Learning Saliency Map From Transformer and Depth

Chenmao Li,Wei Ming,Qiaozhong Huang,Jiamao Li,Dongchen Zhu,Lei Wang
Shanghai Institute of Microsystem and Information Technology, CAS
Shanghai, China

Abstract

Salient object detection(SOD) is widely used in the field of computer vision. By automatically detecting the most prominent targets in an image or video, SOD can help people understand the image or video content faster and more accurately. However, due to the unclear background separation and the lack of fine edge details, there are still many challenges in the practical application of salient object detection technology. We propose Depth-Reformer for salient object detection, which utilizes the Depth Transformer Encoder (DTE) and RGB-Depth Multi-Scale Fuser (RD-MSF) for depth features extraction and multi-level feature fusion to obtain targeted map. Depth-Reformer ranked the 4th place on the track of VizWiz 2023 Salient Object Detection challenge.

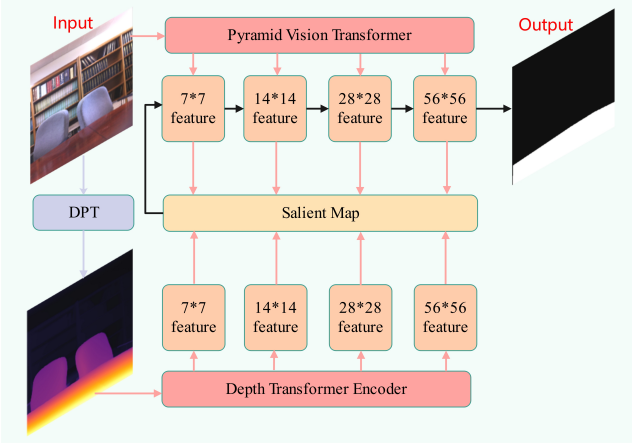

Figure 1. Overview of Depth-Reformer pipeline

1. Introduction

Salient Object Detection(SOD) is the task of generating a binary mask for an image, which can decipher which pixels belong to the foreground target of interest rather than the background. Although significant progress has been made in the field of significance detection, there are still some challenges and problems.

For example, the significance detection task is essentially a subjective issue, and different people may have different understandings and annotations of the significance regions of the same image. Therefore, how to address diversity issues remains a challenging issue. Besides, the salient area may be influenced by multiple factors, such as the image background, the shape and size of foreground objects, lighting conditions, etc. How to accurately detect significance in these situations remains a challenging issue.

In order to effectively separate the front and back backgrounds in salient detection for different group of people, we propose Depth-Reformer model, which integrating depth information on the basis of the Self-Reformer framework [8]. It utilizes the Depth Transformer Encoder (DTE) and RGB-Depth Multi-Scale Fuser (RD-MSF) for depth features extraction and multi-level feature fusion to obtain targeted map, which helps the situation that blind people tend to pay more attention to objects that are closer and more blurry. We will provide a detailed introduction in Section 2 and describe the entire experimental process in Section 3.

2. Method

As shown in Fig. 1,we propose a depth encoder to better incorporate depth features into saliency map detection based on Self-Reformer. For simplicity, we derived the depth features from DPT-Beit-Large-512, which is a Monocular Depth Estimation model [5] and pretrained on 12 datasets with multi-objective optimization. According to the experimental results, the model has accurate and robust monocular depth estimation results in indoor environmental images. Then, we employ a additional Depth Transformer Encoder(DTE), to extract depth features with the same structure as the rgb part structure. Finally, depth features and rgb features are fused in the RGB-Depth Fuser(RDF) to supervise the generation of saliency maps.

Depth Transformer Encoder(DTE). We output features of 7, 14, 28, and 56 pixel sizes using PVT [7] backbone to better extract multi-scale depth information.

Variant	IOU(%)
SR	90.44
SR+Depth	91.02
SR+Depth+Freeze+Crop	91.88

Table 1. Ablation study on *test* set of VizWiz

Afterwards, we inserted a LayerNorm-Linear-GELU module after the multi-scale features. Layer Normalization [1] in transformer [6] can effectively prevent overfitting compared with Batch Normalization, while GELU [4] activation function in transformer structure can better avoid gradient disappearance than ReLU [3].

RGB-Depth Multi-Scale Fuser(RD-MSF). RD-MSF is a module that fuses rgb features with depth features of different scales from bottom to top. For each scale, it includes a BatchNorm-Conv-LeakyReLU module. After the rgb-depth features are concatenated, they are transmitted to the upper level and fused with the larger rgb-depth features, and so on. The final features are input into the ViT [2] to obtain a saliency map of color depth fusion.

3. Experiment

Dataset. The dataset of VizWiz Salient Object Detection contains 19116,6105 and 6779 samples in the *train*,*val*,*test* set, respectively. Evaluation are conducted by the challenge server on the test.

Implementation Details. In our experiments, PVT backbone is pretrained on the DUTS-TR dataset. In the training phase, we freeze the front layers of the backbone to prevent overfitting. At the same time, we conducted random cropping to enhance the diversity of the data, but only retained masks that still contain saliency targets after cropping, otherwise it would change the judgment for saliency objects.

Results. As shown in Tab 1, compared to the Self-Reformer model, ours successfully achieved a 1.44% IoU improvement, thanks to Depth Fusion Encoder and data enhancement strategy to prevent overfitting. Tab 2 shows public results on VizWiz test set, where ours ranked the 4th place.

4. Conclusion

We propose Depth-Reformer for salient object detection, which utilizes the DTE and RD-MSF for depth features extraction and multi-level feature fusion to obtain targeted map. Depth-Reformer ranked the 4th place on the track of VizWiz 2023 Salient Object Detection challenge.

Rank	Team	IOU(%)
1	minivision	94
2	ll-ly	93
3	DeepBlue-AI	93
4	SIMIT-LAB9	92
5	IIAU-csj	90
6	SegLab	90
7	HQ	84
8	Try1Try	75
9	TimZ	68
10	Zero2One	60

Table 2. Results on *test* set of VizWiz

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[3] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. pages 315–323, 2011. 2

[4] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 2

[5] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 1

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[7] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. pages 568–578, 2021. 1

[8] Yi Ke Yun and Weisi Lin. Selfreformer: Self-refined network with transformer for salient object detection. *arXiv preprint arXiv:2205.11283*, 2022. 1