Fast-SCNN: Fast Semantic Segmentation Network

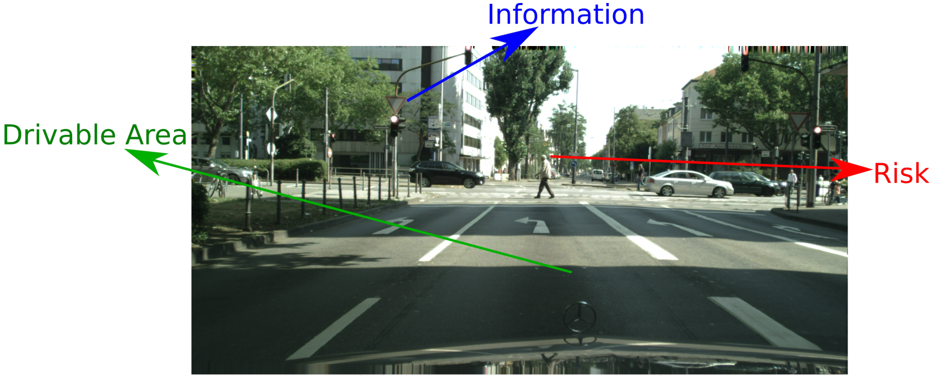
Rudra PK Poudel
Stephan Liwicki
Roberto Cipolla

Cambridge Research Laboratory
Toshiba Research Europe, UK

BMVC 2019

Real-time Semantic Image Segmentation

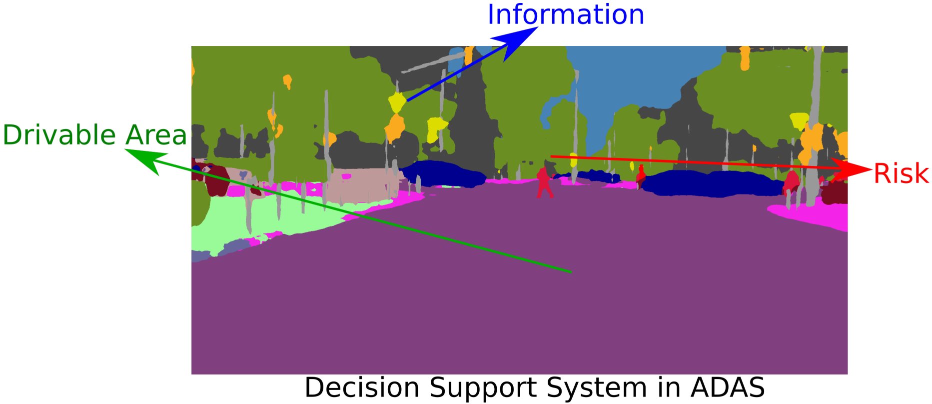
- ***What* am I seeing and *where* is it?**
- Real-time perception is critical for autonomous systems



Decision Support System in ADAS

Real-time Semantic Image Segmentation

- ***What* am I seeing and *where* is it?**
- Real-time perception is critical for autonomous systems

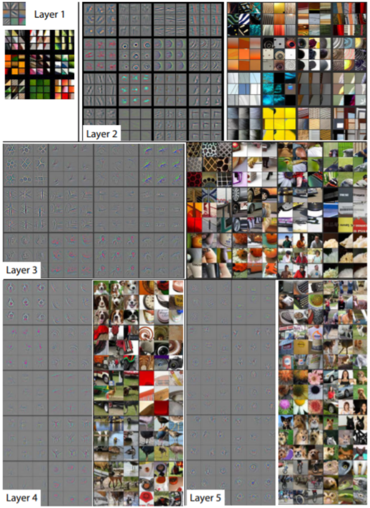


Motivation

- **Problem:** SOTA models are accurate but resource hungry
  - Compute: floating point ops
  - Power consumption
  - Memory

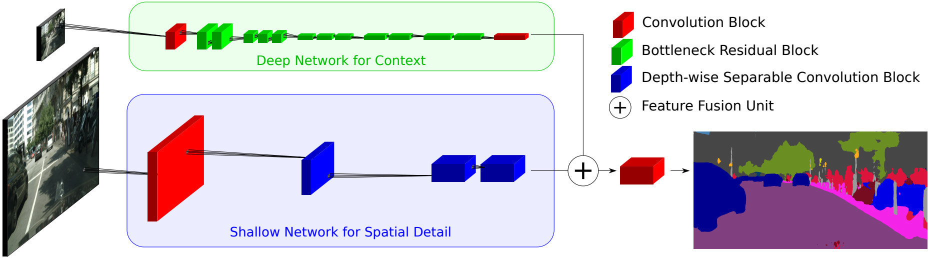
- **Observations:**
  1. First few layers of DCNN extract low-level features (Zeiler et al., 2014)
  2. Larger receptive field (context) is important for accuracy (Poudel et al., 2018)
  3. Spatial details is necessary to preserve boundary (Shelhamer et al. 2016)
  4. SOTA efficient models adapt multi-resolution and multi-branch architecture

Motivation: First Few Layers Learn Low-level Features



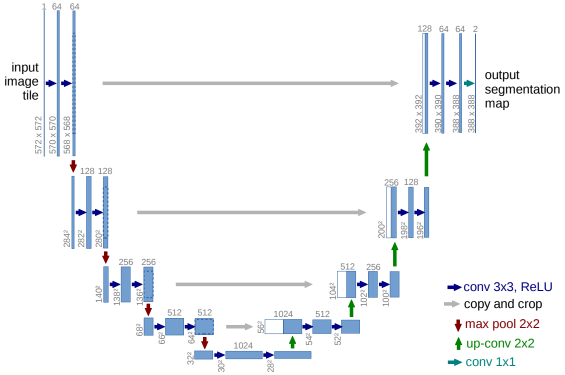
Zeiler et al., ECCV 2014

Motivation: Importance of Larger Receptive Field



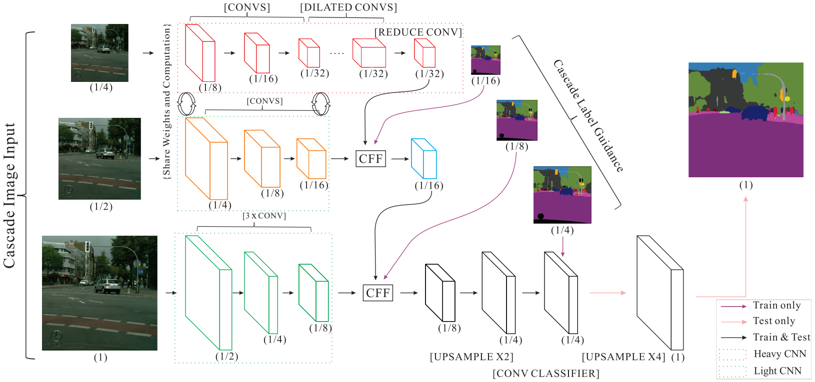
ContextNet (Poudel et al., BMVC 2018)

Motivation: Importance of Spatial Details



U-Net (Ronneberger et al., MICCAI 2015)

Motivation: Efficient Multi-resolution Architectures



ICNet (Zhao et el., ECCV 2018).

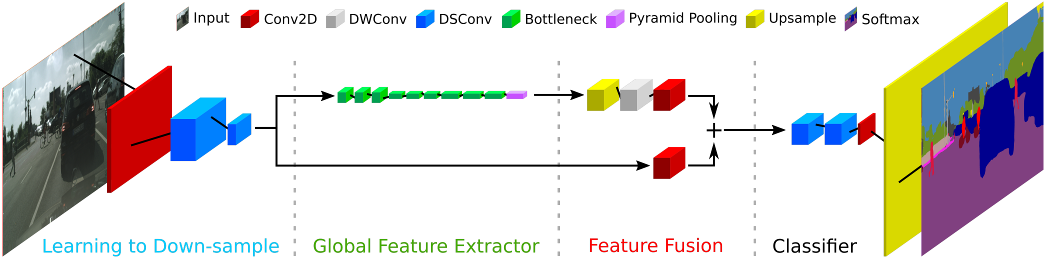
Motivation

- **Problem:** SOTA models are accurate but resource hungry
  - Compute: floating point ops
  - Power consumption
  - Memory

- **Observations:**
  1. First few layers of DCNN extract low-level features (Zeiler et al., 2014)
  2. Larger receptive field (context) is important for accuracy (Poudel et al., 2018)
  3. Spatial details is necessary to preserve boundary (Shelhamer et al. 2016)
  4. SOTA efficient models adapt multi-resolution and multi-branch architecture

Proposed Model: Overview

- **Hypothesis:** jointly learn the low level features of multi-branch networks to increase the model efficiency.

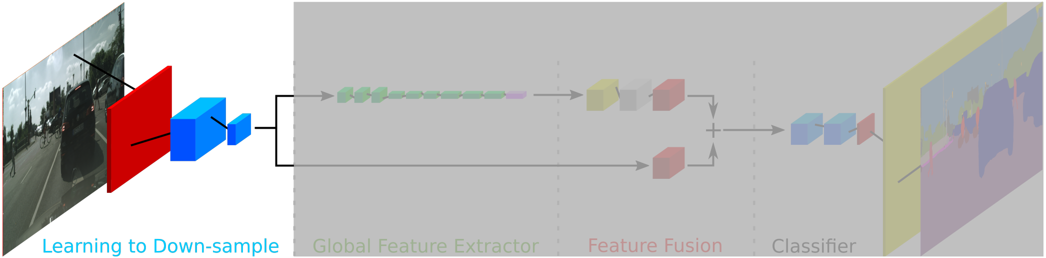


Fast-SCNN

- **Learning to Down-sample** jointly learns the low level features

Proposed Model: Learning to Down-sample

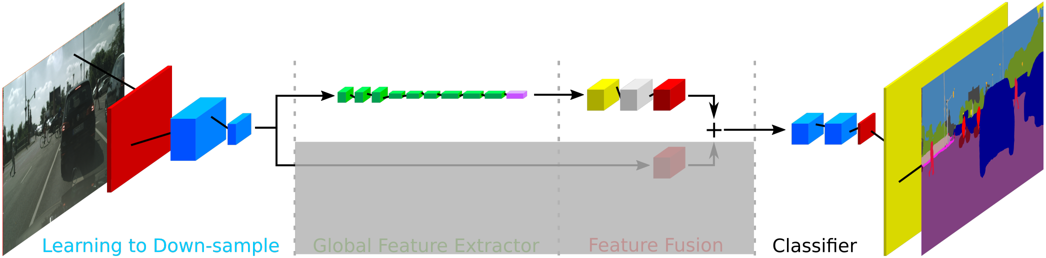
- **Learning to Down-sample** sharing computation of multi-resolution branches improves efficiency



- No need for multiple resizes and memory copies of the original input

Proposed Model: Larger Receptive Field

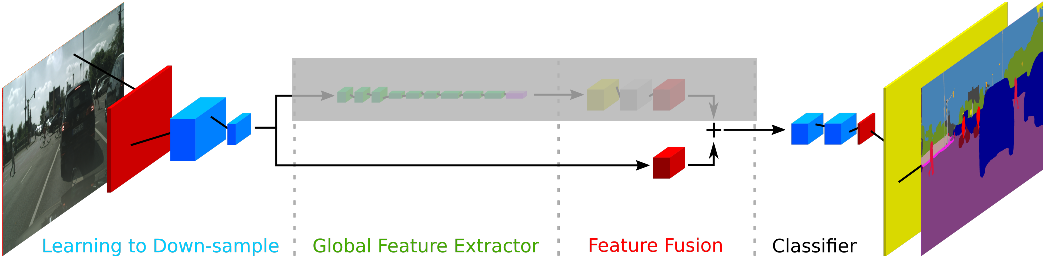
- **Going deeper with convnet** Fast-SCNN can be reduced to convnet



- Early sub-sampling/max-pooling layers increase receptive field and efficiency

Proposed Model: Skip-Connection

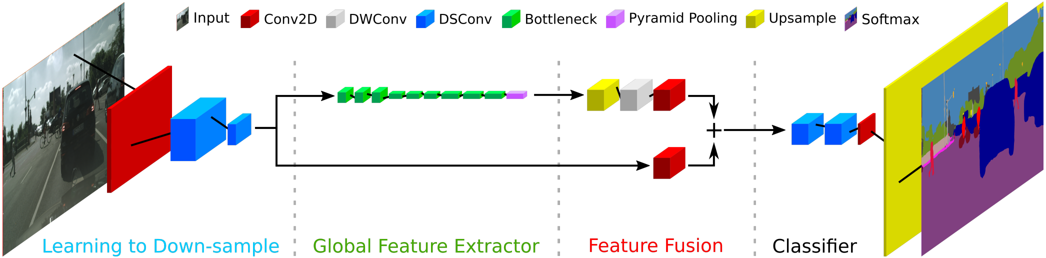
- **Spatial details** skip-connection helps to recover boundary information



- We preferred simple feature fusion module i.e. addition only

Proposed Model: Fast-SCNN

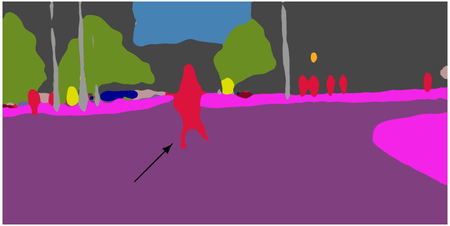
- **Deeper path** at low resolution captures global context information
- **Shallow path** focuses on high resolution segmentation details



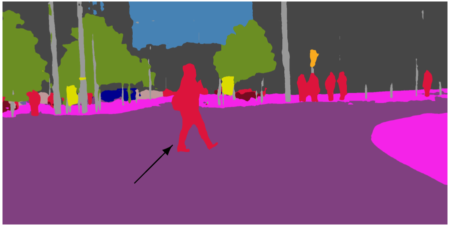
- No need to learn low-level features separately
- Quantization, network pruning and other techniques are also applicable

Proposed Model: Qualitative Validation

Input image



Skip-Connection: No



Skip-Connection: Yes

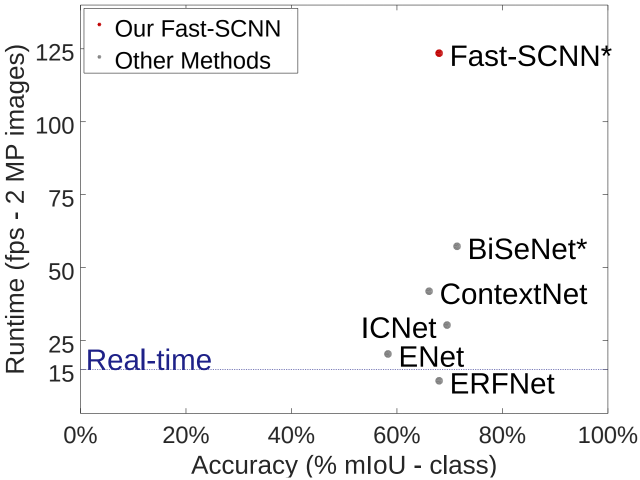
Proposed Model: Qualitative Validation

Input image

Skip-Connection: No

Skip-Connection: Yes

Fast-SCNN: Quantitative Evaluation



$*$ Nvidia Titan Xp (Pascal); *Others* Nvidia Titan X (Maxwell)

Fast-SCNN: Quantitative Evaluation

- Fast-SCNN balances accuracy and speed

	Class mIoU%	Category mIoU%	Params in Millions	FPS on 1024x2048
SegNet	56.1	79.8	29.46	1.6
ENet	58.3	80.4	0.37	20.4
ICNet	69.5	-	6.68	30.3
ERFNet	68.0	86.5	2.1	11.2
ContextNet	66.1	82.7	0.85	41.9
BiSeNet*	71.4	-	5.8	57.3
GUN*	70.4	-	-	33.3
Fast-SCNN*	68.0	84.7	1.11	**123.5**

$*$ Nvidia Titan Xp (Pascal); *Others* Nvidia Titan X (Maxwell)

Fast-SCNN: Input Size Variation

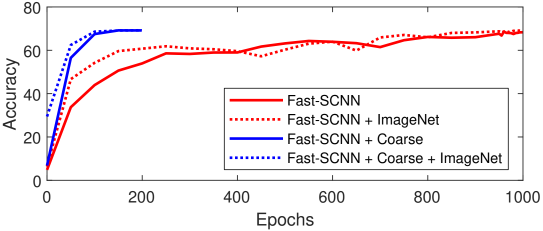
- Fast-SCNN is efficient on smaller as well as larger scale input sizes

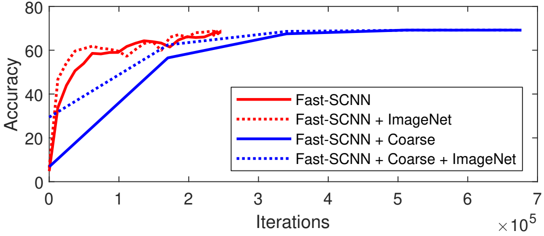
Input Size	Class mIoU%	Frame-Per-Second
1024×2048	68.0	123.5
512×1024	62.8	285.8
256×512	51.9	485.4

Is ImageNet Pre-Training is Necessary?

- Total number of gradient updates is important
- At least in validation and test sets ImageNet pre-training is not important!
- Similar finding on Rethinking ImageNet Pre-training by He et al. (ICCV 2019)

Model	Class mIoU%
Fast-SCNN	68.62
Fast-SCNN + ImageNet	69.15
Fast-SCNN + Coarse	69.22
Fast-SCNN + Coarse + ImageNet	69.19





Fast-SCNN: Qualitative Evaluation

Conclusion

- **Fast-SCNN** is
  - memory, computation and power efficient
  - twice as fast as other state-of-the-art models
  - **above real-time i.e. 123.5 fps on 1024$\times$2048 images**
  - efficient and competitive on smaller as well as larger scale input sizes
- We have shown accuracy without ImageNet pre-training is comparable
- Limitations: accuracy gap with bigger off-line models
- Future work: apply to depth estimation and instance segmentation

References

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B., The Cityscapes Dataset for Semantic Urban Scene Understanding. In CVPR, 2016.

He, K., Girshick, R., DollÃ¡r, P. Rethinking ImageNet Pre-training. In arXiv:1811.08883, 2018.

Poudel, R. P. K., Bonde, U., Liwicki, S., Zach, C., ContextNet: Exploring Context and Detai,l for Semantic Segmentation in Real-time. In BMVC, 2018.

Ronneberger, O. and Fischer, P. and Brox, T., U-Net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015.

Shelhamer, E. and Long, J. and Darrell, T., Fully convolutional networks for semantic segmentation. In PAMI, 2016.

Zeiler, M. D. and Fergus, R., Visualizing and understanding convolutional networks. In ECCV, 2014.

Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J., ICNet for Real-Time Semantic Segmentation on High-Resolution Images, In ECCV 2018.

Questions

Public implementations on PyTorch and TensorFlow are available on Github!

Thank you!