IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2774~2785
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2774-2785

RGB-D salient object detection with local feature and semantic segmentation

Zhang Wang[1], Kim On Chin[1], Rayner Alfred[1], Junyi Chai[2], Rundong Zhang[2], Soo See Chai[3]
[1]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia
[2]College of Science and Technology, Ningbo University, Ningbo, China
[3]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia

Article Info

Article history:

Received Jun 14, 2025
Revised Mar 23, 2026
Accepted May 7, 2026

Keywords:

Attention mechanism
Image processing
Local feature extraction
Multi-scale fusion
RGB-D salient object detection
Semantic segmentation

ABSTRACT

Red, green, blue–depth (RGB-D) salient object detection (SOD) focuses on identifying visually prominent objects by simulating human visual perception. While existing RGB-D SOD methods have demonstrated results, there remain challenges in effectively leveraging extrinsic cues and enhancing feature representation. To address these limitations, novel RGB-D SOD model with local feature extraction and semantic segmentation (LFSS) is introduced, which is built on an encoder-decoder architecture. The encoder preprocesses the input images by merging RGB and depth data through a channel and spatial attention (CSA) module. A local feature extraction module further refines this fusion. The decoder consists of three key modules: i) the multi-feature extraction (MFE) module enhances base features through diverse convolutional operations; ii) the semantic segmentation enhancement (SSE) module optimizes features via spatial pyramid pooling and atrous convolution; and iii) the local/global agreement and edge detection (LGE) module that enables multi-level feature interaction and edge detection. These modules work sequentially to enhance and extract salient objects. LFSS is evaluated on six standard RGB-D SOD datasets (NJU2K, NLPR, STERE, LFSD, SSD, SIP) by four metrics, outperforming the comparison models with up to 1.2% F-measure improvement. LFSS is found to be a versatile model, offering valuable applications in engineering.

Corresponding Author:

Kim On Chin
Faculty of Computing and Informatics, Universiti Malaysia Sabah
Kota Kinabalu, Malaysia
Email: kimonchin@ums.edu.my

1. INTRODUCTION

Salient object detection (SOD) seeks to simulate biological vision to capture the most significant regions of an image. It has been widely applied in various downstream tasks, including object detection [1], image classification [2], semantic segmentation [3], co-saliency detection [4], and object tracking. The SOD models generally accomplish two tasks: i) detecting the most prominent objects and ii) accurately segmenting the regions of these objects. To improve the effectiveness of these tasks, both intrinsic and extrinsic cues are typically utilized in SOD.

Depth maps serve as valuable extrinsic cues that can enhance the performance of saliency models. The advent of depth cameras has made it easier to access depth information from a scene. By extracting red, green, blue (RGB), and depth into a unified representation, fusion-based approaches empower saliency models to utilize depth cues, multi-scale feature fusion, integrity, and contextual information to accurately identify salient objects. Despite these advancements, existing red, green, blue–depth (RGB-D) models [5]

that incorporate extrinsic cues and saliency extraction methods still face several challenges. The fusion of depth maps with RGB images remains a critical area of exploration. Early RGB-D SOD models simply concatenated the depth map with the RGB channels, resulting in a crude and suboptimal approach. Recent models use multi-scale, multi-modal fusion modules [6]–[9], to achieve better integration of RGB and depth information. However, these methods often fail to fully exploit local feature details, indicating room for improvement in effectively utilizing extrinsic cues. Furthermore, current SOD methods based on deep learning typically adopt cross-modal and multi-scale feature extraction, deep boundary-based learning mechanisms, top-down modeling, and context-based modeling. Although these approaches extract saliency from various perspectives, they often fall short in basic feature extraction, semantic segmentation, and edge detection. Therefore, there is still significant potential for enhancing saliency extraction methods.

An RGB-D SOD based on local feature extraction and semantic segmentation (LFSS) is proposed with an encoder-decoder structure to tackle these issues. The encoder part utilizes pyramid vision transformer v2 (PVTv2) [10] to serve as the backbone network and combines a channel/spatial attention module with a local feature extraction module to enhance the integration of RGB and depth information. In the decoder part, the extraction of basic features is optimized mainly through triple convolution. The output features are further enhanced by semantic segmentation, and the effectiveness of local/global agreement detection is improved through edge detection.

The main contributions of this work are as follows: firstly, an encoder is designed that incorporates an image preprocessing module based on the PVTv2 backbone, integrating with a channel and spatial attention (CSA) mechanism, a convolution next extreme (ConvNeXt)-based [11] local feature extraction module, and absolute position encoding (APE) to enable effective cross-modal fusion of RGB and depth features. Then, a semantic segmentation enhancement (SSE) module is developed within the decoder that spatial pyramid pooling [12] is optimized by atrous convolution to enhance multi-scale semantic understanding and boundary recognition. Finally, systematic experiments are conducted on eight state-of-the-art (SOTA) models across six benchmark datasets, proving that the proposed LFSS framework achieves excellent performance in the multi-scale and multi-modal fusion of the RGB-D SOD tasks.

2. RELATED WORKS

2.1. RGB-D salient object detection model

Traditional RGB-D SOD models generally utilized calculations based on color, boundary, contrast, texture, and prior knowledge. For instance, Zhou *et al.* [13] delivered a thorough survey of these traditional methods, detailing region-based, contrast-based, and boundary-based strategies. Ji *et al.* [14] introduced a calibrated RGB-D saliency detection approach to address inter-modal inconsistencies using spatial priors. Chen *et al.* [15] treated depth as a fourth channel via a 3D convolutional neural network (3D-CNN), enabling pre-fusion of RGB and depth cross-modal features in the encoder stage. Sun *et al.* [16] proposed a depth-sensitive detection module combined with automatic multi-modal fusion to generate refined saliency maps. Although these early methods achieved initial success, their precision and robustness remained limited.

Early RGB-D SOD models with deep learning moved beyond handcrafted features. Cong *et al.* [17] presented CIR-Net, which integrates cross-modality interactive modules for adaptive fusion of RGB and depth cues. Zhang *et al.* [18] proposed BTS-Net, a bi-directional network with transfer and selection that purifies noisy depth information through progressive fusion and selection mechanisms. Jia *et al.* [19] introduced SiaTrans, a Siamese transformer network incorporating depth-image quality classification to guide dynamic cross-modal fusion. Most recently, Yi *et al.* [20] proposed GL-DMNet, which uses a dual mutual learning framework with global-partial awareness via position and channel mutual fusion modules, achieving SOTA performance. Despite these advances, key challenges remain unresolved, such as noisy depth maps, real-time processing demands, and poor generalization across diverse scenes.

2.2. RGB-D fusion method and local feature extraction

RGB-D fusion schemes are able to be generally classified into early fusion, late fusion, and multi-scale fusion. Early fusion concatenates the depth channel to the RGB channels as an additional channel. Late fusion uses a multi-scale model based on two streams to fuse RGB and depth. Multi-scale fusion models adopt a multi-level approach where each level outputs a feature map after fusion, and the final saliency is generated through the interaction across multiple scales. Due to the limitations of the first two fusion methods, recent models typically use various multi-scale fusion techniques to improve saliency extraction.

In addition to the fusion method, the ConvNeXt local feature extraction module could also be adapted to further improve the fusion effect of RGB-D models. ConvNeXt applies the local receptive field to self-attention, enabling the model to better concentrate on local details while still considering global context. ConvNeXt provides both the global feature extraction capability of Transformer and the local feature

capture capability of the CNN. Its hierarchical structure also allows the model to obtain information at different levels.

2.3. Enhancement method based on multi-scale semantic segmentation

In the RGB-D SOD, an enhancement scheme is usually applied after the image preprocessing layer to further explore the relationships between features. Spatial pyramid pooling (DeepLabv3+) based on atrous convolution [21] is one such enhancement scheme utilized in semantic segmentation. In semantic segmentation tasks, pyramid pooling and encoder-decoder architectures are frequently used, and DeepLabv3+ combines the advantages of both. Since pyramid pooling can cause the loss of object boundary details, atrous convolution is used to compensate for this shortcoming. The efficiency and accuracy of the model are further improved by using depthwise separable convolution.

DeepLabv3+ is widely applied in many models due to its effectiveness. For example, Chen *et al.* [22] presented a dense residual network (DRNet) based on DeepLabv3+, which integrates a densely connected CNN (DCNN) and a residual network (ResNet) for construction extraction from remote sensing imagery (RSI). Moreover, Du *et al.* [23] presented a model based on DeepLabv3+ and object-based image analysis (OBIA) for remote sensing images' semantic segmentation.

3. METHOD

3.1. Overview of local feature extraction and semantic segmentation

The LFSS is proposed to enhance the effectiveness of the RGB-D saliency task. As illustrated in Figure 1, LFSS adopts an encoder-decoder framework. The encoder preprocesses the RGB-D input images using a dual PVTv2 backbone network with APE. It uses a combination of a CSA module with a local feature extraction module, ConvNext, to achieve better RGB and depth fusion effect, generating multi-scale RGB-D feature representations across four levels.

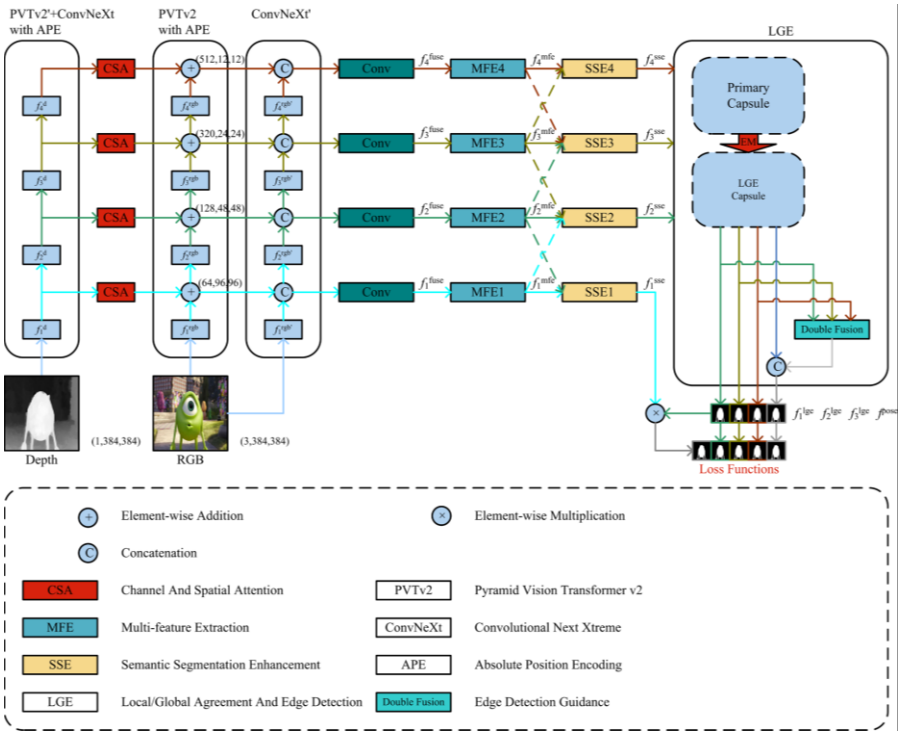

Figure 1. The proposed LFSS framework is based on an encoder-decoder model

The decoder is divided into three modules. Firstly, the multi-feature extraction (MFE) module fuses feature from three distinct convolutional kernels to capture diverse local patterns. Next, the SSE module incorporates atrous convolutions and spatial pyramid pooling to reinforce global contextual understanding and improve boundary recognition. Finally, the local/global agreement with the edge detection (LGE) module promotes cross-scale feature consistency and enhances contour precision through integrated edge detection mechanisms.

3.2. Image preprocessing

The image preprocessing module uses a dual PVTv2 backbone network with pre-trained APE weights for base information on the RGB-D channels. The entire fusion process is performed across four levels. The experiments indicate that adding an additional ConvNeXt module for local feature extraction improves RGB-D fusion accuracy for depth maps. However, the RGB maps are not processed by the ConvNeXt module. Depth features, processed by PVTv2 and ConvNeXt with APE, are passed into the CSA module, and the resulting features are combined with RGB features processed by PVTv2 with APE. Finally, the combined result passes through the ConvNeXt module for another round of local feature extraction to complete the RGB-D fusion. The formula for the fusion is shown as (1).

$$f_i^{fuse} = Conv(ConvNeXt(f_i^{rgb} + f_{csa}(f_i^d))) \quad (1)$$

Where f_i^{rgb} and f_i^d are the PVTv2 + APE–derived features for the RGB and depth streams (the latter including an initial ConvNeXt refinement). The depth-guided attention term is (2) and (3).

$$f_{csa}(f_i^d) = F_s\left(F_c(f_i^d)\right) \quad (2)$$

$$F_c\,(I) = \sigma(MLP(P_c\,(I))) \odot I, \;\; F_s\,(I) = \sigma(Conv(P_s\,(I))) \odot I \quad (3)$$

Where P_c and P_s are channel-wise and spatial max-pooling, $MLP(\cdot)$ is a shared multi-layer perceptron, $Conv(\cdot)$ denotes a 3×3 convolution, σ stands the sigmoid, and $\odot$ indicates element-wise multiplication. Collecting $\{f_i^{fuse}\}_{i=1}^4$ yields the multi-scale fused representation.

- CSA: in traditional RGB saliency methods, a well-chosen fusion method can enhance relevant features and suppress noise when merging two modalities. As shown in Figure 2, the CSA module exemplifies an effective fusion scheme by using both CSA mechanisms. The CSA module achieves efficient and lightweight RGB-D fusion through simple max-pooling-based channel attention and spatial attention, combined with a shared MLP. This approach enhances the fusion of multi-modality information, leading to improved saliency detection.
- ConvNeXt: the role of the ConvNeXt module is to perform local feature extraction in enhancing the RGB-D fusion. It is ineffective to apply the ConvNeXt module to both RGB and depth channels simultaneously. Therefore, the ConvNeXt operation is applied only to the depth map. Afterward, the RGB-D dual-stream features are fused by the CSA module. The RGB feature map, processed through ConvNeXt, is then concatenated with the fused features. This approach has been demonstrated to improve the final saliency accuracy.
- APE: in the PVTv2 encoder backbone network, the image is partitioned into several patches. PVTv2 inherently contains relative position encoding for the pixels within each image patch. However, it does not support APE cross-patches at all four levels. To address this, a truncated normal distribution is adopted for patch position as APE. This method slightly improves the model's accuracy.

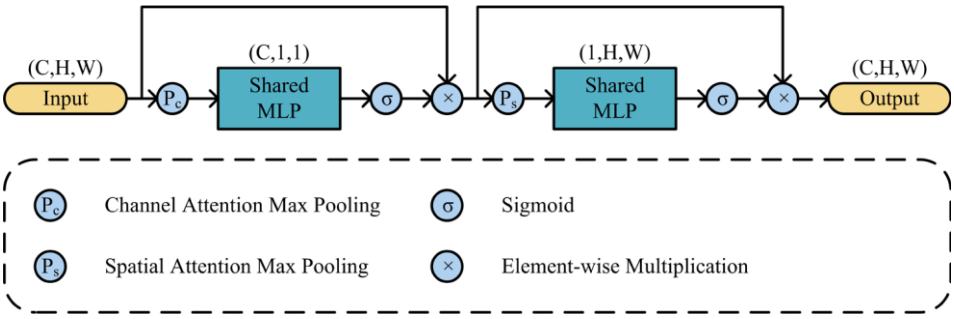

Figure 2. The chart of the CSA RGB-D fusion module

3.3. Multi-feature extraction

As shown in Figure 3, an MFE module containing three different types and sizes of kernels is used to extract diverse features, enriching the layers of information. The formula for the MFE module is shown as (4).

$$f_i^{mfe} = Conv(Concat(Ori(f_i^{fuse}), Asy(f_i^{fuse}), Dyn(f_i^{fuse}))) \quad (4)$$

Where f_i^{mfe} presents the MFE module output feature at the level $i, (i \in \{1,2,3,4\})$, respectively, f_i^{fuse} indicates the yielded features from the RGB-D fusion module, $Ori(\cdot)$ denotes original convolution, $Asy(\cdot)$ denotes asymmetric convolution with three different kernels, $Dyn(\cdot)$ denotes dynamic convolution with a weight computation kernel based on an attention mechanism. The formula for $Asy(\cdot)$ is shown as (5).

$$Asy(I) = (I * K_{3\times3}) \oplus (I * K_{1\times3}) \oplus (I * K_{3\times1}) \quad (5)$$

Where I denotes the input feature, $*$ denotes convolution, $K_{\mathrm{n\times m}}$ denotes n×m kernel shared in the same window, $\oplus$ denotes element-wise addition. The MFE module outputs $F^{mfe} = \{f_1^{mfe}, f_2^{mfe}, f_3^{mfe}, f_4^{mfe}\}$.

A dynamic convolution module was proposed to enable the weights of each pixel to be dynamically adjusted during the convolution process for local feature extraction. This module enhances the convolution operation's ability to perceive local features by incorporating an attention mechanism. This mechanism improves the robustness of the base feature extraction, and experiments have demonstrated its effectiveness in enriching the aggregated features for saliency detection.

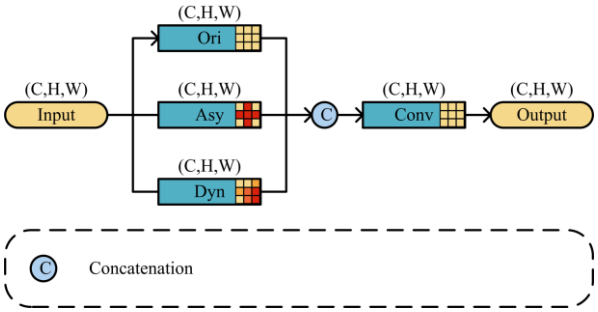

Figure 3. The chart of the MFE module

3.4. Semantic segmentation enhancement

As shown in Figure 4, an up-down-sampling (UDS) module is used to integrate relationships between upper and lower scales to mine information across multiple scales. For the current level, the next level is up-sampled and the previous level is down-sampled so that the three neighboring levels reach the same image size. The features from the three layers are concatenated first and then output. The formula for the sampling module is shown (6). Where f_i^{uds} presents the result of the up-down sampling module of the level $i, (i \in \{1,2,3,4\})$, respectively, f_i^{mfe} signifies the output features from the MFE module.

$$f_i^{uds} = Concat(f_{i-1}^{mfe}, f_i^{mfe}, f_{i+1}^{mfe}) \quad (6)$$

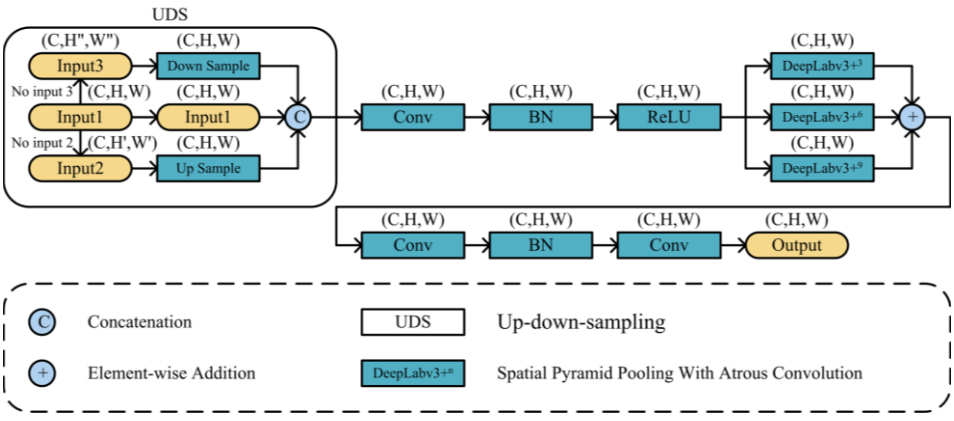

Figure 4. The structure of the SSE module

Next, three DeepLabv3+ modules with different dilation rates of 3, 6, and 9 are used to enhance salient features through image segmentation. The enhancement of the module by the SSE made the edges of the overall salient edge clearer and more defined. The enhancement formula is shown as (7).

$$f_i^{sse} = Conv(BN(Conv(DeepLabv3 +^n (ReLU(BN(Conv(f_i^{uds}))))))) \quad (7)$$

Where f_i^{sse} presents the outcome of the SSE module, f_i^{uds} demonstrate the output feature of the up-down sampling module, $DeepLabv3 +^n$ denotes enhance module based on DeepLabv3+ with a dilation rate of $n, (n \in \{3,6,9\})$, respectively. The SSE module outputs $F^{sse} = \{f_1^{sse}, f_2^{sse}, f_3^{sse}, f_4^{sse}\}$.

3.5. Local and global agreement and edge detection

Next, a capsule network is adopted to address the local and global agreement of salient objects. Additionally, a double fusion module based on edge detection guidance is introduced to enhance the effectiveness of the LGE module. The pose vectors output from the capsule network are concatenated with the edge information processed by double fusion to form the final output pose vectors.

As shown in Figure 5, the source of the LGE is the output from the former SSE module at each level. The primary capsule and the LGE capsule are first built to obtain local and global agreement results. Next, the double fusion module is utilized to extract image edge information, enhancing the pose estimation of the capsule network output. Then, a Smish activation function is carefully designed to enhance optimization during training. Finally, the pose of the capsule network is concatenated with the output of double fusion to produce the output of the LGE module. And, f^{df} presents the output feature of the double fusion module, $\oplus$ denotes element-wise addition, $FSmish(\cdot)$ *and* $Smish(\cdot)$ denote two effective activation functions, $f_1^{lge}, f_2^{lge}, f_3^{lge}$ denote the output features from the LGE module, $DS(\cdot)$ denotes down-sampling, $US(\cdot)$ denotes up-sampling. The formula for the LGE pose is shown as (8).

$$f^{lge\_pose} = Concat(Capsule(Concat(DS(f_2^{sse}), f_3^{sse}, US(f_4^{sse}))), f^{df}) \quad (8)$$

Where f^{lge_pose} presents the output pose of the LGE module, $f_2^{sse}, f_3^{sse}, f_4^{sse}$ denote output features from SSE module, $Capsule(\cdot)$ denotes capsule network. The formula for $f_1^{lge}, f_2^{lge}, f_3^{lge}$ is shown in Figure 1. The LGE module outputs $F^{lge} = \{f_1^{lge}, f_2^{lge}, f_3^{lge}, f^{lge\_pose}\}$.

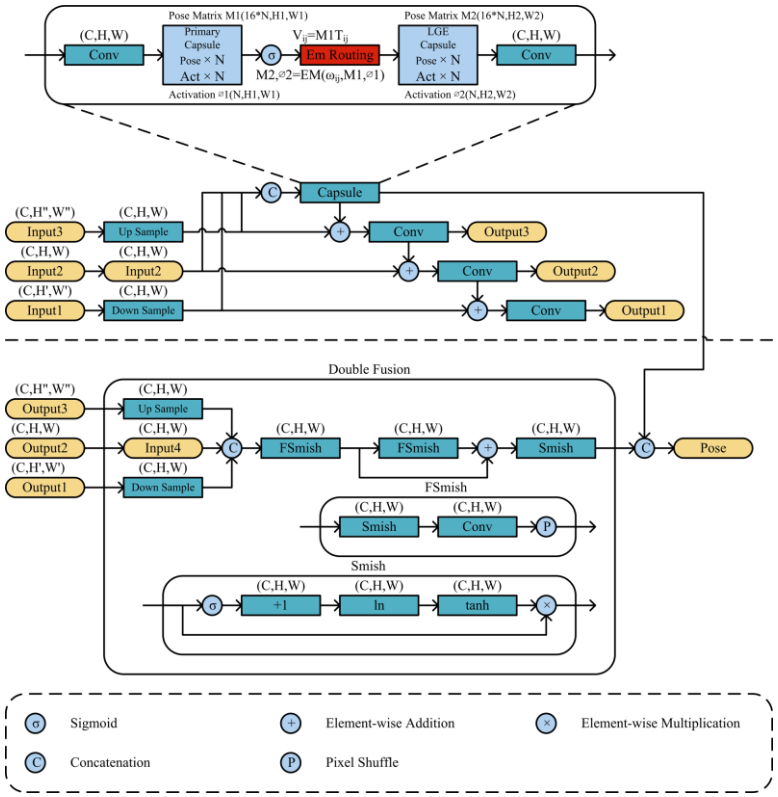

Figure 5. The structure of the LGE module

3.6. Loss functions

The model finally outputs $F^{lfss} = \{f_1^{sse} \otimes f_1^{lge}, f_1^{lge}, f_2^{lge}, f_3^{lge}, f^{lge\_pose}\}$, where f_1^{sse} presents output feature 1 of the SSE module, $\otimes$ denotes element-wise multiplication, F^{lge} refers to the output feature of the LGE module. Binary cross-entropy (BCE) and intersection over union (IoU) loss functions are adopted as supervision strategies. The formula is shown as (9). Five features in F^{lfss} are output as the final predictions under the constraint of the L_{lfss} loss as the supervision strategy, and the summation of these five losses is the final loss.

$$L_{lfss} = L_{bce} + L_{iou} \tag{9}$$

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Experimental settings

S-measure (S_α), maximum E-measure (E_ε), maximum F-measure (F_β), and mean absolute error (MAE) is used for RGB-D SOD's performance comparison. Then, the models are evaluated based on six datasets: NJU2K [1], NLPR [24], STERE [25], LFSD [26], SSD [27], and SIP [28]. 1,485 images from NJU2K and 700 images from NLPR are selected as the training dataset, and the other images in the two datasets are served as the test datasets in the training. All six datasets are subsequently used as test datasets for the final evaluation.

Eight deep-learning models are adopted for comparison purposes: the DF [24], BASNet [29], DSA2F [16], AFNet [30], DMRA [31], PiCANet [32], DCMF [33], and BBS-Net [34]. The whole experiment is conducted on a server equipped with an Intel Core i7-12700K 3.6GHz CPU, and an NVIDIA GeForce RTX 3090 GPU (24 GB video memory). The software environment utilizes Python 3.7.2 on Windows 11, with PyTorch 1.13.0+cu117 serving as the deep learning framework. The input images are uniformly resized to 384×384 pixels. The SGD optimizer is used with an initial learning rate of 0.05, a triangular learning rate schedule (0.0005–0.05), weight decay of 0.0001, and a batch size of 4. The model is trained for 200 epochs. The data and source code are openly available in GitHub at https://github.com/wz-zrd-centin/LFSS.

4.2. Quantitative evaluation

Quantitative results are illustrated in Table 1; the results demonstrate that the proposed model outperforms all comparison models across four evaluation metrics on six benchmark datasets. Among deep learning-based models, our LFSS method achieves the best overall performance. Specifically, the LFSS surpasses the BBS-Net [34], with improvements of up to 1.20%, 1.41%, 2.68%, and 0.09, respectively.

Table 1. Quantitative comparison of eight deep learning models

Dataset		Metric	Deep learning models								This model
			DF	BASNet	DSA2F	AFNet	DMRA	PiCANet	DCMF	BBS-Net	LFSS
NJU2K	[1]	$S_\alpha\uparrow$	.763	.865	.875	.884	.871	.880	.912	.921	.928
		$F_\beta\uparrow$	.804	.861	.871	.885	.865	.868	.907	.920	.933
		$E_\varepsilon\uparrow$	.864	.909	.917	.925	.916	.922	.945	.949	.959
		M↓	.141	.058	.052	.053	.057	.061	.039	.035	.028
NLPR	[24]	$S_\alpha\uparrow$	.802	.904	.880	.759	.879	.901	.918	.930	.934
		$F_\beta\uparrow$	.778	.881	.854	.699	.853	.875	.903	.918	.925
		$E_\varepsilon\uparrow$	.880	.938	.930	.836	.933	.944	.954	.961	.965
		M↓	.085	.033	.038	.070	.039	.038	.027	.023	.019
STERE	[25]	$S_\alpha\uparrow$	.757	.883	.878	.820	.844	.890	.904	.908	.919
		$F_\beta\uparrow$	.757	.875	.875	.812	.841	.876	.895	.903	.916
		$E_\varepsilon\uparrow$	.847	.928	.930	.889	.909	.931	.945	.942	.954
		M↓	.141	.048	.047	.074	.065	.056	.040	.041	.032
LFSD	[26]	$S_\alpha\uparrow$	.783	.802	.828	.817	.807	.827	.858	.864	.874
		$F_\beta\uparrow$	.813	.791	.833	.819	.811	.816	.850	.858	.881
		$E_\varepsilon\uparrow$	.857	.859	.890	.862	.878	.869	.893	.901	.911
		M↓	.145	.097	.083	.092	.095	.098	.074	.072	.066
SSD	[27]	$S_\alpha\uparrow$	.747	.811	.826	.811	.826	.837	.864	.882	.885
		$F_\beta\uparrow$	.735	.796	.802	.785	.799	.798	.841	.859	.877
		$E_\varepsilon\uparrow$	.828	.882	.874	.866	.874	.882	.919	.919	.932
		M↓	.142	.074	.064	.078	.078	.071	.056	.044	.041
SIP	[28]	$S_\alpha\uparrow$	.653	.815	.846	.628	.819	.850	.853	.879	.879
		$F_\beta\uparrow$	.657	.818	.855	.582	.820	.849	.856	.883	.891
		$E_\varepsilon\uparrow$	.759	.882	.902	.744	.877	.910	.904	.922	.927
		M↓	.185	.081	.064	.161	.083	.073	.065	.055	.049

Furthermore, as shown in Figure 6, the red lines presenting our LFSS indicate superior performance over other models. As shown in Figure 7, the model is evaluated in different scenarios. In the first line, the salient object is particularly prominent. In the second line, there are manifold salient objects. In the third line, the salient target is extremely elongated. In the fourth line, the salient target closely resembles the background. In the fifth line, the distinction between the foreground and the background is very low. In the sixth line, the target has a complex shape. These results demonstrated that our LFSS model achieves the highest accuracy among the comparison models.

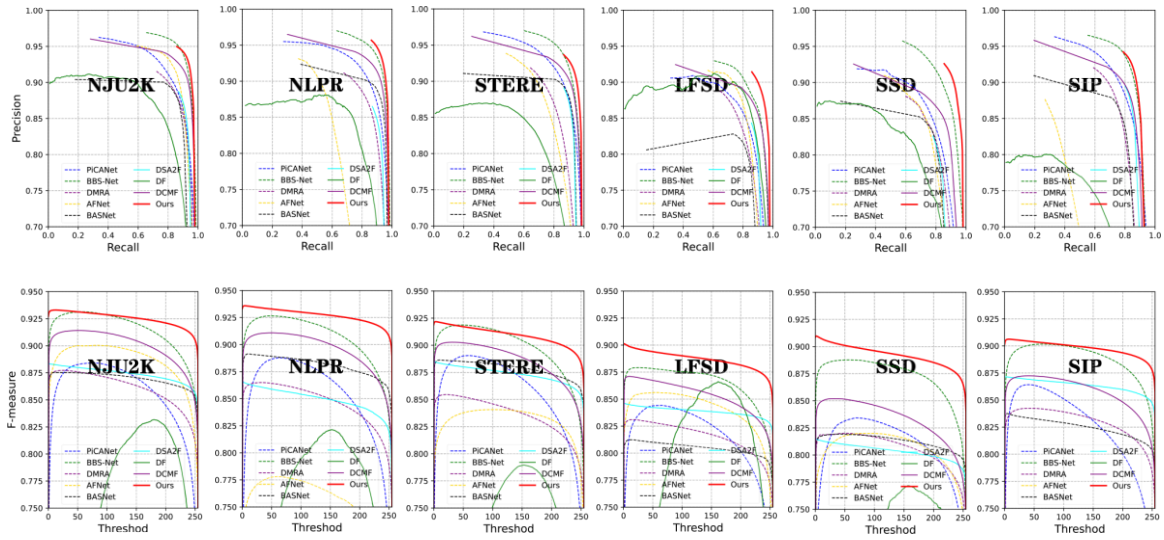

Figure 6. Precision-recall and F-measure curves of proposed method and other SOTA with six popular datasets

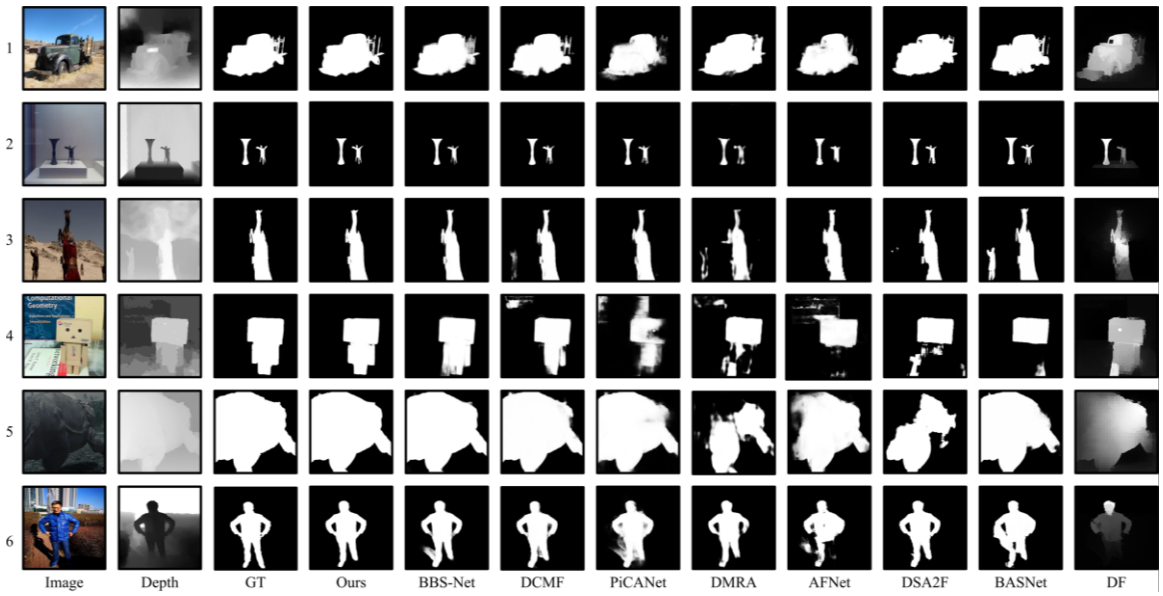

Figure 7. The visual comparison of our method and other SOTAs with six images from popular datasets and the images sorted from best to worst

4.3. Complexity analysis

As shown in Table 2, a complexity analysis of each functional module in LFSS is presented, totaling 161.4 million parameters and 57.3 giga floating-point operations (GFLOPs). The parameter count across modules exhibits a pyramid-shaped distribution, with the bottom-level feature extraction PVTv2 module

having the largest parameter count (124.0 million), the middle-level CSA + ConvNeXt fusion module following next (30.6 million), and the high-level semantic modules MFE, SSE, and LGE having the smallest parameter counts (3.8 million, 1.6 million, and 1.4 million respectively), which in line with the hierarchical representation principle of deep learning. On the RTX 3090 with a theoretical peak FP32 throughput of 35.6 tera floating-point operations per second (TFLOPS), and accounting for operational efficiency losses, LFSS can theoretically achieve 20-40 frames per second (FPS). In real tests, however, the Intel RealSense depth camera (30 FPS) and other I/O constraints lead to the pipeline being maintained at 10 FPS. This value falls within the normal range of 8-20 FPS, which is typical for transformer-intensive RGB-D SOD models. Consequently, LFSS operates well within a manageable computational budget and is readily deployable in engineering applications.

Table 2. The parameter counts and FLOPs (FP32) used by each module of LFSS

Modules	Parameter count (FP32)	FLOPs (FP32)
Full model (LFSS)	161.4 million	57.3 G
PVTv2	124.0 million	34.0 G
CSA+ConvNeXt	30.6 million	22.2 G
MFE	3.8 million	0.5 G
SSE	1.6 million	0.3 G
LGE	1.4 million	0.3 G

5. CONCLUSION

LFSS is proposed, an RGB-D SOD model integrating LFSS. The model employs encoder-decoder architecture. The encoder utilizes a PVTv2 backbone with APE, enhanced by CSA and ConvNeXt modules to fuse RGB-D data while preserving local features. The decoder combines dynamic convolution, semantic segmentation, and edge detection through MFE, SSE, LGE, and other modules to extract salient objects. The performance of the network is tested using six datasets and four metrics. After comparing it to eight deep learning models, the results show that LFSS achieves the highest precision. The model can function as a standalone module and can also be applied to other fields of machine vision to deliver excellent performance. In the field of engineering applications, even with low-quality RGB and depth data and extremely complex backgrounds, LFSS effectively extracts salient objects. At the macro level, the LFSS model accurately identifies the correct number of salient objects, while at the micro level, it captures detailed object edges. Additionally, LFSS is an intuitive and generalizable model that can be readily integrated into other vision systems.

FUNDING INFORMATION

This paper is supported by the General Research Projects of Zhejiang Provincial Department of Education (Grant No. Y202456101).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Zhang Wang	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓
Kim On Chin		✓			✓	✓		✓		✓		✓		✓
Rayner Alfred		✓			✓	✓		✓		✓		✓		✓
Junyi Chai			✓	✓	✓	✓	✓		✓	✓	✓			
Rundong Zhang	✓		✓			✓		✓	✓		✓			
Soo See Chai		✓			✓	✓		✓		✓		✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The author declares that there are no known conflicts of interest associated with this publication. To the best of the author's knowledge, no financial, personal, professional, or institutional relationships exist that could have influenced the design, conduct, interpretation, or reporting of the work in any inappropriate manner. The author also confirms that this manuscript was prepared independently and objectively, without any external pressure or interest that might compromise its integrity.

DATA AVAILABILITY

The data and source code have been made publicly available from GitHub repository at https://github.com/wz-zrd-centin/LFSS.

REFERENCES

[1] A. Borji, M. M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: a survey," *Computational Visual Media*, vol. 5, no. 2, pp. 117–150, 2019, doi: 10.1007/s41095-019-0149-9.

[2] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. V. D. Weijer, "Class-incremental learning: survey and performance evaluation on image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5513–5533, May 2023, doi: 10.1109/TPAMI.2022.3213473.

[3] I. Ulku and E. Akagündüz, "A survey on deep learning-based architectures for semantic segmentation on 2D images," *Applied Artificial Intelligence*, vol. 36, no. 1, 2022, doi: 10.1080/08839514.2022.2032924.

[4] 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, 2022, doi: 10.1109/TPAMI.2021.3051099.

[5] A. Chen, X. Li, T. He, J. Zhou, and D. Chen, "Advancing in RGB-D salient object detection: a survey," *Applied Sciences*, vol. 14, no. 17, 2024, doi: 10.3390/app14178078.

[6] Y. Peng, Z. Zhai, and M. Feng, "SLMSF-Net: a semantic localization and multi-scale fusion network for RGB-D salient object detection," *Sensors*, vol. 24, no. 4, 2024, doi: 10.3390/s24041117.

[7] M. Zhong, J. Sun, P. Ren, F. Wang, and F. Sun, "MAGNet: multi-scale awareness and global fusion network for RGB-D salient object detection," *Knowledge-Based Systems*, vol. 299, 2024, doi: 10.1016/j.knosys.2024.112126.

[8] C. Sun, Q. Zhang, C. Zhuang, and M. Zhang, "BMFNet: bifurcated multi-modal fusion network for RGB-D salient object detection," *Image and Vision Computing*, vol. 147, 2024, doi: 10.1016/j.imavis.2024.105048.

[9] Y. Peng, M. Feng, and Z. Zheng, "RGB-D salient object detection method based on multi-modal fusion and contour guidance," *IEEE Access*, vol. 11, pp. 145217–145230, 2023, doi: 10.1109/ACCESS.2023.3344644.

[10] W. Wang *et al.*, "PVTv2: improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022, doi: 10.1007/s41095-022-0274-8.

[11] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, United States, 2022, pp. 11966–11976, doi: 10.1109/CVPR52688.2022.01167.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015, doi: 10.1109/TPAMI.2015.2389824.

[13] 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, no. 1, pp. 37–69, 2021, doi: 10.1007/s41095-020-0199-z.

[14] W. Ji *et al.*, "Calibrated RGB-D salient object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9466–9476, doi: 10.1109/CVPR46437.2021.00935.

[15] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "RGB-D salient object detection via 3D convolutional neural networks," *The 35th AAAI Conference on Artificial Intelligence (AAAI-21)*, pp. 1063–1071, May 2021, doi: 10.1609/aaai.v35i2.16191.

[16] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1407–1417, doi: 10.1109/CVPR46437.2021.00146.

[17] R. Cong *et al.*, "CIR-Net: cross-modality interaction and refinement for RGB-D salient object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 6800–6815, 2022, doi: 10.1109/TIP.2022.3216198.

[18] W. Zhang, Y. Jiang, K. Fu, and Q. Zhao, "BTS-Net: bi-directional transfer-and-selection network for RGB-D salient object detection," in *I2021 IEEE International Conference on Multimedia and Expo (ICME),* Shenzhen, China, 2021, pp. 1–6, doi: 10.1109/ICME51207.2021.9428263.

[19] X. Z. Jia, C. L. DongYe, and Y. J. Peng, "SiaTrans: Siamese transformer network for RGB-D salient object detection with depth image classification," *Image and Vision Computing*, vol. 127, 2022, doi: 10.1016/j.imavis.2022.104549.

[20] K. Yi, Y. Li, J. Xu, and J. Zhang, "Dual mutual learning network with global-local awareness for RGB-D salient object detection," *Circuits, Systems, and Signal Processing*, vol. 44, no. 10, pp. 7549–7576, Jun. 2025, doi: 10.1007/s00034-025-03143-4.

[21] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *Computer Vision – ECCV 2018 (ECCV 2018): 15th European Conference*, pp. 833–851, 2018, doi: 10.1007/978-3-030-01234-2_49.

[22] M. Chen *et al.*, "DR-Net: an improved network for building extraction from high resolution remote sensing image," *Remote Sensing*, vol. 13, no. 2, pp. 1–19, 2021, doi: 10.3390/rs13020294.

[23] S. Du, S. Du, B. Liu, and X. Zhang, "Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images," *International Journal of Digital Earth*, vol. 14, no. 3, pp. 357–378, 2021, doi: 10.1080/17538947.2020.1831087.

[24] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGB-D salient object detection: a benchmark and algorithms," in *Computer Vision – ECCV 2014 (ECCV 2014): 13th European Conference*, Cham: Springer International Publishing, 2014, pp. 92–109, doi: 10.1007/978-3-319-10578-9_7.
[25] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 454–461, doi: 10.1109/CVPR.2012.6247708.
[26] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2806–2813, doi: 10.1109/CVPR.2014.359.
[27] G. Li and C. Zhu, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, 2018, pp. 3008–3014, doi: 10.1109/ICCVW.2017.355.
[28] 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, May 2021, doi: 10.1109/TNNLS.2020.2996406.
[29] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: boundary-aware salient object detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp. 7471–7481, doi: 10.1109/CVPR.2019.00766.
[30] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019, doi: 10.1109/ACCESS.2019.2913107.
[31] W. Ji *et al.*, "DMRA: depth-induced multi-scale recurrent attention network for RGB-D saliency detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 2321–2336, 2022, doi: 10.1109/TIP.2022.3154931.
[32] N. Liu, J. Han, and M. H. Yang, "PiCANet: learning pixel-wise contextual attention for saliency detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* Salt Lake City, United States, 2018, pp. 3089–3098, doi: 10.1109/CVPR.2018.00326.
[33] F. Wang, J. Pan, S. Xu, and J. Tang, "Learning discriminative cross-modality features for RGB-D saliency detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 1285–1297, 2022, doi: 10.1109/TIP.2022.3140606.
[34] D. P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Computer Vision – ECCV 2020 (ECCV 2020): 16th European Conference*, Springer, 2020, pp. 275–292, doi: 10.1007/978-3-030-58610-2_17.

BIOGRAPHIES OF AUTHORS

Zhang Wang is currently a Ph.D. candidate at the Faculty of Computing and Informatics, Universiti Malaysia Sabah, under the supervision of associate professor Kim On Chin. His research interests include AI-generated content, computer vision, and medical image enhancement. He can be contacted at email: wang_zhang_di23@iluv.ums.edu.my.

Kim On Chin is currently working as an associate professor at the Universiti Malaysia Sabah in the Faculty of Computing and Informatics. His research interests are gaming AI, evolutionary computing, evolutionary robotics, artificial neural networks, image processing, agent technologies, evolutionary data mining, and biometric security systems, with a main focus on fingerprint and voice recognition. He has led several projects related to artificial neuro-cognition for solving real world problems such as mobile based number plate detection and recognition, off-line handwriting recognition, item drop mechanism and auto map generation in gaming AI, as named a few. He can be contacted at email: kimonchin@ums.edu.my.

Rayner Alfred is a professor of computer science and director of CAMIRC at Universiti Malaysia Sabah. His research focuses on artificial intelligence, machine learning, and time series forecasting. He is recognized for contributions to real-world AI deployment in healthcare, public safety, and sustainability, and collaborates actively with government and industry. He can be contacted at email: ralfred@ums.edu.my.

Junyi Chai is currently a student at the Department of Computer Science, College of Science and Technology Ningbo University. His major is artificial intelligence, with research interests in computer vision, heterogeneous data fusion, and digital twin. He can be contacted at email: chaijunyiningbo@icloud.com.

Rundong Zhang is currently working at Ningbo Kobet Technology Co., Ltd. as an AI Engineer. His primary research areas encompass machine vision and the processing of vibration data from bearings. He can be contacted at email: 1500049485@qq.com.

Soo See Chai is presently working as an associate professor at Department of Computing and Software Engineering, Faculty of Computer Science and Information Technology (FCSIT) in Universiti Malaysia Sarawak (UNIMAS). Her research areas are GIS, remote sensing, AI, and image processing. She can be contacted at email: sschai@unimas.my.