Concept-based Explanations for Out-of-Distribution Detectors

Jihye Choi [1] **Jayaram Raghuram** [1] **Ryan Feng** [2] **Jiefeng Chen** [1] **Somesh Jha** [1] **Atul Prakash** [2]

Abstract

Out-of-distribution (OOD) detection plays a crucial role in ensuring the safe deployment of deep neural network (DNN) classifiers. While a myriad of methods have focused on improving the performance of OOD detectors, a critical gap remains in interpreting their decisions. We help bridge this gap by providing explanations for OOD detectors based on learned high-level concepts. We first propose two new metrics for assessing the effectiveness of a particular set of concepts for explaining OOD detectors: 1) *detection completeness* – which quantifies the sufficiency of concepts for explaining an OOD detector's decisions, and 2) *concept separability* – which captures the distributional separation between in-distribution and OOD data in the concept space. Based on these metrics, we propose an unsupervised framework for learning a set of concepts that satisfy the desired properties of high detection completeness and concept separability, and demonstrate its effectiveness in providing concept-based explanations for diverse off-the-shelf OOD detectors. We also show how to identify prominent concepts contributing to the detection results, and provide further reasoning about their decisions.

1. Introduction

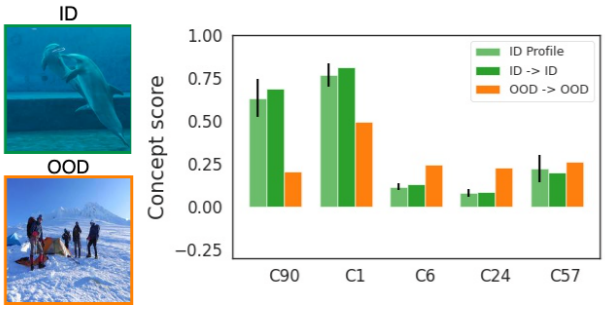
It is well known that machine learning (ML) models can yield uncertain and unreliable predictions on out-of-distribution (OOD) inputs, *i.e.*, inputs from outside the training distribution (Amodei et al., 2016; Goodfellow et al., 2015; Hendrycks et al., 2021). A common line of defense in this situation is to augment the ML model (*e.g.*, a DNN classifier) with a detector that can identify and flag such inputs as OOD (Hendrycks & Gimpel, 2017; Liang et al., 2018). The ML model can then abstain from making predictions on such inputs (Tax & Duin, 2008; Geifman & El-Yaniv, 2019). In many application domains (*e.g.*, medical imaging), it is important to understand both the model's prediction as well the reason for abstaining from prediction on certain inputs (*i.e.*, the OOD detector's decisions). Moreover, abstaining from prediction can often have a practical cost, *e.g.*, due to service denial or the need for manual intervention (Markoff, 2016; Mozannar & Sontag, 2020).

Detecting OOD inputs has received significant attention in the literature, and a number of methods exist that achieve strong detection performance on semantic distribution shifts (Yang et al., 2021; 2022). Much of the focus in learning OOD detectors has been on improving their detection performance (Hendrycks et al., 2019; Liu et al., 2020; Mohseni et al., 2020; Lin et al., 2021; Chen et al., 2021; Sun et al., 2021; Cao & Zhang, 2022). However, the problem of explaining the decisions of an OOD detector and the related problem of designing inherently-interpretable detectors remain largely unexplored (we focus on the former problem). A potential approach could be to run an existing explanation method for DNN classifiers with in-distribution (ID) and OOD data separately, and then inspect the difference between the generated explanations. However, it is unclear whether an explanation method that is effective for ID class predictions will also be effective for OOD detection. For instance, feature attributions, the most popular type of explanation (Sundararajan et al., 2017; Ribeiro et al., 2016), may not capture visual differences in the generated explanations between ID and OOD inputs (Adebayo et al., 2020). Moreover, their explanations based on pixel-level activations may not provide the most intuitive form of explanations for humans.

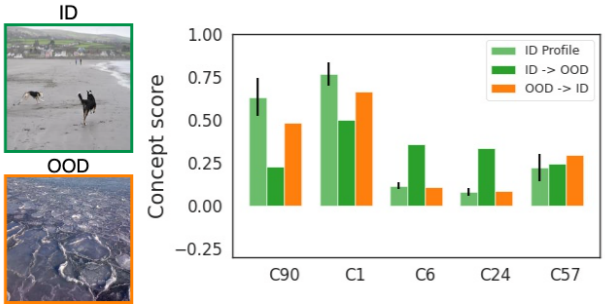
This paper addresses the above research gap by proposing the first method (to our knowledge) for interpreting the decisions of an OOD detector in a human-understandable way. We build upon recent advances in concept-based explanations for DNN classifiers (Ghorbani et al., 2019; Zhou et al., 2018a; Bouchacourt & Denoyer, 2019; Yeh et al., 2020), which offer the benefit of providing explanations in terms of high-level *concepts* for classification tasks. We focus on extending this explanation framework to the problem of OOD detection. As a concrete example, consider Fig. 1 which illustrates our concept-based explanations given inputs which are all classified as the class "Dolphin" by a DNN classifier,

[1]University of Wisconsin - Madison [2]University of Michigan. Correspondence to: Jihye Choi <jihye@cs.wisc.edu>.

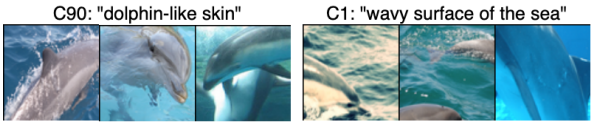
Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).



(a) Correct detection: top dolphin image is correctly detected as ID (dark-green bar), and the bottom image is correctly detected as OOD (orange bar).



(b) Wrong detection: top ID image is detected as OOD (dark-green bar), and the bottom OOD image is detected as ID (orange bar).



(c) Visualization of top-2 important concepts.

Figure 1: **Our concept-based explanation for the Energy OOD detector (Liu et al., 2020).** All input images are classified as "Dolphin" but detected differently. On the x-axis of bar graphs, we present the top-5 important concepts that describe the detector's behavior given images classified as "Dolphin". ID profile (light-green) shows the average concept score pattern for ID images predicted as "Dolphin". We expect ID inputs predicted into this class to have a similar concept score pattern.

but detected as either ID or OOD by an OOD detector. Our method identifies that the concepts such as C90 "dolphin-like skin" and C1 "wavy surface of the sea" are key concepts to understand the OOD detector's decisions to tell apart ID and OOD images predicted as "Dolphin". The user can verify that these concepts are aligned with human intuition and the OOD detector relies on them for making decisions. We also confirm that the OOD detector predicts a certain input as ID when its concept-score patterns are similar to that of normal ID Dolphin images. Likewise, the detector predicts an input as OOD when its concept-score patterns are very different from that of ID inputs from the same class. Our explanations can help a user analyze if an incorrect detection (as in Fig. 1b) is an understandable mistake or a misbehavior of the OOD detector, evaluate the reliability of the OOD detector, and decide upon its adoption in practice.

We aim to design a general interpretability framework that is applicable across a wide range of black-box OOD detectors. Accordingly, a research question we ask is: *without relying on the internal mechanism of an OOD detector, can we identify a good set of concepts that are appropriate for understanding why the OOD detector predicts a certain input to be ID/OOD?* A key contribution of this paper is to show that this can be done unsupervised, without any additional human annotations for interpretation.

In summary, we make the following contributions:

- We motivate and propose new metrics to quantify the effectiveness of concept-based explanations for a black-box OOD detector, namely *detection completeness* and *concept separability* (Sections 2.2, 3.1, and 3.2).
- We propose a concept-learning objective with suitable regularization terms that, given an OOD detector for a DNN classifier, learns a set of concepts with high detection completeness and concept separability (Section 3.3);
- By treating the OOD detector as a black-box, we show that our approach can be applied to explain a variety of existing OOD detection methods. We also provide empirical evidence that concepts learned for classifiers cannot be directly used to explain OOD detectors, whereas concepts learned by our method are effective for explaining both the classifier and OOD detector (Section 4.2).
- By identifying prominent concepts that contribute to an OOD detector's decisions via a modified Shapley value importance score based on the detection completeness, we demonstrate how the discovered concepts can be used to interpret the OOD detector (Section 4.3).

Related Work. In the literature of OOD detection, recent studies have designed various scoring functions based on the representation from the final or penultimate layers (Liang et al., 2018; DeVries & Taylor, 2018), or a combination of different internal layers of a DNN classifier (Lee et al., 2018; Lin et al., 2021; Raghuram et al., 2021). A recent survey on generalized OOD detection can be found in Yang et al. (Yang et al., 2021). Our work aims to provide post-hoc explanations applicable to a wide range of black-box OOD detectors without modifying their internals. Among different interpretability approaches, concept-based explanation (Koh et al., 2020; Alvarez-Melis & Jaakkola, 2018) has gained popularity as it is designed to be better-aligned with human reasoning (Armstrong et al., 1983; Tenenbaum, 1999) and intuition (Ghorbani et al., 2019; Zhou et al., 2018a; Bouchacourt & Denoyer, 2019; Yeh et al., 2020). There have been limited attempts to assess the use of concept-based explanations under data distribution changes such as adversarial manipulation (Kim et al., 2018) or spurious correlations (Adebayo et al., 2020). However, designing concept-based explanations for OOD detection requires further exploration and is the focus of our work.

2. Problem Setup and Background

Notations. Let $\mathcal{X} \subseteq \mathbb{R}^{a_0 \times b_0 \times d_0}$ denote the space of inputs $\mathbf{x}$, where d_0 is the number of channels and a_0 and b_0 are the image size along each channel. Let $\mathcal{Y} := \{1, \cdots, L\}$ denote the space of output class labels y. Let Δ_L denote the set of all probabilities over $\mathcal{Y}$ (the simplex in L-dimensions). We assume that natural inputs to the DNN classifier are sampled from an unknown probability distribution P_{in} over the space $\mathcal{X} \times \mathcal{Y}$. The compact notation $[n]$ denotes $\{1, \cdots, n\}$ for a positive integer n. Boldface symbols are used to denote both vectors and tensors. $\langle \mathbf{x}, \mathbf{x}' \rangle$ denotes the standard inner-product between a pair of vectors. The indicator function $\mathbb{1}[c]$ takes value 1 (0) when the condition c is true (false).

ID and OOD Datasets. Consider a labeled ID training dataset $D_{\text{in}}^{\text{tr}} = \{(\mathbf{x}_i, y_i),\ i = 1, \cdots, N_{\text{in}}^{\text{tr}}\}$ sampled from the distribution P_{in}. We assume the availability of an unlabeled training dataset $D_{\text{out}}^{\text{tr}} = \{\widetilde{\mathbf{x}}_i,\ i = 1, \cdots, N_{\text{out}}^{\text{tr}}\}$ from a different distribution, referred to as the *auxiliary OOD dataset*. Similarly, we define the ID test dataset (from P_{in}) as $D_{\text{in}}^{\text{te}}$, and the OOD test dataset as $D_{\text{out}}^{\text{te}}$. Note that the auxiliary OOD dataset $D_{\text{in}}^{\text{tr}}$ and the test OOD dataset $D_{\text{out}}^{\text{te}}$ are from different distributions. All the OOD datasets are unlabeled since their label space is usually different from $\mathcal{Y}$.

OOD Detector. The goal of an OOD detector is to determine if a test input to the classifier is ID (*i.e.*, from the distribution P_{in}); otherwise the input is declared to be OOD (Yang et al., 2021). Given a trained classifier $\mathbf{f} : \mathcal{X} \mapsto \Delta_L$, the decision function of an OOD detector can be generally defined as $\mathcal{D}_\gamma(\mathbf{x}, \mathbf{f}) = \mathbb{1}[S(\mathbf{x}, \mathbf{f}) \geq \gamma]$, where $S(\mathbf{x}, \mathbf{f}) \in \mathbb{R}$ is the score function of the detector for an input $\mathbf{x}$ and γ is the threshold. We follow the convention that larger scores correspond to ID inputs, and the detector outputs of 1 and 0 correspond to ID and OOD respectively. We assume the availability a pre-trained DNN classifier and a paired OOD detector that is trained to detect inputs for the classifier.

2.1. Projection Into Concept Space

Consider a pre-trained DNN classifier $\mathbf{f} : \mathcal{X} \mapsto \Delta_L$ that maps an input $\mathbf{x}$ to its corresponding predicted class probabilities. Without loss of generality, we can partition the DNN at a convolutional layer ℓ into two parts, *i.e.*, $\mathbf{f} = \mathbf{h} \circ \boldsymbol{\phi}$ where: 1) $\boldsymbol{\phi} : \mathcal{X} \mapsto \mathcal{Z} := \mathbb{R}^{a_\ell b_\ell \times d_\ell}$ is the first half of $\mathbf{f}$ that maps an input $\mathbf{x}$ to the intermediate feature representation [1] $\boldsymbol{\phi}(\mathbf{x})$, and 2) $\mathbf{h} : \mathcal{Z} \mapsto \Delta_L$ is the second half of $\mathbf{f}$ that maps $\boldsymbol{\phi}(\mathbf{x})$ to the predicted class probabilities $\mathbf{h}(\boldsymbol{\phi}(\mathbf{x}))$. We denote the predicted probability of a class y by $f_y(\mathbf{x}) = h_y(\boldsymbol{\phi}(\mathbf{x}))$, and the prediction of the classifier by $\widehat{y}(\mathbf{x}) = \text{argmax}_y\, f_y(\mathbf{x})$.

[1] We flatten the first two dimensions of the feature representation, thus changing an $a_\ell \times b_\ell \times d_\ell$ tensor to an $a_\ell b_\ell \times d_\ell$ matrix, where a_ℓ and b_ℓ are the filter size and d_ℓ is the number of channels.

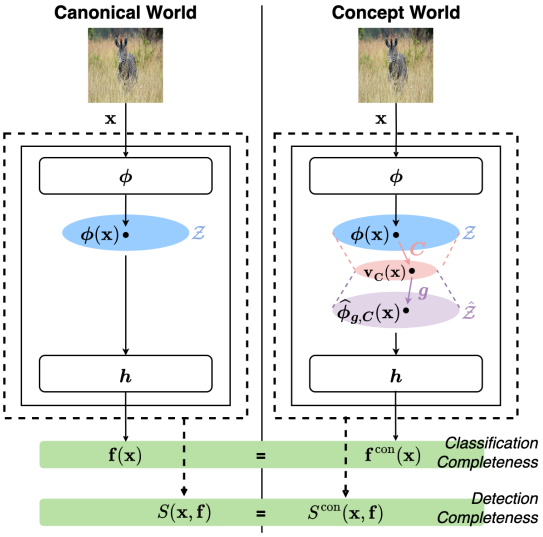


Figure 2: **Our two-world view of the classifier and OOD detector.** In the canonical world, both the classifier and OOD detector are unmodified. In the concept world, the layer representation $\boldsymbol{\phi}(\mathbf{x})$ is projected into the space spanned by the concept vectors and then reconstructed via the non-linear mapping $\mathbf{g}$. The classifier and OOD detector in the concept world are based on this reconstructed layer representation. Given the same input, the outputs from the DNN classifier and OOD detector in both the worlds should be very close to each other (characterized by *Classification Completeness* and *Detection Completeness*, respectively).

Our work is based on the common implicit assumption of linear interpretability in the concept-based explanation literature, *i.e.*, high-level concepts lie in a linearly-projected subspace of the feature representation space $\mathcal{Z}$ of the classifier (Kim et al., 2018). Consider a projection matrix $\mathbf{C} = [\mathbf{c}_1, \cdots, \mathbf{c}_m] \in \mathbb{R}^{d_\ell \times m}$ (with $m \ll d_\ell$) that maps from the space $\mathcal{Z}$ into a reduced-dimension concept space. $\mathbf{C}$ consists of m unit vectors, where $\mathbf{c}_i \in \mathbb{R}^{d_\ell}$ is referred to as the *concept vector* representing the i-th concept (*e.g.*, "stripe" or "oval face"), and m is the number of concepts. We define the *concept score* for $\mathbf{x}$ as the linear projection of the high-dimensional layer representation $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^{a_\ell b_\ell \times d_\ell}$ into the concept space (Yeh et al., 2020), i.e. $\mathbf{v}_{\mathbf{C}}(\mathbf{x}) := \boldsymbol{\phi}(\mathbf{x})\, \mathbf{C} \in \mathbb{R}^{a_\ell b_\ell \times m}$. We also define a mapping from the projected concept space back to the feature space by a non-linear function $\mathbf{g} : \mathbb{R}^{a_\ell b_\ell \times m} \mapsto \mathbb{R}^{a_\ell b_\ell \times d_\ell}$. The reconstructed feature representation at layer ℓ is then defined as $\widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}}(\mathbf{x}) := \mathbf{g}(\mathbf{v}_{\mathbf{C}}(\mathbf{x}))$.

2.2. Canonical World and Concept World

As shown in Fig. 2, we consider a "two-world" view of the classifier and OOD detector consisting of the *canonical world* and the *concept world*, which are defined as follows:

Canonical World. In this case, both the classifier and OOD detector use the original layer representation $\boldsymbol{\phi}(\mathbf{x})$ for their predictions. The prediction of the classifier is $\mathbf{f}(\mathbf{x}) = \mathbf{h}(\boldsymbol{\phi}(\mathbf{x}))$, and the decision function of the detector is $\mathcal{D}_\gamma(\mathbf{x}, \mathbf{h} \circ \boldsymbol{\phi})$ with a score function $S(\mathbf{x}, \mathbf{h} \circ \boldsymbol{\phi})$.

Concept World. We use the following observation in constructing the concept-world formulation: *both the classifier and the OOD detector can be modified to make predictions based on the reconstructed feature representation*, *i.e.*, using $\widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}}(\mathbf{x})$ instead of $\boldsymbol{\phi}(\mathbf{x})$. Accordingly, we define the corresponding classifier, detector, and score function in the concept world as follows:

$$\begin{aligned} \mathbf{f}^{\text{con}}(\mathbf{x}) &:= \mathbf{h}(\widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}}(\mathbf{x})) = \mathbf{h}(\mathbf{g}(\mathbf{v}_{\mathbf{C}}(\mathbf{x}))) \\ \mathcal{D}^{\text{con}}_\gamma(\mathbf{x}, \mathbf{f}) &:= \mathcal{D}_\gamma(\mathbf{x}, \mathbf{h} \circ \widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}}) = \mathcal{D}_\gamma(\mathbf{x}, \mathbf{h} \circ \mathbf{g} \circ \mathbf{v}_{\mathbf{C}}) \\ S^{\text{con}}(\mathbf{x}, \mathbf{f}) &:= S(\mathbf{x}, \mathbf{h} \circ \widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}}) = S(\mathbf{x}, \mathbf{h} \circ \mathbf{g} \circ \mathbf{v}_{\mathbf{C}}). \quad (1) \end{aligned}$$

We further elaborate on this two-world view and introduce the following two desirable properties.

Detection Completeness. Given a fixed algorithmic approach for learning the classifier and OOD detector, and with fixed internal parameters of $\mathbf{f}$, we would ideally like the classifier prediction and the detection score to be indistinguishable between the two worlds. In other words, for the concepts to *sufficiently* explain the OOD detector, we require $\mathcal{D}^{\text{con}}_\gamma(\mathbf{x}, \mathbf{f})$ to closely mimic $\mathcal{D}_\gamma(\mathbf{x}, \mathbf{f})$. Likewise, we require $\mathbf{f}^{\text{con}}(\mathbf{x})$ to closely mimic $\mathbf{f}(\mathbf{x})$ since the detection mechanism of $\mathcal{D}_\gamma$ is closely paired to the classifier. We refer to this property as the *completeness of a set of concepts with respect to the OOD detector and its paired classifier.* As discussed in § 3.1, this extends the notion of classification completeness introduced by Yeh et al. (2020) to an OOD detector and its paired classifier.

Concept Separability. To improve the interpretability of the resulting explanations for the OOD detector, we require another desirable property from the learned concepts: data detected as ID by $\mathcal{D}_\gamma$ (henceforth referred to as *detected-ID* data) and data detected as OOD by $\mathcal{D}_\gamma$ (henceforth referred to as *detected-OOD* data) should be well-separated in the concept-score space. Since our goal is to help an analyst understand which concepts distinguish the detected-ID data from detected-OOD data, we would like to learn a set of concepts that have a well-separated concept score pattern for inputs from these two groups (*e.g.*, the concepts C_{90} and C_1 in Fig. 1 have distinct concept scores).

3. Proposed Approach

Given a trained DNN classifier $\mathbf{f}$, a paired OOD detector $\mathcal{D}_\gamma$, and a set of concepts $\mathbf{C}$, we address the following questions: **1)** *Are the concepts sufficient to capture the prediction behavior of both the classifier and OOD detector?* (see § 3.1); **2)** *Do the concepts show clear distinctions in their scores between detected-ID data and detected-OOD data?* (see § 3.2). We first propose new metrics for quantifying the set of learned concepts, followed by a general framework for learning concepts that possess these properties (see § 3.3).

3.1. Metric for Detection Completeness

Definition 1. Given a trained DNN classifier $\mathbf{f} = \mathbf{h} \circ \boldsymbol{\phi}$ and a set of concept vectors $\mathbf{C}$, the *classification completeness* with respect to $P_{\text{in}}(\mathbf{x}, y)$ is defined as (Yeh et al., 2020):

$$\eta_{\mathbf{f}}(\mathbf{C}) := \frac{\sup_{\mathbf{g}} \mathbb{E}_{(\mathbf{x},y)\sim P_{\text{in}}}\left[\mathbb{1}[y = \text{argmax}_{y'} h_{y'}(\widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}}(\mathbf{x}))]\right] - a_r}{\mathbb{E}_{(\mathbf{x},y)\sim P_{\text{in}}}\left[\mathbb{1}[y = \text{argmax}_{y'} h_{y'}(\boldsymbol{\phi}(\mathbf{x}))]\right] - a_r}$$

where $a_r = 1/L$ is the accuracy of a random classifier.

The denominator of $\eta_{\mathbf{f}}(\mathbf{C})$ is the accuracy of the original classifier $\mathbf{f}$, while the numerator is the maximum accuracy that can be achieved by the concept-world classifier. The maximization is over the parameters of the neural network $\mathbf{g}$ that reconstructs the feature representation from the vector of concept scores.

Definition 2. Given a trained DNN classifier $\mathbf{f} = \mathbf{h} \circ \boldsymbol{\phi}$, a trained OOD detector with score function $S(\mathbf{x}, \mathbf{f})$, and a set of concept vectors $\mathbf{C}$, we define the *detection completeness score* with respect to the ID distribution $P_{\text{in}}(\mathbf{x}, y)$ and OOD distribution $P_{\text{out}}(\mathbf{x})$ as follows:

$$\eta_{\mathbf{f},S}(\mathbf{C}) := \frac{\sup_{\mathbf{g}} \text{AUC}(\mathbf{h} \circ \widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}}) - b_r}{\text{AUC}(\mathbf{h} \circ \boldsymbol{\phi}) - b_r}, \quad (2)$$

where $\text{AUC}(\mathbf{f})$ is the area under the ROC curve of an OOD detector based on $\mathbf{f}$, defined as $\text{AUC}(\mathbf{f}) := \mathbb{E}_{(\mathbf{x},y)\sim P_{\text{in}}} \mathbb{E}_{\mathbf{x}'\sim P_{\text{out}}} \mathbb{1}\big[S(\mathbf{x}, \mathbf{f}) > S(\mathbf{x}', \mathbf{f})\big]$, and $b_r = 0.5$ is the AUROC of a random detector.

The numerator is the maximum achievable AUROC in the concept world using the reconstructed representation from concept scores. In practice, $\text{AUC}(\mathbf{f})$ is estimated using the test datasets $D^{\text{te}}_{\text{in}}$ and $D^{\text{te}}_{\text{out}}$. Both the classification completeness and detection completeness are designed to be in the range $[0, 1]$. However, this is not strictly guaranteed since the classifier or OOD detector in the concept world may empirically have a better (corresponding) metric on a given ID/OOD dataset. Completeness scores close to 1 indicate that the set of concepts $\mathbf{C}$ are close to complete in characterizing the behavior of the classifier and/or OOD detector.

3.2. Concept Separability Score

Concept Scores. In Section 2.1, we introduced a projection matrix $\mathbf{C} \in \mathbb{R}^{d_\ell \times m}$ that maps $\boldsymbol{\phi}(\mathbf{x})$ to $\mathbf{v}_{\mathbf{C}}(\mathbf{x})$, and consists of m unit concept vectors $\mathbf{C} = [\mathbf{c}_1 \cdots \mathbf{c}_m]$. The

inner product between the feature representation and a concept vector is referred to as the *concept score*, and it quantifies how close an input is to the given concept (Kim et al., 2018; Ghorbani et al., 2019). Specifically, the concept score corresponding to concept i is defined as $\mathbf{v}_{\mathbf{c}_i}(\mathbf{x}) := \langle \boldsymbol{\phi}(\mathbf{x}), \mathbf{c}_i \rangle = \boldsymbol{\phi}(\mathbf{x})\, \mathbf{c}_i \in \mathbb{R}^{a_\ell b_\ell}$. The matrix of concept scores from all the concepts is simply the concatenation of the individual concept scores, *i.e.*, $\mathbf{v}_{\mathbf{C}}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})\, \mathbf{C} = [\mathbf{v}_{\mathbf{c}_1}(\mathbf{x}) \cdots \mathbf{v}_{\mathbf{c}_m}(\mathbf{x})] \in \mathbb{R}^{a_\ell b_\ell \times m}$. We also define a dimension-reduced version of the concept scores that takes the maximum of the inner-product over each $a_\ell \times b_\ell$ patch as follows: $\widetilde{\mathbf{v}}_{\mathbf{C}}(\mathbf{x})^T = [\widetilde{v}_{\mathbf{c}_1}(\mathbf{x}), \cdots, \widetilde{v}_{\mathbf{c}_m}(\mathbf{x})] \in \mathbb{R}^m$, where $\widetilde{v}_{\mathbf{c}_i}(\mathbf{x}) = \max_{p,q} |\langle \boldsymbol{\phi}^{p,q}(\mathbf{x}), \mathbf{c}_i \rangle| \in \mathbb{R}$. Here $\boldsymbol{\phi}^{p,q}(\mathbf{x})$ is the feature representation corresponding to the (p, q)-th patch of input $\mathbf{x}$ (*i.e.*, receptive field (Araujo et al., 2019)). This reduction operation is done to capture the most important correlations from each patch, and the m-dimensional concept score will be used to define our concept separability metric as follows.

We would like the set of concept-score vectors from the detected-ID class $V_{\text{in}}(\mathbf{C}) := \{\widetilde{\mathbf{v}}_{\mathbf{C}}(\mathbf{x}),\ \mathbf{x} \in D_{\text{in}}^{\text{tr}} \cup D_{\text{out}}^{\text{tr}} \ :\ \mathcal{D}_\gamma(\mathbf{x}, \mathbf{f}) = 1\}$, and the set of concept-score vectors from the detected-OOD class $V_{\text{out}}(\mathbf{C}) := \{\widetilde{\mathbf{v}}_{\mathbf{C}}(\mathbf{x}),\ \mathbf{x} \in D_{\text{in}}^{\text{tr}} \cup D_{\text{out}}^{\text{tr}} \ :\ \mathcal{D}_\gamma(\mathbf{x}, \mathbf{f}) = 0\}$ to be well separated. Let $J_{\text{sep}}(V_{\text{in}}(\mathbf{C}), V_{\text{out}}(\mathbf{C})) \in \mathbb{R}$ define a general *measure of separability* between the two subsets, such that a larger value corresponds to higher separability. We discuss a specific choice for J_{sep} for which it is possible to tractably optimize concept separability as part of the learning objective in Section 3.3.

Global Concept Separability. Class separability metrics have been well studied in the pattern recognition literature, particularly for the two-class case (Fukunaga, 1990b) [2]. Motivated by Fisher's linear discriminant analysis (LDA), we explore the use of class-separability measures based on the within-class and between-class scatter matrices (Murphy, 2012). The goal of LDA is to find a projection vector (direction) such that data from the two classes are maximally separated and form compact clusters upon projection. Rather than finding an optimal projection direction, we are more interested in ensuring that the concept-score vectors from the detected-ID and detected-OOD data have high separability. Consider the within-class and between-class scatter matrices based on $V_{\text{in}}(\mathbf{C})$ and $V_{\text{out}}(\mathbf{C})$, given by

$$\mathbf{S}_w = \sum_{\mathbf{v} \in V_{\text{in}}(\mathbf{C})} (\mathbf{v} - \boldsymbol{\mu}_{\text{in}})(\mathbf{v} - \boldsymbol{\mu}_{\text{in}})^T + \sum_{\mathbf{v} \in V_{\text{out}}(\mathbf{C})} (\mathbf{v} - \boldsymbol{\mu}_{\text{out}})(\mathbf{v} - \boldsymbol{\mu}_{\text{out}})^T, \quad (3)$$

$$\mathbf{S}_b = (\boldsymbol{\mu}_{\text{out}} - \boldsymbol{\mu}_{\text{in}})(\boldsymbol{\mu}_{\text{out}} - \boldsymbol{\mu}_{\text{in}})^T, \quad (4)$$

[2] In our problem, the two classes correspond to "detected-ID" and "detected-OOD".

where $\boldsymbol{\mu}_{\text{in}}$ and $\boldsymbol{\mu}_{\text{out}}$ are the mean concept-score vectors from $V_{\text{in}}(\mathbf{C})$ and $V_{\text{out}}(\mathbf{C})$ respectively. We define the following separability metric based on the generalized eigenvalue equation solved by Fisher's LDA (Fukunaga, 1990b): $J_{\text{sep}}(\mathbf{C}) := J_{\text{sep}}(V_{\text{in}}(\mathbf{C}), V_{\text{out}}(\mathbf{C})) = \text{tr}\big[\mathbf{S}_w^{-1}\, \mathbf{S}_b\big]$. Maximizing the above metric is equivalent to maximizing the sum of eigenvalues of the matrix $\mathbf{S}_w^{-1}\, \mathbf{S}_b$, which in-turn ensures a large between-class separability and a small within-class separability for the detected-ID and detected-OOD concept scores. We refer to this as a *global concept separability* metric because it does not analyze the separability on a per-class level. The separability metric is closely related to the Bhattacharya distance, which is an upper bound on the Bayes error rate (see Appendix B.1). We define the per-class variations of detection completeness and concept separability in a similar way in Appendix B.2 and B.3.

3.3. Proposed Concept Learning – Key Ideas

Prior Approaches and Limitations. Among post-hoc concept-discovery methods for a DNN classifier with ID data, unlike Kim et al. and Ghorbani et al., that do not support imposing required conditions into the concept discovery, Yeh et al. devised a learning-based approach where classification completeness and the saliency of concepts are optimized via a regularized objective given by

$$\underset{\mathbf{C}, \mathbf{g}}{\text{argmax}} \underset{(\mathbf{x}, y) \sim P_{\text{in}}}{\mathbb{E}} \big[\log h_y(\mathbf{g}(\mathbf{v}_{\mathbf{C}}(\mathbf{x})))\big] + \lambda_{\text{expl}}\, R_{\text{expl}}(\mathbf{C}). \quad (5)$$

Here $\mathbf{C}$ and $\mathbf{g}$ (parameterized by a neural network) are jointly optimized, and $R_{\text{expl}}(\mathbf{C})$ is a regularization term used to ensure that the learned concept vectors have high spatial coherency and low redundancy among themselves (see Yeh et al. (2020) for the definition).

While the objective (5) of Yeh et al. can learn a set of sufficient concepts that have a high classification completeness score, we find that it does not necessarily replicate the per-instance prediction behavior of the classifier in the concept world. Specifically, there can be discrepancies in the reconstructed feature representation, whose effect propagates through the remaining part of the classifier. Since many widely-used OOD detectors rely on the feature representations and/or the classifier's predictions, this discrepancy in the existing concept learning approaches makes it hard to closely replicate the OOD detector in the concept world (see Fig. 3). Furthermore, the scope of Yeh et al. is confined to concept learning for explaining the classifier's predictions based on ID data, and there is no guarantee that the learned concepts would be useful for explaining an OOD detector. To address these gaps, we propose a general method for concept learning that complements prior work by imposing additional instance-level constraints on the concepts, and by considering both the OOD detector and OOD data.

Concept Learning Objective. We define a concept learning objective that aims to find a set of concepts $\mathbf{C}$ and a mapping $\mathbf{g}$ that have the following properties: 1) high detection completeness w.r.t the OOD detector; 2) high classification completeness w.r.t the DNN classifier; and 3) high separability in the concept-score space between detected-ID data and detected-OOD data.

Inspired by recent works on transferring feature information from a teacher model to a student model (Hinton et al., 2015; Zhou et al., 2018b), we encourage accurate reconstruction of $\hat{\mathcal{Z}}$ based on the concept scores by adding a regularization term that is the squared ℓ_2 distance between the original and reconstructed representations $J_{\text{norm}}(\mathbf{C}, \mathbf{g}) = \mathbb{E}_{\mathbf{x}\sim P_{\text{in}}} \|\boldsymbol{\phi}(\mathbf{x}) - \widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}}(\mathbf{x})\|^2$. In order to close the gap between the scores of the OOD detector in the concept world and canonical world on a per-sample level, we introduce the following mean-squared-error (MSE) based regularization:

$$J_{\text{mse}}(\mathbf{C}, \mathbf{g}) = \mathop{\mathbb{E}}_{\mathbf{x}\sim P_{\text{in}}} \big(S(\mathbf{x}, \mathbf{h} \circ \widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}}) - S(\mathbf{x}, \mathbf{f})\big)^2 + \mathop{\mathbb{E}}_{\mathbf{x}\sim P_{\text{out}}} \big(S(\mathbf{x}, \mathbf{h} \circ \widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}}) - S(\mathbf{x}, \mathbf{f})\big)^2. \quad (6)$$

MSE terms are computed with both the ID and OOD data because we want to ensure that the ROC curve corresponding to both the score functions are close to each other (which requires OOD data). Finally, we include a regularization term to maximize the separability metric between the detected-ID and detected-OOD data in the concept-score space, resulting in our final concept learning objective:

$$\underset{\mathbf{C},\mathbf{g}}{\operatorname{argmax}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim P_{\text{in}}} \big[\log h_y(\mathbf{g}(\mathbf{v}_{\mathbf{C}}(\mathbf{x})))\big] + \lambda_{\text{expl}} R_{\text{expl}}(\mathbf{C}) - \lambda_{\text{mse}} J_{\text{mse}}(\mathbf{C}, \mathbf{g}) - \lambda_{\text{norm}} J_{\text{norm}}(\mathbf{C}, \mathbf{g}) + \lambda_{\text{sep}} J_{\text{sep}}(\mathbf{C}). \quad (7)$$

The λ coefficients are non-negative hyper-parameters that are further discussed in Section 4. We note that both $J_{\text{mse}}(\mathbf{C}, \mathbf{g})$ and $J_{\text{sep}}(\mathbf{C})$ depend on the OOD detector [3]. We use the SGD-based Adam optimizer (Kingma & Ba, 2015)) to solve the learning objective. The expectations involved in the objective terms are calculated using sample estimates from the training ID and OOD datasets. Specifically, $D_{\text{in}}^{\text{tr}}$ and $D_{\text{out}}^{\text{tr}}$ are used to compute the expectations over P_{in} and P_{out}, respectively. Our complete concept learning is summarized in Algorithm 1 (Appendix B.4).

4. Experiments

In this section, we conduct experiments to evaluate the proposed method and show that: 1) the learned concepts satisfy the desiderata of completeness and separability across popular off-the-shelf OOD detectors and real-world datasets. 2) the learned concepts can be combined with a Shapley value to provide insightful visual explanations that can help understand the predictions of an OOD detector. The code for our work can be found at `https://github.com/jihyechoi77/concepts-for-ood`.

[3]This dependence may not be obvious for the separability term, but it is clear from its definition.

4.1. Experimental Setup

Datasets. For the ID dataset, we use Animals with Attributes (AwA) (Xian et al., 2018) with 50 animal classes, and split it into a train set (29841 images), validation set (3709 images), and test set (3772 images). We use the MSCOCO dataset (Lin et al., 2014) as the auxiliary OOD dataset $D_{\text{out}}^{\text{tr}}$ for training and validation. For the OOD test dataset $D_{\text{out}}^{\text{te}}$, we follow the literature of large-scale OOD detection (Huang & Li, 2021) and use three different image datasets: `Places365` (Zhou et al., 2017), `SUN` (Xiao et al., 2010), and `Textures` (Cimpoi et al., 2014).

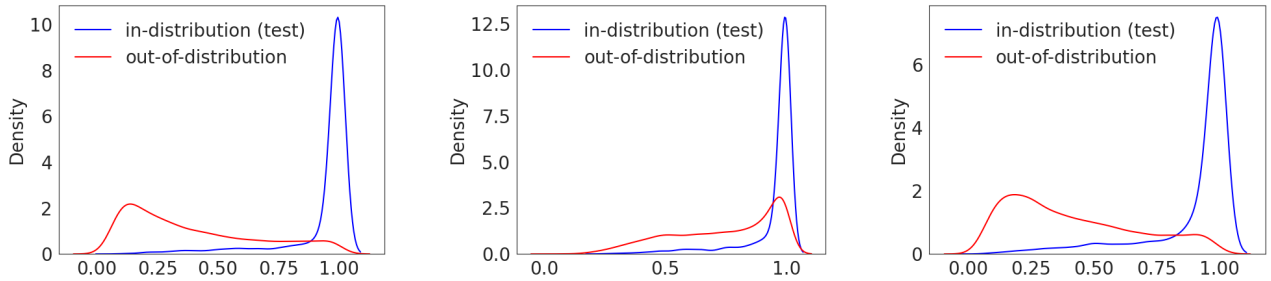
Models. We apply our framework to interpret five prominent OOD detectors from the literature: MSP (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018), Generalized-ODIN (Hsu et al., 2020), Energy (Liu et al., 2020), and Mahalanobis (Lee et al., 2018). The OOD detectors are paired with the widely-used Inception-V3 classifier (Szegedy et al., 2016) (following the setup in prior works (Yeh et al., 2020; Ghorbani et al., 2019; Kim et al., 2018)), which has a test accuracy of 92.13% on the AwA dataset. Additional details on the setup are given in Appendix C.

4.2. Effectiveness of the Concept Learning

Table 1 summarizes the results of concept learning for various combinations of the regularization coefficients ($\lambda_\star$) in Eqn (7), including: **i)** baseline where all the coefficients are set to 0 (first row); **ii)** only the terms directly relevant to detection completeness (*i.e.*, $J_{\text{norm}}(\mathbf{C}, \mathbf{g})$ and $J_{\text{mse}}(\mathbf{C}, \mathbf{g})$) are included (second row); **iii)** only the term responsible for concept separability $J_{\text{sep}}(\mathbf{C})$ is included (third row); **iv)** all the regularization terms are included (fourth row).

From the table, we observe that the regularization terms encourage the learned concepts to satisfy the required desiderata of high completeness and concept separability scores. Having $\lambda_{\text{mse}} > 0$ and $\lambda_{\text{norm}} > 0$ always improves the detection completeness by a large margin (*i.e.*, row 2 compared to row 1), and having $\lambda_{\text{sep}} > 0$ significantly increases the concept separability (*i.e.*, row 3 compared to row 1). Importantly, when all the regularization terms are included, they have the best synergistic effect on the metrics.

Consider the MSP detector for instance. The detection completeness increases from 0.774 to 0.947 with $\lambda_{\text{mse}} = 10, \lambda_{\text{norm}} = 0.1, \lambda_{\text{sep}} = 0$, and the concept separability



(a) Empirical distribution of $S(\mathbf{x}, \mathbf{f})$ from the target detector.

(b) Distribution of $S^{\text{con}}(\mathbf{x}, \mathbf{f})$ using the concepts learned by Yeh et al. (2020).

(c) Distribution of $S^{\text{con}}(\mathbf{x}, \mathbf{f})$ using the concepts learned by our method.

Figure 3: Empirical distribution of: **(a)** MSP detector score $S(\mathbf{x}, \mathbf{f})$ in the canonical world vs. **(b, c)** Reconstructed $S^{\text{con}}(\mathbf{x}, \mathbf{f})$ in the concept world using the concepts learned by Yeh et al. (2020) and our method. Concepts learned by Yeh et al. (2020) have $\eta_{\mathbf{f}} = 0.977$, $\eta_{\mathbf{f},S}(\mathbf{C}) = 0.782$, while the concepts learned by our method ($\lambda_{\text{mse}} = 10, \lambda_{\text{norm}} = 0.1, \lambda_{\text{sep}} = 50$) have $\eta_{\mathbf{f}} = 0.984$, $\eta_{\mathbf{f},S}(\mathbf{C}) = 0.961$. The AwA test set and the `SUN` dataset are used as ID (blue) and OOD (red) respectively.

OOD detector	$(\lambda_{\text{mse}}, \lambda_{\text{norm}}, \lambda_{\text{sep}})$	$\eta_{\mathbf{f}}(\mathbf{C}) \uparrow$	$\eta_{\mathbf{f},S}(\mathbf{C}) \uparrow$	$J_{\text{sep}}(\mathbf{C}, \mathbf{C}') \uparrow$
MSP	$(0, 0, 0)$	0.977 ± 0.0006	0.774 ± 0.0010	0.694 ± 0.0153
	$(10, 0.1, 0)$	$\mathbf{0.994} \pm 0.0004$	$\underline{0.947} \pm 0.0004$	1.892 ± 0.0393
	$(0, 0, 50)$	0.980 ± 0.0005	0.814 ± 0.0008	$\underline{2.533} \pm 0.0714$
	$(10, 0.1, 50)$	$\underline{0.984} \pm 0.0004$	$\mathbf{0.960} \pm 0.0004$	$\mathbf{2.756} \pm 0.0854$
ODIN	$(0, 0, 0)$	0.977 ± 0.0006	0.742 ± 0.0011	0.444 ± 0.0119
	$(10^8, 0.1, 0)$	$\mathbf{0.994} \pm 0.0004$	$\underline{0.951} \pm 0.0004$	1.166 ± 0.0303
	$(0, 0, 50)$	0.987 ± 0.0004	0.899 ± 0.0007	$\underline{1.785} \pm 0.0669$
	$(10^8, 0.1, 50)$	$\underline{0.991} \pm 0.0005$	$\mathbf{0.973} \pm 0.0009$	$\mathbf{1.813} \pm 0.0268$
General-ODIN	$(0, 0, 0)$	0.988 ± 0.0004	0.769 ± 0.0004	0.506 ± 0.0165
	$(10^6, 0.1, 0)$	$\mathbf{0.995} \pm 0.0004$	$\underline{0.951} \pm 0.0006$	1.461 ± 0.0321
	$(0, 0, 50)$	0.981 ± 0.0004	0.859 ± 0.0007	$\underline{1.814} \pm 0.0685$
	$(10^6, 0.1, 50)$	$\underline{0.990} \pm 0.0005$	$\mathbf{0.971} \pm 0.0010$	$\mathbf{1.835} \pm 0.0669$
Energy	$(0, 0, 0)$	0.977 ± 0.0006	0.671 ± 0.0012	0.453 ± 0.0121
	$(1.0.1, 0)$	$\mathbf{0.993} \pm 0.0005$	$\mathbf{0.965} \pm 0.0004$	1.266 ± 0.0319
	$(0, 0, 50)$	$\underline{0.987} \pm 0.0005$	0.779 ± 0.0010	$\mathbf{1.920} \pm 0.0725$
	$(1, 0.1, 50)$	0.980 ± 0.0005	$\underline{0.943} \pm 0.0005$	$\underline{1.839} \pm 0.0662$
Mahalanobis	$(0, 0, 0)$	0.990 ± 0.0007	0.715 ± 0.0011	0.571 ± 0.0110
	$(0.1, 0.1, 0)$	$\mathbf{0.994} \pm 0.0004$	$\underline{0.950} \pm 0.0009$	1.532 ± 0.0351
	$(0, 0, 50)$	0.985 ± 0.0004	0.880 ± 0.0005	$\underline{2.550} \pm 0.0681$
	$(0.1, 0.1, 50)$	$\underline{0.992} \pm 0.0006$	$\mathbf{0.961} \pm 0.0005$	$\mathbf{2.616} \pm 0.0857$

Table 1: Concept learning results with different parameter settings across various OOD detectors, evaluated on AwA test set (ID) and `Places365` (OOD). Hyperparameters are set based on the scale of corresponding regularization terms for a specific choice of the OOD detector. Across the rows (for a given OOD detector), the best value is **boldfaced**, and the second best value is underlined. The 95% confidence intervals are estimated by bootstrapping the test set over 200 trials. Complete results are given in Table 2 in Appendix C.2.

increases from 0.694 to 2.533 with $\lambda_{\text{mse}} = 0, \lambda_{\text{norm}} = 0, \lambda_{\text{sep}} = 50$. However, when all the terms are considered ($\lambda_{\text{mse}} = 10, \lambda_{\text{norm}} = 0.1, \lambda_{\text{sep}} = 50$), we achieve the best result of $\eta_{\mathbf{f},S}(\mathbf{C}) = 0.960$ and $J_{\text{sep}}(\mathbf{C}, \mathbf{C}') = 2.756$. Results on other large-scale OOD datasets and ablation studies on the regularization terms can be found in Appendix C.2.

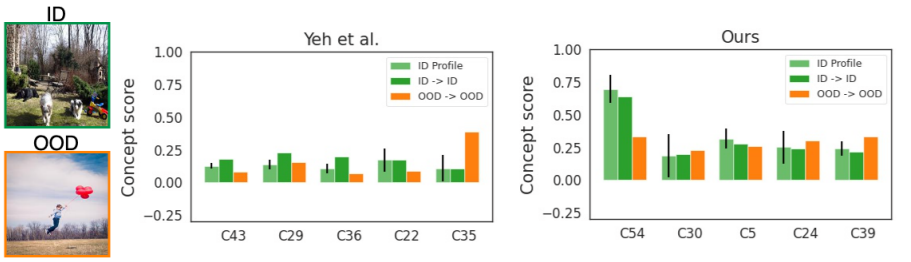
Since the range of the separability score $J_{\text{sep}}(\mathbf{C})$ (or $J^y_{\text{sep}}(\mathbf{C})$) is not well defined, we report a *relative concept separability* score that is easier to interpret, and defined as

$$J_{\text{sep}}(\mathbf{C}, \mathbf{C}') = \text{Median}\left(\left\{\frac{J^y_{\text{sep}}(\mathbf{C}) \; - \; J^y_{\text{sep}}(\mathbf{C}')}{J^y_{\text{sep}}(\mathbf{C}')}\right\}_{y=1}^{L}\right). \tag{8}$$

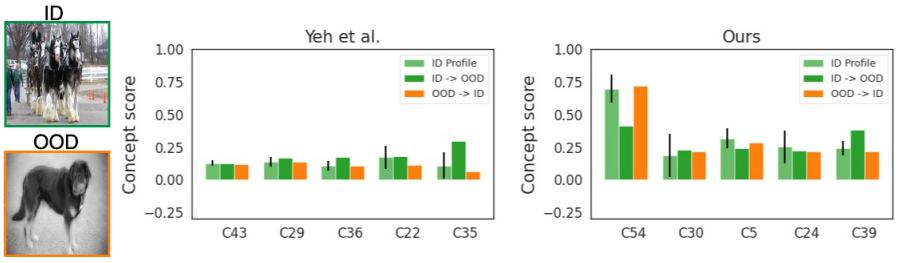
It captures the relative improvement in concept separability using concepts $\mathbf{C}$, compared to a baseline set of concepts $\mathbf{C}'$ obtained by setting $\lambda_{\text{mse}} = \lambda_{\text{norm}} = \lambda_{\text{sep}} = 0$.

Concepts good for the OOD detector are also good for the classifier, but not vice-versa. Recall that the baseline ($\lambda_{\text{mse}} = \lambda_{\text{norm}} = \lambda_{\text{sep}} = 0$) corresponds to the method of Yeh et al. (2020) where only the classifier is considered during the concept learning. For any choice of OOD detector in Table 1 and Table 2 (in Appendix C.2), the concepts learned by our method always achieve higher scores even for classification completeness, compared to the baseline. In contrast, the baseline concepts for only the classifier have the lowest detection completeness and concept separability in all cases. This may not be surprising since the scope of Yeh et al. (2020) does not cover explaining detectors. Nonetheless, such observations provide supporting evidence to motivate the need for our concept learning and novel metrics, indicating that *even if the concepts are sufficient to describe the DNN classifier, the same set of concepts may not be appropriate for the OOD detector.*

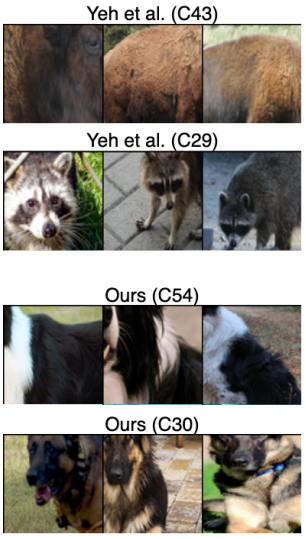
Accurate Reconstruction of Per-sample Behavior. In addition to the above numerical comparisons with respect to the proposed metrics, we found the method of Yeh et al. (2020) to have potential issues in terms of reconstructing the feature representations. This in-turn leads to degraded reconstruction of the *per-sample* behavior of the OOD detector. Comparing Fig. 3a and Fig. 3b, we observe that the concepts of Yeh et al. (2020) lead to a strong mismatch between the score distributions of the OOD detector. In contrast, our method approximates the score distributions more closely (compare Fig. 3a and Fig. 3c). Given that the second



(a) Correct detection: top collie image is correctly detected as ID (dark-green bar), and the bottom image is correctly detected as OOD (orange bar).



(b) Wrong detection: top ID image is detected as OOD (dark-green bar), and the bottom OOD image is detected as ID (orange bar).



(c) Visualization of top-2 important concepts found by the method of Yeh et al. (2020) and our method.

Figure 4: **Concept-based explanations for the Energy OOD detector using concepts learned by Yeh et al. (2020) vs. ours** ($\lambda_{\text{mse}} = 1, \lambda_{\text{norm}} = 0.1, \lambda_{\text{sep}} = 10$). Images are randomly selected from the AwA test set (ID) and `Places` (OOD), and all of them are predicted to the class "Collie". Concept score patterns by Yeh et al. (2020) are not distinctive between detected-ID vs. detected-OOD (*i.e.*, dark-green bar and orange bar are not very different). Whereas, our concepts present very similar patterns to the ID profile (light-green bar) when input is detected as ID, and different from the ID profile when detected as OOD.

half of the classifier and detector remains fixed between the canonical and concept worlds, this observation implies that the reconstructed features fed into the second half of the classifier have to be distorted. Similar observations are made for the Energy detector in Fig. 6 in Appendix C.2. We observe that such inaccurate reconstruction of features poses a similar problem for classifiers as well (more discussion in Appendix C.3). We conclude that the objective of Yeh et al. (2020), which considers only the aggregate statistic of reconstructed accuracy, is not sufficient to recover the per-sample behavior, and augmenting it with our reconstruction error-based regularization term is a straightforward improvement for both the classifier and OOD detector.

4.3. Concept-Based Explanations for OOD Detectors

Finding the Key Concepts. Given a set of learned concepts, we address the question: *how much does each concept contribute to the detection results for inputs predicted to a particular class?* To address this, we follow recent works that have adopted the Shapley value from Game theory literature (Shapley, 1953; Fujimoto et al., 2006) for scoring the importance of a feature subset towards the predictions of a model (Chen et al., 2019; Lundberg & Lee, 2017; Sundararajan & Najmi, 2020). We propose to use our per-class detection completeness metric $\eta^j_{\mathbf{f},S}(\mathbf{C})$ (Eqn. (11) in Appendix B.2) as the characteristic function of the Shapley value. The modified Shapley value of a concept $\mathbf{c}_i \in \mathbf{C}$ with respect to the predicted class $j \in [L]$ is defined as

$$\text{SHAP}(\eta^j_{\mathbf{f},S}, \mathbf{c}_i) := \sum_{\mathbf{C}' \subseteq \mathbf{C} \setminus \{\mathbf{c}_i\}} \frac{\eta^j_{\mathbf{f},S}(\mathbf{C}' \cup \{\mathbf{c}_i\}) - \eta^j_{\mathbf{f},S}(\mathbf{C}')}{m \binom{m-1}{|\mathbf{C}'|}}, \tag{9}$$

where $\mathbf{C}'$ is a subset of $\mathbf{C}$ excluding concept $\mathbf{c}_i$. This Shapley importance score captures the average marginal contribution of concept $\mathbf{c}_i$ towards explaining the decisions of the OOD detector for inputs predicted into class j.

In the rest of the section, we demonstrate how the concepts ranked by the above Shapley importance score can serve as a useful tool for interpreting the OOD detector.

Explaining Detection Errors. Given an OOD detector of interest, we collect inputs that are correctly detected as ID, and average their concept scores (which corresponds to the *ID profile* in Fig. 4). The ID profile quantifies how much each concept matters for the normal ID inputs. Given a test input, either correctly or incorrectly detected, the user could examine how similar or different this input is with respect to the ID concept profile. Fig. 4 illustrates our explanations of the Energy OOD detector's decisions. By visualizing the concepts (see Fig. 4c), we observe that for the predicted class *Collie*, "furry dog skin" (C54) and "oval dog face" (C30) are the key concepts to capture the detector's outputs

to distinguish ID images from OOD images. We also observe that the OOD detector predicts an input as ID when the concept scores show a similar pattern to the ID profile, or predicts an input as OOD when the concept-score pattern is far from the ID profile. For instance, our analysis shows that the bottom input in Fig. 4b is an OOD image from `Places` dataset but detected as ID (false positive) since its score for "furry dog skin" is as high as the usual ID Collie images (which is true in the image). Explaining detection results is crucial for encouraging the adoption of OOD detectors in various decision-making processes. Our example here suggests that certain errors of an OOD detector can be understandable mistakes, which require further reasoning, rather than discarding the model based only on aggregate performance metrics. Additional examples of our concept-based explanations are given in Appendix E.2.

Comparison of Explanations by Yeh et al. and Ours. Lastly, we provide qualitative evidence supporting our argument that: concepts good for the classifier are not necessarily good for the OOD detector. In Fig. 4, given an Energy detector and ID/OOD inputs, we present explanations using concepts learned by Yeh et al. (2020) vs. our method. We observe that Yeh et al. (2020) fails to generate visually-distinguishable explanations between detected-ID and detected-OOD inputs. The separation between the dark-green bars and the orange bars in Fig. 4a and Fig. 4b becomes more visible in our explanations, which enables more intuitive interpretation for human users (this reflects our design goal of concept separability). It is also noteworthy that in Fig. 4c, our concepts that are most important to distinguish ID Collie from OOD Collie (*i.e.*, C54 and C30) are more specific and finer-grained characteristics of Collie, while Yeh et al. (2020) finds concepts that are vaguely similar to the features of a dog, but rather generic (*i.e.*, C43 and C29). This is the reason we require more number of concepts to achieve high detection completeness and concept separability, compared to solely considering the classification completeness [4].

4.4. Explanations For Better OOD Detection.

Our work makes the first effort to reason about the different failure modes of OOD detectors through explanations, rather than just observing aggregated performance metrics (e.g., AUROC or AUPRC). Naturally, the next step would be to utilize such reasoning to modify and improve the OOD detector. We leave the development of concept-based explanations as actionable guidelines for better OOD detection as future work, and describe here a scenario where our explanations can provide direct utility.

[4]In Fig. 4, after concept learning with $m = 100$ and duplicate removal, we found 44 non-redundant concepts for Yeh et al. (2020), and 100 distinct concepts for ours.

We posit that our explanations can provide effective feedback when the failure of the OOD detector originates from misbehavior of the paired classifier, which we confirm to be the most common failure mode of OOD detectors. For instance, consider the top image in Fig. 4b as our input. Its true label is "Horse", but the classifier predicted it to class "Collie". Obviously, the horse image has a different concept-activation pattern from the normal ID Collie profile (compare the dark-green bars with the ID profile of ours in Fig. 4b). To remove such failure cases, the practitioner could identify the key concepts for the prediction of class "Horse" and compare the concept pattern of the input to the normal ID "Horse" profile. It is noteworthy that we can use the same set of concepts here, since our concept-learning objective finds concepts that can effectively explain *both* the classifier and OOD detector. Indeed, we observe that the key concepts for the class "Horse" are "round brown body" and "brown oval face of horse", while the given input is an outlier relative to these concepts. Hence, the practitioner could consider diversifying the training set in the ID "Horse" class to include more examples of horses (*e.g.*, with black and white hair).

5. Conclusion

We develop an unsupervised and human-interpretable explanation method for black-box OOD detectors based on high-level concepts derived from the internal layer representations of a (paired) DNN classifier. We propose novel metrics viz. *detection completeness* and *concept separability* to evaluate the completeness (sufficiency) and quality of the learned concepts for OOD detection. We then propose a concept learning method that is quite general and applies to a broad class of off-the-shelf OOD detectors. Through extensive experiments and qualitative examples, we demonstrate the practical utility of our method for understanding and debugging OOD detectors. We discuss additional aspects of our method such as the auxiliary OOD dataset, human subject study, and societal impact in Appendix A.

Acknowledgements

We thank all the anonymous reviewers for their careful comments and feedback. The work is partially supported by Air Force Grant FA9550-18-1-0166, the National Science Foundation (NSF) Grants CCF-FMitF-1836978, IIS-2008559, SaTC-Frontiers-1804648, CCF-2046710, CCF-1652140, and 2039445, and ARO grant number W911NF-17-1-0405. Choi, Feng, Chen, Jha, and Prakash are partially supported by the DARPA-GARD problem under agreement number 885000. Raghuram is partially supported through the NSF grants CNS-2112562, CNS-2107060, CNS-2003129, CNS-1838733, CNS-1647152, and the US Department of Commerce grant 70NANB21H043.

References

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. Debugging tests for model explanations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/075b051ec3d22dac7b33f788da631fd4-Abstract.html.

Alvarez-Melis, D. and Jaakkola, T. S. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 7786–7795, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Araujo, A., Norris, W., and Sim, J. Computing receptive fields of convolutional neural networks. *Distill*, 2019. doi: 10.23915/distill.00021. https://distill.pub/2019/computing-receptive-fields.

Armstrong, S. L., Gleitman, L. R., and Gleitman, H. What some concepts might not be. *Cognition*, 13(3):263–308, 1983.

Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

Bouchacourt, D. and Denoyer, L. EDUCE: explaining model decisions through unsupervised concepts extraction. *CoRR*, abs/1905.11852, 2019. URL http://arxiv.org/abs/1905.11852.

Cao, S. and Zhang, Z. Deep hybrid models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4733–4743, 2022.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. L-shapley and c-shapley: Efficient model interpretation for structured data. In *7th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2019. URL https://openreview.net/forum?id=S1E3Ko09F7.

Chen, J., Li, Y., Wu, X., Liang, Y., and Jha, S. ATOM: robustifying out-of-distribution detection using outlier mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD*, volume 12977 of *Lecture Notes in Computer Science*, pp. 430–445. Springer, 2021. doi: 10.1007/978-3-030-86523-8_26. URL https://doi.org/10.1007/978-3-030-86523-8_26.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

DeVries, T. and Taylor, G. W. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Fujimoto, K., Kojadinovic, I., and Marichal, J.-L. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55(1):72–99, 2006.

Fukunaga, K. *Introduction to Statistical Pattern Recognition*, chapter 3, pp. 97–103. Academic Press, 2 edition, 1990a.

Fukunaga, K. *Introduction to Statistical Pattern Recognition*, chapter 10, pp. 446–451. Academic Press, 2 edition, 1990b.

Geifman, Y. and El-Yaniv, R. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL http://proceedings.mlr.press/v97/geifman19a.html.

Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 9273–9282, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations (ICLR)*, 2015. URL http://arxiv.org/abs/1412.6572.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Hendrycks, D., Mazeika, M., and Dietterich, T. G. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HyxCxhRcY7.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00823. URL https://doi.org/10.1109/ICCV48922.2021.00823.

Hendrycks, D., Zou, A., Mazeika, M., Tang, L., Li, B., Song, D., and Steinhardt, J. PixMix: Dreamlike pictures comprehensively improve safety measures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16762–16771. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01628. URL https://doi.org/10.1109/CVPR52688.2022.01628.

Hinton, G. E., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://arxiv.org/abs/1503.02531.

Hsu, Y., Shen, Y., Jin, H., and Kira, Z. Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10948–10957. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01096. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Hsu_Generalized_ODIN_Detecting_Out-of-Distribution_Image_Without_Learning_From_Out-of-Distribution_Data_CVPR_2020_paper.html.

Huang, R. and Li, Y. MOS: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8710–8719, June 2021.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning, (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2673–2682. PMLR, 2018. URL http://proceedings.mlr.press/v80/kim18d.html.

Kim, S. S. Y., Meister, N., Ramaswamy, V. V., Fong, R., and Russakovsky, O. HIVE: Evaluating the human interpretability of visual explanations. In *17th European Conference on Computer Vision (ECCV)*, volume 13672 of *Lecture Notes in Computer Science*, pp. 280–298. Springer, 2022. URL https://doi.org/10.1007/978-3-031-19775-8_17.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015. URL http://arxiv.org/abs/1412.6980.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020. URL http://proceedings.mlr.press/v119/koh20a.html.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 7167–7177, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision - ECCV*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Lin, Z., Roy, S. D., and Li, Y. MOOD: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15313–15323, 2021.

Liu, W., Wang, X., Owens, J. D., and Li, Y. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020 (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html.

Lundberg, S. M. and Lee, S. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pp. 4765–4774, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Markoff, J. For now, self-driving cars still need humans. https://www.nytimes.com/2016/01/18/technology/driverless-cars-limits-include-human-nature.html, 2016. The New York Times.

Mohseni, S., Pitale, M., Yadawa, J. B. S., and Wang, Z. Self-supervised learning for generalizable out-of-distribution detection. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pp. 5216–5223. AAAI Press, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/5966.

Mozannar, H. and Sontag, D. A. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL http://proceedings.mlr.press/v119/mozannar20b.html.

Murphy, K. P. *Machine Learning: A Probabilistic Perspective*, chapter 8, pp. 271–274. MIT press, 2012.

Raghuram, J., Chandrasekaran, V., Jha, S., and Banerjee, S. A general framework for detecting anomalous inputs to DNN classifiers. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8764–8775. PMLR, 2021. URL http://proceedings.mlr.press/v139/raghuram21a.html.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Shapley, L. A value for n-person games. *Contributions to the Theory of Games*, (28):307–317, 1953.

Sun, Y., Guo, C., and Li, Y. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 144–157, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/01894d6f048493d2cacde3c579c315a3-Abstract.html.

Sundararajan, M. and Najmi, A. The many Shapley values for model explanation. In *International Conference on Machine Learning*, pp. 9269–9278. PMLR, 2020.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL http://proceedings.mlr.press/v70/sundararajan17a.html.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Tax, D. M. J. and Duin, R. P. W. Growing a multi-class classifier with a reject option. *Pattern Recognition Letters*, 29(10):1565–1570, 2008. doi: 10.1016/j.patrec.2008.03.010. URL https://doi.org/10.1016/j.patrec.2008.03.010.

Tenenbaum, J. B. *A Bayesian framework for concept learning*. PhD thesis, Massachusetts Institute of Technology, 1999.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *CoRR*, abs/2110.11334, 2021. URL https://arxiv.org/abs/2110.11334.

Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., Du, X., Zhou, K., Zhang, W., Hendrycks, D., Li, Y., and Liu, Z. OpenOOD: Benchmarking generalized out-of-distribution detection. In *NeurIPS*, 2022. URL https://proceedings.neurips.cc//paper_files/paper/2022/hash/d201587e3a84fc4761eadc743e9b3f35-Abstract.html.

Yeh, C., Kim, B., Arik, S. Ö., Li, C., Pfister, T., and Ravikumar, P. On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ecb287ff763c169694f682af52c1f309-Abstract.html.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Zhou, B., Sun, Y., Bau, D., and Torralba, A. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 119–134, 2018a.

Zhou, G., Fan, Y., Cui, R., Bian, W., Zhu, X., and Gai, K. Rocket launching: A universal and efficient framework for training well-performing light net. In *Thirty-second AAAI conference on artificial intelligence*, 2018b.

Appendix

In Section A, we discuss additional aspects of our method such as the choice of auxiliary OOD dataset, human subject study, and societal impact. In Section B, we discuss the connection of the proposed concept separability to Bhattacharya Distance, and the per-class variations of detection completeness and concept separability, followed by the overall algorithm for concept learning. In Section C, we provide the detailed setup for the experiments and additional thorough analysis of our concept learning objective. In Section D, we discuss whether our concept learning objective remains effective even when a synthesized auxiliary OOD dataset similar to target ID data is used. In Section E, we illustrate additional examples of our concept-based explanations.

A. Discussion and Societal Impact

Auxiliary OOD Dataset. A limitation of our approach is its requirement of an auxiliary OOD dataset for concept learning, which could be hard to access in certain applications. To overcome that, a research direction would be to design generative models that simulate domain shifts or anomalous behavior and could create the auxiliary OOD dataset synthetically, allowing us additional control on the extent of distributional changes the resulting concepts could deal with (see Appendix D for further discussion).

Human Subject Study. Performing a human-subject (or user) study would be the ultimate way to evaluate the effectiveness of explanations, but remains largely unexplored even for in-distribution classification tasks. We emphasize that designing such a usability test with OOD detectors would be even more challenging due to the characteristics of the OOD detection task, compared to in-indistribution classification tasks. For in-distribution classifiers, users could potentially generate hypotheses about what high-level concepts should attribute to the class prediction, and compare their hypotheses to the provided explanations to determine the classifier's reliability. On the other hand, assessing the reliability of OOD detection involves checking whether a given input belongs to any of the natural distributions of concepts; this is essentially limited to whether users' mental models on such global distributions can be accurately probed via a couple of presented local instances. We believe that designing a thorough probing method for human interpretability on OOD detection would be an interesting yet challenging research quest by itself (Kim et al., 2022) and our paper does not address that.

Societal Impact. Our work helps address the detection results of OOD detectors, giving practitioners the ability to explain the model's decision to invested parties. Our explanations can also be used to keep a data point as an understood mistake by the model rather than throwing it away without further analysis, which could help guide how to improve the OOD detector with respect to the concepts. However, this would also mean that more trust is put back into the human practitioner to not abuse the explanations or misrepresent them.

B. Concept Learning

B.1. Connection to the Bhattacharya Distance

We note that the proposed separability metric in Section 3.2 is closely related to the Bhattacharya distance (Bhattacharyya, 1943) for the special case when the concept scores from both ID and OOD data follow a multivariate Gaussian density. The Bhattacharya distance is a well known measure of divergence between two probability distributions, and it has the nice property of being an upper bound to the Bayes error rate in the two-class case (Fukunaga, 1990a). For the special case when the concept scores from both ID and OOD data follow a multivariate Gaussian with a shared covariance matrix, it can be shown that the Bhattacharya distance reduces to the following separability metric (ignoring scale factors):

$$J_{\text{sep}}(\mathbf{C}) \;:=\; J_{\text{sep}}(V_{\text{in}}(\mathbf{C}), V_{\text{out}}(\mathbf{C})) \;=\; \text{tr}\big[\mathbf{S}_w^{-1}\,\mathbf{S}_b\big]. \tag{10}$$

B.2. Per-class Detection Completeness

We propose a per-class measure for the detection completeness (denoted by $\eta^{y}_{\mathbf{f},S}(\mathbf{C})$), which is obtained by simply modifying $\eta_{\mathbf{f},S}(\mathbf{C})$ in Eqn. (2) based on the subset of ID and OOD data that are predicted into class $y \in [L]$ by the classifier.

Definition 3. Given a trained DNN classifier $\mathbf{f} = \mathbf{h} \circ \boldsymbol{\phi}$, a trained OOD detector with score function $S(\mathbf{x}, \mathbf{f})$, and a set of concept vectors $\mathbf{C}$, the *per-class detection completeness* relative to class $y \in [L]$ with respect to the ID distribution $P_{\text{in}}(\mathbf{x}, y)$

and OOD distribution $P_{\text{out}}(\mathbf{x})$ is defined as

$$\eta^{y}_{\mathbf{f},S}(\mathbf{C}) \;:=\; \frac{\sup_{\mathbf{g}} \text{AUC}^{y}(\mathbf{h} \circ \widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}}) \;-\; b_r}{\text{AUC}(\mathbf{h} \circ \boldsymbol{\phi}) \;-\; b_r}, \tag{11}$$

where $\text{AUC}^{y}(\mathbf{h} \circ \widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}})$ is the AUROC of the detector conditioned on the event that the class predicted by the concept-world classifier $\mathbf{h} \circ \widehat{\boldsymbol{\phi}}_{\mathbf{g},\mathbf{C}}$ is y. We note that the denominator in the above metric is still the global AUROC. The baseline AUROC b_r is equal to 0.5. This per-class detection completeness is used in the modified Shapley value defined in Section 4.3.

B.3. Per-class Concept Separability

In section 3.2, we focused on the separability between the concept scores of ID and OOD data without considering the class prediction of the classifier. However, it would be more appropriate to impose a high separability between the concept scores on a per-class level. In other words, we would like the concept scores of detected-ID and detected-OOD data, that are predicted by the classifier into any given class $y \in [L]$ to be well separated. Consider the set of concept-score vectors from the detected-ID (or detected-OOD) dataset that are also predicted into class y:

$$\begin{aligned} V^{y}_{\text{in}}(\mathbf{C}) \;&:=\; \{\widetilde{\mathbf{v}}_{\mathbf{C}}(\mathbf{x}),\; \mathbf{x} \in D^{\text{tr}}_{\text{in}} \cup D^{\text{tr}}_{\text{out}} \;:\; \mathcal{D}_{\gamma}(\mathbf{x},\mathbf{f}) = 1 \;\text{ and }\; \widehat{y}(\mathbf{x}) = y\} \\ V^{y}_{\text{out}}(\mathbf{C}) \;&:=\; \{\widetilde{\mathbf{v}}_{\mathbf{C}}(\mathbf{x}),\; \mathbf{x} \in D^{\text{tr}}_{\text{in}} \cup D^{\text{tr}}_{\text{out}} \;:\; \mathcal{D}_{\gamma}(\mathbf{x},\mathbf{f}) = 0 \;\text{ and }\; \widehat{y}(\mathbf{x}) = y\}. \end{aligned} \tag{12}$$

We can extend the definition of the global separability metric in Eq. (10) to a given predicted class $y \in [L]$ as follows

$$\begin{aligned} J^{y}_{\text{sep}}(\mathbf{C}) \;&:=\; J_{\text{sep}}(V^{y}_{\text{in}}(\mathbf{C}), V^{y}_{\text{out}}(\mathbf{C})) \;=\; \text{tr}\big[(\mathbf{S}^{y}_{w})^{-1}\,\mathbf{S}^{y}_{b}\big] \\ &=\; (\boldsymbol{\mu}^{y}_{\text{out}} \;-\; \boldsymbol{\mu}^{y}_{\text{in}})^{T}\,(\mathbf{S}^{y}_{w})^{-1}\,(\boldsymbol{\mu}^{y}_{\text{out}} \;-\; \boldsymbol{\mu}^{y}_{\text{in}}). \end{aligned} \tag{13}$$

We refer to these per-class variations as *per-class concept separability*. The scatter matrices $\mathbf{S}^{y}_{w}$ and $\mathbf{S}^{y}_{b}$ are defined similar to Eq. (3), using the per-class subset of concept scores $V^{y}_{\text{in}}(\mathbf{C})$ or $V^{y}_{\text{out}}(\mathbf{C})$, and the mean concept-score vectors from the detected-ID and detected-OOD dataset are also defined at a per-class level.

B.4. Algorithm for Concept Learning

To provide the readers with a clear overview of the proposed concept learning approach, we include Algorithm 1. Note that in line 7 of Algorithm 1, the dimension reduction step in $V_{\text{in}}(\mathbf{C}) = \{\widetilde{\mathbf{v}}_{\mathbf{C}}(\mathbf{x}),\; \mathbf{x} \in D^{\text{tr}}_{\text{in}} \cup D^{\text{tr}}_{\text{out}} \;:\; \mathcal{D}_{\gamma}(\mathbf{x},\mathbf{f}) = 1\}$ and $V_{\text{out}}(\mathbf{C}) = \{\widetilde{\mathbf{v}}_{\mathbf{C}}(\mathbf{x}),\; \mathbf{x} \in D^{\text{tr}}_{\text{in}} \cup D^{\text{tr}}_{\text{out}} \;:\; \mathcal{D}_{\gamma}(\mathbf{x},\mathbf{f}) = 0\}$ involves the maximum function, which is not differentiable; specifically, the step $\widetilde{v}_{\mathbf{c}_i}(\mathbf{x}) = \max_{p,q} |\langle \boldsymbol{\phi}^{p,q}(\mathbf{x}), \mathbf{c}_i \rangle|$. For calculating the gradients (backward pass), we use the `log-sum-exp` function with a temperature parameter to get a differentiable approximation of the maximum function, *i.e.*, $\max_{p,q} |\langle \boldsymbol{\phi}^{p,q}(\mathbf{x}), \mathbf{c}_i \rangle| \approx \alpha \log \left[\sum_{p,q} \exp\left(\frac{1}{\alpha} |\langle \boldsymbol{\phi}^{p,q}(\mathbf{x}), \mathbf{c}_i \rangle|\right)\right]$ as $\alpha \to 0$. In our experiments, we set the temperature constant $\alpha = 0.001$ upon checking that the approximate value of $\widetilde{v}_{\mathbf{c}_i}(\mathbf{x})$ is sufficiently close to the original value using the maximum function.

C. Implementation Details

We ran all our experiments with Tensorflow, Keras and NVDIA GeForce RTX 2080Ti GPUs. We used test-set bootstrapping with 200 runs to obtain the confidence interval for each hyperparameter setting of concept learning.

C.1. Experimental Setting.

OOD Datasets. For the auxiliary OOD dataset for concept learning ($D^{\text{tr}}_{\text{out}}$), we use the unlabeled images from MSCOCO dataset (120K images in total) (Lin et al., 2014). We carefully curate the dataset to make sure that no images contain overlapping animal objects with our ID dataset (*i.e.*, 50 animal classes of Animals-with-Attributes (Xian et al., 2018)), then randomly sample 30K images. For OOD datasets for evaluation ($D^{\text{te}}_{\text{out}}$), we use the high-resolution image datasets processed by Huang and Li (Huang & Li, 2021).

Hyperparameters for Concept Learning. Throughout the experiments, we fix the number of concepts to $m = 100$ (unless specifically mentioned otherwise), and following the implementation of (Yeh et al., 2020), we set $\lambda_{\text{expl}} = 10$ and $\mathbf{g}$ to be a two-layer fully-connected neural network with 500 neurons in the hidden layer. We learn concepts based on feature

Algorithm 1 Learning concepts for OOD detector

INPUT: Entire training set $D^{\text{tr}} = \{D^{\text{tr}}_{\text{in}}, D^{\text{tr}}_{\text{out}}\}$, entire validation set $D^{\text{val}} = \{D^{\text{val}}_{\text{in}}, D^{\text{val}}_{\text{out}}\}$, classifier $\mathbf{f}$, detector $\mathcal{D}_\gamma$.
INITIALIZE: Concept vectors $\mathbf{C} = [\mathbf{c}_1 \cdots \mathbf{c}_m]$ and parameters of the network $\mathbf{g}$.
OUTPUT: $\mathbf{C}$ and $\mathbf{g}$.

Calculate threshold γ for $\mathcal{D}_\gamma$ using D^{val} as the score at which true positive rate is 95%.
for $t = 1, ...T$ epochs **do**
    Compute the prediction accuracy of the concept-world classifier $\mathbf{f}^{\text{con}}$ using $D^{\text{tr}}_{\text{in}}$.
    Compute the explainability regularization term as defined in (Yeh et al., 2020).
    Compute difference of feature representation between canonical world and concept world (i.e. $J_{\text{norm}}(\mathbf{C}, \mathbf{g})$).
    Compute difference of detector outputs between canonical world and concept world using Eqn. (6).
    Compute $V_{\text{in}}(\mathbf{C})$ and $V_{\text{out}}(\mathbf{C})$ using $D^{\text{tr}}, \mathcal{D}_\gamma$ and $\mathbf{C}$.
    Compute separability between $V_{\text{in}}(\mathbf{C})$ and $V_{\text{out}}(\mathbf{C})$ using Eqn. (10) or Eqn. (13).
    Perform a batch-SGD update of $\mathbf{C}$ and g using Eqn. (7) as the objective.
end for

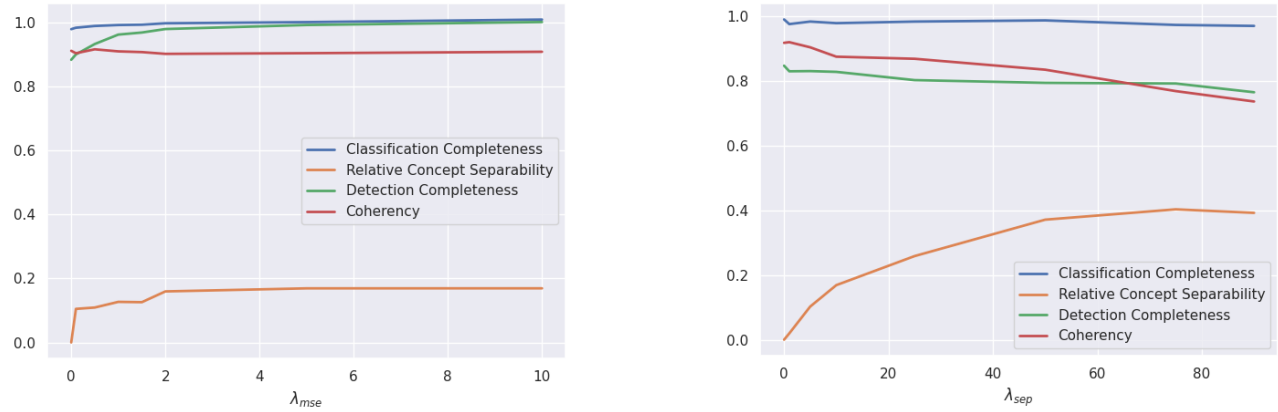
representations from the layer right before the global max-pooling layer of the Inception-V3 model. After concept learning with m concepts, we remove any duplicate (redundant) concept vectors by removing those with a dot product larger than 0.95 with the remaining concept vectors (Yeh et al., 2020).

C.2. Additional Results on the Effectiveness of Our Concept Learning

Ablation Study for Concept Learning. We perform an ablation study that isolates the effect of each regularization term in our concept learning objective (Eqn. 7) towards our evaluation metrics: classification completeness, detection completeness, and relative concept separability. We also observe the coherency among the learned concepts by varying λ_{mse} and λ_{sep}. Coherency of concepts was introduced by Ghorbani *et al.* (Ghorbani et al., 2019) to ensure that the generated concept-based explanations are understandable to humans. It captures the idea that the examples for a concept should be similar to each other, while being different from the examples corresponding to other concepts. For the specific case of the image domain, the receptive fields most correlated to a concept i (*e.g.*, "stripe pattern") should look different from the receptive fields for a different concept j (*e.g.*, "wavy surface of sea"). Yeh et al. (2020) proposed to quantify the coherency of concepts as

$$\frac{1}{m\,K} \sum_{i=1}^{m} \sum_{\mathbf{x}' \in T_{\mathbf{c}_i}} \langle \phi(\mathbf{x}'), \mathbf{c}_i \rangle, \tag{14}$$

where $T_{\mathbf{c}_i}$ is the set of K-nearest neighbor patches of the concept vector $\mathbf{c}_i$ from the ID training set $D^{\text{tr}}_{\text{in}}$.



(a) Ablation study varying λ_{mse}; we set $\lambda_{\text{norm}} = 0.1, \lambda_{\text{sep}} = 0$

(b) Ablation study varying λ_{sep}; we set $\lambda_{\text{mse}} = 0, \lambda_{\text{norm}} = 0$.

Figure 5: **Ablation study with respect to $J_{\text{mse}}(\mathbf{C}, \mathbf{g})$ and $J_{\text{sep}}(\mathbf{C})$.** We fix $m = 100, \lambda_{\text{expl}} = 10$, and the OOD detector used for concept learing and evaluation is Energy (Liu et al., 2020)

We use this metric to quantify how understandable our concepts are for different hyperparameter choices. Figure 5 shows that aligned with our intuition, large λ_{mse} helps to improve the detection completeness. Having non-zero λ_{mse} is also helpful to improve the classification completeness even further, and surprisingly concept separability as well, without sacrificing the coherency of concepts. On the other hand, on the right side of Figure 5, we observe that large relative concept separability with large λ_{sep} comes at the expense of lower detection completeness and coherency. Recall that when visualizing what each concept represents for human's convenience, we apply threshold 0.8 to only presents (see Figure 9). Low coherency with respect to Eqn. 14 (*i.e.*, 0.768 with $\lambda_{\text{sep}} = 75$) means that there is much less number of examples that can pass the threshold, meaning that users can hardly understand what the concepts at hand entail. This observation suggests that one needs to balance between concept coherency and concept separability depending on which property would be more useful for a specific application of concepts.

Effectiveness of the Concept Learning. In Table 2, we present the complete results of concept learning for various combinations of the regularization coefficients across various real-world, large-scale OOD data: `Places`, `SUN` and `Textures`.

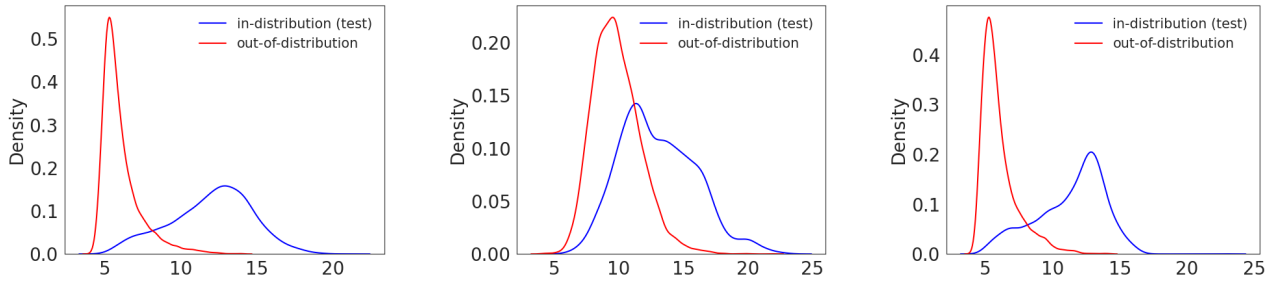
OOD detector	Hyper-parameters	$\eta_{\mathbf{f}}(\mathbf{C}) \uparrow$	Test OOD dataset					
			`Places`		`SUN`		`Textures`	
			$\eta_{\mathbf{f},S}(\mathbf{C}) \uparrow$	$J_{\text{sep}}(\mathbf{C}, \mathbf{C}') \uparrow$	$\eta_{\mathbf{f},S}(\mathbf{C}) \uparrow$	$J_{\text{sep}}(\mathbf{C}, \mathbf{C}') \uparrow$	$\eta_{\mathbf{f},S}(\mathbf{C}) \uparrow$	$J_{\text{sep}}(\mathbf{C}, \mathbf{C}') \uparrow$
MSP	$(0, 0, 0)$	0.977 ± 0.0006	0.774 ± 0.0010	0.694 ± 0.0153	0.782 ± 0.0010	1.088 ± 0.0175	0.593 ± 0.0013	0.765 ± 0.0157
	$(10, 0.1, 0)$	$\mathbf{0.994} \pm 0.0004$	$\underline{0.947} \pm 0.0004$	1.892 ± 0.0393	$\underline{0.946} \pm 0.0004$	3.074 ± 0.0531	$\underline{0.920} \pm 0.0005$	$\underline{3.577} \pm 0.1292$
	$(0, 0, 50)$	0.980 ± 0.0005	0.814 ± 0.0008	$\underline{2.533} \pm 0.0714$	0.816 ± 0.0009	$\underline{4.295} \pm 0.1048$	0.773 ± 0.0010	3.147 ± 0.2076
	$(10, 0.1, 50)$	$\underline{0.984} \pm 0.0004$	$\mathbf{0.960} \pm 0.0004$	$\mathbf{2.756} \pm 0.0854$	$\mathbf{0.961} \pm 0.0005$	$\mathbf{4.442} \pm 0.0830$	$\mathbf{0.937} \pm 0.0004$	$\mathbf{3.587} \pm 0.2145$
ODIN	$(0, 0, 0)$	0.977 ± 0.0006	0.742 ± 0.0011	0.444 ± 0.0119	0.745 ± 0.0010	0.710 ± 0.0156	0.618 ± 0.0013	0.501 ± 0.0121
	$(10^8, 0.1, 0)$	$\mathbf{0.994} \pm 0.0004$	$\underline{0.951} \pm 0.0004$	1.166 ± 0.0303	$\underline{0.958} \pm 0.0004$	2.135 ± 0.0450	$\underline{0.934} \pm 0.0004$	2.793 ± 0.0865
	$(0, 0, 50)$	0.987 ± 0.0004	0.899 ± 0.0007	$\underline{1.785} \pm 0.0669$	0.911 ± 0.0006	$\underline{3.814} \pm 0.0768$	0.793 ± 0.0008	$\underline{3.046} \pm 0.2845$
	$(10^8, 0.1, 50)$	$\underline{0.991} \pm 0.0005$	$\mathbf{0.973} \pm 0.0009$	$\mathbf{1.813} \pm 0.0268$	$\mathbf{0.969} \pm 0.0010$	$\mathbf{4.000} \pm 0.0094$	$\mathbf{0.945} \pm 0.0006$	$\mathbf{3.662} \pm 0.1005$
General-ODIN	$(0, 0, 0)$	0.988 ± 0.0004	0.769 ± 0.0004	0.506 ± 0.0165	0.719 ± 0.0014	0.816 ± 0.0192	0.605 ± 0.0013	0.558 ± 0.1683
	$(10^6, 0.1, 0)$	$\mathbf{0.995} \pm 0.0004$	$\underline{0.951} \pm 0.0006$	1.461 ± 0.0321	$\underline{0.960} \pm 0.0005$	3.007 ± 0.0316	$\underline{0.940} \pm 0.0008$	2.619 ± 0.1077
	$(0, 0, 50)$	0.981 ± 0.0004	0.859 ± 0.0007	$\underline{1.814} \pm 0.0685$	0.803 ± 0.0006	$\underline{4.204} \pm 0.0159$	0.826 ± 0.0008	$\mathbf{4.014} \pm 0.2246$
	$(10^6, 0.1, 50)$	$\underline{0.990} \pm 0.0005$	$\mathbf{0.971} \pm 0.0010$	$\mathbf{1.835} \pm 0.0669$	$\mathbf{0.963} \pm 0.0004$	$\mathbf{4.287} \pm 0.0284$	$\mathbf{0.951} \pm 0.0005$	$\underline{3.695} \pm 0.1921$
Energy	$(0, 0, 0)$	0.977 ± 0.0006	0.671 ± 0.0012	0.453 ± 0.0121	0.682 ± 0.0012	0.675 ± 0.0148	0.557 ± 0.0014	0.521 ± 0.0131
	$(1.0.1, 0)$	$\mathbf{0.993} \pm 0.0005$	$\mathbf{0.965} \pm 0.0004$	1.266 ± 0.0319	$\mathbf{0.963} \pm 0.0004$	2.125 ± 0.0413	$\mathbf{0.960} \pm 0.0003$	2.648 ± 0.0596
	$(0, 0, 50)$	$\underline{0.987} \pm 0.0005$	0.779 ± 0.0010	$\mathbf{1.920} \pm 0.0725$	0.793 ± 0.0009	$\mathbf{3.659} \pm 0.0659$	0.767 ± 0.0010	$\mathbf{4.397} \pm 0.2165$
	$(1, 0.1, 50)$	0.980 ± 0.0005	$\underline{0.943} \pm 0.0005$	$\underline{1.839} \pm 0.0662$	$\underline{0.941} \pm 0.0005$	$\underline{3.421} \pm 0.0619$	$\underline{0.936} \pm 0.0005$	$\underline{3.917} \pm 0.1691$
Mahala-nobis	$(0, 0, 0)$	0.990 ± 0.0007	0.715 ± 0.0011	0.571 ± 0.0110	0.736 ± 0.0011	0.822 ± 0.0165	0.591 ± 0.0011	0.564 ± 0.0203
	$(0.1, 0.1, 0)$	$\mathbf{0.994} \pm 0.0004$	$\underline{0.950} \pm 0.0009$	1.532 ± 0.0351	$\underline{0.960} \pm 0.0010$	2.276 ± 0.0466	$\underline{0.938} \pm 0.0004$	2.915 ± 0.1132
	$(0, 0, 50)$	0.985 ± 0.0004	0.880 ± 0.0005	$\underline{2.550} \pm 0.0681$	0.883 ± 0.0006	$\underline{4.091} \pm 0.1013$	0.774 ± 0.0007	$\underline{4.274} \pm 0.2305$
	$(0.1, 0.1, 50)$	$\underline{0.992} \pm 0.0006$	$\mathbf{0.961} \pm 0.0005$	$\mathbf{2.616} \pm 0.0857$	$\mathbf{0.966} \pm 0.0005$	$\mathbf{4.325} \pm 0.0055$	$\mathbf{0.949} \pm 0.0003$	$\mathbf{4.308} \pm 0.2011$

Table 2: **Results of concept learning with different parameter settings across various OOD detectors and test OOD datasets.** Hyperparameters are in the order of $(\lambda_{\text{mse}}, \lambda_{\text{norm}}, \lambda_{\text{sep}})$. Across the rows (for a given OOD detector and OOD dataset), the best value is **boldfaced**, and second best value is <u>underscored</u>. The 95% confidence intervals are estimated by bootstrapping the test set over 200 trials.

Accurate Reconstruction of OOD Scores In addition to Fig. 3, where we compared the reconstruction accuracy of OOD scores using concepts by Yeh et al. (2020) and ours, Fig. 6 confirms that the same observation also applies to the Energy detector.

C.3. Accurate Reconstruction of Classifier Outputs

We have performed additional experiments to understand if the proposed method can provide improvements in the classification setting. Let $\mathbf{C}_1$ denote the concept matrix learned by the method of Yeh et al. (2020). Let $\mathbf{C}_2$ denote the concept matrix learned by our method with $\lambda_{mse} = \lambda_{sep} = 0$ and $\lambda_{norm} = 0.1$ (set based on the scale of the regularization term J_{norm}). The idea is that we exclude the terms in the concept-learning objective (Eqn. 7) that depend on the OOD detector, but include the ℓ_2 norm based reconstruction error of the layer representation. To evaluate the utility of these two sets of concepts for classification, we calculated the per-sample Hellinger distance between the predicted class probabilities of the original classifier and the concept-world classifier (based on either $\mathbf{C}_1$ or $\mathbf{C}_2$). Fig. 7 compares the empirical distribution of the Hellinger distance for both sets of concepts $\mathbf{C}_1$ and $\mathbf{C}_2$. We observe that the distribution is



(a) Empirical distribution of $S(\mathbf{x}, \mathbf{f})$ from the target detector.

(b) Distribution of $S^{\text{con}}(\mathbf{x}, \mathbf{f})$ using concepts learned by Yeh et al. (2020).

(c) Distribution of $S^{\text{con}}(\mathbf{x}, \mathbf{f})$ using concepts learned by our method.

Figure 6: (a) Energy detector score $S(\mathbf{x}, \mathbf{f})$ in the canonical world vs. (b, c) reconstructed $S^{\text{con}}(\mathbf{x}, \mathbf{f})$ in the concept world, using different set of concepts. Concepts by Yeh et al. (2020) have $\eta_{\mathbf{f}} = 0.977$, $\eta_{\mathbf{f},S}(\mathbf{C}) = 0.682$, while concepts by ours ($\lambda_{\text{mse}} = 1, \lambda_{\text{norm}} = 0.1, \lambda_{\text{sep}} = 50$) have $\eta_{\mathbf{f}} = 0.984$, $\eta_{\mathbf{f},S}(\mathbf{C}) = 0.941$. Comparison is made between AwA test set (ID, blue) vs. SUN (OOD, red).

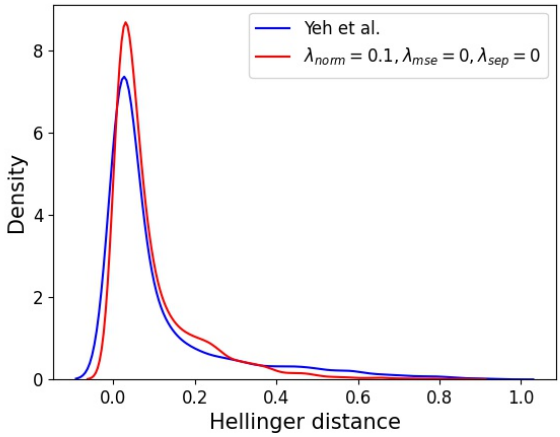


Figure 7: Examples for correct detection

more skewed towards zero with a higher density near zero and a shorter (right) tail in the case of $\mathbf{C}_2$ (red curve) compared to $\mathbf{C}_1$ (blue curve). This suggests that the class predictions are more accurately reconstructed by the concepts learned using our method with only the reconstruction error-based regularization. This can in-turn benefit the concept-based explanations for the classifier.

D. Choice of Auxiliary OOD Dataset in Concept Learning

Under circumstances where having access to auxiliary OOD dataset for concept learning is not feasible, we suggest that one could use generative methods to generate synthetic dataset, or apply data augmentation techniques. Fig. 8 shows an example of AwA image augmented by Hendrycks et al. (2022).

We evaluate the effectiveness of our concept learning objective when such augmented AwA train set is used as auxiliary OOD dataset. Table 3 illustrates that the generated concepts with augmented AwA (*i.e.*, OOD data close to target ID data) have comparable detection completeness and concept separability compared to when MSCOCO (*i.e.*, OOD data far from ID data) was used. But still, further evaluation on generated concept-based explanations with different choice of auxiliary OOD dataset remains as an interesting research question.

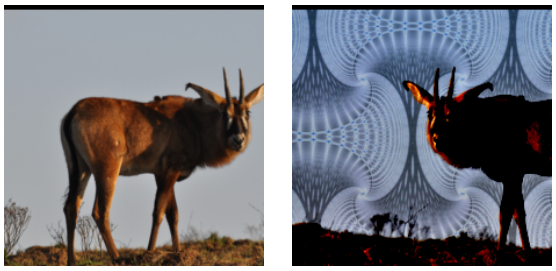


Figure 8: **Random example of augmented AwA dataset. Left:** original image in AwA train set. **Right:** corresponding image augmented using the method of Hendrycks et al. (2022).

OOD detector	Hyper-parameters	$\eta_{\mathbf{f}}(\mathbf{C}) \uparrow$	Test OOD dataset					
			Places		SUN		Textures	
			$\eta_{\mathbf{f},S}(\mathbf{C}) \uparrow$	$J_{\text{sep}}(\mathbf{C}, \mathbf{C}') \uparrow$	$\eta_{\mathbf{f},S}(\mathbf{C}) \uparrow$	$J_{\text{sep}}(\mathbf{C}, \mathbf{C}') \uparrow$	$\eta_{\mathbf{f},S}(\mathbf{C}) \uparrow$	$J_{\text{sep}}(\mathbf{C}, \mathbf{C}') \uparrow$
Energy	$(1, 0.1, 50)$	0.955 ± 0.0006	0.940 ± 0.0005	1.746 ± 0.0712	0.9410 ± 0.0005	3.0703 ± 0.0580	0.927 ± 0.0005	3.417 ± 0.1419

Table 3: Results of concept learning with augmented AwA train set as auxiliary OOD in concept learning.

E. Explanations

E.1. Important Concepts for Each OOD Detector

We show additional examples for the top-ranked concepts by SHAP$(\eta_{\mathbf{f},S}, \mathbf{c}_i)$ in Fig. 9. For each figure with a fixed choice of class prediction, we present receptive fields from ID test set corresponding to top concepts that contribute the most to the decisions of each OOD detector. All receptive fields passed the threshold test that the inner product between the feature representation and the corresponding concept vector is over 0.85.

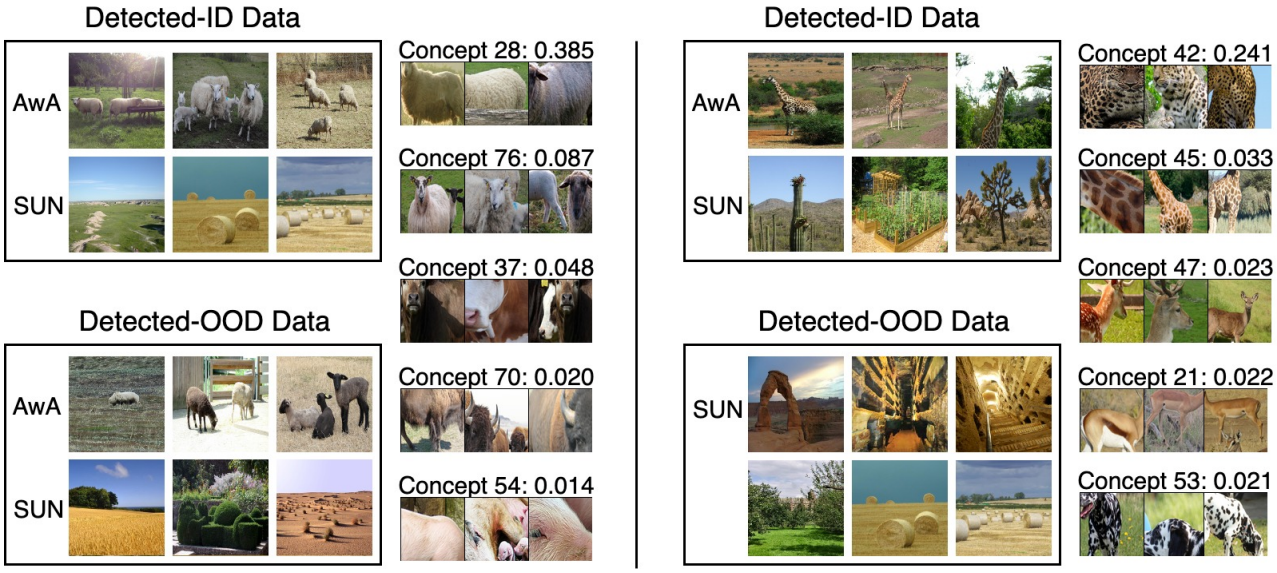


Figure 9: Top-6 important concepts for the Energy OOD detector with respect to class "Sheep" (on the left) and class "Giraffe" (on the right).

Moreover, in Fig. 10, we compare the important concepts discovered by the baseline method (Yeh et al., 2020) (denoted as "baseline") vs. ours. With the baseline, when the learned concepts are solely intended for reconstructing the behavior of the classifier, we observe that interpretation of both the classifier and OOD detector depends on a common set of concepts

(*i.e.*, concepts 32, 10, and 47). On the other hand, the concepts learned by our method focus on reconstructing the behavior of both the OOD detector and the classifier. In this case, we observe that a distinct set of important concepts are selected for classification and OOD detection. We also observe that our method requires more concepts in order to address the decisions of both the classifier and OOD detector. For instance, the number of concepts obtained by our method and the baseline are 78 and 53 (respectively), out of a total 100 concepts after the duplicate removal of concept vectors. In short, when the concepts are only targeted at explaining the DNN classifier (as in the baseline (Yeh et al., 2020)), the behavior of the OOD detector is merely described by the common set of concepts that are important for the DNN classifier. On the other hand, when not only the DNN classifier but also the OOD detector is taken into consideration during concept learning (*i.e.*, our method), we obtain a more diverse and expanded set of concepts, and different concepts play a major role in interpreting the classification and detection results.

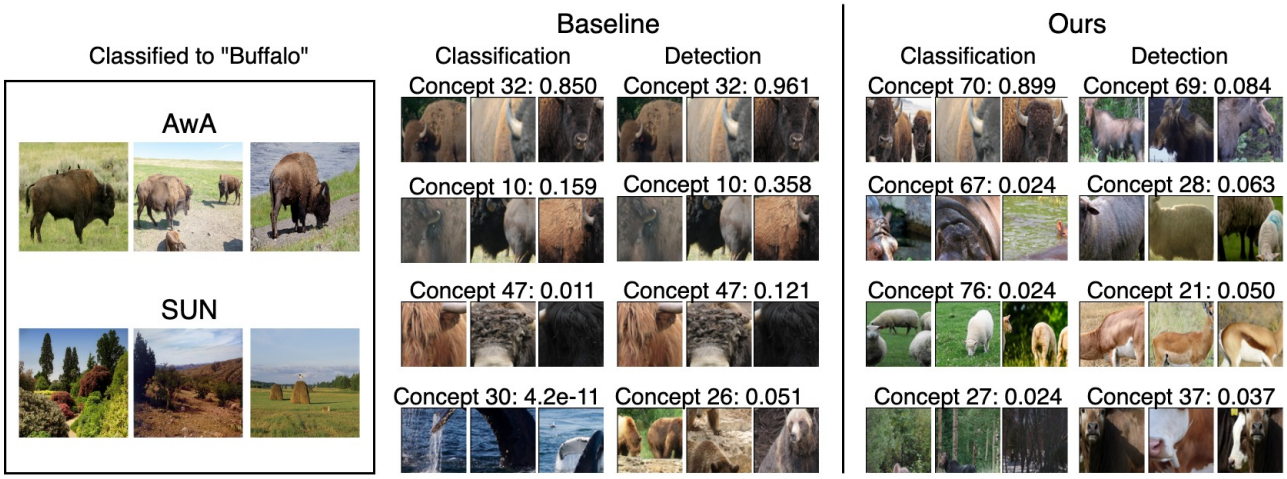
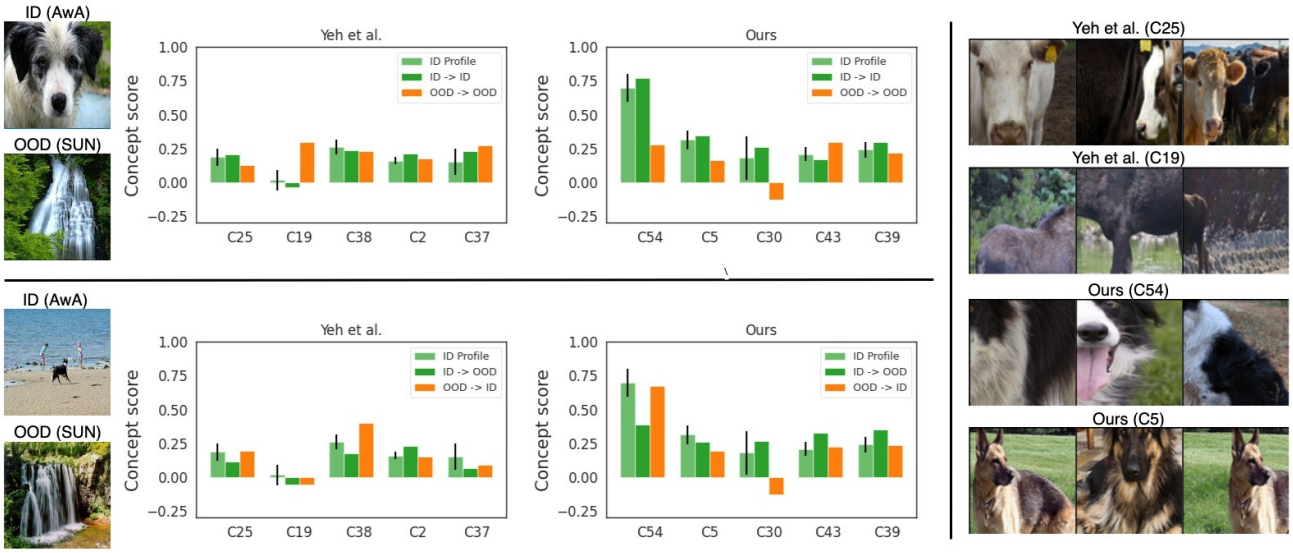


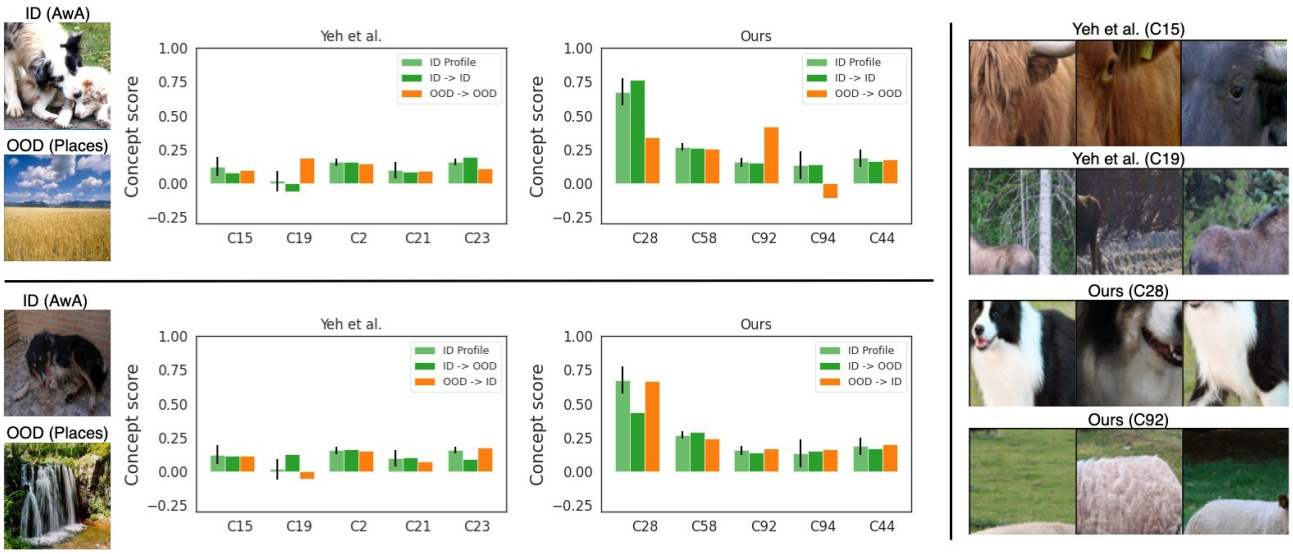
Figure 10: **Most important concepts for the Energy detector with respect to the predicted class "Buffalo".** We demonstrate randomly sampled images that are predicted by the classifier into this class. We compare the top-4 important concepts to describe the DNN classifier (and Energy detector), ranked by the Shapley value based on classification completeness $\text{SHAP}(\eta^j_{\mathbf{f}}, \mathbf{c}_i)$ (and detection completeness $\text{SHAP}(\eta^j_{\mathbf{f},S}, \mathbf{c}_i)$). "Baseline" corresponds to the case when the concepts are learned with $\lambda_{\text{mse}} = \lambda_{\text{norm}} = \lambda_{\text{sep}} = 0$, whereas "Ours" corresponds to the concepts learned with $\lambda_{\text{mse}} = 1, \lambda_{\text{norm}} = 0.1, \lambda_{\text{sep}} = 0$.

E.2. More Examples of Our Concept-Based Explanation

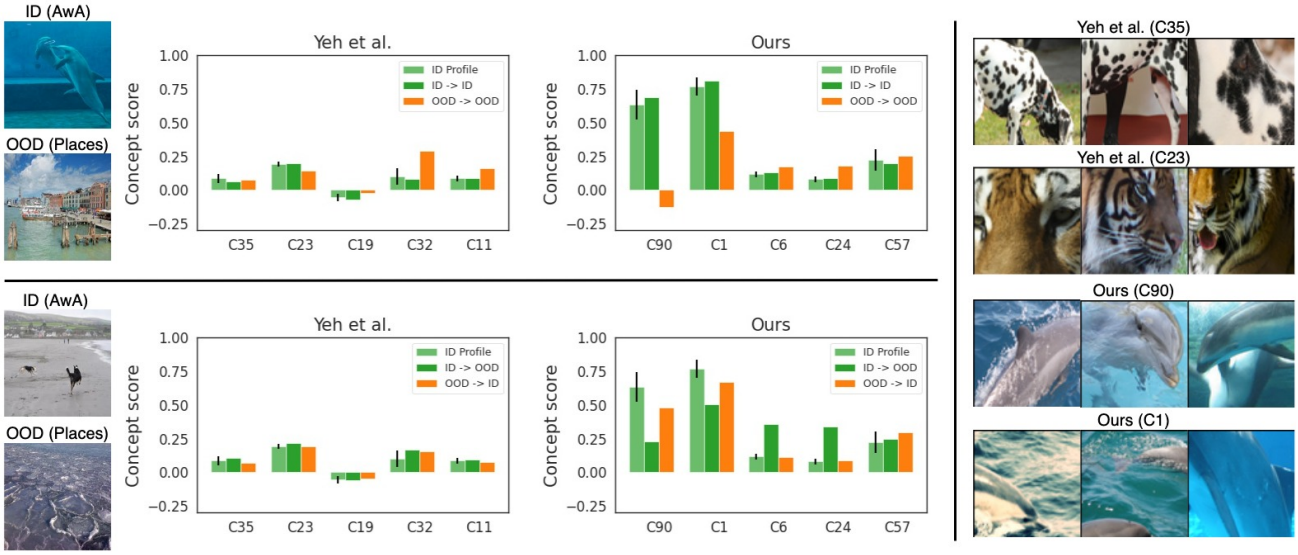
In Fig. 11, we provide additional examples of the concept-based explanations provided by our method and compare it with that of (Yeh et al., 2020).



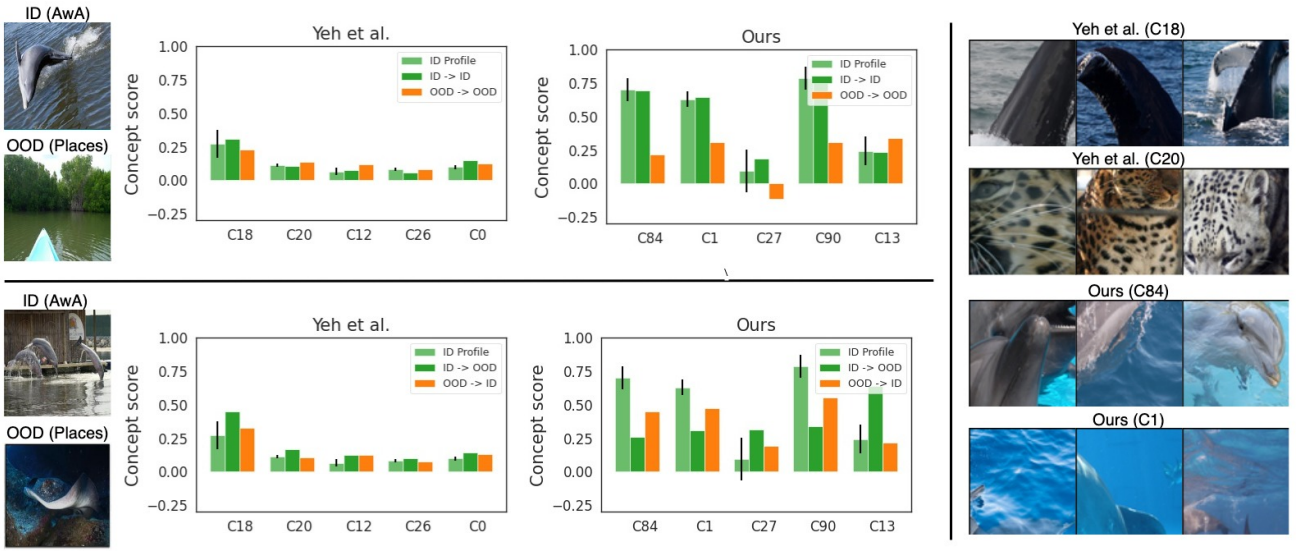
(a) class "Collie", Energy OOD detector. Images randomly selected from AwA test set and SUN.



(b) class "Collie", MSP OOD detector. Images randomly selected from AwA test set and SUN.



(c) class "Dolphin", Energy OOD detector. Images randomly selected from AwA test set and `Places`.



(d) class "Dolphin", MSP OOD detector. Images randomly selected from AwA test set and `Places`.

Figure 11: **Concept-based explanations using concepts identified by Yeh et al. (2020) vs. ours.** ID profile shows the average concept-score pattern for normal ID images.