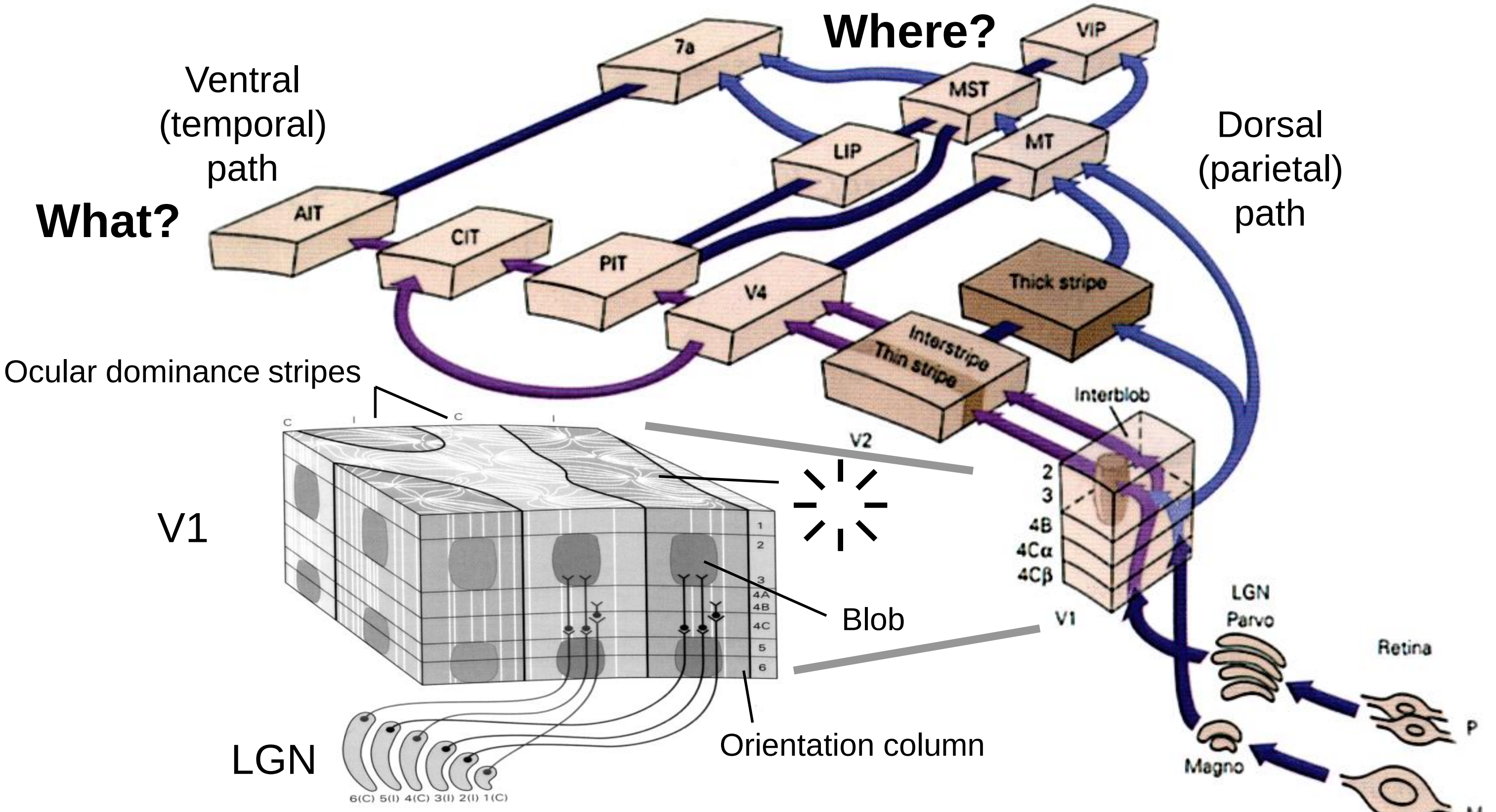
From the Neural Abstraction Pyramid to Semantic RGB-D Perception

Sven Behnke

University of Bonn, Germany

Computer Science Institute VI

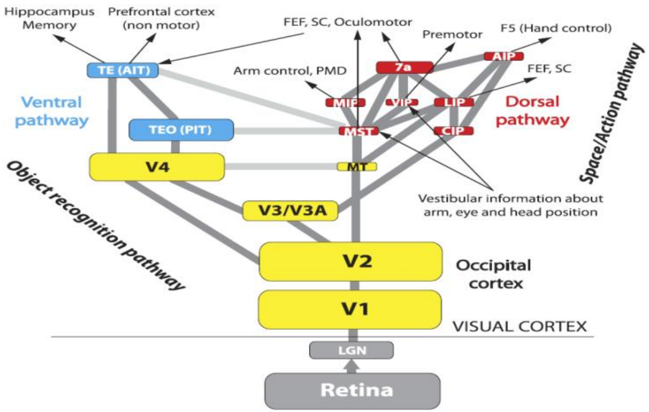
Human Visual System

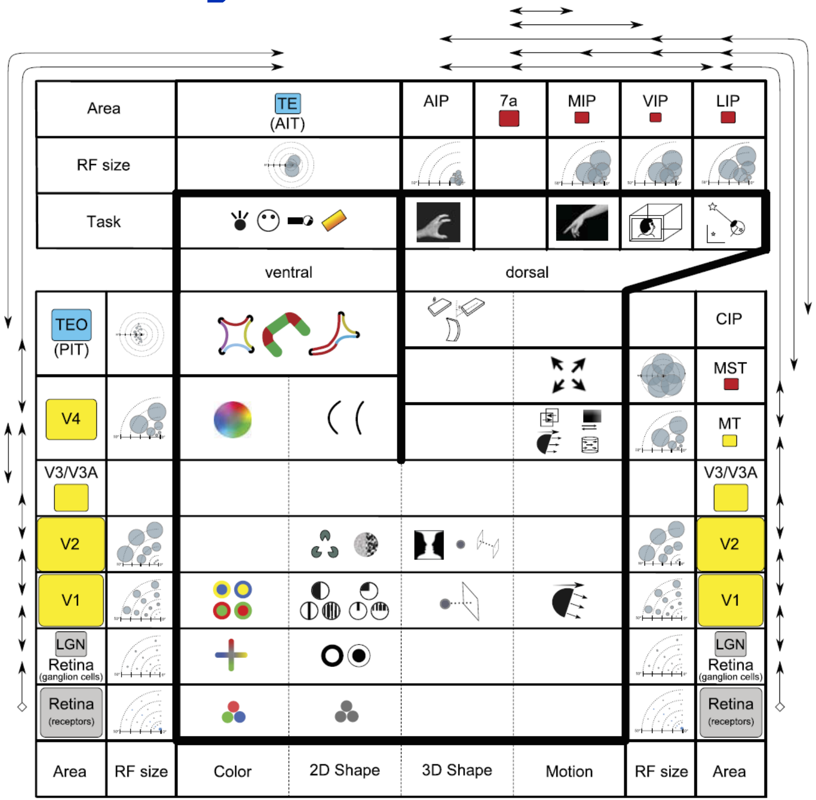


[Kandel et al. 2000]

Visual Processing Hierarchy

- Increasing complexity
- Increasing invariance
- All connections bidirectional
- More feedback than feed forward
- Lateral connections important



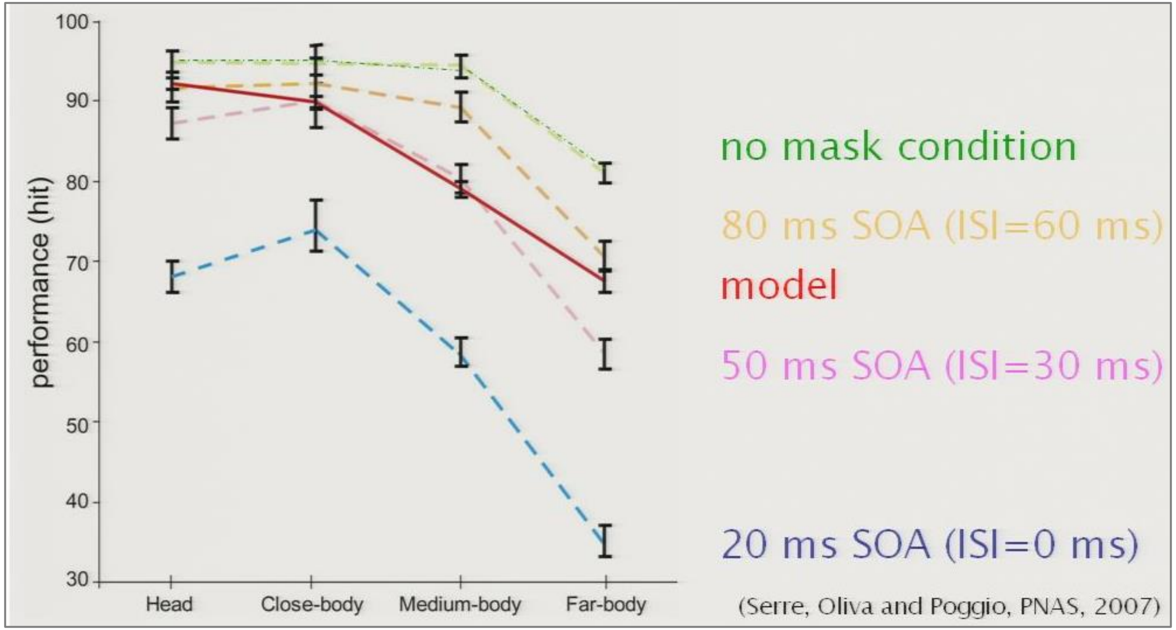


[Krüger et al., TPAMI 2013]

universitätbonn ais

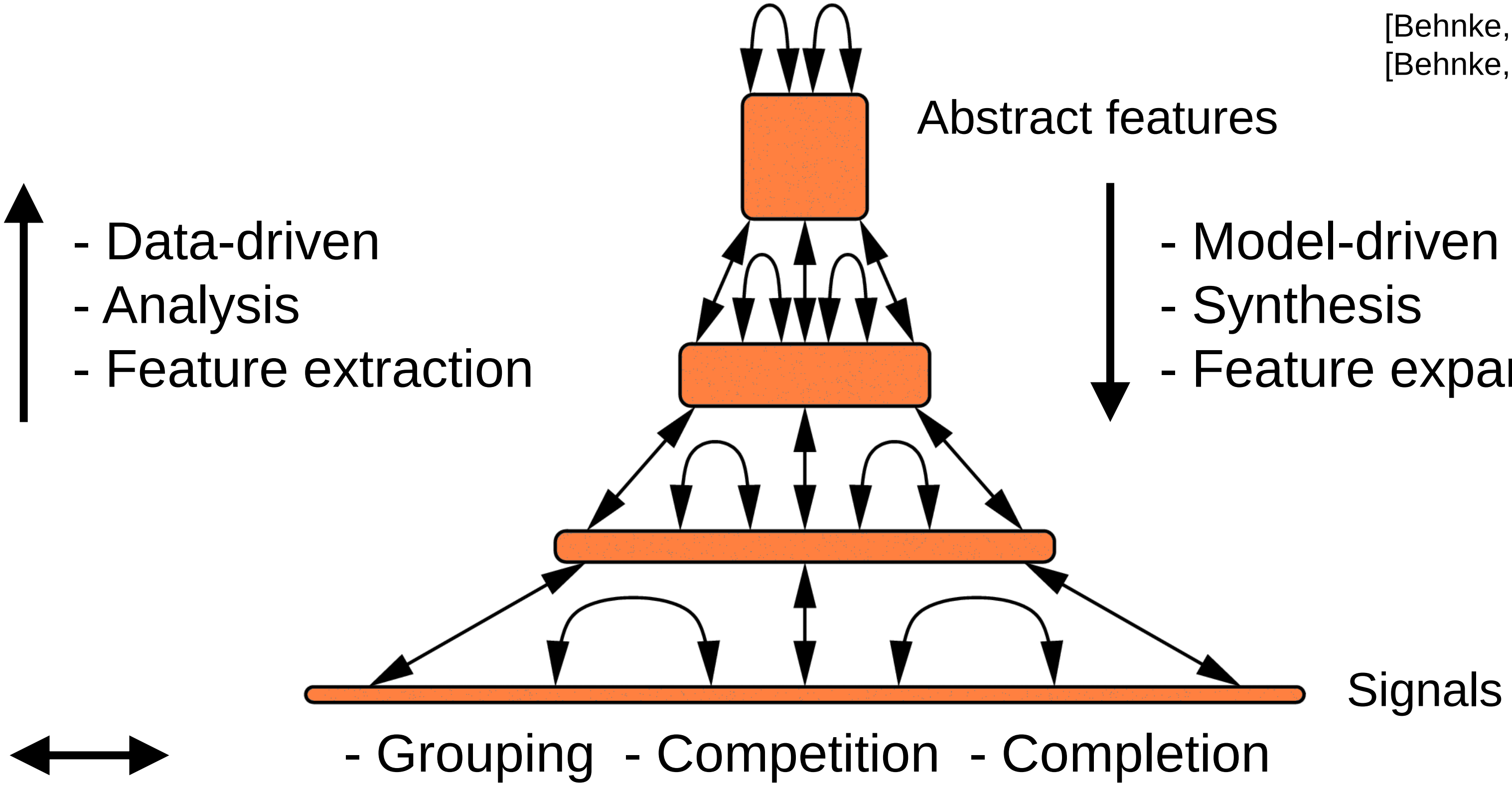
Feed-forward Models Cannot Explain Human Performance

- Performance increases with observation time



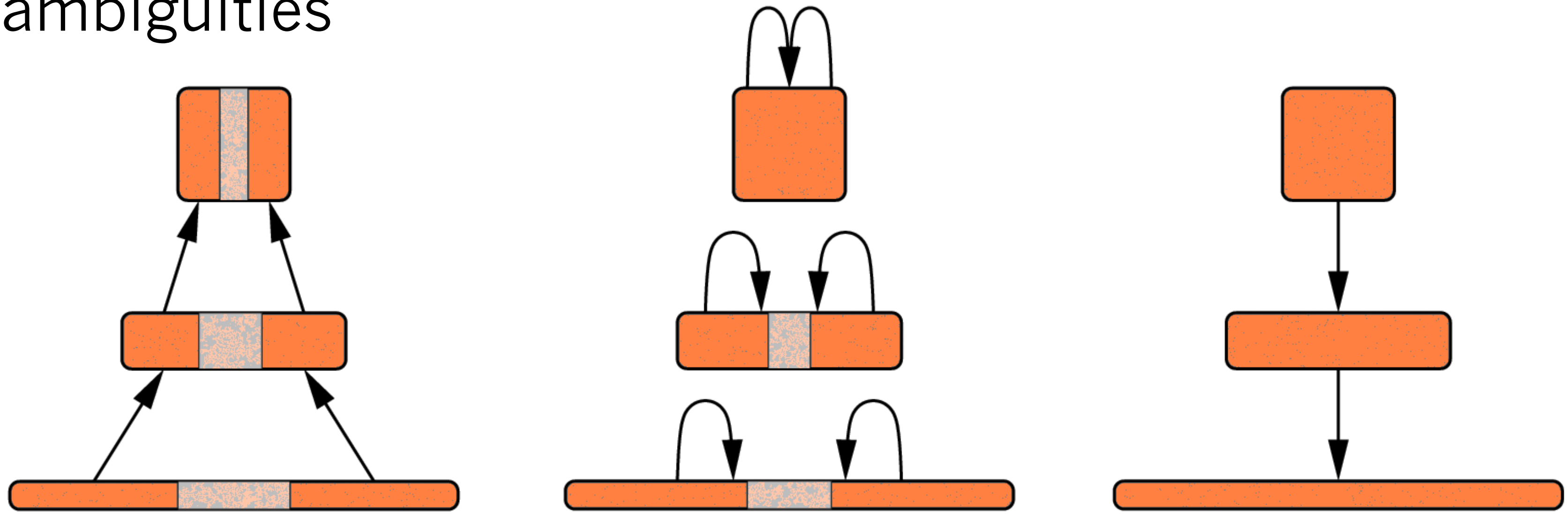
Neural Abstraction Pyramid

[Behnke, Rojas, IJCNN 1998]
[Behnke, LNCS 2766, 2003]



Iterative Image Interpretation

- Interpret most obvious parts first
- Use partial interpretation as context to resolve local ambiguities



Unsupervised Learning a Feature Hierarchy

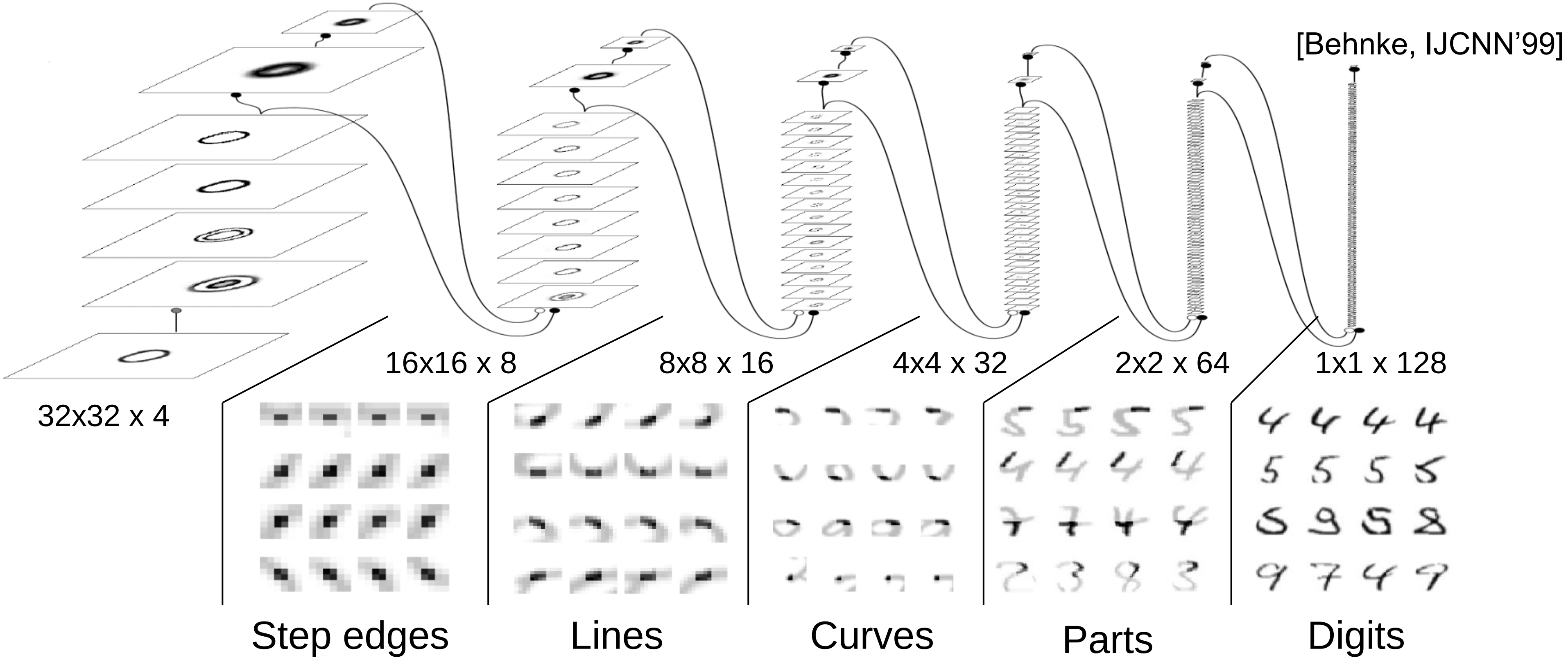
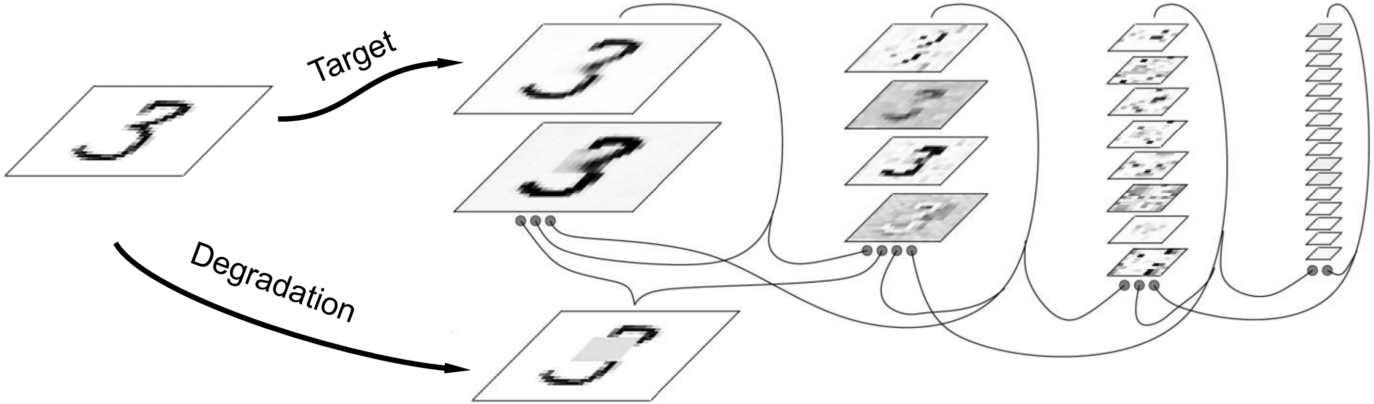
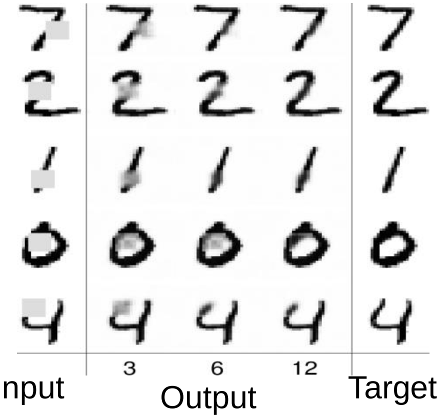


Image Reconstruction

[Behnke, IJCAI'01]





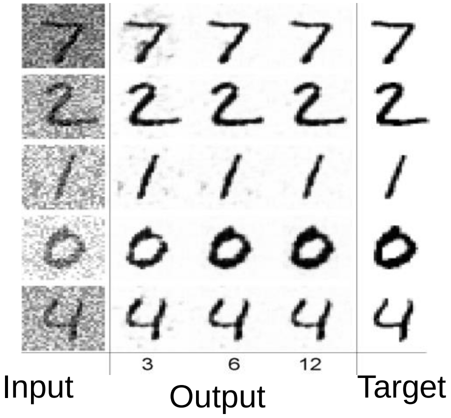
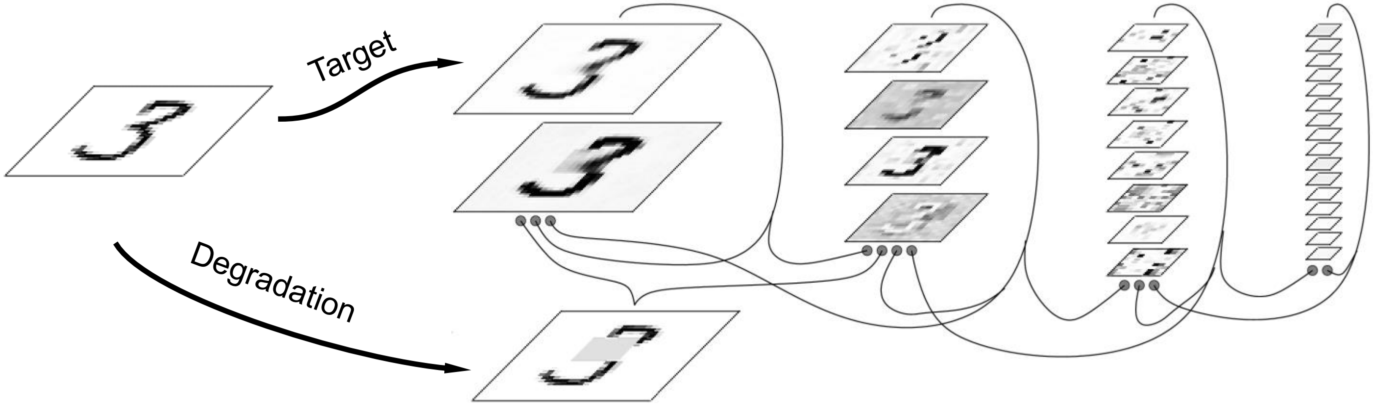
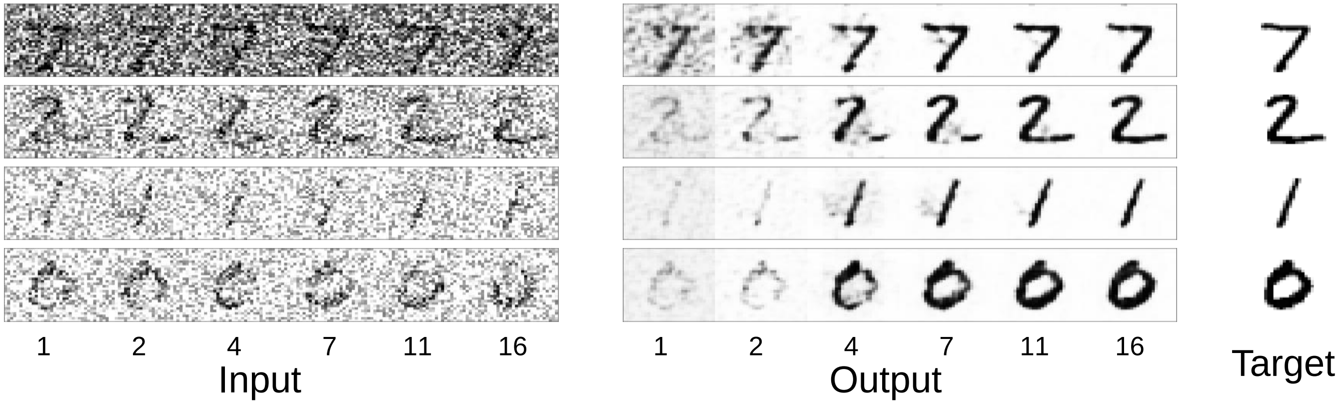


Image Reconstruction

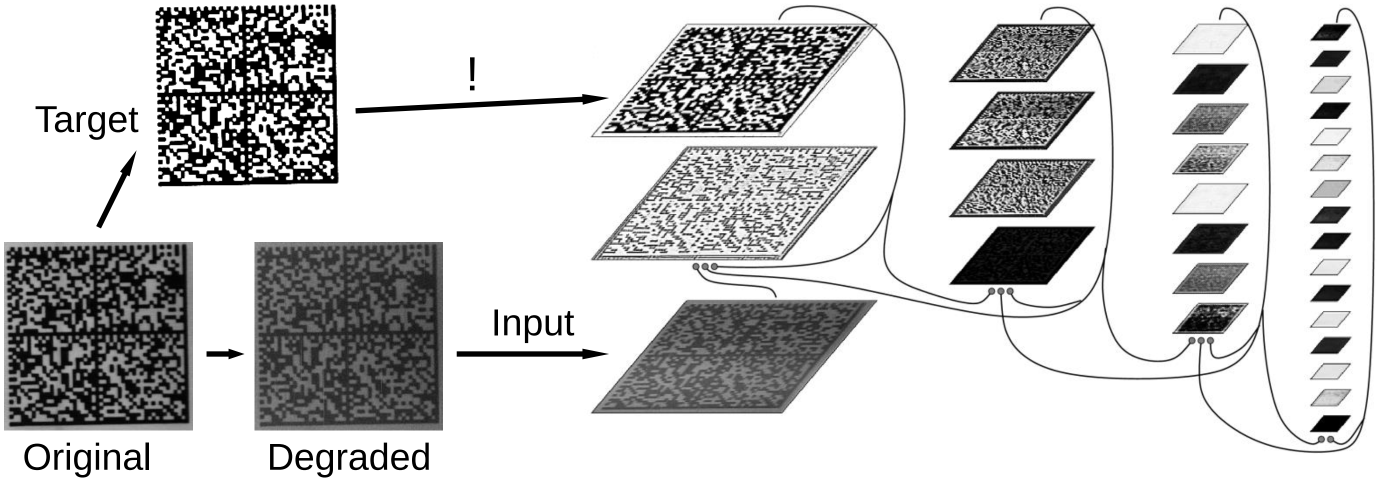
[Behnke, IJCAI'01]

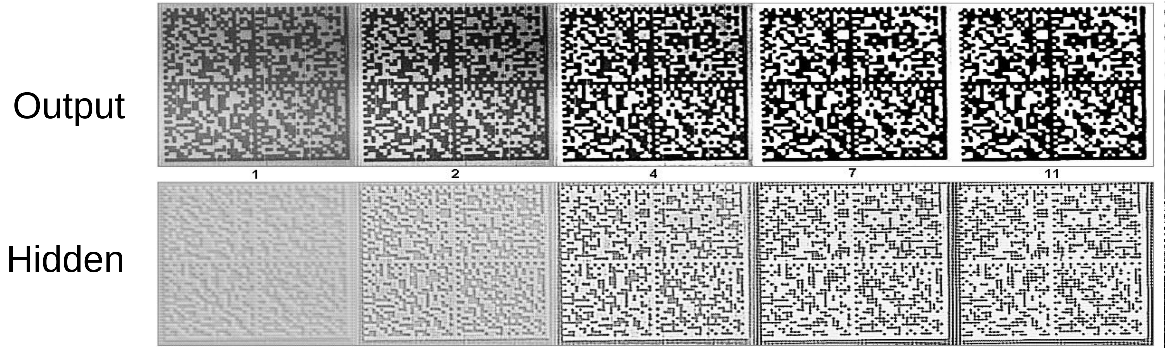


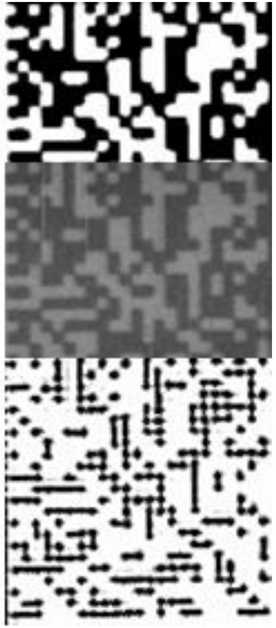


Binarization of Matrix Codes

[Behnke, ICANN 2003]





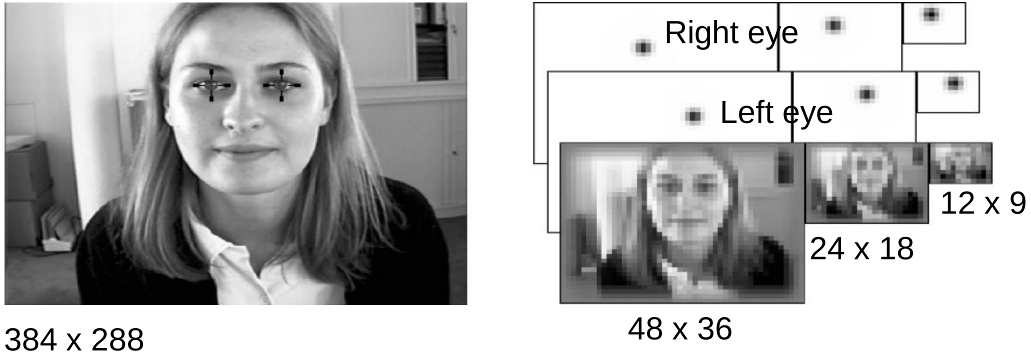


Face Localization

[Behnke, KES'03]

- BioID data set:
  - 1521 images
  - 23 persons

- Encode eye positions with blobs



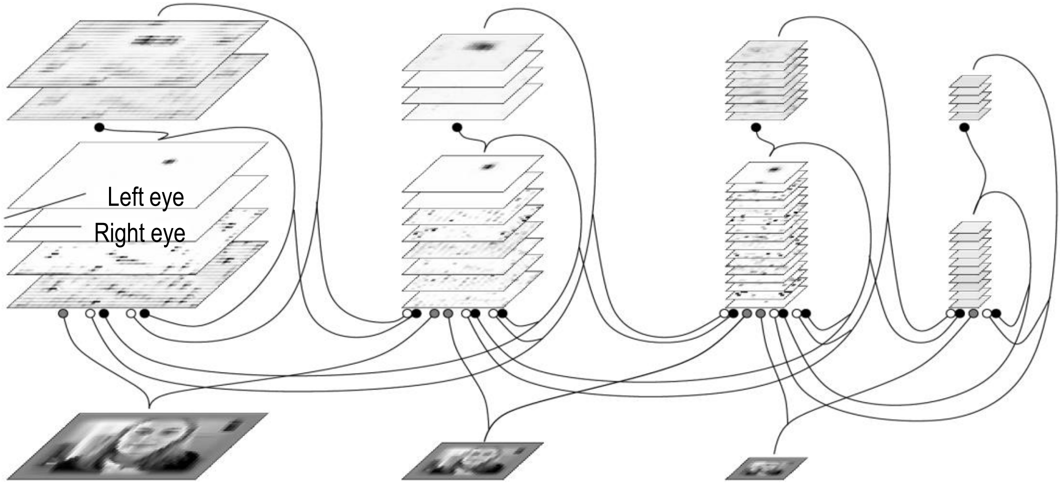
universitätbonn ais

Face Localization

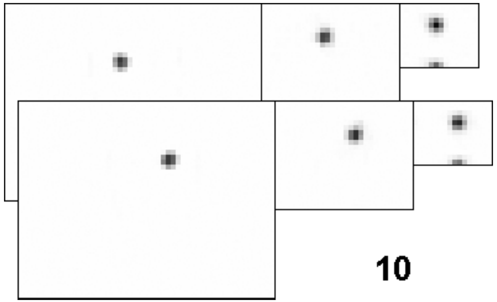
[Behnke, KES'03]

Output

Input

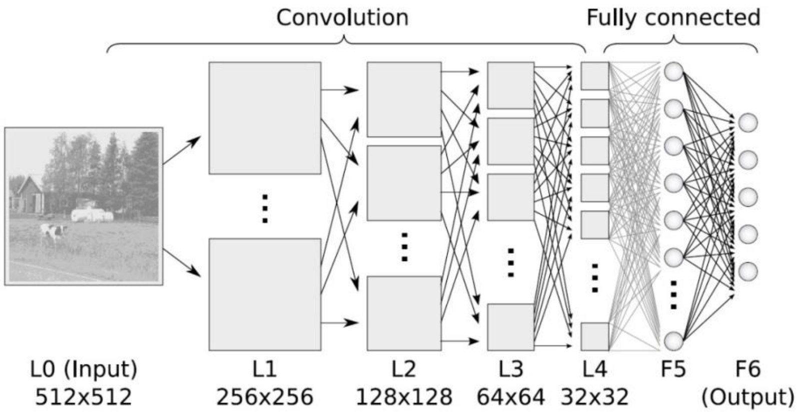


Output

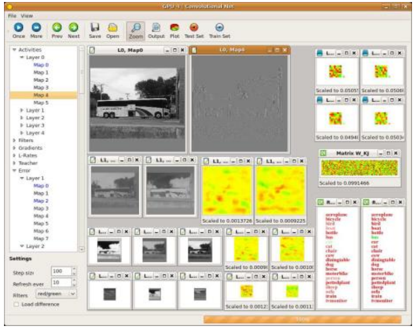


GPU Implementations with NVidia CUDA

- Affordable parallel computers
- General-purpose programming
- Convolutional [Scherer and Behnke, 2009]



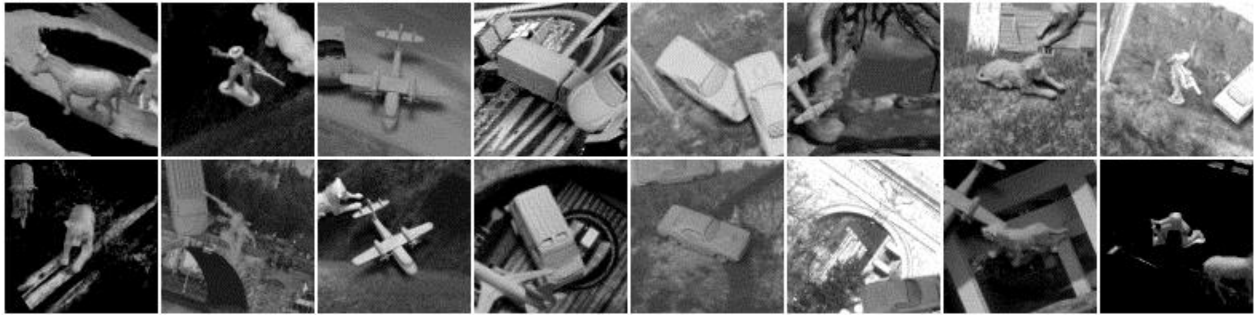
- Local connectivity [Uetz and Behnke, 2009]



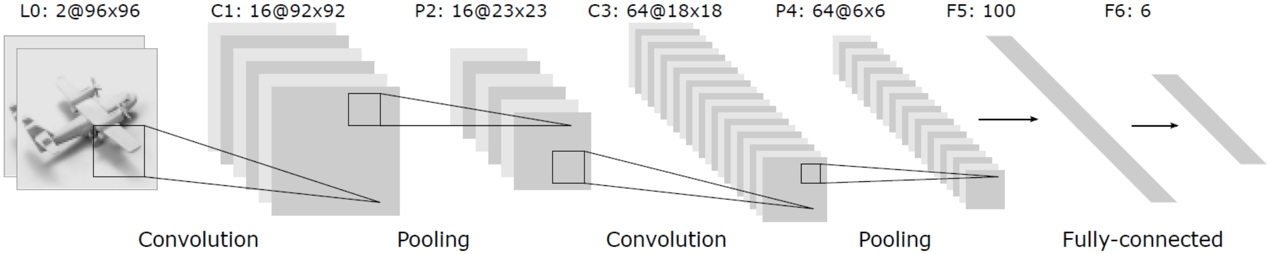
universitätbonn ais

Image Categorization: NORB

- 10 categories, jittered-cluttered



- **Max-pooling**, cross-entropy training on GPU

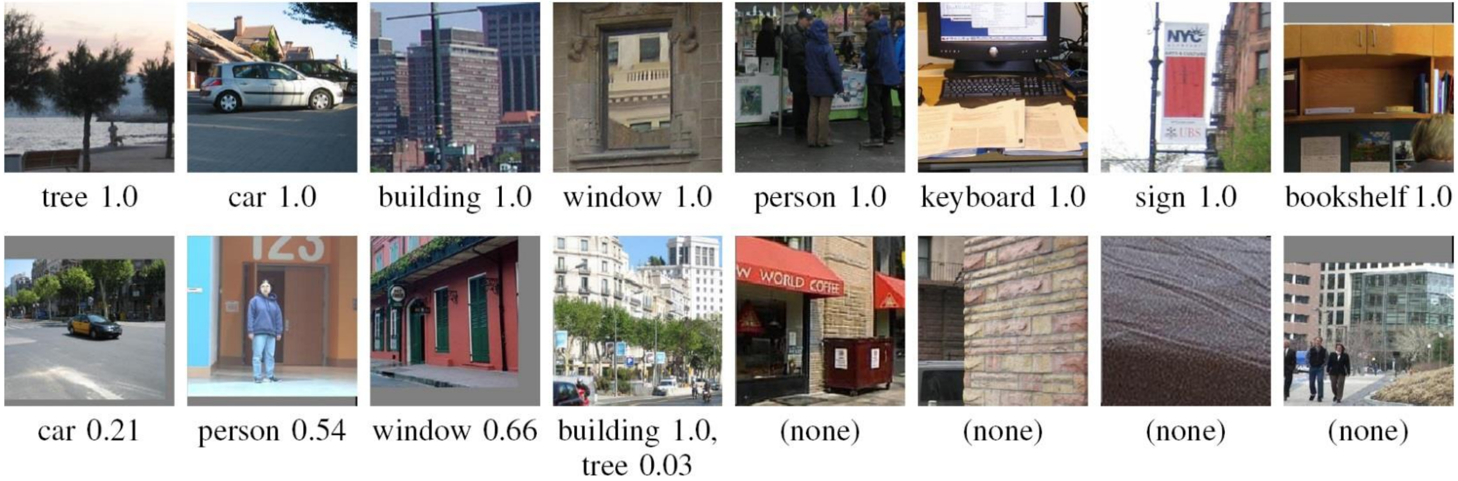


- Test error: 5.6% (LeNet7: 7.8%)

[Scherer, Müller, Behnke, ICANN2010]

Image Categorization: LabelMe

- 50,000 color images (256x256)
- 12 classes, objects scaled and centered + clutter (50%)



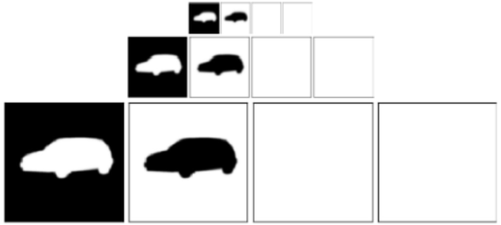
- Error TRN: 3.77%; TST: 16.27%
- Recall: 1,356 images/s

[Uetz, Behnke, ICIS2009]

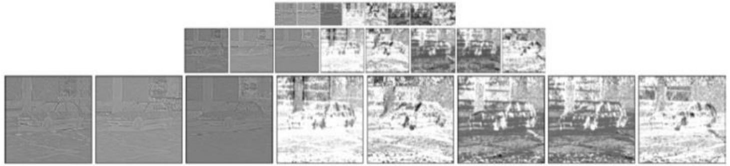
Object-class Segmentation

[Schulz, Behnke, ESANN 2012]

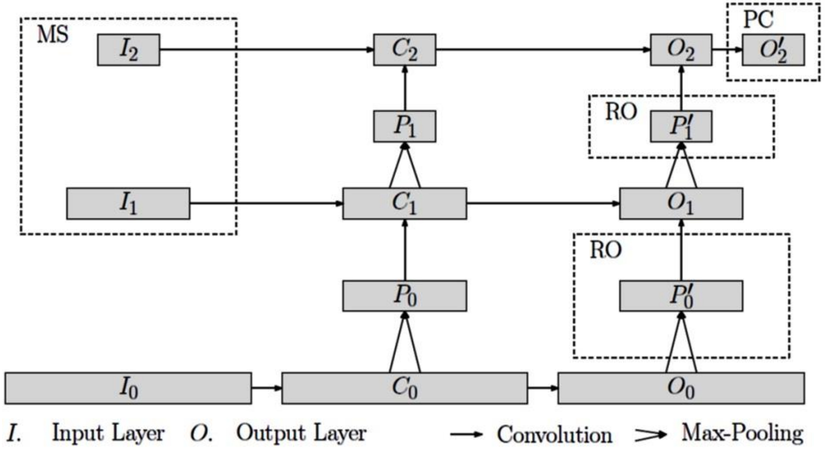
- Class annotation per pixel

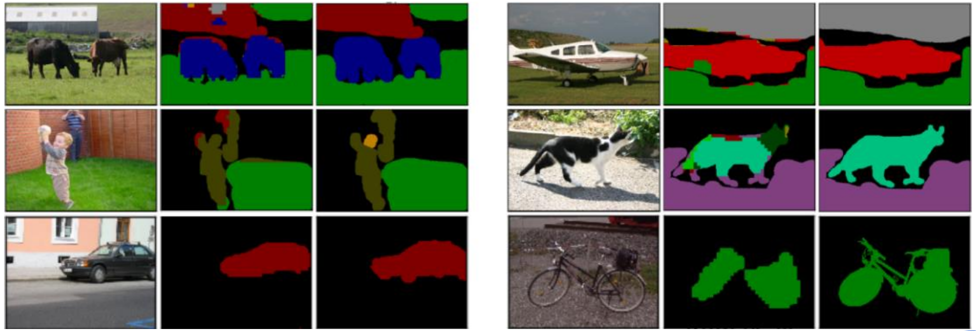


- Multi-scale input channels



- Evaluated on MSRC-9/21 and INRIA Graz-02 data sets





Input Output Truth

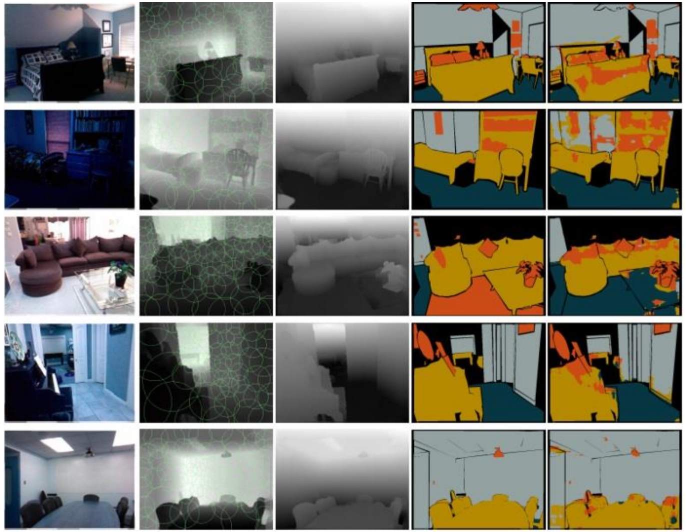
Object Detection in Natural Images

- Bounding box annotation
- Structured loss that directly maximizes overlap of the prediction with ground truth bounding boxes
- Evaluated on two of the Pascal VOC 2007 classes

[Schulz, Behnke, ICANN 2014]

RGB-D Object-Class Segmentation

- Kinect-like sensors provide dense depth (NYU Depth V2)
- Scale input according to depth, compute pixel hight



RGB Depth Height Truth Output

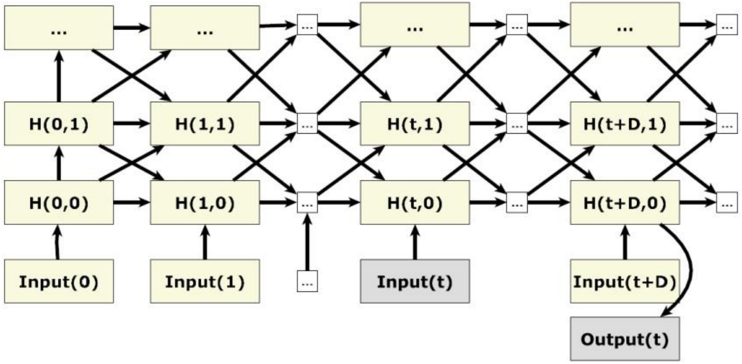
Method	floor	struct	furnit	prop	Class Avg.	Pixel Acc.
CW	84.6	70.3	58.7	52.9	66.6	65.4
CW+DN	87.7	70.8	57.0	53.6	67.3	65.5
CW+H	78.4	74.5	55.6	62.7	67.8	66.5
CW+DN+H	93.7	72.5	61.7	55.5	70.9	70.5
CW+DN+H+SP	91.8	74.1	59.4	63.4	72.2	71.9
CW+DN+H+CRF	93.5	80.2	66.4	54.9	**73.7**	**73.4**
Müller et al.[8]	94.9	78.9	71.1	42.7	71.9	72.3
Random Forest [8]	90.8	81.6	67.9	19.9	65.1	68.3
Couprie et al.[9]	87.3	86.1	45.3	35.5	63.6	64.5
Höft et al.[10]	77.9	65.4	55.9	49.9	62.3	62.0
Silberman [12]	68	59	70	42	59.7	58.6

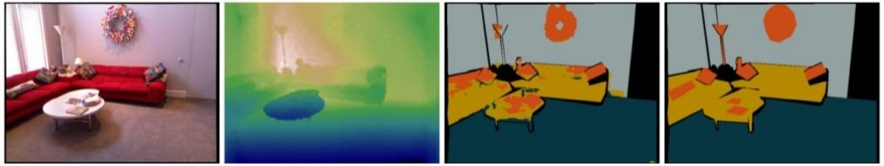
CW is covering windows, H is height above ground, DN is depth normalized patch sizes. SP is averaged within superpixels and SVM-reweighted. CRF is a conditional random field over superpixels [8]. Structure class numbers are optimized for class accuracy.

[Schulz, Höft, Behnke, ESANN 2015]

Neural Abstraction Pyramid for RGB-D Video Object-class Segmentation

- NYU Depth V2 contains RGB-D video sequences
- Recursive computation is efficient for temporal integration





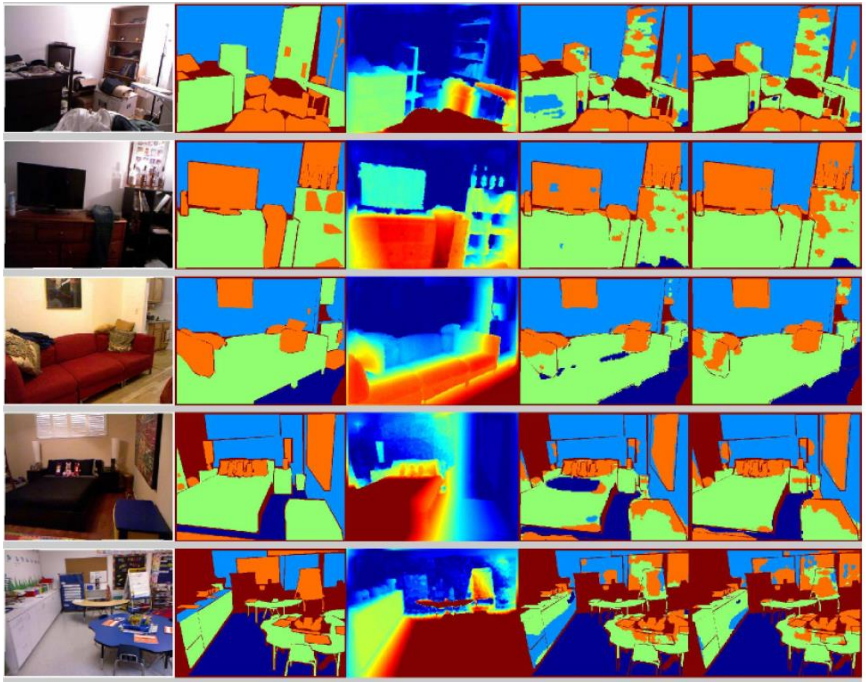
RGB Depth Output Truth

Method	Class Accuracies (%)				Average (%)	
	ground	struct	furnit	prop	Class	Pixel
Höft *et al.* [19]	77.9	65.4	55.9	49.9	62.0	61.1
Unidirectional + MS	73.4	66.8	**60.3**	49.2	62.4	63.1
Schulz *et al.* [20] (no height)	87.7	70.8	57.0	53.6	67.3	65.5
Unidirectional + SW	**90.0**	**76.3**	52.1	**61.2**	**69.9**	**67.5**

[Pavel, Schulz, Behnke, IJCNN 2015]

Geometric and Semantic Features for RGB-D Object-class Segmentation

- New **geometric** feature: distance from wall
- **Semantic** features pretrained from ImageNet
- Both help significantly

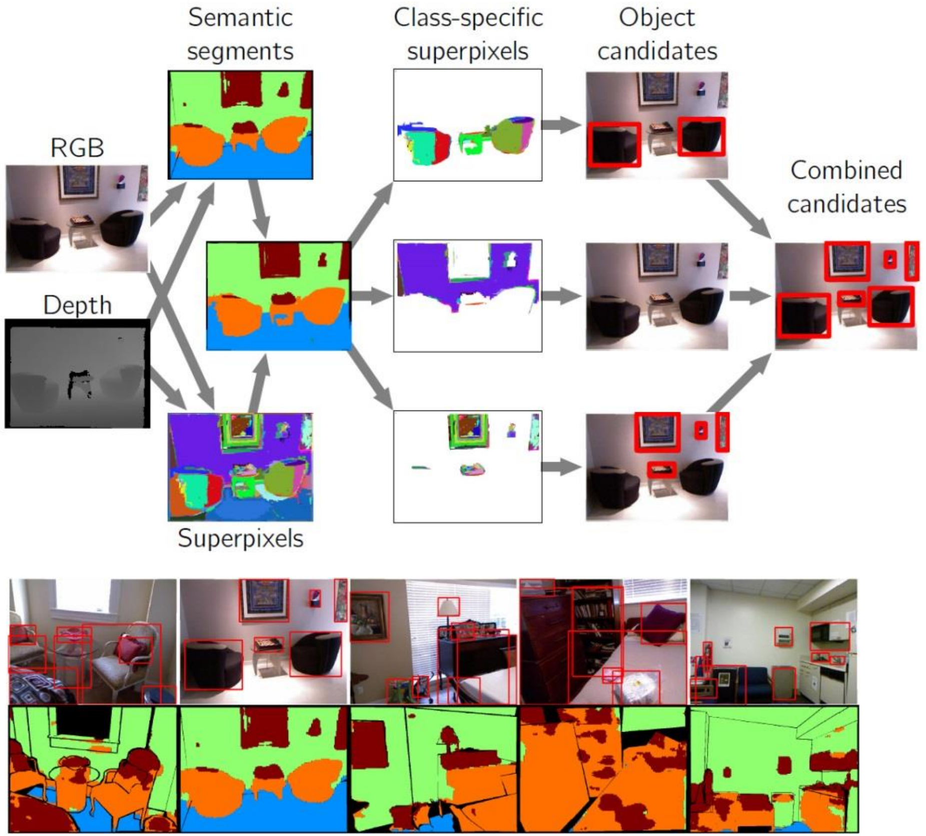


[Husain et al. under review]

Semantic Segmentation Priors for Object Discovery

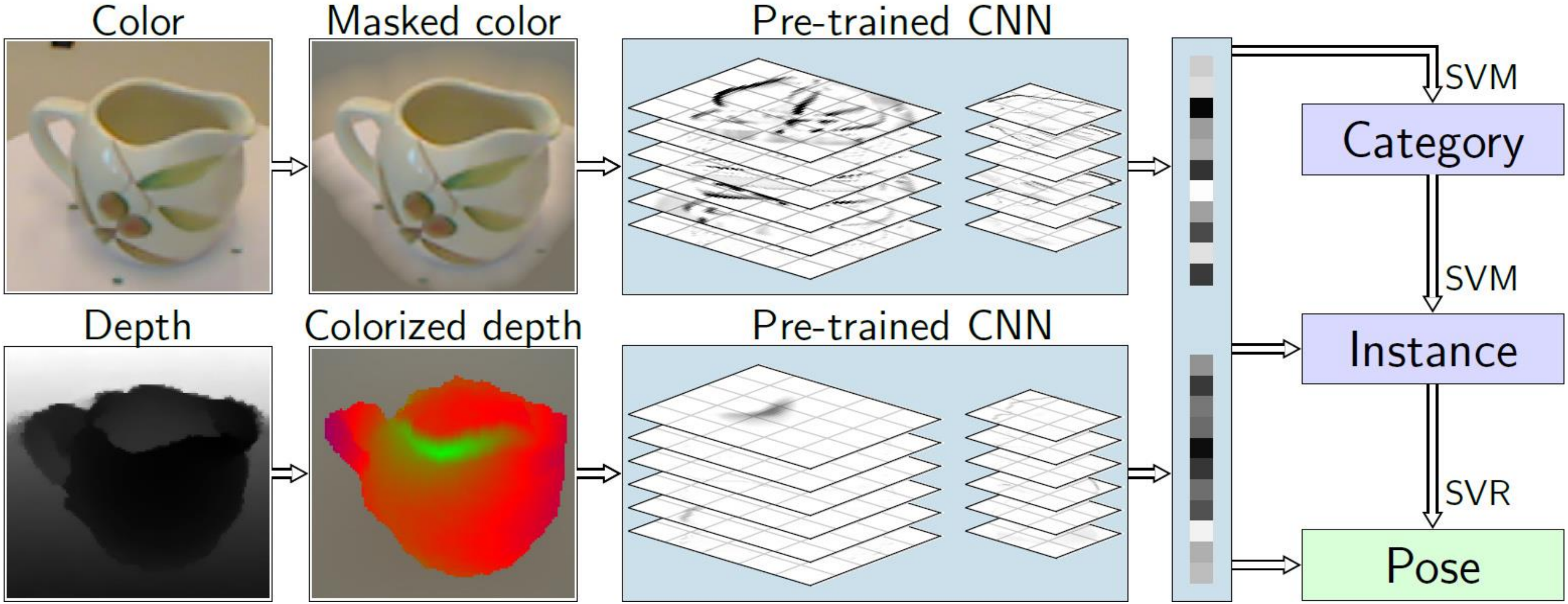
- Combine bottom-up object discovery and semantic priors
- Semantic segmentation used to classify color and depth superpixels
- Higher recall, more precise object borders

[Garcia et al. under review]



universitätbonn ais

RGB-D Object Recognition and Pose Estimation

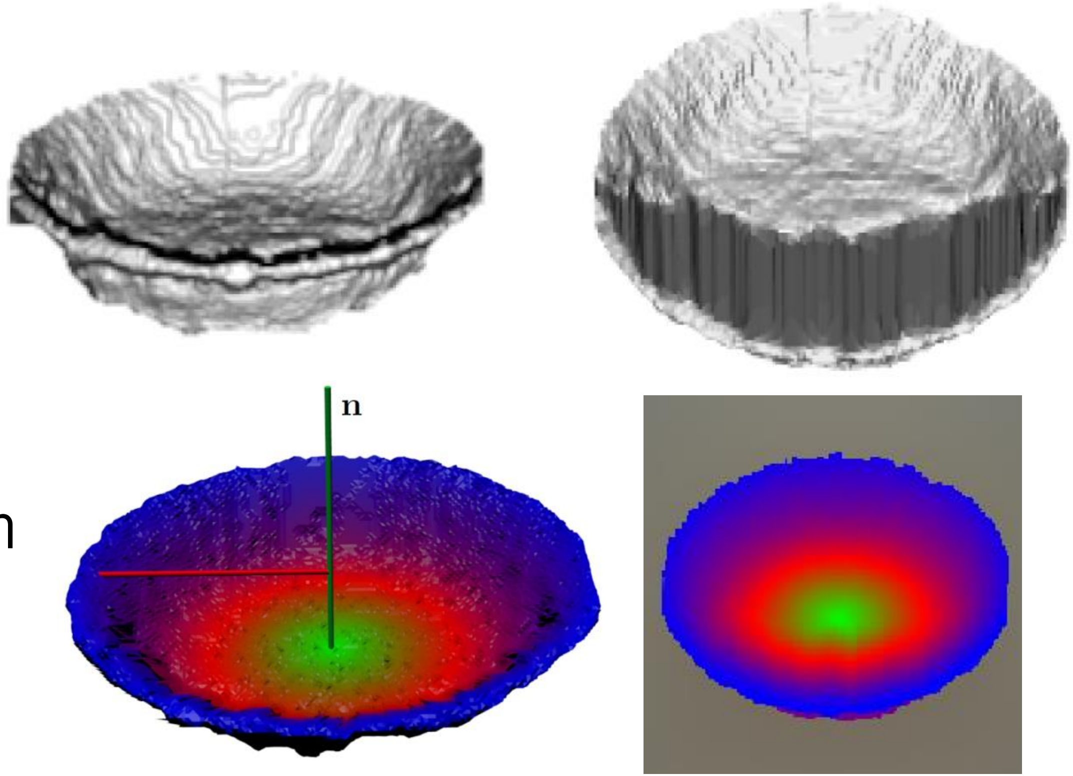


[Schwarz, Schulz, Behnke, ICRA2015]

Canonical View, Colorization

- Objects viewed from different elevation
- Render canonical view

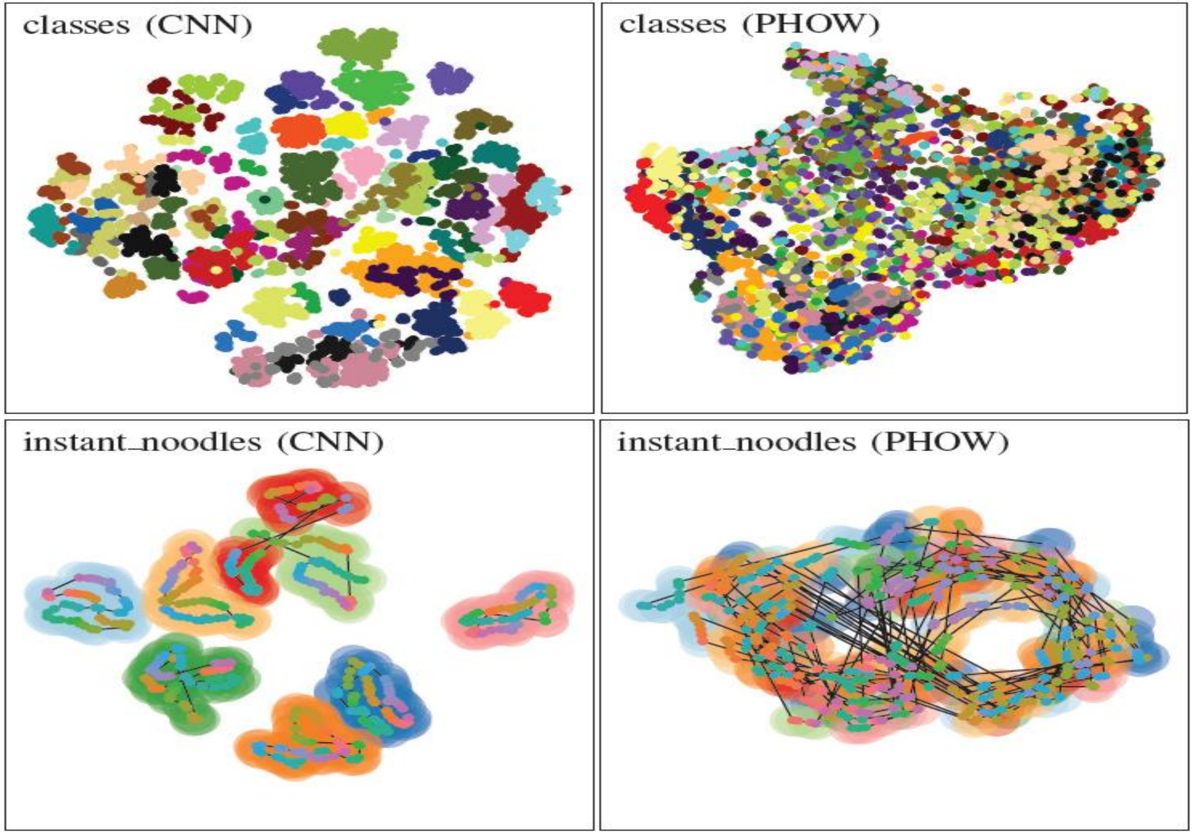
- Colorization based on distance from center vertical



[Schwarz, Schulz, Behnke, ICRA2015]

Pretrained Features Disentangle Data

- t-SNE embedding



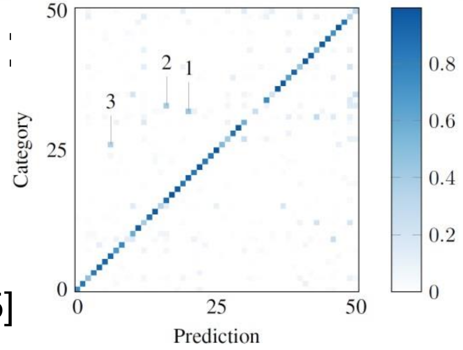
[Schwarz, Schulz, Behnke ICRA2015]

Recognition Accuracy

- Improved both category and instance recognition

Method	Category Accuracy (%) RGB	Category Accuracy (%) RGB-D	Instance Accuracy (%) RGB	Instance Accuracy (%) RGB-D
Lai *et al.* [1]	74.3 ± 3.3	81.9 ± 2.8	59.3	73.9
Bo *et al.* [2]	82.4 ± 3.1	87.5 ± 2.9	**92.1**	92.8
PHOW[3]	80.2 ± 1.8	—	62.8	—
Ours	$\mathbf{83.1 \pm 2.0}$	88.3 ± 1.5	92.0	**94.1**
Ours	$\mathbf{83.1 \pm 2.0}$	$\mathbf{89.4 \pm 1.3}$	92.0	**94.1**

- Confusion:



1: pitcher / coffe mug

2: peach / sponge

[Schwarz, Schulz, Behnke, ICRA2015]

Conclusions

- Deep learning has a long history
  - Recurrence (weight sharing over time) might be as useful as convolutional processing (spatial weight sharing)
  - In Neural Abstraction Pyramid, top-down and lateral connections are an efficient way to incorporate context for resolving local ambiguities
- Depth and geometric features help with scene segmentation and semantic interpretation
- Transfer learning from pretrained features helps a lot

Questions?