Expanded Edge Penalty Loss for Salient Object Detection

Nan Wang[1,2], Yuetian Shi[1,2], Jie Fang[3], Fanchao Yang[1], Geng Zhang[1,†], Siyuan Li[1], Xuebin Liu[1]
[1]*Key Laboratory of Spectral Imaging Technology CAS,Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, P.R.China*
[2]*University of Chinese Academy of Sciences, Beijing, P.R.China*
[3]*School of Communications and Information Engineering & School of Artificial Intelligence, Xi'an University of Posts & Telecommunications, Xi'an, 710121, Shaanxi, China.*
nanwang.ac@gmail.com, yuetian.shi.opt@gmail.com, 2443952262@qq.com, frankyang1987@126.com, gzhang@opt.ac.cn, lsy@opt.ac.cn, lxb@opt.ac.cn

2021 4th International Conference on Information Communication and Signal Processing (ICICSP) | 978-1-6654-0757-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICSP54369.2021.9611977

***Abstract*—As an indispensable preprocessing technique for image understanding, salient object detection aims to extract interesting regions from an image for subsequent processing, which has attracted much attention since its wide range of applications. Recently, with the rapid development of artificial intelligence and machine learning, deep neural network especially deep convolutional neural network based methods have achieved competitive performances because of their strong feature representation capability. However, most of these methods often suffer from coarse boundaries. The main reason is that equal penalty factor is applied to each pixel in the image to optimize the network, but there are huge distinctions in prediction complexity among different ones actually. Specifically, pixels closer to the boundaries are increasingly difficult because of their huge gaps between structural information and semantic label. In these cases, we present an Expanded Edge Penalty Loss (E^2PL) for salient object detection. E^2PL gives bigger penalty factors to pixels distributed in boundary and near-boundary regions, and further dynamically adjusts their contributions to the model optimization. In addition, the experimental results on five public and challenging datasets have validated the superiority and effectiveness of the proposed method.**

Keywords*—RGB images, Salient object detection, coarse object boundaries, expanded edge penalty loss.

I. Introduction

Salient object detection is an important preprocessing technique for many visual understanding tasks, which has attracted much attention during the last decade because of its wide range of practical applications such as image retrieval[1], semantic segmentation[2], camouflaged object detection[3], object classification[4] and age prediction[5].Multi-source Salient object detection [6],[7] are also proposed. Recently, with the rapid development of machine learning and artificial intelligence, some deep fully convolutional neural network based methods have achieved relatively significant performance on salient object detection task. However, most of these methods often suffer from coarse boundaries. The reasons mainly contain two aspects, intrinsic limitations of pooling operations and inappropriate penalty losses. Specifically, the pooling operator expands the receptive field by shrinking feature maps layer by layer, which results in the low spatial resolution of feature maps from deep layers. In addition, most existing methods apply the same penalty factor to each pixel in the whole images to optimize the network while ignoring the huge distinctions in the prediction complexity among different pixels. In these cases, the detection results of these methods can not meet our expectations.

To address the aforementioned issues, this paper presents an expanded edge penalty loss based approach for salient object detection. Specifically, U-Net is used as our backbone network of our framework, which can propagate more detailed spatial information from shallow to deep layers through its skip connections. Besides, self-attention mechanism is incorporated into the backbone network to enhance the feature representation. In addition, an expanded edge weight coefficient matrix based loss is proposed to increase the influence of difficult samples to the model optimization, which alleviates the prediction complexity distinctions of different pixels.

In summary, the contributions of this paper can be listed as follows:

1) We utilize a U-Net based framework for salient object detection, which can propagate spatial details from shallow layers to deep ones through skip connection.
2) We incorporate self-attention mechanism into our framework to enhance the representation, which fully considers interactions among different local regions.
3) We present an expanded edge penalty loss term to optimize the proposed network, which increases sensitivity and effectiveness of the model for difficult regions.

II. Proposed Method

A. Overview

Firstly, an image division strategy is designed to address the problem of variant input sizes. In addition, inspired by the superiority of U-Net[8], dilated convolutional network, and dense convolutional network on many visual understanding tasks, a novel context guided convolutional network is proposed to finalize the detection. Finally, an expanded edge penalty loss is proposed to optimize the model. The overview framework of the proposed method is shown as Fig. 1.

† Corresponding author.

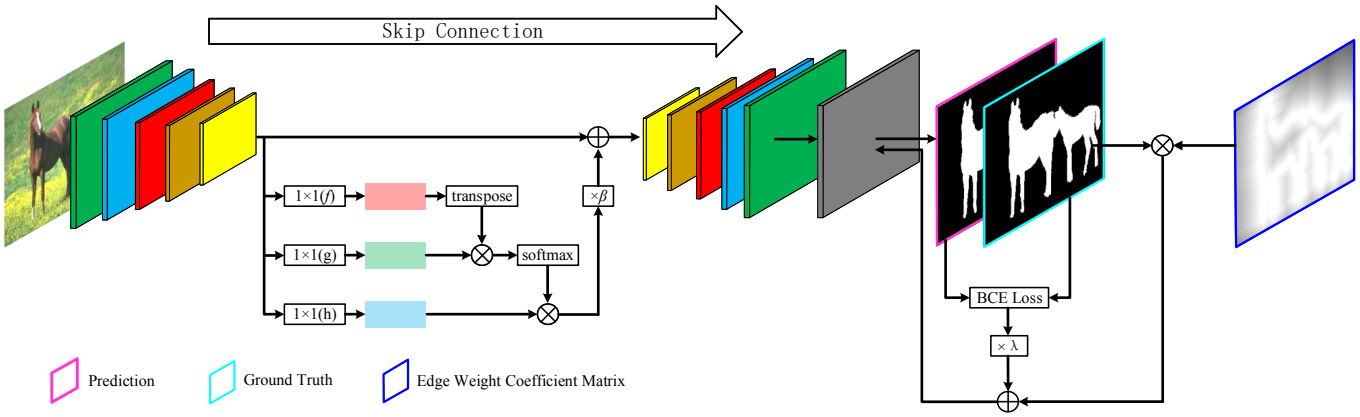

Figure 1: The architecture of the proposed method for salient object detection, which utilizes a dynamical edge weight coefficient matrix to adjust the impacts of different pixels to the model optimization.

B. Network

The proposed network is actually the U-Net equipped with popular self-attention module[19], which is described particularly as follows. U-Net has achieved relatively competitive performances on many image-to-image transformations tasks since its skip connection strategy, which can propagate more spatial structure information from shallower layers to deeper ones effectively. In this case, we utilize U-Net as our baseline network. Besides, we incorporate the popular self-attention mechanism into the U-Net to encode the relationships among different local regions into the features, which can provide auxiliary structure information to each region in a global view. Incidentally, the self-attention mechanism is only followed by the convolution layer with smallest output size since the computation complexity. Additionally, in order to clearly describe the interactive computing between encoder and decoder, we review the self-attention mechanism in details here.

Feature maps from the fifth convolution layer $\mathbf{X} \in \Re^{D\times N}$ are fed into two convolutional branches $f(\cdot)$ and $g(\cdot)$ to obtain new feature maps $f(\mathbf{X}) = \mathbf{W}_f\mathbf{X} + \mathbf{b}_f$ and $g(\mathbf{X}) = \mathbf{W}_g\mathbf{X} + \mathbf{b}_g$. Then the interactions among different local regions are calculated with Equation 1,

$$r_{(j,i)} = \frac{e^{s^{(i,j)}}}{\sum_{i=1}^{N} e^{s^{(i,j)}}}, \tag{1}$$

where N is the number of blocks, and $s^{(i,j)} = f\left(\mathbf{X}^{(i)}\right)^T g\left(\mathbf{X}^{(j)}\right)$. $r^{(j,i)}$ depicts the extent to which the algorithm attends to the i_{th} block when representing the j_{th} one. In addition, the element of self-attention feature $\mathbf{O}_{SA} \in R^{D\times N}$ can be obtained by Equation 2,

$$\mathbf{O}_{SA}^{(j)} = \sum_{i=1}^{N} r^{(i,j)} h\left(\mathbf{x}^{(i)}\right). \tag{2}$$

Finally, the output of the self-attention mechanism can be obtained by Equation 3,

$$\mathbf{Y} = \mathbf{X} + \beta\mathbf{O}_{SA}. \tag{3}$$

From Equation 3 we can see the output of self-attention mechanism contains two terms, self-attention features and original input features. β is a hyperparameter to balance these two terms, which is set to 2 in this paper.

C. Expanded edge penalty loss

Even though conventional binary cross entropy loss has achieved satisfactory performances on salient object detection, which still exist two limitations, 1) sample imbalance problem and 2) sample inconsistency problem. The first problem has been well addressed by some data augmentation strategies and novel classification loss functions while the second one is often ignored. Actually, prediction difficulties for pixels even in the same object are significantly different, ones in boundary and near-boundary regions are more difficult than those in inner flat regions. In this case, we design an expanded edge coefficient matrix based loss function to emphasize the influence of difficult regions when training the model, which is described in details as follows.

First, we utilize Canny descriptor to extract edge of the groundtruth. Second, we calculate the minimum distance between each pixel in the groundtruth to ones in the extracted edge regions. Third, we utilize a negative exponential function to mapping the distance matrix to edge weight coefficient matrix, which reflects the prediction complexity of corresponding pixels. Finally, we combine the edge weight coefficient matrix with the conventional MSE loss to optimize the model, which is defined as Equation 4,

$$\mathcal{L}_1 = \left\|(\mathbf{P} - \mathbf{G}) \odot \mathbf{E}_v\right\|_F^2, \tag{4}$$

where $\mathbf{P}$ and $\mathbf{G}$ respectively denote the predicted saliency map and corresponding groundtruth. In addition, $\odot$ denotes the element-wise multiplication. $\mathbf{E}_v$ denotes the proposed expanded edge weight coefficient matrix, whose $(i,j)_{th}$ is defined as Equation 5,

$$\mathbf{E}_v^{(i,j)} = \exp\left(-\frac{\mathbf{D}^{(i,j)}}{\sigma^2}\right), \tag{5}$$

where $\mathbf{D}^{(i,j)}$ represents the minimum distance between $(i,j)_{th}$ pixel and boundary regions, and σ is a hyperparameter to adjust the sensitivity of distance value, which is set to 1 in this paper. Specifically, $\mathbf{D}^{(i,j)}$ is defined as Equation 6,

$$\mathbf{D}^{(i,j)} = \min_{(w,h)\in \mathrm{E}} \{|i-w| + |j-h|\}, \tag{6}$$

where E represents the coordinate collection of edge regions.

Actually, the overall loss used in this paper is defined as Equation 7,

$$\mathcal{L} = \mathcal{L}_1 + \lambda \mathcal{L}_2, \tag{7}$$

where $\mathcal{L}_1$ is the proposed edge expanded weight coefficient matrix based MSE loss, $\mathcal{L}_2$ is the focal loss[20], λ is used to balance loss terms. $\mathcal{L}_2$ is defined as Equation 8,

$$\begin{aligned}\mathcal{L}_2 = -\frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}(\alpha y_{(w,h)}\left(1-\hat{y}_{(w,h)}\right)^{\gamma}\log \hat{y}_{(w,h)} \\ + (1-\alpha)\, y_{(w,h)}\hat{y}_{(w,h)}^{\gamma}\log\left(1-\hat{y}_{(w,h)}\right)),\end{aligned} \tag{8}$$

where W and H represent the width and height of the image respectively. $y_{(w,h)}$ and $\hat{y}_{(w,h)}$ denotes the label and predicted saliency score of $(w,h)_{th}$ pixel in the image respectively. α and γ are two hyperparameters to balance the samples of salient and other regions, the hyperparameter pair with larger α and smaller γ means the greater emphasis on salient regions. Specifically, α and γ are defined as Equation 9 and Equation 10 respectively.

$$\alpha = \frac{\sum_{w=1}^{W}\sum_{h=1}^{H}\mathrm{I}\{y_{(w,h)}=0\}}{WH}, \tag{9}$$

$$\gamma = \frac{\sum_{w=1}^{W}\sum_{h=1}^{H}\mathrm{I}\{y_{(w,h)}=1\}}{WH}, \tag{10}$$

where $\mathrm{I}\{\cdot\}$ denotes the indictor function, which equals to 1 if the condition satisfies and equals to 0 otherwise.

III. Experiments

This section details the experiments, including datasets, evaluation metrics, contrasting methods and experimental results, which is introduced in subsection A, subsection B, subsection C, and subsection D respectively.

A. Datasets

To validate the superiority of the proposed method, we test it on five public and challenging datasets, including ECSSD[9], PASCAL-S[10], DUT-OMRON[11], HKUIS[12], and DUTS [13]. Specifically, ECSSD contains 1000 images with different complex scenes. PASCAL-S contains 850 images from the validation subset of PASCAL VOC semantic segmentation dataset[14]. DUT-OMRON contains 5168 images, and ones in this set contain one more salient objects with complex backgrounds. HKU-IS contains 4447 images with pixel-wise annotations, and most images in this dataset have multiple disconnected salient objects. DUTS contains 15572 images, and ones are challenging since the variety of locations and scales. In addition, some samples of these five datasets are shown in Fig. 2.

Figure 2: Some samples of different datasets. Specifically, images in the first to fifth column denote samples from ECSSD, PASCAL-S, DUT-OMRON, HKUIS and DUTs datasets respectively. In addition, images in the first row are original RGB ones, and images in the second row are corresponding saliency groundtruths.

B. Evaluation metrics

Three common metrics are used to evaluate the performance of our proposed method, including precision (P), recall (R), and F-measure (F_β). Specifically, MAE is defined as Equation 11,

$$MAE = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\left|M_{(w,h)} - G_{(w,h)}\right|, \tag{11}$$

where M and G represent the binary mask from predicted saliency map through a specific threshold, and G is the corresponding annotated groundtruth. P and R are shown as Equation 12 and Equation 13 respectively.

$$P = \frac{|M\cap G|}{|M|}, \tag{12}$$

$$R = \frac{|M\cap G|}{|G|}. \tag{13}$$

Besides, we also utilize F_β, a weighted harmonic mean of P and R with a non-negative β, to evaluate the method, which is defined as Equation 14.

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

where β is a hyperparameter to balance the importance of P and R. According to the suggestion in [15], β^2 is set to 0.3 to emphasize precision.

C. Contrasting methods

In order to verify the superiority of the proposed method, we compare it with three state-of-the-art methods, including SRM[16], DGRL[17], and PiCANet[18]. Specifically, SRM augments feedforward neural networks with a novel pyramid pooling module and a multi-stage refinement mechanism to obtain finer structures and details. DGRL utilizes a global Recurrent Localization Network (RLN) to exploit contexture information, and utilizes a local Boundary Refinement Network (BRN) to adaptively learn the local contextual information for each spatial position. PiCANet generates an attention map in which each attention weight corresponds to the contextual relevance at each context location to select the useful contextual information.

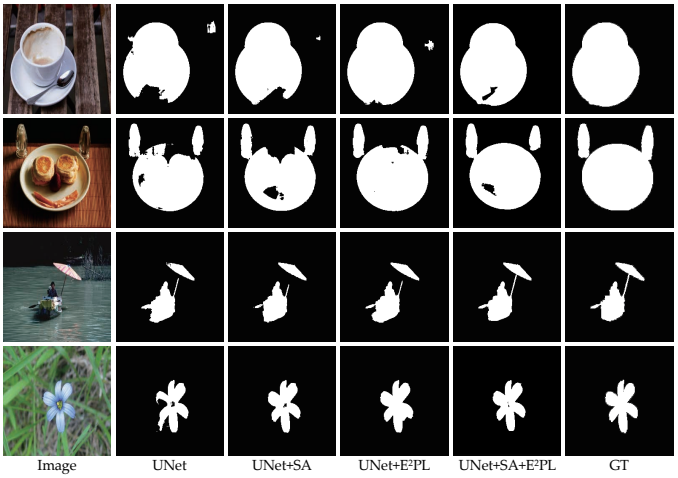

Figure 3: Visualized results of ablation experiments. Images in the first column are original input images, ones in the second to fourth column are detection results of ResNet18-UNet, ResNet18-U$_s$Net, ResNet18-UNet+E^{2}PL, and ResNet18-U$_s$Net+E^{2}PL. In addition, ones in the last column are corresponding groundtruth. Specifically, ResNet18-UNet denotes the ResNet18 based UNet, ResNet18-U$_s$Net denotes the combination of ResNet18UNet and self-attention.

D. Experimental results

This subsection reports the experimental results, including ablation experimental results and contrasting experimental results. Specifically, ablation experiment is to validate the effectiveness of each proposed part, and contrasting experiment is to demonstrate the superiority of our proposed method compared to existing ones.

Table I: Ablation experimental results on ECSSD dataset

S-Att	E^{2}PL	P	R	F_β	MAE
–	–	0.8572	0.7657	0.8357	0.0970
√	–	0.8591	0.7691	0.8380	0.0940
–	√	0.8647	0.7570	0.8387	0.0963
√	√	0.8687	0.8155	0.8575	0.8255

1) Ablation experimental results: This Sec. reports the ablation experimental results, the specific performances are shown in Table I and some visualized samples are shown in Fig. 3. In Table I and Fig. 3, S-Att denotes self-attention mechanism, and E^{2}PL denotes the proposed expanded edge penalty loss. From Table I we can see that, both S-Att and E^{2}PL can improve the detection results to a certain extent, and their combination can improve the performance furthermore. Specifically, when the self-attention module is incorporated into the ResNet18 based U-Net, it obtains 0.28% increment in terms of F_β. In addition, compared with the traditional Focal Loss, when the proposed expanded penalty loss is used to optimize the ResNet18 based U-Net, it obtains 0.36% increment in terms of F_β. Finally, when the self-attention module and the proposed penalty expanded penalty loss are simultaneously incorporated into the framework, they improve F_β by 2.61%. From these perspectives, the proposed expanded edge penalty loss and self-attention mechanism can contribute to each other to a large extent.

In addition, visualized results can also reflect the effectiveness of self-attention mechanism and the proposed expanded edge penalty loss. Specifically, from images in second and third columns we can see that, self-attention mechanism can significantly alleviate the influences of discrete noisy points. In addition, compared saliency results in second and fourth columns, we can find that the proposed expanded edge penalty loss can remain the structural information of boundary regions. Finally, saliency results in the fourth column illustrate the superiority of the combination of self-attention and expanded edge penalty loss, which surpasses the results of applying these two components independently.

Table II: Contrasting experimental results on five datasets

Dataset	Method	F_β	MAE
ECSSD[9]	SRM[16]	0.878	0.060
	DGRL[17]	0.903	**0.045**
	PiCANet[18]	**0.935**	0.047
	ResNet18-U$_s$Net+E^{2}PL	0.858	0.083
	ResNet50-U$_s$Net+E^{2}PL	0.896	0.052
PASCAL-S[10]	SRM[16]	0.864	0.067
	DGRL[17]	0.886	0.059
	PiCANet[18]	0.881	0.087
	ResNet18-U$_s$Net+E^{2}PL	0.875	0.062
	ResNet50-U$_s$Net+E^{2}PL	**0.902**	**0.054**
DUT-OMRON[11]	SRM[16]	0.688	0.073
	DGRL[17]	0.709	0.063
	PiCANet[18]	0.803	0.065
	ResNet18-U$_s$Net+E^{2}PL	0.781	0.074
	ResNet50-U$_s$Net+E^{2}PL	**0.815**	**0.059**
HKUIS[12]	SRM[16]	0.856	0.049
	DGRL[17]	0.882	0.037
	PiCANet[18]	0.881	0.078
	ResNet18-U$_s$Net+E^{2}PL	0.852	0.045
	ResNet50-U$_s$Net+E^{2}PL	**0.894**	**0.032**
DUTS[13]	SRM[16]	0.734	0.063
	DGRL[17]	0.768	0.057
	PiCANet[18]	0.860	**0.051**
	ResNet18-U$_s$Net+E^{2}PL	0.854	0.069
	ResNet50-U$_s$Net+E^{2}PL	**0.878**	0.055

2) Contrasting experimental results: This Sec. reports the contrasting experimental results, the specific performances are shown in Table II. In general, the proposed method achieves relatively performances on most of the testing datasets. For instance, the proposed ResNet50 based method achieves 0.016 increment in terms of F_β and 0.005 decrement in terms of MAE on PASCAL-S dataset. The proposed method achieves 0.018 increment in terms of F_β on DUTs dataset. The reasons mainly include two aspects, 1) the self-attention mechanism can encode the interaction relationships among different local regions in the image, and further improve the representation capability of the model, and 2) the proposed E^{2}PL can enhance the sensitivity of the network for pixels distributed in the boundary and near-boundary regions in the image, and improve the detailed information of the detection predicted saliency map. In addition, the proposed ResNet50 based method achieves better performance than that of ResNet18 based one. This demonstrates that, the intrinsic non-linear representation capability of the backbone is also important for the salient object detection task. Specifically, under the condition of sufficient training samples, deep models can dig out more latent information of the image, compared to shallow

ones.

IV. Conclusion

In this paper, we propose an expanded edge penalty loss based framework for salient object detection, which improves the detection performance through sufficiently considering the interaction relationships among different local regions in the image, and dynamically adjusting the penalty coefficient of each pixel according to its minimum distance to the boundary. Specifically, we incorporate self-attention mechanism into the bottleneck of the model to enhance the dependency relationships of different objects. In addition, we apply bigger penalty factors to pixels that closer to the boundaries, which dynamically adjusts the impacts of different pixels to the model optimization. Finally, the experimental results on five public and challenging datasets have achieved relatively satisfactory performances under different evaluation metrics, especially for pixels distributed in the boundary and near-boundary regions, which can reflect its superiority, compared to the contrasting methods. More research about salient object detection will be studied in future work, including, but not limited to the accuracy and the speed of the proposed method.

Acknowledgement

This work was supported by Youth Innovation Promotion Association CAS.

References

[1] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, and S.-F. Chang, "Mobile product search with bag of hash bits and boundary reranking," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3005–3012.

[2] Q. Hou, P. Jiang, Y. Wei, and M.-M. Cheng, "Self-erasing network for integral object attention," in *Advances in Neural Information Processing Systems*, 2018, pp. 549–559.

[3] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 071–10 081.

[4] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2. IEEE, 2004, pp. II–II.

[5] J. Fang, Y. Yuan, X. Lu, and Y. Feng, "Muti-stage learning for gender and age prediction," *Neurocomputing*, vol. 334, pp. 114–124, 2019.

[6] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "Ecffnet: effective and consistent feature fusion network for rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[7] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "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*, 2020, pp. 9060–9069.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[9] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155–1162.

[10] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

[11] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166–3173.

[12] D. Martin, C. Fowlkes, D. Tal, J. Malik *et al.*, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics." Iccv Vancouver:, 2001.

[13] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 136–145.

[14] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[15] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 733–740.

[16] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4019–4028.

[17] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3127–3135.

[18] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3089–3098.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.