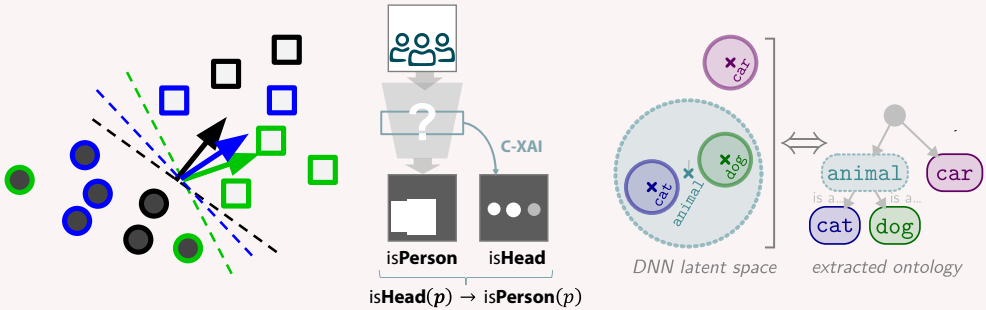


Concept-Based Explainability for Verification and Integration of Symbolic Knowledge in Neural Networks

Jun.-Prof. Dr. Gesina Schwalbe

2026-10-22, Imperial College London

Our Explainable AI Lab

chAI 癌

Sparsh Tiwari



*How to **insert concepts** weakly supervised?*

Raik Dankworth



*How to **extract concepts** and (formally) **verify rules** on them?*

Mahdi Pourghasem



*How to include **neuro-symbolic training objectives**, e.g., utilizing **analogies**?*

Fatemeh Ghorbanigaleshi



*How to explain chemical **NMR spectra**?*

Outline

Motivation

Extract Knowledge: Post-hoc C-XAI
- *Background & Challenges*
- More Structure with Local Concept-based Explanations
- Robust C-XAI

Integrate Knowledge: Ante-hoc C-XAI
- *Background & Challenges*
- Dynamic Orthogonal Concept Bottlenecks

Beyond Concepts: Applications
- Post-hoc Verification of C-XAI Predicates
- Extracting & Manipulating Ontological Commitment of DNNs

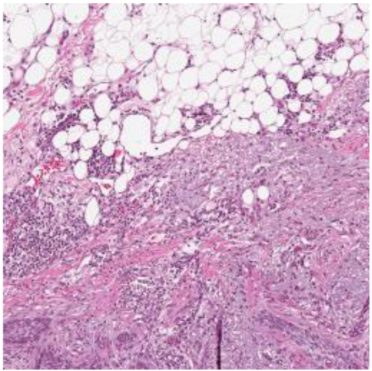
When do we need control over learned knowledge?

Automated Driving



Patrick Fallon/Imago

Cell Imaging for Diagnosis



(Pocevičiūtė et al. 2020)

⇒ **find / understand / prevent failures!**

Challenges for Symbolic Knowledge in Deep Neural Networks

Example (Symbolic Constraint)

$\forall p \in \text{ImageRegions}: \texttt{isHead}(p) \rightarrow \texttt{isPerson}(p)$

- non-symbolic inputs



(Zhang et al. 2018)

Challenges for Symbolic Knowledge in Deep Neural Networks

Example (Symbolic Constraint)

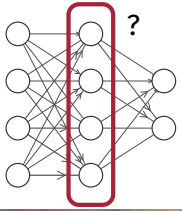
$\forall p \in \text{ImageRegions}: \texttt{isHead}(p) \rightarrow \texttt{isPerson}(p)$

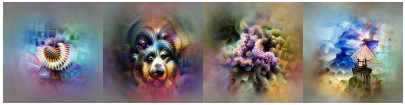
- non-symbolic inputs



(Zhang et al. 2018)

- black-box deep neural networks (DNNs)





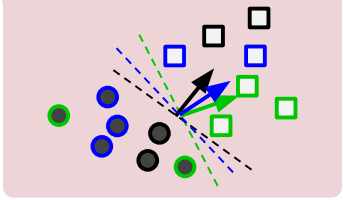
(Olah et al. 2017)

My Research Interests

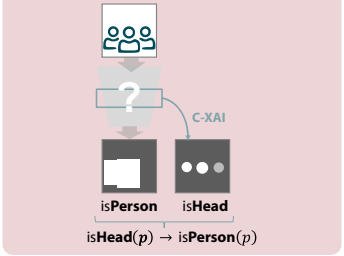
Main Question:

How to *control*, i.e.,

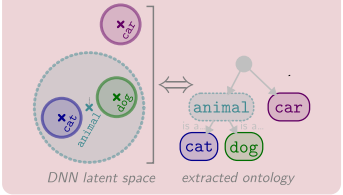
extract



formalize & **verify**



integrate & **correct**



the **symbolic knowledge**, *e.g.*, $\forall p\colon \texttt{isHead}(p) \rightarrow \texttt{isPerson}(p)$,
encoded in **DNNs**, *e.g.*, *their latent spaces* ?

Promising Field: **Concept-based Explainable Artificial Intelligence** (C-XAI)

Concept-based Explainable Artificial Intelligence (C-XAI)

General XAI for DNNs:

Def. (Explanation of DNNs)

...allows a human to *understand* (= align mental model with) (parts of) the DNN or DNN behavior.[1]



(G. Schwalbe and Finzel 2023)

[1] G. Schwalbe and B. Finzel (2023). "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts". In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00867-8

Concept-based Explainable Artificial Intelligence (C-XAI)

General XAI for DNNs:

Def. (Explanation of DNNs)

...allows a human to *understand* (= align mental model with) (parts of) the DNN or DNN behavior.[1]



(G. Schwalbe and Finzel 2023)

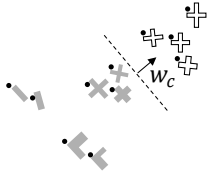
C-XAI: Globally explain DNN representations, i.e., associate

concept data ⟷ **concept model**

semantic concepts, e.g., `isHead`



C-XAI

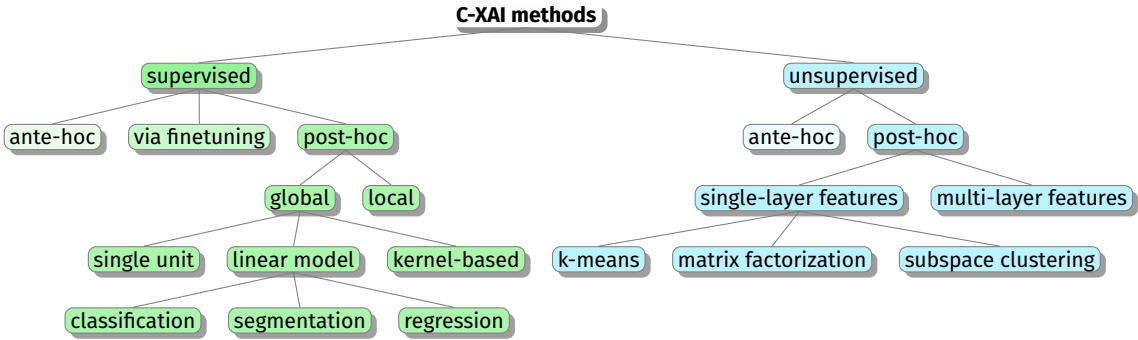


concept vectors w_c resp. regions in DNN latent space

[1] G. Schwalbe and B. Finzel (2023). "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts". In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00867-8

A Taxonomy

Differentiating aspects: **supervision**, **ante-** versus **post-hoc**, arch. **preliminaries**; the **DNN structures** to align concepts with, and type of **concept**.



[1] Jae Hee Lee et al. (2025). "Concept-Based Explanations in Computer Vision: Where Are We and Where Could We Go?" In: *Computer Vision – ECCV 2024 Workshops*. Cham: Springer Nature Switzerland, pp. 266–287. DOI: 10.1007/978-3-031-92648-8_17

Outline

Motivation

Extract Knowledge: Post-hoc C-XAI
- *Background & Challenges*
- More Structure with Local Concept-based Explanations
- Robust C-XAI

Integrate Knowledge: Ante-hoc C-XAI
- *Background & Challenges*
- Dynamic Orthogonal Concept Bottlenecks

Beyond Concepts: Applications
- Post-hoc Verification of C-XAI Predicates
- Extracting & Manipulating Ontological Commitment of DNNs

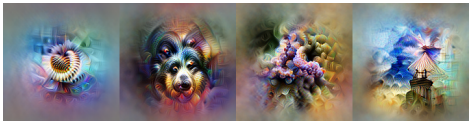
Post-hoc C-XAI: Starting with Single Units

Idea: A filter/neuron η is associated to concept c if its outputs $f_{\to\eta}(x)$ are predictive for c.

Feature visualization (Olah et al. 2017)
(latent $\to$ human language, unsupervised):
Find plausible input(s)

$$x^* = \underset{x \in X}{\operatorname{argmax}} f_{\to\eta}(x)$$

for single neuron / filter η.



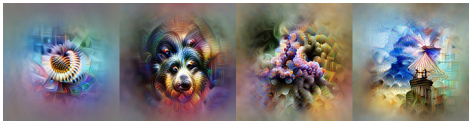
Post-hoc C-XAI: Starting with Single Units

Idea: A filter/neuron η is associated to concept c if its outputs $f_{\to\eta}(x)$ are predictive for c.

Feature visualization (Olah et al. 2017)
(latent $\to$ human language, unsupervised):
Find plausible input(s)

$$x^* = \underset{x \in X}{\operatorname{argmax}} f_{\to\eta}(x)$$

for single neuron / filter η.



Network Dissection (Bau et al. 2017)
(human $\to$ latent language, supervised):
Find single neuron(s)/filter(s)

$$\eta^* = \underset{\eta \in \text{Neurons}}{\operatorname{argmax}} \underset{x \in X_c}{\operatorname{mean}} f_{\to\eta}(x)$$

for concept probing data $X_c \subset X$.

lamps in places net



wheels in object net



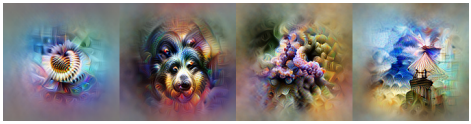
Post-hoc C-XAI: Starting with Single Units

Idea: A filter/neuron η is associated to concept c if its outputs $f_{\to\eta}(x)$ are predictive for c.

Feature visualization (Olah et al. 2017)
(latent $\to$ human language, unsupervised):
Find plausible input(s)

$$x^* = \underset{x \in X}{\operatorname{argmax}} f_{\to\eta}(x)$$

for single neuron / filter η.



Network Dissection (Bau et al. 2017)
(human $\to$ latent language, supervised):
Find single neuron(s)/filter(s)

$$\eta^* = \underset{\eta \in \text{Neurons}}{\operatorname{argmax}} \underset{x \in X_c}{\operatorname{mean}} f_{\to\eta}(x)$$

for concept probing data $X_c \subset X$.

lamps in places net



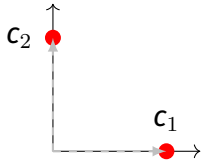
wheels in object net



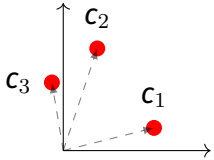
Polysemanticity

Disadvantage of single-neuron representation:
Neglects neuron **polysemanticity**!

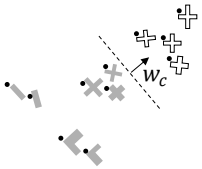
monosemantic neurons:
concepts are unit directions



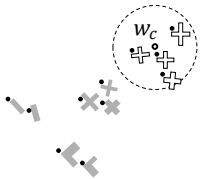
polysemantic neurons:
concepts are arbitrary directions



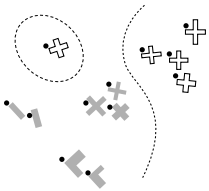
Concept Models and Concept Activation Vectors (CAVs)



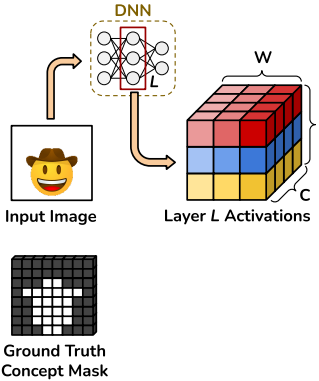
$$\text{Sim}(w_1, w_2) = \frac{w_1 \cdot w_2}{\|w_1\| \cdot \|w_2\|}$$



$$\text{Sim}(w_1, w_2) = \|w_1 - w_2\|_2$$

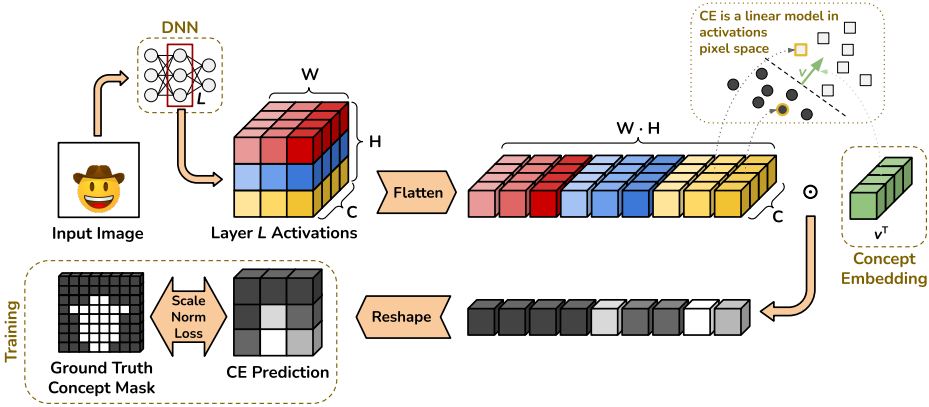


Post-hoc Concept Segmentation



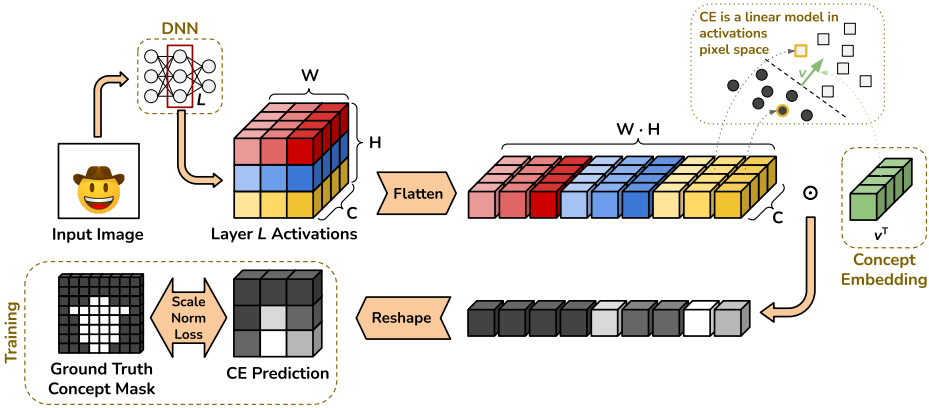
[2] G. Schwalbe (2022b). *Concept Embedding Analysis: A Review.* DOI: 10.48550/arXiv.2203.13909

Post-hoc Concept Segmentation



[2] G. Schwalbe (2022b). *Concept Embedding Analysis: A Review.* DOI: 10.48550/arXiv.2203.13909

Post-hoc Concept Segmentation



quality: $\text{IoU}(x) = \frac{M_{\text{pred}}(x) \cap M_{\text{true}}(x)}{M_{\text{pred}}(x) \cup M_{\text{true}}(x)}$

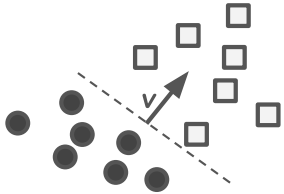
similarity: $\text{CosineSim}(v, v') = \frac{v \cdot v'}{\|v\| \cdot \|v'\|}$

[2] G. Schwalbe (2022b). *Concept Embedding Analysis: A Review.* DOI: 10.48550/arXiv.2203.13909

Beyond Linear C-XAI: Concept Distributions

Global

"*any* `head` against *any* `bg`"



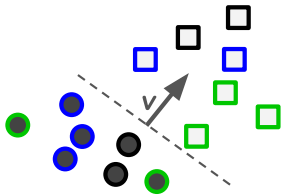
[3] G. Mikriukov, G. Schwalbe, and K. Bade (2025). "Local Concept Embeddings for Analysis of Concept Distributions in Vision DNN Feature Spaces". In: *International Journal of Computer Vision*. DOI: 10.1007/s11263-025-02446-y.

Beyond Linear C-XAI: Concept Distributions

Approach: **Local** Concept Embeddings (**LoCE**[3]) for *individual* images.

Global

"*any* `head` against *any* `bg`"



[3] G. Mikriukov, G. Schwalbe, and K. Bade (2025). "Local Concept Embeddings for Analysis of Concept Distributions in Vision DNN Feature Spaces". In: *International Journal of Computer Vision*. DOI: 10.1007/s11263-025-02446-y.

Beyond Linear C-XAI: Concept Distributions

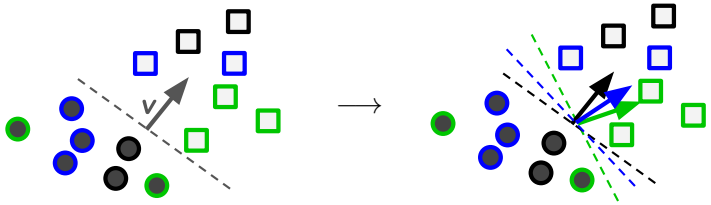
Approach: **Local** Concept Embeddings (**LoCE**[3]) for *individual* images.

Global

"*any* `head` against *any* `bg`"

Local

"*these* `heads` against *this* `bg`"



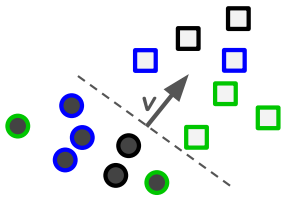
[3] G. Mikriukov, G. Schwalbe, and K. Bade (2025). "Local Concept Embeddings for Analysis of Concept Distributions in Vision DNN Feature Spaces". In: *International Journal of Computer Vision*. DOI: 10.1007/s11263-025-02446-y.

Beyond Linear C-XAI: Concept Distributions

Approach: **Local** Concept Embeddings (**LoCE**[3]) for *individual* images.

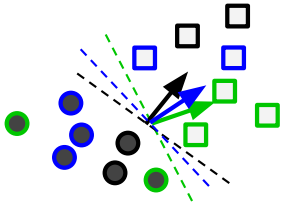
Global

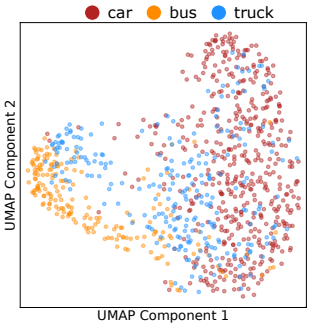
"*any* `head` against *any* `bg`"



Local

"*these* `heads` against *this* `bg`"





⇒ Distribution modeling

⇒ Concept confusion, outliers and subconcepts

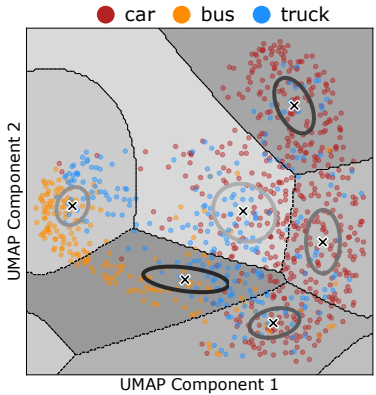
[3] G. Mikriukov, G. Schwalbe, and K. Bade (2025). "Local Concept Embeddings for Analysis of Concept Distributions in Vision DNN Feature Spaces". In: *International Journal of Computer Vision*. DOI: 10.1007/s11263-025-02446-y.

LoCE: Main Results on Concept Distributions

Datasets: **MS COCO** (20 concepts, 2k images, 3k instances); Capybara Dataset (165 images)

Models: **6 DNNs** (det. & cls.; incl. ViTs), $\geq$2 layers each

$\Rightarrow$ Latent representation of visual concepts **overlap**

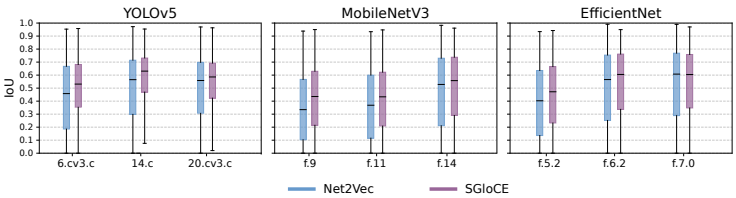


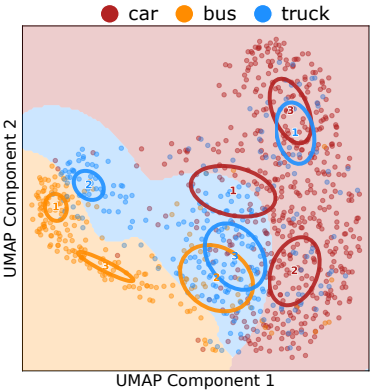
LoCE: Main Results on Concept Distributions

Datasets: **MS COCO** (20 concepts, 2k images, 3k instances); Capybara Dataset (165 images)

Models: **6 DNNs** (det. & cls.; incl. ViTs), $\geq$2 layers each

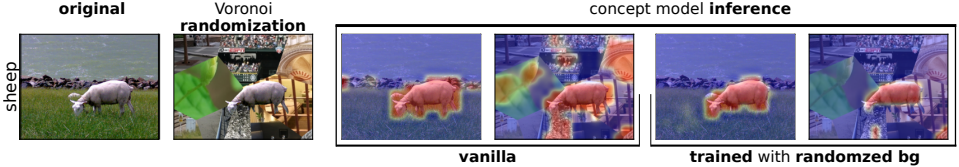
$\Rightarrow$ Latent representation of visual concepts **overlap** and exhibit interesting **sub-concepts**!





Further Challenges: Improving Linear C-XAI Modeling

- **Scalability** via efficient losses & preproc., adaptive patch size (G. Schwalbe 2022a)
- **Uncertainty** calibration (G. Schwalbe, Wirth, et al. 2022)
- **Stability** via ensembling (Rabold, G. Schwalbe, et al. 2020; Mikriukov, G. Schwalbe, Hellert, et al. 2023)
- **Robustness** to background bias via background randomization (Schwalbe et al. 2025)



[3] More challenges in Jae Hee Lee et al. (2025). "Concept-Based Explanations in Computer Vision: Where Are We and Where Could We Go?" In: *Computer Vision – ECCV 2024 Workshops*. Cham: Springer Nature Switzerland, pp. 266–287. DOI: 10.1007/978-3-031-92648-8_17.

Outline

Motivation

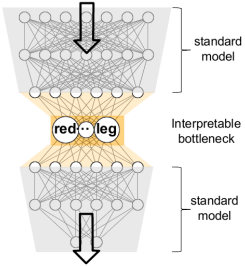
Extract Knowledge: Post-hoc C-XAI
- *Background & Challenges*
- More Structure with Local Concept-based Explanations
- Robust C-XAI

Integrate Knowledge: Ante-hoc C-XAI
- *Background & Challenges*
- Dynamic Orthogonal Concept Bottlenecks

Beyond Concepts: Applications
- Post-hoc Verification of C-XAI Predicates
- Extracting & Manipulating Ontological Commitment of DNNs

Background: Supervised Concept-bottleneck Models

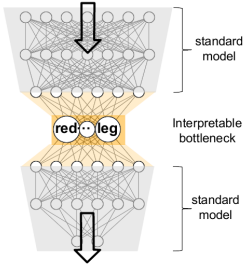
Supervised:



(and also more cool architectures, e.g., contrastive explanations method (CEM) using vector representatives and representing both positive and negative class)

Background: Supervised Concept-bottleneck Models

Supervised:



(and also more cool architectures, e.g., CEM using vector representatives and representing both positive and negative class)

- Concept leakage ⇒ outlier detection

  (De Santis et al. 2026; Marconato et al. 2022).
- Concept completeness ⇒ align with VLM concept spaces

  (Oikarinen et al. 2022)
- Labeling effort

Background: Unsupervised Concept Bottleneck Models

Unsupervised Feature disentanglement:
Features should each capture statistically *independent* latent factors of variation. E.g.:

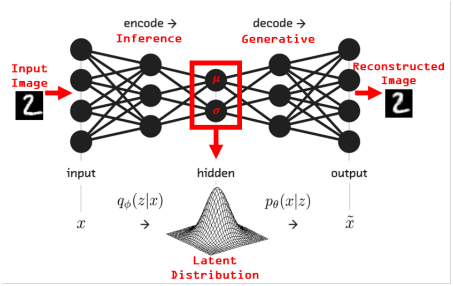
- L_1 regularization
- Penalizing simlarity of distributions, *e.g., via Kullback-Leibler divergence as used in variational autoencoders (VAEs)* (Higgins et al. 2016)

Background: Unsupervised Concept Bottleneck Models

Unsupervised Feature disentanglement:

Features should each capture statistically *independent* latent factors of variation. E.g.:

- L_1 regularization
- Penalizing simlarity of distributions, *e.g., via Kullback-Leibler divergence as used in VAEs* (Higgins et al. 2016)



Caveats of independence:



Human notions are not statistically independent!

⇒ not straightforward to select

- meaningful disentanglement (labels or downstream task required),
- concept subspace dimensionality.

Our research on feature learning: D-OCB

Dynamic Orthogonal Concept Bottleneck for object detection (Sparsh Tiwari et al. 2026):

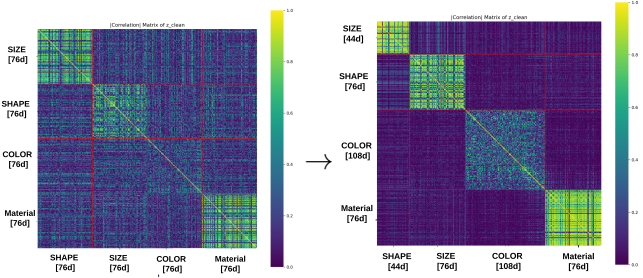
- **weak supervision** by combining CBMs with SlotVAE self-supervision *(further improves using CR-VAE* (Lygerakis et al. 2023) *and MMD* (Zhao et al. 2019) *VAE architectures/training)*

Our research on feature learning: D-OCB

Dynamic Orthogonal Concept Bottleneck for object detection (Sparsh Tiwari et al. 2026):

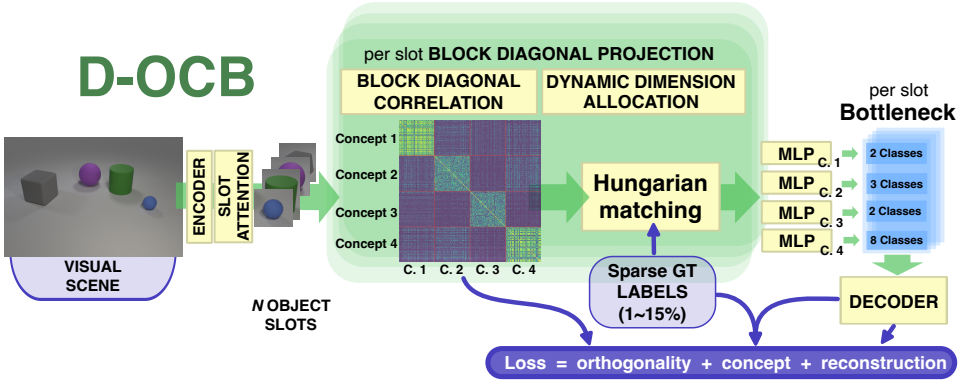
- **weak supervision** by combining CBMs with SlotVAE self-supervision *(further improves using CR-VAE* (Lygerakis et al. 2023) *and MMD* (Zhao et al. 2019) *VAE architectures/training)*
- **dynamically allocate feature subspaces** to concepts via iterated Hungarian matching

Feature Correlations



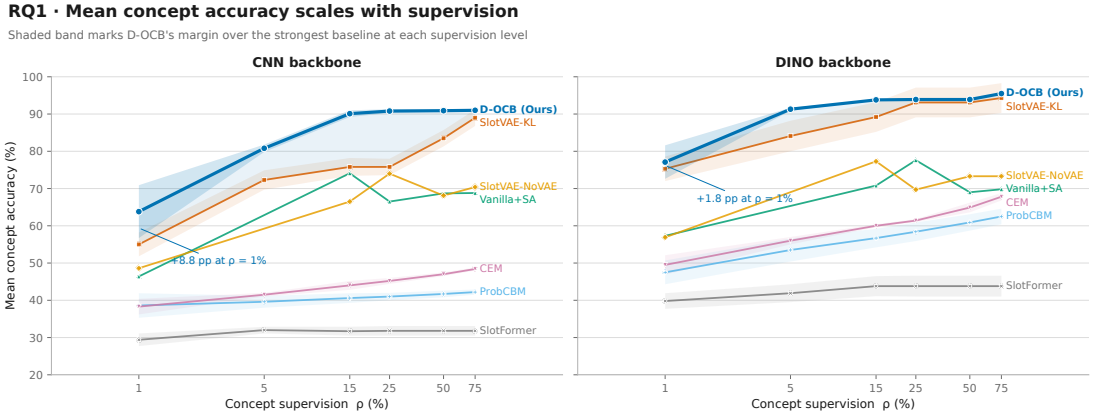
5 epochs **100 epochs**

Our research on feature learning: D-OCB



Sparsh Tiwari et al. (Sept. 1, 2026). "Weakly Supervised Concept Bottleneck Learning for Robust Two Stage Object Centric Visual Reasoning". In: *Proc. 20th Int. Conf. on Neurosymbolic Learning and Reasoning* (Lisbon, Portugal). PMLR. URL: https://openreview.net/forum?id=UFGinErSES

Some D-OCB results



Outlook: Disentangled Feature Learning from Analogies


collie : 
dalmatian = 
lion : 
leopard

(Prade et al. 2021, Fig. 1)

Given: Analogie samples $A : B :: C : D$ ("A is to B as C to D") and countersamples

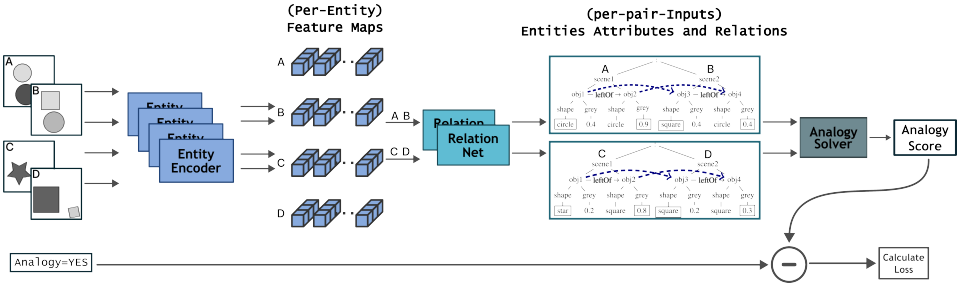
Outlook: Disentangled Feature Learning from Analogies



(Prade et al. 2021, Fig. 1)

Given: Analogie samples $A : B :: C : D$ ("A is to B as C to D") and countersamples

Idea: features optimal to predict whether $A : B :: C : D \iff$ **disentangled**.



Outline

Motivation

Extract Knowledge: Post-hoc C-XAI
- *Background & Challenges*
- More Structure with Local Concept-based Explanations
- Robust C-XAI

Integrate Knowledge: Ante-hoc C-XAI
- *Background & Challenges*
- Dynamic Orthogonal Concept Bottlenecks

Beyond Concepts: Applications
- Post-hoc Verification of C-XAI Predicates
- Extracting & Manipulating Ontological Commitment of DNNs

Outline

Motivation

Extract Knowledge: Post-hoc C-XAI
- *Background & Challenges*
- More Structure with Local Concept-based Explanations
- Robust C-XAI

Integrate Knowledge: Ante-hoc C-XAI
- *Background & Challenges*
- Dynamic Orthogonal Concept Bottlenecks

Beyond Concepts: Applications
- Post-hoc Verification of C-XAI Predicates
- Extracting & Manipulating Ontological Commitment of DNNs

Formalizing human knowledge: Ontologies

Def. (Ontology)

pair of **vocabulary** $\mathcal{V} = \mathcal{C} \cup \mathcal{R}$, and **assumptions** $\mathcal{A}$ on domain D where

- *(concepts)* $\mathcal{C} \subset \mathsf{Fun}(D, \mathbb{B})$
  e.g., class memberships like `isPerson`
- *(instance relations)* $\mathcal{R} \subset \mathsf{Fun}(D \times D, \mathbb{B})$
  e.g., `isCloseTo`
- *(rules)* $\mathcal{A}$ set of assumptions on $\mathcal{V}$,
  e.g., $\forall p :$ `isHead`$(p) \implies$ `isPerson`(p)

(G. Schwalbe, Keser, et al. 2024)

Formalizing human knowledge: Ontologies

Def. (Ontology)

pair of **vocabulary** $\mathcal{V} = \mathcal{C} \cup \mathcal{R}$, and **assumptions** $\mathcal{A}$ on domain D where

- *(concepts)* $\mathcal{C} \subset \mathsf{Fun}(D, \mathbb{B})$
  e.g., class memberships like `isPerson`
- *(instance relations)* $\mathcal{R} \subset \mathsf{Fun}(D \times D, \mathbb{B})$
  e.g., `isCloseTo`
- *(rules)* $\mathcal{A}$ set of assumptions on $\mathcal{V}$,
  e.g., $\forall p :$ `isHead`$(p) \implies$ `isPerson`(p)

(G. Schwalbe, Keser, et al. 2024)

Example (Visual Concept)

Attribute (1-ary logical predicate) applicable to image region, *e.g.,*

materials



,



object parts

Formalizing human knowledge: Ontologies

Def. (Ontology)

pair of **vocabulary** $\mathcal{V} = \mathcal{C} \cup \mathcal{R}$, and **assumptions** $\mathcal{A}$ on domain D where

- *(concepts)* $\mathcal{C} \subset \mathsf{Fun}(D, \mathbb{B})$
  e.g., class memberships like `isPerson`
- *(instance relations)* $\mathcal{R} \subset \mathsf{Fun}(D \times D, \mathbb{B})$
  e.g., `isCloseTo`
- *(rules)* $\mathcal{A}$ set of assumptions on $\mathcal{V}$,
  e.g., $\forall p\colon \texttt{isHead}(p) \implies \texttt{isPerson}(p)$

(G. Schwalbe, Keser, et al. 2024)

Example (Visual Concept)

Attribute (1-ary logical predicate) applicable to image region, *e.g.,*

materials



,

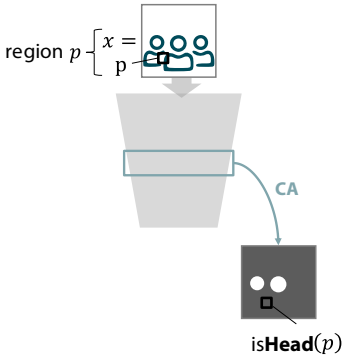


object parts

Note: May also be ***multi-valued logics****,*
e.g., using t-norm logic operators $\vee, \wedge, \neg$ (G. Schwalbe, Wirth, et al. 2022).

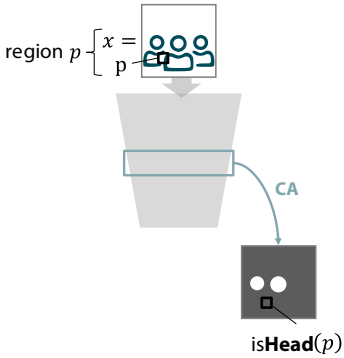
What we can extract from latent space

- **Concepts** & confidence (& quality)
  `isHead`: $D \to [0, 1]$



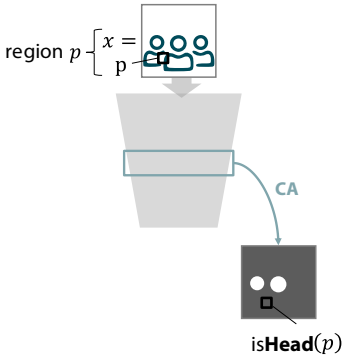
What we can extract from latent space

- **Concepts** & confidence (& quality)
  `isHead`: $D \to [0, 1]$
- 2D **Spatial** relations *(via segmentation)*
  `isCloseTo`: $D \times D \to [0, 1]$



What we can extract from latent space

- **Concepts** & confidence (& quality)
  $\texttt{isHead}\colon D \to [0,1]$
- 2D **Spatial** relations *(via segmentation)*
  $\texttt{isCloseTo}\colon D \times D \to [0,1]$
- Concept **similarity** / overlap *(via LoCEs)*
  $\forall c, d, b\colon \texttt{cat}(c) \wedge \texttt{dog}(d) \wedge \texttt{bus}(b) \to \texttt{Similarity}(c,d) > \texttt{Similarity}(c,b)$
- **Superclass** relationships *(via LoCE clustering)*
  $\forall p\colon \text{is}\texttt{Person}(p) \to \text{is}\texttt{Movable}(p)$

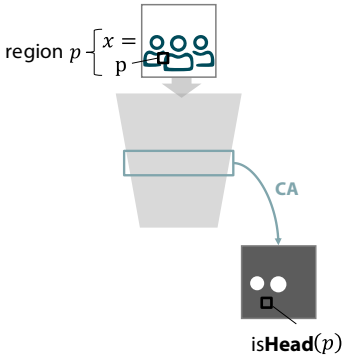


What we can extract from latent space

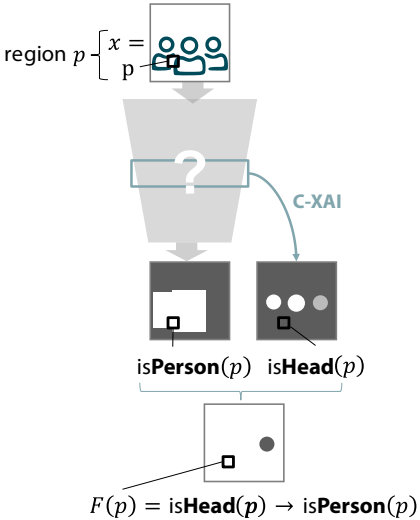
- **Concepts** & confidence (& quality)
  $\texttt{isHead}\colon D \to [0,1]$
- 2D **Spatial** relations *(via segmentation)*
  $\texttt{isCloseTo}\colon D \times D \to [0,1]$
- Concept **similarity** / overlap *(via LoCEs)*
  $\forall c, d, b\colon \texttt{cat}(c) \wedge \texttt{dog}(d) \wedge \texttt{bus}(b) \to \texttt{Similarity}(c,d) > \texttt{Similarity}(c,b)$
- **Superclass** relationships *(via LoCE clustering)*
  $\forall p\colon \text{is}\texttt{Person}(p) \to \text{is}\texttt{Movable}(p)$

Outlook:

- More relations, *e.g., 3D spatial, temporal, relevancy*
- **Invariances**, *e.g., unwanted* ***bias***



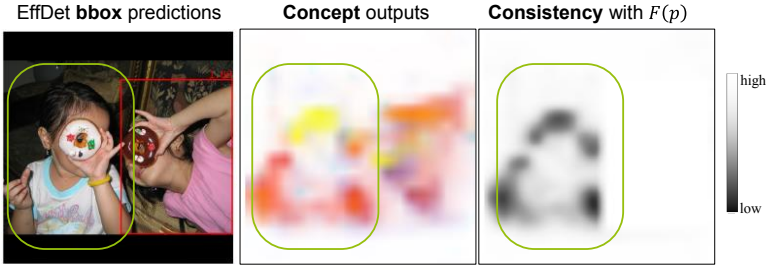
Post-hoc Verification of C-XAI Predicates[4]

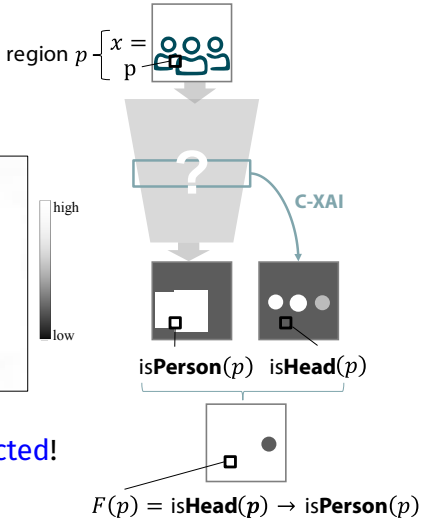


[4] G. Schwalbe, C. Wirth, et al. (2022). *Enabling Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings*. DOI: 10.48550/arXiv.2201.00572.

Post-hoc Verification of C-XAI Predicates[4]

$\forall p \in \text{Pixels}\colon\ (\texttt{isEye}(p) \vee \texttt{isLimb}(p)) \rightarrow \texttt{isPerson}(p)$





$\Rightarrow$ Considerable amount of failures can be detected!

[4] G. Schwalbe, C. Wirth, et al. (2022). *Enabling Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings*. DOI: 10.48550/arXiv.2201.00572.

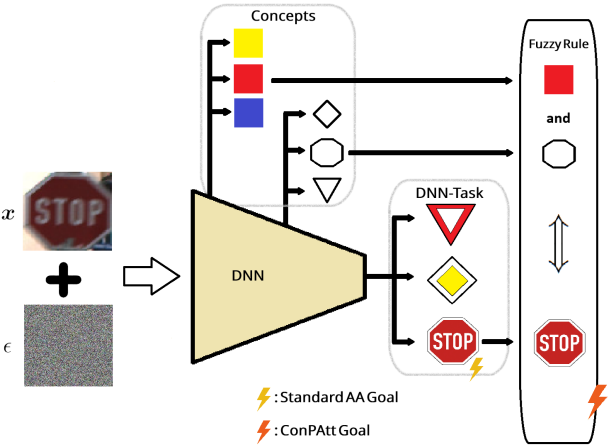
Outlook: Concept-based Adversarial Attacks[5]

$\varphi(x) = \text{stop_sign}(x) \leftrightarrow \text{red}(x) \wedge \text{octagon}(x) \in [0, 1]$

1. Attach outputs.
2. Fuzzify rules φ.
3. Find minimal adversarial perturbation of x that attacks rule consistency

$$\epsilon^* = \underset{\epsilon}{\text{argmin}}(\varphi(x + \epsilon) \leq 0.5)$$

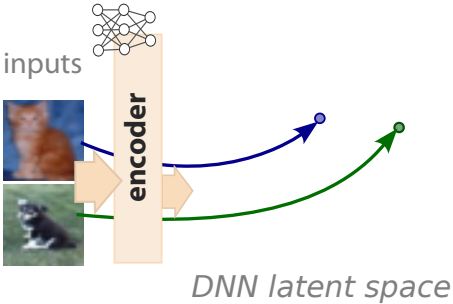
4. *E.g., use samples for adv. training to robustify against biases.*



[5] Raik Dankworth et al. (Oct. 1, 2025). *Attack Logics, Not Outputs: Towards Efficient Robustification of Deep Neural Networks by Falsifying Concept-Based Properties*. DOI: 10.48550/arXiv.2510.03390. arXiv: 2510.03390 [cs]. URL: http://arxiv.org/abs/2510.03390

Rule Discovery: CLIP Training ignores Concept Hierarchies

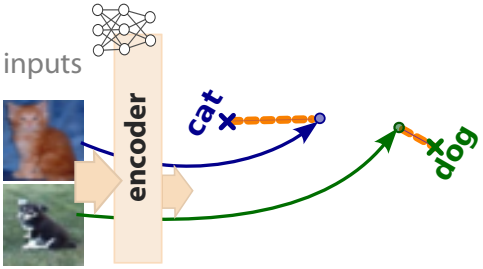
- *So far:* zero-shot **discriminability** of classes



[6] Gesina Schwalbe et al. (Sept. 16, 2026). "Explaining, Verifying and Aligning Semantic Hierarchies in Vision-Language Model Embeddings". In: *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*. Thirty-Fifth International Joint Conference on Artificial Intelligence. Vol. Main Track, pp. 724–734. DOI: 10.24963/ijcai.2026/82. URL: https://www.ijcai.org/proceedings/2026/82

Rule Discovery: CLIP Training ignores Concept Hierarchies

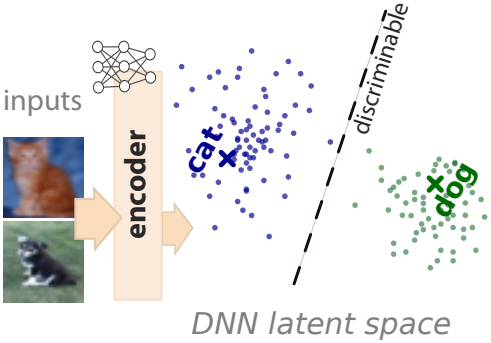
- *So far:* zero-shot **discriminability** of classes



[6] Gesina Schwalbe et al. (Sept. 16, 2026). "Explaining, Verifying and Aligning Semantic Hierarchies in Vision-Language Model Embeddings". In: *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*. Thirty-Fifth International Joint Conference on Artificial Intelligence. Vol. Main Track, pp. 724–734. DOI: 10.24963/ijcai.2026/82. URL: https://www.ijcai.org/proceedings/2026/82

Rule Discovery: CLIP Training ignores Concept Hierarchies

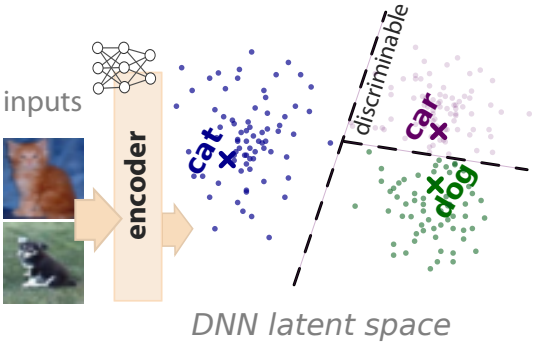
- *So far:* zero-shot **discriminability** of classes



[6] Gesina Schwalbe et al. (Sept. 16, 2026). "Explaining, Verifying and Aligning Semantic Hierarchies in Vision-Language Model Embeddings". In: *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*. Thirty-Fifth International Joint Conference on Artificial Intelligence. Vol. Main Track, pp. 724–734. DOI: 10.24963/ijcai.2026/82. URL: https://www.ijcai.org/proceedings/2026/82

Rule Discovery: CLIP Training ignores Concept Hierarchies

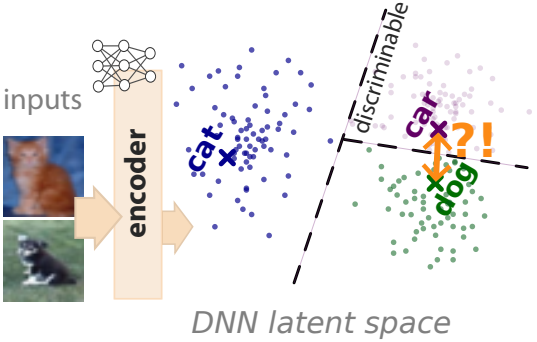
- *So far:* zero-shot **discriminability** of classes



[6] Gesina Schwalbe et al. (Sept. 16, 2026). "Explaining, Verifying and Aligning Semantic Hierarchies in Vision-Language Model Embeddings". In: *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*. Thirty-Fifth International Joint Conference on Artificial Intelligence. Vol. Main Track, pp. 724–734. DOI: 10.24963/ijcai.2026/82. URL: https://www.ijcai.org/proceedings/2026/82

Rule Discovery: CLIP Training ignores Concept Hierarchies

- *So far:* zero-shot **discriminability** of classes
- *New*[6]*:* learned **semantic structure like *hierarchies***
  Why care? ⇒ (1) *inspect*, (2) *use* & (3) *align task-specific bias!*



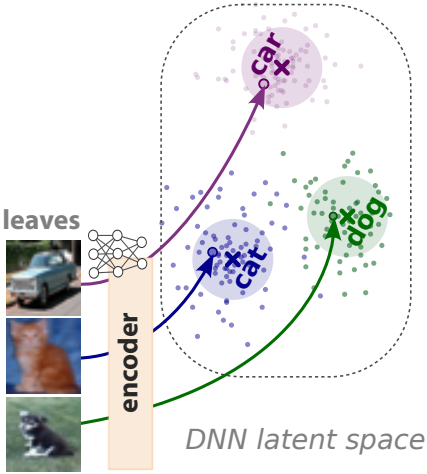
Research Questions

1. What **hierarchy** of leaves is learned?
2. **Agreement** with **human** ontologies and other **encoders**?
3. Can we post-hoc **improve** alignment?

[6] Gesina Schwalbe et al. (Sept. 16, 2026). "Explaining, Verifying and Aligning Semantic Hierarchies in Vision-Language Model Embeddings". In: *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence.* Thirty-Fifth International Joint Conference on Artificial Intelligence. Vol. Main Track, pp. 724–734. DOI: 10.24963/ijcai.2026/82. URL: https://www.ijcai.org/proceedings/2026/82

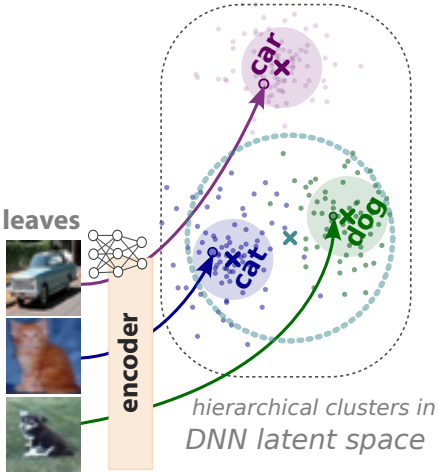
1 Extract Concept Hierarchies via agglomerative hierarchical clustering

Assumption on concept embeddings e: $e(\mathtt{P}) = \underset{\mathtt{C} \in \text{Children}(\mathtt{P})}{\text{mean}}\, e(\mathtt{C})$



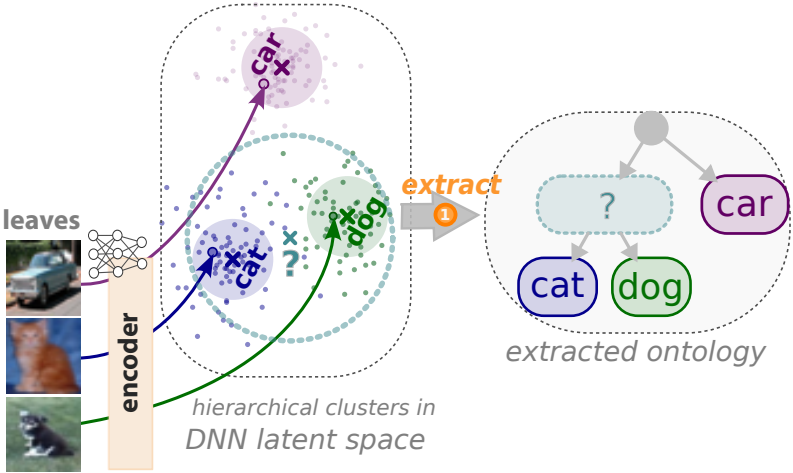
1 Extract Concept Hierarchies via agglomerative hierarchical clustering

Assumption on concept embeddings e: $e(\mathtt{P}) = \underset{\mathtt{C} \in \mathrm{Children}(\mathtt{P})}{\mathrm{mean}}\, e(\mathtt{C})$



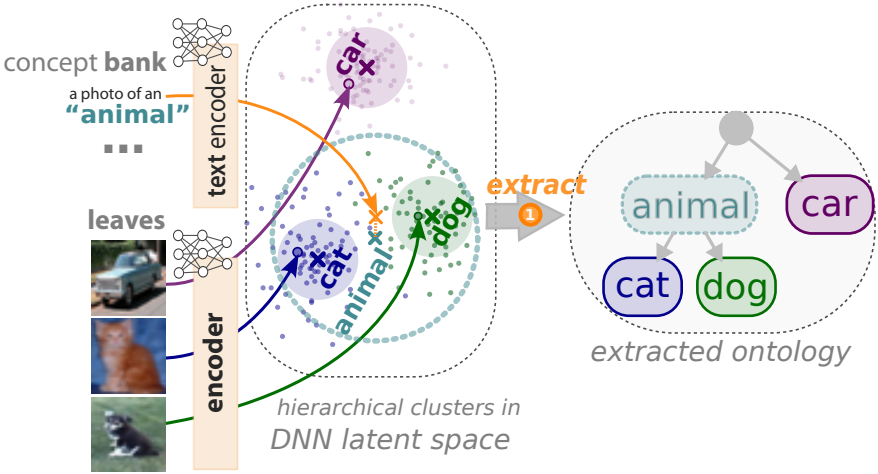
1 Extract Concept Hierarchies via agglomerative hierarchical clustering

Assumption on concept embeddings e: $e(\mathtt{P}) = \underset{\mathtt{C} \in \mathrm{Children}(\mathtt{P})}{\mathrm{mean}}\, e(\mathtt{C})$



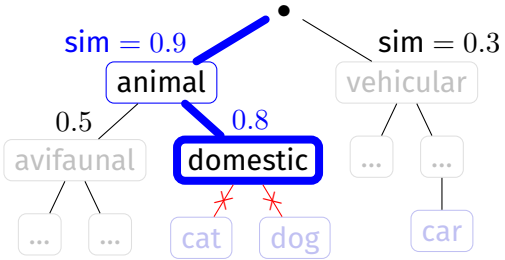
1 Extract Concept Hierarchies via agglomerative hierarchical clustering

Assumption on concept embeddings e: $e(\mathtt{P}) = \underset{\mathtt{C} \in \mathrm{Children}(\mathtt{P})}{\mathrm{mean}}\, e(\mathtt{C})$

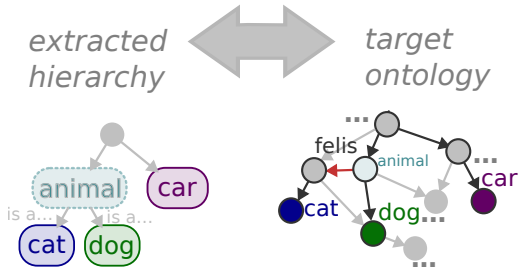


Approach: 2 Use for Inference & Verification

Hierarchical (Wan et al. 2020) & **uncertainty-aware** Inference



Spot & quantify differences



Output: animal, **domestic**.

(Wan et al. 2020) Alvin Wan et al. (Sept. 2020). "NBDT: Neural-backed Decision Tree". In: *Posters 2021 Int. Conf. Learning Representations*. (Visited on 09/16/2021)

Details on Verification: Used Metric

Q: *Do parent-child edges in extracted hierarchy T respect hypernymy of human-defined ontology $\mathcal{O}$?*
$\Rightarrow$ measure distance of parent from childrens' least common ancestor (LCA)

$$\rho(\texttt{P}) \coloneqq \text{LCA}(\{\rho(\texttt{C}) \mid \texttt{C} \text{ child of } \texttt{P}\}) \tag{1}$$

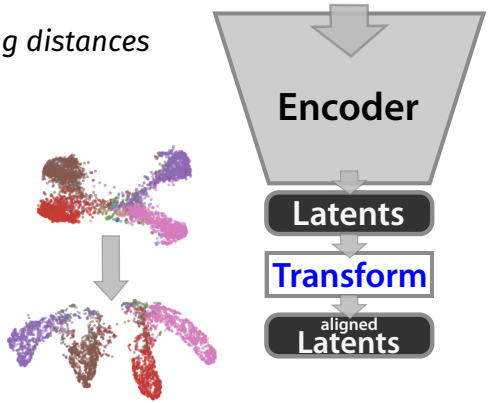
$$S_{\text{edge}}(\texttt{P} \to \texttt{C}) \coloneqq \begin{cases} 0 \text{ if no path in } \mathcal{O} \\ \min(1, \frac{1}{0.5 \cdot \text{dist}(\rho(u), \rho(v))}) \text{ else} \end{cases} \tag{2}$$

$$\textbf{Hierarch. Consistency}(T) \coloneqq \underset{T\text{-edges } \texttt{P} \to \texttt{C}}{\text{mean}} S_{\text{edge}}(\texttt{P} \to \texttt{C}) \tag{3}$$

Approach: 3 Align Hierarchy post-hoc & model-agnostic

Compete

- *maintain **original instance** embedding distances*
- *foster **new class** centroid distances*

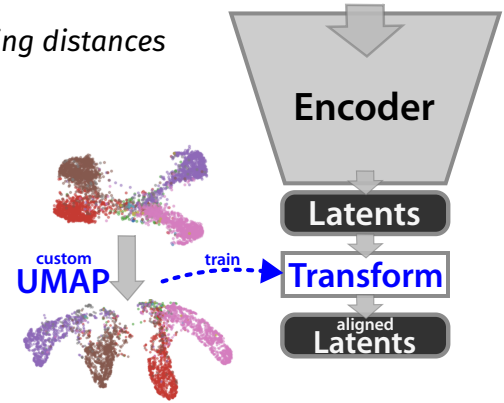


Approach: 3 Align Hierarchy post-hoc & model-agnostic

Compete

- *maintain **original instance** embedding distances*
- *foster **new class** centroid distances*

1. Get pairings via adapted **UMAP** (McInnes et al. 2018) dimensionality reduction method
2. Train fully-connected **DNN**



Findings on Vision Language Models

Experiment Settings: **13 VLMs**, 3 leaf datasets, 3 ontologies $\mathcal{O}$ (SUMO, OpenCyc, Yago); alignment scenarios: leaf swapping, image→text, $\rightarrow \mathcal{O}$.

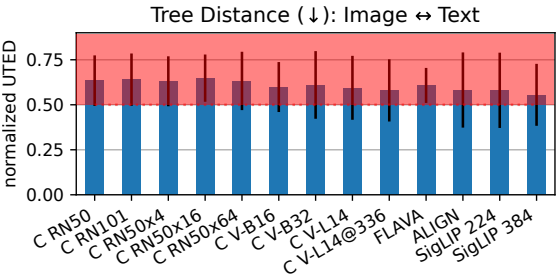
Does discriminability induce ...

...encoder alignment? *No.*

- (2) Text vs. image encoder hierarchies show **high tree edit distance** (beyond random).

...human alignment? *No.*

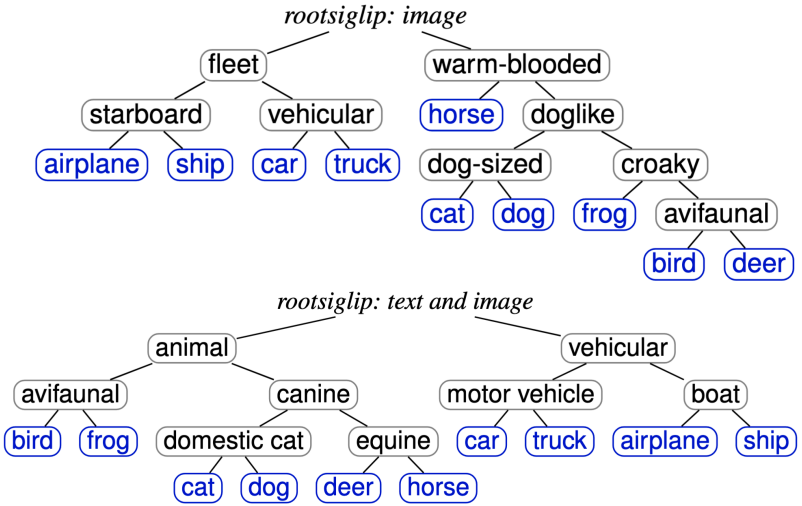
- (2) Zero-shot acc. and hierarch. consistency are significantly **negatively correlated**.
- (3) **Alignment** correction achievable, but **competes** with zero-shot accuracy.



Applications:

- **Knowledge discovery** & assessment
- **Bias removal** / correction
- **Specialization** to task-specific hierarchy

Findings on Vision Language Models



There is more to see (and to do)!

- C-XAI for **outlier** detection (Shoeb, Chan, G. Schwalbe, et al. 2024; Mikriukov, G. Schwalbe, and Bade 2025)
- C-XAI to explore
  - the shape of **adversarial attacks** via C-XAI (Mikriukov, G. Schwalbe, Motzkus, et al. 2024)
  - the impact of **self-supervised pretraining** in latent space (Paulke et al. 2026)
- Integrating symbolic rules into **world models** (Ramachandra et al. 2026)

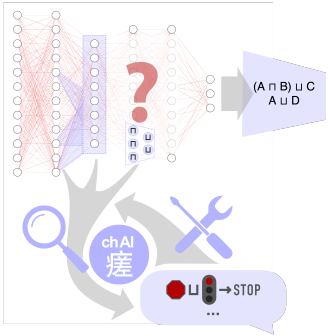
And for more on C-XAI, check out my C-XAI Tutorial.

Summary

My research interests:

understand and **control**
the **symbolic** knowledge encoded in DNNs

- Research field: **C-XAI**
  = associate interpretable concepts with DNN internals
- Current research:
  - **Improving C-XAI** techniques
  - Applications to model **verification** & **correction**



my Junior Research Group
correctable hybrid AI

References I

Bau, David et al. (2017). "Network Dissection: Quantifying Interpretability of Deep Visual Representations". In: *Proc. 2017 IEEE Conf. Comput. Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE Computer Society, pp. 3319–3327. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.354. arXiv: 1704.05796.

Dankworth, Raik and Gesina Schwalbe (Oct. 1, 2025). *Attack Logics, Not Outputs: Towards Efficient Robustification of Deep Neural Networks by Falsifying Concept-Based Properties*. DOI: 10.48550/arXiv.2510.03320. arXiv: 2510.03320 [cs]. URL: http://arxiv.org/abs/2510.03320. Pre-published.

De Santis, Francesco et al. (2026). "V-CEM: Bridging Performance and Intervenability in Concept-Based Models". In: *Explainable Artificial Intelligence*. Ed. by Riccardo Guidotti, Ute Schmid, and Luca Longo. Cham: Springer Nature Switzerland, pp. 48–67. ISBN: 978-3-032-08317-3. DOI: 10.1007/978-3-032-08317-3_3.

Higgins, Irina et al. (Nov. 4, 2016). "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *Posters 5th Int. Conf. Learning Representations*. 5th Int. Conf. Learning Representations (ICLR'2017). URL: https://openreview.net/forum?id=Sy2fzU9gl.

Lee, Jae Hee et al. (2025). "Concept-Based Explanations in Computer Vision: Where Are We and Where Could We Go?" In: *Computer Vision – ECCV 2024 Workshops*. Cham: Springer Nature Switzerland, pp. 266–287. DOI: 10.1007/978-3-031-92648-8_17.

Lygerakis, Fotios and Elmar Rueckert (Nov. 2023). "CR-VAE: Contrastive Regularization on Variational Autoencoders for Preventing Posterior Collapse". In: *2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT)*. 2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT), pp. 427–437. DOI: 10.1109/ACAIT60137.2023.10528405. URL: https://ieeexplore.ieee.org/abstract/document/10528405.

Marconato, Emanuele, Andrea Passerini, and Stefano Teso (Dec. 6, 2022). "GlanceNets: Interpretable, Leak-proof Concept-based Models". In: *Advances in Neural Information Processing Systems*. 2022 Neural Information Processing Systems (NeurIPS'2022). Vol. 35, pp. 21212–21227. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/85b2ff7574ef265f3a4800db9112ce14-Abstract-Conference.html.

McInnes, Leland et al. (Sept. 2, 2018). "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29, p. 861. ISSN: 2475-9066. DOI: 10.21105/joss.00861. URL: https://joss.theoj.org/papers/10.21105/joss.00861.

Mikriukov, G., G. Schwalbe, and K. Bade (2025). "Local Concept Embeddings for Analysis of Concept Distributions in Vision DNN Feature Spaces". In: *International Journal of Computer Vision*. DOI: 10.1007/s11263-025-02446-y.

Mikriukov, G., G. Schwalbe, C. Hellert, et al. (2023). "Evaluating the Stability of Semantic Concept Representations in CNNs for Robust Explainability". In: *Proc. 1st World Conf. eXplainable Artificial Intelligence*. Springer. URL: https://arxiv.org/abs/2304.14864. (*Best Industry Paper Award*).

References II

Mikriukov, G., G. Schwalbe, F. Motzkus, et al. (2024). "Unveiling the Anatomy of Adversarial Attacks: Concept-Based XAI Dissection of CNNs". In: *Explainable Artificial Intelligence*. Ed. by L. Longo, S. Lapuschkin, and C. Seifert. Cham: Springer Nature Switzerland, pp. 92–116. DOI: 10.1007/978-3-031-63787-2_6.

Oikarinen, Tuomas et al. (Sept. 29, 2022). "Label-Free Concept Bottleneck Models". In: The Eleventh International Conference on Learning Representations. URL: https://openreview.net/forum?id=FlCg47MNvBA.

Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert (Nov. 2017). "Feature Visualization". In: *Distill* 2.11, e7. ISSN: 2476-0757. DOI: 10.23915/distill.00007. (Visited on 05/20/2019).

Paulke, Kim et al. (Apr. 2026). "Local Concept Embeddings in the Context of Self-Supervised Learning". In: *ESANN 2026 - Proceedings*. 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'26) (Bruges, Belgium). i6doc.com, p. 806. ISBN: 978-2-87587-095-7. DOI: 10.14428/esann/2026.ES2026-85.

Pocevičiūtė, Milda, Gabriel Eilertsen, and Claes Lundström (2020). "Survey of XAI in Digital Pathology". In: *Lecture Notes in Computer Science* 2020, pp. 56–88. DOI: 10.1007/978-3-030-50402-1_4. arXiv: 2008.06353. (Visited on 02/24/2021).

Prade, Henri and Gilles Richard (2021). "Analogical Proportions: Why They Are Useful in AI". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*. Ed. by Zhi-Hua Zhou. ijcai.org, pp. 4568–4576. DOI: 10.24963/IJCAI.2021/621. URL: https://doi.org/10.24963/ijcai.2021/621.

Rabold, J., G. Schwalbe, and U. Schmid (2020). "Expressive Explanations of DNNs by Combining Concept Analysis with ILP". In: *KI 2020: Advances in Artificial Intelligence*. Lecture Notes in Computer Science. Springer, pp. 148–162. DOI: 10.1007/978-3-030-58285-2_11.

Ramachandra, Ajay, Gesina Schwalbe, and Zoe Falomir (Sept. 16, 2026). "Injecting Qualitative Spatial Reasoning into World Models via Logic Tensor Networks". In: Thirty-Fifth International Joint Conference on Artificial Intelligence. Vol. 5, pp. 4016–4025. DOI: 10.24963/ijcai.2026/447. URL: https://www.ijcai.org/proceedings/2026/447.

Schwalbe, G et al. (2025). *On Background Bias of Post-Hoc Concept Embeddings in Computer Vision DNNs*. DOI: 10.48550/arXiv.2504.08602. (Accepted at 3rd World Conference on Explainable Artificial Intelligence).

References III

Schwalbe, G. (2022a). "Concept Embedding Analysis Based Methods for the Safety Assurance of Deep Neural Networks: Towards Safe Automotive Computer Vision Applications". PhD thesis. Otto-Friedrich-Universität Bamberg, Fakultät Wirtschafts- und angewandte Informatik. DOI: 10.20378/irb-57172.

— (2022b). *Concept Embedding Analysis: A Review*. DOI: 10.48550/arXiv.2203.13909.

Schwalbe, G. and B. Finzel (2023). "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts". In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00867-8.

Schwalbe, G., M. Keser, et al. (Oct. 2024). *Unveiling Ontological Commitment in Multi-Modal Foundation Models*. DOI: 10.48550/arXiv.2409.17109.

Schwalbe, G., C. Wirth, and U. Schmid (2022). *Enabling Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings*. DOI: 10.48550/arXiv.2201.00572.

Schwalbe, Gesina et al. (Sept. 16, 2026). "Explaining, Verifying and Aligning Semantic Hierarchies in Vision-Language Model Embeddings". In: *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*. Thirty-Fifth International Joint Conference on Artificial Intelligence. Vol. Main Track, pp. 724–734. DOI: 10.24963/ijcai.2026/82. URL: https://www.ijcai.org/proceedings/2026/82.

Shoeb, Y. et al. (2024). "Have We Ever Encountered This Before? Retrieving Out-of-Distribution Road Obstacles From Driving Scenes". In: *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7396–7406.

Sparsh Tiwari, Bettina Finzel, and Gesina Schwalbe (Sept. 1, 2026). "Weakly Supervised Concept Bottleneck Learning for Robust Two Stage Object Centric Visual Reasoning". In: *Proc. 20th Int. Conf. on Neurosymbolic Learning and Reasoning* (Lisbon, Portugal). PMLR. URL: https://openreview.net/forum?id=UFGinErSES.

Wan, Alvin et al. (Sept. 2020). "NBDT: Neural-backed Decision Tree". In: *Posters 2021 Int. Conf. Learning Representations*. (Visited on 09/16/2021).

Zhang, Shanshan, Jian Yang, and Bernt Schiele (2018). "Occluded Pedestrian Detection through Guided Attention in CNNs". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, pp. 6995–7003. DOI: 10.1109/CVPR.2018.00731. (Visited on 10/10/2022).

Zhao, Shengjia, Jiaming Song, and Stefano Ermon (2019). "Infovae: Balancing learning and inference in variational autoencoders". In: *Proceedings of the aaai conference on artificial intelligence*. Vol. 33. 01, pp. 5885–5892.