Recover and Match: Open-Vocabulary Multi-Label Recognition through Knowledge-Constrained Optimal Transport
Supplementary Material

1. Additional Illustration on Method

1.1. Derivation of KCOT

This section gives a detailed derivation of the KCOT problem in Sec.4.3. The KCOT problem is formulated as:

$$\begin{aligned} &\min_{\boldsymbol{P}} \sum_{k=1}^{M}\sum_{i=1}^{N} \boldsymbol{P}_{ki}\boldsymbol{C}_{ki} - \lambda_1 H(\boldsymbol{P}) + \lambda_2 D_{KL}(\boldsymbol{P}||\widetilde{\boldsymbol{P}}), \\ &\text{s.t. } \boldsymbol{P}\mathbf{1}^N = \tilde{\boldsymbol{u}},\ \boldsymbol{P}^{\mathsf{T}}\mathbf{1}^M = \boldsymbol{v}. \end{aligned} \tag{1}$$

Based on the definition of KL Divergence and entropy, the minimization objective can be simplified as follows:

$$\begin{aligned} &\sum_{k=1}^{M}\sum_{i=1}^{N} \boldsymbol{P}_{ki}\boldsymbol{C}_{ki} - \lambda_1 H(\boldsymbol{P}) + \lambda_2 D_{KL}(\boldsymbol{P}||\widetilde{\boldsymbol{P}}), \\ =&\sum_{k=1}^{M}\sum_{i=1}^{N} \boldsymbol{P}_{ki}\boldsymbol{C}_{ki} - \lambda_1 H(\boldsymbol{P}) + \lambda_2 \sum_{k=1}^{M}\sum_{i=1}^{N}(\boldsymbol{P}_{ki}\log\frac{\boldsymbol{P}_{ki}}{\widetilde{\boldsymbol{P}}_{ki}}), \\ =&\sum_{k=1}^{M}\sum_{i=1}^{N} \boldsymbol{P}_{ki}\boldsymbol{C}_{ki} - \lambda_1 H(\boldsymbol{P}) + \lambda_2 \sum_{k=1}^{M}\sum_{i=1}^{N}(\boldsymbol{P}_{ki}\log\boldsymbol{P}_{ki}) \\ &- \lambda_2 \sum_{k=1}^{M}\sum_{i=1}^{N}(\boldsymbol{P}_{ki}\log\widetilde{\boldsymbol{P}}_{ki}), \\ =&\sum_{k=1}^{M}\sum_{i=1}^{N} \boldsymbol{P}_{ki}(\boldsymbol{C}_{ki} - \lambda_2\log\widetilde{\boldsymbol{P}}_{ki}) - (\lambda_1+\lambda_2)H(\boldsymbol{P}). \end{aligned} \tag{2}$$

Here, the cost is transformed into $\widetilde{\boldsymbol{C}}_{ki} = \boldsymbol{C}_{ki} - \lambda_2\log\widetilde{\boldsymbol{P}}_{ki}$ and the coefficient of entropy $H(\boldsymbol{P})$ is re-weighted into $\tilde{\lambda} = \lambda_1 + \lambda_2$. As a result, the KCOT problem is simplified to the following form:

$$\begin{aligned} &\min_{\boldsymbol{P}} \sum_{k=1}^{M}\sum_{i=1}^{N} \boldsymbol{P}_{ki}\widetilde{\boldsymbol{C}}_{ki} - \tilde{\lambda} H(\boldsymbol{P}), \\ &\text{s.t. } \boldsymbol{P}\mathbf{1}^N = \tilde{\boldsymbol{u}},\ \boldsymbol{P}^{\mathsf{T}}\mathbf{1}^M = \boldsymbol{v}. \end{aligned} \tag{3}$$

Eq. 3 is equivalent to an entropic OT problem, and can be solved by Sinkhorn algorithm [15] within several iterations. Compared to OT, our KCOT enhances performance without increasing the problem complexity. Moreover, the solution of KCOT does not involve any gradient propagation or updates on the model parameters, which is highly efficient and transferable.

Algorithm 1 Knowledge-Constrained Optimal Transport

Input:
Visual set $\boldsymbol{X}^{\psi}$ and label set $\boldsymbol{T}$;
Teacher plan $\widetilde{\boldsymbol{P}}$ and parameters $\lambda_1, \lambda_2, t_{max}$;
Output:
Optimal transport plan $\boldsymbol{P}^*$;
Calculate the cost matrix $\boldsymbol{C}$ with $\{\boldsymbol{X}^{\psi}, \boldsymbol{T}\}$ in Eq. 7;
Set unbalanced marginal distribution $\tilde{\boldsymbol{u}}$ for visual set according to Eq. 8. Set balanced marginal distribution $\boldsymbol{v} = \mathbf{1}^N/N$ for label set;
Calculate the transformed cost matrix $\widetilde{\boldsymbol{C}}$ in Eq. 11, and calculate the parameter $\tilde{\lambda} = \lambda_1 + \lambda_2$;
Initialize $\boldsymbol{b}^0 = \mathbf{1}$, iteration $t = 0$;
while $t \le t_{max}$ and not converge **do**
  $\boldsymbol{a}^t = \tilde{\boldsymbol{u}}/((\exp(-\widetilde{\boldsymbol{C}}/\tilde{\lambda})\boldsymbol{b}^{t-1})$;
  $\boldsymbol{b}^t = \boldsymbol{v}/((\exp(-\widetilde{\boldsymbol{C}}/\tilde{\lambda})^{\mathsf{T}}\boldsymbol{a}^t)$;
end while
Transport plan $\boldsymbol{P}^* = \text{diag}(\boldsymbol{a}^t)\exp(-\widetilde{\boldsymbol{C}}/\tilde{\lambda})\text{diag}(\boldsymbol{b}^t)$;
return $\boldsymbol{P}^*$;

1.2. Pseudo Code

The procedure of KCOT is summarized in Alg. 1. In our studies, $t_{max} = 100$ ensures convergence in most cases, which exhibits great efficiency.

1.3. Discussion on SAA (How does SAA work?)

Suppose the input sequences are projected into $\{\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\}$. The attention score a_{ij} is determined by $a_{ij} = \boldsymbol{Q}_i\boldsymbol{K}_j^{\mathsf{T}}$. Due to the image-level objective, the pre-trained attention always focuses on *dominant patch* with large attention scores. Our SAA replaces the calculation with $\tilde{a}_{ij} = \boldsymbol{V}_i\boldsymbol{V}_j^{\mathsf{T}}$. Since the angle between V_i and V_i is 0, with similar magnitudes, $\tilde{a}_{ii}$ is always greater than $\tilde{a}_{ij}(j \neq i)$, producing diagonal-style attention maps, which ensures better focus on itself.

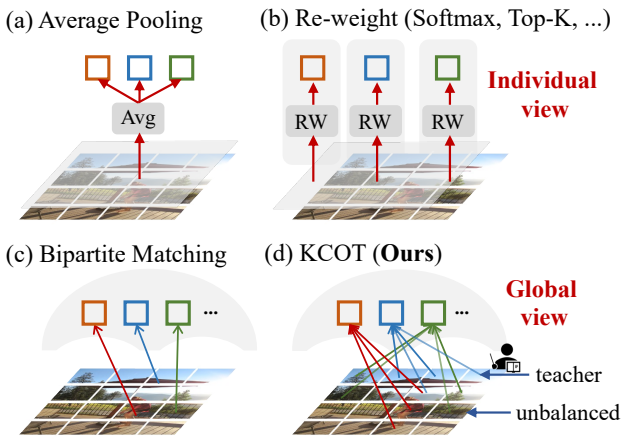

Figure 1. Matching comparisons. (a) Simple average pooling in [4, 19] and (b) independent re-weighting in [5, 6, 9, 11, 16] overlook the matching property, assigning features to each label individually. (c) Bipartite matching [8] finds one-to-one matching. (d) Our KCOT performs one-to-many matching under the guidance of teacher plan and unbalanced marginal. KCOT finds matching from a global view, suppressing meaningless matching.

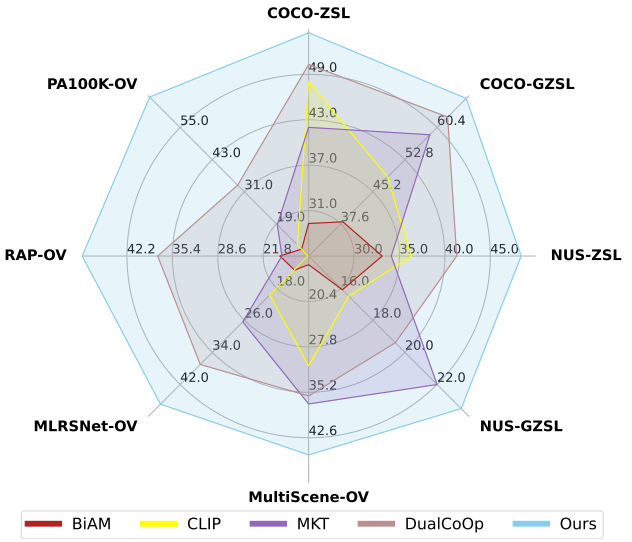

Figure 2. Performance comparisons to state-of-the-art methods on six datasets across different domains.

1.4. Discussion on the matching framework

As shown in Figure 1, the region aggregations of existing methods can be summarized into two types: (a) Average pooling which distributes equal weights to all regions, involving predictions from most irrelevant areas. (b) Independent re-weighting, which can be categorized into two clusters: (i) Softmax re-weighting in [5, 16], which normalizes the image-text similarities for each label individually. (ii) Attention re-weighting, which is a core strategy in traditional ML-ZSL methods [6, 12]. Recently, some works [9, 11] have reused the ideas, which perform cross-attention between label embeddings and regional features. However, the "matching" claimed in [9, 11] is actually an independent re-weighting strategy, since the attention scores are calculated for each label individually. The similarities between regions and all labels are *not* jointly compared, as a result, it always emphasizes particular regions for each label including those non-GTs, resulting in noisy and error-prone region aggregations.

In this work, we reformulate the problem from the *set matching* perspective, where **the matching weights between different elements of the two sets are jointly determined**. As shown in Figure 1 (c), a straightforward way is to find a one-to-one matching (*i.e.*, bipartite matching) between image regions and labels, which can be resolved by Hungarian algorithm as in [1]. However, the excessively sparse matching results in unsatisfactory performance. In contrast, we introduce optimal transport theory and formulate the KCOT problem, which implicitly suppresses matching to irrelevant labels by jointly comparing similarities of all labels, largely enhancing open-vocabulary performance.

1.5. Detailed Architecture of TSS

The input features $\boldsymbol{X}_l \in \mathbb{R}^{M\times d}$ are first reshaped into feature maps $\boldsymbol{X}_l' \in \mathbb{R}^{H'\times W'\times d}$, where $H' \times W' = d$. To aggregate local semantics from different receptive fields, the kernel sizes of two convolutions are set as 3×3 and 1×1, respectively. To reduce the number of parameters and network complexity, the cross-attention between image and text features is parameter-free [3]. The outputs from two streams are concatenated and pooled. Then a depth-wise convolution is applied and the sigmoid is performed to get the spatial masks. Above all, TSS recovers the local semantics through the lens of convolutions and integrates text features to highlight salient image regions.

1.6. Discussion on the MMC loss

We discuss the difference between our MMC loss and the commonly used ASL [14] and ranking loss [2]. For a mini-batch images $\{\boldsymbol{I}_b\}_{b=1}^{B}$, the predicted logit of the i^th label for image $\boldsymbol{I}_b$ is denoted as $s(b,i)$. Suppose $\mathcal{P}$ denotes the positive label set for a **single image**, and $\mathcal{P}_B$ is the positive set for the **mini-batch images**. ASL is defined as:

$$\mathcal{L}_{ASL} = \frac{1}{B}\frac{1}{C}\sum_{b=1}^{B}\sum_{i=1}^{C}\begin{cases}(1-s(b,i))^{\gamma^+}\log(s(b,i)), & i\in\mathcal{P},\\ (s(b,i))^{\gamma^-}\log(1-s(b,i)), & i\notin\mathcal{P},\end{cases} \tag{4}$$

where γ^+ and γ^- are two hyper-parameters to enable the asymmetric focus for positive and negative labels. It is worth noting that ASL treats each label individually. Although this can perform well in closed-set multi-label recognition, it is hard to generalize well to unseen classes.

Datasets	Domain	Training	GZSL Testing	Labels (seen/unseen)
NUS-WIDE	Natural	143K	59K	925 / 81
MS-COCO	Natural	44K	6K	48 / 17
RAP-OV	PAR	25K	8K	35 / 16
PA100K-OV	PAR	59K	10K	16 / 10
MultiScene-OV	RS	24K	7K	20 / 16
MLRSNet-OV	RS	11K	21K	40 / 20

Table 1. Dataset statistics of the open-vocabulary benchmarks. "PAR" denotes Pedestrian Attribute Recognition. "RS" denotes Remote Sensing image classification.

Method	# Prompts	NUS		COCO	
		ZSL	GZSL	ZSL	GZSL
w/o prompt	-	41.2	20.3	52.7	66.7
Deep prompt	2	46.4	22.9	**54.5**	67.1
Deep prompt	4	**48.4**	**23.5**	**54.5**	67.3
Deep prompt	8	47.9	23.4	54.1	**67.7**
Deep prompt	16	47.9	23.3	53.6	67.5
Deep prompt	32	45.0	22.7	53.1	67.5

Table 2. Analysis on the deep label prompting.

Alternatively, contrastive learning is a promising paradigm to improve zero-shot generalization. Ranking loss is a representative approach, which is defined as:

$$\mathcal{L}_{rank} = \sum_{b=1}^{B} \sum_{\substack{i \in \mathcal{P} \\ j \notin \mathcal{P}}} \max(1 + s(b,j) - s(b,i), 0). \quad (5)$$

Different from ASL, ranking loss jointly compares all labels, learning the ordering relationships among these categories. However, it still encounters two problems: (1) the contrastive learning is restricted to a single sample, which hinders the generalization. (2) the performance on large-scale categories is limited (as shown in Table 3), possibly due to the unstable gradients caused by large values in loss.

In contrast, InfoNCE [13] exhibits more stable optimization due to the normalized outputs. Inspired by Supervised Contrastive Learning (SCL) [7], InfoNCE can be extended to multi-label scenarios by defining multiple positive pairs:

$$\mathcal{L} = -\frac{1}{|\mathcal{P}|} \sum_{b=1}^{B} \log \frac{\exp(s(b,i)/\tau')}{\sum_{j=1}^{N} \exp(s(b,j)/\tau')}. \quad (6)$$

In this work, we further extend the negative references into the mini-batch, yielding MMC loss:

$$\mathcal{L}_{\text{MMC}} = -\frac{1}{|\mathcal{P}_B|} \sum_{(b,i) \in \mathcal{P}_B} \log \frac{\exp(s(b,i)/\tau')}{\sum_{b'=1}^{B} \sum_{j=1}^{N} \exp(s(b',j)/\tau')}, \quad (7)$$

Compared to ASL, MMC loss inherits the advantages of contrastive learning, enabling discriminative representations. Compared to ranking loss and InfoNCE, MMC loss allows more diverse entries to be negative references, facilitating generalizable alignments.

Loss Type	NUS		COCO	
	ZSL	GZSL	ZSL	GZSL
ASL	43.2	22.4	53.3	66.9
Ranking	41.3	20.8	53.5	66.4
MMC (w/o batch)	45.6	**24.2**	53.8	**67.7**
MMC	**48.4**	23.5	**54.5**	67.3

Table 3. Additional ablations on MMC loss.

Method	NUS		COCO	
	ZSL	GZSL	ZSL	GZSL
RAM (w/o SAA)	45.2	22.5	52.9	66.8
RAM (w/o TSS)	47.6	23.2	53.6	67.0
RAM	**48.4**	**23.5**	**54.5**	**67.3**

Table 4. Additional ablations on LLA.

2. Datasets and Implementation Details

Construction of OV benchmarks. For NUS-WIDE and MS-COCO, we follow previous works to sample the seen and unseen classes, which results in 925/81 seen and unseen classes for NUS-WIDE, 48/17 seen and unseen classes for MS-COCO.

For PAR benchmarks (*i.e.*, RAPv1 and PA100K), we sample over 30% classes as unseen based on the frequency. To ensure the quantity of training samples, we select the most frequent classes as seen labels, which results in 35/16 seen and unseen classes for RAPv1, 16/10 seen and unseen classes for PA100K.

For RS benchmarks (*i.e.*, MultiScene and MLRSNet), we randomly sample over 30% classes as unseen, which results in 20/16 seen and unseen classes for MultiScene, 40/20 seen and unseen classes for MLRSNet.

Dataset statistics. As summarized in Table 1, the six datasets contain challenges from distinct domains. For RAPv1, PA100K and MLRSNet, we take the official test set for evaluation. For MultiScene dataset, we take the official MultiScene-Clean subset for evaluation which contains 7K manually-annotated images.

Implementation details. We apply several augmentation strategies to training images, including random crop, random flip, gaussian blur and random erasing [17]. During testing, we only perform resize operation. Learning rate is set as 5e-6 for natural images and 5e-5 for other domains using AdamW [10] optimizer, and decays with cosine policy. SGD optimizer with learning rate of 1e-3 is set for all learnable prompts following a common practice [16, 18]. LLA is applied in the last few layers and visual prompts are integrated to modulate global feature. On RS and PAR, λ_2 is set as 0.01 to neutralize the effect of frozen knowledge.

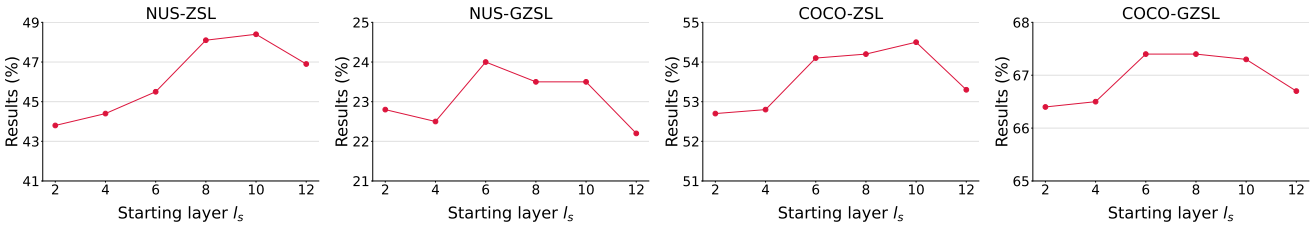

Figure 3. Analysis on the starting layer l_s. F1@3 is reported on NUS-WIDE and MS-COCO. $l_s = 10$ is a better trade-off.

Method	Train Speed (samples/s)	Test Speed (samples/s)	Train Mem. (Byte)	Test Mem. (Byte)	ZSL	GZSL
Baseline	143.5	101.9	5.9G	3.5G	40.6	19.4
Baseline + LLA	103.2	95.7	7.2G	3.8G	41.8	22.0
Δ (by LLA)	**↓40.3**	**↓6.2 (6.1%)**	**↑1.3G**	**↑0.3G (8.6%)**	**↑1.2**	**↑2.6**
Baseline + LLA + KCOT	90.3	94.5	8.5G	3.8G	48.4	23.5
Δ (by KCOT)	**↓12.9**	**↓1.2 (1.3%)**	**↑1.3G**	**↑0.0G (0.0%)**	**↑6.6**	**↑1.5**

Table 5. Efficiency analysis on the proposed LLA and KCOT. "Mem." is short for GPU memory usage. All statistics are obtained with a batch size of 32 on NUS-WIDE.

3. Additional Results

3.1. Analysis on the starting layer l_s

As shown in Figure 3, integrating LLA and textual prompts in early stages (*e.g.*, $l_s < 6$) leads to degraded performance, possibly due to overfitting. $l_s = 6$ achieves the best results on GZSL task while $l_s = 10$ exhibits better performance on ZSL. Overall, $l_s = 10$ is a better trade-off and introduces fewer parameters and computational overhead.

3.2. Analysis on the deep label prompting Q_l

As shown in Table 2, introducing textual prompts greatly improves performance. A limited number of prompts (*e.g.*, 2, 4 and 8) achieves great performance on both ZSL and GZSL. Increasing the number of prompts (*e.g.*, 16 and 32) does not bring further improvements.

3.3. Additional ablations on the LLA

As shown in Table 4, we remove only SAA and TSS, respectively. Notably, both SAA and TSS bring notable improvements and SAA is more important (*e.g.*, with margins of 3.2% and 1.6% on NUS-ZSL and COCO-ZSL, respectively). Note that SAA is parameter-free, which can be seamlessly integrated into more diverse tasks as an effective and efficient recovery.

3.4. Additional ablations on the MMC loss

In Table 3, we provide additional results on both ZSL and GZSL tasks. Notably, the proposed MMC loss achieves impressive improvements on NUS-WIDE, while the improvements on MS-COCO are less significant. The reason is that: with a larger number of categories, visually or semantically similar categories are more likely to overlap in the embedding space. Compared to ASL, contrastive learning is more effective in handling fine-grained distinctions by explicitly pushing different labels apart in the embedding space. Besides, removing batch operation in MMC loss yields better GZSL performance but significantly inferior ZSL performance. We suggest tailoring the choice to specific task, as batch operations are easy to implement.

3.5. Analysis on efficiency

As presented in Table 5, LLA brings certain computational overhead during training while remains lightweight during inference. KCOT delivers remarkable performance gains with almost no extra overhead during inference (*i.e.*, only 1.3% speed drop and nearly 0.0% memory increase), which verifies the efficiency of our proposed LLA and KCOT.

3.6. Additional visualizations

Unbalanced marginal $\tilde{\boldsymbol{u}}$. In Figure 4, we visualize vector $\tilde{\boldsymbol{u}} \in \mathbb{R}^M$ in LPD by reshaping it into $\mathbb{R}^{H' \times W'}$ and resizing to the image size. Note that this does not directly represent the weights of regions, but instead serves as a marginal constraint to the matching. LPD successfully emphasizes foreground areas. Notably, it avoids fully *binarizing* the target regions, allowing most regions to remain partially highlighted, which ensures generalization to unseen classes.

Visualizations of matching results (individual view). As shown in Figure 5, we visualize the matching results $\boldsymbol{P}^* \in \mathbb{R}^{M \times C}$ w.r.t. each label (*i.e.*, $\mathbb{R}^M$, individual view). We first reshape it into $\mathbb{R}^{H' \times W'}$ and resize to the image size. Our method generates precise region-to-label matching under different circumstances, *e.g.*, matching *clock* in both indoor and outdoor, matching *banana* of whole and sliced ones. Notably, for a given category, our approach is capable of matching all corresponding objects in the image. For

instance, it can find all *elephants* when there are multiple in the scene, which is similar for *banana*, *car* and *bicycle*.

Visualizations of matching results (global view). In Figure 6, we visualize the matching results $\boldsymbol{P}^* \in \mathbb{R}^{M \times C}$ w.r.t. all labels (*i.e.*, global view). We compare the proposed KCOT and independent re-weighting, which is widely used in previous works. Re-weighting distributes high-response regions to most labels, which we refer as "extraneous high-response" phenomenon. The reason is that it computes region weights for each label *separately*, inevitably producing some regions with higher weights for every label. In contrast, our KCOT successfully focus on the matching to target labels, facilitating precise and robust predictions.

3.7. Performance comparison

As shown in Figure 2, our method surpasses state-of-the-art methods on six diverse datasets. Notably, previous methods achieve limited performance in specialized domains such as PAR (*i.e.*, PA100K-OV and RAP-OV) and RS (*i.e.*, MultiScene-OV and MLRSNet-OV), while our method exhibits robust performance across different domains.

Figure 4. Visualizations of unbalanced marginal $\tilde{\boldsymbol{u}}$. Brighter color means higher weight. Best viewed in color.

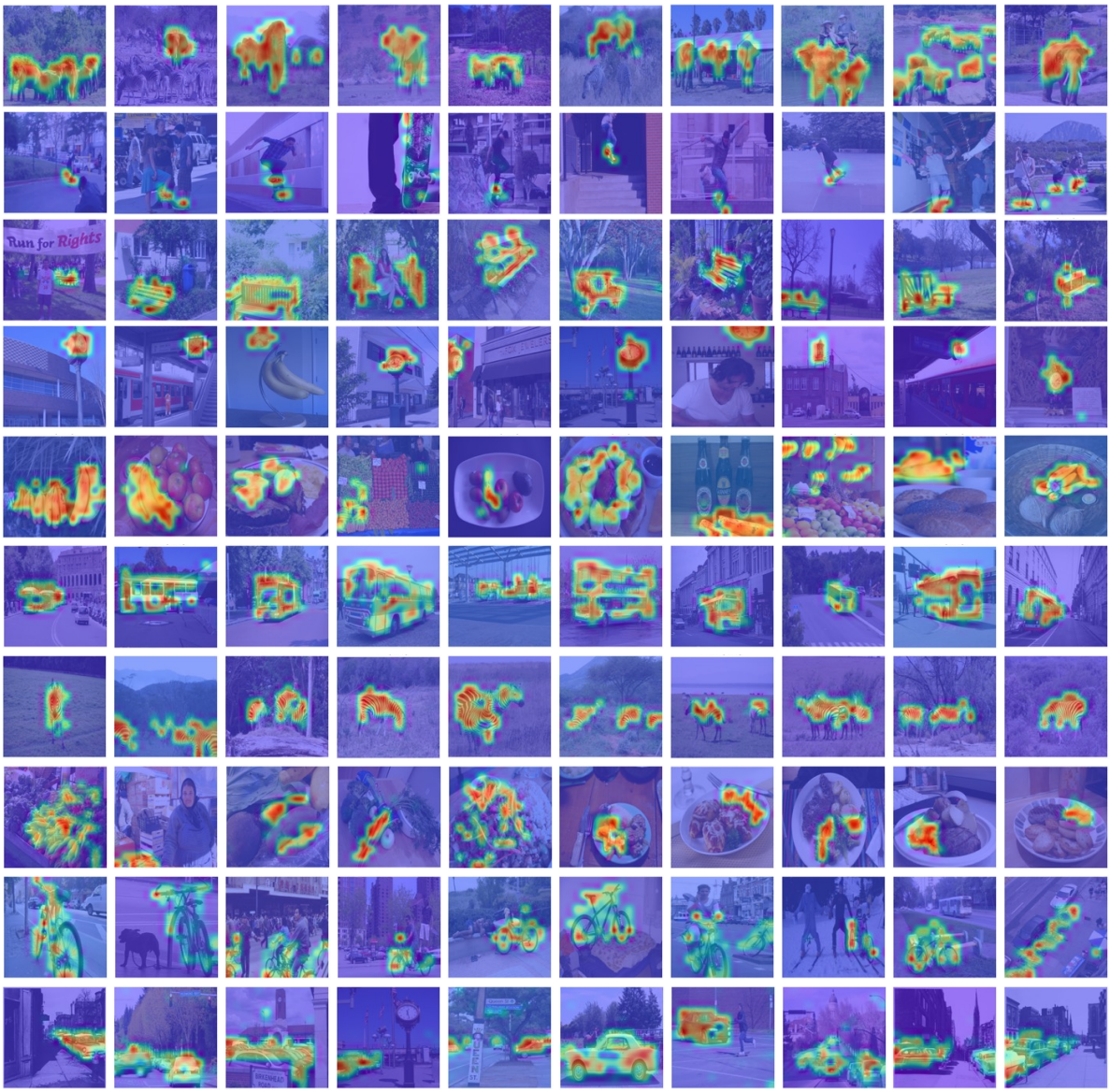

Figure 5. Visualizations of matching results (**individual view**). Each row is the matching weight for one label. From top to bottom are "elephant", "skateboard", "bench", "clock", "banana", "bus", "zebra", "carrot", "bicycle" and "car", respectively.

Figure 6. Visualizations of matching results (**global view**). The x-axis denotes the regions, and the y-axis corresponds to candidate labels. Ground-truth labels are marked in red boxes. Brighter color means higher matching weight.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[2] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013. 2

[3] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 746–754, 2023. 2

[4] Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Xiujun Shu, Bo Ren, and Shu-Tao Xia. Open-vocabulary multi-label classification via multi-modal knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 808–816, 2023. 2

[5] Ping Hu, Ximeng Sun, Stan Sclaroff, and Kate Saenko. Dualcoop++: Fast and effective adaptation to multi-label recognition with limited annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[6] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8776–8786, 2020. 2

[7] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3

[8] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 2

[9] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021. 2

[10] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[11] Leilei Ma, Hongxing Xie, Lei Wang, Yanping Fu, Dengdi Sun, and Haifeng Zhao. Text-region matching for multi-label image recognition with missing labels. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6133–6142, 2024. 2

[12] Sanath Narayan, Akshita Gupta, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Mubarak Shah. Discriminative region-based multi-label zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8731–8740, 2021. 2

[13] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[14] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 82–91, 2021. 2

[15] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 1

[16] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022. 2, 3

[17] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13001–13008, 2020. 3

[18] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

[19] Xuelin Zhu, Jian Liu, Dongqi Tang, Jiawei Ge, Weijia Liu, Bo Liu, and Jiuxin Cao. Query-based knowledge sharing for open-vocabulary multi-label classification. *arXiv preprint arXiv:2401.01181*, 2024. 2