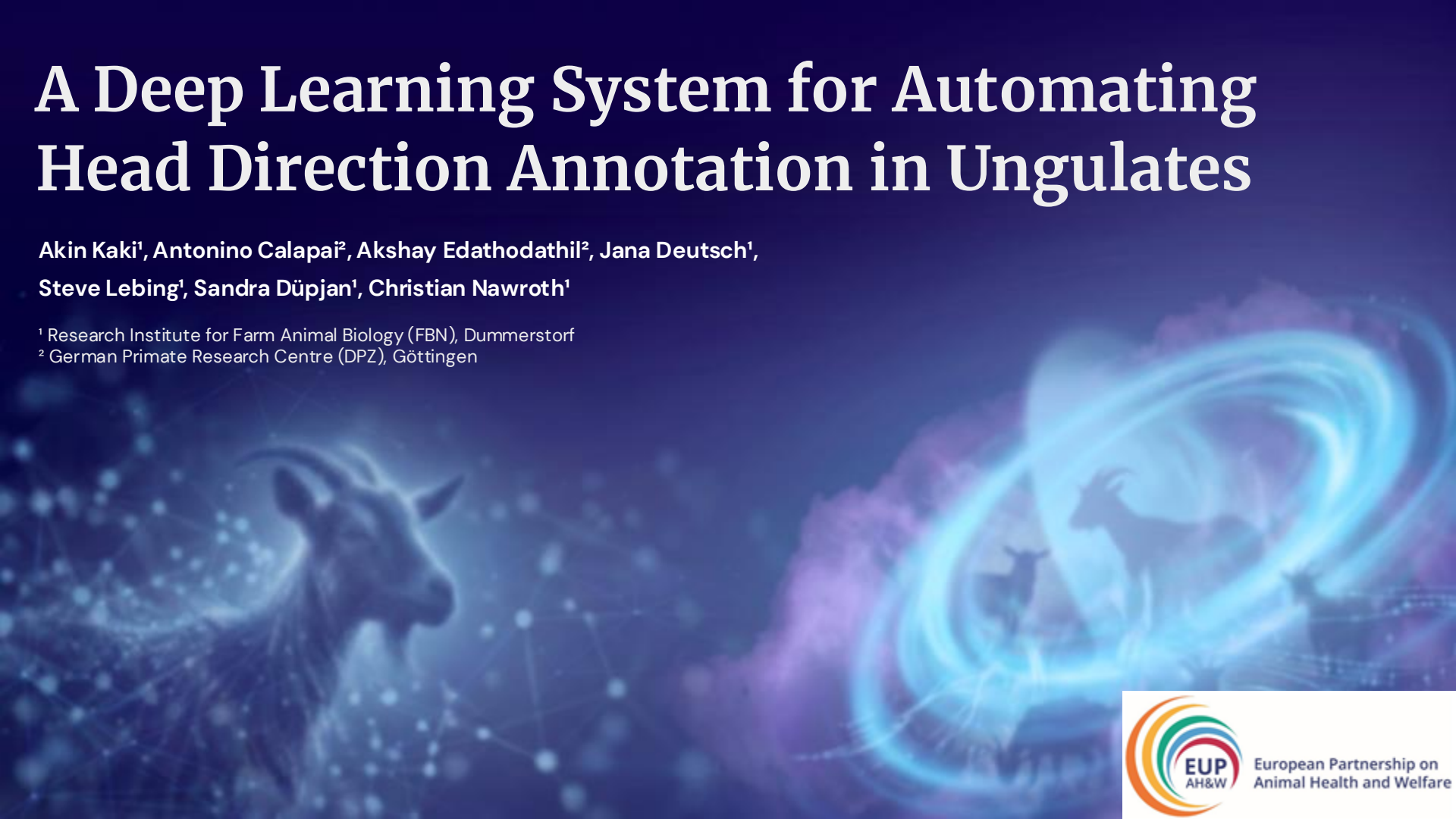
A Deep Learning System for Automating Head Direction Annotation in Ungulates

Akin Kaki[1], Antonino Calapai[2], Akshay Edathodathil[2], Jana Deutsch[1], Steve Lebing[1], Sandra Düpjan[1], Christian Nawroth[1]

[1] Research Institute for Farm Animal Biology (FBN), Dummerstorf
[2] German Primate Research Centre (DPZ), Göttingen



EUP AH&W European Partnership on Animal Health and Welfare

Introduction

Ethological approaches can help to assess how farm animals perceive and engage with their environment

Understanding underlying needs, motivations, and behavioral choices can help to ultimately improve welfare

Often implemented decision-making tasks require time-consuming training and arbitrary behavioural responses

But: Reactive paradigms, such as the **Looking Time Paradigm,** exploit an animal's natural response to a stimulus to gain insights into the cognitive processes at play, with no prior training required

Introduction - Looking Time Paradigm

Core Mechanism - exploits an animal's natural tendency to direct its gaze toward novel or relevant visual information.

- Stimuli can be systematically controlled
- Head direction and looking duration can be assessed
- Allows inferences about the underlying cognitive processes at play

Important
Continuous video documentation provides the primary dataset for experimental analysis.

Problem Statement -The Bottleneck

Assessing data from looking time paradigms currently comes with significant practical challenges, mainly the need for laborious manual coding of data.

The Solution: Deep Learning allows for high-throughput, objective behavior recognition.

- **Free-Moving Subjects:** Animals move unconstrained, leading to varying distances and head positions
- **Technical Obstacles:** Motion blur, low contrast, and lateral eye positioning make traditional gaze tools (e.g., DeepLabCut) struggle.

Data Acquisition & Annotation

Data was sourced from Deutsch et al. 2024, which focused on behavioural response of goats towards various visual stimuli in a looking time paradigm.

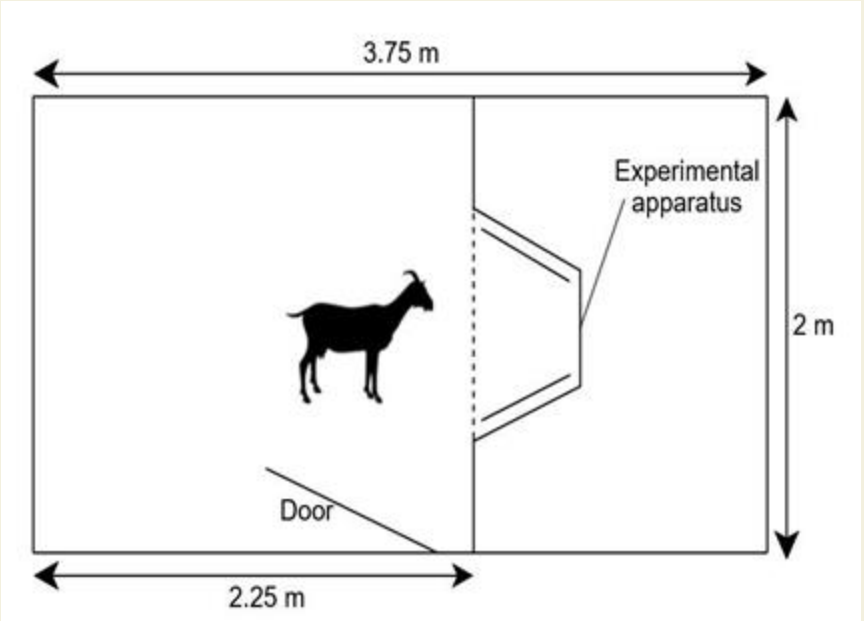
Focal Subject: Goat in a dual-screen testing arena (2.25m × 2m)

Input: Video at 1280 × 720p, 30 FPS

Dataset: 3,500 frames across 6 different individuals with varying head positions were annotated for training.

Classes Annotated:

- **Eyes:** Each visible eye.
- **Nose:** Visible nose area.
- **Horns:** Treated as a single entity (critical for the fallback mechanism)



Model Architectures

We evaluated three architectures to find the best feature extractor.

Baseline 1

Faster R-CNN (ResNet-101)
Standard CNN backbone using FPN and RPN hierarchies.

Baseline 2

YOLOv8-Medium
Selected for efficiency and high-speed detection benchmarks.

Experimental

Faster R-CNN (ViT)
A Vision Transformer backbone leveraging global self-attention.

Why ViT?

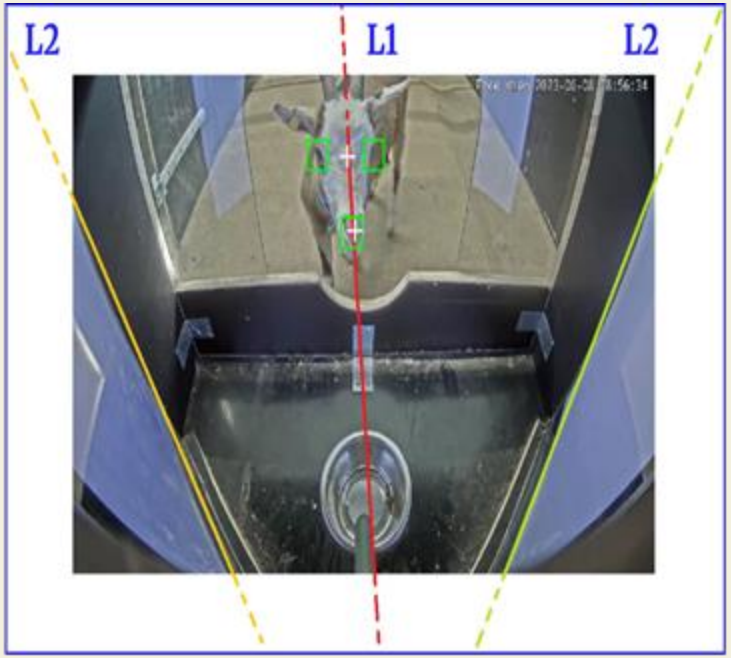
- **Global Attention:** Relates image patches across the entire frame simultaneously.
- **Occlusion Recovery:** Infers hidden features (e.g., a nose) from surrounding context.
- **Structural Insight:** Superior at geometric reasoning for atypical head postures.

Geometric Estimation Pipeline

- Head direction modelled as a 1D vector (L1)
- The screens are modelled as a 1D vector joining the opposite corners (diagonal) (L2)

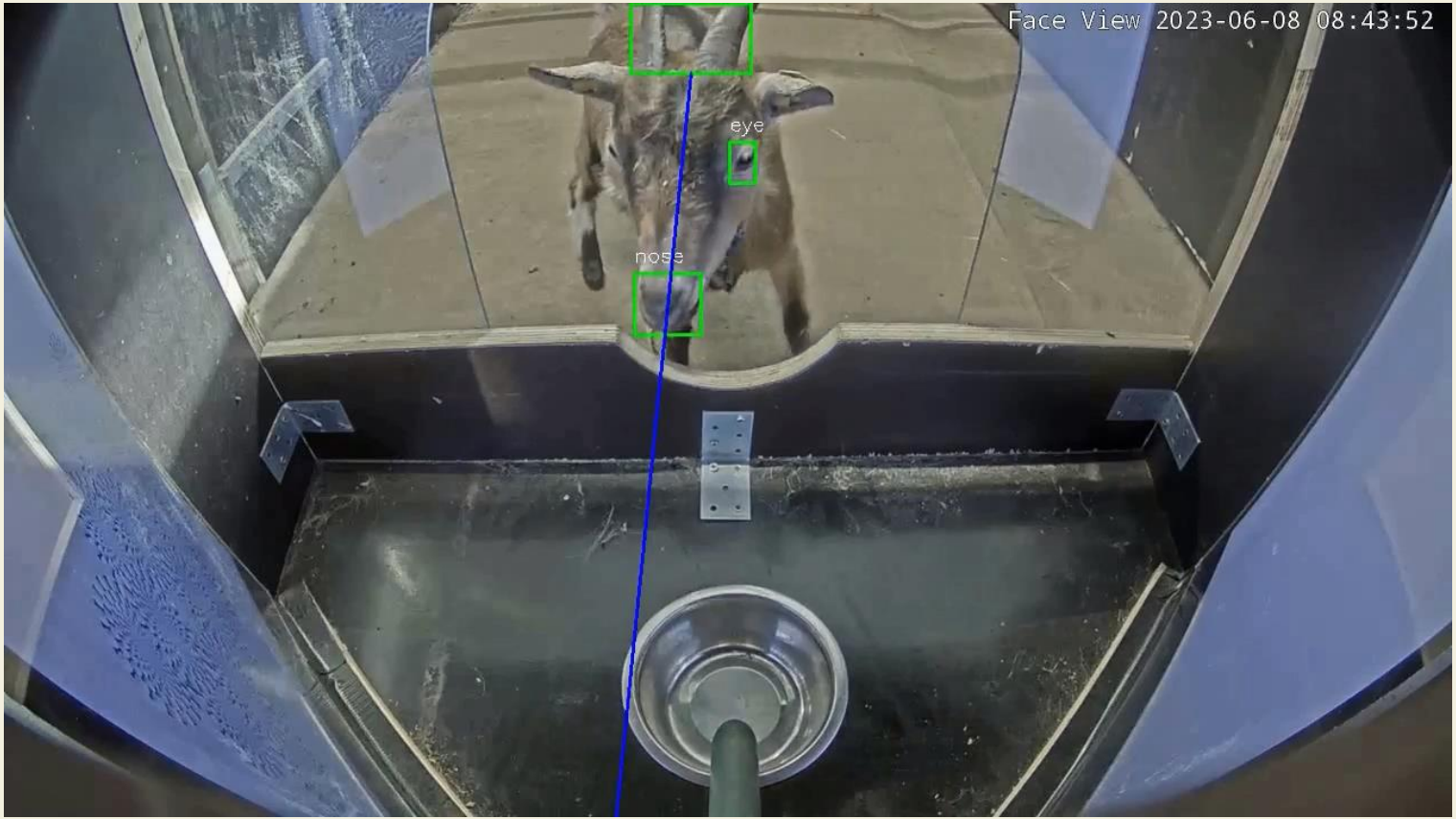
Decision Logic:
A "hit" is only registered if the calculated intersection point falls within the screen's actual bounds (x_{min},y_{min},x_{max},y_{max});

Condition	Outcome
Intersecting “Left” screen	Left
Intersecting “Right” screen	Right
No Intersection/ Default	No Screen



Similar estimation pipeline has been used for manual coding.

Results



Detection Results: Feature Learning

Model performance scores (from 20% validation set).

- **DIoU:** Measures box overlap while ensuring their center points are as close as possible.
- **mAP50:** A lenient accuracy score requiring only a 50% overlap to count as a correct detection.
- **mAP95:** A strict accuracy score that demands near-perfect, tight bounding boxes to score high.

Feature	Model	DIoU	mAP50	mAP95
Nose (Most stable anchor)	Resnet-101(Baseline 1)	0.78	0.89	0.45
	YOLOV8m (Baseline 2)	0.80	0.92	0.48
	ViT (Experimental)	**0.86**	**0.96**	**0.54**

The Faster R-CNN model with ViT as backbone outperformed other models in all the metrics.

Validation Framework

We did not just evaluate the general metrics; we also evaluated the **final functional output** of the pipeline

The Test Set:

- Evaluation was conducted on an independent test set of **2,566 frames**.
- Each frame was manually annotated with the final gaze classification: 'Left Screen', 'Right Screen', or 'No Screen’.

The Workflow:

- Model detects features (Eyes/Nose/Horns).
- Algorithm calculates the vector and intersection.
- System outputs a classification ('Left', 'Right', 'No Screen') .

The system's algorithmic output was compared directly against the manual ground truth for that frame.

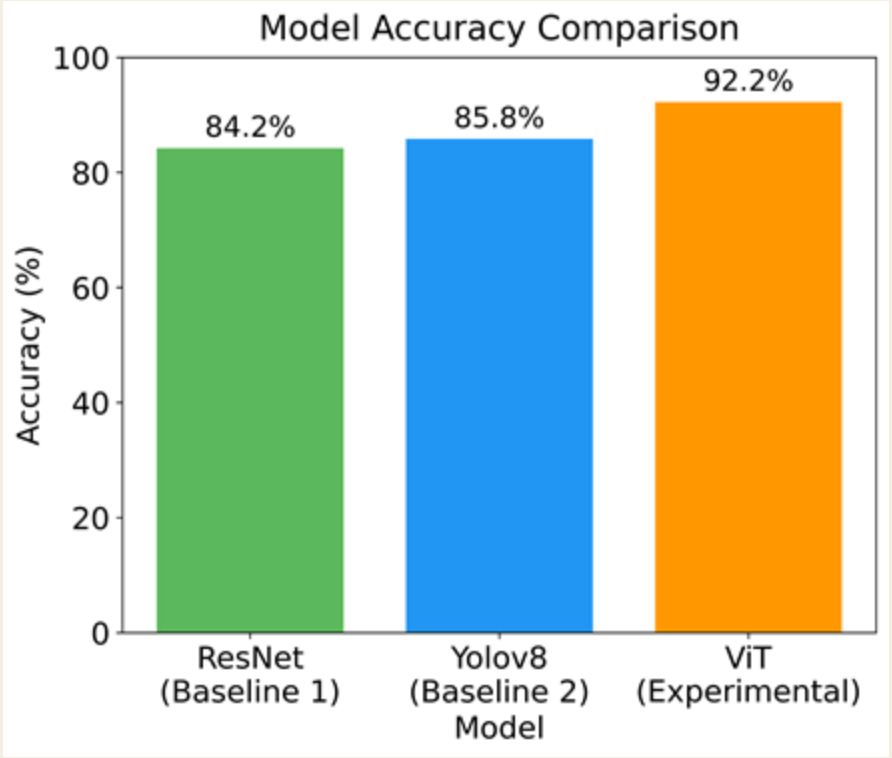
Performance Comparison

Accuracy: ViT Achieved the highest overall accuracy (92.2%)

Potential Explanation (ViT characteristics):

- Global attention of Transformer is better than local focus of CNN.
- Superior spatial inference (e.g. position of nose relative to horns).
- Robust during occlusion.

This suggests that ViT model can be integrated into the pipeline to eliminate bottleneck



Error Analysis (Test Set)

Model	Class	Precision	Recall	F1-Score
Resnet - 101	Left	0.950	0.782	0.858
	Right	0.850	**0.747**	0.795
	No Screen	**0.691**	0.897	0.781
ViT	Left	0.997	0.930	0.962
	Right	0.993	**0.786**	0.877
	No Screen	**0.777**	0.996	0.873
YOLOv8m	Left	0.947	0.804	0.870
	Right	0.881	**0.745**	0.808
	No Screen	**0.692**	0.905	0.784

Error Analysis (Test Set)

- **Model Performance Analysis:**
  - **"Right Screen"** Recall values are lower than others.
  - **"No Screen"** Precision values are lower than others.
  - This suggests all models struggled with "Right Screen" and frequently misclassified it as "No Screen."

- **CNN Performance:**
  - **Succeeds** by localising high-contrast features regardless of overall pose.
  - **Fails** when local details are obscured by bad lighting, blending backgrounds, or motion blur.

- **ViT Performance:**
  - **Succeeds** by using global context to overcome obscured local details.
  - **Fails** when unusual body postures are paired with misleading high-contrast textures.

Key Conclusions

Achievements

ViT-based architecture achieves 92.2% accuracy, outperforming standard CNN baselines by over 6%.

Impact

Enables scalable cognitive research by automating the previously labor-intensive coding bottleneck.

Next Steps

Modeling gaze as a 2D cone field and integrating object tracking for group dynamics.

Thank You

Special thanks to the caretakers and technical staff at FBN.
And a heartfelt thank you to our animal participants.

EUP AH&W European Partnership on Animal Health and Welfare

This work was co-funded buy the European Union's Horizon Europe Project, 101136346 EUPAHW