



Lightweight Model For Saliency Prediction Project Report

Students:

Abdallah Hammad

Mohammad Almatrafi

Fawaz Albahussein

Abdullah Bin-Talib

Abdullah Al-Qahtani

AI Summer School

Mentor: Salma Kharrat

KAUST Academy

Aug 31, 2024

Abstract

Salient object detection (SOD) focuses on identifying the most visually prominent parts of an image, typically those that stand out due to distinct colors. In our project, we present a straightforward SOD method that leverages color separation to enhance feature extraction. This technique reduces computational demands while maintaining high performance, making it ideal for resource-constrained devices. Despite the simplicity and lightweight nature of our model, we achieved results that surpass expectations.

1. Introduction

Salient Object Detection (SOD) represents a crucial task within the realm of computer vision, aimed at the automatic identification and segmentation of the most visually prominent regions in an image. By emulating human visual attention mechanisms, SOD is integral to a variety of applications, such as image and video editing, object tracking, content-aware image resizing, and autonomous navigation. The precise localization of salient objects is essential for enhancing human-computer interaction, optimizing visual search processes, and enabling intelligent systems to concentrate on pertinent information.

Despite notable progress in deep learning-based SOD models, many existing frameworks are characterized by substantial computational demands and elevated memory usage, which impede their implementation on platforms with limited resources. This constraint significantly restricts the practical use of SOD in real-world contexts, including mobile devices, embedded systems, and Internet of Things (IoT) applications. To tackle this issue, the present paper introduces an innovative lightweight architecture for salient object detection.

Our methodology emphasizes the reduction of model complexity while preserving performance levels by employing a channel-wise splitting technique during the input stage. By decomposing the RGB input image into its constituent channels and processing them separately, we aim to effectively extract distinctive features. This design decision is informed by the understanding that salient objects frequently display varying characteristics across different color channels.

The structure of this report is organized as follows: the related work section offers a thorough review of related literature on salient object detection, outlining the challenges and shortcomings of current methodologies. The methodology section elaborates on the proposed lightweight architecture, detailing the channel-wise partitioning approach and the overall network design. The experiments section presents experimental findings and comparisons with leading-edge techniques. Finally, we conclude with a discussion of the implications of our work and potential avenues for future research.

2. Related Work

2.1 Introduction to the Field

Salient object detection (SOD) is a vital component in the broader field of computer vision, primarily focused on identifying the most visually prominent parts of an image. This field has gained significant attention due to its applications in image processing, object recognition, and human-computer interaction. The fundamental idea behind SOD is rooted in human visual perception, where certain regions of an image naturally attract more attention due to their distinctive features such as color, brightness, and texture. The relevance of SOD continues to grow as it plays a crucial role in enhancing the efficiency and effectiveness of various downstream tasks, including image segmentation, object tracking, and visual scene analysis.

2.2 Categorization of Existing Methods

2.2.1 Conventional Methods

The early approaches to SOD were predominantly based on bottom-up processes that mimic the human visual system. Itti et al. [1] introduced one of the pioneering methods, where multi-scale image features are combined into a saliency map through a dynamical neural network inspired by the primate visual system. Graph-based visual saliency (GBVS), proposed by Harel et al. [2], is another notable technique that generates saliency maps by normalizing activation maps across different feature channels. These conventional methods, while effective in their time, are often limited by their inability to capture complex visual patterns and their relatively high computational requirements.

2.2.2 Deep Learning-Based Methods

The advent of deep learning (DL) has revolutionized SOD, enabling the development of models that can automatically learn and extract hierarchical features from images. Early works like eDN and DeepGaze [3][4] marked the beginning of using convolutional neural networks (CNNs) for saliency detection. These methods leverage large-scale image datasets and transfer learning to improve their performance on saliency tasks. Recent models like UNISAL and DeepGaze II [5][6] have pushed the boundaries by integrating attention mechanisms and leveraging pre-trained architectures to achieve state-of-the-art results on standard benchmarks. Despite their success, these models are often resource-intensive, making them less suitable for real-time or resource-constrained applications.

2.2.3 Green Machine Learning Approaches

Green machine learning represents a shift towards developing models that are not only accurate but also computationally efficient and environmentally friendly. The GreenSaliency method, for instance, is designed to perform image saliency detection with a minimal carbon footprint and low computational complexity, making it ideal for deployment on mobile devices and edge computing platforms [7]. This method combines multi-layer hybrid feature extraction with a

multi-path saliency prediction framework, achieving competitive performance with significantly lower resource requirements compared to traditional DL-based methods.

2.3 Identification of Gaps and Challenges

Despite the advancements in SOD, existing methods face several challenges that limit their applicability. Conventional methods, while computationally less demanding, struggle with capturing complex image semantics and require extensive hand-tuning. On the other hand, DL-based approaches, though powerful, are often hampered by their high computational costs and dependency on large datasets for training. These limitations highlight the need for more efficient methods that can maintain high accuracy without the associated computational burden. Additionally, many existing models do not adequately address the interpretability of the generated saliency maps, which is crucial for applications in fields like medical imaging and autonomous driving.

2.4 Comparison with Proposed Work

Our proposed method addresses these challenges by leveraging color separation to enhance feature extraction in a computationally efficient manner. Unlike traditional DL-based methods that require extensive computational resources, our model processes each RGB channel independently, allowing for efficient exploitation of color-specific saliency cues. This approach significantly reduces the computational demands while maintaining high performance, particularly in scenarios where distinct colors play a crucial role in determining saliency. Moreover, by focusing on a lightweight architecture, our method offers a practical solution for deployment in resource-constrained environments, such as mobile devices, without sacrificing accuracy.

3. Methodology / Proposed Method

In this section, we detail the proposed saliency detection model, which is grounded in the influence of color on human visual perception. Recognizing that distinct colors significantly draw human attention, our approach processes each RGB channel of an input image independently to effectively leverage color-specific saliency cues.

3.1. Overview

The model architecture, depicted in Figure 1, employs a multi-level encoder-decoder framework, specifically designed to preserve and refine color-specific features throughout the network. The input image is first decomposed into its three RGB channels, which are processed separately to maintain the unique characteristics of each color channel. This decomposition allows the model to fully exploit the distinct color properties that contribute to saliency.

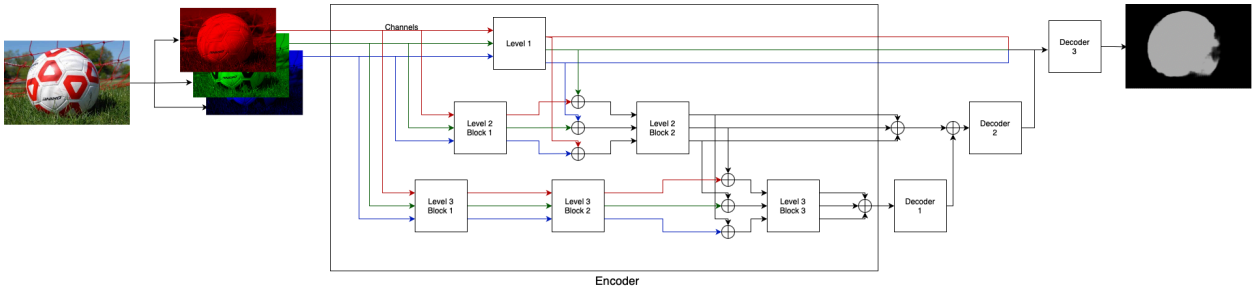


Figure 1. Depiction of our proposed architecture

3.2. Multi-Level Encoder Architecture

The encoder consists of three levels, with each level comprising multiple convolution blocks. Each convolution block includes a convolutional layer, batch normalization, and an activation function, ensuring effective feature extraction and stability during training. The levels within the encoder are interconnected via skip connections, facilitating the preservation of detailed features across different scales.

At each level, the network processes the RGB channels independently, maintaining the integrity of color-specific information. The outputs of these blocks are then fused and passed on to the subsequent level, where the process is repeated with additional refinement.

3.3. Decoder and Feature Fusion

The decoder network mirrors the hierarchical structure of the encoder, progressively refining the features extracted at each level. The refined features from the encoder are concatenated with those in the decoder through skip connections, allowing for precise upsampling and enhanced saliency map generation. The final saliency map is produced with minimal computational overhead by integrating the output of all three decoders, each responsible for different levels of detail refinement.

4. Experiments

4.1 Setup

We implemented our architecture using the PyTorch deep learning framework and trained it using an NVIDIA A100 GPU. We used an AdamW optimizer in training starting from a learning rate set to $1e^{-3}$. All images are resized to a standard 256 x 256. The model is first pre-trained on the DUTS-TE [8] dataset. Then only the decoder is trained on the remaining datasets. The learning rate is adjusted after every epoch using a Cosine Annealing [9] scheduler.

4.2 Datasets

As mentioned before, we pre-trained our model on the DUTS-TE [8] dataset before carrying our experiments on three RGB-D datasets: DUT-RGBD [10], NLPR [11], and NJUD2K [12]. Only

the RGB images were used without the depth. The three datasets were chosen to ease the comparison of our model to other models.

4.3 Evaluation Metrics

We evaluate our model on four widely used metrics: Mean Absolute Error (MAE), F-measure (F_β), S-measure (S_γ), and E-measure (E_α). As a note, the lower the MAE the better while the remaining metrics are better when higher.

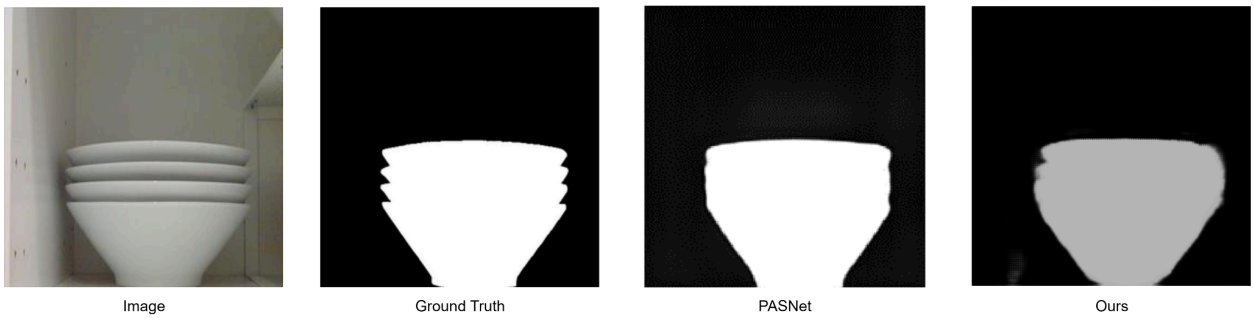


Figure 2. A prediction example using our model against PASNet

4.4 Comparisons

We compare our model with eleven models on the three aforementioned RGB-D datasets. All of the models used in the comparison are models that only use the RGB images and ignore the depth same as our model. The results shown in Table 1 are taken from [13]. Our model performance ranges from 70 to 90 percent of PASNet [13] performance; the top model on the table. It should be noted that our model is less than 0.5 percent the size of PASNet. PASNet has 240 million parameters, while our model only has 600 thousand parameters.

Methods	DUT-RGBD				NJUD2K				NLPR			
	$E_\alpha \uparrow$	$F_\beta \uparrow$	$S_\gamma \uparrow$	MAE↓	$E_\alpha \uparrow$	$F_\beta \uparrow$	$S_\gamma \uparrow$	MAE↓	$E_\alpha \uparrow$	$F_\beta \uparrow$	$S_\gamma \uparrow$	MAE↓
DSS [14]	.	.732	.803	.127	.	.776	.769	.108	.	.755	.838	.076
Amulet [15]	.	.803	.813	.083	.	.798	.827	.085	.	.722	.838	.062
R^3Net [16]	.	.781	.819	.113	.	.775	.770	.092	.	.649	.846	.101
PiCANet [17]	.	.826	.878	.080	.	.806	.872	.071	.	.761	.871	.053
PAGRN [18]	.	.836	.	.079	.	.827	.	.081	.	.795	.	.051
PoolNet [19]	.	.871	.	.049	.	.850	.	.057	.	.791	.	.046
AFNet [20]	.	.851	.	.064	.	.857	.	.056	.	.807	.	.043
CPD [21]	.911	.865	.875	.055	.905	.905	.863	.060	.925	.866	.893	.034
EGNet [22]	.	.866	.872	.059	.	.846	.840	.060	.	.800	.880	.047
MSI-Net [23]	.900	.861	.875	.058	.906	.859	.870	.057	.914	.854	.886	.041
PASNet [13]	.966	.940	.903	.033	.946	.907	.891	.040	.964	.921	.912	.024
Ours	**.799**	**.671**	**.688**	**.144**	**.778**	**.643**	**.693**	**.170**	**.873**	**.723**	**.786**	**.060**

Table 1. Quantitative results taken from [13] and compared to our model. The symbols F_β, S_γ, and E_α refer to the F-measure , S-measure, and E-measure. The red color represents the best result and the blue represents the second best.

5. Conclusion

In this project, we have presented a novel and efficient method for salient object detection that capitalizes on color separation to enhance feature extraction while maintaining low computational complexity. Our approach addresses the limitations of existing deep learning-based SOD models, particularly their high computational demands, by employing a lightweight architecture that processes RGB channels independently. The results from our experiments demonstrate that despite its simplicity, our model performs competitively against state-of-the-art methods, particularly in resource-constrained environments such as mobile devices.

The model's ability to achieve high accuracy with a significantly smaller number of parameters compared to leading models highlights its potential for practical applications where computational resources are limited. Furthermore, the architecture's modular design allows for easy adaptation and optimization, paving the way for future enhancements.

Moving forward, potential avenues for further research include extending this model to incorporate depth information for RGB-D saliency detection and exploring its applicability in real-time video processing scenarios. Additionally, investigating the integration of more advanced attention mechanisms could further improve the model's performance without compromising its efficiency. The proposed method represents a step towards more accessible and efficient saliency detection solutions, contributing to the broader field of computer vision.

6. References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," Advances in Neural Information Processing Systems, vol. 19, 2006.

[3] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze I: Boosting saliency prediction with feature maps trained on ImageNet," arXiv preprint arXiv:1411.1045, 2014.

[4] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2798–2805, 2014.

[5] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," Proceedings of the European Conference on Computer Vision (ECCV), pp. 419–435, 2020.

[6] A. Linardos, M. Kümmerer, O. Press, and M. Bethge, "DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12919–12928, 2021.

[7] Z. Mei, Y.-C. Wang, and C.-C. J. Kuo, "GreenSaliency: A lightweight and efficient image saliency detection method," arXiv preprint arXiv:2404.00253, 2024.

[8] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 136-145, 2017.

[9] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[10] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7254-7263, 2019.

[11] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, Springer International Publishing, pp. 92-109, 2014.

[12] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 1115-1119, 2014.

[13] T. Hussain, A. Anwar, S. Anwar, L. Petersson, and S. W. Baik, "Pyramidal attention for saliency detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2877-2887, 2022.

[14] Q. Hou, M. M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3203-3212, 2017.

[15] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 202-211, 2017.

[16] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P. A. Heng, "R3Net: Recurrent residual refinement network for saliency detection," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, vol. 684-690, Menlo Park, CA, USA: AAAI Press, July 2018.

[17] N. Liu, J. Han, and M. H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3089-3098, 2018.

[18] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3085-3094, 2019.

[19] J. J. Liu, Q. Hou, M. M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3917-3926, 2019.

[20] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1623-1632, 2019.

[21] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3907-3916, 2019.

[22] J. X. Zhao, J. J. Liu, D. P. Fan, Y. Cao, J. Yang, and M. M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8779-8788, 2019.

[23] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9413-9422, 2020.