524	.682	.099	.691	.596	.742	.141	.794	.411	.672	.186
ICCVW17	CDCP* [40]	.794	.530	.633	.159	.751	.522	.618	.181	.785	.512	.591	.114	.751	.596	.666	.149	.721	.411	.494	.224
CVPR18	PCA [36]	.858	.696	.760	.100	.896	.811	.844	.059	.916	.772	.794	.044	.887	.801	.826	.064	.898	.777	.824	.071
TIP19	TANet [160]	.866	.712	.779	.093	.893	.812	.844	.061	.916	.789	.795	.041	.893	.804	.835	.060	.893	.762	.809	.075
ICME19	PDNet [25]	.861	.650	.757	.112	.890	.798	.832	.062	.876	.659	.740	.064	.880	.799	.813	.071	.802	.503	.620	.166
PR19	MPCI [131]	.855	.636	.753	.113	.878	.749	.813	.079	.871	.688	.729	.059	.873	.757	.829	.068	.886	.726	.795	.086
CVPR19	CPFP [32]	.814	.644	.736	.099	.895	.837	.850	.053	.924	.820	.822	.036	.912	.808	.830	.051	.899	.798	.818	.064
CVPR20	JL-DCF [161]	-	-	-	-	-	-	-	-	.954	.882	.878	.022	.919	.857	.869	.040	.919	.844	.873	.051
CVPR20	S2MA [129]	-	-	-	-	-	-	-	-	.938	.852	.853	.030	.907	.825	.855	.051	.911	.825	.849	.058
CVPR20	UCNet [33]	-	-	-	-	-	-	-	-	.953	.878	.890	.025	.922	.867	.885	.039	.913	.836	.868	.051
TNNLS20	D3Net [34]	.847	.668	.756	.097	.913	.860	.863	.047	.943	.854	.857	.030	.920	.845	.855	.046	.902	.808	.835	.063
ECCV20	CMWN [66]	-	-	-	-	.910	.855	.878	.047	.940	.856	.859	.029	.917	.847	.869	.043	.906	.811	.851	.062
ECCV20	BBSNet [22]	.833	.663	.774	.120	.924	.884	.902	.035	.952	.879	.882	.023	.925	.858	.885	.041	.916	.830	.872	.055
Ours$_{NJU+NLPR}$		.890	.766	.804	.071	.924	.893	.902	.035	.957	.892	.891	.021	.927	.873	.885	.039	.920	.848	.875	.051
ICCV19	DMRA [91]	.927	.858	.883	.048	.908	.853	.872	.051	.942	.845	.855	.031	.923	.841	.876	.049	.863	.750	.819	.085
CVPR20	SSF [162]	.946	.894	.914	.034	.913	.871	.886	.043	.949	.874	.875	.026	.921	.850	.867	.046	.911	.829	.851	.056
CVPR20	A2dele [149]	.924	.864	.890	.043	.897	.851	.874	.051	.945	.867	.878	.028	.915	.855	.874	.044	.892	.793	.825	.070
ACMM20	FRDT [130]	.941	.878	.902	.039	.917	.862	.879	.048	.946	.863	.868	.029	.925	.858	.872	.042	.905	.817	.854	.063
ECCV20	DANet [163]	.925	.847	.884	.047	-	-	-	-	.949	.858	.871	.028	.914	.830	.858	.047	.916	.829	.864	.054
ECCV20	HDFNet [164]	.934	.865	.892	.040	.915	.879	.893	.038	.948	.869	.878	.027	.925	.863	.879	.040	.918	.835	.863	.051
ECCV20	CoNet [147]	.947	.896	.908	.034	.911	.856	.872	.047	.934	.850	.848	.031	.928	.874	.885	.037	.909	.814	.842	.063
ECCV20	PGAR [128]	.944	.889	.914	.035	.915	.871	.893	.042	.955	.881	.885	.024	.919	.856	.880	.041	.908	.822	.854	.055
ECCV20	ATSA [165]	.947	.901	.918	.032	.921	.883	.893	.040	.945	.867	.876	.028	.919	.866	.874	.040	.912	.848	.871	.053
Ours$_{DUT+NJU+NLPR}$		.952	.909	.926	.030	.922	.884	.897	.038	.956	.892	.893	.023	.931	.880	.890	.037	.920	.850	.877	.051

3.3.4 Comparison with State-of-the-Arts

The proposed method is evaluated and compared with 27 RGB-D segmentation methods, including 22 deep-learning-based methods and 5 non-deep-learning ones (marked with * in Table 3.3).

Quantitative Evaluation. Table 3.3 lists the quantitative comparison results. Following the main-stream training setups as that of [34] and [2], two different training settings are adopted, the results of which are independently

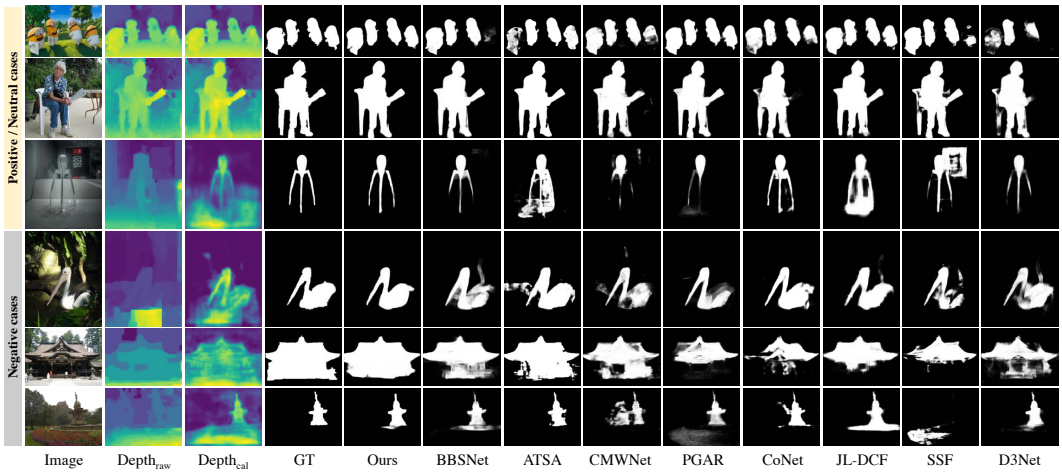


Figure 3.5: Visual comparisons of the proposed model and existing state-of-the-art algorithms.

listed in the first and second block of Table 3.3. Overall, our proposed approach achieves superior performance compared to the state-of-the-art methods with both training setups on the five commonly used SOD datasets.

Qualitative Evaluation. Fig. 3.5 shows some representative samples generated by the proposed methods and several top-ranking RGB-D approaches on several challenging cases, including the long distance, cluttered background, sharp boundary and multiple objects. As shown in the third column of Fig. 3.5, the calibrated depth ($Depth_{cal}$) can provide richer 3D layout cues than the raw depth ($Depth_{raw}$). For the challenging scenes with low-quality depth map resulted by reflection (*e.g.*, the 4^{th} row) and viewing distance (*e.g.*, the 5^{th} and 6^{th} rows), the proposed method can better identify the salient objects by taking advantage of the reliable spatial cues from the calibrated depth map $Depth_{cal}$. Therefore, both quantitative and qualitative evaluations demonstrate the effectiveness of the proposed depth calibration and fusion framework.

3.3.5 Generalization Experiments

Furthermore, to verify the generalization capability of the proposed depth calibration module, we have also applied the calibrated depth on two state-of-the-art SOD models, including D3Net [34] and DMRA [2]. As listed in Table 3.4, by replacing the original depth map with the calibrated depth to train D3Net

Table 3.4: Accuracy of the state-of-the-art RGB-D models trained with our calibrated depth vs. the raw depth. '+Cal' represents the models trained on the calibrated depth.

*	DUTLF-Depth [2]			NJUD [41]		
	F_β^w	F_β	MAE	F_β^w	F_β	MAE
D3Net [34]	0.668	0.756	0.097	0.860	0.863	0.047
D3Net(+Cal)	0.747	0.788	0.081	0.872	0.875	0.043
DMRA [91]	0.858	0.883	0.048	0.853	0.872	0.051
DMRA(+Cal)	0.875	0.899	0.043	0.864	0.883	0.047

and DMRA, noticeable performance gains have been achieved for the DUTLF-Depth dataset and NJUD dataset. The MAE metric has been decreased by 12.5% and 9.1% for D3Net and DMRA, respectively. Therefore, extensive experiments have demonstrated the advantages of the proposed depth calibration strategy.

3.4 Conclusion

In the chapter, a **D**epth **C**alibration and **F**usion (**DCF**) framework is proposed for accurate RGB-D SOD. Firstly, a depth calibration strategy is designed to correct the potential noise from unreliable raw depth. The calibrated depth has been proved to effectively improve the model performance, for both the proposed framework and state-of-the-art RGB-D saliency models. Additionally, a cross reference module is proposed to effectively integrate the complementary cues from RGB and depth features. Extensive experiments demonstrated the superior performance of our approach over 27 state-of-the-art methods.

Chapter 4

Depth-induced Multi-scale Recurrent Attention Network

4.1 Introduction

As shown in Fig. 4.1, it's evident that when facing complex scenarios like indistinguishable foreground and background, RGB-based methods [54], [55], [166], [167] fails to accurately detect complete objects. Introducing depth maps naturally addresses this challenge by providing rich spatial structure and 3D layout information in a scene, which complements conventional RGB images alone. In this project, we mainly concentrate on effectively leveraging RGB-D data to enhance model robustness, particularly in challenging scenes. It's observed that RGB-D methods outperform RGB-based methods in such complex scene where the salient object shares a similar appearance with its surroundings. However, when compared to the ground-truths, the results of existing RGB-D methods are still unsatisfactory. This may be attributed to inadequate fusion of RGB and depth data, as well as limitations in extracting powerful representations from deep networks.

In order to develop effective multimodal fusion schemes and advanced feature extraction networks to enhance segmentation robustness, there are four points to be considered. 1) Direct cross-modal fusion (*e.g.*, addition [23] or concatenation [131]) usually fails to capture the complex interactions. It is thus necessary to deeply explore the complementarity of cross-modality RGB and depth information, and exploit these useful cues for effective detection. 2)

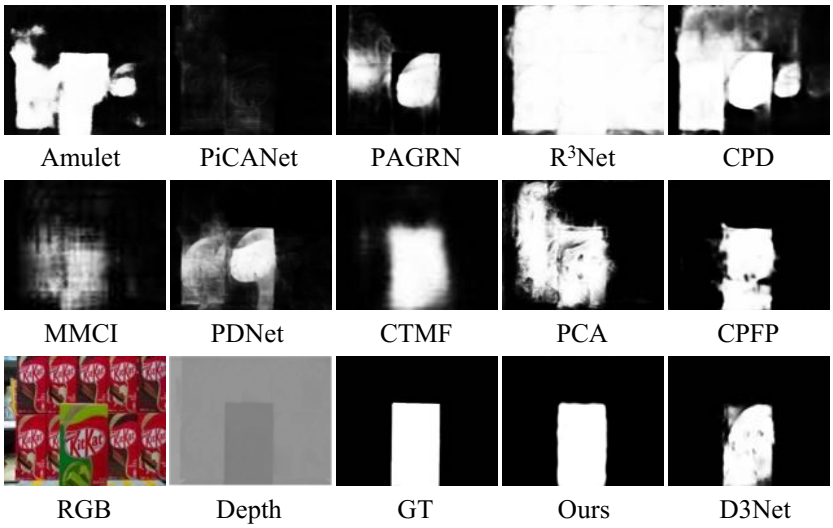


Figure 4.1: Performance of RGB-based and depth-aware SOD methods on a complex scene. RGB-based methods: Amulet [168], PiCANet [167], PAGRN [55], R^3Net [166], CPD [54]. RGB-D SOD methods: MMCI [131], PDNet [25], CTMF [23], PCA [36], CPFP [32], D3Net [34] and our proposed method.

Multiple objects in a scene have large variations in both depth and scale. Exploring the relationship between depth cues and objects with different scales can further provide vital guidance cues for obtaining informative feature representation. 3) Studies show that people perceive visual information using an Internal Generative Mechanism (IGM) [37], [38]. In the IGM, saliency captured by human is not a straight translation of the ocular input, but a result of a series of active inferences of brains, especially in complex scenes. However, the benefits of IGM for comprehensively understanding a scene and capturing accurate saliency regions have never been explored in previous works. Particularly, the fused feature is directly used for prediction while the internal semantic relation in the fused feature is ignored. 4) Deep features in the hierarchical feature representations can provide discriminative semantic information while the shallow features also contain affluent local details for accurately identifying salient objects. Designing an efficient multi-level feature fusion strategy is essential for the saliency detection task.

To this end, we propose a depth-induced multi-scale recurrent attention network for RGB-D SOD, named as DMRA. There are four main components

in our DMRA to achieve effective utilization of RGB-D data. First, we design an effective depth refinement block (DRB) taking advantages of residual connections to fully extract and fuse complementary cross-modal features in both RGB and depth views. Second, we innovatively design a depth-induced multi-scale weighting (DMSW) module. In this module, the relationship between depth information and objects with different scales is explored for the first time in saliency detection task. Ablation analysis shows that utilizing this relevance can improve detection accuracy and facilitate the integration of RGB and depth data. After the two procedures, a fused feature with abundant saliency cues is generated. Third, we design a novel recurrent attention module (RAM) inspired by the IGM of human brain. Our RAM can iteratively generate more accurate saliency results in a coarse-to-fine manner by comprehensively learning the internal semantic relation of the fused feature. Specifically, when inferring the current result, our RAM retrieves the previous memory to aid current decision. This can progressively optimize local details with memory-oriented scene understanding for generating better saliency results. Finally, a bottom-up cascaded hierarchical feature fusion strategy (CHFF) with a channel-specific contextual interaction block (CCIB) is designed to progressively integrate multi-level cross-modality features. Such efficient feature interaction enables the saliency detector to obtain more reliable predictions. Extensive ablation studies are conducted to tease apart the effectiveness and contribution of each component of DMRA.

4.2 Proposed Method

4.2.1 Method Overview

An overview of our DMRA architecture[1], based on a two-stream model as shown in Fig. 4.2, is presented below. The two streams have the same structure, where 5 convolutional blocks of VGG-19 [70] are maintained and the last pooling and fully-connected layers are discarded for making a better fit with our task. The only difference between two streams is that the depth stream

[1]Source code is publicly available at `https://github.com/jiwei0921/DMRA`.

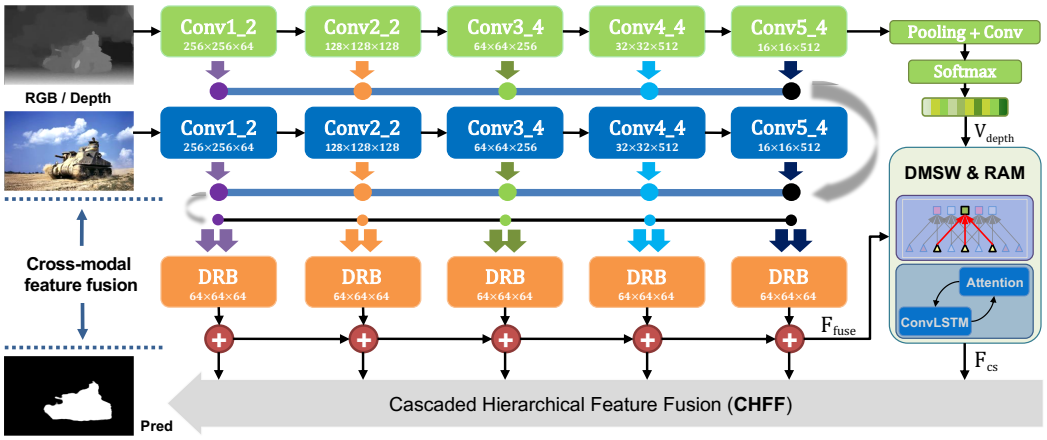


Figure 4.2: The overall architecture of our proposed DMRA, where 'DMSW' and 'RAM' represent the proposed depth-induced multi-scale weighting module and recurrent attention module, respectively. 'Pred' is the predicted saliency of the model.

is further processed to learn a depth vector. We refine and fuse paired side-out features in multiple layers by employing the proposed DRB. Then, the depth vector and the fused feature are fed into a DMSW module, in which multi-scale features generated from the fused feature are integrated based on the guidance from the depth vector. Moreover, we boost our model's performance by a novel RAM which ably combines the attention mechanism and ConvLSTM [169]. Finally, the refined feature from the RAM and multi-level cross-modal features from the two-stream network are fully integrated through our CHFF strategy. The predicted saliency maps are supervised by the ground truths. Our network is trained in an end-to-end manner.

4.2.2 Depth Refinement Block

First of all, considering the complementarity between paired depth and RGB cues in multiple layers, we design a simple yet effective DRB (Depth Refinement Block) using residual connections [85] to fully extract and fuse cross-modal paired complementary information. As illustrated in Fig. 4.3, the inputs f_i^{RGB} and f_i^{depth} represent the side-out features from the RGB and depth streams in the i-th level respectively. We feed f_i^{depth} into a series of weight layers $\Psi(\cdot)$ containing two convolutional layers and two PReLU activation

functions [170] to learn a depth residual $\Delta depth_i = \Psi(f_i^{depth})$. Then, the depth residual is added to the RGB feature by residual connection to learn a fused feature $f_i^{fuse} = f_i^{RGB} + \Delta depth_i$. In this way, complementary clues in the i-th level are fused effectively. Then, we reshape (*i.e.*, up-sample with bilinear interpolation or down-sample with max-pooling operation) f_i^{fuse} to the same resolution. A conventional residual unit [85] $\Re(\cdot)$ is followed for re-scaling feature values and then a 1×1 convolution operation W_i is used to adjust the channel dimension. The final feature in the i-th level is defined as $f_i = W_i * \Re(reshape(f_i^{fuse}))$, which is 1/4 of the input spatial resolution with 64 channels. Finally, all features f_i in multiple layers are summated as $F_{fuse} = \sum_{i=1}^{N} f_i$ in an element-wise manner, where N=5 denotes the total number of convolutional blocks. In this way, we obtain the roughly fused feature with both local spatial details and global semantic information.

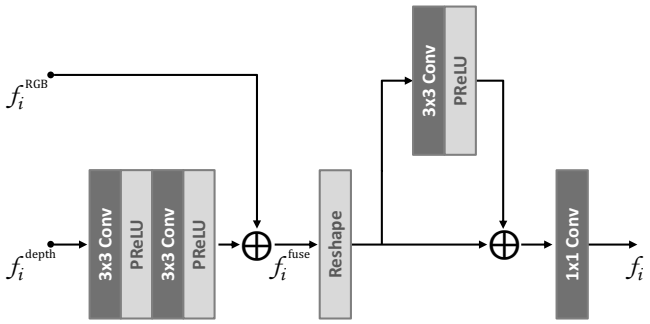


Figure 4.3: Detailed diagram of the Depth Refinement Block (DRB).

4.2.3 Depth-induced Multi-scale Weighting Module

Considering that an image consists of multiple distinct objects with different sizes, scales and laid across different spatial locations in numerous layouts, we propose a depth-induced multi-scale weighting (DMSW) module. In this module, depth cues are further connected with multi-scale features to accurately locate salient objects.

As shown in Fig. 4.4, depth cues with abundant spatial information are further processed to learn a depth vector to guide the weight allocation of multi-scale features. To be specific, in order to capture multi-scale context

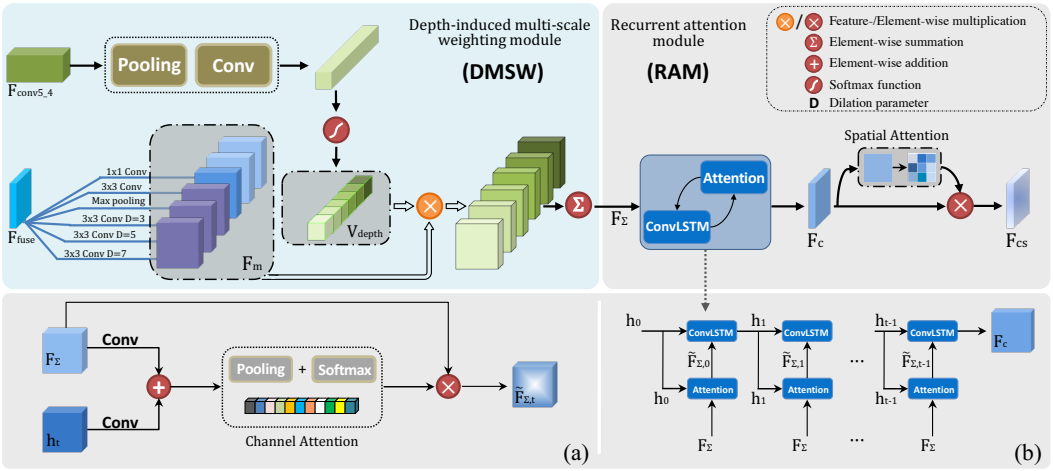


Figure 4.4: Detailed diagram of DMSW and RAM sub-modules. In RAM, (b) is the details of RAM and (a) presents its attention block.

features, we impose a global pooling layer and several parallel convolutional layers with different kernel sizes and different dilation rates on the input feature F_{fuse}. In this way, six multi-scale features F_m ($m = 1, 2, \ldots, 6$) with the same resolution but different contexts are generated. Detailed parameters are shown in Fig. 4.4. Compared with classic convolution operation, dilated convolution can increase the size of the receptive field without sacrificing image resolution and redundant computation [16], [171]. Meanwhile, in order to obtain the corresponding depth vector, a global average pooling layer and a convolutional layer are imposed on F_{conv5_4} in the depth stream. Then we use a softmax function δ to obtain the depth vector $V_{depth} \in \mathbb{R}^{1\times1\times M}$, which can act as the scale factor for weighting each multi-scale feature F_m, where M responds to the maximum of m. Finally, all multi-scale features F_m are weighted based on depth vector V_{depth} and then summated to form the final output F_Σ. Formally, the DMSW module can be defined as:

$$V_{depth} = \delta(W_b * AvgPooling(F_{conv5\_4})), \tag{4.1}$$

$$F_m = \xi(F_{fuse}; \theta_m), \tag{4.2}$$

$$F_\Sigma = \sum_{m=1}^{M} V_{depth}^m \times F_m, \tag{4.3}$$

where $*$ and W_b denote convolution operation and corresponding parameters. $\delta(\cdot)$ represents the softmax function. $\xi(\cdot)$ denotes those parallel convolution or pooling operations and θ_m is the parameters to be learned in the m-th branch. V_{depth}^m represents the weight of the corresponding multi-scale feature F_m and $\times$ means the feature-wise multiplication.

In summary, it is beneficial to introduce depth cues to learn the contribution of multi-scale features for determination of salient objects especially when objects of different sizes appear at different depths. This module can also be regarded as a deeper fusion of RGB and depth information.

4.2.4 Recurrent Attention Module

As discussed in the introduction section, it is essential to fully explore the semantic relation inside the fused feature for accurately segmenting interested objects. Thus we design a novel recurrent attention module (RAM). This module, drawing core ideas from the Internal Generative Mechanism (IGM) of human brain, can comprehensively understand a scene and learn the internal semantic relation of the fused feature. To be specific, in order to infer conspicuous objects, the IGM recurrently deduces and predicts saliency based on memory stored in the brain, while uncertain information that is not important will be discarded.

Inspired by the IGM, we propose the RAM by ably combining attention mechanism and ConvLSTM [169]. As noted in previous studies [110], [172], each channel of a feature map is considered as a "feature detector", and different "feature detectors" will capture various representation features, such as edge and content. The channel attention mechanism is acknowledged to be able to adaptively aggregate various channel-wise features. At each iteration of our RAM module, we opt to learn a channel-wise attention vector to gradually increase the representation power of features, which engages the memory information in previous iteration to help the model focus on more important features and suppress these unnecessary ones. In this way, the RAM can retrieve the previous memory to aid current decision when inferring the saliency result. That is, the RAM recurrently deduces and predicts saliency based on

memory stored in the brain, while uncertain information that is not important will be discarded.

It iteratively learns the spatio-temporal dependencies between different semantics and progressively optimizes detection details with memory-oriented scene understanding. Concretely, for the attention block (see Fig. 4.4(a)), h_t stands for the previous memory for scene understanding and F_Σ is the input feature. The subscript t denotes time step in ConvLSTM. Both h_t and F_Σ are followed by a convolutional layer and then we merge the output features by element-wise summation. Then, a global average pooling and a softmax function are used to generate the channel-wise attention map $Att_c(h_t, F_\Sigma) \in \mathbb{R}^{1\times1\times C}$, in which C denotes the number of channels of F_Σ. By performing element-wise multiplication on $Att_c(h_t, F_\Sigma)$ and F_Σ, a more informative feature $\widetilde{F}_{\Sigma,t}$ is produced. This procedure can be defined as:

$$Att_c(h_t, F_\Sigma) = \delta(AvgPooling(W_0 * h_t + W_1 * F_\Sigma)), \tag{4.4}$$

$$\widetilde{F}_{\Sigma,t} = Att_c(h_t, F_\Sigma) \otimes F_\Sigma, \tag{4.5}$$

where W_* are convolution parameters. $\otimes$ means element-wise multiplication. Next, in Fig. 4.4(b), $\widetilde{F}_{\Sigma,t}$ is fed into ConvLSTM to further learn the spatial correlation between different semantic features. The ConvLSTM is calculated by

$$\begin{aligned}
i_t &= \sigma(W_{xi} * \widetilde{F}_{\Sigma,t} + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i),\\
f_t &= \sigma(W_{xf} * \widetilde{F}_{\Sigma,t} + W_{hf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f),\\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc} * \widetilde{F}_{\Sigma,t} + W_{hc} * h_{t-1} + b_c),\\
o_t &= \sigma(W_{xo} * \widetilde{F}_{\Sigma,t} + W_{ho} * h_{t-1} + W_{co} \circ c_{t-1} + b_o),\\
h_t &= o_t \circ \tanh(c_t),
\end{aligned} \tag{4.6}$$

where $\circ$ denotes the Hadamard product and $\sigma(\cdot)$ is sigmoid function. i_t, f_t and o_t stand for input, forget and output gates, respectively. c_t stores the earlier information. All W_* and b_* are model parameters to be learned. h_0 and c_0 are initialized to 0. After N steps, where we set $N = 3$ in this work, a channel-refined feature $F_c = h_N$ is generated.

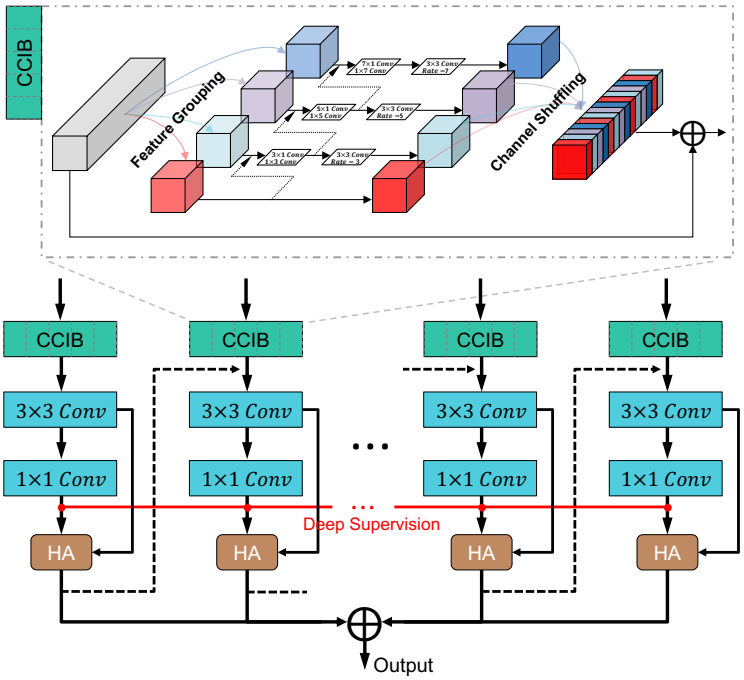


Figure 4.5: The detailed architecture of the proposed cascaded hierarchical feature fusion (CHFF) strategy as well as its key component CCIB. 'HA' stands for the holistic attention.

In addition, we add a common spatial attention block to emphasize the contribution of each pixel for the final saliency prediction. We first learn a spatial-wise attention map $Att_s(F_c) = \sigma(W_s * F_c)$, where $*$ and W_s represent a 1×1 convolution operation and corresponding parameters, respectively. Then $Att_s(F_c) \in \mathbb{R}^{W \times H \times 1}$ and F_c are multiplied in an element-wise manner to get a spatial weighted feature $F_{cs} = Att_s(F_c) \otimes F_c$.

Until now, our model has generated the refined F_{cs} as well as the cross-modal features in multiple layers with different semantic information. Next, we will elaborate on an effective feature fusion strategy to integrate features in multiple layers that further boosts detection accuracy.

4.2.5 Cascaded Hierarchical Feature Fusion Strategy

Given the refined feature F_{cs} and cross-modal features f_i, i = 1, 2, ..., 5 in multiple layers, an effective cascaded hierarchical feature fusion strategy (CHFF) equipped with a channel-specific contextual interaction block (CCIB) is de-

signed to integrate multi-level features in a progressive bottom-up manner. Detailed diagram is illustrated in Fig. 4.5. This strategy is able to promote efficient feature interaction and greatly improve the model's detection performance. For simplicity, we redefine these features as $f_i \in \mathbb{R}^{W\times H\times C}$, i = 1, 2, ..., 6, where $f_6 = F_{cs}$ in the following description.

For each f_i, we first fed it into the CCIB Block to promote channel-level information interaction within layer-wise features and simultaneously improve its scale-level contextual representation capabilities. Specifically, we propose a multi-branch structure with a 'split-transform-merge' strategy that enforces the convolutional operations to efficiently generate more informative and multifarious features. The input feature is first divided into four groups with the same channel numbers using a 1×1 convolution layer. Then we obtain $f_i = [f_i^{c1}, f_i^{c2}, f_i^{c3}, f_i^{c4}]$, where $f_i^{c_k} \in \mathbb{R}^{W\times H\times C/4}$. The feature $f_i^{c_1}$ in the first group is directly reused without further processing to retain more original details; features $f_i^{c2}, f_i^{c3}, f_i^{c4}$ in the other groups are processed by several asymmetric convolutional layers $1 \times k$ & $k \times 1$ ($k = 3, 5, 7$). Compared to the conventional convolutional layers in ResNet [85], the asymmetric convolutional layers could greatly reduce training parameters and enables deeper non-linear transformations. To further encourage hierarchical feature fusion, we progressively fuse features in multiple groups using skip-connection operations. Furthermore, to promote multi-scale feature learning, several dilated convolution layers with the same kernel size (3×3) but different dilation rates ($r = 3, 5, 7$) are also adopted here. Then those multi-scale features $\widetilde{f_i^{c_k}}(k = 1, 2, 3, 4)$ in four groups are concatenated together, followed by a channel-level shuffling operation along the channel dimension to facilitate information flow across different groups. In the end, a skip connection operation and 1×1 convolution operation are adopted to generate the final contextual fused feature F_i. The CCIB procedure could be formulated as:

$$\widetilde{f_i^{c_k}} = \begin{cases} f_i^{c_k} & k = 1; \\ K(f_i^{c_{k-1}} + f_i^{c_k}) & k = 2; \\ K(\mathcal{AS}_{Conv}(f_i^{c_{k-1}}) + f_i^{c_k}) & k = 3, 4, \end{cases} \tag{4.7}$$

$$\widetilde{f}_i = CS(Concat(\widetilde{f_i^{c_k}})), k = 1, 2, 3, 4, \tag{4.8}$$

$$F_i = Conv_{1\times1}(f_i + \widetilde{f}_i), \tag{4.9}$$

where $K(\cdot)$ denotes asymmetric convolution & dilated convolution operations. $\mathcal{AS}_{Conv}(\cdot)$ means asymmetric convolution and $CS(\cdot)$ is the channel shuffle operation.

4.3 Experiments

4.3.1 Datasets

To evaluate the performance of the proposed DMRA, we conduct experiments on five representative RGB-D SOD datasets, including DUTLF-Depth (DUT-D) [91] with 1200 RGB-D pairs, NJUD [41] and NLPR [56] containing 1985 and 1000 paired stereo images, respectively, STERE [150] with 1000 stereoscopic images downloaded from the Internet and LFSD [173] with 100 RGB-D samples. For model training, 800 samples from DUT-D, 1485 samples from NJU2K and 700 samples from NLPR are used the training set. The remaining images and other public datasets are used for testing.

4.3.2 Evaluation Metrics

Five widely-used metrics are adopted to evaluate the model performance, including S-measure (S_λ), E-measure (E_ξ) [156], weighed F-measure (F_β^w) [153], F-measure (F_β) [152] and Mean Absolute Error (MAE) [154].

4.3.3 Ablation Studies

In this section, we perform ablation analysis over each component of the proposed DMRA and further investigate their relative importance and specific contributions.

Performance of DRB. In order to verify the effectiveness of the proposed cross-modal fusion strategy, we evaluate the performance of a common fusion strategy (see Fig. 4.6 (a)) and our DRB fusion strategy (denoted as 'Baseline' and '+DRB', respectively). As shown in Table 4.1 and Fig. 4.7, '+DRB' consistently outperforms 'Baseline' across all datasets. The predictions produced

Table 4.1: Ablation analysis on five RGB-D datasets. Obviously, each component of our DMRA can provide additional accuracy gains.

Index	*	DUT-D [91]		NJUD [41]		NLPR [56]		STERE [150]		LFSD [173]	
		$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$
(a)	Baseline	0.828	0.070	0.820	0.068	0.758	0.051	0.822	0.067	0.822	0.094
(b)	+DRB	0.839	0.065	0.828	0.064	0.774	0.046	0.828	0.064	0.825	0.090
(c)	+DMSW (w/o depth)	0.855	0.061	0.844	0.062	0.805	0.044	0.837	0.061	0.836	0.087
(d)	+DMSW	0.861	0.057	0.850	0.059	0.801	0.042	0.852	0.057	0.836	0.086
(e)	+Attention (Common)	0.869	0.054	0.860	0.055	0.827	0.036	0.859	0.053	0.847	0.081
(f)	+RAM	0.883	0.048	0.872	0.051	0.855	0.031	0.868	0.047	0.849	0.075
(g)	+CCIB	0.899	0.040	0.875	0.047	0.864	0.029	0.871	0.045	0.856	0.074
(h)	+CCIB with HA	0.905	0.037	0.878	0.045	0.869	0.028	0.873	0.045	0.858	0.072
(i)	+CHFF (**DMRA**$^+$)	**0.911**	**0.035**	**0.882**	**0.044**	**0.880**	**0.026**	**0.875**	**0.043**	**0.861**	**0.069**

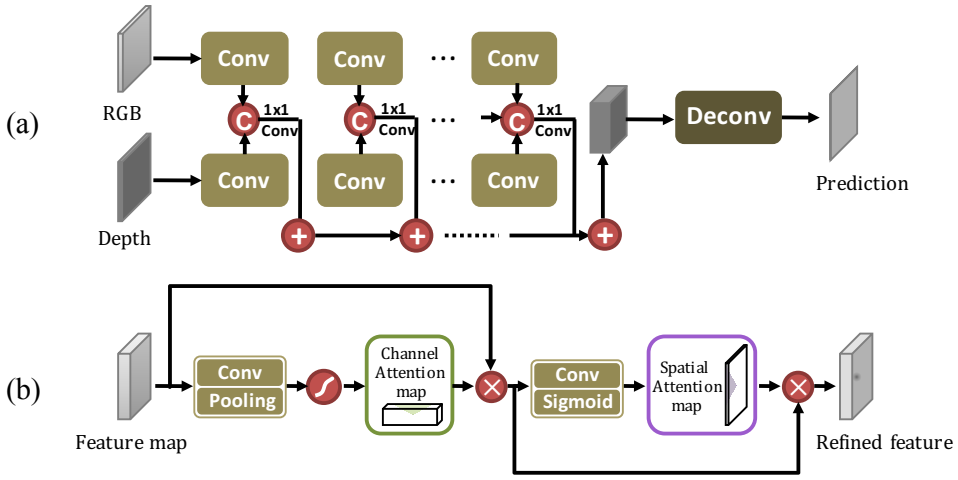


Figure 4.6: Diagrams of ablation analysis. (a) Baseline. 'C' means concatenation operation. (b) Common channel-spatial attention mechanism.

by our DRB have more complete salient regions than 'Baseline' in Fig. 4.8. This advance further confirms the superiority of our DRB in effectively and abundantly extracting and fusing cross-modal complementary information.

Performance of DMSW Module. One of our core claims is that incorporating depth cues with multi-scale features can help locate saliency regions. To give evidence for this claim, we add the DMSW module ('+DMSW') to previous '+DRB' model. Results in Table 4.1 and Fig. 4.7 show that our DMSW module achieves impressive accuracy gains on all datasets by comparing '+DMSW' and '+DRB'. From Fig. 4.8, we can see '+DMSW' can identify more saliency regions compared with '+DRB'. Those results demonstrate the advantage of our DMSW module in sufficiently utilizing depth cues and multi-scale information.

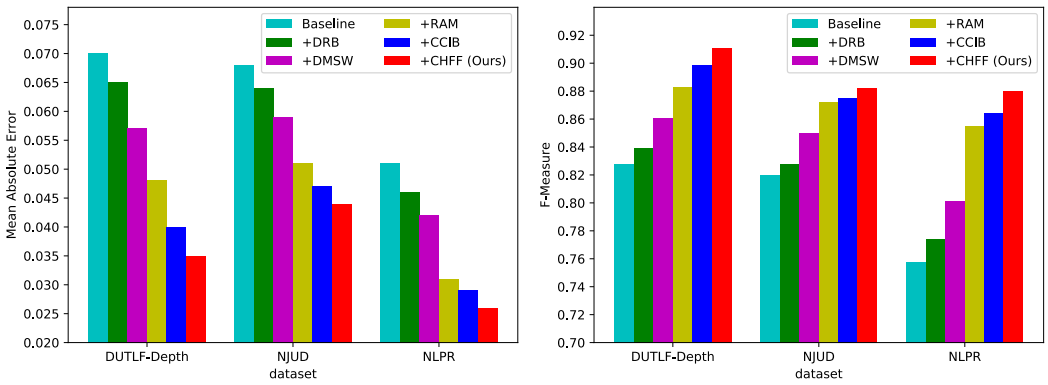


Figure 4.7: Histogram comparisons of different ablation settings in our method.

Moreover, we also verify the benefits of utilizing the relationship between depth cues and multi-scale features by performing a new experiment in Table VI (c), in which features at multiple scales are integrated by a 1×1 convolution operation instead of depth cues (denoted as '+DMSW (w/o depth)'). Experimental results show that removing depth guidance leads to degraded performance on five datasets (the average MAE score is increased by over 5%) by comparing (c) and (d). Those results further demonstrate that the combination of depth information and multi-scale features can further improve the detection accuracy.

Performance of RAM. In this section, we evaluate the performance of our RAM. By comparing visual results in Fig. 4.8, we observe our RAM can further suppress background distractions and substantially optimize detection details. In addition, we replace the RAM with a basic channel-spatial attention block [110] (denoted as '+Attention (Common)') in Fig. 4.6 (b). Results in Table 4.1 suggest that our RAM is superior to '+Attention (common)' and boosts model's performance by a large margin. We attribute this advance to its powerful ability in progressively optimizing detection details with memory-oriented scene understanding. For better understanding, we visualize the internal inspections of the RAM in Fig. 4.9. In case (a), F_{Σ} illustrates the 64-channel feature maps. Each channel represents the feature extracted by a specific "feature detector". In channel attention (CA), we demonstrate that the useful information is emphasized and the unnecessary information is

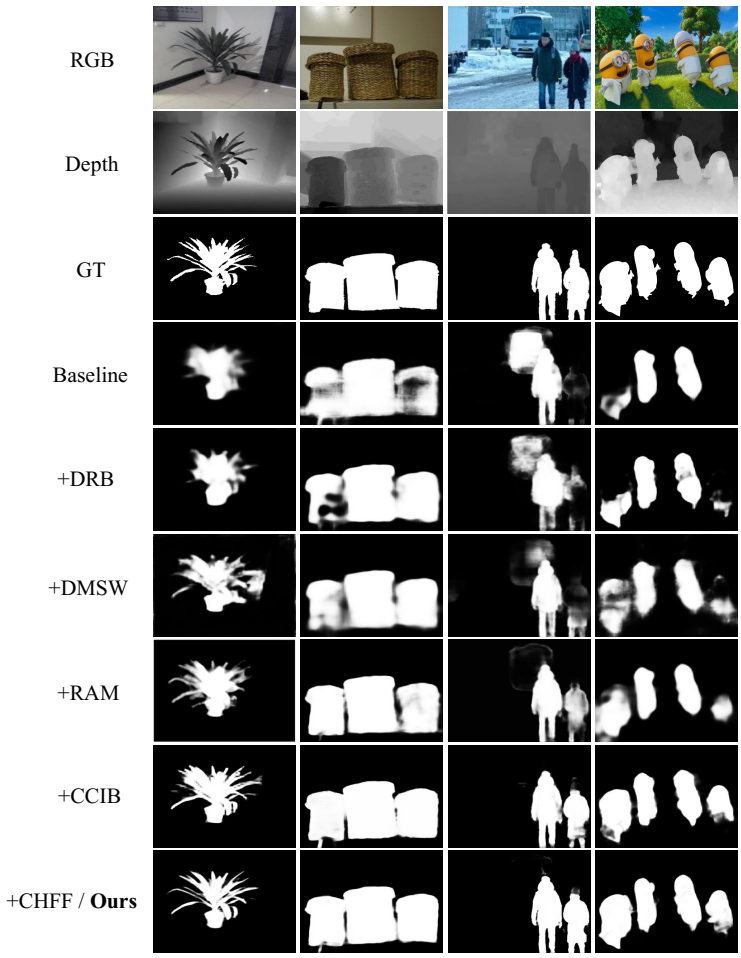


Figure 4.8: The visual results of ablation analysis. GT is ground-truth saliency.

suppressed as the channel attention weight increases for useful feature maps after three iterations, and attention weight decreases for unnecessary feature maps. For example, the 19-th channel feature map in CA is a discriminative feature that is able to better depict the content or location of salient object. Our RAM module assigns higher attention weights (0.1441) to this representative feature at the first iteration. As the iteration proceeds, the RAM module gradually improves the attention weight to as high as 0.2155, which puts more emphasize on this useful feature map. Similarly, as we can observe, the 22-th channel feature map also reveals detailed object silhouette information, which is able to help capture accurate salient object boundaries. Our RAM module also assigns gradually-increased high attention weights to that feature map, to put more emphasize on useful information. On the contrary, the 55-th channel

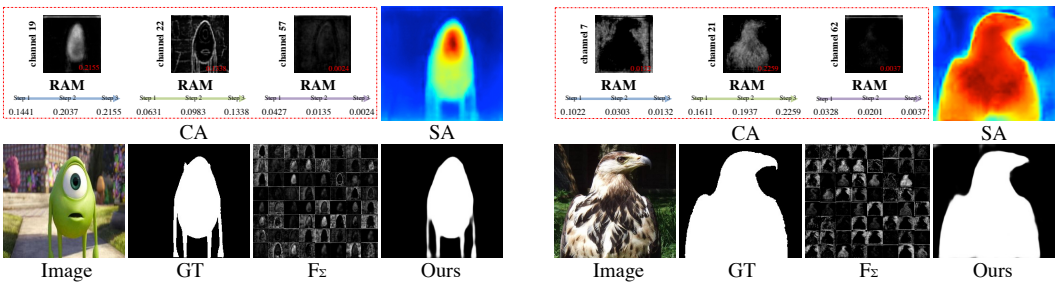


Figure 4.9: The internal inspections of the proposed recurrent attention module (RAM), where we provide two representative visualizations for better understanding.

feature map captures very limited salient object information; as a result, our RAM assigns very low attention weights to that feature, and the weights are gradually decreased by engaging our recurrent attention mechanism to make better use of memory information. In the spatial attention (SA), salient regions are also highlighted for further prediction. Similar phenomenon is also observed in case (b).

Performance of CHFF. To further validate the effectiveness of the proposed feature fusion strategy CHFF and its key component CCIB, we conduct ablated studies in Table 4.1 (g)-(i) and Fig. 4.8. When adding the CCIB to the original DMRA model, the performance is greatly improved (see (f) vs. (g)), and object details also are retrieved effectively as shown in Fig. 4.8 ('+CCIB'). This improvement is benefiting from the effective multi-scale information interaction of channel-specific features in our CCIB. Meanwhile, the cascaded hierarchical feature fusion strategy is capable of fully integrating multi-level contextual features (*i.e.*, our final model DMRA$^+$) and further brings obvious performance improvement over all datasets compared with the original DMRA model. The boundaries and local details of the saliency predictions in Fig. 4.8 are also greatly improved with the proposed CHFF. In addition, when excluding the proposed CHFF strategy (*i.e.*, using only HA and addition aggregation to fuse multi-level features), the degraded performance of model is observed by comparing (h) and (i) in Table 4.1. These results further demonstrate the reasonability and effectiveness of the proposed CHFF.

Table 4.2: Complexity analysis on each component of the proposed method.

*	FLOPs (G)	Param. (MB)	NJUD [41]		NLPR [56]	
			$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$
Baseline	55.15	43.37	0.820	0.068	0.758	0.051
+DRB	46.42	17.30	0.828	0.064	0.774	0.046
+DMSW	0.63	0.16	0.850	0.059	0.801	0.042
+RAM	22.76	1.85	0.872	0.051	0.855	0.031
+CCIB	0.47	0.12	0.875	0.047	0.864	0.029
+CHFF	0.86	0.21	0.882	0.044	0.880	0.026

Model Complexity. Table 4.2 presents the computational cost of each component in the proposed method, where the complexity is evaluated with FLOPs (*i.e.*, the number of floating-point multiplication-adds) and Param. (*i.e.*, the number of parameters involved in the module being used). For these two metrics, smaller is better. As our proposed modules are gradually incorporated into the baseline, noticeable performance gains are consistently achieved in all datasets. We observe that the RAM is a computationally intensive module that boosts the model performance by a large margin and thus is an indispensable part of the framework. Meanwhile, it is worth noting that our CHFF with CCIB achieves appealing performance improvement with fewer computational cost. This is due to our lightweight channel-level information interaction within layer-wise features and effective hierarchical fusion for multi-scale contextual learning.

4.3.4 Comparison with State-of-the-Arts

Quantitative Evaluation. Table 4.3 shows the numerical results in terms of five evaluation metrics on three widely-used datasets. It can be observed that the our model DMRA achieves superior performance compared with other methods. Especially, our model outperforms other methods obviously on DUTLF-Depth, and NJUD datasets, where the images are comparably complicated. This indicates that our model is more powerful in dealing with the complex scenes.

Qualitative Evaluation. We visually compare our method with the most representative methods as shown in Fig. 4.10. From those results, we can

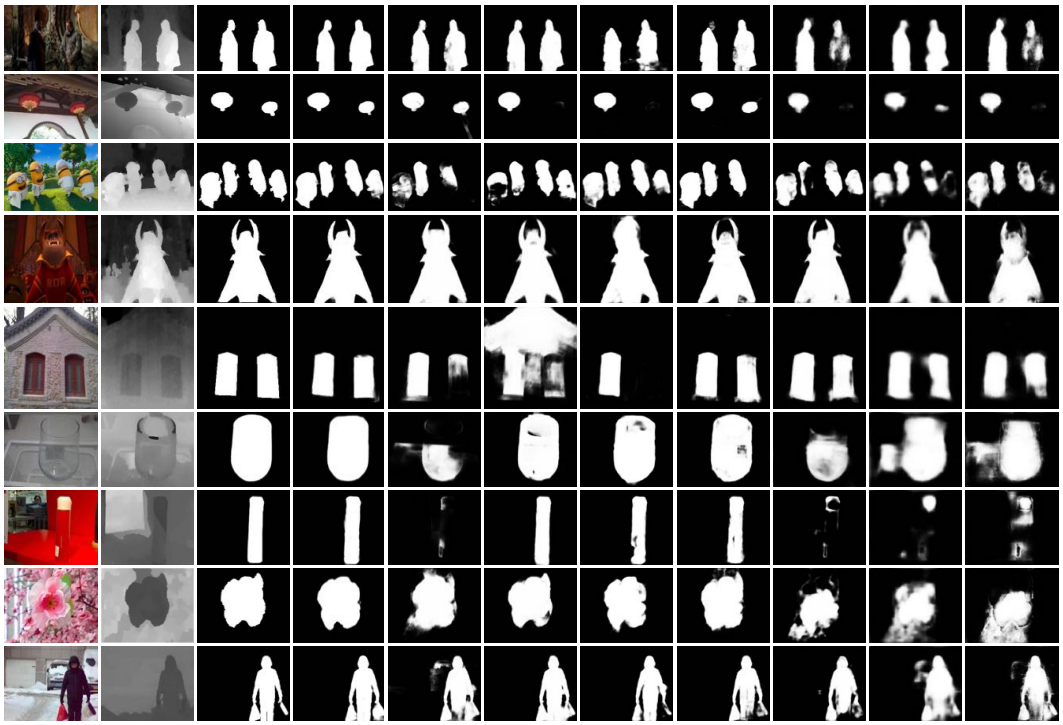


Figure 4.10: Comparisons of our DMRA with state-of-the-art RGB-D Segmentation models.

Table 4.3: Quantitative comparison on DUTLF-Depth, NJUD and NLPR saliency datasets. The best two results are shown in boldface and blue fonts respectively.

Method	DUTLF-Depth [91]					NJUD [41]					NLPR [56]				
	E_γ	S_λ	F_β^w	F_β	MAE	E_γ	S_λ	F_β^w	F_β	MAE	E_γ	S_λ	F_β^w	F_β	MAE
DES [151]	0.733	0.659	0.386	0.668	0.280	0.421	0.413	0.241	0.165	0.448	0.735	0.582	0.259	0.583	0.301
LHM [56]	0.767	0.568	0.350	0.659	0.174	0.722	0.530	0.311	0.625	0.201	0.772	0.591	0.320	0.520	0.119
DCMC [158]	0.712	0.499	0.290	0.406	0.243	0.796	0.703	0.506	0.715	0.167	0.684	0.550	0.265	0.328	0.196
MB [159]	0.691	0.607	0.464	0.577	0.156	0.643	0.534	0.369	0.492	0.202	0.814	0.714	0.574	0.637	0.089
CDCP [40]	0.794	0.687	0.530	0.633	0.159	0.751	0.673	0.522	0.618	0.181	0.785	0.724	0.512	0.591	0.114
DF [71]	0.842	0.730	0.542	0.748	0.145	0.818	0.735	0.552	0.744	0.151	0.838	0.769	0.524	0.682	0.099
CTMF [23]	0.884	0.833	0.690	0.792	0.097	0.864	0.849	0.732	0.788	0.085	0.869	0.860	0.691	0.723	0.056
PCA [36]	0.858	0.801	0.696	0.760	0.100	0.896	0.877	0.811	0.844	0.059	0.916	0.873	0.772	0.794	0.044
PDNet [25]	0.861	0.799	0.650	0.757	0.112	0.890	0.883	0.798	0.832	0.062	0.876	0.835	0.659	0.740	0.064
MMCI [131]	0.855	0.791	0.636	0.753	0.113	0.878	0.859	0.749	0.813	0.079	0.871	0.855	0.688	0.729	0.059
TANet [160]	0.866	0.808	0.712	0.779	0.093	0.893	0.878	0.812	0.844	0.061	0.916	0.886	0.789	0.795	0.041
CPFP [32]	0.814	0.749	0.644	0.736	0.099	0.895	0.878	0.837	0.850	0.053	0.924	0.888	0.820	0.822	0.036
D3Net [34]	0.847	0.775	0.668	0.756	0.097	0.913	0.900	0.860	0.863	0.047	0.943	0.912	0.854	0.857	0.030
S2MA [129]	0.921	0.903	0.868	0.886	0.043	-	-	-	-	-	0.937	0.915	0.857	0.847	0.030
A2dele [149]	0.924	0.886	0.864	0.890	0.043	0.897	0.869	0.851	0.874	0.051	0.945	0.896	0.867	0.878	0.028
DANet [163]	0.925	0.889	0.847	0.884	0.047	-	-	-	-	-	0.949	0.915	0.858	0.871	0.028
CoNet [147]	0.947	0.918	**0.896**	0.908	**0.034**	0.911	0.894	0.856	0.872	0.047	0.934	0.907	0.850	0.848	0.031
FRDT [130]	0.941	0.910	0.883	0.903	0.039	**0.917**	0.898	0.861	0.878	0.048	0.946	0.914	0.863	0.868	0.029
DMRA	**0.948**	**0.919**	0.894	**0.911**	0.035	0.914	**0.905**	**0.869**	**0.882**	**0.044**	**0.952**	**0.926**	**0.882**	**0.880**	**0.026**

observe that our results are closer to the ground truths (GT). For example, other methods are difficult to segment the whole salient regions in multiple object environments (see the 2^{nd} and 3^{rd} rows), while ours can precisely identify the whole objects. And our DMRA is able to more accurately locate and detect the entire conspicuous objects with sharp details than others in more challenging scenes such as low-contrast, transparent object situations (see the 6^{th}-8^{th} rows). Those results further verify the effectiveness and robustness of our proposed method.

4.4 Conclusion

In this work, our proposed DMRA model enhances the performance of RGB-D salient object detection from four aspects: 1) effectively extracts and fuses cross-modal complementary features by using a simple yet effective DRB; 2) innovatively combines depth cues with multi-scale information to accurately locate and identify salient objects; 3) gradually generalizes discriminative saliency features through a novel recurrent attention model; 4) progressively integrates multi-level contextual features by the CHFF feature fusion strategy. We comprehensively validate the effectiveness of each component of our network and show the contributions of these components. Extensive experiments also demonstrate that our method achieves appealing performance on nine public RGB-D saliency datasets.

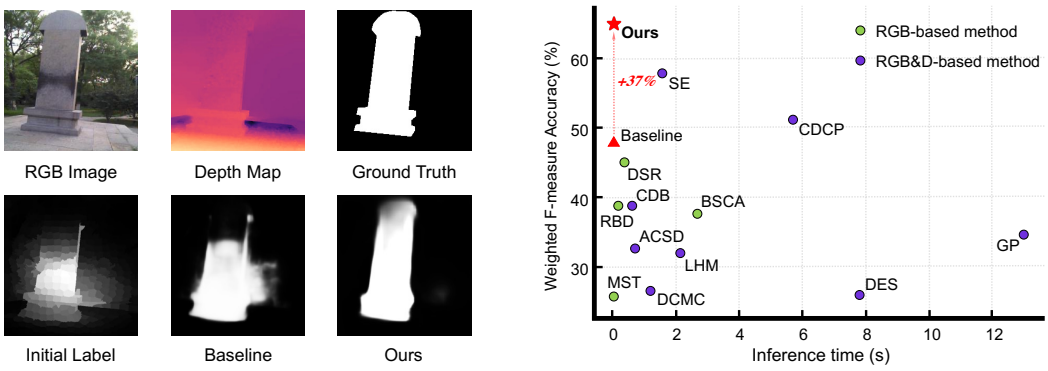
Chapter 5

Deep Unsupervised RGB-D Salient Object Detection

5.1 Introduction

The state-of-the-art RGB-D SOD approaches [67], [127], [142], [147], [174] typically entail an image-to-mask mapping pipeline that is based on the powerful deep learning paradigms of *e.g.*, VGG16 [70] or ResNet50 [85]. This strategy has led to excellent performance. On the other hand, these RGB-D SOD methods are fully supervised, thus demand a significant amount of pixel-level training annotations. This however becomes much less appealing in practical scenarios, owing to the laborious and time-consuming process in obtaining manual annotations. It is therefore natural and desirable to contemplating unsupervised alternatives. Unfortunately, existing unsupervised RGB-D SOD methods, such as global priors [39], center prior [40], and depth contrast prior [41], rely primarily on handcrafted feature representations. This is in stark contrast to the deep representations learned by their supervised SOD counterparts, which in effect imposes severe limitations on the feature representation power that may otherwise benefit greatly from the potentially abundant unlabeled RGB-D images.

These observations motivate us to explore a new problem of *deep unsupervised RGB-D saliency detection*: given an unlabeled set of RGB-D images, deep neural network is trained to predict saliency without any laborious human annotations in the training stage. A relatively straightforward idea is to



(a) Deep unsupervised RGB-D saliency detection

(b) Accuracy and inference speed comparison

Figure 5.1: (a) An illustration of deep unsupervised RGB-D saliency detection. 'Initial label' is generated by a traditional method. 'Baseline' shows the saliency map generated by saliency network trained with initial pseudo-labels. 'Ours' shows our final results. (b) Efficiency and effectiveness comparison over a wide range of unsupervised SOD methods on the NLPR benchmark.

exploit the outputs from traditional RGB-D method as pseudo-labels, which are internally employed to train the saliency prediction network ('baseline'). Moreover, the input depth map may serve as a complementary source of information in refining the pseudo-labels, as it contains cues of spatial scene layout that may help in exposing the salient objects. Nevertheless, practical examination reveals two main issues: *(1) Inconsistency and large variations in raw depth maps*: as illustrated in Fig. 5.1 (a), similar depth values are often shared by a salient object and its surrounding, making it very difficult in extracting the salient regions from depth without explicit pixel-level supervision; *(2) Noises from unreliable pseudo-labels*: unreliable pseudo-labels may inevitably bring false positive into training, resulting in severe damage in its prediction performance.

To address the above challenges, the following *two* key components are considered in our approach. First, a depth-disentangled saliency update (DSU) framework is proposed to iteratively refine & update the pseudo-labels by engaging the depth knowledge. Here a depth-disentangled network is devised to explicitly learn the discriminative saliency cues and non-salient background from raw depth map, denoted as saliency-guided depth D_{Sal} and non-saliency-guided depth D_{NonSal}, respectively. This is followed by a depth-disentangled

label update (DLU) module that takes advantage of D_{Sal} to emphasize saliency response from pseudo-label; it also utilizes D_{NonSal} to eliminate the background influence, thus facilitating more trustworthy supervision signals in training the saliency network. Note the DSU module is not engaged at test time. Therefore, at test time, our trained model takes as input only an RGB image, instead of involving both RGB and depth as input and in the follow-up computation. Second, an attentive training strategy is introduced to alleviate the issue of noisy pseudo-labels; it is achieved by re-weighting the training samples in each training batch to focus on those more reliable pseudo-labels. As demonstrated in Fig. 5.1 (b), our approach works effectively and efficiently in practice. It significantly outperforms existing unsupervised SOD methods on the widely-used NLPR benchmark. Specifically, it improves over the baseline by 37%, a significant amount without incurring extra computation cost. Besides, the test time execution of our approach is at 35 frame-per-second (FPS), the fastest among all RGB-D unsupervised methods, and on par with the most efficient RGB-based methods.

In summary, our main contributions are as follows. To our knowledge, our work is the first in exploring deep representation to tackle the problem of unsupervised RGB-D saliency detection. This is enabled by two key components in the training process, namely the DSU strategy to produce & refine pseudo-labels, and the attentive training strategy to alleviate the influence of noisy pseudo-labels. It results in a light-weight architecture that engages only RGB data at test time (*i.e.,* w/o depth map), achieving a significant improvement without extra computation cost. Empirically, our approach outperforms state-of-the-art unsupervised methods on four public benchmarks. Moreover, it runs in real time at 35 FPS, much faster than existing unsupervised RGB-D SOD methods, and at least on par with the fastest RGB counterparts. Furthermore, our approach could be adapted to work with fully-supervised scenario. As demonstrated, augmented with our proposed DSU module, the empirical results of existing RGB-D SOD models have been notably improved.

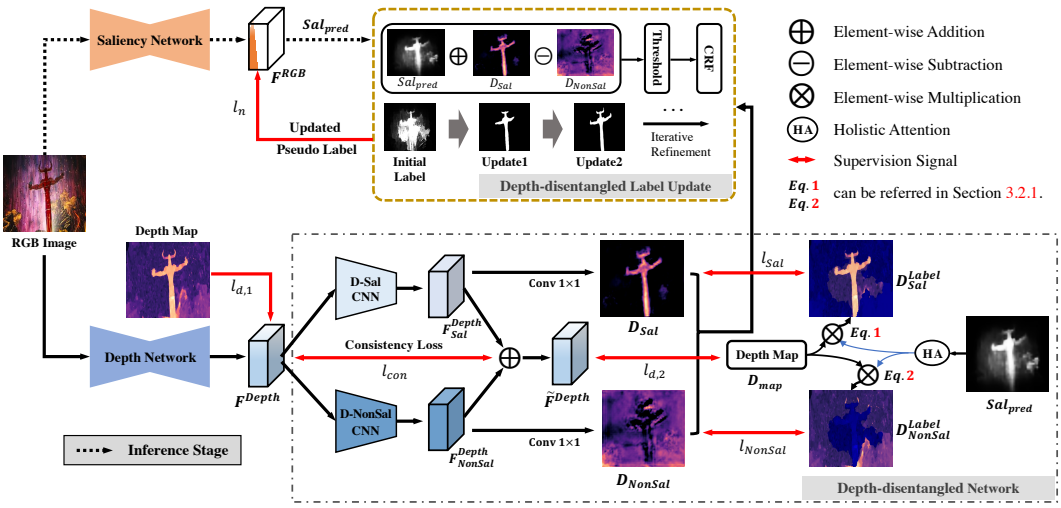


Figure 5.2: Overview of the proposed method. The saliency network is trained with the iteratively updated pseudo-labels. The depth network and depth-disentangled network are designed to decompose raw depth into saliency-guided depth D_{Sal} and non-saliency-guided depth D_{NonSal}, which are subsequently fed into the depth-disentangled label update (DLU) module to refine and update pseudo-labels. The inference stage involves only the black dashed portion.

5.2 Proposed Method

5.2.1 Method Overview

Fig. 5.2 presents an overview of our Depth-disentangled Saliency Update (DSU) framework[1]. Overall our DSU strives to significantly improve the quality of pseudo-labels, that leads to more trustworthy supervision signals for training the saliency network. It consists of three key components. First is a saliency network responsible for saliency prediction, whose initial supervision signal is provided by traditional handcrafted method without using human annotations. Second, a depth network and a depth-disentangled network are designed to decompose depth cues into saliency-guided depth D_{Sal} and non-saliency-guided depth D_{NonSal}, to explicitly depict saliency cues in the spatial layout. Third, a depth-disentangled label update (DLU) module is devised to refine and update the pseudo-labels, by engaging the learned D_{Sal} and D_{NonSal}. The updated pseudo-labels could in turn provide more trustworthy supervisions

[1]Source code is publicly available at `https://github.com/jiwei0921/DSU`.

for the saliency network. Moreover, an attentive training strategy (ATS) is incorporated when training the saliency network, tailored for noisy unsupervised learning by mitigating the ambiguities caused by noisy pseudo-labels. We note that the DSU and ATS are not performed at test time, so does not affect the inference speed. In other words, the inference stage involves only the black dashed portion of the proposed network architecture in Fig. 5.2, *i.e.*, only RGB images are used for predicting saliency, which enables very efficient detection.

5.2.2 Depth-disentangled Network

The depth-disentangled network aims at capturing valuable saliency as well as redundant non-salient cues from raw depth map. As presented in the bottom of Fig. 5.2, informative depth feature F^{Depth} is first extracted from the depth network under the supervision of raw depth map, using the mean square error (MSE) loss function, *i.e.*, $l_{d,1}$. The F^{Depth} is then decomposed into saliency-guided depth D_{Sal} and non-saliency-guided depth D_{NonSal} following two principles: 1) explicitly guiding the model to learn saliency-specific cues from depth; 2) ensuring the coherence between the disentangled and original depth features.

Specifically, in the bottom right of Fig. 5.2, we first construct the spatial supervision signals for the depth-disentangled network. Given the rough saliency prediction Sal_{pred} from the saliency network and the raw depth map D_{map}, the (non-)saliency-guided depth masks, *i.e.*, D_{Sal}^{Label} and D_{NonSal}^{Label}, can be obtained by multiplying Sal_{pred} (or $1 - Sal_{pred}$) and depth map D_{map} in a spatial attention manner. Since the predicted saliency may contain errors introduced from the inaccurate pseudo-labels, we employ a holistic attention (HA) operation [54] to smooth the coverage area of the predicted saliency, so as to effectively perceive more saliency area from depth. Formally, the (non-)saliency-guided depth masks are generated by:

$$D_{Sal}^{Label} = \Psi_{\max}(\mathcal{F}_G(Sal_{pred}, k), Sal_{pred}) \otimes D_{map}, \tag{5.1}$$

$$D_{NonSal}^{Label} = \Psi_{\max}(\mathcal{F}_G(1 - Sal_{pred}, k), 1 - Sal_{pred}) \otimes D_{map}, \tag{5.2}$$

where $\mathcal{F}_G(\cdot, k)$ represents the HA operation, which is implemented using the convolution operation with Gaussian kernel k and zero bias; the size and standard deviation of the Gaussian kernel k are initialized with 32 and 4, respectively, which are then finetuned through the training procedure; $\Psi_{\max}(\cdot, \cdot)$ is a maximum function to preserve the higher values from the Gaussian filtered map and the original map; $\otimes$ denotes pixel-wise multiplication.

Building upon the guidance of D_{Sal}^{Label} and D_{NonSal}^{Label}, F^{Depth} is fed into D-Sal CNN and D-NonSal CNN to explicitly learn valuable saliency and redundant non-salient cues from depth map, generating D_{Sal} and D_{NonSal}, respectively. The loss functions here (*i.e.*, l_{Sal} and l_{NonSal}) are MSE loss. Detailed structures of D-Sal and D-NonSal CNNs are in the appendix. To further ensure the coherence between the disentangled and original depth features, a consistency loss, l_{con}, is employed as:

$$l_{con} = \frac{1}{H \times W \times C} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{C} \| F_{i,j,k}^{Depth}, \tilde{F}_{i,j,k}^{Depth} \|_2, \tag{5.3}$$

where $\tilde{F}^{Depth}$ is the sum of the disentangled F_{Sal}^{Depth} and F_{NonSal}^{Depth}, which denotes the regenerated depth feature; H, W and C are the height, width and channel of F^{Depth} and $\tilde{F}^{Depth}$; $\|\cdot\|_2$ represents Euclidean norm. Here, $\tilde{F}^{Depth}$ is also under the supervision of depth map, using MSE loss, *i.e.*, $l_{d,2}$.

Then, the overall training objective for the depth network and the depth-disentangled network is as:

$$\mathcal{L}_{depth} = \frac{1}{5N} \sum_{n=1}^{N} (l_{d,1}^n + l_{d,2}^n + l_{Sal}^n + l_{NonSal}^n + \lambda l_{con}^n), \tag{5.4}$$

where n denotes the n_{th} sample in a mini-batch with N training samples; λ is set to 0.02 in the experiments to balance the consistency loss l_{con} and other loss terms. In the next step, the learned D_{Sal} and D_{NonSal} are fed into our Depth-disentangled Label Update, to obtain the improved pseudo-labels.

5.2.3 Depth-disentangled Label Update

To maintain more reliable supervision signals in training, a depth-disentangled label update (DLU) strategy is devised to iteratively refine & update pseudo-labels. Specifically, as shown in the upper stream of Fig. 5.2, using the obtained Sal_{pred}, D_{Sal} and D_{NonSal}, the DLU simultaneously highlights the salient regions in the coarse saliency prediction by the sum of Sal_{pred} and D_{Sal}, and suppresses the non-saliency negative responses by subtracting D_{NonSal} in a pixel-wise manner. This process can be formulated as:

$$\mathcal{S}_{temp} = Sal_{pred}^{i,j} + D_{Sal}^{i,j} - D_{NonSal}^{i,j}\Big|_{i\in[1,H];j\in[1,W]}. \tag{5.5}$$

To avoid the value overflow of the obtained $\mathcal{S}_{temp}$ ($i.e.$, removing negative numbers and normalizing the results to the range of [0, 1]), a thresholding operation and a normalization process are performed as:

$$\mathcal{S}_{\mathcal{N}} = \frac{\mathcal{S}_n^{i,j} - min(\mathcal{S}_n)}{max(\mathcal{S}_n) - min(\mathcal{S}_n)}, \text{where } \mathcal{S}_n = \begin{cases} 0, & \text{if } \mathcal{S}_{temp}^{i,j} < 0 \\ \mathcal{S}_{temp}^{i,j}, & \text{others} \end{cases}, i \in [1, H]; j \in [1, W], \tag{5.6}$$

where $min(\cdot)$ and $max(\cdot)$ denote the minimum and maximum functions. Finally, a fully-connected conditional random field (CRF [175]) is applied to $\mathcal{S}_{\mathcal{N}}$, to generate the enhanced saliency map $\mathcal{S}_{map}$ as the updated pseudo-labels.

5.2.4 Attentive Training Strategy

When training the saliency network using pseudo-labels, an attentive training strategy (ATS) is proposed to tailor for the deep unsupervised learning context, to reduce the influence of ambiguous pseudo-labels, and concentrate on the more reliable training examples. This strategy is inspired by the human learning process of understanding new knowledge, that is, from general to specific understanding cycle. The ATS alternates between two steps to re-weigh the training instances in a mini-batch.

To be specific, we first start by settling the related loss functions. For the n_{th} sample in a mini-batch with N training samples, we define the binary cross-entropy loss between the predicted saliency Sal_{pred}^{n} and the pseudo-label

$\mathcal{S}^n_{map}$ as:

$$l_n = -(\mathcal{S}^n_{map} \cdot \log Sal^n_{pred} + (1 - \mathcal{S}^n_{map}) \cdot \log(1 - Sal^n_{pred})). \tag{5.7}$$

Then, the training objective for the saliency network in current mini-batch is defined as an attentive binary cross-entropy loss $\mathcal{L}_{sal}$, which can be represented as follows:

$$\mathcal{L}_{sal} = \frac{1}{\sum_{n=1}^{N} \alpha_n} \sum_{n=1}^{N} (\alpha_n \cdot l_n), \alpha_n = \begin{cases} 1, & \text{step one,} \\ \frac{\sum_{i \in N}^{i \neq n} e^{l_i}}{\sum_{i \in N} e^{l_i}}, & \text{step two,} \end{cases} \tag{5.8}$$

where α_n represents the weight of the n_{th} training sample at current training mini-batch.

The ATS starts from a uniform weight for each training sample in step one to learn general representations across a lot of training data. Step two decreases the importance of ambiguous training instances through the imposed attentive loss; the higher the loss value, the less weight an instance is to get.

In this paper, we define step one and two as a training round (2τ epochs, $\tau = 3$). During each training round, the saliency loss $\mathcal{L}_{sal}$ and depth loss $\mathcal{L}_{depth}$ are optimized simultaneously to train their network parameters. The proposed DLU is taken at the end of each training round to update the pseudo-labels for the saliency network; meanwhile, D^{Label}_{Sal} and D^{Label}_{NonSal} in Eqs.5.1 and 5.2 are also updated using the improved Sal_{pred}.

5.3 Experiments

5.3.1 Datasets

Extensive experiments are conducted over four large-scale RGB-D segmentation benchmarks. NJUD [41] in its latest version consists of 1,985 samples, that are collected from the Internet and 3D movies; NLPR [56] has 1,000 stereo images collected with Microsoft Kinect; STERE [150] contains 1,000 pairs of binocular images downloaded from the Internet; DUTLF-Depth [2] has 1,200 real scene images captured by a Lytro2 camera. We follow the setup of [34] to construct the training set, which includes 1,485 samples from NJUD and

Table 5.1: Ablation study of our deep unsupervised RGB-D SOD pipeline, using the F-measure and MAE metrics.

Index	Model Setups		NJUD		NLPR	
			$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
(a)	Backbone		0.627	0.186	0.570	0.126
(b)	(a) + attentive training strategy		0.646	0.174	0.603	0.112
(c)	DSU strategy	(b) + CRF	0.674	0.160	0.663	0.093
(d)		(b) + DSU (w/o l_{con}&HA)	0.703	0.141	0.716	0.074
(e)		(b) + DSU (w/o l_{con})	0.712	0.137	0.735	0.068
(f)		(b) + DSU (**Ours**)	0.719	0.135	0.745	0.065

700 samples from NLPR, respectively. Data augmentation is also performed by randomly rotating, cropping and flipping the training images to avoid potential overfitting. The remaining images are reserved for testing.

5.3.2 Evaluation Metrics

Here, five widely-used evaluation metrics are adopted: E-measure (E_ξ) [156], weighed F-measure (F_β^w) [153], F-measure (F_β) [152], Mean Absolute Error (MAE or $\mathcal{M}$) [154], and inference time(s) or FPS (Frames Per Second).

5.3.3 Ablation Studies

The focus here is on the evaluation of the contributions from each of the components, and the evaluation of the obtained pseudo-labels as intermediate results.

Effect of each component. In Table 5.1, we conduct ablation study to investigate the contribution of each component. To start with, we consider the backbone (a), where the saliency network is trained with initial pseudo-labels. As our proposed ATS and DSU are gradually incorporated, increased performance has been observed on both datasets. Here we first investigate the benefits of ATS by applying it to the backbone and obtaining (b). We observe increased F-measure scores of 3% and 5.8% on NJUD and NLPR benchmarks, respectively. This clearly shows that the proposed ATS can effectively improve the utilization of reliable pseudo-labels, by re-weighting ambiguous training data. We then investigate in detail all components in our DSU strategy. The addition of the entire DSU leads to (f), which significantly improves the F-

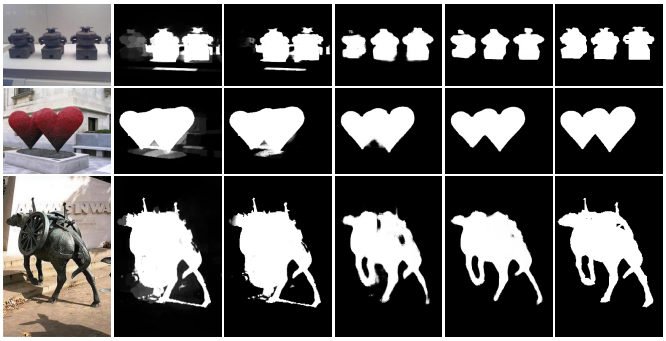


Figure 5.3: Visual examples of the intermediate pseudo-labels used in our approach. 'Initial' shows the initial pseudo-labels generated by traditional handcrafted method. '+CRF' refers to the pseudo-labels after applying fully-connected CRF. Update 1&2 represent the updated pseudo-labels produced in our pipeline over two training rounds. 'GT' means the ground truth, used for reference purpose only.

Table 5.2: Internal mean absolute errors, each is evaluated between current pseudo-labels and the corresponding true labels (only used for evaluation purpose) during the training process.

Pseudo-label Update	Initial	Update 1	Update 2	Update 3	Update 4
Mean absolute error	0.162	0.124	0.117	0.116	0.116

measure metric by 11.3% and 23.5% on each of the benchmarks, while reducing the MAE by 22.4% and 41.9%, respectively. This verifies the effectiveness of the DSU strategy to refine and update pseudo-labels. Moreover, as we gradually exclude the consistency loss l_{con} (row (e)) and HA operation (row (d)), degraded performances are observed on both datasets. For an extreme case where we remove the DLU and only maintain CRF to refine pseudo-labels, it is observed that much worse performance is achieved. These results consistently demonstrate that all components in the DSU strategy are beneficial for generating more accurate pseudo-labels.

We also display the visual evidence of the updated pseudo-labels obtained from the DSU strategy in Fig. 5.3. It is shown that the initial pseudo-labels unfortunately tend to miss important parts as well as fine-grained details. The application of CRF helps to filter away background noises, while salient parts could still be missing out. By adopting our DSU and attentive training

Table 5.3: Comparison of different pseudo-label generation variants. 'CRF' refers to fully-connected CRF. 'OTSU' represents the standard Otsu image thresholding method.

Label Accuray	$E_\xi \uparrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
Initial pseudo-label	0.760	0.526	0.614	0.162
Initial pseudo-label + CRF	0.763	0.578	0.634	0.144
Our DSU	**0.792**	**0.635**	**0.708**	**0.116**
Depth map	0.419	0.284	0.164	0.414
Depth map + OTSU	0.465	0.398	0.429	0.332

strategy, the missing parts could be retrieved in the updated pseudo-labels, with the object silhouette also being refined. These numerical and visual results consistently verify the effectiveness of our pipeline in deep unsupervised RGB-D saliency detection.

Analysis of pseudo-labels. We analyze the quality of pseudo-labels over the training process in Table 5.2, where the mean absolute error scores between the pseudo-labels at different update rounds and the ground-truth labels are reported. It is observed that the quality of pseudo-labels is significantly improved during the first two rounds, which then remains stable in the consecutive rounds. Fig. 5.3 also shows that the initial pseudo-label is effectively refined, where the updated pseudo-label is close to the true label. This provides more reliable guiding signals for training the saliency network.

In Table 5.3, we investigate other possible pseudo-label generation variants, including, initial pseudo-label with CRF refinement, raw depth map, and raw depth map together with OTSU thresholding [176]. It is shown that, compared with the direct use of CRF, our proposed DSU is able to provide more reliable pseudo-labels, by disentangling depth to promote saliency. It is worth noting that a direct application of the raw depth map or together with an OTSU adaptive thresholding of the depth map, may nevertheless lead to awful results. We conjecture this is because of the large variations embedded in raw depth, and the fact that foreground objects may be affected by nearby background stuffs that are close in depth.

Table 5.4: Quantitative comparison with unsupervised SOD (salient object detection) methods. 'Backbone' refers to the saliency feature extraction network [54] adopted in our pipeline, *i.e.*, the one without the two proposed key components. The RGB-based methods are specifically marked by †. UnSOD is shorthand for unsupervised SOD.

	*	Inference Time(s)↓	NJUD				NLPR				STERE				DUTLF-Depth			
			$E_\xi\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$
Handcrafted UnSOD	RBD† [177]	0.189	.684	.387	.556	.256	.765	.388	.590	.211	.730	.443	.610	.223	.733	.447	.619	.222
	MST† [178]	0.030	.670	.291	.436	.281	.762	.257	.491	.199	.681	.312	.447	.269	.678	.254	.401	.279
	BSCA† [179]	2.665	.756	.446	.623	.216	.745	.376	.554	.178	.803	.497	.676	.179	.808	.479	.682	.181
	DSR† [11]	0.376	.739	.436	.594	.196	.757	.451	.545	.120	.785	.486	.645	.165	.797	.478	.640	.164
	ACSD [41]	0.718	.790	.448	.696	.198	.751	.327	.547	.171	.793	.425	.661	.200	.250	.210	.188	.668
	DES [151]	7.790	.421	.241	.165	.448	.735	.259	.583	.301	.673	.383	.592	.297	.733	.386	.668	.280
	LHM [56]	2.130	.722	.311	.625	.201	.772	.320	.520	.119	.772	.360	.703	.171	.767	.350	.659	.174
	GP [39]	12.98	.730	.323	.666	.204	.813	.347	.670	.144	.785	.371	.710	.182	-	-	-	-
	CDB [180]	0.600	.752	.408	.650	.200	.810	.388	.618	.108	.808	.436	.713	.166	-	-	-	-
	SE [181]	1.570	.780	.518	.735	.164	.853	.578	.701	.085	.825	.546	.747	.143	.730	.339	.474	.196
	DCMC [158]	1.210	.796	.506	.715	.167	.684	.265	.328	.196	.832	.529	.743	.148	.712	.290	.406	.243
	MB [159]	-	.643	.369	.492	.202	.814	.574	.637	.089	.693	.455	.572	.178	.691	.464	.577	.156
	CDCP [40]	5.720	.751	.522	.618	.181	.785	.512	.591	.114	.797	.596	.666	.149	.794	.530	.633	.159
Deep UnSOD	USD† [80]	0.0180	.768	.565	.630	.163	.786	.536	.580	.119	.796	.572	.670	.146	.795	.545	.650	.157
	DeepUSPS† [81]	0.0292	.771	.576	.647	.159	.809	.622	.639	.088	.806	.632	.682	.124	.798	.573	.654	.149
	Backbone	0.0286	.759	.510	.627	.186	.760	.479	.570	.126	.794	.555	.666	.158	.798	.512	.644	.167
	Δ gains	-	↑5%	↑17%	↑15%	↓27%	↑16%	↑37%	↑31%	↓48%	↑8%	↑22%	↑16%	↓37%	↑7%	↑27%	↑18%	↓36%
	Ours	0.0286	**.797**	**.597**	**.719**	**.135**	**.879**	**.657**	**.745**	**.065**	**.857**	**.678**	**.774**	**.099**	**.854**	**.650**	**.763**	**.107**

5.3.4 Comparison with State-of-the-Arts

Our approach is compared with 15 unsupervised SOD methods, *i.e.*, without using any human annotations. Their results are either directly furnished by the authors of the respective papers, or generated by re-running their original implementations. In this paper, we make the first attempt to address deep-learning-based unsupervised RGB-D SOD. Since existing unsupervised RGB-D methods are all based on handcrafted feature representations, we additionally provide several RGB-based methods (*e.g.*, USD and DeepUSPS) for reference purpose only. This gives more observational evidences for the related works. These RGB-based methods are specifically marked by † in Table 5.4.

Quantitative results are listed in Table 5.4, where our approach clearly outperforms the state-of-the-art unsupervised SOD methods in both RGB-D and RGB only scenarios. This is due to our DSU framework that leads to trustworthy supervision signals for saliency network. Furthermore, our network design leads to a light-weight architecture in the inference stage, shown as the black dashed portion in Fig. 5.2. This enables efficient & effective detection of salient objects and brings a large-margin improvement over the backbone

network without introducing additional depth input and computational costs, as shown in Table 5.4. Qualitatively, saliency predictions of competing methods are exhibited in Fig. 5.4. These results consistently proves the superiority of our method.

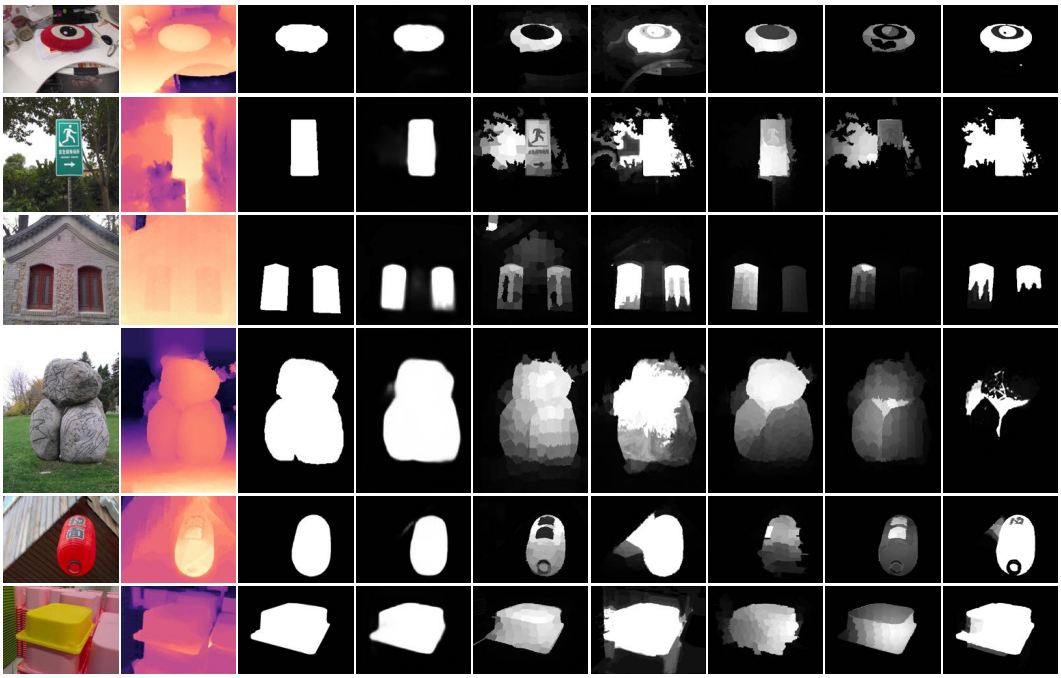


Figure 5.4: Qualitative comparison with unsupervised saliency detection methods. GT denotes ground-truth for reference.

5.3.5 Generalization Experiments

Application to Fully-supervised Setting. To show the generic applicability of our approach, a variant of our DSU is applied to several cutting-edge fully-supervised SOD models to improve their performance. This is made possible by redefining the quantities $D_{Sal}^{Label} = \mathcal{S}_{GT} \otimes D_{map}$ and $D_{NonSal}^{Label} = (1-\mathcal{S}_{GT}) \otimes D_{map}$, with $\mathcal{S}_{GT}$ being the ground-truth saliency. Then the saliency network (*i.e.*, existing SOD models) and the depth-disentangled network are retrained by $\mathcal{S}_{GT}$ and the new D_{Sal}^{Label} and D_{NonSal}^{Label}, respectively. After training, the proposed DLU is engaged to obtain the final improved saliency. In Table 5.5, we report the original results of four SOD methods and the new results

of incorporating our DSU strategy on two popular benchmarks. It is observed that our supervised variants have consistent performance improvement comparing to each of existing models. For example, the average MAE score of four SOD methods on NJUD benchmark is reduced by 18.0%. We attribute the performance improvement to our DSU strategy that can exploit the learned D_{Sal} to facilitate the localization of salient object regions in a scene, as well as suppress the redundant background noises by subtracting D_{NonSal}.

Table 5.5: Applying our DSU to existing fully-supervised RGB-D SOD methods.

*	**NJUD**		**NLPR**	
	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
DMRA [91]	0.872	0.051	0.855	0.031
+ **Our DSU**	0.893	0.044	0.879	0.026
CMWN [66]	0.878	0.047	0.859	0.029
+ **Our DSU**	0.901	0.041	0.882	0.025
FRDT [130]	0.879	0.048	0.868	0.029
+ **Our DSU**	0.903	0.038	0.901	0.023
CPD [54]	0.873	0.045	0.866	0.028
+ **Our DSU**	0.909	0.036	0.907	0.022

5.4 Conclusion

This paper tackles the new task of deep unsupervised RGB-D salient object segmentation. Our key insight is to internally engage and refine the pseudo-labels. This is realized by two key modules, the depth-disentangled saliency update in iteratively fine-tuning the pseudo-labels, and the attentive training strategy in addressing the issue of noisy pseudo-labels. Extensive empirical experiments demonstrate the superior performance and realtime efficiency of our approach. Furthermore, to demonstrate the applicability of our idea in more general settings, a supervised variant of our approach is also evaluated with superior performance in the supervised scenario of object segmentation.

Part II

RGB-Thermal Semantic Segmentation

In low-light conditions or when confronted with intense oncoming headlights, distinguishing pedestrians and vehicles can be exceptionally difficult, even for human vision. In contrast, thermal infrared imaging becomes invaluable in such situations, providing a crucial supplement to traditional RGB images by detecting infrared radiation emitted by objects above absolute zero.

This advantage has sparked increased interest in RGB-Thermal Semantic Segmentation, also known as Multispectral Semantic Segmentation (MSS), which utilizes a combination of RGB and thermal (RGB-T) imagery to overcome the limitations faced by conventional RGB models in poor lighting conditions. For instance, in autonomous driving applications [9], this task enables the perception of pedestrians and moving cars based on their heat signatures, complementing the limited appearance cues provided by RGB imagery alone under adverse lighting conditions. Despite its significance, this field remains nascent, primarily due to the lack of extensive datasets. In this part, we aim to support the advancement of RGB-Thermal Semantic Segmentation, addressing both still images and video content in Chapter 6 and Chapter 7, respectively.

This part is based on the following publications:

- **Chapter 6**: [4] W. Ji, J. Li, C. Bian, Z. Zhang, L. Cheng. "SemanticRT: A large-scale dataset and method for robust semantic segmentation in multispectral images". In Proceedings of the 31st ACM International Conference on Multimedia (ACM MM), 2023.

- **Chapter 7**: [5] W. Ji, J. Li, C. Bian, Z. Zhou, J. Zhao, A. Yuille, L. Cheng. "Multispectral video semantic segmentation: a benchmark dataset and baseline". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

Chapter 6

RGB-Thermal Image Semantic Segmentation

6.1 Introduction

The problem of semantic segmentation concerns associating each image pixel with a predefined class label. Achieving robust and reliable semantic segmentation in various lighting conditions is crucial for many real-life applications such as autonomous safe driving and nighttime rescue [182]. In Part I, we have explored the utility of depth maps in aiding the differentiation of target objects from cluttered backgrounds. However, despite notable progress, their effectiveness is predominantly observed in favorable weather conditions. When confronted with adverse scenarios such low-light, or complete darkness, it is difficult for depth sensors to capture accurate spatial information due to unreliable light reflection from object surfaces. In this work, our focus shifts to the exploration of emerging thermal infrared imaging that is able to capture infrared radiation emitted from any object with a temperature above absolute zero [42]. As showcased by the two exemplar RGB images in Fig. 6.1 (a), it could be exceedingly challenging even for human eye to discern the pedestrians highlighted by the green boxes, when in a scenario of low-light night or facing a strong coming headlight. With the limited appearance cues, RGB-based methods often struggle to detect objects in their entirety, as presented in Fig. 6.1 (d). On the contrary, thermal and RGB images together result in a more accurate segmentation in such low-light circumstances. This has natu-

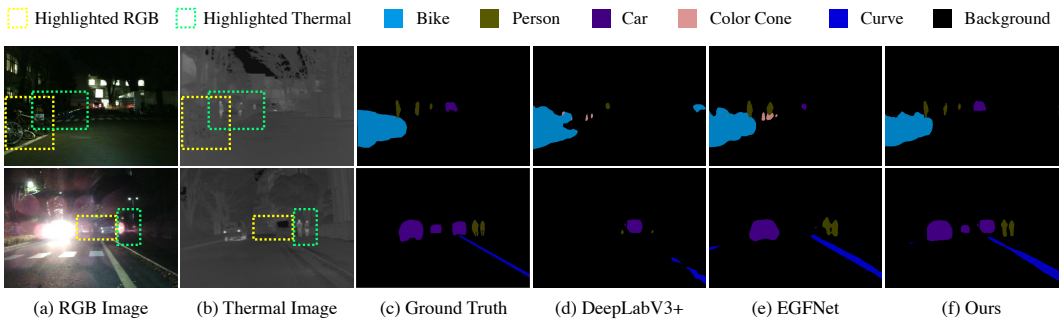


Figure 6.1: Two multispectral semantic segmentation examples under adverse illumination conditions. The complementary nature of RGB and thermal images are highlighted using yellow and green boxes, respectively. The RGB-only method, DeepLabV3+ [15], is susceptible to incorrect segmentation or even missing target objects entirely. In contrast, multispectral segmentation methods, *e.g.*, EGFNet [51] and our proposed method, which incorporate thermal infrared information, is able to effectively identify the segments. Furthermore, our results are visually better aligned to the ground truths than the state-of-the-art EGFNet [51].

rally led to a growing interest in multispectral semantic segmentation (MSS), where a pair of RGB and thermal (RGB-T) images is used as an input. This line of research has seen a range of real-world applications, from autonomous safe driving [9], night patrol [43], and fire rescue [44], to object tracking [45].

Indeed, this field is still in its early stages of development, primarily due to the absence of large-scale benchmarks. Table 6.1 outlines benchmark datasets in the MSS field. The pioneering benchmark, MFNet [42], offers 1,569 RGB-T images, accompanied with pixel-wise annotations that support the training and evaluation of MSS models. Another dataset, PST900 [46], contains 894 pairs of images captured in underground tunnels and caves, which could serve to validate the generalization capabilities of MSS models. Contrasting to the RGB-based semantic segmentation datasets such as Cityscapes [8] and PASCAL-Context [47] that contain 5,000-10,103 finely annotated images, for the MSS community, existing benchmark datasets are considerably smaller – the largest benchmark containing only 1,569 images. This has imposed a severe limit toward developing better MSS models. Additionally, existing datasets typically lack diversity in scene contents & categories, and feature low image resolutions. This may impede practical development in the MSS field.

Table 6.1: An comparison of our SemanticRT dataset with existing MSS benchmark datasets.

Dataset	Year	#Numbers	# Semantic Classes	Resolution (> 95%)
MFNet [42]	IROS'17	1,569	9	640×480
PST900 [46]	ICRA'20	894	5	1280×720
SemanticRT	-	11,371	13	1280×1024

To tackle these challenges, we curate in this work a large-scale SemanticRT dataset, comprising 11,371 RGB and thermal infrared image pairs, accompanied with high-quality, pixel-wise annotations over 13 categories. It also covers a diverse range of scenarios (*e.g.*, road, park, campus, street) in both daytime and nighttime settings. The majority (over 95%) of these RGB-T image pairs are of high-resolution (1280×1024).

With access to these rich multispectral cues, existing MSS methods have developed plausible solutions to unify the two types of information, by concatenating or summing multimodal features from separate encoders [42], [48], [49], direct incorporation of thermal images as an additional input channel [46], or weighted attention fusions [50], [51]. It is observed in Fig. 6.1 (d) *vs.* (e) that the incorporation of thermal infrared cues indeed leads to enhanced performance in low light scenes. However, when compared to the ground-truths, the results are still unsatisfactory with these *implicit* fusion strategies. The key challenges stem from two aspects. First, existing efforts often indiscriminately aggregate two modal cues extracted from individual feature extractor, which could bring an overemphasis on shared high-intensity information, and eventually dilute the useful modality-specific cues. This may weaken their discriminative power in scene representations. Second, existing methods tend to overlook the essential variations between RGB and thermal images caused by inherent imaging differences, as highlighted in the boxed regions of Fig. 6.1 (a)&(b). Hence, how to explicitly identify the innate multi-modal complementary characteristics and model their specificity is still an open problem.

This motivates us to propose a new *explicit complement modeling* (ECM) framework for the MSS task. Our approach features a cross-referenced discrepancy learning structure to explicitly extract inter-modality *complements* cues,

and a dedicated scheme to sufficiently incorporate them along the feature fusion and encoding processes. These two processes are termed as complements-aware feature fusion (CA-Fuse) and complements-aided feature encoding (CA-Encode), respectively. The promising results on benchmark datasets clearly demonstrate its effectiveness.

There are two main contributions in this work:

- A large-scale pixel-wise MSS dataset, SemanticRT, is constructed. It covers a wide range of real-life urban scenarios under diverse lighting conditions. To our best knowledge, it is the largest MSS dataset to date (7× larger than MFNet), and is expected to facilitate benchmarking & robust training of new MSS algorithms.

- An explicit complement modeling (ECM) scheme is developed to better exploit the complementary characteristics of both the RGB and thermal modalities. Extensive experiments are carried out with three benchmark datasets. Empirical results demonstrate the superiority of our approach.

6.2 Proposed SemanticRT Dataset

In this section, we describe the construction process of SemanticRT dataset[1], and analyze its statistical results.

6.2.1 Dataset Collection and Annotation

Our goal is to collect a large-scale dateset featuring high-quality RGB-thermal images that cover a wide diversity of scenarios. To this end, we initially gather over 16,000 RGB-T images pairs from multiple sources, including OSU [183], INO [184] and LLVIP [185]. The collected image samples span a variety of locations (*e.g.*, parks, campus, and streets), and are captured under a wide range of illumination conditions (*e.g.*, daytime, nighttime, darkness, dim light, and reflective environments.). To ensure the quality of our dataset, we meticulously remove unqualified samples, including blurred images, similar images,

[1]Dataset is publicly available at `https://github.com/jiwei0921/SemanticRT`.

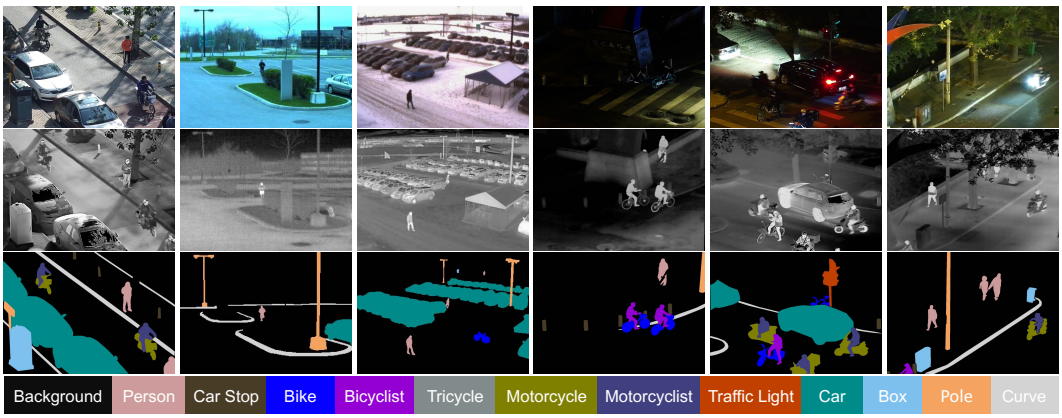
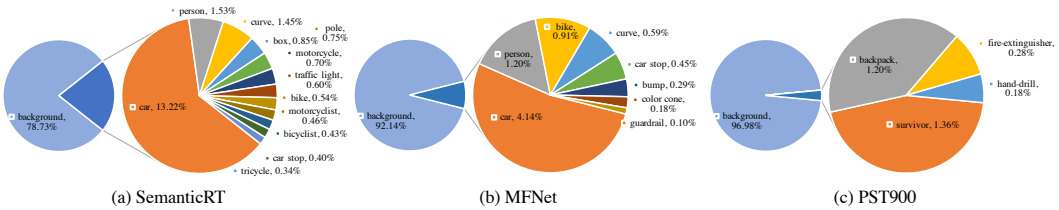


Figure 6.2: Exemplar images and annotations from the proposed SemanticRT dataset. Rows 1-3: RGB images, corresponding thermal images, and their pixel-wise semantic annotations. Left: daytime scenarios; Right: nighttime scenarios.

and unaligned images. After this selection process, we finalize the Semamti-cRT dataset, which comprises 11,371 high-quality RGB-T image pairs. Several visual examples are illustrated in Fig. 6.2.

Creating ground-truth annotations for RGB-T semantic segmentation is difficult. Unlike annotating RGB images, annotating a large-scale MSS dataset poses additional challenges, as many complex scenes make it difficult to accurately identify less visible semantic regions, even for human annotators. To address these challenges, we establish a professional team dedicated to producing high-quality annotations. Our annotation consists of three main steps. First, a team of experts meticulously reviews all images, ultimately identifying 13 semantic categories: car stop, bike, bicyclist, motorcycle, motorcyclist, car, tricycle, traffic light, box, pole, curve, person, and background (*i.e.*, unlabeled pixels). The selection criteria for these categories are based on their frequency, relevance to applications, and compatibility with existing datasets. Second, we train six annotators to recognize and annotate various semantic objects. To assist annotators, we display both RGB and thermal images side-by-side on a single screen, synchronizing the annotation traces on both images. This aids human annotators when dealing with challenging images, such as those captured in dim light or darkness. Third, two additional inspectors carefully examine the initial annotations on an item-by-item basis, identifying any



(a) SemanticRT (b) MFNet (c) PST900

Figure 6.3: Comparative pie chart visualizations: distribution of semantic categories for pixel-Level annotations in the newly-proposed SemanticRT dataset, MFNet dataset [42], and PST900 dataset [46]. Our SemanticRT showcases a more diverse range of semantic categories and a higher proportion of annotated pixels compared to the other two datasets.

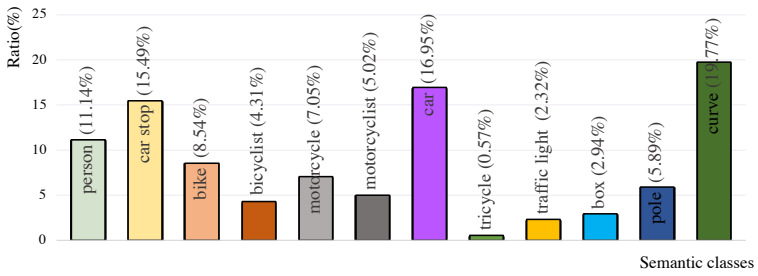
mislabeled samples and sending them back to the annotators for corrections and re-verification. Through these efforts, we successfully obtain a large-scale dataset with high-quality annotations.

6.2.2 Dataset Split

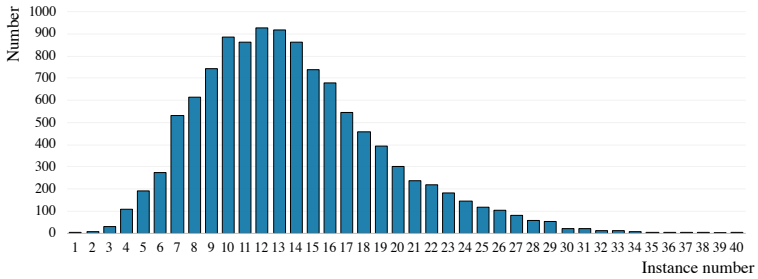
The entire dataset is randomly divided into three parts: 6,830, 1,705, and 2,836 samples for training, validation, and testing, respectively. To thoroughly assess the performance of various MSS models, we introduce six distinct test subsets based on representative image attributes. Specifically, the entire test set is denoted as *Test-All*. We divide the test set into *Test-Day* (daytime scenarios) and *Test-Night* (nighttime scenarios). *Test-MC* is constructed that comprises multi-class images (*i.e.*, images with 10 or more classes per image). *Test-MO* encompasses multi-object scenarios (*i.e.*, images with 20 or more instances per image). We also generate a low-contrast test subset, referred to as *Test-LC*, which contains images with the bottom 15% global color contrast scores [186] between foreground and background regions. Our dataset is the first to provide such informative test subsets, enabling a comprehensive evaluation of MSS algorithms.

6.2.3 Statistical Analysis

Table 6.1 provides an overview of statistical results of the proposed SemanticRT dataset and existing MFNet [42] and PST900 [46] datasets. As shown, the largest available dataset for MSS prior to our work (*i.e.*, MFNet [42]) contains



(a) Ratio of instances per class



(b) Distribution of instances per image

Figure 6.4: Instance-level statistical analysis of the SemanticRT dataset. (a) Percentage distribution of instance occurrences across object categories. (b) Frequency distribution of object instances per image.

only 1,569 samples, while our SemanticRT dataset comprises 11,371 samples. Our dataset is significantly larger, being more than 7 times the size of the MFNet benchmark and 12 times larger than PST900 [46]. Besides its large-scale nature, the proposed SemanticRT also features higher image resolution (1280×1024) compared to the 640×480 resolution found in MFNet, making it more suitable for contemporary high-resolution image processing applications. In Fig. 6.3, we further analyze the ratio of annotated pixels per class in each of these MSS datasets. It can be observed that, compared to previous MFNet and PST900 benchmarks, the proposed SemanticRT dataset includes more diverse semantic classes (13 in SemanticRT *vs.* 9 in MFNet and 5 in PST900) and a higher proportion of labeled foreground pixels (21.27% in SemanticRT *vs.* 7.86% in MFNet and 3.02% in PST900). This indicates that our SemanticRT dataset is more informative and presents greater challenges for MSS. In addition to pixel-level analysis, we also conduct instance-level analysis for SemanticRT. In Fig. 6.4(a), we present the ratio of instances per class in the SemanticRT dataset. Overall, there are more than 156k labeled instances in

the SemanticRT dataset, and the top three frequent classes are curve, car, and car stop. In Fig. 6.4(b), we illustrate the distribution of instances per image. As seen, 80% of images in SemanticRT contain more than 10 instances, demonstrating its richness in content. The proposed SemanticRT dataset is expected to encourage further advancements in related research fields.

6.3 Proposed Method

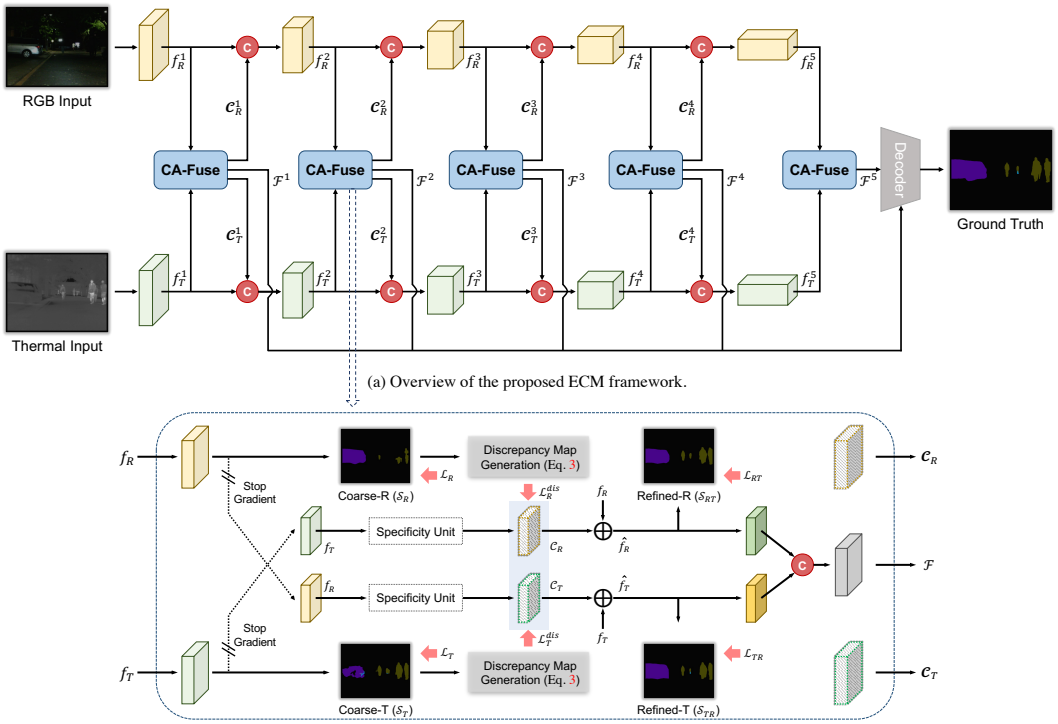


Figure 6.5: An overview of the proposed Explicit Complement Modeling (ECM) framework.

6.3.1 Method Overview

Fig. 6.5 illustrates an overview of our Explicit Complement Modeling (ECM) framework[2]. Overall our ECM strives to explicitly model inter-modality specificity cues and engage them into a robust cross-modal feature encoding and fusion process, so as to effectively aggregate complementary information in both

[2]Source code is publicly available at `https://github.com/jiwei0921/SemanticRT`.

RGB and thermal views. Building upon a two-stream feature extraction network, our ECM consists of two key components, *i.e.*, complements-aware feature fusion (CA-Fuse) and complements-aided feature encoding (CA-Encode). The CA-Fuse first adopts a well-designed cross-referenced discrepancy learning structure to explicitly extract modality-specific useful cues (referred to as *complements*), and then incorporates them into the cross-modal feature fusion process. The CA-Encode further engages the *complements* cues into the two-stream feature extraction process, thereby promoting sufficient multi-modal data interaction and propagation. Finally, the *complements*-furnished features are decoded by a segmentation decoder to render the predicted masks. We detail each component in the following section.

6.3.2 Proposed CA-Fuse

The critical aspect of CA-Fuse lies in how to explicitly identify the innate multi-modal complementary characteristics and model their specificity. Intuitively, a single-modality feature (*e.g.*, RGB appearance feature alone) possesses limited semantic information, which tends to only produce a relatively coarse segmentation prediction. By integrating the corresponding cross-modality feature (*e.g.*, RGB feature combined with its complementary thermal counterpart), the segmentation prediction is anticipated to be significantly refined. This intuition inspires us to design a cross-referenced discrepancy learning structure, attempting to approximate the desired inter-modality *complements* as the residuals between the coarse segmentation prediction and the refined segmentation prediction. As shown in Fig. 6.5, the RGB stream and thermal stream have identical structure, so in the following contents we take the RGB stream as an example to show the details.

Due to the intrinsic limitations of visible sensors, suboptimal lighting conditions often result in less-than-ideal outcomes when using the RGB data alone. Consequently, a single RGB feature[3], represented as f_R, often yields a coarse segmentation map, $\mathcal{S}_R$, as illustrated in Fig. 6.5(b). Concretely, we generate this map using a coarse segmentation branch $\Phi_{coarse}(\cdot)$, which is composed of

[3]We omit the layer superscript i for simplicity.

a 1×1 convolutional layer operating on f_R. We train this coarse branch using a cross-entropy segmentation loss $\mathcal{L}_R$. Mathematically, this process can be expressed as:

$$\begin{aligned} \mathcal{S}_R &= \Phi_{coarse}(f_R), \\ \mathcal{L}_R &= \mathcal{L}_{CE}(\mathcal{S}_R, \mathcal{G}). \end{aligned} \tag{6.1}$$

where $\mathcal{G}$ represents the ground-truth map.

Thereafter, we seek to extract the missing information in a single RGB feature from its associated thermal infrared counterpart. Thermal imaging cameras are able to perceive under varying lighting conditions. Thus, by merging the RGB feature with the RGB-*complement* derived from the thermal feature, the model should generate a more complete and precise segmentation map. In practice, we employ a specificity unit, $\mathcal{R}_{unit}(\cdot)$, composed of three vanilla convolutional blocks, to adaptively convert the information-redundant thermal feature f_T into an information-specific RGB-*complement* $\mathcal{C}_R$. By adding $\mathcal{C}_R$ to f_R, we then obtain an enhanced RGB feature, $\hat{f}_R$, and subsequently deduce a more refined segmentation map, $\mathcal{S}_{RT}$. This is achieved by $\Phi_{refine}(\cdot)$ with a 1×1 convolutional layer. We train this refined branch using another cross-entropy segmentation loss $\mathcal{L}_{RT}$. Formally, these operations can be described as:

$$\begin{aligned} \mathcal{C}_R &= \mathcal{R}_{unit}(f_T), \\ \hat{f}_R &= f_R + \mathcal{C}_R, \\ \mathcal{S}_{RT} &= \Phi_{refine}(\hat{f}_R), \\ \mathcal{L}_{RT} &= \mathcal{L}_{CE}(\mathcal{S}_{RT}, \mathcal{G}). \end{aligned} \tag{6.2}$$

To further reduce redundancy in the *complement* representation $\mathcal{C}_R$, we introduce a new discrepancy regularization loss $\mathcal{L}_R^{dis}$. Ideally, an optimal $\mathcal{C}_R$ should be able to bridge the gap between the coarse segmentation map $\mathcal{S}_R$ and the ground-truth map $\mathcal{G}$. As such, we calculate the distance between $\mathcal{S}_R$ and $\mathcal{G}$, which acts as an intermediate supervision signal to further regularize the learning of the *complement* representation. We achieve this by first computing the ground-truth discrepancy map, D_{GT}, using a normalized Kullback–Leibler

divergence function, expressed as:

$$\begin{aligned} D_{GT}^{(x,y)} &= \text{Norm}(KL(\mathcal{G}^{(x,y)}||\mathcal{S}_R^{(x,y)})) \\ &= \text{Norm}(\sum_{m\in M} \mathcal{G}^{(x,y)}(m) \log\,(\frac{\mathcal{G}^{(x,y)}(m)}{\mathcal{S}_R^{(x,y)}(m)})). \end{aligned} \tag{6.3}$$

In this expression, (x, y) represents a specific pixel; M and m are the total number of classes and a specific class, respectively; and Norm$(\cdot)$ is a min-max normalization function performed over the spatial dimension to rescale the discrepancy values into the range of (0, 1). Numerically, the discrepancy ground-truth map will yield a large value when the predicted semantic class distribution significantly deviates from the ground-truth distribution, and a small value when the distributions are more closely aligned.

Based on this ground-truth map, we then strive to constrain $\mathcal{C}_R$ to focus on capturing the discrepancy information. We accomplish this by applying on feature $\mathcal{C}_R$ a discrepancy head $\Phi_{dis}(\cdot)$ with a 1×1 convolutional layer to predict a discrepancy map, which is supervised using our discrepancy regularization loss, as follows:

$$\mathcal{L}_R^{dis} = ||D_{GT} - \Phi_{dis}(\mathcal{C}_R)||, \tag{6.4}$$

Here $||\cdot||$ represents the mean squared error (MSE) loss. Simultaneously, to prevent any potential conflicts between the discrepancy loss $\mathcal{L}_R^{dis}$ and segmentation losses $\mathcal{L}_R$, $\mathcal{L}_{RT}$, we stop the gradients in the RGB branch from back-propagating to the cross-referenced thermal encoder. The effectiveness of this stop gradient operation has been empirically verified.

Lastly, we generate the fused cross-modal feature $\mathcal{F}$ by concatenating the refined RGB and thermal features, $\hat{f}_R$ and $\hat{f}_T$, which have been enriched by the explicitly captured *complements*, $\mathcal{C}_R$ and $\mathcal{C}_T$. Formally, this process is represented as,

$$\mathcal{F} = Conv_{3\times3}(\hat{f}_R||\hat{f}_T), \tag{6.5}$$

where $Conv_{3\times3}$ is a 3×3 convolutional operation.

6.3.3 Proposed CA-Encode

After explicitly capturing the *complements* features $\mathcal{C}_R$ and $\mathcal{C}_T$, we further explore their potential in the two-stream feature encoding process using our CA-Encode. For each layer i, we enrich the raw output of the encoder with the complements features $\mathcal{C}_R$ and $\mathcal{C}_T$ by performing feature concatenation followed by 3×3 convolution operations. This results in two refined RGB and thermal features:

$$f_R^{out} = Conv_{3\times3}(f_R||\mathcal{C}_R); f_T^{out} = Conv_{3\times3}(f_T||\mathcal{C}_T), \tag{6.6}$$

which are propagated to the next encoding layer for more accurate and efficient encoding of the two modalities. In comparison to conventional independent two-stream encoders that are used in previous MSS models, our complement-aided encoders can alleviate the information deficiency issue in each individual feature extractor, and further boost the segmentation performance.

6.3.4 Segmentation Decoder

The final stage of ECM involves a segmentation decoder to predict segmentation masks from the cross-modality features in five layers, *i.e.*, $\{\mathcal{F}^i\}_{i=1}^5$. Specifically, we adopt a ASPP-based U-shape decoder to integrate the multi-layer features through feature interpolation and concatenation in a top-to-bottom pathway. The resulting feature is then processed to predict the final segmentation mask using a 3×3 convolutional layer. To train the network, we adopt the weighted cross-entropy and Lovasz losses in this work.

6.4 Experiments

6.4.1 Datasets

Empirical analyses are carried out on two existing MSS benchmarks and our newly-proposed SemanticRT dataset. Concretely, *MFNet* [42] dataset contains 1,569 annotated RGB-thermal pairs captured in urban scenario, with nine classes including background. This dataset is split into 784/392/393 for train/validation/test, respectively. The images have a resolution of 640×480

pixels. The test set is further divided into daytime (205) and nighttime (188) scenes for evaluation. *PST900* [46] dataset in its latest version contains 597 and 288 RGB-thermal pairs in train and test splits, respectively. It has five classes (including background). The image size is set to 640×1280 pixels for training. The proposed *SemanticRT* includes 11,371 RGB-thermal pairs with pixel-level semantic annotations of 13 semantic classes. This dataset is divided into 6,830/1,705/2,836 samples for train/validation/test, respectively. During training, the image is down-sampled to 640×512 pixels to balance the memory cost. To enable comprehensive evaluation, we introduce six test subsets based on different image attributes. They are *'Test-All'* with $2,836$ samples, *'Test-Day'* with 242 samples, *'Test-Night'* with $2,594$ samples, *'Test-MC'* with 328 samples, *'Test-MO'* with 170 samples, and *'Test-LC'* with 425 samples.

6.4.2 Evaluation Metrics

Following the standard protocol, we adopt mIoU (mean Intersection over Union) metric for evaluation. To be specific, for image k, the IoU_i are computed by

$$IoU_i = \frac{\sum_{k=1}^{K} \theta_{ii}^k}{\sum_{k=1}^{K} \theta_{ii}^k + \sum_{k=1}^{K} \sum_{j=1, j\neq i}^{N} \theta_{ji}^k + \sum_{k=1}^{K} \sum_{j=1, j\neq i}^{N} \theta_{ij}^k}, \tag{6.7}$$

where θ_{ii}^k is the number of pixels for class i correctly classified as class i, θ_{ji}^k is the number of pixels for class j incorrectly classified as class i, and θ_{ij}^k is the number of pixels for class i incorrectly classified as class j. The K and N stand for the numbers of images and classes in test set, respectively. Thus, we can obtain the $mIoU$ that separately represent the arithmetic average values of IoU across all semantic classes (including background class).

6.4.3 Ablation Studies

In this section, we provide detailed analysis on the effectiveness of our core model designs, using MFNet dataset.

Multispectral Information. We first investigate the benefits of multispectral information in Table 6.2 (a) & (b). As shown, the model trained with

Table 6.2: Ablation studies of ECM embedded in different backbone layers on the MFNet dataset. The upper section displays models without complements modeling, the middle section illustrates single-layer complements modeling. The bottom block shows our ECM with multiple layer complements modeling. Our ECM achieves the best result.

Settings	Index	Description	Layer1	Layer2	Layer3	Layer4	Layer5	mIoU (%)
w/o Complements Modeling	(a)	RGB-only	-	-	-	-	-	52.1
	(b)	RGB-Thermal	-	-	-	-	-	54.3
Single-layer Complements Modeling	(c)	ECM-single-1	✓					56.6
	(d)	ECM-single-2		✓				56.3
	(e)	ECM-single-3			✓			55.8
	(f)	ECM-single-4				✓		55.3
	(g)	ECM-single-5					✓	54.9
Multi-layer Complements Modeling	(h)	ECM-multi (**Ours**)	✓	✓	✓	✓	✓	58.0

RGB images along achieves an mIoU score of 52.1%. By incorporating the thermal infrared branch (*i.e.*, RGB+Thermal), we observe a notable performance increase of 2.2%, even when using a direct concatenation of multi-layer cross-modal RGB and thermal features. This reveals the advantages of utilizing multispectral information to enhance semantic segmentation.

Effectiveness of ECM Framework. We next validate the designs of our ECM framework, with results summarized in Table 6.2 & 6.3. From Table 6.3, we notice that incorporating explicit complement modeling at either the feature fusion stage (*i.e.*, CA-Fuse) or the feature encoding stage (*i.e.*, CA-Encode) leads to significant performance improvements (*i.e.*, mIoU from 54.3% to 56.3% for CA-Fuse, and from 54.3% to 57.1% for CA-Encode), compared to the "Baseline" without explicit complement modeling. When both modules are employed, a higher score of 58.0% mIoU is attained. These results evidence that explicitly integrating complements information during both the feature fusion and feature encoding stages contributes to more robust cross-modal semantic feature learning. In Table 6.2(c)-(h), we further assess the impact of ECM (including both CA-Fuse and CA-Encode) across various network layers. It is evident that ECMs at different network layers are all beneficial, leading to performance improvements ranging from 0.6% to 2.3%. When ECM is incorporated at an earlier stage, it tends to produce relatively higher performance. This could be attributed to the fact that cross-modal features in a shallower

Table 6.3: Ablation studies of our CA-Fuse and CA-Encode in ECM on the MFNet dataset.

Index	Modules				mIoU (%)
	Tab. 6.2b (Baseline)	Complements Modeling	+CA-Fuse	+CA-Encode	
(a)	✓	✗			54.3
(b)	✓	✓	✓		56.3
(c)	✓	✓		✓	57.1
(d)	✓	✓	✓	✓	58.0

Table 6.4: Ablation studies of various operations that guarantee effective complements modeling on the MFNet dataset.

Index	Operations					mIoU (%)
	Loss $\mathcal{L}_{R/T}^{dis}$	Cross-reference	RGB-Comp $\mathcal{C}_R$	T-Comp $\mathcal{C}_T$	Stop Gradient	
(a)	✓	✓	✓	✓	✓	58.0
(b)	✗					57.1
(c)		✗				55.9
(d)			✗			56.9
(e)				✗		56.7
(f)					✗	56.3

layer have larger feature resolution and possess more detailed complementary information. Additionally, fusing complements features at an earlier stage can also facilitate learning in subsequent network stages. Notably, our ECM embedded across all five network layers (*i.e.*, ECM-multi) achieves the highest performance with a 58.0% mIoU.

Operations in Complements Modeling. Table 6.4 provides a deep investigation into key operations in the complements learning process. Overall we observe consistent performance drops when each operation is individually removed, which implies that all our design elements are essential to guarantee effective complements modeling. The optimal complements learning, in turn, contributes to the superior performance of our CA-Fuse and CA-Encode. To be more specific, in Table 6.4(a) *vs.* (b), the elimination of discrepancy regularization loss $\mathcal{L}_{R/T}^{dis}$ results in a 0.9% decrease in mIoU. This showcases the benefits of explicitly guiding the learning of complements features using this loss. In Table 6.4(c), replacing the cross-referenced structure with a self-connected residual structure causes a large performance drop (*i.e.*, from 58.0% to 55.9%). This indicates that a single-modality feature is insufficient to provide the missing scene representations on their own, further emphasizing the superiority of our cross-referenced design. In Table 6.4(d)&(e), the performance decrease

Table 6.5: Generalizations of ECM using different backbones.

*	Models	Backbones	**FLOPs** (G)↓	**Params** (M)↓	**FPS**↑	**mIoU (%)**
CNN-based	ECM	ResNet-50	139.1	54.6	25.3	57.4
	ECM	ResNet-152	230.2	123.9	16.1	58.0
Transformer-based	ECM	Segformer-B2	155.4	58.2	17.8	58.6
	ECM	Segformer-B5	256.2	172.5	8.1	60.7

Table 6.6: Quantitative segmentation results on the MFNet test set [42]. '-' means unavailable results in the original papers. 3c and 4c mean taking RGB or RGBT as input, respectively.

Methods	Publications	**mIoU (%)**		
		Daytime	Nighttime	Overall
DFN(3c) [112]	CVPR'18	44.2	44.6	47.5
DFN(4c) [112]	CVPR'18	43.9	51.8	52.0
MFNet [42]	IROS'17	36.1	36.8	39.7
RTFNet [48]	RAL'19	45.8	54.8	53.2
PSTNet [46]	ICRA'20	-	-	48.4
FuseSeg [49]	TASE'20	47.8	54.6	54.5
FEANet [119]	IROS'21	-	-	55.3
MFFENet [187]	TMM'21	47.9	56.7	55.5
ABMDR [50]	CVPR'21	46.7	55.5	54.8
EGFNet [51]	AAAI'22	47.3	55.0	54.8
EGGFNet [188]	TITS'23	47.1	55.9	55.3
Ours	-	**49.0**	**57.3**	**58.0**

upon removing either the RGB-complement or Thermal-complement proves their significance in achieving more robust segmentation. Lastly, the stop gradient operation in Table 6.4(f) impacts the mIoU score by 2.7%, demonstrating its crucial role in facilitating stable network training.

Applications on Different Backbones. To verify the scalability of the proposed method, we further apply our ECM to various backbones, including CNN-based ones and Transformer-based ones, as shown in Table 7.8. It is observed that our ECM achieves appealing performance regardless of the used backbone networks, demonstrating its good generalization ability. Notably, our ECM with Segformer-B5 backbone [115] boosts the mIoU to 60.7%. For a fair comparison with previous RGB-T models [48], [51], we adopt ResNet-152 as our default backbone.

6.4.4 Comparison with State-of-the-Arts

Quantitative Results. *MFNet [42].* Table 6.6 lists the scores for daytime, nighttime, and all scenarios on the MFNet test set. Our method consistently

Table 6.7: Quantitative segmentation results on the PST900 test set [46]. "Fire-Ext" represents "Fire-Extinguisher".

Methods	Background	Fire-Ext	Back-Pack	Hand-Drill	Survivor	**mIoU (%)**
DLav3+(3c) [15]	98.81	47.31	68.30	54.73	47.29	63.29
DLav3+(4c) [15]	98.86	49.13	66.74	61.46	51.81	65.60
RTFNet [48]	99.02	51.93	74.17	7.07	70.11	60.46
PSTNet [46]	98.85	70.12	69.20	53.60	50.03	68.36
MFFENet [187]	99.34	79.76	76.61	66.79	63.01	77.10
EGFNet [51]	99.26	71.29	83.05	64.67	74.30	78.51
EGGFNet [188]	99.21	75.65	75.11	74.64	65.51	78.02
Ours	99.43	79.14	83.58	84.75	75.45	**84.47**

Table 6.8: Quantitative segmentation results on each test subset of the new SemanticRT dataset.

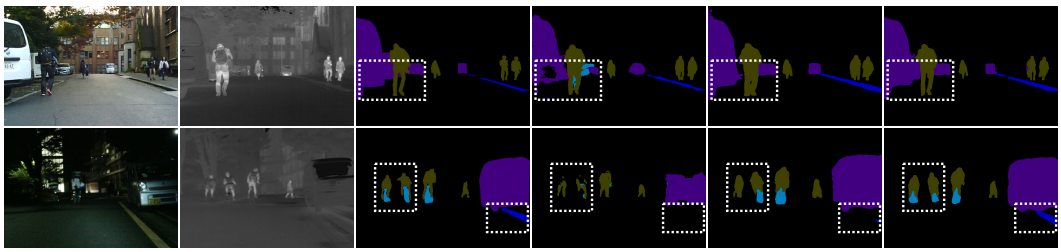
Methods	Test-Day	Test-Night	Test-MC	Test-MO	Test-LC	**Test-All**
DLav3+(3c) [15]	58.78	72.33	72.19	66.46	67.48	71.61
DLav3+(4c) [15]	65.73	76.38	76.73	72.03	72.36	75.80
MFNet [42]	61.64	74.66	74.39	68.92	70.14	74.08
RTFNet [48]	64.66	75.96	76.66	71.49	72.14	75.48
SAGate [189]	63.75	77.84	77.92	73.48	73.66	77.17
PSTNet [46]	51.48	69.33	71.00	60.75	64.05	67.98
EGFNet [51]	63.88	78.08	78.16	73.84	73.96	77.44
Ours	**70.33**	**79.81**	**80.09**	**75.47**	**76.21**	**79.26**

outperforms state-of-the-art models under diverse lighting conditions, with a considerable margin (*e.g.*, 58.0% *vs.* 55.3% for our ECM *vs.* EGGFNet [188]), demonstrating the effectiveness of our approach. *PST900 [46].* Table 6.7 presents comparison results on the PST900 dataset. Our approach again surpasses existing SOTA models, further showcasing the good generalization ability of our ECM. *SemanticRT.* We retrain existing MSS models with released codes using the newly-provided SemanticRT dataset, and report performance comparison of our method against five competitors in Table 6.8 (6 test subsets) and Table 6.9 (13 semantic classes). As we find that our ECM yields superior scores on all test subsets and almost all semantic classes, demonstrating its effectiveness in accurately parsing diverse challenging scenarios.

Qualitative Results. Fig. 6.6 depicts qualitative comparisons of our ECM against MFNet [42] and EGFNet [51] on the MFNet dataset. As seen, our method is capable of producing more accurate segmentation results under challenge scenarios.

Table 6.9: Quantitative segmentation results on each class of the new SemanticRT test set.

*	PSTNet	MFNet	RTFNet	EGFNet	**Ours**
Background	95.03	96.31	96.40	96.57	96.55
Car Stop	71.08	78.32	79.64	78.62	80.19
Bike	62.25	65.87	67.96	71.26	75.04
Bicyclist	58.48	64.07	67.41	70.86	75.50
Motorcycle	47.33	60.02	63.69	68.36	71.39
Motorcyclist	55.19	58.36	61.55	66.08	70.43
Car	85.41	89.70	90.39	90.52	90.26
Tricycle	44.18	62.10	65.96	71.51	74.01
Traffic Light	75.72	80.73	78.26	80.36	80.85
Box	83.00	83.93	85.91	85.41	85.61
Pole	71.65	77.14	78.02	76.49	77.23
Curve	62.15	66.18	67.22	66.92	68.28
Person	72.21	80.29	78.90	83.74	85.02
Test-All (mIoU)	67.98	74.08	75.48	77.44	**79.26**



(a) RGB Image (b) Thermal Image (c) Ground Truth (d) MFNet (e) EGFNet (f) Ours

Figure 6.6: Qualitative segmentation results on the MFNet test set.

6.5 Conclusion

In this chapter, our primary focus is on achieving resilient segmentation results across diverse lighting conditions. To address this challenge comprehensively, we present an explicit complement modeling (ECM) framework that explicitly captures modality-specific informative cues and seamlessly integrates them into a robust cross-modal feature fusion and encoding process. Our extensive empirical results underscore the effectiveness of the proposed ECM, demonstrating its superiority.

Chapter 7

RGB-Thermal Video Semantic Segmentation

7.1 Introduction

With the popularity of thermal imaging sensors, a growing demand in using thermal images for semantic segmentation has been witnessed. A number of RGBT models have been subsequently developed, to engage both RGB and thermal images as input for semantic segmentation especially with complex scenes [42], [46], [50], [51], [187]. This may be attributed to the fact that thermal infrared imaging is relatively insensitive to illumination conditions, as it works by recording infrared radiations of an object above absolute zero temperature [190]. It is worth noting that current multispectral segmentation methods are based on static images. The lack of mechanism to account for the temporal contexts may limit their performance when working with video inputs containing dynamic scenes, which are omnipresent in our daily lives. This leads us to explore in this paper a relatively new task of Multispectral Video Semantic Segmentation, or in short MVSS, with a specific focus on RGBT video inputs. Fig. 7.1 illustrates several exemplar multispectral video sequences and their ground-truth semantic annotations. As shown, the RGB frames and thermal frames provide rich and often complementary information for identifying moving foreground objects and static background scenes in low-light night or facing strong headlights. The new task opens up possibilities for applications that require a holistic view of video segmentation under chal-

Figure 7.1: Multispectral Video Semantic Segmentation. Examples of three typical real-life video sequences under diverse conditions are given (*daytime* (left), *nighttime* and *overexposure* (middle), *rainy* and *low-light* (right)), where RGB images, thermal infrared images, and their pixel-level semantic annotations are shown through the first to the third rows, respectively.

lenging conditions, *e.g.*, autonomous safe driving, nighttime patrol, and fire rescue. To our knowledge, this is the first work to address such multispectral video semantic segmentation problem.

In the deep learning era, benchmark datasets have become the critical infrastructure upon which the computer vision research community relies to advance the state-of-the-arts. Thanks to the publicly available benchmarks, such as MFNet [42], PST900 [46], Cityscapes [8], and CamVid [191], the related tasks of multispectral semantic segmentation (MSS) and video semantic segmentation (VSS) have evidenced notable progresses. Meanwhile, these existing datasets provide as input either single pairs of RGB and thermal images, or RGB only video sequences. There unfortunately lacks a suitable dataset to train and evaluate learning based models for the proposed MVSS task. This leads us to curate a high-quality and large-scale MVSS dataset, referred to as MVSeg, that contains diverse situations. Specifically, our MVSeg dataset comprises 738 synchronized and calibrated RGB and thermal infrared video sequences, with a total of 52,735 RGB and thermal image pairs. Among them, 3,545 image pairs are densely annotated with fine-grained semantic segmentation labels, consisting of a rich set of 26 object categories in urban scenes. In particular, as showcased in Fig. 7.1, our MVSeg dataset involves many challenging scenes with adverse lighting conditions. It is expected to provide a

sufficiently realistic benchmark in this field.

Furthermore, a dedicated baseline model is developed for this new task, which is called Multispectral Video semantic segmentation NETwork or simply MVNet. Our MVNet possesses two key components in addressing the main challenges of MVSS task. Considering the high complexity of processing large-volume multispectral video data, a prototypical MVFuse module is devised to attend to rich contextual multispectral video features with a moderate memory footprint. A novel MVRegulator loss is further introduced, which regularizes the feature learning process to reduce cross-spectral modality difference and promote better exploitation of multispectral temporal data. Comprehensive experiments on various state-of-the-art semantic segmentation models are also carried out at the MVSeg dataset. Experimental results demonstrate the significance of multispectral video data for semantic segmentation, and verify the effectiveness of our MVNet model. We expect the MVSeg dataset and the MVNet baseline will facilitate future research activities toward the MVSS task.

7.2 Proposed MVSeg Dataset

Benchmark datasets have become the critical infrastructure upon which the computer vision research community relies to advance the state-of-the-arts. To benchmark this new multispectral video semantic segmentation (MVSS) field, we curate a high-quality and large-scale MVSS dataset, referred to as MVSeg. Here we focus on describing the construction of the MVSeg dataset[1], and analyzing the statistical results.

7.2.1 Dataset Construction

Data collection. Our goal is to collect a large-scale dataset with calibrated visible (RGB) and thermal infrared video sequences, covering a diverse set of challenging scenes, with high-quality dense annotations. We gathered RGB-

[1]Dataset is publicly available at `https://jiwei0921.github.io/Multispectral-Video-Semantic-Segmentation/resources/dataset.txt`.

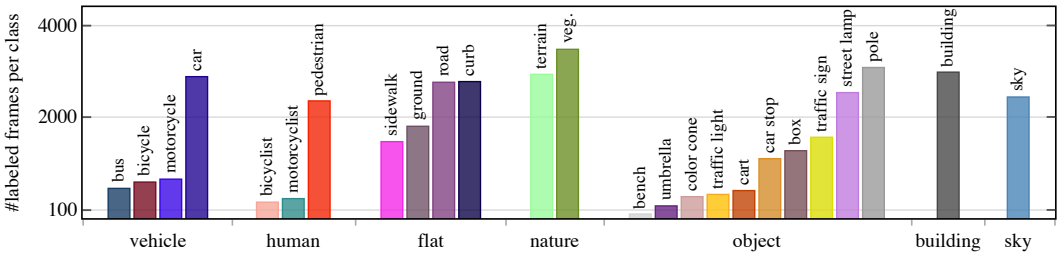


Figure 7.2: Statistics on the number of finely annotated frames (y-axis) per class and root category. The background class is not shown.

thermal videos from multiple sources in related works, including OSU [183], INO [184], RGBT234 [192], and KAIST [193], and manually selected 738 high-quality video shots (5 seconds on average) to build our MVSeg dataset. Most of these videos are at the resolution of 480×640. This dataset covers many complex scenes during daytime, nighttime, normal weather conditions (*e.g.*, sunny and cloudy), and adverse weather conditions (*e.g.*, rainy, snowy and foggy). We illustrate several visual examples in Fig. 7.1.

Classes and annotations. To identify object classes of interest, we carefully reviewed all paired videos of both RGB and thermal modes, and collected all object classes that appeared in the dataset. Then 26 object classes of interest were selected for annotation, which were grouped into 8 root categories, including vehicle, human, flat, nature, object, building, sky, and background (unlabeled pixels), as illustrated in Fig. 7.2. Guided by [8], criteria for selecting classes were based on their frequency, relevance to the applications, practical considerations for annotation efforts, and promoting compatibility with existing datasets, *e.g.*, [8], [42].

Labeling the MVSeg dataset poses greater challenges compared to RGB-based segmentation datasets. Firstly, the MVSeg dataset contains many challenging scenes recorded under adverse conditions, which complicates the identification of less visible objects and the differentiation of their silhouettes. To assist annotators, we display RGB and thermal image pairs side by side, synchronizing their annotation traces to provide useful reference information. Secondly, we strive for consistent annotations between adjacent frames in a video by presenting a "global" view of annotated frames within each video.

Table 7.1: High-level statistics of our MVSS dataset and existing MSS/VSS datasets. 'Seq.' means providing sequential video frames; 'TIR' means providing thermal infrared images. * Data annotations are obtained by human and models jointly.

Dataset	Seq.	TIR	#Videos(Frames)	#GTs	#Classes
Cityscapes [8]	✓		- (150k)	5,000	30
CamVid [191]	✓		5 (40k)	701	32
VSPW* [88]	✓		3,536 (252k)	252k	124
MFNet [42]		✓	-	1,569	9
PST900 [46]		✓	-	894	5
MVSeg	✓	✓	738 (53k)	3,545	26

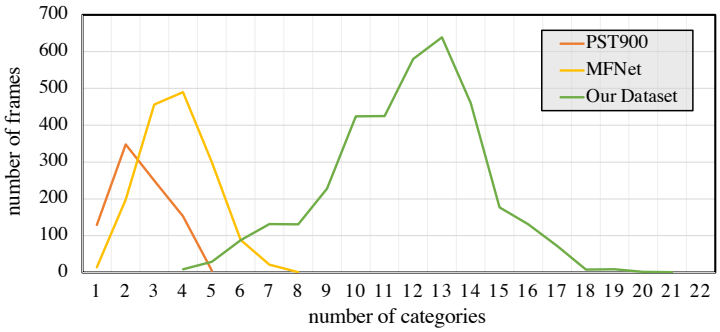


Figure 7.3: The number of categories per frame across existing semantic segmentation datasets with RGB and thermal pairs.

This allows inspectors to more easily spot missing objects and inconsistent annotations. Despite these efforts, the annotation and quality control process for the MVSeg dataset still remains time-consuming, averaging over 50 minutes per video frame due to the intricate nature of dense pixel-level semantic labeling and the challenging scenes it encompasses.

7.2.2 Statistical Analysis

Table 7.1 shows an overview of the statistical results of the proposed MVSeg dataset and related MSS/VSS datasets. Our MVSeg dataset contains 738 multispectral videos at a frame rate of 15 f/s, including 53K image pairs in total and 3,545 annotated image pairs of 26 categories. Similar to other VSS datasets (Cityscapes [8] and CamVid [191]), we annotate one frame for every 15 frames. We may notice that our MVSeg dataset and the MSS datasets (MFNet [42] and PST900[46]) have fewer annotated GTs than VSS datasets. This is reasonable due to the scarcity of calibrated multispectral images/videos

Table 7.2: The pixel percentage per root category across existing multispectral (RGBT) semantic segmentation dataset, where '-' means no such classes.

Dataset	vehicle	human	flat	nature
MFNet [42]	5.05%	1.20%	0.59%	-
PST900 [46]	-	1.36%	-	-
Our Dataset	6.79%	0.91%	37.15%	33.22%
Dataset	object	building	sky	bkg.
MFNet [42]	1.02%	-	-	92.14%
PST900 [46]	1.66%	-	-	96.98%
Our Dataset	1.77%	11.76%	7.36%	1.04%

and the difficulty of annotating such data. Meanwhile, our MVSeg dataset has comparable or richer object categories compared to MSS & VSS datasets. Fig. 7.2 illustrates the detailed object sub- and root-categories in MVSeg, and plots the number of frames in each category. It shows that the distribution is unbalanced between each class, similar to any other semantic segmentation datasets. The common categories, *e.g.*, car and pedestrian, appear in most of frames. Table 7.2 lists the pixel-wise annotate rate for each root-category in the multispectral-based datasets. It is shown that existing MSS datasets [42], [46] only label a small fraction of pixels in a scene (7.86% and 3.02%, respectively). In comparison, our MVSeg dataset has a high pixel-wise annotation rate of 98.96%, which is more meaningful for understanding the entire scene. Finally, we display the distribution of categories in video frames in Fig. 7.3. It is shown that most frames in [42], [46] only contain 4 or 2 categories, whereas that result is 13 categories in our MVSeg.

7.3 Proposed Method

7.3.1 Technical Motivation

To date, various network architectures have been developed for the tasks of MSS and VSS. In the former task, many advanced feature fusion techniques have been designed to fuse features extracted from multispectral images based on two-stream encoders. The latter task focuses more on exploiting temporal associations in video sequence, such as optical flow warping [194] or space-time attention [195]. The use of either multispectral or temporal information has

demonstrated their individual advantages in improving segmentation accuracy & robustness. However, there is no research touching the joint learning of both *multispectral* and *temporal* contexts which are both essential for MVSS.

Drawing ideas from recent MSS/VSS models, a straightforward solution for a MVSS model is to, *first* extract features from different spectra data using two-stream encoders, *then* build an external memory to hold the rich temporal & multispectral features, and *finally* extend the conventional space-time attention to an advanced spectrum-space-time version, where pixels of query features attend to all pixels of memory features, including these of RGB and thermal modalities as well as these of past video frames. In this way, we can definitely exploit the rich source of multispectral video features, and learn a joint relationship from multispectral and temporal contexts for semantic segmentation.

However, there are two certain challenges associated with this straightforward solution. ❶ *The first challenge* is how to keep the computational and memory costs moderate when processing large amounts of multispectral video data. As revealed, conventional attention block that performs all-to-all matching of feature maps is memory-consuming and computationally expensive [195]; it is unsuitable and unaffordable for MVSS, as multispectral video streams usually come in sequentially and need to be processed on time. This requires us to devise more elegant strategies for efficient MVSS. ❷ *The second challenge* comes from the inherent modality differences between RGB and thermal modes. Due to imaging differences, RGB data usually provide rich visible appearance information, while thermal data present more invisible thermal radiations of objects. Such modality differences will cause the feature embeddings of RGB and thermal frames to be distributed in different embedding spaces, leading to suboptimal cross-spectral feature attending and affecting the full exploitation of cross-spectral complementarity. Therefore, we should properly address the modality difference issue to make better use of multispectral complementary information. In Sec. 7.3.2, we introduce a well-designed MVSS baseline, called MVNet, which addresses the two challenges for MVSS.

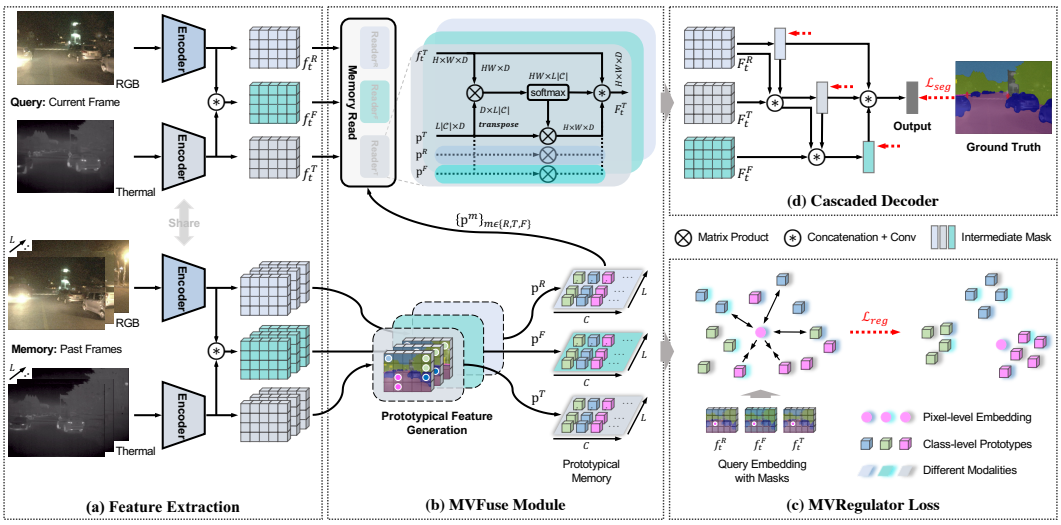


Figure 7.4: Illustration of the proposed MVNet. The input is a multispectral video clip, which contains one Query pair of RGB and thermal images, as well as L Memory pairs at past frames. The MVNet consists of four parts: (a) Feature Extraction to obtain the multispectral video features; (b) an MVFuse Module to furnish the query features with the rich semantic cues of memory frames; (c) an MVRegulator Loss to regularize the multispectral video embedding space; and (d) a Cascaded Decoder to generate the final segmentation mask.

7.3.2 Proposed MVNet Baseline

Fig. 7.4 presents an overview of the proposed MVNet[2]. Starting from the input multispectral video, its pipeline consists of four parts: (a) feature extraction; (b) an MVFuse module to address challenge ❶; (c) an MVRegulator loss to address challenge ❷; and (d) a cascaded decoder to generate the final segmentation mask.

Feature Extraction: The multispectral video input contains a Query pair of RGB and thermal images at current frame t, and L Memory pairs at past frames. They are denoted as $\{I_d^m\}_{d\in U, m\in\{R,T\}}$, where d represents the time subscript of a certain frame in the set of $U=\{t-L,\ldots,t-1,t\}$, and m denotes the modality type in $\{R,T\}$.

These image pairs are fed into two-stream encoders to extract RGB and thermal features, respectively. To enrich the features, we fuse the outputs of different spectra by concatenation and 1×1 convolution, resulting in a series of

[2]Source code is publicly available at `https://github.com/jiwei0921/MVSS-Baseline`.

fused features. These RGB, thermal, and fused features, together constitute a rich source of multispectral temporal cues for MVSS. We represent these features as $\{\mathbf{f}_d^m \in \mathbb{R}^{H\times W\times D}\}_{d\in U, m\in\mathcal{M}}$, where $H\times W$ represents the spatial size, D is the channel dimension, and $\mathcal{M} = \{R, T, F\}$.

MVFuse: An MVFuse module is then developed in Fig. 7.4b to furnish the Query features by engaging the rich yet cumbersome features of Memory frames. This is realized by two key designs: a *memory-efficient* prototypical memory and a *computationally-efficient* memory read block.

To preserve as many representative "pixels" as possible with minimal memory consumption, we build a prototypical memory that stores only a small number of the most representative categorical features of memory frames. Specifically, for each memory feature $\mathbf{f}_d^m$, we derive $|\mathcal{C}|$ class-level prototypical features, by average pooling all the embeddings of pixels belonging to each category $c \in \mathcal{C}$. The estimated semantic masks are employed here to provide the required pixel category information of memory frames. Therefore, the memory features are summarized into a condensed set of prototypical features. We group the prototypical features of each modality as $\{\mathbf{p}^m \in \mathbb{R}^{L|\mathcal{C}|\times D}\}_{m\in\mathcal{M}}$.

Afterwards, we devise an efficient Memory Read block, which enables a fast and efficient access of relevant semantic cues from prototypical memory to refine query features. This is achieved via an all-to-prototype attention. Taken the *query* feature $\mathbf{f}_t^T$ as an example, we match it against all *key*s in prototypical memory. As shown in Fig. 7.4b, the inner product between the reshaped $\mathbf{f}_t^T$ and $\mathbf{p}^m$ are calculated as correlation maps, and transformed to weighting maps using a Softmax layer, expressed as:

$$\mathbf{w}^m = \text{Softmax}(\mathbf{f}_t^T \otimes \mathbf{p}^m), m \in \mathcal{M}. \tag{7.1}$$

Here we process the attending of each modality separately, due to their different characteristics. The learned weighting maps are then used to selectively retrieve relevant information from memory, and update the query feature by:

$$\mathbf{F}_t^T = \Phi(\text{Concat}[\{\mathbf{w}^m\mathbf{p}^m\}_{m\in\mathcal{M}} \cup \{\mathbf{f}_t^T\}]), \tag{7.2}$$

where Concat$[\cdot]$ denotes feature concatenation along channel dimension, and $\Phi(\cdot)$ is a 1×1 convolution operation to reduce the channel number to the

original feature size.

Our MVFuse module finally outputs three informative features $\mathbf{F}_t^R$, $\mathbf{F}_t^T$, and $\mathbf{F}_t^F$ ($\mathbb{R}^{H\times W\times D}$) that have equipped with rich temporal and multispectral contexts, by modeling both cross-spectral and cross-frame relationships. In practice, we find that this strategy is not only more efficient (reducing the complexity from $\mathcal{O}(L(HW)^2)$ to $\mathcal{O}(L(HW)\times|\mathcal{C}|)$, where $|\mathcal{C}|\ll HW$), but also more effective (increasing mIoU by 0.3%) than conventional attention that densely models pixel-to-pixel relationships. This may be partly due to the way of dense pixel matching may introduce some unnecessary or wrong correlations between regions with similar semantic but different classes, whereas our prototypical memory can degrade the side effects of ambiguous pixels and preserve the most typical representations.

MVRegulator: Inspired by the contrastive loss in unsupervised representation learning [196], [197], we further design a tailored MVRegulator loss for MVSS. Intuitively, features from different spectra or video frames but with the same object class should be closer to each other than any other features with different object classes in the same video.

Specifically, for a query pixel $\mathbf{f}_t^m(i,j)$ at position (i,j) of modality m with its groundtruth semantic label $\bar{c}$, the positive set $\mathcal{P}$ includes prototypical features also belonging to the class $\bar{c}$, and its negative set $\mathcal{N}$ consists of prototypical features belonging to the other classes $\mathcal{C}/\bar{c}$. We include prototypical features of Query frame into the contrastive sets to consider within-frame contrasts. Formally, the MVRegulator loss is defined as:

$$\mathcal{L}_{reg}^m(i,j)=\frac{1}{|\mathcal{P}|}\sum_{\mathbf{p}^+\in\mathcal{P}}-\log\frac{\exp(\mathbf{f}\cdot\mathbf{p}^+/\tau)}{\exp(\mathbf{f}\cdot\mathbf{p}^+/\tau)+\sum_{\mathbf{p}^-\in\mathcal{N}}\exp(\mathbf{f}\cdot\mathbf{p}^-/\tau),} \tag{7.3}$$

$$\mathcal{L}_{reg}=\frac{1}{|\mathcal{M}|\times H\times W}\sum_{m\in\mathcal{M}}\sum_{(i,j)\in[H,W]}\mathcal{L}_{reg}^m(i,j). \tag{7.4}$$

Here $\mathcal{L}_{reg}^m(i,j)$ is the partial loss for query pixel $\mathbf{f}_t^m(i,j)$ (simplified as $\mathbf{f}$ in Eq. 7.3), τ denotes the temperature parameter, and all the embeddings are l_2-normalized.

With $\mathcal{L}_{reg}$, the model is able to not only reduce modality differences between different spectra, but also promote intra-class compactness & inter-class

separability. We would note that the MVRegulator loss is performed only during training, so it does not affect the inference time.

Cascaded Decoder: The final stage of the MVNet involves a cascaded decoder to predict segmentation mask based on $\mathbf{F}_t^R$, $\mathbf{F}_t^T$, and $\mathbf{F}_t^F$. Instead of direct prediction, we propose to cascadely integrate these features, and impose multiple supervisions on each level and the final result. This strategy is able to further promote multi-modal feature interaction and help filter unnecessary information redundancy. The segmentation loss in the decoder is then computed by the sum of these supervisions as:

$$\mathcal{L}_{seg} = \mathcal{L}_{wCE} + \sum\nolimits_{m \in \mathcal{M}} \mathcal{L}_{wCE}^{m}, \tag{7.5}$$

where we adopt the weighted cross-entropy loss $\mathcal{L}_{wCE}$ suggested by [50], [51], [198] for training. The overall training objective of the MVNet is thus defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{reg}, \tag{7.6}$$

where λ is a weighting parameter for balancing the losses.

7.4 Experiments

7.4.1 Datasets

Empirical analyses are carried out on *MVSeg* dataset. It is split into training, validation, and test sets, which consist of 452/84/202 videos with 2,241/378/926 annotated image pairs, respectively. The entire test set is also divided into daytime and nighttime scenes (134/68 videos), to make a comprehensive evaluation. During training, all images are resized to 320×480.

7.4.2 Evaluation Metrics

Following the standard protocol, we adopt mIoU (mean Intersection over Union) metric for evaluation.

7.4.3 Ablation Studies

To investigate the effect of our core designs, we conduct ablation studies on the test set of MVSeg, with results presented in Table 7.3-7.5. Throughout

Table 7.3: Quantitative results of ablation study. 'TIR' means thermal infrared image. #Params refers to model parameters. #Mem means GPU memory usage during training. The inference time (ms) per frame is calculated under the same input scale.

*	Model Setups	#Param (M)	#Mem (G)	Times (ms)	mIoU (%)
(a)	RGB	41.6	4.6	8.1	51.59
(b)	RGB+TIR (direct fusion)	85.5	7.1	15.5	52.53
(c)	RGB+TIR (cascade fusion)	87.5	7.6	15.9	52.87
(d)	(c)+MVFuse$_{stm}$	96.1	45.7	32.6	53.74
(e)	(c)+MVFuse$_{lma}$	95.6	25.3	25.3	53.95
(f)	(c)+MVFuse$_{proto}$	88.4	18.7	18.4	54.03
(g)	(f)+MVRegulator$_{uni}$	88.4	18.8	18.4	54.26
(h)	(f)+MVRegulator (Ours)	88.4	18.8	18.4	54.52

the ablation experiments, we use DeepLabv3+ [15] as the backbone encoder.

Multispectral Information. We first investigate the benefits of multispectral information in Table 7.3(a)&(b). As shown, the model trained with RGB images along achieves an mIoU score of 51.59%; adding the thermal infrared (TIR) branch brings a substantial performance gain of 0.94% even using a simple direct fusion strategy (*i.e.*, direct concatenation). This reveals the benefits of leveraging multispectral information to improve semantic segmentation.

Cascaded Decoder. We then validate the efficacy of our cascaded decoder by using it to replace the direct fusion strategy. As shown in Table 7.3(c), the cascaded decoder leads to an mIoU gain of 0.34%, thanks to the advantages of our cascaded decoder to better filter & fuse complementary information from RGB and thermal modes.

MVFuse Module. We deeply investigate the design of our MVFuse module in Table 7.3(d)-(f). Based on "model (c)", we examine three MVFuse variants, *i.e.*, MVFuse$_{stm}$, MVFuse$_{lma}$, and our proposed MVFuse$_{proto}$, which differ in the design of memory and attention, while remaining all other settings the same. Technically, MVFuse$_{stm}$ performs an all-to-all matching attention between query and memory frames with a large pixel-wise memory; MVFuse$_{lma}$ reads only the spatial neighborhood regions of each position in query frame from pixel-wise memory. The results suggest that, **i)** leveraging multispectral video data is indeed useful, since all MVFuse variants yield increased mIoU scores compared to the single-frame baseline (c), ranging from 0.87% to 1.16%;

Table 7.4: Ablation on the impact of memory size using mIoU(%).

Memory Size	$M=0$	$M=1$	$M=2$	$M=3$	$M=4$	$M=5$
Ours, $S=3$	52.87	53.96	54.21	54.52	54.57	54.52

Table 7.5: Ablation on the impact of sample rate using mIoU(%).

Sample Rate	$S=1$	$S=2$	$S=3$	$S=4$	$S=5$
Ours, $M=3$	54.25	54.47	54.52	54.41	54.39

and **ii)** our MVFuse$_{proto}$ module is more favored, since it performs better, has smaller model size, faster inference, and requires less GPU memory, compared to MVFuse$_{stm}$ and MVFuse$_{lma}$. We attribute this to the superiority of our memory-efficient prototypical memory to preserve as many representative "pixels" as possible in the video, and our computationally-efficient memory read block to engage the rich multispectral temporal knowledge.

MVRegulator Loss. We evaluate the MVRegulator loss in Table 7.3(g)&(h). As shown, integrating our MVRegulator loss improves mIoU score by 0.49% (*i.e.*, 54.03%→54.52%), without introducing any extra model parameters or affecting inference time, which demonstrates its effectiveness to generate a more structured feature embedding space. We also derive an MVRegulator$_{uni}$ variant, which removes the cross-spectral contrast in Eq. 7.3. As seen, the mIoU score degrades, further showcasing the necessity of addressing the modality differences issue in MVSS.

Memory Frames Selection. This part examines the impact of memory size M and sample rate S for memory frame selection. As shown in Table 7.4, adding memory frames consistently improves mIoU scores compared to the single-frame baseline (*i.e.*, $M=0$). When using more memory frames (*i.e.*, $M=3$), we see a clear performance increase (*i.e.*, 52.87%→54.52%). Raising M further beyond 3 gives marginal returns in performance. As a result, we set $M=3$ for a better trade-off between accuracy and memory cost. Then we fix memory size $M=3$, and experiment with different sample rate S. As shown in Table 7.5, best result is achieved when using a moderate sample rate $S=3$. We set M and S to 3 in MVNet, which can efficiently make use of past video frames without holding on too old information.

Table 7.6: Quantitative evaluation on the test set of MVSeg dataset. The notation † and ‡ mean the VSS and MSS models, respectively.

Method	Backbone	mIoU(%)
CCNet [111]	ResNet-50	51.70
OCRNet [103]	ResNet-50	52.38
STM† [199]	ResNet-50	52.51
LMANet† [195]	ResNet-50	52.73
MFNet‡ [42]	Mini-inception	51.63
RTFNet‡ [48]	ResNet-152	52.77
EGFNet‡ [51]	ResNet-152	53.44
FCN [21]	ResNet-50	50.67
MVNet$_{FCN}$	ResNet-50	53.90 (+3.23)
PSPNet [20]	ResNet-50	51.38
MVNet$_{PSPNet}$	ResNet-50	54.36 (+2.98)
DeepLabv3+ [15]	ResNet-50	51.59
MVNet$_{DeepLabv3+}$	ResNet-50	**54.52** (+2.93)

7.4.4 Benchmark Results

We first benchmark MVSS by performing comprehensive experiments on various segmentation methods, including *image-based SS models* (CCNet [111], OCRNet [103], FCN [21], PSPNet [20], and DeepLabv3+ [15]), *MSS models* (MFNet [42], RTFNet [48], and EGFNet [51]), *VSS models* (STM [199] and LMANet [195]), and our proposed *MVSS model* - MVNet, using the MVSeg dataset.

Table 7.6 presents the segmentation results on the test set of MVSeg. Since there is no prior work directly applicable to the new MVSS task, we first present the closely-related SS/MSS/VSS methods to provide a reference level. We reproduce these methods using their published codes with default setups. In our MVSS model, one important expectation compared to its image-level counterpart is whether the MVSS model improves per-frame segmentation accuracy by properly utilizing multispectral temporal features. To verify it, we apply our method to three popular image-based segmentation networks, including FCN [21], PSPNet [20], and DeepLabv3+ [15], to thoroughly validate the proposed algorithm. It is shown that our approach improves the performance of base networks by solid margins (*e.g.*, 51.38%→54.36% for PSPNet), suggesting that leveraging the multispectral temporal contexts is indeed beneficial for semantic segmentation, which has remained relatively untapped.

Table 7.7: Quantitative results on daytime and nighttime scenarios of MVSeg dataset, respectively, evaluated using mIoU (%) metric.

Method	*Daytime*	*Nighttime*
CCNet [111]	54.59	38.38
OCRNet [103]	55.42	38.79
STM† [199]	55.22	38.19
LMANet† [195]	56.52	38.54
MFNet‡ [42]	54.63	39.14
RTFNet‡ [48]	56.62	39.26
EGFNet‡ [51]	56.89	40.10
FCN [21]	53.02	37.40
MVNet$_{FCN}$	57.19 (+4.17)	40.05 (+2.65)
PSPNet [20]	54.62	37.29
MVNet$_{PSPNet}$	57.73 (+3.11)	39.53 (+2.24)
DeepLabv3+ [15]	55.17	38.13
MVNet$_{DeepLabv3+}$	**57.80** (+2.63)	**40.48** (+2.35)

Moreover, our MVNet shows a good generalization ability, which achieves consistently improved segmentation performance, independent of base networks.

To further evaluate the methods, we test them on daytime and nighttime scenarios, with results reported in Table 7.7. Again, our approach brings impressive gains over three strong baselines on both daytime and nighttime scenarios. For example, our MVNet$_{DeepLabv3+}$ yields mIoU scores of 57.80% and 40.48% on daytime and nighttime scenes, respectively, which shows promising gains of 2.63% and 2.35% over its counterpart DeepLabv3+. This further demonstrates the advantages of our MVNet to segment target objects under diverse lighting conditions.

Table 7.8: Analysis of our proposed MVNet with different backbones. ⋆ denotes transformer-based models that have input image with size 480×480.

Methods	Backbone	#Param(M)	#Mem(G)	Times(ms)	mIoU(%)
DeepLabv3+ [15]	ResNet-50	41.6	4.6	8.1	51.59
Ours (DeepLabv3+)	ResNet-50	88.4	18.8	18.4	54.52
SegFormer⋆ [115]	MiT-B1	13.8	4.0	7.9	51.11
Ours (SegFormer)⋆	MiT-B1	33.7	14.1	17.6	54.25
SegFormer⋆ [115]	MiT-B2	24.8	4.6	13.7	53.07
Ours (SegFormer)⋆	MiT-B2	56.1	18.6	29.8	55.22

In Table 7.8, we show the results of the proposed MVNet using CNN-based and transformer-based backbone as feature extraction network. Specifically, we apply our method to the transformer-based image segmentation network, SegFormer (MiT-B1&-B2) [115], to provide a more thorough validation. The

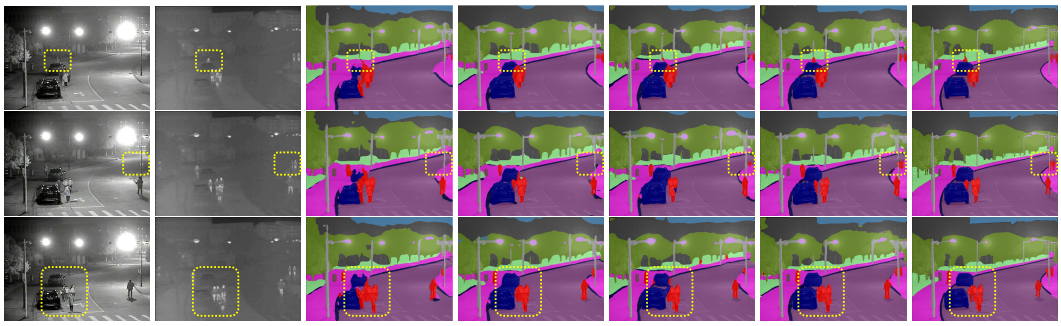


Figure 7.5: Qualitative results on the MVSeg dataset. We highlight the details with the yellow boxes. Best viewed in color and zoom in.

model parameters, GPU memory usage during training, and inference time (ms) per frame are also included. These results consistently verify the benefits of incorporating multispectral temporal contexts for semantic segmentation as well as the superiority of our approach.

Fig. 7.5 visualizes the segmentation results of a challenging nighttime scene with dim light. Compared with the competing methods, the results from our MVSS model (*i.e.*, $\text{MVNet}_{DeepLabv3+}$) are more accurate.

7.5 Conclusion

In this work, we have presented a preliminary investigation on the new task of semantic segmentation of multispectral video inputs. Specifically, we have provided a new challenging and finely annotated MVSeg dataset, developed a simple but efficient baseline framework (*i.e.*, MVNet), conducted comprehensive benchmark experiments, and highlighted several potential challenges and future directions. The above contributions provide an opportunity for the community to design new algorithms for robust MVSS.

Chapter 8

Conclusion, Discussion and Future Work

In this chapter, we begin by summarizing our primary contributions. Following this, we discuss the recent trends concerning large foundational models, *e.g.*, segment anything model. Finally, we outline some compelling issues and future research avenues within the field.

8.1 Conclusion

In conclusion, this thesis represents a significant leap forward in the realm of visual scene segmentation, a pivotal field with wide-ranging implications. Our efforts in developing innovative deep learning algorithms and resources have led to substantial enhancements in the robustness and efficiency of segmentation models. These improvements enable more effective performance in intricate visual settings, marking a notable advancement over prior methods.

Our systematic exploration across RGB-D salient object detection and RGB-Thermal semantic segmentation have not only advanced the state of the art but also provided practical solutions that can be readily applied in real-world scenarios. Among our key contributions is a groundbreaking depth calibration strategy that effectively mitigates latent biases in original depth inputs, an advanced DMRA network that elevates multimodal fusion and contextual understanding, and a pioneering approach to deep unsupervised RGB-D saliency detection that significantly reduces the need for human annotations.

Additionally, the introduction of two large-scale datasets, SemanticRT and MVSeg, alongside the development of the ECM and MVNet networks, have catalyzed progress in RGB-Thermal image and video semantic segmentation.

By offering a holistic suite of methodologies for addressing the multifaceted challenges in visual scene segmentation, our research stands as a testament to the substantial progress achievable through thoughtful innovation. We have ensured that all source codes, pre-trained models, and datasets are publicly accessible, further amplifying the impact of our work.

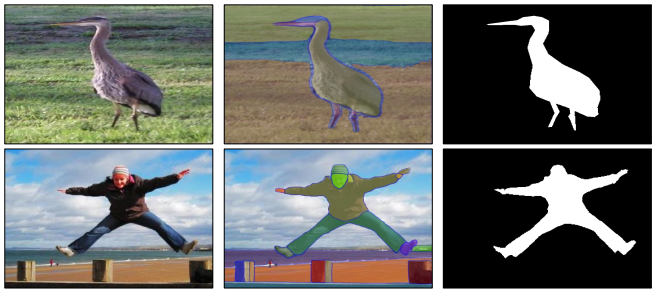
8.2 Discussion

The discussion about foundation model is based on our study published in [6].

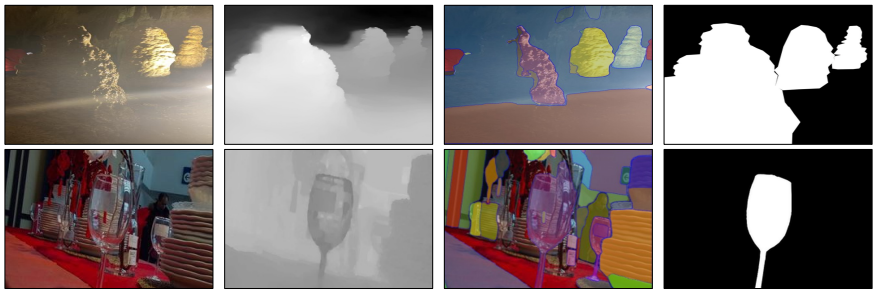
Recently, large foundation models [133], [200]–[202] are transforming the landscape with powerful zero-shot capabilities, which can be attributed to the sufficient pre-training on web-scale datasets and their superior ability to generalize to various downstream tasks. Regarding the context of this thesis, a standout work is Segment Anything Model (SAM) [133], which has gained great success for the zero-shot image segmentation. The strength of SAM lies in its interactive segmentation paradigm: the model segments the region of interest following the user-given prompts, such as a point, a bounding box, or free text-like descriptions.

The emergence of SAM has undoubtedly demonstrated strong generalization across various images and objects, opening up new possibilities and avenues for intelligent image analysis and understanding. Actually, a dedicated dataset for pre-training is hard to encompass the vast array of unusual real-world scenarios and imaging modalities, particularly for computer vision community with a variety of conditions (*e.g.*, transparent, low-light, darkness), or employing various input modalities (*e.g.*, depth, thermal), and with numerous real-world applications. Thus, we make an investigation on how well SAM can infer or generalize to diverse scenarios, particularly in transparent objects, complex surroundings and adverse lighting conditions concerned in this thesis.

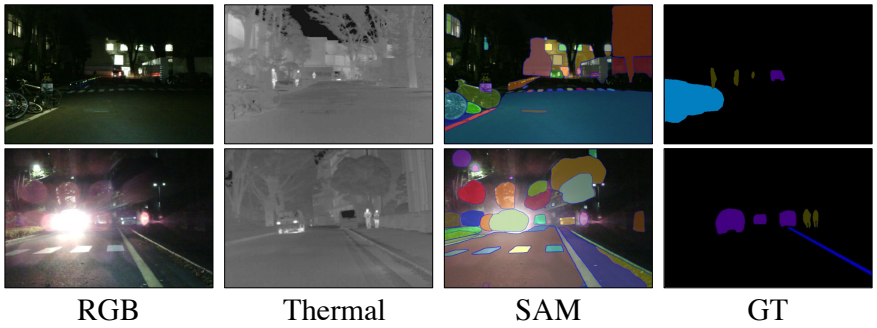
As visualized in Fig. 8.1, we can observe SAM is able to generalize well to



(a) Common Scenarios



(b) Low-contrast/cluttered background, and transparent objects



(c) Poor lighting condition, and glare from the headlights

Figure 8.1: Application of SAM [133] on common scene and complex scene segmentation.

typical natural image scenarios, especially when target regions distinct prominently from their surroundings. This emphasizes the superiority of the promptable SAM's model design and the strength of its massive and diverse training data source. However, when we apply SAM to segment these complex scenes, it is shown that SAM struggles to accurately detect the target objects. This is not surprising as SAM is specifically designed for conventional RGB-based segmentation, relying solely on limited information from single-modality RGB data input. This failure also highlights the importance of advancing segmentation models in complex scene understanding, and the pressing need to ex-

plore multimodal foundational models to fully engage complementary imaging modalities, *e.g.*, depth data and thermal infrared maps.

8.3 Future Work

While we make the substantial strides in this thesis, we acknowledge the complexity of segmentation in real-world environments and the ongoing challenges that remain. Looking forward, we identify several challenges and promising avenues for future research:

(1) Fine-grained unsupervised representation. In Chapter 5, we delve into deep unsupervised representations for RGB-D SOD. Given the sparse nature of pseudo labels in this context, models may struggle to precisely delineate fine-grained object boundaries. One potential approach is to incorporate an auxiliary edge constraint [13], [57], such as integrating an edge detection loss on low-level features to accentuate object details. Further exploration of effective fine-grained constraints or techniques [81] for refining pseudo labels is also promising to enhance segmentation accuracy.

(2) Robustness of missing modalities. In Part I and Part II, leveraging multiple modalities (*e.g.*, depth or thermal) has demonstrated significant benefits in enhancing segmentation performance. However, it is conceivable that not all modalities may be available during test time. Hence, this arises a critical need to enhance the robustness of models in dealing with missing modalities. To adapt this context, we plan to involve multimodal knowledge distillation [149] on model training phase for providing valuable cues for missing modality.

(3) Semi-supervised multimodal learning. We have explored multispectral video input for segmentation task in Chapter 7. However, we note that pixel-wise semantic labeling is very labor-extensive and costly, posing a challenge for large-scale multispectral video annotation. To tackle the label-hungry problem, we plan to introduce semi-supervised learning [203] designed specifically for processing both few labeled and extensive unlabeled multispectral video data.

(4) Advanced network architecture in MVSS. While the engagement of mul-

tispectral videos brings significant improvement as demonstrated in Chapter 7, the research of MVSS is still in its initial stage. By drawing ideas from the well-studied semantic segmentation of RGB images, the accuracy of MVSS model can be further advanced. For example, we may integrate the multi-scale learning technique [15], [171] into cross-spectral and cross-frame fusion to improve the contextual representability of MVSS models. Moreover, delving into the characteristics of the modality, such as the varying intensities of thermal information exhibited by vehicles in different states, represents a promising avenue for future research.

(5) Evaluation metrics in MVSS. Due to the challenging scenes in MVSeg benchmark, the popular TC metric [204] that evaluates temporal consistency based on optical flow warping may not correctly reflect the performance of MVSS models. As studied in [5], the estimated optical flows of complex night-time scenes is not meaningful, which cannot well represent the motions of objects in the scene, *e.g.*, the less-visible driving cars in dim night. Thus, how to design suitable metrics for MVSS is still an open issue.

(6) Long-tailed distribution problem. In the realm of natural images, dealing with a long-tailed distribution of category frequencies in large datasets [8] is a common and unavoidable challenge. In Chapter 6, our approach utilizes a commonly-used re-sampling method [198] to address this issue. Looking ahead, it's worth noting that delving further into the long-tailed problem is a promising avenue for future research.

(7) SAM-driven label-efficient learning. The SAM has demonstrated exceptional performance and versatility, making it a promising tool for various related tasks. Researchers can further leverage pre-trained SAM to empower semi-supervised/weakly-supervised segmentation tasks [205], [206], using the model in combination with suitable point prompts, bounding box prompts, or scribble prompts to generate pseudo-labels.

(8) Foundation models in complex scene. As previously discussed, the current Segment Anything Model (SAM) has achieved significant success in zero-shot image segmentation due to extensive pre-training on web-scale datasets. However, it exhibits limited effectiveness in addressing challenging scenarios

highlighted in this thesis, such as low-contrast scenes and low-light nighttime cases. To enhance the robustness and generalization capability of foundational models, it is intriguing to incorporate complementary sources [2], [4], such as depth and thermal data. We intend to investigate this promising direction in future research endeavors.

(9) Potential advancements enabled by vision-language models. Recent years witness a rapid development in vision-language models [207] (VLMs), prompting consideration of their potential benefits for visual segmentation tasks. One promising direction involves leveraging VLMs to enhance weakly-supervised segmentation tasks [174], wherein image-level tags or captions enabled by VLMs can serve as weak supervision signals for training segmentation networks. This approach avoids the necessity for extensive pixel-wise dense annotations, thereby alleviating huge annotation efforts.

References

[1] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu, and L. Cheng, "Calibrated rgb-d salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9471–9481.

[2] W. Ji, G. Yan, J. Li, Y. Piao, S. Yao, M. Zhang, L. Cheng, and H. Lu, "Dmra: Depth-induced multi-scale recurrent attention network for rgb-d saliency detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 2321–2336, 2022.

[3] W. Ji, J. Li, Q. Bi, C. Guo, J. Liu, and L. Cheng, "Promoting saliency from depth: Deep unsupervised RGB-d saliency detection," in *International Conference on Learning Representations*, 2022, pp. 1–22.

[4] W. Ji, J. Li, C. Bian, Z. Zhang, and L. Cheng, "Semanticrt: A large-scale dataset and method for robust semantic segmentation in multispectral images," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3307–3316.

[5] W. Ji, J. Li, C. Bian, Z. Zhou, J. Zhao, A. L. Yuille, and L. Cheng, "Multispectral video semantic segmentation: A benchmark dataset and baseline," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1094–1104.

[6] W. Ji, J. Li, Q. Bi, T. Liu, W. Li, and L. Cheng, "Segment anything is not always perfect: An investigation of sam on different real-world applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023, pp. 1–11.

[7] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, 2022.

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[9] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2015, pp. 234–241.

[11] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2013, pp. 2976–2983.

[12] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2343–2350.

[13] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *European Conference on Computer Vision*, 2012, pp. 29–42.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advance in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision*, 2018, pp. 801–818.

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[18] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[19] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015, pp. 1520–1528.

[20] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[22] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *European Conference on Computer Vision*, 2020, pp. 275–292.

[23] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "Cnns-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 48, no. 11, pp. 3171–3183, 2018.

[24] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *IEEE International Conference on Image Processing*, 2019, pp. 1440–1444.

[25] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "PDNet: Prior-model guided depth-enhanced network for salient object detection," in *IEEE International Conference on Multimedia and Expo*, 2019, pp. 199–204.

[26] R. Patterson, L. Moe, and T. Hewitt, "Factors that affect depth perception in stereoscopic displays," *Human Factors*, vol. 34, no. 6, pp. 655–667, 1992.

[27] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[28] W. Williem and I. Kyu Park, "Robust light field depth estimation for noisy scene with occlusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4396–4404.

[29] D. N. Bhat and S. K. Nayar, "Stereo in the presence of specular reflection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1995, pp. 1086–1092.

[30] C. Li, H. Fu, R. Cong, Z. Li, and Q. Xu, "Nui-go: Recursive non-local encoder-decoder network for retinal image non-uniform illumination removal," in *ACM International Conference on Multimedia*, 2020, pp. 1478–1487.

[31] R. S. Allison, B. J. Gillam, and E. Vecellio, "Binocular depth discrimination and estimation beyond interaction space," *Journal of Vision*, vol. 9, no. 1, pp. 10–10, 2009.

[32] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for RGBD salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3927–3936.

[33] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8582–8591.

[34] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2075–2089, 2020.

[35] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, "Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4296–4307, 2020.

[36] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3051–3060.

[37] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 989–1005, 2009.

[38] X. Zhang, X. Li, Y. Feng, H. Zhao, and Z. Liu, "Image fusion with internal generative mechanism," *Expert Systems With Applications*, vol. 42, no. 5, pp. 2382–2391, 2015.

[39] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Ying Yang, "Exploiting global priors for RGB-D saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 25–32.

[40] C. Zhu, G. Li, W. Wang, and R. Wang, "An innovative salient object detection using center-dark channel prior," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2017, pp. 1509–1515.

[41] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *IEEE International Conference on Image Processing*, 2014, pp. 1115–1119.

[42] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017, pp. 5108–5115.

[43] N. Sharma, A. Arora, A. P. Singh, and J. Singh, "The role of infrared thermal imaging in road patrolling using unmanned aerial vehicles," *Unmanned Aerial Vehicle: Applications in Agriculture and Environment*, pp. 143–157, 2020.

[44] C. Yuan, Z. Liu, and Y. Zhang, "Fire detection using infrared images for uav-based forest fire surveillance," in *International Conference on Unmanned Aircraft Systems*, 2017, pp. 567–572.

[45] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, "Visible-thermal uav tracking: A large-scale benchmark and new baseline," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8886–8895.

[46] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "Pst900: Rgb-thermal calibration, dataset and segmentation network," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 9441–9447.

[47] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition"*, 2014, pp. 891–898.

[48] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[49] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion," *IEEE Transactions on Automation Science and Engineering*, 2020.

[50] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, and J. Han, "Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2633–2642.

[51] W. Zhou, S. Dong, C. Xu, and Y. Qian, "Edge-aware guidance fusion network for rgb thermal scene parsing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[52] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3239–3259, 2021.

[53] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: State of the art," *International Journal of Multimedia Information Retrieval*, vol. 9, no. 3, pp. 171–189, 2020.

[54] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3907–3916.

[55] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 714–722.

[56] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *European Conference on Computer Vision*, 2014, pp. 92–109.

[57] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8779–8788.

[58] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3917–3926.

[59] R. Valenti, N. Sebe, and T. Gevers, "Image saliency by isocentric curvedness and color," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2009, pp. 2185–2192.

[60] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[61] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183–3192.

[62] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5455–5463.

[63] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265–1274.

[64] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 678–686.

[65] Z. Wu, G. Allibert, F. Meriaudeau, C. Ma, and C. Demonceaux, "Hidanet: Rgb-d salient object detection via hierarchical depth awareness," *IEEE Transactions on Image Processing*, vol. 32, pp. 2160–2173, 2023.

[66] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for RGB-D salient object detection," in *European Conference on Computer Vision*, 2020, pp. 665–681.

[67] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for RGB-D salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3528–3542, 2021.

[68] J. Li, W. Ji, S. Wang, W. Li, and L. Cheng, "Dvsod: Rgb-d video salient object detection," in *Advances in Neural Information Processing Systems*, 2023, pp. 8774–8787.

[69] M. Zhang, S. Yao, B. Hu, Y. Piao, and W. Ji, "C2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection," *IEEE Transactions on Multimedia*, vol. 25, pp. 5142–5154, 2023.

[70] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.

[71] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2274–2285, 2017.

[72] M. Lee, C. Park, S. Cho, and S. Lee, "Spsn: Superpixel prototype sampling network for rgb-d salient object detection," in *European Conference on Computer Vision*, Springer, 2022, pp. 630–647.

[73] C. Chen, J. Wei, C. Peng, and H. Qin, "Depth-quality-aware salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 2350–2363, 2021.

[74] A. Luo, X. Li, F. Yang, Z. Jiao, H. Cheng, and S. Lyu, "Cascade graph neural networks for RGB-D salient object detection," in *European Conference on Computer Vision*, 2020.

[75] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang, "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–13, 2020.

[76] G. Liao, W. Gao, Q. Jiang, R. Wang, and G. Li, "MMNet: Multi-stage and multi-scale fusion network for RGB-D salient object detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2436–2444.

[77] J. Zhao, Y. Zhao, J. Li, and X. Chen, "Is depth really necessary for salient object detection?" In *ACM International Conference on Multimedia*, 2020.

[78] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *European Conference on Computer Vision*, Springer, 2012, pp. 101–115.

[79] D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4048–4056.

[80] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley, "Deep unsupervised saliency detection: A multiple noisy labeling perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9029–9038.

[81] T. Nguyen, M. Dax, C. K. Mummadi, N. Ngo, T. H. P. Nguyen, Z. Lou, and T. Brox, "DeepUSPS: Deep robust unsupervised saliency prediction via self-supervision," in *Advances in Neural Information Processing Systems*, 2019, pp. 204–214.

[82] K.-J. Hsu, Y.-Y. Lin, Y.-Y. Chuang, *et al.*, "Co-attention cnns for unsupervised object co-segmentation," in *International Joint Conferences on Artificial Intelligence*, 2018, pp. 748–756.

[83] K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, X. Qian, and Y.-Y. Chuang, "Unsupervised cnn-based co-saliency detection with graphical optimization," in *European Conference on Computer Vision*, 2018, pp. 485–501.

[84] C.-C. Tsai, K.-J. Hsu, Y.-Y. Lin, X. Qian, and Y.-Y. Chuang, "Deep co-saliency detection via stacked autoencoder-enabled fusion and self-trained cnns," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1016–1031, 2019.

[85] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[86] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[87] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[88] J. Miao, Y. Wei, Y. Wu, C. Liang, G. Li, and Y. Yang, "Vspw: A large-scale dataset for video scene parsing in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4133–4143.

[89] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 633–641.

[90] M. Zhang, J. Li, W. Ji, Y. Piao, and H. Lu, "Memory-oriented decoder for light field salient object detection," *Advance in Neural Information Processing Systems*, vol. 32, pp. 898–908, 2019.

[91] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7254–7263.

[92] J. Li, W. Ji, M. Zhang, Y. Piao, H. Lu, and L. Cheng, "Delving into calibrated depth for accurate rgb-d salient object detection," *International Journal of Computer Vision*, vol. 131, no. 4, pp. 855–876, 2023.

[93] M. Paul, C. Mayer, L. V. Gool, and R. Timofte, "Efficient video semantic segmentation with labels propagation and refinement," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2873–2882.

[94] J. Li, T. Yang, W. Ji, J. Wang, and L. Cheng, "Exploring denoised cross-video contrast for weakly-supervised temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 914–19 924.

[95] M. Zhang, J. Liu, Y. Wang, Y. Piao, S. Yao, W. Ji, J. Li, H. Lu, and Z. Luo, "Dynamic context-sensitive filtering network for video salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1553–1563.

[96] M. Arsalan, M. Owais, T. Mahmood, S. W. Cho, and K. R. Park, "Aiding the diagnosis of diabetic and hypertensive retinopathy using artificial intelligence-based semantic segmentation," *Journal of clinical medicine*, vol. 8, no. 9, p. 1446, 2019.

[97] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[98] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng, "Learning calibrated medical image segmentation via multi-rater agreement modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 341–12 351.

[99] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015, pp. 1377–1385.

[100] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 1451–1460.

[101] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[102] H. Zhang, H. Zhang, C. Wang, and J. Xie, "Co-occurrent features in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 548–557.

[103] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *European Conference on Computer Vision*, 2020, pp. 173–190.

[104] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *European Conference on Computer Vision*, 2018, pp. 405–420.

[105] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," in *European Conference on Computer Vision*, 2018, pp. 269–284.

[106] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1925–1934.

[107] M. Zhang, W. Ji, Y. Piao, J. Li, Y. Zhang, S. Xu, and H. Lu, "Lfnet: Light field fusion network for salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 6276–6287, 2020.

[108] S. Liu, D. Huang, *et al.*, "Receptive field block net for accurate and fast object detection," in *European Conference on Computer Vision*, 2018, pp. 385–400.

[109] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3684–3692.

[110] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *European Conference on Computer Vision*, 2018, pp. 3–19.

[111] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 603–612.

[112] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1857–1866.

[113] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[114] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[115] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Advance in Neural Information Processing Systems*, vol. 34, 2021, pp. 12 077–12 090.

[116] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[117] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, 2019.

[118] I. Ulku and E. Akagündüz, "A survey on deep learning-based architectures for semantic segmentation on 2d images," *Applied Artificial Intelligence*, pp. 1–45, 2022.

[119] F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, J. Chen, J. Hu, X. Guo, and T. L. Lam, "Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021, pp. 4467–4473.

[120] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[121] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 7790–7802, 2021.

[122] J. Liu, W. Zhou, Y. Cui, L. Yu, and T. Luo, "Gcnet: Grid-like context-aware network for rgb-thermal semantic segmentation," *Neurocomputing*, vol. 506, pp. 60–67, 2022.

[123] H. Zhou, C. Tian, Z. Zhang, Q. Huo, Y. Xie, and Z. Li, "Multispectral fusion transformer network for rgb-thermal urban scene semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[124] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advance in Neural Information Processing Systems*, 2017.

[125] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representation*, 2020.

[126] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "RGB-D salient object detection: A survey," *Computational visual media*, pp. 1–33, 2021.

[127] Z. Chen, R. Cong, Q. Xu, and Q. Huang, "DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection," *IEEE Transactions on Image Processing*, 2020.

[128] S. Chen and Y. Fu, "Progressively guided alternate refinement network for RGB-D salient object detection," in *European Conference on Computer Vision*, 2020, pp. 520–538.

[129] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 756–13 765.

[130] M. Zhang, Y. Zhang, Y. Piao, B. Hu, and H. Lu, "Feature reintegration over differential treatment: A top-down and adaptive fusion network for RGB-D salient object detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 4107–4115.

[131] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognition*, vol. 86, pp. 376–385, 2019.

[132] G. Li, Z. Liu, and H. Ling, "Icnet: Information conversion network for rgb-d based salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 4873–4884, 2020.

[133] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[134] L. Tang, H. Xiao, and B. Li, "Can sam segment anything? when sam meets camouflaged object detection," *arXiv preprint arXiv:2304.04709*, 2023.

[135] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: An experimental study," *Medical Image Analysis*, vol. 89, p. 102 918, 2023.

[136] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8554–8564.

[137] R. Mechrez, E. Shechtman, and L. Zelnik-Manor, "Saliency driven image manipulation," *Machine Vision and Applications*, vol. 30, no. 2, pp. 189–202, 2019.

[138] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," *International Conference on Machine Learning*, pp. 597–606, 2015.

[139] N. Liu, L. Li, W. Zhao, J. Han, and L. Shao, "Instance-level relative saliency ranking with graph reasoning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8321–8337, 2021.

[140] L. Yang, J. Han, D. Zhang, N. Liu, and D. Zhang, "Segmentation in weakly labeled videos via a semantic ranking and optical warping network," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4025–4037, 2018.

[141] L. Shao and M. Brady, "Specific object retrieval based on salient regions," *Pattern Recognition*, vol. 39, no. 10, pp. 1932–1948, 2006.

[142] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "RGB-D salient object detection with cross-modality modulation and selection," in *European Conference on Computer Vision*, 2020.

[143] S. Giancola, M. Valenti, and R. Sala, *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies*. Springer, 2018.

[144] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1780–1789.

[145] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3487–3495.

[146] H. Chen, Y. Deng, Y. Li, T.-Y. Hung, and G. Lin, "RGBD salient object detection via disentangled cross-modal fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 8407–8416, 2020.

[147] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate rgb-d salient object detection via collaborative learning," in *European Conference on Computer Vision*, 2020, pp. 52–69.

[148] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[149] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9060–9069.

[150] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 454–461.

[151] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *International Conference on Internet Multimedia Computing and Service*, 2014, pp. 23–27.

[152] R. Achanta, S. S. Hemami, F. J. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[153] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014, pp. 248–255.

[154] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[155] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 4558–4567.

[156] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 698–704.

[157] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9788–9798.

[158] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Processing Letters*, vol. 23, no. 6, pp. 819–823, 2016.

[159] C. Zhu, G. Li, X. Guo, W. Wang, and R. Wang, "A multilayer backpropagation saliency detection algorithm based on depth mining," in *International Conference on Computer Analysis of Images and Patterns*, 2017, pp. 14–23.

[160] H. Chen and Y. Li, "Three-stream attention-aware network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2825–2835, 2019.

[161] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3052–3062.

[162] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for RGB-D saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3472–3481.

[163] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time rgb-d salient object detection," in *European Conference on Computer Vision*, 2020, pp. 646–662.

[164] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for RGB-D salient object detection," in *European Conference on Computer Vision*, 2020, pp. 235–252.

[165] M. Zhang, S. X. Fei, J. Liu, S. Xu, Y. Piao, and H. Lu, "Asymmetric two-stream architecture for accurate RGB-D saliency detection," in *European Conference on Computer Vision*, 2020, pp. 374–390.

[166] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R^3Net: Recurrent residual refinement network for saliency detection," in *Proceedings of International Joint Conference on Artificial Intelligenc*, 2018, pp. 684–690.

[167] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3089–3098.

[168] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 202–211.

[169] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, pp. 802–810, 2015.

[170] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015, pp. 1026–1034.

[171] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *International Conference on Learning Representations*, 2016.

[172] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014, pp. 818–833.

[173] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2806–2813.

[174] J. Li, W. Ji, Q. Bi, C. Yan, M. Zhang, Y. Piao, H. Lu, *et al.*, "Joint semantic mining for weakly supervised rgb-d salient object detection," *Advance in Neural Information Processing Systems*, vol. 34, pp. 11 945–11 959, 2021.

[175] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3203–3212.

[176] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[177] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2814–2821.

[178] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2334–2342.

[179] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 110–119.

[180] F. Liang, L. Duan, W. Ma, Y. Qiao, Z. Cai, and L. Qing, "Stereoscopic saliency model using contrast and depth-guided-background prior," *Neurocomputing*, vol. 275, pp. 2227–2238, 2018.

[181] J. Guo, T. Ren, and J. Bei, "Salient object detection for RGB-D image via saliency evolution," in *IEEE International Conference on Multimedia and Expo*, 2016, pp. 1–6.

[182] B. Maheswari and S. Reeja, "Thermal infrared image semantic segmentation for night-time driving scenes based on deep learning," *Multimedia Tools and Applications*, pp. 1–26, 2023.

[183] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 162–182, 2007.

[184] INO, *Video analytics dataset*, `https://www.ino.ca/en/technologies/video-analytics-dataset/`, 2012.

[185] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3496–3504.

[186] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *European Conference on Computer Vision*, 2018, pp. 186–202.

[187] W. Zhou, X. Lin, J. Lei, L. Yu, and J.-N. Hwang, "Mffenet: Multiscale feature fusion and enhancement network for rgb-thermal urban road scene parsing," *IEEE Transactions on Multimedia*, 2021.

[188] W. Zhou, Y. Lv, J. Lei, and L. Yu, "Embedded control gate fusion and attention residual learning for rgb-thermal urban scene parsing," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[189] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *European Conference on Computer Vision*, 2020, pp. 561–577.

[190] R. Gade and T. B. Moeslund, "Thermal cameras and applications: A survey," *Machine Vision and Applications*, vol. 25, no. 1, pp. 245–262, 2014.

[191] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[192] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "Rgb-t object tracking: Benchmark and baseline," *Pattern Recognition*, vol. 96, p. 106 977, 2019.

[193] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1037–1045.

[194] S. Jain, X. Wang, and J. E. Gonzalez, "Accel: A corrective fusion network for efficient semantic segmentation on video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8866–8875.

[195] M. Paul, M. Danelljan, L. Van Gool, and R. Timofte, "Local memory attention for fast video semantic segmentation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021, pp. 1102–1109.

[196] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020, pp. 1597–1607.

[197] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[198] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[199] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9226–9235.

[200] J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, and Y. Jin, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[201] Y. Liu, L. Kong, J. CEN, R. Chen, W. Zhang, L. Pan, K. Chen, and Z. Liu, "Segment any point cloud sequences by distilling vision foundation models," in *Advances in Neural Information Processing Systems*, 2023, pp. 37 193–37 229.

[202] J. Ma and B. Wang, "Towards foundation models of biological image segmentation," *Nature Methods*, vol. 20, no. 7, pp. 953–955, 2023.

[203] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2613–2622.

[204] Y. Liu, C. Shen, C. Yu, and J. Wang, "Efficient semantic video segmentation with per-frame inference," in *European Conference on Computer Vision*, 2020, pp. 352–368.

[205] C. He, K. Li, Y. Zhang, G. Xu, L. Tang, Y. Zhang, Z. Guo, and X. Li, "Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping," in *Advances in Neural Information Processing Systems*, 2023, pp. 30 726–30 737.

[206] X. Yang and X. Gong, "Foundation model assisted weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 523–532.

[207] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 13 041–13 049.