PUB-SalNet: A Pre-trained Unsupervised Self-Aware Backpropagation Network for Biomedical Salient Segmentation

Feiyang Chen[1,+], Ying Jiang[1,+], Xiangrui Zeng[1], Jing Zhang[2], Xin Gao[3] and Min Xu[1,*]

[1] Compututational Biology Department,
Carnegie Mellon University

[2] Department of Computer Science,
University of California Irvine

[3] Computational Bioscience Research Center (CBRC),
Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division,
King Abdullah University of Science and Technology

[+] Equal contribution
[*] Corresponding author

Abstract

Salient segmentation is a critical step in biomedical image analysis, aiming to cut out regions that are most interesting to humans. Recently, supervised methods have achieved promising results in biomedical areas, but they depend on annotated training data sets, which requires labor and proficiency in related background knowledge. In contrast, unsupervised learning makes data-driven decisions by obtaining insights directly from the data themselves. In this paper, we propose a completely unsupervised self-aware network based on pre-training and attentional backpropagation for biomedical salient segmentation, named as PUB-SalNet. Firstly, we aggregate a new biomedical data set from several simulated Cellular Electron Cryo-Tomography (CECT) data sets featuring rich salient objects, different SNR settings and various resolutions, which is called SalSeg-CECT. Based on the SalSeg-CECT data set, we then pre-train a model specially designed for biomedical tasks as a backbone module to initialize network parameters. Next, we present a U-SalNet network to learn to selectively attend to salient objects. It includes two types of attention modules to facilitate learning saliency through global contrast and local similarity. Lastly, we jointly refine the salient regions together with feature representations from U-SalNet, with the parameters updated by self-aware attentional backpropagation. We apply PUB-SalNet for analysis of 2D simulated and real images and achieve state-of-the-art performance on simulated biomedical data sets. Furthermore, our proposed PUB-SalNet can be easily extended to 3D images. The experimental results on the 2d and 3d data sets also demonstrate the generalization ability and robustness of our method.

Unsupervised learning; Saliency segmentation; Biomedical image processing; Pre-trained methods

1 Introduction

Biomedical image segmentation has drawn attention due to its widespread applications in computer-aided diagnosis and intelligent medical programs [1], among which salient segmentation refers to pixel-level annotation for regions of interest (e.g. organelle, substructures, and lesions) on biomedical images (e.g. Cellular Electron Cryo-Tomography (CECT) 3D images, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI)). An example of semantic segmentation [2] and salient segmentation

with unsupervised methods on CECT images is shown in Figure. 1. We discover that traditional segmentation cannot handle CECT images properly due to heterogeneous salient objects, various SNR and low resolution of the data set, while salient segmentation can efficiently and effectively capture objects of interest to people and mask off irrelevant regions. However, accurate salient segmentation is challenging due to different shapes and sizes of the region of interest and diversity of images produced by various biomedical imaging devices.

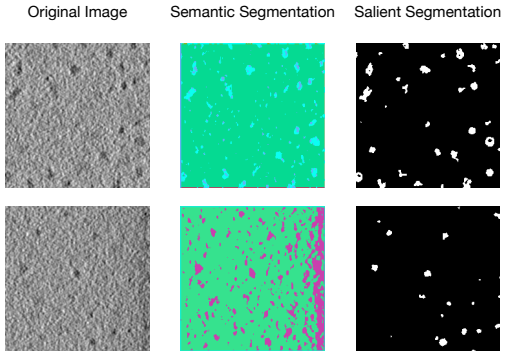

Figure 1: An unsupervised example of semantic segmentation and salient segmentation on CECT images.

Recently, current salient segmentation methods on biomedical images [3] mostly use improved convolutional neural networks (CNNs) and its variants, which rely on supervised learning that require labor-intensive annotation of large data sets by experts. Furthermore, such methods are strictly limited by the quality of data sets. They are vulnerable to problems of model generalization and extensibility when facing adversarial training. In contrast, unsupervised learning not only derives insights directly from the data but also uses them to make data-driven decisions. It is more practical and robust for some complex tasks in biomedical areas, such as saliency detection and image segmentation. Therefore, it is crucial to work out an effective unsupervised salient segmentation method for biomedical images.

However, in recent years few works have looked into unsupervised salient segmentation on biomedical images due to the complexity of biological structures. [4] systematically reviewed current unsupervised models for biomedical image segmentation. More recently, a unified unsupervised approach based on clustering and deep representation learning was designed by [5]. [6] proposed a teacher-student unsupervised learning system. The teacher performs unsupervised object discovery and at the same time, multiple students with various network architectures are trained to ensure a better diversity. These methods are either based on clustering with posterior selection or dependant on carefully optimized network hyper-parameters, which indicates that they are in need of human interference and are not completely unsupervised.

In this paper, we propose the PUB-SalNet, which is a completely unsupervised network based on pre-training and attentional backpropagation. In order to build a high-performance automatic biomedical salient segmentation model to improve computer-aided diagnosis and other biomedical image analyzing tasks, we design a processing pipeline with three major modules: 1). A pre-training method specially designed for biomedical images; 2). The U-SalNet model, which selectively attends to salient objects via fusing two attention mechanisms into U-Net; 3). An unsupervised self-aware backpropagation method based on superpixels, which iteratively updates the parameters of U-SalNet. Through extensive experiments, we demonstrate that the proposed PUB-SalNet outperforms all existing unsupervised methods and achieves state-of-the-art performance on the simulated biomedical data sets. Furthermore, PUB-SalNet can also be easily extended to 3D images. The experimental results on 2d and 3d biomedical data sets show generalization ability and robustness of the proposed method. Our main contributions are summarized as follows:

- We propose a novel PUB-SalNet model for biomedical salient segmentation, which is a completely unsupervised method based on pre-training and attentional backpropagation.
- We aggregate a new biomedical data set called SalSeg-CECT, featuring rich salient objects, dif-

ferent SNR settings and various resolutions, which also serves for pre-training and fine-tuning for other complex biomedical tasks.

- Extensive experiments show that the proposed PUB-SalNet achieves state-of-the-art performance and can be easily extended to 3D images, demonstrating generalization ability and robustness of our method.

The rest of the paper is organized as follows. We review related works in the next section. In Section 3 we describe the PUB-SalNet and the completed processing pipeline for salient segmentation. Quantitative and qualitative experiments are discussed in detail in Section 4. Lastly, in Section 5 we conclude our method and future works.

2 Related Work

2.1 Pre-trained methods in biomedical images

Many works have shown that the pre-training method along with adequate fine-tuning is superior to training from scratch, also being less dependent on the size of the training set [7]. However, in the biomedical area, it is extremely challenging to build an effective pre-trained model due to the difficulty of data acquisition and annotation by experts. Only a few pre-trained models are related to biomedical images, among which the most famous is [8]'s MedicalNet. They collect data from some medical challenges and build the 3DSeg-8 data set with diverse modalities to train MedicalNet, and then transfer it to other segmentation and classification tasks and achieve state-of-the-art performance. But their work is demanding of high-quality medical images with various scan regions, target organs, and pathologies, which is not applicable to cellular image analysis. With recent breakthroughs in CECT 3D imaging technology [9], it is now possible for researchers to deeply look into and comprehend the macromolecular structure of the cell, which is more meaningful to biomedical fundamental studies. In order to solve the challenges above, we present a pre-trained model intended for biomedical images featuring a low SNR, low resolution and rich salient objects.

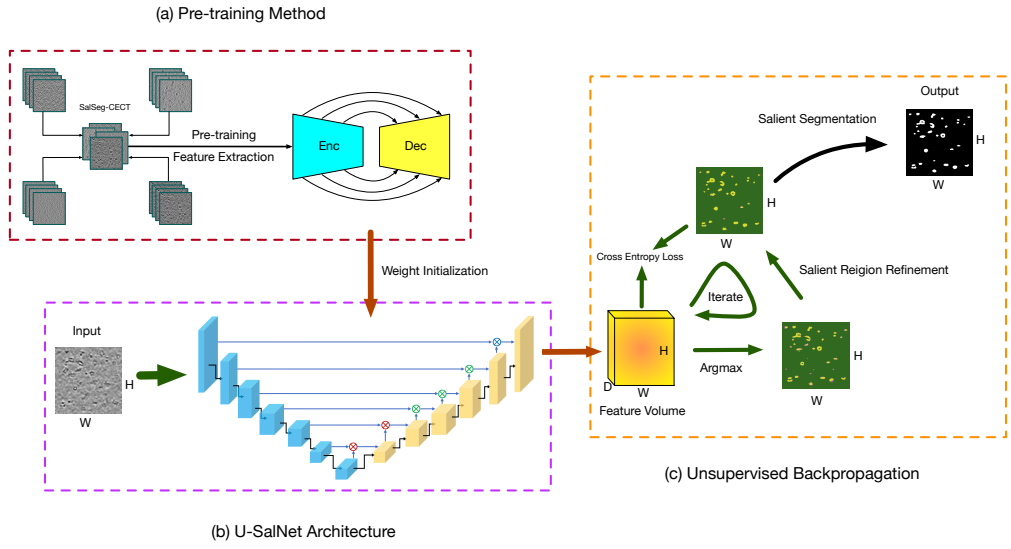

Figure 2: The overview framework of our proposed method. The processing pipeline consists of three main steps: (a). Pre-training on the SalSeg-CECT data set; (b). Prediction using the U-SalNet model; (c). Unsupervised attentional backpropagation iterating on single images. The Enc and Dec stands for the encoder and decoder. The ⊗, ⊗ and ⊗ denotes the global attention mechanism, local attention mechanism and convolutional decoding, respectively.

2.2 Unsupervised biomedical image segmentation

Unsupervised segmentation for biomedical images is very promising yet challenging. [10] concatenates two fully convolutional networks together into an autoencoder. The encoder produces a k-way pixel-wise prediction while the decoder reconstructs the image. Both the normalized cutting loss of the segmentation map and the reconstruction error are jointly minimized during training. However, we find the training very difficult to converge due to the inappropriate combination of the loss functions. [11] proposes an unsupervised skin lesion segmentation method to combine color and brightness saliency maps into enhanced fusion saliency. Although it shows good results on dermoscopy images, it relies too much on coloring and contrast information and cannot effectively perform salient segmentation on grey-scale images (such as CECT). [2] optimizes the pixel labels using a common CNN network while their parameters are iteratively updated by gradient descent to unify labels within a superpixel. However, their model is trained every time on a natural image and displays randomness in predictions. It cannot utilize knowledge from a large training set. Also, it does not work well when applied to noisy biomedical images. In order to solve these problems and adapt to salient segmentation on biomedical images, we load weights pre-trained on an assembled biomedical data set and present the U-SalNet model to extract significant features.

2.3 Salient segmentation

Current salient segmentation methods on biomedical images treat the problem as a binary (namely salient and non-salient) segmentation task, identifying the label (foreground or background) of pixels. Recently, [12] proposes an attention gate (AG) model to focus on significant targets for medical image analysis. Through AG, the model can ignore the background in images while mark out the salient objects meaningful to the medical segmentation task. Specifically, AG extracts local information from a denser layer in the decoder and then uses it as a gating signal for the current layer to combine low-level features from the encoder network with the decoded features. However, it does not explicitly make use of global interference when producing attention maps, resulting in their saliency maps ignoring global contrast mechanisms. Different from their work, our proposed U-SalNet model applies both global and local self-attention to better integrate saliency information.

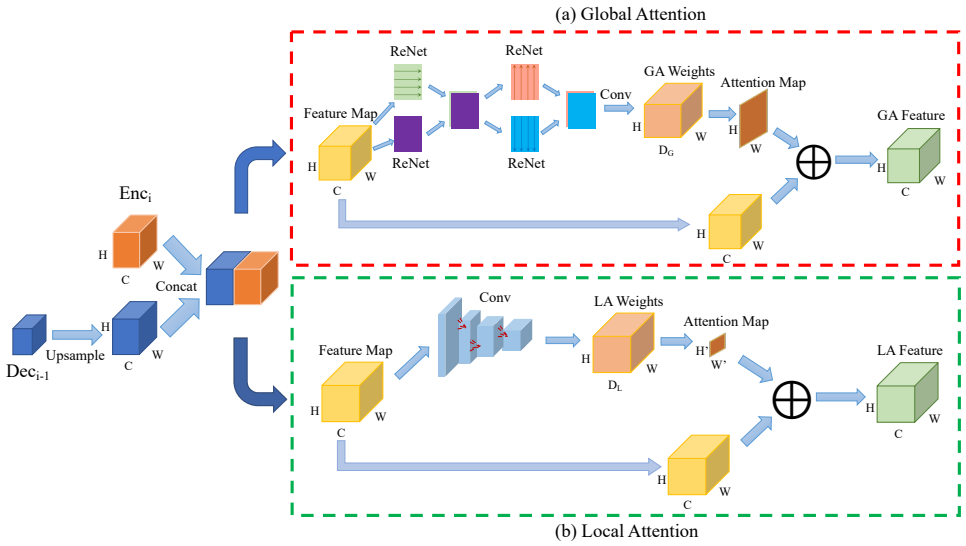

Figure 3: The architecture of our U-SalNet model. (a) Global Attention and (b) Local Attention corresponds to ⊗ and ⊗ from Figure. 2(b), respectively. GA and LA stands for Global Attention and Local Attention. Conv means the convolution operation. ⊕ stands for weighted summation over the feature map.

3 PUB-SalNet

In this section, we describe our proposed PUB-SalNet. The goal of our work is to build a high-performance automatic salient segmentation model appropriate for biomedical tasks without ground-truth labels to improve computer-aided diagnosis. To reach this target, we design a processing pipeline with three major steps, as shown in Figure. 2. In the first step, we aggregate a new biomedical data set called SalSeg-CECT. We then pre-train a deep feature extraction model specially designed for biomedical images, which can be used as a backbone module to initialize model parameters and to boost other tasks without data annotations. In the second step, we present the U-SalNet model on the basis of the U-Net architecture to learn to selectively attend to salient segmentation objects. The network includes global attention and local attention to facilitate learning saliency through global contrast and local similarity. In the last step, we jointly refine the salient regions together with feature representations from U-SalNet, with the parameters updated by unsupervised attentional backpropagation. Details of each step are explained in the following sections.

3.1 Pre-training method

Inspired by [8]'s work, we aggregate a large data set from several simulated CECT data sets with rich salient objects, different SNR settings and various resolutions, which is called SalSeg-CECT and will be described in Section 4. Based on the SalSeg-CECT data set, we then pre-train a model specially designed for biomedical images, as is shown in the red dashed-line box in Figure. 2(a). Our goal is to learn robust feature representations which can benefit training on biomedical data by utilizing a pre-trained network on the SalSeg-CECT data set. In this work, for deep feature extraction on biomedical data sets, we adopt the common encoder-decoder architecture to train our backbones of the network. Particularly, we choose the U-Net model as the basic structure on 2D images, and the V-Net for 3D volumes. The significant differences of SalSeg-CECT images from natural images come from the low SNR, limited tilt projection range (the missing wedge effect) and crowded nature of intracellular structures. Therefore, our pre-trained method is different from the current common pre-trained models. To the best of our knowledge, we are the first to pre-train a model on biomedical CECT data.

3.2 U-SalNet architecture

Our U-SalNet model includes two attentional mechanisms: global attention and local attention, which are integrated into the U-Net architecture, as is shown in the purple dashed-line box in Figure. 2(b). U-SalNet aims to selectively segment salient objects from the background by generating an attention map at each pixel. We apply the two modes of attention to refining salient regions in biomedical images. The detailed U-SalNet architecture is shown in Figure. 3. We first upsample each level of feature maps in the decoder branch and concatenate them with their corresponding levels of features in the encoder branch. After concatenating the outputs from encoder and decoder, we get a feature map $\Gamma \in \mathbb{R}^{C\times W\times H}$ as the input to the attention module, where C, W, H denote the channels, width, and height, respectively.

For global attention, as is shown in Figure. 3(a). To generate attention over the whole feature map Γ for each pixel, we first apply four ReNet models[13] to sweep across a feature volume both horizontally and vertically along two directions to concentrate global information. Next, a convolution operation is performed to transform this feature map to D_G channels, where $D_G = W \times H$. At the same time, the feature vector $x^{w,h}$ at each pixel (w, h) is normalized via softmax to obtain global attention weights $\Phi^{w,h}$, as is shown in Equation (1). Here $i, j \in \{1, ..., D_G\}$.

In order to generate the global attention feature Γ_{att}, as is shown in Equation (2) and (3), the features at all locations in the whole feature map Γ are summed with weights according to $\Phi_i^{w,h}$. $\Phi_i^{w,h}$ refers to the salient correlation between the object pixel (w, h) and the pixel at the i^{th} location (w_i, h_i). $Conv_i \in \mathbb{R}^C$ is the Conv feature at (w_i, h_i) in Γ. Γ_{att} has the same size with Γ.

For the local attention, as is shown in Figure. 3(b), we consider a local feature cube $\Gamma'^{w,h} \in \mathbb{R}^{W'\times H'\times C}$ centered at (w, h) with width W' and height H'. Φ' is derived from convolution layers with a reception field of $W' \times H'$ in the original feature map for each location. Similar to global attention, the features in $\Gamma'^{w,h}$ are weightedly summed by $\Phi'^{w,h}$ to construct Γ'_{att}, as is shown in Equation

(4).

$$\begin{aligned}\Phi_i^{w,h} &= \text{softmax}\left(x^{w,h}\right) \\ &= \frac{\exp\left(x_i^{w,h}\right)}{\sum_{j=1}^{D_G}\exp\left(x_j^{w,h}\right)},\end{aligned} \tag{1}$$

$$\Gamma_{att}^{w,h} = \sum_{i=1}^{D_G}\Phi_i^{w,h}Conv_i, \tag{2}$$

$$Conv_i = \int_{\mathbb{R}^C} f(\tau)g(t-\tau)d\tau \tag{3}$$

$$\Gamma'^{w,h}_{att} = \sum_{i=1}^{D_L}\Phi'^{w,h}_i Conv'^{w,h}_i, \tag{4}$$

$$Loss = -(y_t\log(y_p) + (1-y_t)\log(1-y_p)). \tag{5}$$

It's worth noticing that although our attention mechanism is similar to [14]'s work, they only consider the serial combination of global and local attention. Our U-SalNet focuses more on fusing two attention modules through convolutional decoding. In addition, their model adopts deep supervision, as is shown in Equation (5), where y_t denotes the true saliency map and y_p is the predicted saliency map. Also, their experiments are all based on natural images and do not apply to biomedical images. On the contrary, our proposed U-SalNet architecture not only applies to complex biomedical images featuring a low SNR, low resolution and rich salient objects but also achieves complete unsupervision. We will further describe the unsupervised attentional backpropagation algorithm in the next section.

3.3 Unsupervised backpropagation

The proposed unsupervised backpropagation algorithm is shown in Algorithm 1. For the salient segmentation task, we consider two aspects: 1) predicting salient objects through current network parameters; 2) training network parameters through current salient predictions. Accordingly, we can obtain salient regions by a forward pass through the neural network and use gradient descent to optimize the backward pass of the network at the same time. In order to update the parameters, we adopt stochastic gradient descent (SGD) with momentum to backpropagate results of the cross-entropy loss calculated between outputs of the model $\{Y_n'\}$ and the refined salient object labels $\{y_n'\}$. In the backpropagation step, we use the SLIC algorithm [15] implemented in scipy for the $GetSuperpixels$ method. $\{y_n'\}$ is unified within every superpixel in an image, which is obtained through voting of labels inside a single superpixel and taking the majority of labels as the result. Different from [2]'s work, their prediction results are random and often fails on biomedical images. For solving these problems, our network parameters are initialized with the pre-trained method mentioned above. Furthermore, two attentional modules of the U-SalNet are combined to further refine the results, which achieves self-awareness. Finally, the forward-backward process is iterated I times to generate the final prediction of salient objects $\{y_n\}$. The probabilities $\{Y_n'\}$ was trained in a self-supervised manner using $\{y_n'\}$. The orange dashed-line box in Figure. 2(c) illustrates the proposed algorithm to train our U-SalNet model.

4 Experiments

4.1 Datasets setting

Our SalSeg-CECT data set includes 36,000 2D and 72,000 3D CECT images, which are generated with 4 levels of SNR (0.1, 0.5, 1.0 and 1.5), various resolutions and missing wedge effects. The procedure of simulation uses the same simulator as in [24], which simulates tomographic images by imitating the actual tomography reconstruction process using macromolecular complexes of known densities. For 2D

Algorithm 1 Unsupervised Backpropagation Algorithm

Require: Original biomedical image
Ensure: Salient segmentation results

$(W, b) = Init()$ // Initialize backbone parameters
$(W', b', nClass) = Init()$ // Initialize classifier parameters
$\{S_k\}_{k=1}^{K} = GetSuperpixels(\{p_n\}_{n=1}^{N})$
for $iter = 1 \rightarrow I$ **do**
  if $nClass > 2$ **then**
    $\{F_n\}_{n=1}^{N} = GetFeatures(\{p_n\}_{n=1}^{N}, \{W, b\})$
    $\{GA_n\}_{n=1}^{N} = GlobalAttention(\{F_n\}_{n=1}^{N})$
    $\{LA_n\}_{n=1}^{N} = LocalAttention(\{F_n\}_{n=1}^{N})$
    $\{Y_n\}_{n=1}^{N} = \{W'(GA_n \bigoplus LA_n) + b'\}_{n=1}^{N}$
    $\{Y'_n\}_{n=1}^{N} = BatchNorm(\{Y_n\}_{n=1}^{N})$
    $\{y_n\}_{n=1}^{N} = \{\text{argmax } Y'_n\}_{n=1}^{N}$ //predict salient labels
    for $p = 1 \rightarrow P$ **do**
      $y_{\max} = \text{argmax}|y_n|_{n \in S_p}$
      $y'_n = y_{\max}$ **for** $n \in S_p$
    end for
    $L = CrossEntropyLoss(\{Y'_n, y'_n\}_{n=1}^{N})$
    $\{W, b\}, \{W', b'\} = Update(L)$
  end if
end for

images, tomograms are sliced into grayscale images in 3 dimensions with resulting width and height of 200 pixels. Our augmentation includes random flipping and cropping. For 3D images, tomograms are segmented into volumes of $64 \times 64 \times 64$ and their values normalized before being fed into models. The test set is generated independently with macromolecular complexes different from the training set with SNR = 0.5 and SNR = 1.5. To verify the ability of generalization of PUB-SalNet, we also apply it to the ISBI data set [25] and visualize our results in Figure. 6.

4.2 Implementation Details

All our networks are implemented with PyTorch. 3 NVIDIA GTX 1080 Ti GPU with 11GB GPU memory each are used for pre-training and testing. In the pre-training step, we choose batch size = 4 for both 2D and 3D settings. For 2D, we use the SGD optimizer with learning rate=0.01, momentum=0.9 and weight decay = 0.0005. While for 3D, the learning rate is $1e^{-5}$ and momentum is 0.99. In case of memory explosion, we apply global attention twice followed by local attention three times and a convolutional decoding layer in the decoder of U-SalNet. The default number and compactness of superpixels are 10000 and 100. The maximum number of backpropagation iterations is set to 1000. The label of the majority of pixels is regarded as "non-salient". The initial number of classes to be decreased is set to 100 as default. If the algorithm does not converge to 2 classes of labels after 1000 iterations, all the labels except for the non-salient type will be counted as "salient".

4.3 Evaluation Metrics

To compare the quantitative results generated by different methods, here we use four popular metrics to evaluate our model against other unsupervised methods.

Region Similarity F. To measure the similarity of matching regions from two salient segmentation maps, F is defined as:

Data set	SNR=0.5				SNR=1.5			
Method \ Metric	ε	F	E	S	ε	F	E	S
Itti [16]	<u>0.1277</u>	<u>0.4759</u>	0.3811	0.4445	<u>0.1206</u>	<u>0.6396</u>	0.4639	0.4781
LC [17]	0.1626	0.3277	0.4466	<u>0.4846</u>	0.1463	0.4615	0.4369	0.5022
SR [18]	0.1340	0.2535	0.3020	0.4406	0.1316	0.3439	0.2911	0.4423
IG [19]	0.2843	0.1713	0.4775	0.4262	0.2978	0.1848	0.4739	0.4322
SIG [20]	0.2623	0.2647	<u>0.4959</u>	0.4781	0.2310	0.3387	<u>0.5134</u>	0.5177
VA [21]	0.2843	0.1713	0.4775	0.4262	0.2978	0.1848	0.4739	0.4322
SVA [20]	0.2625	0.2647	0.4957	0.4779	0.2305	0.3414	0.5129	<u>0.5186</u>
VBP [22]	0.1295	0.3049	0.4033	0.4527	0.1224	0.4588	0.4053	0.4717
SalGAN [23]	0.1427	0.1984	0.3126	0.4411	0.1585	0.2367	0.4090	0.4629
PUB-SalNet	**0.0914**	**0.6573**	**0.7036**	**0.6494**	**0.0762**	**0.7426**	**0.7522**	**0.7209**
Improvement	↓ 28.43%	↑ 38.12%	↑ 41.88%	↑ 34.01%	↓ 36.82%	↑ 16.10%	↑ 46.51%	↑ 39.00%

Table 1: Comparison of performance of ten unsupervised methods with four metrics on the simulated biomedical test sets. ε stands for Mean Absolute Error (MAE), F for region similarity, E for the enhanced alignment measure, and S for structural similarity. Lower is better for ε, and higher is better for the other three metrics. The results are calculated according to Equation (6), (7), (8). The best performance of each metric is in **bold** and the second best is <u>underlined</u>. The improvements of our PUB-SalNet over the best of other methods in relative percentage is shown in the last row.

Data set	SNR=0.5				SNR=1.5			
Method \ Metric	ε	F	E	S	ε	F	E	S
B	0.1461	0.1628	0.3692	0.4230	0.1433	0.1628	0.3960	0.4223
U+B	0.2870	0.1628	0.4834	0.3585	0.2677	0.1628	0.5130	0.3693
P+B	0.1063	0.5631	0.5906	0.5661	0.0949	0.6551	0.5947	0.5979
P+U	0.1104	0.6214	0.6306	**0.6506**	0.0973	**0.7544**	0.7465	**0.7617**
P+U+B	**0.0914**	**0.6573**	**0.7036**	0.6494	**0.0762**	0.7426	**0.7522**	0.7209

Table 2: Quantitative comparisons between different combination of modules from our PUB-SalNet model. **B** stands for a single unsupervised backpropagation module; **U+B** stands for U-SalNet architecture with **B**; **P+B** means **B** based on the pre-training method; **P+U** means U-SalNet based on the pre-training method, note that this is actually not an unsupervised method; **P+U+B** is our proposed PUB-SalNet.

$$F = \frac{\left(1+\beta^2\right) \text{ Precision } \times \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}} \tag{6}$$

where $\beta^2 = 0.3$ to balance between $recall$ and $precision$.

Pixel-wise Accuracy ε. F does not consider true negative saliency predictions. We define the normalized ($[0, 1]$) mean absolute error (MAE) between predicted salient segmentation maps and ground truth masks as:

$$\varepsilon = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \|M(x,y) - G(x,y)\| \tag{7}$$

where W and H are the width and height of images, respectively.

Enhanced Alignment Measure E. Proposed by [26], using the enhanced alignment matrix ϕ_{FM} to measure the two properties (pixel-level matching and image-level statistics) of a binary map, E is defined as:

$$Q_{FM} = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} \phi_{FM}(x,y) \tag{8}$$

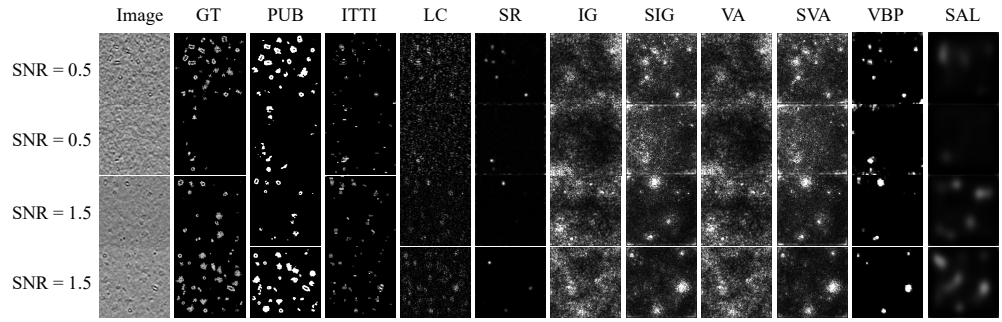

Figure 4: Qualitative visual results of ten unsupervised methods on the simulated biomedical data set with SNR = 0.5 and 1.5. GT stands for ground truth images, PUB is PUB-SalNet, and the other nine methods are referenced in Table 1.

Data set	SNR=0.5				SNR=1.5			
Metric / Method	ε	F	E	S	ε	F	E	S
PUB-SalNet-B**20**	0.0984	0.6347	0.6964	0.6428	0.0793	0.7239	0.7447	0.7124
PUB-SalNet-B**40**	0.0961	0.6396	0.6945	0.6443	0.0766	0.7318	0.7320	0.7107
PUB-SalNet-B**60**	0.0945	0.6437	0.6710	0.6358	0.0774	0.7218	0.7294	0.7032
PUB-SalNet-B**80**	0.0943	0.6543	**0.7221**	**0.6598**	0.0783	0.7327	0.7340	0.7065
PUB-SalNet-B**100**	**0.0914**	**0.6573**	0.7036	0.6494	**0.0762**	**0.7426**	**0.7522**	**0.7209**

Table 3: The quantitative comparison of parameter sensitivity analysis under four metrics. ε, F, E and S are the same as Table 1. For PUB-SalNet-B**X**, **X** stands for the initial number of classes to be decreased, as is in Function *Init()* parameter in Algorithm 1.

where h and w are the height and width of the map, respectively.

Structural Similarity S**.** S proposed by [27] evaluates the structural similarity by considering both regions and objects. Since saliency of potential spacial structures is crucial to biomedical images, we additionally use S to comprehensively evaluate the structural similarity of biomedical images.

4.4 Quantitative Evaluation

Table 1, Table 2 and Table 3 show quantitative evaluation results on the simulated biomedical test set, which we will detailedly discuss in the following three subsections.

4.4.1 Comparison with state-of-the-art

As is shown in Table 1, we compare our PUB-SalNet model to 9 other state-of-the-art unsupervised methods. We demonstrate through experimental results that our proposed PUB-SalNet outperforms all existing unsupervised methods with a great margin (such as 46.51% for E-Measure on $SNR = 1.5$ and 41.88% on $SNR = 0.5$) and achieves new state-of-the-art performance.

4.4.2 Ablation Study

To demonstrate the effectiveness of the proposed PUB-SalNet model, we compare quantitative results of different combinations of modules from our method, as is shown in Table 2. **B** stands for back-propagation from [2]'s work, which serves as our baseline because it is a classic unsupervised image segmentation method using deep learning. The experimental results in Table 2 shows that three parts of our PUB-SalNet functions together and are all indispensable. It is even competitive compared to the supervised method.

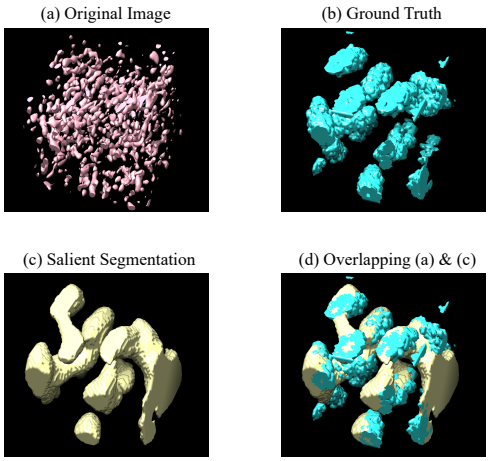

Figure 5: Visualization of 3D salient segmentation by PUB-SalNet on a 3D subvolume of size $64 \times 64 \times 64$ from the CECT test set. The pictures are obtained using UCSF Chimera, which displays the isosurface of the 4 corresponding 3D images. (d) demonstrates that the predicted salient region greatly overlaps with the ground truth macro-molecular structure.

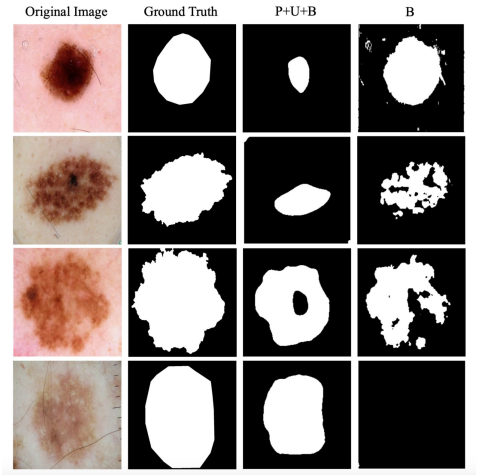

Figure 6: Case study of 2D salient segmentation by PUB-SalNet and the B module on the ISBI Challenge [25].

Data set	ISBI 2017 Skin		
Metric / Method	ε	F	E
B	**0.3136**	**0.3378**	0.4140
P+U+B	0.3498	**0.3378**	**0.4674**

Table 4: The performance comparison of the strong baseline model (B, for backpropagation only) and our proposed PUB-SalNet under three metrics on the ISBI Challenge [25]. ε, F and E are the same as in Table 1. The best performance of each metric is in **bold**.

4.4.3 Parameter Sensitivity Analysis

In order to demonstrate the robustness of our proposed model, we also construct an experiment of parameter sensitivity analysis. The quantitative comparison under four metrics is shown in Table 3. Our experiments show a deviation within 1%-3% on the evaluation of the four metrics, which indicates that the parameters of our model have little influence to salient segmentation results.

4.5 Qualitative Evaluation

Figure 4 demonstrates the saliency maps predicted by nine unsupervised saliency detection methods on our testing data set with $SNR = 0.5$ and $SNR = 1.5$. Traditional algorithms sometimes are able to detect multiple salient objects due to their superior capabilities of capturing low-level contrasts of features. When facing grayscale biomedical data sets with low SNR, deep learning methods are not as promising as we expected, which highlights blurry regions with richer contextual information. The performance of PUB-SalNet outperforms all other unsupervised methods on salient segmentation. We also present our results on the 3D CECT test set in Figure 5, which proves the generalization ability of our model. Our model is capable of detecting salient objects under various settings and can be effectively extended to 3D image processing tasks.

4.6 Case Study on the ISBI Challenge

A case study of 2D salient segmentation by PUB-SalNet and the B module on ISBI 2017 Challenge on Skin Leision Analysis[25] is shown in Figure 6. To the best of our knowledge, we are the first to conduct unsupervised salient segmentation on the ISBI challenge. Comparison of the strong baseline model (B, for backpropagation only) and our proposed PUB-SalNet under three metrics is shown in Table 4. B module outperforms P+U+B by predicting shapes and edges with more accuracy. However, with the lack of global features, it falsely captures small differences in color within a piece of illness. It can also be easily perturbed by impurities or foreign matters, as is shown in the first two examples. P+U+B can produce smoother results, although sometimes fails to match the target in shape. P+U+B benefits from abundant semantic information and produces better results in the last two rows, while B focuses on wrong patches of color and cannot detect saliency on a global scale.

5 Conclusion

In this paper, we propose a completely unsupervised self-aware network based on pre-training and attentional backpropagation for biomedical salient segmentation, namely PUB-SalNet. The experimental results on the 2D and 3D data also display the generalization ability and robustness of our method. In the future, we will integrate our salient segmentation method into other complex biomedical tasks, such as biomedical image registration and quantification of uncertainty in segmentation.

6 Acknowledgement

This work was supported in part by U.S. National Institutes of Health (NIH) grant P41 GM103712. This work was supported by U.S. National Science Foundation (NSF) grant DBI-1949629. XZ was s upported by a fellowship from Carnegie Mellon University's Center for Machine Learning and Health. This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/2602-01 and URF/1/3007-01.

References

[1] June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: general overview. *Korean journal of radiology*, 18(4):570–584, 2017.

[2] Asako Kanezaki. Unsupervised image segmentation by backpropagation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1543–1547. IEEE, 2018.

[3] Anil Kumar Reddy, Sai Vikas, R Raghunatha Sarma, Gurudat Shenoy, and Ravi Kumar. Segmentation and classification of ct renal images using deep networks. In *Soft Computing and Signal Processing*, pages 497–506. Springer, 2019.

[4] Khalid Raza and Nripendra Kumar Singh. A tour of unsupervised deep learning for medical image analysis. *arXiv preprint arXiv:1812.07715*, 2018.

[5] Takayasu Moriya, Holger R Roth, Shota Nakamura, Hirohisa Oda, Kai Nagara, Masahiro Oda, and Kensaku Mori. Unsupervised segmentation of 3d medical images based on clustering and deep representation learning. In *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 10578, page 1057820. International Society for Optics and Photonics, 2018.

[6] Ioana Croitoru, Simion-Vlad Bogolin, and Marius Leordeanu. Unsupervised learning of foreground object segmentation. *International Journal of Computer Vision*, pages 1–24, 2019.

[7] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[8] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[9] Miitu KM Honkanen, Hanna Matikka, Juuso TJ Honkanen, Abhisek Bhattarai, Mark W Grinstaff, Antti Joukainen, Heikki Kröger, Jukka S Jurvelin, and Juha Töyräs. Imaging of proteoglycan and water contents in human articular cartilage with full-body ct using dual contrast technique. *Journal of Orthopaedic Research®*, 37(5):1059–1070, 2019.

[10] Xide Xia and Brian Kulis. W-net: A deep model for fully unsupervised image segmentation. *arXiv preprint arXiv:1711.08506*, 2017.

[11] Kai Hu, Si Liu, Yuan Zhang, Chunhong Cao, Fen Xiao, Wei Huang, and Xieping Gao. Automatic segmentation of dermoscopy images using saliency combined with adaptive thresholding based on wavelet transform. *Multimedia Tools and Applications*, pages 1–18, 2019.

[12] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019.

[13] Francesco Visin, Kyle Kastner, Kyunghyun Cho, Matteo Matteucci, Aaron Courville, and Yoshua Bengio. Renet: A recurrent neural network based alternative to convolutional networks. *arXiv preprint arXiv:1505.00393*, 2015.

[14] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018.

[15] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels. Technical report, 2010.

[16] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998.

[17] Yun Zhai and Mubarak Shah. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 815–824. ACM, 2006.

[18] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. Ieee, 2007.

[19] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017.

[20] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[21] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[22] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry Jackel, Urs Muller, and Karol Zieba. Visualbackprop: efficient visualization of cnns. *arXiv preprint arXiv:1611.05418*, 2016.

[23] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[24] Long Pei, Min Xu, Zachary Frazier, and Frank Alber. Simulating cryo electron tomograms of crowded cell cytoplasm for assessment of automated particle picking. *BMC bioinformatics*, 17(1):405, 2016.

[25] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016.

[26] 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.

[27] 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.