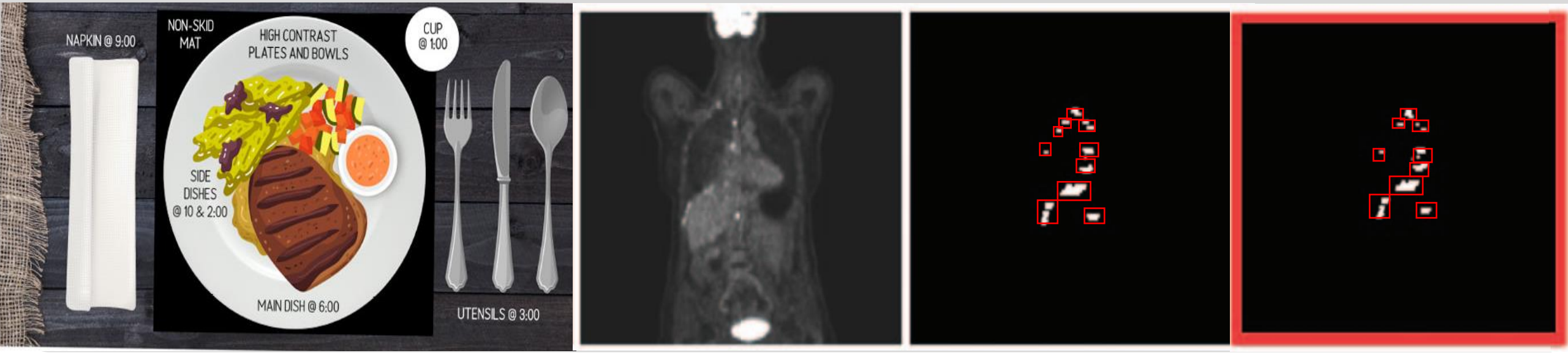
Karlsruher Institut für Technologie

cv:hci @ KIT

Practical Course: Computer Vision for Human-Computer Interaction

SS 2024

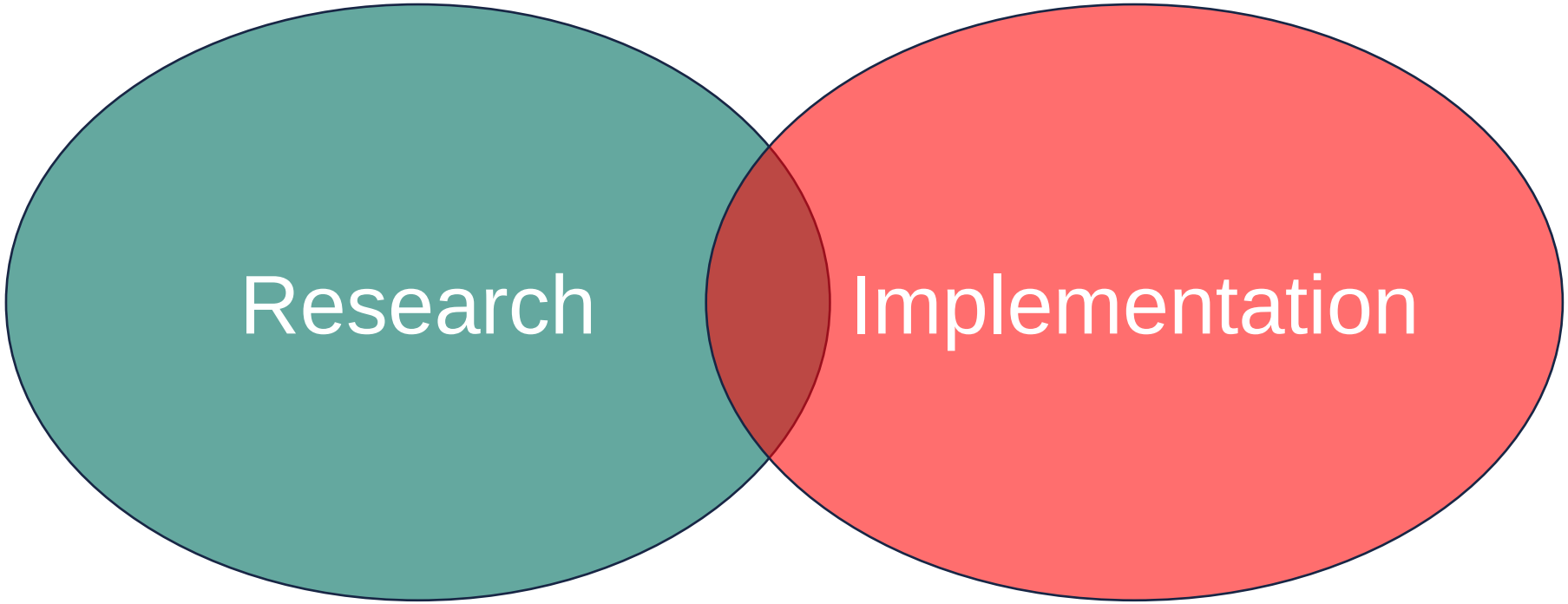
M.Sc. Zdravko Marinov
Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology (KIT)



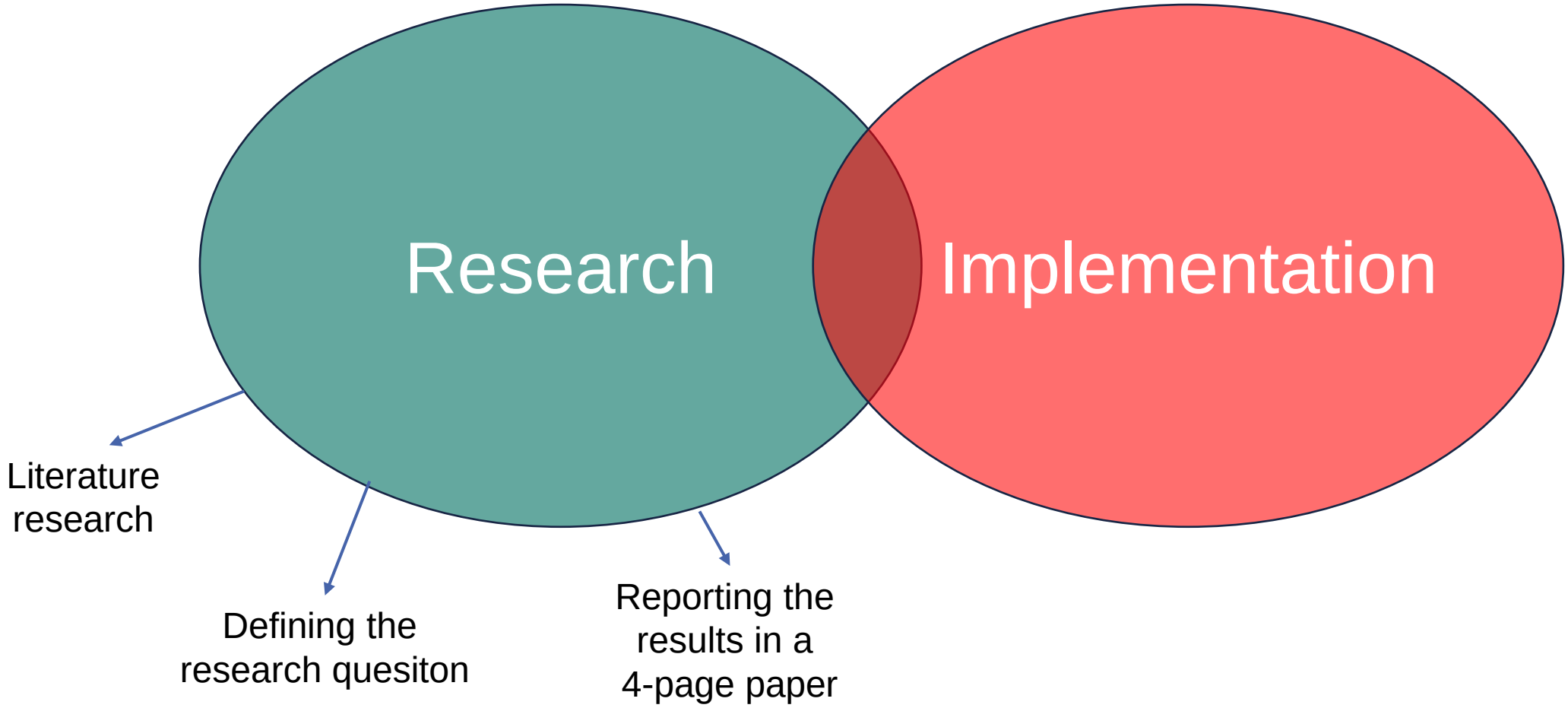
What will you learn?

- 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

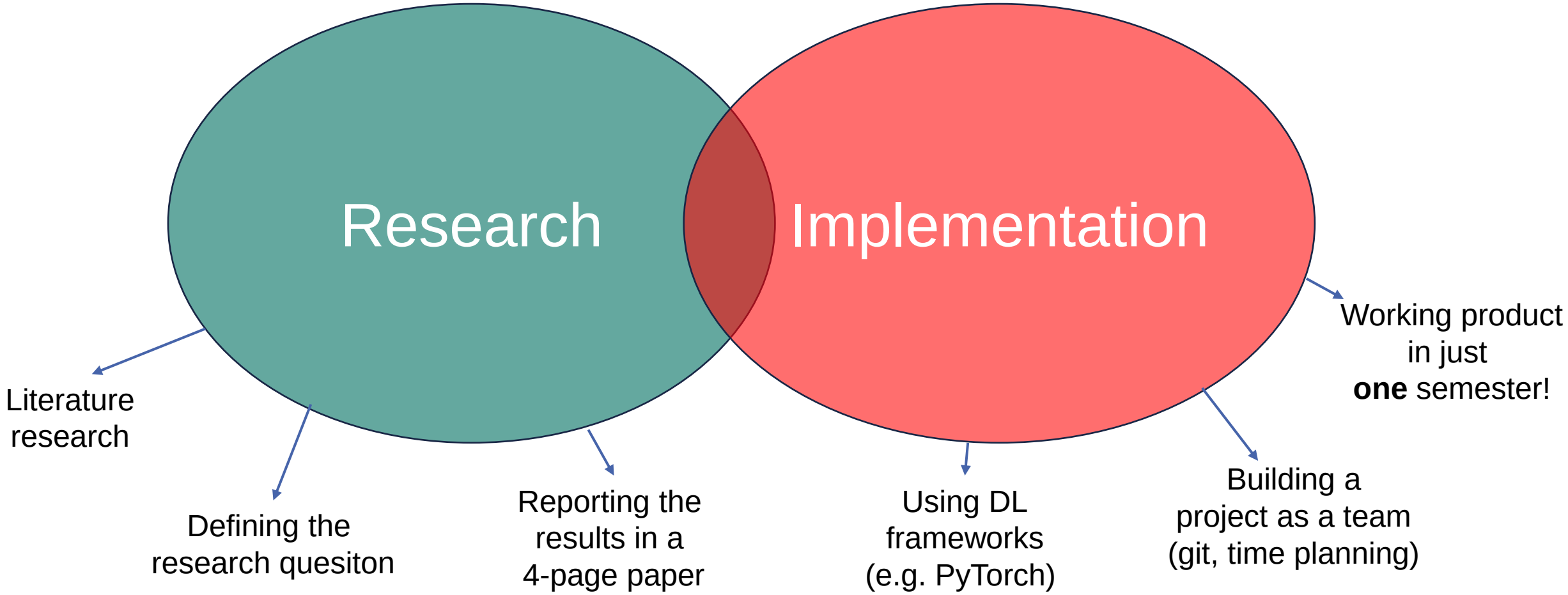
Scope of the Practical Course



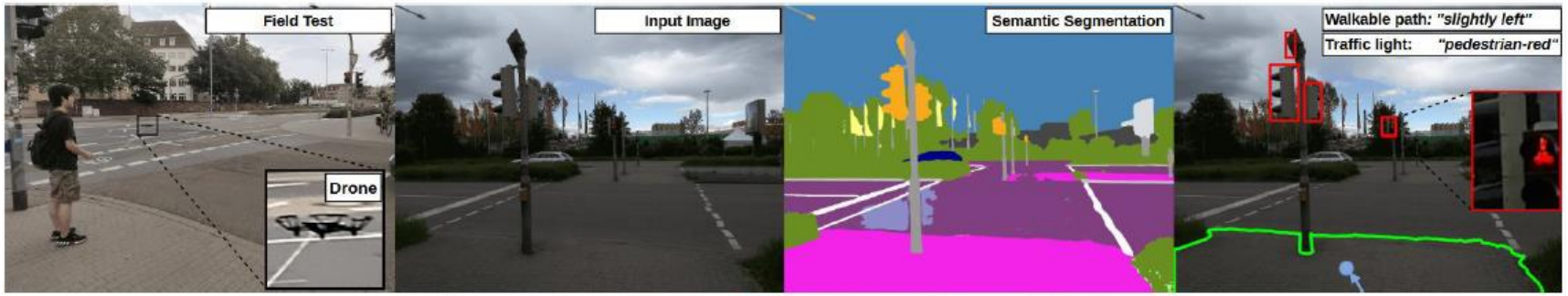
Scope of the Practical Course



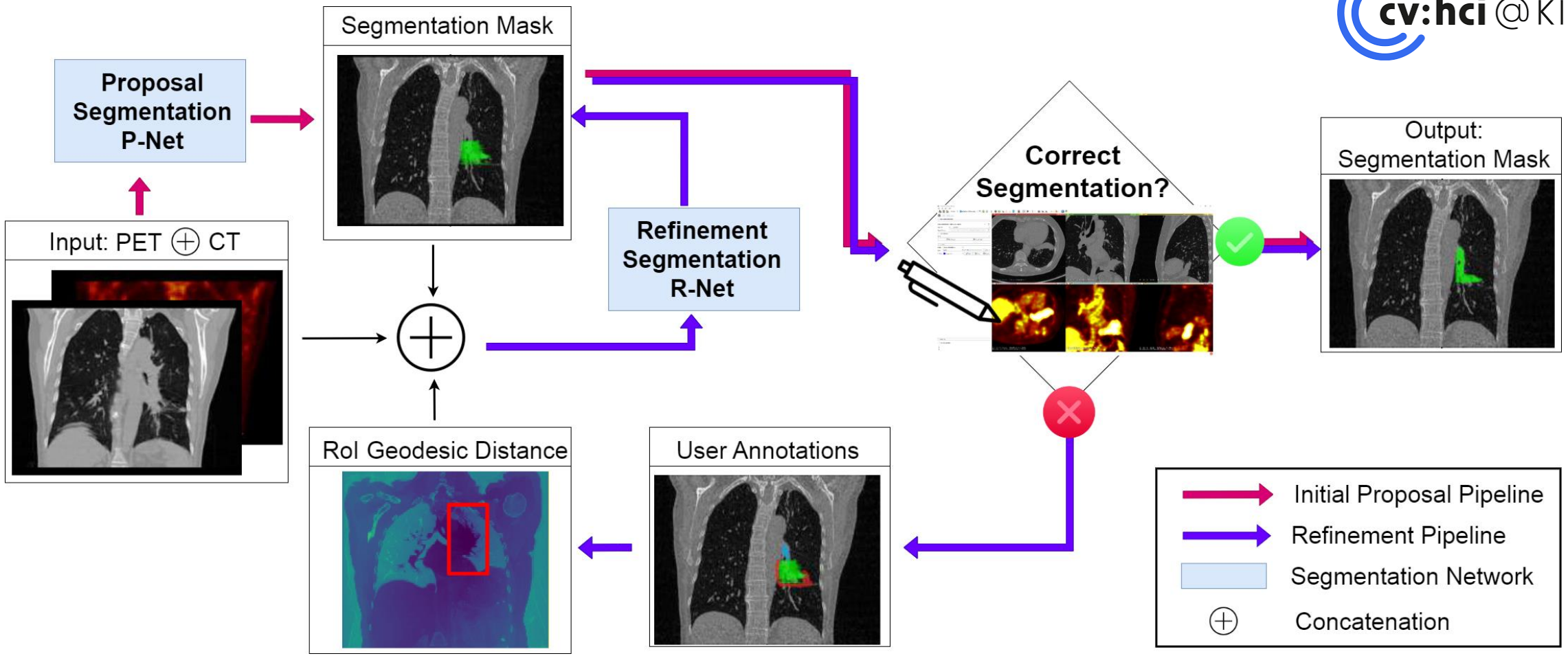
Scope of the Practical Course



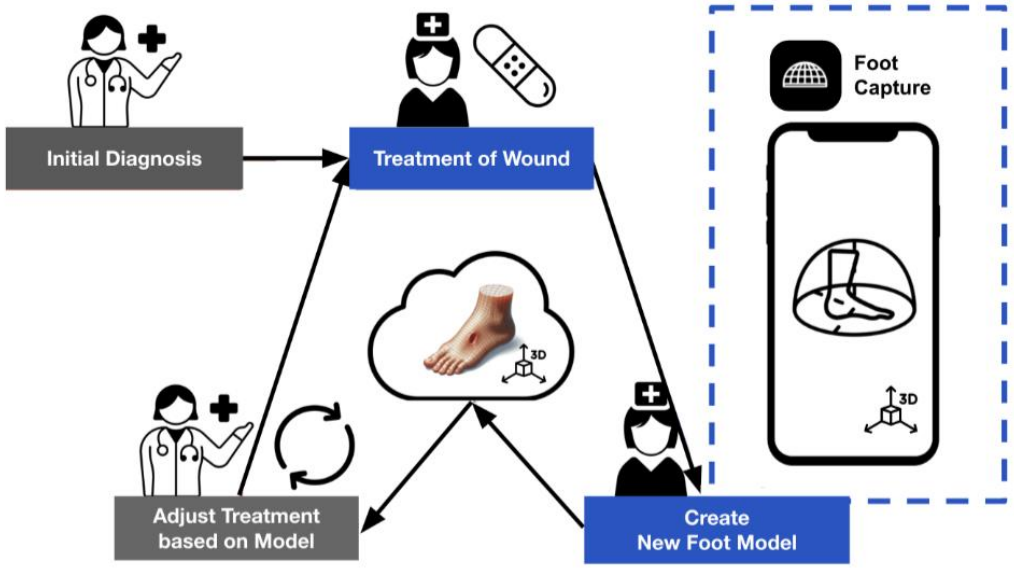
Examples from previous semesters: SS21 – Flying Guide Dog, ROBIO 2021

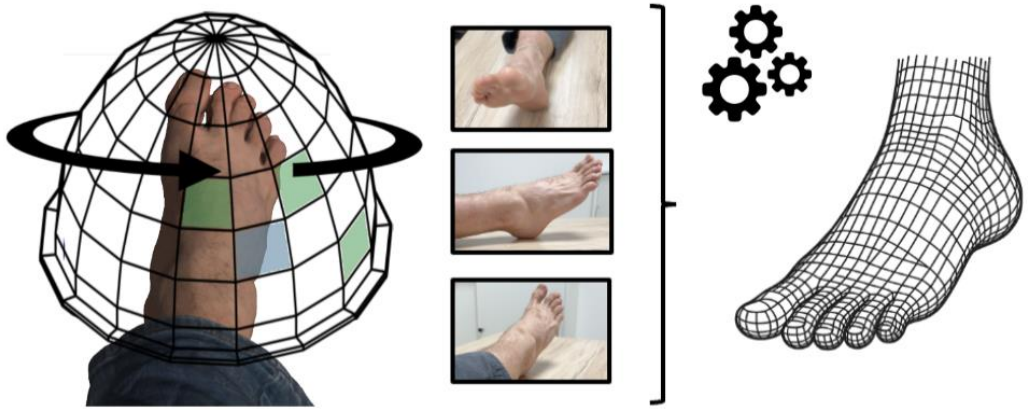


Examples from previous semesters: SS22 – Interactive PET/CT annotation, ISBI 2023



Examples from previous semesters:
SS23 – AR-guided 3D Foot Object Acquisition





General Information

Weekly meeting (MS Teams)

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

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 (5min/student)
  - 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 2024

- **A:** Cancer Detection in volumetric PET/CT images
- **B:** User-friendly Visual In-Context Learning
- **C:** High-Quality Document Image Capturing
- **D:** What's on my plate? An AI-based system to describe the food on a plate for blind people
- **E:** Skeletal Mamba for driver activity recognition
- **F:** Universal click-based interactive segmentation of medical images

6 topics distributed across 6 teams x 3 students

TOPIC A

Supervisors:
Alexander Jaus (alexander.jaus@kit.edu)
Zdravko Marinov (zdravko.marinov@kit.edu)

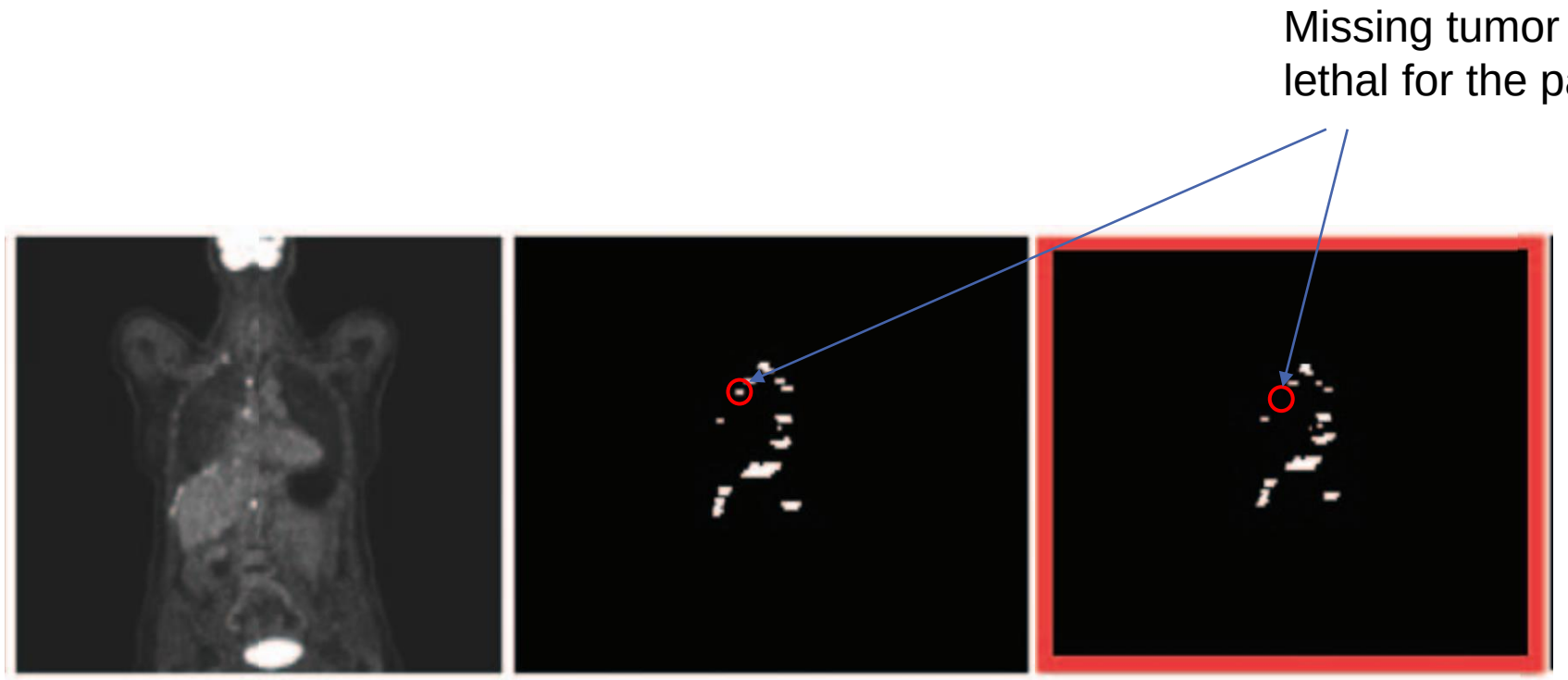
Topic A: Cancer Detection in volumetric PET/CT images

- The detection of cancer is a crucial task for radiologists.
- Recent works tackle cancer detection mostly as a **semantic segmentation** task

PET Image | Ground-truth | Prediction (99% IoU)

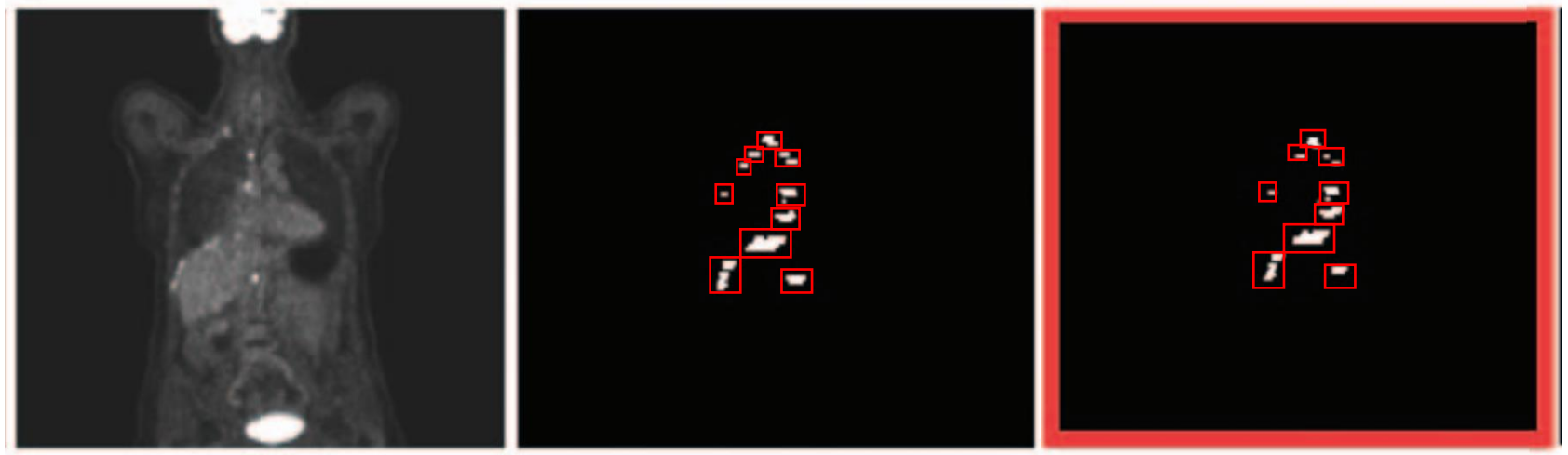
Topic A: Cancer Detection in volumetric PET/CT images

- The detection of cancer is a crucial task for radiologists.
- Recent works tackle cancer detection mostly as a **semantic segmentation** task



Topic A: Cancer Detection in volumetric PET/CT images

- We opt to formulate this task as an instance segmentation task in which large and small and large tumours are of equal importance



PET Image | Ground-truth | Prediction (80% recall)

Topic A: Cancer Detection in volumetric PET/CT images

Task

- Extend an existing PET/CT dataset [1] with semantic segmentation annotations to an instance-aware dataset by treating each connected component as a separate instance.
- Evaluate existing semantic and instance-aware segmentation models [2, 3, 4] on the novel dataset in various metrics (e.g. mAP)
- Improve and rework existing models to prioritize the discovery of cancer over perfect segmentation

Topic A: Cancer Detection in volumetric PET/CT images

Resources

[1] Gatidis, Sergios, et al. "A whole-body fdg-pet/ct dataset with manually annotated tumor lesions." Scientific Data 9.1 (2022): 601 [link]

[2] Isensee, Fabian, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." Nature methods 18.2 (2021): 203-211 [link]

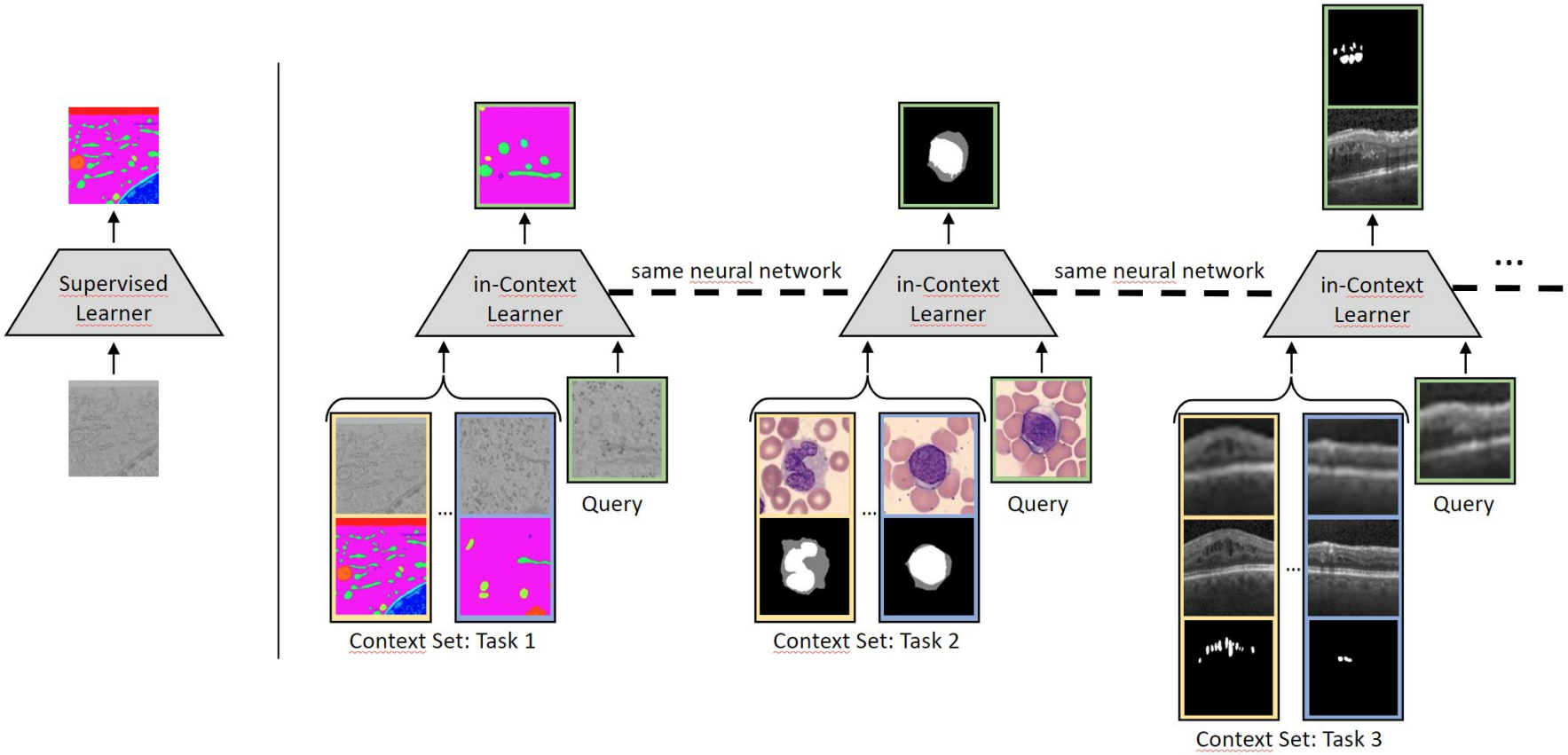
[3] Baumgartner, Michael, et al. "nnDetection: a self-configuring method for medical object detection." Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. Springer International Publishing, 2021 [link]

[4] Jaeger, Paul F., et al. "Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection." Machine Learning for Health Workshop. PMLR, 2020. [link]

TOPIC B

Supervisors:
Simon Reiß (simon.reiss@kit.edu)
Zdravko Marinov (zdravko.marinov@kit.edu)

Topic B: User-friendly Visual In-Context Learning



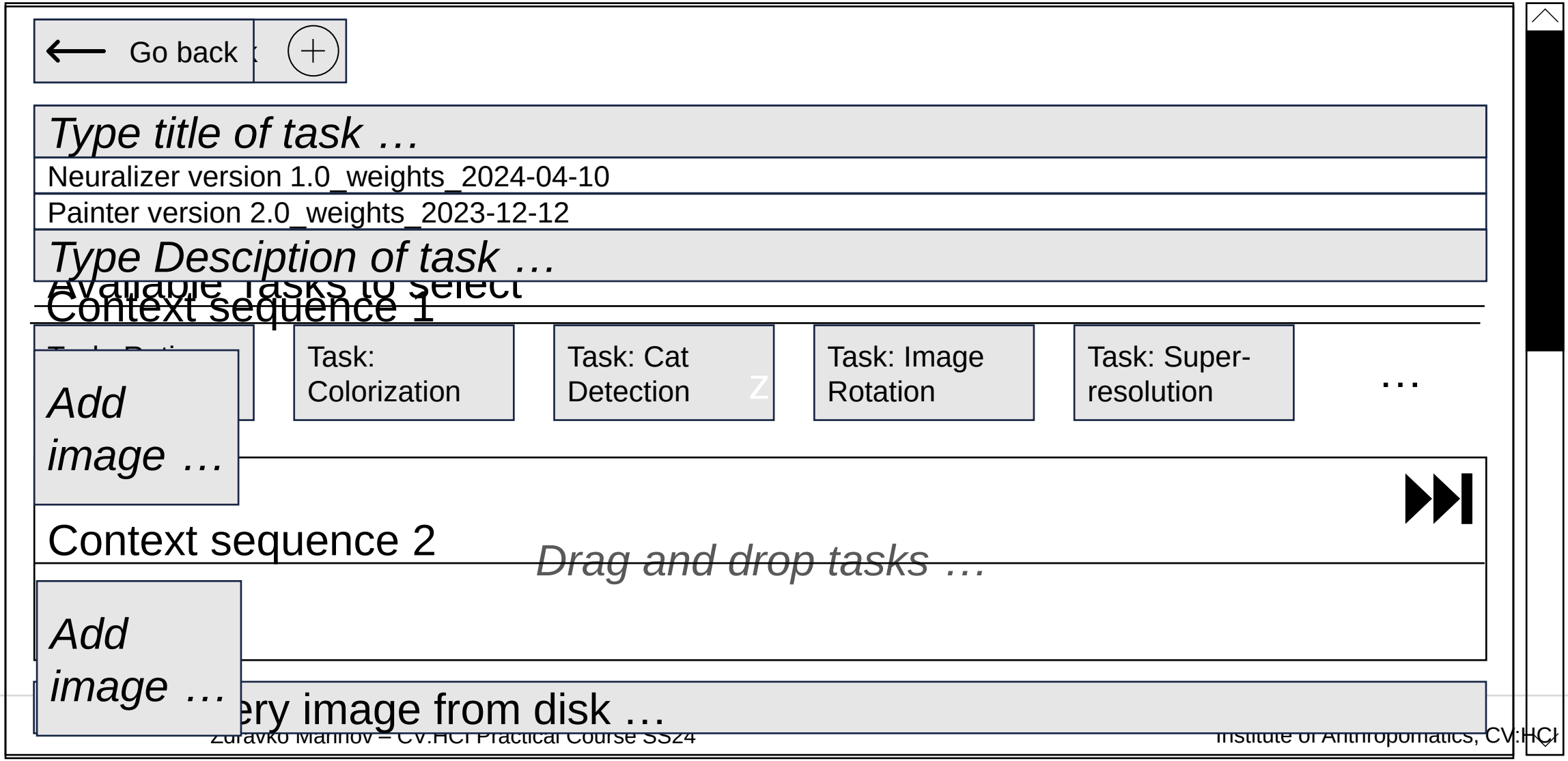
Example of In-Context Learning

Topic B: User-friendly Visual In-Context Learning

Task

- Enable simple interactions with in-Context Learners.
- Design and implement a web-based user interface
  - Integration of new tasks
  - Enable composite tasks
- Integrate existing pre-trained In-Context models
- Evaluate the usability of the interface in a user study

Topic B: User-friendly Visual In-Context Learning



Topic B: User-friendly Visual In-Context Learning

← Go back

Retinal Fluid Segmentation

Type Desciption of task …

Context sequence 1

Add image …

Context sequence 2

Add image …

Topic B: User-friendly Visual In-Context Learning

← Go back

Retinal Fluid Segmentation

The task involves finding pockets of fluid within oct-scans of the human retina, specifically subretinal fl…

Context sequence 1

Add image …

Context sequence 2

Add image …

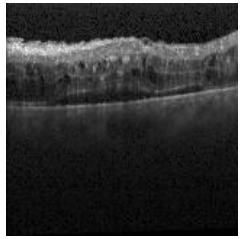
Topic B: User-friendly Visual In-Context Learning

← Go back

Retinal Fluid Segmentation

The task involves finding pockets of fluid within oct-scans of the human retina, specifically subretinal fl…

Context sequence 1



Add image …

Context sequence 2

Add image …

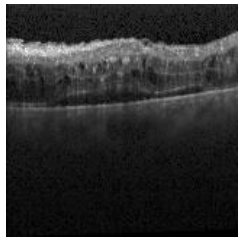
Topic B: User-friendly Visual In-Context Learning

← Go back

Retinal Fluid Segmentation

The task involves finding pockets of fluid within oct-scans of the human retina, specifically subretinal fl…

Context sequence 1



Add image …

Context sequence 2

Add image …

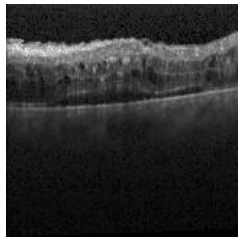
Topic B: User-friendly Visual In-Context Learning

← Go back

Retinal Fluid Segmentation

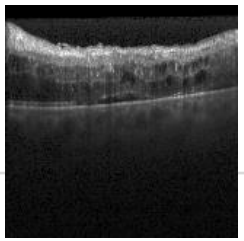
The task involves finding pockets of fluid within oct-scans of the human retina, specifically subretinal fl…

Context sequence 1



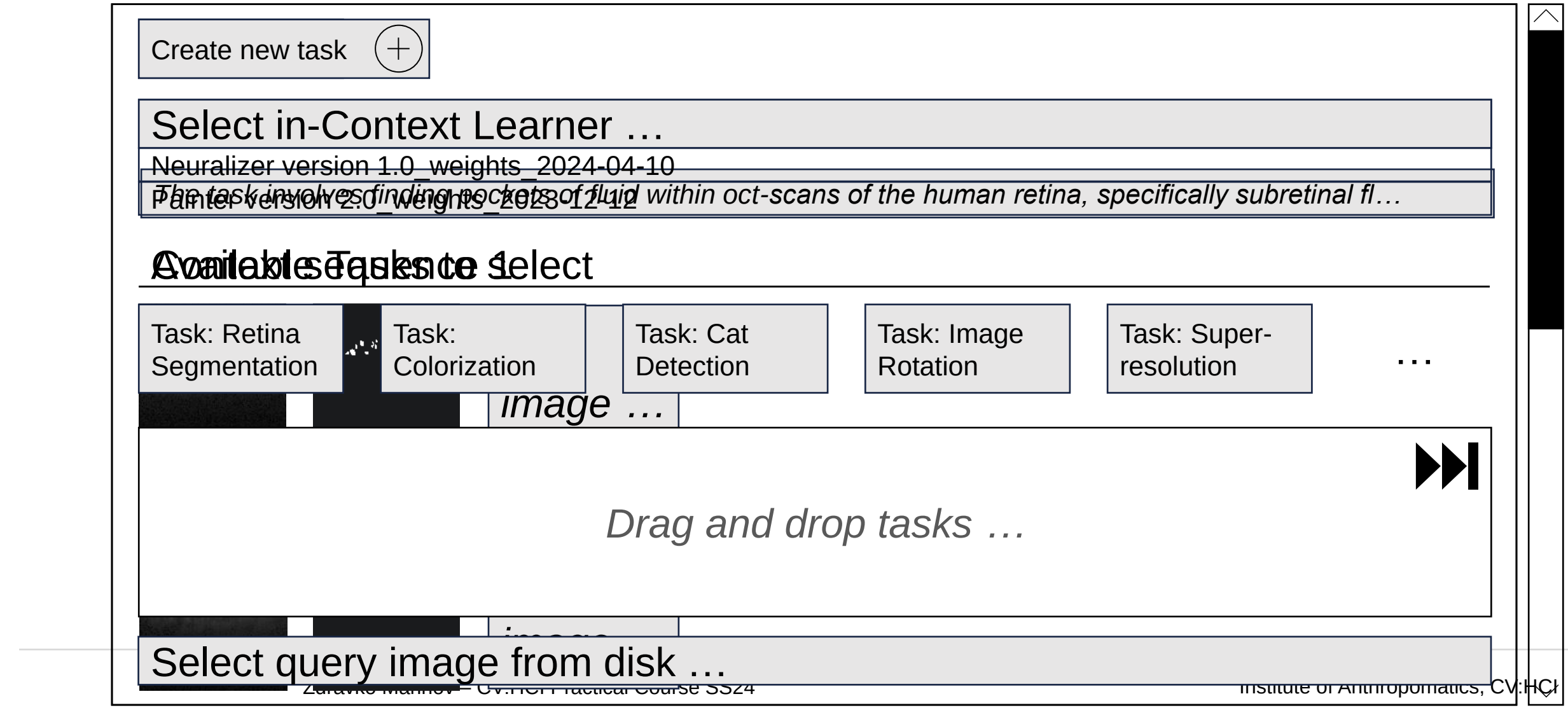
Add image …

Context sequence 2

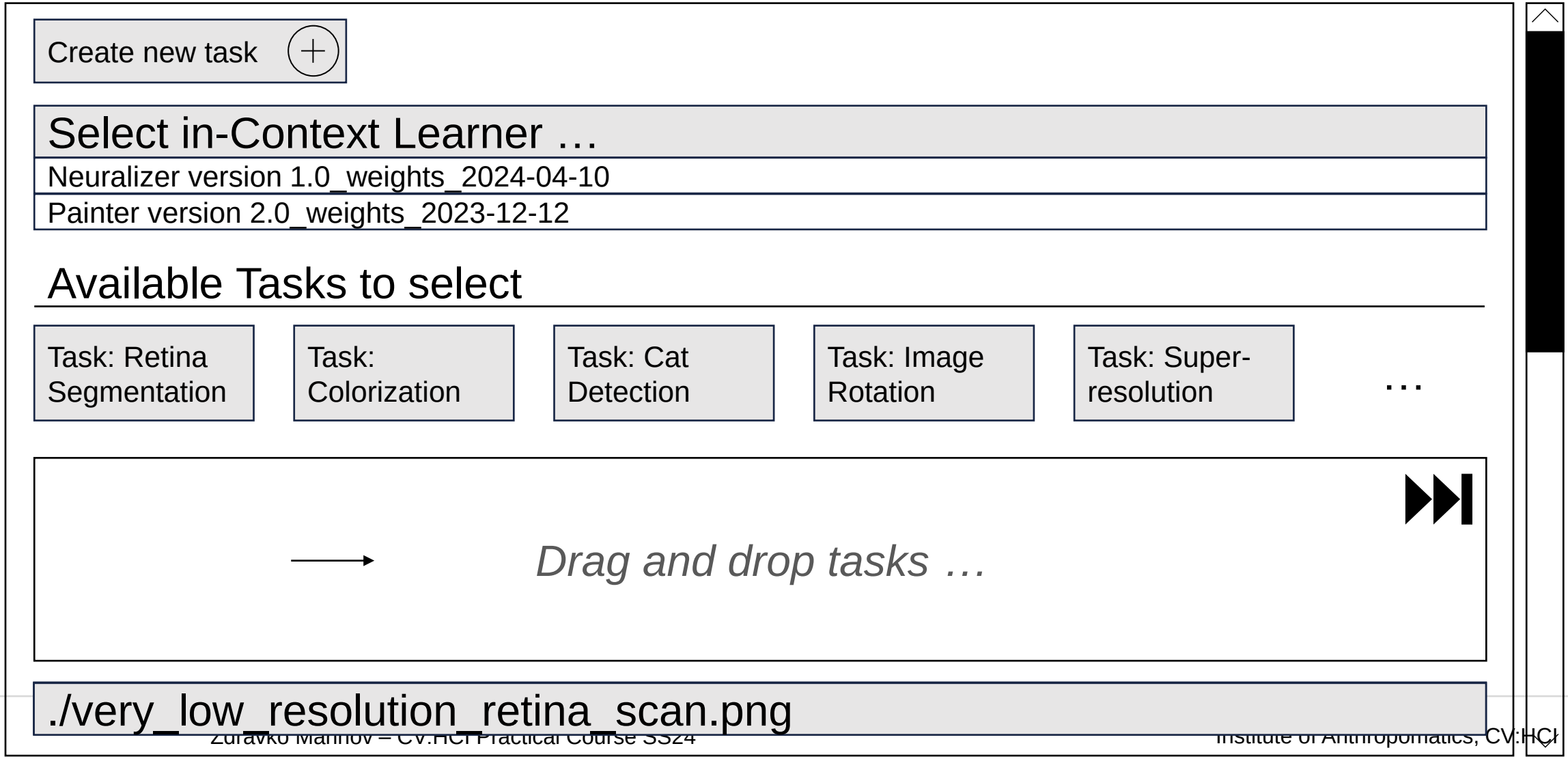


Add image …

Topic B: User-friendly Visual In-Context Learning



Topic B: User-friendly Visual In-Context Learning



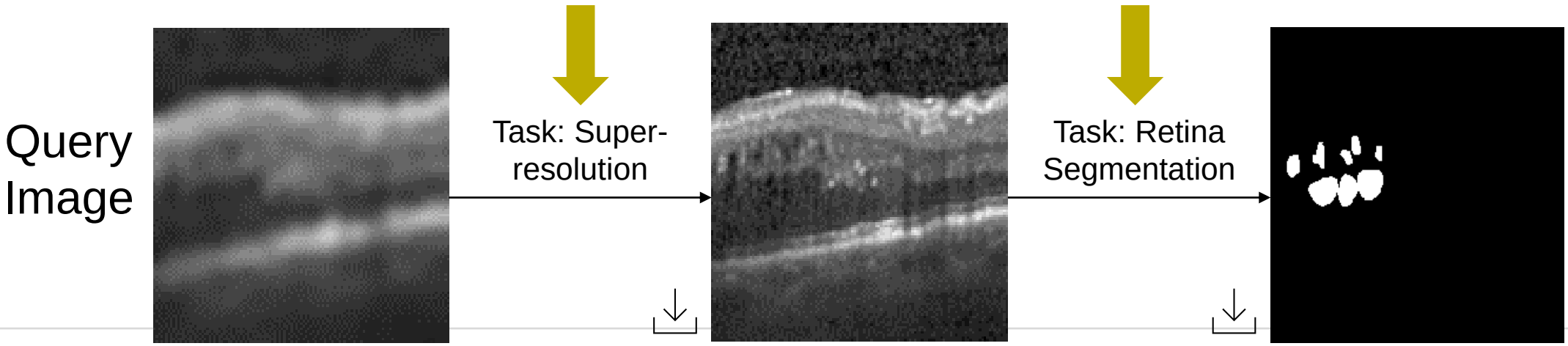
Topic B: User-friendly Visual In-Context Learning

Task: Super-resolution → Task: Retina Segmentation

ag and drop tasks …

./very_low_resolution_retina_scan.png

Compositional output



Topic B: User-friendly Visual In-Context Learning

Resources

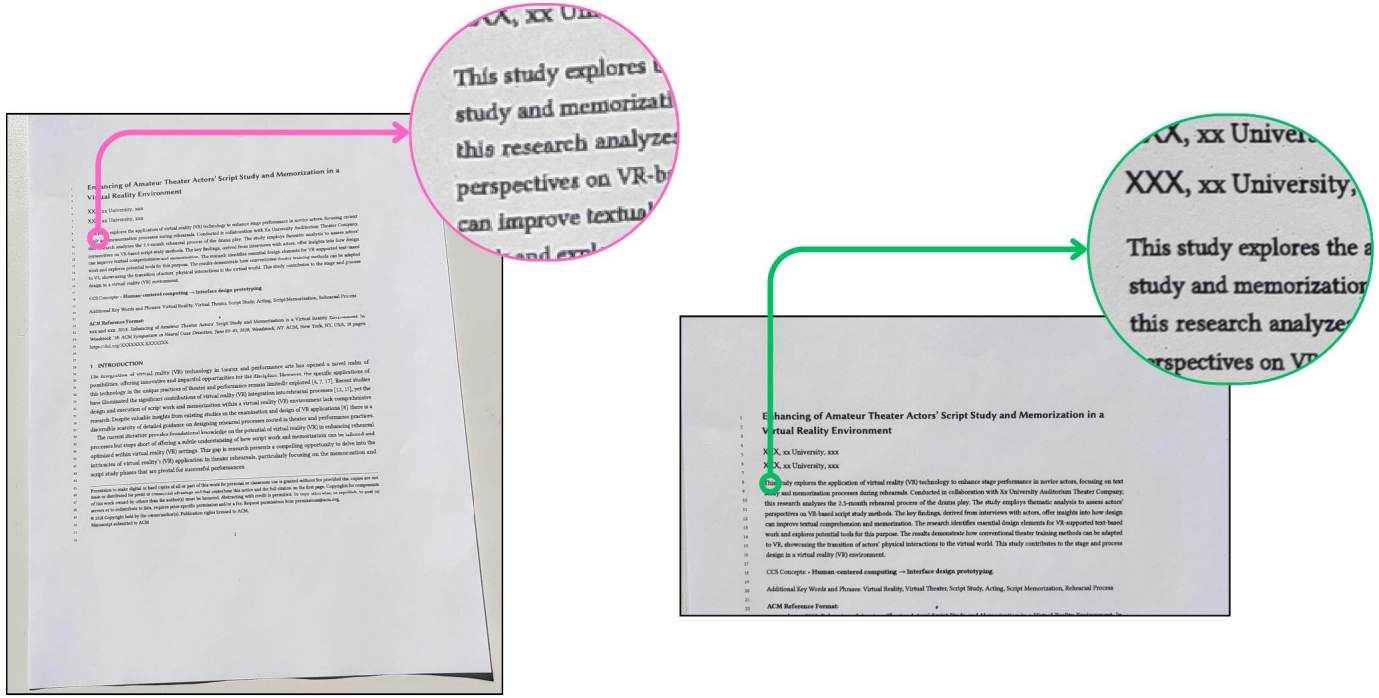
- [1] Czolbe, Steffen, and Adrian V. Dalca. "Neuralizer: General neuroimage analysis without re-training." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
- [2] Bar, Amir, et al. "Visual prompting via image inpainting." *Advances in Neural Information Processing Systems* 35 (2022): 25005-25017.
- [3] Bai, Yutong, et al. "Sequential modeling enables scalable learning for large vision models." *arXiv preprint arXiv:2312.00785* (2023).
- [4] Wang, Xinlong, et al. "Images speak in images: A generalist painter for in-context visual learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

TOPIC C

Supervisors:
Omar Moured (omar.moured@kit.edu)
Yufan Chen (yufan.chen@kit.edu)

Topic C: High-Quality Document Capturing

- Capturing documents with dense content, such as two-column layouts, may result in a reduced pixel count per character, as demonstrated in (a)
- Simply increasing resolution is not sufficient due to the blurring effect; employing a super-resolution model or capturing from a closer distance may be more effective, as shown in (b).



(a) (b)

Topic C: High-Quality Document Capturing

Task

- Select an appropriate dataset, e.g., M6Doc [4], or compile a custom dataset.
- Train and evaluate two methodologies:
  - Pre-trained super-resolution models [1, 2]
  - Document stitching approach [3]
- Evaluate the effectiveness of the aforementioned methodologies subtasks such as:
  - Document layout analysis

Process

- Gather and prepare a diverse set of document images for the experiments
- Experiment with two SOTA super-resolution approaches [1, 2]
- Experiment with the "Document Stitching" approach [3]
  - Using feature keypoints
- Investigate appropriate evaluation metrics specialized for High Quality Documents

Topic C: High-Quality Document Capturing

Resources

[1] Image super-resolution: A comprehensive review, recent trends, challenges and applications - ScienceDirect

[2] Scene Text Telescope: Text-Focused Scene Image Super-Resolution

[3] Image Stitching using OpenCV — A Step-by-Step Tutorial | by Paulson Premsingh | Medium

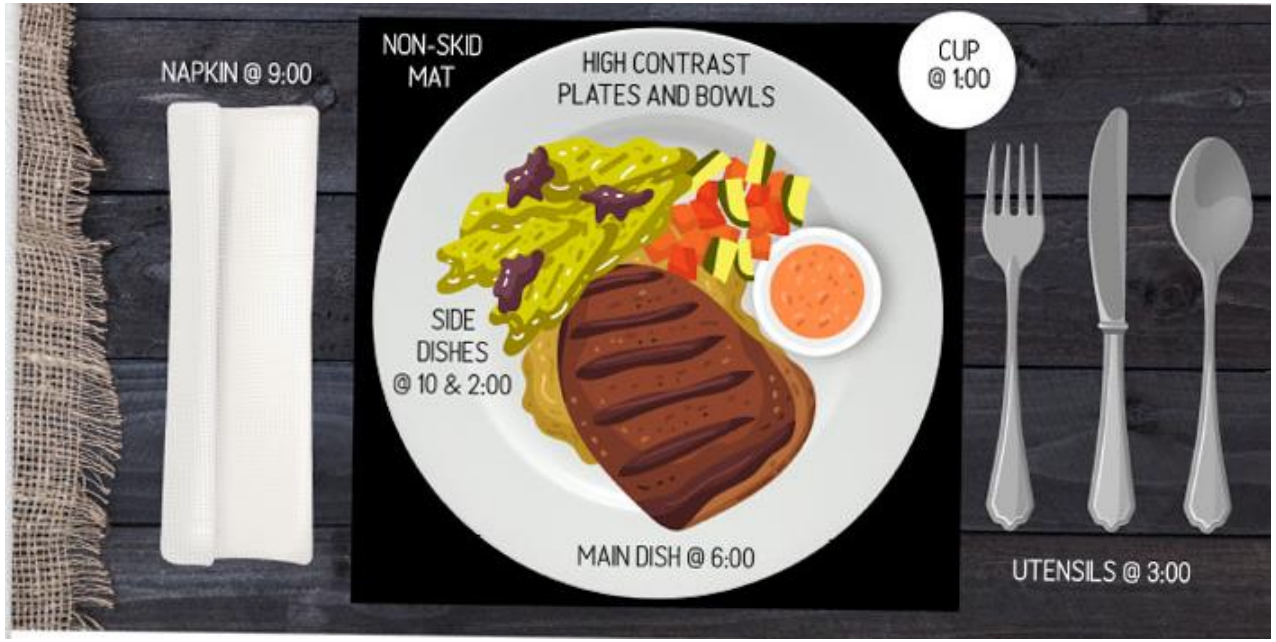
[4] https://github.com/HCIILAB/M6Doc

TOPIC D

Supervisors:
Ruiping Liu (ruiping.liu@kit.edu)
Karin Müller (karin.mueller2@kit.edu)

Topic D: What's on my plate? An AI-based system to describe the food on a plate for blind people

- Food placed on a plate for blind people can be described clockwise, e.g. potatoes from 9 to 12, vegetables from 1 to 3, and meat from 5 to 8.
- Usually a blind person knows what they ordered. Thus, a description using categories like meat, vegetables, rice etc. would be sufficient. More important is the clockwise description.



Source: https://therapyinsights.com/wp-content/uploads/2021/12/dining-with-low-vision.jpg

Topic D: What's on my plate? An AI-based system to describe the food on a plate for blind people

Task

- Develop or customize a food object detection model
  - Determine a dataset for the task, such as UNIMIB2016 [1], or create and annotate a dataset using photos captured within our cafeteria.
  - Pretrain the model using the Food2K dataset (image classification) [2]
- Integrate the model in a Jetson Nano and smart glasses equipped with stereo cameras.
- Assess food depth and detect its presence within the wearer's field of view.
- Provide auditory output with the location of detected food items.

Topic D: What's on my plate? An AI-based system to describe the food on a plate for blind people

Resources

[1] UNIMIB2016 http://www.ivl.disco.unimib.it/activities/food-recognition/

[2] Food2K http://123.57.42.89/FoodProject.html

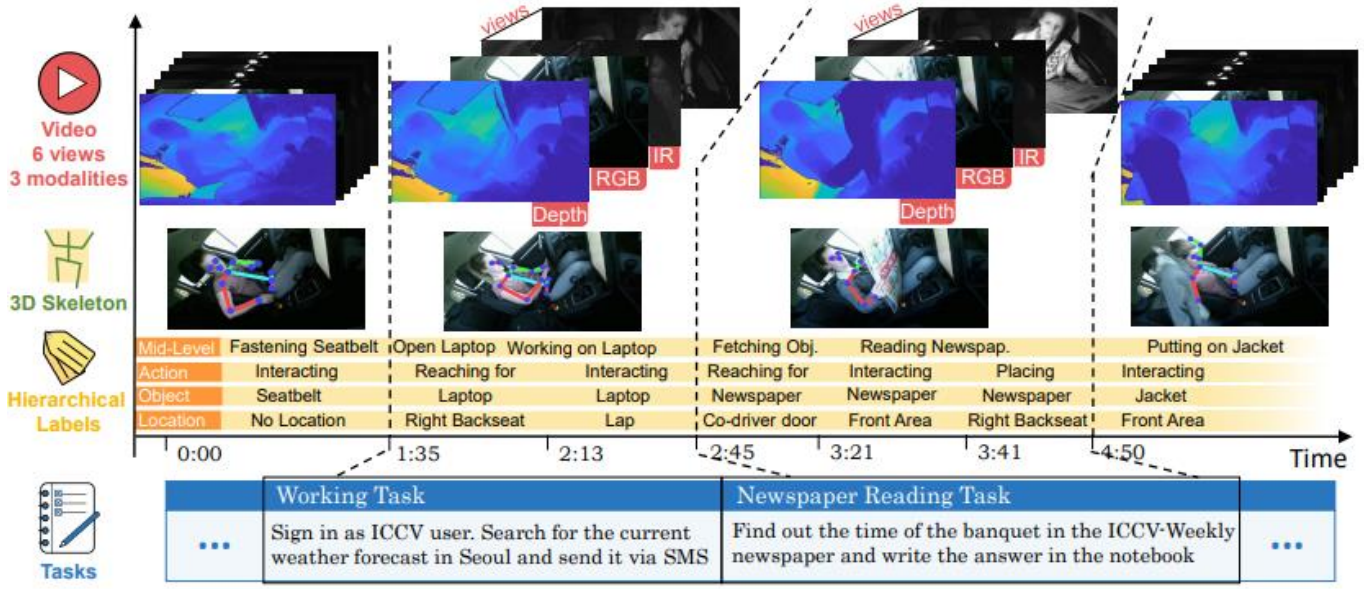
[3] Large scale visual food recognition, Min et al., Arxiv, 2021

TOPIC E

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

Topic E: Skeletal Mamba for driver activity recognition

- Driver activity recognition can be estimated using multiple sensors in the cockpit
- Skeleton poses are a reliable modality to classify the activity
  - Previous work mainly focuses on 3D-ConvNets and ViTs
- Visual state space models [1] have demonstrated a remarkable performance in multiple tasks
  - We aim to explore them in the field of driver activity recognition



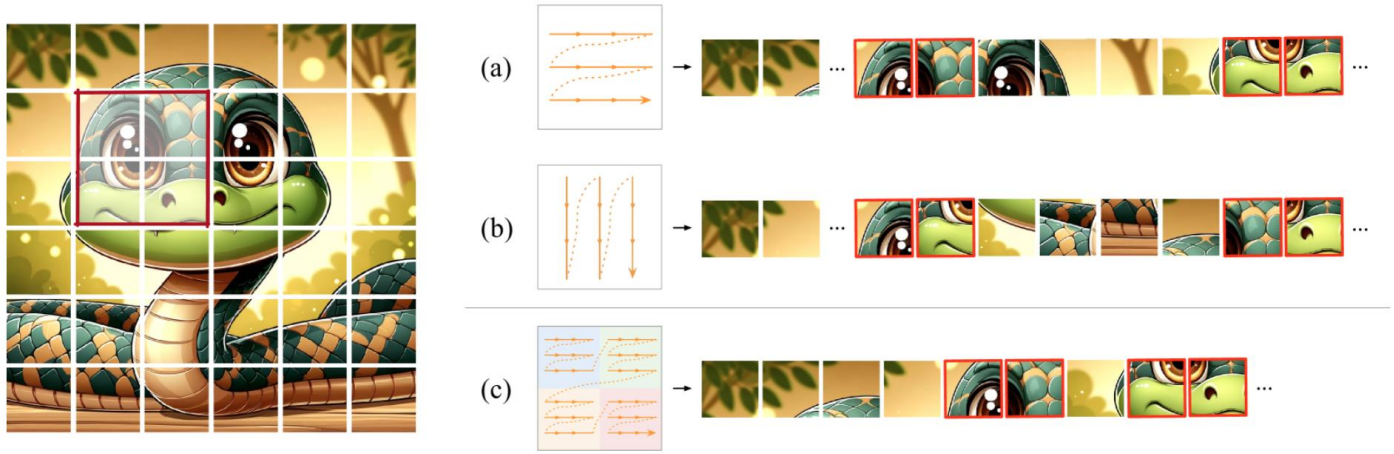
Topic E: Skeletal Mamba for driver activity recognition

Task

- Benchmark existing scanning methods [2] and propose a scanning method for skeleton data
- Analyze the resulting embeddings with tSNE [4]

Dataset

- Drive&Act [3]



Examples of scanning methods

Source : https://arxiv.org/html/2403.09338v1

Topic E: Skeletal Mamba for driver activity recognition

Resources

[1] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., ... & Liu, Y. (2024). Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*.

[2] Li L, Wang H, Zhang W, et al. STG-Mamba: Spatial-Temporal Graph Learning via Selective State Space Model[J]. arXiv preprint arXiv:2403.12418, 2024.

[3] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).

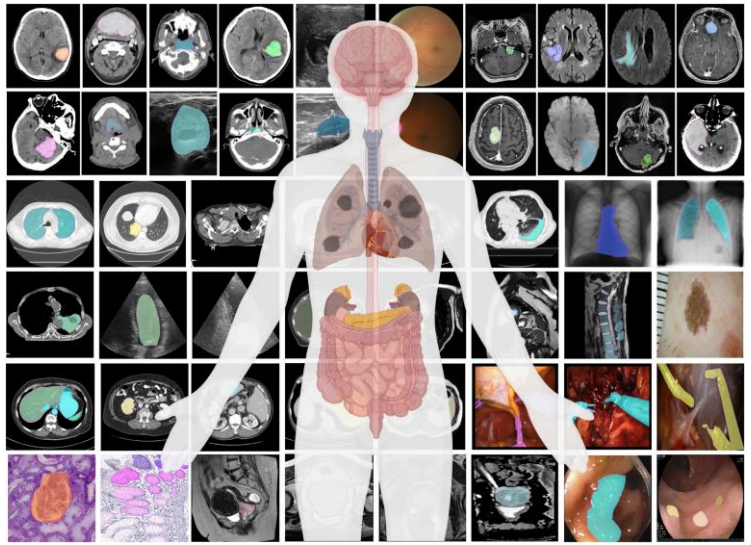
[4] Martin M, Roitberg A, Haurilet M, et al. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 2801-2810.

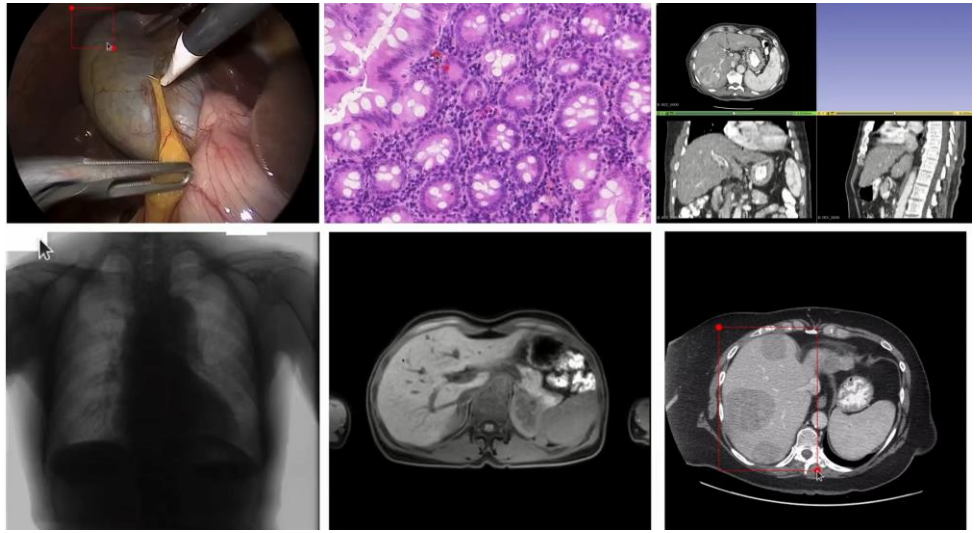
TOPIC F

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

Topic F: Universal click-based interactive segmentation of medical images

- Interactive segmentation uses clicks, scribbles, bounding boxes, and other interactions to guide a segmentation model toward the target obect
- Recently, MedSAM [1] released a universal model which works on multiple medical imaging modalities
  - However, it only uses bounding boxes
- We aim to extend MedSAM to clicks and implement an intuitive annotation interface





Topic F: Universal click-based interactive segmentation of medical images

Task

- Fine-tune MedSAM [1] using simulated clicks (e.g. in the center of the target object) on the MedSAM dataset
- Implement an annotation interface using clicks
  - Based on PyQt
  - Similar to the one implemented for bounding boxes
    - https://github.com/bowang-lab/MedSAM/blob/main/gui.py
- Compare MedSAM's performance with clicks and bounding boxes

Topic F: Universal click-based interactive segmentation of medical images

Resources

[1] Ma, Jun, et al. "Segment anything in medical images." Nature Communications 15.1 (2024): 654.

Topic Selection

- Find a team of three people (i.e. through the MS-Teams chat)
- Each team sends us a ranking of the presented topics until 22nd 23:59 of April per Email at zdravko.marinov@kit.edu (1 – most preferred; 6 – least preferred)
  - Example: A2, B4, C1, D3, E5, F6
- 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 (Potential process)
  - Week 0 [15.04.24]: Introduction and topic selection
  - Week 1: Read related work and present ideas on how to approach the problem
  - Week 2: Implementation
  - …
  - Week 15 [22.07.24] (Monday 14:00-16:00): Final Presentations
- Weekly meeting for discussion and status updates with corresponding supervisor
  - Set a consistent date for weekly meetings
- Register Projektpraktikum with KIT's Studienbüro (Modulhandbuch M-INFO-102966, Teilleistung T-INFO 105943)
  - **Deadline**: 29.04.2024
  - 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] and MS Teams