

Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection

Hongmei Song¹, Wenguan Wang¹ [0000-0002-0802-9567],
Sanyuan Zhao¹, Jianbing Shen^{1,2}, and Kin-Man Lam³

¹ Beijing Lab of Intelligent Information Technology, School of Computer Science,
Beijing Institute of Technology, China

² Inception Institute of Artificial Intelligence, Abu Dhabi, UAE

³ The Hong Kong Polytechnic University, Kowloon, Hong Kong

{songhongmei, zhaosanyuan, shenjianbing}@bit.edu.cn

wenguanwang.ai@gmail.com enkmlam@polyu.edu.hk

<https://github.com/shenjianbing/PDB-ConvLSTM>

Abstract. This paper proposes a fast video salient object detection model, based on a novel recurrent network architecture, named Pyramid Dilated Bidirectional ConvLSTM (PDB-ConvLSTM). A Pyramid Dilated Convolution (PDC) module is first designed for simultaneously extracting spatial features at multiple scales. These spatial features are then concatenated and fed into an extended Deeper Bidirectional ConvLSTM (DB-ConvLSTM) to learn spatiotemporal information. Forward and backward ConvLSTM units are placed in two layers and connected in a cascaded way, encouraging information flow between the bi-directional streams and leading to deeper feature extraction. We further augment DB-ConvLSTM with a PDC-like structure, by adopting several dilated DB-ConvLSTMs to extract multi-scale spatiotemporal information. Extensive experimental results show that our method outperforms previous video saliency models in a large margin, with a real-time speed of **20 fps** on a single GPU. With unsupervised video object segmentation as an example application, the proposed model (with a CRF-based post-process) achieves state-of-the-art results on two popular benchmarks, well demonstrating its superior performance and high applicability.

1 Introduction

Video saliency detection aims at finding the most interesting parts in each video frame that mostly attract human attention. It can be applied as a fundamental module in many visual tasks, such as video object segmentation, scene rendering, object tracking, and so on. Similar to visual saliency detection in static images, research on video saliency detection can also be divided into two categories, i.e., eye fixation prediction [41, 39] and salient object detection [49, 47]. The purpose of eye fixation prediction is to locate the focus of human eyes when looking at

Hongmei Song and Wenguan Wang contributed equally.

Corresponding author: Sanyuan Zhao.

