Towards Human-Understandable Multi-Dimensional Concept Discovery

Arne Grobrügge[1] Niklas Kühl[2] Gerhard Satzger[1] Philipp Spitzer[1*]

[1]Karlsruhe Institute of Technology, Germany [2]University of Bayreuth, Germany

Abstract

Concept-based eXplainable AI (C-XAI) aims to overcome the limitations of traditional saliency maps by converting pixels into human-understandable concepts that are consistent across an entire dataset. A crucial aspect of C-XAI is completeness, which measures how well a set of concepts explains a model's decisions. Among C-XAI methods, Multi-Dimensional Concept Discovery (MCD) effectively improves completeness by breaking down the CNN latent space into distinct and interpretable concept subspaces. However, MCD's explanations can be difficult for humans to understand, raising concerns about their practical utility. To address this, we propose Human-Understandable Multi-dimensional Concept Discovery (HU-MCD). HU-MCD uses the Segment Anything Model for concept identification and implements a CNN-specific input masking technique to reduce noise introduced by traditional masking methods. These changes to MCD, paired with the completeness relation, enable HU-MCD to enhance concept understandability while maintaining explanation faithfulness. Our experiments, including human subject studies, show that HU-MCD provides more precise and reliable explanations than existing C-XAI methods. The code is available at https://github.com/grobruegge/hu-mcd.

1. Introduction

In recent years, industry and research have witnessed an exponential growth of Machine Learning (ML), particularly accelerated by the advancements of Deep Learning (DL). ML models have proven to solve various problems given sufficient data. However, these models are not failure-free. Previous work has identified various failure cases, including vulnerability to adversarial attacks [35] or biased data [6] and reliance on spurious features [39] that leads to shortcut learning [14]. In high-stake sectors, such as autonomous driving and healthcare, failures could lead to catastrophic outcomes [26, 32]. To prevent such risks, ensuring reliable transparency of ML models is essential. With the recent European regulations—including the General Data Protection Regulation (GDPR) [19] and the European AI act [24]—this need for transparency has made eXplainable Artificial Intelligence (XAI) a central topic in ML research [36]. XAI methods aim to provide transparency of ML models by providing insights into their decision-making processes.

Due to their convenience, *post-hoc* explanations that do not require modifying the underlying model architecture are widely used in the computer vision domain, particularly *local* XAI methods that provide explanations for single predictions [1, 33, 45]. For instance, saliency maps highlight areas in images that are significant for the model's predictions. However, there is a growing consensus that these methods do not provide understandable explanations. Adebayo et al. [2] find that these maps are independent of both the model and the data-generating process, questioning the reliability of such approaches. Moreover, local explanation methods are vulnerable to human confirmation bias [37]—the tendency of humans to favor information that confirms their preexisting assumptions while disregarding contradictory evidence. Finally, the human interpretation of local explanations, such as attribution maps on individual instances, poses challenges and may lead humans to draw contradictory conclusions [16, 20, 21]. Highlighting the location of important regions within an image—*where* the model "looks"—is thus not sufficient for humans to interpret the reasoning of a model. Humans also require the semantic content—*what* the model "sees" [12].

To address the shortcomings of local methods, Concept-based XAI (C-XAI) has emerged as a promising line of research within the area of *global post-hoc* explanations—which take a more holistic approach and explain the overall decision logic of ML models [30]. *Concepts* refer to patterns learned by the model that can be associated with high-level and human-understandable visual attributes. For instance, in the medical field, the recognition of such patterns is crucial to assist clinicians in improving diagnostic accuracy. Lucieri et al. [25] utilize C-XAI in a medical scenario and demonstrate its use in dermatology. While early methods used pre-defined concept datasets [4, 20, 47]

*Correspondence: philipp.spitzer@kit.edu

against which the model is evaluated, more recent work has developed frameworks for automatic concept discovery [11, 15, 37, 44]. First, these frameworks identify visual attributes in images of specific classes for a given task. Then, they cluster similar attributes to form meaningful concepts related to that task. However, they have a conflicting relationship between discovering human-**understandable** concepts and **faithfully** quantifying their significance to model predictions, often prioritizing one over the other.

For instance, Automatic Concept-based Explanations (ACE) [15] uses image segmentation clustering for concept discovery. While obtaining promising results in terms of understandability, the segments must be inpainted and rescaled to meet the model input requirements, resulting in noise that distorts the model's predictions. Furthermore, to ensure understandable concepts, ACE uses several heuristics to exclude outliers but does not consider the degree of information loss, thus raising concerns regarding the faithfulness of its explanations. More recent work has addressed some of the limitations of ACE. In particular, Invertible Concept-based Explanations (ICE) [44] replaces image segments with hidden feature maps, and Concept Recursive Activation FacTorization for Explainability (CRAFT) [11] utilizes quadratic image patches to circumvent inpainting requirements. Nonetheless, a key challenge remains to guarantee faithful explanations: the quantification of the **completeness** of a concept set, *i.e.*, the extent to which these concepts are sufficient to explain the model's predictions. Multi-Dimensional Concept Discovery (MCD) [37] generalizes upon ICE and incorporates a completeness relation, highlighting the superior faithfulness of their concepts compared to previous methods. However, MCD does not quantitatively assess the understandability of its discovered concepts.

In this work, we propose a novel framework that provides both understandable and faithful explanations: Human-Understandable MCD (HU-MCD). To discover concepts that are human-understandable, we use the Segment Anything Model (SAM) [22]. To overcome the noise introduced by the use of rescaled or inpainted images, we employ a novel input masking scheme tailored for Convolutional Neural Networks (CNNs) [3]. Subsequently, we adopt the MCD framework, which enables both local and global concept importance scoring to quantify the significance of each concept for the model's predictions. Unlike ACE, HU-MCD incorporates a *completeness* relation, allowing to account for potential information loss during concept discovery, further enhancing the explanations' faithfulness.

We evaluate HU-MCD on the ImageNet1k [9] dataset and demonstrate that HU-MCD outperforms state-of-the-art methods in both the understandability of discovered concepts and the faithfulness in attributing their importance to the model. Furthermore, we benchmark HU-MCD using Concept Deletion (C-Deletion) and Concept Insertion (C-Insertion), thereby demonstrating that the concept importance scores faithfully represent the model's reasoning.

Overall, our main contributions are threefold: (1) We introduce HU-MCD—a framework for automatic completeness-aware concept-based explanations that uses SAM for human-understandable concept discovery. By using SAM, the manual labeling effort in real-world settings can be reduced. (2) To ensure explanations that faithfully relate to the model's decision-making, we use an input masking scheme tailored for CNNs that effectively mitigates noise introduced by the segmentation masks. (3) We design a human subject study and conduct extensive experiments on established benchmarks to verify the understandability and faithfulness of the concepts generated by HU-MCD. Thereby, HU-MCD takes further steps towards aligning AI decisions with legal regulations (e.g., EU AI Act).

2. Related Work

Supervised Post-Hoc Concept Analysis. Recent research demonstrated that CNNs can encapsulate human-understandable concepts without being explicitly trained on these concepts [46]. This discovery led to the development of several XAI methods, aiming to discover these concepts and measure their influence on model predictions [4, 13, 20, 47]. Notably, the Testing with Concept Activation Vectors (TCAV) framework [20] introduces the notion of Concept Activation Vectors (CAVs)—the weights of a linear classifier used to separate activations corresponding to a specific concept from those corresponding to random data within the activation maps of a neural network's final convolutional layer. CAVs provide a formalized method for representing and quantifying concepts, enabling the interpretation of model behavior in terms of human-understandable features.

Unsupervised Post-Hoc Concept Discovery. While the methods above rely on the availability of a human-defined concepts dataset, subsequent work aimed to eliminate this dependency by automatically discovering concepts. ACE [15] uses superpixel segmentation of class images, clusters their embeddings, and groups similar segments as examples of a concept, which are then analyzed using TCAV. However, the clustering step requires that images are cropped, mean-padded, and resized to the model's input size. These image manipulations distort the aspect ratio, introduce noise, and discard the overall scale ratio. Additionally, ACE applies several heuristics to discard irrelevant segments and clusters but does not account for the information loss during this process. More recent work addresses these shortcomings [11, 37, 41, 44]. Yeh et al. [41] builds on ACE by introducing the notion of *completeness*—the extent to which concept scores serve as sufficient

statistics for recovering the model's prediction. However, they provide limited qualitative results and lack a rigorous human-subject study comparing their approach to similar work. ICE [44] applies Non-negative Matrix Factorization (NMF) on feature maps to identify concepts by disentangling frequently appearing directions within the feature space. CRAFT [11] combines ACE and ICE by applying NMF to feature vectors of image sub-regions, thereby eliminating the necessity for a baseline value to inpaint masked regions for image segments but still requiring rescaling. Unlike ICE, CRAFT uses Sobol indices instead of TCAV but lacks a completeness relation. Instead of representing concepts as a single direction in the feature space, MCD [37] allows concepts to lie on a hyperplane spanned across different convolutional channel directions, thus generalizing ICE. This is realized by Sparse Subspace Clustering (SSC) for feature vector clustering and a subsequent Principle Component Analysis (PCA) for cluster basis derivation. Observing the projection into the subspace not covered by the concepts allows for defining a global completeness score directly on the model's parameters, which differs from the original completeness definition [41]. Similar to ICE, MCD aims to mitigate the noise introduced by rescaling and inpainting image segments. However, their methodology lacks an evaluation of the understandability of the concept. Finally, Segment Any Concept (SAC) [34] provides *local* post-hoc concept explanation. Their definition of concepts differs from the methodologies discussed earlier, focusing exclusively on individual image regions within single images rather than on shared patterns across multiple instances.

Self-Interpretable Concept Models. All the methods above make no modification to the underlying model architecture. An alternative approach is to re-design the architecture such that the decision process is inherently linked to concepts' representations. For instance, Concept Bottleneck Models (CBM) [23] introduce a concept bottleneck layer, where single neurons are explicitly linked to pre-defined concepts which has inspired several subsequent studies [10, 42]. While CBMs require concept annotation for the training dataset, self-interpretable models can also be designed in an unsupervised manner [8, 27, 40].

3. Proposed Method

Building on the objectives of discovering human-*understandable* concepts and *faithfully* attributing their contribution to the model's prediction strategy, we propose Human-Understandable Multi-Dimensional Concept Discovery (HU-MCD). As shown in Figure 1, we distinguish two stages for HU-MCD to align with existing literature [12]. First, we discover relevant concepts by segmenting class images using SAM [22], a foundation model for instance segmentation. We cluster them using SSC based on their representation within the feature space of the model. To avoid introducing noise that might distort the model's prediction by rescaling and inpainting image segments, we employ an input masking scheme specifically designed for CNNs [3]. Second, we adapt the MCD framework proposed by Vielhaben et al. [37], which allows for both local and global concept importance scoring and incorporates a completeness relation by decomposing the model's feature space into multi-dimensional concepts.

3.1. Concept Discovery

We choose a set of images that encapsulate the concepts against which the model will be tested. The selection of samples is not constrained, giving users the flexibility to choose class-specific samples or utilize the entire training set to derive class-agnostic concepts. Inspired by ACE, we use a segmentation algorithm to obtain a dataset comprising distinct image regions acting as concept candidates. In particular, we use SAM [22], as it provides precise and comprehensive instance segmentation with strong zero-shot generalization demonstrated across a broad spectrum of tasks [34]. Using the masks provided by SAM, we select, for each image, the most granular decomposition, considering all masks covering at least 1% of the image.

Concepts are inherently linked to the hidden representation of intermediate feature layers. Previous work has demonstrated that state-of-the-art CNNs learn to represent different features of the data by mapping distinct concepts to different regions of the embedding space [43]. Thus, we employ SSC to ensure that the data points within different clusters lie on a union of distinct low-dimensional subspaces embedded within a higher-dimensional space [37], effectively grouping perceptually similar segments as entities of the same concept. Instead of opting for an arbitrary cluster count (as seen in prior studies such as 10 in Zhang et al. [44] or 25 in Ghorbani et al. [15] and Fel et al. [11]), we leverage the robust segmentation capabilities of SAM and determine the number of clusters based on the average number of segments per image.

Processing image segments presents the challenge of passing irregularly shaped regions through CNN models while extracting feature embeddings. This task is complex due to the fixed-size input requirement of CNN architectures, necessitating rescaling and/or the incorporation of baseline colors to fill masked-out regions, as implemented in ACE. However, many baseline colors are not truly neutral [18], potentially introducing artifacts that can bias the model's predictions. Approaches to address this problem include classical imputation algorithms [5] as suggested by Vielhaben et al. [37] or deep generative models [7]. While these can be effective, they may inadvertently reveal hidden information by recovering masked-out regions or require expensive model training.

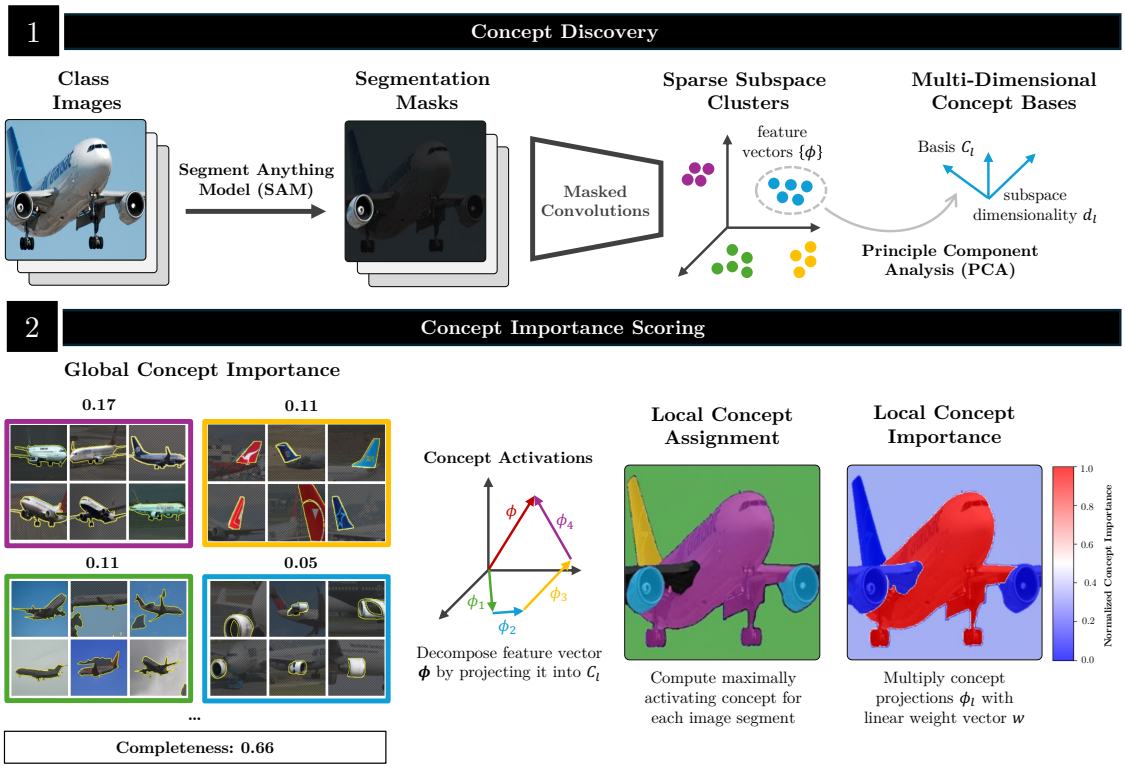

Figure 1. Overview of HU-MCD.

To improve upon ACE, we aim to avoid introducing spurious cues from the mask shape or color. Recent studies indicate that Vision Transformers are less susceptible to masking patterns and colors [18], partially because they can omit patch tokens. In a similar spirit, Balasubramanian and Feizi [3] propose a *layer masking* scheme for CNNs: rather than inserting a baseline color, both the input image and an accompanying mask are propagated through each layer. This effectively simulates running the CNN on an irregularly shaped input—ignoring any activations arising purely from masked regions. By capitalizing on the hierarchical nature of CNNs (whose stacked convolutional layers expand their receptive fields gradually), the masking scheme retains only those values dependent on the unmasked input while discarding those reliant solely on the masked-out areas. Empirically, this leads to better preservation of model accuracy and more *faithful* explanations.

A key challenge in this setup are convolutions near mask edges. Discarding edge convolutions can cause the unmasked portion to rapidly diminish, whereas propagating edges may introduce artifacts by inadvertently revealing masked regions to the model. To mitigate such effects, Balasubramanian and Feizi [3] propose *neighborhood padding*, where the unmasked boundary pixels are padded with an average of their adjacent (non-masked) neighbors. This padding is iterated until it reaches the convolution kernel size. However, given that SAM often produces masks closely aligned with true object edges, extensive padding would remove relevant shape information. For example, an accurately segmented mask of a car tire—once padded to the kernel size—could yield a uniform fill, discarding shape detail entirely. We thus propagate the *un*masked image along with the mask to the first convolutional layer and only apply the masking scheme thereafter. For instance, in a ResNet50 model (with a 7×7 kernel in the first convolution), this grants the kernel access to a narrow band of context around the mask, preventing over-aggressive removal of relevant shape cues. If a particular mask covers more than 25% of the image, we shrink it by the convolution kernel size to avoid exposing large portions of an object's outline when only the background was intended to be masked. Although not flawless, this pragmatic strategy addresses most typical scenarios: small objects remain intact, and large background regions avoid inadvertently revealing the object shape.

3.2. Concept Importance Scoring

To quantify how identified concepts influence the predictions of a model, we adapt the MCD framework. Initially, concepts represent collections of segments associated with similar visual patterns. To achieve independence from individual segments, MCD employs PCA on the hidden rep-

resentations of cluster members, retaining the top principal components as a representative subspace basis, capturing recurring activation patterns. Repeating this process across all clusters yields a set of subspaces that collectively form a comprehensive basis for the feature space. An additional orthogonal complement subspace captures residual information not represented by the identified concepts, ensuring a complete representation. Building upon this decomposition, Vielhaben et al. [37] introduced two complementary metrics: *concept activation*, which quantifies a concept's presence, and *concept relevance*, which assesses the significance of a concept in predicting class labels.

In the original MCD implementation, no image segmentation is utilized; instead, an entire image is processed through the network to generate a feature map, treating each spatial location as a separate feature vector. MCD then decomposes these vectors to compute concept activation and relevance scores, resulting in grid-aligned heatmaps upscaled for visualization [31]. This approach, however, tends to yield ambiguous and block-like regions, as the grid alignment may not correspond to actual object boundaries or meaningful parts. In contrast, our proposed method integrates SAM to guide the discovery of semantically meaningful image regions, thus addressing these interpretability limitations (see Section 4). Consequently, we adapted the original metrics to operate explicitly on image segments, as detailed below.

Concept Activation. Concept activation measures the presence of a concept within a given image segment. Specifically, each concept corresponds to a distinct low-dimensional subspace within the network's hidden layer. By projecting an image segment's hidden representation onto this subspace, we quantify how strongly a concept is expressed. Applying this procedure across all segments generates a concept activation map highlighting regions where each concept is predominantly active. By identifying the concept with the maximum activation score for each segment, we split images into distinct concept regions. Concept prototypes—segments exhibiting the highest activation scores within a sample set—provide intuitive visualizations.

Local Concept Relevance. While concept activation indicates the *presence* of a concept, it does not directly reveal the concept's influence on the model's prediction. To address this, Vielhaben et al. [37] decompose the final hidden layer representation—followed only by a linear mapping to scalar class scores (*logits*)—into contributions from each concept subspace. This decomposition generates *local concept relevance* scores, which quantify each concept's impact on the classification decision at the instance level. Importantly, summing these relevance contributions precisely reconstructs the full logit value, satisfying a completeness criterion. Thus, local relevance scores measure how individual concepts contribute positively or negatively toward a classifier's decision. Applying this analysis segment-wise results in concept relevance heatmaps.

Global Concept Relevance. In contrast to local relevance, *global* concept relevance quantifies concept importance at the class level by projecting the final classification layer's weight vector onto the respective concept subspaces. Similar to local relevance, summing global relevance scores across concepts fully reconstructs the classifier's predictive capability, fulfilling a global completeness criterion based on the model's parameters.

A key advantage of this decomposition framework is its inherent *completeness relation*, ensuring that summing either local or global concept relevance values reproduces the original model outputs (logits or weights). Thus, HU-MCD combines *human-interpretability*—by leveraging SAM-generated fine-grained, semantically meaningful segments—and *faithful* interpretation—through adopting MCD's completeness properties. Furthermore, our input masking strategy supports model fidelity by minimizing noise from baseline colors or mask boundaries.

4. Evaluation

We evaluate HU-MCD's (1) *understandability* of the automatically discovered concepts from a human perspective and (2) *faithfulness* of the concept importance scores in explaining the model prediction. We run our experiments on the ImageNet1k dataset [9] using a selection of ten classes that roughly align with CIFAR10 classes[1] as proposed by Vielhaben et al. [37]. As model architecture, HU-MCD uses ResNet50 with the weights provided by the Python library *timm* [38]. For segmentation, we use SAM with the pre-trained Vision Transformer Huge (ViT-h), the largest of three available Image encoders. We then compare our method with ACE [15] and MCD [37].

Implementation Details. We use SAM on 400 images per class, randomly sampled from the ImageNet1k training set, generating a comprehensive dataset of concept candidates. The generalization of the discovered concept is then verified using class images obtained from the ImageNet1k validation set. In the human-subject study, we implement MCD to ensure the discovery of at least five concepts for each class to ensure a fair comparison to the other methods.

4.1. Understandability

Human Subject Study Design. Motivated by Zhang et al. [44], we use task prediction [17] for evaluation; participants are given one test image with one concept highlighted and five concept explanations from the same class as candidates (as shown in Fig. 2). They are then asked to select the candidate to which the test image most likely belongs. To account for ambiguity, participants can choose up to three or

[1] airliner, beach wagon, hummingbird, Siamese cat, ox, golden retriever, tailed frog, zebra, container ship, police van

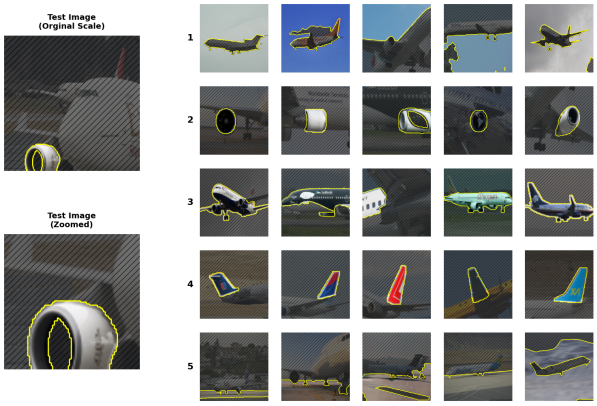

Figure 2. Survey sample of the human subject experiment generated by HU-MCD. Participants are asked to assign the test image on the left to the group on the right which is most similar by only considering the highlighted region.

no candidates. Additionally, they are asked to rate the given concept candidates by stating whether they are *recognizable* and assigning a 1-2 word description for each candidate they rate as recognizable.

The study used a within-subject design involving 15 examples from three methods (ACE, MCD, and HU-MCD) across five random CIFAR-10-like classes, presented in random order. Each participant encountered three sequential examples of each class, with method order randomized. Explanations were generated following method-specific details, retaining the top-10 most influential concepts for each class representing each concept by the top-10 prototypical samples. Five concepts were randomly selected as candidates for each test image. Ten random samples were created per class and method, for a total of 300 different samples. Each participant had a unique selection and order of samples and conducted a short tutorial.

The experiment was conducted online using SoSci Survey, and participants were recruited through Prolific. Attention checks were included to ensure participant comprehension and exclude those who failed the checks. Participants with a prediction accuracy below 20% (random choice) were excluded. A total of 41 participants completed the survey, and each experiment lasted approximately 30 minutes. Participants were compensated with £3 plus an incentive of 3 pennies per correct assignment, up to £3.45. Participant demographics were 61% male, 37% female, 2% unspecified, with ages ranging from 18 to 68 ($\mu = 36$).

Metrics. Similar to Zhang et al. [44], we report the percentage of correctly identified concept explanations. The underlying assumption is that better concept explanations enable participants to associate the highlighted region within the test image to its corresponding concepts more accurately (among five candidates). High prediction accuracy

		Prediction Accuracy ↑	Percentage Recognizable Concepts ↑	Inner-Concept Description Similarity ↑	Intra-Concept Description Similarity ↓
Results	HU-MCD	**70.24%**	**67.12%**	**0.49**	**0.28**
	ACE	42.93%	45.66%	0.39	0.29
	MCD	31.22%	50.34%	0.41	0.38
ANOVA test p-values		< 0.001	< 0.001	< 0.001	< 0.001
T-test p-values	HU-MCD vs. ACE	< 0.001	< 0.001	< 0.001	0.0155
	HU-MCD vs. MCD	< 0.001	0.001	0.008	< 0.001
	ACE vs. MCD	0.029	0.3773	0.4133	< 0.001

Table 1. Results of the human-subject study to validate the *understandability* of the concept explanations generated by HU-MCD. All results involving HU-MCD are statistically significant ($p < 0.05$).

indicates the coherence of individual concepts by requiring concept prototypes of the same concepts to be perceptually similar while being dissimilar to prototypes of other concepts. By asking participants to select all recognizable concepts, we further ensure the understandability of each generated concept explanation. Comparing the 1-2 word descriptions across participants serves as a supplementary indicator of the explanatory quality of the generated concepts. We use pre-trained GloVe [28] word vector representations for each description[2], and we then compute the average pairwise cosine similarity to assess the consistency of concept descriptions across participants (*i.e.*, inner-concept description similarity). Understandable concept explanations should result in different participants assigning similar descriptions to the same concept. Additionally, descriptions across different concepts should differ to indicate that the concepts characterize different attributes of a class. Thus, we also calculate the pairwise cosine similarity of the description embeddings across concepts within one class for each method (*i.e.*, intra-concept description similarity).

Experimental Results. As shown in Table 1, explanations generated by HU-MCD demonstrate a higher prediction accuracy over both ACE and MCD. This confirms that HU-MCD generates a diverse set of concepts and that prototypes of single concepts are perceived as perceptually similar. The superior understandability of the concept explanations is further supported by the percentage of concepts marked as recognizable. Interestingly, MCD shows a notable gap between prediction accuracy and the percentage of recognizable concepts. We observed that this is because the concepts generated by MCD for each class are highly similar, making it possible to recognize them individually but difficult to distinguish them from each other. To underscore this, we additionally evaluated the percentage of recognizable concepts, considering only concepts for which participants provide unique descriptions within each question. This yields a proportion of 37,83% (- 7.83% in com-

[2]We use the *glove-wiki-gigaword-300* model loaded via the Gensim library [29]

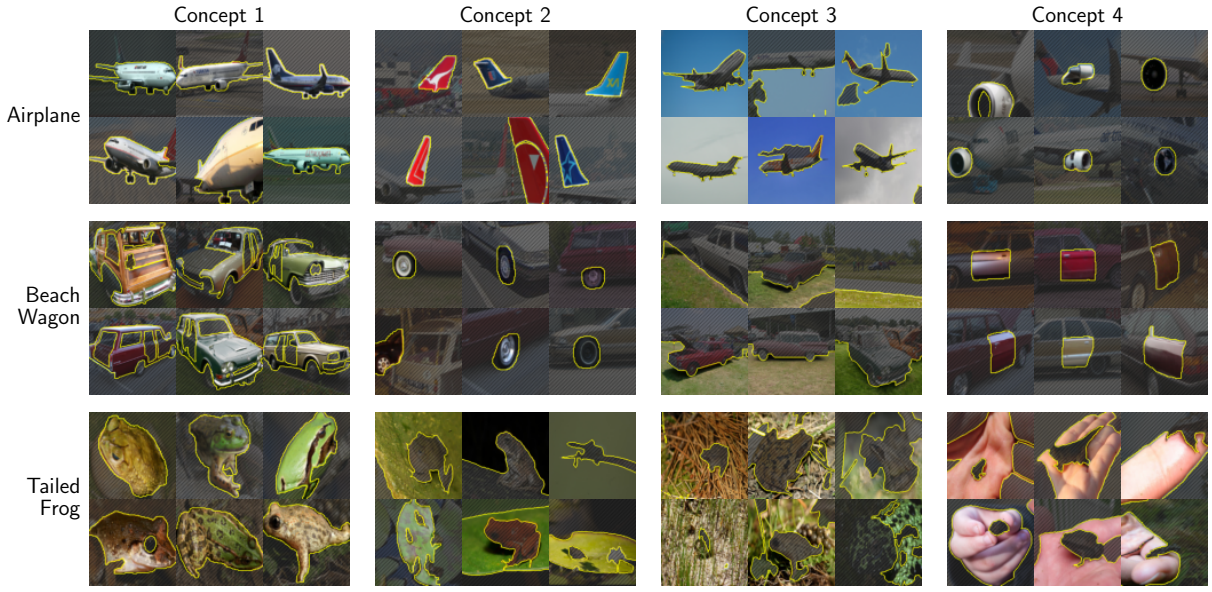

Figure 3. Concept examples for three of the ten *CIFAR-10* alike classes generated by HU-MCD.

parison to the values reported in Table 1) of *uniquely* recognizable concepts for ACE, 38.43 % (- 11.91%) for MCD and 61.03% (- 6.09%) for HU-MCD. Notably, MCD shows the most substantial decline, indicating a lack of clear distinction among its discovered concepts, which limits their effectiveness in explaining the model's prediction behavior.

Finally, HU-MCD achieves the highest description similarity for the same concepts across participants while maintaining distinct descriptions for different concepts within a class. This finding supports the understandability of HU-MCD concepts, as they are perceived similarly by multiple participants. In contrast, MCD shows minimal differences between intra- and inter-concept description similarity, suggesting that participants struggle to consistently identify the meaning of individual concepts and distinguish them. The reason for the improved distinguishability of concepts can be attributed to the usage of SAM, which is trained on human-annotated segmentation masks and thus produces interpretable segments.

Case Study. Figure 3 displays sample concepts identified by HU-MCD, represented by prototypical image segments from the ImageNet1k validation set. It clearly demonstrates the high quality of the segmentation masks, along with the perceptual similarity among prototypes representing the same concept and the distinctiveness between different concepts within the same class, each highlighting unique attributes of the class. These results align with findings from the human-subject study, confirming that HU-MCD's concept explanations are human-understandable. HU-MCD identifies not only entire objects but also parts and contextual information, allowing humans to select and assess concept significance and completeness for any subset of discovered concepts. Notably, HU-MCD identifies concepts indicating spurious correlations among class images, such as *human hands* (concept 4) for the "*tailed frog*." Quantifying the significance of such concepts in model predictions provides a valuable tool for systematic investigations into spurious correlations, thereby addressing model biases.

4.2. Faithfulness

Metrics. To validate the faithfulness of HU-MCD explanations and compare them to prior work, we use the Concept Deletion (C-Deletion) and Concept Insertion (C-Insertion) benchmarks [15]. C-Deletion identifies the smallest set of concepts whose removal results in an incorrect prediction, while C-Insertion identifies the smallest set sufficient for a target class prediction. For each sample, concepts are flipped (masked or unmasked) in decreasing order of local relevance, and the results are aggregated into a single line plot. Specifically, C-Deletion starts with the unmasked image and gradually masks concepts, while C-Insertion starts with the masked image and gradually reveals concepts. Faithful concept relevance scores result in a sharp decrease (C-Deletion) or increase (C-Insertion) in prediction accuracy with the number of flipped concepts. We report the average model prediction accuracy as a function of the fraction of occluded pixels, as proposed by Vielhaben et al. [37].

Experimental Setup. To obtain concept masks and local importance scores for validation images, we segment each image using SAM, compute latent activation with in-

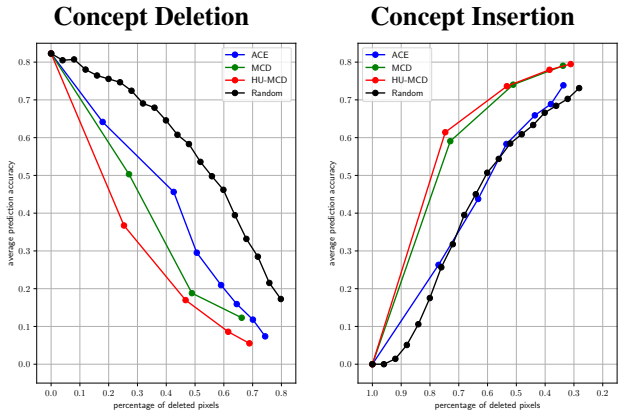

Figure 4. We delete (left) or insert (right) concepts in decreasing order of concept importance and measure the impact on model prediction accuracy, averaged over all validation images of ten *ImageNet1k* classes. Each point represents a discovered concept. Faithful concept importance scores are supposed to result in a sharp decline (left) or ascent (right).

put masking, and apply SSC. Each segment is assigned to a unique concept by selecting the maximum concept activation score and excluding segments aligned with the orthogonal complement, as these are assumed not to encapsulate substantial conceptual significance. The concept importance is calculated by averaging the local relevance score of the corresponding segments, resulting in ordered local concept masks. For ACE and MCD, we follow their original implementation details, noting that ACE uses TCAV scores instead of local concept importance scores. Results are then averaged over 500 images, using 50 ImageNet1k validation images for each of the ten CIFAR10-like classes, and only concepts present in at least 75% of images across all classes are flipped to ensure meaningful averages.

Experimental Results. The results for both C-Deletion and C-Insertion are displayed in Fig. 4. HU-MCD outperforms ACE and MCD in both benchmarks. This result emphasizes the faithfulness of the concept importance scores generated by HU-MCD in representing the model's reasoning process. Interestingly, the C-Insertion benchmark proves to be more challenging than C-Deletion, given that the model begins with a fully masked image, resulting in a near-random performance for ACE. Although the gap narrows, HU-MCD consistently achieves accuracy that is either superior to or comparable with MCD.

5. Limitations

HU-MCD demonstrated promising results for the explanations' *understandability* and *faithfulness*. However, we acknowledge certain limitations. First, non-coherent concepts may arise due to segmentation errors, clustering inaccuracies, or limitations in the similarity metric. Such occurrences may be due to inherent methodological constraints or discrepancies between the model's and humans' perception of similarity. The user study shows that such occurrences are infrequent. Future work can entail an exploration of hyperparameters to overcome this challenge.

Second, our experiments are conducted on visual data, and the user study includes only ten classes. As emphasized by previous work [41], the general idea of concept-based explanations also applies to other data types, such as texts. This adoption challenge provides a promising avenue for future research to adapt HU-MCD to other modalities and expand its usability. Additionally, future research can extend the evaluation of HU-MCD to a wider variety as well as more fine-grained classes.

Finally, a technical limitation in the current setting is that the experiments are restricted to layers followed solely by linear operations to calculate concept relevance scores. However, concept activations can reveal the learned structures within the feature space for any layer. As proposed by Vielhaben et al. [37], future research could approximate the remainder of the model with a linear model, thereby enabling the quantification of concept relevance at different layers.

6. Conclusion

In this work, we introduced HU-MCD, a novel framework designed to extract human-understandable concepts from ML models automatically. HU-MCD satisfies two key criteria that previous research has neglected: (1) providing human-understandable concepts *and* (2) faithfully attributing their importance to the model's predictions. For the first time, we use the SAM to extract understandable concepts automatically. Our approach extends prior concept discovery methods that use segmentation techniques by implementing a novel input masking scheme, which addresses noise introduced by inpainting and rescaling requirements. By representing concepts as multi-dimensional linear subspaces within the hidden feature space of a trained ML model, HU-MCD enables the decomposition of activations into unique concept attributions. This facilitates the calculation of concept importance scores both globally (per-class) and locally (per-image). Additionally, HU-MCD incorporates a completeness relation, quantifying the extent to which concepts sufficiently explain the model's predictions. This distinguishes it from most existing work in the field and offers the possibility to analyze models from a more human-centered perspective. We conduct extensive experiments on common benchmarks as well as a user study demonstrating both the faithfulness and understandability of the explanations generated by HU-MCD.

Acknowledgment

We thank Maria-Paola Forte from Max Planck Institute for Intelligent Systems for her valuable feedback on this work. We also thank the Karlsruhe Digital Service Research & Innovation Hub (KSRI) team for the valuable feedback and consultation on the study design.

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018. 1

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 1

[3] Sriram Balasubramanian and Soheil Feizi. Towards improved input masking for convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1855–1865, 2023. 2, 3, 4

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 1, 2

[5] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages I–I. IEEE, 2001. 3

[6] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 1

[7] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2019. 3

[8] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 5

[10] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems*, pages 21400–21413. Curran Associates, Inc., 2022. 3

[11] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2711–2721, 2023. 2, 3

[12] Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[13] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8730–8738, 2018. 2

[14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1

[15] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019. 2, 3, 5, 7

[16] Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, 2020. Association for Computational Linguistics. 1

[17] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5:1096257, 2023. 5

[18] Saachi Jain, Hadi Salman, Eric Wong, Pengchuan Zhang, Vibhav Vineet, Sai Vemprala, and Aleksander Madry. Missingness bias in model debugging. In *International Conference on Learning Representations*, 2022. 3, 4

[19] Margot E Kaminski and Jennifer M Urban. The right to contest ai. *Columbia Law Review*, 121(7):1957–2048, 2021. 1

[20] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 1, 2

[21] Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Hive: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision*, pages 280–298. Springer, 2022. 1

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3

[23] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020. 3

[24] Mauritz Kop. Eu artificial intelligence act: The european approach to ai. *Stanford - Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust and IPR Developments*, 2021. 1

[25] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. Exaid: A multimodal explanation framework for computer-aided diagnosis of skin lesions. *Computer Methods and Programs in Biomedicine*, 215:106620, 2022. 1

[26] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. The impact of imperfect xai on human-ai decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–39, 2024. 1

[27] Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 6

[29] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, 2010. 6

[30] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pages 1–59, 2023. 1

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 5

[32] Philipp Spitzer, Joshua Holstein, Katelyn Morrison, Kenneth Holstein, Gerhard Satzger, and Niklas Kühl. Don't be fooled: The misinformation effect of explanations in human-ai collaboration. *arXiv preprint arXiv:2409.12809*, 2024. 1

[33] Philipp Spitzer, Niklas Kühl, Marc Goutier, Manuel Kaschura, and Gerhard Satzger. Transferring domain knowledge with (x) ai-based learning systems. *Proceedings of the 32nd European Conference on Information Systems (ECIS)*, 2024. 1

[34] Ao Sun, Pingchuan Ma, Yuanyuan Yuan, and Shuai Wang. Explain any concept: Segment anything meets concept-based explanation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[36] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11): 4793–4813, 2020. 1

[37] Johanna Vielhaben, Stefan Blücher, and Nils Strodthoff. Multi-dimensional concept discovery (mcd): A unifying framework with completeness guarantees. *Transactions on Machine Learning Research*, 2023. 1, 2, 3, 5, 7, 8

[38] Ross Wightman. Pytorch image models. `https://github.com/huggingface/pytorch-image-models`. Last Accessed: Nov. 2024. 5

[39] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. 1

[40] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023. 3

[41] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33: 20554–20565, 2020. 2, 3, 8

[42] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3

[44] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11682–11690, 2021. 2, 3, 5, 6

[45] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021. 1

[46] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *CoRR*, abs/1412.6856, 2014. 2

[47] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018. 1, 2