Enhancing Saliency Object Detection through Attention and U2Net Configurations

Aadesh S. Varude [1] **Mihir M. Kulkarni** [1] **Anagha R. Dangle** [1] **Vaibhav N. Kadam** [1]

Abstract

Owing to the powerful feature extraction capabilities of convolutional neural networks (CNNs), deep learning methods have recently made significant strides in the field of salient object detection. Despite this progress, many current deep models struggle to effectively learn informative contextual features, resulting in suboptimal performance when faced with complex scenes. In our proposed approach, we explore the application of the U2Net-attention framework to enhance this performance of Salient Object Detection (SOD). Specifically, we analyze standard and modified attention blocks in conjunction with various U2Net architectures, such as attention-inside, attention-outside, and a combination of both. To provide a comprehensive evaluation of the proposed models, we employ a qualitative assessment of SOD using all attention-enhanced models and the attention maps generated. Please find our video presentation **link**

1. Introduction

Saliency refers to the quality of being prominent or visually striking in an image or scene. In the context of computer vision, saliency detection is a method used to identify and highlight the most relevant objects or regions within an image. Early salient object detection (SOD) algorithms primarily relied on heuristic priors, such as color, contrast, and texture, to generate saliency maps. However, these hand-crafted features struggle to capture high-level semantic relations and context information, limiting their ability to detect complete salient objects. Recently, deep learning-based methods, particularly convolutional neural networks (CNNs), have significantly improved SOD performance by fusing multi-scale features and detecting prominent objects from global and coarse perspectives. Despite these advances, challenges remain in achieving accurate and complete saliency detection, especially in cases involving complex object structures, messy backgrounds, low contrast between foreground and background, or drastic changes in object contours.

Our proposed method integrates saliency detection and attention mechanisms to boost salient object detection performance and efficiency. By incorporating saliency information as a weighted input or guidance map, the model can prioritize significant regions, resulting in better object localization, reduced computational complexity, and increased interpretability.

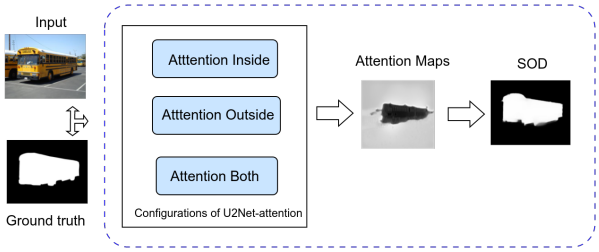

Figure 1. Overview of U2Net-Attention Salient Object Detection.

1.1. Research contributions

The specific contribution of the project is as follows:

- Implement standard and modified attention blocks.
- Analyze performance for a fixed training dataset and epochs
- Comparative assessment of Attention-augmented U2Net and regular U2Net
- Examine attention maps generated at different stages of the training process.

2. Related works

A comprehensive survey of traditional SOD algorithms reveals that most recent research in salient object detec-

[1]Worcester Polytechnic Institute. Correspondence to: Aadesh S. Varude <avarude@wpi.edu>, Mihir M. Kulkarni <mmkulkarni@wpi.edu>, Anagha R. Dangle <adangle@wpi.edu>, Vaibhav N. Kadam <vkadam@wpi.edu>.

CS/DS 541 Deep Learning Final Project, Worcester Polytechnic Institute.

tion has been dominated by CNN-based methods due to their powerful feature extraction capabilities. Several approaches have been proposed to address the challenges associated with SOD, including fusing local and global contexts, utilizing attention mechanisms, and employing multi-mechanism fusion networks. Oktay *et al*., [1] introduced an innovative attention block for the UNet model, significantly enhancing the performance and accuracy of medical image segmentation. Building upon this, Qin *et al*., [2] developed the U2-Net, a robust and efficient architecture for salient object detection in images. The nested U-structure of the U2-Net enables deeper feature extraction, resulting in superior performance compared to other state-of-the-art models.

Ren *et al*., [3] proposed an approach that merges local and global contexts to improve SOD performance. Although this method surpasses traditional SOD techniques, it faces difficulties in incorporating spatial contexts and accurately locating salient objects in complex cases. More recent SOD approaches involve the application of multi-mechanism fusion networks, exemplified by the U2Net-ECA-AS model. This model combines residual structures with the U2Net model to enable automatic learning of crack features.

3. Proposed methodology

In our proposed U2Net architecture, we introduce a novel attention block to optimize its performance. Precisely, we present two distinct types of attention blocks and integrate each one of them in three different configurations as follows:

1. The first approach involves the construction of a U2Net comprising a block similar to Attention UNet (a single block of U2net architecture with attention is shown in Fig 3), albeit with differences in its configuration.

2. In the second configuration of U2Net, we incorporate the output of each block with attention while upsampling the encoder part. This process entails substituting the addition operation in the conventional U2Net with an attention block (Fig 4).

3. The final configuration is a combination of both the above mention configurations i.e. we replace all the UNet-like blocks in Fig4 with Attention UNet block Fig3.

3.1. Attention Blocks

3.1.1. Standard Attention Block

The attention block Fig2 described below is based on the attention U-Net paper [1]. It takes the encoded feature maps 'x' from the encoder pathway and the upsampled feature maps 'g' from the decoder pathway, which contain more encoded features. The attention block generates a weight matrix, this weight matrix is multiplied by the input feature maps 'x' to obtain the attention-weighted feature maps.

3.1.2. Modified attention block

In this adapted block Fig2, we first downsample the height and width of 'x'. We then concatenate them with 'g', and apply the same processing functions as in the previous attention block. The resulting attention weights are upsampled and multiplied with 'x', before being fed into the network for further processing. The primary aim of doing this is to preserve the 'g' features and let the attention get focused more on the center part of the image.

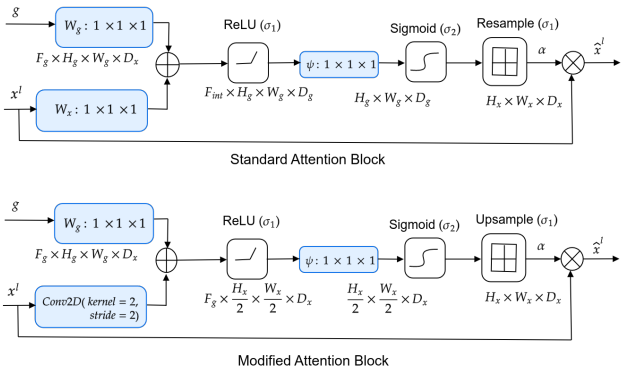

Figure 2. Attention Blocks

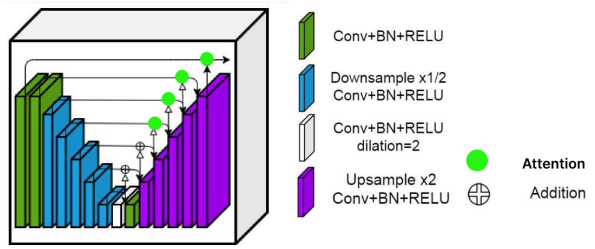

Figure 3. Attention UNet like block.

4. Experiments

4.1. Attention Maps

During training, we generated and analyzed attention maps by obtaining the psi output (i.e., attention maps) from the attention blocks after every 1000 iterations, depending on the block configuration shown in Fig 2. Fig 6 illustrates three different outputs: the first column of images represents the output after 1000 iterations, the middle column after 3000 iterations, and the final column after 5000 iterations. Our observations indicate that the attention maps

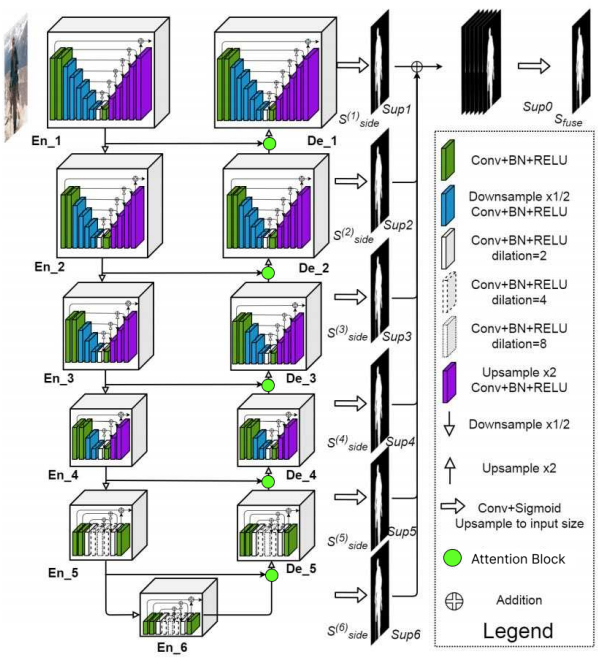

Figure 4. Attention U2Net Outside.

effectively guide the model to concentrate on significant parts of the image, such as cars and buses, as demonstrated in the images.

4.2. Data description

For training and testing saliency detection models, we utilized a subset of the DUTS dataset consisting of 5,716 and 1,000 images, respectively. The dataset's diverse content, including objects, scenes, and backgrounds from various sources such as ImageNet DET training/val sets and the SUN dataset, provides a realistic and complex environment for effective model training and evaluation.

4.3. Evaluation metrics

We adopt the standard MAE and F-measure (with β^2 is equal to 0.3) as evaluation terms which are described in U2Net paper [2]. These are defined as follows:

$$F_\beta = \frac{\left(1+\beta^2\right) \cdot \text{ Precision } \cdot \text{ Recall}}{\beta^2 \cdot \text{ Precision } + \text{ Recall}}$$

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

where $P(i,j)$ and $G(i,j)$ represent the values in the predicted map and ground truth, respectively.

4.4. Implementation Details

Adhering to the training process and data augmentation techniques outlined in the U2Net paper, we preprocess each image by resizing it to 320x320, applying random vertical flipping, and cropping to 288x288. In the absence of a backbone, we employ the Adam optimizer for network training, utilizing default hyperparameters (learning rate lr=0.001, betas=(0.9, 0.999), eps=1e-08, weight decay=0). The training proceeds until the loss converges without the use of validation, leveraging a batch size of 32 across 30 epochs. During the testing phase, input images undergo resizing to 320x320 before being fed into the network to generate saliency maps, which are subsequently resized to match the original image dimensions. The entire training process is executed on an Nvidia A100 GPU.

5. Results and Analysis

In Figure 6, the last image demonstrates that the right person's sword is not visible in U2Net, whereas AttentionU2net with standard blocks successfully captures both swords. The 'both configuration' performs marginally better than the other two approaches. However, the modified configuration struggles to show the upper part of the sword, as it enforces attention on the centre part of the image feature at every layer.

For the first and second images of the dog and elephant, respectively, only minor differences are observed among all the images. The standard attention in the 'both configuration' provides more detail in the dog's ears and the elephant's trunk. The modified attention does not perform as well in these cases since its attention is more focused on the center of the image, which is more evident in the third image where the sword in the center is visible.

Table 1 presents the values for MAE and F-beta, here we can observe that the standard attention block outperforms the standard U2Net whereas the modified block performance is comparable to U2net as the modified block focus on the centre of the image hence for this dataset the results are comparable to U2Net. Further, the loss graphs in Fig 7 compare U2Net to three configurations of the Modified Block Attention U2Net. Attention U2Net demonstrates better performance in terms of loss compared to U2Net. After 30 epochs, U2Net has a loss value of 0.9753, while the attention configurations achieve lower losses of 0.9244, 0.8991, and 0.9313. In a similar manner, the standard attention U2Net marginally outperforms U2Net loss.

When plotting the loss graph, it reveals that the modified attention block demonstrates better training performance compared to the standard attention block, as evidenced by a faster decrease in loss initially. However, after 25 epochs, their performance converges and becomes similar. A slight

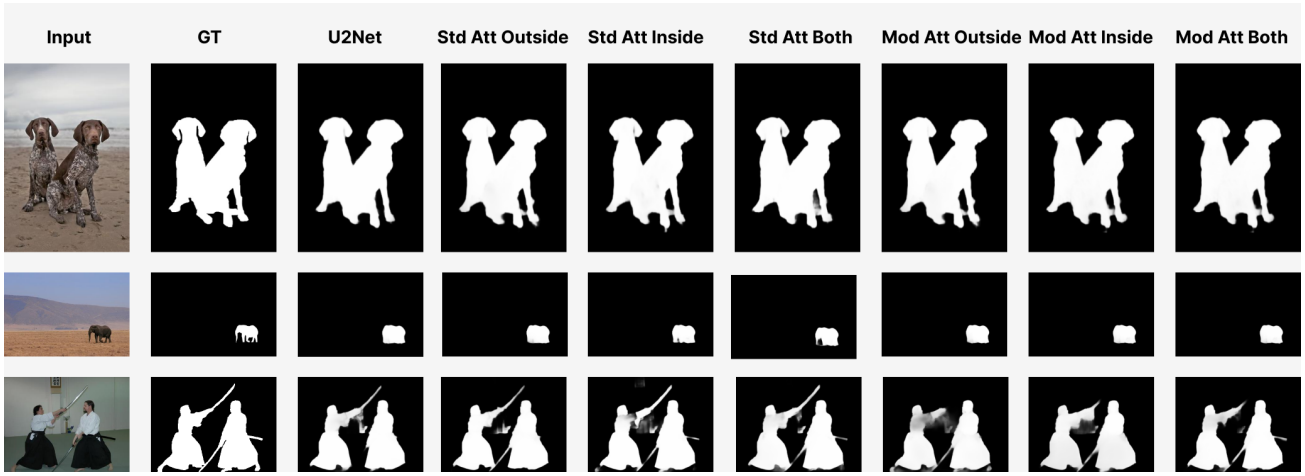

Figure 5. Qualitative comparison of models for SOD.

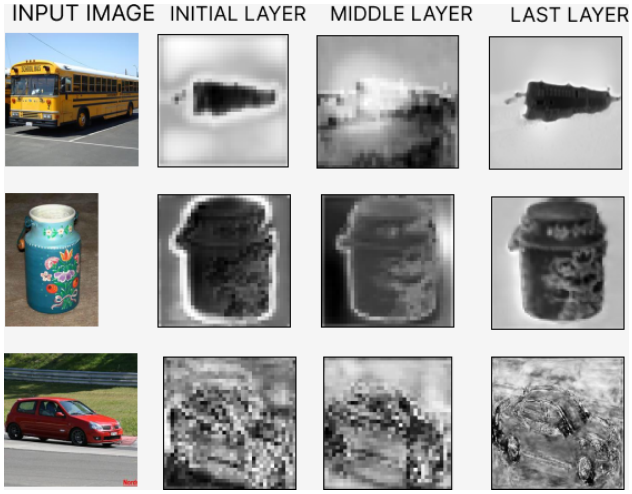

Figure 6. Generated attention maps.

advantage is observed for the modified attention block over the standard attention block. The best results are highlighted in bold.

Table 1. Comparative Evaluation Metrics for Different Models

	Standard		**Modified**	
	F-beta	MAE	F-beta	MAE
U2Net	0.865	0.045	–	–
U2Net_Inside	0.875	0.043	0.8348	0.051
U2Net_Outside	0.868	0.045	0.8196	0.053
U2net_both	**0.887**	**0.039**	0.8421	0.050

Figure 7. Comparison of training loss for different U2Net with modified attention models.

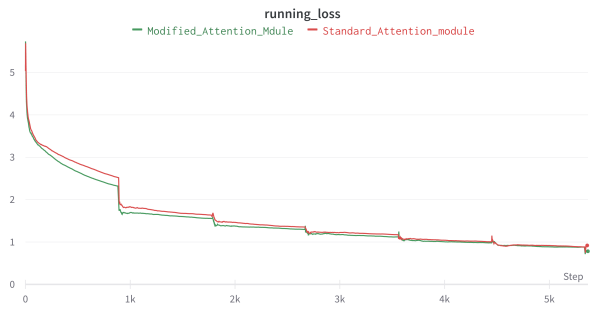

Figure 8. Comparison of training loss modified and standard attention block for U2Net both configuration.

6. Conclusion and Future work

Based on the findings of the study, it is apparent that the attention mechanism has proven to be a highly effective tool in directing the model's focus toward pertinent objects within the scene. The modified approach to attention involves downsampling the original feature vector, which has the potential to enhance the model's focus on the central features of the image. Moreover, it is anticipated that the attention mechanism will continue to evolve and become more sophisticated as the model undergoes further training epochs.

References

[1] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[2] 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.

[3] Qinghua Ren, Shijian Lu, Jinxia Zhang, and Renjie Hu. Salient object detection by fusing local and global contexts. *IEEE Transactions on Multimedia*, 23:1442–1453, 2021.

[4] Xuebinqin. Xuebinqin/u-2-net: The code for our newly accepted paper in pattern recognition 2020: "u2-net: Going deeper with nested u structure for salient object detection.".

[5] LeeJunHyun. Leejunhyun/image-segmentation: Pytorch implementation of unet, r2unet, attention unet, and attention r2u-net.