CLASS-AWARE MARGIN OPTIMIZATION: RESOLVING KL DIVERGENCE-CLASSIFICATION INCONSISTENCY IN DECOUPLED KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Logit-based knowledge distillation (KD) for classification offers a cost-efficient alternative to feature-based and relation-based methods. However, logit-based knowledge distillation methods fail to explicitly model the discriminative margins between target and non-target categories. This limitation primarily stems from their overemphasis on globally minimizing KL divergence during distillation, which rigidly enforces distribution-level similarity while neglecting the inconsistency between KL divergence and classification. To address this limitation, we propose a **C**lass-**A**ware **M**argin **O**ptimization method for **D**ecoupled **K**nowledge **D**istillation (**CAMO-DKD**). CAMO-DKD enforces the margin constraint by aligning the student's decoupled predictions with the teacher's decoupled distribution through a margin-sensitive loss. This enhances the separation between the target class and competing non-target categories, thereby preventing boundary collapse during distillation. The proposed CAMO-DKD method can enhance any logit-based KD approach in a plug-and-play manner. Additionally, we provide a detailed theoretical analysis demonstrating the generalization capabilities of CAMO-DKD. Experiments on CIFAR-100, ImageNet and COCO show consistent improvements, while outperforming state-of-the-art KD methods.

1 INTRODUCTION

Deep neural networks (DNNs) have become extensively employed across multiple domains including computer vision Yang et al. (2025); Russakovsky et al. (2015); Yuan et al. (2020); Fong & Motani (2025), data mining Yang et al. (2023); Hao et al. (2024); Gindra et al. (2025), and machine learning Li et al. (2023a); Wang et al. (2024); Huang et al. (2025). Improved performance typically increases structural complexity and training costs, particularly in the large-model era where billion-parameter architectures demand substantial computational resources. This paradigm has driven lightweight network architectures to gain prominence as a prevalent development trend in DNN research Gou et al. (2021); Chen et al. (2025a); Yu et al. (2025). Among compression methodologies, knowledge distillation operates as a transfer learning mechanism designed to distill knowledge from large-scale teacher models into compact student networks.

Knowledge distillation was spawned three methodological branches: logit-based Zheng & Yang (2024); Zhou et al. (2021); Wei et al. (2024); Sun et al. (2024a); Xiang et al. (2025); Hossain et al. (2025), feature-based Cao et al. (2025); Guo et al. (2023); Chen et al. (2022); Feng et al. (2025); Cao et al. (2025); Chen et al. (2025b), and relation-based approaches Chae & Heo (2025); Huang et al. (2022); Peng et al. (2019); Yang et al. (2025); He et al. (2025). For instance, KDAS Chae & Heo (2025) finds that faster knowledge distillation can be achieved by using data with a large amount of high-quality knowledge in distillation and proposes an efficient adaptive sampling method. MoVE-KD Cao et al. (2025) distills the unique proficiencies of multiple vision encoders into a single encoder model to mitigate conflicts and retain the unique characteristics of each teacher encoder. Despite their sophistication, feature/relation-based methods suffer from computational intensity compared to logit-level distillation, driving recent research resurgence in streamlined logit-based paradigms Sun et al. (2024a); Xiang et al. (2025); Zheng & Yang (2024); Zhou et al. (2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

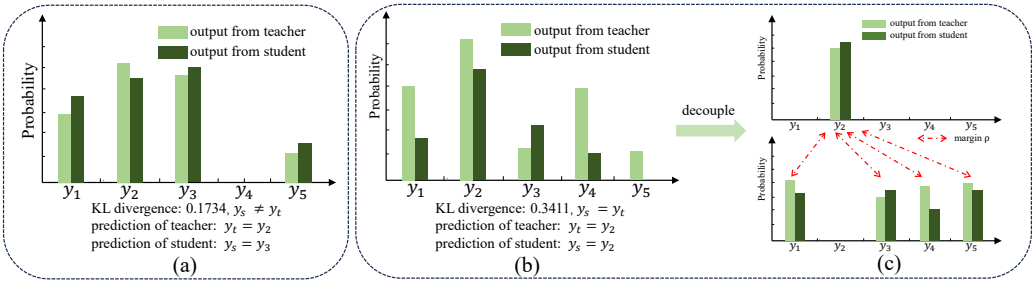

Figure 1: Illustration of the inconsistency between KL divergence and classification, along with the proposed class-aware margin. (a): Smaller KL-divergence, yet the student's predicted label is incorrect. (b): Larger KL-divergence, but the student model's predicted label (y_2) matches the teacher model's label (y_2). (c): A margin is incorporated into the DKD process to address the inconsistency issue. A margin ρ is maintained between the highest and the rest labels.

Decoupled Knowledge Distillation (DKD) Zhao et al. (2022) establishes that logit-level distillation achieves competitive performance with feature-based counterparts at reduced computational overhead. However, DKD confronts a critical methodological limitation - inconsistency between KL divergence and classification Wang & Geng (2021); Gao et al. (2018); Wang & Geng (2019); Lv et al. (2024); Gao et al. (2025). In some fine-grained classification tasks, the margin between the target and non-target labels are too small to discriminate. For example, simply analyzing from facial images, "Shiba Inus" and "Akita Inus" look very similar in appearance and the predicted probabilities of the labels may very close, which easily leads to confusion during the knowledge distillation stage by KL-divergence. The fundamental conflict stems from KL divergence's distributional alignment objective versus classification's discriminative learning goal. Despite target label decoupling, proximity between target/non-target class confidences can still induce prediction divergence. Fig.1 demonstrates this inconsistency scenario: Smaller KL divergence (Fig.1 (a)) yields mismatched classification (y_3 *vs.* teacher's y_2), while larger divergence (Fig.1 (b)) preserves classification consistency (y_2). Therefore, the inconsistency between KL divergence minimization and classification objectives can induce the transfer of misleading knowledge from the teacher network. The KL divergence-classification inconsistency is a long-standing unresolved issue in logit-based knowledge distillation.

To mitigate the discrepancy, we propose a **C**lass-**A**ware **M**argin **O**ptimization method for **D**ecoupled **K**nowledge **D**istillation (**CAMO-DKD**). CAMO-DKD introduces class-specific margin constraints during distillation to enforce discriminative separation between target/non-target categories (Fig.1 (c)), resolving the KL divergence-classification misalignment. Our class-aware margin optimation distillation objective employs L_1-norm for superior probability difference measurement across labels. Theoretically, we establish that both target and non-target classes contribute to KD's generalization bound (Theorems 1-3). Experiments demonstrate plug-and-play enhancement of existing logit-based methods, achieving consistent performance on CIFAR-100, ImageNet and COCO datasets.

In summary, our contributions are as follows:

- We conduct a theoretical analysis showing that both target and non-target classes are necessary for the generalization of DKD.
- We propose a class-aware margin optimization method to address the inconsistency problem between KL divergency and classifcation in DKD.
- We perform extensive experiments under various teacher-student pairs, demonstrating the effectiveness of the proposed class-aware margin optimization method for DKD.

2 Related Work

The concept of knowledge distillation was introduced by Hinton *et al.* Hinton (2015), who posited that softened labels encapsulate implicit knowledge. Subsequently, a variety of logit-based knowl-

edge distillation techniques have been developed Yang et al. (2023); Zhang et al. (2023); Jin et al. (2023); Sun et al. (2024a); Kim et al. (2024); Li et al. (2023c); Xiang et al. (2025); Zhou et al. (2025).

Feature-based KD approaches Yang et al. (2021); Chen et al. (2022); Guo et al. (2023); Chen et al. (2025a); Cao et al. (2025); Yu et al. (2025) focus on aligning the intermediate layers of the student model with those of the teacher model. For example, Align-KD Feng et al. (2025) is proposed to guide the student model to learn the cross-modal matching that occurs at the shallow layer, which has been applied in migrating VLMs to edge devices. CAT-KD Guo et al. (2023) re-examines the architecture of mainstream CNN models and highlights the importance of identifying class-discriminative regions in the input for effective classification. CKD Wencheng et al. (2024) presents a contrastive knowledge distillation approach, which can be formulated as a sample-wise alignment problem with intra- and inter-sample constraints. SMoDi Chen et al. (2025a) treats each stage of LIC models as an independent sub-task, mirroring the teacher model's task decomposition to the student, thereby simplifying knowledge transfer. Nevertheless, a notable drawback of feature-based KD is its relatively high computational complexity and reduced computational efficiency.

Relation-based methods represent another category of knowledge distillation techniques Chae & Heo (2025); Hao et al. (2024); Zheng et al. (2024); Park et al. (2019); Yang et al. (2025); He et al. (2025), which explore the relationships between different layers or data samples in greater depth. For example, He *et al.* He et al. (2025) observes existing KD loss cannot perform well when most of samples are difficult in the distillation dataset and proposed to dynamically adjust the distillation dataset based on the difficulty of samples. CPKD Yang et al. (2025) introduces a knowledge prototype as a regulatory mechanism for the teacher's learning process and effectively addresses key challenges in anomaly detection, including insufficient learning and overadaptation. However, the main limitations of relation-based KD remain tied to concerns regarding complexity and cost efficiency.

Recently, logit-based knowledge distillation approaches have rekindled the interest of researchers. In addition to DKD Zhao et al. (2022), DOT Zhao et al. (2023), TTM Zheng & Yang (2024), WKD Lv et al. (2024), Logit standardization KD Sun et al. (2024b) introduces a plug-and-play Z-score pre-processing step for logit standardization prior to applying softmax and KL divergence. SDD Wei et al. (2024) decouples the global logit output into multiple local logit outputs and establishes distillation pipelines for them. This helps the student to mine and inherit fine-grained and unambiguous logit knowledge. EKD Xiang et al. (2025) distills both the expectation of the teacher distribution and the distribution itself into the student. The expectation captures the macroscopic characteristics of the distribution, while the distribution itself conveys microscopic information about the classification boundaries. TeKAP Hossain et al. (2025) enhances the distillation at both the feature and logit levels. Nevertheless, all logit-based KD methods rely on KL divergence to assess the overall similarity between the outputs of teacher and student models, resulting in the inconsistency problem with the classification Wang & Geng (2021); Gao et al. (2018); Wang & Geng (2019); Lv et al. (2024); Gao et al. (2025). In particular, in fine-grained classification, the logits values among subclasses are extremely close, which may lead to a scenario where the KL metric yields an extremely small value but the classification result is incorrect as shown in Fig.1. To address this, we propose a class-aware margin optimization method in this paper to resolve the inconsistency issue.

3 METHODOLOGY

3.1 NOTATIONS

The training dataset is denoted as $\mathcal{D}$, containing N samples $\{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$, where the input space is denoted by $\mathcal{X} \subset \mathbb{R}^q$. For each sample $\boldsymbol{x} \in \mathcal{X}$, a probability distribution $y_{\boldsymbol{x}} = [p_{\boldsymbol{x}}^1, ..., p_{\boldsymbol{x}}^k]$ can be obtained from either the teacher or student model, where k is the size of the label space $\mathcal{Y}$. Given a training set with n samples $S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_1), ..., (\boldsymbol{x}_n, y_n)\}$, logit-based knowledge distillation aims to minimize the difference between the outputs of the teacher and student models.

In contrast to other logit-based KD methods that rely on KL divergence as the loss function, we employ the L_1-norm to further quantify the differences between the probabilities of each label.

$$\mathcal{L}(y_t||y_s) = \sum_{i,j} |p_{s_i}^j - p_{t_i}^j|. \tag{1}$$

Let $y \in \mathcal{Y}$ be the random label variable. For the output of the student model, the predicted probability distribution, denoted as $y_{\boldsymbol{x}}$, is the conditional probability distribution function, *i.e.* $p^j_{\boldsymbol{x}} = \mathbb{P}(y = y^j_{\boldsymbol{x}}|\boldsymbol{x})$.

3.2 Decouple the Target Label Distribution

As discussed in Section 1, knowledge distillation encounters the challenge of inconsistency when KL divergence is used as the KD loss. To tackle this issue, inspired by LDL-HR Wang & Geng (2021), DILDL Gao et al. (2025), DKD Zhao et al. (2022) and WKD Lv et al. (2024), we first decompose the KD process into target label distillation and non-target label distillation. Subsequently, a class-aware constraint factor is integrated into the decoupled KD loss.

Firstly, we define the degenerated probability distribution. Specifically, for the sample $\boldsymbol{x}$, the output of the teacher model is defined as $\bar{y}_{\boldsymbol{x}} = [\bar{p}^1_{\boldsymbol{x}}, \bar{p}^2_{\boldsymbol{x}}, ..., \bar{p}^k_{\boldsymbol{x}}]^T$, where $\bar{p}^j_{\boldsymbol{x}}$ is defined by

$$\bar{p}^j_{\boldsymbol{x}} = \begin{cases} 1 & if \quad y_j = \arg\max_i \bar{p}^i_{\boldsymbol{x}}, \\ 0 & otherwise. \end{cases} \tag{2}$$

where y_j represents the label with the highest probability, and the decoupled degenerated probability distribution assigns a probability of 1 to the target label while assigning 0 to all other labels. When learning the decoupled target label, the KD process primarily emphasizes the target label provided by the teacher model. As a result, the decoupled target label approach effectively mitigates the inconsistency problem.

Unlike DKD, which employs KL divergence as the KD loss, we adopt the L_1-norm loss to measure the difference between the degenerated probability distributions of the teacher and student models. We assume that the logits of the student model are denoted as z_s, and the output of the student model is obtained by applying the softmax function to these logits.

$$p^l_s = \frac{exp(z^l_s)}{\sum_{j=1}^k exp(z^j_s)}, \quad for \quad l \in [k]. \tag{3}$$

Then the KD can be cast as the following optimization problem,

$$\min_{\boldsymbol{W}} \sum_{i,j} \left| p^j_s - \bar{p}^j_t \right|, \tag{4}$$

where $\boldsymbol{W}$ are parameters of the model. We theoretically prove that the decoupled degenerated probability distribution is supported by theoretical guarantees.

Theorem 3.1. *Let $\bar{y}_{\boldsymbol{x}}$ be the decoupled target label distribution as Eq. (2) and p_s be the learned output of student, and f be the classifier defined as $f(x) = \arg\max_{y_s \in \mathcal{Y}} p_s^{y_j}$. Then the expected 0/1 loss of f satisfies the following bound*

$$\mathbb{P}(f(x) \neq y_t) - L^*_1 \leq \mathbb{E}_{\boldsymbol{x} \sim S} \left[\sum_j |p_s^{y_j} - \bar{p}_t^{y_j}| \right]. \tag{5}$$

Proof. The predicted probability distribution of the student model is assumed to represent a conditional probability distribution function. We define the classifier accordingly, ensuring its outputs align with Bayes classifier Devroye et al. (2013); Wang & Geng (2021).

$$f^*(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} \bar{p}^{y_j}_{\boldsymbol{x}}, \tag{6}$$

expected 0/1 loss of which is the Bayes error, *i.e.,* $L^*_1 = \mathbb{P}(f^*(\boldsymbol{x}) \neq y)$. Fix any $\boldsymbol{x}$, then we have

$$\begin{aligned} \mathbb{P}(f(\boldsymbol{x}) \neq y|\boldsymbol{x}) &= 1 - \mathbb{P}(f(\boldsymbol{x}) = y|\boldsymbol{x}) \\ &= 1 - \sum_{y_j : y_j = f(\boldsymbol{x})} \mathbb{P}(y = y_j|\boldsymbol{x}). \end{aligned} \tag{7}$$

Let $f(\boldsymbol{x}) = y_l$ and $f^*(\boldsymbol{x}) = y_k$. Then, it follows that

$$\mathbb{P}(f(\boldsymbol{x}) \neq y|\boldsymbol{x}) - \mathbb{P}(f^*(\boldsymbol{x}) \neq y|\boldsymbol{x}) = p_s^{y_k} - p_s^{y_l}. \tag{8}$$

If $y_l = y_k$, the $p_s^{y_k} - p_s^{y_l} = 0$. If $y_l \neq y_k$, the $p_s^{y_k} - p_s^{y_l}$ is bounded by $\bar{p}_s^{y_k} - \bar{p}_s^{y_l}$. Notice that $\bar{p}_s^{y_k} \geq \bar{p}_s^{y_l}$ and $p_t^{y_k} \leq p_t^{y_l}$.

Lemma 3.2. *Let $a \geq b$ and $c \geq d$, then $|a-b| \leq |a-c| + |b-d|$.*

In line with the definition of decoupled target label distribution, $y_k \neq y_l$. when $y_k = y_l$, the right-hand equals 0. Therefore, according to Lemma 3.2, we have,

$$\bar{p}_s^{y_k} - \bar{p}_s^{y_l} \leq |\bar{p}_s^{y_k} - p_t^k| + |\bar{p}_s^{y_l} - p_t^l|, \tag{9}$$

which equals to,

$$\mathbb{P}(f(\boldsymbol{x} \neq y|\boldsymbol{x})) - \mathbb{P}(f^*(\boldsymbol{x}) \neq y|\boldsymbol{x}) \leq \sum_{j=1}^{m} |p_s^{y_j} - \bar{p}_t^{y_j}|. \tag{10}$$

We finish the proof of Theorem 1.

The Theorem 1 suggests that the expected 0/1 loss of a student classifier will approach the Bayes error when the output of the student model is sufficiently close to the decoupled probability distribution under the L_1-norm loss.

3.3 Class-Aware Margin Optimization in DKD

To further enhance classification performance and address the inconsistency problem more effectively, we incorporate a margin optimization into the decoupled knowledge distillation loss. The underlying motivation is to ensure that the predicted score of the student model for the correct label is larger than those for other labels by a margin $\rho > 0$. Consequently, the loss function in Equation (4) can be reformulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{W},\boldsymbol{\xi}} \sum_{i,j} \left| p_s^j - \bar{p}_t^j \right| + \sum_{i,j:y_j \neq y_s} \frac{\xi_{i,j}}{\rho} \\ &s.t. \quad p_t^{y_t} - p_s^j \geq \rho - \xi_{i,j}, \\ &\qquad \xi_{i,j} > 0, \forall i \in [n], \forall j \in \{l | l \in [k], y_l \neq y_t\}, \end{aligned} \tag{11}$$

where ξ is analogous to the slack variable in SVM. y_t is the label with highest description degree. i is the index of the instance. The margin constraint encourages the score of the predicted label (the one with the highest probability) to be greater than those of other labels. The introduction of margin optimization can impose a soft constraint on the distillation process to address the issue where minimizing the KL divergence causes the scores of other incorrect labels to exceed those of the correct label, thereby improving the classification accuracy of the student network.

3.4 Decouple the Non-target Label Distribution

Equation (11) focuses solely on the label with the highest probability. To further improve the generalization of knowledge distillation, the probabilities of the remaining labels are also decoupled. Specifically, the $f^{'}$ is defined as the sub-optimal classifier:

$$f^{'}(\boldsymbol{x}) = \arg \max_{j \in \mathcal{Y} \setminus \{y_s\}} \bar{p}_{\boldsymbol{x}}^j. \tag{12}$$

The sub-optimal classifier outputs the label with the highest probability from the remaining probability distribution. Similar to the decoupled target class, we employ the L_1-norm distance to measure the difference between the probability distributions of non-target labels from the teacher and student models. Consequently, Eq. (11) can be reformulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{W},\boldsymbol{\xi}} \sum_{i=1}^{n} \sum_{j=1}^{m} \left| p_s^j - \bar{p}_t^j \right| + \sum_{i,j:y_j \neq y_s} \frac{\xi_{i,j}}{\rho} \\ &\quad + \sum_{i=1}^{n} \sum_{j:y_j \neq y_s} |p_t^j - p_s^j| \\ &s.t. \quad p_t^{y_t} - p_s^j \geq \rho - \xi_{i,j}, \\ &\qquad \xi_{i,j} > 0, \forall i \in [n], \forall j \in \{l | l \in [k], y_l \neq y_t\}. \end{aligned} \tag{13}$$

Define the expected 0/1 loss of $f^{'}$ by $L_2^* = \mathbb{P}(f^{'}(\boldsymbol{x}) \neq y)$. We prove the Theorem 3 to show the generalization of decoupled non-target labels, *w.r.t.* learning the rest probability distribution of non-target labels.

Theorem 3.3. *Let $f(x)$ be a learned KD function, and f is defined as $f(x) = \arg\max_{\bar{y}\in\mathcal{Y}} p_s^{\bar{y}}$. $\bar{y}$ is the decoupled label. Then, the expected 0/1 loss of f satisfies the following bound*

$$\mathbb{P}(f(\boldsymbol{x}) \neq y) \leq L_2^* + \mathbb{E}_{\boldsymbol{x}\sim S}\left[\sum_{\bar{y}\neq y_{\boldsymbol{x}}} |p_t^{\bar{y}} - p_s^{\bar{y}}|\right]. \tag{14}$$

Proof. Let $f(\boldsymbol{x}) = y_l$ and $f^{'}(\boldsymbol{x}) = y_k$. Then,

$$\mathbb{P}(f(\boldsymbol{x})) \neq y|\boldsymbol{x}) - \mathbb{P}(f^{'}(\boldsymbol{x}) \neq y|\boldsymbol{x}) = p_s^{y_k} - p_s^{y_l}. \tag{15}$$

If $y_l = y_k$, $p_s^{y_k} - p_s^{y_l} = 0$. If $y_l \neq y_k$, then $p_s^{y_k} \geq p_s^{y_l}$ and $p_t^{y_k} \geq p_t^{y_l}$. According to Lemma 3.2, the right-hand side of the above equation is bounded by $|p_s^{y_k} - p_t^{y_k}| + |p_s^{y_l} - p_t^{y_l}|$, where $y_l \neq y_t$ and $y_k \neq y_t$. Then, we can obtain the following equation

$$\mathbb{P}(f(\boldsymbol{x}) \neq y|\boldsymbol{x}) - \mathbb{P}(f^{'}(\boldsymbol{x}) \neq y|\boldsymbol{x}) \leq \sum_{j:y_j\neq y_t} |p_t^{y_j} - p_s^{y_j}|. \tag{16}$$

which completes the proof of Theorem 3.

Theorem 3 asserts that the expected 0/1 loss of the student classifier can be bounded by the sum of two components. The first component is the expected 0/1 loss of the sub-optimal classifier, while the second is the expected $L1$-norm distance between the probability distributions of the student and teacher models for non-target labels. Theorem 2 highlights that even if the student's sub-optimal classifier does not reach the Bayes error, its performance remains constrained by that of the sub-optimal classifier.

3.5 Optimization

To solve the Eq. (13), following LDL-HR Wang & Geng (2021), we define $\ell_\rho(\boldsymbol{x}) = \max\{0, 1 - \boldsymbol{x}/\rho\}$. The final class-aware margin optimization for decoupled knowledge function is defined as,

$$\begin{aligned}\mathcal{L}_{CAMO} = &\sum_{i,j}\left|p_s^j - \bar{p}_t^j\right| + \sum_{i,j:y_j\neq y_s} \ell_\rho(p_t^{y_t} - p_s^j) + \\ &\sum_{i,j:y_j\neq y_s} |p_s^j - p_t^j|.\end{aligned} \tag{17}$$

where the first term focuses on decoupling the target class, the third term aims to decouple the non-target probability distribution, and the second term is designed to add a class-aware margin between the target label and the non-target labels. A parameter β is introduced to balance the proposed CAMO loss with other losses (e.g., KD Hinton (2015), DKD Zhao et al. (2022), WKD Lv et al. (2024), TTM Zheng & Yang (2024), SDD Wei et al. (2024), EKD Xiang et al. (2025), Cross-entropy, etc.).

4 Experiment

4.1 Datasets and Baselines

CIFAR-100 Krizhevsky & Hinton (2009) is a widely used image classification dataset, consisting of 50k training images and 10k validation images. ImageNet Russakovsky et al. (2015) is a large-scale dataset for image classification, featuring 1,000 categories. The ImageNet dataset includes approximately 1.28 million training images and 50k validation images. The object detection datasets MS-COCO Lin et al. (2014) is also used to validate the proposed method.

The proposed CAMO-DKD can be integrated into any logit-based method via a plug-and-paly manner. We choose KD Hinton (2015), DKD Zhao et al. (2022), DIST Huang et al. (2022), MLKD Jin et al. (2023), CTKD Li et al. (2023b), Logit-Standard KD (LS_KD) Sun et al. (2024a), WKD Lv et al. (2024), TTM Zheng & Yang (2024), SDD Wei et al. (2024), TeKAP Hossain et al. (2025) and EKD Xiang et al. (2025) to evaluate the effectiveness of our proposed CAMO-DKD method. Additionally, we also compare our approach with several feature-based baselines, including FitNet Romero et al. (2015), AT Zagoruyko & Komodakis (2017), OFD Heo et al. (2019), RKD Park et al. (2019), CRD Tian et al. (2020), SRRL Yang et al. (2021), ReviewKD Chen et al. (2021), SimKD Chen et al. (2022), and CAT-KD Guo et al. (2023).

Table 1: Performance on the CIFAR-100 validation set for identical architecture pairs of teachers and students. We integrate our proposed CAMO-DKD into existing logit-based knowledge distillation methods, with performance gains highlighted in subscript.

Type	Teacher Student	ResNet32x4 ResNet8x4	ResNet56 ResNet20	ResNet110 ResNet32	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	VGG13 VGG8
Feature-based	FitNet Romero et al. (2015)	73.50	69.21	71.06	73.58	72.24	71.02
	OFD Heo et al. (2019)	74.95	70.98	73.23	75.24	74.33	73.95
	AT Zagoruyko & Komodakis (2017)	73.44	70.55	72.31	74.08	72.77	71.43
	RKD Park et al. (2019)	71.90	69.61	71.82	73.35	72.22	71.48
	CRD Tian et al. (2020)	75.51	71.16	73.48	75.48	74.14	73.94
	SRRL Yang et al. (2021)	75.92	71.44	73.80	75.96	74.75	74.40
	ReviewKD Chen et al. (2021)	75.63	71.89	71.34	76.12	75.09	74.84
	SimKD Chen et al. (2022)	78.08	71.05	73.92	75.53	74.53	74.89
	CAT-KD Guo et al. (2023)	76.91	71.62	73.62	75.60	74.82	74.65
Logit-based	KD Hinton (2015)	73.33	70.66	73.08	74.92	73.54	72.98
	KD+Ours	75.01$_{+1.38}$	71.26$_{+0.60}$	73.56$_{+0.48}$	75.47$_{+0.55}$	73.99$_{+0.45}$	73.55$_{+0.57}$
	DKD Zhao et al. (2022)	76.32	71.97	74.11	76.24	74.81	74.68
	DKD+Ours	76.71$_{+0.39}$	72.35$_{+0.38}$	74.07$_{-0.04}$	76.78$_{+0.54}$	75.02$_{+0.21}$	74.92$_{+0.24}$
	CTKD Li et al. (2023b)	72.93	70.00	72.63	73.81	71.11	70.65
	CTKD+Ours	74.39$_{+1.46}$	70.52$_{+0.52}$	73.15$_{+0.52}$	74.23$_{+0.42}$	72.07$_{+0.96}$	72.29$_{+1.64}$
	KD+LS_KD Sun et al. (2024a)	76.62	71.43	74.17	76.11	74.37	74.36
	KD+LS_KD Sun et al. (2024a)+Ours	76.80$_{+0.18}$	71.95$_{+0.52}$	74.26$_{+0.09}$	76.15$_{+0.04}$	74.79$_{+0.42}$	74.42$_{+0.06}$
	MLKD Jin et al. (2023)	76.78	72.08	74.20	76.37	75.06	74.85
	MLKD+Ours	77.21$_{+0.43}$	72.59$_{+0.51}$	74.41$_{+0.21}$	76.41$_{+0.04}$	75.73$_{+0.67}$	74.86$_{+0.01}$
	SDD-DKD Wei et al. (2024)	75.43	-	-	75.02	74.62	74.31
	SDD-DKD Wei et al. (2024)+Ours	75.86$_{+0.43}$	-	-	75.92$_{+0.90}$	74.77$_{+0.15}$	74.79$_{+0.48}$
	TTM Zheng & Yang (2024)	75.63	71.27	73.74	76.37	73.74	73.90
	TTM Zheng & Yang (2024)+Ours	76.39$_{+0.76}$	71.75$_{+0.48}$	73.92$_{+0.18}$	76.57$_{+0.20}$	75.32$_{+1.58}$	74.58$_{+0.68}$
	WTTM-CRDZheng & Yang (2024)	76.12	72.00	73.81	76.60	74.75	74.63
	WTTM-CRDZheng & Yang (2024)+Ours	76.56$_{+0.44}$	72.41$_{+0.41}$	74.21$_{+0.40}$	77.45$_{+0.85}$	75.03$_{+0.28}$	74.82$_{+0.19}$
	MLKD+LS_KD Sun et al. (2024a)	76.68	71.34	74.31	76.22	75.37	74.11
	MLKD+LS_KD Sun et al. (2024a)+Ours	77.14$_{+0.45}$	71.81$_{+0.47}$	74.51$_{+0.20}$	76.30$_{+0.08}$	75.87$_{+0.50}$	74.89$_{+0.78}$
	TeKAP Hossain et al. (2025)	74.49	70.57	73.36	75.63	73.74	73.42
	TeKAP Hossain et al. (2025)+Ours	74.71$_{+0.22}$	70.89$_{+0.32}$	73.31$_{-0.05}$	75.97$_{+0.34}$	74.01$_{+0.27}$	73.84$_{+0.42}$
	EKD Xiang et al. (2025)	77.09	69.21	72.24	75.03	73.23	74.31
	EKDXiang et al. (2025)+Ours	77.28$_{+0.19}$	70.56$_{+1.35}$	72.62$_{+0.38}$	75.86$_{+0.83}$	74.11$_{+0.88}$	74.95$_{+0.64}$

4.2 Results

Table 1 demonstrates the remarkable effectiveness and versatility of the proposed CAMO-DKD when integrated into existing logit-based KD methods. For the basic KD framework, integrating CAMO-DKD yields performance improvements ranging from +0.38 (ResNet56/ResNet20) to +1.46 (ResNet32x4/ResNet8x4) across different model pairs. When combined with advanced logit-based methods like DKD, CTKD, MLKD, and TTM, CAMO-DKD still delivers tangible gains. Notably, the TTM+Ours configuration achieves a substantial +1.58 improvement on the WRN-40-2/WRN-40-1 pair, which is one of the largest gains observed in the experiments. Even for the well-optimized MLKD method, CAMO-DKD provides consistent increments (e.g., +0.43 on ResNet32x4/ResNet8x4, +0.67 on WRN-40-2/WRN-40-1), validating its compatibility with state-of-the-art logit-based distillation techniques.

Table 2 shows the performance of different teacher-student architecture pairs on the CIFAR-100 dataset. For every logit-based method (including KD, DKD, CTKD, MLKD, etc.), the "Method+Ours" variant achieves higher accuracy than the original method, with the improvement amplitude marked in red subscripts.

4.3 Object Detection Experiment

Table 3 evaluates various knowledge distillation methods on the MS-COCO object detection benchmark in terms of *AP*. "KD+Ours" achieves +0.66 AP and +0.55 AP_{50} gains over vanilla KD under the ResNet101/ResNet18 setting, while "DKD+Ours" obtains the highest AP_{50} of 57.03 and 61.31 under the ResNet101/ResNet18 and ResNet101/ResNet50 settings.

Table 2: Performance on the CIFAR-100 validation set for different architecture pairs of teachers and students. We incorporate our proposed CAMO-DKD into existing logit-based knowledge distillation methods, with performance improvements highlighted in subscript.

Type	Teacher Student	ResNet32x4 ShuffleNet-V1	ResNet50 MobileNet-V2	ResNet32x4 ShuffleNet-V2	WRN-40-2 ShuffleNet-V1	VGG13 MobileNet-V2
Feature-based	FitNet Romero et al. (2015)	73.59	63.16	73.54	73.73	64.14
	OFD Heo et al. (2019)	75.98	69.04	76.82	75.85	69.48
	AT Zagoruyko & Komodakis (2017)	68.72	58.58	72.73	64.26	59.40
	RKD Park et al. (2019)	72.28	64.43	73.21	72.21	64.52
	CRD Tian et al. (2020)	75.11	69.11	75.65	76.05	69.73
	SRRL Yang et al. (2021)	75.66	69.45	76.40	76.61	69.14
	ReviewKD Chen et al. (2021)	77.45	69.89	77.78	77.14	70.37
	SimKD Chen et al. (2022)	77.63	69.97	78.39	76.92	69.44
	CAT-KD Guo et al. (2023)	77.37	71.36	78.41	75.81	69.13
Logit-based	KD Hinton (2015)	74.07	67.35	74.45	74.83	67.37
	KD+Ours	74.36$_{+0.29}$	68.36$_{+1.01}$	75.14$_{+0.69}$	76.18$_{+1.35}$	68.29$_{+0.92}$
	DKD Zhao et al. (2022)	76.45	70.35	77.07	76.70	69.71
	DKD+Ours	76.88$_{+0.43}$	70.65$_{+0.30}$	77.33$_{+0.26}$	77.02$_{+0.32}$	69.90$_{+0.19}$
	CTKD Li et al. (2023b)	74.57	67.02	76.34	76.07	67.58
	CTKD+Ours	74.83$_{+0.26}$	67.81$_{+0.79}$	76.52$_{+0.18}$	76.26$_{+0.19}$	67.69$_{+0.11}$
	KD+LS_KD Sun et al. (2024a)	75.21	69.02	75.56	76.37	68.61
	KD+LS_KD Sun et al. (2024a)+Ours	75.46$_{+0.25}$	69.63$_{+0.61}$	76.12$_{+0.56}$	76.61$_{+0.24}$	69.35$_{+0.74}$
	MLKD Jin et al. (2023)	76.79	70.87	78.33	77.09	69.12
	MLKD+Ours	77.31$_{+0.52}$	70.94$_{+0.07}$	78.58$_{+0.25}$	77.24$_{+0.15}$	70.51$_{+1.39}$
	SDD-DKD Wei et al. (2024)	77.48	71.07	78.08	76.94	69.84
	SDD-DKD Wei et al. (2024)+Ours	77.77$_{+0.29}$	71.29$_{+0.22}$	78.42$_{+0.34}$	77.23$_{+0.29}$	70.22$_{+0.38}$
	TTM Zheng & Yang (2024)	73.11	68.10	75.48	76.19	67.07
	TTM Zheng & Yang (2024)+Ours	74.48$_{+1.37}$	69.15$_{+1.05}$	75.93$_{+0.45}$	77.16$_{+0.97}$	68.31$_{+1.24}$
	WTTM-CRDZheng & Yang (2024)	74.18	70.11	76.94	77.32	69.89
	WTTM-CRDZheng & Yang (2024)+Ours	75.09$_{+0.91}$	70.72$_{+0.61}$	77.31$_{+0.37}$	77.87$_{+0.55}$	69.97$_{+0.08}$
	MLKD+LS_KD Sun et al. (2024a)	76.48	69.70	77.60	75.46	69.29
	MLKD+LS_KD Sun et al. (2024a)+Ours	77.17$_{+0.69}$	69.70$_{+1.22}$	78.47$_{+0.87}$	77.83$_{+2.37}$	70.89$_{+1.60}$
	TeKAP Hossain et al. (2025)	74.35	69.10	75.07	75.78	68.40
	TeKAP Hossain et al. (2025)+Ours	75.30$_{+0.95}$	69.70$_{+0.60}$	75.57$_{+0.50}$	76.37$_{+0.59}$	68.49$_{+0.09}$
	EKD Xiang et al. (2025)	76.25	69.23	76.22	75.24	69.50
	EKDXiang et al. (2025)+Ours	76.62$_{+0.37}$	70.74$_{+1.51}$	77.93$_{+1.71}$	75.74$_{+0.50}$	70.24$_{+0.74}$

Table 3: Performance of object detection on MS-COCO.

Teacher/Student	ResNet101/ResNet18		ResNet101/ResNet50	
Method	AP	AP_{50}	AP	AP_{50}
FitNet Romero et al. (2015)	34.13	54.16	38.76	59.62
FGFI Wang et al. (2019)	35.44	55.51	39.44	60.27
ReviewKD Chen et al. (2021)	**36.75**	56.72	**40.36**	<u>60.97</u>
KD Hinton (2015)	33.97	54.66	38.35	59.41
KD+Ours	34.63$_{+0.66}$	55.21$_{+0.55}$	38.87$_{+0.52}$	59.88$_{+0.47}$
DKD Zhao et al. (2022)	35.05	56.60	39.25	60.90
DKD+Ours	35.29$_{+0.24}$	**57.03**$_{+0.43}$	<u>39.57</u>$_{+0.32}$	**61.31**$_{+0.41}$
WKD-L Lv et al. (2024)	35.24	56.73	-	-
WKD-L Lv et al. (2024)+Ours	<u>35.57</u>$_{+0.33}$	<u>56.96</u>$_{+0.23}$	-	-

4.4 Ablation Study

We conduct ablation studies to analyze the effectiveness of each component of the proposed method. Table 4 evaluates the impact of incorporating target label distillation, non-target label distillation, and margin optimization into knowledge distillation methods on the CIFAR-100 validation set. Across all student-teacher architecture pairs (including ResNet, WRN, ShuffleNet series), the full configuration integrating target label, non-target label, and margin consistently achieves the highest top-1 accuracy. In summary, the ablation study confirms that the joint utilization of target label, non-target label, and margin components is the key to achieving optimal distillation performance, and each component contributes non-trivially to the overall accuracy improvement.

4.5 Parameter Sensitivity Analysis

In this section, we analyze the impact of the margin value ρ and the balance parameter β on distillation performance. Fig. 2 presents the sensitivity analysis of ρ and β on CIFAR-100, where CAMO-

Table 4: Ablation study on the CIFAR-100 validation set in terms of top-1 accuracy. The ✓indicates that the corresponding component is included.

Student	Target label	Non-target label	Margin	Top-1
ResNet32x4 as the teacher				
ResNet8x4	✓	✓		74.90$_{-0.11}$
KD Hinton (2015)+Ours	✓		✓	74.26$_{-0.75}$
		✓	✓	74.72$_{-0.29}$
	✓	✓	✓	**75.01**
ResNet8x4	✓	✓		76.04$_{-0.76}$
SDD-NKD Wei et al. (2024)+Ours	✓		✓	76.23$_{-0.57}$
		✓	✓	76.45$_{-0.35}$
	✓	✓	✓	**76.80**
ShuffleNet-V1	✓	✓		76.43$_{-0.}$
EKD Xiang et al. (2025)+Ours	✓		✓	75.93$_{-0.}$
		✓	✓	76.10$_{-0.}$
	✓	✓	✓	**77.13**
WRN-40-2 as the teacher				
WRN-16-2	✓	✓		74.96$_{-0.51}$
KD Hinton (2015)+Ours	✓		✓	75.17$_{-0.30}$
		✓	✓	75.43$_{-0.04}$
	✓	✓	✓	**75.47**
ShuffleNet-V2	✓	✓		77.69$_{-0.17}$
WTTM-CRDZheng & Yang (2024)+Ours	✓		✓	77.53$_{-0.33}$
		✓	✓	77.47$_{-0.39}$
	✓	✓	✓	**77.86**
WRN-40-1	✓	✓		73.86$_{-0.25}$
TeKAP Hossain et al. (2025)+Ours	✓		✓	73.43$_{-0.68}$
		✓	✓	73.73$_{-0.38}$
	✓	✓	✓	**74.11**

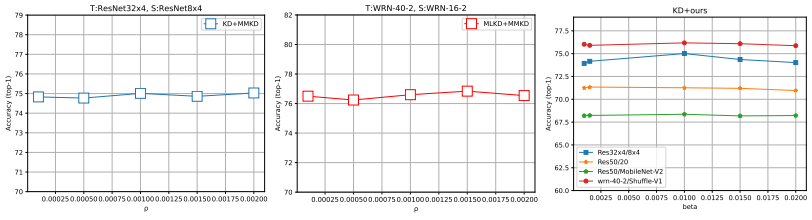

Figure 2: Sensitivity analysis results of the margin ρ and the parameter β.

DKD maintains stable performance across a wide range of parameter values. Appropriate parameter choices can further improve the performance of certain KD models, demonstrating the robustness of our method. **The implementation details, limitation discussion, proof, and more experiment results are shown in the Technical Supplement.**

5 Conclusion

In this paper, we propose a novel class-aware margin optimization method to address the inconsistency problem in DKD. Specifically, we introduce a margin between target and non-target labels, enabling the model to focus on learning target labels while leveraging correlations among non-target labels. Furthermore, our theoretical analysis highlights the necessity of both target and non-target classes for KD generalization, providing a valuable theoretical contribution to the DKD field. Extensive experiments demonstrate the effectiveness of our method across same and different network structures. In the future, we plan to explore adaptive margin setting strategies to further enhance the robustness and performance of our approach.

AI use statement

No generative AI tools were used in the preparation of this work.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or sensitive information. Therefore, we do not identify any ethical concerns related to this research.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed CAMO-DKD framework, optimization objectives, and experimental settings in the main paper and Technical Supplement. Theoretical analyses, proofs, implementation details, and additional experimental results are included in the Technical Supplement.

REFERENCES

Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning Ma, and Shanghang Zhang. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19846–19856, 2025.

Byungchul Chae and Seonyeong Heo. What to distill? fast knowledge distillation with adaptive sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2407–2416, 2025.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11933–11942, 2022.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5008–5017, 2021.

Yunuo Chen, Zezheng Lyu, Bing He, Ning Cao, Gang Chen, Guo Lu, and Wenjun Zhang. Knowledge distillation for learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4996–5006, 2025a.

Zhaowei Chen, Borui Zhao, Yuchen Ge, Yuhao Chen, Renjie Song, and Jiajun Liang. Asymmetric decision-making in online knowledge distillation: Unifying consensus and divergence. *arXiv preprint arXiv:2503.06685*, 2025b.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4178–4188, 2025.

Kei Sen Fong and Mehul Motani. Feat-kd: Learning concise representations for single and multi-target regression via tabnet knowledge distillation. In *Forty-second International Conference on Machine Learning*, 2025.

Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2018.

Yongbiao Gao, Xiangcheng Sun, Miaogen Ling, Chao Tan, Yi Zhai, and Guohua Lv. Decoupled imbalanced label distribution learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 5199–5207, 2025.

Rushin H Gindra, Giovanni Palla, Mathias Nguyen, Sophia J Wagner, Manuel Tran, Fabian J Theis, Dieter Saur, Lorin Crawford, and Tingying Peng. A large-scale benchmark of cross-modal learning for histology and gene expression in spatial transcriptomics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1182–1192, 2025.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11868–11877, 2023.

Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, volume 36, 2024.

Changyi He, Yifu Ding, Jinyang Guo, Ruihao Gong, Haotong Qin, and Xianglong Liu. Da-kd: Difficulty-aware knowledge distillation for efficient large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the International Conference on Computer Vision*, pp. 1921–1930, 2019.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Md Imtiaz Hossain, Sharmen Akhter, Choong Seon Hong, and Eui-Nam Huh. Single teacher, multiple perspectives: Teacher knowledge augmentation for enhanced knowledge distillation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2022.

Zizheng Huang, Haoxing Chen, Jiaqi Li, Huijia Zhu, Weiqiang Wang, Limin Wang, et al. Stochastic layer-wise shuffle for improving vision mamba training. In *Forty-second International Conference on Machine Learning*, 2025.

Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24276–24285, 2023.

Minsoo Kim, Sihwa Lee, Janghwan Lee, Sukjin Hong, Du-Seong Chang, Wonyong Sung, and Jungwook Choi. Token-scaled logit distillation for ternary weight generative language models. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2024.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4), 2009.

Xiaorong Li, Shipeng Wang, Jian Sun, and Zongben Xu. Variational data-free knowledge distillation for continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12618–12634, 2023a.

Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1504–1512, 2023b.

Zhihui Li, Pengfei Xu, Xiaojun Chang, Luyao Yang, Yuanyuan Zhang, Lina Yao, and Xiaojiang Chen. When object detection meets knowledge distillation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10555–10579, 2023c.

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

Jiaming Lv, Haoyuan Yang, and Peihua Li. Wasserstein distance rivals kullback-leibler divergence for knowledge distillation. 2024.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the International Conference on Computer Vision*, pp. 5007–5016, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *Proceedings of the International Conference on Learning Representations*, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15731–15740, 2024a.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15731–15740, 2024b.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *Proceedings of the International Conference on Learning Representations*, 2020.

Jing Wang and Xin Geng. Classification with label distribution learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2019.

Jing Wang and Xin Geng. Learn the highest label and rest label description degrees. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3097–3103, 2021.

Qiufeng Wang, Xu Yang, Haokun Chen, and Xin Geng. Vision transformers as probabilistic expansion from learngene. In *Forty-first International Conference on Machine Learning*, 2024.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4933–4942, 2019.

Shicai Wei, Chunbo Luo, and Yang Luo. Scaled decoupled distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15975–15983, 2024.

Zhu Wencheng, Zhou Xin, Zhu Pengfei, Wang Yu, and Hu Qinghua. Ckd: Contrastive knowledge distillation from a sample-wise perspective. *arXiv2404.14109*, 2024.

Liangyu Xiang, Junyu Gao, and Changsheng Xu. Evidential knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2814–2824, 2025.

Bokang Yang, Zhe Zhang, and Jie Ma. Co-progression knowledge distillation with knowledge prototype for industrial anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13008–13016, 2025.

Jing Yang, Brais Martinez, Adrian Bulat, Georgios Tzimiropoulos, et al. Knowledge distillation via softmax regression representation learning. In *Proceedings of the International Conference on Learning Representations*, 2021.

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 *Proceedings of the International Conference on Computer Vision*, pp. 17185–17194, 2023.

Kairong Yu, Chengting Yu, Tianqing Zhang, Xiaochen Zhao, Shu Yang, Hongwei Wang, Qiang Zhang, and Qi Xu. Temporal separation with entropy regularization for knowledge distillation in spiking neural networks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8806–8816, 2025.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3903–3911, 2020.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of the International Conference on Learning Representations*, 2017.

Rongzhi Zhang, Jiaming Shen, Tianqi Liu, Jialu Liu, Michael Bendersky, Marc Najork, and Chao Zhang. Do not blindly imitate the teacher: Using perturbed loss for knowledge distillation. *arXiv preprint arXiv:2305.05010*, 2023.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11953–11962, 2022.

Borui Zhao, Quan Cui, Renjie Song, and Jiajun Liang. Dot: A distillation-oriented trainer. In *Proceedings of the International Conference on Computer Vision*, 2023.

Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching. In *International Conference on Learning Representations*, 2024.

Yujie Zheng, Chong Wang, Chenchen Tao, Sunqi Lin, Jiangbo Qian, and Jiafei Wu. Restructuring the teacher and student in self-distillation. *IEEE Transactions on Image Processing*, 2024.

Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. In *Proceedings of the International Conference on Learning Representations*, 2021.

Yu Zhou, Dian Zheng, Qijie Mo, Renjie Lu, Kun-Yu Lin, and Wei-Shi Zheng. Decoupled distillation to erase: A general unlearning method for any class-centric tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20350–20359, 2025.

APPENDIX

1. SUPPLEMENTARY PROOF

Lemma 0.1. *Let $a \geq b$ and $c \geq d$, then $|a - b| \leq |a - c| + |b - d|$.*

Proof. If $c \leq b$ or $d \geq a$, the inequality is obvious. If $c \geq b$ and $a \geq d$, then $|a-b| = |a-c|+|c-d| \leq |a - c| + |b - d|$.

2. DISCUSSION

DKD Zhao et al. (2022) is the research work most closely related to ours. DKD decomposes knowledge distillation into target class knowledge and non-target class knowledge distillation. There are three key distinctions between our method and DKD. First, we adopt the L_1-norm as the loss function for decoupled knowledge distillation. Compared to KL-divergence, our proposed approach provides a more direct measure of the differences between probability values, making it more effective in addressing the inconsistency between KL-divergence and classification. Second, we introduce a class-aware margin optimization method to establish a margin between the target label and the non-target distribution, further mitigating the inconsistency issue. Third, we offer a theoretical analysis to elucidate the generalization capability of decoupled knowledge distillation. Thus, our method can be regarded as an enhancement of the decoupled knowledge distillation framework from both theoretical and algorithmic perspectives.

3. IMPLEMENTATION DETAILS

Following the same experimental settings as logit-standard KD Sun et al. (2024), we train the models using SGD Sutskever et al. (2013). The number of epochs is set to 240 for CIFAR-100 and 100 for the ImageNet dataset. The initial learning rate is configured as 0.01 for MobileNets and ShuffleNets, and 0.05 for other architectures. The margin parameter ρ is fixed at 0.001, while the balance parameter β is set to 0.01 for DKD and KD-based methods, and 0.001 for other models. We conduct experiments using both identical and different architecture pairs for teachers and students. For the CIFAR-100 dataset, top-1 accuracy (%) is used as the evaluation metric, whereas both top-1 and top-5 accuracy (%) are reported to validate the effectiveness on ImageNet dataset. All experiments are implemented on the PyTorch, utilizing eight NVIDIA RTX 3090 GPUs for CIFAR-100 and four NVIDIA A100 GPUs for ImageNet.

4. MORE EXPERIMENTAL RESULTS

We also conduct experiment on Tiny-ImageNet. Tiny-ImageNet consists of 200 classes and the image size is 61x64. The model is trained for 200 epochs with learning rates decayed by 0.1 at 60th, 120th, and 160th epoch. The initial learning rate is 0.05 for a 64 batch size. The results are shown in Table 1.

From the results, we can conclude that the integration of our proposed technique significantly enhances the performance of both KD and DKD. For KD+ours, top-1 accuracy improves by 1.36% (from 58.35% to 59.71%) for ResNet18/MobileNet-V2 and by 1.13% (from 62.26% to 63.39%) for ResNet18/ShuffleNet-V2. For DKD+ours, the improvements are even more pronounced, with gains of 1.39% (from 62.04% to 63.43%) and 0.96% (from 65.06% to 66.02%) for the two pairs, respectively. These consistent improvements demonstrate the effectiveness of our method in enhancing the distillation process. In addition, DKD+ours achieves the best performance across both teacher-student pairs. The experimental results highlight the effectiveness of our proposed method in improving the performance of existing KD techniques. When combined with advanced methods like DKD, our approach achieves the highest accuracy, demonstrating its potential to solve the inconsistency issue between KL-divergence and classification.

Table 2 presents a detailed analysis of the ablation experiment results for KL-divergence and L1-norm loss functions, which are integrated with the proposed method on top of baseline knowledge distillation frameworks (KD, DKD, CTKD). The experiment is conducted on four distinct teacher-

Table 1: Experiment results on the Tiny ImageNet dataset.

Teacher/Student	ResNet18/MobileNet-V2	ResNet18/ShuffleNet-V2
Method	top-1	top-1
KD	58.35	62.26
AT	57.18	62.45
CRD	61.18	63.97
DKD	62.04	65.06
KD+Ours	$59.71_{+1.36}$	$63.39_{+1.13}$
DKD+Ours	$\mathbf{63.43}_{+1.39}$	$\mathbf{66.02}_{+0.96}$

Table 2: Ablation performance of KL-divergence and L_1-norm.

Teacher/CIFAR-100 Student/CIFAR-100	ResNet32x4 ResNet8x4	ResNet56 ResNet20	Resnet32x4 ShuffleNet-V1	ResNet50 MobileNet-V2
KD Hinton (2015)+Ours (KL)	$73.42_{-1.59}$	$71.10_{-0.16}$	$74.01_{-0.35}$	$67.31_{-1.05}$
KD Hinton (2015)+Ours (L1-norm)	**75.01**	**71.26**	**74.36**	**68.36**
DKD Zhao et al. (2022)+Ours (KL)	$76.36_{-0.35}$	$71.88_{-0.47}$	$76.52_{-0.36}$	$70.38_{-0.27}$
DKD Zhao et al. (2022)+Ours (L1-norm)	**76.71**	**72.35**	**76.88**	**70.65**
CTKD Li et al. (2023)+Ours (KL)	$73.34_{-1.05}$	$70.22_{-0.30}$	$74.64_{-0.19}$	$67.39_{-0.42}$
CTKD Li et al. (2023)+Ours (L1-norm)	**74.39**	**70.52**	**74.83**	**67.81**

Table 3: Performance on the CIFAR-100 validation set for different architecture pairs of teachers and students. We incorporate our proposed CAMO-DKD into existing logit-based knowledge distillation methods.

Type	Teacher Student	ResNet32x4 ShuffleNet-V1	ResNet50 MobileNet-V2	ResNet32x4 ShuffleNet-V2	WRN-40-2 ShuffleNet-V1	VGG13 MobileNet-V2
Logit-based	DIST Huang et al. (2022)	76.34	68.66	77.35	75.86	68.33
	SDD-NKD Wei et al. (2024)	76.68	68.87	76.97	76.60	69.11
	SDD-NKD Wei et al. (2024)+Ours	$76.80_{+0.12}$	$70.30_{+1.43}$	$77.42_{+0.45}$	$76.95_{+0.35}$	$69.86_{+0.75}$

Table 4: Performance on the CIFAR-100 validation set for identical architecture pairs of teachers and students. We integrate our proposed CAMO-DKD into existing logit-based knowledge distillation methods.

Type	Teacher Student	ResNet32x4 ResNet8x4	ResNet56 ResNet20	ResNet110 ResNet32	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	VGG13 VGG8
Logit-based	DIST Huang et al. (2022)	76.31	71.75	72.85	75.60	74.73	73.89
	SDD-NKD Wei et al. (2024)	76.23	-	-	76.00	74.04	74.09
	SDD-NKD Wei et al. (2024)+Ours	$76.80_{+0.57}$	-	-	$76.86_{+0.86}$	$74.41_{+0.37}$	$74.29_{+0.20}$

student architecture pairs over the CIFAR-100 dataset. Across all baseline distillation methods and teacher-student pairs, the L1-norm variant achieves higher accuracy than the KL-divergence variant, with the latter showing measurable performance degradation in every scenario. For example, For KD+Ours, the L1-norm variant yields a maximum accuracy improvement of 1.59% (from 73.42% to 75.01%) on the ResNet32x4/ResNet8x4 pair, and a minimum improvement of 0.16% (from 71.10% to 71.26%) on the ResNet56/ResNet20 pair. The ablation experiment conclusively verifies that L1-norm is a superior loss function compared to KL-divergence when integrated with the proposed method in knowledge distillation.

Tables 3, 4, 5, and 6 show more experiment results for identical/different architecture teacher/student pairs and ablation study.

Table 5: Ablation performance on the CIFAR-100 in top-1 accuracy. The ✓means that the model includes the corresponding part.

Student	Target label	Non-target label	Margin	top-1
ResNet32x4 as the teacher				
ShuffleNet-V1 (DKD Zhao et al. (2022)+Ours)	✓	✓		$73.84_{-3.04}$
	✓		✓	$73.95_{-2.93}$
		✓	✓	$75.16_{-1.72}$
	✓	✓	✓	**76.88**
ShuffleNet-V2 (DKD Zhao et al. (2022)+Ours)	✓	✓		$76.86_{-0.47}$
	✓		✓	$76.89_{-0.44}$
		✓	✓	$77.12_{-0.21}$
	✓	✓	✓	**77.33**
ShuffleNet-V2 (TTM Zheng & Yang (2024)+Ours)	✓	✓		$75.35_{-0.58}$
	✓		✓	$75.49_{-0.44}$
		✓	✓	$75.04_{-0.89}$
	✓	✓	✓	**75.93**
WRN-40-2 as the teacher				
ShuffleNet-V1 (DKD Zhao et al. (2022)+Ours)	✓	✓		$76.73_{-0.29}$
	✓		✓	$76.70_{-0.32}$
		✓	✓	$76.90_{-0.12}$
	✓	✓	✓	**77.02**
WRN-40-1 (DKD Zhao et al. (2022)+Ours)	✓	✓		$74.64_{-0.38}$
	✓		✓	$74.89_{-0.13}$
		✓	✓	$74.66_{-0.36}$
	✓	✓	✓	**75.02**
Shuffle-V1 (EKD Xiang et al. (2025)+Ours)	✓	✓		$74.51_{-1.23}$
	✓		✓	$75.37_{-0.37}$
		✓	✓	$75.11_{-0.63}$
	✓	✓	✓	**75.74**

5. Parameter Sensitivity Analysis of ρ and β

We provide additionally results of parameter sensitivity analysis. Fig. 1 illustrates the top-1 accuracy for various margin values of ρ on the CIFAR-100 dataset, with four methods selected as the basis for sensitivity analysis of ρ. Similarly, the third column of Fig. 1 presents the sensitivity analysis of the parameter β.

From these curves, we observe the following: 1) Overall, CAMO-DKD exhibits stable performance across a wide range of margin values ρ and balance parameter β. 2) Appropriate choices of ρ and β can yield slight performance improvements for certain KD models. These results further highlight the robustness and effectiveness of the proposed CAMO-DKD method.

6. Limitation Discussion

One major limitation of CAMO-DKD is its reliance on manually tuned fixed hyperparameters ρ (margin) and β (loss balance weight), which require laborious grid search for different datasets and teacher-student pairs without an adaptive adjustment strategy, while the uniform margin ρ fails to handle imbalanced inter-class distinguishability in fine-grained tasks. Another drawback is the increased computational overhead during training: the composite loss combining target-class L_1 loss, non-target-class L_1 loss and margin regularization adds extra probability difference calculations compared with vanilla DKD, leading to slower convergence and higher memory cost, which hinders its deployment on resource-limited edge devices.

7. Training Time

Table 7 illustrates that adding a margin constraint to knowledge distillation methods (e.g., KD+margin, DKD+margin, MLKD+margin) results in only a modest increase in training time

Table 6: Performance on the ImageNet validation set, evaluated in terms of top-1 and top-5 accuracy.

Teacher/Student	ResNet34/ResNet18		ResNet50/MobilNet-V1	
Accuracy	top-1	top5	top-1	top-5
OFD Heo et al. (2019)	70.81	89.98	71.25	90.34
AT Zagoruyko & Komodakis (2017)	70.69	90.01	69.56	89.33
RKD Park et al. (2019)	71.34	90.37	91.32	90.62
CRD Tian et al. (2020)	71.17	90.13	71.37	90.41
SimKD Chen et al. (2022)	71.59	90.48	72.25	90.86
ReviewKD Chen et al. (2021)	71.61	90.51	72.56	91.00
CAT-KD Guo et al. (2023)	71.26	90.45	72.24	91.13
KD Hinton (2015)	70.66	89.88	68.58	88.98
KD+Ours	71.42$_{+0.76}$	90.04$_{+0.16}$	71.74$_{+3.16}$	90.26$_{+1.28}$
DKD Zhao et al. (2022)	71.70	90.41	72.05	91.05
DKD+Ours	<u>72.14</u>$_{+0.44}$	90.59$_{+0.18}$	73.02$_{+0.97}$	91.40$_{+0.35}$
MLKD Jin et al. (2023)	71.13	90.67	73.38	91.55
MLKD+Ours	71.40$_{+0.27}$	91.42$_{+0.75}$	<u>73.50</u>$_{+0.12}$	<u>91.58</u>$_{+0.03}$
KD+LS_KD Sun et al. (2024)	71.42	90.29	72.18	90.80
KD+LS_KD Sun et al. (2024) + Ours	71.70$_{+0.28}$	90.20$_{-0.09}$	72.91$_{+0.73}$	91.00$_{+0.20}$
MLKD+LS_KD Sun et al. (2024)	71.46	90.47	72.53	91.11
MLKD+LS_KD Sun et al. (2024)+Ours	71.81$_{+0.35}$	90.62$_{+0.15}$	**73.51**$_{+0.98}$	**91.58**$_{+0.47}$
SDD-DKDWei et al. (2024)	71.87	91.02	72.94	90.97
SDD-DKDWei et al. (2024)+Ours	72.10$_{+0.23}$	<u>91.55</u>$_{+0.53}$	73.22$_{+0.28}$	91.36$_{+0.39}$
SDD-NKDWei et al. (2024)	72.01	91.14	72.87	90.84
SDD-NKDWei et al. (2024)+Ours	**72.39**$_{+0.38}$	**91.63**$_{+0.49}$	73.11$_{+0.24}$	91.21$_{+0.37}$
EKD Xiang et al. (2025)	71.60	90.47	72.09	90.93
EKD Xiang et al. (2025)+Ours	71.81$_{+0.21}$	90.64$_{+0.17}$	72.33$_{+0.24}$	91.07$_{+0.14}$

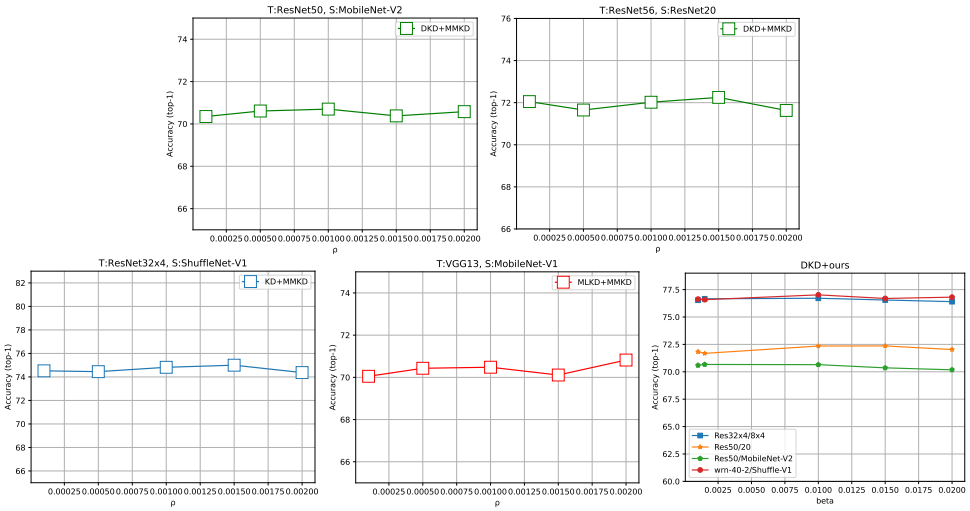

Figure 1: More sensitivity analysis results of the margin ρ and the parameter β.

compared to their margin-free counterparts (KD, DKD, MLKD), but the overhead remains at a manageable level. For instance, in KD and KD+margin comparison, the time overhead ranges from 0.180h (ResNet32x4/ResNet8x4: 0.573h to 0.753h) to 0.204h (ResNet56/ResNet20: 0.343h to 0.547h). The relative overhead ratio is approximately 30–59%, which is acceptable given the accuracy gains demonstrated in previous experiments. The time increase is between 0.176h (ResNet32x4/ResNet8x4: 0.586h to 0.762h) and 0.203h (ResNet56/ResNet20: 0.358h to 0.561h)

Table 7: The comparison of training time (hours).

Teacher/student	KD	KD+margin	DKD	DKD+margin	MLKD	MLKD+margin
ResNet32X4/ResNet8x4	0.573	0.753	0.586	0.762	1.179	1.425
ResNet56/ResNet20	0.343	0.547	0.358	0.561	0.690	0.853
ResNet32x4/ShuffleV1	0.791	0.915	0.824	1.055	1.498	1.726

in DKD and DKD+margin pair. The overhead is close to that of KD, indicating that the margin mechanism's computational cost is independent of the baseline distillation framework.

The training time of both original and margin-augmented methods is closely related to the complexity of the teacher-student pair. The ResNet56/ResNet20 pair has the shortest training time across all methods (0.343h to 0.853h), as both teacher and student networks are shallow and have low computational complexity. The ResNet32x4/ShuffleNet-V1 pair requires longer training time (0.791h to 1.726h) compared to homogeneous ResNet pairs.

In conclude, the margin mechanism achieves a favorable accuracy-time trade-off, especially for complex distillation methods like MLKD, making it a practical enhancement strategy for real-world deployment scenarios. For instance, in the ResNet32X4/ResNet8x4 pair, KD+margin increases training time from 0.573 to 0.753 hours (a 31% increase), while DKD+margin rises from 0.586 to 0.762 hours (30%). Similarly, for MLKD+margin, the increase from 1.179 to 1.425 hours (21%) remains relatively small. This suggests that while margin-based methods introduce additional computational overhead, the impact on training time is not substantial, making them a viable option for potentially improved performance with only a marginal time cost.