Chapter 6

SL-Net: Self-Learning and Mutual Attention based Distingushed Window for RGBD Complex Salient Object Detection

Significant improvement has been noticed in salient object detection by multi-modal cross-complementary fusion between Depth and RGB features. The multi-modal feature extracting backbone of existing networks cannot extract complex RGB and color images effectively, which limits the performance of salient object detection in complex and challenging situations. In this thesis, a composite backbone with a mutual attention-based distinguished window is proposed to enhance the salient

region and minimize the non-salient region. The distinguished window based on the channel-wise, spatial, mutual, and feature-level attention is inserted in each encoder stage to enhance the saliency features. Finally, a novel self-learning-based decoder which is capable of utilizing multi-level features is designed to get the accurately dense prediction. The multi-level fusion is guided by deep global localized features. The performance of salient object detection could significantly be enhanced in this way. The details introduction and research gaps is discussed in next section 6.1.

6.1 Introduction

The salient object detection inspired from human visual attention mechanism. It aims to identify and predict the most prominent and conspicuous object in an image irrespective of size, texture, color, and complex background. Significant development has been noticed during recent years due to the wide range of its applications; such as online visual tracking [186], semantic segmentation [177], object classification [18], re-identification [187], video saliency [188], [189], and content-based image editing [179].

Most existing RGBD SODs [43], [110], [109], [133] used Depth and RGB as separate inputs in deep CNN, which further explored complementary features during the fusion process. The RGB and depth modality have complementary information. RGB modality has ample informationa about regional, color, textural, spatial, and high-level semantic and contextual cues. Similarly, the Depth modality

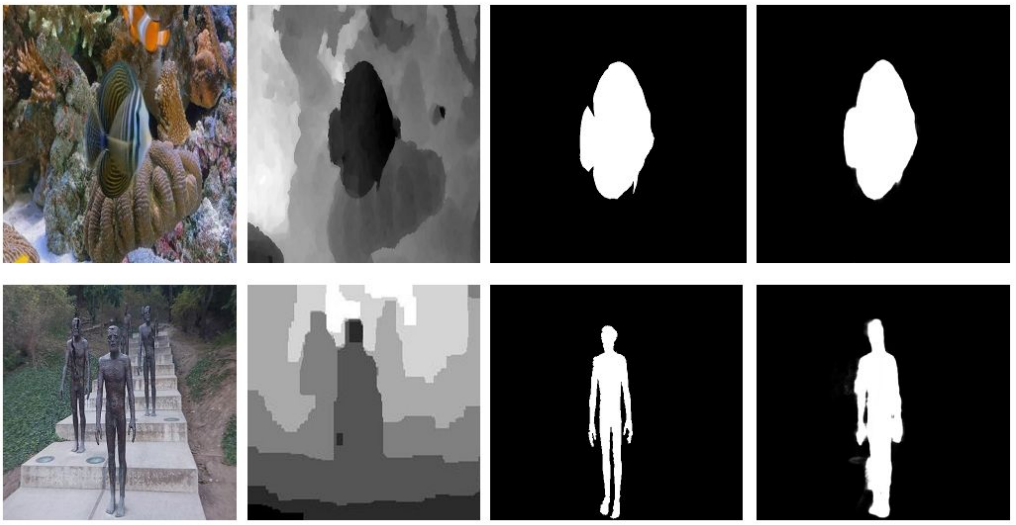

FIGURE 6.1: The importance of mutual attention mechanism to distinguish the salient object in complex and clutter background.

has the affluent geometrical, structural, 3D spatial, edge, and boundary informations, which are complementary to each other. The current and efficient methods [133], [129], [137], [131], [134], [136], used both modalities separately during the encoding stage, followed by straightforward complimentary fusion processes to fuse and predict saliency. These models gradually improved the performance while stuck in complex and cluttered backgrounds. There are three vital issues during the encoding stage and utilizing both modalities to improve the performance. 1 > How to enhance the salient regions and minimize the irrelevant features in the non-salient regions during the encoding stage? 2 > How to model the geometrical structure and boundary information of depth modality to purify the RGB features during the encoding stage? 3 > How to develop the multi-model multi-stage fusion strategy to provide equal importance to low-level purified encoder features, high-level semantic

features, and global localized features?

In the recent approaches in RGBD saliency [133], [129], [137], [131], depth channel fed into CNN to extract saliency features and finally utilized in the fusion process. Consequently, Depth based geometrical information is not fully utilized. The geometrical information provides the guided reference window for identifying salient and non-salient regions. In some complex scenarios and low depth images, which are shown Fig.6.1, the RGB information is vital to overcome these limitations and stop further propagation of irrelevant features in the following stages of the encoder. The importance of geometrical information to distinguish saliency is our motivation to propose the depth guided, mutual attention based distinguished window ($MADW$). This distinguished window is a reference surface for salient and non-salient regions. It is applied before each VGG encoder layer to enhance saliency in salient regions and minimize irrelevant disturbance in non-salient regions.

Additionally, the effective feature fusion principle is needed to interact the global localized deep features with enhanced encoded features and high-level semantic features. The existing models [133], [180], [137] , [130], [129] only fused same stage encoded RGB and Depth features by using simple element-wise multiplication and addition operation followed by some enhancement mechanism. The common limitations of these models [110], [109], [133], [129], [137], [131], [134], [136] [180], [130] are the inefficient fusion of two different modalities to predict saliency correctly. Furthermore, the multi-stage, multi-source, and multi hierarchy, RGB and depth informations are exceptionally heterogeneous. Subsequently, it makes the fusion

process complicated. The limitations mentioned above in the fusion process are the basis for designing self-learnable, *Self Learning-Based Dense Decoder-SDD*, among local enhanced encoded, global localized, and high-level semantic features.

Attention based Model: Attention Mechanism is widely used in other applications like Image captioning [190], Language Modeling [183] and 3D block matching [191]. It assigned a different weight to provide distinguishable essence to different regions. The spatial attention [64], channel-wise attention [144], and self-attention mechanism [178] are prevailing and prominent mechanisms to improve the performance in various applications. The spatial attention removes the regional and spatial discrepancy, while the channel-wise attention mechanism distinguishes the channel-wise features. In contrast, a non-local network-based self-attention mechanism is used to compute the long-range dependency. Some other attention mechanism such as dual-attention is also used. Recent models S2MA [137] introduced attention mechanisms to overcome the regional disparity to compute saliency. The non-local attention maps were initially used in language modeling. The self-attention-based learning model for 2D or 3D application is proposed by Wang *et al.* [183]. The initial model of the non-local network is based on the 3D block matching BM3D [191]. Similarly, another model based on the attention map, CMSA [190], is utilized to segment the object by using a given input string in natural language. These models improved the performance and showed the importance of an attention map. The above models used the attention map to enhance the saliency features while lacking the full utilization of the attention map during features generation, localization, and

features fusions.

Consequently, the Self Learning-Based Dense Decoder-SDD with Mutual Attention-based Distinguished window-MADW accomplishes the state-of-the-art performance. To address the aforementioned limitations and challenges, our proposed model, Mutal Attention based Distingused Window and Self Learning based RGBD, provides the following distinct contributions to improve performance of SOD.

- A composite backbone is proposed for the encoder to improve the encoded saliency features by adding depth-guided, mutual attention-based distinguished window-MADW before each stage of encoder.
- We design a depth-guided mutual attention-based distinguished window to remove the discrepancy in non-salient regions and enhance RGB features in salient regions, using a spatial, channel, mutual, and feature-level attention mechanism.
- A novel self-learning-based dense decoder is proposed to integrate enhanced encoded features, global localized features, and high-level semantic features.
- In conclusion, extensive experiments have been conducted with seven publicly available datasets to demonstrate performance improvements with other state-of-the-art methods.

To target the limitation mentioned above, we propose the Mutual Attention-based Distinguished window and self-learning-based fusion model, $SL - Net$, to predict

exact salient object detection. This distinguished window is formulated using a channel-wise, spatial, features level, and mutual attention mechanism to automatically enhance the salient regions and minimize the non-salient regions during the encoding stage. Therefore, more enriched features are available for decoding to predict accurate salient object. The proposed model used two streams encoder to produce enhanced saliency features and minimized irrelevant features during the encoding. The mutual attention window guides the deep-CNN features to enhance and accurately identify the global localized features. The decoder stream is composed of novel Self Learning-based Dense Decoder-SDD. It is used to progressively fuse enhanced features through global localized and high-level semantic features. The proposed SDD and mutual attention map predict the exact saliency in complex and clutter images. This chapter describes the deep learning based 3D salient object detection model. Section 6.2 describes and defines the proposed method $SL-Net$ in detail. Section 6.3 discusses the Experiment Set-Up and demonstrates the Performance of Self-Learning with other state-of-the-art methods. Section 6.4 describes the conclusion and the future scope of improvements in the Self-Learning model.

6.2 The proposed method

6.2.1 Overview

The proposed model has two encoders and one decoder stream. Two encoder streams produce the RGB and Depth features, while the decoder stream fuse both stream features to produce saliency. Most of the existing models either used separate encoder and decoder with simple fusion model to combine the saliency features or used the cross-complementary fusion of side outputs [134], [124] straightforwardly to find the saliency. Therefore, existing fusion models are based on the element-wise multiplication and addition operations, which are insufficient in complex and clutter backgrounds. Because, in these type of networks, irrelevant features propagate further in the following stages, and some essential, regional, spatial and textural features have not been enhanced during feature extraction in the encoder.

These existing drawbacks are the motivation behind proposing a stage-wise, depth-guided, Mutual Attention Based Distinguished Window-MADW, to enhance the essential features and minimize the irrelevant features in the encoding stage. This process produces enhanced and more accurate saliency features for the decoder to produce exact salient objects in complex and cluttered images. RGB features rich in color, regional, spatial, and texture information. It is superior over depth modality to finding semantic features. However, the high-level semantic features of depth

modality have a relatively simple structure, sound in object localization, and minimize the non-salient regions. Therefore, a Mutual Attention Based Distinguished Window-MADW is only used in the color features in the encoding stage. This attention map is also used to produce deep global localized features. The deep global localized features are utilized in each SDD module to provide the reference window to enhance the essential features and minimize the complex background. The global localized features, high-level semantic features, and enhanced encoded features from both modalities are finally utilized by the proposed self Learning-based dense decoder to fuse multi-stages, and multi-resolution hierarchical features.

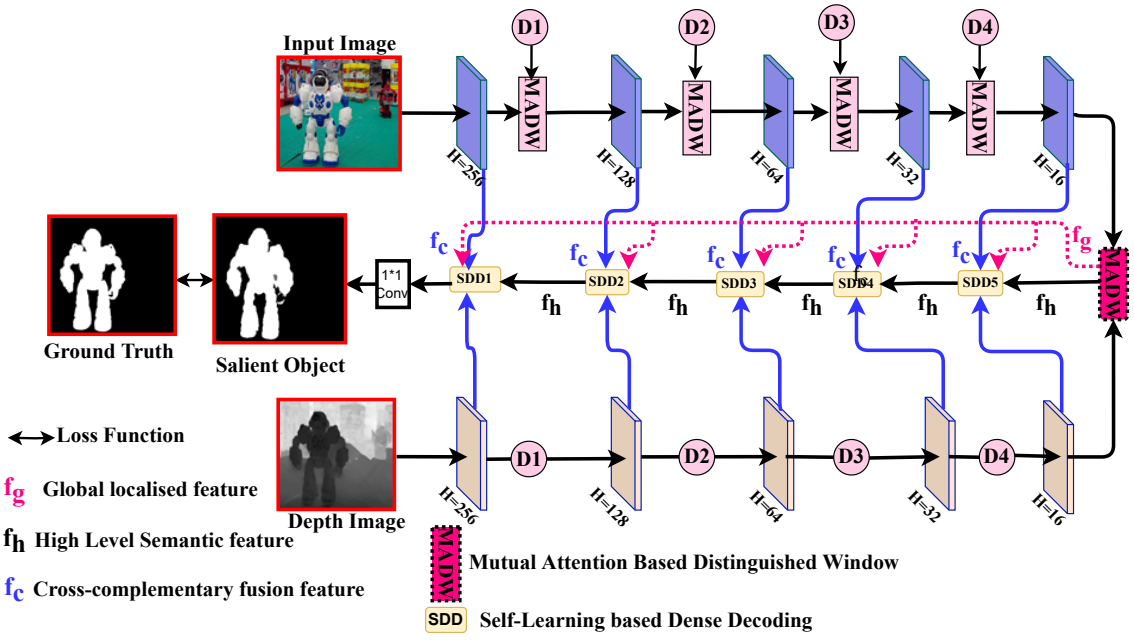

FIGURE 6.2: The illustration of the proposed framework $SL-Net$.

6.2.2 Enhanced Encoded Feature through Composite Backbone

Let us define the composite backbone network with five convolution and attention blocks, MADW,(i.e., $Conv1\_2$, MADW, $Conv2\_2$, MADW, $Conv3\_3$, MADW, $Conv4\_3$, MADW, and $Conv5\_3$) in color stream. While the Depth stream has only convolution blocks without $MADW$(i.e., $Conv1\_2$, $Conv2\_2$, $Conv3\_3$, $Conv4\_3$, and $Conv5\_3$). The Depth stream has simple geometrical information, which is essential for better localization of the salient object. The outputs produced by these blocks are denoted as C^i_{RGB} and D^i_{depth}, and their side outputs are F^i_{RGB} and F^i_{depth} for both RGB and Depth stream respectively, where $1 \leq i \leq 5$. The composite backbone network is designed with five layers of $VGG-16$ network along with distinguished attention windows to enhance the salient regions and minimize the non-salient region.

6.2.3 Mutual Attention Based Distinguished Window-MADW

The RGB and Depth maps have cross-complementary features, which are essential to detect a complete salient object. This objective has been achieved by a guided composite backbone network to produce enhanced encoded features. Depth modality contains details about border, edge, shape, and structure. At the same time,

color modality contains color, texture, region, and high-level semantic and contextual features. Therefore, both modalities are essential in complex and clutter backgrounds. The existing models used a simple VGG network to produce CNN features and side outputs, which are incapable of enhancing the spatial distribution loss and correlation among different channel features. A depth map has distinguishing characteristics, which are essential to incorporate in encoder stages. The distinguishing characteristics are missing in most exiting models in encoding stages, which is a milestone in improving the performance. These are formulated as a guided mutual attention-based distinguished window. The Depth guided, channel-wise [153], spatial, and mutual attention maps provide the reference window to remove irrelevant features and enhance the salient regions. The proposed attention map also solves the low depth issues. Because Depth guided mutual attention maps enhance the feature generation in color modality during the encoding stage.

The Depth modality-based distinguished window purifies the RGB-based features maps. The depth features are processed by a series of operations using a spatial attention mask. The spatial mask (3×3) is followed by a large spatial window (7×7). It is suggested by Xu et al. [152] where two (7×7) spatial masks are used, while in our proposed method, (3×3) window is followed by (7×7). The (3×3) window enriches the saliency features while the (7×7) window emphasizes the regional saliency in a large receptive field. The channel-wise, spatial attention

is formulated in Eq. 6.1 and shown in Fig. 6.3 as follows:

$$S_w(D^i_{depth}) = \vartheta(\psi_{7\times7}(\psi_{3\times3}(MaxPool(D^i_{depth})) \tag{6.1}$$

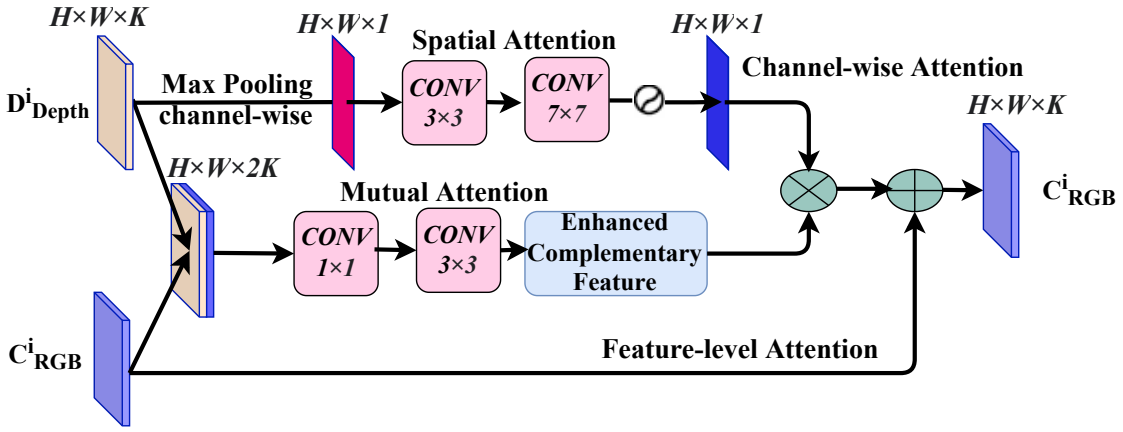

FIGURE 6.3: The illustration of the proposed Mutual Attention Based Distinguished Window-MADW.

Where ψ is a convolution operation with specified size, and ϑ is channel-wise max pool operation with sigmoid function. $S_w(D^i_{depth})$ is a spatial attention window that highlights the border, region, edge, and shape.

The cross-complementary features at each encoding stage are essential to remove the irrelevant features and minimize the background in a complex scenario. The fused modality-based mutual attention map is proposed to guide the features generations at each encoding stage. Therefore, the enhanced encoded features will utilize in the decoding stage. Firstly, reduce the channels in concatenated features from both modalities by applying 1×1 convolution mask. Then after 3×3 convolution mask enhance the more details in combined features. The mutual attention feature maps

M_{att} is defined in Eq. 6.2 as follows:

$$M_{att}(C^i_{RGB}, D^i_{Depth}) = \vartheta\left(\psi_{3\times3}\left(\psi_{1\times1}\left(\left|C^i_{RGB}, D^i_{Depth}\right|\right)\right)\right) \tag{6.2}$$

Where, $|.|$ is the concatenation of feature maps. The multiplication of spatial attention weight with mutually enhanced features reduces irrelevant features. The RGB-based residual connection highlights and restores the essential features in the encoding stage. Finally, i^{th} stage enhanced and purified feature is defined in Eq. 6.3 as follows:

$$C^i_{RGB} = M_{att}(C^i_{RGB}, D^i_{Depth}) \otimes S_w(D^i_{depth}) \oplus C^i_{RGB}) \tag{6.3}$$

Where, $\otimes$ and $\oplus$ are element-wise multiplication and addition operation.

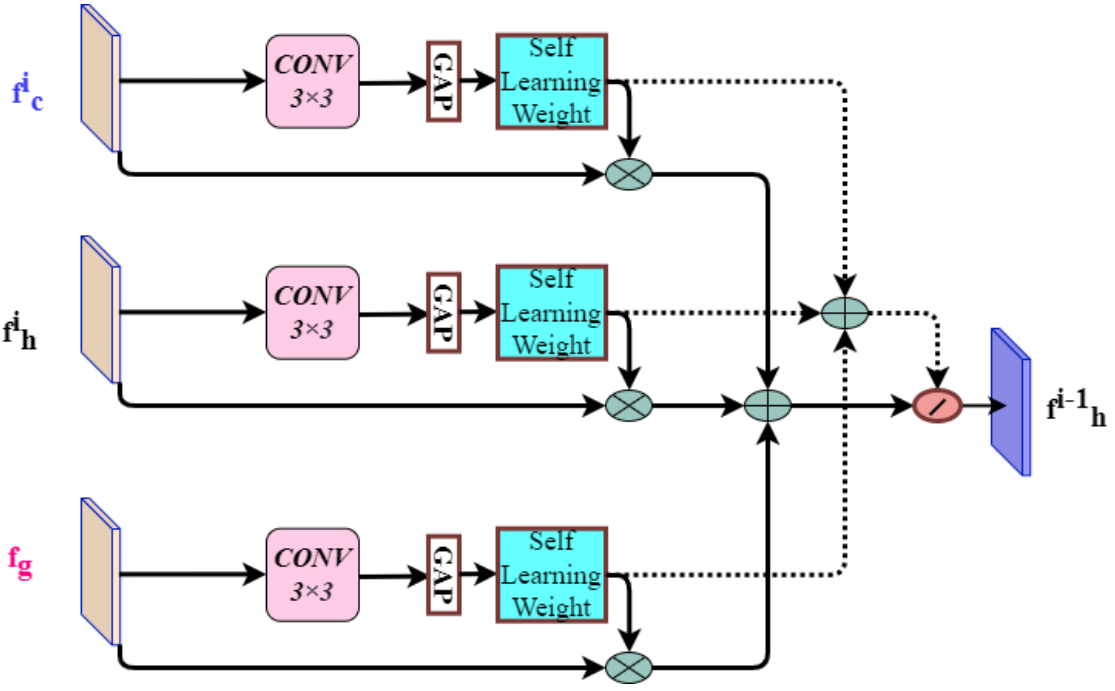

FIGURE 6.4: The design and process of Self-Learning based Dense Decoder- SDD

6.2.4 Self-Learning based Dense Decoding-SDD

The Multi-level distinguished saliency feature through the mutual attention-based distinguished window in the encoding stage is utilized in the proposed self learning-based dense decoding model to learn essential cross-complementary features. The cross-complementary features learn from cross modalities are important contributing steps to predict the exact salient object. The self-learning-based Dense Decoding has four main components. 1 > Global Localized Feature 2 > Cross-complementary fusion(CF) 3 >Dense Decoding 4 > Self-Learning based Aggregation Module. The global localized feature is utilized in each stage of the decoder. The cross-complementary fusion combines the enhanced features of both modalities. The self-Learning-based aggregation model learns the importance of global features, local cross-complementary features, and high-level semantic features. The proposed model of dense decoder describes in Fig. 6.4 in the following stage as follows:

6.2.4.1 Global Localized Feature

The last level of encoder stream in both modalities has high-level semantic and localized information, which are essential features to localize the salient objects. Therefore, it is utilized in all the decoder stages, which provides the reference window for localizing the salient object. The global localized features is utilised as high-level semantic features for the SDD5 module only. The last layer of deep features in

the encoder is processed using Cross-complementary fusion(CF) followed by Dense Decoding.

6.2.4.2 Cross-complementary fusion(CF)

The enhanced side outputs similar to [134], [124] at each stage are utilized in each decoder to reconstruct the saliency features. These purified side outputs of color and depth modality have been utilized and fused to produce the enhanced local features. In this model, the varied resolution features are compressed into smaller (fixed size equal to k) to minimize the number of channels. The compressed features have been achieved by convolution operation with k channels, and its size is 3×3, and with the stride, size is 1. These depth and color features are combined using addition and multiplication operations among cross modalities. The processed features in RGB and $depth$ modality are denoted as F^i_{rgb}, F^i_{depth}, each with equal k channels. The output of the CF module is defined in Eq. 6.4 and in Fig. 6.4 as follows:

$$C^i_f\left(F^i_{rgb}, F^i_{depth}\right) = (F^i_{rgb} \otimes F^i_{depth}) \oplus (F^i_{rgb} \oplus F^i_{depth}) \tag{6.4}$$

In this cross-complementary fusion, " $\otimes$ " and " $\oplus$ " are defined as element-wise multiplication and addition respectively. These operations have characteristics that exploit the commonality and complementary features to increase saliency. The output of the CF model in each stage is fed into a dense decoder.

6.2.4.3 Self Learning Aggregation model

The existing decoding model [49], [131], [137], [130], [64] utilized the side outputs that enhanced the decoder to predict the saliency. Nevertheless, these methods provide the encoding features into corresponding decoding layers. At the same time, stage-wise different encoding and decoding features and global deep localized features have been ignored. For example, the deep-layer encoding features can provide localized informations, and the low level provides semantic guidance into the decoding process. Subsequently, purified cross complementary enhanced encoded features from dense connection (f_c) are aggregated with global localized features (f_g) and high-level semantic features (f_h) in the Self-learning aggregation model. The self-learning-based normalization is first time used in local saliency coherence [192]. In contrast, it is utilized as a self-learning coefficient-based aggregation model of three different features maps in our proposed model. These three multilevel information aggregated using self-learning based dense connection which is different from traditional Unet [193] like structure. The detailed implementation plane is shown in Fig. 6.4. The self-learning weight of each input feature is computed by 3×3 convolution layers followed by global average pooling (GAP). These self-learning weights from three components are aggregated and normalised through SDD^i at i^{th} stage to produce high-level semantic feature at $i - 1^{th}$ stage till $i = 1$ output level $SDD1$ in the proposed aggregation model. Here, $i \in \{5, 4, 3, 2, 1\}$ indexes for different stages. It

is defined in Eq. 6.5 as follows:

$$f_{i-1}^{h} = \frac{\sum_{x\in(c,h,l)} \overbrace{GAP(\psi_{3\times3}(f_i^x))}^{\textbf{Self-learning weight}} \; f_i^x}{\sum_{x\in(c,h,l)} GAP(\psi_{3\times3}(f_i^x))} \tag{6.5}$$

Where a global localized feature is invariant and the same for all stages in the decoder. It is upsampled multiple times to make stage-wise the same resolutions by using linear interpolation for each SDD module. The Self-learning weight highlights the importance of each feature. Finally, the 1×1 convolution mask is applied with the Sigmoid function to produce the final salient object.

6.2.5 Loss Function

The whole network is trained using all training data by standrard binary cross-entropy ($\mathfrak{E}_{loss}$), and the IOU-loss [194]. The IOU-loss, ($\mathfrak{I}_{loss}$) loss function is emphasized the global structural similarity. The loss function, $\mathfrak{L}$, is computed with their respective saliency map Sm and ground truth map Gt. The total loss function is defined as:

$$\mathfrak{L}(Sm, Gt) = \mathfrak{E}_{loss}(Sm, Gt) + \mathfrak{I}_{loss}(Sm, Gt) \tag{6.6}$$

$$\mathfrak{E}_{loss}(Sm, Gt) = -\sum_{k}((Gt_k log(Sm_k) + (1 - Gt_k) log(1 - Sm_k))) \tag{6.7}$$

$$\Im_{loss}(Sm, Gt) = 1 - \frac{\sum_{k \in Gt} Sm_k Gt_k}{\sum_{k \in Gt}(Sm_k + Gt_k - Sm_k \times Gt_k)} \tag{6.8}$$

Where k is defined as level pixel index in ground truth image.

6.3 Experiment and Result Analysis

The proposed framework, $SL - Net$, has two encoders and one decoder network. Encoder streams enhance the saliency feature using a Mutual Attention-based Distinguished Window-MADW before each stage of the VGG-16 network. The decoder stream utilized enhanced encoded, high-level semantics, and global localized features to predict exact salient object detection. The network parameters, Data-set, evaluating parameters, implementation details, and other parameters are described here.

6.3.1 Data-Set

We conduct extensive experiments on seven publicly available RGBD benchmark datasets NJUD dataset [110], NLPR dataset [98] STEREO [97] SSD [163] RGBD-135 [100] and DUT-RGBD [43] and LFSD for complex salient object detection.For a fair comparison to the state-of-the-art method, the same data pattern of [127] for training and testing is used here. In the DUT-RGBD dataset, we use the same data pattern as used in DANet [127].The training set contains random 1400 images

from the $NJUD-2000$ dataset and 650 samples from $NLPR$, following the similar pattern of most state-of-the-art methods [64], [127], [126], [116]. The validation set contains 150 images in which 50 image pairs are from NJUD and 50 image pairs from $RGBD-135$. The rest of the images from all datasets are used as training images. The augmentation of the training set is used to reduce overfitting by randomly flipping and rotating the training images.

6.3.2 Evaluation metrics

The proposed method $SL-Net$ is evaluated with others State-of-the-art methods by using recent evaluation metrics. These metrics are (1) S-Measure, (2) F-Measure, (3) Mean Absolute Error (MAE), and (4) E-measure(E_ψ).

6.3.3 Implementation Details

The proposed method $SL-Net$ is implemented on the PyTorch [195] framework. The VGG-16 model [182] has been used to design the composite backbone network to extract features in the encoder. The training and testing images of all the datasets are resized to 256×256. Depth stream, the gray-scale input image is converted into three channels of color image using color mapping [184] technique. All the implementations are performed on an NVIDIA 1080Ti GPU accelerates. The training process of the proposed network is performed in an end-to-end manner, using a widely used Adam optimizer [185] with initial learning rate $\alpha = 0.0003$, and $\beta = (0.5, 0.999)$ and

weight decay is 0.001. The SGD optimizer is used to optimize network parameters with an initial learning rate of 0.025, the momentum of 0.9, and weight decay of 0.0003. The approximate computational testing time of an RGBD image pair is 0.021s. The proposed network $SL-Net$ has a self-Learning, based Dense Decoder to explore optimal cross-complementary fusion. The composite backbone network is configured with five convolution layers along with MSDW(in color stream) separately and rest pooling and other layers have been ignored. The size of convolution operations in all CF, which is part of SDD modules is (3×3) and filter size is $k = 64$. Maintaining each feature's resolution at SDD high-level semantic features is one time up-sampled at each SDD module. A simple bilinear interpolation is used in up-sampling operations. In the last stage of the decoder $SDD1$ module, the resolution of an output image is the same 256×256.

6.3.4 Comparison and Result Analysis

The proposed method, $SL-Net$, is compared with fifteen recent, top-performing, and closely related state-of-the-art methods with four recent evaluation metrics. The following state-of-art methods CAS-GNN [142], DANet [127], PGAR [129], cmMS [134], CoNet [196], UCNET [136], JL-DCF [131], S2NET [137], D3NET [130], CPFP [64], TANet [180], AFNet [133], CTMF [126], PCFNet [116], DF [125], are compared on seven publicly available datasets. These deep learning-based methods are recent, efficient, and closely related to the proposed method. We execute their source code with the same default settings and other related parameters as suggested

by corresponding authors for fair comparisons. The publicly available saliency maps for methods as mentioned above is used for result analysis. The result analysis through visual and quantitative comparison is demonstrated as follows.

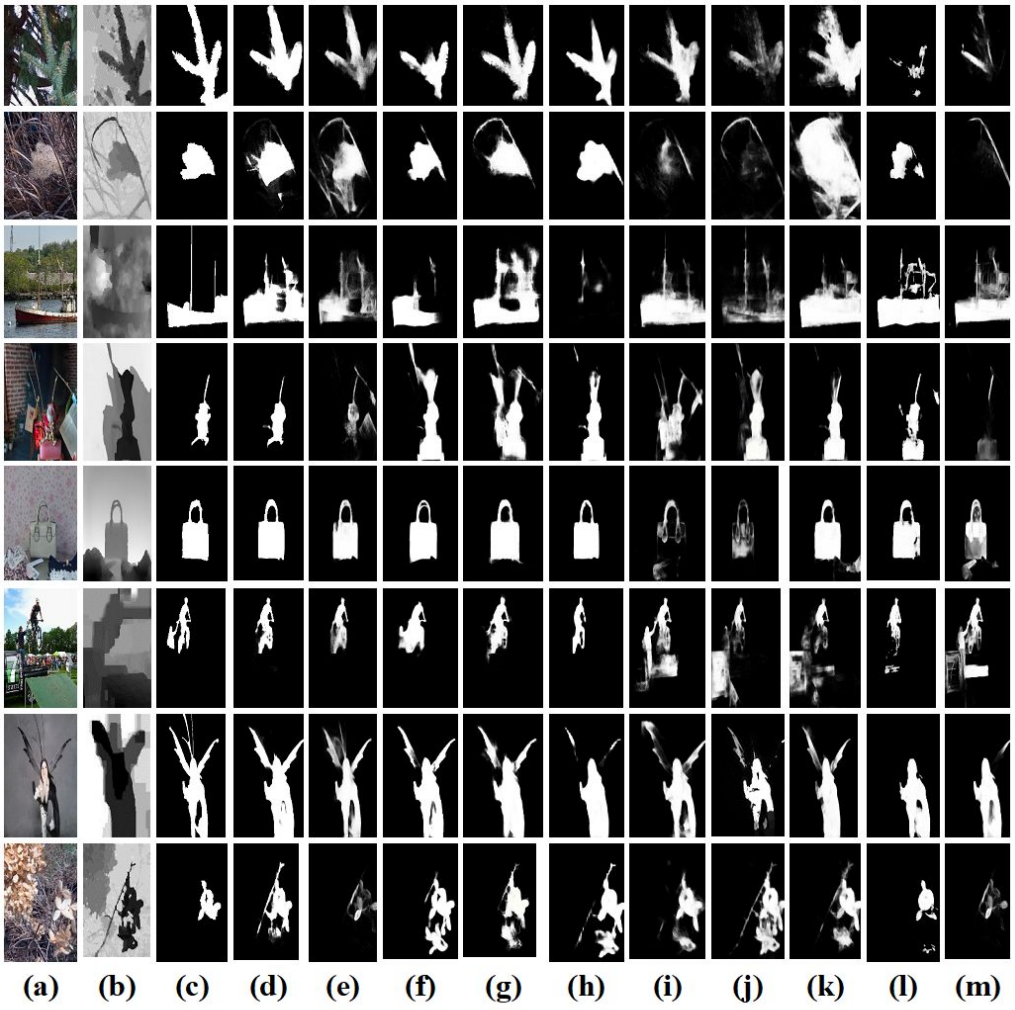

FIGURE 6.5: Visual Demonstration of proposed method SL-Net with other closely and recent State-of-art-methods.

Visual Comparison: The visual assessment is shown in Fig. 6.5 intuitively demonstrates the noticeable performance of proposed methods in complex and clutter backgrounds. As per observation from Fig. 6.5 the proposed model show better

saliency with other methods in the complex and challenging scenario, such as crowd-based objects (i.e., the sixth image), complex object and background(i.e., the last image), low quality depth map (i.e., the third image). Although, the proposed method not only detects the salient object but also preserves the object border, internal salient regions consistency, and structural integrity. For example, in 2^{nd} and 8^{th} images, most methods produce multiple non-salient regions as salient objects. At the same time, our method shows better saliency because it preserves structural integrity and internal consistency. It is achieved by enhanced encoded features guided by the Mutual Attention Based Distinguished Window-MADW. The 3^{th}, 6^{th} and 7^{th} images have a inferior quality depth map, in which our method predicts the exact salient object while other methods fail. In addition, the 1^{st}, 4^{st} and 8^{th} images have confusing backgrounds, and objects have similar characteristics to the backgrounds. In these situations, our model predicts exact salient objects with sharper object borders because of the proposed self-Learning-based SDD model, which utilizes deep global localized, enhanced encoded features and high-level semantic features.

TABLE 6.1: The quantitative comparison of proposed framework on seven benchmark RGBD datasets with four recent evaluation parameters.

Data-Set	Metric	OUR	CAS-GNN [142]	DANet [127]	PGAR [129]	cmMS [134]	CoNet [196]	UCNET [136]	JLDCF [131]	S2Net [137]	D3Net [130]	CPFP [64]	TANet [180]	PCFNET [116]	CTMF [126]	AFNet [133]	DF [125]
NJUD [110]	$F_\beta \uparrow$	0.916	0.893	0.877	0.883	0.914	0.872	0.895	0.903	0.889	0.900	0.877	0.874	0.872	0.845	0.775	0.804
	$S_\alpha \uparrow$	0.919	0.911	0.897	0.909	0.900	0.894	0.897	0.903	0.894	0.900	0.879	0.878	0.877	0.849	0.772	0.763
	$E_\psi \uparrow$	0.935	0.922	0.926	0.916	0.914	0.912	0.936	0.944	0.930	0.950	0.926	0.925	0.924	0.913	0.853	0.864
	$MAE \downarrow$	0.037	0.036	0.046	0.042	0.044	0.047	0.043	0.043	0.053	0.041	0.053	0.060	0.059	0.085	0.100	0.141
NLPR [98]	$F_\beta \uparrow$	0.917	0.888	0.865	0.885	0.913	0.850	0.886	0.875	0.902	0.897	0.867	0.863	0.841	0.825	0.771	0.778
	$S_\alpha \uparrow$	0.932	0.919	0.908	0.930	0.899	0.907	0.920	0.925	0.915	0.912	0.888	0.886	0.874	0.860	0.799	0.802
	$E_\psi \uparrow$	0.955	0.951	0.945	0.955	0.945	0.936	0.951	0.952	0.953	0.953	0.932	0.941	0.925	0.929	0.879	0.880
	$MAE \downarrow$	0.024	0.025	0.031	0.024	0.027	0.031	0.025	0.022	0.030	0.030	0.036	0.041	0.044	0.056	0.058	0.085
STEREO [97]	$F_\beta \uparrow$	0.914	0.876	0.868	0.880	0.908	0.885	0.885	0.869	0.882	0.891	0.871	0.861	0.860	0.831	0.823	0.757
	$S_\alpha \uparrow$	0.911	0.899	0.901	0.907	0.889	0.908	0.903	0.905	0.890	0.899	0.879	0.874	0.875	0.848	0.825	0.757
	$E_\psi \uparrow$	0.935	0.929	0.921	0.919	0.922	0.922	0.919	0.946	0.932	0.938	0.925	0.923	0.925	0.912	0.887	0.847
	$MAE \downarrow$	0.038	0.039	0.043	0.041	0.042	0.041	0.039	0.042	0.051	0.046	0.051	0.060	0.064	0.086	0.075	0.141
SSD [163]	$F_\beta \uparrow$	0.880	0.840	0.831	-	0.865	0.806	-	-	0.848	0.834	0.766	0.810	0.807	0.729	0.687	0.735
	$S_\alpha \uparrow$	0.884	0.872	0.869	-	0.874	0.853	-	-	0.868	0.857	0.807	0.839	0.841	0.776	0.711	0.747
	$E_\psi \uparrow$	0.920	0.915	0.909	-	0.911	0.896	-	-	0.909	0.910	0.852	0.897	0.894	0.865	0.807	0.828
	$MAE \downarrow$	0.044	0.047	0.050	-	0.052	0.059	-	-	0.052	0.058	0.082	0.063	0.062	0.099	0.118	0.142
DES [100]	$F_\beta \uparrow$	0.905	0.885	0.891	0.880	-	0.861	0.905	0.885	0.935	0.885	0.846	0.827	0.804	0.844	0.728	0.766
	$S_\alpha \uparrow$	0.911	0.898	0.905	0.913	-	0.910	0.934	0.929	0.973	0.898	0.872	0.858	0.842	0.863	0.770	0.752
	$E_\psi \uparrow$	0.966	0.943	0.961	0.939	-	0.945	0.967	0.965	0.961	0.916	0.923	0.910	0.893	0.932	0.881	0.870
	$MAE \downarrow$	0.020	0.026	0.028	0.026	-	0.027	0.019	0.022	0.021	0.031	0.038	0.046	0.049	0.055	0.068	0.093
DUTO-RGBD [43]	$F_\beta \uparrow$	0.919	0.912	0.884	0.914	0.932	0.908	-	0.882	0.901	0.867	0.795	0.790	0.771	0.823	0.659	0.744
	$S_\alpha \uparrow$	0.911	0.891	0.889	0.920	0.885	0.918	-	0.900	0.903	0.882	0.818	0.808	0.801	0.831	0.762	0.705
	$E_\psi \uparrow$	0.933	0.932	0.929	0.944	0.940	0.941	-	0.931	0.937	0.889	0.859	0.861	0.856	0.899	0.796	0.823
	$MAE \downarrow$	0.040	0.043	0.047	0.035	0.036	0.034	-	0.049	0.043	0.061	0.076	0.093	0.100	0.097	0.122	0.145
LFSD [100]	$F_\beta \uparrow$	0.881	0.832	0.822	0.852	0.888	0.848	0.859	0.854	0.835	0.810	0.826	0.796	0.775	0.787	0.744	0.813
	$S_\alpha \uparrow$	0.872	0.846	0.849	0.853	0.846	0.862	0.865	0.862	0.837	0.825	0.828	0.801	0.786	0.788	0.738	0.783
	$E_\psi \uparrow$	0.905	0.877	0.879	0.889	0.891	0.897	0.897	0.882	0.873	0.862	0.872	0.847	0.827	0.857	0.815	0.857
	$MAE \downarrow$	0.062	0.074	0.079	0.074	0.72	0.071	0.066	0.070	0.094	0.095	0.088	0.111	0.119	0.127	0.133	0.145

TABLE 6.2: The ablation study of each component in the $SL-Net$

Setting				DUT-RGBD [43]				NJU2K [110]				LFSD [100]			
BASE MODEL	SDD	MADW	$f_g+f_h+f_c$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\psi\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\psi\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\psi\uparrow$	$MAE\downarrow$
✓				0.6590	0.7435	0.7963	0.1221	0.7759	0.7724	0.8053	0.1003	0.7445	0.7385	0.7990	0.1453
✓	✓			0.7060	0.7665	0.8203	0.0899	0.7965	0.7975	0.8105	0.0832	0.7775	0.7905	0.8442	0.1012
✓		✓		0.8221	0.8392	0.8235	0.0812	0.8395	0.8325	0.8425	0.0772	0.8235	0.8352	0.8610	0.0962
✓	✓	✓		0.8772	0.8809	0.8931	0.0531	0.8892	0.8785	0.8890	0.0560	0.8566	0.8490	0.8850	0.0710
✓	✓	✓	✓	0.9192	0.9112	0.9334	0.0405	0.9066	0.9152	0.9296	0.0381	0.8717	0.8625	0.9054	0.0622

Quantitative Comparison: The quantitative analysis from Table 6.1 objectively illustrates that the proposed model achieves remarkable improvements on all datasets. The improvements are visible through S-measure, E-measure, and F-measure while declining in MAE significantly. The improvements in the proposed model are shown here with recent benchmarks and top-performing methods- CAS-GNN [142], DANet [127], PGAR [129], cmMS [134]. These improvements have been achieved through three-level feature enhancements by proposed attention maps, Mutual Attention Based Distinguished Window-MADW, and SDD model. The quantitative analysis validates the effectiveness of the proposed attention model, which demonstrates the capability of generalization..

6.3.5 Ablation Analysis

Extensive experiments are performed for ablation analysis to investigate the contributions of each component in performance improvements. The mutual attention-based distinguished window-MADW produces enhanced encoded, deep global localized, and high-level semantic features. The effectiveness of Self-Learning based

Dense Decoder-SDD is also analyzed to show the importance of SDD in the cross-complementary fusion of the above three features. In order to validate the effectiveness of the proposed MADW and SDD module, we perform a series of experiments using four evaluating parameters with a defined BASE MODEL. This base model has simple VGG layers without MADW and a simple fusion of side-outputs of each layer without using SDD and deep global localized features, which is similar to AFNet [133]. This strategy shows the contributions of each component. The validation of the effectiveness of each component is analyzed as follows.

6.3.5.1 Effectiveness of Mutual attention based distinguished window-MADW

To verify the effectiveness of the proposed Attention map, MADW, which applies before each layer of VGG to improve the encoded saliency features and minimize the non-salient regions. Deep global localized features are also produced by MADW, which guide the fusion process in each stage. The improvements of using MADW and three enhanced features, f_c, f_g, f_c, are clearly visible in Table 6.2. The MADW improved the BASE model with a large margin, which shows the effectiveness of the proposed attention map MADW. The addition of f_c, f_g, f_c of Table 6.2 shows noticeable improvements in all parameters because of spatial, channel-wise, mutual, and feature level attention mechanisms. It minimizes the non-salient regions and improves the internal consistency of salient regions. The successive contributions in saliency computations are shown in Table 6.2, which validates the effectiveness of

each component of SL-Net on complex RGBD-Datasets. The visual contributions of each step are shown in Fig. 6.6. The purified deep global localized features and high-level semantic features are the same and shown in Fig. 6.6.

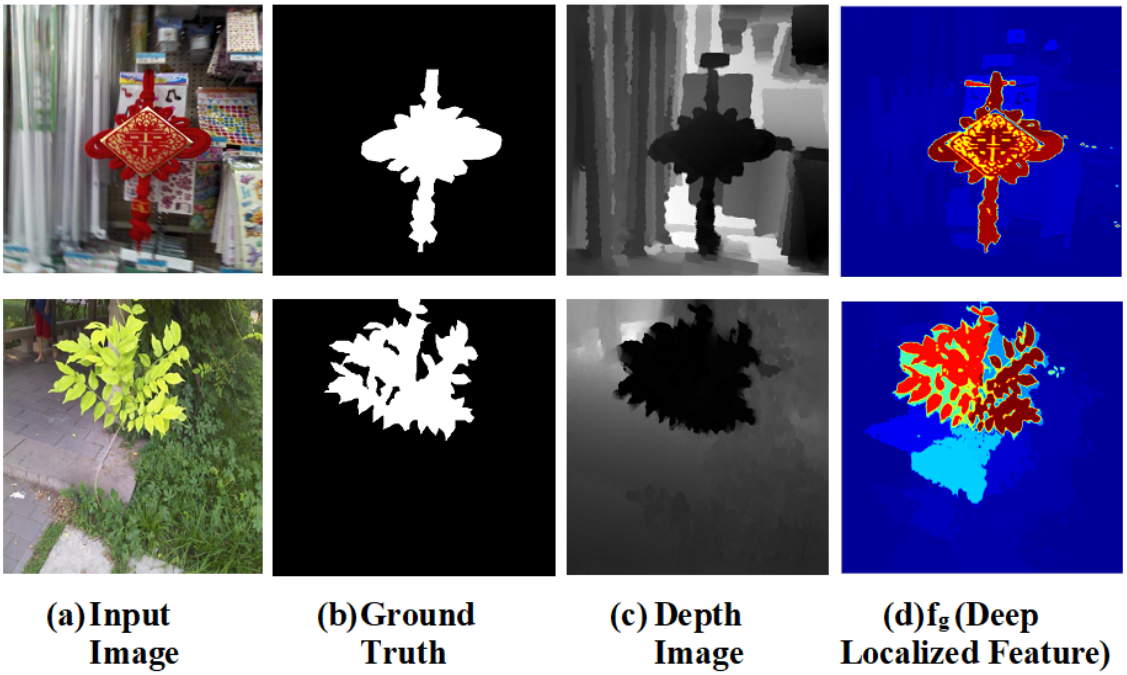

FIGURE 6.6: Visual illustration of the global localized features.

6.3.5.2 Effectiveness of SDD and Fusion model

The effectiveness of the proposed SDD model is compared with the BASE MODEL. The results of the SDD fusion process are shown in Table 6.2, which demonstrate the improvements with the components mentioned above. The SDD module used deep global localized features as a reference plane to distinguish the salient and non-salient regions for saliency enhancement and discrepancy minimization, respectively. The validation of the proposed Self-Learning based SDD, which is defined in Eq. 6.5

TABLE 6.3: Validation of stage-wise improvements using proposed Mutual Attention based Distingushed Window-MADW

Stage No.	DUT-RGBD [43]				NJU2K [110]				LFSD [100]			
	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\psi \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\psi \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\psi \uparrow$	$MAE \downarrow$
D4	0.8901	0.8905	0.9004	0.0449	0.8893	0.8944	0.9305	0.0544	0.8367	0.8376	0.8738	0.0942
D3+D4	0.8905	0.8907	0.9070	0.0483	0.8899	0.8949	0.9315	0.0501	0.8388	0.8389	0.8809	0.0890
D2+D3+D4	0.9022	0.9034	0.9156	0.0442	0.9054	0.9078	0.9148	0.0448	0.8567	0.8570	0.8906	0.0719
D1+D2+D3+D4	0.9192	0.9112	0.9334	0.0405	0.9166	0.9192	0.9296	0.0371	0.8817	0.8725	0.9054	0.0622

is also compared with the basic fusion model and shown in Table 6.4 . In Table 6.4, basic fusion model for cross-complementary features is validate using element-wise addition, multiplication and concatenation operation. The result analysis also includes the JL-DCF fusion model based on joint learning and densely-cooperative fusion. The results are computed using 256×256 images. The significant improvements in the results by using SDD module is shown in Table 6.2 and Table 6.4, with all parameters which validate the effectiveness of SDD module to predict exact salient object detection.

6.3.5.3 Effectiveness of Proposed Composite model

Finally, the validation of the proposed structure is illustrated in Table 6.3 and Fig. 6.7. The ablation analysis checks the effectiveness of adding MADW at multiple stages in the RGB encoder stream. D_i (where $i \in \{4, 3, 2, 1\}$) is used to denote the stage number. The stage-wise improvements from Table 6.3 show the effectiveness of adopting composite networks in performance improvements and the relevancy of adding MADW before each stage in the RGB stream. The successive improvements shown in Table 6.2, and Table 6.3 validate the proposed composite networks.

TABLE 6.4: The ablation analysis of Cross-complementary fusion using *SDD* module and basic fusion strategy.

Fusion Model	DUT-RGBD [43]				NJU2K [110]				LFSD [100]			
	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\psi \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\psi \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\psi \uparrow$	$MAE \downarrow$
ADD	0.8789	0.8809	0.9000	0.0469	0.8904	0.8990	0.9090	0.0449	0.8800	0.8600	0.8811	0.0755
MUL	0.8810	0.8911	0.9010	0.0464	0.8999	0.9010	0.9105	0.0440	0.8609	0.8605	0.8821	0.0750
Cat	0.8821	0.9024	0.9105	0.0438	0.9054	0.9022	0.9148	0.0438	0.8617	0.8610	0.8861	0.0736
JL-DCF	0.8822	0.9005	0.9314	0.0495	0.9036	0.9032	0.9446	0.0435	0.8546	0.8623	0.8825	0.0703
SDD	0.9192	0.9112	0.9334	0.0405	0.9166	0.9192	0.9296	0.0371	0.8817	0.8725	0.9054	0.0622

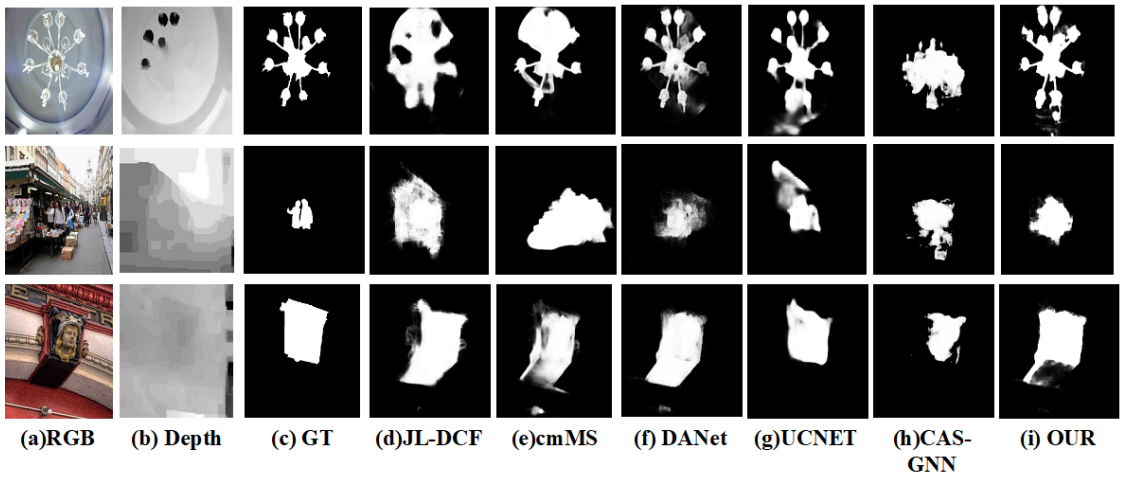

FIGURE 6.7: Visual illustration of failure case in incomplete depth maps and complex and cluttered background.

6.3.6 Discussion of Failure case and limitations

The drastic improvements in the recent SODs model mentioned in Section **??** have also failed in very complex and challenging situations. The primary issues of producing incomplete and inaccurate salient objects are low and incomplete depth maps, complex and clutter backgrounds, and challenging structures. The visual demonstration of the above challenging situations is shown in Fig. 6.7. In this Fig. 6.7, our proposed model accurately localizes the salient object and produces better saliency than other recent and top-performing methods, even in failure cases also.

TABLE 6.5: The comparison of own proposed deep learning based models.

Fusion Model	DUT-RGBD [43]				NJU2K [110]				LFSD [100]			
	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\psi \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\psi \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\psi \uparrow$	$MAE \downarrow$
CCL-Net	0.899	0.901	0.912	0.047	0.910	0.910	0.918	0.040	0.868	0.867	0.880	0.070
HFL-Net	0.890	0.901	0.910	0.049	0.910	0.908	0.910	0.041	0.850	0.860	0.861	0.074
DGMA-Net	0.918	0.927	0.909	0.044	0.913	0.917	0.914	0.041	0.869	0.865	0.886	0.070
CSA-Net	0.921	0.921	0.920	0.039	0.912	0.911	0.934	0.040	0.871	0.862	0.905	0.062
SL-Net	0.919	0.911	0.933	0.040	0.916	0.919	0.929	0.037	0.881	0.872	0.905	0.062

6.3.7 Comparison of own proposed methods

The comparative study of our own proposed deep learning-based models is shown in table 6.5 using four evaluating parameters and three datasets. The $CCL-Net$ proposed to efficiently exploit the cross-complementary features.$HFL-Net$ is used to utilize the cross, non, and intra-complementary features.$DGMA-Net$ effectively used depth-guided mutual attention maps to improve the deep localized features. $CSA-Net$ proposed two-stage additive Cross-complementary Self Attention maps based on a Non-Local network to exploit long-range contextual dependency. Finally, $SL-Net$ proposed a composite backbone by proposing an attention,$MADW$, based encoder to enhance the encoded features. The result analysis from Table 6.5 demonstrates the effectiveness of the proposed model in performance improvements. These models aim to address most of the mentioned research gaps and improve performance.

6.4 Conclusion

This study proposes a novel mutual attention-based distinguish window, MADW, to enhance encoded features. A multi-stage mutual attention map-based encoder could be integrated to produce enhanced encoded, and deep global localized feature maps to address the challenge of a complex image, low depth map, and complex backgrounds. It is proposed to obtain better feature representation. This attention map is formulated with spatial, channel-wise, mutual, and feature-level attention mechanisms. Furthermore, the deep global localized feature map is the fusion process in the decoder to localize the salient object using the proposed attention map MADW. It is used as a reference plane to distinguish the salient and non-salient regions during the fusion process in SDD. Thus, the boundaries of detected objects could be better preserved. A systematic ablation study is conducted on publicly available datasets, and the experimental results have verified the effectiveness of each component of the proposed $SL - Net$. Moreover, the results of comparative experiments have demonstrated that for salient object detection in complex and clutter images, the proposed $SL - Net$ is better than other state-of-the-art methods. However, some drawbacks still exist for the proposed network. For one thing, the feature fusion strategy for CNN and SDD needs to be further considered to achieve better saliency accuracy. In future work, an efficient feature fusion module should be designed to improve the fusion process of the feature maps extracted by CNN and guided by an attention map.