electronics

MDPI

Article

MFDANet: Multi-Scale Feature Dual-Stream Aggregation Network for Salient Object Detection

Bin Ge [1,*], Jiajia Pei [1], Chenxing Xia [1,2] and Taolin Wu [1]

[1] College of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China; 2021201228@aust.edu.cn (J.P.); cxxia@aust.edu.cn (C.X.); 2021201297@aust.edu.cn (T.W.)
[2] Institute of Energy, Hefei Comprehensive National Science Center, Hefei 230031, China
* Correspondence: bge@aust.edu.cn

Abstract: With the development of deep learning, significant improvements and optimizations have been made in salient object detection. However, many salient object detection methods have limitations, such as insufficient context information extraction, limited interaction modes for different level features, and potential information loss due to a single interaction mode. In order to solve the aforementioned issues, we proposed a dual-stream aggregation network based on multi-scale features, which consists of two main modules, namely a residual context information extraction (RCIE) module and a dense dual-stream aggregation (DDA) module. Firstly, the RCIE module was designed to fully extract context information by connecting features from different receptive fields via residual connections, where convolutional groups composed of asymmetric convolution and dilated convolution are used to extract features from different receptive fields. Secondly, the DDA module aimed to enhance the relationships between different level features by leveraging dense connections to obtain high-quality feature information. Finally, two interaction modes were used for dual-stream aggregation to generate saliency maps. Extensive experiments on 5 benchmark datasets show that the proposed model performs favorably against 15 state-of-the-art methods.

Keywords: salient object detection; context information extraction; dense connection; dual-stream aggregation

check for updates

Citation: Ge, B.; Pei, J.; Xia, C.; Wu, T. MFDANet: Multi-Scale Feature Dual-Stream Aggregation Network for Salient Object Detection. *Electronics* **2023**, *12*, 2880. https://doi.org/10.3390/electronics12132880

Academic Editors: Christos J. Bouras and Hüseyin Kusetogullari

Received: 28 April 2023
Revised: 21 June 2023
Accepted: 26 June 2023
Published: 29 June 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the field of computer vision, salient object detection (SOD) is a critical task that involves identifying and extracting the most visually prominent objects in images and videos. Through its ability to identify regions of interest, SOD has found broad applications as a pre-processing step in various computer vision tasks. These include but are not limited to object recognition [1], semantic segmentation [2], robot navigation [3], and image retrieval [4].

Early SOD approaches [5–8] relied on handcrafted crafted, low-level features to generate saliency maps. However, the introduction of deep learning, specifically fully convolutional networks (FCNs) [9], has dramatically advanced SOD research by addressing the limitations of earlier methods in capturing high-level semantic information. Unlike prior approaches, deep learning models are capable of capturing both local and global features, resulting in significantly improved predictions.

Effective context information extraction is a current area of active research in SOD. Many current SOD methods [10–13] incorporate context information through the use of dilated convolutions with different filling rates to capture multi-scale context, which improves the salient object detection performance. However, these methods suffer from issues that may affect their ability to learn salient object information. Specifically, when using high filling rates to extract context information, dilated convolutions may lose local information and fail to capture correlations between the extracted context features.

Fully integrating features of different levels is crucial for improving the performance of SOD. While current SOD methods [14–17] mainly focus on constructing decoders for the feature fusion of different levels, they often overlook the correlation between features across different levels. Using a single interaction method may result in suboptimal feature representations, necessitating the exploration of new methods to better integrate the features.

MFDANet is a multi-scale feature dual-stream aggregation network designed for detecting salient objects. It comprises two components, the residual context information extraction (RCIE) module and the dense dual-stream aggregation (DDA) module. The RCIE module employs dilated convolutions and asymmetric convolutions with different filling rates to capture multi-scale context information while avoiding a loss of local information. The use of residual learning enables effective feature extraction through the avoidance of gradient vanishing. The DDA module utilizes dense connections and two interaction modes to aggregate different features, facilitating interaction among complementary information across different levels of features. This successfully reduces the loss of semantic information and produces high-quality feature representations.

In summary, the main contributions of this paper are summarized as follows:

(1) The proposed MFDANet network effectively addresses the issues in SOD by incorporating the RCIE and DDA modules to capture multi-scale context information and aggregate different level features.
(2) The RCIE module introduces a novel approach to extracting context information using dilated and asymmetric convolutions, which improves the discriminative power of features.
(3) The DDA module introduces two interaction modes for feature aggregation, allowing for information interactions between different level features and resulting in high-quality feature representations.
(4) Extensive experiments on 5 benchmark datasets demonstrate the superiority of the MFDANet method against 15 state-of-the-art SOD methods across diverse evaluation metrics.

2. Related Work

Salient object detection is used to detect and segment salient objects in real-world scenes. Traditional handcrafted saliency detection methods rely on low-level features and heuristic priors to detect salient objects, such as color contrast [5], texture contrast [6], center prior [7], and background prior [8]. However, salient object detection approaches that solely rely on handcrafted features and heuristic priors may have a limited performance in capturing high-level semantic information and understanding image content, leading to low-quality saliency maps.

The emergence of deep learning has led to the widespread adoption of convolutional neural networks (CNNs) [18] for a variety of computer vision applications, resulting in a remarkable performance [19–21]. By effectively capturing and utilizing multi-level features, CNNs have greatly enhanced the accuracy of saliency map detection. However, the traditional CNN architecture with fully connected layers can hinder computational efficiency and cause a loss of spatial structural information for salient objects, which can result in low-quality saliency maps. The introduction of fully convolutional networks (FCNs) has effectively solved these issues. FCNs can efficiently process inputs of arbitrary size and produce output feature maps with the same spatial resolution as the input, making them ideal for saliency detection.

2.1. Extract Multi-Scale Context Information

Currently, extracting multi-scale context information is an important research topic in salient object detection. For example, Wang et al. [12] proposed a global recurrent localization network that collects context information using a recurrent mechanism to iteratively refine convolutional features. They also used a boundary refinement mechanism to improve

the prediction map based on the spatial relationship between each pixel and its neighbors. Similarly, Qin et al. [13] proposed U2-Net, a deep network architecture that uses ReSidual U-blocks to combine receptive fields of different sizes, capturing context information from various scales and learning features at different levels. These methods demonstrate the significance of multi-scale contextual information in salient object detection.

2.2. Aggregate Multi-Level Features

Aggregating multi-level features is a beneficial approach for improving the performance of salient object detection. For example, Zhe et al. [15] proposed a cascaded partial decoder structure that discards shallow features while refining features using the generated saliency map. This approach discriminates features and quickly integrates features from different levels to achieve a faster and more accurate detection of salient objects. Similarly, Zhao et al. [18] employed a one-to-one guidance module to aggregate local edge information and global position information, leading to enhanced salient edge features and a better overall quality of the saliency map. Liu et al. [22] designed a feature aggregation module (FAM) that effectively integrates coarse-level semantic information with fine-level features from the top-down path, leading to a more accurate detection of salient objects. Zhang et al. [23] proposed integrating shallow and deep features to enhance salient object detection accuracy using forward and backward information propagation. These approaches demonstrate the importance of multi-level feature aggregation in improving the performance of salient object detection.

Although the previously mentioned methods have achieved good performance, some may not fully extract features or consider the correlation between features, or may only use a single interaction method, which can lead to suboptimal saliency maps. To address these issues, a new MFDANet network structure was designed. It includes a residual context extraction module and a dense dual-stream aggregation module that can fully capture multi-scale context information. Additionally, it utilizes dense connections through element-wise addition and multiplication in both streams to aggregate multiple levels of features and generate high-quality feature representations. This approach enables an effective and precise detection of salient objects in an efficient manner. By taking advantage of these advanced techniques, we can significantly improve the accuracy of salient object detection.

3. Proposed MFDANet Method

This section discusses the overall architecture of the salient object detection method, MFDANet, which is based on a multi-scale feature dense dual-stream aggregation network as proposed in Section 3.1. The two main components, the RCIE module and DDA module, are then described in Sections 3.2 and 3.3, respectively. Finally, in Section 3.4, the loss function used to train the network is introduced.

3.1. Overall Architecture

Figure 1 illustrates the overall architecture of the proposed MFDANet network. The MFDANet method aims to accurately detect salient objects in complex scenes by utilizing a multi-scale feature dual-stream aggregation network. The network architecture consists of two main parts: the RCIE module and the DDA module. The RCIE module extracts context information and employs multi-level features effectively, while the DDA module aggregates multi-scale features to provide comprehensive and detailed information for salient object detection.

The process begins by utilizing a ResNet-50 network that excludes the fully connected layers to extract initial features from the input image. This results in the creation of five initial features, denoted as $F = \{F_i | i = 1, 2, 3, 4, 5\}$. The RCIE module is then used to capture and utilize context information to generate five rich context information features $R = \{R_i | i = 1, 2, 3, 4, 5\}$. We use the DDA module to enhance the interaction between multi-level features, which generates five high-quality feature representations $D = \{D_i | i = 1, 2, 3, 4, 5\}$ to achieve an accurate prediction of salient objects.

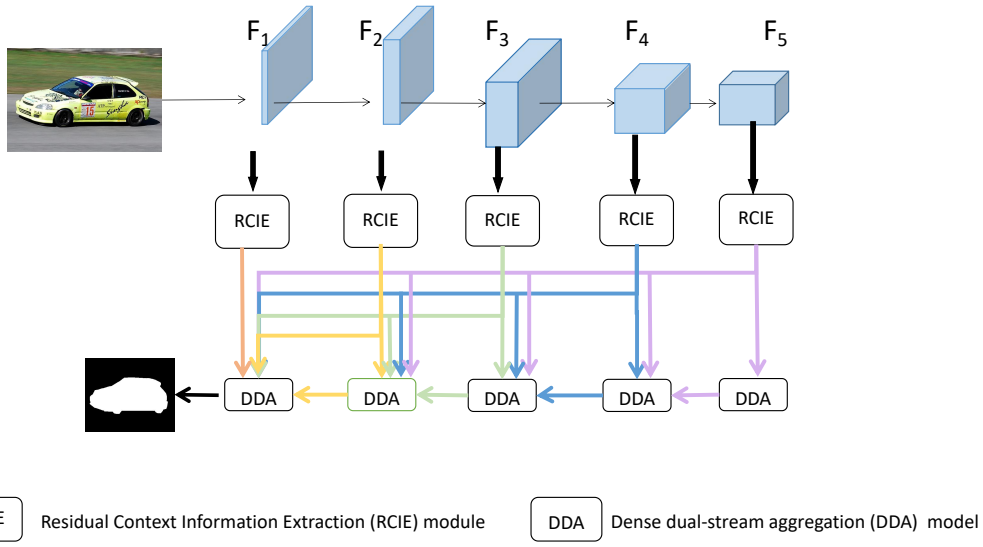


Figure 1. The framework of the proposed MFDANet method, which consists of a backbone network, a residual context information extraction (RCIE) module, and a dense dual-stream aggregation (DDA) module.

3.2. Residual Context Information Extraction (RCIE) Module

Effectively extracting contextual information is crucial for detecting salient objects. However, increasing the filling rate in dilated convolution, which is a commonly used technique to extract contextual information at different scales, may result in the occurrence of the checkerboard artifact, which can cause a loss of local information and compromise the accuracy of detection. Inspired by the RFB module in the paper [11], we designed an RCIE module to address the above issues as well as to extract richer and more effective contextual information (Figure 2). The RCIE module utilizes dilated convolutions with varying filling rates and asymmetric convolutions to capture and utilize context information at different scales. The parallel use of multiple dilated convolutions effectively enlarges the receptive field while avoiding the loss of local information caused by increasing filling rates.

Moreover, the input of dilated convolution in the current stage is the summation of the output of the dilated convolution in the previous stage and the feature maps with reduced dimensionality. The input of asymmetric convolution in the current stage is the summation of the output of the asymmetric convolution in the previous stage and the output of the dilated convolution in the current stage. These approaches maintain the correlation between features at different stages and reduce the complexity of the network, leading to lower computation and memory consumption.

The convolution operation's receptive field varies across the stages, and, during the nth stage specifically, multi-scale context information is extracted using an asymmetric convolution operation with a kernel size of $1 \times (2n-1)$ and $1 \times (2n-1)$; then, the dilated convolution is performed with a kernel size of 3×3 and different filling rates of 1, 2, 4, and 6. By capturing context features $M = \{M_n | n = 1, 2, 3, 4\}$ at different scales using the RCIE module, these features are concatenated M_n and $\rho(F_i)$ and then the dimensionality is reduced while introducing a residual connection to avoid the problem of gradient disappearance. Finally, five multi-level context features $R = \{R_i | i = 1, 2, 3, 4, 5\}$ are obtained, which contain rich and salient information, and can be summarized in the following equation:

$$R_i = \sum(C(\rho((F_i), \rho(Cat((M_1, \ldots, M_n)))))), n = 4 \quad (1)$$

In this equation, "$\sum$", "C", and "Cat" represent element-wise addition, convolution operation, and concatenation, respectively. The variable "ρ" denotes the dimensionality reduction operation, which sets the convolution kernel to $1 \times 1 \times 148$.

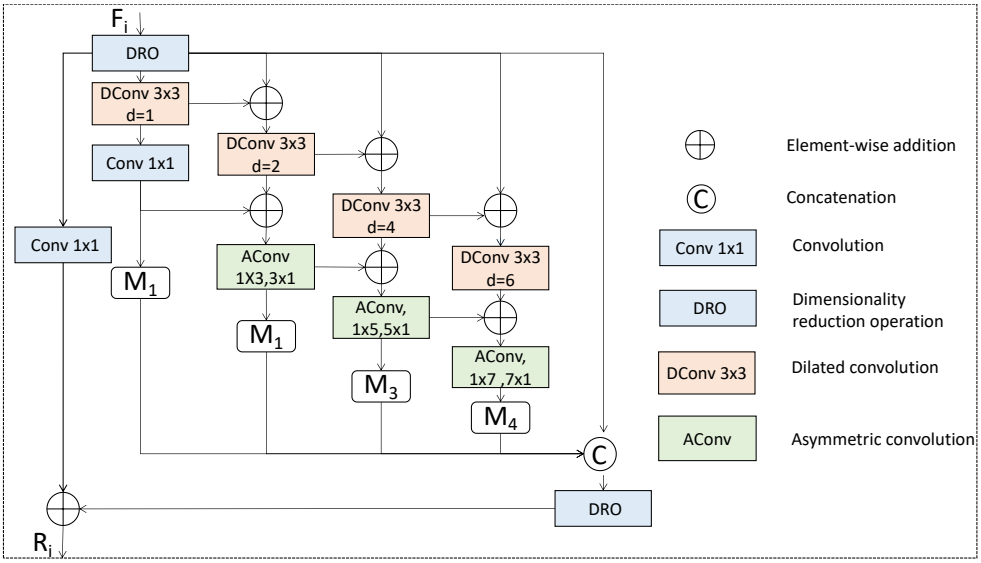


Figure 2. Details of residual context information extraction (RCIE) module.

3.3. Dense Dual-Stream Aggregation (DDA) Module

High-level features are rich in semantic information, while low-level features contain detailed information. Aggregating high-level and low-level features with a single interaction mode is not efficient for exchanging information, hindering the efficient generation of saliency maps. Moreover, deep semantic feature information tends to gradually weaken during the decoding stage's bottom-up transmission process, causing salient objects to lose high-level semantic information guidance as they pass through multiple convolutional layers. This, in turn, can lead to a decreased model detection performance. Therefore, it is crucial to explicitly connect the correlation between different level features and enable effective interactions among multiple pieces of information in multi-level features to improve the accuracy of detecting salient objects. Inspired by the FPN, we designed the DDA module to address the problem, which enables efficient interactions with multi-level features, resulting in powerful feature representations and an increased model stability and accuracy.

The details of the DDA module are shown in Figure 3. The DDA module uses the features R_i, R_{i+1}, R_{i+2} generated by the RCIE module and the features D_i from the previous level as inputs for the interaction of different level features to generate high-quality feature representations. Specifically, the DDA module consists of densely connected feature interactions, which include element-wise addition and multiplication. Firstly, two interaction modes are used to fuse different level features to generate feature T_i^1 and T_i^2, which is then concatenated with another feature and dimension-reduced to generate features with different saliency information. Channel attention is then applied to alleviate the noise in the features.

To further enhance the diversity of information and ensure the generation of robust feature representations, we introduced residual connections to the module. The resulting feature set is denoted as $D = \{D_i | i = 1, 2, 3, 4, 5\}$, and enables accurate inference on salient objects. The complete process can be expressed as follows:

$$\begin{cases} T_4^1 = \sum(D_5, R_4, R_5) \\ T_4^2 = D_5 \otimes R_4 \otimes D_5 \\ T_3^1 = \sum(D_4, R_3, R_4, R_5) \\ T_3^2 = D_4 \otimes R_3 \otimes R_4 \otimes D_5 \\ T_2^1 = \sum(D_3, R_2, R_3, R_4, R_5) \\ T_2^2 = D_3 \otimes R_2 \otimes R_3 \otimes R_4 \otimes D_5 \\ T_1^1 = \sum(D_2, R_1, R_2, R_3, R_4, R_5) \\ T_1^2 = D_2 \otimes R_1 \otimes R_2 \otimes R_3 \otimes R_4 \otimes D_5 \\ T_i = \sum CA\left(\rho\left(Cat\left(T_i^1, T_i^2\right)\right), \rho\left(Cat\left(T_i^1, T_i^2\right)\right)\right) \end{cases} \quad (2)$$

$$D_i = \begin{cases} \rho(\sum(CA(R_i), R_i)) & i = 5 \\ \rho(T_i) & i = 1, 2, 3, 4 \end{cases} \tag{3}$$

In this equation, the symbols "$\sum$", "$\otimes$", "CA", "Cat" represent element-wise addition, element-wise multiplication, channel attention operation, and concatenation, respectively. The variable "ρ" denotes the dimensionality reduction operation, which sets the convolution kernel to $1 \times 1 \times 148$.

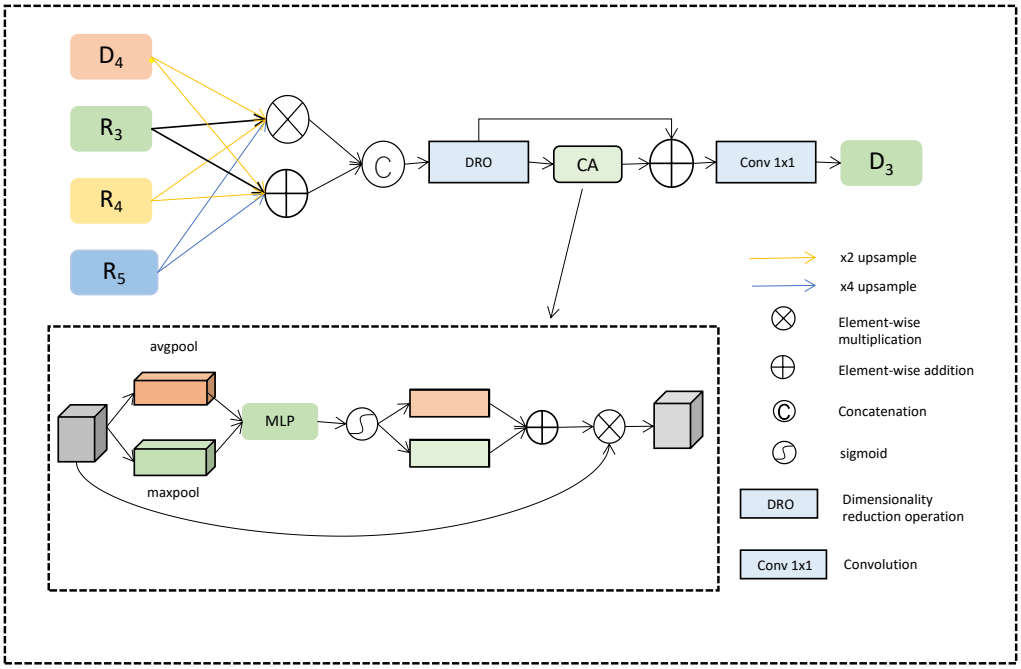


Figure 3. Details of the dense dual-stream aggregation module with D_3 as an example.

3.4. Loss Function

To obtain a complete saliency map with clear boundaries, the loss function of this model is defined as:

$$L_{all} = L_{bce} + L_{iou}, \tag{4}$$

Here, L_{bce} and L_{iou} denote the binary cross-entropy (BCE) loss function [24] and the IoU loss function [25], respectively. The BCE loss function is commonly used for segmentation and binary classification tasks and is defined as follows:

$$L_{bce} = -\sum_{(m,n)} [Y(m,n) \log (X(m,n)) + (1 - Y(m,n)) \log (1 - X(m,n))], \tag{5}$$

where X and Y represent the predicted salient object and ground truth, respectively. $X(m,n) \in [0,1]$ and $Y(m,n) \in [0,1]$ denote the probability of the predicted salient objects and the ground-truth label at position (m,n), respectively. In order to focus on more complete saliency regions, the IoU loss function is defined as follows:

$$L_{iou} = 1 - \frac{\sum_{(m,n)} [X(m,n) \times Y(m,n)]}{\sum_{(m,n)} [Y(m,n) + X(m,n) - Y(m,n) \times X(m,n)]}, \tag{6}$$

The proposed MFDANet method was trained using the Adam optimizer, which generated the final saliency map.

4. Experiment Analysis

4.1. Datasets

The proposed method was thoroughly evaluated for its effectiveness on five publicly available datasets, namely ECSSD [26], PASCAL-S [27], HKU-IS [28], DUT-OMRON [6], and DUTS-TE [29]. In addition, the DUTS dataset consists of two subsets, the DUTS-TR, which contains 10,553 images used for training, and the DUTS-TE, with 5019 images used for testing.

The ECSSD dataset is composed of 1000 images with diverse and intricate background scenes, while the PASCAL-S dataset encompasses 850 images that exhibit varying types and complex structures of salient objects. The HKU-IS dataset contains 4447 images that feature multiple salient objects, whereas the DUT-OMRON dataset consists of 5168 images that exhibit complex structures and diverse types of salient objects.

4.2. Experimental Details

We implemented our method, MFDANet, using the PyTorch framework and conducted training and testing on a PC machine equipped with an NVIDIA GTX 3090 GPU. To capture initial multi-level features, we utilized a pre-trained ResNet-50 network. During the training process, we randomly cropped and resized the input images to 320×320 and applied random horizontal flipping and rotation to increase the model's robustness. To optimize the loss, we used the Adam optimizer with an initial learning rate of 5×10^{-5}, a momentum of 0.9, and a weight decay of 5×10^{-4}, with a batch size of 16. The training process was carried out for 100 epochs.

4.3. Evaluation Metrics

In this experiment, six evaluation metrics were used to assess the performance of MFDANet, which includes the PR curve, F-measure (F_β) curve, mean absolute error (MAE), average F-measure (AF_β), weighted F-measure (WF_β), and E-measure (E_m).

PR curve: The precision–recall (PR) curve is a graphical plot of the correlation between precision and recall metrics for various threshold values within the saliency score range of [0, 255].

F-measure (F_β): F_β is a combination of precision and recall, and it provides a balanced evaluation of both metrics. It is defined as follows:

$$F_\beta = \frac{(1+\beta^2) \times N_{Precision} \times N_{recall}}{\beta^2 \times N_{Precision} + N_{recall}}, \tag{7}$$

where $\beta^2 = 0.3$ as suggested by previous work [5] to emphasize precision more than recall.

Mean absolute error (MAE) measures the average difference between the ground truth and the saliency map. MAE is a popular metric for evaluating salient object detection algorithms, and is defined as follows:

$$MAE = \frac{1}{W \times H} \sum_{m=1}^{W} \sum_{n=1}^{H} |S(m,n) - G(m,n)|, \tag{8}$$

where H and W are the height and width of the input image, respectively. S represents the saliency map outputted by the network, and G is the ground truth.

E-measure (E_m): The E_m estimates the similarity between the predicted saliency map and the ground truth at object level and pixel level. It is defined as follows:

$$E_m = \frac{1}{W \times H} \sum_{m=1}^{W} \sum_{n=1}^{H} A(x), \tag{9}$$

where x is the alignment matrix and $A(x)$ represents the augmented alignment matrix

4.4. Comparison with State-of-the-Art Experiments

To demonstrate the effectiveness of the proposed MFDANet method, a quantitative and qualitative comparison was conducted with 15 state-of-the-art salient object detection methods, including DGRL [10], BPMB [23], EGNet [17], F3Net [16], CPD [14], PoolNet [22], AFNet [30], TSPOANet [21], GateNet [13], CAGNet [31], ITSD [32], GCPANet [33], SUCA [19], DSRNet [34], and ICON [35]. To ensure a fair comparison, we chose ResNet-50 as the backbone network for all the compared state-of-the-art SOD methods, and all the compared saliency maps were either directly provided by the authors or generated from publicly available source codes. Moreover, the same evaluation code was used to generate different evaluation metrics for all the saliency maps. In addition, we used the same evaluation code to compute the results for all saliency maps.

4.4.1. Quantitative Comparison

Table 1 shows the quantitative comparison results between our proposed MFDANet method and 15 other state-of-the-art salient object detection methods on 5 public benchmark datasets, using 4 evaluation indicators: MAE, AF_β, WF_β, and E_m). MFDANet outperforms other methods in terms of these four evaluation metrics, with improvements observed across all benchmark datasets.

For instance, MFDANet achieves an AF_β score of 0.934 on the ECSSD dataset, which surpasses that of ICON [35] (0.928). Similarly, on the PASCAL-S dataset, MFDANet achieves an AF_β score of 0.838, exceeding that of the F^3Net [16] (0.835), and, on the HKU-IS dataset, MFDANet achieves an AF_β score of 0.923, higher than that of ICON (0.810). On the DUT-OMRON dataset and the DUTS-TE dataset, MFDANet achieves AF_β scores of 0.783 and 0.859, respectively, outperforming ICON (0.772) and ICON(0.840), respectively. Specifically, we compared our proposed method with other methods using four evaluation metrics on five datasets. Our results show that the proposed method achieves a higher accuracy and recall in detecting salient objects, and the performance of AF_{beta} improves by 0.6%, 0.3%, 1.43%, 1.42%, and 2.26% on the ECSSD, PASCAL-S, HKU-IS, DUT-OMRON, and DUTS-TE datasets, respectively. Furthermore, our results show that the MAE scores of the proposed method decrease by 3.70%, 1.92%, and 2.8% on the HKU-IS, DUT-OMRON, and DUTS-TE datasets, respectively, indicating its superior ability to locate salient objects and make accurate predictions. It also shows that the proposed MFDANet method is able to correctly predict more pixels of prominent objects. In addition, the significant improvements achieved by the proposed method on W_β and E_m evaluation metrics indicate its excellent performance in detecting and segmenting salient objects.

In addition, Figure 4 displays the PR curves and F-measure curves on five publicly available SOD datasets. Our proposed method (red ones) shows outstanding results compared to other SOD methods in terms of PR curves and F-measure curves. These outcomes serve as evidence of MFDANet's robustness in comparison to other approaches.

Table 1. The proposed MFDANet compared with 15 state-of-the-art methods on 4 evaluation metrics: MAE (↓), AF_β (↑), WF_β(↑), and E_m (↑) across 5 datasets. The best result for each row is highlighted in bold.

Method		DGRL [10]	BDMPM [23]	EGNet [17]	F3Net [16]	CPD [14]	PoolNet [22]	AFNet [30]	TSPOANet [21]	GateNet [13]	CAGNet [31]	ITSD [32]	GCPANet [33]	SUCA [19]	DSRNet [34]	ICON [35]	Ours
Publish		$CVPR_{18}$	$CVPR_{18}$	$ICCV_{19}$	$AAAI_{19}$	$CVPR_{19}$	$CVPR_{19}$	$CVPR_{19}$	$ICCV_{19}$	$ECCV_{20}$	PR_{20}	$CVPR_{20}$	$AAAI_{20}$	TMM_{21}	$TCSVT_{21}$	$TPAMI_{22}$	
ECSSD(1000)	MAE ↓	0.046	0.045	0.037	0.033	0.037	0.039	0.042	0.046	0.040	0.037	0.035	0.035	0.040	0.039	**0.032**	**0.032**
	AF_β ↑	0.893	0.869	0.920	0.925	0.917	0.915	0.908	0.900	0.916	0.921	0.895	0.919	0.916	0.910	0.928	**0.934**
	WF_β ↑	0.871	0.871	0.903	0.912	0.898	0.896	0.886	0.876	0.894	0.902	0.911	0.903	0.894	0.891	0.918	**0.921**
	E_m ↑	0.935	0.916	0.947	0.946	0.950	0.945	0.941	0.935	0.943	0.944	0.932	0.952	0.943	0.942	**0.954**	0.953
PASCAL-S(850)	MAE ↓	0.077	0.073	0.074	**0.061**	0.071	0.075	0.070	0.077	0.067	0.066	0.066	0.062	0.067	0.067	0.064	0.064
	AF_β ↑	0.794	0.758	0.817	0.835	0.820	0.815	0.815	0.804	0.819	0.833	0.785	0.827	0.819	0.819	0.833	**0.838**
	WF_β ↑	0.772	0.774	0.795	0.816	0.794	0.793	0.792	0.775	0.797	0.808	0.812	0.808	0.797	0.800	0.818	**0.820**
	E_m ↑	0.863	0.840	0.877	0.895	0.887	0.876	0.885	0.871	0.884	0.896	0.863	**0.897**	0.884	0.883	0.893	0.896
HKU-IS(4447)	MAE ↓	0.041	0.039	0.031	0.028	0.033	0.032	0.036	0.038	0.033	0.030	0.031	0.031	0.033	0.035	0.029	**0.027**
	AF_β ↑	0.875	0.871	0.902	0.910	0.895	0.900	0.888	0.882	0.899	0.909	0.899	0.898	0.899	0.893	0.910	**0.923**
	WF_β ↑	0.851	0.859	0.886	0.900	0.879	0.883	0.869	0.862	0.880	0.893	0.894	0.889	0.880	0.873	0.902	**0.909**
	E_m ↑	0.943	0.938	0.955	0.958	0.952	0.955	0.948	0.946	0.953	0.950	0.953	0.956	0.953	0.951	0.959	**0.963**
DUT-OMRON(5168)	MAE ↓	0.066	0.064	0.053	0.053	0.056	0.056	0.057	0.061	0.055	0.054	0.061	0.056	-	0.061	0.057	**0.052**
	AF_β ↑	0.711	0.692	0.756	0.766	0.747	0.739	0.739	0.716	0.746	0.752	0.756	0.748	-	0.727	0.772	**0.783**
	WF_β ↑	0.688	0.681	0.738	0.747	0.719	0.721	0.717	0.697	0.729	0.728	0.750	0.734	-	0.711	0.761	**0.765**
	E_m ↑	0.847	0.839	0.874	0.876	0.873	0.864	0.860	0.850	0.868	0.862	0.867	0.869	-	0.855	0.879	**0.886**
DUTS-TE(5019)	MAE ↓	0.054	0.049	0.039	0.035	0.043	0.040	0.046	0.049	0.040	0.040	0.041	0.038	0.044	0.043	0.037	**0.036**
	AF_β ↑	0.755	0.746	0.815	0.840	0.805	0.809	0.793	0.776	0.807	0.837	0.804	0.817	0.802	0.791	0.838	**0.859**
	WF_β ↑	0.748	0.761	0.816	0.835	0.795	0.807	0.785	0.767	0.809	0.817	0.824	0.821	0.803	0.794	0.837	**0.846**
	E_m ↑	0.873	0.863	0.907	0.918	0.904	0.904	0.895	0.885	0.903	0.913	0.898	0.913	0.903	0.892	0.919	**0.927**

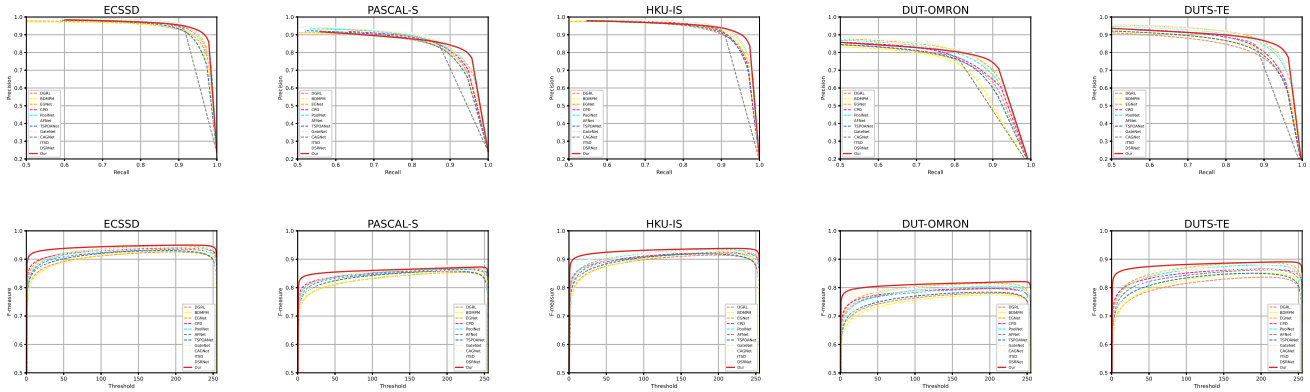


Figure 4. Quantitative comparisons of the proposed approach MFDANet and other state-of-the-art baseline methods on five datasets. The first and second rows are the PR curves and F-measure curves of different methods respectively.

4.4.2. Qualitative Analysis

Figure 5 presents a visual comparison between our proposed MFDANet method and 11 state-of-the-art SOD methods. The saliency maps generated by our approach demonstrate a remarkable similarity to the ground truth across a wide range of scenes.

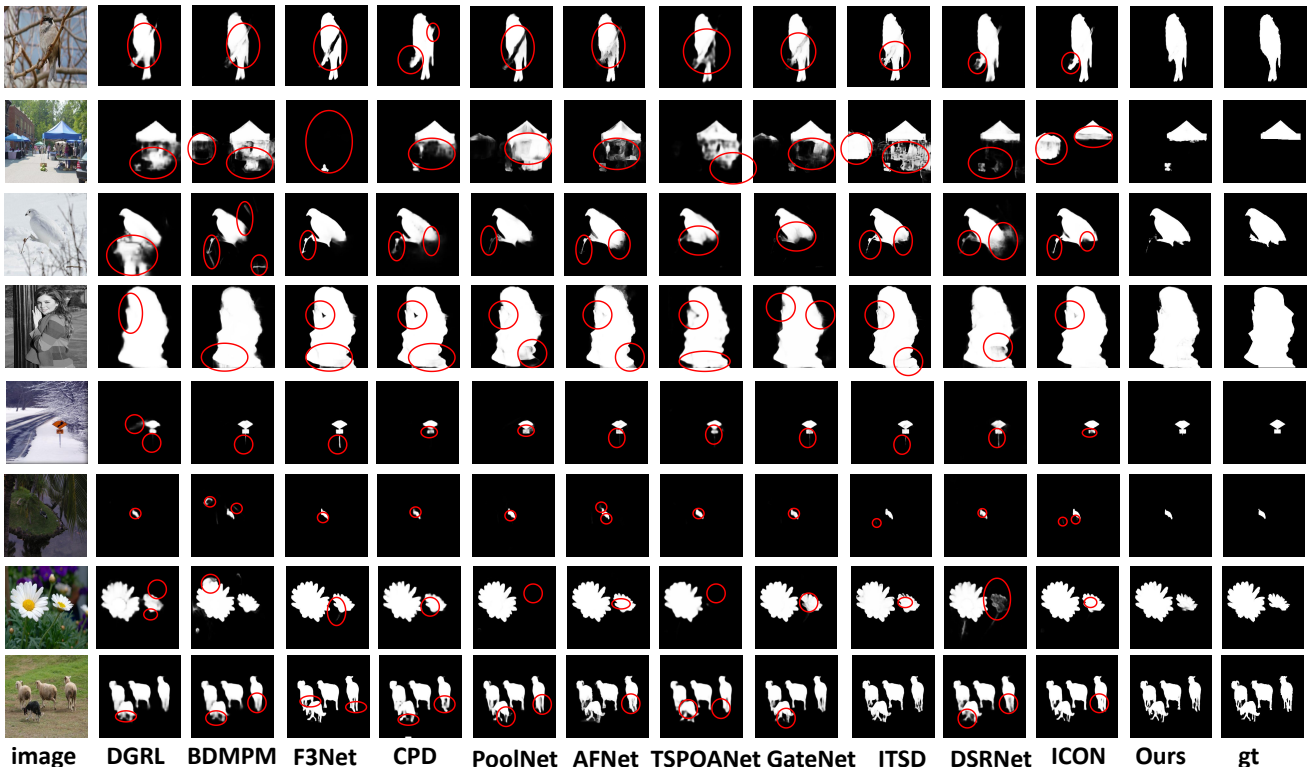


Figure 5. Qualitative comparison between our MFDANet method and 11 recent state-of-the-art SOD methods.

For objects with different scales and shapes, MFDANet effectively highlights salient objects in complex backgrounds with different scales and shapes (rows 1–2), is robust in dealing with low foreground–background contrast (rows 3–4) and small objects (rows 5–6), and generates effective saliency maps even in images with multiple identical semantic information (rows 7–8). Compared to other state-of-the-art methods, MFDANet generates more complete and accurate saliency maps in a variety of scenes.

4.5. Ablation Experiment

We conducted ablation experiments aimed at evaluating the effectiveness of each module in MFDANet, including the RCIE module and the DDA module. For this purpose, we incrementally added the RCIE module and the DDA module using the ResNet-50 network and the FPN as a baseline and trained them on the DUTS-TR dataset. Then, we tested the performance of MFDANet on five public benchmark datasets and present the results in Table 2. The ablation experiments effectively highlight the contribution of each module to the overall performance of MFDANet and validate our design choices.

Table 2. Quantitative results on five public SOD datasets: the best results are marked in bold black. The symbol "↑/↓" indicates that the larger the result, the better the result, and the smaller the result, the better the result, respectively.

Datasets		Res + FPN	Res + RFB + FPN	Res + DDA	Res + RCIE + FPN	Res + RCIE + DDA (add)	Res + RCIE + DDA (mul)	Res + RCIE + DDA (cat)	Res + RCIE + DDA
ECSSD(1000)	MAE ↓	0.055	0.050	0.040	0.046	0.033	0.036	**0.031**	0.032
	AF_β ↑	0.873	0.879	0.903	0.887	0.929	0.922	0.931	**0.934**
	WF_β ↑	0.856	0.870	0.893	0.883	0.916	0.906	0.919	**0.921**
	E_m ↑	0.917	0.920	0.944	0.925	0.955	0.948	**0.956**	0.953
PASCAL-S(850)	MAE ↓	0.078	0.075	0.067	0.074	**0.064**	**0.064**	0.065	**0.064**
	AF_β ↑	0.772	0.769	0.816	0.771	0.836	0.834	0.837	**0.838**
	WF_β ↑	0.764	0.777	0.797	0.785	0.817	0.813	0.818	**0.820**
	E_m ↑	0.861	0.858	0.894	0.086	0.902	0.897	**0.900**	0.895
HKU-IS(4447)	MAE ↓	0.047	0.043	0.032	0.039	**0.027**	0.030	**0.027**	**0.027**
	AF_β ↑	0.878	0.886	0.889	0.899	0.914	0.911	0.917	**0.923**
	WF_β ↑	0.832	0.848	0.884	0.868	0.905	0.899	0.908	**0.909**
	E_m ↑	0.943	0.946	0.954	0.954	0.961	0.958	0.962	**0.963**
DUT-OMRON(5168)	MAE ↓	0.065	0.062	0.056	0.057	0.052	0.054	**0.051**	0.052
	AF_β ↑	0.711	0.724	0.739	0.744	0.771	0.757	0.777	**0.783**
	WF_β ↑	0.669	0.689	0.724	0.717	0.754	0.737	0.762	**0.765**
	E_m ↑	0.852	0.859	0.869	0.867	0.882	0.873	0.885	**0.886**
DUTS-TE(5019)	MAE ↓	0.051	0.047	0.040	0.044	0.036	0.037	**0.036**	**0.036**
	AF_β ↑	0.769	0.781	0.811	0.793	0.845	0.843	0.850	**0.859**
	WF_β ↑	0.748	0.775	0.811	0.794	0.837	0.834	0.824	**0.846**
	E_m ↑	0.889	0.895	0.913	0.897	0.928	0.926	**0.928**	0.927

4.5.1. Effectiveness of RCIE Module

The RCIE module was designed to capture multi-scale context information, as shown in columns 1, 2, and 4 of Table 2, and can significantly enhance the initial feature performance. In particular, we compared the fourth column (ResNet-50 + RCIE + FPN) against the first column (ResNet-50 + FPN) and the second column (ResNet-50 + RFB + FPN). Our results show that our method's MAE score decreased by (19.57% and 8.70%), (5.41% and 1.35%), (20.5% and 10.26%), (14.04% and 8.77%), and (15.91% and 6.82%) on the ECSSD, PASCAL-S, HKU-IS, DUT-OMRON, and DUTS-TR datasets, respectively. Similarly, we observed significant improvements in performance on WF_β, AF_β, and E_m evaluation metrics, The results show that using the RCIE module helps to reduce the computational cost without sacrificing performance, along with indicating the effectiveness of the RCIE module in extracting contextual information and helping to detect salient objects.

4.5.2. Effectiveness of DDA Module

The DDA module adopts two interaction modes and densely connects multi-level features to generate high-quality feature representations. In order to confirm the efficacy of the DDA module, we compared the third column (ResNet-50 + DDA) with the first column (ResNet-50 + FPN). Our results show that our method's MAE score decreased by 37.5%, 16.42%, 46.88%, 16.07%, and 27.5% on the ECSSD, PASCAL-S, HKU-IS, DUT-OMRON, and DUTS-TR datasets, respectively, along with improvements on the other three evaluation metrics. Furthermore, when comparing the fourth column (ResNet-50 + RCIE + FPN) with the eighth column (ResNet-50 + RCIE + DDA) in terms of MAE, we observed that our method's MAE score decreased by 43.75%, 15.63%, 44.44%, 9.62%, and 22.22% on the ECSSD, PASCAL-S, HKU-IS, DUT-OMRON, and DUTS-TR datasets, respectively, along

with an improvement on the other three evaluation metrics. In order to confirm the efficacy of the DDA module, internal structure designs were experimentally tested, including "+DDA (add)", "+DDA (mul)", and "+DDA (cat)", where add, mul, and cat correspond to element-wise addition, element-wise multiplication, and concatenation, respectively. Their performance is shown in columns 5, 6, and 7 of Table 2. It can be seen from the table that the interaction modes in the DDA module outperform a single interaction, further confirming the effectiveness of the DDA module.

5. Discussion

We proposed a multi-scale feature dual-stream aggregation network for salient object detection. Our method extracts multi-scale contextual information using dilated convolutions, asymmetric convolutions, and residual connections. We also used dense connections and the dual-stream aggregation of multi-level features in two modes, achieving excellent performance in terms of both accuracy and robustness.

Compared to other SOD methods, our proposed MFDANet addresses the problem of information loss caused by a single interaction mode, which can limit the model's accuracy and robustness. Additionally, MFDANet achieves competitive performance in terms of parameter number (Param(M)) and model memory (M), as shown in Table 3. While our method does not have the fastest inference speed, we plan to explore lightweight network architectures in future work to improve inference speed while maintaining a high accuracy.

Table 3. Our proposed MFDANet method is compared with other SOD methods in terms of three metrics: parameter number (Param(M)), inference speed (FPS), and model memory (M).

Method	Input Size	Param (M)	Inference Speed (FPS)	Model Memory (M)
EGNet	384 × 384	108.07	9	437
PoolNet	384 × 384	68.26	17	410
AFNet	224 × 224	37.11	26	128
GataNet	384 × 384	128.63	30	503
DSRNet	400 × 400	75.29	15	290
Ours	320 × 320	32.51	25	123

6. Conclusions

We proposed a salient object detection method based on the multi-scale feature dual-stream aggregation network. Firstly, we used different filling rates of dilated convolution, asymmetric convolution, and residual connection to fully extract multi-scale contextual information and alleviate problems such as checkerboard artifacts and local feature loss caused by different filling rates of dilated convolution. Secondly, we used dense connections and dual-stream aggregation with two interaction techniques to generate high-quality saliency maps. This approach solves the information loss problem caused by a single interaction method, improves the model's generalization ability, and reduces computational complexity. Our proposed method outperforms 15 state-of-the-art methods in terms of accuracy and robustness, as demonstrated in both qualitative and quantitative evaluations.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, and visualization, J.P.; writing—review and editing, B.G. and C.X.; supervision, T.W.; project administration and funding acquisition, B.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (2020YFB1314103); the Natural Science Research Project of Colleges and Universities in Anhui Province (KJ2020A0299); National Natural Science Foundation of China (62102003); Anhui Postdoctoral Science Foundation (2022B623); and the Natural Science Foundation of Anhui Province (2108085QF258).

Data Availability Statement: All datasets utilized in this article are open source and publicly available for researchers to use. Interested individuals can obtain the datasets using the following link: http://mmcheng.net/socbenchmark/ (accessed on 28 April 2023).

Acknowledgments: Special thanks to reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rutishauser, U.; Walther, D.; Koch, C.; Perona, P. Is Bottom-Up Attention Useful for Object Recognition? In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), with CD-ROM, Washington, DC, USA, 27 June–2 July 2004; IEEE Computer Society: Los Alamitos, CA, USA, 2004; pp. 37–44. [CrossRef]
2. Wang, W.; Shen, J.; Sun, H.; Shao, L. Video Co-Saliency Guided Co-Segmentation. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 1727–1736. [CrossRef]
3. Craye, C.; Filliat, D.; Goudou, J. Environment exploration for object-based visual saliency learning. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, 16–21 May 2016; Kragic, D., Bicchi, A., Luca, A.D., Eds.; IEEE: Toulouse, France, 2016; pp. 2303–2309. [CrossRef]
4. He, J.; Feng, J.; Liu, X.; Cheng, T.; Lin, T.; Chung, H.; Chang, S. Mobile product search with Bag of Hash Bits and boundary reranking. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE Computer Society: Los Alamitos, CA, USA, 2012; pp. 3005–3012. [CrossRef]
5. Cheng, M.; Mitra, N.J.; Huang, X.; Torr, P.H.S.; Hu, S. Global Contrast Based Salient Region Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 569–582. [CrossRef] [PubMed]
6. Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; Yang, M. Saliency Detection via Graph-Based Manifold Ranking. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; IEEE Computer Society: Los Alamitos, CA, USA, 2013; pp. 3166–3173. [CrossRef]
7. Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; Li, S. Salient Object Detection: A Discriminative Regional Feature Integration Approach. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; IEEE Computer Society: Los Alamitos, CA, USA, 2013; pp. 2083–2090. [CrossRef]
8. Han, J.; Zhang, D.; Hu, X.; Guo, L.; Ren, J.; Wu, F. Background Prior-Based Salient Object Detection via Deep Reconstruction Residual. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 1309–1321. [CrossRef]
9. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; IEEE Computer Society: Los Alamitos, CA, USA, 2015; pp. 3431–3440. [CrossRef]
10. Wang, T.; Zhang, L.; Wang, S.; Lu, H.; Yang, G.; Ruan, X.; Borji, A. Detect Globally, Refine Locally: A Novel Approach to Saliency Detection. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; Computer Vision Foundation/IEEE Computer Society: Washington, DC, USA, 2018; pp. 3127–3135. [CrossRef]
11. Liu, S.; Huang, D.; Wang, Y. Receptive Field Block Net for Accurate and Fast Object Detection. In *Lecture Notes in Computer Science, Proceedings of the Computer Vision—ECCV 2018—15th European Conference, Munich, Germany, 8–14 September 2018*; Proceedings, Part XI; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: New York, NY, USA, 2018; Volume 11215, pp. 404–419. [CrossRef]
12. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaïane, O.R.; Jägersand, M. U^2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
13. Zhao, X.; Pang, Y.; Zhang, L.; Lu, H.; Zhang, L. Suppress and Balance: A Simple Gated Network for Salient Object Detection. In *Lecture Notes in Computer Science, Proceedings of the Computer Vision—ECCV 2020—16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part II; Vedaldi, A., Bischof, H., Brox, T., Frahm, J., Eds.; Springer: New York, NY, USA, 2020, Volume 12347, pp. 35–51. [CrossRef]
14. Wu, Z.; Su, L.; Huang, Q. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; Computer Vision Foundation/IEEE: Washington, DC, USA, 2019; pp. 3907–3916. [CrossRef]
15. Liu, N.; Han, J.; Yang, M. PiCANet: Learning Pixel-Wise Contextual Attention for Saliency Detection. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; Computer Vision Foundation/IEEE Computer Society: Washington, DC, USA, 2018; pp. 3089–3098. [CrossRef]
16. Wei, J.; Wang, S.; Huang, Q. F^3Net: Fusion, Feedback and Focus for Salient Object Detection. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, 7–12 February 2020; AAAI Press: Menlo Park, CA, USA, 2020; pp. 12321–12328.
17. Zhao, J.; Liu, J.; Fan, D.; Cao, Y.; Yang, J.; Cheng, M. EGNet: Edge Guidance Network for Salient Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Toulouse, France, 2019; pp. 8778–8787. [CrossRef]

18. Wang, L.; Lu, H.; Ruan, X.; Yang, M. Deep networks for saliency detection via local estimation and global search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; IEEE Computer Society: Los Alamitos, CA, USA, 2015; pp. 3183–3192. [CrossRef]
19. Li, J.; Pan, Z.; Liu, Q.; Wang, Z. Stacked U-Shape Network with Channel-Wise Attention for Salient Object Detection. *IEEE Trans. Multim.* **2021**, *23*, 1397–1409. [CrossRef]
20. Xu, B.; Liang, H.; Liang, R.; Chen, P. Locate Globally, Segment Locally: A Progressive Architecture with Knowledge Review Network for Salient Object Detection. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, 2–9 February 2021; AAAI Press: Menlo Park, CA, USA, 2021; pp. 3004–3012.
21. Liu, Y.; Zhang, Q.; Zhang, D.; Han, J. Employing Deep Part-Object Relationships for Salient Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Toulouse, France, 2019; pp. 1232–1241. [CrossRef]
22. Liu, J.; Hou, Q.; Cheng, M.; Feng, J.; Jiang, J. A Simple Pooling-Based Design for Real-Time Salient Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019. Computer Vision Foundation/IEEE: Washington, DC, USA, 2019; pp. 3917–3926. [CrossRef]
23. Zhang, L.; Dai, J.; Lu, H.; He, Y.; Wang, G. A Bi-Directional Message Passing Model for Salient Object Detection. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; Computer Vision Foundation/IEEE Computer Society: Los Alamitos, CA, USA, 2018; pp. 1741–1750. [CrossRef]
24. de Boer, P.; Kroese, D.P.; Mannor, S.; Rubinstein, R.Y. A Tutorial on the Cross-Entropy Method. *Ann. Oper. Res.* **2005**, *134*, 19–67. [CrossRef]
25. Máttyus, G.; Luo, W.; Urtasun, R. DeepRoadMapper: Extracting Road Topology from Aerial Images. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 3458–3466. [CrossRef]
26. Yan, Q.; Xu, L.; Shi, J.; Jia, J. Hierarchical Saliency Detection. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; IEEE Computer Society: Washington, DC, USA, 2013; pp. 1155–1162. [CrossRef]
27. Li, Y.; Hou, X.; Koch, C.; Rehg, J.M.; Yuille, A.L. The Secrets of Salient Object Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, 23–28 June 2014; IEEE Computer Society: Washington, DC, USA, 2014; pp. 280–287. [CrossRef]
28. Li, G.; Yu, Y. Visual saliency based on multiscale deep features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 5455–5463. [CrossRef]
29. Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; Ruan, X. Learning to Detect Salient Objects with Image-Level Supervision. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 3796–3805. [CrossRef]
30. Feng, M.; Lu, H.; Ding, E. Attentive Feedback Network for Boundary-Aware Salient Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; Computer Vision Foundation/IEEE: Washington, DC, USA, 2019; pp. 1623–1632. [CrossRef]
31. Mohammadi, S.; Noori, M.; Bahri, A.; Majelan, S.G.; Havaei, M. CAGNet: Content-Aware Guidance for Salient Object Detection. *Pattern Recognit.* **2020**, *103*, 107303. [CrossRef]
32. Zhou, H.; Xie, X.; Lai, J.; Chen, Z.; Yang, L. Interactive Two-Stream Decoder for Accurate and Fast Saliency Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; Computer Vision Foundation/IEEE: Washington, DC, USA, 2020; pp. 9138–9147. [CrossRef]
33. Chen, Z.; Xu, Q.; Cong, R.; Huang, Q. Global Context-Aware Progressive Aggregation Network for Salient Object Detection. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, 7–12 February 2020; AAAI Press: Menlo Park, CA, USA, 2020; pp. 10599–10606.
34. Wang, L.; Chen, R.; Zhu, L.; Xie, H.; Li, X. Deep Sub-Region Network for Salient Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 728–741. [CrossRef]
35. Zhuge, M.; Fan, D.; Liu, N.; Zhang, D.; Xu, D.; Shao, L. Salient Object Detection via Integrity Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3738–3752. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.