The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

Interpretable Failure Detection with Human-Level Concepts

Kien X. Nguyen, Tang Li, Xi Peng

Department of Computer and Information Sciences
University of Delaware
Newark, DE, USA
{kxnguyen, tangli, xipeng}@udel.edu

Abstract

Reliable failure detection holds paramount importance in safety-critical applications. Yet, neural networks are known to produce overconfident predictions for misclassified samples. As a result, it remains a problematic matter as existing confidence score functions rely on category-level signals, the logits, to detect failures. This research introduces an innovative strategy, leveraging human-level concepts for a dual purpose: to reliably detect *when* a model fails and to transparently interpret *why*. By integrating a nuanced array of signals for each category, our method enables a finer-grained assessment of the model's confidence. We present a simple yet highly effective approach based on the ordinal ranking of concept activation to the input image. Without bells and whistles, our method significantly reduce the false positive rate across diverse real-world image classification benchmarks, specifically by 3.7% on *ImageNet* and 9% on *EuroSAT*.

Introduction

Vision-language models have demonstrated impressive capability across diverse visual recognition domains (Radford et al. 2021; Jia et al. 2021; Singh et al. 2021; Li et al. 2022, 2023). However, when it comes to safe deployment in high-stake applications, it is of paramount importance for a model to be self-aware of its own shortcomings. For instance, in monitoring for natural disasters such as floods or wildfires, the AI system must signal for human intervention upon encountering scenarios where its confidence is low. Such self-awareness ensures that preemptive measures can be taken to mitigate disaster impacts on communities and ecosystems. Therefore, it is imperative not only to detect failures accurately but also to understand the reasons behind them.

Traditional methods (Hendrycks and Gimpel 2016; Granese et al. 2021; Zhu et al. 2023a,b; Liang, Li, and Srikant 2018) rely on category-level information to detect misclassifications, performing confidence estimation on the class logits. However, neural networks are known to produce overconfident predictions for misclassified samples due to factors like spurious correlations (Arjovsky et al. 2019; Sagawa et al. 2019), thus existing confidence scoring functions (CSFs) fall short in such cases. Besides, the model confidence depicted through category-level information impedes the ability for humans to interpret *why* it fails. To this end, we ask the following question: *"What other sources of information can we leverage to enhance failure detection?"*

We present a novel perspective on detecting failures by leveraging human-level concepts, or visual attributes. With the flexibility to incorporate free-form language to VLMs (*i.e.* CLIP), we can represent a category with a set of predefined concepts (Menon and Vondrick 2023; Oikarinen et al. 2023; Li, Ma, and Peng 2024a,b). Instead of only prompting the model *"Do you recognize a camel?"*, we collectively ask *"Do you recognize humps on back?"*, or *"Do you recognize shaggy coat?"*. The purpose is to measure the model's confidence in the object's detailed visual attributes in addition to the holistic category. We thus achieve a more *accurate* confidence estimate to detect failures more effectively (Fig. 1).

Ideally, a VLM that can recognize a image of a camel should also recognize all the associated visual attributes, such as *humps on back*, *shaggy coat*, *etc.* Such visual attributes should yield higher confidence scores compared to those associated with the absent categories. Conversely, if the model shows high confidence in concepts from multiple unrelated categories at the same time, it could indicate a failure in its recognition process. Based on such intuition, we present a simple but effective approach using the **O**rdinal **R**anking of **C**oncept **A**ctivation (ORCA) to detect failures. Additionally, these human-understandable concepts allow users to understand the reasons behind such failures, thereby aiding them in refining the training process.

We rigorously validate our method's efficacy in detecting incorrect samples across both natural and remote sensing image benchmarks, which mirror the complexity in real-world scenarios. ORCA demonstrates a significant capability to mitigate the issue of overly confident misclassifications. In summary, our contributions are threefold:

1. We leverage human-level concepts to detect *when* and interpret *why* a model fails using vision-language models.
2. We present a simple but effective approach, called ORCA, to estimate more reliable confidence via the ordinal ranking of the concepts' activation.
3. We empirically demonstrate that the concept-based methods enhance failure prediction performance across a wide range of classification benchmarks.

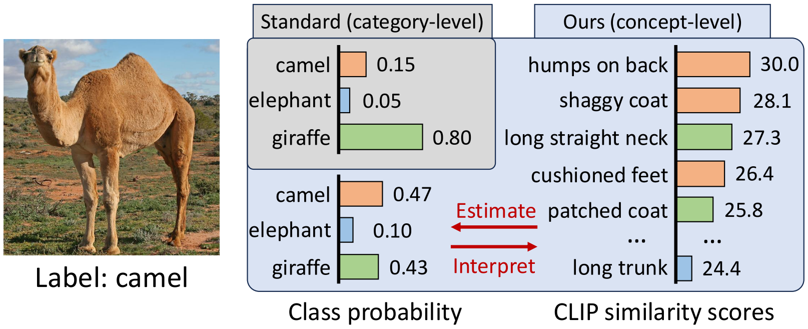


Figure 1: Comparison between standard (MSP) and our approaches. MSP relies solely on class logits to predict failures, which is problematic in detecting overconfident but incorrect predictions. To tackle this problem, we propose to deconstruct each category into its associated human-level concepts for a *finer-grained* estimate of confidence.

Related Work

Failure Detection. Failure detection, or misclassification detection, is a burgeoning area of research within the realm of artificial intelligence. Detecting when machine learning models produce incorrect or misleading predictions has significant implications for safety, reliability, and transparency in various domains. Existing research in this field falls into two main categories: (1) retraining or fine-tuning of neural networks (Moon et al. 2020; Zhu et al. 2023a,b), and (2) the design of novel confidence score functions (Granese et al. 2021; Hendrycks and Gimpel 2016). The former approach involves retraining or fine-tuning neural networks with specific objectives aimed at improving the model's capability to recognize its own failures. Zhu et al. (Zhu et al. 2023b) employs a training objective that seeks flat minima to mitigate overconfident predictions. While these approaches have shown promise, they often require extensive computational resources and access to the entire model, which may not be feasible for large VLMs. Researchers have also turned their attention to the design of new CSFs (Granese et al. 2021). Despite these efforts, the most robust CSF remains the MSP (Jaeger et al. 2023). However, a downside of category-level CSFs is their inability to detect overconfident but incorrect predictions, which is problematic. In this work, we deconstruct category-level into concept-level signals to achieve a more nuanced estimate of the model's confidence.

A closely related sub-field is *confidence calibration* (Minderer et al. 2021; LeVine et al. 2023; Mukhoti et al. 2020; Pereyra et al. 2017), where the goal is to adjust a model's predicted probabilities to ensure that they accurately reflect the true likelihood of those predictions being correct. However, Zhu et al. (Zhu et al. 2023b) has empirically shown that calibration methods frequently yield no benefits or even detrimentally affect failure prediction. Similarly, some works (Jaeger et al. 2023; Bernhardt, Ribeiro, and Glocker 2022) also emphasizes the importance of *confidence ranking* over *confidence calibration* in failure detection. Some other related sub-fields are *predictive uncertainty estimation* (Gal and Ghahramani 2016; Blundell et al. 2015; Lakshminarayanan, Pritzel, and Blundell 2017; Mukhoti et al. 2023), *out-of-distribution detection* (Zhu et al. 2023a; Liang, Li, and Srikant 2018; Dinari and Freifeld 2022; Lee et al. 2018), *open-set recognition* (Vaze et al. 2022; Geng, Huang, and Chen 2021) and *selective classification* (Geifman and El-Yaniv 2017; Fisch, Jaakkola, and Barzilay 2022).

Human-level Concepts in Vision-Language Models. Concept-based models (CBMs) aim to open the black box of neural networks. Concept bottleneck networks are pioneers for interpretable neural networks, with each neuron in the concept bottleneck layer representing a concept at the human level (Koh et al. 2020; Yuksekgonul, Wang, and Zou 2022). With the flexibility to employ free language in vision language models, such as CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), FLAVA (Singh et al. 2021), and BLIP (Li et al. 2022, 2023), concepts of human level can be naturally integrated into the prediction mechanism (Menon and Vondrick 2023; Yang et al. 2022; Oikarinen et al. 2023). This work can be viewed as a variant of CBMs for failure detection, which has never been considered before. We show the approach better predicts failures and, as a byproduct, helps interpret *why* a model fails.

Backgrounds

Overview on Failure Detection. We consider failure detection on the multi-class classification task. Let $\mathcal{X} \in \mathbb{R}^d$ be the input space and $\mathcal{Y} = \{1, 2, \ldots, C\}$ be the label space, where d is the dimension of the input vector. Given a data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with N data points independently sampled from the joint probability distribution $\mathcal{X} \times \mathcal{Y}$, a standard neural network $f : \mathcal{X} \to \mathcal{Y}$ outputs a probability distribution over the C categories. For an input $\mathbf{x}$, f outputs $\hat{\boldsymbol{p}} = \hat{P}(y|\mathbf{x}; \theta)$ as the class probabilities, where θ denotes the network's parameters. In the context of failure detection, we consider a pair of functions (f, g), where $g : \mathcal{F} \times \mathcal{X} \to \mathbb{R}$ is the confidence scoring function, and $f \in \mathcal{F}$. With a predefined threshold $\tau \in \mathbb{R}^+$, the failure detection output is defined as:

$$(f, g)(\mathbf{x}) = \begin{cases} \hat{P}(y|\mathbf{x}; \theta), & \text{if } g(f, \mathbf{x}) \geq \tau \\ \text{detect}, & \text{otherwise.} \end{cases} \quad (1)$$

Failure detection is initiated when $g(f, \mathbf{x})$ falls below a threshold τ. Ideally, a confidence scoring function should output higher confidence scores for correct predictions and lower confidence scores for incorrect predictions. Despite efforts in designing CSFs, Jaeger et al. (Jaeger et al. 2023) has shown that the standard Maximum Softmax Prediction remains the best CSF across a wide range of datasets and network architectures. Mathematically, MSP is defined as:

$$g(f, \mathbf{x}) = \max_{c \in \mathcal{Y}} \hat{P}(y = c|\mathbf{x}; \theta) \tag{2}$$

which returns the maximum output signal after the softmax activation function on the network output layer.

Failure Detection with VLM. CLIP (Radford et al. 2021), a vision-language model, is pre-trained on a large-scale dataset comprising of 400 million image-text pairs. CLIP uses contrastive learning to align the image and text pairs. During inference, we calculate the model's logits as the cosine similarity score between the input image embedding and the corresponding text embeddings. Given an input image $\mathbf{x}$, the embedding is denoted as $f_{\text{img}}(\mathbf{x}) \in \mathbb{R}^m$. In addition, C text labels represents the category names $\{\mathbf{t}_c\}_{c=1}^{C}$, where $f_{\text{txt}}(\mathbf{t}_c) \in \mathbb{R}^m$ is the embeddings and $m \ll d$. For each category, we calculate the corresponding logit as:

$$s_c = 100 \times \frac{f_{\text{img}}(\mathbf{x}) \cdot f_{\text{txt}}(\mathbf{t}_c)}{\|f_{\text{img}}(\mathbf{x})\| \|f_{\text{txt}}(\mathbf{t}_c)\|} \tag{3}$$

where $\|\cdot\|$ is the L_2 norm. The softmax function then converts the logits into probabilities:

$$\hat{p}_c = \frac{\exp(s_c)}{\sum_{j=1}^{C} \exp(s_j)} \tag{4}$$

where $\hat{p}_c \in \hat{\boldsymbol{p}}$. $f(\mathbf{x}) = \operatorname{argmax}_{c \in \mathcal{Y}} \hat{p}_c$ is the prediction, and $g(f, \mathbf{x}) = \max_{c \in \mathcal{Y}} \hat{p}_c$ can be regarded as the model confidence for a given input $\mathbf{x}$ using MSP.

Methods

Traditional methods rely on the category-level signals to estimate the model's confidence. This leads to unreliable confidence estimate as neural networks are prone to overconfident misclassification. To address this issue, we suggest exposing the model to diverse viewpoints via human-level concepts. Rather than inquiring about the model's certainty regarding an image being a camel, we also query its confidence regarding specific attributes like the presence of humps on the camel's back, a shaggy coat, *etc.*

Recent advancements in VLMs enable such integration of human-level concepts as free-form language into the pipeline (Menon and Vondrick 2023; Yang et al. 2022; Oikarinen et al. 2023). In this section, we describe the integration of the work by Menon and Vondrick (Menon and Vondrick 2023) which employs concept aggregation to establish a baseline concept-based method for failure detection. Subsequently, we introduce ORCA, our novel approach that captures the interaction among concept activations through ordinal ranking, enhancing the reliability of failure detection.

Human-Level Concepts for Failure Detection

Given K concepts per category, we define $\mathcal{A}$ as a collection of all concepts, where $|\mathcal{A}| = C \times K$. We obtain the vector of similarity scores (or logits), $S_{\text{conc}} = [s_{1,1}, \ldots, s_{1,K}, s_{2,1}, \ldots, s_{C,K}]$, between the image embedding and all the concepts using Eq. 3. DescCLIP then calculates the mean similarity score among all concepts for each category c to retrieve the logits and output the prediction:

$$f(\mathbf{x}) = \operatorname{argmax}_{c \in \mathcal{Y}} \frac{1}{K} \sum_{k=1}^{K} s_{c,k} \tag{5}$$

Finally, we apply the softmax function (Eq. 4) on the logits to get the class probabilities and employ MSP to obtain the model's confidence score.

Ordinal Ranking of Concept Activation

DescCLIP's concept aggregation leads to a coarse-grained confidence estimation procedure. We propose a fine-grained approach that models the interaction among concepts via ordinal ranking to estimate confidence more reliably.

Ideally, if a model is confident about predicting a category $\hat{c}$ then the concepts associated with $\hat{c}$ should yield the strongest activations. In other words, the similarity scores of all concepts belonging to $\hat{c}$, $\{s_{\hat{c},k}\}_{k=1}^{K}$, should belong to the top-K ranking. Conversely, we would see a mixture of concepts from different categories in the top-K ranking if the model is likely to make an incorrect prediction. With such information, we can separate correct and incorrect predictions more reliably. Next, we describe two variants of our proposed method: baseline and rank-aware ORCA. In brevity, the former builds upon simple counting mechanisms, while the latter weighs the concept contributions to the confidence estimate based on their ranks.

Baseline ORCA. We first sort S_{conc} in descending order and retrieve the set of the top-K concepts, denoted as an ordered set $\mathcal{A}_{\text{top-}K}$. After that, we derive the confidence based on the number of different categories whose concepts belong in $\mathcal{A}_{\text{top-}K}$. The rationale is straightforward: the model is at a higher risk of failure as there are more categories featuring in $\mathcal{A}_{\text{top-}K}$. The prediction is determined as follows:

$$f(\mathbf{x}) = \operatorname{argmax}_{c \in \mathcal{Y}} |\mathcal{A}_{\text{top-}K} \cap \mathcal{A}_c|, \tag{6}$$

where $\mathcal{A}_c$ denotes the set of concepts of an arbitrary category c's concepts, and $|\cdot|$ denotes the set cardinality. The confidence of the prediction is the ratio between the number of the predicted category's concepts appearing in $\mathcal{A}_{\text{top-}K}$ over K:

$$g(f, \mathbf{x}) = \frac{|\mathcal{A}_{\text{top-}K} \cap \mathcal{A}_{\hat{c}}|}{K} \tag{7}$$

where $\hat{c} = f(\mathbf{x})$ is the prediction. We dub this variant ORCA-B in the text.

Rank-aware ORCA. While ORCA-B provides a fundamental approach, its reliance solely on rudimentary counting mechanisms limits its ability to capture nuanced distinctions. To enhance our approach, we introduce a rank-aware variant that uses ordinal ranking information to deliver more accurate failure detection. In detail, we construct a rank-aware weight vector $\mathbf{w}$ where the value of each element is proportional to the ordinal ranking. First, we define the ordinal ranking vector $\mathbf{r} = [K, K-1, \ldots, 1]$ with K elements in descending order. Then, we apply a logarithmic weighting

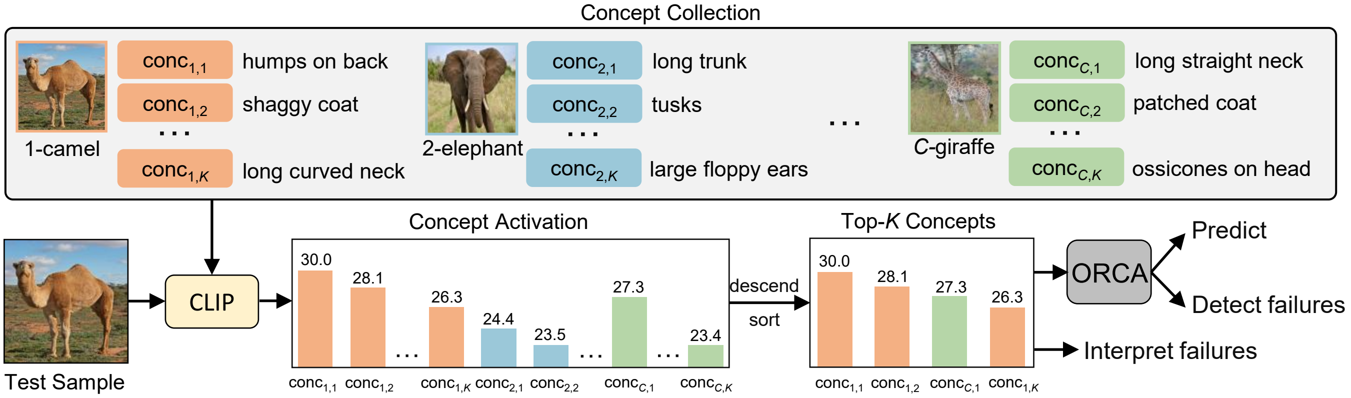


Figure 2: Overview of the ORCA framework. We first prompt GPT-3.5 to construct the concept collection $\mathcal{A}$. We then pass the image and all the concepts into CLIP to retrieve the concept similarity scores, represented by the number above each bar, and sort them in descending order. Based on the top-K responses, we analyze the interaction among concept activations through ordinal ranking to predict the model's failures, and interpret why it fails. "Detect failures" is triggered when the confidence falls below a predefined threshold. Best viewed in color.

function to assign each rank in $\mathbf{r}$ a weight $w_i \in \mathbf{w}$, resulting in a decreasing vector whose elements sum up to 1. Logarithmic ensures a smooth distribution of weights among the ranks of each concept, enabling a more nuanced estimation of the confidence level. Specifically, the logarithmic scaling equation is defined as $w_i = \frac{\log(1+r_i)}{\sum_{j=1}^{K} \log(1+r_j)}$, with the normalization of each weight w_i in $\mathbf{w}$. Finally, for each category c with its concepts featuring in $\mathcal{A}_{\text{top-}K}$, we calculate the prediction and the confidence of the model as follows:

$$f(\mathbf{x}) = \text{argmax}_{c \in \mathcal{Y}} \sum_{k=1}^{K} \mathbb{I}(a_k \in \mathcal{A}_c) \cdot w_k \tag{8}$$

$$g(f, \mathbf{x}) = \max_{c \in \mathcal{Y}} \sum_{k=1}^{K} \mathbb{I}(a_k \in \mathcal{A}_c) \cdot w_k \tag{9}$$

where a_k is the k^{th} concept in the ordered set $\mathcal{A}_{\text{top-}K}$, and $\mathbb{I}(\cdot)$ denotes the indicator function that returns 1 if the condition is true. We refer to this variant as ORCA-R.

Experiment

Datasets. We evaluate ORCA on a wide variety of datasets: **1. Natural Image Benchmark** (1) *CIFAR-10/100* (Krizhevsky 2009) is a popular image recognition benchmark spanning across 10/100 categories. (2) *ImageNet-1K* (Deng et al. 2009) a well-known benchmark in computer vision, containing 1000 fine-grained categories, with 1,281,167 training and 50,000 validation samples. This benchmark contains fine-grained categories that are visually similar, making the failure detection task more challenging. **2. Satellite Image Benchmark** (3) *EuroSAT* (Helber et al. 2017) is a satellite RGB image dataset, containing 10 categories of land usage, such as forest, river, residential buildings, industrial buildings, *etc*. The dataset comprises of 27,000 geo-referenced samples. (4) *RESISC45* (Cheng, Han, and Lu 2017) is a public benchmark for Remote Sensing Image Scene Classification. It contains 31,500 images, covering 45 scene categories with 700 images in each categories.

Baselines. We compare ORCA to 3 models in combination with 3 CSFs, yielding a total of 9 baselines. Note that we only compare with post-hoc CSFs because our methods do not require any training.
1. Models (1) *Zero-shot* (Radford et al. 2021): The prediction of zero-shot CLIP relies on the text category name as introduced in the original paper. We compute the logits using Eq. 3 and apply CSFs to calculate the model's confidence. (2) *Ensemble* (Radford et al. 2021): This model ensembles multiple templates into zero-shot classification, effectively acting as an ensemble method. We average the similarity scores from multiple templates for each category before extracting the softmax logits. (3) *DescCLIP* (Menon and Vondrick 2023): As describe in Sec. , DescCLIP averages the similarity scores of all the concepts for each category; we then applies CSFs to estimate the confidence score.
2. CSFs (1) *MSP* (Hendrycks and Gimpel 2016): The confidence score is measured by taking the maximum value of the softmax responses. (2) *ODIN* (Liang, Li, and Srikant 2018): This CSF is a temperature-scaled version of MSP. We use the default temperature $T = 1000$ and do not use perturbation for fair comparison. (3) *DOCTOR* (Granese et al. 2021): Different from MSP, DOCTOR fully exploits all available information contained in the soft-probabilities of the predictions to estimate the confidence.

Implementation Details. We utilize CLIP's ResNet-101 and ViT-B/32 backbones to perform zero-shot prediction on the benchmarks and calculate the performance metrics. For dataset with few categories, such as *CIFAR-10* and *EuroSAT*, we use different prompts to retrieve diverse collections of concepts from the large language model GPT-3.5 (Brown et al. 2020; Peng et al. 2023) and manually select the top 10 visual concepts that are the most distinctive among categories. An example of our prompt is as follows, with more details in the Supplementary:

```
Q: What are some distinctive visual
    concepts of [CATEGORY]?
```

Dataset	Method	ResNet-101			ViT-B/32		
		AUROC ↑	FPR95 ↓	ACC ↑	AUROC ↑	FPR95 ↓	ACC ↑
CIFAR10 ($K = 10$)	Zero-shot + MSP	85.98	62.98	78.01	88.92	58.66	88.92
	+ ODIN	83.65	65.50	78.01	84.49	65.36	88.92
	+ DOCTOR	86.56	63.76	78.01	88.58	62.32	88.92
	Ensemble + MSP	**86.35**	63.53	80.97	**89.25**	57.03	89.70
	+ ODIN	83.39	67.95	80.97	83.66	63.34	89.70
	+ DOCTOR	85.67	66.53	80.97	88.68	58.87	89.70
	DescCLIP + MSP	85.84	64.68	80.70	89.28	58.77	88.80
	+ ODIN	80.92	68.34	80.70	82.61	66.83	88.80
	+ DOCTOR	84.99	67.92	80.70	88.80	61.64	88.80
	ORCA-B	84.90	66.09	80.98	87.34	**50.52**	89.34
	ORCA-R	85.93	**62.68**	80.60	89.00	52.70	**90.00**
CIFAR100 ($K = 20$)	Zero-shot + MSP	80.72	73.40	48.50	81.15	71.09	58.42
	+ ODIN	77.21	75.13	48.50	76.93	71.08	58.42
	+ DOCTOR	79.68	75.36	48.50	81.57	69.40	58.42
	Ensemble + MSP	79.22	73.43	48.66	81.44	70.88	63.91
	+ ODIN	75.59	76.00	48.66	75.73	73.87	63.91
	+ DOCTOR	77.96	76.47	48.66	80.02	74.06	63.91
	DescCLIP + MSP	80.22	73.39	52.90	82.54	67.38	66.70
	+ ODIN	75.86	75.35	52.90	75.72	73.11	66.70
	+ DOCTOR	79.09	74.96	52.90	81.30	70.83	66.70
	ORCA-B	80.35	**70.46**	52.16	83.35	67.35	66.00
	ORCA-R	**80.46**	72.38	**53.11**	**83.40**	**67.00**	66.50
ImageNet ($K = 25$)	Zero-shot + MSP	78.93	74.05	56.67	79.44	72.91	58.37
	+ ODIN	70.59	80.75	56.67	70.48	80.07	58.37
	+ DOCTOR	78.38	75.90	56.67	79.01	74.17	58.37
	Ensemble + MSP	78.58	74.37	56.73	79.66	72.89	59.22
	+ ODIN	70.29	80.98	56.73	70.61	80.55	59.22
	+ DOCTOR	77.98	76.25	56.73	78.34	76.24	59.22
	DescCLIP + MSP	80.09	72.99	61.94	80.77	71.34	63.20
	+ ODIN	69.92	81.53	61.94	70.80	80.14	63.20
	+ DOCTOR	79.68	73.95	61.94	80.50	71.96	63.20
	ORCA-B	80.24	**71.13**	62.11	80.77	**69.19**	63.02
	ORCA-R	**80.57**	72.41	**62.29**	**80.91**	71.70	**63.20**

Table 1: Performance on *CIFAR-10/100* and *ImageNet*. AUROC, FPR@95TPR (FPR95), and ACC are percentages. With ACC taken into account, **bold** indicate the best results, underlined denote ours with the second best results.

```
A: Some distinctive visual concepts of
    [CATEGORY] are:
```

For datasets with a larger number of categories, we use the concept collection provided by Yang et al. (Yang et al. 2022). This collection contains up to 500 concept candidates per category; we then select the top concepts that yield the highest average similarity score with the images within each category to form $\mathcal{A}$. We include the number of concepts used for each dataset in Table 1 and 2.

Evaluation Metrics

Failure detection accuracy (AUROC). This evaluation protocol, a threshold-independent performance evaluation, measures the area under the receiver operating characteristic curve as CSFs inherently perform binary classification between correct and incorrect predictions. A higher value denotes better ability to predict failures.

False positive rate (FPR@95TPR). This metric denotes the false positive rate or the probability that a misclassified sample is predicted as a correct one when the true positive rate is at 95%. It is a fraction that the model falsely assigns higher confidence values to incorrect samples, reflecting the tendency to be overly confident in incorrect predictions.

Classification accuracy (ACC). A classifier with low accuracy might produce easy-to-detect failures (Jaeger et al. 2023) and benefit from a high AUROC. Ideally, we wish a model to yield a high AUROC and ACC, and a low FPR simultaneously.

Results on Natural Image Benchmarks

We report the performance of all methods on the three evaluation metrics on the natural image benchmarks on ResNet-101 and ViT-B/32 and provide the following *observations*:

Observation 1: Concept-based methods demonstrate better failure detection.

Table 1 shows DescCLIP and ORCA consistently achieves higher AUROC compared to Zero-shot and Ensemble, especially on datasets with a large number of categories, such as *CIFAR-100* and *ImageNet*. The augmentation to multiple signals per category helps concept-based methods obtain a finer-grained analysis for better failure detection. On a different note, Ensemble boosts the Zero-shot's ACC but still results in a lower AUROC and higher FPR on the large-scale datasets for both backbones. Ensemble, in the same principles as concept-based methods, augments the number of signals; however, we hypothesize the *lack of diversity* in those signals deteriorates the separability between correct and incorrect samples.

Observation 2: Our method reduces overconfident but incorrect predictions.

In Table 1, we observe that our methods consistently reduce the false positive rate across datasets and for both backbones. Both variants of ORCA decrease the FPR@95TPR substantially while keeping AUROC and ACC competitive. On *ImageNet*, ORCA-B achieves the best performance on this metric, outperforming the zero-shot model and DescCLIP by 3.72% and 2.15% respectively using ViT-B/32. We hypothesize that allowing the model to recognize an object from different angles provides more reliable confidence assessment, enabling faithful failure detection while also achieving superior predictive accuracy.

Results on Satellite Image Benchmarks

We report the performance on *EuroSAT* and *RESISC45* on ResNet-101 and ViT-B/32. Note that all results are zero-shot performance. We discuss the following *observation*:

Observation 3: Our method boosts both predictive and failure detection accuracy on remote sensing benchmarks.

Table 2 shows that ORCA-R consistently outperforms all baselines on all evaluation metrics. Compared to DescCLIP + MSP on *EuroSAT*, ORCA-R enjoys a 3.6% improvement in AUROC and 6.25% in FPR while boosting the overall accuracy by 1.49%. On *RESISC45*, while ORCA-R's improvement on AUROC and ACC is marginal, it significantly reduces FPR. Additionally, these datasets represent out-of-distribution data for CLIP, underscoring ORCA's enhanced reliability and robustness against such distributional variations.

Ablation Studies

We conduct two ablation studies on the effect of the number of concepts and the choice of the weighting function used for ORCA-R in this section.

Ablation on number of concepts. We use the ViT-B/32 backbone on *CIFAR-100* and $K = \{5, 10, 15, 20\}$ for this experiment. We study the effect of the number of concepts on the performance on AUROC and FPR@95TPR of DescCLIP + MSP, ODIN, DOCTOR and ORCA-R. Fig. 3 shows that the FPR of ORCA-R is consistently lower than those of the other baselines across various K. We also see an increasing (decreasing) trend in AUROC (FPR) as the number of concepts rises. This signifies a finer-grained assessment both enables better failure detection and alleviates the problem of assigning high confidence to incorrect predictions.

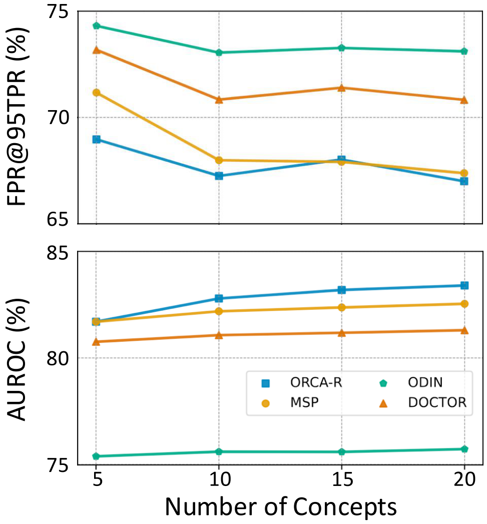


Figure 3: Failure detection accuracy (AUROC) and false positive rate (FPR@95TPR) across different numbers of concepts on *CIFAR-100*. Overall, we can an increase in the number of concepts boosts the performance in both metrics.

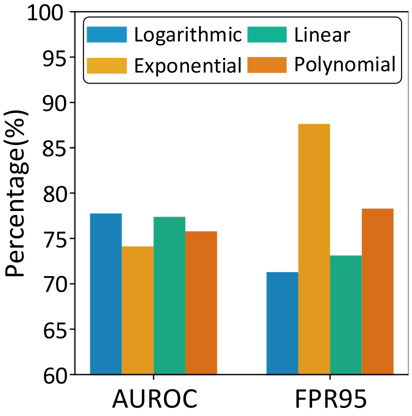
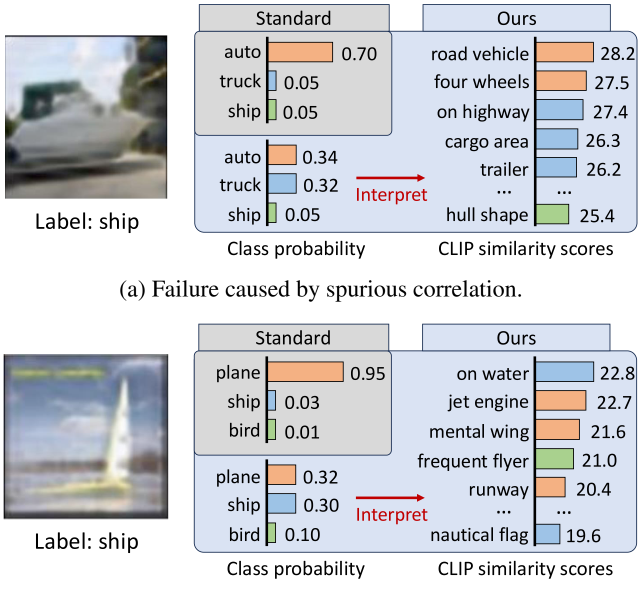


Figure 4: Failure detection capabilities of each weighting function on *EuroSAT*, where `Logarithmic` consistently outperforms others.

Ablation on choice of weighting function. We examine how various weighting functions influence the failure detection efficacy of ORCA-R. Fig. 4 (left) visualizes the weight distribution on the top-10 concepts among the weighting functions. In Figure 4 (right), `Logarithmic` outperforms others, contrasting with `Exponential`, which exhibits the least effectiveness. `Logarithmic` ensures a balanced distribution of weights, recognizing the importance of higher-ranked concepts while also accounting for lower-ranked ones. Conversely, `Exponential` significantly overweighs the highest-ranked concept, neglecting the contributions of those ranked lower.

Dataset	Method	ResNet-101			ViT-B/32		
		AUROC ↑	**FPR95** ↓	**ACC** ↑	**AUROC** ↑	**FPR95** ↓	**ACC** ↑
EuroSAT ($K = 10$)	Zero-shot + MSP	61.73	88.98	30.30	76.42	80.24	41.11
	+ ODIN	61.35	89.38	30.30	75.54	79.28	41.11
	+ DOCTOR	60.76	89.85	30.30	76.67	79.30	41.11
	Ensemble + MSP	54.69	92.21	31.90	66.83	89.19	48.73
	+ ODIN	55.10	93.09	31.90	65.73	90.09	48.73
	+ DOCTOR	53.73	94.09	31.90	61.14	90.63	48.73
	DescCLIP + MSP	64.89	86.39	33.13	73.93	77.54	48.51
	+ ODIN	64.16	87.16	33.13	71.74	78.34	48.51
	+ DOCTOR	62.79	89.05	33.13	72.74	79.85	48.51
	ORCA-B	$\underline{67.86}$	**86.43**	$\underline{34.11}$	76.20	77.80	$\underline{49.74}$
	ORCA-R	**69.01**	**86.43**	**34.76**	**77.55**	**71.29**	**50.00**
RESISC45 ($K = 10$)	Zero-shot + MSP	68.13	87.04	37.66	77.92	80.35	55.57
	+ ODIN	62.60	89.48	37.66	71.66	84.85	55.57
	+ DOCTOR	67.57	87.19	37.66	76.95	82.17	55.57
	Ensemble + MSP	68.87	85.39	39.79	78.40	80.14	56.68
	+ ODIN	62.67	89.57	39.79	71.99	85.31	56.68
	+ DOCTOR	67.88	87.29	39.79	77.54	82.34	56.68
	DescCLIP + MSP	73.44	79.78	43.16	77.47	82.25	58.33
	+ ODIN	69.47	84.61	43.16	71.49	86.21	58.33
	+ DOCTOR	72.95	79.89	43.16	76.81	84.88	58.33
	ORCA-B	71.88	90.41	**46.22**	$\underline{77.71}$	86.31	59.10
	ORCA-R	**74.28**	**80.31**	$\underline{45.13}$	**78.24**	**76.52**	**59.10**

Table 2: Performance on *EuroSAT* and *RESICS45*. AUROC, FPR@95TPR (FPR95), and ACC are percentages. With ACC taken into account, **bold** indicate the best results, underlined denote ours with the second best results.



(a) Failure caused by spurious correlation.

(b) Failure caused by cross-category resemblance.

Figure 5: Failure interpretation with human-level concepts. We show the confidence scores of the top 3 categories (left histograms) and similarity scores of the top 10 concepts (right histograms) from *CIFAR-10*. Standard methods might output overconfident misclassifications due to: (a) *spurious correlation* and (b) *cross-category resemblance*. Concept-level signals not only achieves better failure detection capability in such scenarios but also enables further interpretation of *why* the model fails. "auto" is short for "automobile."

Failure Interpretation

ORCA not only achieves superior failure detection but also enables failure interpretation with human-level concepts. We discuss two scenarios that cause the model to output overconfident values on misclassified samples: *spurious correlation* and *cross-category resemblance* (Fig. 5).

In the former scenario (Fig. 5a), the presence of a road (a spurious feature) leads the model to misclassify the ship as a land vehicle, automobile or truck. We demonstrate that a standard model struggles to identify such failures, resulting in a high confidence score for automobile. In contrast, ORCA leverages human-level concepts, offering more nuanced signals for a refined assessment of the model's confidence. For instance, strong responses from concepts like "road vehicle" and "four wheels" for automobile, and "cargo area" and "trailer" for truck, contribute to a significantly lower confidence. Furthermore, we can easily interpret *why* the model makes such a prediction through concepts.

In the latter scenario (Fig. 5b), the ship (sailboat) bears a resemblance to an airplane from a distance. The similarity between the sky and water also creates an illusion of the object being airborne. The top-K concepts from our method exhibit strong responses to concepts associated with airplanes and birds. Analyzing this information allows us to confidently deduce that the model misclassifies the image as an airplane due to the sky-like background and the object's resemblance to an airplane.

Code — https://github.com/Nyquixt/ORCA

Acknowledgments

This work is supported by the DoD DEPSCoR Award AFOSR FA9550-23-1-0494, the NSF CAREER Award No. 2340074, the NSF SLES Award No. 2416937, and the NSF III CORE Award No. 2412675. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the supporting entities.

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization. *ArXiv*, abs/1907.02893.

Bernhardt, M.; Ribeiro, F. D. S.; and Glocker, B. 2022. Failure Detection in Medical Image Classification: A Reality Check and Benchmarking Testbed. *TMLR*.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 1613–1622. JMLR.org.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T. J.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. *ArXiv*, abs/2005.14165.

Cheng, G.; Han, J.; and Lu, X. 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 105(10): 1865–1883.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. 248–255.

Dinari, O.; and Freifeld, O. 2022. Variational- and metric-based deep latent space for out-of-distribution detection. In Cussens, J.; and Zhang, K., eds., *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, 569–578. PMLR.

Fisch, A.; Jaakkola, T.; and Barzilay, R. 2022. Calibrated Selective Classification. arXiv:2208.12084.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Balcan, M. F.; and Weinberger, K. Q., eds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, 1050–1059. New York, New York, USA: PMLR.

Geifman, Y.; and El-Yaniv, R. 2017. Selective Classification for Deep Neural Networks. arXiv:1705.08500.

Geng, C.; Huang, S.-J.; and Chen, S. 2021. Recent Advances in Open Set Recognition: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10): 3614–3631.

Granese, F.; Romanelli, M.; Gorla, D.; Palamidessi, C.; and Piantanida, P. 2021. DOCTOR: A Simple Method for Detecting Misclassification Errors. In *Neural Information Processing Systems*.

Helber, P.; Bischke, B.; Dengel, A. R.; and Borth, D. 2017. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12: 2217–2226.

Hendrycks, D.; and Gimpel, K. 2016. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *ArXiv*, abs/1610.02136.

Jaeger, P. F.; Lüth, C. T.; Klein, L.; and Bungert, T. J. 2023. A Call to Reflect on Evaluation Practices for Failure Detection in Image Classification. In *The Eleventh International Conference on Learning Representations*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning*.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept Bottleneck Models. *ArXiv*, abs/2007.04612.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6405–6416. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. arXiv:1807.03888.

LeVine, W.; Pikus, B.; Raja, P.; and Gil, F. A. 2023. Enabling Calibration In The Zero-Shot Inference of Large Vision-Language Models. arXiv:2303.12748.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *ArXiv*, abs/2301.12597.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*.

Li, T.; Ma, M.; and Peng, X. 2024a. Beyond Accuracy: Ensuring Correct Predictions With Correct Rationales. *arXiv preprint arXiv:2411.00132*.

Li, T.; Ma, M.; and Peng, X. 2024b. Deal: Disentangle and localize concept-level explanations for vlms. In *European Conference on Computer Vision*, 383–401. Springer.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. *ICLR*.

Menon, S.; and Vondrick, C. 2023. Visual Classification via Description from Large Language Models. *ICLR*.

Minderer, M.; Djolonga, J.; Romijnders, R.; Hubis, F. A.; Zhai, X.; Houlsby, N.; Tran, D.; and Lucic, M. 2021. Revisiting the Calibration of Modern Neural Networks. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Moon, J.; Kim, J.; Shin, Y.; and Hwang, S. 2020. Confidence-Aware Learning for Deep Neural Networks. *ArXiv*, abs/2007.01458.

Mukhoti, J.; Kirsch, A.; van Amersfoort, J.; Torr, P. H.; and Gal, Y. 2023. Deep Deterministic Uncertainty: A New Simple Baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24384–24394.

Mukhoti, J.; Kulharia, V.; Sanyal, A.; Golodetz, S.; Torr, P. H. S.; and Dokania, P. K. 2020. Calibrating deep neural networks using focal loss. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Oikarinen, T. P.; Das, S.; Nguyen, L. M.; and Weng, T.-W. 2023. Label-Free Concept Bottleneck Models. *ArXiv*, abs/2304.06129.

Peng, A.; Wu, M.; Allard, J.; Kilpatrick, L.; and Heidel, S. 2023. GPT-3.5 Turbo Fine-Tuning and API Updates. https://openai.com/blog/gpt-3-5-turbo/.

Pereyra, G.; Tucker, G.; Chorowski, J.; Łukasz Kaiser; and Hinton, G. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. arXiv:1701.06548.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *ArXiv*, abs/1911.08731.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2021. FLAVA: A Foundational Language And Vision Alignment Model. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15617–15629.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Open-Set Recognition: A Good Closed-Set Classifier is All You Need. In *International Conference on Learning Representations*.

Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; and Yatskar, M. 2022. Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification. *CVPR*.

Yuksekgonul, M.; Wang, M.; and Zou, J. Y. 2022. Post-hoc Concept Bottleneck Models. *ArXiv*, abs/2205.15480.

Zhu, F.; Cheng, Z.; Zhang, X.-Y.; and Liu, C.-L. 2023a. OpenMix: Exploring Outlier Samples for Misclassification Detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12074–12083.

Zhu, F.; Cheng, Z.; Zhang, X.-Y.; and Liu, C.-L. 2023b. Rethinking Confidence Calibration for Failure Prediction. *ArXiv*, abs/2303.02970.