2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 978-1-6654-6946-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CVPR52688.2022.01142

Pyramid Grafting Network for One-Stage High Resolution Saliency Detection

Chenxi Xie[1], Changqun Xia[*2], Mingcan Ma[1], Zhirui Zhao[1], Xiaowu Chen[1,2], Jia Li[1,2]
[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
[2]Peng Cheng Laboratory, Shenzhen, China
{xiechenxi,mingcanma,zhiruizhao,chen,jiali}@buaa.edu.cn, xiachq@pcl.ac.cn

Abstract

Recent salient object detection (SOD) methods based on deep neural network have achieved remarkable performance. However, most of existing SOD models designed for low-resolution input perform poorly on high-resolution images due to the contradiction between the sampling depth and the receptive field size. Aiming at resolving this contradiction, we propose a novel one-stage framework called Pyramid Grafting Network (PGNet), using transformer and CNN backbone to extract features from different resolution images independently and then graft the features from transformer branch to CNN branch. An attention-based Cross-Model Grafting Module (CMGM) is proposed to enable CNN branch to combine broken detailed information more holistically, guided by different source feature during decoding process. Moreover, we design an Attention Guided Loss (AGL) to explicitly supervise the attention matrix generated by CMGM to help the network better interact with the attention from different models. We contribute a new Ultra-High-Resolution Saliency Detection dataset UHRSD, containing 5,920 images at 4K-8K resolutions. To our knowledge, it is the largest dataset in both quantity and resolution for high-resolution SOD task, which can be used for training and testing in future research. Sufficient experiments on UHRSD and widely-used SOD datasets demonstrate that our method achieves superior performance compared to the state-of-the-art methods.

1. Introduction

Salient object detection (SOD) [1, 5] aims at identifying and segmenting the most attractive objects in a certain scene. As a pre-processing step, it is widely applied in various computer vision tasks, such as light field segmentation [21, 41], instance segmentation [47] and video object segmentation [13, 42].

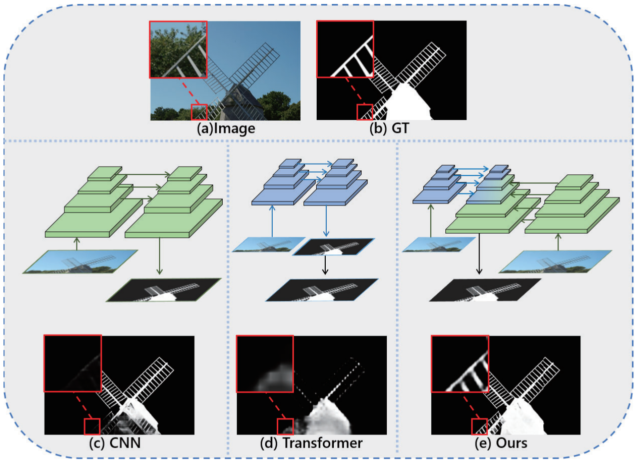

Figure 1. Comparison of the results of the different methods. (a) Input image. (b) Ground truth mask. (c) Directly input to Resnet-18 based FPN. (d) Downsample then input to Swin transformer based FPN. (e) Ours.

Recently, deep neural networks based salient object detection methods have made remarkable achievements [3, 9, 14, 19, 26, 29]. However, most of existing SOD methods perform well within a specific input low-resolution range (*e.g.*, $224 \times 224, 384 \times 384$). With the rapid development of image capture devices (*e.g.*, smartphone), the resolution (*e.g.*, 1080p, 2K and 4K) of the images accessible to people has far exceeded the range to which existing saliency detection method can be adapted directly. As shown in Fig. 1 (c), we fed the high-resolution image directly into the commonly used network with Resnet-18 as the backbone, and comparing ground truth Fig. 1 (b) shows that the segmentation result is incomplete and many detail regions are lost. In order to reduce computational consumption and memory usage, existing methods often downsample the input images

[*]Correspondence should be addressed to Changqun Xia (Email: xiachq@pcl.ac.cn). The code and dataset are available at https://github.com/iCVTEAM/PGNet.

978-1-6654-6946-3/22/$31.00 ©2022 IEEE
DOI 10.1109/CVPR52688.2022.01142

and then upsample the output results to recover original resolution, as illustrated in Fig. 1 (d). This challenge is due to the fact that most low-resolution SOD networks are designed in an Encoder-Decoder style, and as the input resolution increases dramatically, the size of features extracted increases, but the receptive field determined by the network is fixed, making the relative receptive field small, ultimately resulting in the inability to capture global semantics that are vital to SOD task. Since direct processing cannot handle the challenges posed by high resolution, a number of methods have emerged in recent years specifically designed for high-resolution input. There are two representative high-resolution SOD methods (HRSOD [40], DHQSOD [30]). HRSOD divides the whole process into global stage, local stage and reorganization stage, where the global stage provides guidance on both the local stage and the crop process. And DHQSOD disentangle the SOD task into classification task and regression task, where the two task is connected by their proposed trimap and uncertainty loss. They generate relatively good saliency maps with sharp boundaries.

However, both of the above methods use a multi-stage architecture, dividing SOD into semantic(in low resolution) and detailed (in high resolution) phases. This has led to two new problems: (1) Inconsistent contextual semantic transfer between stages. The intermediate maps obtained in the previous stages are input into the last stage, while the errors are also passed on. Further more, the refinement in the last stage will likely inherit or even amplify the previous errors as there is not enough semantic support, which implies that the final saliency maps are heavily dependent on the performance of the low-resolution network. (2) Time consuming. Compared to the one-stage method, the multi-stage method are not only difficult to parallel but also have the potential problem of increasing number of parameters, which makes it slow.

Based on the above defects of existing high-resolution methods , we propose a new perspective that since the specific features in a single network cannot settle the paradox of receptive field and detail retention simultaneously, instead we can separately extract two sets of features of different spatial sizes and then graft the information from one branch to the other. In this paper, we rethink the dual-branch architecture and design a novel one-stage deep neural network for high-resolution saliency detection named Pyramid Grafting Network (PGNet). As illustrated in Fig. 1 (e), we use both Resnet and Transformer as our Encoders, extracting features with dual spatial sizes in parallel. The transformer branch first decode the features in the FPN style, then pass the global semantic information to the Resnet branch in the stage where the feature maps of two branches have similar spatial sizes. We call this process feature grafting. Eventually, the Resnet branch completes the decoding process with the grafted features. Compared to classic FPNs, we have constructed a higher feature pyramid at a lower cost. To better graft features cross two different types of models, we design the Cross-Model Grafting Module (CMGM) based on the attention mechanism and propose the Attention Guided Loss to further guide the grafting. Considering that supervised deep learning method requires a large amount of high quality data, we have provided a 4K resolution SOD dataset (UHRSD) with the largest number to date in an effort to promote future high-resolution salient object detection research.

Our major contributions can be summarized as follows:

- We propose the first one-stage framework named PGNet for high-resolution salient object detection, which uses staggered connection to capture both continuous semantics and rich details.
- We introduce the Cross-Model Grafting Module to transfer the information from transformer branch to CNN branch, which allows CNN to not only inherit global information but also remedy the defects common to both. Moreover, we design the Attention Guided Loss to further promote the feature grafting.
- We contribute a new challenging Ultra High-Resolution Saliency Detection dataset (UHRSD) containing 5,920 images of various scenes at over 4K resolution and corresponding pixel-wise salient annotations, which is the largest high-resolution saliency dataset available.
- Experimental results on both existing datasets and ours demonstrate our method outperforms state-of-the-art methods in accuracy and speed.

2. Related work

During the past decades, a large amount traditional methods have been proposed to solve saliency detection problem [12, 38, 39]. However, these methods only focus on the low-level feature and ignore the rich semantic information resulting in unstable performance in complex scenarios.More details can be found in [1].

2.1. Deep Learning-Based Saliency Detection

Recently, remarkable progress has been made in saliency detection due to the application of deep neural network [18, 33, 36, 37, 44]. Hou *et al.* [11] and Chen *et al.* [4] use deep convolutional networks as Encoder to extract multi-level features and design various modules to fuse them in an FPN style. Ma *et al.* [23] and Xu *et al.* [37] avoid semantic dilution while suppressing loss of detail by experimenting with various feature connection paths. In addition, Wei *et al.* [33] generate saliency maps with sharp boundary by explicitly supervising edge pixels. The extensive use of transfomer in vision has also led to new advances in saliency detection. Liu *et al*. [20] take use of T2T-vit as backbone and design a multi-tasking decoder with a pure transformer architecture to perform RGB and RGB-D saliency detection. However, these methods are designed for low-resolution scenes and cannot be directly applied to high-resolution scenes.

2.2. High-Resolution SOD

Nowadays, focusing on high-resolution SOD methods is already trending. Zeng *et al*. [40] propose a paradigm for high-resolution salient object detection using GSN for extracting semantic information, and APS guided LRN for optimizing local details and finally GLFN for prediction fusion. Also they contributed the first high-resolution salient object detection dataset (HRSOD). Tang *et al*. [30] propose that salient object detection should be disentangled into two tasks. They first design LRSCN to capture sufficient semantics at low resolution and generate the trimap. By introducing the uncertainty loss, the designed HRRN can refine the trimap generated in first stage using low-resolution dataset. However, both of them use multi-stage architecture, which has led to slow inference, making it difficult to meet some real-world application scenarios. And a more serious problem is the semantic incoherence between networks. Thus we aim to design a one-stage deep network to get rid of the above defects.

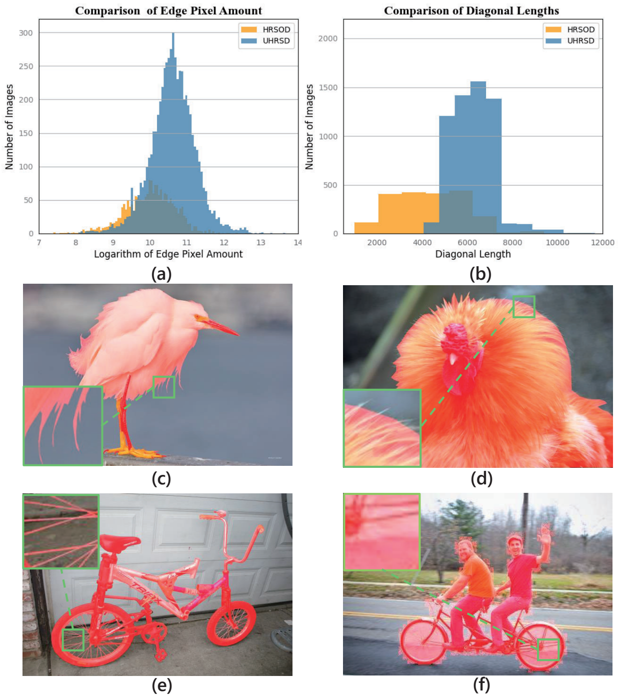

Figure 2. Comparison of the results of the different methods. (a) Comparison of the logarithm of edge pixel amount between our UHRSD and HRSOD. (b) Comparison of the diagonal length between our UHRSD and HRSOD [40] (c) Sample from our UHRSD. (d) Sample from HRSOD. (e) Sample from our UHRSD. (f) Sample from DUTS-TE. Best viewed by zooming in.

3. UHR Saliency Detection Dataset

Available SOD datasets. The existing common SOD datasets usually are in low-resolution (below 500 × 500). What's more, they have the following drawbacks for training high-resolution networks and evaluating high-quality segmentation results. Firstly the low resolution of the images results in insufficient detail information. Secondly, the quality of the edges of annotations is poor [40] .Lastly, the finer level of annotations is dissatisfied, especially for hard-case annotations which are handled perfunctorily as shown in Fig. 2 (f). The only available high-resolution dataset known is HRSOD [40]. However, the number of high-resolution images in HRSOD is limited.

UHRSD dataset. For supervised learning, training data is obviously important. Before this, the only available high-resolution training set was only 1,610 images, and we experimentally discovered that training only on it was easy to overfit its data distribution, which significantly impacted the model's generalization ability. If the low-resolution datasets are mixed together for training, a lot of noise will be introduced to affect the performance of the high-resolution model. To relief the lack of high-resolution datasets for SOD, we contribute the Ultra High-Resolution for Saliency Detection (UHRSD) dataset with a total of 5,920 images in 4K(3840 × 2160) or higher resolution, including 4,932 images for training and 988 images for testing. A total of 5,920 images were manually selected from websites (*e.g.* Flickr Pixabay) with free copyright. Our dataset is diverse in terms of image scenes, with a balance of complex and simple salient objects of various size. Multiple participants in the constructing process to ensure accuracy of salient annotations. Fig. 2 illustrates the superiority of our UHRSD. As shown in histogram Fig. 2 (a) (b), UHRSD datset is much larger than HRSOD datset and to the best of our knowledge is the largest dataset available. The large scale considerably alleviates the issues mentioned above when training high-resolution deep neural networks. In addition, the histogram Fig. 2(b) shows that the size of images in UHRSD far exceeds that of the existing high-resolution dataset. Not only that, Fig. 2(a) shows the number of pixels at the edges of our images also far surpasses the existing high-resolution dataset by a large margin, which means that UHRSD has richer and more challenging edge details. Lastly, through the comparison among Fig. 2(c)-(f), it's evident that UHRSD also has a finer level of annotation for the hard cases than both existing high-resolution dataset and low-resolution dataset.

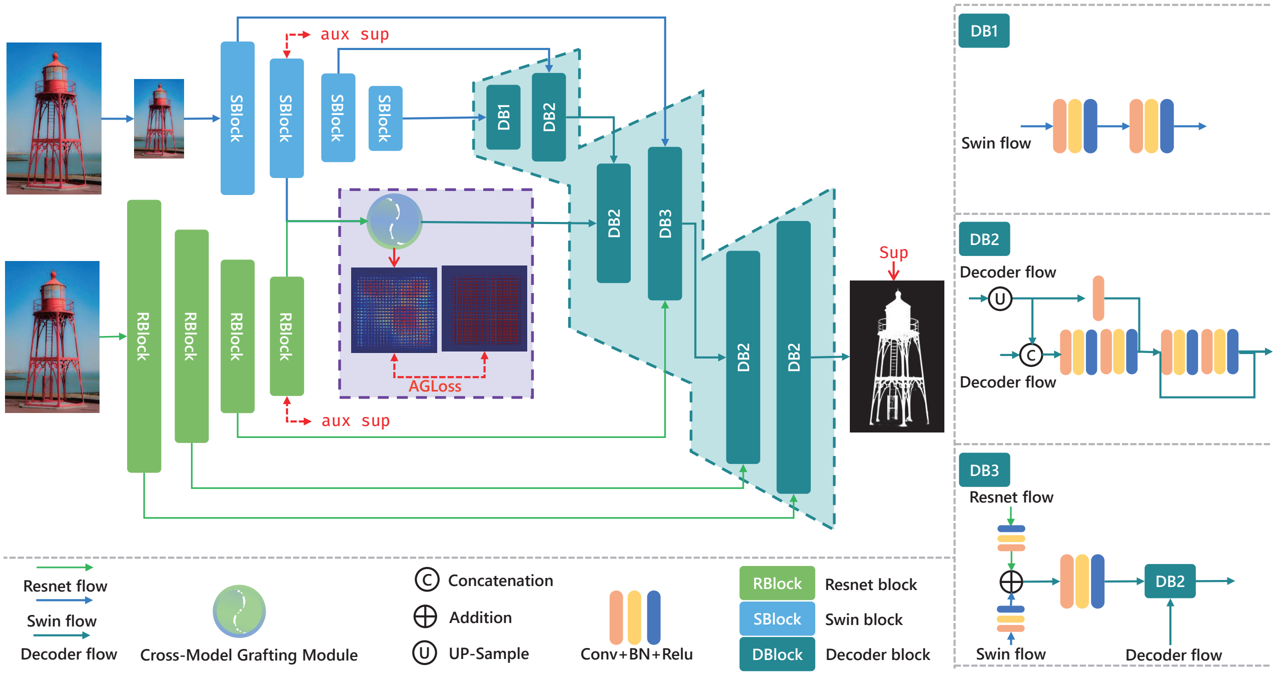

Figure 3. An overview of proposed Pyramid Grafting Network. Dual branches use Resnet and Swin transformer as encoder respectively. The DBn is the Decoder block with n input features, and the specific structure are shown on the right side. The two auxiliary supervisions are used to supervise the RP and SP mentioned in Sec. 4.4.

4. Methodology

4.1. Staggered Grafting Framework

The architecture of proposed network is shown in Fig. 3. As can be seen, the network consists of two encoders and a decoder. To better perform the respective tasks of the two encoders, Swin transformer and Resnet-18 are chosen as encoders. This choice of combination was made for the consideration of balancing efficiency and effectiveness. On the one hand, the transformer encoder is able to get accurate global semantic information in the low-resolution case, and the convolutional encoder can get rich detail with the high-resolution input. On the other hand, variability in the features extracted by different models may be complementary to identify saliency more accurately.

During the encoding process, two encoders are fed with images of different resolutions in order to capture global semantic information and detailed information respectively in parallel. The decoding phase can be divided into three sub-stages, first Swin decoding, followed by grafted feature decoding and finally Resnet decoding in a staggered structure. The feature decoded in the second sub-stage is produced from Cross-Model Grafting Module (CMGM), where the global semantic information is grafted from Swin branch to Resnet branch. Also the CMGM process a matrix named CAM to be supervised. Reviewing the whole process, we construct a higher feature pyramid through two lower pyramid using staggered connection structure as shown in Fig. 1. In other word, the network achieves deeper sampling depth at low computational cost to adapt to the challenge caused by high-resolution input.

4.2. Feature Extractors

Countering the massive computational consumption and memory usage generated by high-resolution input, we choose Resnet-18 [10] and Swin-B [22] as our backbones to balance performance and efficiency. For Resnet-18 encoder, five feature maps will be generated, which we denote the set as $\mathbb{R}$. The feature map extracted by top 7×7 layer offers limited performance gains but consume huge computational effort, especially for high-resolution input. Thus the utilized features in $\mathbb{R}$ can be denoted as $\{\boldsymbol{R}_i | i = 2, 3, 4, 5\}$. Due to the down-sampling in every stage, for input size $H \times W$, the size of feature $\boldsymbol{R}_i$ is $\frac{H}{2^i} \times \frac{W}{2^i} \times (C \times 2^{i-1})$, where $(C \times 2^i)$ is the channel of features. We remove the last stage while adopt the patch embedding feature of Swin transformer, which generates 4 features denoted as $\{\boldsymbol{S}_i | i = 1, 2, 3, 4\}$. Due to the nature that the embedding dim is fixed in transformer, the input size is 224×224 and the feature size in $\mathbb{S}$ is$\left\{\frac{56}{2^{i-1}} \times \frac{56}{2^{i-1}} \times (64 \times 2^i)\right\}$ for $i = 1, 2, 3$ and $14 \times 14 \times 512$ for $\boldsymbol{S_4}$. The spatial size of $\boldsymbol{R}_5$ is close to $\boldsymbol{S}_2$, hence we chose to graft the features here.

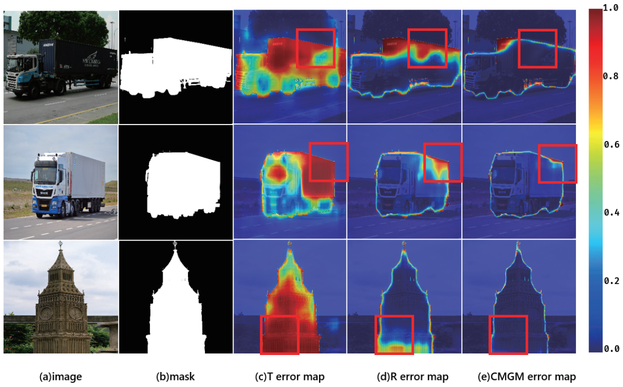

Figure 4. Examples of error elimination through CMGM. (a)(b) are the images and ground truth masks. (c) and (d) are thee salient error map generated from Transformer branch and Resnet branch respectively. (e) shows the error map generated from CMGM.

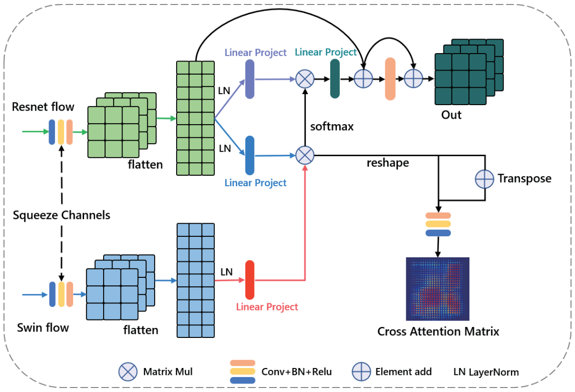

Figure 5. Architecture of Cross-Model Grafting Module.

4.3. Cross-Model Grafting Module

We propose Cross-Model Grafting Module(CMGM) to graft the feature f_{R_5} and f_{S_2} extracted by two different encoders. For feature f_{S_2}, due to the transformer's ability to capture information over long distances, it has global semantic information that is important for saliency detection. In contrast, CNNs perform well at extracting local information thus f_{R_5} have relatively rich details. However, due to the contradiction between feature size and receptive field, in f_{R_5} there will be many noises in the background. For a salient prediction of a certain region, the predictions generated from different features can be roughly summarized as three cases: (a) Both right, (b) Some of them right and (c) Both wrong. Existing fusion method using element-wise operation such as addition and multiplication may work for the first two cases. However, the element-wise operation and the convolutional operation focus on only limited local information, resulting fusion methods hardly remedy for common errors. Compared with the feature fusion, CMGM re-calculates the point-wise relationship between Resnet feature and Transformer feature, transferring the global semantic information from Transformer branch to Resnet branch so as to remedy the common errors. We calculate the error map by $E = |G - P| \in [0,1]$,where G is the ground truth and P is the salient prediction map generated by different branchs or CMGM. As shown in Fig. 4, the CMGM remedy the common error as expected.

Specifically, in CMGM it first flatten the $f_{R_5} \in \upsilon^{H\times W\times C}$ to $f'_R \in \upsilon^{1\times C\times HW}$, and do the same to f_{S_2} to get f'_S. Inspired by the multi-head self-attention mechanism, we apply layer normalization and linear projection on them respectively to get f^q_R, f^v_R and f^k_S. We obtain $\boldsymbol{Z}$ by matrix multiplication, the process can be expressed as follows:

$$\boldsymbol{Y} = \mathrm{softmax}(f^q_R \times {f^k_S}^{\mathrm{T}}), \tag{1}$$

$$\boldsymbol{Z} = \boldsymbol{Y} \times f^v_R, \tag{2}$$

then we input $\boldsymbol{Z}$ to the linear projection layer and reshape it back to size of $\upsilon^{H\times W\times C}$ before feeding into convolutional layer. Two shortcut connections were performed in the process as shown in Fig. 5. In addition, during the cross attention process, we generate Cross Attention Matrix based on $\boldsymbol{Y}$, which can be expressed as :

$$\mathrm{CAM} = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}(\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}))), \tag{3}$$

The detailed usage of CAM can be found in Sec. 4.4.

4.4. Attention Guided Loss

In order for CMGM to better serve the purpose of transferring information from the Transformer branch to the Renset branch, we design the Attention Guided Loss to supervise the Cross Attention Matrix explicitly. We argue that the Cross Attention Matrix should be similar to the attention matrix generated from ground truth, because the salient features should have a higher similarity, in other words the dot product should has a larger activation value. As shown in Fig. 6 given a salient map $\boldsymbol{M}$ with size $H \times W$, we first flatten it to $\boldsymbol{M}'$ with size $1 \times HW$. Then we apply matrix multiplication on $\boldsymbol{M}'$ to obtain corresponding attention matrix $\boldsymbol{M}^a$. The process can be denoted as $\boldsymbol{M}^a = \mathcal{F}(\boldsymbol{M})$ and the value of $\boldsymbol{M}^a_{xy}$ can be expressed as

$$\boldsymbol{M}^a_{xy} = {\boldsymbol{M}'}^{\mathrm{T}}_x \times \boldsymbol{M}'_y, \tag{4}$$

Then we use the transformation $\mathcal{F}(\cdot)$ to construct G^a, RP^a, SP^a, where G is the ground truth map, RP and SP are salient prediction map generated from feature $\boldsymbol{R}_5$ and $\boldsymbol{S}_2$ respectively. We propose the Attention Guided Loss based on weighted binary cross entropy (wBCE) to supervise the Cross Attention Matrix CAM generated from CMGM shown in Fig. 5. The BCE [6] can be written as:

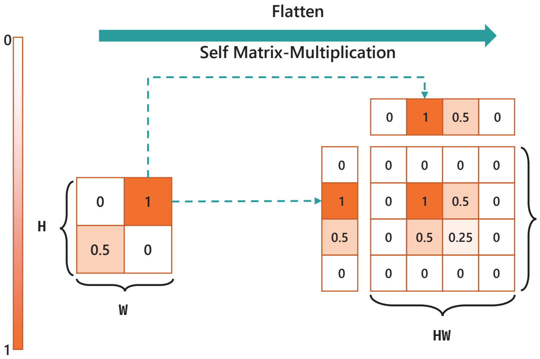

Figure 6. The construction of attention matrix. The operation is used to create the target and weights for proposed AGL.

$$\ell_{bce}(G_{xy},P_{xy}) = \begin{cases}\log(P_{xy}) & G_{xy}=1\\ \log(1-P_{xy}) & G_{xy}=0\end{cases}, \tag{5}$$

where G_{xy} is the ground truth label of the pixel (x, y), and P_{xy} is the predicted probability in predicted map and both of them are in range$[0, 1]$. Then our $\mathcal{L}_{AG}$ can be expressed as:

$$\mathcal{L}_{AG} = -\frac{\sum\limits_{i=1}^{H}\sum\limits_{j=1}^{W}(1+\beta\omega_{ij})\cdot \ell_{bce}(G^a_{ij},\mathrm{CAM}_{ij})}{\sum\limits_{i=1}^{H}\sum\limits_{j=1}^{W}(1+\beta\omega_{ij})}, \tag{6}$$

where β is a hyperparameter to adjust impact of the weight ω Eq. (7). In the Eq. (6), the ℓ_{bce} on each pixel is assigned with a weight $\beta\omega_{ij}$. The use of weight ω serves two purposes: (1) The degree of positive and negative sample imbalance is squared due to the matrix multiplication.(2) As described in Sec. 4.3, we want to remedy the errors common to both of branches. When $\beta\omega$ equals 0, the Eq. (6) becomes usual binary cross entropy loss $\mathcal{L}_{bce}$. The weight ω can be calculated by:

$$\omega_{ij} = \frac{1}{2}(|\ (G^a_{ij}-RP^a_{ij})\ | + |\ (G^a_{ij}-SP^a_{ij})\ |)+1, \tag{7}$$

where RP^a and SP^a are the attention matrix of RP and SP defined above.

What's more, we also apply the widely-used IoU loss [24] to pay more attention to the global structure of the image as suggested by [27]. The IoU loss $\mathcal{L}_{iou}$ can be referred to supplementary materials. In the end, our total loss can be expressed as follow:

$$\mathcal{L}_{total} = \mathcal{L}^P_{b+i}+\mathcal{L}_{AG}+\frac{1}{8}(\mathcal{L}^{auxiliary}_{b+i}), \tag{8}$$

where $\mathcal{L}_{b+i} = \mathcal{L}_{bce} + \mathcal{L}_{iou}$, and $\mathcal{L}^{auxiliary}_{b+i}$ is $\mathcal{L}_{b+i}$ applied on the RP and SP.

5. Experiments

5.1. Datasets and Evaluation Metrics

High-Resolution Datasets. The high-resolution datasets available are UHRSD (4,932 images for training and 988 for testing) , HRSOD [40] (1,610 images for training and 400 for testing). Followed by [30, 40], we also use the DAVIS-S for evaluation.

Low-Resolution Datasets. DUTS-TR [31] is used to train the model. In addition, we also evaluate our method on widely-used benchmark datasets: ECSSD [38] with 1,000 images, DUT-OMRON [39] with 5,168 images, PASCAL-S [17] with 850 images, DUTS-TE [31] with 5,019 images and HKU-IS [16] with 4,447 images.

Evaluation Metrics. We use following metrics to evaluate the performance of all methods. Firstly, Mean Absolute Error (MAE), defined as Eq. (9) where P is the prediction map and G is the ground truth. The second is Max F-measure (F^{Max}_{β}), which can be calculated by $F_\beta = \frac{(1+\beta^2)\cdot\text{precision}\cdot\text{recall}}{\beta^2\cdot\text{precision}\cdot\text{recall}}$, where β^2 is set to 0.3 as suggested in [2]. Then we adopt Structural similarity Measure (S_m) [7] and E-measure(E_ξ) [8] as many other methods [23, 32]. At last, to better evaluate the boundary quality which is important in High-resolution Saliency Detection [30, 40], we adopt the Boundary Displacement Error (BDE) to evaluate the result of high-resolution datasets, where lower values means better boundary quality.

$$\mathrm{MAE} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}|P_{ij}-G_{ij}|. \tag{9}$$

5.2. Implementation Details

We use Pytorch [25] to implement our model and two RTX 2080Ti GPUs are used for accelerating training. We choose Resnet-18 [10] and Swin-B_224 [22] as the backbone for convolutional branch and transformer branch respectively. The whole network is trained end-to-end by using stochastic gradient descent (SGD). We set the maximum learning rate to 0.003 for Swin backbone and 0.03 for others. The learning rate first increases then decays during the training process, what's more Momentum and weight decay are set to 0.9 and 0.0005 respectively. Batchsize is set to 16 and maximum epoch is set to 32. For data augmentation, we use random flip, crop and multi-scale input images [27, 30, 44]. In order to make fair comparisons and fully demonstrate the attributes of our UHRSD, we take three combinations of available datasets to train our model: (1) DUTS-TR (2) DUTS-TR+HRSOD-TR (3) UHRSD-TR+HRSOD-TR. During testing, each image is resized to 1024×1024 and then fed into the network without any post-processing(*e.g.* CRF [15]).

Table 1. Quantitative comparisons with state-of-the-art SOD models on five benchmark datasets in terms of max F-measure, MAE , E-measure, S-measure and BDE. The best two results are shown in red and green, respectively. D: trained on DUTS-TR, HD: trained on DUTS-TR and HRSOD-TR, UH: trained on UHRSD-TR and HRSOD-TR . The best two results are in red and green fonts.

Method	**HRSOD-TE**					**DAVIS-S**					**UHRSD-TE**					**DUT-OMRON**				**DUTS-TE**			
	F_β^{Max}	MAE	E_ξ	S_m	BDE	F_β^{Max}	MAE	E_ξ	S_m	BDE	F_β^{Max}	MAE	E_ξ	S_m	BDE	F_β^{Max}	MAE	E_ξ	S_m	F_β^{Max}	MAE	E_ξ	S_m
CPD$_{19}$	.867	.041	.891	.881	62.066	.871	.029	.921	.893	33.971	.894	.055	.884	.878	32.587	.797	.056	.866	.825	.865	.043	.887	.869
SCRN$_{19}$	.880	.042	.887	.888	75.696	.893	.027	.911	.902	46.592	.904	.051	.880	.887	40.176	.811	.056	.863	.837	.888	.040	.888	.885
DASNet$_{20}$	.893	.032	.925	.897	69.310	.902	.020	.949	.911	26.761	.914	.045	.892	.889	35.044	.827	.050	.877	.845	.895	.034	.908	.894
F3Net$_{20}$	.900	.035	.913	.897	65.757	.915	.020	.940	.914	44.760	.909	.046	.887	.890	39.612	.813	.053	.871	.838	.891	.035	.902	.888
GCPA$_{20}$	.889	.036	.898	.898	74.900	.922	.020	.934	.929	39.160	.912	.047	.886	.896	35.947	.812	.056	.860	.839	.888	.038	.891	.891
ITSD$_{20}$	.896	.036	.912	.898	87.946	.899	.022	.922	.909	68.256	.911	.045	.895	.897	41.174	.821	.061	.863	.840	.883	.041	.895	.885
LDF$_{20}$	.904	.032	.919	.904	58.714	.911	.019	.947	.922	35.447	.913	.047	.891	.888	33.775	.820	.051	.873	.838	.898	.034	.910	.892
CTD$_{21}$	.905	.032	.921	.905	63.907	.904	.019	.938	.911	42.832	.917	.043	.898	.897	33.835	.826	.052	.875	.844	.897	.034	.909	.893
PFS$_{21}$	.911	.033	.922	.906	63.537	.916	.019	.946	.923	30.612	.918	.043	.896	.897	37.387	.823	.055	.875	.842	.896	.036	.902	.892
HRSOD-DH$_{19}$	.905	.030	.934	.896	88.017	.899	.026	.955	.876	44.359	-	-	-	-	-	.743	.065	.831	.762	.835	.050	.885	.824
DHQSOD-DH$_{21}$	.922	.022	.947	.920	46.495	.938	.012	.947	.920	14.266	-	-	-	-	-	.820	.045	.873	.836	.900	.031	.919	.894
Our PGNet																							
Ours-D	.931	.021	.944	.930	46.923	.936	.015	.947	.935	34.957	.931	.037	.904	.912	32.300	.835	.045	.887	.855	.917	.027	.922	.911
Ours-DH	.937	.020	.946	.935	45.292	.950	.012	.975	.948	14.463	.935	.036	.905	.912	32.008	.835	.046	.887	.858	.919	.028	.925	.912
Ours-UH	.945	.020	.946	.938	57.147	.957	.010	.979	.954	12.725	.949	.026	.916	.935	30.019	.772	.058	.884	.786	.871	.038	.897	.859

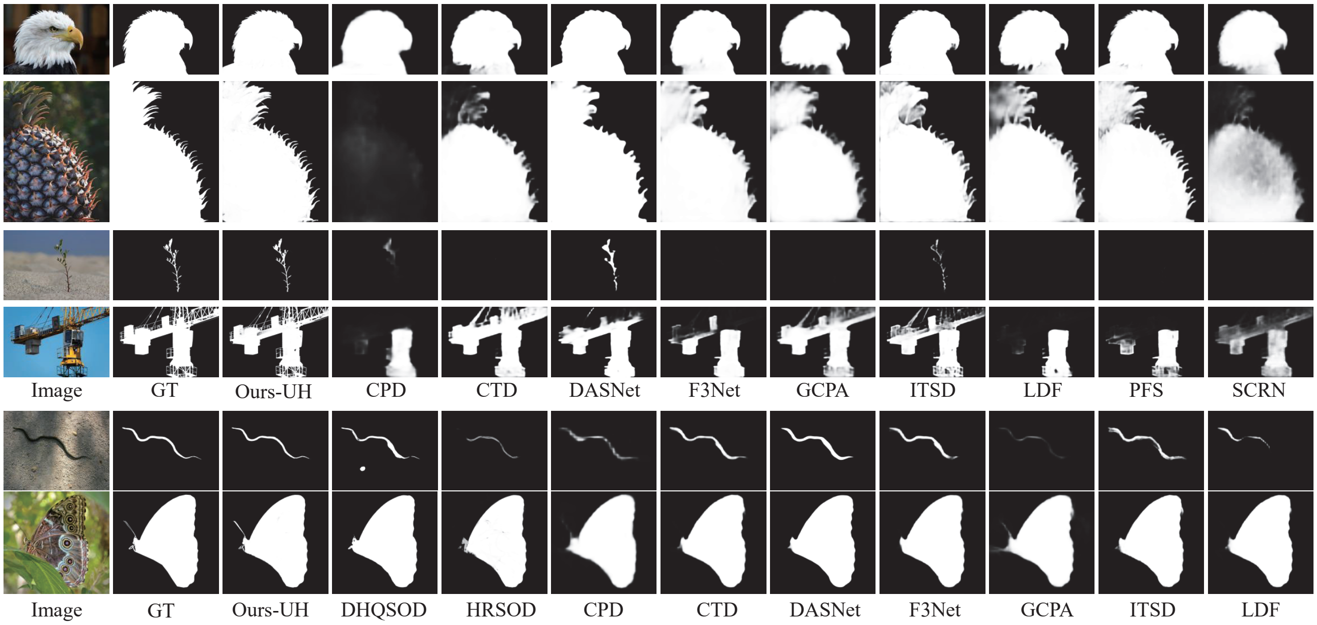

Figure 7. Visual comparison between our method and SOTA methods. The first four lines are from our UHRSD-TE and the next two lines are from HRSOD-TE. Best viewed by zooming in.

5.3. Comparison with the State-of-the-arts

We compare our proposed PGNet with 11 SOTA methods, including CPD [34], SCRN [35], DASNet [43], F3Net [32], GCPA [4], ITSD [46], LDF [33], CTD [45], PFS [23], HRSOD [40], DHQSOD [30], where HRSOD and DHQSOD are designed for high-resolution salient object detection. All of the above methods use Resnet-50 [10] as the backbone except for HRSOD which uses VGG16 [28]. And all of them are trained on DUTS-TR [31] dataset, except for the marked ones like HRSOD-DH and DHQSOD-DH, which are trained on the mixed dataset (HRSOD [40] and DUTS-TR). For a fair comparison, we use either the available implementations or the saliency maps provided by the authors. It's worth noting that the vacant lines in Tab. 1 are caused by the fact that one of them is not available so far and the other not being consistent with our test environment.

Table 2. Comparison of different architectures and compositions.

Composition	HRSOD-TE			
	F_β^{Max}	MAE	E_ξ	S_m
baseline_Resnet-18	.878	.051	.875	.871
baseline_Swin	.915	.027	.937	.921
baseline_R+S+CMGM	.940	.023	.944	.936
baseline_R+S+CMGM+AGL	.945	.020	.946	.938

Table 3. Performance with the different grafted features. $\boldsymbol{R}_i$ denotes the ith feature of $\mathbb{R}$ defined in Sec. 4.2, and $\boldsymbol{S}_i$ is similar.

Feature Pair	HRSOD-TE			UHRSD-TE		
	F_β^{Max}	MAE	E_ξ	F_β^{Max}	MAE	E_ξ
$R_5 - S_4$	.913	.029	.922	.935	.031	.907
$R_5 - S_3$	.939	.022	.937	.947	.026	.912
$R_5 - S_2$	.945	.020	.946	.949	.026	.916
$R_5 - S_1$	.937	.022	.935	.947	.026	.910

Quantitative Comparison. As mentioned above, for fair comparison we use three settings of train set. As can be seen in Tab. 1, the results of training on either only DUTS-TR or mix of DUTS-TR and HRSOD-TR exceed the SOTA by a large margin on both high-resolution and low-resolution test sets. When using the mixed dataset DUTS-HRSOD, our method has significantly improved on high-resolution datasets. There may be discrepancy between the distribution of high-resolution and low-resolution data. This is further supported by the results of training on the UHRSD-HRSOD mixed dataset, where the performance of the high-resolution dataset is significantly improved, especially for UHRSD-TE. This demonstrates that the annotation bias of high-resolution datasets differing from low-resolution datasets has a promotional effect on supervised high-resolution saliency detection method, which is the reason why high-resolution training data with high-quality annotation is in great demand.

Visual Comparison. To exhibit the characteristics of high-resolution dataset and the superiority of our method on it, Fig. 7 shows representative examples of visual comparison of ours with respect to others. As can be seen, our method can capture details well and produce clear boundary (row 1 and 2). More than the high quality boundary, another significant aspect of high-resolution SOD is the ability to segment objects with small sturctures that are easily to overlook in low-resolution cases (row 3, 5 and 6). This also demonstrates the superiority that our method makes the process one-stage. What's more, our method works well even in some extremly complex scenarios such as row 4.

5.4. Ablation Study

To better illustrate the nature of proposed method for high-resolution images, the ablation studies are based on the settings of **Ours-UH**, which is trained on the mixed dataset UHRSD-TR and HRSOD-TR.

Ablation Study for Compositions. To prove the effectiveness of proposed feature grafting method including the CMGM and AGL, we report the quantitative performance in Tab. 2. The baseline_Resnet-18 and baseline_Swin represent the widely-used U-shape network with Resnet-18 backbone and Swin backbone respectively. As can be seen in row 3, our proposed staggered architecture and Cross-Model Grafting Module inherits the strengths of both models. What's more, under the guiding role of AGL, performance has been further improved.

Ablation Study for Grafting Position. To investigate the impact of grafting position on the network performance, we conduct a series of experiments with different grafting feature pairs. As shown in Tab. 3, starting with the alignment of the last stage of two encoders, the performance gradually improves as the number of staggered layers increase until reaching the best at the pair $\boldsymbol{R}_5 - \boldsymbol{S}_2$. This may be due to the spatial size of feature maps. When the sizes are close, the spatial information in the features extracted from two models corresponds to each other, which in turn promotes the feature grafting.

6. Limitation

Our method is simple and fast for one-stage high-resolution saliency detection, but the training process is still quite demanding on GPU memory usage, resulting in a high cost of training. What's more, though our method has already a superior input resolution compared to previous SOD methods, the input resolution is not unlimited. For excessive resolution such as 4K, images need to be downsampled first before input.

7. Conclusion

In this paper, we propose the Pyramid Grafting Network for one-stage high-resolution salient object detection. The proposed staggered grafting patterns effectively exploit the advantages of each of the two existing different encoder. In addtion the proposed Cross-Model Grafting Modules and Attention Guided Loss cooperate with each other to inherit the advantages and remedy the common defects of the CNN and transformer. It is worth noting that we contribute the first 4K resolution SOD dataset for advancing future studies in high-resolution SOD. Extensive experiments demonstrates that our method not only outperforms the state-of-the-art methods but also is able to produce high-resolution saliency predictions fast and accurately.

Acknowledgments: This work was supported by the National Natural Science Foundation of China under Grant 62132002, Grant 61922006 and Grant 62102206.

References

[1] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational visual media*, pages 1–34, 2019. 1, 2

[2] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722, 2015. 6

[3] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–250, 2018. 1

[4] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. *arXiv preprint arXiv:2003.00651*, 2020. 2, 7

[5] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014. 1

[6] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005. 5

[7] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017. 6

[8] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018. 6

[9] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *European Conference on Computer Vision*, pages 275–292. Springer, 2020. 1

[10] K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. 2016 ieee conf comput vispattern recognit. 2016: 770-778 https://doi. org/10.1109. CVPR, 2016. 4, 6, 7

[11] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3203–3212, 2017. 2

[12] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998. 2

[13] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4922–4933, 2021. 1

[14] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated rgb-d salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9471–9481, 2021. 1

[15] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. 6

[16] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015. 6

[17] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014. 6

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[19] Jiang-Jiang Liu, Qibin Hou, and Ming-Ming Cheng. Dynamic feature integration for simultaneous detection of salient object, edge and skeleton. *arXiv preprint arXiv:2004.08595*, 2020. 1

[20] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4722–4732, 2021. 3

[21] Nian Liu, Wangbo Zhao, Dingwen Zhang, Junwei Han, and Ling Shao. Light field saliency detection with dual local graph learning and reciprocative guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4712–4721, 2021. 1

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 4, 6

[23] Mingcan Ma, Changqun Xia, and Jia Li. Pyramidal feature shrinking for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2311–2318, 2021. 2, 6, 7

[24] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3438–3446, 2017. 6

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[26] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020. 1

[27] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019. 6

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[29] Jinming Su, Jia Li, Yu Zhang, Changqun Xia, and Yonghong. Tian. Selectivity or invariance: Boundary-aware salient object detection. In *ICCV*, 2019. 1

[30] Lv Tang, Bo Li, Yijie Zhong, Shouhong Ding, and Mofei Song. Disentangled high quality salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3580–3590, 2021. 2, 3, 6, 7

[31] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017. 6, 7

[32] Jun Wei, Shuhui Wang, and Qingming Huang. F^{3}net: Fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12321–12328, 2020. 6, 7

[33] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13025–13034, 2020. 2, 7

[34] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2019. 7

[35] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7264–7273, 2019. 7

[36] Changqun Xia, Jia Li, Xiaowu Chen, Anlin Zheng, and Yu Zhang. What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4142–4150, 2017. 2

[37] Binwei Xu, Haoran Liang, Ronghua Liang, and Peng Chen. Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3004–3012, 2021. 2

[38] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1155–1162, 2013. 2, 6

[39] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 2, 6

[40] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7234–7243, 2019. 2, 3, 6, 7

[41] Miao Zhang, Jingjing Li, Ji Wei, Yongri Piao, and Huchuan Lu. Memory-oriented decoder for light field salient object detection. *Advances in neural information processing systems*, 32, 2019. 1

[42] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1553–1563, 2021. 1

[43] Jiawei Zhao, Yifan Zhao, Jia Li, and Xiaowu Chen. Is depth really necessary for salient object detection? In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1745–1754, 2020. 7

[44] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8779–8788, 2019. 2, 6

[45] Zhirui Zhao, Changqun Xia, Chenxi Xie, and Jia Li. Complementary trilateral decoder for fast and accurate salient object detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4967–4975, 2021. 7

[46] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9141–9150, 2020. 7

[47] Ziqi Zhou, Zheng Wang, Huchuan Lu, Song Wang, and Meijun Sun. Multi-type self-attention guided degraded saliency detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13082–13089, 2020. 1