Intervening in Black Box: Concept Bottleneck Model for Enhancing Human Neural Network Mutual Understanding

Nuoye Xiong[1], Anqi Dong[2,*], Ning Wang[1], Cong Hua[1], Guangming Zhu[1],
Lin Mei[3], Peiyi Shen[1], Liang Zhang[1†]
[1]Xidian University, China
[2]KTH Royal Institute of Technology, Sweden
[3]Donghai Laboratory, China
{nyx, ningwang, chua}@stu.xidian.edu.cn
{gmzhu, pyshen, liangzhang}@xidian.edu.cn
{meilin}@donghailab.com
{anqid}@kth.se

Abstract

Recent advances in deep learning have led to increasingly complex models with deeper layers and more parameters, reducing interpretability and making their decisions harder to understand. While many methods explain black-box reasoning, most lack effective interventions or only operate at sample-level without modifying the model itself. To address this, we propose the Concept Bottleneck Model for Enhancing Human-Neural Network Mutual Understanding (CBM-HNMU). CBM-HNMU leverages the Concept Bottleneck Model (CBM) as an interpretable framework to approximate black-box reasoning and communicate conceptual understanding. Detrimental concepts are automatically identified and refined (removed/replaced) based on global gradient contributions. The modified CBM then distills corrected knowledge back into the black-box model, enhancing both interpretability and accuracy. We evaluate CBM-HNMU on various CNN and transformer-based models across Flower-102, CIFAR-10, CIFAR-100, FGVC-Aircraft, and CUB-200, achieving a maximum accuracy improvement of 2.64% *and a maximum increase in average accuracy across* 1.03%*. Source code is available at: https://github.com/XiGuaBo/CBM-HNMU.*

1. Introduction

Interpretable machine learning has become a crucial research direction, particularly in image classification, where understanding model decisions is essential for trust and reliability. Neural networks for image classification can be broadly categorized into three types by their interpretability: (i) Black-box models, where decision-making is opaque and difficult to trace; (ii) Gray-box models, which offer partial transparency, allowing some explanation of decisions; and (iii) White-box models, which are fully transparent with completely traceable decision processes [1, 14]. While black-box models often achieve superior performance, their lack of interpretability poses significant challenges.

Many studies aim to achieve white-box-like interpretability while maintaining black-box models' performance. Methods such as Grad-CAM [31], Saliency Maps [33], and Feature Attack [8] use *feature attribution*, helping to explain its focus and reasoning in making predictions. Alternatively, methods include ACE [12], MOCE [19], IBD [46], and CRAFT [10] focus on *visual concept* explanations, linking model decisions to human-understandable patterns or attributes in data. These approaches provide more structured interpretation by associating model reasoning with recognizable concepts instead of abstract features. However, most existing methods focus on passive explanations rather than allowing active intervention to modify black-box behavior. This limitation weakens their ability to systematically refine model decision-making.

The Concept Bottleneck Model (CBM) [20] introduced an intervenable framework for concept-based reasoning, expanding the scope of interpretability methods. However, its reliance on *concept bottleneck* representations often leads to lower classification accuracy, compared to black-box models with the same backbone. Post-hoc CBM [45] was later introduced to mitigate this limitation by integrating black-box residual connections. Yet, this modification weakens overall interpretability and, even with interventions, Post-

*Co-first contributing author.
†Corresponding author.

hoc CBM still falls short of achieving the classification accuracy of original black-box model. More critically, its intervention mechanism is in the same vein as CBM and is constrained to adjusting the concept weight matrix, that heavily dependent on human priors. As a result, biases inherent in the backbone [47] cannot be effectively corrected.

Beyond CBM-based approaches, numerous studies employ interpretable surrogate structures and human-understandable concept space to explain black-box behavior without compromising performance [11, 27]. While these methods enable transparent reasoning, answering questions like "Why is it/is it not", they focus on explanation rather than intervention. Beyond CBM [24] addresses this limitation by using probe functions to map intermediate black-box representations to a human-understandable concept space. This approach identifies biases and aligns reasoning pathways by modifying the concept distribution based on prior knowledge. However, its interventions are restricted to the sample level, without directly modifying black-box parameters, and requires additional data to train the probe function for concept space projection.

Despite advances in interpretability of black-box models, existing methods primarily diagnose errors but offer limited mechanisms for correction. To address this gap, we propose the Concept Bottleneck Model for Enhancing Human Neural Network Mutual Understanding (CBM-HNMU). The contributions can be summarized as

1. **Integrated Explanation Framework:** CBM-HNMU integrates feature attribution and multimodal concept-based explanations to enhance interpretability in black-box inference. It aligns CBM with black-box hidden layer concepts and class predictions for more transparent reasoning and less approximation bias.
2. **Beyond Sample-Level Automatic Intervention:** CBM-HNMU introduces black-box model modification and error correction beyond individual samples. By computing global concept contributions based on class-level gradient, it autonomously identifies detrimental concepts for CBM intervention. Knowledge in intervention refines black-box model's parameters through knowledge distillation, correcting classification errors.
3. **Label-free Human-AI Mutual Understanding:** CBM-HNMU utilizes natural language concept bottlenecks generated by ChatGPT-3.5-Turbo [3] and cross-modal probe based on OpenAI-CLIP [29] to enhance interpretability and performance without requiring additional human annotations.

2. Related Works

2.1. Model Approximation

Model approximation employs external interpretable models to replicate and explain black-box decision-making, offering insights without directly modifying original model. For example, Ge et al. [11] utilized graph neural networks and visual concepts from ACE to develop the Visual Reasoning Explanation Framework (VRX). This framework facilitates human-to-network interaction through Knowledge Distillation (KD) [18], addressing the question, "Why is it?" in model decisions. Similarly, DiConStruct [27] employs a distillation-based framework that integrates an approximate black-box graph reasoning network as a surrogate model. This approach efficiently approximates black-box predictions while providing causal explanations [2, 41].

2.2. Model Intervention

Model intervention encompasses techniques that go beyond identifying biases or errors in black-box models, allowing for direct adjustments that improve both accuracy and interpretability. For example, Moayeri et al. [26] used attribution methods to identify spurious correlations in models trained on Hard-ImageNet and applied foreground constraints to refine model attention. Yet, attribution methods typically indicate only "where" the model focuses, without explaining "what" specific features are influential. As a result, explanations are often limited to the foreground of the object, preventing fine-grained attribution to critical features.

In some cases, biases stem from inherent object characteristics [36, 37]. For instance, models may confuse "horse" and "zebra" due to shape similarity, but emphasizing texture and color, like "black and white stripes", improves classification. Identifying biases, whether in foreground or background features, is crucial for implementing human-understandable interventions that refine model behavior.

Post-hoc methods allow limited intervention, while CBM-based approaches can be refined to the point of concept. CB2M [35] employs a memory buffer mechanism to automate and generalize CBM interventions, requiring user input on only a small number of samples. However, its intervention remains sample-level and still relies on prior user intervention. CCGM [6] structures the concept bottleneck using a causal concept graph, modeling causal relationships in concept-based reasoning. While it provides intervention methods to explain the model's causal reasoning, it is limited to single-modal natural language concepts. Marcinkevics et al. [24] used probe functions to map black-box hidden-layer features into a concept bottleneck, enabling interpretation through language-based concepts and associated scores. This approach combines concept editing with human prior knowledge to address spurious correlations, and thus correct black-box errors.

3. Methodology

The feature attribution or concept explanation provided by existing post-hoc methods is actually intended to allow people to understand how neural networks work. At the same

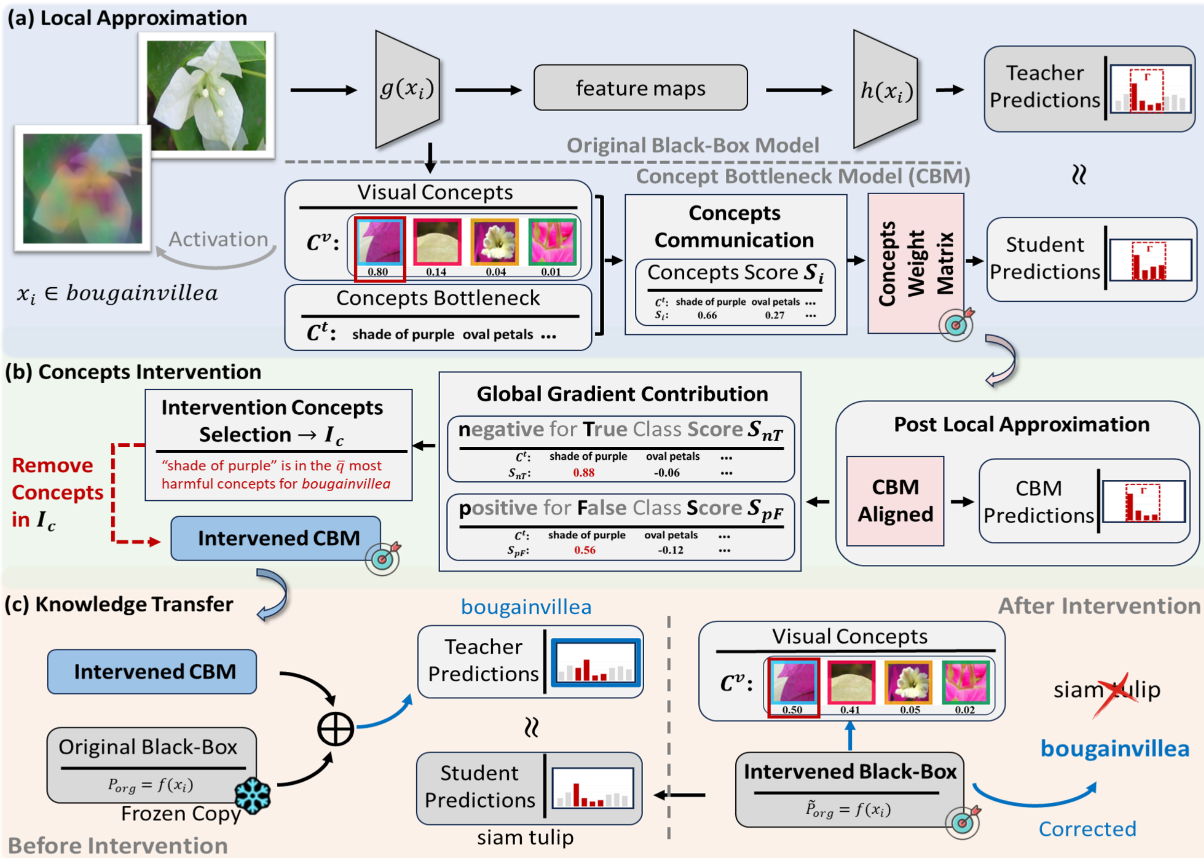

Figure 1. **The Intervention Process**. **(a) Local approximation**: The CBM inference structure first distills an interpretable basis, that approximates the original black-box model's class predictions on confused classes (Γ); **(b) Concept intervention**: Next, the classification bias is corrected by intervening with concepts that are detrimental to the decision-making and selected based on the concept score of S_{nT} and S_{pF} in the CBM obtained from Step **(a)**; **(c) Knowledge transfer**: Finally, the corrective adjustments made to the CBM in Step **(b)** are transferred back to the black-box model to improve reasoning and reduce biases. We refer to Figure 2 as an illustrative example to demonstrate intervention interpretation and predicted corrective effects.

time, making the neural network understand human knowledge and correct its errors is the intention of the human neural network mutual understanding framework.

In this section, we formally define CBM-HNMU framework, illustrated in **Figure 1**. We further explain how this framework facilitates mutual understanding between humans and neural networks, as well as how it enables the correction of erroneous predictions in black-box models.

3.1. Problem Setup

We first introduce the black-box classifier, denoted by the functional $f(x) = h(g(x))$. Here, the feature activation map is denoted by $g(x) : \mathbb{R}^{m\times m\times 3} \mapsto \mathbb{R}^p$, where p is the dimensionality of the feature representation and m is the standard size (number of pixels) of input image x. The function $g(x)$ characterizes the feature extraction process in a hidden layer of the black-box classifier, while $h(\cdot)$ serves as the classification head, mapping the extracted feature representation $g(x)$ to the final logits.

Additionally, we define the dataset used for training, validation, and inference as the set of doublet of every image and its corresponding class, i.e., $D_{\text{train}} := \{(x, y)\}^{N_{\text{train}}}, D_{\text{val}} := \{(x, y)\}^{N_{\text{val}}}, D_{\text{test}} := \{(x, y)\}^{N_{\text{test}}}$ where N_{train}, N_{val}, and N_{test} denote the number of sample images in the training, validation, and test sets, respectively. Each sample consists of an image x and its corresponding class label y_i, $\forall\ i = 1, 2, \ldots, N_{\text{class}}$, and N_{class} represents the total number of classes.

3.2. Confused Classes Selection

We are now ready to investigate whether interpretable structures can approximate the reasoning processes of black-box models using a constrained set of concepts. However, constructing an inference model with a fixed, limited concept bottleneck that accurately captures the black-box model's reasoning across all classes in a dataset remains a significant challenge. To this end, we begin by identifying specific classes within each dataset and model that are frequently misclassified as one another. We define confused samples as instances where the model erroneously predicts one similar class as another. Using the black-box classifier trained on D_{train}, we evaluate its performance on D_{val} and record the number of misclassified samples for each class pair.

Next, we rank these class pairs based on the frequency of misclassification and extract the pairs exhibiting the highest confusion. Finally, we identify the set of confused classes, denoted as $\Gamma \subseteq \{1, 2, \ldots, N_{\text{class}}\}$, whose corresponding cardinality is known as the number of confused classes, $|\Gamma| = N_\Gamma \geq 2$.* These classes are considered for intervention and corrective adjustments in the black-box model.

3.3. Concepts Communication

Once we identify the classes requiring intervention, we can begin constructing a human-neural network mutual understanding framework for the confused classes, Γ. It is essential to align their understanding to establish a communication bridge between humans and neural networks. Methods such as ACE, IBD, MOCE, and CRAFT, among others, provide extraction techniques that help us gain insights into the black-box model. Among them, CRAFT performs efficient unsupervised concept extraction based on non-negative matrix factorization (NMF) [13] and provides feature attribution while avoiding additional annotations. Thus, in our subsequent experiments, we adopt CRAFT as a black-box-driven visual concept extraction method, where each concept is represented as $C_k^v \in \mathbb{R}^{m \times m \times 3}$ for $k = 1, 2, \ldots, n$. Given an input image x_i from the training/testing/validation set, the corresponding visual concepts are extracted using CRAFT, resulting in

$$C^v(i) = \text{CRAFT}\Big(g(x_i)\Big), \tag{1}$$

so that $C^v : \{C_1^v, C_2^v, \ldots, C_n^v\}$, with n denotes the number of visual concepts used for mapping concept bottleneck.

The most intuitive and easily interpretable concepts for humans are expressed through natural language. By leveraging a concept bottleneck generated using ChatGPT-3.5-Turbo, we can efficiently construct a transparent reasoning framework. To design prompt templates for generating concept bottlenecks across all datasets in our experiments, we directly adopt the methodology introduced in LaBo [43]. Consequently, we obtain a concept bottleneck $C^t \in \mathbb{C}$ for each dataset. For example, in the Flower-102 dataset, the bottleneck is represented as $C^t = \{\text{shade of purple, ovalpetals}, \ldots\}$, with $|C^t| = N_c$ the total number of natural language concepts.

To efficiently establish connection between C^v and C^t, CBM-HNMU utilizes CLIP as an intermediary. The visual concepts extracted from the black-box model's intermediate layer and the natural language concepts in the concept bottleneck are mapped into a shared representation space $\mathbb{R}^{1 \times d}$ ($d = 512$). This transformation is performed using CLIP's image encoder $E_{\text{img}}(C_k^v) : \mathbb{R}^{m \times m \times 3} \mapsto \mathbb{R}^{1 \times d}$ for $k = 1, 2, \ldots, n$, and its text encoder $E_{\text{text}}(C^t) : \mathbb{C} \mapsto \mathbb{R}^{N_c \times d}$.

*Confused set must contain at least two elements for confusion to occur.

Subsequently, concept score $S_i \in \mathbb{R}^{1 \times N_c}$ is computed for each input image $x_i \in D_{\text{train,val,test}}$ by

$$\begin{aligned} S_i = \text{CS}(g(x_i), C^t) &= \frac{1}{n} \sum_{k=1}^{n} E_{\text{img}}(C_k^v(i)) \times E_{\text{text}}(C^t)^{\text{T}} \\ &= \frac{1}{n} \sum_{k=1}^{n} E_{\text{img}}(\text{CRAFT}(g(x_i))_k) \times E_{\text{text}}(C^t)^{\text{T}}, \end{aligned} \tag{2}$$

where $\times$ denotes the matrix multiplication and $\text{CS}(\cdot)$ the function representing the acquisition process of the concept score S_i. The number of visual concepts used to map the concept bottleneck is specified by n. The notation $\text{CRAFT}(g(x_i))_k = C_k^v(i)$ refers to the k-th visual concept extracted from the hidden layer of the black-box model. The computed concept score S_i provides a unified representation of the model's interpretation of an input image x_i in terms of human-understandable concepts. From the score distribution across natural language concepts, humans can gain insight into how the black-box model internally processes and understands different aspects of the image.

3.4. Local Approximation

As illustrated in **Figure 1.(a)**, we first obtain the concept score vector S_i, which is interpretable for both humans and black-box models. To construct the inference component of CBM, we use a concept weight matrix $W \in \mathbb{R}^{N_\Gamma \times N_c}$. This matrix is used to build a classification header similar to that of the original model.

For each input image x_i, the corresponding concept score vector S_i is passed through CBM, producing the classification output $P_{\text{cbm}} : \mathbb{R}^{m \times m \times 3} \mapsto \mathbb{R}^{1 \times N_\Gamma}$ as

$$P_{\text{cbm}}(x_i) = S_i \times W^{\text{T}} = \text{CS}(g(x_i), C^t) \times W^{\text{T}}. \tag{3}$$

Simultaneously, the original black-box model generates its own predictions $P_{\text{org}} : \mathbb{R}^{m \times m \times 3} \mapsto \mathbb{R}^{1 \times N_{\text{class}}}$ that reads

$$P_{\text{org}}(x_i) = f(x_i) = h(g(x_i)). \tag{4}$$

Next, we extract the original black-box predictions $P_{\text{org}}^\Gamma \in \mathbb{R}^{1 \times N_\Gamma}$ from confused classes Γ, and align them with CBM's output using the ℓ_2-norm. The local approximation loss function is then defined as

$$\text{Loss}_{\text{lp}} = \frac{1}{N_\Gamma} \| P_{\text{cbm}}(x_i) - P_{\text{org}}^\Gamma(x_i) \|_2, \tag{5}$$

where P_{org}^Γ represents the subset of the black-box predictions corresponding to the confused classes.

Our goal is to reduce the bias and complexity of model approximation by utilizing similar concept expressions and structures. This allows the CBM to take input aligned with the visual concepts extracted from the black-box's hidden layer, and make predictions through a simple linear layer that mimics the black-box behavior.

Lewis and Catlett [22], Esteva et al. [9], Yuan et al. [44], Guo and Greiner [15], among others, have provided valuable insights into misleading concept selection and modification. As shown in **Figure 1.(b)**, our intervention method is based on gradient contribution. The potential concept errors within locally confused classifications may fall into the following two categories

i) Concept C_j^t with **negative** impact on the correct (**True**) class predicted by CBM, that quantified by S_{nT}.

ii) Concept C_j^t with **positive** influence on CBM's incorrect (**False**) prediction of other classes, quantified by S_{pF}.

These two errors are represented by concept scores and quantified using contribution matrices $S_{\text{nT}}, S_{\text{pF}} \in \mathbb{R}^{N_\Gamma \times N_c}$, which capture the statistical impact of each concept on global model behavior for each confusion class.

To construct the concept contribution matrices S_{nT} and S_{pF}, consider a sample $x_i \in D$ with ground-truth label y_i. Let $y_i^* := \max_j [P_{\text{cbm}}(x_i)]_{1,j}$ denotes predicted class under the concept bottleneck model and $W \in \mathbb{R}^{N_\Gamma \times N_c}$ the concept weight matrix, where each row vector $w_k \in \mathbb{R}^{1\times N_c}$ corresponds to confusion class k.

To assess the influence of concepts on model behavior, we define a gradient-based attribution function

$$G(w_k, P_k(x_i)) := \frac{\partial P_k(x_i)}{\partial w_k} \odot w_k, \tag{6}$$

where $\odot$ denotes element-wise multiplication and $P_k(x_i) = [P_{\text{cbm}}(x_i)]_{1,k}$ is the predicted probability for class k.

Based on $S_i \in \mathbb{R}^{1\times N_c}$ obtained from Eq. (2), the per-sample contribution scores are defined as follows:

$$S_{\text{nT}}(i)_{y_i} := \begin{cases} -S_i \odot G(w_{y_i}, P_{y_i}(x_i)), & \text{if } y_i \in \Gamma \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{7a}$$

$$S_{\text{pF}}(i)_{y_i^*} := \begin{cases} S_i \odot G(w_{y_i^*}, P_{y_i^*}(x_i)), & \text{if } y_i^* \neq y_i \\ \mathbf{0}, & \text{otherwise} \end{cases}. \tag{7b}$$

Here, S_{nT} captures how the activated concepts suppress the true class, while S_{pF} measures how they reinforce the incorrect predicted class. To mitigate the impact of unfavorable concepts, we accumulate the sample-wise scores over the validation set to form global contribution matrices. Concepts in each confused class are ranked by their total influence in S_{nT} and S_{pF}, and the top $\bar{q}$ most detrimental concepts are selected for removal during intervention.

We summarize the concept intervention procedure in Algorithm 1, which identifies and prunes detrimental concepts based on their contribution to prediction errors.

After identifying the intervention index set I_c from D_{val}, we update the concept weight matrix $W \in \mathbb{R}^{N_\Gamma \times N_c}$ by zeroing out the selected concept weights for each confused class. Specifically, for each $k \in \Gamma$ and each concept $j \in I_c^k$, we set $W[k,j] = 0$. The resulting $\overline{W}$ defines the intervened concept weight matrix of the CBM.

Algorithm 1 Gradient-Based Concept Intervention

Input: Validation set D_{val}, concept weight matrix W, encoders E_{img}, E_{text}

Initialize contribution matrices $S_{\text{nT}} \leftarrow 0$, $S_{\text{pF}} \leftarrow 0$
for $i = 1$ to N_{val} **do**
  Compute S_i, $P_{\text{cbm}}(x_i)$, and predicted class $y_i^\star$
  Update $S_{\text{nT}}(i)_{y_i}$ and $S_{\text{pF}}(i)_{y_i^\star}$ using Eqs. (7a) and (7b), then accumulate S_{nT} and S_{pF}
end for
for each confused class $k \in \Gamma$ **do**
  Select top $q/2$ concepts $I_{\text{nT}}^k > 0$ from S_{nT}^k.
  Select top $q/2$ concepts $I_{\text{pF}}^k > 0$ from S_{pF}^k.
  Merge selected concepts as $I_c^k = \{I_{\text{nT}}^k\} \bigcup \{I_{\text{pF}}^k\}$.
end for
return Intervention indices $I_c \in \mathbb{R}^{N_\Gamma \times \bar{q}}$ with $\bar{q} \leq q$

3.5. Knowledge Transfer

After the intervened CBM ($\overline{W}$) is updated, the next step is to transfer the refined knowledge back into the original black-box model. As illustrated in **Figure 1.(c)**, since CBM provides a localized approximation of the black-box model, we retain the learned knowledge for all classes except those involved in confusion. To achieve this, we construct a teacher $P_t \in \mathbb{R}^{1\times N_{\text{class}}}$ by combining two sources of predictions

i) Frozen copy of the original black-box predictions for the unaffected classes denoted as $P_{\text{org}}^{\Gamma^{\complement}} \in \mathbb{R}^{1\times N_{\Gamma^{\complement}}}$, where $\Gamma^{\complement} := \{1, \ldots, N_{\text{class}}\} \setminus \Gamma$ and $N_{\Gamma^{\complement}} = N_{\text{class}} - N_\Gamma$.

ii) Post-intervention CBM predictions $P_{\text{cbm}}^\star \in \mathbb{R}^{1\times N_\Gamma}$, that refine model's reasoning for the confused classes.

The combined output P_t serves as a supervisory signal for reverse distillation, guiding the original black-box model $P_s \in \mathbb{R}^{1\times N_{\text{class}}}$ to integrate the improvements introduced by CBM while preserving its original predictions for non-confused classes. Given the distillation temperature T_1 of $P_{\text{cbm}}^\star$, the teacher's prediction reads

$$\begin{aligned} P_t &= \text{combine}(P_{\text{cbm}}^\star, P_{\text{org}}^{\Gamma^{\complement}}) \\ &= \text{pr} \cdot \text{softmax}(P_{\text{cbm}}^\star / T_1) \oplus \text{softmax}(P_{\text{org}})^{\Gamma^{\complement}}, \end{aligned} \tag{8}$$

where $\text{softmax}(P_{\text{org}})^{\Gamma^{\complement}}$ and $P_{\text{cbm}}^\star$ represents the predicted probability of the confusion classes by the original black-box model with frozen parameters, and the logits output of the post-intervention CBM, receptively. The operator $\oplus$ denotes the concatenation of output elements.

To ensure that the model's predictions for non-intervention classes remain unchanged, we introduce the

Models	Flower-102		CIFAR-10		CIFAR-100		CUB-200		FGVC-Aircraft		AVG IMP
	w/o INT	$w/$ INT	w/o INT	$w/$ INT	w/o INT	$w/$ INT	w/o INT	$w/$ INT	w/o INT	$w/$ INT	
NFResNet50	94.28	**95.36** $\pm$0.019	80.92	**83.56** $\pm$0.005	73.58	**73.92** $\pm$0.037	62.41	**62.67** $\pm$0.057	64.63	**64.94** $\pm$0.028	$\uparrow$ **0.93**
Vit_Small	97.24	**98.20** $\pm$0.019	91.51	**92.84** $\pm$0.154	81.26	**82.43** $\pm$0.031	74.75	**75.39** $\pm$0.029	69.88	**70.94** $\pm$0.251	$\uparrow$ **1.03**
GCVit	93.58	**95.20** $\pm$0.050	80.20	**80.97** $\pm$0.054	72.38	**72.55** $\pm$0.034	76.99	**77.64** $\pm$0.070	69.81	**71.41** $\pm$0.171	$\uparrow$ **0.96**

Table 1. Comparison of global classification accuracy without (w/o) or with ($w/$) intervention (INT). All experimental models using CBM-HNMU for black-box intervention have objective generalization performance improvements on the D_{test} compared with the baseline. Results are presented as average $\pm$ standard deviation (AVG $\pm$ STD) over three random tests. The average improvement (AVG IMP) represents the mean accuracy increase for each row.

probability residual coefficient pr, defined as

$$\text{pr} = 1 - |\text{softmax}(P_{\text{org}})^{\Gamma^{\complement}}|_1, \tag{9}$$

whereby the residual probabilities of the intervention classes from the original black-box model onto the post-intervention CBM are redistributed.

The student's prediction P_s is given by

$$P_s = \text{softmax}(\tilde{P}_{\text{org}}/T_2), \tag{10}$$

where $\tilde{P}_{\text{org}}$ represents the black-box predictions for the subset of parameters requiring intervention, and T_2 is the distillation temperature for the original black-box model.

Finally, we perform black-box intervention by minimizing the cross-entropy loss between P_t and P_s, i.e.,

$$\text{Loss}_{\text{kt}} = \sum_{k=1}^{N_{\text{class}}} P_t^k \cdot \log(P_s^k), \tag{11}$$

over the validation dataset D_{val}. This ensures that refined knowledge from CBM is effectively transferred back to the black-box model.

4. Experiments

4.1. Experiments Setups

Dataset. CBM-HNMU is evaluated on Flower-102 [28], CIFAR-10, CIFAR-100 [21], FGVC-Aircraft [23], and CUB-200 [40]. CIFAR-10 and CIFAR-100 represent object datasets of varying scales with high discriminative power, while Flower-102, CUB-200, and FGVC-Aircraft contain objects with fewer distinguishing features, necessitating more fine-grained classification.

Baselines. We evaluate CBM-HNMU across classic and commonly used CNN and Transformer architectures, including NFResNet50 [25], ResNeXt26 [42], BotNet26 [34], RexNet100 [16], ViT-Small [7], GCVit [17], CaiT-Small [39], ConVit-Base [5], and DeiT-Base [38].

Evaluation Protocol. All baselines are initialized with ImageNet-1K [30] pre-trained weights and trained for 50 epochs on different D_{train}. Confusion class selection, local approximation, concept intervention, and intervention knowledge transfer are performed on validation set D_{val}, while evaluations before and after black-box intervention are conducted on testing set D_{test}. The local approximation step is with learning rate of $1e^{-4}$ and runs for 200 epochs, and intervention knowledge transfer is trained with a learning rate of $3e^{-7}$ for 10 epochs.

4.2. Experiments Results

We evaluate the improvements in both global accuracy and fine-grained class prediction, resulting from the CBM-HNMU intervention across multiple datasets. In addition, we examine how the visual concepts captured by black-box model evolve before and after the intervention, using visualization techniques and natural language concepts to interpret the process. An ablation study further illustrates how the quality of black-box approximation and the overall intervention effect vary with the number of intervention concepts. Furthermore, we examine the impact of CBM-based classification methods (CBM and Post-hoc CBM) and the influence of the number of confused classes (N_Γ) on intervention effectiveness. Finally, we conducted a user study to confirm that the chosen intervention concepts (i) align with human visual perception and (ii) help correct classification errors.

Performance Improvements. As shown in **Table 1**, the three architectures in our experiments achieved global accuracy improvements across five datasets after intervention with CBM-HNMU. To further verify the effect of CBM-HNMU, we conduct the identical comparative test using another three architectures on Flower-102, FGVC-Aircraft, and CUB-200. The results are presented in **Table 2**. As a result, CBM-HNMU achieved the highest accuracy

Models	Flower-102		CUB-200		FGVC-Aircraft		AVG IMP
	w/o INT	$w/$ INT	w/o INT	$w/$ INT	w/o INT	$w/$ INT	
ResNeXt26	88.47	**90.85** $\pm$0.363	59.32	**59.60** $\pm$0.113	64.27	**64.72** $\pm$0.212	$\uparrow$ **1.04**
BotNet26	89.20	**91.30** $\pm$0.050	56.89	**57.09** $\pm$0.066	53.70	**54.51** $\pm$0.079	$\uparrow$ **1.04**
RexNet100	90.26	**90.77** $\pm$0.082	55.40	**57.39** $\pm$0.024	52.60	**52.98** $\pm$0.079	$\uparrow$ **0.96**

Table 2. Global classification accuracy with (w/) and without (w/o) intervention (INT) of another three architectures. Results are reported as the average $\pm$ standard deviation (AVG $\pm$ STD) over three random tests, and AVG IMP as the average improvement for each row.

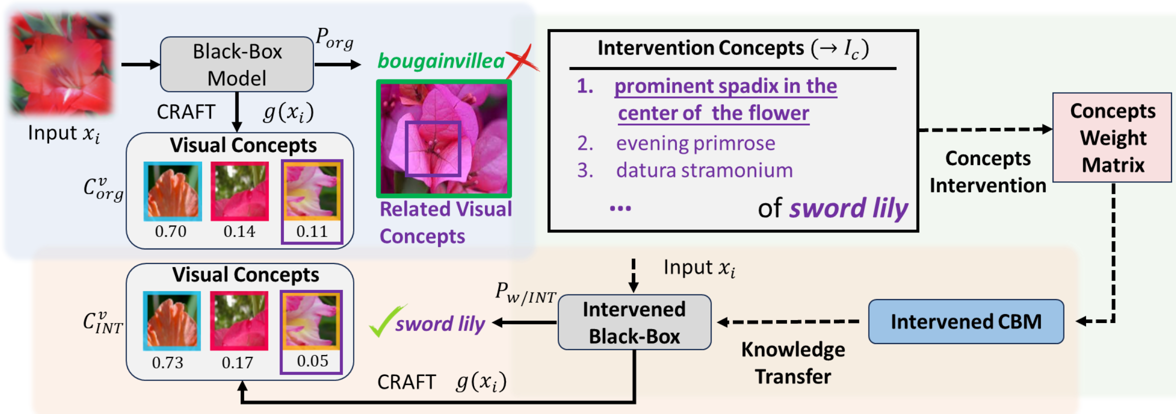

Figure 2. **Comparative visualization of intervention concepts and black-box attributions before and after intervention.** It shows the visual results of NFResNet50 on Flower-102. Through **concept intervention**, CBM-HNMU identify and remove $\overline{q}$ harmful concept (I_c) from confusing class (**"sword lily"**) based on S_{nT} and S_{pF}. After intervention knowledge is transferred back through CBM, the black-box model corrects the misclassified **"bougainvillea"**. Analyzing the visual concept changes in the black-box model's middle layer before and after intervention, we observe a decrease in the "total Sobol indices" of the Related Visual Concept, which is corresponding to **"prominent spadix in the center of the flower"**. Compared with the bougainvillea sample marked by green border, we find that the flower center features of both classes are similar, which likely contributed to the misclassification. The intervention reduces the model's reliance on such subtle distinctions, leading to more accurate predictions. For additional visual explanations, see **Appendix. H**.

improvement with NFResNet50 on CIFAR-10, increasing from 80.92% to 83.56%, a gain of 2.64%. Additionally, Vit_Small achieved the highest average improvement (AVG IMP) across the five datasets, with an increase of 1.03%.

Table 3 shows the number of corrected samples and the coverage of confusion classes in different intervention models on D_{test} of Flower-102. The notation "$-n$" indicates that n samples originally misclassified into a confusion class are corrected after intervention. Conversely, "$+n$" denotes n samples initially misclassified into any class that are reassigned to the confusion class after intervention. Most corrected samples in D_{test} are related to the intervened classes, either because they were previously predicted as one of the confusion classes or because their true labels belong to the confusion set.

Models	Flower-102 (D_{test})		
	Confusion Class (Γ)	Corrected	Coverage
NFResNet50	bougainvillea (-6,$+0$), camellia (-1,$+2$), mallow (-), gaura (-0,$+2$), cyclamen (-), sweet pea (-9,$+0$), sword lily (-0,$+5$)	34	25
BotNet26	bougainvillea (-0,$+3$), camellia (-22,$+0$), hibiscus (-12,$+0$), mallow (-0,$+1$), petunia (-0,$+1$)	59	39
GCVit	petunia (-18,$+0$), camellia (-6,$+0$), mallow (-), hibiscus (-4,$+0$), morning glory (-)	51	28

Table 3. Statistics of fine-grained class interventions in Flower-102. "Corrected" refers to the total number of error samples corrected by each model on D_{test}, and "Coverage" represents the number of corrected samples associated with the confused classes.

Intervention Visualization. Herein, We analyze how the intervened black box's performance improvement relates to changes in natural language and visual-related concepts. **Figure 2** presents comparative results illustrating the intervention process on Flower-102 dataset, highlighting visual concept explanations before and after intervention. Samples are randomly selected from a subset, where the black-box model's original classification errors on the test set are corrected following intervention. Each corrected sample includes at least one pre- or post-correction class related to the confused classes, ensuring a clear visual link to the intervention concept (refer to "Coverage" in **Table 3**).

We observe that intervened natural language concepts often shift black-box visual concept explanations. In the visualization, the black-box model misclassifies sword lily as bougainvillea, relying on feature "prominent spadix in the center of the flower," that shared by both classes. The corresponding concept score drops from 0.11 to 0.05 after intervention, indicating reduced reliance on this feature. CBM-HNMU refocuses the model from the spadix to petal-related features, demonstrating that it does not merely transfer knowledge, but actively corrects black-box errors by aligning human and neural network reasoning.

Ablation Study. In **Figure 3**, after ablating the number of confused classes across the three datasets, when the proportion of intervention classes is less than 25%, local approximation bias and the intervention improvement are relatively stable. When all classes are intervened, approximation bias and intervention improvement deteriorate significantly.

To further validate the effectiveness of our intervention concept selection method, we conducted a random concept intervention for reverse knowledge transfer, applying CBM-HNMU on Flower-102, CUB-200, and FGVC-Aircraft using NFResNet50 and ViT. The results in **Table 4**, show that randomized intervention (often used as a baseline for CBM

Models	Flower-102		CUB-200		FGVC-Aircraft	
	w/ OURS	w/ RAND	w/ OURS	w/ RAND	w/ OURS	w/ RAND
NFResNet50	95.36	94.72 ($\downarrow$ 0.64)	62.67	62.51 ($\downarrow$ 0.16)	64.94	64.75 ($\downarrow$ 0.19)
Vit_Small	98.20	97.81 ($\downarrow$ 0.39)	75.39	75.23 ($\downarrow$ 0.16)	70.94	69.97 ($\downarrow$ 0.97)

Table 4. Comparison of global classification accuracy using CBM-HNMU with intervention method (w/OURs) and randomized concept intervention (w/RAND). Black-box model optimized by CBM-HNMU outperforms random intervention under the same parameter settings.

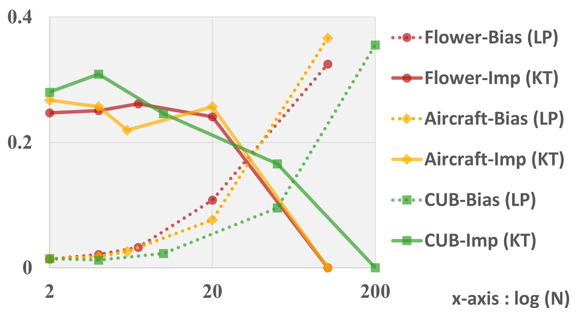

Figure 3. The dotted line represents the normalized category prediction bias of the local approximate (LP) CBM and the original NFResNet-50 model across three datasets (D_{test}) under different numbers of intervention classes. The solid line shows the normalized accuracy improvement (Imp) after intervention knowledge transfer (KT) back to the original model.

intervention methods) performs poorly. Our findings align with this expectation, showing that CBM-HNMU achieves an average accuracy improvement of 0.42 over random intervention, with the highest improvement of 0.97.

Performance Comparison with CBM-Related Methods. We evaluate the performance of the black-box model after CBM-HNMU intervention using NFResNet-50 and compare it with CBM (same backbone and concept bottleneck) and Post-hoc CBM across three datasets. As shown in **Table 5**, CBM-HNMU achieves best classification performance. Additionally, we observe that Post-hoc CBM improves performance over CBM by incorporating the backbone's residual expression. However, it still incurs accuracy loss compared to the corresponding black box model.

Models \ Dataset	Flower-102	CUB-200	FGVC-Aircraft
Black-Box	94.28	62.41	64.63
CBM	92.57	54.53	56.31
PCBM	93.58	58.20	64.45
CBM-HNMU	**95.36**	**62.67**	**64.94**

Table 5. Models optimized with CBM-HNMU intervention show significant improvements over corresponding baseline, CBM, and Post-hoc CBM across three datasets (NFResNet50 as backbone).

User Study. We tested whether the concepts suppressed by CBM-HNMU are the same cues that humans find misleading. The study spanned the *Flower-102*, *CUB-200*, and *FGVC-Aircraft* test sets. For each pair of confused class we clustered its images into ten groups and retained one exemplar per cluster. Thirty paid, non-expert volunteers examined each exemplar along with its intervention concept and scored the concept as (i) true-class cue (CT), (ii) false-class cue (CF), or (iii) a semantically related cue shared by both classes (CR). As shown in Fig. 4, the mean confidence for all three score exceeds 0.5, and the CT and CF scores are nearly identical. This parity, together with the high CR confidence, indicates that CBM-HNMU targets exactly the visual hints that mislead both the network and human observers, validating the choice of intervention concepts.

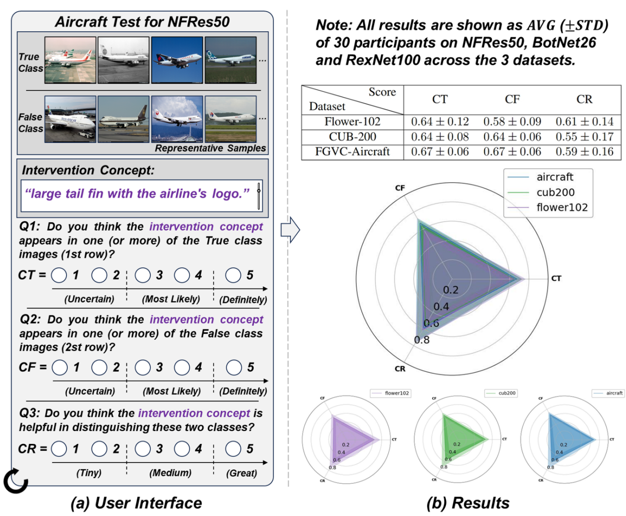

Figure 4. The interface was polished, but prompts and evaluation criteria remained unchanged. Scores were normalized, with values above 0.5 indicating consistency with the evaluation rule. The results show that our intervention concepts align with human perception and aid in classifying confused classes.

5. Conclusion

We propose a novel human neural network mutual understanding framework that combines visual concepts and natural language concept bottleneck. Based on CBM-HNMU, we can intervene in the black box based on understanding the black box prediction basis to improve the generalization performance of the black box while maintaining interpretability. By combining LLMs with pre-trained VLM and unsupervised concept extraction method, CBM-HNMU can complete automated black-box intervention beyond the sample level without additional labor costs and provide enriched intervention explanations.

Acknowledgments

This work was supported by grants from the Natural Science Foundation of Shanxi Province (2024JCJCQN-66), Science and Technology Commission of Shanghai Municipality (NO.24511106900), Key RD Program of Zhejiang (2024SSYS0091) and is partially supported by the National Natural Science Foundation of China under grant Nos. 62072358 and 62072352.

References

[1] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose Maria Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion*, 99:101805, 2023. 1

[2] Sander Beckers. Causal Explanations and XAI. In *CLEaR*, 2022. 2

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv: 2005.14165*, 2020. 2

[4] Kushal Chauhan, Rishabh Tiwari, Jana von Freyberg, Pradeep Shenoy, and Krishnamurthy Dj Dvijotham. Interactive Concept Bottleneck Models. In *AAAI Conference on Artificial Intelligence*, 2022.

[5] Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. *arXiv: 2103.10697*, 2021. 6

[6] Gabriele Dominici, Pietro Barbiero, Mateo Espinosa Zarlenga, Alberto Termine, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. Causal Concept Graph Models: Beyond Causal Opacity in Deep Learning. *arXiv: 2405.16507*, 2024. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv: 2010.11929*, 2020. 6

[8] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial Robustness as a Prior for Learned Representations. *arXiv: 1906.00945*, 2019. 1

[9] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin M. Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017. 5

[10] Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. CRAFT: Concept Recursive Activation FacTorization for Explainability. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2711–2721, 2022. 1

[11] Yun Ping Ge, Yao Xiao, Zhi Xu, Meng Zheng, Srikrishna Karanam, Terrence Chen, Laurent Itti, and Ziyan Wu. A Peek Into the Reasoning of Neural Networks: Interpreting with Structural Visual Concepts. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2195–2204, 2021. 2

[12] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards Automatic Concept-based Explanations. In *Neural Information Processing Systems*, 2019. 1

[13] D. Green and Stephen Bailey. Algorithms for Non-Negative Matrix Factorization on Noisy Data With Negative Values. *IEEE Transactions on Signal Processing*, 72:5187–5197, 2023. 4

[14] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)*, 51:1 – 42, 2018. 1

[15] Yuhong Guo and Russell Greiner. Optimistic Active-Learning Using Mutual Information. In *International Joint Conference on Artificial Intelligence*, 2007. 5

[16] Dongyoon Han, Sangdoo Yun, Byeongho Heo, and Young Joon Yoo. Rexnet: Diminishing Representational Bottleneck on Convolutional Neural Network. *arXiv: 2007.00992*, 2020. 6

[17] Ali Hatamizadeh, Hongxu Yin, Jan Kautz, and Pavlo Molchanov. Global Context Vision Transformers. In *International Conference on Machine Learning*, 2022. 6

[18] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the Knowledge in a Neural Network. *arXiv: 1503.02531*, 2015. 2

[19] Seonggyeom Kim and Dong-Kyu Chae. What Does a Model Really Look at?: Extracting Model-Oriented Concepts for Explaining Deep Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:4612–4624, 2024. 1

[20] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. *arXiv: 2007.04612*, 2020. 1

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[22] David D. Lewis and Jason Catlett. Heterogeneous Uncertainty Sampling for Supervised Learning. In *International Conference on Machine Learning*, 1994. 5

[23] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-Grained Visual Classification of Aircraft. *arXiv: 1306.5151*, 2013. 6

[24] Ricards Marcinkevics, Sonia Laguna, Moritz Vandenhirtz, and Julia E. Vogt. Beyond concept bottleneck models: How to make black boxes intervenable? *arXiv: 2401.13544*, 2024. 2

[25] Tanish Mittal, Preyansh Agrawal, Esha Pahwa, and Aarya Makwana. NFResNet: Multi-scale and U-shaped networks for deblurring. *arXiv: 2212.05909*, 2022. 6

[26] Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard ImageNet: Segmentations for Objects with Strong Spurious Cues. In *Neural Information Processing Systems*, 2022. 2

[27] Ricardo Moreira, Jacopo Bono, Mário Cardoso, Pedro Saleiro, M'ario A. T. Figueiredo, and P. Bizarro. DiConStruct: Causal Concept-based Explanations through Black-Box Distillation. In *CLEaR*, 2024. 2

[28] Maria-Elena Nilsback and Andrew Zisserman. Automated Flower Classification over a Large Number of Classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 6

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 2021. 2

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 6

[31] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128:336 – 359, 2016. 1

[32] Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A Closer Look at the Intervention Procedure of Concept Bottleneck Models. *arXiv: 2302.14260*, 2023.

[33] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR*, abs/1312.6034, 2013. 1

[34] A. Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, P. Abbeel, and Ashish Vaswani. Bottleneck Transformers for Visual Recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16514–16524, 2021. 6

[35] David Steinmann, Wolfgang Stammer, Felix Friedrich, and Kristian Kersting. Learning to Intervene on Concept Bottlenecks. *arXiv: 2308.13453*, 2023. 2

[36] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv: 1312.6199*, 2013. 2

[37] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. *CVPR 2011*, pages 1521–1528, 2011. 2

[38] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2020. 6

[39] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with Image Transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 32–42, 2021. 6

[40] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *authors.library.caltech.edu*, 2011. 6

[41] Greta Warren, Mark T. Keane, and Ruth M. J. Byrne. Features of Explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI. *arXiv: 2204.10152*, 2022. 2

[42] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2016. 6

[43] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19187–19197, 2022. 4

[44] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple Instance Active Learning for Object Detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5326–5335, 2021. 5

[45] Mert Yuksekgonul, Maggie Wang, and James Y. Zou. Post-hoc Concept Bottleneck Models. *arXiv: 2205.15480*, 2022. 1

[46] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable Basis Decomposition for Visual Explanation. In *European Conference on Computer Vision*, 2018. 1

[47] Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating Modality Prior-Induced Hallucinations in Multimodal Large Language Models via Deciphering Attention Causality. In *The Thirteenth International Conference on Learning Representations*, 2025. 2