

Concept-Based Explainable Artificial Intelligence: Metrics and Benchmarks

Halil Ibrahim Aysel¹ Xiaohao Cai¹ Adam Prugel-Bennett¹

Abstract

Concept-based explanation methods, such as concept bottleneck models (CBMs), aim to improve the interpretability of machine learning models by linking their decisions to human-understandable concepts, under the critical assumption that such concepts can be accurately attributed to the network’s feature space. However, this foundational assumption has not been rigorously validated, mainly because the field lacks standardised metrics and benchmarks to assess the existence and spatial alignment of such concepts. To address this, we propose three metrics: the *concept global importance metric*, the *concept existence metric*, and the *concept location metric*, including a technique for visualising concept activations, i.e., *concept activation mapping*. We benchmark post-hoc CBMs to illustrate their capabilities and challenges. Through qualitative and quantitative experiments, we demonstrate that, in many cases, even the most important concepts determined by post-hoc CBMs are not present in input images; moreover, when they are present, their saliency maps fail to align with the expected regions by either activating across an entire object or misidentifying relevant concept-specific regions. We analyse the root causes of these limitations, such as the natural correlation of concepts. Our findings underscore the need for more careful application of concept-based explanation techniques especially in settings where spatial interpretability is critical.

1. Introduction

In recent years interest in explainable artificial intelligence (XAI) methods has grown substantially because of the desire to exploit the success of newly developed machine learning methods to new areas of our lives (Goebel et al., 2018; Buhrmester et al., 2021; Van der Velden et al., 2022; Ali et al., 2023; Hassija et al., 2024). In an attempt to make XAI

more understandable to the layman there has been a growing drive to develop techniques that provide explanations in terms of human-understandable concepts (Bau et al., 2017; Kim et al., 2018; Koh et al., 2020; Havasi et al., 2022; Aysel et al., 2023; Shin et al., 2023). One of the big challenges of concept-based XAI methods is that of paramount importance yet lacks a systematic study to ensure that the concepts identified as important to making a decision properly align with human understanding of the concepts.

In this paper, we propose three new metrics for measuring this alignment. The first is the *concept global importance metric* (CGIM), measuring the concept alignment for each image in a class. The second is the *concept existence metric* (CEM), measuring whether the concepts identified as important for making a classification exist in an image. For example, if the horn is identified as the most important concept for deciding the image is a rhinoceros then we should expect the horn to be visible in the image. The third metric is the *concept location metric* (CLM), measuring whether the excitable region of the feature maps used to determine an important concept is close to the location where we would expect the concept to be. In the example above, we would expect the heatmap representing the area of the feature map that corresponds to the horn concept should be located around the horn. Using these three metrics, we create a benchmark problem using the Caltech–UCSB Bird (CUB) dataset (Wah et al., 2011), and test the performance of concept-based XAI methods.

To illustrate the usefulness of our metrics, we examine a prominent example of a concept-based XAI system known as the post-hoc concept-bottleneck models (CBMs) (Yuksekgonul et al., 2023). This method is designed to provide explanations of classifiers based on deep neural network (DNN). The method is a synthesis of two approaches – traditional CBMs (Koh et al., 2020) and concept activation vectors (CAVs) (Kim et al., 2018) – for concept-based explanations. Traditional CBMs are a relatively straightforward approach to introducing human-understandable concepts into XAI. In traditional CBMs, we start from a network trained to classify a set of classes and replace the final few layers with a new set of layers that are trained to predict human-understandable concepts, which provides a “concept bottleneck”. From this concept representation, a fully connected layer is trained to predict the classes. Given a new

¹Electronics and Computer Science, University of Southampton, Southampton, United Kingdom. Correspondence to: Halil Ibrahim Aysel <hia1v20@soton.ac.uk>.

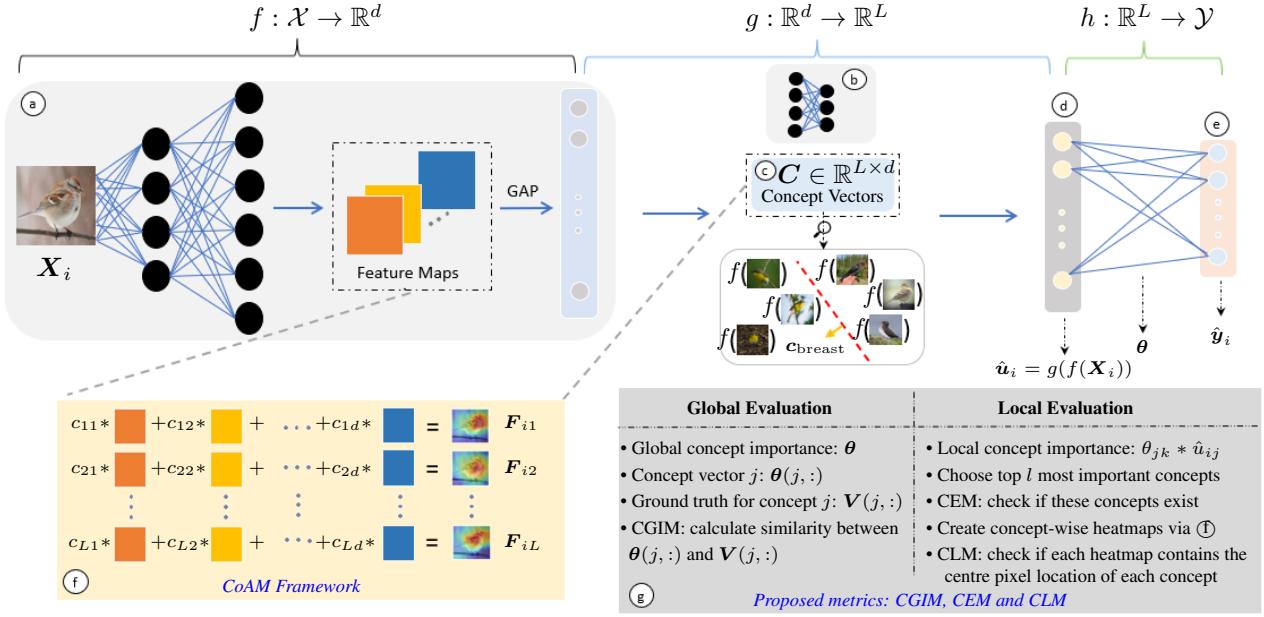


Figure 1. Overview of CAVs, CBMs, post-hoc CBMs and the proposed techniques. Feature extractor ④, concept prediction block ⑤, CAVs ⑥, concept bottleneck ⑦, classifier ⑧, and our proposed CoAM framework ⑨. A traditional (without concept bottleneck) classification model consists of ④ + ⑧, and ⑥ is the introduced post-hoc to explain its predictions via CAVs. ④ + ⑤ + ⑦ + ⑧ forms the steps for traditional CBMs training, whereas ④ + ⑥ + ⑦ + ⑧ forms the post-hoc CBMs. Our proposed CoAM framework is ⑨, weighing pre-GAP feature maps with CAVs for concept visualisation. ⑩ presents the example steps of our proposed metrics.

image, it is then straightforward to see which concepts are important in making the prediction (Koh et al., 2020). The disadvantage of traditional CBMs is that in order to train the network it requires that every image is annotated with the set of concepts that are visible in the image. Although there exists a few datasets where such annotations are given, generally it would be prohibitively expensive to annotate a large dataset.

There has therefore been a drive to find cheaper methods to learn concepts. One example proposed by Aysel et al. (Aysel et al., 2023) is to use annotations for the classes rather than individual images. A second family of models that were developed to provide concept-based explanations is known as CAV methods (Kim et al., 2018). These methods take a pre-trained network and probe the internal representation to determine the directions in that representation that align with human-understandable concepts. One approach for doing this is to take two subsets of the images, one class where the concept is present and the other class where the concept is absent. From examining the difference in the representations between the two classes we can determine CAVs. This method is an example of a “post-hoc” XAI method as it seeks to explain the decisions of a pretrained network without changing that network.

The post-hoc CBMs (Yuksekgonul et al., 2023) combine CAVs with the traditional CBMs. They take a pre-trained network and feed each channel in the last convolution layers

into a global average pooling (GAP) layer. They use the GAP representation to learn a set of CAVs. To do so, for each concept, they choose m positive and negative example images which they then train a support vector machine (SVM) to separate (m is of the order of 100 images). Each SVM discriminant vector is taken as a CAV. These CAVs are then used to determine the degree to which a concept is present in an image. From this, a concept bottleneck can be trained. This is the post-hoc CBMs that we study in this paper. The network is illustrated in the top row of Figure 1, and the bottom row illustrates the new metrics that we propose to evaluate the alignment of the concepts with human understanding of the concepts.

Although post-hoc CBMs sacrifice some performance accuracy in predicting classes compared to traditional DNNs (i.e., the ones without a concept bottleneck), they provide a relatively cheap way to obtain human-interpretable concepts. However, for this explanation to be useful, the concepts need to be accurately aligned with human understanding of the concepts. We use the new metrics and new benchmarks to evaluate this alignment. As we will see, the alignment is surprisingly poor, which highlights the necessity of introducing new metrics for assessing this alignment. The main contributions of the paper are as follows.

- We propose the *concept activation mapping* (CoAM) to visualise concept activations.
- We propose three quantitative metrics: i) CGIM, to

- test the global concept alignment by XAI methods; ii) CEM, to test whether a concept being identified by XAI methods exists in the image; and iii) CLM, to test whether a concept being identified by XAI methods is spatially aligned with the human concept.
- We benchmark the post-hoc CBMs (Yuksekgonul et al., 2023) using the proposed metrics to evaluate the alignment of concept-based XAI techniques on a benchmark dataset and discuss their advantages and limitations.

2. Preliminary

Let \mathcal{X} be the set of images, \mathcal{U} be the set of concept labels, and $\mathcal{Y} = \{1, 2, \dots, K\}$ be the set of K class labels. Let $\mathcal{S} = \{(\mathbf{X}_i, \mathbf{u}_i, \mathbf{y}_i, \Lambda_i, \mathcal{P}_i) \mid \mathbf{X}_i \in \mathcal{X}, \mathbf{u}_i \in \mathcal{U}, \mathbf{y}_i \in \mathcal{Y}, i = 1, 2, \dots, N\}$ be the training set with N samples, where $\mathbf{u}_i \in \{0, 1\}^L$ is the concept label vector with L different concepts for image $\mathbf{X}_i \in \mathbb{R}^{M_1 \times M_2 \times M_3}$ ($M_3 = 3$ for RGB images), $\mathbf{y}_i \in \mathbb{R}^K$ (a one-hot vector) denotes the class label of image \mathbf{X}_i , Λ_i is the set containing the indexes of activated concepts for image \mathbf{X}_i (i.e., the indexes of the components in \mathbf{u}_i with value 1), and $\mathcal{P}_i = \{p_{i1}, \dots, p_{iL}\}$ is the set holding centre pixel coordinates p_{ij} of concept j with $j = 1, \dots, L$ for image \mathbf{X}_i . Let $f : \mathcal{X} \rightarrow \mathbb{R}^d$ be a d -dimensional feature extractor, which can be any trained DNN such as ResNet (He et al., 2016) or VGG (Simonyan & Zisserman, 2014). From block @ in Figure 1, we see that the feature vector $f(\mathbf{X}_i)$ consists of the post-GAP features (i.e., the features right after the GAP layer). Let $\mathbf{E}_i \in \mathbb{R}^{H \times W \times d}$ represent the pre-GAP feature maps (i.e., the features right before the GAP layer), where H, W and d denote the height, width and depth (i.e., the number of channels). The k -th channel of \mathbf{E}_i is represented as $\mathbf{E}_i(:, :, k) \in \mathbb{R}^{H \times W}$.

CAVs. Following Kim et al. (2018) and Yuksekgonul et al. (2023), for $j = 1, \dots, L$, to generate the CAV $\mathbf{c}_j \in \mathbb{R}^d$ for the j -th concept, two sets of image embeddings through f are needed, i.e., $\mathcal{N}_j^{\text{pos}}$ for positive examples and $\mathcal{N}_j^{\text{neg}}$ for negative ones. In detail, set $\mathcal{N}_j^{\text{pos}}$ consists of embeddings of N_p images (positive examples) that contain the j -th concept, and set $\mathcal{N}_j^{\text{neg}}$ consists of embeddings of N_n randomly chosen images (negative examples) that do not contain the concept. Sets $\mathcal{N}_j^{\text{pos}}$ and $\mathcal{N}_j^{\text{neg}}$ are then used to train an SVM with \mathbf{c}_j being the obtained normal vector to the hyperplane separating sets $\mathcal{N}_j^{\text{pos}}$ and $\mathcal{N}_j^{\text{neg}}$. All together, these L number of CAVs form a concept bank $\mathbf{C} = (\mathbf{c}_1, \dots, \mathbf{c}_L)^\top \in \mathbb{R}^{L \times d}$. For an image \mathbf{X}_i , the feature vector $f(\mathbf{X}_i)$ is to be projected onto the concept space by \mathbf{C} , i.e., $\mathbf{C}f(\mathbf{X}_i) \in \mathbb{R}^L$, which is the concept value vector $\hat{\mathbf{u}}_i$ to be fed to the classifier.

Traditional vs. post-hoc CBMs. After the feature vector $f(\mathbf{X}_i)$ is obtained for image \mathbf{X}_i , the traditional CBMs predict concepts by the concept prediction block, while the post-hoc CBMs project the feature vector $f(\mathbf{X}_i)$ onto the

concept space using the concept bank \mathbf{C} , see Figure 1. Let

$$\hat{\mathbf{u}}_i = (\hat{u}_{i1}, \hat{u}_{i2}, \dots, \hat{u}_{iL})^\top = g(f(\mathbf{X}_i)) \quad (1)$$

be the obtained concept vector for image \mathbf{X}_i and g be the projection function. Then, for traditional CBMs, as the ground-truth concept label vector \mathbf{u}_i for image \mathbf{X}_i is available, g (i.e., the concept prediction block) is achieved/trained by minimising the binary cross-entropy loss function $\mathcal{L}_g = \sum_i \mathcal{L}_g(\hat{\mathbf{u}}_i, \mathbf{u}_i)$. In contrast, for post-hoc CBMs, where the ground-truth concept label vector \mathbf{u}_i is not available, the obtained concept vector $\hat{\mathbf{u}}_i$ corresponding to \mathbf{X}_i is directly obtained by setting $g(f(\mathbf{X}_i)) = \mathbf{C}f(\mathbf{X}_i)$.

Finally, the obtained concept vector $\hat{\mathbf{u}}_i$ is used to predict the final classes via a single classification layer $h : \mathbb{R}^L \rightarrow \mathcal{Y}$ for both the traditional and post-hoc CBMs. In detail, $h(\hat{\mathbf{u}}_i) = \boldsymbol{\theta}^\top \hat{\mathbf{u}}_i + b$, where $\boldsymbol{\theta} \in \mathbb{R}^{L \times K}$ holds the weights and b is the bias. Function h is trained for the final classification with the categorical cross-entropy loss function $\mathcal{L}_h = \sum_i \mathcal{L}_h(\hat{\mathbf{y}}_i, \mathbf{y}_i)$, where $\hat{\mathbf{y}}_i = h(g(f(\mathbf{X}_i)))$ is the class prediction of \mathbf{X}_i .

Global vs. local concept importance. After training the model is used to make a prediction for every test image, and then rank the concepts and present the highest l of them as explanations. In this regard, it is crucial to differentiate between the global and local importance of concepts for a task as they may play key roles in different scenarios. Global importance is the overall effect of concepts for a given class. For instance, in the post-hoc CBMs setting, the classifier h is a single layer with weights $\boldsymbol{\theta}$ mapping concept values to the final classes (also see the right of Figure 1) and each parameter of this layer is proposed as the *global importance* of a concept that they weigh for an examined class. By analysing each parameter, say θ_{jk} , one can assess the overall effect of the concept j for class k . Moreover, tuning these parameters may allow the model to be debugged as proposed in Yuksekgonul et al. (2023). The *local importance* of concepts on the other hand is their influence on individual image predictions rather than on the entire class. The CBM and its variants focus on *local concept interventions* (Kim et al., 2018), which is the process of tweaking the predicted/projected concept values in $\hat{\mathbf{u}}_i$ at the concept bottleneck layer, i.e., @ in Figure 1, to flip a single class prediction when needed. An effective way to determine what concept values to intervene on is an active area of research (Steinmann et al., 2024; Vandenhirtz et al., 2024; Shin et al., 2023).

One, however, should note that the magnitude of the concept values at the bottleneck layer is not the same as the *local importance*. This is because a class prediction score by h is the $\boldsymbol{\theta}$ -weighted sum of concept values, and the parameters of $\boldsymbol{\theta}$ may greatly increase or decrease the individual concept effects on the final classification. Therefore, defining the

concept importance solely based on their values in $\hat{\mathbf{u}}_i$ is misleading. We will address this issue in our proposed methodology in Section 3.

3. Proposed Methodology

There is a significant gap in the field regarding the evaluation of the explainability power of the well-known concept-based methodologies. To fill this gap and assess the existence and correctness of the concepts given as highly important by XAI techniques, we propose our CoAM (concept activation mapping) framework (see ① in Figure 1 for an overview), which allows concept visualisation. Moreover, we also propose the CGIM (concept global importance metric) to test the global concept alignment by XAI methods, the CEM (concept existence metric) to evaluate the existence of the concepts, and the CLM (concept location metric) to reveal whether the highly important concepts correspond to the correct regions in a given test image.

3.1. Concept Activation Mapping

We propose the CoAM framework, which generates concept activation maps revealing the parts of an image that correspond to the concepts. As we know, for post-hoc CBMs, the pre-GAP feature maps $\mathbf{E}_i \in \mathbb{R}^{H \times W \times d}$ (which contain spatial information) for the examined image \mathbf{X}_i become the post-GAP feature vector $f(\mathbf{X}_i)$ after the GAP layer, which is then linked to the CAVs, $\mathbf{c}_j = (c_{j1}, \dots, c_{jd})^\top, j = 1, \dots, L$.

Our introduced concept activation maps, say \mathbf{F}_{ij} , for \mathbf{X}_i corresponding to the j -th concept for $j = 1, \dots, L$, are calculated by

$$\mathbf{F}_{ij} = \frac{1}{d} \sum_{k=1}^d c_{jk} \mathbf{E}_i(:, :, k) \in \mathbb{R}^{H \times W}, \quad (2)$$

i.e., weighing the pre-GAP feature maps of \mathbf{X}_i by the j -th CAV; see block ② in Figure 1. The CoAM framework is also summarised in Algorithm 1, with \mathbf{F}_i being the output, where $\mathbf{F}_i(:, :, j) = \mathbf{F}_{ij}$ for $j = 1, 2, \dots, L$. Since the size of each \mathbf{F}_{ij} is significantly smaller than that of \mathbf{X}_i , to visualise the concept activation maps in a better way and for localisation assessment, we upsample them to the original image size of \mathbf{X}_i , denoted by $\hat{\mathbf{F}}_{ij}$, and overlay them on \mathbf{X}_i . This will tell us what parts of the input image contribute to the individual concepts. Algorithm 2 in the Appendix gives the details of the final feature visualisation pseudo-code.

3.2. Concept Global Importance Metric

We firstly introduce the *global importance score* of concept j for class k as θ_{jk} [the (j, k) -th entry of $\boldsymbol{\theta}$], i.e., the weight in the classifier h mapping the j -th concept to the k -th class, for $j = 1, 2, \dots, L$ and $k = 1, 2, \dots, K$. Let $\mathbf{V} \in$

$\mathbb{R}^{L \times K}$ be the ground-truth concept matrix for all the classes provided by annotators, where the entry of its j -th row and k -th column V_{jk} is the ground-truth value of the j -th concept for the k -th class. One might consider directly comparing θ_{jk} and V_{jk} for global evaluation of the correctness of θ_{jk} . However, this is inappropriate because these values are on different scales; in particular, \mathbf{V} contains values between 0 and 1, while $\boldsymbol{\theta}$ can take any real value as it represents layer weights. To address this issue, we propose to compare the entire j -th row vectors $\boldsymbol{\theta}(j, :)$ and $\mathbf{V}(j, :)$ by calculating their similarity for $j = 1, 2, \dots, L$.

Our *first type* CGIM is defined as

$$\rho_j^{\text{CGIM}_1} := \phi(\boldsymbol{\theta}(j, :), \mathbf{V}(j, :)), \quad j = 1, 2, \dots, L, \quad (3)$$

where ϕ is the function for similarity calculation. In this paper, we use the cosine similarity (measuring the alignment between two vectors regardless of their magnitudes) for ϕ . Therefore, $\rho_j^{\text{CGIM}_1}$ is a similarity score between -1 and 1 for the j -th concept. Ideally, $\rho_j^{\text{CGIM}_1}$ is expected to be close to 1 if the obtained $\boldsymbol{\theta}(j, :)$ is meaningful.

Analogous to the first type CGIM in Eqn (3), we also introduce the concept global explanations based on the average say $\hat{\mathbf{u}}_k^*$ of the concept vectors $\hat{\mathbf{u}}_i$ of $\forall \mathbf{X}_i \in \mathcal{X}_T^k$, where \mathcal{X}_T^k is the set that consists of all the test images with correct predicted class $1 \leq k \leq K$ and $|\mathcal{X}_T^k| = N_k$. Then, form $\hat{\mathbf{U}}^* = (\hat{\mathbf{u}}_1^*, \hat{\mathbf{u}}_2^*, \dots, \hat{\mathbf{u}}_K^*) \in \mathbb{R}^{L \times K}$, i.e., the obtained average concept matrix. Our *second type* CGIM is then defined as

$$\rho_j^{\text{CGIM}_2} := \phi(\hat{\mathbf{U}}^*(j, :), \mathbf{V}(j, :)), \quad j = 1, 2, \dots, L. \quad (4)$$

If we consider both the weight matrix $\boldsymbol{\theta}$ and the obtained average concept matrix $\hat{\mathbf{U}}^*$, we have our *third type* CGIM, which is defined as

$$\rho_j^{\text{CGIM}_3} := \phi(\hat{\mathbf{U}}_\theta^*(j, :), \mathbf{V}(j, :)), \quad j = 1, 2, \dots, L, \quad (5)$$

where $\hat{\mathbf{U}}_\theta^* = \boldsymbol{\theta} \odot \hat{\mathbf{U}}^*$ with \odot being the pointwise multiplication operator.

The above proposed CGIM scores $\rho_j^{\text{CGIM}_1}$, $\rho_j^{\text{CGIM}_2}$, and $\rho_j^{\text{CGIM}_3}$ are for each concept $1 \leq j \leq L$. They can also be readily modified analogously so that we can calculate CGIM scores for each class $1 \leq k \leq K$, i.e.,

$$\rho_k^{\text{CGIM}_1} := \phi(\boldsymbol{\theta}(:, k), \mathbf{V}(:, k)), \quad (6)$$

$$\rho_k^{\text{CGIM}_2} := \phi(\hat{\mathbf{U}}^*(:, k), \mathbf{V}(:, k)), \quad (7)$$

$$\rho_k^{\text{CGIM}_3} := \phi(\hat{\mathbf{U}}_\theta^*(:, k), \mathbf{V}(:, k)). \quad (8)$$

3.3. Concept Existence Metric

We now define the *local importance score* of concept j for class k as $\theta_{jk}\hat{u}_{ij}$; note that \hat{u}_{ij} is the obtained j -th

Algorithm 1 Concept Activation Mapping (CoAM)

```

1: Input: Pre-GAP feature maps  $E_i \in \mathbb{R}^{H \times W \times d}$  for  $X_i$ ,  

   and concept bank  $C \in \mathbb{R}^{L \times d}$   

2: Output: Concept activation map  $F_i \in \mathbb{R}^{H \times W \times L}$   

3: for each concept  $j$  in  $C$  do  

4:   Compute the weighted map  $F_{ij}$  with Eqn (2)  

5:   Set  $F_i(:, :, j) = F_{ij}$   

6: end for  

7: return  $F_i$ 
```

concept value of test image X_i and θ_{jk} is the weight in classifier h linking the j -th concept and the k -th class prediction. We rank the total L concepts for test image X_i based on their contribution to the final classification prediction k using the local importance score $\theta_{jk}\hat{u}_{ij}$, and let $q_i = (q_{i1}, q_{i2}, \dots, q_{iL})^\top$ represent the ranked indexes of the concepts for X_i . Therefore, if $q_{is} = m$, it means the m -concept is ranked at the s place for $s = 1, 2, \dots, L$ based on the descending order of the magnitude of $\theta_{mk}\hat{u}_{im}$ among $\{\theta_{jk}\hat{u}_{ij}\}_{j=1}^L$.

Recall that Λ_i is the set containing the indexes of activated concepts for image X_i . Our CEM is defined as

$$\rho_l^{\text{CEM}} := \frac{1}{l} \sum_{j=1}^l \mathbf{1}_{\Lambda_i}(q_{ij}), \quad (9)$$

assessing if the first $l \leq L$ concepts (i.e., the first l components) in q_i exist in the examined image X_i , where $\mathbf{1}_{\Lambda_i}$ is an indicator function defined as

$$\mathbf{1}_{\Lambda_i}(x) = \begin{cases} 1, & \text{if } x \in \Lambda_i; \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

Obviously, the CEM ρ_l^{CEM} is an accuracy score between 0 and 1 evaluating the existence of highly important concepts in image X_i , thanks to the set Λ_i containing the indexes of activated concepts. CEM reveals the reliability of explanations generated by a trained model; for example, $\rho_l^{\text{CEM}} = 0$ means none of the l highly important concepts exists in the examined image, whereas $\rho_l^{\text{CEM}} = 1$ means all of the l highly important concepts exist in the examined image. We remark that ρ_l^{CEM} can also be obtained in the same manner by using θ_{jk} or \hat{u}_{ij} instead of $\theta_{jk}\hat{u}_{ij}$ as the local importance score for comparison purpose.

3.4. Concept Location Metric

After checking whether the obtained important concepts of image X_i exist in the ground-truth set Λ_i with CEM and generating concept activation maps with CoAM, we now propose CLM to assess whether the obtained concepts of image X_i correspond to the correct region in X_i .

Note that this check could be rigorously done by calculating the intersection over union (IoU) score if a ground-truth segmentation map per concept is available. However, the absence of these ground-truth maps makes this way impractical. In contrast, it will be much easier to mark some pixels, e.g. the coordinate information of the centre pixel for each important semantic area in an image, and then link the coordinate information to each concept. One useful label available for this purpose is the coordinate information of the centre pixel for each concept, i.e., \mathcal{P}_i , for image X_i .

The proposed CLM checks whether the concept-wise activation heatmap \bar{F}_{ij} for concept j generated by CoAM contains the ground-truth centre location p_{ij} . For the $l \leq L$ most important concepts of X_i obtained in q_i , our CLM is defined as

$$\rho_l^{\text{CLM}} := \frac{1}{l} \sum_{j=1}^l \mathbf{1}_{\Omega_{ij}}(p_{ij}), \quad (11)$$

where Ω_{ij} is the visual region of concept j of X_i . Obviously, ρ_l^{CLM} is an accuracy score between 0 and 1 evaluating the alignment between the obtained individual concept heatmaps and their actual region in the image X_i . In particular, $\rho_l^{\text{CLM}} = 0$ means none of the l highly important concepts corresponds to the correct region in the image, whereas a $\rho_l^{\text{CLM}} = 1$ score means all the l highly important concepts correspond to the correct region in the image. Finally, we remark that there are many ways to generate the visual region Ω_{ij} . In this paper, we use thresholding on the concept-wise activation heatmap \bar{F}_{ij} with threshold τ to obtain Ω_{ij} .

4. Experiments

In this section, we present benchmark results and evaluate the performance of the post-hoc CBMs using our proposed metrics. The benchmark fine-grained bird classification dataset, Caltech–UCSD Birds (CUB) (Wah et al., 2011), with concept annotations such as *wing color*, *beak shape* and *feather pattern* is employed for the experiments. It consists of 200 different classes and 112 binary concept labels for around 11,800 images. Additionally, the central pixel locations of 12 different body parts are provided and used for concept localisation assessment by the proposed CLM. Following Yuksekgonul et al. (2023), we employ a ResNet-18 (He et al., 2016) trained on the CUB dataset¹ as the feature extractor f . CAVs are calculated as explained in Section 3 to create a concept bank C (also see \mathcal{C} in Figure 1). Finally, a single layer h with weights $\theta \in \mathbb{R}^{112 \times 200}$ is trained for the classification.

¹The trained CUB model is available at <https://github.com/osmr/imgclsmob>.

4.1. Post-hoc CBMs Reproduction

By employing the same model as the feature extractor and following the same steps for CAVs and classifier training, we reproduce the results of post-hoc CBMs (Yuksekgonul et al., 2023) with various hyperparameter combinations. The details of the post-hoc CBMs reproduction is given in the Appendix.

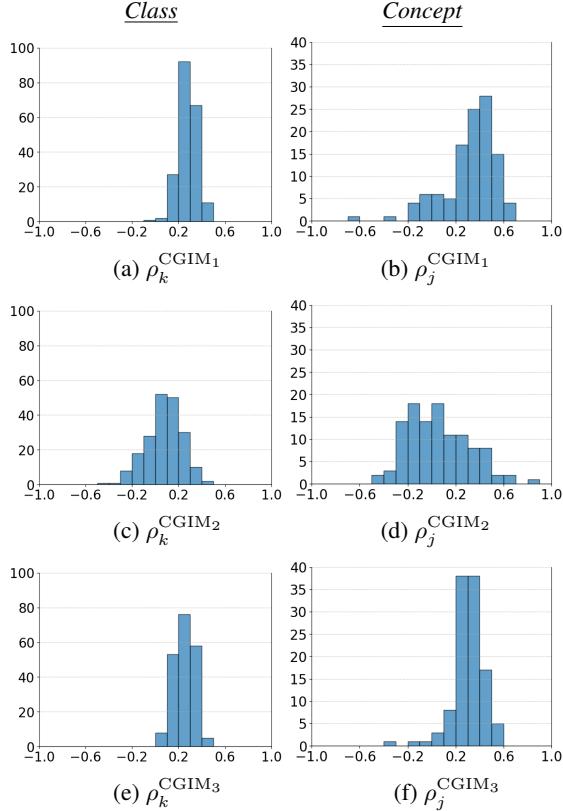


Figure 2. Histograms of the CGIM scores of the post-hoc CBMs. Plots on the left and right columns show the results for classes and concepts, respectively. A full list of the CGIM scores can be found in Table 6 for the concepts and in Tables 7 and 8 for the classes in the Appendix.

4.2. Global Importance Evaluation

We now investigate the quality of the global explanations of the post-hoc CBMs. Recall that the entries of $\theta \in \mathbb{R}^{112 \times 200}$ are considered as the global importance scores, determining the importance of a concept for an examined class. Ideally, these weights should closely align with human annotations in $V \in \mathbb{R}^{112 \times 200}$, i.e., the so-called ground truth. Intuitively, we expect the CGIM scores $\rho_j^{\text{CGIM}_1}$, $\rho_j^{\text{CGIM}_2}$, and $\rho_j^{\text{CGIM}_3}$ of θ , \hat{U}^* , and \hat{U}_θ corresponding to V for each concept $1 \leq j \leq 112$ (and analogously for each class $1 \leq k \leq 200$) to be close to 1 if the obtained θ , \hat{U}^* , and \hat{U}_θ are meaningful.

The calculated CGIM scores of the post-hoc CBMs for each concept $1 \leq j \leq 112$ and for each class $1 \leq k \leq 200$ are respectively presented in Table 6 and Tables 7 and 8 in the Appendix. To better visualise and interpret results, Figure 2 showcases the histograms of the obtained CGIM scores across a range between -1 (maximum dissimilarity) and 1 (maximum similarity), regarding individual classes and concepts. Again, in an ideal scenario, it would be expected the CGIM scores $\rho_k^{\text{CGIM}_1}$ and $\rho_j^{\text{CGIM}_1}$ in the top row of Figure 2 to be a single prominent bar at the value of 1 , or at the very least, a clear accumulation of bars towards the right end of the histogram (approaching 1), if the results of the post-hoc CBMs are meaningful/correct. Obviously, this is not the case. For example, the class and concept histograms in Figure 2 (a)–(b) show that many bars are distributed across the range from -1 to 1 with a noticeable number of values on the negative side, indicating a tendency towards negative correlation for some classes and concepts, which is contrary to the expected accumulation near 1 . The class and concept histograms in terms of $\rho_k^{\text{CGIM}_2}$ and $\rho_j^{\text{CGIM}_2}$ in Figure 2 (c)–(d) and $\rho_j^{\text{CGIM}_3}$ and $\rho_k^{\text{CGIM}_3}$ in Figure 2 (e)–(f) again disclose the same issue of the post-hoc CBMs.

A deeper analysis is also conducted through investigating the specific concepts and classes with significantly low or negative CGIM scores presented in Tables 6, 7 and 8 in the Appendix. For instance, the $\rho_j^{\text{CGIM}_1}$ score for concept ($j = 51$) *black eye colour* in Table 6 is a large negative value, i.e., -0.63 . Similarly, the $\rho_j^{\text{CGIM}_1}$ score close to 0 for class ($k = 18$) *spotted catbird* and class ($k = 25$) *pelagic cormorant* in Table 7 indicates that these classes share no similarities to their ground truth. For the first time, the low and negative valued CGIM scores occurring for many concepts and classes raise concerns about the reliability and quality of the explanations of the post-hoc CBMs.

4.3. Concept Existence Evaluation

After analysing the global importance evaluation based on classifier's weights and average concept predictions, we, in this section, focus on the local importance analysis. The first step in this regard is to assess the concept existence qualitatively and quantitatively.

4.3.1. QUALITATIVE OBSERVATIONS

When a set of concepts is presented as highly important for a prediction by a trained model, it is essential to qualitatively verify whether these concepts really exist in the image. In Figure 3, we present random images from the test set with the top 5 most important concepts for their prediction outputted by the reproduced post-hoc CBMs. As shown in Figure 3, many of those highly important concepts do not actually exist in the given images. For instance, for an *American Redstart* in the first column, the most important

Table 1. Concept existence assessment of the reproduced post-hoc CBMs under CEM for the top l most important concepts.

Image	CEM based on	$l = 1$	$l = 3$	$l = 5$
Entire test set	θ_{jk}	39.2	37.9	37.1
	\hat{u}_{ij}	84.3	80.1	77.2
	$\theta_{jk}\hat{u}_{ij}$	49.3	44.3	41.2
Correct class set	θ_{jk}	48.5	46.8	44.8
	\hat{u}_{ij}	85.4	82.1	80.7
	$\theta_{jk}\hat{u}_{ij}$	55.4	49.1	45.8

concept is given as *white throat*; this is incorrect because the bird has a *black throat*, which can be clearly seen in the input image. Similarly, for the *brown pelican* image in the second column, the fifth most important concept is given as *shorter than head bill*; this is not the case as the pelican has a much longer bill than its head.

4.3.2. QUANTITATIVE TEST BY CEM

We calculate the CEM score over the entire test set. The full results are presented in Table 1 in terms of ranking the importance of the concepts based on i) the weights of the classifier θ_{jk} , ii) the projected concept values \hat{u}_{ij} , and iii) their combination $\theta_{jk}\hat{u}_{ij}$, for the top l most important concepts with l set to 1, 3, and 5. The results show that the CEM score based on \hat{u}_{ij} is significantly higher than the others, which is intuitive at first glance as the highest values after concept projection are highly likely to be present in the ground-truth label. However, as detailed in Section 3, the concept values in \hat{u}_i do not independently determine the final class prediction; instead, these values are weighted by their respective weights in θ , which can significantly alter their overall impact. Relying solely on the projected concept values in \hat{u}_i may therefore lead to misleading conclusions. Hence, we build our argument based on $\theta_{jk}\hat{u}_{ij}$ rather than solely on \hat{u}_{ij} or θ_{jk} . Strikingly, as shown in Table 1, the single most important concept (i.e., when $l = 1$) only exists in the images around 55% of the times when the image is correctly classified. This score drops to 49% when the test is done on the entire test set. Moreover, the CEM score is even lower when l is set to 3 and 5.

4.4. Concept Localisation Evaluation

4.4.1. QUALITATIVE OBSERVATIONS

By visualizing concept heatmaps for different concepts using our proposed CoAM, we identified several recurring patterns. Figure 4 in the Appendix presents some examples of class and concept visualisation by using our CoAM. In many cases, the concept activation maps cover broad image regions, often extending beyond the expected concept areas. For instance, when detecting the concept *grey leg* in a *white breast nuthatch* image, the concept map covers the entire body of the bird rather than focusing on the specific region around the *leg*, as shown in the first row in Figure

Table 2. Concept localisation assessment of the reproduced post-hoc CBMs under CLM for the top l most important concepts.

Value α for Ω_{ij}	CLM based on	$l = 1$	$l = 3$	$l = 5$
$\alpha = 1$	θ_{jk}	10.8	13.6	13.8
	\hat{u}_{ij}	14.9	14.6	14.6
	$\theta_{jk}\hat{u}_{ij}$	13.3	13.5	12.9
$\alpha = 3$	θ_{jk}	29.6	30.1	31.2
	\hat{u}_{ij}	39.2	33.7	33.1
	$\theta_{jk}\hat{u}_{ij}$	33.4	32.2	31.8
$\alpha = 6$	θ_{jk}	52.3	50.8	51.6
	\hat{u}_{ij}	54.5	56.4	55.9
	$\theta_{jk}\hat{u}_{ij}$	59.0	55.0	53.9

4. Moreover, many of the fine-grained concepts such as *crown* or *tail pattern* are often not correctly localised. For instance, the heatmap highlights a region around the *leg* for *blue crown* concept as given in the last row of Figure 4.

4.4.2. QUANTITATIVE TEST BY CLM

To be able to calculate the CLM score, the centre pixel coordinates for individual concepts are needed. In the CUB dataset, the centre pixel coordinates are only available for 12 broader body parts such as *beak*, *throat*, and *leg*. Fortunately, most of the 112 concepts are related to one of the 12 body parts, allowing us to match each concept to its closest body part and hence exploit the corresponding body-part coordinates for concepts. For instance, we match the *hooked seabird beak* concept with the *beak* part and the *solid wing* concept with the *wing*; see Tables 4 and 5 in the Appendix for details. We ignore concepts that are not related to a specific body part such as overall size, shape and colour information, which leaves us with 89 out of 112 concepts for the CLM evaluation.

Recall that after obtaining the activation map for the j -th concept \bar{F}_{ij} of X_i , CLM checks if the centre pixel location p_{ij} falls into the highest activated region Ω_{ij} . Here Ω_{ij} is formed by the $\alpha(M_1 M_2)/12$ number of pixels in terms of the largest pixel intensities in \bar{F}_{ij} , where α is a hyperparameter that allows changing the region's size. For example, $\alpha = 1$ means the 1/12 of the image is scanned, which is the size of a rough area for each of the 12 body parts such as *beak*, *back* and *throat* as given in Tables 4 and 5.

Table 2 gives the CLM scores for different choices of α . For $\alpha = 1$, only 13.3% of the time the centre pixel for a concept falls into the highly activated region Ω_{ij} . Even increasing α to 6, which means half of the image is scanned, the centre pixels of the individual concepts are still not in the correct concept locations 41% of the time.

5. Discussion

Learning human-understandable concepts is a challenging task. Often concepts are highly correlated with other features. For example, although hooves clearly relate to the feet

American Redstart	Brown Pelican	Ivory Gull	Ovenbird	Bronzed Cowbird	Magnolia Warbler
					
white throat (✗)	perching-like shape (✗)	white forehead (✓)	striped breast (✓)	white belly (✗)	yellow underpart (✓)
white back (✗)	yellow nape (✗)	black breast (✗)	yellow throat (✗)	black upper tail (✓)	mainly yellow (✓)
plain head (✓)	multicolour back (✗)	all purpose bill (✗)	buff upper tail (✗)	buff upper tail (✗)	grey under tail (✗)
black throat (✓)	mainly yellow (✗)	white under tail (✓)	striped wing (✗)	buff breast (✗)	black crown (✗)
mainly white (✗)	shorter-than-head bill (✗)	perching-like shape (✗)	brown forehead (✓)	black back (✓)	plain head (✗)

Figure 3. Randomly selected test images from different classes and the top 5 most important concepts for their classification by the post-hoc CBMs. In particular, symbols ✓ and ✗ are for concept existence and absence in the ground-truth label, respectively.

of animals, a group of hooved animals often share other common features (e.g., they are often quadrupeds that feed on grass). Although these other features might help to identify the concepts, it is not very helpful to be told that an important concept for determining that the image represents a cow is hooves if the hooves are not visible in the image. It is therefore important to check that the human-understandable concepts exist in the image and when they exist the network is finding them in the correct location. By providing metrics and benchmarks we hope this will provide an important stimulus to develop models with improved alignment.

Concept-based XAI methodologies showcase either global or local explainability of their proposed techniques, depending on their model setting. For instance, with the traditional CBMs, the quality of concept predictions can be assessed just like the final class predictions since the concept labels should be readily available for them to work in the first place. On the other hand, when concept labels are not available as in the post-hoc CBMs case, the main evaluation is on the classifier weights (i.e., θ) as the global expicator. This evaluation is further supported by model editing experiments. However, none of these experiments make a comparison between these weights and the ground-truth labels (i.e., V for the CUB dataset), hindering their reliability. Therefore, we propose CoAM and CGIM to visualise and evaluate the global explainability of concept-based XAI methodologies more rigorously and expose the alignment between the global concept explanations and ground-truth labels. Moreover, when the ground-truth concept labels are not available, for methods such as post-hoc CBMs, they are unable to do concept predictions and instead project concepts to the concept space, which would prevent a direct concept prediction evaluation even if the test set had ground-truth labels for each concept. Our CEM and CLM allows local concept evaluation regardless of the training settings of methodologies, i.e., whether they predict or project concepts to the concept space.

We should note, however, that the metrics we provide do not directly measure the usefulness of the concepts as an ex-

planation. Rather they act as sanity checks that the concepts are correctly identified in the images. Also, success on the benchmark is not necessarily the top objective of a network; for example, the motivation of the post-hoc CBMs (Yuksekgonul et al., 2023) was to provide a low-cost means of building traditional CBMs. A traditional CBM that uses per-image annotation will surely have a superior performance on our benchmark, but it may be too costly to train this on other datasets.

Our findings raise important questions about the utility of current concept-based explanation methodologies in providing spatially grounded explanations for image-based tasks. While these models offer some degree of interpretability by linking decisions to human-understandable concepts, their failure to predict and localise concepts correctly can lead to misleading interpretations. This highlights the importance of more rigorous evaluation criteria such as CGIM, CEM and CLM and the development of models that prioritise both concept prediction accuracy and spatial interpretability.

6. Conclusion

In this paper, we proposed three novel metrics, i.e., CGIM, CEM and CLM, for concept-based XAI systems. CGIM provides a way to measure the global concept alignment ability of concept-based XAI techniques. CEM and CLM are introduced for local importance evaluation, testing if highly important concepts proposed by XAI techniques exist and can be correctly localised in a given test image, respectively. Employing these three metrics, we benchmarked post-hoc CBMs on the CUB dataset. Our experiments demonstrated significant limitations in current post-hoc methods, with many concepts and classes found to be weakly or even negatively correlated with their ground-truth labels by CGIM. Moreover, many concepts presented as highly important are not found to be present in test images by CEM, and their concept activations fail to align with the expected regions of the input images by CLM. As the field of XAI continues to evolve, it is essential to ensure that methods not only provide understandable concepts but also accurately predict

and localise these concepts within input data. Future work may focus on improving both the concept prediction and spatial localisation capabilities of concept-based XAI methods, ensuring that they can offer reliable and interpretable insights across diverse applications.

7. Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023.
- Aysel, H. I., Cai, X., and Prugel-Bennett, A. Multilevel Explainable Artificial Intelligence: Visual and Linguistic Bonded Explanations. *IEEE Transactions on Artificial Intelligence*, 2023.
- Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.
- Buhrmester, V., Münch, D., and Arens, M. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*, 3(4):966–989, 2021.
- Chattpadhyay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847. IEEE, 2018.
- Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019.
- Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., and Holzinger, A. Explainable AI: the new 42? In *International Cross-domain Conference for Machine Learning and Knowledge Extraction*, pp. 295–303. Springer, 2018.
- Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., and Hussain, A. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74, 2024.
- Havasi, M., Parbhoo, S., and Doshi-Velez, F. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35:23386–23397, 2022.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pp. 2668–2677. PMLR, 2018.
- Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept Bottleneck Models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.
- Oikarinen, T., Das, S., Nguyen, L. M., and Weng, T.-W. Label-free Concept Bottleneck Models. In *The Eleventh International Conference on Learning Representations*, 2023.
- Ramaswamy, H. G. et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 983–991, 2020.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- Shin, S., Jo, Y., Ahn, S., and Lee, N. A closer look at the intervention procedure of concept bottleneck models. In *International Conference on Machine Learning*, pp. 31504–31520. PMLR, 2023.
- Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Steinmann, D., Stammer, W., Friedrich, F., and Kersting, K. Learning to Intervene on Concept Bottlenecks. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46556–46571. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/steinmann24a.html>.

Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., and Viergever, M. A. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022.

Vandenhirtz, M., Laguna, S., Marcinkevičs, R., and Vogt, J. E. Stochastic Concept Bottleneck Models. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024. URL <https://openreview.net/forum?id=8jG3Y0xX7b>.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. Score-CAM: Score-weighted Visual Explanations for Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 24–25, 2020.

Yuksekgonul, M., Wang, M., and Zou, J. Post-hoc Concept Bottleneck Models. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

A. Related Work

This section recalls concept-based methodologies for XAI and examines existing variants of class activation mapping (CAM) highlighting the need for a dedicated approach to concept visualisation (which can be addressed by our CoAM).

Network dissection (Bau et al., 2017) is one of the well-known concept-based approaches, where individual neurons in a network are examined to identify their correspondence to human-understandable concepts like *edges*, *textures*, or *objects*. By aligning neuron activations with segmentation-annotated images, network dissection quantifies how well a model’s internal representations map to meaningful concepts. However, this method is computationally expensive and data-intensive, requiring large and richly labelled datasets to accurately associate neurons with interpretable concepts. Despite its valuable insights, these limitations have prompted the development of more efficient and flexible methods, such as CAVs and CBMs.

Testing with CAVs (TCAV) framework (Kim et al., 2018) introduced CAVs to explain model predictions based on high-level human-interpretable concepts. CAVs represent directions in the latent space of a model corresponding to specific concepts, allowing for sensitivity analysis. By perturbing an input in the direction of a concept vector, TCAV measures how much the model’s prediction depends on that specific concept, offering quantitative insights into the reliance on different concepts for a given task. TCAV has been applied in several fields to assess whether models depend on sensitive attributes like gender or race when making decisions. Recent adaptations have improved the computational efficiency and robustness of CAVs when applied to large-scale models (Ghorbani et al., 2019). However, TCAV can only unveil the global effect of concepts on examined classes and not on individual samples. Therefore, it is unable to directly assess the concept predictions or provide spatial concept localisation for individual images.

CBMs (Koh et al., 2020) offers a significantly different approach to interpretability. They enforce that intermediate representations of the model correspond to human-understandable concepts, such as attributes (e.g., *colour*, *shape*, *part*) of objects in an image. By constraining the model to predict based on these explicit concepts, CBMs inherently provide an interpretable mechanism for understanding decisions. This makes it easier to debug and correct errors by diagnosing the model’s performance on individual concepts. Recent work in CBMs has focused on improving robustness, especially when concept labels are noisy or incomplete. For instance, in Label-free CBMs (Oikarinen et al., 2023), a method was proposed using unsupervised techniques to learn concept bottlenecks, thereby extending the applicability of CBMs to scenarios where manual labelling is expensive or impractical. Despite their interpretability, CBMs typically lack the ability to provide spatial visualisations, limiting their usefulness in tasks that require precise localisation of important concepts.

The multilevel XAI method in (Aysel et al., 2023) offers solutions for both expensive annotation needs and single-level output drawbacks of CBMs. The cost-effective solution to CBMs is achieved by only requiring class-wise concept annotations rather than per-image. Moreover, the multilevel XAI method provides concept-wise heatmaps by-product handling the single-level limitation of CBMs. To be more precise, different from other CBM approaches, the explanations by the multilevel XAI method are not only raw concept values, but also each concept comes with its saliency map that highlights the region in the image activated by that concept. The authors in Aysel et al. (2023) have also shown the possibility of concept intervention on the input dimension, which is much more intuitive than the concept dimension. To give an example, in other CBMs, one may tweak the concept value, say, “white” at the bottleneck layer to flip the prediction, say, from polar bear to grizzly bear. In the multilevel XAI method, one can convert the white colour region in the image to brown to achieve the same flipping, which is more intuitive and reliable.

A breakthrough in visual explanations came with the introduction of CAM (Zhou et al., 2016), which provides spatial localisation by computing class-specific activation maps that highlight the regions of an image most relevant for a given prediction. CAM operates by utilising the output of GAP layers in CNNs, enabling the generation of heatmaps that represent regions crucial for the final classification. This approach was generalised in Grad-CAM (Selvaraju et al., 2017), which makes use of the gradients flowing into the final convolutional layer to visualise where the model “looks” when making a decision. Grad-CAM extends CAM to more general architectures without requiring specific layers like GAP. However, Grad-CAM does not always provide sharp localisation, especially when multiple objects are present in the image. Grad-CAM++ (Chattopadhyay et al., 2018) addresses this limitation by refining the localisation to better handle multiple instances of objects, offering a more fine-grained interpretation. Further extensions include Score-CAM (Wang et al., 2020), which eliminates the dependency on gradients, instead using the activations themselves to weigh different regions of the input. This addresses some of the instability associated with gradient-based methods but comes with increased computational overhead. Other advancements like Ablation-CAM (Ramaswamy et al., 2020) explore removing parts of the model and input to measure their impact on predictions, thus improving interpretability.

Table 3. Classification accuracy of the reproduced post-hoc CBMs with different settings for the parameters λ , N_p , and N_n .

$\lambda \backslash N_p = N_n$	50	100
0.001	26.7	52.2
0.01	34.1	44.9
0.1	29.1	41.5
1	25.5	59.1
10	25.3	58.7
Traditional model w/o bottleneck	75.4	

B. Limitations

There are drawbacks to the metrics CEM and CLM that we propose. The CEM can only be used on datasets where we have per-image annotations of the concepts for a test set (note that for our case, only a small number of concepts like the top $l \ll L$ are required per-image, and therefore is cheap). This limits its use to a very small number of datasets. Having a metric limited to one (or a small number of datasets) runs the risk that models are developed that overfit to that particular dataset. The CLM requires knowledge of the location of the concepts. In fact, the concept locations were not given and we had to do a “best guess” approximation of whether the concepts found in the “saliency maps” overlap with the real concept locations. It is also debatable whether the heatmaps we obtained by weighting the feature maps before doing GAP correctly capture the location of the concepts. In our judgment, this seems as fair an estimate of the position as we can make. We feel there is considerable value in visualising the location of a concept through the use of heatmaps. In [Aysel et al. \(2023\)](#), the authors built saliency maps for each concept, but there they aligned each feature map to a concept which prevented cross-contamination between concept locations. By providing visualisations of the parts of the image that activated the concept, it made it much easier to assess the alignment of concepts in that model. We have attempted to provide a similar visualisation for the post-hoc CBMs ([Yuksekgonul et al., 2023](#)), although as this is not part of the design of that model the visualisation may not be perfect. Finally, reducing the assessment of alignment to a couple of numbers loses a lot of fine-grain detail. As we illustrated, we can get a better understanding of the failure of the network by examining the performance in more detail, for example, by plotting histograms of the CBMs results to identify particularly poor concepts, or by visualising the locations of the features to understand what concepts might be being learnt.

Despite those drawbacks, we believe that proposing a new benchmark for assessing concept alignment has the potential to concentrate the effort of researchers on improving the performance of concept-based XAI systems. As we have illustrated, the performance of post-hoc CBMs is surprisingly poor. Without doing a systematic analysis of this alignment, it is easy to overlook this problem and believe that an XAI system is more powerful than it actually is. Our hope is that by introducing new metrics and benchmarks we can improve the accuracy of future concept-based XAI systems.

C. Post-hoc CBMs Reproduction – Details

By employing the same model as the feature extractor and following the same steps for CAVs and classifier training, we reproduce the results of post-hoc CBMs ([Yuksekgonul et al., 2023](#)) with various hyperparameter combinations. There are two hyperparameters to tune during the SVM training for CAV learning, i.e., N_p and N_n (the number of positive and negative images per concept), which we set to 50 and 100, respectively. The other hyperparameter is the regularisation parameter λ in SVM, which controls the trade-off between maximising the margin that separates classes and minimising classification errors on the training data. A low λ value allows the model to prioritise a wider margin, even if some data points are misclassified, making the model more robust to noise and potentially improves its generalisation of new data. In contrast, a high λ value forces the SVM to minimise the training error, making it less tolerant of misclassifications and resulting in a narrower margin. While a high λ can lead to more accurate training performance, it may also increase the risk of overfitting, as the model becomes more sensitive to individual data points. Thus, λ helps balance the SVM’s complexity and flexibility, impacting its ability to generalise well.

We train SVM with λ values ranging from 0.001 to 10. Table 3 shows the classification accuracy of the classifier h with various concept banks obtained by these hyperparameter combinations. For the experiments in the main paper, we employ the model with the best classification accuracy 59.1%, which is achieved when $N_p = N_n = 100$ and $\lambda = 1$. This result is very close to the accuracy 58.8% reported in the seminal work ([Yuksekgonul et al., 2023](#)). Note that there is more than 15% accuracy loss in comparison to the traditional model, i.e., the one without a concept bottleneck (i.e., $\textcircled{a} + \textcircled{c}$ in Figure 1), for

the sake of obtaining an interpretable model via concept bottleneck.

D. Concept Localisation Evaluation – Figures and Tables

Figure 4 presents some examples of class and concept visualisation by using our CoAM. The matching between the concept groups and the body parts for the CUB dataset is given in Tables 4 and 5.

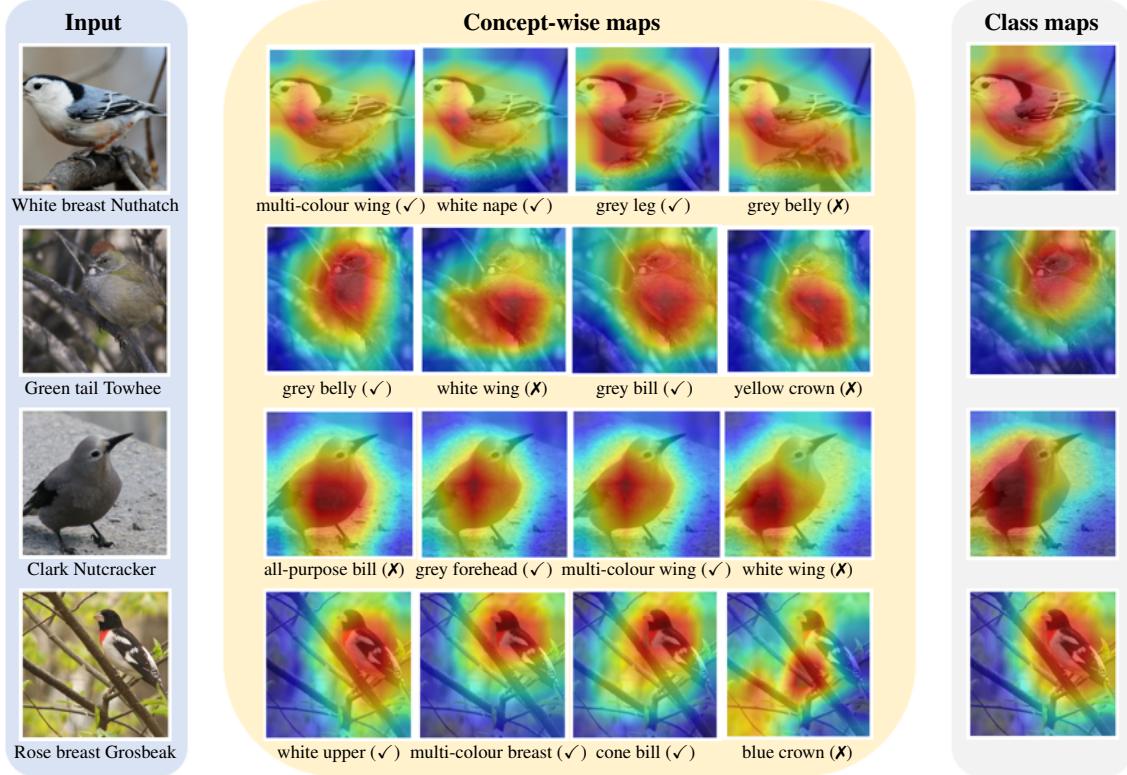


Figure 4. Class and concept visualisation with our CoAM. All images (on the left) are correctly classified and their class-wise saliency maps are given on the right. The four most important concepts under CEM for the given classifications and their individual saliency maps are given in the middle. In particular, symbols ✓ and ✗ are for concept existence and absence in the ground-truth label, respectively.

Table 4. The number of concepts, grouped based on their types, and mapped to the body parts (see Table 5 for more details).

<i>Part \ Type</i>	Color	Pattern	Shape	Total
Back	6	3	—	9
Beak	3	2	4	9
Belly	6	1	—	7
Breast	6	3	—	9
Crown	6	—	—	6
Head	6	2	—	8
Eye	1	—	—	1
Leg	3	—	—	3
Wing	6	4	2	12
Nape	6	—	—	6
Tail	10	3	1	14
Throat	5	—	—	5
Others	18	—	5	23
Total	82	18	12	112

Table 5. Details of the concepts and the body parts they are mapped to.

Part \ Type	Color	Pattern (length for beak)	Shape
Back	brown, grey, yellow, black, white, buff	<i>solid, striped, multi-coloured</i>	—
Beak	grey, black, buff	<i>same-as-head, shorter-than-head</i>	dagger, hooked-seabird, all-purpose, cone
Belly	brown, grey, yellow, black, white, buff	<i>solid</i>	—
Breast	brown, grey, yellow, black, white, buff	<i>solid, striped, multi-coloured</i>	—
Crown	blue, brown, grey, yellow, black, white	—	—
Head	blue, brown, grey, yellow, black, white	<i>eyebrow, plain</i>	—
Eye	black	—	—
Leg	grey, black, buff	—	—
Wing	brown, grey, yellow, black, white, buff	<i>solid, spotted, striped, multi-coloured</i>	rounded, pointed
Nape	brown, grey, yellow, black, white, buff	—	—
Tail	brown, grey, black, white, buff	<i>solid, striped, multi-colored</i>	notched
Throat	grey, yellow, black, white, buff	—	—

E. Concept Visualisation Algorithm

Concept activation maps in \mathbf{F}_i for image \mathbf{X}_i are obtained as summarised in Algorithm 1. These maps are at smaller resolutions since they are at a late layer of the trained DNN. Algorithm 2 below demonstrates the steps to upsample these maps to the input size and we use them to mask the input image \mathbf{X}_i .

Algorithm 2 Feature Visualisation in CoAM

```

1: Input:
    • Boolean flag coloured for generating coloured heatmaps.
    • Threshold value threshold for binary heatmaps.
    • Opacity level  $\beta$  for superimposed heatmaps.
    • Input image  $\mathbf{X}_i \in \mathbb{R}^{M_1 \times M_2 \times M_3}$ .
    • Concept activation map  $\mathbf{F}_i \in \mathbb{R}^{H \times W \times L}$  of  $\mathbf{X}_i$ . //  $L$  is the number of concepts

2: Output: Set of superimposed images  $\bar{\mathcal{S}} \in \mathbb{R}^{M_1 \times M_2 \times M_3 \times L}$ .


---


3: Initialize an empty list of superimposed images  $\bar{\mathcal{S}}$ 

4: for each spatial projection map  $j$  in  $\mathbf{F}_i$  do
5:   heatmap = resize ( $\mathbf{F}_{ij}, (M_1, M_2)$ ) // Generate heatmap with size of  $(M_1, M_2)$ 
6:   if coloured then
7:     jet_heatmap = apply_colormap (heatmap, "jet") // Convert the heatmap to an RGB image
8:     superimposed_img =  $\beta \cdot \text{jet\_heatmap} + \mathbf{X}_i$  // Overlay heatmap on the original image  $\mathbf{X}_i$ 
9:     Append superimposed_img to  $\bar{\mathcal{S}}$ 
10:  else
11:    binary_heatmap = binary_threshold (heatmap, threshold)
12:    superimposed_img =  $\mathbf{X}_i \odot \text{binary\_heatmap}$  // Overlay heatmap on the original image  $\mathbf{X}_i$ 
13:    Append superimposed_img to  $\bar{\mathcal{S}}$ 
14:  end if
15: end for
16: return  $\bar{\mathcal{S}}$ 

```

F. CGIM Scores

Figure 2 presents the histograms of the CGIM scores. The full list of CGIM scores for every concept and every class is presented below in Table 6 and in Tables 7 and 8, respectively.

Table 6. Full list of CGIM scores for concepts in CUB dataset (Wah et al., 2011) with reproduced post-hoc CBMs (Yuksekgonul et al., 2023).

<i>CGIM</i> <i>Concept</i>	$\rho_j^{\text{CGIM}_1}$	$\rho_j^{\text{CGIM}_2}$	$\rho_j^{\text{CGIM}_3}$	<i>CGIM</i> <i>Concept</i>	$\rho_j^{\text{CGIM}_1}$	$\rho_j^{\text{CGIM}_2}$	$\rho_j^{\text{CGIM}_3}$
1: Dagger beak	0.54	0.05	0.41	57: Yellow forehead colour	0.41	0.22	0.47
2: Hooked seabird beak	0.45	0.08	0.34	58: Black forehead colour	0.39	0.25	0.35
3: All-purpose beak	0.37	0.67	0.46	59: White forehead colour	0.52	-0.13	0.07
4: Cone beak	0.47	0.23	0.43	60: Brown under tail colour	-0.01	-0.16	0.23
5: Brown wing colour	0.22	-0.05	0.28	61: Grey under tail colour	0.36	-0.21	0.30
6: Grey wing colour	-0.13	-0.25	0.31	62: Black under tail colour	-0.11	0.35	0.24
7: Yellow wing colour	0.42	-0.04	0.42	63: White under tail colour	0.02	-0.22	0.23
8: Black wing colour	0.42	-0.02	0.33	64: Buff under tail colour	0.46	-0.21	0.22
9: White wing colour	-0.05	0.08	0.30	65: Brown nape colour	0.22	0.00	0.39
10: Buff wing colour	0.22	0.12	0.30	66: Grey nape colour	0.27	0.34	0.38
11: Brown upper-part colour	0.09	0.03	0.26	67: Yellow nape colour	0.35	0.23	0.52
12: Grey upper-part colour	-0.36	0.16	0.23	68: Black nape colour	0.37	0.00	0.20
13: Yellow upper-part colour	0.46	0.09	0.45	69: White nape colour	0.38	0.16	0.28
14: Black upper-part colour	-0.04	0.48	0.30	70: Buff nape colour	0.49	-0.24	0.14
15: White upper-part colour	0.32	-0.17	0.23	71: Brown belly colour	0.56	-0.15	0.22
16: Buff upper-part colour	0.27	-0.13	0.20	72: Grey belly colour	0.52	-0.12	0.41
17: Brown underpart colour	0.47	0.02	0.36	73: Yellow belly colour	0.20	0.43	0.27
18: Grey underpart colour	0.19	0.04	0.37	74: Black belly colour	0.48	0.21	0.46
19: Yellow underpart colour	0.40	0.55	0.51	75: White belly colour	0.01	0.33	0.21
20: Black underpart colour	0.55	0.05	0.29	76: Buff belly colour	0.48	-0.31	-0.07
21: White underpart colour	-0.04	0.03	0.26	77: Rounded wing shape	0.05	0.03	0.38
22: Buff underpart colour	0.18	-0.10	0.20	78: Pointed wing shape	0.51	-0.16	0.34
23: Solid breast pattern	0.38	0.39	0.40	79: Small size	-0.09	0.37	0.28
24: Striped breast pattern	0.33	0.00	0.33	80: Medium size	0.23	0.06	0.29
25: Multi-coloured breast pattern	0.51	-0.19	0.25	81: Very small size	0.57	-0.21	0.26
26: Brown back colour	0.43	0.04	0.23	82: Duck-like shape	0.42	0.41	0.57
27: Grey back colour	0.38	-0.21	0.18	83: Perching-like shape	-0.18	0.47	0.11
28: Yellow back colour	0.31	0.03	0.32	84: Solid back pattern	0.53	-0.13	0.20
29: Black back colour	0.22	0.60	0.38	85: Striped back pattern	0.33	-0.09	0.33
30: White back colour	0.34	-0.23	0.20	86: Multi-coloured back pattern	0.25	-0.38	0.39
31: Buff back colour	0.01	-0.24	0.27	87: Solid tail pattern	0.64	0.41	0.45
32: Notched tail shape	0.21	-0.10	0.31	88: Striped tail pattern	0.42	-0.30	0.21
33: Brown upper tail colour	0.33	-0.10	0.23	89: Multi-coloured tail pattern	0.25	-0.41	0.21
34: Grey upper tail colour	0.08	-0.19	0.23	90: Solid belly pattern	0.35	0.47	0.37
35: Black upper tail colour	0.44	0.50	0.50	91: Brown primary colour	0.12	0.19	0.26
36: White upper tail colour	-0.13	0.19	0.25	92: Grey primary colour	0.45	0.00	0.27
37: Buff upper tail colour	0.51	-0.01	0.39	93: Yellow primary colour	0.12	0.44	0.31
38: Head pattern eyebrow	0.41	-0.12	0.37	94: Black primary colour	0.47	0.11	0.39
39: Head pattern plain	0.67	-0.05	0.34	95: White primary colour	0.58	-0.05	0.26
40: Brown breast colour	0.39	0.19	0.23	96: Buff primary colour	0.32	-0.29	0.26
41: Grey breast colour	0.58	-0.05	0.35	97: Grey leg colour	0.55	0.22	0.43
42: Yellow breast colour	0.47	0.26	0.46	98: Black leg colour	0.43	-0.09	0.29
43: Black breast colour	0.22	0.00	0.25	99: Buff leg colour	0.17	0.00	0.27
44: White breast colour	0.32	-0.19	0.30	100: Grey bill colour	0.43	0.06	0.41
45: Buff breast colour	0.32	-0.26	0.17	101: Black bill colour	0.36	0.39	0.38
46: Grey throat colour	0.38	-0.11	0.37	102: Buff bill colour	0.52	-0.43	-0.13
47: Yellow throat colour	0.26	0.23	0.29	103: Blue crown colour	0.37	0.27	0.50
48: Black throat colour	0.52	-0.03	0.28	104: Brown crown colour	0.39	0.16	0.32
49: White throat colour	0.45	0.14	0.32	105: Grey crown colour	0.39	-0.24	0.19
50: Buff throat colour	0.42	-0.29	0.22	106: Yellow crown colour	0.25	0.20	0.36
51: Black eye colour	-0.63	0.82	-0.33	107: Black crown colour	0.45	0.18	0.39
52: Head size beak	0.40	0.26	0.38	108: White crown colour	0.42	-0.13	0.01
53: Shorten than head size beak	-0.08	0.43	0.23	109: Solid wing pattern	0.61	0.39	0.60
54: Blue forehead colour	0.26	0.36	0.50	110: Spotted wing pattern	0.48	0.04	0.49
55: Brown forehead colour	0.43	-0.20	0.32	111: Striped wing pattern	0.24	-0.12	0.37
56: Grey forehead colour	0.62	-0.20	0.06	112: Multi-coloured wing pattern	0.27	0.17	0.40

Table 7. Full list of CGIM scores for classes in CUB dataset (Wah et al., 2011) with reproduced post-hoc CBMs (Yuksekgonul et al., 2023).

<i>CGIM</i> <i>Class</i>	$\rho_k^{\text{CGIM}_1}$	$\rho_k^{\text{CGIM}_2}$	$\rho_k^{\text{CGIM}_3}$	<i>CGIM</i> <i>Class</i>	$\rho_k^{\text{CGIM}_1}$	$\rho_k^{\text{CGIM}_2}$	$\rho_k^{\text{CGIM}_3}$
1: Black footed Albatross	0.24	-0.20	0.03	51: Horned Grebe	0.27	-0.24	0.24
2: Laysan Albatross	0.30	0.04	0.11	52: Pied billed Grebe	0.32	-0.12	0.32
3: Sooty Albatross	0.28	-0.15	0.17	53: Western Grebe	0.25	-0.04	0.24
4: Groove billed Ani	0.28	0.37	0.31	54: Blue Grosbeak	0.20	0.20	0.31
5: Crested Auklet	0.25	0.18	0.26	55: Evening Grosbeak	0.40	0.24	0.35
6: Least Auklet	0.26	0.07	0.14	56: Pine Grosbeak	0.14	-0.01	0.20
7: Parakeet Auklet	0.25	0.12	0.20	57: Rose breasted Grosbeak	0.31	0.17	0.31
8: Rhinoceros Auklet	0.36	0.00	0.22	58: Pigeon Guillemot	0.36	0.05	0.20
9: Brewer Blackbird	0.29	0.21	0.28	59: California Gull	0.27	0.16	0.24
10: Red winged Blackbird	0.27	0.36	0.38	60: Glaucous winged Gull	0.37	0.04	0.26
11: Rusty Blackbird	0.17	-0.09	0.32	61: Heermann Gull	0.26	0.12	0.32
12: Yellow headed Blackbird	0.36	0.32	0.41	62: Herring Gull	0.27	0.01	0.17
13: Bobolink	0.27	0.23	0.34	63: Ivory Gull	0.33	0.33	0.30
14: Indigo Bunting	0.15	0.16	0.23	64: Ring billed Gull	0.34	0.20	0.29
15: Lazuli Bunting	0.16	-0.07	0.27	65: Slaty backed Gull	0.25	-0.04	0.23
16: Painted Bunting	0.28	0.01	0.36	66: Western Gull	0.16	0.15	0.20
17: Cardinal	0.22	0.00	0.17	67: Anna Hummingbird	0.16	-0.15	0.17
18: Spotted Catbird	-0.02	-0.38	0.14	68: Ruby throated Hummingbird	0.28	-0.22	0.17
19: Gray Catbird	0.27	0.17	0.28	69: Rufous Hummingbird	0.19	-0.09	0.16
20: Yellow breasted Chat	0.39	0.01	0.33	70: Green Violetear	0.20	-0.14	0.12
21: Eastern Towhee	0.31	0.00	0.33	71: Long tailed Jaeger	0.24	-0.16	0.13
22: Chuck will Widow	0.23	0.05	0.32	72: Pomarine Jaeger	0.18	-0.28	0.36
23: Brandt Cormorant	0.27	0.02	0.16	73: Blue Jay	0.30	0.02	0.36
24: Red faced Cormorant	0.22	0.15	0.24	74: Florida Jay	0.32	-0.06	0.24
25: Pelagic Cormorant	0.05	0.06	0.16	75: Green Jay	0.32	0.13	0.38
26: Bronzed Cowbird	0.22	0.19	0.32	76: Dark eyed Junco	0.22	-0.05	0.20
27: Shiny Cowbird	0.33	0.25	0.24	77: Tropical Kingbird	0.28	0.13	0.36
28: Brown Creeper	0.30	0.11	0.27	78: Gray Kingbird	0.17	0.06	0.29
29: American Crow	0.28	0.38	0.28	79: Belted Kingfisher	0.25	0.01	0.19
30: Fish Crow	0.29	0.44	0.33	80: Green Kingfisher	0.27	-0.04	0.23
31: Black billed Cuckoo	0.26	0.04	0.21	81: Pied Kingfisher	0.14	0.41	0.19
32: Mangrove Cuckoo	0.23	-0.22	0.17	82: Ringed Kingfisher	0.21	-0.18	0.17
33: Yellow billed Cuckoo	0.26	-0.02	0.29	83: White breasted Kingfisher	0.39	0.13	0.28
34: Gray-crowned Rosy Finch	0.33	0.14	0.21	84: Red legged Kittiwake	0.30	0.17	0.25
35: Purple Finch	0.05	0.00	0.16	85: Horned Lark	0.31	0.13	0.23
36: Northern Flicker	0.35	0.02	0.29	86: Pacific Loon	0.48	-0.13	0.17
37: Acadian Flycatcher	0.32	-0.15	0.33	87: Mallard	0.27	-0.14	0.23
38: Great Crested Flycatcher	0.37	0.00	0.30	88: Western Meadowlark	0.43	0.05	0.38
39: Least Flycatcher	0.14	-0.15	0.16	89: Hooded Merganser	0.46	-0.03	0.38
40: Olive sided Flycatcher	0.21	-0.20	0.25	90: Red breasted Merganser	0.24	-0.10	0.29
41: Scissor tailed Flycatcher	0.25	-0.01	0.25	91: Mockingbird	0.22	-0.03	0.10
42: Vermilion Flycatcher	0.12	-0.01	0.07	92: Nighthawk	0.32	-0.12	0.26
43: Yellow bellied Flycatcher	0.37	-0.22	0.18	93: Clark Nutcracker	0.32	0.20	0.31
44: Frigatebird	0.25	0.04	0.18	94: White breasted Nuthatch	0.24	0.11	0.34
45: Northern Fulmar	0.14	0.12	0.16	95: Baltimore Oriole	0.27	0.07	0.30
46: Gadwall	0.32	-0.04	0.36	96: Hooded Oriole	0.27	0.17	0.19
47: American Goldfinch	0.38	0.37	0.37	97: Orchard Oriole	0.30	0.08	0.17
48: European Goldfinch	0.45	0.13	0.32	98: Scott Oriole	0.43	0.23	0.43
49: Boat tailed Grackle	0.32	0.08	0.25	99: Ovenbird	0.23	0.00	0.13
50: Eared Grebe	0.26	-0.10	0.14	100: Brown Pelican	0.22	-0.45	0.06

Table 8. Full list of CGIM scores for classes in CUB dataset (Wah et al., 2011) with reproduced post-hoc CBMs (Yuksekgonul et al., 2023) (Table 7 continued).

<i>CGIM</i> <i>Class</i>	$\rho_k^{\text{CGIM}_1}$	$\rho_k^{\text{CGIM}_2}$	$\rho_k^{\text{CGIM}_3}$	<i>CGIM</i> <i>Class</i>	$\rho_k^{\text{CGIM}_1}$	$\rho_k^{\text{CGIM}_2}$	$\rho_k^{\text{CGIM}_3}$
101: White Pelican	0.19	0.03	0.05	151: Black-capped Vireo	0.25	0.02	0.27
102: Western Wood Pewee	0.26	0.05	0.23	152: Blue-headed Vireo	0.35	0.03	0.30
103: Sayornis	0.25	-0.16	0.13	153: Philadelphia Vireo	0.19	0.00	0.20
104: American Pipit	0.27	0.09	0.27	154: Red-eyed Vireo	0.20	0.00	0.32
105: Whip poor will	0.24	-0.21	0.34	155: Warbling Vireo	0.34	0.10	0.14
106: Horned Puffin	0.37	0.24	0.35	156: White-eyed Vireo	0.31	0.05	0.25
107: Common Raven	0.32	0.24	0.33	157: Yellow-throated Vireo	0.21	0.16	0.13
108: White-necked Raven	0.35	0.22	0.35	158: Bay-breasted Warbler	0.32	-0.12	0.22
109: American Redstart	0.28	0.12	0.24	159: Black-and-white Warbler	0.35	0.15	0.30
110: Geococcyx	0.24	-0.10	0.24	160: Black-throated Blue Warbler	0.26	0.02	0.27
111: Loggerhead Shrike	0.33	0.28	0.31	161: Blue-winged Warbler	0.26	0.36	0.35
112: Great Grey Shrike	0.28	0.19	0.31	162: Canada Warbler	0.27	0.06	0.17
113: Baird's Sparrow	0.24	0.12	0.31	163: Cape May Warbler	0.28	-0.05	0.30
114: Black-throated Sparrow	0.33	0.08	0.22	164: Cerulean Warbler	0.10	-0.10	0.21
115: Brewer's Sparrow	0.25	0.23	0.12	165: Chestnut-sided Warbler	0.27	0.02	0.35
116: Chipping Sparrow	0.28	0.02	0.21	166: Golden-winged Warbler	0.40	0.20	0.39
117: Clay-colored Sparrow	0.22	0.07	0.16	167: Hooded Warbler	0.26	0.17	0.43
118: House Sparrow	0.36	0.11	0.24	168: Kentucky Warbler	0.12	0.17	0.14
119: Field Sparrow	0.14	0.07	0.19	169: Magnolia Warbler	0.41	0.19	0.35
120: Fox Sparrow	0.22	0.10	0.22	170: Mourning Warbler	0.33	0.06	0.24
121: Grasshopper Sparrow	0.30	0.08	0.34	171: Myrtle Warbler	0.34	0.01	0.22
122: Harris's Sparrow	0.22	-0.01	0.31	172: Nashville Warbler	0.27	0.18	0.33
123: Henslow's Sparrow	0.34	0.10	0.27	173: Orange-crowned Warbler	0.21	0.00	0.18
124: Le Conte's Sparrow	0.36	0.09	0.29	174: Palm Warbler	0.18	-0.05	0.09
125: Lincoln Sparrow	0.28	0.22	0.21	175: Pine Warbler	0.23	0.23	0.19
126: Nelson's Sharp-tailed Sparrow	0.31	-0.13	0.17	176: Prairie Warbler	0.20	0.16	0.24
127: Savannah Sparrow	0.35	0.13	0.32	177: Prothonotary Warbler	0.33	0.37	0.42
128: Seaside Sparrow	0.15	-0.24	0.07	178: Swainson's Warbler	0.31	0.05	0.20
129: Song Sparrow	0.37	0.20	0.29	179: Tennessee Warbler	0.17	0.00	0.20
130: Tree Sparrow	0.38	0.10	0.20	180: Wilson's Warbler	0.20	0.25	0.34
131: Vesper Sparrow	0.16	0.12	0.17	181: Worm-eating Warbler	0.33	0.02	0.32
132: White-crowned Sparrow	0.40	0.12	0.33	182: Yellow Warbler	0.33	0.37	0.34
133: White-throated Sparrow	0.18	0.00	0.27	183: Northern Waterthrush	0.23	0.08	0.24
134: Cape Glossy Starling	0.40	0.14	0.34	184: Louisiana Waterthrush	0.27	0.01	0.19
135: Bank Swallow	0.17	-0.15	0.19	185: Bohemian Waxwing	0.29	0.23	0.28
136: Barn Swallow	0.37	0.09	0.22	186: Cedar Waxwing	0.40	0.12	0.25
137: Cliff Swallow	0.29	-0.13	0.12	187: American Three-toed Woodpecker	0.32	0.14	0.20
138: Tree Swallow	0.33	0.17	0.30	188: Pileated Woodpecker	0.31	0.27	0.26
139: Scarlet Tanager	0.30	0.10	0.09	189: Red-bellied Woodpecker	0.20	0.10	0.27
140: Summer Tanager	0.21	0.10	0.16	190: Red-cockaded Woodpecker	0.28	0.26	0.26
141: Arctic Tern	0.23	0.17	0.26	191: Red-headed Woodpecker	0.39	0.22	0.33
142: Black Tern	0.22	-0.07	0.27	192: Downy Woodpecker	0.27	0.24	0.27
143: Caspian Tern	0.12	0.20	0.26	193: Bewick Wren	0.20	0.25	0.30
144: Common Tern	0.17	0.11	0.26	194: Cactus Wren	0.39	0.17	0.29
145: Elegant Tern	0.22	0.23	0.25	195: Carolina Wren	0.33	0.27	0.30
146: Forsters Tern	0.17	0.25	0.24	196: House Wren	0.26	0.30	0.28
147: Least Tern	0.38	0.20	0.33	197: Marsh Wren	0.33	0.07	0.21
148: Green tailed Towhee	0.30	0.06	0.27	198: Rock Wren	0.24	0.19	0.22
149: Brown Thrasher	0.34	0.10	0.30	199: Winter Wren	0.20	0.32	0.16
150: Sage Thrasher	0.37	0.12	0.25	200: Common Yellowthroat	0.30	-0.04	0.30