Multi-label Cluster Discrimination for Visual Representation Learning

Xiang An[1], Kaicheng Yang[1], Xiangzi Dai[1], Ziyong Feng[1], and Jiankang Deng[⋆2]

[1] DeepGlint
xiangan@deepglint.com
[2] Huawei Noah's Ark Lab
jiankang.deng@gmail.com

Abstract. Contrastive Language Image Pre-training (CLIP) has recently demonstrated success across various tasks due to superior feature representation empowered by image-text contrastive learning. However, the instance discrimination method used by CLIP can hardly encode the semantic structure of training data. To handle this limitation, cluster discrimination has been proposed through iterative cluster assignment and classification. Nevertheless, most cluster discrimination approaches only define a single pseudo-label for each image, neglecting multi-label signals in the image. In this paper, we propose a novel Multi-Label Cluster Discrimination method named MLCD to enhance representation learning. In the clustering step, we first cluster the large-scale LAION-400M dataset into one million centers based on off-the-shelf embedding features. Considering that natural images frequently contain multiple visual objects or attributes, we select the multiple closest centers as auxiliary class labels. In the discrimination step, we design a novel multi-label classification loss, which elegantly separates losses from positive classes and negative classes, and alleviates ambiguity on decision boundary. We validate the proposed multi-label cluster discrimination method with experiments on different scales of models and pre-training datasets. Experimental results show that our method achieves state-of-the-art performance on multiple downstream tasks including linear probe, zero-shot classification, and image-text retrieval.

Keywords: Visual Representation Learning, Instance Discrimination, Cluster Discrimination, Multi-label Learning

1 Introduction

Language-supervised visual pre-training, *e.g.*, CLIP [34] and ALIGN [19], has been established as a simple yet effective methodology for visual representation learning. Empowered by image-text contrastive learning, pre-trained CLIP models exhibit remarkable versatility and transferability across various downstream

⋆ Corresponding author.

tasks (*e.g.*, linear probe, zero-shot classification, and image retrieval). As illustrated in Fig. 1a, CLIP aligns the visual and textual signals of each instance into a unified semantic space by cross-modal instance discrimination. Nevertheless, the instance discrimination method used by CLIP can hardly encode the semantic structure of training data, because instance-wise contrastive learning always treats two samples as a negative pair if they are from different instances, regardless of their semantic similarity. When a large number of instances are selected into the mini-batch to form the contrastive loss, negative pairs that share similar semantics will be undesirably pushed apart in the embedding space.

To handle the limitations of instance discrimination, cluster discrimination methods (*e.g.*, DeepCluster [7], SeLa [5], ODC [48], SwAV [8], CoKe [33], and UNICOM [4]) have been proposed for deep unsupervised learning through jointly learning image embeddings and cluster assignments. Learning representations with clusters will pull similar instances together, which is beneficial for capturing semantic structures in data. However, most cluster discrimination approaches only define a single pseudo-label for each image as depicted in Fig. 1b. By contrast, natural language supervision proposed in CLIP can provide richer forms of labels for a single image, *e.g.*, objects, scenes, actions, and relations, at multiple levels of granularity.

As can be seen from Fig. 2, a web image frequently contains multiple classification targets, such as objects [44] or attributes [32]. The existence of multiple objects in the image requires laborious cropping [2, 23] to construct single-label annotations, while some scenario elements and attributes in the image are hard to disentangle to obtain single-label instances [32, 51]. These real-world challenges pose so-called multi-label classification where an image is equipped with multiple labels beyond a single label.

In this paper, we aim to boost the visual representation power of the CLIP model by introducing a novel Multi-Label Cluster Discrimination (MLCD) approach. In the clustering step, we follow UNICOM [4] to conduct one step of offline clustering by using the features predicted by a pre-trained CLIP model. Due to the limited discrimination power of the CLIP model [34], the single pseudo-label may not cover all of the visual signals (*e.g.*, objects or attributes) in the image. To this end, we further perform a similarity-based sorting against k class centers and select the top l class centers as the positive class centers for that image. In the discrimination step, we follow the Circle loss [40] to design a multi-label loss to effectively deal with multiple labels. The vanilla version of the multi-label loss exploits relative similarity comparisons between positive and negative classes. More specifically, the optimization seeks to narrow the gap between the intra-class similarities $\{s_i\}$ and the inter-class similarities $\{s_j\}$ by reducing all possible $(s_j - s_i)$. However, optimizing $(s_j - s_i)$ usually leads to a decision boundary allowing ambiguity [40]. To this end, we introduce another two optimization targets (*i.e.*, decreasing s_j and increasing s_i) into the loss function. Introducing the additional two items enables an elegant separation of positive class loss and negative class loss (Eq. 5), which can alleviate the ambiguity on the decision boundary. To alleviate inter-class conflict and save the computation

time on the classifier layer, we also employ PartialFC [3] and randomly sample part of the negative class centers during each iteration.

The main contributions of our paper are the following:

1. We propose a novel multi-label cluster discrimination method for visual representation learning on large-scale data. In the clustering step, we employ one step of offline k-means to predict multiple labels for each training sample. In the discrimination step, we explore multi-label classification, which considers multiple supervision signals for a single image and learns better semantic structure in data.
2. To avoid ambiguity during the optimization of $(s_j - s_i)$, we add additional optimization targets by maximizing the within-class similarity s_i, as well as to minimizing the between-class similarity s_j. By doing so, the loss from positive class labels and negative class labels can be elegantly separated.
3. The proposed multi-label cluster discrimination significantly boosts the representation power compared to the instance discrimination-based model (*e.g.*, OpenCLIP [11] and FLIP [24]) and the cluster discrimination-based model (*e.g.*, UNICOM [4]) on the downstream tasks (*e.g.*, linear probe, zero-shot classification, zero-shot retrieval).

2 Related Work

Visual Representation Learning. Visual representation pre-training methods can be mainly divided into three categories: (1) supervised learning by using manually annotated class labels (*e.g.*, ImageNet-1K/-21K [13] and JFT-300M/-3B [15,46]), (2) weakly-supervised learning by employing hashtags [29,38] or text descriptions [19,24,34], and (3) unsupervised learning [7,9,18] by designing appropriate pretext tasks (*e.g.*, solving jigsaw puzzles [31], invariant mapping [10], and masked image inpainting [17]). Even though fully supervised pre-training can learn a strong semantic signal from each training example, manual label annotation is time-consuming and expensive thus supervised learning is less scalable. In this paper, we focus on annotation-free pre-training which can be easily scaled to billions of web images to learn visual representation for downstream tasks.

Instance and Cluster Discrimination. Instance discrimination [9, 18, 34] is usually implemented by the contrastive loss to pull images from the same instance as well as push away images from different instances. Among these instance discrimination methods, language-supervised visual pre-training, *e.g.*, CLIP [16, 34, 45], is a simple yet powerful approach to take advantage of rich forms of labels at multiple levels of granularity for a single image. Even though CLIP [34] has recently demonstrated impressive success, instance-wise contrastive learning always treats different instances as negative pairs thus it can hardly capture the full semantic information from the training data.

To explore potential semantic structures in the training data, cluster discrimination [5,7,8,22,33,48] is proposed with two iterative steps: (1) the clustering step to assign a single class label for each sample, and (2) the classification step to learn a classifier to predict the assigned pseudo label. In cluster discrimination

methods, each cluster contains more than one instance, visually similar instances will be pulled closer and thus cluster discrimination can better capture semantic structures from data. However, multiple visual elements can exist in one single image and the single label used by cluster discrimination may not cover all visual signals.

Multi-label Classification. Multi-label classification [41, 49] assigns a set of multiple labels for each instance. Compared with single-class classification, where each instance is assigned with a single label, multi-label classification [43, 44, 50] is more challenging [27, 28]. Considering multiple labels are drawn from k categories, the multi-label classification can be decomposed into k binary classification tasks. However, the binary cross-entropy loss involves issues regarding imbalance [35]. Through analyzing the intrinsic loss functions of the classification loss and the metric loss [42], Sun *et al*. [40] formulate a unified multi-label loss function to exploit relative comparison between positive and negative classes. Nevertheless, the relative comparison $(s_j - s_i)$ allows ambiguity for convergence. Su *et al*. [39] introduce a threshold into the multi-label loss and design the Threshold-bounded Log-sum-exp and Pairwise Rank-based (TLPR) loss, hoping that the logits of positive categories will be larger than the threshold and the logits of negative categories will be smaller than the threshold. However, the TLPR loss is only designed for clean multi-label datasets and is not suitable for large-scale multi-label datasets with heavy noises. In this paper, we only employ one step of offline clustering to predict multiple labels for each image and then design a robust multi-label classification disambiguation loss to achieve good feature representation when training on the automatically clustered large-scale data.

3 Method

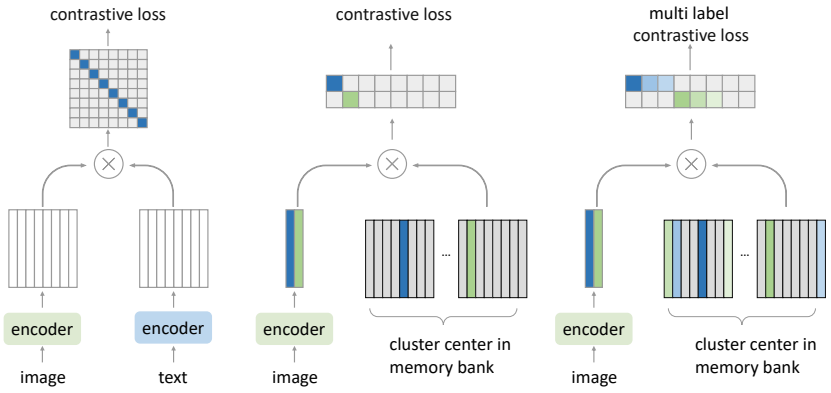
Given a training set $X = \{x_1, x_2, ..., x_n\}$ including n images, visual representation learning aims at learning a function f that maps images X to normalized embeddings $E = \{e_1, e_2, ..., e_n\}$ with $e_i = f(x_i)$, such that embeddings can describe the semantic similarities between different images.

3.1 Preliminaries

Instance Discrimination achieves semantic embedding by minimizing a contrastive loss function represented as:

$$\mathcal{L}_{\text{ID}} = -\log \frac{\exp(e_i'^T e_i)}{\sum_{j=1}^{k} \exp(e_j'^T e_i)}, \tag{1}$$

where $\exp(\cdot)$ denotes the exponential function, and e_i and e_i' denote the normalized image and text embeddings for the instance i in CLIP [34]. Meanwhile, e_j' contains one positive text representation for i and $(k-1)$ negative text representations sourced from different instances. As illustrated in Fig. 1a, the instance discrimination based CLIP model jointly trains an image encoder and a text



(a) Instance Discrimination **(b)** Cluster Discrimination **(c)** Multi-Label Cluster Discrimination

Fig. 1: Comparisons of instance discrimination, cluster discrimination, and the proposed multi-label cluster discrimination. (a) Instance discrimination treats each image-text pair as a unique instance, failing to capture the semantic structure within the training data. (b) Cluster discrimination improves the semantic embedding by grouping similar instances but struggles with multi-label signals in a single image. (c) The proposed multi-label cluster discrimination addresses this challenge by assigning multiple class labels to each sample, capturing different granularities of visual signals (*e.g.*, objects or attributes) in one image.

encoder to predict the correct image-text pairings from a batch of training examples.

Cluster Discrimination is composed of two primary stages: the clustering process and the discrimination process. During the clustering phase, every instance is assigned one pseudo-class label. This label is later employed as a guiding factor for training a classifier in the subsequent discrimination phase. For the normalized embedding feature $e_i = f(x_i)$, the clustering process determines a centroid matrix $W \in \mathbb{R}^{d\times k}$ and assigns the cluster label y_i for each image x_i. This is achieved by

$$\min_{W\in\mathbb{R}^{d\times k}} \frac{1}{n}\sum_{i=1}^{n} \min_{y_i\in\{0,1\}^k} \|e_i - W y_i\|_2^2 \quad \text{s.t.} \quad y_i^\top \mathbf{1_k} = \mathbf{1}, \tag{2}$$

where n is the number of training samples, e_i is the normalized feature embedding obtained by using the image encoder f, and the centroid w_i belonging to centroid matrix $W \in R^{d\times k}$ is considered the normalized prototype of i-th cluster. y_i, falling within the set $\{0,1\}^k$, stands as a single label assignment restricted by the condition $y_i^\top \mathbf{1_k} = \mathbf{1}$, where $\mathbf{1_k}$ is 1-vector with a length of k.

Then, the training data, denoted as $\{x_i\}_{i=1}^n$, is divided into k classes represented by prototypes $W = \{w_i\}_{i=1}^k$. Utilizing the pseudo labels and centroids derived from the clustering phase, the process of cluster discrimination can be executed by minimizing a conventional softmax classification loss, formulated as:

Fig. 2: Illustration of the multiple visual elements (*e.g.*, objects or attributes) in images from the automatically clustered LAION-400M dataset.

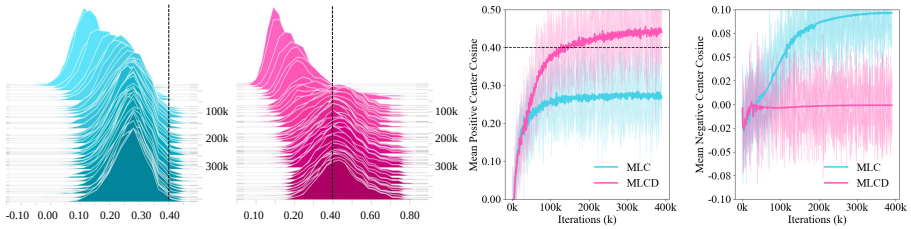
$$\mathcal{L}_{\text{CD}} = -\log \frac{\exp(w_{y_i}^T e_i)}{\sum_{j=1}^{k} \exp(w_j^T e_i)} = -\log \frac{\exp(s_i)}{\sum_{j=1}^{k} \exp(s_j)}$$
$$= \log(1 + \sum_{j=1, j\neq i}^{k} \exp(s_j - s_i)), \tag{3}$$

where e_i is the normalized embedding corresponding to the image x_i, and x_i is categorized under the class symbolized by the normalized prototype w_{y_i}. For a more straightforward representation, we define the intra-class similarity $w_{y_i}^T e_i$, and the inter-class similarity, $w_j^T e_i$ as s_i and s_j, respectively. Based on Eq. 3, in the discrimination phase that employs classification, s_j and s_i are paired to optimize the reduction of the difference $(s_j - s_i)$. As depicted in Fig. 1b, the cluster discrimination based UNICOM model [4] trains an image encoder to predict the one-hot pseudo label for each image from a batch of training examples.

3.2 Multi-label Cluster Discrimination

Clustering. Considering the time consumption of iterative clustering and discrimination [7], An *et al.* [4] implemented a single step of offline clustering with the aid of the pre-trained CLIP model (*i.e.*, ViT-L/14) and efficient feature quantization [20]. On the large-scale LAION-400M dataset, it only takes around 10 minutes to cluster one million classes. Despite the straightforwardness of the clustering step, the automatically clustered large-scale dataset inevitably confronts intra-class purity and inter-class conflict problems due to the specific definition of class granularity.

In the realm of clustering algorithms, there often exists a trade-off between maintaining high within-class purity and ensuring low inter-class conflict. In the



(a) s_i of MLC **(b)** s_i of MLCD **(c)** Mean s_i **(d)** Mean s_j

Fig. 3: Intra-class and inter-class similarity score comparisons between MLC and MLCD. Here, MLC and MLCD are trained on the LAION-400M dataset with the ViT-B/32 as the backbone and a batch size of 32K. (a) and (b) showcase histograms that compare the distributions of positive cosine similarities $\{s_i\}$ between MLC and MLCD, with MLCD clearly showing tighter sample alignment to positive class centers. (c) demonstrates that MLCD consistently achieves higher mean positive cosine values than MLC over iterations, indicating enhanced intra-class compactness. (d) demonstrates MLCD's effectiveness in reducing mean negative cosine values compared to MLC, which indicates a more orthogonal relationship between samples and their negative class centers. This greater orthogonality facilitated by MLCD contributes to enhanced class separability. These figures highlight MLCD's advanced capability in refining feature spaces for more distinct representation compared to MLC.

context of contrastive learning, the issue of inter-class conflict can be significantly alleviated by reducing the number of sampled negative instances within the mini-batch and adopting a suitable semi-hard mining technique. In this paper, we follow UNICOM [4] to prioritize intra-class purity (*i.e.*, clustering one million level classes from 400 million images) and employ margin-based PatialFC [3,14] to alleviate inter-class conflict (*i.e.*, randomly sampling part of the negative class centers during each iteration).

Multi-label Classification. As illustrated in Fig. 2, a single image can encompass several visual components (*e.g.*, objects or attributes). This implies that the single-class label may not cover all visual cues present in the image. To consider the different granularities of visual information for each sample, we perform a similarity-based sorting against one million class centers, selecting the top l class centers as the positive class centers for that sample. During training, this sample will be directed to move closer to these l positive class centers, while simultaneously distancing from the other $k-l$ negative class centers. As shown in Fig. 1c, our method assigns multiple class labels to each training example, capturing different granularities of visual signals in one image.

The corresponding similarity scores are represented as $\{s_i\}\,(i = 1, 2, \cdots, l)$ and $\{s_j\}\,(j = 1, 2, \cdots, k-l)$, respectively. To minimize each s_j $(\forall j \in \{1, 2, \cdots, k-l\})$ as well as to maximize s_i $(\forall i \in \{1, 2, \cdots, l\})$, we employ a multi-label classification strategy [25,40]. This is achieved by

$$\mathcal{L}_{\text{MLC}} = \log(1 + \underbrace{\sum_{j=1}^{k-l}\sum_{i=1}^{l}\exp(s_j - s_i)))}_{contrastive} = \log(1 + \underbrace{\sum_{j\in\Omega_n}\exp(s_j)\sum_{i\in\Omega_p}\exp(-s_i)}_{contrastive}), \quad (4)$$

where Ω_n and Ω_p denote the negative and positive class set to simplify the representation. Eq. 4 iterates through every similarity pair to reduce $(s_j - s_i)$. To alleviate inter-class conflict as in [3,4], we also employ negative class sampling into Eq. 4. Therefore, the loss is changed from $\log(1+\sum_{j\in\Omega_n}\exp(s_j)\sum_{i\in\Omega_p}\exp(-s_i))$ to $\log(1+\sum_{j\in\Omega'_n}\exp(s_j)\sum_{i\in\Omega_p}\exp(-s_i))$, where $|\Omega'_n| = |\Omega_n| * r$, and $r \in [0,1]$ is the negative class sampling ratio. Ω'_n is a subset of Ω_n that is randomly sampled during each loss calculation step.

Multi-label Classification Disambiguation. Optimizing $(s_j - s_i)$ usually leads to a decision boundary of $s_j - s_i = m$ (m is the margin). However, this decision boundary allows ambiguity as indicated in Circle loss [40]. For example, $\{s_j, s_i\} = \{0.1, 0.4\}$ and $\{s'_j, s'_i\} = \{0.5, 0.8\}$ both achieve the margin $m = 0.3$. However, the gap between s_i and s'_j is only 0.1, compromising the separability of the feature space.

As we expect to maximize the within-class similarity s_i and to minimize the between-class similarity s_j, we further introduce these two items into the multi-label classification loss:

$$\begin{aligned}\mathcal{L}_{\text{MLCD}} &= \log(1 + \underbrace{\sum_{j\in\Omega'_n}\exp(s_j)\sum_{i\in\Omega_p}\exp(-s_i)}_{contrastive} + \underbrace{\sum_{j\in\Omega'_n}\exp(s_j)}_{negative} + \underbrace{\sum_{i\in\Omega_p}\exp(-s_i)}_{positive}) \\ &= \log(1 + \sum_{i\in\Omega_p}\exp(-s_i)) + \log(1 + \sum_{j\in\Omega'_n}\exp(s_j)),\end{aligned} \tag{5}$$

where Ω_p symbolizes the collection of positive class labels for each sample, s_i encapsulates the score associated with each positive class, Ω'_n denotes the collection of negative class labels for each sample, and s_j corresponds to the score for each negative class. In Eq. 5, loss from positive class labels $\log(1 + \sum_{i\in\Omega_p}\exp(-s_i))$ and loss from negative class labels $\log(1 + \sum_{j\in\Omega'_n}\exp(s_j))$ are elegantly separated. In Fig. 3a and Fig. 3b, we compare the dynamic distributions of s_i of MLC (Eq. 4) and MLCD (Eq. 5) during training steps. Besides, Fig. 3c illustrates the average s_i from MLC and MLCD during training. As we can see, the item designed for maximizing the within-class similarity s_i in Eq. 5 can significantly increase the intra-class cosine similarities, enhancing the intra-class compactness. In Fig. 3d, the item designed for minimizing the between-class similarity s_j can effectively suppress the inter-class cosine similarities, enforcing the inter-class discrepancy.

4 Experiments

4.1 Experimental Setting

Our models are pre-trained on the LAION-400M dataset [36] with the same model configurations as CLIP. The training process consists of 32 epochs, utilizing a batch size of 32K on 80 NVIDIA A100 GPUs. To expedite the training, we employ mixed-precision computation [30] and flash attention [12], while leveraging the DALI library for efficient data loading and pre-processing. We use the

Table 1: Linear probe performance of various pre-trained models on 26 datasets. †: Results reported in CLIP paper. ‡: Results we reproduced. Entries in green are the best results using LAION-400M. Here, all methods employ the same backbone of ViT-L/14.

CASE	DATA	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Cal101	Flowers	MNIST	FER2013	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	K700	CLEVR	HM	SST	AVG
CLIP†	WIT-400M	95.2	98.0	87.5	77.0	81.8	90.9	69.4	89.6	82.1	95.1	96.5	99.2	99.2	72.2	99.8	98.2	94.1	92.5	64.7	42.9	85.8	91.5	72.0	57.8	76.2	80.8	84.2
CLIP‡	WIT-400M	95.3	98.1	87.2	77.8	81.5	90.7	68.0	89.7	80.9	94.9	96.0	99.2	99.2	72.3	99.8	96.7	94.5	92.9	65.9	41.9	85.3	91.0	70.6	59.6	61.8	79.8	83.5
OPNCLIP‡	LAION-400M	93.3	97.9	87.9	78.0	81.0	93.6	64.4	91.7	83.0	93.3	95.5	98.8	99.2	66.5	99.2	97.1	92.4	92.5	77.5	32.5	84.3	88.1	64.0	59.8	57.6	71.9	82.3
UNICOM	LAION-400M	93.4	98.5	90.8	82.4	80.0	94.6	74.5	91.4	82.2	94.2	95.7	99.3	99.2	68.7	98.5	96.7	92.6	92.7	77.8	33.4	85.4	87.4	66.7	60.3	57.4	72.4	83.3
Ours	LAION-400M	94.3	98.9	92.0	83.4	82.1	94.8	79.6	92.5	84.6	95.3	97.2	99.3	99.3	72.4	99.3	99.1	94.7	92.5	78.2	34.5	86.0	90.0	68.5	60.1	57.9	73.4	84.6

AdamW optimizer with a learning rate of 0.001 and weight decay of 0.2. To assess the performance of zero-shot classification and zero-shot image-text retrieval tasks, we employ contrastive learning to train a text encoder from scratch for 32 epochs with a frozen image encoder following Locked-image Tuning (LiT) [47]. The structure of the text encoder is also identical to CLIP. In the following experiments, unless otherwise specified, the model used is ViT-L/14, the number of classes (k) is one million, the ratio of sampled negative class centers (r) is 0.1, and the number of positive labels (l) assigned to each image is 8.

4.2 Linear Probe

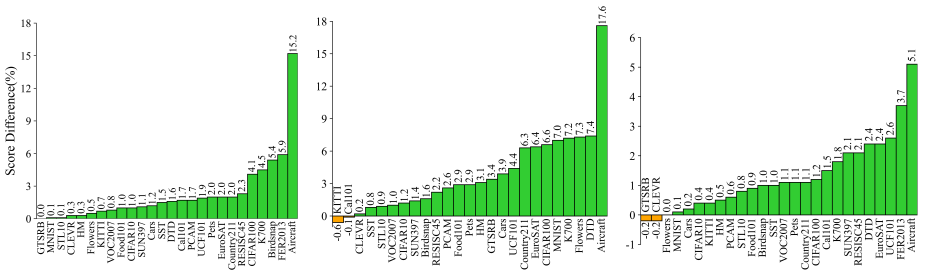
Following the same evaluation setting as CLIP, we report the linear probe performance of our method on 26 datasets. As depicted in Tab. 1, inherent biases exist in different pre-training data. The WIT dataset is beneficial for action-related datasets (*e.g.*, Kinetics700, UCF101), while LAION exhibits superior proficiency in object datasets (*e.g.*, Cars, Birdsnap). Nevertheless, our method still achieves an average improvement of 1.1% compared to CLIP. To isolate the confounding effects of pre-training data, we compare our model with OPENCLIP and UNICOM by using the LAION-400M dataset as the training data. As shown in Fig. 4a, our method outperforms OPENCLIP on 25 datasets, demonstrating an average improvement of 2.3%. In Fig. 4c, our model surpasses UNICOM on 23 datasets and achieves an average improvement of 1.3%, confirming the effectiveness of the proposed multi-label loss.

4.3 Zero-shot Classification

In Tab. 2, we present a comparison of our method with state-of-the-art approaches in zero-shot classification on 25 datasets. The prompt templates and class names are consistent with previous works [24]. As depicted in Fig. 4b, our method surpasses OpenCLIP on 23 datasets with 3.9% average performance improvement. Although FLIP uses masking to save memory footprint to learn more samples per iteration, our method demonstrates better results on 15 out of 25 datasets in Tab. 2 and achieves a significant performance boost of 1.5% on average.

Table 2: Zero-shot classification performance on 25 datasets. †: Results reported in CLIP paper. ‡: Results reported in FLIP paper. Entries in green are the best results using LAION-400M. Here, all methods employ the same backbone of ViT-L/14.

CASE	DATA	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Cal101	Flowers	MNIST	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	K700	CLEVR	HM	SST	AVG
CLIP†	WIT-400M	92.9	96.2	77.9	48.3	67.7	77.3	36.1	84.1	55.3	93.5	92.6	78.7	87.2	99.3	59.9	71.6	50.3	23.1	32.7	58.8	76.2	60.3	24.3	63.3	64.0	66.9
CLIP‡	WIT-400M	91.0	95.2	75.6	51.2	66.6	75.0	32.3	83.3	55.0	93.6	92.4	77.7	76.0	99.3	62.0	71.6	51.6	26.9	30.9	51.6	76.1	59.5	22.2	55.3	67.3	65.6
OpenCLIP‡	LAION-400M	87.4	94.1	77.1	61.3	70.7	86.2	21.8	83.5	54.9	90.8	94.0	72.1	71.5	98.2	53.3	67.7	47.3	29.3	21.6	51.1	71.3	50.5	22.0	55.3	57.1	63.6
FLIP‡	LAION-400M	89.3	97.2	84.1	63.0	73.1	90.7	29.1	83.1	60.4	92.6	93.8	75.0	80.3	98.5	53.5	70.8	41.4	34.8	23.1	50.3	74.1	55.8	22.7	54.0	58.5	66.0
Ours	LAION-400M	90.3	95.3	83.7	62.9	72.1	90.1	39.4	84.5	62.3	93.7	93.9	79.4	78.5	99.1	59.7	69.9	50.7	28.7	27.9	53.7	75.7	57.7	22.2	58.4	57.9	67.5



(a) MLCD vs. OpenCLIP on linear **(b)** MLCD vs. OpenCLIP on zero-shot **(c)** MLCD vs. UNICOM on linear

Fig. 4: Linear probe and zero-shot comparisons on different downstream datasets. The Y-axis shows the performance difference. Green bars indicate our model outperforms the baselines, while the orange bars depict our model is surpassed by the baselines.

4.4 Zero-shot Retrieval

Tab. 3 reports zero-shot image-text retrieval results on Flickr30k and MSCOCO. In comparison to OpenCLIP, our model achieves 60.8%/44.5% I2T/T2I retrieval Recall@1 on the MSCOCO dataset, which is 2.8%/3.2% higher than OpenCLIP. Similarly, our model demonstrates significant improvements of 1.8%/3.9% on the Flickr30k dataset. Furthermore, compared to FLIP, our model exhibits either competitive or superior retrieval performance.

4.5 ImageNet Classification and Robustness Evaluation

We evaluate performance on ImageNet [13] under three distinct settings: fine-tuning, linear probe, and zero-shot. As shown in Tab. 4, our ViT-L/14 model achieves better performance on all settings, outperforming OpenCLIP by 0.9% under the finetuning setting, and surpassing FLIP by 1.0% under the zero-shot setting. These improvements indicate that multi-label cluster discrimination can better encode the semantics of data than instance discrimination. Following FLIP [24], we conduct a robustness evaluation as shown in Tab. 4. In comparison to the models pre-trained on LAION, our method demonstrates superior robustness compared to both OpenCLIP and FLIP. It is worth noting that the performance gap between our model pre-trained on LAION and CLIP pre-trained on WIT arises from the statistical differences in pre-training data.

Table 3: Zero-shot image-text retrieval on the test splits of Flickr30k and MSCOCO. ‡: Results reported in FLIP paper. Entries in green are the best results using LAION-400M. Here, all methods employ the same backbone of ViT-L/14.

		Text retrieval						Image retrieval					
		Flickr30k			MSCOCO			Flickr30k			MSCOCO		
CASE	DATA	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP‡	WIT-400M	87.8	99.1	99.8	56.2	79.8	86.4	69.3	90.2	94.0	35.8	60.7	70.7
OpenCLIP‡	LAION-400M	87.3	97.9	99.1	58.0	80.6	88.1	72.0	90.8	95.0	41.3	66.6	76.1
FLIP‡	LAION-400M	89.1	98.5	99.6	60.2	82.6	89.9	75.4	92.5	95.9	44.2	69.2	78.4
Ours	LAION-400M	89.1	98.4	99.5	60.8	83.2	91.3	75.9	93.1	96.8	44.5	69.6	79.9

Table 4: ImageNet results under finetuning, linear probe, zero-shot, and zero-shot robustness evaluation settings. ‡: Results reported in FLIP paper. Entries in green are the best results using LAION-400M. Here, all methods employ the same backbone of ViT-L/14.

CASE	DATA	Finetune	Linear Probe	Zero Shot	IN-V2	IN-A	IN-R	ObjectNet	IN-Sketch
CLIP‡	WIT-400M	-	83.9	75.3	69.5	71.9	86.8	68.6	58.5
OpenCLIP‡	LAION-400M	86.2	82.1	72.8	64.0	48.3	84.3	58.8	56.9
FLIP‡	LAION-400M	-	-	74.6	66.8	51.2	86.5	59.1	59.9
Ours	LAION-400M	87.1	84.6	75.6	68.9	56.4	85.1	62.7	60.4

4.6 Ablation Study

Number of Classes. The number of classes (k) plays a crucial role in balancing inter-class conflict and intra-class purity. In Tab. 5a, we observe that as the number of classes increases from 100K to 1M, there is a gradual increase in intra-class purity, leading to an improved performance on ImageNet. However, as the number of classes continues to increase from 1M to 5M, inter-class conflicts gradually escalate, resulting in a deteriorated performance.

Inter-class sampling Ratio. The inter-class sampling ratio (r) influences the number of negative samples and directly affects the likelihood of encountering inter-class conflicts. A sample ratio of 0.01 yields a linear probe performance of only 73.4% due to the limited number of negative samples, which adversely affects the representation learning. Conversely, a sample ratio of 1.0 substantially increases the probability of encountering inter-class conflicts. Tab. 5b presents that the superior linear probe performance of 75.2% is achieved when employing a sample ratio of 0.1.

Multi-label Assignment. We explore two different approaches to obtain multi-labels. Firstly, we artificially assign a predetermined number of labels to each sample. Tab. 5c presents linear probe results on ImageNet with different numbers of positive centers. Consequently, we observe a gradual improvement in performance as the number of positive centers increases from 1 to 8. However, as the number of positive centers continues to increase, the inclusion of excessive positive centers introduces noise labels, leading to a degradation in performance. Additionally, we have also investigated the use of sample-cluster similarity thresholds to obtain multiple labels. This approach results in varying numbers of positive centers associated with each sample. However, as shown in Tab. 5d, the performance of applying adaptive positive centers is generally lower

Table 5: Ablation experiments. The model backbone used here is ViT-B/32. Pre-training is executed on the LAION-400M dataset for a duration of 5 epochs. Performance assessment is undertaken using a linear probe on the ImageNet validation set.

Num Classes	100K	200K	500K	1M	2M	5M
IN1K	66.9	71.1	74.4	75.2	74.9	74.7

(a) The number of **classes** in training set.

Sampling Ratio	0.01	0.05	0.1	0.2	0.5	1.0
IN1K	73.4	75.1	75.2	74.9	68.3	63.2

(b) The **ratio** of negative class centers.

Positive Centers	1	2	4	8	16	32
IN1K	71.4	72.9	73.2	75.2	72.1	68.7

(c) The effect of **multi labels** per sample.

Positive Threshold	0.95	0.93	0.91	0.89	0.87	0.85
IN1K	72.2	72.7	73.3	72.4	68.7	63.2

(d) The effect of **positive thresholds**.

Table 6: Ablation experiments of the proposed contrastive loss decomposition. Pre-training is executed on the LAION-400M dataset by 32 epochs. The model backbone used here is ViT-B/32. Results are reported on the ImageNet validation dataset.

CASE	DATA	Finetune	Linear Probe	Zero Shot
MLC	LAION-400M	80.9	76.9	63.9
MLCD	LAION-400M	81.2	78.1	64.5

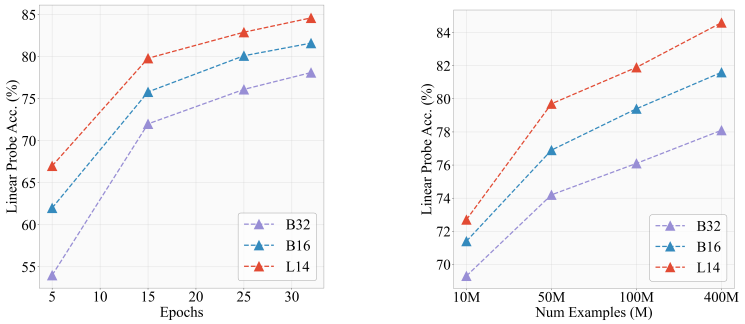
compared to that of using fixed assignment of positive centers (Tab. 5c). This indicates that the global similarity threshold is hard to search while the fixed assignment strategy benefits from the prior that the daily image statistically contains several visual concepts.

Effectiveness of MLCD Compared to MLC. In Tab. 6, we compare the performance of the vanilla MLC (Eq. 4) and the proposed MLCD (Eq. 5) on the ImageNet. Both MLC and MLCD employ the negative class center sampling with a ratio of 0.1. MLCD outperforms MLC in all three settings, confirming the effectiveness of the two additional optimization targets.

Scalability. In Fig. 5a and Fig. 5b, we validate the scalability of our method. Scaling up the ViT model and incorporating more data can significantly enhance our model's performance.

Effectiveness of MLCD on Different Training Data. In Tab. 7, we compare the linear probe performance of the proposed multi-label cluster discrimination approach (*i.e.*, MLCD) and the single-label cluster discrimination method (*e.g.*, UNICOM) on LAION-400M and COYO-700M. The hyper-parameter settings on COYO-700M follow the best settings on LAION-400M as explored in Tab. 5. As we can see from the results, the proposed MLCD consistently outperforms UNICOM by 2.2% and 1.6% when using LAION-400M and COYO-700M as the training data. In addition, the COYO-700M supports superior performance on action-related evaluation, achieving 3.3% improvement on Kinetics700 by using MLCD.

Effectiveness of MLCD in Vision Language Model. Tab. 8 compares the performance of replacing the vision tower in LLaVA-1.5 [26] from the CLIP model with our MLCD model. We validate the effectiveness of our MLCD under both Qwen2-7B and Qwen2-72B [1, 6] settings across 14 test datasets. To align



(a) Performance vs. Epochs **(b)** Performance vs. Num examples

Fig. 5: (a) The convergence curves of different ViTs. (b) The scalability curves of different ViTs under varying dataset scales. Larger ViTs and datasets lead to better model performance.

Table 7: Comparisons of linear probe performance across 26 different datasets for models trained on LAION-400M and COYO-700M datasets. Here, all methods employ the same backbone of ViT-B/32.

CASE	DATA	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Cal101	Flowers	MNIST	FER2013	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	K700	CLEVR	HM	SST	AVG
UNICOM	LAION-400M	85.8	96.8	86.6	70.2	74.6	93.3	70.7	88.3	78.0	93.1	94.6	98.5	98.7	64.3	97.8	96.8	90.6	90.0	76.4	22.5	82.9	84.2	57.2	52.6	52.4	62.1	79.2
MLCD	LAION-400M	87.8	97.5	88.2	72.4	77.6	93.8	71.4	91.9	80.4	93.2	96.9	98.8	99.3	66.4	98.6	98.6	92.1	90.5	77.7	30.9	83.4	86.3	60.9	54.1	57.9	70.4	81.4
UNICOM	COYO-700M	88.1	95.4	85.8	71.4	76.6	93.1	72.7	88.1	81.7	93.3	95.6	97.5	99.3	70.3	98.7	97.8	91.5	89.9	76.7	30.4	82.1	86.3	61.8	57.4	64.3	69.1	81.3
MLCD	COYO-700M	90.2	96.9	86.8	72.1	77.4	93.5	74.7	90.4	83.5	93.6	97.7	98.8	99.3	70.9	99.1	99.0	92.7	90.1	77.5	33.7	84.4	87.5	64.2	59.2	68.4	73.4	82.9

the experimental settings as in LLaVA-1.5, our model is fine-tuned for one epoch at a resolution of 336×336 after training at a resolution of 224×224. It can be observed that our method, MLCD, outperforms CLIP on most of the test datasets. However, there is a noticeable drop in performance on OCR-related benchmarks, such as TextVQA [37] and AI2D [21], under both 7B and 72B settings. To this end, we will incorporate additional OCR models for clustering to enhance our OCR capabilities in the future.

Semantic Visualization. In Fig. 6, we show the results of the Principal Component Analysis (PCA) performed on the patch features extracted by our MLCD model. We fine-tune our ViT-L/14 model on the LAION-400M dataset by one epoch using the resolution of 448×448. As the patch size is 14×14, we can obtain $32 \times 32 \times 1024$ spatial-wise tokens for each image. Then, we build a PCA projection from $32 \times 32 \times 1024$ to $32 \times 32 \times 3$. After we threshold the first component, we only keep patches with a positive value. As we can see from Fig. 6, the unsupervised foreground/background detector, based on detecting the highest variance direction, can separate the salient objects from the background. Afterward, we map the three PCA projection parameters into three different colors (*i.e.*, [R, G, B]). As shown in Fig. 6, objects from the same category exhibit color consistency, and objects from different categories present distinguishable colors,

Table 8: Evaluation of different visual towers (*i.e.*, CLIP and MLCD) used in VLM. The evaluation settings and test datasets align with LLaVA-1.5. The MLCD model (ViT-L/14) used here has employed training data from both LAION-400M and COYO-700M.

LLM	Vision Tower	VQAv2	GQA	VisWiz	SQA	TVQA	L-Wild	AI2D	MathV	HBI	MMMU	cMMMU	MMBench		SEED-Bench			MME	
		Val	Eval	Val	Img	Val	Test	Test	Mini	ALL	Val	Val	EN	CN	All	Img	Vid	Per	Cog
Qwen2-7B	CLIP	77.99	62.66	48.58	72.24	48.98	58.70	64.86	33.60	39.96	40.70	33.70	72.03	70.29	64.25	69.40	44.72	1512	335
Qwen2-7B	Ours	78.32	63.56	46.27	74.22	42.52	58.90	62.82	33.60	39.46	42.30	33.10	73.88	71.47	65.79	71.05	45.89	1558	384
Qwen2-72B	CLIP	79.47	63.81	67.14	76.10	62.31	65.41	72.41	38.30	45.10	39.70	37.45	76.63	75.39	66.54	72.28	44.71	1596	378
Qwen2-72B	Ours	79.51	66.80	67.37	74.69	57.32	66.00	71.41	46.5	45.21	44.70	41.20	78.59	77.24	68.67	76.53	45.91	1633	383

Fig. 6: PCA visualization of patch features extracted by our MLCD model. We fine-tuned the ViT-L/14 model on LAION-400M for one epoch at the resolution of 448×448, which allows each image to have 32×32 tokens for visualization. For each image, PCA is conducted on the extracted patch features to three principal components, which are subsequently normalized to the range of [0, 255] and mapped into the RGB space. Patches displaying similar colors indicate semantic similarities, reflecting that they embody analogous elements or attributes.

which indicates that the proposed multi-label cluster discrimination method can effectively capture multiple semantic signals from one image.

5 Conclusions

In this paper, we propose a novel multi-label cluster discrimination method to cope with multiple visual signals existing in one image. Compared to the vanilla version of the multi-label loss, which seeks to narrow the relative gap between inter-class similarities and intra-class similarities, our method introduces another two optimization targets (*i.e.*, decreasing inter-class similarities and increasing intra-class similarities) into the loss function. Introducing these two items enables the elegant separation of losses from positive and negative classes and alleviates the ambiguity on the decision boundary. Extensive experimental results show that the proposed multi-label cluster discrimination loss is effective for providing better transferrable features on multiple downstream tasks than both instance and cluster discrimination methods.

References

1. Qwen2 technical report (2024), https://qwenlm.github.io/blog/qwen2/ 12
2. Abdelfattah, R., Guo, Q., Li, X., Wang, X., Wang, S.: Cdul: Clip-driven unsupervised learning for multi-label image classification. In: ICCV (2023) 2
3. An, X., Deng, J., Guo, J., Feng, Z., Zhu, X., Yang, J., Liu, T.: Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In: CVPR (2022) 3, 7, 8
4. An, X., Deng, J., Yang, K., Li, J., Feng, Z., Guo, J., Yang, J., Liu, T.: Unicom: Universal and compact representation learning for image retrieval. In: ICLR (2023) 2, 3, 6, 7, 8
5. Asano, Y.M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. In: ICLR (2020) 2, 3
6. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv:2309.16609 (2023) 12
7. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV (2018) 2, 3, 6
8. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020) 2, 3
9. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: NeurIPS (2020) 3
10. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2021) 3
11. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: CVPR (2023) 3
12. Dao, T.: FlashAttention-2: Faster attention with better parallelism and work partitioning. In: ICLR (2024) 8
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 3, 10
14. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019) 7
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 3
16. Gu, T., Yang, K., An, X., Feng, Z., Liu, D., Cai, W., Deng, J.: Rwkv-clip: A robust vision-language representation learner. arXiv:2406.06973 (2024) 3
17. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022) 3
18. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020) 3
19. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021) 1, 3
20. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. IEEE Transactions on Big Data (2019) 6
21. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: ECCV (2016) 13

22. Li, J., Zhou, P., Xiong, C., Hoi, S.: Prototypical contrastive learning of unsupervised representations. In: ICLR (2020) 3
23. Li, M., Wang, D., Liu, X., Zeng, Z., Lu, R., Chen, B., Zhou, M.: Patchct: Aligning patch set and label set with conditional transport for multi-label image classification. In: ICCV (2023) 2
24. Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. In: CVPR (2023) 3, 9, 10
25. Li, Y., Song, Y., Luo, J.: Improving pairwise ranking for multi-label image classification. In: CVPR (2017) 7
26. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024) 12
27. Liu, W., Tsang, I.W., Müller, K.R.: An easy-to-hard learning paradigm for multiple classes and multiple labels. JMLR (2017) 4
28. Liu, W., Wang, H., Shen, X., Tsang, I.W.: The emerging trends of multi-label learning. TPAMI (2021) 4
29. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: ECCV (2018) 3
30. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al.: Mixed precision training. arXiv:1710.03740 (2017) 8
31. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016) 3
32. Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., Shrivastava, A.: Learning to predict visual attributes in the wild. In: CVPR (2021) 2
33. Qian, Q., Xu, Y., Hu, J., Li, H., Jin, R.: Unsupervised visual representation learning by online constrained k-means. In: CVPR (2022) 2, 3
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 1, 2, 3, 4
35. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: ICCV (2021) 4
36. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv:2111.02114 (2021) 8
37. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: CVPR (2019) 13
38. Singh, M., Gustafson, L., Adcock, A., de Freitas Reis, V., Gedik, B., Kosaraju, R.P., Mahajan, D., Girshick, R., Dollár, P., van der Maaten, L.: Revisiting weakly supervised pre-training of visual perception models. In: CVPR (2022) 3
39. Su, J., Zhu, M., Murtadha, A., Pan, S., Wen, B., Liu, Y.: Zlpr: A novel loss for multi-label classification. arXiv:2208.02955 (2022) 4
40. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: CVPR (2020) 2, 4, 7, 8
41. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (2007) 4
42. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: CVPR (2019) 4
43. Xia, X., Deng, J., Bao, W., Du, Y., Han, B., Shan, S., Liu, T.: Holistic label correction for noisy multi-label classification. In: ICCV (2023) 4

44. Yang, H., Tianyi Zhou, J., Zhang, Y., Gao, B.B., Wu, J., Cai, J.: Exploit bounding box annotations for multi-label object recognition. In: CVPR (2016) 2, 4
45. Yang, K., Deng, J., An, X., Li, J., Feng, Z., Guo, J., Yang, J., Liu, T.: Alip: Adaptive language-image pre-training with synthetic caption. In: ICCV (2023) 3
46. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: CVPR (2022) 3
47. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: CVPR (2022) 9
48. Zhan, X., Xie, J., Liu, Z., Ong, Y.S., Loy, C.C.: Online deep clustering for unsupervised representation learning. In: CVPR (2020) 2, 3
49. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. TKDE (2013) 4
50. Zhao, J., Yan, K., Zhao, Y., Guo, X., Huang, F., Li, J.: Transformer-based dual relation graph for multi-label image recognition. In: ICCV (2021) 4
51. Zhu, K., Fu, M., Wu, J.: Multi-label self-supervised learning with scene images. In: ICCV (2023) 2