Is Deep Learning All You Need for Unsupervised Saliency Detection of Biomedical Images?

Feiyang Chen, Ying Jiang, Xiangrui Zeng, Min Xu
Carnegie Mellon University
{feiyangc, yingjian, xiangruz}@andrew.cmu.edu, mxu1@cs.cmu.edu

Abstract

Pre-trained networks have recently achieved great success in computer vision. At present, most deep learning-based saliency detection methods use pre-trained networks to extract features, regardless of supervised or unsupervised. However, we found that when unsupervised saliency detection is performed on grayscale biomedical images, pre-trained networks such as VGG cannot effectively extract significant features. We suggest that VGG is not able to learn salient information from grayscale biomedical images and its performance greatly depends on RGB cues and quality of the training set. To verify our hypothesis, we construct an adversarial data set featuring a low signal-to-noise ratio (SNR), low resolution and rich salient objects and conduct a series of probing experiments. What's more, in order to further explore what VGG has learned, we visualize intermediate feature maps. To the best of our knowledge, we are the first to investigate the reliability of deep learning methods for unsupervised saliency detection on grayscale biomedical images. It's worth noticing that our adversarial data set also provides a more robust evaluation of saliency detection and may serve as a standard benchmark in future work on this task.

1. Introduction

Saliency detection is a task of simulating human visual behavior to detect distinctive salient objects of an image (i.e. areas of human interest). For example, a face in a photo, a petal in an outdoor natural image and so on[21], as is shown in Figure 1. This is a challenging task for the machine and even humans can hardly determine which goals are salient in an image, as evidenced by [3]'s work.

With the development of deep learning[22] and the resurgence of Convolutional Neural Networks (CNNs)[23], CNN models are gradually becoming the mainstream in saliency detection[7]. One of the most commonly used techniques with regard to CNNs is pre-training [39]. Currently, the majority of pre-training methods (e.g. VGG[34]) make use of RGB images with high quality. Yet, in reality, grayscale images are also popular and easier to obtain in many fields, particularly in the biomedical area [46]. Figure 2 shows four slices of a 3D Cellular Electron Cryo-Tomography (CECT) image. CECT [26] is a cutting-edge 3D imaging technique that visualizes the subcellular organization of a sub-region inside a single cell at sub-molecular resolution and in its near-native state. What's more, grayscale images have advantages over RGB images for saliency detection. They lead not only to a speedup in model training and in detection time but also in preserving rich saliency information when compared to color images, as explained by [41].

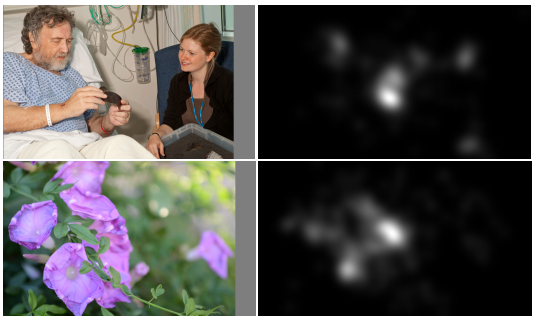

Figure 1. Two example images from CAT2000 [8] data set. The left column is the original image and the right column is the saliency map.

However, biomedical images are more or less distorted. For example, CECT images come with low SNR, limited tilt projection range (the missing wedge effect) and crowded nature of intracellular structures, which makes saliency object detection for latent structures very difficult under such settings. On the one hand, laborious manual annotation of images and knowledge of experts is needed to discover potentially salient regions, which makes unsupervised learning essential to such tasks since it enables automatic detection of salient objects without relying on available data annotations. On the other hand, all existing pre-trained net-

works (e.g. VGG) only extract features from RGB images. Unfortunately, variant types of biomedical grayscale images have very different complex structures from natural images. Such structures may not be properly encoded in the original VGG. Therefore, we doubt whether VGG trained on RGB images is capable of extracting features aside from color cues from such datasets through unsupervised methods. This motivates a question: what has VGG learned about saliency object detection in grayscale biomedical images?

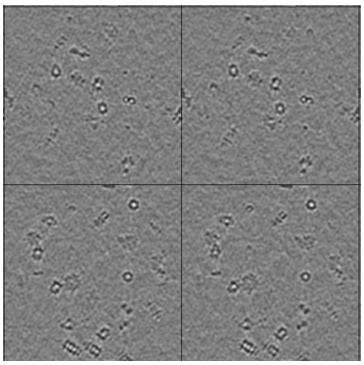

Figure 2. An example of slices of simulated tomogram constructed using [29]'s method.

To investigate VGG's effectiveness in unsupervised saliency detection, we study how pre-training methods help to extract features in detecting an interesting object to people. [43] extracts deep features from the output of the last convolution layer of VGG16 or FCN32s[25], which is because features from the last layer of CNNs encode semantic abstraction of objects and are robust to appearance variations. Since the deep features are too concentrated in the last layer, they use dilated convolution in the last three convolution blocks to obtain lager feature maps. [45] utilized a fully convolutional neural network (FCN) due to its superior capability in feature learning and feature representation as a latent saliency prediction module. They are both methods specifically designed for RGB images. To verify our hypothesis, we constructed an adversarial data set featuring low SNR, low resolution and richness in salience objects. Through probing experiments to isolate such effects, we demonstrate in this work that VGG's performance greatly depends on RGB cues and the quality of the training set. The adversarial data set, therefore, provides a more robust evaluation of saliency and should be considered to be adopted as the standard in biomedical saliency detection in future works. In this paper, our contribution is summarized as follows:

1. We construct an adversarial data set featuring low SNR, low resolution and rich salient objects for salient detection tasks, providing a more robust evaluation and could be adopted as a standard in future work on biomedical saliency detection.

2. Through probing experiments on our adversarial data set, we demonstrate that VGG's performance greatly depends on RGB cues and the quality of the training set.

3. To the best of our knowledge, we are the first to conduct comprehensive experiments on unsupervised saliency detection methods with both traditional and deep learning in the biomedical area. Our work provides a novel insight for improving the robustness of biomedical saliency detection.

The rest of the paper is organized as follows. In the following section, we introduce task descriptions and baselines. In Section 3 we describe the VGG pre-train network's traits. We will illustrate how to construct our adversarial test set in Section 4. Probing experiments are discussed in detail in Section 5. We review our related works in Section 6 and conclude in Section 7.

2. Task Description and Baselines

Saliency detection can usually be formulated as a binary segmentation problem[24] that separates a salient object from the background, here we adopted a continuous measure which is the same as in [6] using fixation density maps. Let $i = 1, ..., n$ index each pixel in the input image plane. The ground truth is a grayscale value map scaled to $[0, 1]$, indicating how salient each pixel is relative to the most salient pixel in the picture. The model outputs the belief for each pixel in its saliency. All pixels' predicted values forms a grayscale map D, where $|D| = n$. Unsupervised learning restricts that we do not have available data annotations, and ground truth images will not be used except for evaluating the results of the saliency map.

According to [7], a good saliency detection model should meet the following three criteria: 1) good detection: the probability of missing real salient regions and falsely marking the background as a salient region should be lower, 2) high resolution: saliency maps should have high or full resolution to accurately locate salient objects and retain original image information, and 3) computational efficiency: as front-ends to other complex processes, these models should detect salient regions quickly.

Inspired by the principals, we select some of the simplest yet effective traditional algorithms, such as Spectral Residual Approach [16], LC Algorithm [44] and Itti's algorithm [18]. We compare them with deep learning methods, including Vanilla Gradient [33], Integrated gradient [37], Visual Backprop [5] along with their corresponding smooth-gradient versions [35], and the best unsupervised deep learning model of [27]. For all of our deep learning

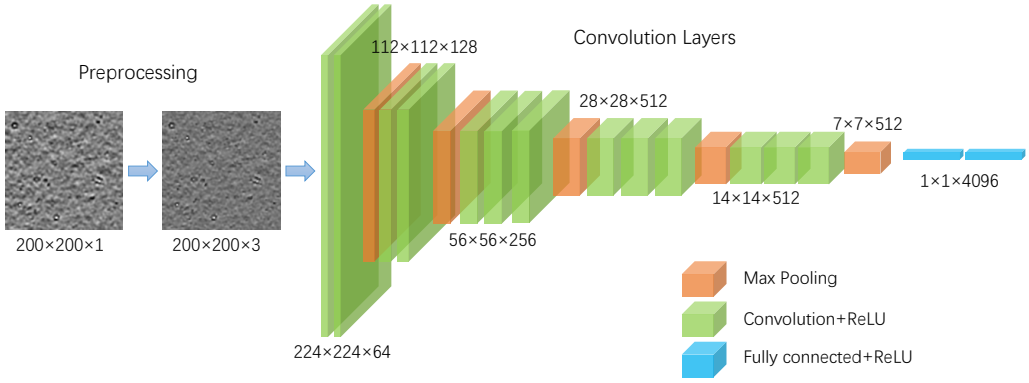

Figure 3. The Classic VGG16 network architecture we adopted from [34]'s work for deep feature extraction. VGG is also the most popular pre-trained network architecture in current saliency detection tasks.

experiments we use grid search to select hyperparameters, dropout regularization [36], and Adam [19] for optimization. The methodologies will be further described in Section 6.

3. VGG Pre-Trained Network

Currently, the VGG16 network is one of the most popular pre-trained methods in the saliency detection task. Its architecture is shown in 3. VGG16 is a CNN model proposed by [34], which achieves 92.7% top-5 test accuracy in ImageNet [10], an RGB data set of over 14 million images belonging to 1000 classes. In complement with the low-level details captured by classic algorithms, the pre-trained VGG [34] network is often utilized as a global feature extractor to help to detect globally salient regions [43, 17, 33, 37, 5]. In our experiments, we extract feature volumes from the output of the second fully-connected layer of VGG-16, since features from the deeper layers encode better semantic abstraction of objects and are robust to appearance variations [43].

However, although VGG16 is trained with high-quality RGB images, it has not been tested with grayscale images. The training set may induce the network to rely too much on image resolution and direct color information. [41] has pointed out that saliency information is preserved in low-resolution grayscale images, but applying pre-trained saliency models to grayscale images does not produce very promising outcomes. This leads to our assumption that pure structural information may not be effectively learned aside from RGB color cues by the network when trained on color images only. In this paper, for probing experiments, we transform the grey images into 3 channels as input to the VGG network. We visualize the active maps of the middle layers of VGG to see what it learned. We find out that VGG's performance greatly depends on RGB cues and the quality of the training set.

4. An Adversarial Test Set

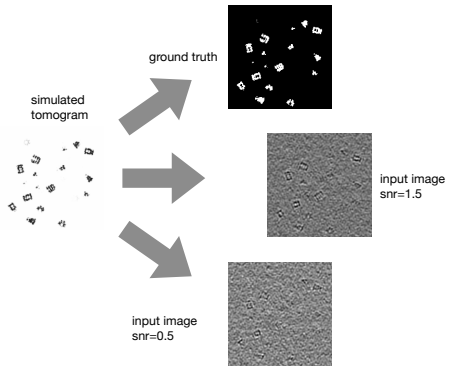

Figure 4. An illustration of the process of creating our adversarial data set.

[41] explains the biological and computational motivation for low-resolution grayscale images (LG). They show through a range of human eye-tracking and computational modeling experiments that saliency information is preserved in those images, and using LG images leads to significant speedups in model training and detection times. They propose LG images for fast and efficient saliency detection. Given that pre-trained networks such as VGG are mainly trained on high resolution colored images, we construct a data set derived from a simulated Cellular Electron Cryo-Tomography (CECT) data set featuring grayscale, crowded content and low signal-to-noise ratio (SNR). In principle, a CECT image, called a tomogram, captures structural information of cellular components in the field of view. However, several factors, such as low SNR, the limited tilt projection range (missing wedge effect) and the crowded nature of intra-cellular structures make systematical recognition and extraction of structures in a tomogram a difficult task for computational analysis. With the help of a ma-

ture simulation procedure proposed in [29], in which known densities of macromolecular complexes are used to simulate tomograms by mimicking the actual tomographic image reconstruction process, we can construct tomograms with a high degree of simulation and the ground truth of macromolecular structures. The definition of saliency is clear on such a data set, which is the spots containing meaningful structural information. Since the ground truth is accurate and easy to obtain, with the traits of tomography finely preserved, the simulated CECT data set is highly desirable for exploratory tasks on biomedical images.

Algorithm 1 The Generation of our Adversarial Data Set Procedure

procedure TOMOMINER RECONSTRUCTION
  $op = SetSimulationParameters()$
  for i:[0,n] **do**
    $v = getData(t_i)$ // get the i^{th} density map
    vb = $Reconstruction(v, op)$ // use parameters in op to reconstruct a tomogram vb
  end for
  for i:[0,vb.shape[2]] **do**
    $im = vb[:, :, slice\_count]$ // get the $slice\_count^{th}$ slice from vb
    $im\_v = im[N.isfinite(im)]$ // Value normalization within each slice
    if $im\_v.max() > im\_v.min()$: **then**
      $im = (im - im\_v.min())/(im\_v.max() - im\_v.min())$
    end if
  end for
end procedure
procedure SET SIMULATION PARAMETERS
  $model = \{mwa = 30, SNR = 0.5\}$
  $ctf = \{pix\_size = 1.0, Dz = -5.0, voltage = 300, Cs = 2.0, sigma = 0.4\}$
end procedure
procedure ROTATION ANGLE
  $loc\_proportion = 0.1$
  $loc\_max = Shape(v) \times loc\_proportion$
  $angle = Random\_Rotation\_Angle\_zyz()$
  $loc\_r = (Random(3) - 0.5) \times loc\_max$
  $vr = Rotate(v, angle, loc\_r)$
end procedure

Our data set consists of 12,000 grayscale images converted from slices of simulated CECT tomograms, which are generated with 2 different levels of SNR (0.5 and 1.5). The generation of our adversarial data set is shown in Algorithm 1, where 'mwa' stands for 'missing wedge angle', 'ctf' for 'contrast transfer function', and 'Dz' for 'defocus'. The simulation procedure uses the same simulator as in [29]. We choose 5 distinct structures from the Protein Databank (PDB) [4] and simulate tomograms of 200^3 voxels with missing wedge angle 30°. To obtain the ground truth intensity maps, we set a threshold $x = 0$, discard the positive values, flip the sign of the negative ones which are produced by stimulation of variant intensities of light passing through different parts of the cell, and normalize the values. We also perform anisotropic diffusion denoising [32] with $\kappa = 70$ and $n = 10$ to reduce the noise and to better preserve subtle, edge-like structures. As is an usual technique in cell image pre-processing. Finally, tomograms are sliced to grayscale images in 3 dimensions with an resulting width and height of 200 pixels. The process of construction is demonstrated in 4. The sliced images will be resized to 224*224 and then be fed into different models.

5. Probing Experiments

5.1. Grayscale and RGB Conversion

Images are sometimes converted from 24-bit sRGB (Standard Red Green Blue) to 8-bit grayscale since it is faster and more efficient to merge the three channels before performing subsequent operations. Color to grayscale conversion is a lossy operation since it can result in reduced brightness, which affects the sensibility of models to saliency [15, 14]. In order to avoid such systematic errors, the conversion must preserve at least the luminance characteristics of the original stimulus (i.e. the luminance of the grayscale pixels must be the same as the original color image).

According to [41], to convert a 24-bit sRGB gamma-compressed color image I_{HC} into an 8-bit grayscale representation of its luminance I_{HG}, the gamma compression function must first be removed by gamma expansion to convert the image into a linear RGB color space [31]. Appropriately weighted sums can be applied to linear color components R_{linear}, G_{linear}, B_{linear}. For the sRGB color space, the gamma extension is defined as

$$C_{linear} = \begin{cases} \frac{C_{sRGB}}{12.92} & C_{sRGB} \leq 0.04045 \\ \left(\frac{C_{sRGB}+0.055}{1.055}\right)^{2.4} & C_{sRGB} > 0.04045 \end{cases} \quad (1)$$

Where C_{sRGB} represents any of the three gamma-compressed sRGB primary colors (R_{sRGB}, G_{sRGB}, and B_{sRGB}, each in the range $[0, 1]$), and C_linear is the corresponding linear intensity value (R_{linear}, G_{linear}, and B_{linear}, also in the range $[0, 1]$). Then, I_{HG} is calculated as a weighted sum of three linear intensity values, which is given by

$$I_{HG} = 0.2126 \times R_{\text{linear}} + 0.7152 \times G_{\text{linear}} + 0.0722 \times B_{\text{linear}} \quad (2)$$

These three coefficients represent the intensity (luminance) perception of a standard observer to the light of the

Data set	SNR=0.5				SNR=1.5			
Method \ Metric	ε	F	E	S	ε	F	E	S
Itti	**0.1277**	**0.4759**	0.3811	0.4445	**0.1206**	**0.6396**	0.4639	0.4781
LC	0.1626	0.3277	0.4466	**0.4846**	0.1463	0.4615	0.4369	0.5022
SR	0.1340	0.2535	0.3020	0.4406	0.1316	0.3439	0.2911	0.4423
integrated_grad	0.2843	0.1713	0.4775	0.4262	0.2978	0.1848	0.4739	0.4322
smoothed_int_grad	0.2623	0.2647	**0.4959**	0.4781	0.2310	0.3387	**0.5134**	0.5177
vanilla_grad	0.2843	0.1713	0.4775	0.4262	0.2978	0.1848	0.4739	0.4322
smoothed_van_grad	0.2625	0.2647	0.4957	0.4779	0.2305	0.3414	0.5129	**0.5186**
visual_bp	0.1295	0.3049	0.4033	0.4527	0.1224	0.4588	0.4053	0.4717
SalGAN	0.1427	0.1984	0.3126	0.4411	0.1585	0.2367	0.4090	0.4629

Table 1. The performance of different unsupervised saliency detection models (traditional algorithms and deep learning-based methods) under four metrics. 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 the opposite for the other three metrics. The results are calculated according to Eqn. (3), (4), (5) over our adversarial data set. The best performance of each metric is in **bold**.

precise Rec. 709 [30] additive primary colors that are used in the definition of sRGB [41].

5.2. Saliency Detection Evaluation Metrics

In our evaluation framework, given a predicted saliency map M generated by an unsupervised saliency detection model and a ground truth mask G, the evaluation metrics are to tell how well are the results generated by different models. Here, we use four evaluation metrics [11] to evaluate saliency models on our constructed adversarial data set.

Region Similarity F. To measure how well the regions of the two maps match, we use the $F-measure$, defined as:

$$F = \frac{\left(1+\beta^2\right) \text{ Precision } \times \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}} \tag{3}$$

where $\beta^2 = 0.3$ is suggested by [2] to balance between the $recall$ and $precision$. However, the dark (all-zero) regions in the ground truth are not well defined in $F-measure$ when calculating $recall$ and $precision$. Under this circumstance, different foreground maps get the same result 0, which is apparently unreasonable. Thus, $F-measure$ is not suitable for measuring the results of non-salient object detection. However, both metrics of ε and F are based on pixel-wise errors and often ignore the structural similarities. Behavioral vision studies have shown that the human visual system is highly sensitive to structures in scenes [12]. In many applications, it is desirable that the results of the saliency detection model retain the structure of objects.

Pixel-wise Accuracy ε. The region similarity evaluation measure does not consider the true negative saliency assignments. As a remedy, we compute the normalized ($[0,1]$) mean absolute error (MAE) between M and G, defined as:

$$\varepsilon = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \|M(x,y) - G(x,y)\| \tag{4}$$

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

Structural Similarity S. $S-measure$ proposed by [12] evaluates the structural similarity by considering both regions and objects. Since the saliency of potential spacial structures is crucial to biomedical images, we additionally use $S-measure$ to evaluate the structural similarity between M and G.

Enhanced Alignment Measure E. Using the enhanced alignment matrix [13] ϕ_{FM} to capture the two properties (pixel-level matching and image-level statistics) of a binary map, we define $E-measure$ as:

$$Q_{FM} = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} \phi_{FM}(x,y) \tag{5}$$

where h and w are the height and width of the map, respectively. We use this measure to evaluate the foreground map (FM) and noise in images and to correctly rank the maps consistent with the application rank [13].

5.3. Quantitative Comparison

We perform a comparison of nine unsupervised saliency detection methods on our constructed adversarial data sets, including three traditional algorithms and six deep learning-based methods. All methods of deep learning are pre-trained using VGG16, and we have not used transfer learning with regards to the unsupervised setting. We applied the 4 metrics described above for evaluation. The experimental results are shown in Table 1.

Table 1 shows that on the $SNR = 0.5$ adversarial data set, traditional algorithms significantly outperform deep learning-based methods with a 2%-16% decrease in MAE.

Method / Datasets	LC	Itti	SR	integrated_grad	smoothed_int_grad	vanilla_grad	smoothed_van_grad	visual_bp	SalGAN
SNR=0.5	0.2495	0.5323	<u>0.1019</u>	**0.0793**	0.9894	0.0963	0.9878	9.0667	1.1028
SNR=1.5	0.2785	0.4774	<u>0.0721</u>	**0.0794**	0.9784	0.0813	0.9888	8.8969	1.2822

Table 2. Time performance of the 9 methods measured in seconds per image. The platform contains 2 Intel Core I7 CPUs and 2 NVIDIA GTX 1080 GPUs. LC and SR algorithms are implemented in Matlab and do not utilize GPU. Itti is in python and does not utilize GPU. Others are in python and use GPU. The best classic method is <u>underlined</u> and the best deep learning methods are in **bold**.

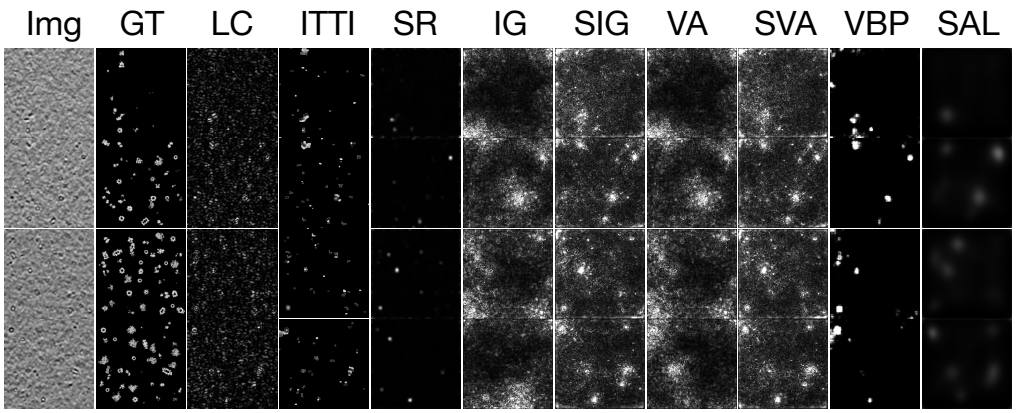

Figure 5. Results of the nine algorithms on the simulated data set with SNR = 0.5. Img is the short term for input image, GT = ground truth, LC = output of LC algorithm (same for Itti, SR), IG = integrated gradient, SIG = smoothed IG, VA = vanilla gradient, SVA = smoothed VA, VBP = visual back propagation, SAL = SalGAN

The $F-Measure$ and $S-Measure$ are improved by 17%-30% and 1%-6%, respectively. While on the $SNR = 1.5$ data set, traditional algorithms achieved even greater improvement over deep learning methods, with 30%-45% improve in $F-Measure$. The results support our hypothesis that VGG cannot automatically detect generally salient regions as well as classic methods. At the same time, it is worth noticing that deep learning-based models are superior to traditional algorithms in $E-Measure$. We believe that it may be related to the definition of the $E-Measure$, since $E-Measure$ applies the enhanced alignment matrix, which could compensate for the effect of low SNR and the limited tilt projection range (missing wedge effect). All in all, this is consistent with our assumptions, VGG is not as effective as we expected.

In addition to the above four metrics, we also analyze the time consumption per image of traditional algorithms and deep learning-based methods on an adversarial test set with 2,000 images, as is shown in Table 2. Traditional saliency detection algorithms prove to be time-efficient, even when compared to deep learning methods with GPU acceleration.

5.4. Qualitative Comparison

Figure 5 and Figure 6 show the saliency map predicted by nine unsupervised saliency detection methods on our adversarial data set with $SNR = 0.5$ and $SNR = 1.5$, which demonstrate traditional algorithms consistently outperform the deep learning-based methods. For our adversarial datasets with a low SNR, traditional algorithms always successfully detect multiple salient objects due to the superior capabilities of capturing low-level feature contrasts and predicting high values at object boundaries, while deep learning-based methods focus on assigning high saliency to regions that contain semantic information, which is not straightly conveyed due to low SNR in our adversarial data sets. This further proves our hypothesis that VGG's performance greatly depends on RGB cues and the quality of the training set. When facing grayscale biomedical data sets with low SNR, it is not as effective as we expected.

5.5. Visualizing what VGG pre-trained network learned

To answer our question in the introduction, what VGG has learned about saliency object detection in grayscale biomedical images, we display the feature maps output by various convolution and pooling layers in a network. Given a certain input, the output of a layer is often called its "activation", indicating the output of the activation function. This gives us a view of how an input feature is decomposed to the different filters learned by the network. As is shown in Figure 7, we can clearly find out that the VGG network has not learned any salient features from our grayscale biomedical image. The majority it learns is noise data, which is so even in the deepest block5_conv3 layer. From Figure 7 (c), (d) and (e), we found that VGG has produced completely false salient targets by comparison. Furthermore, in order

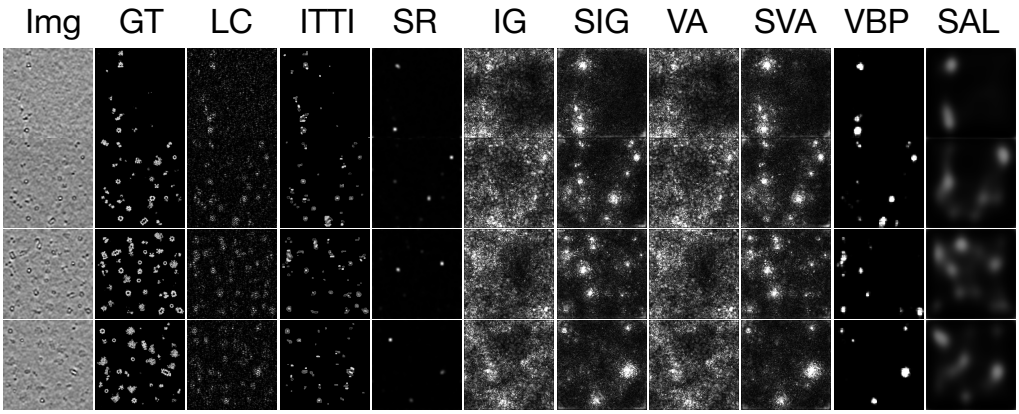

Figure 6. Results of the 9 algorithms on the simulated data set with SNR = 1.5. Img = input image, GT = ground truth, LC = output of LC algorithm (same for Itti, SR), IG = integrated gradient, SIG = smoothed IG, VA = vanilla gradient, SVA = smoothed VA, VBP = visual back propagation, SAL = SalGAN.

to more clearly visualize what VGG pre-trained network has learned, we show the feature maps of VGG16's intermediate layers on our adversarial data in Figure 8. This also verifies our hypothesis that VGG cannot actually learn any salient information on grayscale biomedical maps and is not capable of taking advantage of deep global features in the unsupervised saliency detection task.

6. Related Work

Detecting and segmenting salient objects from natural scenes, usually termed as salient object detection, has been a central problem in computer vision. According to [7], it is usually implemented as a process that includes two stages: 1) detecting the most salient object in a picture and 2) segmenting the accurate region of that object, although there are rarely models which explicitly distinguish between these two stages. Most of the current models attempt to find out the most salient object in the first stage, although theoretically their prediction maps can be used to find several objects within a picture. The second stage indicates that accuracy is only measured by the most salient object. Some behavioral [28] and computational [9] investigations used eye fixations as an approach to verify saliency predictions. Later Liu et al. [24, 24] and Achanta et al. [1] defined saliency detection as a binary segmentation problem.

Early models are typically based on general computational frameworks and psychological theories of bottom-up attention based on center-surround mechanisms [38, 40, 20]. Itti's algorithm [18] presented a visual attention system which combines multi-scale image features into a single topographical saliency map and then selects attended locations in order of decreasing saliency with a dynamical neural network. [44] designed a spatiotemporal video attention detection technique for detecting the attended regions that correspond to both interesting objects and actions in video sequences, which proposed the effective LC algorithm and shed new light on saliency detection. [42] proposed a bottom-up method based on global contrast with clustering. Using the mean shift filter, a few colors of the original image can be decreased, which is useful for the computation of saliency maps. The SR algorithm [16] extracts the spectral residual of an image in the spectral domain, and proposes a fast method to construct the corresponding saliency map in the spatial domain and requires no prior knowledge of the objects.

CNN-based methods [33, 37, 5, 35], trained with rich prior information, eliminate the need for hand-crafted features and alleviate the dependency on center bias knowledge, which has been adopted by many researchers. For example, [27] is a novel data-driven saliency prediction method trained with an adversarial loss function. Also, many supervised and unsupervised architectures adopted pre-trained networks as a feature extractor[43, 17, 33, 37, 5] in order to combine global features learned by deep layers, which are sensitive to large reception fields, with local details gained by shallow layers.

However, many state-of-the-art computational saliency models are relatively complex and inefficient and are only shown to be effective on RGB images. According to [7], a good saliency detection model should not only achieve high precision and resolution but also high computational efficiency: as front-ends to other complex processes, these models should detect salient regions quickly. Processing grayscale images lead to significant speedups than RGB images, and grayscale also preserves rich saliency information when compared to color images, which is experimented by [41]. Hence we believe that on grayscale biomedical images with a crowded context, traditional methods provide a much more efficient approach towards saliency detection. Prior to our work, [46] proposed a PCA method focusing on picking

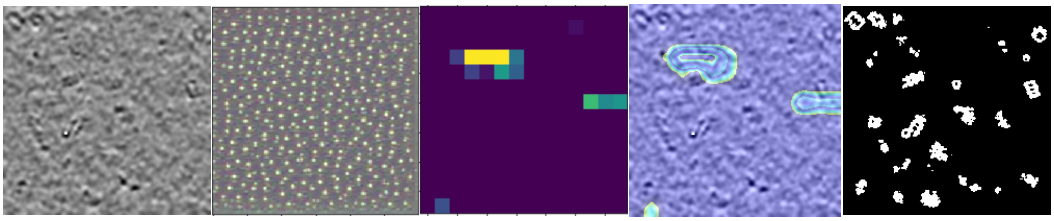

Figure 7. The process of visualizing intermediate activations, (a). Input image, (b). The image in layer block5_conv3, (c). Heatmap (for visualization purpose, we also normalize the heatmap between 0 and 1) in layer block5_conv3, (d). The original image superimposed with the heatmap, (e) Ground truth image.

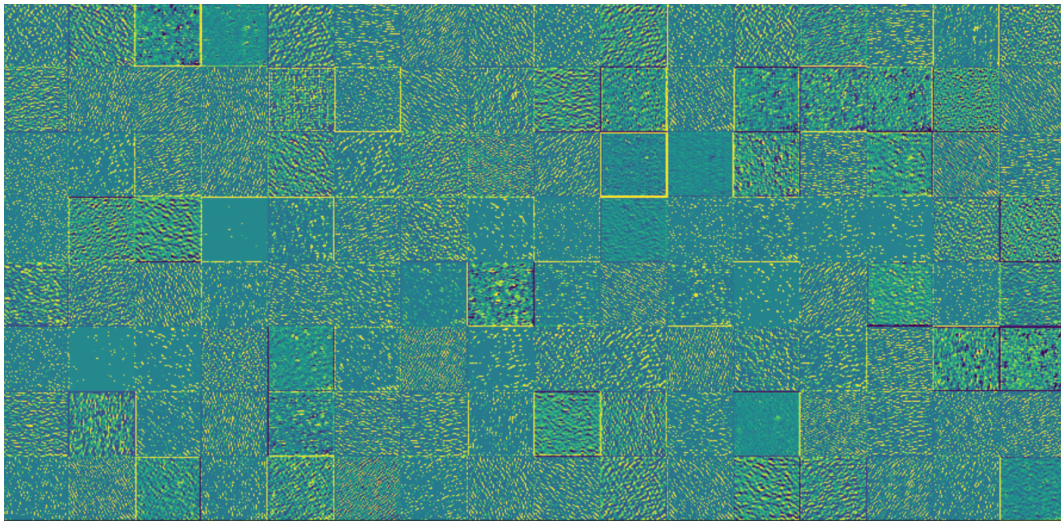

Figure 8. Feature maps of VGG16's intermediate layers on our adversarial data. The input image is the same as Figure 7(a).

out significant patterns which stands out from a large content, but they didn't not involve deep learning-based methods or the theoretical basis of the traditional algorithms.

7. Conclusion

In this paper, we discuss the unsupervised saliency detection task in biomedical imaging domain, but we find that pre-trained networks such as VGG is not as powerful as we expected when we apply it to a typical grayscale biomedical images data set.

Although previous works have confirmed that VGG is indeed a very strong learner as a pre-trained network, we doubt whether VGG is effective in unsupervised saliency detection for data sets with low SNR and no color information. To verify our hypothesis, we construct an adversarial data set featuring low SNR, low resolution and rich significant objects for the saliency detection task. Through probing experiments on our adversarial data set and visualizing feature maps from a pre-trained VGG network, we demonstrate in this study that VGG cannot reliably detect salient objects from grayscale biomedical images and its performance greatly depends on RGB cues and quality of the training set.

Furthermore, to the best of our knowledge, we are the first to conduct comprehensive experiments on unsupervised saliency detection methods with both classic and deep learning components on a biomedical related data set. Our work provides a novel insight for improving the performance of saliency detection models for biomedical image analysis. We believe that the adversarial data set should be adopted as a standard in future work on the saliency detection task. We hope that providing a more robust evaluation would help to fuel more productive research on this problem.

References

[1] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk. Salient region detection and segmentation. In *International conference on computer vision systems*, pages 66–75. Springer, 2008.

[2] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *IEEE Interna-*

tional Conference on Computer Vision and Pattern Recognition (CVPR 2009), number CONF, pages 1597–1604, 2009.

[3] M. Amirul Islam, M. Kalash, and N. D. Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7142–7150, 2018.

[4] H. M. Berman, T. Bhat, P. E. Bourne, Z. Feng, G. Gilliland, H. Weissig, and J. Westbrook. The pdb and the challenge of structural genomics.

[5] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, U. Muller, and K. Zieba. Visualbackprop: efficient visualization of cnns. *arXiv preprint arXiv:1611.05418*, 2016.

[6] A. Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[7] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li. Salient object detection: A survey. *Computational Visual Media*, pages 1–34, 2014.

[8] A. Borji and L. Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015.

[9] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2006.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 186–202, 2018.

[12] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017.

[13] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018.

[14] H.-P. Frey, C. Honey, and P. König. What's color got to do with it? the influence of color on visual attention in different categories. *Journal of Vision*, 8(14):6–6, 2008.

[15] S. Hamel, N. Guyader, D. Pellerin, and D. Houzet. Contribution of color information in visual saliency model for videos. In *International Conference on Image and Signal Processing*, pages 213–221. Springer, 2014.

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

[17] K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, X. Qian, and Y.-Y. Chuang. Unsupervised cnn-based co-saliency detection with graphical optimization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 485–501, 2018.

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

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[21] A. Kroner, M. Senden, K. Driessens, and R. Goebel. Contextual encoder-decoder network for visual saliency prediction. *arXiv preprint arXiv:1902.06634*, 2019.

[22] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[23] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[24] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2010.

[25] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[26] V. Lučić, A. Rigort, and W. Baumeister. Cryo-electron tomography: the challenge of doing structural biology in situ. *The Journal of cell biology*, 202(3):407–419, 2013.

[27] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[28] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123, 2002.

[29] L. Pei, M. Xu, Z. Frazier, and F. Alber. Simulating cryo electron tomograms of crowded cell cytoplasm for assessment of automated particle picking. *BMC bioinformatics*, 17(1):405, 2016.

[30] C. Poynton and B. Funt. Perceptual uniformity in digital image representation and display. *Color Research & Application*, 39(1):6–15, 2014.

[31] C. A. Poynton. Rehabilitation of gamma. In *Human Vision and Electronic Imaging III*, volume 3299, pages 232–249. International Society for Optics and Photonics, 1998.

[32] V. S. Prasath. Image denoising by anisotropic diffusion with inter-scale information fusion. *Pattern Recognition and Image Analysis*, 27(4):748–753, 2017.

[33] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

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

[38] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[39] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In *European conference on computer vision*, pages 825–841. Springer, 2016.

[40] J. M. Wolfe, K. R. Cave, and S. L. Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance*, 15(3):419, 1989.

[41] S. Yohanandan, A. Song, A. G. Dyer, and D. Tao. Saliency preservation in low-resolution grayscale images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018.

[42] C.-y. Yu, W.-s. Zhang, and C.-l. Wang. A saliency detection method based on global contrast. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(7):111–122, 2015.

[43] Y. Zeng, M. Feng, H. Lu, G. Yang, and A. Borji. An unsupervised game-theoretic approach to saliency detection. *IEEE Transactions on Image Processing*, 27(9):4545–4554, 2018.

[44] Y. Zhai and M. 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.

[45] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9029–9038, 2018.

[46] B. Zhou, Q. Guo, K. Wang, X. Zeng, X. Gao, and M. Xu. Feature decomposition based saliency detection in electron cryo-tomograms. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2467–2473. IEEE, 2018.