

# Positive Divide and Negative Discrepancy: A New Perspective on Multi-Label Logit Distillation

Cong Li   Gong Cheng\*

School of Automation, Northwestern Polytechnical University, China

{conglee4017, chenggong1119}@gmail.com

## Abstract

Logit knowledge distillation (KD) has achieved competitive performance in the field of single-label learning. However, recklessly migrating the existing methods to the multi-label learning (MLL) encounters an intractable dilemma of either disturbing model optimization or crippling the effectiveness of logit KD, leading to compromised results. In this paper, we present a novel and powerful logit distillation approach for MLL dubbed 3D, consisting of the positive divide distillation (PDD) and negative discrepancy distillation (NDD). Specifically, PDD proposes to divide the positive labels and explicitly cast the single multi-label KD problem, per sample, as multiple single-label KD tasks, each playing a specific role in distilling dark knowledge from one positive label for students. Across images, NDD is devised to model and distill their predictive discrepancy over the intersection of the negative labels, encouraging the student to inherit the known inductive biases from the teacher. NDD also introduces an asymmetric scaling scheme to expand the discrepancy value ranges, thereby stimulating the supervision signals and guaranteeing better distillation. Accompanied by extensive experiments across MS-COCO, PASCAL-VOC, and NUS-WIDE, our 3D harvests compelling results (up to 4%-5% mAP gains) and consistently beats (by 1%-2% on average) the prior state-of-the-arts in both offline KD and self-KD. Our work unleashes the great potential of logit KD for MLL and provides a fresh insight for future research.

## 1. Introduction

Knowledge distillation (KD), debuted and popularized in single-label classification [10], is an attractive solution for obtaining a compact student model by absorbing knowledge from a large teacher model. Recently, applying KD to the more challenging yet realistic multi-label learning (MLL) tasks has received popularity [29, 30, 33, 34]. Among them, the primary research attention has been put on modeling

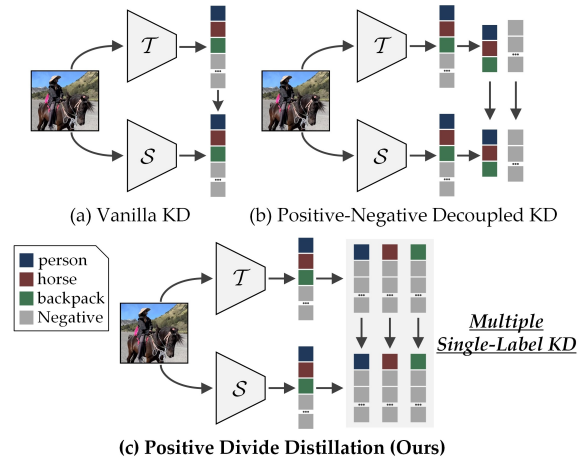


Figure 1. Comparison of the logit distillation schemes for MLL. The existing approaches typically execute distillation over the (a) coupled labels or (b) decoupled positive and negative labels, in contrast, (c) our PDD proposes the *positive divide* concept to cast the single multi-label KD problem as *multiple single-label KD* tasks, each of which individually transfers the dark knowledge for one positive label between the teacher and student.

the label correlations, but these intricate solutions usually induce high computational costs, methodology complexity, and heavily rely on the coarse label embeddings derived by peculiar training strategies (e.g., the class-aware attention or Transformer encoder [5, 13, 15, 18]). Logit KD [10], requiring marginal training overhead, is more generic and has achieved huge success in single-label learning [11, 28, 38]. However, logit KD only brings minor (see Tab. 1 and Tab. 7) gains in MLL [30, 34], and how to effectively distill logit in MLL is still unclear and under-exploited. In this paper, we, for the first time, undertake the task to unleash the great potential of logit distillation for MLL.

Typically, as shown in Fig. 1a, the vanilla logit KD [10] directly processes the logit in a coupled manner, leading to a *single multi-label KD* in MLL. Recently, related works [34] proposed to activate the logit with the `sigmoid` per label, aligning the optimization protocol of MLL. Such a scheme

\*Corresponding author

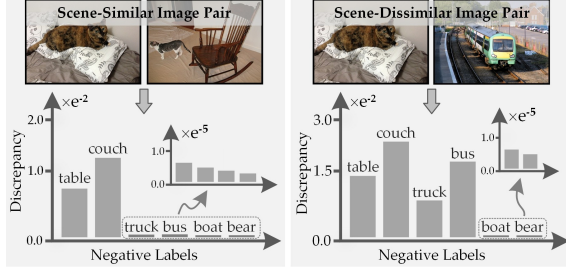


Figure 2. The motivation for NDD. For paired images, regardless of scene-similar or scene-dissimilar, their *predictive discrepancy* on the *negative labels* are distributed diversely and structurally, reflecting the model’s behaviors and inductive biases. Our NDD strives to distill this prior knowledge from the teacher into student.

is widely advocated and adopted in MLL-related works [22, 29, 34, 35, 37]. However, paradoxically, sigmoid misses the opportunity to utilize the innate logit relations among classes (e.g., inter-class discriminability), which are the primary attribute of logit KD [10, 11, 28, 38]. To make matters worse, this negative impact is exacerbated under current MLL networks, in which each class logit is generated through an independent classifier [5, 15, 18, 34, 41]. We also notice that there are methods [36, 38], *initially designed for the single-label learning*<sup>1</sup>, decoupled the target and non-target (i.e., the positive and negative in MLL) labels, as in Fig. 1b. Putting their careful treatments on the weights-tuning aside, however, due to the *exclusivity property* of softmax adopted in these methods, the KD loss tends to be dominated by *only one* positive label, leading to incomplete distillation in MLL and even hurting students (see Tab. 7). *It seems that the existing logit KD approaches face a dilemma of either disturbing the student optimization or hindering the distillation effectiveness in MLL.*

In this paper, we present a pure and novel logit distillation approach for MLL, consisting of two KD components. (i) Intuitively, all positive labels should be independently and correctly classified by a robust MLL model. This motivates us to cast the single multi-label distillation as several single-label distillation tasks, enabling the student to inherit the knowledge regarding how to classify each positive label from the teacher. Towards this goal, we propose positive divide distillation (PDD), as illustrated in Fig. 1c. In detail, PDD frames each single-label KD with the corresponding positive label and all negative labels, and accordingly performs distillation in a softmax driven manner. By doing so, the learning signals from the teacher side explicitly describe the innate logit relations between labels, enhancing the student’s inter-label differences and unleashing the potential of logit KD for MLL. Particularly, we find that PDD brings a desired effect in MLL [6, 13, 20, 31] that increases

the positive-negative classification margins, hence enhancing the student’s discriminability and performance.

(ii) We further conceive a cross-image negative discrepancy distillation (NDD). NDD stems from a well-justified fact that the positive labels have the causal effect on the predictions of negative labels [5, 15, 17, 18, 30]. Since different images contain diverse positive labels, delving deeper into this effect, we discover that the *predictive discrepancy of the negative label between paired images* is distributed diversely and structurally across labels. For example, as depicted in Fig. 2 (right), the exclusively context- or visual-relevant *negative labels* (e.g., *couch* for an indoor scene and *bus* for a city-street scene) generally show more significant predictive discrepancy across images, while the discrepancy on the *negative labels* (say *boat* and *bear*) irrelevant to both scenes is fairly small. Such known inductive dependencies contain the prior knowledge of models’ decision-making and should be exploited for distilling students.

For concrete realizations, NDD computes the predictive residual component over the intersection of *negative labels of a paired images*. In addition, considering the cross-image discrepancy value of the negative labels is often small (see Fig. 2), we introduce an asymmetric scaling scheme, subject to the label-specific relative predictive values across images, to amplify the value ranges of discrepancy and deliver better distillation eventually. With loss reformulation, we derive an upper bound of NDD that is equivalent to a composite of an intra-image mean square error and an extra regularization exerted across images, thereby aiding the student in shaping its whole predictive space towards the teacher.

Overall, we dub the proposed approach positive divide and negative discrepancy distillation (3D), which is super efficient, performant, and generic. To evaluate 3D, we carry out extensive experiments over three benchmark datasets, i.e., MS-COCO [16], NUS-WIDE [7], and PASCAL-VOC [8] spanning across various teacher-student KD combinations. Distillation results evidence the effectiveness (boosting the students by 4%~5% in mAP) and superiority (e.g., beats MDKD [29] by 1%~2% on average) of our method. To sum up, we make the following contributions:

- We introduce 3D, a novel approach consisting of two distillation terms PDD and NDD that are respectively built upon the idea of *positive divide* and *negative discrepancy*, offering a new perspective on multi-label logit KD.
- By dividing the positive labels, PDD formulates multiple single-label KD tasks per sample, which breaks the bottleneck of the traditional single multi-label KD methods.
- The NDD measures the negative predictive discrepancy across images and reveals its great potential to serve as a profitable knowledge pattern for distilling students.
- Our 3D achieves superiority results and consistently surpasses the previous works by a large margin, establishing a new state-of-the-art on both offline KD and self-KD.

<sup>1</sup>For a thorough comparison, we extend their ideas to the MLL context.

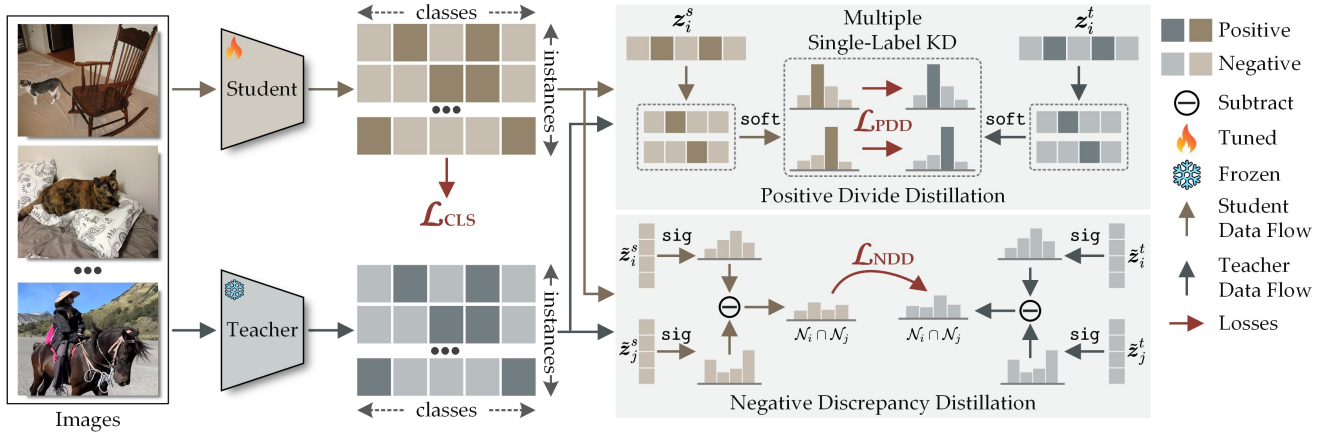


Figure 3. Overview of the proposed 3D for MLL, which is composed of two logit distillation losses: Positive Divide Distillation ( $\mathcal{L}_{PDD}$ ) and Negative Discrepancy Distillation ( $\mathcal{L}_{NDD}$ ).  $\mathcal{L}_{CLS}$  is the task classification loss, supervised by the ground-truth.

## 2. Related Work

### 2.1. Multi-Label Learning

Multi-label learning (MLL) has received ever-growing attention from research communities [4–6, 25, 29, 30, 34]. Representative MLL approaches typically fall into three categories: loss re-balancing [25], label relation modeling [4], and label-specific embeddings decoupling [15]. Specifically, ASL [25] assigned different weights to re-balance the loss contribution from the positive and negative samples, and TWL [13] introduced a scheme to compute the loss not only at each sample but also for each class. For label relations modeling, some efforts utilize multi-source prior information, *i.e.*, processing the label-occurrence graph with a GCN network [4]. Recently, the performant solutions aim to extract label embedding for each semantic class *per sample*, by, for instance, using class attention maps [18], textual features [41], or a Transformer encoder [15]. In this paper, we investigate KD under the MLL context.

### 2.2. Multi-Label Knowledge Distillation

Knowledge distillation (KD), debuted and popularized in the single-label learning task, aims to produce a compact yet powerful model by inheriting knowledge from a high-capacity one [10, 28, 38]. Regarding what to distill, the existing works can be roughly divided into two groups: logit-based [10, 11, 36, 39] and feature-based [2, 3, 14, 23, 40]. The former executes KD with the model predictions, and the latter extracts knowledge from the intermediate features.

There are several papers that have studied KD under the MLL context [24, 27, 29, 30, 32, 34]. Specifically, UD [27] leveraged the predictive uncertainty and calibrated model to handle the difficult labels. To enhance the logit KD in MLL, a rising idea in [34] is to perform multi-label KD through a set of binary classification distillation problems, which is

also adopted in [29]. However, as discussed above, such a logit KD scheme is still problematic for disturbing the models training in MLL and the lack of dark knowledge utilization. The label correlation is also exploited for multi-label distillation. L2D [34] distilled the student with intra-class and inter-class structural knowledge with the label-wise embeddings. Building upon this, MDKD [29] further constructed multi-order (from second-order to sixth-order) label pair dependencies for distillation. Those methods only consider the positive labels and rely on peculiar networks, leading to trivial improvement and limited generalization. Recently, MSKD [30] made the first attempt to conduct KD without a teacher model (*i.e.*, self-KD [26, 36]) in MLL, which proposed to extract the local knowledge from cropped patches and achieved balanced distillation with rectified KL-divergence. Unlike the existing works, this paper formulates multi-label KD as multiple single-label KD and probes the cross-image causal discrepancy for distillation, spawning a novel and general logit KD method.

## 3. Methodology

### 3.1. Preliminary

**Notations.** Given an image  $\mathbf{x}$ , we indicate its corresponding label  $\mathbf{y} = [y_1, y_2, \dots, y_C] \in \mathbb{R}^C$  form  $C$  classes, where  $y_i = 1$  indicates that  $\mathbf{x}$  has objects of the  $i$ -th class (*i.e.*, positive label), and  $y_i = 0$  otherwise (*i.e.*, negative label). For a specific image, we denote the class index set of associated positive labels as  $\mathcal{P}$  and negative labels as  $\mathcal{N}$ . We have  $\mathcal{P} \cap \mathcal{N} = \emptyset$ ,  $1 \leq |\mathcal{P}| < C$ , and  $|\mathcal{P} \cup \mathcal{N}| = C$ , where  $|\cdot|$  returns the number of elements in a set. We tag the logit as  $\mathbf{z} = [z_1, z_2, \dots, z_C] \in \mathbb{R}^{1 \times C}$  with  $z_i$  as the logit of the  $i$ -th class. With an activation function,  $\mathbf{z}$  is further converted to the predictive probability  $\mathbf{p} \in \mathbb{R}^{1 \times C}$ . We use the superscripts  $s$  and  $t$  to discriminate the student and teacher.

**Existing Logit KD for MLL.** Formally, the existing logit KD approaches minimize the following objective:

$$\mathcal{L}_{\text{KD}} = \ell(\delta(\mathbf{z}^s), \delta(\mathbf{z}^t)) = \ell(\mathbf{p}^s, \mathbf{p}^t) \quad (1)$$

where  $\ell(\cdot)$  and  $\delta(\cdot)$  refer to the loss and activation function, respectively. In classical KD [10],  $\ell(\cdot)$  is the KL-divergence ( $\ell_{\text{KL}}$ ) and  $\delta$  is softmax:

$$\mathbf{p}_i = \text{softmax}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{c=1}^C e^{z_c}} \quad (2)$$

In [29, 34], the authors proposed multi-label logit distillation (MLD) to compute the KD loss with the Binary Cross Entropy and activated the logit with sigmoid:

$$\mathbf{p}_i = \text{sig}(z_i) = \frac{1}{1 + e^{-z_i}} \quad (3)$$

### 3.2. Positive Divide Distillation

To effectively exploit the knowledge from all positive labels, in this paper, we cast the single multi-label KD as multiple single-label KD tasks, each of which handles one positive label during distillation. The outcome of our design is a novel positive divide distillation (PDD). Specifically, as in Fig. 3, PDD re-formulates the per-sample logit outputs  $\mathbf{z} \in \mathbb{R}^{1 \times C}$  into a group of logits  $\{\tilde{\mathbf{z}}_k \in \mathbb{R}^{1 \times (1+|\mathcal{N}|)}\}_{k=1}^{|\mathcal{P}|}$ , in which each  $\tilde{\mathbf{z}}_k$  consists of one positive logit  $z_i$  ( $i \in \mathcal{P}$ ) and all negative logits  $\{z_j; j \in \mathcal{N}\}$ :

$$\begin{aligned} \tilde{\mathbf{z}}_k^s &= [z_{\mathcal{P}[k]}^s, z_{\mathcal{N}[1]}^s, z_{\mathcal{N}[2]}^s, \dots, z_{\mathcal{N}[|\mathcal{N}|]}^s] \\ \tilde{\mathbf{z}}_k^t &= [z_{\mathcal{P}[k]}^t, z_{\mathcal{N}[1]}^t, z_{\mathcal{N}[2]}^t, \dots, z_{\mathcal{N}[|\mathcal{N}|]}^t] \end{aligned} \quad (4)$$

where the subscripts  $\mathcal{P}[l]$  and  $\mathcal{N}[l]$  return the  $l$ -th element in the set of  $\mathcal{P}$  and  $\mathcal{N}$ , respectively. Then, we apply softmax on  $\tilde{\mathbf{z}}_k^s$  and  $\tilde{\mathbf{z}}_k^t$  to explicitly capture the innate logit relations among classes, and generate the corresponding probability:

$$\begin{aligned} \bar{\mathbf{p}}_{k,i}^s &= \text{softmax}(\tilde{\mathbf{z}}_k^s; \tau)_i = \frac{e^{\tilde{z}_k^s / \tau}}{e^{z_{\mathcal{P}[k]}^s / \tau} + \sum_{l=1}^{|\mathcal{N}|} e^{z_{\mathcal{N}[l]}^s / \tau}} \\ \bar{\mathbf{p}}_{k,i}^t &= \text{softmax}(\tilde{\mathbf{z}}_k^t; \tau)_i = \frac{e^{\tilde{z}_k^t / \tau}}{e^{z_{\mathcal{P}[k]}^t / \tau} + \sum_{l=1}^{|\mathcal{N}|} e^{z_{\mathcal{N}[l]}^t / \tau}} \end{aligned} \quad (5)$$

where  $i \in [1, 1+|\mathcal{N}|]$  is the index in  $\bar{\mathbf{p}}_k^s$  and  $\bar{\mathbf{p}}_k^t$ , and  $\tau$  is the commonly used temperature factor. Then, by minimizing the most plainest  $\ell_{\text{KL}}(\cdot)$  between the teacher and student, we express the PDD loss as:

$$\mathcal{L}_{\text{PDD}} = \frac{1}{|\mathcal{P}|} \sum_{k=1}^{|\mathcal{P}|} \ell_{\text{KL}}(\bar{\mathbf{p}}_k^s, \bar{\mathbf{p}}_k^t) \quad (6)$$

Note that,  $\ell_{\text{KL}}(\cdot)$  is our default choice, and PDD could be further boosted by cooperating with more advanced distillation schemes developed in the single-label learning task, which is an appealing advantage of the proposed method.

**Enhancing Classification Margin.** (i) One may reasonably argue that the classical KD [10] also models the innate relations between labels. However, applying softmax in the global logit (see Eq. (2)) diminishes the logit discrimination in the MLL context. Specifically, with Eq. (2), the teacher's predictive probability difference of a label pair  $(i, j)$  is:

$$\Delta_{\text{KD}} = \left| \frac{e^{z_i^t} - e^{z_j^t}}{\sum_{l \in \mathcal{P}} e^{z_l^t} + \sum_{l \in \mathcal{N}} e^{z_l^t}} \right| \quad (7)$$

The corresponding difference value of our PDD is:

$$\Delta_{\text{PDD}} = \left| \frac{e^{z_i^t} - e^{z_j^t}}{e^{z_{\mathcal{P}[k]}^t} + \sum_{l \in \mathcal{N}} e^{z_l^t}} \right| \quad (8)$$

Clearly, we have  $e^{z_{\mathcal{P}[k]}^t} \ll \sum_{l \in \mathcal{P}} e^{z_l^t}$ , and a fairly small  $\sum_{l \in \mathcal{N}} e^{z_l^t}$ . Then, comparing Eq. (7) and Eq. (8), it is obvious that  $\Delta_{\text{PDD}} \gg \Delta_{\text{KD}}$ , indicating that our approach enhances the inter-label differences compared to the vanilla KD. Despite extremely sparse, *if and only if*  $|\mathcal{P}| = 1$ , i.e., *this sample only contains one positive label*,  $\Delta_{\text{PDD}} = \Delta_{\text{KD}}$ , in which cases our PDD also degenerates to KD.

(ii) Now, considering a special case of the label pair  $(i, j)$  with  $\forall i \in \mathcal{P}, \forall j \in \mathcal{N}$ , our PDD meets one advocated [1, 6, 13, 20] learning goal in MLL:

$$p_i > p_j, \text{ s.t., } \forall i \in \mathcal{P}, \forall j \in \mathcal{N}, \quad (9)$$

and stronger models (the teacher model in our case) tend to show higher positive-negative classification margins, i.e.,  $\Delta = \min_{i \in \mathcal{P}} p_i - \max_{j \in \mathcal{N}} p_j$ . As in Eq. (4), in our PDD, *each positive label is assigned to compete with all the negative labels*, consequently, PDD implicitly regulates the student to mimic such a desired property from the teacher, improving the discriminative ability and performance.

### 3.3. Negative Discrepancy Distillation

It is well-known that the response scores of the negative labels are heavily conditioned on the positive labels [5, 15]. Since each image contains different positive labels, the predictive discrepancy of the negative label would be diverse and structurally, as in Fig. 2, revealing the models' behaviors and inductive biases. In this paper, we encourage the student to distill such prior knowledge from the teacher, further shaping their whole predictive space.

Formally, given paired images  $(\mathbf{x}_1, \mathbf{x}_2)$ , we denote  $\tilde{\mathcal{N}} = \mathcal{N}_1 \cap \mathcal{N}_2$ , indicating the intersection of corresponding  $(\mathcal{N}_1, \mathcal{N}_2)$ . Then, we get the corresponding logit pair  $(\tilde{\mathbf{z}}_1 \in \mathbb{R}^{1 \times |\tilde{\mathcal{N}}|}, \tilde{\mathbf{z}}_2 \in \mathbb{R}^{1 \times |\tilde{\mathcal{N}}|})$  by retrieving  $\tilde{\mathcal{N}}$ , and define the predictive discrepancy between  $\tilde{\mathbf{z}}_1$  and  $\tilde{\mathbf{z}}_2$  as  $\boldsymbol{\rho} \in \mathbb{R}^{1 \times |\tilde{\mathcal{N}}|}$ :

$$\boldsymbol{\rho}_{i,\gamma} = \text{sig}\left(\frac{\tilde{z}_{1,i}}{\gamma}\right) - \text{sig}(\tilde{z}_{2,i}) \quad (10)$$

$$\gamma = \begin{cases} \gamma^+, & \tilde{z}_{1,i} > \tilde{z}_{2,i} \\ \gamma^-, & \tilde{z}_{1,i} < \tilde{z}_{2,i} \end{cases} \quad (11)$$

where  $\gamma > 0$  is an adaptive scaling factor to adjust the distribution, and we provide its detailed discussions in later. Then, on the basis of Eq. (11), we analogously build  $\rho^s$  using  $(\tilde{z}_1^s, \tilde{z}_2^s)$  for the student, and  $\rho^t$  using  $(\tilde{z}_1^t, \tilde{z}_2^t)$  for the teacher. As a result, the distillation loss is written as:

$$\mathcal{L}_{\text{NDD}} = \sum_{i=1}^B \sum_{j=i+1}^B \ell_{\text{MSE}}(\rho_{i,j}^s, \rho_{i,j}^t) \quad (12)$$

where  $B$  is batch size and  $\ell_{\text{MSE}}(\cdot)$  is the Mean Square Error. **Selection of  $\gamma^+$  and  $\gamma^-$ .** As in Fig. 2, we observe that the cross-image prediction discrepancy of the negative label is generally small, suggesting that the directly measured (*i.e.*,  $\gamma = 1.0$  in Eq. (11))  $\rho$  are not guaranteed to provide discriminative guidance for the students. To deal with this issue, we introduce the above label-specific asymmetric scaling strategy, subject to the paired relative predictive values (see Eq. (11)), to amplify the absolute value of prediction discrepancy. Formally, our goal is to always have:

$$|\rho_{i,\gamma \neq 1.0}| > |\rho_{i,\gamma = 1.0}| \quad (13)$$

**Assumption 3.1** *In the training split<sup>2</sup>, the well-converged teacher model could give high-confidence and correct predictions for negative labels, *i.e.*,  $\tilde{z}_i^t < 0$ ,  $i \in [1, |\tilde{\mathcal{N}}|]$ .*<sup>3</sup>

**Lemma 3.2** *Under Assumption 3.1,  $\text{sig}(\tilde{z}_i/\gamma)$  monotonically increases when  $\gamma > 1.0$ , and monotonically decreases when  $\gamma < 1.0$ . Specifically, as  $\gamma \rightarrow \infty$ ,  $\text{sig}(\tilde{z}_i/\gamma) \rightarrow 0.5$ , and as  $\gamma \rightarrow 0.0$ ,  $\text{sig}(\tilde{z}_i/\gamma) \rightarrow 0.0$ .*

We omit the simple proof of Lemma 3.2. Then, according to Lemma 3.2, we need to set  $\gamma^+ > 1.0$  and  $\gamma^- < 1.0$  to meet the requirement outlined in Eq. (13). Fig. 4 gives the visual illustration for a better understanding.

In practice, to reduce the complexity of tuning hyperparameters, we unify  $\gamma^+$  and  $\gamma^-$  with a base  $\gamma_b$  as follows, and it works surprisingly well.

$$\gamma^+ = \gamma_b, \quad \gamma^- = \frac{1}{\gamma_b}, \quad \text{s.t., } \gamma_b > 1.0 \quad (14)$$

**Loss Reformulation.** Here<sup>4</sup>, with  $\tilde{\mathbf{p}}_1 = \text{sig}(\tilde{z}_1) > 0$  and  $\tilde{\mathbf{p}}_2 = \text{sig}(\tilde{z}_2) > 0$ , we mathematically have:

$$\begin{aligned} \mathcal{L}_{\text{NDD}} &< \ell_{\text{MSE}}(\tilde{\mathbf{p}}_1^s, \tilde{\mathbf{p}}_1^t) + \ell_{\text{MSE}}(\tilde{\mathbf{p}}_2^s, \tilde{\mathbf{p}}_2^t) \\ &\quad + \underbrace{2(\tilde{\mathbf{p}}_1^s \cdot \tilde{\mathbf{p}}_2^t + \tilde{\mathbf{p}}_2^s \cdot \tilde{\mathbf{p}}_1^t)}_{\mathcal{L}_{\text{aux}}} \end{aligned} \quad (15)$$

<sup>2</sup>We highlight that the training split is the one used for distillation.

<sup>3</sup>Assumption 3.1 is rational and legitimate in KD. As reported in the *Supplementary Material*, the teacher (ResNet101) achieves  $> 97\%$  accuracy and  $< 0.03$  averaged predictive scores for the negative labels, on both MS-COCO [16], PASCAL-VOC [8], and NUS-WIDE [7].

<sup>4</sup>More detailed derivation of Eq. (15) is in the *Supplementary Material*, and we omit  $\gamma_b$  for simplicity, without loss of generality.

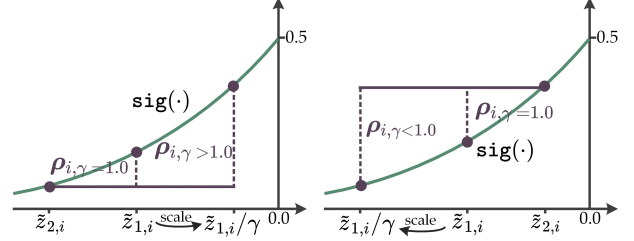


Figure 4. Impact of  $\gamma$ . Left: In case of  $\tilde{z}_{1,i} > \tilde{z}_{2,i}$ , we need  $\gamma > 1.0$ . Right: In the case of  $\tilde{z}_{1,i} < \tilde{z}_{2,i}$ , we need  $\gamma < 1.0$ .

With this derivation, one could observe that the upper bound of  $\mathcal{L}_{\text{NDD}}$  is equivalent to the composite of intra-image predictive imitation and an auxiliary regularization  $\mathcal{L}_{\text{aux}}$  exerted across images. In particular, since  $\tilde{\mathbf{p}}_1^s, \tilde{\mathbf{p}}_1^t, \tilde{\mathbf{p}}_2^s$ , and  $\tilde{\mathbf{p}}_2^t$  are *always greater than 0*, the  $\mathcal{L}_{\text{aux}}$  explicitly enforces the student to **decrease** (*i.e.*, approaching 0, to be more specifically) the predictions of the negative labels, thereby producing more discriminative students after distillation.

Moreover, the peculiar behavior of  $\mathcal{L}_{\text{aux}}$  also explains why our NDD only selects the negative labels for distillation, because the model in MLL is trained to **increase** the predictions of the positive labels, violating the optimization direction of  $\mathcal{L}_{\text{aux}}$  and leading to trivial (see Tab. 8) results.

### 3.4. Student Optimization

Overall, as in Fig. 3, the training loss of the student is:

$$\mathcal{L} = \mathcal{L}_{\text{CLS}} + \alpha \mathcal{L}_{\text{PDD}} + \beta \mathcal{L}_{\text{NDD}} \quad (16)$$

where  $\mathcal{L}_{\text{CLS}}$  is the original classification loss supervised by the ground truth, and  $\alpha$  and  $\beta$  are the KD weights.

## 4. Experiments

### 4.1. Setup

Our experiments are conducted three standard datasets, *i.e.*, MS-COCO [16], NUS-WIDE [7], and PASCAL-VOC [8], over the offline KD and self-KD. We report the mean average precision (mAP) over all classes, average per-class F1-score (CF1), overall F1-score (OF1), and the class-wise AP to evaluate the performance.

All the experimental settings are strictly aligned with prior works [29, 30, 34], and we attach the details in the *Supplementary Material* due to the page limit.

### 4.2. Comparison to State-of-the-Arts

**Results on Offline KD.** We compare the proposed approach 3D with state-of-the-art KD methods over several teacher-student pairs. Tab. 1 and Tab. 2 summarizes the performance on MS-COCO [16], and Tab. 3 lists the results NUS-WIDE [7]. It is observed that the distilled students harvest outstanding  $4\% \sim 5\%$  absolute

Table 1. Comparison KD performance on MS-COCO [16]. Teacher and student are in the same architectures.

| Method        | ResNet101 → ResNet34 |              |              | WRN101 → WRN50 |              |              | VGG16 → VGG11 |              |              | SwinS → SwinT |              |              |
|---------------|----------------------|--------------|--------------|----------------|--------------|--------------|---------------|--------------|--------------|---------------|--------------|--------------|
|               | mAP                  | OF1          | CF1          | mAP            | OF1          | CF1          | mAP           | OF1          | CF1          | mAP           | OF1          | CF1          |
| Teacher       | 73.62                | 73.89        | 68.61        | 74.70          | 75.56        | 70.73        | 72.71         | 74.11        | 68.63        | 81.70         | 80.48        | 77.12        |
| Student       | 70.31                | 72.49        | 66.82        | 74.45          | 75.43        | 70.61        | 70.02         | 72.49        | 66.77        | 79.59         | 79.18        | 75.42        |
| RKD [23]      | 70.13                | 72.44        | 66.78        | 74.70          | 75.71        | 70.84        | 70.13         | 72.39        | 66.73        | 79.59         | 79.18        | 75.42        |
| PKT [24]      | 70.43                | 72.64        | 66.68        | 74.54          | 75.47        | 70.58        | 70.11         | 72.47        | 66.80        | 79.64         | 79.09        | 75.39        |
| Review KD [3] | 70.39                | 72.62        | 66.76        | 74.03          | 75.29        | 70.36        | 70.00         | 72.35        | 66.82        | 79.81         | 79.18        | 75.55        |
| MSE [32]      | 70.54                | 72.75        | 66.85        | 74.53          | 75.60        | 70.71        | 70.26         | 72.54        | 66.99        | 79.67         | 79.20        | 75.25        |
| PS [27]       | 70.86                | 72.66        | 67.12        | 75.12          | 76.05        | 71.63        | 70.65         | 72.89        | 67.60        | 79.96         | 79.64        | 76.62        |
| L2D [34]      | 72.87                | 74.45        | 69.43        | 76.61          | 77.08        | 72.79        | 72.81         | 74.59        | 69.49        | 81.59         | 81.03        | 77.86        |
| MDKD [29]     | 73.35                | 74.13        | 68.89        | 77.01          | 77.39        | 72.95        | 72.53         | 74.24        | 69.17        | 82.33         | 82.17        | 77.56        |
| MLD           | 70.68                | 72.69        | 67.19        | 74.92          | 75.75        | 71.21        | 70.74         | 72.81        | 67.46        | 80.11         | 79.68        | 76.44        |
| 3D (Ours)     | <b>74.80</b>         | <b>75.84</b> | <b>70.75</b> | <b>77.69</b>   | <b>77.94</b> | <b>73.37</b> | <b>73.97</b>  | <b>75.37</b> | <b>70.23</b> | <b>83.16</b>  | <b>82.11</b> | <b>79.27</b> |

Table 2. Comparison KD performance on MS-COCO [16]. Teacher and student are in the different architectures.

| Method        | ResNet101 → MBV2 |              |              | SwinT → ResNet34 |              |              | ResNet101 → VGG16 |              |              | SwinT → MBV2 |              |              |
|---------------|------------------|--------------|--------------|------------------|--------------|--------------|-------------------|--------------|--------------|--------------|--------------|--------------|
|               | mAP              | OF1          | CF1          | mAP              | OF1          | CF1          | mAP               | OF1          | CF1          | mAP          | OF1          | CF1          |
| Teacher       | 73.62            | 73.89        | 68.61        | 79.43            | 78.77        | 75.07        | 73.62             | 73.89        | 68.61        | 74.08        | 75.05        | 72.20        |
| Student       | 71.85            | 73.68        | 68.40        | 70.31            | 72.49        | 66.82        | 70.02             | 72.49        | 66.77        | 71.85        | 73.68        | 68.40        |
| RKD [23]      | 71.76            | 73.68        | 68.40        | 70.00            | 72.34        | 66.64        | 70.08             | 72.35        | 66.72        | 72.48        | 73.48        | 67.48        |
| PKT [24]      | 71.88            | 73.60        | 68.35        | 69.99            | 72.35        | 66.56        | 70.11             | 72.47        | 66.80        | 71.45        | 73.08        | 68.43        |
| Review KD [3] | 71.92            | 73.73        | 68.48        | 70.29            | 72.39        | 66.58        | 70.00             | 72.33        | 66.62        | 72.49        | 73.32        | 69.18        |
| MSE [32]      | 71.91            | 73.68        | 68.28        | 70.33            | 72.57        | 66.72        | 70.07             | 72.50        | 66.85        | 72.00        | 73.90        | 68.87        |
| PS [27]       | 72.11            | 73.89        | 68.42        | 70.94            | 72.93        | 67.57        | 70.30             | 72.61        | 67.10        | 71.43        | 72.98        | 68.78        |
| L2D [34]      | 73.24            | 74.85        | 69.72        | 73.42            | 74.97        | 70.20        | 72.14             | 74.08        | 68.78        | 71.05        | 73.98        | 68.18        |
| MDKD [29]     | 73.40            | 74.45        | 69.41        | 74.01            | 74.99        | 68.49        | 72.65             | 74.66        | 68.42        | 73.98        | 74.89        | 70.80        |
| MLD           | 72.17            | 73.84        | 68.52        | 71.14            | 72.99        | 67.63        | 70.48             | 72.77        | 67.10        | 71.55        | 72.76        | 68.79        |
| 3D (Ours)     | <b>74.05</b>     | <b>75.41</b> | <b>70.26</b> | <b>75.37</b>     | <b>76.49</b> | <b>71.83</b> | <b>73.72</b>      | <b>75.15</b> | <b>70.18</b> | <b>74.63</b> | <b>75.93</b> | <b>71.29</b> |

improvements, which even outstrip the teachers in most cases. More importantly, our 3D consistently outperforms those sophisticatedly-designed methods (*i.e.*, L2D [34] and MDKD [29]) by considerable margins (up to 1%  $\sim$  2% in mAP) and establishes a new state-of-the-art in the multi-label logit KD community, strongly demonstrating the superiority and effectiveness of the proposed method. *These results verify that our 3D could serve as a new baseline for future logit distillation in MLL.* More results on PASCAL-VOC [8] are attached in the *Supplementary Material*.

**Results on Self-KD.** Following MSKD [30], we apply our 3D to self-KD, where the students distill knowledge from themselves progressively. As reported in Tab. 4, once again, our 3D could deliver compelling performance in the setting of self-KD, and substantially surpasses its previous counterparts in all evaluation metrics. For example, 3D beats the state-of-the-art MSKD [30] by 3.15% and 5.40% in mAP over ResNet34 [9] and SwinTiny [19], respectively.

**Training Efficiency.** We highlight that 3D is more practical than the prior carefully designed methods [29, 34]. As com-

Table 3. Comparison KD performance on NUS-WIDE [7].

| Method        | ResNet101 → ResNet34 |              |              | SwinT → MBV2 |              |              |
|---------------|----------------------|--------------|--------------|--------------|--------------|--------------|
|               | mAP                  | OF1          | CF1          | mAP          | OF1          | CF1          |
| Teacher       | 55.32                | 75.56        | 61.13        | 59.73        | 77.30        | 65.44        |
| Student       | 53.41                | 75.10        | 60.08        | 54.49        | 75.72        | 61.74        |
| RKD [23]      | 53.62                | 75.20        | 59.91        | 54.76        | 75.69        | 61.74        |
| PKT [24]      | 53.55                | 75.08        | 60.35        | 54.59        | 75.69        | 61.74        |
| Review KD [3] | 53.52                | 75.23        | 60.44        | 54.85        | 75.84        | 61.75        |
| MSE [32]      | 53.52                | 75.13        | 59.94        | 54.86        | 75.80        | 61.69        |
| PS [27]       | 54.14                | 75.43        | 60.79        | 55.18        | 75.91        | 62.35        |
| L2D [34]      | 55.31                | 76.17        | 62.79        | 56.91        | 76.92        | 63.89        |
| MDKD [29]     | 55.08                | 76.15        | 62.58        | 58.46        | 76.69        | 63.44        |
| MLD           | 54.44                | 75.36        | 60.73        | 55.36        | 76.00        | 62.52        |
| 3D (Ours)     | <b>57.59</b>         | <b>76.71</b> | <b>64.24</b> | <b>58.70</b> | <b>77.32</b> | <b>64.67</b> |

pared in Tab. 5, our 3D requires the lowest training overhead. This is because our 3D only distills with the logits, without relying on the complex and multi-order label dependencies across batch samples and classes.

Table 4. Self-KD performance on MS-COCO [16]. The results are obtained with the training configurations in MSKD [30].

| Method    | ResNet34     |              |              | SwinT        |              |              |
|-----------|--------------|--------------|--------------|--------------|--------------|--------------|
|           | mAP          | OF1          | CF1          | mAP          | CF1          | OF1          |
| Student   | 66.26        | 59.78        | 65.99        | 71.94        | 64.98        | 68.24        |
| DLB [26]  | 65.25        | 56.53        | 64.52        | 70.77        | 62.54        | 68.59        |
| USKD [36] | 61.06        | 54.28        | 66.48        | 70.54        | 64.70        | 58.62        |
| UD [27]   | 63.71        | 58.50        | 59.27        | 70.86        | 64.77        | 69.62        |
| MSKD [30] | 67.02        | 63.84        | 67.19        | 73.62        | 68.18        | 71.61        |
| 3D (Ours) | <b>70.17</b> | <b>67.03</b> | <b>72.35</b> | <b>79.02</b> | <b>79.36</b> | <b>75.11</b> |

Table 5. Comparison of training efficiency. The training time is measured per batch and #Num. is number of Hyper-parameters.

| Method    | #Num. | Training time (s) | GPU Memory (G) | mAP          |
|-----------|-------|-------------------|----------------|--------------|
| L2D [34]  | 3     | 1.23              | 7.12           | 72.87        |
| MDKD [29] | 10    | 1.45              | 8.39           | 73.35        |
| 3D (Ours) | 4     | 1.08              | 6.99           | <b>74.80</b> |

Table 6. Effect of each components over MS-COCO [16]. The teacher and student are ResNet101 and ResNet34.

| $\mathcal{L}_{CLS}$ | $\mathcal{L}_{PDD}$ | $\mathcal{L}_{NDD}$ | mAP          | CF1          | OF1          |
|---------------------|---------------------|---------------------|--------------|--------------|--------------|
| ✓                   |                     |                     | 70.31        | 72.49        | 66.82        |
| ✓                   | ✓                   |                     | 74.13        | 75.30        | 70.51        |
| ✓                   |                     | ✓                   | 73.45        | 74.79        | 69.79        |
| ✓                   | ✓                   | ✓                   | <b>74.80</b> | <b>75.84</b> | <b>70.75</b> |

### 4.3. Diagnostic Experiment

**Distillation Objectives.** We examine the individual effectiveness of each distillation term on MS-COCO [16], with the ResNet101-ResNet34 KD pair. As reported in Tab. 6, the following observations could be made. *(i)* Both  $\mathcal{L}_{PDD}$  and  $\mathcal{L}_{NDD}$  greatly improve the student by 3.82% and 3.14% in mAP, respectively. Those results demonstrate that our new perspective of positive divide and negative discrepancy certainly brings practical benefits and their advantages for distilling logit in MLL. *(ii)* Remarkably, only applying PDD (74.13%) or NDD (73.45%), the results already surpass the recent strong works L2D (72.87%) [34] and MDKD (73.35%) [29], further confirming the superiority of our methods. *(iii)* By synergizing these two components, our 3D significantly stimulates the student performance and achieves the best distillation results, indicating that the proposed PDD and NDD complement each other. We ascribe this phenomenon to the fact that PDD and NDD, respectively, focus on the distillation of positive and negative labels, hence fulfilling comprehensive knowledge transferring for students and leading to excellent distillation results.

**Ablative Studies on PDD.** *(i)* We notice that, in *single-label learning*, there are works (DKD [38] and NKD [36]) to divide the logits into two parts: the target one and non-

Table 7. Ablative studies on PDD over MS-COCO [16].  $\Delta$  means the performance gains over the un-distilled students.

| Method     | Res101 $\rightarrow$ Res34 |          | SwinS $\rightarrow$ SwinT |          |
|------------|----------------------------|----------|---------------------------|----------|
|            | mAP                        | $\Delta$ | mAP                       | $\Delta$ |
| Student    | 70.31                      | –        | 79.59                     | –        |
| KD [10]    | 70.71                      | 0.40     | 80.02                     | 0.43     |
| DKD [38]   | 70.17                      | -0.14    | 79.23                     | -0.36    |
| NKD [36]   | 70.57                      | 0.26     | 79.82                     | 0.23     |
| PDD        | 74.13                      | 3.82     | 82.89                     | 3.30     |
| +DKD [38]  | 74.35                      | 4.04     | 83.14                     | 3.55     |
| +DIST [11] | 74.62                      | 4.32     | 83.37                     | 3.78     |
| +WKD [21]  | 74.22                      | 3.91     | 83.19                     | 3.60     |

Table 8. Ablative studies on NDD over MS-COCO [16].  $\Delta$  means the performance gains over the un-distilled students.

| Method                                                                    | Res101 $\rightarrow$ Res34 |          | SwinS $\rightarrow$ SwinT |          |
|---------------------------------------------------------------------------|----------------------------|----------|---------------------------|----------|
|                                                                           | mAP                        | $\Delta$ | mAP                       | $\Delta$ |
| Student                                                                   | 70.31                      | –        | 79.59                     | –        |
| All Labels ( <i>i.e.</i> , w/ Pos.)                                       | 71.28                      | 0.97     | 80.47                     | 0.88     |
| $\mathcal{N}_1 \cap \mathcal{N}_2$ ( <i>i.e.</i> , w/o Pos.)              | 73.45                      | 3.14     | 82.24                     | 2.65     |
| $\tilde{z}_{1,i}/\gamma_b - \tilde{z}_{2,i}$                              | 70.59                      | 0.28     | 79.68                     | 0.09     |
| $\sqrt{\text{sig}(\tilde{z}_{1,i})} - \sqrt{\text{sig}(\tilde{z}_{2,i})}$ | 71.73                      | 1.43     | 80.70                     | 1.11     |
| $\text{sig}(\tilde{z}_{1,i})\gamma_b - \text{sig}(\tilde{z}_{2,i})$       | 71.23                      | 0.92     | 80.57                     | 0.98     |
| $\text{sig}(\tilde{z}_{1,i}/\gamma_b) - \text{sig}(\tilde{z}_{2,i})$      | 73.45                      | 3.14     | 82.24                     | 2.65     |

target one. We extend this idea to the MLL context. As in Tab. 7, these methods only have minor improvement (even harmful) in MLL, and our PDD beats them by a large margin. *(ii)* Moreover, thanks to our positive divide scheme, the yielded multiple single-label KD tasks in PDD could be further boosted by incorporating the methods (like DKD, DIST [11], and WKD [21]) *designed for the single-label learning*. Note that, examining which method is more compatible with PDD goes beyond our main concerns. Nevertheless, those results once again support the effectiveness, superiority, and generalization of our motivation and method that casts multi-label KD as multiple single-label KD tasks.

**Ablative Studies on NDD.** *(i)* Recall that, NDD only selects the negative labels for distillation. Here, we give the empirical evidence. As in Tab. 8, it is observed that involving the positive labels in NDD results in undesired performance degradation, echoing the theoretical discussions regarding the loss re-formulation in Sec. 3.3 (*cf.* Eq. (15)). *(ii)* Moreover, as shown, we find that pre-processing the logits with `sigmoid` achieves better results, presumably because the predictive discrepancy  $\rho$  (*cf.* Eq. (10)) is naturally normalized and bounded after `sigmoid`, thereby promising a more stable distillation course. *(iii)* The impact of different scaling strategies are also studied. The results demonstrate that the proposed label-specific asymmetric scaling (*cf.* Eq. (11)) yields a more pronounced improvement, justifying its effectiveness and superiority.

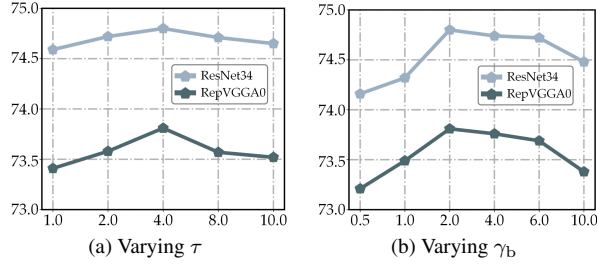


Figure 5. Impact of  $\tau$  and  $\gamma_b$  on MS-COCO [16]. The students share the ResNet101 as the teacher. mAP is reported.

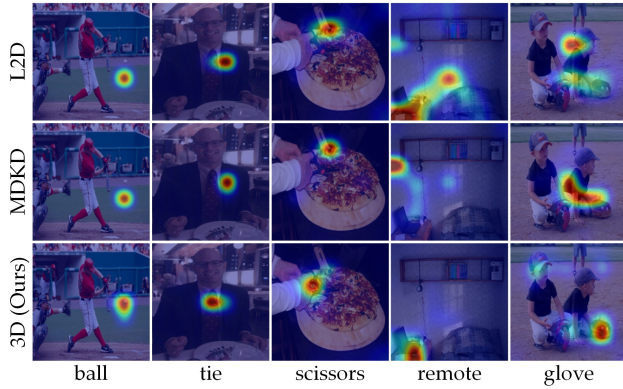


Figure 6. Comparison of the class activation maps on MS-COCO [16]. The texts below the images are the query categories.

**Impact of Hyper-parameters.** Fig. 5 studies the impacts of the tuning factors involved in PDD and NDD. (i) As shown in Fig. 5a, our method is fairly robust to the temperature factor  $\tau$  (cf. Eq. (5)) and consistently leads to significant distillation performance. (ii) Fig. 5b shows the results by varying the scaling factor  $\gamma_b$  (cf. Eq. (11) and Eq. (13)) in NDD. In general, a moderate  $\gamma_b \in [2, 6]$  produces stable results. One could also observe a clear performance degradation when  $\gamma_b < 1.0$ . The essential reason for this performance trend is that  $\gamma_b < 1.0$  (i.e.,  $\gamma^+ < 1.0$  and  $\gamma^- > 1.0$ ) will cause the predictive discrepancy  $\rho^t$  to be less discriminative (i.e.,  $|\rho^t|$  is approaching a full zero vector, see Fig. 4), in which case our NDD is not guaranteed to offer helpful guidance for students. These results back up our analysis in Sec. 3.3. (iii) In the *Supplementary Material*, we provide the sensitivity studies on the loss weights  $\alpha$  and  $\beta$  (cf. Eq. (16)), showing that our 3D is also not sensitive to their choices.

**Class Activation Map.** Fig. 6 compares the class activation maps (using LayerCAM [12]) with [29, 34]. It is observed that our method exhibits more precisely activated areas of the query classes (e.g., the tiny *ball* and *remote*), which directly translates to better KD performance.

**Classification Margin.** Fig. 7 plots the averaged positive-negative classification margin (see Sec. 3.2). As shown, the proposed PDD and NDD observably enlarge the margins

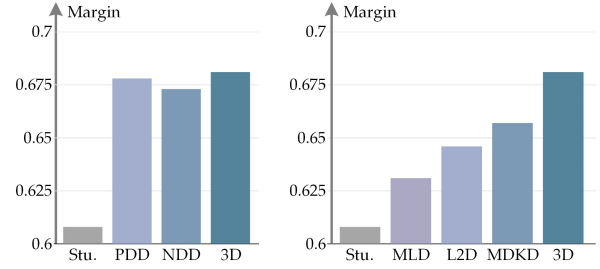


Figure 7. The positive-negative classification margin on COCO.

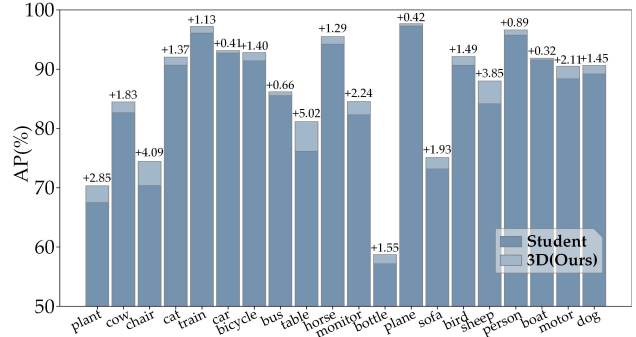


Figure 8. Per-category performance on PASCAL-VOC [8].

(left), and our 3D achieves the highest value compared to prior works (right) and thus leads to better distilled students. **Per-category Performance.** On PASCAL-VOC [8], Fig. 8 shows the class-wise AP scores, manifesting that the performance gains of our 3D are holistic, covering all categories. In particular, meeting the both KD goals of NDD and PDD, the co-occurrent categories *chair* and *table* obtain sizeable 5.02% and 4.09% improvements. The categories (like *plane* (0.42%) and *boat* (0.32%)) rarely appear together with others, accordingly damping the advantages of our methods and leading to less significant AP gains.

*Additional experiments, visualizations, and relevant discussions are provided in the Supplementary Material.*

## 5. Conclusion

In this paper, we present a novel viewpoint to execute logit distillation in MLL. Specifically, we propose positive divide distillation (PDD) to cast the single multi-label KD problem as multiple single-label KD tasks, and negative discrepancy distillation (NDD), for paired images, to model and distill their predictive discrepancy over the intersection of the negative labels. The outcome of our designs is a new and powerful logit KD approach for MLL, dubbed 3D. On both offline KD and self-KD, comprehensive experiments show that 3D achieves the state-of-the-art performance and latency compared with previous advanced methods, well illustrating its superiority. We hope this paper will contribute to future logit distillation research in MLL.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62376223, and in part by the Fundamental Research Funds for the Central Universities. We also thank Suyu Hao for the helpful presentation suggestions.

## References

- [1] Nan Cao, Teng Zhang, and Hai Jin. Partial multi-label optimal margin distribution machine. In *IJCAI*, pages 2198–2204, 2021. 4
- [2] Weihao Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. In *NeurIPS*, pages 15394–15406, 2022. 3
- [3] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, pages 5008–5017, 2021. 3, 6
- [4] 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. 3
- [5] Zhao-Min Chen, Xin Jin, Yisu Ge, and Sixian Chan. In pursuit of causal label correlations for multi-label image recognition. In *NeurIPS*, pages 51634–51654, 2024. 1, 2, 4
- [6] Jiacheng Cheng and Nuno Vasconcelos. Towards calibrated multi-label deep neural networks. In *CVPR*, pages 27589–27599, 2024. 2, 3, 4
- [7] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM Int. Conf. Image. Video Retr.*, pages 1–9, 2009. 2, 5, 6
- [8] 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*, pages 98–136, 2015. 2, 5, 6, 8
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6
- [10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS*, pages 1–12, 2015. 1, 2, 3, 4, 7
- [11] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *NeurIPS*, pages 1–12, 2022. 1, 2, 3, 7
- [12] Peng Tao Jiang, Chang Bin Zhang, Qibin Hou, Ming Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps. *IEEE TIP*, pages 1–12, 2021. 8
- [13] Takumi Kobayashi. Two-way multi-label loss. In *CVPR*, pages 7476–7485, 2023. 1, 2, 3, 4
- [14] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*, pages 1–12, 2017. 3
- [15] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *CVPR*, pages 16478–16488, 2021. 1, 2, 3, 4
- [16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 2, 5, 6, 7, 8
- [17] Ruyang Liu, Hao Liu, Ge Li, Haodi Hou, Tinghao Yu, and Tao Yang. Contextual debiasing for visual recognition with causal mechanisms. In *CVPR*, pages 12755–12765, 2022. 2
- [18] Ruyang Liu, Jingjia Huang, Thomas H Li, and Ge Li. Causality compensated attention for contextual biased visual recognition. In *ICLR*, pages 1–12, 2023. 1, 2, 3
- [19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 6
- [20] Ziyi Liu, Zengmao Wang, and Bo Du. Multi-marginal contrastive learning for multi-label subcellular protein localization. In *CVPR*, pages 20626–20635, 2022. 2, 4
- [21] Jiaming Lv, Haoyuan Yang, and Peihua Li. Wasserstein distance rivals kullback-leibler divergence for knowledge distillation. In *NeurIPS*, pages 65445–65475, 2024. 7
- [22] Amin Parchami-Araghi, Moritz Böhle, Sukrut Rao, and Bernt Schiele. Good teachers explain: Explanation-enhanced knowledge distillation. In *ECCV*, pages 293–310, 2024. 2
- [23] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *ICCV*, pages 3967–3976, 2019. 3, 6
- [24] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, pages 268–284, 2018. 3, 6
- [25] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, pages 82–91, 2021. 3
- [26] Yiqing Shen, Liwu Xu, Yuzhe Yang, Yaqian Li, and Yandong Guo. Self-distillation from the last mini-batch for consistency regularization. In *CVPR*, pages 11943–11952, 2022. 3, 7
- [27] Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Handling difficult labels for multi-label image classification via uncertainty distillation. In *ACM MM*, pages 2410–2419, 2021. 3, 6, 7
- [28] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *CVPR*, pages 15731–15740, 2024. 1, 2, 3
- [29] Jingchao Wang, Zhengnan Deng, Tongxu Lin, Wenyuan Li, Shaobin Ling, and Junyu Lin. Beyond direct relationships: Exploring multi-order label pair dependencies for knowledge distillation. In *ACM MM*, pages 1–9, 2024. 1, 2, 3, 4, 5, 6, 7, 8
- [30] Xucong Wang, Pengkun Wang, Shurui Zhang, Miao Fang, and Yang Wang. Multi-label self knowledge distillation. In *AAAI*, pages 21330–21338, 2025. 1, 2, 3, 5, 6, 7
- [31] Xi-Zhu Wu and Zhi-Hua Zhou. A unified view of multi-label performance measures. pages 3780–3788, 2017. 2
- [32] Jiazhi Xu, Sheng Huang, Fengtao Zhou, Luwen Huangfu, Daniel Zeng, and Bo Liu. Boosting multi-label image clas-

- sification with complementary parallel self-distillation. In *IJCAI*, pages 1–9, 2022. [3](#), [6](#)
- [33] Longrong Yang, Xianpan Zhou, Xuwei Li, Liang Qiao, Zheyang Li, Ziwei Yang, Gaoang Wang, and Xi Li. Bridging cross-task protocol inconsistency for distillation in dense object detection. In *ICCV*, pages 17175–17184, 2023. [1](#)
- [34] Penghui Yang, Ming-Kun Xie, Chen-Chen Zong, Lei Feng, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Multi-label knowledge distillation. In *ICCV*, pages 17271–17280, 2023. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [35] Penghui Yang, Chen-Chen Zong, Sheng-Jun Huang, Lei Feng, and Bo An. Dual-head knowledge distillation: Enhancing logits utilization with an auxiliary head. In *KDD*, pages 3530–3541, 2025. [2](#)
- [36] Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *ICCV*, pages 17185–17194, 2023. [2](#), [3](#), [7](#)
- [37] Sung-Hoon Yoon, Hoyong Kwon, Jaeseok Jeong, Daehee Park, and Kuk-Jin Yoon. Diffusion-guided weakly supervised semantic segmentation. In *ECCV*, pages 393–411, 2024. [2](#)
- [38] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, pages 11953–11962, 2022. [1](#), [2](#), [3](#), [7](#)
- [39] Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching. In *ICLR*, pages 1–12, 2024. [3](#)
- [40] Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Mingzhe Rong, Aijun Yang, and Xiaohua Wang. Complementary relation contrastive distillation. In *CVPR*, pages 9260–9269, 2021. [3](#)
- [41] Xuelin Zhu, Jian Liu, Weijia Liu, Jiawei Ge, Bo Liu, and Juxin Cao. Scene-aware label graph learning for multi-label image classification. In *ICCV*, pages 1473–1482, 2023. [2](#), [3](#)