Identifying Confusion Trends in Concept-based XAI for Multi-Label Classification

Authors: **Haadia Amjad**, Kilian Göller, Steffen Seitz, Carsten Knoll, Ronald Tetzlaff

Chair of Fundamentals of Electrical Engineering,
TUD | Dresden University of Technology,
Dresden, Germany

E-mail: haadia.amjad@tu-dresden.de
EXPLAINABILITY 2025, Barcelona, Spain

Technische Universität Dresden

SECAI

IARIA

Haadia Amjad

Haadia Amjad has been a Research Associate (and doctoral candidate) at TU Dresden since January 2024. She has a Master’s degree in Computer Science (October 2023).

With a total of 6 years of experience in **Computer Vision**, 3 years in **Explainable AI**, and overall 4 years of technical consultancy in start-ups, she aims for new horizons in applicative research development.

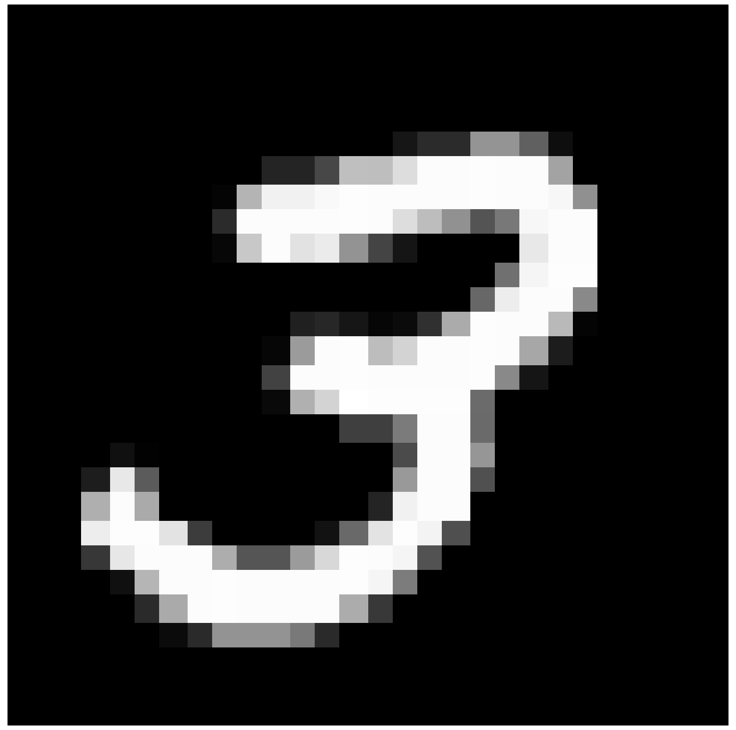
Research Focus: Explainable AI, AI Alignment, Computer Vision.

Motivation

- Deep Neural Networks (DNNs) are widely used for high-risk applications.
- Beyond accuracy, **understandability** and **trust** are crucial.
- Many XAI methods exist, but do we know *how to use them effectively*?

Motivation - The Case Study

- Real-world images are complex, with **multiple overlapping labels** and **background noise**.
- These conditions can lead to **confusion** - wrong or ambiguous label predictions.
- **Concept-based XAI (CXAI)** offers insights into what a model learns, not just where it looks.



MNIST

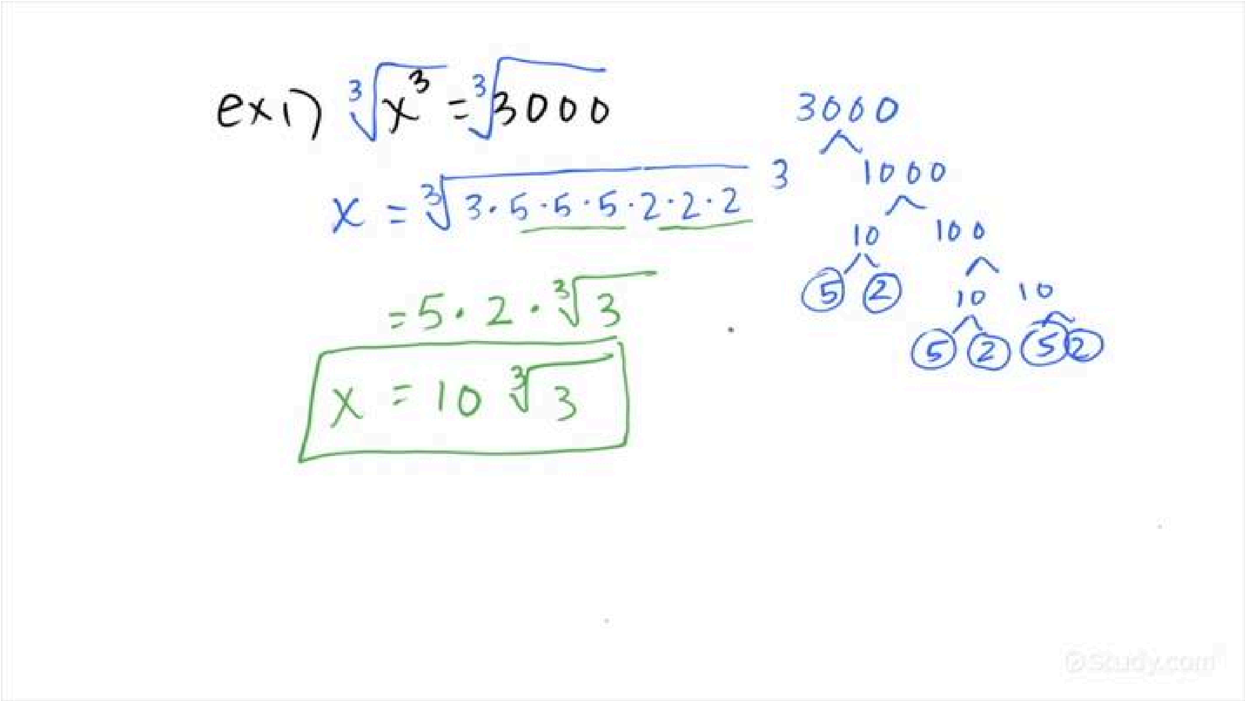
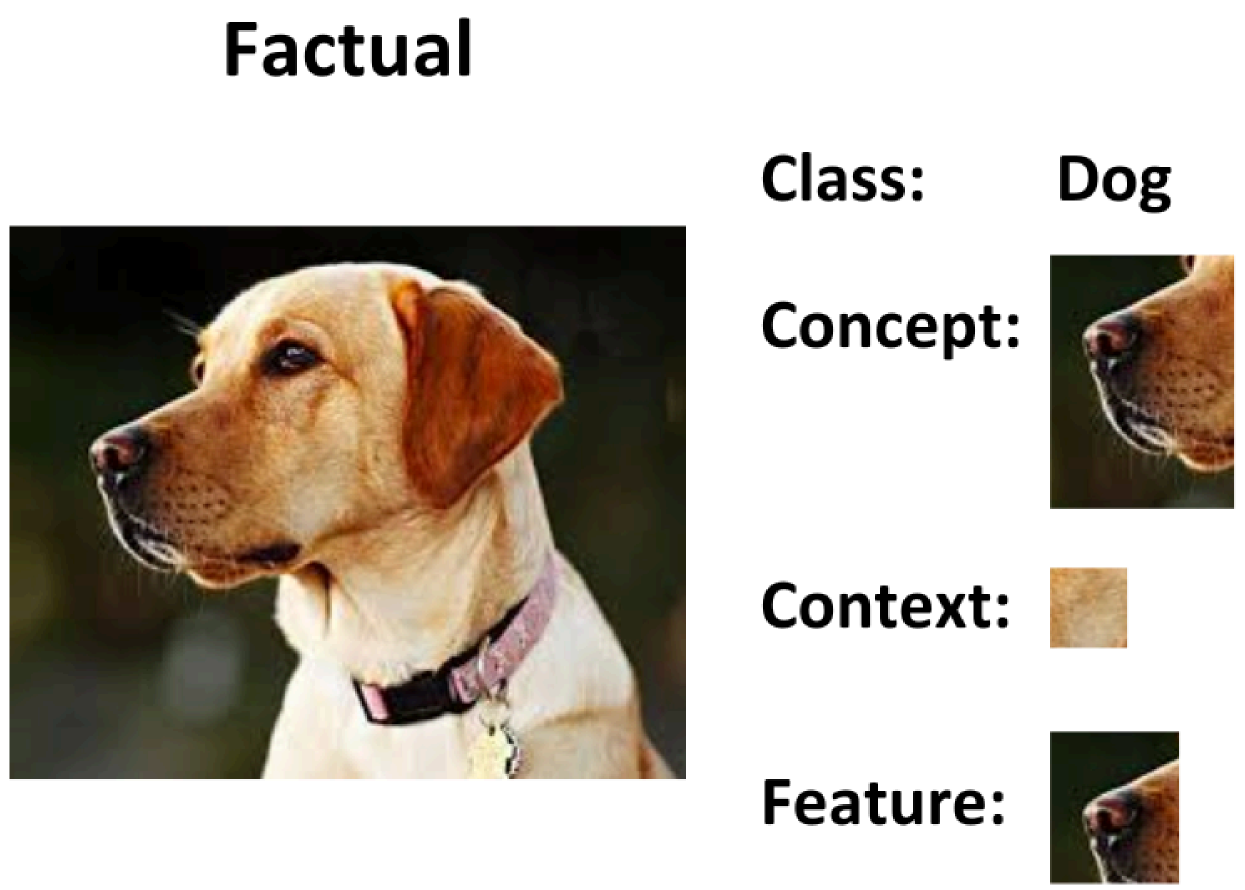
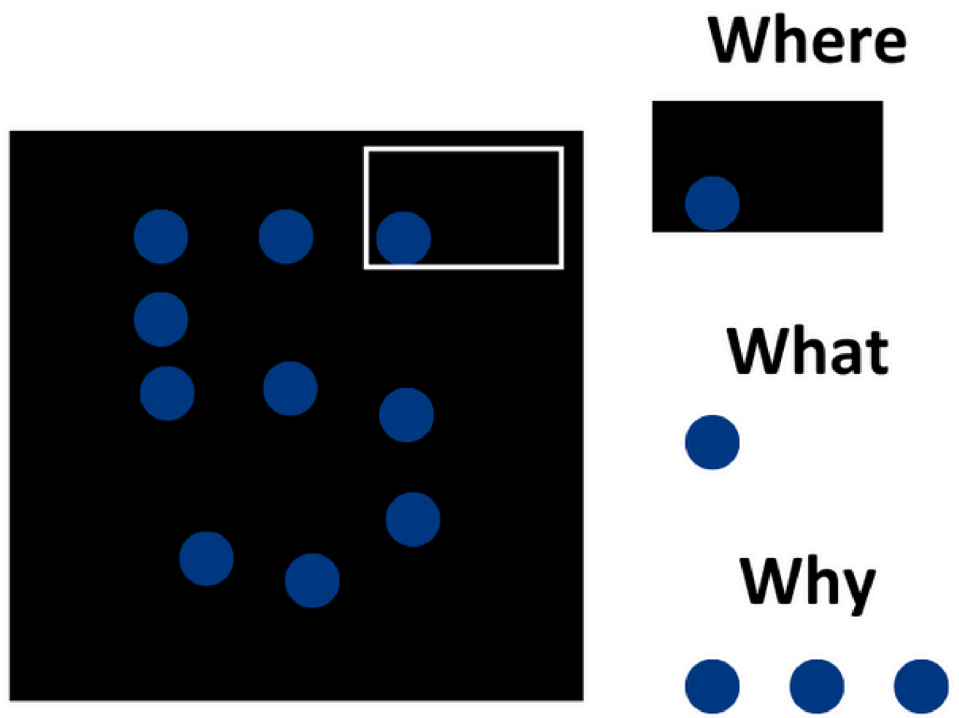


Image from study.com

Concept-based XAI (CXAI)

Concept-based Explainable AI (CXAI) explains deep models through **human-understandable** concepts rather than features. (*what* + *where*)





Research Questions

- **RQ1:** What do advanced CXAI methods reveal about model behaviour on real-world multi-label datasets?

- **RQ2:** What evaluation strategies can validate the trustworthiness of CXAI explanations in real-world applications?

- **RQ3:** Can CXAI identify plausible confusion patterns arising from visual, contextual, or dataset-induced factors?

Main Contributions

- Demonstrated that CXAI methods (CRP, CRAFT) **expose learning weaknesses** in multi-label classification DNNs.
- Found that **greater concept distinctiveness correlates with less confusion.**
- Showed that **environmental concepts expose dataset biases** affecting model generalizability.

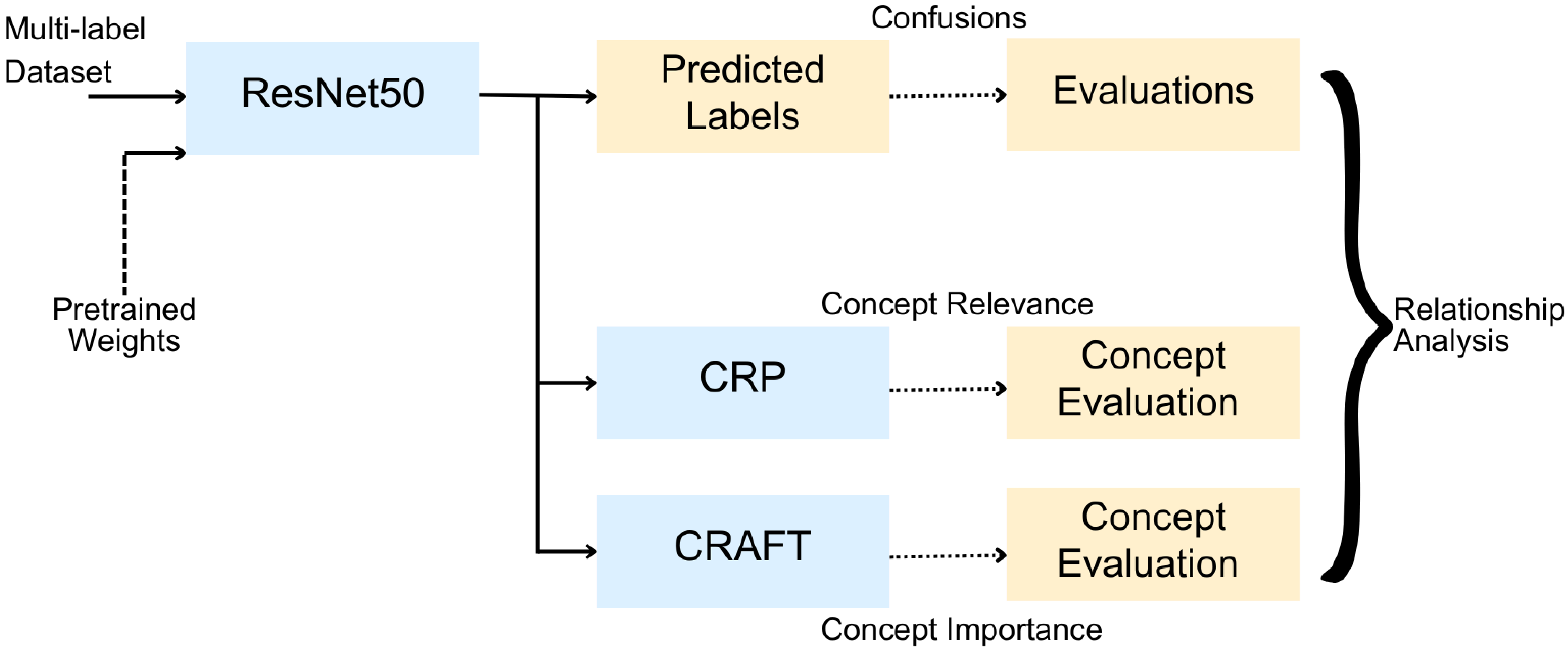
Experimental Setup

- Trained two DNNs, **VGG-16** and **ResNet50**, on the 20 most frequent labels of **MS-COCO 2017**.
  - Dataset: MS-COCO
    - 80 classes (2017)
    - Train: 106,200 images **(90%** of Train2017)
    - Test: 11,800 images (**10%** of Train2017)
    - Validation: 5000 (**100% of Val2017**)
- Evaluated two CXAI methods:
  - **CRP (Concept Relevance Propagation)** – measures **concept relevance** for target classes.
  - **CRAFT (Concept Recursive Activation Factorization)** – measures **concept importance** overall.
- Each model was tested under two scenarios:
  - **Scenario 1:** Well-performing model
  - **Scenario 2:** Poor-performing model

Metrics Terminology

- **Concept Distinctiveness** - How unique a concept's representation is (0–1 scale).
- **Concept Error** - Use of incorrect or irrelevant concepts during prediction.
- **Mutual Information (MI)** - Shared information between labels or concepts (reveals dependency).
- **Jaccard Similarity** - Label co-occurrence measure.

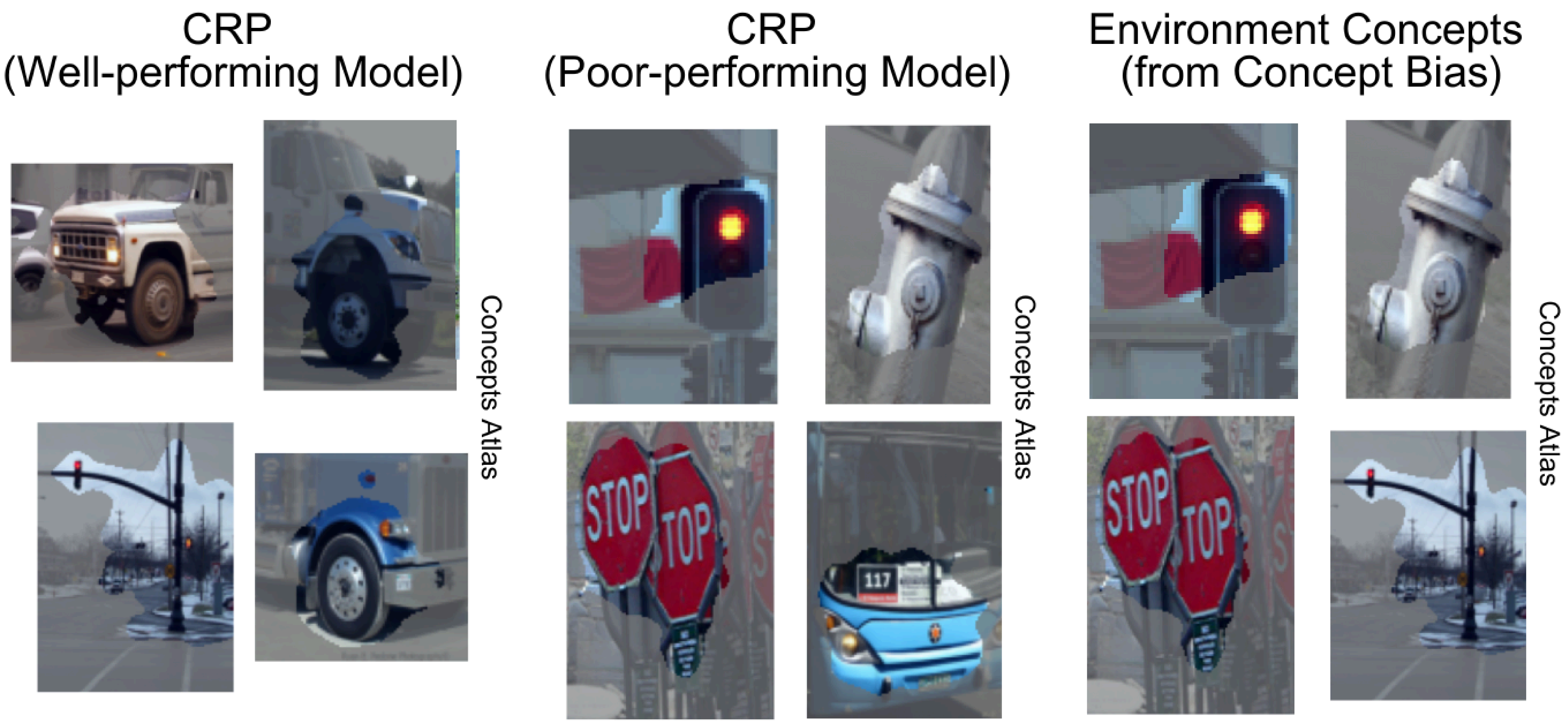
Research Pipeline



Result 1 (RQ1): Confusion in Labels Can Be Understood by Their Explanations

- Confusion arises between co-occurring or visually similar classes (e.g., person-car-chair).
- MI analysis confirms **models learn contextual dependencies, not isolated features**.
- Example: *"person"* strongly linked to *"handbag"* and *"backpack."*
- **CXAI reveals** whether confusion stems from *visual overlap, dataset co-occurrence, or mislearning.*
- Low distinctiveness & high concept error align with higher confusion.

Result 1 (RQ1): Confusion in Labels Can Be Understood by Their Explanations



Result 2 (RQ2): Distinctiveness Reduces Conceptual Confusion

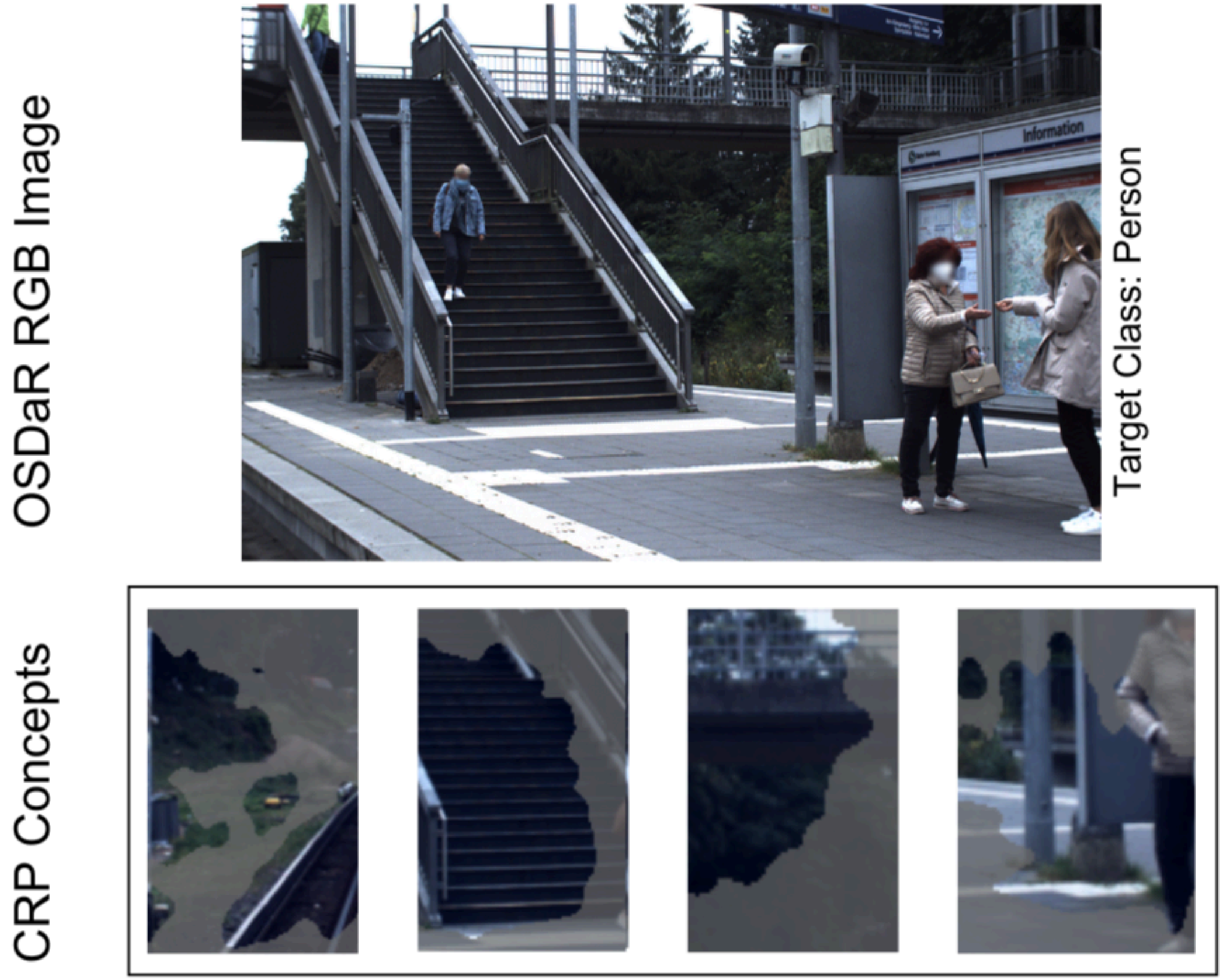
- **Well-performing models exhibit higher concept distinctiveness** - clear class boundaries.
- Poor-performing models blur conceptual separations, relying on contextual bias.
- Correlation: High distinctiveness -> low confusion -> better generalization.

Concept-based XAI method	Model category	Highest ranking distinctiveness class	Highest distinctiveness score	Second Highest distinctiveness score	Lowest ranking distinctiveness class	Lowest distinctiveness score	Second lowest distinctiveness score
CRP	well performing model	person	car	skateboard	sports ball	tennis racket	baseball bat
	poor performing model	giraffe	baseball glove	sports ball	baseball glove	sports ball	baseball bat
CRAFT	well performing model	cow	cellphone	car	parking meter	car	traffic light
	poor performing model	car	kite	giraffe	tie	traffic light	zebra

Result 2 (RQ2): Distinctiveness Reduces Conceptual Confusion

Result 3 (RQ3): Environmental Concepts Reveal Dataset Biases

- **Environmental** (non-target) **concepts show how datasets mislead models.**
- Indicates strong *dataset-induced bias* even in high-accuracy models.
- On the **OSDaR23 dataset, CRP** revealed "person" identified via staircases or platforms - <u>unlabeled features</u>)



Implications

Concept-based analysis helps audit model reasoning:

- Useful for **safety-critical AI**, e.g., medical imaging, autonomous systems.
- Aids **dataset debugging**: spotting spurious correlations and environmental factors early.
- Moves toward **interpretable generalization**, not just performance.

Future Work

- Extend analysis to newer architectures (e.g., Vision Transformers).
- Expand to **more diverse, domain-specific datasets.**
- Incorporate **human-in-the-loop evaluation** for concept validity.
- Automate confusion detection via **CXAI-based metrics.**

Summary and Conclusion

- Confusion in multi-label classification mirrors conceptual confusion inside models.
- **CXAI (CRP & CRAFT)** effectively uncovers:
    - Learning weaknesses
    - Lack of concept separation
    - Dataset-induced biases

- **Higher concept distinctiveness → lower confusion → better generalization.**
- CXAI is a valuable diagnostic lens for understanding how models think.

Thank You

Haadia Amjad
E-mail: haadia.amjad@tu-dresden.de

References

1. R. Tagiew et al., “Osdar23: Open sensor data for rail 2023,” in International Conference on Robotics and Automation Engineering, 2023, pp. 270–276.
2. A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, “Layer-wise relevance propagation for neural networks with local renormalization layers,” in International Conference on Artificial Neural Networks, Springer International Publishing, 2016, pp. 63–71.
3. R. R. Selvaraju et al., “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
4. K. K. Singh et al., “Don’t judge an object by its context: Learning to overcome contextual bias,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11 070–11 078.
5. V. V. Ramaswamy, S. S. Y. Kim, R. Fong, and O. Russakovsky, “Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10 932–10 941.
6. R. Achtibat et al., “From attribution maps to human-understandable explanations through concept relevance propagation,” Nature Machine Intelligence, vol. 5,pp. 1006–1019, 2023.
7. T. Fel et al., “Craft: Concept recursive activation factorization for explainability,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2711–2721.

Spurious Correlation vs Environmental Concepts

Spurious Correlation:

- An unintended relationship learned by a model between features and labels.
- Example: associating *“snow”* with *“wolf”* simply because wolves often appear in snowy images.
- It reflects *false or misleading statistical dependencies* in the data.

Environmental Concepts:

- Non-target features or background elements that a model uses during prediction.
- Example: recognizing a *“pen”* through the presence of a *“hand”* or *“desk.”*
- These arise from *contextual bias* in the dataset and influence model explanations at the concept level.

Model Performance

- Scenario 1 (well-performing model):
  - ResNet50: Accuracy: 82.85%, Recall: 85.50,
    - Precision: 58.84, F1 Score: 60.84
  - VGG-16: Accuracy: 84.26%, Recall: 86.91,
    - Precision: 59.74, F1 Score: 58.84
- Scenario 2 (poor-performing model):
  - ResNet50: Accuracy: 58.24%, Recall: 77.04,
    - Precision: 53.82, F1 Score: 42.92
  - VGG-16: Accuracy: 52.85%, Recall: 74.50,
    - Precision: 53.62, F1 Score: 46.12

TABLE III. PERCENTAGE OF CO-OCCURRENCE OF TARGET LABEL WITH OTHER LABELS (TOP 20 FREQUENTLY ANNOTATED LABELS

Class Name	1st Class	%	2nd Class	%
person	car	13.29	backpack	7.85
car	person	69.54	backpack	8.43
motorcy -cle	person	79.55	car	39.32
truck	person	65.15	car	59.80
boat	person	65.69	car	8.66
traffic light	car	61.22	person	59.19
bench	person	73.75	car	14.63
bird	person	24.56	boat	7.29
sheep	person	24.07	dog	7.59
backpack	person	91.06	car	18.69
umbrella	person	86.87	handbag	28.81
handbag	person	90.95	backpack	24.62
kite	person	92.84	car	11.54
bottle	person	53.65	cup	34.65
cup	person	52.76	dining table	50.92
bowl	dining table	47.76	person	40.73
banana	person	41.37	bowl	23.05
potted plant	person	44.07	chair	38.61
dining table	person	49.58	chair	43.29
book	dining table	75.61	cup	52.97

CXAI Methods in Brief

CRAFT

- Derived from Grad-CAM: uses Non-Negative Matrix Factorization and Sobol indices.
- Produces *importance maps* showing major learned concepts and sub-concepts.

CRP

- Based on Layer-wise Relevance Propagation.
- Produces *relevance maps* identifying where and what contributes to class predictions.
- Focuses on concept-level relevance rather than pixel-level saliency.

Concept Distinctiveness Equation - Screenshot from Our Paper

$$D(C_i, C_j) = 1 - \frac{v_{C_i} \cdot v_{C_j}}{|v_{C_i}||v_{C_j}|} \tag{1}$$

Here, v_{C_i} and v_{C_j} are the concept vectors for concepts C_i and C_j, respectively. Concept vectors are directions in activation space that capture distinct features [23].