Computer vision for laboratory assistance tools

Ilya Agapov, Klaus Flöttmann, Francois Lemery, Frank Mayet, Bianca Veglia

4th Beam Dynamics Mini-Workshop on Machine Learning for Particle Accelerators, *Gyeongju, South Korea*

ICFA MaLAPA 2024

Computer Vision

"Vision is the act of knowing what is where by looking." - Aristotle

- The goal of computer vision is to compute geometric and semantic properties of the three-dimensional world from digital images.
- Overall, computer vision plays a crucial role in a wide range of applications. Its ability to analyse and interpret visual data enables machines to perceive and understand the world around them, leading to numerous practical benefits and advancements in technology.

Computer vision is essential in many fields and applications for several reasons

- <u>Perception and Understanding</u>: Computer vision enables machines to perceive and understand the visual world, similar to how humans do. Recognize objects, scenes, and patterns, which is crucial for tasks such as object detection, image classification, and scene understanding.
- <u>Automation</u>: This includes quality control in manufacturing, sorting and categorizing items in logistics, and monitoring processes in various industries. By automating these tasks, computer vision improves efficiency, reduces errors, and frees up human resources for more complex tasks.
- <u>Augmented Reality (AR) and Virtual Reality (VR)</u>: Fundamental to AR and VR technologies, which overlay digital information onto the real world or create immersive virtual environments.
- <u>Medical Imaging</u>: Tasks such as MRI and CT image analysis, cancer detection, and diagnostic support. Algorithms can analyse medical images to detect abnormalities, assist radiologists in interpretation, and aid in treatment planning.

Computer vision is essential in many fields and applications for several reasons (cont’d)

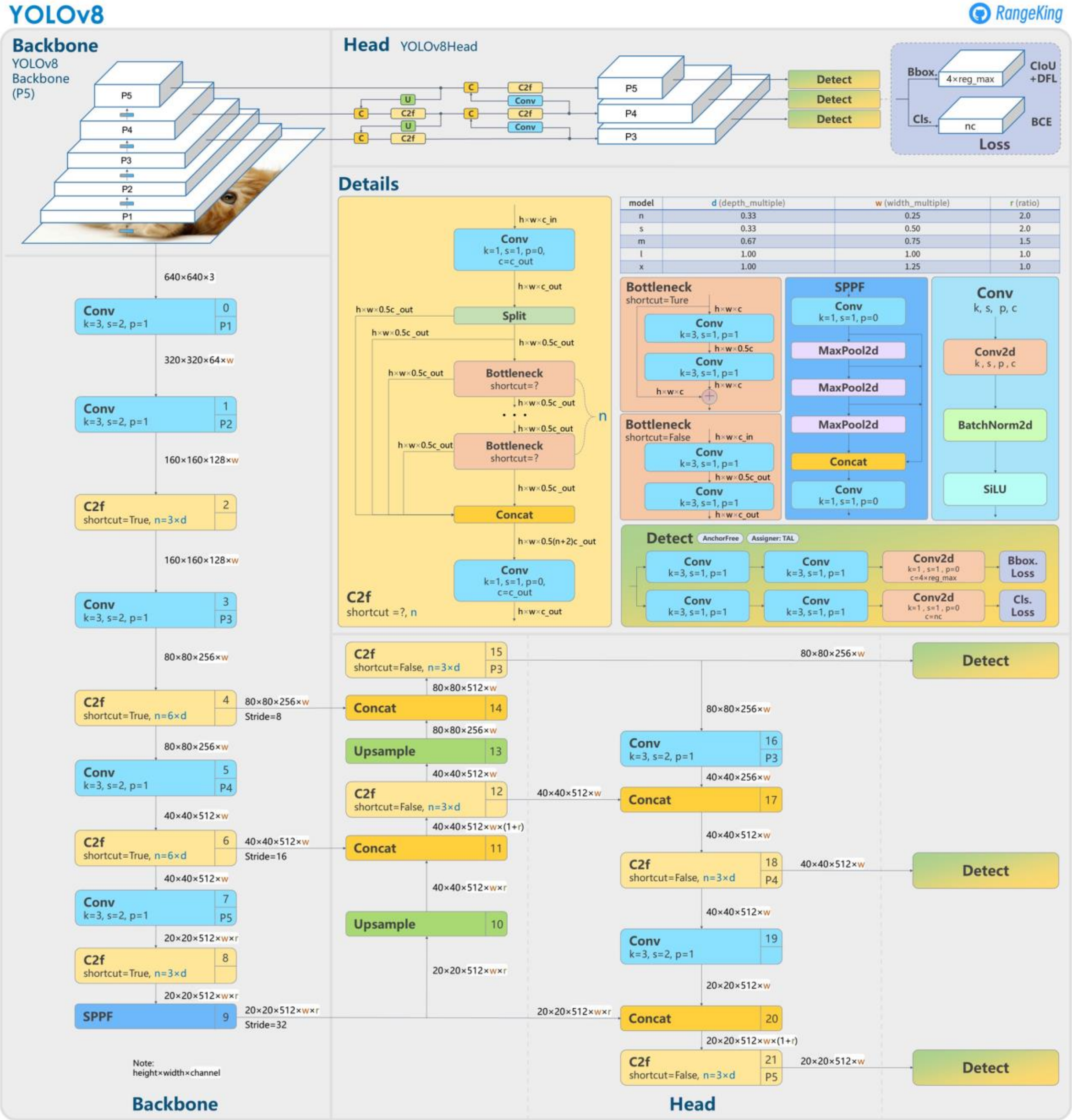
- Surveillance and Security: applications include video surveillance, facial recognition, and anomaly detection.
- Autonomous Vehicles: Computer vision is a critical component of autonomous vehicles, enabling them to perceive and interpret the surrounding environment. Computer vision systems in autonomous vehicles can detect and track objects such as pedestrians, vehicles, and traffic signs, allowing the vehicle to navigate safely and make informed driving decisions.
- Environmental Monitoring: environmental monitoring and analysis tasks such as satellite image analysis, wildlife tracking, and agriculture monitoring. By analysing visual data from various sources, it can help monitor environmental changes, assess biodiversity, and support conservation efforts.
- Human-Computer Interaction: Gesture recognition, eye tracking, and facial expression analysis are examples of applications that enhance human-computer interaction in areas such as gaming, user interfaces, and assistive technologies.

YOLO

(You Only Look Once)

YOLO is a family of computer vision models

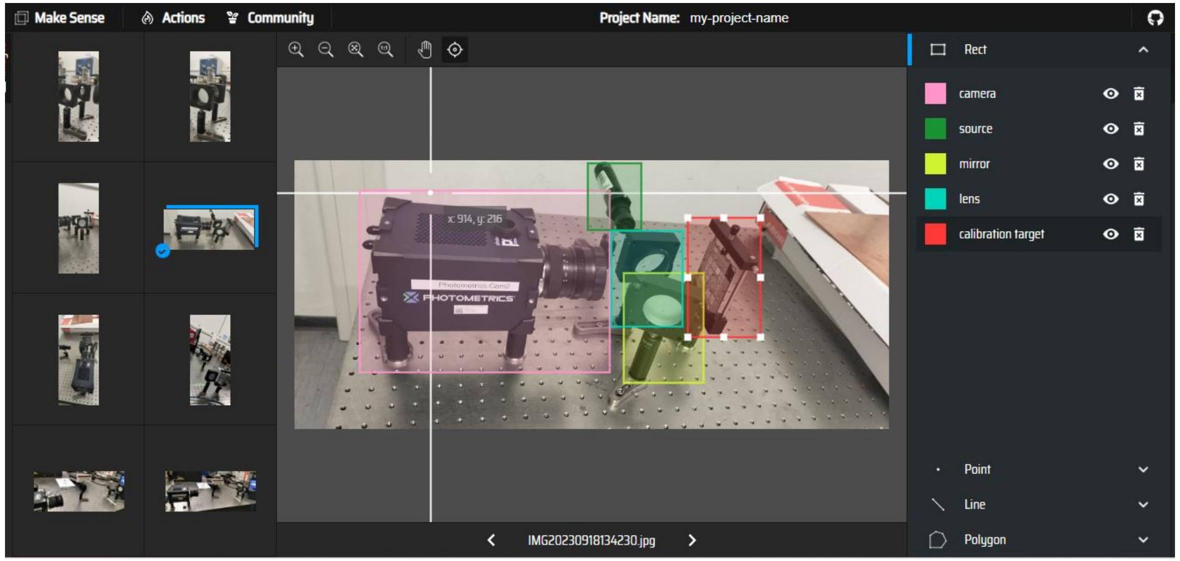
- Initially introduced (in 2016) as the first object detection model that combined bounding box prediction and object classification into a single end to end differentiable network. It was written and is maintained in a framework called Darknet. From v5 models are written in the PyTorch framework.
- In treating the detection task as a single shot regression approach for identifying bounding boxes, YOLO models are often very fast and very small – often making them faster to train and easier to deploy, especially to edge devices.
- YOLOv8 augments images during training online. One of those augmentations is called **mosaic augmentation**. This involves stitching four images together, forcing the model to learn objects in new locations, in partial occlusion, and against different surrounding pixels.



Data problem

Large amount of images properly annotated.

To train an object detector, we need to supervise its learning with bounding box annotations. If the objects and environment are uncommon, the data need to be labelled MANUALLY.

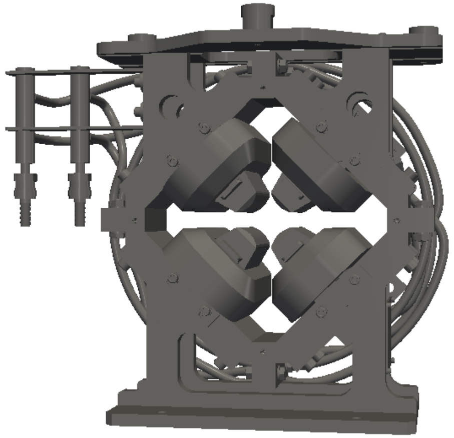


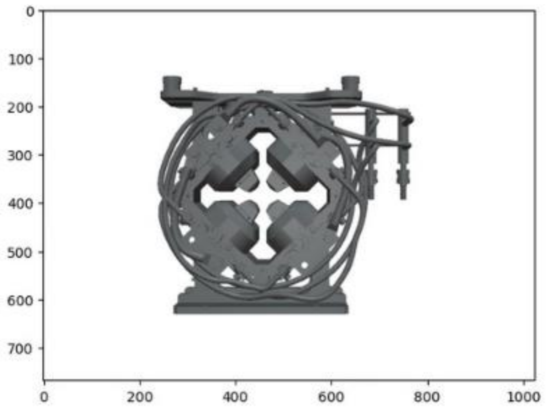
Limited Data Availability: Collecting a large and diverse dataset for object detection can be challenging, especially for custom or specialized domains. Limited data can lead to overfitting, where the model fails to generalize well to unseen data.

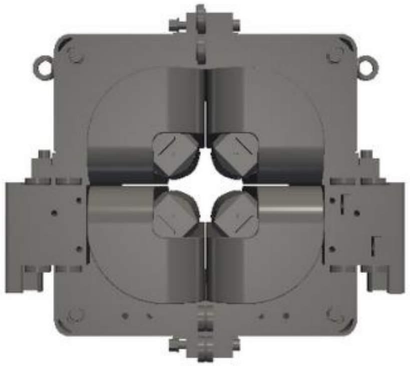
Annotation Quality: Annotating objects in images with bounding boxes or segmentation masks requires manual effort and expertise. Poorly annotated data can negatively impact model training and performance.

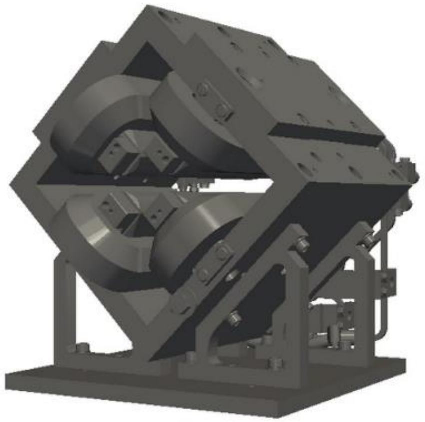
CAD images

- CAD models can be used to produce large amount of images (same object different view angle) and allow for automated labelling.
- When model trained in combination with real-life images it improved its prediction accuracy.

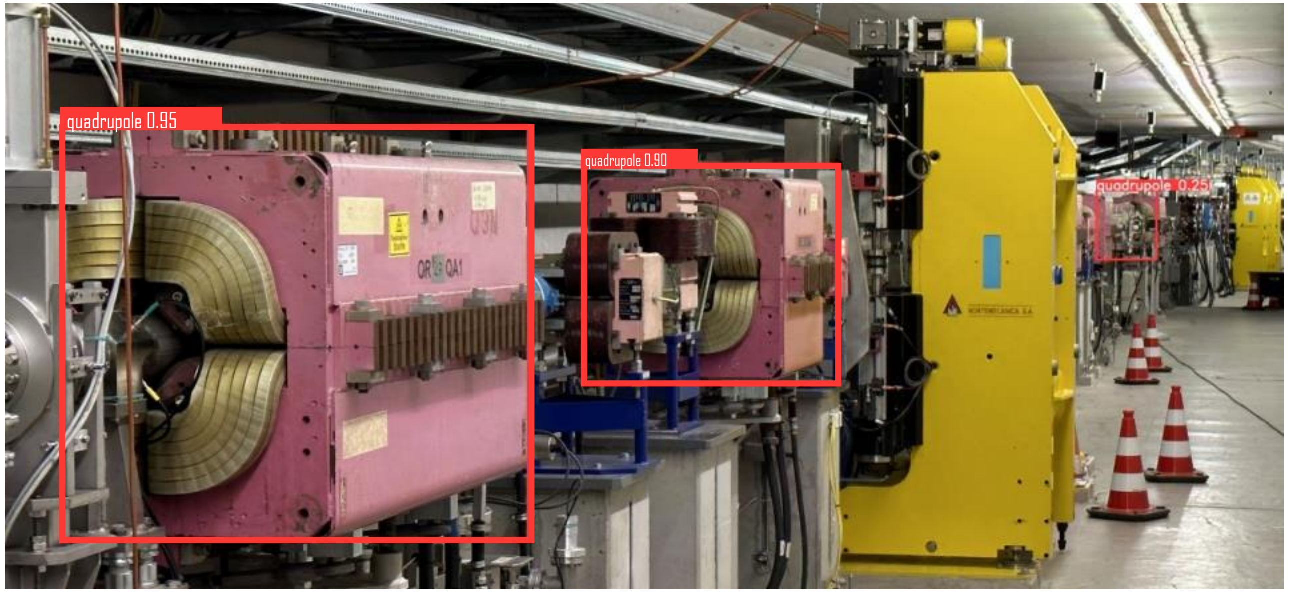




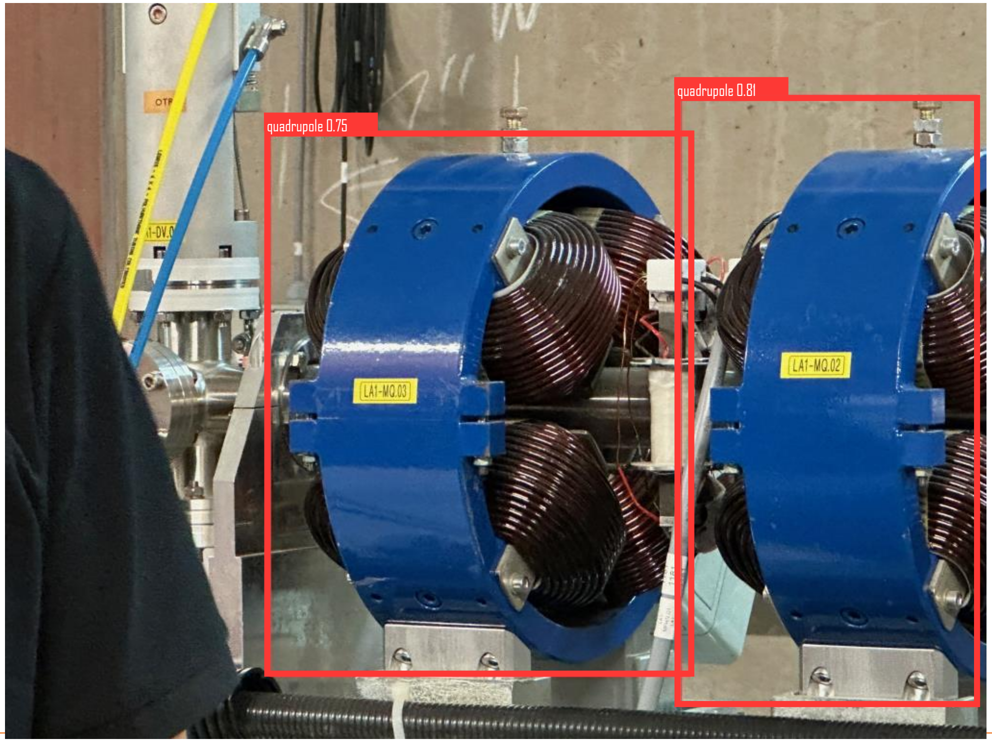


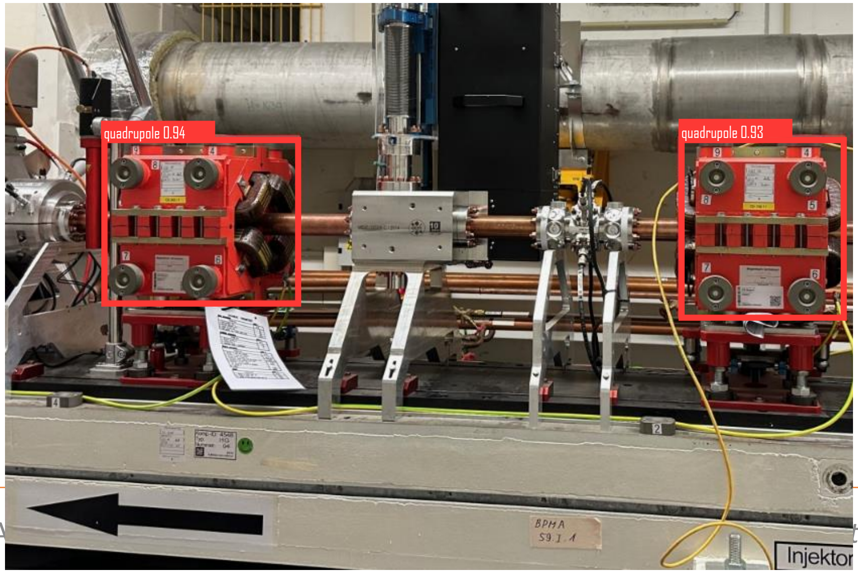


Detect

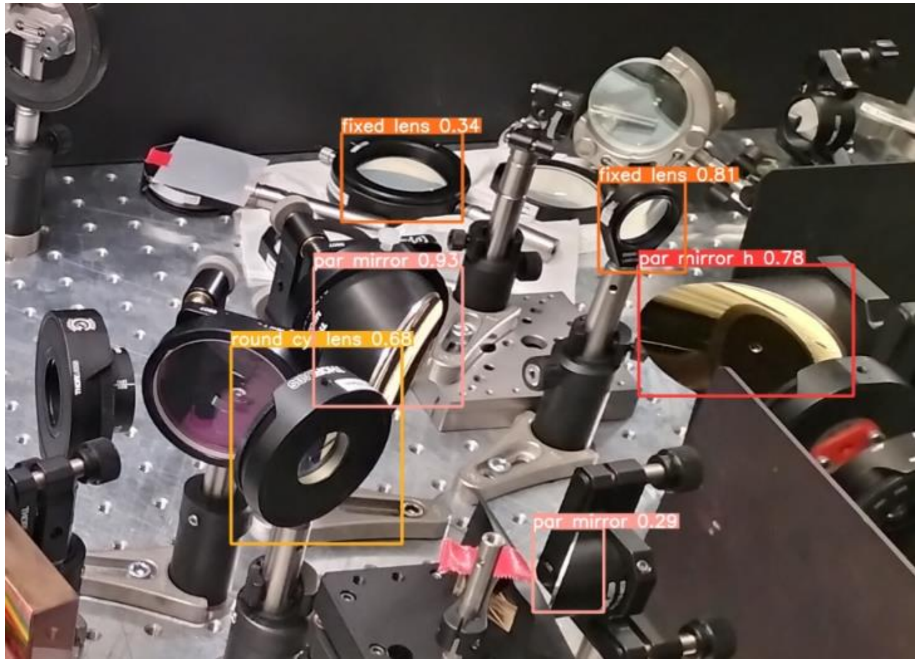


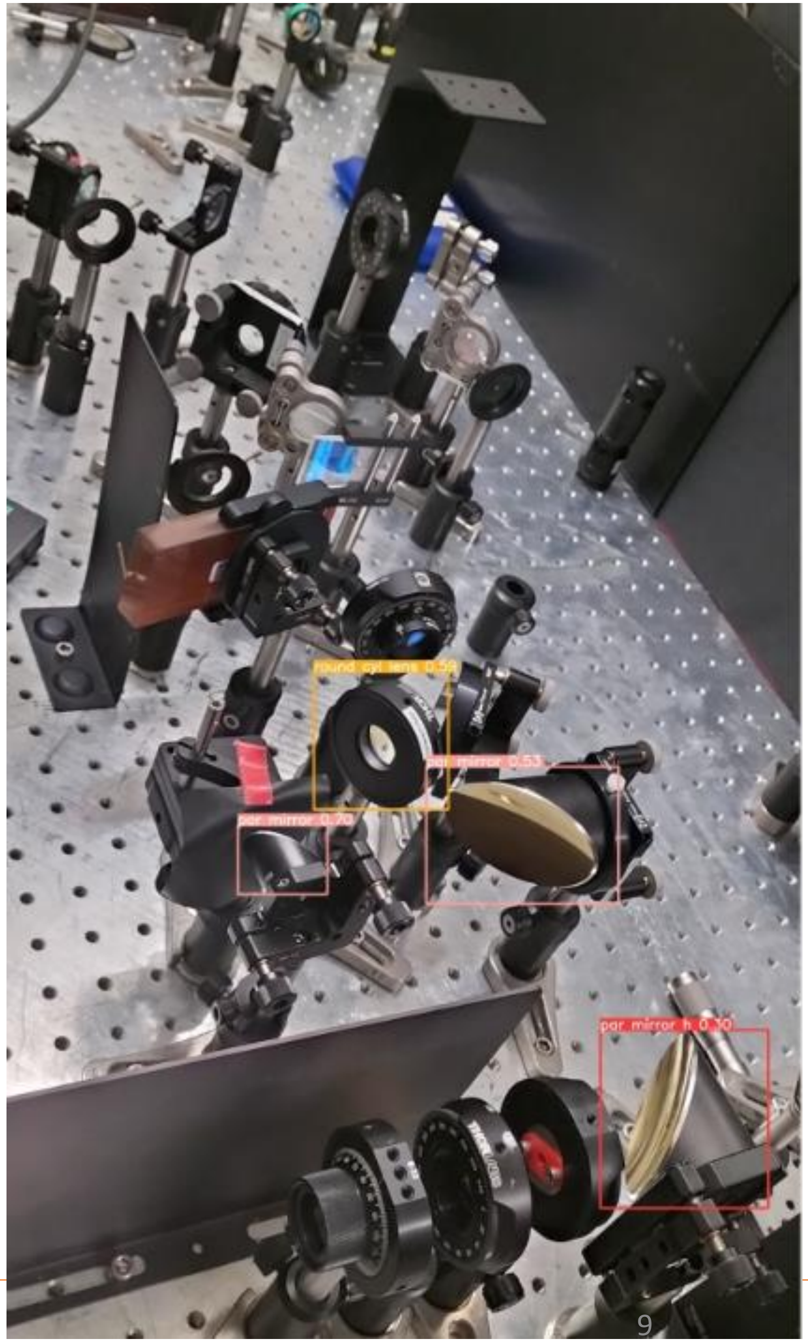
200 views of the different CAD models + 90 real images: VERY GOOD PREDICTIONS!



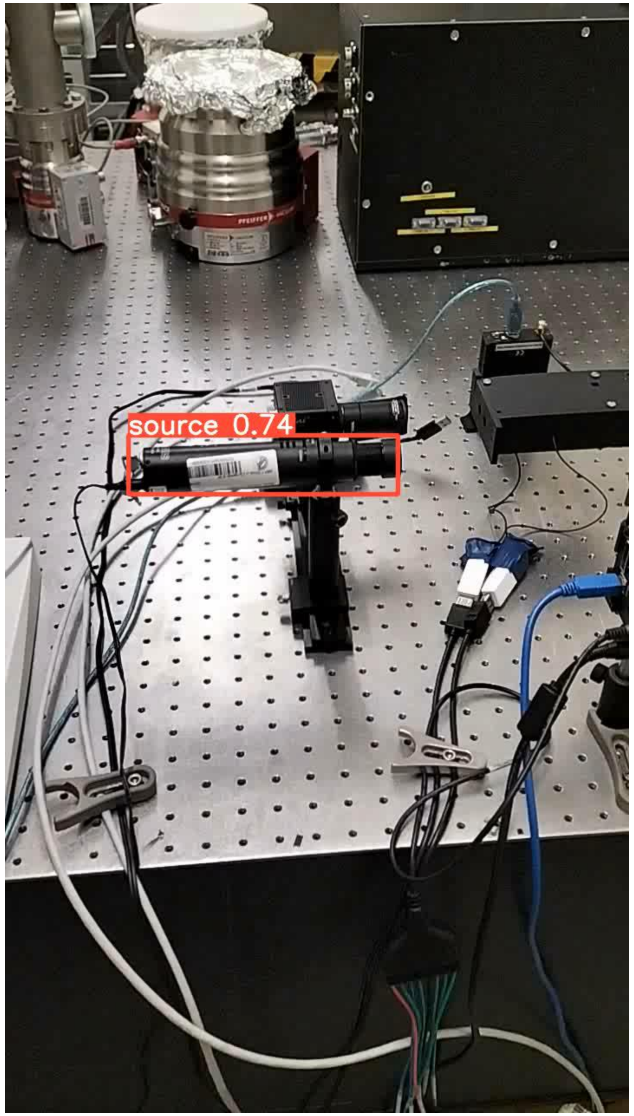


REGAE laser optics (4 classes, 144 CAD images each, only 5 photos)…ok but not great.

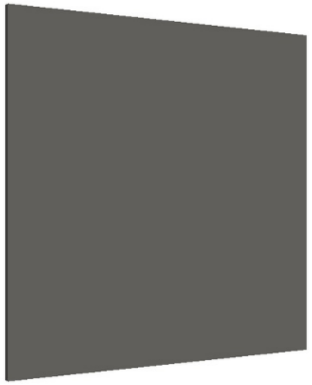




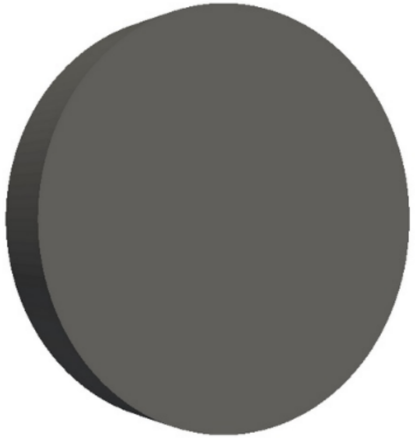
Identify



Some objects cannot be univocally recognized.
For example optics elements are hard to tell apart and the same type of mounting could be used for different objects (CAD models won't help here).



Calibration target model (?!?)

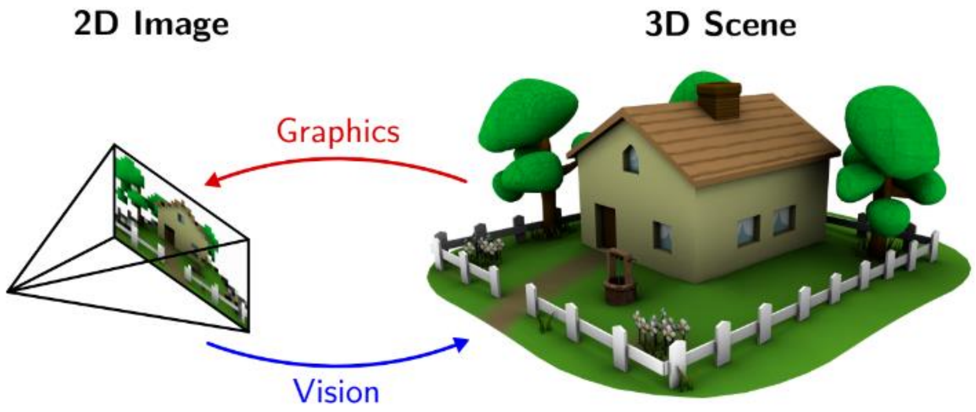


Mirror or lens?

Space

To develop a tool assisting in various lab tasks not only object detection but also spatial relations between objects.

- Measuring the distance of each pixel relative to the camera. Depth is extracted from either monocular (single) or stereo (multiple views of a scene) images.
- Traditional methods use multi-view geometry to find the relationship between the images.
- Newer methods can directly estimate depth by minimizing the regression loss, or by learning to generate a novel view from a sequence.

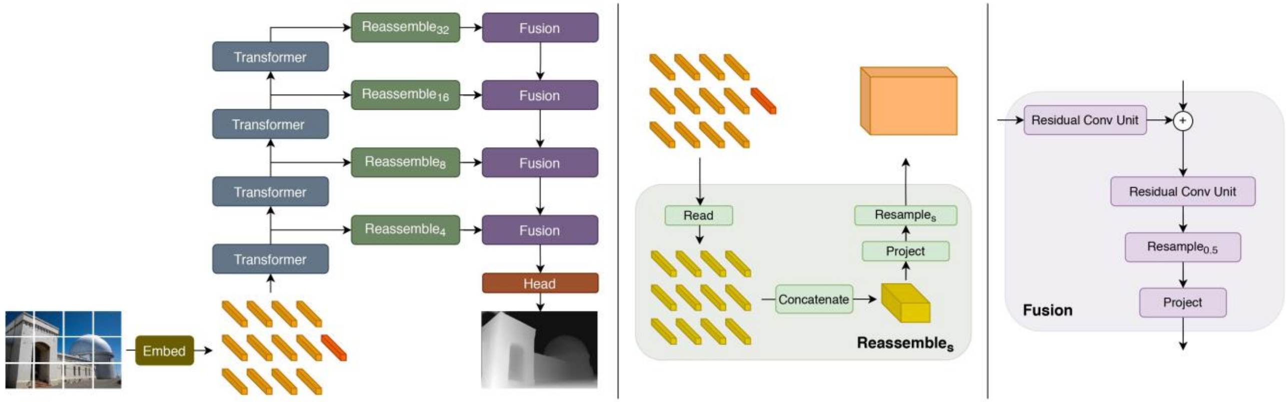


Computer Vision is an ill-posed inverse problem:

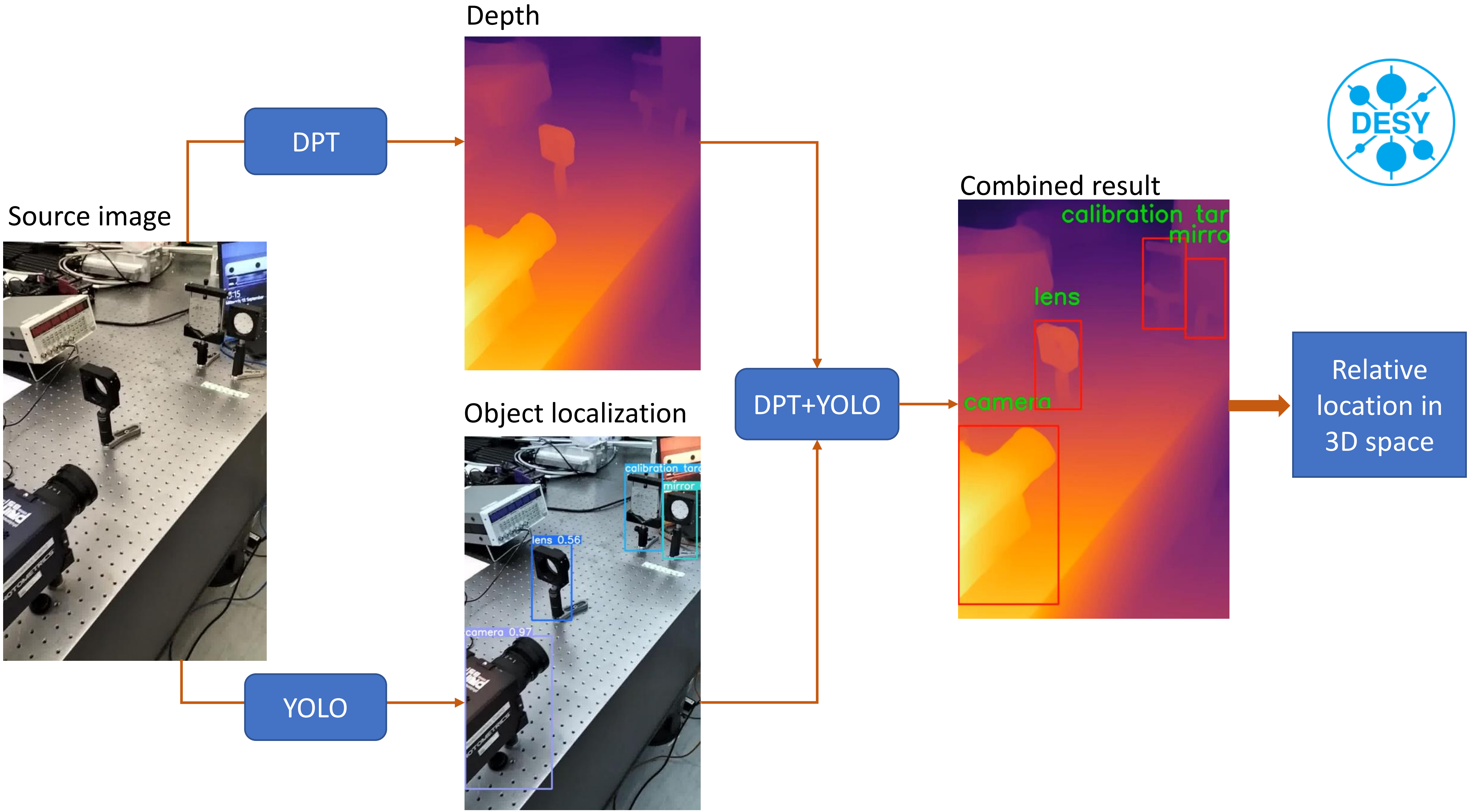
- ► Many 3D scenes yield the same 2D image
- ► Additional constraints (knowledge about world) required

Depth Estimation

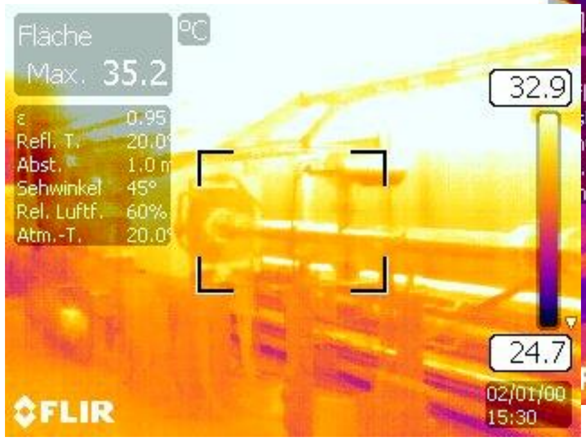
- Using DPT *(Dense Prediction Transformers)*, a segmentation model released by Intel in March 2021 that applies *vision transformers* (ViT) to images.

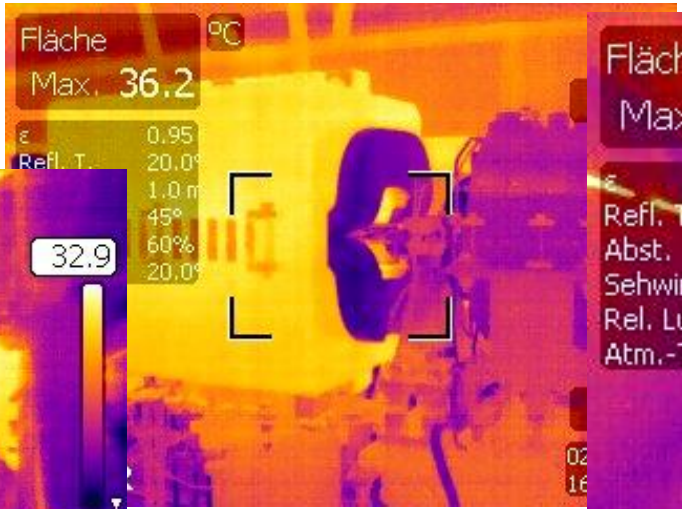


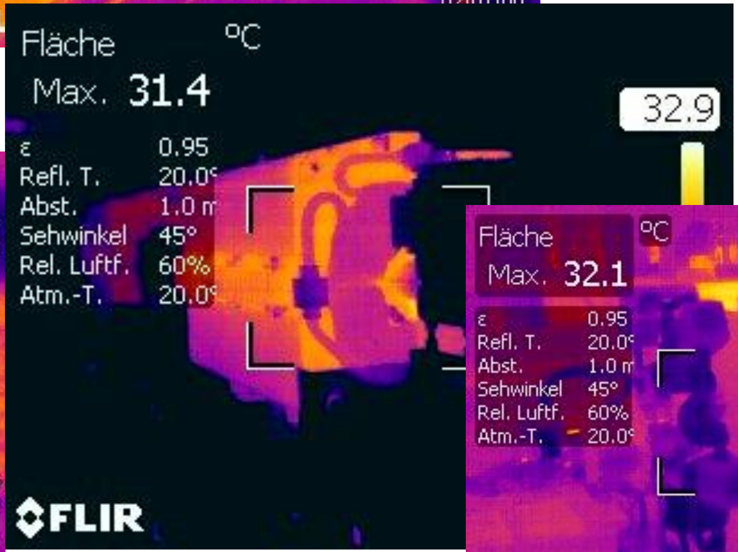
- DPT models are trained on 1.4 million images for monocular depth estimation. DPT uses the ViT as backbone and adds a neck + head on top for monocular depth estimation.


Depth
DPT
Source image
YOLO
Object localization
lens 0.56
camera 0.97
mirror
DPT+YOLO
Combined result
calibration tar
mirro
lens
camera
Relative location in 3D space

Thermal imaging



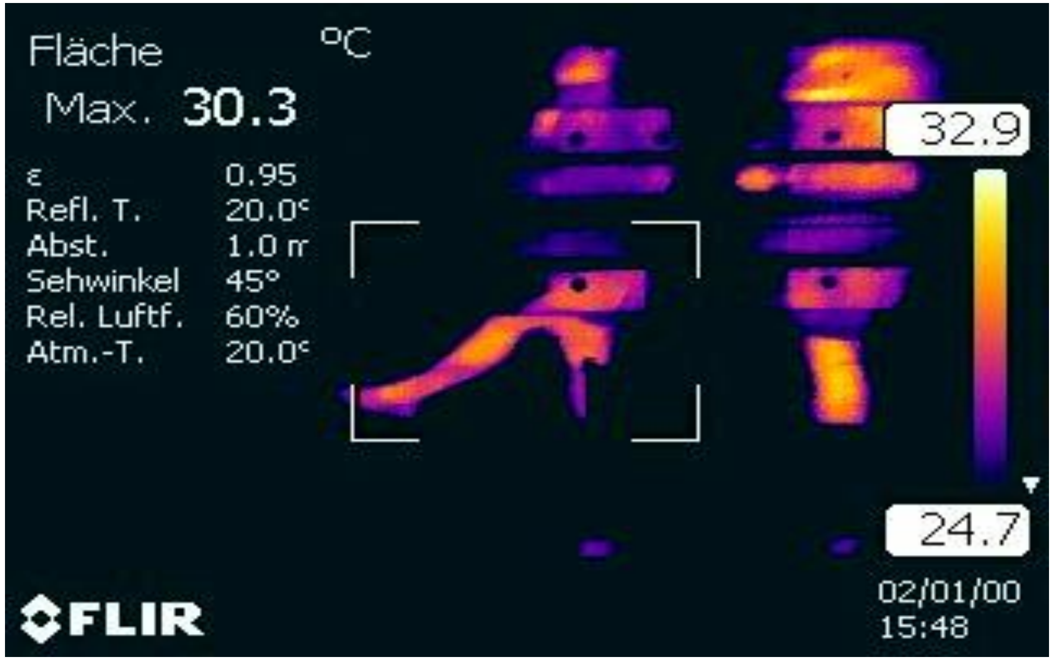




Summary and outlook

- First approach with CV for lab assistance tools.
- Object detection requires large training datasets. For unequivocal identification physically labelling the objects might be the best option. Plus it allows to access additional information.
- Reconstructing 3D space from bi-dimensional images seems possible with DPT (pre-trained).
- Extracting the relevant information from the depth predictions to facilitate lab tasks.
- Fine tuning of the model to obtain reliable 3D mapping for applications.

Thank you!



Contact

DESY. Deutsches Elektronen-Synchrotron

www.desy.de

Bianca Veglia
MPY
bianca.veglia@desy.de