ContextNet: Exploring Context and Detail for Semantic Segmentation in Real-time

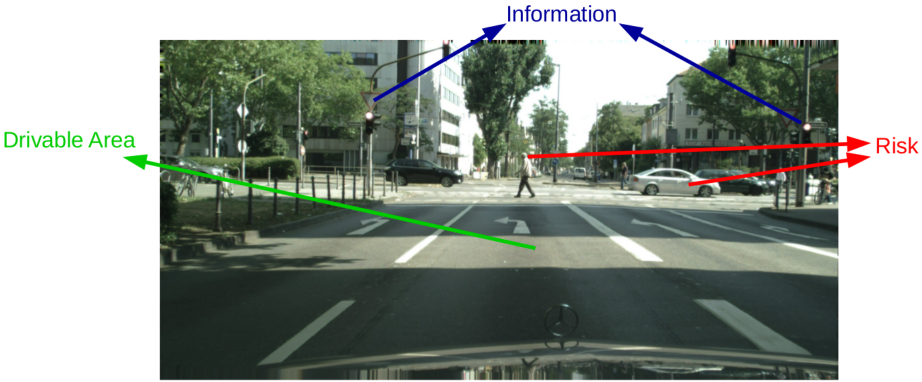
Rudra PK Poudel
Ujwal Bonde
Stephan Liwicki
Christopher Zach

Computer Vision Group
Toshiba Research Europe

TRE 2018

Real-time Semantic Image Segmentation

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



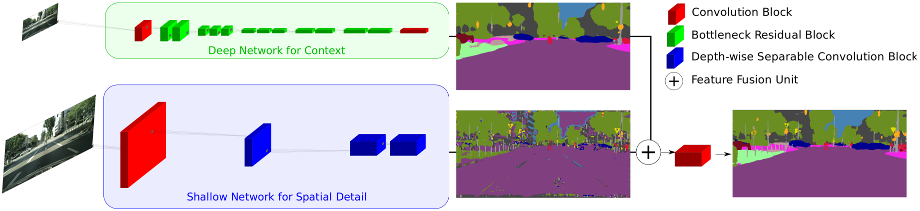
Decision Support System in ADAS

Motivation

- **Problem:** SOTA models are accurate but not real-time
- **Observations:**
  - Deeper models improve accuracy (He et al., 2015)
  - Multi-scale information fusion is beneficial (Burt et al. 1987)
  - Downside: increased cost
    - Floating point ops
    - Memory usage
    - Power consumption
- **Hypothesis:** efficient semantic segmentation based on
  - what (global context), and
  - where (spatial detail)
- **Aim:** real-time system for low resource (embedded) devices

Proposed Model: Overview

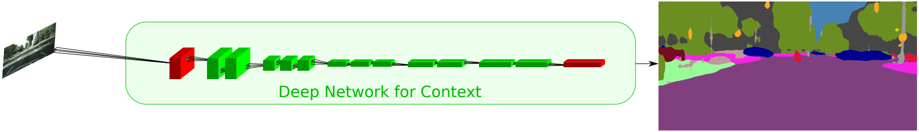
- **Context branch** at low resolution captures global context information
- **Detail branch** focuses on high resolution segmentation details



ContextNet

Proposed Model: Context Branch

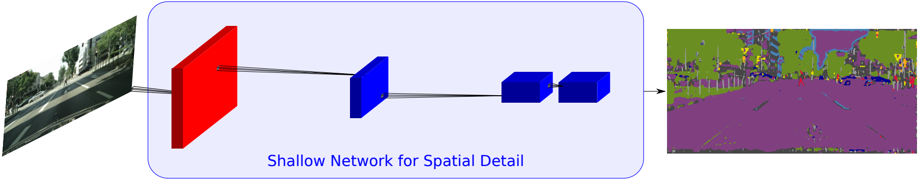
- **Context branch** at low resolution captures global context information



- No need for high resolution images to know what is there
- Lower resolution input reduces the computational cost

Proposed Model: Detail Branch

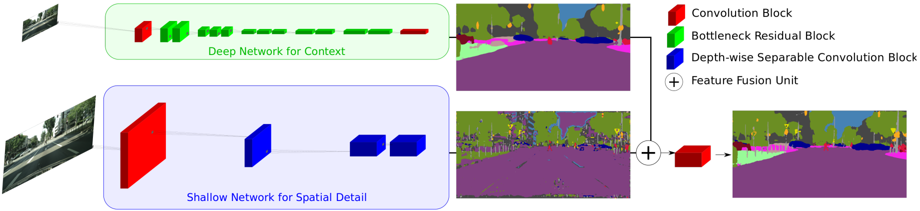
- **Detail branch** focuses on high resolution segmentation details



- No need for very deep network to detect segmentation boundary

Proposed Model: Combined Branchs

- **Context branch** at low resolution captures global context information
- **Detail branch** focuses on high resolution segmentation details

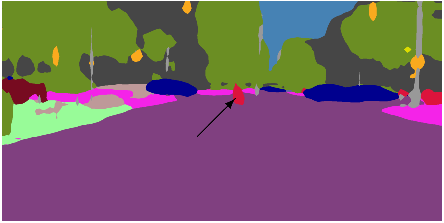


- Losses at context and detail branches help to learn auxiliary tasks
- Efficiently learning global context and spatial detail separately to reduce cost

Proposed Model: Qualitative Validation

Input image

ContextNet: using Both Branches



Context Branch

Detail Branch

Proposed Model: Qualitative Validation

Input image

ContextNet: using Both Branches

Context Branch

Detail Branch

Proposed Model: Qualitative Validation

Input image

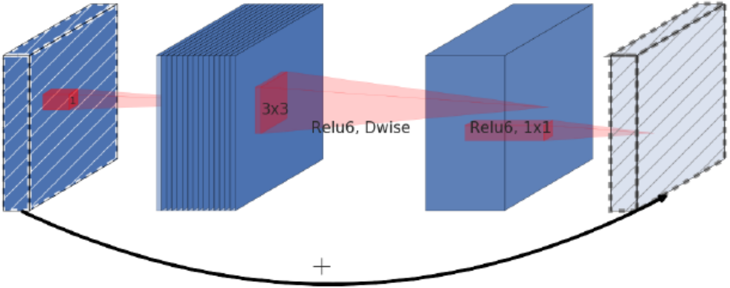
ContextNet: using Both Branches

Context Branch

Detail Branch

Network Design

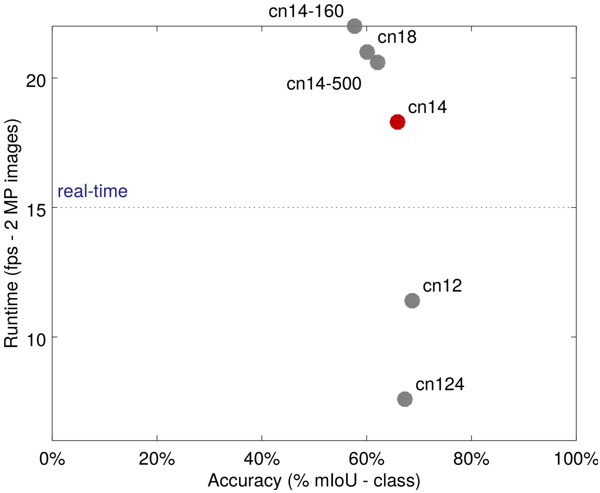
- Depthwise Convolution
  - Factorizes standard convolution to spatial and 1x1 conv(s)
  - Fewer number of parameters
  - Fewer number of floating point operations



Bottleneck residual block (Sandler et al., 2018)

Network Design

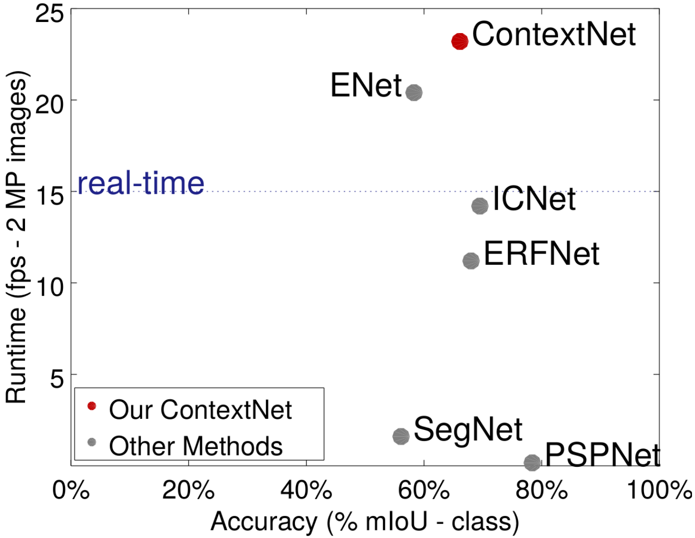
- Multi-scale features fusion
  - Two branches (cn14) balances between accuracy and runtime
  - cn14 with 160K params get 57.7% mIoU in Cityscapes (Cordts et al., 2016)



Network Pruning

- Pruning:
  - Start with “wider” network
  - Pruning to obtain “skinnier” network
- Pruning strategy improves accuracy compared to direct training!
- Lottery ticket hypothesis (Frankle et al., 2018):
  - More feature channels $\implies$ more chances of success

ContextNet: Quantitative Evaluation

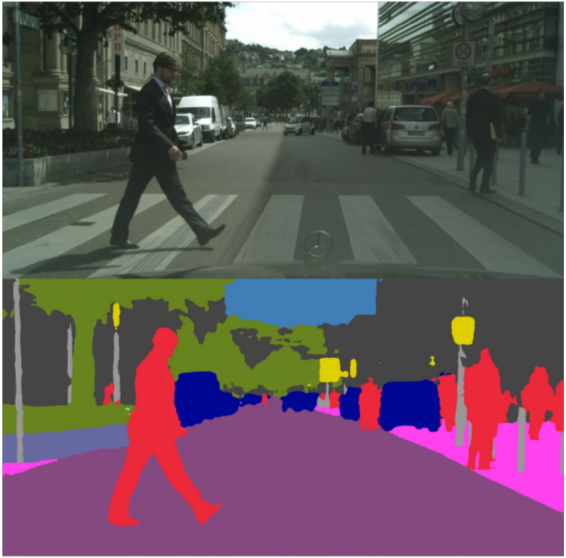


ContextNet: Quantitative Evaluation

- Runtime measured on Nvidia Titan X (Maxwell, 3072 CUDA cores)
- ContextNet balances accuracy and speed

	Class mIoU%	Category mIoU%	Parameters in Millions	1024x2048
SegNet	56.1	79.8	29.46	1.6
ENet	58.3	80.4	0.37	20.4
ICNet*	69.5	-	6.68	14.2
ERFNet	68.0	86.5	2.1	11.2
ContextNet	66.1	82.7	0.85	**23.2**

ContextNet: Qualitative Evaluation



road 0	
sidewalk 1	
building 2	
wall 3	
fence 4	
pole 5	
traffic light 6	
traffic sign 7	
vegetation 8	
terrain 9	
sky 10	
person 11	
rider 12	
car 13	
truck 14	
bus 15	
train 16	
motorcycle 17	
bicycle 18	

Conclusion

- **ContextNet:**
  - Efficiently learn global and local context separately
  - Runs in real-time for 2 megapixels images
    - 2048x1024 images @ >16 fps in Nvidia Jetson TX2
- Our pruning strategy increases accuracy
- Limitations: accuracy gap with bigger off-line models

References

Burt, P.J. and Adelson, E.H., The Laplacian pyramid as a compact image code. In Readings in Computer Vision, 1987.

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.

Frankle, J. and Carbin, M., The lottery ticket hypothesis: Training pruned neural networks. In arXiv:1803.03635, 2018.

He, K., Zhang, X., Ren, S. and Sun, J., Deep residual learning for image recognition. In arXiv:1512.03385, 2015.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C., Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. In arXiv:1801.04381, 2018.

Questions?

Thank you for your attention!