Refined BASNet for Salienct Object Detection

Sreeja Gaddamidi
University of Massachusetts Amherst
sgaddamidi@umass.edu

Jiye Choi
University of Massachusetts Amherst
jiyechoi@umass.edu

Abstract

Salient Object Detection (SOD) is to highlight object regions with more visual attention in images. As a research on Computer Vision has attracted much attention, SOD is one of crucial roles in the Computer Vision. In particular, significant progress has been made by applying Fully Convolution Neural Networks (FCNs) technique in SOD. However, many models still unable to highlight clear boundaries of objects in images. In this paper, we aimed to refine a SOD model that predicts outputs similar to ground truth images focusing on capturing boundaries of objects. We improved refinement in the Boundary-Aware Segmentation Network[15] by adding residual connections in the network, which improved the performance of the model and also reduced number of parameters further reducing training and inference time. We performed all our experiments on MSRA-B dataset[8]. As BASNet proposed, we adopted the hybrid loss function using Binary Cross Entropy (BCE) loss, Structual SIMilarity (SSIM) loss, and IoU loss in order to improve on three categories, overall pixels, local details and foreground. We presented performance analysis of both the models. We have achieved slight improvement which generated qualitatively closer outputs to the ground truth model than the BASNet model.

1. Introduction

1.1. Introduction

Human visual mechanism allocates attention to more salient and important regions, similar to this, Salient Object Detection (SOD) tries to identify visually attractive object regions in images. Early model of SOD are usually done in two stages, 1) detecting the salient object, 2)identifying the accurate region of that object. However, recent Deep Learning approaches combine both these stages. Identifying salient regions in images can help other high-level tasks by reducing the model parameters (if we use a pre-trained SOD, that features will not have to be separately learnt by the high level model), and improve efficiency. Even though the history of SOD is relatively short, SOD improves many computer vision tasks like image captioning, visual tracking, image and video segmentation, and so on.

After Fully Convolution Neural Networks(FCNs) appeared in SOD, FCNs is one of the most popular architectures of SOD. From traditional models to FCNs, SOD several significant achievements have been addressed in past years. However, most methods mainly focus on the global contrast of the image. As a result, capturing accurate boundaries of objects and fine structures in image are still remained as challenges. To overcome limitation of capturing boundaries, we implement a baseline SOD model, Boundary-Aware network,namely **BASNet** [15] which allows to predict clearer boundaries. The model proposed encoder-decoder architecture which allows the model to learn not only global overall contexts but also local detail contexts. In addition, To improve the drawback of Cross Entropy loss, which tends to perform poorly in differentiating boundary pixels, we use the hybrid loss, Binary Cross Entropy (BCE) loss, Structual SIMilarity (SSIM) loss, and IoU loss as suggested in [15]. The adoption allowed our model to detect pixel-wise details as well as contexts in foreground. Furthermore, we modified the architecture of **BASNet**,and evaluated its performance with MSRA-B [9] dataset.

1.2. Problem statement

As we mentioned in section 1.1, as most SOD methods focus on the global contrast of the image, they have lack of understanding on pixel-wise (local) details of images. It leads less accurate salient object detection results due to inaccurate boundary detection. To overcome this limitation, we adopt a model, **BASNet**, for Salient object detection, which achieves accurate salient object segmentation with high quality boundaries [14].

Qualitatively, we expect our output close to the ground truth(mask) images. Especially, we expect our output with higher accurate boundaries of objects. To evaluate our method, we use three widely used metrics, Precision-Recall(PR)curve, F-measure, Mean Absolute Error(MAE). We describe more details about these metrics in section 3.

This paper is organized as follows: Section 2 describes works related to Salient object detection. Our proposed model is given in the Section 3. Dataset, experiments and results are described in Section 4. Finally Section 5 concludes the paper.

2. Related Works

2.1. Early models

Most traditional methods compute a saliency map by identifying salient pixels or blocks from images, and then, merge those pixels into segment of the entire salient object. Early Borji *et al* [1] and Wang *et al* [18] well-defined and categorized classic models and history of SOD.

2.2. Deep Learning based models

While early models tend to rely on hand-craft features, as Convolutional Neural networks(CNNs), particularly, Fully Convolutional Neural networks (FCNs), has appeared, the need of hand-craft features has been removed. The main benefit of CNN-based is that these models achieve capturing highlight of salient regions and refining their boundaries since the large receptive fields contribute to identify the most salient region and the small receptive fields provide local information which helps refine the boundaries. [1] With these reasons, the numerous number of research is inspired by CNN-based models and it has achieved great performances. Tang *et al* [16] introduced a chained multi-scale fully CNNs. Liu *eta al*. [12] proposed a capsule method based on FCNs.

However, there are still difficulties to detect accurate bound aries. To overcome this, many models have been proposed. Hu *eta al*. [8] introduced a deep Level Set networks to produce compact and uniform saliency maps that output more accurate boundaries. Zhang *et al*.[21] proposed a novel network architecture namely symmetrical fully convolutional network. It allows to learn complementary visual features and predict accurate saliency maps under the guidance of lossless feature reflection. [21]. Tu *et al* [17] proposed ENFNet learning with edge guided feature to predict accurate SOD. Hou *et al*. [7] propoposed an approach using short connections to the skip-layer stuructures within Holisitcally-Nested Edge Detector(HED). [20]

2.3. Encoder-decoder Architectures

In order to capture low-level local contexts details as well as high-level global details, many models adopted encoder-decoder architecture. Hacha *et al*. [5] proposed deep convolutions symmetric encoder-decoder aiming to predict saliency maps of human's visual attention when the participants were solving quiz questionnaire images. For saliency prediction on dynamic scenes, LI *et al*. [11] developed a novel spatio temporal 3D convolutional encoder decoder network. Kroner *et al* [10] proposed an approach formed in encoder-decoder architecture that allows to highlight multi-scale features in parallel using multiple convolutional layer at different dilation rates.Our model also adopted this architecture to learn more details of objects for both the global contrast context and local pixel wise details.

3. Approach

3.1. Proposed Model

In this section, we describe our model which is inspired from **BASNet**[15] model. Since BASNet is a huge model with many parameters, we experimented and changed the model architecture to improve its computational efficiency and performance. Furthermore, our baseline model was not experimented or trained on MSRA-B dataset previously. The model consists of two modules, the prediction module and the refinement module as shown in the Figure 1 and Figure 2 respectively. The prediction module learns and produces saliency maps from input images and the refinement module learns residuals between the saliency map obtained from the previous module and ground truth and refines the saliency maps.

The **Prediction module** consists of input, encoder, bridge, decoder and coarse map output. This encoder-decoder architecture learns not only the high level global context but also low level local details. The Encoder contains total seven stages, an input convolution layer and six stages comprised of basic res-blocks [6]. A bridge stage is added in between the encoder and the decoder networks to further capture the global information from larger receptive fields and coarse features. The Decoder part has six stages with each stage having three convolution with a batch normalization and a ReLU activation function similar to encoder and bridge stages. In order to avoid over-fitting, last layer of each stage is supervised by ground truth i.e., saliency maps generated at each of the decoder stages are upsampled to the input size and used for calculating the final loss. The input of each decoder stage is the concatenation of its corresponding stage in the encoder and the upsampled output from its previous stage.

The **Refinement module** is the final module where we can get our final saliency map. It is designed as Encoder-Decoder network with a bridge stage similar to the prediction module. The refinement module consists of four stages in the Encoder and four stages in the Decoder with one convolution layer per each stage. Each layer has 64 filters of size 3×3 followed by a batch normalization and a ReLU activation function. The refinement module refines the predicted map by learning the residuals between the ground truth and the coarse saliency map. To further improve the refinement of region and boundary drawbacks in coarse saliency maps, we removed the concatenation connections

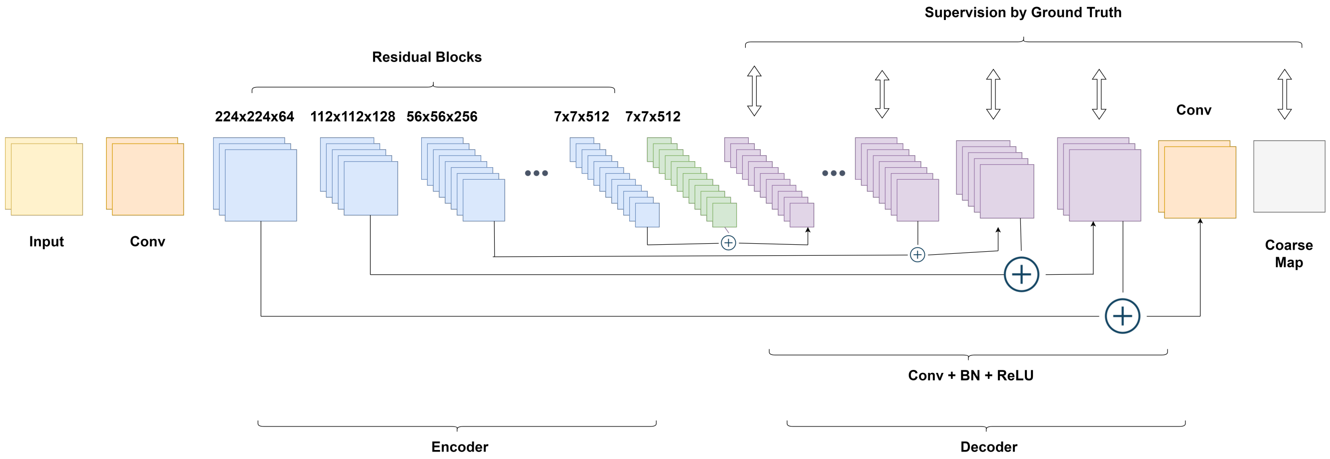

Figure 1. Architecture of the prediction module in our model

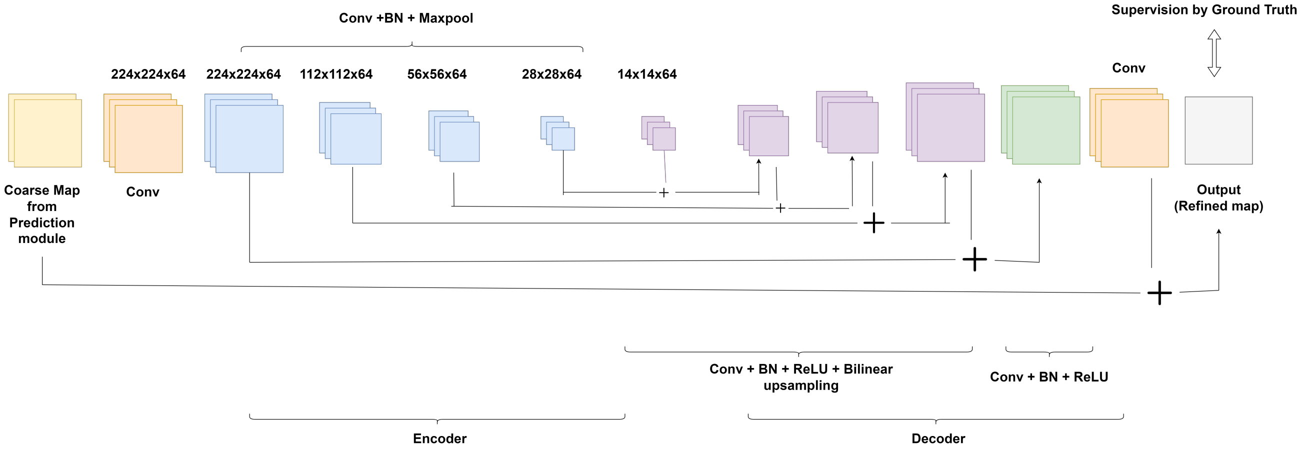

Figure 2. Architecture of the refinement module in our model

between encoder and decoder as in prediction module and replaced it with residual connections. As a result, this improved not only improved our model's performance but also reduced parameters of the model as compared to BASNet.

3.2. Loss Function

The training loss of the model is weighted sum of loss at each of the Prediction module' decoder outputs and the final output at the refinement module. The loss at each output is calculated as sum of Binary Cross Entropy (BCE) loss [4], Structual SIMilarity (SSIM) loss [19]and IoU loss [13]. This results in clear boundaries and high quality regional segmentation.

$$L = L_{BCE} + L_{SSIM} + L_{IoU} \tag{1}$$

BCE is used for binary classification and segmentation. It helps in convergence on overall pixels not only foreground but also background.

$$L_{BCE} = -\sum_{h,w}[G(h,w)\log(S(h,w)) + (1-G(h,w))\log(1-S(h,w))] \tag{2}$$

where S(h, w) is the predicted probability of being salient object and G(h, w) $\in \{0, 1\}$ is the ground truth label of the pixel (h, w).

SSIM captures the structural information in salient object. It helps to detect accurate boundary by focusing on local details and weighting more on the boundary. For two corresponding patches x and y of size N × N from the predicted probability map S and the binary ground truth mask G respectively, the SSIM of x and y is given as

$$L_{SSIM}(x,y) = 1 - \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3}$$

where $\mu x, \mu y$ and $\sigma x, \sigma y$ are the mean and standard deviations of x and y respectively, σxy is their covariance, C1 = 0.012 and C2 = 0.032 are used to avoid dividing by zero.

IoU is used for object detection and segmentation. It helps to emphasize the foreground in the input image.

$$L_{IoU} = 1 - \frac{\sum_{h=1}^{H}\sum_{w=1}^{W} G(h,w)(S(h,w)}{\sum_{h=1}^{H}\sum_{w=1}^{W}[G(h,w)+(S(h,w)-G(h,w)(S(h,w)]} \quad (4)$$

where S(h, w) is the predicted probability of being salient object and G(h, w) $\in \{0, 1\}$ is the ground truth label of the pixel (h, w).

4. Experiment

4.1. Dataset

We are using MSRA-B saliency detection dataset [9]. This dataset contains 5000 total images including natural scenes, animals, indoor and outdoor, etc. We did a 80/20 split on the dataset to obtain train and test set i.e., there are 4000 training images and 1000 test images. A sample pair of image and ground truth is shown in Table 1.

Image

Ground Truth

Table 1. Sample image and ground truth from MSRA-B

4.2. Evaluation Metrics

As we mentioned in section 1.2, we have three metrics to evaluate our results: Precision-Recall curve, F-measure, Mean Absolute Error.

Precision-Recall(PR) curve. For a saliency map, we binarize the map. Comparing the binary mask with ground-truth mask, we can get the precision and recall. Here, M is binary mask, and G is ground truth mask.

$$Precision = \frac{|M \cap G|}{|M|}, Recall = \frac{|M \cap G|}{|G|}. \text{ [2]}$$

Plotting precision-recall pairs at various thresholds from 0 to 1, we can compute PR curve.

F-measure. It is a harmonic mean of average precision and average recall. F-measure is fomulated as:

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall}$$

To weight precision more than recall, we set $\beta^2 = 0.3$ as suggested in [3].

Mean Absolute Error(MAE). It computes the average difference between a saliency map S and its ground truth G. Let h and w be the height of saliency map and the width of saliecny map respectively. Here, S(i,j) and G(i,j) are pixel coordinates corresponding to the height and width respectively on the saliency map and the ground truth. MAE is formulated as:

$$MAE = \frac{1}{h \times w}\sum_{x=1}^{h}\sum_{y=1}^{w}|S(x,y) - G(x,y)|$$

4.3. Experiments

We started our experiments by performing hyperparameter tuning on the BASNet model, we experimented with various learning rates, epsilon values in Adam optimizer. Due to the resource limitation we had to limit out batch size to 2. We later tied to make changes to the model architecture, we added dropout layer along with batch normalization layer, but it did not make much difference in the output however, the computation time increased by a small value. With the motivation that refinement layer is trying to get smoother predicted output using residuals from coarse map from prediction module, we changed all the concatenation connections to residual connections, this improved the results form the model, and also reduced the number of parameters by 2%, We later tried make similar changes to the prediction module, this reduced the number of parameters by almost 10% (nearly 8 million) but did not improve the performance of the model. We performed hyperparameter tuning on all the experimented architectures.

4.4. Results

We conducted all our experiments on eight-core PC with an intel i7-11800H 2.3 GHz CPU (with 32GB RAM) and a RTX 3060 GPU (with 6GB memory) is used for both training and testing. The BASNet model took 1255.5s for 1 epoch on 4000 images. On average the inference for a 256×256 image takes 0.127s, and our approach took 1098.6s for 1 epoch, and 0.113s for testing one image.

Results with finetuning pre-trained BASNet on MSRA-B data. We experimented pre-trained **BASNet** and tested with our dataset. This model outputs relatively accurate result close to the ground truth masks with the clear boundary. Our visualization result is shown in Table 2.

Table 3 indicates the average MAE, max F score and mean F score after we fine tuned **BASNet** for 4 epochs. The PR and F-measure curves on the result are shown in Table 4.

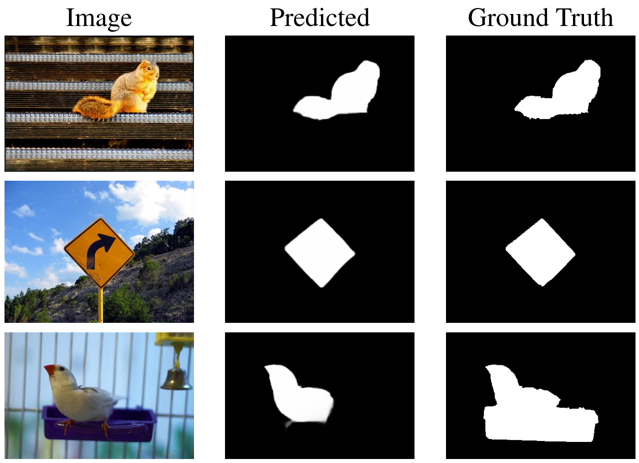

Table 2. Sample results trained with BASNet.

Model	avgMAE	maxF	meanF
BASNet	0.038	0.92	0.905

Table 3. Results after finetuning BASNet for 4 epochs on MSRA-B dataset

Results with our approach. This section we compare results for our approach and **BASNet**(not pre trained) after training for 5 epochs. As the model and the dataset are very huge, because of limitation in resources we were able to perform our experiments for 5 epochs. The AvgMAE of BASNet model after 2 epochs is 0.151, maxF is 0.762 and meanF is 0.644, and the AvgMAE for our model after 2 epochs is 0.119, maxF is 0.763 and meanF is 0.704. Comparing these with the results after 5 epochs from the Table 5 we can suggest that with increased epochs the model performance is also expected to be increased and eventually perform better than pretrained BASNet. As shown in Figure 3, mostly our model produced outputs closer to the ground truth images with clearer boundaries of objects than the results of the **BASNet** model. In particular, our images from the fourth row and last row have more accurate boundaries with our model. Moreover, for the first row right image, our model predicted more closer to the ground truth image than pre-trained **BASNet** even though it has blurred boundary. Using evaluation metrices mentioned in section 4.2, we compared our model to BASNet. The table of the results and illustrations of PR curves and F-measures are described in Table 5 and Table 6 respectively. From the Table 5, we have improved the average MAE by 0.003. In addition, we have improved mean F by 0.004 as well. In the model, mentioned as first model in the Table 5 we changed all the concatenation connections to residual connections, this reduced the model size by a huge factor but as seen from table did not improve the model.

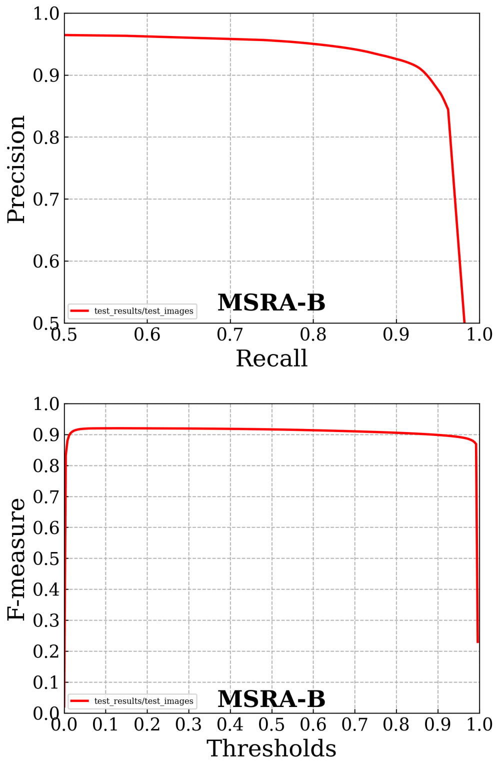

Table 4. Illustration of PR curves and F-measure curves of pre-trained BASNet

Model	avgMAE	maxF	meanF
BASNet	0.096	0.789	0.755
First Model	0.105	0.805	0.755
Our	0.093	0.812	0.759

Table 5. Results from our initial model and final model

5. Conclusion

In this paper, we aimed to refine a model to predict accurate saliency map with clearer boundary of objects from input images. Our model is inspired by a encoder-decoder architecture model, **BASNet**, which captures not only global high level context but also local pixel wise context. The approach the model proposed allows to get more accurate result with clear boundary of an object. Based on this idea, we changed some part of architecture of **BASNet** by removing the concantenation connection between the encoder and the decoder. Instead, we added residual connections which improves the performance of the model and reduces the train-

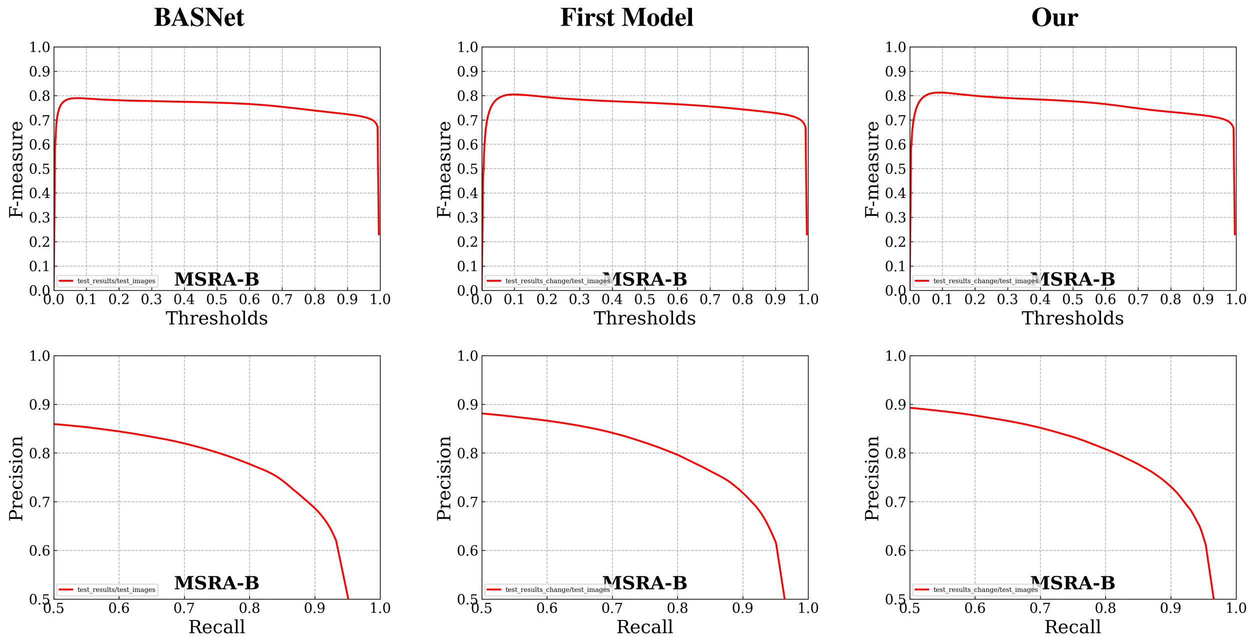

Table 6. Illustration of PR curves and F-measure curves of BASNet, First experiment model, Our model

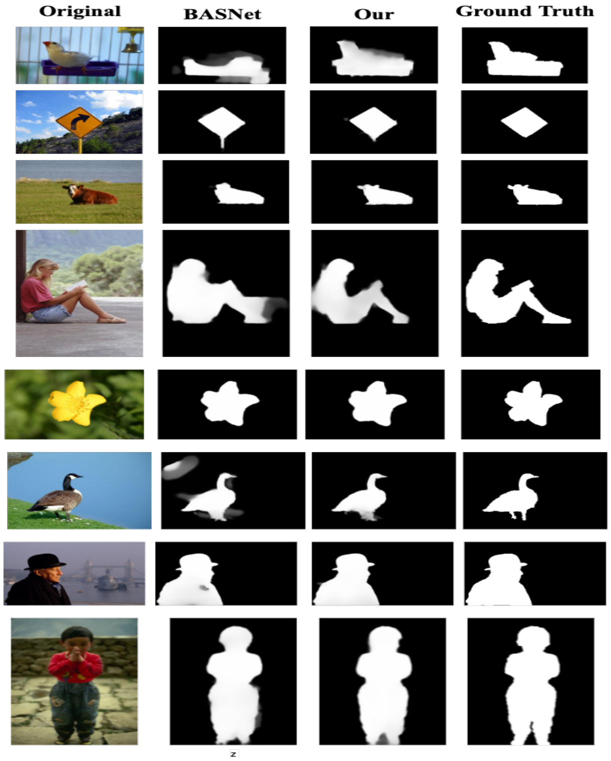

Figure 3. Results from BASNet model and Our model with our dataset

ing and testing time .

As results of our experiment, we have slightly improved our saliency map that predicted more similar to their ground truth images than our baseline model **BASNet** without pretrain. By replacing concatenation connection with residual connections, we have reduced the number of parameters of the model. Therefore, it has the less computation time. However, our model still produces some poor results with blurred boundaries. As future suggestions, we suggest to increase the number of epochs so that the model can improve its predicted results and it is expected to give better results compared to pretrained BASNet.

References

[1] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational Visual Media*, 5(2):117–150, 2019. 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, Dec 2015. 4

[3] M. Cheng, N. J. Mitra, X. Huang, P. S. Torr, and S. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 37(03):569–582, mar 2015. 4

[4] 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, Jan. 2005. 3

[5] Tomasz Hachaj, Anna Stolińska, Magdalena Andrzejewska, and Piotr Czerski. Deep convolutional symmetric encodermdash;decoder neural networks to predict studentsrsquo; visual attention. *Symmetry*, 13(12), 2021. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 2

[7] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[8] Ping Hu, Bing Shuai, Jun Liu, and Gang Wang. Deep level sets for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2

[9] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2083–2090, 2013. 1, 4

[10] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020. 2

[11] Hao Li, Fei Qi, and Guangming Shi. A novel spatio-temporal 3d convolutional encoder-decoder network for dynamic saliency prediction. *IEEE Access*, 9:36328–36341, 2021. 2

[12] Yi Liu, Qiang Zhang, Dingwen Zhang, and Jungong Han. Employing deep part-object relationships for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[13] Gellert Mattyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3

[14] Xuebin Qin, Deng-Ping Fan, Chenyang Huang, Cyril Diagne, Zichen Zhang, Adrià Cabeza Sant'Anna, Albert Suàrez, Martin Jägersand, and Ling Shao. Boundary-aware segmentation network for mobile and web applications. *CoRR*, abs/2101.04704, 2021. 1

[15] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019. 1, 2

[16] Youbao Tang and Xiangqian Wu. Salient object detection with chained multi-scale fully convolutional network. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 618–626, New York, NY, USA, 2017. Association for Computing Machinery. 2

[17] Zhengzheng Tu, Yan Ma, Chenglong Li, Jin Tang, and Bin Luo. Edge-guided non-local fully convolutional network for salient object detection. *CoRR*, abs/1908.02460, 2019. 2

[18] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, and Haibin Ling. Salient object detection in the deep learning era: An in-depth survey. *CoRR*, abs/1904.09146, 2019. 2

[19] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. *Conference Record of the Asilomar Conference on Signals, Systems and Computers*, 2:1398–1402, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers ; Conference date: 09-11-2003 Through 12-11-2003. 3

[20] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. *CoRR*, abs/1504.06375, 2015. 2

[21] Pingping Zhang, Wei Liu, Huchuan Lu, and Chunhua Shen. Salient object detection by lossless feature reflection. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1149–1155. International Joint Conferences on Artificial Intelligence Organization, 7 2018. 2