ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation

Sachin Mehta[1], Mohammad Rastegari[2], Anat Caspi[1], Linda Shapiro[1], and Hannaneh Hajishirzi[1]

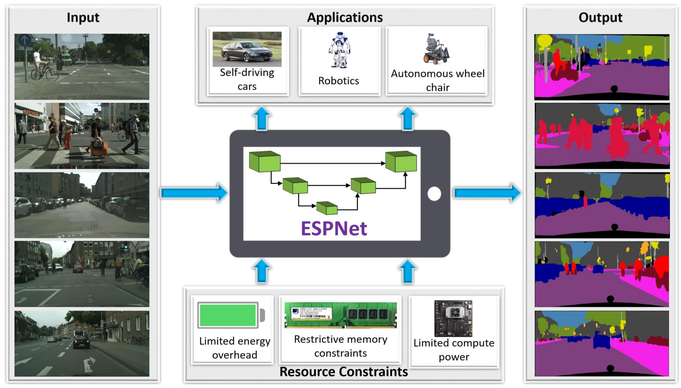
[1]**University of Washington, Seattle, WA** [2]**Allen Institute of Artificial Intelligence and XNOR.AI**

Email: {***sacmehta,*** caspian, shapiro, hannaneh}@cs.washington.edu mohammadr@allenai.org

Project Webpage

Introduction

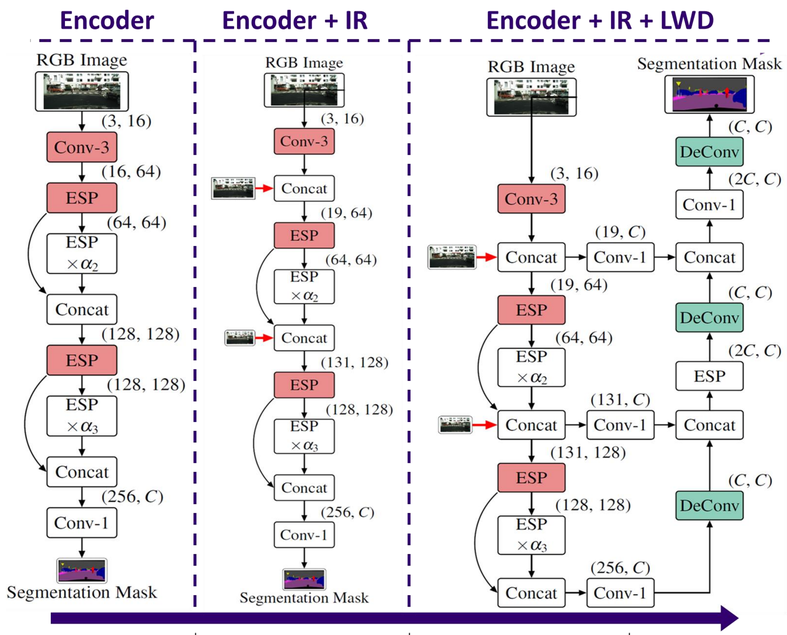
- Real-world applications demand **online processing** of data locally on ***resource constrained edge devices.***



- We introduce **ESPNet**, a deep learning based segmentation network for edge devices which is ***fast, light-weight, memory efficient,*** and ***power efficient***.

ESPNet

- Encoder-decoder with ESP as the basic building block
- For efficiency and accuracy:
  - **Input-reinforcement (IR)**: concatenate image with feature maps at different spatial-levels
  - **Light-weight decoder (LWD):** bottom-up decoding with low-dimensional feature maps from the encoder



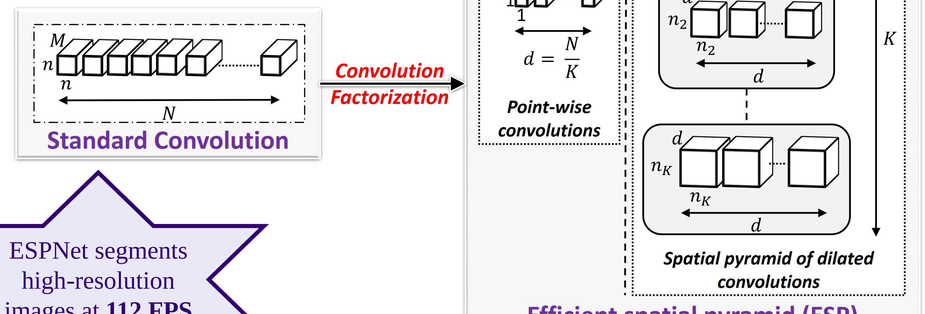
0.47	0.49	**0.56**	**mIOU**
0.186 M	0.187 M	0.202 M	**# params**

Efficient spatial pyramid (ESP) module.

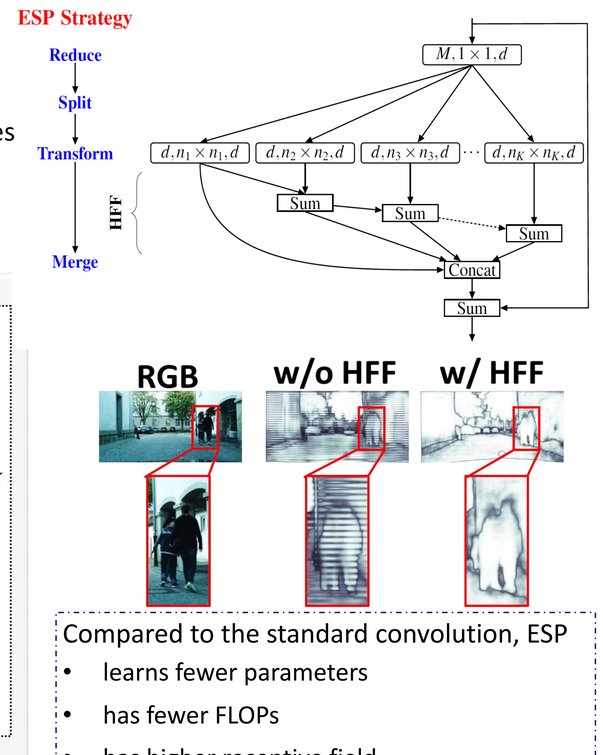
- **Reduce:** project high-dimensional feature maps to low-dimensional space
- **Split** and **Transform**: learn representations in parallel with different dilation rates
- **Merge**: concatenate ***hierarchically fused feature*** maps to produce output

Hierarchical feature fusion (HFF)

- removes the gridding artifacts caused by dilated convolutions
- does not increase the computational complexity of the ESP module



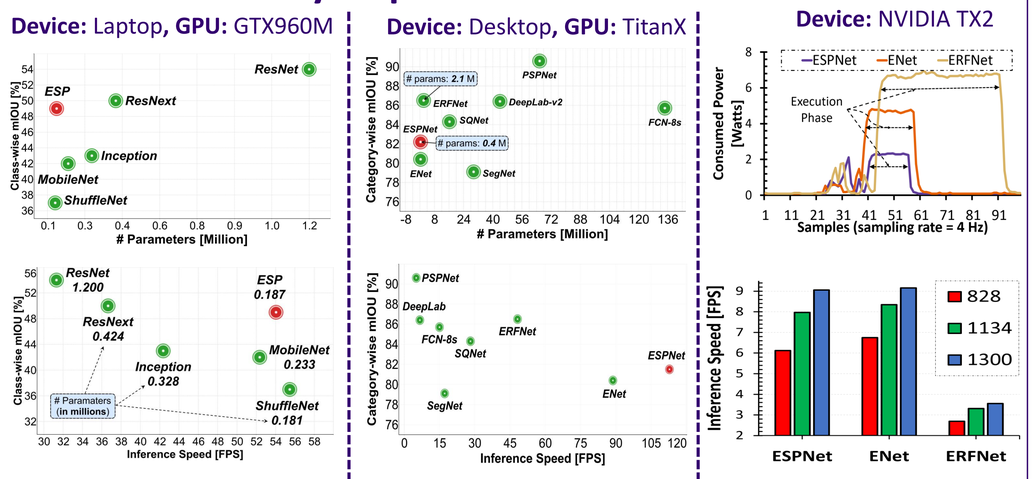
ESPNet segments high-resolution images at **112 FPS**



Compared to the standard convolution, ESP

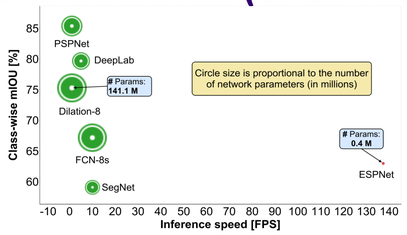
- learns fewer parameters
- has fewer FLOPs
- has higher receptive field

Results on the CityScapes dataset

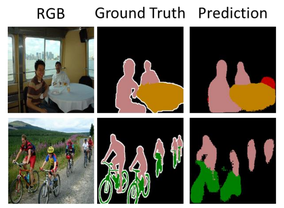


- ESP outperforms **MobileNet (7%) and ShuffleNet (12%)** while learning a similar number of parameters with comparable inference speed
- ESPNet is more accurate, faster, and more power efficient than ENet
- ESPNet is **22x faster** and **180x smaller** than PSPNet, while its 8% less accurate

PASCAL VOC 2012 (GPU: TitanX)



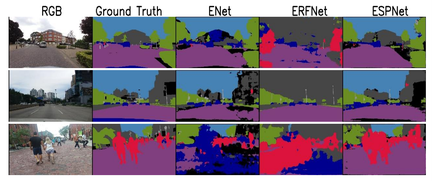
- ESPNet delivers competitive performance **without pre-training on the ImageNet**.
- ESPNet outperforms **SegNet (4%)** with **81x** fewer parameters.



Unseen dataset

ESPNet has more generalization power than ENet and ERFNet on an unseen dataset, **Mapillary**.

	mIOU
ENet	0.33
ERFNet	0.25
ESPNet	**0.40**



Breast biopsy WSI dataset

ESPNet achieves the same accuracy as the SOTA method while learning 9.5x fewer parameters.