This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Discovering Fine-Grained Visual-Concept Relations by Disentangled Optimal Transport Concept Bottleneck Models

Yan Xie[1]* Zequn Zeng[1]* Hao Zhang[1]†
Yucheng Ding[1] Yi Wang[1] Zhengjue Wang[2] Bo Chen[1] Hongwei Liu[1]
[1]National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an, 710071, China
[2]State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, 710071, China
{yanxie0904, zzequn99, zhanghao_xidian}@163.com, bchen@mail.xidian.edu.cn

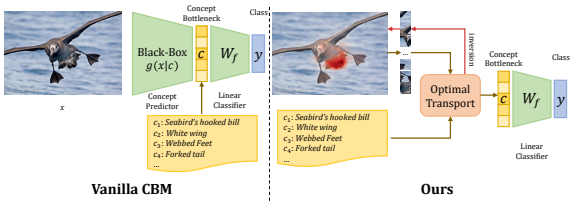
Abstract

Concept Bottleneck Models (CBMs) try to make the decision-making process transparent by exploring an intermediate concept space between the input image and the output prediction. Existing CBMs just learn coarse-grained relations between the whole image and the concepts, less considering local image information, leading to two main drawbacks: i) they often produce spurious visual-concept relations, hence decreasing model reliability; and ii) though CBMs could explain the importance of every concept to the final prediction, it is still challenging to tell which visual region produces the prediction. To solve these problems, this paper proposes a Disentangled Optimal Transport CBM (DOT-CBM) framework to explore fine-grained visual-concept relations between local image patches and concepts. Specifically, we model the concept prediction process as a transportation problem between the patches and concepts, thereby achieving explicit fine-grained feature alignment. We also incorporate orthogonal projection losses within the modality to enhance local feature disentanglement. To further address the shortcut issues caused by statistical biases in the data, we utilize the visual saliency map and concept label statistics as transportation priors. Thus, DOT-CBM can visualize inversion heatmaps, provide more reliable concept predictions, and produce more accurate class predictions. Comprehensive experiments demonstrate that our proposed DOT-CBM achieves SOTA performance on several tasks, including image classification, local part detection and out-of-distribution generalization. Codes are available in supplementary material.

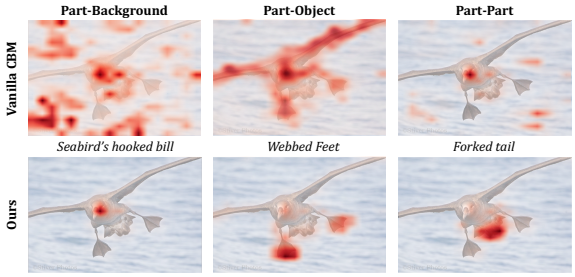
1. Introduction

Explainable artificial intelligence aims to allow human users to comprehend and trust the output created by machine learning algorithms [6, 14]. Compared with post-hoc explanation methods [37] for black-box models, designing explicit interpretability during modeling and learning processes is beneficial to obtaining more natural explanations [8].

The Concept Bottleneck Model (CBM) [17], a representative interpretable image classification model, first predicts a pre-defined set of concepts from images and then



(a) Brief comparison about the coarse-grained design of vanilla CBM and fine-grained design of our proposed DOT-CBM.



(b) Inversion heatmap visualization.

Figure 1. Comparison between vanilla CBM and our proposed DOT-CBM, including architecture design and inversion heatmap for concept predictions. (a) Due to the black-box mapping from images to concepts, Vanilla CBM needs Grad-CAM [37] techniques to locate concept predictions back to image space while DOT-CBM can provide an explicit inversion heatmap visualization. (b) Due to lack of fine-grained alignment, vanilla CBM produces spurious correlations in three levels of granularity (part-background, part-object and part-part level) that mislocalizes the local concept to the background, the whole object, and incorrect local region.

*Equal contribution. †Corresponding authors

maps these concept descriptions to the final output class. Therefore, CBM is good at interpreting the relationship between the input image and output class predictions via intermediate human-understandable concepts. However, vanilla CBM brings these good properties at the expense of classification performance [27]. For this problem, a series of CBM variants are proposed from two lines, construct a more comprehensive and discriminative concept bank [31, 45, 46], and sophisticated architecture design [29, 38, 39] for better concept representation learning.

Although these variants narrow the performance gap between CBMs and their black-box counterparts, their modeling method between images and concepts still has two disadvantages, as shown in Fig. 1a. Firstly, they predict multiple local concepts (*i.e*. local attributes like "head") from a global image feature, which only learns a coarse-grained relationship between the full image and all concepts. Secondly, the interpretability of prevailing CBMs mainly comes from the linear concept-class relations. The mapping from the original image to all concepts is still modeled as the black-box neural networks, which still need some post-hoc analysis [26, 27] to visualize the learned visual-concept relations. As shown in Fig. 1b, these two disadvantages often bring some visual-concept spurious correlations, where many concept predictions are frequently mis-localized in the images. These spurious correlations inherently undermine the interpretability and generalizing capabilities of CBMs.

Empirically, we categorized these visual-concept spurious correlations into three levels of granularity: *i) part-background spurious correlation.* This refers to a misattribution of the concept predictions of local parts to the background. *ii) part-object spurious correlation.* The concepts of local parts are wrongly associated with the entire object without further distinctions. *iii) part-part spurious correlation.* In this case, the local concepts are mislocalized to incorrect image regions. We attribute these spurious correlation problems to two key factors. The first one is the lack of fine-grained alignment between the local image regions and corresponding textual concepts, *i.e*. the fine-grained visual-concept relations. The second one is the overfitting of the data bias such as the statistically high co-occurrence between class and background and between concept and concept, yielding confusion to learn the difference between them. Besides, it is intuitive that the explicit interpretability modeling for the concept predictions from the images is conceivably effective in addressing these visual-concept spurious correlations.

To this end, we propose a novel **D**isentangled **O**ptimal **T**ransport **C**oncept **B**ottleneck **M**odels, namely **DOT-CBM**. Compared with current CBMs, our proposed DOT-CBM can provide holistic explicit interpretability on not only the concept-class relations but also the visual region-concept relations. Specifically, we introduce a disentangled optimal transport (DOT) framework to ensure the intra-modality disentanglement and inter-modality alignment between the image patches and textual concepts. The transport assignment solution of OT framework, is an explicit correlation matrix between image patches and pre-defined concepts, facilitating bidirectional outputs including concept predictions on concept space and concept inversion mask on the image space. Furthermore, to tackle the overfitting to the data bias, we introduce two discrete prior distributions on the patch set and the concept set to penalize the incorrect shortcut alignment between patches and concepts.

In summary, our contributions are as follows:

- We empirically investigate the visual-concept spurious correlation phenomenon and pinpoint the two key factors of these limitations, *i.e*. lack of fine-grained visual-concept alignment and overfitting to the data bias.
- We propose a novel disentangled optimal transport framework to model the fine-grained visual-concept alignment and address the overfitting problems. The transport design between image patches and textual concepts provides explicit relations between image and concept predictions, which can produce concept predictions on the concept spaces and concept inversion mask on the image space.
- Extensive experiments on image classification, part detection, and out-of-distribution generalization benchmarks demonstrate our proposed DOT-CBM can achieve significant improvements over current state-of-the-art CBMs on classification accuracy, interpretability and generalizing capabilities to new distributions.

2. Related Work

Concept Bottleneck Models are a family of models [17] that first predict concepts from input images, then use these concepts to predict a downstream label. This design allows CBMs to (1) provide concept-based explanations via their predicted concepts, and (2) improve their test performance when deployed with experts via concept interventions [2, 11, 40]. Similar to previous interpretable models, CBMs also face the issue of balancing task accuracy and interoperability [10, 27]. Recent works [31, 46] has leveraged extra knowledge graph [38, 46], expert rules [18] or Large Language Models (LLMs) [22, 31, 38, 45, 47] to overcome the limitations of manually defined concept sets and has utilized pre-trained Vision-Language Models (VLMs) [24, 42, 45, 49, 50] to obtain concept predictions. These advancements have elevated the performance of CBMs to levels approaching those of black-box models.

The interpretability of prevailing CBMs primarily arises from linear relationships between concepts and classes. However, the mapping from the image to the concepts is still handled by black-box neural networks, requiring post-hoc analysis for visualization of the learned visual-concept

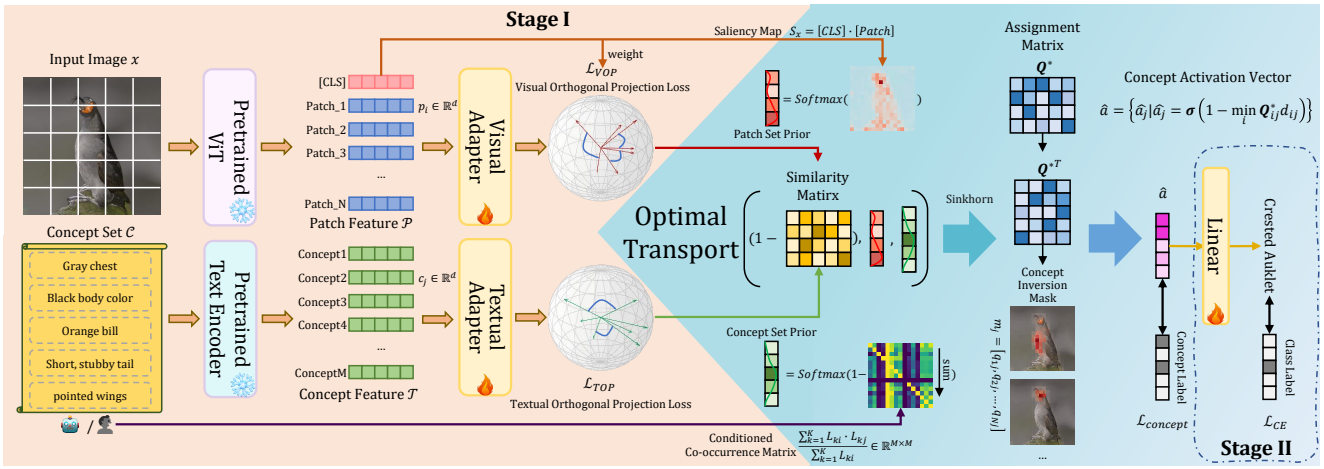


Figure 2. Overview of our proposed DOT-CBM. The overall CBM framework proceeds from left to right. In the first stage, the model transforms input images into concept activation vectors. In forward pass, local embeddings from a pre-trained Vision Transformer (ViT) and a text encoder are processed through learnable adapters to generate two feature sets (Sec. 3.2). Two orthogonal projection losses are applied to both feature sets to constrain Adapter training. An Optimal Transport (OT) algorithm optimizes the Assignment Matrix, representing the explicit correlation between concepts and image patches. Concept activation values are derived by combining this matrix with a cost matrix, and the loss is supervised by concept labels for training (Sec. 3.3).To address data bias, we use the Saliency Map from the pre-trained ViT for rough foreground-background separation as the visual prior in OT. A Conditioned Co-occurrence Matrix, based on frequency statistics, serves as the prior for concepts, enhancing the model's ability to distinguish co-occurring concepts (Sec. 3.4). In the second stage, consistent with the vanilla CBM framework, concept activation values are used to predict final class labels via a linear classification network, which is trained using class labels (Sec. 3.5).

relations. This lack of transparency often leads to visual-concept spurious correlations, where concept predictions are frequently mis-localized in the image. We propose a disentangled OT to address these limitations.

Optimal Transport aims to find the most efficient way to map one distribution onto another by minimizing a transportation cost. In the context of machine learning, OT is widely applied in a variety of domains, including unsupervised domain adaptation [36], label representation [12, 52], cross-modal semantics [3, 4, 20, 21, 23, 25], and beyond.

In this work, we further extend OT to a disentangled OT framework, to enforce the fine-grained alignment between image patches and their corresponding concepts while also ensuring disentanglement within each modality.

3. Method

In this section, we begin with a brief review of the CBM problem setup in Sec. 3.1 including two stages of optimization, *i.e.* patches to concepts ($f_{\mathcal{X}\to\mathcal{C}}$) and concepts to classes ($f_{\mathcal{C}\to\mathcal{Y}}$). In the first stage, to address the limited interpretation of previous CBMs in the patches-to-concepts process, we introduce a disentangled optimal transport alignment framework between patches and concepts. Concretely, in Sec. 3.2, we propose two orthogonal projection losses to disentangle the patch features and concept features. In Sec. 3.3, we introduce an optimal transport (OT) based framework to model the complex relationship between patches and concepts, which can predict concepts from patches and meanwhile inverse the concept to patches to provide explanations of concept prediction. In Sec. 3.4, we integrate two prior distribution modeling to address the data bias problem. In the second stage, in Sec. 3.5, we follow vanilla CBMs and predict classes from concepts. Overall, our proposed DOT-CBM achieves holistic interpretability whether on intermediate concept predictions or on final class predictions. The whole pipeline is shown in Fig. 2.

3.1. Problem Formulation

Traditional black-box visual models aim to fit the conditional distribution $P(y|x)$ given an a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$, where $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$ are images and class labels, respectively. Though effective, they still lack transparency to understand the underlying decision-making process to drive the final predictions.

Concept Bottleneck Models (CBMs) [17] offer an interpretable solution by first predicting human-interpretable high-level concepts before making the final prediction. Formally, the dataset is extended to $\mathcal{D} = \{x_i, y_i, \boldsymbol{a}_i\}_{i=1}^N$, where $\boldsymbol{a}_i = \{a_{ij} | a_{ij} \in \{0, 1\}\}_{j=1}^M$ represents a binary concept activation label. $a_{ij} = 1$ indicates the presence of a specific high-level concept c_j in image x_i, while $a_{ij} = 0$ means the absence. c_j is concept from the pre-defined concept set $\mathcal{C} = \{c_j\}_{j=1}^M$. The optimization process of CBMs can be divided into two-stage [26, 35]: *1)* they first leverage a con-

cept predictor $f_{\mathcal{X}\to\mathcal{C}}: \mathbb{R}^{H\times W\times C} \to \mathbb{R}^M$ to map the input data x to the concept activation vector $\boldsymbol{a} \in \mathbb{R}^M$. *2)* Subsequently, a target predictor $f_{\mathcal{C}\to\mathcal{Y}}: \mathbb{R}^M \to \mathbb{R}^{N_y}$ maps the concept activation vector $\boldsymbol{a}$ to the final label y. This structure enhances the interpretability of the model, as the intermediate concept predictions provide interpretability into the model's final decision. However, the interpretability from image to concept is un-explored.

3.2. Disentanglement on Local representations

Disentangled visual patch features $\widetilde{\mathcal{P}}$. As previously discussed, current CBMs typically associate multiple local concepts with a global image feature, which leads to coarse-grained correspondences. In this paper, we adopt the pre-trained vision transformer (ViT) [1] model, which has been approved to exhibit better spatial location information extraction than CNN [34], to extract local patch features to represent the local image regions.

Specifically, as shown in the left part of Fig. 2, ViT uniformly divides the full image $x \in \mathbb{R}^{H\times W\times C}$ into N non-overlapping patches $\{I_{pi} \in \mathbb{R}^{P_h\times P_w}\}_{i=1}^N$. After linearly projecting these patches into a d-dimension embedding space and flattening all patch embeddings into a sequence, we can get an initialized patch embedding matrix $X_P \in \mathbb{R}^{N\times d}$. To extract the global image feature, ViT concatenates a learnable special token [CLS] with X_P. Besides, in order to preserve the spatial information of images, ViT adds a learnable position embedding $X_{pos} \in \mathbb{R}^{(N+1)\times d}$ and we get the input embeddings X for ViT:

$$X_P = [I_{p1}\cdot W_p, ..., I_{pN}\cdot W_p], \tag{1}$$

$$X = [[\text{CLS}], X_P] + X_{pos}, \tag{2}$$

where $W_p \in \mathbb{R}^{(P_h\cdot P_w\cdot C)\times d}$ is the patch embedding projection weights. Then, the pre-trained ViT E_I to map the input embeddings X into the global feature g_I, and patch features $\mathcal{P} = \{p_i\}_{i=1}^N$:

$$[g_I, \mathcal{P}] = E_I(X), \tag{3}$$

We empirically find that the patch features within the object regions are highly similar which is good for object representation, but bad for distinguishing local parts within the object region. Thus, we fit a learnable visual adapter A_v on the $\mathcal{P}$ and map the patch features within the foreground region to a disentangled unit hyperspere by introducing a visual orthogonal projection (VOP) regularizer as:

$$S_I = g_I \cdot P, \tag{4}$$

$$\langle A_v(p_i), A_v(p_j)\rangle = \frac{A_v(p_i)\cdot A_v(p_j)}{\|A_v(p_i)\|\|A_v(p_j)\|}, \tag{5}$$

$$\mathcal{L}_{\text{VOP}} = \mathop{\mathbb{E}}_{(p_i,p_j)\sim\mathcal{P}}[S_I(i)\cdot S_I(j)\cdot\langle A_v(p_i), A_v(p_j)\rangle] \tag{6}$$

where S_I is the saliency map to distinguish foreground and background, $\langle\cdot,\cdot\rangle$ is the cosine similarity. Finally, we get the disentangled patch features $\widetilde{\mathcal{P}} = \{\widetilde{p_i} = A_v(p_i)\}_{i=1}^N$

Disentangled textual concept features $\widetilde{\mathcal{T}}$. For concept feature extraction, following [22, 29, 45, 47], we utilize a pre-trained text encoder E_T to compute concept embeddings $\mathcal{T} = \{t_j = E_T(c_j)\}_{j=1}^M$. Similarly, we fit a learnable textual adapter A_t and introduce a textual orthogonal projection (TOP) regularizer to disentangle the concept features in a unit hyper-sphere, as:

$$\mathcal{L}_{\text{TOP}} = \mathop{\mathbb{E}}_{(t_i,t_j)\sim\mathcal{T}}[\langle A_t(t_i), A_t(t_j)\rangle], \tag{7}$$

Consequently, we get the disentangled concept features $\widetilde{\mathcal{T}} = \{\widetilde{t_j} = A_t(t_j)\}_{j=1}^M$.

3.3. Patch-concept alignment via optimal transport

Now, we have the disentangled patch feature set $\widetilde{\mathcal{P}}$ and disentangled concept feature set $\widetilde{\mathcal{T}}$, we need to model the relationship $\mathbf{Q}$ between these two sets and derive the concept activation vector $\hat{\boldsymbol{a}}$ based on the input image. As a complex many-to-many alignment problem, we propose an optimal transport (OT) based binary cross-entropy loss to tackle these problems. Our goal is to correctly align the patch features and concept features with the same semantics.

Assume that these two sets $\widetilde{\mathcal{P}}$ and $\widetilde{\mathcal{T}}$ follow two discrete probability distributions $\boldsymbol{\Theta} \in \mathbb{R}^N$ and $\boldsymbol{\Gamma} \in \mathbb{R}^M$ on the d-dimension embedding space, respectively:

$$\boldsymbol{\Theta} = \sum_{i=1}^N \theta_i \delta_{p_i}, \quad \boldsymbol{\Gamma} = \sum_{j=1}^M \gamma_j \delta_{c_j}, \tag{8}$$

where $\boldsymbol{\theta} \in \Sigma^N$ and $\boldsymbol{\gamma} \in \Sigma^M$, the simplex of $\mathbb{R}^N$ and $\mathbb{R}^M$, denote the probability values of the discrete states with satisfaction that $\sum_{i=1}^N \theta_i = 1$ and $\sum_{j=1}^M \gamma_j = 1$. δ_e refers to a point mass located at coordinate $e \in \mathbb{R}^d$. The prior modeling of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are shown in the Sec. 3.4.

To get the transport assignment $\mathbf{Q}$ between $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}$, the objective of OT is formulated as the optimization problem:

$$\begin{aligned} &\min_{\mathbf{Q}\in\mathcal{Q}} \sum_{i,j} q_{ij} d_{ij} - \varepsilon H(\mathbf{Q}) \\ &\text{s.t. } \mathcal{Q} = \left\{\mathbf{Q} \in \mathbb{R}_+^{N\times M} \mid \mathbf{Q}\mathbf{1}_M = \boldsymbol{\theta}, \mathbf{Q}^\top \mathbf{1}_N = \boldsymbol{\gamma}\right\}. \end{aligned} \tag{9}$$

where $\mathcal{Q}$ is the set of doubly stochastic matrices $\mathbf{Q} = [q_{ij}]_{N\times M}$, denotes the corresponding transport assignment where q_{ij} represents the probabilities of aligning p_i with t_j. $\mathbf{1}^M$ is the M-dimensional vector of ones. $d_{ij} = d(p_i, t_j) = (1 - \langle p_i, t_j\rangle) \geq 0$ is the transport cost between the two points p_i and s_j and we use the cosine distance to compute cost matrix $\mathbf{D}$. The optimal $\mathbf{Q}^*$ of (9) has a simple normalized exponential matrix solution by Sinkhorn fixed point iterations [7],

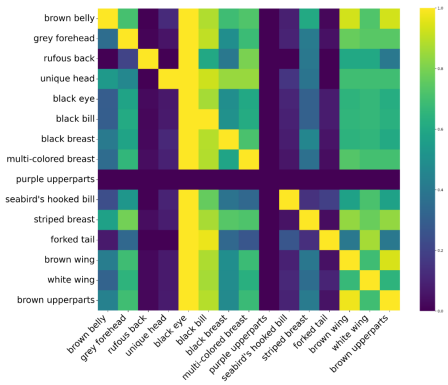


Figure 3. Conditioned co-occurrence matrix of concept labels.

$$\mathbf{Q}^* = \text{Sinkhorn}(\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\gamma}), \tag{10}$$

After obtaining the optimal transport assignment $\mathbf{Q}^*$, the OT distance between the two sets is expressed as:

$$d_{OT} = \sum_{i,j} q_{ij} d_{ij} \tag{11}$$

For the concept embedding t_j, we only consider the minimum distance $d_{OT_j} = \min_i (q_{ij} d_{ij})$ between this concept embedding t_j and all patch embeddings to facilitate the concept activation value prediction ($x \rightarrow \mathcal{C}$), the predicted concept activation vector $\hat{\boldsymbol{a}}$ are stated as:

$$\hat{\boldsymbol{a}} = \{\hat{a_j} | \hat{a_j} = \boldsymbol{\sigma}(1 - d_{OT_j})\}_{j=1}^{M}. \tag{12}$$

where $\boldsymbol{\sigma}$ is the sigmoid function. Moreover, we can inverse specific concepts to the image patches based on $\mathbf{Q}^*$, and get the inversion mask m_j of concept c_j, as:

$$m_j = [q_{1j}, q_{2j}, ..., q_{Nj}]. \tag{13}$$

By utilizing OT distance between patches and concepts as the patch-concept transport, our OT-based binary cross-entropy loss $\mathcal{L}_{\text{concept}}$ are as follows:

$$\mathcal{L}_{\text{concept}} = -\mathbb{E}[\boldsymbol{a} \log(\hat{\boldsymbol{a}}) + (1 - \boldsymbol{a}) \log(1 - \hat{\boldsymbol{a}})] \tag{14}$$

Now, we can get the total objective of stage 1 ($f_{\mathcal{X} \rightarrow \mathcal{C}}$), as:

$$\mathcal{L}_{f_{\mathcal{X} \rightarrow \mathcal{C}}} = \lambda_1 \mathcal{L}_{\text{VOP}} + \lambda_2 \mathcal{L}_{\text{TOP}} + \lambda_3 \mathcal{L}_{\text{concept}}, \tag{15}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the hyperparameters.

The obtained assignment $\mathbf{Q}^* = [q_{i,j}]_{N \times M}$ indicates the relationship between patches and concepts, where $q_{i,j}$ represent correlation between patch p_i and concept c_j. Consequently, based on $\mathbf{Q}^*$, we can obtain concept activation $\boldsymbol{a}$ (Eq.(12)) based on the image features ($x \rightarrow \mathcal{C}$) and meanwhile inverse the concept to the original image patches (Eq.(13)) ($\mathcal{C} \rightarrow x$) to provide explicit interpretation.

3.4. Prior distributions Modeling

The prior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ of OT are typically uniform distributions, which is sub-optimal for our task. Considering the part-background spurious correlation and part-part spurious correlation problem mentioned in Fig. 1b, we propose two prior distribution modeling in this section to address these limitations, respectively.

Patch set prior $\boldsymbol{\theta}$. We attribute the part-background spurious correlation phenomenon to the high co-occurrence of some classes and backgrounds, such as "seabird" and "water". To solve this biased association between object concepts and the background, we use saliency map S_I from Eq. (4) to encourage the transports between the patch features within foreground regions and concepts and penalize the transports of background regions. Specifically, the prior distribution of patch set $\boldsymbol{\theta}$ are formulated as:

$$\boldsymbol{\theta} = \text{Softmax}(S_I), \tag{16}$$

Concept set prior $\boldsymbol{\gamma}$. As illustrated in Fig. 3, some concepts are highly co-occurrent. These concept label biases lead to the model's confusion between these concepts, yielding the part-part spurious correlation phenomenon. To address this problem, we employ the statistical co-occurrence matrix of the training dataset to facilitate the prior distribution modeling of $\boldsymbol{\gamma}$, aiming to penalize those concepts with high co-occurrence with other concepts.

Specifically, for a dataset with M concepts and K classes, class-concept label form a matrix $L = [l_{i,j}]_{K \times M}$, where $l_{i,j}$ represents the activation value of the j-th concept in the concept label for the i-th class. First, we count the co-occurrence frequencies of each pair of concepts. Then, we normalize these counts by the total number of times each concept is activated, resulting in the conditional co-occurrence matrix. The conditional co-occurrence vector r_i is calculated as follows:

$$r_i = \sum_{j=1}^{M} \left(\frac{\sum_{k=1}^{K} L_{ki} \cdot L_{kj}}{\sum_{k=1}^{K} L_{ki}} \right) \tag{17}$$

The concept set prior distribution $\boldsymbol{\gamma}$ is then derived from the conditional co-occurrence vector as follows:

$$\boldsymbol{\gamma} = 1 - \text{Softmax}(r_i) \tag{18}$$

This prior modeling ensures that those concepts with high co-occurrence with other concepts will be penalized.

3.5. Concept $\rightarrow$ class prediction

Now, we already finished the first stage which obtains the intermediate concept activation predictions $\boldsymbol{a}$. In the second stage, our goal is to learn the association between concept and final class prediction similar to vanilla CBM. Specifically, we can map the $\boldsymbol{a}$ to final output $\hat{y}$, as:

$$\hat{y} = W_F \cdot \boldsymbol{a}, \tag{19}$$

The classification objective of stage 2 ($f_{\mathcal{C}\to\mathcal{Y}}$) is the cross-entropy loss $\mathcal{L}_{\text{CE}}(y, \hat{y})$.

4. Experiment

To evaluate the effectiveness of our proposed DOT-CBM, we conduct experiments on the image classification task and part detection task. To further demonstrate our method's abilities to reduce spurious correlations, we also conduct experiments on out-of-distribution generalization tasks. For the concept-to-class linear layer, we omit explicit sparse regularization (e.g., L1 penalties) during training. Despite this, the model spontaneously exhibits significant sparsity in the weight matrix due to two orthogonal losses. Comprehensive training details, training efficiency results, quantitative evaluations of sparsity and additional experiments are listed in the Appendix.

4.1. Datasets and Metrics

For the image classification task. We evaluate our method on the CUB [41] and AwA2 [43] datasets, which are widely adopted in the interpretable community. The CUB dataset contains 11,788 images of 200 bird subcategories, and we use the official train-test split for fine-grained image classification. The AwA2 dataset includes 37,322 images across 50 animal categories, and we follow the same train-test split as in [17]. Additionally, we conduct experiments on the larger and more complex ImageNet-1K[9] and CIFAR-100[19] datasets. On these classification datasets, we leverage the classification accuracy to measure the performance of our method. We also conducted experiments on a more complex dataset (Places365), where the complexity arises from discriminative content being image backgrounds. Quantitative and qualitative results are listed in Appendix.

For the part detection task. To evaluate the accuracy of the inversion mask produced by DOT-CBM, we evaluate our method on some part detection benchmarks, including CUB dataset, Part-ImageNet [13] and PASCAL-Part [5] datasets, which are widely used for part segmentation tasks. We also test our model on the RIVAL-10 dataset [28], which is specifically designed for evaluating model interpretability. On the CUB dataset, we obtain the bounding boxes for local parts that are commonly used in part detection tasks[44, 51]. The RIVAL-10 dataset consists of over 26K images retrieved from the ImageNet dataset according to the CIFAR-10 [19] categories, and it includes part attribute segmentation masks for the main subjects in the images. The Part-ImageNet and PASCAL-Part datasets are part segmentation datasets that provide fine-grained attribute segmentation masks. We follow [30] and preprocess these datasets by focusing on the animal data and merging overly fine-grained part attribute masks. The detailed data preprocessing procedures are described in the Appendix. We use the bounding boxes of these masks as the ground truth of part detection. On these datasets, we use the $mAP_{0.5}$ metric to evaluate the precision of predicted concept inversion masks.

4.2. Implementation Details

We employ the pre-trained DINOv2 ViT-L/14 [32] as the image encoder and pre-trained CLIP text encoder [33] as the text encoder. The feature dimension in both encoders is $d = 768$. We resize all images to $H \times W = 252 \times 252$ as input resolution in both training and testing phases throughout all experiments. For both adapter A_v and A_t, we use a three-layer MLP with ReLU activation functions and a dropout rate of 0.4. The optimization of DOT-CBM is done by AdamW with an initialized learning rate of $1e-4$, and decreased every ten epochs by a factor of 0.5. We set weight decay of 10-5. Batch sizes are set as 128 with maximally 200 training epochs. All experiments are conducted on a single RTX 3090 GPU and our model is implemented in PyTorch. We set $\lambda_1 = 0.15, \lambda_2 = 0.2, \lambda_3 = 0.8$ for CUB dataset. Hyperparameters setting on other datasets and the range ablation study are in the Appendix.

4.3. Main Results

Baselines. We compare our proposed DOT-CBM with several state-of-the-art CBM baselines from the literature. *Vanilla-CBM* [17] is the first CBM model, proposing the design of the intermediate concept layer. Vanilla-CBM introduces three training strategies, *i.e*. independent, sequential and joint training, in which the independent training can achieve the best interpretation [26]. Our two-stage optimization belongs to the independent training strategy. Additionally, we compare our performance with recent CBM variants including *CEM* [48], *LaBo* [45], *SparseCBM* [38], *CoopCBM* [39]. Due to the fact that pre-defined concepts and image backbone will significantly affect the final performance. For a fair comparison, we re-implement all baselines with the high-quality concepts produced by the CDL [47] and a pre-trained ViT backbone to extract image features, the same as our method. The performance of all re-implemented baselines exceeds the best performance reported in their papers.

Image classification results. As shown in Table 1, we compare our proposed DOT-CBM with all baselines on four common image classification benchmarks. The classification results show that the DOT-CBM achieves superior performance on three benchmarks. Specifically, the performance improvements over the state-of-the-art (SOTA) are 0.91 points on ImageNet, 3.29 points on CUB, and 1.08 points on CIFAR100. Our method only lags behind CoopCBM by 0.25 points on the AwA2 dataset.

Part detection results. A key property of our proposed DOT-CBM model is able to provide explicit explanations for the concept predictions, via the locate the concept to the

Method	Classification Accuracy (↑)				Part Detection $mAP_{0.5}$ (↑)			
	ImageNet	CUB	CIFAR100	AWA2	PartImageNet	CUB	RIVAL	PASCAL-Parts
Vanilla-CBM [17]	79.17	78.32	80.04	93.15	26.94	27.02	26.76	17.32
CEM [48]	81.29	80.47	81.23	95.92	29.19	30.86	30.83	20.85
LaBo [45]	82.93	81.30	84.10	96.92	38.27	39.28	40.34	30.17
SparseCBM [38]	82.85	82.07	84.75	95.56	39.84	40.23	40.68	33.71
CoopCBM [39]	82.73	82.10	84.66	**97.08**	41.17	45.16	43.82	35.75
DOT-CBM	**83.84** (+0.91)	**85.39** (+3.29)	**85.83** (+1.08)	96.83 (-0.25)	**50.12** (+8.95)	**53.47** (+8.31)	**50.93** (+7.11)	**44.18** (+8.43)

Table 1. Performance comparison of Classification Accuracy (%) and Part Detection ($mAP_{0.5}$).

Method	CUB		Dogs	
	ID	OOD	ID	OOD
Vanilla-CBM [17]	86.9	27.7	87.3	29.4
CEM [48]	84.0	34.0	83.8	36.5
LaBo [45]	85.3	39.4	85.9	41.7
Sparse-CBM [38]	84.5	35.8	85.8	38.6
Coop-CBM [39]	85.8	36.2	86.1	40.3
DOT-CBM w/o. prior	86.5	36.9	87.0	40.8
DOT-CBM w. prior	**87.2** (+0.3)	**49.7** (+10.3)	**89.1** (+1.8)	**52.4** (+10.7)

Table 2. Performance comparison of OOD generalization.

original image space, *i.e*. the concept inversion mask derived from the transport assignment. This concept inversion mask in the image space indicates the most influential image region contributing to the concept predictions. By contrast, all previous CBMs, using a black-box neuron network to model the mapping from global image features to concept predictions, can only inverse the concept prediction to the image space via the post-hoc Grad-CAM. To evaluate the the accuracy of concept inversion mask, we conduct experiments on the part detection datasets and use the ground-truth bounding boxes of the parts to assess the localization performance. To measure the similarity with ground-truths, we need to process our concept inversion mask output and heatmap outputs of baselines (derived from Grad-CAM) into rectangle boxes format. Specifically, we set the threshold at 0.6 of the maximum peak value output of the mask. The local regions above this threshold are considered as the corresponding regions of the concept in the original image. We take the bounding box that encloses this region as the predicted bounding box for the concept.

From Table. 1, we can observe that DOT-CBM significantly surpasses all baselines (average 8 points improvements), demonstrating the superiority of our design for fine-grained patch-concept alignment. The Coop-CBM achieves performance that is second only to ours due to the introduction of an orthogonality loss similar to our proposed textual orthogonal project $\mathcal{L}_{\text{TOP}}$, which further proves the validity of our proposed disentangled OT framework. Consequently, DOT-CBM can inherently produce accurate and explicit explanations for intermediate concept predictions. We also conduct comparison with other ViT-appropriate attribution method. Results shown in Appendix.

4.4. Out-of-distribution generalization

As discussed in the previous sections, there exists data bias in many datasets, leading to shortcut learning. For example, classifying a dog as a huskie based on the snowy background due to the high co-occurrence in the training set, which is analogous to the part-background spurious correlation discussed in Fig. 1b. To validate the effectiveness of our method with background shifts, we follow [39] to segment the foreground object in CUB[41] and Dogs[15] datasets using SAM [16] with bounding box annotations and add a colored background to all images. Each class is associated with a randomly generated color background with an 80% probability in the training set. The in-distribution (ID) test set contains images with a similar color background probability as the training set, while the color similarity of out-of-distribution (OOD) test set is reduced to 30%. Results are shown in Table 2. As we can see, DOT-CBM can achieve comparable results on ID set and surpass the performance of baselines on the OOD set by a large margin, demonstrating that DOT-CBM can effectively decrease the biased association between object concepts and background. To investigate whether the improvements of tackling part-background spurious correlation are derived from the patch set prior modeling as mentioned in Sec. 3.4, we conduct ablations for the patch set prior and demonstrate the efficacy of patch set prior $\boldsymbol{\theta}$.

4.5. Ablation Studies

In this section, we investigate the effects of each component in our framework. Ablation results are listed in Table 3. The baseline has the same trainable networks as DOT-CBM, including two adapters and a linear classification layer. The baseline replaces the OT assignment with a dot product of patch features and concept features.

Effect of OT design (OT). For the comparison between the first row and second row in Table. 3, we can find the

OT	Prior	LrD	Classification Accuracy (↑)				Part Detection $mIOU$ (↑)			
			ImageNet	CUB	CIFAR100	AWA2	PartImageNet	CUB	RIVAL	PASCAL-Parts
✗	✗	✗	81.60	80.92	82.73	96.07	0.35	0.41	0.37	0.28
✔	✗	✗	82.03	81.75	82.91	96.24	0.42	0.49	0.42	0.34
✔	✔	✗	83.22	84.92	85.49	96.47	0.50	0.61	0.53	0.46
✗	✗	✔	82.73	82.10	84.66	97.08	0.47	0.58	0.49	0.41
✔	✔	✔	83.84	85.39	85.83	96.83	0.52	0.66	0.54	0.49

Table 3. Ablation studies on image classification and part detection. OT, Prior, and LrD denote the optimal transport design, prior distribution modeling, and local representation disentanglement.

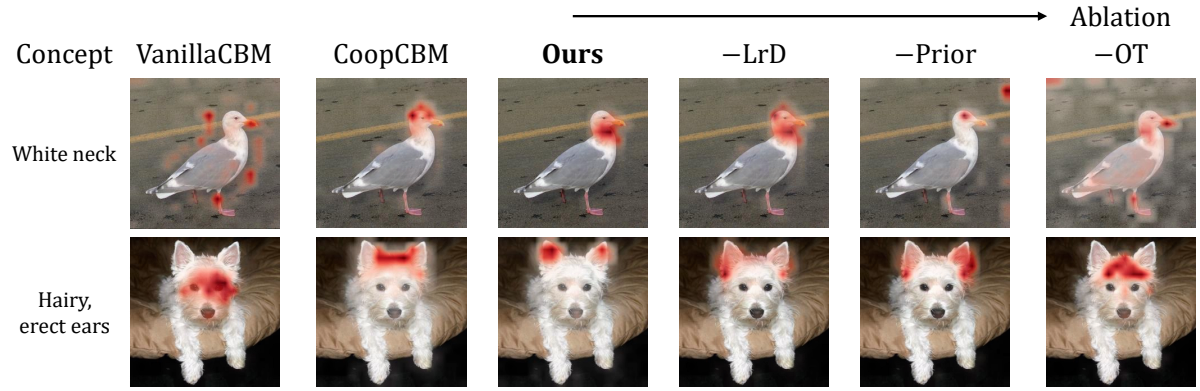


Figure 4. Qualitative results for the concept inversion. VanillaCBM and CoopCBM adopt Grad-CAM to provide a concept inversion heatmap, while our proposed DOT-CBM can generate the inversion mask on its own.

incorporation of OT distance, mentioned in Sec. 3.3, can slightly improve the classification accuracy and largely enhance the part detection performance. OT provides a better distance metric space for modeling the complex many-to-many correlation between concept features and patch features, compared to the direct dot product similarity used in the baseline model.

Effect of prior distribution modeling (Prior). The OT design above, setting the two prior distributions $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ as a uniform distribution. To mitigate the shortcut learning brought by high occurrence between concept and background and between concept and concept, we add the background penalty patch set prior and the co-occurrence penalty concept set prior, mentioned in Sec. 3.4. Comparing the second and third rows of Table. 3, it demonstrates that these two prior distributions can further improve the classification and interpretability.

Effect of local representation disentanglement (LrD). After integrating the local representation disentanglement into the baseline (the fourth row of Table. 3), the improvements are larger than a single incorporation of OT distance (the second row of Table. 3). The overall DOT-CBM, with a combination of the OT, Prior, and LrD, yields more improvements than individual modules alone.

Qualitative results. Fig. 4 has shown a qualitative comparison between our method and the baseline methods Vanilla-CBM and CoopCBM. It is observed that our method is able to more accurately locate the region of the concept in the original image compared to the baseline methods. Meanwhile, We also present qualitative results by sequentially ablating each component of our framework. When the local representation disentanglement is ablated, the model's localization accuracy decreases to some extent. Further removal of the priors leads to the emergence of part-background spurious correlations. When the OT design is removed, the model produces incorrect localization results.

5. Conclusion.

In this paper, we introduced the Disentangled Optimal Transport Concept Bottleneck Model (DOT-CBM) to address key limitations in existing Concept Bottleneck Models. While CBMs offer interpretable models by mapping image features to human-understandable concepts, they suffer from misalignment between the image and concept, leading to spurious correlations that hinder both interpretability and generalization. To overcome these challenges, DOT-CBM introduces a disentangled optimal transport framework that ensures fine-grained alignment between image patches and corresponding textual concepts, thereby providing a more robust and interpretable model. By employing a dual penalty system, we mitigate the overfitting problem caused by data bias. Empirical results demonstrate that DOT-CBM outperforms previous CBM variants in both classification accuracy and interpretability, offering an effective solution to the problem of visual-concept spurious correlation.

6. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant U21B2006; in part by Shaanxi Youth Innovation Team Project; in part by the Fundamental Research Funds for the Central Universities QTZX24003 and QTZX22160; in part by the 111 Project under Grant B18039; Hao Zhang acknowledges the support of NSFC (62301384); Excellent Young Scientists Fund (Overseas); Foundation of National Key Laboratory of Radar Signal Processing under Grant JKW202308. Zhengjue Wang acknowledges the support of NSFC (62301407).

References

[1] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 4

[2] Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5948–5955, 2023. 2

[3] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. 3

[4] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR, 2020. 3

[5] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014. 6

[6] Molnar Christoph. *Interpretable machine learning: A guide for making black box models explainable*. Leanpub, 2020. 1

[7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 4

[8] Barry De Ville. Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6):448–455, 2013. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[10] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in Neural Information Processing Systems*, 35:21400–21413, 2022. 2

[11] Mateo Espinosa Zarlenga, Katie Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[12] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28, 2015. 3

[13] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 6

[14] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011. 1

[15] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011. 7

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 7

[17] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020. 1, 2, 3, 6, 7

[18] Andrei V Konstantinov and Lev V Utkin. Incorporating expert rules into neural networks in the framework of concept-based learning. *arXiv preprint arXiv:2402.14726*, 2024. 2

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

[20] John Lee, Max Dabagia, Eva Dyer, and Christopher Rozell. Hierarchical optimal transport for multimodal distribution alignment. *Advances in neural information processing systems*, 32, 2019. 3

[21] Miaoge Li, Dongsheng Wang, Xinyang Liu, Zequn Zeng, Ruiying Lu, Bo Chen, and Mingyuan Zhou. Patchct: Aligning patch set and label set with conditional transport for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15348–15358, 2023. 3

[22] Tang Li, Mengmeng Ma, and Xi Peng. Deal: Disentangle and localize concept-level explanations for vlms. In *European Conference on Computer Vision*, pages 383–401. Springer, 2025. 2, 4

[23] Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Zujie Wen, and Xi Peng. Multi-granularity correspondence learning from long-term noisy videos. *arXiv preprint arXiv:2401.16702*, 2024. 3

[24] Jiawei Ma, Po-Yao Huang, Saining Xie, Shang-Wen Li, Luke Zettlemoyer, Shih-Fu Chang, Wen-Tau Yih, and Hu Xu. Mode: Clip data experts via clustering. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26354–26363, 2024. 2

[25] Shweta Mahajan, Teresa Botschen, Iryna Gurevych, and Stefan Roth. Joint wasserstein autoencoders for aligning multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[26] Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021. 2, 3, 6

[27] Nesta Midavaine, Gregory Hok Tjoan Go, Diego Canez, Ioana Simion, and Satchit Chatterji. [re] on the reproducibility of post-hoc concept bottleneck models. *Transactions on Machine Learning Research*. 2

[28] Mazda Moayeri, Phillip Pope, Yogesh Balaji, and Soheil Feizi. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6

[29] Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text2concept: Concept activation vectors directly from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3744–3749, 2023. 2, 4

[30] Keval Morabia, Jatin Arora, and Tara Vijaykumar. Attention-based joint detection of object and semantic part. *arXiv preprint arXiv:2007.02419*, 2020. 6

[31] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023. 2

[32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[34] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021. 4

[35] Naveen Raman, Mateo Espinosa Zarlenga, Juyeon Heo, and Mateja Jamnik. Do concept bottleneck models obey locality? In *XAI in Action: Past, Present, and Future Applications*, 2023. 3

[36] Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on artificial intelligence and statistics*, pages 849–858. PMLR, 2019. 3

[37] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[38] Andrei Semenov, Vladimir Ivanov, Aleksandr Beznosikov, and Alexander Gasnikov. Sparse concept bottleneck models: Gumbel tricks in contrastive learning. *arXiv preprint arXiv:2404.03323*, 2024. 2, 6, 7

[39] Ivaxi Sheth and Samira Ebrahimi Kahou. Auxiliary losses for learning generalizable concept-based models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 6, 7

[40] Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *International Conference on Machine Learning*, pages 31504–31520. PMLR, 2023. 2

[41] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6, 7

[42] Tiansheng Wen, Yifei Wang, Zequn Zeng, Zhong Peng, Yudi Su, Xinyang Liu, Bo Chen, Hongwei Liu, Stefanie Jegelka, and Chenyu You. Beyond matryoshka: Revisiting sparse coding for adaptive representation. *arXiv preprint arXiv:2503.01776*, 2025. 2

[43] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 6

[44] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems*, 33:21969–21980, 2020. 6

[45] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023. 2, 4, 6, 7

[46] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022. 2

[47] Yuan Zang, Tian Yun, Hao Tan, Trung Bui, and Chen Sun. Pre-trained vision-language models learn discoverable visual concepts. *arXiv preprint arXiv:2404.12652*, 2024. 2, 4, 6

[48] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *arXiv preprint arXiv:2209.09056*, 2022. 6, 7

[49] Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23465–23476, 2023. 2

[50] Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. Meacap: Memory-augmented zero-

shot image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14100–14110, 2024. 2

[51] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1143–1152, 2016. 6

[52] Peng Zhao and Zhi-Hua Zhou. Label distribution learning by optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 3