Multi-label Self Knowledge Distillation

Xucong Wang [1], Pengkun Wang [12*], Shurui Zhang [3], Miao Fang [3], Yang Wang [12*]

[1]University of Science and Technology of China, Hefei 230236, China
[2]Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou 215123, China
[3]Northeastern University at Qinhuangdao, Qinhuangdao 066000, China
xuco@mail.ustc.edu.cn, {pengkun, angyan}@ustc.edu.cn, 202315279@stu.neuq.edu.cn, fangmiao@neuq.edu.cn

Abstract

Self-Knowledge Distillation (SKD) leverages the student's own knowledge to create a virtual teacher for distillation when the pre-trained bulky teacher is not available. Whilst existing SKD approaches demonstrate gorgeous efficiency in single-label learning, to directly apply them to multi-label learning would suffer from dramatic degradation due to the following inherent imbalance: *targets with unified labels but multifarious visual scales are crammed into one image, resulting in biased learning of major targets and disequilibrium of precision-recall*. To address this issue, this paper proposes a novel SKD method for multi-label learning named Multi-label Self-knowledge Distillation (MSKD), incorporating three Spatial Decoupling mechanisms (i.e. Locality-SD (L-SD), Reconstruction-SD (R-SD), and Step-SD (S-SD)). L-SD exploits relational dark knowledge from regional outputs to amplify the model's perception of visual details. R-SD reconstructs global semantics by integrating regional outputs from local patches and leverages it to guide the model. S-SD aligns outputs of the same input at different steps, aiming to find a synthetical optimizing direction and avoid the overconfidence. In addition, MSKD combines our tailored loss named MBD for balanced distillation. Exhaustive experiments demonstrate that MSKD not only outperforms previous approaches but also effectively mitigates biased learning and equips the model with more robustness.

Code — https://github.com/asaxuc/MSKD

Introduction

Multi-Label Learning (MLL) has exemplified eye-catching applications in various downstream tasks, like action recognition (Zhang et al. 2020b), recommendation (Schultheis et al. 2022; Zhang et al. 2020a), and user profiling (Wang et al. 2021). Different from Single-Label Learning (SLL), the major challenge of MLL lies in learning severe non-injective mapping between visual targets and labels, which means the model is required to perceive all targets within an image equally, regardless of their cardinality, size, or location. While existing MLL models which propose to model visual-label correlations (Li et al. 2023; Wang et al. 2020; Chen et al. 2019) or design class-specific decoders (Liu et al. 2021; Ridnik et al. 2023) have made remarkable success, most of them are costly and struggled to reach prevailing lightweight demand on both efficiency and accuracy. In this case, Knowledge Distillation (Hinton, Vinyals, and Dean 2015), the latest benchmark of model compression which asks a lightweight student to mimic its pre-trained bulky teacher to receive comparable performance, is widely applied to MLL and exemplifies prominent efficiency. (Yang et al. 2023a; Xu et al. 2022; Song et al. 2021).

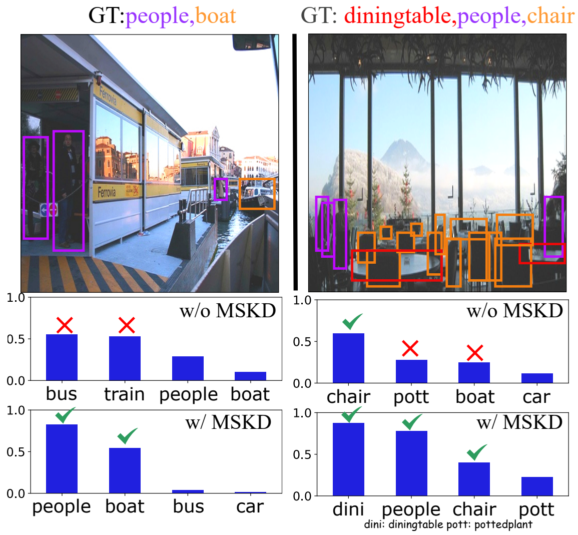

Figure 1: Top-4 predictions of ResNet34 trained w \ w/o our MSKD, ranked in descending order. MSKD endows the model with significant distinguishing capability and robustness to space-intricately distributed samples.

However, the fact is that pre-trained teachers aren't always accessible in latency-sensitive circumstances. In this case, SKD offers a competent solution by allowing students to directly learn from their self-teachers. Notably SKD is a non-trivial problem since no explicit peer model is provided. Existing SKD approaches dig out self-teacher from various layers (Yang et al. 2023b), samples (Yang et al. 2022), and training stages (Yang et al. 2019; Kim et al. 2021) as the self-teacher to form knowledge transferring, which has shown

[*]Corresponding author.

prominent efficiency in SLL. However, they perform unsatisfyingly in MLL due to the tough awareness of inherent imbalance, especially when faced with sophisticated backbones or datasets. Recently, a few SKD methods attempted to approach MLL but work poorly and are restrictive. For example, (Song et al. 2021) proposed the Uncertainty Distillation (UD) scheme to avoid over-training on difficult labels; MulSupCon (Zhang and Wu 2024) proposed a supervised contrastive learning for MLL which selects positive sets based on their overlapping degree with the anchor. However, both of them primarily leverage the coarse regularization over the entire image semantics and show highly limited performance, where further consideration of the inherent imbalance problem of MLL is neglected.

Therefore, an urgent problem to be solved is: **How to design an exactly efficient and generic SKD for MLL, which could alleviate the inherent imbalance?** We answer the question by proposing our **M**ulti-label **S**elf-**K**nowledge **D**istillation (MSKD), incorporating three decoupling mechanisms attached with our tailored distillation for MLL. To begin, we extract randomly cropped and resized regions from original image. Intuitively, outputs of those regions contain more limited but specific semantics which fall to oblivion while processing the whole image, thus our first mechanism L-SD lies in utilizing them to amplify the corresponding features in the overall image. Moreover, we leverage these outputs to reversely identify the uniqueness of each patch and generate corresponding pseudo labels, then employ nonparametric graph propagation on them to capture spatial correlations for semantic integration. Accordingly we formulate our R-SD between integrated patches and vanilla output logits. Another mechanism S-SD requires the model to find a synthetical optimizing direction by aligning it's outputs at near training steps, which would not only mitigate overfitting tendencies towards fitful difficult or miss-labeled targets but also enhance model's stability.

In addition, we design a balanced distillation function tailored for MLL named MBD. Inspired by ASL (Ridnik et al. 2021), MBD employs reformulated softmax and dynamic KL-Divergence to mitigate the imbalance distillation between positive and negative logits.

Comprehensive experiments exemplify MSKD's state-of-art performance. Figure 1 provides an intuitive visualization of MSKD's effects, indicating that MSKD enhances the model's ability to discriminate imbalanced samples and strengthens decision boundaries. Our contributions are then summarized as follows:

- *New insight*: To the best of our knowledge, our work is the first study to expand SKD methods to MLL and proposes a tailored benchmark SKD for it named MSKD.
- *New advisable distillation framework*: MSKD combines three kinds of spatial decoupling mechanisms to address inherent imbalance problems of MLL, supported by our tailored distillation loss named MBD.
- *Compelling empirical results*: we conduct exhaustive tests on multiple benchmark datasets. Results show our MSKD's state-of-art performance as well as superior robustness under sophisticated circumstances.

Related Works

Multi-Label Learning

Multi-Label Learning (MLL) has attracted prevalent interest from research communities due to its widespread application in real-world scenarios. Existing MLL works can be roughly divided into the following categories: Loss Rebalancing (LB), Relation Modeling (RM), and Class Specific Decoding (CSC). LB-based approaches (Lin 2017) endeavor to mitigate the biased label supervision resulted from unmatching cardinality of positive and negative labels in the dataset. For example, ASL (Ridnik et al. 2021) punishes easy negative learning and emphasizes positive learning by introducing flexible exponent hyperparams and special threshold for negative probabilities. RM-based SKD utilizes multi-source prior information like the co-occurrence matrix, label embedding, or knowledge graphs (Lee et al. 2018) to assist the modeling of deep semantics across labels. Typically, ML-GCN (Chen et al. 2019) extracts label embedding and condition probability as the node features and edges separately and constructs a labeled graph over the image. Then multiple GCN layers are employed to model the label relation, and outputs are fused with visual features to generate integrated prediction. PatchCT (Li et al. 2023) wisely utilizes the optimal transportation theory to learn the visual-language interactions. CSC-based SKD aims at designing class-specific decoder architectures for better modeling the distinguished class semantics. For example, Query2Label (Liu et al. 2021) adapts DETR (Carion et al. 2020) from object detection and relies on distinguished learnable queries to perform class-specific prediction.

Self-Knowledge Distillation

Firstly introduced by Geoffrey Hinton, Knowledge Distillation (KD) aims to transfer the abundant knowledge of a bulky teacher into a lightweight student which is more suitable for real-time applications. However, access to a pretrained teacher model is often impractical, thus it's vital to dig out a self-teacher to guide the model in such circumstances, which is Self-Knowledge Distillation (SKD). As a general semi-supervised method, SKD mainly consists of the following categories: Architecture Modification (AM), Consistency Regularization (CM), and Label Smoothing (LS). AM-based SKD employs auxiliary modules to draw extra updating flows, such as BYOT (Zhang et al. 2019) which employs an auxiliary classifier for each block, and USKD (Yang et al. 2023b) which introduces an extra classifier for the middle layer to regularize the whole model. CM-based SKD focuses on improving the model's robustness in multiple dimensions, for example, CS-KD (Yun et al. 2020) endeavors to close the logits between the same class. PS-KD (Kim et al. 2021) and DLB (Shen et al. 2022) try to keep the model's robustness in different training stages. In a more general perspective, several contrastive learning approaches would be also treated as CM if the label supervision is applied. LS is regarded as a specific SKD category where smoothed labels could be viewed as the virtual teacher. For example, Zipf's LS (Liang et al. 2022) employs Zipf's distribution to guide the model.

While numerous SKD methods have been proposed, few have effectively addressed the specific challenges of MLL. Reluctantly, (Song et al. 2021) proposes an uncertainty-based self-distillation scheme (UD), but limited progress is achieved due to the undistinguished calibration branch, and its cumbersome training pipeline deviates efficiency principle of SKD. (Pan et al. 2022) introduces a self-distillation scheme between visual encoder and label encoder, however, it's largely limited to specific backbones and prior knowledge. Others like MulSupCon (Zhang and Wu 2024) devise a supervised contrastive learning mechanism that imposes dynamical weights based on the overlap between two samples. However, the essential bottleneck of MLL concerning the aforementioned imbalance problem has not been considered. To the best of our knowledge, we first try to directly utilize SKD theory to handle the inherent imbalance in MLL and craft a tailored and pioneering SKD method for it.

Methodology

Preliminary. Firstly we define some notations for a C-class MLL task. Given batch of samples $(\mathcal{X}, \mathcal{Y}) = [(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \cdots, (\boldsymbol{x}_i, \boldsymbol{y}_i), \cdots, (\boldsymbol{x}_B, \boldsymbol{y}_B)]$, where B denotes the batch size and $\boldsymbol{y}_i \in \{0, 1\}^C$ denote the label, where each element indicates whether the corresponding target exists in image $\boldsymbol{x}_i$ or not. We denote the positive label set of $\boldsymbol{x}_i$ as $\mathcal{M}_i$ ($\mathcal{M}_i \subset \boldsymbol{y}_i$) in extra. For any feature extractor $h(;\boldsymbol{\phi})$ and classifier $d(;\boldsymbol{\rho})$, we denote $\boldsymbol{f}_i = h(\boldsymbol{x}_i;\boldsymbol{\phi})$, $\boldsymbol{q}_i = d(\boldsymbol{f}_i;\boldsymbol{\rho})$). The optimizer endeavors to discover the optimal $\tilde{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\rho}}$ that minimize the expected Binary Cross-Entropy loss on $(\mathcal{X}, \mathcal{Y})$:

$$\mathcal{L}_{\mathrm{BCE}} = \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i)} \boldsymbol{y}_i \log(\sigma(\boldsymbol{q}_i)) + (1 - \boldsymbol{y}_i) \log(1 - \sigma(\boldsymbol{q}_i)), \tag{1}$$

where σ is the sigmoid function. Next, we introduce our tailored SKD for dealing with the inherent imbalance problem.

Multi-label Self Knowledge Distillation

The general framework of MSKD is illustrated in Figure 2. It incorporates three SD mechanisms, namely Locality-SD, Reconstruction-SD, and Step-SD, arranged in three flows of a single training step t.

Locality SD. We propose to randomly crop S patches (each patch is denoted as $\mathcal{O}_i^s$) from the original image $\boldsymbol{x}_i$ and resize them to magnify the details of the image and dilate the influence of large visual targets. Cropped patches are then fed into $h(;\boldsymbol{\phi}^t)$ to generate regional feature maps $\boldsymbol{f}_i^s$ and output logits $\boldsymbol{o}_i^s$, i.e $\boldsymbol{f}_i^s = h(\mathcal{O}_i^s;\boldsymbol{\phi}^t)$, and $\boldsymbol{o}_i^s = d(f_i^s;\boldsymbol{\rho}^t)$. We suggest treating these regional semantics as the self-teacher and leveraging them to amplify the corresponding regions of the original feature map. To achieve this, we firstly employ Region-of-Interest Pooling to obtain the feature map regions corresponding to $\mathcal{O}_i^s$ in the original feature map $\boldsymbol{f}_i^s$, then use classifier $d(;\boldsymbol{\rho}^t)$ to generate outputs:

$$\boldsymbol{r}_i^s = d(\mathrm{ROI}(\boldsymbol{f}_i;\mathcal{O}_i^s);\boldsymbol{\rho}^t). \tag{2}$$

A naive idea to form regularization is directly aligning $\boldsymbol{r}_i^s$ with $\boldsymbol{o}_i^s$. However, it would incur excessive learning of large targets and offset the locality regularization brought by patches. Instead, we conduct relation distillation between them to utilize the batch-wise dark knowledge and region-wise dark knowledge. Given $\boldsymbol{r} \in R^{B\times S\times C}$ and $\boldsymbol{o} \in R^{B\times S\times C}$, we firstly calculate inter-patch and inter-batch similarities (i.e., $\mathrm{sim}_p(\cdot)$ and $\mathrm{sim}_b(\cdot)$) for both of them:

$$\begin{aligned}(\mathrm{sim}_{\mathrm{p}}(\boldsymbol{r}_i))_{jk} &= ||\boldsymbol{r}_i^j - \boldsymbol{r}_i^k||_2,\\ (\mathrm{sim}_{\mathrm{b}}(\boldsymbol{r}))_{jk} &= ||\mathrm{avg}_s(\boldsymbol{r}_j^s) - \mathrm{avg}_s(\boldsymbol{r}_k^s)||_2,\end{aligned} \tag{3}$$

where $\mathrm{sim}_{\mathrm{p}}(\boldsymbol{r}_i) \in R^{S\times S}$, $\mathrm{sim}_{\mathrm{b}}(\boldsymbol{r}) \in R^{N\times N}$. $(\cdot)_{jk}$ denotes the $(j$,$k)$ element; $||\cdot||_2$ denotes normalized 2-D Euclidean distance. Notably, $\boldsymbol{r}_i^s$ is averaged when calculating $\mathrm{sim}_{\mathrm{b}}(\boldsymbol{r})$ since separate alignment of the same locations in different samples is not expected. The same operation in Eq. 3 is conducted on $\boldsymbol{o}_i$. To distill instance-wise and patch-wise dark knowledge from the self-teacher, we employ Huber-Loss like (Yang et al. 2023a) as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{LSD}} = \frac{1}{B}(\sum_i &\mathrm{HuberLoss}(\mathrm{sim}_{\mathrm{p}}(\boldsymbol{r}_i), \mathrm{sim}_{\mathrm{p}}(\mathrm{sg}(\boldsymbol{o}_i)))) \\ &+ \mathrm{HuberLoss}(\mathrm{sim}_{\mathrm{b}}(\boldsymbol{r}), \mathrm{sim}_{\mathrm{b}}(\mathrm{sg}(\boldsymbol{o}))),\end{aligned} \tag{4}$$

where $\mathrm{sg}(\cdot)$ means stop-gradient.

Reconstruction SD. For certain, resized random patches magnify visually subtle targets and provide different perspectives of viewing image semantics. With this regard, we seek to synthetically utilize them to endow the model with more sensibility to finer details. To be specific, we propose a graph-based reconstruction module named Graph Propagation (GP), to generate dynamic weights for different regional logits and reconstruct global semantics for regularization. For a random patch $\mathcal{O}_i^s$ of image $\boldsymbol{x}_i$, use $\boldsymbol{o}_i^s$ to denote its output logits, GP firstly formulates pseudo label $\boldsymbol{u}_i^s$ with:

$$(\boldsymbol{u}_i^s)_j = \begin{cases} 1, & i, j, s \in \mathrm{argmax}(\{o_{ij}^s - q_{ij}\}, \beta) \\ -1, & i, j, s \in \mathrm{argmin}(\{o_{ij}^s - q_{ij}\}, \beta) \\ 0, & else \end{cases} \tag{5}$$

where o_{ij}^s denotes the j-th class in $\boldsymbol{o}_i^s$, the same meaning for j in q_{ij}. Note that, the argmax operation is batch-wise and β is set to $|\boldsymbol{y}|$ in default. Eq. 5 highlights top-β classes whose prediction probabilities show the largest difference with that of the overall image, i.e. 'overachiever' classes, as representatives of the patch. Intuitively, these pseudo labels indicate the distinctiveness and specificity of patches with respect to the original image in certain training stages: positive pseudo labels exemplify more equal 'existing' uniqueness of targets in the patch since the imbalance effect is dilated by the random cropping; Negative pseudo labels indicate the peculiarity of 'non-existing' by minus value, which would be transferred to the general "existing" of corresponding classes in other patches after normalization. As a result, the pseudo labels are believed to indicate reliability of every class of each logit in representing a patch or being a self-teacher, and are more likely to concentrate on visually subtle targets.

Next, we consider forming graph propagation on these pseudo labels to further obtain relation-aware representations. Specifically, $\boldsymbol{o}_i^s$ is selected to construct edges for its

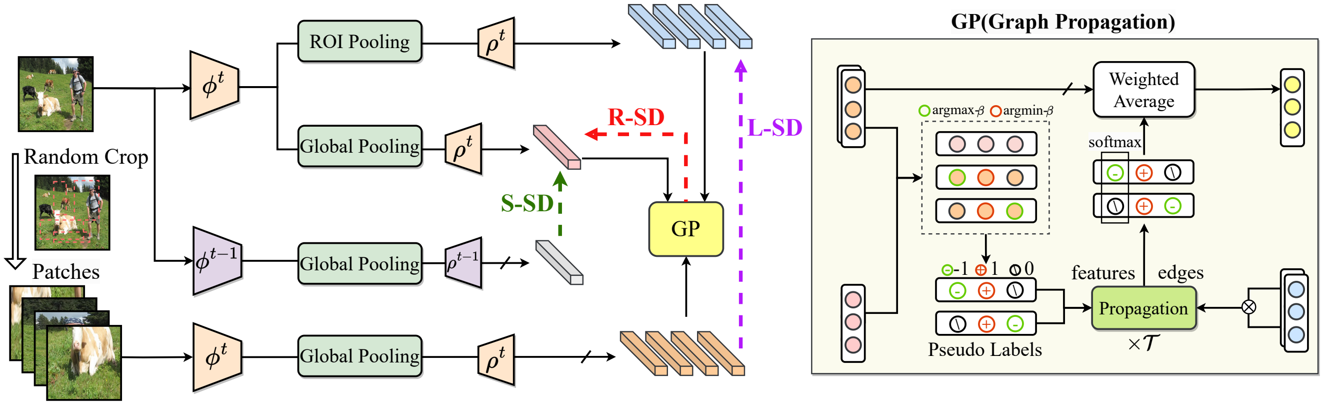

Figure 2: Overview of MSKD. ϕ^t represents parameters of the feature encoder in iteration t. ρ^t represents the parameters of the classifier in iteration t. The slash (/) over arrows means stop-gradient. Label supervision is omitted.

abundant spatial dark knowledge. We refer to Geom-GCN (Pei et al. 2020) and construct edge weights with normalized 2-D Euclidean distance:

$$(A_i^s)_{jk} = \frac{||\boldsymbol{o}_i^j - \boldsymbol{o}_i^k||_2}{\max\{||\boldsymbol{o}_i^j - \boldsymbol{o}_i^k||_2; (j,k) \in B^2\}}. \quad (6)$$

then a nonparametric propagation is employed to endow pseudo labels with deep spatial correlation:

$$u_i^s \leftarrow (D_i^s)^{-\frac{1}{2}}(A_i^s + I)(D_i^s)^{-\frac{1}{2}} u_i^s \quad (7)$$

where I is identity matrix, and $D_i^s = \text{diag}[\sum_k (A_i^s)_{:,k} + (I)_{:,k}]$ is the diagonal degree matrix of A_i^s. Initial feature u_i^s is updated $\mathcal{T}$ times based on Eq. 7 to capture deep representations (denoted as $\overline{\boldsymbol{e}}_i^s$). Finally, $\overline{e}_i^s$ is normalized with Softmax, and then weighs the average of regional logits o_i^s. The averaged regional logits reconstruct the global semantics from a finer locality perspective which the model fails to capture, hence our R-SD leverages it to guide the model:

$$\mathcal{L}_{\text{RSD}}^{(i)} = \mathcal{L}_{\text{MBD}}(\sum_s (\text{softmax}(\overline{\boldsymbol{e}}_i^s) \odot \text{sg}(\boldsymbol{o}_i^s)), \boldsymbol{q}_i), \quad (8)$$

where $\odot$ is hadamard product, $\mathcal{L}_{\text{MBD}}$ is our proposed MBD loss for more balanced distillation and will be introduced in the next subsection. Notably, softmax applied on $\overline{e}_i^s$ makes classes that frequently appear in different patches to obtain averaged weights, enabling the awareness of multifarious disjoint patterns and improving the model's robustness.

Step SD. Since the inherent imbalance problem of MLL gives a natural optimization challenge, we would like to find an optimized direction for each iteration and avoid models from being severely influenced by fitful difficult or miss-labeled samples. In this case, our S-SD takes parameters $\boldsymbol{\phi}^{t-1}$ of the model h in last iteration step $t-1$, and feed current input $\boldsymbol{x}_i$ into both $h(;\boldsymbol{\phi}^{t-1})$ and $h(;\boldsymbol{\phi}^t)$. Denote their output logits as q^{t-1} and q^t respectively, S-SD hopes that the model would find an integrated optimization direction by shortening the discrepancy between q_i^{t-1} and q_i^t:

$$\mathcal{L}_{\text{SSD}}^{(i)} = \mathcal{L}_{\text{MBD}}(\boldsymbol{q}_i^t, \text{sg}(\boldsymbol{q}_i^{t-1})) \quad (9)$$

here $\mathcal{L}_{\text{MBD}}$ is also employed to avoid the overwhelming distillation on negative logits.

Balanced Distillation Loss for MLL: MBD

Theoretical analysis. Let's begin with the limitations of employing softmax-based KL-Div in MLL. Using $\tilde{\boldsymbol{p}}_i$ and $\tilde{\boldsymbol{q}}_i$ to denote the prediction distribution of teacher and student after softmax, $\tilde{q}_{it}$ to denote the random t-th term of $\tilde{\boldsymbol{q}}_i$, then:

$$\mathcal{L}_{\text{sfx+KL}}^{(i)} = \sum_j^C \tilde{p}_{ij} \log(\frac{\tilde{p}_{ij}}{\tilde{q}_{ij}}). \quad (10)$$

by taking the gradient of $\mathcal{L}_{\text{sfx+KL}}^{(i)}$ with respect to $\tilde{q}_{it}$ we get (see **Appendix C** for detailed steps):

$$\nabla_{\tilde{q}_{it}} \mathcal{L}_{\text{sfx+KL}}^{(i)} = -\frac{\tilde{p}_{it}}{\tilde{q}_{it}} + (\sum_{j\neq t} \tilde{p}_{ij} \frac{1}{\tilde{q}_{it} \cdot \sum_{j\neq t} e^{q_{ij}}}). \quad (11)$$

after reformulation, we get:

$$\nabla_{\tilde{q}_{it}} \mathcal{L}_{\text{sfx+KL}}^{(i)} = (-\tilde{p}_{it} + \frac{\sum_{j\neq i} \tilde{p}_{ij}}{\sum_{j\neq i} e^{q_{ij}}}) \frac{1}{\tilde{q}_{it}}. \quad (12)$$

without loss of generality, assume $\sum_{j\neq i} e^{q_{ij}} \gg \sum_{j\neq i} \tilde{p_{ij}}$. Due to the existence of multi-labels, the discrepancy between $\tilde{p}_{it}$ and $\sum_{j\neq i} \tilde{p}_i$ largely shrinks, which means the optimization strategy would push both positive and negative logits to the same direction. This will lead to two **issues**:

- Negative logits would be largely mis-leaded;
- Positive and negative distillation is indistinguishable, resulting in overwhelming learning of the latter one and making the model conservative to positive prediction.

To deal with above issues, we propose MBD which simultaneously adopts reformulated softmax and reformulated KL-Div Loss for balanced distillation.

Reformulated Softmax (RS). To solve *issue 1*, our Reformulated Softmax (RS) borrows the calibration branch from (Song et al. 2021) and applies softmax on multiple one-versus-negative combinations, to highlight the distinctive group knowledge of the positive logits:

$$\tilde{q}_{ij} = \frac{1}{|\mathcal{M}_i|} \sum_{t\in\mathcal{M}_i} \frac{e^{q_{ij}}}{e^{q_{it}} + \sum_{j\notin\mathcal{M}_i} e^{q_{ij}}} \quad (13)$$

where $|\mathcal{M}_i|$ denotes number of elements in $\mathcal{M}_i$. RS generates probabilities for one-positive-all-negative combinations rather than directly take all positive logits into consideration, which substantially sharpen the discrepancy between positive and negative distillation and avoids the misleading of corrupted $\tilde{p}_{it}$. From another perspective, RS avoids the mutual influence between optimization of different positive logits and unleashes the distinctiveness and equality of each one, since the maximation of posterior distribution no longer contradicts with the simultaneous propulsion of all positive logits, where more pareto optimality could be discovered.

Reformulated Distillation (ReD). To deal with *issue 2*, we rectify the KL-Divergence with separate coefficients respectively for emphasizing positive distillation and punishing negative ones. Reusing $\tilde{\boldsymbol{p}}_i$, $\tilde{\boldsymbol{q}}_i$ in the former subsection, our Reformulated Distillation (ReD) would be written as:

$$\mathcal{L}_{\text{ReD}} = \sum_{j \in \mathcal{M}_i} w^+_{ij} \tilde{p}_{ij} \log(\frac{\tilde{p}_{ij}}{\tilde{q}_{ij}}) + \sum_{j \notin \mathcal{M}_i} w^-_{ij} \tilde{p}_{ij} \log(\frac{\tilde{p}_{ij}}{\tilde{q}_{ij}}) \tag{14}$$

w^+_{ij} and w^-_{ij} are trainable and formulated as:

$$\begin{aligned} w^+_{ij} &= |(1-\omega) + \tilde{p}_{ij} \cdot \omega - \tilde{q}_{ij}|^{\gamma^+} \\ w^-_{ij} &= |\tilde{p}_{ij} \cdot \omega - \tilde{q}_{ij}|^{\gamma^-} \end{aligned} \tag{15}$$

with a little ambiguity, here $|\cdot|$ means absoulute value. ω controls the balance and is set to 0.5 in default. γ^+ and γ^- re-balance positive-negative learning and $\gamma^+ \ll \gamma^-$.

Clearly, w^+_{ij} and w^-_{ij} are proportional with distance between both teacher outputs / student outputs and ground-truth / student outputs. Intuitively, the rationality lies in:

- By measuring distance of outputs between teachers and students, w^+_{ij} and w^-_{ij} slow the distillation of easy logits (true-positives and true-negatives), and focus on conquering hard logits (false-positives and false-negatives).
- By measuring distance of outputs between ground-truth and students, w^+_{ij} and w^-_{ij} provide a synthetical evaluation of the fidelity of both teacher and student, which would largely avoid the misleading of teacher, especially in SKD which always employs on-the-fly guidance of an unreliable teacher.

the MBD loss then can be formulated as:

$$\mathcal{L}^{(i)}_{\text{MBD}} = \mathcal{L}_{\text{ReD}}(\text{RS}(\boldsymbol{p}_i/\tau), \text{RS}(\boldsymbol{q}_i/\tau)) \cdot \tau^2 \tag{16}$$

where τ is the temperature. Following (Chen et al. 2020), τ^2 is multiplied to ensure that the relative contribution of distillation loss remains roughly unchanged when combined with other losses. Further analysis of how our MBD works and ablation studies are attached to **Appendix D**.

Training Pipeline

Finally, the overall loss function is the weighted combination of our three decoupling losses and the BCE loss:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \iota \cdot \mathcal{L}_{\text{LSD}} + \sum_i (\kappa \cdot \mathcal{L}^{(i)}_{\text{RSD}} + \lambda \cdot \mathcal{L}^{(i)}_{\text{SSD}}) \tag{17}$$

where ι, κ, and λ are hyper-params to balance each loss's contribution. We will discuss them in ablation studies.

Experiments

Experiment Settings

Datasets and Baselines. Three benchmark datasets are employed in our experiments: Pascal-VOC 2007 (Everingham 2009), MS-COCO (Lin et al. 2014), and MIRFLICKR (Huiskes and Lew 2008). See **Appendix A** for a detailed introduction. For baselines, see **Appendix B** for the introduction of comparison methods.

Evaluation Metrics. We report five commonly used metrics for evaluation: (i) mean Averaged Precision (mAP), (ii) Precision of top-1 predictions (P@1), (iii) Recall of top-1 predictions (R@1), (iv) macro-F1 score (CF1), and (v) micro-F1 score (OF1). Employing both (ii) and (iii) would offer a more comprehensive view of performance on both positive and negative samples.

Implementations. We employ three backbones: ResNet-34, MobileNet_v2, and Swin-Transformer Tiny (Swin-T), which are the representations of CNN-based classical models, CNN-based lightweight models, and ViT-based models respectively. Especially for Swin-T, we employ two heads and adopt an attention dropout of 0.4. For all experiments we train for 80 epochs. Images are random-augmented (Cubuk et al. 2020) and resized to 224×224. We employ SGD as the optimizer and the momentum and weight decay are set to 0.9 and $5e^{-4}$ respectively, combined with a cosine-annealing scheduler for learning rate (LR). For Pascal-VOC 2007 and MIRFLICKR, a batch size of 64 is employed and the initial learning rate is set to 0.01. For COCO, we set the batch size as 128, and the initial learning rate as 0.1. Full parameter settings are listed in **Appendix F**.

Experiment Results

(I) Does MSKD outperforms existing state-of-arts? We show comparison results with SOTA based on three backbones and datasets in Table 1 and 2. We observe that:

- Most previous works perform minor improvements. In Table 1, PS-KD only receives 0.02% mAP improvement compared with Vanilla on ResNet34. mAP of DLB and DDGSD even decrease by -0.31% and -1.28%, accompanied by severe precision-recall imbalance, which also makes the F1 score deteriorated. We argue that the over-emphasis on consistency in DLB and DDGSD may make the model conservative to positive prediction. UD and USKD outperform the baseline mAP by 0.79% / 2.94% respectively on ResNet34. However, they both received negative results on Swin-T, with mAP dropped by 1.91% / 0.13% and precision dropped by 5.29%, revealing their limited applications. Moreover, almost all methods fail on large datasets like MIRFLICKR and COCO as shown in Table 2, which may be ascribed to their naive and incompetent self-teacher.
- Our MSKD exemplifies superior performance and balanced Precision-Recall. On Pascal VOC 2007, MSKD outperforms the previous state-of-art in mAP by 2.84% on average, with precision boosts while the recall still slightly increases. In particular, a significant increment of all metrics with MobileNet_v2 is observed, proving that

Methods	ResNet34					MobileNet_v2					Swin-T				
	mAP	P@1	R@1	CF1	OF1	mAP	P@1	R@1	CF1	OF1	mAP	P@1	R@1	CF1	OF1
Vanilla	82.51	79.46	72.01	75.93	78.56	72.88	76.04	46.91	58.02	59.11	87.89	77.41	87.02	81.93	82.15
TF-KD	82.67	76.90	74.22	75.53	78.68	71.56	74.95	50.39	57.44	59.75	88.39	73.32	88.17	79.40	80.07
PS-KD	82.53	79.14	72.14	75.48	78.39	72.99	76.44	46.34	57.70	58.32	84.42	77.98	78.47	78.22	78.38
DLB	82.20	87.51	61.51	72.24	75.58	74.42	83.47	37.27	51.53	54.49	88.20	78.94	85.96	82.30	83.46
DDGSD	81.23	**87.99**	59.23	70.80	74.70	81.51	**89.75**	52.07	65.90	66.83	86.19	**83.96**	76.78	80.21	81.46
USKD	85.44	87.32	72.40	79.16	78.66	80.00	84.15	65.13	73.42	74.05	87.76	72.15	88.73	79.59	79.55
MulCon	83.46	79.26	72.64	76.34	78.21	72.86	76.18	46.99	58.12	60.30	82.43	78.31	72.93	75.53	75.58
UD	83.30	80.93	72.72	76.60	78.09	80.89	81.17	68.27	74.24	76.51	85.98	71.41	85.85	77.97	78.75
MSKD	**86.83**	85.42	**74.98**	**79.37**	**80.29**	**82.62**	85.38	**69.43**	**76.58**	**78.77**	**89.16**	80.50	**88.77**	**84.43**	**85.79**

Table 1: Comparison experiments on Pascal VOC 2007 based on ResNet34, MobileNet_v2, and Swin-T.

Dataset	Methods	ResNet34					Swin-T				
		mAP	P@1	R@1	CF1	OF1	mAP	P@1	R@1	CF1	OF1
MIRFLICKR	Vanilla	76.33	77.98	60.68	68.88	78.85	80.69	76.99	**73.76**	75.34	83.16
	DLB	75.46	76.39	58.97	66.56	72.07	80.46	81.69	65.01	72.40	80.78
	USKD	74.68	78.71	59.44	67.63	73.53	79.64	78.52	64.71	70.95	80.43
	UD	75.46	76.39	58.97	66.56	78.48	80.93	80.32	66.51	74.64	82.80
	MSKD	**78.54**	**82.90**	**60.75**	**71.09**	**80.80**	**81.77**	**81.56**	70.20	**75.46**	**83.48**
COCO	Vanilla	66.26	73.64	49.01	59.78	65.99	71.94	76.09	56.70	64.98	68.24
	DLB	65.25	**78.21**	44.26	56.53	64.52	70.77	75.19	52.93	62.54	68.59
	USKD	61.06	76.68	42.01	54.28	66.48	70.54	**76.84**	54.95	64.70	58.62
	UD	63.71	73.80	48.45	58.50	59.27	70.86	75.44	56.59	64.77	69.62
	MSKD	**67.02**	74.96	**54.84**	**63.84**	**67.19**	**73.62**	76.17	**61.71**	**68.18**	**71.61**

Table 2: Extended comparisons on MIRFLICKR and COCO. For backbones, ResNet34 / Swin-T are selected.

$\mathcal{L}_{\text{LSR}}$	$\mathcal{L}_{\text{TSR}}$	$\mathcal{L}_{\text{RSR}}$	mAP	P@1	R@1	CF1	OF1
			82.51	79.46	72.01	75.93	76.76
✓			83.45	83.06	73.00	76.80	77.78
	✓		83.64	87.79	67.00	73.36	74.09
		✓	84.18	**87.94**	67.68	73.87	73.62
✓	✓		83.54	86.13	74.16	79.37	80.63
✓		✓	85.79	87.53	69.69	77.62	78.32
✓	✓✗	✓✗	84.43	85.70	68.51	76.95	77.58
✓	✓	✓	**86.83**	85.42	**74.98**	**79.97**	**81.17**

Table 3: Ablation study of three mechanisms and MBD. ✓means to employ the loss. ✓✗means to employ the loss but MBD is replaced with softmax+KL-Div.

MSKD dramatically reverses the corrupted optimization. Also, unlike stagnant peer works, MSKD performs remarkably on MIRFLICKR and COCO, with an average increment of 1.5% in mAP and 2% in F1. MSKD's success may lie in its proactive solution to inherent imbalance which was highly neglected previously. By decoupling the intricate global semantics with cropping and merging, models applied with MSKD obtain more specific representations of every single target.

layers	mAP	P@1	R@1	CF1	OF1
1	86.12	**85.53**	71.73	78.56	79.95
2	**86.83**	85.42	**74.98**	**79.37**	**80.29**
3	83.66	85.85	65.49	74.33	75.70

Table 4: Ablation study of graph propagation times $\mathcal{T}$.

(II) Does MSKD work well in downstream tasks? We further perform an image retrieval task to discover MSKD's application ability for downstream tasks. Following (Yang et al. 2023a), we employ k-nn algorithm to retrieve top-5 correlated images and depict our results in Figure 3. It's obvious that our MSKD retrieves more accurate images than UD, indicating MSKD gives even more concentration to small targets and learns to distinguish classes well.

(III) How do components and hyper-parameters affect the results? We answer this problem with following ablation studies:

- Components. Table 3 reports the ablation study results of three mechanisms and MBD on Pascal VOC with ResNet34 as the backbone. It can be observed that each mechanism performs well in mAP with an average improvement of 1%. Plus, we find that: $\mathcal{L}_{\text{LSR}}$ is better in precision-recall balancing but not distinguished in mAP, while $\mathcal{L}_{\text{TSR}}$ and $\mathcal{L}_{\text{RSR}}$ are the opposite. This mutual compensation forms a generalized improvement across all metrics when these mechanisms are

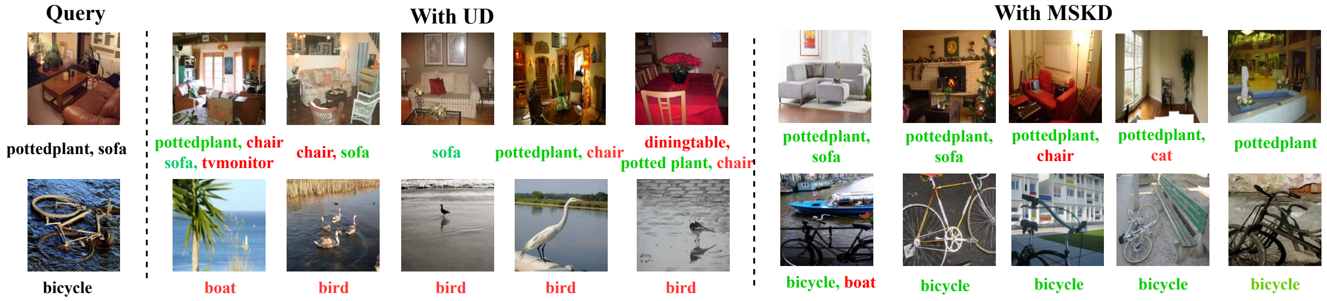

Figure 3: Performance of proposed MSKD on Image Retrieval. The first column is query images with labels. Each row of the following columns exemplifies the retrieved images and labels, sorted by relevance in descending order. Labels marked with green and red denote that they are included / not included in query labels respectively.

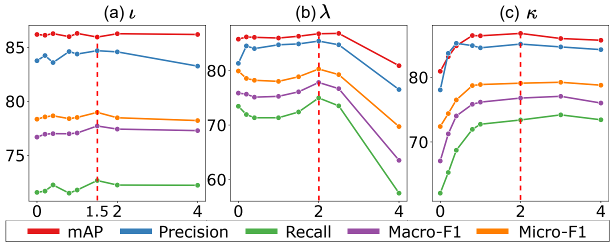

Figure 4: Ablation study of hyperparams ι, λ and κ.

combined, as depicted in the last row. In extra, to employ MBD provides further advancements in all metrics.

- Coefficients ι, λ, and κ. Results are shown in Figure 4 where optima are marked with red lines. It's obvious that:

- Change of ι causes minor effects, possibly for the small magnitude of $\mathcal{L}_{\text{LSD}}$. We set ι as 1.5 where local maximum locates.
- When λ grows larger than 2, all metrics are corrupted. We suggest that S-SD is poisoned by over-emphasizing the homogeneity. Slight improvements in all metrics are observed when λ changes from 0 to 1, so we set λ to 2.
- All metrics deteriorate when κ approaches 0. There exists about 0.5% decline in precision compared with the local maximum when λ is 4, which would ascribe to over-utilization of unreliable semantics in the initial stages. We set κ to 2.0 as the trade-off.

- Propagation Times $\mathcal{T}$. We empirically set $\mathcal{T}$ from 1 to 3 to investigate the ablation results. As demonstrated in Table 4, generally MSKD reaches its best performance when $\mathcal{T}$ is 2, slightly outperforming that when $\mathcal{T}$ is 1, which be attributed to better expressivity of two-layer propagation. However, all metrics suffer from dramatic decrease when $\mathcal{T}$ is set to 3, where over-smoothing may get severe.

(IV) Whether MSKD truly alleviates the inherent imbalance? To validate this we collect the area of all visual targets in Pascal VOC 2007 utilizing their bounding boxes, then classify each target into 22 disjoint area ranges accordingly (since all images are of 224×224). Then we calculate correctly-predicted targets with MSKD, UD, and Vanilla and demonstrate them by each area set, as shown in Figure 5. While UD has minor effects in promoting models to identify small targets and even poisons the large-scale recognition, MSKD receives up to 12% improvements in recognizing small targets and significantly shortens the distance to total numbers (the goal) while keeping undisturbed in large targets. See **Appendix E** for further analysis of how MSKD works.

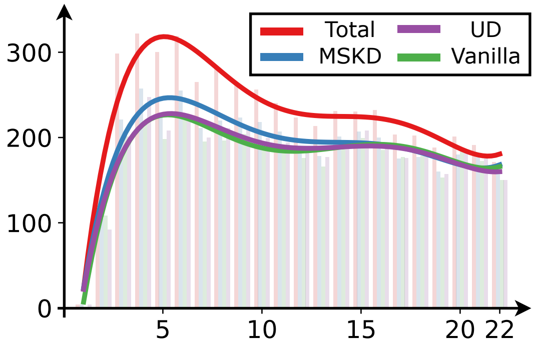

Figure 5: Number of total targets, correctly predicted targets w/ MSKD, UD or Vanilla on Pascal VOC 2007 shown by region scales, of which area is bounded by pow of adjacent x-axis values multiply 10. For example, x-axis value 3 represents regions of which area is less than $(10 \times 3)^2$ but more than $(10 \times 2)^2$. Fitting curve is plotted in extra to show the distributions along different areas.

Conclusion

We propose to extend effective SKD methods from SLL to MLL and design a tailored SKD named MSKD for the first time. Faced with the inherent imbalance of visual targets and labels in MLL, MSKD incorporates three spatial decoupling mechanism where detail semantics in each image is magnified in nunanced perspectives, supported by tailored distillation loss to propel the unbiased self-supervision. Experiments validate the superior performance as well as the general applicability of MSKD. In future we would like to further enhance MSKD with other self-teachers.

Acknowledgments

This work was supported by the Natural Science Foundation of China Youth Project (No. 62402472), the Natural Science Foundation of Jiangsu Province of China Youth Project (No. BK20240461), the National Natural Science Foundation of China (No.62072427, No.12227901), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR- 005), and Academic Leaders Cultivation Program, USTC.

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, D.; Mei, J.-P.; Wang, C.; Feng, Y.; and Chen, C. 2020. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 3430–3437.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5177–5186.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

Everingham, M. 2009. The PASCAL visual object classes challenge 2007. In *http://www. pascal-network. org/challenges/VOC/voc2007/workshop/index. html*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huiskes, M. J.; and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 39–43.

Kim, K.; Ji, B.; Yoon, D.; and Hwang, S. 2021. Self-knowledge distillation with progressive refinement of targets. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6567–6576.

Lee, C.-W.; Fang, W.; Yeh, C.-K.; and Wang, Y.-C. F. 2018. Multi-label zero-shot learning with structured knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1576–1585.

Li, M.; Wang, D.; Liu, X.; Zeng, Z.; Lu, R.; Chen, B.; and Zhou, M. 2023. Patchct: Aligning patch set and label set with conditional transport for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15348–15358.

Liang, J.; Li, L.; Bing, Z.; Zhao, B.; Tang, Y.; Lin, B.; and Fan, H. 2022. Efficient one pass self-distillation with zipf's label smoothing. In *European conference on computer vision*, 104–119. Springer.

Lin, T. 2017. Focal Loss for Dense Object Detection. *arXiv preprint arXiv:1708.02002*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, S.; Zhang, L.; Yang, X.; Su, H.; and Zhu, J. 2021. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*.

Pan, Q.; Yan, M.; Li, G.; Li, J.; and Ma, Y. 2022. Self-knowledge distillation from target-embedding AutoEncoder for multi-label classification. In *2022 IEEE International Conference on Knowledge Graph (ICKG)*, 210–216. IEEE.

Pei, H.; Wei, B.; Chang, K. C.-C.; Lei, Y.; and Yang, B. 2020. Geom-gcn: Geometric graph convolutional networks. *arXiv preprint arXiv:2002.05287*.

Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 82–91.

Ridnik, T.; Sharir, G.; Ben-Cohen, A.; Ben-Baruch, E.; and Noy, A. 2023. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 32–41.

Schultheis, E.; Wydmuch, M.; Babbar, R.; and Dembczynski, K. 2022. On missing labels, long-tails and propensities in extreme multi-label classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1547–1557.

Shen, Y.; Xu, L.; Yang, Y.; Li, Y.; and Guo, Y. 2022. Self-distillation from the last mini-batch for consistency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11943–11952.

Song, L.; Wu, J.; Yang, M.; Zhang, Q.; Li, Y.; and Yuan, J. 2021. Handling difficult labels for multi-label image classification via uncertainty distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2410–2419.

Wang, H.; Li, Z.; Huang, J.; Hui, P.; Liu, W.; Hu, T.; and Chen, G. 2021. Collaboration based multi-label propagation for fraud detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2477–2483.

Wang, Y.; He, D.; Li, F.; Long, X.; Zhou, Z.; Ma, J.; and Wen, S. 2020. Multi-label classification with label graph superimposing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12265–12272.

Xu, J.; Huang, S.; Zhou, F.; Huangfu, L.; Zeng, D.; and Liu, B. 2022. Boosting multi-label image classification with complementary parallel self-distillation. *arXiv preprint arXiv:2205.10986*.

Yang, C.; An, Z.; Zhou, H.; Cai, L.; Zhi, X.; Wu, J.; Xu, Y.; and Zhang, Q. 2022. Mixskd: Self-knowledge distillation from mixup for image recognition. In *European Conference on Computer Vision*, 534–551. Springer.

Yang, C.; Xie, L.; Su, C.; and Yuille, A. L. 2019. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2859–2868.

Yang, P.; Xie, M.-K.; Zong, C.-C.; Feng, L.; Niu, G.; Sugiyama, M.; and Huang, S.-J. 2023a. Multi-label knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 17271–17280.

Yang, Z.; Zeng, A.; Li, Z.; Zhang, T.; Yuan, C.; and Li, Y. 2023b. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17185–17194.

Yun, S.; Park, J.; Lee, K.; and Shin, J. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13876–13885.

Zhang, D.; Zhao, S.; Duan, Z.; Chen, J.; Zhang, Y.; and Tang, J. 2020a. A multi-label classification method using a hierarchical and transparent representation for paper-reviewer recommendation. *ACM Transactions on Information Systems (TOIS)*, 38(1): 1–20.

Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; and Ma, K. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3713–3722.

Zhang, P.; and Wu, M. 2024. Multi-Label Supervised Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16786–16793.

Zhang, X.-Y.; Shi, H.; Li, C.; and Li, P. 2020b. Multi-instance multi-label action recognition and localization based on spatio-temporal pre-trimming for untrimmed videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12886–12893.