Bidirectional Fusion Enhancement in CDINet for Optimized Salient Object Detection

Yuxi Wang*, Yang Xu

Abstract— **Salient object detection has emerged as a critical research direction in image processing technology. To address the challenges of inference efficiency and feature fusion in dual-modal salient object detection, this paper proposes an enhanced multi-modal salient object detection method. Specifically, a SimAM module (a parameter-free attention mechanism) is integrated into the early convolutional layers of CDINet to enhance key feature extraction, thereby improving the subsequent layers' ability to capture critical information. Furthermore, the bidirectional feature fusion (BiFPN) mechanism is adopted in the decoder stage, where deep separable convolution replaces standard convolution to reduce computational load, achieve efficient feature fusion, and improve multi-scale detection capabilities. Lastly, focal loss and weighted cross-entropy loss functions are employed to enhance the model's handling of unbalanced data. In contrast, the intersection-over-union (IoU) loss function is combined to refine boundary prediction accuracy. Experimental results demonstrate that the F-measure of the improved model (SblCDINet) surpasses that of the original CDINet model by 1.15%, 3.93%, 1.01%, 3.10%, and 1.66% on five datasets(DUTS, LFSD, STERE, NLPR, NLU2K), respectively, with a reduction in computational complexity of 19.8%. Compared to other models, the SblCDINet model not only enhances inference efficiency and feature fusion but also effectively reduces computational complexity. The experimental data further corroborate the efficacy and superiority of this approach in multi-modal image saliency target detection.**

Index Terms—**Significance target detection, Computer vision, Attention mechanism, Feature fusion, BiFPN strategy.**

I. Introduction

WITH the rapid development of the Internet and information technology, the generation and usage of image data have increased dramatically. Especially in the field of Salient Object Detection (SOD), thanks to the color independence, illumination invariance, and location uniqueness of Depth images, it provides valuable supplementary information for salient object extraction in complex environments. As the core task of image understanding, Salient Object Detection (SOD) has shown important value in the fields of autonomous driving, augmented reality (AR), and medical image analysis. Recently, the introduction of multimodal data, such as RGB-D images, has provided new research dimensions for SOD tasks. Compared with traditional RGB images, the Depth modality has the characteristics of illumination invariance and uniqueness of spatial structure, which can effectively supplement the target semantic information in complex scenes (such as low light and haze environments) [1]. However, existing RGB-D SOD algorithms still face two major challenges: (1) insufficient mining of high-level semantic information in the process of cross-modal feature fusion; (2) The model complexity increases due to the inefficiency of multi-scale feature interaction. How to construct an efficient and robust cross-modal fusion framework has become the focus of research in this field.

Although RGB-D images have shown significant advantages in salient object detection, it is still a challenge for SOD to accurately segment the boundaries of complex objects in clutched backgrounds, especially when the objects have complex shapes or edges. As a result, multi-modal (e.g. RGB-Depth/RGB-Thermal) fusion algorithms have shown great potential to improve semantic segmentation in complex scenes (e.g., indoor/low-light conditions) [2].

Current RGB-D SOD methods mainly focus on cross-modal feature fusion and attention mechanism. Cross-modal fusion is a technique that integrates information from RGB and depth data to improve the accuracy and robustness of object detection [3].In recent years, attention mechanisms have demonstrated significant effectiveness in capturing crucial differences in feature space and channels across various computer vision tasks [4]. Internationally, Xiao et al. [5] proposed a Deep Guided Fusion Network (DGFNet) to enhance the guiding effect of depth features on RGB modalities through a channel weighting strategy. Although this proves the auxiliary of depth feature information for salient object detection, cross-modal fusion methods mainly rely on low-level features such as color, texture, and edge. Their limited attention to high-level semantic information often leads to poor performance in complex scenarios. With the rise of deep learning, neural network-based saliency detection methods have become a prominent research hotspot. The deep network structure can provide more discriminative semantic features and greatly enhance detection performance in diverse and complex environments. Yuan's team [6] designed a collaborative mechanism between self-attention and cross-attention, which improved cross-modal semantic consistency but caused a surge in computational complexity due to multi-level feature stacking. Chen et al. [7] used a context-based attention mechanism to dynamically adjust the contribution optimization fusion process of RGB and depth modalities, which improved the robustness of cross-modal object detection, but it may be difficult to transfer to different tasks depending on the training equipment. Domestic scholars have also made remarkable progress

Manuscript received March 4, 2025; revised Jun 5, 2025.

This work was supported by the National Natural Science Foundation of China (61775169), the Education Department of Liaoning Province (LJKZ0310), the Excellent Young Talents Program of Liaoning University of Science and Technology (2021YQ04)

Yuxi Wang is a postgraduate student of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (corresponding author to provide phone: 86-18524332357; e-mail: 2692769200@qq.com).

Yang Xu is a Professor at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (phone: 86-13889785726; e-mail: xuyang 1981@aliyun.com).

in this field: The Cross-modal Differential Interaction Network (CDINet) proposed by Zhang et al. [8] realizes the differential interaction between modalities through the bidirectional induction module (RDE/DSE), and achieves SOTA performance on the public data set. However, its Dense Decoding and Reconstruction (DDR) structure has parameter redundancy, which restricts the real-time application. In addition, the existing methods generally use binary Cross Entropy (BCE) loss function, which makes it difficult to solve the class imbalance problem in complex boundary regions. The above studies show that the existing models have not yet achieved an effective balance between feature fusion efficiency, computational complexity, and boundary accuracy.

To solve the above problems, this paper proposes an enhanced RGB-D SOD model SblCDINet, whose innovation is mainly reflected in the following three aspects: Firstly, the DDR module is reconstructed based on BiFPN (Bidirectional Feature Pyramid Network), and the standard convolution is replaced by depthwise separable convolution, which reduces the computational complexity of the model by 19.8% while maintaining the ability of multi-scale feature expression. Secondly, the SimAM (Simple Paramet-free Attention Module) module is embedded in the early convolutional layer, and the cross-modal key features are dynamically strengthened through three-dimensional attention weights, which significantly improves the detection accuracy of small targets and complex boundaries (F-measure is increased by 1.44%). Finally, Focal Loss, Weighted Cross Entropy (WBCE), and IoU Loss were fused, the dynamic weight allocation strategy ($\lambda_1 + \lambda_2 + \lambda_3 = 1$) was used to balance the class sensitivity and boundary matching degree, and the MAE index was reduced by 0.5% on LFSD and other datasets. Experiments show that SblCDINet outperforms existing models on five benchmark datasets, such as DUTS and NJU2K, and its inference speed reaches 153 FPS (256×256 resolution), which provides a new solution for real-time saliency detection. This study not only promotes the development of lightweight multimodal fusion theory but also lays a technical foundation for practical applications such as autonomous driving environment perception and AR/VR scene reconstruction.

II. Related Work

CDINet combines RGB modality in different ways of unidirectional and bidirectional interactions, emphasizing that the interaction between the two modalities should be carried out in an independent and differentiated manner. Low-level RGB features can help Depth features distinguish different objects in the same depth range, while high-level Depth features can further enrich RGB semantics and effectively suppress the interference of complex backgrounds. On this basis, CDINet proposes the RGB-induced Detail Enhancement (RDE) module, as shown in Fig. 1. This module achieves the fusion of two modal visual features by a two-layer cascade convolution to generate the fusion feature pool FPOOL, as shown in Equation 1. Where $i \in \{1, 2\}$ denotes the underlying coding layer feature layer, $[f_r^i, f_d^i]$ denotes the channel level splicing operation of RGB features and depth features, and $convN()$ denotes a convolutional layer with a kernel size of N×N.

$$f_{pool}^i = conv3(conv1([f_r^i, f_d^i])) \tag{1}$$

For Depth features, a spatial attention template is generated through a series of operations, and finally mask is multiplied with the feature pooling information to reduce the interference of irrelevant RGB features to obtain the required supplementary information from the depth modality. The whole process can be described as Equation 2:

$$f_{out}^i = \sigma(conv7(cprv7(\max pool(f_d^i))))^* f_{pool}^i + f_d^i \tag{2}$$

Where $\sigma()$ and $\text{maxpool}()$ denote maxpool operation in channel dimension and sigmoid function, respectively, and * denotes element-by-element multiplication. The feature f_{out}^i will be used as input to the next layer of the depth branch. By using the splicing operation instead of passing the RGB features directly to the depth branch, the common detail information between the two modalities can be enhanced while weakening the irrelevant features.

In the Depth-induced Semantic Enhancement (DSE) module, the weight vectors are learned through a global average pooling (GAP) layer, two fully connected layers (FC), and a sigmoid function as shown in Equations 3 and 4:

$$C_{weight} = \sigma(FC(GAP(f_{rs}^i))) \tag{3}$$

$$D_{att}^i = C_{weight} \times f_{rs}^i \tag{4}$$

In the Dense Decoding Reconstruction (DDR) module, as shown in Fig.2. The features f_{out}^i generated at each layer in the encoding phase constitute a list of jump connections, which are labelled as $f_{skip}^i (i \in \{1, 2, 3, 4, 5\})$ for easy differentiation. Semantic block B is generated using higher-level encoder features to constrain the jump connection information of the current corresponding encoder layer. This design enhances the effectiveness of feature fusion and ensures that the decoding process can make full use of the rich contextual information The semantic block B is defined as follows:

$$B^i = conv3(conv1([up(f_{skip}^{i+1}), \cdots, up(f_{skip}^5)])) \tag{5}$$

Where $up()$ denotes the up-sampling operation by bilinear interpolation. The obtained f_{skip}^i combines the decoded features of the previous layer and gradually restores the image details by up-sampling and successive convolution operations. Finally, the decoded features of the last layer are used to generate the predicted saliency map by the sigmoid activation function.

III. Improvement strategies for the algorithmic model

In this article, two primary optimizations are implemented to address the various metrics of the detection of RGB-D image saliency, utilizing the characteristics of the CDINet algorithm: optimization of the DDR module and adjustment of the loss function. The CDINet decoder is tasked with decoding the depth features extracted by the encoder to achieve effective saliency detection. The first part focuses on optimizing the DDR dense reconstruction decoder. As the fundamental feature extractor for the detection task, the

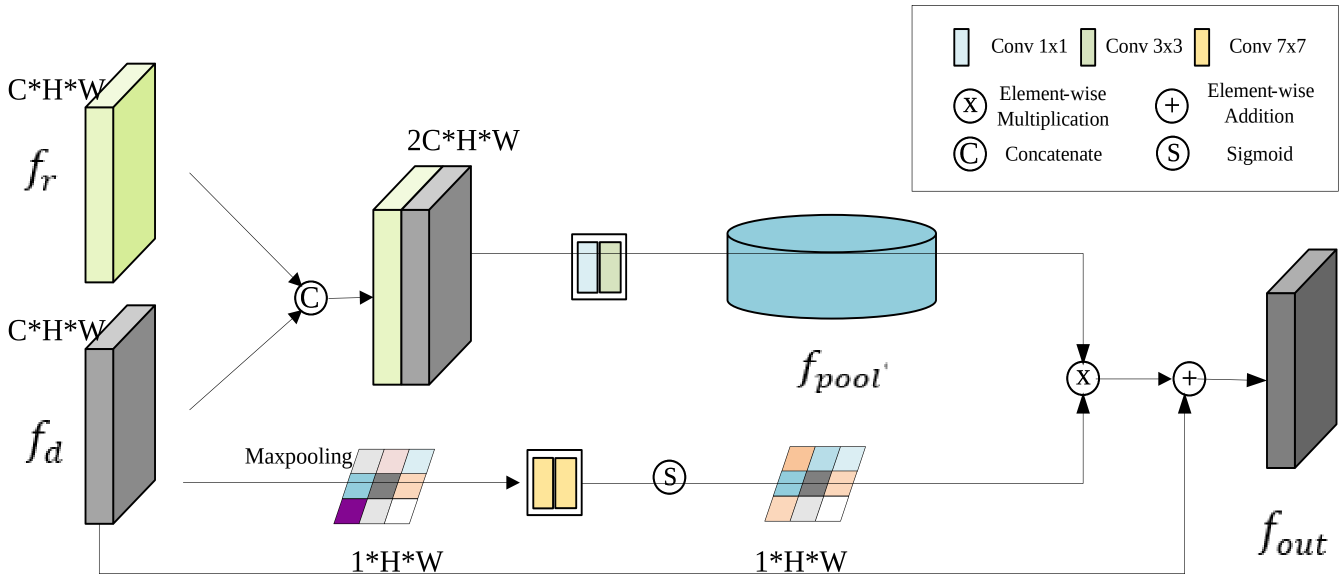


Fig. 1. The Architecture for RGB-induced Detail Enhancement.

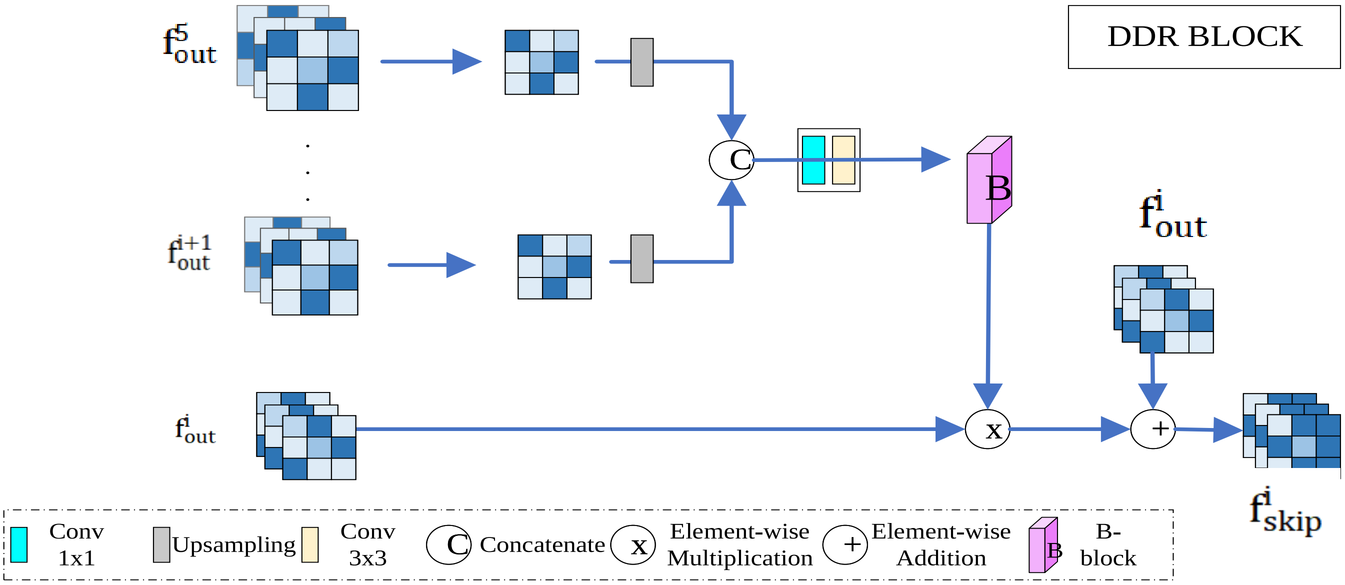


Fig. 2. The architecture of Dense Decoding Reconstruction.

decoder's primary role is to map high-dimensional encoder features to lower-dimensional outputs while preserving as much critical information as possible. To enhance the decoder's performance, this paper incorporates a bidirectional feature pyramid network (BiFPN) into the design, thereby improving the feature fusion effect. The enhanced model structure, referred to as SblCDINet, is illustrated in Fig. 3, showcasing the refined design of the DDR module and its application in the saliency detection task.

A. The SimAM module is improved to enhance feature extraction

In deep learning and Convolutional Neural Networks (CNN), the attention mechanism has been proven to be an important tool to improve model performance. However, many attention modules introduce additional parameters or computational complexity, especially in the channel or spatial dimension, which poses challenges for lightweight models deployed in resource-constrained environments. To solve this problem, a simple parameter-free attention module (SimAM [9]) is integrated into the improved CDINet model to enhance feature interaction. Through the analysis, the RDE block in the CDINet model focuses on enhancing the interaction of RGB and depth features. However, the observation of Fig.2 in Chapter II of this paper shows that the attention mechanism that can selectively focus on important spatial regions or channels of the feature map is missing in this block. As shown in Fig.4(a),(b), traditional attention mechanisms usually compute one-dimensional channels or two-dimensional spatial weights. In contrast, SimAM estimates 3D attention weights (Fig. 4c), enabling finer-grained control over feature maps by simultaneously modeling spatial and channel dimensions. In this paper, we adopt the SimAM module to improve the feature

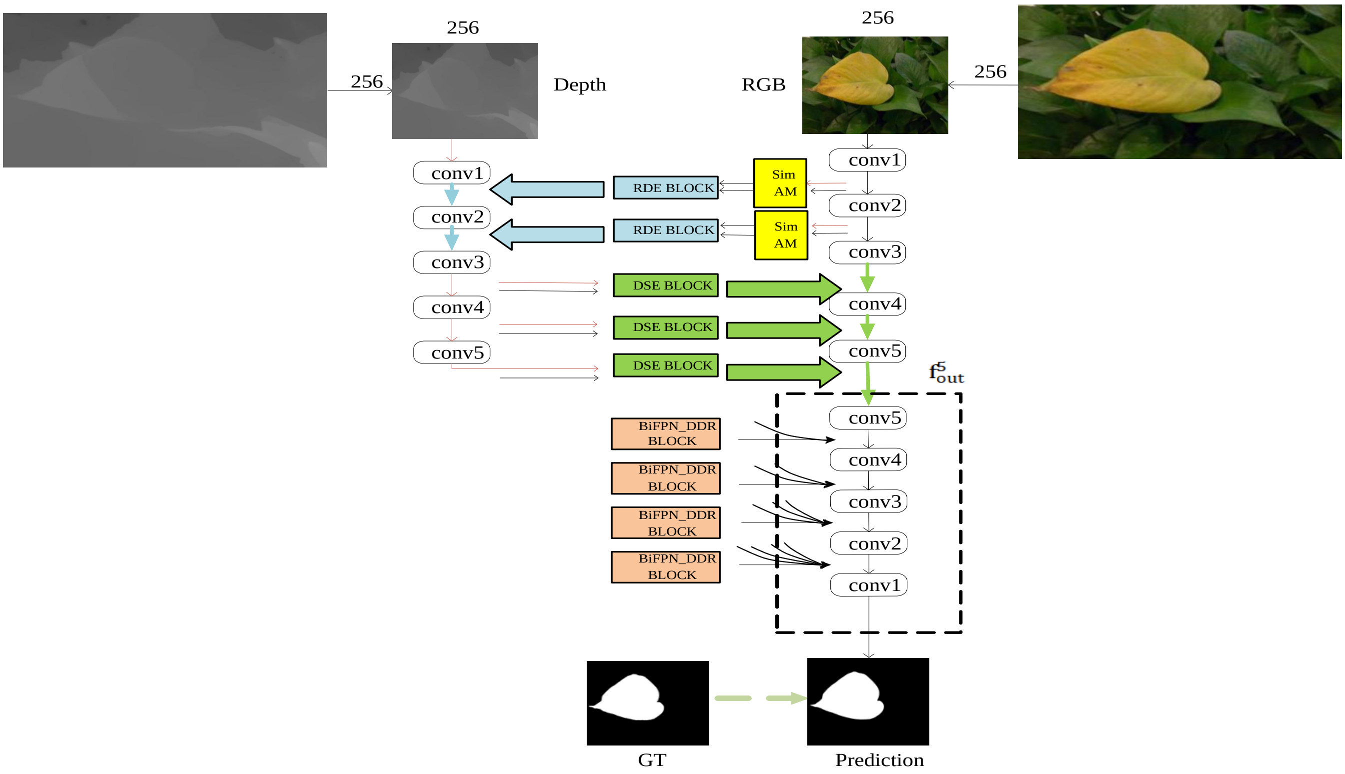


Fig. 3. Schematic diagram of the overall structure of SblCDINet.

representation ability of the model in the RGB-D SOD task, while keeping the computational cost low. Specifically, SimAM first applies global average pooling over the channel dimension of the entire feature map to capture global statistics. Subsequently, the interaction relationship between channels is calculated by a one-dimensional convolution operation, which combines information from both spatial and channel dimensions to generate the final 3D attention weights. These weights are normalized by a Sigmoid activation function and used to adaptively reweight the original feature map to enhance the expressive power of key features.

Compared with the traditional attention mechanism, the core advantage of SimAM is that it can effectively measure the importance of each feature in spatial and channel dimensions without additional parameters or large-scale calculations. This mechanism is especially suitable for the RDE (RGB-Depth Enhancement) module. In the process of fusion of RGB and Depth information, SimAM can highlight the information expression of salient regions and improve the ability to capture the boundary and detail of the object by modeling the difference and complementarity of the features. In addition, the calculation method of SimAM is optimized, the power calculation is replaced by the square operation (i.e. x^2), and the eigenvalues are processed by a two-step smoothing term (defined as the smoothing function) to $S(x) = \frac{x}{1+x}$ enhance the numerical stability. Experimental results show that after integrating the improved SimAM, the performance of the model is improved on multiple RGB-D SOD benchmark datasets, while maintaining the lightweight advantage in computational overhead.

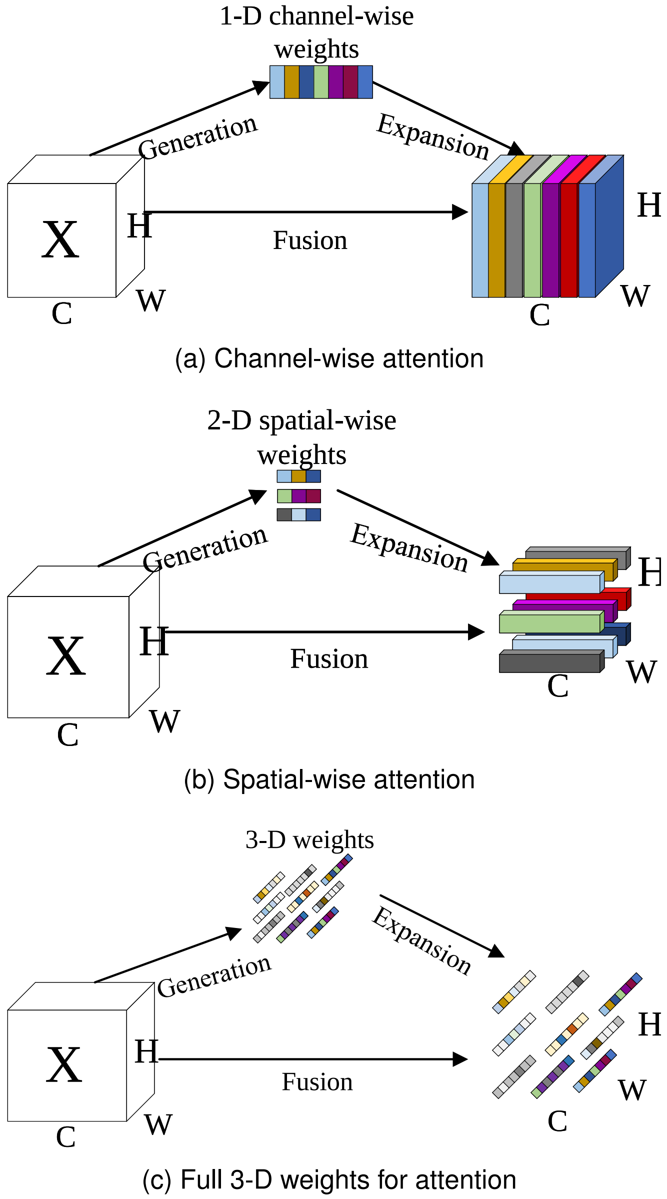
B. Bi-directional multi-scale feature fusion of BiFPN module is improved

Bidirectional Feature Pyramid Network (BiFPN [10], Bidirectional Feature Pyramid Network) introduces an efficient method for dealing with multi-scale features by enabling bi-directional connections between neighboring layers in the feature pyramid. Unlike traditional FPNs that contain only top-down feature refinement paths, BiFPN allows information to flow in two directions top-down and bottom-up, and uses learnable weights to ensure optimal feature fusion, as shown in Feature Fusion Equation 6. This approach enhances the ability of the model to aggregate features from different levels, resulting in a more efficient multi-scale feature representation.

$$f_{filsed} = \omega_1 \times f_{top-bottom} + \omega_2 \times f_{bottom-up} \quad (6)$$

Where $f_{top-bottom}$ is the top-down feature map, $f_{bottom-up}$ is the bottom-up feature map, ω_1 and ω_2 are learnable parameters, which are automatically optimized by back-propagation. Specifically, a differentiable weight fusion mechanism is adopted, which dynamically adjusts the weights by minimizing a loss function during the training process without manual setting. The weights are initialized with a uniform value of 0.5 and eventually converge to the optimal balance of contributions to multi-scale features, the weight coefficients.

In the CDINet model, inspired by the multi-scale feature fusion of BiFPN, a lightweight version of this module is integrated into the Dense Decoding Reconstruction (DDR) process, mainly to improve the multi-scale feature processing and fusion capability and efficiency in the decoding stage, which optimizes the model's training speed and computational resource shown in Fig.5, the



(a) Channel-wise attention

(b) Spatial-wise attention

(c) Full 3-D weights for attention

Fig. 4. Comparison of different attention steps. (a) one-dimensional channel weight (b), two-dimensional spatial weight (c), and 3D attention weight of SimAM.

BiFPN DDR block utilizes depth-separable convolution ($DepthwiseConv()$) to efficiently extract spatial features as shown in Equation 7, followed by point-by-point convolution ($PointwiseConv()$) to adjust the number of channels for each input feature map as shown in Equation 8. These operations allow the model to perform feature refinement across multiple scales while significantly reducing computational cost.

$$f_{depthwise} = DepthwiseConv(f_{in}, k_{3\times3}) \tag{7}$$

$$f_{pointwise} = \text{PointwiseConv}(f_{depthwise}, k_{1\times1}) \tag{8}$$

$$f_{BiFPN} = Upsample(f_{pointwise}) \tag{9}$$

Where f_{in} denotes the input feature map, $k_{N\times N}$ is the convolution kernel of $N \times N$ size, and $Upsample()$ denotes the up-sampling operation via bilinear interpolation. By adding the multi-scaleup-sampling operation Equation 9 to the DDR, the BiFPN layer can refine and reconstruct the feature map layer by layer, which makes the final output feature map optimized in terms of both spatial resolution and semantic information. Experiments have demonstrated that after introducing the multi-scale feature fusion mechanism of BiFPN into the dense decoding and reconstruction (DDR) module, the whole network not only improves the fusion ability of multi-scale features but also refines the features with different resolutions in the decoding process. This design is especially critical for multi-scale target localization in saliency target detection tasks.

$$f_{final} = Conv3(Concat(f_{BiFPN}, f_{lateral})) \tag{10}$$

Thus, the final output f_{final} is shown in Equation 10, where $f_{lateral}$ the feature map from the neighboring layers, $Concat()$ is the splicing operation at the channel level, and $ConvN()$ is the convolution operation with a convolution kernel size of N. The more effective optimization of SblCDINet is reflected in the light weight of the model and the improvement of the computational efficiency. BiFPN significantly reduces the number of parameters and FLOPs (floating point operation counts) through the combination of deeply separable convolution and point-by-point convolution, which drastically reduces the amount of computation under the premise of guaranteeing the performance of the model. In addition, the up-sampling operation adopts bilinear interpolation to gradually restore the feature map to its original size, which effectively improves the detailing performance of the reconstruction. With this design, the model achieves a good balance between multi-scale feature processing and decoding efficiency. The incorporation of this modified BiFPN into the DDR module in this paper has had a profound impact on the performance of the CDINet model. Due to the reduction in computational complexity and more efficient use of resources, the training time for 100 epochs is reduced from around 5 hours to around 3.6 hours. In addition, both the parameter counts and ELORs (floating point operations) of the model are significantly reduced without affecting the performance. As a result of the experimental exploration, the improvement of the DDR module highlights the advantages of using the BiFPN style multi-scale feature fusion in improving the efficiency and accuracy of the saliency target detection task.

C. Optimization of the loss function

The original model employs the Binary Cross Entropy (BCE) loss function, which can result in ambiguous boundaries during training. To achieve clearer boundaries in the saliency map and enhance the robustness and detection accuracy of the model, a hybrid loss function is utilized. Traditional BCE loss tends to bias training towards the majority class in imbalanced datasets; therefore, Focal Loss [11] is introduced to mitigate the impact of easily classified samples and emphasize the contribution of challenging samples. Furthermore, the Weighted Binary Cross-Entropy (WBCE) loss function [12] is applied to diminish the influence of an excessive number of negative samples on the optimization process, ensuring the model focuses more on salient regions. Additionally, to improve the shape matching of the predicted mask, IoU Loss is adopted, directly optimizing the intersection between the predicted and true masks, thereby enhancing overall detection accuracy and

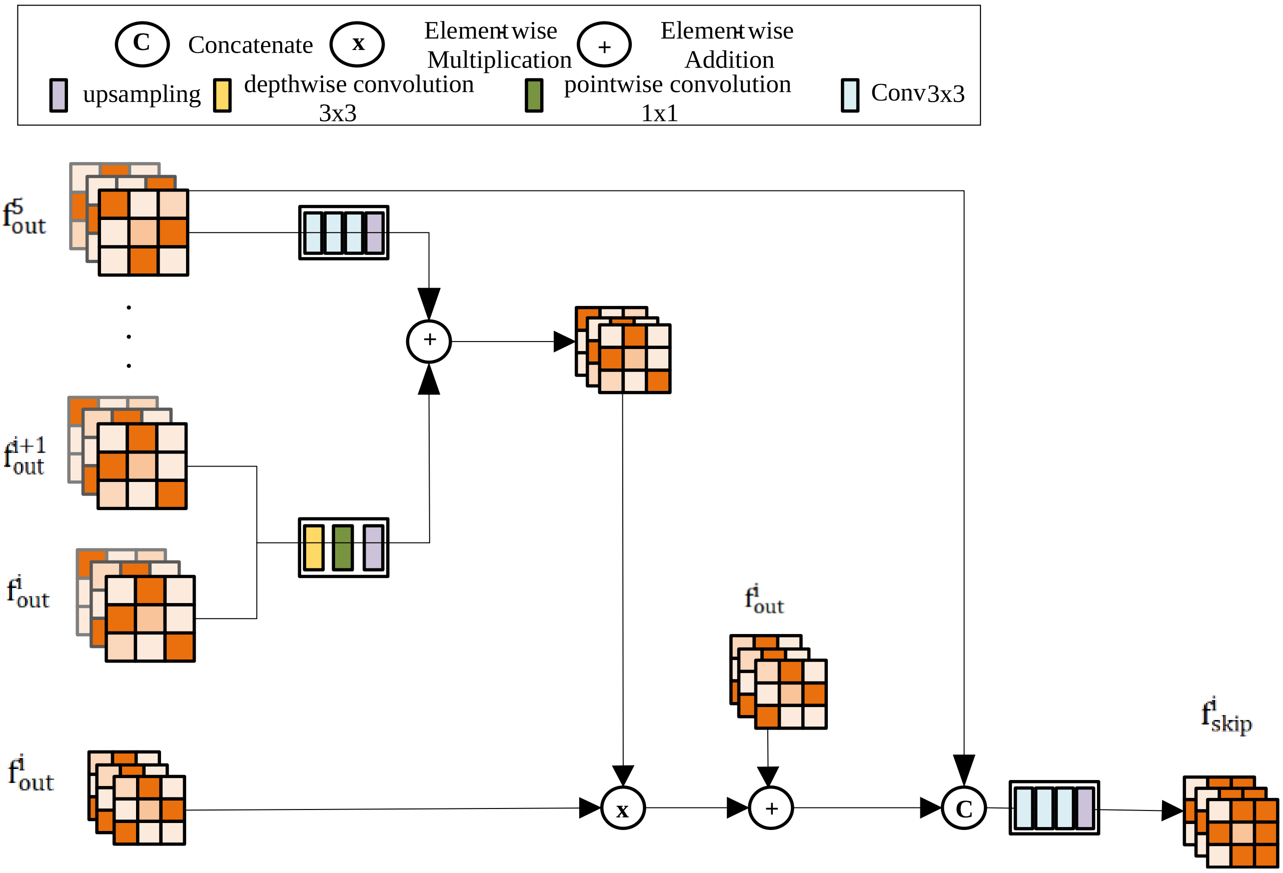


Fig. 5. The architecture of BiFPN Dense Decoding Reconstruction.

edge quality.

(1) Focal Loss function (Focal Loss)

Focal loss function (Focal) is proposed to solve the problem of category imbalance, where α is the weights to balance the positive and negative samples and is the focus parameter to regulate the weights of easy-to-category samples. Here the focal loss is calculated for each sample and then averaged. The formula is expressed as:

$$Loss_{fl} = -\alpha(1 - p_i)^{\gamma} \log(p_i) \tag{11}$$

(2) Weighted Binary Cross-Entropy Loss function (Weighted Binary Cross-Entropy Loss)

Binary Cross Entropy (BCE) loss is a very widely used loss in binary classification and segmentation, n is the number of samples, y_i is the true label of the i'th sample, and $p(y_i)$ is the probability that the model predicts a positive class. BCEWithLogitsLoss is used here instead of BCELoss because the former combines the Sigmoid activation function and the binary cross-entropy loss, which can be more stable. The pos-weight parameter is used to deal with the problem of category imbalance, and its value is usually set to be the inverse of the ratio of the number of positive samples to the total number of samples. Its formula is expressed as:

$$Loss_{BCE} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(p(y_i)) + (1 - y_i)\log(1 - p(y_i)) \tag{12}$$

(3) Intersection and integration ratio loss (IOU Loss)

In target detection networks, target localization heavily relies on a module that performs bounding box regression [13]. The Intersection and Merger Ratio Loss (IOU) was originally used as a measure of similarity between two sets and later used as a standard evaluation metric for target detection and segmentation. Also known as Jaccard loss, it is used to measure the similarity between predicted and true regions. Where $\hat{y}$ is the predicted mask, y is the true mask, $Intersection()$ denotes the intersection of the two, and $Union()$ denotes the concatenation of the two. Here the predictions are converted into probabilities by the Sigmoid function, then intersection and concatenation are calculated, and finally, the loU loss is calculated. The formula is given below.

$$Loss_{IoU} = 1 - IoU = 1 - \frac{\text{Intersection}(\hat{y}, y)}{Union(\hat{y}, y)} \tag{13}$$

During training, these three loss functions were combined to train the model with one total loss. This combination utilizes the strengths of each loss function to improve the accuracy of the saliency effect map, with a particular focus on hard-to-detect regions through the use of a focal loss function and accurate boundary prediction using the IOU loss function, as shown in Equation 14:

$$\lambda_1 Loss_{fl} + \lambda_2 Loss_{BCE} + \lambda_3 Loss_{IoU} = NewLoss \tag{14}$$

Where $\lambda_1, \lambda_2, \lambda_3$ is a dynamically allocated positive integer($\lambda_1 + \lambda_2 + \lambda_3 = 1$). In the experiments, we evaluated the impact of different combinations of loss function weights $(\lambda_1, \lambda_2, \lambda_3)$ on model performance. The results indicate that when the weight of Focal Loss (λ_1) is excessively high, its influence is amplified, making the model more sensitive to difficult samples but potentially leading to overfitting. When the weight of WBCE Loss (λ_2) is too large, it enhances the model's focus on foreground regions, which may compromise its ability to suppress background regions. An overly high weight for IoU Loss (λ_3) improves mask shape optimization but might degrade classification performance. Based on validation set experiments, we adjusted the hyperparameters such that their sum equals 1 and selected the optimal weight combination to achieve the best F-measure and overall performance. In this study, we adopted a combined strategy of Focal Loss, WBCE Loss, and IoU Loss. The primary advantages of this approach are as follows: Focal Loss addresses class imbalance, WBCE Loss enhances learning of salient objects, and IoU Loss improves mask shape matching quality. By reasonably allocating the weights, the loss function optimizes both classification accuracy and mask shape, thereby consistently improving model performance across multiple datasets and enhancing the overall performance of SOD tasks.

IV. Analysis and discussion of experimental results

A. Introduction to the dataset

In the field of saliency target detection, with the continuous progress of technology, numerous high-quality RGB-D datasets have been proposed to validate and enhance the performance of detection models. In this paper, in the experimental stage, the model is trained and evaluated on five widely recognized public datasets as follows:

(1) DUTS dataset [14]: this dataset covers 1200 images containing both indoor and outdoor complex scenes and provides the corresponding depth information image for each image.

(2) NLPR dataset [15]: this dataset consists of 1000 pairs of images, including RGB images of indoor and outdoor scenes such as supermarkets, campuses, streets, and their corresponding depth images.

(3) NJU2K dataset [16]: this dataset contains 1985 RGB images and their corresponding depth images from the web, 3D movies, and shots from the Fuji W3 camera.

(4) STEREO dataset [17]: this dataset consists of 797 stereo images, which were collected from an online image library, and depth maps were predicted by analyzing the left and right views.

(5) LFSD dataset [18]: this dataset contains 100 RGB-D images captured by the Lytro 1 light-field camera with manually labeled precision data.

Out of these datasets, a total of 2985 RGB-D images are selected as the training set in this paper, which is assigned as 1485 images from the NJU2K dataset,700 images from the NLPR dataset, and800 images from the DUTS dataset. The test set, on the other hand, consists of the remaining images from the DUTS, NLPR, NJU2K, STEREO, and LFSD datasets. Such a dataset assignment aims to ensure that the model can be adequately trained through diverse scenarios and effectively evaluate its performance on multiple different datasets. In this paper, three commonly used standard evaluation metrics are used for quantitative evaluation: structural measure [19]($S-\text{measure}$), F-measure [20] (F_β, where β=0.3), and mean absolute error (MAE). To ensure a fair comparison, this paper used either reported results in other papers or reproduced results under the same recommended data setting.

B. Model implementation details

The experimental setup involves both hardware facilities and software configuration. The hardware platform used in this study is equipped with two NVIDIA GeForce RTX 2080 Ti GPUs, each with 45 GB of graphics memory. On the software side, all experiments are based on the VGG16 model pre-trained on the ImageNet dataset. In this paper, the Adam optimizer is chosen to initialize the backbone parameters of the model, and the initial learning rate is set to 1e-4. During the training process, 100 training cycles (epochs) are set, and 4 images are processed in each batch (batch size of 4). In addition, every 40 epochs, the learning rate is decayed by a factor of 5.

The entire model is trained in an end-to-end manner without any preprocessing steps. After about 3.63 hours of training, a final model with 100 epochs completed can be obtained. In the testing phase, using the above GPU configurations, inference is performed on images of size 256 x 256 pixels with an average processing speed of 153 FPS and a standard deviation of 15 FPS. These parameters and settings ensure the high efficiency of the experiments and the accuracy of the model.

C. Analysis of experimental results

To validate the impact of the proposed modules (SimAM, BiFPN-DDR, and hybrid loss function), we conducted comprehensive experiments. Table I presents the ablation study results, while Table II compares the performance of SblCDINet with state-of-the-art saliency detection models.

To verify the effectiveness of SblCDINet, this paper evaluates the independent contributions of the optimization modules (SimAM, BiFPN-DDR, and hybrid loss function) by gradually splitting them. The first row in Table I shows the MaxF values of the baseline model (AbsoluteAbs-CDINet) on the DUT, LFSD, and STERE datasets in this experimental environment are 0.9281, 0.8676, and 0.8798, respectively, and the MAE values are 0.0314, 0.0708, and 0.040,9, respectively. It is used as a reference to calculate the improvement range. On DUT and STERE datasets, after the introduction of the SimAM module, MaxF is increased to 0.9307 (+0.28%) and 0.8813 (+0.17%) respectively, but MAE is increased to 0.0323 (+2.87%), and 0.0690 (+19.8%). It indicates that SimAM may reduce the generalization of simple scenes due to excessive attention to complex features. However, on the LFSD dataset, the MaxF is increased to 0.8813 and the MAE is reduced to 0.0690, which verifies the effectiveness of SimAM in complex boundary detection.

After the DDR part of the CDINet model fuses the BiFPN mechanism, on the DUT dataset, the MaxF is increased to 0.9332, and the MAE is reduced to 0.0305, indicating that the bidirectional multi-scale fusion strategy effectively reduces parameter redundancy through depthwise separable convolution, while preserving low-level details. On the LFSD dataset, MaxF is increased to 0.8755 (+0.91%), but MAE is increased to 0.0714 (+0.85%), which reflects that the fusion of multi-scale features in deep fuzzy regions needs to be

TABLE I
COMPARISON OF ABLATION EXPERIMENTS ON DUT AND LFSD DATASETS

Composition	DUT			LFSD			STERE		
	MaxF ↑	MAE ↓	S-measure ↑	MaxF ↑	MAE ↓	S-measure ↑	MaxF ↑	MAE ↓	S-measure ↑
Acu-baseline-CDINet	0.9281	0.0314	0.9212	0.8676	0.0708	0.8587	0.8798	0.0409	0.9031
baseline-CDINet+SimAM	0.9307	0.0323	0.9198	0.8813	0.0690	0.8635	0.8813	0.0490	0.9035
baseline-CDINet+BiFPN_DDR	0.9332	0.0305	0.9206	0.8755	0.0714	0.8614	0.8895	0.0314	0.8614
baseline-CDINet+Loss	0.9310	0.0314	0.9191	0.8677	0.0651	0.8602	0.8677	0.0351	0.9002
baseline-CDINet+SimAM+ BiFPN_DDR+Loss (ours)	**0.9388**	**0.0283**	**0.9221**	**0.8820**	**0.0655**	**0.8654**	**0.9069**	**0.0374**	**0.9033**

further optimized. On the STERE dataset, the MaxF reaches 0.8895 and the S-measure reaches 0.8614, which verifies the robustness of BiFPN-DDR for multi-object detection in stereo scenes.

Experimental results show that after adding SimAM, BiFPN-DDR and a hybrid loss function to CDINet, the MaxF of the complete improved model (SblCDINet) on DUT, LFSD, and STERE datasets reached 0.9388 (+1.15%), 0.8820 (+1.66%), and 0.9069 (+3.10%), respectively. MAE decreased to 0.0283 (-9.87%) and 0.0655 (-7.49%), and S-measure increased to 0.9221, 0.8654, and 0.9033 synchronously. The synergistic effect of Multi-scale feature fusion (BiFPN-DDR) and attention mechanism (SimAM) can significantly enhance the detection ability of the model for complex boundaries and small objects. The hybrid loss function effectively alleviates the class imbalance problem and optimizes the boundary accuracy of the saliency map through the dynamic weight allocation strategy. SblCDINet achieves robust saliency detection in multiple scenes while maintaining high inference efficiency (153 FPS), which provides a reliable solution for autonomous driving and AR/VR applications.

The ablation experiments quantify the contribution of the individual optimization modules and demonstrate their effect on the overall performance improvement. Fig. 6 demonstrates the detection results of the improved model SblCDINet on the public test set. It can be seen that the improved model performs well in small object detection and complex background processing, and can accurately identify salient objects. In particular, the images in the second row and third column in Fig.6 show that the SblCDINet proposed in this paper is able to successfully segment the subtle gaps in the character's hand when he/she puts his/her hand in his/her pocket, thanks to the spatial information provided by the depth map, which is not even reflected in the real labels. It is also demonstrated in Fig.6 that the improved model of this paper is also able to recognize better in the case of multiple targets, which fully demonstrates the clarity and meticulousness of the improved model in dealing with the edges of salient objects. Therefore, it can be concluded that depth information is particularly effective in providing spatial information and texture-free foreground-background separation, which helps in SOD-related tasks. In order to demonstrate the superiority of the SblCDINet algorithm more comprehensively, this paper compares it with 10 state-of-the-art and representative saliency target detection models. The tests were conducted on three datasets, namely, LFSD, NLPR, and STERE, and the detection performance of the models was measured using the F-measure, the MAE, and the S-measure as evaluation metrics.

The decreasing trend of the loss value (Loss)during the training process of the two different models is shown in Fig. 7. The blue curve represents the loss value of the original CDINet model, and the orange curve represents the improved model in this paper. It can be observed from the figure that compared to the CDINet model, the loss value of the improved model decreases faster in the early stage of training, which indicates that the improved model can learn effective features faster in the early stage and has better initial convergence. In the middle and late stages of training, the loss value of the improved model tends to be stable, and the overall loss value is lower, which indicates that the training process of the model is more stable, and it is not easy to have oscillation or overfitting: The loss value of the improved model is always lower than that of the CDINet model throughout the training process, which indicates that the improved model has a better fitting effect and generalization performance on this task.

Table II shows the performance comparison between the SblCDINet model and the current mainstream saliency detection models on LFSD, NLPR, and STERE datasets. The results in Table 2 show that the accuracy of SblCDINet on LFSD, NLPR, and STERE datasets reaches 88.2%, 92.6%, and 90.7%, respectively. Compared with the benchmark model Acu-CDINet, the accuracy is relatively increased by 1.6%, 1.2%, and 0.9%. MAE decreased by 7%, 4.2% and 7.3%. The S-measure increases by 0.7%, 0.2%, and 0.2% relative to the baseline model. These improvements benefit from the innovation of SblCDINet in feature extraction and fusion methods. The synergistic effect of the Multi-scale Feature Fusion module (BiFPN-DDR) and the 3D Attention Mechanism (SimAM) significantly enhances the semantic understanding ability of the model for complex scenes, thus achieving stable performance improvement in multiple datasets.

The proposed algorithm has better performance than other algorithms in the three data sets. The experimental results show that the optimization algorithm significantly improves the effectiveness of the model by introducing a multi-layer feature fusion module, an attention mechanism, and an adjusted loss function. In the comparative tests on multiple data sets, the advantages of the improved scheme have been fully verified, and its effect is demonstrated by visual means, as shown in FIG. 8. It can be seen from the figure that the proposed method can more clearly detect salient objects in various application scenarios, especially when the discrimination between the saliency map and the background is low, and it is superior to other algorithms.

The algorithm in this paper performs better relative to other algorithms in all three datasets. Several experimental results show that the optimized algorithm significantly improves the effectiveness of the model by introducing a

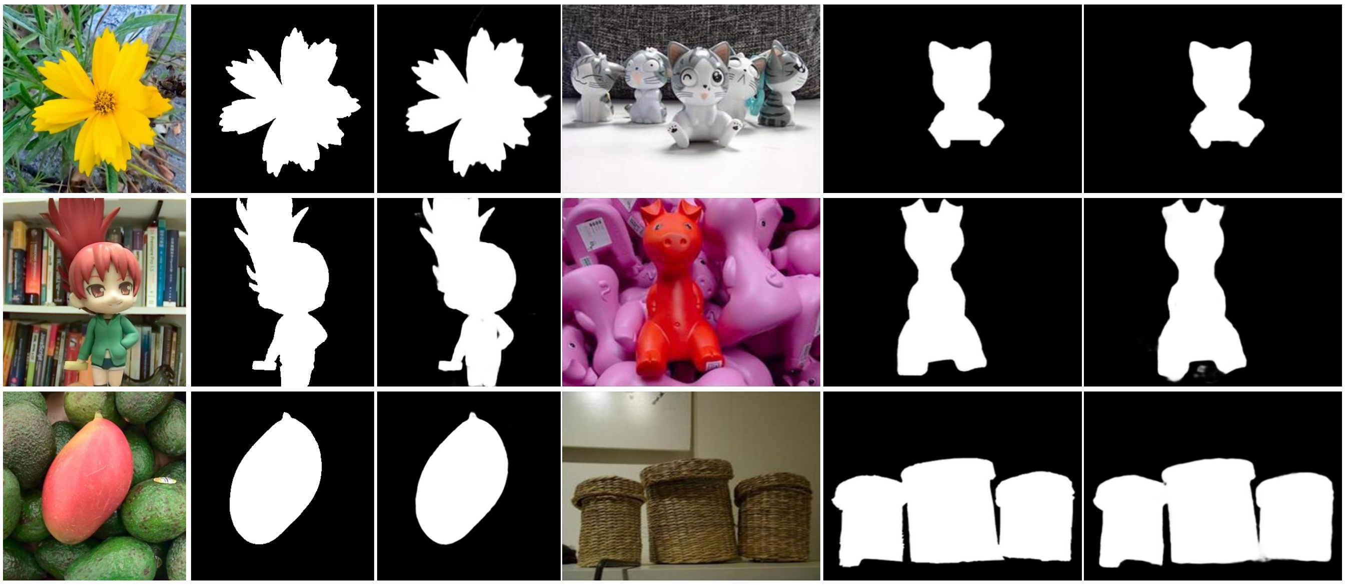


Fig. 6. Test effect diagram.

TABLE II
COMPARISON OF EXPERIMENTAL RESULTS ON LFSD, NLPR, AND STERE DATASETS

Model	Venue	LFSD			NLPR			STERE		
		maxF↑	MAE↓	S-m↑	maxF↑	MAE↓	S-m↑	maxF↑	MAE↓	S-m↑
JL-DCF [21]	CVPR	0.821	0.103	0.817	0.891	0.029	0.909	0.874	0.050	0.885
PGAR [22]	ECCV	0.839	0.081	0.844	0.915	0.024	0.929	0.900	0.042	0.905
DANet [23]	ECCV	0.841	0.103	0.837	0.901	0.028	0.915	0.819	0.071	0.841
D3Net [24]	TNNLS	0.806	0.102	0.816	0.896	0.029	0.911	0.849	0.057	0.868
ASIFNet [25]	TCyb	0.860	0.080	0.852	0.890	0.029	0.907	0.880	0.048	0.882
Acu-CDINet	ACM	0.868	0.071	0.859	0.915	0.024	0.925	0.899	0.041	0.903
Ours	—	**0.882**	**0.066**	**0.865**	**0.926**	**0.023**	**0.927**	**0.907**	**0.038**	**0.905**

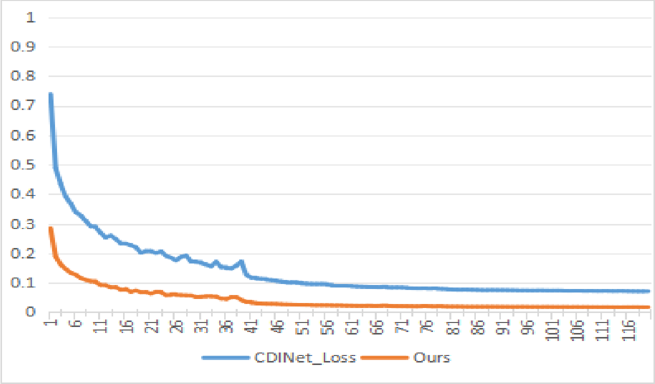


Fig. 7. Comparison of convergence of model loss function.

multi-layer feature fusion module, an attention mechanism, and an adjusted loss function. The advantages of the improved scheme are fully verified in comparison tests on multiple datasets, and its effectiveness is demonstrated by visual means, as shown in Fig. 8. From the figure, it can be seen that the method in this paper can detect salient objects more clearly in various application scenarios, especially when the salient map is poorly differentiated from the background is better than other algorithms.

The computational effort of the original model is 96,007,095,296, while the computational effort of the improved model is reduced to 77,001,786,368which reduces the computational complexity of SblCDINet by about 20% compared to the original CDINet model. This reduction stems from the integration of a lightweight, dense reconstruction decoder for multi-scale feature fusion, inspired by the Bidirectional Feature Pyramid Network (BiFPN). The inclusion of optimizes the BiFPN effectively, feature reuse and fusion strategy, reducing the computational cost while maintaining the high performance of the model.

In addition, experiments on the NJU2K dataset show that the MaxF of the improved model SblCDINet reaches 0.9239 while maintaining structural consistency, which is 1.01% higher than that of the baseline model (0.9147). The effectiveness of multi-scale feature fusion and attention mechanisms in complex scenes is verified. The S-measure of SblCDINet remains 0.9157 (which is equal to the baseline model), indicating that the model has strong robustness in retaining the target structure information, and there is no structural distortion introduced by module optimization. The MAE increased slightly from 0.0329 of the baseline model to 0.0371 (the absolute difference was 0.0042). It is speculated that the model introduces slight noise in the boundary refinement of low-contrast regions, which can be further optimized by adjusting the weight of the loss function in the future.

SblCDINet not only improves the detection accuracy but also maintains the structure consistency, which confirms the stability of its improvement strategy. The small fluctuation of MAE suggests that it is necessary to balance accuracy and error sensitivity in subsequent work, for example, by dynamically adjusting the IoU Loss weight to suppress

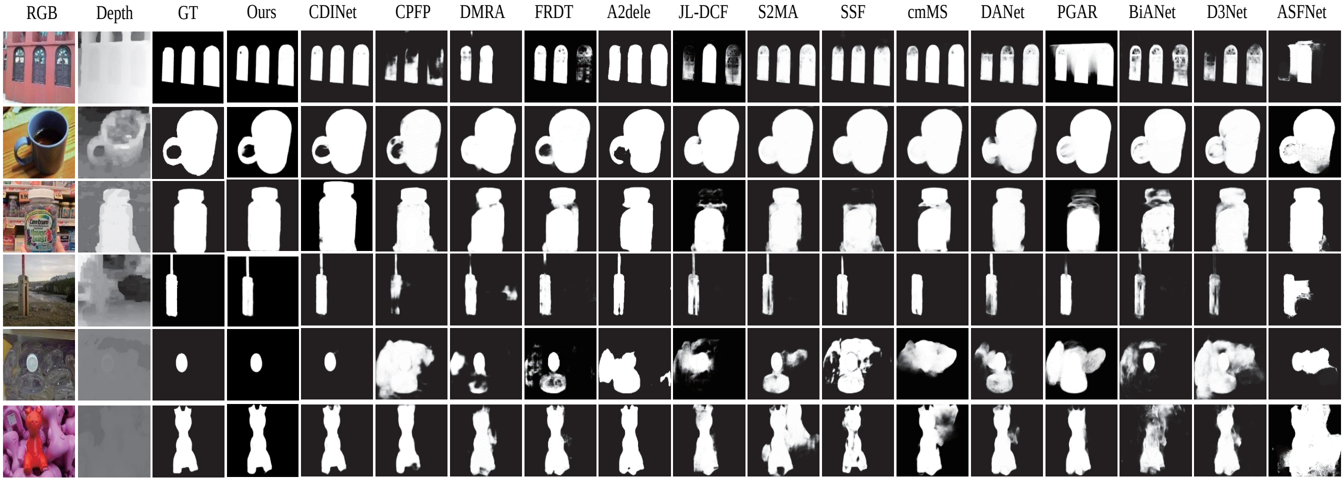


Fig. 8. Visualization comparison between this method and advanced methods.

background noise.

V. Conclusion

In this paper, we propose an innovative RGB-D image saliency detection method aimed at enhancing the feature processing capability. Within the existing network structure, we elaborately design and integrate a multi-scale feature fusion module, which achieves deep extraction of features through auxiliary jump connections and effectively fuses cross-level feature information, thus significantly enhancing the representation power of features. In addition, the introduced cyclic attention module, which is borrowed from the human brain's attention mechanism, significantly improves the efficiency of scene information processing. This module optimizes the localization accuracy of salient objects by reducing the saliency weights of background pixels. The improved model, compared with the original CDINet algorithm, improves all three metrics on each dataset, with the following details of improvement in accuracy: DUTS:1.1%, NJU2K:1%, NLPR:1.1%, STEREO:0.8%, and LFSD:1.4%, with a reduced number of parameters and a significant optimization of training time. The model proposed in this paper still effectively maintains the balance of the three evaluation metrics while reducing the model size. The future research direction will be devoted to further simplifying the network structure without sacrificing the accuracy of the model, with the expectation that the model will be able to quickly identify road signs and obstacles in the field of automatic driving thus significantly improving safety and reaction speed; in the field of medical image analysis, effectively separating lesion regions and providing doctors with more accurate diagnostic references. In addition, the application of the method in more fields, such as augmented reality is explored, to promote the wide application and continuous optimization of saliency detection technology.

References

[1] Y. Liu, Z. Kaihua, F. Jiaqing, *et al.*, "Camouflage object detection by progressively aggregating multi-scale scene context features," *Chinese Journal of Computer Science*, vol. 45, no. 12, pp. 2637–2651, 2022.

[2] S. Chen and Z. Meng, "Pedestrian detection algorithm based on multi-modal feature fusion," *Computer Engineering and Design*, vol. 45, no. 10, pp. 3017–3025, 2024.

[3] Y. Peng, Z. Zhai, and M. Feng, "Rgb-d salient object detection based on cross-modal and cross-level feature fusion.," *IEEE Access*, 2024.

[4] Z. Zhao, Z. Li, J. Miao, K. Wu, and J. Wu, "Global context-enhanced network for pixel-level change detection in remote sensing images.," *IAENG International Journal of Computer Science*, vol. 51, no. 8, pp. 1060–1070, 2024.

[5] F. Xiao, Z. Pu, J. Chen, and X. Gao, "Dgfnet: Depth-guided cross-modality fusion network for rgb-d salient object detection," *IEEE Transactions on Multimedia*, vol. 26, pp. 2648–2658, 2023.

[6] Y. Yuan, W. Liu, P. Gao, Q. Dai, and J. Qin, "Unified unsupervised salient object detection via knowledge transfer," *arXiv preprint arXiv:2404.14759*, 2024.

[7] H. Chen, F. Shen, D. Ding, Y. Deng, and C. Li, "Disentangled cross-modal transformer for rgb-d salient object detection and beyond," *IEEE Transactions on Image Processing*, 2024.

[8] C. Zhang, R. Cong, Q. Lin, L. Ma, F. Li, Y. Zhao, and S. Kwong, "Cross-modality discrepant interaction network for rgb-d salient object detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2094–2102, 2021.

[9] Z. Cai, X. Qiao, J. Zhang, Y. Feng, X. Hu, and N. Jiang, "Repvgg-simam: An efficient bad image classification method based on repvgg with simple parameter-free attention module," *Applied Sciences*, vol. 13, no. 21, p. 11925, 2023.

[10] J. Liu and H. Wang, "An improved yolov5s method of vehicle target detection in remote sensing images," in *Fourth International Conference on Signal Image Processing and Communication (ICSIPC 2024)*, vol. 13253, pp. 173–179, SPIE, 2024.

[11] A. Demir, E. Massaad, and B. Kiziltan, "Topology-aware focal loss for 3d image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 580–589, 2023.

[12] R. Maurya, P. Thirwarni, T. Gopalakrishnan, and M. Karnati, "Combining focal loss with cross-entropy loss for pneumonia classification with a weighted sampling approach," in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, vol. 2, pp. 1–5, IEEE, 2024.

[13] H. Zhuang and W. Liu, "Underwater biological target detection algorithm and research based on yolov7 algorithm," *IAENG International Journal of Computer Science*, vol. 51, no. 6, pp. 594–601, 2024.

[14] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7254–7263, 2019.

[15] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pp. 92–109, Springer, 2014.

[16] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 1115–1119, IEEE, 2014.

[17] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 454–461, IEEE, 2012.

[18] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light

field," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2806–2813, 2014.

[19] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4548–4557, 2017.

[20] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2014.

[21] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3052–3062, 2020.

[22] S. Chen and Y. Fu, "Progressively guided alternate refinement network for rgb-d salient object detection," in *European Conference on Computer Vision (ECCV)*, pp. 520–538, Springer, 2020.

[23] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time rgb-d salient object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pp. 646–662, Springer, 2020.

[24] 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.

[25] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang, "Asif-net: Attention steered interweave fusion network for rgb-d salient object detection," *IEEE Transactions on Cybernetics*, vol. 51, no. 1, pp. 88–100, 2020.