Hindawi
Journal of Sensors
Volume 2024, Article ID 9818214, 1 page
https://doi.org/10.1155/2024/9818214

Retraction

Retracted: Application and Analysis of RGB-D Salient Object Detection in Photographic Camera Vision Processing

Journal of Sensors

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Q. Fu, "Application and Analysis of RGB-D Salient Object Detection in Photographic Camera Vision Processing," *Journal of Sensors*, vol. 2022, Article ID 5125346, 10 pages, 2022.

Hindawi
Journal of Sensors
Volume 2022, Article ID 5125346, 10 pages
https://doi.org/10.1155/2022/5125346

Research Article

Application and Analysis of RGB-D Salient Object Detection in Photographic Camera Vision Processing

Qiang Fu

Qingdao Vocational and Technical College of Hotel Management, Qingdao, Shandong Province, 266100 Shandong, China

Correspondence should be addressed to Qiang Fu; fuqiang@qchm.edu.cn

Received 23 August 2022; Revised 5 September 2022; Accepted 11 September 2022; Published 24 September 2022

Academic Editor: Sweta Bhattacharya

To identify the most visually salient regions in a set of paired RGB and depth maps, in this paper, we propose a multimodal feature fusion supervised RGB-D image saliency detection network, which learns RGB and depth data by two independent streams separately, uses a dual-stream side-supervision module to obtain saliency maps based on RGB and depth features for each layer of the network separately, and then uses a multimodal feature fusion module to fuse the latter 3 layers of RGB and depth high-dimensional information to generate high-level significant prediction results. Experiments on three publicly available datasets show that the proposed network outperforms the current mainstream RGB-D saliency detection models with strong robustness due to the use of a dual-stream side-surveillance module and a multimodal feature fusion module. We use the proposed RGB-D SOD model for background defocusing in realistic scenes and achieve excellent visual results.

1. Introduction

The purpose of image saliency detection is to extract regions of an image that are of more interest to humans by simulating human visual characteristics through intelligent algorithms, and it is promising for a wide range of applications in various computer vision tasks, such as image retrieval, image compression, and visual tracking [1]. More and more saliency detection research works in recent years have designed a large number of deep Convolutional Nerve Networks (CNNS) for RGB saliency target detection and achieved better performance [2]. Compared with traditional methods, deep learning can automatically extract features from a large amount of data. However, these RGB saliency detection models may not be able to distinguish salient targets from the background when the salient targets and the background are similar.

In fact, depth data contains clear target shapes and rich spatial structures, which can provide many additional saliency cues compared to RGB data, which provides detailed appearance and texture information. In addition, the perceptual robustness of depth sensors (e.g., Microsoft Kinect or Intel RealSense) to illumination changes greatly helps to extend the application scenarios of saliency detection. Therefore, for RGB-D saliency detection tasks, how to fully fuse RGB and depth information is the key issue. For how to clearly form the complementary information between the two modalities of RGB and depth and fully fuse them, most previous RGB-D fusion networks exploring the cross-modal complementarity of RGB and depth data are divided into two types of single-stream network architectures and dual-stream network architectures [3]. The single-stream network architecture considers that depth data can be used as an undifferentiated channel in tandem with RGB data to obtain salient maps by learning RGB and depth features together through the network. In the paper study [4], after superpixel segmentation of the input RGB data and depth data, the significant feature vector of each superpixel region is calculated, and then, the calculated significant value feature vector is used as the input of the network, which generates the significant graph by closely coupling the RGB information and depth information by combining the saliency features of the superpixels. The dual-stream network architecture learns RGB data and depth data separately through two independent streams and then learns the joint representation of RGB and depth features through a shared network layer added at an early or late stage to obtain the final saliency map. The study [5] inputs RGB data and depth data as two small networks, trains them separately, and then forms a fusion network with the generated RGB and depth features through multipath and multimodal interactions to train them together. The study [6] proposed a CNN-based framework to automatically fuse RGB and depth data to obtain salient maps. A late fusion network model is proposed in study [7] to capture the higher-order features of both RGB and depth modalities to generate the saliency map. The study [8] argues that this deep CNN feature that only fuses RGB and depth modalities is unlikely to capture the complementary information of cross modalities well. Therefore, a progressive complementary-aware fusion network is proposed to effectively utilize the cross-modal complementary information at multiple levels. It is widely believed that features at different levels are complementary and they abstract the scene at different scales. However, not all levels of cross-modal information are complementary.

Therefore, for the problem of how to fuse different levels of cross-modal information, a dual-stream network structure is used in this paper. Firstly, RGB and depth maps are used as network inputs for two VGG16Nets [9]. Furthermore, a dual-stream side-supervision module is used to significantly predict the supervision of RGB and depth streams to speed up the network convergence and help the network learn the features of each layer better. In order to fully utilize and fuse the semantic information of RGB and depth at different layers of the network, the final significant prediction results are obtained by adopting a high-level guidance of the network to the lower layers, from global to local. Among them, a multimodal feature fusion module is constructed to generate multiscale multimodal fusion features for the high-dimensional multimodal information in the last three layers of the network, so as to obtain the network high-level significant prediction results, while the multimodal feature fusion module is not used for the fusion of the features in the first two layers of the network because the low-level features of the network contain the target detail information. Significant prediction results will appear noisy. In order to eliminate the negative effect brought by low-dimensional depth features, this paper chooses not to include low-dimensional depth features in the low-level feature fusion. Experiments on widely used datasets show that the model in this paper outperforms the current mainstream RGB-D saliency detection model and has strong robustness. It can accurately detect salient target regions.

2. Related Work

2.1. RGB Salient Object Detection. Early 2D saliency target detection methods typically rely on hand-crafted features and heuristic priors such as image contrast, color, texture, and other low-level visual cues. Obviously, hand-crafted features are insufficient to capture high-level semantic information, so approaches based on these features are not universally applicable and can only achieve salient target detection in limited scenes.

Recently, benefiting from the development of convolutional neural networks (CNNs), some work has made great progress in using CNNs to learn deep features. Some deep learning-based saliency methods divide images into small blocks or superpixels and extract single or multiple scale features from each block or superpixel to determine whether an image region is salient or not. Although better performance than traditional methods has been obtained, processing images in a block-by-block manner ignores the underlying spatial information of the entire image, which limits the accuracy of complete salient target detection. The study [10] used a fully connected CNN to extract features and combine global and local features to predict the saliency map. Reference [11] proposes a cyclic CNN with a prediction map guided by a previous cyclic step. Reference [12] used a dropout technique to learn deep uncertain convolutional features in the network to enhance its generalization ability. However, since these methods only employ features extracted at the deeper layers of the CNN, they tend to miss details in salient objects captured mainly at the shallow layers. Several recent works have improved the quality of saliency object detection by further aggregating features across multiple CNN layers to exploit more global and local contextual information simultaneously in the inference process. Among them, study [13] explored the semantic properties and visual contrast of salient objects. Reference [14] created short connections to aggregate features in different layers. Reference [15] derives a resolution-based feature combination module and a boundary-preserving optimization strategy. Reference [16] iteratively aggregated deep features to exploit the complementary saliency information between multilevel features and features in each individual layer. Later, [17] used residual learning to alternately define deep and shallow features. Reference [18] formulated a bidirectional message passing model to selectively aggregate features to improve saliency target detection accuracy. Reference [19] designed an attention-guided network to progressively select and integrate multiple levels of information to predict saliency targets. Reference [20] designed a symmetric CNN to learn complementary saliency information and proposed weighted structural loss to enhance the boundaries of salient objects. Reference [21] explored global and local spatial relationships in deep networks to locate salient objects and define object boundaries. Despite the increasing detection quality, the exploration of global spatial contexts (especially in shallow layers) is still strictly limited by the convolution operator in CNNs, which are essentially local spatial filters. In recent years, 2D-based saliency target detection algorithms have developed rapidly, and even some algorithms have been applied in industry, but there are still challenges to tackle, such as unclear edge prediction of salient objects, incomplete prediction of transparent and reflective objects, and missed detection of small objects. For specific tasks, we should design corresponding models according to the characteristics of the data.

2.2. RGB-D Salient Object Detection. In the past, most traditional saliency target detection methods relied on hand-extracted features to capture local details and global contextual information separately or simultaneously. However, the lack of high-level semantic information limits their detection capability in complex scenes. Obviously, handcrafted features are not sufficient to capture high-level semantics, so approaches based on these features are not universally applicable and can only achieve salient target detection in limited scenes. Recently, thanks to the ability of convolutional neural networks to extract high-level semantic features and low-level detail features in a multiscale space, salient target detection has been rapidly developed. These neural network-based methods have made a qualitative leap in experimental results compared to traditional manual feature-based methods. Massive RGB-based salient target detection has focused on using color RGB images to identify salient objects, with good results. Although many RGB-based saliency target detection methods have achieved attractive performance, these methods may still fail to accurately detect salient regions when dealing with complex scenes because of the poor predictive power of the appearance feature contributions in RGB data. Examples include low-contrast scenes, transparent objects, similar foreground and background, multiple objects, and complex backgrounds. In these environments, it is difficult to determine salient targets by referring to RGB color images alone. With the advent of consumer-grade depth cameras such as Kinect cameras, light field cameras, and LiDAR, depth cues with large amounts of geometric and structural information have been widely used for salient object detection (SOD). To better mine salient information in challenging scenes, several CNN-based methods combine depth information with RGB letters to obtain more accurate results. Long-standing research has produced the practice and theory of extracting RGB and depth representations equally for symmetric two-stream structures. Reference [22] designed a symmetric structure for automatically fusing the features of depth and RGB views to obtain the final salient map. Reference [23] used a two-stream CNN-based model to introduce crossmodel interactions in multiple layers by direct summation. Recently, several asymmetric structures have been proposed to handle different data types. Reference [24] used enhanced depth information as an auxiliary cue and a pyramid decoding structure to obtain more accurate salient regions. Reference [25] proposed a structure that consists of a backbone network for processing RGB values and a subnetwork that makes full use of depth cues, which fuses depth-based features into the backbone network by direct cascading. However, simple fusion strategies like direct cascading or summation are not well suited for locating salient objects due to the infinite possibilities of their locations in the real world. Taken together, these approaches ignore the fact that depth cues contribute differently to salient object prediction in various scenarios. In addition, existing RGB-D methods inevitably suffer from loss of detail information when employing convolution steps and pooling operations in RGB and depth streams. An intuitive solution is to use hopping connections or short connections to reconstruct the detail information. Although these strategies mentioned above bring satisfactory improvements, they still struggle to accurately predict the complete structure.

3. Methodology

The method in this paper uses two VGG16Nets as the backbone base network, as shown in Figure 1. RGB and depth maps are used as inputs to extract RGB and depth features to form RGB streams and depth streams, respectively. Since the high-level features of the network acquire the high-dimensional semantic information of the salient targets and ignore the boundary information of the targets, therefore, in this paper, we adopt the high-level guidance to the low level, from deep to shallow and from global to local, to obtain the saliency map and multimodal fusion saliency map of each layer based on RGB and Depth features, respectively; and optimize the network parameters under the supervision of the truth map; and finally take the output of the saliency map of RGB stream as the final prediction result. For the network layer, the RGB and depth features are concatenated, and the significant output of the truth map supervised by its side and the significant output of the upper layer are used as the guidance to obtain the high level significant prediction results using the multimodal feature fusion module. For the network layer, the high-level feature fusion method is not used in the first two layers of feature fusion because the low-level features are more concerned with local information. And the low-layer depth information is not good to affect the final prediction results, so the depth flow supervision of the lower two layers is removed, and the significant maps of each layer are guided by the upper layer.

3.1. Two-Stream Lateral Supervision. According to [10, 11], it can be concluded that network supervision can promote network convergence speed and generate better hierarchical representations to meet the feature requirements at each stage. Considering that single convolution will cause the number of channels to plummet and lose more information, the method of gradually reducing the number of feature channels by using three convolutions is used, while ref. [11, 12] fully consider that the deep high-dimensional feature output retains more target and location information and ignores the target detail information, while the low-dimensional features focus more on local and boundary information. Therefore, the number of channels is reduced to 64 for each layer of VGG16Net backbone base network after three convolution operations and then combined with the saliency map output of the higher layer and then through convolution, deconvolution, and convolution operations to produce the saliency map of this layer. The network learning process is supervised with the true value map, which can help the network to learn the features of each layer better. The above operations are processed separately for the RGB stream and depth stream, as shown in Figure 2, and are specifically expressed in the following equation:

$$
\begin{aligned}
P^{m\mathrm{r}} &= \delta\left(\mathrm{Conv}\left(\mathrm{Dec}\left(\mathrm{Conv}\left(3\mathrm{Conv}\left(R^{m}\right)\cdot P^{m+1}\right)\right)\right)\right),\\
P^{m\,\mathrm{d}} &= \delta\left(\mathrm{Conv}\left(\mathrm{Dec}\left(\mathrm{Conv}\left(3\mathrm{Conv}\left(D^{m}\right)\cdot P^{m+1}\right)\right)\right)\right).
\end{aligned}
\tag{1}
$$

Based on the literature [13, 14] and the experiments in this paper, it can be concluded that the depth features of

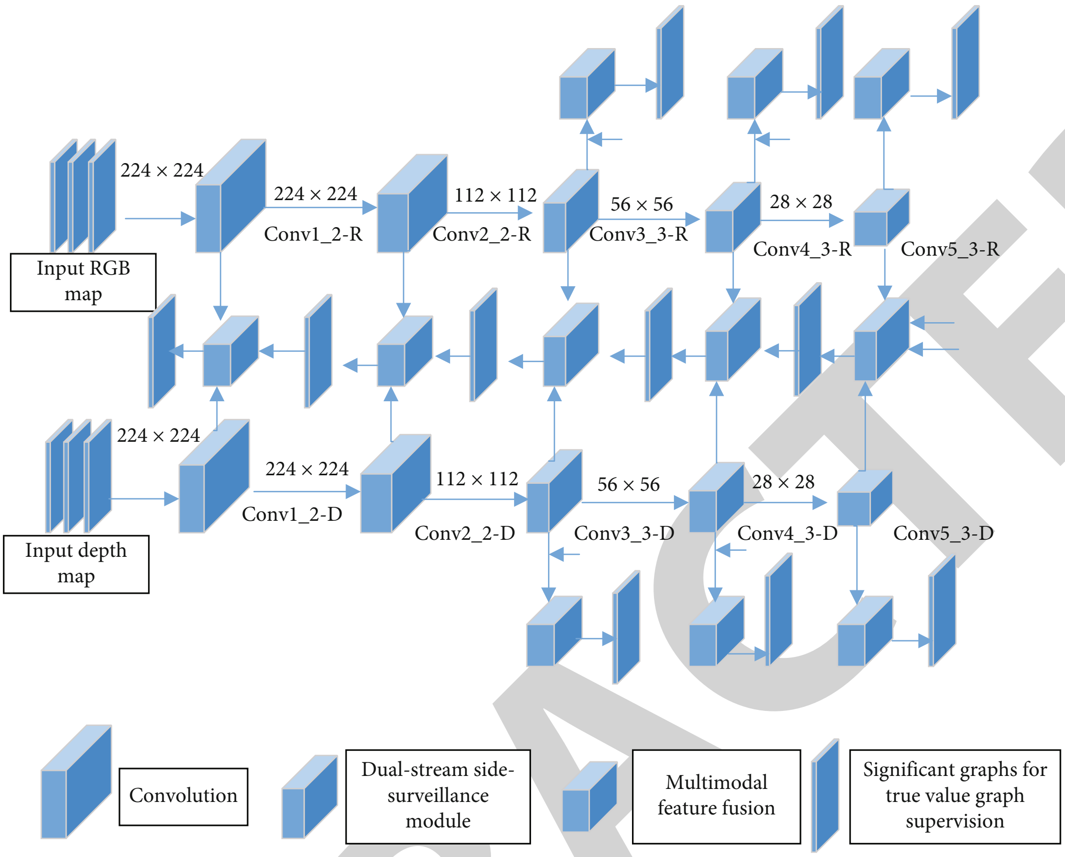
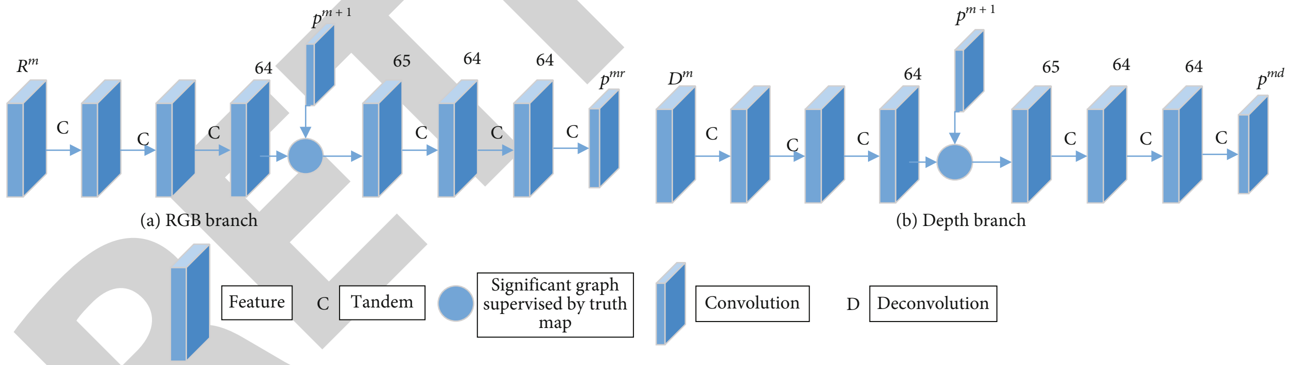


Figure 1: Flowchart of our method.



(a) RGB branch

(b) Depth branch

Figure 2: Dual-stream lateral supervision module.

the first two layers have low confidence; that is, the depth features containing more local information do not play a positive role in the detection of salient targets in the whole RGB-D image, and therefore, the extraction and supervision of the first two layers of features in the depth stream are eliminated. Also, it was found during the experiments that the indicated high-level saliency map works better when taking the saliency map of the fused high-level RGB stream and depth stream; therefore, the fused saliency map is used for the calculation of the side features of the high 3 layers of the dual streams in both the first and fifth rows on the right side of Figure 2.

3.2. Multimodal Feature Fusion. Considering that the features generated at the higher levels of the network have complete key information, simply generating the salient graph using one scale of convolution operation may pass the noise in some bad feature graphs to the salient prediction output without restriction. Therefore, in this paper, we propose a multimodal feature fusion method for RGB and depth features of layer 1 of the backbone network VGG16Net, as shown in Figure 3, where the two features are concatenated in series and the feature channels are reduced exponentially by the convolution operation,P^{m+1},P^{mr},$P^{m\mathrm{d}}$, and F^m. However, lacking the guidance of high-level information or

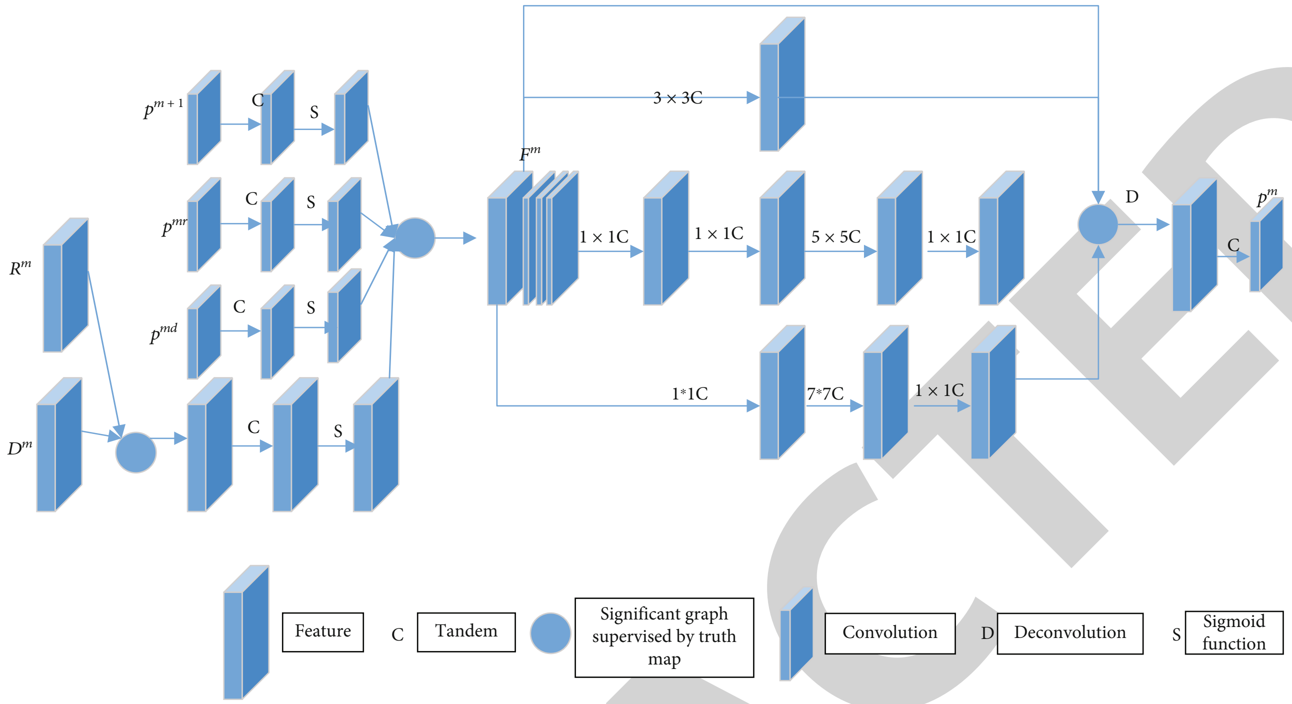


Figure 3: Multimodal feature fusion method.

output, the error between the side-by-side output and the true value map optimized directly by the supervised approach becomes larger and its effectiveness is also poor. Similar to the two-stream lateral supervision module, the high-level output and the supervised RGB stream output and depth stream output are used to provide semantic and positional information for the multimodal features to speed up the convergence of the network, optimize the target boundary, and obtain more suitable multimodal fusion features, which are calculated as follows (see Figure 3):

$$F^m = \delta(\mathrm{Conv}(R^m \cdot D^m)) \cdot \delta\left(\mathrm{Conv}\left(P^{m+1}\right)\right) \cdot \delta(\mathrm{Conv}(P^{mr})) \cdot \delta\left(\mathrm{Conv}\left(P^{m\,\mathrm{d}}\right)\right), \tag{2}$$

where F^m denotes the excitation function sigmoid, which normalizes feature values and significant values to the same interval to prevent significant output maps from being ignored.

RGB and depth multimodal feature fusion can complement and fuse the robust hierarchical feature representations across modal information and pave the way for generating hierarchical outputs based on multimodal features, compared with the way to process single RGB features and single depth features separately. After forming multimodal fusion features, this paper uses the multiscale convolution module to mine stronger fusion feature representations. The multiscale convolution module extracts multiscale contextual information, and its purpose is to obtain a spatial response mapping so as to adaptively weight the feature mapping at each location and to make each given input locate the most concerned part by learning weights for each pixel, thus making it more applicable to scenes with complex backgrounds. The multiscale convolution module uses 4 scales of convolution kernels (1 × 1, 3 × 3, 5 × 5, and 7 × 7), and the convolution layers with different kernel sizes have different sizes of perceptual fields to obtain feature information at different scales. At the same time, since larger convolution kernels correspond to more network parameters, this paper modifies the multiscale convolution module proposed in [16] by adding 1 × 1 convolution layer before and after the 5 × 5 convolution layer and 7 × 7 convolution layer to reduce the number of channels and then restore the number of channels to reduce the network parameters and then form multiscale multimodal fusion features by channel tandem.

$$f^m_{\mathrm{Ini}} = \mathrm{Conv}_{1\times1}(F^m),$$

$$f^m_{\mathrm{cat}} = f^m_{\mathrm{Ini}^m} \cdot \mathrm{Conv}_{3\times3}(f^m_{\mathrm{Ini}}) \cdot \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{5\times5}(\mathrm{Conv}_{1\times1}(f^m_{\mathrm{Ini}^m}))) \cdot \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{7\times7}(\mathrm{Conv}_{1\times1}(f^m_{\mathrm{Ini}}))). \tag{3}$$

Then, the significant output of the corresponding multiscale multimodal fusion feature is

$$P^m = \delta(\mathrm{Conv}(\mathrm{Dec}(f^m_{\mathrm{cat}}))). \tag{4}$$

The multimodal fusion features of the backbone network VGG16Net better combine the high-dimensional features of RGB and depth and better characterize the salient object features after processing by multiscale convolution. Because multiscale convolution corresponds to different convolution kernels, the larger the convolution kernel is, the larger its corresponding perceptual field is, and the more global information is seen. Low-dimensional features retain more information about target details, and a large convolution kernel may destroy its integrity. Therefore, this paper does not use the multiscale convolutional module in high-level fusion to process fused features in the first two layers of low-dimensional semantic information fusion part. The salient graph produced by the fusion network is similarly supervised by the truth graph to learn better multimodal fusion features. As mentioned in Section 2.1, the low-level depth information is not very reliable and will affect the final results, so in this paper, the depth stream supervision of the lower two layers is removed, and the saliency map of each layer is guided by the upper layer, and the final output of the saliency map of the RGB stream is taken as the final prediction.

4. Experimental Results

4.1. Dataset. This paper evaluates this model on three of the most widely used datasets. The NLPR1000 dataset contains 1000 RGB images and depth maps and their corresponding truth maps, containing 11 indoor and outdoor scenes with over 400 objects. The NJU2000 dataset contains 2003 stereo RGB images and their corresponding hand-labeled truth maps, whose depth maps are generated by the optical flow method. The STEREO data contains 797 RGB images and corresponding truth maps (GT), which were collected mainly from the Internet and 3D movies, and their depth maps were generated by the optical flow method. For a fair comparison, similar to the literature [8], their same training and test sets are used for training and evaluation. To solve the problem of insufficient training set, in this paper, the training set is subjected to a data enhancement operation; i.e., the original image is flipped, and the boundary 1/10 cropping operation is performed to retain the main target information, and the training set is increased by a factor of 16.

4.2. Evaluation Criteria. Evaluation criteria are used to evaluate the performance of different significant target detection methods. In this paper, five assessment criteria are used to evaluate the goodness of the model and other models. PR curve is generated by binarizing the significance map through a series of thresholds and then comparing it with the true value map. InF-measure, for the precision rate (Pre) and the detection rate (Rec), they are negatively correlated, and in order to balance the effect between them, F -measure is used to evaluate the experimental effect. The formula is

$$F_\beta = \frac{\left(1+\beta^2\right)\cdot \mathrm{Pre}\cdot \mathrm{Rec}}{\beta^2\cdot \mathrm{Pre}+\mathrm{Rec}}. \tag{5}$$

MAE: mean absolute error (MAE) evaluates the mean value of the absolute error between the significant and true value maps pixel by pixel. Its calculation formula is

$$\mathrm{MAE} = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H}|S(x,y)-G(x,y)|}{W\cdot H}, \tag{6}$$

where H and W are the length and width of the image, respectively, and $S(x,y), G(x,y)$ denote the significant and true values of the pixels, respectively.

S *-measure*: the structural similarity assessment criterion evaluates both the regional similarity and the target similarity between the salient and true value maps, which is defined as

$$S_\lambda = \lambda\cdot S_0 + (1-\lambda)\cdot S_r. \tag{7}$$

E *-measure*: E-measure measures statistical information at the image level and local pixel matching information. In order to have a fair comparison with other methods, all evaluation criteria were tested using the code provided in the literature [23].

4.3. Experimental Details. In this paper, experiments were conducted using Python and Caffe toolbox with GTX Titan-x GPUs (12 GB) machine configuration. The experimental training impulse, learning rate, weight decay rate, and minimum batch size are set to 0.99, 1e-10, 0.0005, and 1. The network structure of this paper is based on two pretrained VGG16Net networks, and the final network model is obtained by using them as the initial weights and fine-tuning the model training iterations for 10 cycles, totaling 160,000 times, which takes about 8 h.

4.4. Experimental Comparison. This model is compared with other TAN, PCFN, MMCI, and DF models under the above evaluation criteria, and the significant graphs are provided by the corresponding papers or generated by their provided codes. The model in this paper is compared with four representative deep learning-based models on PR curves as shown in Figure 4. From the figure, it can be seen that the model in this paper has significantly improved relative to these four models and generally outperformed them on other evaluation criteria. Table 1 shows the experimental results of this model on three datasets based on four evaluation criteria, F-measure, MAE, S-measure, and E-measure; compared with other models, the higher the value of F, S, and E, the better, and the smaller the value of MAE, the better.

4.5. Experimental Comparison of the Lateral Supervision Module. The experimental comparison of the dual-stream lateral supervision module in Section 2.1 is shown in Table 2. NDS (No Deep Supervised) means that the above supervision module is not used for the lateral output supervision, and only one convolution is used to make the number of lateral feature channels to 1 for supervision. From the experimental results in Table 2, it can be seen that the supervision module in this paper can retain feature information better in the supervision process and generate better hierarchical feature representations to meet the feature requirements at each stage.

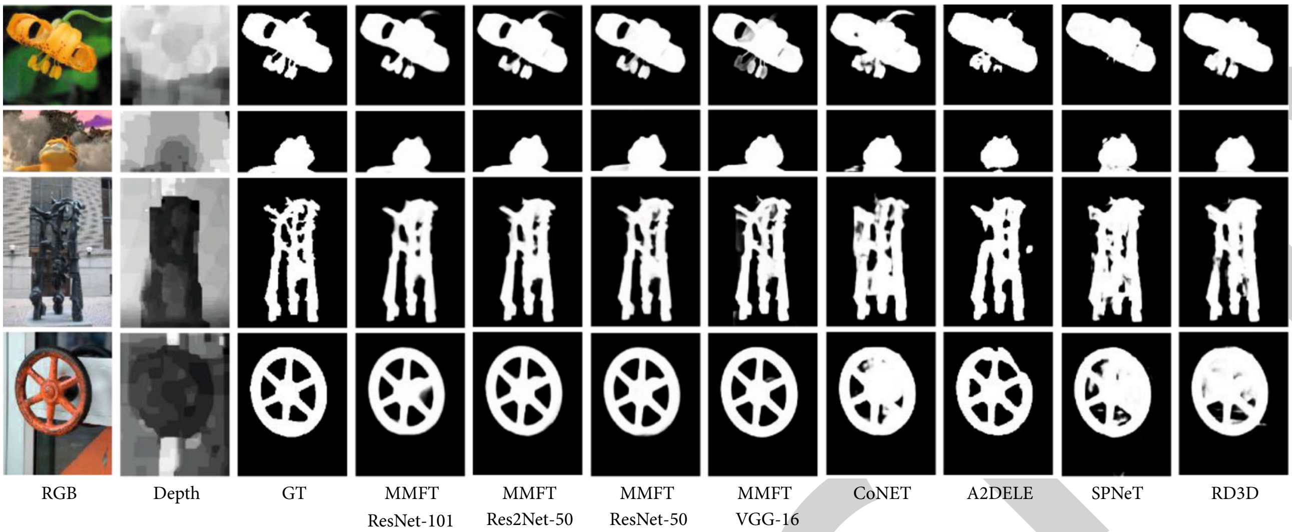


FIGURE 4: Visualization comparison of the model in this paper with the four models.

TABLE 1: Comparison with other models on *F*-measure, MAE, *S*-measure, and E-measure.

Algorithm	NLPR1000				NJU2000				STEREO			
	F	MAE	*S*	*E*	*F*	MAE	*S*	*E*	*F*	MAE	*S*	*E*
TAN	0.7955	0.0111	0.8862	0.9163	0.8440	0.0606	0.8786	0.8933	0.8488	0.0592	0.8774	0.9106
PCFN	0.7947	0.0436	0.8735	0.9162	0.8442	0.0592	0.8774	0.8967	0.8452	0.0608	0.8802	0.9055
MMCI	0.7398	0.0592	0.8556	0.8718	0.8123	0.0793	0.8586	0.8777	0.8122	0.0798	0.8558	0.8897
DF	0.7349	0.0892	0.7907	0.8600	0.7704	0.1405	0.7995	0.8384	0.7655	0.1396	0.7668	0.8425
Model of this paper	0.8628	0.0319	0.9116	0.9565	0.8589	0.0543	0.8856	0.8955	0.8623	0.0518	0.8896	0.9132

TABLE 2: Experimental comparison results of the effectiveness of the dual-stream lateral supervision module.

Algorithm	NLPR1000				NJU2000				STEREO			
	F	MAE	*S*	*E*	*F*	MAE	*S*	*E*	*F*	MAE	*S*	*E*
NDS	0.8356	0.0342	0.9081	0.9335	0.8501	0.0567	0.8847	0.8903	0.8525	0.0551	0.8877	0.9065
Model of this paper	0.8628	0.0319	0.9118	0.9465	0.8579	0.0542	0.8853	0.8955	0.8623	0.0518	0.8893	0.9132

TABLE 3: Experimental comparison results of multiscale module effectiveness.

Algorithm	NLPR1000				NJU2000				STEREO			
	F	MAE	*S*	*E*	*F*	MAE	*S*	*E*	*F*	MAE	*S*	*E*
BN	0.8487	0.0342	0.9058	0.9397	0.8505	0.0565	0.8812	0.8926	0.8571	0.0546	0.8849	0.9092
Model of this paper	0.8628	0.0317	0.9118	0.9465	0.8577	0.0542	0.8851	0.8955	0.8623	0.0518	0.8892	0.9132

4.6. Experimental Comparison of Multimodal Feature Fusion. The experimental comparison of multimodal feature fusion in Section 2.2 is shown in Table 3. BN indicates the results obtained by using only the normal convolution method without the improved multiscale convolution module. From the comparison results, it is found that multiscale convolution has an important role in significant computational results under the four evaluation criteria of *F*-measure, MAE, *S*-measure, and *E*-measure.

4.7. Experimental Comparison of Low-Dimensional Depth Features. In the experiments, it is found that depth features located in low dimensions are not good and affect the final results; as shown in Figures 5 and 6, depth 1 to depth 5 represent the significant maps corresponding to each stage of depth flow network, RGB1 to RGB5 represent the significant maps corresponding to each stage of RGB flow network, and Conv1 to Conv5 represent the significant maps generated by combining multimodal

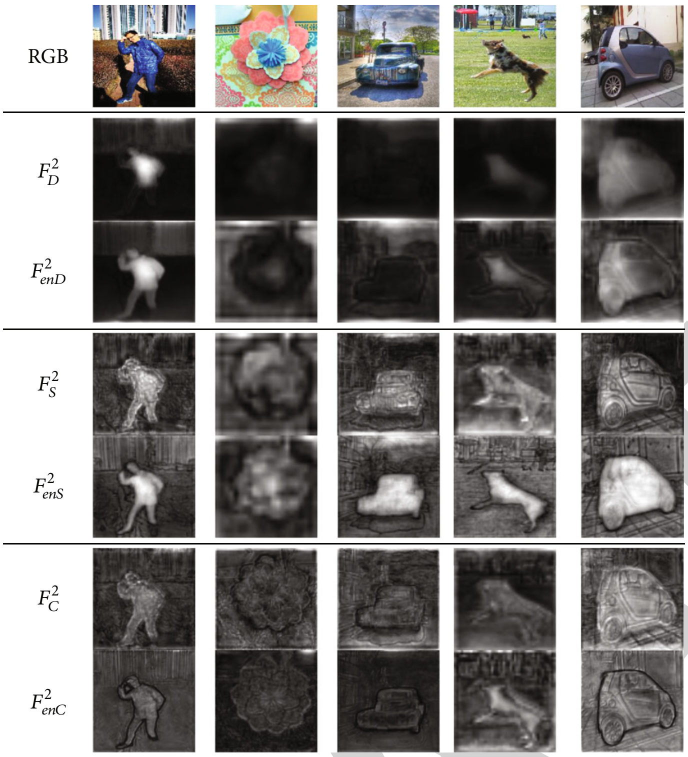


Figure 5: Visualization of the RGB-D fusion module.

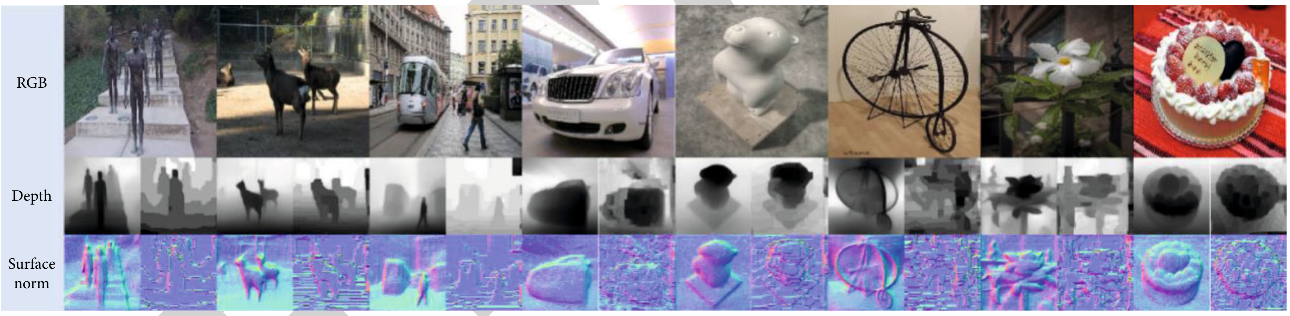


Figure 6: Visualization of the depth prediction.

Table 4: Experimental comparison results of low-dimensional depth features.

Algorithm	NLPR1000				NJU2000				STEREO			
	F	MAE	S	E	F	MAE	S	E	F	MAE	S	E
DY	0.8716	0.1088	0.8188	0.9477	0.8552	0.1312	0.8415	0.8786	0.8336	0.1278	0.8542	0.8985
Model of this paper	0.8628	0.0317	0.9118	0.9465	0.8577	0.0542	0.8851	0.8955	0.8621	0.0518	0.8895	0.9132

features. From Figure 5, it can be seen that due to the low-dimensional depth features of the depth flow network, the influence of depth 1 and depth 2 leads to noise in the low-level output, i.e., the final significant prediction results. Therefore, in order to eliminate the negative effect brought by low-dimensional depth features, this paper chooses not to include low-dimensional depth features in the low-level feature fusion, i.e., not to include depth 1 and depth 2. As shown in Figure 6, without the influence of low-dimensional depth features, this paper uses RGB1 as the final result, and from the visualization comparison, it can be seen that the effect has been significantly improved and the noise of negative impact. Table 4 shows the specific data comparison.

5. Conclusion

In this paper, we propose a CNN-based RGB-D saliency detection network, which consists of two modules to assist the network in guiding the lower levels from global to local and from deep to shallow, to obtain better saliency prediction results. The lateral supervision module facilitates the convergence speed of the network and generates better hierarchical representations to meet the requirements of each stage of features, while the multimodal feature fusion module obtains multiscale texture information of the high-level targets of the network and complements and fuses the robust high-level features with cross-modal information, proposing a different fusion method for the low-level features of the network than the high-level features. The proposed method can fuse the low-level features and high-level features extracted by the model. Information is not good leading to noise in the final prediction results, so the depth stream features of the lower two layers are removed from the final feature fusion. The experimental results on three widely used datasets show that the experimental results of this paper's method are generally better than the current mainstream algorithms and have stronger robustness. Future research can consider how to use depth information to make it more effective to assist RGB information to obtain better RGB-D saliency prediction results. The method in this paper can be applied to AI automatic bokeh of cameras in realistic scenes to speed up people's creative progress.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declared that there are no conflicts of interest regarding this work.

References

[1] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20–28, 2021.

[2] M. Xin and Y. Wang, "Research on image classification model based on deep convolution neural network," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, p. 11, 2019.

[3] X. Xie, X. Pan, W. Zhang, and J. An, "A context hierarchical integrated network for medical image segmentation," *Computers and Electrical Engineering*, vol. 101, article 108029, 2022.

[4] E. Kremic and A. Subasi, "Performance of random forest and SVM in face recognition," *International Arab Journal of Information Technology*, vol. 13, no. 2, pp. 287–293, 2016.

[5] J. Xia, N. Falco, J. A. Benediktsson, P. Du, and J. Chanussot, "Hyperspectral image classification with rotation random forest via KPCA," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 4, pp. 1601–1609, 2017.

[6] D. R. Nayak, R. Dash, and B. Majhi, "Brain MR image classification using two-dimensional discrete wavelet transform and AdaBoost with random forests," *Neurocomputing*, vol. 177, pp. 188–197, 2016.

[7] X. Xie, X. Pan, F. Shao, W. Zhang, and J. An, "MCI-Net: multiscale context integrated network for liver CT image segmentation," *Computers and Electrical Engineering*, vol. 101, article 108085, 2022.

[8] Y. Zhang, G. Cao, X. Li, and B. Wang, "Cascaded random forest for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 4, pp. 1082–1094, 2018.

[9] R. Chen, Z. Cai, and W. Cao, "MFFN: an underwater sensing scene image enhancement method based on multiscale feature fusion network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[10] Y. Li and R. Chen, "SE–RWNN: an synergistic evolution and randomly wired neural network-based model for adaptive underwater image enhancement," *IET Image Processing*, vol. 14, no. 16, pp. 4349–4358, 2020.

[11] M. Mathur and N. Goel, "Enhancement algorithm for high visibility of underwater images," *IET Image Processing*, vol. 16, no. 4, pp. 1067–1082, 2022.

[12] S. Raveendran, M. D. Patil, and G. K. Birajdar, "Underwater image enhancement: a comprehensive review, recent trends, challenges and applications," *Artificial Intelligence Review*, vol. 54, no. 7, pp. 5413–5467, 2021.

[13] X. Xie, W. Zhang, H. Wang et al., "Dynamic adaptive residual network for liver CT image segmentation," *Computers and Electrical Engineering*, vol. 91, article 107024, 2021.

[14] T. Li, S. Rong, W. Zhao et al., "Underwater image enhancement using adaptive color restoration and dehazing," *Optics Express*, vol. 30, no. 4, pp. 6216–6235, 2022.

[15] Y. Feng, J. Tang, B. Su, Q. Su, and Z. Zhou, "Point cloud registration algorithm based on the grey wolf optimizer," *Ieee Access*, vol. 8, pp. 143375–143382, 2020.

[16] D. Wu, Y. Lei, M. He, C. Zhang, and L. Ji, "Deep reinforcement learning-based path control and optimization for unmanned ships," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 7135043, 8 pages, 2022.

[17] F. Kulwa, C. Li, X. Zhao et al., "A state-of-the-art survey for microorganism image segmentation methods and future potential," *IEEE Access*, vol. 7, pp. 100243–100269, 2019.

[18] N. Kleefeldt, K. Bermond, I. S. Tarau et al., "Quantitative fundus autofluorescence: advanced analysis tools," *Translational Vision Science & Technology*, vol. 9, no. 8, pp. 2–2, 2020.

[19] L. Cheng, S. Chen, X. Liu et al., "Registration of laser scanning point clouds: a review," *Sensors*, vol. 18, no. 5, p. 1641, 2018.

[20] S. Palanisamy, B. Thangaraju, O. I. Khalaf, Y. Alotaibi, S. Alghamdi, and F. Alassery, "A novel approach of design and analysis of a hexagonal fractal antenna array (HFAA) for next-generation wireless communication," *Energies*, vol. 14, no. 19, p. 6204, 2021.

[21] Z. Dong, F. Liang, B. Yang et al., "Registration of large-scale terrestrial laser scanner point clouds: a review and benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 163, pp. 327–342, 2020.

[22] S. N. Alsubari, S. N. Deshmukh, A. A. Alqarni et al., "Data analytics for the identification of fake reviews using supervised learning," *CMC-Computers, Materials & Continua*, vol. 70, no. 2, pp. 3189–3204, 2022.

[23] M. M. Shanoer and F. M. Abed, “Evaluate 3D laser point clouds registration for cultural heritage documentation,” *The Egyptian Journal of Remote Sensing and Space Science*, vol. 21, no. 3, pp. 295–304, 2018.

[24] Y. Zou and T. Chen, “Laser vision seam tracking system based on image processing and continuous convolution operator tracker,” *Optics and Lasers in Engineering*, vol. 105, pp. 141–149, 2018.

[25] B. Zhang, S. Liu, and Y. C. Shin, “In-process monitoring of porosity during laser additive manufacturing process,” *Additive Manufacturing*, vol. 28, pp. 497–505, 2019.