Spatiotemporal Integration in Recurrent Deep Neural Networks

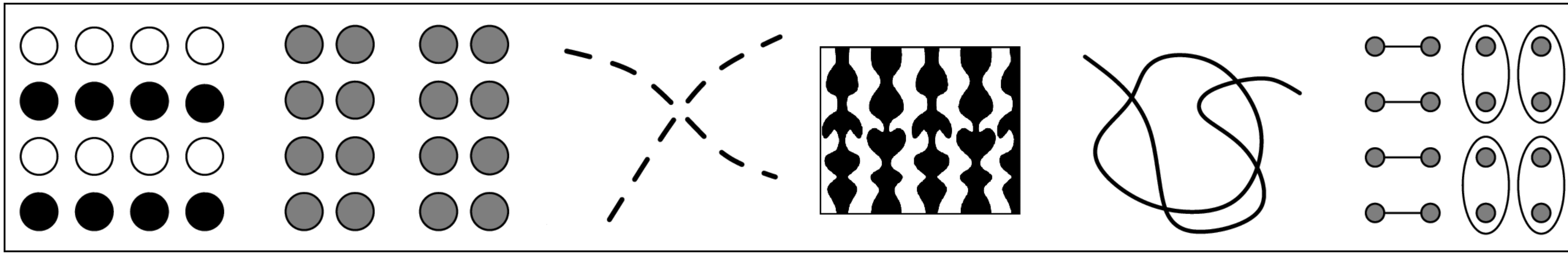
Sven Behnke

University of Bonn
Computer Science Institute VI
Autonomous Intelligent Systems

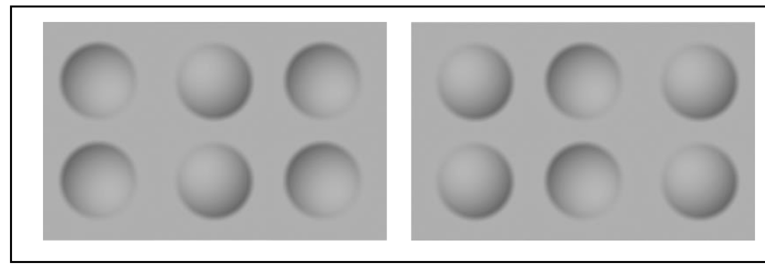
Performance of the Human Visual System

Psychophysics

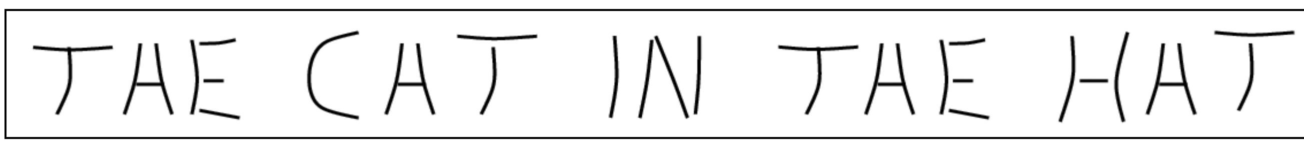
- Gestalt principles



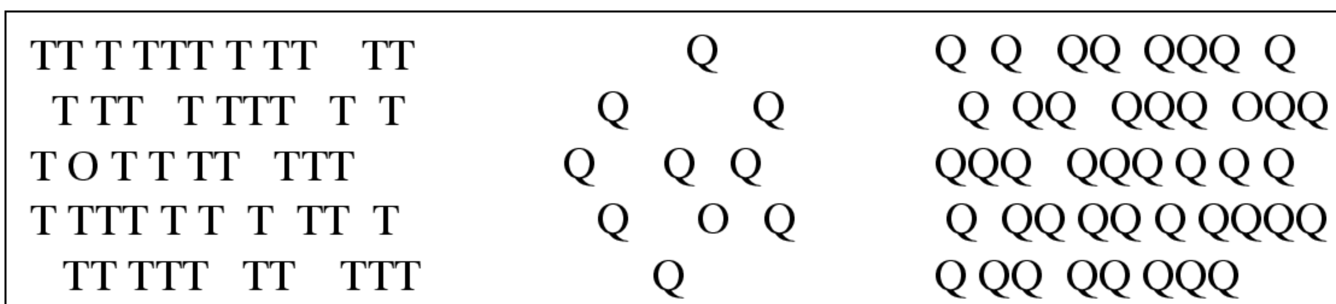
- Heuristics



- Context

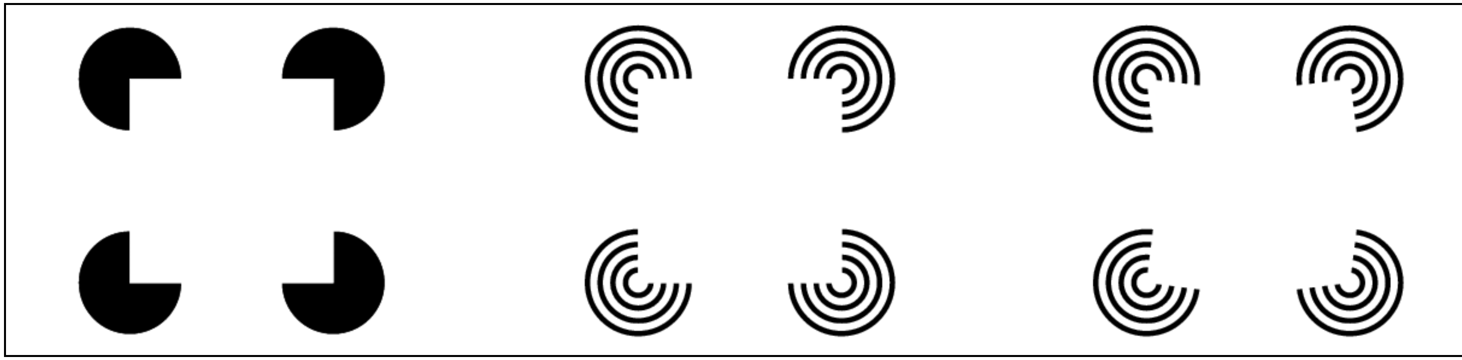


- Attention



Visual Illusions

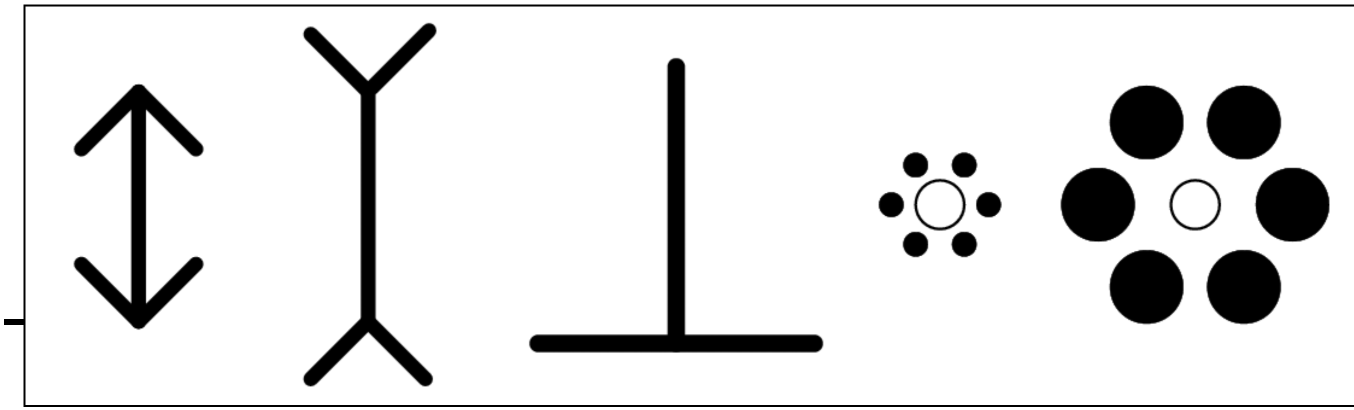
Kanizsa Figures



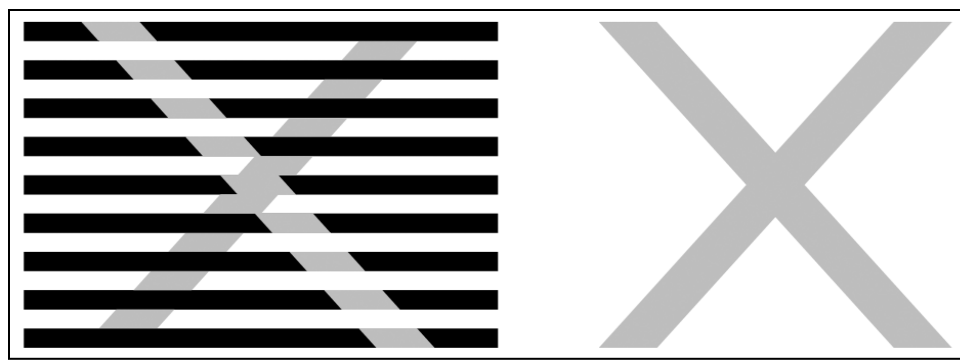
Müller-Lyer

horizontal/vertical

Ebbinghaus-Titchener

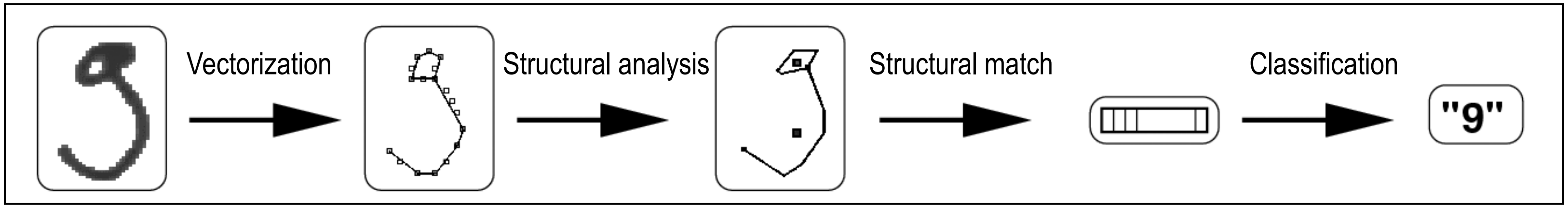


Munker-White

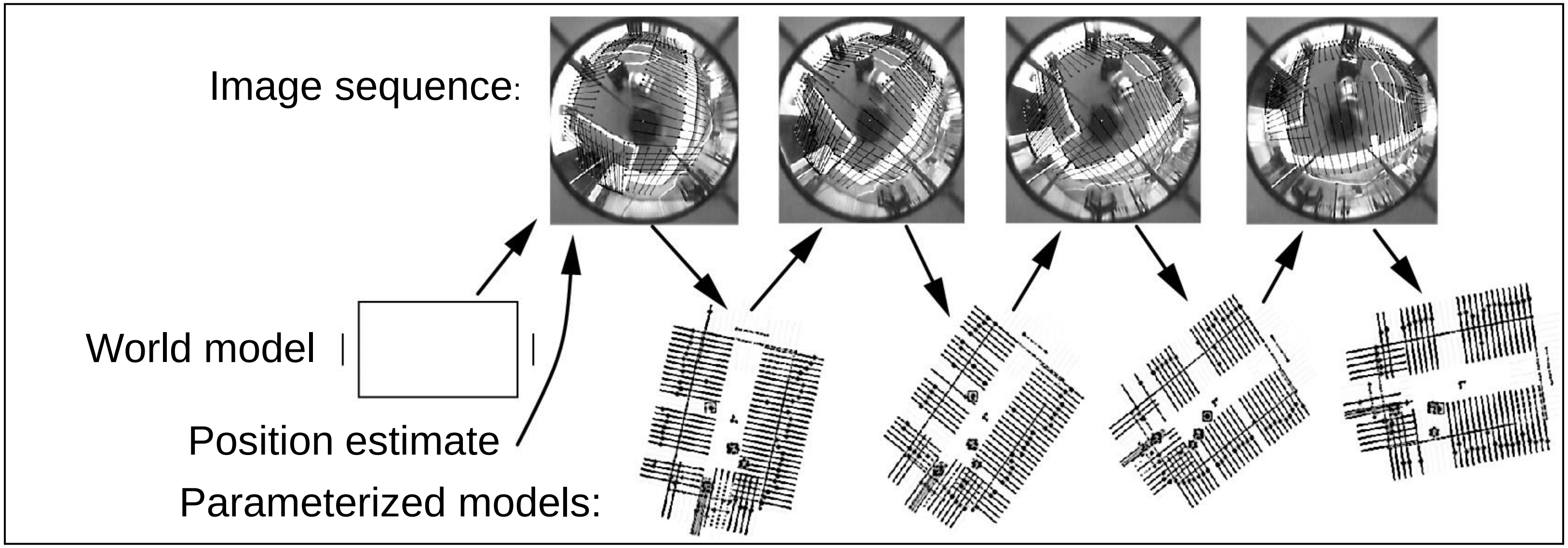


Computer Vision

- Data driven



- Model driven



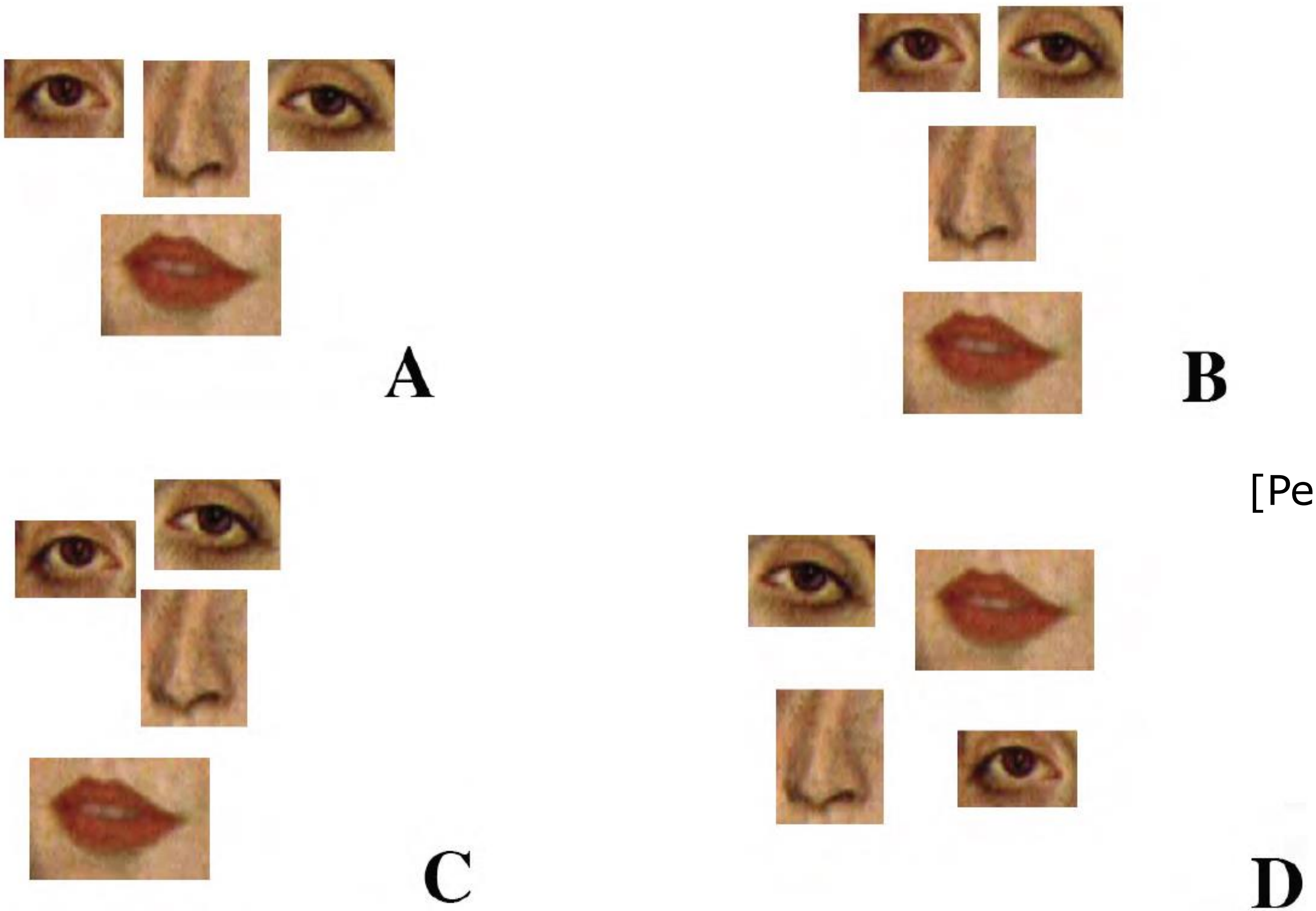
- Interface problem

Observations

In the world around us it mostly holds that:

- Neighboring things have something to do with each other
  - Spatially
  - Temporally
- There is hierarchical structure
  - Objects consist of parts
  - Parts are composed of components, …

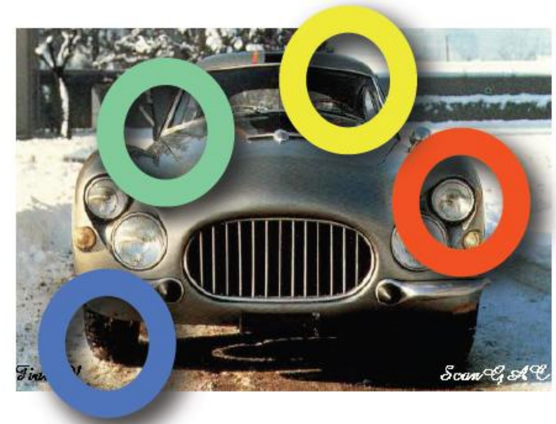
Spatial Arrangement of Facial Parts



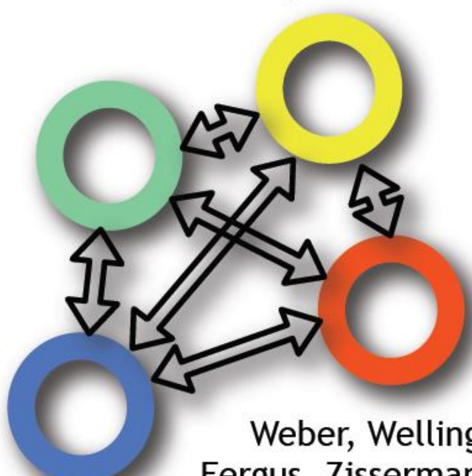
[Perona]

Face Perception

Horizontal and Vertical Dependencies

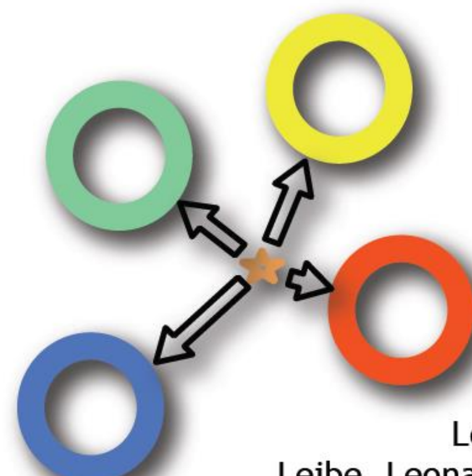


Constellation Model:
Fully connected shape model



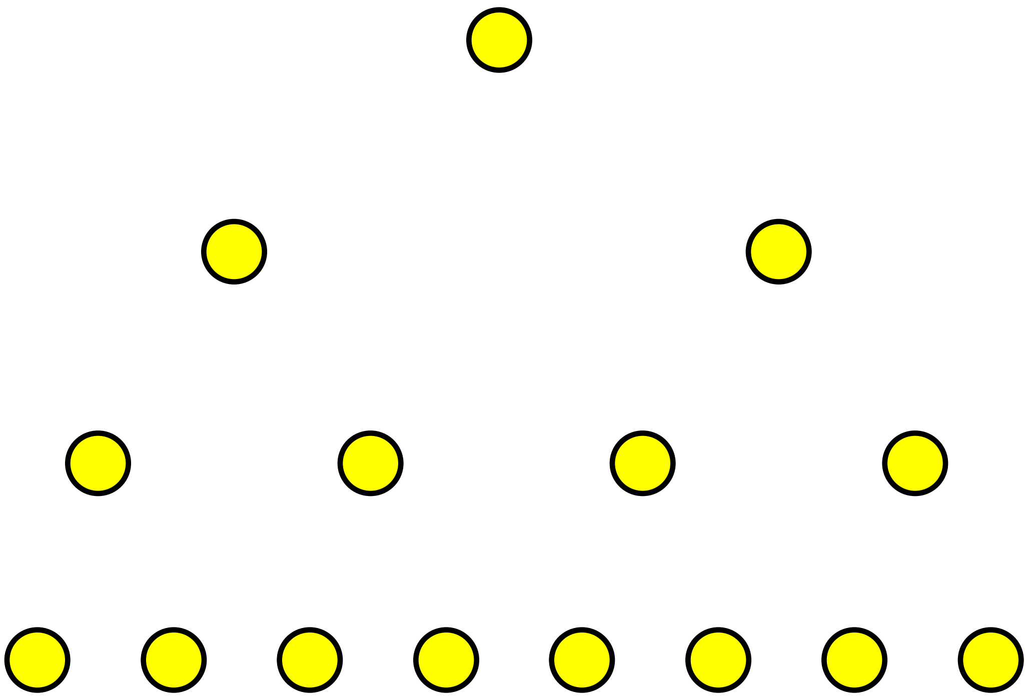
Weber, Welling, Perona ’00
Fergus, Zisserman, Perona ’03

Implicit Shape Model:
Star-Model w.r.t. Reference Point



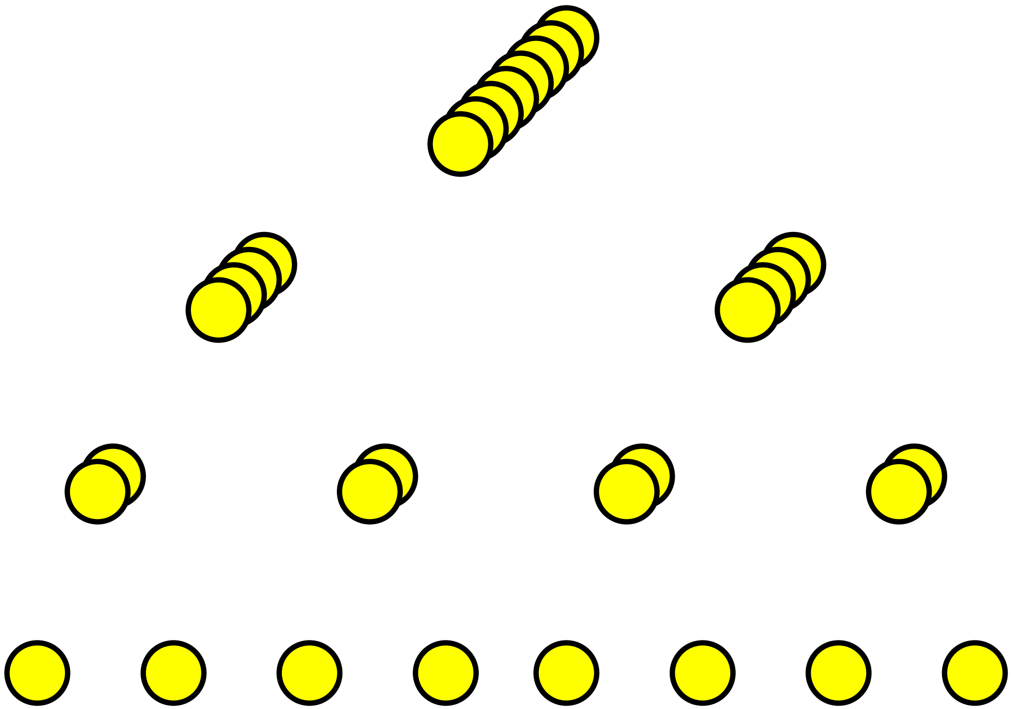
Leibe, Schiele ’03
Leibe, Leonardis, Schiele ’04

Multi-Scale Representation



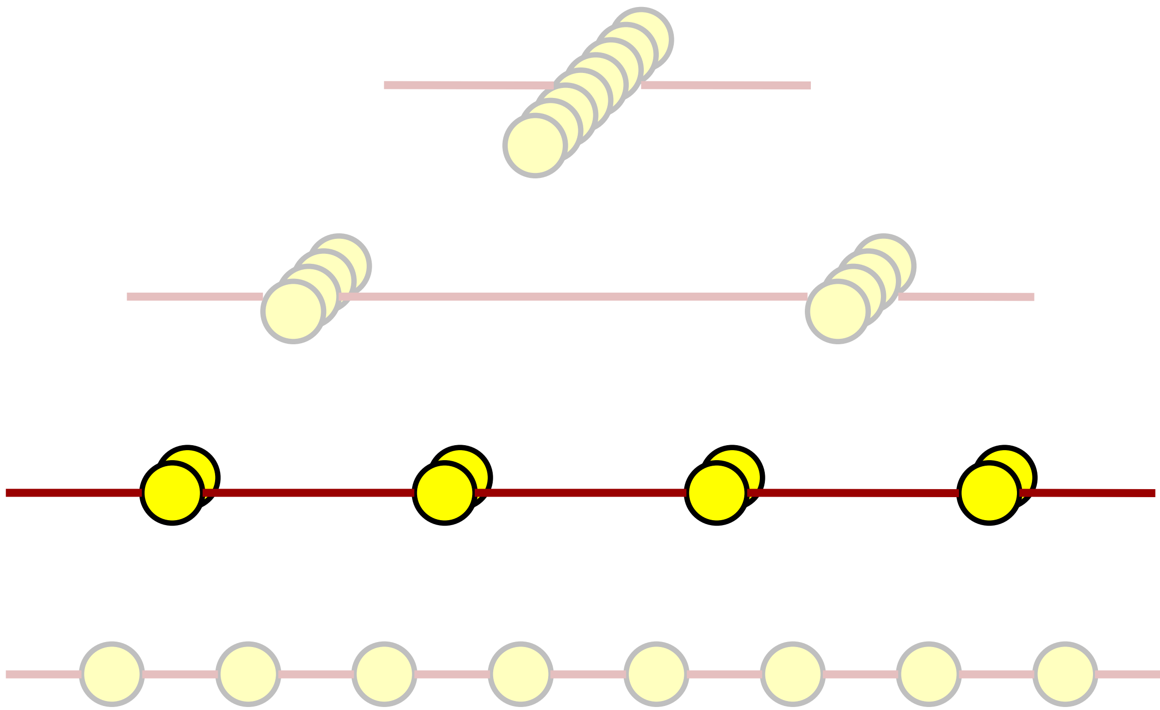
- Image pyramids are not expressive enough

Increasing Number of Features with Decreasing Resolution



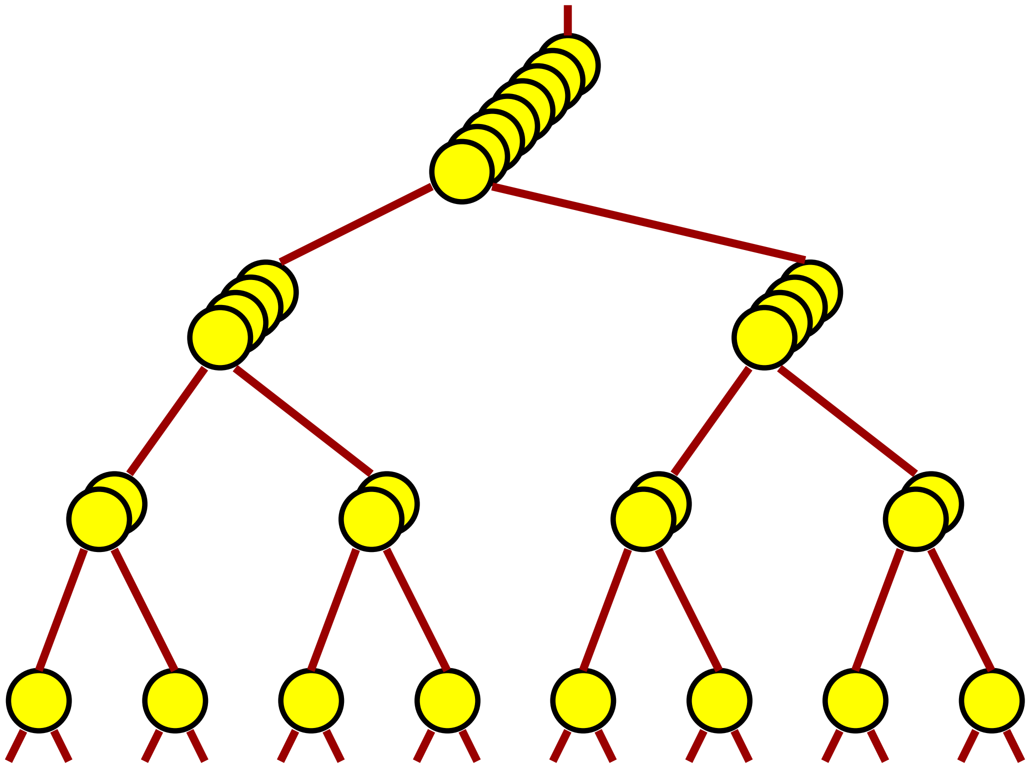
- Rich representations also in the higher layers

Modeling Horizontal Dependencies



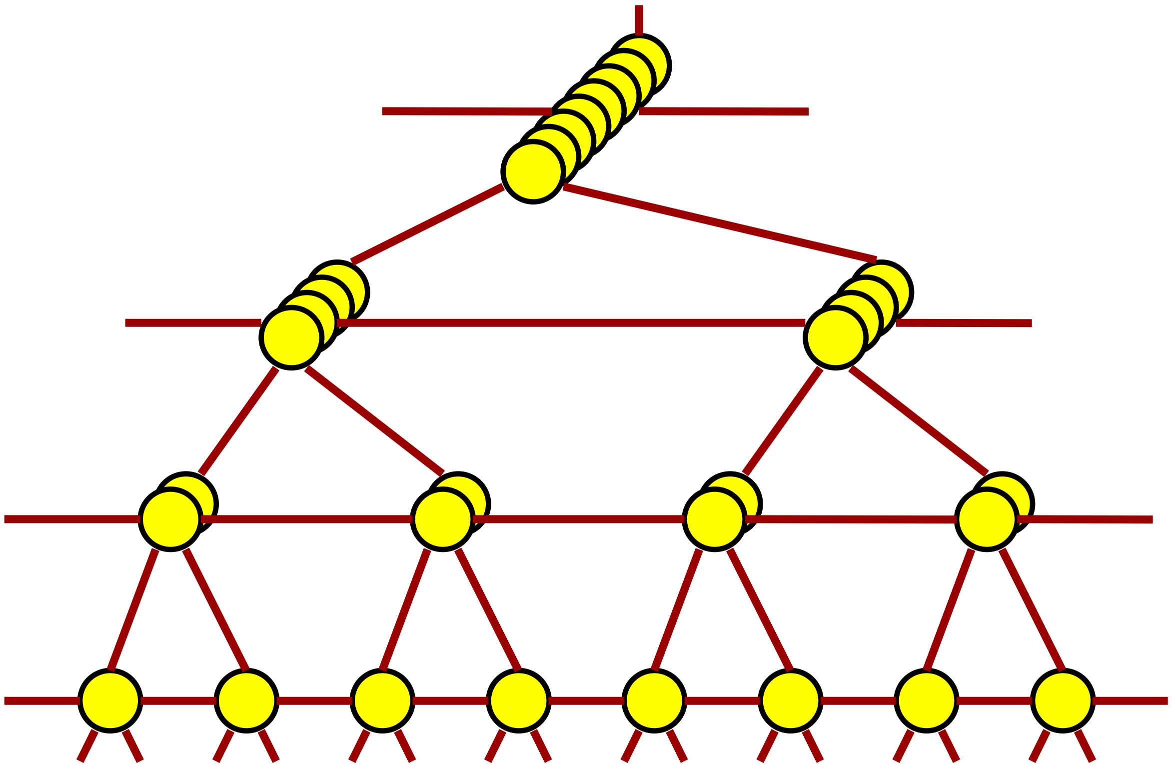
- 1D: HMM, Kalman Filter, Particle Filter
- 2D: Markov Random Fields
- Decision for level of description problematic
- Ignores vertical dependencies, flat models do not scale

Modeling Vertical Dependencies



- Structure graphs, etc.
- Ignores horizontal dependencies

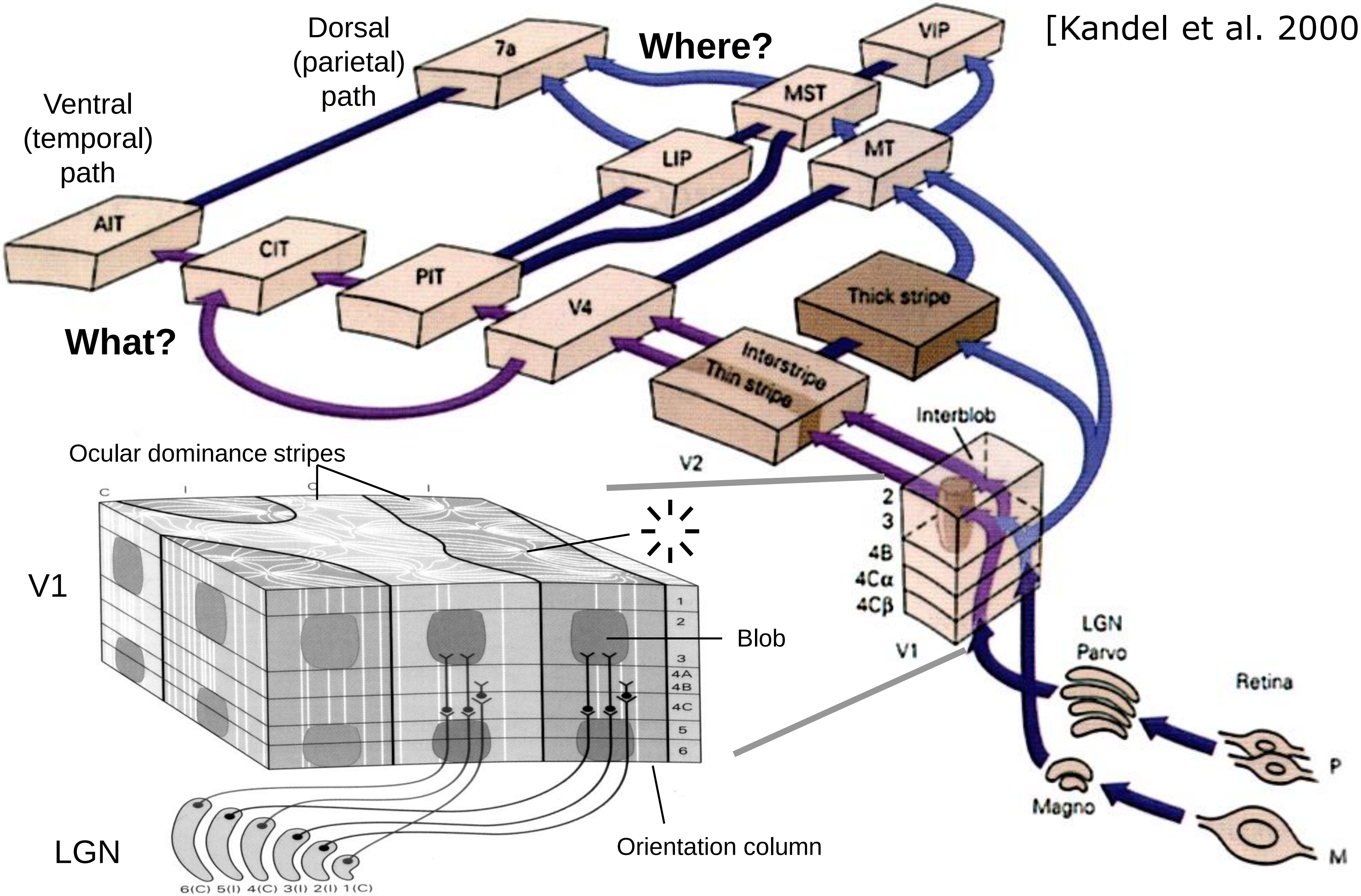
Horizontal and vertical Dependencies



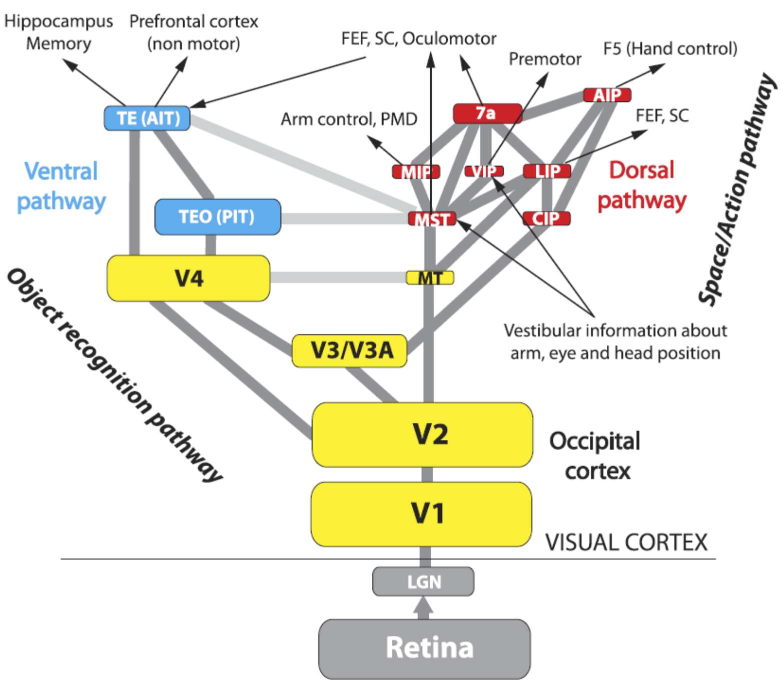
- Problem: Cycles make exact inference impossible
- Idea: Use approximate inference

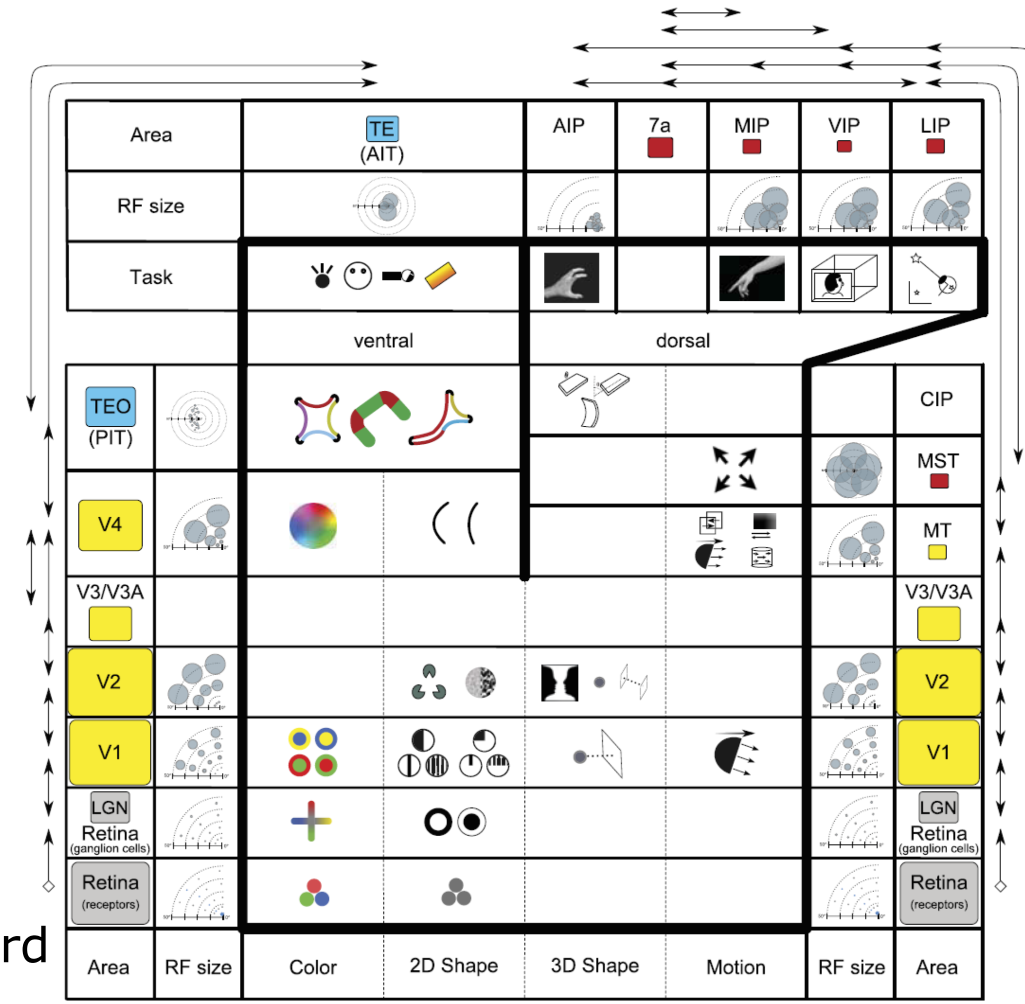
Human Visual System

[Kandel et al. 2000]



Visual Processing Hierarchy





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

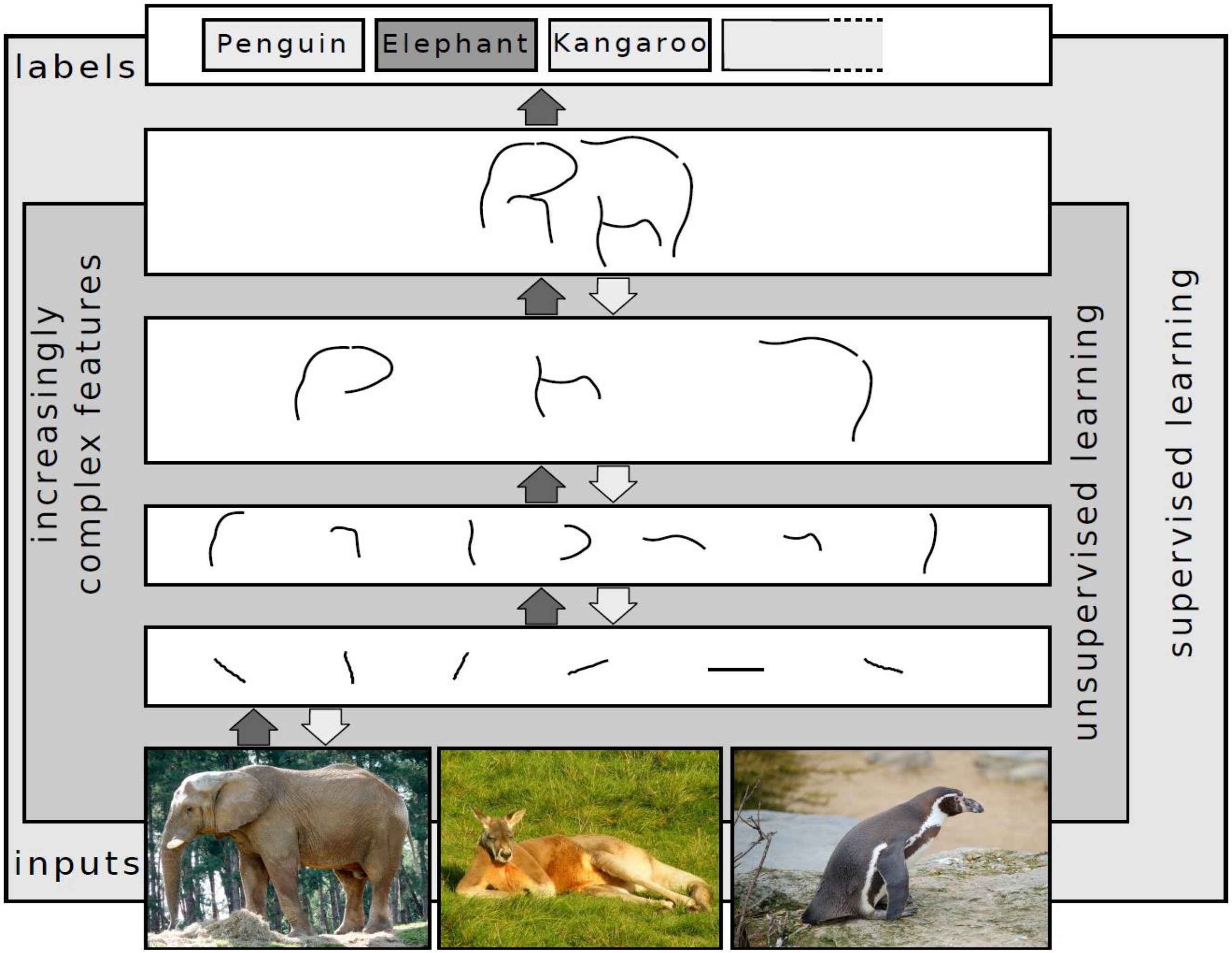
[Krüger et al., TPAMI 2013]

Deep Learning Definition

- Deep learning is a set of algorithms in machine learning that attempt to **learn layered models of inputs**, commonly neural networks.
- The layers in such models correspond to **distinct levels of concepts**, where
  - higher-level concepts are defined from lower-level ones, and
  - the same lower-level concepts can help to define many higher-level concepts.

[Bengio 2009]

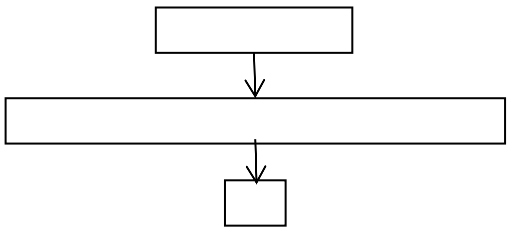
Layered Representations



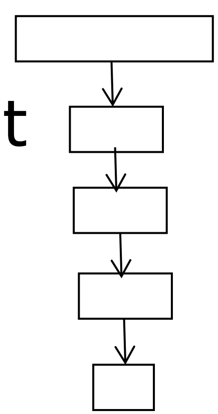
[Schulz and Behnke, KI 2012]

Flat vs. Deep Networks

- A neural network with a **single hidden layer** that is wide enough can compute any function (Cybenko, 1989)
  - Certain functions, like parity, may require exponentially many hidden units (in the number of inputs)
  - Compare to conjunctive / disjunctive normal form of Boolean function

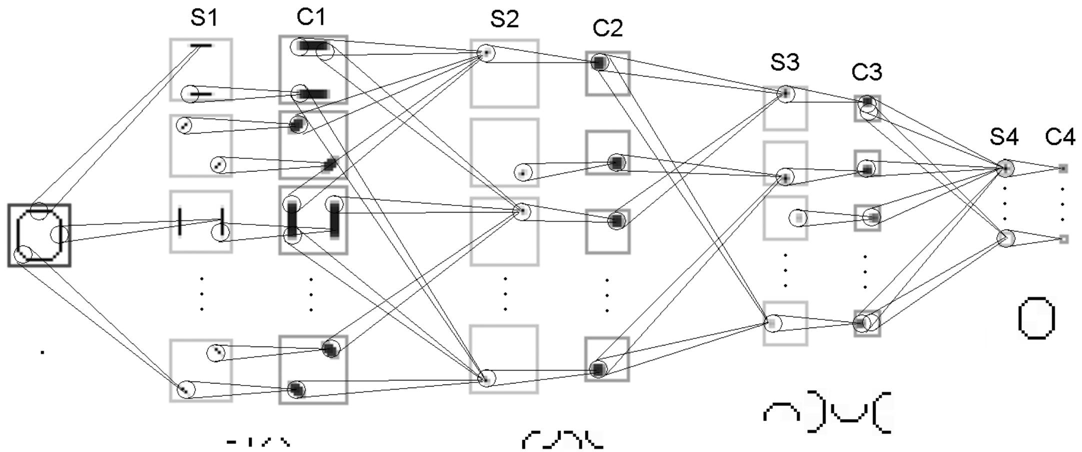


- **Deep networks** (with multiple hidden layers) may be exponentially more efficient
  - Parity example:
    - As many hidden layers as inputs
    - Compute carry bit sequentially

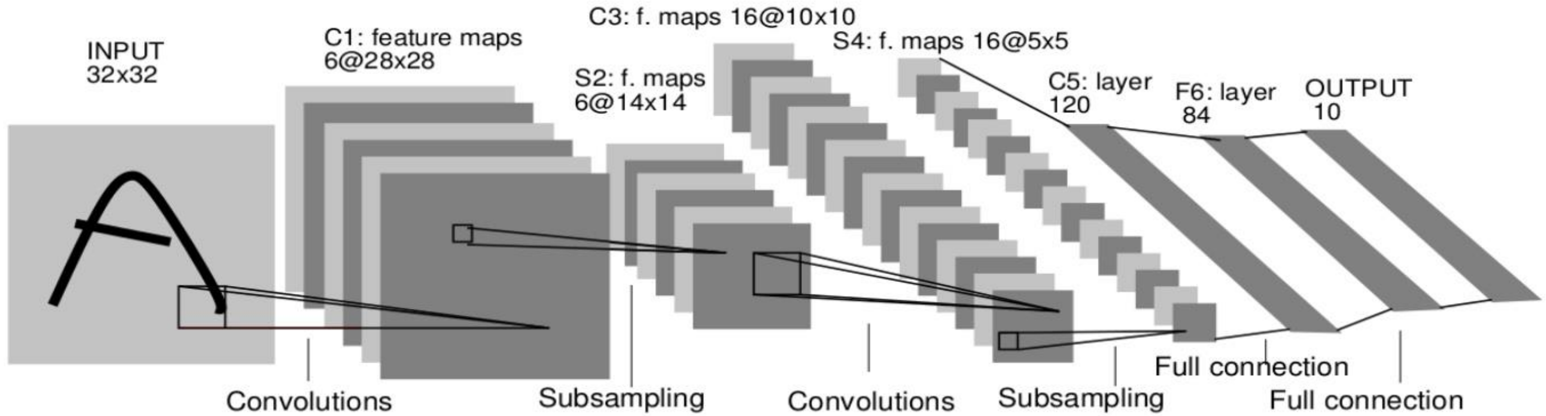


Convolutional Models

- Neocognitron: Fukushima 1980

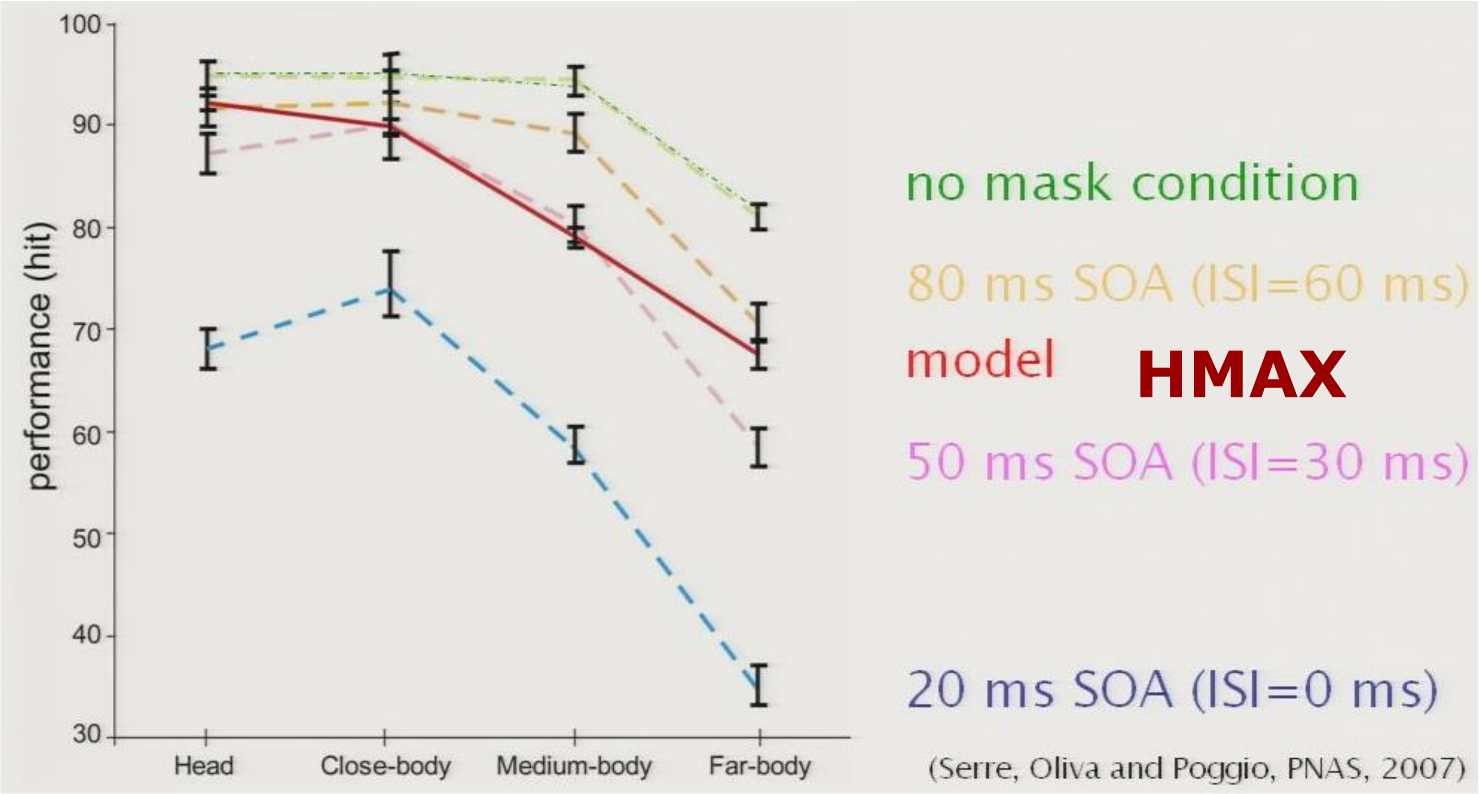


- Supervised training of convolutional networks: LeCun 1989



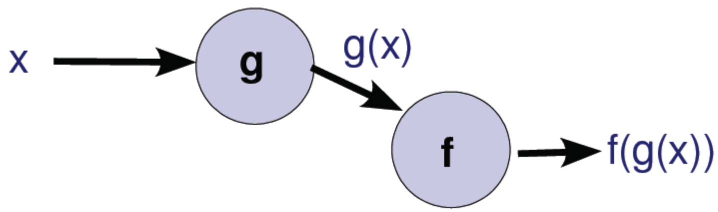
Feed-forward Models Cannot Explain Human Performance

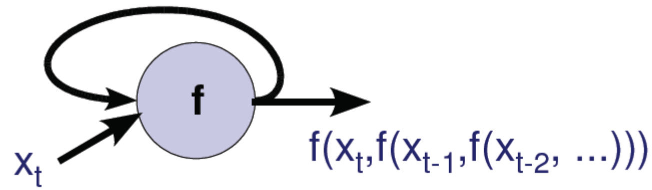
- Performance increases with observation time



Feed Forward

vs. Recurrent



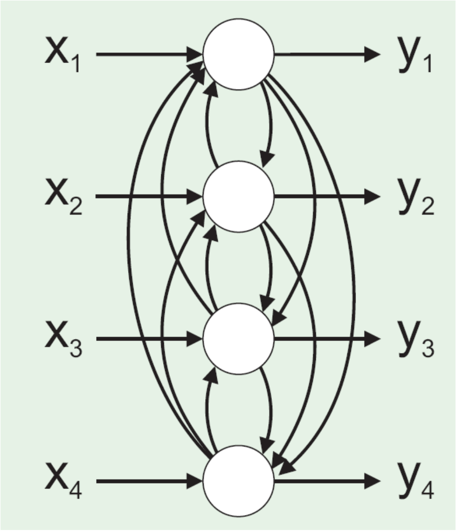


- Connectivity without cycles
- Composition of simple functions
- A node can only by computed if its inputs are available
- Reuse of partial results
- Order of computation determined by directed connectivity

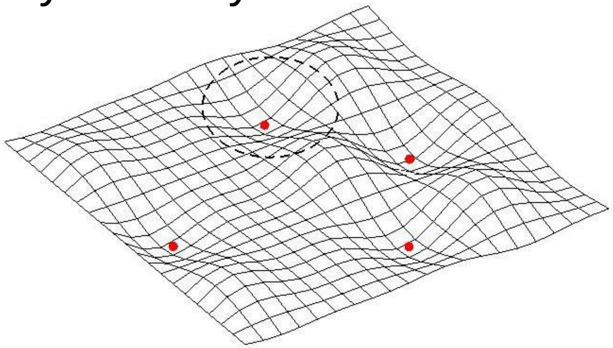
- Connectivity with cycles
- Explicit modeling of computation time necessary
- Computation needs one unit of time
- Input at time t yields output at time t+1
- Order of computation not any longer determined by connectivity

Hopfield-Networks

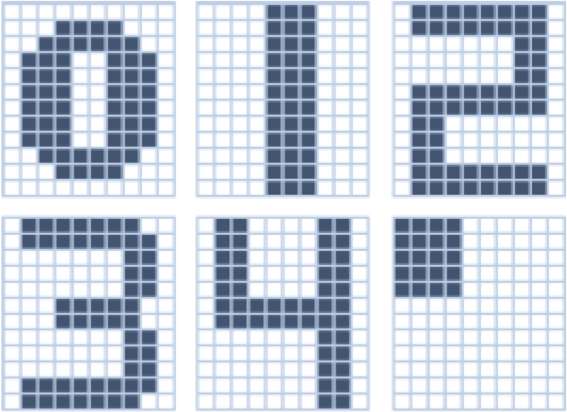
[Hopfield 1982]



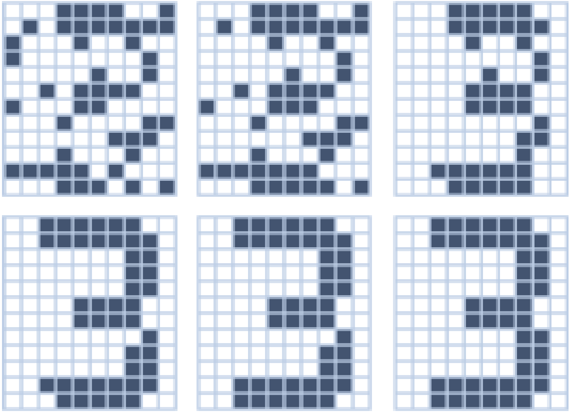
- Fully connected binary units
- Symmetric weights
- Non-linear dynamic system minimizes energy



Stored patterns



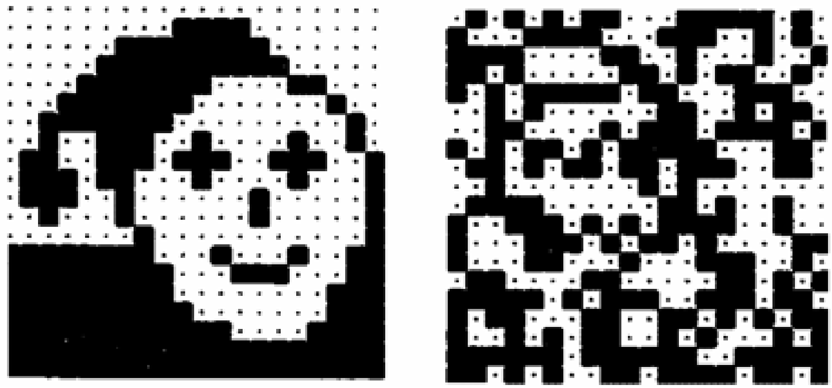
Iterative denoising



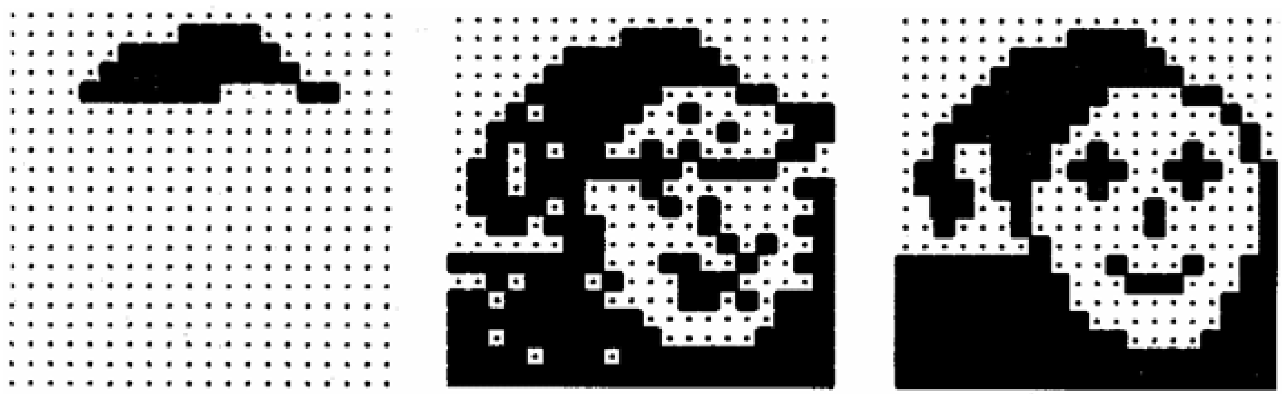
Completion of Patterns

- Associative memory

Stored patterns



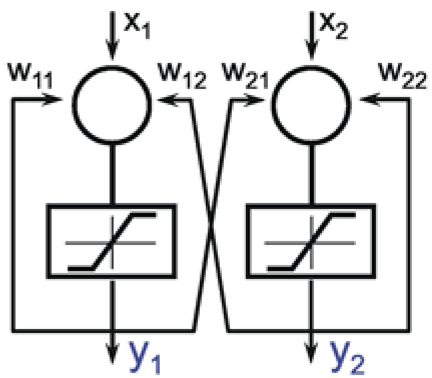
Pattern completion

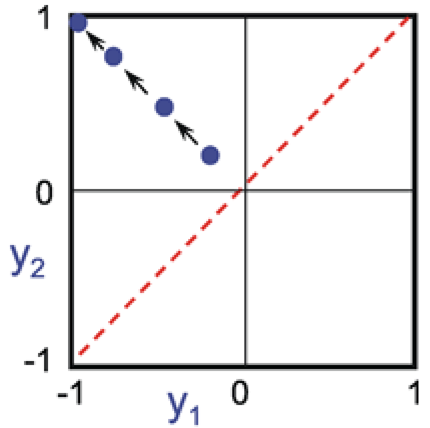


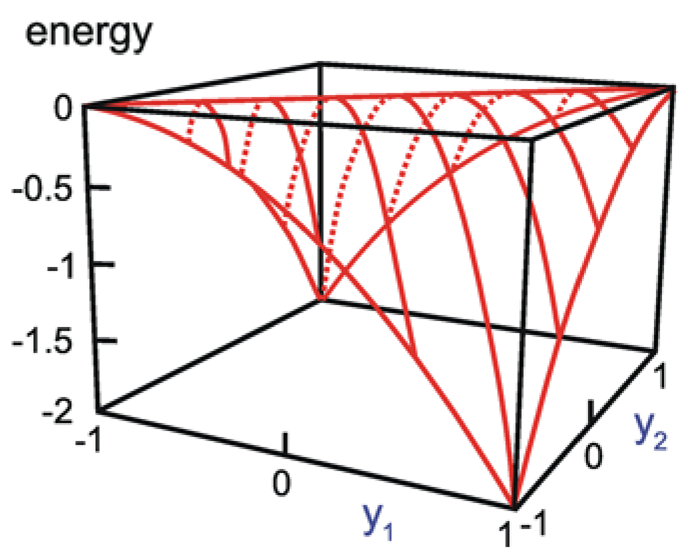
Hopfield-Networks with Continous Activation

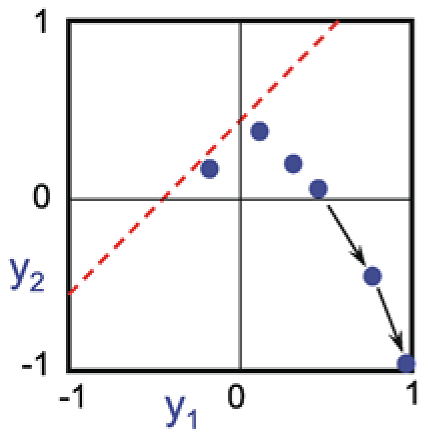
$w_{21} = w_{12} = -1$

$w_{11} = w_{22} = 1$







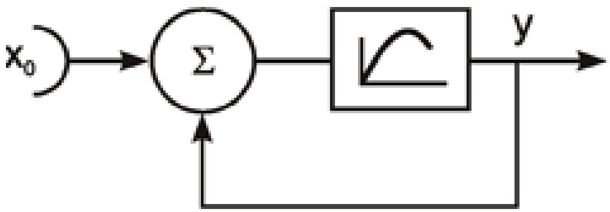


$\theta_1 = 0.5$

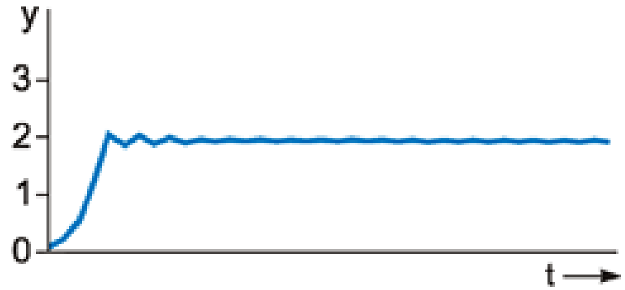
$\theta_2 = 0$

Oszillation and Chaos

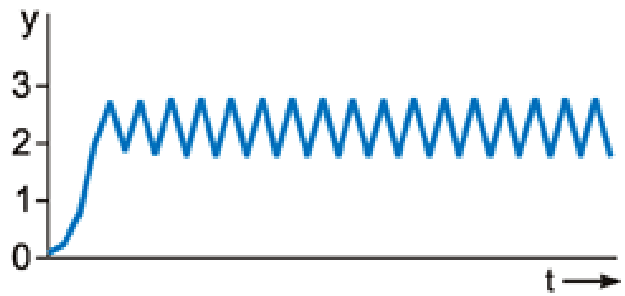
$$y = a_1 x - a_2\, x^2$$



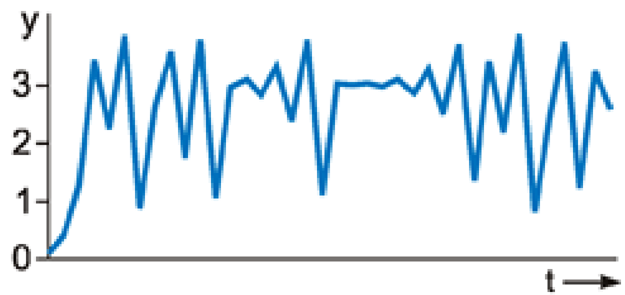
$x_0 = 0.1$



$a_2 = 2.5$



$a_2 = 3.2$

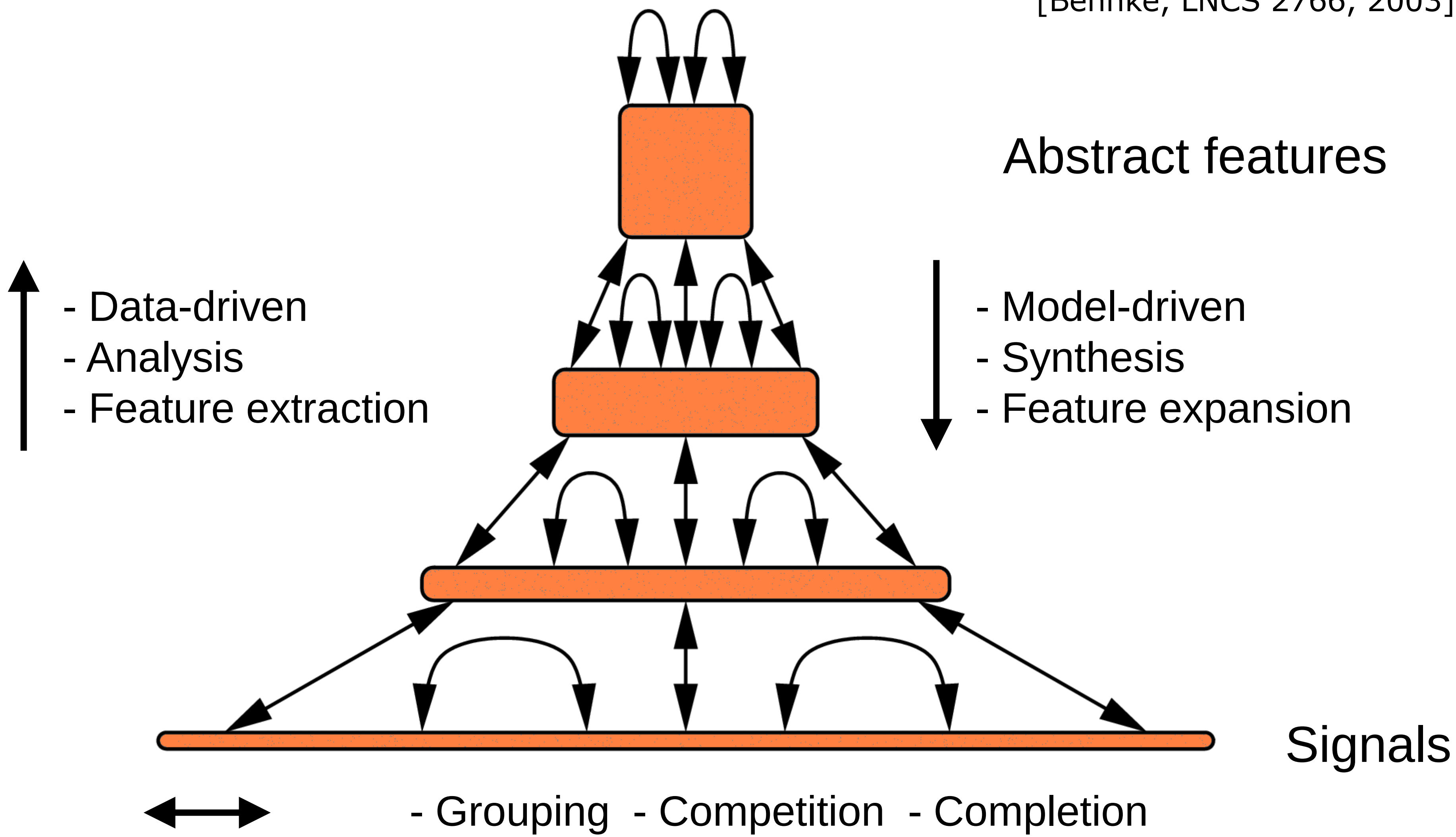


$a_2 = 3.8$

- Behavior depends on choice of parameters

Neural Abstraction Pyramid

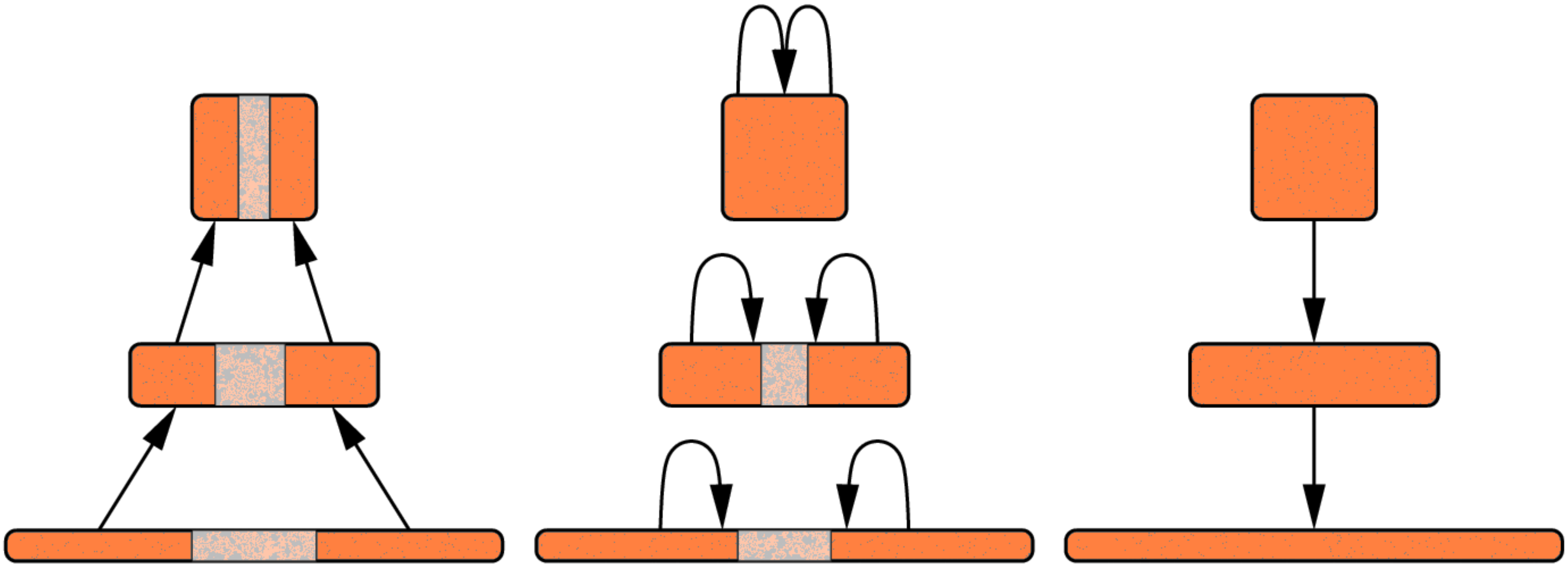
[Behnke, LNCS 2766, 2003]



Iterative Interpretation

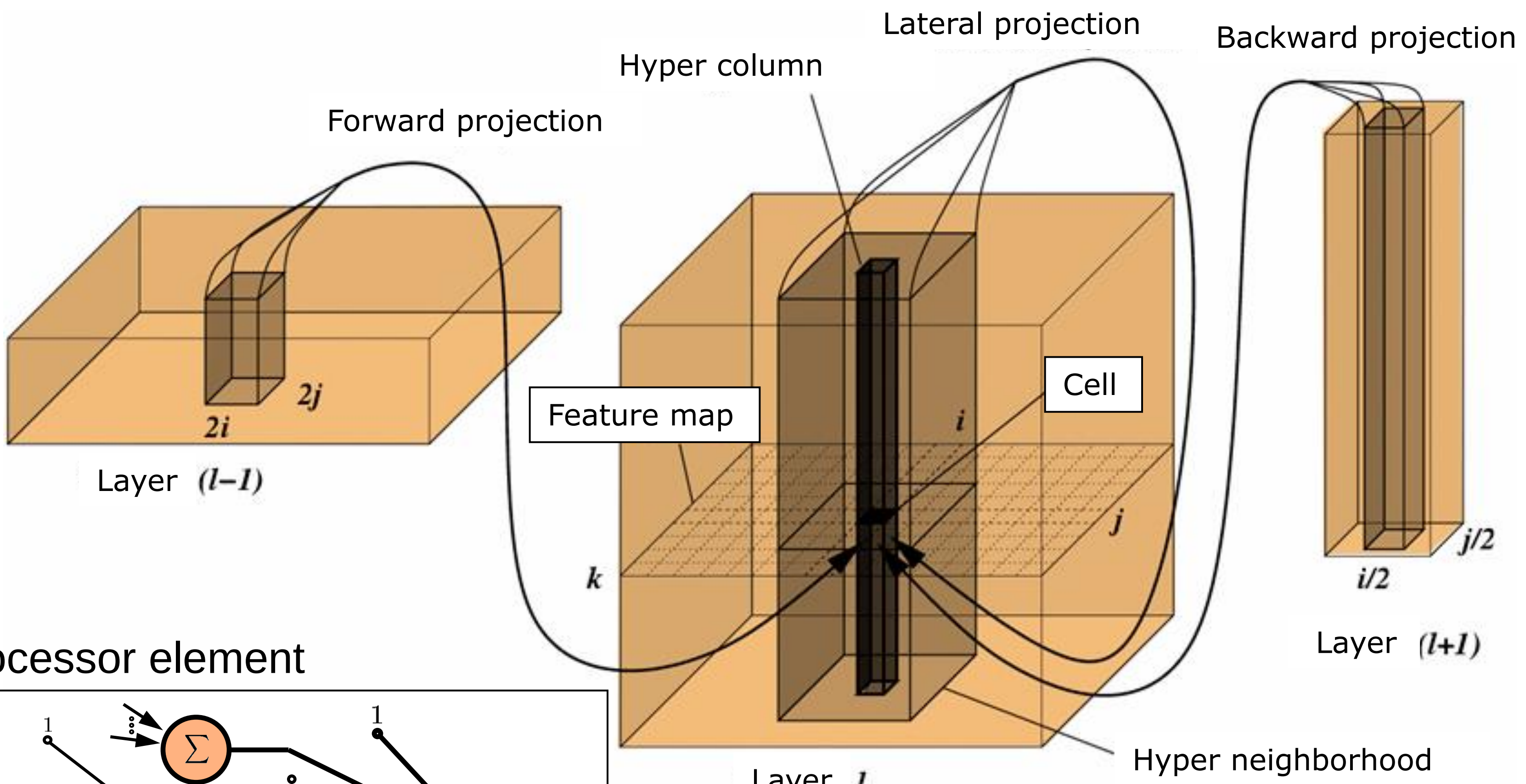
[Behnke, LNCS 2766, 2003]

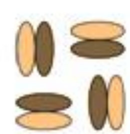
- Interpret most obvious parts first



- Use partial interpretation as context to resolve local ambiguities

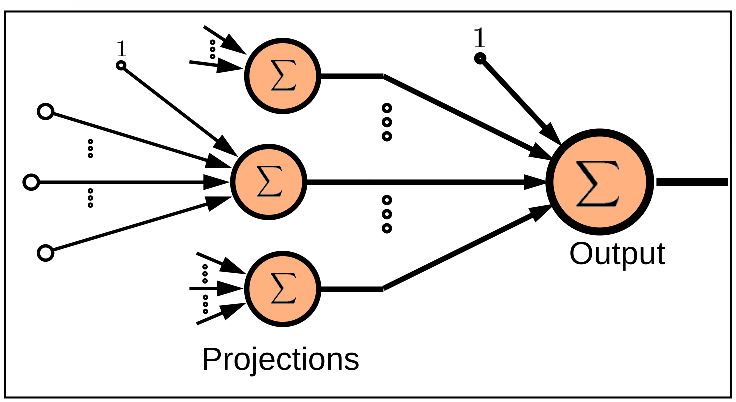
Local Recurrent Connectivity





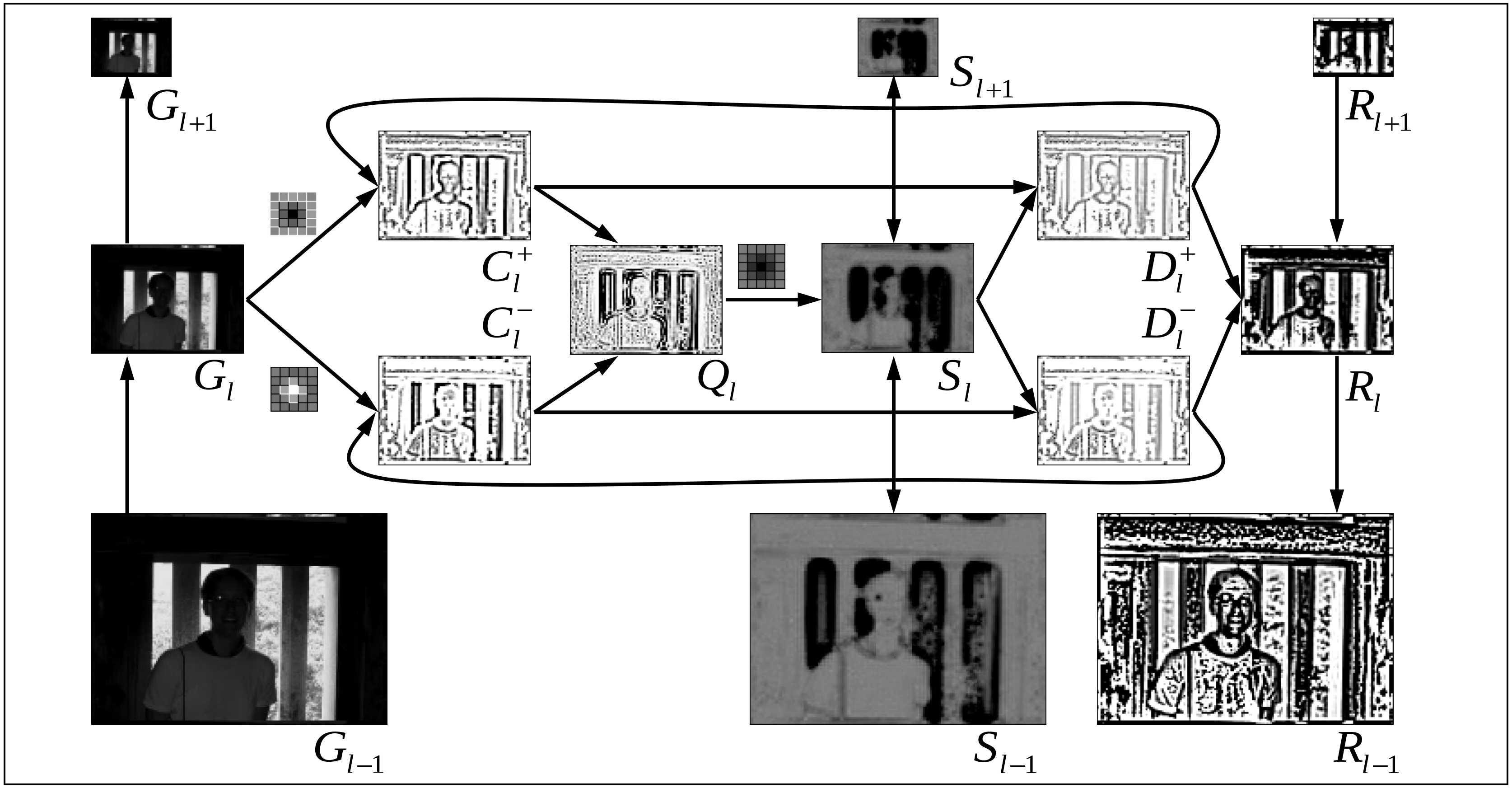
More abstract

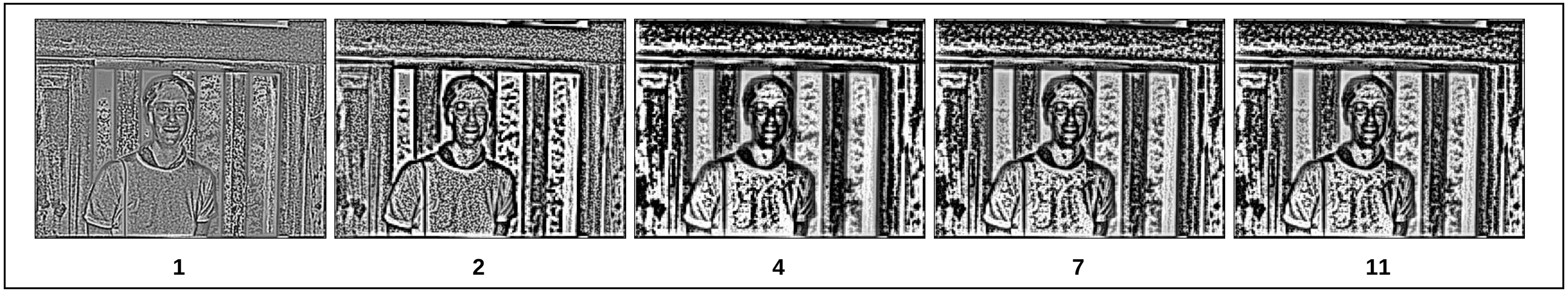
Processor element



[Behnke, LNCS 2766, 2003]

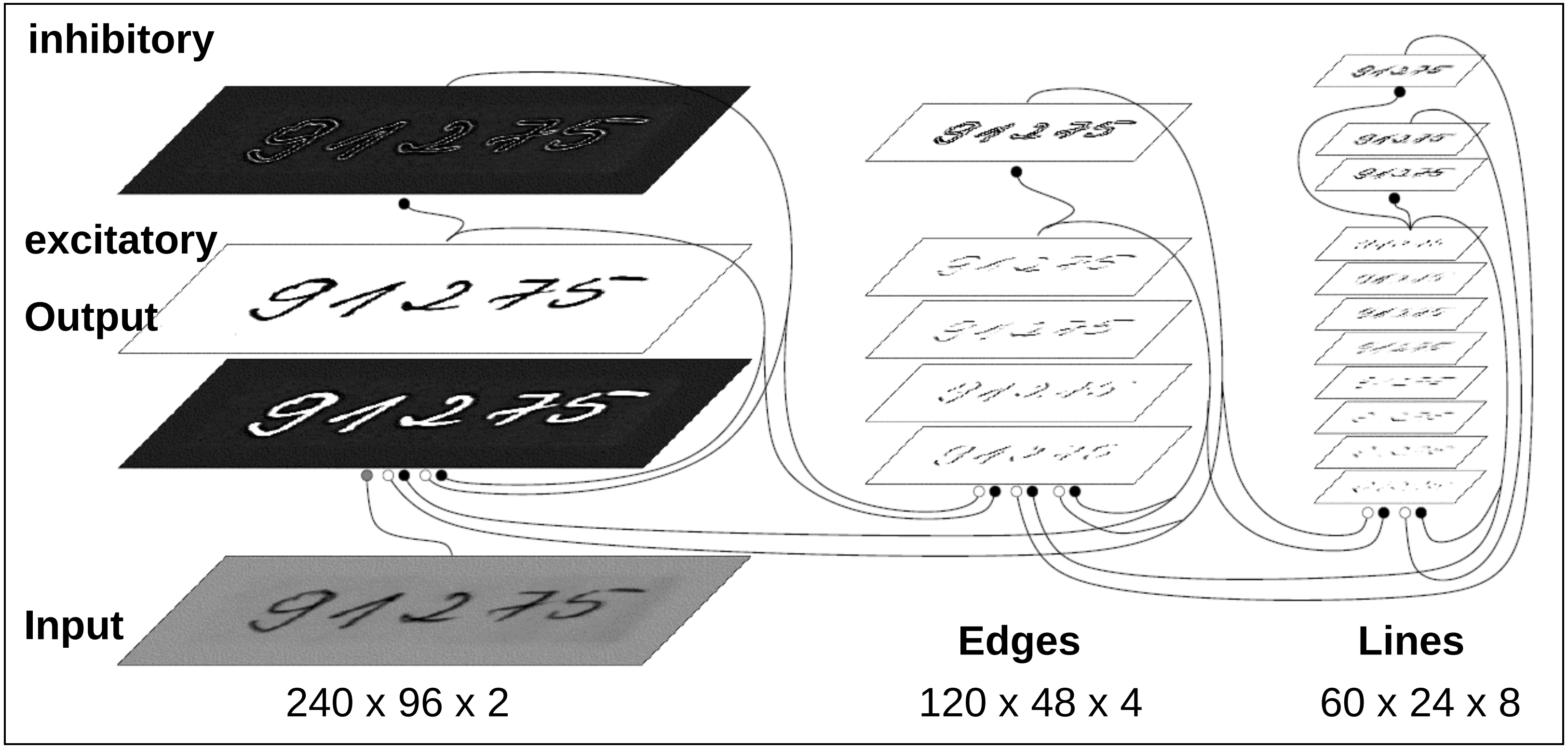
Contrast Normalization



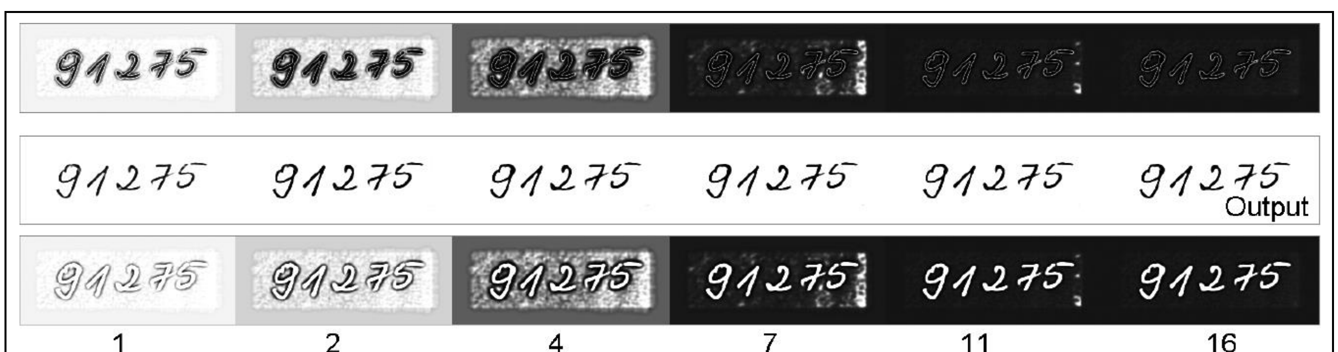


Binarization of Handwriting

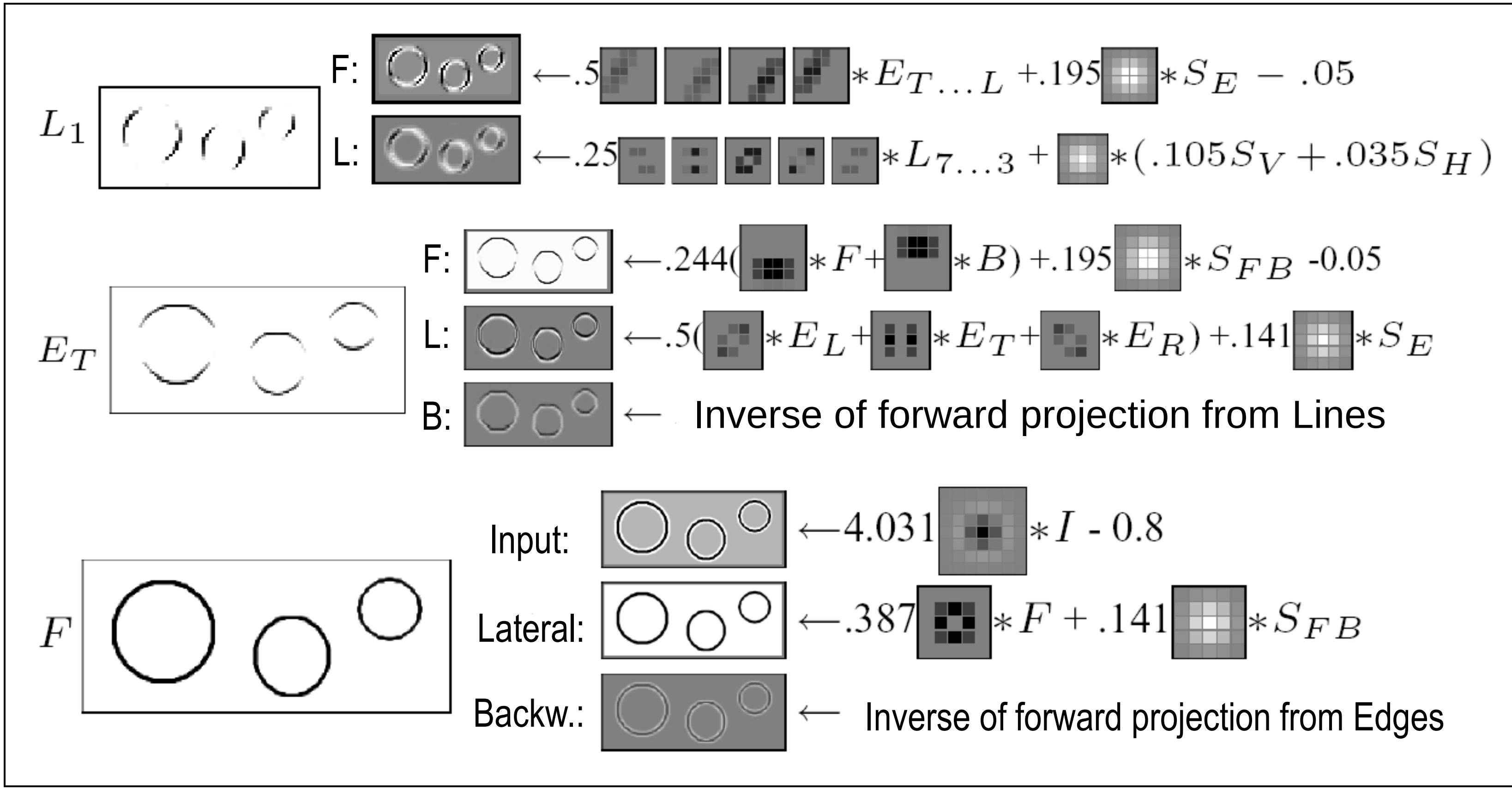
[IJCNN’98]



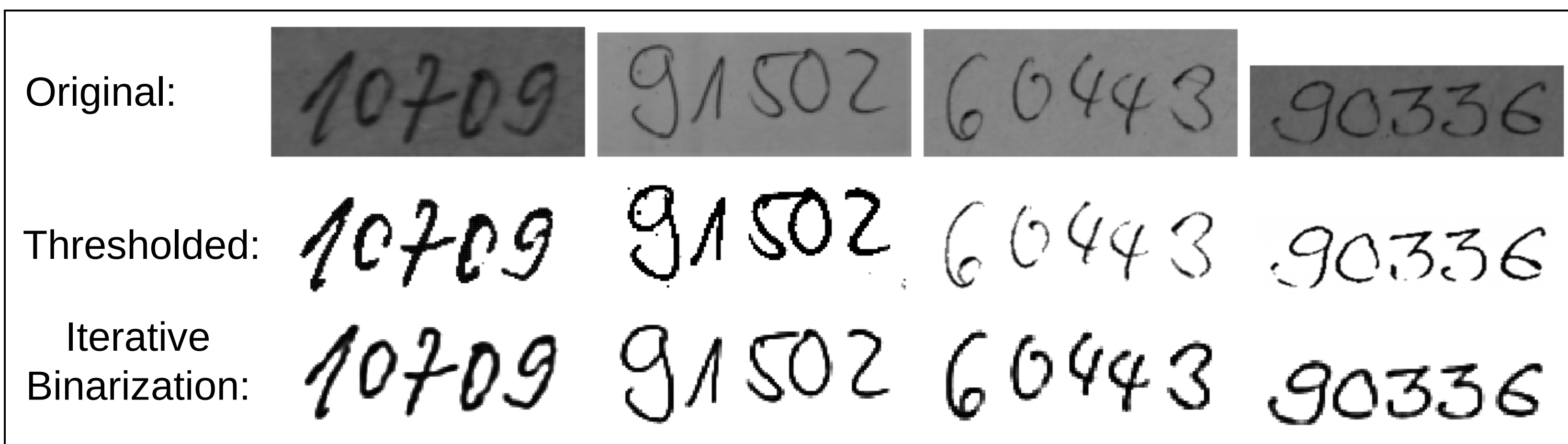
Iterative refinement:



universitätbonn ais

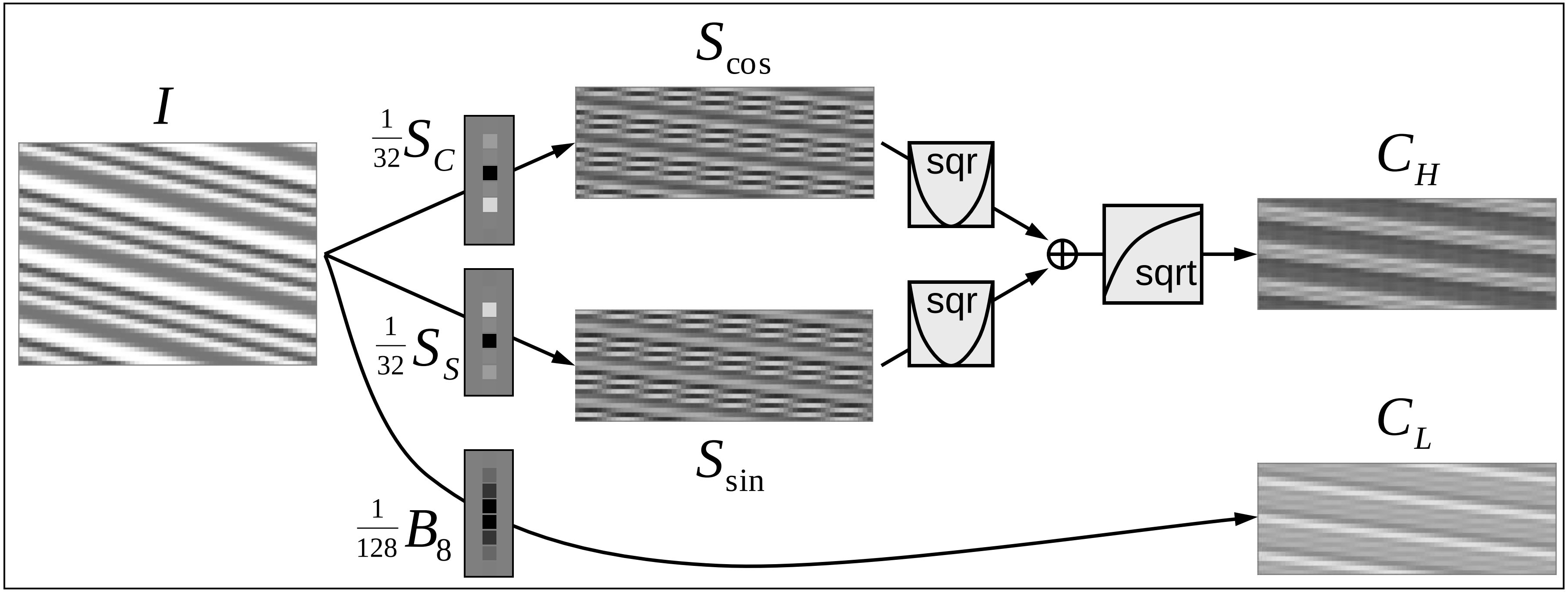


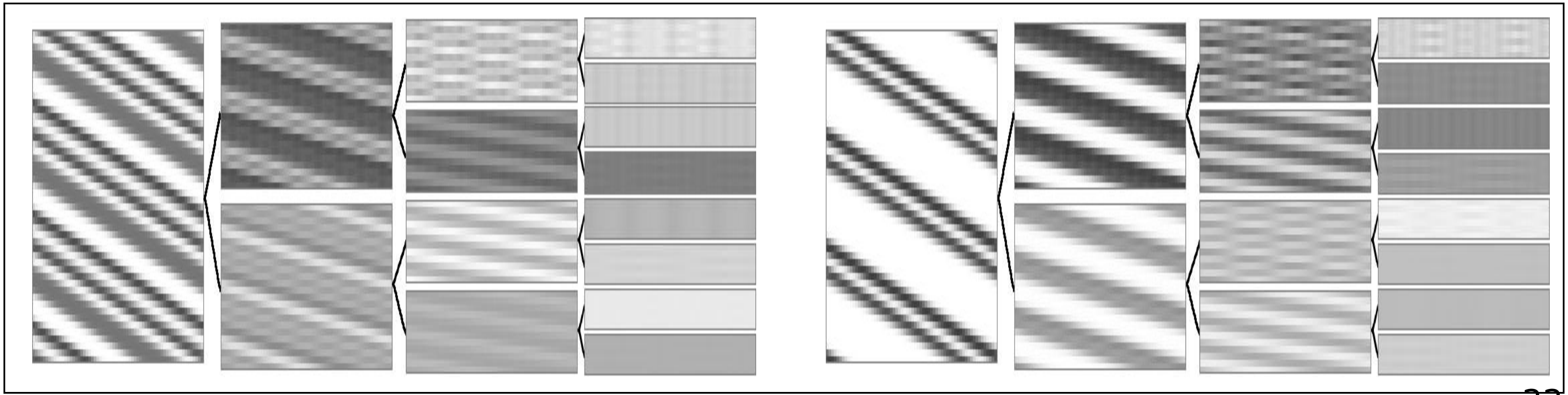
L_1
F: $\leftarrow .5 * E_{T \dots L} + .195 * S_E - .05$
L: $\leftarrow .25 * L_{7 \dots 3} + * (.105 S_V + .035 S_H)$
E_T
F: $\leftarrow .244(* F + * B) + .195 * S_{FB}$ -0.05
L: $\leftarrow .5(* E_L + * E_T + * E_R) + .141 * S_E$
B: $\leftarrow$ Inverse of forward projection from Lines
F
Input: $\leftarrow 4.031 * I$ - 0.8
Lateral: $\leftarrow .387 * F + .141 * S_{FB}$
Backw.: $\leftarrow$ Inverse of forward projection from Edges



Original:
Thresholded:
Iterative Binarization:

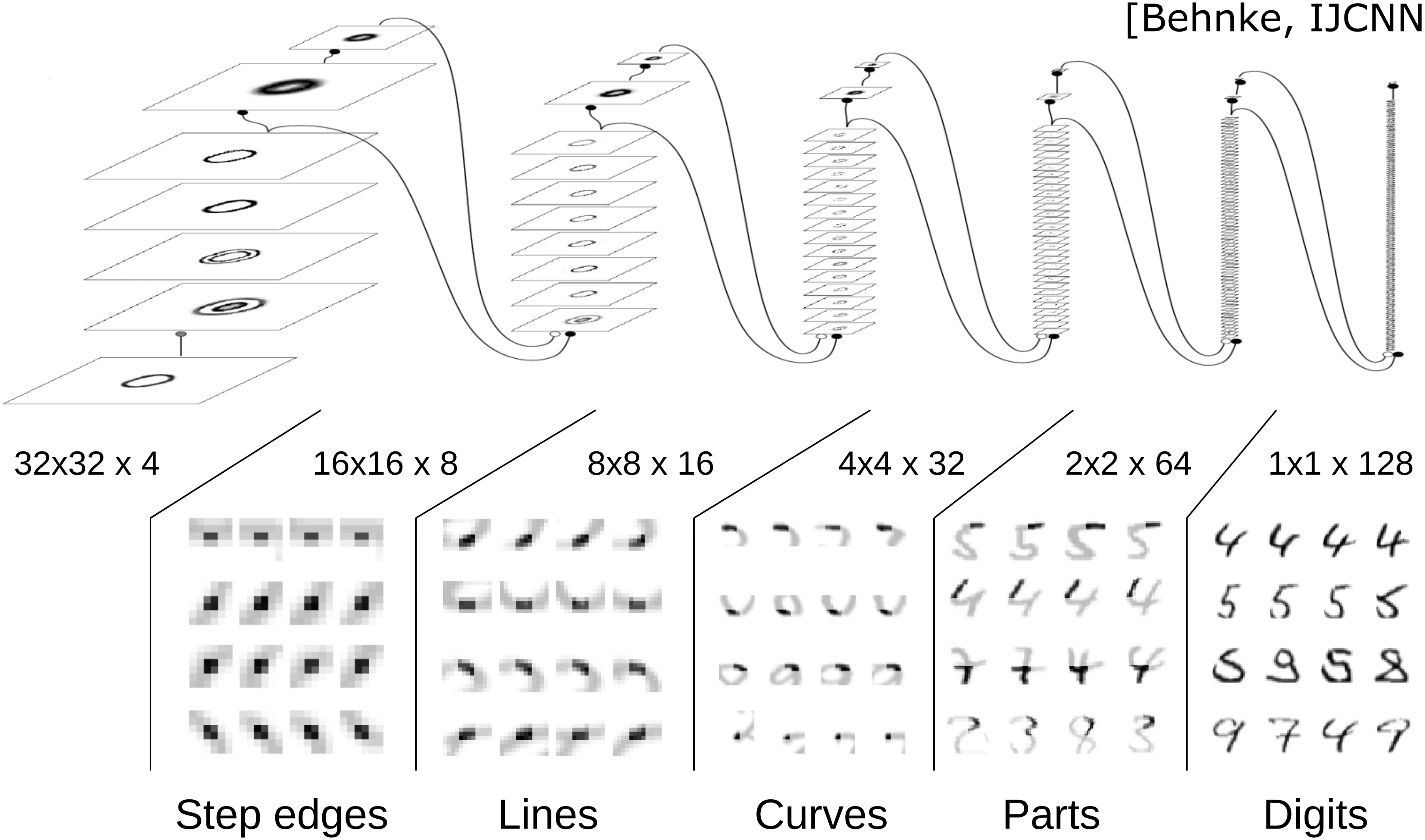
Creating Shift Invariance





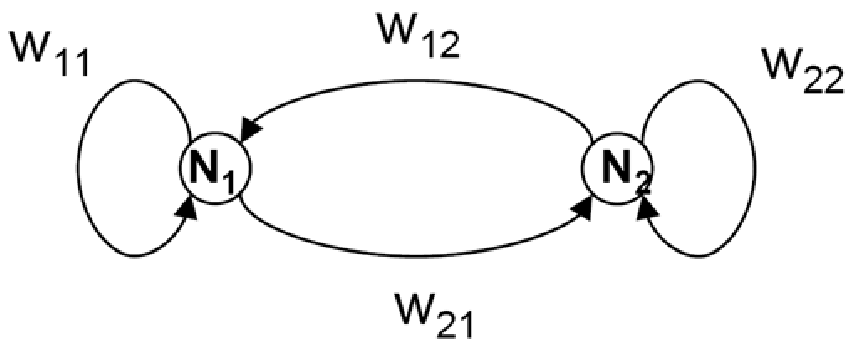
Learning a Feature Hierarchy

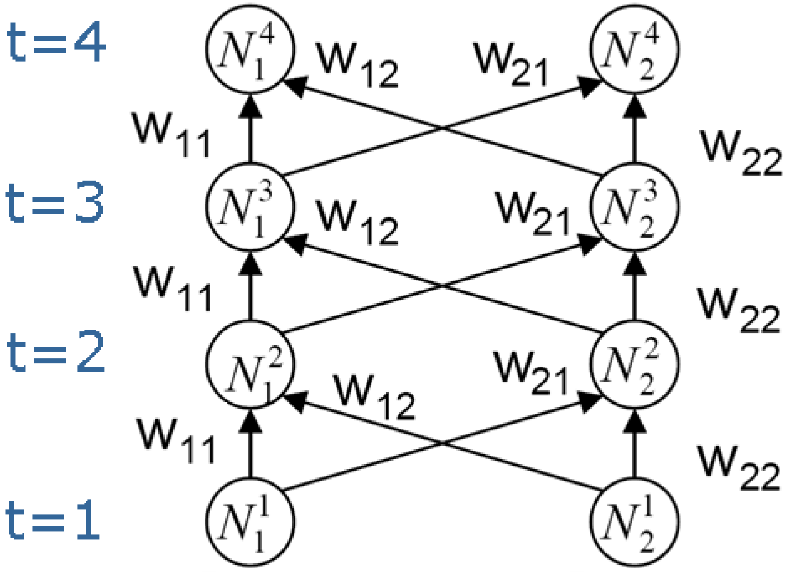
[Behnke, IJCNN'99]



Backpropagation Through Time (BPTT)

- Unfolding along time axis -> deep network
- Weight-sharing -> Average updates

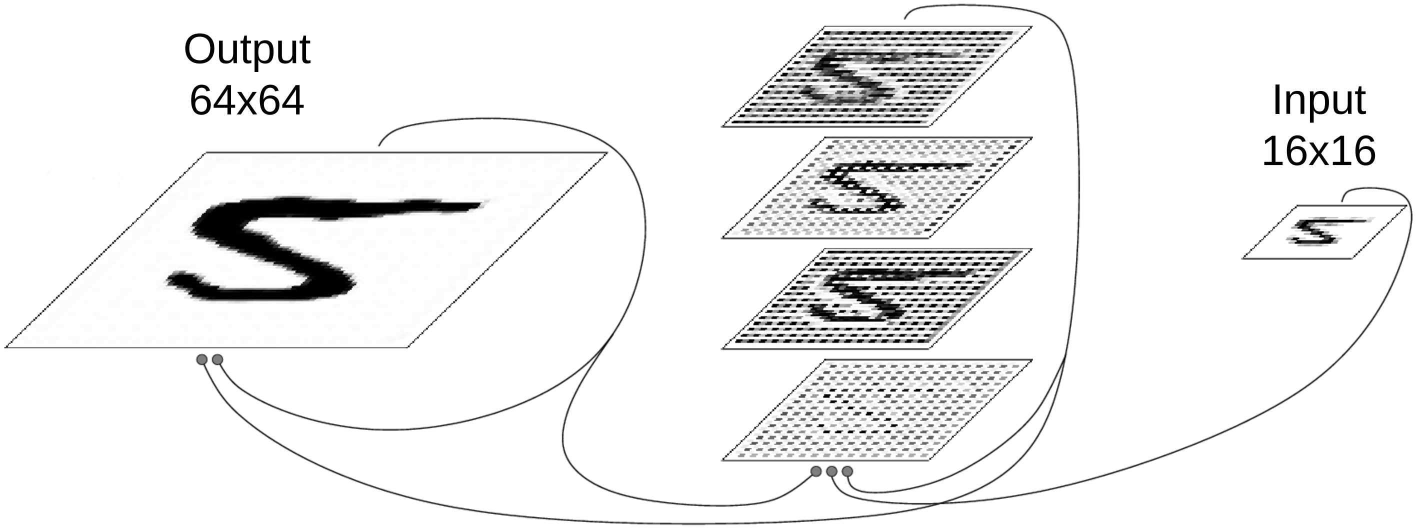


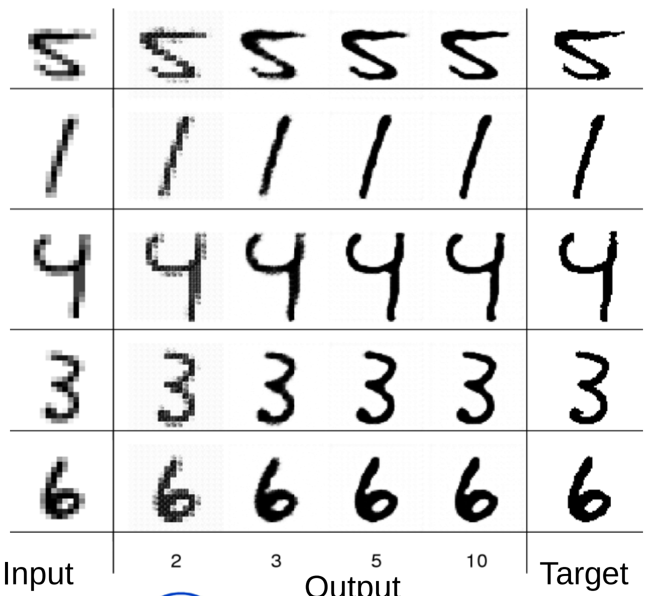


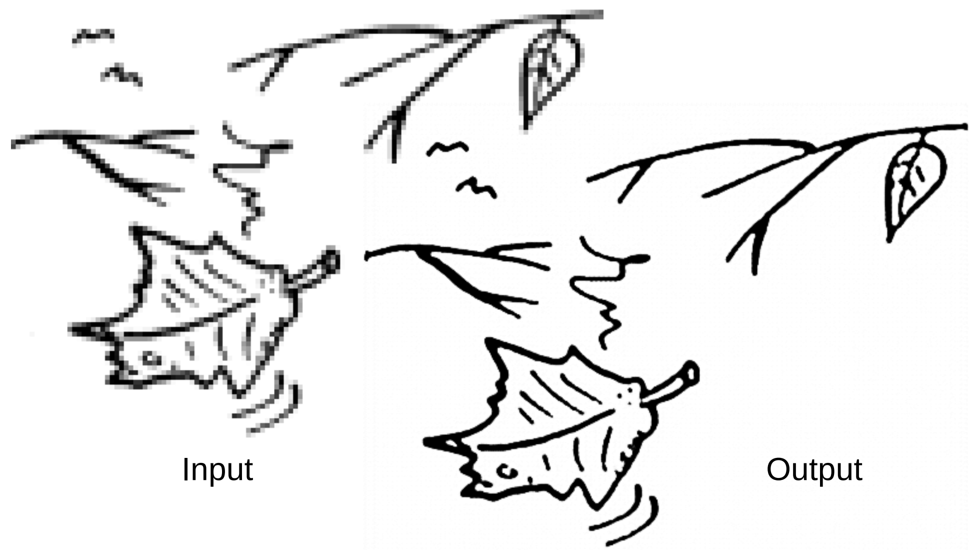
$$o_i\,(t+1) = f\,(net_i(t)) = f\Big(\sum_j w_{ij} o_j(t) + x_i(t)\Big)$$

Superresolution

[Behnke, IJCAI'01]

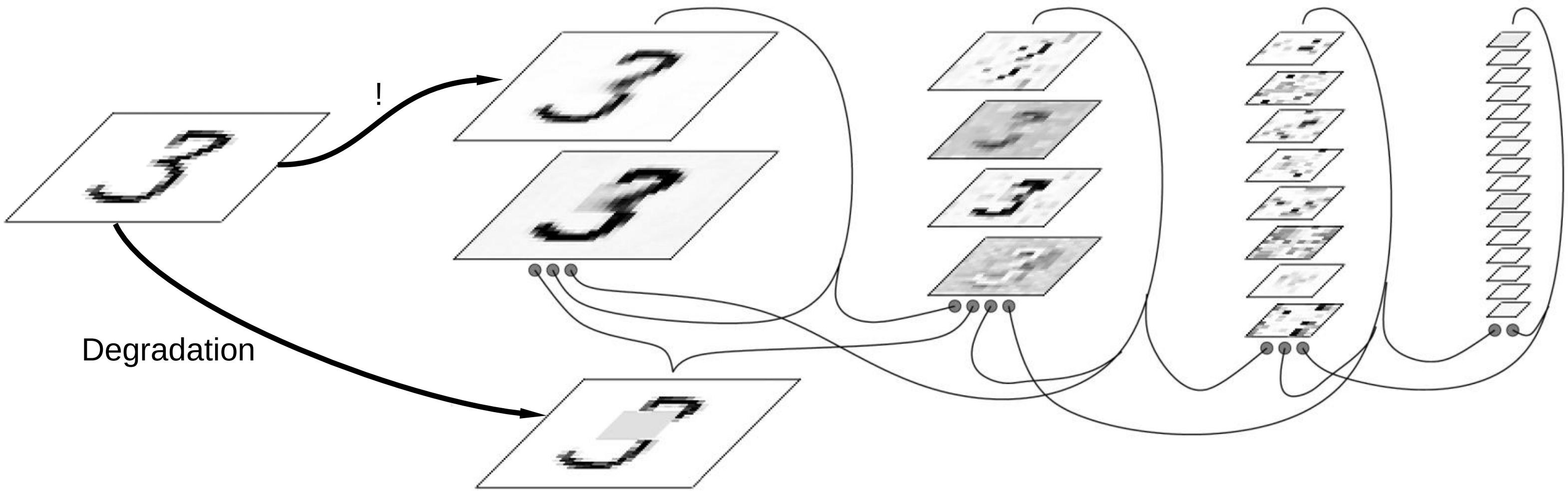


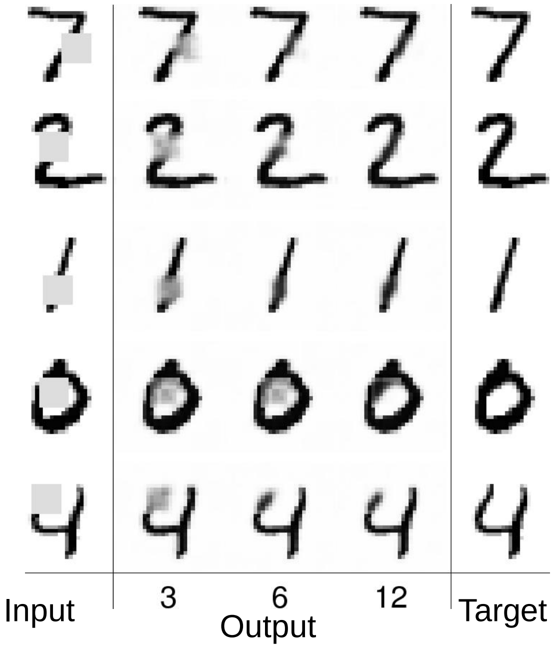


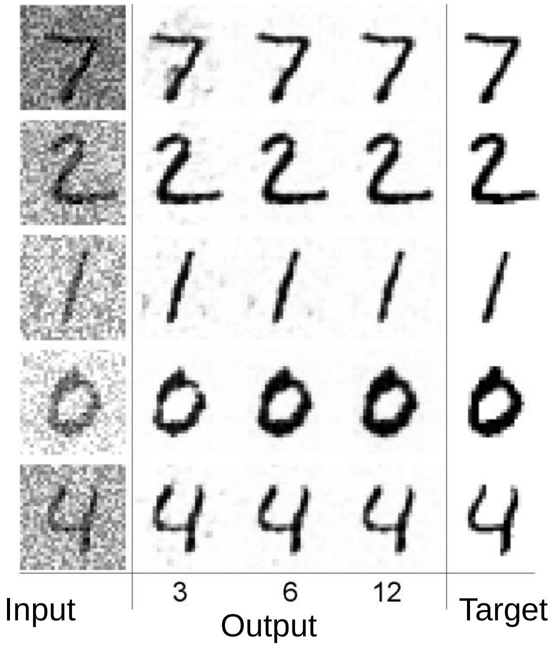


Digit Reconstruction

[Behnke, IJCAI'01]

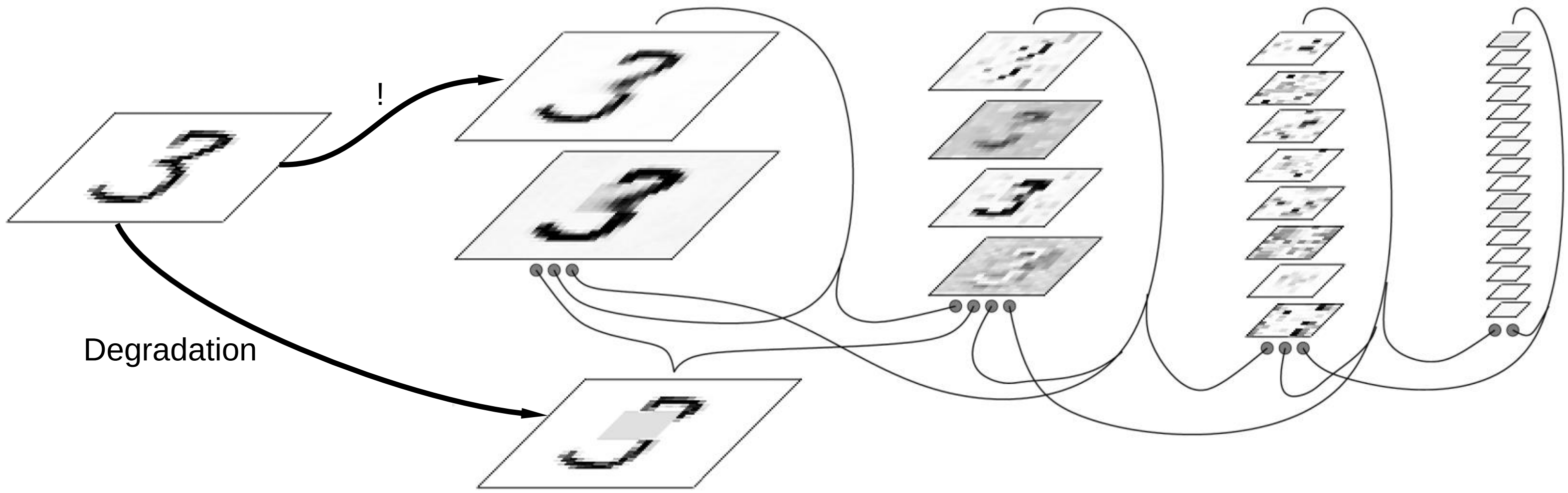


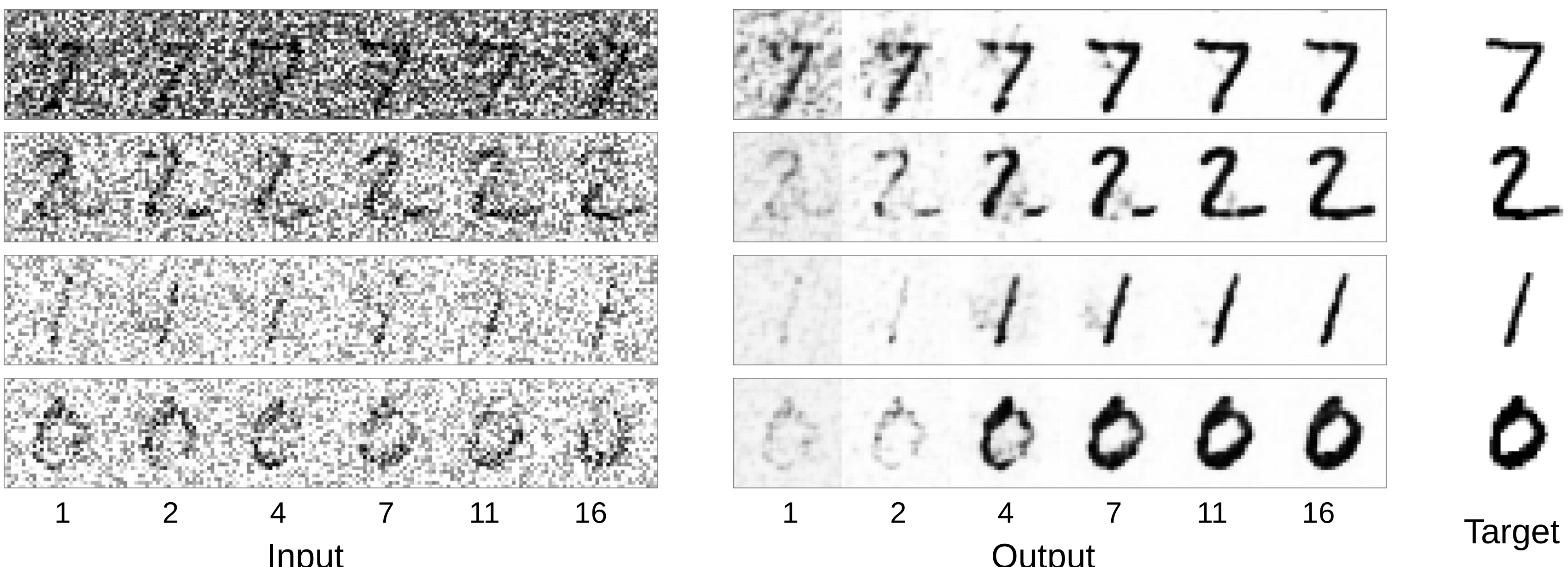




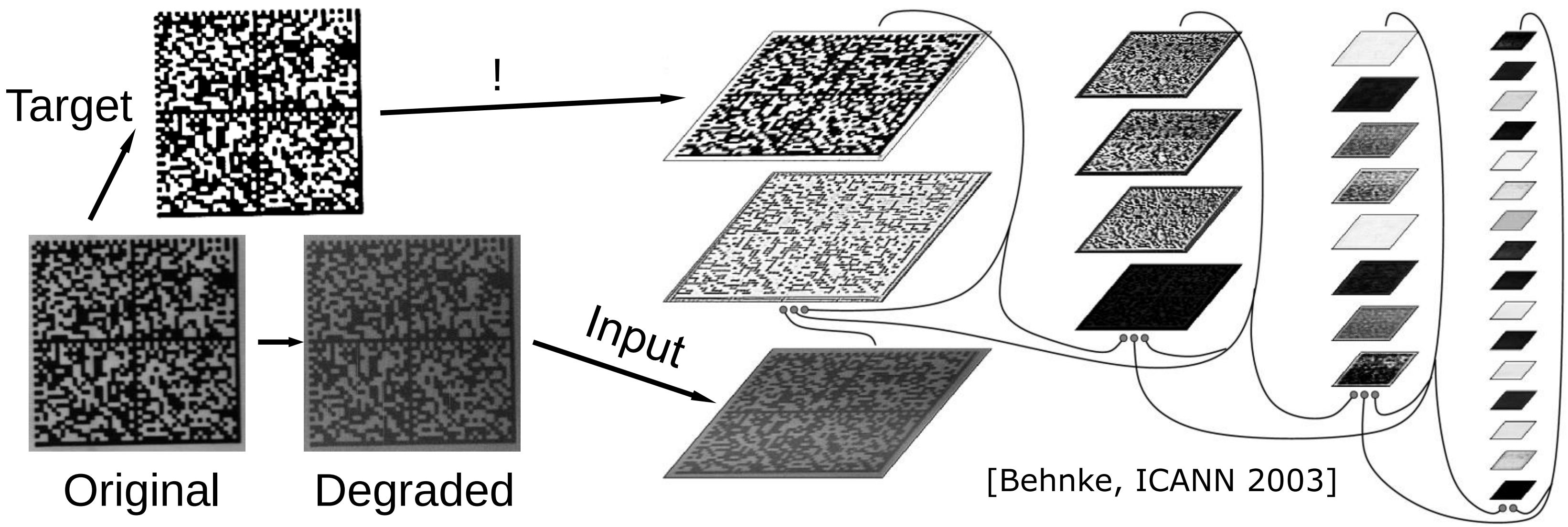
Digit Reconstruction

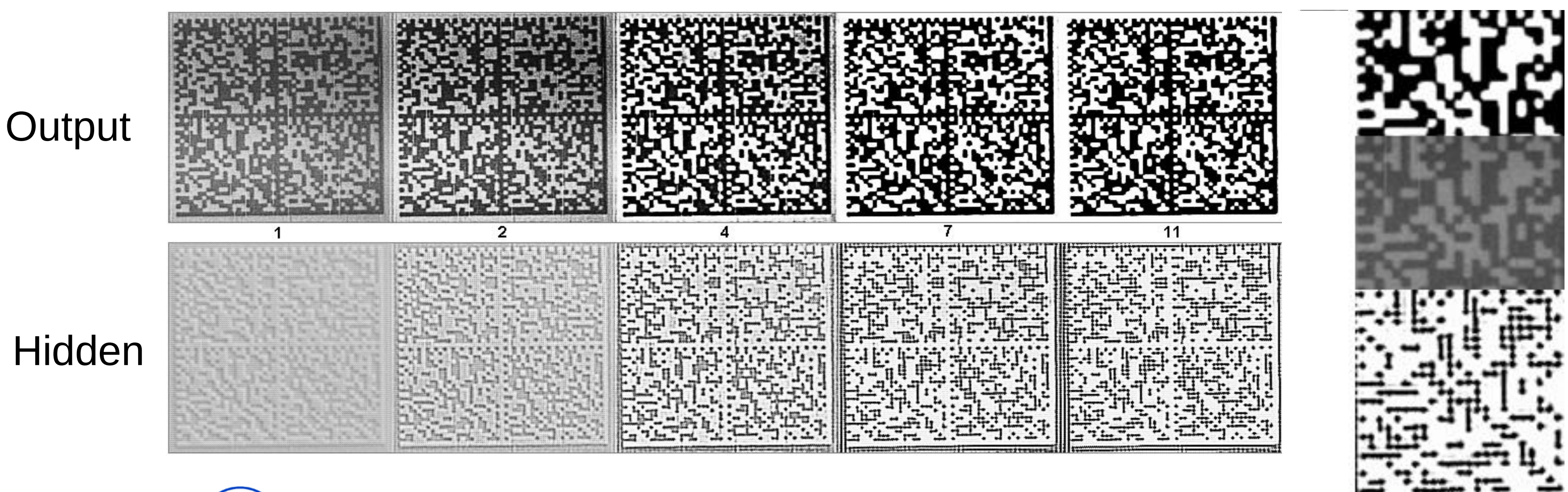
[Behnke, IJCAI'01]



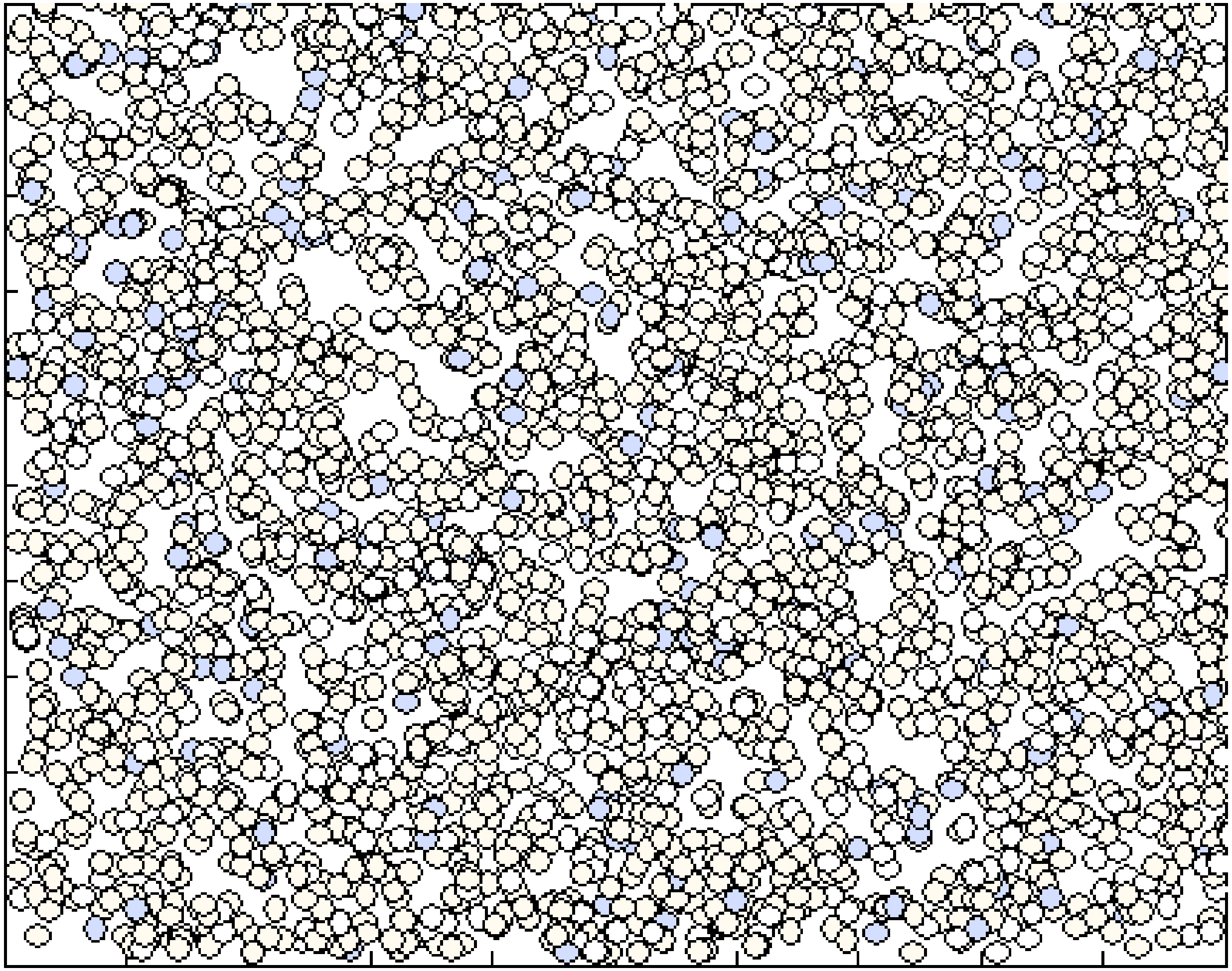


Binarization of Matrix Codes





Continous Attractor



- Local excitation and global inhibition
- Stable activity blobs can be shifted

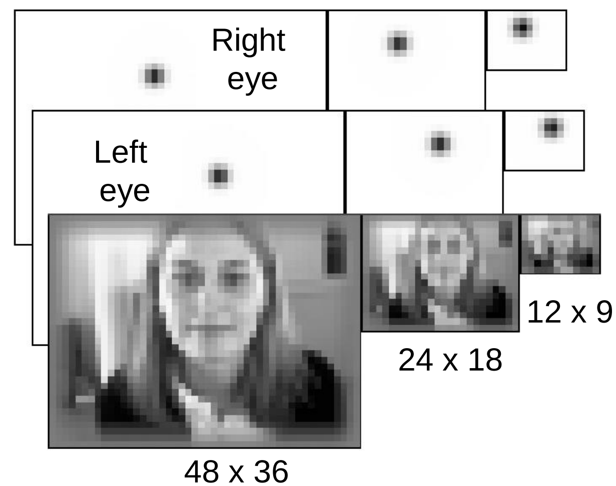
Face Localization

[Behnke, KES'03]

- BioID data set:
  - 1521 images
  - 23 persons

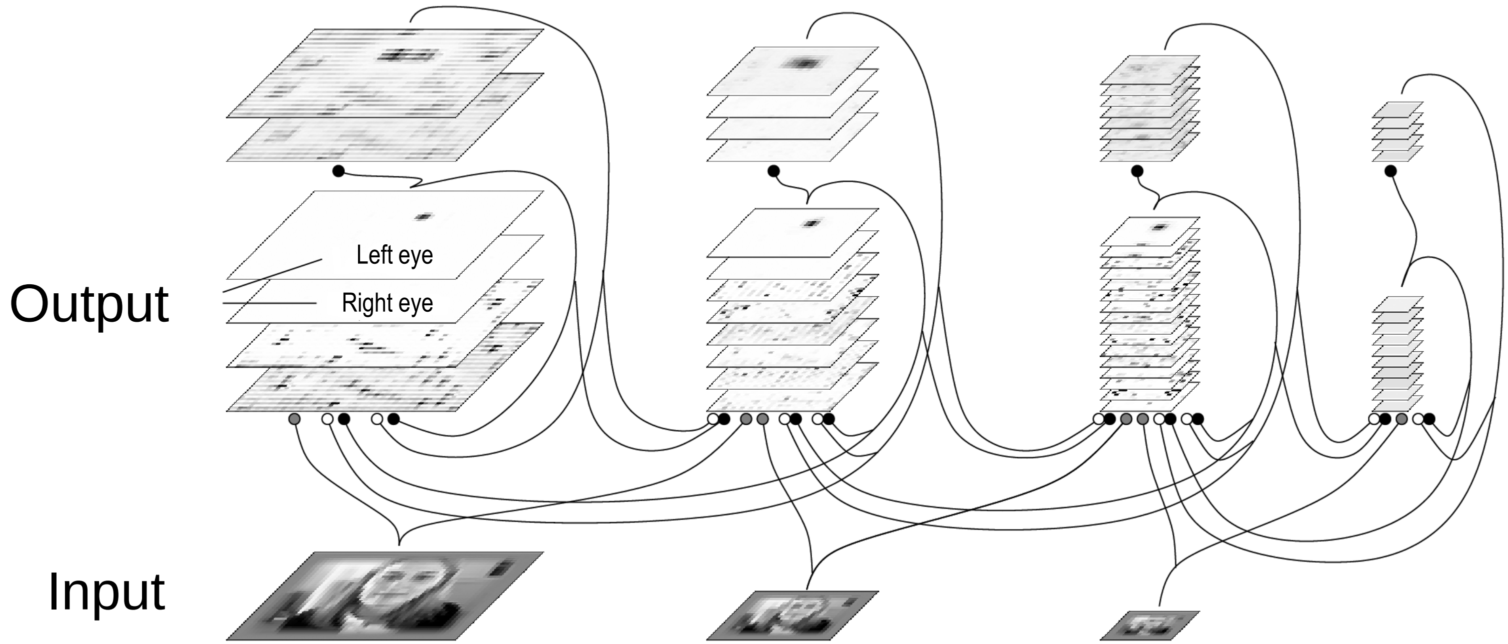
- Encode eye positions with blobs

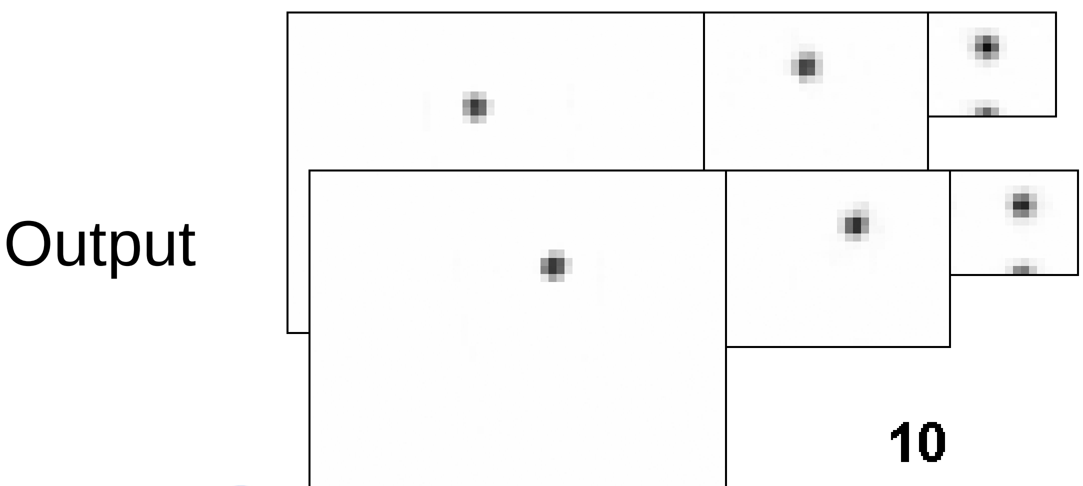
384 x 288



Face Localization

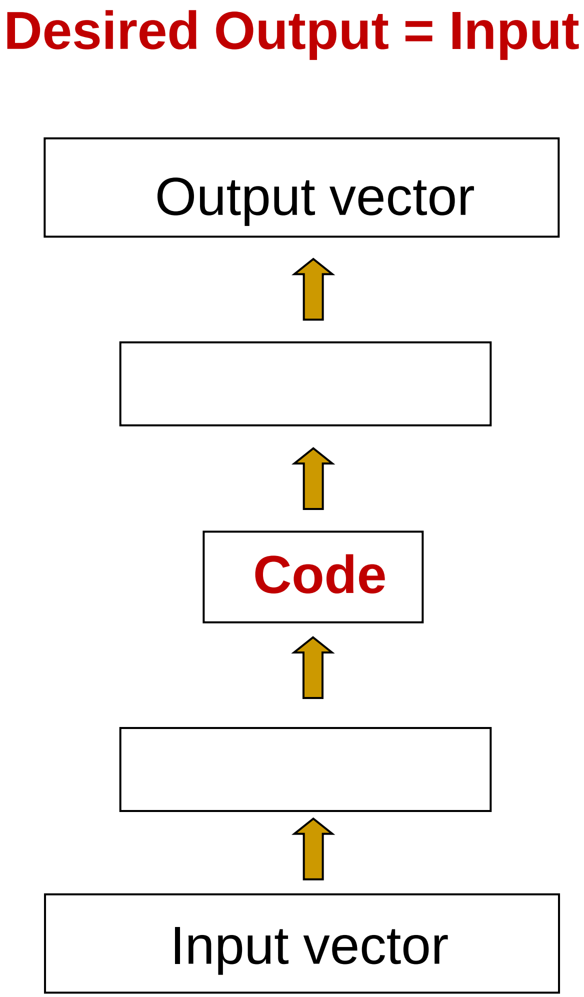
[Behnke, KES’03]





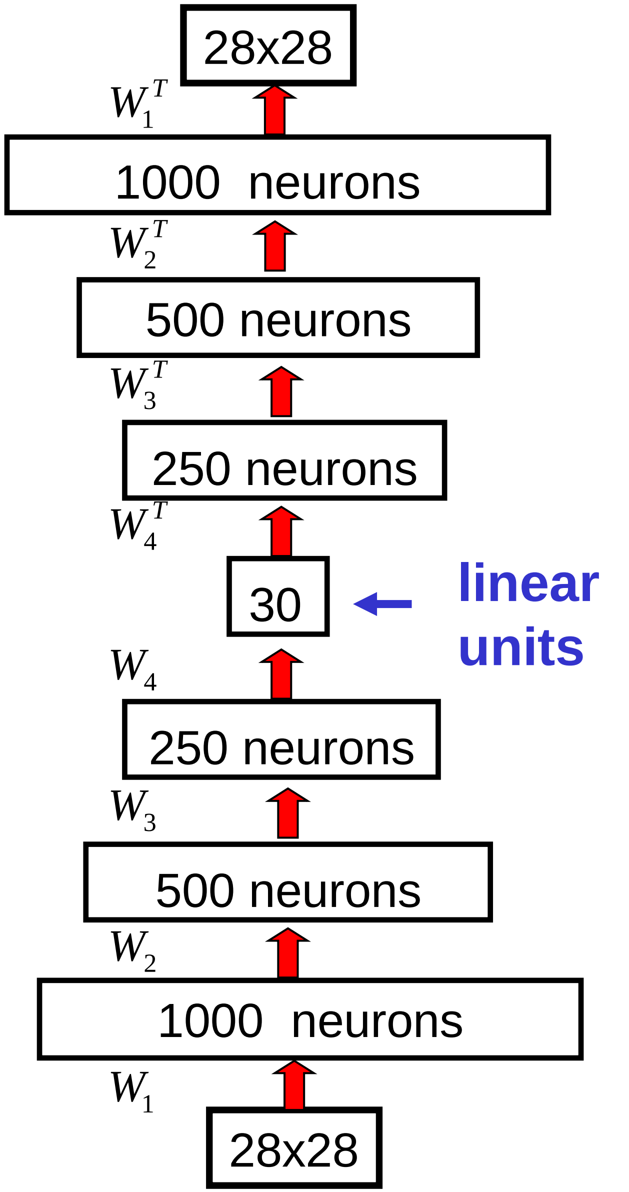
Auto-Encoder

- Try to push input through a bottleneck
- Activities of hidden units form an efficient code
  - There is no space for redundancy in the bottleneck
- Extracts frequently independent features (factorial code)



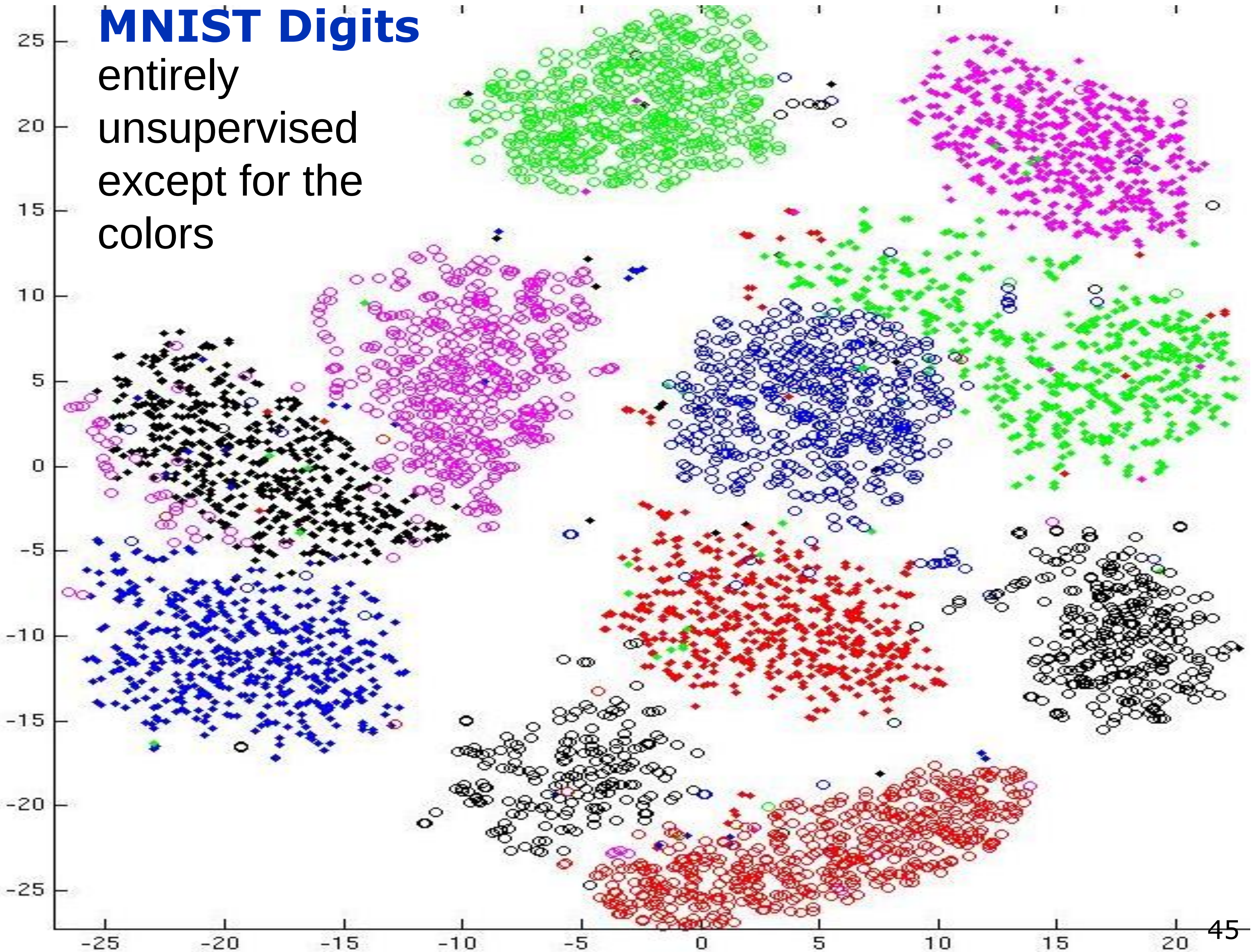
Deep Autoencoders
(Hinton & Salakhutdinov, 2006)

- Multi-layer autoencoders for non-linear dimensionality reduction
- Difficult to optimize deep autoencoders using backpropagation
- Greedy, layer wise training
- Unrolling
- Supervised fine-tuning



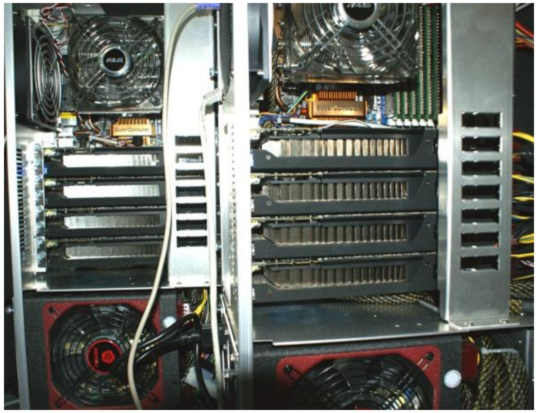
MNIST Digits

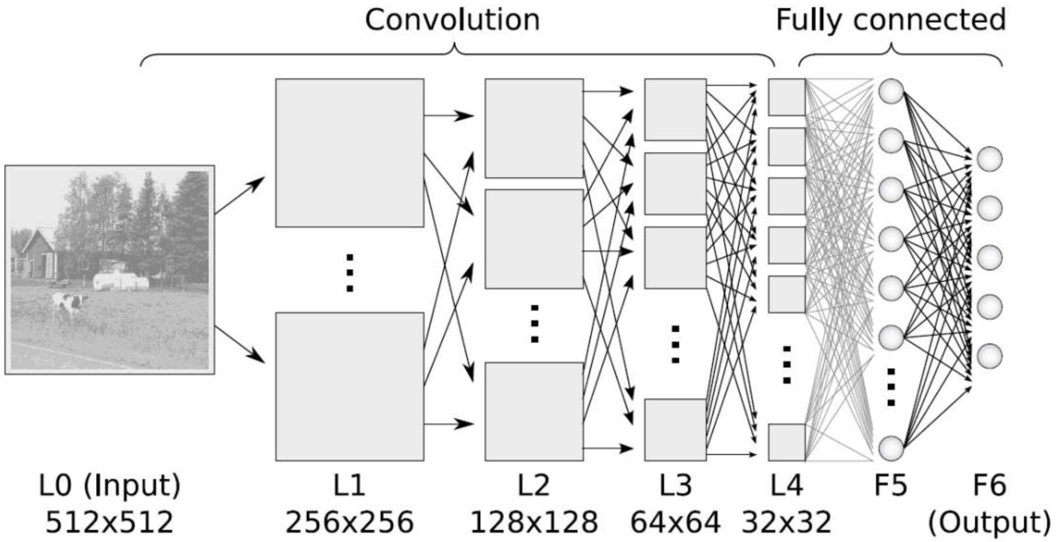
entirely unsupervised except for the colors

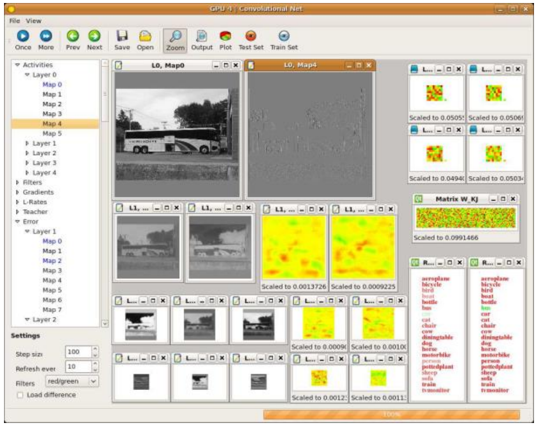


GPU Implementations (CUDA)

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



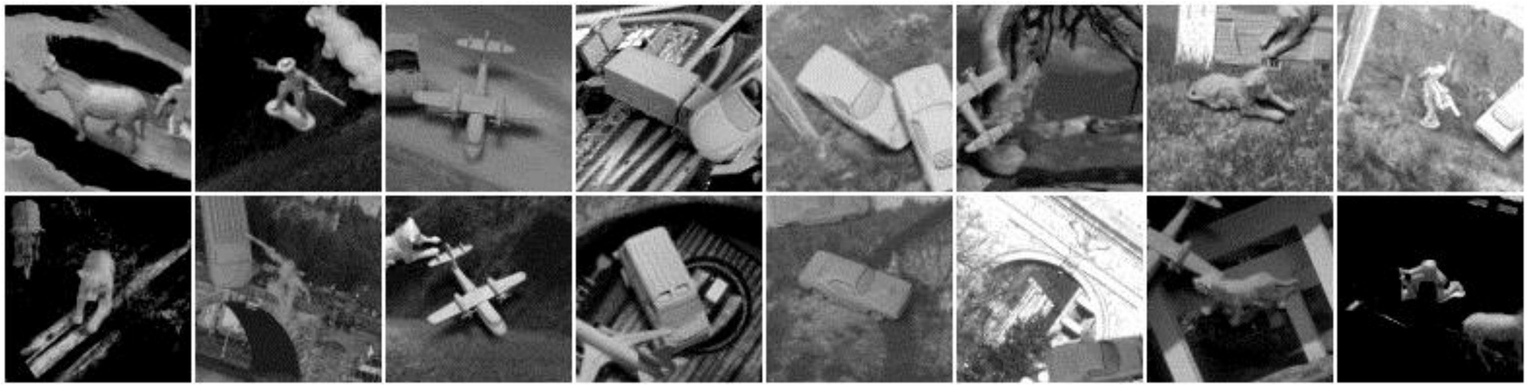




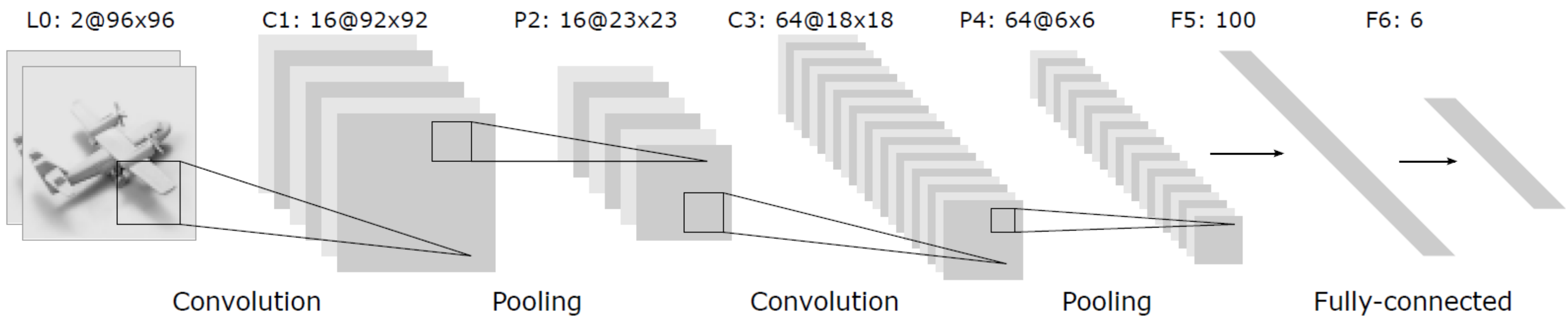
- Local connectivity [Uetz & Behnke, 2009]

Image Categorization: NORB

- 10 categories, jittered-cluttered



- **Max-Pooling**, cross-entropy training

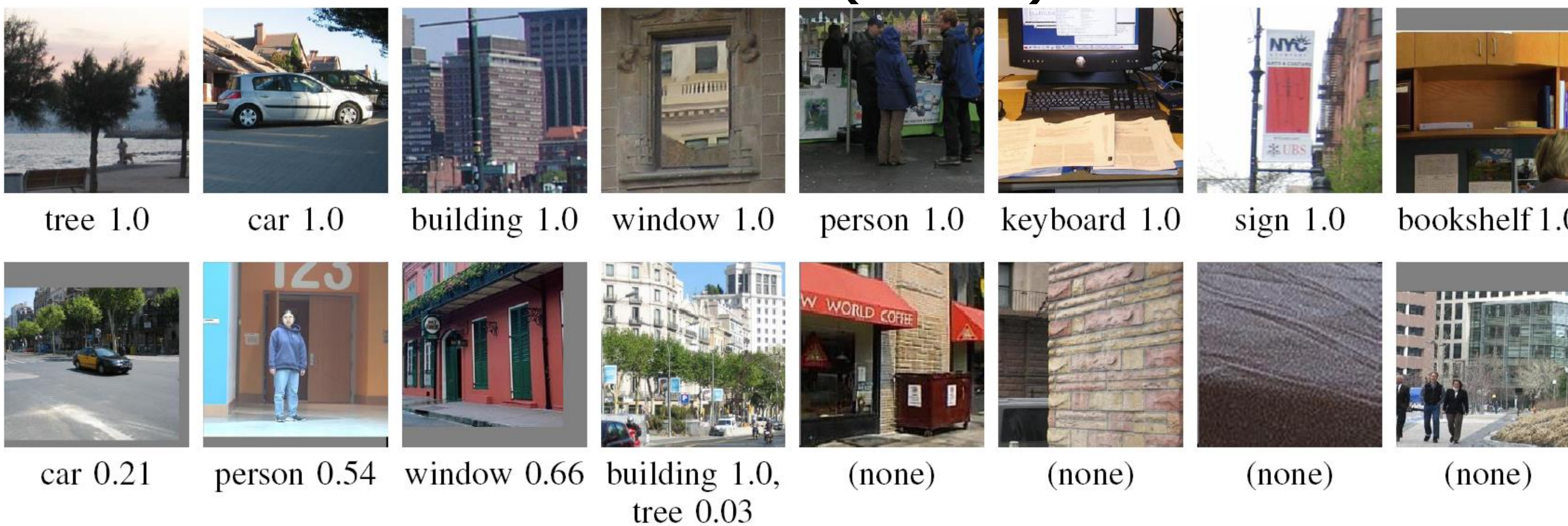


- Test error: 5,6% (LeNet7: 7.8%)

[Scherer, Müller, Behnke, ICANN'10]

Image Categorization: LabelMe

- 50,000 color images (256x256)
- 12 classes + clutter (50%)

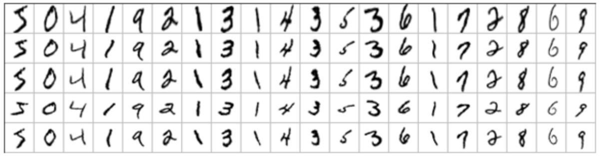


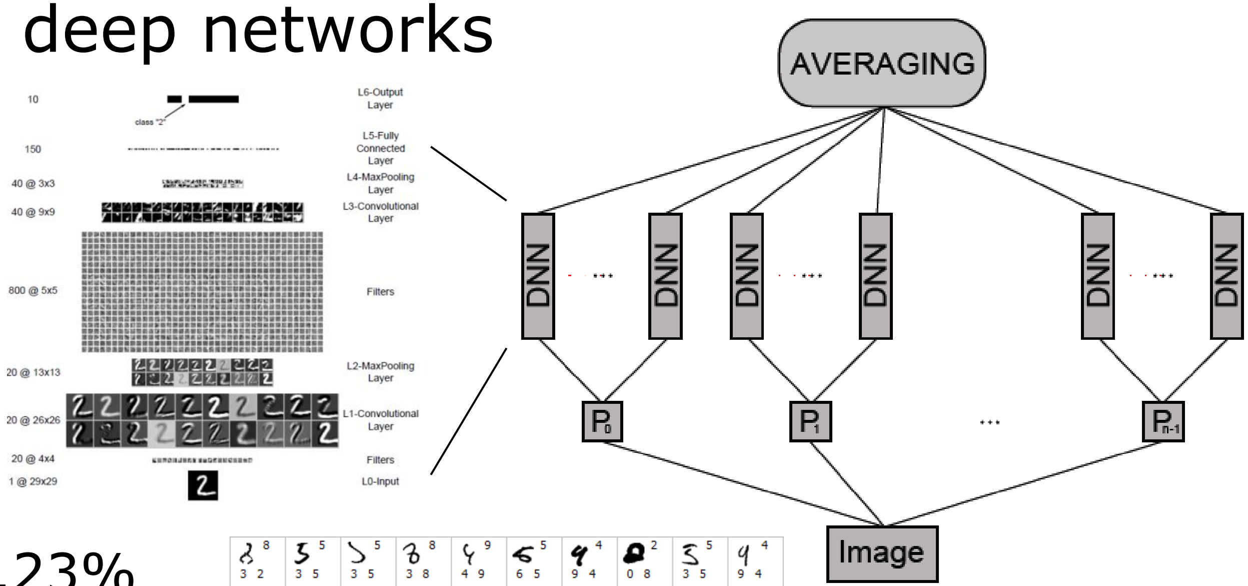
- Error TRN: 3.77%; TST: 16.27%
- Recall: 1,356 images/s

[Uetz, Behnke, ICIS2009]

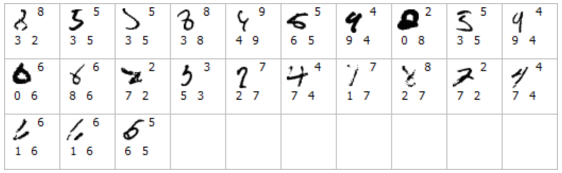
Multi-Column Deep Convolutional Networks

- Different preprocessings
- Trained with distortions
- Bagging deep networks





- MNIST: 0.23%
- NORB: 2.7%
- CIFAR10: 11.2%
- Traffic signs: 0.54% test error



[Ciresan et al. CVPR 2012]

ImageNet Challenge

- 1.2 million images
- 1000 categories, no overlap
- Subset of 11 million images from 15.000+ categories
- Hierarchical category structure (WordNet)

Golf cart (motor vehicle, self-propelled vehicle, wheeled vehicle, …

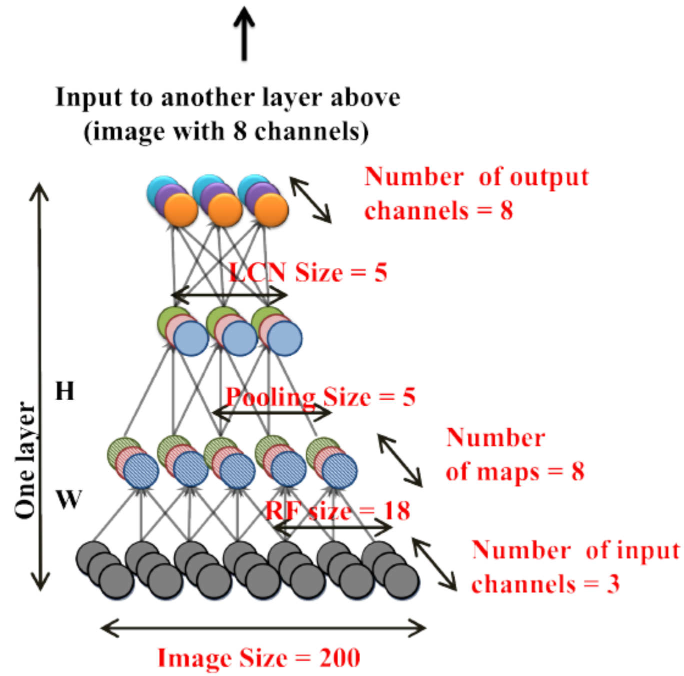
Egyptian cat (domestic cat, domestic animal, animal)

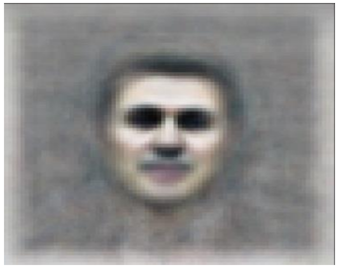
- Task: recognize object category
- Low penalty for extra detections
- Hierarchical error computation

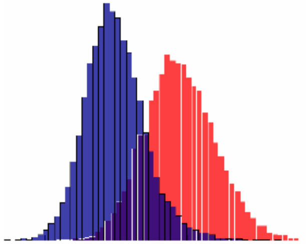
Large Unsupervised Feature Learning

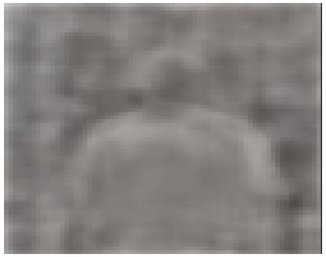
- 9 layer model
- Locally connected
- Sparse auto-encoder
- L2 pooling
- Local contrast normalization
- 1 billion connections
- Trained on 10 million images
- Unsupervised learned detectors

3x







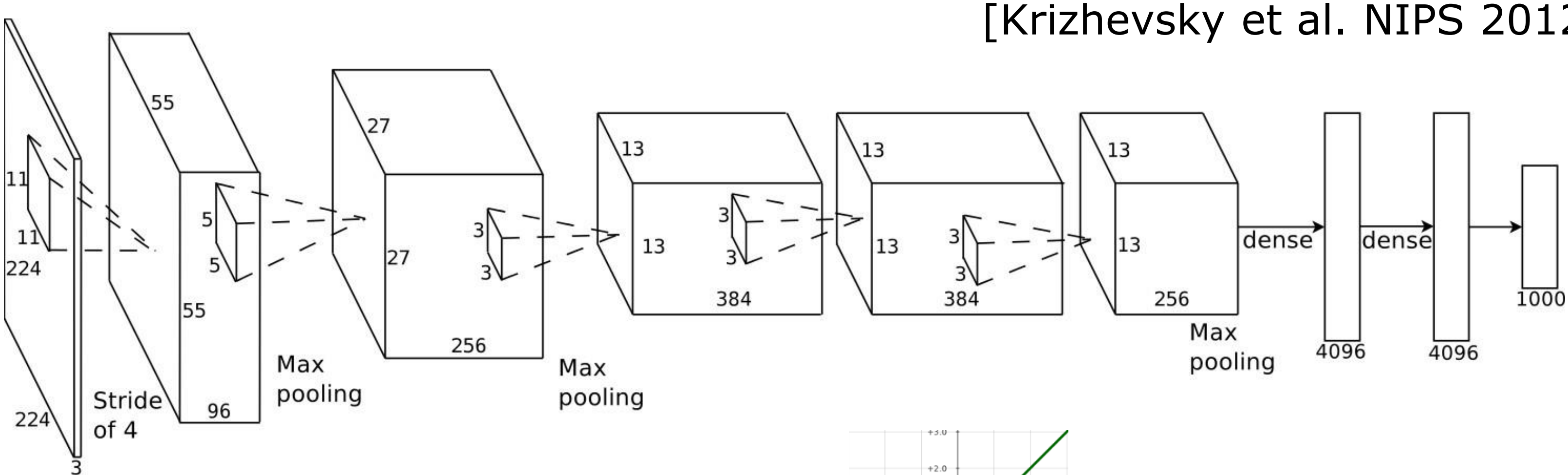


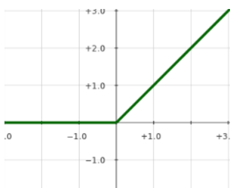
- Supervised ImageNet 2011 results (14M images, 22K categories): 15.8%

[Le et al. 2012]

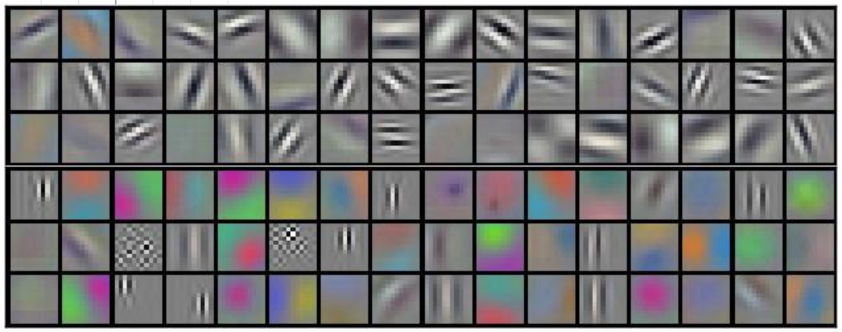
Large Convolutional Network

[Krizhevsky et al. NIPS 2012]



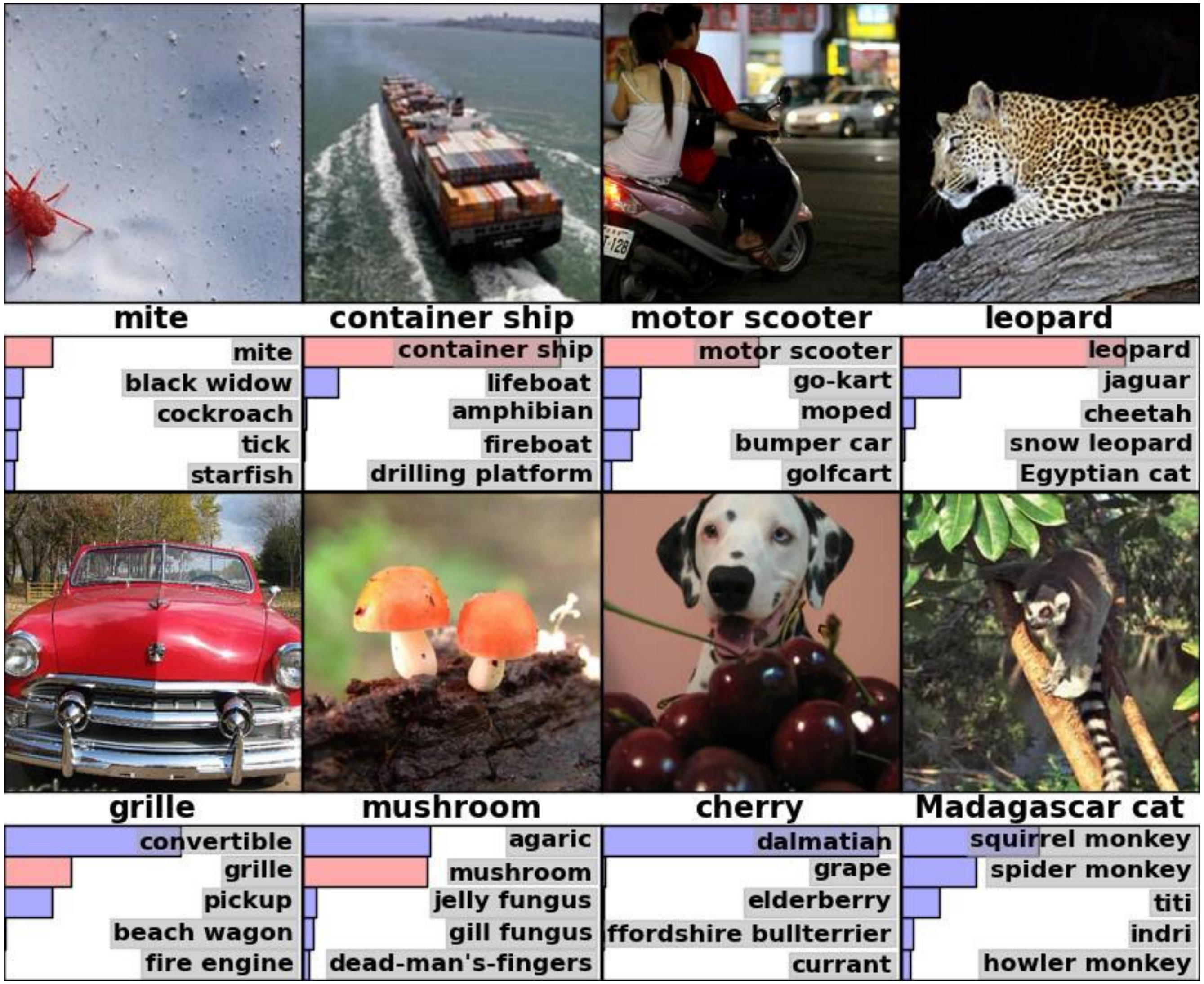


- Rectifying transfer functions
- 650,000 neurons
- 60,000,000 parameters
- 630,000,000 connections
- Trained using dropout and data augmentation
- Testing 10 sub-images
- ILSVRC-2012: top-5 error 15.3%



96 learned low-level filters

Validation Classification



[Krizhevsky et al. NIPS 2012]

Surpassing Human Performance


GT: horse cart
1: horse cart
2: minibus
3: oxcart
4: stretcher
5: half track


GT: birdhouse
1: birdhouse
2: sliding door
3: window screen
4: mailbox
5: pot


GT: forklift
1: forklift
2: garbage truck
3: tow truck
4: trailer truck
5: go-kart


GT: letter opener
1: drumstick
2: candle
3: wooden spoon
4: spatula
5: ladle


GT: coucal
1: coucal
2: indigo bunting
3: lorikeet
4: walking stick
5: custard apple


GT: komondor
1: komondor
2: patio
3: llama
4: mobile home
5: Old English sheepdog

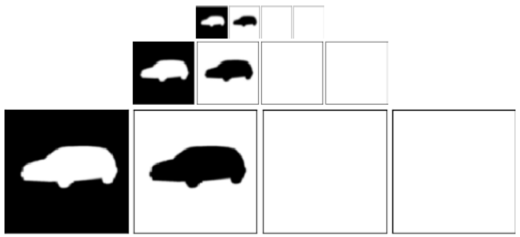

GT: yellow lady's slipper
1: yellow lady's slipper
2: slug
3: hen-of-the-woods
4: stinkhorn
5: coral fungus


GT: spotlight
1: grand piano
2: folding chair
3: rocking chair
4: dining table
5: upright piano

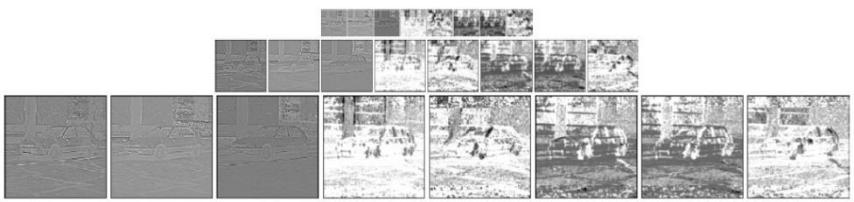
[He et al. 2015]

Object-class Segmentation

- Class annotation per pixel

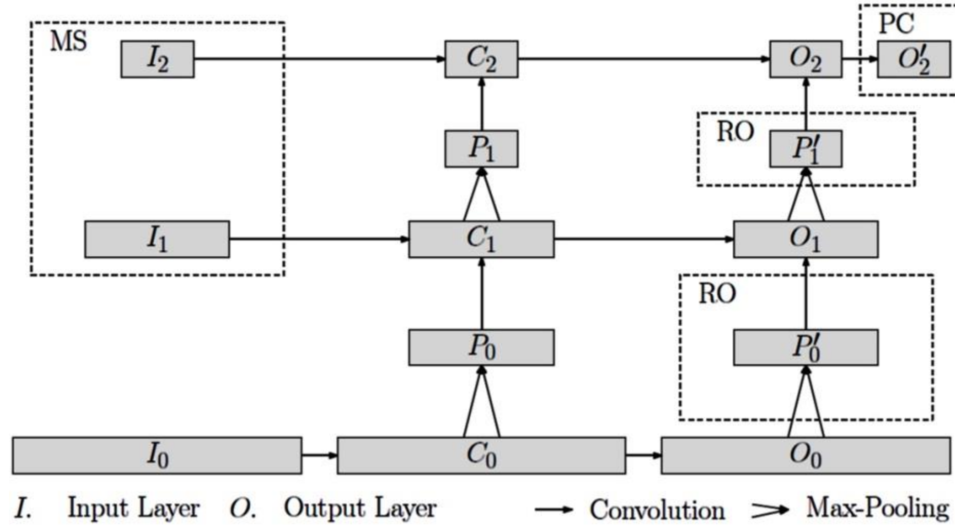


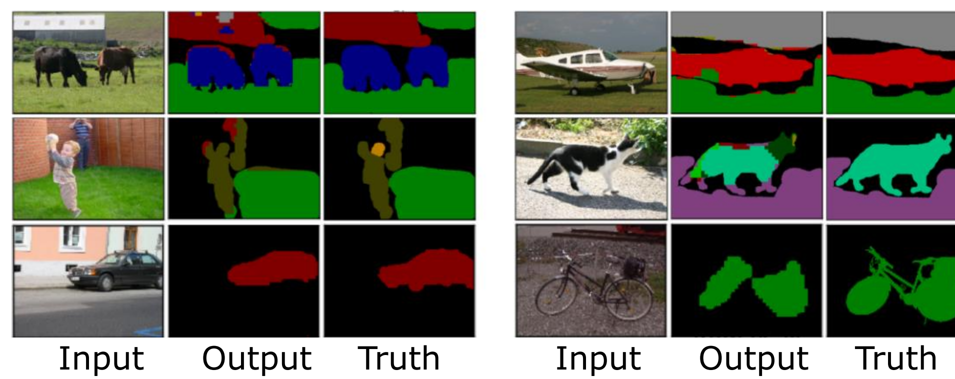
- Multi-scale input channels



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

[Schulz, Behnke 2012]





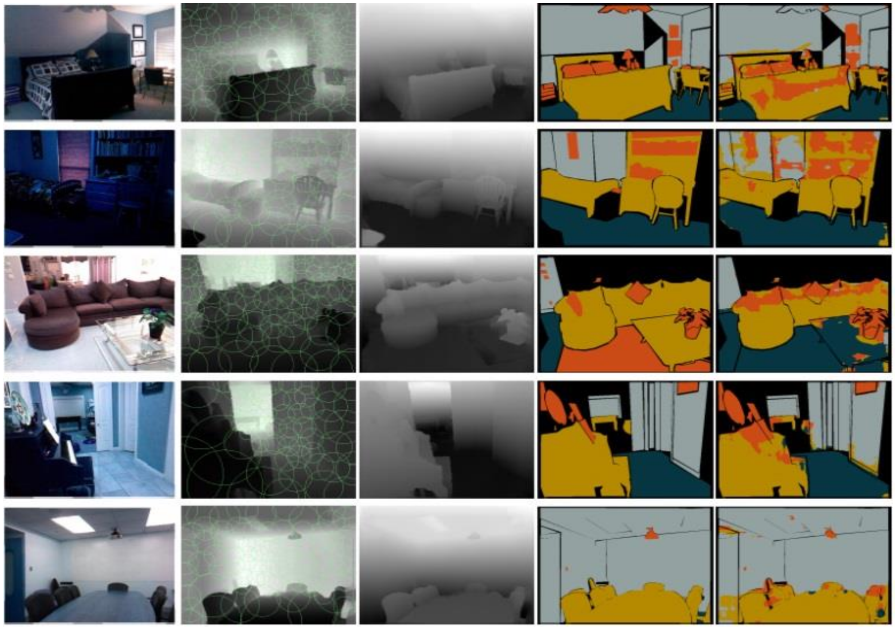
Object Detection in Images

- 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
- Scale input according to depth, compute pixel height



NYU Depth V2

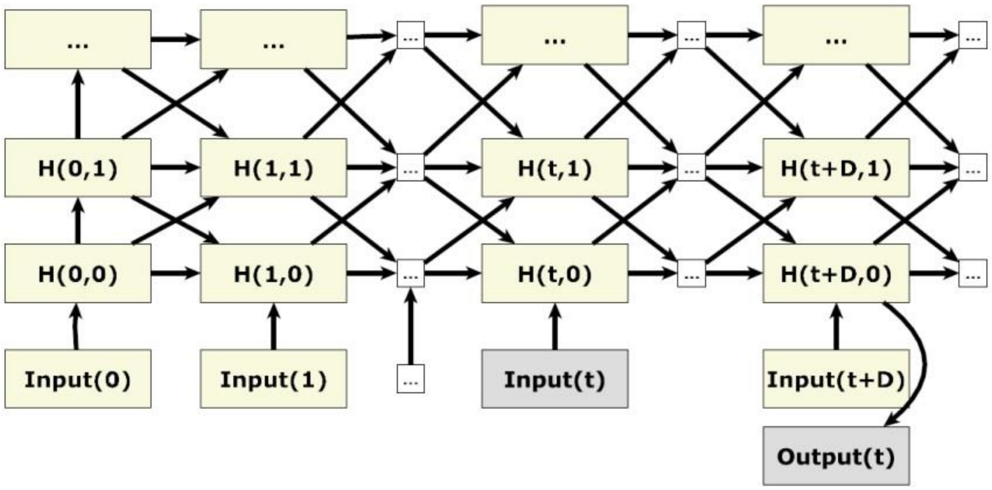
Method	floor	struct	furnit	prop	Class Avg.	Pixel Acc.
CW	84.6	70.3	58.7	52.9	66.6	65.4
CW+DN	87.7	70.8	57.0	53.6	67.3	65.5
CW+H	78.4	74.5	55.6	62.7	67.8	66.5
CW+DN+H	93.7	72.5	61.7	55.5	70.9	70.5
CW+DN+H+SP	91.8	74.1	59.4	63.4	72.2	71.9
CW+DN+H+CRF	93.5	80.2	66.4	54.9	**73.7**	**73.4**
Müller et al.[8]	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

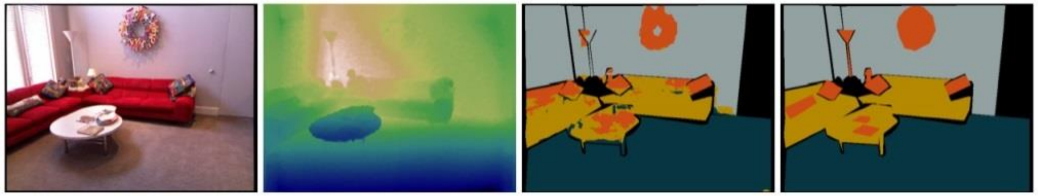
CW is covering windows, H is height above ground, DN is depth normalized patch sizes. SP is averaged within superpixels and SVM-reweighted. CRF is a conditional random field over superpixels [8]. Structure class numbers are optimized for class accuracy.

[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

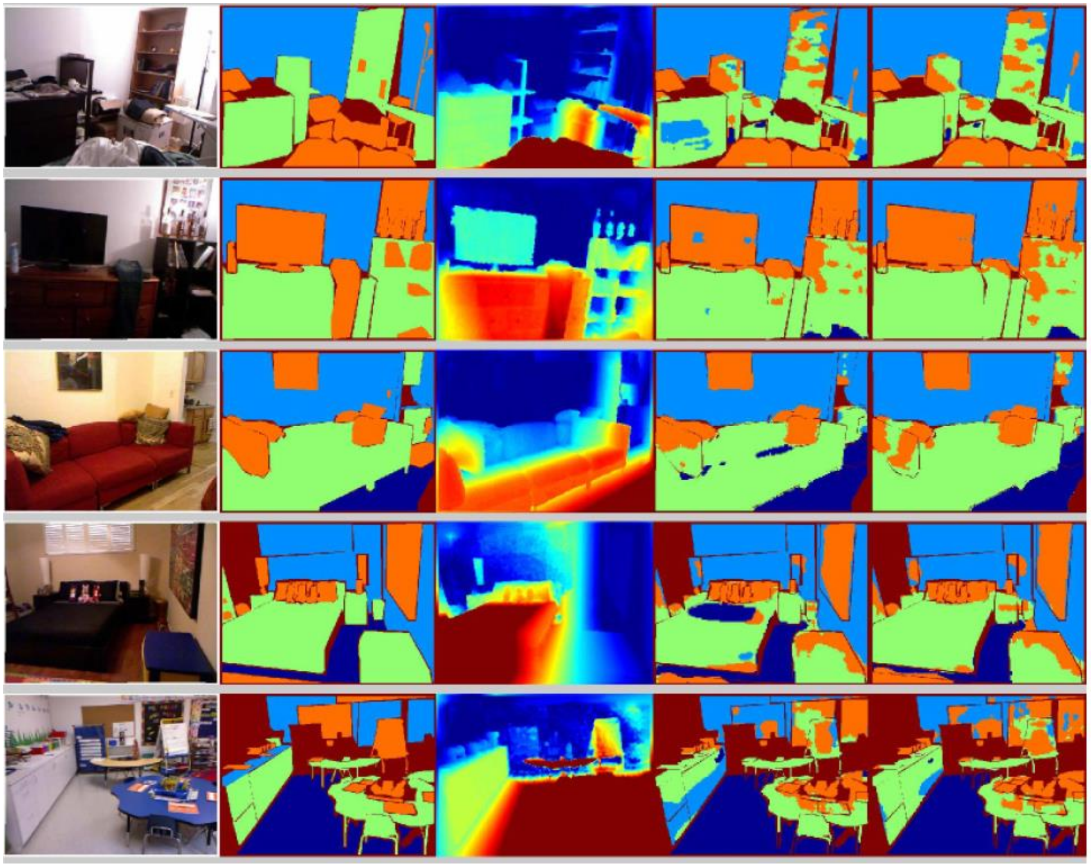
[Pavel, Schulz, Behnke, Neural Networks 2017]

Method	Class Accuracies (%)				Average (%)	
	ground	struct	furnit	prop	Class	Pixel
ours (CW, RGB-D only)	78.6	49.2	48.7	48.3	56.2	52.0
ours (CW)	95.8	74.6	54.2	**64.0**	72.1	68.6
ours (WI+CW)	94.9	76.8	65.5	60.8	74.5	73.1
ours (WI)	94.3	83.7	72.0	54.9	76.2	76.4
ours (WI+CW+CRF)	95.4	78.9	67.3	60.8	75.6	74.6
ours (WI+CRF)	94.2	83.9	**72.0**	56.3	**76.6**	**77.2**
all-frames	**97.2**	70.0	51.1	56.0	68.6	64.6
Schulz et al. (2015a) (CNN+CRF)	93.6	80.2	66.4	54.9	73.7	73.4
Müller and Behnke (2014) (RF+CRF)	94.9	78.9	**71.1**	42.7	71.9	72.3
Stückler et al. (2013) (RF+SLAM)	90.8	81.6	67.9	19.9	65.0	68.3
Couprie et al. (2013) (CNN)	87.3	**86.1**	45.3	35.5	63.5	64.5
Silberman et al. (2012) (RF)	68.0	59.0	70.0	42.0	59.6	58.6

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

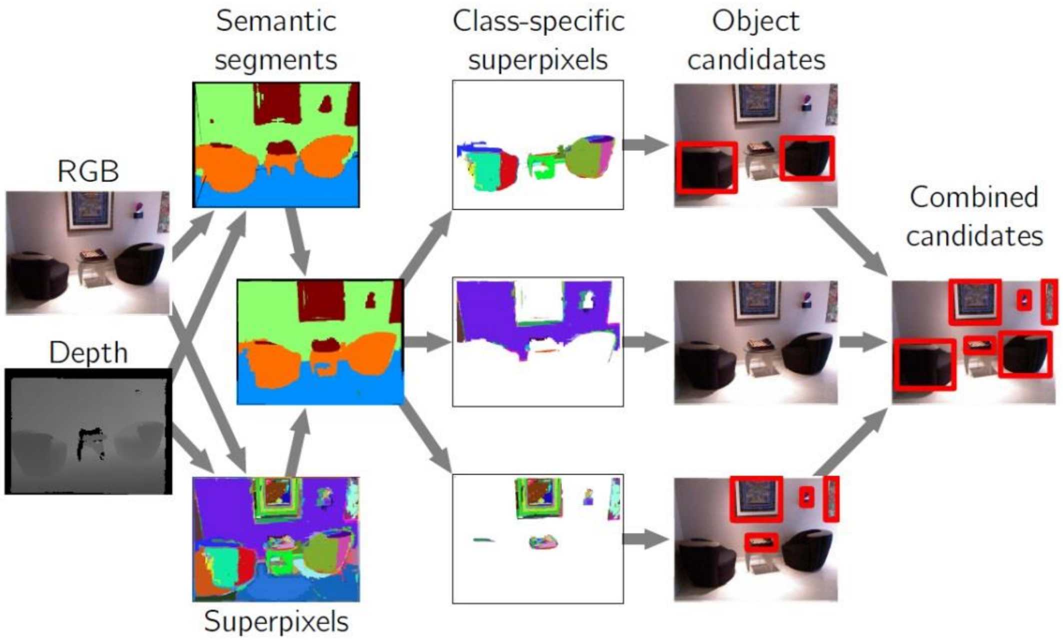
- New **geometric** feature: distance from wall
- **Semantic** features pretrained from ImageNet
- Both help significantly

[Husain et al. RA-L 2016]



Semantic Segmentation Priors for Object Discovery

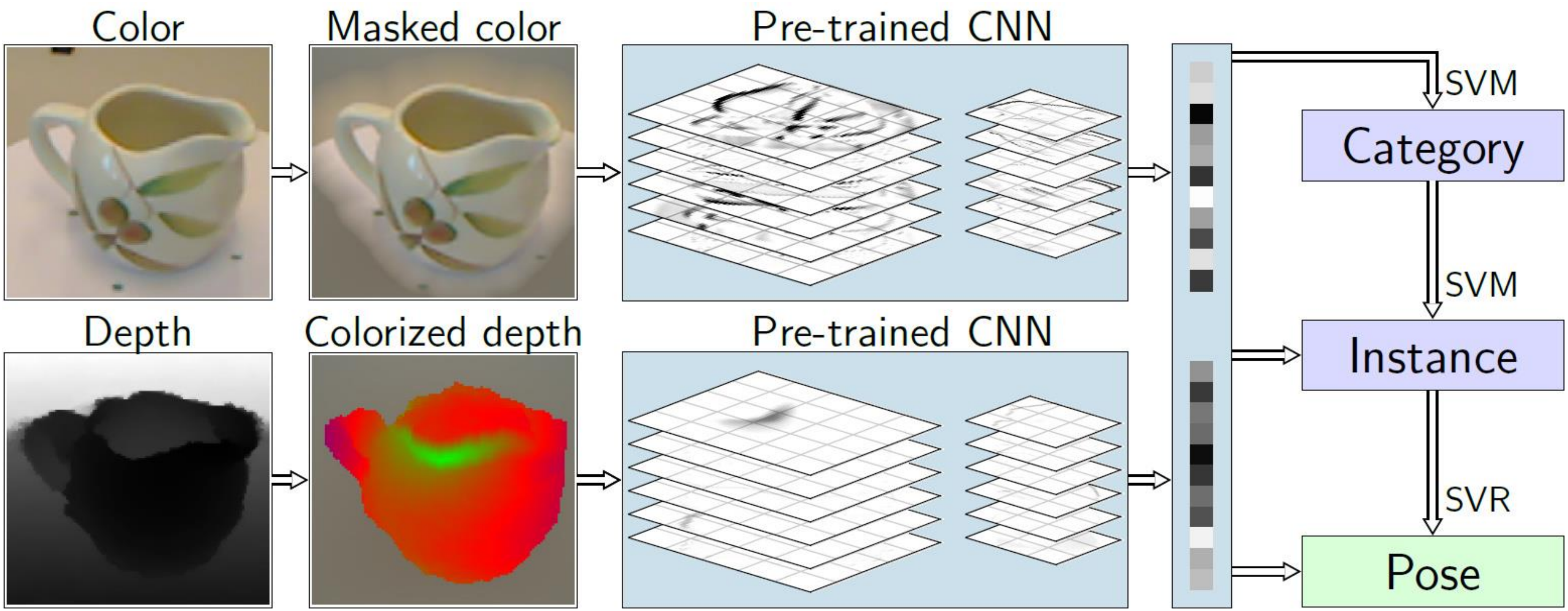
- Combine bottom-up object discovery and semantic priors
- Semantic segmentation used to classify color and depth superpixels
- Higher recall, more precise object borders



[Garcia et al. ICPR 2016]

RGB-D Object Recognition and Pose Estimation

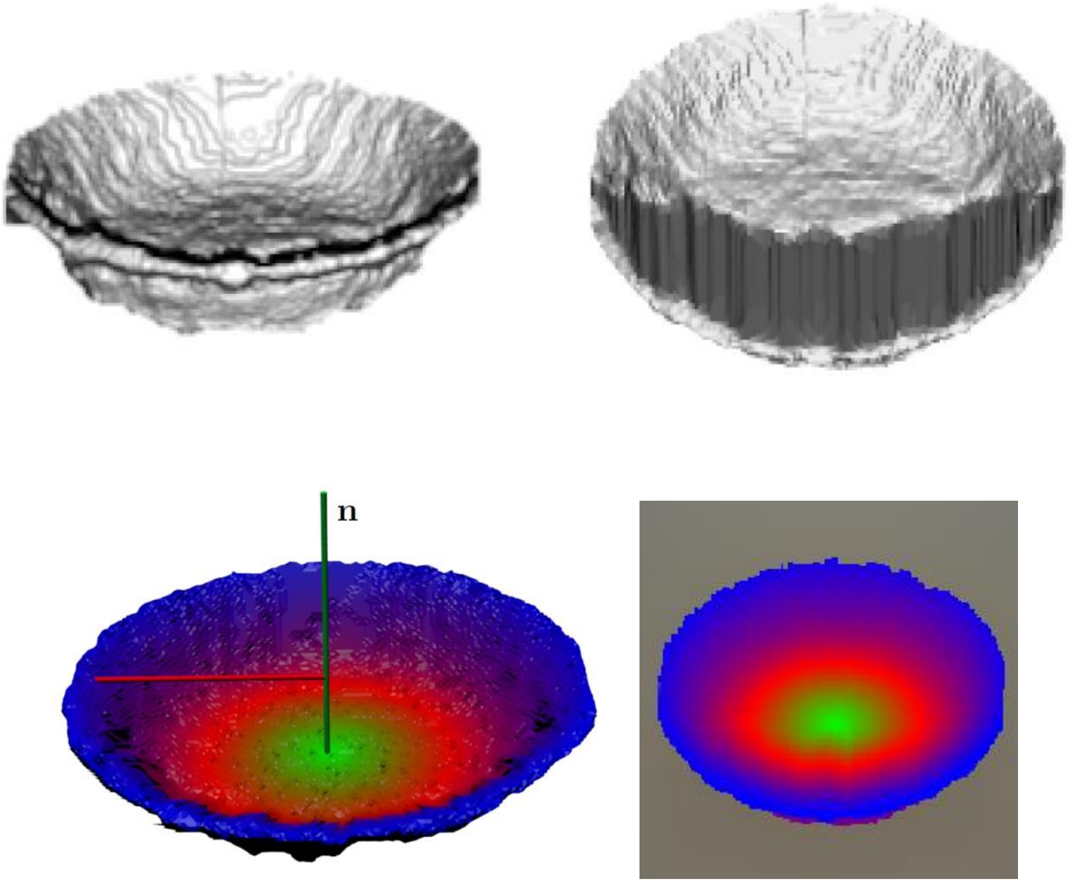
- Use pretrained features from ImageNet



[Schwarz, Schulz, Behnke, ICRA2015]

Canonical View, Colorization

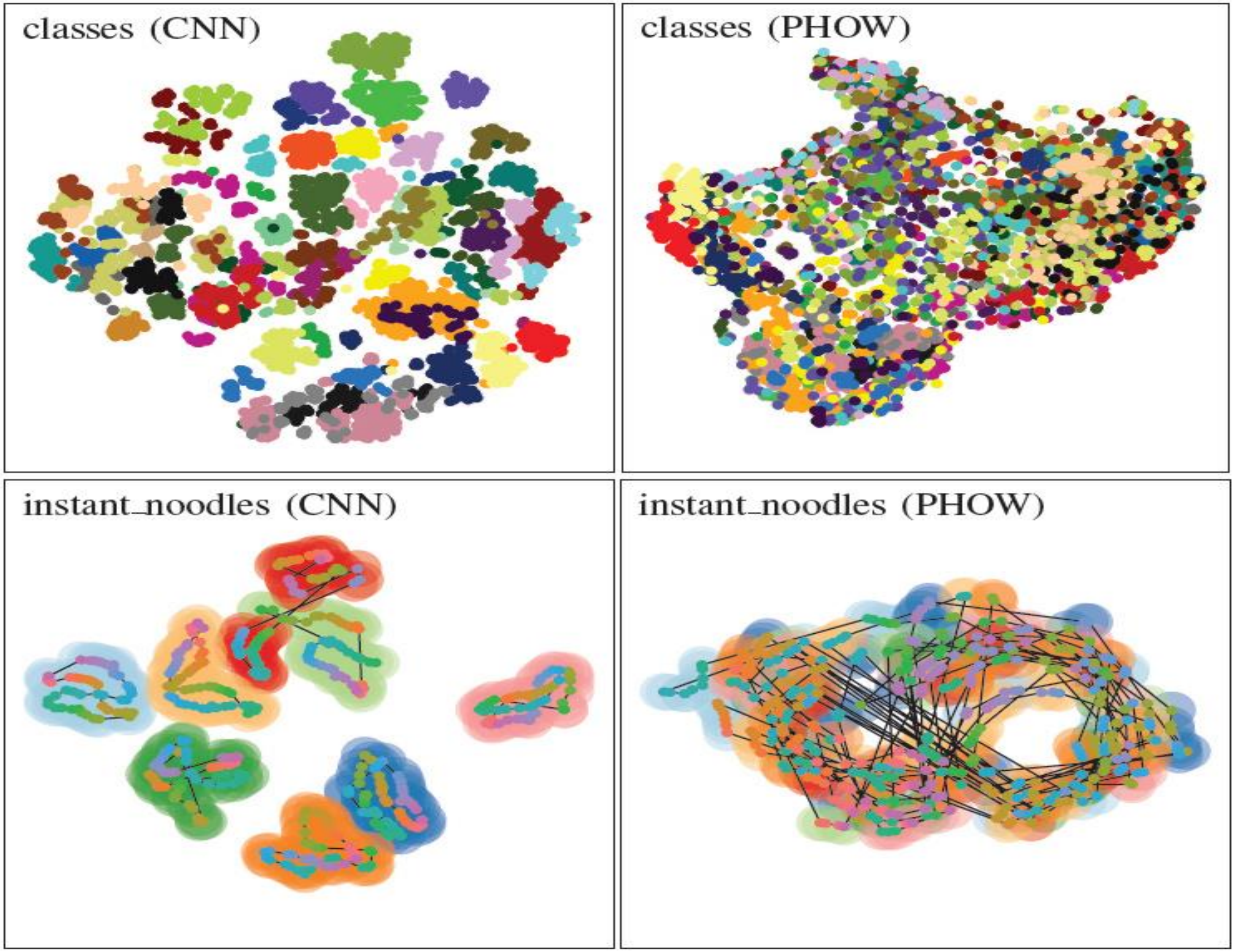
- Objects viewed from different elevation
- Render canonical view
- Colorization based on distance from center vertical



[Schwarz, Schulz, Behnke, ICRA2015]

Features Disentangle Data

- t-SNE embedding



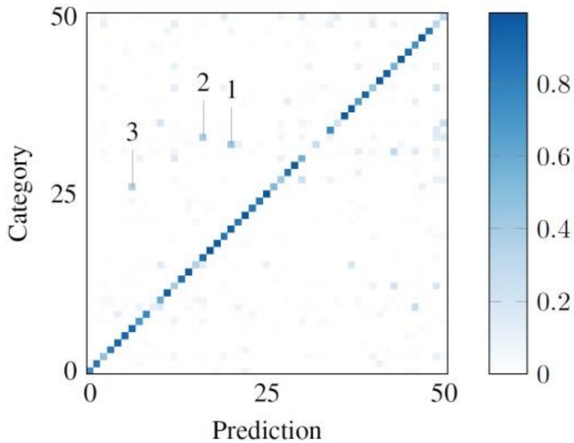
[Schwarz, Schulz, Behnke ICRA2015]

Recognition Accuracy

- Improved both category and instance recognition

Method	Category Accuracy (%)		Instance Accuracy (%)	
	RGB	RGB-D	RGB	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	**83.1 ± 2.0**	88.3 ± 1.5	92.0	**94.1**
Ours	**83.1 ± 2.0**	**89.4 ± 1.3**	92.0	**94.1**

- Confusion

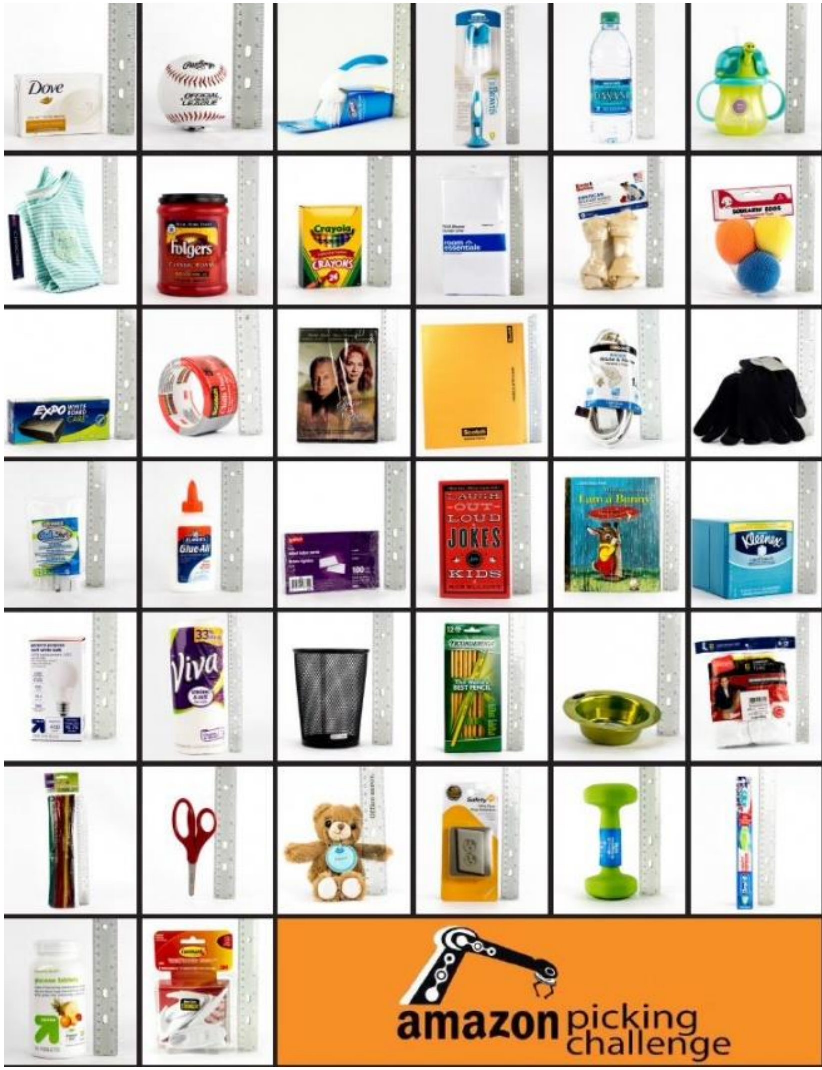


1: pitcher / coffe mug 2: peach / sponge

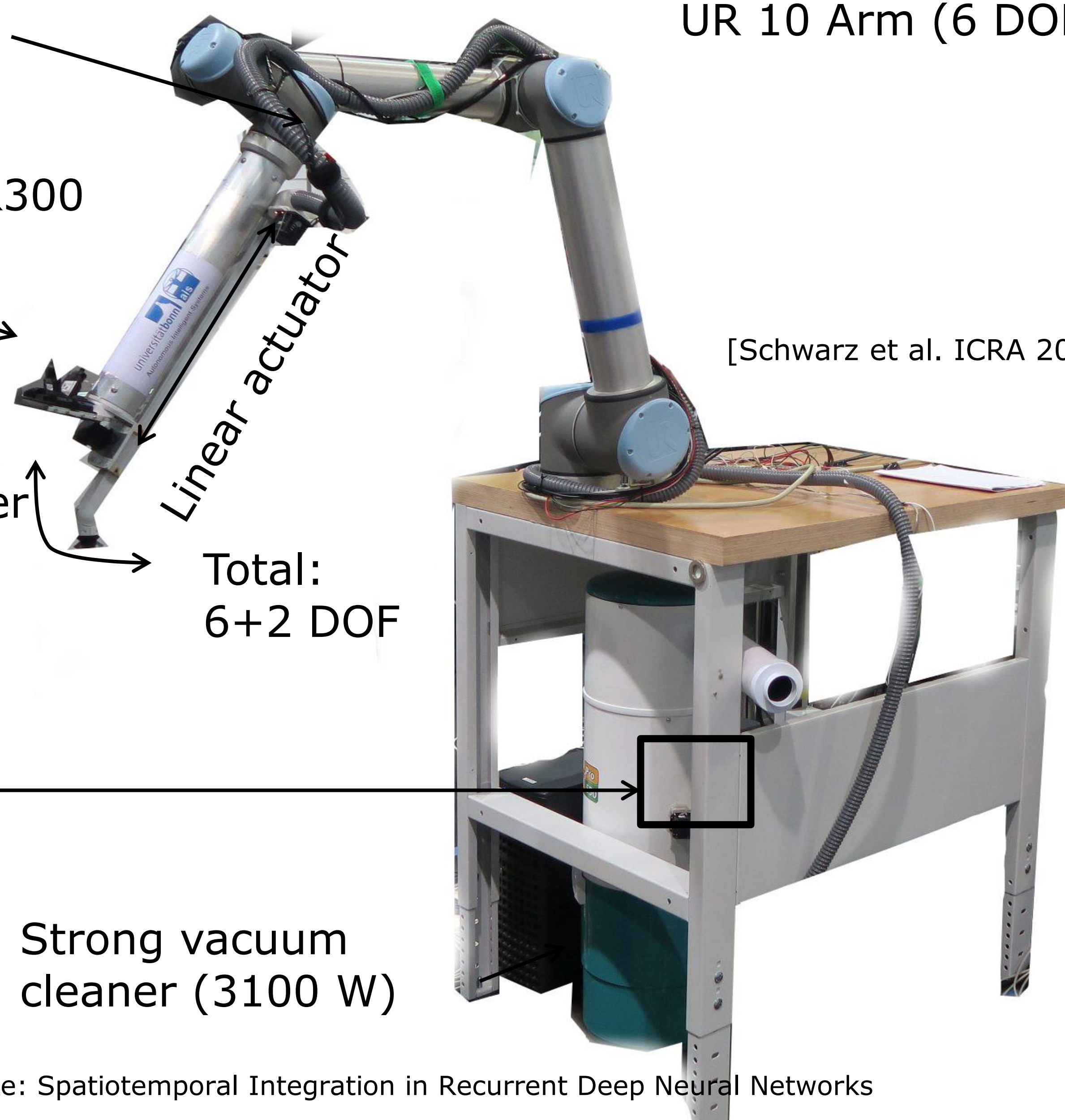
[Schwarz, Schulz, Behnke, ICRA2015]

Amazon Picking Challenge 2016

- Large variety of objects
- Different properties
  - Transparent
  - Shiny
  - Deformable
  - Heavy
- Stowing task
- Picking task



System



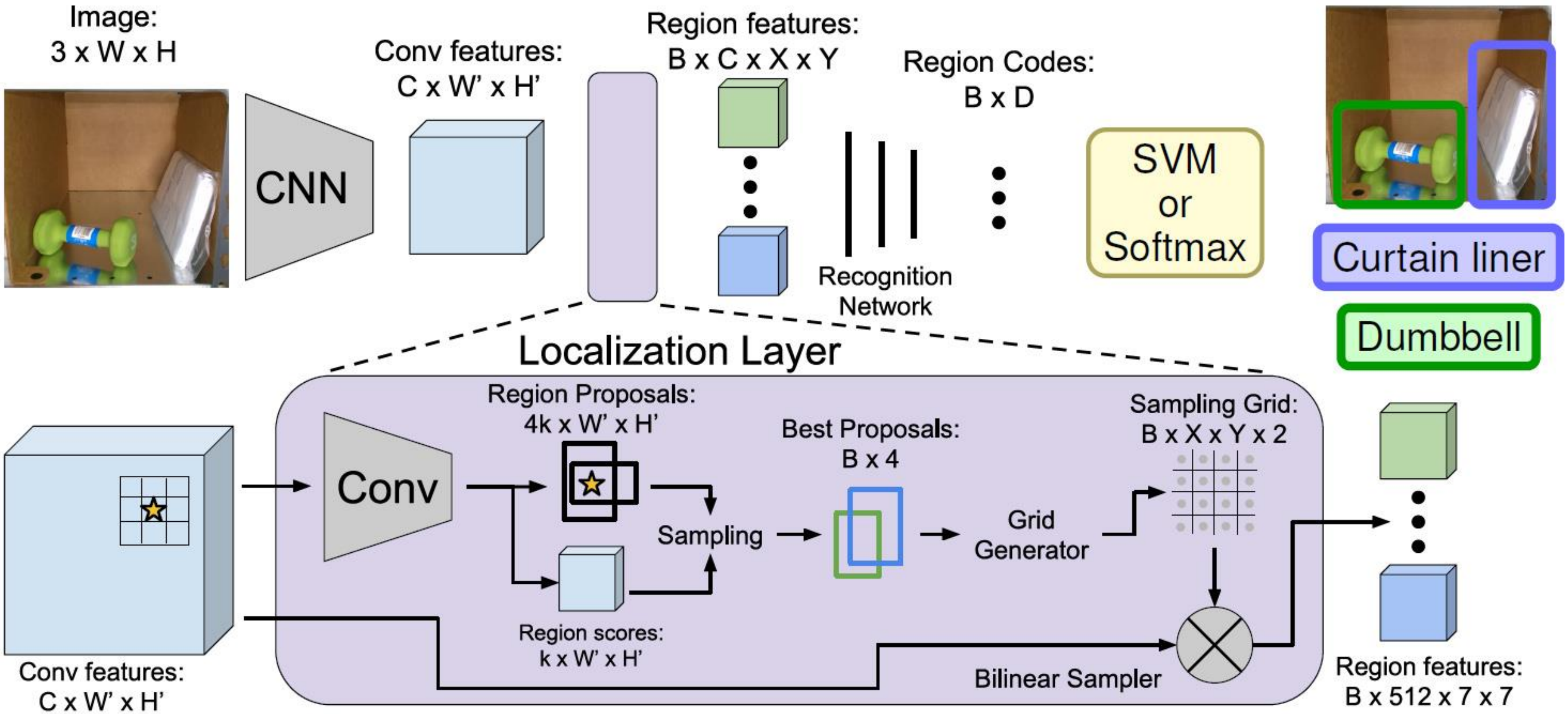
Suction strength control

RGB-D Cameras

- 2x Intel RealSense SR300
- Fusion of three depth estimates per pixel (including RGB stereo)

[Schwarz et al. ICRA 2017]

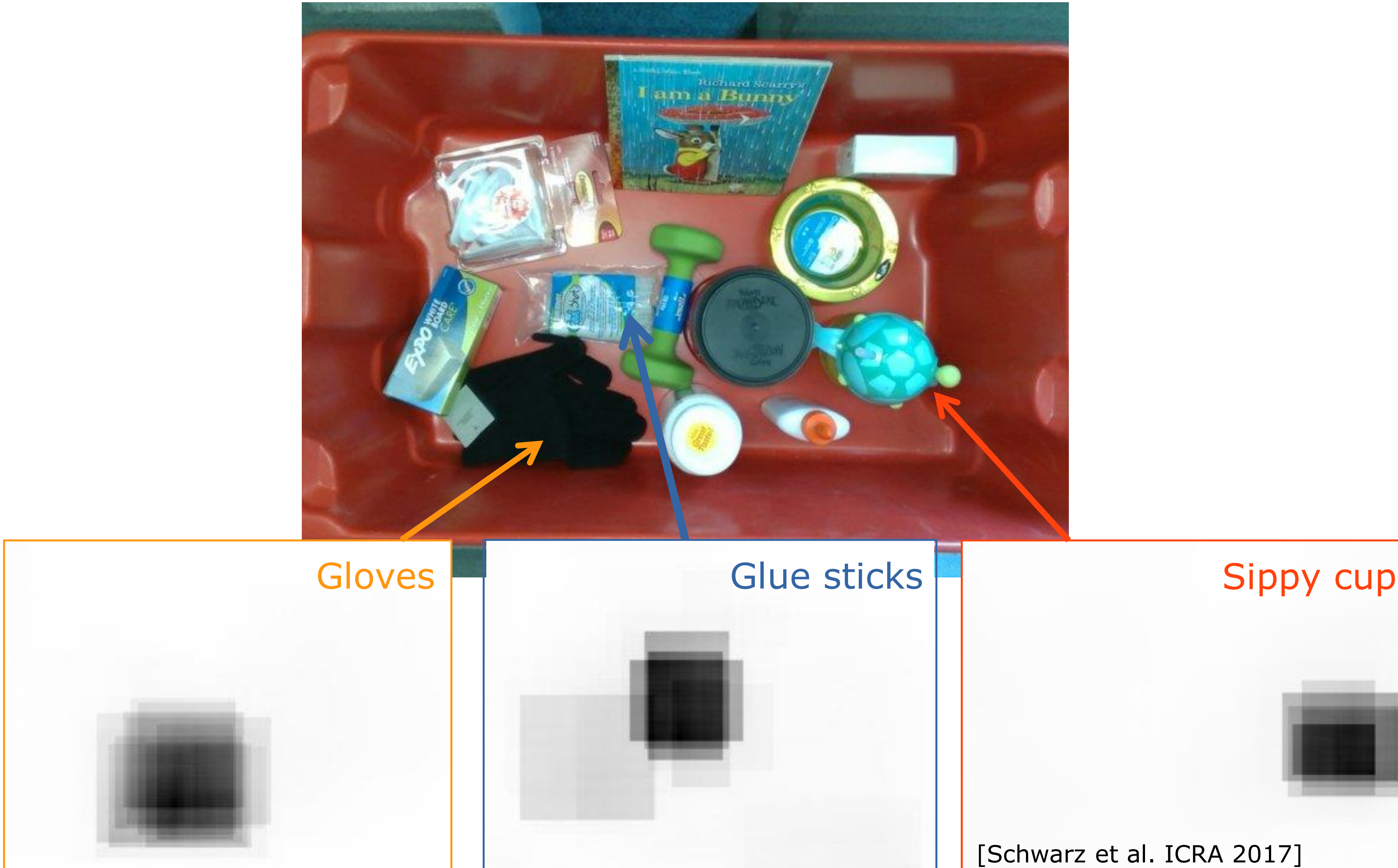
Object Detection



[Adapted from Johnson et al. CVPR 2016]

[Schwarz et al. ICRA 2017]

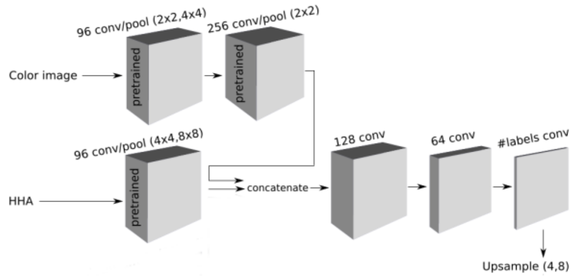
Example Detections



[Schwarz et al. ICRA 2017]

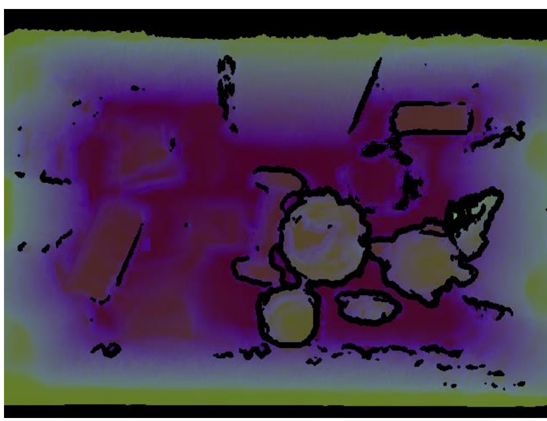
Semantic Segmentation

- Deep Convolutional Network



RGB

Result

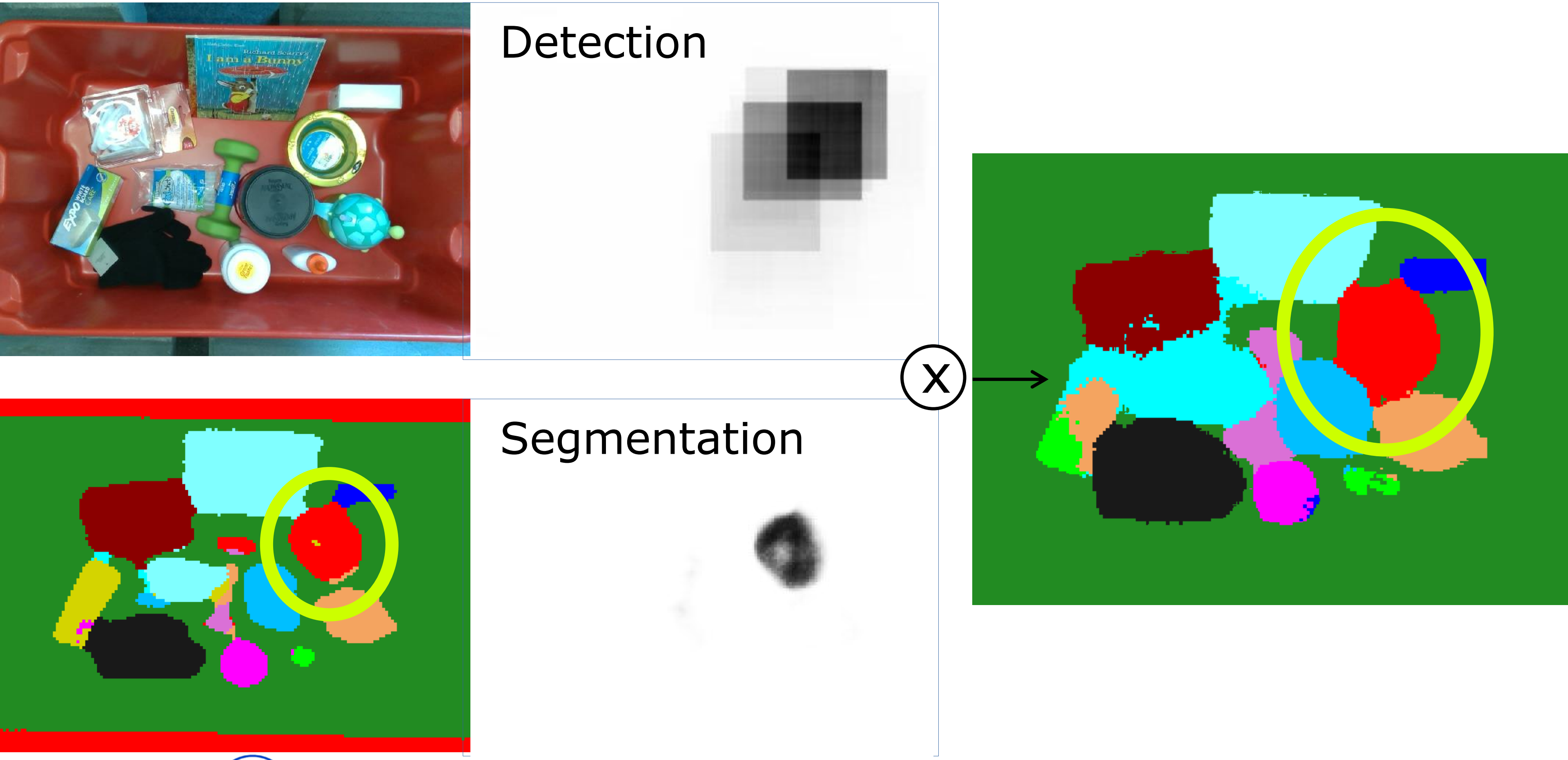


HHA

[Husain et al. RA-L 2016]

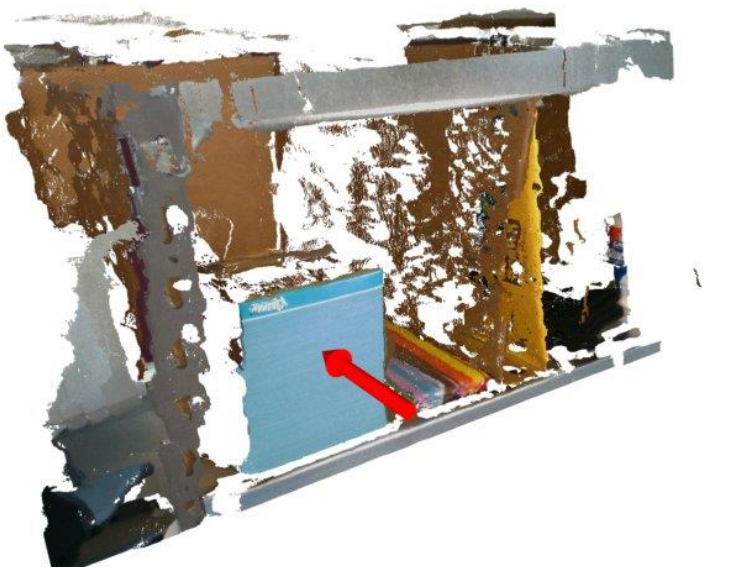
Combined Detection and Segmentation

- Pixel-wise multiplication



Grasp Pose Selection

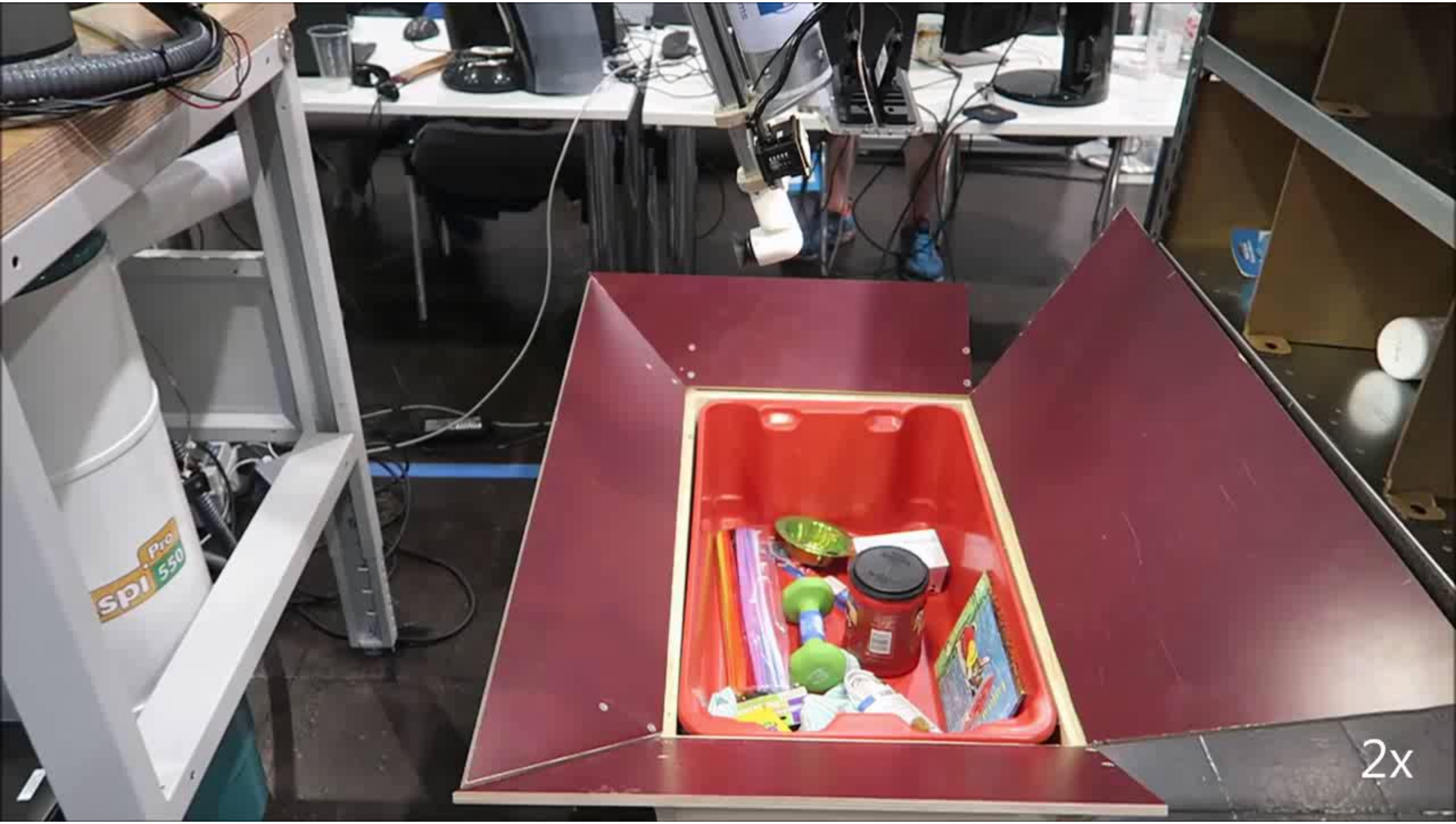
- Center grasp for "standing" objects:
  - Find support area for suction close to bounding box center



- Top grasp for "lying" objects:
  - Find support area for suction close to horizontal bounding box center

[Schwarz et al. ICRA 2017]

Example Stowing Top Grasp



[Schwarz et al. ICRA 2017]

Example Picking Grasps

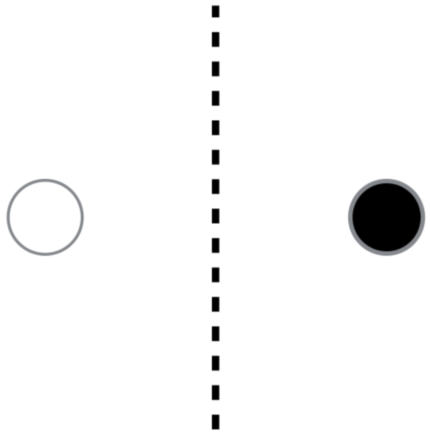
[Schwarz et al. ICRA 2017]

Problems of Supervised Training

- Non-convex optimization (local minima)
- Overfitting => Regularization
- Availability of labels

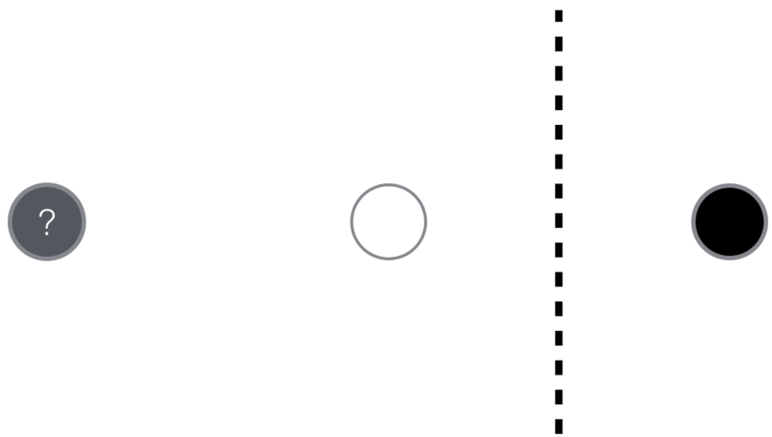
Semi-supervised Learning

- Given two labeled examples



Semi-supervised Learning

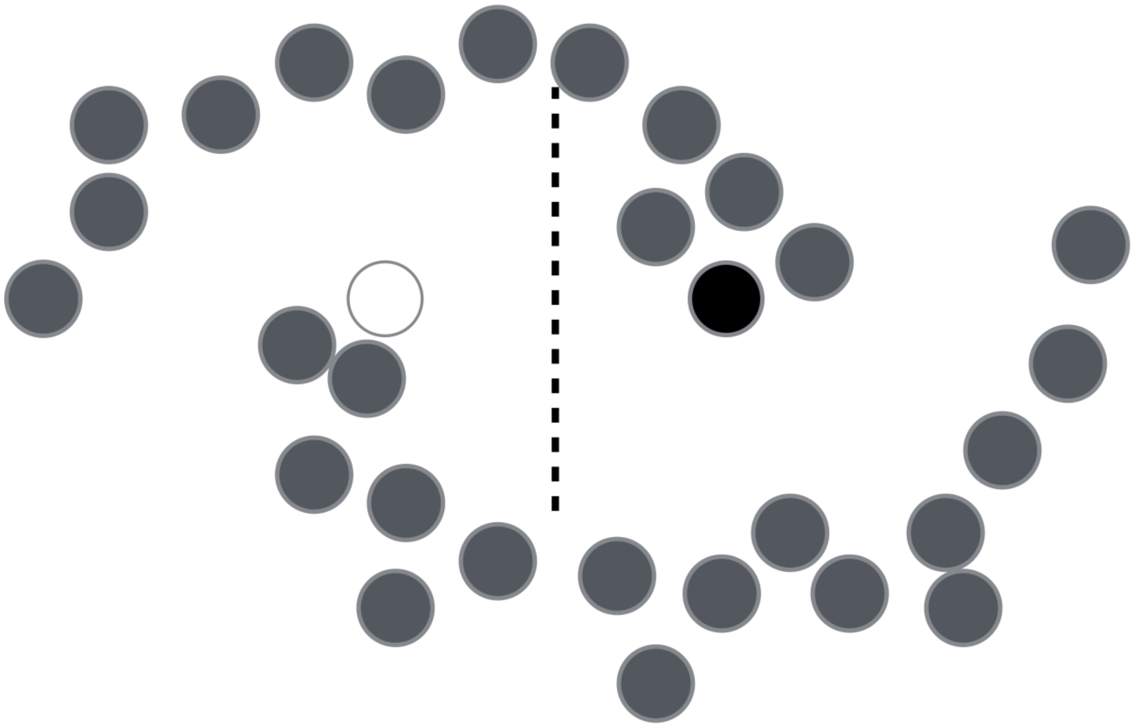
- How to label this point?



[Raiko]

Semi-supervised Learning

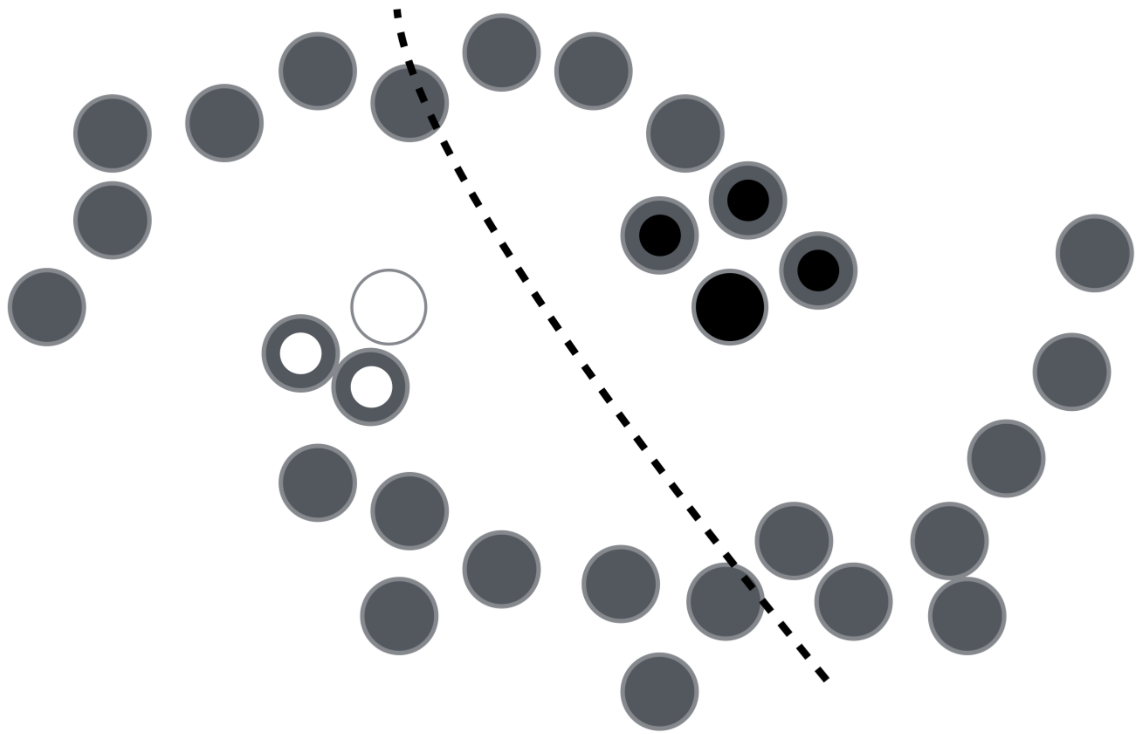
- What if you see all the unlabeled data?



[Raiko]

Semi-supervised Learning

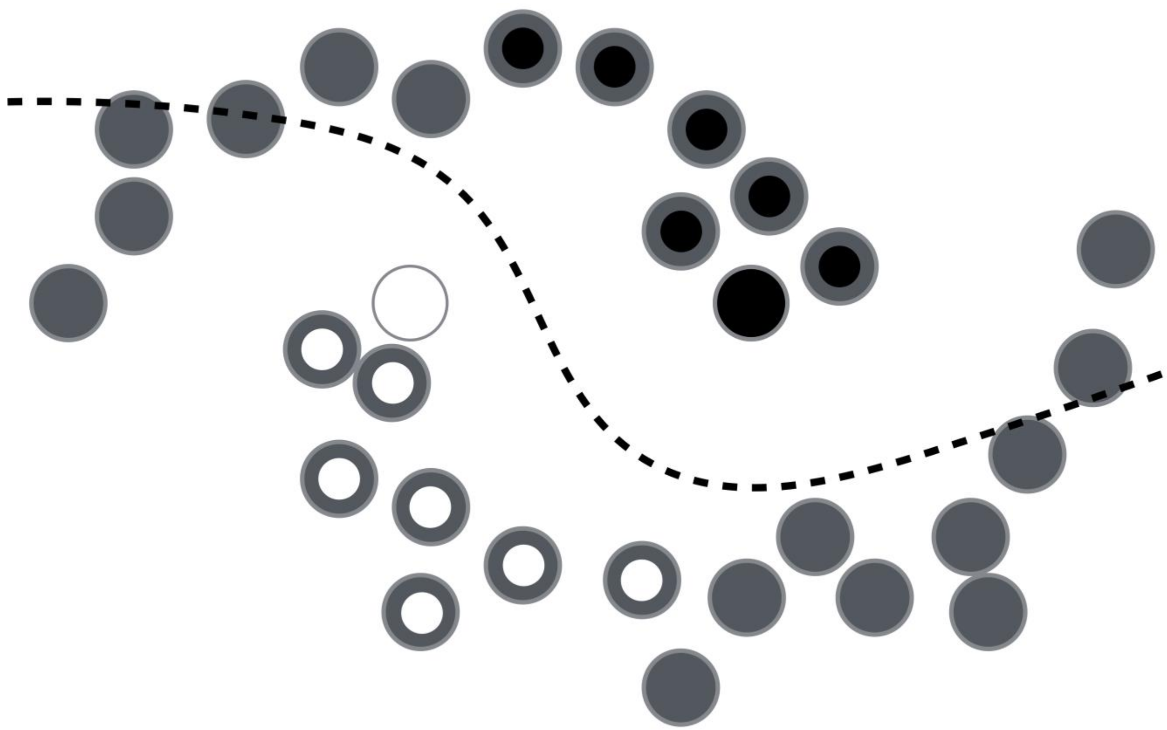
- Labels homogeneous in densely populated space



[Raiko]

Semi-supervised Learning

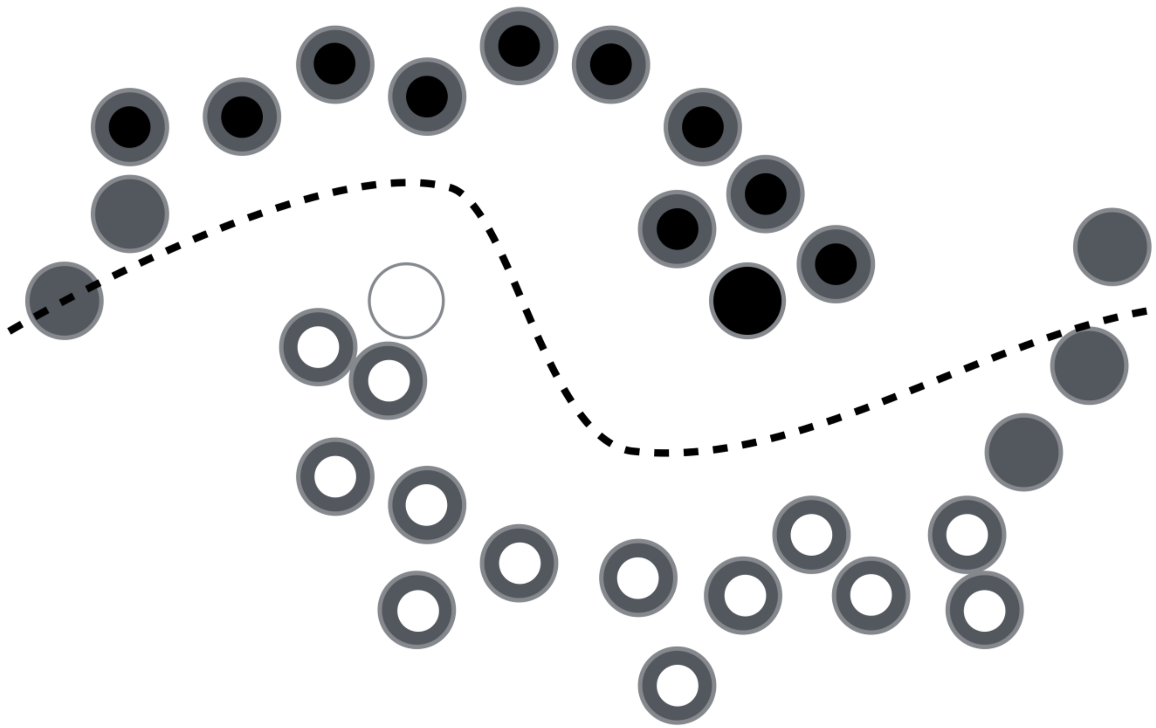
- Labels homogeneous in densely populated space



[Raiko]

Semi-supervised Learning

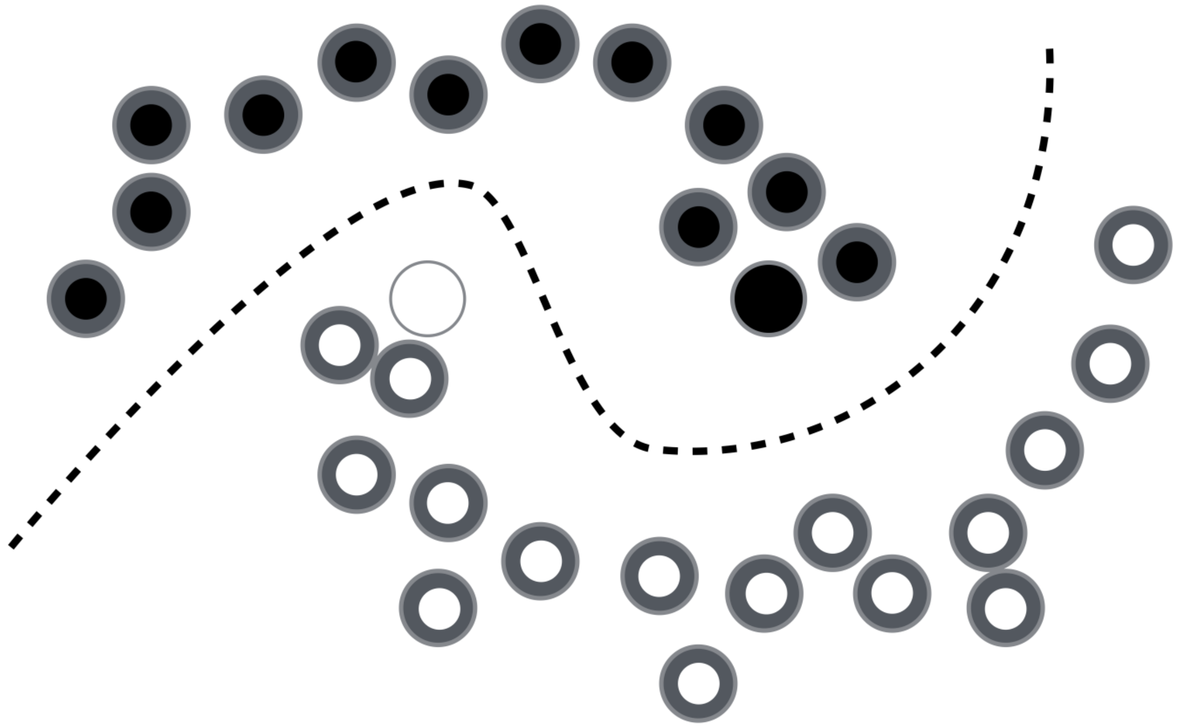
- Labels homogeneous in densely populated space



[Raiko]

Semi-supervised Learning

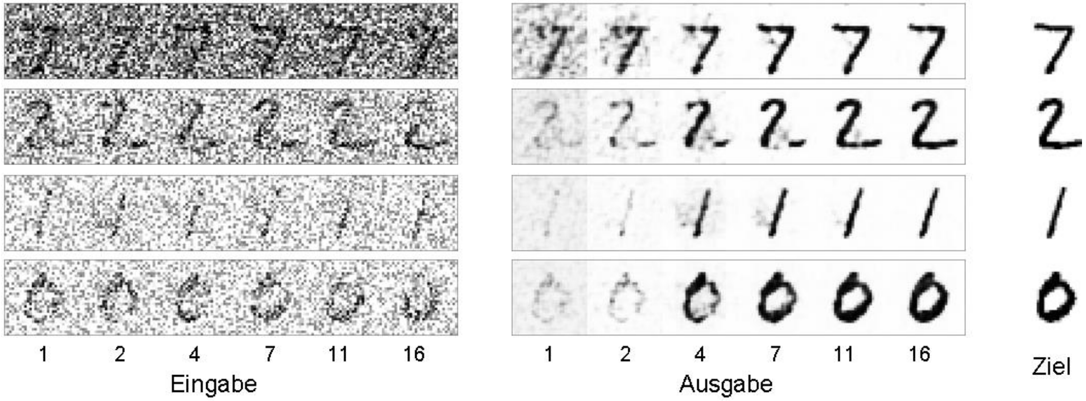
- Labels homogeneous in densely populated space



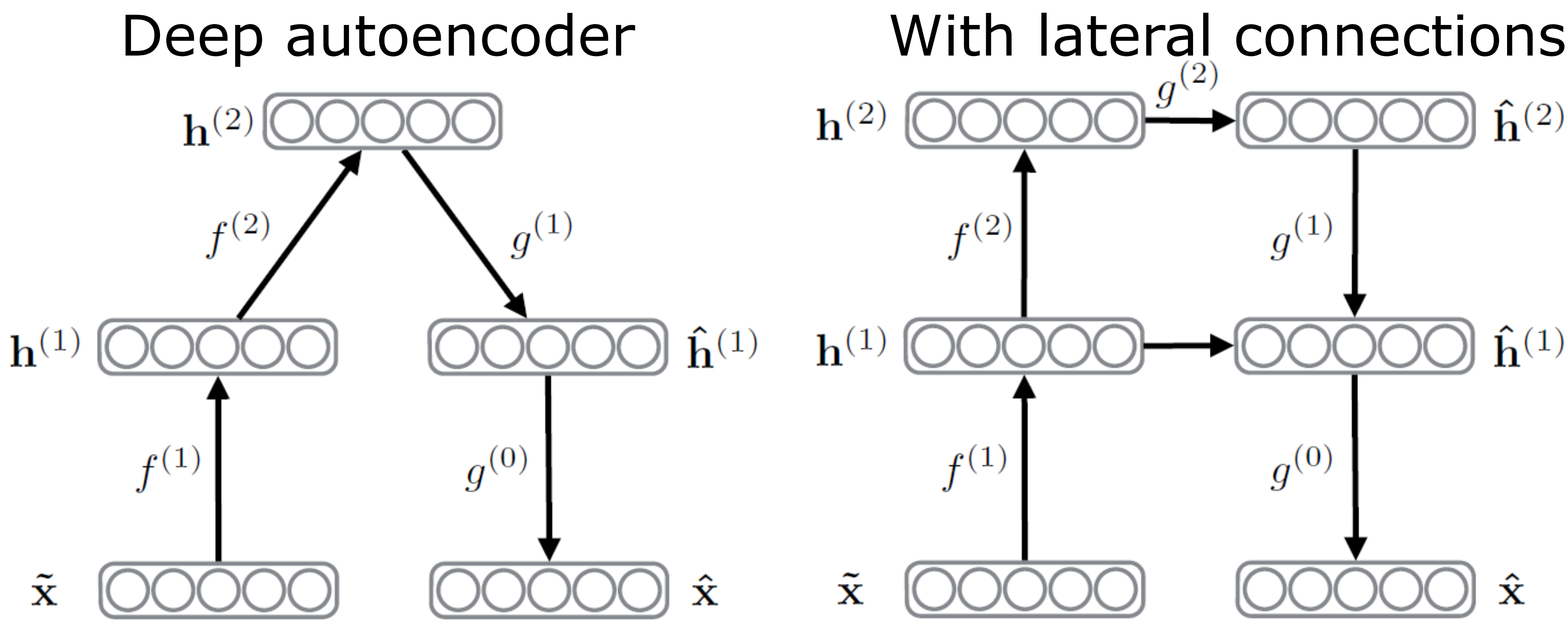
[Raiko]

Problems of Supervised Training

- Non-convex optimization (local minima)
- Overfitting => Regularization
- Availability of labels
  - One idea: generate target output without human annotation
    - Vestibulo occular reflex:
      minimize movement on the retina
    - Predictions: wait a little
    - Reconstruction: degrade the original



Adding Lateral Connections to Autoencoders

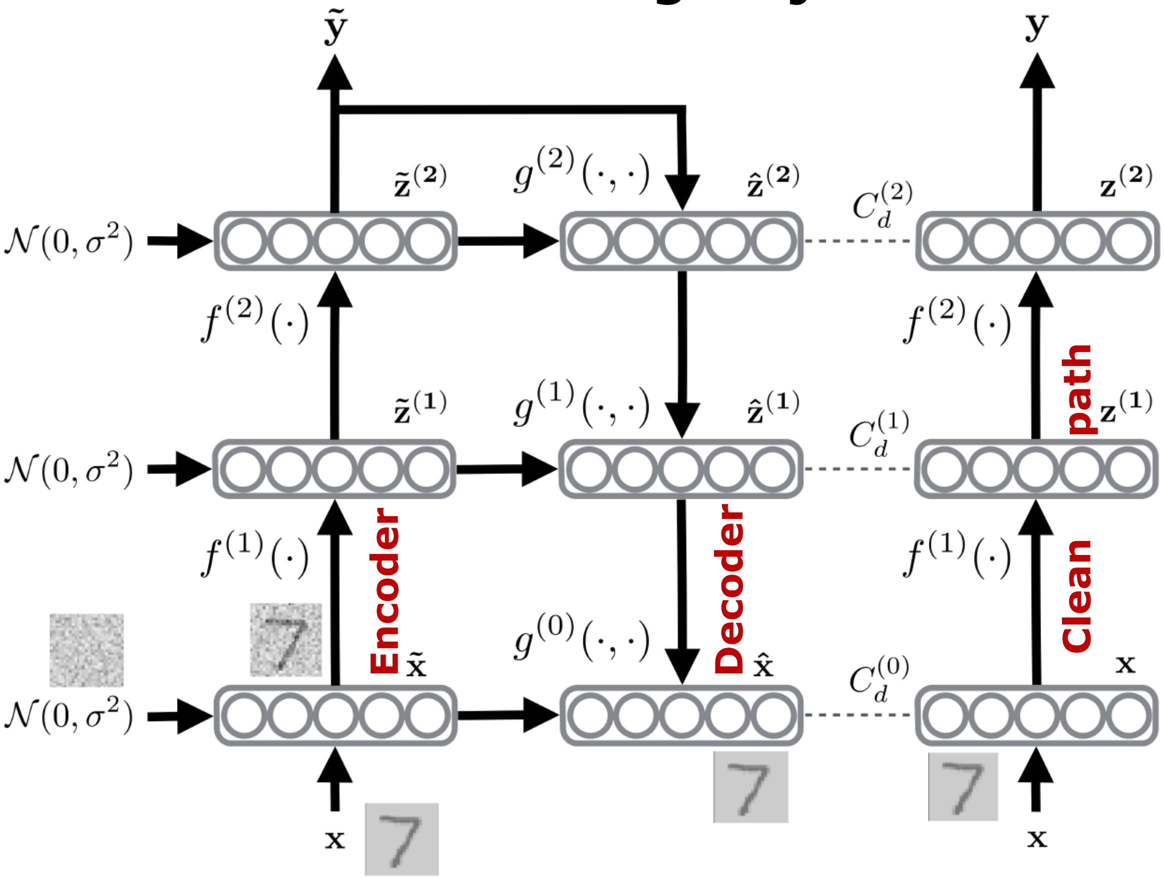


- No need to encode low-level features at higher layers

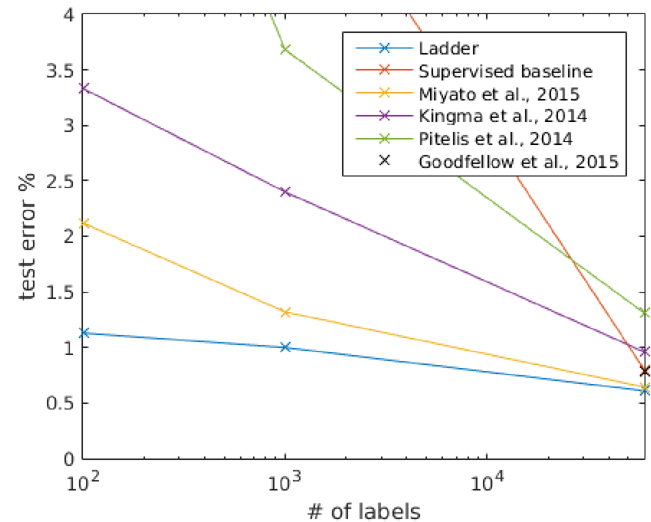
[Rasmus et al. 2015]

Semi-Supervised Learning with Ladder Networks

- Lateral connections between encoder and decoder
- Local denoising objectives on each layer



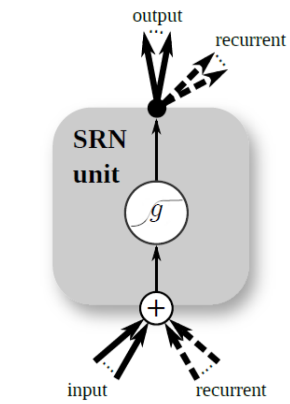
MNIST results



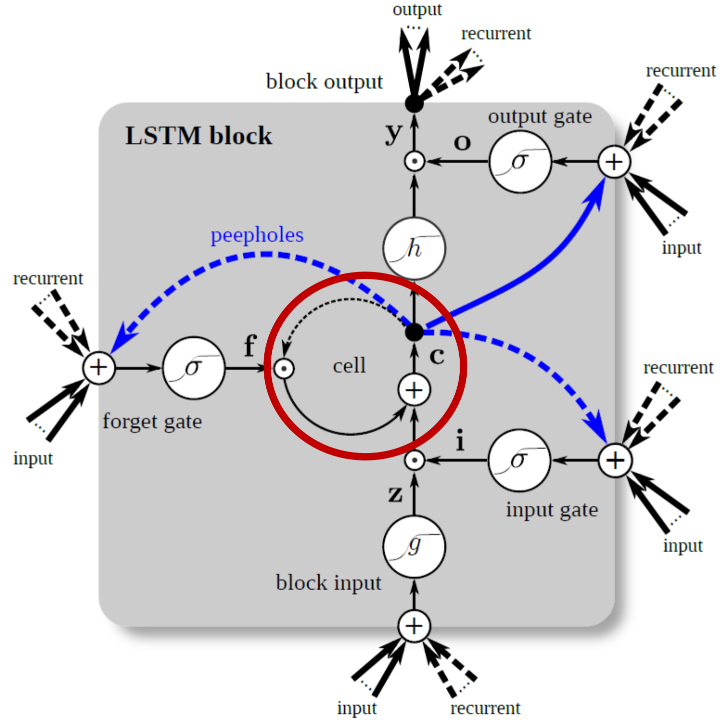
[Rasmus et al. NIPS 2015]

Long Short-Term Memory (LSTM)

- Simple recurrent neural networks (SRN) may suffer from **vanishing or exploding gradient** problem



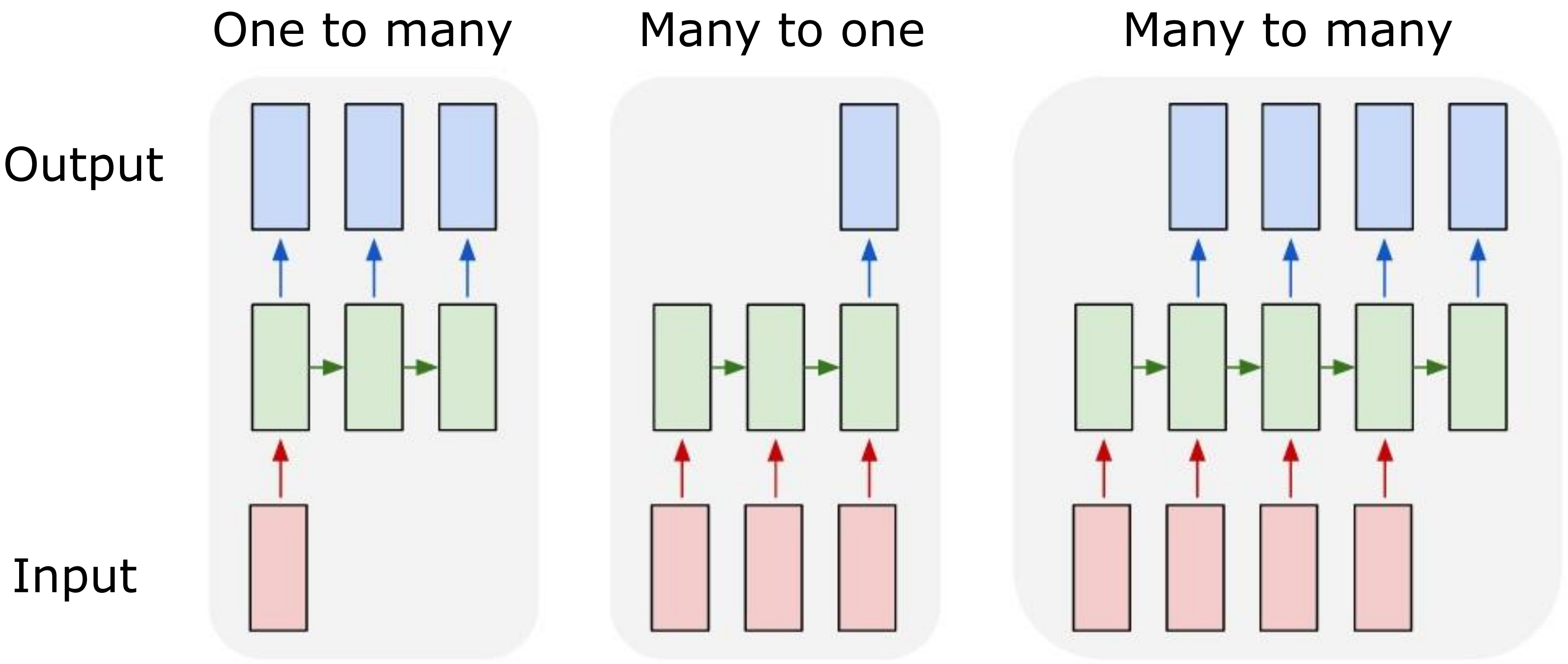
- LSTM hard-wires a **memory cell** and controls input, forgetting, and output with gates



[Hochreiter & Schmidhuber 1997, Graves & Schmidhuber 2005, Greff et al. 2015]

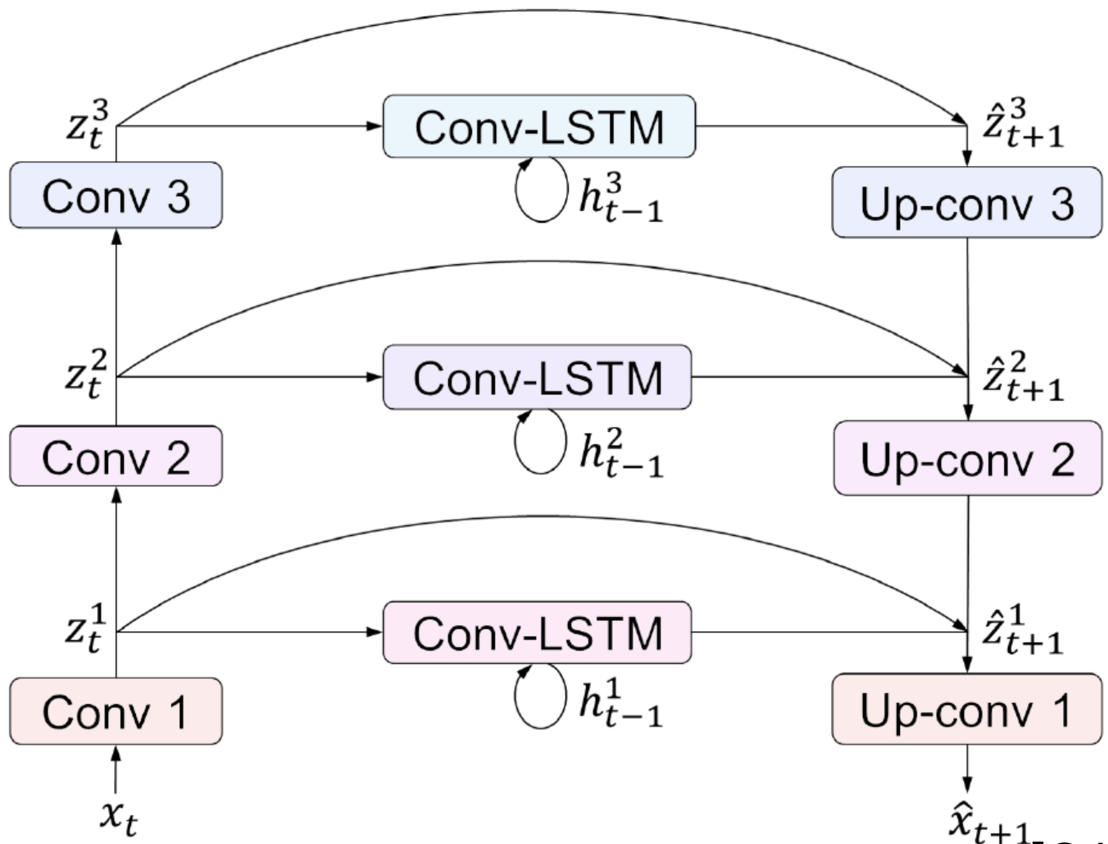
Processing of Sequences

- Recurrent neural networks are suitable for processing sequences



Video Ladder Network

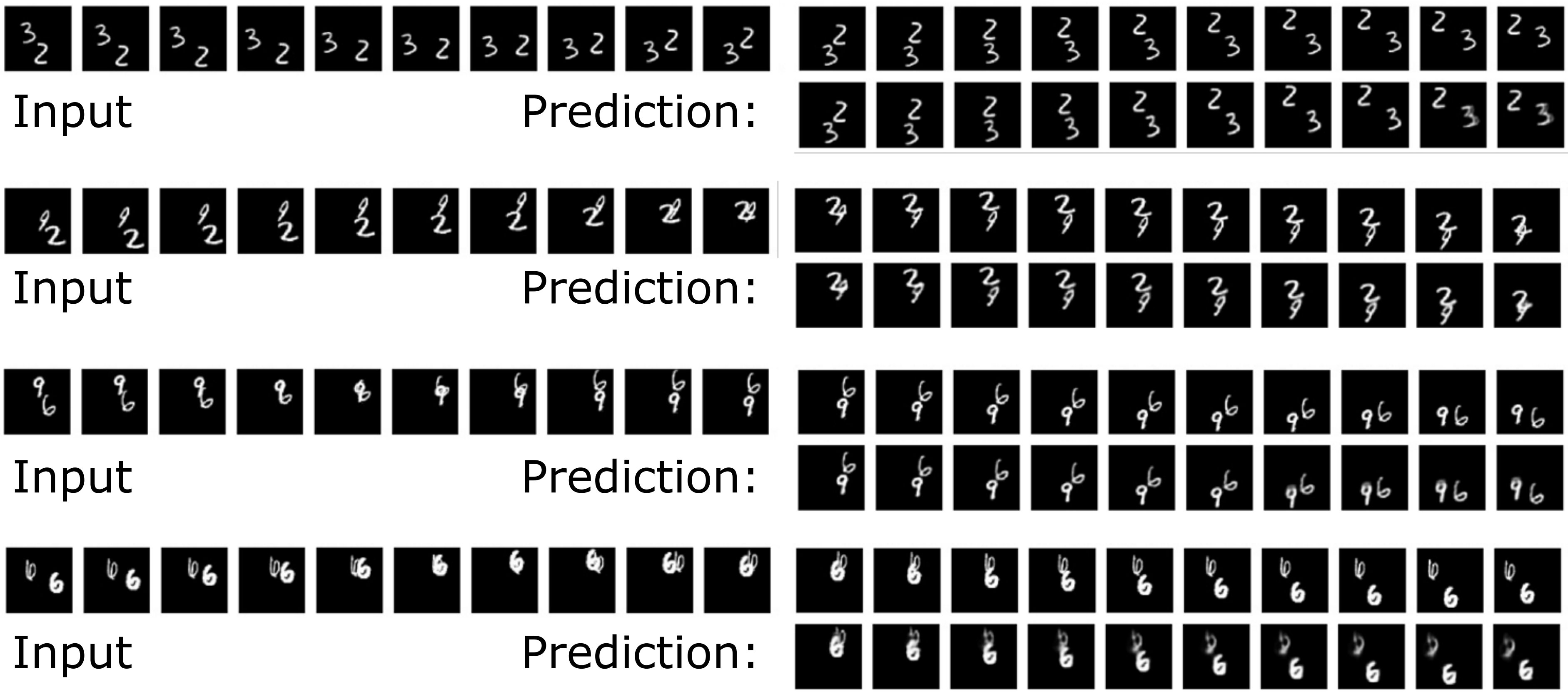
- Learning prediction with recurrent and feed-forward lateral connections



[Cricri et al. 2016]

Moving MNIST Digits

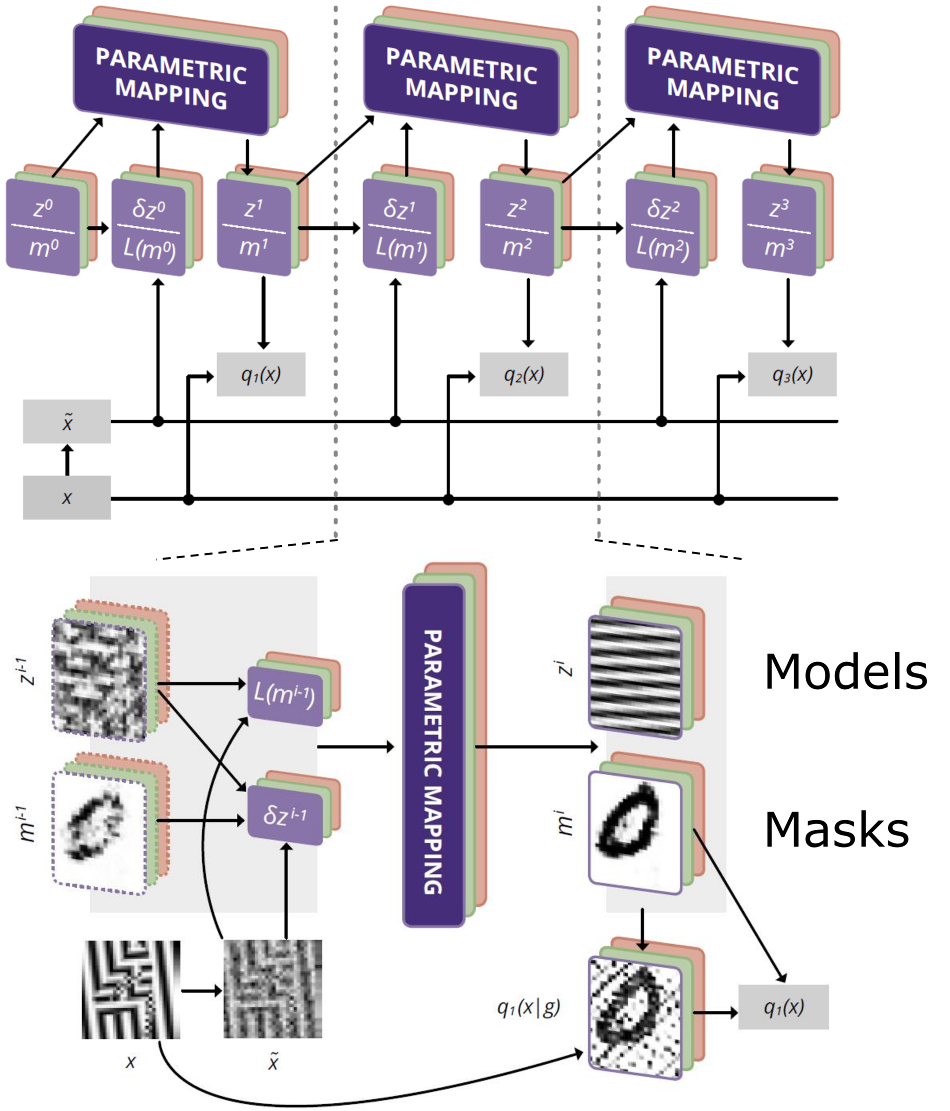
- Two random digits moving in random directions with constant speed, bouncing at border



[Cricri et al. 2016]

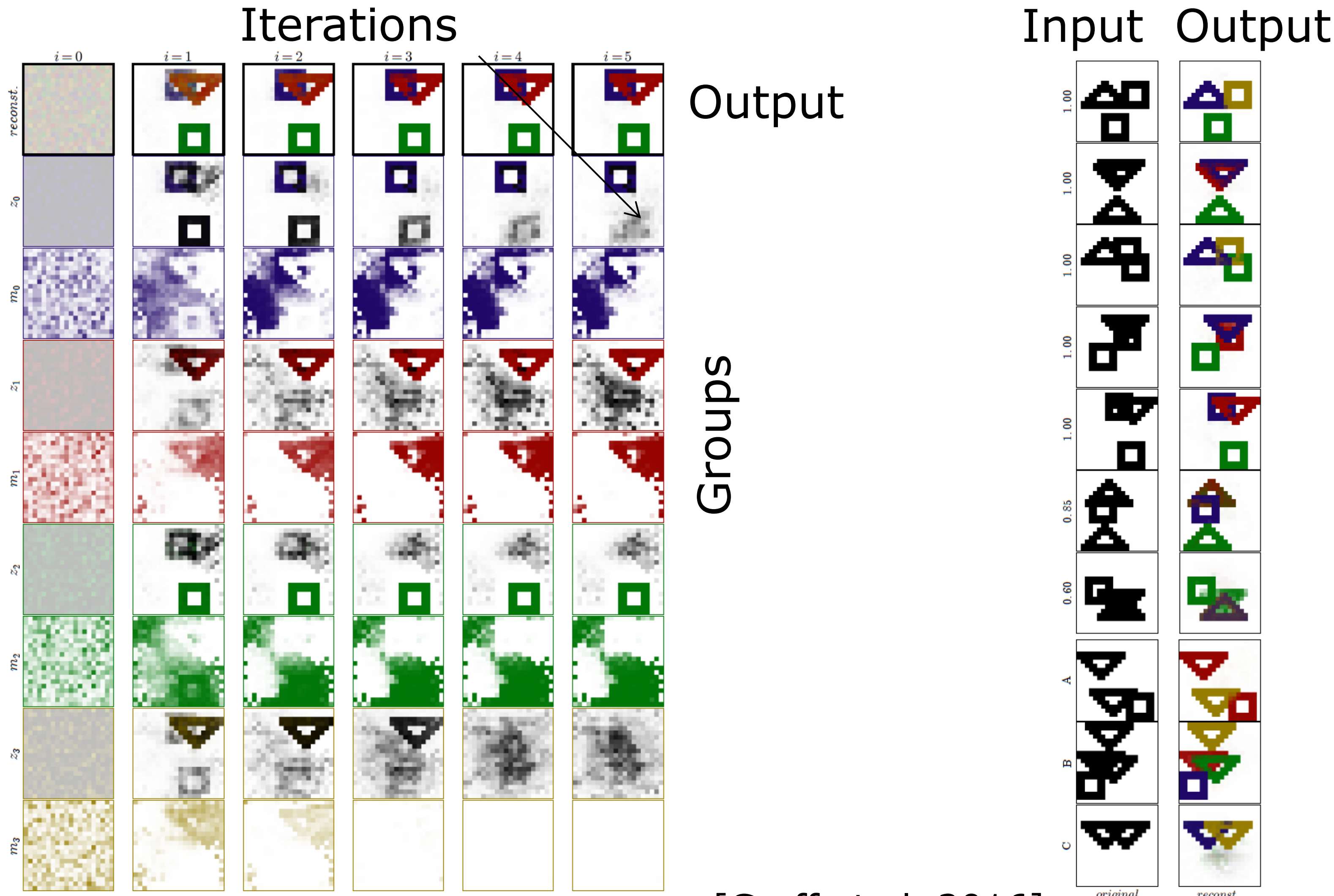
Tagger: Deep Unsupervised Perceptual Grouping

- Hard-wired group-wise modeling and iterative reconstruction
- Trained on reconstruction (denoising) loss
- Learns unsupervised grouping



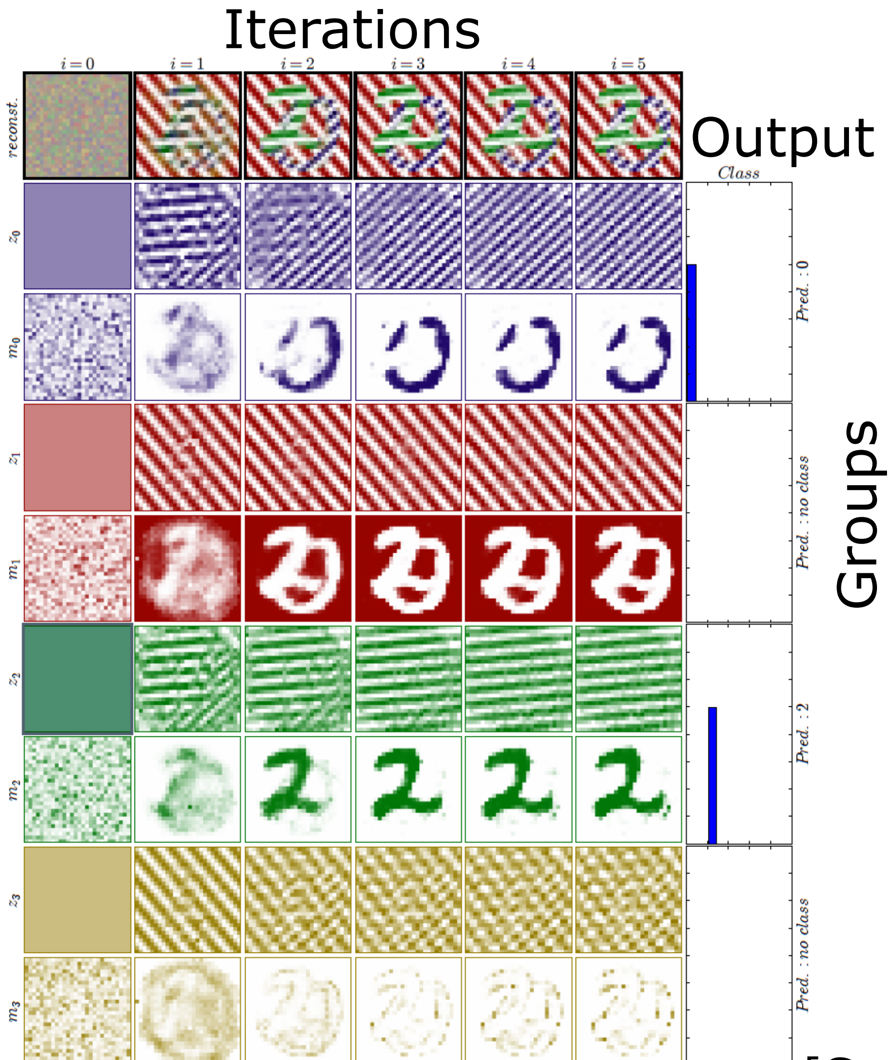
[Greff et al. 2016]

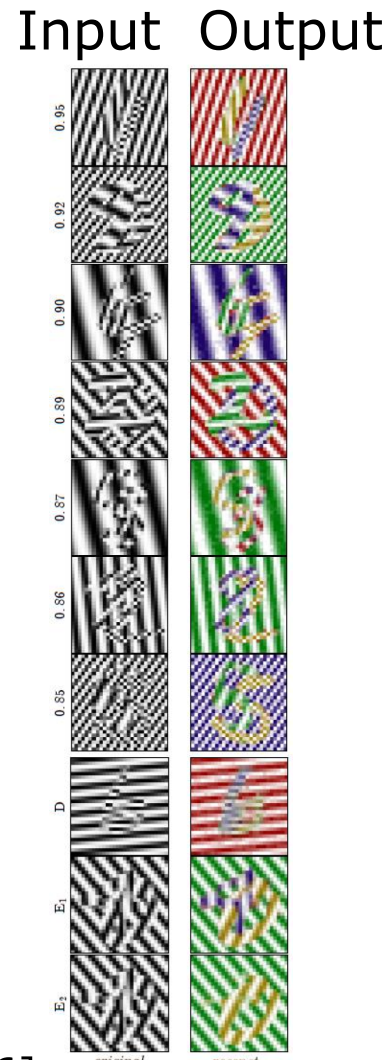
Iterative Grouping of Shapes



[Greff et al. 2016]

Grouping of Textured Digits

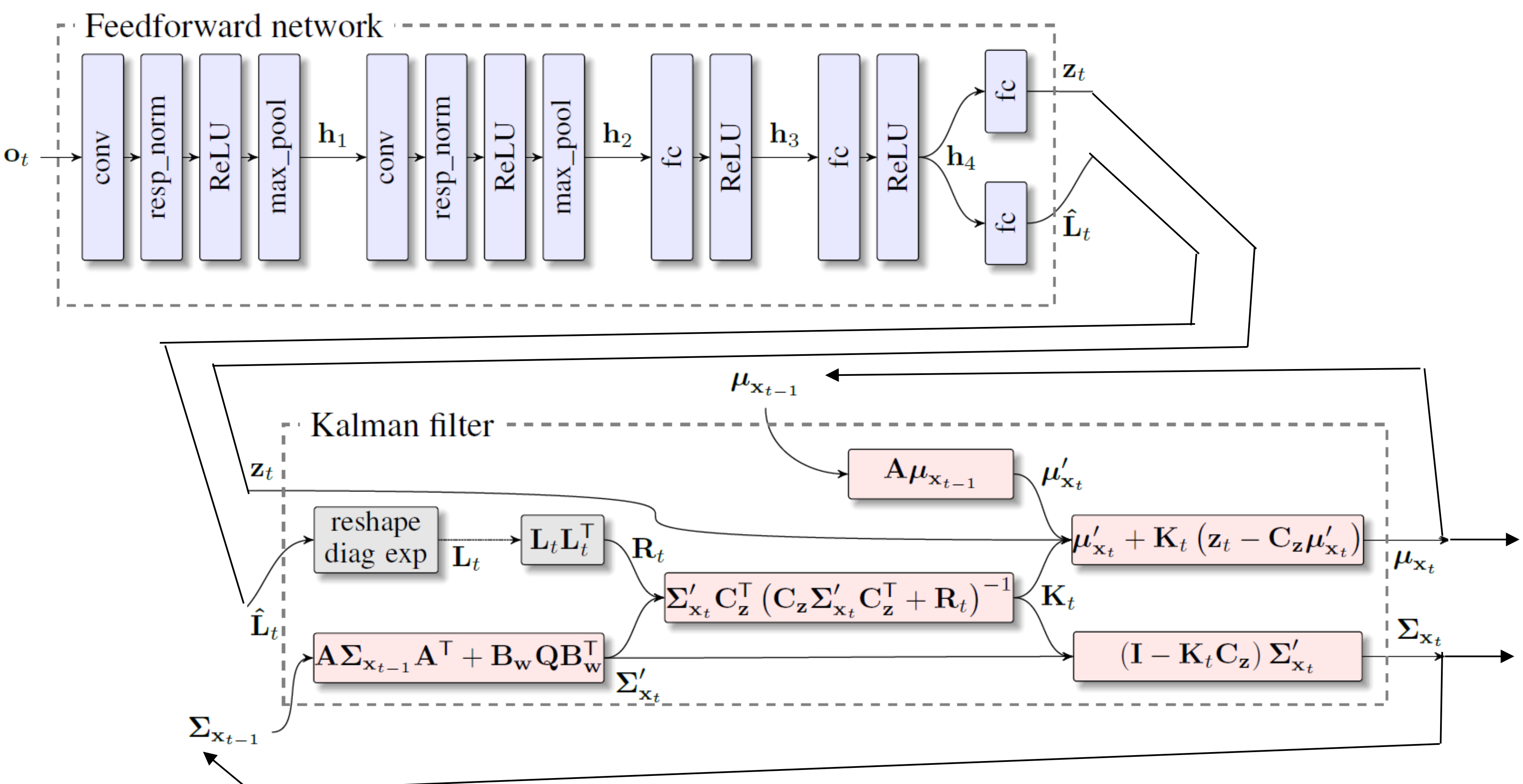




[Greff et al. 2016]

Backprop Kalman Filter

- End-to-end training learns latent z_t



[Haarnoja et al. NIPS 2016]

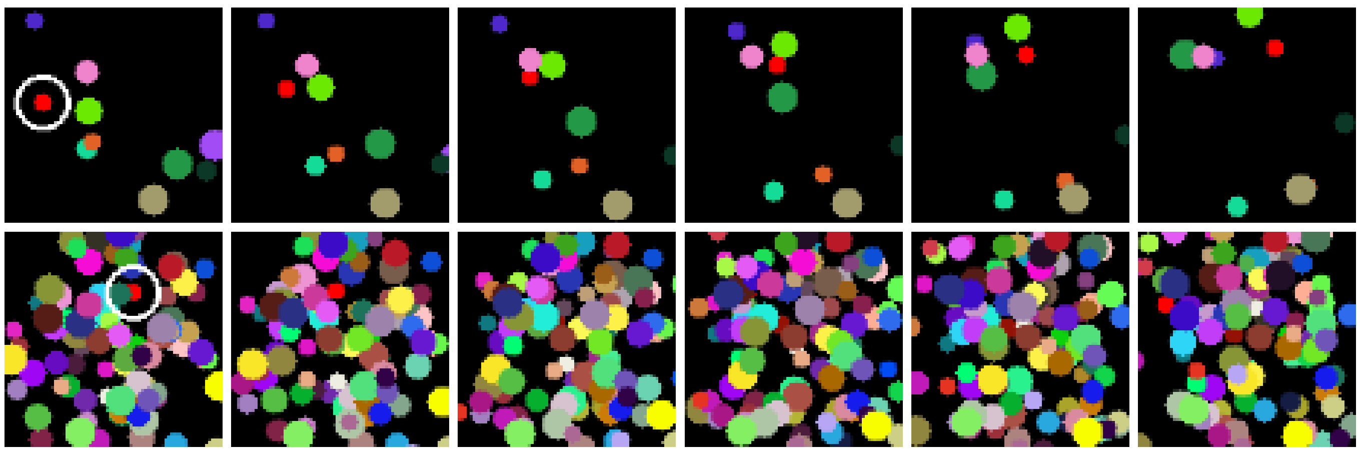
Visual State Estimation

- Tracking a red disc among distractors

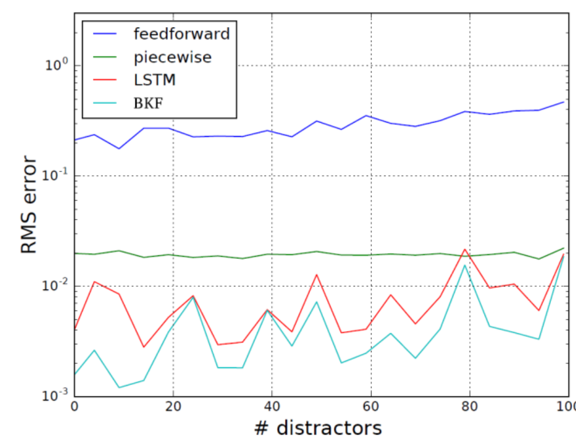
Time

Easy

Hard



State Estimation Model	# Parameters	RMS test error $\pm\sigma$
feedforward model	7394	0.2322 ± 0.1316
piecewise KF	7397	0.1160 ± 0.0330
LSTM model (64 units)	33506	0.1407 ± 0.1154
LSTM model (128 units)	92450	0.1423 ± 0.1352
BKF (ours)	**7493**	$\mathbf{0.0537} \pm \mathbf{0.1235}$



[Haarnoja et al. NIPS 2016]

KITTY Visual Odometry

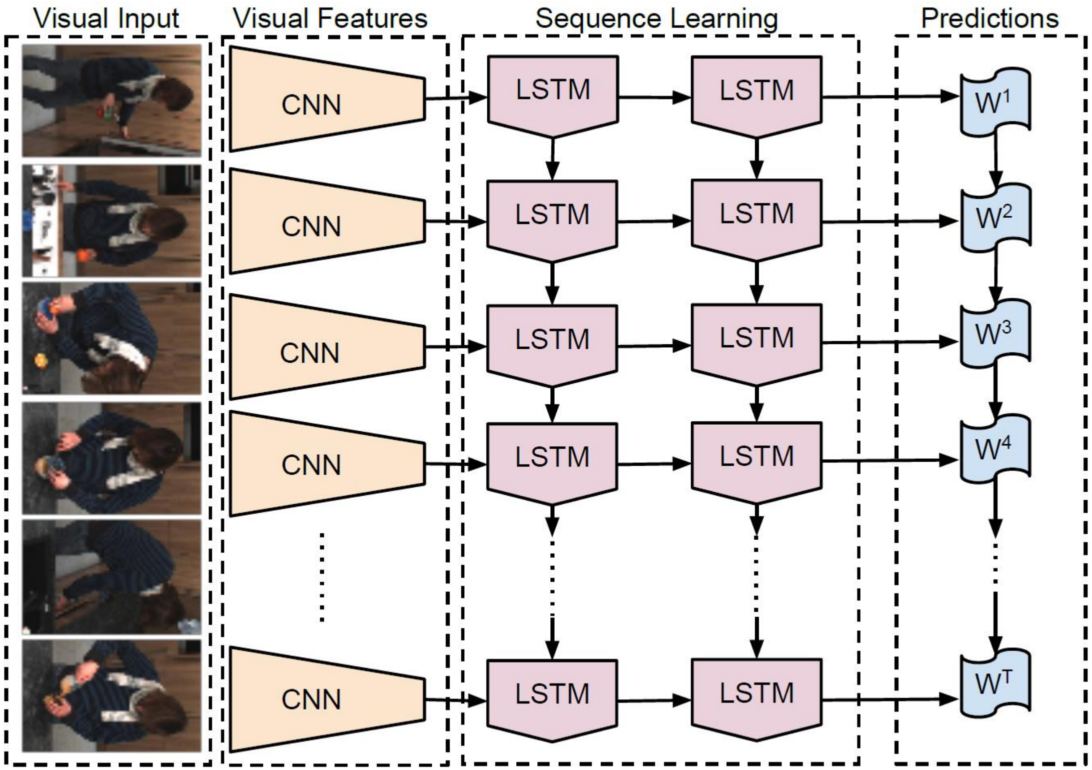
- Feedforward network of four convolutional and two fully connected layers (~500.000 weights) estimates velocities from image pairs

	Test 100			Test 100/200/400/800		
# training trajectories	3	6	10	3	6	10
Translational Error [m/m]						
piecewise KF	0.3257	0.2452	0.2265	0.3277	0.2313	0.2197
LSTM model (128 units)	0.5022	0.3456	0.2769	0.5491	0.4732	0.4352
LSTM model (256 units)	0.5199	0.3172	0.2630	0.5439	0.4506	0.4228
BKF (ours)	**0.3089**	**0.2346**	**0.2062**	**0.2982**	**0.2031**	**0.1804**
Rotational Error [deg/m]						
piecewise KF	0.1408	0.1028	0.0978	0.1069	0.0768	0.0754
LSTM model (128 units)	0.5484	0.3681	0.3767	0.4123	0. 3573	0.3530
LSTM model (256 units)	0.4960	0.3391	0.2933	0.3845	0.3566	0.3221
BKF (ours)	**0.1207**	**0.0901**	**0.0801**	**0.0888**	**0.0587**	**0.0556**

[Haarnoja et al. NIPS 2016]

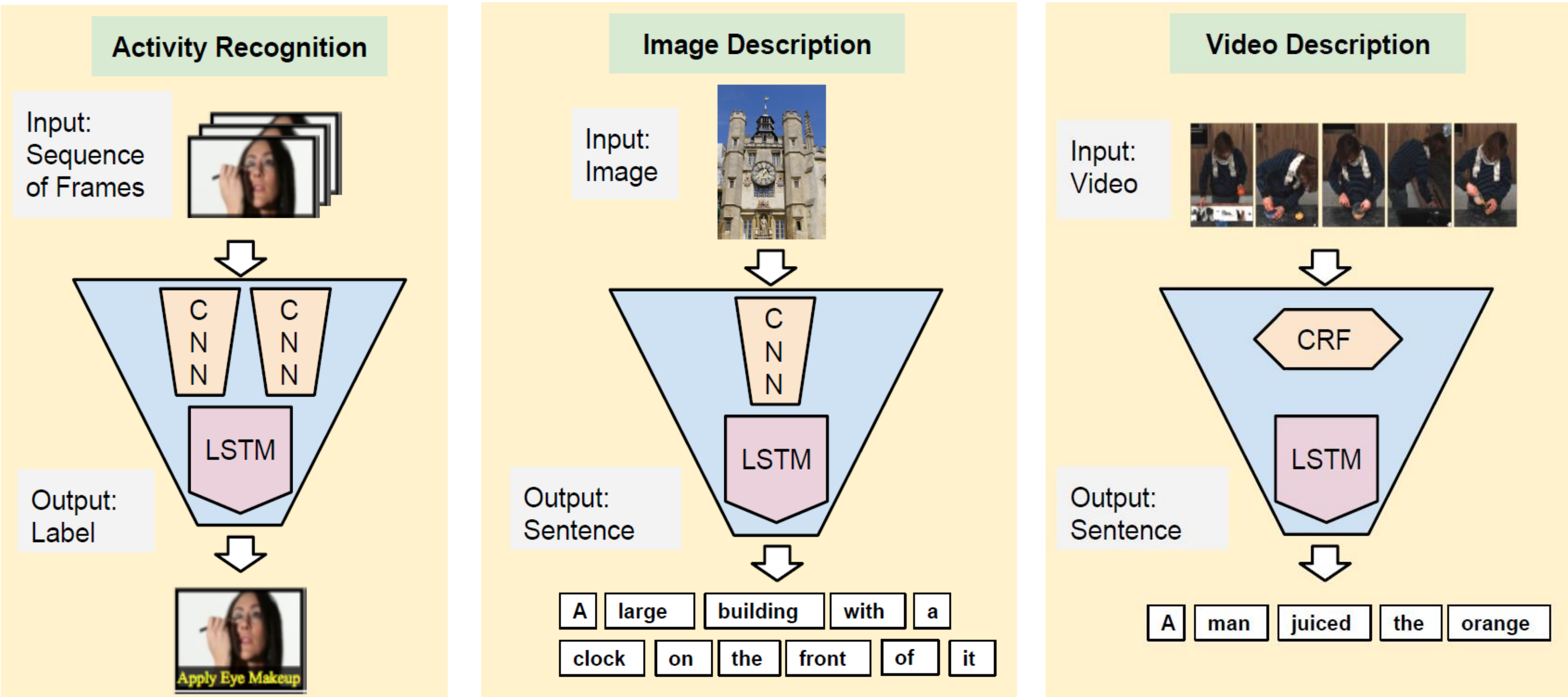
Image and Video Description

- Combining CNN feature extraction and LSTM sequence modelling



[Donahue et al. CVPR 2015]

Task-specific Model Variants



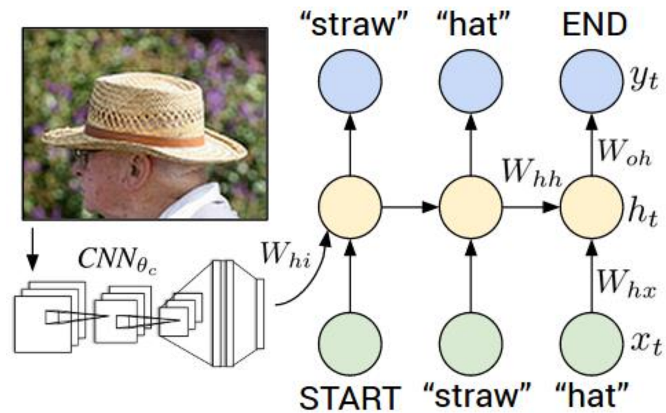
- Jointly learning feature extraction and sequential dynamics improves performance

[Donahue et al. CVPR 2015]

Generating Image Captions

- Multimodal recurrent neural network generative model

  [Karpathy, Fei-Fei 2015]



man in black shirt is playing guitar.

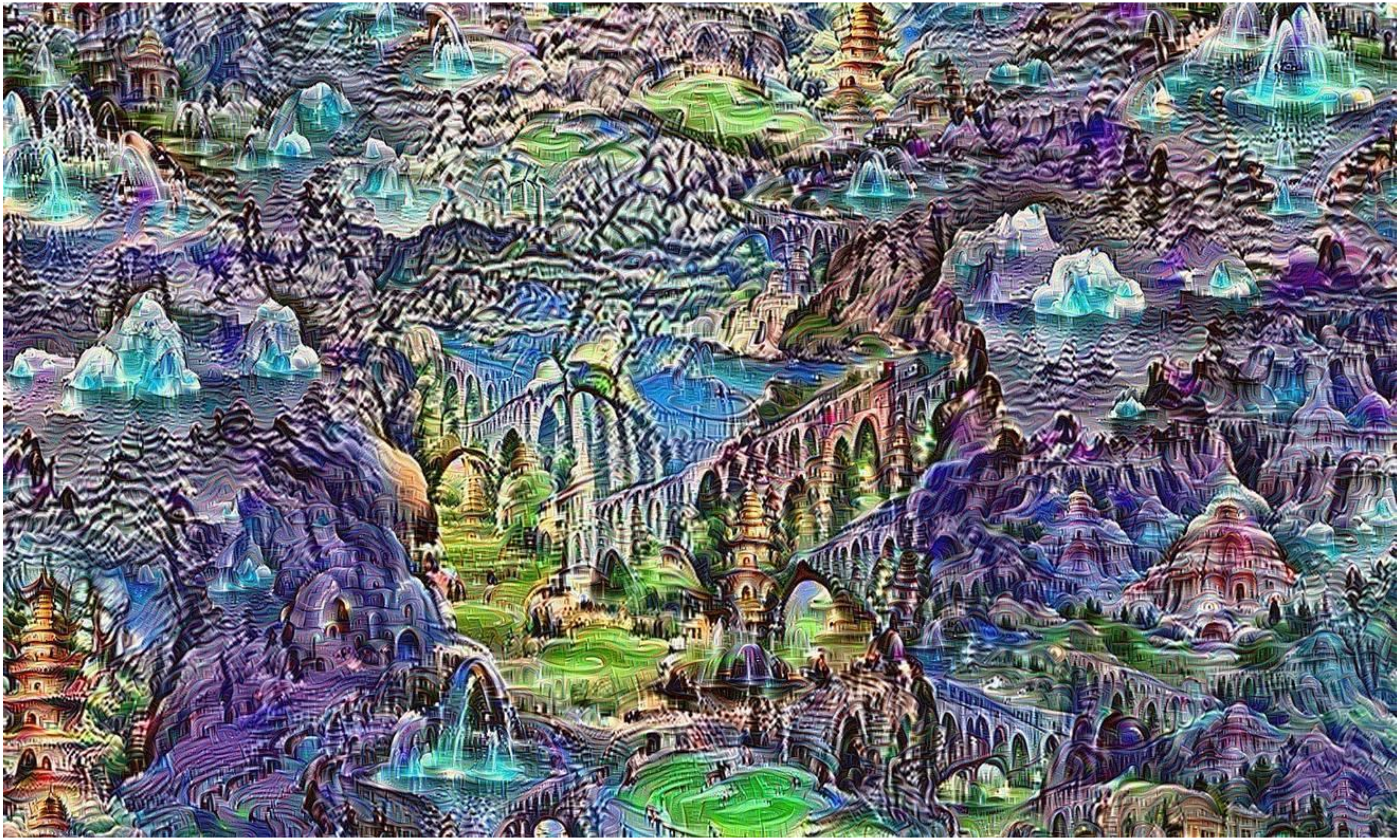
construction worker in orange safety vest is working on road.

two young girls are playing with lego toy.

Generating Image Captions

[Vinyals et al. 2015]

Dreaming Deep Networks



[Mordvintsev et al 2015]

Painting Style Transfer

Original

Turner

van Gogh

Munch

[Gatys et al. 2015]

Conclusion

- Flat models do not suffice
- Jump from signal to symbols too large
- Deep learning helps here:
  - **Hierarchical, locally connected** models
  - **Non-linear** feature extraction
- **Structure** of learning machine does matter
- Proposed architectures map well to **GPUs**
- **Iterative interpretation** uses partial results as context to resolve ambiguities
- Many questions open
  - Graphical models vs. neural networks
  - Structured vs. unstructured modelling
  - Stability of recurrent networks