

Concept-Based Explanations in Computer Vision: Where Are We and Where Could We Go?

Jae Hee Lee^{1*}, Georgii Mikriukov², Gesina Schwalbe³, Stefan Wermter¹,
and Diedrich Wolter³

¹ University of Hamburg, Germany

{jae.hee.lee, stefan.wermter}@uni-hamburg.de

² Anhalt University of Applied Sciences, Germany

georgii.mikriukov@student.hs-anhalt.de

³ University of Lübeck, Germany

{gesina.schwalbe, diedrich.wolter}@uni-luebeck.de

Abstract. Concept-based XAI (C-XAI) approaches to explaining neural vision models are a promising field of research, since explanations that refer to concepts (i.e., semantically meaningful parts in an image) are intuitive to understand and go beyond saliency-based techniques that only reveal relevant regions. Given the remarkable progress in this field in recent years, it is time for the community to take a critical look at the advances and trends. Consequently, this paper reviews C-XAI methods to identify interesting and underexplored areas and proposes future research directions. To this end, we consider three main directions: the choice of concepts to explain, the choice of concept representation, and how we can control concepts. For the latter, we propose techniques and draw inspiration from the field of knowledge representation and learning, showing how this could enrich future C-XAI research.

Keywords: Concept-Based Explainable AI · Concept Embedding Analysis · Concept Control · Neuro-Symbolic AI · Knowledge Representation

1 Introduction

As the capabilities of deep learning models grow and as our society uses them more, it becomes increasingly important to *understand* how they work [105] and how to *control* them in effective ways [37]: Understanding how a model works is the basis for trusting the model [53] and for its verification against ethical, privacy, or safety requirements [31]. Control is imperative to effectively enforce requirements by design, during maintenance, or through manual ad-hoc intervention. Understanding and control have formed the basis for a wealth of explainable artificial intelligence (XAI) methods for computer vision (CV) [22, 113, 131].

In XAI for CV, early post-hoc explainability approaches have focused on areas of importance of features in the input image relevant for a vision model's

* Author names are in alphabetic order.

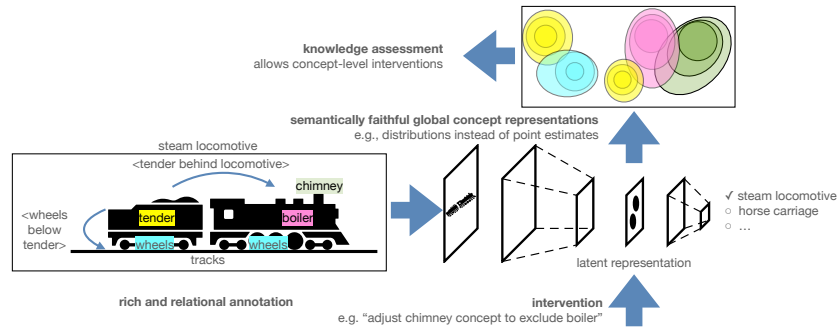


Fig. 1: Overview of envisaged methodology for model understanding and control. Using rich and relational concept annotations (e.g., grounded in an ontology) of visual model inputs, intuitive concepts and relations are associated with global, expressive, and semantically faithful concept representations in the model’s latent space (e.g., distributions). This allows interactive knowledge verification and local or global control, e.g., adjusting the concept representation to globally separate the concept **boiler** from the concept **chimney**.

decision [6,67]. However, these approaches do not explain what happens internally in the model. Concept-based XAI (C-XAI) [59,94,112] overcomes this shortcoming by explaining how a vision model represents input in its intermediate layers using semantically meaningful concepts that can be understood by users. Finding concept-based descriptions of internal representations is needed to gain more insight into the internal information processing of the model [99], since concepts can act as a Rosetta Stone, i.e., as a common alphabet between users and the model. Such concepts can be task-related objects (e.g., **head**, **beak**) or scene properties (e.g., **red**, **sunny**), and are not necessarily part of the output labels.

Goal and contributions. Our overall aim is to give XAI researchers a good starting point to dive into the subtopic of C-XAI and a guide to interesting next steps to advance the field further. Previous C-XAI surveys [59,94,99,112] have focused mainly on motivating and introducing C-XAI methods, focusing less on discussing the future of C-XAI research. In this paper, in addition to **reviewing the state of the art in C-XAI**, our objective is to draw attention to important challenges of C-XAI, which are largely neglected in the literature. We **discuss the state of the art and open challenges** in extracting new **concept types**, devising **concept representations** that go beyond the initial vector-based approach, and applying **concept control** mechanisms, where each **challenge** (a trophy icon followed by bold text) is accompanied by **proposals** (a light bulb icon followed by italicized text) on how to tackle it.

Our particular position in this paper is that C-XAI can benefit from the established field of knowledge representation and reasoning (KR) [15,46]. This includes verifying and controlling the **ontological commitment** [16] of a vision

model, that is, whether it has learned the right concepts and their relations to each other. Answering these questions empowers the pipeline in Fig. 1.

Scope. We will only briefly touch on the mature C-XAI directions for investigating concept relevance scores [41, 42, 56, 93, 127], and leave aside the promising applications of combination with feature importance methods [3, 83], and guidance for creating counterfactual explanations [82]. Similarly, with regard to challenges, we do not discuss in detail the already well-known issues of concept completeness [18, 108, 127], concept leakage [47, 49, 55, 73, 75, 76], lack of causality of concepts (fluffy & ear not necessarily implies the composite fluffy ear) [70], as well as cost and availability of concept labels [14, 85]. Instead, we identify and highlight the so far underexplored directions for potential advancements. Also, there is the question of how to evaluate the quality of C-XAI methods [29], which goes along with the general issues in XAI to define meaningful functionally-, human- and application-grounded metrics [25, 64, 113] and is not considered here.

In the next section, we will give a compact review of C-XAI and relevant background (Sec. 2), and then present our results on open challenges along the dimensions of concept type (Sec. 3), representation (Sec. 4), and control (Sec. 5).

2 Background on C-XAI for Computer Vision

Concept-based explainable AI seeks to enhance the interpretability of AI models by connecting their internal representations with human-understandable *concepts*. The definition of the concept used here varies across the literature.

2.1 What is a concept?

Poeta et al. [94] define concepts as “human-interpretable high-level features of the input data that are important for the model’s decision-making process”. This definition highlights the connection to features as used in feature importance methods. In semantic contexts, a concept is generally a notion that can be described using natural language [112], for example, a synonym set in the lexical database WordNet [4] (e.g., *ear*, *fluffy*) or a combination thereof (e.g., *fluffy ear*). This definition focuses on close alignment with human language and is the most commonly used definition in C-XAI [12, 35, 56]. We use this as the default notion in this paper. Extending this to image analysis, a concept can be considered as a meaningful region within an image [35, 39, 112]. So far, concepts considered are only vaguely interrelated and do not capture the rich structure of an ontological language that allows to define complex concepts from a set of basic ones.

In symbolic AI literature, particularly within KR, concepts are viewed as what can be modeled as logic predicates, characterized by their relations to other concepts [8, 44]. This structural approach emphasizes the logical relationships and hierarchy among concepts but is underexplored in C-XAI [114].

2.2 An Overview of C-XAI Directions for CV

In this section, we briefly review the state-of-the-art of C-XAI methods, to set the scene for later discussion (see Fig. 4 in Appendix A.1 for a taxonomy). For further details, the reader is referred to more elaborate surveys [59, 94, 99, 112].

C-XAI aims to associate the mentioned human-understandable concepts with a representation allocated in a neural model’s latent space(s), i.e., the intermediate output space of the model. We use the terms *concept representation*, or equivalently, *concept embeddings*, as a collective notion to encompass the variety of existing methods to represent concepts. In order to be understandable, a model must operate using the same conceptual “alphabet” as humans. Ideally, black-box models should also internally represent and use concepts that match those from the catalog of human cognition.

Concept representations can be *post-hoc* extracted (i.e., after training a model) or *ante-hoc* enforced (i.e., explainable by design) [59, 94, 99, 112]. We can also categorize C-XAI into *supervised* and *unsupervised* methods [112]: Supervised methods utilize pre-defined concept specifications, such as labeled concept examples, to check whether a neural model encodes information about a concept in question. Unsupervised methods instead aim to identify what concepts a model has learned; considerations here are what qualifies a representation as that of a concept, typically cluster centers [39, 132] or linear basis directions [132] (cf. Fig. 2); and how to ensure human interpretability of the found concepts, e.g., by constraining found concepts to be connected image regions [39]. In the following, we review existing variants for concept representation and discuss supervised and unsupervised C-XAI methods in detail.

Concept Representation Variants. A concept representation consists of two parts: the representation of the human-interpretable part (usually via examples [13], in vision-language models also via text [65, 91]) and the associated latent representation, which is typically given by the parameters of the function that associates a concept with its latent representation. For example, TCAV [56] defines a concept via images with binary classification labels, and the association function as a binary classifier of latent vectors. In its simplest form, a concept is associated with a single unit of the network (neuron [57] or filter [12] in a given layer). A more general perspective represents concepts by weight vectors with one weight per network unit of interest, taking the role of directions or centroids in latent space (see Fig. 2). Such techniques were shown to capture better the distributed way [21] of how information is stored in a model and were established in TCAV [56] and Net2Vec [35]. Since then, more complex representations include clusters [39, 96], and kernel functions [20]. It should be noted that nonlinear association functions are also investigated, such as generalized linear models [5, 74], or normalizing flows [30, 102]. However, this sacrifices the interpretability of the association [56]. The selection of the association function is task-specific, focusing on aspects such as the concept type [112] like spatial localization (e.g., image classification [56], segmentation [35]), and concept values (e.g., binary [56], regression [42], multi-

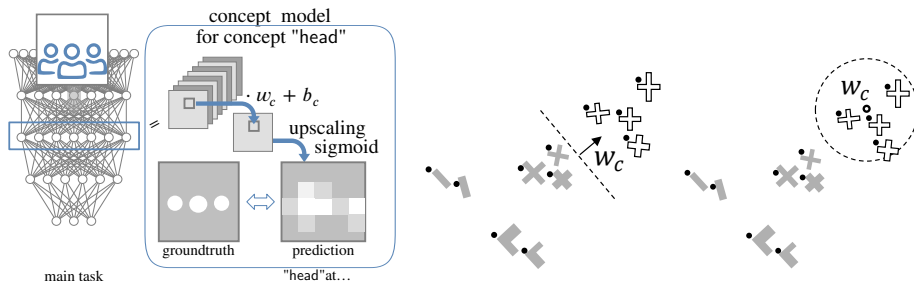


Fig. 2: Illustration of Net2Vec [35] for associating a concept with a linear separator with weight vector w_c in (activation pixel) latent space (*left*), and illustration of typical concept representation variants (*center*: direction-based, *right*: cluster-based).

class [54]); and constraints on the involved latent representations like being non-negative [132], unit vectors [12] or even a complete orthogonal basis [18, 57, 128].

Following an initial [56] and still prevalent [41, 42, 93, 127] application of C-XAI, some authors also demand as part of the concept representation an importance score [96]. This score tells how much the concept participates in the model’s decision process [95, 96], which is similar to feature importance [10].

Supervised Concept Analysis. Supervised concept embedding analysis methods associate predefined concepts with the units of the neural model. First approaches matched concept segmentations to the most similarly activated convolutional neural network (CNN) filters [12]. Fong et al. in Net2Vec [35] and Kim et al. in TCAV [56] soon after trained linear models, for concept segmentation and concept classification respectively, to separate concept from non-concept activations, with their weight vector serving as concept embedding vector. This is up to now the basis for essentially all post-hoc supervised techniques: Their linear models were extended to linear regression [42, 43], kernel-based methods producing region-based concepts [20], and from global to image-local explanations by training on concept data subsets [81, 130].

By contrast, ante-hoc (or explainable-by-design) approaches typically use the simple representation again and associate single units in a layer with concepts. They were first introduced as concept bottleneck models (CBMs) [57, 69]. This was later improved by denoising techniques to model concept interdependencies [11, 47], semisupervised training strategies for label efficiency [14, 85], concept hierarchies [77], binary [47] and multidimensional [29] concept representations; and combined with unsupervised methods [108] to overcome the well-known challenge of choosing a complete set of concepts, that is, one sufficient for the task [18, 108, 127]. Furthermore, CBMs are criticized for concept-leakage [49, 55, 73, 75, 76]: The vector produced by all concept neurons may learn to encode not only information about the given concepts but also “leak” other information to achieve higher accuracy.

Unsupervised Concept Analysis. Unsupervised concept analysis methods identify the most important concepts in the feature space without labeling information. They achieve concept completeness by design, but at the cost of possibly uninterpretable concept formations. The manual labeling effort for assigning labels to the found concepts is still necessary to finally establish the concept association. Techniques to identify prevalent features include standard clustering of activations obtained from a probing dataset, as first done for image-level concepts [38, 39]; and via (multi-layer) activation pixel clustering for image-region concepts [95, 96]. This was shown to be subsumed by matrix factorization techniques such as k-means clustering, classical PCA, or non-negative matrix factorization [33, 34, 60, 119, 132],

There also exist ante-hoc methods that, similar to CBMs, have a bottleneck layer. Instead of assigning one neuron per concept, they learn to encode concepts as prototype vectors. Comparing these with the intermediate representations produces the concept scores. This case-based reasoning approach was first introduced in ProtoPNet [17, 63] and continued in its successors on object detection [32], with prototype sharing across output classes [104], and with preferable cosine similarity instead of L_2 distance for prototype comparison [122].

2.3 Ontological Commitment in Knowledge Representations

We will here show how the notion of ontological commitment from the field of knowledge representations naturally translates to C-XAI requirements, which are further elaborated later. For everyday concepts, humans typically have an understanding of a concept based on *what other concepts are related and via which relations*. To connect this to neural models, note that also the model’s internals are supposed to be a (learned) knowledge representation. Thus, both concepts and their relations induce *constraints* on valid model (intermediate) predictions [114]. For example, consider object existence constraints from object-to-part relations [40, 114]: since the **head** is part of a **person**, the presence of a **head** should imply the presence of a **person**. Similarly for hierarchical class subsumption [103]: since a **human** is a **movable object**, the detection of a **person** implies it may be **movable**. Aside from these constraints, we also expect explanations to be more intuitive for humans to understand if they use humans’ cognitive catalog of concepts and relations (also known as cognitive chunks [25]). Therefore, an implicit requirement for concept representations is that they support reasoning with concepts and capture human prior knowledge about the task. The respective kind of reasoning is determined by the so-called *ontological commitment* [23]. Ontological commitment refers to the catalog of defined *concepts* (1-ary logic predicates) as well as *relations* (binary, possibly n-ary predicates), for example, `IsSimilarTo`, `IsSubclassOf`, `IsPartOf`, `IsCloseTo`. The term *commitment* signals the choice of admissible concepts, and significantly influences what kinds of inferences are possible or easy. For example, if **dog** and **cat** are organized as subconcepts of concept **pet**, then their co-occurrence with humans is easier to predict than choosing a zoology-motivated taxonomy (cf. Fig. 3b).

3 Types of Concepts

At the heart of the problem definition in C-XAI lies the questions of what concepts to extract and where to extract them from. In the visual domain, multiple concept types have already been considered [112]: image-level scene attributes (e.g., `sunny`) [12] and image qualities (e.g., `contrast`) [1]; as well as attributes of image regions such as object (e.g., `person`) and object part classes (e.g., `beak`) [12, 57], and object attributes such as material, texture [56], and color [109].

Apart from a few exceptions [77], the concepts are based on layered neural networks or spatial alignment in unimodal CNNs. Hence, post-hoc C-XAI has in CV so far been applied to classifiers [35, 56], regression [43], object detectors [80, 111], and only recently for the first time to video models [52, 106]; applications to language models [133], generative adversarial networks [13], and quite recently to diffusion models [36, 51] suggest that more architectures could be covered. An example is the Vision Transformer (ViT), which has recently become a popular CV architecture. Its self-attention mechanism, however, is difficult to interpret. Rigotti et al. [101] propose the Concept-Transformer, which extends attention from low-level features to high-level concepts, providing plausible and faithful explanations.

Open Challenges

C-XAI Research so far seems to be limited to the mentioned static attributes of images and image regions that are extracted from CNNs. This neglects concepts arising from *temporal* or *other sensory features*, as well as *other architectures* such as ViTs [26]. Since they could take important roles in future critical applications, we argue that more research is needed on concept extraction in these fields.

Temporal and Multimodal Concepts. What remains largely unexplored is **the identification of concepts for temporal and other sensory features**, despite being of interest for many important robotic applications like automated driving. These have an inherent temporal and multisensory resolution, which is inevitable for reliable prediction of trajectories, e.g., to differentiate advertisements on trucks from true pedestrians. Meanwhile, research of C-XAI in videos is very sparse, with the first TCAV-based work still concentrating on objects instead of movement patterns as concepts [52, 106]. Similarly, the investigation of C-XAI in multimodal models has just started, but with a focus on vision-language models to utilize the language input for concept definition [65, 91]. *Since it is possible to disentangle multimodal representations into single-modal ones [72], this might be an attack point for the transferral of C-XAI techniques to multimodal non-language models.* Generally, it would be interesting to see how and what concepts are represented in video and/or multimodal models, in order to enable in-depth debugging. Furthermore, **the so-far unused temporal resolution in spatio-temporal concepts might open up new ways of self-supervised concept**

extraction. It is well known that motion cues such as optical flow arising from temporal consistency in real-world videos are valuable information for object segmentation [124, 126]. This has, to our knowledge, not yet been used to analyze trained latent representations of video processing models. *Latent representations that occur in spatio-temporal regions with stable optical flow, such as on a moving object, might be interpreted as learned object properties.* That would, for example, allow self-supervised part-object extraction, to validate whether force exertion (e.g., locomotive tugging tender Fig. 1, swarm-like behavior), connectedness (e.g., arms typically do not detach), or even shadows on 3D objects are adequately modeled.

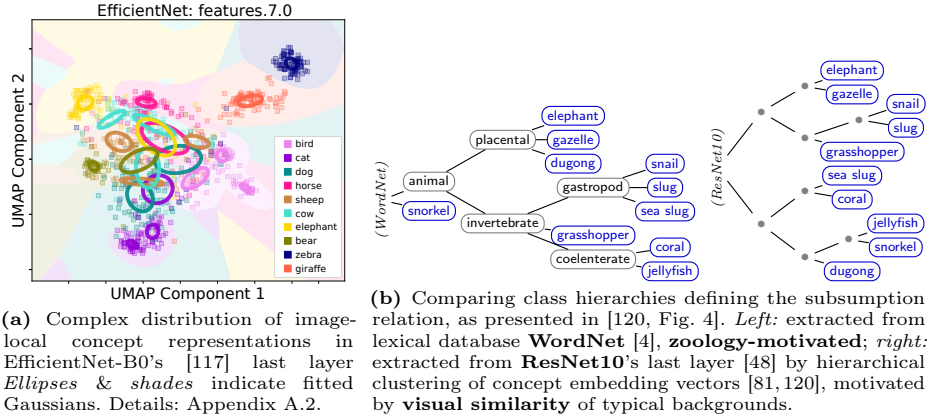
Concepts in New Architectures. As reviewed above, it is not yet clear **how to associate concepts in new architectures.** ViTs, for example, break with the direct association of neurons with spatial locations in the input. This, however, is utilized in nearly all C-XAI methods for extraction of subimage concepts: A concept in a spatial position must be reflected in the activation spatially aligned to that position, as already done in the base C-XAI methods [32, 35, 69, 132]. An alternative would be to use full-layer concept vectors, and during inference allocate them to individual image input regions by feature importance techniques, as done in [71] but with mediocre success. *A combination that leverages the coarse patch-wise processing of vision transformers together with feature attribution methods may be a promising direction.* Stassin et al. [116] discuss adapting existing XAI techniques to Transformers by converting embeddings into pseudo-activation maps, with a particular interest in applying this approach to the MLP layers. Another approach is *training a sparse autoencoder* on the activations of a layer, which is so far used in understanding large language models [50] and could be transferred to the vision domain. Similarly to ViTs, explaining diffusion models for image generation has only just sparked interest, both ante-hoc [51] as well as post-hoc [36, 92], although diffusion models are already being used in turn for concept discovery [118]. That could be an entry point for the diffusion model analysis. In summary, we see many opportunities to advance our understanding of novel model types via C-XAI.

4 Concept Representation

Concept representation encompasses two directions: How a specific concept is represented and which concepts are represented.

4.1 Basic Types of Concept Representations.

Using single neurons (i.e., unit vectors in latent space) as concepts [57] makes it easy to quantify concept attribution via neurons’ activation magnitude. However, this can be overly simplistic [78] because it overlooks the distributed nature of neural network representations [21, 35, 56]. Standard now are vector-based concept representations that hold weights for each neuron [39, 56] or filter [35, 81, 132] of



(a) Complex distribution of image-local concept representations in EfficientNet-B0’s [117] last layer *Ellipses & shades* indicate fitted Gaussians. Details: Appendix A.2.

(b) Comparing class hierarchies defining the subsumption relation, as presented in [120, Fig. 4]. *Left*: extracted from lexical database **WordNet** [4], **zoology-motivated**; *right*: extracted from **ResNet10**’s last layer [48] by hierarchical clustering of concept embedding vectors [81, 120], motivated by **visual similarity** of typical backgrounds.

Fig. 3: Illustration of the ontological commitment (Fig. 3b, *right*), and complex concept distribution (Fig. 3a, *left*) in actual vision model’s latent spaces.

one or several [96] layers. They require optimization but provide more accurate concept embeddings [35, 56]. So far, only a few exceptions generalize this from global point estimates to (nonlinear) subspaces [30], latent space regions [20], or hierarchies of (local or global) point estimates [77, 81, 120]. In the following subsection, we will argue the immediate shortcomings of the currently prevalent vector-based representations.

4.2 Ontological Commitment of Concept Representations.

The available background commonsense knowledge regarding concept definitions (e.g., `IsPartOf(head, person)`) is essential for pinning down semantics. Manually crafted, large ontologies often aim to capture the ontological commitment of human common sense. Notable examples are WordNet [4], Cyc [62], SUMO [84], or ConceptNet [115]. To connect these sources of information to C-XAI one has to ground ontology concepts in network activation. As a first step, individual concepts are grounded in network activation, but it is desirable to extend this approach to capture more expressive ontological languages, e.g., [88]. With respect to grounding individual concepts, recall that, e.g., in TCAV [56] and Net2Vec [35] the cosine similarity was used as a measurement for semantic similarity of latent concept representations, and vector addition as semantic combination of concepts (a kind of logical AND). This can now be considered as relations in the ontological language of their chosen vector-based concept representation. Probing what concept representation is a combination of others (e.g., `wood + green ≈ tree` [35]) or is similar to others (e.g., `brown hair ~ black hair` [56]) extracts the constraints and hence the ontological commitment of what the model has learned. This commitment, however, does not necessarily coincide with human intuition but can encode unwanted biases like `apron ~ female`. Therefore, the important goal

of C-XAI to uncover faulty learned knowledge reformulates as to *verify, validate, and control the ontological commitment and conceptualization of vision models*.

Unfortunately, little investigation has been devoted so far to the ontological commitment of types of concept representations. Vanilla (point-estimate) vector embeddings allow measurement of semantic similarity via cosine distance [35] but are criticized for their inability to model richer concept relations [45, 123]. Spatial calculi [110] become applicable to concept segmentations for modeling object-part-relations on image regions [114] or region-based concept representations (instead of point estimates) to model subsumption relations [81, 120] (extractable via hierarchical clustering, cf. Figs. 3b, 3a). Donadello et al. [24] showed that neural networks are also capable of learning more complex relations. This is in accordance with the findings that deep neural networks employ simple reasoning steps on concepts across several layers, the subnetworks encoding these also called circuits [86].

Open Challenges

We will now first argue, why the prevalent vector-based representations fall short of capturing some basic interesting information about concepts. This is then extended to the perspective of ontological commitment, where proposals are made to find richer relations between concept representations for better model validation and verification.

Questioning Point Estimates as Concept Representations. A challenge posed by the prevalent vector-based approaches is their two inherent assumptions that we will question in the following: (i) Concepts can meaningfully be approximated by *linear* trajectories in latent space pointing from *less concept* to *more concept* [56, 90], and (ii) this direction can be expressed by a point estimate.

🔥 **Linear point estimated representations are too simplistic, concepts should be modeled by regions or distributions.** This can be argued from two perspectives. For one, while point estimates might be sufficient for small models and datasets with few clearly distinct concepts [57], Mikriukov et al. [80] showed that this can break down at scale, as illustrated in Fig. 3a: Concepts in larger object detectors are smeared over the latent space at different densities, start overlapping, and even break down into distinct subconcepts. Such relevant information cannot be captured by point estimates. Instead, region-based [20, 89, 95] or density-based [81] approaches can capture spread (or even density and thus outliers), non-connectedness (i.e., sub-concepts), and overlap (i.e., concept confusion or concept commonalities) of concepts in the model’s latent spaces. 🍷 *Future research could involve generalizing local C-XAI approaches [81, 130], fitting Gaussian mixture models to sets of such local concept vectors* and investigating factors that influence concept spread. Furthermore, 🍷 *the region-based approach poses an interesting direction*. For example, representing concepts as *cones* could be promising, as they naturally come with negation, intersection (**AND**) and union (**OR**) on concepts [61, 88], as picked up again below.

As a second argument, we would like to draw attention to 🚩 **modeling the rate of change**: So far, C-XAI-based approaches only considered the general direction towards more concept regions but not the rate of change when traversing the trajectory. It might, however, be interesting information whether the model assumes a rapid change (turning point) like one would expect for glasses versus broken glasses; a somewhat smooth transition, like non-smiling to smiling [30]; or a truly linear change of an object’s representation in latent space when continuously modifying rotation or color. And lastly, it is not taken for granted that local approximation by a trajectory with 1D curvature (i.e., a straight line) is strong enough to capture all concept information of interest. Several approaches now discard this assumption by using the highly non-straight trajectories of traversing concepts in generative model latent spaces [30, 118]. Validating and addressing the linearity assumption is essential for developing more flexible, potentially non-linear concept representations that accurately capture the complexities of real-world data.

Richer Ontological Commitment of Concept Representations. So far, C-XAI research has mostly focused on the ability of a model to grasp a concept as intended. Unfortunately, little work beyond this is devoted to 🚩 **systematic investigation of the ontological commitment in trained models, in particular the relations that they can model**. In the context of knowledge embeddings, on the other hand, there exist principled embedding approaches that capture rich concept relations, including subsumption [45, 87, 123], or negation [88].

In consideration of models that exhibit reasoning capabilities, further requirements may arise. This has already been shown for the usual approach of geometric containment in latent space to represent concept subsumption, i.e., points inside a region represent the instances of some concept. In order to allow reasoning, concept regions cannot be arbitrarily shaped [61]. Put differently, the geometry of concepts and their relations in latent space is tightly coupled with the reasoning capabilities that can be achieved. Little of geometry-reasoning interdependency has been revealed so far. A promising direction to solve this issue is to 🍷 *investigate whether vision models use some of the known principled embedding approaches such as from spatial reasoning [28, 46, 88] and how to extend existing C-XAI approaches to extract these.*

Another open challenge is to 🚩 **develop tools for verification of a model’s ontological commitments**. Options for achieving this are to (a) find accurate representations of known relations in the model, or (b) verify a given relation function commitment (like the cosine similarity) against expected behavior. An approach to the first challenge could be 🍷 *considering so-called reification of relations [87], an idea from knowledge graph embeddings that flexibly represents relations themselves as concepts*. For the second challenge, 🍷 *both the rich common sense ontologies should be taken into account, complemented by densely labeled visual datasets labeling both concepts and local relations*, similar to scene graph datasets [58]. Note that this might require considerable efforts in the community to 🍷 *define task-specific sub-ontologies, and to develop more specialized and con-*

trollable datasets and testing environments, such as 3D-generated scenes with automatic annotations [97,98] or generative AI-produced data. Apart from that, understanding what relations a deep neural network of given depth can accurately model, and investigating whether it does model the relations of interest, are an important step for fully understanding the automatic reasoning applied by the model.

5 Concept Control

Concept-based explanations not only allow us to understand a model, but also provide us with means to change the model to achieve a specific objective, e.g., improving the generalizability of the model. Among several ways of changing a model (e.g., improving the quality of the training set, choosing a better model architecture, or directly modifying the intermediate representations of an input), we see high potential in targeted modification of the intermediate representations of model inputs, which we refer to as *concept control* or *concept intervention*.

5.1 Modification of Latent Representations

In the C-XAI literature, concepts are often represented as vectors in an embedding space [2, 56, 85, 129]. Given a set of concepts $\mathcal{C} = \{c_1, \dots, c_n\}$, we can regard the corresponding concept vectors $w_{c_1}, \dots, w_{c_n} \in \mathbb{R}^d$ as a generating set or, by abusing the language, a basis of interpretable linear subspace of the embedding space \mathbb{R}^d . The i^{th} coordinate of a representation with respect to that *concept basis* captures the strength of the presence of concept c_i in the representation. This allows us to intervene on a concept c_i in an intuitive way: increasing (or decreasing) the i^{th} coordinate leads to increasing (or decreasing) the presence of concept c_i in the representation. Koh et al. [57] were to our knowledge the first to apply a concept intervention. This was done on the CBM architecture, where a complete layer is trained such that each single neuron, which corresponds to a unit vector in the embedding space, is associated with a given concept in an ante-hoc supervised manner. An issue here is the need for ground-truth concept labels, which are often unavailable.

Several approaches circumvent the above issue by interpreting the concept vectors in a post-hoc manner. Abid et al. [2] and Yuksekogunul et al. [129] use the CAVs of a pre-trained vision model as a concept basis and identify concepts that need to be added (or strengthened) or removed (or suppressed). To apply local interventions, Abid et al. learn counterfactual explanations, that is, identifying the concepts that should have been added or removed so that the model predicts the correct label. In contrast to that, the post-hoc CBM approach [129] intervenes globally by removing a spurious concept to predict a class, e.g., removing the concept **dog** to predict the concept **table** in the test set when **dog** is spuriously correlated with **table** in the training set, but not in the test set. Further methods include the editing of classifiers [107] that can control the behavior of an image classifier by using only a single example, P-CLArC [7], which projects out concept

directions, and RR-ClArC [27], which regularizes CAVs during training to guide the model to become less reliant on biases. Similarly to CBMs, post-hoc C-XAI methods require a predefined set \mathcal{C} of concepts.

Open Challenges

We identify three underexplored areas: imposing logic constraint, application of concept control, and mitigation of side effects which are explained below.

Imposing Logical Constraints. Logical constraints can be imposed on concepts and allow for tight neurosymbolic integration [24, 59, 114]. Such logical constraints can be used to guarantee the consistency of the model’s reasoning [114] and to align the model with a knowledge base or with different criteria by the user [24] (e.g., criteria related to ethics, privacy, and safety). C-XAI brings in the benefit that one can directly act on concepts in the embedding space of intermediate layers of a pretrained model instead of on the model’s output (e.g., removing skin color, which is usually not an output label, from an intermediate layer for fairness assessment). Controlling intermediate representations promises to achieve higher coverage of concepts and performance and greater flexibility.

We highlight two future challenges on how such constraints can be enforced: (i) by 🍌 **guiding the model training** and (ii) 🍌 **globally modifying intermediate representations**. For the first challenge, one can use a 🍷 *multi-task training routine that simultaneously or alternately updates the concept models* to maintain a correct association of the concepts, and updates the model parameters according to both the main task and the constraints on the concepts. Constraints may be formulated and approximated using regularization terms, as proposed in the semantic loss formulation in [9, 125]. In contrast to this “soft” approach to model weights for the first challenge, the second challenge can be tackled by inserting intermediate processing steps that will modify the intermediate input representations to comply with the constraints, for example, by 🍷 *linear projection or linear skew*. A proof of concept is shown in [100], but this was on simplistic unit-vector concepts. Generally, 🍌 **it remains to be shown that logical constraints can be applied to diverse image datasets and with varying expressivity of the logical constraints**, for example, allowing relations or functions in addition to concepts in the logical constraints.

Note that the logical constraints imposed on the model need to be compatible with the rules that can be extracted from the same model (cf. Sec. 4). More precisely, depending on the expressivity [114] of the logic language used to extract the rules from the model, logical constraints of high or lower expressive power can be imposed on the model. Therefore, investigating the reasoning of the model can affect concept control.

Applications of Concept Control. The motivation for concept control can come from various applications, such as model editing and debugging, increasing the robustness of models against adversarial attacks or distribution shifts [66], or

🔥 **avoiding catastrophic forgetting (i.e., retaining previously learned knowledge) in new tasks in a lifelong learning scenario** [121]. For example, a self-driving car that was trained to recognize humans based on specific clothes may fail in areas with a different climate or culture. Although model editing and debugging are performed on a model after training, catastrophic forgetting can already be mitigated at training time by regularizing by 🍷 *penalizing deviation of the model’s ontological commitment across different tasks*.

Evaluating C-XAI methods is still an open problem [94]. Approaching C-XAI from the perspective of concept control with 🍷 *concrete objectives in applications, such as model correction [27], can be an effective way to evaluate C-XAI methods*. Therefore, 🔥 **identifying what potential applications can benefit from concept control and comprehensively evaluating the controllability of C-XAI methods in such applications** can be beneficial for the C-XAI research.

Mitigating Side Effects. Concepts can depend on each other; for example, in most cases the concept `car` co-occurs in an image with `wheel` which again includes the concept `round`. Thus, modifying a concept globally can affect other concepts, for example, replacing `round` with `rectangular` affects concepts `wheel` and `car`. This side effect, which is also known as the ripple effect in the language model editing literature [19], can be a big impediment to controlling concepts. 🔥 **Identifying the side effects of a specific concept control mechanism and avoiding the side effects** are therefore important open challenges. 🍷 *Inspirations from C-XAI approaches to natural language processing* (e.g., [19]) could be a starting point for next steps.

6 Conclusion

In this paper, we have examined the current state and open challenges in C-XAI for CV, focusing on concept types, expressive representations, and use of control. We identified three currently underexplored areas with high potential to advance the field:

- (i) **Expand the types of concepts** that can be extracted and analyzed to temporal ones, as well as recent model architectures like ViTs.
- (ii) Inspired by knowledge representation, develop **richer concept representations** that go beyond simple point-estimate vector embeddings and capture the complexity and relations of concepts learned by CV models.
- (iii) On the application side, improve the techniques for concept control by **imposing logical constraints** directly on the model’s internal representations.

Addressing these challenges, C-XAI methods can provide deeper insights into the inner workings of vision models and enable more fine-grained, interactive control over their behavior. This will be crucial for the verification and maintainability of critical CV applications. We hope to have provided a good starting point for researchers new to the field, as well as helpful inspiration for the community to advance it further.

Acknowledgements

Jae Hee Lee and Stefan Wermter gratefully acknowledge support from the German Research Foundation DFG for the project CML TRR169.

References

1. Abid, A., Yuksekgonul, M., Zou, J.: Meaningfully debugging model mistakes using conceptual counterfactual explanations. In: Proc. 39th Int. Conf. Machine Learning. pp. 66–88. PMLR (2022)
2. Abid, A., Yuksekgonul, M., Zou, J.: Meaningfully debugging model mistakes using conceptual counterfactual explanations. In: Proc. 39th International Conference on Machine Learning. pp. 66–88. PMLR (2022)
3. Achtabat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., Lapuschkin, S.: From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence* **5**(9), 1006–1019 (2023)
4. Al-Halimi, R., Berwick, R.C., Burg, J.F.M., Chodorow, M., Fellbaum, C., Grabowski, J., Harabagiu, S., Hearst, M.A., Jones, D.A., Kazman, R., Kohl, K.T., Landes, S., Leacock, C., Miller, G.A., Miller, K.J., Moldovan, D., Nomura, N., Priss, U., Resnik, P., St-Onge, D., Teng, R., van de Riet, R.P., Voorhees, E.: *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication, A Bradford Book (1998)
5. Alvarez-Melis, D., Jaakkola, T.: Towards Robust Interpretability with Self-Explaining Neural Networks. In: *Advances in Neural Information Processing Systems*. vol. 31. Curran Associates, Inc. (2018)
6. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Gradient-Based Attribution Methods. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 169–191. Lecture Notes in Computer Science, Springer International Publishing (2019)
7. Anders, C.J., Weber, L., Neumann, D., Samek, W., Müller, K.R., Lapuschkin, S.: Finding and removing Clever Hans: Using explanation methods to debug and improve deep models. *Information Fusion* **77**, 261–295 (2022)
8. Baader, F.: Description logics. In: *Reasoning Web: Semantic Technologies for Information Systems*, 5th International Summer School 2009. LNCS, vol. 5689, pp. 1–39. Springer-Verlag (2009)
9. Badreddine, S., d’Avila Garcez, A., Serafini, L., Spranger, M.: Logic Tensor Networks. *Artificial Intelligence* **303**, 103649 (2022)
10. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.R.: How to explain individual classification decisions. *Journal of Machine Learning Research* **11**, 1803–1831 (2010)
11. Bahadori, M.T., Heckerman, D.: Debiasing concept-based explanations with causal analysis. In: *Posters of the 2021 Int. Conf. Learning Representations* (2020)
12. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: *Proc. IEEE conference on computer vision and pattern recognition*. pp. 6541–6549 (2017)
13. Bau, D., Zhu, J.Y., Strobel, H., Zhou, B., Tenenbaum, J.B., Freeman, W.T., Torralba, A.: GAN dissection: Visualizing and understanding generative adversarial networks. In: *Posters 2021 Int. Conf. Learning Representations* (2018)

14. Belém, C., Balayan, V., Saleiro, P., Bizarro, P.: Weakly Supervised Multi-task Learning for Concept-based Explainability. arXiv preprint arXiv:2104.12459 (2021)
15. Brachman, R., Levesque, D.H.: Knowledge Representation and Reasoning. Morgan Kaufmann (2014)
16. Bricker, P.: Ontological Commitment. In: The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, winter 2016 edn. (2016)
17. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.: This looks like that: Deep learning for interpretable image recognition. In: Advances in Neural Information Processing Systems 32. vol. 32, pp. 8928–8939 (2019)
18. Chen, Z., Bei, Y., Rudin, C.: Concept whitening for interpretable image recognition. *Nature Machine Intelligence* **2**, 772–782 (2020)
19. Cohen, R., Biran, E., Yoran, O., Globerson, A., Geva, M.: Evaluating the Ripple Effects of Knowledge Editing in Language Models. *Transactions of the Association for Computational Linguistics* **12**, 283–298 (2024)
20. Crabbé, J., van der Schaar, M.: Concept activation regions: A generalized framework for concept-based explanations. In: Advances in Neural Information Processing Systems. vol. 35, pp. 2590–2607 (2022)
21. Craven, M.W., Shavlik, J.: Visualizing learning and computation in artificial neural networks. *Int. J. Artif. Intell. Tools* (1992)
22. Das, A., Rad, P.: Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. arXiv preprint arXiv:2006.11371 (2020)
23. Davis, R., Shrobe, H., Szolovits, P.: What is a knowledge representation? *AI Magazine* **14**(1) (1993)
24. Donadello, I., Serafini, L., d’Avila Garcez, A.S.: Logic tensor networks for semantic image interpretation. In: Proc. 26th Int. Joint Conf. Artificial Intelligence. pp. 1596–1602. ijcai.org (2017)
25. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
26. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (2020)
27. Dreyer, M., Pahde, F., Anders, C.J., Samek, W., Lapuschkin, S.: From Hope to Safety: Unlearning Biases of Deep Models via Gradient Penalization in Latent Space. *Proceedings of the AAAI Conference on Artificial Intelligence* **38**(19), 21046–21054 (2024)
28. Dylla, F., Lee, J.H., Mossakowski, T., Schneider, T., Delden, A.V., Ven, J.V.D., Wolter, D.: A Survey of Qualitative Spatial and Temporal Calculi: Algebraic and Computational Properties. *ACM Comput. Surv.* **50**(1), 7:1–7:39 (2017)
29. Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lió, P., Jamnik, M.: Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off. *Advances in Neural Information Processing Systems* **35**, 21400–21413 (Dec 2022)
30. Esser, P., Rombach, R., Ommer, B.: A disentangling invertible interpretation network for explaining latent representations. In: Proc. 2020 IEEE Conf. Comput. Vision and Pattern Recognition. pp. 9220–9229. IEEE (2020)
31. European Commission: Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts (2021)

32. Feifel, P., Bonarens, F., Koster, F.: Reevaluating the safety impact of inherent interpretability on deep neural networks for pedestrian detection. In: Proc. 2021 IEEE/CVF Conf. Comput. Vision and Pattern Recognition. pp. 29–37 (2021)
33. Fel, T., Boutin, V., Béthune, L., Cadene, R., Moayeri, M., Andéol, L., Chalvidal, M., Serre, T.: A Holistic Approach to Unifying Automatic Concept Extraction and Concept Importance Estimation. *Advances in Neural Information Processing Systems* **36**, 54805–54818 (2023)
34. Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., Serre, T.: CRAFT: Concept recursive activation factorization for explainability. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2711–2721 (2023)
35. Fong, R., Vedaldi, A.: Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In: Proc. 2018 IEEE Conf. Comput. Vision and Pattern Recognition. pp. 8730–8738. IEEE Computer Society (2018)
36. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified Concept Editing in Diffusion Models. In: Proc. IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5111–5120 (2024)
37. d’Avila Garcez, A., Lamb, L.C.: Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review* **56**(11), 12387–12406 (2023)
38. Ge, Y., Xiao, Y., Xu, Z., Zheng, M., Karanam, S., Chen, T., Itti, L., Wu, Z.: A peek into the reasoning of neural networks: Interpreting with structural visual concepts. In: Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition. pp. 2195–2204 (2021)
39. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. *Advances in neural information processing systems* **32** (2019)
40. Giunchiglia, E., Stoian, M., Khan, S., Cuzzolin, F., Lukasiewicz, T.: ROAD-R: The Autonomous Driving Dataset with Logical Requirements. In: *IJCLR 2022 Workshops* (2022)
41. Goyal, Y., Shalit, U., Kim, B.: Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165* **abs/1907.07165** (2019)
42. Graziani, M., Andrearczyk, V., Marchand-Maillet, S., Müller, H.: Concept attribution: Explaining CNN decisions to physicians. *Computers in Biology and Medicine* **123**, 103865 (2020)
43. Graziani, M., Andrearczyk, V., Müller, H.: Regression concept vectors for bidirectional explanations in histopathology. In: *Understanding and Interpreting Machine Learning in Medical Image Computing Applications, first International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018*. pp. 124–132. *Lecture Notes in Computer Science*, Springer International Publishing (2018)
44. Guarino, N.: *Formal Ontologies and Information Systems*. In: Proc. FOIS’98. pp. 3–15. IOS Press (1998)
45. Gutiérrez-Basulto, V., Schockaert, S.: From knowledge graph embedding to ontology embedding? An analysis of the compatibility between vector space representations and rules. In: *Principles of Knowledge Representation and Reasoning: Proc. Sixteenth International Conference, KR 2018, Tempe, Arizona, 30 October – 2 November 2018*. pp. 379–388. AAAI Press (2018)
46. van Harmelen, F., Lifschitz, V., Porter, B., et al.: *Handbook of Knowledge Representation. Foundations of Artificial Intelligence*, Elsevier, 1 edn. (2007)
47. Havasi, M., Parbhoo, S., Doshi-Velez, F.: Addressing leakage in concept bottleneck models. In: Proc. 36th International Conference on Neural Information Processing Systems. NIPS ’22, Curran Associates Inc., Red Hook, NY, USA (2024)

48. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. 2016 IEEE Conf. Comput. Vision and Pattern Recognition. pp. 770–778 (2016)
49. Hoffmann, A., Fanconi, C., Rade, R., Kohler, J.: This looks like that... does it? shortcomings of latent space prototype interpretability in deep networks. arXiv preprint arXiv:2105.02968 (2021)
50. Huben, R., Cunningham, H., Smith, L.R., Ewart, A., Sharkey, L.: Sparse Autoencoders Find Highly Interpretable Features in Language Models. In: The Twelfth International Conference on Learning Representations (2024)
51. Ismail, A.A., Adebayo, J., Bravo, H.C., Ra, S., Cho, K.: Concept Bottleneck Generative Models. In: The Twelfth International Conference on Learning Representations (Oct 2023)
52. Ji, Y., Wang, Y., Kato, J.: Spatial-Temporal Concept Based Explanation of 3D ConvNets. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15444–15453 (2023)
53. Kaur, D., Uslu, S., Rittichier, K.J., Durrezi, A.: Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys* **55**(2), 39:1–39:38 (2022)
54. Kazhdan, D., Dimanov, B., Jamnik, M., Liò, P., Weller, A.: Now you see me (CME): Concept-based model extraction. In: Proc. 29th ACM Int. Conf. Information and Knowledge Management Workshops. CEUR Workshop Proceedings, vol. 2699. CEUR-WS.org (2020)
55. Kazhdan, D., Dimanov, B., Terre, H.A., Jamnik, M., Liò, P., Weller, A.: Is disentanglement all you need? comparing concept-based & disentanglement approaches. arXiv preprint arXiv:2104.06917 (2021)
56. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)
57. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
58. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* **123**(1), 32–73 (2017)
59. Lee, J.H., Lanza, S., Wermter, S.: From neural activations to concepts: A survey on explaining concepts in neural networks. arXiv preprint arXiv:2310.11884 (2024)
60. Leemann, T., Kirchhof, M., Rong, Y., Kasneci, E., Kasneci, G.: When are post-hoc conceptual explanations identifiable? In: Proc. Thirty-Ninth Conference on Uncertainty in Artificial Intelligence. pp. 1207–1218. PMLR (2023)
61. Leemhuis, M., Özçep, Ö.L.: Conceptual orthospaces—convexity meets negation. *International Journal of Approximate Reasoning* **162**, 109013 (2023)
62. Lenat, D.B.: Building Large Knowledge-Based Systems. Addison-Wesley Pub. Co. (1989)
63. Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: Proc. Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. pp. 3530–3537. AAAI’18/IAAI’18/EAAI’18, AAAI Press (2018)

64. Li, X.H., Shi, Y., Li, H., Bai, W., Cao, C.C., Chen, L.: An experimental study of quantitative evaluations on saliency methods. In: KDD '21: Proc. 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 3200 – 3208 (2021)
65. Liang, P.P., Lyu, Y., Chhablani, G., Jain, N., Deng, Z., Wang, X., Morency, L.P., Salakhutdinov, R.: MultiViz: Towards Visualizing and Understanding Multimodal Models. In: The Eleventh International Conference on Learning Representations (2022)
66. Liang, W., Zou, J.: MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts and Training Conflicts. In: International Conference on Learning Representations (2022)
67. Liang, Y., Li, S., Yan, C., Li, M., Jiang, C.: Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing* **419**, 168–182 (2021)
68. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proc. 13th European Conf. Computer Vision - Part V. Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer International Publishing (2014)
69. Losch, M., Fritz, M., Schiele, B.: Interpretability beyond classification output: Semantic bottleneck networks. In: Proc. 3rd ACM Computer Science in Cars Symp. Extended Abstracts (2019)
70. Lovering, C., Pavlick, E.: Unit Testing for Concepts in Neural Networks. *Transactions of the Association for Computational Linguistics* **10**, 1193–1208 (Nov 2022)
71. Lucieri, A., Bajwa, M.N., Dengel, A., Ahmed, S.: Explaining AI-based decision support systems using concept localization maps. In: Neural Information Processing. pp. 185–193. Communications in Computer and Information Science, Springer International Publishing (2020)
72. Lyu, Y., Liang, P.P., Deng, Z., Salakhutdinov, R., Morency, L.P.: DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations. In: Proc. 2022 AAAI/ACM Conference on AI, Ethics, and Society. pp. 455–467. AIES '22, Association for Computing Machinery (2022)
73. Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., Pan, W.: Promises and pitfalls of black-box concept learning models. arXiv preprint arXiv:2106.13314 (2021)
74. Marcinkevičs, R., Vogt, J.E.: Interpretable Models for Granger Causality Using Self-explaining Neural Networks. In: International Conference on Learning Representations (2020)
75. Marconato, E., Passerini, A., Teso, S.: GlanceNets: Interpretable, Leak-proof Concept-based Models. In: Advances in Neural Information Processing Systems. vol. 35, pp. 21212–21227 (2022)
76. Marconato, E., Passerini, A., Teso, S.: Interpretability is in the mind of the beholder: A causal framework for human-interpretable representation learning. arXiv preprint arXiv:2309.07742 (2023)
77. Marcos, D., Fong, R., Lobry, S., Flamary, R., Courty, N., Tuia, D.: Contextual semantic interpretability. In: Proc. 15th Asian Conf. Comput. Vision Revised Selected Papers, Part IV. Lecture Notes in Computer Science, vol. 12625, pp. 351–368. Springer (2020)
78. Margeloiu, A., Ashman, M., Bhatt, U., Chen, Y., Jamnik, M., Weller, A.: Do concept bottleneck models learn as intended? In: Proc. of ICLR 2021: Workshop on Responsible AI (2021)

79. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2020)
80. Mikriukov, G., Schwalbe, G., Hellert, C., Bade, K.: Evaluating the stability of semantic concept representations in cnns for robust explainability. In: World Conference on Explainable Artificial Intelligence. pp. 499–524. Springer (2023)
81. Mikriukov, G., Schwalbe, G., Hellert, C., Bade, K.: Gcpv: Guided concept projection vectors for the explainable inspection of cnn feature spaces. arXiv preprint arXiv:2311.14435 (2023)
82. Motzkus, F., Hellert, C., Schmid, U.: Cola-dce – concept-guided latent diffusion counterfactual explanations. arXiv preprint arXiv:2406.01649 (2024)
83. Motzkus, F., Mikriukov, G., Hellert, C., Schmid, U.: Locally testing model detections for semantic global concepts. arXiv preprint arXiv:2405.17523 (2024)
84. Niles, I., Pease, A.: Towards a standard upper ontology. In: Proc. International Conference on Formal Ontology in Information Systems - Volume 2001. pp. 2–9. FOIS '01, Association for Computing Machinery (2001)
85. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free Concept Bottleneck Models. In: The Eleventh International Conference on Learning Representations (2023)
86. Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., Carter, S.: Zoom In: An Introduction to Circuits. Distill **5**(3), e00024.001 (2020)
87. Özçep, Ö.L., Leemhuis, M., Wolter, D.: Knowledge graph embeddings with ontologies: Reification for representing arbitrary relations. In: KI 2022: Advances in Artificial Intelligence, 45th German Conference on AI, Trier, Germany, September 19–23, 2022, Proceedings. pp. 146–159. Springer International Publishing (2022)
88. Özçep, Ö.L., Leemhuis, M., Wolter, D.: Embedding ontologies in the description logic ALC by axis-aligned cones. Journal of Artificial Intelligence Research (JAIR) **78**, 217–267 (2023)
89. Özçep, Ö.L., Leemhuis, M., Wolter, D.: Embedding Ontologies in the Description Logic ALC by Axis-Aligned Cones. Journal of Artificial Intelligence Research **78**, 217–267 (2023)
90. Pahde, F., Dreyer, M., Weber, L., Weckbecker, M., Anders, C.J., Wiegand, T., Samek, W., Lopuschkin, S.: Navigating neural space: Revisiting concept activation vectors to overcome directional divergence. arXiv preprint arXiv:2202.03482 (2024)
91. Parekh, J., Khayatan, P., Shukor, M., Newson, A., Cord, M.: A concept-based explainability framework for large multimodal models. arXiv preprint arXiv:2406.08074 (2024)
92. Park, J.H., Ju, Y.J., Lee, S.W.: Explaining generative diffusion models via visual analysis for interpretable decision-making process. Expert Systems with Applications **248**, 123231 (2024)
93. Pfau, J., Young, A.T., Wei, J., Wei, M.L., Keiser, M.J.: Robust Semantic Interpretability: Revisiting Concept Activation Vectors. arXiv preprint arXiv:2104.02768 (2021)
94. Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T., Baralis, E.: Concept-based explainable artificial intelligence: A survey. arXiv preprint arXiv:2312.12936 (2023)
95. Posada-Moreno, A.F., Müller, K., Brillowski, F., Solowjow, F., Gries, T., Trimpe, S.: Scalable concept extraction in industry 4.0. In: World Conference on Explainable Artificial Intelligence. pp. 512–535. Springer (2023)
96. Posada-Moreno, A.F., Surya, N., Trimpe, S.: ECLAD: Extracting Concepts with Local Aggregated Descriptors. Pattern Recognition **147**, 110146 (2024)

97. Raistrick, A., Lipson, L., Ma, Z., Mei, L., Wang, M., Zuo, Y., Kayan, K., Wen, H., Han, B., Wang, Y., Newell, A., Law, H., Goyal, A., Yang, K., Deng, J.: Infinite photorealistic worlds using procedural generation. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12630–12641 (2023)
98. Raistrick, A., Mei, L., Kayan, K., Yan, D., Zuo, Y., Han, B., Wen, H., Parakh, M., Alexandropoulos, S., Lipson, L., Ma, Z., Deng, J.: Infinigen indoors: Photorealistic indoor scenes using procedural generation. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21783–21794 (2024)
99. Räuker, T., Ho, A., Casper, S., Hadfield-Menell, D.: Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. In: 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). pp. 464–483 (2023)
100. Ribeiro, M.d.S., Leite, J.: On modifying a neural network’s perception. arXiv preprint arXiv:2303.02655 (2023)
101. Rigotti, M., Mikšović, C., Giurgiu, I., Gschwind, T., Scotton, P.: Attention-based interpretability with concept transformers. In: International conference on learning representations (2021)
102. Rombach, R., Esser, P., Ommer, B.: Making sense of CNNs: Interpreting deep representations and their invariances with INNs. In: Proc. 16th Europ. Conf. Computer Vision (ECCV 2020). Lecture Notes in Computer Science, vol. 12362, pp. 647–664. Springer (2020)
103. Roychowdhury, S., Diligenti, M., Gori, M.: Image classification using deep learning and prior knowledge. In: Workshops of the 32nd AAAI Conf. Artificial Intelligence. AAAI Workshops, vol. WS-18, pp. 336–343. AAAI Press (2018)
104. Rymarczyk, D., Struski, L., Tabor, J., Zieliński, B.: ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification. In: Proc. 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 1420–1430. KDD ’21, Association for Computing Machinery (2021)
105. Saeed, W., Omlin, C.: Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems* **263**, 110273 (2023)
106. Saha, A., Gupta, S., Ankireddy, S.K., Chahine, K., Ghosh, J.: Exploring explainability in video action recognition. arXiv preprint arXiv:2404.09067 (2024)
107. Santurkar, S., Tsipras, D., Elango, M., Bau, D., Torralba, A., Madry, A.: Editing a classifier by rewriting its prediction rules. In: Advances in Neural Information Processing Systems. vol. 34, pp. 23359–23373. Curran Associates, Inc. (2021)
108. Sawada, Y., Nakamura, K.: Concept Bottleneck Model With Additional Unsupervised Concepts. *IEEE Access* **10**, 41758–41765 (2022)
109. Schauerte, B.: Google-512, color term learning data set. <https://cvhci.anthropomatik.kit.edu/~bschauer/datasets/google-512/> (2010)
110. Schockaert, S., De Cock, M., Cornelis, C., Kerre, E.E.: Fuzzy region connection calculus: Representing vague topological information. *International Journal of Approximate Reasoning* **48**(1), 314–331 (2008)
111. Schwalbe, G.: Verification of size invariance in DNN activations using concept embeddings. In: Artificial Intelligence Applications and Innovations. pp. 374–386. IFIP Advances in Information and Communication Technology, Springer International Publishing (2021)
112. Schwalbe, G.: Concept embedding analysis: A review. arXiv preprint arXiv:2203.13909 (2022)
113. Schwalbe, G., Finzel, B.: A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery* (2023)

114. Schwalbe, G., Wirth, C., Schmid, U.: Enabling verification of deep neural networks in perception tasks using fuzzy logic and concept embeddings. *arXiv preprint arXiv:2201.00572* (2022)
115. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In: *Proc. of AAAI Conference on Artificial Intelligence* (2017)
116. Stassin, S., Corduant, V., Mahmoudi, S.A., Siebert, X.: Explainability and evaluation of vision transformers: An in-depth experimental study. *Electronics* **13**(1), 175 (2023)
117. Tan, M., Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: *Proc. 36th International Conference on Machine Learning*. pp. 6105–6114. PMLR (2019)
118. Varshney, P., Lucieri, A., Balada, C., Dengel, A., Ahmed, S.: Generating counterfactual trajectories with latent diffusion models for concept discovery. *arXiv preprint arXiv:2404.10356* (2024)
119. Vielhaben, J., Bluecher, S., Strodthoff, N.: Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees. *Transactions on Machine Learning Research* (2023)
120. Wan, A., Dunlap, L., Ho, D., Yin, J., Lee, S., Petryk, S., Bargal, S.A., Gonzalez, J.E.: NBDT: Neural-backed decision tree. In: *Posters 2021 Int. Conf. Learning Representations* (2020)
121. Wang, L., Zhang, X., Su, H., Zhu, J.: A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 1–20 (2024)
122. Willard, F., Moffett, L., Mokel, E., Donnelly, J., Guo, S., Yang, J., Kim, G., Barnett, A.J., Rudin, C.: This looks better than that: Better interpretable models with protopnext. *arXiv preprint arXiv:2406.14675* (2024)
123. Xiong, B., Potyka, N., Tran, T.K., Nayyeri, M., Staab, S.: Faithful embeddings for \mathcal{EL}^{++} knowledge bases. In: *The Semantic Web – ISWC 2022*. pp. 22–38. Springer International Publishing, Cham (2022)
124. Xiong, Y., Ren, M., Zeng, W., Urtasun, R.: Self-supervised representation learning from flow equivariance. In: *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 10191–10200 (2021)
125. Xu, J., Zhang, Z., Friedman, T., Liang, Y., Broeck, G.: A Semantic Loss Function for Deep Learning with Symbolic Knowledge. In: *Proc. 35th International Conference on Machine Learning*. pp. 5502–5511. PMLR (2018)
126. Yang, C., Lamdouar, H., Lu, E., Zisserman, A., Xie, W.: Self-supervised video object segmentation by motion grouping. In: *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 7177–7188 (2021)
127. Yeh, C.K., Kim, B., Arik, S., Li, C.L., Pfister, T., Ravikumar, P.: On completeness-aware concept-based explanations in deep neural networks. In: *Advances in Neural Information Processing Systems* 33. vol. 33, pp. 20554–20565 (2020)
128. Yuksekogonul, M., Wang, M., Zou, J.: Post-hoc Concept Bottleneck Models. In: *ICLR 2022 Workshop on PAIR2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data* (2022)
129. Yuksekogonul, M., Wang, M., Zou, J.: Post-hoc Concept Bottleneck Models. In: *The Eleventh International Conference on Learning Representations* (2023)
130. Zhang, Q., Wang, W., Zhu, S.C.: Examining CNN representations with respect to dataset bias. In: *Proc. 32nd AAAI Conf. Artificial Intelligence*. pp. 4464–4473. AAAI Press (2018)
131. Zhang, Q., Zhu, S.C.: Visual interpretability for deep learning: A survey. *Frontiers of IT & EE* **19**(1), 27–39 (2018)

132. Zhang, R., Madumal, P., Miller, T., Ehinger, K.A., Rubinstein, B.I.: Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In: Proc. AAAI Conference on Artificial Intelligence. vol. 35, pp. 11682–11690 (2021)
133. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M.: Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.* **15**(2), 20:1–20:38 (2024)

A Appendix

A.1 Taxonomy of C-XAI Methods

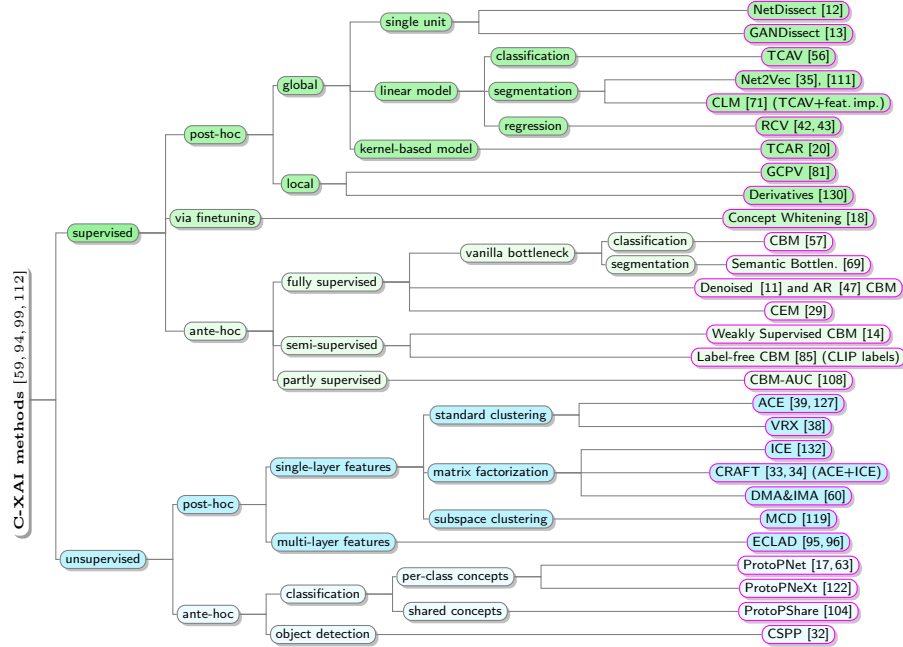


Fig. 4: Detailed taxonomy of state-of-the-art C-XAI methods.

A.2 Details on the Illustration of Concept Distribution

The illustration of the concept distribution in EfficientNet-B0 from Fig. 3a, shown in its separate steps in Fig. 5, was obtained as follows:

1. The local concept vectors are obtained using the **GCPV** optimization technique suggested in [81] with the difference that the optimization objective was changed to pseudo-BCE-loss like in [35]: Given an image together with a concept label, a linear classifier is optimized to correctly classify the activation map pixels of this single image as a concept or background. The normal vector of this linear mapping is then taken as the concept embedding vector for this concept local to this image. This is essentially a local version of Net2Vec [35], where the resulting vectors represent concepts in the context (background) for each sample.
2. This procedure was applied to the concepts belonging to the supercategory “animal” of the MS COCO dataset [68], and the activations of the last layer of `features.7.0` of **EfficientNet-B0** [117].

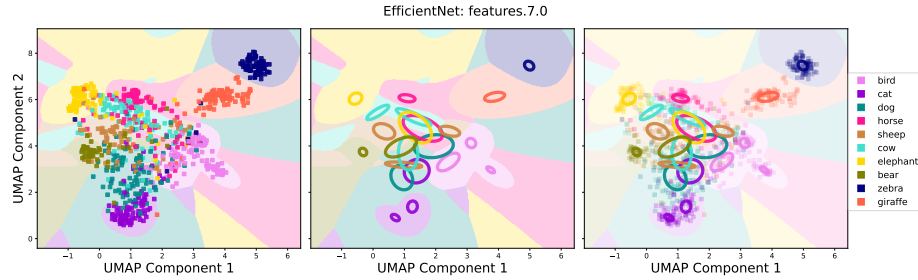


Fig. 5: Some creation steps of Fig. 3a, from *left to right*: (1) Given the local concept embedding vectors, apply UMAP for dimensionality reduction; (2) fit Gaussian mixture models and determine boundaries of standard deviations; (3) add background shading to indicate most probable concepts (here shown for all 3 graphics), and overlay everything.

3. The reduced dimensionality points shown are the density-preserving L_2 -distance **UMAP** [79] 2d-mapping of the set of embedding vectors of the local image concept (of all concepts).
4. On each concept’s vectors separately, a **multivariate Gaussian mixture model** was trained to capture a 2d representation of their distribution. The number of components was determined using the Bayesian information criterion (BIC), using 1 to 3 components and ignoring outliers.
 Note that despite doing the (non-distance preserving) UMAP mapping first and only after that the distribution approximation, the diagrams can still be considered informative, as UMAP preserves the local density information, and thus separation of modes.
5. **Ellipses** are used to visualize the per-concept fitted multivariate Gaussian distributions: They demarcate the boundary at one standard deviation of each of the Gaussian components.
6. The **background color** marks are the most probable concept at that point according to the fitted Gaussians.