

# ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation

Sachin Mehta<sup>1</sup>[0000-0002-5420-4725], Mohammad Rastegari<sup>2</sup>, Anat Caspi<sup>1</sup>,  
Linda Shapiro<sup>1</sup>, and Hannaneh Hajishirzi<sup>1</sup>

<sup>1</sup> University of Washington, Seattle, WA, USA

fsacmehta, caspi an, shapi ro, hannanehG@cs.washington.edu

<sup>2</sup> Allen Institute for AI and XNOR.AI, Seattle, WA, USA

mohammadr@allenai.org

**Abstract.** We introduce a fast and efficient convolutional neural network, ESPNet, for semantic segmentation of high resolution images under resource constraints. ESPNet is based on a new convolutional module, efficient spatial pyramid (ESP), which is efficient in terms of computation, memory, and power. ESPNet is 22 times faster (on a standard GPU) and 180 times smaller than the state-of-the-art semantic segmentation network PSPNet, while its category-wise accuracy is only 8% less. We evaluated ESPNet on a variety of semantic segmentation datasets including Cityscapes, PASCAL VOC, and a breast biopsy whole slide image dataset. Under the same constraints on memory and computation, ESPNet outperforms all the current efficient CNN networks such as MobileNet, ShuffleNet, and ENet on both standard metrics and our newly introduced performance metrics that measure efficiency on edge devices. Our network can process high resolution images at a rate of 112 and 9 frames per second on a standard GPU and edge device, respectively. Our code is open-source and available at <https://sacmehta.github.io/ESPNet/>.

## 1 Introduction

Deep convolutional neural network (CNN) models have achieved high accuracy in visual scene understanding tasks [1–3]. While the accuracy of these networks has improved with their increase in depth and width, large networks are slow and power hungry. This is especially problematic on the computationally heavy task of semantic segmentation [4–10]. For example, PSPNet [1] has 65.7 million parameters and runs at about 1 FPS while discharging the battery of a standard laptop at a rate of 77 Watts. Many advanced real-world applications, such as self-driving cars, robots, and augmented reality, are sensitive and demand on-line processing of data locally on edge devices. These accurate networks require enormous resources and are not suitable for edge devices, which have limited energy overhead, restrictive memory constraints, and reduced computational capabilities.

Convolution factorization has demonstrated its success in reducing the computational complexity of deep CNNs [11–15]. We introduce an efficient convolutional module, ESP (efficient spatial pyramid), which is based on the convolutional factorization



































