— *Supplementary Material* —
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

A. Experimental Settings

A.1. Datasets

(i) MS-COCO [11], covering 80 categories, is the standard benchmark for multi-label image classification. MS-COCO contains 115K and 5K images for training and validation. (ii) PASCAL-VOC [6] consists of 5K and 4K images, across 20 categories, for training and validation. (iii) NUS-WIDE [1], which is a large-scale dataset commonly used for image classification with 81 concept categories, includes 161K and 107K annotated images for training and validation, respectively.

A.2. Network Architectures

Our work considers several architectures, including ResNet [7], Wide ResNet (WRN) [19], VGG [15], RepVGG [5], MobileNetV2 (MBV2) [14], ShuffleNetV2 (SHNV2) [20], and Swin Transformer [12], under different model sizes. All models are pre-trained on ImageNet-1K [3], as in [16, 18].

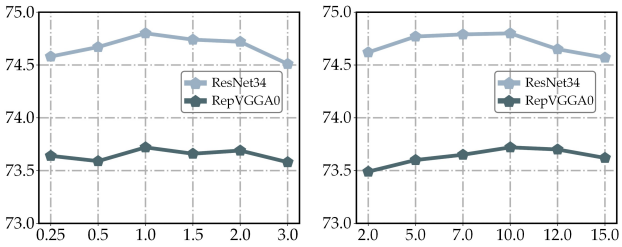
A.3. Evaluation Metrics

As [16, 18], we report the mean average precision (mAP) over all classes, average per-class F1-score (CF1), and overall F1-score (OF1) to evaluate the performance.

A.4. Implementation Details

For a fair comparison, we follow the L2D [18] and MSKD [17] codebase. Specifically, we respectively train the teachers and students for 30 and 80 epochs with the Adam optimizer [10]. The one-cycle policy is used with a maximal learning rate of 1e-4 and a weight decay of 1e-4. The batch size is 64 and the input size is 224×224. For each training image, we apply a weak augmentation consisting of random horizontal flipping and a strong augmentation consisting of Cutout [4] and RandAugment [2]. In particular, the models used in MSKD [17] remove the additional MLP head.

*Corresponding author



(a) Varying $\alpha \times 1e2$ ($\beta = 10.0$) (b) Varying β ($\alpha = 1e2$)

Figure B.1. Impact of α and β on MS-COCO [11]. The students share the ResNet101 as the teacher. mAP is reported.

Regarding hyper-parameters, we set $\alpha = 100.0$, $\beta = 10.0$, $\tau = 4.0$, and $\gamma_{\text{b}} = 2.0$ for all experiments.

B. Additional Experiments and Visualizations

B.1. More Distillation Results

Tab. B.1 and Tab. B.2 summarize more distillation results on MS-COCO [11] and PASCAL-VOC [6]. It is observed that the proposed 3D consistently surpasses the previous methods on both various teacher-student combinations, strongly illustrating its effectiveness and superiority.

B.2. Reversed Distillation

Following L2D [18] and MDKD [16], we perform reversed distillation on MS-COCO [11] by setting ResNet101 as the student and ResNet34 as the teacher. As reported in Tab. B.3, our method still outstrips the previous competitors and obtains the top performance.

B.3. Sensitivity Analysis on Loss Weights

We introduce two hyper-parameters α and β to balance the proposed KD losses. Fig. B.1 investigates their influences. As shown, the performance variation is not large against the loss weights, and different combinations of α and β could result in stable and superior distillation performance.

Table B.1. Additional multi-Label image classification KD performance on MS-COCO [11].

Method	VGG13 → MBV2			ResNet32×4 → SHNV2			WRN40-2 → WRN16-2			ResNet56 → ResNet20		
	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1
Teacher	74.08	75.05	72.20	79.44	79.04	75.98	75.42	77.11	74.68	72.18	74.26	68.09
Student	71.85	73.68	68.40	72.58	73.13	70.15	73.19	75.68	72.17	69.25	71.45	66.28
RKD	72.48	73.48	67.48	72.00	73.27	70.49	73.18	75.62	72.15	69.48	71.98	67.18
PKT	71.45	73.08	68.43	73.12	73.18	70.48	73.89	76.14	72.89	70.15	72.16	66.40
Review KD	72.49	73.32	69.18	73.15	73.89	70.14	74.01	76.08	72.69	70.45	72.68	67.34
MSE	72.00	73.90	68.87	72.14	73.64	71.00	74.22	76.15	73.05	69.40	72.44	67.48
PS	71.43	72.98	68.78	72.45	73.81	70.78	73.04	75.65	73.48	69.77	71.18	66.34
L2D	71.05	73.98	68.18	72.66	73.22	71.34	72.98	76.89	72.03	71.45	73.09	67.54
MDKD	73.98	74.89	71.80	74.07	75.13	72.66	75.34	77.20	73.89	73.37	74.08	68.29
MLD	74.92	75.75	71.21	72.79	73.88	70.91	73.32	75.74	73.72	70.28	72.46	67.29
3D (Ours)	**74.34**	**75.27**	**71.98**	**74.98**	**75.82**	**73.59**	**76.12**	**77.99**	**74.92**	**74.37**	**75.66**	**69.55**

Table B.2. Additional Multi-Label image classification KD performance on PASCAL-VOC [6].

Method	ResNet50 → RepVGGA0			SwinT → ResNet18			ResNet50 → MBV2			SwinT → MBV2		
	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1
Teacher	86.73	84.92	81.21	91.43	89.81	87.63	86.73	84.92	81.21	91.43	89.81	87.63
Student	83.79	83.36	79.83	84.01	83.60	79.42	86.12	85.01	81.76	86.12	85.01	81.76
RKD	84.26	84.29	80.70	83.27	83.05	79.55	86.22	84.97	81.76	85.68	85.31	81.57
PKT	83.93	83.79	80.03	83.45	83.25	79.64	86.10	84.84	81.66	85.67	85.22	81.68
Review KD	84.07	83.62	80.34	83.37	83.08	78.93	85.87	85.04	81.73	85.69	85.10	81.56
MSE	84.01	84.05	80.52	83.60	83.06	79.46	86.20	84.94	81.84	85.80	85.51	81.98
PS	84.80	84.46	81.13	83.97	83.75	79.86	86.26	85.47	82.06	86.07	85.73	82.39
L2D	86.26	85.85	82.55	85.87	85.67	82.17	87.32	86.48	83.26	87.37	86.88	83.68
MDKD	86.56	85.10	81.46	85.99	85.71	82.11	87.49	86.49	83.27	87.29	86.71	83.55
MLD	85.07	84.91	81.55	84.61	84.26	80.78	86.38	85.67	82.43	86.11	85.98	82.55
3D (Ours)	**87.27**	**86.23**	**82.65**	**86.53**	**85.98**	**82.62**	**87.92**	**86.78**	**83.59**	**87.78**	**87.12**	**84.16**

Table B.3. Results of reversed distillation on MS-COCO [11]. The teacher-student KD pair is ResNet34-ResNet101.

Method	mAP	OF1	CF1
Teacher	70.19	72.30	66.50
Student	73.98	75.01	70.12
RKD	74.03	74.96	70.01
PKT	73.95	74.94	69.98
Review KD	74.02	74.96	70.07
MSE	74.21	75.12	70.18
PS	74.70	75.78	71.08
L2D	75.51	76.25	71.75
MDKD	76.61	76.98	72.03
MLD	74.64	75.78	71.10
3D (Ours)	**77.48**	**77.63**	**72.42**

C. Additional Discussions

C.1. Explanation on Assumption 3.1

Assumption 3.1 stated that the teacher could elicit high-confidence and correct predictions for the negative labels. Here, we show the empirical evidence. Tab. C.4 summarizes the pre-trained teacher's predictions. As summarized, the teacher model could elicit high-confidence and correct predictions, per label, for almost all training samples, verifying that the Assumption 3.1 is rational and legitimate.

C.2. Details on Loss $\mathcal{L}_{\text{NDD}}$ Reformulation

Here, we give the detailed derivation of the $\mathcal{L}_{\text{NDD}}$ reformulation (*i.e.,* Eq. (15)) in Sec. 3.3, and the notations are the same in Sec. 3.3 of the manuscript. First, given two vectors $\mathbf{u} \in \mathbb{R}^{1\times N}$ and $\mathbf{v} \in \mathbb{R}^{1\times N}$, we have:

$$\ell_{\text{MSE}}(\mathbf{u}, \mathbf{v}) = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{u}_i - \mathbf{v}_i)^2 \tag{1}$$

Table C.4. The predictive scores, per label, generated from a well-trained teacher (ResNet101) model. All the metrics are calculated and averaged on the training split, *which is the one used for distillation.* P. and N. stand for positive and negative, respectively.

Data	MS-COCO	PASCAL-VOC	NUS-WIDE
Accuracy	98.67	96.27	98.90
P. Score	0.75	0.71	0.69
N. Score	0.01	0.03	0.01

Moving forward and supposing $N = |\mathcal{N}_1 \cap \mathcal{N}_2|$, $\mathcal{L}_{\text{NDD}}$ can be computed as:

$$
\begin{aligned}
\mathcal{L}_{\text{NDD}} &= \ell_{\text{MSE}}(\boldsymbol{\rho}^s, \boldsymbol{\rho}^t) \\
&= \ell_{\text{MSE}}\big((\tilde{\mathbf{p}}^s_1 - \tilde{\mathbf{p}}^s_2), (\tilde{\mathbf{p}}^t_1 - \tilde{\mathbf{p}}^t_2)\big) \\
&= \frac{1}{N}\sum_{i=1}^{N}\Big((\tilde{\mathbf{p}}^s_{1,i} - \tilde{\mathbf{p}}^s_{2,i}) - (\tilde{\mathbf{p}}^t_{1,i} - \tilde{\mathbf{p}}^t_{2,i})\Big)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\Big((\tilde{\mathbf{p}}^s_{1,i} - \tilde{\mathbf{p}}^t_{1,i})^2 + (\tilde{\mathbf{p}}^s_{2,i} - \tilde{\mathbf{p}}^t_{2,i})^2 \\
&\quad + 2(\tilde{\mathbf{p}}^s_{1,i} \cdot \tilde{\mathbf{p}}^t_{2,i} + \tilde{\mathbf{p}}^s_{2,i} \cdot \tilde{\mathbf{p}}^t_{1,i}) \\
&\quad - 2\tilde{\mathbf{p}}^s_{1,i} \cdot \tilde{\mathbf{p}}^s_{2,i} - \underbrace{2\tilde{\mathbf{p}}^t_{1,i} \cdot \tilde{\mathbf{p}}^t_{2,i}}_{\text{no gradient}}\Big) \\
&< \ell_{\text{MSE}}(\tilde{\mathbf{p}}^s_1, \tilde{\mathbf{p}}^t_1) + \ell_{\text{MSE}}(\tilde{\mathbf{p}}^s_2, \tilde{\mathbf{p}}^t_2) \\
&\quad + \underbrace{2(\tilde{\mathbf{p}}^s_1 \cdot \tilde{\mathbf{p}}^t_2 + \tilde{\mathbf{p}}^s_2 \cdot \tilde{\mathbf{p}}^t_1)}_{\mathcal{L}_{\text{aux}}}
\end{aligned}
\tag{2}
$$

References

[1] 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. 1

[2] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, pages 702–703, 2020. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[4] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1

[5] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *CVPR*, pages 13733–13742, 2021. 1

[6] 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. 1, 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[8] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *NeurIPS*, pages 1–12, 2022.

[9] 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.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, pages 1–12, 2015. 1

[11] 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. 1, 2

[12] 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. 1

[13] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, pages 5191–5198, 2020.

[14] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 1

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, pages 1–12, 2014. 1

[16] 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

[17] Xucong Wang, Pengkun Wang, Shurui Zhang, Miao Fang, and Yang Wang. Multi-label self knowledge distillation. In *AAAI*, pages 21330–21338, 2025. 1

[18] 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

[19] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, pages 1–9, 2016. 1

[20] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, pages 6848–6856, 2018. 1