南京航空航天大學

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

Yan Xie1* Zequn Zeng1* Hao Zhang1† Yucheng Ding1 Yi Wang1 Zhengjue Wang2 Bo Chen1 Hongwei Liu1
1National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an, 710071, China
2State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, 710071, China

CVPR 2025

Introduction

普通CNN：

图片 → 卷积层 → 分类结果

CBM：

概念预测器 g 分类器 f

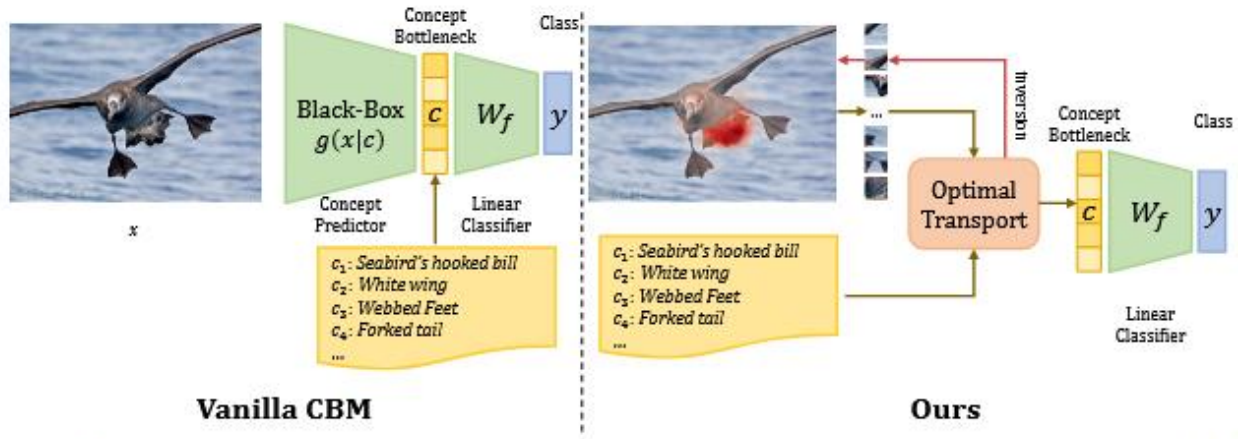
$$x \xrightarrow{g} c \xrightarrow{f} y$$

概念向量

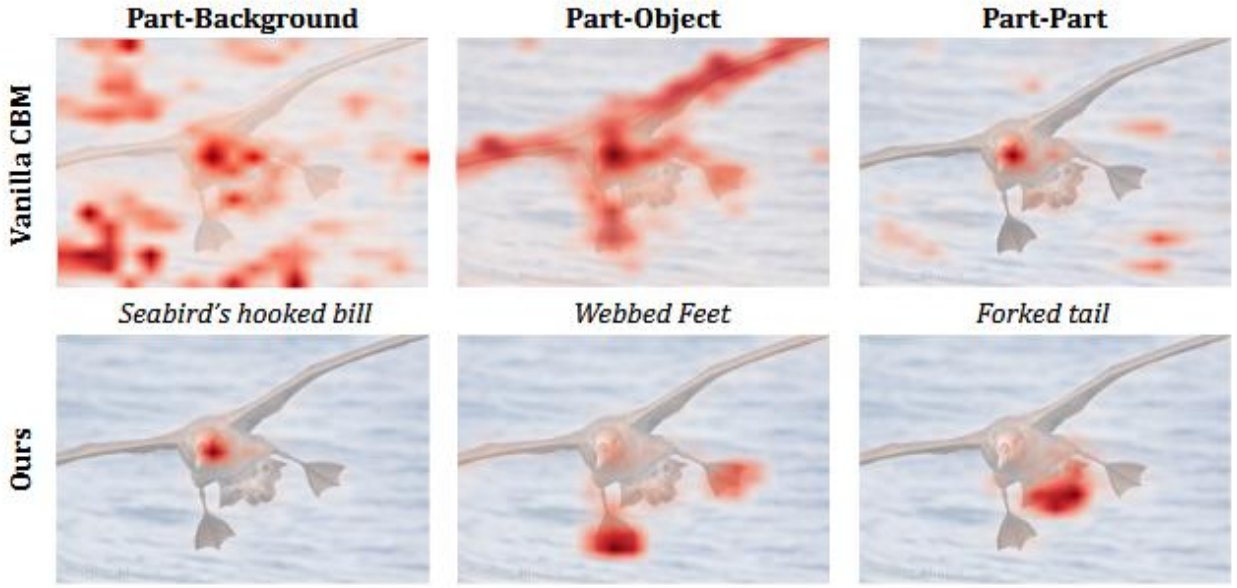
可解释性强

问题 1：概念标注成本极高

问题 2：概念–输入不对齐



(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.

Method

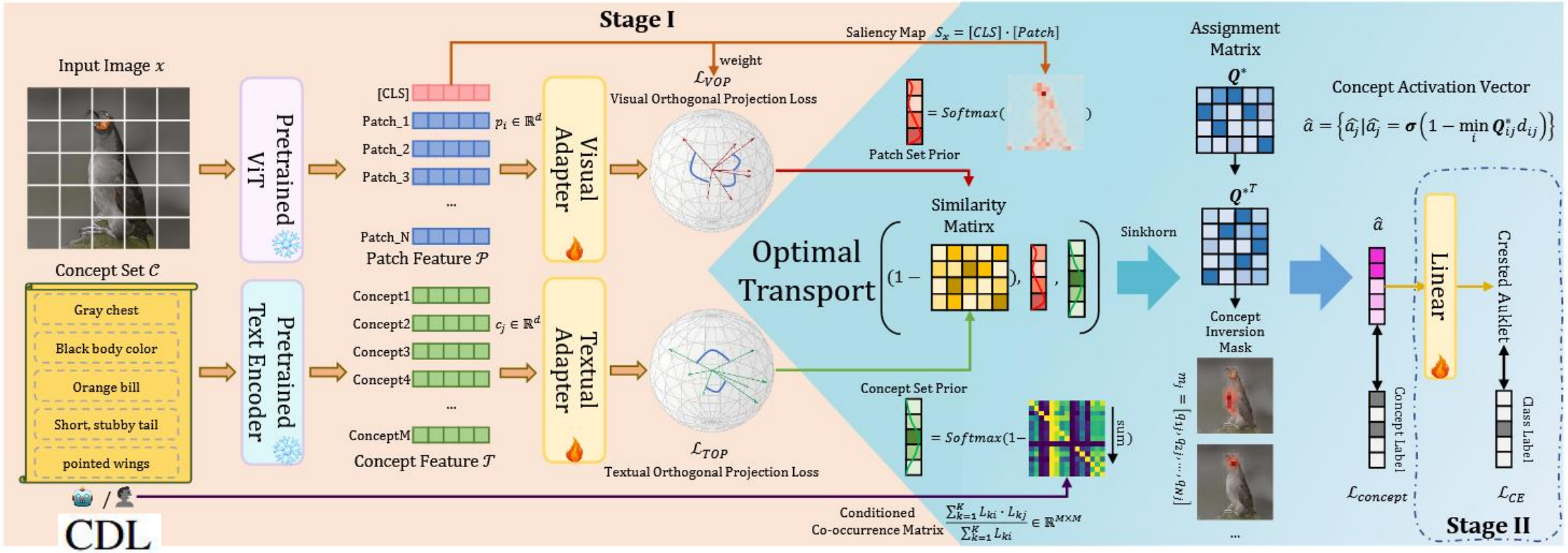
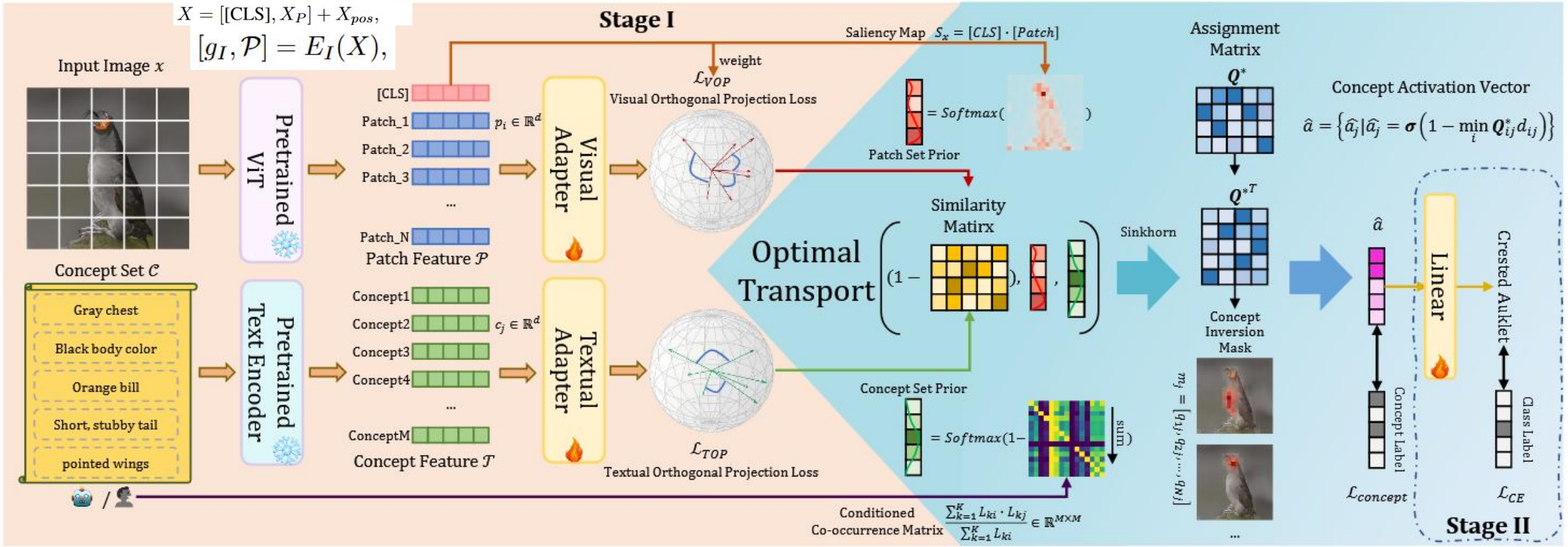


Figure 2. Overview of our proposed DOT-CBM. The overall CBM framework proceeds from left to right. In the first stage, the model transforms input images into concept activation vectors. In forward pass, local embeddings from a pre-trained Vision Transformer (ViT) and a text encoder are processed through 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).

Method Disentanglement on Local representations

南京航空航天大學

$$X_P = [I_{p1} \cdot W_p, ..., I_{pN} \cdot W_p],$$
$$X = [[\text{CLS}], X_P] + X_{pos},$$
$$[g_I, \mathcal{P}] = E_I(X),$$



$$S_I = g_I \cdot P,$$

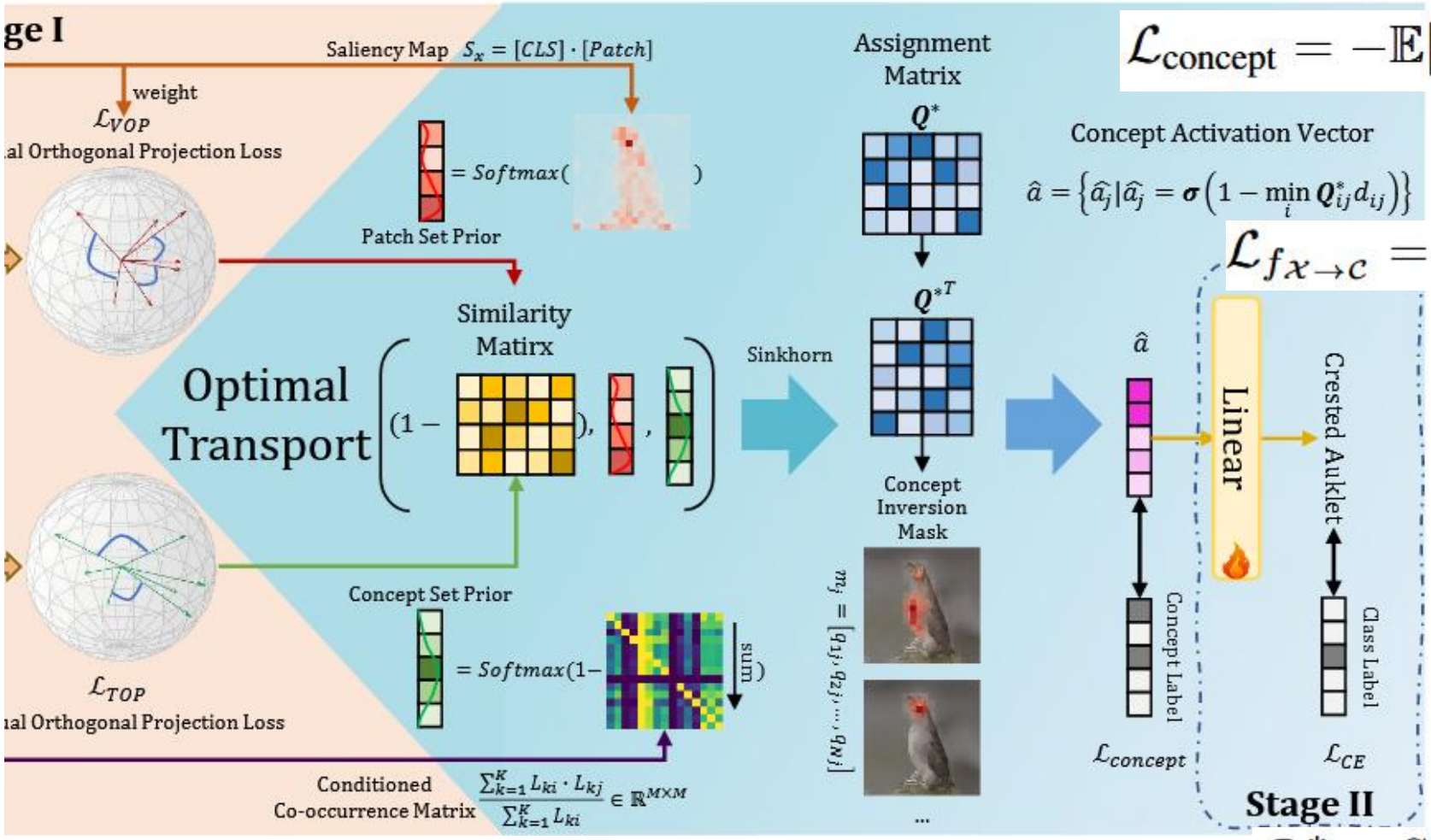
$$\langle A_v(p_i), A_v(p_j) \rangle = \frac{A_v(p_i) \cdot A_v(p_j)}{\|A_v(p_i)\| \|A_v(p_j)\|},$$

$$\mathcal{T} = \{t_j = E_T(c_j)\}_{j=1}^{M}$$

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

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

Method Patch-concept alignment via optimal transport



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

$$\mathcal{L}_{f_{\mathcal{X}\to\mathcal{C}}} = \lambda_1\mathcal{L}_{\text{VOP}} + \lambda_2\mathcal{L}_{\text{TOP}} + \lambda_3\mathcal{L}_{\text{concept}}$$

$$m_j = [q_{1j}, q_{2j}, \ldots, q_{Nj}]$$

Q = [qij]N×M dij = d(pi, tj) = (1 − <pi, tj>) ≥ 0

$$\boldsymbol{\Theta} = \sum_{i=1}^{N} \theta_i \delta_{p_i}, \quad \boldsymbol{\Gamma} = \sum_{j=1}^{M} \gamma_j \delta_{c_j}, \quad \min_{\mathbf{Q}\in\mathcal{Q}} \sum_{i,j} q_{ij} d_{ij} - \varepsilon H(\mathbf{Q})$$

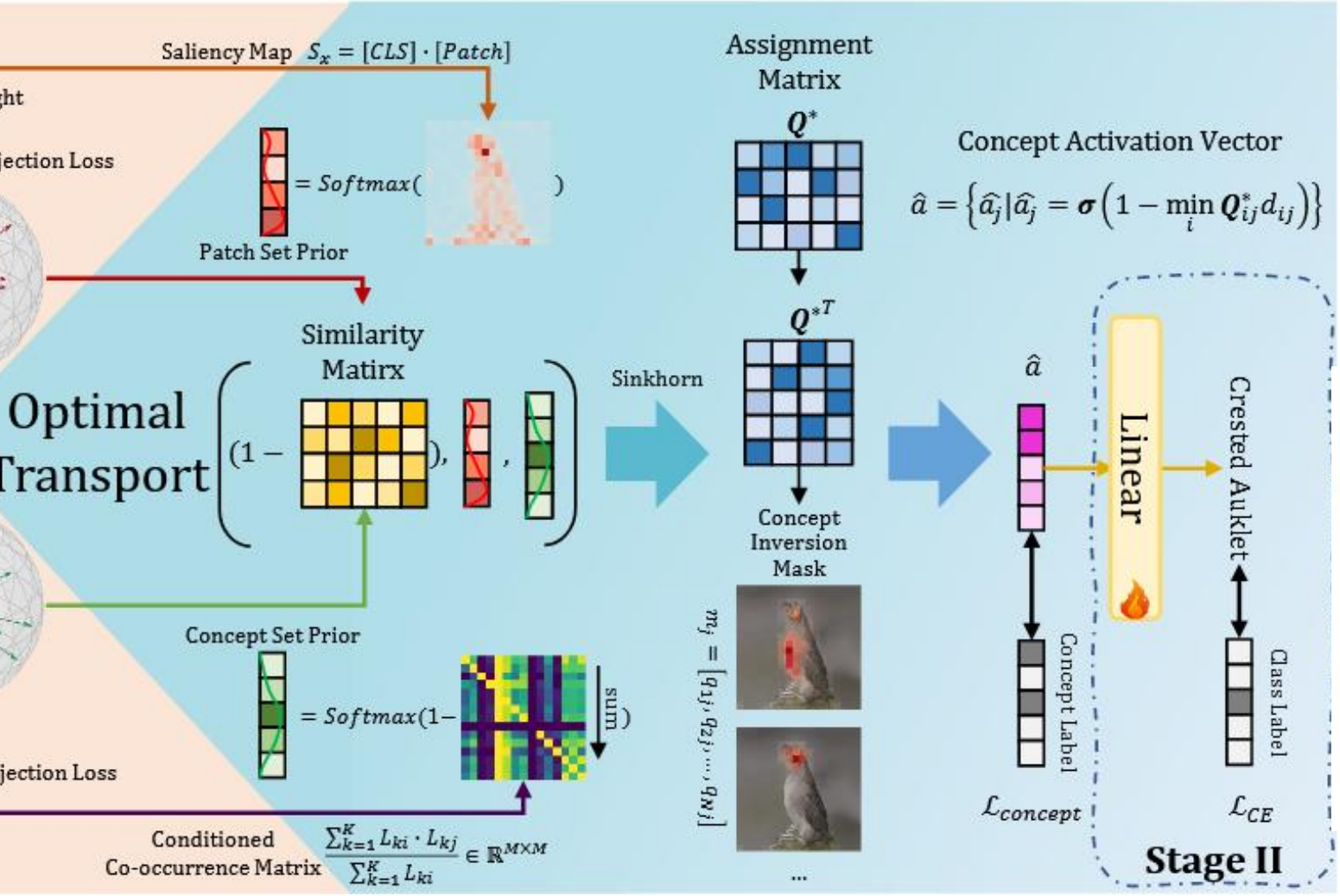
$$\text{s.t. } \mathcal{Q} = \{\mathbf{Q} \in \mathbb{R}_+^{N\times M} \mid \mathbf{Q}\mathbf{1}_M = \boldsymbol{\theta}, \mathbf{Q}^\top \mathbf{1}_N = \boldsymbol{\gamma}\}$$

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

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

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

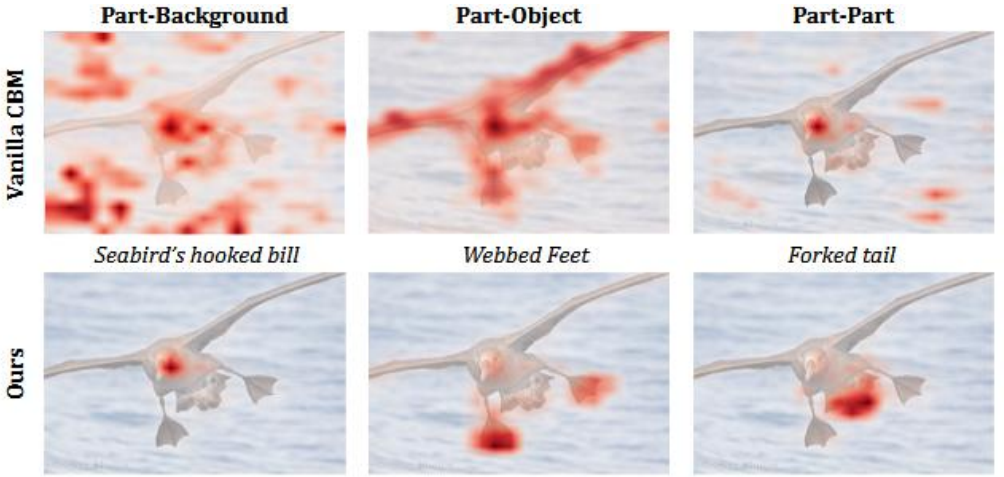
Method Prior distributions Modeling



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

$$S_I = g_I \cdot P,$$

$$\boldsymbol{\theta} = \mathrm{Softmax}(S_I), \textbf{Patch set prior } \boldsymbol{\theta}$$



(b) Inversion heatmap visualization.

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

Concept set prior γ

$$\boldsymbol{\gamma} = 1 - \mathrm{Softmax}(r_i)$$

class-concept label form a matrix $L = [l_{i,j}]_{K \times M}$

Experiment

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.

Experiment

OT	Prior	LrD	Classification Accuracy (↑) ImageNet	CUB	CIFAR100	AWA2	Part Detection $mIOU$ (↑) 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.

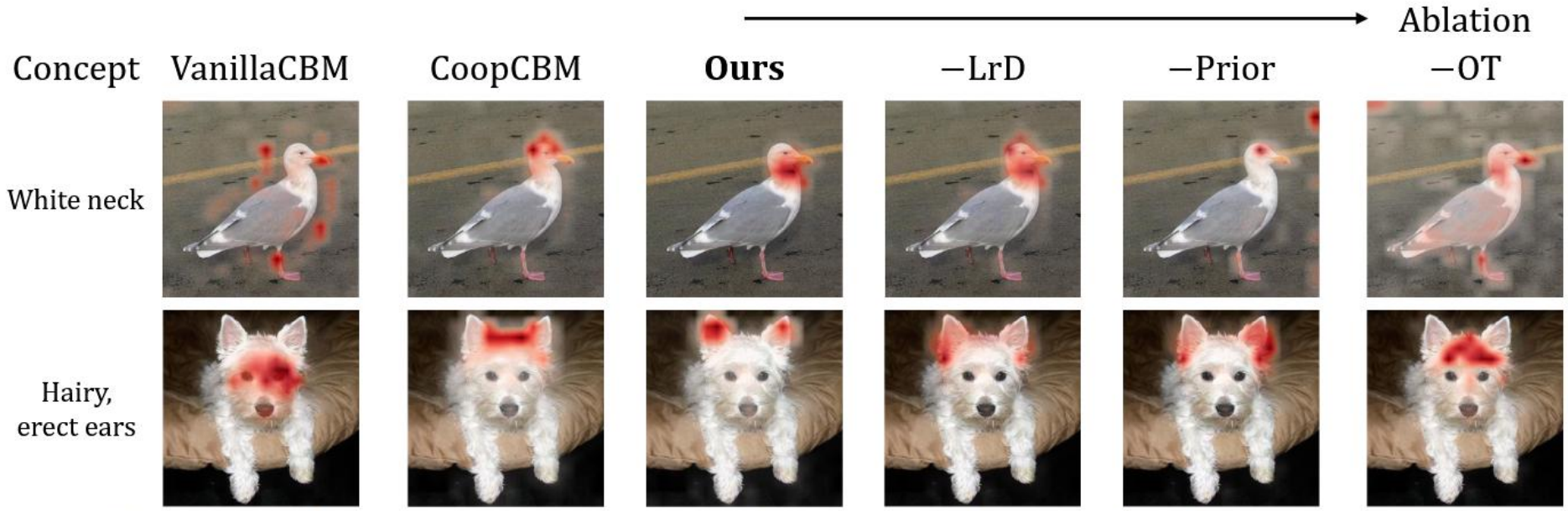


Figure 4. Qualitative results for the concept inversion. VanillaCBM and CoopCBM adopt Grad-CAM to provide a concept inversion heatmap, while our proposed DOT-CBM can generate the inversion mask on its own.

Thanks