Tutorial for Salient object detection

顯著目標偵測

Author: Yu-Yao Huang

Edit by Jian-Jiun Ding

黃煜堯 著

丁建均教授編輯

2023.1

Abstract

Salient object detection (SOD) is the pre-process of several computer vision techniques such as visual tracking, image captioning, image segmentation and so on. By differentiate the most visually distinctive regions of an image, one can focus on these regions and design suitable algorithms for other computer vision tasks.

There are a variety of implementations for the SOD. The earliest but classic methods use hand-crafted features as cues to predict desired results. With technological progress, many neural-network-based methods are proposed. In this tutorial, we will introduce several approaches to implement salient object detection. Besides, we will also introduce some common datasets for the SOD and some widely used evaluation indices. The experiments with these algorithms and datasets will be presented in this tutorial.

We attempt to make the tutorial as concise as possible. Hope that all readers can learn something about salient object detection after reading this tutorial. Based on the knowledge of this tutorial, one can have their own idea and perspective of salient object detection.

Content

Chapter 1. Introduction

Salient object detection (SOD) is to find notable and significant objects in an image. When human beings see an image, we will attempt to segment an image into several regions or objects. For instance, the simplest method is to divide the image into the foreground regions and the background regions. Selecting the most attention-attractive region out of all regions is called as salient object detection. Besides, unlike the object detection using bounding boxes, the SOD requires predicting regions with clear boundaries. Owing to its essential role in computer vision, the SOD is widely used in broad applications, such as image and video segmentation, visual tracking, AR/VR and so on.

There are lots of methods about salient object detection. According to input data, we can divide existing SOD methods into three categories: (1) RGB images, (2) RGB-D images, (3) light field images. Each category has their advantages and disadvantages. RGB images is the basic type of input data for all computer vision tasks because one can easily obtain massive RGB images. Moreover, in consideration of practicality, the most widely-used data type users can get is RGB images. [1] and [2] are two of classic algorithms using RGB images as input data. However, when accuracy rate keeps increasing with technological advancement, researchers hope for a breakthrough for salient object detection. Depth map is one of the ways researches experiment. Depth map can provide appearance saliency cues which RGB images cannot convey when the scene is too complex. [3] can been seen as one of the most popular approaches with RGB-D images. Despite the fact that RGB-D images can offer depth information to boost performance of models, it is hard to get high-quality depth maps or only can get noisy depth maps. Light field images are proposed to resolve this problem. Light field images consist of an all-focus central view and a focal stack. This data type contains abundant spatial information and clear boundaries with high resolution. Therefore, light field images become one of the most prevalent type of input data in the field of salient object detection. [4] is the state-of-the-art method adopting light field images as inputs.

From the perspective of algorithms, we can divide all approaches into two topics: (1) traditional methods, (2) neural networks. First of all, the earliest but classic methods use hand-crafted features as cues to implement salient object detection. These hand-crafted features include color, contrast, edge, texture, etc. [5] is one of the common methods. Besides, [6] use super-pixel segmentation as pre-process to predict saliency maps. Nevertheless, traditional methods have heavily constraint and limitation. After the convolution neural network (CNN) proposed, neural networks become the most

prevalent way in computer vision tasks. This condition becomes more serious after the fully convolution network (FCN) [7] be put forward. With flexibility and robustness, there are numerous different architectures of neural networks. [8], [9] and [10] are all approaches with neural networks but adopt different architectures.

[11] interprets the evolution history of salient object detection and introduce each popular framework of SOD. In this tutorial, we refer to [11] and will focus on pros and cons of each methods instead of detailed algorithms or development of salient object detection. In addition, to save time, we only focus on the most classic and the newest algorithms.

In Chapter 2, we will first introduce some traditional methods to make readers have basic knowledge about the SOD. In Chapter 3, the content contains several neural-network-based algorithms. In Chapter 4, we introduce some common datasets for SOD and how we evaluate performance of the SOD. And then, we conduct experiments with some SOTA methods of salient object detection. Last but not least, in Chapter 5, we will summarize all prevalent architecutres and suggest future outlook.

Chapter 2. Traditional Methods

2.1. Blocked-based methods

Blocked-based methods are the earliest methods of salient object detection. Limited by computation resource, all methods adopt a low-level approach by evaluating contrast between the regions in the images. Contrast of two regions are determined by features like colors, edges, intensities and so on. [5] and [12] are the approaches based on this kind of methods.

In [5], the authors use frequency maps as breakthrough. Set two thresholds: a low-frequency threshold and a high-frequency threshold. On the one side, to emphasize the largest salient objects and highlight whole salient regions, one needs to set low-frequency threshold as low as possible. On the other side, to preserve clear boundaries, high-frequency threshold is needed to be as high as possible. However, if it becomes too high, it will be interfered by noise or blocking artifacts.

According to mentioned analysis in the previous paragraph, the authors of [5] decide a wide range of frequency between low-frequency threshold ω_{lc} and ω_{hc}. Afterwards, use combinations of multiple difference of Gaussians (DoG) as proposed band-pass filter. We can view DoG as an edge detector. Therefore, if we combine all

difference of Gaussians, it is equivalent to combine all edge detectors with different scales. Besides, it is important how we choose parameters of DoG. After experiments, set σ_1 as infinity and σ_2 as a small gaussian kernel to filter noise. At the end, the authors set Euclidean distance of average images and Gaussian blurred images as saliency value to get the final saliency object prediction.

The algorithms mentioned above is totally based on features in regions of images without any neural networks or super-pixels. This kind of methods are basic but essential. From the perspective of salient object detection, one can only use traditional feature extraction with some techniques of salient object detection to derive adequate results.

2.2. Super-pixel-based methods

In case of limited computation resource, people hope that computation cost could be as minimal as possible. Therefore, instead of taking whole image in consideration, one can take a pre-process called the super-pixel segmentation before other processes. The super-pixel segmentation is the approach which classifies regions by uniform color, brightness, etc. After the super-pixel segmentation, one can view a super-pixel as an unit to conduct following task. By this skill, computation cost can be greatly reduced. Furthermore, due to the development of the super-pixel segmentation, boundaries of object in images will be perfectly preserved. These properties are good for salient object detection so some researchers had their eyes on the super-pixel-based methods.

One of the common super-pixel-based methods is [6]. Pixel-wise contrasts make computation cost larger. In [6], the authors introduce a novel analysis of contrast: Region Contrast (RC). The RC combines spatial relationship with region contrasts. First of all, segment an image into several regions and then calculate color contrasts between regions. After calculating, define saliency values by summing up weighted contrasts between each region with other regions. It is noteworthy that weighted coefficients are decided by the spatial distance between two regions. By the algorithm mentioned above, we can get saliency maps with clear boundaries. Meanwhile, save time and cost less computation resource.

[25] is also a super-pixel-based method. This method introduces the generic knowledge of object into salient object detection. The authors predict saliency values by evaluate contrast between multi-scale local regions which is segmented by multiple algorithms of super-pixel segmentations. After getting saliency values, combine these regional values and predict the final salient object detection. This hierarchical structures is often adopted in later papers and this paper is the basis of this kind of methods.

Note that the super-pixel segmentation is a widely-used pre-process of many compute vision tasks. However, due to its irregular shape and size, it is hard to directly implement super-pixel segmentation on neural-network-based architectures. Hence, if we want to adopt the super-pixel segmentation, we must introduce some converting mechanisms to make super-pixels be suitable for neural networks like CNN or FCN.

2.3. Extrinsic cues models

As the title implies, instead of only defining saliency values by information of an image itself, models introduce something extrinsic like ground-truth of datasets, sequence of videos, similar images and so on. It is obvious that if one can get substantial relevant data, one can derive nearly perfect saliency maps. Through this statement is too ideal to be realized, it is an effective way to improve performance of salient object detection. Even now, using extrinsic cues as external input of neural networks or using extrinsic cues as information of post-process is often took on by several state-of-the-art algorithms.

In this section, we only focus on traditional models with extrinsic cues such as [13] and [14]. Take [14] for instance. Given a group of related images, this paper proposed a novel algorithm for interactive co-segmentation of a foreground object. Assuming that there are several stone images, some of images are scribbled. The scribbled regions are foreground and background respectively. By these cues, the authors propose an algorithm which guides outputs of the co-segmentation. What is special about this method is that instead of making methods unsupervised via extrinsic cues, the authors design an algorithm which allow users to scribble on images. This small change makes the whole method simpler and highly parallelizable energy function. However, human beings will make mistakes sometimes. Therefore, the proposed method includes an automatic recommendation system that would suggest users where the most likely foreground candidate is.

The method mentioned in the previous paragraph is a way of salient object detection. Nonetheless, the SOD is often as a pre-process of other computer vision tasks. Due to this property, we hope algorithms of the SOD could be as efficient as possible. Although [14] can form a virtuous circle, this process cost too much time, especially over great number of images.

Chapter 3. Machine Learning Method

3.1. Convolutional Neural Networks

The traditional salient object detection extract features manually and evaluate contrasts and various prior knowledge one by one. At the end, combine all of features to predict the final results. With the advance of science and technology, the Convolutional Neural Network (CNN) become the most prevalent and the most robust way in computer vision. The reason is that the CNN can extract features under multi-scale and multi-layer. Automatically take local and global features into account and do extremely complex operations. In this instance, performance of every computer vision task reached a new level.

[1] is the standard salient object detection with the convolutional neural network. The authors design two models for global context and local context respectively. For the global-context model, use a super-pixel-centered window padded with mean pixel value. For the local-context model, use a closer-focused super-pixel-centered window. On the one hand, the global-context branch robustly models saliency with few large errors. On the other hand, the local-context branch focuses on details to refine the saliency prediction of the centered super-pixel. At the end of the model, combine the features from the global-context model and the local-context model and predict final results. With the interaction of global information and local information, one can get a precise saliency map. The model of this paper is very simple and easy to understand. This is the basic of the neural networks in salient object detection. The following methods can be regarded as an extension of this method.

Another architecture of the convolutional neural network is [2]. This method makes the whole process end-to-end. First of all, the images are thrown into a coarse model to extract global features by automatically learning various global structured saliency cues. In the paper, the authors state that the global features include contrast, compactness, and their optimal combination. After this step, the authors propose a new hierarchical convolutional neural network. This architecture can progressively refine the details of saliency maps step by step via integrating local context information. With this two-stage strategy, the model can cleverly combine global information and local information. Besides, due to its end-to-end property, this method achieves a real-time speed with high accuracy rating.

The convolutional neural networks are all the rage after introducing. This kind of methods give new direction of compute vision tasks including salient object detection.

3.2. Fully Convolutional Networks

The methods mentioned above can prove that the convolutional neural network is the right way of salient object detection. However, this obviously cannot meet researchers' requirements. Therefore, the fully convolutional network (FCN) [7] be proposed.

Because the CNN uses fully connected layer at the end of models to predict probabilities, it is hard to apply to semantic segmentation tasks or other pixel-wise tasks. The reason is that predicted probabilities are one-dimensional and loss of spatial information cause models not to conduct pixel-wise classification. The main idea of the FCN is to replace fully connected layers with convolutional layers and allow any size of images as input. Due to these properties, FCN can implement on semantic segmentation tasks or other pixel-wise tasks and make them end-to-end. These methods are also applied to salient object detection. However, FCN like VGG16, ResNet etc. cannot directly be adopted on the SOD because it is not design for this kind of tasks. Therefore, the common way is to propose a novel architecture with the FCN as backbones.

[8] is not only basic but also innovated method. How to extract features and combine features with global information and local information has been the hard challenges for salient object detection. In [8], the authors propose a two-level nested U-structure. This nested structure allows the model to capture more contextual information from multi-scale due to its multi-size receptive fields. In addition, if we want models be robust, the usual and direct way is making it deeper. Nevertheless, this operation causes computation cost larger. On the contrary, the model proposed in [8] increases the depth of the whole architecture without significantly increasing the computational cost owing to the introduced pooling operation. Unlike other models, two-level nested U-structure is the extension of basic neural networks with tidy architecture. [8] shows that the model of neural networks can be very flexible. The neurons can be connected not only in series or in parallel but also in vertical perspectives. With this three-dimensional structure, models can get more robust.

[23] is an unsupervised salient object detection. One of the problems of salient object detection is the large requirements of images. However, common unsupervised methods cannot reach high performance as supervised ones. It is because how to generate pseudo labels and extract useful features from pseudo labels are very challenging. The authors generate a pixel-wise uncertainty map by repeatedly analyzing the pros and cons of different pseudo labels. Each handcrafted method has its own advantages and disadvantages. The authors attempt to find the high-performance part

of pseudo labels and discard the low-performance part of pseudo labels. By this technique, model can refinement pseudo ground-truth by uncertainty maps and get the better feedback to the model. By this, the model can be trained well and predict accurate results. Besides, the authors propose a merge-and-split module to parallelly analyze all pixel-wise labels. This efficient way helps whole process save time. Unsupervised approaches have always been a challenge. This method provides a new way of thinking.

Until now, the FCN still be the most prevalent methods in many fields of compute vision. Its robust performance breaks through again and again with the development of hardware like CPUs and GPUs. Nowadays, still many novel methods and algorithms are proposed.

3.3. Transformers

The CNN and the FCN achieve new heights. To solve the problem of extraction of multi-scale information and combination, global pooling layer, non-local module, and layer-to-layer connecting etc. are proposed. These approaches are just operating and comparing between certain layers instead of really global information. Hence, Transformer has been introduced to salient object detection in recent years.

The Vision Transformers (ViT) [15] is the substitute of CNN in many computer vision tasks. The original field of transformer is NLP. To convert it into the field of computer vision, the authors propose two methods: (1) change inputs, (2) change the mechanism of self-attention.

At first, the model split images into several patches. Given patches, convert them into D-dimensional vector by linear projection of flattened patches and then throw them into transformer. This operation is called as Patch Embedding. This is because images are a continuum and pixels are units with low information. Both of them are not appropriate to be inputs of transformer. Therefore, authors convert images into vector sequence by linear projection of flattened sequences of image patches. Besides, as transformers requiring position embedding in the field of NLP, the transformer in the field of CV also needs to add patch embedding with 1-dimensional position embedding.

Furthermore, a transformer consists of several multiheaded self-attention (MSA) and multi-layer perceptron (MLP). After throw inputs into transformer, they will pass by layer normalization (LN) and go into MSA. Adding another input with the results of MSA by residual connection can get the intermediate results. After then, pass by one layer of LN and MLP and implement residual connection again. We will get the final results of a transformer. This operation will conduct N times. Tokens of images will be

predicted. One can perform any down-stream tasks with these tokens of images.

[16] is the extension of vision transformer. It connects an inverse transformer after the original transformer. By this operation, this model can be viewed as an encoder and a decoder. For an encoder, transformer embeds an image into tokens which carry information of images. For a decoder, reverse the process of encoder and convert tokens back into images. In [16], the authors set the ground-truth of models as saliency maps. Therefore, the whole architecture will fit salient object detection. The encoder will find proper tokens which is suitable for the SOD. In addition, this paper not only use images as inputs but also use depth maps as inputs. The distance of object often be the essential but crucial features in salient object detection. If accurate depth maps can be derived, it must helpful to predict saliency maps. Adding depth maps to models is now be a popular operation to supply one more information to models which can boost performance in evidence. This paper is one of the few papers about the transformer in salient object detection. [16] only improves a little part of models but still gets state-of-the-art performance. This research shows that the Transformer has excellent prospects.

Vision transformer and the models with same frameworks have flexible domain of prediction due to less inductive biases. The reason is that the CNN has the properties of locality and spatial invariance which raise inductive biases. Moreover, size of receptive field by depth of layer of transformer is always the same regardless of depth. This may make transformer have better performance than convolutional neural network. However, transformer still has its own disadvantages. One of its disadvantages is that the performance of transformer highly depends on size of datasets. If users do not have enough large datasets, it is hard to train transformers up to saturation. There is no denying that transformers will be the one way of salient object detection in the future.

3.4. Others

No one restricts an architecture to only one method. Therefore, it is well to combine more than one method in one model. [22] and [24] is the instance for this classification.

[23] is a basic method which combine the super-pixel-based method and the CNN. This method split the model into two parts. The first part predicts a roughly saliency map by global information. In addition, using super-pixel segmentation to provide some basic local context information to refine the predictions. The second part use the former coarse saliency map as inputs. Refine the details by the CNN and then output the final salient object detection. This method is very simple. However, it still tells us that the traditional methods still have their own value. Using as pre-process or post-process is

not a complex operation but may improve the performance of a model.

[24] is a method which looks very simple at the first glance. From the perspective of extracting features, the transformer is better than the CNN. However, due to the large receptive field size, segmentation results of the transformer are incomplete and many details are lost. The authors propose a one-stage framework called Pyramid Grafting Network (PGNet). PGNet are the model which combines CNN and transformer. Extract features by the transformer and then graft the features from transformer branch to CNN branch. Let each branch plays to their strengths. This method shows that there is no single method that can do the best, but that one can combine the advantages of various methods to make models better.

Table 1 shows the comparison of the methods mentioned above. This table will discuss in terms of performance, predicting speed, computation cost and data requirement. The methods using neural network like the FCN and the Transformer are with the best performance. Comparatively, the traditional methods are with worse performance. Besides, the CNN cannot reach the high performance as the FCN due to its unsuitability for salient object detection. The time complexity of these methods is under the same situation. This time, the CNN is with the same time complexity with the FCN and Transformer. This time complexity is about testing time. From the perspective of training time, the Transformer takes more time than FCN. However, neural-network-based methods still have their own disadvantages. They require more computation resources and data requirements. This is because training models is a big work. Therefore, this problem is every developer and researcher attempt to overcome. By the way, data requirements of the Transformer are much higher than the CNN and the FCN. Last but not least, traditional methods are not outdated. The techniques of pre-process or post-process has been useful. On the one hand, one can implement traditional methods as pre-process to alleviate loadings of training model and computation resources. On the other hand, one can apply post-process to improve results, such as boundaries refinement or bounding box filtering. Therefore, while focusing on new approaches, do not forget about traditional ones.

	Performance	Training Time	Predicting Time	Computation cost	Data Requirement
Block-based	low		slow	low	low
Super-pixel-based	low		slow	low	low
Extrinsic cues	low		slow	low	low
CNN	middle	fast	fast	high	high
FCN	high	fast	fast	high	high
Transformer	high	slow	fast	high	high

Table 1

Chapter 4. Experiments

4.1. General datasets

There are so many prevalent datasets for salient object detection, such as PASCAL-S [17], DUTS [18], ECSSD [19], DUT-OMROM [20] and so on. The datasets for salient object detection are demanded to include the following properties: (1) a huge number of images and the corresponding ground-truth, (2) clear boundaries of objects. Apart from what is just mentioned, there are some additional requirements. One of them is the hope that make the resolution of images as higher as possible. Another one is the hope that types of objects in datasets could be as much as possible.

PASCAL-S is a dataset for salient object detection consisting of a set of 850 images from PASCAL VOC 2010 validation set with multiple salient objects on the various scenes. For mean absolute error (MAE), this kind of evaluation indices on PASCAL-S is lower than 5 percent. Besides, for S-measure, this kind of evaluation indices achieves up to 87.5 percent.

DUTS is a salient object detection dataset which contains 10553 training images and 5019 test images. On the one hand, training images are selected from the ImageNet DET training sets and validation sets. On the other hand, test images are collected from the ImageNet DET test sets and the SUN datasets. There are diverse objects and challenging scenarios in the datasets which make it credible. For mean absolute error (MAE), this kind of evaluation indices on DUTS is lower than 2.75 percent. Besides, for S-measure, this kind of evaluation indices achieves up to 88 percent.

4.2. Evaluation indices

The common evaluation indices include mean absolute error (MAE), S-Measure, F-Measure and E-Measure. In this chapter and the following experiments, we will introduce MAE and S-Measure and conduct experiments with these two evaluation indices.

The mean absolute error (MAE) is to find the error between a predicted image S and ground-truth G. The formula is below:

$$\mathrm{MAE} = \frac{1}{W * H} \sum_{x=1}^{W} \sum_{y=1}^{H} \|S(x,y) - G(x,y)\|$$

The W and H in the formula is the width of the image and the height of the image

respectively. This is the basic but effective methods to measure the performance of an algorithm. The more powerful an algorithm is, the lower a MAE is.

The full name of S-measure is structure-measure [21]. How to evaluate performance of the saliency maps is important for the development of salient object detection. Other evaluation indices are based on pixel-wise errors and often ignore the similarities of structure. Therefore, the authors proposed a new and efficient measurement which simultaneously evaluates region-aware and object-aware structural similarity between a predicted image and a ground-truth image. The more powerful an algorithm is, the higher S-measure is.

4.3. U^2-Net Going Deeper with Nested U-Structure for Salient Object Detection

Both training and testing are conducted on a six-core, twelve-thread personal computer with an AMD Ryzen 5600x 3.7 GHz CPU, 32GB RAM and a NVIDIA GeForce RTX 3070 GPU with 8 GB memory. Table 2 shows the evaluation indices of the U^2-Net with the PASCAL-S dataset and the DUTS dataset. Besides, Fig. 1 and Fig. 2 demonstrate the results of the U^2-Net.

	MAE	**S-measure**
PASCAL-S	0.083	0.76
DUTS	0.085	0.788

Table 2

Fig 1

Fig 2

4.4. Visual Saliency Transformer

Both training and testing are conducted on a six-core, twelve-thread personal computer with an AMD Ryzen 5600x 3.7 GHz CPU, 32GB RAM and a NVIDIA GeForce RTX 3070 GPU with 8 GB memory. Table 3 shows the evaluation indices of the Visual Saliency Transformer with the PASCAL-S dataset and the DUTS dataset. Besides, Fig. 3 and Fig. 4 demonstrate the results of the Visual Saliency Transformer.

	MAE	**S-measure**
PASCAL-S	0.066	0.89
DUTS	0.069	0.88

Table 3

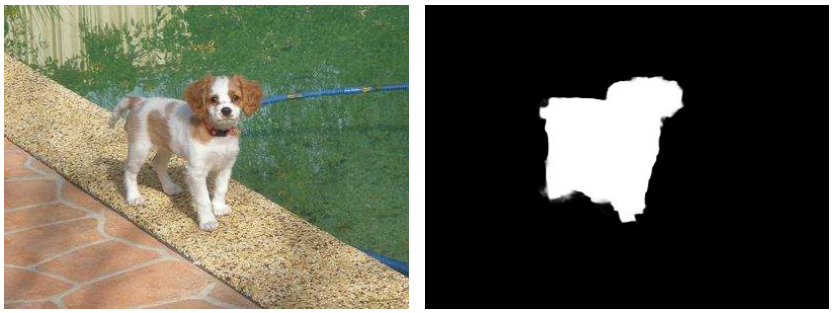

Fig 3

Fig 4

Chapter 5. Conclusion

Salient object detection is a very essential but crucial field in the computer vision. In this tutorial, introduce traditional methods like block-based methods, super-pixel-based methods and extrinsic methods. The differences of these methods are from the perspective of regions of interest and sources of features. Also, introduce neural-network-based methods like CNN-based methods, FCN-based methods and Transformer-based methods. The differences of these neural-network-based methods are the main structure of models.

Each classification has their own advantages and disadvantages which are mentioned in the previous paragraphs. One need to focus on the modern techniques but also need to learn from the thoughts of traditional methods. It is because there are no methods come from nowhere. All methods are improved by earlier methods. Only when one knows every detail and structure of each problem and solution, one can develop their own ideas and approaches. Hope this tutorial can give readers a basic knowledge of salient object detection and an inspiration of salient object detection. If one is interested in a particular paper, please refer to the reference in this tutorial for more details.

Reference

[1]. R. Zhao, W. Ouyang, H. Li and X. Wang, "Saliency detection by multi-context deep learning," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1265-1274, doi: 10.1109/CVPR.2015.7298731

[2]. N. Liu and J. Han, "DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 678-686, doi: 10.1109/CVPR.2016.80.

[3]. Ji, W., Li, J., Bi, Q., Guo, C., Liu, J., & Cheng, L. (2022). Promoting Saliency from Depth: Deep Unsupervised RGB-D Saliency Detection. International Conference on Learning Representations.

[4]. M. Feng, K. Liu, L. Zhang, H. Yu, Y. Wang and A. Mian, "Learning from PixelLevel Noisy Label : A New Perspective for Light Field Saliency Detection," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1746-1756, doi: 10.1109/CVPR52688.2022.00180

[5]. R. Achanta, S. Hemami, F. Estrada and S. Susstrunk, "Frequency-tuned salient region detection," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1597-1604, doi: 10.1109/CVPR.2009.5206596.

[6]. M. -M. Cheng, G. -X. Zhang, N. J. Mitra, X. Huang and S. -M. Hu, "Global contrast based salient region detection," CVPR 2011, 2011, pp. 409-416, doi: 10.1109/CVPR.2011.5995344.

[7]. J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431-3440, doi: 10.1109/CVPR.2015.7298965.

[8]. Qin, Xuebin and Zhang, Zichen and Huang, Chenyang and Dehghan, Masood and Zaiane, Osmar and Jagersand, Martin, “U2-Net: Going Deeper with Nested U-Structure for Salient Object Detection,” Pattern Recognition 2020, pp. 107404, arXiv:2005.09007

[9]. N. Liu, N. Zhang, K. Wan, L. Shao and J. Han, "Visual Saliency Transformer," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4702-4712, doi: 10.1109/ICCV48922.2021.00468.

[10]. M. Feng, K. Liu, L. Zhang, H. Yu, Y. Wang and A. Mian, "Learning from Pixel-Level Noisy Label : A New Perspective for Light Field Saliency Detection," 2022

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1746-1756, doi: 10.1109/CVPR52688.2022.00180.

[11]. Borji, A., Cheng, MM., Hou, Q. et al. Salient object detection: A survey. Comp. Visual Media 5, 117–150 (2019). https://doi.org/10.1007/s41095-019-0149-9

[12]. Achanta, R., Estrada, F., Wils, P., Süsstrunk, S. (2008). Salient Region Detection and Segmentation. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds) Computer Vision Systems. ICVS 2008. Lecture Notes in Computer Science, vol 5008. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79547-6_7

[13]. M. Wang, J. Konrad, P. Ishwar, K. Jing and H. Rowley, "Image saliency: From intrinsic to extrinsic context," CVPR 2011, 2011, pp. 417-424, doi: 10.1109/CVPR.2011.5995743.

[14]. D. Batra, A. Kowdle, D. Parikh, J. Luo and T. Chen, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3169-3176, doi: 10.1109/CVPR.2010.5540080.

[15]. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929.

[16]. N. Liu, N. Zhang, K. Wan, L. Shao and J. Han, "Visual Saliency Transformer," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 4702-4712, doi: 10.1109/ICCV48922.2021.00468.

[17]. Y. Li, X. Hou, C. Koch, J. M. Rehg and A. L. Yuille, "The Secrets of Salient Object Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 280-287, doi: 10.1109/CVPR.2014.43.

[18]. L. Wang et al., "Learning to Detect Salient Objects with Image-Level Supervision," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 3796-3805, doi: 10.1109/CVPR.2017.404.

[19]. Tran, R., Patrick, D., Geyer, M., & Fernandez, A. (2020). SAD: Saliency-based Defenses Against Adversarial Examples. ArXiv, abs/2003.04820.

[20]. C. Yang, L. Zhang, H. Lu, X. Ruan and M. -H. Yang, "Saliency Detection via

Graph-Based Manifold Ranking," 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013, pp. 3166-3173, doi: 10.1109/CVPR.2013.407.

[21]. DengPing Fan, MingMing Cheng, YunLiu, et al. Structure-measure: A new way to evaluate foreground maps[C]. IEEE ICCV, 2017.

[22]. Chen T, Lin L, Liu L, et al. Disc: Deep image saliency computing via progressive representation learning[J]. IEEE transactions on neural networks and learning systems, 2016, 27(6): 1135-1149.

[23]. Yi Ke Yun, Takahiro Tsubono: “Recursive Contour Saliency Blending Network for Accurate Salient Object Detection”, 2021; [http://arxiv.org/abs/2105.13865 arXiv:2105.13865].

[24]. Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, Jia Li: “Pyramid Grafting Network for One-Stage High Resolution Saliency Detection”, 2022; [http://arxiv.org/abs/2204.05041 arXiv:2204.05041].

[25]. H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng. Automatic salient object segmentation based on context and shape prior. In BMVC, 2011.