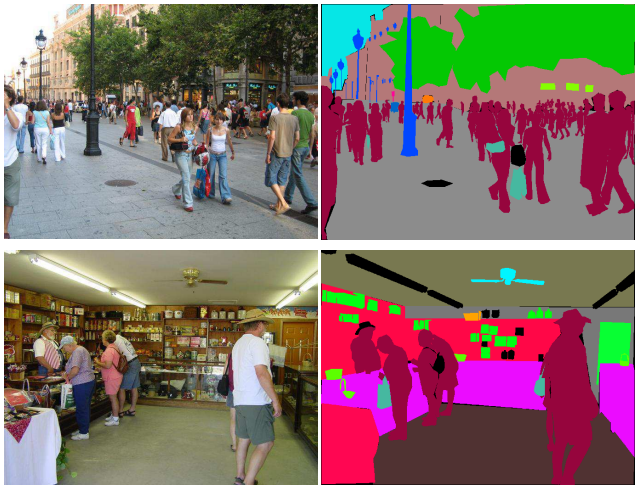
Pyramid Scene Parsing Network

Hengshuang Zhao[1] Jianping Shi[2] Xiaojuan Qi[1] Xiaogang Wang[1] Jiaya Jia[1]
[1]The Chinese University of Hong Kong [2]SenseTime Group Limited
{hszhao, xjqi, leojia}@cse.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk, shijianping@sensetime.com

Abstract

Scene parsing is challenging for unrestricted open vocabulary and diverse scenes. In this paper, we exploit the capability of global context information by different-region-based context aggregation through our pyramid pooling module together with the proposed pyramid scene parsing network (PSPNet). Our global prior representation is effective to produce good quality results on the scene parsing task, while PSPNet provides a superior framework for pixel-level prediction. The proposed approach achieves state-of-the-art performance on various datasets. It came first in ImageNet scene parsing challenge 2016, PASCAL VOC 2012 benchmark and Cityscapes benchmark. A single PSPNet yields the new record of mIoU accuracy 85.4% on PASCAL VOC 2012 and accuracy 80.2% on Cityscapes.



(a) Image (b) Ground Truth

Figure 1. Illustration of complex scenes in ADE20K dataset.

1. Introduction

Scene parsing, based on semantic segmentation, is a fundamental topic in computer vision. The goal is to assign each pixel in the image a category label. Scene parsing provides complete understanding of the scene. It predicts the label, location, as well as shape for each element. This topic is of broad interest for potential applications of automatic driving, robot sensing, to name a few.

Difficulty of scene parsing is closely related to scene and label variety. The pioneer scene parsing task [23] is to classify 33 scenes for 2,688 images on LMO dataset [22]. More recent PASCAL VOC semantic segmentation and PASCAL context datasets [8, 29] include more labels with similar context, such as chair and sofa, horse and cow, etc. The new ADE20K dataset [43] is the most challenging one with a large and unrestricted open vocabulary and more scene classes. A few representative images are shown in Fig. 1. To develop an effective algorithm for these datasets needs to conquer a few difficulties.

State-of-the-art scene parsing frameworks are mostly based on the *fully convolutional network* (FCN) [26]. The deep *convolutional neural network* (CNN) based methods boost dynamic object understanding, and yet still face challenges considering diverse scenes and unrestricted vocabulary. One example is shown in the first row of Fig. 2, where a *boat* is mistaken as a *car*. These errors are due to similar appearance of objects. But when viewing the image regarding the context prior that the scene is described as *boathouse* near a river, correct prediction should be yielded.

Towards accurate scene perception, the knowledge graph relies on prior information of scene context. We found that the major issue for current FCN based models is lack of suitable strategy to utilize global scene category clues. For typical complex scene understanding, previously to get a global image-level feature, spatial pyramid pooling [18] was widely employed where spatial statistics provide a good descriptor for overall scene interpretation. Spatial pyramid pooling network [12] further enhances the ability.

Different from these methods, to incorporate suitable global features, we propose *pyramid scene parsing network* (PSPNet). In addition to traditional dilated FCN [3, 40] for pixel prediction, we extend the pixel-level feature to the specially designed global pyramid pooling one. The local and global clues together make the final prediction more reliable. We also propose an optimization strategy with