Chapter 5

CSA-Net:Deep Cross Complementary Self Attention and Modality-Specific Preservation for Saliency Detection

The multi-modality or multi-stream-based convolution neural network is the recent trend in saliency computation, which is receiving tremendous research interest. The previous models used modality-based independent fusion or cross-modality-based complementary fusion to find saliency that leads to incurring inconsistency or distribution loss of salient points and regions. Most existing models did not effectively utilize accurate localization of high-level semantic and contextual features.

The proposed model collectively uses the above two methods and a precise deep localization model to target the above-mentioned challenges. Specifically, CSA-Net comprises four essential features: non-complementary, cross-complementary, intra-complementary, and deep localized improved high-level features. The designed 2×3 encoder and decoder streams produce these essential features and assure modality-specific saliency preservation. The cross and intra- complementary fusion are deeply guided by proposed novel, cross-complementary self-attention to produce fused saliency. The attention map is computed by two-stage additive fusion based on a Non-Local network. A novel, Optimal Selective Saliency, has been proposed to find two similar saliencies among three steam-wise saliencies. The details introduction of proposed models and releted method is discussed in next section 5.1.

5.1 Introduction

Visual salient object detection generates conspicuous and prominent objects in the complex and clutter background image, which is similarly identified by the human visual system. With the emergence of depth-sensing technologies and real-time applications, many researchers explore RGBD saliency in 3D space. The depth information is the most vital parameter in complex images where only color modality is insufficient to identify saliency. The saliency computations are commonly used as integral part in many vision-related applications, such as semantic segmentation

[177], object classification [18], video summarization and segmentation [20], [178], and content-based image editing [179].

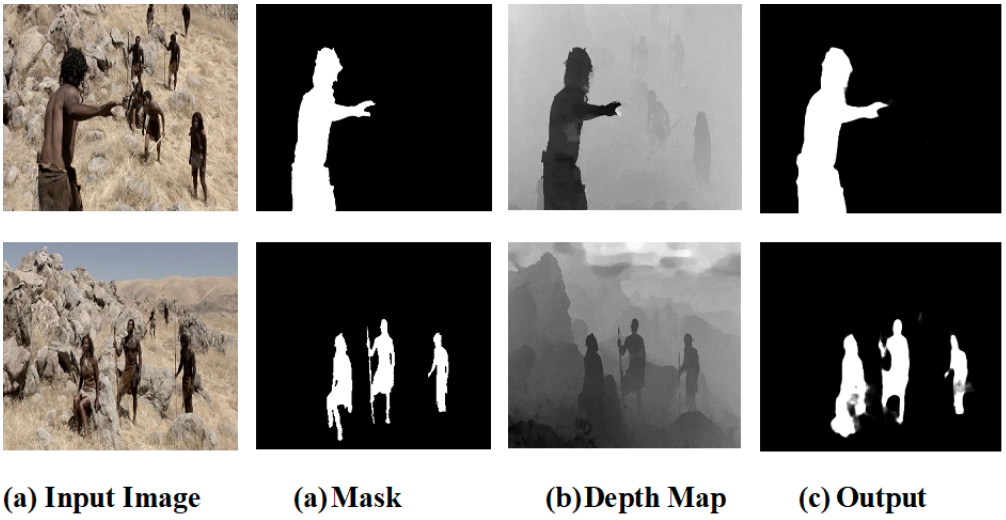

FIGURE 5.1: The 2×3 encoder and decoder streams utilize, deeply guided proposed attention map $CSA - Net$ to find the exact salient object.

In particular, we define three objectives to mitigate the challenges in the existing models of RGBD saliency computation: (1) Interior saliency enhancement (2) Exterior saliency minimization, and (3) Border regions saliency preservation. Afterward, we discuss our motivation to achieve these objectives by 2×3 encoder and decoder network. Interior saliency enhancement is defined as "increasing the saliency for non-identified regions in the salient regions. Similarly, exterior saliency reduction is defined " as minimizing the discrepancy of non-salient regions in the background. Furthermore, border region saliency preservation aims to increase the saliency in the border region of the salient object. The primary motivation of the three-stream decoder network is to explore all possible features to predict the salient objects correctly. Two networks, $(stream_1)$ and $(stream_3)$ based on color and

depth modalities, have a target to compute saliency to preserve modality-dependent features. It is defined as non-complementary features in this model. These networks target non-complementary features like purely color contrast-based features, depth contrast-based features, and regional color features. These networks minimize discrepancies in exterior, structural and spatial regions. Common modality-based ($stream_2$) has the target to explore cross and intra- complementary features. The intra-complementary features are defined as ***essential features between deep to shallow levels in multi-resolution space***. It is guided by proposed non Local-Network-based Attention maps. These features are high-level contextual, semantic, structural, spatial, global, and local features to produce salient objects similar to the ground truth level. These cross complementary features target the interior as well as border region saliency enhancements.

The learning capabilities of discriminative features in CNN are the most important characteristics, and it is the basis for designing multi-stream models. There are various two-stream based, improved models [126], [133], [137] have been proposed and improved the performance. Similarly, the most recent, multi-stream network, [180], [130], [47] have been utilized in complex scenarios to achieve the next level benchmark. These models have more powerful feature extraction capabilities. These models have produced high-level semantic features independently from two streams and fused them using the middle or late fusion strategy. Some recent models [180], [130] have used multi-stream models to exploit high-level correlation among cross modalities and various fusion strategies. These contemporary architectures improved

the performance at another level. However, these recent models face the following challenges: 1) discrimination of complementary and non-complementary features in successive integration in middle-level strategy 2) Most of the recent methods like [180], [130], [47], [137], [131], and [181] are either only exploit the complementary features or non-complementary feature but do not utilised both features. It leads to border regions and interior saliency discrepancies. 3) The late fusion strategy model remains inefficient for fusing the final saliency.

The proposed model CSA-Net uses a middle and late fusion strategy to address the limitations mentioned above. This joint fusion strategy integrates the low-resolution to the high-resolution, intra-complementary features, guided by the cross-complementary self-attention maps. Further, it finds a correlation for fusion between cross-complementary features. The non-complementary features are separately fused in decoders to preserve the modality-dependent saliency features. The late strategy uses selective saliency to fuse the stream-wise final saliency based on the similarities among the three saliencies. With the specially designed fusion strategy and network structures, this model can predict saliency maps similar to human perception, illustrated in Fig. 5.1. The proposed model predicts complete objects with exact object boundary, minimized background, and smooth saliency even in complex, clutter, and challenging background with illumination disturbance, low depth, object shadow, having multiple objects. The main contributions of the proposed methods are summarized as follows:

- In this model, the 2×3 encoder and decoder network has been proposed to produce non-complementary and cross-complementary features. Two independent streams ($stream_1$ and $stream_3$) are dedicated for color and depth. These streams have a target to produce modality-dependent saliency, and another decoder has a target to produce fused saliency, based on the cross and intra-complementary features.

- We design a two-stage additive, *cross complimentary self attention map* first time as per our knowledge, to deeply supervise the cross and intra-complementary fusion, simultaneously preserving the modality-based saliency.

- we design a novel *Optimal Selective Saliency*, late fusion model, to combine the two optimal saliencies among three-stream-wise saliencies.

- The comprehensive experiment on seven complex datasets and using recent evaluation matrix, demonstrate remarkable improvement with state-of-the-art methods.

The proposed CSA-Net targets to produce non-complementary, cross-complementary, intra-complementary, and deep localized improved high-level features. The 2×3 encoder and decoder architecture has been designed for the above features. Non-complementary features are essential in color and regional, while the cross and intra-complementary features established a substantial correlation to differentiate in complex and cluttered backgrounds. The complementary fusion begins with highly improved and accurate deep localized features produced by the proposed deep CSA

attention map. The innovative design and deep-CSA module improved the performance remarkably, as described and compared with state-of-the-art methods.

The rest of the chapter is organized in the following sections. The Section 5.2 describes the proposed method $CSA-Net$ in detail. Section 5.3 discusses the Experiment Set-Up, datasets, Network parameters in details. Section 5.4 demonstrates the Performance of $CSA-Net$ with other state-of-the-art saliency detection methods. Section 5.5 describes the conclusion and the future scope of improvements in this model.

5.2 The Proposed Model

Motivation: The shallow level RGB features have ample information to distinguish objects by regional structural and spatial attributes. At the same time, deep-level features have essential information to localize the object correctly. The previous models fused all these features using element-wise multiplications additions or concatenation. So these features have not been utilized for their own specific relevance. The limitations mentioned above are our biggest motivations to design 2×3 architecture to target efficient utilization of each specific feature for a specific purpose.

The proposed model has 2×3 encoder and decoder stream, based on RGB ($Stream_1$), Depth ($Stream_3$), and Cross-modality ($Stream_2$). The RGB and Depth have encoders as well as decoder streams. The encoder is based on the most preferred

backbone network, VGG-16 [182]. The unimodal non-complementary features are produced by RGB ($Stream_1$) and Depth stream ($Stream_3$). It is used to produce and preserve the modality-dependent feature through the Modality-specific saliency fusion model(MSF). The fused stream $Stream_2$ is a decoder stream. It uses the side output(raw saliencies) from each stream. The proposed attention map supervises the cross-complementary fusion in $Stream_2$ from the deepest level. It is modified the NL [183] module to enhance the cross-complementary fusion, which is shown in Fig. 5.2. It fuses the deep cross-complementary and intra-complementary features to produce long-range global contextual dependency. This attention map is a beacon for the cross-complementary fusion process in $Stream_2$ from the deepest and coarsest features. These three streams produce stream-wise saliencies. These stream-wise saliencies are collectively utilized in the proposed late fusion strategy, *Optimal Selective Saliency*, to predict accurate and robust salient objects.

The 2×3 encoder and decoder network architecture of the proposed model and learning module is visualized in Fig. 5.3. The input image in each stream should be uniform and have the same resolution. In the color stream, the input image has three channels of $224 * 224 * 3$. While in the depth stream, the color mapping [184] technique is used to convert the grayscale depth map into three channels.

5.2.1 Modality-specific Saliency Fusion Model(MSF)

The main objective of this model is to preserve the modality dependant saliency features. It is implemented using separate depth and color-based encoder and decoder via $stream_1$ and $stream_3$, respectively. The non-complementary features generated by the $VGG-16$ in the color and depth stream are produced separately and simultaneously, demonstrated in $stream_1$ and $stream_3$, respectively in Fig. 6.3. Let us use five convolution blocks $Conv1\_2$, $Conv2\_2$, $Conv3\_3$, $Conv4\_3$, and $Conv5\_3$, in backbone features generation. The outputs produced by these blocks are denoted as C_i. Their corresponding saliency feature is S_i. These stage-wise side saliencies(raw saliencies) [124] are utilized into $Stream_2$ for cross complementary raw saliencies fusion. We add a convolution block and up-sampling layer with resolution $224 * 224$ at the end of RGB and Depth stream. These features are fused the feature S_i with the deep fused feature $\Im_{i+1}$ at i and $i+1$ scale respectively. Finally, these two streams produce their saliency map S^{rgb} and S^{depth} separately and simultaneously. These streams have modality-dependent fusion to preserve the modality-dependent saliency. Let us define modalities $m = (RGB, Depth)$ in following Eq. 5.1. The formulation of features generation and saliency prediction is defined in Eq. 5.1as follows:

$$
\begin{aligned}
S_i &= \psi\left(\varphi\left(\varphi\left(C_i\right)\right)\right) \qquad 1 \leq i \leq 5 \\
\Im_i &= \begin{cases} \varphi(\|\Im_{i+1}, s_i\|) & 1 \leq i < 5 \\ S_i & i = 5 \end{cases} \\
S^m &= Sig(k_s * \Im_i + bias)
\end{aligned}
\tag{5.1}
$$

$\psi(...)$ denotes the up-sampling function to make raw saliency with the same resolution that uses bilinear interpolation. $\varphi(...)$ describes the operation of convolution with 64 channels. A non-linear activation function follows it. In this convolution operation the kennel size is 3×3 and with the stride size is 1. k_s is 1×1 kernel and $bias$ is bias parameter. $Sig(...)$ is the Sigmoid function, while $*$,and $\|...\|$representsthe convolution and channel-wise concatenation operations, respectively.

The $stream_1$ and $stream_3$ produce color and depth based saliency S^{rgb} and S^{depth} to prserve the specificity by the Eq. 5.1 respectively. In $stream_2$, network learns Cross complementary and Intra-complementary features and produced S^{fused} saliency guided by Cross-Complementary Self-Attention map. The S^{rgb}, S^{depth} and S^{fused} are used in late fusion strategy by proposed Optimal Selective Saliency Fusion model. These models are described below.

5.2.2 Cross -Complementary Self-Attention -CSA

The main objective of using this cross additive self mutual attention map is to start the fusion process between color and depth features to explore hierarchical, cross, and intra-complementary features. CSA module starts the fusion from deep and guides the multi-stage complementary fusion in $stream_2$. The detailed block diagram is shown in Fig.5.2.The proposed model of two-stage additive, ***Cross - complementary Self-Attention-CSA*** is based on the Non-local Network-NL

[183] module to compute the mutual self-attention map. This cross-modality self-attention is a beacon from deep to the shallow level complementary fusion process. Let us define the two modality dependent deep feature maps as X^i, where $i = (rgb, depth)$ with specified dimension $X \in \mathbb{R}^{h \times w \times k}$. Where *h,w, and k* are denoted as height, width, and channel number, respectively. The embedding process of the NL module converts the feature maps X^i into θ, ϕ and g space having $k1$ channels. The affinity or similarity matrices are defined in Eq. 5.2 as follows:

$$f^i(X^i) = \Theta^i(X^i)\Phi^i(X^i)^T \tag{5.2}$$

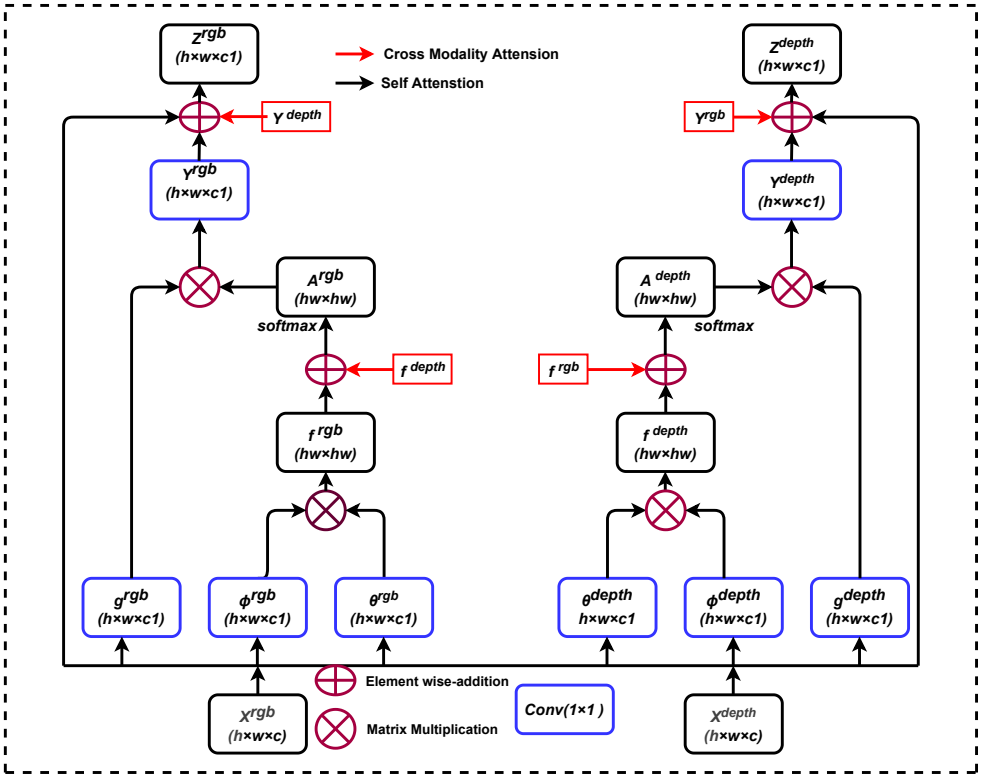

FIGURE 5.2: The proposed two-stage ***Cross -complementary Self-Attention-CSA*** model is based on the Non-Local Network.

where $\Theta(X) = XW_\Theta, \Phi(X) = XW_\Phi, g(X) = XW_g$ and $\Theta(X)$, $\Phi(X)$ and $g(X)$ are embedding weights. The embeddings are implemented with (1×1) convolution, which is shown in Fig. 5.2. The similarity or affinity function, $f(X) \in \mathbb{R}^{hw \times hw}$, is computed in its own space(modality). It is represented the similarity between i^{th} and j^{th} spatial location in feature space X. The cross-modality-based attention map A^f is computed by cross fusion of multi-modality affinity matrix by a simple element-wise matrix addition operation. It is definedin Eq. 5.3 as:

$$A^f(X^{rgb}, X^{depth}) = softmax(f^{rgb}(X^{rgb}) + f^{depth}(X^{depth})) \tag{5.3}$$

This attention map guides the long-range contextual dependency in two modalities for further cross-modality based fusion. It is computed in Eq. 5.4 as:

$$Y^i = A^i(X^i)g^i(X^i) \tag{5.4}$$

Where $i = (rgb, depth)$ and the Cross fused attention maps are defined for both modalities. Finally, the modality-wise Attention map is measured by the simple addition of residual signal of both modalities based on Y^i in original feature space X^i. At this stage, the NL module has been modified to improve the performance. The modification has been taken place to compute an additive cross attention map. It is calculated by the simple addition of stream-wise self-attentive maps. It is

defined in Eq. 5.5as follows:

$$Z^{rgb} = (Y^{rgb} + Y^{depth})W_Z^{rgb} + X^{rgb}$$
$$Z^{depth} = (Y^{depth} + Y^{rgb})W_Z^{depth} + X^{depth} \tag{5.5}$$

Where $W_z \in \mathbb{R}^{K1\times K}$ is the weight of the convolution layer((1×1)). it is used to project the attentive features back into their original feature space. Finally, the Cross Modality Attention maps Z^{rgb} and Z^{depth} fed into CFF_6(Eq. 5.6) to guide and propagate decoding of the cross and intra-complementary features.

5.2.3 Complementary Features Fusion Model

The Cross and non-complementary learned from cross modalities are necessary steps to predict the exact salient object through coarse localization by attention map. The coarse localization and guided fusion are achieved through CSA maps. The multi-resolution hierarchical features generated in color and depth channel is utilized in $stream_2$ for Cross -complementary Self-Attention-CSA. The process of $stream_2$ describes in the following two stages is as follows:

5.2.3.1 Cross-Complementary Features Fusion(CFF)

The CNN backbone network in color $stream_1$ and depth $stream_3$ produces saliency features like [124] (side-output). These saliencies have multi-resolutions and multi-channels (more channels in deep feature). The saliency features from depth and

color stream from each stage (total five sages and one for CSA output from $CFF1$ to $CFF6$) are fused. In this model, the varied resolution features are compressed into smaller (fixed size equal to k) and exact sizes. The output of CSA is fed into $CFF6$ to start the decoding process. The processed saliency features in RGB and Depth modality are denoted as Sf_{rgb}, Sf_{depth}, each with equal k channels. The output of the CFF module is defined in Eq. 5.6 as:

$$CFF^{k}\left(Sf_{rgb}, Sf_{depth}\right) = \left(Sf_{rgb}^{k} \otimes Sf_{depth}^{k}\right) \oplus \left(Sf_{rgb}^{k} \oplus Sf_{depth}^{k}\right) \tag{5.6}$$

In this cross-view 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 the saliency, respectively. The output of the CFF model from each stage is successively fused from deep coarse localization to the final saliency prediction deeply guided from the attention map. The fused features $CFF5$ to $CFF1$ are fed into dense decoder that have a dense connection to purify the saliency. The same computation of Eq. 5.6 is also performed in CFF6, where (Sf_{rgb}, Sf_{depth}) is replaced with $\left(Z^{rgb}, Z^{depth}\right)$ from Eq. 5.5 to start the guiding the fusion process. These dense connections are used to unify the multi scales features at multi-stages.

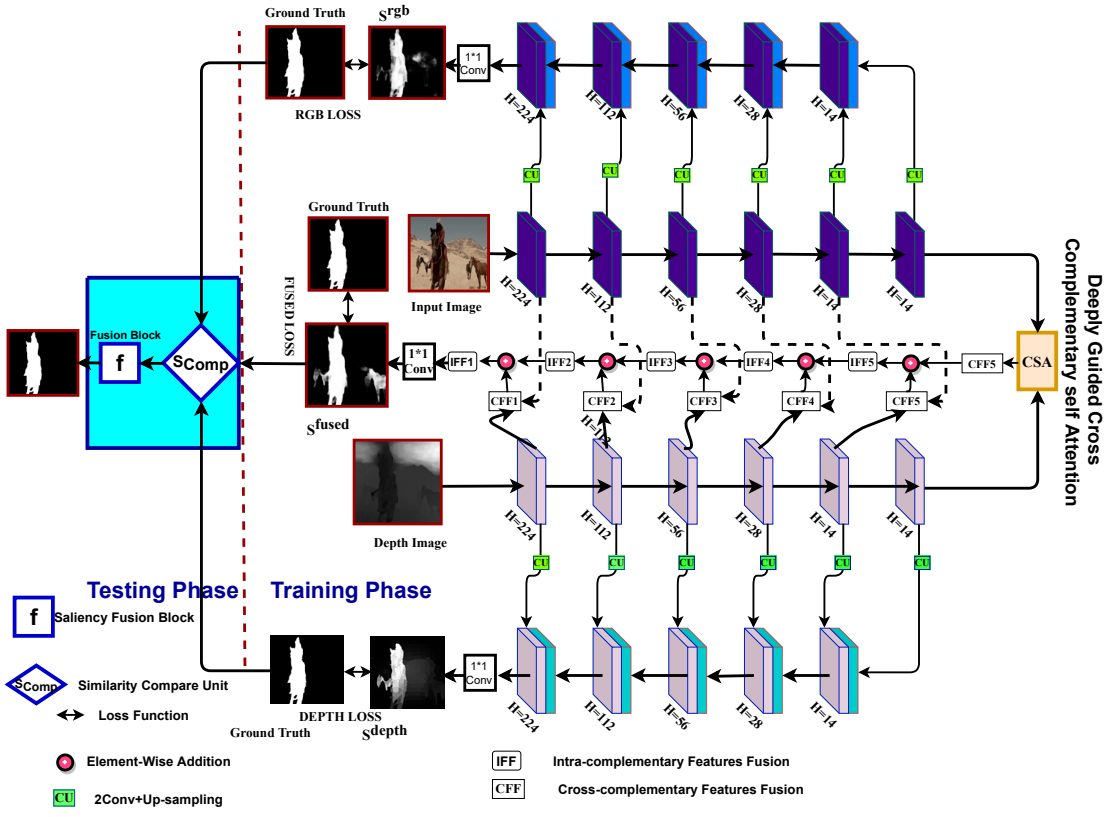

FIGURE 5.3: The proposed architecture of 2×3 encoder and decoder, $CSA-Net$ models.

5.2.3.2 Intra-Complementary Features Fusion(IFF)

In this fusion model, the feature maps of all preceding layers are used as inputs for each layer to find the intra-complimentary features. The successive saliency enhancement is achieved at each stage with the contributions from all previous stages and further propagate with the next layer. The resultant final enhanced fused saliency S^{fused} from $Stream_2$ features are used in the late fusion strategy. This fusion model enhances the interior saliency from deep level to output (image)

level. This model is mathematically formulated in Eq. 5.7 and shown in Fig. 5.4 with width (w), hight(h) and feature map I with channel k. It is defined as follows:

$$f_{w,h}(I,k) = f_{Con(1,1)}(I,k/4) \oplus f_{Con(3,3)}((f_{Con(1,1)}(I,k/2),k/4) \\ \oplus f_{Con(5,5)}((f_{Con(1,1)}(I,k/4),k/4) \oplus f_{Max-pool(3,3)}((f_{Con(1,1)}(I,k/4) \tag{5.7}$$

Where $Con(i,j)$ and $Max-pool(i,j)$ are convolution operation and max-pooling operation respectively, having stride 1 to maintain the spatial feature resolution. " $\oplus$" defines the simple concatenation. In this module, the multi-level convolutions filter with size values of (i,j) are 1×1, 3×3, 5×5 and max-pooling layer are used. This formulation is similar to the original Inception module [182], with one difference. In this method, the same channel number k is maintained in input and output.

5.2.4 Optimal Selective Saliency Fusion Model-OSS

To generate the optimal saliency map, we design a novel Selective Saliency method. This method is applicable only in Test Phase. This late fusion strategy selects the two optimal saliency maps among three-stream-wise saliencies maps. This fusion strategy eliminates the low-quality saliency map among three-stream saliency maps. Depth stream ($stream_3$) generates a low-quality saliency map in the low depth image [130], [97], [43] and color stream($stream_1$) saliency maps generate the low-quality saliency in images where a salient object cannot distinguish with color, regional,

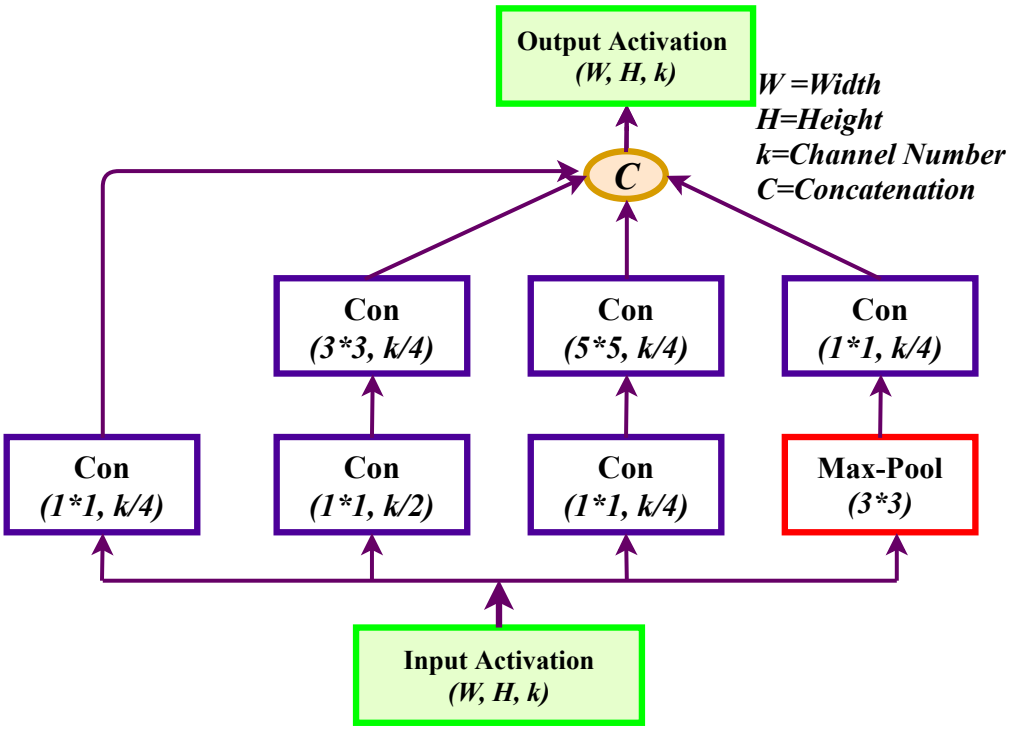

FIGURE 5.4: The block diagram of the Intra-complementary features Fusion model(IFF).

structural, spatial, and global features. The output of the last purified module in each stream having the finest features is fed to a $(1 \times 1, 1)$ convolution layer to generate the final streamwise S^{rgb}, S^{depth}, S^{fused} saliency maps. These saliency maps are further supervised with the same resolution ground truth map during the training process. For the notational convenience, we use a, b, c at the place of S^{rgb}, S^{fused}, and S^{depth} in Eq. 5.8 and in Fig. 5.3. Final saliency map S_{Comp} produces with following late and final fusion function f. It is defined as:

$$S_{Comp}(a,b,c) = \begin{cases} f(a,b) & if\ ((\vartheta(a,b) < \vartheta(b,c)) \wedge (\vartheta(a,b) < \vartheta(c,a))) \\ f(b,c) & elseif\ ((\vartheta(b,c) < \vartheta(a,b)) \wedge (\vartheta(b,c) < \vartheta(c,a))) \\ f(c,a) & otherwise \end{cases} \tag{5.8}$$

The final saliency map S is updated fusion of two optimal saliency maps (x, y) used in Eq. 5.9 among three stream-wise saliencies (a, b, c) from Eq.5.8 .

$$S = f(x,y) = \eta(x \oplus y) \otimes (x \otimes y) \tag{5.9}$$

Where $\vartheta(a, b)$ is defined as Minimum Absolute Error between two saliency maps and defined in Eq. 5.14. if $\vartheta(a, b)is = 0$ means both saliency maps are equal and similar. $\otimes$ and $\oplus$ are element-wise multiplication and element-wise addition respectively. η is used to normalized the values of saliency in range $(0, 1)$.

5.2.5 Loss Function

The total loss function is composed of stream-wise saliency loss in Eq. 5.10. The color, depth, and fused saliency loss functions are computed with their respective saliency map S^{rgb}, S^{depth}, S^{fused} and ground truth map Gt. The total loss function is defined as:

$$\xi_{total}(S, Gt) = \sum_{i \in (rgb, depth, fused)} \xi(S^i, Gt) \tag{5.10}$$

The loss function in Eq. 5.10 is defined as standard cross-entropy loss. It is widely used to compute the difference between the saliency map and the ground truth result. It is defined as follows:

$$\xi(S, Gt) = -\sum_{i} [Gt_i log(S_i) + (1 - Gt_i) log(1 - S_i)] \tag{5.11}$$

Where i is defined as pixel index in ground truth image.

5.3 Experimental Details:

The proposed model has 2×3 stream networks. Color ($stream_1$) and Depth ($stream_3$) are independent, parallel, and with similar network attributes, while fused, ($stream_2$) stream has combined features and characteristics from color and depth modality. The related parameters, Experimental setup, Data-set, implementations detail, and Network Architecture are described in detail in the following section.

5.3.1 Data-Set

The extensive experiments have been conducted on seven publicly available RGB-D benchmark datasets: STEREO-1000 [97], SSD-80 [163],NJUD-2000 [110], RGBD-135 [100], NLPR-1000 [98], DUTO-RGBD [144], and LFSD [100]. Most of the

contemporary models have been used the same data pattern similar to [133] for training and testing. The same data pattern for training and testing is used in this model for a fair comparison to the state-of-the-art methods. The training set contains 1400 images from the $NJUD-2000$ and 650 samples from $RGBD-1000$ dataset. The validation set contains 150 images in which 100 image pairs are from NJUD and 50 image pairs from $RGBD-135$. All images from $NJUD-2000$, $STEREO-1000$ and $SSD-80$ $RGBD-135$, and $NLPR-1000$ including training dataset is used in testing phase.

5.3.2 Implementation Details

The network architecture of the proposed model has 2×3 encoder and decoder networks. The computation of two-stream, $stream_1$ (Color) and $stream_3$ (Depth), are independent and parallel and have similar network characteristics. At the same time, $stream_2$ (fused) has shared the stage-wise features from the above two streams, which are fused using CSA attention maps from deep. The backbone network of this model is the VGG-16 model [182]. The existing pre-trained parameters of mostly preferred method DSS [124] are used to initialize the backbone network. In VGG-16, the backbone network is configured with five convolution layers $Conv1\_2$, $Conv2\_2$, $Conv3\_3$, $Conv4\_3$, and $Conv5\_3$ separately in-depth and color stream. These convolution blocks produce stage-wise saliency features (side outputs) in color and depth stream, which is also used to explore the complementary features in the fused stream. The additional two convolutional blocks have been added in each

stream as the deepest block. These blocks produce coarse features with the spatial size of 14×14, which is shown in Fig. 5.2. The size of convolution layers in all CFF modules is (3×3) and filter size is $k = 64$. We have used the stride of 1 in $Max - Pool_5$ to enhance the resolution of the coarsest feature maps. Up-sampling operation by multiple factors in $Stream2$ is performed in each stage to maintain the same resolution. The fusion is guided by the proposed Cross -Complementary Self-Attention -CSA module. A simple bilinear interpolation is used in up-sampling operations. In the final $IFF1$ module, the resolution of an output image is the same 224×224 as the input image. Up-sampling has been done in the outputs of each stage from $IFF5$ by a factor of $2, 4, 8$, and 16. This process converts the stage-wise saliency features into the same resolution. Data augmentation is a vital step in all models of saliency computation using deep learning. Due to the limited training data set, we have used horizontal flipping of training images to double the size of the training set.

5.3.3 Training Details

The training process of the proposed network is performed in an end-to-end manner. The Pytorch package has been used as implementing platform of the proposed model. In this training process, we used Adam optimizer [185] for optimizing the training process. An NVIDIA 1080Ti GPU is used to accelerate the training process with 40,000 iterations. This optimization has batch size 8, with an initial learning rate of 0.00001, a momentum of 0.9, and a weight decay of 0.0001. All input images

FIGURE 5.5: The Visual comparison of the proposed model with other State-of-the-art-methods.

are resized during the complete training or learning process and have equal size of 224×224. The total approximate training time is around 18 hours/16 hours, which contains 40 epochs in VGG-16 configuration.

5.3.4 Evaluation Metrics

In the comprehensive evaluation of proposed model with others State-of-the-art methods. We use recent evaluation metrics, which are used in recent comparison. These metrics are (1) S-Measure, (2) F-Measure, (3) Mean Absolute Error (MAE), and (4) E-measure(E_ψ). These metrics are defined as below:

5.3.4.1 S-Measure

S-measure [164] is a recent metric used to compare the structural similarity and dissimilarity, which is defined in Eq. 5.12. This metric computes region-aware S_{reg} and object-aware S_{obj} structural similarity between computed saliency map and ground truth map. This metric is defined as:

$$S_{measure} = \alpha S_{obj} + (1 - \alpha) S_{reg} \tag{5.12}$$

where $\alpha \in [0, 1]$ is set to 0.5.

5.3.4.2 F-Measure

The comprehensive evaluation of the proposed method is demonstrated with F-measure [158]. This metric is used to compute the relevancy of parameters like Precision and Recall. In this metric, Precision and Recall are combined as weighted harmonic, which is defined in Eq. 5.13 as follows:

$$F - Measure = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision \times Recall} \tag{5.13}$$

We use β^2 =0.3 F-Measure in Eq.5.13 for uniform comparison because the same value is preferred in the majority of saliency methods.

5.3.4.3 Mean Absolute Error(MAE)

MAE directly approximates the conformity between saliency maps and ground-truth maps. The mean absolute error(MAE) [62] is the preferred metric in successive steps validation and demonstration of successive steps contribution. MAE in Eq. 5.14 is defined in normalized range $[0, 1]$ saliency map S and the ground-truth binary mask Gt, which is defined as follows:

$$MAE = \frac{1}{n} \sum_{k \in n} (S(k) - Gt(k)) \tag{5.14}$$

5.3.4.4 E-Measure(E_ψ)

E-Measure is recently defined as Enhanced alignment measure, and the detail definition and formulation is available here [165]. This measure is based on cognitive vision studies. It uses image-level statistics(mean) and local level pixel matching information. To demonstrate a comprehensive evaluation, we use maximum value of E-measure.

TABLE 5.1: The quantitative comparison of proposed framework on seven benchmark RGBD datasets with four recent evaluation parameters.

Data-Set	Metric	OUR	JLDCF [131]	S2NET [137]	D3NET [130]	CPFP [64]	TANet [180]	PCF [116]	CTMF [126]	AFNet [133]	DF [125]	MDSF [105]	CDS [38]	DCME [109]	LBE [102]	DES [43]
STEREO [97]	$F_\beta^m \uparrow$	0.911	0.901	0.882	0.891	0.874	0.861	0.860	0.831	0.823	0.757	0.728	0.716	0.740	0.633	0.566
	$S_\alpha \uparrow$	0.909	0.905	0.890	0.899	0.879	0.871	0.875	0.848	0.825	0.757	0.719	0.711	0.731	0.660	0.582
	$E_\psi^m \uparrow$	0.945	0.946	0.932	0.938	0.925	0.923	0.925	0.912	0.887	0.847	0.809	0.851	0.819	0.787	0.670
	$MAE \downarrow$	0.039	0.042	0.051	0.046	0.051	0.060	0.064	0.086	0.075	0.141	0.176	0.122	0.118	0.250	0.193
SSD [163]	$F_\beta^m \uparrow$	0.920	-	0.818	0.834	0.766	0.810	0.807	0.729	0.687	0.735	0.793	0.768	0.711	0.619	0.581
	$S_\alpha \uparrow$	0.934	-	0.868	0.857	0.807	0.839	0.841	0.776	0.714	0.747	0.805	0.712	0.704	0.621	0.602
	$E_\psi^m \uparrow$	0.970	-	0.909	0.910	0.852	0.897	0.894	0.865	0.807	0.828	0.858	0.804	0.786	0.736	0.606
	$MAE \downarrow$	0.041	-	0.052	0.058	0.082	0.063	0.062	0.099	0.118	0.142	0.095	0.118	0.169	0.278	0.295
NJUD [110]	$F_\beta^m \uparrow$	0.912	0.903	0.819	0.900	0.877	0.874	0.872	0.845	0.775	0.804	0.755	0.779	0.715	0.748	0.704
	$S_\alpha \uparrow$	0.911	0.903	0.899	0.900	0.878	0.878	0.877	0.849	0.772	0.763	0.748	0.711	0.686	0.695	0.690
	$E_\psi^m \uparrow$	0.934	0.944	0.911	0.950	0.923	0.925	0.924	0.913	0.853	0.864	0.838	0.803	0.799	0.803	0.754
	$MAE \downarrow$	0.040	0.043	0.053	0.041	0.053	0.060	0.059	0.085	0.100	0.141	0.157	0.160	0.172	0.153	0.189
RGBD-135 [100]	$F_\beta^m \uparrow$	0.925	0.919	0.935	0.885	0.846	0.827	0.804	0.844	0.728	0.766	0.746	0.786	0.666	0.788	0.666
	$S_\alpha \uparrow$	0.930	0.929	0.973	0.898	0.872	0.858	0.842	0.863	0.770	0.752	0.711	0.791	0.707	0.703	0.682
	$E_\psi^m \uparrow$	0.970	0.968	0.961	0.916	0.923	0.910	0.893	0.932	0.881	0.870	0.851	0.832	0.773	0.890	0.770
	$MAE \downarrow$	0.019	0.022	0.021	0.031	0.038	0.046	0.049	0.055	0.068	0.093	0.122	0.129	0.111	0.208	0.153
NLPR [98]	$F_\beta^m \uparrow$	0.923	0.916	0.902	0.897	0.867	0.863	0.841	0.825	0.771	0.778	0.793	0.768	0.648	0.745	0.681
	$S_\alpha \uparrow$	0.925	0.925	0.915	0.912	0.888	0.886	0.874	0.860	0.799	0.802	0.805	0.782	0.724	0.762	0.702
	$E_\psi^m \uparrow$	0.967	0.962	0.953	0.953	0.932	0.941	0.925	0.929	0.879	0.880	0.885	0.821	0.793	0.855	0.700
	$MAE \downarrow$	0.022	0.022	0.030	0.030	0.036	0.041	0.044	0.056	0.058	0.085	0.095	0.098	0.117	0.081	0.125
DUTO-RGBD [43]	$F_\beta^m \uparrow$	0.921	-	0.901	-	0.795	0.790	0.771	0.823	0.659	0.740	0.775	0.768	0.411	0.692	0.504
	$S_\alpha \uparrow$	0.921	-	0.903	-	0.818	0.808	0.801	0.831	0.762	0.736	0.748	0.711	0.499	0.695	0.513
	$E_\psi^m \uparrow$	0.920	-	0.937	-	0.859	0.861	0.856	0.899	0.796	0.823	0.838	0.833	0.654	0.800	0.654
	$MAE \downarrow$	0.039	-	0.043	-	0.076	0.093	0.100	0.097	0.122	0.144	0.157	0.160	0.243	0.220	0.289
LFSD [100]	$F_\beta^m \uparrow$	0.871	0.862	0.835	0.815	0.826	0.796	0.779	0.791	0.744	0.817	0.779	0.753	0.817	0.726	0.682
	$S_\alpha \uparrow$	0.862	0.854	0.837	0.824	0.828	0.801	0.794	0.796	0.738	0.791	0.694	0.688	0.753	0.736	0.659
	$E_\psi^m \uparrow$	0.905	0.893	0.873	0.856	0.872	0.847	0.835	0.865	0.815	0.840	0.819	0.799	0.856	0.804	0.701
	$MAE \downarrow$	0.062	0.078	0.094	0.106	0.088	0.111	0.112	0.119	0.133	0.167	0.197	0.209	0.155	0.208	0.225

5.4 Comparison and Result Analysis

We compare the results of proposed model with following Fourteen State-of-art methods, JL-DCF [131], S2NET [137], D3NET [130], CPFP [64] AFNet [133],TANet [180], CTMF [126], PCANet [116], DF [125] are closely related and purely deep learning-based RGBD methods. While, CDS [38], MDSF [105], DCMC [109], DES [43], and LBE [102] are latest and traditional methods based on low level handcrafted features. The Most recent, closely related, and widely referenced Fourteen State-of-the-Art saliency methods are selected for experimental analysis. Note that the above-used saliency maps are either produced by running source codes or precomputed and publicly posted by corresponding authors. We use the same experimental results, same default settings, and other related parameters as suggested in their models. The result analysis is demonstrated through visual comparison as well as quantitative comparison. The following observations from the result analysis are summarized below.

Quantitative Comparison:The quantitative analysis from Table 6.1 illustrates that the proposed model achieves a noteworthy improvement on all datasets. The improvements are visible through S-measure, E-measure, and F-measure while declining in MAE significantly. The proposed model has been achieved noticeable improvements with recently bench-marking and top-performing methods- JL-DCF [131], S2NET [137], D3NET [130], CPFP [64]. These improvements have been achieved because 2×3 stream networks can learn essential complementary features deeply

guided by CSA attention maps and preserve modality-dependent saliency. The CFF fusion model exploits all holistic features to predict exact salient objects. This quantitative analysis endorses the effectiveness of the proposed model and demonstrates the capability of generalization.

Visual Comparison: The visual assessment is shown in Fig. 5.5 to show the visual superiority of the proposed model. Most of the images shown in Fig.5.5 have complex backgrounds, low depth images, small and multiple objects which are not easily recognized with necked eyes. We have selected some illustrative examples to demonstrate the effectiveness of the proposed model, which is shown in Fig. 5.5. The 2^{nd}, 3^{rd}, and 5^{th} row image of Fig. 5.5 has a very complex salient object with non-visible depth map. Even some parts of salient objects are damaged and similar to the background. In 1^{st}, 4^{th}, and 7^{th}row image is highly complex which cannot be distinguished in color modality. Also, the 2^{th} row image contains isolated, multicolor, and multiple salient objects having some salient regions similar to the background region. For these above-mentioned challenging situations, our proposed model produces correct salient objects with no background with proper structure of salient object similar to ground truth image. While in the same situation, top contemporary models are unlikely to produce the consistent salient object. AF-Net, CTMF, PCAnet, and TANet produce salient objects with irregular borders and unstructured saliency. The most recent and accurate models, like S2NET, D3NET, and JL-DCF, produce better saliency while inconsistent on the border regions of salient objects.

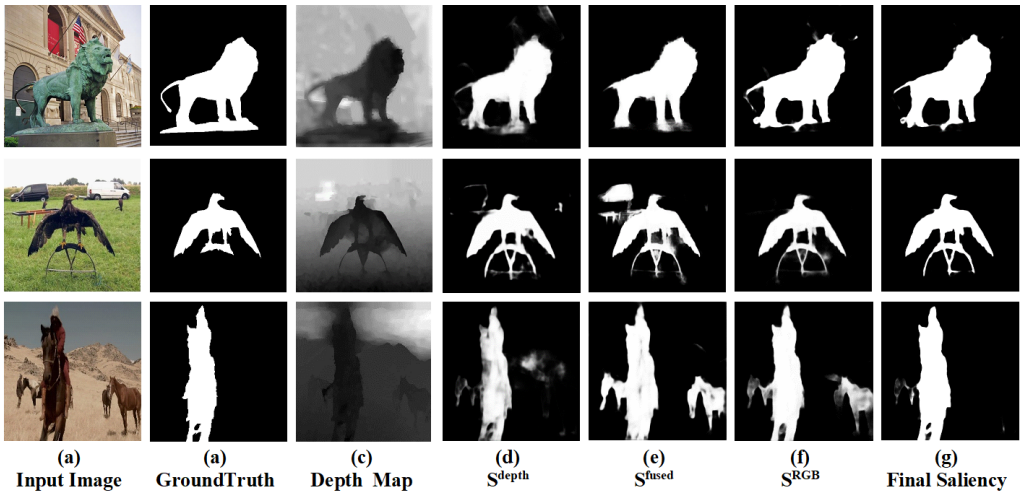

FIGURE 5.6: The successive validation of three-stream networks in Complex Images with inferior and low depth images.

5.4.1 Ablation Analysis

5.4.1.1 Validation of Network Structure and Fusion model

The 2×3 encoder and decoder of the proposed model is validated on the seven publicly available RGBD datasets with four recent evaluation parameters. This validation is essential to demonstrate the contribution of each stream in the performance improvements. In the complex background, noisy, and low depth images, any one of the three-stream cannot distinguish the salient regions individually. The gradual improvements in three-stream saliencies are shown in Table 5.2 and Fig. 5.6. Finally, the optimal Selective Saliency is based on Minimum Absolute Error, has achieved remarkable improvements in the final selection of the two similar saliencies for the late fusion, which is shown in Table 5.2 and Fig. 5.6. A single

TABLE 5.2: The validation of the effectiveness of Network Architecture and Streams-wise saliency using Mean Absolute Error-MAE.

Data-Set	Metric	S^{depth}	S^{rgb}	S^{fused}	f(D,C)	f(D,F)	f(C,F)	f(D,C,F)	**Final Saliency**
STEREO [97]	$F_\beta^m \uparrow$	0.7852	0.8354	0.8664	0.8413	0.8239	0.8529	0.8844	0.9114
	$S_\alpha \uparrow$	0.7548	0.8228	0.8764	0.8344	0.8655	0.8745	0.8789	0.9092
	$E_\psi^m \uparrow$	0.8086	0.8326	0.8865	0.8450	0.8732	0.9105	0.9288	0.9453
	$MAE \downarrow$	0.0601	0.0500	0.0431	0.0445	0.0403	0.0300	0.0260	0.0391
SSD [163]	$F_\beta^m \uparrow$	0.7563	0.8126	0.8669	.08235	0.8703	0.8970	0.8995	0.9203
	$S_\alpha \uparrow$	0.8021	0.8546	0.8743	0.8509	0.8790	0.8860	0.9007	0.9344
	$E_\psi^m \uparrow$	0.8120	0.8743	0.9123	0.8677	0.8870	0.9123	0.9340	0.9702
	$MAE \downarrow$	0.0450	0.0339	0.0206	0.0367	0.0292	0.0250	0.0230	0.0419
NJU2K [110]	$F_\beta^m \uparrow$	0.7780	0.8459	0.8796	0.8538	0.8769	0.8890	.8905	0.9120
	$S_\alpha \uparrow$	0.7978	0.8453	0.8878	0.8600	0.8760	0.8790	0.8905	0.9113
	$E_\psi^m \uparrow$	0.8394	0.8507	0.9054	0.8790	0.8907	0.9070	0.9124	0.9342
	$MAE \downarrow$	0.0670	0.0559	0.0547	0.0509	0.0487	0.0460	0.0431	0.0405
RGBD-135 [100]	$F_\beta^m \uparrow$	0.7983	0.8329	0.8740	0.8709	0.8870	0.8930	0.9067	0.9250
	$S_\alpha \uparrow$	0.8176	0.8780	0.8890	0.8701	0.8890	0.8745	0.9076	0.9300
	$E_\psi^m \uparrow$	0.8183	0.8876	0.9287	0.8790	0.9334	0.9342	0.9670	0.9701
	$MAE \downarrow$	0.0411	0.0355	0.0307	0.0339	0.0345	0.0290	0.0250	0.0196
NLPR [98]	$F_\beta^m \uparrow$	0.8012	0.8568	0.8890	0.8590	0.8760	0.8890	0.9076	0.9230
	$S_\alpha \uparrow$	0.7743	0.8489	0.8848	0.8456	0.8790	0.8900	0.9225	0.9250
	$E_\psi^m \uparrow$	0.8065	0.8532	0.9278	0.8725	0.9006	0.9129	0.9430	0.9670
	$MAE \downarrow$	0.0455	0.0369	0.0334	0.0409	0.0350	0.0298	0.0212	0.0221
DUTO-RGBD [43]	$F_\beta^m \uparrow$	0.7509	0.8012	0.8780	0.8209	0.8450	0.8800	0.9023	0.9215
	$S_\alpha \uparrow$	0.8143	0.8676	0.8890	0.8456	0.8790	0.8900	0.9067	0.9218
	$E_\psi^m \uparrow$	0.7789	0.8456	0.8876	0.8456	0.8789	0.8987	0.9086	0.9207
	$MAE \downarrow$	0.0598	0.0520	0.0491	0.0550	0.0487	0.0430	0.0420	0.0398
LFSD [100]	$F_\beta^m \uparrow$	0.7200	0.7937	0.8249	0.8010	0.8305	0.8440	0.8500	0.8713
	$S_\alpha \uparrow$	0.7016	0.7743	0.8282	0.7908	0.8309	0.8341	0.8501	0.8622
	$E_\psi^m \uparrow$	0.7600	0.8237	0.8212	0.8090	0.8220	0.8437	0.8760	0.9055
	$MAE \downarrow$	0.1206	0.0823	0.0765	0.0889	0.0809	0.0725	0.0709	0.0622

stream architecture has no generalization capability. It loses the modality dependant saliency. The top-performing methods like JL-DCF, S2-NET, D3NET, and CPFP focus on cross-complementary fusion-based saliency. Although, the proposed model improves by using the cross and infra-complementary features and modality dependant non-complementary features.

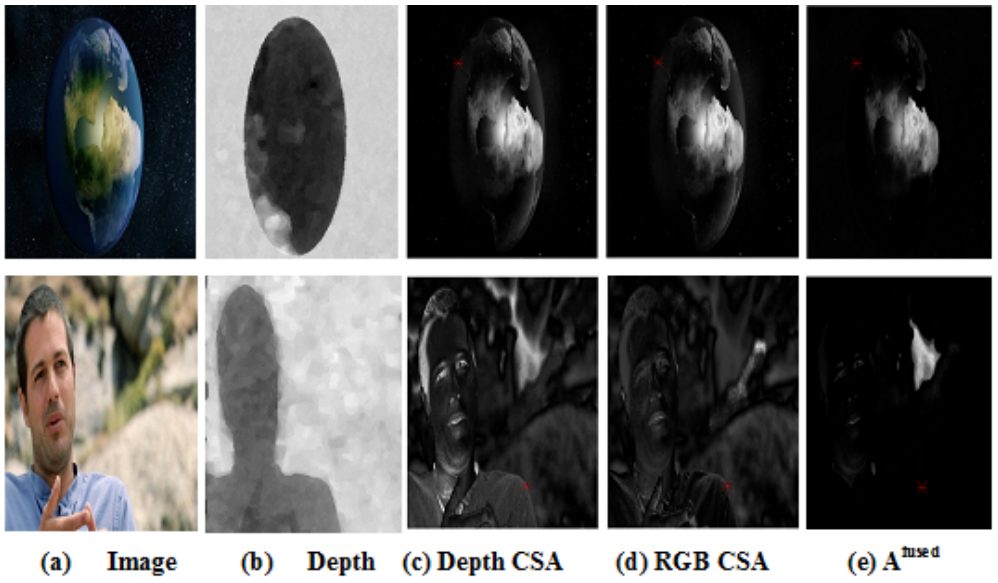

FIGURE 5.7: The visual demonstration and validation of attention map learned through proposed deeply guided, two-stage additive, Cross -complementary Self-Attention(CSA-Net).

5.4.1.2 Effectiveness of CSA-Net Model

The proposed architecture produces a holistic feature space. To validate the utilization of adequate features through the proposed two-stage fusion process is essential. The depth stream produces saliency with internal and external saliency discrepancies, although the color stream produces saliency with non-salient regions. Similarly, the fused stream produces saliency with no internal saliency discrepancy while producing with some non-salient regions. These discrepancies are shown in Fig. 5.5. At the same time, final saliency has minimum saliency discrepancy with no non-salient regions because final fusion is based on two best similar saliencies selected by the OSS model. The CSA attention map has improved these limitations, which is strongly supported through the results of Table 5.2, Table 5.3 and Fig.5.5, Fig.5.7.

TABLE 5.3: The ablation study of each component in the CSA-Net module.

Setting				DUTO-RGBD [43]				NJU2K [110]			
MSF+IFF+CFF	NL	S2MA	CSA	$F_\beta^m \uparrow$	$S_\alpha \uparrow$	$E_\psi^m \uparrow$	$MAE \downarrow$	$F_\beta^m \uparrow$	$S_\alpha \uparrow$	$E_\psi^m \uparrow$	$MAE \downarrow$
✓				0.8650	0.8535	0.8763	0.0651	0.8559	0.8645	0.8705	0.0693
✓	✓			0.8801	0.8792	0.9035	0.0512	0.8775	0.8725	0.9025	0.0572
✓	✓	✓		0.9072	0.9099	0.9231	0.4316	0.9092	0.8985	0.9299	0.0460
✓	✓		✓	0.9215	0.9218	0.9207	0.0398	0.9122	0.9113	0.9342	0.0405

The significant improvements in the results shown in Table 5.2 and Table 5.3 are observed in all datasets with all parameters validating the importance of the deeply guided attention maps CSA-Net. The successive contributions in all components and parameter settings are shown in Table 5.3, which validates the effectiveness of each component of $CSA-Net$ on complex RGBD-Datasets. The visual contribution of each step is shown in Fig. 5.5. Depth CSA, in Fig. 5.7, restructured the depth map using two-stage additive CSA(The process is shown in Fig. 5.2 in section 5.2) fusion to make it effective in the low and noisy depth maps.

5.4.1.3 Failure Cases and Analyses

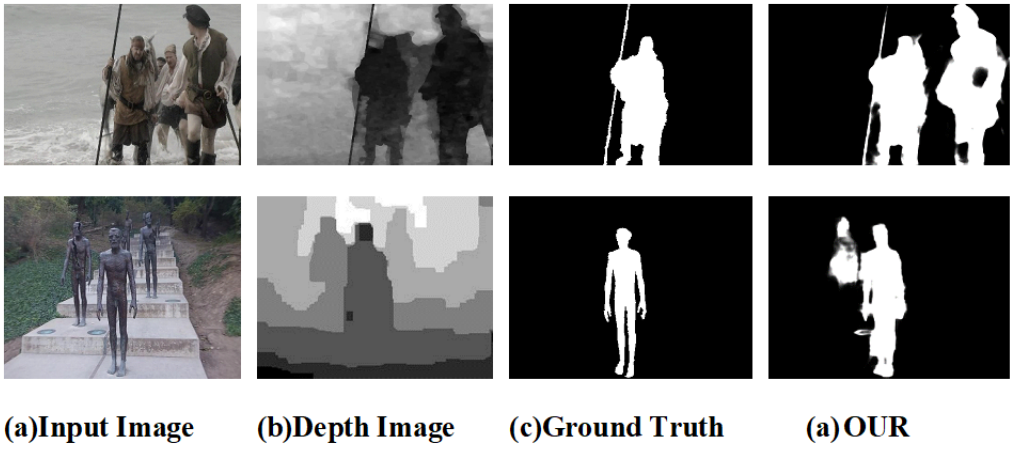

FIGURE 5.8: The visual demonstration of failure cases.

As per observations, our proposed model improves the performance while there are two challenging situations where this model shows some discrepancies in producing saliency. As shown in Fig.5.8 first row has produced more than one salient object because the depth map is inaccurate. All three objects are together and share similar characteristics in the RGB modality. Although proposed deep CSA module precisely localizes and predicts the salient object. In the second image, similar situations while objects are distinguishable only through the distance between the objects, which is not captured in the depth map exactly; therefore, some additional objects produce as salient objects. Thus, an accurate depth map plays a vital role in predicting the exact saliency in complex and cluttered situations. The proposed Cross-complementary Self-Attention(CSA-Net) enhances the deep localization of complex images. At the same time, it minimizes the salient regions which are similar to the background leading to some internal discrepancies, which are compensated by modalities preservation and intra-complementary fusions.

5.5 Conclusion

The proposed model $CSA-Net$ with 2×3 encoder and decoder produces a Holistic Feature Space. The Feature space includes all essential features, which are utilized to generate modality-specific and cross-complimentary fusion-based saliencies. The two independent encoders and decoder produce modality-dependent saliency and stage-wise cross-complementary fusion guided by the CSA module. The Fusion

model combines saliency using middle-level and late fusion strategies to explore relevant features from a Holistic feature Space. An innovative three-stream decoder based on a two-stage encoder efficiently locates the salient object with the exact object border and removes the background. The creative, progressive learning is designed to generate these features into three-stream decoder networks, and the two independent fusion strategy remarkably improves the performance. The acquisition of the deep-CSA module accurately localizes the salient object, which is missing in most contemporary methods. Our model has generalization capability, further incorporating another model to enhance saliency computation. As per discussion in failure cases, distance-based distinguishing characteristics in complex and challenging, low, and inaccurate depth maps should be formulated in the future directions for further enhancements and accurate detection of the salient object.