IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 July 2023, accepted 2 August 2023, date of publication 29 August 2023, date of current version 8 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3309636

RESEARCH ARTICLE

Hybrid Attention Mechanism and Forward Feedback Unit for RGB-D Salient Object Detection

HAITANG LI[1], YIBO HAN[2], PEILING LI[3], XIAOHUI LI[4], AND LIJUAN SHI[5]

[1]School of Information Engineering, Zhoukou Vocational College of Arts and Science, Zhoukou 466000, China
[2]Shanghai Tianma Microelectronics Company Ltd., Shanghai 201201, China
[3]College of Information Engineering, Zhengzhou University of Science and Technology, Zhengzhou 450052, China
[4]International Joint Research Laboratory for Cooperative Vehicular Networks of Henan, Zhengzhou 450046, China
[5]College of Big Data and Artificial Intelligence, Zhengzhou University of Science and Technology, Zhengzhou 450052, China

Corresponding author: Haitang Li (lihaitang0412@163.com)

This work was supported by the Key Research Project Plan of Henan Universities under Grant 22A630036.

ABSTRACT RGB-D saliency object detection (SOD) is an important pre-processing operation for various computer vision tasks and has received much attention in recent years. However, how to extract more effective features and how to effectively fuse RGB and depth modality features are still challenges that restrict the development of SOD. In this paper, we propose an effective network architecture called FFMA-Net: 1) We replace the backbone network of the baseline with a ResNet34 model to extract more effective features from the input data; 2) We design the HAM module to refine the features extracted by the ResNet34 model at different stages to ensure the effectiveness of features from each stage; 3) We propose the FFU module to perform multi-scale fusion of features from different stages, resulting in more semantic-rich features that are crucial for the decoding stage of the model. Finally, our model performs better than the latest methods on six RGB-D datasets on all evaluation metrics, especially in terms of F-measure metric, which shows significant improvement with approximately 5% on both SSD and LFSD datasets.

INDEX TERMS RGB-D salient object detection, forward feedback unit, hybrid attention mechanism.

I. INTRODUCTION

The human visual system is equipped with an attention mechanism that allows us to effortlessly concentrate on the most prominent objects or regions within a scene. In computer vision, the task of saliency object detection [1], [2] aims to automatically identify the most significant region or object in a given scene. In addition to perceiving the color appearance, texture features, and physical size of an object, the human visual system also possesses the capability to perceive its depth. This depth perception contributes to our understanding of three-dimensionality of the object and enriches our perception by providing more comprehensive information.

With the continuous advancements in camera technology, depth cameras like Microsoft Kinect are being used to capture depth maps of visual targets. In contrast to RGB images, depth maps provide valuable geometric structure, boundary information, and internal consistency of the visual target. On the other hand, RGB images offer detailed color appearance, texture, and other relevant information. By effectively incorporating RGB and depth information, saliency object detection (SOD) models can tackle more challenging visual scenes, including those with cluttered backgrounds or low contrast. Consequently, the research on RGB-D image saliency object detection has gained considerable attention from scholars and has witnessed significant progress in recent years [12], [19], [20], [21], [22], [23]. The utilization of RGB-D image saliency object detection extends to various fields, such as co-saliency object detection [3], [4], [5], image retrieval [6], video segmentation [7], super-resolution [8], [9], [10], visual tracking [11], depth estimation [13], super pixel segmentation [14], remote sensing SOD [15], [16], [17], light field saliency object detection [18], and more. However, in traditional or deep learning-based methods, the main emphasis has been on fusing and interacting depth features with RGB features. Unfortunately, a crucial issue has often been overlooked: the variability in the quality of the depth maps within the dataset. When high-quality depth maps

The associate editor coordinating the review of this manuscript and approving it for publication was Joewono Widjaja.

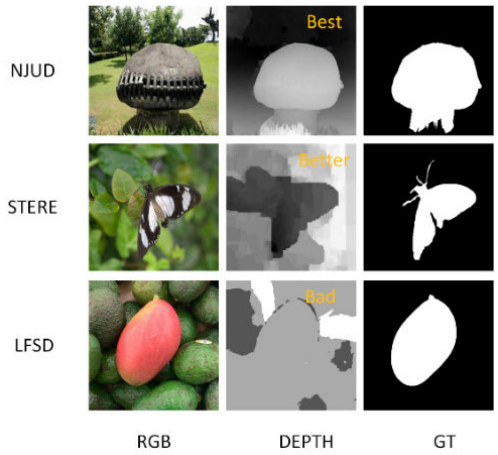

FIGURE 1. Examples of different quality depth maps are shown. The first row is a better depth map, the second row is second, and the third row is worse. A good quality depth map can display boundary information and accurate positioning more clearly.

are accessible, they contain clearer boundary and internal consistency information that is essential for accurate RGB-D salient object detection.

For low-quality depth maps, their lack of reliable information can impede the effective fusion of RGB and depth features, consequently impacting saliency object detection performance. Thus, it is crucial to ensure the quality of depth maps when conducting RGB-D image saliency object detection. Figure 1 illustrates several examples of depth maps with differing qualities. The labeled best depth map exhibits clearer boundaries and more accurate target position information, whereas the labeled poor depth map fails to provide reliable information and may even disrupt the fusion of RGB and depth features, thereby compromising saliency object detection performance.

The depth information of visual targets has been proven to be highly effective in enhancing salient object detection, as demonstrated in previous studies. However, the existing research primarily emphasizes the successful fusion of depth features and RGB features. As a result, this paper explores two additional aspects to build upon this foundation. Firstly, it focuses on extracting more efficient feature representations during the feature encoding stage. In RGB-D salient object detection, the input comprises both RGB and depth information from the target image. However, the quality of captured depth maps can vary due to the use of different devices for depth map acquisition. When low-quality depth maps are used for feature extraction, they may adversely affect the model's inference and predictions. Therefore, it is crucial to extract more effective features during the feature encoding stage in order to enhance the model's performance in inference and predictions. Secondly, this paper introduces the fusion of feature representations at different scales, going beyond the previous emphasis on fusing depth map and RGB map features exclusively. Generally, larger-scale features convey more detailed information about the target, such as color appearance and texture features. On the other hand, smaller-scale features provide more semantic information. By fusing information from multiple scales, this paper aims to further enhance the expressive power of the features.

Inspired by the previous discussion, this paper presents a novel architecture that enhances feature extraction capabilities compared to the baseline model. We introduce the Mixed Attention Mechanism module in the encoding stage to further refine the extracted feature information from the backbone network. These refined features enhance accuracy and effectiveness, crucial for effective model learning. Additionally, we incorporate Forward Feedback Unit modules into adjacent feature encoding stages, allowing the fusion of information at different scales. The fusion of multi-scale features is well-known to provide richer information as features at different scales possess distinct characteristics. We effectively transfer the fused multi-scale features to the decoding stage, benefiting the model's decoding process. Moreover, ablation experiments support the effectiveness of our proposed Forward Feedback Unit module. Specifically, we replace the backbone network of the baseline model with the ResNet34 model, which offers stronger feature extraction capabilities compared to VGG16. This enables the extraction of more diverse and effective feature representations. Considering the varying quality of depth maps in the dataset and the redundant nature of features extracted by the backbone network, we design the Mixed Attention Mechanism module. By refining RGB and depth features in the encoding stage, we ensure the correctness and effectiveness of features at each stage. Finally, our novel Forward Feedback Unit module facilitates the multi-scale fusion of features from different encoding stages. This fusion process combines the distinctive information carried by features at various scales, resulting in richer feature information, particularly beneficial for detection and classification tasks.

Overall, our contributions in this paper can be summarized as follows:

- We designed a novel Hybrid Attention Mechanism module to refine and correct features in the encoding stage, ensuring the effectiveness and accuracy of features at each encoding stage.
- We proposed a Forward Feedback Unit module, which can fuse multi-scale encoded features to achieve richer semantic information, essential for feature decoding.
- To further improve the performance of the proposed model, we explored the effects of different backbone networks on the model. In this section, we replaced the backbone of Baseline with ResNet34 and ResNet50 models to enhance the feature extraction ability of the model.

The structure of this paper is briefly outlined as follows: Section II introduces related work. Section III describes the overall architecture of the model, the implementation of the Hybrid Attention Mechanism module, and the Forward Feedback Unit module. Section IV provides a detailed analysis of

experimental performance and results. Section V concludes this work.

II. RELATED WORKS

In this section, we mainly introduced some related work of the proposed model in this paper, including RGB-D salient object detection, forward feedback unit module, and attention mechanism.

A. RGB-D SALIENT OBJECT DETECTION

There are primarily two categories of RGB-D salient object detection methods: traditional detection methods and deep learning-based detection methods. Traditional detection methods primarily utilize manually extracted features for object detection, such as image contrast [26], target shape [27], local background closure [23], [24], etc. However, the performance of traditional methods is typically unsatisfactory when dealing with complex visual scenes due to the limited expressive power of handcrafted features. In recent years, deep learning-based RGB-D salient object detection methods have made significant progress, benefiting from the rapid development of deep learning in the field of computer vision.

For instance, Fu et al. [28] designed a Siamese network for joint learning and devised a dense collaborative fusion strategy to fuse features. Zhang et al. [30] integrated cross-modal feature information through a tightly connected structure and then built a dynamic filtering network. Liu et al. [29] designed a residual fusion module to integrate depth decoded features into the RGB branch during the feature decoding stage. Zhang et al. [30] primarily studied the interaction between RGB and depth modalities in cross-modal settings and proposed an inconsistent interaction mode, i.e., the interaction between RGB modality and depth modality. Zhao et al. [31] introduced depth quality-aware control to mitigate the impact of low-quality depth maps while performing interactions in cross-modal settings. Zhao et al. [31] trained an enhanced contrastive network primarily to improve the quality of depth maps, resulting in clearer and more consistent depth maps and their regions, which in turn leads to better performance during model inference. Similarly, Chen et al. [32] proposed an enhancement and fusion framework that first generates a guidance map to address the low-quality issues of depth maps. Wang et al. [62] proposed a cross-modal network design along with a multi-cross attention module, which combines spatial attention and channel attention in a multi-cross manner to better utilize the rich detailed information of salient objects, thus improving the performance of the model. Kanwal et al. [63] introduced a local detail enhancement module that effectively captures intra-modal features at lower levels of the base network using a novel operation-level shuffling channel attention module, thereby enhancing the performance of the model. In comparison to the aforementioned works, the hybrid attention mechanism proposed in this paper has the advantages of having a simpler structure and achieving better results.

B. FORWARD FEEDBACK UNIT MODULE

Recently, the forward feedback unit module has been widely adopted in various network architectures to tackle computer vision tasks. For example, in semantic segmentation [35], the aim is to extract high-level linguistic information by employing topological loss. The obtained higher-level language information is then fed back to the shallow network to correct low-level semantic details. Consequently, this approach transforms significant outputs into input information for image-related problems, effectively addressing visual classification challenges within computer vision tasks.

The forward feedback unit module offers an efficient way to integrate two tensors of different sizes. In deep learning models, this mechanism enables the reuse of tensors with the same size across multiple dimensions, resulting in reduced parameter count and improved inference calculation speed. This feedback mechanism can be applied to various network models associated with visual tasks [33], [34], making it versatile for fulfilling different computer vision objectives such as object detection, semantic segmentation, instance segmentation, among others. In semantic segmentation tasks specifically, there are endeavors to extract high-level language information utilizing topological loss, with the potential for feeding back the resulting higher-level information to the shallow network for refining low-level semantic information.

C. ATTENTION MECHANISM

Traditional convolutional neural networks (CNNs) are primarily focused on extracting features from input data, often overlooking the important relationships and dependencies between these features. However, incorporating these dependence relationships can greatly enhance the performance of CNNs. As a result, researchers have increasingly directed their attention towards exploring and leveraging these dependencies.

Several mechanisms have been introduced to address this issue. The spatial attention mechanism [36] enables a neural network model to automatically identify regions of interest within an image. The channel attention mechanism [37], on the other hand, learns the importance of each channel and assigns corresponding weights to enhance the overall model performance. The self-attention mechanism [38] is designed to capture long-range dependencies in feature information, leading to improved model performance. Researchers have also developed text attention mechanisms [39] for multi-modal reasoning and matching, as well as recurrent attention mechanisms [40] that iteratively generate more accurate saliency results.

Furthermore, hybrid attention mechanisms have been proposed to combine different types of attention. For instance, Woo et al. [41] introduced BAM and CBAM attention

mechanisms, which integrate spatial and channel attention mechanisms in series or parallel to achieve superior results.

Building upon these insights, this paper presents a novel attention mechanism model that combines channel attention mechanisms, spatial attention mechanisms, and NAM [42] units in series. By utilizing parameter-free calculations offered by the NAM unit, the proposed attention mechanism model strikes a balance between model parameters and computational efficiency.

III. METHOD

In this section, we first elaborate on the overall structure of the proposed method in Section III-A. Then, we provide a detailed description of our designed hybrid attention mechanism module in Section III-B. Finally, we introduce the forward feedback mechanism in Section III-C.

A. OVERALL FRAMEWORK OF MAFF-NET

Figure 2 illustrates the overall structure of the proposed method, which consists of three main parts: the encoder, decoder, and feature fusion module. Firstly, the RGB image and depth map are input into twin encoders to obtain multi-level feature representations. At each stage of feature encoding, a hybrid attention mechanism module is added to refine and rectify the corresponding stage features. Secondly, the refined and rectified features are fused and computed by the feature integration module [57]. We introduce a forward feedback mechanism module to the feature integration module, effectively fusing the current-stage features with the next-stage features to obtain semantically rich feature representations. Finally, the feature representations obtained by the encoder are passed to the decoder for inference calculations, resulting in the final prediction outputs. The RGB images and depth maps input to the model are uniformly scaled to 352*352*3 and 352*352*1, respectively. Next, we will provide a detailed introduction to the hybrid attention mechanism module and the forward feedback mechanism module.

To ensure computational efficiency, we employ the relatively shallow ResNet34 network as the backbone network, which has been pre-trained on the ImageNet dataset. We retain only the convolutional layers of the ResNet34 network, removing the last max pooling layer and fully connected layer. Firstly, the RGB image and depth map are input into the ResNet34 backbone network for feature extraction, yielding five stages of RGB features denoted as RGB_F^i and depth features denoted as $DEPTH\_F^i$, where i is the index of the feature level ranging from 1 to 5. Considering the information redundancy of different stage RGB features and depth features, along with potential errors in depth features, we incorporate a hybrid attention mechanism module after each stage to rectify and refine the feature, ensuring its effectiveness and correctness. Different scales of features capture different information, with low-level encoded features containing color, texture, and detail information, and high-level encoded features containing semantic and category information. Therefore, we design a forward feedback unit module after cross-modal feature fusion to combine the current-scale feature with the next-scale feature, resulting in fused features that contain richer information.

The decoder structure of our proposed model corresponds one-to-one with the encoder structure and consists of five stages. The first two decoding blocks of the decoder comprise two convolutional layers and one transposed convolution, while the last three decoding blocks consist of three convolutional layers and one transposed convolution. During the decoding stage, the fused features obtained in the encoding stage are passed to the corresponding decoding block for feature concatenation. This allows the decoder to restore features and perform progressive inference, leading to more accurate prediction maps. Additionally, the outputs of the five decoding stages of the proposed model are S1, S2, S3, S4, and S5. Finally, we consider the output of the first decoding block as our final prediction map.

B. HYBRID ATTENTION MECHANISM MODULE

Convolutional neural networks (CNN) are widely used in various computer vision tasks as a powerful tool for feature extraction, but they have inherent weaknesses. For example, the features extracted by CNN may not fully describe the content of an image, which can lead to classification errors or detection false alarms. Additionally, CNN may be sensitive to distortions, lighting, and scale changes, resulting in insufficient feature representation. These shortcomings can negatively impact the performance of the entire network model.

Inspired by attention mechanism theory, we propose a new attention mechanism architecture called the hybrid attention mechanism module to address these issues. The hybrid attention mechanism module consists of three parts: channel attention mechanism, spatial attention mechanism, and the Normalization-based Attention Module. We combine them organically using a concatenation approach, compare six different concatenation methods in our ablation experiments, with the most prominent being the combination of channel attention mechanism [37], spatial attention mechanism [38], and NAM [42].

The computational process of the hybrid attention mechanism module is shown in Figure 3. Specifically, the input data is first processed by the channel attention mechanism to obtain feature weights, which are then point-multiplied with the input data. Next, the feature weights from the channel attention mechanism are passed into the spatial attention mechanism, where feature weights are calculated again and point-multiplied with the previous result. Finally, the output of the spatial attention mechanism is passed into the NAM module for calculation of feature weights, which are then again point-multiplied with the previous result to obtain the final feature information. This entire computation process can be expressed using the following formula: We first compute the channel attention mechanism for the input features. The model can adaptively learn the importance of each channel, selectively weighting and merging features from different

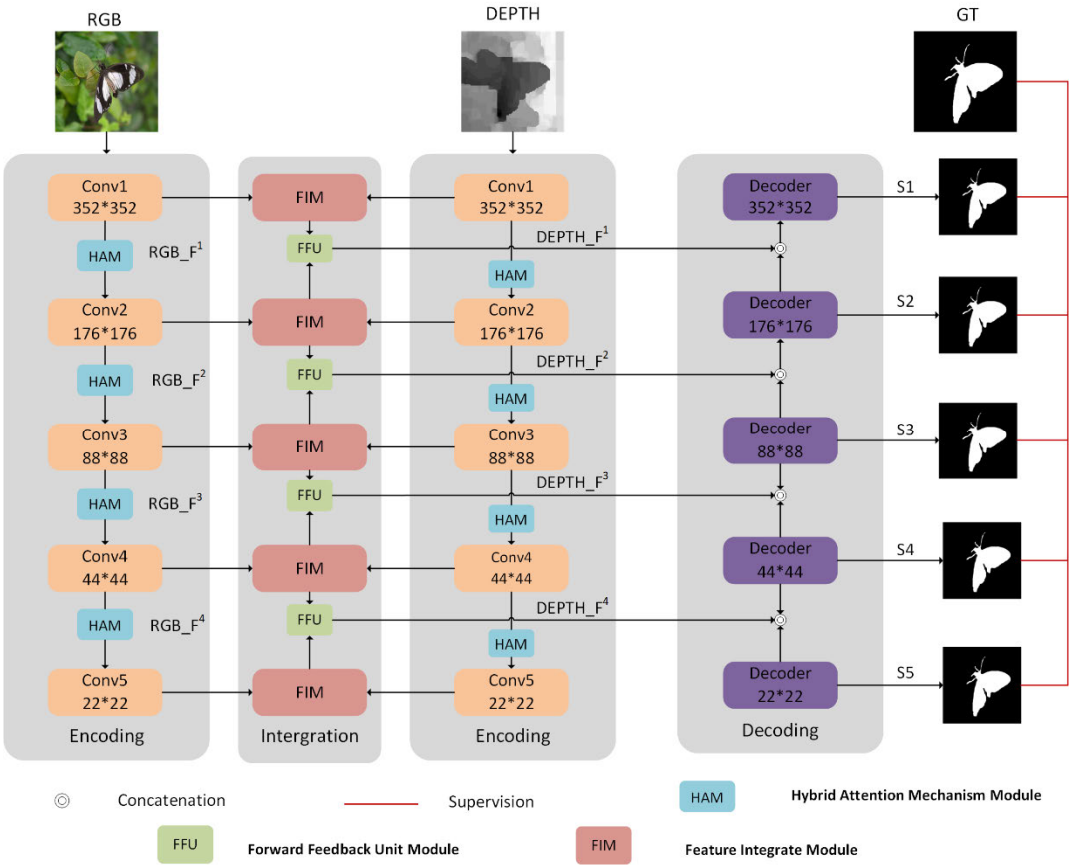

FIGURE 2. The overall architecture of the proposed HAFF-Net. The overall network structure is mainly composed of three stages: feature encoding, feature fusion, and feature decoding. The HAM module mainly extracts and preserves the features obtained by the backbone network to ensure their effectiveness, while the FFU mainly performs multi-scale feature fusion on the basis of modal feature fusion. Finally, after the features are fused, they are passed to the decoding module for decoding processing.

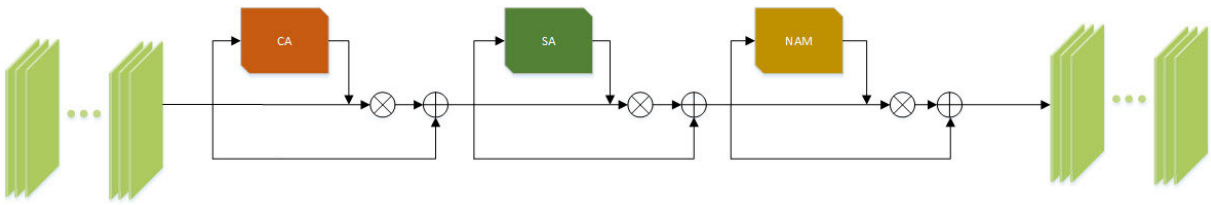

FIGURE 3. The detail of the HAM.

channels. Next, we incorporate the spatial attention mechanism on top of the previous computation. This mechanism calculates weights to determine the contribution of each position or region to the task and focuses more on areas with higher weights. Finally, we utilize the Normalization-based Attention Module (NAM) to compute attention on the features. NAM applies normalization techniques to better handle the distribution of input data during attention computation. As a result, the Mixed Attention Mechanism module enhances the model's emphasis on important features, thereby improving performance and robustness.

$$f_{rd}^{i,c} = CA\left(f_{rd}^{i}\right) \otimes f_{rd}^{i} + f_{rd}^{i} \tag{1}$$

$$f_{rd}^{i,cs} = SA\left(f_{rd}^{i,c}\right) \otimes f_{rd}^{i,c} + f_{rd}^{i,c} \tag{2}$$

$$f_{rd}^{i,css} = NAM\left(f_{rd}^{i,cs}\right) \otimes f_{rd}^{i,cs} + f_{rd}^{i,cs} \tag{3}$$

In the equations above, f_{rd}^{i} the result of merging depth and RGB features in each encoding stage, the symbol $\otimes$ denotes matrix multiplication, while CA and SA represent the channel attention mechanism and spatial attention mechanism respectively, and NAM represents the Normalization-based Attention Module.

C. FORWARD FEEDBACK UNIT MODULE

RGB-D salient object detection utilizes a convolutional neural network with an Encoder-Decoder structure for processing. The Encoder incorporates a sequence of convolution operations, pooling operations, and activation calculations, primarily employed to extract features from the input data.

However, the deep structure of the network may lead to the neglect or loss of crucial feature information during these convolutions, pooling, and activation computations. This becomes especially pronounced when considering the input for salient object detection, which comprises RGB and depth information. Notably, the quality of the depth map can vary depending on the dataset, and in the case of inadequate depth maps, it may contain insufficient effective feature information and even unwanted interference. Consequently, this can cause the model to potentially forfeit essential feature information throughout the feature extraction process.

To solve this problem, we propose a new module called the Forward Feedback Unit module. The Forward Feedback Unit module mainly consists of convolution operations, up sampling operations, and pooling operations. Specifically, it has two inputs, and the shapes of these two input tensors are different. The Forward Feedback Unit module performs a series of convolution operations, up sampling operations, and pooling operations on these two input tensors to effectively fuse and enhance them. The shape of the fused tensor is the same as one of the input tensors. Finally, we describe in detail the calculation process of the Forward Feedback Unit module, here we take F^1 and F^2 as examples. The Forward Feedback Unit module consists of two branches. For the first branch: firstly, F^1 is input into the Forward Feedback Unit for convolution calculation, secondly, the result of the calculation is point-multiplied with f_{result}^{i-1}, and finally, the result of the multiplication is added to F^1 by point addition. For the second branch: firstly, F^2 is input into the Forward Feedback Unit for global average pooling and convolution calculation. Assuming that the result obtained is f_{result}^{i-1}, it is then subjected to upsampling calculation, and assuming that the result obtained is f_{up}^{i-1}. Finally, we point-add the results of the two branches to obtain the final result. The entire computation process of the Forward Feedback Unit module can be expressed using the following formula.

$$f_{result}^{i-1} = conv_{1\times1}\left(GAP\left(f^{i-1}\right)\right) \tag{4}$$

$$f_{up}^{i-1} = up\left(f^{i-1}\right) \tag{5}$$

$$f_{final} = conv_{3\times3}\left(f^{i}\right) \otimes f_{result}^{i-1} + f^{i} + f_{up}^{i-1} \tag{6}$$

Here, in the first branch of the Forward Feedback Unit Module, a 3*3 convolution is applied to f^{i-1}, the purpose of this convolution is mainly to resize f^{i-1}, resulting in $conv_{3\times3}\left(f^{i}\right)$. Next, we fuse $conv_{3\times3}\left(f^{i}\right)$ and f^{i-1} by performing matrix multiplication, which merges their features. Finally, we enhance the features by adding the preliminary fused result with $conv_{3\times3}\left(f^{i}\right)$, yielding the output of the first branch. Moving on to the second branch of the module, we start by applying global average pooling *GAP* to the features. Then, we use a 1*1 convolution to calculate the weights for f^{i-1} based on the pooled result. Subsequently, we perform matrix multiplication to fuse f^{i} with the resulting weighted feature map, denoted as f_{result}^{i-1}. To match the dimensions, we up sample f^{i-1} using linear interpolation. The upsampled result serves as the output of the second branch. Finally, we combine the outputs of the two branches through matrix addition to achieve feature fusion at different scales. In the formulas, f^{i-1} and f^{i} represent features from two distinct encoding stages. *GAP* represents global average pooling, *up* denotes up sampling, the symbol $\otimes$ denotes matrix multiplication, $conv_{1\times1}$ and $conv_{3\times3}$ refer to 1*1 convolution and 3*3 convolution operations, respectively.

IV. EXPERIMENT

We first introduce six commonly used RGB-D salient object detection datasets and four commonly used evaluation metrics. Then we describe in detail our experimental settings and the experimental environment. Afterwards, we compare our proposed method with the 21 latest methods, and finally, we conduct a series of ablation experiments to demonstrate the effectiveness of our proposed modules and methods.

A. DATASETS

We evaluated the performance of our model on six commonly used RGB-D datasets and compared it with 21 other state-of-the-art methods. These datasets include STERE, LFSD, NLPR, RGBD135, NJUD, and SSD. STERE

Reference [43] comprises 1000 pairs of stereo images obtained from the Internet. Depth maps corresponding to these stereo images were estimated, encompassing various outdoor scenes and objects. LFSD [44] is a light field SOD dataset consisting of 100 indoor and outdoor images with accompanying depth maps, predominantly featuring simple foreground objects. The RGBD135 [45] dataset contains 135 images mostly involving relatively uncomplicated foreground objects and visual scenes, with good quality depth maps. NJUD [43] is a collection of 1985 stereo images, 3D movies, and photos sourced from the Internet and stereoscopic films. It showcases different objects and complex scenes, with depth maps estimated from stereo images. SSD [46] is a smaller-scale dataset comprising 80 stereo movie frames, including various movie scenes with people, animals, buildings, etc., serving as foreground objects. SIP [22] is a recently released dataset containing 929 images and high-quality depth maps, each with a resolution of 744 × 992 pixels.

B. EVALUTION METRICS

To quantitatively evaluate the performance of our proposed method, we use four metrics: S-measure, F-measure, precision-recall curves, and Mean Absolute Error to assess the final performance of the model.

Precision-Recall curves: P-R curve is a performance evaluation curve plotted with Precision (P) as the horizontal axis and Recall (R) as the vertical axis. Generally, the following formulas are used to calculate Precision (P) and Recall (R).

$$Precision = \frac{|H \bigcap G|}{|M|} \tag{7}$$

$$Recall = \frac{|H \bigcap G|}{|G|} \tag{8}$$

F-measure: Sometimes the results evaluated by the P-R curve may be unreliable. Therefore, F-measure is a widely used comprehensive evaluation index, which mainly calculates the final result by calculating the precision and recall scores. We can use the following formula to calculate the F-measure.

$$F_\beta = \frac{\left(1+\beta^2\right) Precision \times Recall}{\beta^2 Precision + Recall} \tag{9}$$

The precision and recall in the formula represent the precision score and recall score, and β^2 is set to 0.3 to emphasize precision.

Mean Absolute Error (MAE): MAE calculates the average pixel absolute error between the predicted image and the corresponding ground truth. The formula is as follows.

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} |P(i,j) - G(i,j)| \tag{10}$$

In the formula, H and W represent the height and width of an image respectively. $P(i, j)$ represents the predicted image and $G(i, j)$ represents the ground truth.

S-measure: The emphasis of S-measure is to measure the structural similarity between the predicted image and the ground truth while considering both region-aware and object-aware structural similarities. The specific calculation process is shown in the following formula.

$$S = \alpha S_o + (1-\alpha) S_k \tag{11}$$

The parameter α in the formula is set to 0.5 to balance the region-aware similarity S_k and object-aware similarity S_o.

C. EXPERIMENTAL SETTINGS

The MAFF-Net model proposed in this study is designed for salient object detection, a crucial task in computer vision. The training of the model was conducted on a synthetic dataset consisting of 2050 samples, including 1400 samples from the NJUD dataset and 650 samples from the NLPR dataset. During the training process, the input image size was uniformly scaled to 352 × 352, and various data augmentation techniques were utilized such as flipping, cropping, and rotation to prevent overfitting. The model architecture was implemented in Pytorch and trained on a Tesla V100-PCIE-16GB GPU. To reduce computational cost and improve efficiency, ResNet34 was selected as the backbone network, and the final pooling layer and fully connected layer were removed. Pretrained weights from the ImageNet dataset were applied to improve the initialization of the network. For optimizing the performance of the designed model, the stochastic gradient descent (SGD) algorithm was used with a batch size of 4, learning rate of 1e-4, weight decay of 0.1, and a total of 60 epochs. Additionally, to avoid overfitting during long-term training, the model was saved every 5 epochs. After training, the model achieved an average FPS of 21.

D. COMPARISON WITH STATE-OF-ART METHODS

We compare our method with 21 other state-of-the-art methods on six widely used benchmarks, including PCF [47], AFNet [48], CPFP [31],DMRA [40], MMCI [49], TANet [50], D3Net [22], JLDCF [51], S2MA [29], PGAR [52], ICNet [19], DASNet [53], UCNet [54],DCF [55], DSA2F [56], CCAFNet [61], HAINet [57], LIANet [58], CDINet [30], CIRNet [59], RD3DNet [60]. Some of the above methods are trained with subsets of NJU2K and NLPR, while others are trained with subsets of NJU2K, NLPR and DUTLF-depth.

1) VISUAL COMPARISON

Figure 5 presents a visual comparison of the inference results between our proposed model and seven other state-of-the-art models. Upon observing Figure 5, it is evident that our model achieves the best saliency detection performance compared to the other seven models. The saliency maps generated by our model exhibit clearer contours and internal consistency in most scenes. Specifically, in the fourth and sixth rows where the visual scenes feature complex backgrounds, our model produces saliency maps with sharper boundaries compared to the other models, thereby demonstrating its effectiveness in challenging scenarios.

Furthermore, we showcase visual comparisons of feature activation on the LFSD and STERE datasets. The LFSD dataset consists of 100 images, most of which contain prominently distinguishable target regions. By employing the mixed attention mechanism for feature activation, the pixel-level feature representation and regional target perception are enhanced, as exemplified in the left three images of the fourth row in Figure 6. As for the STERE dataset, which is the first stereo image SOD dataset comprising 1000 pairs of stereo images collected from the Internet, the image quality varies, and some visual scenes have distracting elements in the target region. Nevertheless, by utilizing the proposed mixed attention mechanism for feature activation, the perception of the target region is also enhanced, as shown in the right three images of the fourth row in Figure 6.

In summary, through comparisons with seven other state-of-the-art SOD models and visual comparisons of feature activations on multiple datasets, the HAFF-Net model proposed in this study achieves better salient object detection performance in different scenarios, demonstrating its potential for practical applications and broad prospects.

2) COMPARISON WITH ADVANCED METHODS

Table 1 presents the quantitative comparison results between the proposed model and 21 other models. Figure 7 depicts the P-R curves of the proposed model and six other models: CDINET, HAINET, SPNET, LIANET, CCAFNET, and CIRNET. Overall, our proposed model outperforms these models on six public datasets, especially in terms of the MAE metric, which shows a significant decrease across all six datasets. Our model demonstrates a significant improvement in all

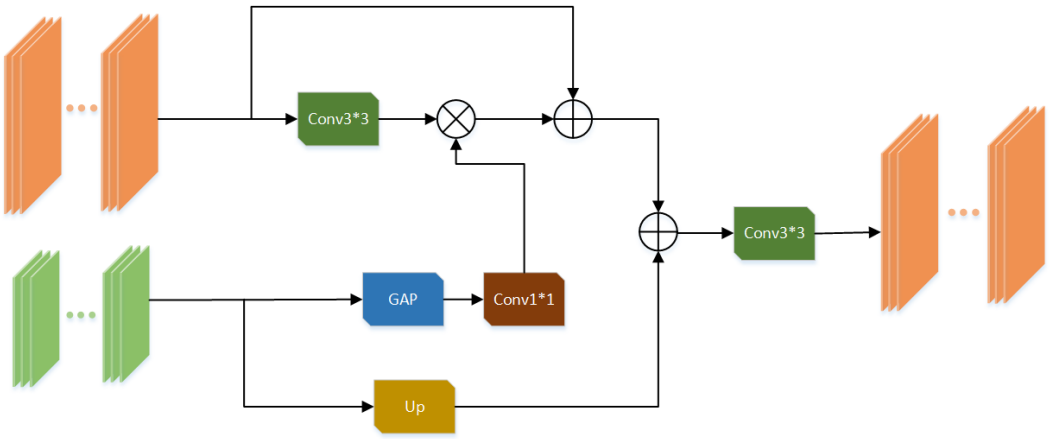

FIGURE 4. The detail of the FFUM network.

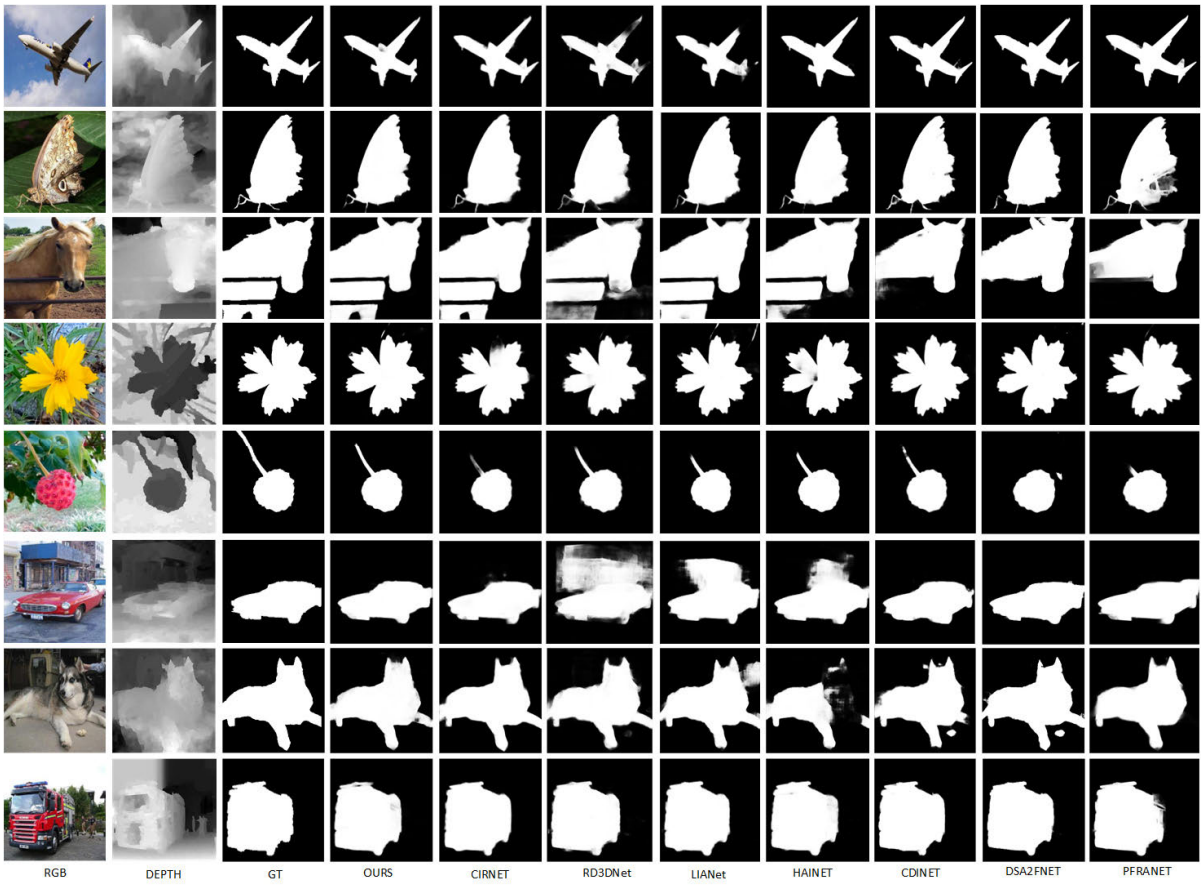

FIGURE 5. Visual comparison with other 9 advanced methods.

evaluation metrics on the NJUD and SSD datasets, particularly achieving an approximately 5% increase in the F-measure metric. On the SIP, RGBD135, and STERE datasets, our model performs similarly to the best-performing models. There is only a slight difference in the F-measure metric for the SIP and STERE datasets, while the RGBD135 dataset shows a slightly higher difference in the E-measure metric. On the LFSD dataset, our model maintains consistency with the best model in terms of the MAE metric but shows a slight difference in the other three metrics. Additionally, we compared the FPS values of each model in Table 1, indicating that our proposed model achieves the best balance between performance and FPS.

E. ABLATION STUDIES

In this section, we conducted a series of ablation experiments to evaluate the effectiveness and importance of the proposed modules in our paper. The ablation experiments were performed with a batch size of 4, a learning rate of 1e-4, weight decay of 0.1, and a total of 60 epochs.

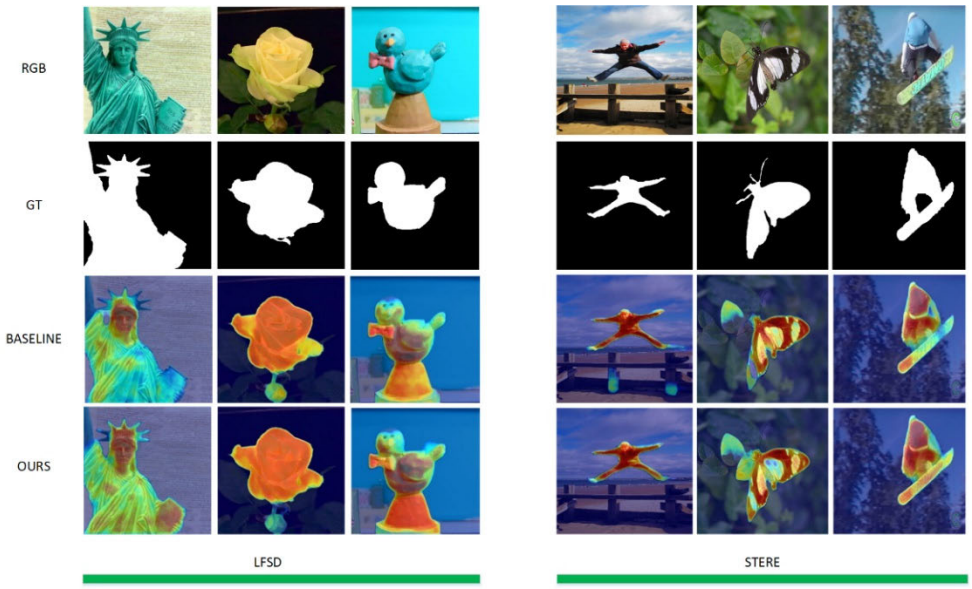

FIGURE 6. Visualization of feature activations.

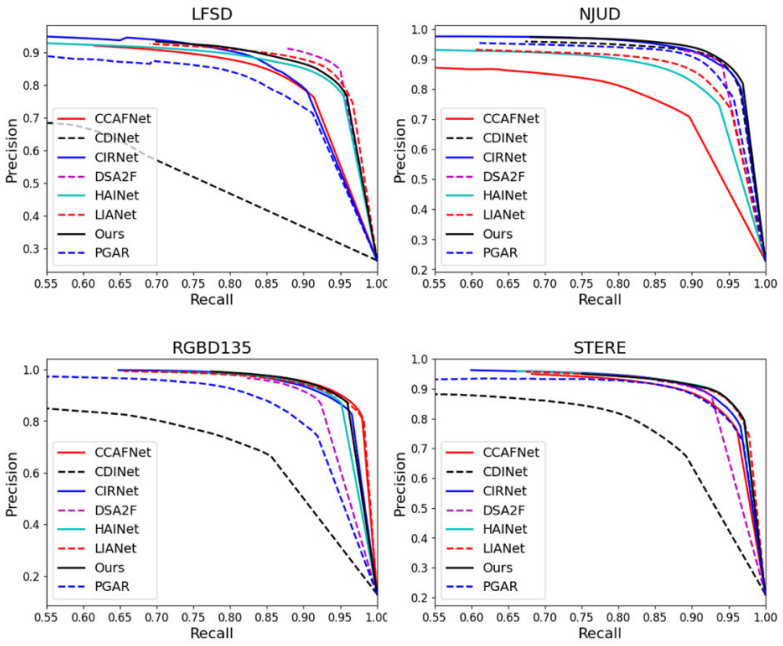

FIGURE 7. Quantitative comparisons with 7 state-of-the-art methods over 4 datasets.

We saved the trained model at intervals of 5 epochs. The proposed modules that we introduced in our paper include the Hybrid Attention Mechanism Module, the Forward Feedback Unit Module, and the replacement of the baseline backbone network with ResNet34. These modules were designed to enhance the performance of object detection. To assess the impact of these modules, we compared the results of the ablation experiments. The specific outcomes highlighted the effectiveness of the proposed modules and the improvements achieved. Table 2 provides a detailed overview of the effectiveness of these modules and the corresponding enhancements observed.

1) COMPONENT VALIDITY

By examining Table 2, where B represents the baseline, we utilized VGG16 as the backbone network and performed fusion operations on the original RGB and depth maps for decoding, following the approach described in HAINet [57]. Rows 2, 3, and 4 of the table highlight different combinations of components. Analyzing the second row of Table 2,

TABLE 1. Quantitative comparisons on six widely used datasets. We select four commonly used evaluation indicators for the evaluation of the model, including F-measure (FM, the bigger, the better), S-measure (SM, the bigger, the better), MAE (the smaller, the better), and E-measure (EM, the bigger, the better). The best results are marked in red font.

Models	FPS	NJUD	RGBD135	SIP	SSD	LFSD	STERE
		FM/SM/MAE/EM	FM /SM/MAE/EM	FM /SM/MAE/EM	FM /SM/MAE/EM	FM /SM/MAE/EM	FM /SM/MAE/EM
PCF(CVPR2018)	17	.872/.877/.059/.924	.804/.842/.049/.893	.838/.842/.071/.901	.807/.841/.062/.894	.775/.786/.119/.827	.860/.875/.064/.925
AFNet(ACCESS2019)	33	.775/.772/.100/.853	.728/.770/.068/.881	.712/.720/.118/.819	.687/.714/.118.807	.744/.738/.133/.815	.823/.825/.075/.887
CPFP(CVPR2019)	6	.877/.878/.053/.923	.846/.872/.038/.923	.851/.850/.064/.903	.766/.807/.082/.852	.828/.828/.088/.872	.874/.879/.051/.925
DMRA(ICCV2019)	10	.886/.886/.051/.927	.888/.900/.030/.943	.821/.806/.085/.875	.844/.857/.058/.906	.856/.847/.075/.884	.886/.886/.047/.938
MMCI(patcog2019)	20	.852/.858/.079/.915	.822/.848/.065/.928	.818/.833/.086/.897	.781/.813/.082/.882	.771/.787/.132/.839	.863/.873/.068/.927
TANet(TIP2019)	14	.874/.878/.060/.925	.827/.858/.046/.910	.830/.835/.075/.895	.810/.839/.063/.897	.796/.801/.111/.847	.861/.871/.060/.923
D3Net(TNNLS2020)	20	.900/.902/.046/.939	.885/.898/.031/.946	.861/.860/.063/.902	.834/.857/.059/.897	.810/.825/.095/.853	.891/.899/.046/.938
JL-DCF(CVPR2020)	9	- - - -	.933/.929/.021/.963	.898/.873/.053/.914	- - - -	.870/.854/.074/.882	.909/.901/.040/.919
S2MA(CVPR2020)	9	.889/.894/.053/.930	.935/.941/.021/.963	.877/.872/.057/.919	.848/.868/.052/.909	.835/.837/.094/.873	.882/.890/.051/.922
PGAR(ECCV2020)	24	.893/.909/.042/.916	.869/.913/.026/.939	.883/.876/.055/.908	- - - -	.852/.853/.074/.879	.880/.901/.041/.919
ICNet(TIP2020)	13	.891/.894/.052/.901	.925/.920/.027/.959	.857/.854/.069/.906	- - - -	.880/.858/.071/.881	.897/.891/.054/.911
DASNet(MM2020)	40	- - - -	.926/.905/.025/.942	.900/.877/.051/.917	.871/.875/.045/.891	.840//.825/.095/.853	.904/.899/.046/.920
UC-Net(CVPR2020)	17	- - - -	.930/.913/.025/.954	.896/.868/.051/.913	- - - -	.872/.854/.076/.878	.908/.901/.039/.921
DCF(CVPR2021)	26	.923/.911/.038/.924	.909/.904/.024/.950	.899/.875/.052/.920	.867/.864/.049/.898	.867/.841/.075/.883	.911/.902/.039/.929
CDINet（2021）	25	.913/.912/.038/.945	.741/.801/.075/.862	.689/.731/.128/.813	.861/.873/.048/.915	.614/.670/.164/.756	.798/.827/.075/.884
DSA2F(CVPR2021)	34	.917/.904/.039/.937	.930/.916/.023/.955	.891/.862/.057/.911	- - - -	.871/.854/.076/.881	.910/.897/.039/.928
CCAFNet（2021）	58	.796/.818/.076/.880	.935/.931/.018/.973	.880/.876/.054/.916	.736/.775/.080/.855	.832/.827/.087/.876	.886/.891/.044/.934
HAINet(TIP2021)	20	.872/.876/.050/.914	.911/.915/.023/.948	.897/.884/.048/.925	.727/.770/.098/.846	.854/.850/.078/.888	.906/.906/.038/.944
LIANet（2022）	28	.875/.884/.049/.925	.918/.930/.019/.965	.890/.880/.052/.922	.697/.755/.110/.811	.868/.867/.070/.901	.904/.904/.040/.945
CIRNet（2022）	22	.913/.906/.043/.949	.919/.925/.021/.964	.882/.877/.054/.914	.727/.757/.097/.838	.849/.830/.088/.875	.901/.903/.042/.940
RD3DNet（2022）	32	.860/.878/.052/.915	.921/.929/.022/.969	.885/.882/.054/.916	.701/.767/.095/.830	.874/.871/.074/.902	.897/.905/.045/.938
Ours	21	.922/.913/.035/.952	.935/.931/.018/.967	.891/.882/.047/.926	.765/.794/.079/.870	.864/.859/.070/.896	.906/.907/.036/.945

it becomes evident that incorporating each module into the baseline leads to a corresponding improvement in performance. Particularly noteworthy is the fourth row of the table, where we integrated all modules and enhancements into the baseline, resulting in significant performance improvements across the three datasets. Specifically, compared to the baseline, the F-measure metric increased by approximately 5% on the NJUD and SSD datasets, while the MAE metric decreased by around 20%. Overall, our proposed modules and improvements presented in this paper offer substantial enhancements for the performance of the model.

2) DIFFERENT HYBRID ATTENTION

To further investigate the performance of the Hybrid Attention Mechanism (HAM) module, we conducted an ablation experiment analysis on different combinations of the HAM module. The HAM module comprises the spatial attention mechanism, channel attention mechanism, and NAM. We explored six combinations by concatenating these components in different orders. Our experiments were performed on the LFSD, NJUD, and SSD datasets, and each combination represented a set of experiments. After analyzing the results, we found that the first combination of the HAM module achieved the best performance on the NJUD and SSD datasets. Compared to the baseline, it showed approximately a 2% improvement in F-measure, S-measure, and E-measure. Additionally, the MAE metric decreased by 22%. However, the third combination only performed well on the LFSD dataset. It exhibited a decrease in F-measure, S-measure, and E-measure by 2% on the NJUD and SSD datasets, along with a 20% increase in MAE.

The second, fourth, fifth, and sixth combinations did not yield satisfactory results on all three datasets. When compared to the best combination (first row), these groups exhibited a decline in F-measure, S-measure, and E-measure, while the MAE metric increased. Therefore, overall, the first combination of the HAM module demonstrated the best performance.

3) THE IMPACT OF DIFFERENT BACKBONE ABOUT THE MODEL

We conducted experiments to examine the influence of different backbone networks on the performance of our object detection model. We replaced the original backbone

TABLE 2. Ablation experiments of different components on NJUD, RGBD135 and SSD datasets. B is the baseline. HAMM and FFUM are hybrid attention mechanism module and forward feedback unit module respectively. CResNet34 represents that we replaced the baseline backbone network with a ResNet34 model. The best results are marked in bold font.

#	Settings	Params.	NJUD	RGBD135	SSD
			FM/SM/MAE/EM	FM /SM/MAE/EM	FM /SM/MAE/EM
1	B	59823374	.872/.872/.050/.914	.911/.915/.023/.948	.727/.770/.098/.846
2	B + HAMM	59867090	.893/.893/.045/.932	.920/.923/.021/.960	.725/.760/.100/.834
3	B + HAMM + FFUM	65617746	.904/.901/.040/.940	.927/.928/.019/.966	.737/.769/.092/.842
4	B + HAMM + FFUM+ResNet34	68143544	**.922/.913/.035/.952**	**.935/.931/.018/.967**	**.765/.794/.079/.870**

TABLE 3. Show that the hybrid attention mechanism module consists of three parts in series, namely, spatial attention mechanism, channel attention mechanism respectively. Here, we discuss six different combinations of the hybrid attention mechanism, with the best results highlighted in bold.

#	Settings	Params.	LFSD	NJUD	SSD
			FM /SM/MAE/EM	FM /SM/MAE/EM	FM /SM/MAE/EM
1	CA+SA+NAM	68143544	.864/.859/.070/.896	**.922/.913/.035/.952**	**.765/.794/.079/.870**
2	SA+CA+ NAM	68143544	.860/.857/.073/.892	.888/.890/.043/.933	.744/.781/.085/.855
3	NAM +SA+CA	68143544	**.871/.867/.067/.902**	.897/.899/.042/.935	.740/.782/.092/.840
4	CA+ NAM +SA	68143544	.866/.861/.074/.896	.899/.894/.044/.934	.704/.762/.098/.804
5	SA+ NAM +CA	68143544	.852/.849/.077/.886	.910/.907/.036/.946	.745/.781/.095/.835
6	NAM +CA+SA	68143544	.849/.849/.849/.887	.883/.885/.046/.046	.621/.718/.117/.756

TABLE 4. Ablation experiments of different components on NJUD, RGBD135 and SSD datasets. B is the baseline. HAMM and FFUM are mixed attention mechanism, forward feedback unit respectively. ResNet34 and ResNet50 stand for the different backbone of the baseline. The best results are marked in bold font.

#	Settings	Params.	NJUD	RGBD135	SSD
			FM /SM/MAE/EM	FM /SM/MAE/EM	FM /SM/MAE/EM
1	B + HAMM + FFUM	65617746	.904/.901/.040/.940	.927/.928/.019/.966	.737/.769/.092/.842
2	ResNet34+ HAMM + FFUM	68143544	.922/.913/.035/.952	**.935/.931/.018/.967**	.765/.794/**.079/.870**
3	ResNet50+ HAMM + FFUM	199993400	**.939/.932/.026/.963**	.921/.921/.020/.954	.765/.794/.083/.856

network with ResNet34 and ResNet50. In the first row of the table, we presented the results achieved by adding the HAM and FFU modules on top of the baseline model across three datasets. The second row details the outcomes obtained when utilizing ResNet34 as the new backbone network.

Our findings indicate a significant enhancement in performance on all three datasets after adopting ResNet34 as the backbone network. Notably, the SSD dataset exhibited an approximate 3% increase in F-measure, S-measure, and E-measure, accompanied by a 14% decrease in MAE compared to the values from the first row.

On the other hand, using ResNet50 as the backbone network resulted in improved performance solely on the NJUD dataset. However, evaluation metrics on the RGBD135 and SSD datasets experienced varying degrees of decline compared to the first row. Additionally, the parameter count of the model increased approximately threefold. Consequently,

considering all factors, the utilization of ResNet34 as the backbone network proved to deliver the best overall performance.

V. CONCLUSION

In this paper, we propose a new RGB-D salient object detection model, called HAFF-Net. The network is based on an encoder-decoder structure, unlike most existing RGB-D SOD methods, we first introduce a Forward Feedback Unit (FFU) module after the fusion of RGB and depth features to fuse feature information from different stages and enhance semantic expression of the fused features. The fused feature representation is then passed to the decoder for decoding and prediction. We also introduce Hybrid Attention Mechanism (HAM) modules at each decoding stage to refine the decoded features by filtering out interfering factors and enhancing the discriminability of features. Comprehensive experiments and ablation studies demonstrate that our proposed model is highly competitive with other state-of-the-art RGB-D SOD methods and achieves a good balance between accuracy and speed.

REFERENCES

[1] A. Borji, M. M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[2] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Comput. Vis. Media*, vol. 5, no. 2, pp. 117–150, Jun. 2019.

[3] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, and C. Hou, "An iterative co-saliency framework for RGBD images," *IEEE Trans. Cybern.*, vol. 49, no. 1, pp. 233–246, Jan. 2019.

[4] Y. Zhang, L. Li, R. Cong, X. Guo, H. Xu, and J. Zhang, "Co-saliency detection via hierarchical consistency measure," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[5] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE Trans. Image Process*, vol. 27, no. 2, pp. 568–579, Jan. 2018.

[6] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-based image retrieval by salient contour reinforcement," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1604–1615, Aug. 2016.

[7] D. Fan, W. Wang, M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8554–8564.

[8] C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han, "Hierarchical features driven residual learning for depth map super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2545–2557, May 2019.

[9] Q. Tang, R. Cong, R. Sheng, L. He, D. Zhang, Y. Zhao, and S. Kwong, "BridgeNet: A joint learning network of depth map super-resolution and monocular depth estimation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2148–2157.

[10] L. He, H. Zhu, F. Li, H. Bai, R. Cong, C. Zhang, C. Lin, M. Liu, and Y. Zhao, "Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9225–9234.

[11] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. ICML*, Jul. 2015, pp. 597–606.

[12] 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 Trans. Neural Netw. Learn. Syst.*, early access, Jun. 3, 2020, doi: 10.1109/TNNLS.2020.2996406.

[13] L. Wang, J. Zhang, Y. Wang, H. Lu, and X. Ruan, "CLIFFNet for monocular depth estimation with hierarchical embedding loss," in *Proc. ECCV*, 2020, pp. 316–331.

[14] H. Li, R. Cong, S. Kwong, C. Chen, Q. Xu, and C. Li, "Stereo superpixel: An iterative framework based on parallax consistency and collaborative optimization," *Inf. Sci.*, vol. 556, pp. 209–222, May 2021.

[15] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested network with two-stream pyramid for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9156–9166, Nov. 2019.

[16] R. Cong, Y. Zhang, L. Fang, J. Li, Y. Zhao, and S. Kwong, "RRNet: Relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5613311.

[17] C. Li, R. Cong, C. Guo, H. Li, C. Zhang, F. Zheng, and Y. Zhao, "A parallel down-up fusion network for salient object detection in optical remote sensing images," *Neurocomputing*, vol. 415, pp. 411–420, Nov. 2020.

[18] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2806–2813.

[19] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, 2020.

[20] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for RGB-D salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 665–681.

[21] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, "Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4296–4307, 2020.

[22] D.-P. Fan, Z. Lin, Z. Zhang, and M. Zhu, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Networks Learn.*, vol. 32, no. 5, pp. 2075–2089, May 2021.

[23] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3469–3478.

[24] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2343–2350.

[25] C. Lang, T.-V. Nguyen, H. Katti, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Proc. ECCV*, Florence, Italy, Oct, 2012, pp. 101–115.

[26] H. Peng, W. Xiong, W. Hu, and R. Ji, "RGB-D salient object detection: A benchmark and algorithms," in *Proc. ECCV*, Zurich, Switzerland, Sep. 2014, pp. 92–109.

[27] A. Ciptadi, T. Hermans, and J.-M. Rehg, "An in depth view of saliency," in *Proc. Brit. Mach. Vis. Conf.* Georgia Institute of Technology, 2013, pp. 1–11.

[28] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3049–3059.

[29] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13753–13762.

[30] 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 *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2094–2102.

[31] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for RGBD salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3922–3931.

[32] Q. Chen, K. Fu, Z. Liu, G. Chen, H. Du, B. Qiu, and L. Shao, "EF-Net: A novel enhancement and fusion network for RGB-D saliency detection," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107740.

[33] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.

[34] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, "Image super-resolution via dual-state recurrent networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1654–1663.

[35] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote. Sens.*, vol. 63, no. 6, pp. 691–699, 1997.

[36] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[37] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2048–2057.
[38] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[39] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 2156–2164.
[40] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7253–7262.
[41] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[42] Y. Liu, Z. Shao, Y. Teng, and N. Hoffmann, "NAM: Normalization-based attention module," 2021, *arXiv:2111.12419*.
[43] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 454–461.
[44] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1605–1616, Aug. 2017.
[45] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, Jul. 2014, pp. 23–27.
[46] G. Li and C. Zhu, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3008–3014.
[47] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3051–3060.
[48] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019.
[49] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognit.*, vol. 86, pp. 376–385, Feb. 2019.
[50] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2825–2835, Jun. 2019.
[51] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for RGB-D salient object detection and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5541–5559, Sep. 2022.
[52] S. Chen and Y. Fu, "Progressively guided alternate refinement network for RGB-D salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 520–538.
[53] J. Zhao, Y. Zhao, J. Li, and X. Chen, "Is depth really necessary for salient object detection?" in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1745–1754.
[54] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8579–8588.
[55] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu, and L. Cheng, "Calibrated RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9466–9476.
[56] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1407–1417.
[57] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3528–3542, 2021.
[58] Y. Han, L. Wang, A. Du, and S. Jiang, "LIANet: Layer interactive attention network for RGB-D salient object detection," *IEEE Access*, vol. 10, pp. 25435–25447, 2022.
[59] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao, "CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 6800–6815, 2022.
[60] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "RGB-D salient object detection via 3D convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 1063–1070.
[61] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "CCAFNet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images," *IEEE Trans. Multimedia*, vol. 24, pp. 2192–2204, 2022.
[62] F. Wang, R. Wang, and F. Sun, "DCMNet: Discriminant and cross-modality network for RGB-D salient object detection," *Exp. Syst. Appl.*, vol. 214, Mar. 2023, Art. no. 119047.
[63] S. Kanwal and I. A. Taj, "CVit-Net: A conformer driven RGB-D salient object detector with operation-wise attention learning," *Exp. Syst. Appl.*, vol. 225, Sep. 2023, Art. no. 120075.

HAITANG LI received the M.S. degree from the Henan University of Technology. He is currently with the Zhoukou Vocational College of Arts and Science. His research interests include computer vision and image processing.

YIBO HAN received the M.S. degree from the School of Information Science and Engineering, Xinjiang University. She is currently with Shanghai Tianma Microelectronics Company Ltd. Her research interests include image processing and object detection.

PEILING LI received the M.S. degree from the Henan University of Technology, in 2020. She is currently with the Zhengzhou Institute of Science and Technology. Her research interests include computer vision and image processing.

XIAOHUI LI received the bachelor's degree from Henan University, in 2019. He is currently with the International Joint Research Laboratory for Cooperative Vehicular Networks of Henan. His research interests include image processing and 3D vision technology.

LIJUAN SHI received the M.S. degree from Northwest University. She is currently an Associate Professor with the School of Big Data and Artificial Intelligence, Zhengzhou University of Science and Technology. Her research interests include multi-source information fusion and complex system computation and simulation.