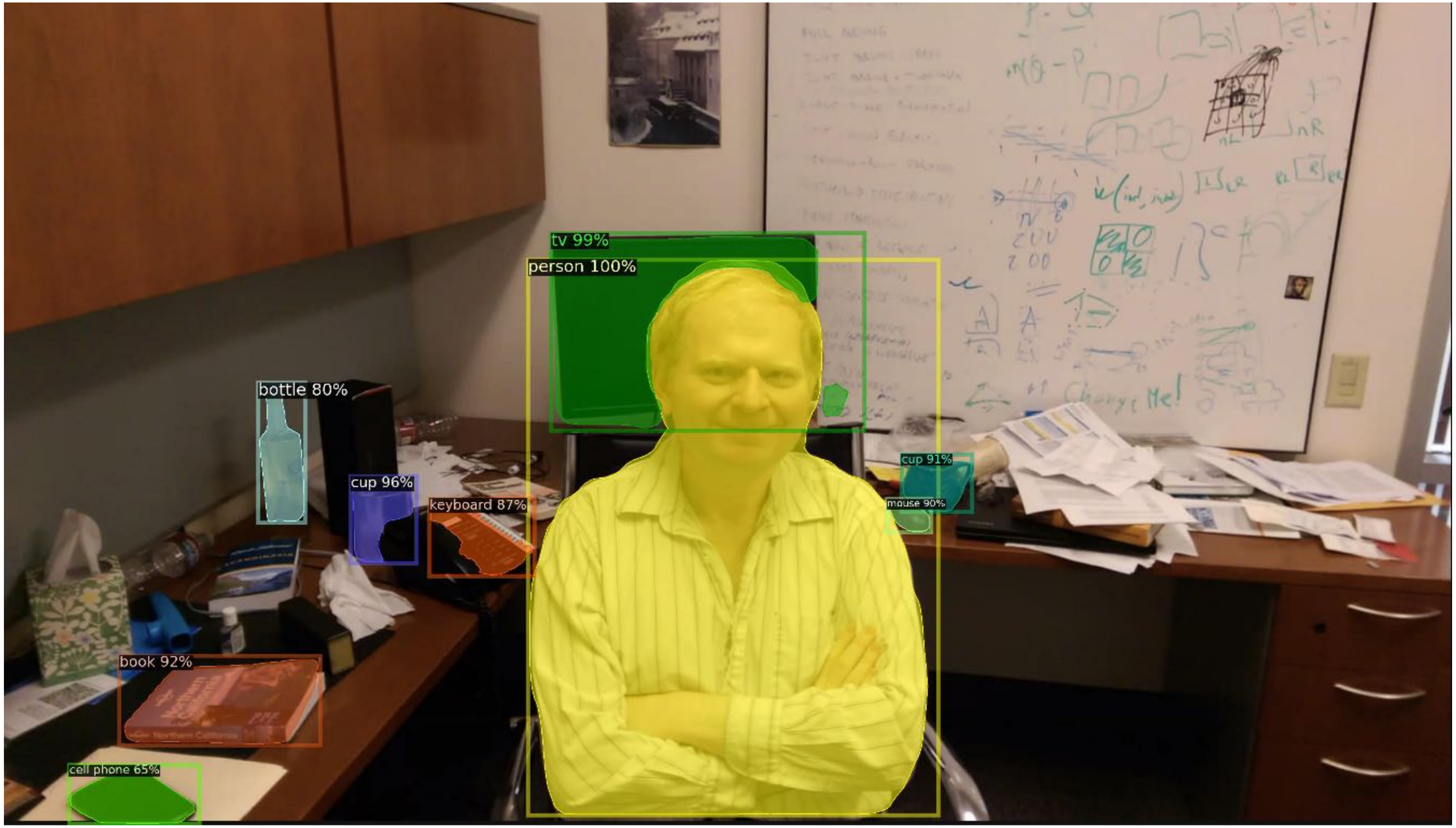
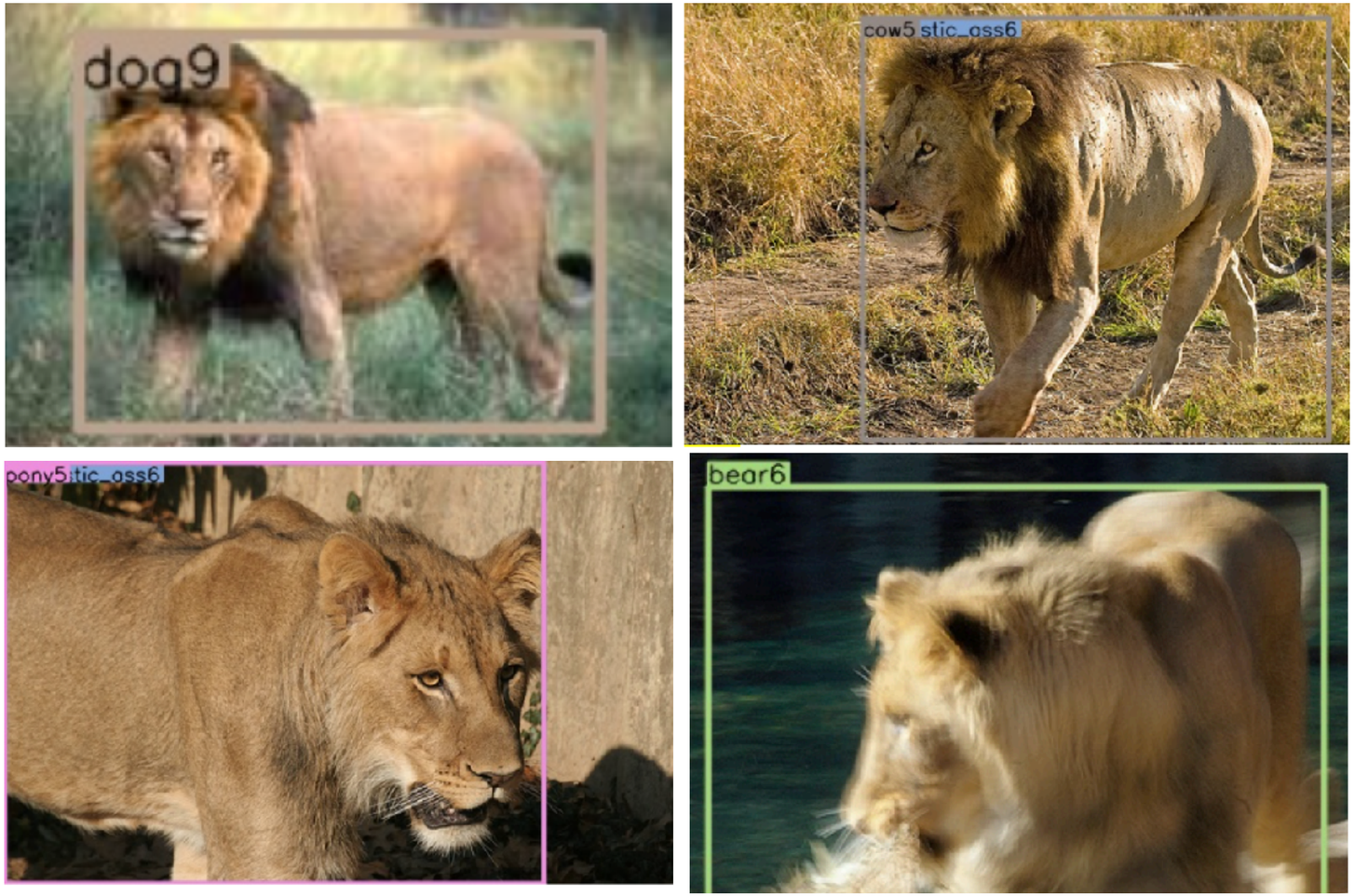
Open-Vocabulary Recognition

Philipp Krähenbühl, UT Austin

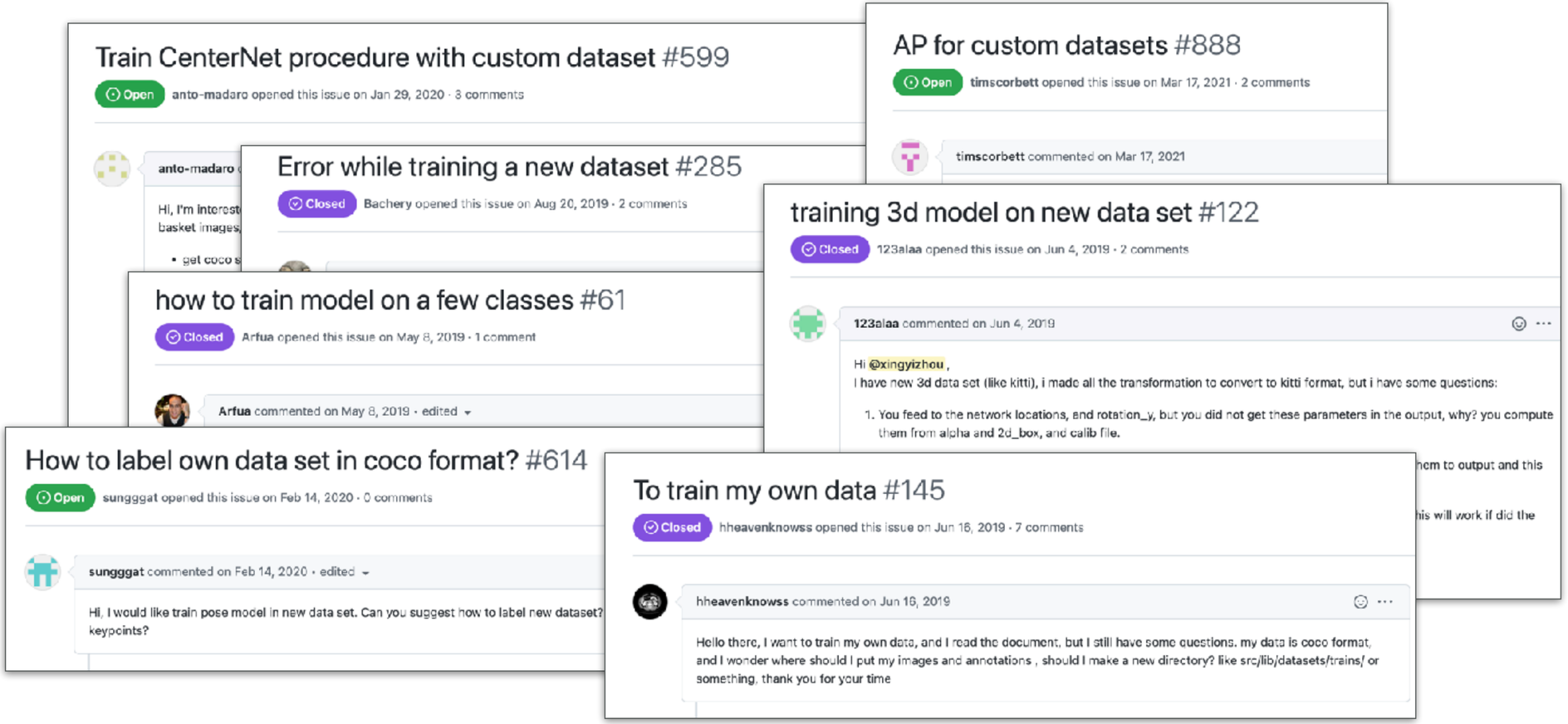
How well do detectors work?



How well do detectors work?



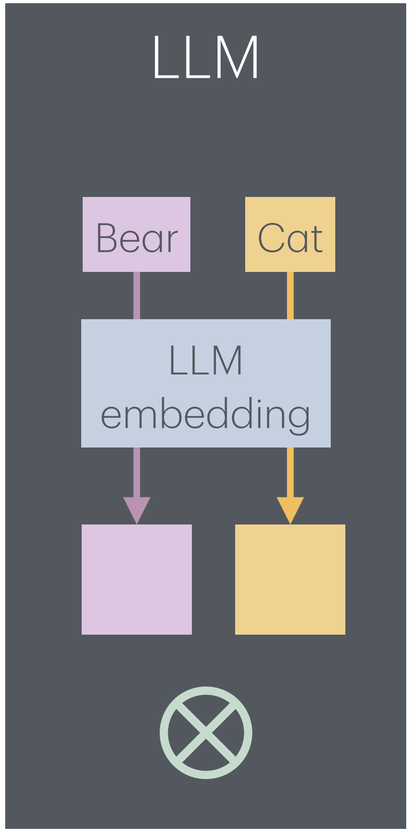
How well do detectors work?



CLIP

Classifier

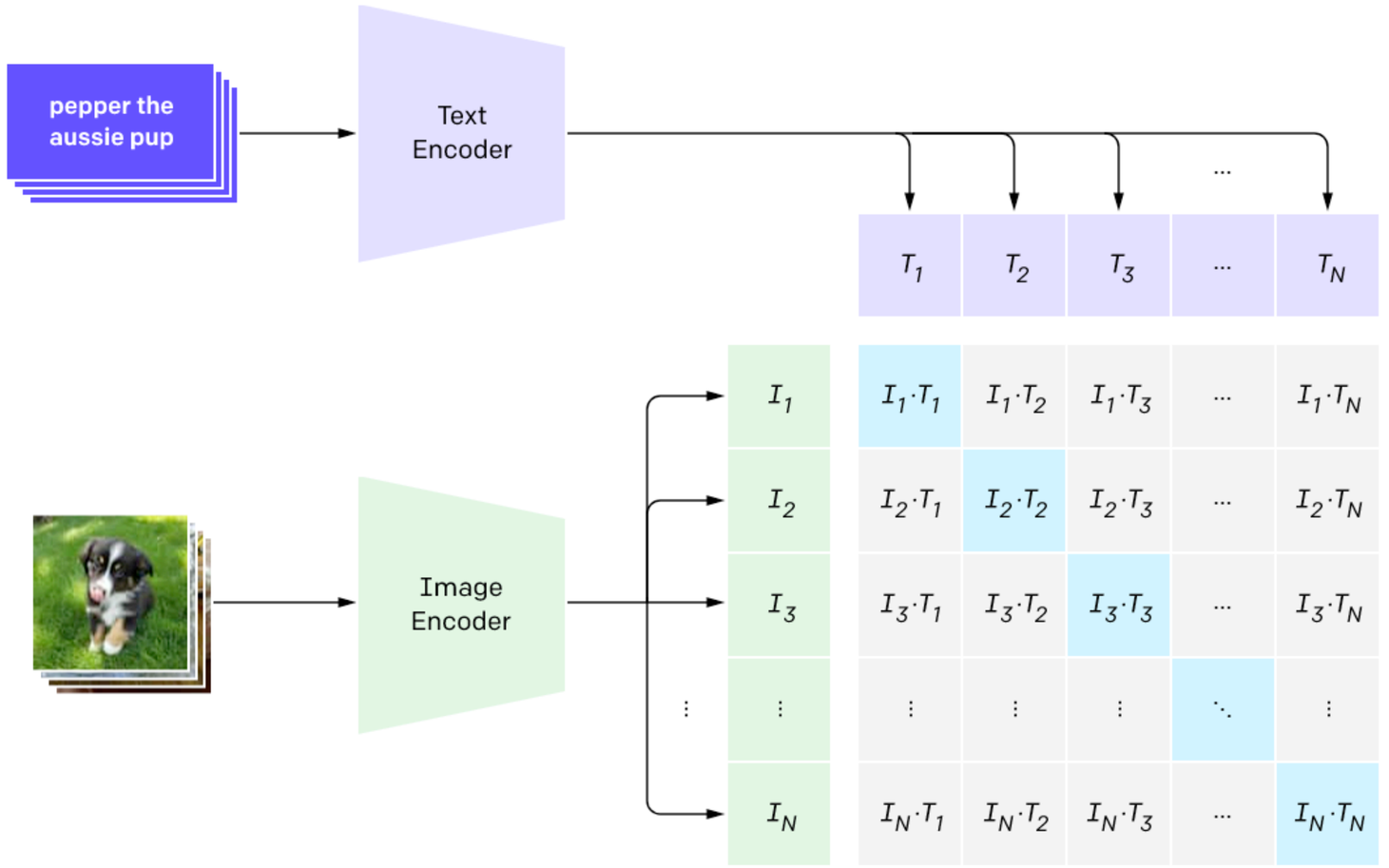
feature f



class
probabilities

Learning Transferable Visual Models From Natural Language Supervision, Radford et al. 2021

CLIP



Learning Transferable Visual Models From Natural Language Supervision, Radford et al. 2021

Image credit: OpenAI

CLIP - Training

Training dataset : The Internet

internet images

alt-text / captions

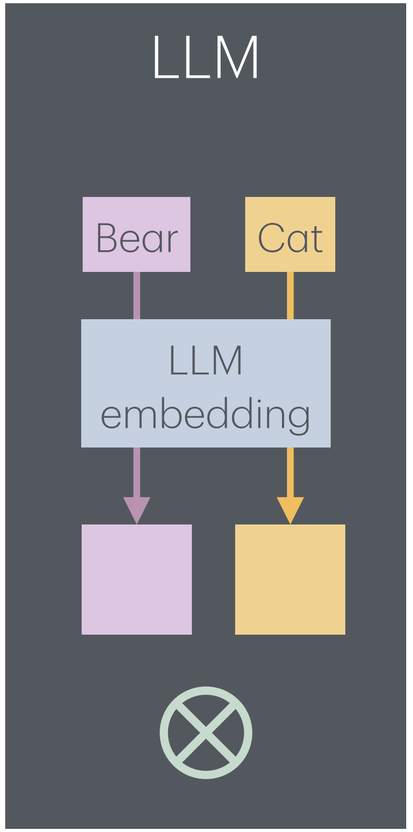
Learning Transferable Visual Models From Natural Language Supervision, Radford et al. 2021

Image credit: OpenAI

CLIP

Classifier

feature f



class
probabilities

Learning Transferable Visual Models From Natural Language Supervision, Radford et al. 2021

LLMs + Detection

Detector Stage 1

Box, score

Detector Stage 2

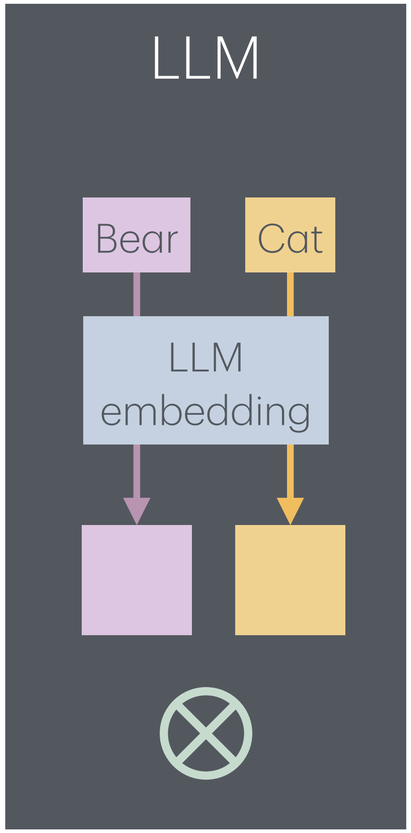
f_1

f_2

f_3

...

Refined box, feature, (optional) mask, (optional) tracking,



per box class probabilities

Open-vocabulary Object Detection via Vision and Language Knowledge Distillation, Gu et al. 2023

LLMs + Detection

Training data

LLM
Detector

COCO
Common Objects in Context

LVIS

~1000 classes
with enough
supervision

Open-vocabulary Object Detection via Vision and Language Knowledge Distillation, Gu et al. 2023

DETIC

LLM
Detector

COCO
Common Objects in Context

LVIS

IMAGENET

The Internet

Detecting Twenty-thousand Classes using Image-level Supervision, Zhou et al, 2022

DETIC

Detector Stage 1

Box, score

Detector Stage 2

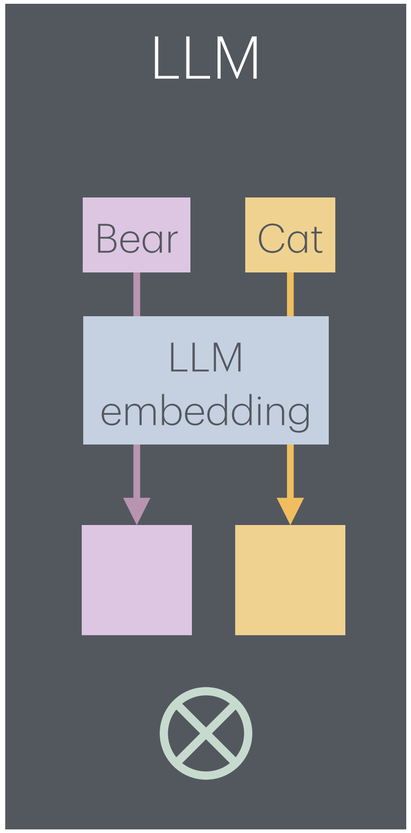
f_1

f_2

f_3

...

Refined box, feature,
(optional) mask,
(optional) tracking,



per box class probabilities

Detecting Twenty-thousand Classes using Image-level Supervision, Zhou et al, 2022

DETIC

LLM
Detector

IMAGENET

The Internet

Detecting Twenty-thousand Classes using Image-level Supervision, Zhou et al, 2022

DETIC - Results

Guglielmo Iozzia
@Guglielmolozzia

#Detic from #Facebook Research: source code for the "Detecting twenty-thousand classes using image-level supervision" paper. It can run on different vocabularies, including custom ones. #DeepLearning #ComputerVision #Python #PyTorch

12:15 PM · Jan 12, 2022 · Twitter Web App

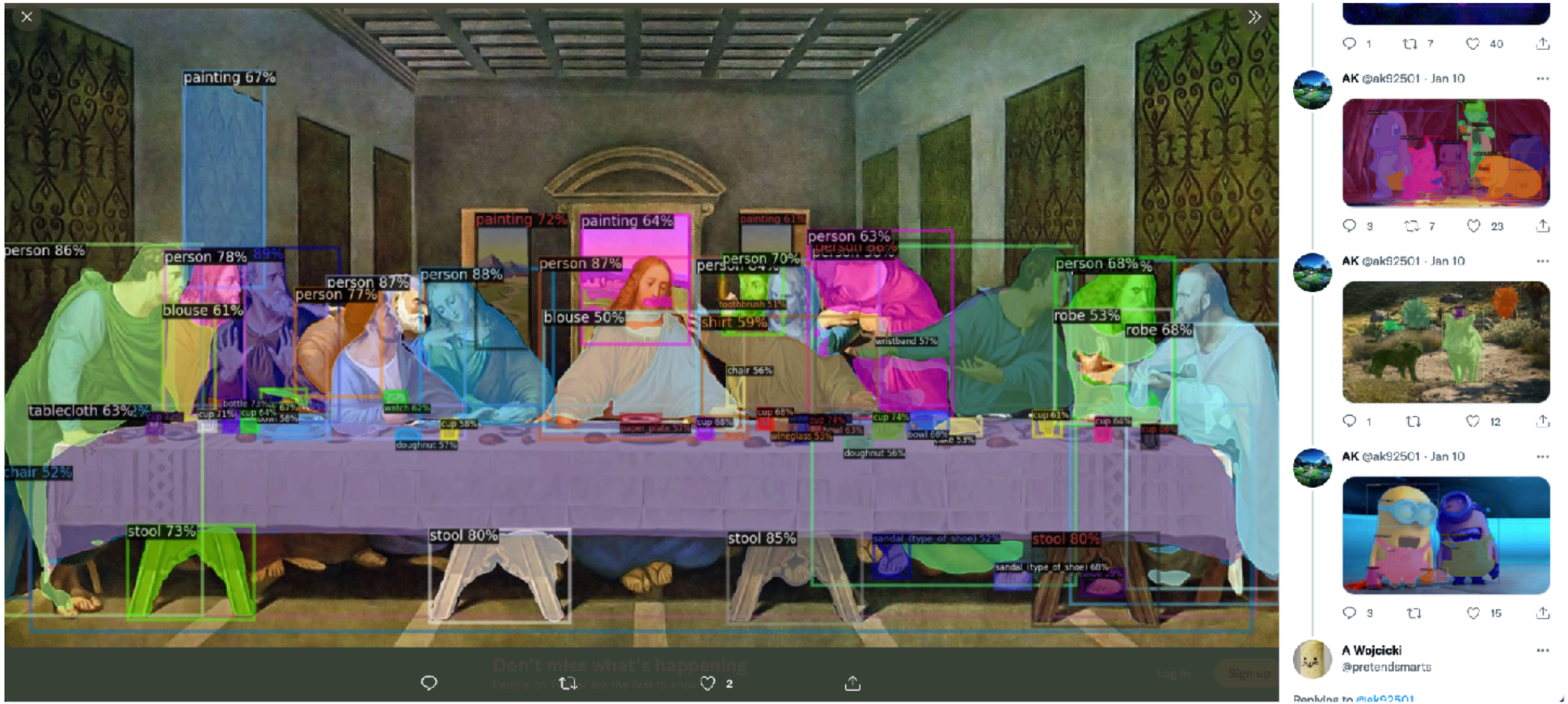
1 Retweet

Guglielmo Iozzia @Guglie... · 4h
Replying to @Guglielmolozzia
Official #GitHub repository:
github.com/facebookresear...
#Python #PyTorch #opensource
#DeepLearning #ComputerVision

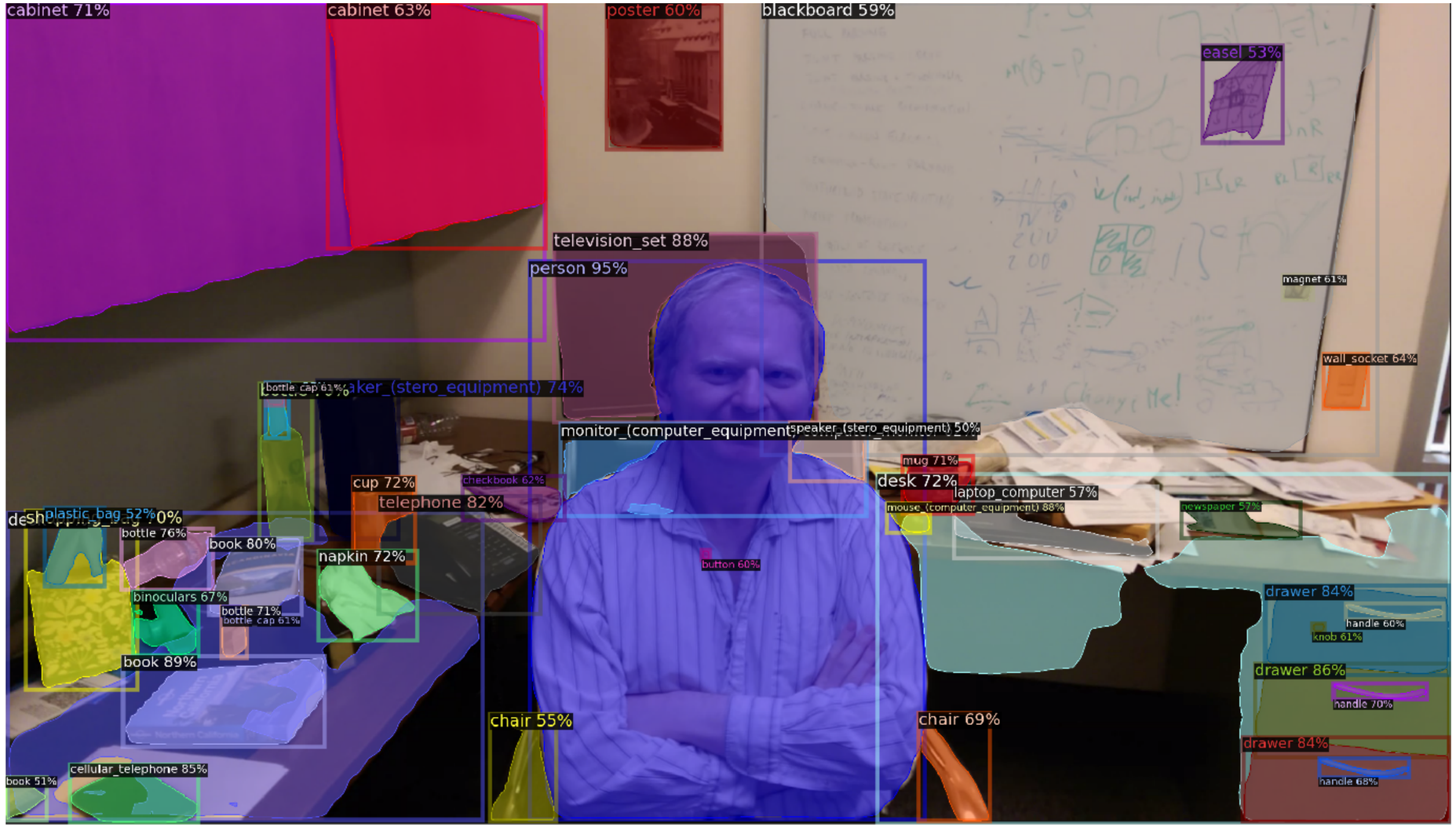
facebookresearch/
Detic

github.com
GitHub - facebookresearch/Detic:
Code release for "Detecting ...

DETIC - Results



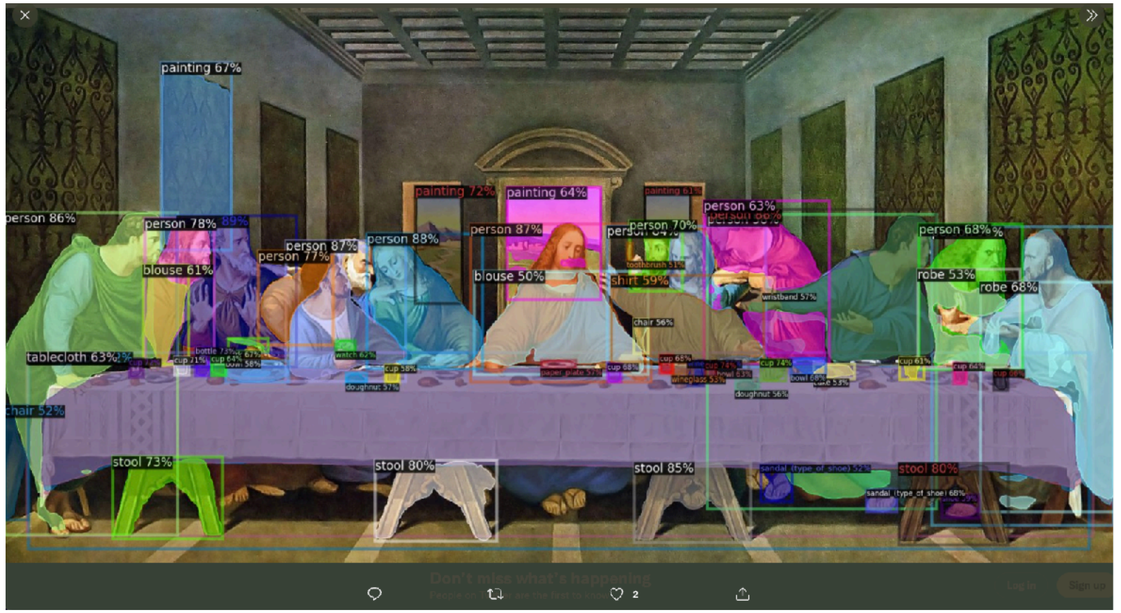
DETIC - Results



Open-Vocabulary Recognition

Trend 1

- **Language embeddings become norm in detection**
  - Much of the detection literature moved to open-vocabulary or large-vocabulary detection



Open-Vocabulary Recognition

Trend 2

- **Zero-shot evaluation become norm**
  - Classes are slowly replaced by word or sentence embeddings
  - Training and test vocabulary no longer needs to align

Open-Vocabulary Recognition

Trend 3

- **Recognition slowly replaced by Vision Language Models**
  - Recognition dataset provide great supervision for training
  - Too rigid for deployment

References

- [1] Learning Transferable Visual Models From Natural Language Supervision, Radford et al. 2021
- [2] Open-vocabulary Object Detection via Vision and Language Knowledge Distillation, Gu et al. 2023
- [3] Detecting Twenty-thousand Classes using Image-level Supervision, Zhou et al, 2022