Text-Guided Salient Object Detection

Zixian Xu[1], Luanqi Liu[1,*], Yingxun Wang[1], Xue Wang[1] and Pu Li[2]
[1]*Qilu Institute of Technology, Shandong, China*
[2]*Guangzhou College of Technology and Business, Guangzhou, China*

Keywords: Salient Object Detection, Natural Language.

Abstract: Salient object detection (SOD), a core task in the field of computer vision, is dedicated to accurately identifying the salient objects in images. Unlike previous research methods, this study recognizes the key role of textual information in salient object detection and thus proposes a unique text-based range control method for salient object detection. In this method, we introduce the semantic labels from the CoSOD3K dataset into a pre-trained text-driven semantic segmentation model to align the textual and image feature information. Subsequently, the image features are analyzed for saliency through a salient object detection network. Through the SFE (Salient Feature Extractor) module, we fuse the extracted salient features with the semantically aligned features to derive the saliency detection results. Experimental results show that the robustness and efficiency of our framework surpass existing salient object detection methods. Users can guide the detection process through natural language interaction, expanding applications such as image editing and data annotation, and to some extent solving challenges like complex backgrounds, multi-scale issues, and blurry boundaries. This offers the potential for new breakthroughs in the field of salient object detection.

1 INTRODUCTION

The goal of computer vision is to enable machines to "see" and "understand" their environment, with salient object detection being one of its important tasks. The aim of this task is to identify salient, eye-catching objects within images. These objects attract the observer's attention due to their distinctiveness or differences in context.

Traditional salient object detection methods mainly rely on low-level visual cues or deep learning techniques to extract and analyse image features. However, these methods often face difficulties when dealing with complex backgrounds, multi-scale issues, and blurred boundaries. Moreover, they frequently overlook the value of text information in enhancing detection performance.

To address this issue, we propose a new solution a text-based salient object detection range control method. In this approach, we incorporate the semantic labels from the CoSOD3k dataset (Fan D P, 2021) into a pre-trained text-driven semantic segmentation model to align text information with image feature information. Then, we utilise a salient object detection network to conduct saliency analysis on the image features.

Through the SFE module, we fuse the extracted saliency features with the semantically aligned features to derive the saliency detection results. Experimental results show that our framework outperforms existing salient object detection methods in terms of robustness and efficiency. Additionally, the detection process can be guided by natural language interaction, opening up new possibilities for applications such as image editing, data annotation, and more.

With this study, we aim to provide an effective solution to the challenges of salient object detection in complex backgrounds, multi-scale issues, and blurred boundaries, paving the way for new breakthroughs and opportunities in the field of salient object detection.

Our contributions can be summarized as follows:

1. We propose a novel text-guided salient object detection framework that integrates natural language information to guide the detection process, expanding possible applications such as image editing and data annotation.

2. This research introduces the SFE module, which combines salient features with semantically aligned features and uses upsampling techniques to derive saliency detection results. This innovative

Xu, Z., Liu, L., Wang, Y., Wang, X. and Li, P.
Text-Guided Salient Object Detection.
DOI: 10.5220/0012284400003807

In *Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2023)*, pages 381-385
ISBN: 978-989-758-677-4

approach improves the robustness and efficiency of salient object detection.

3. We have conducted extensive experiments, demonstrating that our approach exhibits superior capability in addressing common challenges encountered in real-world images, such as complex backgrounds, multi-scale issues, and blurred boundaries.

2 RELATED WORK

Salient object detection, a salient research field within computer vision, has accumulated a wealth of studies. In early research (Yang C, 2013), salient object detection relied primarily on low-level visual features such as color, texture, and shape. These methods performed quite well in some simple scenarios but fell short in handling images with complex backgrounds or multi-scale issues.

In recent years, advancements in deep learning have brought new opportunities for salient object detection. Many studies utilize Convolutional Neural Networks (CNNs) to automatically extract rich features from images to enhance the performance of salient object detection (Ji Y, 2021). However, most of these studies depend on the internal information of images, overlooking higher-level semantic information.

Simultaneously, some studies have started exploring how to incorporate semantic information into salient object detection, such as improving salient object detection using semantic segmentation (Wang L, 2018). Although these works have utilized semantic information to a certain extent, they have not fully capitalized on the potential of text information.

Therefore, this study focuses primarily on how to effectively merge text information and image features to boost the performance of salient object detection. We propose a new text-based salient object detection method that improves the performance of salient object detection by aligning text information and image features and using the SFE module for feature fusion and upsampling. Our method shows clear advantages in facing the challenge of salient object detection.

3 PROPOSED METHOD

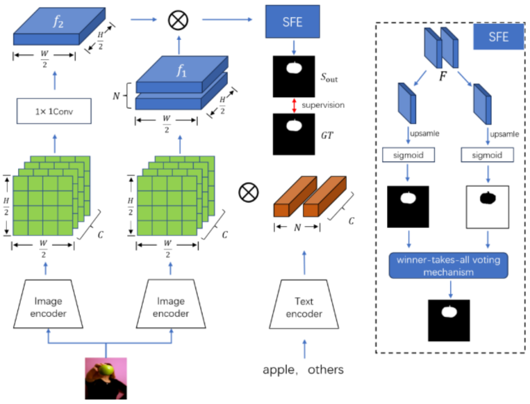

Figure 1. The core framework of this paper.

Our framework is founded on a language-driven semantic segmentation model that embeds text labels and image pixels into a common space. Building on this, we introduced an additional image encoder， the generated encoding information processed through a 1 × 1 convolution layer, resulting in a single-channel encoding with dimensions matching the common space. This encoding is subsequently broadcasted throughout the common space, with each feature being assigned a salient score. Following this, the SFE module employs depthwise convolution to process the space and performs upsampling on each channel, this generates the saliency detection results corresponding to each text label. As shown in Figure 1.

3.1 Text Encoder and Image Encoder

To ensure an effective alignment of text encoding and image encoding, we opted to utilize the associated encoders from LSeg (Boyi Li,). Specifically, the text encoder of framework is grounded on the CLIP (Radford A, 2021) pre-trained model, which outputs a set of vectors that are invariant to the sequence of input labels and can flexibly accommodate a varying number, N, of labels. On the flip side, framework adopts the DPT (Ranftl R, 2021) structure in its image encoding, a strategy that allows the image encoder to deeply harness features and semantics within images. By leveraging the pre-trained models of LSeg that already achieve text and image alignment, the training workload for our project has been significantly reduced.

3.2 Saliency Score Broadcasting

After processing through the image encoder, the input image is embedded into a feature matrix I of size $H \times W \times C$, the text label is encoded to produce a T matrix, with a size of $N \times C$. we correlate them by the inner product, creating a tensor f_1 of size $H \times W \times N$. which is defined as follow:

$$f_1 = I \cdot T \tag{1}$$

In order to assign saliency scores to each feature channel within this common space, we use a 1×1 convolution to mapping the image encoding to a feature matrix f_2 of size $H \times W \times 1$. Then, we broadcast it into the common space, weighting the salient features in each channel. The final common space F is defined as:

$$F = f_1 \cdot f_2 \tag{2}$$

3.3 SFE Module

Each channel of the common space F represents the features of an object associated with a specific text. During the upsampling process, interactions between channels should not occur. We opted for depthwise convolution, which allows for individual upsampling of each channel, returning to the original input resolution, and producing saliency detection results via a sigmoid function. In the final step, we use a winner-takes-all voting mechanism (Shin G, 2022) to determine the most accurate saliency detection outcome across the channels. As a result, we need a BCE loss to supervise this process, defined as follows:

$$BCE(x, y) = -\frac{1}{n}\sum_{i=1}^{n}[y \log x + (1-y)\log(1-x)] \tag{3}$$

The variable x represents the saliency map output by the network, while y is the ground truth label. The purpose of salient object detection is to find salient areas in the input image rather than make judgments based on specific semantics. Therefore, when assigning saliency scores, we only need to use the dot product to allocate the learned saliency results to each channel of the encoding. This ensures that the learning of saliency scores is carried out over the entire input image without deviating from the objective of salient object detection, and it reasonably allocates scores to salient regions corresponding to each semantic piece of information. The common space F are supervised by the following loss for the decoded saliency map S_{out}:

$$L = BCE(S_{out}, z_i) \tag{4}$$

Where S_{out} is the saliency map output by our framework, and z_i is the ground-truth labels, i represents the image number in the dataset.

4 EXPERIMENT

4.1 Dataset

Our saliency training set, CoSOD3k, contains 3316 images spread across 13 categories. Each category is accompanied by its specific text labels and saliency outcomes. For joint salient object detection, we perform evaluations on three datasets: CoSal2015 (Wang C, 2019), CoCA (Zhang Z, 2020), and CoSOD3k.

4.2 Evaluation Metrics

We adopt three criteria in our experiment, including F-measure (F_β (Achanta R, 2009)), Mean Absolute Error (MAE), and E-measure (E_m) to quantitatively evaluate the performance of our method. F-measure is calculated by:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{5}$$

where β 2 is set to 0.3 as in (Achanta R, 2009). MAE is computed by:

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{W} \times \boldsymbol{H}}\sum_{x=1}^{W}\sum_{y=1}^{H}|\boldsymbol{p}_i, \boldsymbol{g}_i| \tag{6}$$

where p_i and g_i are prediction and ground truth.

E-measure capture global statistics and local pixel matching information with an alignment matrix ϕFM as:

$$\boldsymbol{E_m} = \frac{\mathbf{1}}{\boldsymbol{W} \times \boldsymbol{H}}\sum_{x=1}^{W}\sum_{y=1}^{H}\boldsymbol{\phi FM(i, j)} \tag{7}$$

where $\boldsymbol{H}$ and $\boldsymbol{W}$ are the height and width of the image, respectively.

4.3 Quantitative Results

Table 1 presents the scores of our method compared to other Salient object detection methods. Our approach achieves saliency detection by simultaneously detecting multiple images and using their shared class labels as textual cues for saliency. By leveraging text to determine specific semantic information, our method effectively detects salient objects. Experimental results demonstrate the effectiveness of our approach.

Table 1. Quantitative comparison of our method with other methods. ↓ and ↑ denote the smaller the better and the larger the better.

Methods		CoSal2015			CoSOD3k			CoCA	
	F_β ↑	MAE ↓	E_m ↑	F_β ↑	MAE ↓	E_m ↑	F_β ↑	MAE ↓	E_m ↑
UMLF	0.7298	0.2691	0.6841	0.6895	0.2774	0.6541	0.7512	0.2514	0.7154
DIM	0.6363	0.3126	0.6243	0.5603	0.3267	0.6012	0.6571	0.2915	0.6814
CSMG	0.8340	0.1309	0.7915	0.7641	0.1478	0.7353	0.8411	0.1219	0.9061
Our	**0.8715**	**0.0695**	**0.8244**	**0.8527**	**0.0721**	**0.8122**	**0.8815**	**0.0581**	**0.8426**

4.4 Qualitative Results

From figure 2, we can observe that our method achieves more accurate detection results for salient objects. This is primarily because our approach fully utilizes textual cues to assist in identifying salient objects within the images. When dealing with multiple images, by analyzing their shared class labels, our method can discern genuinely salient portions and effectively filter out background noise and irrelevant objects, ensuring superior performance in the experiments.

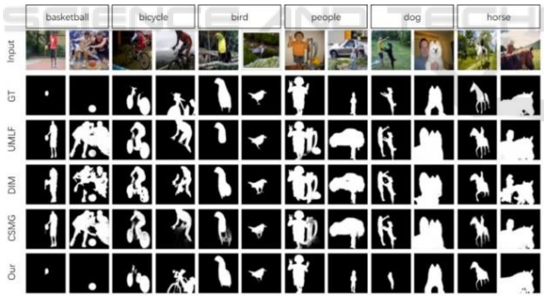

Figure 2. The saliency detection results under specific labels.

4.5 Impact of Different Text Labels on Results

As shown in Figure 3, In the same input image, the detection range produced by the saliency detector changes when different object labels are given. Notably, due to the flexibility of the text encoder, our detection results can also map to text labels that haven't been trained on, such as 'animal' in (c). Although it hasn't been trained on this label, it can still locate the 'dog' and 'horse' in the image which belong to the 'animal' category. This demonstrates that the model can successfully segment other objects using an extended label set.

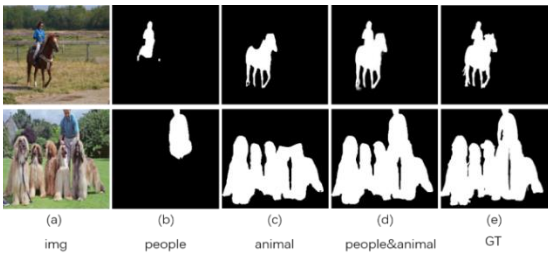

Figure3. For the different detection results produced by different label frameworks

5 CONCLUSION

Salient object detection is a fundamental task in computer vision, aiming to identify prominent objects in images. This study introduces a novel approach that harnesses semantic labels and a pre-trained text-driven model to enhance the accuracy and controllability of salient object detection. Our method surpasses existing techniques in terms of robustness and efficiency and allows users to guide the detection process through natural language. It holds the potential to address challenges such as blurry boundaries and complex backgrounds, paving the way for breakthroughs in salient object detection.

REFERENCES

Fan D P, Li T, Lin Z, et al. Re-thinking co-salient object detection(J), IEEE transactions on pattern analysis and machine intelligence, 2021, 44(8): 4339-4354. https://doi.org/10.1109/tpami.2021.3060412

Yang C, Zhang L, Lu H, et al. Saliency detection via graph-based manifold ranking(C), Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3166-3173. https://doi.org/10.1109/cvpr.2013.407

Ji Y, Zhang H, Zhang Z, et al. CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances (J), Information Sciences, 2021, 546: 835-857. https://doi.org/10.1016/j.ins.2020.09.003

Wang L, Wang L, Lu H, et al. Salient object detection with recurrent fully convolutional networks(J), IEEE transactions on pattern analysis and machine intelligence, 2018, 41(7): 1734-1746. https://doi.org/10.1109/tpami.2018.2846598

Boyi Li, Kilian Q Weinberger, Serge Belongie, et al. Language-driven Semantic

Segmentation(C), International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2201.03546

Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision(C), International conference on machine learning. PMLR, 2021: 8748-8763. https://doi.org/10.48550/arXiv.2103.00020

Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction(C), Proceedings of the IEEE/CVF international conference on computer vision. 2021: 12179-12188. https://doi.org/10.1109/iccv48922.2021.01196

Wang C, Zha Z J, Liu D, et al. Robust deep co-saliency detection with group semantic (C), Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 8917-8924. https://doi.org/10.1609/aaai.v33i01.33018917

Zhang Z, Jin W, Xu J, et al. Gradient-induced co-saliency detection(C), Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer International Publishing, 2020: 455-472. https://doi.org/10.1007/978-3-030-58610-2_27

Achanta R, Hemami S, Estrada F, et al. Frequency-tuned salient region detection(C), 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009: 1597-1604. https://doi.org/10.1109/cvpr.2009.5206596

Shin G, Albanie S, Xie W. Unsupervised salient object detection with spectral cluster voting(C), Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3971-3980. https://doi.org/10.1109/cvprw56347.2022.00442