ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 17, No. 2, 2022, *pp*.102-117

LTDNet: A lightweight two-stage decoder network for RGB-D salient object detection

Jian Wang, Wenbing Chen[1]

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China

(Received August 17, 2022, accepted October 20, 2022)

Abstract: Most existing models of RGB-D salient object detection (SOD) utilize heavy backbones like VGGs and ResNets which lead to large model size and high computational costs. In order to improve this problem, a lightweight two-stage decoder network is proposed. Firstly, the network utilizes MobileNet-V2 and a customized backbone to extract the features of RGB images and depth maps respectively. In order to mine and combine cross-modality information, cross reference module is used to fuse complementary information from different modalities. Subsequently, we design a feature enhancement module to enhance the clues of the fused features which has four parallel convolutions with different expansion rates. Finally, a two-stage decoder is used to predict the saliency maps, which processes high-level features and low-level features separately and then merges them. Experiments on 5 benchmark datasets comparing with 10 state-of-the-art models demonstrate that our model can achieve significant improvement with smallest model size.

Keywords: salient object detection, RGB-D, lightweight, efficient

1. Introduction

Salient object detection (SOD) aims to locate and segment the most eye-catching objects in a scene by simulating the human visual attention mechanism. SOD has been developed rapidly due to its wide application in image processing and computer vision, such as visual tracking [1], image segmentation [2], face recognition [3], medical segmentation [4] and so on. In the past years, the development of deep learning has driven SOD to achieve promising performance. Most existing SOD methods focus RGB images. However, it is difficult to get outstanding result in complex senses, such as camouflaged objects, similar texture, complex backgrounds, transparent objects, low-contrast.

With the popularity of depth device, depth sensor has been widely introduced into different fields to capture depth maps, which can provide additional clues for RGB images, such as object edges, 3D distribution, spatial structure. Many recent works [5-9] have been proposed and demonstrated that it is effective to improve efficiency and performance using auxiliary depth maps to assist RGB images for SOD. Although RGB-D SOD has achieved extraordinary results [10-16] in recent years, most methods use cumbersome networks as backbones which bring large model size and high computational costs, such as Resnets, VGGs. This makes it difficult to apply these methods to the devices with poor computing power.

In this paper, we propose a lightweight two-stage decoder network (LTDNet) for RGB-D SOD, which possesses smaller size and lower computational costs. We employ MobileNet-V2 to extract the features of RGB, which reduces the computational cost and network size significantly. For depth stream backbone, we

[1] Corresponding author. *E-mail address*: 001101@nuist.edu.cn.

design a lightweight network, which has only 0.89MB for a 3×352×352 input, to extract feature instead of VGGs or MobileNets. In order to retain the salient information of two modalities, we utilize a module named cross reference module (CRM) [17] to fuse the most salient information of depth features and RGB features. Subsequently, we utilize a feature refine module (FRM) to enhance the fused features. We use parallel dilated convolutions with different expansion rates to extract the large-scale information of fused features. Finally, considering the details of high-level features and the global semantic information of low-level features, a lightweight two-stage decoder is used to predict the salient object maps.

The main contributions of this paper can be summarized as follows:

- We propose an efficient two-stage decoder to combine different levels features. The decoder can fuse the detailed information of high-level features and the global semantic information of low-level features in two steps instead of top-down strategy.
- We design a feature refine module to enhance feature with larger receptive fields and channel attention. Four parallel dilated convolutions with different expansion rates can effectively extract the large-scale context information of features.
- We design a lightweight but efficient depth stream backbone instead of using the same backbone for RGB and depth. The customized backbone has fewer parameters but fits the model better.
- Compared with 10 state-of-the-art RGB-D SOD models on 5 datasets, our LTDNet shows outstanding performance both in terms of FPS and accuracy of evaluation indicators.

2. Related Work

2.1. Traditional RGB-D Salient Object Detection

The additional depth information is beneficial to more efficient and accurate localization and segmentation of salient objects in RGB images. Early methods utilize hand-crafted features for RGB-D SOD, such as boundary, contrast, shape attributes, 3D layout priors, anisotropic center-surround difference prior and so on. In [18], Peng et al. proposed a multi-contextual contrast model and built the first large-scale RGB-D dataset named NLPR for RGB-D SOD. In [19] Feng et al proposed local background enclosure features to solve the false positives due to areas of high contrast in background regions. In [20] Ren et al. obtained a saliency map by combining background, depth, region contrast, and orientation priors. In [21], Cong et al. proposed a depth-guided transformation model consisting of multilevel RGBD saliency initialization, depth-guided saliency refinement, and saliency optimization with depth constraints. However, traditional methods rely on hand-crafted features that lack high-level semantic representations and robustness in complex scenes.

2.2. Deep Learning-Based RGB-D Salient Object Detection

With the rapid development of deep learning, various methods based on convolutional neural networks (CNNs) have emerged. DF [22] is the first method to introduce deep learning into RGB-D SOD. Qu et al. fed hand-crafted features into a special-designed CNN model to fuse low-level salient features into hierarchical features and automatically detect salient objects in RGB-D images. In [23], Shigematsu et al. adopted two independent convolutional networks to process RGB images and hand-crafted deep features and fused the features to achieve salient maps. In CTMF [24], Han et al. utilized CNNs to transfer the structure of RGB images to be applicable for depth maps and fuses the high-level representations automatically to obtain saliency maps. In PCF [25], Chen et al. proposed a complementarity-aware fusion module to integrate complementary information from both modalities. In JL-DCF [10], taking depth map as a special case of RGB map, Fu et al. employed a shared CNN for RGB and depth feature extraction and presented joint learning and densely cooperative fusion to fuse multi-scale features effectively. In BBS-Net [13], Deng proposed a bifurcated backbone strategy to divide the multi-level features into teacher features and student features, and utilized a