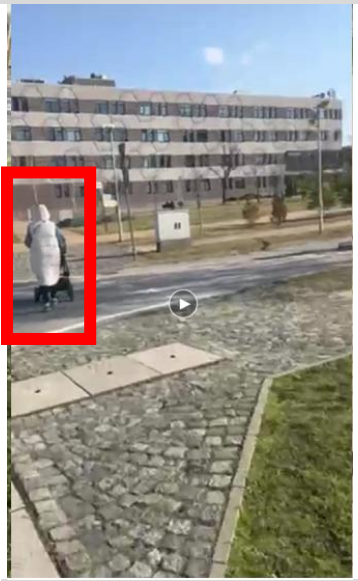
Karlsruher Institut für Technologie

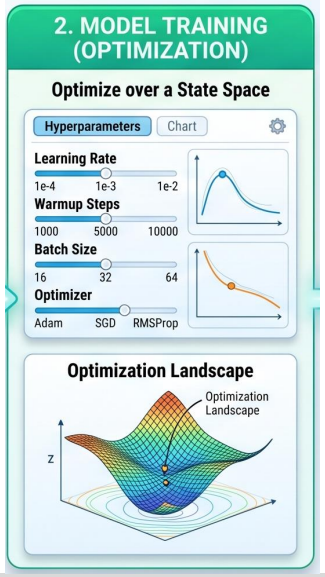
cv:hci @ KIT

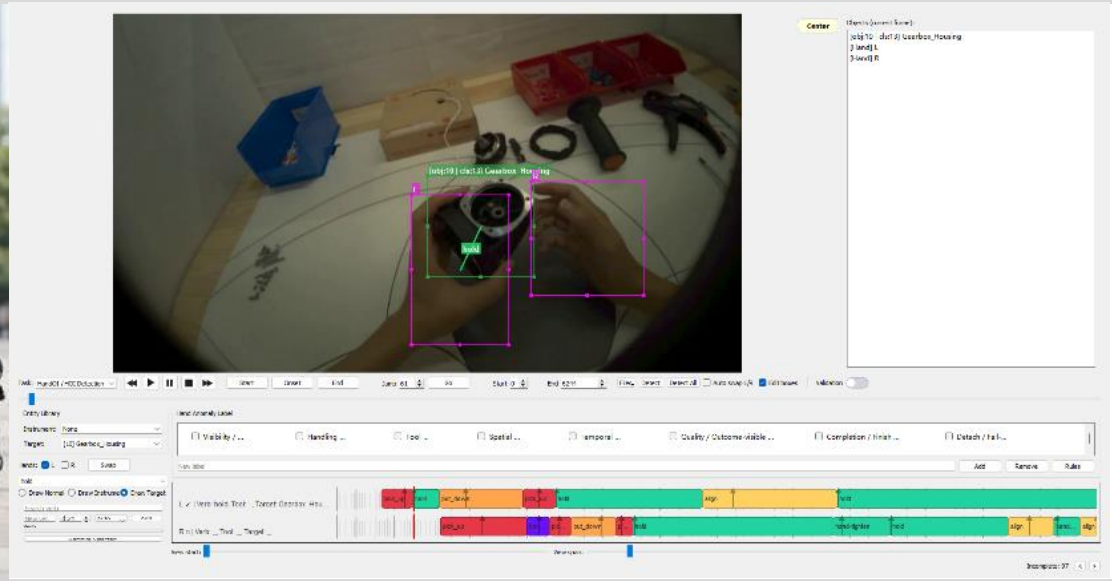
Practical Course: Computer Vision for Human-Computer Interaction

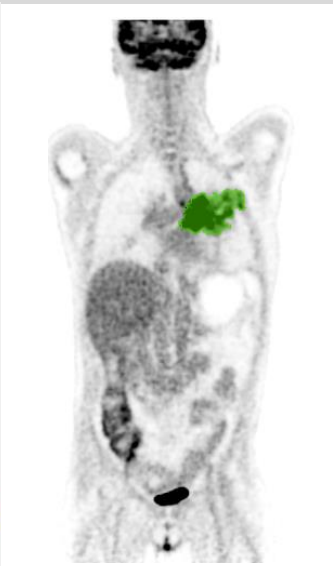
SS 2026

Dr.-Ing. Zdravko Marinov
Institute for Anthropomatics and Robotics (IAR), Karlsruhe Institute of Technology (KIT)





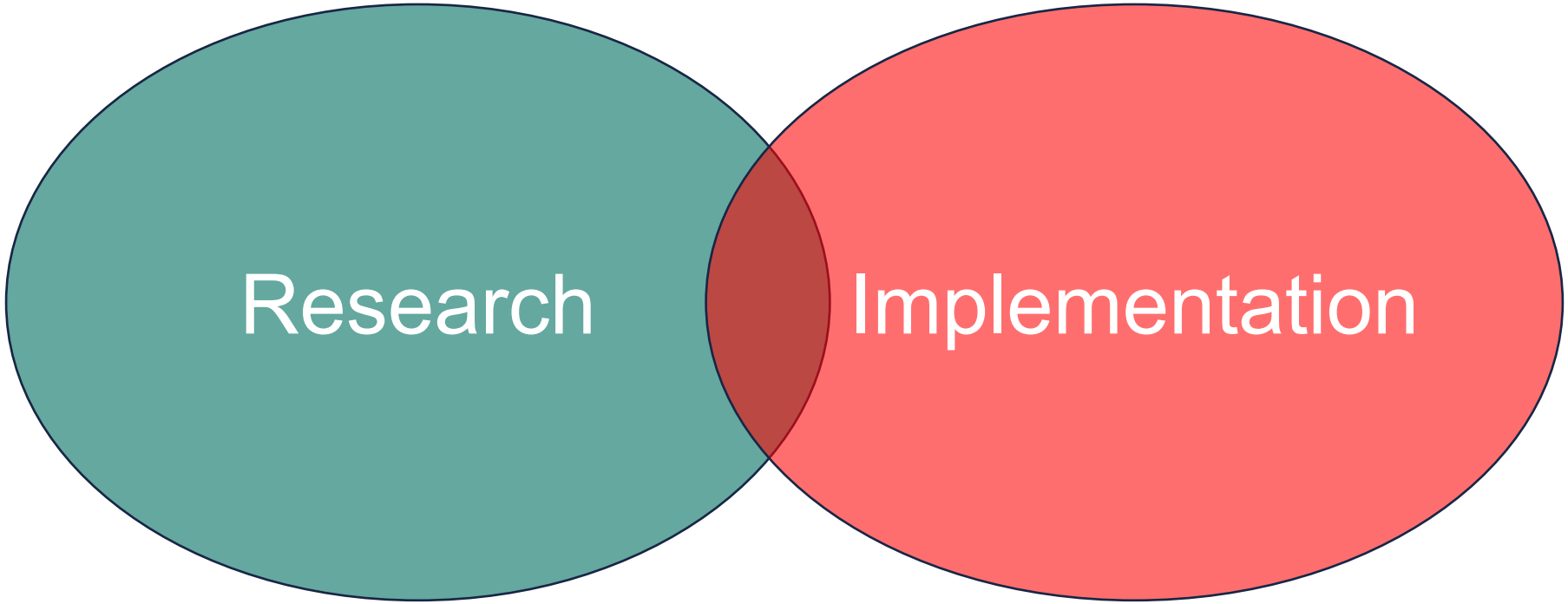




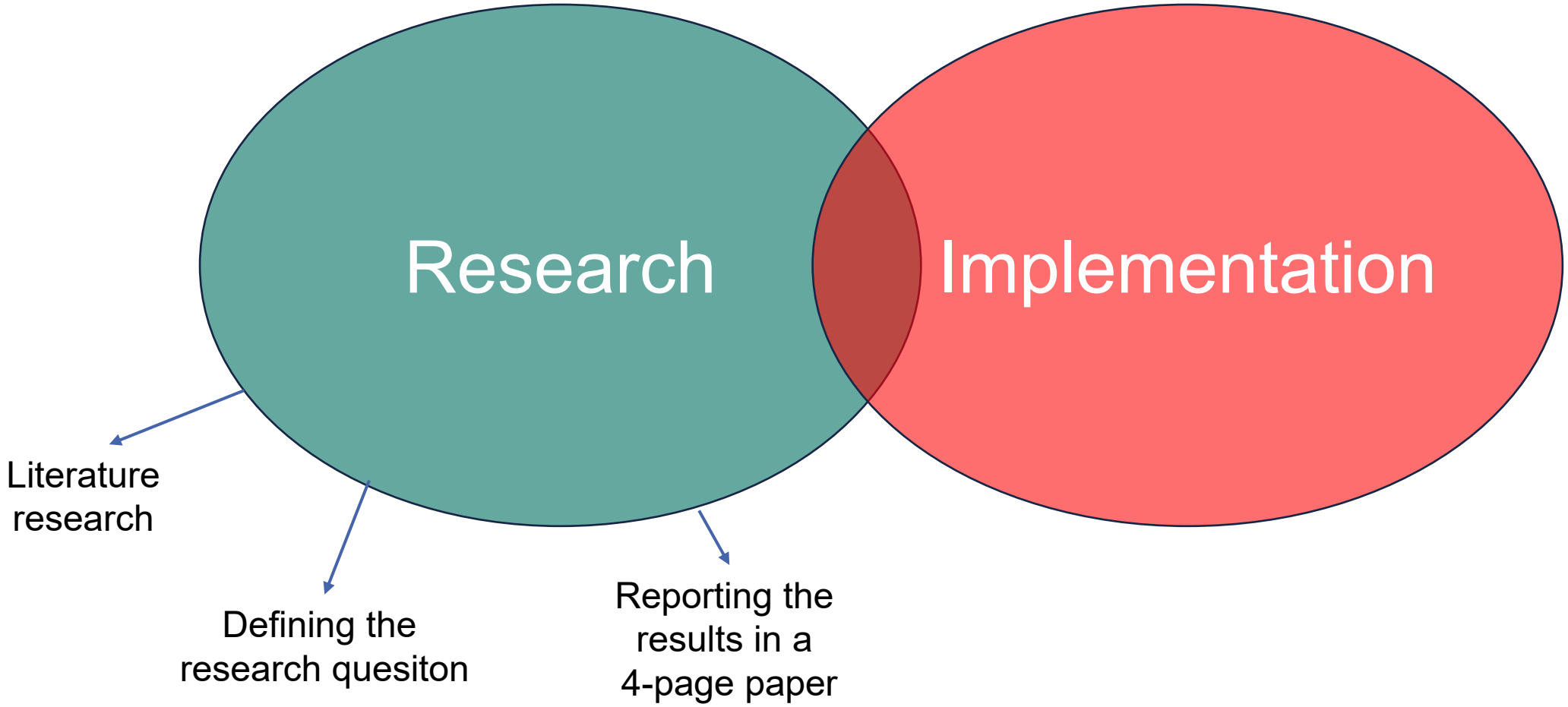
What will you learn?

- Apply algorithms from lectures and papers
- Hands-on experience
- Get comfortable with machine learning tools
- Learn about current problems and applications in machine learning and vision
- Find solutions to difficult problems

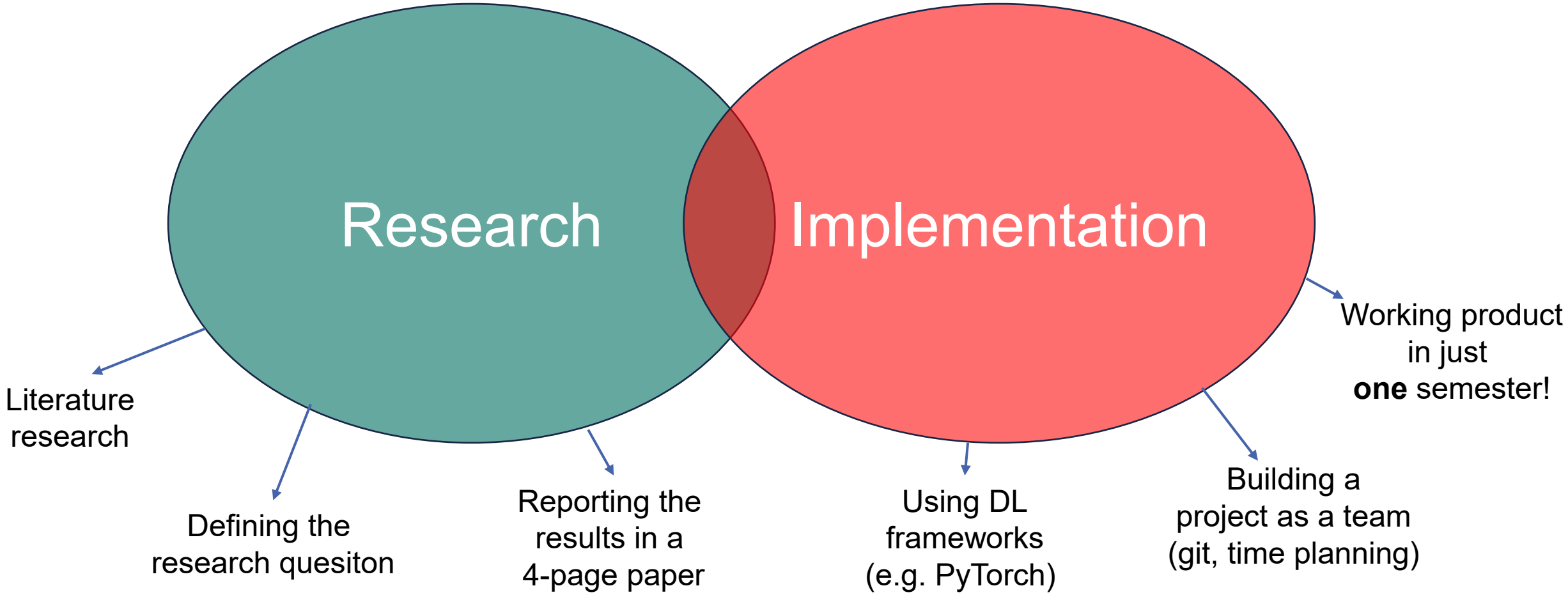
Scope of the Practical Course



Scope of the Practical Course

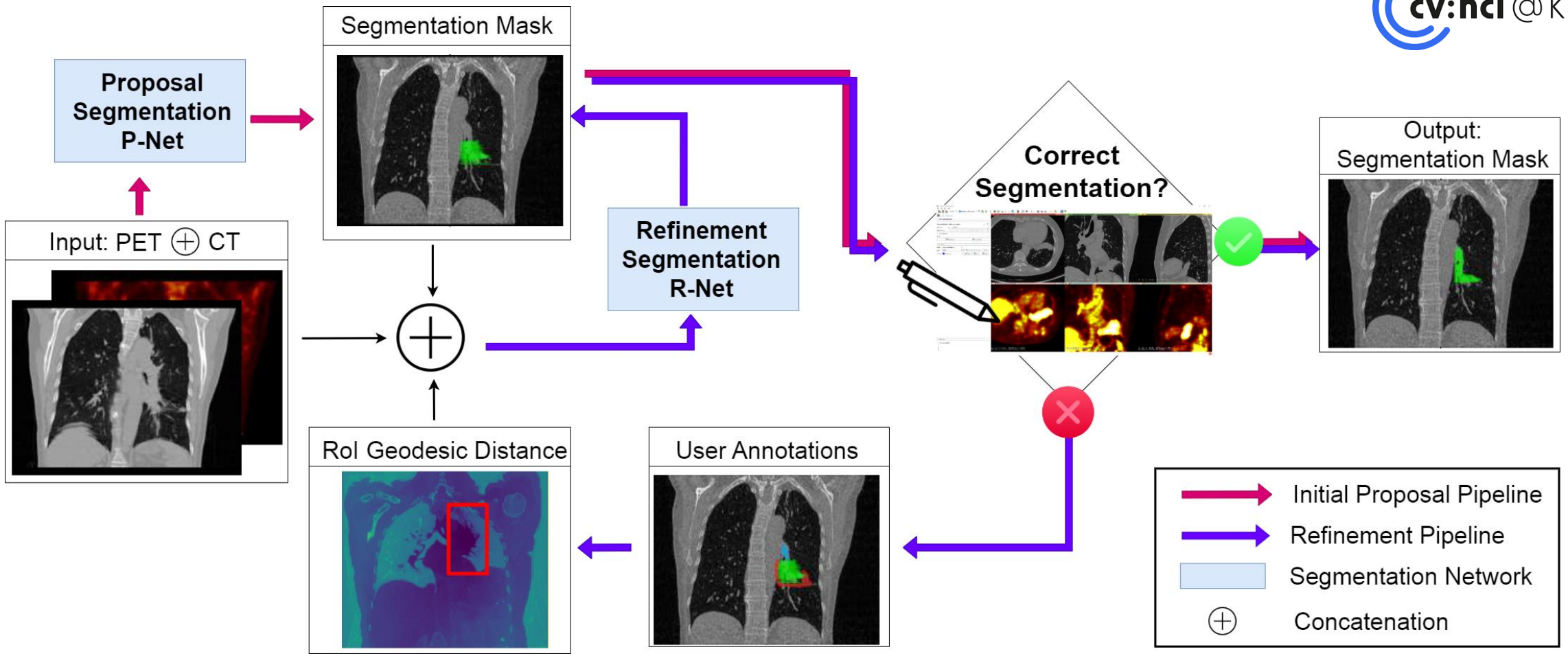


Scope of the Practical Course

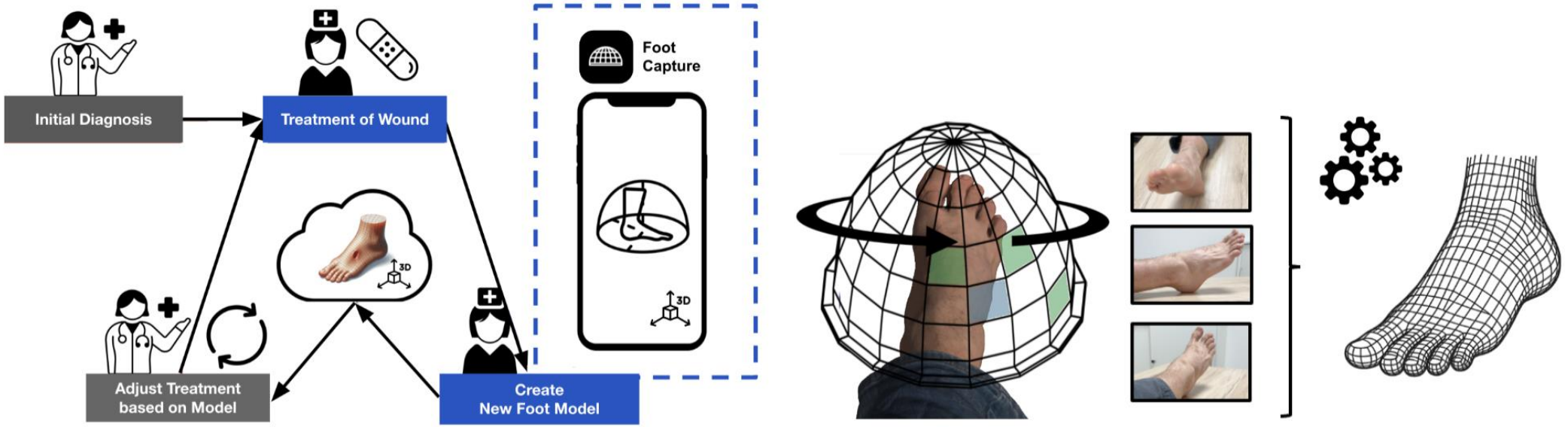


Examples from previous semesters: SS21 – Flying Guide Dog, ROBIO 2021

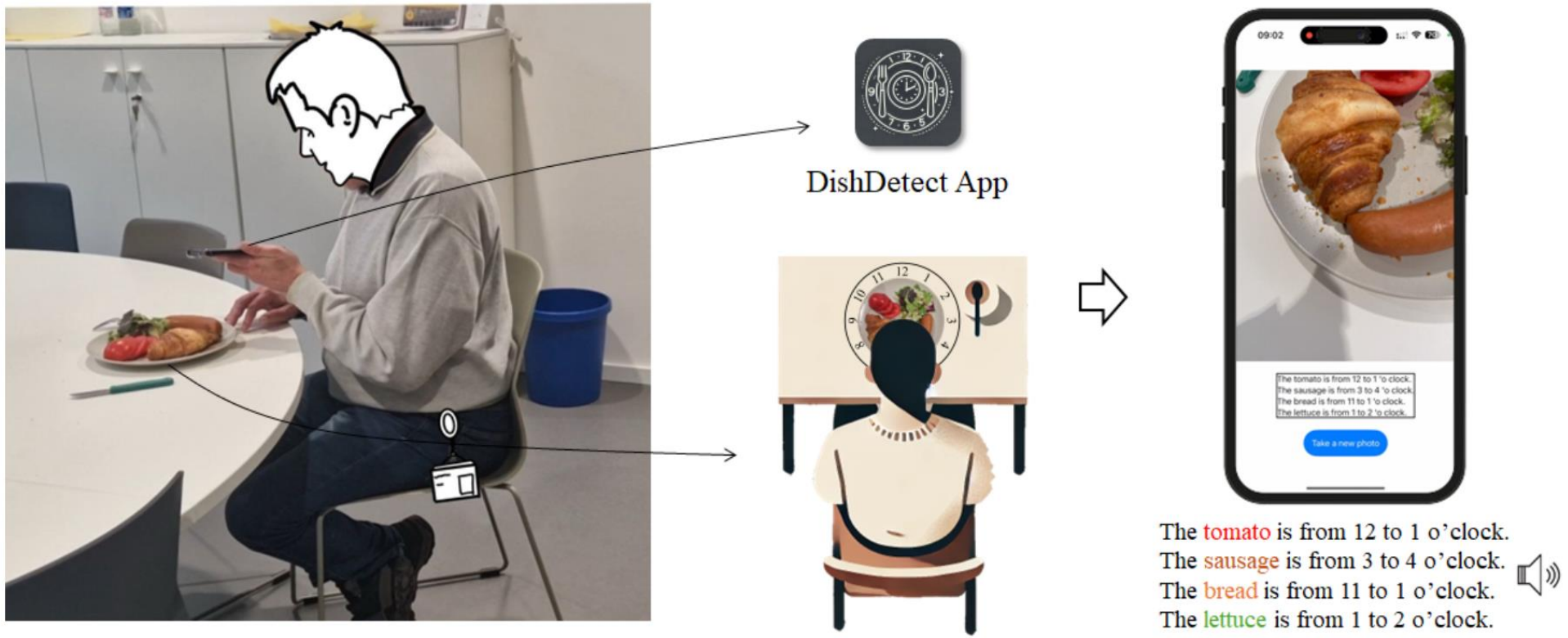
Examples from previous semesters: SS22 – Interactive PET/CT annotation, ISBI 2023



Examples from previous semesters:
SS23 – AR-guided 3D Foot Object Acquisition (MIDL 2024)



Examples from previous semesters:
SS24 – DishDetect: Mobile app for food description (CHI 2025)



General Information

Weekly meeting (MS Teams)

- Compulsory Attendance
- Talk about intermediate results & problems
- Ask for help and guidance
- Weekly goal: stay on “track”

2-3 Students per Team

- Use version control (e.g. git)
- Internal git repos provided via the SCC’s GitLab (https:/git.scc.kit.edu/)
- Divide work into separate tasks and distribute within group

At the end of the Practical Course…

- Final presentation of each group (1/3 grading)
  - 15 minute talk per group
  - The presentation should be about:
    - Goals and usefulness of your topic
    - Your proposed approach
    - Results
- Written report describing the topic/approach/results (1/3 grading)
  - 4-pages in standard paper format
    - Abstract/Introduction/Method/Results/Conclusion
    - References do not count in the 4-pages!
    - Written in a conference template
- Working implementations of your algorithms (1/3 grading)
  - A Readme-file describing how the code can be used to reproduce the results
    - If the team agrees → make code publicly available to the community

Topics SS 2026

- **A:** Benchmarking MLLM for highly shaky video grounding and QA
- **B:** Egocentric Assembly Training Assessment
- **C:** A Benchmark for Heterogeneous Multi-Agent Coordination in Assistive Embodied AI
- **D:** Agentic Research
- **E:** Interactive Segmentation of Whole-body PET/CT Lesions

TOPIC A

Supervisors:
Dr.-Ing. Kunyu Peng (kunyu.peng@kit.edu)

Topic A: Benchmarking MLLM for highly shaky video grounding and QA

Goal: Evaluate multimodal large language models (MLLMs) on *shaky, real-world running videos*

Evaluation categories:

- Object and attribute QA
  - Basic perception of entities in the video
- Action and event QA
  - Answer questions about actions in the scene
- Temporal Reasoning
  - Understand *order* and short-term memory
- Trajectory Reasoning
  - Use route data (turns, path direction)
- Robustness Testing
  - Handle blur, shake, night conditions

Question: What is the woman wearing white coat doing?
Answer: She is pushing a baby cart

Topic A: Benchmarking MLLM for highly shaky video grounding and QA

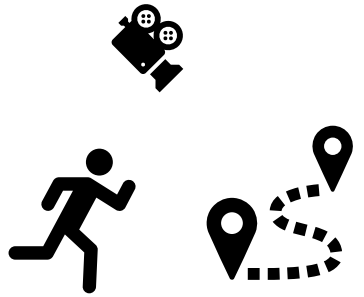
Potential questions for the benchmark:

- **[Object]:** How many pedestrians are visible at the intersection?
- **[Action]:** What is the person near the curb doing?
- **[Temporal]:** Is the stroller seen before or after the road crossing?
- **[Trajectory]:** Does the runner turn left or right after passing the pole?
- **[Robustness]:** Is this segment recorded during daytime or nighttime?

Topic A: Benchmarking MLLM for highly shaky video grounding and QA

Tasks:

- Build a **benchmark for MLLMs** on *shaky, real-world videos*
- Collect 10+ hours of first-person running videos in Karlsruhe (day & night)
- Record movement trajectories using mobile apps (e.g., Google Keep or similar GPS tracking apps)
- Include **different motion levels** (walking vs. running on same routes)
- Create annotation pipeline for:
  - **Video QA** (objects, actions, locations, motion)
- Evaluate robustness of MLLMs to:
  - Shake, blur, ego-motion
  - Rapid viewpoint changes
  - Lighting variations (day/night)



Topic A: Benchmarking MLLM for highly shaky video grounding and QA

Resources:

- Wen, Di, et al. "Go beyond Earth: Understanding human actions and scenes in microgravity environments." *arXiv preprint arXiv:2506.02845* (2025).
  - https://arxiv.org/abs/2506.02845

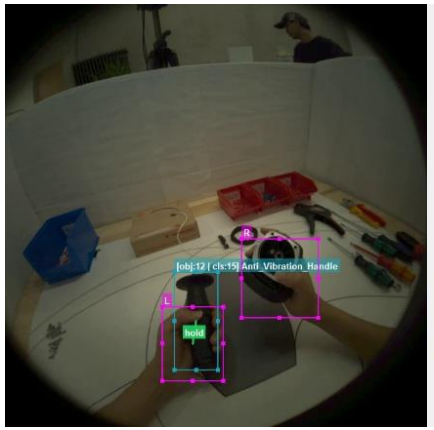
TOPIC B

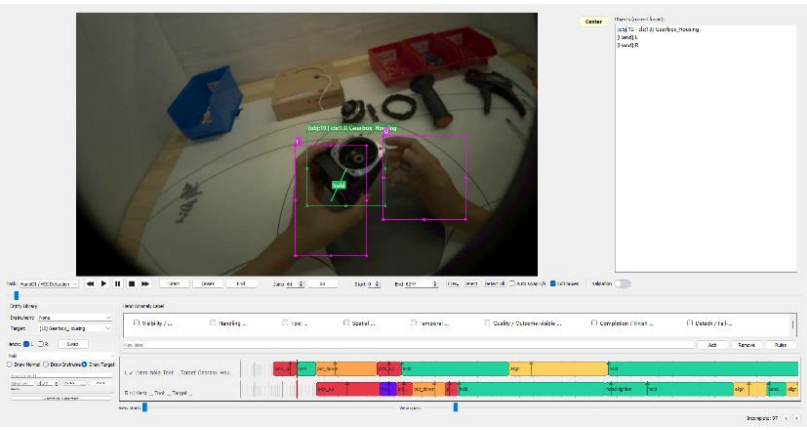
Supervisors:
Di Wen (di.wen@kit.edu)

Topic B: Egocentric Assembly Training Assessment
via Bimanual HOI Recognition and Multi-label Anomaly Detection

Goal: Build a system for **automatic worker training assessment from egocentric videos**

- **Input**: First-person assembly video
- Analyze assembly execution over time
- Detect common errors:
  - Missing steps
  - Wrong order
  - Incorrect tool use
  - Incomplete actions
- Compare against **expert reference procedure**
- Extract **visual evidence (key frames/clips)**
- **Output:** Interactive Error Debrief Report
  - Timestamped mistakes
  - Replayable evidence
  - Overall performance score

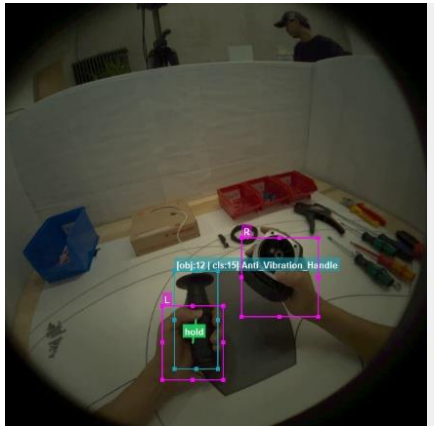


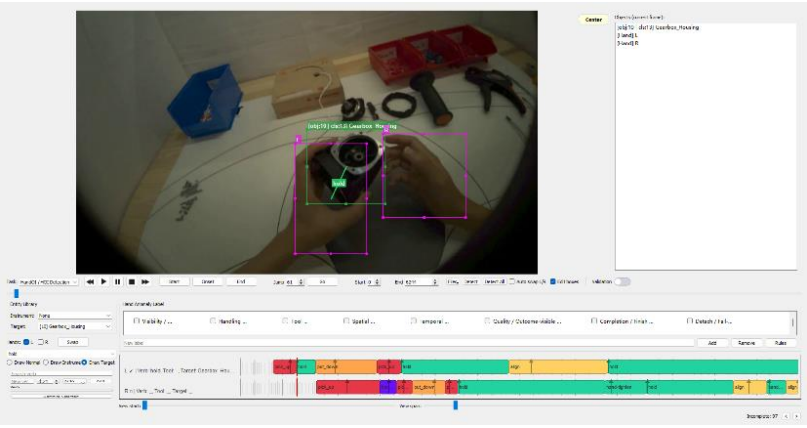


Topic B: Egocentric Assembly Training Assessment
via Bimanual HOI Recognition and Multi-label Anomaly Detection

Tasks

- Understand dataset annotations from provided industrial dataset
  - Step labels & boundaries, Error/anomaly labels, Tool/object info
- Implement a lightweight a temporal model
  - Predict framewise step labels, detect anomalies
- Implement procedure alignment model
  - Compare worker vs. expert sequence
  - Identify deviations (missing step, wrong order, etc.)
- Extract visual evidence
  - Key frames showing errors based on tools / hand interactions
- Design an interactive report interface that visualizes:
  - Timeline of steps
  - Highlighted errors
  - Replayable clips
  - Final score





Topic B: Egocentric Assembly Training Assessment
via Bimanual HOI Recognition and Multi-label Anomaly Detection

Resources

- Our collected and annotated egocentric industrial assembly/disassembly dataset.
- Optional reference datasets for literature study and protocol inspiration:
  - Assembly101 [4], MECCANO [5], EgoPER [6], IndustReal [7]

Related Work

[1] Abu Farha, Y., & Gall, J. MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[2] Ishikawa, Y., Kasai, S., Aoki, Y., & Kataoka, H. Alleviating Over-Segmentation Errors by Detecting Action Boundaries. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021.

[4] Sener, F., Chatterjee, D., Shelepov, D., et al. Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[5] Ragusa, F., Furnari, A., Livatino, S., & Farinella, G. M. The MECCANO Dataset: Understanding Human-Object Interactions From Egocentric Videos in an Industrial-Like Domain. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021.

[6] Lee, S.-P., Lu, Z., Zhang, Z., Hoai, M., & Elhamifar, E. Error Detection in Egocentric Procedural Task Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[7] Schoonbeek, T. J., Houben, T., Onvlee, H., de With, P. H. N., & van der Sommen, F. IndustReal: A Dataset for Procedure Step Recognition Handling Execution Errors in Egocentric Videos in an Industrial-Like Setting. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024.

TOPIC C

Supervisors:
Ruiping Liu (ruiping.liu@kit.edu)
Dr.-Ing. Kunyu Peng (kunyu.peng@kit.edu)

Topic C: A Benchmark for Heterogeneous Multi-Agent Coordination in Assistive Embodied AI

Goal: Evaluate **robot teamwork** for assistive tasks (mobility + errands)

- Scenario: **Blind user + multiple robots**
- Two agents work **in parallel**:
  - **Guide robot** → navigates user safely
  - **Assistant robot** → fetches & delivers objects
- Tasks include:
  - **Navigation** in complex indoor spaces
  - **Object manipulation** from language instructions
  - **Coordination & timing** between agents
- User simulated via **LLM-based interaction**
- Output: Successful **joint task completion** (guide + delivery)

Topic C: A Benchmark for Heterogeneous Multi-Agent Coordination in Assistive Embodied AI

Tasks:

- Implement navigation (guide agent):
  - Safe path planning, collision avoidance, user-friendly movement
  - Isaac Sim platform: physically grounded robot simulation
- Implement manipulation (assistant agent):
  - Understand speech instructions, find & grasp object, deliver to user (**fetch!**)
- LLM-driven blind user simulation:
  - Natural instructions, simple dialogue feedback with other agents
  - Mimic limited perception (no exact location)
- Evaluate agentic system for real-world transfer:
  - Task success, efficiency, safety & coordination
  - Examine how this can be transferred to the real-world

Topic C: A Benchmark for Heterogeneous Multi-Agent Coordination in Assistive Embodied AI

Resources:

- Wang et al. GRUtopia: Dream General Robots in a City at Scale, ICLR 2025.
- ISAAC Simulator: https://github.com/isaac-sim/IsaacSim

TOPIC D

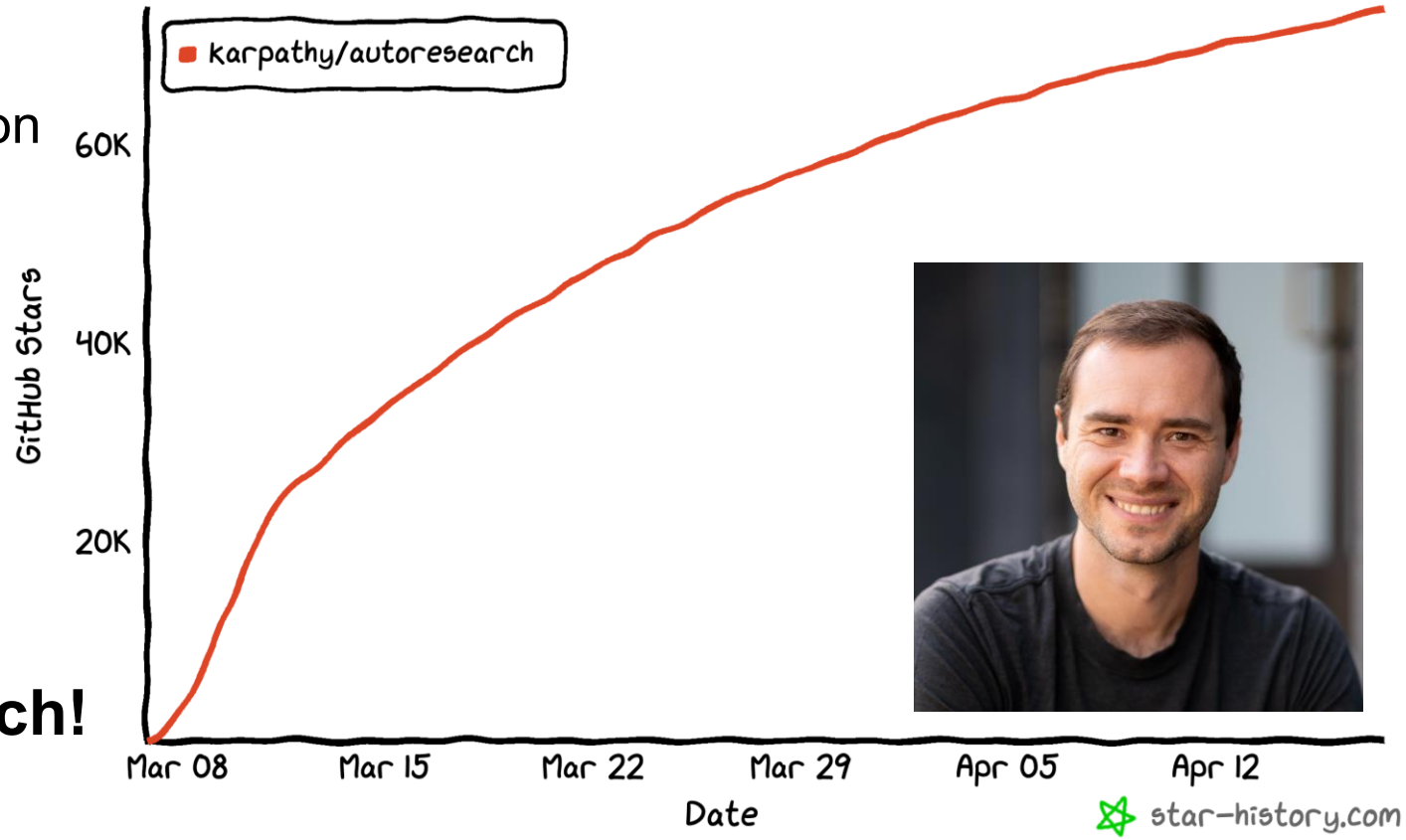
Supervisors:
Dr.-Ing. Alexander Jaus (alexander.jaus@kit.edu)
Jiale Wei (jiale.wei@kit.edu(

Topic D: Agentic Research

Goal: Explore the capabilities of modern LLMs to drive the innovation across potentially three pillars of Model Development:

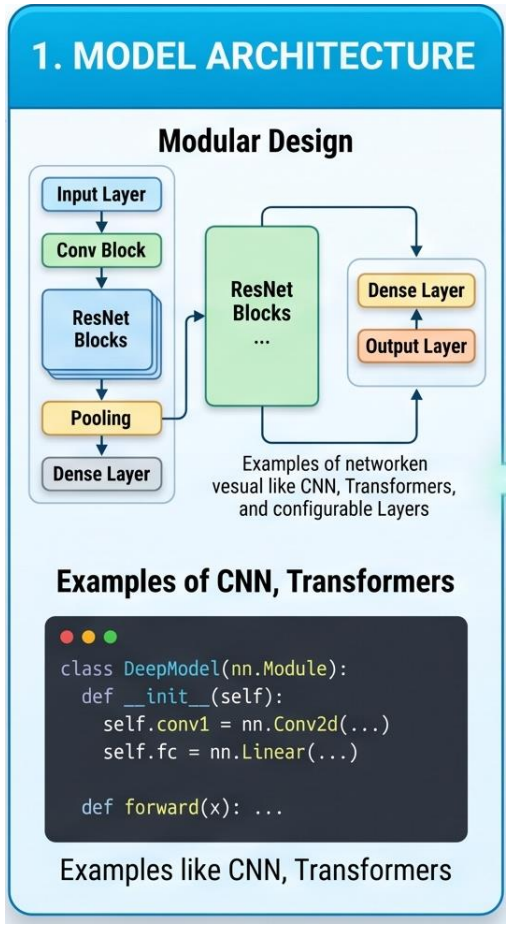
1) Model-architecture
2) Training hyperparameters
3) Data pipeline

Let the agents do the research!



https://github.com/karpathy/autoresearch

Topic D: Agentic Research

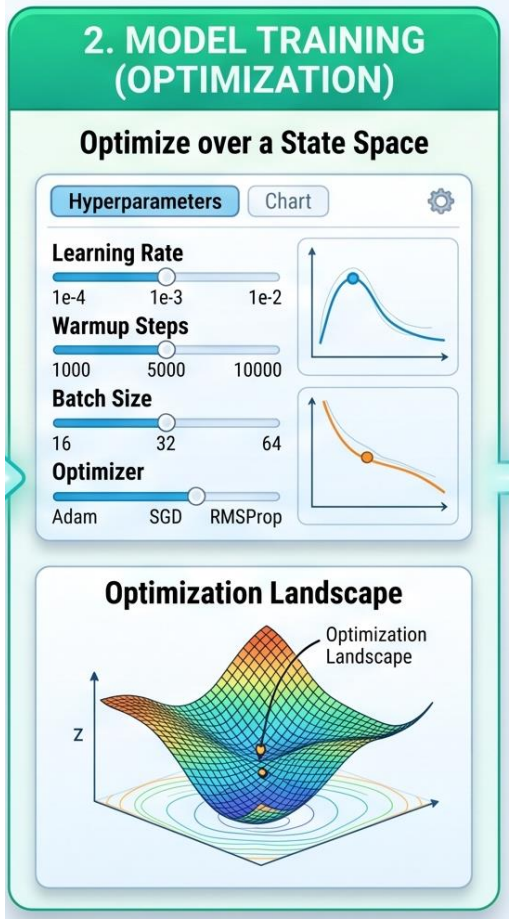


Model architecture:

- State-space: configurable blocks
- Free-form Python code

Agents may either either optimize in a constrained setting or write free-form Python code

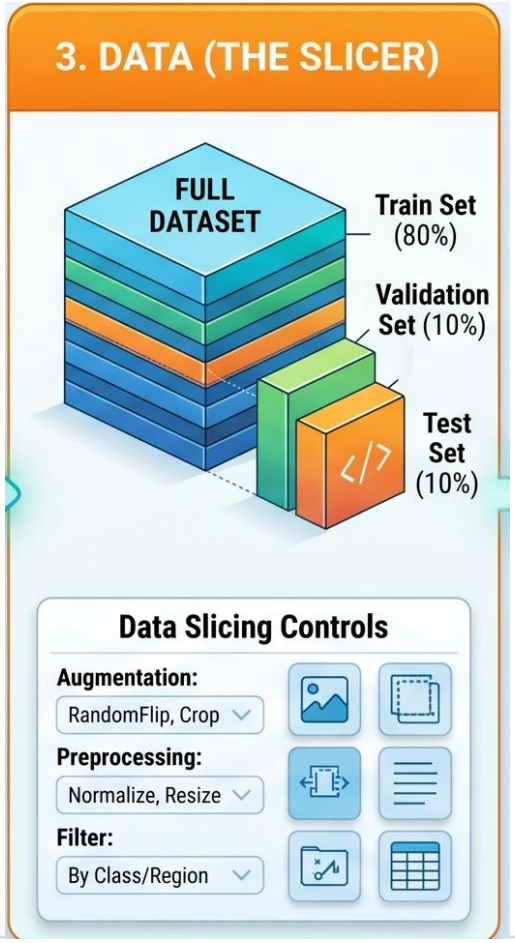
Topic D: Agentic Research



Model hyperparameters:

- A grid the agent may traverse
- Possible baselines:
  - Brute-force grid search (upper bound)
  - Evolutionary algorithms
  - Bayesian optimization

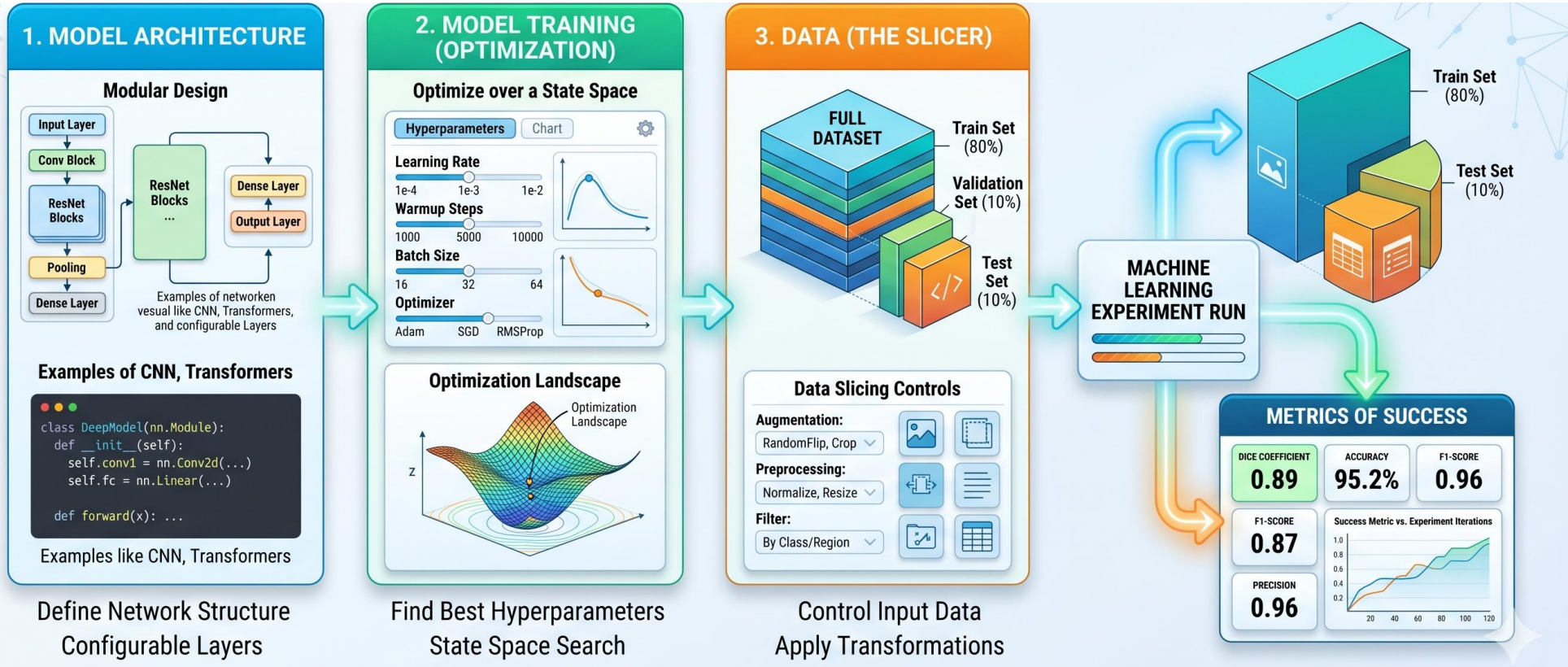
Topic D: Agentic Research



Training data:

- Augmentation, preprocessing, data filtering
- Easy: (yes/no data augmentation)
- Difficult (data filtering → larger search space)

Topic D: Agentic Research



Topic D: Agentic Research

Resources:

- Autoresearch: https://github.com/karpathy/autoresearch

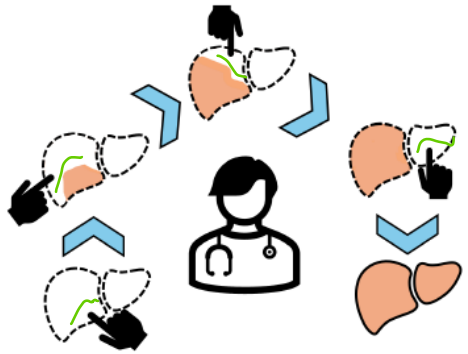
TOPIC E

Supervisors:
Dr.-Ing. Zdravko Marinov (zdravko.marinov@kit.edu)

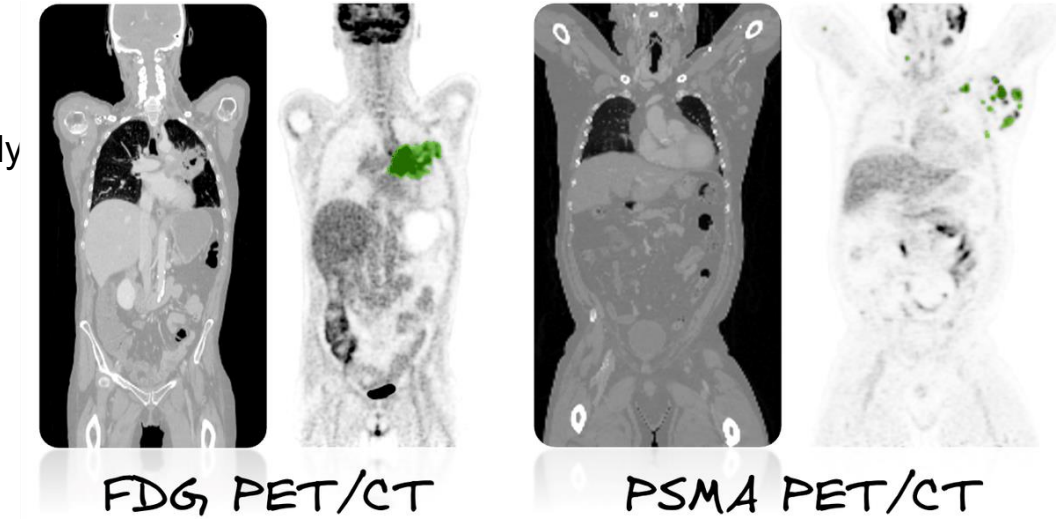
Topic E: Interactive Segmentation of Whole-body PET/CT Lesions

Goal: Train a scribble-based interactive segmentation model for whole-body PET/CT lesions

- PET/CT tumor segmentation is difficult
  - Automatic approaches miss small lesions and confuse anatomy and pathology
- Interactive segmentation:
  - User provides scribbles to correct model output
  - Difficult lesions are detected and user can continue correcting indefinitely



Scribble-based Interactive Segmentation

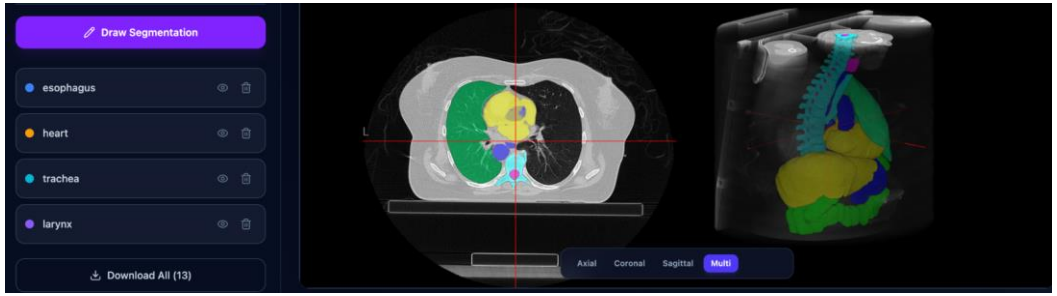


Source: https://autopet-v.grand-challenge.org/

Topic E: Interactive Segmentation of Whole-body PET/CT Lesions

Tasks:

- Participate in the autoPET V challenge
  - Train a scribble-based interactive segmentation model
    - Utilize scribble corrections to improve model predictions
  - Use Docker to create submission and compete worldwide with 100+ other participants
  - Out-of-competition since Zdravko Marinov is a co-organizer
    - Will be included in final journal publication if in top teams
- Implement a PET/CT web-based annotation tool (using NiiVue)
  - Use Dockerized model and run inference
  - Let user draw scribbles and predict with model
  - Load and save cases via centralized inference server



Topic E: Interactive Segmentation of Whole-body PET/CT Lesions

Resources:

- autoPET V challenge: https://autopet-v.grand-challenge.org/
- NiiVue 3D medical image visualization tool: https://github.com/niivue/niivue

Related Work:

- Marinov, Zdravko, et al. "Deep interactive segmentation of medical images: A systematic review and taxonomy." *IEEE transactions on pattern analysis and machine intelligence* 46.12 (2024): 10998-11018.
- Wong, Hallee E., et al. "Scribbleprompt: fast and flexible interactive segmentation for any biomedical image." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.
- Gatidis, Sergios, et al. "A whole-body FDG-PET/CT dataset with manually annotated tumor lesions." Scientific Data 9.1 (2022): 601.

Topic Selection

- Find a team of 2-3 people (i.e. through the MS-Teams chat)
- Each team sends us a ranking of the presented topics **until 27th 23:59 of April** per Email at zdravko.marinov@kit.edu (1 – most preferred; 5 – least preferred)
  - Example: A2, B4, C1, D3, E5
  - **Late e-mails lead to exclusion of the practical course**
- If you cannot find a team, you can also send personal preferences

- Students will be assigned to the respective topics based on their preferences and the order of registration

Organization

- Meeting schedule
  - Week 0 [20.04.26]: Introduction and topic selection
  - Week 1: Read related work and present ideas on how to approach the problem
  - Week 2: Implementation
  - …
  - **Week 14 [27.07.26] (Monday 14:00-16:00): Final Presentations**
  - **Week 14 [01.08.26]: Deadline written reports (01/08/2026, 23:59, CET)**
- Weekly meeting for discussion and status updates with corresponding supervisor
  - Set a consistent date for weekly meetings
- Register Projektpraktikum with KIT's Studienbüro
  - **Deadline: 04.05.2026, 23:59, CET**
  - If you are not registered by the deadline, you are not considered for the course!
- For these slides, other information, announcements and updates → check website [coursemember/321meins]