CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Multi-Level Cross-Modal Features Representation and Fusion Network for Object Detection using Deep Learning

by

Samra Kanwal

A dissertation submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Engineering
Department of Electrical Engineering

2024

Multi-Level Cross-Modal Features Representation and Fusion Network for Object Detection using Deep Learning

By

Samra Kanwal

(DEE181002)

Dr. Andrew Ware, Professor

University of South Wales, UK

(Foreign Evaluator 1)

Dr. José Valente de Oliveria, Senior Researcher

University of Lisboa, Portugal

(Foreign Evaluator 2)

Dr. Imtiaz Ahmad Taj

(Research Supervisor)

Dr. Noor Muhammad Khan

(Head, Department of Electrical Engineering)

Dr. Imtiaz Ahmad Taj

(Dean, Faculty of Engineering)

DEPARTMENT OF ELECTRICAL ENGINEERING

CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

ISLAMABAD

2024

Copyright © 2024 by Samra Kanwal

All rights reserved. No part of this dissertation may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.

To my beloved parents, for your unwavering love and endless support, this scholarly work is dedicated to you with heartfelt gratitude.

CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD

Expressway, Kahuta Road, Zone-V, Islamabad
Phone:+92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: https://www.cust.edu.pk

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the dissertation, entitled "**Multi-Level Cross-Modal Features Representation and Fusion Network for Object Detection using Deep Learning**" was conducted under the supervision of **Dr. Imtiaz Ahmed Taj**. No part of this dissertation has been submitted anywhere else for any other degree. This dissertation is submitted to the **Department of Electrical Engineering, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Electrical Engineering.** The open defence of the dissertation was conducted on **July 02, 2024**.

Student Name : Samra Kanwal (DEE181002)

The Examination Committee unanimously agrees to award PhD degree in the mentioned field.

Examination Committee :

(a) External Examiner 1: Dr. Abdul Ghafoor
Professor
MCS, NUST, Islamabad

(b) External Examiner 2: Dr. Usman Akram
Professor
CEME, NUST, Islamabad

(c) Internal Examiner : Dr. Nadeem Anjum
Professor
CUST, Islamabad

Supervisor Name : Dr. Imtiaz Ahmad Taj
Professor
CUST, Islamabad

Name of HoD : Dr. Noor Muhammad Khan
Professor
CUST, Islamabad

Name of Dean : Dr. Imtiaz Ahmad Taj
Professor
CUST, Islamabad

AUTHOR'S DECLARATION

I, **Samra Kanwal** (Registration No. **DEE181002**), hereby state that my dissertation titled, "**Multi-Level Cross-Modal Features Representation and Fusion Network for Object Detection using Deep Learning**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my PhD Degree.

(**Samra Kanwal**)

Dated: 02 July, 2024

Registration No : DEE181002

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the dissertation titled "**Multi-Level Cross-Modal Features Representation and Fusion Network for Object Detection using Deep Learning**" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete dissertation has been written by me.

I understand the zero-tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled dissertation declare that no portion of my dissertation has been plagiarized and any material used as reference is properly referred/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled dissertation even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD degree and that HEC and the University have the right to publish my name on the HEC/ University Website on which names of students are placed who submitted plagiarized dissertation.

(**Samra Kanwal**)

Dated: 02 July, 2024

Registration No : DEE181002

List of Publications

It is certified that following publication(s) have been made out of the research work that has been carried out for this dissertation:-

Journals

1. **S. Kanwal** and I.A. Taj, "CVit-Net: A conformer driven RGB-D salient object detector with operation-wise attention learning," *Expert Systems with Applications*, vol. 225, pp. 120075, 2023.
2. **S. Kanwal** and I.A. Taj, "Incomplete RGB-D Salient Object Detection: Conceal, Correlate and Fuse," *Pattern Recognition*, vol. 155, pp. 110700, 2024.

(Samra Kanwal)
Registration No: DEE181002

Acknowledgement

All praises be to Allah, the Most Gracious, the Most Merciful, whose divine guidance and blessings have illuminated my path throughout this journey.

I am deeply grateful to my husband, Hasnain Iftikhar, without whom this research would not have been possible. Your unwavering love, understanding, and encouragement kept me motivated during the entire course of this study. Your patience, support, and belief in my abilities have been my greatest source of strength.

To my dear sister, Sahar Afshan, thank you for being my confidante and providing me with valuable insights and feedback throughout this academic pursuit. Your encouragement and enthusiasm inspired me to push my boundaries.

I extend my sincere appreciation to my supervisor, Dr. Imtiaz Ahmad Taj, for his invaluable guidance, expertise, and mentorship. His continuous support, constructive feedback, and belief in my potential have been instrumental in shaping this research.

I am forever indebted to all my teachers, and I acknowledge Allah's divine plan in placing them in my educational journey. Their teachings have enriched my knowledge and nurtured my growth.

I am thankful to Mr. Mohsin Ullah for his support and valuable suggestions. I would also like to thank all the other members of Vision and Pattern Recognition (VisPRS) research group for their support and guidance. Their diverse perspectives and shared experiences enriched my research and made this journey more enjoyable. Finally, I would like to thank Higher Education Commission for funding my PhD studies under Indigenous Ph.D. Fellowship Program and giving me an opportunity to enhance my research skills and qualification.

(Samra Kanwal)

Abstract

Salient object detection (SOD) models are trained to recognize and locate the most conspicuous object(s) that capture human attention. The growing potential of multi-modalities based deep learning methods has motivated computer vision community to build multi-modal saliency detection models. Recently, the availability of depth sensors such as Microsoft Kinect and Time-of-Flight etc. has contributed to the accessibility of depth images at a low cost, therefore, depth is added as an additional modality with RGB modality for saliency detection. However, the rise of RGB-D salient object detection has introduced several challenges such as (i) What are the real complementarity of RGB and depth modalities and how to determine them, (ii) Is low quality depth images degrade the saliency detection accuracies, if yes, then how to handle it and (iii) How to fuse RGB and depth at multi-levels of abstraction. This dissertation is dedicated to two distinct depth quality-aware RGB-D saliency detection learning framework. The first contribution develops a depth quality-aware salient object detection model, that utilizes the synergies of RGB and depth modalities and assesses the depth quality explicitly. The proposed model considers certain design choices. Specifically, the two parallel streams hybrid Conformer backbone network is used. Keeping in view the intrinsic characteristics of RGB and depth modality, the CNN stream of Conformer is fed with RGB modality while Transformer stream of Conformer is fed with depth modality. It is explored that shallow layers of backbone delineate modality-specific features while deep layers possess task-oriented characteristics. Consequently, shallow and deep features are handled separately in Local Detail Enhancement Module (LDE) and Global Detail Enhancement Module (GDE), respectively. The intra-modality features in LDE are utilized to investigate the depth quality in a supervised manner. A novel operation-wise shuffle channel attention module is proposed that correlates the edge maps of RGB and depth with saliency edge map. GDE module captures the global context using proposed reverse attention and presents affluent saliency cues. A light-weight decoder combines the outputs of LDE and GDE for final saliency map. Proposed model is tested on six benchmark datasets and result evaluation shows notable performance gain against

22 state-of-the-art (SOTA) RGB-D SOD models. The imperfections in depth images are further investigated and it was figured out that some depth images are so blurred that they need to be discarded. Earlier approaches, rectify severely noisy depth images by adding some depth correction weights or by augmenting estimated depth with raw depth. These methods rely on RGB images for depth rectification, thereby, lack generalization, specifically for low contrast and cluttered background RGB images. Contrary to former methods, a novel incomplete RGB-D modality learning paradigm is formulated. Depth quality assessment model detects and discards the low quality depth leaving to missing depth images for some corresponding RGB. This is believed that this is first saliency detection framework that works with missing depth. Therefore, it can also be applied to scarcity of depth modality in contrast to RGB modality. A robust RGB-D common latent correlation model is proposed which follows three steps. (i) Distinct conceal blocks to capture low-level (i.e. modality-specific) and high level (i.e. saliency) features in Shallow Common Latent Representation (SCLR) block and Deep Common Latent Representation (DCLR) block, respectively. (ii) Correlate block that presents common latent correlation representation for RGB and depth. (iii) And multi-level fusion block for final saliency generation. A thorough validation against 14 latest state-of-the-art (SOTA) models demonstrate the supremacy of proposed model.

Contents

List of Figures

List of Tables

Abbreviations

ASPP	Atrous Spatial Pyramid Pooling
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DL	Deep Learning
E-measure	Enhanced Alignment Measure
FC	Fully Connected
FCN	Fully Convolution Neural Networks
FLOPS	Floating Point Operations per Second
FN	False Negative
FP	False Positive
FPN	Feature Pyramid Networks
FPS	Frames per Second
GAN	Generative Adversarial Network
GAP	Global Average Pooling
GPU	Graphics Processing Unit
GT	Ground Truth
JL-DCF	Joint Learning and Densely Cooperative Fusion
KD	Knowledge Distillation
LiDAR	Light Detection and Ranging
MAE	Mean Absolute Error
MHSA	Multi-Head Self-Attention
ML	Machine Learning
MLP	Multi-layer Perceptron
MOS	Mean Opinion Score

MSE	Mean Squared Error
OS	Operating System
PLCC	Pearson Linear Correlation Coefficient
PR	Precision and Recall
RAM	Random-access Memory
RMSE	Root Mean Squared Error
ReLU	Rectified Linear Unit
RGB-D	Red Green Blue-Depth
RNN	Recurrent Neural Network
SIP	Salient Person
S-measure	Structure Measure
SOD	Salient Object Detection
SOTA	State-of-the-art
SRCC	Spearman Rank Order Correlation Coefficient
ViT	Visual Transformer

Chapter 1

Introduction

1.1 Background

Object detection is a seminal task within the realm of computer vision that deals with the identification and localization of objects within an image or video. Salient object detection (SOD) is a specialized area of object detection task that deals with identifying and locating objects that garners the greatest human visual attention [1]. Unlike generic object detection, which aims to find all objects in a scene, SOD aims at highlighting visually prominent object. There are vast number of applications in computer vision that utilize saliency detection frameworks as pre-processing step, such as, image retrieval [2], object tracking [3, 4], object segmentation [5], image understanding [6], action recognition [7, 8], image captioning [9], medical image segmentation [10–12], camouflaged object detection [13], person re-identification [14, 15] and video summarization [16] etc. The recognition task in SOD models need to detect one or multiple salient objects in image while localization task need to segment out the detected objects with precise boundaries [17]. For image-based salient object detection model formulation, a saliency map $S \in [0,1]^{W \times H}$ is generated by feeding an input image $I \in \mathbb{R}^{C \times W \times H}$ to a model f. Where, C, W and H represents number of channels, width of image and height of image respectively. The goal of training of model f is to minimize the distance d between S and provided ground truth $G \in [0,1]^{W \times H}$. Broadly speaking, SOD

models are categorized into two: (i) Conventional SOD models and (ii) Deep-learning based SOD models. A good saliency detection model should possess the following criteria:

1. Detection Accuracy: low probabilities of missing salient object and minimum chances of false positive prediction.
2. Precise localization of salient object.
3. Resource efficient.

Generally, a salient region is defined as a segment of an image that stands out predominantly from its surroundings, capturing the attention of a human observer [18]. Conventional SOD models utilize heuristic details, such as contrast, focus, spatial distribution etc. for salient object detection [19–22]. Mainly, RGB (Red-Green-Blue) visual image displays distinctive patterns, color distribution and textural details for various objects in an image. Traditional SOD models extract informative hand-crafted patterns from RGB image. The prediction accuracies of these models are unsatisfactory, specifically, for complex scenarios [18]. With the advancement in deep learning, remarkable progress in SOD technology is observed. The deep SOD models can handle millions of parameters, hence, can be effectively applied on challenging scenes. Initial deep models were designed for single modality [23]. In recent years, it was observed that richer information can be extracted from multiple modalities and noticeable performance gain can be achieved with multi-modal system development. For RGB salient object detection task, various factors (such as low contrast, varying illumination conditions, appearance changes, indistinguishable foreground etc.) pose challenge in detection model [18, 24–26]. To overcome these challenges, data from other complementary modalities (such as depth, thermal, audio, LiDAR etc.) can be utilized. In recent studies, depth (D) modality is established as a RGB complementary source to address SOD's fundamental problems [1, 24, 26, 27]. The intrinsic characteristics obtained from depth modality includes distance of objects from viewpoint, edges, spatial and geometrical details of objects. Hence, endorsement of depth modality in RGB-D SOD

frameworks helps to discern foreground from complex background. The rationale behind the endorsement of multi-modality data is to exploit the synergies of RGB and depth [18]. The joint learning of RGB and depth provides a robust detection model especially in challenging scenes such as illumination changes, cluttered background, indistinguishable patterns etc. [25].

1.1.1 Evolution of RGB-D Salient Object Detection Techniques

RGB-D Salient Object Detection deals with extracting, processing and fusing feature representations from RGB and depth modality in a such a way that the commonalities and complementarities of two modalities are exploited to strengthen the detection performance. Traditional models used handcrafted features extracted from RGB and depth modalities such as 3D structure and shape [19], depth contrast [20, 21, 28] and contextual contrast [22] etc. Conventional RGB-D SOD models rely on low-level features and lack to capture global context required for saliency detection task. Therefore, with the advancement in deep learning the cumbersome task of feature extraction is vanished and millions of parameters are trained to solve complex task. Eventually, deep learning based RGB-D SOD models needs to perform two tasks (i) multi-modal feature alignment and (ii) multi-modal feature fusion. Effective performance of RGB-D SOD models heavily depends on backbone network. Most studies used convolutional neural network (CNN) based backbone such as VGG [29] and ResNet [30] etc. while other adopted Transformer based models such as Swin Transformer [31], T2T-ViT [32] etc.

Mainly three types of fusion strategies are incorporated in detection models.
Early Fusion: Low-level features of each modality is fused in initial stage of detection model. Early fusion either concatenate all modalities in channel dimension before feeding input to model [33, 34] or fuse the extracted low-level features from shallow layers of backbone network [35]. Two early fusion schemes are presented in Fig. 1.1. This fusion strategy is mainly adopted in single-stream networks. Although early fusion is a resource efficient scheme, however, lacks uniformity due

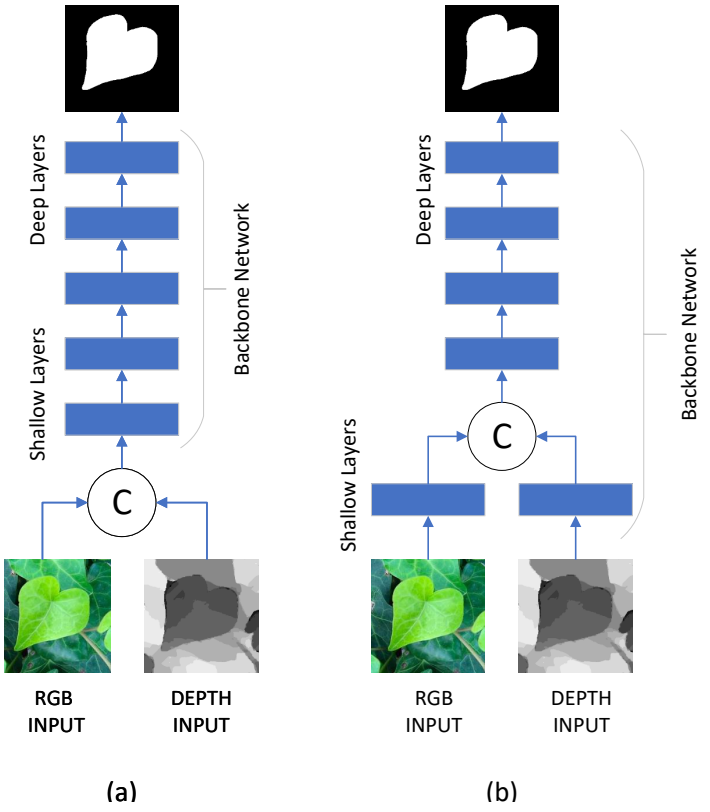


FIGURE 1.1: Early Fusion: (a) Concatenation of RGB-D Inputs (b) Concatenation of shallow RGB-D features.

to diversity of RGB and depth features at shallow level.

Late Fusion: Two stream backbone network is adopted and deep features are fused in later stages of detection model. Late fusion methods either concatenate high level features extracted from backbone network [18, 36] or separate RGB and depth saliency maps are generated and concatenated for final prediction map [37].Two late fusion schemes are presented in Fig. 1.2. These fusion schemes ignore the intrinsic correlation of RGB and depth as modality-specific features are totally ignored.

Multi-Level Fusion: Considering the fact that shallow layers are modality-specific and task agnostic while characteristics of deep layers are opposite to shallow characteristics [1], multi-level fusion scheme came to existence. Therefore, to subjugate the limitations of early and late fusion strategies, fusion of RGB and depth features at multiple levels of backbone network is accomplished. RGB and depth input are fed to separate backbone stream. Then either multi-levels cross-modal features are interacted to form separate saliency maps, which are later on

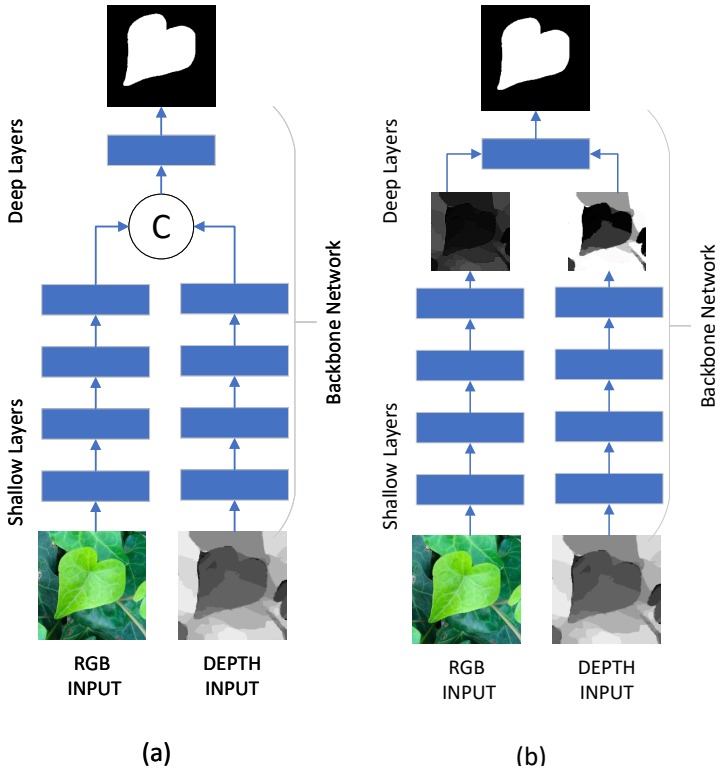


FIGURE 1.2: Late Fusion: (a) Concatenation of deep RGB-D features (b) Concatenation of RGB and depth saliency maps.

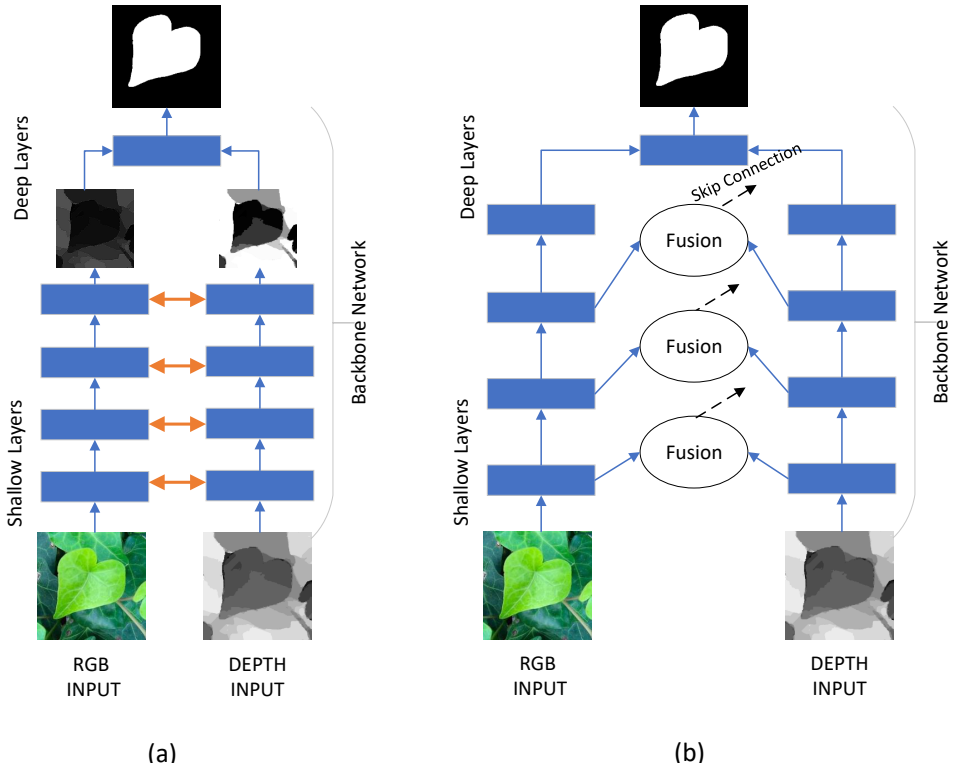


FIGURE 1.3: Multi-Level Fusion: (a) Multi-level interaction of RGB-D features (b) Multi-level fusion of RGB-D features.

fused for final prediction [25] as illustrated in Fig. 1.3 (a) or fusion is performed at multi-levels and combined at decoder [24, 38, 39], demonstrated in Fig. 1.3 (b).

1.1.2 Datasets for Salient Object Detection

With rapid progression in salient object detection frameworks, various datasets have been developed. Previously presented SOD datasets contain simple images which are center biased [40]. Recent datasets lean towards more challenging scenes. Popular datasets are summarized in Table 1.1 Following RGB and RGB-D SOD datasets are adopted by current studies [1, 17, 18, 24–26, 33]:

RGB SOD Datasets: Popular RGB SOD Datasets have included complex background, multiple objects and diverse scenes.

- **DUTS [41]:** This dataset contains 10,553 and 5,019 RGB images with corresponding ground truth saliency maps for training and testing respectively. Train set is formed by collecting images from ImageNet [42] train/val dataset while test samples are from ImageNet-test and SUN [43] datasets.
- **ECSSD [44]:** This dataset contains 1000 images from real-world scenes, having cluttered background.

RGB-D SOD Datasets: Introduction of multi-modal RGB-D SOD frameworks, directed the construction of RGB-D datasets. Popular RGB-D datasets are given below. They contain RGB and depth pair along with corresponding binary ground truth saliency maps.

- **NJU2K [45]:** This dataset contains 2000 stereoscopic RGB-D pairs. This a diverse dataset with images collected from movies and internet. It also includes images taken using Fuji W3 stereo camera.
- **NLPR [22]:** This dataset comprises of 1000 images with depth images obtained using Microsoft Kinect. Various scenes from outdoor and indoor locations focusing on different scales and contrast are introduced in the dataset.

TABLE 1.1: SOD Benchmark Datasets Characteristics.

Sr. No.	Dataset	Size	Number of Salient Objects	Object Type	Sensor	Resolution
RGB SOD Datasets						
1	DUTS [41]	15,572	Multiple	Indoor/ Outdoor	-	$[100-500] \times [100-500]$
2	ECSSD [44]	1000	Multiple	Outdoor	-	$(139-400) \times (139-400)$
RGB-D SOD Datasets						
1	NJU2K [45]	2000	Single	Movies/ Internet/ Camera pictures	Fuji W3	$[231-1213] \times [274-828]$
2	NLPR [22]	1000	Multiple	Indoor/ Outdoor	Microsoft Kinect	640×480, 480×640
3	LFSD [46]	100	Single	Indoor/ Outdoor	Lytro light field	360×360
4	SIP [33]	929	Multiple	Person	Huawei Mate10	992×744
5	STERE [47]	1000	Single	Internet	Stero	$[251-1200] \times [222-900]$
6	RGBD135 [21]	135	Single	Indoor	Microsoft Kinect	640×480

- **LFSD [46]:** 100 RGB-D pair collected by Lytro light field camerais included in this dataset.
- **SIP [33]:** This dataset includes 929 RGB-D pair taken by Huawei Mate10. Salient-in-Person include multiple human objects in each high-resolution image.
- **STERE [47]:** It includes a collection of 1000 stereoscopic RGB-D image pairs from internet.
- **RGBD135 [21]:** Also known as DES contain 135 indoor images. Depth images are obtained by Microsoft Kinect.

1.1.3 Performance Evaluation Metrics

To evaluate the extent of similarity between the predicted saliency map (S) and ground truth (GT) map, various metrics are used including Precision-Recall Curve (PR-Curve), Structural Measure (S-measure), F-measure, Enhance-alignment Measure (E-measure) and Mean Absolute Error (MAE). In these evaluation metrics, Precision-Recall, F-measure and MAE calculates pixel-level inaccuracies of prediction with respect to GT. While object-level errors are found using S-measure and E-measure. A brief introduction of evaluation metrics are given below.

- **Precision-Recall [48]:**

  To calculate PR a binary mask B is formed from S and following formula is applied.

$$\begin{aligned} Precision &= \frac{B \cap G}{B}, \\ Recall &= \frac{B \cap G}{G}. \end{aligned} \tag{1.1}$$

  A threshold ranging from 0-255 is applied on S and for each thresholded S, PR is calculated. These PR are combined to form a curve as shown in Fig. 1.4.

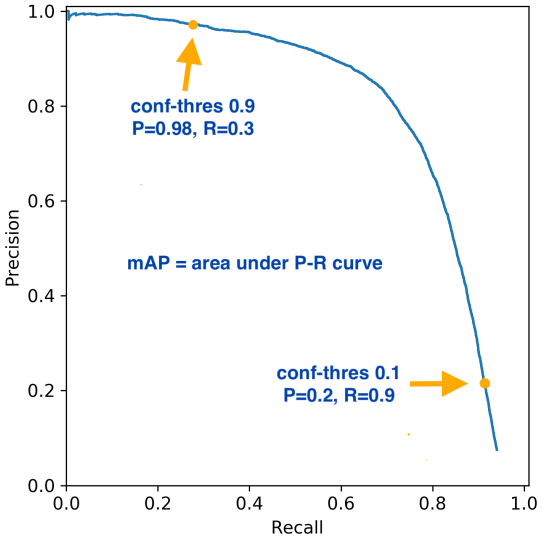


FIGURE 1.4: Precision-Recall Curve

Area under the PR-curve is used to compare different models. High recall means low false negative (FN) rate while high precision relates to low false positive (FP) rate. Thereby, PR-curve represents the trade-off between FN and FP.

- **Structural Measure [49]:**

  The structural similarity between S and G is given by S-measure (S_α). This metrics presents a weighted expression between object-aware structural similarity S_o and region-aware structural similarity S_r.

$$S_\alpha = \alpha S_o + (1 - \alpha) S_r, \tag{1.2}$$

  here, α is balancing factor.

- **F-measure [50]:**

  This metric yields weighted harmonic mean of precision and recall. Different thresholds are applied to obtain set of F-measure, which are used to represent maximum of F-measure (F_β^{max}). In Eq. 1.3, β value 0.3 is chosen to emphasize on precision.

$$F_\beta = \frac{(1 + \beta^2) Precision \times Recall}{\beta^2 \times Precision + Recall}. \tag{1.3}$$

- **E-measure [51]:**

  Enhance-alignment Measure considers image-level statistics and local pixel matching jointly, given in Eq. 1.4

$$E_\phi = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi(i, j), \tag{1.4}$$

  here, ϕ is enhanced-alignment matrix.

- **Mean Absolute Error [52]:**

  It computes pixel-wise mean absolute error between S and G. Unlike PR and

F-measure, MAE considers true negative pixels also.

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S(i,j) - G(i,j)|, \tag{1.5}$$

here, W and H are weight and width of image.

1.1.4 Applications and Impact of Salient Object Detection

Salient object detection plays a vital role as a pre-processing step in wide variety of applications. Various domains need the interpretation of visual content. Salient object detection helps to obtain useful insights from images or videos. Some of the applications of salient object detection in several fields are listed below.

Computer Vision Domain:

A large number of computer vision applications need salient object detection models. Some of these are listed below:

- **Image Retrieval:** In computer vision field, salient object detection is used in image retrieval applications for relevant search results based on detected salient object.
- **Image Captioning:** For image captioning task, capturing salient object helps to obtain descriptive annotations. Image understanding models apply salient object detectors to extract meaningful interpretation of visual images.
- **Surveillance:** In surveillance applications, saliency detection is utilized by object tracking, camouflaged object detection models, person re-identification and object detection frameworks.
- **Medical Imaging:** Salient object detection also contributes in medical imaging field to identify crucial anatomical structures.
- **Segmentation:** Salieny detection enhances the target objects in Unsupervised object segmentation and semantic segmentation applications.

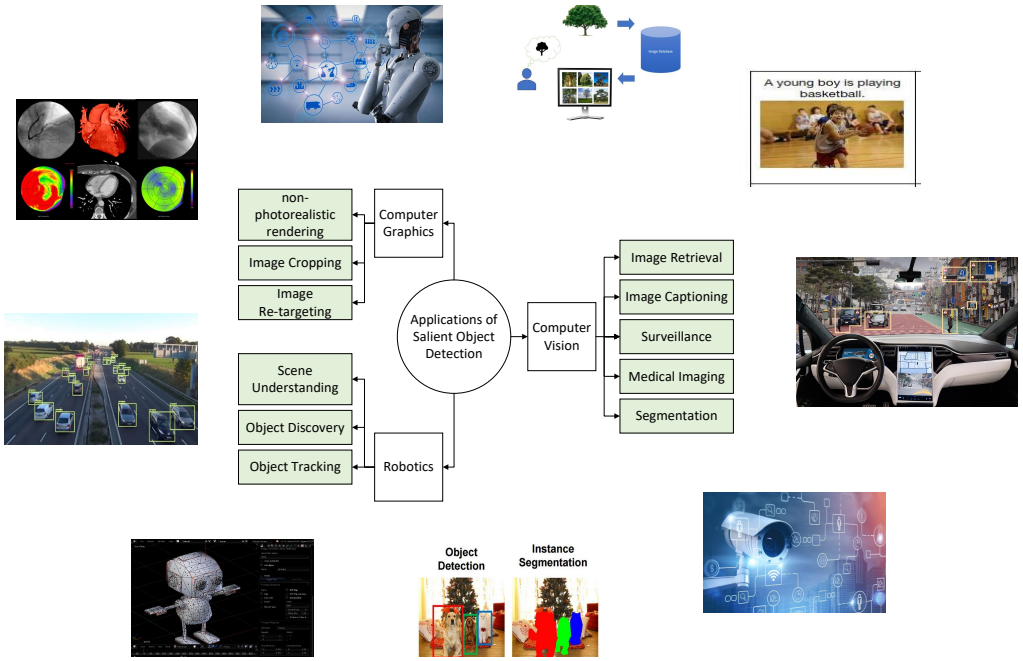


FIGURE 1.5: Applications of Salient Object Detection.

Computer Graphics Domain:

A vast number of computer graphics application opt salient object detectors for non-photorealistic rendering, image cropping, image re-targeting, etc.

Robotics Domain:

In robotics field, it aids to identify hazards thus improves scene understanding. It helps to enhance human-robot interaction. Overall, salient object detection has huge impact in diverse fields.

Fig. 1.5 illustrated the summary of above mentioned applications

1.2 Research Objectives

The SOD models that incorporate depth modality as additional source along with color, presents more precise and robust saliency predictions. However, they are confronted with three major challenges. Firstly, depth images are often contaminated with noise and error during acquisition process. The underlying premise of accurate depth data is impossible in real world due to constraints of depth sensors and scene conditions, eventually, low quality depth has negative effect on performance of detection model. Secondly, intra-modality features of RGB and depth

are very distinct. The real complementarity selection of two modalities is still a challenge. Thirdly, how to effectively fuse inter-modality features. Although RGB gives color/texture details and depth provides geometric cues, however, there exists a strong correlation between two modalities for a given scene. Focus of this research is to develop a multi-modal salient object detection framework capable of extracting advantageous depth cues and fusing real complementarities of RGB and depth.

Main objectives of this research are:

1. Develop a depth quality-aware salient object detection (SOD) framework that explicitly assesses and integrates depth quality by correlating RGB and depth edge maps.
2. Create a robust model for salient object detection that can handle incomplete RGB-D data by:
   (a) Identifying and discarding severely noisy depth images.
   (b) Effectively detecting salient objects with missing depth information or using complete RGB-D pairs.

1.3 Research Contributions

Main contributions of this research work are:

1. Development of depth quality aware salient object detection framework. Considering the fact that edge maps of depth maps can contribute to detect depth quality explicitly. With the increase in distortions, edge details of depth images are lost. Also the misalignment of RGB and depth images also lead to inaccurate predictions. Edge maps of RGB and depth are correlated in a supervised manner for a resilient RGB-D salient object detection model implementation. Evaluation of proposed work demonstrates noticeable performance gain against state-of-the-art SOTA model. My contributions have been published in [53].

2. Development of depth quality assessment regression model, to distinguish good and bad quality depths. This no reference depth quality assessment model assigns a quality score to depth images. Afterwards, a threshold is applied to predict it as good or bad quality image. The depth quality assessment model is also tested on a database for stereoscopic image quality assessment. Our proposed model shows improved performance as compared to SOTA model. My contributions have been published in [54].

3. Development of incomplete RGB-D salient object detection network. Severely noisy depth images are discarded and resultant data contain complete RGB-D pair and RGB with missing depth. A robust representation is formulated that detects salient object exploiting the common latent RGB-D correlation representation. Results shows that proposed model outperforms SOTA models. My contributions have been published in [54].

4. Incomplete RGB-D salient object detection network is applied for the case of severally missing depth modality. So that large amount of RGB data as compared to scarce depth data can be effectively utilized. Results illustrate the robustness of proposed work for severally missing depth. My contributions have been published in [54].

1.4 Thesis Organization

This research thesis is composed of five chapters.

Chapter 1 starts with the introduction of preliminary knowledge about object detection and saliency detection. This is followed by the evolution of multi-model RGB and depth salient object detection. It describes the overview of traditional and deep SOD methods. It also presents different fusion schemes of multi-modal data. It outlines the benchmark datasets used for the implementation and evaluation of SOD models along with performance evaluation metrics. The chapter describes the applications, research objectives and research contributions of the thesis.

Chapter 2 reviews the existing work. The literature survey discusses the current work based on backbone network. It also outlines major fusion strategies of RGB-D salient object detection models. Limitations of existing work are presented. The chapter gives a research gap analysis and problem statement of the proposed model.

Chapter 3 gives a comprehensive framework of the proposed depth-aware saliency detection model and provides its results. A comparison of proposed methodology with SOTA models is presented. It gives comprehensive ablation studies of the proposed model.

Chapter 4 introduces the implementation of incomplete RGB-D salient object detection model. It first presents proposed depth quality assessment regression model, then describes the proposed methodology of common latent RGB-D correlation framework for saliency prediction. The results of incomplete RGB-D salient object detection model are discussed and compared with SOTA models. It also covers comprehensive ablation studies.

Chapter 5 concludes this research thesis and provides future directions.

Chapter 2

Literature Review

A comprehensive literature survey of RGB-D salient object detection is presented in this chapter. Existing approaches are categorized according to conventional/deep models, backbone network architecture, fusion strategies and complementarity cue selection schemes. the limitations of earlier models are also discussed. Chapter also give research gap, problem statement and proposed methodology. Organization of literature review is given in Fig. 2.1.

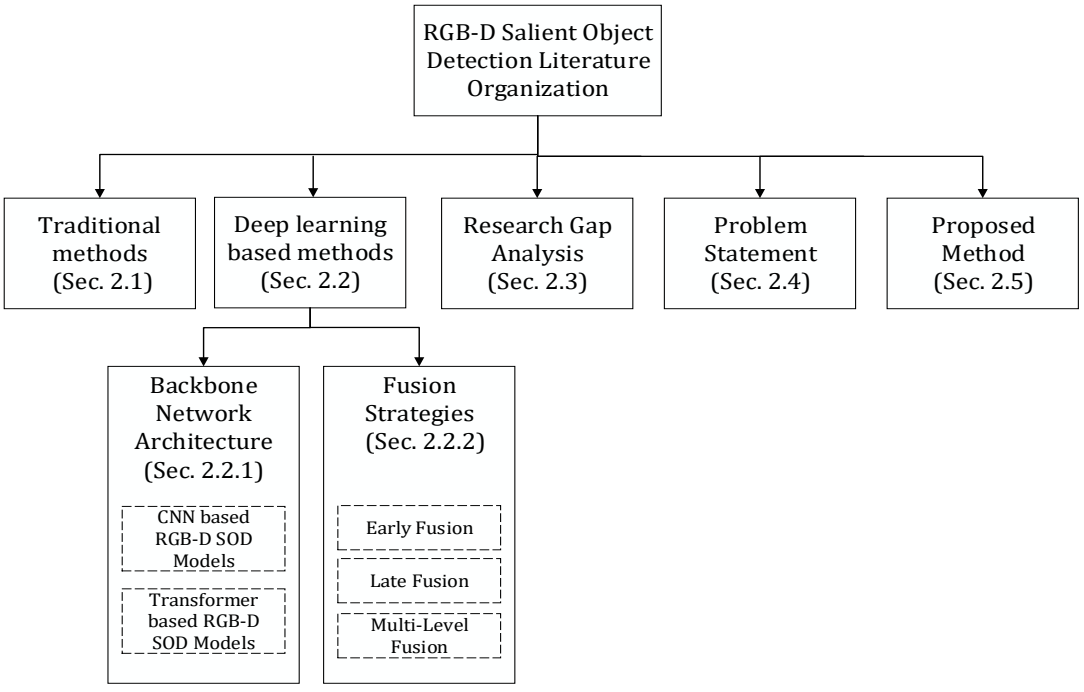


FIGURE 2.1: Organization of Literature Review.

2.1 Traditional Methods

Initially, color and textural patterns from visual data was explored for saliency detection task. Considering that human exploit multiple senses to interpret the information from environment, multi-modality learning approach for machines came into being. With the availability of depth sensors such as Microsoft Kinect and Time-of-Flight, depth images are adjoined with RGB images for salient object detection task. Over the past several years, hand-crafted geometric and appearance cues from depth and RGB modality performed saliency detection task. Notable conventional RGB-D SOD models are categorized according to low-level depth feature measurements, saliency measures and foreground extraction [55]. Fig. 2.2 summarises the methods and brief description is given below.

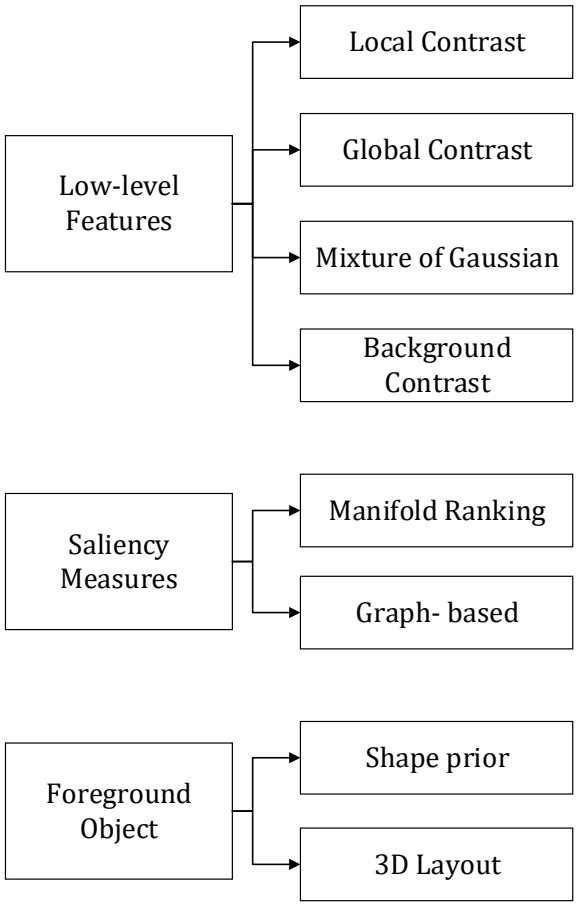


FIGURE 2.2: Conventional RGB-D SOD methods categorization.

Depth prior is used in many SOD models to improve saliency prediction. Lang et al. in [56] integrates depth and saliency cues by learning mixture of Gaussian distribution. Recently, Xiaoqin et al. in [57] utilize depth prior along with gradient and edge prior for multi-modality saliency detection. In [20], region contrast is utilized for 3D saliency and fused with 2D saliency for improved performance. Niu et al. in [47] utilize global disparity contrast. Another work proposed local, global and background contextual contrast based depth saliency method [22]. In [19], structural and shape features are extracted from depth images. In [28], background and orientation prior are used for 3D saliency prediction. Method proposed in [58] exploited local background enclosure instead of depth contrast for salient object detection. Xue et al. proposed different depth features such as distance, spatial location and surface normals etc. and integrate them with RGB in a dual manifold ranking scheme [59]. In [60] graph-based segmentation is used to highlight salient region.

Although, incorporating heuristic depth cues with RGB using traditional methods have proved their efficiency for SOD task, however, these methods lack generalization ability due to limited manifestation of hand-crafted features. With the introduction of AlexNet [61] convolutional neural network (CNN) image classification model, researches started to investigate their usefulness for high-level tasks.

2.2 Deep Learning Based Methods

Deep learning uses multiple interconnected layers of nodes to capture hierarchical representations of data. It mimics the neurological structure of the human brain to interpret data, allowing the construction of computational models. In deep neural networks, feature extraction occurs in an automated manner. Before 2000, the introduced deep neural networks such as Recurrent Neural Networks (RNNs) [62] and LeNet [63] were potentially limited due to hardware constraints. In 2006, Hinton et al. [64] proposed Deep Belief Networks (DBNs) with several Restricted Boltzmann machine (RBM) type structures stacked together for a deep network formation. The first convolutional neural network (CNN) proposed by Krizhevsky

et al. [61] in 2012, opens the door for research community to utilize deep learning based models for complex tasks. Deep learning based methods can learn complex representations and maneuver millions of parameters. CNN consists of convolution and pooling operations along with activation functions. The hierarchical structure of CNN can learn abstract manifestation of input data automatically. Various network architectures VGG [29], ResNet [30], GoogLeNet [65] etc. were developed to automatically learn discriminative features from input data. These network architectures were initially developed for image recognition task but later on adopted for complex high-level computer vision tasks such as object detection, segmentation, image understanding, robotics and surveillance. For salient object detection task, the pre-trained CNN networks are used for feature extraction. The extracted features are enhanced and refined through additional layers.

The recent studies on RGB-D SOD are almost exclusively using deep architectures and optimizing the detection accuracies using differing fusion schemes, attention based modelling and advanced architectures. The subsequent sections discuss deep learning based methods, classified based on feature extraction and fusing strategies.

2.2.1 Backbone Network Architecture

The backbone network plays a crucial role for feature manifestation, thereby, influences the saliency detection performance. At the beginning, several deep learning based multi-modal SOD model was built upon CNN backbone network. However, with the advent of Visual Transformer (ViT) [66], researchers started to investigate Transformer or Hybrid backbone networks. Extensive survey on the adoption and impact of various backbone networks are discussed below.

CNN based RGB-D Salient Object Detection Models

Convolutional neural networks (CNN) provides a strong hierarchical feature representation with shallow layers embedded with modality-specific features while deep

layers presents global context. Initially, CNN based multi-modal fusion is not performed in end-to-end manner. The first systematic SOD model having CNN network was proposed in 2017 [36]. Before that, saliency cues are fed to CNN network or some post-processing is applied. Subsequently, advantages of CNN is not fully exploited. The prominent CNN based RGB-D salient object detection models are presented below

- **LP-Net** This research model highlights RGB and depth features, combines the multi-modality features and feed them to CNN. Afterwards CNN output is refined by Laplacian. This method consists of three steps. In the first step of method, various saliency priors are extracted such as, local contrast, global contrast, background prior and spatial positioning etc. CNN was used only for deep fusion. Just like, traditional methods where several low level saliency cues are used, this method feed CNN with prior welldesigned knowledge. In second step, hyper features are extracted using CNN. The saliency features from first step are fed as input to second step where binary logistic regression problem is formulated and saliency prediction is obtained. The third step is Laplacian propagation, where the spatially inconsistent saliency is mapped to optimal saliency generation [67].

- **CA-Fuse** Earlier methods, either concatenate RGB and depth features or fuse modality-specific predictions. In this model, following a progressive hierarchy a well structured complementarity aware fusion scheme is proposed. This model develops a multi-level supervision scheme, which combines the RGB and depth residual with complementarity-aware supervisions. A backward prediction denseconnection (BPDC) module further adds the global context to saliency detection model [68].

- **CPFP** The intuition behind this research work stemmed from the sub-optimal outcomes observed when employing ImageNet pre-trained backbone network for both RGB and depth modality. Reason of this is the inherent difference of RGB and depth. Therefore it was proposed that contrast prior can be used to enhance the original depth map in a CNN network.

A novel fluid pyramid integration is proposed to fuse cross-modal features. Performance of this method is superior than multi-level fusion [69].

- **Att-CMCL** This model employs a three stream network, one for modality-specific RGB features, second for modality-specific depth features and third stream is distillation stream. Earlier methods implement two stream network for RGB and depth and fuse cross-modalities at later stages. While Att-CMCL focus on complementarity of two modalities in bottom-up and top-down paths. This methods propose a novel channel attention mechanism employed at cross-modal cross-level fusion [70].

- **LSF** This model learns modality-specific representation of RGB and depth via distillation network, then selects real complementarities and fuses in a progressive way. This model trains a RGB teacher network with color images from RGB and RGB-D datasets. Then a hierarchical knowledge distillation scheme is devised to transfer knowledge from pre-trained RGB modality teacher network to depth student network. This approach adopts L2 loss for five levels of hierarchy. Then progressive top-down fusion combines cross-modalities and skip connections for final prediction. This model can be used for zero-shot learning [1].

- **A2dele** This model develops an attentive depth distiller, that transfers saliency information from depth stream to RGB stream. Unlike, custom methods which utilize two independent streams for RGB and depth modalities, this model propose that depth should be learned earlier so that depth ambiguity during testing can be hindered. In test phase, only RGB input can be used. It remedies low quality depth with small number of parameters. Depth stream is trained using VGG encoder and privileged knowledge is transferred to RGB stream. RGB stream acts as student network and adaptive distillation is used to mitigate negative impact of low quality depth [71] .

- **CIR-Net** This model proposes an encoder decoder architecture, where encoder consists of attention guided interaction module between RGB and

depth modalities. Decoder consists of convergence aggregation via gated fusion. A middleware enhancement structure is constructed between encoder and decoder. CIR-Net is a three stream network. RGB and depth encoder streams learn features from ResNet50 backbone network and extracted multi-level features are fed to progressive attention-guided integration unit and third stream i.e. RGB-D stream is generated. Further to reduce redundancy, middleware refinement is used for top encoder features. In decoder stage, convergence aggregation structure combines the RGB and depth to RGB-D decoding stream in progressive manner. Although middleware structure is pluggable to three stream network, however, model complexity increases [72].

- **LIANet** This model proposes a two stream network for RGB and depth. It consists of encoder with attention to obtain relevant features. The encoded features are enhanced for saliency detection via layered interaction fusion module. Model proposes a novel RGB-Depth-RGB modulation feedback mechanism. Two stream network is defined for RGB and depth modality in encoder stage and simple attention is used to enhance channel and spatial features. Layered interaction fusion module is most crucial part of this methodology that interacts encoder and decoder features. The main role of layered interaction fusion module is to filter low-quality depth and enhance RGB object details [73].

- **HiDAnet** This research work develops a novel granularity based attention. In order to effectively capture geometric priors, multi-Otsu method is applied to generate granularity based multiple regions. In this way, multiple objects far from viewpoint can be detected. Furthermore, edges are sharpened with channel attention. Encoder consists of parallel RGB and depth streams for feature extraction. Granularitybased attention is proposed based on the observation that geometric priors defined by multi-granularities correlates with saliency cues at multi-level. Local Efficient Channel Attention (L-ECA) is proposed to generate depth aware encoded features. In decoder

stage, multi-input is adaptively concatenated with efficient channel attention model [74].

- **M^{2}RNet** This work proposes a novel multi-scale refinement network, which consists of nested dual attention module to select informative features. The two adjacent feature maps are combined in adjacent interactive aggregation module and for further saliency supervision loss called joint optimization loss is proposed. Encoder consists of RGB and depth stream for independent feature extraction and combined. RGB-D is fed to nested dual attention module to enhance the features. The coordination of RGB-depth and contamination in depth is handled using channel and spatial attention. Additionally, Adjacent Interactive Aggregation Module focus on difference of deep and shallow features properties [75].

Transformer based RGB-D Salient Object Detection Models

Mostly, RGB-D saliency detection is performed by extracting modality-specific features from independent CNN based backbone streams and then fusing these features for final prediction. Low-level details of any modality can be easily leveraged with CNN but it fails to capture global context. Some methods used dilation convolution to increase the receptive field but the intrinsic nature of CNN cannot fully apprehend long-range dependencies. Global context can help to predict salient regions in an image, subsequently, researchers started to investigate other backbone networks. After the development of Visual Transformer (ViT) [66], many SOD models utilize Transformer as backbone network. Some popular methods are given below.

- **SwinNet** This research model utilizes Swin Transformer [31] for RGB and depth features extraction. CNN is limited by capturing long-range dependencies while transformer is inefficient to represent local details. Swin Transformer combine the CNN and Transformer structure in one framework. Swin Transformer is used as backbone network in two stream structure. The hierarchical multi-modal features are aligned using attention mechanism. Along

with that shallow layers are trained with edge guided module to refine edges [17].

- **VST** This work is based on pure transformer. A novel token upsampling technique is develop to increase the resolution. Token based decoder network performs two tasks. One is to predict salient object using task-related tokens and other is to refine boundaries using patch-task-attention. This model utilize T2T-ViT [32] Transformer as backbone network to obtain RGB and depth encoder patch tokens. Transformer converter then combines the encoder patch tokens. The decoder converts the patch tokens to saliency map [76].

- **TriTransNet** The intuition behind this work is to introduce triplet transformer embedding module in U-Net ResNet network. It consists of transition layer to align features for fusion. These features are fed to triplet transformer embedding module for enhancement. In decoder, enhanced feature and low two layers are combined. TriTransNet consists of three main modules. First the encoder modules purify the low quality depth and performs RGB-D fusion through attention. Second feature enhancement module selects three deep features and align them so that they can be fed to triplet transformer. Last is the three stream decoder that receives the enhanced features from feature enhancement module and combines deep and shallow features for final saliency map [26].

- **MutualFormer** The model performs token mixer for intra-modality features using self-attention and modality-mixer for inter-modality fusion using cross-diffusion attention. Final aggregation results in a modality-invariant saliency output. Based on self similarity, modality-specific tokens are generated and at the same time the correlation of RGB and depth is studied using cross-diffusion attention. Moreover, MutualFormer also enhance the most important features of each modality using Focal Feature Extractor. Due to focus on global context, this model has limitations in capturing local details [77].

- **SiaTrans** This model performs depth image quality classification along with saliency detection task. It adopts Siamese transformer network for weight sharing in both encoder and decoder. Cross modality fusion is designed to choose between RGBD or RGB, in case of poor depth, based on classification. SiaTrans opt encoder decoder architecture. Encoder is developed on Siamese transformer while decoder has CNN architecture. Last layer of backbone generates coarse saliency for both modalities. These two coarse saliencies are combined to form RGB-D coarse saliency in CMF module which consists of attention. It also produces depth quality classification labels. Adaptive attention is used in decoder to fuse the features and form the final prediction [27].

- **CAVER** This research work builds a transformer based information propagation path (TIPP) on CNN backbone. Matrix operation in attention is optimized by Patch-wise token re-embedding. Model predicts the global context in an efficient way. CNN backbone is used to extract features from RGB and depth modality. And the intermediate features are fed to corresponding transformer stage. TIPP consists of cascaded integration units which absorbs the modality-specific local features to obtain global context. CAVER also presents parameter-free patch-wise token re-embedding strategy to reduce complexity [78].

2.2.2 Fusion Strategies

The complementarities of multi-modalities need to be fused efficiently and accurately. Existing work can be categorized in three fusion schemes, as given below:

1. Early Fusion

Early fusion scheme is simplest approach used to combine multi-modalities. As, early fusion concatenates low-level or raw cross-modal features, therefore, it generates a rich representation from diverse data. Although, there exist a strong

correlation between cross-modalities, the intrinsic characteristics of them are heterogeneous. Additionally, a large representation may result degradation of accuracy. A review of popular single stream models are presented below.

- **SSRC** This model exploits a single stream recurrent network. Four channel input is used to generate coarse prediction which is fed back in top-down manner. The top down hierarchy sharpens the boundaries by processing raw depth, current CNN level features and coarse saliency prediction. This early fusion method concatenates the RGB and depth before feeding to backbone network. Then backbone generates a coarse saliency map. Multiple Depth Recurrent Convolution Neural Networks (DRCNN) follow a top-down progressive structure. Each stage of backbone is connected in a side-out fashion with the DRCNN. Each DRCNN takes following inputs: Coarse saliency map, raw depth, prediction from one step forward stage and feature map from current stage. These four inputs are concatenated and recurrent connection learns the features. The saliency maps from each stage is combined for final saliency map [34].

- **D^{3}Net** This work proposes three stream network. Input to three streams are RGB, Depth and fusion of RGB-D. Three feature pyramid networks produce three saliency maps S_{rgb}, S_{depth} and S_{rgbd}. In test phase, a depth depurator unit selects the optimized saliency map and thus filters low quality depth. In test phase, a depth depurator unit (DDU) selects the optimized saliency map and thus filters low quality depth. DDU consists of gate connection and MAE metric is used to find the distance between S_{depth} and S_{rgbd}. A threshold is used to select the multi-modal or RGB stream. The three stream network increases the model complexity. This research work develops a new dataset SIP of type 'person in wild' using Huawei Mate10 sensor. It contains high quality depth images [33].

- **UCNet** In this research work, uncertainty is captured in human annotations by generating a set of saliency maps instead of single prediction. The

model utilizes generative adversarial network (GAN). The training and testing pipelines are different. Training pipeline consists of generator and inference module. The generator model produces the saliency based on input latent uncertainty in saliency region while inference module infers the latent uncertainty. Testing pipeline utilizes prior distribution of latent uncertainty and generates several predictions for each input [79].

2. Late Fusion

This fusion scheme combines the cross-modalities in later stages of detection framework. Bypassing the low-level details, late fusion merge the independent predictions. Following models are based on late fusion scheme.

- **MV-CNN** This work proposes the first CNN based end-to-end RGB-D salient object detection model. In this model pre-trained RGB network supervises depth stream. The fully connected layer gives saliency probabilities which are then mapped to saliency prediction. Two stream network is adopted one for RGB modality and one for depth modality. A third stream MV-CNN is proposed to fuse the two modalities. This layer fuses the fully connected layers of RGB and depth stream. Global structural loss is utilized to capture global context [36].

- **AF** This work proposes a model based on observation that salient object is prominent in at least one modality. It fuses the predicted saliency maps from cross-modality streams. Due to limitations of each modality, this scheme fails at various challenging scenes. AF consists of two stream network, each stream fed with one modality. Through progressive aggregation at each stage of backbone, a saliency map is generated. RGB stream generates saliency map with RGB input. Depth stream generates saliency map with depth input. The last stage of RGB and depth streams are combined to form a switch map. A novel saliency fusion expression is proposed based on RGB saliency map, depth saliency map and switch map for final prediction [37].

- **HANet** In this research, a asymmetric two stream topology is explored, which fuses middle and high level features. The model distinguish adjacent salient and non salient objects with the attention mechanism. A large backbone network is used to extract detailed RGB modality features and a shallow backbone is used to extract comparatively simple features from depth modality. For RGB stream fully connected layer is removed and multi-level features are processed in side-output fashion. A weighted attention map from RGB and depth streams predicts the final saliency map [18].

3. Multi-Level Fusion

At each hierarchical position of backbone network, the extracted features contain distinct rich information. Therefore, multi-level fusion scheme is developed to effectively combine all representations. Following are SOTA models employing multi-level fusion.

- **DPANet** The model formulates potentiality of depth map explicitly using ground truth saliency map and generates pseudo labels. It then trains a regressor along with two stream network and fuses the RGB and depth at multi-level decoder. The encoder architecture of DPANet represents multi-modality features in side-output fashion and feed them to gated multi-modality attention (GMA) module. There are four GMA modules and they enhance RGB and multi-scale depth features. Decoder combines these multi-scale features progressively and separately for each modality. Lastly feature fusion block combines RGB and depth final decoded features to generate final saliency map. The last layer of backbone network is fed to global average pooling module to generate RGB and depth feature vector which is fed to regression module. The pseudo labels are used to train this regressor to handle the effect of low quality depth [80] .

- **BBSNet** This work proposes a Bifurcated Backbone Strategy that separates low and high features. Informative depth representations are obtained from depth enhancement module using attention mechanism while high features

are processed for global context. Two stream network consists of backbone for feature extraction where shallow and deep features are processed separately. To mitigate the impact of low quality depth, a module for depth enhancement is designed at multi-scales. This module have channel and spatial attention in series to attend strong features. The multi-modality features are combined in cascaded decoder. The global context is enhanced for deep features and using shallow features local details are effectively used [81].

- **BTSNet** This model observes the difference of representations at low and high level of backbone network. Thus, proposes a bidirectional transfer and select (BTS) module using encoder decoder framework. RGB and depth backbone network features of different stages are fed at each level with the output of previous level proposed BTS. At the end of each backbone stream atrous spatial pyramid pooling (ASPP) is added. BTS consists of spatial attention and channel attention. The decoder treats low and high features separately. All low features are added to form combined low feature and similarly for deep features . Multi-modality features are concatenated and then low and high multi-modal features are added and multiplied. Model consists of three prediction heads one for RGB, second for depth and third for RGB-D [39].

- **ICNet** This model adopts a Siamese encoder for information conversion (ICM) between cross-modalities and Cross-modal Depth-weighted Combination (CDC) block to enhance RGB features based on depth. They are fused in decoder stage. ICM adopts correlation to prop out salient object in each modality and concatenation to extract commonality. CDC is applied at multiple stages of backbone. CDC maps complementary features of multi-modality. Decoder generates multi-scale saliencies using CDC output and combines them for final saliency [82].

- **JLDCF** This model utilizes Siamese network for feature learning, which greatly reduces the model size. Then complementarity and commonalities of two modalities is observed at multiple levels and decoded using densely corporative fusion. Siamese backbone is used for multi-modal input and

side-outputs are extracted for multiscale features processing. Channel information is compressed to reduce the model complexity and then multi-modal feature fusion through addition and multiplication. Before fusion batch split is applied. Cross-modal fusion at multi-levels are progressively combine through inception module and addition. Model can be opted for other computer vision tasks [25].

Depth Quality-aware Models: Significant progress in saliency detection is observed with RGB and depth modality. However, exploiting multi-modalities has manifold drawbacks. (i) Adding depth branch increases complexity. (ii) Depth maps suffer from noise and redundant information during acquisition process. Low quality depth are due to errors in sensors or various scene conditions. Using these low quality depth reduces the accuracy. (iii) Depth is a scarce modality as compared to RGB. Existing depth quality-aware works are discussed below.

- **DASNet** This model consists of three modules. First saliency detection module takes RGB as input and generates a saliency map under supervision of ground truth. Second depth awareness module takes depth image as ground truth and produce a predicted depth with RGB input. Third, error weighted correction utilize predicted and original depth to produce depth error weight matrix, which is utilized in final saliency refinement. In test phase, depth is not used and prediction is based on RGB only [83].

- **CDNet** This model utilizes saliency informative depth along with original depth and RGB for salient object detection. For depth estimation, RGB is input to VGG encoder and decoder network generates estimated depth with target of original depth. In this process, low quality depth with little saliency information can degrade the results, that is why IoU values between original depth and ground truth saliency map is used. And for training only those samples are used, which have high IoU. For final prediction, estimated and original depth features are dynamically selected and fused with RGB [84].

- **DCF** The model proposes depth calibration scheme, which calibrates the original depth based on estimated depth. Two stream network is trained with RGB-D and ground truth. The resultant saliency maps are sorted and with higher IoU values with GT are selected for calibration network training. The calibrated depth maps are used for final saliency prediction. The calibrated depth and corresponding RGB is fed to two stream network. Cross Reference Module (CRM) combines deep three layers of RGB and depth stream in progressive manner. CRM extracts most discriminative features from RGB and depth modality and concatenates them, which is trained using triplet loss. Three decoders are used. One for RGB saliency generation from RGB stream, second for depth saliency and third from CRM. All are combined for final prediction [85].

- **FCFNet** In this research work, an image generation network trained on RGB and saliency based depth images to generate pseudo depth is developed. These pseudo depth are used to calibrate original depth. Afterwards, designed fusion module generates the final saliency map. Image generation stage is designed for low quality depth images and 2-step selection is used to detect high quality depth images to supervise image generation. In step 1, intersection over union between depth saliency map and ground truth is computed. In step 2, true positive rate of RGB saliency and depth saliency is calculated. And all those depth images are selected which have high true positive than RGB true positives. All RGB-D pairs having saliency consistent depths are selected for depth image generation. Next stage is salient object detection stage, which utilize raw depth and generated depth for calibrated depth generation. These depth are fused with RGB for final prediction [86].

- **DCMNet** The model separates low quality depth and performs enhancement on them. A novel multi-cross attention scheme is proposed, which performs channel and spatial attention in multi-cross way. Rectified depth images are generated based on saliency cues from ground truth and RGB.

Then the rectified depth and RGB are fed to two stream network. The multi-scale multi-modal features are enhanced and fused in proposed multi-cross attention module (MCAM) [87].

Critical Survey of CNN Based RGB-D SOD Models

A review of deep learning based SOTA models are presented. Based on extensive survey, the essential characteristics, strengths and limitations of existing CNN based RGB-D SOD models are presented in Table 2.1. Table is split according to backbone network and presents the architecture, number of streams and fusion strategy.

TABLE 2.1: Summary of essential characteristics of CNN based RGB-D Salient Object Detection Models

Models & Publ.	**Streams & Fusion Scheme**	**Architecture**	**Highlights and Limitations**
		VGG Backbone Network	
MV-CNN [36] IEEE 2017	Three Streams Late Fusion	Transfer Learning	**Highlights:** RGB-D transfer learning in a supervised manner via global contrast loss **Limitations:** (i) Modality-specific representation in shallow layers is ignored. (ii)Blurry saliency maps
LP-Net [67] TIP 2017	Single Stream Early Fusion	Simple CNN	**Highlights:** RGB-D saliency feature vector fed to CNN and refined by Laplacian propagation. **Limitations:** (i) not trained end-to-end. (ii) Fail at low contrast
CA-Fuse [68] CVF 2018	Two Streams Multi-level Fusion	BottomUp CNN	**Highlights:** Complementarity-aware fusion **Limitations:** Quality of depth not considered

Continued on next page

Table 2.1 – continued from previous page

Models & Publ.	**Streams & Fusion Scheme**	**Architecture**	**Highlights and Limitations**
CPFP [69] CVF 2019	Single Stream Contrast Prior Fused	Fluid Pyramid Integration	**Highlights:** Novel depth contrast loss and multi-scale integration. **Limitations:** Missing depth where depth contrast fail.
Att-CMCL [70] TIP 2019	Three Streams Multi-level Fusion	BottomUp TopDown Attentive Distillation	**Highlights:** Attention-aware fusion. **Limitations:** Complex model.
AF [37] IEEE 2019	Two Streams Late Fusion	BottomUp	**Highlights:** Switch-map between RGB and Depth streams. **Limitations:** Poor performance when objects are not distinguishable in RGB and depth modalities.
SSRC [34] Elsevier 2019	Single Stream Early Fusion	RNN	**Highlights:** RNN merges features from top to bottom. **Limitations:** Noisy depth gives inaccurate results.
D^3Net [33] TNNLS 2020	Three Streams Early Fusion	Feature Pyramid Network	**Highlights:** Depth depurator unit filters low quality depth in inference. **Limitations:** (i) Biased towards RGB for low quality depth. (ii) Larger parameter size.
A2dele [71] CVF 2020	Two Streams Late Fusion	Distiller & Attention	**Highlights:** Depth distiller mitigates low quality effects. **Limitations:** More informative depth features can be used.

Continued on next page

Table 2.1 – continued from previous page

Models & Publ.	**Streams & Fusion Scheme**	**Architecture**	**Highlights and Limitations**
ICNet [82] TIP 2020	Two Streams Multi-level Fusion	Siamese	**Highlights:** Attention based Encoder decoder network with concatenation-convolution and correlation-convolution **Limitations:** Poor performance for low quality depth.
LSF [1] IJCV 2021	Two Streams Multi-level Fusion	Teacher-Student Framework	**Highlights:** (i) RGB teacher model distil knowledge to depth student model (ii) progressive complementarities (iii) zero-shot learning **Limitations:** depth quality not considered.
CDNet [84] TIP 2021	Two Streams Multi-level Fusion	Encoder-Decoder	**Highlights:** Saliency informative depth is utilized for low quality depth **Limitations:** Biased towards RGB can lead to incorrect estimated depth
LIANet [73] Access 2022	Two Streams Multi-level Fusion	Feedback Mechanism	**Highlights:** Layered interactive fusion and RGB-depth- RGB feedback for noisy depth. **Limitations:** Uncertainty in objects.
FCFNet [86] TCSVT 2023	Two Streams Multi-level Fusion	U-Net	**Highlights:** Pseudo depth is calibrated and used in fusion. **Limitations:** Pseudo depth is based on RGB.
M^2RNet [75] PR 2023	Two Streams Late Fusion	Encoder-Decoder	**Highlights:** (i) Attention to reinforce good features. (ii) Adjacent integration. (iii) Joint optimization loss. **Limitations:** Lacks generalization.

Continued on next page

Table 2.1 – continued from previous page

Models & Publ.	**Streams & Fusion Scheme**	**Architecture**	**Highlights and Limitations**
		ResNet Backbone Network	
HANet [18] Appli. Sci. 2020	Two Streams Late Fusion	Res2Net for RGB Inception-v4-ResNet2 for depth	**Highlights:** Asymmetric attention network **Limitations:** Low quality depth not considered
DPANet [80] TIP 2020	Two Streams Multi-level Fusion	Encoder-Decoder	**Highlights:** Saliency based depth quality assessment **Limitations:** Poor performance if (i) Conflict between saliency and depth (ii) long distance objects
DASNet [83] ACM 2020	Single Stream Multi-level Fusion	Encoder-Decoder	**Highlights:** Channel-Aware fusion for distinct features. **Limitations:** Prediction is biased towards RGB
UCNet [79] PAMI 2021	-	GAN	**Highlights:** Attention based GAN with DenseA-SPP **Limitations:** Need improvement for illumination changes.
BBSNet [81] TIP 2021	Two Streams Multi-level fusion	Encoder-Decoder	**Highlights:** (i)Attention for depth enhancement. (ii)Cascaded refinement based fusion. **Limitations:** Depth quality ignored.
BTSNet [39] ICME 2021	Two Streams Multi-level Fusion	Encoder-Decoder	**Highlights:** (i)Transfer and select. (ii)Low and high features combined separately. **Limitations:** Complex.

Continued on next page

Table 2.1 – continued from previous page

Models & Publ.	**Streams & Fusion Scheme**	**Architecture**	**Highlights and Limitations**
JLDCF [25] PAMI 2021	Two Streams Multi-level Fusion	Siamese	**Highlights:** (i)Joint learning using Siamese network (ii)Cross-modal corporation fusion. **Limitations:** Poor performance for low contrast RGB and non-salient regions are adjoined.
DCF [85] CVF 2021	Two Streams Multi-level Fusion	Encoder-Decoder	**Highlights:** Low quality depth images are calibrated in plug-n-play manner. **Limitations:** Lacks generalization ability.
CIR-Net [72] TIP 2022	Three Streams Multi-level Interaction	ResNet	**Highlights:** (i) Interaction modules between cross-modalities in encoder and decoder. (ii) Middleware refinement network. **Limitations:** Poor detection at (i) Multiple and small salient objects. (ii) non salient (iii)complex background.
DCM-Net [87] ESA 2023	Two Streams Multi-level Fusion	Res2Net	**Highlights:** (i) DDRM for depth rectification. (ii) Progressive feature refinement and fusion. **Limitations:** Depth stream architecture needs to be enhanced.
HiDA-Net [74] TIP 2023	Two Streams Multi-level Fusion	U-Net	**Highlights:** Granularity-based attention. **Limitations:** Low depth quality ignored.
BFLNet [88] PR 2024	Two Streams Multi-level Fusion	Encoder Decoder	**Highlights:** (i)Asymmetric feature extraction. (ii)Bidirectional feature fusion. (iii)Dual Consistency Loss function **Limitations:** Low depth quality ignored.

Critical Survey of Transformer Based RGB-D SOD Models

Table 2.2 presents the summary of properties of Transformer based RGB-D SOD Models. The highlights and limitations of recent SOTA models are discussed. Some Transformer based models are build on pure transformer network while others utilize CNN-Transformer Hybrid network as backbone. Transformers can capture long-range dependencies more effectively than convolution neural networks. Hybrid networks can leverage the capabilities of transformer to obtain global contextual information along with the capturing of local details from convolution neural networks. Hybrid networks have CNN-Transformer blocks in early, late, sequential or parallel patterns.

TABLE 2.2: Summary of essential characteristics of Transformer based RGB-D Salient Object Detection Models

Models & Publ.	**Backbone**	**Architecture**	**Highlights and Limitations**
SwinNet [17] TCSVT 2021	Swin	Encoder-Decoder	**Highlights:** (i) Two stream network with attention based fusion. (ii) Edge-aware decoder for boundaries. **Limitations:** Complex.
VST [76] CVF 2021	T2T-ViT	Encoder Multi-task Decoder	**Highlights:** (i)Multi-level token fusion. (ii) Task-related tokens for saliency prediction. (iii) Patch-task-attention for boundary detection **Limitations:** Depth quality ignored.
TriTransNet [26] ACM 2021	Triplet transformer embedding with ResNet	U-Net	**Highlights:** Transformer embedded with CNN for feature extraction. **Limitations:** Depth purification module needs to be improved.
MutualFormer [77] 2021	Hybrid	Encoder-Decoder	**Highlights:** Intra-modality and inter-modality tokens used. **Limitations:** Error in fine-grained detection.

Continued on next page

Table 2.2 – continued from previous page

Models & Publ.	**Backbone**	**Architecture**	**Highlights and Limitations**
SiaTrans [27] IVC 2022	T2T-ViT	Siamese Transformer Network	**Highlights:** Fusion and decoder module selects RGB-D or RGB stream. **Limitations:** RGB stream gives poor results for very low quality depth.
CAVER [78] TIP 2023	Hybrid	Transformer based information propagation path	**Highlights:** Cross modal view-mixed transformer. **Limitations:** Poor performance for cluttered background.
CMAFE [89] IET 2024	Swin	Encoder-Decoder	**Highlights:** Dual-attention to filter noise from two modalities. **Limitations:** Poor result for low-light scenes.

2.3 Research Gap Analysis

Extensive survey on multi-modality salient object detection presented in literature review shows that deep learning has obtained great success to manifest RGB and depth features for saliency detection task.

Main limitations of existing methods are given below:

- The intra-modality features of RGB and depth are distinct. RGB gives color and textural details while depth modality provides geometric cues, 3D layout, spatial details and object edges etc. The disparity of two modalities is not well explored, specifically, with feature extraction from backbone network.
- Mostly, CNN are used as backbone network for feature extraction, which have inherent ability to focus only on local details. Recently, some models are build on Transformer as backbone network to capture global context. However, transformer lacks the ability to fully represent local details. For

RGB-D SOD task, both local and global details are required to fully comprehend the scene.

- In current literature, the intrinsic correlation of RGB and depth is utilized only for effective multi-modality fusion. However, no contribution in literature is available that exploit multi-modality correlation for complex scenarios and less instructive images.

- Quality of depth images are not always the same. The noise and redundant information is introduced inevitability due to depth sensors and various image capturing conditions such as occlusion, reflection, and viewing distance. The low-quality depth adversely effect the detection accuracy. Existing methods, ignore the image quality in fusion process. Some researchers have exploited multi-modular approach for depth quality enhancement and cross-modal feature fusion. This two step strategy leads to sub-optimal solution.

- Some depth maps are so blurred and noisy that they need to be discarded. No contribution in existing literature is available for incomplete RGB-D modality saliency learning problem, where we have complete RGB images but some depth images are missing.

The research gaps mentioned above lead us to following research questions:

1. Can we integrate depth quality enhancement and cross-modal feature fusion to develop an end-to-end framework for RGB-D SOD?

2. How can we solve the incomplete multi-modality learning problem, when we have complete RGB images but with some missing depth images?

2.4 Problem Statement

"Quality of depth map in RGB-D salient object detection varies due to depth sensors. Low quality depth map introduce noise, sparse or redundant information,

causing negative effect on detection. Some depth maps are so blurred that they need to be discarded, creating incomplete multi-modal learning problem." Thus, the main focus of this research is "To develop a model that can learn modal-specific information and cross-modal complement fusion for salient object detection having incomplete RGB-D data."

2.5 Proposed Methodology

The main focus of this research is to develop depth quality-aware salient object detection method with improved detection efficiency and reasonable efficacy. In Fig. 2.3 the proposed research methodology can be visualized. The proposed methodology consists of two distinct methods. The two methods introduce depth quality aware saliency detection explicitly and implicitly. The explicit depth quality-aware model correlates the saliency edge map, RGB edge map and depth edge map. The intuition behind is two-fold: (1) Spatial distortion of low quality depth is evident in depth edge map. (2) RGB and depth edge misalignment reduces detection accuracy. The implicit depth quality-aware model do not fuse low-quality depth. Initially, each depth image is fed to depth quality assessment module to predict the quality score. Low quality depth images are discarded and saliency detection model is trained using incomplete RGB and depth i.e. training is done with two types of data (1) complete RGB and depth pair, (2) RGB present depth missing.

2.6 Conclusion

A comprehensive review of recent salient object detection models, presented in this chapter, reveals that deep learning has achieved significant advancements in leveraging RGB and depth features for the task of saliency detection. The chapter categorized the different aspects of saliency detection, including traditional vs deep models, CNN vs transformer backbone network, fusion schemes and depth aware models. Based on the literature review, it is evident that multi-modality

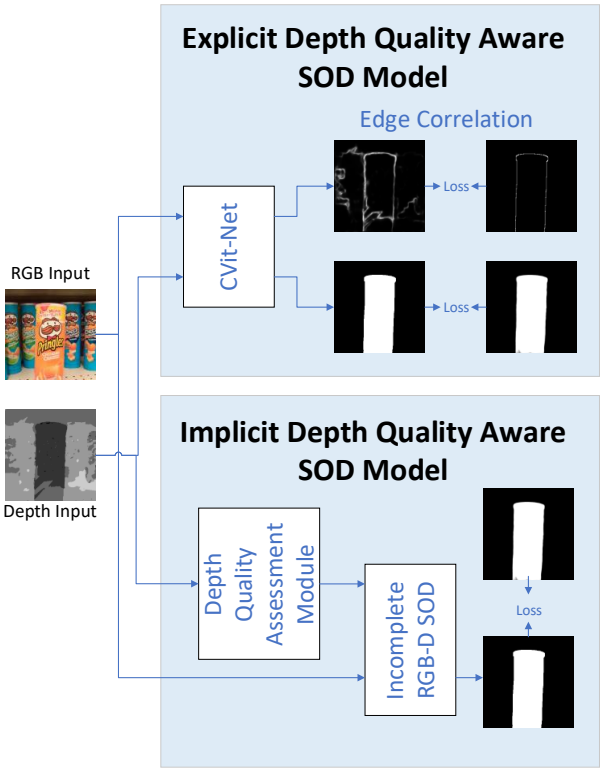


FIGURE 2.3: Proposed Research Methodology.

SOD modeling faces three major challenges. (i) The first challenge is to consider the differences between modalities for intra-modality feature selection. (ii) The development of a task-oriented optimal multi-modality fusion scheme poses a significant challenge. (iii) How can we mitigate the detrimental impact on model performance caused by low-quality depth maps? The existing approaches do not address all of these challenges, leading to ambiguous learning. Therefore, the primary goal of this research is to develop a Salient Object Detection (SOD) model capable of effectively handling and minimizing the impact of low-quality depth, as well as selecting descriptive RGB-D features and achieving effective fusion. Two distinct approaches for a depth quality-aware SOD model have been proposed. The first model explicitly evaluates depth quality using edge correlation, while the second approach introduces a depth quality assessment regression module to eliminate low-quality depth images. This is followed by the proposal of a novel incomplete RGB-D modality SOD learning approach. In the following chapters, main contribution of proposed methodology will be elaborated.

Chapter 3

Depth-Aware Saliency Detection

This chapter introduces the first contribution of the research, highlighting the innovative aspects and novel findings of depth quality aware salient object detection (SOD) framework. Several reasons contribute to the ambiguity of multi-modal (RGB and depth modalities) fusion in SOD models. Firstly, the complementarity features of RGB and depth modalities are not properly selected. Secondly, the quality of depth images is not considered which eventually mitigate the detection performance. Thirdly, the fusion strategy of cross-modalities is not very persuasive. To overcome these challenges, a two-way feature manifestation scheme is proposed to exploit the intrinsic properties of RGB and depth modalities. To reduce the disruption caused by low quality depth images, a novel edge guidance module is proposed that correlates the RGB and depth edge features in a supervised manner. The integration of depth quality enhancement and cross-modal feature fusion occur in end-to-end manner. In this chapter, experimental setup of proposed explicit depth quality-aware salient object detection model (CVit-Net) is also presented. A detail analysis of results obtained on benchmark datasets is provided. Qualitative and quantitative analysis of proposed models with respect to state-of-the-art (SOTA) models is presented. A detail ablation investigation to confirm the importance of each module is presented. The proposed model demonstrates promising outcome.

3.1 Multi-modality Feature Processing for Saliency Detection

Salient object detection involves identifying the most visually conspicuous object(s) within images or videos. The integration of depth images (captured by sensors like Microsoft Kinect and Time-of-Flight) along with RGB images has become increasingly prevalent in various computer vision and robotics tasks. Robustness and effectiveness of salient object detection models can by enhanced by utilizing the synergies of RGB and depth modality [1].

Conventional RGB-based SOD methods fail when the salient object and background exhibit similar appearances. Unlike the color and texture details extracted from the RGB modality, the depth modality offers valuable geometric cues that prove advantageous in numerous challenging scenarios, including illumination variations, low-contrast, and complex backgrounds. In addition to traditional and convolutional neural networks (CNN) based SOD models, some researchers have proposed co-saliency [90] and generative adversarial network (GAN) models [37] to enhance detection accuracy.

For the efficient processing of multi-modal data, three crucial factors require consideration [1]:

1. intra-modality features
2. inter-modality features correlation
3. quality of RGB and depth images

Previously proposed models fail to address all the mentioned aspects in a single comprehensive model. The subsequent sections methodically explore these three aspects of RGB-D SOD architecture.

3.1.1 Intra-modality Features

Knowing the modality-specific representation of various modalities is crucial for development of multi-modal SOD model. The depth modality captures spatial details and object locations, while the RGB modality provides color and texture features. For depth modality, global contextual representation provides legitimate cues for salient object detection. While rich information can be disentangled for saliency detection using local features representations. The disparity of two modalities is handled previously through following different architectures.

3.1.1.1 Modality-specific Features Extraction using CNN

Considerable amount of work published so far, have used convolutional neural networks (CNN) for various computer vision tasks. For RGB-D salient object detections, various models [18, 24, 25, 38, 39], used encoder-decoder architecture. The innate forte of CNN is extraction local details. Although some work [25, 39] suggest the use of dilated convolution to capture global details with CNN but proficiency of these models are not up-to the mark. The global semantic details are not well projected due to spatial loss in pooling operation of CNN.

3.1.1.2 Modality-specific Features Extraction using Transformer

To overcome the paucity of capturing long range dependencies the Transformer architecture [91] was proposed for natural language processing task. After that [66] proposed visual tranformer (ViT) to process images. As for SOD task global semantic information is very important to subjugate visually prominent object, therefore, several new models [17, 26, 27, 76, 77, 92] of RGB-D SOD utilize transformer as backbone network. The Transformer architecture is good for the manifestation of long range dependencies but lacks to capture local details. However, the intrinsic nature of RGB modality suggest to incorporate local information in saliency detection task. Subsequently, RGB features are not well represented using Transformer based backbone network.

3.1.1.3 Modality-specific Features Extraction using Conformer

Considering the intrinsic properties of RGB and depth modalities, recently proposed the Conformer [93] architecture is opted for modality-specific features extraction. The Conformer network captures local and global details simultaneously. The Conformer architecture has two streams: one is CNN stream and other is Transformer stream. Instead of using two Conformers for two modalities, a single Conformer is employed and consign RGB modality to CNN stream and depth modality to Transformer stream. In this way, not only modality-specific features are well represented but also the number of parameters are reduced. To illustrate the value of concurrently recording local and global features in a Conformer network, two scenarios are shown in Fig. 3.1. When salient and non-salient objects can't be separated because of a complicated background and low contrast, as is the scenario in the first test case in Fig. 3.1, our Conformer-based model effectively captures the salient object. However, CNN-based joint learning and densely cooperative fusion (JL-DCF) [25] model, do not perform well due to preponderance of CNN for extracting local details than global contextual representation. Moreover, for low textured multiple objects at equal distance from viewpoint as shown in second test case of Fig. 3.1, the Transformer based model [26] fails to separate the objects, however, my proposed Conformer-based network utilizes both local and global representation.

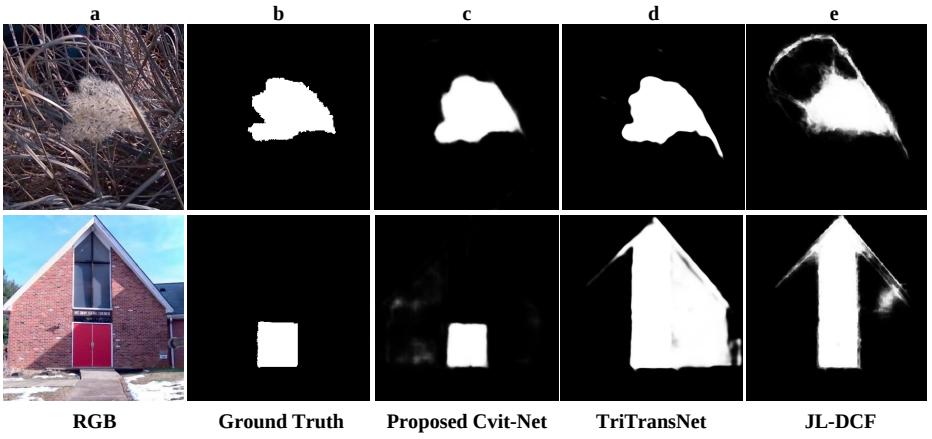


FIGURE 3.1: Comparison of saliency maps obtained from CNN-based model JL-DCF [25], transformer-based model TriTransNet [26] and proposed Conformer-based model.

3.1.2 Inter-modality Features Correlation

The second challenge for an efficient multi-modal salient object detection model design, is the selection of appropriate fusion strategy that can boost the inter-modality features correlation. Recent research has demonstrated that multi-level (middle) [17, 25, 39] fusion strategies outperform more traditional early fusion strategies [33–35] and late fusion strategies [18, 36, 37]. However, the features extracted from the shallow layers of backbone network are rich in modality-specific information while deep features contain saliency cues. Therefore, fusion of relevant patterns is crucial. Considering the discrepancy of shallow and deep features, we opt for a two-way feature aggregation strategy. Shallow layers are handled in local detail enhancement (LDE) module using bottom-up approach while deep layers are handled in global detail enhancement (GDE) module in a top-down way.

3.1.3 Quality of RGB and Depth Images

In contrast to uni-modal SOD models, structural information extracted from depth modality when combined with color and textural details extracted from RGB modality, plays a crucial role in boosting the performance of SOD model. However,depth is not a panacea, as noise and errors are introduced during acquisition process. Utilizing low quality depth maps significantly reduces the detection accuracy. Therefore, it is necessary to assess the depth quality before fusing it with RGB modality. Previously, only few models [24, 33] focus to consider depth quality before fusion. Thus, we evaluate the quality of depth image data using edge guidance and introduce adaptive aggregation to emphasize the inter-modality correlation. Our proposed model learns the correlation of RGB and depth features in a supervised manner and thus explicitly learns the depth quality. As depicted in Fig. 3.2, the edge map of corresponding depth map explicitly guide about the quality of depth map. The edge map of good quality depth maps preserves all structural details as shown in Fig. 3.2 (a). In case (b) of Fig. 3.2, the depth image quality is somewhat deteriorated, and hence, its corresponding edge image

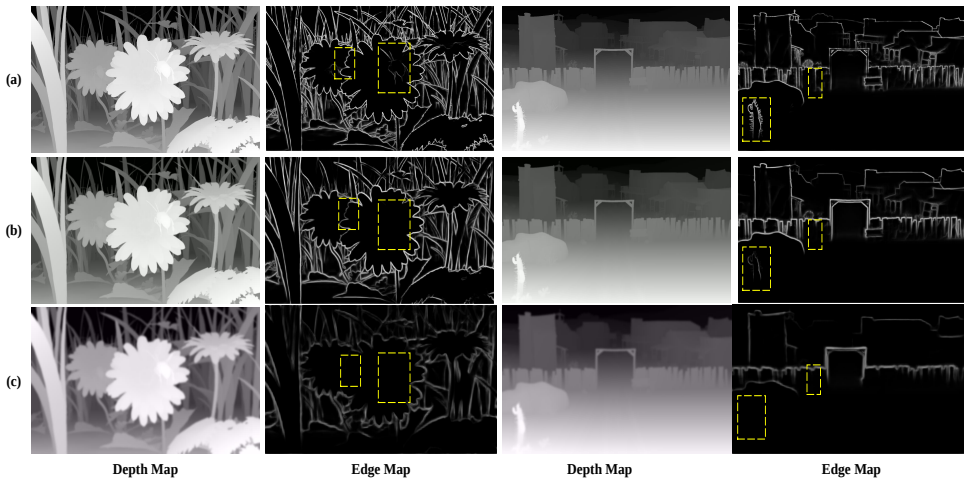


FIGURE 3.2: (a) Good quality depth maps and their edge maps (b) Less noisy depth maps and their edge maps (c) More noisy depth maps and their edge maps.

losses some details. While in case (c) of Fig. 3.2, very poor quality depth map fails to retain most of geometric cues in edge map.

3.1.4 Research Contribution Overview

Driven by above challenges, we propose a novel RGB-D Salient Object Detection Model (CVit-Net) that utilizes global contextual representation and multi-modal local discrepant representation in parallel top/down and bottom-up manner. Brief description of key contributions are as follows:

1. We opt encoder-decoder architecture for multi-modal saliency detection task. To fully comprehend local details along with effectively capturing long range dependencies, a recently proposed Conformer network [93] is used as encoder. Conformer is a two stream network. RGB enhanced feature representations are obtained from CNN stream of Conformer, while for depth manifestation Transformer stream of Conformer is used.

2. We handled shallow RGB-D features in Local Detail Enhancement (LDE) module. We propose a novel operation-wise shuffle channel attention scheme that correlates edge features of both modalities in a supervised manner. The motivation behind this is to assess depth quality explicitly. Moreover, we

streamlined the model by incorporating end-to-end training for edge guidance using ground truth edge maps. This simplifies the model as in real world scenarios depth quality labels are not usually available.

3. We handled deep RGB-D features in Global Detail Enhancement (GDE) module. GDE module apprehend global details using reverse attention in recurrent top/down approach.

4. To integrate the representations obtained from LDE and GDE module, we designed a lightweight decoder that combines multi-level LDE and GDE outputs considering the features alignment legitimation.

3.2 Proposed Method

Keeping in view the three challenging aspects of RGB-D salient object detection framework, we propose a novel model CVit-Net using Conformer as backbone. The overall architecture is presented in Fig. 3.3. The proposed model is an enocder-decoder framework. In encoder part, the RGB and depth features are extracted using Conformer network. Conformer consists of two streams, one is CNN stream and other is Transformer stream. CNN stream is used to extract RGB features, which are rich in textural details. And Transformer stream is used to extract depth features. Shallow layers are handled separately in local detail enhancement (LDE) module while features from deep layers are processed in global detail enhancement (GDE) module. The quality of depth images is assessed explicitly in LDE module. LDE module consists of novel operation-wise shuffle channel attention module, which correlates the RGB and depth edge features in a supervised manner. Features are derived from the RGB modality using a CNN-based architecture that incorporates dilated convolutions and pooling with diverse kernel sizes and dilation rates. This approach enhances the receptive field of the extracted representations. The depth modality employs shuffle channel attention, which seeks to identify the inter-channel interaction of depth information in order to uncover heuristic depth signals.Therefore, inspired by [94], in order to handle any type of

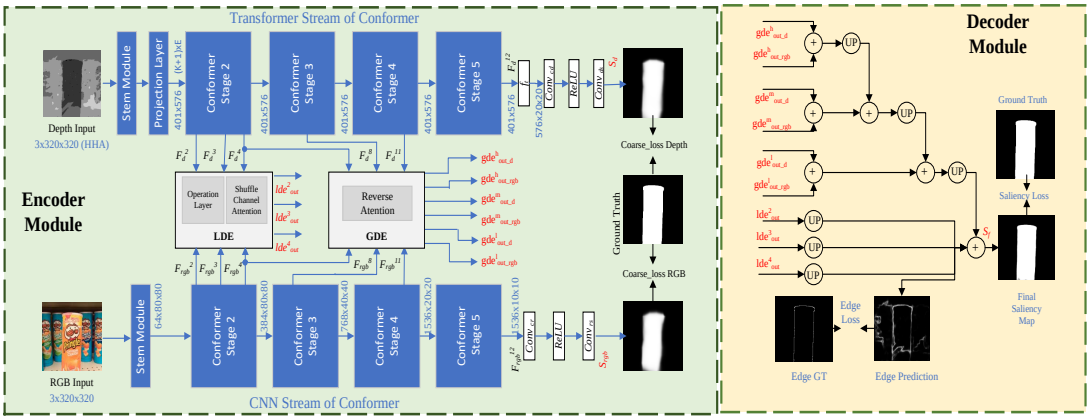


FIGURE 3.3: Encoder is shown on the left which consists of Conformer backbone for feature extraction. Shallow features of Conformer encoder are fused in LDE module. Deep features are enhanced in GDE module. Decoder is shown on the right. 'UP' represents upsampling and '+' represents element-wise addition.

distortion without reference depth map, the proposed innovative operation-wise shuffle channel attention network executes a number of operations concurrently, weighted by shuffle channel attention. GDE module consists of reverse attention mechanism for semantic saliency detection. A lightweight decoder combines the outputs of LDE and GDE module for final saliency map prediction. The training pipeline of proposed model is given in Algorithm 1

3.2.1 Raw Feature Extraction through Backbone Network

Beforehand, it is explored that high-level computer vision tasks need a suitable backbone network for informative features extraction. Mostly, classification networks like VGG [29], Resnet, Visual transformer, Swin transformer etc. are used as backbone feature extractor framework. The pre-trained modules of above mentioned classification networks are fine-tuned for object detection task. The Conformer for Visual Recognition is incorporated as the backbone network because of its capability to simultaneously learn both local and global details. We choose to use Conformer-B with 9 heads in the multi-head self-attention (MHSA-9) block of the transformer, along with a feature pyramid structure of convolutions, organized into 4 stages. RGB and depth images are fed as input to stem module of Conformer. Stem module generates two outputs, one is fed to CNN branch and other

Algorithm 1
Training pipeline of CVit-Net for salient object detection

Input:
(1) Training dataset $D = \{(X_c, Y_c)\}_{c=1}^{N}$, Ground truth saliency maps $GT = \{S_c\}_{c=1}^{N}$, Ground truth edge maps obtained from ground truth saliency maps $Edge\_GT = \{EdgeGT_c\}_{c=1}^{N}$ where c indexes images and N is training dataset size (2) Number of epochs *Epochs* (3) Hyper-parameter of edge loss α_2, α_3, α_4 (4) Learning-rate λ

Output:
Final saliency map S_f , RGB coarse saliency map S_{rgb}, Depth coarse saliency map S_d, Edge saliency maps $\{S_E i\}_{i=2}^{4}$, where i represents backbone layer number.

1: Initialize backbone Conformer network $\mathcal{N}$ with pretrained Conformer-B [93] weights for image classification.

2: **for** $e \leftarrow 1$ `to Epochs` **do**

i) Sample input images $\{(X_c, Y_c)\}_{c=1}^{N}$

ii) Sample corresponding ground truth saliency maps $\{S_c\}_{c=1}^{N}$ and ground truth edge maps $\{EdgeGT_c\}_{c=1}^{N}$

iii) $F_d^i, F_{rgb}^i \leftarrow \mathcal{N}(X, Y; w)$, obtain i^{th} feature maps from Conformer network $\mathcal{N}$

iv) Sample $\{rgb_{l_j}^i\}_{j=1}^{5}$ and $\{A_{d_j}^i\}_{j=1}^{5}$ following Eq. 1 and 3 respectively.

v) $lde_{out}^i \leftarrow c((A_{d_1}^i \times rgb_{l_1}^i), (A_{d_2}^i \times rgb_{l_2}^i), (A_{d_3}^i \times rgb_{l_3}^i), (t_4(A_{d_4}^i \times rgb_{l_4}^i), (t_5(A_{d_5}^i \times rgb_{l_5}^i)))$, Generate encoder output signals from Local Detail Enhancement (LDE) module

vi) Obtain S_{rgb} and S_d following Eq. 5 and 6 respectively.

vii) $gde_{out_m}^i \leftarrow (1 - Sigmoid(UP(S_m))) \times E_m^i$, Generate encoder output signals from Global Detail Enhancement (GDE) module

viii) $S_f \leftarrow \sum_i UP_i(lde_{out}^i) + UP((gde_{out_d}^l + gde_{out_{rgb}}^l) + UP((gde_{out_d}^m + gde_{out_{rgb}}^m) + (UP(gde_{out_d}^h + gde_{out_{rgb}}^h))))$, Decode (v) and (vii) output signals and Predict final saliency map

ix) $\mathcal{L}_t \leftarrow \mathcal{L}_f(S_f) + \alpha_i \mathcal{L}_E i(S_E i) + \mathcal{L}_d(S_d) + \mathcal{L}_r(S_{rgb})$, loss function

x) Update model parameters using $\mathcal{L}_t$

3: **end for**

is fed to Transformer branch. Instead of combining multi-model RGB-D input or using two separate Conformer for each modality, we fed RGB images to CNN stream and depth images to Transformer stream. Initial RGB features are captured using convolution operation with stride 2, kernel size 7x7 and max pooling operation with kernel size 3x3. The depth input image is converted to 20x20 patch embeddings, using linear projection layer. Initial depth features are processed by same stem module as of RGB followed by convolution of kernel size 4x4 and stride 4 to generate patch embeddings. Fully-connected layer of Conformer is removed. Features are extracted from last layer of each stage of Conformer in a side-out manner. While for shallow features processing, all three layers in stage c2 are used for low level details enhancement. In proposed model these features are denoted as low, mid, high and coarse features. The features' resolution of CNN branch in form of $H \times W, C$ (Height, Width and Channels) and Transformer branch in form of $E, (K + 1)$ (K, 1 and E are number of image patches, class tokens and embedding dimensions) for each stage is presented in Table 3.1. The resolution and channel number of the side-outs in the CNN branch vary, with increasing channels and decreasing resolution, which does not align with the side-outs of the transformer stream. We ignore the class tokens as it is not relevant to our SOD task.

TABLE 3.1: Output features shape of Conformer-B and configuration used: multi-head self attention (MHSA) with 9 number of heads, patch size of 16, channel ratio of 6 and embedding dimensions of 576.

Stage	**Features Notation**	**CNN Branch Output**	**Transformer Branch Output**
c1	stem	160x160,64	160x160,64
		80x80,64	80x80,64
c2	low	80x80,384	401,576
c3	mid	40x40,768	401,576
c4	high	20x20,1536	401,576
c5	coarse	10x10,1536	401,576

Proposed CVit-Net model describes i^{th} side-out layer feature from CNN and Transformer streams with F_{rgb}^{i} and F_{d}^{i} respectively.

3.2.2 Local Detail Enhancement Module

Intuitively, depth images captures structural details and object location. As a result, in shallow layers, the depth map presents more distinct feature representations in contrast to RGB. However, there is a possibility of increased depth estimation error when salient and non-salient objects are in close proximity to each other or when non-salient objects are closer to the viewpoint. Additionally, depth images also suffer from noise during acquisition. Hence, fallacious feature manifestations are obtained if low quality depth maps are used for detection. There are also several reasons of depth image quality degradations and the reference depth maps are not usually available in real world scenarios. Therefore, a novel operation-wise shuffle channel attention is proposed in local detail enhancement module. The detail architecture is provided in Fig. 3.4. Operation layer consists of five parallel operations to supplement depth with affluent RGB manifestations. Besides, shuffle channel attention subjugates depth degradations. Overall, LDE module follows following steps:

- Three side-outs of Conformer backbone (described by i=2,3,4) are selected for shallow RGB and depth features extraction.
- Five parallel operations including convolution and pooling operation with different kernel and stride values are implemented on CNN stream of Conformer.
- For each operation-wise learning of RGB modality, depth features are extracted using shuffle channel attention with different number of groups.
- The hierarchical shallow RGB and depth representations are multiplied and concatenated. The intuition behind this is to effectively select the complementarity of cross-modalities.
- The low level details are trained in a supervised manner. For this purpose edge ground truth are generated from saliency ground truth using Sobel operator [95].

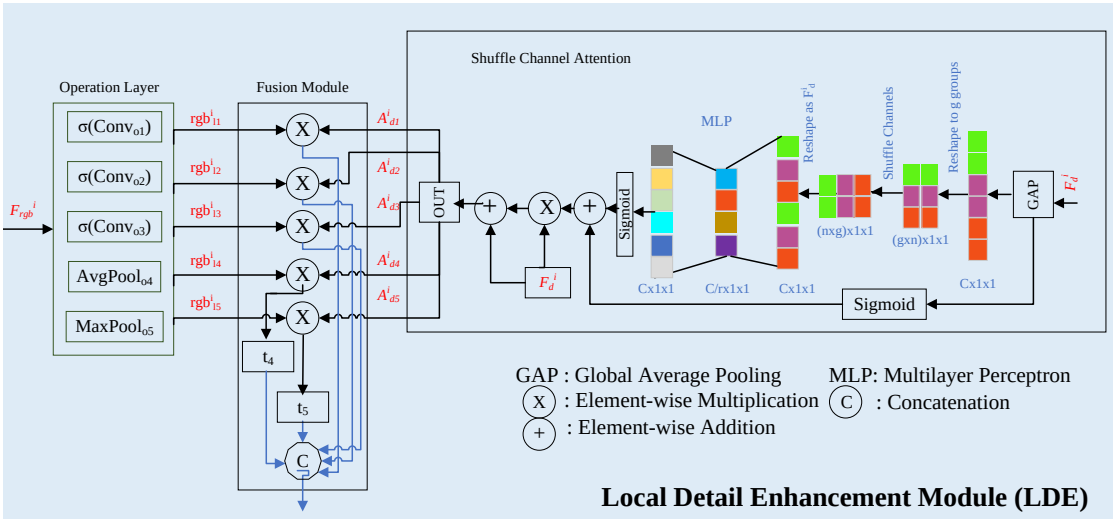


FIGURE 3.4: Detailed architecture of proposed Local Detail Enhancement (LDE) Module.

The motivation behind LDE module is to correlate edge details of RGB and depth modality to mitigate the impact of low quality depth images. The adaptive training of RGB and depth boundary cues guides the SOD model about depth quality in an explicit manner.

3.2.2.1 Extraction of Color/Textural Encoded Patterns from RGB Modality

Following the conventional computer vision modelling approach that involves exploiting a backbone network to extract raw RGB features, we leverage the CNN branch within the Conformer network. Typically, feature maps acquired through convolution operations with sliding window kernels tend to lack a robust representation of effective global contextual information. Therefore, to obtain effective global details, receptive field of RGB features are enhanced using parallel convolutional layers with different kernel sizes, accompanied by dilated convolution. As a result, the computational cost is reduced. The modality-specific patterns of RGB images are processed in five parallel operation layers shown in Eq. (3.1). The first layer consists of convolutional operation $Conv_{o1}$ having kernel size (7x7), stride 1, padding 6 and dilation rate 2, followed by ReLU activation σ. The convolutional operation $Conv_{o2}$ in second layer have kernel size(5x5), stride 1, padding 4

and dilation rate 2. The third layer has convolutional operation $Conv_{o3}$ of kernel size(3x3), stride 1, padding 2 and dilation rate 2. Fourth and fifth layers are pooling layers. $AvgPool_{o4}$ and $MaxPool_{o5}$ corresponds to Average and Max Pooling operations with kernel size (4x4) and stride 4. These operations are performed on three shallow conformer side-outs levels represented by i, where i is (2 to 4).

$$rgb^i_{l_1} = \sigma(Conv_{o1}(F^i_{rgb})), \tag{3.1a}$$

$$rgb^i_{l_2} = \sigma(Conv_{o2}(F^i_{rgb})), \tag{3.1b}$$

$$rgb^i_{l_3} = \sigma(Conv_{o3}(F^i_{rgb})), \tag{3.1c}$$

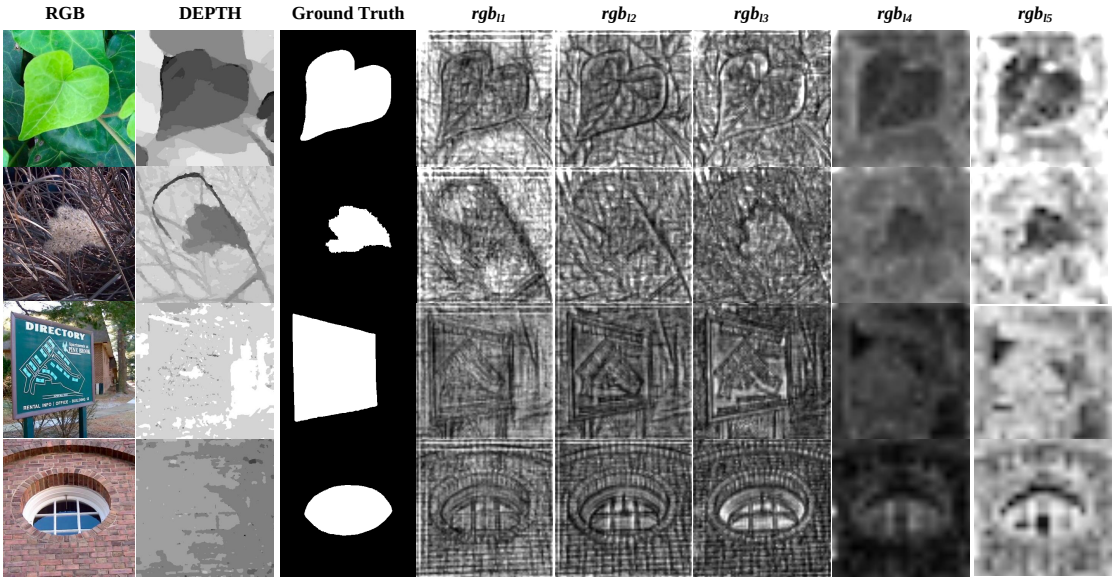
$$rgb^i_{l_4} = AvgPool_{o4}(F^i_{rgb}), \tag{3.1d}$$

$$rgb^i_{l_5} = MaxPool_{o5}(F^i_{rgb}). \tag{3.1e}$$

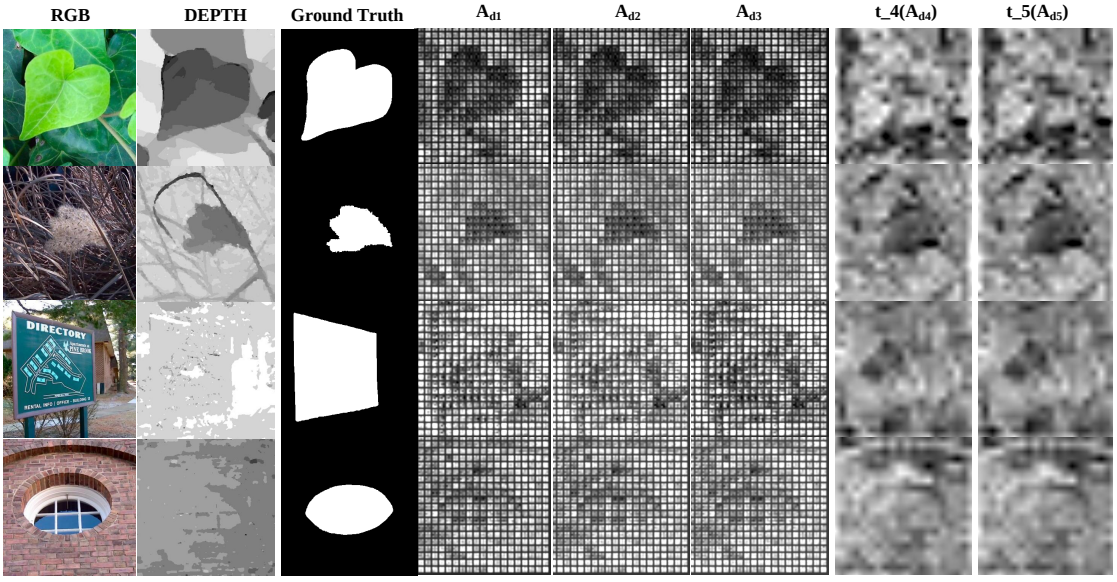
It is apparent in visual representation of encoded RGB patterns (given in Fig. 3.5 (a)) that parallel operation can capture rich local details. Convolutional operations help emphasize boundary cues based on texture or color variations, while pooling operations strengthen appearance details that are advantageous for emphasizing salient objects. For low quality depth images (Row 3 and 4 of Fig. 3.5 (a)), concurrent operations on RGB data generate diverse patterns and acquire informative cues to supplement depth modality.

3.2.2.2 Extraction of Depth Quality Aware Attentive Maps from Depth Modality

Although, shallow layers provide rich geometrical details of depth modality,however, noisy depth maps loose several boundary cues and contribute in redundant information. Therefore, it is crucial to extract depth quality aware representations. With regard to the mentioned scenario, Shuffle Channel Attention is proposed to depict 'what' features matters. Analogous to various Channel Attention algorithms, Shuffle Channel Attention approach comprehend the spatial information by applying Global Average Pooling. The dimension of resultant feature map is $C \times 1 \times 1$. Afterthat, C channels are divided into g groups, such that, each g group has n Channels (presented as $(g \times n \times 1 \times 1)$). Channels are then shuffled



(a) Feature Maps after Convolution and Pooling Operations on RGB Modality



(b) Shuffle Channel Attention Maps of Depth Modality

FIGURE 3.5: Demonstration of five parallel operation results on RGB and depth modality.

to maximize the information exchange among g groups. The new representation has $n \times g \times 1 \times 1$ dimension. Afterwards, shuffle channels are reshaped to original dimension of $C \times 1 \times 1$. Mathematically, shuffle channels operation $SC(.)$ is presented in Eq. 3.2

$$\mathcal{D}^i_{sc} = SC(GAP(F^i_d)), \tag{3.2}$$

Furthermore, $SC(.)$ operation is forwarded to multi-layer perceptron (MLP) network. MLP is designed to have a single hidden layer. Parallel to shuffle channel

operation branch, a residual connection is provided to optimize training process. Each parallel branch is equipped with Sigmoid activation to generate attention map. Original depth features are multiplied and added with attention map as shown in Eq. 3.3

$$A^i_{d_j} = [\ \{Sigmoid(MLP(\mathcal{D}^i_{sc})) + Sigmoid(GAP(F^i_d))\} \times F^i_d]\ + F^i_d, \tag{3.3}$$

where j represent layer (1 to 5) and i represent Conformer side-out levels (2 to 4). To validate the performance of proposed shuffle channel attention, the intermediate features from five shuffle channel operations are shown in Fig. 3.5 (b). Group size of (4,8,16,32,64) are used in five parallel operations. It can be visualized that, even for low quality depths with almost zero informative details, proposed scheme can extract salient object effectively.

3.2.2.3 Fusion of Shallow Features in LDE Module

As mentioned earlier, shallow RGB features are extracted using five parallel operations including convolution and pooling operations. While, shallow depth features are extracted from shuffle channel attention module. This operation-wise attention learning is performed on three low level layers of Conformer backbone represented with superscript i. In our implementation, we have selected 2 to 4 side-out levels of Conformer. The complementarity of RGB and depth is selected by adaptively fusing parallel multi-modal representations. As described in Eq. 3.4, all layers are concatenated after element-wise multiplication of realigned depth and RGB features.

$$\begin{aligned} lde^i_{out} = c((A^i_{d_1} \times rgb^i_{l_1}), (A^i_{d_2} \times rgb^i_{l_2}), (A^i_{d_3} \times rgb^i_{l_3}), \\ (t_4(A^i_{d_4} \times rgb^i_{l_4}), (t_5(A^i_{d_5} \times rgb^i_{l_5}))), \end{aligned} \tag{3.4}$$

3.2.3 Global Detail Enhancement Module

It is well established that, features extracted from shallow layers of backbone network exhibit modality-specific properties while deeper layer representations are

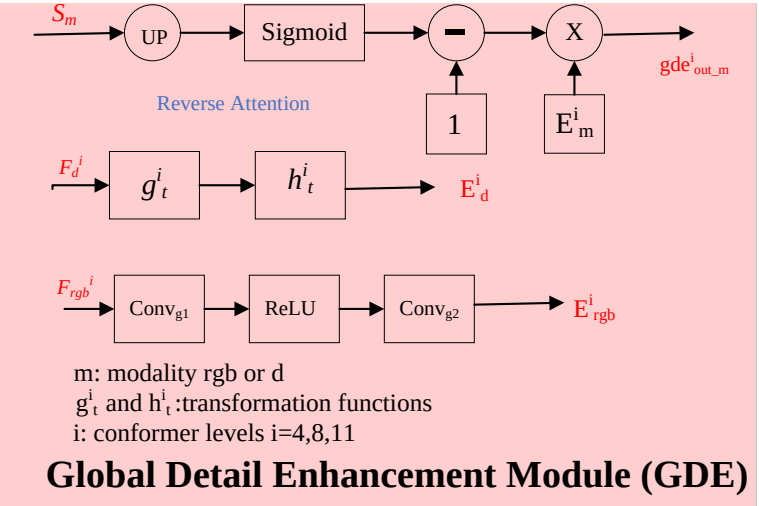


FIGURE 3.6: Detailed architecture of proposed Global Detail Enhancement (GDE) Module.

more task agnostic. Therefore, in proposed model, deep RGB and depth features are handled separately from shallow features in Global Detail Enhancement (GDE) Module. GDE module captures global context aware information from 4,8 and 11 side-outs levels of Conformer backbone. It then performs reverse attention on these and coarse level 12 side-out of Conformer representation in parallel manner. Hence, coarse target localizations helps subsequent level to learn fine details. Subsequently, it combines the enhanced hierarchical features in top-down manner.

The detailed architecture of GDE module can be visualized in Fig. 3.6. The GDE module is designed to generate saliency maps at three levels, represented as low, middle and high level. The reverse attention is performed in two steps:

- Coarse target localization generation.
- Reverse attention learning for saliency maps generation using transformed F^i_{rgb} , F^i_d, S_{rgb} and S_d

3.2.3.1 Coarse Target Localization Generation

For reverse attention scheme opted in GDE module, firstly coarse localization is carried out on high level RGB and depth features in a supervised manner. The raw features from last layer of CNN and transformer streams of Conformer are

supervised using salient ground truth. Then the enhanced target localization is fed to previous layer for attention learning. In the training procedure of coarse RGB and depth features several convolution and transformations are applied. The applied loss function is described by RGB coarse loss $\mathcal{L}_{rgb}$ and depth coarse loss $\mathcal{L}_d$ explained in section 3.7. The Eqs. 3.5 and 3.6 shows the relationship between the raw RGB F^c_{rgb} and raw depth F^c_d with the generated coarse RGB saliency maps S_{rgb} and coarse depth saliency maps S_d.

$$S_{rgb} = Conv_{rs}(\sigma(Conv_{cr}(F^c_{rgb}))), \tag{3.5}$$

$$S_d = Conv_{ds}(\sigma(Conv_{cd}(f_t(F^c_d)))), \tag{3.6}$$

where superscript c=12. In Eq. 3.6 f_t represents feature transformation function used to align the transformer output with CNN output. Firstly, 1x1 convolution ($Conv_{cr}$ and $Conv_{cd}$) followed by σ ReLU activation , is applied on F^c_{rgb} and F^c_d. The resultant features are further enhanced for supervised learning using 1x1 convolution $Conv_{rs}$ and 3x3,stride 2 convolution $Conv_{ds}$ operations on last RGB and depth features of Conformer respectively.

3.2.3.2 Reverse Attention Learning

For reverse attention learning F^i_{rgb} and F^i_d from high level side-outs of Conformer are utilized. In proposed model GDE $i = 4, 8, 11$ side-outs are considered. As in Eq. 3.7 F^i_{rgb} is converted to single channel feature map using two convolution operations $Conv_{g1}$ and $Conv_{g2}$ (1x1,1) followed by σ ReLU activation.

$$E^i_{rgb} = Conv_{g2}(\sigma(Conv_{g1}(F^i_{rgb}))), \tag{3.7}$$

The enhanced depth features E^i_d are obtained at each saliency generation stage by applying g^i_t and h^i_t transformation functions given in Eq. 3.8. g^i_t and h^i_t converts patch embeddings of depth modality to match the RGB modality in spatial resolution. To do so, transposed convolutions is performed to up-sample depth feature maps with kernel size 4x4 and strides of 1,2,4 for high, middle and low

level stage respectively. Along with that convolutional layer of (1x1,1) is used to reduce channel dimension to 1.

$$E_d^i = h_t^i(g_t^i(F_d^i)). \tag{3.8}$$

For reverse attention S_{rgb} and S_d are upsampled using transposed convolution. Stride of 2,4 and 8 for high, middle and low level stage of GDE are implemented. Thus a different upsampled version of RGB and depth coarse saliency followed by Sigmoid are used in each reverse attention stage. As shown in Eq. 3.9, transformed RGB and depth coarse saliency maps are inverted and then multiplied by transformed side-out of RGB and depth feature map. This allows global contextual details to spatially compensate with the intermediate finer details.

$$gde_{out_m}^i = (1 - Sigmoid(UP(S_m))) \times E_m^i, \tag{3.9}$$

where m is rgb or d for RGB and depth modality, respectively.

3.2.3.3 Visualization and Discussion on Side-outs Inference of RGB (E_{rgb}^i) and Depth (E_d^i)

Proposed reverse attention mechanism integrates (S_d, S_{rgb}) from 12^{th} layer with low, middle and high level stage features. As each stage is responsible to capture specific details, the parallel complementary details extraction boost the saliency detection performance. Fig. 3.7 represents side-outs inference of RGB(E_{rgb}^i) and depth(E_d^i) at different layers of backbone. Where, E_{rgb}^{11} and E_d^{11} represents high level saliency maps at layer 11. At this layer, rich semantic details are obtained, however, the resolution of feature maps is low. The side-outs E_{rgb}^8 and E_d^8 give enhanced representation at middle layer 8. While E_{rgb}^4 and E_d^4 side-outs feature maps capture finer details with extended spatial resolution. Therefore, the intrinsic properties of low to high level features are when combined with RGB and depth coarse saliency maps in reverse attention mechanism, the enriched global context is achieved.

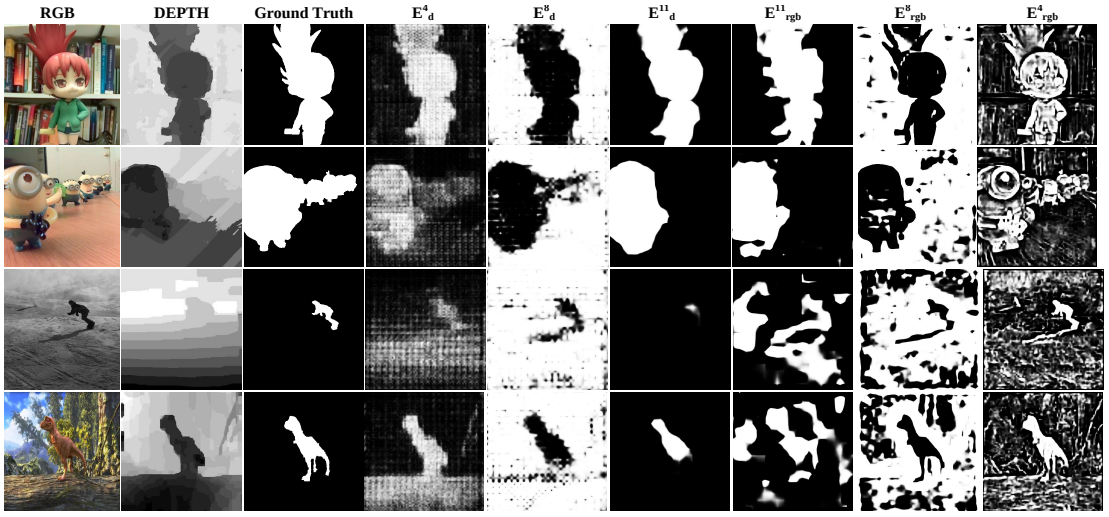


FIGURE 3.7: Side-outs inference of RGB(E^i_{rgb}) and depth(E^i_d) at different levels without combining with deeper levels in reverse attention module.

3.2.4 Integration of LDE and GDE in Decoder Module

To reduce the complexity of model, decoder only integrates the outputs of LDE and GDE module after proper re-alignment. Given in Eq. 3.10, the depth and RGB outputs from high level stage of reverse attention module is added and then upsampled using Transposed convolution with stride of 2. The resultant feature maps are added with middle level depth and RGB representation of GDE. Concurrently, they are upsampled and combined with low level reverse attention stage features and again upsampled. Thus, the hierarchical reverse attention features are added to upsampled version of LDE output. The final output resolution matches the ground truth.

$$S_f = \sum_i UP_i(lde^i_{out}) + UP((gde^l_{out_d} + gde^l_{out_{rgb}}) + UP((gde^m_{out_d} + gde^m_{out_{rgb}}) + (UP(gde^h_{out_d} + gde^h_{out_{rgb}})))). \tag{3.10}$$

3.2.5 Loss Function

The loss function is defined as:

$$\mathcal{L}_t = \mathcal{L}_f(S_f) + \alpha_i \mathcal{L}_E i(S_E i) + \mathcal{L}_d(S_d) + \mathcal{L}_r(S_{rgb}), \tag{3.11}$$

The generated Final S_f, Coarse Depth S_d and Coarse RGB S_{rgb} saliency prediction maps respectively, are trained using ground truth saliency map, using binary cross-entropy loss ($\mathcal{L}_f(S_f)$, $\mathcal{L}_d(S_d)$ and $\mathcal{L}_r(S_{rgb})$). For edge saliency map S_Ei training is done using generated edge maps from ground truth maps using Sobel operator. Here, i=2,3,4 for three LDE module outputs and $\mathcal{L}_Ei(S_Ei)$ is defined as smooth L1 loss. α_i is tuneable parameter.

3.3 Detail Evaluation of CVit-Net

3.3.1 Datasets and Evaluation Metrics

For the training of proposed CVit-Net, 1500 RGB and depth pairs from NJU2K [45] and 700 from NLPR [22] dataset is used as followed by [1, 25, 39, 79]. The proposed model is evaluated on following benchmark datasets: remaining samples of NJU2K and NLPR, LFSD [46], SIP [33], STERE [47] and RGBD135 [21]. For quantitative results evaluation, precision–recall [48], Structural measure (Smeasure) [49], maximum F-measure (Fmax) [50], maximum E-measure (Emax) [51], and Mean Absolute Error (MAE) [52] are adopted. For qualitative analysis, various scenarios reported by SOTA models [1, 33, 68, 71, 73] are used.

3.3.2 Experiment Settings

The implementation of proposed CVit-Net is done on Pytorch. The specification of workstation used for training is given in Table 3.2. RGB and depth input images are resized to 320 × 320 resolution. The single channel depth images are replicated to three channels. Depth input is normalized between 0 to 255 range. Additional data augmentations are added for a generalized solution, including random crop and normalization. The initialization of backbone network is done using Conformer-B [93] pretrained on ImageNet. Other coefficients are initialized by default setting of Pytorch. The empirically selected learning rate of 5e-5 is

used. The total training time is 24h and the proposed model converges in 100 epochs using Adam optimizer. A batch size of 4 is used.

TABLE 3.2: Specification of workstation used for training of CVit-Net.

OS	Linux:Ubuntu 18.04.6 LTS
GPU	Single NVIDIA GeForce RTX 3060 Ti
GPU	Memory 8 GB
CPU	Intel® Core™i7-10700K
RAM	16 GB
System Type	64 bit
Cuda Version	11.8
Interpreter	Python 3.12

Conformer-base backbone configuration In the implementation of CVit-Net for salient object detection, fully connected layers of Conformer-base backbone network is removed. In the proposed model 12 conv-tran layers of backbone are utilized. Other configuration parameters are listed in Table 3.3.

TABLE 3.3: Conformer-B Configuration.

Patch size	16
Channel ratio	6
Embedding dimensions	576
Depth	12
Number of Heads	9
MLP ratio	4

3.3.3 Quantitative Results

To verify the performance of the proposed CVit-Net, the model is compared with 22 SOTA models including SSRCNN[34], D3Net[33], ICNet[82], CMWNet[38], PGARNet[24], BBSNet[81], ASIF-Net[96], CVAE[79], JL-DCF[25], BTSNet[39], RD3D[97], BiANet[98], SSL[99], DIGRNet[100], SPSN[101], LIANet[73], DCMNet [87] CNN-based model and TriTransNett[26], VST[76], MFormer[77], SiaTrans[27], DFTR[102] transformer-based model. . The saliency maps of these models are provided by their corresponding authors. The quantitative comparison of results with these SOTA models is shown in Table 3.4 against S-measure ($S_\alpha \uparrow$), F-measure

($F_\beta^{max}\uparrow$), E-measure ($E_\zeta^{max}\uparrow$) and mean absolute error ($MAE\downarrow$). The upward arrow accompanying the first three metrics indicates that higher values correspond to favorable outcomes, while the downward arrow paired with MAE denotes that lower values correspond to favorable outcomes. The aforementioned metrics offer valuable insights into the precision, comprehensiveness, and overall efficacy of the models in accurately identifying prominent objects within images.

Performance Metrics Analysis:

An in-depth quantitative analysis of the performance of the proposed CVit-Net is presented below:

PR-Curve: The precision-recall curve of proposed CVit-Net and SOTA models is presented in Fig. 3.8. It can be observed that the proposed model has shorter curve in all datasets, indicating high recall.

S-measure (S_α)**:** The significant performance improvement of the proposed model in the S-measure metric for object-level errors indicates a high level of structural similarity and spatial information correlation. Results presented in Table 3.4 shows that proposed CVit-Net achieves a performance boost of 1.57% in NLPR dataset, 1.04% in LFSD dataset, 0.77% in SIP dataset, 0.56% in STERE dataset and 0.95% in RGBD135 dataset in comparison to SOTA models.

F-measure (F_β^{max})**:** The F-measure combines precision and recall into a unified metric, thereby offering a well-rounded evaluation of a model's performance. Precision assesses the correctness of the identified salient regions, indicating how many of the detected regions are truly relevant. Recall, on the other hand, evaluates the model's capacity to capture all the salient regions that exist in the ground truth, indicating the completeness of the detection process. Results presented in Table 3.4 shows that proposed CVit-Net achieves a performance boost of 0.21% in NLPR dataset, 0.53% in LFSD dataset,1.19% in SIP dataset,0.47% in STERE dataset and 0.94% in RGBD135 dataset in comparison to SOTA models. Due to the inherent imbalance between salient and non-salient regions in saliency detection datasets, F-measure emerges as a robust metric for evaluating salient object detection (SOD) models. Specifically, a significant improvement of 0.47% in STERE

dataset can be observed, which is described as highly imbalanced RGB-D SOD dataset. It can be observed in Fig. 3.9, that, the proposed CVit-Net accurately identified salient regions while accounting for the scarcity of salient regions in the provided examples.

E-measure (E_{ζ}^{max})**:** The proposed model demonstrates superior delineation, localization, and segmentation accuracy of salient object regions, as evaluated by the E-measure metric. The model outperforms state-of-the-art (SOTA) models by 1.04% in NJU2K dataset, 0.82% in NLPR dataset, 0.57% in LFSD dataset, 1.15% in SIP dataset, 0.10% in STERE dataset and 0.50% in RGBD135 dataset.

Mean Absolute Error (MAE): It can be demonstrated that the proposed model outperforms SOTA models. Specifically, mean absolute error has been decreased by 15% in SIP dataset, 13.4% in LFSD dataset, 9.1% in STERE dataset, 5.6% in NLPR dataset and 3.6% in RGBD135 dataset.

3.3.4 Qualitative Results

In order to evaluate proposed CVit-Net visually, the 9 different test cases which include 8 complex scenarios are identified. These challenging scenarios have been selected as pointed in various published studies [1, 24, 25, 39, 79, 80, 98]. Visual comparison with SOTA models is provided in Fig. 3.9. In this figure, each row describes different scenarios, provided with saliency maps of proposed and SOTA models.

1. **Depth quality is so poor that no object can be visualized:** This test case is shown in row (a) and (b) of Fig. 3.9. In this complex scenario proposed model successfully captured the salient object. Although saliency map of CVit-Net in row (b) are not only very close to ground truth saliency but the proposed model expressed even small details like the legs of bird, while other SOTA models have included the additional information.

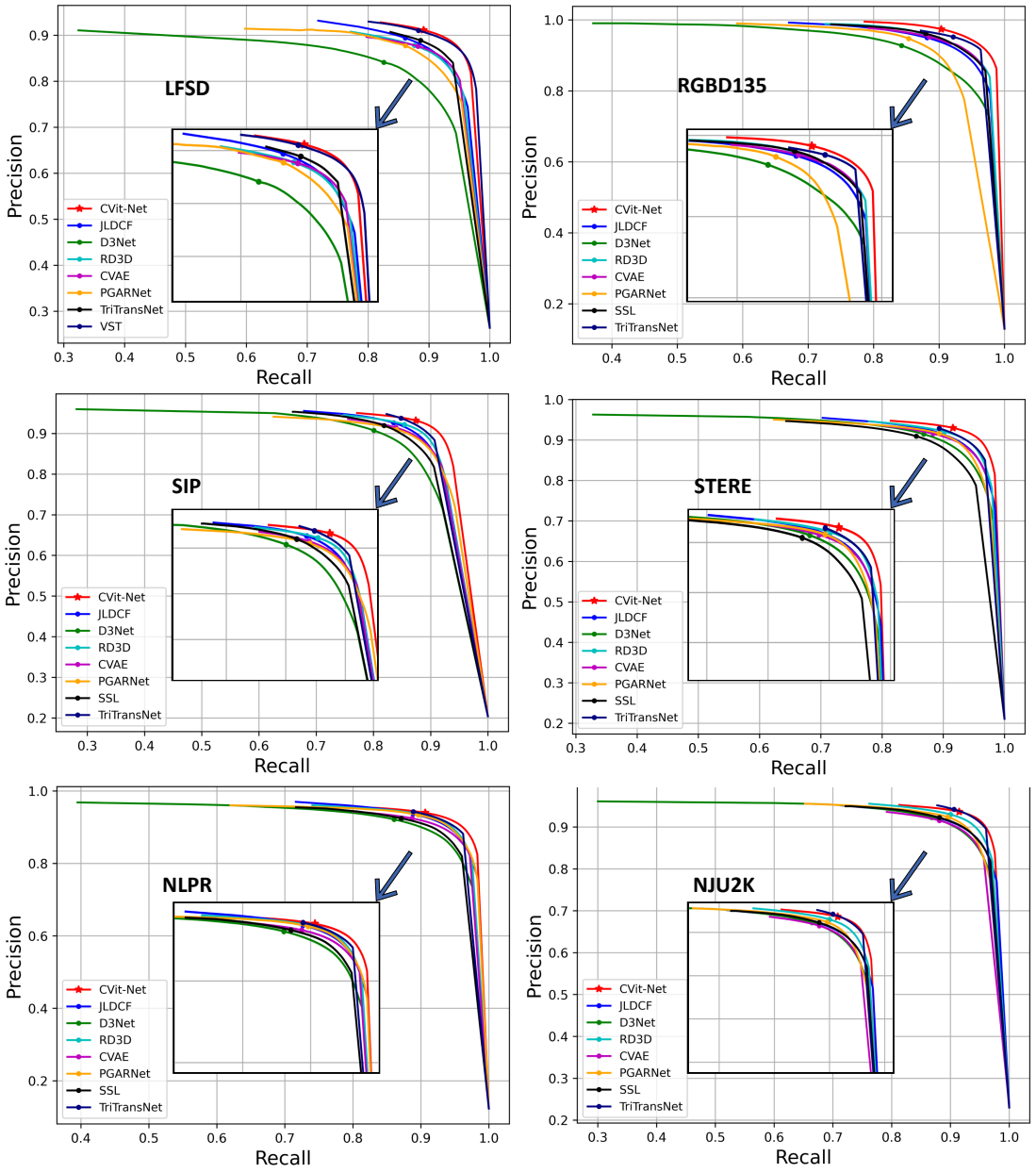


FIGURE 3.8: Precision-recall curves of SOTA methods and proposed CVit-Net across 6 Datasets.

2. **Depth has missed small details:** Fig. 3.9 row (c) is image of chandelier and its depth map has missed small details, still proposed model is able to capture the whole object effectively.

3. **Non-salient object adjoined in depth image:** Bird and the rock in row (d) of Fig. 3.9 are adjoined in depth image. Also legs of the bird and rock have same color in RGB image. Even so, CVit-Net outperforms other SOTA models.

TABLE 3.4: Quantitative comparison of proposed CVit-Net with 22 SOTA CNN and transformer based RGB-D SOD modelson 6 benchmark datasets. Best results are represented by 'Red' color and second best results are represented by 'Blue' color. '-' indicates result is not available.

Datasets	**Quantitative Evaluation on NJU2K [45], NLPR [22]and LFSD [46] Datasets**											
	NJU2K				**NLPR**				**LFSD**			
Models/Metrics	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$
CNN-based Models												
SSRCNN [34]	0.8827	0.8746	0.9087	0.0511	0.8932	0.8526	0.9354	0.0357	0.8587	0.848	0.8859	0.0763
D3Net [33]	0.895	0.889	0.932	0.051	0.906	0.885	0.946	0.034	0.825	0.81	0.862	0.095
ICNet [82]	0.894	0.891	0.926	0.052	0.923	0.908	0.952	0.028	0.868	0.871	0.903	0.071
CMWNet [38]	0.903	0.902	0.936	0.046	0.917	0.903	0.951	0.029	0.876	0.883	0.912	0.066
PGARNet [24]	0.909	0.893	0.916	0.042	0.93	0.885	0.955	0.024	0.853	0.852	0.889	0.074
BBSNet [81]	0.921	0.92	0.949	0.035	0.93	0.918	0.961	0.023	0.864	0.858	0.901	0.072
ASIFNet [96]	0.8887	0.9007	-	0.0471	0.8844	0.9002	-	0.0298	0.8391	0.8723	-	0.0754
CVAE [79]	0.902	0.893	0.937	0.039	0.917	0.893	0.952	0.025	0.868	0.857	0.904	0.065
JL-DCF [25]	0.903	0.903	0.944	0.043	0.925	0.916	0.962	0.022	0.862	0.866	0.901	0.071
BTSNet [39]	0.921	0.924	0.954	0.036	0.934	0.923	0.965	0.023	0.867	0.874	0.906	0.07
RD3D [97]	0.916	0.914	0.947	0.036	0.93	0.919	0.965	0.022	-	-	-	-

.....Continued on next page.....

TABLE 3.4: Quantitative comparison of proposed CVit-Net with 22 SOTA CNN and transformer based RGB-D SOD modelson 6 benchmark datasets. Best results are represented by 'Red' color and second best results are represented by 'Blue' color. '-' indicates result is not available.

Datasets	**NJU2K**				**NLPR**				**LFSD**			
Models/Metrics	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$
BiANet[98]	0.915	0.92	0.948	0.039	0.925	0.914	0.961	0.024	-	-	-	-
SSL [99],	0.909	0.923	0.939	0.038	0.922	0.923	0.96	0.025	-	-	-	-
DIGRNet[100]	0.933	0.939	0.966	0.026	0.931	0.923	0.964	0.021	0.869	0.869	0.905	0.067
SPSN [101]	0.918	0.92	0.95	0.032	0.923	0.91	0.958	0.023	-	-	-	-
LIANet [73]	0.904	0.911	0.911	0.042	-	-	-	-	0.862	0.884	0.889	0.07
DCMNet [87]	-	0.899	0.92	0.036	-	0.883	0.954	0.024	-	0.867	0.906	0.064
Transformer-based Models												
TriTransNet [26]	0.92	0.919	0.925	0.03	0.928	0.909	0.96	0.02	-	-	-	-
VST [76]	0.922	0.92	0.951	0.035	0.932	0.92	0.962	0.024	0.882	0.889	0.921	0.061
Mformer [77]	0.922	0.923	0.954	0.032	0.932	0.925	0.965	0.021	0.872	0.879	0.911	0.062
SiaTrans [27]	0.923	0.921	0.956	0.035	0.929	0.918	0.964	0.024	0.871	0.876	0.907	0.069
DFTR [102]	0.922	0.923	0.954	0.034	0.941	0.934	0.972	0.018	-	-	-	-
CVit-Net	0.924	0.926	0.966	0.03	0.956	0.936	0.98	0.017	0.8913	0.8937	0.9263	0.0528

.....Continued on next page.....

TABLE 3.4: Quantitative comparison of proposed CVit-Net with 22 SOTA CNN and transformer based RGB-D SOD modelson 6 benchmark datasets. Best results are represented by 'Red' color and second best results are represented by 'Blue' color. '-' indicates result is not available.

Datasets	**Quantitative Evaluation on SIP [33],STERE [47] and RGBD135 [21] Datasets**											
	SIP				**STERE**				**RGBD135**			
Models/Metrics	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$
CNN-based Models												
SSRCNN [34]	-	-	-	-	0.8822	0.8632	0.9146	0.0499	-	-	-	-
D3Net [33]	0.86	0.861	0.909	0.063	0.899	0.891	0.938	0.046	0.898	0.885	0.946	0.031
ICNet [82]	-	-	-	-	0.903	0.898	0.942	0.045	0.92	0.913	0.96	0.027
CMWNet [38]	0.867	0.874	0.913	0.062	0.905	0.901	0.944	0.043	0.934	0.93	0.969	0.022
PGARNet [24]	0.876	0.854	0.908	0.055	0.907	0.88	0.919	0.041	0.913	0.88	0.939	0.026
BBSNet [81]	0.879	0.883	0.922	0.055	0.908	0.903	0.942	0.041	0.933	0.927	0.966	0.021
ASIFNet [96]	-	-	-	-	0.8778	0.8953	-	0.0474	-	-	-	-
CVAE [79]	0.883	0.877	0.927	0.045	-	-	-	-	0.937	0.929	0.975	0.016
JL-DCF [25]	0.879	0.885	0.923	0.051	0.905	0.901	0.946	0.042	0.929	0.919	0.968	0.022
BTSNet [39]	0.896	0.901	0.933	0.044	0.915	0.911	0.949	0.038	0.943	0.94	0.979	0.018
RD3D [97]	0.885	0.889	0.924	0.048	0.911	0.906	0.947	0.037	0.935	0.929	0.972	0.019

.....Continued on next page.....

TABLE 3.4: Quantitative comparison of proposed CVit-Net with 22 SOTA CNN and transformer based RGB-D SOD modelson 6 benchmark datasets. Best results are represented by 'Red' color and second best results are represented by 'Blue' color. '-' indicates result is not available.

Datasets	**SIP**				**STERE**				**RGBD135**			
Models/Metrics	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$
BiANet [98]	0.883	0.89	0.925	0.052	0.904	0.898	0.942	0.043	0.931	0.926	0.971	0.021
SSL [99]	0.888	0.909	0.927	0.046	0.904	0.914	0.939	0.039	0.936	0.944	0.978	0.017
DIGRNet [100]	0.882	0.892	0.925	0.051	0.914	0.916	0.957	0.035	0.936	0.932	0.972	0.019
SPSN [101]	0.892	0.899	0.934	0.042	0.907	0.9	0.943	0.035	0.937	0.936	0.974	0.016
LIANet[73]	0.884	0.912	0.923	0.048	0.906	0.914	0.929	0.037	0.929	0.938	0.969	0.021
DCMNet [87]	-	0.883	0.926	0.047	-	-	-	-	-	-	-	-
Transformer-based Models												
TriTransNet [26]	0.886	0.892	0.924	0.043	0.908	0.893	0.927	0.033	0.943	0.936	0.981	0.014
VST [76]	0.904	0.915	0.944	0.04	0.913	0.907	0.951	0.038	0.943	0.94	0.978	0.017
MFormer [77]	0.894	0.902	0.932	0.043	-	-	-	-	-	-	-	-
SiaTrans [27]	0.899	0.913	0.945	0.041	0.914	0.907	0.951	0.038	0.936	0.932	0.975	0.02
DFTR [102]	0.904	0.913	0.946	0.04	0.918	0.914	0.951	0.034	-	-	-	-
CVit-Net	0.911	0.926	0.957	0.034	0.9232	0.9203	0.958	0.03	0.952	0.953	0.9859	0.0135

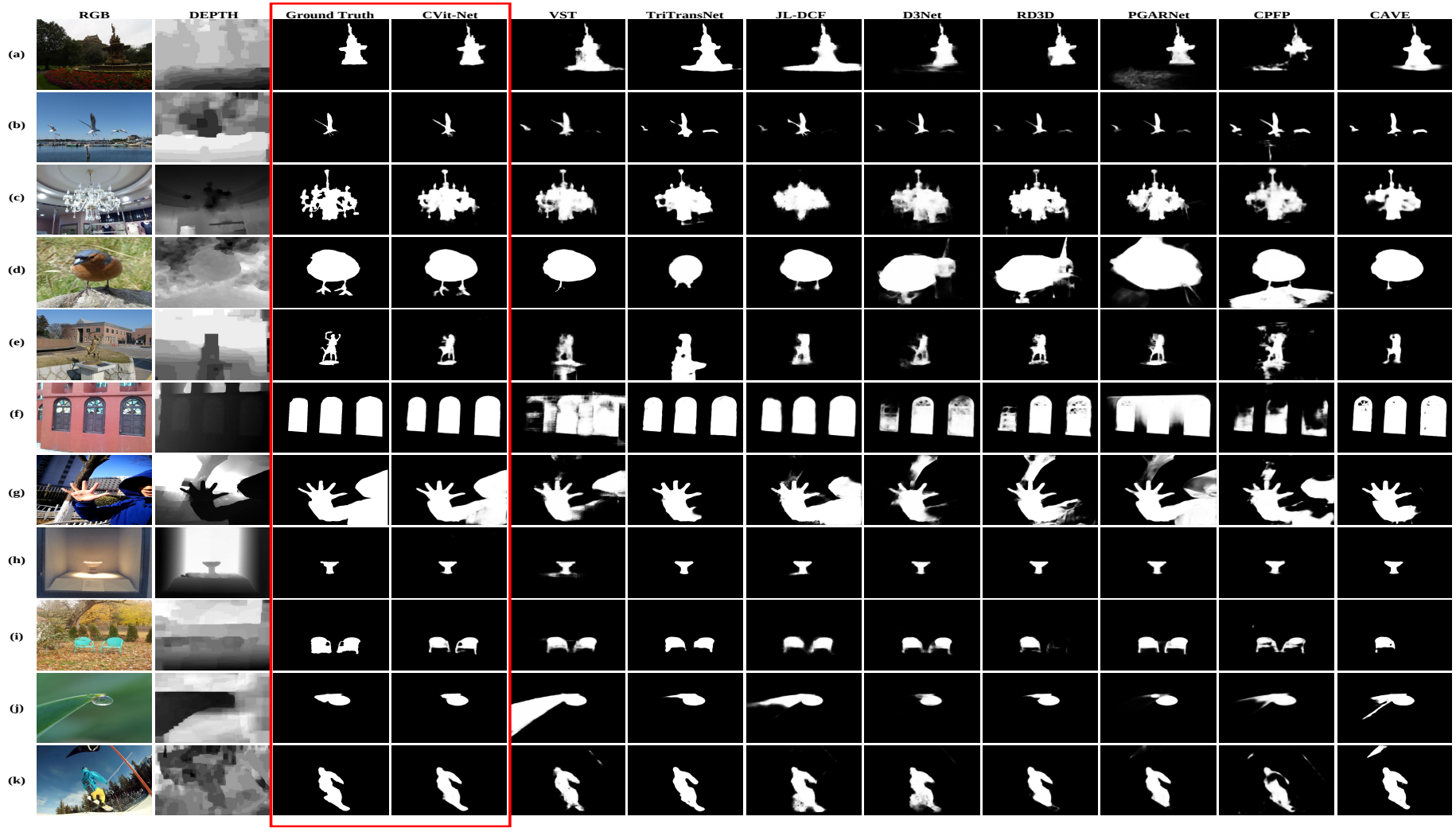


FIGURE 3.9: Visual comparison of SOTA methods and proposed CVit-Net in different challenging scenes.

4. **Depth highlights additional details:** Fig. 3.9 row(e) represents complex scenario of low contrast RGB with poor quality depth. In this challenging scene, CVit-Net perform pretty well as compared to other models.

5. **Good contrast RGB and low contrast depth:** Row (f) of Fig. 3.9 depicts that the proposed model can represent objects with low contrast depth.

6. **Good contrast RGB and depth:** Although row (g) of Fig. 3.9 is simple test case, however, many models fail to capture the details of fingers while CVit-Net show impressive result.

7. **Small objects:** Row (h) of Fig. 3.9 shows that proposed model can effectively capture small object.

8. **Multiple objects:** Another challenging case is multiple salient objects in a scene presented in Fig. 3.9 row (i). Proposed CVit-Net well represented all salient objects.

9. **Complex background:** Fig. 3.9 row (j) represent indistinguishable foreground and background scenario due to same color appearance while row(k) represents cluttered background test case. It can be visualized that the proposed model notably performed well as compared to other SOTA models.

3.3.5 Ablation Studies

In this section, a comprehensive ablation studies is presented to verify the contribution of each module of proposed CVit-Net. To conduct ablation analysis, four datasets SIP and RGBD135 (having good quality depth maps), LFSD and STERE (having poor quality depth map) are selected. The full implementation of CVit-Net used as a reference model is denoted as 'Model F'. All ablated models are compared with 'Model F' and visual comparison is presented in Fig. 3.10.

Effectiveness of reverse attention module: The global contextual information from deep layers of Conformer are processed in reverse attention module of

CVit-Net. To verify the effectiveness of reverse attention module, the local detail enhancement module is removed and model is trained with only global detail enhacement module, denoted as 'Model A'. The reverse attention utilize coarse saliency from last hierarchy of Conformer and generate output in top-down manner. Comparative analysis of 'Model F' and 'Model A' is provided in Table 3.5. It is evident that results of LFSD and STERE (known for their poor quality depth images) are very close to reference 'Model F', which depicts that reverse attention itself is powerful technique to handle low quality depth map. Visual comparison is given in Fig. 3.10.

TABLE 3.5: Ablation study about the role of reverse attention module in RGBD SOD. The best result is in **Bold**. 'Model A' is model having only reverse attention module and 'Model F' is full implementation of CVit-Net.

Datasets/Metrics	Model	$S_\alpha \uparrow$	$F_{\max} \uparrow$	$E_{\max} \uparrow$	MAE $\downarrow$
SIP	Modal A	0.8958	0.9018	0.9328	0.0432
	Modal F	**0.911**	**0.926**	**0.957**	**0.034**
STERE	Modal A	0.9232	0.9203	0.956	0.032
	Modal F	**0.9232**	**0.9203**	**0.958**	**0.03**
LFSD	Modal A	0.8867	0.8866	0.9171	0.0595
	Modal F	**0.8913**	**0.8937**	**0.9263**	**0.0528**
RGBD135	Modal A	0.9499	0.9462	0.9806	0.0151
	Modal F	**0.952**	**0.953**	**0.9859**	**0.0135**

Effectiveness of Depth Modality using 'RGB only': Does depth modality really enhances the model performance? To answer this question, a model is trained, denoted as 'Model B', with only RGB input. The results of 'Model B' and reference model 'Model F' are compared and provided in Table 3.6. Results validates the role of depth modality in enhancing the detection performance in almost every dataset. With the exception in STERE dataset, where only slight improvement is observed in 'Model F' when compared with 'Model B'. This is justifiable as STERE data has very poor quality depth maps. In next paragraph, the proposed model is evaluated using 'Depth only' input. To conclude the effectiveness of RGB input over depth due to presence of rich information in RGB modality, illustrated in 1^{st} and 2^{nd} row of Fig. 3.10 additional details are captured due to rich textured RGB image.

TABLE 3.6: Effectiveness analysis of multi-modal RGB-D input compared to uni-modal input for SOD task. The best result is in **Bold**.

Datasets/Metric/Models		Depth only 'Model C'	RGB only 'Model B'	'Model F'
SIP	$S_\alpha \uparrow$	0.8701	0.8823	**0.911**
	$F_{\max} \uparrow$	0.885	0.8871	**0.926**
	$E_{\max} \uparrow$	0.9165	0.9283	**0.957**
	MAE $\downarrow$	0.0552	0.049	**0.034**
STERE	$S_\alpha \uparrow$	0.7674	0.924	**0.9232**
	$F_{\max} \uparrow$	0.7409	0.9198	**0.9203**
	$E_{\max} \uparrow$	0.8576	0.9573	**0.958**
	MAE $\downarrow$	0.1015	0.0306	**0.03**
LFSD	$S_\alpha \uparrow$	0.7612	0.8699	**0.8913**
	$F_{\max} \uparrow$	0.7636	0.8691	**0.8937**
	$E_{\max} \uparrow$	0.8277	0.9073	**0.9263**
	MAE $\downarrow$	0.1175	0.0676	**0.0528**
RGBD135	$S_\alpha \uparrow$	0.9186	0.908	**0.952**
	$F_{\max} \uparrow$	0.9109	0.8968	**0.953**
	$E_{\max} \uparrow$	0.9676	0.9442	**0.9859**
	MAE $\downarrow$	0.0244	0.0254	**0.0135**

Replacement of Shuffle Channel Attention with Self Attention: The proposed shuffle channel attention correlates the structural information of depth edge saliency. To validate its importance, Shuffle Channel Attention is replaced with the self attention mostly used in Transformer network and the results are demonstrated in Table 3.7. It can be observed that it performed better than 'reverse attention only' given in Table 3.5 but still inferior to CVit-Net. This proves that Shuffle Channel Attention is best possible choice for depth images.

Without edge loss: In the proposed model, the edge details of RGB and depth are correlated with saliency edges in supervised manner. To validate the importance of edge loss, the model is trained without edge supervision and denote it as 'Model E'. Results of 'Model E' and reference model 'Model F' is presented in Table 3.8. The predictions obtained from SIP and RGBD135 are more inferior in 'Model E' than the predictions of STERE and LFSD when compared with reference 'Model F'. It can be concluded that for good quality depth maps edge guidance is more worthwhile in SOD task, while for low quality depth maps it is slightly superior than model without edge guidance. As shown in Fig. 3.10 results,

TABLE 3.7: Ablation study about the role of Operation-wise shuffle channel attention in proposed CVit-Net. Channel shuffle attention is replaced with self attention and the new model is called 'Model D'. The best result is in **Bold**.

Datasets/Metrics/Models		'Model D'	'Model F'
SIP	$S_\alpha \uparrow$	0.9022	**0.911**
	$F_{\max} \uparrow$	0.9093	**0.926**
	$E_{\max} \uparrow$	0.9414	**0.957**
	MAE $\downarrow$	0.0389	**0.034**
STERE	$S_\alpha \uparrow$	0.923	**0.9232**
	$F_{\max} \uparrow$	0.9202	**0.9203**
	$E_{\max} \uparrow$	0.9577	**0.958**
	MAE $\downarrow$	0.0312	**0.03**
LFSD	$S_\alpha \uparrow$	0.8817	**0.8913**
	$F_{\max} \uparrow$	0.8818	**0.8937**
	$E_{\max} \uparrow$	0.9163	**0.9263**
	MAE $\downarrow$	0.0603	**0.0528**
RGBD135	$S_\alpha \uparrow$	0.9457	**0.952**
	$F_{\max} \uparrow$	0.9403	**0.953**
	$E_{\max} \uparrow$	0.9757	**0.9859**
	MAE $\downarrow$	0.0158	**0.0135**

edge loss plays a vital role in accuracy of model. In Fig. 3.10, 1^{st} row having low contrast depth depicts that without edge loss results deteriorate more.

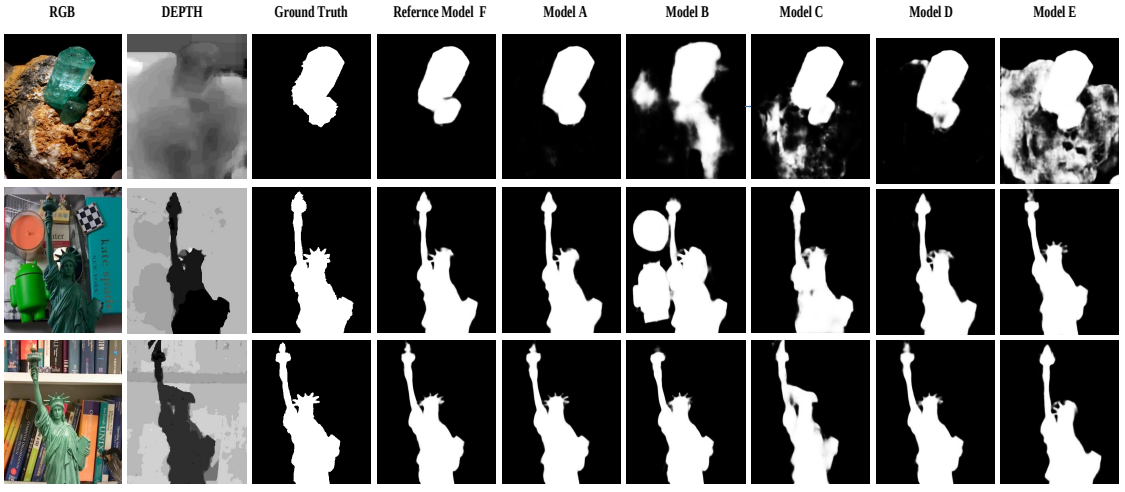


FIGURE 3.10: Visual examples for ablation studies. Reference 'Model F' is full implementation of CVit-Net.

Model complexity analysis: In Table 3.9, a detailed analysis of time-space complexity of proposed CVit-Net is provided and its comparison with other SOTA models is presented. In the main contribution of research, operation-wise shuffle

TABLE 3.8: Ablation study about the role of edge guidance in proposed CVit-Net. 'Model E' is trained without edge loss. The best result is in **Bold**.

Datasets/Metrics/Models		'Model E'	'Model F'
SIP	$S_\alpha \uparrow$	0.8963	**0.911**
	$F_{\max} \uparrow$	0.901	**0.926**
	$E_{\max} \uparrow$	0.9355	**0.957**
	MAE $\downarrow$	0.0423	**0.034**
STERE	$S_\alpha \uparrow$	0.9238	**0.9232**
	$F_{\max} \uparrow$	0.919	**0.9203**
	$E_{\max} \uparrow$	0.9557	**0.958**
	MAE $\downarrow$	0.0319	**0.03**
LFSD	$S_\alpha \uparrow$	0.8911	**0.8913**
	$F_{\max} \uparrow$	0.8897	**0.8937**
	$E_{\max} \uparrow$	0.9233	**0.9263**
	MAE $\downarrow$	0.0547	**0.0528**
RGBD135	$S_\alpha \uparrow$	0.943	**0.952**
	$F_{\max} \uparrow$	0.9397	**0.953**
	$E_{\max} \uparrow$	0.9746	**0.9859**
	MAE $\downarrow$	0.0161	**0.0135**

channel attention framework is proposed, which implements dilated convolution to increase receptive field with fewer parameters. Mostly, single stream methodology is opted to reduce the computational cost. But early or late fusion in single stream have lower accuracies. Therefore, the choice of Conformer network has twofold benefits: a multi-level fusion strategy can be utilized and using a single Conformer network reduces the number of parameters . The number of parameters of the proposed CVit-Net are 90.23M and inference speed in terms of frames per second (FPS) is about 12 FPS. Mostly, backbone network highly influences the performance of object detection models. Therefore, in Table 3.9 it can be observed that computational complexity of the CVit-Net is mostly due to backbone network.

The performance of Conformer-S and Conformer-B as backbone network in the proposed CVit-Net is also compared. Although, the number of parameters of Conformer-B based CVit-Net is almost double of the CVit-Net with Conformer-S, but it can be observed a 39% decrease in MAE metric in SIP dataset and 26% decrease in MAE metric in STERE dataset. Furthermore, the MAE metric and model size of other SOTA CNN and transformer based models are reported. It can

TABLE 3.9: Ablation study about model complexity on three efficiency metrics (Number of parameters, model size in MB and FPS computed on NVIDIA P100) and MAE accuracy metric.

Methods	**Backbone**	**Param.(M)**	**Size(MB)**	**FPS**	**SIP(MAE)**	**STERE(MAE)**
JL-DCF [25]	Resnet101	143.5	547.8	1.36	0.051	0.042
SwinNet [17]	SwinTransformer	198.7	785.7	10	0.035	0.033
CVit-Net (Backbone)	Conformer-B	83.2	318	-	-	-
CVit-Net (Backbone+LDE)	Conformer-B	87.2	333.13	-	-	-
CVit-Net (Backbone+LDE +GDE)	Conformer-B	90.2	344	-	-	-
CVit-Net (Backbone+LDE +GDE+Decoder)	Conformer-B	90.23	344.3	12	0.034	0.03
CVit-Net (Backbone+LDE +GDE+Decoder)	Conformer-S	41.5	158.5	16	0.056	0.041

be seen that the proposed model achieves lowest MAE value with fewer number of parameters.

Trade-off between accuracy and efficiency: Accuracy refers to how well a model can identify and locate salient object while efficiency defines how much time and computational resources it will take to do so. To visualize the efficiency and efficacy comparison with SOTA models, max F-Measure is averaged over six benchmark datasets (as provided in various published studies [74, 103]) and plots of max F-measure against FPS and number of parameters are provided in Fig 3.11 and 3.12, respectively. Trade-off visualization of computational efficiency in terms of frames per second and max F-measure shows a balance for proposed CVit- Net. Although two SOTA models have high FPS than the proposed model, but with highly compromised accuracy. For the graph between number of parameters versus max-Fmeasure, proposed CVit-Net, holds a favourable position in the topleft quadrant of the chart, signifying a beneficial balance between accuracy and efficiency. A comparison of max F-Measure, MAE, and Model Size of different models are shown in Fig. 3.13. The size of the circle represents the model size. It's worth noting that superior models are situated in the top left corner, characterized by a larger max F-measure and a smaller MAE. Methods with a smaller size demonstrate inferior performance, highlighting the efficiency and accuracy of

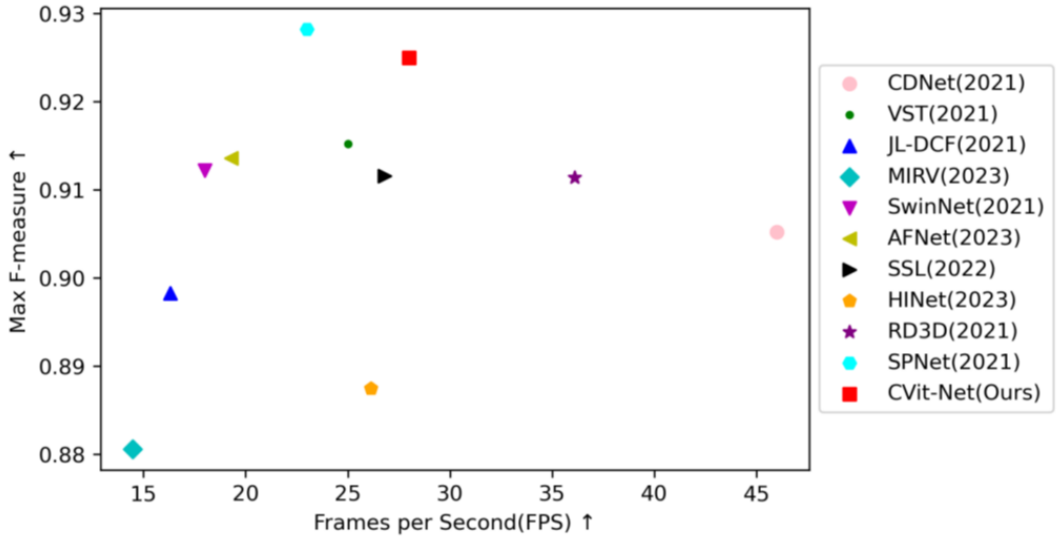


FIGURE 3.11: FPS vs max F-measure.

the proposed CVit-Net approach. All the efficiency parameters are computed on machine with specifications provided in Table 3.3.

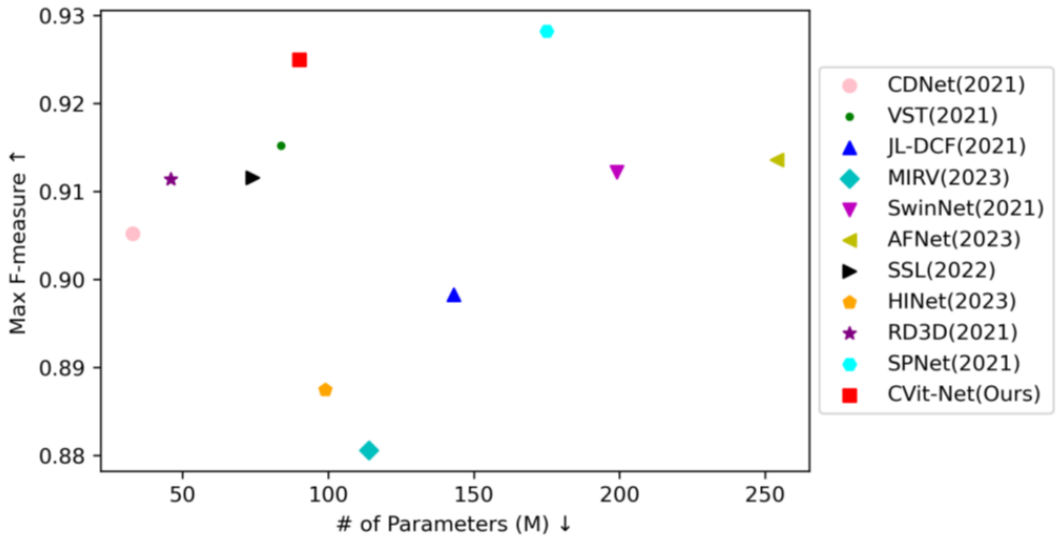


FIGURE 3.12: Number of Parameters vs max F-measure.

Failure cases: The failure cases of the proposed methodology are carefully analysed. It is observed that our Global Detail Enhancement (GDE) module covers a large receptive field. Consequently, large salient foreground detection can lead to false positive predictions. Specifically, when salient and less-salient objects are prominent in both modalities. Therefore, in the future work, saliency ranking can be considered to minimize the false positive predictions. Some examples of failure cases are shown in Fig. 3.14.

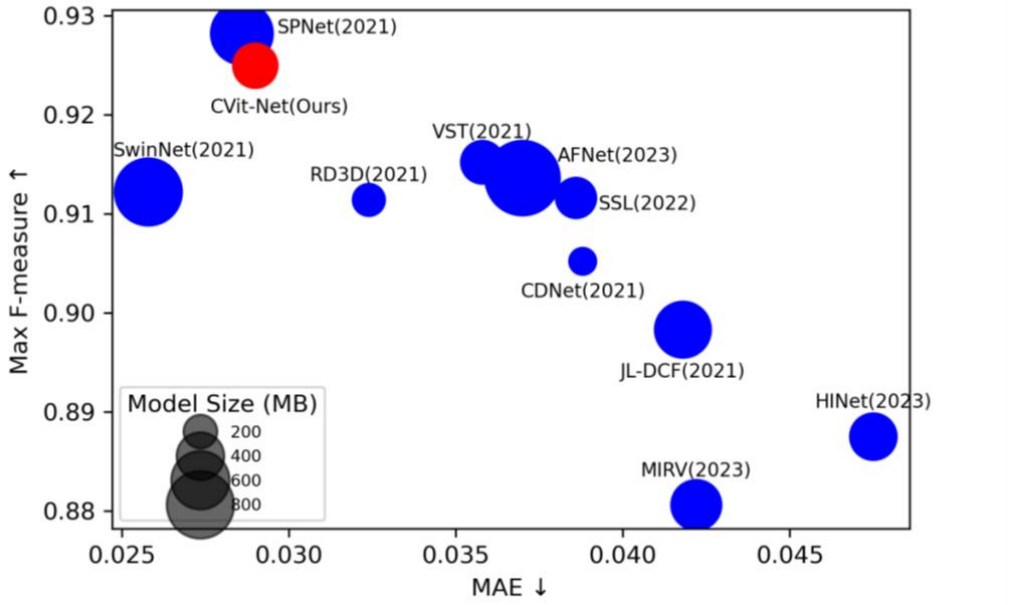


FIGURE 3.13: Average Max F-measure, MAE and Model SIze.

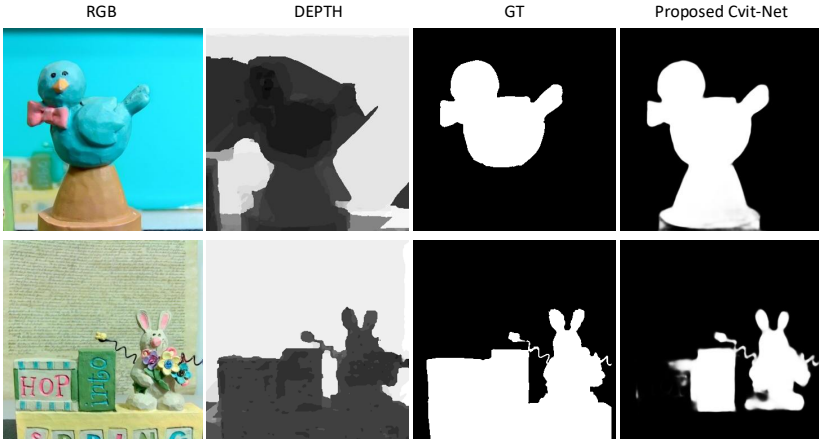


FIGURE 3.14: Failure cases.

3.4 Conclusion

In this chapter, explicit depth-aware salient object detection model was developed to mitigate the impact of low quality depth maps on detection accuracy. The proposed CVit-Net effectively subjugate the local details via Local Detail Enhancement (LDE) module and global details via Global Detail Enhancement (GDE) module at the same time, which also benefit the modality-specific characteristics of RGB and depth. For feature extraction, Conformer encoder is used as the backbone, which consists of one CNN stream and one transformer stream.

Keeping in mind the discrepancy of RGB and depth modality, rich textural details of RGB modality are extracted from CNN stream while straight forward geometrical details of depth modality are extracted from transformer stream. Additionally, novel operation-wise shuffle channel attention is proposed to make full use of complementarity of two modalities, which plays vital role in edge guidance network and enhance the local details along with explicitly assessing low quality depth map. Moreover, salient object area details are captured in reverse attention module in coarse to fine manner. A light-weight decoder network is designed to refine the affluent saliency maps generated from GDE module, with the edge maps generated from LDE module. The comprehensive evaluation of proposed CVit-Net on six publicly available datasets demonstrated its efficacy in comparison to the state-of-the-art RGB-D SOD models. Specifically, MAE metric has been decreased by 15% in SIP dataset, 13.4% in LFSD dataset, 9.1% in STERE dataset, 5.6% in NLPR dataset and 3.6% in RGBD135 dataset. CVit-Net also advances the state-of-the-art RGB-D SOD models by an average of 1.0% (S-measure), 0.7% (max F-measure) and 0.63% (max E-measure) across five datasets. Furthermore, the visual comparison has been reported in 9 different complex scenarios, selected according to the reported challenges in literature, and the comparison of the proposed CVit-Net with SOTA RGB-D SOD models shows a notable performance improvement. Ablation analysis shows that in the absence of novel operation-wise shuffle channel attention and supervised edge guidance, the average $F_{\beta}^{max}/S_{\alpha}$ drop by 1.1% /0.7% and 1.2% /0.6% respectively across four datasets. The asymmetric feature representation methodology presented in this work, will allow to effectively fuse scarce depth data with large number of RGB data in the future work.

Chapter 4

Incomplete RGB-D Modality for Saliency Detection

This chapter introduces the second major contribution of research, implementing implicit depth-aware salient object detection model. Multi-modality learning has significantly improved detection accuracies, specifically, for inconsistent foreground-background, illumination changes, cluttered background and similar textures etc. RGB and depth are combined to overcome challenges posed by using only RGB modality. However, depth has its own limitations. Noisy depth degrades the performance. Earlier implicit depth quality aware models utilize weighting approach or estimate saliency cue dependent depth images from RGB. Then the rectified depths are used in saliency detection models. These methods lack generalization ability, specifically, for low contrast RGB and cluttered background. To address the challenges in existing work, a novel depth quality assessment model is proposed to know the depth quality in advance. Based on the quality score, low quality depth are discarded and salient object detection model is trained with two types of data (i) complete RGB and depth pair and (ii) RGB present depth missing. The novel framework is called incomplete RGB-D modality salient object detection. Fig. 4.1 illustrates vanilla depth-quality aware models and incomplete multi-modality model. Fig. 4.1 (a) shows DASNet [83] model which utilizes RGB modality to generate depth correction matrix and adjusts the contribution of depth

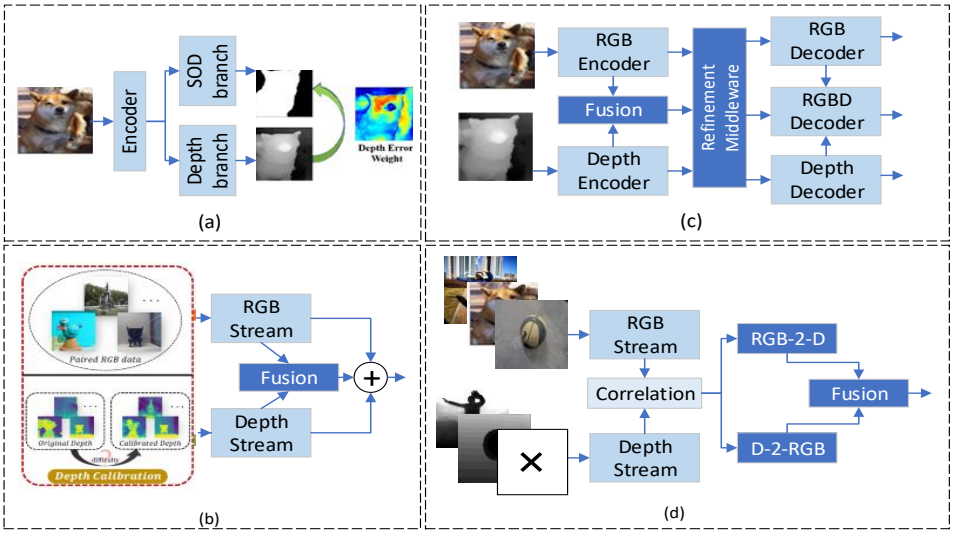


FIGURE 4.1: Depth quality aware SOD models comparison. (a) DASNet [83]. (b) DCF [85]. (c) CIR-Net [72]. (d) Proposed framework.

for noisy depth data. In Fig. 4.1 (b) DCF [85] model is represented that calibrates the original depth using estimated depth, original depth and weighting approach. In Fig. 4.1 (c) CIR-Net [72] model enhances the representation of two modalities using refinement layer. Unlike these model, my proposed model presented in Fig. 4.1 (d) discards low quality depth and detects salient object using complete RGB-D pair or with missing depth. Therefore, the proposed model is robust to data scarcity and evaluation results show notable performance gain.

Research Contribution Overview

The main contributions of implicit depth-quality aware SOD is as follows:

1. A new learning paradigm called incomplete multi-modality saliency learning for salient object detection is proposed. The proposed model is robust to missing depth due to noisy depth or depth scarcity. To assess the quality of depth images, each depth is passed to proposed depth quality assessment regression (DQAR) module and based on its score, low quality depths are discarded.

2. The saliency detection model operates on two types of training data, RGB with high quality depth and RGB with missing depth.

3. The Conformer as backbone network is utilized and shallow and deep features are processed separately in proposed Shallow Common Latent Representation (SCLR) Module and Deep Common Latent Representation (DCLR) module, respectively.

4.1 Proposed Model

The proposed incomplete multi-modality saliency detection framework consists of two modules.

1. Depth quality assessment module: Existing methods rectify low quality depth using RGB and saliency cues. Some depth images are so blurred and noisy that meaningful information cannot be collected. Subsequently, the appropriate solution is to discard the depth images. To distinguish between low and high quality depth, a novel depth quality assessment regression (DQAR) module is proposed, which generates a quality score for each depth image. Large value of quality score represents depth images of high quality and vice versa. In the proposed model, an empirically selected threshold is used to apprehend depth quality. All depth images with quality score below than threshold are treated as low quality and discarded.

2. Salient object detection module: This module is trained with two types of training data based on output of DQAR module, complete RGB-D pair and RGB with missing depth. A Conformer network is used to extract RGB and depth features and separate processing modules for spatial and contextual feature representations are proposed. The learning pipeline consists of three steps as depicted in Fig. 4.2. Three step learning process is as follows:

   (a) Conceal: The extracted features from backbone network are concealed for saliency detection task. The shallow features are concealed keeping in view that they exhibit modality-specific characteristics and are task agnostic. Characteristics of deep features are opposite to shallow one,

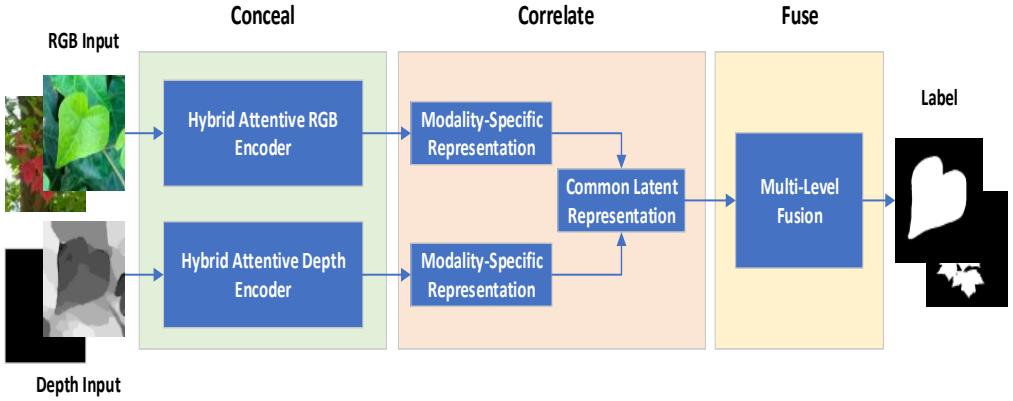


FIGURE 4.2: Proposed incomplete multi-modality SOD learning framework.

therefore, hybrid attentive encoding scheme is used. The discriminative RGB and depth features are forwarded to next module for further processing.

(b) Correlate: Although RGB and depth have disparities but as they represent same scene therefore a strong correlation also exists between them. The correlation between cross-modalities is utilized to perform incomplete RGB-D saliency detection. For this purpose, modality specific information is mapped to set of independent parameters and multi-modality is correlated.

(c) Fuse: Fusion is just a simple block with feature alignment and summation. In this way, effective fusion is done.

In the following subsections, the implementation of each module and parts will be discussed in detail.

4.1.1 Depth Quality Assessment Module

The depth images have inherent uncertainty in their quality due to number of reasons. During image acquisition assorted deteriorations can affect depth quality, due to inaccurate depth measurement or various scene conditions such as occlusion, lighting effect, appearance changes etc. The low quality depth has negative

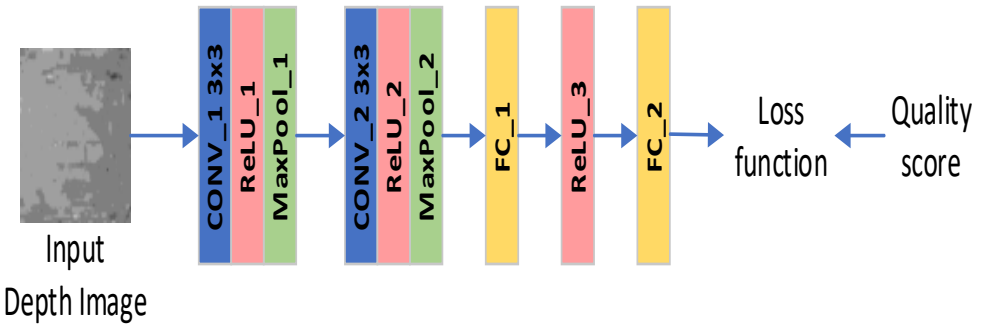


FIGURE 4.3: Proposed incomplete multi-modality SOD learning framework.

effect on detection accuracy. To overcome these challenges, a novel depth quality assessment regression module is proposed, shown in Fig. 4.3.

Mostly in real world, ground truth depth images or subjective score such as mean opinion score (MOS) are not available. Subsequently, no reference image quality assessment methods came into existence. In the proposed model, a set of distorted images from high quality SIP [33] dataset is generated. Then root mean score error metric is used to generate a quality score for each distorted images with respect to reference image. This score acts as a pseudo ground truth in training stage. Common degradations such as noise, motion blur, occlusion, incomplete depth, depth discontinuities and depth maps quantization of depth images are added to SIP dataset. Distorted depth images along with quality score labels train DQAR model. Framework of DQAR module consists of 2 convolution layers (kernel size:3x3 , stride:1), followed by activation function and pooling layer. In proposed model Rectified Linear Unit(ReLU) is used as activation and Max Pooling (kernel size:2x2) is applied. Two fully connected layers are added to generate quality score. Between linear layers ReLU non-linearity is applied. Model is trained using Squared L2 Norm loss function. The proposed DQAR converges within 40 epochs. Adam optimizer is used and learning rate of 0.001 is selected.

For inference, MCL-3D [104]: a database for stereoscopic image quality assessment is used. This dataset consists of 648 images with their subjective MOS labels. MCL-3D dataset covers 9 different scenes captured from 3 views. 6 different distortions are present in this dataset and there are 4 distortion levels. To validate the performance of DQAR model, comparison with state-of-the-art depth

quality assessment models including BIQA [105], SEP [106], BPR [107] and DSS [108] is provided in Table 4.1. Evaluation metrics of Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SRCC) and Root mean squared error (RMSE) are used. Table 4.1 demonstrates that proposed DQAR advances by 7.5%(Pearson linear correlation coefficient) and 18.6%(Spearman rank order correlation coefficient) when comparing to the latest depth quality assessment SOTA models. Root mean squared error of proposed DQAR has been decreased by 78%. Fig. 4.4 presents some examples of depth images with their quality score. It has been noticed that good quality depth images have high quality score α_q and usually salient object prominently stand out from its surroundings.

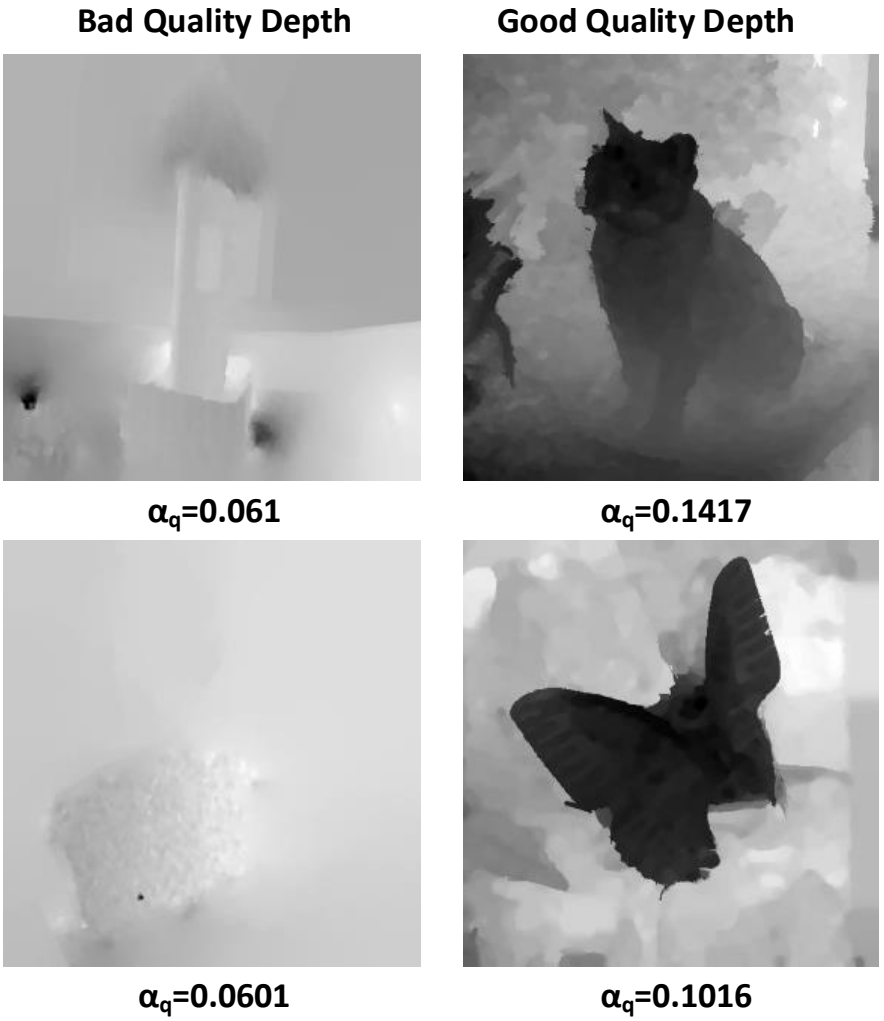


FIGURE 4.4: Some depth image examples and their quality score α_q computed by proposed DQAR.

TABLE 4.1: Proposed DQAR prediction performance comparison by PLCC, SRCC and RMSE evaluation metrics. The top two results are highlighted in 'Red' and 'Blue', respectively.

Model	Pearson linear correlation coefficient (PLCC)	Spearman rank order correlation coefficient (SRCC)	Root mean squared error (RMSE)
BIQA [105]	0.8094	0.6176	0.9307
SEP [106]	0.8744	0.7983	0.7722
BPR [107]	0.5938	0.5539	1.2637
DSS [108]	0.9263	0.835	0.5881
Proposed (DQAR)	0.996	0.9908	0.129

4.1.2 Salient Object Detection Module

In this section, the detail architecture of salient object detection module will be discussed. The proposed salient object detection framework works for incomplete RGB-D modality saliency learning. First of all the incomplete RGB-D problem formulation will be presented. Afterwords, the implementation details of all parts of proposed model will be discussed.

4.1.2.1 Incomplete RGB-D Problem Formulation

Given two modalities, $X \in \mathbb{R}^{C\times H\times W}$ and $Y \in \mathbb{R}^{C\times H\times W}$ with labels $Z \in \mathbb{R}^{H\times W}$ represents RGB modality and depth modality with saliency maps respectively. The proposed incomplete RGB-D modality learning framework consists of two types of training data, complete pair of RGB and depth, RGB present depth missing, denoted as $T_1 = \{(X^{(i)}, Y^{(i)}, Z^{(i)})\}_{i=1}^{n_1}$ and $T_2 = \{(X^{(i)}, Z^{(i)})\}_{i=1}^{n_2}$, respectively. Total number of data samples can be obtained using $T_c = \{(T_1+T_2)\}_{i=1}^{n_c}$ expression. Here $n_c = n_1 + n_2$ represents size of data T_c. The missing depth samples are represented by zero vector and then training data is fed to backbone network. We have used Conformer [93] as a backbone network. Conformer is a parallel hybrid two stream network . One is CNN stream and other is Transformer stream. Conformer network has 12 repeated convolution and transformer blocks and they are categorized in 5 stages. The dimension of CNN branch is $C \times H \times W$, where

C, H and W represents channel, height and width of feature map, respectively. From c1 to c5 stage number of channels increases while resolution decreases. The dimension of Transformer branch is $(K+1)\times E$, where $K, 1$ and E denotes number of image patches, class token and dimensionality of embeddings, respectively [93]. The framework is shown in Fig. 4.5 where Conformer [93] network is used as backbone for feature extraction. Mostly, for multi-modalities each modality is fed to separate backbone network. However, in proposed model single Conformer network is used by leveraging CNN stream of Conformer to extract color and textural details from RGB images and feeding depth image to Transformer stream for efficient geometrical cues. The use of single Conformer network reduces the number of parameters. Moreover, the intrinsic nature of RGB modality gives edges, textures and color details, therefore, efficiency of CNN framework for local details extraction can be well exploited. While depth modality provides structural cues, therefore, we opt Transformer to represent global contextual details from depth images. The hierarchical features from the CNN stream f_r and Transformer stream f_d are then leveraged in a side-output fashion i, denoted as F_r^i and F_d^i with parameters φ_r and φ_d respectively. The extracted feature map pair ($F_r^i \in \mathbb{R}^{C\times H\times W}$, $F_d^i \in \mathbb{R}^{(K+1)\times E}$) for each RGB and depth pair (X, Y) are defined by $F_r^i = f_r(X;\varphi_r)$ and $F_d^i = f_d(Y;\varphi_d)$. For dataset T_2, with missing depth, raw depth data Y is initialized by zero vector. F_r^i side-output has increasing channel and decreasing resolution. Instead of using all layers, typically only the last layer of each stage of the backbone network is selected for multi-level fusion. This ensures a good balance between computational efficiency and the extraction of semantically meaningful features. However proposed model utilizes $(i = 2, 3, 4)$ for shallow feature processing in SCLR block and the resultant feature maps is represented by $SCorr^i$. For deep features processing in DCLR block utilize feature maps at layers$(i = 8, 11)$ and generate output $DCorr^i$ as shown in Fig. 4.5. Furthermore features at layer $(i = 12)$ are used for supervised coarse saliency map generation.

The number of blocks in stage c2 of Conformer is three, which have been represented as side-output features (F_r^2, F_d^2), (F_r^3, F_d^3), and (F_r^4, F_d^4). These features are classified as low-level features, comprising modality-specific details. All the three

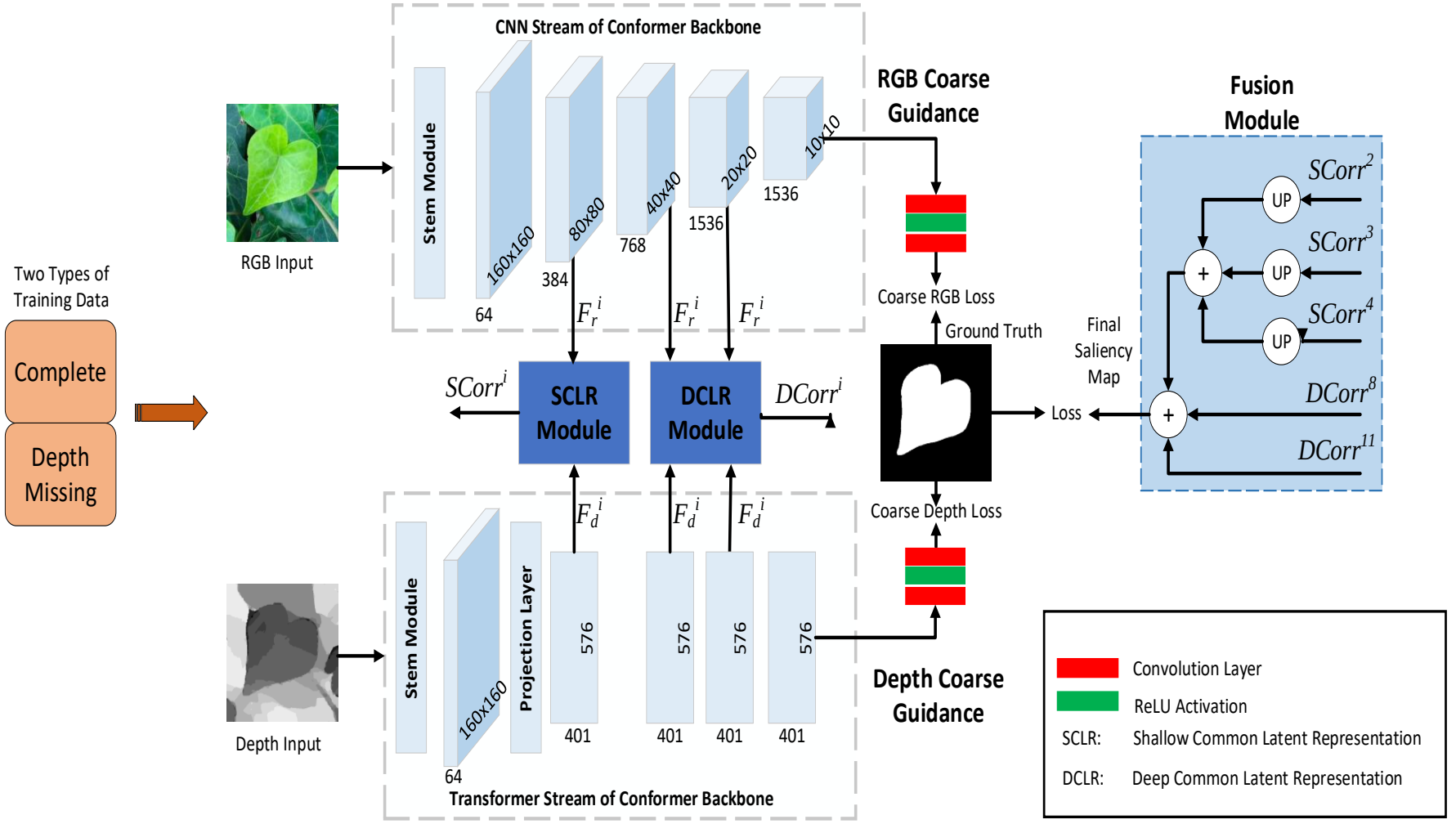


FIGURE 4.5: Proposed framework.

blocks of c2 stage are used in SCLR block for two reasons.

1. The shallow layers carry rich textural and edge details. Specifically, for cluttered background, information from all blocks of c2 stage is beneficial.

2. For missing depth, low-level depth features gradually learn from RGB modality. To ensure that no important features are missed, all blocks from c2 have been selected. Later, the hierarchical feature maps of all 12 blocks will be illustrated to visualize (F_r^2, F_d^2), (F_r^3, F_d^3), and (F_r^4, F_d^4) for complete and incomplete RGB-D pairs in ablation studies.

For DCLR block, last layer of mid and high stage of Conformer network captures affluent global context while the selection of last layer for supervised coarse saliency map generation filters out distractions and highlight only the most salient objects in the scene. Transformer hierarchical manifestations with F_d^i side-output has fixed dimension. Therefore, a transformation$F_{dt}^i = \tau(F_d^i)$ is applied to align the F_r^i and F_d^i side-outputs. After transformation depth stream manifestations are presented as $F_{dt}^i \in \mathbb{R}^{E\times\sqrt{K}\times\sqrt{K}}$. From previous studies [1, 39], it is observed shallow and deep features characteristics have disparities. Hence, proposed model process shallow and deep features separately in Shallow Common Latent Representation(SCLR) block and Deep Common Latent Representation(DCLR), respectively. There are 12 levels of Conformer network with 5 stages (details provided in Table 3.1 of Chapter 3). Proposed model selects $i = 2, 3, 4$ as shallow feature, $i = 8, 11$ as deep features and last layer $i = 12$ for coarse guidance. In SCLR block, RGB modality X is fed to encoder p_r and depth modality Y is input to encoder p_d. The encoders conceal the modality-specific representation, given by $p_r(F_r^i; \omega_r^i)$ and $p_d(F_{dt}^i; \omega_d^i)$. The encoder p_r is parameterized by ω_r^i, which represents learned parameters that are optimized during the training process to minimize the loss function. While the encoder p_d is parameterized by ω_d^i. The framework of p_r and p_d focus the intrinsic properties of RGB and depth. The local details such as edges and textures etc. are aimed to be concealed by p_r while p_d conceals geometric cues. Architectural details of p_r and p_d will be discussed later in this chapter. To correlate the cross-modalities, firstly modality-specific features manifestations obtained from

encoders, given by $p_r(F_r^i;\omega_r^i)$and $p_d(F_{dt}^i;\omega_d^i)$, are mapped to independent variables $\theta_{r1}^i, \theta_{r2}^i, \theta_{d1}^i, \theta_{d2}^i$. Later on, common latent representations from two modalities m_1 and m_2 is acquired by Eq. 4.1. This equation provides a softmax normalized correlation representation.

$$\mathrm{NCorr}_{m_1}^i(F_{m_1}^i;\omega_{m_1}^i) = \sigma(\theta_{m_1 1}^i \odot p_{m_2}(F_{m_2}^i;\omega_{m_2}^i) + \theta_{m_1 2}^i), \tag{4.1}$$

where σ is softmax function. In Deep Common Latent Representation Block, encoder q_r and q_d are applied for RGB modality X and depth modality Y , respectively. These encoders exhibit task-specific properties. Therefore, the output of DCLR encoders captures global contextual characteristics and are presented by $q_r(F_r^i;\omega_r^i)$and $q_d(F_{dt}^i;\omega_d^i)$. Further mapping to θ parameters and correlation representation is similar to SCLR.

4.1.2.2 Shallow Common Latent Representation Block

The architectural diagram of Shallow Common Latent Representation(SCLR) Block is shown in Fig. 4.6. The three modules of SCLR are implemented as follows:

Conceal:

The innate quality of shallow features of any backbone network is incorporation of modality-specific features. Subsequently, two distinct encoders are used to conceal RGB and depth modality such that intrinsic properties of each modality can be preserved in all conceivable manners. The focus of concealing shallow RGB (F_r^i) and depth features (F_{dt}^i) from backbone Conformer network through encoders p_r and p_d is to preserve edges of salient object. For this purpose, operation-wise attention learning similar to [53] is applied. Although, the main focus of shallow features processing is to pop-out boundaries of salient object, but, without having a coarse global contextual information the real saliency cannot be efficiently captured. To overcome this challenge, five parallel operations are performed on (F_r^i) and (F_{dt}^i). The five parallel operations in the CNN stream consists of three

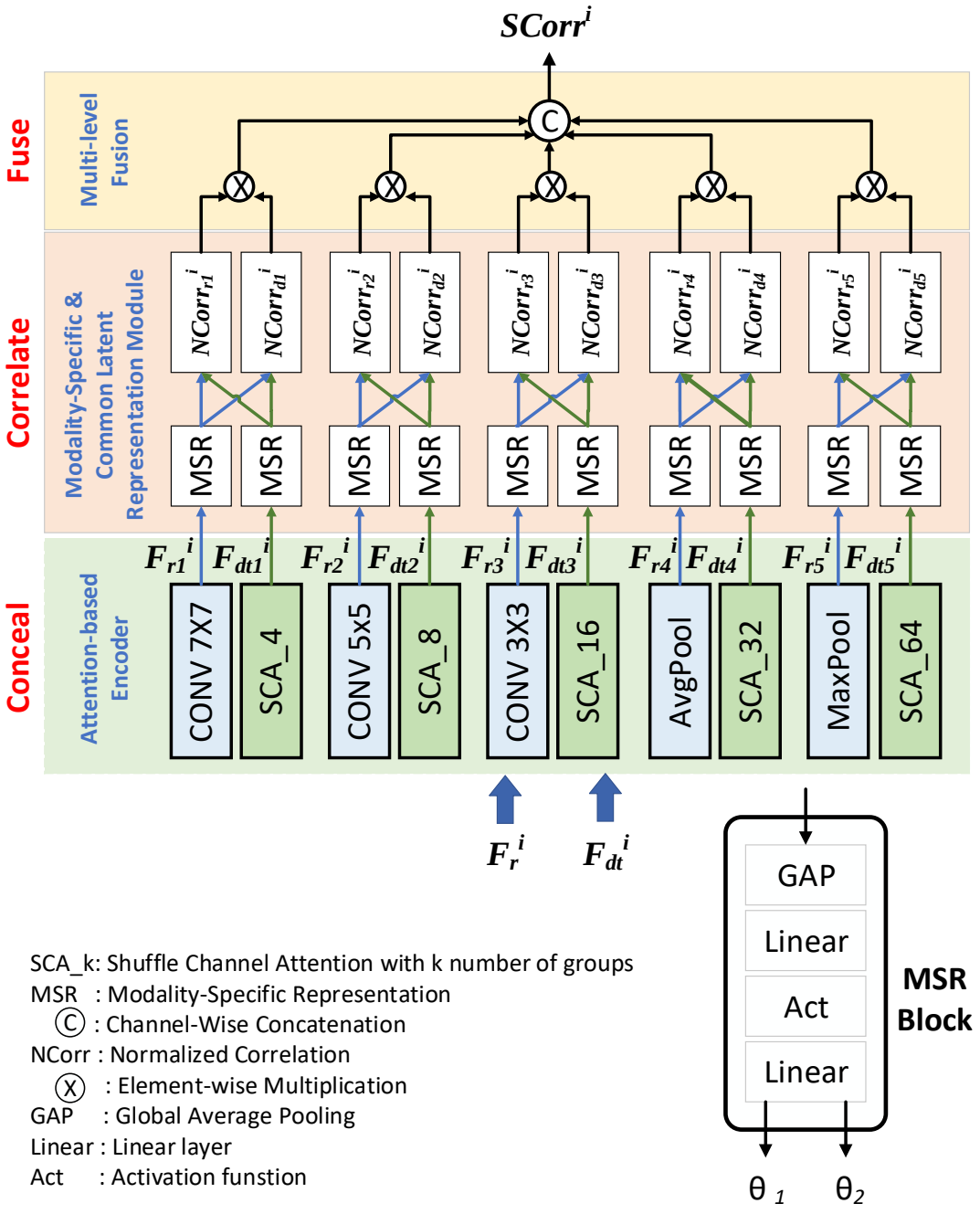


FIGURE 4.6: Shallow Common Latent Representation Block.

convolutional operation with different dilation rate and two pooling operations, given by Eq. 4.2.

$$\begin{aligned} F^i_{r1}, F^i_{r2}, F^i_{r3} &= C(F^i_r; \omega^i_{7\times7}), C(F^i_r; \omega^i_{5\times5}), C(F^i_r; \omega^i_{3\times3}) \\ F^i_{r4}, F^i_{r5} &= Pool_{avg}(F^i_r; \omega^i_{4\times4}), Pool_{max}(F^i_r; \omega^i_{4\times4}), \end{aligned} \tag{4.2}$$

here, $C(*;\omega^i_{\mathrm{k}\times\mathrm{k}})$ represents convolution operation with parameter $\omega^i_{\mathrm{k}\times\mathrm{k}}$ i.e. $(\mathrm{k}\times\mathrm{k})$ kernel size.

Similar to p_r, p_d encoder block also comprises of five parallel operations. However, the intrinsic characteristics of depth modality is to showcase structural details, therefore, shuffle channel attention with different number of groups are applied to depth modality. This scheme effectively captures depth with varying quality, from good to average or some times missing depth. Eq. 4.3 gives mathematical representation.

$$\begin{aligned} F^i_{dt1}, F^i_{dt2}, F^i_{dt3} &= SCA(F^i_{dt};\mathbf{g}^i_4), SCA(F^i_{dt};\mathbf{g}^i_8), SCA(F^i_{dt};\mathbf{g}^i_{16}) \\ F^i_{dt4}, F^i_{dt5} &= SCA(F^i_{dt};\mathbf{g}^i_{32}), SCA(F^i_{dt};\mathbf{g}^i_{64}), \end{aligned} \tag{4.3}$$

where, $SCA(*;\mathbf{g}^i_{\mathrm{o}})$ represents shuffle channel attention with group $\mathbf{g}$ of size o. From Fig. 4.7, it is evident the encoders preserve the modality-specific information to the maximum extent.

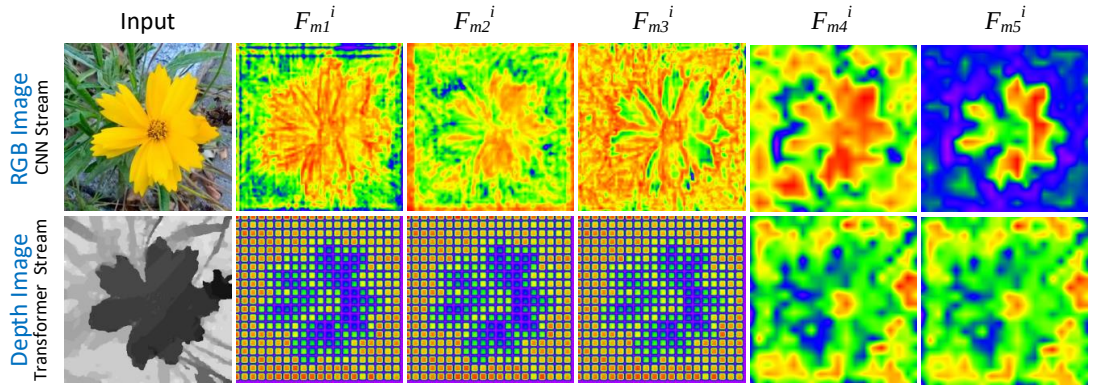


FIGURE 4.7: Five parallel RGB and depth concealed features preserves modality-specific representation.

Correlate:

The distinct traits of concealed RGB and depth features are forwarded to modalityspecific representation (MSR) block. Here, they are mapped to a set of independent parameters, denoted as$\theta^i_{r1}, \theta^i_{r2}, \theta^i_{d1}, \theta^i_{d2}$. In MSR, the first step is global

average pooling operation which can retain informative spatial cues and channel-wise statistics. The θ parameters are obtained through the two fully connected layers with ReLU activation. The boundaries activated feature representation are converted into two distinct parameters, given in Eq. 4.4.

$$\theta^i_{m1}, \theta^i_{m2} = (ReLU(GAP(F^i_m) \cdot W_{l1} + b_{l1})) \cdot W_{l2} + b_{l2}, \tag{4.4}$$

where, m represents modality, GAP represents global average pooling, W is weight and b is bias for linear layers. Afterwords, normalized correlation of RGB and depth is obtained using Eq. 4.1.

Fuse:

From previous "correlate" block, two representations are obtained i.e. RGB-to-depth correlation and depth-to-RGB correlation. The common latent representation of these two are highlighted using multiplication operation. Five parallel layers are concatenated in fusion module. Eq. 4.5 presents the fusion module.

$$\begin{aligned} SCorr^i = & concat((NCorr^i_{r1} \otimes NCorr^i_{d1}), (NCorr^i_{r2} \otimes NCorr^i_{d2}), \\ & (NCorr^i_{r3} \otimes NCorr^i_{d3}), (NCorr^i_{r4} \otimes NCorr^i_{d4}), \\ & (NCorr^i_{r5} \otimes NCorr^i_{d5})). \end{aligned} \tag{4.5}$$

4.1.2.3 Deep Common Latent Representation Block

The architectural diagram of Deep Common Latent Representation (DCLR) block is given in Fig. 4.8. The description of each module of DCLR is given below:

Conceal:

As deep features of backbone network encapsulates task-oriented properties, therefore, salient object can be localized using these features. The architectures of q_r and q_d are different from p_r and p_d. As the role of deep features is to represent saliency cues, therefore same encoders are used to conceal deep RGB (F^i_r) and depth (F^i_{dt}) features. $q_r(F^i_r; \omega^i_r)$and $q_d(F^i_{dt}; \omega^i_d)$ first utilize spatial attention to

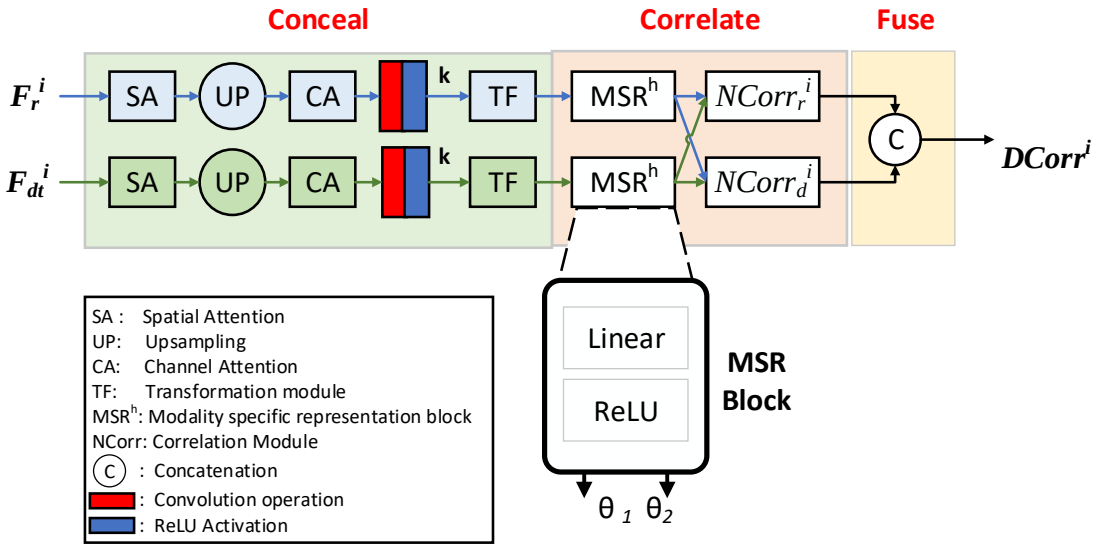


FIGURE 4.8: Deep Common Latent Representation Block.

focus more on foreground and suppress background necessary for salient object localization. After that channel attention finds the inter-channel dependencies. To reduce the model complexity, convolution followed by ReLU non-linearity is applied and k channels are obtained. As there can be multiple salient objects with varying size and shape, inception module [65] is inserted to capture saliency at granularities. The conceal module is presented in Eq.4.6.

$$q_m = TF(ReLU(Conv(CA(UP(SA(F_m^i; \omega_m^i)))))), \tag{4.6}$$

where, m,SA,UP,CA and TF represents modality, spatial attention,upsampling,channel attention and transformation using inception module, respectively. The resultant feature map conceals object localization which can be depicted in Fig. 4.9. Irrespective of modality, deep features are task-oriented and proposed conceal module effectively captures salient object in both modalities.

Correlate:

Similar to SCLR, DCLR also consists of modality-specific representation block which tends to map concealed features to set of independent parameters given in

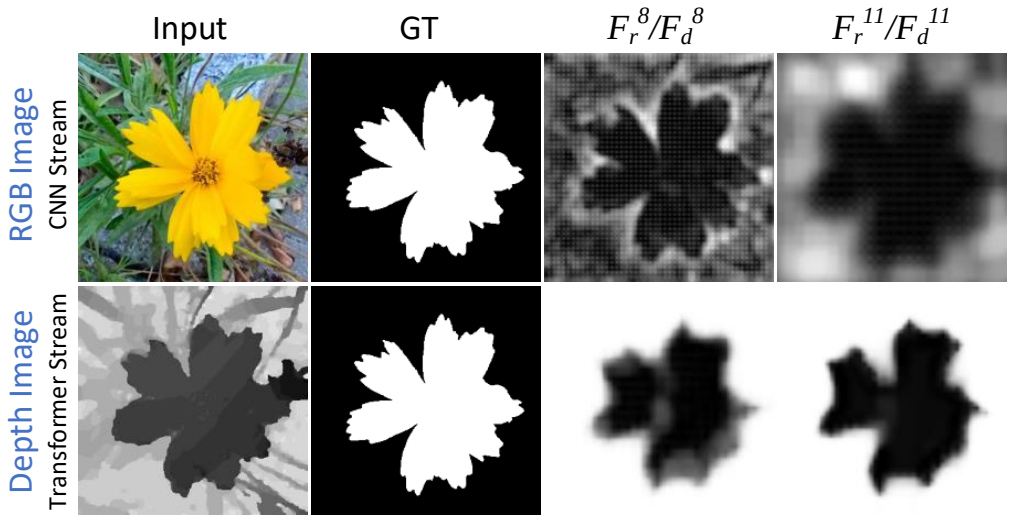


FIGURE 4.9: Deep features concealing salient object.

Eq. 4.7. Correlation expressions are obtained using Eq. 4.1.

$$\theta^i_{m1}, \theta^i_{m2} = ReLU(F^i_m \cdot W_g + b_g). \tag{4.7}$$

Where, θ^i_{m1} and θ^i_{m2} are two parameters for $m1$ modality and $m2$ modality respectively at i^{th} layer.

Fuse:

Two correlation expressions representing RGB-to-Depth and Depth-to-RGB correlation are concatenated in fusion module of DCLR. The equation representing fusion in DCLR is given in Eq. 4.8.

$$DCorr^i = concat(NCorr^i_r, NCorr^i_d). \tag{4.8}$$

4.1.2.4 Coarse Guidance

To further emphasize the positing of salient object, last layer of Conformer backbone guides the coarse localization. To achieve this, ground truth saliency map is used to supervise the prediction at last hierarchy. The loss function used in coarse guidance module is famous binary cross-entropy loss. The last layer is modified

by applying convolution and ReLU activation given in Eq. 4.9.

$$S_{m_c} = C(ReLU(C(F_m^i; \omega_{m_c}^i))). \tag{4.9}$$

Where, C refers to convolution operation with kernel size (1x1), m refers to modality and S is saliency map. For RGB modality m is represented by r and for depth modality m is represented by d. Therefore, S_{r_c} and S_{d_c} are coarse saliency maps of RGB and depth modalities, respectively.

4.1.2.5 Fusion

A simple fusion module is fused that just combines the outputs of correlated features $SCorr^i$ and $DCorr^i$ acquired from SCLR and DCLR blocks, respectively. Before addition operation on these features, they are aligned to same channel and spatial resolution. The final saliency map S_f is obtained using Eq. 4.10.

$$S_f = UP(SCorr^2) + UP(SCorr^3) + UP(SCorr^4) + DCorr^8 + DCorr^{11}, \tag{4.10}$$

here, upsampling operation UP is performed using transposed convolution (stride:4). Correlated output from SCLR module is obtained at $i = 2, 3$, and 4 backbone levels while in DCLR, $i = 8$ and 11 levels are used.

4.1.2.6 Loss Function

The total loss in the proposed model comprises of two components: the coarse guidance loss for RGB and depth modality, denoted as $\mathcal{L}(S_{r_c})$ and $\mathcal{L}(S_{d_c})$, respectively and the final loss, denoted as $\mathcal{L}(S_f)$ and is defined in Eq. 4.11.

$$\mathcal{L}_t = \mathcal{L}(S_f) + \mathcal{L}(S_{r_c}) + \mathcal{L}(S_{d_c}), \tag{4.11}$$

where, $\mathcal{L}$ represents widely adopted binary cross-entropy loss.

4.2 Detail Evaluation of INC-CorrNet

INC-CorrNet is multi-modality salient object detection model two types of data (i) both RGB and depth available, (ii) RGB available but depth missing. Extensive validation of proposed INC-CorrNet representing incomplete RGB-D saliency detection framework is discussed below.

4.2.1 Datasets and Evaluation Metrics

RGB-D SOD datasets: Following the recent SOTA models [1, 25, 39, 79] same 1500 RGB and depth pairs from NJU2K [45] and 700 from NLPR [22] dataset is used. Before feeding depth map to SOD model, their quality is checked and low quality depth are discarded. Therefore, model utilize all RGB data but only acceptable quality depth data. Remaining samples of NJU2K and NLPR are used for testing. Furthermore, 4 benchmark datasets including LFSD [46], SIP [33], STERE [47] and RGBD135 [21] for salient object detection are also used to test the model performance.

RGB and RGB-D SOD datasets: The proposed model is also trained with DUTS [41] RGB training samples along with RGB-D training datasets to check the robustness of model with severely missing modality.

Results Evaluation Metrics: For quantitative results evaluation, precision–recall [48], Structural measure (Smeasure) [49], maximum F-measure (Fmax) [50], maximum E-measure (Emax) [51], and Mean Absolute Error (MAE) [52] are adopted. For qualitative analysis, various scenarios reported by SOTA models [1, 33, 68, 71, 73] are used.

4.2.2 Experiment Settings

This proposed model INC-CorrNet model is implemented on workstation with specifications presented in Table 4.2. The backbone configuration and other parameters are stated in Table 4.3. The input consists of an RGB image and its

corresponding depth map, both resized to a resolution of 320x320 pixels for complete RGB-D pair. For missing depth either due to noise or data scarcity, depth images are initialized by zero vector. Since the depth input is a single channel, depth is replicated to three channels to match the RGB input. The depth values are normalized to the range [0-255]. During the training phase, random crop and normalization data augmentation techniques are applied. The backbone network's weights are initialized with those of Conformer-B [93], pretrained on ImageNet. Other coefficients are initialized using the default settings in PyTorch. A learning rate of 5e-5 with the Adam optimizer is chosen empirically to optimize model performance. Stochastic Gradient Descent (SGD) is opted as the learning approach. The model undergoes a complete training cycle, totaling 115 epochs.

TABLE 4.2: Specification of workstation used for training of INC-CorrNet.

OS	Linux:Ubuntu 18.04.6 LTS
GPU	Single NVIDIA GeForce RTX 3060 Ti
GPU	Memory 8 GB
CPU	Intel® Core™i7-10700K
RAM	16 GB
System Type	64 bit
Cuda Version	11.8
Interpreter	Python 3.12

TABLE 4.3: Conformer-B Configuration.

Patch size	16
Channel ratio	6
Embedding dimensions	576
Depth	12
Number of Heads	9
MLP ratio	4

4.2.3 Quantitative Results

In this section, quantitative results analysis will be presented against five evaluation metrics stated above. And a thorough comparison with 14 state-of-the-art (SOTA) models including, D3Net [33], CalibD [85] , SwinNet [17] , CDNet [84] , SiaTrans [27] , CIR-Net [72] , DIGR-Net [100] , DTMINet [109] , SERANet [110],

MIRV [111] , DCMNet [87] , FCFNet [86] , HiDAnet [74] , M2RNet [75]. The saliency maps for these models have been provided by their respective authors. Quantitative comparison of proposed INC-CorrNet with recent SOTA models is presented in Table 4.4. Notable performance gain of INC-CorrNet is observed, specifically, significant decrease in mean absolute error is marked by 3.8%, 8.3%, 14.4%, 5.7%, 22.5%, 0.8% on NJU2K, NLPR, LFSD, SIP, STERE and RGBD135 datasets, respectively, as compared to SOTA models. Also average increase of 1.1% (S-measure), 1.32% (max F-measure) and 1.3% (max E-measure) across five datasets has been observed. It can be inferred that proposed INC-CorrNet has achieved top performance in almost all metrics, with only two instances where it ranks as the second best. The Precision-Recall Curve presented in Fig. 4.10 shows that INC-CorrNet have high recall because of shorter curves across all datasets.

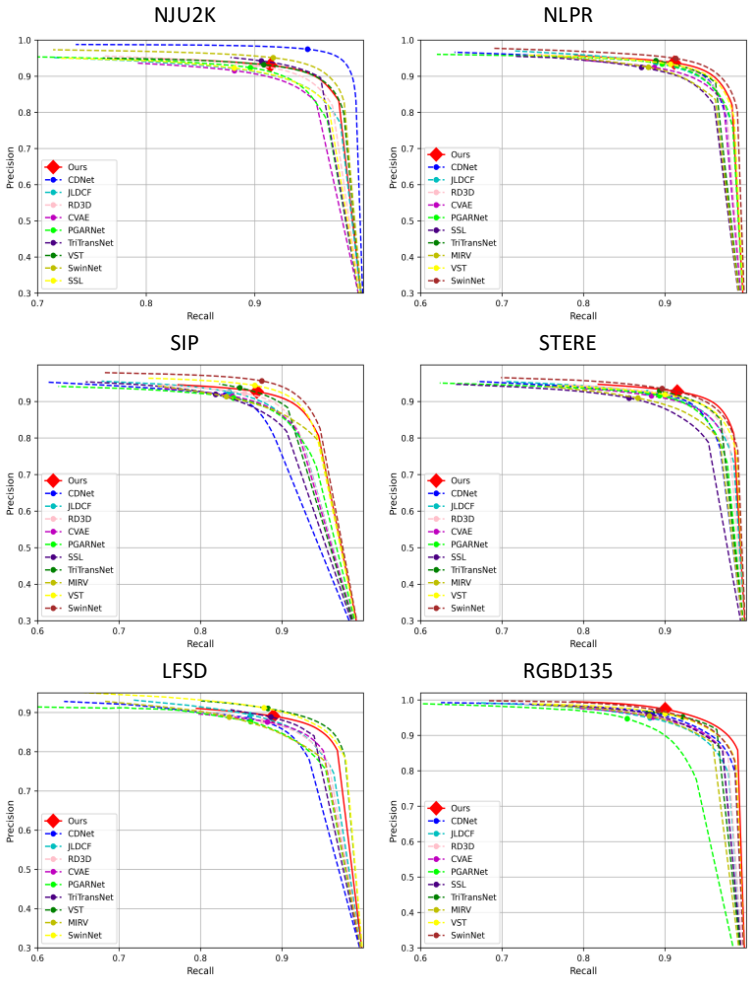


FIGURE 4.10: Precision-recall curves of proposed model and SOTA models for six benchmark datasets.

TABLE 4.4: Quantitative comparison of proposed INC-CorrNet model with 14 SOTA RGB-D SOD models on 6 benchmark datasets. Best results are represented by 'Red' color and second best results are represented by 'Blue' color. '-' indicates result is not available.

Models / Mertic	D3Net (2020) [33]	CalibD (2021) [85]	SwinNet (2021) [17]	CDNet (2021) [84]	SiaTrans (2022) [27]	CIR-Net (2022) [72]	DIGR-Net (2022) [100]	DTMI-Net (2023) [109]	SERA-Net (2023) [110]	MIRV (2023) [111]	DCM-Net (2023) [87]	FCF-Net (2023) [86]	HiDAnet (2023) [74]	M2RNet (2023) [75]	INC-CorrNet (2023)
Quantitative Evaluation on NJU2K [45] Dataset.															
$S_\alpha \uparrow$	0.9	0.922	0.935	0.918	0.923	0.925	0.933	0.929	0.916	0.89	–	0.918	0.926	0.91	0.9334
$F_\beta^{max} \uparrow$	0.9	0.884	0.922	0.919	0.921	0.9277	0.939	0.933	0.926	0.88	0.899	0.923	0.939	0.922	0.939
$E_\zeta^{max} \uparrow$	0.95	0.897	0.934	0.95	0.956	-	0.966	-	0.942	0.929	0.92	0.953	0.954	0.904	0.959
$MAE \downarrow$	0.041	0.038	0.027	0.036	0.035	0.035	0.026	0.029	0.037	0.046	0.036	0.034	0.029	0.049	0.025
Quantitative Evaluation on NLPR [22] Dataset.															
$S_\alpha \uparrow$	0.912	0.956	0.941	0.931	0.929	0.9334	0.931	0.939	0.932	0.914	-	0.924	0.93	0.918	0.9664
$F_\beta^{max} \uparrow$	0.897	0.892	0.908	0.92	0.918	0.9241	0.923	0.929	0.931	0.895	0.883	0.911	0.929	0.921	0.953
$E_\zeta^{max} \uparrow$	0.853	0.893	0.967	0.964	0.964	-	0.964	-	0.961	0.953	0.954	0.96	0.961	0.941	0.9841
$MAE \downarrow$	0.03	0.023	0.018	0.025	0.024	0.0227	0.021	0.019	0.022	0.025	0.024	0.024	0.021	0.033	0.0165

.....Continued on next page.....

TABLE 4.4: Quantitative comparison of proposed INC-CorrNet model with 14 SOTA RGB-D SOD models on 6 benchmark datasets. Best results are represented by 'Red' color and second best results are represented by 'Blue' color. '-' indicates result is not available.

Models / Mertic	D3Net (2020) [33]	CalibD (2021) [85]	SwinNet (2021) [17]	CDNet (2021) [84]	SiaTrans (2022) [27]	CIR-Net (2022) [72]	DIGR-Net (2022) [100]	DTMI-Net (2023) [109]	SERA-Net (2023) [110]	MIRV (2023) [111]	DCM-Net (2023) [87]	FCF-Net (2023) [86]	HiDAnet (2023) [74]	M2RNet (2023) [75]	INC-CorrNet (2023)
Quantitative Evaluation on LFSD [46] Dataset.															
$S_\alpha \uparrow$	0.825	-	-	0.858	0.871	0.8753	0.869	-	-	0.849	-	0.875	-	0.842	0.898
$F_\beta^{max} \uparrow$	0.81	-	-	0.861	0.876	0.8828	0.869	-	-	0.844	0.867	0.876	-	0.861	0.8968
$E_\zeta^{max} \uparrow$	0.862	-	-	0.896	0.907	-	0.905	-	-	0.889	0.906	0.913	-	0.874	0.9332
$MAE \downarrow$	0.095	-	-	0.073	0.069	0.0677	0.067	-	-	0.072	0.064	0.061	-	0.088	0.0522
Quantitative Evaluation on SIP[33] Dataset.															
$S_\alpha \uparrow$	0.86	0.92	0.911	-	0.899	0.8884	0.882	0.908	0.895	0.876	-	-	0.892	0.882	0.9269
$F_\beta^{max} \uparrow$	0.861	0.85	0.912	-	0.913	0.8959	0.892	0.918	0.919	0.863	0.883	-	0.919	0.902	0.9321
$E_\zeta^{max} \uparrow$	0.909	0.877	0.943	-	0.945	-	0.925	-	0.93	0.924	0.926	-	0.927	0.921	0.9571
$MAE \downarrow$	0.063	0.051	0.035	-	0.041	0.0523	0.051	0.037	0.045	0.049	0.047	-	0.043	0.049	0.033

.....Continued on next page.....

TABLE 4.4: Quantitative comparison of proposed INC-CorrNet model with 14 SOTA RGB-D SOD models on 6 benchmark datasets. Best results are represented by 'Red' color and second best results are represented by 'Blue' color. '-' indicates result is not available.

Models / Mertic	D3Net (2020) [33]	CalibD (2021) [85]	SwinNet (2021) [17]	CDNet (2021) [84]	SiaTrans (2022) [27]	CIR-Net (2022) [72]	DIGR-Net (2022) [100]	DTMI-Net (2023) [109]	SERA-Net (2023) [110]	MIRV (2023) [111]	DCM-Net (2023) [87]	FCF-Net (2023) [86]	HiDAnet (2023) [74]	M2RNet (2023) [75]	INC-CorrNet (2023)
Quantitative Evaluation on STERE [47] Dataset.															
$S_\alpha \uparrow$	0.899	0.931	0.919	0.906	0.914	-	0.914	0.922	0.909	-	-	0.906	0.911	-	0.933
$F_\beta^{max} \uparrow$	0.891	0.88	0.893	0.898	0.907	-	0.916	0.92	0.918	-	-	0.906	0.921	-	0.931
$E_\zeta^{max} \uparrow$	0.938	0.89	0.923	0.942	0.951	-	0.957	-	0.943	-	-	0.947	0.946	-	0.961
$MAE \downarrow$	0.046	0.037	0.033	0.04	0.038	-	0.035	0.031	0.038	-	-	0.038	0.035	-	0.024
Quantitative Evaluation on RGBD135 [21] Dataset.															
$S_\alpha \uparrow$	0.898	-	0.945	0.936	0.936	-	0.936	-	-	0.928	-	0.939	0.946	0.934	0.953
$F_\beta^{max} \uparrow$	0.885	-	0.926	0.928	0.932	-	0.932	-	-	0.921	-	0.939	0.952	0.937	0.9539
$E_\zeta^{max} \uparrow$	0.946	-	0.98	0.969	0.975	-	0.972	-	-	0.965	-	0.98	0.98	0.971	0.9867
$MAE \downarrow$	0.031	-	0.016	0.02	0.02	-	0.019	-	-	0.019	-	0.017	0.013	0.019	0.0129

4.2.4 Qualitative Results

A qualitative results comparison is provided in Fig. 4.11, where some challenging samples are selected based on criteria highlighted in various published studies [1, 25, 33, 39, 68, 75, 86, 98]. One of the complex test case scenario is cluttered background. And if corresponding depth is also of low quality, then it becomes more challenging. In Fig. 4.11 row (a), this complex scenario is presented. It is evident that proposed INC-CorrNet performs best in this challenging scene. Additional depth input source enhances SOD performance for similar appearance in foreground and background. But for low quality depth as in row (b) and (c) ,detection difficulty increases. As shown by predicted saliency maps of INC-CorrNet and other SOTA models, the proposed model excels in both comprehensiveness and clarity when compared to other methods. Several other challenging scenes such as the case where salient and non-salient object are adjacent (row (d)), low contrast RGB (row (e)), low contrast depth(row (f)), good quality RGB and depth (row (g))and multiple small object with complex background (row (h)) are also evaluated. The accuracy of proposed INC-CorrNet is much better than other models.

4.2.5 Analysis of Model Robustness

To validate the robustness of the proposed model with severely missing depth modality INC-CorrNet is tested on DUTS RGB [41] test dataset with only RGB data and results are illustrated in Fig. 4.12. Two complex scenarios, where existing RGB-D SOD models based on augmenting estimated depth with raw depth, fail are complex background and similar background-foreground appearance. Examples (a) and (c) are of complex background while examples (b) and (d) are of similar background-foreground appearance. It can be observed that the proposed INC-CorrNet shows a resilient solution for missing depth and captures the salient objects with accuracy. Therefore, it can be concluded that the proposed model presents a robust solution to incomplete RGB-D saliency detection application.

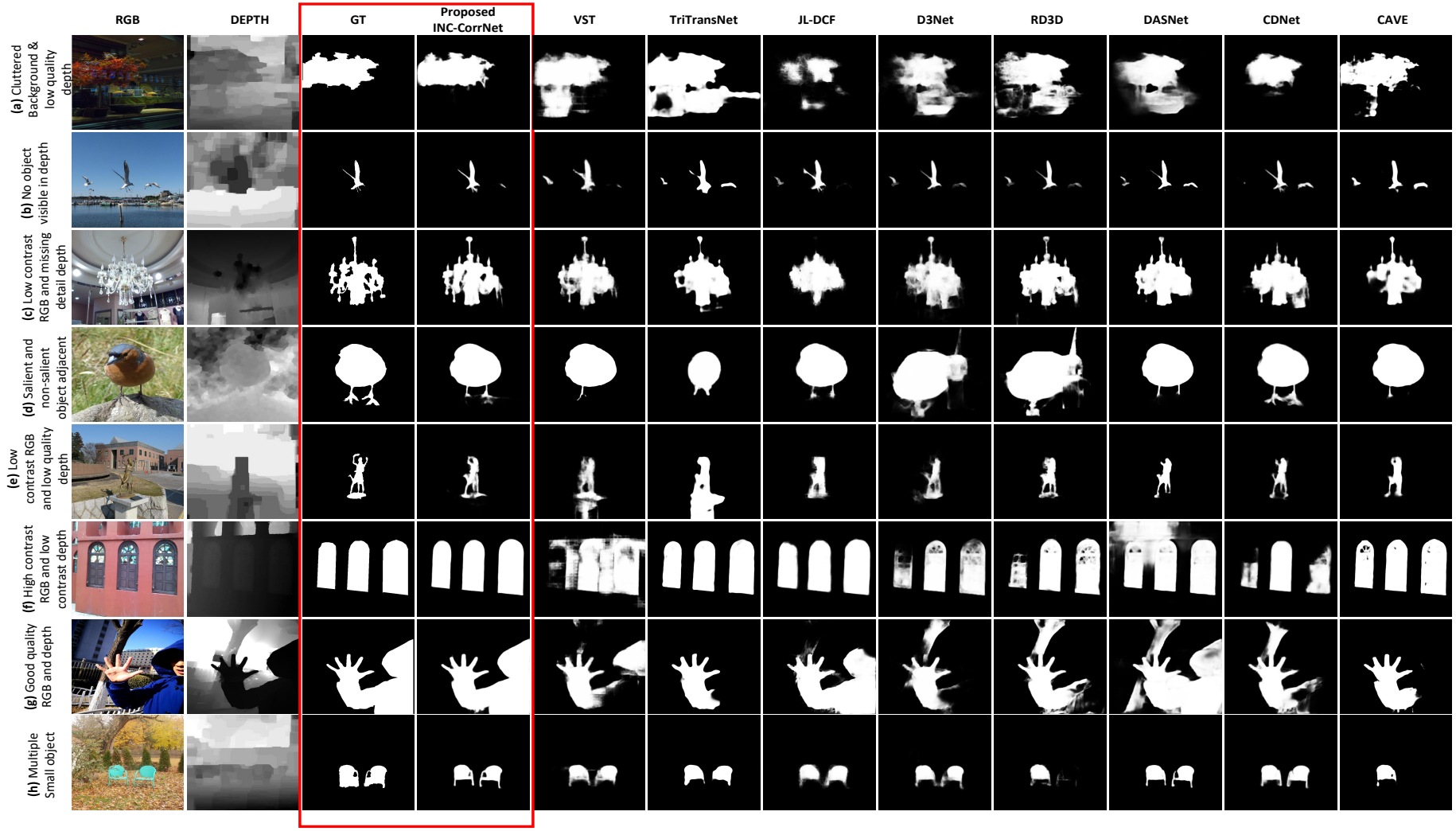


FIGURE 4.11: Visual comparison of proposed model with SOTA models for various challenging scenarios.

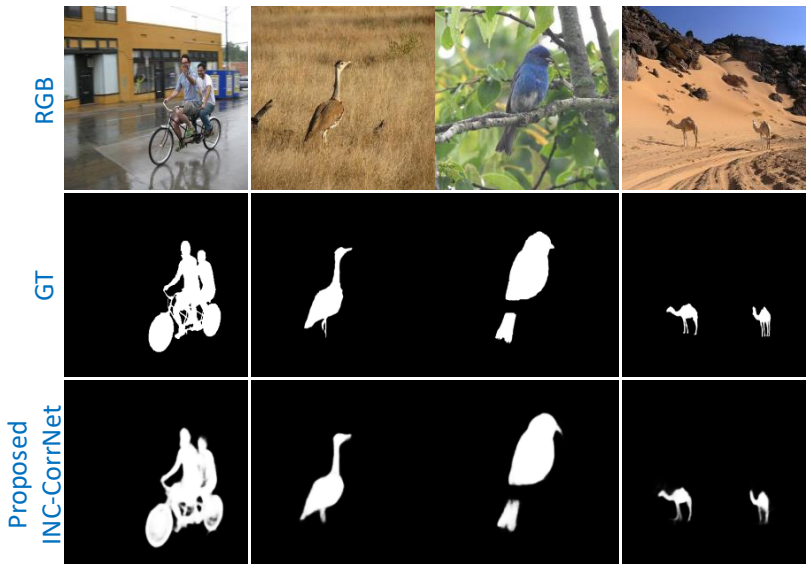


FIGURE 4.12: Visual examples of saliency predictions by proposed model with missing depth.

4.2.6 Trade-off between Accuracy and Efficiency

In Table 4.5, efficiency comparison of proposed model with SOTA model against Frames per second (FPS), number of parameters and model size is presented. All these values are computed on machine specified in Table 4.2. Accuracy gauges a model's ability to precisely identify and pinpoint significant objects, whereas efficiency defines the amount of time and computational resources required to accomplish this task.

To visualize the trade-off between accuracy and efficiency of proposed and SOTA models, averaged max F-Measure value over six benchmark datasets (as provided in various published studies [74, 103]) are used and plotted against Frames per second (FPS) and number of parameters in Fig 4.13 and 4.14, respectively. Trade-off visualization of computational efficiency in terms of frames per second and max F-measure shows a balance for proposed INCCorrNet. Although some SOTA models have high FPS than proposed model, but with highly compromised accuracy. Moreover, most of SOTA models have similar FPS as INC-CorrNet. For the graph between number of parameters versus max-Fmeasure, proposed INC-CorrNet, holds a favourable position almost in the top-left quadrant of the chart,

TABLE 4.5: Efficiency comparison of proposed INC-CorrNet with SOTA models on Frame per Second (FPS), number of parameters, and model size.

Model	**RGB Input**	**Depth Input**	**FPS**	**Param. (M)**	**Model Size (MB)**
CDNet [84]	3x224x224	1x224x224	46	32.9	125.6
VST [76]	3x224x224	3x224x224	25	83.83	320.3
JL-DCF [25]	3x320x320	3x320x320	16.3	143	547.8
MIRV [111]	3x352x352	3x352x352	14.47	114	436.5
SwinNet [17]	3x384x384	3x384x384	18	199.1	785.7
AFNet [112]	3x352x352	1x352x352	19.3	254	971
SSL [99]	3x352x352	3x352x352	26.8	74.1	283.1
HINet [113]	3x352x352	1x352x352	26.13	98.9	377.76
RD3D [97]	2x3x352x352		36.1	46	179.1
SPNet [114]	3x352x352	1x352x352	23	175	669.2
INC-CorrNet	3x320x320	3x320x320	22	105	403

signifying a beneficial balance between accuracy and efficiency. A comparison of max F-Measure, MAE, and Model Size of different models is shown in Fig. 4.15. The size of the circle represents the model size. It's worth noting that superior models are situated in the top left corner, characterized by a larger max F-measure and a smaller MAE. Methods with a smaller size demonstrate inferior performance, highlighting the efficiency and accuracy of the proposed INC-CorrNet approach.

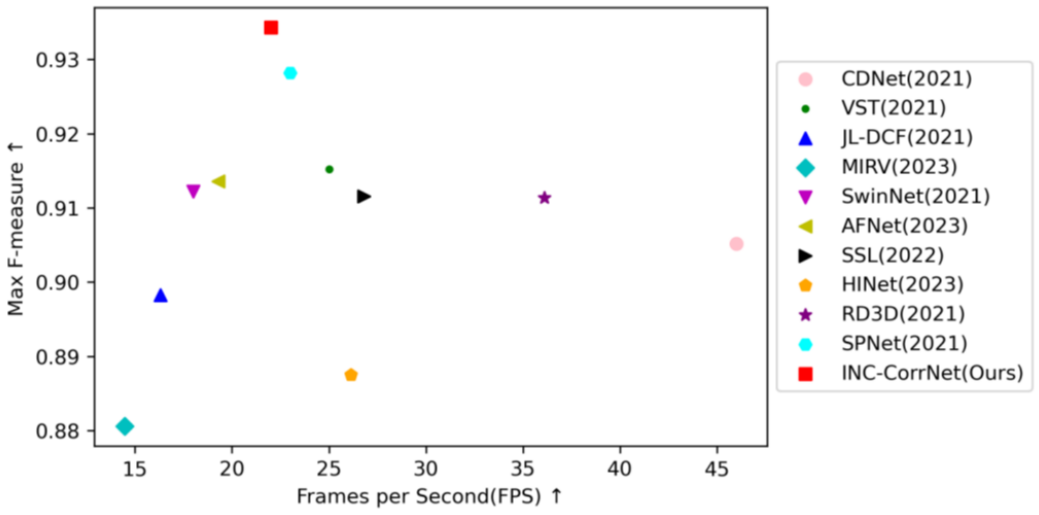


FIGURE 4.13: FPS vs max F-measure.

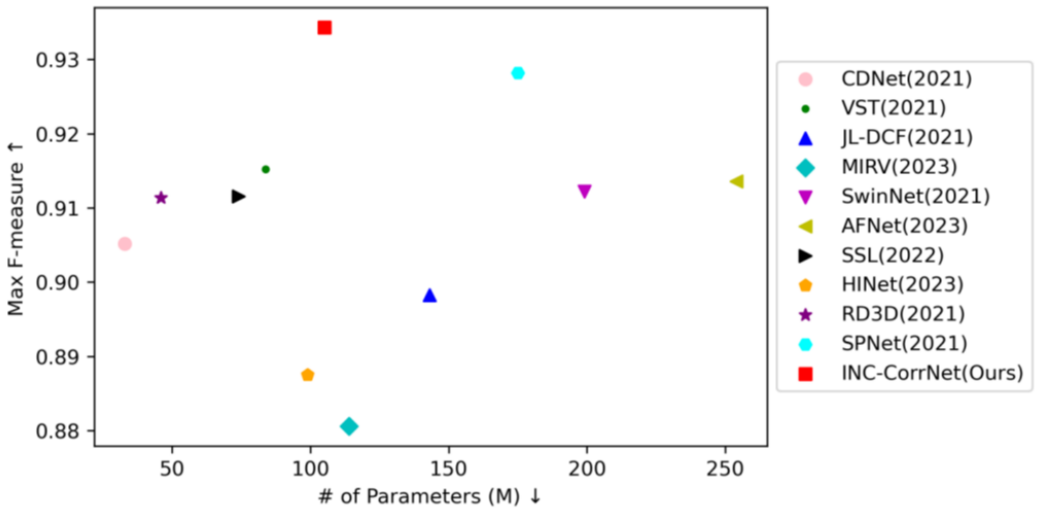


FIGURE 4.14: Number of Parameters vs max F-measure.

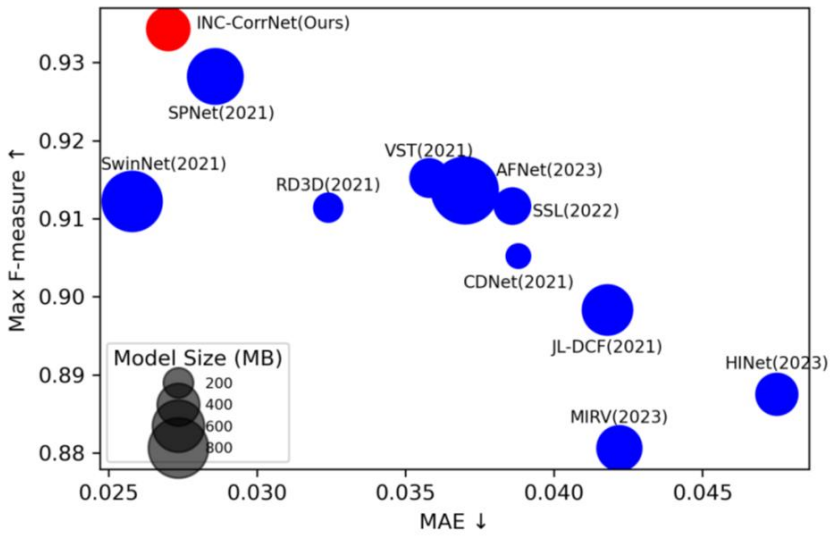


FIGURE 4.15: Average Max F-measure, MAE and Model SIze.

4.2.7 Ablation Studies

In this section, a comprehensive ablation studies will be discussed, to examine the impact each module of INC-CorrNet. The full implementation of INC-CorrNet is

denoted as 'Model A', comprising DQAR + SCLR + DCLR (RGB-D) and trained on NJU2K [45] and NLPR [22] RGB-D training samples.

Model performance with severely missing depth:

To validate the incomplete RGB-D modality for saliency learning with severely missing depth, a 'Model B' is trained. This model comprises of DQAR + SCLR + DCLR (RGB + RGBD) components and trained using NJU2K RGB-D [45], NLPR RGB-D [22] and DUTS RGB [41] training sample. Low quality depth images from NJU2K and NLPR are discarded, thus, model is trained with large number of missing depth. Although 'Model A' and 'Model B' represent full implementation of INC-CorrNet but later have large number of missing depth. The comparative analysis of 'Model A' and 'Model B' is shown in Table 4.6 which depicts that MAE has been decreased by 11.3%, 4.2%, 2.5% on LFSD, SIP and STERE datasets, respectively. However for NLPR and RGBD135 datasets slight increase in MAE value by 3% and 3.8%, respectively is observed. These two datasets contain high quality depth images and low contrast RGB image. Due to the substantial absence of depth modality, the training process is biased towards RGB patterns, posing challenges in effectively training the model for low-contrast images.

Validity of Correlation Representation:

The primary strength of the proposed model lies in leveraging correlation representation, enabling the discovery of latent correlations between RGB and depth modalities to enhance the model's robustness in scenarios involving missing depth. To validate the effectiveness of proposed correlation representation, correlation expression is removed from full implementation of proposed model. Hence 'Model C': DQAR + w/o correlation, represents model with correlation expression removed from INC-CorrNet. The quantitative evaluation and visual comparison between model A and C are presented in Table 4.7 and Fig. 4.16, respectively. It is observed that performance of Model C degrades by 42%, 22%, 9.2%, 42%, 39.5%, 34.8% on NJU2K, NLPR, LFSD, SIP, STERE and RGBD135 datasets, respectively for MAE metric. This not only underscores the effectiveness of the proposed component but also validates the assumption that a robust correlation exists in the latent

TABLE 4.6: Quantitative evaluation of ablation studies regarding performance of INC-CorrNet under severely missing depth. 'Model A': DQAR + SCLR + DCLR (RGB-D) presents the reference model and 'Model B': DQAR + SCLR + DCLR (RGB + RGBD) presents the model with severely missing depth.

Dataset/Models/Metrics		$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$
NJU2K	Model A	0.9334	0.939	0.959	0.025
	Model B	0.9371	0.9419	0.9609	0.025
NLPR	Model A	0.9664	0.953	0.9841	0.0165
	Model B	0.9643	0.9492	0.9806	0.0175
LFSD	Model A	0.898	0.8968	0.9332	0.0522
	Model B	0.9056	0.9048	0.9389	0.0463
SIP	Model A	0.9269	0.9321	0.9571	0.033
	Model B	0.9306	0.9343	0.9583	0.0316
STERE	Model A	0.933	0.931	0.961	0.024
	Model B	0.9365	0.9338	0.9619	0.0234
RGBD135	Model A	0.953	0.9539	0.9867	0.0129
	Model B	0.9505	0.9531	0.9845	0.0134

representation between RGB and depth modalities. The marginal improvement in the performance of 'Model A' on the LFSD dataset can be attributed to the limited number of samples and challenging nature of this dataset, characterized by only 100 low-quality images with intricate backgrounds.

Validity of SCLR:

The correlation of two modalities highly depends on the modality-specific latent representation. The shallow layers of backbone network exhibit modality-specific and task agnostic characteristics as oppose to deep layers. Therefore, in SCLR and DCLR of the proposed model, different encoders are used. To validate this point, modality oriented encoder in SCLR block is replaced with task oriented encoder used in DCLR. Hence, 'Model D': DQAR + DCLR, represents model with same encoder in SCLR and DCLR, i.e. modality oriented encoder in SCLR block is replaced with task oriented encoder used in DCLR. The qualitative and quantitative

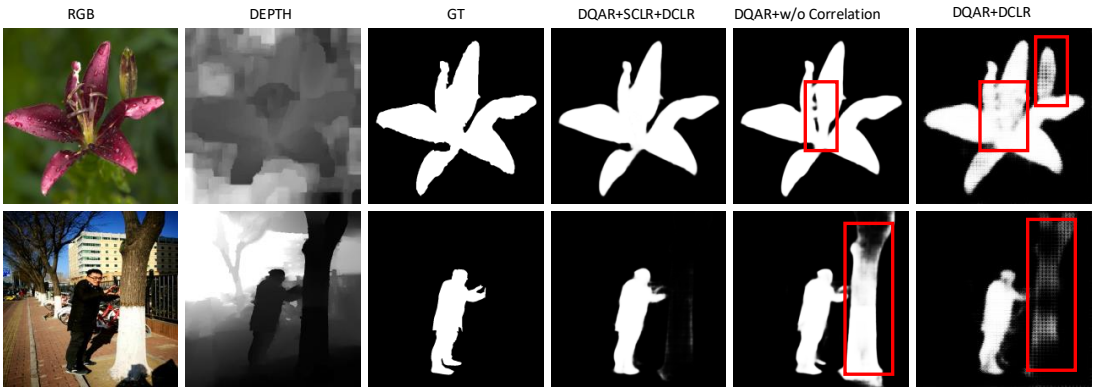


FIGURE 4.16: Visual comparison among full implementation of proposed model ('Model A': DQAR+SCLR+DCLR), proposed model without correlation representation ('Model C': DQAR+w/o Correlation) and proposed model with same encoder for shallow and deep features ('Model D': DQAR+DCLR).

evaluation of 'Model A' and 'Model D' are presented in Table 4.8 and Fig. 4.16, respectively. The introduction of modality oriented encoder in SCLR block ('Model A') has led to enhancement in performance. The MAE metric has been decreased by 47%, 49%, 32%, 32%, 46.7%, 57.7% on NJU2K,NLPR,LFSD,SIP,STERE and RGBD135 datasets, respectively.

Validity of DQAR:

To demonstrate the influence of low quality depth maps on detection accuracy, the DQAR block is removed. Hence, 'Model E': w/o DQAR, presents model without DQAR module. Consequently, the depth images will be directly input into the proposed model to accomplish saliency prediction.The findings in Table 4.9 indicates that the absence of the DQAR block leads to a more substantial performance degradation across all datasets compared to the removal of any other block. Fig. 4.17 provides some visual examples for ablation studies about validity of DQAR. The inaccurate predictions obtained without DQAR realizes the impact of noisy depth maps.

Effectiveness of Conformer backbone network:

Despite the availability of various hybrid backbone networks combining CNN and transformer architectures, opting for Conformer backbone network offers numerous

TABLE 4.7: Quantitative evaluation of ablation studies regarding validity of correlation representation. 'Model A': DQAR + SCLR + DCLR (RGB-D) presents the reference model and 'Model C': DQAR+w/o Correlation presents the proposed model without correlation representation.

Dataset/Models/Metrics		$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$
NJU2K	Model A	0.9334	0.939	0.959	0.025
	Model C	0.9179	0.9198	0.9521	0.0356
NLPR	Model A	0.9664	0.953	0.9841	0.0165
	Model C	0.9363	0.927	0.9663	0.0202
LFSD	Model A	0.898	0.8968	0.9332	0.0522
	Model C	0.8866	0.8892	0.9231	0.057
SIP	Model A	0.9269	0.9321	0.9571	0.033
	Model C	0.8858	0.8898	0.9277	0.0469
STERE	Model A	0.933	0.931	0.961	0.024
	Model C	0.9171	0.9111	0.9523	0.0335
RGBD135	Model A	0.953	0.9539	0.9867	0.0129
	Model C	0.9364	0.935	0.967	0.0174

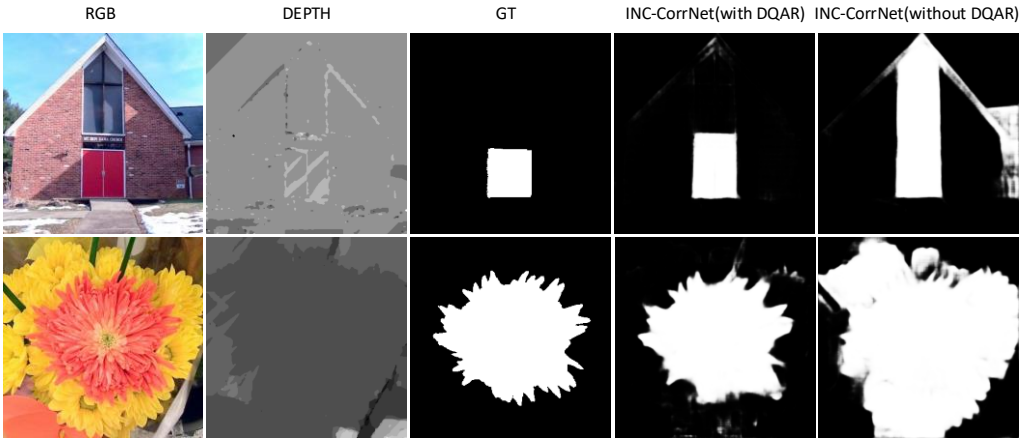


FIGURE 4.17: Visual examples for ablation studies about validity of DQAR.

advantages.

1. High performing RGB-D salient object detection models select multi-level fusion scheme using two stream network.The rationale behind this selection

TABLE 4.8: Quantitative evaluation of ablation studies regarding validity of SCLR. 'Model A': DQAR + SCLR + DCLR (RGB-D) presents the reference model and 'Model D': DQAR + DCLR represents the model with the same encoder in SCLR and DCLR.

Dataset/Models/Metrics		$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$
NJU2K	Model A	0.9334	0.939	0.959	0.025
	Model D	0.9026	0.9002	0.9434	0.0473
NLPR	Model A	0.9664	0.953	0.9841	0.0165
	Model D	0.9103	0.8927	0.9513	0.0325
LFSD	Model A	0.898	0.8968	0.9332	0.0522
	Model D	0.861	0.8494	0.888	0.0766
SIP	Model A	0.9269	0.9321	0.9571	0.033
	Model D	0.8931	0.9004	0.9349	0.0486
STERE	Model A	0.933	0.931	0.961	0.024
	Model D	0.9063	0.8982	0.9431	0.0451
RGBD135	Model A	0.953	0.9539	0.9867	0.0129
	Model D	0.9075	0.8989	0.9532	0.0305

is disparities in cross-modalities. RGB modality gives color and texture details while depth modality provides structural information. Therefore, two stream network can be exploited to obtain modality-specific feature representations. As Conformer already has two parallel streams, therefore, a single Conformer network is utilized by feeding RGB modality to CNN stream and depth modality to Transformer stream. This type of architecture is not available in any hybrid backbone network. Other existing hybrid network follows integration of CNN and Transformer in early, late or sequential pattern. The utilization of any of these backbone architectures necessitates the deployment of two backbones for each modality, thereby contributing to an increase in the overall model size. Hence, number of parameters can be greatly reduced using single Conformer backbone. The parameters and computational complexity using FLOPs of several backbone networks are as shown in Table 4.11.

TABLE 4.9: Quantitative evaluation of ablation studies regarding validity of DQAR. 'Model A': DQAR + SCLR + DCLR (RGB-D) presents the reference model and 'Model E': w/o DQAR, presents the model without DQAR module.

Dataset/Models/Metrics		$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\zeta^{max} \uparrow$	$MAE \downarrow$
NJU2K	Model A	0.9334	0.939	0.959	0.025
	Model E	0.8994	0.9	0.9385	0.0474
NLPR	Model A	0.9664	0.953	0.9841	0.0165
	Model E	0.9222	0.9079	0.954	0.0281
LFSD	Model A	0.898	0.8968	0.9332	0.0522
	Model E	0.8588	0.8568	0.8952	0.0772
SIP	Model A	0.9269	0.9321	0.9571	0.033
	Model E	0.8624	0.8523	0.9075	0.0638
STERE	Model A	0.933	0.931	0.961	0.024
	Model E	0.9051	0.8985	0.9428	0.0418
RGBD135	Model A	0.953	0.9539	0.9867	0.0129
	Model E	0.9394	0.9357	0.9727	0.0178

TABLE 4.10: Effectiveness of Conformer backbone network.

Backbone Network	Architecture	FLOPs (G)	#Params (M)
ResNet101 [30]	CNN	15.6	44.5
ViT-B [66]	Transformer	111	86
LeViT [115]	Hybrid	2353	39
ResT [116]	Hybrid	42.4	87
Conformer-B [93]	Hybrid	46.6	83.3

2. The CNN's adeptness at capturing local details can be effectively utilized by inputting RGB modality into the CNN branch. This approach enables the extraction of abundant information from the RGB input. While the Transformer stream efficiently captures the global context from depth structural information. As illustrated in Fig. 4.18, the CNN stream primarily emphasizes capturing local details, whereas the Transformer stream is focused on capturing global contextual information.

3. In the parallel hybrid Conformer network, the CNN and Transformer streams exchange significant information at the end of each stage through the feature coupling unit (FCU). As a result, in scenarios where the depth modality is missing and depth is initialized as zero vector, CNN imparts crucial insights to the Transformer, effectively bridging the semantic disparity among cross-modal features. Fig. 4.18 depicts two samples: one comprising a complete

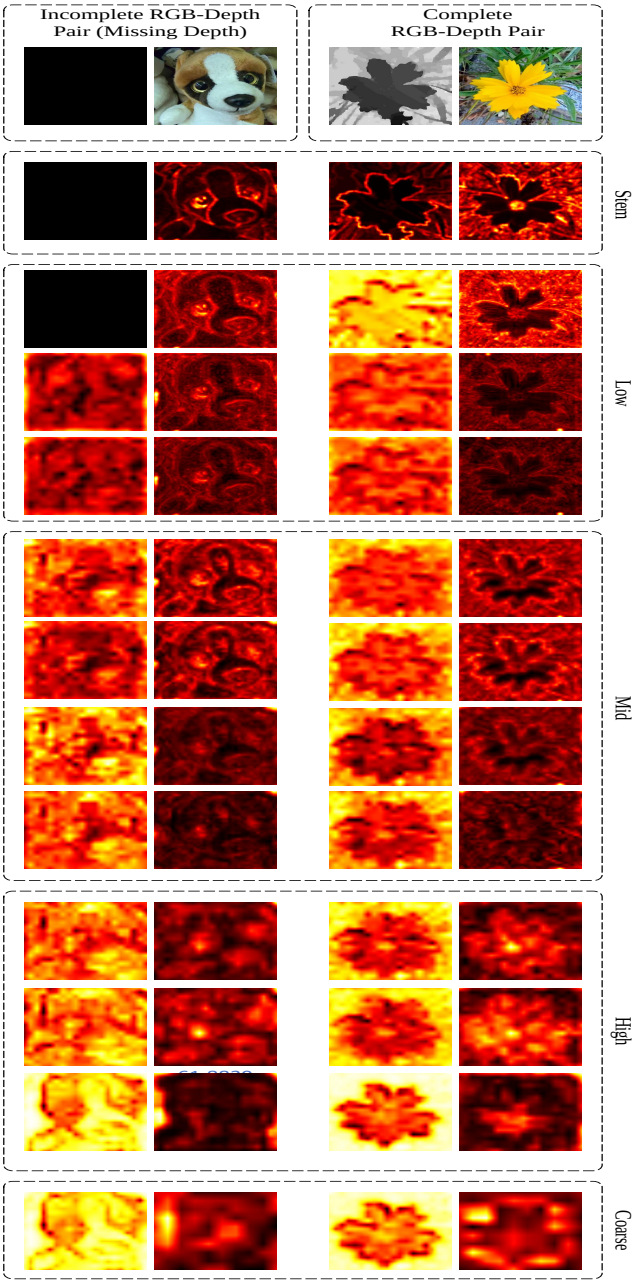


FIGURE 4.18: Hierarchical feature maps for complete and incomplete RGB-Depth pair .

RGB-Depth pair and the other an incomplete RGB-Depth pair with missing depth. Additionally, hierarchical feature maps corresponding to these samples are presented. It can be observed that in case of missing depth after block 2, transformer features learn significant information.

4.2.8 Analysis of Failure Cases

After carefully examining the shortcomings of proposed methodology and examples of failure cases are presented in Fig. 4.19. First two rows of Fig. 4.19 are representative examples of contextual salience with cluttered background.

In first row, man walking on street is near the viewpoint and also prominent

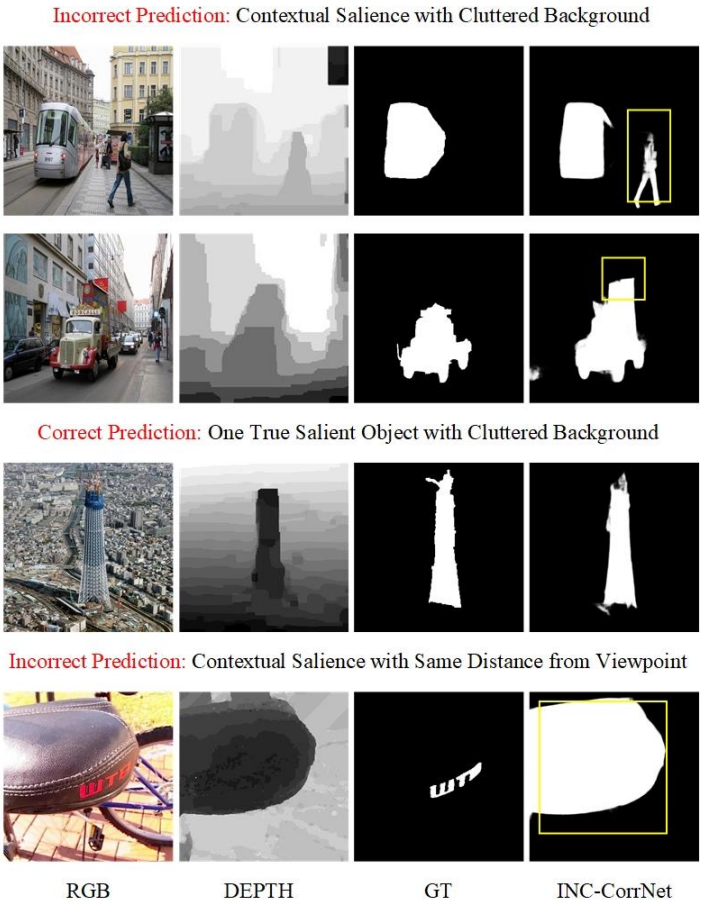


FIGURE 4.19: Visual analysis of failure cases.

in the scene and in second row, the connectivity of boards with the vehicle make them salient. Therefore, contextual saliency lead to false positive predictions and

require saliency ranking to predict true positives only. As can be described by third row of Fig. 4.19, that, one true salient object can be identified correctly by proposed method in cluttered background. In fourth row of Fig. 4.19, although the seat of bicycle is prominent in both RGB and depth but red logo identified as salient object in ground truth map has high contrast with the seat of bicycle. Therefore, this ambiguity can also be minimized using saliency ranking.

4.2.9 Model Complexity Analysis

In Table 4.11, a comprehensive complexity analysis of each module of proposed INC-CorrNet is presented. The number of parameters of the proposed model is 105M and model size is 403MB. It can be observed that model complexity highly depend on backbone network. Therefore, using a single Conformer plays active role to reduce number of parameters.

TABLE 4.11: Ablation study regrading model complexity. Complexity analysis of each component of proposed INC-CorrNet.

Complexity analysis of each component of proposed INC-CorrNet.				
Model	Backbone	Backbone + Coarse Guidance	Backbone + Coarse Guidance +SCLR	Backbone + Coarse Guidance +SCLR+DCLR
Model Size (MB)↓	318	322.9	338	403
No. of Param. (M)↓	83.5	84.5	88.5	105

4.3 Conclusion

In this chapter, two components of implicit depth quality-aware model was presented. First was the depth quality assessment module to distinguish between

good and bad quality depth. Based on the depth quality knowledge, only good quality depths were utilized in second task which is salient object detection. Hence, a distinctive RGB-D salient object detection framework aimed at alleviating the influence of low-quality depth maps on detection accuracy has been developed. In contrast to existing algorithms, the proposed approach involves formulating the learning process for incomplete modality salient object detection by classifying and discarding low-quality depth images. This marks the pioneering instance of an incomplete multi-modality salient object detection model proficient in elucidating the shared latent correlation representation across RGB and depth modalities. The proposed model comprises three stages: conceal, correlate, and fuse. Furthermore, these stages are independently applied in both deep and shallow layers of the backbone network. The SCLR block comprises of encoders crafted to extract distinct feature representations from RGB and depth images, enabling the model to understand the contributions of these features to the object boundaries prediction. While encoders in DCLR block enable the selective enhancement of feature representations through channel attention and spatial attention for salient object detection. Additionally, the proposed correlation model is employed to unveil latent correlations among feature representations from the RGB and depth modalities. This enhances detection robustness for missing depth modality. Then shallow and deep correlated representations are combined in fusion block. The comprehensive experiments are conducted to assess the efficacy of the proposed method. Both quantitative and qualitative analyses with SOTA models on six widely used datasets underscore the effectiveness of the proposed approach. To scrutinize the influence of the proposed components in the network, several ablation experiments are conducted. The MAE metric for the proposed model has significantly decreased by 3.8%, 8.3%, 14.4%, 5.7%, 22.5%, 0.8% on NJU2K, NLPR, LFSD, SIP, STERE and RGBD135 datasets, respectively, as compared to SOTA models. Notable performance gains have been realized with average increase of 1.1% (S-measure), 1.32% (max F-measure) and 1.3% (max E-measure) across five datasets. This work is intended to be used as a catalyst to improve representations from incomplete RGB-D modalities and advance saliency detection tasks.

Chapter 5

Conclusion and Future Work

Multi-modality salient object detection is the identification and highlighting of visually distinct objects across various data modalities in an image. The development of wide range of sensors to represent other modalities has contributed to use cross-modalities in artificial intelligence applications. Hence, extracting relevant features from the discrepancy of multi-modalities is a trending research area. Although efforts have been made, multi-modality salient object detection still presents unresolved challenges and continues to face new obstacles.

5.1 Summary and Contribution

In this research work, two salient object detection frameworks based on deep learning have been developed. 1) A depth quality-aware and optimal cross-modal fusion network for a saliency detection. 2) Identify and discard severely noisy depth images and develop a model which is able to detect salient object with RGB missing depth along with complete RGB-D pair. Thus to implement incomplete RGB-D salient object detection framework. This chapter summarizes the accomplishments of this research work and concludes the research achievements. Future directions of this research work is also presented in this chapter.

5.1.1 Research Summary

Everyday life has enormous amount of unstructured data that can take various forms, called modality. The most widely available visual content is represented as RGB modality. In recent years, with the advent of depth sensors such as Microsoft Kinect and Time-of-Flight etc., depth modality is used in conjunction with RGB modality for salient object detection. However, depth quality encounters degradations due to inaccurate measurements from depth sensors and several environmental conditions. Low quality depth degrades the performance of saliency detection. Therefore, SOD framework should consider the depth quality along with real complementarity features selection and fusion. Three major challenges in multimodal SOD framework implementation have been discussed in this research. Major contributions involve optimal selection of discriminant features from cross-modalities, effective multi-modal fusion and depth quality awareness. Promising results have been obtained for saliency detection task.

5.1.2 Research Contributions

1. **CVit-Net depth quality-aware RGB-D saliency detection model:** In the first contribution, an end-to-end depth quality aware SOD model CVit-Net is developed. CVit-Net correlates the edge saliency map with RGB edge map and depth edge map to reduce the impact of low quality depth images in salient object detection application using novel operation-wise shuffle channel attention. Contribution of low level features in deep learning models is different from deep features. Therefore, the proposed CVit-Net extracts low level details in Local Detail Enhancement (LDE) module and high level details in Global Detail Enhancement (GDE) module. The raw RGB and depth features are extracted using backbone network. In proposed model, first time Conformer network is used as backbone. The intuition behind using Conformer is its distinct two streams i.e. CNN and Transformer streams, beneficial in parameter reduction and local and global context extraction. RGB is a detailed modality specifying many low level features

such as edges, patterns, color etc. while depth possesses straight forward structural cues. RGB is fed to CNN stream and depth is fed to Transformer stream for modality-specific feature extraction. Edge details from LDE are combined with saliency details extracted through reverse attention from GDE. The comprehensive evaluation shows that mean absolute error has been decreased by 15%, 13.4%, 9.1%, 5.6% and 3.6% in SIP, LFSD, STERE, NLPR and RGBD135 datasets, respectively. While increase in S-measure,max F-measure and max E-measure is observed by an average of 1.0% , 0.7% and 0.63% , respectively, across five datasets. Hence proposed explicit depth quality-aware SOD model outperforms SOTA models.

2. **INC-CorrNet incomplete multi-modality saliency detection:** The second contribution of this research work focuses on obstinate condition for depth quality-aware SOD model. That is, the quality of some depth maps is so poor that they are discarded and leading to incomplete modality saliency learning problem. The proposed model consists of two modules. One is depth quality assessment module which assigns a quality score to depth images and based on a threshold value, low quality depths are discarded. Second is salient object detection module which is trained using two types of data, complete RGB and depth and RGB present depth missing. This is the first incomplete multi-modality salient object detector, providing a robust common latent correlation model. Proposed model consists of three steps: conceal, correlate and fuse. Similar to first contribution the Conformer is opted as backbone and treat low and high level features separately in SCLR and DCLR modules respectively. The proposed model is robust to missing depth due to noisy depth or due to scarcity of depth modality as compared to RGB modality. Thorough evaluation demonstrates the efficacy of proposed model. Specifically, the MAE metric for the proposed model has significantly decreased by 3.8%,8.3%, 14.4%, 5.7%, 22.5%, 0.8% on NJU2K, NLPR, LFSD, SIP, STERE and RGBD135 datasets, respectively, as compared to SOTA models. Notable performance gains have been realized with average increase of 1.1% (S-measure), 1.32% (max F-measure) and 1.3%

(max E-measure) across five datasets. This work aims to enhance saliency detection tasks through improved representations from incomplete RGB-D modalities.

5.1.3 Future Directions

The outcomes of the research conducted in this thesis on RGB-D salient object detection using deep learning are promising in several aspects. Some potential future directions are listed below.

- Saliency Ranking: Additional depth source improves detection accuracy in complex scenarios where RGB fails. They include illumination variation, low contrast, cluttered background and appearance changes etc. However, due to constraint of depth quality, depth quality-aware SOD model is proposed. When features are extracted from, low quality depth and RGB with cluttered background, multiple salient objects can be captured. However, even for the case of RGB with cluttered background, some objects constitute to be highly salient while others may be relatively less salient. Distance from the viewpoint also defines saliency in some cases. Therefore, due to lack of universal agreement about salient object, saliency ranking should be used to avoid false positive predictions.

- Co-attention for incomplete RGB-D modality SOD: Co-attention has been widely applied to extract the correlation of multi-modalities. One of the reason of performance degradation in SOD is misalignment between RGB and depth edges. Therefore, in future work, co-attention can be opted for incomplete RGB-D modality saliency learning to better extract the correlation of two modalities. And it can also be applied to align the edges of RGB and depth for complete modality saliency learning.

- Task-Driven Datasets: Most of the salient object detection datasets consist of varied indoor and outdoor scenes and only SIP dataset is task-driven comprised of salient-in-person dataset. Therefore, one of the potential direction

for future is to develop task-driven datasets. For example in car driving assistant model, a dataset of road-sign will be helpful. For wildlife monitoring task, specific animal specie dataset may be required. For industrial inspection application, specific product dataset will be beneficial for defect detection.

- Feasibility to Real-world Applications: Although the inference speed of proposed model is acceptable for variety of applications such as i) certain image processing tasks, such as image filtering and feature extraction etc. ii) remote sensing iii) non-Interactive computer graphics tasks such as rendering complex scenes or generating high-quality animations offline. However, for real-world applications this speed is not feasible. Therefore, in future model compression, knowledge distillation or light-weight backbone network adoption can be used to increase the speed of model.

- Zero-shot learning: The applications of salient object detection are very vast. And enormous amount of labelled data for each salient object detection application is not feasible. Therefore, in future zero-shot saliency learning model can be developed so that a generalized solution can be provided for new object categories that are not present in training phase.

Bibliography

[1] H. Chen, Y. Li, Y. Deng, and G. Lin, "Cnn-based rgb-d salient object detection: Learn, select, and fuse," *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2076–2096, 2021.

[2] G. Liu and D. Fan, "A model of visual attention for natural image retrieval," in *2013 international conference on information science and cloud computing companion*. IEEE, 2013, pp. 728–733.

[3] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *International conference on machine learning*. PMLR, 2015, pp. 597–606.

[4] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1007–1013.

[5] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, "Salient object segmentation via effective integration of saliency and objectness," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1742–1756, 2017.

[6] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 4, pp. 862–875, 2014.

[7] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Dense saliency-based spatiotemporal feature points for action recognition," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 1454–1461.

[8] K. R. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1896–1909, 2016.

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[10] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2020, pp. 263–273.

[11] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE transactions on medical imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[12] Y.-H. Wu, S.-H. Gao, J. Mei, J. Xu, D.-P. Fan, R.-G. Zhang, and M.-M. Cheng, "Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 3113–3126, 2021.

[13] A. Mondal, "Camouflaged object detection and tracking: A survey," *International Journal of Image and Graphics*, vol. 20, no. 04, p. 2050028, 2020.

[14] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 356–370, 2016.

[15] N. Martinel, C. Micheloni, and G. L. Foresti, "Kernelized saliency-based person re-identification through multiple metric learning," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5645–5658, 2015.

[16] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proceedings of the tenth ACM international conference on Multimedia*, 2002, pp. 533–542.

[17] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4486–4497, 2021.

[18] Y. Chen and W. Zhou, "Hybrid-attention network for rgb-d salient object detection," *Applied Sciences*, vol. 10, no. 17, p. 5806, 2020.

[19] A. Ciptadi, T. Hermans, J. M. Rehg *et al.*, "An in depth view of saliency." in *BMVC*, 2013, pp. 1–11.

[20] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, "Depth really matters: Improving visual salient region detection with depth." in *BMVC*, 2013, pp. 1–11.

[21] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proceedings of international conference on internet multimedia computing and service*, 2014, pp. 23–27.

[22] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*. Springer, 2014, pp. 92–109.

[23] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 478–487.

[24] S. Chen and Y. Fu, "Progressively guided alternate refinement network for rgb-d salient object detection," in *European conference on computer vision*. Springer, 2020, pp. 520–538.

[25] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for rgb-d salient object detection and beyond," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5541–5559, 2021.

[26] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, and B. Tang, "Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 4481–4490.

[27] X. Jia, C. DongYe, and Y. Peng, "Siatrans: Siamese transformer network for rgb-d salient object detection with depth image classification," *Image and Vision Computing*, vol. 127, p. 104549, 2022.

[28] 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 conference on computer vision and pattern recognition workshops*, 2015, pp. 25–32.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] 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.

[32] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 558–567.

[33] 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.

[34] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao, "Salient object detection for rgb-d image by single stream recurrent convolution neural network," *Neurocomputing*, vol. 363, pp. 46–57, 2019.

[35] 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.

[36] 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 cybernetics*, vol. 48, no. 11, pp. 3171–3183, 2017.

[37] N. Wang and X. Gong, "Adaptive fusion for rgb-d salient object detection," *IEEE Access*, vol. 7, pp. 55 277–55 284, 2019.

[38] 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*. Springer, 2020, pp. 665–681.

[39] W. Zhang, Y. Jiang, K. Fu, and Q. Zhao, "Bts-net: Bi-directional transfer-and-selection network for rgb-d salient object detection," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[40] 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.

[41] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 136–145.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[43] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3485–3492.

[44] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended cssd," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 4, pp. 717–729, 2015.

[45] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 1115–1119.

[46] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2806–2813.

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

[48] 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.

[49] 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 international conference on computer vision*, 2017, pp. 4548–4557.

[50] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1597–1604.

[51] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv preprint arXiv:1805.10421*, 2018.

[52] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 733–740.

[53] S. Kanwal and I. A. Taj, "Cvit-net: A conformer driven rgb-d salient object detector with operation-wise attention learning," *Expert Systems with Applications*, vol. 225, p. 120075, 2023.

[54] ——, "Incomplete rgb-d salient object detection: Conceal, correlate and fuse," *Pattern Recognition*, vol. 155, p. 110700, 2024. [Online]. Available: https://doi.org/10.1016/j.patcog.2024.110700

[55] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "Rgb-d salient object detection: A survey," *Computational Visual Media*, vol. 7, pp. 37–69, 2021.

[56] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*. Springer, 2012, pp. 101–115.

[57] X. Zhang, Y. Xu, T. Wang, and T. Liao, "Multi-prior driven network for rgb-d salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[58] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for rgb-d salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2343–2350.

[59] H. Xue, Y. Gu, Y. Li, and J. Yang, "Rgb-d saliency detection via mutual guided manifold ranking," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 666–670.

[60] L. Jiang, A. Koch, and A. Zell, "Salient regions detection for indoor robots using rgb-d data," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 1323–1328.

[61] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[62] M. I. Jordan, "Serial order: A parallel distributed processing approach," in *Advances in psychology*. Elsevier, 1997, vol. 121, pp. 471–495.

[63] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[64] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[65] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[66] 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," *arXiv preprint arXiv:2010.11929*, 2020.

[67] 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.

[68] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3051–3060.

[69] 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.

[70] 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.

[71] 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.

[72] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao, "Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 6800–6815, 2022.

[73] Y. Han, L. Wang, A. Du, and S. Jiang, "Lianet: Layer interactive attention network for rgb-d salient object detection," *IEEE Access*, vol. 10, pp. 25 435–25 447, 2022.

[74] 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.

[75] X. Fang, M. Jiang, J. Zhu, X. Shao, and H. Wang, "M2rnet: Multi-modal and multi-scale refined network for rgb-d salient object detection," *Pattern Recognition*, vol. 135, p. 109139, 2023.

[76] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4722–4732.

[77] X. Wang, B. Jiang, X. Wang, and B. Luo, "Mutualformer: Multi-modality representation learning via cross-diffusion attention," *arXiv e-prints*, pp. arXiv–2112, 2021.

[78] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Caver: Cross-modal view-mixed transformer for bi-modal salient object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 892–904, 2023.

[79] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, and N. Barnes, "Uncertainty inspired rgb-d saliency detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5761–5779, 2021.

[80] 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*, vol. 30, pp. 7012–7024, 2020.

[81] Y. Zhai, D.-P. Fan, J. Yang, A. Borji, L. Shao, J. Han, and L. Wang, "Bifurcated backbone strategy for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 8727–8742, 2021.

[82] 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.

[83] J. Zhao, Y. Zhao, J. Li, and X. Chen, "Is depth really necessary for salient object detection?" in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1745–1754.

[84] W.-D. Jin, J. Xu, Q. Han, Y. Zhang, and M.-M. Cheng, "Cdnet: Complementary depth network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3376–3390, 2021.

[85] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu *et al.*, "Calibrated rgb-d salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9471–9481.

[86] Q. Zhang, Q. Qin, Y. Yang, Q. Jiao, and J. Han, "Feature calibrating and fusing network for rgb-d salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[87] F. Wang, R. Wang, and F. Sun, "Dcmnet: Discriminant and cross-modality network for rgb-d salient object detection," *Expert Systems with Applications*, vol. 214, p. 119047, 2023.

[88] Y. Niu, S. Zhou, Y. Dong, L. Wang, J. Wang, and N. Zheng, "Bidirectional feature learning network for rgb-d salient object detection," *Pattern Recognition*, p. 110304, 2024.

[89] L. Meng, M. Yuan, X. Shi, L. Zhang, Q. Liu, D. Ping, J. Wu, and F. Cheng, "Rgb depth salient object detection via cross-modal attention and boundary feature guidance," *IET Computer Vision*, vol. 18, no. 2, pp. 273–288, 2024.

[90] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 5, pp. 865–878, 2016.

[91] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[92] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Transcmd: cross-modal decoder equipped with transformer for rgb-d salient object detection," *arXiv preprint arXiv:2112.02363*, vol. 3, no. 9, 2021.

[93] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 367–376.

[94] P. Charitidis, G. Kordopatis-Zilos, S. Papadopoulos, and I. Kompatsiaris, "Operation-wise attention network for tampering localization fusion," in *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2021, pp. 1–6.

[95] 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/CVF international conference on computer vision*, 2019, pp. 8779–8788.

[96] 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. 51, no. 1, pp. 88–100, 2020.

[97] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "Rgb-d salient object detection via 3d convolutional neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1063–1071.

[98] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for rgb-d salient object detection," *IEEE transactions on image processing*, vol. 30, pp. 1949–1961, 2021.

[99] X. Zhao, Y. Pang, L. Zhang, H. Lu, and X. Ruan, "Self-supervised pretraining for rgb-d salient object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 3463–3471.

[100] X. Cheng, X. Zheng, J. Pei, H. Tang, Z. Lyu, and C. Chen, "Depth-induced gap-reducing network for rgb-d salient object detection: An interaction, guidance and refinement approach," *IEEE Transactions on Multimedia*, 2023.

[101] 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.

[102] H. Zhu, X. Sun, Y. Li, K. Ma, S. K. Zhou, and Y. Zheng, "Dftr: Depth-supervised fusion transformer for salient object detection," *arXiv preprint arXiv:2203.06429*, 2022.

[103] N. T. Thu, M. D. Hossain, and E.-N. Huh, "Ec2net: Efficient attention-based cross-context network for near real-time salient object detection," *IEEE Access*, vol. 11, pp. 39 845–39 854, 2023.

[104] R. Song, H. Ko, and C. Kuo, "Mcl-3d: A database for stereoscopic image quality assessment using 2d-image-plus-depth source," *arXiv preprint arXiv:1405.1403*, 2014.

[105] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, 2014.

[106] L. Li, X. Chen, J. Wu, S. Wang, and G. Shi, "No-reference quality index of depth images based on statistics of edge profiles for view synthesis," *Information Sciences*, vol. 516, pp. 205–219, 2020.

[107] S. Xiang, L. Yu, and C. W. Chen, "No-reference depth assessment based on edge misalignment errors for t+ d images," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1479–1494, 2015.

[108] Y. Huang, L. Li, H. Zhu, and B. Hu, "Blind quality index of depth images based on structural statistics for view synthesis," *IEEE Signal Processing Letters*, vol. 27, pp. 685–689, 2020.

[109] C. Zeng, S. Kwong, and H. Ip, "Dual swin-transformer based mutual interactive network for rgb-d salient object detection," *Neurocomputing*, vol. 559, p. 126779, Nov. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2023.126779

[110] T. Ikeda and M. Ikehara, "Rgb-d salient object detection using saliency and edge reverse attention," *IEEE Access*, vol. 11, p. 68818–68825, 2023. [Online]. Available: http://dx.doi.org/10.1109/access.2023.3292880

[111] A. Li, Y. Mao, J. Zhang, and Y. Dai, "Mutual information regularization for weakly-supervised rgb-d salient object detection," *IEEE Transactions on*

Circuits and Systems for Video Technology, vol. 34, no. 1, p. 397–410, Jan. 2024. [Online]. Available: http://dx.doi.org/10.1109/tcsvt.2023.3285249

[112] T. Chen, J. Xiao, X. Hu, G. Zhang, and S. Wang, "Adaptive fusion network for rgb-d salient object detection," *Neurocomputing*, vol. 522, pp. 152–164, 2023.

[113] H. Bi, R. Wu, Z. Liu, H. Zhu, C. Zhang, and T.-Z. Xiang, "Cross-modal hierarchical interaction network for rgb-d salient object detection," *Pattern Recognition*, vol. 136, p. 109194, 2023.

[114] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, "Specificity-preserving rgb-d saliency detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4681–4691.

[115] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jegou, and M. Douze, "Levit: A vision transformer in convnet's clothing for faster inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 12 259–12 269.

[116] Q. Zhang and Y.-B. Yang, "Rest v2: simpler, faster and stronger," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 440–36 452, 2022.