applied sciences

MDPI

Article

Camouflaged Object Detection Based on Deep Learning with Attention-Guided Edge Detection and Multi-Scale Context Fusion

Yalin Wen [1,*], Wei Ke [1,2,*] and Hao Sheng [1,3,4,*]

1 Faculty of Applied Sciences, Macao Polytechnic University, Macau 999078, China
2 Engineering Research Centre of Applied Technology on Machine Translation and Artificial Intelligence, Ministry of Education, Macao Polytechnic University, Macau 999078, China
3 State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
4 Zhongfa Aviation Institute of Beihang University, Hangzhou 310000, China
* Correspondence: yalin.wen@mpu.edu.mo (Y.W.); wke@mpu.edu.mo (W.K.); shenghao@buaa.edu.cn (H.S.)

Abstract: In nature, objects that use camouflage have features like colors and textures that closely resemble their background. This creates visual illusions that help them hide and protect themselves from predators. This similarity also makes the task of detecting camouflaged objects very challenging. Methods for camouflaged object detection (COD), which rely on deep neural networks, are increasingly gaining attention. These methods focus on improving model performance and computational efficiency by extracting edge information and using multi-layer feature fusion. Our improvement is based on researching ways to enhance efficiency in the encode–decode process. We have developed a variant model that combines Swin Transformer (Swin-T) and EfficientNet-B7. This model integrates the strengths of both Swin-T and EfficientNet-B7, and it employs an attention-guided tracking module to efficiently extract edge information and identify objects in camouflaged environments. Additionally, we have incorporated dense skip links to enhance the aggregation of deep-level feature information. A boundary-aware attention module has been incorporated into the final layer of the initial shallow information recognition phase. This module utilizes the Fourier transform to quickly relay specific edge information from the initially obtained shallow semantics to subsequent stages, thereby more effectively achieving feature recognition and edge extraction. In the latter phase, which is focused on deep semantic extraction, we employ a dense skip joint attention module to enhance the decoder's performance and efficiency, ensuring accurate capture of deep-level information, feature recognition, and edge extraction. In the later stage of deep semantic extraction, we use a dense skip joint attention module to improve the decoder's performance and efficiency in capturing precise deep information. This module efficiently identifies the specifics and edge information of undetected camouflaged objects across channels and spaces. Differing from previous methods, we introduce an adaptive pixel strength loss function for handling key captured information. Our proposed method shows strong competitive performance on three current benchmark datasets (CHAMELEON, CAMO, COD10K). Compared to 26 previously proposed methods using 4 measurement metrics, our approach exhibits favorable competitiveness.

Keywords: camouflaged object detection; EfficientNet; salient object detection; deep learning

Citation: Wen, Y.; Ke, W.; Sheng, H. Camouflaged Object Detection Based on Deep Learning with Attention-Guided Edge Detection and Multi-Scale Context Fusion. *Appl. Sci.* **2024**, *14*, 2494. https://doi.org/10.3390/app14062494

Academic Editors: Douglas O'Shaughnessy and Lorenzo J. Tardón

Received: 30 January 2024
Revised: 4 March 2024
Accepted: 7 March 2024
Published: 15 March 2024

1. Introduction

Object detection [1], recognized as one of the most foundational, time-honored, and complex challenges in computer vision, has received significant attention in the past 20 years, and related technical fields have also developed rapidly. The advancement of object detection technology is driving progress in computer vision very fast. Due to the completely different evaluation indicators and standards for different detection tasks, the

difficulty level varies. So, in recent years, while traditional object detection has flourished, there have also been challenges in other computer vision tasks, including but not limited to the following directions: observing changes in object rotation and size, clearly locating objects, salience object detection [2], and the recently popular Camouflaged Object detection (COD). There are a series of related problems in the direction of COD that we need to focus on and commit to solving.

Before the level of advancement in the current computer vision application technology, traditional methods were to use manually annotated features, which were defined as the attention of pixels to each other. So, the traditional methods of manually annotating feature models have a receptive domain of all pixels, and the characteristics of the object depend on all pixels besides the pixels of this object. However, traditional methods have two major drawbacks—poor detection performance and slow processing speed—which greatly limit the accuracy of detecting objects with highly similar textures and colors to the background. With the ongoing introduction of new CNN models and methods, in solving the problem of camouflaged object detection, the performance of constructed neural models is much better than that of traditional manual annotation methods.

Convolutional networks can use GPUs for efficient computation to improve processing speed and propose a series of innovative model structures, such as VGG [3] and ResNet50 [4], which are already sufficiently mature. There has also been a focus on designing efficient feature fusion processors, the corresponding preferences of which can be set for feature extraction from images.However, at present, there are still two obvious problems with CNN-based feature extraction models. First, such models are not sensitive to global position information. This is because the extraction of information based on CNN feature models has a translation function, which is considered an advantage of CNN feature models. However, this weight sharing is a disadvantage for processing camouflaged object detection objects. Second, the information features extracted by the convolutional model only consider a portion of the extracted pixels during each operation. Generally speaking, the receptive fields of the model, built according to the convolutional neural network, will increase with the number of layers of the built model, that is, the deeper the layers are, the larger the range of receptive fields will be. Most skeleton networks built based on CNN now exchange information loss for a wider range of receptive fields. In order to try to recover the lost structural information, the general model structure needs to add a decoder to recover the lost information.

However, there is currently no suitable method to recover all unnoticed information. Furthermore, the skeletal network constructed using CNN must encompass the entire image within its receptive field range. However, previous experiments show that this was only in theory, and the range of receptive field perception is not as large as expected. If we can build a network that can sense a large range of receptive fields while retaining relevant details, it may be useful for processing the task of camouflaged object detection depending on contextual information. Due to the significant limitation of simple convolutional networks, researchers have found in recent years that Transformers have great potential in handling contextual information problems.

Camouflaged object detection is also an emerging research field in semantic recognition, which has been proposed upon the combination of deep learning and computer vision to process images and videos. Because it is very difficult to distinguish between the disguised object and the environment, it is, consequently, very difficult to process images with such contents. A camouflaged object detection network based on deep learning is generally structured into two parts, namely, a module for locating or obtaining camouflage object features and a module for segmenting object detection. The two modules interact with each other to fuse features and obtain the final camouflage component. Our proposed approach, based on EfficientNet-B7 and SWIN-T, combines a new overall architecture for feature extraction, including feature encoding with an attention flow mechanism and feature fusion, as well as feature decoding with dual tasks. The main improvements are listed as follows:

- The feature fusion multi-path joint idea is added in the encoder section, which can better utilize cross-layer links to fuse multiple layers of information. Deep networks leveraging Transformers more effectively account for the contextual information of objects, upgrade the correctness of positions, and reduce noise.
- A novel fusion module is introduced, establishing a connection through the attention mechanism between the encoding and decoding phases. It merges local and global messages, addressing the challenge of disparate fusion techniques across two modes. It proposes an attention weight that is completely different from the single mode, and combines salient features and residual connections to achieve cross-modal guided fusion, that is, the joint dense hop module.
- In the decoding section, the single task is changed to a dual task, shifting from only the output of region segmentation maps to the concatenated region segmentation maps and contour segmentation maps, to better achieve camouflaged object detection tasks and improve accuracy.

To enhance the distinction between foreground objects and similar background objects, our approach focuses on identifying visual sensory mechanisms employed by organisms in camouflage, along with other deceptive strategies used in images, thereby increasing the likelihood of object detection. We have developed a comprehensive model for camouflaged object detection, which is grounded in the integration of joint dense jumps and dual-task outputs. This model has demonstrated notable effectiveness on three widely recognized benchmark datasets. Compared to 19 previously proposed methods, our approach achieved better results across all 4 reference metrics.

2. Related Work

The purpose of ordinary object detection [2] is to discover and locate the types and positions of all objects of interest in images or videos, which is a major problem in computer vision. Object detection can be practiced in multiple areas, for example, facial recognition, pedestrian detection, vehicle detection, and remote sensing detection. The task of object detection generally consists of two parts [5]: classification and localization. Classification refers to determining which predefined category an object belongs to, such as cats, dogs, or human beings. Positioning refers to determining the position of an object in an image or video, usually represented by a rectangular box called a bounding box. The disadvantage of object detection is that the object may appear at any position in the image, with different sizes, shapes, poses, and occlusion situations, and the background may be complex. Therefore, object detection requires strong feature extraction and representation capabilities, as well as effective search and filtering strategies. Existing ordinary object detection methods are, thus, not effective in detecting camouflage objects.

For salient object detection (SOD), Wang [6] proposed a new way in light fields. It combines a multi-modal and multi-level feature aggregation network, utilizing the complementary information of multi-view images, focus stacking, and depth maps in light field data to process objects. The information within the light field object was enriched, and it achieved good results, improving the performance. Chen [7] proposed another way called BPFINet to handle the problem of salient object detection, which selected the most prominent object from the image. The method uses U-shaped structure features to gradually combine the low/high-level information with global information layer by layer. It also introduces a self-developed BAM module to improve the certainty of SOD and its accuracy. The creator of ERBANet [8] proposed a model for SOD that enhanced boundary perception and region partitioning, and utilized attention enhancement modules to improve detection performance, accuracy, and robustness. They proposed a model that solved the issue of SOD by using the depth of spatial dilated convolution based on spatial attention. This model can efficiently extract object information and mix feature extraction to expand the range of receptive fields, thereby improving the achievement and accuracy of SOD.

Cui [9] proposed a model that solved the issue of SOD by using a depth spatial dilated convolution based on spatial attention. This model can efficiently extract object information and mix feature extraction to expand the range of receptive fields, thereby improving the achievement and accuracy of SOD. Kervadec's [10] main contribution was to provide a new loss function for solving image segmentation problems. Starting from the contour of the image, using image optimization techniques, the L_2 distance on the image contour is called region integration. The proposed new loss function obtained by utilizing image contours can be used alone or in combination with other loss functions. They also proved that the loss function could improve model performance and enhance training stability. Ning [11] proposed a network structure for multi-channel SOD based on FPN, introducing a global feature module and utilizing control gates and pooling networks. They integrated the FPN network and SINet network into their own network, processed features of different scales through feature fusion, and then convoluted the two single-channel character maps to achieve dimensional maps. This method can extract some large and observed information from multiple channels and retain big parts in the feature map. Li [12] introduced a novel approach to salient object detection (SOD) through the development of CFPN, a method designed to effectively integrate features from various levels, thereby enhancing the accuracy and completeness of detection. This model initially converts features from distinct levels into graphs that encapsulate both shallow and deep semantic information, subsequently redistributing this information. In an effort to minimize losses, each layer acquires additional semantic information from other layers. Building on this, the model is then able to capture the contextual information of the effectively processed information.

Ullah [13] proposed a new model for SOD. This new model mainly comprises three parts: global context, multi-scale aggregation, and adaptive enhancement. The authors conducted experiments on six public datasets, and the results showed that they were better than previous results across all indicators, obtaining good results with robustness. Ji [14] proposed a salient object detection method to improve the depth calibration and the fusion of RGB-D images, which could effectively enhance performance. The proposal of deep calibration as a learning strategy can enhance the testing effectiveness of salient object detection. A simple and effective fusion cross-model can handle both shallow and deep semantic information, which means it can handle more complex scenes. Chen [15] aimed to deal with the issue of SOD through 3D convolutional neural networks. They established a new order and attempted preprocessing during the encoder part. The effective fusion in the decoder part can better interpret the information of shallow and deep semantics. This method has achieved good results on six publicly available datasets, outperforming other methods.

In the direction of Camouflaged Object detection (COD), Chen [16] proposed a Global-Context-Aware Progressive Aggregation Network (GPANet) for SOD, which highlights the most eye-catching objects from images. This network utilizes global contextual information and mixed information to improve the localization and segmentation achievement of salient objects. It can also be used for camouflaged object detection, enhancing the perception ability of camouflaged objects by using depth maps as additional input. Wang [4] proposed a camouflage object segmentation method with interference mining (PFNet) for segmenting camouflage objects from complex scenes. This method imitates the predation process in nature, first locating potential objects from a global perspective through a positioning module (PM) and then gradually refining the segmentation results by focusing on fuzzy areas through a focusing module (FM).

He [17] proposed a COD method based on mutual graph learning (MGL) to improve the performance of COD on RGB-D images. This method utilizes two graph-based interaction modules, one for obtaining semantic guidance knowledge from edge information and the other for enhancing object localization by combining real edge priors. This method achieves repeated mutual learning and mutual assistance by clearly inferring the relationship between two tasks. Li [12] proposed a cross-layer feature pyramid network (CFPN) for SOD. This network allocates features from different levels to all involved levels, allowing

each layer's features to have both semantic and detailed information from other layers, thereby reducing the loss of important information.

In order to further explain the challenge of the COD problem, Lv [18] proposed a triple-task learning framework that can simultaneously perform three tasks: localization, segmentation, and sorting of camouflage objects. It indicates the saliency level of camouflage objects relative to the background. The authors believed that binary segmentation settings could not fully understand the concept of camouflage, and saliency models could help design more complex camouflage techniques. They also used an eye tracker to generate localization maps and generate ranking-based training and testing datasets based on instance-level labels. They also contributed the largest COD test set and comprehensively analyzed the performance of the COD model. The experimental results indicated that the triple-task learning framework had reached a new state-of-the-art level, making the COD network more interpretable. Lv [19] believed that for camouflaged object detection tasks, it was not only necessary to segment the camouflaged object but also to evaluate the significance of the background in order to design an overall strategy for camouflaged object detection. Moreover, the author of the article also contributed a lot of COD information to analyze the abilities of camouflaged object detection models.

Chen [20] proposed a new network structure for detecting hidden objects in natural scenes. This network structure utilizes the Attention-induced Cross-level Fusion Module (ACFM) and Dual-brand Global Context Module (DGCM) to fuse multiple layers of features and generate accurate predictions. They conducted experiments on three commonly used datasets, demonstrating the effectiveness and superiority of C2FNet. The article also demonstrated the potential of C2FNet in the application of polyp segmentation.

Bi [21] mainly focused on four aspects. A comprehensive introduction and analysis were conducted on 39 existing COD models, from 1998 to 2021, including traditional manual feature-based models and deep learning-based models, and classification and a comparison were made based on different detection mechanisms and strategies. A detailed introduction and evaluation were conducted on four widely used COD datasets, including the construction methods, object categories, number of images, annotation methods, and difficulty. Some improvement suggestions were proposed for the datasets. They also compared the performance of existing COD models on four datasets, including mathematical comparison, image comparison, and effect evaluation, and then analyzed the detection performance of these models on different types of objects. The limitations and challenges of COD were discussed as well, and some future research directions were proposed, such as utilizing multimodal information, designing more complex camouflage techniques, and developing larger and more diverse datasets.

Hu [22] proposed a context aggregation-based small object detection method to improve the detection performance of YOLOv3-based detectors for small objects. This method utilizes multi-scale contextual information, including global scene information, local object information, and intermediate region information, to enhance the feature representation of small objects. The method has been tested on multiple public datasets, demonstrating its effectiveness and superiority. Liu [23] proposed a feature-enhanced context-aware network (FECANet) for small sample semantic segmentation to solve the problem of small sample semantic segmentation. The network utilizes multiple layers of contextual information, including global semantic information, local edge information, and intermediate spatial information, to enhance the feature representation of the object. The network also utilizes a self-attention mechanism and a multi-scale fusion module to improve the selectivity and adaptability of features. It has been tested on multiple public datasets, confirming its validity and nobility.

The EfficientNets model [24] provides a simple and efficient composite model based on balancing network depth, width, and decomposition rate after scaling existing ConvNets models, and further validates the validity of this method on MobileNets and ResNet. EfficientNet-B7 achieved the most advanced top-1 accuracy of 84.4% and a top-5 accuracy of 97.1% on ImageNet. At the same time, it is 8.4 times smaller than ConvNet and its inference

speed is 6.1 times faster. Moreover, it was experimentally confirmed that EfficientNets has excellent transfer ability, achieving extremely high accuracy on CIFAR-100 (91.7%), Flower (98.8%), and another three learning datasets, with a parameter count that is an order of magnitude lower.

On the other hand, the biggest drawback of CNNs in processing image features for a long time is the loss of relevant structural information in the process of extracting images, and in terms of receptive fields, the actual receptive fields obtained by using CNNs alone will be much smaller than theoretical expectations. The current work focuses on how to design a robust algorithm to accurately detect and segment camouflaged objects in given images. There is limited research on the current state and future prospects of camouflaged object detection (COD), especially regarding statistics and analysis of deep learning-based algorithms.

Liang [25] comprehensively introduces all the research achievements in COD using deep learning. They provide in-depth analyses and discussions on model structures, learning paradigms, datasets, evaluation metrics, and performance comparisons, detailing all COD models involved and offering deep analysis and comparisons of some of the more effective models. In the context of sensor-based model research for camouflaged object detection, Transformers [26] were initially used as the backbone for encoders. SETR [27] marked the beginning of applying Transformers to semantic segmentation, and the introduction of Swin Transformer further accelerated the innovation of method models in the field of computer vision. SwinNet [28], by integrating RGB-D and RGB-T modes, achieved a cross-modal fusion model for salient object detection (SOD). It designed a dual-stream SwinT and used spatial alignment and channel recalibration modules to optimize features. DMT [29] was the first to introduce self-attention to the study of differentiating foreground and background information in images. HitNet [30] introduces a new cross-scale recursive method, significantly improving model evaluation metrics.

Researchers then focused on exploring the combination of Transformers and CNNs to efficiently complement feature extraction and fusion, incorporating the idea of self-attention. UGTR [31] ingeniously combined CNNs and Transformers, explicitly using probabilistic representation models to learn the uncertainty of objects stored within the Transformer framework. CamoFormer [32] designed a COD algorithm inspired by multi-head self-attention, setting SwinT as its backbone and employing a multi-head self-attention masking strategy for feature aggregation. Currently, there is scant research using this approach, indicating substantial opportunities for further investigation in this field.

Therefore, camouflaged object detection models that rely solely on completing CNN tasks often cannot search for information from a global perspective and cannot fully capture all contextual information. In 2017, Vaswani [33] proposed a Transformer for natural language processing that utilized self-attention mechanisms to find and obtain certain dependency relationships, which were long-distance and could better obtain global information.

However, learning contextual features directly from the entire image is very unwise and meaningless, as it is often disturbed by various obvious information. Therefore, like other camouflaged object research, we propose a method of exploring context based on attention to background features, fully combining the above advantages to achieve an improvement in the benchmark rate of camouflaged object detection.

3. Methodology

Through the above literature review, we can see that based on the outstanding achievements of Transformers in the field of natural language processing and its excellent transform ability, we plan to introduce it into the task of camouflaged object detection. However, for some intensive prediction tasks, the position encoding of Transformers performs poorly in modeling accurate "spatial" information. Considering the above factors, we want to combine the advantages of CNN and Transformer models; therefore, we utilize the EfficientNet-B7 model as the benchmark model to process camouflaged object detection data, and on this basis, we replace the B6 and B7 modules of EfficientNet-B7 with Swin-T.

3.1. Model Overview

We choose the EfficientNet-B7 model as the benchmark model to process camouflaged object detection data, and merge the original seven modules into four modules. In addition to the initial convolution module, the output resolution is changed. To this end, we represent the output of each encoder as E_1, and record the shallowest semantic information obtained from the initial use of the encoder as E_1, enhancing its information to improve efficiency. In the decoder stage, we design a joint dense skip attention module that can aggregate various levels of features separately and merge them, and then finally merge the output of the encoder and decoder. In the joint dense skip attention module, we write the output results of the remaining three encoders as E_2, E_3, and E_4, and integrate these three inputs together. Each layer extracts different semantic information, and the deeper the level, the richer the semantic information extracted. Therefore, in order to obtain clearer edge information and richer semantic information, a joint dense skip attention module is used. Finally, we merge the output of the encoder and decoder to complement the shallow and deep semantic information, and use this supplementary information to compensate for the difference between the encoder and decoder that extract the shallow semantic information. The specific model is shown in Figure 1.

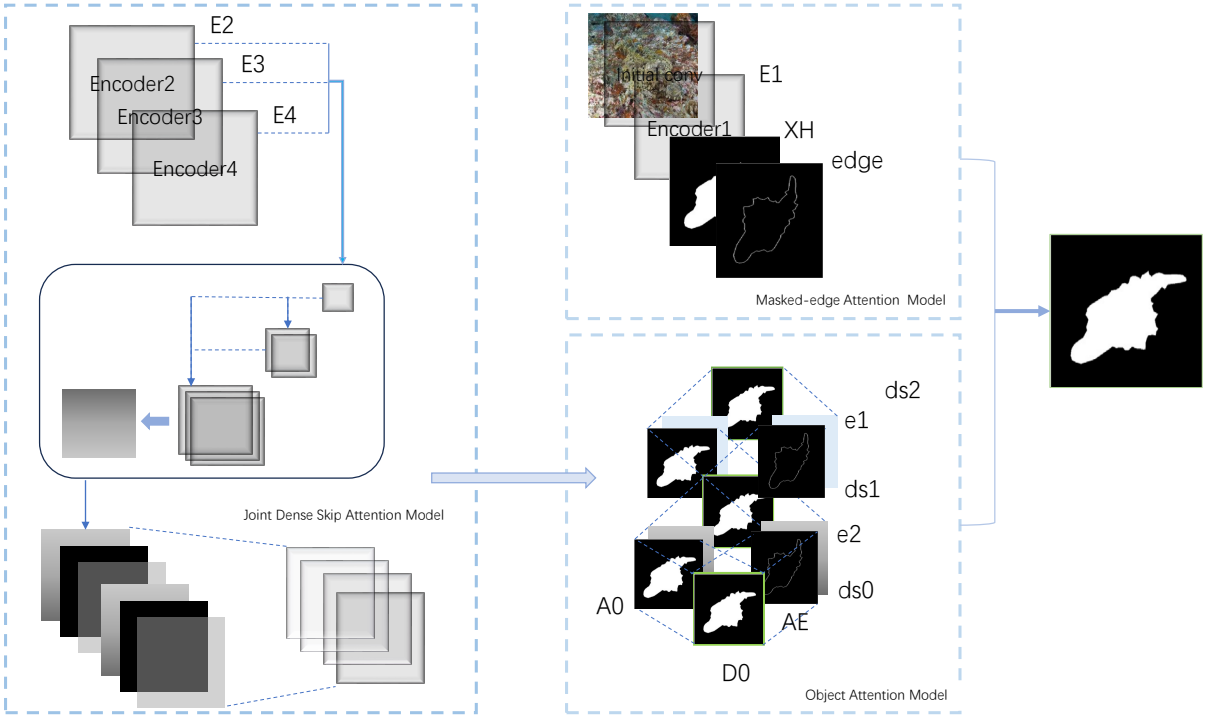

Figure 1. Overall framework of the proposed model. Masked-edge Attention Module is explained in Section 3.2.1. Joint Dense Skip Attention Module is explained in Section 3.2.2. Object Attention Module is explained in Section 3.2.3.

3.2. Camouflaged Object Detection Based on Attention Enhancement

Adding attention is highly effective in enhancing camouflaged object detection and identifying clearer objects. To improve the computational efficiency, we propose an attention-enhanced camouflage object module to track the edges and objects of camouflage objects. Thus, our camouflaged object detection task is divided into two sub-tasks, namely, tracking objects and edges.

3.2.1. Masked-Edge Attention Module

First, the module detects edge information. We mainly use the Fourier transform to quickly extract more obvious shallow semantic information, namely, edge information, E_1, and enhance the information of E_1. However, the edge extraction method we use can

only extract shallow semantic information, which cannot clearly represent deeper semantic information. Deeper semantic information requires deeper decoder outputs to obtain different edge information. Therefore, we use the Fourier transform to quickly extract shallow semantic information E_1 by FFT [1,34,35] and FFT^{-1}, and the first encoder's representation is divided into shallow and deep layers, as shown in Equation (1):

$$A_h = FFT^{-1}\left(f_r^h(FFT(A))\right) \tag{1}$$

where A is the input feature, and $FFT(\cdot)$ and $FFT^{-1}(\cdot)$ represent the fast Fourier transform and its inverse transform, respectively. Furthermore, f_r^h acts as a high-pass filter, designed to remove all frequencies outside of a specific radius r, retaining only those within it.

In addition, representing deep semantic information, r refers to the radius of the circle. In order to obtain edge information, we use a high-pass filter to obtain all the required deep edge semantic information. Because A_h contains background noise in the field of conversion, we need to eliminate the noise by expanding the receptive field $\mathcal{RFB}(\cdot)$ and generating the boundary feature $E = \mathcal{RFB}(A_h)$. Finally, we reinforce the edges $A_E = A + E$ to obtain clear edge detection information.

3.2.2. Joint Dense Skip Attention Module

Our main purpose in establishing the joint dense jump attention module is to obtain richer semantic expressions from the contextual channels and spaces, that is, to aggregate them at multiple levels. In the following equations, $f(\cdot)$ represents convolutional operation, and $cat(\cdot)$ represents channel-wise feature concatenation. The output result of each encoder is denoted as E_2, E_3, and E_4, corresponding to the number of channels of sizes 32, 64, and 128, expressed as in Equations (2)–(4):

$$E_2' = E_2 \otimes f(Up(E_3)) \otimes f(Up(Up(E_4))) \tag{2}$$

$$E_3'' = f(\text{cat}[E_3 \otimes f(Up(E_4)), f(Up(E_4))]) \tag{3}$$

$$E_2'' = f(Up(E_3'')) \tag{4}$$

We obtain an aggregated representation, which is the scale of E_2, through A, $A = f(cat[E_2', E_2'']) \in \mathbb{R}^{(32+64+128)\times H_2 \times W_2}$. UP refers to upsampling; for instance, Up(E_3) signifies upsampling the features of the E_3 layer, typically resizing them to match the dimensions of the E_2 layer. After the aggregation, it still represents the contextual information. The current research on COD has separated the channel and spatial attention modules for decoding and expanding the receptive field. So, we first distinguish the relatively important channel contextual information:

$$\alpha_c = \sigma\left(\frac{\exp\left(\mathcal{P}_q(\widetilde{A})\left(\mathcal{P}_k(\widetilde{A})\right)^\top\right)}{\sum\exp\left(\mathcal{P}_q(\widetilde{A})\left(\mathcal{P}_k(\widetilde{A})\right)^\top\right)}\mathcal{P}_v(\widetilde{A})\right) \tag{5}$$

where $\tilde{A} \in \mathbb{R}^{C\times 1\times 1}$ represents channel level pooling representation, and $\mathcal{P}(\cdot)$ represents the use of 1×1 convolutional kernel operation, utilizing the softmax function and self-attention, combined with the sigmoid function to distinguish important channels and obtain deep semantic information, as shown in Equation (5). In order to obtain more detailed deep semantic information, we perform aggregation transformation on it and set the confidence channel weights as follows: $A_c = (A \otimes \alpha_c) + A$. Subsequently, we rely on α_c and the confidence ratio γ. The distribution retains the confidence channels, as shown in Equation (6):

$$\tilde{A}_c = A_c \otimes \text{ mask} \begin{cases} \text{mask } = 1, & \text{if } \alpha_c > P^{-1}(\gamma) \\ \text{mask } = 0, & \text{otherwise} \end{cases} \tag{6}$$

We use $P^{-1}(\gamma)$ to denote the γ quantile of α_c. Then, the refined $\tilde{A}_c$ is computed spatially to discriminate the camouflaged object detection and generate the first decoder representation $D_0 \in \mathbb{R}^{1\times H_2\times W_2}$, and the announcement is as follows Equation (7):

$$D_0 = \frac{\exp\left(\mathcal{Z}_q\left(\widetilde{A}_c\right)\left(\mathcal{Z}_k\left(\widetilde{A}_c\right)\right)^{\top}\right)}{\sum\exp\left(\mathcal{Z}_q\left(\widetilde{A}_c\right)\left(\mathcal{Z}_k\left(\widetilde{A}_c\right)\right)^{\top}\right)}\mathcal{Z}_v\left(\widetilde{A}_c\right) + \mathcal{Z}_v\left(\widetilde{A}_c\right) \tag{7}$$

We use a convolution operation to map the input features to a new space A_c. We generate a decoder D_0 using A_c, which is of the same size as A_c. The features projected by $\mathcal{Z}(\cdot)$ through convolution operation onto A_c. We upsample D_0 to the size of the original input to obtain DS_0, which is a graph used to provide deep supervision.

3.2.3. Object Attention Module

The primary aim of designing this module is to minimize distributional disparities between the encoder and decoder representations while using the fewest parameters possible. We have introduced an Object Attention Module as an integral part of the decoder. In comparison to existing research [36], we maintain the decoder as a single channel, denoted as D, to enhance efficiency. We leverage the Object Attention Module to track objects and complement edge information from each decoder ($D_i \in \mathbb{R}^{1\times H\times W}$). To better refine camouflaged objects, we compute object weights α_O according to the following formula $\alpha_O = \sigma(D_i)$. However, α_O may not always fully detect the entire object with clearly defined edge regions. Therefore, we generate complementary edge weights α_E to cover areas that remain undetected. For each pixel x_{ij} in the decoder, we invert the detected regions and employ a specific formula Equation (8) to eliminate background noise associated with missing region detection("d" means denoising ratio):

$$\alpha_E = \begin{cases} 0, & \text{if } \left(-\sigma\left(a_{ij}\right)+1\right) > d \\ -\sigma\left(a_{ij}\right)+1, & \text{otherwise} \end{cases} \tag{8}$$

We incorporate the encoder output $E_{i\in\{2,1\}}$ and decoder feature $D_{i\in\{0,1\}}$, as shown in Equation (9). To reduce discrepancy, we exploit a receptive field operation $\mathcal{RFB}(\cdot)$ and upsample $D_i + 1$ to generate $DS_i + 1$.

$$D_{i+1} = \mathcal{RFB}((\alpha_O \otimes E_{2-i}) + (\alpha_E \otimes E_{2-i})) \tag{9}$$

3.3. Loss Function

To calculate the loss for each pixel, COD tasks typically use binary cross entropy loss (LBCE). However, when the background pixels are far more than the current scene pixels, the model tends to ignore the foreground. Wei [37] used a weighted binary cross entropy loss (L_wBCE) that assigned different weights to pixels based on the difference between them and their neighborhoods. In addition, they also introduced weighted IOU loss (L_wBCE), which could achieve global constraints, and was proven effective in salient object detection.

$$L = L_wBCE + L_wIOU. \tag{10}$$

L_wBCE is the weighted binary cross-entropy loss. Binary cross-entropy loss is commonly used in binary classification problems to measure the difference between the probability distribution predicted by the model and the actual labels. "Weighted" implies that each sample can have a different level of importance in the calculation of the loss. L_wIOU is the weighted Intersection over Union (IoU) loss. The Intersection over Union is a commonly used metric to evaluate the performance of object detection models, especially in terms of bounding box predictions. It calculates the overlap area between the predicted bounding boxes and the actual bounding boxes. The weighted version means that different predictions can have varying weights in the loss calculation. Overall, this formula combines

binary cross-entropy loss and Intersection over Union loss, which we use for camouflaged object detection tasks where it is necessary to consider both classification accuracy and spatial location accuracy. Through this combination, we can simultaneously optimize our model's ability to classify and locate camouflaged objects.

4. Experiment

4.1. Experiment Setup

Datasets . There are currently three main datasets for camouflaged object detection: CHAMELEON [38], CAMO [39], and COD10K [40]. The method proposed in this article needs to demonstrate good generalization ability and robustness in different camouflage scenarios. Therefore, we used three mainstream public datasets for evaluation criteria, which included multiple categories of camouflage objects, such as animals, plants, and artificial objects, with different camouflage difficulties, such as background matching, destructive colors, and edge blurring, to fully evaluate the advantages and disadvantages of our method. This article follows a previous work [40] and uses the training set of CAMO [39] and COD10K [40] as the training set (4040 images). The remaining images will be used as the test set.

Evaluation Metrics. In order to evaluate the performance of the method proposed in this article, four widely used standard metrics were used, namely, a structural metric $S_\alpha \uparrow$ [41], which takes into account the structural similarity between saliency maps and truth values; an adaptive E-metric $E_\phi^{ad} \uparrow$ [42], which takes into account the degree of edge alignment between saliency maps and true values; a weighted F-measure $F_\beta^w \uparrow$ [3], which considers the balance of accuracy and recall between saliency maps and true values; and the mean absolute error $M \downarrow$, which takes into account pixel-level differences between saliency maps and true values.

Experimental Process. We used PyTorch [43] to implement our proposed model. We trained and tested our model on an Ubuntu 20.04 system using two NVIDIA GeForce RTX 3070 GPUs. During the model training phase, we set the size of the input image to 320×320 and performed data augmentation with random horizontal flipping and color jitter. We used the pre-trained ENet-B7 [24] model on ImageNet to initialize the parameters of the encoder network, and adopted a poly [44] strategy to adjust the learning rate, where the initial learning rate was 0.001 and the exponent was 0.9. In the testing phase, we first performed network inference on the image and then restored the size of the output image to the original size of the input image. We used bilinear interpolation in both size transformations. We did not use any post-processing methods to improve the final output performance.

4.2. Comparison to the State-of-the-Art

To demonstrate the effectiveness of the proposed model, we compared it with 26 existing newer methods: FPN [45], PSPNet [46], Mask RCNN [47], HTC [48], and MSRCNN [49], DSC [50], BDRAR [51], UNet++ [52], and PraNet [53], PiCANet [54], BASNet [55], CPD [56], PFANet [57], EGNet [36], F3Net [37], GCPANet [16], MINet-R [58], SINet [40], PFNet [59], UGTR [31], RankNet [19], BgNet [60], ERRNet [61], FEDER [17], MRRNet [62], and BCNet [63]. Table 1 shows the qualitative results of our model and some models in recent years.

Table 1 shows the quantitative results of our proposed method compared to 26 other common methods on three benchmark datasets. Numerically, it can be seen that our method is quite competitive. Compared to the three methods from 2023, our M value is quite competitive. On the Chameleon dataset, our method improved by 0.002 over FEDER, by 0.005 over MRRNet, and by 0.006 over BCNet. On the Camo dataset, our model improved by 0.002 over FEDER, by 0.007 over MRRNet, and by 0.007 over BCNet. On the COD10k dataset, our model improved by 0.004 over MRRNet and by 0.009 over BCNet, obtaining good results.

Table 1. Our proposed method was compared with 26 state-of-the-art methods in related fields across four evaluation metrics: structural measure S_α, adaptive E-measure E_ϕ^{ad}, weighted F-measure F_β^w, and mean absolute error M. Structural measures S_α, adaptive E-measures E_ϕ^{ad}, and weighted F-measures F_β^w are positively correlated with results, and mean absolute error M is negatively correlated with results, meaning that the larger the first three test results, the better, and the smaller the last result, the better, as indicated by arrows. We evaluated prediction graphs of all methods on three benchmark datasets using the same code. ○ represents object detection method. • represents semantic segmentation method. ⋆ represents instance segmentation methods. △ represents shadow detection methods. § represents medical image segmentation methods. † represents SOD methods. * represents COS methods. The experimental results show that our method outperforms other methods significantly across all datasets and evaluation metrics, with the perfect conclusion highlighted in **bold**. The ↑ means that a larger value indicates better performance, the ↓ means that a smaller value indicates better performance.

Methods	Year	Chameleon (76 Images)				Camo Test (250 Images)				COD10K-Test (2026 Images)			
		$S_\alpha\uparrow$	$E_\phi^{ad}\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi^{ad}\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi^{ad}\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$
FPN ○ [45]	2017	0.794	0.835	0.590	0.075	0.684	0.791	0.483	0.131	0.697	0.711	0.411	0.075
PSPNet • [46]	2017	0.773	0.814	0.555	0.085	0.663	0.778	0.455	0.139	0.678	0.688	0.377	0.080
Mask RCNN ⋆ [47]	2017	0.643	0.780	0.518	0.099	0.574	0.716	0.430	0.151	0.613	0.750	0.402	0.080
UNet++ § [52]	2018	0.695	0.808	0.501	0.094	0.599	0.740	0.392	0.149	0.623	0.718	0.350	0.086
DSC △ [50]	2018	0.850	0.888	0.714	0.050	0.736	0.830	0.592	0.105	0.758	0.788	0.542	0.052
PiCANet † [54]	2018	0.769	0.836	0.536	0.085	0.609	0.753	0.356	0.156	0.649	0.678	0.322	0.090
BDRAR △ [51]	2018	0.779	0.881	0.663	0.064	0.759	0.825	0.664	0.093	0.753	0.836	0.591	0.051
HTC ⋆ [48]	2019	0.517	0.490	0.204	0.129	0.476	0.442	0.174	0.172	0.548	0.521	0.221	0.088
MSRCNN ⋆ [49]	2019	0.637	0.688	0.443	0.091	0.617	0.670	0.454	0.133	0.641	0.708	0.419	0.073
BASNet † [55]	2019	0.687	0.742	0.474	0.118	0.618	0.719	0.413	0.159	0.634	0.676	0.365	0.105
CPD † [56]	2019	0.853	0.878	0.706	0.052	0.726	0.802	0.550	0.115	0.747	0.763	0.508	0.059
PFANet † [57]	2019	0.679	0.732	0.378	0.144	0.659	0.735	0.391	0.172	0.636	0.619	0.286	0.128
EGNet † [36]	2019	0.848	0.879	0.702	0.050	0.732	0.827	0.583	0.104	0.737	0.777	0.509	0.056
F3Net † [37]	2020	0.854	0.899	0.749	0.045	0.779	0.840	0.666	0.091	0.786	0.832	0.617	0.046
GCPANet † [16]	2020	0.876	0.891	0.748	0.041	0.778	0.842	0.646	0.092	0.791	0.799	0.592	0.045
PraNet § [53]	2020	0.860	0.898	0.763	0.044	0.769	0.833	0.663	0.094	0.789	0.839	0.629	0.045
MINet-R † [58]	2020	0.844	0.919	0.746	0.040	0.749	0.835	0.635	0.090	0.759	0.832	0.580	0.045
SINet * [40]	2020	0.869	0.899	0.740	0.044	0.751	0.834	0.606	0.100	0.771	0.797	0.551	0.051
PFNet * [59]	2020	0.882	0.942	0.810	0.033	0.782	0.852	0.695	0.085	0.800	0.868	0.660	0.040
UGTR * [31]	2021	0.888	0.921	0.804	0.031	0.784	0.858	0.747	0.086	0.817	0.850	0.670	0.036
RankNet * [19]	2021	0.890	0.936	0.835	0.030	0.787	0.859	0.756	0.080	0.804	0.883	0.699	0.037
BgNet * [60]	2022	0.885	0.936	0.822	0.032	0.804	0.872	0.766	0.075	0.804	0.866	0.676	0.039
ERRNet * [61]	2022	0.877	0.923	0.815	0.036	0.760	0.830	0.722	0.088	0.780	0.856	0.656	0.044
FEDER * [17]	2023	0.887	0.943	0.847	0.030	0.804	0.877	0.786	0.071	0.822	0.901	0.740	0.032
MRRNet * [62]	2023	0.882	0.924	0.812	0.033	0.811	0.870	0.766	0.076	0.822	0.869	0.695	0.036
BCNet * [63]	2023	0.883	0.932	0.821	0.034	0.799	0.875	0.759	0.076	0.800	0.858	0.673	0.041
Ours *	2024	**0.853**	**0.914**	**0.776**	**0.028**	**0.802**	**0.865**	**0.727**	**0.069**	**0.802**	**0.890**	**0.676**	**0.032**

4.3. Ablation Experiment

We conducted ablation experiments to validate the effectiveness of our proposed new model, as shown in Table 2 and Figure 2.

From the results chart, combined with the accompanying chart, it is evident that our proposed method has achieved good results in the camouflaged object detection task. Comparing the final results with the GT chart, it can be found that details such as fish fins are also well discovered. From Table 2, we can see that each operation of our model has improved all evaluation indicators on the three benchmark datasets. On the CHAMELEON dataset, our method increased S_α from an initial 0.694 to 0.853, an increase of 0.159. We increased E_ϕ^{ad} from an initial 0.833 to 0.914, an increase of 0.081. We increased F_β^w from 0.526 to 0.776, an increase of 0.25, and M from 0.076 to 0.028, an increase of 0.051. On the CAMO dataset, S_α increased from an initial 0.634 to 0.802, an increase of 0.168. We

also increased E_{ϕ}^{ad} from an initial 0.745 to 0.865, representing an improvement of 0.12; F_{β}^{w} increased from 0.457 to 0.727, representing an increase of 0.27; and M increased from 0.127 to 0.069, representing a decrease of 0.058. On the COD10K dataset, our method increased S_{α} from the initial 0.659 to 0.802, increased by 0.143. E_{ϕ}^{ad} increased from an initial 0.775 to 0.890, representing an increase of 0.115; F_{β}^{w} increased from 0.420 to 0.676, representing an increase of 0.256; and M increased from 0.066 to 0.032, representing a decrease of 0.034. It can be seen that our proposed method is effective.

Table 2. Ablation analysis: "B" represents our baseline EfficientNet-B7, "Q" represents the model after replacing the Swin Transformer module in B6 and B7 of "B", and "W" represents the addition of dense and attention mechanisms. The results demonstrate that each step yields effective outcomes, thereby substantiating the efficacy of our model.The ↑ means that a larger value indicates better performance, the ↓ means that a smaller value indicates better performance.

	Networks	CHAMELEON (76 Images)				CAMO Test (250 Images)				COD10K-Test (2026 Images)			
		$S_{\alpha}\uparrow$	$E_{\phi}^{ad}\uparrow$	$F_{\beta}^{w}\uparrow$	$M\downarrow$	$S_{\alpha}\uparrow$	$E_{\phi}^{ad}\uparrow$	$F_{\beta}^{w}\uparrow$	$M\downarrow$	$S_{\alpha}\uparrow$	$E_{\phi}^{ad}\uparrow$	$F_{\beta}^{w}\uparrow$	$M\downarrow$
(a)	B (our)	0.694	0.833	0.526	0.076	0.634	0.745	0.457	0.127	0.659	0.775	0.420	0.066
(b)	Q (B + T)	0.796	0.903	0.687	0.047	0.685	0.720	0.553	0.109	0.719	0.829	0.639	0.048
(h)	B + W (D + A)	0.822	0.901	0.730	0.038	0.742	0.826	0.638	0.092	0.758	0.855	0.605	0.043
(l)	Ours	0.853	0.914	0.776	0.028	0.802	0.865	0.727	0.069	0.802	0.890	0.676	0.032

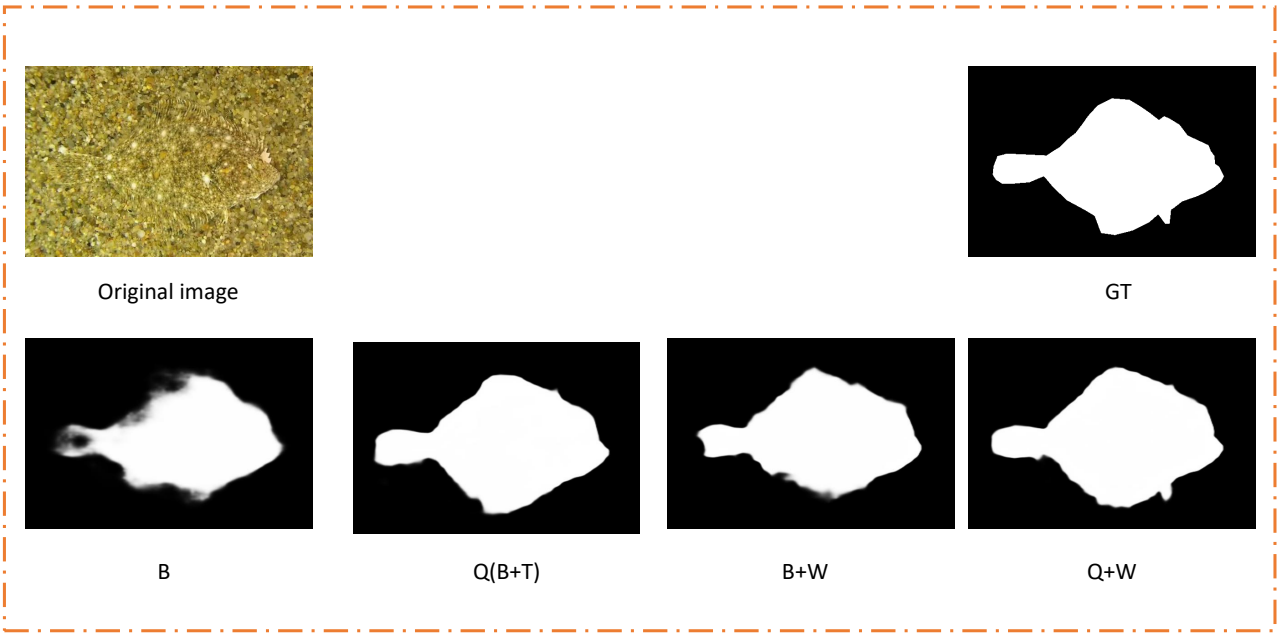

Figure 2. The results of the ablation experiment.

5. Conclusions

In this work, our goal was to tackle the challenge of accurately segmenting camouflaged objects. We developed a new method for detecting camouflaged objects, in which we replaced the B6 and B7 layers of Effenet-B7 with the SWIN-T module and combined dense skip attention with object processing. In camouflaged object detection tasks, our proposed new model achieved good results on the benchmark datasets. In the camouflaged object task, our newly proposed model demonstrated good achievement on the benchmark datasets. In the next work, we plan to use this model in the field of polyp segmentation and continue to improve its performance to achieve better results.

Author Contributions: Conceptualization, Y.W., W.K. and H.S.; methodology, Y.W., W.K. and H.S.; software, W.K. and H.S.; validation, Y.W., W.K. and H.S.; formal analysis, Y.W., W.K. and H.S.; investigation, Y.W.; resources, W.K. and H.S.; data curation, Y.W., W.K. and H.S.; writing—original draft preparation, Y.W., W.K. and H.S.; writing—review and editing, Y.W., W.K. and H.S.; visualization, Y.W., W.K. and H.S.; supervision, W.K. and H.S.; project administration, W.K. and H.S.; funding acquisition, W.K. and H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the National Key R & D Program of China (No. 2019YFB21 01600), the National Natural Science Foundation of China (No. 61872025), the Macao Science and Technology Development Fund (File No. 0122/2023/AMJ), and the Open Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE-2021ZX-03).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

References

1. Lachkar, A.; Gadi, T.; Benslimane, R.; D'orazio, L.; Martuscelli, E. Textile woven-fabric recognition by using Fourier image-analysis techniques: Part I: A fully automatic approach for crossed-points detection. *J. Text. Inst.* **2003**, *94*, 194–201. [CrossRef]
2. Papageorgiou, C.; Poggio, T. A trainable system for object detection. *Int. J. Comput. Vis.* **2000**, *38*, 15–33. [CrossRef]
3. Margolin, R.; Zelnik-Manor, L.; Tal, A. How to evaluate foreground maps? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014, pp. 248–255.
4. Wang, C.; Zuo, K.; Zhang, S.; Lei, H.; Hu, P.; Shen, Z.; Wang, R.; Zhao, P. PFNet: Large-Scale Traffic Forecasting with Progressive Spatio-Temporal Fusion. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 14580–14597. [CrossRef]
5. Ragland, K.; Tharcis, P. A survey on object detection, classification and tracking methods. *Int. J. Eng. Res. Technol.* **2014**, *3*, 622–628.
6. Wang, X.; Chen, S.; Wei, G.; Liu, J. TENet: Accurate light-field salient object detection with a transformer embedding network. *Image Vis. Comput.* **2023**, *129*, 104595. [CrossRef]
7. Chen, T.; Hu, X.; Xiao, J.; Zhang, G. BPFINet: Boundary-aware progressive feature integration network for salient object detection. *Neurocomputing* **2021**, *451*, 152–166. [CrossRef]
8. Yao, Z.; Wang, L. ERBANet: Enhancing region and boundary awareness for salient object detection. *Neurocomputing* **2021**, *448*, 152–167. [CrossRef]
9. Cui, W.; Zhang, Q.; Zuo, B. Deep saliency detection via spatial-wise dilated convolutional attention. *Neurocomputing* **2021**, *445*, 35–49. [CrossRef]
10. Kervadec, H.; Bouchtiba, J.; Desrosiers, C.; Granger, E.; Dolz, J.; Ayed, I.B. Boundary loss for highly unbalanced segmentation. In Proceedings of the International Conference on Medical Imaging with Deep Learning, London, UK, 8–10 July 2019; PMLR: Cambridge, MA, USA, 2019; pp. 285–296.
11. Ning, L.; Jincai, H.; Yanghe, F. Construction of multi-channel fusion salient object detection network based on gating mechanism and pooling network. *Multimed. Tools Appl.* **2022**, *81*, 12111–12126. [CrossRef]
12. Li, Z.; Lang, C.; Liew, J.H.; Li, Y.; Hou, Q.; Feng, J. Cross-layer feature pyramid network for salient object detection. *IEEE Trans. Image Process.* **2021**, *30*, 4587–4598. [CrossRef]
13. Ullah, I.; Jian, M.; Hussain, S.; Lian, L.; Ali, Z.; Qureshi, I.; Guo, J.; Yin, Y. Global context-aware multi-scale features aggregative network for salient object detection. *Neurocomputing* **2021**, *455*, 139–153. [CrossRef]
14. Ji, W.; Li, J.; Yu, S.; Zhang, M.; Piao, Y.; Yao, S.; Bi, Q.; Ma, K.; Zheng, Y.; Lu, H.; et al. Calibrated RGB-D salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9471–9481.
15. Chen, Q.; Liu, Z.; Zhang, Y.; Fu, K.; Zhao, Q.; Du, H. RGB-D salient object detection via 3D convolutional neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; Volume 35, pp. 1063–1071.
16. Chen, Z.; Xu, Q.; Cong, R.; Huang, Q. Global context-aware progressive aggregation network for salient object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10599–10606.
17. He, C.; Li, K.; Zhang, Y.; Tang, L.; Zhang, Y.; Guo, Z.; Li, X. Camouflaged object detection with feature decomposition and edge reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 22046–22055.
18. Lv, Y.; Zhang, J.; Dai, Y.; Li, A.; Barnes, N.; Fan, D.P. Towards deeper understanding of camouflaged object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 3462–3476. [CrossRef]

19. Lv, Y.; Zhang, J.; Dai, Y.; Li, A.; Liu, B.; Barnes, N.; Fan, D.P. Simultaneously localize, segment and rank the camouflaged objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11591–11601.
20. Chen, G.; Liu, S.J.; Sun, Y.J.; Ji, G.P.; Wu, Y.F.; Zhou, T. Camouflaged object detection via context-aware cross-level fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6981–6993. [CrossRef]
21. Bi, H.; Zhang, C.; Wang, K.; Tong, J.; Zheng, F. Rethinking camouflaged object detection: Models and datasets. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 5708–5724. [CrossRef]
22. Hu, H.; Bai, S.; Li, A.; Cui, J.; Wang, L. Dense relation distillation with context-aware aggregation for few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10185–10194.
23. Liu, H.; Peng, P.; Chen, T.; Wang, Q.; Yao, Y.; Hua, X.S. Fecanet: Boosting few-shot semantic segmentation with feature-enhanced context-aware network. *IEEE Trans. Multimed.* **2023**, *25*, 8580–8592. [CrossRef]
24. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; PMLR: Cambridge, MA, USA, 2019; pp. 6105–6114.
25. Liang, Y.; Qin, G.; Sun, M.; Wang, X.; Yan, J.; Zhang, Z. A systematic review of image-level camouflaged object detection with deep learning. *Neurocomputing* **2023**, *566*, 127050. [CrossRef]
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
27. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
28. Liu, X.; Zhang, T.T.; Zhou, J.; Xiang, B.; Liu, J.T. A novel 3-D lead-iodide polymer based on the linkage of rare binuclear [Pb 2 I] 3+ cations and anionic bis (pyrazinyl)-trizole bridges. *Dalton Trans.* **2021**, *50*, 4486–4489. [CrossRef]
29. Li, L.; Han, J.; Zhang, N.; Liu, N.; Khan, S.; Cholakkal, H.; Anwer, R.M.; Khan, F.S. Discriminative Co-Saliency and Background Mining Transformer for Co-Salient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7247–7256.
30. Hu, X.; Wang, S.; Qin, X.; Dai, H.; Ren, W.; Luo, D.; Tai, Y.; Shao, L. High-resolution iterative feedback network for camouflaged object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, London, UK, 30–31 May 2023; Volume 37, pp. 881–889.
31. Yang, F.; Zhai, Q.; Li, X.; Huang, R.; Luo, A.; Cheng, H.; Fan, D.P. Uncertainty-guided transformer reasoning for camouflaged object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4146–4155.
32. Yin, B.; Zhang, X.; Hou, Q.; Sun, B.Y.; Fan, D.P.; Van Gool, L. Camoformer: Masked separable attention for camouflaged object detection. *arXiv* **2022**, arXiv:2212.06570.
33. Kulkarni, S.V.; Khaparde, S.A. *Transformer Engineering: Design, Technology, and Diagnostics*; CRC Press: Boca Raton, FL, USA, 2017.
34. Shanmugam, K.S.; Dickey, F.M.; Green, J.A. An optimal frequency domain filter for edge detection in digital pictures. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *1*, 37–49. [CrossRef]
35. Xu, B. Identifying fabric structures with fast Fourier transform techniques. *Text. Res. J.* **1996**, *66*, 496–506.
36. Zhao, J.X.; Liu, J.J.; Fan, D.P.; Cao, Y.; Yang, J.; Cheng, M.M. EGNet: Edge guidance network for salient object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 8779–8788.
37. Wei, J.; Wang, S.; Huang, Q. F^3Net: Fusion, feedback and focus for salient object detection. In Proceedings of the AAAI Conference On Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34; pp. 12321–12328.
38. Skurowski, P.; Abdulameer, H.; Błaszczyk, J.; Depta, T.; Kornacki, A.; Kozieł, P. Animal camouflage analysis: Chameleon database, 2017. *Unpubl. Manuscr.* **2018**, *2*, 7.
39. Le, T.N.; Nguyen, T.V.; Nie, Z.; Tran, M.T.; Sugimoto, A. Anabranch network for camouflaged object segmentation. *Comput. Vis. Image Underst.* **2019**, *184*, 45–56. [CrossRef]
40. Fan, D.P.; Ji, G.P.; Sun, G.; Cheng, M.M.; Shen, J.; Shao, L. Camouflaged object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2777–2787.
41. Fan, D.; Cheng, M.; Liu, Y.; Li, T.; Borji, A. A new way to evaluate foreground maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Volume 245484557.
42. Fan, D.P.; Ji, G.P.; Qin, X.; Cheng, M.M. Cognitive vision inspired object segmentation metric and loss function. *Sci. Sin. Informationis* **2021**, *6*, 6.
43. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.P.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst* **2019**, *32*, 8026.
44. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
45. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

46. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
47. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
48. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. Hybrid task cascade for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4974–4983.
49. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask scoring r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 6409–6418.
50. Hu, X.; Zhu, L.; Fu, C.W.; Qin, J.; Heng, P.A. Direction-aware spatial context features for shadow detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7454–7462.
51. Zhu, L.; Deng, Z.; Hu, X.; Fu, C.W.; Xu, X.; Qin, J.; Heng, P.A. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 121–136.
52. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; Proceedings 4; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
53. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 263–273.
54. 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.
55. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M. Basnet: Boundary-aware salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–21 June 2019; pp. 7479–7489.
56. Wu, Z.; Su, L.; Huang, Q. Cascaded partial decoder for fast and accurate salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–21 June 2019; pp. 3907–3916.
57. Zhao, T.; Wu, X. Pyramid feature attention network for saliency detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–21 June 2019; pp. 3085–3094.
58. Pang, Y.; Zhao, X.; Zhang, L.; Lu, H. Multi-scale interactive network for salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9413–9422.
59. Zhang, J.; Shao, J.; Chen, J.; Yang, D.; Liang, B.; Liang, R. PFNet: An unsupervised deep network for polarization image fusion. *Opt. Lett.* **2020**, *45*, 1507–1510. [CrossRef] [PubMed]
60. Chen, T.; Xiao, J.; Hu, X.; Zhang, G.; Wang, S. Boundary-guided network for camouflaged object detection. *Knowl. Based Syst.* **2022**, *248*, 108901. [CrossRef]
61. Ji, G.P.; Zhu, L.; Zhuge, M.; Fu, K. Fast camouflaged object detection via edge-based reversible re-calibration network. *Pattern Recognit.* **2022**, *123*, 108414. [CrossRef]
62. Yan, X.; Sun, M.; Han, Y.; Wang, Z. Camouflaged object segmentation based on matching–recognition–refinement network. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**. [CrossRef] [PubMed]
63. Xiao, J.; Chen, T.; Hu, X.; Zhang, G.; Wang, S. Boundary-guided context-aware network for camouflaged object detection. *Neural Comput. Appl.* **2023**, *35*, 15075–15093. [CrossRef]