Correlative and Discriminative Label Grouping for Multi-Label Visual Prompt Tuning

Supplementary Material

1. Supplementary Method

1.1. Divide All Classes into Subsets

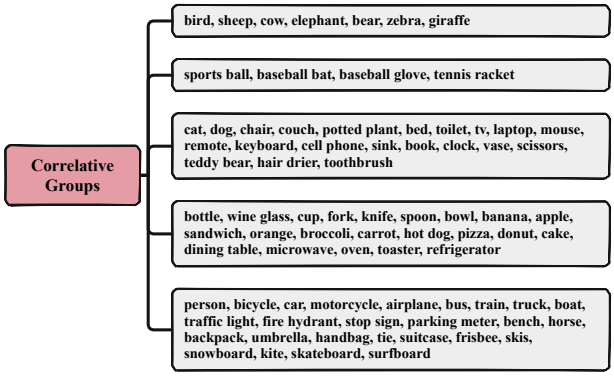
In this section, we present a grouping strategy for multi-label image classification (MLC), which divides the labels into several subgroups. Each subgroup is dedicated to addressing different label relationships. And, a similar method has been shown in BootMLC [16]. Each of these simpler subtasks is processed individually and in parallel, with the correlations among the labels being modeled within each subtask. These are then integrated to formulate a comprehensive solution to the original task. In our setting, we decompose the modeling of label correlation into co-occurrence and dis-occurrence. Concretely, we construct a co-occurrence graph $\mathcal{G}^+$ and a dis-occurrence graph $\mathcal{G}^-$ to encode the correlative representations between labels and the discriminative representations of each label. Firstly, we count the co-occurrence of label pairs to obtain the co-occurrence matrix $\mathbf{S} \in \mathbb{R}^{K\times K}$, and $\mathbf{S}_{i,j}$ represents the probability of occurrence of label $\boldsymbol{y}_j$ when label $\boldsymbol{y}_i$ is present. Subsequently, a smoothing operation and a symmetrization are employed to derive the affinity matrix as follows:

$$\mathbf{M} = \begin{cases} \mathbf{M}^+ = (\sqrt[\tau]{\mathbf{S}} + \sqrt[\tau]{\mathbf{S}}^\top)\,/\,2, & \mathcal{G} = \mathcal{G}^+ \\ \mathbf{M}^- = \mathbf{I} - (\sqrt[\tau]{\mathbf{S}} + \sqrt[\tau]{\mathbf{S}}^\top)\,/\,2, & \mathcal{G} = \mathcal{G}^- \end{cases}$$

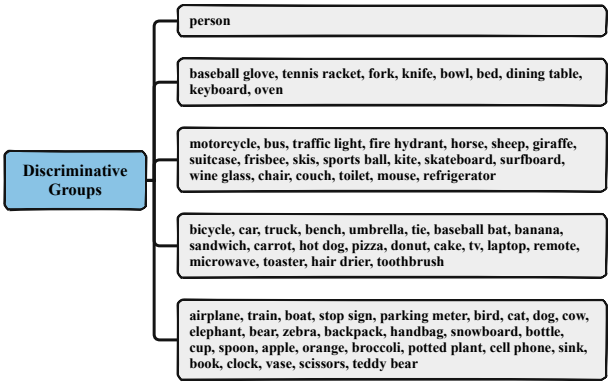
where τ is a positive hyper-parameter to adjust the distribution of co-occurrence matrix $\mathbf{S}$, and the $\sqrt[\tau]{\mathbf{S}}^\top$ denotes its transpose. The symmetrization ensures an undirected graph, with bidirectional connection strengths. The affinity matrix $\mathbf{M}^+$ and $\mathbf{M}^-$ are utilized to encode the co-occurrence relationship and dis-occurrence between categories respectively. Then we leverage them to generate the Laplacian matrix and treat the decomposition problem as a spectral clustering [14] problem as follows:

$$\hat{\mathbf{F}} \leftarrow \arg\min_{F} \mathrm{Trace}(\mathbf{F}^T\mathbf{L}_{\mathrm{syn}}\mathbf{F}), \quad \text{s.t. } \mathbf{F}^T\mathbf{F} = \mathbf{I},$$

where $\mathbf{L}_{\mathrm{syn}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{M}\mathbf{D}^{-\frac{1}{2}}$ represents the normalized Laplacian matrix and $\mathbf{D}$ is the degree matrix of graph $\mathcal{G}$. $\mathbf{F}$ is the learned graph embedding of vertices (categories), and $\hat{\mathbf{F}}$ indicates the top-k minimum eigenvectors of $\mathbf{L}_{\mathrm{syn}}$. We cluster the $\hat{\mathbf{F}}$ via the k-means algorithm into clusters $\{\mathcal{C}_t\}_{t=1}^T$. Ultimately, the original task can be decomposed into T sub-tasks $\{\mathcal{T}_t\}_{t=1}^T$ according to the clustered class subset. Correspondingly, sub-tasks $\{\mathcal{T}_t^+\}_{t=1}^T$ derived from the graph $\mathcal{G}^+$, acts as a guide for the model to learn the shared representations under co-occurrence relationship, whereas sub-tasks $\{\mathcal{T}_t^-\}_{t=1}^T$ generated from graph $\mathcal{G}^-$ promotes the model to focus on learning discriminative representations for each class by masking the correlations among labels. In Fig. 1, we present an example of grouping classes on COCO.



(a) correlative groups on COCO.



(b) discriminative groups on COCO.

Figure 1. An example of grouping all classes in both co-occurrence and dis-occurrence graphs within the COCO, with each graph divided into 5 groups.

1.2. Optimizing Objective in MLC

Most MLC methods often use Binary Cross-Entropy (BCE) as the loss function, which converts the problem into several binary classification tasks:

$$\mathcal{L}_{\mathrm{BCE}}(f(\boldsymbol{x}), \boldsymbol{y}) = \sum_{k=1}^{L} [y_k \ell_1(f_k(\boldsymbol{x})) + (1-y_k)\ell_0(f_k(\boldsymbol{x}))].$$

Here, $\ell_1(f_k) = -(1-f_k)\log(f_k)$ and $\ell_0(f_k) = -f_k\log(1-f_k)$ represent the losses calculated on positive and negative labels. To simplify the notation for clarity, we let $f(\boldsymbol{x})$ represent the probability distribution of example $\boldsymbol{x}$ and $f_k(\boldsymbol{x})$ denote the probability of the k-th class.

In MLC, the imbalance between positive and negative classes in each instance is a common issue. To address this, we adopt the asymmetric loss (ASL) [12] as the multi-label classification loss, as suggested by prior researchs [9, 10, 15]. ASL dynamically down-weights the easy negative samples and directs the optimization process towards the positive samples:

$$\mathcal{L}_{\text{ASL}}(f(\boldsymbol{x}), \boldsymbol{y}) = \sum_{k=1}^{K} [y_k \ell_1(f_k(\boldsymbol{x})) + (1-y_k)\ell_0(f_k(\boldsymbol{x}))],$$
$$\text{s.t.} \quad \ell_1(f_k) = -(1-f_k)^{\lambda_1} \log(f_k)\,,$$
$$\ell_0(f_k) = -(f_k)^{\lambda_0} \log(1-f_k)\,.$$

ℓ_1 and ℓ_0 calculate the positive and negative class losses, respectively. $\lambda_1 \geq 0$ and $\lambda_0 \geq 0$ are the hyperparameters of the positive and negative classes. Additionally, experiments demonstrate that ASL outperforms BCE. In this work, we set λ_0 to 2 and λ_1 to 0.

1.3. Pseudo-Code of Proposed Method

In order to describe our proposed algorithm more clearly, we summarize it in the form of pseudo-code in Algorithm 1.

2. Supplementary Experiments

2.1. Dataset

In Tab. 1, we present four key characteristics of three benchmark datasets, including the number of training images, the number of test images, and the average number of labels per image. Pascal VOC 2007 [5] is a popular multi-label dataset containing 20 object categories, divided into a `trainval` set with 5,011 samples and a `test` set with 4,952 samples. MS-COCO 2014 [8] is another widely used multi-label dataset with 80 common categories, consisting of 82,081 training examples and 40,504 validation examples. Following prior majority work [2, 10], we use all of its validation examples as the `test` set, along with the 82,081 training images as the `train` set. NUS-WIDE [3] is a web image dataset containing 81 categories, with all images sourced from Flickr. In our experiments, we select 126,034 training images as the `train` set and 84,226 test images as the `test` set. Visual Genome [7] is a knowledge base and dataset containing 108,249 images covering 80,138 categories, each of which is annotated by humans with visual concepts. Given that most categories have very few samples and many categories share similar semantic concepts, we further processed this dataset. Following previous work [15], we merged categories with the same meaning and excluded categories with fewer than 500 images. Finally, we obtained a dataset called `VG256`, comprising 256 classes and 108,249 images, with 70% of the images used as the `train` set and 30% as the `test` set. Objects365 (`O365`) [13] is a large-scale object detection dataset with more than 600,000 images covering 365 different categories of everyday objects. It aims to provide a more comprehensive and diverse object recognition scenario. Compared with `COCO`, `O365` provides more categories and images, which is more in line with real scenes and is, therefore, more suitable for multi-label image learning. Similar to the `COCO` dataset, annotation information is unavailable for the `test` set. Consequently, we designated the `train` set, which comprises 1,742,292 images, for training, and all validation examples as `test` set, containing 193,588 images, for testing.

2.2. Evaluation Metrics

In this work, beyond mean average precision (mAP), the standard metrics reported in the experimental section include overall precision (OP), overall recall (OR), overall F1 score (OF1), as well as per-category precision (CP), per-category recall (CR), and per-category F1 score (CF1). These metrics are computed as follows:

$$\text{OP} = \frac{\sum_i \text{TP}_i}{\sum_i \text{TP}_i + \text{FP}_i}\,, \qquad \text{OR} = \frac{\sum_i \text{TP}_i}{\sum_i \text{TP}_i + \text{FN}_i}\,,$$
$$\text{CP} = \frac{1}{C}\sum_i \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}\,, \qquad \text{CR} = \frac{1}{C}\sum_i \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}\,,$$
$$\text{OF1} = \frac{2 \times \text{OP} \times \text{OR}}{\text{OP} + \text{OR}}\,, \qquad \text{CF1} = \frac{2 \times \text{CP} \times \text{CR}}{\text{CP} + \text{CR}}\,,$$

where TP_i is true positive of class i, FP_i is false positive of class i, FN_i is false negative of class i. Among the metrics, OF1 and CF1 are the most significant, as they take both recall and precision into account, offering a more comprehensive evaluation. Moreover, with the exception of mAP, note that these results may be sensitive to the chosen threshold.

2.3. Compared to State-of-the-Art Results

Performance on Objects365. To evaluate our method on a more comprehensive and realistic dataset, we compared our approach with state-of-the-art (SOTA) methods on the `O365`. The results are presented in Table 2, demonstrating that our method improves mAP, CF1 and OF1 by 0.9-4.1%, 0.4-3.1% and 0.6-2.5%, respectively. The effectiveness of our proposed method is verified on a larger dataset with stronger label relationships.

2.4. Diagnostic Experiments

Number of Groups. Due to the limitations in the length of the main body, we present a comprehensive evaluation of the number of groups as shown in Fig. 3, which are conducted on `COCO` and `VOC07`, including metrics such as mAP, CF1, and OF1. The overall results indicate that increasing the number of groups has improved the performance of our method, although some minor fluctuations were observed.
Number of Experts. Due to space constraints in the main body, we present a comprehensive evaluation of the study

Algorithm 1: Multi-Label Visual Prompt Tuning for Multi-Label Image Classification

Input & Prepare: Given a multi-label image dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ with K classes, and their label co-occurrence graph $\mathcal{G}^+$ and dis-occurrence graph $\mathcal{G}^-$, where $\mathbf{1} = \mathcal{G}^+ + \mathcal{G}^-$.

Grouping classes into multiple groups $\{\mathcal{C}_t^+\}_{t=1}^{T} \leftarrow \text{GraphPartition}(\mathcal{G}^+)$ and $\{\mathcal{C}_t^-\}_{t=1}^{T} \leftarrow \text{GraphPartition}(\mathcal{G}^+)$ `// Apply a clustering algorithm on the graph` $(\mathcal{G}^+)$ `and` $\mathcal{G}^-$`)`

Freeze the ViT parameters and add a set of learnable prompt tokens ($\mathbf{P}^+$ / $\mathbf{P}^-$) for $\{\mathcal{C}_t^+\}_{t=1}^{T}$ and $\{\mathcal{C}_t^-\}_{t=1}^{T}$, respectively, $\text{ViT}(\boldsymbol{x}) = \text{ViT}([\mathbf{P}^+, \mathbf{P}^-], \boldsymbol{x})$ `// Using VPT technology to build a visual encoder model`

for $k = 1$ *to MaxEpoch* **do**

  Obtain group-level representations $\mathbf{Z}^+ \cup \mathbf{Z}^- = \text{ViT}(\boldsymbol{x})$

  Obtain label-aware representations $\{\mathbf{c}_k^+\}_{k=1}^{K} = \text{MoE}(\mathbf{Z}^+)$, $\{\mathbf{c}_k^-\}_{k=1}^{K} = \text{MoE}(\mathbf{Z}^-)$ `// Using MoE`

  Calculate logits $\hat{\boldsymbol{y}}^+ = \{\hat{y}_k^+\}_{k=1}^{K} = \text{Classifier}^+(\{\mathbf{c}_k^+\}_{k=1}^{K})$, $\hat{\boldsymbol{y}}^- = \{\hat{y}_k^-\}_{k=1}^{K} = \text{Classifier}^-(\{\mathbf{c}_k^-\}_{k=1}^{K})$

  Update model $f(\cdot)$ with $\mathcal{L}_{\text{ASL}}(\hat{\boldsymbol{y}}^+, \boldsymbol{y}) + \mathcal{L}_{\text{ASL}}(\hat{\boldsymbol{y}}^-, \boldsymbol{y})$

end

Output: The trained multi-label visual prompt tuning model $f(\cdot)$.

Predict: $\hat{\boldsymbol{y}} = 0.5 \cdot (\hat{\boldsymbol{y}}^+ + \hat{\boldsymbol{y}}^-) = f(\boldsymbol{x})$

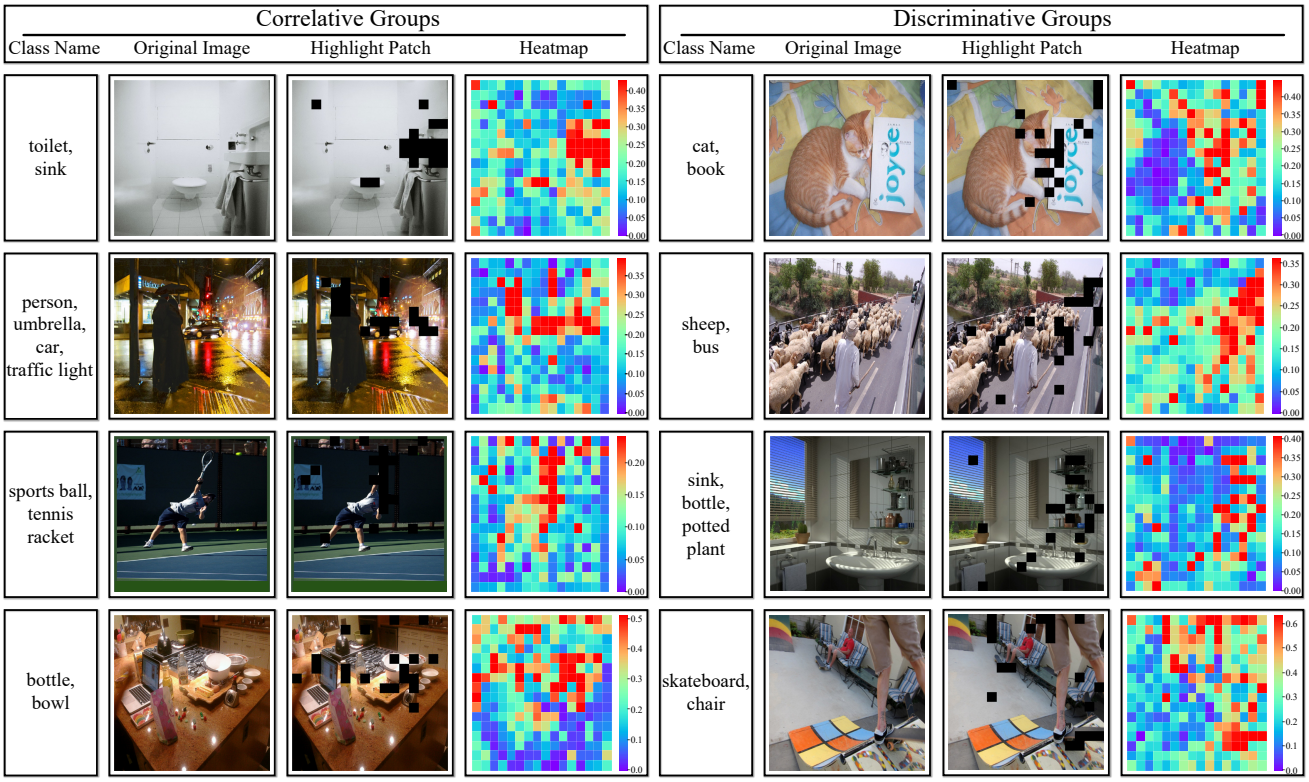


Figure 2. Visualization of group heatmap from the final layer of the ViT on COCO. The left side is the correlative groups, and the right side is the discriminative groups. As the main body, we also highlight the top-20 image patches based on heatmap scores.

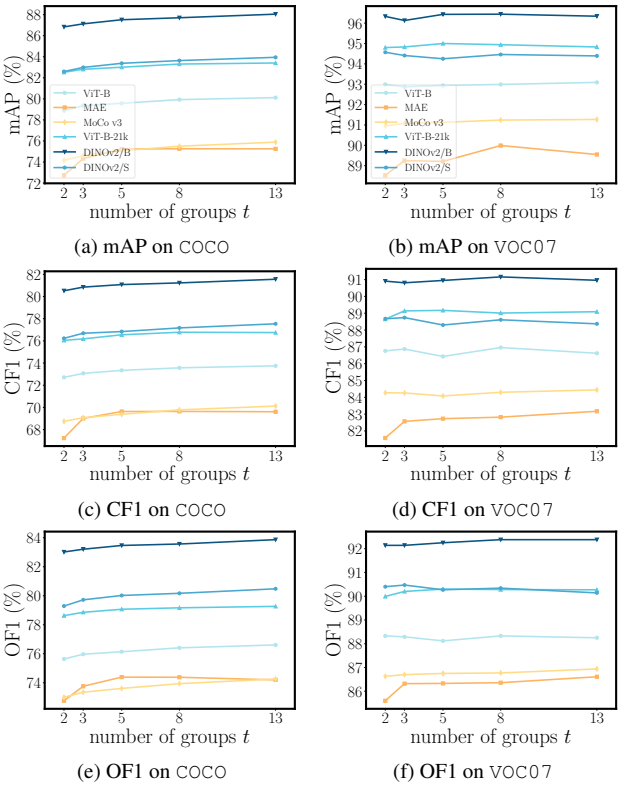
Table 1. Statistics for the popular benchmark dataset, including the number of training images, test images, categories, and average number of labels per image.

Dataset	# Train	# Test	# Classes	# Avg. Pos.
Pascal VOC 2007 (VOC07)	5,011	4,952	20	1.5
MS-COCO 2014 (COCO)	82,081	40,504	80	2.9
NUS-WIDE (NUS)	126,034	84,226	81	2.4
Visual-Genome (VG256)	75,773	32,475	256	7.3
Objects365 (O365)	1,742,292	193,588	365	6.1

on the number of experts as shown in Fig. 4, which are conducted on COCO and VOC07, including metrics such as mAP, CF1, and OF1. The results show that, on the COCO dataset, our method performs better as the number of experts increases. In contrast, on the VOC07 dataset, the performance improvement with more experts fluctuates significantly. One possible explanation is that the smaller number of categories in each group does not require more prompt tokens to capture label relationships, so the transition from group-aware representation to label-aware representation does not require as many experts.

Table 2. Comparison of our method with SOTA models on O365 at 224 × 224 resolution. All metrics are in %.

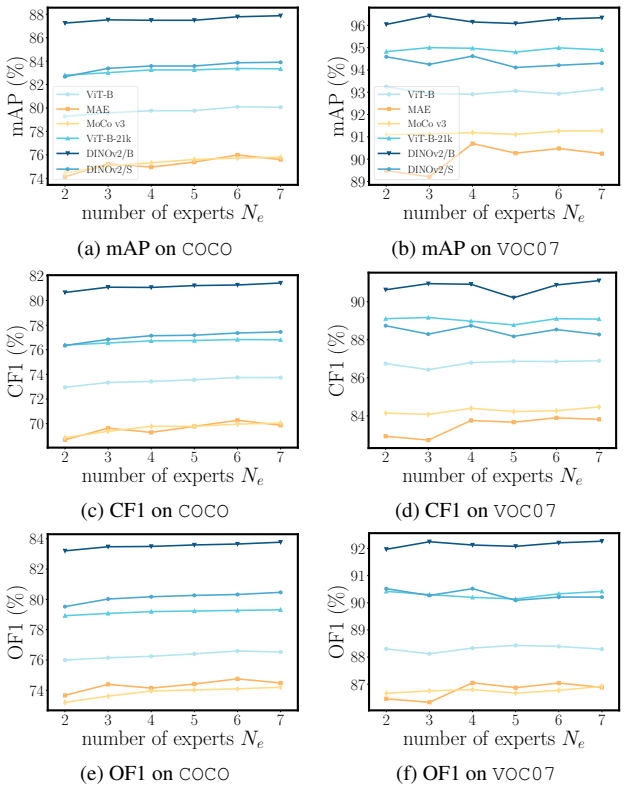
Method	Backbone	Resolution: 224 × 224 mAP	CP	CR	CF1	OP	OR	OF1
VPT	ViT-B	37.6	39.3	39.4	39.4	60.4	60.6	60.5
GateVPT		36.3	38.4	38.4	38.4	58.8	59.0	58.9
E2VPT		37.9	39.6	39.6	39.6	60.5	60.7	60.6
Ours		40.0	41.3	41.4	41.3	61.7	61.8	61.7
VPT	MAE	31.5	34.3	34.4	34.4	60.2	60.4	60.3
GateVPT		26.8	29.9	30.0	30.0	56.9	57.1	57.0
E2VPT		29.6	32.4	32.5	32.4	59.4	59.5	59.4
Ours		31.1	33.9	33.9	33.9	60.6	60.7	60.7
VPT	MoCO v3	31.8	34.5	34.5	34.5	58.1	58.3	58.2
GateVPT		31.8	34.5	34.5	34.5	58.1	58.3	58.2
E2VPT		31.8	34.4	34.5	34.4	58.0	58.3	58.2
Ours		33.6	35.7	35.7	35.7	59.1	59.2	59.2
VPT	ViT-B-21k	44.3	45.0	45.1	45.1	64.1	64.3	64.2
GateVPT		42.9	44.0	44.0	44.0	62.8	63.1	62.9
E2VPT		44.1	44.9	44.9	44.9	64.1	64.3	64.2
Ours		45.2	45.5	45.5	45.5	64.7	64.8	64.8
VPT	DINOv2/B	49.3	49.3	49.3	49.3	69.6	69.8	69.7
GateVPT		48.1	48.4	48.5	48.4	68.3	68.5	68.4
E2VPT		49.2	49.2	49.3	49.3	69.5	69.7	69.6
Ours		52.2	51.5	51.5	51.5	70.8	71.0	70.9
VPT	DINOv2/S	40.1	41.5	41.6	41.6	64.2	64.4	64.3
GateVPT		39.0	40.8	40.8	40.8	62.7	62.9	62.8
E2VPT		40.2	41.8	41.8	41.8	64.2	64.4	64.3
Ours		41.4	42.7	42.8	42.7	65.2	65.3	65.3



(a) mAP on COCO (b) mAP on VOC07
(c) CF1 on COCO (d) CF1 on VOC07
(e) OF1 on COCO (f) OF1 on VOC07

Figure 3. The performance curve varies with the increase in the number of groups, in the 3 evaluation metrics: mAP, CF1, and OF1.

Effect of Group Strategy. To further validate the effectiveness of the grouping strategy, we selected label pairs with co-occurrence probabilities greater than 0.2 from on COCO. As shown in Fig. 5, the we proposed ML-VPT outperforms the VPT method in terms of CTPR and CFPR for approximately 88.4% and 85.3% of the label pairs respectively. The higher CTPR indicates that ML-VPT has a stronger ability to capture correlational features, while the lower CFPR suggests that ML-VPT effectively balances both correlational and discriminative features, thereby reducing the risk of overfitting. Notably, the label pairs in the figure are sorted based on the VPT results for clarity and aesthetic purposes.



(a) mAP on COCO (b) mAP on VOC07
(c) CF1 on COCO (d) CF1 on VOC07
(e) OF1 on COCO (f) OF1 on VOC07

Figure 4. The performance curve varies with the increase in the number of experts, in the mAP, CF1, and OF1.

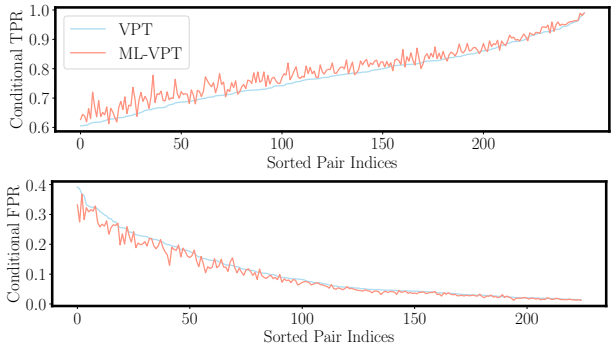
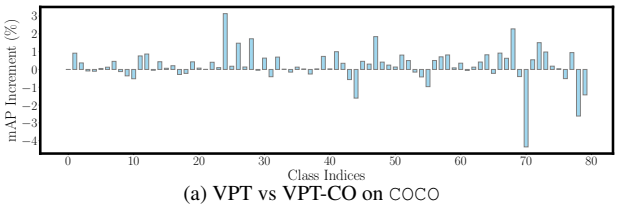


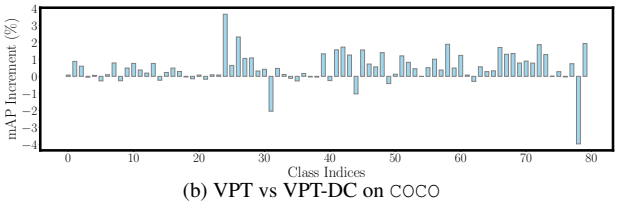
Figure 5. Comparison between VPT and ML-VPT (ours) in terms of conditional TPR and FPR, the label pairs with co-occurrence probabilities larger than 0.2 on COCO is selected.

Delve into Grouping Strategy. Our method divides classes into correlative and discriminative groups to balance their relationships. For comparison, we also conduct experiments with only correlative grouping (VPT-CO) and only discriminative grouping (VPT-DC), as shown in Figure 6. While both VPT-CO and VPT-DC show an overall improvement

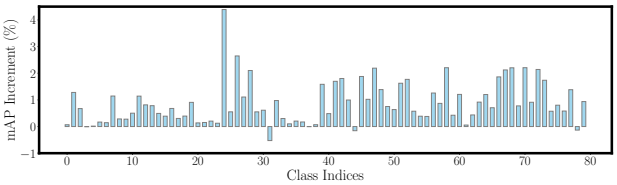
in mean Average Precision (mAP) over VPT, they lead to significant accuracy drops for certain classes. Our grouping strategy (GVPT) avoids this issue, providing strong evidence that balancing correlative and discriminative groups is both effective and reasonable.



(a) VPT vs VPT-CO on COCO



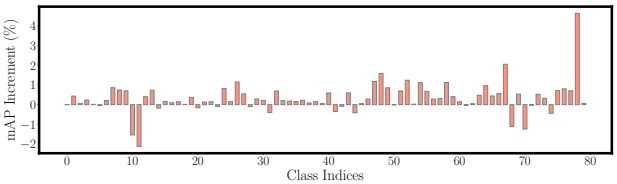
(b) VPT vs VPT-DC on COCO



(c) VPT vs GVPT on COCO

Figure 6. Per-class mAP increment for VPT-CO (with only correlative groups), VPT-DC (with only discriminative groups), and GVPT (both correlative and discriminative groups).

Effect of Mixture-of-Experts In this work, mixture of experts(MoE) is employed to allocate group-aware representations to label-aware representations, aiming to improve classification performance. To evaluate the effectiveness of the MoE strategy, we present the mAP increment when MoE is incorporated, compared to its absence, as depicted in Figure 7, MoE demonstrates a beneficial effect for 93.75% of the labels, with only 5 labels exhibiting a slight decrement.



(a) GVPT vs ML-VPT on COCO

Figure 7. Per-class mAP increment for Mixture-of-experts.

Gating Network Strategy: Label-Aware or Group-Aware. To compare the performance of building gating networks at both the label-aware and group-aware, we conducted the following experiments on the COCO dataset using various pre-trained models. The reported results represent the average performance of each method across these models. As illustrated in Tab. 3, the label-aware gating network strategy outperforms the group-aware strategy. This superiority is attributed to the label-aware gating network's ability to select group-aware representations that are most appropriate for the current class, based on the image-specific. In contrast, the group-aware strategy does not account for this selection. Note that in this work, we choose the label-aware gating networks.

Table 3. Comparison of two ways to build gating networks on COCO. All metrics are in %.

Method	Avg. mAP	Avg. CF1	Avg. OF1
Group Level	79.96	73.93	77.32
Label Level	80.60	74.41	77.64

Randomization Grouping strategies. The results, which are averaged across multiple pre-trained models (including ViT [4], ViT-21k [4], MAE [6], MoCo v3 [1], DINOv2/S [11], and DINOv2/B [11]), are presented in Tab. 4. Our grouping strategy outperforms all others, including random grouping. Note that in this experiment, we do not consider using MoE. We hypothesize that random grouping might somewhat balance relevance and discrimination relationships; however, it is not the optimal strategy.

Table 4. Comparison between multiple grouping strategies on COCO. All metrics are in %.

Method	Avg. mAP	Avg. CF1	Avg. OF1
Random-Group	78.39	72.67	76.40
CO-Group	77.79	72.21	75.97
DC-Group	77.96	72.30	76.08
CO-Group&DC-Group	78.84	73.00	76.58

2.5. More Case Study

Visualization of Group Heatmap. To demonstrate that our method effectively learns group-aware representations (group-aware representations) through our grouping strategy, we present a visualization of the group heatmap in Fig. 2. These results show that our method can model the relationship between relevant labels and discriminative labels.

Weights Assigned to Experts. As shown in Tab. 5, the proposed MoE effectively assigns distinct weights to different classes across various images. For instance, although bottle and broccoli are grouped within the same group, the weights required for these two classes by the three experts differ significantly.

Table 5. Weights assigned to experts for different classes on COCO. From left to right, they are the image ID, class name, and the weight assigned by the corresponding expert to the corresponding class.

Image ID	Classes Name	Expert 1	Expert 2	Expert 3
000000001000	person	3.0e-06	9.9e-01	3.2e-04
000000010056	car	9.5e-01	2.2e-02	2.5e-02
000000100000	cat	9.9e-01	3.3e-05	1.1e-04
000000100132	fork	4.4e-04	1.9e-03	9.9e-01
000000100238	bottle	5.1e-01	3.1e-01	1.8e-01
000000100624	car	2.3e-05	9.9e-01	2.9e-03
000000100582	fork	4.3e-01	3.1e-01	2.5e-01
000000100811	person	2.5e-02	6.1e-01	3.6e-01
000000113294	boat	1.5e-01	8.5e-01	6.8e-05
000000214919	broccoli	2.7e-05	1.8e-04	9.9e-01

References

[1] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 5

[2] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *CVPR*, pages 5177–5186, 2019. 2

[3] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*, Santorini, Greece., 2009. 2

[4] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[5] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111:98–136, 2015. 2

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 5

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 2

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 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] Leilei Ma, Dengdi Sun, Lei Wang, Haifeng Zhao, and Bin Luo. Semantic-aware dual contrastive learning for multi-label image classification. In *ECAI*, pages 1656–1663, 2023. 2

[11] 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. 5

[12] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *CVPR*, pages 82–91, 2021. 2

[13] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *CVPR*, pages 8430–8439, 2019. 2

[14] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17:395–416, 2007. 1

[15] Ming-Kun Xie, Jia-Hao Xiao, Pei Peng, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Counterfactual reasoning for multi-label image classification via patching-based training. In *ICML*, 2024. 2

[16] Jiazhi Xu, Sheng Huang, Fengtao Zhou, Luwen Huangfu, Daniel Zeng, and Bo Liu. Boosting multi-label image classification with complementary parallel self-distillation. In *IJCAI*, 2022. 1