sensors

MDPI

Article

Multiscale Cascaded Attention Network for Saliency Detection Based on ResNet

Muwei Jian [1,2,*], Haodong Jin [1,†], Xiangyu Liu [1,†] and Linsong Zhang [1]

[1] School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China
[2] School of Information Science and Technology, Linyi University, Linyi 276012, China
* Correspondence: jianmuwei@ouc.edu.cn
† These authors contributed equally to this work.

Abstract: Saliency detection is a key research topic in the field of computer vision. Humans can be accurately and quickly mesmerized by an area of interest in complex and changing scenes through the visual perception area of the brain. Although existing saliency-detection methods can achieve competent performance, they have deficiencies such as unclear margins of salient objects and the interference of background information on the saliency map. In this study, to improve the defects during saliency detection, a multiscale cascaded attention network was designed based on ResNet34. Different from the typical U-shaped encoding–decoding architecture, we devised a contextual feature extraction module to enhance the advanced semantic feature extraction. Specifically, a multiscale cascade block (MCB) and a lightweight channel attention (CA) module were added between the encoding and decoding networks for optimization. To address the blur edge issue, which is neglected by many previous approaches, we adopted the edge thinning module to carry out a deeper edge-thinning process on the output layer image. The experimental results illustrate that this method can achieve competitive saliency-detection performance, and the accuracy and recall rate are improved compared with those of other representative methods.

Keywords: ResNet; multiscale cascade extraction module; attention module; saliency detection

Citation: Jian, M.; Jin, H.; Liu, X.; Zhang, L. Multiscale Cascaded Attention Network for Saliency Detection Based on ResNet. *Sensors* **2022**, *22*, 9950. https://doi.org/10.3390/s22249950

Academic Editors: Bin Fan and Wenqi Ren

Received: 20 November 2022
Accepted: 14 December 2022
Published: 16 December 2022

1. Introduction

The task of visual saliency detection was created to allow computer systems to mimic the capabilities of the human visual system (HVS) for quickly extracting salient objects from a scene. These saliency regions in an image/video usually contain the object of interest to the observer and those areas that can gain HVS attention in real life. With the in-depth study of convolutional neural networks, saliency detection has been widely applied as an effective technique for preprocessing numerous content-based tasks in computer vision, such as image recognition, image segmentation, image retrieval, and pedestrian/object detection [1–3].

Early vision work was classified based on viewpoint acquisition mechanisms into cognitive [4–8], Bayesian [9], spectral analysis [10], information-theoretic [11], graphical [12], decision-theoretic [13], and pattern classification models [14]. With the progress of saliency detection, image-oriented detection methods have formed more complete detection systems, which can be separated into two main groups. One is the task-driven top-down detection methods, which often require a training process of task-dependent and specific prior knowledge; the other one is data-driven and subconscious, bottom-up detection methods, which mainly use underlying visual cues such as color, contrast, and shape for saliency object. In addition, in pace with the advancement and development of imaging devices, depth information acquisition is becoming easier and more convenient to manipulate, which has created the groundwork for the rise and progress of RGBD image saliency-detection algorithms [15]. Compared with research on traditional 2D image saliency

detection [16,17], the research on RGBD image saliency-detection algorithm started late and has achieved certain satisfactory results. However, researchers have not reached a consensus on the mechanism through which the effect of depth information on the human perceptual system is achieved and how to effectively explore depth information; thus, further in-depth research is still needed.

Even though many of saliency-detection methods have achieved notable results, they are still not satisfactory in removing background interference, maintaining unabridged edges, and other slight details. To address the shortcomings of these conventional methods, we developed an image saliency-detection network using the classical convolutional neural network model as the basic framework, and designed an efficient saliency-detection model based on a multiscale cascaded attention network. In summary, the main contributions of this study are characterized as follows:

(1) We employ a multiscale cascade block and a lightweight channel attention module between the typical encoding–decoding networks for optimizing the performance of image saliency detection based on ResNet34.
(2) A multiscale cascaded attention model is devised to rationally use the multiscale extraction module for high-level semantic features of the image, while the attention module is used for the joint refinement of low- and high-level semantic features to enhance the precision of saliency detection.
(3) To solve the problem of blurred edges that has been neglected in many existing methods, we applied the edge refinement module to the output layer image for clear edge refinement.

The remainder of this paper is structured as follows: We first describe the present status of the associated work in Section 2. The designed network architecture and loss function are outlined in Section 3. Additionally, Section 4 provides the outcomes of our experiments. Finally, Section 5 presents our conclusions and discussion.

2. Related Work

2.1. Traditional Saliency Detection Methods

Traditional saliency-detection methods can be coarsely separated into spatial-domain-based and transform domain-based modeling frameworks. Spatial-domain-based detection approaches are usually studied based on image processing theory, with the output of saliency detection being results generated by low-level cues (contrast, chrominance, luminance, texture, etc.). These methods usually perform pixel-level saliency region extraction by calculating the difference between the pixels in the salient region and the surrounding background pixels; thus, they depend on the size of the selected window and the threshold value for saliency discrimination. A typical strategy based on low-level features is to extract the salient regions by optimizing thresholds. The AC algorithm [18] is a numerically computed saliency mapping generation algorithm, where the local comparison between the input image's area R1 and its neighborhood R2 at various scales determines the saliency value. Later, with the increases in data volume and the accuracy requirement on extracted images, the optimized thresholding approach was replaced by other advanced methods due to its limitation of not being applicable in many images with complex textures. Cheng et al. [19] created different approaches combined with global contrast, named histogram-based contrast (HC) and regional contrast (RC) salient object models.

In recent years, spatial-domain-based methods are more often implemented based on image component analysis strategy, the core tenet of which is analyzing the principal component and independent component, and using other spatial variation methods to explore the correlation between image foreground and background pixels to achieve salient region extraction. For instance, Goferman et al. [20] devised a saliency method focused on context, which can detect the salient regions in representative scenes rather than just salient objects. Additionally, in the spatial domain, the graph-theory-based saliency-detection model usually splits the inputs into diverse blocks and regards each of them as nodes. Then, weighted edges between blocks of pixels, depending on visual characteristics such as

color, luminance, and orientation, are integrated to determine the graphical mapping [21]. Harel et al. [12] designed a method named graph-theory-based algorithm, which simulates the visualization principle in the feature extraction process. Specifically, in the stage of generating the saliency map, Markov chains are introduced, and the central surround difference is calculated with a graph model. The saliency map is then obtained by a purely mathematical calculation.

The spatial-domain-based methods can achieve satisfactory results on images with certain differences between the foreground and background, but the results are not ideal for many images without significant differences in the spatial domain. In order to tackle the limitations of these spatial domain methods, many transform-domain methods based on Fourier transform and wavelet/Gabor transform have been exploited in the field of saliency detection. Transform-domain-based methods generally include wavelet transform [22], wavelet frame transform [23], curvelet transform, projection transform, etc. Although transform operators such as Fourier transform can more accurately describe the global and macroscopic features, the results of this method are not acceptable for local or unsmooth information. Hou and Zhang [10] developed a spectral residual algorithm, which is a typical task in the field of frequency-domain-based saliency detection. In detail, this strategy considers the possibility of distributing an image's substantial content as salient and redundant information. The log spectrum distribution exhibits a consistent trend for various data, and the curve complies with the local linearity requirement.

Based on previous studies, Guo et al. [24] exploited a novel method, the phase spectrum of quaternion Fourier transform (PQFT), which abandons the magnitude spectrum and only utilizes the input image's phase spectrum following Fourier transformation. Saliency mapping similar to that of the SR method is obtained by Fourier inverse transform. The Fourier-transformed phase spectrum is expanded. After that, Achanta et al. [25] furthered the FT algorithm and devised the maximum symmetric surround method for saliency detection. This method varies the center surrounding bandwidth according to the separation between a pixel's point and the edge of an image. Thus, the algorithm uses the average of the most likely symmetric neighboring areas rather than calculating the average of the global feature vectors generated by the FT method. Although traditional techniques have taken some steps in the domain of saliency detection, they still cannot adapt to the numerous high-complexity and low-quality data.

2.2. Deep-Learning-Based Saliency-Detection Methods

As it is difficult to improve the effect of traditional saliency detection methods, deep learning-based approaches have received the attention of scholars in recent years. Since the earliest BP networks [26], a group of saliency detection frameworks have appeared in the field of machine learning. After the accelerated advancement of neural networks [27–30], more models based deep learning have emerged. Since 2015, saliency detection has been processed by convolutional neural networks (CNNs). Unlike the traditional techniques based on comparison of visual cues, CNN-based methods effectively reduce the need to design manual features and greatly improve the computing efficiency, so these methods have been extensively used by many scientific scholars [31–33].

CNN-based models typically contain many neurons with adjustable parameters and variable structural field sizes. The neurons have a large receptive range to provide global information, which makes it possible to identify the regions of salient objects in the scenario more effectively. Compared with traditional methods, CNN and its optimization methods have become the most mainstream methods at the current research stage due to their excellent extraction accuracy and computational efficiency. Wang et al. [34] suggested a visual attention module based on global saliency feature information for visual saliency-detection networks, which focuses on both superficial refined layers with locally salient responses and deep coarse layers with globally salient information. Cornia et al. [35] designed an architecture that incorporates neural attention mechanisms to generate saliency maps. Zhu et al. [36] designed a multiscale adversarial feature learning (MAFL) model for

saliency detection. Recently, Wei et al. [37] introduced a deep saliency-detection framework using full convolutional networks (FCNs) to solve the cosalient object discovery and detection problem. He et al. [38] devised a new superpixel-based framework called SuperCNN, which can better extract the interior representations of saliency and hierarchical contrast features independent of the region size by using a multiscale network structure. Later, Hou et al. [39] designed a new saliency-detection method stemming from holistically nested edge detection (HED) by adding a skip layer structure, where high-level features guide low-level features, thus forming an efficient end-to-end salient-object-detection method. Hui et al. [40] exploited a multiguided saliency-detection model using the intrinsic relationship between different features. To further improve the performance and robustness, a novel pixel-by-pixel contrast loss function was developed and integrated with the cross-entropy loss function to jointly supervise the training process. Recently, as a key advance in deep learning, a transformer-based network is applied to salient object detection. Liu et al. [41] introduced a pure transformer into saliency detection to make a convolution-free model called visual saliency transformer (VST). For better extract low- and high-level information, Hussain et al. [42] designed a parallel architecture to integrate both transformer and CNN features, which are fed into a pyramidal attention module.

3. Proposed Method

Regarding the problems commonly experienced in the current research, in this section, we propose a multiscale cascaded attention network for salient object detection. Firstly, the image data et is preprocessed in real time; i.e., after the object is locked, grayscale images matching the region are generated and fused, so as to effectively eliminate the background noises and improve the accuracy of the salient region detection. The devised model is able to reduce the interference of redundant objects by accordingly processing the multiobject images. Then, the preprocessed images are put into the multiscale cascaded attention network for saliency detection. The extraction part of the consists of an encoder, extraction block, and composition, in which the extraction of low-level features by the encoder and the extraction of high-level semantic features by the multiscale cascade-attention module are jointly utilized to enhance the performance of saliency detection for the whole and detailed parts. Finally, the extracted information is integrated using the decoder network to obtain the final saliency-detection map. We compared our method with nine advanced methods on three public datasets (DUTS [43], ECSSD [44], and HKU-IS [45]), and the experiments demonstrated that this method is advantageous in terms of overall metrics and visual details. The general structure of the network is shown in Figure 1.

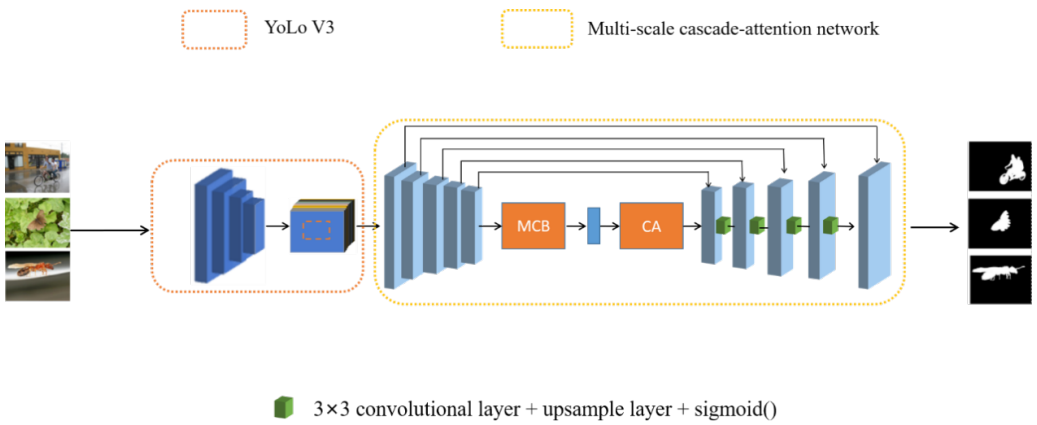

Figure 1. Overall architecture of the proposed network.

3.1. Network Architecture

The overall framework of the proposed method is presented in Figure 1. The input image is preprocessed with the YoLoV3 network [46] to eliminate most of the interference from the background, the processed image is fed into the U-shaped backbone network

for feature extraction and processing, and then the extracted feature map is upsampled to generate the final saliency image. The following subsections provide a detailed description of the entire network.

3.1.1. Object Locking and Extraction from Images

In this subsection, the preprocessing stage to highlight the salient object as well as to remove the background interference is systematically introduced. During image preprocessing, the initial image is directly input to the preprocessing module, and then the processed intermediate image is used as the input of the saliency extraction network for subsequent processing. With regard to object tracking, Redmon et al. [46] improved the YoLo network and designed a model that assigns only one bounding box to each object. Compared with C-RNN, this network significantly improves the processing speed of the YoLoV3 network because the coordinates of the bounding box are directly predicted and localized using the convolutional extraction of the features, followed by the fully connected layer. The YoLoV3 network is a further improvement of the YoLo series of algorithms, which retains the advantages of the previous algorithms while improving the accuracy. YoLoV3 has strengthened performance and increased speed, so has become one of the preferred detection algorithms in the engineering community due to its powerful real-time performance and concise network structure. In applications, the coordinates of the center of the selected object can be obtained through the YoLoV3 network, and the generated center coordinates are used to produce a matching grayscale image, which is then fused with the original input image. For the natural images in the test set, the relative position and area information of the detected objects are accurately acquired during the preprocessing stage. After inputting an image to the module, it is first sent to the preprocessing module, and the specific formula for determining the bounding box is formulated as follows:

$$b_x = \sigma(t_x) + C_x, \tag{1}$$

$$b_y = \sigma(t_y) + C_y, \tag{2}$$

$$b_w = P_w e^{t_w}, \tag{3}$$

$$b_h = P_h e^{t_h}, \tag{4}$$

where t_x, t_y, t_w, and t_h denote the four coordinates used to predict the object bounding box in the YoLo network; C_x and C_y represent the horizontal and vertical offsets of the network where the center of the object is located and the coordinates of the upper left corner of the object image, respectively; and P_w and P_h indicate the width and height of the corresponding bounding box, respectively.

After we obtain the parameters of the object bounding box, we generate the corresponding grayscale image based on the extracted coordinate information of the object and merge the grayscale image with the original image to filter out the purposeless background information. Regarding the visual perception mechanism mentioned above, when the observer extracts the object of interest in the scene, if the observer closely watches the object, the gaze distance decreases, and the field of view becomes smaller, but the object will be clearer. On the contrary, as the gaze distance increases, the observer's field of view gradually increases, but the clarity of the salient object becomes increasingly blurred. At the same time, if the gaze distance remains the same, the objects around the region of interest decrease in sharpness as the distance from the central object increases. Based on this principle, we generate a grayscale image matching the coordinate of the object frame, and add the α channel, which denotes transparency. The value of α is calculated by the Gaussian function, as shown in the following calculation formula:

$$\alpha(x_i,\ y_i) = e^{\frac{\sqrt{(x_c - x_i)^2 + (y_c - y_i)^2}}{s \times s}}, \tag{5}$$

where $(x_c,\ y_c)$ are the horizontal and vertical coordinates of the center of the object in the bounding box, respectively; $(x_i,\ y_i)$ represent the position of the corresponding pixel in the image; and the value of s is dynamically set according to the width and height of the bounding box. Thereafter, α can be further estimated. When the pixel in the generated grayscale image lies within the object bounding box $Area_b$, the corresponding value is the actual value of the pixel, and the value of α for that point is constant. When the pixel of the grayscale image is not located in the boxed area, the value is set to 255. The final preprocessed image resulting from the fusion of the grayscale image with the input image is expressed with the following equation:

$$\alpha(x_i,\ y_i) = \begin{cases} \alpha, & (x_i,\ y_i) \in Area_b \\ 255, & others \end{cases}, \tag{6}$$

Some samples of the preprocessed images are displayed in Figure 2. Occasionally, multiple objects may occur during the process. In this situation, we detect the horizontal and vertical coordinates of each coordinate point of all objects in an image, and select the largest and smallest horizontal and vertical coordinates to determine the two coordinate points. These two coordinate points form a bounding box, which can also effectively remove other distracting backgrounds.

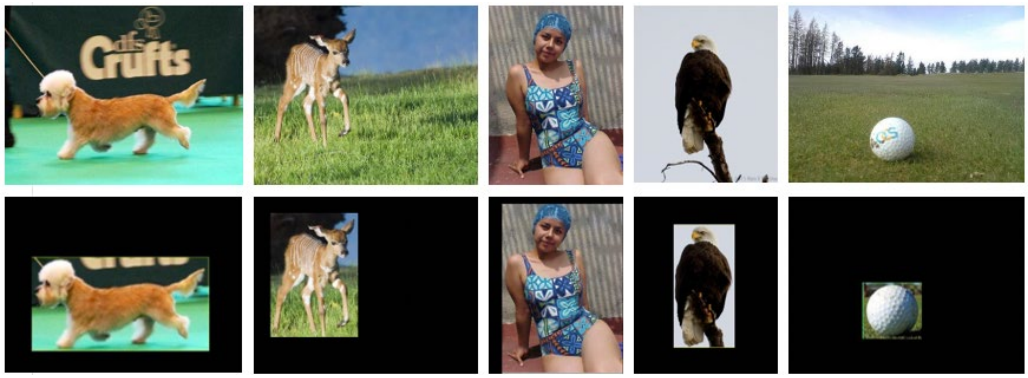

Figure 2. Some sample images with salient objects after preprocessing module.

3.1.2. Feature Encoder Module

The core framework of the network devised for saliency detection in this study includes three main modules: the feature encoder module, the contextual feature extraction module, and the feature decoder module. In this subsection, the feature encoder is introduced. As shown in Figure 3, based on the design of U-Net, the encoder is replaced with a pretrained ResNet34. The designed structure preserves the feature extraction modules and the average pooling layer, but discards the final fully connected layer of the origin network. For better expression, the dimensional transformation of the feature map is clearly represented.

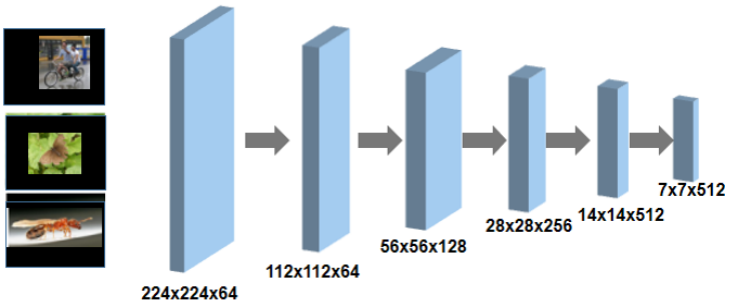

Figure 3. The designed architecture of the feature encoder module.

3.1.3. Contextual Feature Extraction Module

This subsection focuses on the second part of the network structure—the contextual feature extraction module. This module consists of two distinct parts. One is a multiscale cascade block to perform multiscale feature extraction, while the other part is a lightweight channel attention module to perform feature refinement. This entire module is intended to enhance the semantic information of the context and allows the generation of higher-level feature maps.

(1) Dilated Convolution

Deep convolutional layers have been shown to be an efficient for generating visual feature representations for tasks such as semantic segmentation and object detection. However, the pooling layer may cause the original image's semantic information to be lost. In order to overcome this limitation, we employ dilated convolution for this process to enhance the efficiency of computation, and this operation is formulated as below:

$$y(i) = \sum_k x(i + rk)w(k), \tag{7}$$

where x represents the input feature map, y expresses the output feature map, w indicates the filter, and r denotes the dilation rate when sampling the image. Typically, the standard convolution is a special case of $r = 1$. In contrast, in the multiscale extraction module, the dilated convolution allows us to change this rate to adaptively modify the receptive field of the filter. This process is illustrated in Figure 4.

(2) Multiscale Cascade Block (MCB)

Both the inception structure and ResNet network are typical and representative frameworks based on deep learning. The inception structure widens the network architecture using various receptive fields, whereas ResNet uses a skip connection method to prevent gradients from explosion and disappearance. Therefore, the multiscale cascade block applies the inception structure to splice with the decoder's ResNet network as a way of inheriting the advantages of both.

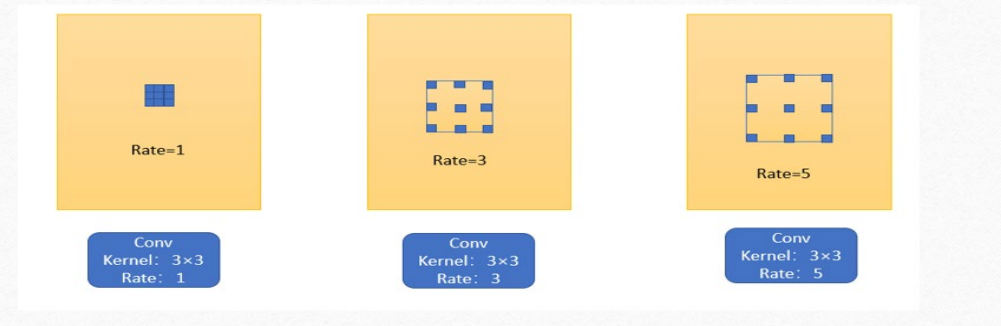

Figure 4. Schematic diagram of dilated convolution.

As shown in Figure 5, the proposed multiscale cascade block has four cascade branches. The convolutions with various dilation rates are sequentially added from top to bottom: each branch contains 3, 7, 9, and 19 perceptual fields, respectively. We use a 1×1 convolution for the activation of every branch. Afterward, the original features and other multiscale features are summed. In this module, the convolution with a larger receptive field is extracted for larger objects and generates more contracted features. The convolution of smaller receptive fields achieves better results for the extraction of small objects. Additionally, through combining the convolution with various dilation rates, the multiscale cascade block can simultaneously extract the salient features of diverse objects of different sizes.

(3) Channel Attention (CA) Module

To further strengthen the accuracy of the saliency detection results, we also added a lightweight channel attention module for saliency detection optimization through assigning greater weights to the high response channels.

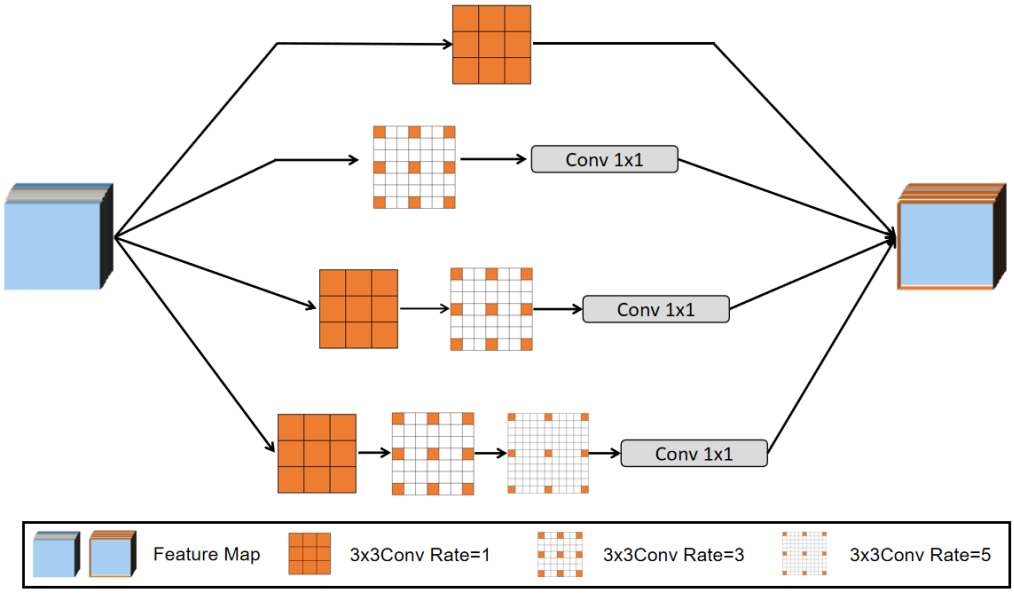

Figure 5. The devised architecture of multiscale cascade block.

3.1.4. Feature Decoder Module

In this study, we exploited the feature decoder module to recover the high-level semantic features obtained by the feature encoder and the contextual feature extraction modules. Our decoding network is almost symmetrical with the first half of the encoder. Each stage in the decoder includes three convolutional layers with normalization and ReLU activation functions. The input of each stage is a feature mapping of the connection between its previous stage and its upsampled output of the corresponding stage in the encoder, and then the multichannel output of each decoder is fed into a 3×3 convolutional layer, followed by a bilinear upsampling and a sigmoid function. Moreover, there is a separate supervision in each stage as a means of achieving intensive deep supervision during the training process and increasing the accuracy of the saliency-detection results. The structure of the decoder is illustrated in Figure 6.

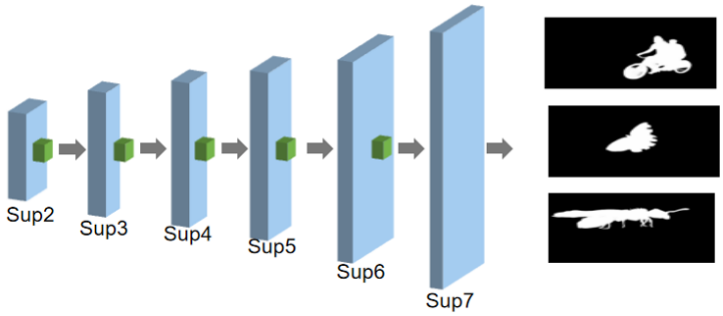

Figure 6. The devised architecture of the feature decoder module.

3.2. Loss Function

During the training process of the framework described above, the computed saliency map with the labeled dataset is learned by the loss estimation between them [35,47]. In binary classification studies, the binary cross-entropy (BCE) loss function has been

frequently used. It also an objective function typically used to measure the difference between the predicted saliency maps and ground truth in saliency detection tasks, which has widely achieved good results. Therefore, a BCE loss function was employed in this study. The expression of binary cross-entropy loss is formulated as:

$$L(\theta, w) = l_{predict}(\theta, w_{predict}), \tag{8}$$

where θ represents the set of all network parameters, w indicates the weight of the corresponding layer, and l is the binary cross-entropy loss function, which can be employed to equalize the generated saliency value $Y \in (0,1)^N$ and its corresponding labeled image $G \in (0,1)^N$ as follows:

$$L = -\sum_{i=1}^{N} \{(1-a) g_i \log y_i + a(1-g_i) \log(1-y_i)\} \tag{9}$$

where $N = H \times W$ represents the image size, $g_i \in G$ and $y_i \in Y$, accordingly.

4. Experiments

4.1. Implementation Details

In this study, the deep-learning framework for the experiments was built on Pytorch, and other detailed environment configurations are indicated in Table 1.

Table 1. Configuration details of the experimental implementation.

Experimental Implementation	Configuration
Operating System	Win10
Python	3.7
Pytorch	1.5.0
CUDA	9.0
GPU	NVIDIA-GTX1080ti

4.2. Qualitative Analysis

In this part of the study, visual comparisons of saliency-detection results were evaluated. To validate the effectiveness of the devised saliency-detection network, we selected nine existing saliency-detection methods for contradistinctive experiments, namely Amulet [8], DCL [9], DHS [10], MDF [11], NLDF [12], UCF [13], RAS [14], R3Net [48], and DGRL [49]. Three different datasets, including DUTS [43], ECSSD [44], and HKU-IS [45], were tested. In Figure 7, the first column represents the example image, the second column expresses the ground truth of the highlighted salient objects, and the third column displays the saliency map produced by the proposed method. The subsequent columns are the saliency results for each different method. The probability that each pixel point in the image belongs to the foreground is represented its corresponding pixel value. It can be seen from Figure 7 that our devised model produced saliency results nearer to the ground truth.

In Figure 7, the first three rows are salient object detection comparisons on the DUTS dataset; the fourth and fifth rows are detection results in terms of the ECSSD dataset and the HKU-IS dataset, respectively. It can be distinctly observed that the results obtained by the proposed method in the first and third rows are more accurate than those of the comparative methods in the overall extraction of the deer and human. Moreover, the experimental data in the third row show that our method also obtained better result in the details, such as the tentacles of flying insects and the tiny legs, compared with the other nine approaches used for comparison. In addition, the fifth row reflects that our method was not only effective in extracting single objects, but could also maintain efficient saliency detection for multiple objects. In summary, the proposed method is more effective than the other nine comparative approaches in terms of visual comparison, and our results resembled the ground truth. The results of the comparison indicated that the devised model is able to generate more accurate salient maps.

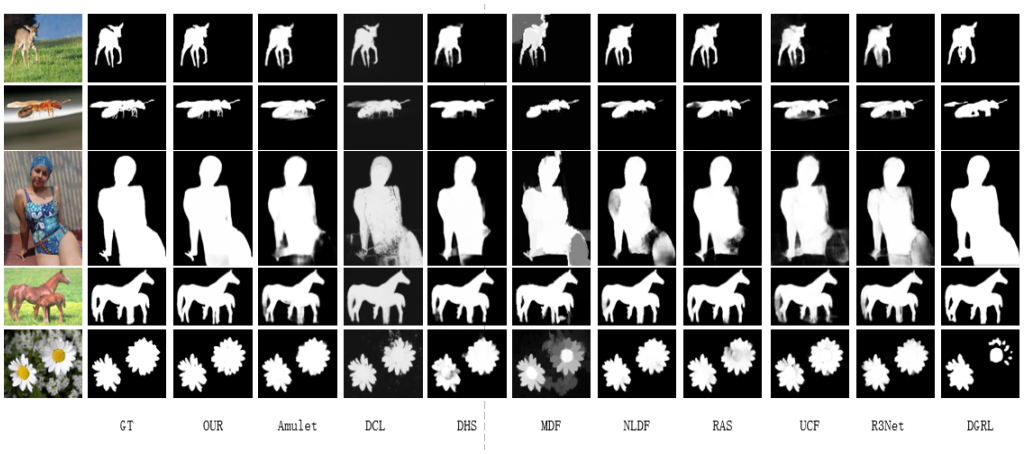

Figure 7. The visual comparison of each model on different datasets.

4.3. Quantitative Analysis

Next, we quantitatively analyzed the experimental results by comparing evaluation metrics. The results of the precision–recall (P-R) curves of each model on three distinct image datasets are presented in Figure 8. Specifically, Figure 8a displays the P-R curves of the individual algorithm on the DUTS dataset, Figure 8b reveals the P-R curves of each model on the ECSSD dataset, and Figure 8c shows the P-R curves of diverse approaches on the HKU-IS dataset. Figure 8 shows that the proposed method has a greatly improved accuracy and recall rate compared with those of the other methods.

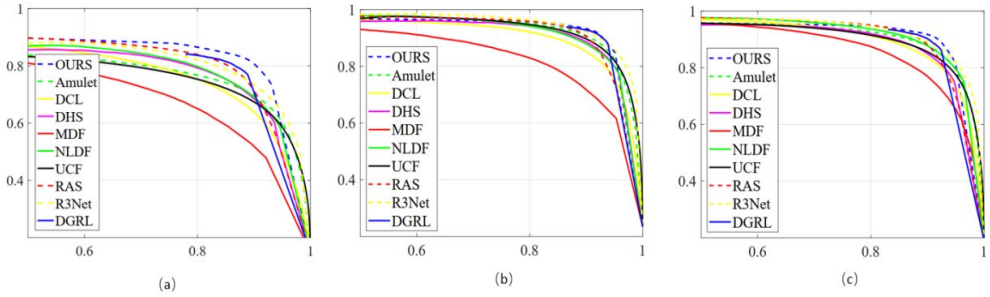

Figure 8. The P-R curves of each method according to different datasets. (**a**) the P-R curves of the individual algorithm on the DUTS dataset; (**b**) the P-R curves of each model on the ECSSD dataset; (**c**) the P-R curves of diverse approaches on the HKU-IS dataset.

Additionally, Table 2 presents the $F^{\max}$ (F measure) evaluation metric, F^{ω} (weighted F-measure score) and mean absolute error (MAE) score of the different methods on the three datasets. Table 2 shows that the proposed method has an improved $F^{\max}$ on all of three datasets compared with those of the other nine methods. Our method was able to achieve $F^{\max}$ values of 0.832, 0.932, and 0.917 and F^{ω} scores of 0.736, 0.865, and 0.844 on DUTS, ECSSD, and HKU-IS, respectively.

MAE can reflect the accuracy of the model in terms of the error rate of the detection results, and Table 2 shows that our method obtains the smallest MAE values among all the methods, which were reduced to 0.052, 0.041, and 0.035 on the three data sets. Although one of the methods used for comparison achieved the same value as our method, this is enough to reflect the devised model is capable of achieving promising results. The results of the many experiments indicated that the designed image saliency-detection framework is feasible and effective.

Table 2. Comparison of the *F*-measure evaluation metric and MAE score on different dataset.

Methods	DUTS			ECSSD			HKU-IS		
	$F^{max}\uparrow$	$F^{\omega}\uparrow$	$MAE\downarrow$	$F^{max}\uparrow$	$F^{\omega}\uparrow$	$MAE\downarrow$	$F^{max}\uparrow$	$F^{\omega}\uparrow$	$MAE\downarrow$
Amulet [8]	0.778	0.657	0.085	0.915	0.841	0.059	0.895	0.813	0.052
DCL [9]	0.782	0.606	0.088	0.890	0.802	0.088	0.885	0.736	0.072
DHS [10]	0.807	0.698	0.067	0.832	0.841	0.059	0.890	0.806	0.053
MDF [11]	0.730	0.509	0.094	0.783	0.605	0.105	0.861	0.726	0.129
NLDF [12]	0.812	0.710	0.066	0.905	0.839	0.063	0.902	0.838	0.045
UCF [13]	0.771	0.588	0.117	0.911	0.789	0.078	0.886	0.751	0.074
RAS [14]	0.831	0.727	0.060	0.920	0.809	0.056	0.913	0.821	0.045
R3Net [48]	0.828	0.715	0.059	0.931	0.832	0.046	0.916	0.837	0.038
DGRL [49]	0.829	0.708	0.050	0.922	0.813	0.041	0.910	0.842	0.036
DSS [16]	0.825	0.732	0.057	0.915	0.858	0.052	0.913	0.836	0.039
PiCANet [50]	0.851	0.748	0.054	0.931	0.863	0.042	0.921	0.847	0.042
CSNet [51]	0.819	0.712	0.074	0.916	0.837	0.066	0.899	0.813	0.059
Ours	0.832	0.736	0.052	0.932	0.865	0.041	0.917	0.844	0.035

5. Conclusions

In this study, a multiscale cascaded-attention framework was developed for saliency detection, which overcomes the shortcomings of existing methods, such as the edges of salient objects being not clear enough and the presence of background interfering with the saliency map. The main network framework of the proposed method was inspired by U-Net, while the ResNet was designed as a U-shaped network for saliency detection optimization. Specifically, a multiscale cascade block (MCB) and a lightweight channel attention (CA) module were jointly added between the encoding and decoding networks for optimization. Eventually, the visual attention mechanism was exploited for feature extraction, and integration was performed to refine the saliency-detection results. The experimental results illustrated that the designed method produces competitive saliency-detection performance and has higher accuracy and recall than other methods. Recently, as a key advance in deep learning, the transformer-based network has also started to be applied to salient object detection. Limited by the experimental equipment, our framework was designed based on the convolutional neural network. Therefore, exploring transformer-based saliency methods is our future research direction.

Author Contributions: M.J.: contributed to the conceptualization, methodology, validation, data analysis, and writing of the paper; H.J. and X.L.: reviewed the work, writing of the paper grammar and spelling checking, and suggested additional experiments. L.Z.: assisted in data acquisition, review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) (6197612 and, 62072213); Taishan Young Scholars Program of Shandong Province; and Key Development Program for Basic Research of Shandong Province (ZR2020ZD44).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Jian, M.; Wang, J.; Yu, H.; Wang, G.; Meng, X.; Yang, L.; Dong, J.; Yin, Y. Visual saliency detection by integrating spatial position prior of object with background cues. *Expert Syst. Appl.* **2021**, *168*, 114219. [CrossRef]
2. Jian, M.W.; Wang, J.; Dong, J.Y.; Cui, C.R.; Nie, X.S.; Yin, Y.L. Saliency detection using multiple low-level priors and a prop-agation mechanism. *Multimed. Tools Appl.* **2020**, *79*, 33465–33482. [CrossRef]

3. Lu, X.; Jian, M.; Wang, X.; Yu, H.; Dong, J.; Lam, K.-M. Visual saliency detection via combining center prior and U-Net. *Multimedia Syst.* **2022**, *28*, 1689–1698. [CrossRef]
4. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
5. Le Meur, O.; Le Callet, P.; Barba, D.; Thoreau, D. A coherent computational approach to model bottom-up visual attention. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 802–817. [CrossRef]
6. Mathe, S.; Sminchisescu, C. Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 842–856. [CrossRef]
7. Mathe, S.; Sminchisescu, C. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 1923–1931.
8. Mathe, S.; Sminchisescu, C. Actions in the Eye: Dynamic Gaze Datasets and Learnt Saliency Models for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1408–1424. [CrossRef]
9. Zhang, L.; Tong, M.H.; Marks, T.K.; Shan, H.; Cottrell, G.W. SUN: A Bayesian framework for saliency using natural statistics. *J. Vis.* **2008**, *8*, 32. [CrossRef]
10. Hou, X.; Zhang, L. Saliency Detection: A Spectral Residual Approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
11. Bruce, N.; Tsotsos, J. Saliency based on information maximization. *Adv. Neural Inf. Process. Syst.* **2005**, *18*, 155–162.
12. Harel, J.; Koch, C.; Perona, P. Graph-based visual saliency. *Adv. Neural Inf. Process. Syst.* **2006**, *19*, 545–552.
13. Gao, D.; Vasconcelos, N. Discriminant saliency for visual recognition from cluttered scenes. *Adv. Neural Inf. Process. Syst.* **2004**, *17*, 481–488.
14. Judd, T.; Ehinger, K.; Durand, F.; Torralba, A. Learning to Predict Where Humans Look. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2010; pp. 2106–2113.
15. Ren, G.; Yu, Y.; Liu, H.; Stathaki, T. Dynamic Knowledge Distillation with Noise Elimination for RGB-D Salient Object Detection. *Sensors* **2022**, *22*, 6188. [CrossRef]
16. Duan, F.; Wu, Y.; Guan, H.; Wu, C. Saliency Detection of Light Field Images by Fusing Focus Degree and GrabCut. *Sensors* **2022**, *22*, 7411. [CrossRef]
17. Yang, J.; Wang, L.; Li, Y. Feature Refine Network for Salient Object Detection. *Sensors* **2022**, *22*, 4490. [CrossRef]
18. Achanta, R.; Estrada, F.; Wils, P.; Süsstrunk, S. Salient Region Detection and Segmentation. In *International Conference on Computer Vision Systems*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 66–75.
19. Cheng, M.M.; Mitra, N.J.; Huang, X.; Torr, P.H.; Hu, S.M. Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 569–582. [CrossRef]
20. Goferman, S.; Zelnik-Manor, L.; Tal, A. Context-aware saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1915–1926. [CrossRef]
21. Aiello, W.; Chung, F.; Lu, L. A Random Graph Model for Massive Graphs. In Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, Portland, OR, USA, 21–23 May 2000; pp. 171–180.
22. Jian, M.; Lam, K.-M.; Dong, J.; Shen, L. Visual-Patch-Attention-Aware Saliency Detection. *IEEE Trans. Cybern.* **2014**, *45*, 1575–1586. [CrossRef]
23. Jian, M.; Zhang, W.; Yu, H.; Cui, C.; Nie, X.; Zhang, H.; Yin, Y. Saliency detection based on directional patches extraction and principal local color contrast. *J. Vis. Commun. Image Represent.* **2018**, *57*, 1–11. [CrossRef]
24. Guo, C.; Ma, Q.; Zhang, L. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
25. Achanta, R.; Susstrunk, S. Saliency Detection Using Maximum Symmetric Surround. In Proceedings of the IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; pp. 2653–2656. [CrossRef]
26. Hecht-Nielsen, R. Theory of the Backpropagation Neural Network. In *Neural Networks for Perception*; Academic Press: Cambridge, MA, USA, 1992; pp. 65–93.
27. Ren, W.; Zhang, J.; Pan, J.; Liu, S.; Ren, J.S.; Du, J.; Cao, X.; Yang, M.-H. Deblurring Dynamic Scenes via Spatially Varying Recurrent Neural Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3974–3987. [CrossRef]
28. Ren, W.; Liu, S.; Zhang, H.; Pan, J.; Cao, X.; Yang, M.H. Single Image Dehazing via Multi-scale Convolutional Neural Networks with Holistic Edges. *Int. J. Comput. Vis.* **2016**, *128*, 240–259. [CrossRef]
29. Fan, B.; Zhou, J.; Feng, W.; Pu, H.; Yang, Y.; Kong, Q.; Wu, F.; Liu, H. Learning Semantic-Aware Local Features for Long Term Visual Localization. *IEEE Trans. Image Process.* **2022**, *31*, 4842–4855. [CrossRef] [PubMed]
30. Fan, B.; Yang, Y.; Feng, W.; Wu, F.; Lu, J.; Liu, H. Seeing through Darkness: Visual Localization at Night via Weakly Supervised Learning of Domain Invariant Features. *IEEE Trans. Multimedia* **2022**, 1. [CrossRef]
31. Luo, A.; Li, X.; Yang, F.; Jiao, Z.; Cheng, H.; Lyu, S. Cascade Graph Neural Networks for RGB-D Salient Object Detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 8–14 September 2020; pp. 346–364. [CrossRef]
32. Feng, M.; Lu, H.; Ding, E. Attentive Feedback Network for Boundary-Aware Salient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1623–1632.

33. Liu, J.J.; Hou, Q.; Cheng, M.M.; Feng, J.; Jiang, J. A simple Pooling-Based Design for Realtime Salient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3917–3926.
34. Wang, W.; Shen, J. Deep visual attention prediction. *IEEE Trans. Image Process.* **2017**, *27*, 2368–2378. [CrossRef] [PubMed]
35. Cornia, M.; Baraldi, L.; Serra, G.; Cucchiara, R. Predicting Human Eye Fixations via an LSTM-Based Saliency Attentive Model. *IEEE Trans. Image Process.* **2018**, *27*, 5142–5154. [CrossRef] [PubMed]
36. Wei, L.; Zhao, S.; Bourahla, O.E.F.; Li, X.; Wu, F.; Zhuang, Y. Deep Group-Wise Fully Convolutional Network for Co-Saliency Detection with Graph Propagation. *IEEE Trans. Image Process.* **2019**, *28*, 5052–5063. [CrossRef]
37. Zhu, D.; Dai, L.; Luo, Y.; Zhang, G.; Shao, X.; Itti, L.; Lu, J. MAFL: Multi-Scale Adversarial Feature Learning for Saliency Detection. In Proceedings of the 2018 International Conference on Control and Computer Vision, New York, NY, USA, 15–18 June 2018; pp. 90–95.
38. He, S.; Lau, R.W.; Liu, W.; Huang, Z.; Yang, Q. SuperCNN: A Superpixelwise Convolutional Neural Networkfor Salient Object Detection. *Int. J. Comput. Vis.* **2015**, *115*, 330–344. [CrossRef]
39. Hou, Q.; Cheng, M.M.; Hu, X.; Borji, A.; Tu, Z.; Torr, P.H. Deeply Supervised Salient Object Detection with Short Connections. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3203–3321.
40. Hui, S.; Guo, Q.; Geng, X.; Zhang, C. Multi-Guidance CNNs for Salient Object Detection. *ACM Trans. Multimed. Comput. Commun. Appl.* **2022**. *Early Access*. [CrossRef]
41. Liu, N.; Zhang, N.; Wan, K.; Shao, L.; Han, J. Visual Saliency Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4722–4732.
42. Hussain, T.; Anwar, A.; Anwar, S.; Petersson, L.; Baik, S.W. Pyramidal Attention for Saliency Detection. *arXiv* **2022**, arXiv:2204.06788. [CrossRef]
43. 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 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 136–145.
44. Yan, Q.; Xu, L.; Shi, J.; Jia, J. Hierarchical Saliency Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 23–28 June 2013; pp. 1155–1162.
45. Li, G.; Yu, Y. Visual Saliency Based on Multiscale Deep Features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5455–5463.
46. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
47. Jian, M.; Wang, J.; Yu, H.; Wang, G.-G. Integrating object proposal with attention networks for video saliency detection. *Inf. Sci.* **2021**, *576*, 819–830. [CrossRef]
48. Deng, Z.; Hu, X.; Zhu, L.; Xu, X.; Qin, J.; Han, G.; Heng, P.A. R3net: Recurrent residual refinement network for saliency detection. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 684–690.
49. 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 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3127–3135.
50. Liu, N.; Han, J.; Yang, M.H. Picanet: Learning Pixel-Wise Contextual Attention for Saliency Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3089–3098.
51. Gao, S.-H.; Tan, Y.-Q.; Cheng, M.-M.; Lu, C.; Chen, Y.; Yan, S. Highly Efficient Salient Object Detection with 100K Parameters. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 8–14 September 2020; pp. 702–721. [CrossRef]