KI ABSICHERUNG
Safe AI for Automated Driving

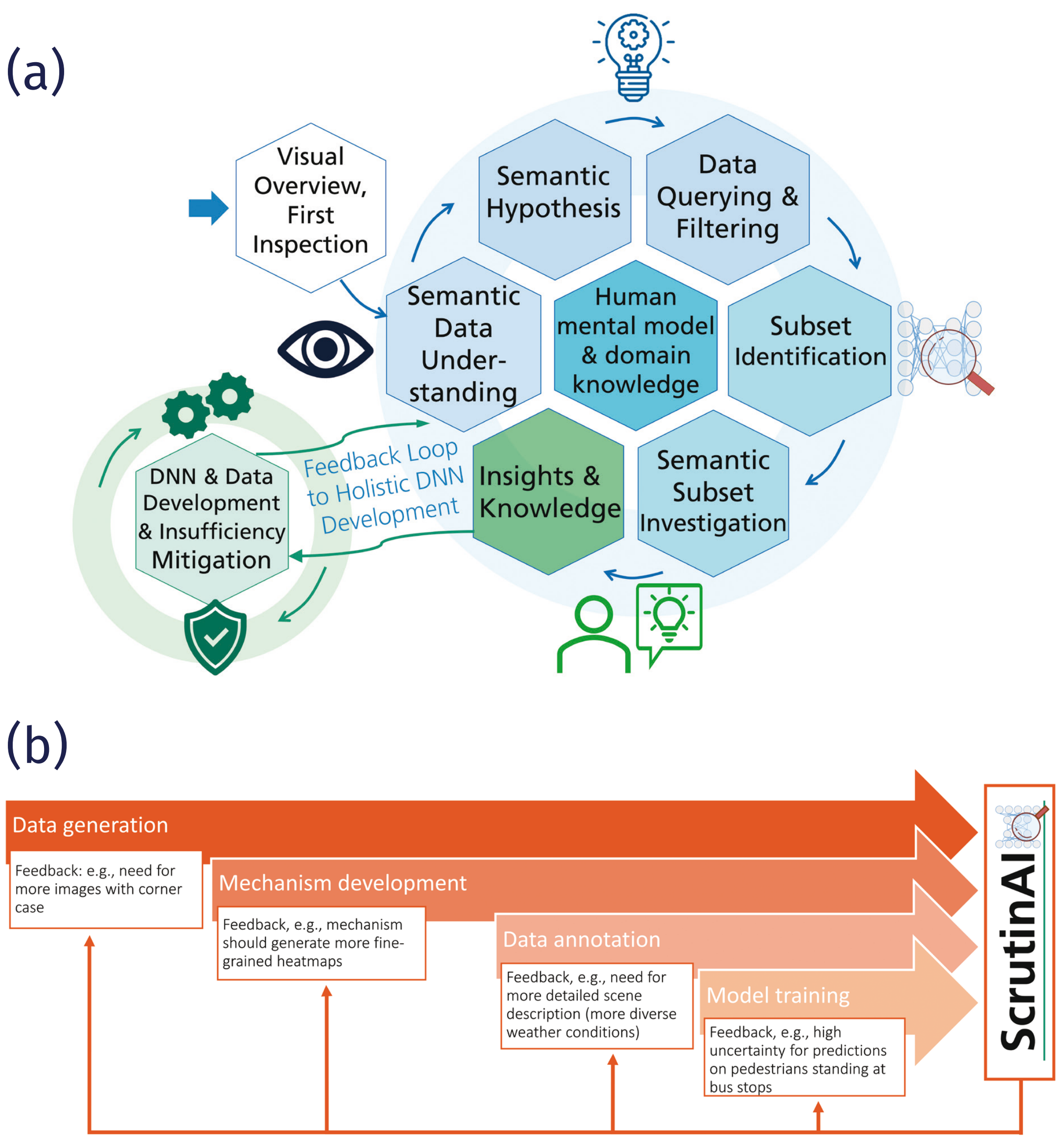
Visual Exploration and Semantic Analysis of DNN Weaknesses with ScrutinAI

Elena Haedecke, Michael Mock

DNN specific Safety Concern: Incomprehensible Behavior

Investigating DNN predictions

The performance of a DNN usually depends on a combination of different dimensions and cannot be explained by looking at a single dimension of the input. Rather, several dimensions have to be investigated as potential influencing factors. In order to efficiently arrive at relevant insights within the large amount of data and dimensions, tool support is needed that guides the human in terms of visual analytics (VA). The goal is to allow the analyst and/or the auditor to use the insights gained during the analysis process to find evidence for the origins or the absence of specific insufficiencies, ultimately contributing towards an overall safety argumentation. Thus, the workflow of the tool focuses on efficiently guiding the human analyst on utilizing domain knowledge and concentrating on semantic aspects of the model and data (see Figure 1 (a)).


(a)

(b)

Visual Overview, First Inspection
Semantic Hypothesis
Data Querying & Filtering
Semantic Data Under-standing
Human mental model & domain knowledge
Subset Identification
Insights & Knowledge
Semantic Subset Investigation
DNN & Data Development & Insufficiency Mitigation
Feedback Loop to Holistic DNN Development
Data generation
Feedback: e.g., need for more images with corner case
Mechanism development
Feedback, e.g., mechanism should generate more fine-grained heatmaps
Data annotation
Feedback, e.g., need for more detailed scene description (more diverse weather conditions)
Model training
Feedback, e.g., high uncertainty for predictions on pedestrians standing at bus stops
ScrutinAI

Figure 1: Workflow (a) and feedback loop (b).

Safety Hypothesis:

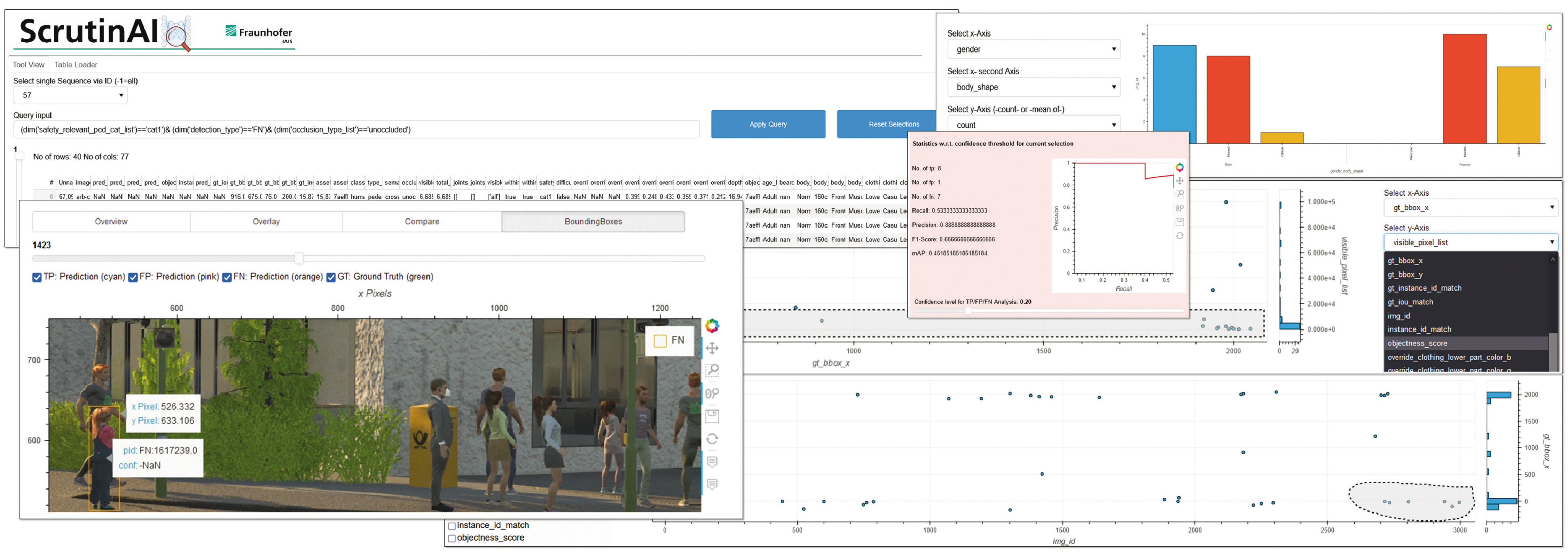
The method addresses the safety concern Incomprehensible behavior. It enables the analyst and/or the auditor to identify presence and causes of incorrect predictions by visually exploring the DNN decisions from different perspectives with the interactive tool ScrutinAI (see Figure 2).

Deriving local and global explanations

To identify noticeable patterns and systematic weaknesses of the DNN and to investigate their causes, common performance metrics are linked to human-understandable semantic concepts defined by domain experts. This helps to establish semantic hypotheses, which are the basis for further, more in-depth investigation of subsets of interest, that potentially represent weaknesses of the model. By examining individual predictions, local explanations can be derived, and by considering larger subsets of data, inferences can be made about the global behavior of the model. During this investigation, the analyst either finds sufficient and strong evidence to verify or falsify the hypothesis, or proceeds with deeper analysis. Insights and knowledge gained during the investigation are fed back to the other stakeholders to mitigate the shortcomings of the DNN (see Figure 1 (b)).

References:

Haedecke et al.: ScrutinAI: A Visual Analytics Approach for the Semantic Analysis of Deep Neural Network Predictions, 2022. EuroVis Workshop on Visual Analytics (2022)


ScrutinAI

Tool View | Table Loader

Select single Sequence via ID (-1=all): 57

Query input: (dim('safety_relevant_ped_cat_list')=='cat1')& (dim('detection_type')=='FN')& (dim('occlusion_type_list')=='unoccluded')

No of rows: 40 No of cols: 77

Select x-Axis: gender
Select x- second Axis: body_shape
Select y-Axis (-count- or -mean of-): count

Statistics w.r.t. confidence threshold for current selection
No. of tp: 8
No. of fp: 1
No. of fn: 7
Recall: 0.5333333333333333
Precision: 0.8888888888888888
F1-Score: 0.6666666666666666
mAP: 0.4518518518518518

☑ TP: Prediction (cyan) ☑ FP: Prediction (pink) ☑ FN: Prediction (orange) ☑ GT: Ground Truth (green)

x Pixel: 526.332
y Pixel: 633.106
pid: FN:1617239.0
conf: -NaN

Figure 2: ScrutinAI visually represents model inputs, outputs, and metadata through linked, interactive elements. The dynamic adjustment of parameters, textual queries or the graphical selection of interesting data subsets allow an investigation of the influence and interdependence of attributes.

Fraunhofer
IAIS

For more information contact:
elena.haedecke@iais.fraunhofer.de
michael.mock@iais.fraunhofer.de

KI Absicherung is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

KI FAMILIE

https://www.ki-absicherung-projekt.de   @KI_Familie   KI Familie

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag

Image Sources:
Figure 1: Fraunhofer IAIS; Figure 2: Fraunhofer IAIS / MackeVision