sensors

MDPI

*Article*

# Swin Transformer-Based Edge Guidance Network for RGB-D Salient Object Detection

**Shuaihui Wang, Fengyi Jiang and Boqian Xu ***

Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; wangshuaihui@ciomp.ac.cn (S.W.); jiangfengyi@ciomp.ac.cn (F.J.)
* Correspondence: xuboqian@ciomp.ac.cn

**Abstract:** Salient object detection (SOD), which is used to identify the most distinctive object in a given scene, plays an important role in computer vision tasks. Most existing RGB-D SOD methods employ a CNN-based network as the backbone to extract features from RGB and depth images; however, the inherent locality of a CNN-based network limits the performance of CNN-based methods. To tackle this issue, we propose a novel Swin Transformer-based edge guidance network (SwinEGNet) for RGB-D SOD in which the Swin Transformer is employed as a powerful feature extractor to capture the global context. An edge-guided cross-modal interaction module is proposed to effectively enhance and fuse features. In particular, we employed the Swin Transformer as the backbone to extract features from RGB images and depth maps. Then, we introduced the edge extraction module (EEM) to extract edge features and the depth enhancement module (DEM) to enhance depth features. Additionally, a cross-modal interaction module (CIM) was used to integrate cross-modal features from global and local contexts. Finally, we employed a cascaded decoder to refine the prediction map in a coarse-to-fine manner. Extensive experiments demonstrated that our SwinEGNet achieved the best performance on the LFSD, NLPR, DES, and NJU2K datasets and achieved comparable performance on the STEREO dataset compared to 14 state-of-the-art methods. Our model achieved better performance compared to SwinNet, with 88.4% parameters and 77.2% FLOPs. Our code will be publicly available.

**Keywords:** RGB-D salient object detection; edge guidance; transformer; cross-modal interaction



**Citation:** Wang, S.; Jiang, F.; Xu, B. Swin Transformer-Based Edge Guidance Network for RGB-D Salient Object Detection. *Sensors* **2023**, *23*, 8802. https://doi.org/10.3390/s23218802

Academic Editors: Man Qi and Matteo Dunnhofer

Received: 31 August 2023
Revised: 9 October 2023
Accepted: 24 October 2023
Published: 29 October 2023



## 1. Introduction

Salient object detection (SOD) is an important preprocessing method in computer vision tasks, with applications in video detection and segmentation [1], semantic segmentation [2], object tracking [3], etc.

CNN-based models for RGB SOD have yielded great performance in localizing salient objects [4–8]. However, it is still difficult to localize the salient object accurately in scenes such as those with low contrast or objects with a cluttered background. CNN-based RGB-D SOD models, which employ features from RGB images and depth maps, have attracted growing interest and presented promising performance [9–23]. However, some issues still limit the performance of existing CNN-based RGB-D SOD models.

The first issue is that CNN-based models cannot effectively capture long-range dependencies. Long-range semantic information plays an important role in identifying and locating salient objects [24]. Due to the intrinsic locality of convolution operations, CNN-based models cannot effectively extract global context information. In addition, the empirical receptive field of CNN is much smaller than the theoretical receptive field, especially on high-level layers [25].

The second issue is that depth maps are often noisy. The performance of RGB-D SOD models relies on reliable RGB images and depth maps. Misleading information in depth maps degrades the performance of RGB-D SOD models.

Global context information helps reduce errors created via poor depth maps. Transformers can extract features and model long-range dependencies, and Transformer-based methods have achieved outstanding performance in various computer vision tasks [26–29]. However, Transformers are less effective in capturing local features. The Swin Transformer [29], combining the advantages of Transformers and CNN, has been shown to have a powerful feature extraction ability. Considering the above challenges, the Swin Transformer is suitable as a feature extractor for RGB-D SOD tasks.

Swin Transformer-based models are relatively weak in their ability to model local context information. Therefore, Swin Transformer-based models should pay more attention to local feature information.

Based on the investigation above, we propose a novel Swin Transformer-based edge guidance network (SwinEGNet) that enhances feature locality to boost the performance of RGB-D SOD. We employed the Swin Transformer as the backbone to extract features from RGB images and depth maps for capturing long-range dependencies. We introduced a depth enhancement module (DEM) and a cross-modal interaction module to enhance local features. Unlike other methods, we employed edge clues to enhance depth features rather than edge clues as decoder guidance to directly refine the final prediction map. We designed the edge extraction module (EEM) to extract edge information and the depth enhancement module (DEM) to enhance depth features. Furthermore, we used a cross-modal interaction module to effectively integrate information from global and local contexts. To effectively explore the features of each layer, we employed a cascaded decoder to progressively refine our saliency maps.

Our main contributions are summarized as follows:

- A novel edge extraction module (EEM) is proposed, which generates edge features from the depth features.
- A newly designed edge-guided cross-modal interaction was employed to effectively integrate cross-modal features, where the depth enhancement module was employed to enhance the depth feature and the cross-modal interaction module was employed to encourage cross-modal interaction from global and local aspects.
- A novel Swin Transformer-based edge guidance network (SwinEGNet) for RGB-D SOD is proposed. The proposed SwinEGNet was evaluated with four evaluation metrics and compared to 14 state-of-the-art (SOTA) RGB-D SOD methods on six public datasets. Our model achieved better performance with less parameters and FLOPs than SwinNet, as shown in Figure 1. In addition, a comprehensive ablation experiment was also conducted to verify the effectiveness of the proposed modules. The experiment results showed the outstanding performance of our proposed method.

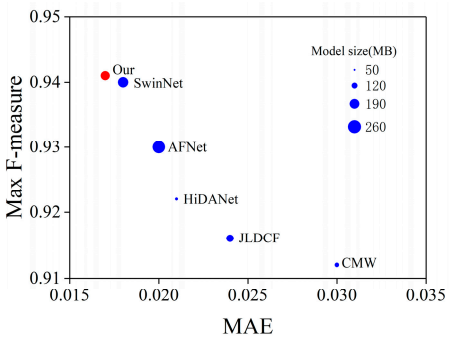


**Figure 1.** Max F-measure, MAE, and model size of different methods on the NLPR dataset. Our model achieves better performance with a smaller model size.

The remainder of this paper is structured as follows: The current status of RGB-D salient object detection is presented in Section 2. The overall architecture, detailed structure, and loss function of the proposed network are outlined in Section 3. The results of our experiments are provided in Section 4. Finally, our conclusions are presented in Section 5.

## 2. Related Work

CNN-based RGB-D salient object detection: Benefitting from the development of deep learning and depth sensors, many CNN-based RGB-D SOD methods have recently been proposed. Compared to RGB SOD methods, RGB-D SOD models employ depth clues as complementary information and have shown outstanding performance in salient object detection. Most RGB-D SOD models adopt CNN-based networks to extract features and focus on cross-modal fusion strategies to improve salient object detection performance. Various frameworks and fusion strategies have been proposed to effectively merge cross-modal cross-scale features [14,17,21–23,30,31]. Zhang et al. [30] designed an asymmetric two-stream network, where a flow ladder module is introduced to the RGB stream to capture global context information and DepthNet for the depth stream. Zhang et al. [17] proposed a multistage cascaded learning framework for RGB-D saliency detection, which minimizes the mutual information between RGB images and depth maps to model complementary information. Chen et al. [22] designed a triplet encoder network that processes RGB, depth, and fused features separately to suppress the background noise in the depth map and sharpen the boundaries of high-level features. Li et al. [14] designed a hierarchical alternate interaction module that progressively and hierarchically integrates local and global contexts. Wu et al. [21] proposed layer-wise, trident spatial, and attention mechanisms to fuse robust RGB and depth features against low-quality depths. Wu et al. [23] employed a granularity-based attention module to leverage the details of salient objects and introduced a dual-attention module to fuse the cross-modal cross-scale features in a coarse-to-fine manner.

To address the insufficiency of obtaining global semantic information of CNN-based networks, Liu et al. [7] proposed using a receptive field block to enhance feature discriminability and robustness by enlarging the receptive field. Dilated convolutions can enlarge the receptive field of CNN without loss of resolution. As a result, Yu et al. [32] presented modules based on dilated convolutions to aggregate multiscale information. Liu et al. [8] designed a global guidance module for RGB SOD that utilizes the revised pyramid pooling module to capture global semantic information.

Transformer-based RGB-D salient object detection: The Transformer was first employed for machine translation and gradually introduced in computer vision tasks. Dosovitskiy et al. [26] proposed the first Vision Transformer (ViT), Wang et al. [28] proposed a progressive shrinking pyramid Transformer (PVT), and Liu et al. [29] designed the Swin Transformer. Subsequently, researchers employed the Transformer as the backbone network to improve the detection performance of RGB-D SOD. Liu et al. [33] developed a unified model based on ViT for both RGB and RGB-D SOD. Zeng et al. [34] employed the Swin Transformer as the encoding backbone to extract features from RGB images and depth maps. Liu et al. [35] employed PVT as a powerful feature extractor to extract global context information and designed a lightweight CNN-based backbone to extract spatial structure information in depth maps. Pang et al. [36] proposed using a novel top-down information propagation path based on the Transformer to capture important global clues to promote cross-modal feature fusion. Liu et al. [37] proposed using a cross-modal fusion network based on SwinNet for RGB-D and RGB-T SOD. Roy et al. [38] employed the Swin Transformer as the encoder block to detect multiscale objects.

## 3. Methodologies

In this section, we present the proposed Swin Transformer-based edge guidance network (SwinEGNet). We provide an overview of our method and describe its main components in detail, including the feature encoder, edge extraction module, edge-guided cross-modal interaction module, cascaded decoder, and loss function.

### *3.1. The Overall Architecture*

As illustrated in Figure 2, we present a Swin Transformer-based edge guidance network (SwinEGNet). Inspired by [37], we employed edge clues to guide salient object detection. However, unlike [37], edge clues were incorporated into cross-modal interaction blocks

to enhance depth features rather than being employed as decoder guidance to refine the final prediction map. The proposed SwinEGNet adopts the encoder–decoder structure. As shown in Figure 2, SwinEGNet consists of a feature encoder, edge extraction module (EEM), edge-guided cross-modal interaction module (EGCIM), and cascaded decoder. Firstly, RGB images and depth maps are fed into two independent Swin Transformers for feature extraction, and an EEM is proposed to extract edge features. Then, these features are fed into EGCIM for depth feature enhancement and feature fusion, where the depth enhancement module (DEM) is responsible for depth feature enhancement and the cross-modal interaction module (CIM) is responsible for feature fusion. Finally, the fused features are fed into the decoder block for saliency maps. The cascaded decoder was employed to effectively explore the features of the four layers and progressively refine the saliency maps.

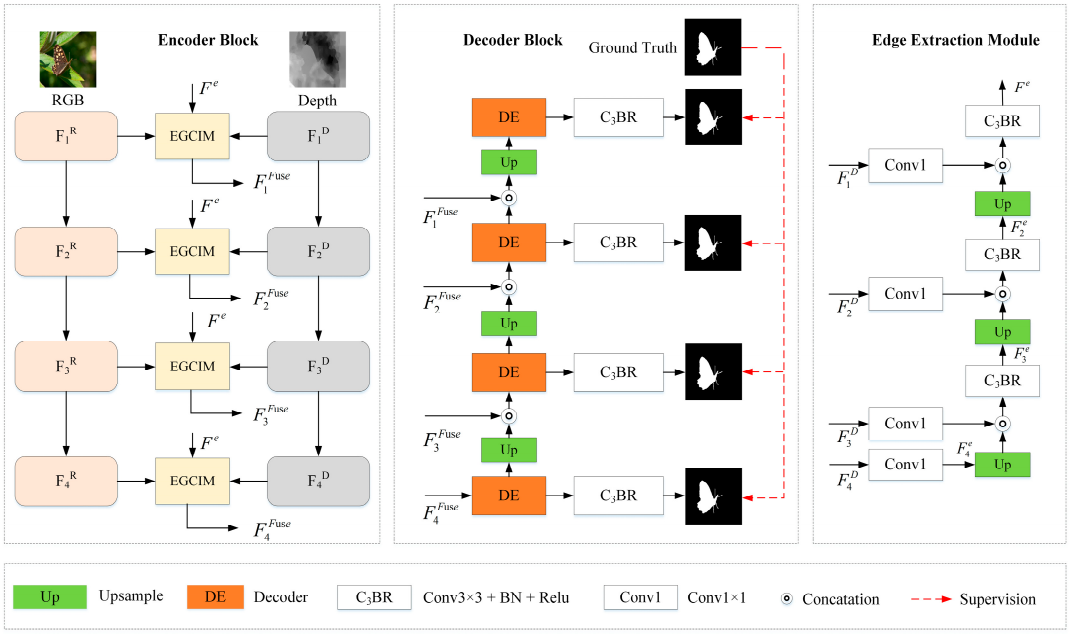


**Figure 2.** An overview of the proposed SwinEGNet. It consists of a feature encoder, an edge extraction module (EEM), an edge-guided cross-modal interaction module (EGCIM), and a cascaded decoder.

### *3.2. Feature Encoder*

In contrast to other Transformers, the Swin Transformer computes multihead self-attention within a local window instead of the whole input to model locality relationships. Furthermore, it employs a shifted window operation to model long-range dependence across windows. Therefore, the Swin Transformer is suitable for feature extraction because it incorporates the merits of the Transformer and CNN. Considering the performance and computational complexity, we adopted the Swin-B Transformer as the backbone to extract features from RGB images and depth maps, which accept an input size of 384 × 384.

RGB images and depth maps are fed into two independent Swin Transformers for feature extraction. Considering the first layer contains redundant noisy information, the extracted features of the last four layers are employed for feature fusion. The features can be expressed as follows:

$$F_i^R = trans(I_R), i = 1, 2, 3, 4 \quad (1)$$

$$F_i^D = trans(I_D), i = 1, 2, 3, 4 \quad (2)$$

where $F_i^R$ denotes the RGB feature; $F_i^D$ denotes the depth feature, $trans(\cdot)$ denotes the Transformer; and $I_R$ and $I_D$ denote the input RGB image and depth image, respectively.

### *3.3. Edge Extraction Module*

To extract edge features, we propose an edge extraction module (EEM). The extracted edge features are fed into EGCIM to enhance the depth feature. The details of the proposed EEM are illustrated in Figure 3.

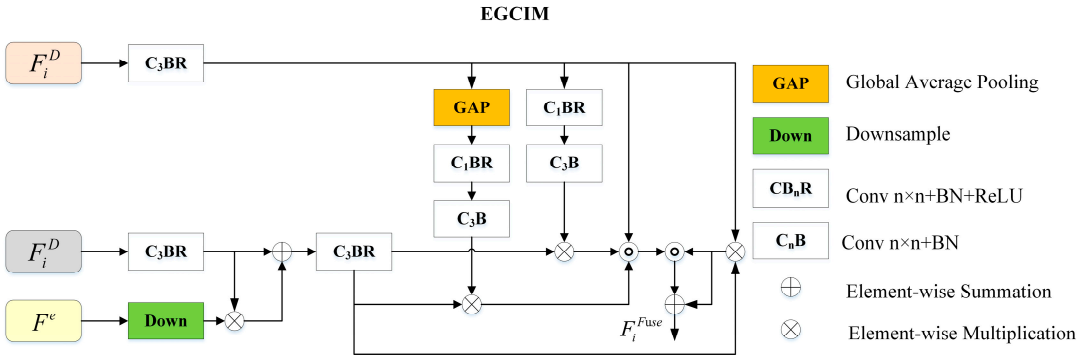


**Figure 3.** Details of the proposed edge-guided cross-modal interaction module (EGCIM).

Shallow layers contain low-level information such as structure clues, while deep layers contain global semantic information. They are all helpful in extracting edge information. In contrast to other methods that employ parts of the depth features for edge prediction, we employed all depth features for edge extraction, and the edge features were progressively refined in a coarse-to-fine manner.

In particular, the depth features $F_i^D (i = 1, 2, 3, 4)$ are fed into a $1 \times 1$ convolutional layer for channel reduction. Then, features $F_i^D (i = 2, 3, 4)$ perform the upsample operation to generate the same size features as $F_{i+1}^D$. The edge feature $F_4^e$ can be expressed as follows:

$$F_4^e = Up\left(Conv_1(F_4^D)\right) \tag{3}$$

where $Up(\cdot)$ denotes the upsample operation.

Next, the edge feature performs a concatenation operation and a $3 \times 3$ convolutional layer with a BatchNorm and a ReLU function to generate the edge feature $F_{i-1}^e$, which can be expressed as follows:

$$F_i^e = C_3BR\left(Cat\left(Conv_1(F_i^D), F_{i+1}^e\right)\right), i = 1, 2, 3 \tag{4}$$

where $C_3BR(\cdot)$ denotes a $3 \times 3$ convolutional layer with a BatchNorm and a ReLU function, and $Cat(\cdot)$ denotes concatenation operation. The edge feature $F_1^e$ is the final edge feature $F^e$. The final edge feature $F^e$ will be fed into EGCIM for depth enhancement.

### *3.4. Edge-Guided Cross-modal Interaction Module*

To enhance depth features and encourage cross-modal feature interaction, we designed an edge-guided cross-modal interaction module (EGCIM) to integrate features from both modalities, including a depth enhancement module (DEM) and a cross-modal interaction module (CIM).

Depth enhancement module: Though Transformer-based methods sufficiently capture global context information, they are relatively weak at capturing local context information compared to CNN-based methods. Therefore, it is necessary to utilize local clues like edge information to enhance the depth features. We designed a depth enhancement module (DEM) to enhance the depth features, which introduces edge information extracted from the depth features to these features for depth enhancement. The detailed structure of DEM is shown in Figure 3.

The depth features $F_i^D$ and edge features $F^e$ at a certain hierarchy $i = 1, 2, 3, 4$, $F_i^D$ performs the convolution operation with a BatchNorm and a ReLU function for channel

reduction, and $F^e$ performs the downsample operation to gain the same size as $F_i^D$. Then, the depth features $F_i^D$ and edge features of the same size are fused using multiplication and addition operations. The enhanced depth features can be expressed as follows:

$$F_i^{DE} = C_3BR\Big(C_3BR(F_i^D) + C_3BR(F_i^D) \times Down(F^e)\Big) \tag{5}$$

where + denotes the addition operation, and $Down(\cdot)$ denotes the downsample operation. The enhanced depth features $F_i^{DE}$ will be fed into CIM for feature fusion.

Cross-modal interaction module: We used a cross-modal interaction module (CIM) to effectively combine RGB and depth modalities. The CIM contains a global attention branch and a local attention branch to enhance globality and locality. In addition, a residual connection is adopted to combine the fused features with RGB features for the preservation of the RGB images' original information. The local information of the depth features enhances the RGB features to sharpen the details of salient objects, and the global context information of the depth features enhances the RGB features to locate the salient object.

As shown in Figure 3, the RGB features are fed into a $3 \times 3$ convolutional layer with a BatchNorm and a ReLU activation function for channel reduction. There are three branches for feature fusion: the first branch employs global average pooling (GAP) to capture global context information, the second branch employs $1 \times 1$ convolution to obtain local information, and the third branch aims to keep the original information of RGB features. Then, we carry out multiplication, concatenation, and addition operations for fusion. The fused features can be expressed as follows:

$$F_i^{Fuse} = C_3BR\Big(C_3BR\Big(Cat\Big(C_3BR(F_i^R), F_i^g, F_i^l, F_i^o\Big)\Big) + F_i^o\Big) \tag{6}$$

$$F_i^{Fuse} = C_3BR(F_i^R) \times F_i^{DE} \tag{7}$$

$$F_i^g = C_3BR(F_i^R) \times C_1B\Big(C_1BR\Big(GAP(F_i^{DE})\Big)\Big) \tag{8}$$

$$F_i^l = C_3BR(F_i^R) \times C_1B\Big(C_1BR(F_i^{DE})\Big) \tag{9}$$

where $GAP(\cdot)$ represents the global average pooling operation, $C_1B$ represents a convolution operation with a BatchNorm function, $C_1BR$ represents a convolution operation with a BatchNorm function and a ReLU function, and $F_i^{Fuse}$ represents the fused features.

### *3.5. Cascaded Decoder*

The cascaded encoder can effectively leverage the multilevel features and eliminate the noise in low-level features, which improves the accuracy of salient maps. Moreover, deep-layer supervision performs better than single supervision [13]. Therefore, we employed a cascaded decoder for the final prediction map, as shown in Figure 3. The decoder has four decoding levels corresponding to the four-level cross-modal feature interaction. Consequently, the prediction map is refined progressively. Each decoder contains two $3 \times 3$ convolution layers with a BatchNorm and a ReLU function, a dropout layer, and an upsample layer. The initial prediction map $S_4$ is fed into the decoder and concatenates with the previous prediction map $S_{n-1}$ for refinement. The prediction features $S_i$ can be donated as follows:

$$S_i = \begin{cases} C_3BR(Up(S_{i+1}), S_i), i = 1, 2, 3 \\ C_3BR(F_i^{Fuse}), i = 4 \end{cases} \tag{10}$$

where $D(\cdot)$ represents the decoder operation, $S_n$ represents the prediction map, and $Up(\cdot)$ represents the upsample operation. Next, features $S_i$ perform convolution operations to obtain the prediction map, and $S_1$ is the final prediction map.

### *3.6. Loss Function*

Detection loss is composed of the weighted binary cross-entropy (BCE) loss $L^{\omega}_{BCE}$ and the weighted intersection-over-union (IOU) loss $L^{\omega}_{IoU}$ [39], which has been invalidated in salient object detection. The detection loss can be formulated as follows:

$$L_d = L^{\omega}_{BCE} + L^{\omega}_{IoU} \tag{11}$$

$L^{\omega}_{IoU}$ and $L^{\omega}_{BCE}$ pay more attention to the structure of SOD and the hard pixels to highlight the importance of the hard pixel. As illustrated in Figure 2, four-level supervisions are applied to supervise the four side-output maps. Each map $S_i$ is upsampled to the same size as the ground truth map. Thus, the total loss function L can be expressed as follows:

$$L = \sum_{i=1}^{4} (L^{i}_{d}(S_i, G) \tag{12}$$

## 4. Experiments

### *4.1. Datasets and Evaluation Metrics*

Datasets: We evaluated the proposed method on six widely used benchmark datasets: STEREO (1000 image pairs) [40], NJU2K (2003 image pairs) [41], NLPR (1000 image pairs) [42], LFSD (100 image pairs) [43], SIP (929 image pairs) [44], and DES (135 image pairs) [45]. For a fair comparison, our training settings were the same as the existing works [12], which consisted of 1485 samples from the NJU2K dataset and 700 samples from the NLPR dataset. The remaining images from NLPR, DES, and NJU2K, and the whole SIP, STEREO, and LFSD were used for testing.

Evaluation metrics: We adopted four widely used evaluation metrics for quantitative evaluation, including S-measure ($S_\alpha$, $\alpha = 0.5$) [46], maximum F-measure ($F_m$) [47], maximum E-measure ($E_m$) [48], and mean absolute error (MAE, $M$) [49]. S-measure evaluates the structural similarity between the saliency map and ground truth, which is defined as follows:

$$S = \alpha S_o + (1 - \alpha) S_r \tag{13}$$

where $\alpha$ is a trade-off parameter set to 0.5, $S_o$ represents the object perception, and $S_r$ represents the regional perception. F-measure focuses on region-based similarity that considers precision and recall, which is defined as follows:

$$F_\beta = \left(1 + \beta^2\right) \frac{P \times R}{\beta^2 \times P + R} \tag{14}$$

where $P$ denotes precision, $R$ denotes recall, and $\beta^2$ is a trade-off parameter set to 0.3. We used the maximum F-measure as the evaluation metric. MAE assesses the average absolute error at the pixel level, which is defined as follows:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S(i,j) - G(i,j)| \tag{15}$$

where $W$ and $H$ represent the width and height of the image, respectively. $S$ represents the saliency maps, and $G$ represents the ground truth. E-measure is employed to capture image-level statistics and local pixel matching, which is defined as follows:

$$E_m = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi_{FM}(i,j) \tag{16}$$

where $\phi_{FM}$ represents the enhanced alignment matrix. For a fair comparison, we used the evaluation tools provided by [15].

### *4.2. Implementation Details*

We implemented our model on PyTorch with one NVIDIA A4000 GPU. The Swin Transformer that has been pretrained on ImageNet was employed as our backbone network. The parameters of the Swin-B model were initialized with the pretrained parameters, and the remaining parameters were initialized with PyTorch default settings. The Adam optimizer was employed to train the proposed model with a batch size of 5, a momentum of 0.9, and a weight decay of 0.1. The initial learning rate was $1 \times 10^{-4}$, which was then divided by 10 for every 60 epochs. All images were resized to $384 \times 384$ for training and testing. The single-channel depth image was replicated to a three-channel image, which was the same as the RGB image. Data augment strategies, including random flipping, rotating, and border clipping, were employed to augment the training data. The model was trained for 120 epochs.

### *4.3. Comparison with SOTAs*

Quantitative comparison: We compared the proposed network with 14 SOTA CNN-based methods and Transformer-based methods, which were CMW [13], JLDCF [50], HINet [51], DSA2F [20], CFIDNet [52], $C^2$DFNet [53], SPSNet [19], AFNet [22], HiDANet [23], MTFormer [54], VST [43], TANet [35], and SwinNet [37]. The compared saliency maps were directly provided by the authors or generated via their released codes. The quantitative comparison under four evaluation metrics on six datasets is shown in Table 1. As shown in Table 1, our SwinEGNet performed the best on LFSD, NLPR, and DES datasets and competitively performed on NJU2K, STEREO, and SIP datasets. In particular, SwinEGNet performed outstandingly on the LFSD dataset, which is considered a challenging dataset. Compared to the second model DSA2F, the improvements of S-measure, F-measure, E-measure, and MAE were about 0.011, 0.006, 0.005, and 0.002, respectively. On the NJU2K dataset, the performance of our method was comparable with SwinNet. On the STEREO dataset, our method performed the best in $E_m$.

**Table 1.** Quantitative comparison of SOTA methods under four evaluation metrics: S-measure ($S_a$), max F-measure ($F_m$), max E-measure ($E_m$), and MAE ($M$). ↑ denotes that higher is better, and ↓ denotes that lower is better. The best two results are shown in red and green fonts, respectively.

| | Metric | CMW | JLDCF | HINet | HAINet | DSA2F | CFIDNet | $C^2$DFNet | SPSNet | AFNet | HiDANet | MTFormer | VST | TANet | SwinNet | Our |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LFSD | $S_m\uparrow$ | 0.876 | 0.854 | 0.852 | 0.854 | 0.882 | 0.869 | 0.863 | - | 0.89 | - | 0.872 | 0.89 | 0.875 | 0.886 | 0.893 |
| | $F_m\uparrow$ | 0.899 | 0.862 | 0.872 | 0.877 | 0.903 | 0.883 | 0.89 | - | 0.9 | - | 0.879 | 0.903 | 0.892 | 0.903 | 0.909 |
| | $E_m\uparrow$ | 0.901 | 0.893 | 0.88 | 0.882 | 0.920 | 0.897 | 0.899 | - | 0.917 | - | 0.911 | 0.918 | - | 0.914 | 0.925 |
| | $M\downarrow$ | 0.067 | 0.078 | 0.076 | 0.08 | 0.054 | 0.07 | 0.065 | - | 0.056 | - | 0.062 | 0.054 | 0.059 | 0.059 | 0.052 |
| NLPR | $S_m\uparrow$ | 0.917 | 0.925 | 0.922 | 0.924 | 0.918 | 0.922 | 0.928 | 0.923 | 0.936 | 0.93 | 0.932 | 0.931 | 0.935 | 0.941 | 0.941 |
| | $F_m\uparrow$ | 0.912 | 0.916 | 0.915 | 0.922 | 0.917 | 0.914 | 0.926 | 0.918 | 0.93 | 0.929 | 0.925 | 0.927 | 0.943 | 0.94 | 0.941 |
| | $E_m\uparrow$ | 0.94 | 0.962 | 0.949 | 0.956 | 0.95 | 0.95 | 0.957 | 0.956 | 0.961 | 0.961 | 0.965 | 0.954 | - | 0.968 | 0.969 |
| | $M\downarrow$ | 0.03 | 0.022 | 0.026 | 0.024 | 0.024 | 0.026 | 0.021 | 0.024 | 0.02 | 0.021 | 0.021 | A0.024 | 0.018 | 0.018 | 0.017 |
| NJU2K | $S_m\uparrow$ | 0.903 | 0.903 | 0.915 | 0.912 | 0.904 | 0.914 | 0.908 | 0.918 | 0.926 | 0.926 | 0.922 | 0.922 | 0.927 | 0.935 | 0.931 |
| | $F_m\uparrow$ | 0.913 | 0.903 | 0.925 | 0.925 | 0.916 | 0.923 | 0.918 | 0.927 | 0.933 | 0.939 | 0.923 | 0.926 | 0.941 | 0.943 | 0.938 |
| | $E_m\uparrow$ | 0.925 | 0.944 | 0.936 | 0.94 | 0.935 | 0.938 | 0.937 | 0.949 | 0.95 | 0.954 | 0.954 | 0.942 | - | 0.957 | 0.958 |
| | $M\downarrow$ | 0.046 | 0.043 | 0.038 | 0.038 | 0.039 | 0.038 | 0.039 | 0.033 | 0.032 | 0.029 | 0.032 | 0.036 | 0.027 | 0.027 | 0.026 |
| STEREO | $S_m\uparrow$ | 0.913 | 0.903 | 0.892 | 0.915 | 0.898 | 0.91 | 0.911 | 0.914 | 0.918 | 0.911 | 0.908 | 0.913 | 0.923 | 0.919 | 0.919 |
| | $F_m\uparrow$ | 0.909 | 0.903 | 0.897 | 0.914 | 0.91 | 0.906 | 0.91 | 0.908 | 0.923 | 0.921 | 0.908 | 0.915 | 0.934 | 0.926 | 0.926 |
| | $E_m\uparrow$ | 0.93 | 0.944 | 0.92 | 0.938 | 0.939 | 0.935 | 0.938 | 0.941 | 0.949 | 0.946 | 0.947 | 0.939 | - | 0.947 | 0.951 |
| | $M\downarrow$ | 0.042 | 0.043 | 0.048 | 0.039 | 0.039 | 0.042 | 0.037 | 0.035 | 0.034 | 0.035 | 0.038 | 0.038 | 0.027 | 0.033 | 0.031 |
| DES | $S_m\uparrow$ | 0.937 | 0.929 | 0.927 | 0.939 | 0.917 | 0.92 | 0.924 | 0.94 | 0.925 | 0.946 | - | 0.946 | - | 0.945 | 0.947 |
| | $F_m\uparrow$ | 0.943 | 0.919 | 0.937 | 0.949 | 0.929 | 0.937 | 0.937 | 0.944 | 0.938 | 0.952 | - | 0.949 | - | 0.952 | 0.956 |
| | $E_m\uparrow$ | 0.961 | 0.968 | 0.953 | 0.971 | 0.955 | 0.938 | 0.953 | 0.974 | 0.946 | 0.98 | - | 0.971 | - | 0.973 | 0.98 |
| | $M\downarrow$ | 0.021 | 0.022 | 0.22 | 0.017 | 0.023 | 0.022 | 0.018 | 0.015 | 0.022 | 0.013 | - | 0.017 | - | 0.016 | 0.014 |
| SIP | $S_m\uparrow$ | 0.867 | 0.879 | 0.856 | 0.879 | 0.861 | 0.881 | 0.871 | 0.892 | 0.896 | 0.892 | 0.894 | 0.903 | 0.893 | 0.911 | 0.9 |
| | $F_m\uparrow$ | 0.889 | 0.885 | 0.88 | 0.906 | 0.891 | 0.9 | 0.895 | 0.91 | 0.919 | 0.919 | 0.902 | 0.924 | 0.922 | 0.936 | 0.93 |
| | $E_m\uparrow$ | 0.9 | 0.923 | 0.888 | 0.916 | 0.909 | 0.918 | 0.913 | 0.931 | 0.931 | 0.927 | 0.932 | 0.935 | - | 0.944 | 0.935 |
| | $M\downarrow$ | 0.063 | 0.051 | 0.066 | 0.053 | 0.057 | 0.051 | 0.052 | 0.044 | 0.043 | 0.043 | 0.043 | 0.041 | 0.041 | 0.035 | 0.04 |

Qualitative comparison: We qualitatively compared seven representative methods on challenging scenes. The first scene had a similar foreground and background (first row), the second scene had poor depth map (second row and third row), the third scene had a complex background (fourth row and fifth row), the fourth scene had a small object (sixth row), the fifth scene had multiple objects (seventh row and eighth row), and the sixth scene had a fine structure (ninth row). As shown in Figure 4, our method obtained the best detection results. For the first scene, the foreground and background of the RGB image

were similar, but the depth map provided correct information. Our method located salient objects better than other methods thanks to the power of EEM and EGCIM. For the second scene, though the depth map provided incorrect information, our method successfully located salient objects by eliminating misleading information of the poor depth map. For the fourth scene, our method fused the RGB feature and depth feature the best. For the fifth scene, our method not only located the salient objects but also maintained the sharp boundaries. These all indicate the effectiveness of our model.


**Figure 4.** Visual comparison of our method and seven SOTAS, including CMW, DSA2F, CFIDNet, $C^2$DFNet, AFNet, VST, and SwinNet.

### *4.4. Ablation Study*

We conducted comprehensive ablation studies on LFSD and STEREO datasets to evaluate the effectiveness of the proposed modules in our proposed model.

Effectiveness of Swin Transformer backbone: We replaced the feature encoder with ResNet50 to verify the effectiveness of the Swin Transformer backbone. As shown in Table 2, the Transformer-based model showed better performance in all the evaluation benchmarks and metrics, especially on the LFSD dataset. We show the visual comparison of ResNet50 and Swin Transformer in Figure 5. The ResNet50 was inferior to the Swin Transformer. This validates the effectiveness of the Swin Transformer backbone for the RGB-D SOD.

**Table 2.** Effective analysis of the proposed modules on two datasets. The best results are shown in bold.

| Models | LFSD | | | | STEREO | | | |
|---|---|---|---|---|---|---|---|---|
| | $M\downarrow$ | $S_m\uparrow$ | $F_m\uparrow$ | $E_m\uparrow$ | $M\downarrow$ | $S_m\uparrow$ | $F_m\uparrow$ | $E_m\uparrow$ |
| **Ours** | **0.052** | **0.893** | **0.909** | **0.925** | **0.031** | **0.919** | **0.926** | **0.951** |
| ResNet50 | 0.084 | 0.835 | 0.864 | 0.868 | 0.044 | 0.893 | 0.0.9 | 0.927 |
| w/o EGCIM | 0.067 | 0.87 | 0.887 | 0.902 | 0.035 | 0.913 | 0.922 | 0.946 |
| w/o DEM | 0.064 | 0.875 | 0.893 | 0.906 | 0.032 | 0.917 | 0.925 | 0.949 |
| w/o CIM | 0.066 | 0.869 | 0.887 | 0.901 | 0.033 | 0.914 | 0.923 | 0.947 |

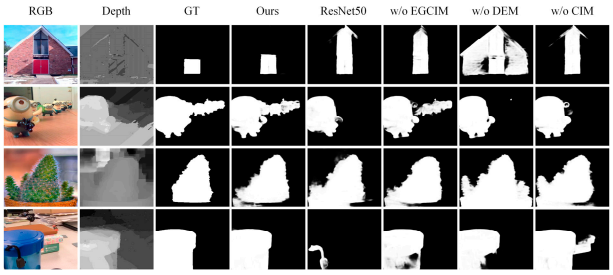


**Figure 5.** Visual comparison of the ablation study.

Effectiveness of EGCIM: To explore the effectiveness of EGCIM, we replaced EGCIM with a multiplication operation. In Table 2, we quantitatively demonstrate the contribution of the EGCIM. The performance of our model degraded without the help of EGCIM. This validates the effectiveness of the edge-guided cross-modal interaction module.

Effectiveness of DEM in EGCIM: To verify the effectiveness of DEM in EGCIM, we removed DEM from our full model. In Table 2, we quantitatively demonstrate the contribution of DEM. As shown in Table 2, the depth enhancement module improved the performance of the proposed model, especially on the LFSD dataset. The MAE, S-measure, F-measure, and E-measure are improved by about 0.012, 0.018, 0.013, and 0.009 in the LFSD dataset, respectively.

Effectiveness of CIM in EGCIM: We replaced CIM with a multiplication operation to verify the effectiveness of CIM in EGCIM. In Table 2, we quantitatively demonstrate the contribution of CIM. As shown in Table 2, the performance degradation caused by removing CIM supports our claim that the cross-modal interaction module can effectively fuse the RGB and depth features.

### *4.5. Complexity Analysis*

We conducted a complexity comparison with the other five models on the number of parameters and FLOPs, as shown in Table 3. The performance of the CNN-based models was relatively poor compared to the Transformer-based models. Our model performed better with fewer parameters and lower computational costs compared to SwinNet. The parameters and FLOPs of our model were 175.6 M and 96 G, respectively. Our model achieved comparable performance to SwinNet, yielding 88.4% parameters and 77.2% FLOPs.

**Table 3.** Complexity comparison and performance on LFSD and NLPR datasets. The best two results are shown in red and green fonts, respectively.

| Backbone | Model | Num_Parameters ↓ | FLOPs ↓ | LFSD $F_m$ ↑ | NLPR $F_m$ ↑ |
|---|---|---|---|---|---|
| CNN | CMW | 85.7 M | 208 G | 0.899 | 0.912 |
| | HiDANet | 59.8 M | 73.6 G | 0.877 | 0.922 |
| | JLDCF | 143.5 M | 211.1 G | 0.862 | 0.916 |
| | AFNet | 242 M | 128 G | 0.902 | 0.93 |
| Transformer | SwinNet | 198.7 M | 124.3 G | 0.903 | 0.94 |
| | **Ours** | 175.6 M | 96 G | 0.909 | 0.941 |

### *4.6. Failure Cases*

We show some failure cases on the challenging scenes in Figure 6: the first scene with multiple objects (first row and second column), and the second scene with poor depth map (third row and fourth row). As shown in the first scene, our model could not accurately locate multiple objects with complex backgrounds. Global feature relations are important for locating multiple salient objects. Multihead self-attention within a local window enhanced the locality, but it also limited the long-range model ability of the Swin Transformer. The second scene indicates that our model could not locate salient objects well in some scenes with poor depth maps. In addition to the low quality of depth maps,

there were misalignments between RGB images and depth maps at the pixel level. It is difficult to effectively fuse features for direct pixel-wise fusion. We will conduct further research in the future.

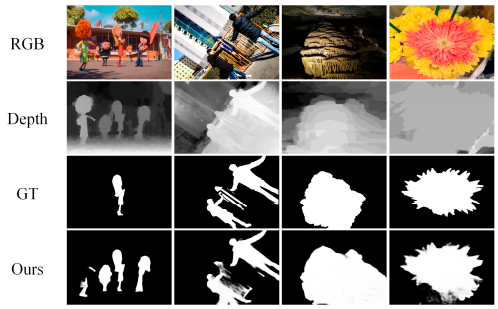


**Figure 6.** Visualization of failure cases in challenging scenes.

## 5. Conclusions

In this paper, we propose a novel Swin Transformer-based edge guidance network for RGB-D SOD. We employed the Swin Transformer as the backbone to extract features from RGB images and depth maps for capturing the long-range dependencies. Additionally, we proposed using the edge extraction module (EEM), the depth enhancement module, and the cross-modal interaction module (CIM) to enhance the local features. The EEM extracts edge features from the depth features, and the DEM employs edge information to enhance the depth features. The CIM effectively fuses RGB features and depth features from global and local contexts. With all these modules working together, our SwinEGNet model can accurately localize salient objects in various complex scenarios with sharp boundaries. Countless comparison studies and ablation experiments demonstrated that the proposed SwinEGNet showed outstanding performance on six widely used RGB-D SOD benchmark datasets. As an independent module, EEM can be applied to related tasks. In the future, we will extend our model to RGB-T salient object detection.

**Author Contributions:** Conceptualization, B.X.; methodology, S.W.; software, S.W.; validation, F.J.; supervision, B.X.; data curation, S.W.; formal analysis, S.W.; investigation, S.W.; resources, B.X.; writing—original draft, S.W.; writing—review and editing, F.J. and B.X; visualization, F.J. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is funded by the National Natural Science Foundation of China under grant number 62205334.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

## References

1. Fan, D.-P.; Wang, W.; Cheng, M.; Shen, J. Shifting more attention to video salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
2. Shimoda, W.; Yanai, K. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016.
3. Mahadevan, V.; Vasconcelos, N. Saliency-based discriminant tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Miami, FL, USA, 20–25 June 2009.
4. Ma, C.; Huang, J.B.; Yang, X.K.; Yang, M.H. Hierarchical convolutional features for visual tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3074–3082.
5. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
6. Wang, X.; Ma, H.; Chen, X.; You, S. Edge preserving and multiscale contextual neural network for salient object detection. *IEEE Trans. Image Process.* **2018**, *27*, 121–134. [CrossRef] [PubMed]

7. Liu, S.; Huang, D.; Wang, Y. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
8. Liu, J.; Hou, Q.; Cheng, M.; Feng, J.; Jiang, J. A simple pooling-based design for real-time salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
9. Zhao, J.; Cao, Y.; Fan, D.-P.; Cheng, M.; Li, X.; Zhang, L. Contrast prior and fluid pyramid integration for RGBD salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
10. Piao, Y.; Rong, Z.; Zhang, M.; Ren, W.; Lu, H. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
11. Chen, S.; Fu, Y. Progressively guided alternate refinement network for RGB-D salient object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Virtual, 23–28 August 2020; pp. 520–538.
12. Fan, D.-P.; Zhai, Y.; Borji, A.; Yang, J.; Shao, L. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In Proceedings of the European Conference on Computer Vision (ECCV), Virtual, 23–28 August 2020; pp. 275–292.
13. Li, G.; Liu, Z.; Ye, L.; Wang, Y.; Ling, H. Cross-modal weighting network for RGB-D salient object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Virtual, 23–28 August 2020; pp. 665–681.
14. Li, G.; Liu, Z.; Chen, M.; Bai, Z.; Lin, W.; Ling, H. Hierarchical Alternate Interaction Network for RGB-D Salient Object Detection. *IEEE Trans. Image Process.* **2021**, *30*, 3528–3542. [CrossRef] [PubMed]
15. Zhou, T.; Fu, H.; Chen, G.; Zhou, Y.; Fan, D.-P.; Shao, L. Specificity-preserving RGB-D Saliency Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 4681–4691.
16. Chen, H.; Li, Y.; Su, D. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. *Pattern Recognit.* **2019**, *86*, 376–385. [CrossRef]
17. Zhang, J.; Fan, D.-P.; Dai, Y.; Yu, X.; Zhong, Y.; Barnes, N.; Shao, L. RGB-D saliency detection via cascaded mutual information minimization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 4318–4327.
18. 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. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 9471–9481.
19. Lee, M.; Park, C.; Cho, S.; Lee, S. SPSN: Superpixel prototype sampling network for RGB-D salient object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
20. Sun, P.; Zhang, W.; Wang, H.; Li, S.; Li, X. Deep RGB-D Saliency Detection with Depth-Sensitive Attention and Automatic Multi-Modal Fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021.
21. Wu, Z.; Gobichettipalayam, S.; Tamadazte, B.; Allibert, G.; Paudel, D.P.; Demonceaux, C. Robust RGB-D fusion for saliency detection. In Proceedings of the 2022 International Conference on 3D Vision (3DV), Prague, Czechia, 12–15 September 2022; pp. 403–413.
22. Chen, T.; Xiao, J.; Hu, X.; Zhang, G.; Wang, S. Adaptive fusion network for RGB-D salient object detection. *Neurocomputing* **2023**, *522*, 152–164. [CrossRef]
23. Wu, Z.; Allibert, G.; Meriaudeau, F.; Ma, C.; Demonceaux, C. HiDAnet: RGB-D Salient Object Detection via Hierarchical Depth Awareness. *IEEE Trans. Image Process.* **2023**, *32*, 2160–2173. [CrossRef] [PubMed]
24. Pang, Y.; Zhao, X.; Zhang, L.; Lu, H. Caver: Cross-modal view mixed transformer for bi-modal salient object detection. *IEEE Trans. Image Process.* **2023**, *32*, 892–904. [CrossRef] [PubMed]
25. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
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 Inform. Process. Syst.* **2023**, *30*, 5998–6008.
27. Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; He, Z. A Survey of visual transformers. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**. *early access*. [CrossRef] [PubMed]
28. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
29. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 10012–10022.
30. Zhang, M.; Fei, S.; Liu, J.; Xu, S.; Piao, Y.; Lu, H. Asymmetric two-stream architecture for accurate rgb-d saliency detection. In Proceedings of the European Conference on Computer Vision (ECCV), Virtual, 23–28 August 2020; pp. 374–390.
31. Jiang, B.; Chen, S.; Wang, B.; Luo, B. MGLNN: Semi-supervised learning via multiple graph cooperative learning neural networks. *Neural Netw.* **2022**, *153*, 204–214. [CrossRef] [PubMed]
32. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122. [CrossRef]

33. Liu, N.; Zhang, N.; Wan, K.; Han, J.; Shao, L. Visual Saliency Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 4722–4732.
34. Zeng, C.; Kwong, S. Dual Swin-Transformer based Mutual Interactive Network for RGB-D Salient Object Detection. *arXiv* **2022**, arXiv:2206.03105. [CrossRef]
35. Liu, C.; Yang, G.; Wang, S.; Wang, H.; Zhang, Y.; Wang, Y. TANet: Transformer-based Asymmetric Network for RGB-D Salient Object Detection. *IET Comput. Vis.* **2023**, *17*, 415–430. [CrossRef]
36. Pang, Y.; Zhao, X.; Zhang, L.; Lu, H. Transcmd: Cross-modal decoder equipped with transformer for rgb-d salient object detection. *arXiv* **2021**, arXiv:2112.02363. [CrossRef]
37. Liu, Z.; Tan, Y.; He, Q.; Xiao, Y. Swinnet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection. *IEEE Trans. Circ. Syst. Video Technol.* **2021**, *32*, 4486–4497. [CrossRef]
38. Roy, A.M.; Bhaduri, J. DenseSPH-YOLOv5: An automated damage detection model based on DenseNet and Swin-Transformer prediction head-enabled YOLOv5 with attention mechanism. *Adv. Eng. Inform.* **2023**, *56*, 102007. [CrossRef]
39. Wei, J.; Wang, S.; Huang, Q. $F^3$net: 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; pp. 12321–12328.
40. Niu, Y.; Geng, Y.; Li, X.; Liu, F. Leveraging stereopsis for saliency analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Province, RI, USA, 16–21 June 2012; pp. 454–461.
41. Ju, R.; Ge, L.; Geng, W.; Ren, T.; Wu, G. Depth saliency based on anisotropic center-surround difference. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 1115–1119.
42. Peng, H.; Li, B.; Xiong, W.; Hu, W.; Ji, R. RGBD salient object detection: A benchmark and algorithms. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 5–12 September 2014; pp. 92–109.
43. Li, N.; Ye, J.; Ji, Y.; Ling, H.; Yu, J. Saliency detection on light field. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Zurich, Switzerland, 5–12 September 2014; pp. 2806–2813.
44. Fan, D.-P.; Lin, Z.; Zhang, Z.; Zhu, M.; Cheng, M. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2075–2089. [CrossRef] [PubMed]
45. Cheng, Y.; Fu, H.; Wei, X.; Xiao, J.; Cao, X. Depth enhanced saliency detection method. In Proceedings of the International Conference on Internet Multimedia Computing and Service, Xiamen, China, 10–12 July 2014; pp. 23–27.
46. Fan, D.-P.; Cheng, M.M.; Liu, Y.; Li, T.; Borji, A. Structure-measure: A new way to evaluate foreground maps. In Proceedings of the IEEE international conference on computer vision (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4548–4557.
47. Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 22–25 June 2009; pp. 1597–1604.
48. Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.; Borji, A. Enhanced-alignment measure for binary foreground map evaluation. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 13–19 July 2018; pp. 698–704.
49. Perazzi, F.; Krähenbühl, P.; Pritch, Y.; Hornung, A. Saliency filters: Contrast based filtering for salient region detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 733–740.
50. Fu, K.; Fan, D.-P.; Ji, G.; Zhao, Q. JL-DCF: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3052–3062.
51. Bi, H.; Wu, R.; Liu, Z.; Zhu, H. Cross-modal Hierarchical Interaction Network for RGB-D Salient Object Detection. *Pattern Recognit.* **2023**, *136*, 109194. [CrossRef]
52. Chen, T.; Hu, X.; Xiao, J.; Zhang, G.; Wang, S. CFIDNet: Cascaded Feature Interaction Decoder for RGB-D Salient Object Detection. *Neural Comput. Applic.* **2022**, *34*, 7547–7563. [CrossRef]
53. Zhang, M.; Yao, S.; Hu, B.; Piao, Y.; Ji, W. $C^2$DFNet: Criss-Cross Dynamic Filter Network for RGB-D Salient Object Detection. *IEEE Trans. Multimed.* **2022**. *early access*. [CrossRef]
54. Wang, X.; Jiang, B.; Wang, X.; Luo, B. Mutualformer: Multimodality representation learning via mutual transformer. *arXiv* **2021**, arXiv:2112.01177. [CrossRef]