Lab CudaVision
Learning Vision Systems on Graphics Cards (MA-INF 4308)

Introduction Session

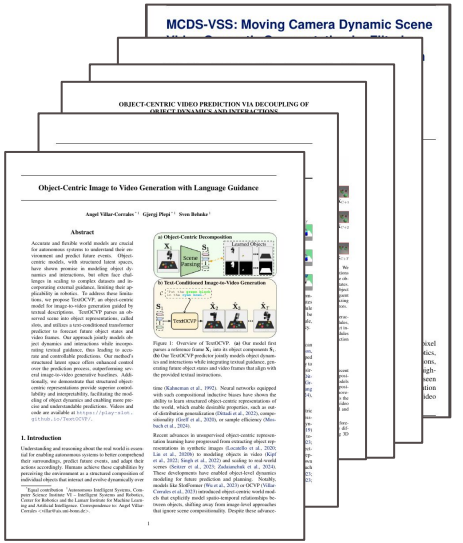
17.04.2026

PROF. SVEN BEHNKE, LUCA EICHLER

Contact: eichler@ais.uni-bonn.de

Original slides by Angel Villar-Corrales

- Member of research staff at AIS since 02.2021
- Before:
  - M. Sc. at FAU Erlangen-Nürnberg
  - B. Sc. at University of Vigo (Spain)
- Research interests:
  - Object-Centric Learning
  - Video Prediction and World Modelling
  - Computer Vision and Deep Learning



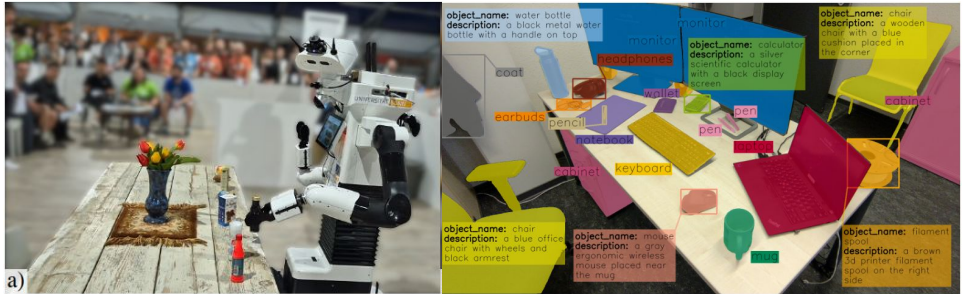
About Me

- Member of research staff at AIS since 03.2026
- Before:
  - M. Sc. at University of Bonn
  - B. Sc. at University of Bonn
- Work and Research interests:
  - World Modelling
  - Improving cognitive capabilities of robots
  - Vision Systems for RoboCup

Motivation

Extracting Visual Information

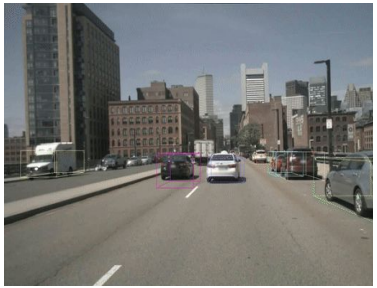
Service Robots

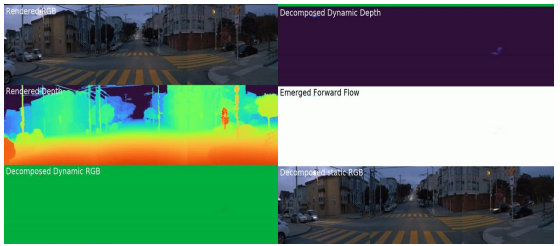


Manufacturing

Video Analytics

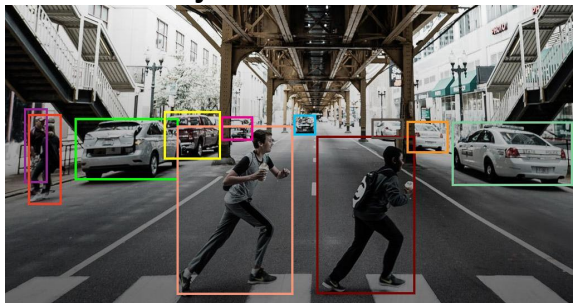
Scene Understanding





Applications in Vision

Object Detection



Semantic Segmentation

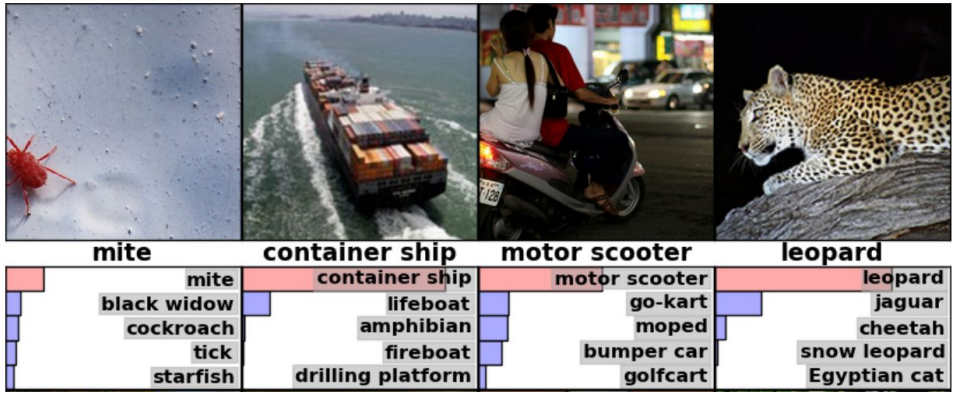
Human Pose Estimation

Image Synthesis & Manipulation

ImageNet Challenge

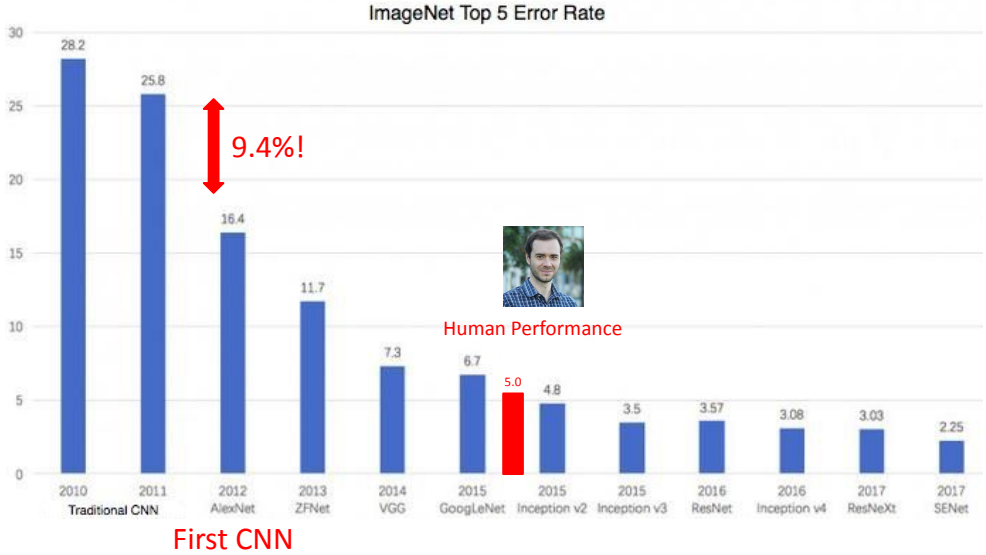
- ≈14 million natural images labelled into ≈1000 classes
- **2012:** Deep learning breakthrough by Krizhevsky et al.

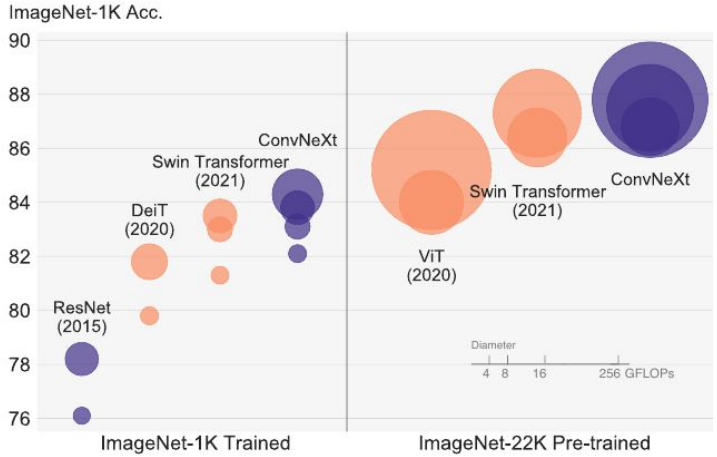
IMAGENET



[Krizhevsky et al. NeurIPS 2012]

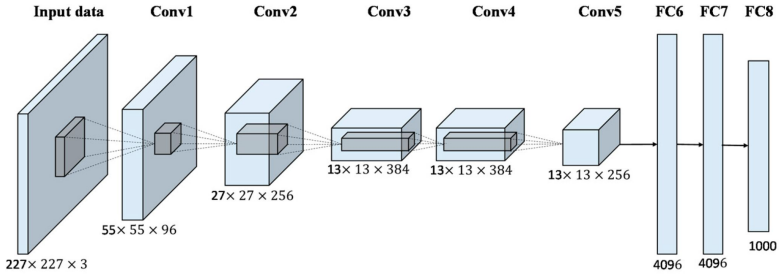
ImageNet SOTA



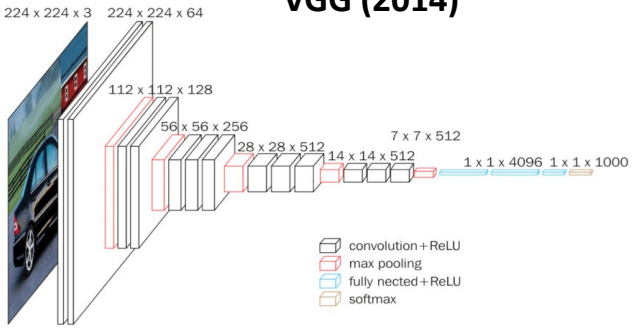


Architectures

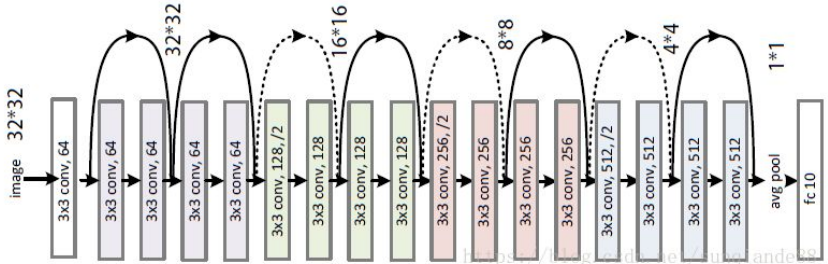
AlexNet (2012)



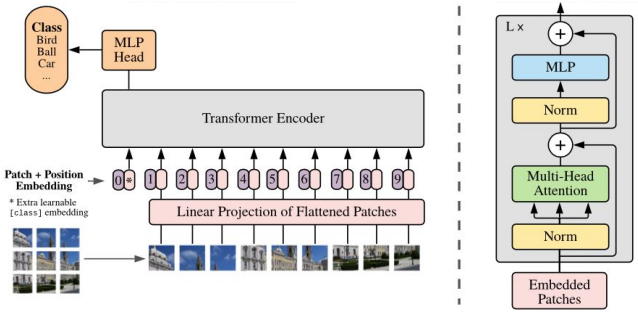
VGG (2014)



ResNet (2015)

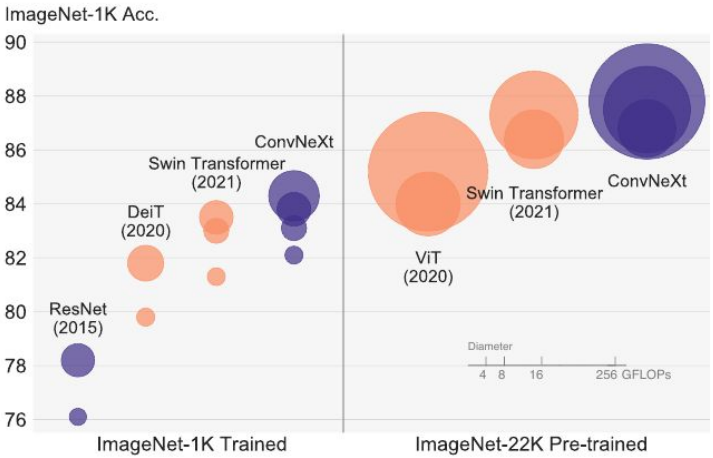


Vision Transformers (2020)

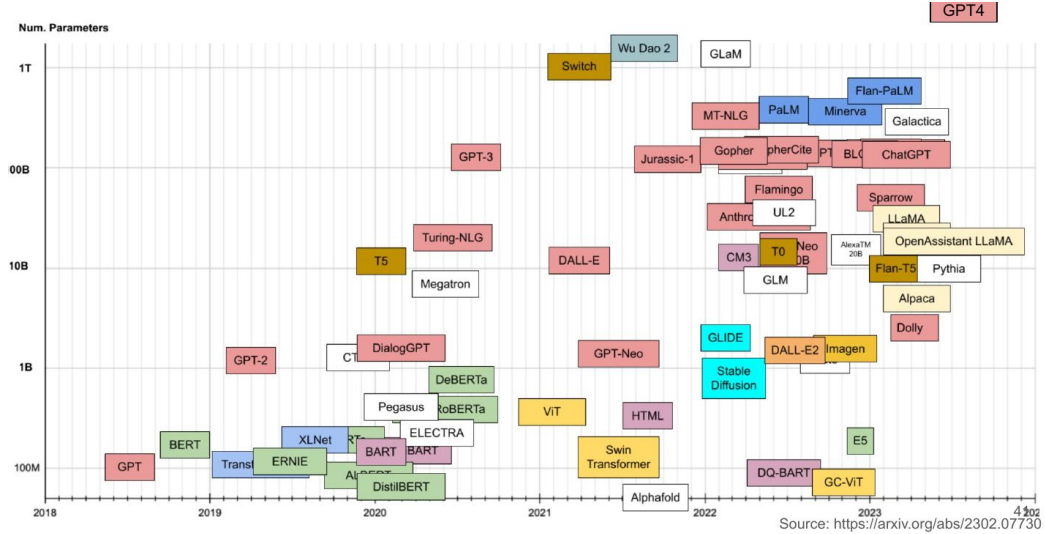


Model Landscape

(Former) State-of-the-art on ImageNet

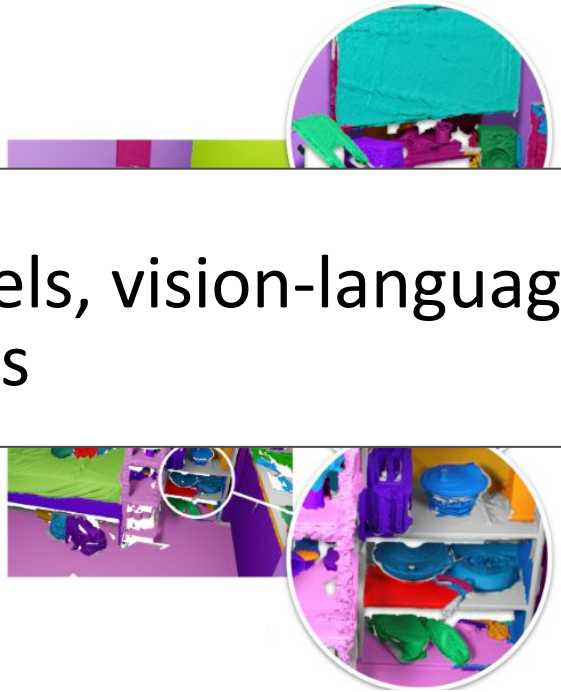


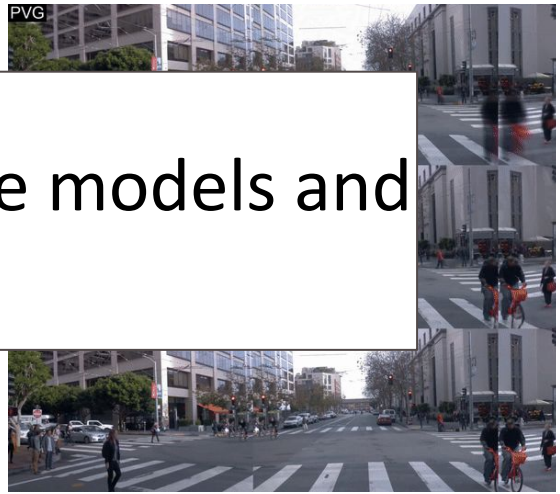
Large Language Models Landscape



(Beyond) The age of Deep Learning

The age of…

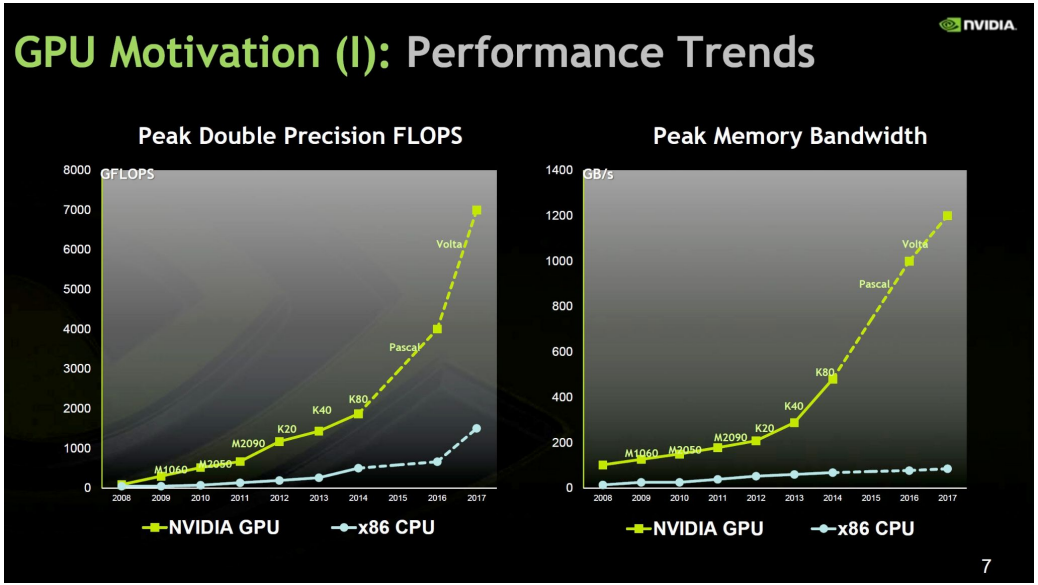




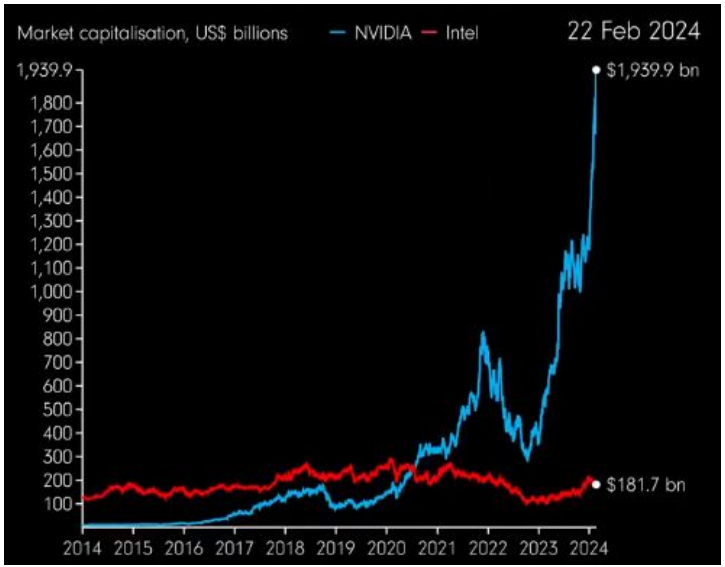
…foundation models, vision-language models and multimodal systems

(Beyond) The age of Deep Learning

CPU vs. GPU Performance

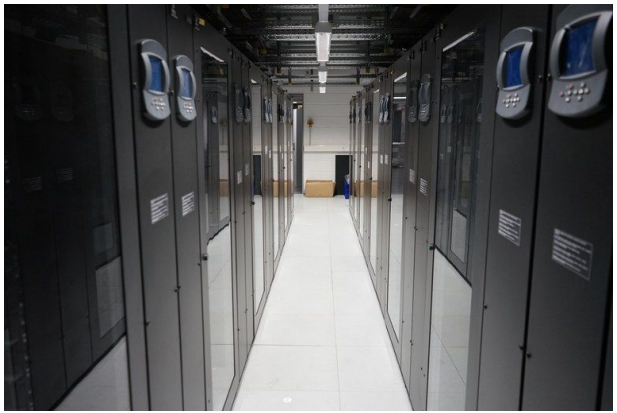


Nvidia Stock Price



Some of our GPUs

- HPC Marvin Cluster
  - 424 in Top500 list 2024



- Lamarr GPU Clusters
  - 2 clusters
  - 8/9 Compute Nodes
  - 8x A100 w/ 80GB

- Cuda Computers
  - 9 GPU computers
  - RTX 2080/3090 GPUs
  - 12 - 24 GB

In this Lab...

In this Lab...

- Implementing deep learning algorithms for visual pattern recognition
  - Python programming language
  - PyTorch framework
  - Deep learning: from basics to more advanced topics
- 8 Lab sessions (30%) in 13 weeks
- Final project (70%) in lecture free period (2 months)
  - Code/Results
  - Technical Report (8-10 pages)

Organization

- Meeting time: (Bi)Weekly 2 hours meeting, in-person
  - Discuss solution to previous assignment
  - Review some theory
  - Run sample code
  - Provide next assignment
  - Questions

- Accessible GPUs (Informatik ID):
  - Room 0.057 accessible during working hours
  - 9 GPU computers (cuda3 - cuda12) with RTX 2080/3090/Titan
  - Free online resources (Google Colab/ Kaggle kernels)

Assignments

- Each session covers one topic (e.g. Transfer Learning, GANs, …)
- Take-home assignment
  - Similar to what we do during the session
  - Due shortly before follow up session
- Assignments & project can be done in groups (max 3. people)
  - Highly recommended!
- Send me a mail with the name of your partner or tell me if you need one

Topics Covered

1. PyTorch basics, Optimization, Fully Connected and Convolutional Networks
2. Popular Architectures and Transfer Learning
3. Recurrent Neural Networks (RNNs & LSTM)
4. Autoencoders (AEs), Denoising and Variational AEs
5. Generative Adversarial Networks (GANs)
6. Diffusion Models
7. Transformers and Attention-Mechanisms
8. Introduction to Final Project

Registration

- 12 slots assigned at random (more on this later)
- Please fill this form before 22.04.2026 (select the time slots that work for you)

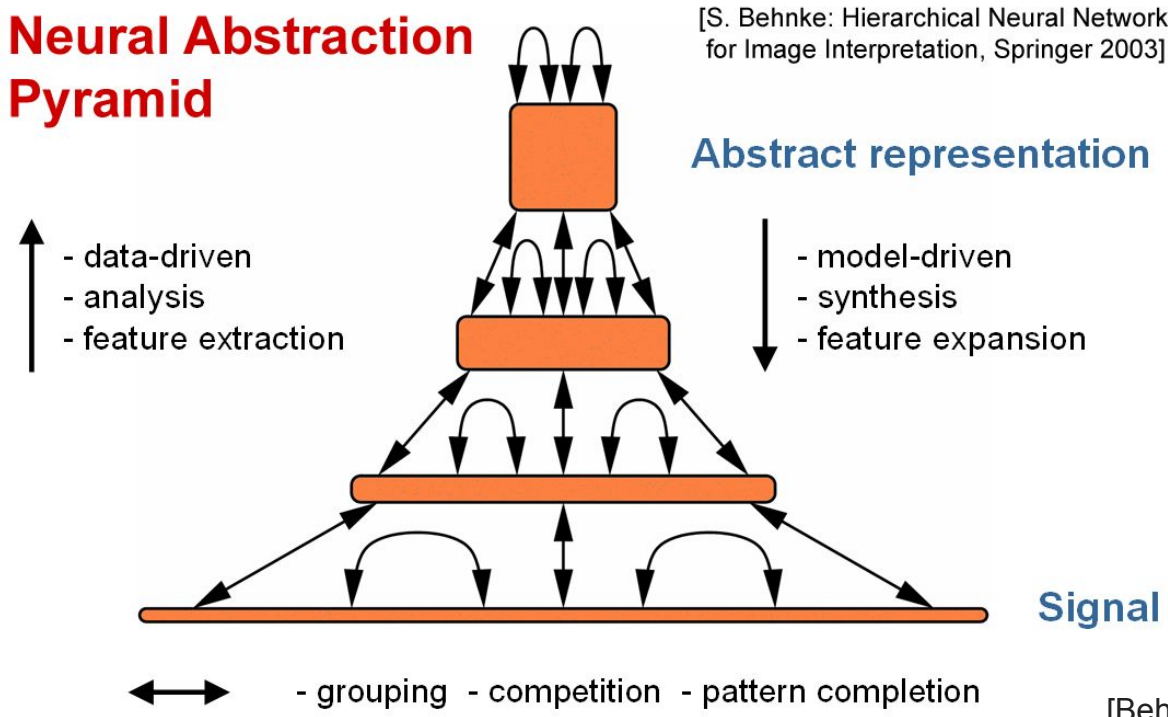
  https://forms.gle/GSg5YNAZsVJ4GkRs7

- Registration (contact me first):
  - **Uni Bonn**: in BASIS
  - **Bonn-Rhein-Sieg & Others**: Contact me

	Monday	Tuesday	Wednesday	Thursday	Friday
10-12					
12-14					
14-16					
16-18					

Some of our Research...

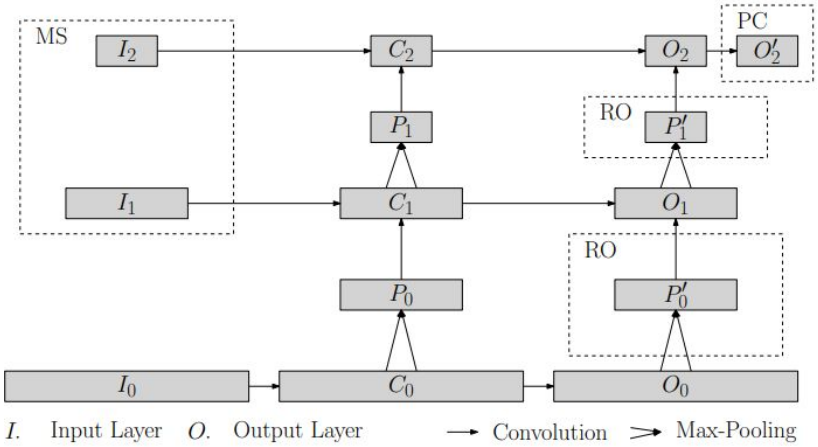
Neural Abstraction Pyramid



[Behnke. IJCNN 1998]

Object-class Segmentation

- Multi-scale CNN for RGB-pixel segmentation

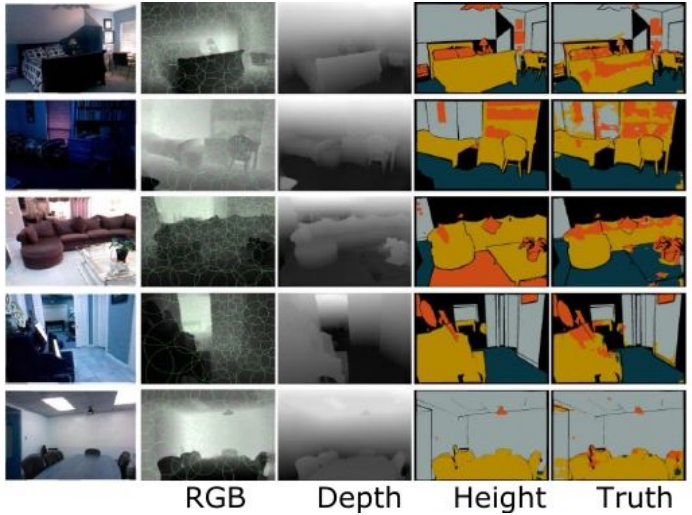


Input Output Ground Truth

[Schulz and Behnke. ESANN 2012]

RGB-D Object Segmentation

- Use of kinect-like sensors to obtain depth values



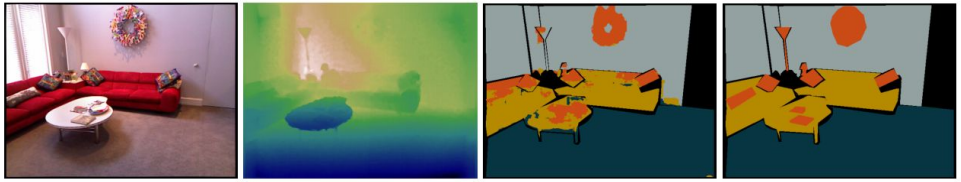
Method	floor	struct	furnit	prop	Class Avg.	Pixel Acc.
CW	84.6	70.3	58.7	52.9	66.6	65.4
CW+DN	87.7	70.8	57.0	53.6	67.3	65.5
CW+H	78.4	74.5	55.6	62.7	67.8	66.5
CW+DN+H	93.7	72.5	61.7	55.5	70.9	70.5
CW+DN+H+SP	91.8	74.1	59.4	63.4	72.2	71.9
CW+DN+H+CRF	93.5	80.2	66.4	54.9	**73.7**	**73.4**
Müller et al.[8]	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 averaged within superpixels and SVM-reweighted. CRF is a conditional random field ov superpixels [8]. Structure class numbers are optimized for class accuracy.

[Schulz, Höft and Behnke. ESANN 2015]

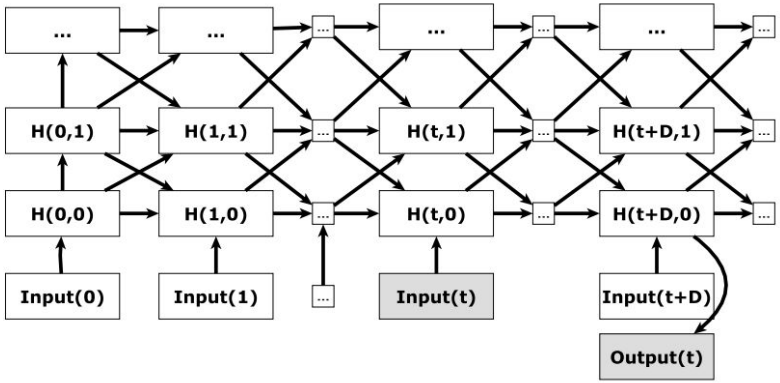
Object Segmentation from RGB-D Video

- Video processing with multi-scale Convolutional RNNs
- Iterative refinement through different time steps



(a) RGB frame (b) Depth (c) Prediction (d) Ground truth

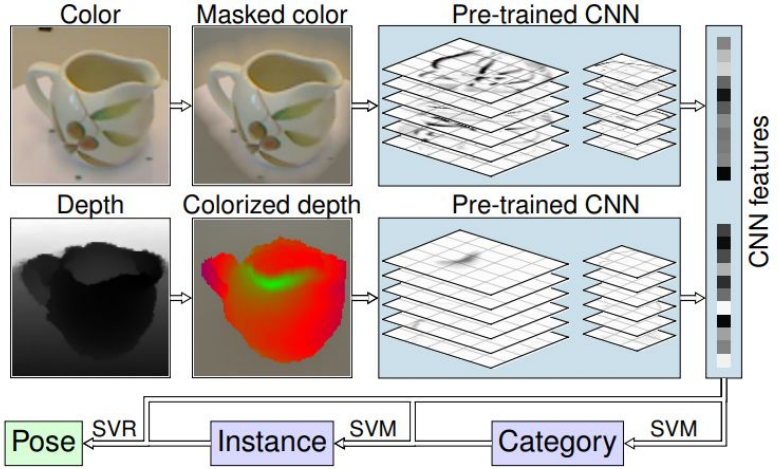
Method	Class Accuracies (%)				Average (%)	
	ground	struct	furnit	prop	Class	Pixel
Unidirectional + SW	90.0	76.3	52.1	**61.2**	69.9	67.5
Schulz *et al.* [20]	93.6	80.2	66.4	54.9	**73.7**	**73.4**
Müller and Behnke [22]	**94.9**	78.9	**79.7**	55.1	71.9	72.3
Stückler *et al.* [21]	90.8	81.6	67.9	19.9	65.0	68.3
Couprie *et al.* [23]	87.3	**86.1**	45.3	35.5	63.5	64.5
Höft *et al.* [19]	77.9	65.4	55.9	49.9	61.1	62.0
Silberman *et al.* [17]	68	59	70	42	59.6	58.6



[Pavel, Schulz, and Behnke. IJCNN 2015, Neural Networks 2017]

Computer Vision with Pretrained Features

- Object recognition and pose estimation
- Pretrained features from ImageNet
- Improved classification and estimation performance



Evaluation on the Washington RGB-D Objects dataset

Method	Category Accuracy (%)		Instance Accuracy (%)	
	RGB	RGB-D	RGB	RGB-D
Lai *et al.* [12]	74.3 ± 3.3	81.9 ± 2.8	59.3	73.9
Bo *et al.* [14]	82.4 ± 3.1	87.5 ± 2.9	**92.1**	92.8
PHOW[18]	80.2 ± 1.8	—	62.8	—
Ours	$\mathbf{83.1 \pm 2.0}$	$\mathbf{89.4 \pm 1.3}$	92.0	**94.1**

[Schwarz, Schulz and Behnke. ICRA 2015]

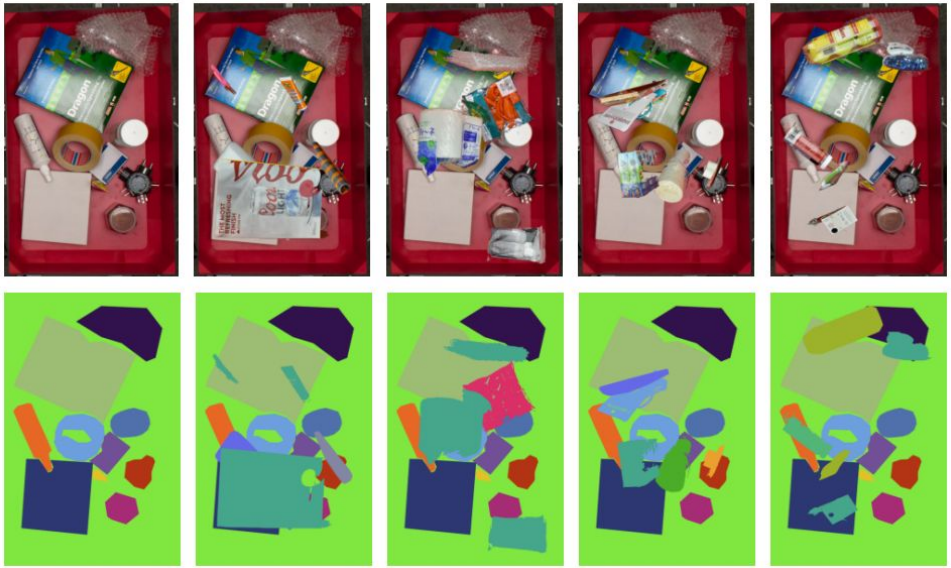
Amazon Bin-Picking Challenge

- Picking a large variety of objects
- Placing them on a shelf or packing boxes
- NimbRo team came in 2nd
- Computer vision challenge

[Schwarz et al. ICRA 2017]

Object Capture and Scene Synthesis

- Capture data with a turn table
- Rendered realistic scenes



Object Detection and Segmentation

- RefineNet architecture [1]
- Trained on rendered data

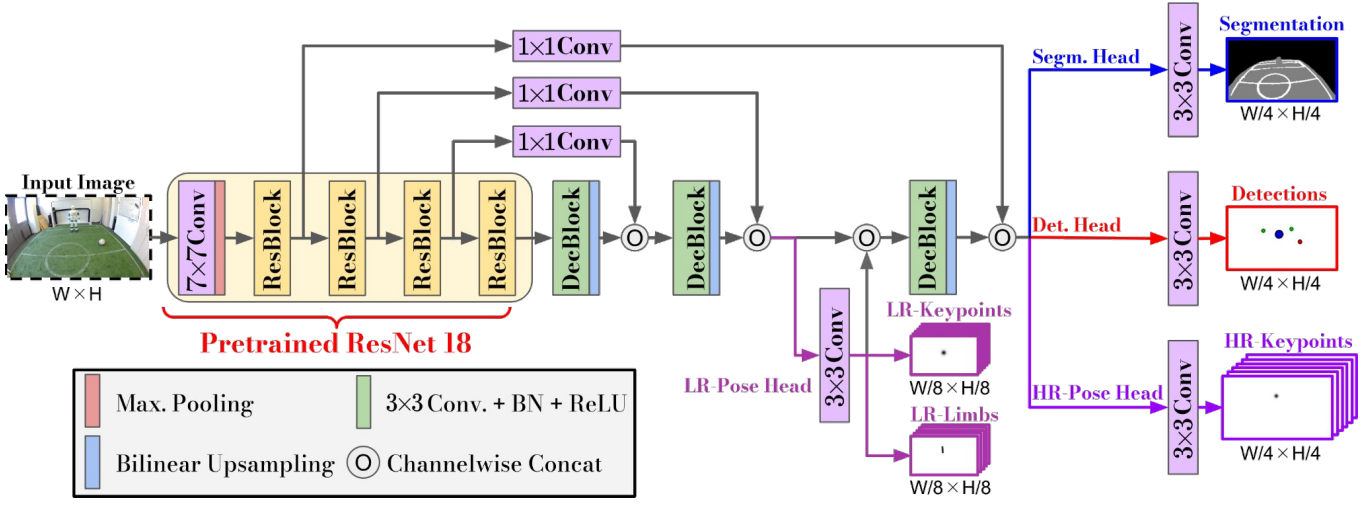
[1] Lin et al. CVPR 2016

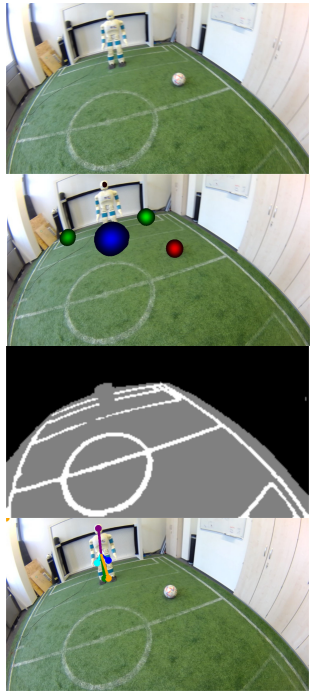
Soccer Robots

- NimbRo participates in humanoid soccer robot competitions (RoboCup)
- Challenging perception scenario

Soccer Scene Understanding

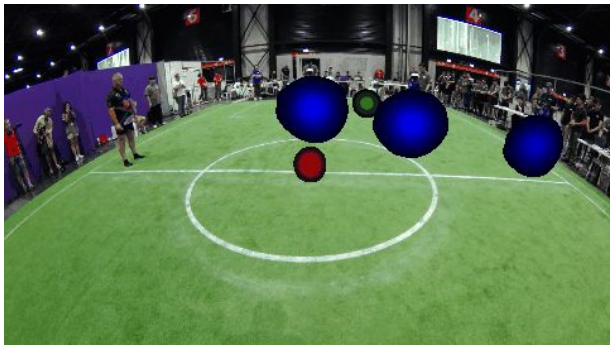
- End-to-end convolutional model
- Robot and ball detection
- Soccer field segmentation

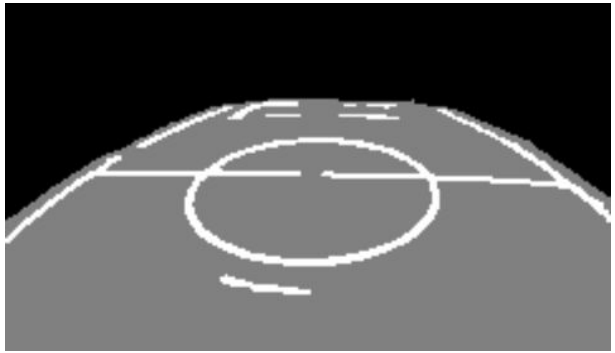




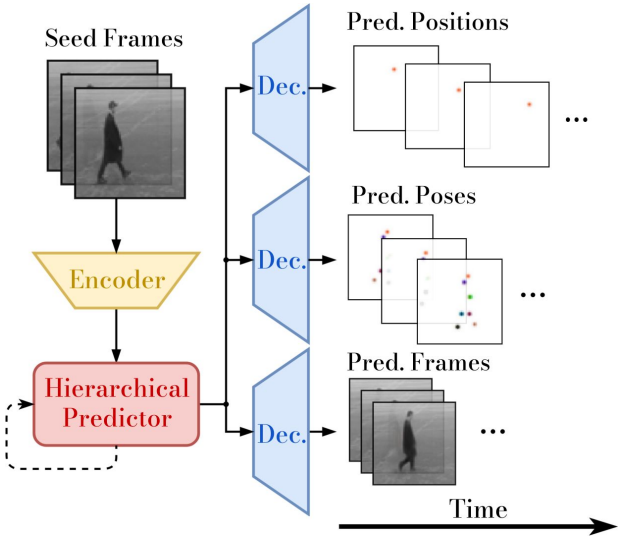
[Villar-Corrales et al. RoboCup 2023]

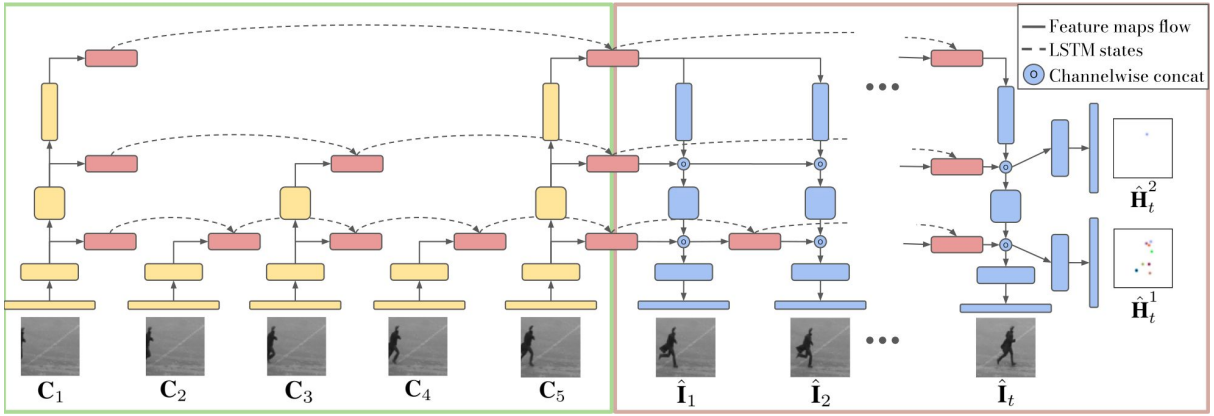
Scene Understanding





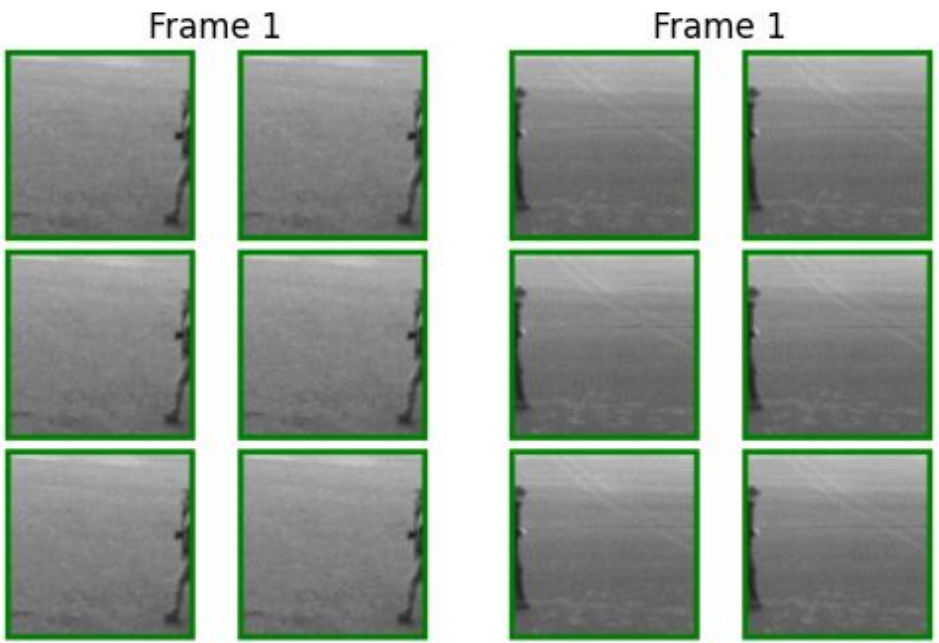
Future Frame Video Prediction





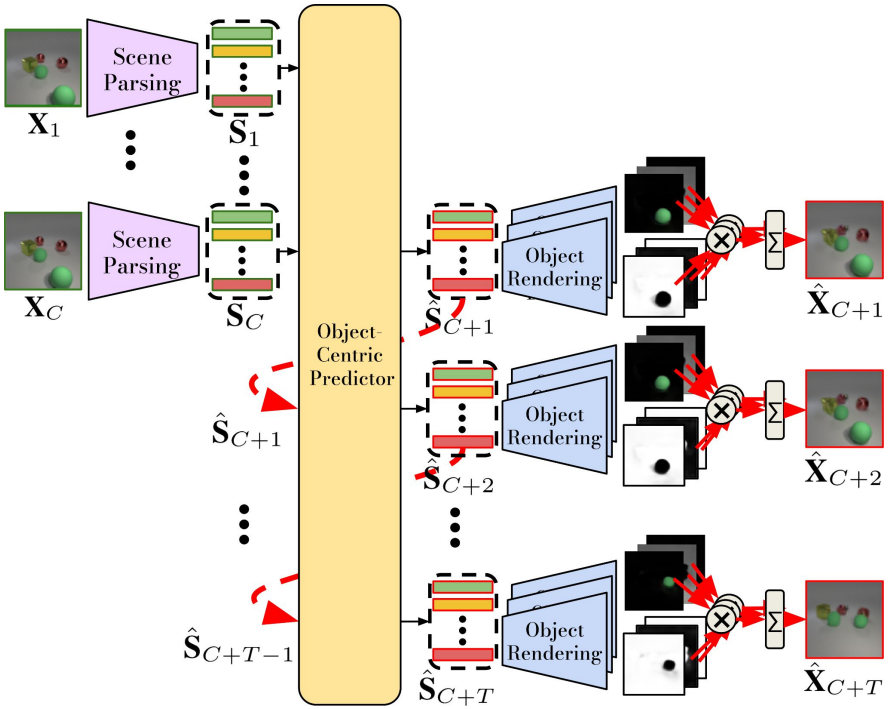
[Villar-Corrales et al. BMVC 2022]

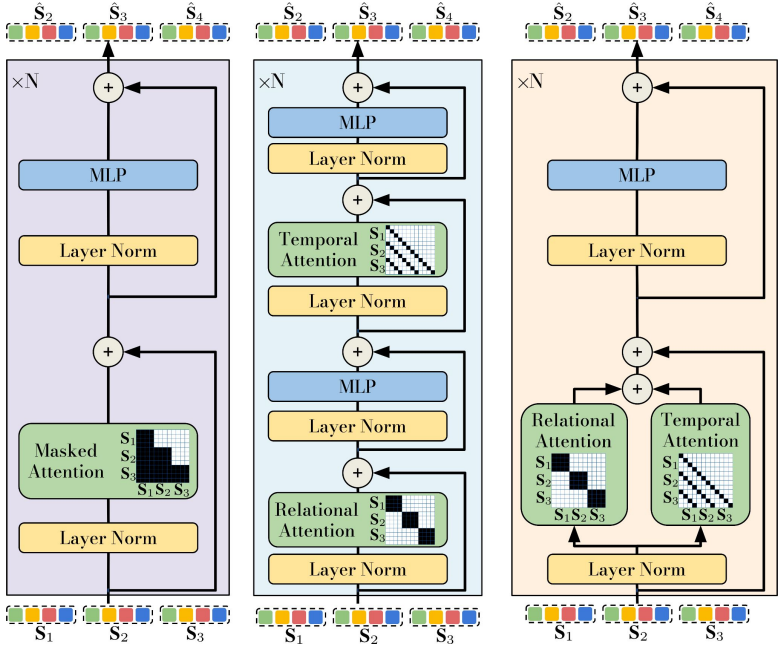
Future Frame Video Prediction



[Villar-Corrales et al. BMVC 2022]

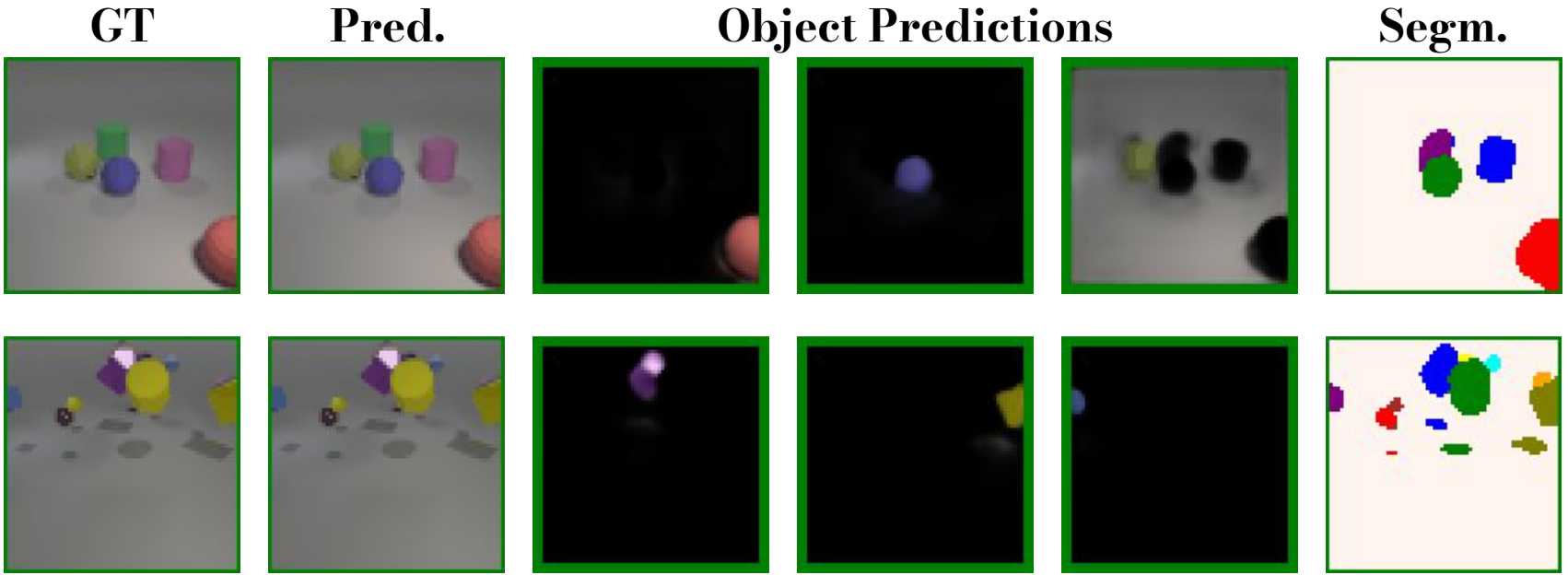
Object-Centric Video Prediction





[Villar-Corrales et al. ICIP 2023]

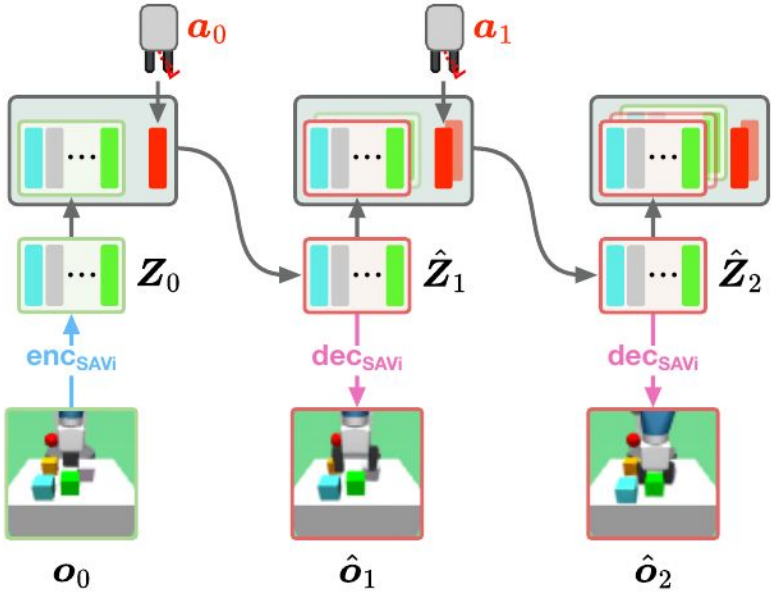
Object-Centric Video Prediction



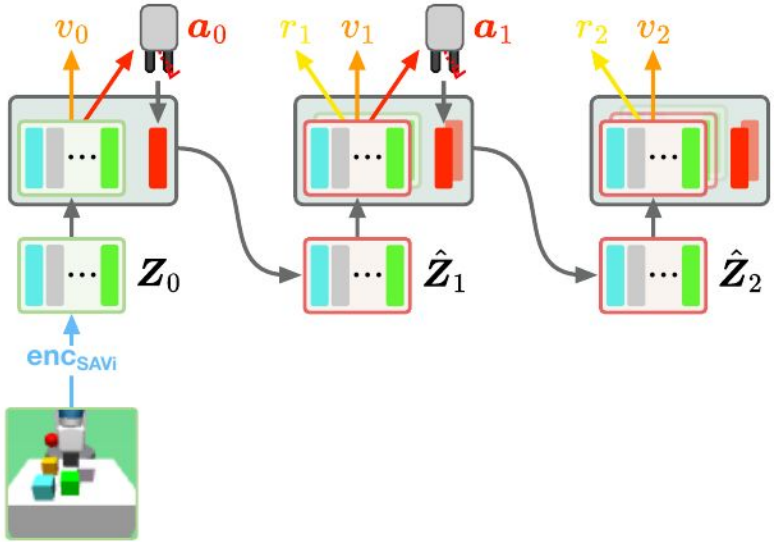
[Villar-Corrales et al. ICIP 2023]

Learning Object-Centric Latent Dynamics

- Learning an object-centric world model
  - Action-conditional prediction

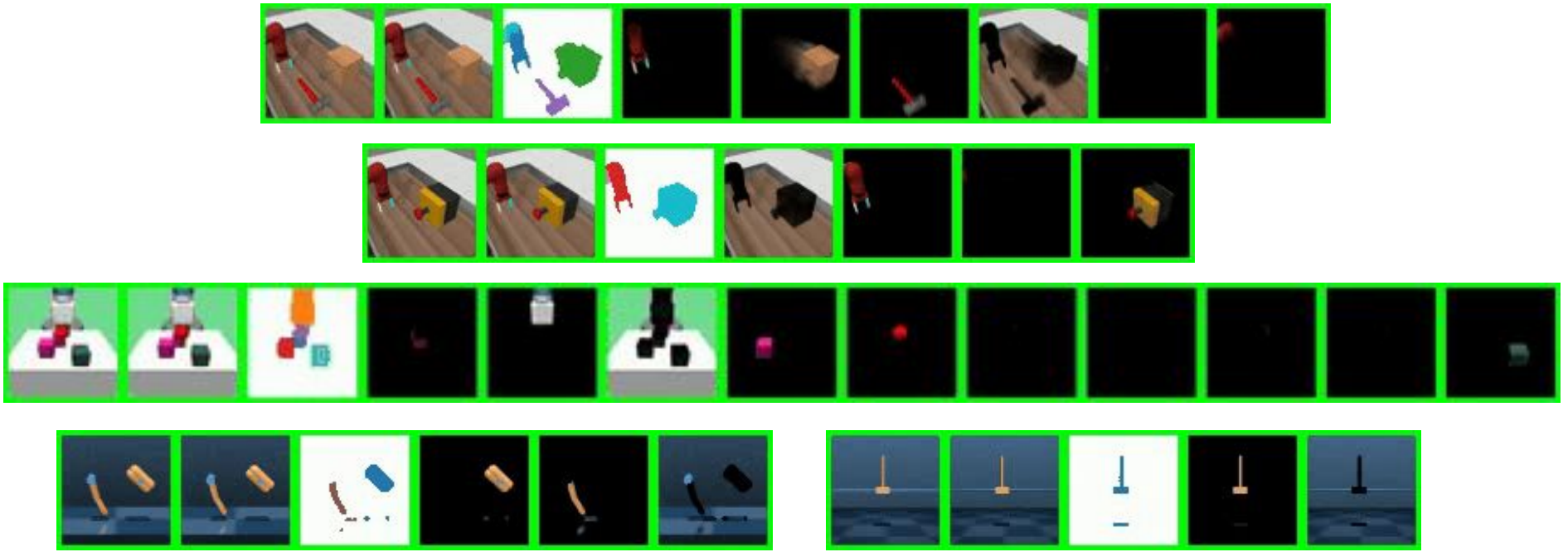


- Learning behaviors via imaged rollouts
  - Actor and critic networks



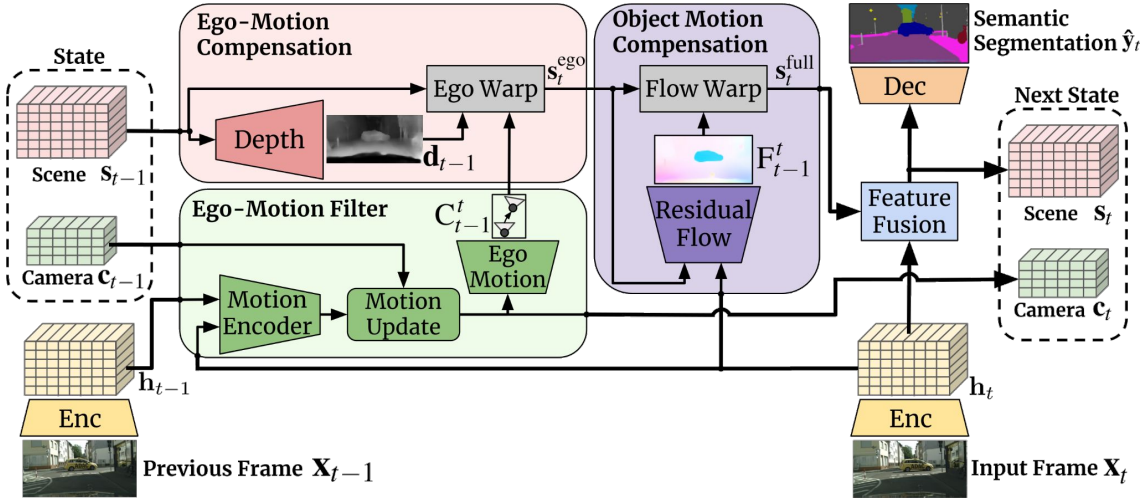
[Mosbach, Ewertz, Villar-Corrales and Behnke. Under Review 2025]

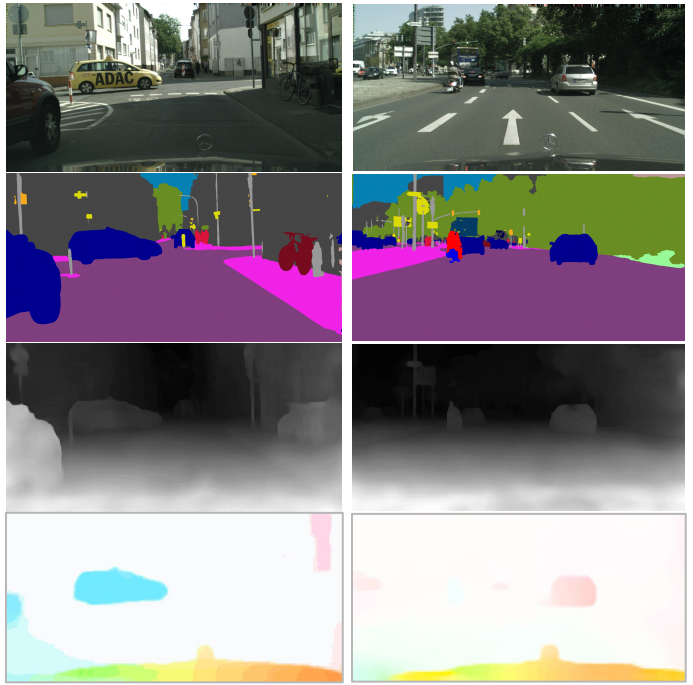
Learning Object-Centric Latent Dynamics



[Mosbach, Ewertz, Villar-Corrales and Behnke. Under Review 2024]

Structured Video Segmentation





[Villar-Corrales et al. BMVC 2024]

And much more...

UAV Perception

3D Deep Learning

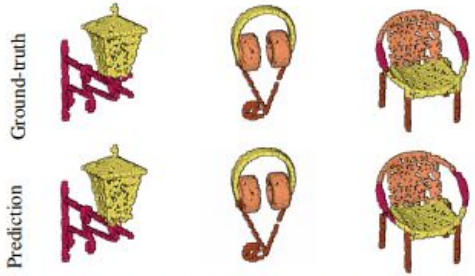
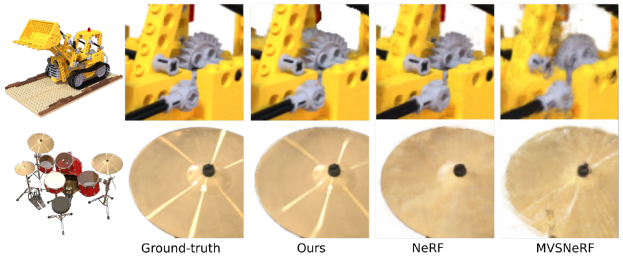
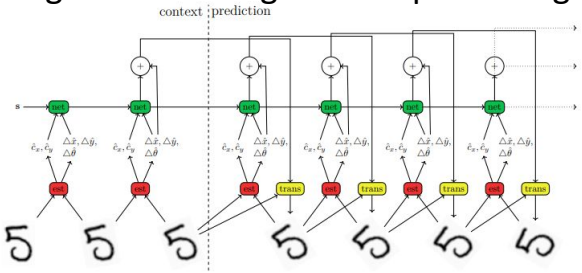


Fig. 8: ShapeNet [39] results of our method.

Neural Representations & Rendering



Signal Processing and Deep Learning



Scene Synthesis

Once Again...

Slot Assignment Selection

- 12 slots for students
  - Assigned at random

Registration

- 12 slots assigned at random (more on this later)
- Please fill this form **before** 22.04.2025 (select the time slots that work for you)

  https://forms.gle/GSg5YNAZsVJ4GkRs7

- Registration (contact me first):
  - **Uni Bonn**: in BASIS
  - **Bonn-Rhein-Sieg & Others**: Contact me

	Monday	Tuesday	Wednesday	Thursday	Friday
10-12					
12-14					
14-16					
16-18					

Questions?