MLFNet: Rgb-d Salient Object Detection Based on Multi-Level Feature Perception

Shanshan Wang, Huiying Ru*, Bin Li, Zhao Liu

***Abstract*—This paper proposes a salient object detection model (MLFNet) based on extracting multi-level features from RGB images and depth images. The MLFNet model consists of two parallel learning networks specific to RGB images and depth images, as well as a cross-modal shared learning network. To effectively capture the salient object feature information, a multi-scale feature extraction module (FEM) is constructed in two image-specific encoders in this paper. It adopts dilated convolutions with multiple kernel sizes to perceive both global and local features of the image, eliminating the differences of salient objects, thereby enabling the model to effectively locate salient objects. To effectively integrate the specific features of RGB images and depth images, this paper constructs a dual-attention adaptive fusion module (DASM) in the shared encoder. It adopts channel attention and spatial attention to learn the specific features of RGB images and depth images, and then adaptively fuses the two image features through scalar values, thereby enhancing the shared feature output. In the shared decoder, this paper constructs a dynamic feature fusion module (DFM). It enhances the final saliency detection results by further integrating the specific features of RGB images and depth images to enrich the shared features. This paper validates the effectiveness of the MLFNet model on four RGB-D SOD benchmark datasets. Its highest accuracy rate for significance detection reached 94.3%, and the average accuracy rate was 93.1%. The experimental results show that the MLFNet proposed in this paper can detect complete salient objects in complex visual scenes. It has better robustness and accuracy compared with the existing models.**

***Index Terms*—salient object detection, encoder-decoder, feature fusion, adaptive.**

I. Introduction

SALIENT object detection (SOD) aims to simulate the human visual attention mechanism and extract the objects that the human eye particularly focuses on through a complete detection model [1]. In recent years, salient object detection has played a significant role in visual tasks such as person re-identification, weakly supervised semantic segmentation, and image quality assessment [2–4]. With the development of the field of computer vision, the demand for the preservation and transmission of key information in images has become increasingly prominent. Recent studies have shown that image saliency detection models based on deep learning have achieved remarkable results. However, the research objects of these models are mainly single-modal visible light (Red-Green-Blue, RGB) images. RGB images can provide color or texture information of the target to be detected, but they lose the three-dimensional spatial information of the target, which makes the model unable to obtain accurate salient object detection results in complex backgrounds such as low contrast. Researchers have found that introducing depth images can enable models to obtain more depth information, thus making the RGB-D image salient object detection model based on visible light and depth the mainstream approach. Depth images are different from RGB images in that they can provide spatial prior information of the target to be detected in the scene, thereby improving the accuracy of salient object detection. However, due to the limitations of imaging conditions, RGB images and depth images will contain a large amount of interfering information. Therefore, how to eliminate interfering information and effectively integrate the specific features of the two images is of great significance [5].

At present, in the field of image salient object detection based on deep learning, according to the two principles of image information fusion, it is mainly divided into three categories: RGB-D image salient object detection methods based on pixel-level fusion, decision-level fusion and feature-level fusion [6]. The RGB-D SOD method based on pixel-level fusion adopts a single-branch network structure. At the input end, it first fuses the RGB image and the depth image, then extracts the specific features of the image through a multi-level feature learning network from the fused image, and finally conducts saliency detection through a saliency object generation network. The RGB-D SOD method based on decision-level fusion adopts a dual-branch network structure. It uses the specific feature learning networks of RGB images and depth images to extract the features of the two types of images respectively. Then, the generated saliency maps of each are fused using a decision-level fusion strategy to produce the final saliency map. The RGB-D SOD method based on feature-level fusion also adopts a dual-branch structure and fuses the specific features of different levels of the two images through a cross-modal fusion module, thereby further enhancing the final saliency detection results. Since RGB-D SOD methods based on pixel-level fusion and decision-level fusion directly extract and fuse the original features from the backbone network without deeply considering the importance of multi-level features, these methods are prone to be restricted by low-quality modality data and redundant cross-modal features. In addition, the dual-branch network structure of feature-level fusion can fully integrate the multi-level features of the two images and mine the global context information in the scene. Therefore, the RGB-D SOD method based on feature-level fusion has received increasing attention.

Manuscript received January 7, 2025; revised May 17, 2025.

Shanshan Wang is the deputy director of Scientific Innovation Center of Metal Structure Engineering Branch of China 22MCC Group Co. LTD, Tangshan 064099, China. (e-mail: 1813001787@qq.com).

Huiying Ru is a teacher at Hebei University of Architecture, Zhangjiakou 075000, China. (Corresponding author, e-mail: BL1813001787@163.com).

Bin Li is a master's student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: 294386202@qq.com).

Zhao Liu is a master's student at the School of Electronic and information Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: 761566057@qq.com).

Due to the fact that the interference information contained in low quality RGB images and depth images can have a negative impact on the saliency detection results. To address this issue, this paper constructs a multi-scale feature extraction module FEM that incorporates multiple convolutional kernels and embeds it into the specific feature learning networks for RGB images and depth images. FEM learns features through atrous convolutions with multiple kernel sizes to capture image context information at different scales, thereby extracting more abundant salient features. To address the issue that the two types of image features cannot be effectively fused, this paper constructs a unified dual-attention adaptive fusion module, DASM. It adopts channel attention and spatial attention to extract image feature information and cross-fuse them, thereby enhancing the specific features of RGB images and depth images. Then, scalar value adaptive fusion of the two image features is introduced to enhance the output of the shared network. To fully utilize the specific features of the two types of images, this paper constructs a dynamic feature fusion module DFM. It fuses by integrating specific features of RGB images and depth images into a shared decoder to enrich the shared features, thereby generating the final saliency detection result.

II. RELATED WORK

Multimodal feature fusion is a key issue in RGB-D salient object detection, aiming to effectively integrate the specific features of RGB images and depth images to enhance the final saliency detection performance.The existing fusion methods can be classified into pixel-level fusion methods, feature-level fusion methods and decision-level fusion methods.

A. pixel-level fusion methods

Pixel-level fusion-based methods typically adopt a single-branch network structure as the foundation and concatenate RGB images and depth images through certain fusion strategies to obtain multi-channel input RGB-D images. Although simple and efficient, due to the differences in the two types of image feature information, the pixel-level fusion method cannot fully utilize the complementarity of RGB images and depth images. Li et al. [7] fused RGB images and depth images through a time series cascade approach to improve the simple cascade strategy. Then, 3D convolution is fully utilized to learn the complementary information between RGB images and depth images, and the receptive field of 3D convolution is dynamically adjusted through the spatial prior information contained in the depth image to optimize 3D convolution, thereby enhancing the feature output of RGB-D images. Chen et al. [8] also adopted the time series to cascade two images, thereby generating pseudo RGB-D images. It adopts an encoder-decoder structure to extract and fuse features, and designs a new channel attention module to enhance the learning ability of multimodal features, thereby improving the final saliency detection results. Zhou et al. [9] proposed a multi-view saliency detection model, which converts RGB images into multiple perspectives for sampling. Then, the generated multi-view RGB images are fused with depth images to map the feature information of RGB images into 3D space for predicting salient objects.

B. feature-level fusion methods

The method based on feature-level fusion aims to achieve the fusion of features from RGB images and depth images through certain strategies. It can achieve better results than image-level fusion, so current saliency detection models usually adopt feature-level fusion methods. Early feature fusion methods typically employed addition or concatenation strategies, which, although simple to operate, were difficult to fully utilize the complementary information of RGB images and depth images. Therefore, how to design a more efficient feature fusion strategy is the current research focus. For instance, Zhang et al. [10] proposed a novel bidirectional transfer-select module, which established an interaction relationship of feature information between RGB images and depth images through a bidirectional structure, thereby achieving effective fusion of specific features of the two types of images. Zhou et al. [11] proposed a feature fusion and reshaping network, which achieves recursive interaction between RGB image features and depth image features of adjacent layers through an interaction fusion module and a multi-scale pyramid module, thereby fully mining the context information of RGB-D images. Since high-level features of images contain rich abstract semantic cues and low-level features have abundant detailed information, a unified fusion strategy cannot fully utilize the multi-level features of the two images. Therefore, Yao et al. [12] designed a dual dilation merging module for high-level features to fully learn the context information of high-level features. A feature fusion enhancement module was designed for low-level features to selectively fuse the channel information and spatial information contained in the features. Zhu et al. [13] proposed an adaptive collaborative fusion network, which adopts a two-stage fusion strategy to integrate low-level features and high-level features, thereby enhancing the accuracy of salient object detection.

C. decision-level fusion methods

The decision-level fusion method differs from the pixel-level and feature-level fusion methods. It adopts a dual-branch network structure to obtain the saliency detection results of RGB images and depth images respectively, and then fuses the saliency detection results of the two images through a certain strategy. Wang et al. [14] proposed a dynamic salient object detection framework. It first predicts the saliency maps of the two types of images, RGB and depth images, respectively, through two independent feature extraction networks for RGB images and depth images. Then, it dynamically generates a weight map using a shared sub-network to fuse the saliency maps of the RGB and depth images. In addition, Wang et al. [15] proposed a new multi-level feature fusion method. It generates the fusion weights of RGB images and depth images through the method of reinforcement learning, thereby guiding the fusion of the saliency results of the two types of images.

III. METHOD

A. Specific feature learning network

1) Structure: Fig. 1 shows the overall framework of the MLFNet proposed in this paper. MLFNet is composed of a multi-branch stream network, namely the specific learning

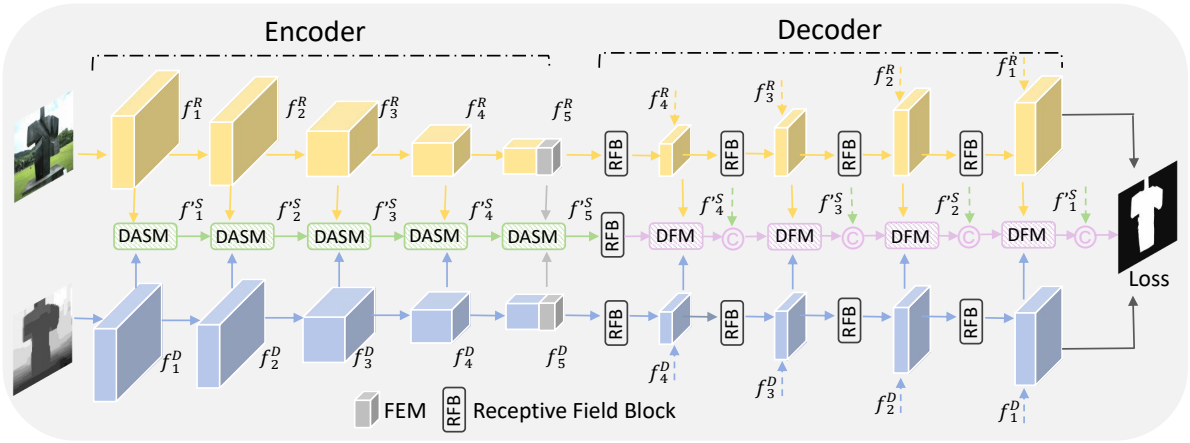

Fig. 1: Architecture of FFENet. It includes a multi-scale feature extraction module (FEM Section 3.1.2), a dual-attention adaptive fusion module (DASM Section 3.2.2), and a dynamic feature fusion module (DFM Section 3.2.3). The FEM in this paper effectively eliminates the differences of salient objects in different visual scenes. This paper's DASM enhances the shared feature output. The enhanced features are used to generate the final saliency detection results through DFM.

networks for RGB images and depth images, as well as the cross-modal shared learning network. Firstly, a specific sub-network receives the RGB image (H) and the depth image (D), and then extracts the feature information of the RGB image and the depth image through two image-specific encoders. The encoders of the specific sub-networks for RGB images and depth images are based on ResNet-50 [16], from which five multi-scale features can be extracted [17], namely R = $\{f_m^R, \mathrm{m} = 1, \cdots, 5\}$ and D = $\{f_m^D, \mathrm{m} = 1, \cdots, 5\}$. After the two image-specific encoders extract multi-scale features f_5^R and f_5^D, they are passed to the two image-specific decoders to generate their respective saliency prediction maps. The specific decoders for RGB images and depth images are constructed using the U-Net structure [18]. They enhance the saliency detection results by integrating multi-scale features learned by the encoder through residual connections.

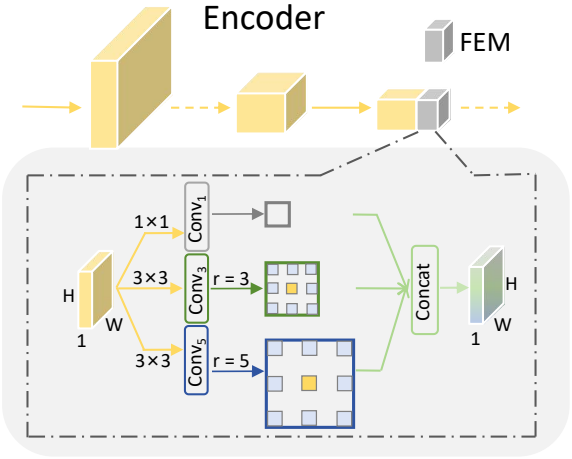

Fig. 2: The proposed a multi-scale feature extraction module.

Due to the significant differences in shape, size and position of the targets to be detected, some local feature information will be lost when two types of image features are extracted by a specific sub-network. To address this issue, this paper constructs a multi-scale feature extraction module, FEM. It is naturally embedded into the specific sub-networks of RGB images and depth images, capturing the salient object information contained in the two types of images through dilated convolutions with various kernel sizes, thereby eliminating the feature differences of the targets to be detected in different scenarios.

2) Multi-scale Feature Extraction Module: Current salient object detection methods typically employ convolutional layers and pooling layers when extracting specific features from RGB images and depth images. However, in different visual scenes, salient objects vary in size, shape and position, which leads to the loss of some feature information during the sampling process. To address this issue, this paper constructs a multi-scale feature extraction module, FEM. It adopts dilated convolutions with multiple different kernel sizes to learn the features of RGB images and depth images, in order to capture the multi-scale features of the two types of images and thereby eliminate the differences between salient objects.

As shown in Fig. 2, after the specific encoders for RGB images and depth images extract high-level features f_5^R and f_5^D, they are passed to the FEM module for sampling. To extract the rich contextual information of the two types of images, this paper adopts convolution with multiple dilation values for feature sampling. In this process, this paper adopts dilated convolution to expand the receptive field, thereby effectively capturing the salient features contained in the two types of images. Then, the captured multi-scale features are concatenated along the channel direction, and 1×1 convolution is used to process the multi-scale feature maps to avoid introducing redundant information. In addition, this paper concatenates the original features with the generated multi-scale feature maps through average pooling to obtain more comprehensive feature information. The formula for processing feature maps using dilated convolutions with multiple dilation values is as follows:

$$FEM(f) = \sigma(Conv_1(f), Conv_3(f), Conv_5(f)) \quad (1)$$

Where $\sigma(\cdot)$ represents the Sigmoid activation function. Conv1, Conv3 and Conv5 represent dilated convolutions with dilation values of 1, 3 and 5 respectively.

B. Shared learning network

1) Structure: As shown in Fig. 1, the features extracted by the RGB image and depth image encoders are fused in the shared encoder. After the shared encoder fuses to generate deep-level features f'^{S}_{5}, they are passed to the shared decoder to generate the saliency map. The shared decoder is also built based on the U-Net structure and uses residual connections to associate the multi-scale features extracted by the encoder, thereby enhancing the output of the shared features. To effectively integrate the specific features of RGB images and depth images, this paper constructs a dual-attention adaptive fusion module (DASM) in the shared encoder. It adopts channel attention and spatial attention to extract specific features from RGB images and depth images. Then, the extracted features are cross-fused to achieve bidirectional feature calibration. In addition, this paper introduces the scalar value adaptive fusion calibrated features to further enhance the accuracy of the shared features. The enhanced features are fused and passed to the shared decoder, where the final saliency detection result is generated through the dynamic feature fusion module (DFM).

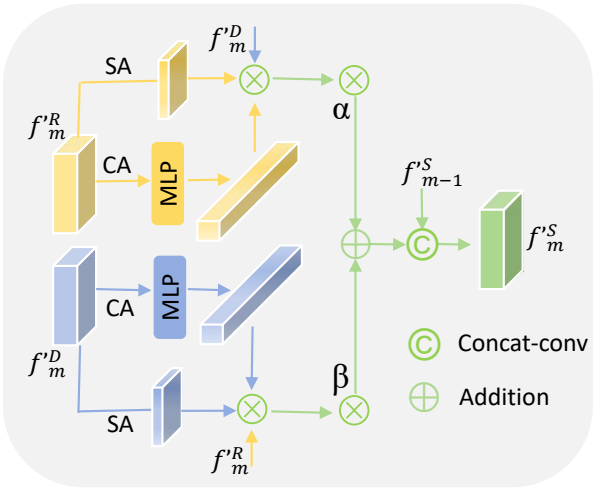

Fig. 3: The proposed a dual-attention adaptive fusion module.

2) Dual-Attention Adaptive Fusion Module: The visual information provided by RGB images and the spatial prior information provided by depth images help the model to more accurately locate the salient object regions. However, in complex scenarios, most salient object detection methods assume that the features of the two modalities are perfectly aligned, which often affects the final detection performance. To address this issue, this paper constructs a unified dual-attention adaptive fusion module, DASM. Firstly, it extracts two types of image feature information through channel attention and spatial attention. Then, it cross-fuses the extracted RGB image and depth image feature information to achieve bidirectional calibration of the shared encoder feature output. Then, scalar values are introduced to adjust the feature contributions of different branches, so as to adaptively fuse the two types of image features and thereby enhance the shared feature output.

As shown in Fig. 3, let the RGB image and depth image features contained in the mth layer be f^{R}_{m} and f^{D}_{m}. This paper first uses 1×1 convolution to compress the number of feature channels of the mth layer of the two images to $C_m/2$. Then, the specific features of the two images are processed through a 3×3 convolutional layer with a Sigmoid activation function, thereby obtaining the normalized feature maps f'^{R}_{m} and f'^{D}_{m}. The formulas for obtaining the two types of normalized feature maps of images are as follows:

$$f'^{R}_{m} = Conv_{3\times3}\left(Conv_{1\times1}\left(f^{R}_{m}\right)\right) \tag{2}$$

$$f'^{D}_{m} = Conv_{3\times3}\left(Conv_{1\times1}\left(f^{D}_{m}\right)\right) \tag{3}$$

After obtaining the normalized feature maps of the two images, this paper suppresses irrelevant information through channel attention and focuses on the salient object regions by using spatial attention, thereby enabling the model to effectively distinguish salient objects from the background. In addition, this paper adopts element-wise multiplication to cross-fuse the attention feature maps of the two images, so as to fully utilize the complementary features of RGB images and depth images, thereby enhancing the representation ability of the shared network. The formulas for learning feature maps using channel attention and spatial attention are as follows:

$$M_c = \sigma(MLP(f'_{avg}) + MLP(f'_{max})) \tag{4}$$

$$M_s = \sigma(Concat(f'_{avg}, f'_{avg})) \tag{5}$$

The formula for cross-fusing the specific features of two images is as follows:

$$f'^{RD}_{m} = M_s\left(f'^{R}_{m}\right) \otimes M_c\left(f'^{R}_{m}\right) \otimes f'^{D}_{m} \tag{6}$$

$$f'^{DR}_{m} = M_s\left(f'^{D}_{m}\right) \otimes M_c\left(f'^{D}_{m}\right) \otimes f'^{R}_{m} \tag{7}$$

Subsequently, in order to fully integrate the features of different branches, this paper introduces two scalar values, α and β, and further utilizes adaptive addition to generate the shared feature output f'^{S}_{m}. The formula for shared feature output is as follows:

$$f'^{S}_{m} = Concat((\alpha \times f'^{RD}_{m} + \beta \times f'^{DR}_{m}), f'^{S}_{m-1}) \tag{8}$$

The DASM feature output of the mth layer will be associated with the output result of the (m-1)th layer, thereby generating the final shared feature output.

3) Dynamic Feature Fusion Module: In the shared decoder, this paper constructs a multi-level feature fusion module DFM. It improves the accuracy of salient object detection by integrating the rich shared features learned from two image-specific decoders.

The goal of multi-level learning is to extract and integrate information from different levels of abstraction: high-level features have rich abstract semantic information, while low-level features have rich fine-grained information. Therefore, this paper designs a unified multi-level feature fusion module, DFM. It enhances the feature output of the shared encoder by integrating the features learned in the specific decoders of RGB images and depth images into the shared decoder. As shown in Fig. 4, the m-th layer of the shared decoder has the shared feature f'^{S}_{m}, as well as the features f^{R}_{m} and f^{D}_{m} learned by the specific decoders for RGB images and depth images respectively. The entire process is shown in the figure. First, the shared feature f'^{S}_{m} is concatenated with the specific feature f^{R}_{m} of the RGB image at the same

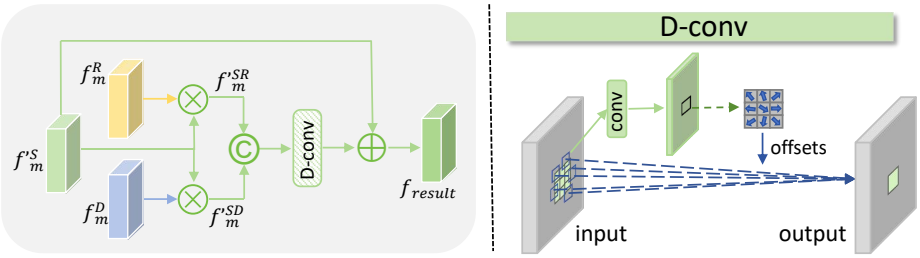

Fig. 4: The proposed a dynamic feature fusion module.

level and the specific feature f_m^D of the depth image, thereby enriching the shared feature representation by leveraging the specific features of the two images. Then, this paper adopts Deformable Convolution (D-conv) [19] to learn the enhanced shared features. By introducing learnable offsets in the receptive field, the sampling points of the convolutional kernel are shifted, making the receptive field more concentrated on the salient object region. After obtaining the output features of the Deformable Conv, the shared features ${f'}_m^S$ of the previous layer are associated through residual addition, thereby outputting the final result of salient object detection. The formula for generating the final result of salient object detection ${f'}_{result}$ is as follows:

$$f_{result} = D - conv(Concat({f'}_m^{SR}, {f'}_m^{SD})) + {f'}_m^S \quad (9)$$

Unlike DASM, DFM enriches the shared features by fusing the specific features extracted from the RGB image and depth image decoders, while DASM forms the shared features by fusing the specific features of the two images in the shared encoder.

IV. EXPERIMENTS

A. Datasets

This paper mainly evaluates the proposed RGB-D salient object detection model on four widely used datasets, including NJU2K, NLPR, SIP and DES. Among them, the training set is composed of 1,485 pairs of RGB-D images from NJU2K and 700 pairs of RGB-D images from NLPR.

B. Evaluation metrics

This paper evaluates the performance of the model through the four most commonly used performance metrics in the field of RGB-D image salient object detection, including: Mean Absolute Error (MAE), F-measure (F_β), S-measure (S_α), and E-measure (E_m). In an RGB-D image salient object detection model, larger F_β, S_α and E_m, as well as a smaller MAE, are expected.

Mean Absolute Error:It is used to calculate the average pixel absolute error between the normalized predicted saliency map and the ground truth map. Its calculation formula is:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |p(x,y) - G(x,y)| \quad (10)$$

Among them, $p(x,y)$ is the saliency result map detected by the model, and $G(x,y)$ is the real saliency result map manually annotated.

F-measure:It is mainly used to evaluate the comprehensive predictive performance of the model. F_β is expressed as the weighted harmonic mean of precision and recall, and its calculation formula is:

$$F_\beta = \frac{\left(1+\beta^2\right) Precision \times Recall}{\beta^2 Precision + Recall} \quad (11)$$

Among them, the balance parameter $\beta^2 = 0.3$.

S-measure:It is mainly used to evaluate the structural similarity between the ground truth map and the predicted saliency map, and its calculation formula is:

$$S_\alpha = \alpha S_o + (1-\alpha) S_r \quad (12)$$

Among them, S_o represents target perception, S_r represents regional perception, and $\alpha \in [0, 1]$ is a weighting parameter, which is set to 0.5 by default.

E-measure::It integrates local pixel evaluation and image-level evaluation, mainly calculating the statistical characteristics of the ground truth map and the predicted saliency map at the image level and the pixel matching degree in local regions. Its calculation formula is:

$$E_m = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi(i,j) \quad (13)$$

Here, W and H represent the width and height of the image respectively. $\phi(i,j)$ indicates a diagonal matrix.

C. Quantitative Comparison

To fully verify the effectiveness of the MLFNet model, this paper compares it with five outstanding salient object detection models, including SPNet [20], SPSN [21], RD3D [22], AFNet [23], and PopNet [24]. The specific results are shown in TABLE I. The model proposed in this paper achieved the best F-measure on all four datasets. For DES, NLPR, and NJU2K, the key metric MAE achieved the best results among the comparison models. The experimental results show that the MLFNet proposed in this paper can effectively fuse the features of RGB images and depth images to achieve the best overall performance in the key MAE and F-measure indicators. Fig. 5 presents the comparison results of the F-measure curves of the MLFNet model proposed in this paper with those of other salient object detection models such as SPNet, SPSN, RD3D, AFNet, and PopNet.

When the threshold is low, the model tends to predict more samples as positive, which may lead to a higher recall rate and a lower precision rate. As the threshold increases, the model becomes more cautious in predicting the positive class, which in turn affects the F-measure value. As shown in Fig. 5, the model proposed in this paper effectively improves the detection accuracy and achieves the best F-measure curve.

TABLE I: Quantitative comparison with RGB-D SOD models.↑ (↓) denotes that the higher (lower) is better. This paper use the mean absolute error (M), F_β, S_m, and E_m as evaluation metrics.

Dataset	DES				NLPR				NJU2K				SIP			
Metric	M↓	F_β↑	S_m↑	E_m↑	M↓	F_β↑	S_m↑	E_m↑	M↓	F_β↑	S_m↑	E_m↑	M↓	F_β↑	S_m↑	E_m↑
SPNet	.017	.939	.935	.970	.022	.928	.928	.958	.033	.927	.918	.944	.047	.912	.887	.923
RD3D	.018	.931	.934	.971	.022	.919	.927	.957	.033	.920	.916	.947	.046	.898	.885	.924
SPSN	.016	.941	.938	.972	.023	.910	.923	.958	.032	.912	.912	.943	.042	.896	.890	.934
AFNet	.022	.920	.922	.948	.020	.925	.936	.968	.032	.928	.926	.958	.043	.909	.896	.939
PopNet	.018	.937	.939	.957	.021	.923	.924	.952	.029	.928	.917	.951	.044	.909	.889	.925
Ours	.016	.943	.931	.963	.020	.930	.931	.960	.029	.933	.927	.954	.044	.916	.896	.931

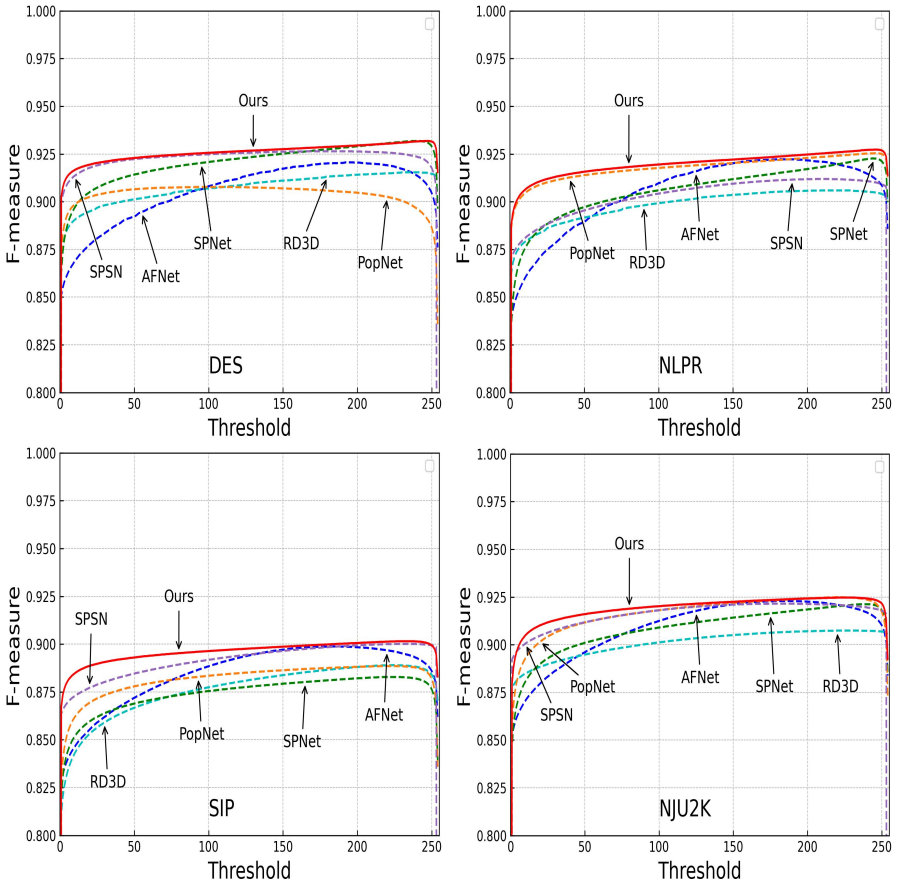

Fig. 5: F-measure curves for different thresholds, for DES, NLPR, SIP, and NJU2K.

D. Ablation studies

Table II presents the results of the ablation experiments of the benchmark model adopted in this paper, along with the addition of the FEM module, DASM module, and DFM module. This paper adopts the key indicators MAE and F-measure for evaluation.

TABLE II: Quantitative evaluation for ablation studies.

				NJU2K	
Baseline	FEM	DASM	DFM	M↓	F_β ↑
✓				0.033	0.927
✓	✓			0.032	0.927
✓		✓		0.031	0.931
✓			✓	0.031	0.929
✓	✓	✓	✓	0.029	0.933

The data in TABLE II indicate that: 1) The FEM module utilizes dilated convolutions with multiple kernel sizes to extract image features, capturing more comprehensive contextual information and thereby enhancing the saliency detection results. 2) The DASM module learns the features of RGB images and depth images through two types of attention, and effectively fuses the features of the two images in an adaptive manner, thereby achieving the highest F-measure index. 3) In the shared decoder, the DFM module employs dynamic convolution to extract salient features and integrates specific features of RGB images and depth images to enrich the shared features, thereby enhancing the saliency detection results. 4) By further integrating the three modules, the model in this paper achieved the best results.

E. Qualitative Comparison

Fig. 6 presents the visualization comparison results of the MLFNet proposed in this paper with RGB-D SOD models such as SPNet, SPSN, RD3D, AFNet, and PopNet on four benchmark datasets. The first column, the second column

Fig. 6: The visual comparison results between the model in this paper and the other models: SPNet, RD3D, SPSN, AFNet, and PopNet.

and the ninth column are respectively the RGB images, depth images and ground truths in the dataset. The closer the visualization results are to the ground truth, the better the detection results of the model are. Columns 3 to 8 respectively present the visualization results of SPNet, SPSN, RD3D, AFNet, PopNet and the model proposed in this paper.

By comparing the visualization results of lines 1 and 2, it can be observed that the model proposed in this paper can effectively locate the salient regions in indoor scenes, thereby detecting complete salient objects. By comparing the visualization results of lines 3 to 4, it can be observed that the salient objects generated by the model in this paper can effectively eliminate the interference of background information. By comparing the visualization results of lines 5 and 6, it can be observed that the model proposed in this paper can describe the hollow parts of the salient objects in detail. By comparing the visualization results in lines 7 to 9, it can be observed that the model proposed in this paper also has a significant effect on multiple salient objects in outdoor scenes.

The experimental results show that: The model in this paper can accurately locate the salient regions, effectively distinguish salient objects from the background, and the generated saliency detection results are closer to the ground truth provided by the dataset. Compared with other models, the model proposed in this paper has good robustness and applicability in complex visual scenes.

V. CONCLUSION

Salient object detection, as an important preprocessing step in computer vision, can effectively filter out the key information in images, and thus has received increasing attention. However, in complex scenarios, the feature information contained in low-quality images often has a negative impact on saliency detection. Therefore, even with spatial information as an aid, the existing methods still struggle to efficiently locate the salient regions. To effectively enhance the accuracy of saliency detection, this paper proposes an end-to-end MLFNet saliency object detection model. It effectively eliminates the influence of low-quality feature information by perceiving multi-level features of RGB images and depth images, thereby significantly improving the accuracy of saliency detection.

Compared with the baseline model, MLFNet captures the context information of the target to be detected in the two images through the multi-scale feature extraction module,

thereby effectively eliminating the differences in target features. Then, through the cross-learning of specific features of RGB images and depth images by channel attention and spatial attention, and the adaptive fusion of the specific features of the two images, the shared features are enhanced. Finally, a dynamic feature fusion module is adopted to further integrate the specific features of RGB images and depth images, thereby enhancing the final saliency detection results. During the experiment, the key indicators were effectively improved. The experimental results show that the MLFNet model proposed in this paper has good stability in various visual scenes. Compared with other RGB-D SOD models, the model proposed in this paper can effectively integrate the specific features of RGB images and depth images, and the generated saliency detection results are closer to the ground truth provided by the dataset.

Although significant achievements have been made in salient object detection that integrates depth information, there are still many aspects that deserve further research: 1) RGB-D SOD algorithms based on deep learning are highly dependent on the quantity and quality of pre-training datasets. To address this issue, the latest imaging and annotation technologies can be adopted to collect high-quality datasets, thereby supporting the research of saliency detection algorithms. 2) Research different supervision strategies, such as unsupervised learning, weakly supervised learning, semi-supervised learning and other algorithms, to reduce the dependence of feature learning networks on data.

REFERENCES

[1] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2075–2089, 2020.

[2] X. Zhao, M. Li, G. Zhang, N. Li, and J. Li, "Object detection method based on saliency map fusion for uav-borne thermal images," *Acta Automatica Sinica*, vol. 47, no. 9, pp. 2120–2131, 2021.

[3] L. Song-Tao, L. Zhen-Xing, and J. Ning, "Target segmentation of infrared image using fused saliency map and efficient subwindow search," *Acta Automatica Sinica*, vol. 44, no. 12, pp. 2210–2221, 2018.

[4] H. Li, R. Yuan, J. Chen, Q. Li, and C. Hu, "Research on double attention mechanism high-resolution network for human pose estimation.," *Engineering Letters*, vol. 33, no. 2, pp. 338–347, 2025.

[5] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3528–3542, 2021.

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

[7] F. Li, J. Zheng, and Y.-F. Zhang, "Depth-guided deformable convolutions for rgb-d saliency object detection," in *2021 6th International Conference on Communication, Image and Signal Processing (CCISP)*, pp. 234–239, IEEE, 2021.

[8] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "rgb-d salient object detection via 3d convolutional neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 1063–1071, 2021.

[9] J. Zhou, L. Wang, H. Lu, K. Huang, X. Shi, and B. Liu, "Mvsalnet: Multi-view augmentation for rgb-d salient object detection," in *European Conference on Computer Vision*, pp. 270–287, Springer, 2022.

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

[11] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "Irfr-net: Interactive recursive feature-reshaping network for detecting salient objects in rgb-d images," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[12] C. Yao, L. Feng, Y. Kong, S. Li, and H. Li, "Double cross-modality progressively guided network for rgb-d salient object detection," *Image and Vision Computing*, vol. 117, p. 104351, 2022.

[13] J. Zhu, "Acfnet: Adaptively-cooperative fusion network for rgb-d salient object detection," *ArXiv Preprint ArXiv:2109.04627*, 2021.

[14] N. Wang and X. Gong, "Adaptive fusion for rgb-d salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019.

[15] X. Wang, T. Sun, R. Yang, C. Li, B. Luo, and J. Tang, "Quality-aware dual-modal saliency detection via deep reinforcement learning," *Signal Processing: Image Communication*, vol. 75, pp. 158–167, 2019.

[16] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[17] N. Liang and W. Liu, "Small target detection algorithm for traffic signs based on improved rt-detr.," *Engineering Letters*, vol. 33, no. 1, pp. 140–147, 2025.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241, Springer, 2015.

[19] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 764–773, 2017.

[20] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, "Specificity-preserving rgb-d saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4681–4691, 2021.

[21] M. Lee, C. Park, S. Cho, and S. Lee, "Spsn: Superpixel prototype sampling network for rgb-d salient object detection," in *European Conference on Computer Vision*, pp. 630–647, Springer, 2022.

[22] Q. Chen, Z. Zhang, Y. Lu, K. Fu, and Q. Zhao, "3-d convolutional neural networks for rgb-d salient object detection and beyond," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 3, pp. 4309–4323, 2022.

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

[24] Z. Wu, D. P. Paudel, D.-P. Fan, J. Wang, S. Wang, C. Demonceaux, R. Timofte, and L. Van Gool, "Source-free depth for object pop-out," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1032–1042, 2023.