Boosting over Deep Convolutional Channel Features for Scene Perception

Arthur Costea

Research Center for Image Processing and Pattern Recognition

Technical University of Cluj-Napoca

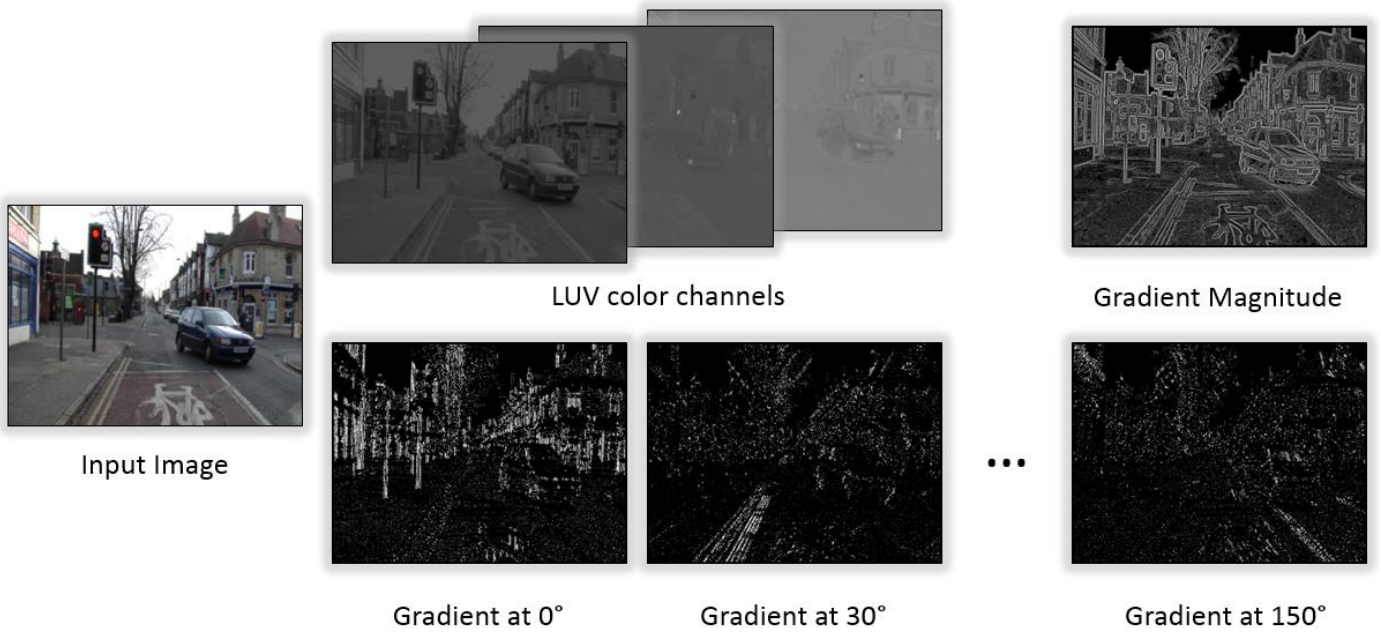
2017 IEEE International Conference on Intelligent Computer Communication and Processing
September 7-9, Cluj-Napoca, Romania

Overview

- Perception tasks:
  - Object detection
  - Semantic segmentation
- Proposed solution:
  - Channel-like image features
    - Multiresolution Filtered Channels
    - Multimodal Channels
    - Deep Convolutional Channels
  - Boosting over channel features
    - Easy fusion of different features types
    - Low computational costs

Image Channels

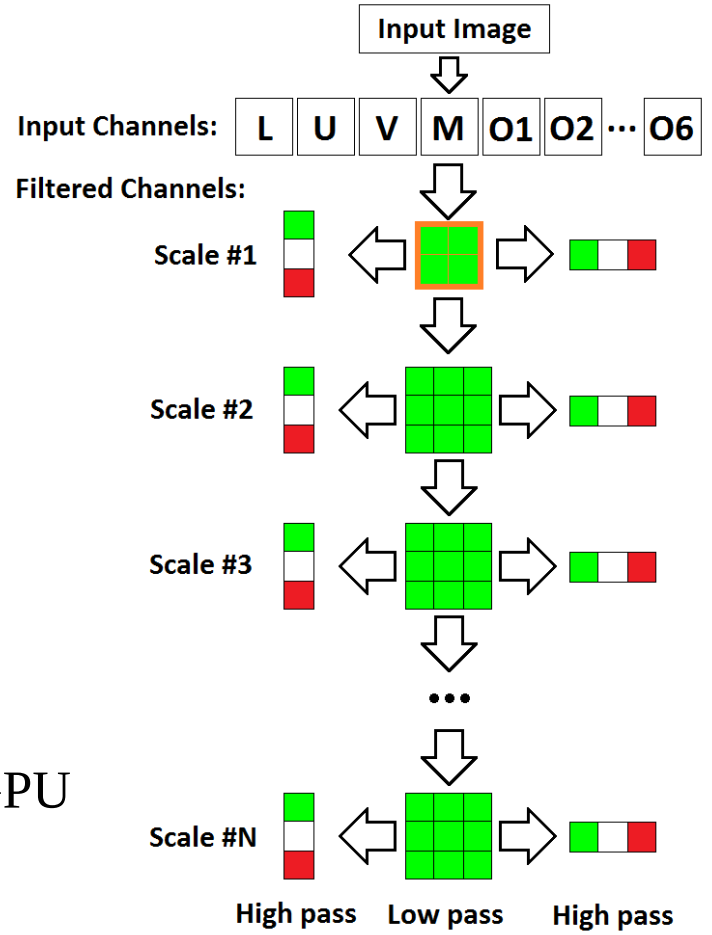
- 10 LUV + HOG image channels [Dollar et al. 2009]:
  - 3 LUV channels
  - 1 gradient magnitude
  - 6 oriented gradient magnitudes



Multiresolution Filtered Channels

- **Multiresolution** filtering scheme:
  - **Low pass** and **high pass** filters
  - Applied iteratively at **multiple scales**
  - 7 scales => (5 x 3) x 10 = 150 channels

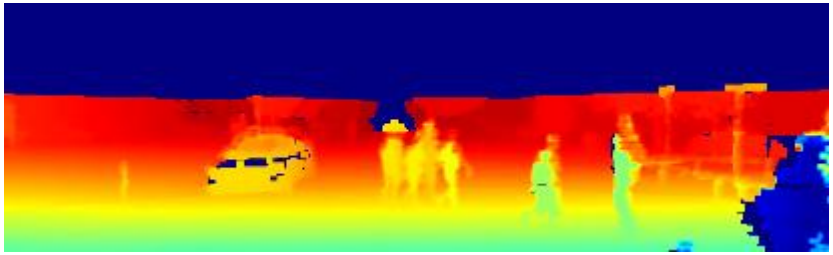
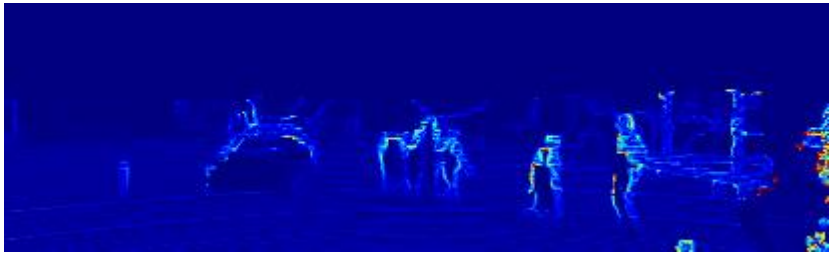
- **Efficient implementation**:
  - **< 3 ms** for a 640 x 480 pixel image on GPU



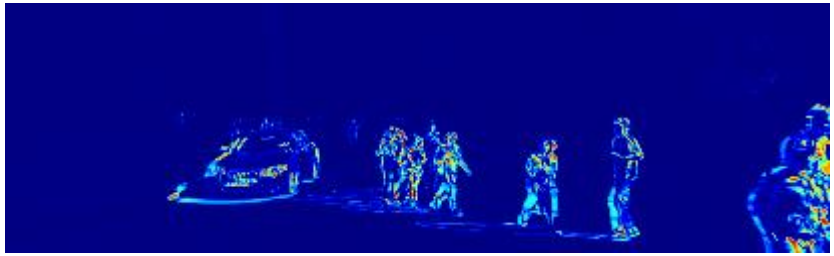
Multimodal Sensorial Input

Color

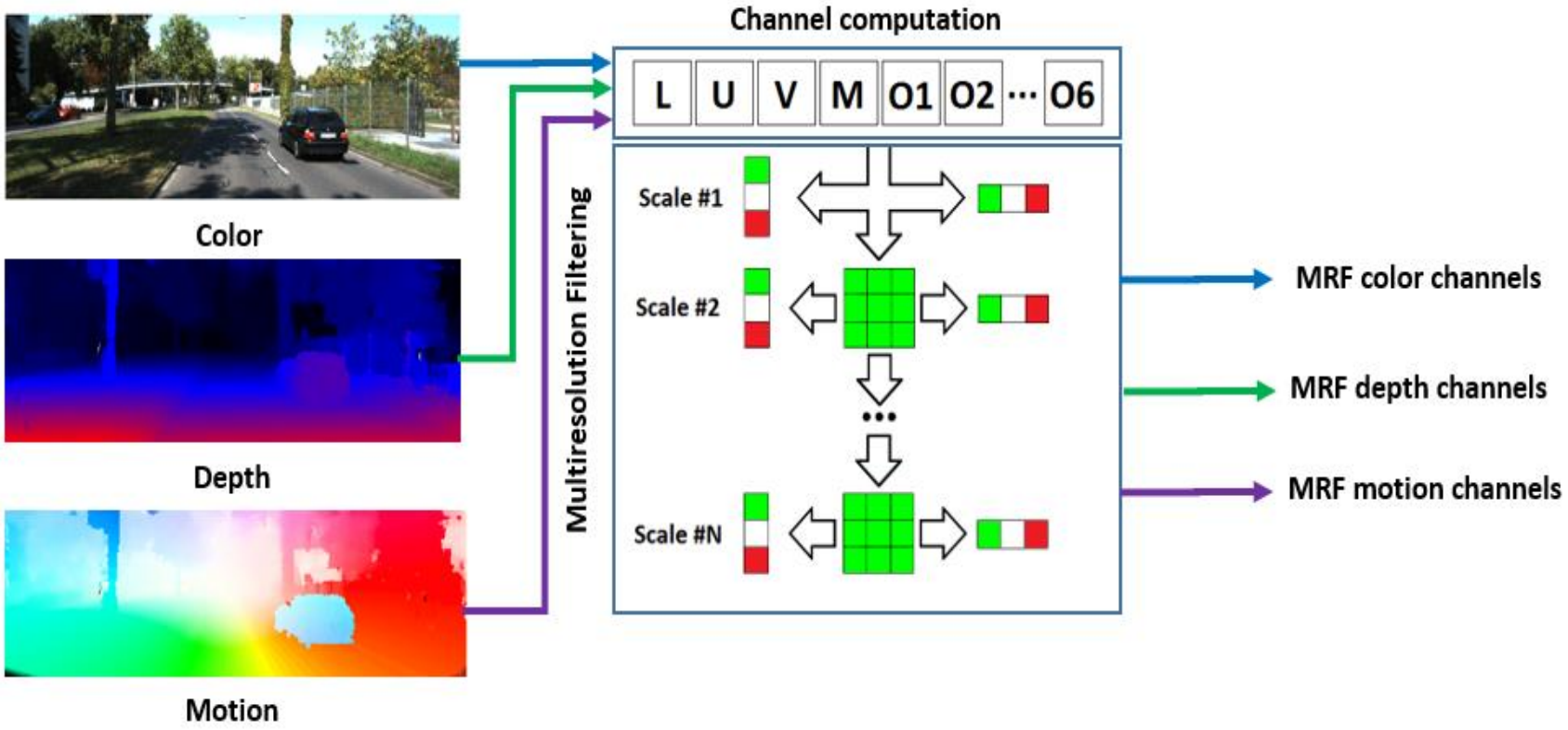

Depth

Motion

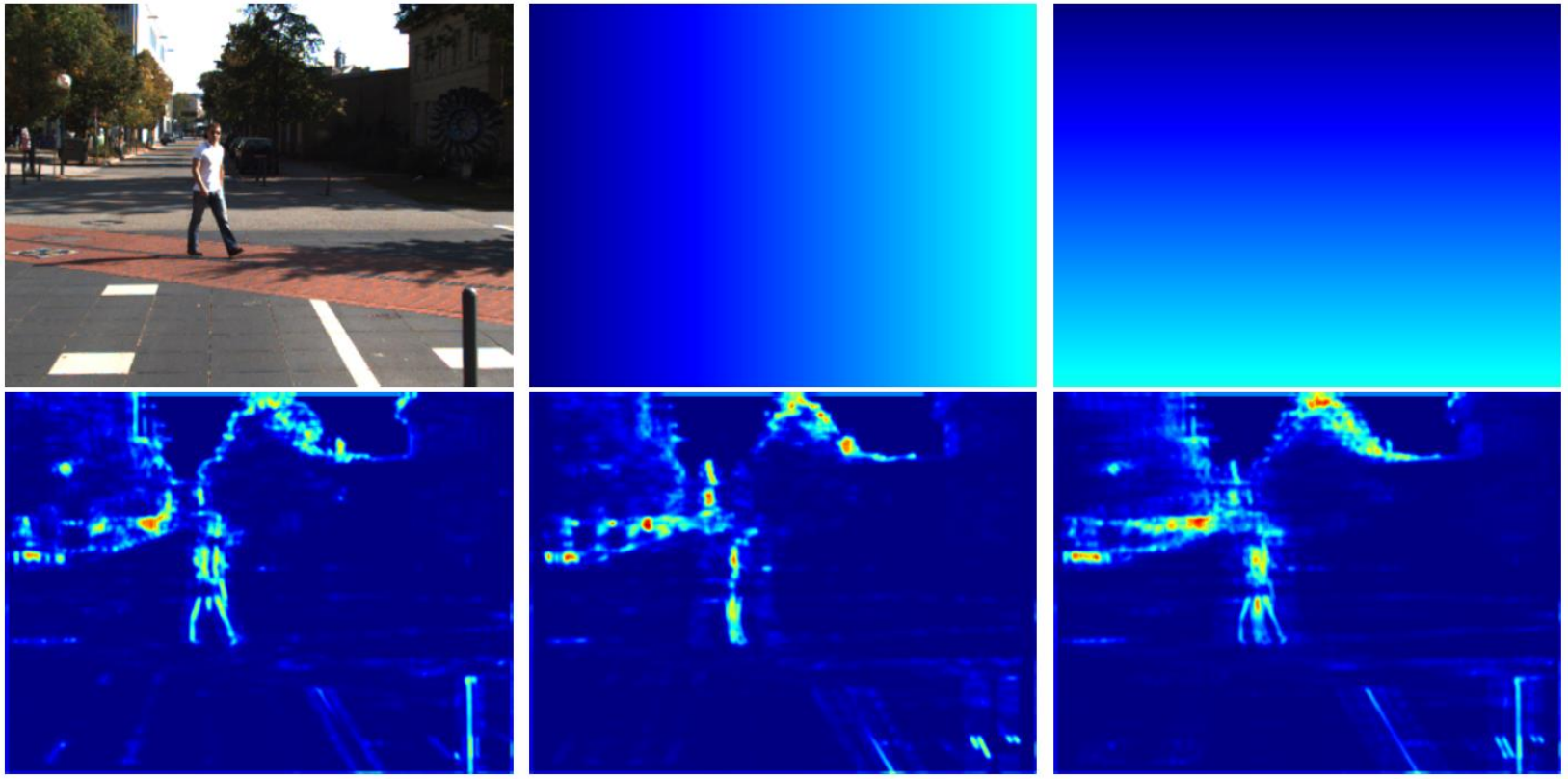



Multimodal Multiresolution Channels



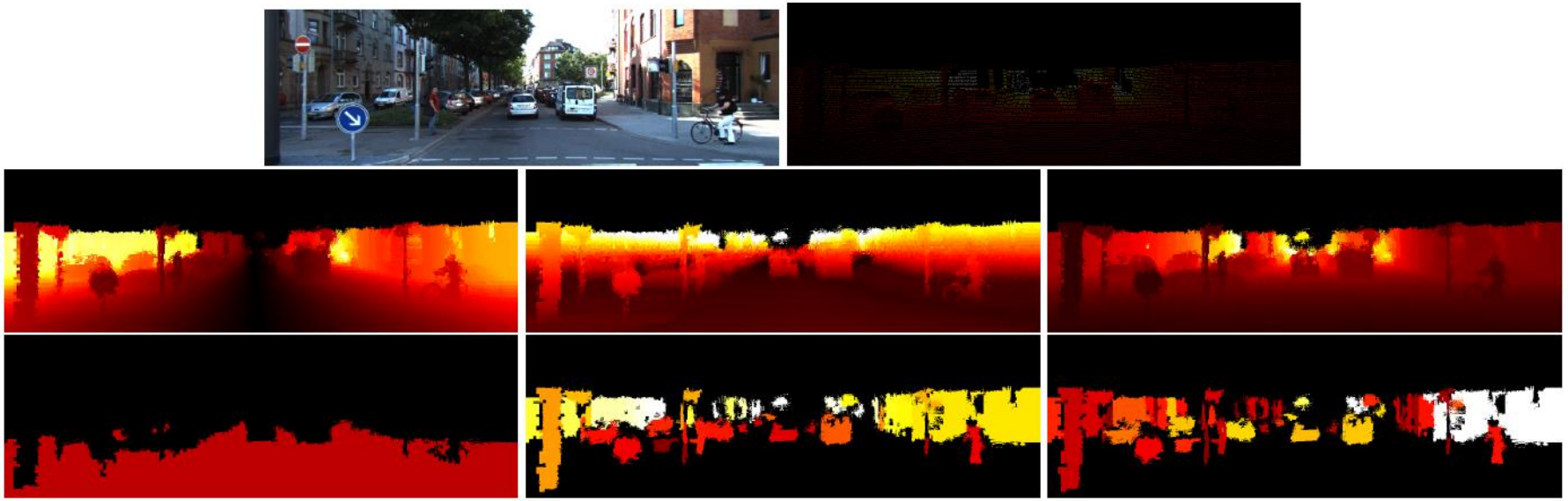
2D context channels

2D spatial and symmetry channels:

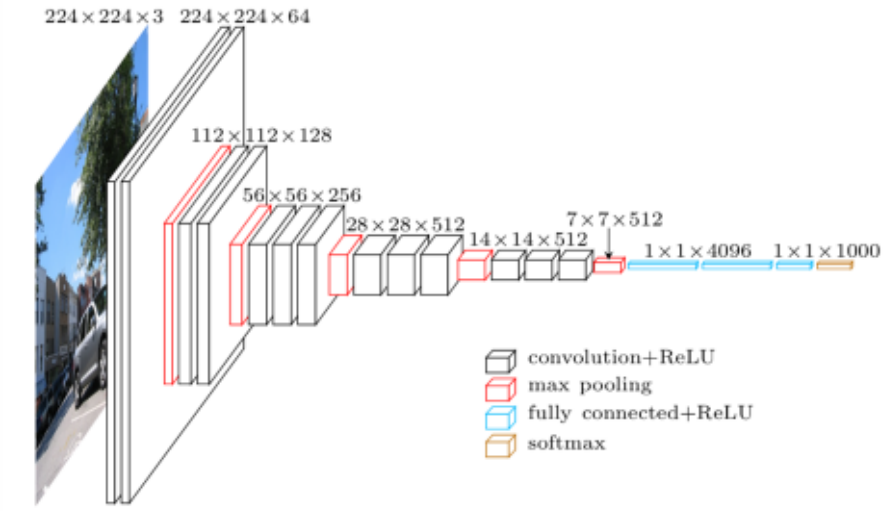


3D Context Channels

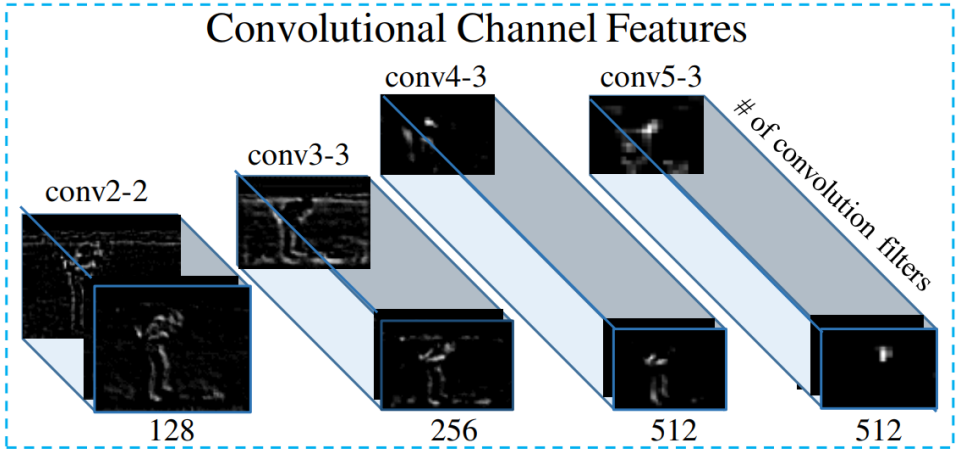
- 3D Context channels:
  - Spatial channels: X, Y, Z
  - Ground Plane
  - Geometric channels: height, width, size



Deep Convolutional Channels

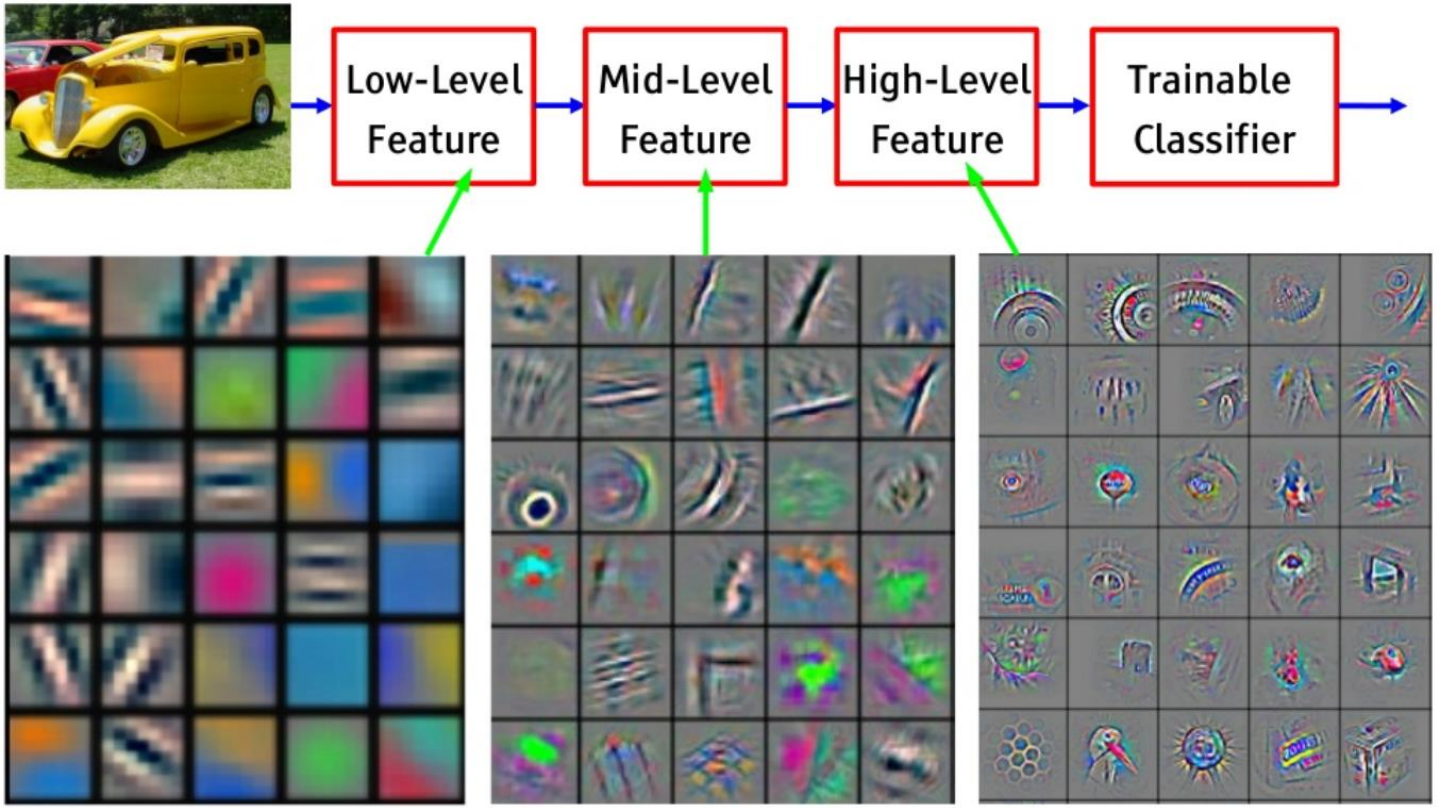


Input Image



Deep Convolutional Channels

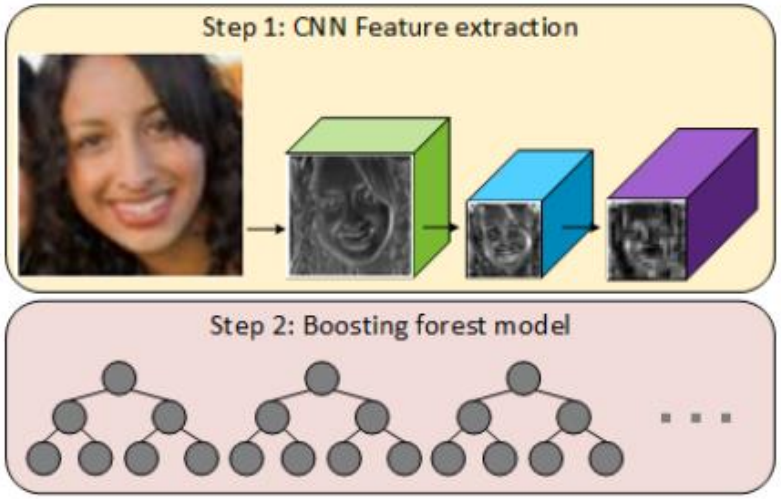
Convolutional net feature visualization [Zeiler & Fergus 2013]



Deep Convolutional Channels

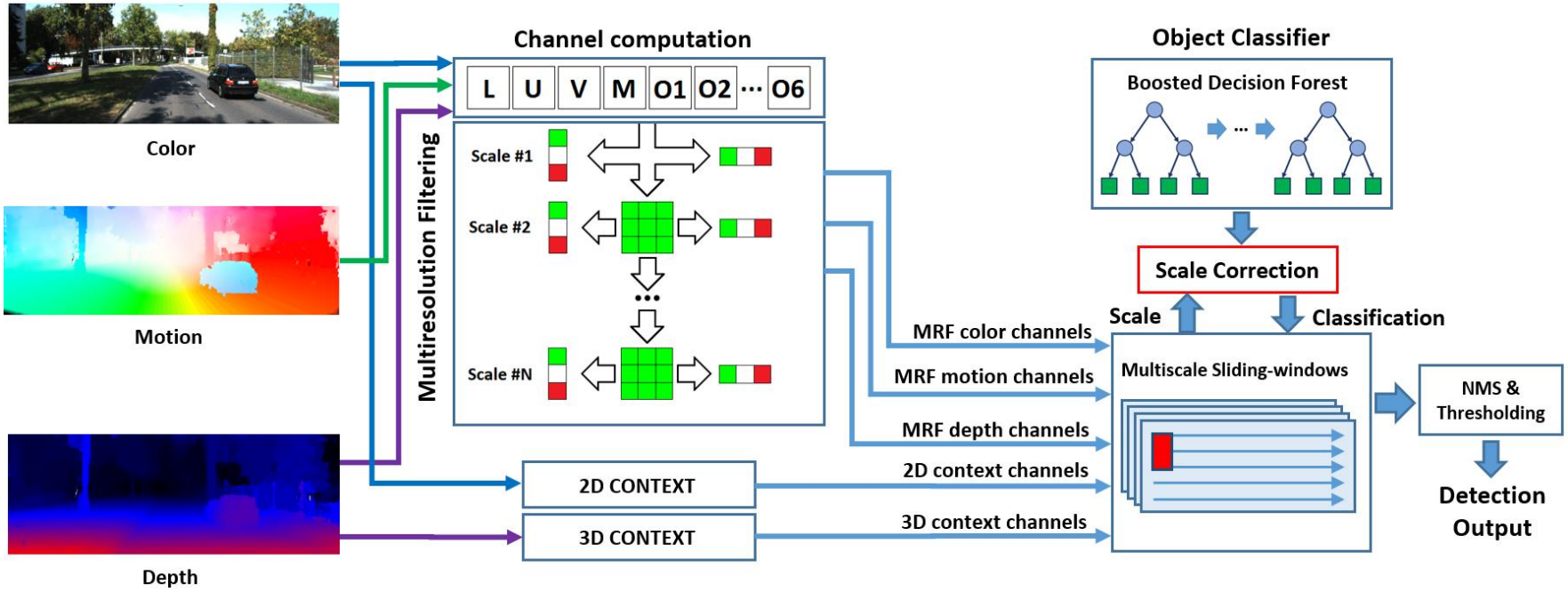
Convolutional channel features [Yang et al. 2015]:

- best results for pedestrian detection using the standard VGG16 pre-trained model
- VGG16 was trained for 2 weeks on ImageNet (over 1 million images, 1000 classes)



	Output layer	#Output maps	Filter size	#Ds	Miss Rate(%)
ACF	-	10	3	4	**41.22**
LDCF	-	40	7	4	**38.66**
ANet-s1	conv1	96	11	4	61.65
	conv2	256	5	4	51.52
	conv3	384	3	4	**43.73**
	conv4	384	3	4	48.37
	conv5	256	3	4	53.37
VGG-16	conv2-2	128	3	4	53.86
	conv3-3	256	3	4	31.28
	conv4-3	512	3	8	**27.66**
	conv5-3	512	3	16	51.52
VGG-19	conv2-2	128	3	4	51.25
	conv3-4	256	3	4	33.56
	conv4-4	512	3	8	**30.17**
	conv5-4	512	3	16	55.55
GNet	conv2	192	3	4	45.06
	icp1	256	-	8	38.44
	icp2	480	-	8	**31.66**
	icp3	512	-	16	35.99
GNet-s1	conv2	192	3	4	49.39
	icp1	256	-	4	41.85
	icp2	480	-	4	**32.18**
	icp3	512	-	8	32.87

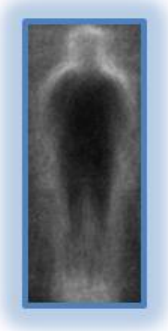
Multiscale object detection

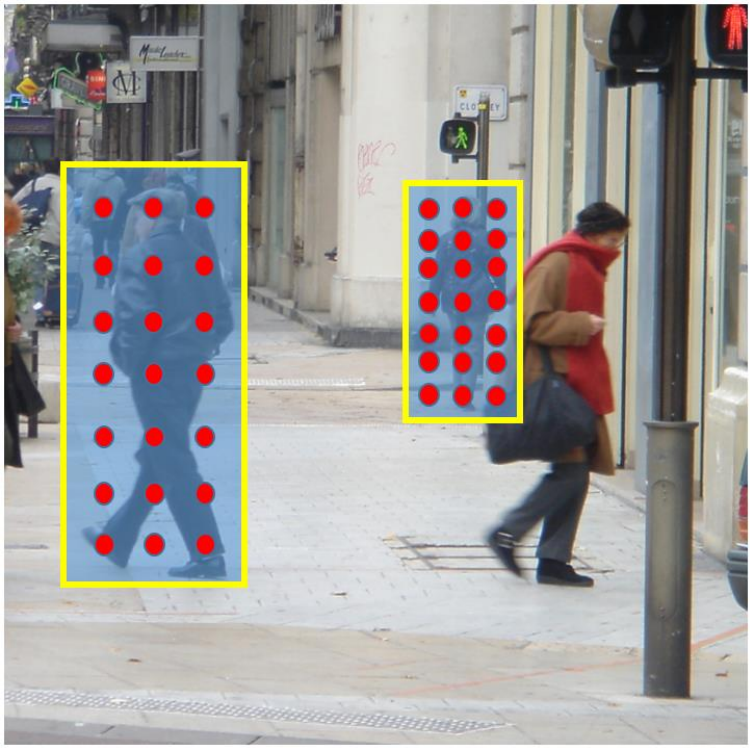


Multiscale object detection

Multiscale sliding window :

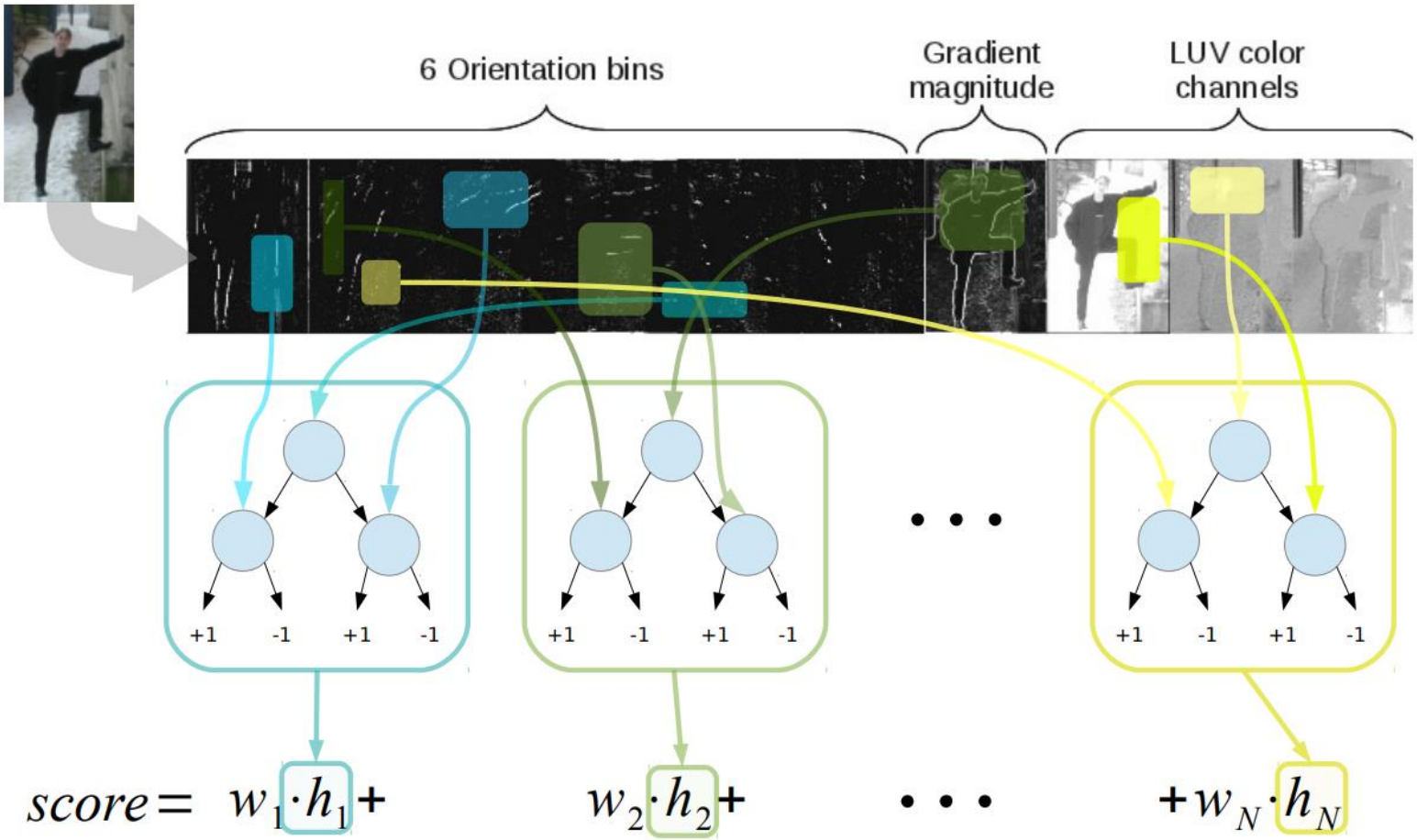
- **Single** image feature scale
- **Single** pedestrian classifier model
- Feature sampling adapted to window size



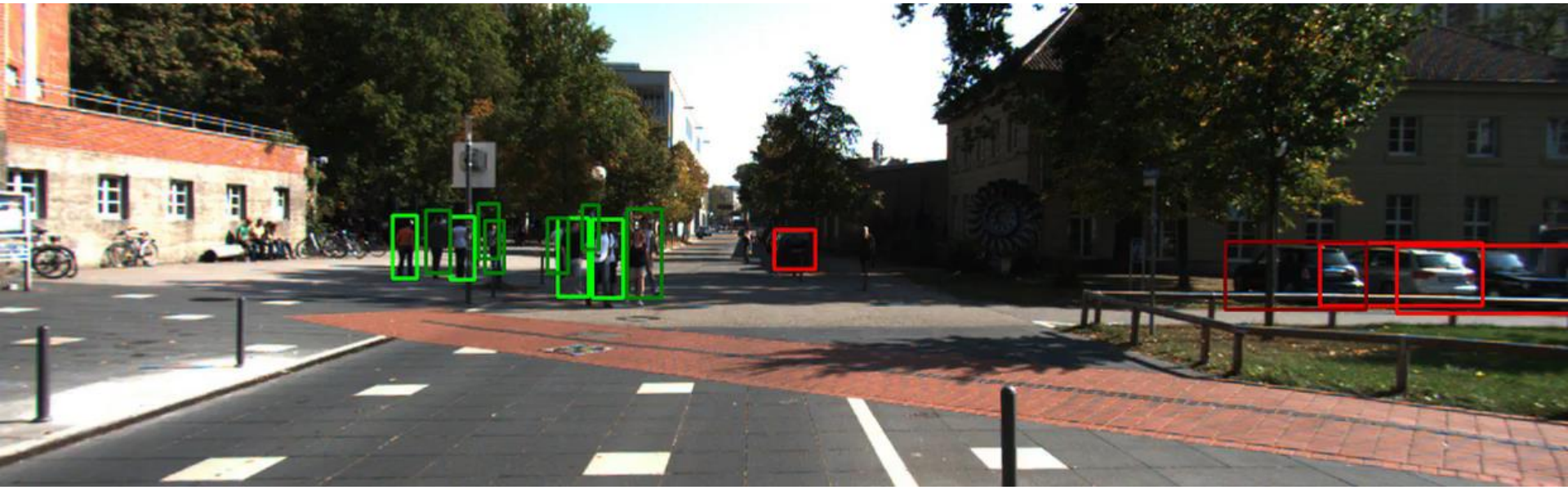


Boosting based classification

[Dollar et al. 2009, Benenson 2016]

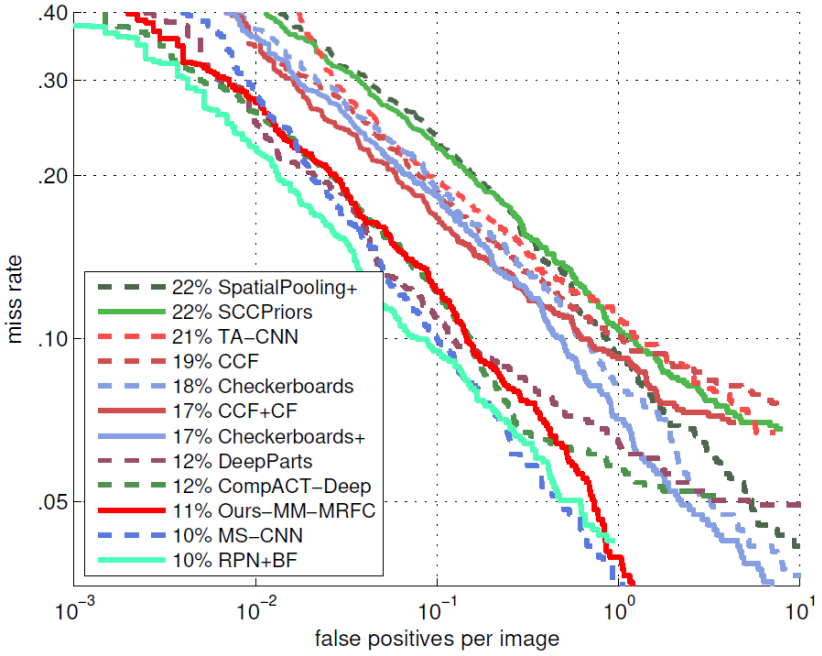


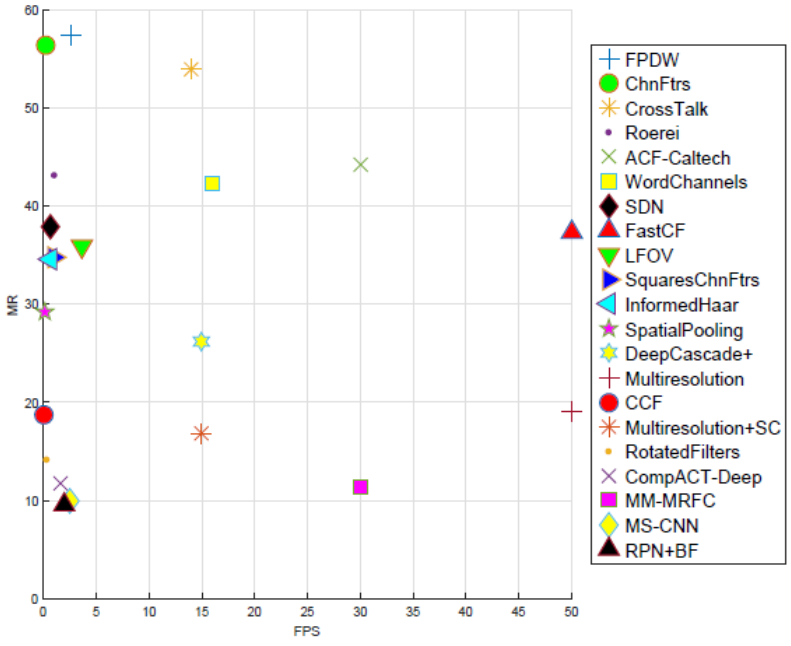
Detection Demo (KITTI)



Experimental results

- Caltech – Pedestrian benchmark
  - 11.41 % avg. MR at 30 FPS
  - 9.58 % avg. MR at 25 FPS using deep conv. chnl. features

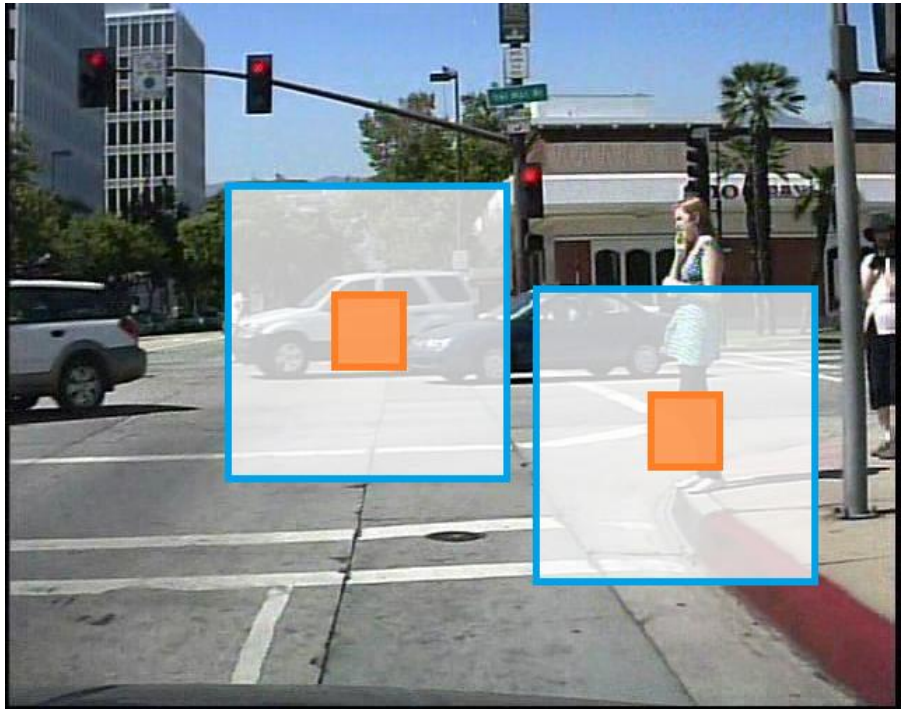




Semantic Segmentation

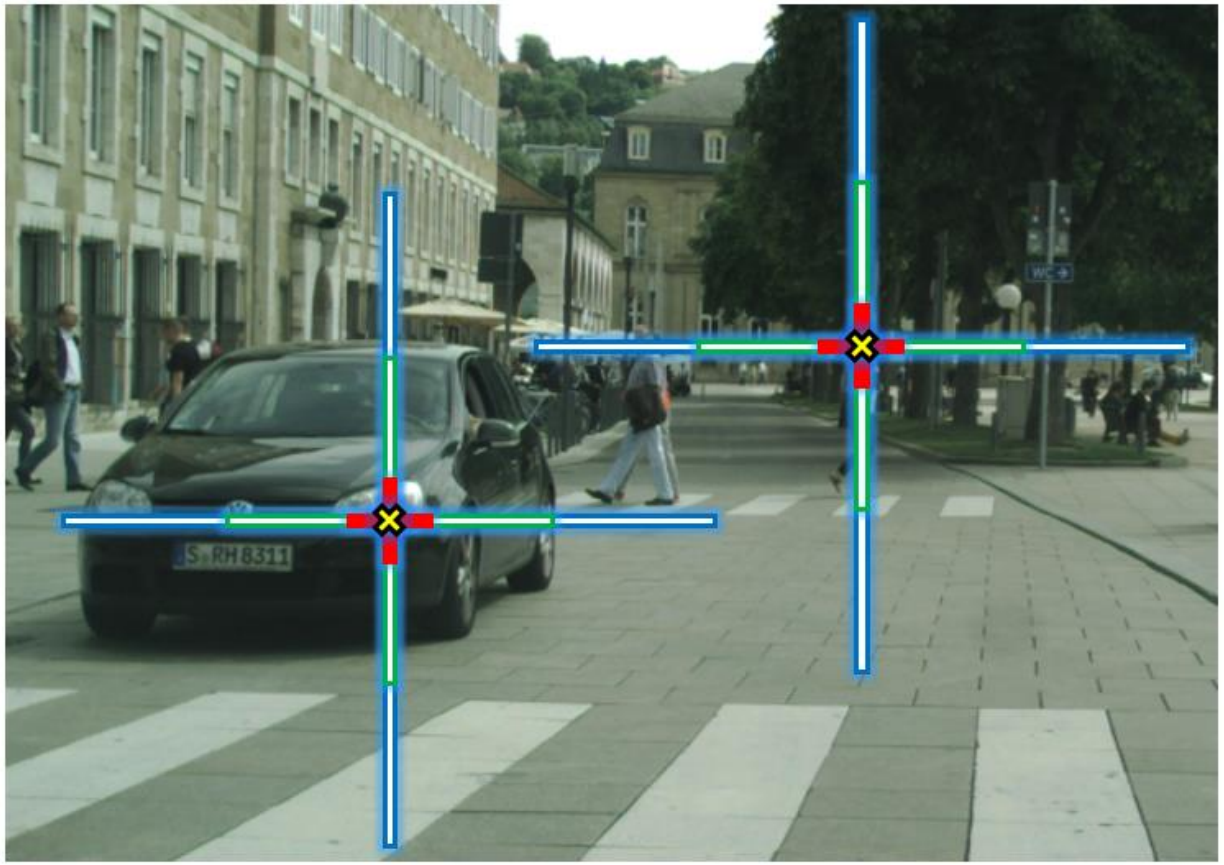
Similar classification scheme for pixels:

- Boosting over Multiresolution Channel features
- **Short** range features => **local structure**
- **Long** range features => **context**

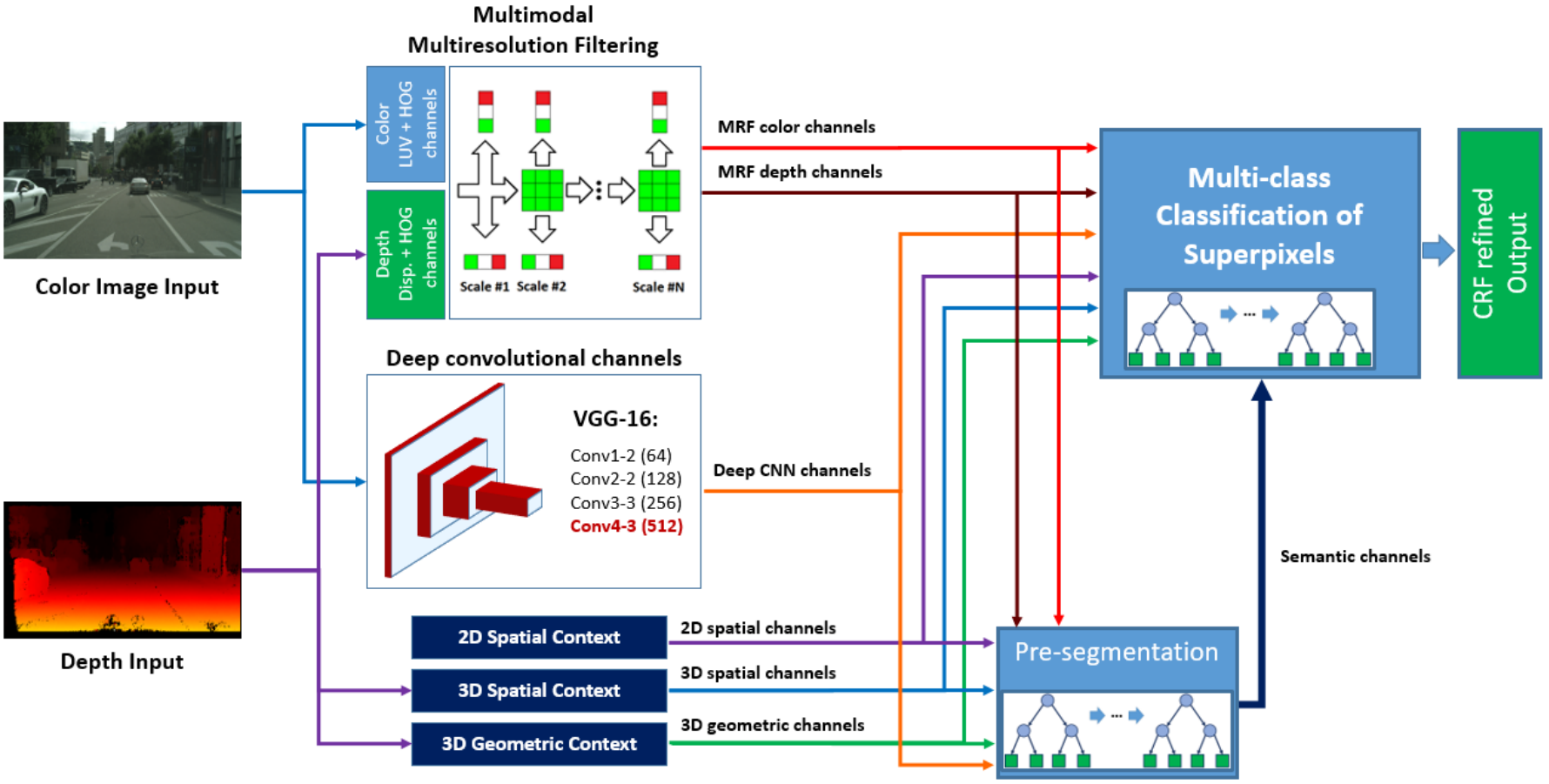


Semantic Segmentation

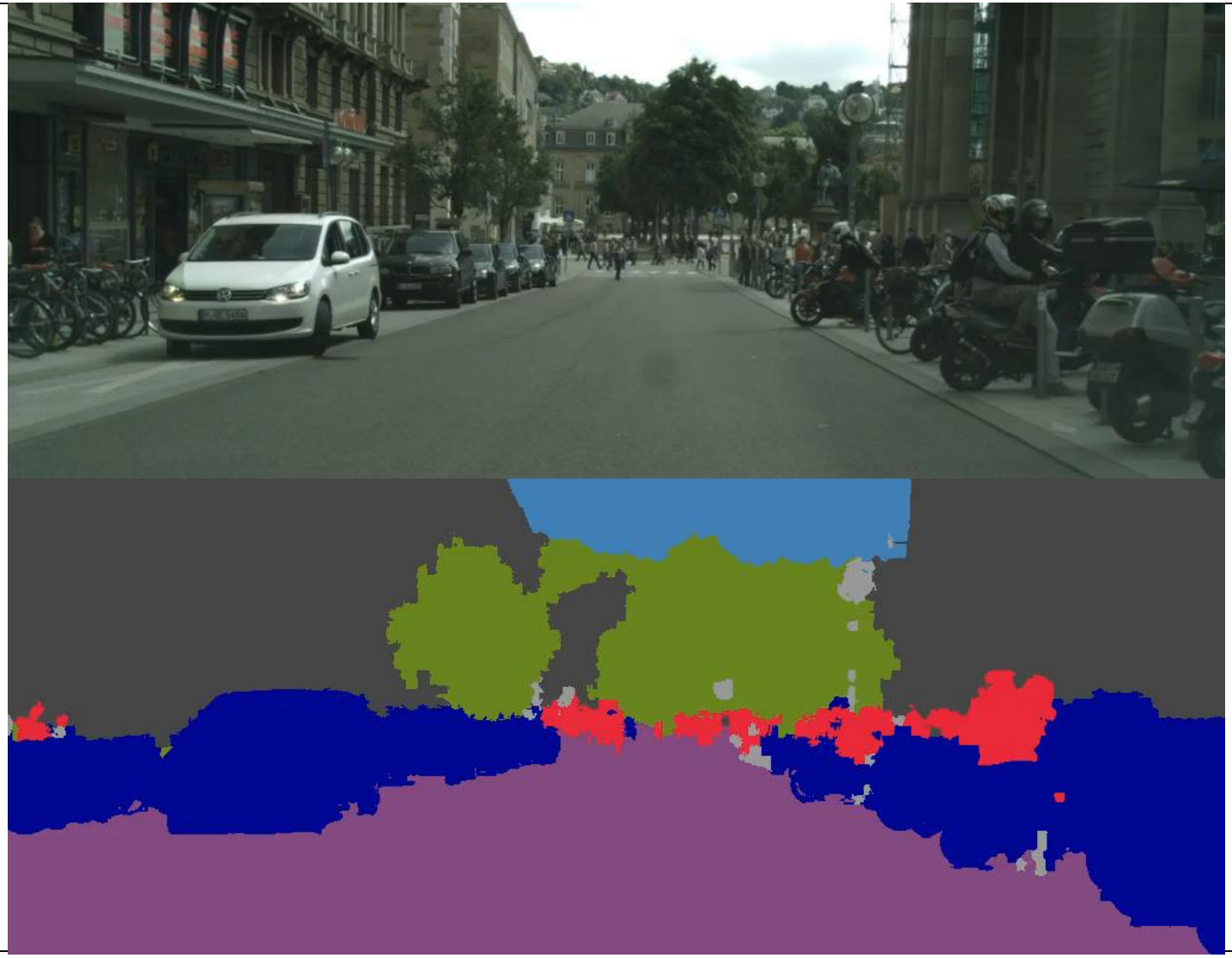
Simplified multi-range classification features:



Semantic Segmentation



Segmentation Demo (Cityscapes)

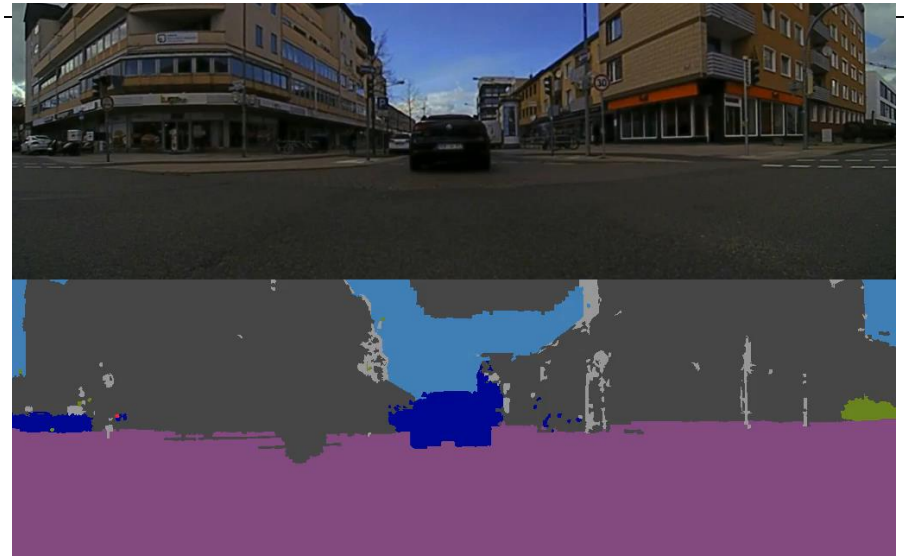


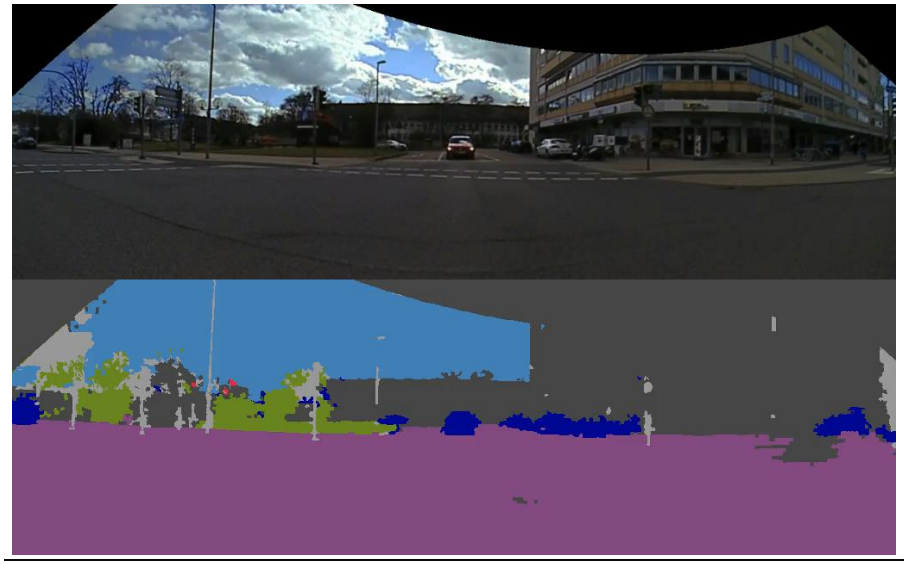
Experimental results

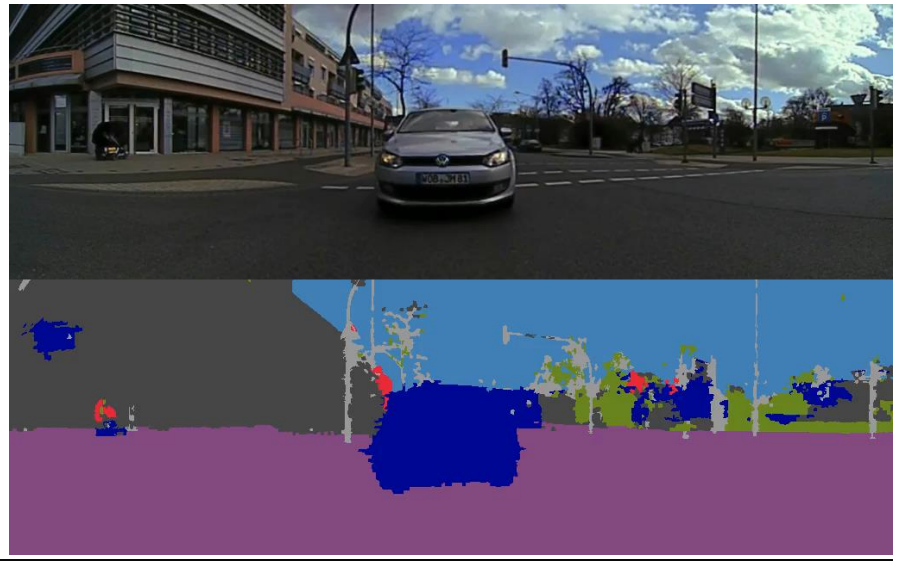
Segmentation performance using different features
(validation set – 7 classes)

	Mean IoU	Mean Acc.	Global Acc.
Low level: multimodal MRFC	71.5	81.8	90.8
+ Intermediate level: CNN channels	73.2	82.7	91.2
+ High level: 2D + 3D channels	75.1	84.6	92.1
+ Pyramidal Context	76.8	86.7	92.5
Pre-segmentation	61.2	76.8	85.4
Final segmentation	78.8	87.3	93.6
Final segmentation + CRF	79.9	88.6	94.3

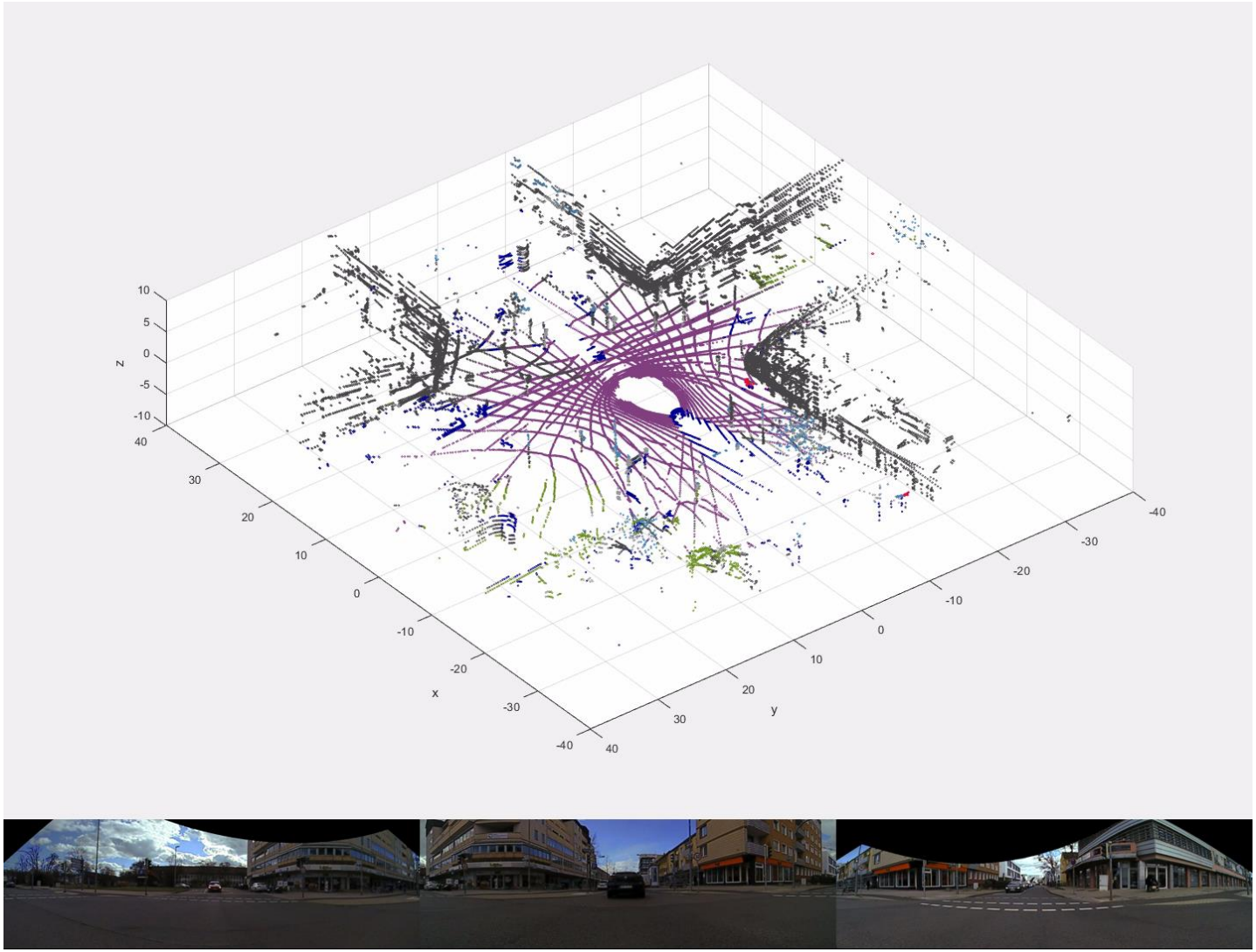
Semantic 3D perception







Semantic 3D perception



Conclusions

- Boosting over channel features:
  - enables easy fusion of different feature types
  - computational cost friendly
  - easy tuning
- Deep convolutional channels
  - captures features in a hierachical manner
  - deep neural nets are evolving quickly

    ERFNet – runs at 20 ms

Acknowledgment:
This work was supported by the EU H2020 project, UP-Drive under grant nr. 688652

Thank you for your attention!

Questions?