

Studies Advanced in Salient Object Detection based on Deep Learning

Xiji Zhang *

Finance and Economics University of Jiangxi, Jiangxi China

* Corresponding author email: 2202004487@stu.jxufe.edu.cn

Abstract. A very important job in the area of computer vision is salient object detection, its goal is to check salient or obvious objects in images or videos. Thanks to rapidly developing networks of convolutive neurons, especially generative adversarial networks, CGAN has become a popular framework for salient object detection, which has promoted breakthroughs in detection precision and velocity of protruding objects. Focusing on the topic of salient object detection, in this paper, we systematically introduce the research progress of salient object detection. Specifically, we first introduce methods for detecting protracted objects based on general object detection algorithms (such as Fast R-CNN, YOLO, and SSD), including their basic ideas, key steps, advantages and disadvantages, etc. We then focus on CGAN-based salient object detection methods, including method classification, typical method introduction, and application examples. Finally, we quantitatively compare the performance of different saliency detection algorithms on common datasets, and discuss the existing research problems and feasible solutions of the topic.

Keywords: Salient Object Detection; Generating Adversarial Networks; Convolutional Neural Network; Recurrent Neural Network; Computer Vision.

1. Introduction

A very important task in the field of computer vision is salient object detection., aiming at detect salient or salient objects in images or videos. Saliency object detection is used extensively in a wide range of applications, such as video surveillance, image retrieval, natural scene understanding [1]. However, salient object detection remains a difficult task because of the large amount of background noise and the variability of protruding objects in different images.

Early salient object detection methods were mostly inspired by manually crafted features. Itti et al. have successfully migrated visual saliency from cognitive psychology and neurology to the field of computer vision, making salient content that can be calculated. Liu et al. [2] regarded the significance detection problem as a binary segmentation problem for the first time, using local features, regional features, and global features to represent salient objects. Thereafter, the salient detection gradually separated from the biological model and turned to the computational model based on the underlying features, which greatly improves the accuracy of the detection results. Most traditional salient object detection methods are embodied in edge detection, which extracts edge information by filtering and differentiating images using a Laplacian operator or other forms of second-order differentiation [3]. Edge detection helps to distinguish the salient objects in the images from the background. In general, the traditional salient object detection is highly targeted, and generally only applicable to scenes with clear subjects, and object color and structure is monotonous. Meanwhile, they are often sensitive to the angle, light and size. Due to poor generalization ability and inability to describe and adapt complex scene and object structures, the study of such saliency detection methods is stuck and stuck.

In recent years, with the improvement of computer computing power and the successful application of deep learning-related models in many visual tasks, Deep learning has gradually become the main method in this field, with an unstoppable trend. Although it takes a lot of time in manual annotation and sometime in network training and parameter adjustment, once the network model is trained, it can be applied to salient object detection in more complex scenarios, and many of its functions are significantly better than traditional salient object detection methods. It uses deep neural networks to model the data. Deep learning methods have high accuracy in the field of salient object

detection and the ability to learn from data is also a characteristic of it. This makes deep learning methods the first choice for many people when performing salient object detection. According to the framework, the existing salient object detection based on deep learning may be subdivided into three categories: CNN-based, RNN-based and GAN-based.

Focusing on the three aspects, in this paper, we review the salient object detection methods based on CGAN, including the classification of methods, the introduction of typical methods and examples of application. First, we will introduce the traditional salient object detection methods. Then, we will introduce the salient row object detection based on CGAN, including its operating principle, advantages and disadvantages, etc. Finally, we will introduce some application examples and provide some cases in practical applications.

2. Methods

2.1 Salient Object Detection based on CNN

The main applications of CNN can be divided into image classification, semantic segmentation and object detection. Here, we will introduce the classical general object detection methods and salient object detection algorithms extended.

Fast R-CNN is a region-based convolutional neural network method proposed by Ross Girshick in 2015. The Fast R-CNN method first uses the convolutive neuronal array to extract image characteristics, then uses the region proposal network to generate multiple candidate regions, and finally classifies each candidate region using a classifier. The Fast R-CNN method has high precision, and it uses the region-proposed method to reduce the search space and improve the speed. The complete name of YOLO is you look only once, which refers to the category and location of objects within the image. Because you only need to see it once.

YOLO can also be considered a region-free method. Compared to region-based methods, YOLO is characterized by not needing to find regions that are likely to contain objects ahead of time. In other words, a classic, foundational region-based method works like this: people use computer graphics (or deep learning) to analyze an image, find some places where the computer thinks there may be objects, crop them down, and put them into a picture classifier, classified by the classifier.

SSD, short for Single Shot MultiBox Detector, which was proposed by Wei Liu at ECCV 2016, is also an object detection algorithm. Until now, it is also one of the main detection frameworks. Its speed advantage is more evident when compared to faster RCNN network. The core of SSD is using a network of convolutional neurons to extract data from the image. And then using a classifier to detect objects on the image. Similar to the YOLO method, the SSD method has a faster speed because it also performs only one scan. However, the SSD method has a higher accuracy rate than the YOLO method.

2.2 Salient Object Detection based on RNN

When the input object is a sequence of cultural inputs that clearly has contextual features, such as predicting the next frame in a video, it is obvious that these outputs depend on previous inputs. This means that the network must have some sort of memory so that it can understand the content it has seen before given the current input. To address this problem of memory in networks, a special kind of structured neural network known as recurrent neural network (RNN) gradually emerged.

RNN (Recurrent Neural Network) is an artificial neural network similar to other neural networks, but it has the nature of a temporal cycle, allowing it to process data in the form of a sequence. This makes the RNN well suited for processing sequential data, such as temporal data, speech data, and video data. A striking feature of the RNN is that its weights are shared, meaning that it uses the same weights to handle all the elements in the input sequence. This allows the RNN to capture long-range dependencies in the sequence that is intractable to other types of neural networks. There are many different variants of RNN, including LSTM (long and short-term memory network) and GRU (gated cycle unit). The purpose of these variants is to improve the ability of the RNN to better handle long-

term dependence. Unlike CNN, RNN is characterized by processing image or numerical data, and because the network itself has the memory ability to learn the correlated data types. For example, for language translation or text translation, the words before and after in a sentence usually have a certain relationship, but the CNN network cannot learn this relationship, and RNN has memory, so the performance will be better. Because text can be understood through RNN, other applications such as entering an image, but output a sentence about the narrative of the image.

RNN-based salient object detection methods mainly include time-based modeling methods, sequence-based modeling methods and attention-based modeling methods. For time-based modeling methods, RNNs are used to simulate continuous changes in time to help detect objects. For example, one approach is to use an RNN to predict changes in the next frame of the image, and then use those predictions to help detect objects. In the sequence-based modeling method, RNN is used to simulate the appearance and disappearance of objects in the sequence. For example, one approach is to use an RNN to predict which frames in a sequence contain objects, and then use those predictions to help detect objects. Whereas, attention-based modeling methods use RNN to simulate the attention mechanism to help the detector focus on important parts of the sequence. For example, one approach is to use an RNN to generate attention weights, and then use these weights to weight each frame in the sequence to help detect objects.

2.3 Salient Object Detection based on GAN

Compared to the conventional CNN and RNN, Generative Adversarial Network (GAN) is a revolution in deep learning, which has produced some major technologic can produce high-quality images. GAN has produced some major technological breakthroughs. The core of the Generative Adversarial Network has two parts: a generator(G) network, which is responsible for generating synthetic data, and a discriminator(D) network, which is responsible for determining the authenticity of the input data. The goal of the G network is to continually improve the synthetic data it produces to the point where the D network is unable to distinguish between the synthetic and real data, However, the D network should also continually improve itself in order to make more accurate determinations. Because of this relationship between the two networks, it creates a form of competition, thus the name Generative Adversarial Network (GAN) The core idea of GAN is an iterative the struggle between the G and the D. The G generates fake data, and then feeds both the generated fake data and real data into the D. The D has to decide which is true and which is false. The G and D are optimized alternately until balanced.

The basic idea of implementing saliency detection with GAN network is as follows: (1) Use the G to generate false saliency images. The G is trained to give images with fake saliency, and then the D is trained to distinguish real saliency images from false saliency images. Ultimately, the D can be used to detect objects in salient images. (2) Evaluate the authenticity of the saliency image using a D. The D is trained to distinguish real saliency images from false saliency images. Ultimately, the D can be used to evaluate the authenticity of the saliency image and detect objects in the input image. (3) The G and the D collaborate to generate salient images. The G and the D collaborate to generate salient images, and the G tries to generate images that can pass the D. Finally, the generated saliency images can be used to detect objects.

CGAN is a variant of the generative adversarial network (GAN), where both generators and discriminators are composed of convolutional layers. CGAN-based salient object detection methods usually include the following steps:

Data preprocessing: The purpose of this step is to convert image data into a format suitable for the training of generative and discriminative models. This usually includes preprocessing of images, encoding of labels, and preprocessing of other relevant information. (2) Model design: The goal of this step is to compose and fine-tune the structural spaces of the generator and discriminator models. Convolutional neural networks (CNN) or other deep learning models are usually used as generative and discriminative models. (3) Model training: The purpose of this step is to train training the above two models, so that the generated virtual image is more and more real, while the discriminative model

is more and more difficult to distinguish between the real image and the virtual image. A back-propagation algorithm is usually used to train the model. (4) Model evaluation: The purpose of this step is to evaluate the performance of the model. Commonly used evaluation indicators include accuracy, recall, F1 score, average accuracy and average recall. (5) Model application: The purpose of this step is to apply the trained model to the actual salient object detection task. Often, using models to identify important objects within an image. requires the use of image segmentation algorithms or other visual techniques.

Compared to conventional methods of detecting protruding objects, the advantages of the methods based on CGAN are as follows. First, it can use additional conditional information to generate virtual images similar to the real image, which helps to make the model have higher adaptability to new samples and less overfitting phenomenon. Second, the CGAN-based salient object detection method can utilize the powerful representation ability of deep learning skills to enhance the accuracy and recall of the model. Finally, the CGAN-based salient object detection methods can be cultivated on complex datasets to improve the model performance.

3. Experiments

3.1 Common Dataset

Common datasets for obvious object detection mainly include the MSRA, SOD, MSRA10K, HKU-IS, DUTS, SED, ECSSD, DUTO-OMRON, PASCAL-S, whose detailed information are list in Table 1.

Table 1. Commonly used data sets for salient object detection

Dataset name	Number of images	Annotation form	Significant objects in each image
ASD([3],2009)	1 000	Pixel by pixel	1
MSRA-A([4],2011)	20 840	Calibration box	1
MSRA-B([4],2011)	5 000	Calibration box	1
MSRA10K([4],2011)	10 000	Calibration box	1
SED2([5],2019)	100	Pixel by pixel	2
DUT-OMRON([6],2013)	5 168	Calibration box, pixel by pixel human eye attention point	1 or n
ECSSD([7], 2021)	1 000	Pixel by pixel	1 or n
PASCAL-S([8], 2022)	850	Calibration box, pixel by pixel human eye attention point	1 or n
HKU-IS([9], 2015)	4 447	Pixel by pixel	1 or n
XPIE([10], 2017)	10 000+8 598	Pixel by pixel	0, 1 or n
DUTS([11], 2022)	10 553+5 019	Pixel by pixel	0, 1 or n
ILSO([12],2017)	1 000	Pixel-level instance annotation contour annotation	0, 1 or n
SOC([13], 2018)	6 000+	Pixel-level instance annotation contour annotation	0, 1 or n

3.2 Evaluation Metrics

To evaluate the capability indicators of different methods, several metrics are developed, which mainly includes the Precision, Recall, F1 score, mAP and mAR.

(1) Precision: This is a common measure to measure the performance of a detection algorithm. It represents the proportion of detected objects that meet the standard to the total number of objects. The higher accuracy indicates the better ability of the model.

(2) Recall: This is a common measure to measure the performance of a detection algorithm, which represents the proportion of detected objects that meet the standard to the actual number of objects that exist. A higher recall rate indicates a better performance of the model.

(3) F1 score: This is a mixture of accuracy and recall rate. The higher the value, the better ability of the detection algorithm.

(4) Mean Average Precision (mAP): This is a common measure of the performance of a detection algorithm. It represents the average accuracy over all images. The higher the average accuracy indicates that the greater the ability of the model.

(5) Mean Average Recall (mAR): This is a common measure of the performance of the detection algorithm. It represents the average recall over all images, and it represents the average proportion of all targets that the model can correctly detect. The higher average recall indicates the superior ability of the model.

3.3 Performance Comparison and Analysis

Table 2 presents the comparison of the performance of the traditional salient object detection algorithm on multiple datasets. F_β Indicates the maximum F-measure value, the larger the F_β , the better the algorithm performance; MAE(M) represents the average absolute error, the smaller the better.

Table 2. Performance comparison of different salient object detection

Methods	MSRA-B		ECSSD		SED2		DUT-OMRON	
	F_β	M	F_β	M	F_β	M	F_β	M
GMR([6],2013)	0.839	0.126	0.736	0.189	0.643	0.163	0.591	0.189
UFO([14], 2013)	0.862	0.15	0.739	0.207	0.729	0.18	0.593	0.173
HDCT([18], 2014)	0.877	0.143	0.74	0.199	0.758	0.162	0.643	0.164
ST([16], 2014)	0.896	0.122	0.777	0.193	0.768	0.145	0.635	0.182
RBD([17], 2014)	0.884	0.108	0.757	0.173	0.75	0.13	0.647	0.144
RC([18], 2015)	0.875	0.137	0.776	0.187	0.649	0.148	0.621	0.189
MBD([19], 2015)	0.89	0.107	0.785	0.172	0.759	0.137	0.636	0.168

The above methods achieve excellent results in detecting some simple pictures, while for some complex and missing pictures, the results are not very ideal. for example, the HDCT algorithm has a MAE value of 0.143 on the MSRA-B dataset, but the MAE value on the ECSSD dataset is 0.199. This is because the algorithm over-relies on the underlying features of the image, lead to the inability to judge the reality of the changing scene. Table 3 Comparison of the ability of the significance detection methods for deep learning on multiple datasets. Red represents the CNN based algorithm and blue represents the GAN based algorithm.

Table 3. Performance comparison of different salient object detection

Methods	DUT-OMRON		ECSSD		HKU-IS		PASCAL-S		DUTS-test	
	F_β	M	F_β	M	F_β	M	F_β	M	F_β	M
LEGS ([20], 2019)	0.669	0.133	0.868	0.115	0.770	0.118	0.752	0.157	0.612	0.137
MDF ([9], 2015)	0.694	0.092	0.833	0.108	0.860	0.129	0.764	0.145	0.657	0.114
ELD ([21], 2011)	0.719	0.091	0.865	0.081	0.844	0.071	0.767	0.121	0.737	0.092
cGANw ([22],2018)	0.622	0.143	0.779	0.113	0.747	0.109	0.693	0.164	0.674	0.142
ALN ([23], 2020)	-	-	0.926	0.045	0.913	0.038	0.833	0.101	0.819	0.065

According to the content of the above table, we can see that the CNN-based algorithm is generally better stable than the model trained by the CGAN-based algorithm in one data set. Like the MAE of LEGS in the DUT-OMRON data set is 0.133, which is higher than the other two MDF and ELD, while the MAE value of cGANw is 0.143, which is obviously higher than LEGS. However, the ALN algorithm is an obvious exception. It is also based on the significance detection algorithm of generating the adversarial network, but it is more stable than the model trained by the above three CNN-based algorithms in the same data set. It is obvious to see that the F_β values of the models trained by ALN in the ECSSD, HKU-IS, PASCAL-S and DUTS-test are higher than those of the three algorithms based on CNN, and the MAE values are also lower than those of LEGS, MDF and ELD. For example, the F_β value of ALN in the PASCAL-S data set is 0.833, higher than 0.752 (LEGS), 0.764 (MDF), and 0.767 (ELD); the MAE is 0.101, lower than 0.157 (LEGS), 0.145 (MDF), and 0.121 (ELD).

In general, the model trained by the CNN-based algorithm in the same data set is more stable than the GAN model, but GAN can indeed increase the accuracy of the model (such as ALN) in specific aspects, making the model more efficient.

4. Discussion

The CGAN-based salient object detection method is an effective detection method that has obvious advantages in improving the detection effect. First, since the CGAN model relies on random noise z , the generated spurious samples may not be realistic in some cases, thus affecting the learning effect of the discriminant D . Second, because the generation process of the CGAN model is nonlinear, it is difficult to effectively interpret the generated spurious samples, thus limiting the application of the model. Finally, because the training process of the CGAN model depends on adversarial training, it is difficult to guarantee the stability of the training process, which affects the accuracy of the model. In addition, it requires a large amount of training data, which may lead to longer training times. Second, CGAN-based salient object detection methods may be difficult to cope with complex image scenes and may produce poor generated image quality. Finally, because the CGAN-based salient object detection method relies on the training of generative and discriminative models, the problem of model generalization and insufficient.

In the future, we hope to be able to improve the CGAN-based method to solve the above problems, making this method more practical for salient object detection. Specifically, we hope to introduce a new random noise distribution or improve the structure of generator G to make the generated false samples more realistic, so as to improve the learning effect of discriminator D . In addition, we hope to make the generation process of the CGAN model more interpretable, so as to expand the application field of the model. Finally, we hope to optimize the training process to make the balance of the training process is not easily disrupted, so as to improve the adaptivity of the model. The specific improvement ideas are as follows:

(1) Use more data for training: The more training data, the better the model can generalize. Therefore, one can try to train with more data to (1) increase the generalizability of the model.

(2) Use more complex models: the more complex the model, the more information you can learn. Therefore, more complex models can be tried to enhance the model's performance on new data.

(3) Using data augmentation: Data augmentation is an effective way to increase the difference of waiting data and, consequently, enhance the generalizability of the model. For example, images can be rotated, flipped, zoom, etc., to generate more training data.

(4) Use regularization: Regularization is an effective way to protect the model from overfitting, and thus make the model able to handle new data better when encountered. For example, regularization methods such as tenure value decay, Dropout, etc.

(5) Using multi-task learning: Multi-task learning is an effective way to train multiple tasks at once, thus improving the ability of the model to recognize and adapt to new data. For example, challenges

such as image feature classification and object detection can be performed simultaneously using CGAN-based saliency object detection methods.

5. Conclusion

This paper mainly introduces the methods, datasets, evaluation criteria and their application in many fields based on CGAN. We first introduce traditional saliency detection methods, and then introduce several classes of commonly used deep learning-based, salient object detection algorithms. Then, we divided it into three categories and took the GAN out for comparison. Next, we analyze the CGAN-based salient object detection method and look into possible future directions.

References

- [1] Jiang Fengling, Kong Bin, Qian Jing, et al. Review on Salient Object Detection [J]. Measurement & Control Technology, 2021.
- [2] Liu Nian, Han Junwei, Zhang Dingwen, et al. Predicting eye fixations using convolutional neural networks [C]// Proc of IEEE International Conference on CVPR. USA: Las Vegas, Nevada, 2015: 362-370.
- [3] Achanta R, Hemami S, Estrada F, et al. Frequency-tuned salient region detection [C]// Proc of IEEE International Conference on CVPR. USA: FL, 2009: 1597-1604.
- [4] Liu Tie, Yuan Zejian, Sun Jian, et al. Learning to detect a salient object [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2011, 33 (2): 353-367.
- [5] Songlong Zhang. Image salience detection based on fully convolutional neural networks[J]. Computer software and computer applications University of The South 2019(12).
- [6] Yang Chuan, Zhang Lihe, Lu Huchuan, et al. Saliency detection via graph-based manifold ranking [C]// Proc of IEEE International Conference on CVPR. USA: Portland, Oregon, 2013: 3166-3173.
- [7] Yihan Zhang. Deep learning-based image saliaunce detection [J]. Hebei Normal University, 2021 (09).
- [8] Liting Tang. Research and application of the improved significance detection algorithm[J]. Computer software and computer applications Jiangsu University of Science and Technology 2022(10).
- [9] Li Guanbin, Yu Yizhou. Visual saliency based on multiscale deep features [C]// Proc of IEEE International Conference on CVPR. MA: Boston, 2015: 5455-5463.
- [10] Xia Changqun, Li Jia, Chen Xiaowu, et al. What is and what is not a salient object? Learning salient object detector by ensembling linear exemplar regressors [C]// Proc of IEEE International Conference on CVPR.
- [11] Jianzhe Wang Research on the saliency object detection algorithm based on convolutional neural networks[J]. Computer software and computer applications 2022(06).
- [12] Li Guanbin, Xie Yuan, Lin Liang, et al. Instance-level salient object segmentation [C]// Proc of IEEE International Conference on CVPR. USA: Honolulu, Hawaii.
- [13] Fan Dengping, Cheng Mingming, Liu Jiangjiang, et al. Salient objects in clutter: Bringing salient object detection to the foreground [C]// European Conference on Computer Vision. Cham: Springer, 2018: 186-202.
- [14] Jiang Peng, Ling Haibin, Yu Jingyi, et al. Salient region detection by ufo: Uniqueness, focusness and objectness [C]// Proc of IEEE International Conference on CVPR. Australia: NSW, 2013: 1976-1983.
- [15] Kim J, Han D, Tai Y, et al. Salient region detection via high-dimensional color transform [C]// Proc of IEEE International Conference on CVPR. Ohio: Columbus, 2014: 883-890.
- [16] Liu Zhi, Zou Wenbin, Le Meur O. Saliency tree: A novel saliency detection framework [J]. IEEE Trans on Image Processing, 2014, 23 (5): 1937-1952.
- [17] Borji A, Cheng Mingming, Jiang Huaizu, et al. Salient object detection: A survey [J]. Computational Visual Media, 2014, 5 (2): 117-150.
- [18] Cheng Mingming, Mitra N J, Huang Xiaolei. Global contrast based salient region detection [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2015.

- [19] Zhang Jianming, Sclaroff S, Lin Zhe, et al. Minimum barrier salient object detection at 80 fps [C]// Proc of the IEEE International Conference on Computer Vision. Chile: Santiago, 2015: 1404-1412.
- [20] Wang Tiantian. Research on the image saliency detection algorithm[J]. Computer software and computer applications. Dalian University of Technology. 2019(02).
- [21] Wang Yu. An evolutionary algorithm study based on multi-method fusion[J]. Computer software and computer applications. University of Science and Technology of China. 2011(09).
- [22] Ji Yuzhu, Zhang Haijun, Wu Q M. Saliency detection via conditional adversarial image-to-image network [J]. Neurocomputing, 2018, 316: 357-368.
- [23] Wu Yong, Liu Zhi, Zhou Xiaofei. Saliency detection using adversarial learning networks [J]. Journal of Visual Communication and Image Representation, 2020, 67: 102761.