applied sciences

MDPI

Article

Advancing Model Explainability: Visual Concept Knowledge Distillation for Concept Bottleneck Models

Ju-Hwan Lee [1,*], Dang Thanh Vu [2], Nam-Kyung Lee [3], Il-Hong Shin [3] and Jin-Young Kim [1,*]

[1] Department of Intelligent Electronics and Computer Engineering, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 61186, Republic of Korea
[2] Research Center, AISeed Inc., 77 Yongbong-ro, Buk-gu, Gwangju 61186, Republic of Korea; dtvu1707@gmail.com
[3] Electronics and Telecommunications Research Institute, Media Intelligence Research Section, 218 Gajeong-ro, Yuseong-gu, Daejeon 34129, Republic of Korea; nklee@etri.re.kr (N.-K.L.); ssi@etri.re.kr (I.-H.S.)
* Correspondence: juhwanlee@jnu.ac.kr (J.H.L.); beyondi@jnu.ac.kr (J.-Y.K.)

Abstract: This study explores the integration of concept bottleneck models (CBMs) with knowledge distillation (KD) while preserving the locality characteristics of the CBM. Although KD proves effective in model compression, compressed models often lack interpretability in their decision-making process. We enhance comprehensive explainability by maintaining CBMs' inherent interpretability through our novel approach to knowledge distillation. We introduce visual concept knowledge distillation (VICO-KD), which transfers both explicit and implicit visual concepts from the teacher to the student model while preserving the local interpretability of the CBM, enabling accurate classification and clear visualization of evidence. VICO-KD demonstrates superior performance on benchmark datasets compared to Vanilla-KD, ensuring the student model learns visual concepts while maintaining the local interpretation capabilities of the teacher CBM. Our methodology shows competitive performance against existing concept models, and the student model, trained via VICO-KD, exhibits enhanced performance compared to the teacher during interventions. This study highlights the effectiveness of combining a CBM with KD to improve both interpretability and explainability in compressed models while maintaining locality properties.

Keywords: concept bottleneck models; knowledge distillation; explainable AI; interpretability

check for updates

Academic Editors: Raheleh Jafari, Alexander Gegov and Farzad Arabikhan

Received: 2 December 2024
Revised: 23 December 2024
Accepted: 26 December 2024
Published: 7 January 2025

Citation: Lee, J.-H.; Vu, D.T.; Lee, N.-K.; Shin, I.-H.; Kim, J.-Y. Advancing Model Explainability: Visual Concept Knowledge Distillation for Concept Bottleneck Models. *Appl. Sci.* **2025**, *15*, 493. https://doi.org/10.3390/app15020493

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

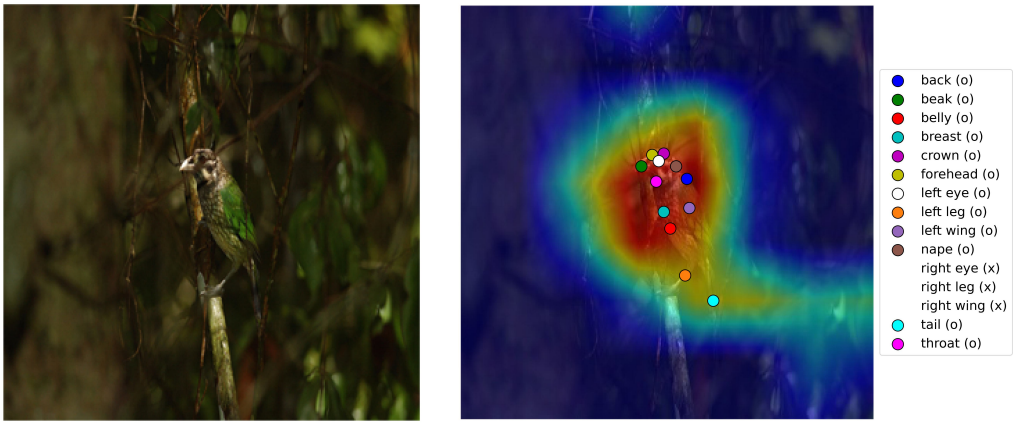
1. Introduction

Mission-critical systems, such as those in healthcare, autonomous driving, and defense, often demand reliable predictions. These systems also face hardware limitations, necessitating lightweight models rather than cumbersome ones. Knowledge distillation (KD) has emerged as a promising method for crafting lightweight models. By transferring knowledge from a cumbersome model (teacher) to a smaller, simpler model (student), KD empowers the student model to attain comparable performance with reduced computational complexity [1]. However, model compression using KD poses challenges in comprehending the network's decision-making process in relation to the teacher model [2].

While post hoc methods are commonly employed to grasp the decision-making process, they occasionally yield unintended explanations [3]. Despite the increasing volume of research literature on applying compressed models, discussions on the validity of these models in terms of explainability are rare.

In contrast to post hoc methods, an ante hoc approach exists that learns using human-understandable domain knowledge. This approach mitigates misinterpretation and distortion of explanations by integrating human-understandable domain knowledge during learning. A notable interpretable model among ante hoc approaches is the concept bottleneck model (CBM) [4], which predicts human-understandable concepts upon receiving input and subsequently predicts the final task based on these inferred concepts. CBMs can be integrated with any end-to-end model, offering inherent interpretability that is absent in standard CNN models. The primary advantage of a CBM lies in its ability to provide intuitive interpretation through learning human-level concepts.

In this study, we emphasize the inherent interpretability of CBM and integrate them with knowledge KD to address the limitations of post hoc interpretability methods, aiming for explainability, as illustrated in Figure 1. Specifically, we focus on image classification tasks to demonstrate the effectiveness of our approach. To our knowledge, it has not yet been demonstrated that KD can be used to train CBMs. Therefore, we investigate the effectiveness of KD in training three types of bottleneck models, as proposed by [4], when combined with KD. We introduce the representation of implicit knowledge through visual concept knowledge distillation (VICO-KD) to transfer the teacher's visual concept knowledge to the student while preserving the locality characteristics of the CBM. Our proposed VICO-KD applies to image classification models, leveraging explicit concepts as is traditional for CBMs and providing correct implicit concepts to users. This results in a transparent model that can accurately classify images while highlighting the visual evidence underlying its decision, presenting inferences in a user-friendly manner. This transparency is beneficial for understanding the model's learning process, diagnosing errors, and building trust in the model's predictions.



(**a**) Original image (**b**) Explanations to users

Figure 1. Enhancing interpretability and explainability through visual concept knowledge distillation in CBMs.

We demonstrate improvements of 10.65%, 2.9%, and 0.3% on the CUB-200-2011 (CUB) [5], Animals with Attributes2 (AwA2) [6], and Large-scale CelebFaces Attributes (CelebA) [7] datasets, respectively, compared to Vanilla-KD, using benchmark datasets with concept annotations. When visualizing and comparing CBMs that have used Vanilla-KD for learning and saliency maps, we confirm that they learn visual concepts similar to the teacher CBM. This proves our method can teach simple, explicit, and complex visual concepts. We also validate that our approach is competitive with existing concept models. Finally, we perform an intervention, a critical function of CBMs, on the student trained by VICO-KD and verify that it performs even better than the teacher.

The main contributions of this study are as follows:

1. We conduct a comprehensive empirical study investigating the effectiveness of knowledge distillation across three distinct bottleneck architectures on benchmark datasets.
2. We propose visual concept knowledge distillation (VICO-KD), a novel framework that facilitates the transfer of visual concepts while preserving CBM's locality characteristics, thereby enhancing model explainablity and interpretability during compression.
3. Through extensive experiments on three benchmark datasets, we demonstrate that VICO-KD not only achieves superior performance compared to concept model approaches but also maintains robust interpretability.

This paper is organized as follows: In Section 2, recent studies on CBMs and KD are reviewed, confirming the absence of previous research on CBMs with KD. Section 3 details the VICO-KD approach. Section 4 presents qualitative and quantitative evaluations of knowledge distillation exploration and VICO-KD performance. Section 5 addresses the limitations of this study, while Section 6 provides a summary and conclusion.

2. Related Works

2.1. Concept Bottleneck Models (CBMs)

In tasks such as facial recognition [8] and animal identification [6], CBMs initially deviated from the conventional end-to-end network structure. These networks generally performed less well than standard networks, revealing a trade-off between accuracy and interpretability within concepts. However, the emergence of CBMs [4] introduced an intervenable approach comparable to standard networks based on end-to-end structures. CBMs were designed with three types of bottleneck structures: independent, sequential, and joint bottlenecks. These designs enable interventions that are not feasible in standard networks, facilitating updates to standard end-to-end networks' predictive performance.

In a critical examination of CBMs, Margeloiu, Andrei, et al. [9] questioned whether a CBM truly learns and operates as intended. They thoroughly analyzed the three structures of CBM using post hoc methods, assessing the concepts' relevance and impact. Shin et al. [10] presented the first case study that thoroughly analyzed interventions in CBMs. They introduced various methods for selecting intervention concepts and the effects of interventions, performing a deep analysis of how these evolve across different scenarios. Ref. [11] focused on the locality of concept-based learning within CBMs, pointing out the challenges of localized learning.

Alternative modifications of CBMs have been proposed [12–17]. Notably, Yuksekgonul et al. [12] suggested a method to convert any pretrained model into a data-efficient CBM to address performance issues arising from limited concept-annotated training datasets. Additionally, Kim et al. [13] introduce ProbCBM, an interpretable network that utilizes probabilistic embeddings to provide explanations based on concepts and their corresponding uncertainties. In contrast to these new approaches to modifying CBMs, Pittino et al. [18] implemented these modifications in actual applications.

As shown above, previous studies have mainly focused on analyzing the operational aspects of CBMs and providing solutions to the need for concept annotations. However, our study takes a different approach by viewing CBMs from a model compression perspective, aiming to leverage KD as a practical methodology. Similar attempts have been made by Magister et al. [19], who investigated concept distillation in graph neural networks (GNNs), and Sousa et al. [20], who proposed a model-agnostic distillation technique. However, these KD approaches were limited to graph-structured data and tabular data. Therefore, we propose VICO-KD, which targets CNN-based networks, as the most commonly used architecture, achieving both model efficiency and interpretability in visual tasks such as image classification.

2.2. Knowledge Distillation (KD)

The KD method proposed by Gutstein et al. [21] was generalized by Hinton et al. [1]. This involves the use of softened labels, which is the most common approach to leveraging the knowledge of logits (i.e., logit-based KD) [22–27]. However, the majority of studies on logit-based knowledge distillation focus more on proposing effective regularization and optimization methods rather than on new approaches. Zhang et al. [22] suggest a mutual learning strategy, where the student and teacher are trained simultaneously. Furlanello et al. [23] propose a self-distillation method, where learning occurs within neural networks of the same size. Mirzadeh et al. [26] introduce a method that utilizes an intermediary-sized assistant teacher to bridge the gap between the teacher and student networks. Nonetheless, logit-based knowledge can be challenging to generalize in certain datasets [25] (e.g., ImageNet [28]), and issues have been raised regarding the difficulty for the student to learn efficiently when there is a significant capacity difference between the teacher and student networks [29].

Alternatively, a method exists that directly transfers the representations of intermediate layers, known as feature-based KD, to the student [30–38]. This approach aims to guide the student's learning process by leveraging the outputs of the teacher's hidden layers. Initially proposed by Romero et al. [30], this method entails directing the student's learning based on the outputs of the teacher's hidden layers. Essentially, the student network learns to predict feature representations based on the intermediate representations of the teacher network. Although this approach demands more computational resources than the traditional logit-based KD, it has demonstrated promising results and generalization capabilities. Among the various methods of transferring feature knowledge from intermediate layers in KD, utilizing the attention mechanism [39,40] has proven highly effective. However, these attention-based feature distillation methods lack a clear explanation of the role of attention.

As discussed above, KD has undergone extensive study in the context of CNNs, primarily focusing on transferring knowledge at both the logit and feature levels. However, the application of KD techniques to CBMs has not been explored. Considering the promising results of KD in CNNs, investigating the potential of these methods in the context of CBMs could yield substantial improvements in their performance and explainability. Hence, this study aims to comprehensively explore training CBMs through KD, specifically proposing a method for transferring visual concept knowledge to enhance the interpretability of these models.

3. Methodology

3.1. Preliminaries

3.1.1. Concept Bottleneck Models

Concept bottleneck models (CBMs) are structured as $f(g(x))$, where for an input $x \in \mathbb{R}^d$, the model predicts intermediate concepts $c \in \mathbb{R}^k$, and based on these predicted concepts it predicts the final target $y \in \mathbb{R}$. Specifically, $g : \mathbb{R}^d \to \mathbb{R}^k$ is a function that maps the input to a concept space, and $f : \mathbb{R}^k \to \mathbb{R}$ is a function that maps the concepts to the final prediction.

During the training process of CBMs, the dataset $\{x^{(i)}, y^{(i)}, c^{(i)}\}_{i=1}^{n}$ is utilized, comprising input data x, corresponding concepts c, and final targets y. The objective of the model is to minimize two loss functions. The first loss function, $L_{C_j} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}+$, quantifies the disparity between the predicted and actual values for the j-th concept. In contrast, the second loss function, $L_Y : \mathbb{R} \times \mathbb{R} \to \mathbb{R}+$, assesses the discrepancy between the final prediction and the ground truth. Three main types of CBM are introduced in [4]:

1. **Independent bottleneck**: Concept and final predictions are conducted independently in this model. During training, the ground truth concepts c are utilized to train $f(c)$, but not to train $\hat{g}(x)$. During testing, the output of $f(\hat{g}(x))$ is employed.
2. **Sequential bottleneck**: This model predicts concepts from the input data and then predicts the final target based on these predicted concepts. The output of $g(x)$ is directly utilized as the input to f.
3. **Joint bottleneck**: Similar to the sequential bottleneck model, this model predicts concepts c from the input data x and then predicts the final target y based on these predicted concepts. However, it additionally considers the loss function for concept prediction. In other words, it simultaneously minimizes the concept prediction loss $\mathcal{L}_C$ and the final target prediction loss $\mathcal{L}_Y$ during training. To balance the contributions of these two loss functions, a weight $\lambda > 0$ is introduced, defining the total loss function as $\mathcal{L}_{joint} = \mathcal{L}_Y + \lambda \cdot \mathcal{L}_C$. Consequently, the joint bottleneck model performs concept and final target predictions simultaneously, considering the relationship between the two prediction results during training.

3.1.2. Knowledge Distillation

This study focuses on logit-based knowledge distillation (KD) to train the three CBM types. Although feature-based KD presents another approach, it falls outside the scope of this research.

Let us consider a classification task involving K classes and denote the logit output of a single input as $z \in \mathbb{R}^K$. The class probability is computed as

$$\hat{y}_i = \frac{e^{z_i/\tau}}{\sum_{k=1}^{K} e^{z_k/\tau}}, \quad i = 1, \ldots, K, \tag{1}$$

where $\hat{y}_i$ is the probability value for the i-th class, and we compute the softmax function with a temperature-scaling hyperparameter τ. By scaling the outputs, we aim to minimize the Kullback–Leibler (KL) divergence between the teacher's and the student's outputs:

$$\mathcal{L}_{KL} = KL(\hat{y}^T || \hat{y}^S) = \sum_i \hat{y}_i^T \log \frac{\hat{y}_i^T}{\hat{y}_i^S}, \tag{2}$$

where $\mathcal{L}_{KL}$ represents the logit-level KL loss, and $\hat{y}_i^T$ and $\hat{y}_i^S$ denote the probabilities for the i-th category in the teacher and student models, respectively.

The most significant difference in applying KD to CBMs lies in how the logit distillation method is applied differently in each bottleneck model. The functions f and g are learned separately in the independent bottleneck student model. The loss $\mathcal{L}_{KD}$ is solely used to compute $\mathcal{L}_{C_j}$, representing the loss for predicting the j-th concept. Distillation from the teacher for $\mathcal{L}_Y$, representing the loss for the final task output, is unnecessary as it employs the ground-truth concept c. The sequential bottleneck model predicts concepts from the input x and then predicts the task from the predicted concepts. Therefore, only $\mathcal{L}_{KD}$ needs to be calculated in $\mathcal{L}_Y$. Lastly, the joint bottleneck model learns the functions f and g jointly and performs distillation separately for both $\mathcal{L}_{C_j}$ and $\mathcal{L}_Y$ due to the presence of weights connecting the concept bottleneck and the final output.

3.2. Visual Concept Knowledge Distillation

Incorporating explainability is crucial to ensuring a comprehensive understanding of the operation of compressed CBMs. When training CBMs with inherent interpretability using KD, the probability distributions of concept predictions between the teacher and student models tend to become highly similar. However, as discussed earlier, relying solely on post hoc explanations may result in discrepancies between the teacher and student

models' explanations, leading to incomplete explanations presented to the user. To tackle this challenge, our proposed VICO-KD method encourages the student model to learn discriminative regions similar to the teacher model.

In the general KD process, there can be a mixture of information about the foreground and background [41]. Therefore, guiding the student with attribute values the teacher deems important for prediction becomes necessary. To achieve this, we induce alignment of attribute values between the student and teacher based on saliency maps. Given the bottleneck structure of CBMs, concepts are predicted from the input, and task prediction is conducted based on the predicted concepts. From an interpretability perspective, identifying pixels in the image associated with essential concepts for task prediction is crucial. VICO-KD applies the rectified linear unit (ReLU) activation function to consider only the pixels that contribute to the prediction.

Let $A^t \in \mathbb{R}^{C\times H\times W}$ and $A^s \in \mathbb{R}^{C\times H\times W}$ denote the activation maps of the teacher at layer l_t and the student model at layer l_s layer, respectively, where C is the number of channels, H is the height, and W is the width. For each class $k \in 1,\ldots,K$, the saliency maps are computed as follows:

$$\begin{aligned} U^t_{c,k} &= \sum_{h=1}^{H}\sum_{w=1}^{W} \frac{\partial y^t_k}{\partial A^t_{c,h,w}} \odot A^t_{c,h,w} \\ U^s_{c,k} &= \sum_{h=1}^{H}\sum_{w=1}^{W} \frac{\partial y^s_k}{\partial A^s_{c,h,w}} \odot A^s_{c,h,w}, \end{aligned} \tag{3}$$

where y^t_k and y^s_k are the outputs for class k from the teacher and student models, respectively, and $\odot$ denotes the element-wise multiplication. Grad-CAM is calculated for each channel $c \in 1,\ldots,C$.

To consider only the important attributions, we apply the ReLU function:

$$\begin{aligned} M^t_{c,k} &= \mathrm{ReLU}\left(U^t_{c,k}\right) \\ M^s_{c,k} &= \mathrm{ReLU}\left(U^s_{c,k}\right). \end{aligned} \tag{4}$$

Here, $M^t_{c,k} \in \mathbb{R}^{H\times W}$ and $M^s_{c,k} \in \mathbb{R}^{H\times W}$ represent the class-discriminative saliency maps [42] for the teacher and student models, respectively, highlighting the important regions for predicting class k in channel c.

To ensure that the student model learns similar attributions to the teacher model, we compute the L2 loss between the saliency maps:

$$\mathcal{L}_{VC} = \frac{1}{C\cdot K}\sum_{k=1}^{K}\sum_{c=1}^{C}\left\| M^t_{c,k} - M^s_{c,k}\right\|_2^2, \tag{5}$$

This loss function encourages the student model to have similar saliency maps to the teacher model for each class and channel, effectively transferring the visual concept knowledge from the teacher to the student.

3.3. Loss Function

Logit-level KD utilizes the intermediate layer logits from the auxiliary classifiers, as described in [43]. In our study, we introduce the concept of an auxiliary bottleneck. As depicted in Figure 2, the intermediate layers of the teacher and student models are connected to their respective auxiliary task predictors. Subsequently, the KL divergence is calculated between the output distributions of these predictors to quantify the discrepancy between the teacher and student models at the auxiliary bottleneck level.

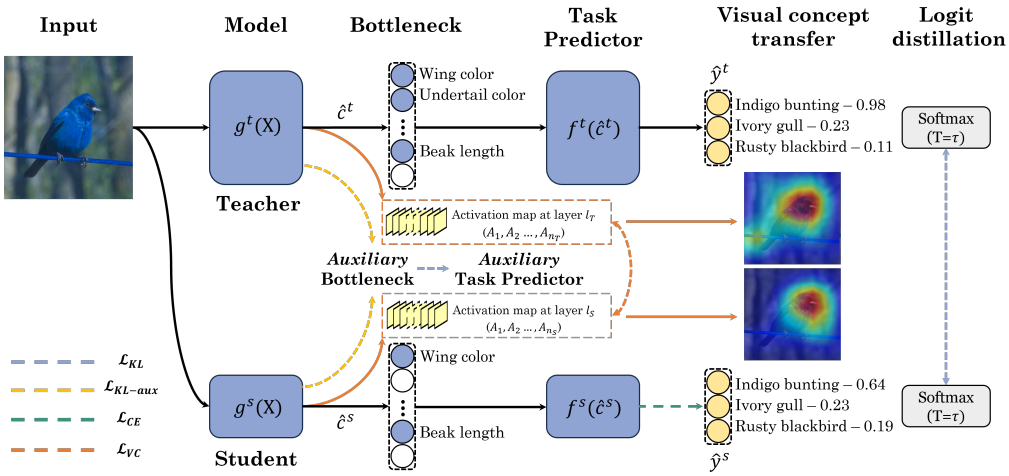


Figure 2. VICO-KD framework. The frameworks encompass three main streams: (1) logit-level KD stream: Illustrated in blue, this stream conducts logit-level KD to align the output distributions of the student and teacher. (2) The independent learning stream of the student: Represented in green, this stream focuses on the independent learning of the student. (3) The visual concept transferring stream: Illustrated in orange, this stream facilitates the transfer of visual explainability from the teacher to the student.

The total logit-level KD loss in a CBM is computed as

$$\mathcal{L}_{logit} = \alpha \cdot (\mathcal{L}_{KL} + \mathcal{L}_{Aux-KL}) + (1 - \alpha) \cdot \mathcal{L}_{CE}, \tag{6}$$

where $\mathcal{L}_{KD}$ represents the KL divergence loss between the final output logits of the teacher and student models, $\mathcal{L}_{Aux-KL}$ represents the KL divergence loss computed using auxiliary classifiers at the intermediate layers, and $\mathcal{L}_{CE}$ is the cross-entropy loss. The coefficient α is utilized to balance the contributions of these loss terms.

VICO-KD is calculated solely at the logit level, and the overall loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{logit} + w \cdot \mathcal{L}_{VC}, \tag{7}$$

where $\mathcal{L}_{VC}$ represents the loss term for transferring visual concept knowledge from the teacher to the student model, and w is a weight parameter that regulates the intensity of this transfer. By minimizing $\mathcal{L}_{VC}$, the student CBM learns to align its visual concept activation with the teacher CBM.

4. Experiments

4.1. Dataset and Experimental Setup

To investigate the effectiveness of the KD method in CBM, we conducted experiments using benchmark datasets with concept annotations, namely, CUB, AwA2, and CelebA. All experiments were conducted on a single NVIDIA Tesla V100 GPU with 32 GB of memory, and the experiments were carried out using the PyTorch framework.

Dataset. We conducted experiments on the CUB, AwA2, and CelebA datasets. For the CUB dataset, we followed the preprocessing steps described in [4]. This dataset comprises 11,788 images of 200 bird species, with 312 attributes annotated for each image. These attributes represent concepts related to the visual characteristics of birds, such as wing color and beak size. We removed noisy attributes and retained only 112 attributes for our experiments. AwA2 contains 37,322 images of 85 animal species, with 85 attributes for each animal. We selected 50 animal species and 50 semantic attributes for our experiments. CelebA consists of 202,599 images of human faces, with 10,177 identities and 40 binary

attribute annotations per image. We utilized a subset of 256 identities and five binary attributes for our experiments on CelebA.

Model Architecture. For CUB and AwA2 datasets, we trained CBM using the bottleneck model training strategy and adopted the same hyperparameters as detailed in [4]. The architecture comprised an Inception-V3 [44] pretrained on ImageNet for the $x \rightarrow c$ mapping and a 3-layer MLP for the $c \rightarrow y$ mapping. Both teacher and student models share the same bottleneck structure, but the student model has a smaller capacity. Specifically, for the student model, we halved the channels in each inception module of the inception-V3 network. In contrast, for CelebA, we utilized the ResNet series to construct the teacher–student models, employing ResNet50 [45] for the teacher and ResNet18 for the student. For quantitative evaluation of the benchmark datasets, we utilize classification and 0-1 errors (for binary concepts) as task and concept errors, respectively.

Experimental setup. During training, both the teacher and student models underwent standard data augmentation techniques, including color jitter, random resized crop, and random horizontal flip. The teacher model underwent training for 1000 epochs, while the student model required a longer training duration of 3000 to ensure stable convergence. To determine the optimal hyperparameters, we conducted a grid search over the following values: optimizer $\in \{\text{SGD}, \text{Adam}\}$; learning rate $\in \{0.001, 0.01\}$; $\mathcal{L}_{KL} \in \{0.5, 1.0, 2.0\}$; $\mathcal{L}_{Aux-KL} \in \{0.5, 1.0, 2.0\}$; and w for $\mathcal{L}_{VC}$, ranging from 0.5 to 4.0 in increments of 0.5.

Evaluation metric. The performance of VICO-KD is evaluated using two metrics: Task accuracy and concept accuracy. These metrics are defined as follows:

$$\text{Task Accuracy} = \frac{1}{N}\sum_{i=1}^{N} 1(\hat{y}_i = y_i), \tag{8}$$

$$\text{Concept Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M} 1(\hat{c}_{i,j} = c_{i,j}), \tag{9}$$

where N is the total number of test samples, and M denotes the number of concepts. Here, $\hat{y}_i$ and y_i are the predicted and ground truth task labels for the i-th sample, respectively, while $\hat{c}_{i,j}$ and $c_{i,j}$ represent the predicted and ground truth concept labels for the j-th concept of the i-th sample.

Task accuracy quantifies the proportion of test samples for which the model correctly predicts the target labels. Concept accuracy measures the model's ability to predict the intermediate concept labels accurately, averaged over all concepts and test samples.

4.2. Experimental Results

4.2.1. Experimental Results on CUB-200-2011

In this section, we conduct a comparative analysis to evaluate the performance of our proposed VICO-KD method against other concept models and Vanilla-KD (without VICO) on the CUB dataset. Furthermore, we explore the influence of varying the strength of the weight hyperparameter (w) in VICO-KD on the model's performance.

Performance comparison. We employed the joint bottleneck in our proposed VICO-KD approach, which is considered the best-performing logit-level KD method. We conducted a comparative analysis, comparing the performance of our method with other concept models, including CEM [46], ProbCBM [13], PCBM [12], and CME [47]. Table 1 summarizes the results, indicating that VICO-KD achieves competitive performance across accuracy metrics. Particularly noteworthy is that VICO-KD outperforms other concept models regarding task accuracy (y), with improvements ranging from 0.004 to 0.152. These performance gains can be attributed to effectively incorporating visual concepts in the

knowledge distillation process, enabling the student model to learn more discriminative and semantically meaningful features.

Table 1. Experimental results on CUB-200-2011.

Model	Task (y)	Concept (c)
Teacher CBM		
Independent	0.767 ± 0.006	0.968 ± 0.001
Sequential	0.759 ± 0.005	0.968 ± 0.001
Joint	0.803 ± 0.015	0.966 ± 0.004
Concept model		
PCBM	0.635 ± 0.002	-
ProbCBM (224×224)	0.680 ± 0.004	0.949 ± 0.001
ProbCBM (299×299)	0.708 ± 0.006	0.956 ± 0.001
CEM	0.759 ± 0.002	0.954 ± 0.001
CME	0.783 ± 0.210	0.508 ± 0.001
Student CBM		
Vanilla-KD (Joint)	0.689 ± 0.024	0.948 ± 0.002
VICO-KD (Joint)	$\mathbf{0.787 \pm 0.002}$	$\mathbf{0.962 \pm 0.001}$

In contrast, compared to the Vanilla-KD model trained without visual concepts, there is a notable difference of 0.098 in task accuracy and 0.014 in concept prediction accuracy. This signifies that our proposed method for training the CBM transfers knowledge at the concept and visual concept levels, effectively enhancing task performance. This improvement can be attributed to the explicit integration of visual concepts in the knowledge distillation process. Transferring knowledge at the class label and visual concept levels allows the student model to acquire more meaningful and discriminative features that align better with human-interpretable concepts.

Weight impact on visual concept transfer. We investigated how the weight hyperparameter w influences the transfer of visual concepts from a teacher CBM to a student CBM, with w values ranging from 0.5 to 4.0, as depicted in Figure 3. Our results demonstrate that increasing w from 0.5 to 2.0 enhances task and concept accuracies, with peak task accuracy at 0.7865 and concept accuracy at 0.9619, both occurring at $w = 2.0$. This indicates optimal concept learning and model performance at this weight value. However, weights above 2.0 lead to declines in performance, indicating potential overfitting and reduced generalization when too much emphasis is placed on visual concept transfer.

Nevertheless, concept accuracy remains consistently high, underscoring the model's robustness. Our findings emphasize the significance of tuning w to strike a balance between task performance and concept learning, recommending a weight of 2.0 for optimal results in visual concept transfer.

Visualization of saliency maps. To gain deeper insights into the effectiveness of VICO-KD in transferring visual concepts from the teacher to the student model, we utilize the Grad-CAM technique to visualize and compare the saliency maps of the teacher CBM, VICO-KD, and Vanilla-KD, as depicted in Figure 4.

Upon analyzing the saliency maps, a striking difference between the teacher CBM and Vanilla-KD becomes apparent. The teacher CBM consistently focuses on the most discriminative parts of the object, aligning well with human intuition. This indicates that the teacher model has learned to attend to the salient visual concepts crucial for accurate object classification. In contrast, Vanilla-KD often attends to regions less relevant to the object of interest, indicating that it may struggle to capture the teacher's decision-making process fully.

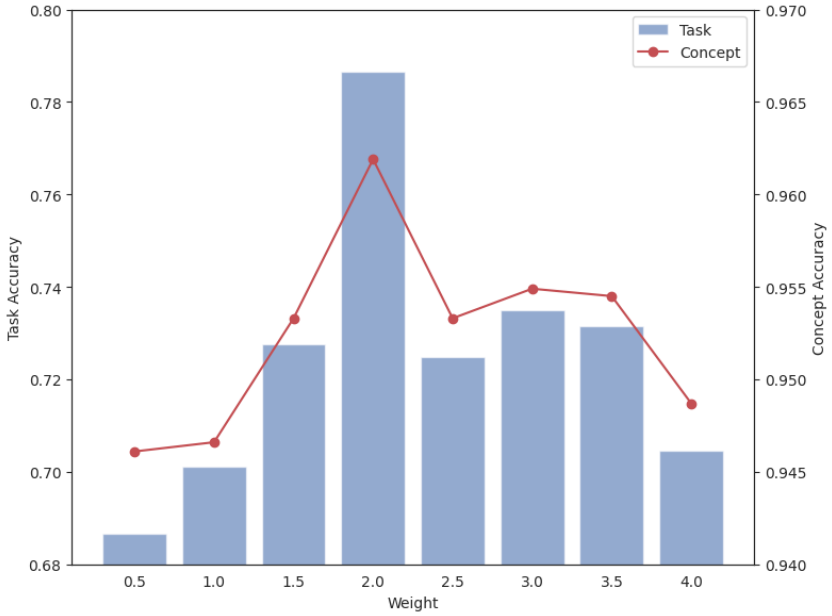
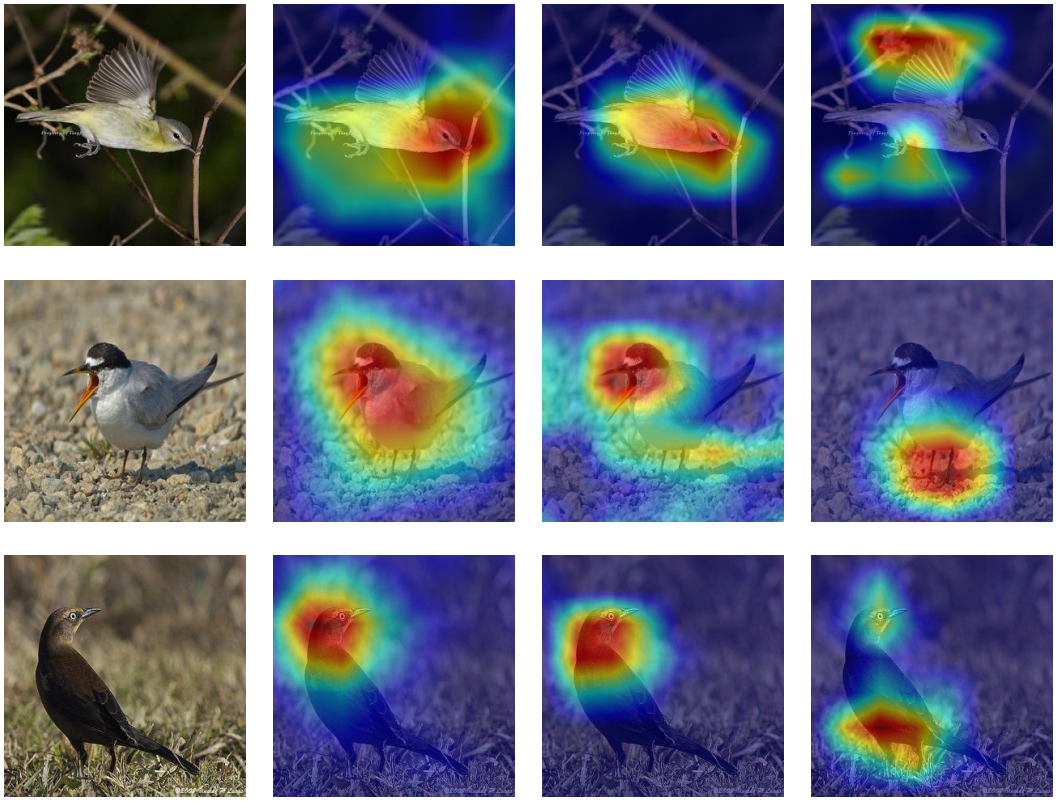


Figure 3. Experimental results on the impact of weight on visual concept transfer.



(**a**) Original image (**b**) Teacher CBM (**c**) VICO-KD (**d**) Vanilla-KD

Figure 4. Visualization comparison results for saliency maps. The top row shows saliency maps for the Philadelphia vireo, the middle row for the tropical kingbird, and the bottom row for the rusty blackbird. The proposed VICO-KD method exhibits explainability similar to that of the teacher CBM. However, Vanilla-KD focuses on parts unrelated to the object, such as the background or legs.

The saliency maps of the student CBM trained using our proposed VICO-KD method demonstrate a notable similarity to those of the teacher CBM. The regions of interest highlighted by VICO-KD consistently align with the salient visual concepts attended to by the teacher model. This finding provides compelling evidence that VICO-KD effectively

transfers visual concepts from the teacher to the student model, enabling the student to acquire a more robust and interpretable representation of the visual task.

In Figure 4, third row, it is evident that both the teacher and VICO-KD student focus on the crucial part of the bird, from the head to the neck, which serves as trustworthy evidence for recognition. This further underscores the effectiveness of VICO-KD in transferring the teacher's visual concept knowledge to the student model, resulting in enhanced performance and interpretability.

4.2.2. Experimental Results on AwA2

Performance comparison. Table 2 presents the experimental outcomes on the AwA2 dataset. Among the teacher CBM models, the joint training approach attains the highest task performance. Utilizing this joint model as the teacher, we contrast the performance of the Vanilla-KD method with the proposed VICO-KD approach. The findings illustrate that the VICO-KD method notably surpasses the Vanilla-KD approach, successfully transferring visual-concept knowledge from the teacher to the student model.

Table 2. Experimental results on AwA2 and CelebA.

	AwA2		CelebA	
Model	**Task (y)**	**Concept (c)**	**Task (y)**	**Concept (c)**
Teacher CBM				
Independent	0.876 ± 0.004	0.944 ± 0.007	0.681 ± 0.002	0.921 ± 0.004
Sequential	0.886 ± 0.004	0.945 ± 0.003	0.699 ± 0.015	0.919 ± 0.007
Joint	0.889 ± 0.005	0.948 ± 0.005	0.743 ± 0.017	0.933 ± 0.010
Student CBM				
Vanilla-KD (Joint)	0.833 ± 0.021	0.935 ± 0.001	0.680 ± 0.004	0.878 ± 0.003
VICO-KD (Joint)	0.862 ± 0.006	0.943 ± 0.001	0.683 ± 0.004	0.887 ± 0.001

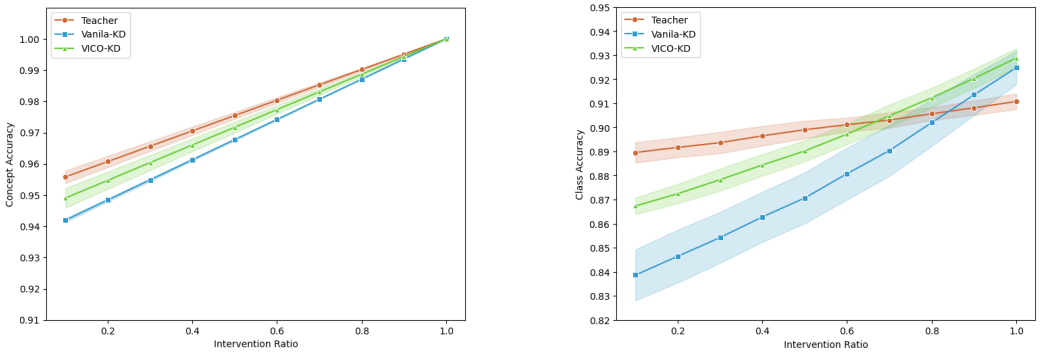
Intervention. The intervention experiment, as outlined in [4], examines the impact of correcting misclassified concepts on the performance of the teacher model, Vanilla-KD, and VICO-KD. As depicted in Figure 5, augmenting the intervention ratio enhances concept and task accuracy for all three methods. Nevertheless, VICO-KD consistently surpasses the teacher model and Vanilla-KD across all intervention ratios, underscoring its superior capacity to utilize corrected concept information and transfer this knowledge to enhance final task performance. Moreover, the performance disparity between VICO-KD and the other methods widens as the intervention ratio escalates, underscoring its scalability and robustness. In essence, the intervention experiment conducted on the AwA2 dataset underscores the efficacy of VICO-KD in enhancing both concept and task accuracies, outperforming the teacher model and Vanilla-KD across diverse intervention ratios.

4.2.3. Experimental Results on CelebA

Performance comparison. Like the AwA2 dataset, the teacher CBM joint model achieves the highest task (y) performance on the CelebA dataset. Employing this joint model as the teacher, the proposed VICO-KD method enhances the student model's performance compared to the Vanilla-KD approach. However, the performance gain is less significant than that observed on the AwA2 dataset, indicating that the efficacy of the proposed method may vary depending on the dataset and task complexity.

Visual concept transfer for domain knowledge with different correlation characteristics. We conducted experiments to validate the effectiveness of visual concept transfer in VICO-KD through concept correlation analysis. Logistic regression analysis categorized features into two groups: features with weak correlations ('Narrow Eyes', 'Bangs', 'Pale

Skin', 'Pointy Nose', 'Oval Face') and features with strong correlations ('Goatee', 'Heavy Makeup', 'Male', 'No Beard', 'Wearing Lipstick').



(**a**) Post-intervention concept accuracy (**b**) Post-intervention class accuracy

Figure 5. Experimental results on the impact of weight on visual concept transfer.

We performed visual concept transfer experiments by assigning concept sets with different correlation characteristics to the teacher and student models. Table 3 shows that VICO-KD outperforms Vanilla-KD across all teacher–student pair scenarios. Notably, the performance improvements in the THSE and TESH scenarios demonstrate VICO-KD's ability to effectively transfer visual cues even when concepts have weak correlations. This suggests that VICO-KD complements the relationships between weakly correlated concepts through visual features, enabling richer feature representations.

Table 3. The experimental results of inexpressive concept predictors. The TS pair configurations encompass various combinations of concept set groups and teacher–student pairs. These configurations are delineated as teacher(easy) → student(easy) (TESE), teacher(hard) → student(hard) (THSH), teacher(hard) → student(easy) (THSE), and teacher(easy) → student(hard) (TESH).

TS Pair	Model	Task (y)	Concept (c)
TESE	Vanilla-KD	0.833 ± 0.021	0.935 ± 0.001
	VICO-KD	0.862 ± 0.006	0.943 ± 0.001
THSE	Vanilla-KD	0.564 ± 0.003	0.789 ± 0.002
	VICO-KD	0.654 ± 0.008	0.847 ± 0.008
TESH	Vanilla-KD	0.299 ± 0.016	0.756 ± 0.050
	VICO-KD	0.373 ± 0.007	0.830 ± 0.001
THSH	Vanilla-KD	0.351 ± 0.005	0.775 ± 0.003
	VICO-KD	0.406 ± 0.005	0.841 ± 0.000

5. Discussion and Limitations

Performance preservation under model compression. Our VICO-KD method demonstrates the ability to maintain model performance while substantially reducing computational requirements. By applying the VICO-KD approach, we achieved significant model compression without compromising accuracy. In our Inception-V3 based experiments on the CUB and AwA2 datasets, we reduced the total number of parameters from 23.8 million to 5.95 million—a 75% reduction—by halving the channel count in each Inception module. In our ResNet-based experiments on the CelebA dataset, we distilled a ResNet50 teacher model (23.5 million parameters) into a ResNet18 student model (11.2 million parameters), achieving a 52% parameter reduction while preserving the model's core performance characteristics.

Spurious correlations in knowledge distillation. Recent studies have revealed that CBM concept predictors can exploit spurious correlations between concepts rather than

their intrinsic characteristics [11]. This issue can be exacerbated during knowledge distillation, particularly when transferring knowledge from teacher to student models.

While VICO-KD attempts to address this by complementing concept relationships through visual features, there is room for improvement. Future research should explore whether alternative knowledge distillation methods, such as feature-based KD, could better mitigate spurious correlations. Additionally, studying the relationship between explicit and implicit concepts could lead to more robust knowledge transfer methods that reduce reliance on spurious correlations. The presence of spurious correlations raises fundamental questions about the interpretability and reliability of distilled knowledge in CBMs. Understanding and addressing these correlations is crucial for developing more trustworthy and interpretable models through knowledge distillation.

Alternative visual feature transfer methods. While our current approach uses saliency maps for visual concept transfer, attention maps could potentially offer an alternative mechanism. Attention maps, which directly capture where the model focuses during inference, might provide more direct insights into the model's decision-making process compared to gradient-based saliency maps. However, attention-based approaches present their own challenges: (1) they would require architectural modifications to incorporate attention mechanisms in both teacher and student models, potentially limiting the method's generality; (2) the relationship between attention patterns and concept learning is less understood compared to saliency maps' relationship with feature importance; and (3) attention maps might capture different aspects of visual concepts than saliency maps, possibly leading to complementary rather than superior results. Future work could explore a hybrid approach that leverages both attention and saliency information for more robust visual concept transfer.

Extension to other modalities. While our current work focuses on visual concept knowledge distillation, recent work has demonstrated the potential of extending concept-based models to other modalities, particularly text data. For instance, Text Bottleneck Models (TBMs) [48] have shown promising results by automatically discovering and measuring text concepts using large language models without human curation. However, adapting VICO-KD to text data presents unique challenges that differ from visual domains: (1) text concepts are often more abstract and contextual compared to visual attributes (e.g., "customer satisfaction" vs. "red color"); (2) the mechanism for concept measurement would need to be fundamentally different since saliency maps are specific to visual data; and (3) maintaining interpretability with sequential data structures poses additional challenges. Future research could explore bridging these gaps by developing text-specific concept transfer mechanisms, perhaps leveraging techniques like attention visualization or embedding space analysis while preserving the core principle of concept-based knowledge transfer.

6. Conclusions

This study proposes a novel knowledge distillation method, VICO-KD, designed for training CBMs. Our approach harnesses the potential of visual concepts to bolster the transfer of knowledge from a teacher model to a student model, thereby enhancing task accuracy, concept prediction performance, and model interpretability. By explicitly integrating visual concepts into the knowledge distillation framework, VICO-KD facilitates the student model's acquisition of more meaningful and discriminative features, aligning them more closely with human-interpretable concepts.

Our proposed VICO-KD method exhibits substantial enhancements over Vanilla-KD when applied to benchmark datasets annotated with concepts. Specifically, it achieves performance improvements of 10.65%, 2.9%, and 0.3% on the CUB, AwA2, and CelebA datasets, respectively. By visualizing and comparing saliency maps, we verify that the CBM trained using VICO-KD acquires visual concepts closely resembling those learned by the

teacher CBM. This observation underscores the capacity of our method to capture not only simple explicit concepts but also more intricate visual concepts, thereby further validating its efficacy.

Through thorough experimentation on benchmark datasets, we substantiate the effectiveness of VICO-KD when compared to existing concept models and logit-based knowledge distillation. Our method attains state-of-the-art performance while ensuring a consistent learning process and fostering more resilient and interpretable visual representations. Consequently, VICO-KD constitutes a considerable advancement in the evolution of interpretable and efficient computer vision models, laying the groundwork for more transparent and dependable AI systems.

While our VICO-KD method demonstrates promising results, future research should focus on addressing potential limitations. Specifically, we aim to develop more robust techniques for mitigating dataset biases and enhancing model fairness. This includes implementing advanced data augmentation strategies, creating comprehensive bias detection frameworks, and designing methodologies that can systematically identify and correct representational skews in concept-based machine learning models.

Moreover, we recognize the critical need to improve model resilience against real-world challenges. Future work will explore adversarial robustness training, noise tolerance mechanisms, and uncertainty quantification techniques. These efforts will not only refine the VICO-KD approach but also contribute to developing more transparent, reliable, and interpretable artificial intelligence systems.

Author Contributions: Conceptualization, J.-H.L.; methodology, J.-H.L.; software, J.-H.L.; investigation, J.-H.L. and I.-H.S.; resources, J.-Y.K.; data curation, J.-H.L.; writing—original draft preparation, J.-H.L.; writing—review and editing, D.T.V. and J.-Y.K.; visualization, J.-H.L.; project administration, N.-K.L.; funding acquisition, N.-K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2024-00156287, 30%) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-02068, Artificial Intelligence Innovation Hub, 30%), Korea government (MSIT) (2022-0-00215, 40%).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The CUB-200-201 dataset can be found at https://www.vision.caltech.edu/datasets/cub_200_2011/ (accessed on 1 March 2024). The Animals with Attributes 2 (AWA2) dataset is available at https://cvml.ista.ac.at/AwA2/ (accessed on 1 March 2024), And Large-scale CelebFaces Attributes (CelebA) dataset can be found at https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (accessed on 1 March 2024).

Conflicts of Interest: Author Dang Thanh Vu was employed by the company AISeed Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
2. Park, G.; Yang, J.Y.; Hwang, S.J.; Yang, E. Attribution preservation in network compression for reliable network interpretation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 5093–5104.
3. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity checks for saliency maps. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.

4. Koh, P.W.; Nguyen, T.; Tang, Y.S.; Mussmann, S.; Pierson, E.; Kim, B.; Liang, P. Concept bottleneck models. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 5338–5348.
5. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *The Caltech-Ucsd Birds-200-2011 Dataset*; California Institute of Technology: Pasadena, CA, USA, 2011.
6. Lampert, C.H.; Nickisch, H.; Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 951–958.
7. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
8. Kumar, N.; Berg, A.C.; Belhumeur, P.N.; Nayar, S.K. Attribute and simile classifiers for face verification. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009; pp. 365–372.
9. Margeloiu, A.; Ashman, M.; Bhatt, U.; Chen, Y.; Jamnik, M.; Weller, A. Do concept bottleneck models learn as intended? *arXiv* **2021**, arXiv:2105.04289.
10. Shin, S.; Jo, Y.; Ahn, S.; Lee, N. A closer look at the intervention procedure of concept bottleneck models. *arXiv* **2023**, arXiv:2302.14260.
11. Raman, N.; Zarlenga, M.E.; Heo, J.; Jamnik, M. Do Concept Bottleneck Models Obey Locality? *arXiv* **2024**, arXiv:2401.01259.
12. Yuksekgonul, M.; Wang, M.; Zou, J. Post-hoc concept bottleneck models. *arXiv* **2022**, arXiv:2205.15480.
13. Kim, E.; Jung, D.; Park, S.; Kim, S.; Yoon, S. Probabilistic Concept Bottleneck Models. *arXiv* **2023**, arXiv:2306.01574.
14. Zhang, R.; Du, X.; Yan, J.; Zhang, S. Decoupling Concept Bottleneck Model. 2022. Available online: https://ieeexplore.ieee.org/document/10740789 (accessed on 1 December 2024).
15. Sawada, Y.; Nakamura, K. Concept bottleneck model with additional unsupervised concepts. *IEEE Access* **2022**, *10*, 41758–41765. [CrossRef]
16. Chauhan, K.; Tiwari, R.; Freyberg, J.; Shenoy, P.; Dvijotham, K. Interactive concept bottleneck models. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 5948–5955.
17. Oikarinen, T.; Das, S.; Nguyen, L.M.; Weng, T.W. Label-Free Concept Bottleneck Models. *arXiv* **2023**, arXiv:2304.06129.
18. Pittino, F.; Dimitrievska, V.; Heer, R. Hierarchical concept Bottleneck models for vision and their application to explainable fine classification and tracking. *Eng. Appl. Artif. Intell.* **2023**, *118*, 105674. [CrossRef]
19. Magister, L.C.; Barbiero, P.; Kazhdan, D.; Siciliano, F.; Ciravegna, G.; Silvestri, F.; Jamnik, M.; Liò, P. Concept distillation in graph neural networks. In *Proceedings of the World Conference on Explainable Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 233–255.
20. Sousa, J.B.; Moreira, R.; Balayan, V.; Saleiro, P.; Bizarro, P. ConceptDistil: Model-Agnostic Distillation of Concept Explanations. *arXiv* **2022**, arXiv:2205.03601.
21. Gutstein, S.; Fuentes, O.; Freudenthal, E. Knowledge transfer in deep convolutional neural nets. *Int. J. Artif. Intell. Tools* **2008**, *17*, 555–567. [CrossRef]
22. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep mutual learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4320–4328.
23. Furlanello, T.; Lipton, Z.; Tschannen, M.; Itti, L.; Anandkumar, A. Born again neural networks. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1607–1616.
24. Yang, C.; Xie, L.; Su, C.; Yuille, A.L. Snapshot distillation: Teacher-student optimization in one generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2859–2868.
25. Cho, J.H.; Hariharan, B. On the efficacy of knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4794–4802.
26. Mirzadeh, S.I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; Ghasemzadeh, H. Improved knowledge distillation via teacher assistant. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 5191–5198.
27. Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; Liang, J. Decoupled knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11953–11962.
28. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
29. Wang, L.; Yoon, K.J. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3048–3068. [CrossRef] [PubMed]
30. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
31. Huang, Z.; Wang, N. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv* **2017**, arXiv:1707.01219.

32. Yim, J.; Joo, D.; Bae, J.; Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4133–4141.
33. Kim, J.; Park, S.; Kwak, N. Paraphrasing complex network: Network compression via factor transfer. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
34. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3967–3976.
35. Tian, Y.; Krishnan, D.; Isola, P. Contrastive representation distillation. *arXiv* **2019**, arXiv:1910.10699.
36. Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; Choi, J.Y. A comprehensive overhaul of feature distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1921–1930.
37. Heo, B.; Lee, M.; Yun, S.; Choi, J.Y. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 3779–3787.
38. Yue, K.; Deng, J.; Zhou, F. Matching guided distillation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 312–328.
39. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
41. Cheng, X.; Rao, Z.; Chen, Y.; Zhang, Q. Explaining knowledge distillation by quantifying the knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12925–12935.
42. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
43. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
44. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
46. Espinosa Zarlenga, M.; Barbiero, P.; Ciravegna, G.; Marra, G.; Giannini, F.; Diligenti, M.; Shams, Z.; Precioso, F.; Melacci, S.; Weller, A.; et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 21400–21413.
47. Kazhdan, D.; Dimanov, B.; Jamnik, M.; Liò, P.; Weller, A. Now you see me (CME): Concept-based model extraction. *arXiv* **2020**, arXiv:2010.13233.
48. Ludan, J.M.; Lyu, Q.; Yang, Y.; Dugan, L.; Yatskar, M.; Callison-Burch, C. Interpretable-by-Design Text Classification with Iteratively Generated Concept Bottleneck. *arXiv* **2023**, arXiv:2310.19660.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.