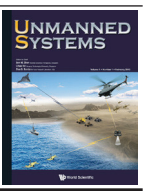
Unmanned Systems, Vol. 14, No. 4 (2026) 899–912

DOI: 10.1142/S2301385026500251

UNMANNED SYSTEMS
https://www.worldscientific.com/worldscinet/us



DGSNet: An RGB-D Salient Object Detection Network Based on a Deep Guidance Strategy

Zhiqiang Lu*, Jianlin Guo, Luwang Li

School of Artificial Intelligence, Henan University, Zhengzhou 450046, P. R. China

The existing RGB-D salient object detection (SOD) networks utilize multilevel RGB and depth features to detect salient objects. However, most models do not fully exploit the distinct advantages provided by features located at different levels and lack mechanisms for regulating the impact of depth information. This paper introduces a novel RGB-D SOD network based on a deep guidance strategy that contains two branches. The deep features extracted by one branch serve as guidance for the other branch, which produces the final detection results. The network incorporates a novel self-attention-weighted fusion mechanism, which significantly enhances the robustness exhibited in scenarios with low-quality depth maps. Additionally, a global enhancement module (GEM) for handling dense contextual information efficiently utilizes rich global data. The experimental results obtained on six commonly used benchmark datasets demonstrate that the proposed network model outperforms several existing models across five widely recognized evaluation metrics.

Keywords: Salient object detection; self-attention; neural networks; multipath attention module.

1. Introduction

Salient object detection (SOD) is aimed at replicating the human visual system for locating the regions in a scene that attracts the most attention [1]. Specifically, it focuses on identifying visually distinctive areas in images and is widely used in semantic segmentation, tracking, image/video compression, and retrieval tasks [2]. The traditional SOD methods rely on handcrafted features to compute low-level details for saliency prediction purposes, but they do not fully utilize high-level semantic information [3]. Although the algorithmic design and feature extraction processes have improved, the spatial localization and background segmentation procedures remain limited by the lack of depth information [4]. With the advent of depth sensors, RGB-D SOD has attracted significant research interest. Depth images provide illumination invariance and internal consistency, offering complementary spatial information for RGB images [5]. Employing effective interaction strategies between these modalities can increase the accuracy of saliency maps [6]. The early RGB-D SOD methods employed early or late fusion strategies to merge features derived from different modalities. While these methods have achieved some progress, they often fail to fully capture the complex complementary information between RGB and depth images [7]. The contemporary methods utilize intermediate fusion strategies, where RGB and depth features are separately computed, integrated, and then input into a decoder for the final prediction step [8]. However, low-quality depth maps can lead to inaccurate results [9]. Some intermediate fusion methods directly fuse features obtained from different convolutional layers, which can diminish the distinguishability of the fused features, especially when addressing low-quality input images or local regions [10]. Shallow features acquired from initial convolutional layers are sensitive to general structures and local details, whereas deeper features capture more complex content, offering greater discriminative power [11].

Received 17 December 2024; Revised 11 March 2025; Accepted 11 March 2025; Published 20 June 2025. This paper was recommended for publication in its revised form by editorial board member, Shiyu Zhao.
Email Addresses: *luzq@henu.edu.cn
*Corresponding author.

To address these issues, we propose a depth guidance strategy network (DGSNet), which is a framework that leverages deep features to guide a shallow network to generate results. DGSNet combines a guidance strategy similar to knowledge distillation with an effective attention mechanism and a global enhancement module (GEM). Our key contributions include the following:

(i) The network employs a strategy wherein the deep feature branch guides the shallow feature branch network, effectively leveraging and integrating features from different levels. This approach yields enhanced performance without substantially increasing the incurred computational costs.
(ii) A multimodal cross-self-attention mechanism (MAM) with depth information control is introduced. This mechanism implements context-aware control for the acquired depth information to achieve self-attention-weighted dual-stream data fusion.
(iii) A novel GEM, which is capable of conducting dense contextual information searches, is proposed. This module comprehensively understands the overall structure and semantic relationships within an image, thereby preserving more details.

2. Related Works

In recent years, RGB-D SOD technology has been extensively researched. To effectively leverage the multimodal information contained in RGB-D data, researchers have proposed various algorithms. Below, we discuss the related research in the context of the methods and strategies used in this paper.

2.1. *Deep feature-guided dual-stream network frameworks*

Network models can be classified into single-stream and multistream architectures based on the number of encoders used. Single-stream networks, which have low weights and offer high computational efficiency, include models such as the one proposed by Fu *et al*. [12], where RGB-D images are processed via a single encoder–decoder architecture, and the model presented by Piao *et al*., which efficiently implements the inference process with a single-stream network via depth distillation [13]. However, these single-stream models often face limitations in complex scenarios, particularly in tasks involving cross-modal correlation analyses.

Dual-stream architectures, on the other hand, explicitly model RGB and depth cues via two parallel encoders, as seen in Chen *et al*. approach, which uses a cascaded complementary strategy to better integrate features, leading to significant improvements [14]. Extending this concept, Zhou *et al*. introduced a three-stream network that includes two independent-modality backbones and a cross-modal distillation branch to enhance the ability to learn complementary information [15]. While three-stream networks provide more precise detection results than two-stream networks do, they are also more complex and incur higher computational costs [16].

To ensure the effective integration of the features obtained from both streams of a two-streak network, which is crucial for preventing information losses and suboptimal performance, knowledge distillation is a valuable method. Liu *et al*. utilized this approach in a saliency detection framework that leverages scene knowledge distillation, enhancing the accuracy and generalizability of the saliency detection process for remote sensing images [17].

Therefore, the proposed DGSNet method employs a dual-stream architecture and uses a deep feature-guided knowledge distillation strategy. This approach retains the rich feature representation capabilities of dual-stream networks while enhancing the fusion strategy through guided knowledge distillation, thus maximizing the performance of the dual-stream network framework.

2.2. *Self-attention with depth information control*

Self-attention mechanisms have demonstrated great effectiveness in visual tasks involving complex multimodal data [18]. In RGB-D SOD scenarios, these mechanisms provide enhanced feature preservation, calibration, and fusion capabilities. Ji *et al*. utilized channel-directed self-attention to achieve a better feature alignment effect during fusion [19]. Liu *et al*. employed mutual and nonlocal self-attention for intermodal exchanges, enhancing the alignment among deeper layers [20]. Woo *et al*. combined spatial and channel attention mechanisms to aggregate features acquired from different stages [21], whereas Song *et al*. developed a triple attention architecture to achieve optimized detection effect by integrating quality factors, channel differences, and spatial contrasts [22]. Additionally, Li *et al*. combined convolutional networks with transformers to leverage self-attention for capturing long-range dependencies, leading to improved detection results [23]. Sun *et al*. designed a network that boosts features through multihead self-attention while fusing RGB and depth features [24].

However, self-attention requires high-quality depth data; low-quality depth data can compromise attention maps. To mitigate this issue, Li *et al*. designed a two-stage network that predicts low-resolution depth maps in the first stage to reduce the impact of the process on self-attention [25]. The novel MAM mechanism proposed in this paper enhances the self-attention performance attained under low-quality depth conditions by using a depth quality

index to capture global dependencies and integrate multimodal features in a more robust manner.

2.3. *Global enhancement strategies with dense context information searches*

In computer vision, the use of contextual information is crucial for attaining improved performance [26]. Chen *et al*. introduced atrous spatial pyramid pooling (ASPP) to capture multiscale context information for improving the performance of their semantic segmentation method [27]. Zhao *et al*. developed a pyramid pooling module (PPM) that aggregates local and global context information to improve the accuracy of scene parsing [28]. Liu *et al*. combined ResNet-50 with the PPM in a superresolution-assisted learning (SRAL) framework to enhance the multiscale feature extraction process and achieve improved inference efficiency [29]. Wang *et al*. integrated multiscale context information to attain improved SOD performance [30], whereas Liu *et al*. developed a global context block (GCB) with a linear attention mechanism to efficiently explore multiscale and global contexts [31]. Zhang *et al*. reached improved parsing accuracy by aggregating context information at different levels and improved the single-image superresolution procedure by integrating multilevel context information [32, 33]. Mei *et al*. proposed a method that improves detection capabilities in real-world scenarios by capturing extensive context information [34]. Hu *et al*. developed direction-aware contextual features for shadow detection and introduced a spatial attenuation context (SAC) network to achieve enhanced SOD performance [35, 36]. Yan *et al*. proposed a GCABlock module that combines two attention mechanisms to leverage multilevel and multiscale feature information, enhancing the performance achieved in SOD tasks [37].

A novel large-scale global context information mechanism that employs a dense sampling approach is proposed in this paper to integrate local features with multiscale context information, significantly enhancing the discriminability of features. The effectiveness and robustness of this approach are validated through the experimental results produced when handling complex visual tasks.

3. Methodology

This paper provides detailed explanations of the DGSNet model and the deep guidance strategy in Sec. 3.1, outlining the processing workflow implemented by the various network modules. Section 3.2 introduces a depth quality module (DQM), which generates multilevel depth quality indices, and an MAM, which incorporates depth information control. Section 3.3 describes the GEM, which enhances the global features through dense contextual information.

3.1. *DGSNet*

The core concept of DGSNet is that it iteratively refines the constructed model, progressively enhancing the details of low-level features. High-level features contain crucial global information that aids in the location of salient objects, whereas low-level features are rich in edge and texture details. Utilizing the unique advantages of the features contained in each layer, DGSNet effectively eliminates the noise derived from the initial multimodal features and constructs the final predicted image through a hierarchical refinement process. As illustrated in Fig. 1, the convolutional features of DGSNet are divided into two groups: $Q_1 = \{\text{Conv3}, \text{Conv4}, \text{Conv5}\}$ and $Q_2 = \{\text{Conv1}, \text{Conv2}, \text{Conv3}\}$. Conv3 is a shared layer used in both Q_1 and Q_2. Group Q_2, which includes high-level abstract features with semantic information, generates guidance features that influence and promote the feature fusion process for group Q_1, resulting in saliency detection.

We refer to Q_2 as deep features and to Q_1 as shallow features. The strategy of using deep features Q_2 to guide the shallow features Q_1 is called the deep guidance strategy, which aims to more effectively integrate multimodal features. As discussed in [38], deep features contain semantic information with strong discriminative power, providing effective guidance for conducting salient object localization, whereas shallow features carry rich detail information, which plays an important role in edge optimization. Therefore, our deep guidance strategy uses the initial saliency map P generated by the deep features to guide the process of improving the final saliency detection result S. Through this mechanism, the model can iteratively refine the detailed information contained in the low-level features. By fully leveraging the unique properties of features at different levels, this strategy effectively removes noisy information from the low-level multimodal features and generates the final predicted image through a cascading refinement process.

At the beginning of the network, the features R_1 and D_1 extracted from the first layer are sent to the DQM to calculate the corresponding depth quality index. The RGB and depth information extracted from each Conv layer, along with the depth quality index, are fed into the MAM for attention-weighted feature enhancement, resulting in preliminary mixed features (f_1, f_2, f_3, f_4, f_5). In the first stage, f_3, f_4 and f_5 are sent to the GEM for global feature enhancement, resulting in f_3^*, f_4^* and f_5^*, respectively. These three enhanced features are then decoded and fused to produce the preliminary salient result P, which is expressed as follows:

$$f_i^* = \text{GEM}(f_i), \quad i = \{3, 4, 5\}, \tag{1}$$

$$P = \text{Conv}_1\{B\text{Conv}_3[DA(f_3^* \mid f_4^* \mid f_5^*)]\}. \tag{2}$$

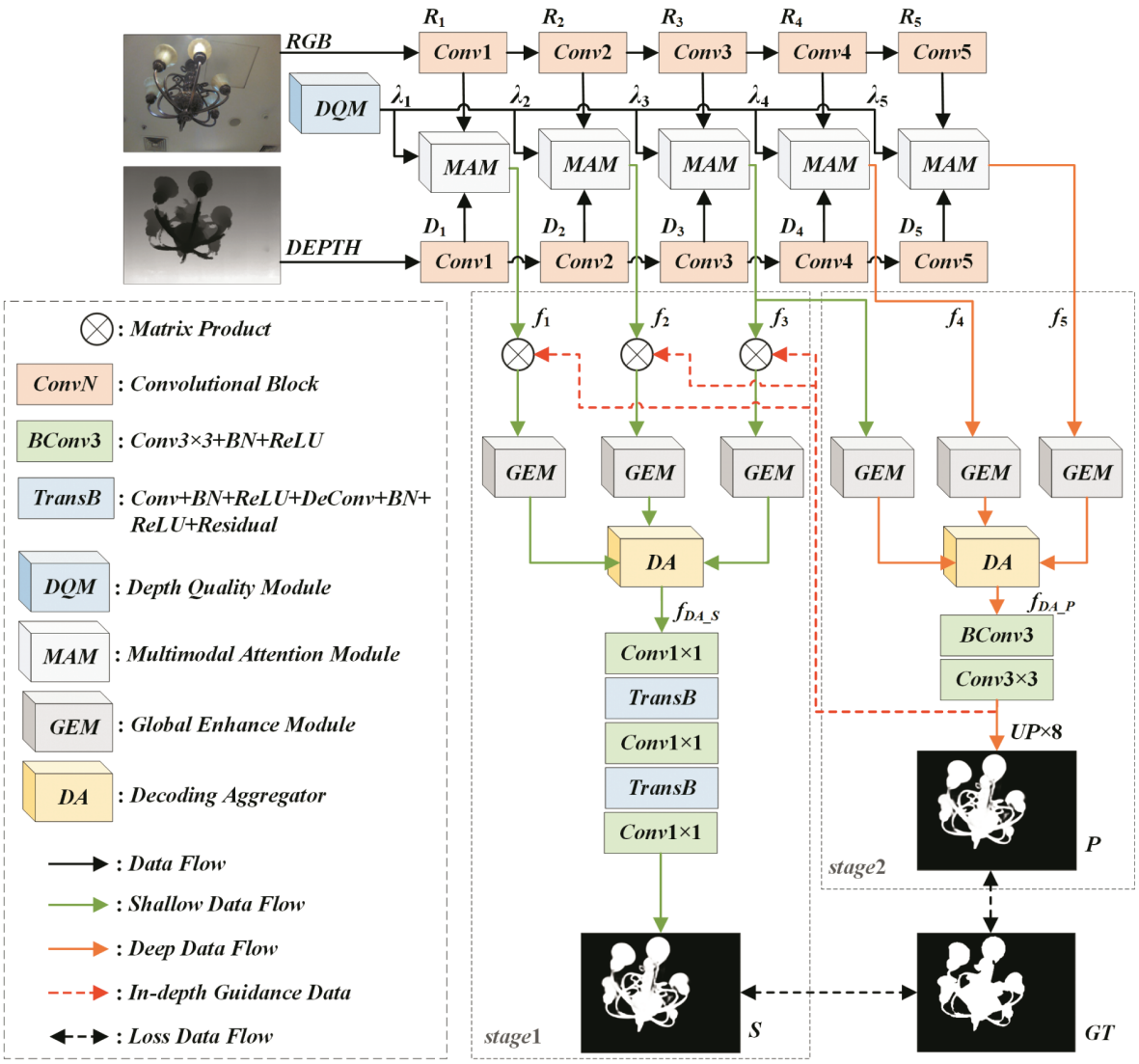


Fig. 1. The overall framework of the proposed method.

In this context, Conv1 is a standard 1×1 convolution block. *Bconv*3 is a composite operation involving a 3×3 convolution block, normalization, and an activation function, which reduces the number of channels to 1. The GEM is detailed in §3.3 of this paper.

In Eq. (1), DA refers to a decoder aggregator used in both stages of the network model. Its structure is shown in Fig. 2. The fundamental principle behind the DA is to multiply and concatenate each input feature with all higher-level features to produce the final result:

$$f_i^{*'} = f_i^* \odot \prod_{i+1}^{n} \mathrm{Conv}(F_{\mathrm{up}}(f_{i+1}^*)). \tag{3}$$

In this context, $i = 1, n = 3$ or $i = 3, n = 5$. Conv() denotes a standard 3×3 convolution operation, whereas F_{up} refers to an upsampling operation that is used to match different resolutions. $\odot$ signifies the Hadamard product. The DA result is achieved by concatenating the dimensions of the features:

$$f_{DA} = f_i^{*'} \copyright \mathrm{Conv}\{F_{\mathrm{up}}[f_{i+1}^{*'} \copyright F_{\mathrm{up}}(f_{i+2}^{*'})]\}. \tag{4}$$

In the first stage of the DA process, $i = 3$, and in the second stage of the DA procedure, $i = 1$. The preliminary salient result P obtained from Eq. (1) is used as deep guidance features to improve the aggregation process applied to the shallow features. This is represented as shown below:

$$f_i' = f_i \otimes P, \quad i \in \{1, 2, 3\}, \tag{5}$$

where f_i' represents the result of f_i after being guided by the deep features. As in stage 1, f_1', f_2' and f_3' are enhanced with the global information contained in the GEM and then sent

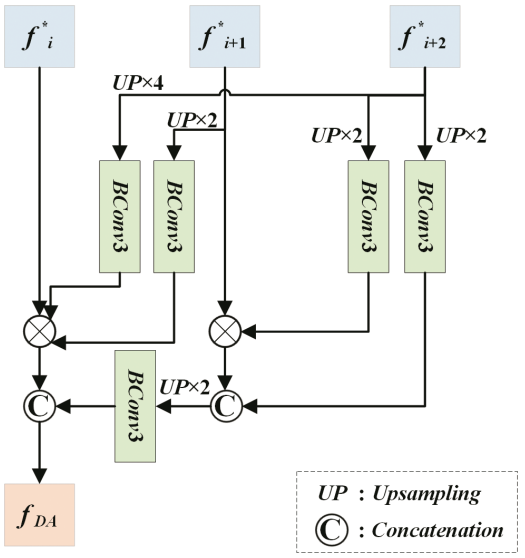


Fig. 2. The DA processes the feature outputs derived from the three GEMs by upsampling them multiple times to match their sizes before sending them to the convolution layer. Finally, matrix multiplication and dimension concatenation are performed to obtain an aggregated output.

to the DA for aggregation, resulting in the final salient result S. This process is expressed as follows:

$$f_{DA\_S} = [DA(\mathrm{GEM}(f_1'), \mathrm{GEM}(f_2'), \mathrm{GEM}(f_3'))], \tag{6}$$

$$S = B\mathrm{Conv}_1(\mathrm{Trans}B \circ B\mathrm{Conv}_1)^2 \cdot f_{DA\_S}. \tag{7}$$

(Trans $B \circ B$Conv1)^2 denotes the process where the input first goes through BConv1, followed by a transposed convolution Trans$B$. This sequence is repeated twice, and then the result passes through BConv1 once more to produce the output S. The only difference between BConv1 and BConv3 concerns their convolution kernel sizes, which are 1×1 and 3×3, respectively. Trans$B$ involves operations such as convolution, normalization, and deconvolution.

Additionally, a loss function is used to optimize the network during the learning stage:

$$L = \alpha \ell_{ce}(P, G) + (1-\alpha)\ell_{ce}(S, G). \tag{8}$$

In this loss function, ℓ_{ce} refers to the commonly used binary cross-entropy loss function, and $\alpha \in [1, 3]$ is a parameter that is used to adjust the weights of the two parts of the loss. S represents the final prediction result output by the network. G represents the ground-truth binary saliency map. The process of calculating G is as follows:

$$\ell_{ce}(S, G) = G \log S + (1-G)\log(1-S). \tag{9}$$

3.2. *Self-attention mechanism with depth information control*

A self-attention mechanism emphasizes important information by analyzing the relationships between features. However, when the quality of the available depth information is low, this mechanism may mistakenly focus on inaccurate data, compromising the performance of an RGB-D SOD model. To address this issue, we combine the DQM with a cross-modal self-attention module, enabling the self-attention mechanism to effectively manage depth information and enhance the robustness of the model.

3.2.1. *Depth quality module (DQM)*

In previous works, directly fusing RGB and depth information in cases with reliable depth data was not an effective strategy. Therefore, we design a method for evaluating a depth quality index, which is specifically intended to accurately model the contribution of depth information. The assumption is that when addressing low-quality depth data at primary levels, the network should prioritize RGB features rather than simply averaging multimodal data.

As shown in Fig. 3, the DQM takes the outputs of the encoder layer, namely, $R_1 \in R^{C\times H\times W}$ and $D_1 \in \mathrm{R}^{C\times H\times W}$, as its inputs. These features contain more modality-specific

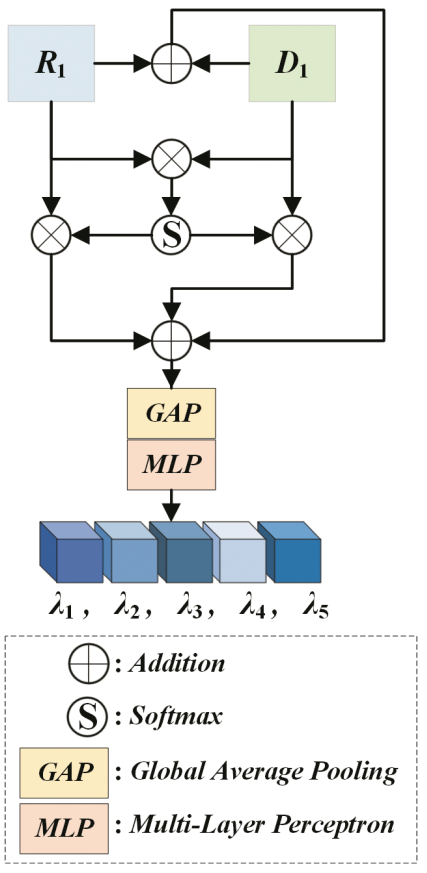


Fig. 3. The DQM takes paired RGB and depth features from the first layer Conv1 as inputs and outputs confidence values λ_i to adjust the depth contributions during the fusion process.

and heterogeneous cues than homogeneous semantic-level features. To evaluate the similarity between R_1 and D_1, contextual awareness is used instead of direct pixel-level multiplication to reduce the degree of feature misalignment and focus on measuring bias. Specifically, R_1 and D_1 are processed through a 1×1 convolution layer and then flattened to generate $R_1' \in R^{C\times HW}$ and $D_1' \in R^{C\times HW}$, respectively. The processed features are then calculated through matrix multiplication. The softmax function is used to adjust the weights for normalizing the resulting attention map. The normalized weight map is multiplied by the flattened R_1 and D_1 to enhance the cross-modality perception capabilities of the model.

The attention maps of R_1 and D_1 are then combined through addition. The similarity matrix is represented as follows:

$$\text{Attention}(R_1', D_1') = \text{softmax}\left(\frac{R_1' D_1'^T}{\sqrt{c}}\right)(R_1' + D_1'). \quad (10)$$

After generating the similarity matrix, the goal is to accurately identify biases in the depth measurements. This is done by using global average pooling (GAP) to extract feature vectors, which are then fed into a multilayer perceptron (MLP) to calculate confidence scores. Different values are computed to guide the process of fusing features at various scales. The resulting adaptive weights, denoted as $\lambda \in \mathrm{R}^5$, are defined as follows:

$$\lambda_i = \text{MLP}\{\text{GAP}[\text{Attention}(R_1', D_1')]\}, \quad i \in \{1, 2, 3, 4, 5\}. \quad (11)$$

The resulting weights form a five-dimensional tensor, with each dimension $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 being sent to a different layer of the MAM. These weights control the contributions of the depth features, transforming the original $R_i + D_i$ into $R_i + \lambda_i D_i$ for fusion purposes. This approach does not rely on the quality of the input depth data and more effectively combines multimodal features with contextual awareness, enhancing the performance of the self-attention mechanism.

3.2.2. *Multimodal cross-self-attention module*

The multimodal cross-self-attention module (MAM) enables second-order interactions between attention maps, as illustrated in Fig. 4. The MAM captures the positional relationships within the feature maps of individual data sources and interacts with the two-dimensional spatial feature maps of RGB and depth images. With the added depth quality control scheme provided by the DQM, the resulting performance is further improved.

When RGB and depth information are input into the model, two sets of matrices V_R, Q_R, K_R and V_D, Q_D, and K_D are extracted from R and D, respectively, via three 1×1 convolution kernels Conv_R and Conv_D with identical parameters. Next, matrix multiplication is performed between K_R and the transpose of Q_R, followed by activation using the softmax function to generate a self-attention map S_{att_R}. The deep self-attention map S_{att_D} is obtained through a similar process and is defined as follows:

$$S_{att_R} = \text{softmax}(Q_R^T \otimes K_R), \quad (12)$$

$$S_{att_D} = \text{softmax}(Q_D^T \otimes K_D). \quad (13)$$

$\odot$ denotes the matrix multiplication operation. These steps capture the internal correlations between features and effectively model long-range dependencies, viewing the entire input image as a whole to enhance the contextual associations between distant pixels.

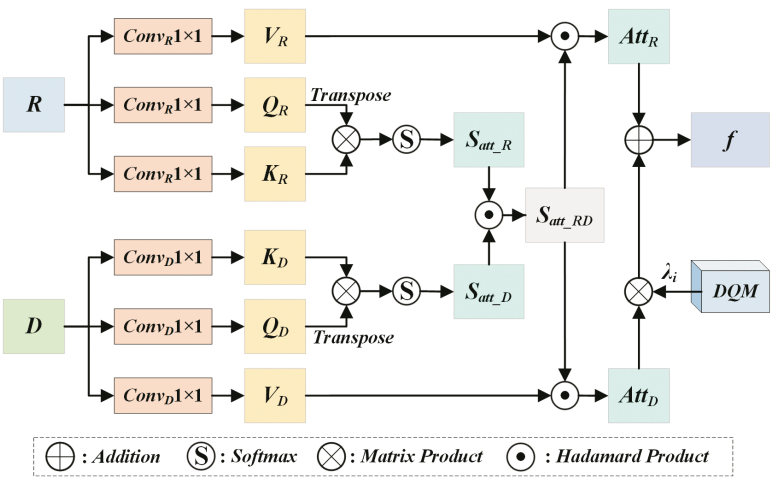


Fig. 4. The MAM computes attention weights for the RGB and depth modalities via "query", "key" and "value" matrices. It then applies cross-attention to derive the initial mixed attention. Finally, a depth quality index is used to weight these attention values, producing the final attention feature map.

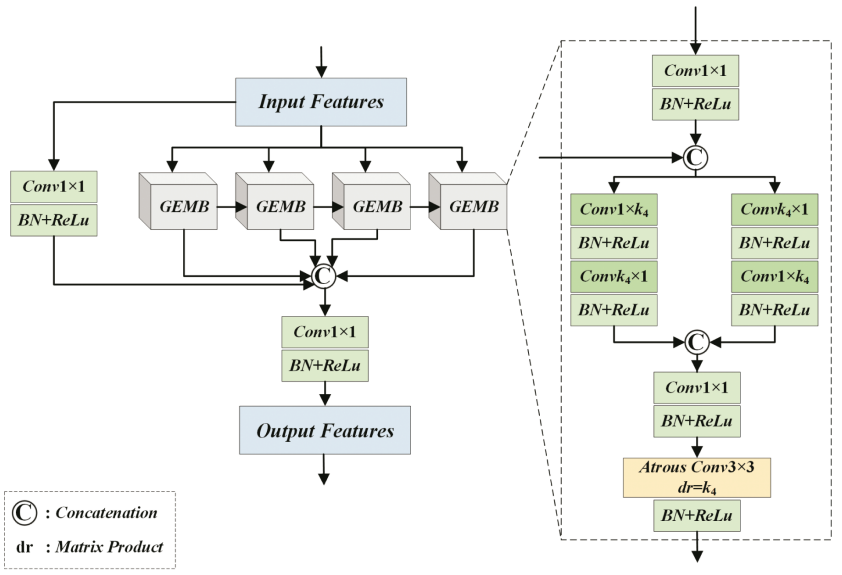


Fig. 5. The GEM consists of four GEM blocks (GEMBs), which integrate local convolutions with atrous convolutions. By using an innovative dense sampling strategy, the GEM effectively captures rich contextual information from a broad receptive field.

Next, by cross-fusing the nonlinear features of the self-attention maps derived from the RGB and depth images, a joint attention map is generated. The high-level features of S_{att_R} and S_{att_D} are combined via the Hadamard product operation to create an attention map. This map is used to weight the feature map V_R acquired from the RGB data, resulting in a weighted feature map Att_R. Similarly, the weighted feature map for the depth data is Att_D. This process is defined as follows:

$$Att_R = (S_{att_R} \odot S_{att_D}) \odot V_R, \tag{14}$$

$$Att_D = (S_{att_R} \odot S_{att_D}) \odot V_D. \tag{15}$$

The advantage of this method is that it enables the exploration of the hidden interactions between features and captures the second-order correlations between them. After the attention map is obtained, instead of directly fusing the RGB and depth information, the depth information is controlled to mitigate the impact of low-quality data. Thus, the depth quality index λ obtained from the DQM is used to adjust the weights of the depth information at different levels before fusing them to produce the final result f. This process is as follows:

$$\begin{aligned} f &= Att_R + \lambda \cdot Att_D \\ &= S_{att_R} \odot S_{att_D} \odot V_R + \lambda \cdot S_{att_R} \odot S_{att_D} \odot V_D. \end{aligned} \tag{16}$$

3.3. *Global enhancement module*

In SOD, contextual information is crucial, as high-level features reveal relationships and low-level features provide details. Traditional techniques such as large-kernel and dilated convolutions often struggle to balance efficiency and effectiveness. To address this, a GEM is developed. The GEM enhances the discriminability of features by capturing multiscale contextual information, thus improving the ability of the network to effectively learn salient features.

Figure 5 illustrates the intricate structure of the meticulously designed GEM. The GEM is composed of four GEM blocks (GEMBs) that integrate local and dilated convolutions, employing an innovative dense sampling strategy to effectively gather extensive contextual information. Each GEMB starts by reducing the number of input feature channels through a convolution layer and then applies two sets of parallel depthwise separable convolutions. These configurations efficiently capture local area information. A standard convolution layer follows, gathering rich contextual information from an enlarged receptive field. Each convolution step is accompanied by batch normalization (BN) and rectified linear unit (ReLU) activation.

To accommodate salient objects of various sizes, multiple parallel GEMBs with different kernel sizes (3, 5, 7, 9) are configured within the GEM. Their outputs are merged via a 1×1 convolution to comprehensively represent features, enhancing the ability of the model to detect salient objects at various scales. However, the actual receptive field is often smaller than the theoretical one, meaning that even the largest GEMB kernel may not fully cover all the salient areas. To mitigate this issue, the information flow channels between adjacent GEMBs allow each block to process its direct features and integrate contextual information from the previous block, enabling the subsequent GEMBs to refine information over a larger area. This design allows the GEM to effectively perceive and process dense contextual

information on a broader scale, enhancing the overall SOD performance of the network.

4. Experiments and Results

4.1. *Experimental setup, datasets, and evaluation metrics*

Datasets: To validate the proposed network model, tests were conducted on six complex RGB-D SOD benchmark datasets: NJU2K, NLPR, STERE, DES, SSD, and SIP. These datasets vary in size, ranging from 80 to 1985 images, and cover wide ranges of resolutions and scenes. They were sourced from the internet, 3D movies, cameras, and smartphones. Table 1 provides a summary of these datasets.

Specifically, the training set comprised of 1485 samples from the NJU2K dataset and 700 samples from the NLPR dataset. The test set included the remaining images from NJU2K (500) and NLPR (300), as well as the entire STERE (1000), DES (135), SSD (80), and SIP (929) datasets.

Experimental Setup: In the experiments, the input RGB and depth images were preprocessed by resizing them to 352 × 352 pixels. During the training phase, various data augmentation techniques, including random flipping, rotation, and cropping, were used to prevent overfitting. The training set comprised a total of 2185 samples drawn from portions of the NJU2K and NLPR datasets. During the testing phase, the model was separately tested on the STERE, DES, SSD, and SIP datasets, as well as the remaining parts of the NJU2K and NLPR datasets. In the evaluation phase, the model was tested and assessed on six complete public datasets.

The backbone of the network was a ResNet-50 model pretrained on ImageNet for parameter initialization purposes. The final pooling layer and fully connected layer of ResNet-50 were removed, and the outputs of the five intermediate convolutional blocks were used as inputs for the subsequent operations. The RGB and depth branches did not share weights, whereas the other parameters were initialized via the default settings of PyTorch. The model was trained using the adaptive moment estimation (Adam) optimizer with a batch size of 8, an initial learning rate of 1e4, and a learning rate decay parameter of one-tenth every 60 epochs. DGSNet was trained on a computer with a single NVIDIA GTX 4090 GPU, and it converged within 200 epochs.

***Com* Evaluation Metrics:** In terms of evaluation metrics, five widely recognized standards were selected to evaluate the performance of the tested models, including the E-measure (E_ξ), S-measure (S_α), F-measure (F_β), mean absolute error (MAE) and precision-recall (PR) curve. In this case, the E-measure (E_ξ) was used to evaluate the degrees of matching exhibited by the image at the local pixel level and at the overall image level. It is defined as follows:

$$E_\xi = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} \xi(x, y), \tag{17}$$

where w and h are the width and height of the significant figure, respectively. ξ is the enhanced alignment matrix.

The S-measure (S_α), on the other hand, is a structural metric that combines region-aware structural similarity (S_r) and object-aware structural similarity (S_o).

Here, $\in [0, 1]$ is a hyperparameter that balances S_r and S_o. It is set to 0.5 as the default setting.

The F-measure (F_β) provides a weighted harmonic mean of the precision and recall values for conducting a comprehensive performance evaluation. Its definition is:

$$F_\beta^i = \frac{(1 + \beta^2) P^i \times R^i}{\beta^2 \times P^i + R^i}, \tag{18}$$

where P^i and R^i are the precision and recall corresponding to a threshold of $i \in \{1, 2, \ldots, 255\}$, respectively. β controls the tradeoff between P^i and R^i. β^2 is set to 0.3.

The MAE metric denotes the average absolute error per pixel between the salient map and the ground truth, which is calculated as follows:

$$\mathrm{MAE} = \frac{1}{N} |S - G|, \tag{19}$$

where S and G are the predicted saliency map and the true binary map, respectively. N denotes the total number of pixels.

The PR curve is generated from a series of PR pairs that are computed from a binarized significance map with a threshold ranging from 0 to 255. Specifically, precision (P)

Table 1. Dataset descriptions.

No.	Dataset	Size	Train size	Test size	Year	Resolution	Device
1	STERE	1000	0	1000	2012	[251–1200]*[222–900]	Internet
2	NJU2K	1985	1485	500	2014	[231–1213]*[274–828]	Fuji W3 stereo camera
3	NLPR	1000	700	300	2014	640*840,480*640	Remolded Kinect
4	DES	135	0	135	2014	640*480	Microsoft Kinect
5	SSD	80	0	80	2017	960*1080	Stere movies
6	SIP	929	0	929	2020	744*992	Huawei Mate 10

Table 2. Quantitative comparison with the SOTA models. ↑ (↓) denotes that higher (lower) values are better. We use the mean absolute error (M), adapted F-measure (F_β), S-measure (S_α), and adapted E-measure (E_ξ) as evaluation metrics. (The top scores are indicated in red, and the second-best scores are indicated in blue.)

Dataset	Metric	AFNet	AISBNet	HFILNet	C2DFNet	EGANet	HiDANet	TBINet	Ours
DES	$S_\alpha \uparrow$	0.9418	0.9440	0.9450	0.9217	0.9336	0.9456	0.9363	0.9462
	$F_\beta \uparrow$	0.9297	0.9500	0.9370	0.9108	0.9058	0.9358	0.9241	0.9408
	$E_\xi \uparrow$	0.9818	0.9790	0.9800	0.9639	0.9675	0.9831	0.9747	0.9828
	$M \downarrow$	0.0152	0.0140	0.0140	0.0199	0.0209	0.0136	0.0173	0.0134
NJU2K	$S_\alpha \uparrow$	0.9260	0.9260	0.9330	0.9082	0.9210	0.9256	0.9249	0.9345
	$F_\beta \uparrow$	0.9178	0.9400	0.9310	0.8974	0.9021	0.9218	0.9162	0.9328
	$E_\xi \uparrow$	0.9298	0.9590	0.9390	0.9188	0.9239	0.9360	0.9302	0.9351
	$M \downarrow$	0.0299	0.0260	0.0250	0.0387	0.0351	0.0293	0.0291	0.0247
NLPR	$S_\alpha \uparrow$	0.9311	0.9320	0.9430	0.9279	0.9305	0.9296	0.9355	0.9398
	$F_\beta \uparrow$	0.9072	0.9340	0.9150	0.8945	0.8821	0.9077	0.9101	0.9243
	$E_\xi \uparrow$	0.9609	0.9660	0.9690	0.9574	0.9520	0.9601	0.9634	0.9702
	$M \downarrow$	0.0202	0.0190	0.0164	0.0217	0.0234	0.0212	0.0191	0.0162
SIP	$S_\alpha \uparrow$	0.8936	0.9000	0.9050	0.8715	0.8789	0.8921	0.8934	0.9073
	$F_\beta \uparrow$	0.8880	0.9150	0.9130	0.8636	0.8716	0.8929	0.8923	0.9171
	$E_\xi \uparrow$	0.9288	0.9370	0.9400	0.9154	0.9155	0.9255	0.9308	0.9408
	$M \downarrow$	0.0426	0.0380	0.0350	0.0529	0.0548	0.0438	0.0412	0.0352
SSD	$S_\alpha \uparrow$	0.8676	0.8710	0.9450	0.8718	0.8823	0.8707	0.8770	0.8928
	$F_\beta \uparrow$	0.8429	0.8610	0.8680	0.8450	0.8494	0.8452	0.8554	0.8729
	$E_\xi \uparrow$	0.9084	0.9170	0.9340	0.9097	0.9120	0.9083	0.9111	0.9168
	$M \downarrow$	0.0467	0.0450	0.0360	0.0478	0.0438	0.0472	0.0418	0.0350
STERE	$S_\alpha \uparrow$	0.9113	0.9020	0.9020	0.9023	0.9083	0.9115	0.9102	0.9202
	$F_\beta \uparrow$	0.8966	0.9030	0.9030	0.8803	0.8850	0.8969	0.8913	0.9032
	$E_\xi \uparrow$	0.9326	0.9330	0.9330	0.9255	0.9254	0.9334	0.9330	0.9334
	$M \downarrow$	0.0357	0.0290	0.0280	0.0385	0.0411	0.0353	0.0348	0.0290

and recall (R) are computed as follows:

$$P = \frac{|S' \cap G|}{|S'|}, \quad R = \frac{|S' \cap G|}{|G|}, \tag{20}$$

where S' is a binarized mask of the prediction map S according to the preset threshold.

In the experiments, the E-measure and F-measure were determined through adaptive evaluations. The detailed evaluation codes are available at http://dpfan.net/ d3netbenchmark/.

4.2. *Comparative experimental analysis*

4.2.1. *Quantitative results*

To analyze the effectiveness of the proposed model, we compared it with seven SOTA deep models (EGANet [39], C2DFNet [40], TBINet [41], HiDANet [42], HFILNet [43], AISBNet [44] and AFNet [45]). We used their published results for comparison purposes, and for the approaches without published results, we used the authors' open-source codes to perform our own experiments and obtain salient results. Table 2 presents a series of quantitative analysis results. The findings indicate that our method outperformed the existing models on some challenging datasets, such as NLPR, SSD, and STERE, demonstrating its excellent robustness in terms of handling depth biases. In addition, DGSNet achieved competitive performance on the other datasets with lower depth noise, such as DES and SIP. The PR curves of these models are shown in Fig. 6 to further illustrate the superior performance of our model.

4.2.2. *Visual comparison*

Figure 7 presents a comparison among the saliency maps generated by various methods under complex conditions, such as those with intricate foreground–background relationships and low-quality depth data. The figure clearly shows that the saliency masks produced by our method more accurately reflect the ground truth. The DGSNet model demonstrated superior precision in terms of handling low-quality depth information, which is attributable to the control process of the DQM. Additionally, the MAM enhanced the activations at the object boundaries, improving the ability of the network to leverage geometric information for saliency detection purposes. These visual comparisons highlight the effectiveness of the proposed GEM, which retained global information features more effectively than the other methods did.

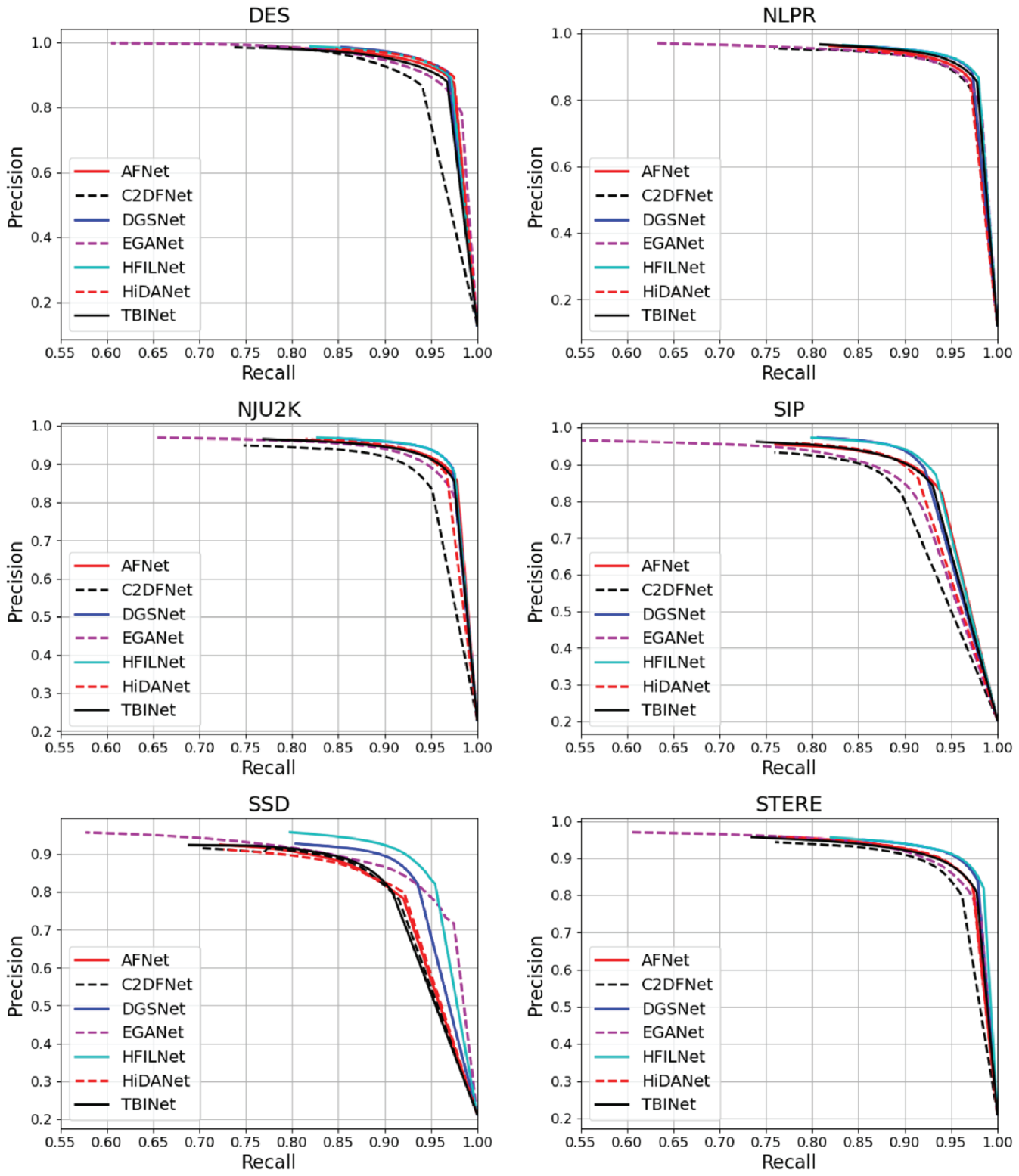


Fig. 6. The closer the PR curve is to the top-right corner of a graph, the better the corresponding performance metric.

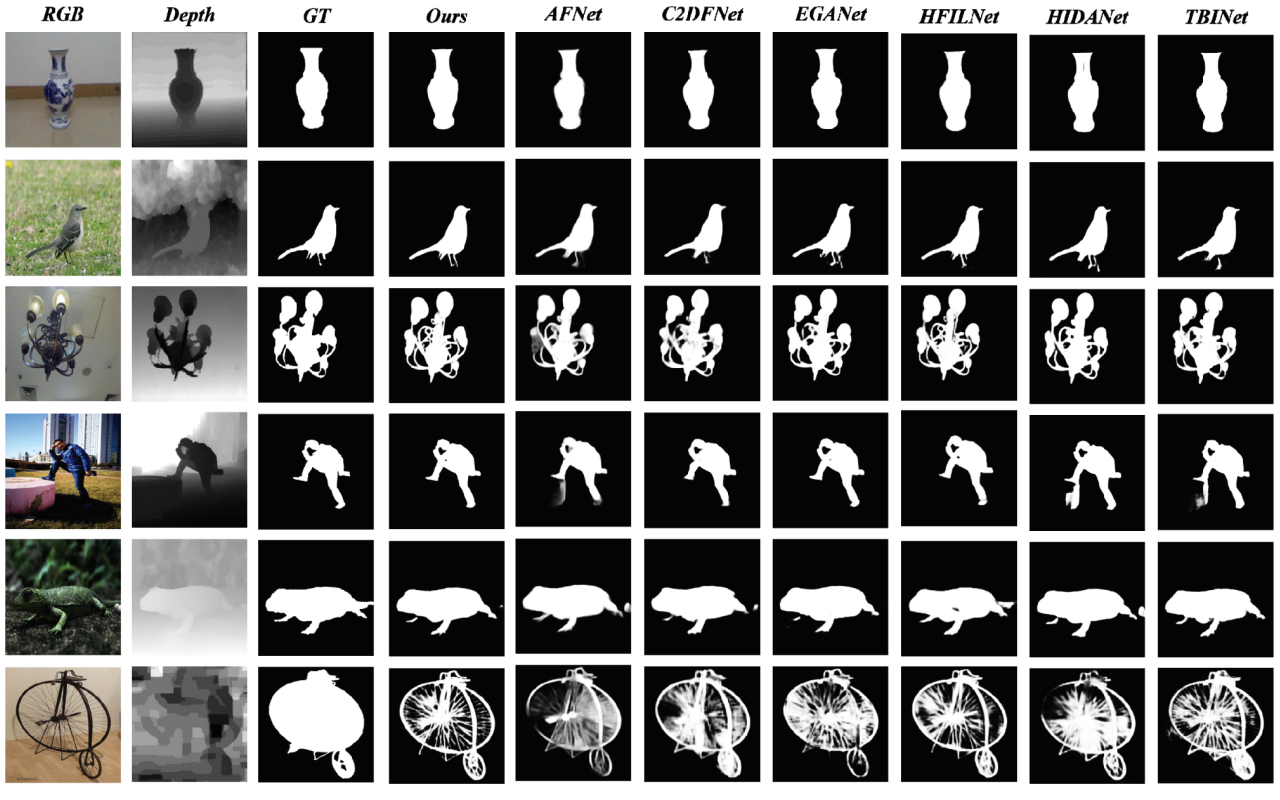


Fig. 7. Visualizations produced by different network models on the benchmark datasets.

4.3. *Ablation study*

4.3.1. *Parameter and time consumption comparisons*

As shown in Table 3, the proposed DGSNet required a parameter count of 106 M, which was significantly lower than those of several state-of-the-art methods, such as AFNet, HiDANet, and HFILNet. Despite having fewer parameters, DGSNet achieved good performance and could maintain more than 60 frames per second (FPS), satisfying the requirements of real-time processing. Moreover, the reduction in the number of parameters not only made the model more lightweight but also resulted in lower computational costs for both the training and inference processes.

Table 3. The S_α and MAE metrics used to evaluate each network model. "Ours (4)" refers to changing our 5-layer feature extraction network to a 4-layer version, whereas "Ours (6)" denotes modifying the 5-layer feature extraction network to a 6-layer version. The best results are highlighted in bold.

Models	Size (MB)	FPS	NJU2K		NLPR	
			$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
AFNet	242	50	0.9260	0.0299	0.9311	0.0202
AISBNet	104	50	0.9260	0.0260	0.9320	0.0190
HFILNet	350	35	0.9330	0.0250	0.9430	0.0164
C2DFNet	112	62	0.9082	0.0387	0.9279	0.0217
EGANet	103	34	0.9210	0.0351	0.9305	0.0234
HiDANet	169	57	0.9256	0.0293	0.9296	0.0212
TBINet	72	80	0.9249	0.0291	0.9355	0.0191
Ours (4)	92	69	0.9020	0.0351	0.9280	0.0242
Ours (6)	133	35	0.9340	**0.0245**	0.9395	0.0170
Ours	106	63	**0.9345**	0.0247	**0.9398**	**0.0162**

In addition to comparisons with other models, we also tested the performance achieved after changing the number of network layers. For the 4-layer network, we used the last two layers as deep guidance features and the first two layers as shallow features. However, the experimental results showed that although the 4-layer structure improved the inference speed, the accuracy of the model decreased. For the 6-layer network, we used the last three layers as deep guidance features and the first three layers as shallow features, but this did not result in a performance improvement. In fact, it made the model more cumbersome and slowed its training speed.

From the experimental data, it is evident that the 5-layer network, while slightly slower in terms of inference speed than the 4-layer model, achieved significantly improved performance. Therefore, sacrificing a small amount of speed is worthwhile. While the 6-layer structure could refine the feature extraction process, it introduced too much noise, resulting in no obvious performance improvements.

In conclusion, despite having fewer parameters, the 5-layer DGSNet variant strikes an effective balance between model complexity and processing time. The efficiency of DGSNet in terms of the number of parameters and time makes it a promising solution, especially in real-time or resource-constrained environments where speed is critical but accuracy must not be sacrificed.

4.3.2. *Ablation analysis of the model components*

To evaluate the impacts of individual components on the performance of the model, a series of ablation studies were conducted. These studies involved progressively removing or substituting specific components within the model, with the resulting performance assessed across the same datasets (NLPR, NJU2K, and STERE) and under uniform training conditions, including the same learning rate, batch size, and training duration. The outcomes were evaluated via the S-measure, E-measure, F-measure, and MAE, with the results summarized and statistically analyzed in Table 4.

One experiment focused on the role of the DQM by removing it from the network. This change led to equal initial weights for the RGB and depth information during the fusion process. According to the experimental results in Table 4, although the model still performed reasonably well without the inclusion of the DQM module, its detection performance was poor when the quality of the depth input was low. After the DQM module was introduced, the model demonstrated more stable and accurate performance under the influence of λ, with improvements in the attained performance metrics, especially in the case with low-quality depth inputs.

Another test examined the MAM by replacing it with an attention strategy from HiDANet. This replacement resulted in a decrease in overall performance, highlighting the importance of MAM's second-order interactions in attention maps and its capability to retain multiscale features through low-to-high-level feature fusion.

Finally, replacing our GEM with the GCM from HiDANet also led to reduced performance. The GEM proved more effective at retaining global information and managing dense information due to its design, thus demonstrating its superiority with respect to enhancing performance.

Table 4. Statistical results produced by the key model component in ablation studies. B represents the baseline performance, which was achieved when RGB-D features were combined via simple addition without any attention mechanisms. The modules marked with √ symbols are those included and used in DGSNet, whereas those without √ symbols are modules that were replaced or removed (bold font represents the best results).

B	DGSNet			Index			
	DQM	MAM	GEM	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$
√	√			0.039	0.915	0.904	0.935
√		√		0.035	0.918	0.907	0.940
√			√	0.035	0.923	0.910	0.943
√	√	√		0.035	0.920	0.908	0.941
√	√		√	0.034	0.924	0.910	0.943
√		√	√	0.035	0.921	0.908	0.941
√	√	√	√	**0.031**	**0.928**	**0.919**	**0.947**

5. Conclusion

This paper presents an innovative RGB-D SOD framework, DGSNet, with three main contributions.

(1) DGSNet features a dual-stream network architecture with a knowledge distillation strategy. It uses deep features to guide the shallow feature fusion process, effectively leveraging and integrating multilevel feature information to enhance the generalizability of the model.
(2) By utilizing multilevel context awareness to obtain depth information quality indices and applying them to a MAM, the process of fusing RGB and depth information becomes more flexible, significantly improving the robustness of the model in terms of handling low-quality depth information.
(3) A GEM is introduced to address global contextual information. The applied dense sampling strategy significantly reduces the loss of global features. Combined with those of a simple and efficient multilevel cascaded decoding aggregation method, the final detection results are improved.

These designs enable the DGSNet model to generate high-quality saliency masks in various complex scenarios, maintaining high detection accuracy and robustness even in cases with poor-quality depth information. The experimental results demonstrate that these innovative modules play crucial roles in enhancing the performance of the overall model. However, DGSNet still has areas for improvement. For example, its running speed needs to be enhanced, and the detected boundary delineation results are not ideal in cases with complex object edge variations. Additionally, the robustness of the model must be further improved.

Our future work will focus not only on continuing to improve DGSNet but also on applying it to various robotic platforms. The use of RGB-D multimodal information to perceive environments significantly enhances the perception capabilities of models, making SOD algorithms highly applicable for use in real-world scenarios.

Acknowledgment

This work was supported in part by the Key Project in Science and Technology of Henan Province (No. 222102220060).

ORCID

Zhiqiang Lu
https://orcid.org/0000-0001-8137-5506
Jianlin Guo
https://orcid.org/0009-0007-8058-7014
Luwang Li
https://orcid.org/0009-0002-4879-2206

References

[1] 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, *Knowl.-Based Syst.* **299** (2024) 112126.

[2] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S. Lu and D.-P. Fan, Bilateral attention network for rgb-d salient object detection, *IEEE Trans. Image Process.* **30** (2021) 1949–1961.

[3] S. Yang, W. Lin, G. Lin, Q. Jiang and Z. Liu, Progressive self-guided loss for salient object detection, *IEEE Trans. Image Process.* **30** (2021) 8426–8438.

[4] Z. Feng, W. Wang, W. Li, G. Li, M. Li and M. Zhou, Mfur-net: Multimodal feature fusion and unimodal feature refinement for rgb-d salient object detection, *Knowl.-Based Syst.* **299** (2024) 112022.

[5] Z. Liu, Y. Wang, Z. Tu, Y. Xiao and B. Tang, Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network, in *Proc. of the* 29*th ACM International Conf. on multimedia* (2021), pp. 4481–4490.

[6] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen and L. Shao, Rgb-d salient object detection: A survey, *Comput. Vis. Media* **7** (2021) 37–69.

[7] A. Luo, X. Li, F. Yang, Z. Jiao, H. Cheng and S. Lyu, Cascade graph neural networks for rgb-d salient object detection, in *Computer Vision–ECCV* 2020: 16*th European Conference*, *Glasgow*, *UK*, *August* 23–28, 2020, *Proceedings*, Part XII 16 (Springer, 2020), pp. 346–364.

[8] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu and Y. Fu, Reverse attention-based residual network for salient object detection, *IEEE Trans. Image Process.* **29** (2020) 3763–3776.

[9] S. Kanwal and I. Ahmad Taj, Incomplete rgb-d salient object detection: Conceal, correlate and fuse, *Pattern Recognit.* **155** (2024) 110700.

[10] Y. Yang, Q. Qin, Y. Luo, Y. Liu, Q. Zhang and J. Han, Bi-directional progressive guidance network for rgb-d salient object detection, *IEEE Trans. Circuits Syst. Video Technol.* **32**(8) (2022) 5346–5360.

[11] Z. Zhang, Z. Zhou, H. Meng and M. Yang, Lightweight multi-feature fusion network for finger vein recognition, *IEEE Sens. J.* **24** (2024) 32455–32467.

[12] K. Fu, D.-P. Fan, G.-P. Ji and Q. Zhao, Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection, in *Pro. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2020), pp. 3052–3062.

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

[14] H. Chen and Y. Li, Progressively complementarity-aware fusion network for rgb-d salient object detection, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 3051–3060.

[15] W. Zhou, J. Yuan, J. Lei and T. Luo, Tsnet: Three-stream self-attention network for rgb-d indoor semantic segmentation, *IEEE Intell. Syst.* **36**(4) (2020) 73–78.

[16] C.-H. Wang, Y.-C. Tseng, T.-H. Chiang and Y.-A. Chen, Learning multi-scale representations with single-stream network for video retrieval, in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (2023), pp. 6166–6176.

[17] Y. Liu, Z. Xiong, Y. Yuan and Q. Wang, Transcending pixels: Boosting saliency detection via scene understanding from aerial imagery, *IEEE Trans. Geosci. Remote Sens.* **61** (2023) 1–16.

[18] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang and Y. Zhao, Cirnet: Cross-modality interaction and refinement for rgb-d salient object detection, *IEEE Trans. Image Process.* **31** (2022) 6800–6815.

[19] W. Ji, *et al*., Calibrated rgb-d salient object detection, in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 9471–9481.

[20] N. Liu, N. Zhang and J. Han, Learning selective self-mutual attention for rgb-d saliency detection, in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (2020), pp. 13756–13765.

[21] S. Woo, J. Park, J.-Y. Lee and I. So Kweon, Cbam: Convolutional block attention module, in *Proc. of the European Conf. on Computer Vision* (*ECCV*) (2018), pp. 3–19.

[22] P. Song, W. Li, P. Zhong, J. Zhang, P. Konuisz, F. Duan and N. Barnes, Synergizing triple attention with depth quality for rgb-d salient object detection, *Neurocomputing* **589** (2024) 127672.

[23] B.-S. Li and H.-F. Hsiao, A hybrid convolutional and transformer network for salient object detection, In 2023 *IEEE International Conference on Visual Communications and Image Processing* (*VCIP*) (IEEE, 2023), pp. 1–5.

[24] F. Sun, P. Ren, B. Yin, F. Wang and H. Li, Catnet: A cascaded and aggregated transformer network for rgb-d salient object detection, *IEEE Trans. Multimed.* **26** (2023) 2249–2262.

[25] W. Li and R. Jia, A self-attention based network for low resolution multi-view stereo, in 2023 3*rd International Conference on Consumer Electronics and Computer Engineering* (*ICCECE*) (IEEE, 2023), pp. 646–649.

[26] I. Elezi, Exploiting contextual information with deep neural networks, arXiv:2006.11706, 2020.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4) (2017) 834–848.

[28] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, Pyramid scene parsing network. in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, 2017), pp. 2881–2890.

[29] Y. Liu, Z. Xiong, Y. Yuan and Q. Wang, Distilling knowledge from super-resolution for efficient remote sensing salient object detection, *IEEE Trans. Geosci. Remote Sens.* **61** (2023) 1–16.

[30] L. Wang, R. Chen, L. Zhu, H. Xie and X. Li, Deep sub-region network for salient object detection, *IEEE Trans. Circuits Syst. Video Technol.* **31**(2) (2020) 728–741.

[31] Y. Liu, Y. Yuan and Q. Wang, Uncertainty-aware graph reasoning with global collaborative learning for remote sensing salient object detection, *IEEE Geosci. Remote Sens. Lett.* **20** (2023) 1–5.

[32] S. Zhang, G.-J. Qi, X. Cao, Z. Song and J. Zhou, Human parsing with pyramidical gather-excite context, *IEEE Trans. Circuits Syst. Video Technol.* **31**(3) (2020) 1016–1030.

[33] J. Zhang, C. Long, Y. Wang, X. Yang, H. Mei and B. Yin, Multi-context and enhanced reconstruction network for single image super resolution, in 2020 *IEEE International Conference on Multimedia and Expo* (*ICME*) (IEEE, 2020), pp. 1–6.

[34] H. Mei, X. Yang, Y. Wang, Y. Liu, S. He, Q. Zhang, X. Wei and R. W. H. Lau, Don't hit me! glass detection in realworld scenes, in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (2020), pp. 3687–3696.

[35] X. Hu, L. Zhu, C.-W. Fu, J. Qin and P.-A Heng, Direction-aware spatial context features for shadow detection, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* (2018), pp. 7454–7462.

[36] X. Hu, C.-W. Fu, L. Zhu, T. Wang and P.-A. Heng, Sac-net: Spatial attenuation context for salient object detection, *IEEE Trans. Circuits Syst. Video Technol.* **31**(3) (2020) 1079–1090.
[37] L. Yan, R. Yan, G. Geng, M. Zhou and R. Chen, Salient object detection in optical remote sensing images based on global context mixed attention, *J. Indian Soc. Remote Sens.* **52**(7) (2024) 1489–1499.
[38] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng and J. Jiang, A simple pooling-based design for real-time salient object detection **2019** (2019) 3917–3926.
[39] L. Wei and G. Zong, Ega-net: Edge feature enhancement and global information attention network for rgb-d salient object detection, *Inf. Sci.* **626** (2023) 223–248.
[40] M. Zhang, S. Yao, B. Hu, Y. Piao and W. Ji, C2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection, *IEEE Trans. Multimedia* **25** (2022) 5142–5154.
[41] Y. Wang and Y. Zhang, Three-stage bidirectional interaction network for efficient rgb-d salient object detection, in *Proc. of the Asian Conf. on Computer Vision* (2022), pp. 3672–3689.
[42] Z. Wu, G. Allibert, F. Meriaudeau, C. Ma and C'. Demonceaux, Hidanet: Rgbd salient object detection via hierarchical depth awareness, *IEEE Trans. Image Process.* **32** (2023) 2160–2173.
[43] H. Gao, Y. Su, F. Wang and H. Li, Heterogeneous fusion and integrity learning network for rgb-d salient object detection, *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(7) (2024) 1–24.
[44] K. Zuo, H. Xiao, H. Zhang, D. Chen, T. Liu, Y. Li and H. Wen, Improving rgb-d salient object detection by addressing inconsistent saliency problems, *Knowl.-Based Syst.* **299** (2024) 111996.
[45] T. Chen, J. Xiao, X. Hu, G. Zhang and S. Wang, Adaptive fusion network for rgb-d salient object detection, *Neurocomputing* **522** (2023) 152–164.

Zhiqiang Lu received his Ph.D. degree in mechanical engineering from Xi'an University of Science and Technology, China, in 2023. Currently, he is an Associate Professor in the School of Artificial Intelligence at Henan University, China. His research primarily focuses on 3D scene recognition, deep learning models, and humanoid robot motion planning.

Jianlin Guo is currently a master's student at Henan University, with a specialization in computer vision and RGB-D salient object detection. His research interests include image processing and object detection in computer vision, with a particular focus on enhancing the accuracy and efficiency of RGB-D salient object detection.

Luwang Li is currently a master's student at Henan University, specializing in computer vision with a research focus on RGB-D Salient Object Detection. He integrates image processing and deep learning methodologies to advance the frontiers of multimodal scene understanding.