Boundary Aware PoolNet for Salient Object Detection

Binghan Chen[1*], **31520211154024**
Mengzhao Chen[1*], **31520211154026**
Shaocong Chen[1*], **31520211154001**
Xunchao Li[1*], **31520211154010**
Yixin Qian[1*], **31520211154002**
[1]School of Informatics
Xiamen University, Xiamen, China
binghanchen@stu.xmu.edu.cn

Abstract

Salient Object Detection(SOD) aims to identify the most obvious and significant objects in the images. Traditional SOD methods rely too much on hand-crafted features, so they are gradually abandoned. Existing CNN-based methods are still facing challenges in complex scenes, such as blurry boundaries and unbalanced confidence. To solve the above problems, we propose Boundary Aware PoolNet(BAPoolNet) by improving PoolNet with deep supervision and hybrid loss. Deep supervision makes the shallow part of the network supervised more effectively and helps to locate the salient objects more accurately. Hybrid loss allows the network to predict the location and boundary of the salient objects from the perspective of the pixel/patch/image-level respectively. Experiment results show our BAPoolNet exceeds other methods without bells and whistles.

1 Introduction

Salient Object Detection(Gupta et al. 2020) can identify the most obvious and significant objects in the images. Unlike object detection, SOD divides the images into foreground pixels and background pixels, where the foreground represent the salient objects. SOD can extract effective information from massive images, which can provide convenience for other tasks in the computer vision, such as image classification, semantic segmentation, image super-resolution, image retrieval, and so on. Therefore, SOD can promote the development of computer vision.

The rise of deep learning provides a new paradigm for SOD. Researchers have proposed a large number of CNN-based SOD methods. CNN-based SOD methods are far superior to traditional methods in terms of salient object's location accuracy and pixel accuracy. However, existing methods also have some shortcomings, such as performance needs to be improved in complex backgrounds and model complexity needs to be reduced.

PoolNet is a U-Net structured network based on full convolutional network(FCN) and feature pyramid network(FPN). PoolNet contains a global guidance module(GGM) and a feature aggregation module(FAM). GGM uses short connection to achieve lossless transmission and further processing of the deep features. FAM ensures the effective fusion of deep and shallow features.

However, PoolNet has two drawbacks. First, PoolNet only uses the final output for gradient descent supervision and does not fully utilize the output of other layers. 2nd, PoolNet uses a binary cross-entropy loss function during training. So the pixels near the boundary of the salient object usually have a low confidence score, resulting in blurred boundary of the salient object.

Considering the above-mentioned shortcomings of PoolNet, we conduct the following improvements:

- **Deep supervision**. The losses of the multi-layer features in the top-down path of PoolNet are summed for gradient descent, which helps more accurately predict the locations and boundaries of salient targets.
- **Hybrid loss**. The binary cross-entropy loss(BCE loss), structural similarity loss(SSIM loss), and Intersection over Union loss(IoU loss) are combined together, which helps to predict the location and boundary of salient objects from pixel/local/global level.

2 Related Work

Most SOD methods adopt FCN as basic architecture to achieve saliency learning in an end-to-end manner. Typical architectures can be further classified into: single-stream network, multi-stream network, side-fusion network, and bottom-up/top-down network.

Now bottom-up/top-down network is the most popular architecture, such as PoolNet, so we take PoolNet as our baseline.

2.1 Single-stream Network

Single-stream network is the most standard architecture, which has a stack of convolution layers, intermediated with pooling and non-linear activation operations. It takes a whole image as input, and directly outputs a pixel-wise probabilistic map highlighting salient objects. UCF(Zhang et al. 2017) makes use of an encoder-decoder network architecture for finer-resolution saliency prediction. It incorporates a reformulated dropout in the encoder for learning

[*]These authors contributed equally.

uncertain features, and a hybrid up-sampling scheme in the decoder for avoiding checkerboard artifacts.

2.2 Multi-stream Network

Multi-stream network, typically consists of multiple network streams to explicitly learning multi-scale saliency features from multi-resolution inputs. Multi-stream outputs are fused to form a final prediction. MSRNet(Li et al. 2017) has three streams to process three scaled versions of input images. The three outputs are finally fused through a learnable attention module.

2.3 Side-fusion Network

Side-fusion network fuses multi-layer responses of a backbone network together for SOD prediction, making use of the inherent multi-scale representations of the CNN hierarchy . Side-outputs are typically supervised by the ground-truth, leading to a deep supervision strategy(Xie and Tu 2015). DSS(Hou et al. 2017) adds short connections from deeper side-outputs to shallower ones. Thus higher level features help lower side-outputs better locate salient regions, and lower-level features can enrich deeper side-outputs with finer details.

2.4 Bottom-up/top-down network

Bottom-up/top-down network refines rough saliency maps in the feed-forward pass by gradually incorporating spatial-detail-rich features from lower layers, and produces the finest saliency maps at the top-most layer . DGRL(Wang et al. 2018) purifies low-level features before combining them with the high-level ones. The combined features are refined recurrently in a top down pathway. The final output is enhanced by a boundary refinement submodule. PiCANet(Liu, Han, and Yang 2018) hierarchically embeds global and local pixelwise contextual attention modules into the top-down pathway of a U-Net(Ronneberger, Fischer, and Brox 2015) structure.

3 Boundary Aware PoolNet

Based on FCN(Long, Shelhamer, and Darrell 2015) and FPN, PoolNet achieves excellent trade-off between speed and quality. Considering the disadvantages of PoolNet, we plan to improve it from two aspects.

3.1 Overall Architure

FPN(Lin et al. 2017) can handle multi-scale objects. In the top-down path of PoolNet, shallower layers focus on locating salient objects while deep layers focus on refining details. However, PoolNet only supervises the prediction of the final layer. Therefore, As shown in Figure 1, we may supervise multiple layers in the top-down path. With this deep supervision strategy(Lee et al. 2014), the shallower and deeper layers of the network are effectively supervised simultaneously.

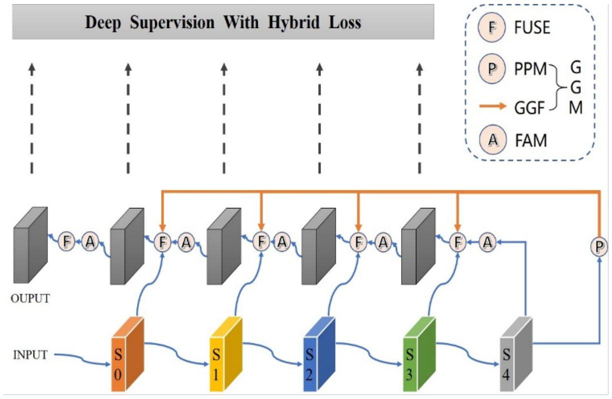

Figure 1: The model structure of Boundary Aware PoolNet

3.2 PoolNet

PoolNet uses BCE loss(Boer et al. 2005) in the training process, so the pixels near the boundary of the salient object usually have low confidence, which results the blurred boundary of the salient object. Inspired by BASNet(Qin et al. 2019), we combine BCE loss, SSIM loss(Wang, Simoncelli, and Bovik 2003), and IoU loss(Mattyus, Luo, and Urtasun 2017) into a hybrid loss function. With this hybrid loss function, PoolNet can learn the salient objects from the pixel/local/global level respectively.

3.3 Boundary Aware Training

We define our training loss as a summation of all output layers form the top-down path:

$$L = \sum_{k=1}^{K} \alpha_k l^k$$

where l^k is the loss of k-th output, α_k denotes a coefficient for l^k, we set $K = 5$ as shown in Figure 1.

For the sake of locating the salient object precisely and obtaining a high-quality salient object boundary, we will define it as a hybrid loss:

$$l^k = l^k_{BCE} + l^k_{SSIM} + l^k_{IOU}$$

where l^k_{BCE}, l^k_{SSIM} and l^k_{IOU} denote BCE loss (Boer et al. 2005), SSIM loss(Wang, Simoncelli, and Bovik 2003) and IoU loss (Mattyus, Luo, and Urtasun 2017), representing pixel, local and global level loss functions respectively. Combining the loss functions of the three levels, the binary cross entropy loss provides a gentle gradient for each pixel. The structure similarity loss is based on the image structure to make the boundary loss greater. The use of intersection is more than the loss. Focus on the salient objects in the image.

Binary Cross Entropy In binary classification and segmentation tasks, the binary cross entropy loss function (Binary Cross Entropy Loss, BCE Loss) (Boer et al. 2005) is the most commonly used loss function, and its formula is:

$$l_{BCE} = \sum_{(r,c)} [G(r,c)log(S(r,c)) + (1 - G(r,c)log((1 - S(r,c))]$$

where $G(r, c)$ is the label of the pixel and $S(r, c)$ is the probability that the pixel predicted by the algorithm is a salient object.

BCE Loss (Boer et al. 2005) is a pixel-level loss function. It doesn't consider the ground truth of surrounding pixels, which means the weights of foreground pixels and background pixels are the same. It is helpful for the convergence of all pixels.

Structural Similarity Loss SSIM, as known as Structural Similarity Loss (Wang, Simoncelli, and Bovik 2003) was used for image quality evaluation when it was proposed. For each image, it can extract the inner structure information, which means it can be used as part of the hybrid loss to obtain structure information of the salient object. Suppose we have label two vectors $X = \{x_i : i = 1, 2, ..., N^2\}$ and $Y = \{y_j : j = 1, 2, ..., N^2\}$. the Structural Similarity Loss can be define as follow:

$$l_{SSIM} = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where μ_x, μ_y and σ_x, σ_y are means and variances of x and y respectively, σ_{xy} is covariance. C_1 and C_2 are constants and set to be 0.01^2 and 0.03^2.

SSIM is a local-level loss function, which considers the local neighbors of each pixel and has a higher weight on the boundary. Even if the probability of being predicted as the foreground is the same, the loss of pixels near the boundary in the image is higher than the loss of pixels near the non-object boundary in the image.

Intersection over Union Loss Intersection over Union Loss (IoU Loss) (Mattyus, Luo, and Urtasun 2017) was used to measure the similarity of two sets when it was proposed, and was later used as a standard evaluation index for object detection and segmentation. In order to ensure that it can be differentiated, this paper defines its formula as:

$$l_{IOU} = 1 - \frac{\sum_{r=1}^{H}\sum_{c=1}^{W} S(r,c)G(r,c)}{\sum_{r=1}^{H}\sum_{c=1}^{W}[S(r,c) + G(r,c) - S(r,c)G(r,c)]}$$

where $G(r, c)$ is the label of the pixel and $S(r, c)$ is the probability that the pixel predicted as a salient object.

IOU is a global level loss function, which can give more attention to the salient objects in the image.

4 Experiments

4.1 Implementation Details

The PyTorch was used for algorithm implementation, training, and evaluation. The DUTS(Wang et al. 2017) dataset was used for model training and testing. The training set contains 10553 images and the test set contains 5019 images. The training set and test set contain a large number of scenarios for salient object detection.

In addition to the improvements made to PoolNet in this paper, other implementation details and experimental details of BAPoolNet (Boundary Aware PoolNet) are consistent with PoolNet. The BAPoolNet proposed in this paper is trained for 24 epochs, and the Adam optimizer is used during the training process (the weight decay value is 5×10-4, the initial learning rate is 5×10-5, and the learning rate is divided by 10 when the epoch is 15) . The model's backbone network is ResNet50, and its parameters are initialized by the corresponding model pre-trained on the ImageNet(Krizhevsky, Sutskever, and Hinton 2012) dataset. The other parameters in the model are initialized randomly by normal distribution. In terms of data enhancement, this paper only flips the training set images horizontally with a probability of 50%. During training and testing, the size of the model's input image remains unchanged. In addition to qualitative comparison, we use PR curve, F-measure curve, max F-measure and MAE to measure our model.

4.2 Qualitative Comparasion

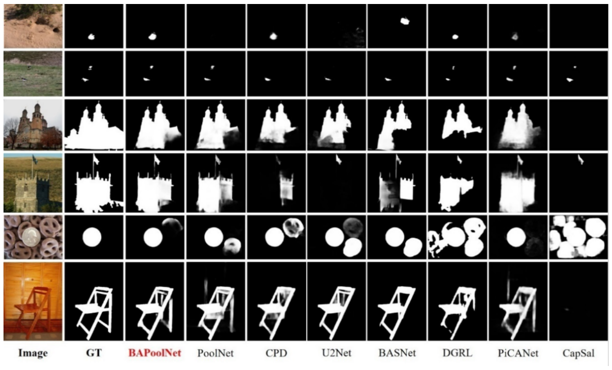

Figure 2: Qualitative comparison of the performance of different methods

In order to visually illustrate the performance difference between the BAPoolNet proposed in this paper and other methods, Fig.2 shows and compares the prediction results of these methods on different categories of images. The BAPoolNet proposed in this paper can capture various types of salient objects, and predict accurate location and boundary with high confidence.

The salient objects in the 1st and 2nd rows of Fig.2 are small objects. The salient object in the 1st row of images is very small, the difference between the foreground and the background is implicit, and the 2nd row of images has multiple salient objects. It can be seen that even when the difference between the salient object and the background is very implicit, BAPoolNet can more accurately capture the location of the salient object with a small area than the other methods.

The salient objects in the 3rd and 4th rows of Fig.2 have large area and rich details. As can be seen from the figure, BAPoolNet captures the most complete and large area of the salient objects compared to the other methods, and successfully captures their rich details.

The background of the 5th row of Fig.2 is complicated, and multiple objects in the background are very similar to the salient object structure. As can be seen from the figure,

Method	Conference	Backbone	Size(MB)	DUTS-TE	
				MAE↓	max F_β↑
CapSal	CVPR19	ResNet-101	-	0.063	0.826
PiCANet	CVPR18	ResNet-50	197.2	0.050	0.860
DGRL	CVPR18	ResNet-50	646.1	0.049	0.828
BASNet	CVPR19	ResNet-34	348.5	0.047	0.860
U2Net	CVPR20	RSU	176.3	0.044	0.873
CPD	CVPR19	ResNet-50	183.0	0.043	0.865
PoolNet	CVPR19	ResNet-50	260.0	0.040	0.880
BAPoolNet	-	ResNet-50	260.7	**0.035**	**0.892**

Table 1: Quantitative comparison of the performance of different models

BAPoolNet captures the salient object more accurately than the other methods and has fewer confusion.

The structure of the salient object in the 6th row of Fig.2 is very complex, and part of it is similar to the object structure in the background (vertical cylinder). As can be seen from the figure, BAPoolNet completely captures the entire salient object compared to the other methods and the predicted structural details are more accurate.

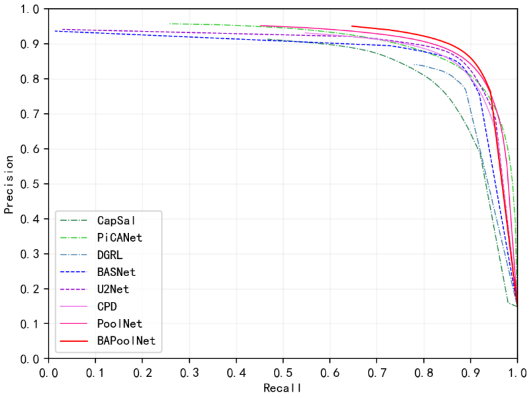

Figure 3: PR Curves

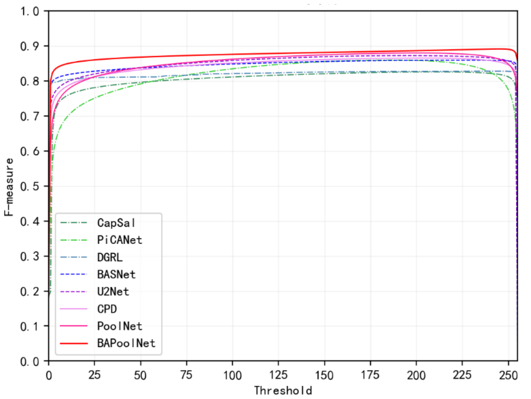

Figure 4: F-measure Curves

4.3 Quantitative Comparison

4.3.1 MAE & Max F_β

Tab.1 shows the quantitative results of BAPoolNet and other methods. In this table, the BAPoolNet's MAE and F-measure marked in red are the best, so it shows that the performance of the BAPoolNet proposed in this paper on the DUTS-TE dataset exceeds other methods.

4.3.2 PR Curves

In addition to quantitative comparison, we also compares the PR curves of BAPoolNet and other methods, as shown in Figure 3. It can be seen from the figure that the PR curve (red solid line) of the BAPoolNet on the DUTS-TE data set is significantly better than other methods.

4.3.3 F-measure Curves

We also compares the F-measure curves of BAPoolNet and other methods, as shown in Figure 4. It can be seen from the figure that the F-measure curve (red solid line) of BAPoolNet on the DUTS-TE dataset is significantly better than the previous method with the best performance.

5 Conclusion

In this paper, we propose Boundary Aware PoolNet to improve PoolNet with deep supervision and hybrid loss. On the one hand, deep supervision helps to locate the salient objects more accurately. On the other hand, hybrid loss allows the network to predict the location and boundary of the salient objects from the perspective of the pixel/patch/image-level respectively. Comparative experiments show that our method substantially exceeds other methods.

References

Boer, P.; Kroese, D. P.; Mannor, S.; and Rubinstein, R. Y. 2005. A Tutorial on the Cross-Entropy Method. *Annals of Operations Research*, 134(1): 19–67.

Gupta, A. K.; Seal, A.; Prasad, M.; and Khanna, P. 2020. Salient Object Detection Techniques in Computer Vision—A Survey. *Entropy*, 22(10): 1174.

Hou, Q.; Cheng, M.-M.; Hu, X.; Borji, A.; Tu, Z.; and Torr, P. H. 2017. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3203–3212.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105.

Lee, C. Y.; Xie, S.; Gallagher, P.; Zhang, Z.; and Tu, Z. 2014. Deeply-Supervised Nets. *Eprint Arxiv*, 562–570.

Li, G.; Xie, Y.; Lin, L.; and Yu, Y. 2017. Instance-level salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2386–2395.

Lin, T. Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4): 640–651.

Mattyus, G.; Luo, W.; and Urtasun, R. 2017. DeepRoadMapper: Extracting Road Topology from Aerial Images. In *2017 IEEE International Conference on Computer Vision (ICCV)*.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7479–7489.

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

Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; and Ruan, X. 2017. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 136–145.

Wang, T.; Zhang, L.; Wang, S.; Lu, H.; Yang, G.; Ruan, X.; and Borji, A. 2018. Detect globally, refine locally: A novel approach to saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3127–3135.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *Proc IEEE Asilomar Conference on Signals*.

Xie, S.; and Tu, Z. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, 1395–1403.

Zhang, P.; Wang, D.; Lu, H.; Wang, H.; and Yin, B. 2017. Learning uncertain convolutional features for accurate saliency detection. In *Proceedings of the IEEE International Conference on computer vision*, 212–221.