EA-KD: Entropy-based Adaptive Knowledge Distillation

Chi-Ping Su[1,*], Ching-Hsun Tseng[2,*], Bin Pu[3,†], Lei Zhao[4], Jiewen Yang[3], Zhuangzhuang Chen[3], Shin-Jye Lee[1,†]

[1]National Yang Ming Chiao Tung University, [2]The University of Manchester,
[3]Hong Kong University of Science and Technology, [4]Hunan University

Abstract

*Knowledge distillation (KD) enables a smaller "student" model to mimic a larger "teacher" model by transferring knowledge from the teacher's output or features. However, most KD methods treat all samples uniformly, overlooking the varying learning value of each sample and thereby limiting effectiveness. In this paper, we propose **E**ntropy-based **A**daptive **K**nowledge **D**istillation (EA-KD), a simple yet effective plug-and-play KD method that prioritizes learning from valuable samples. EA-KD quantifies each sample's learning value by strategically combining the entropy of the teacher and student output, then dynamically reweights the distillation loss to place greater emphasis on high-entropy samples. Extensive experiments across diverse KD frameworks and tasks–including image classification, object detection, and large language model (LLM) distillation–demonstrate that EA-KD consistently enhances performance, achieving state-of-the-art results with negligible computational cost. Our code is available at* https://github.com/cpsu00/EA-KD.

1. Introduction

The growing size of state-of-the-art (SOTA) deep learning models poses challenges for deployment in resource-constrained settings. Knowledge Distillation (KD) [11] offers a solution by training a smaller "student" model to mimic a larger "teacher" model, using both ground-truth and the teacher's "knowledge" (*e.g.* logits or feature representations) to achieve similar performance in a compact form. KD has been widely applied across domains, including computer vision (CV) [1, 31, 40], large language models (LLMs) [9, 13, 15, 35], and medical imaging [8, 21–23].

Advanced KD methods have explored diverse knowledge forms and structural modifications to refine knowledge transfer [2, 11, 14, 26, 30, 33, 47]. However, most methods distill uniformly across all samples, operating under the assumption that each sample has equal importance and overlooking their varying learning value. Recent advancements, such as Instance-T [20], have shown that assigning unique temperatures to each sample outperforms the uniform temperature method Global-T. This highlighted the benefits of adapting to each sample's distinct characteristics. Building on this, we hypothesize that emphasizing samples rich in *valuable knowledge*[1] can further optimize the distillation process. This mirrors how human students learn better when key points are highlighted by teachers. Such a strategy enables the student model to focus on more informative samples, leading to improved performance. However, as we will show, the uniform distillation scheme in most KD methods often overlooks these critical samples, thereby limiting the efficiency of knowledge transfer.

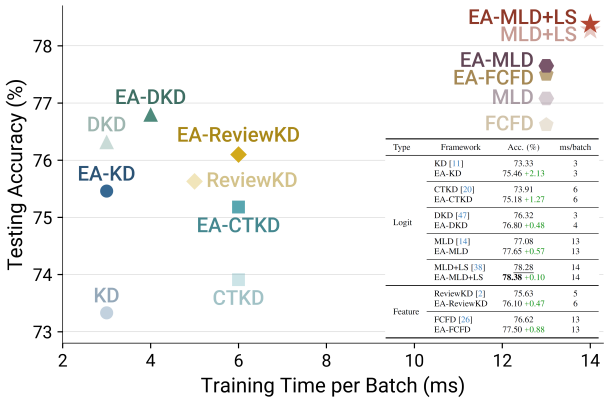
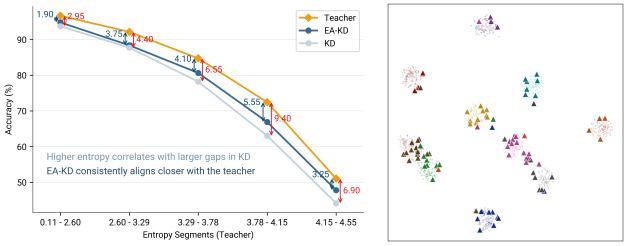


Figure 1. **Training Time per Batch (ms) *vs*. Accuracy (%) of EA-Methods and Baselines on CIFAR-100.**

The question then arises: ***How can we identify the most valuable samples for learning?*** Entropy, the core concept in information theory that quantifies the uncertainty or information of a random variable [36], may be well-suited for this role. Prior methods have incorporated entropy in KD to adjust weighting [17] or refine logit predictions [49]; however, these approaches are typically restricted to specific scenarios (*e.g.*, multi-teacher or logit-based KD) and primarily rely on teacher entropy, leaving the broader potential of entropy in KD underexplored. Therefore, we propose leveraging entropy to quantify the learning value of each

*Equal contribution.
†Bin Pu and Shin-Jye Lee are corresponding authors (eebinpu@ust.hk, camhero@gmail.com).

[1]Throughout this paper, we consider samples with high entropy as containing valuable knowledge.



(a) Accuracy *vs*. Entropy Segments. (b) t-SNE of Teacher.

Figure 2. **High-Entropy Samples in KD.** (a) Higher teacher entropy correlates with larger accuracy gaps in KD, while EA-KD maintains closer alignment. (b) The top 10% high entropy samples (denoted by triangles) cluster near decision boundaries, representing critical knowledge essential for classification.

Table 1. **Entropy-based Reweighting on KD.** Reweighting with entropy ($H^{\mathcal{T}}$ and $H^{\mathcal{S}}$) improves student's performance, while inverted reweighting ($H_{\text{ub}} - w$) reduces accuracy.

Loss Function	Reweighting Factor w	
	$H^{\mathcal{T}}$	$H^{\mathcal{S}}$
L_{KD} [11]	73.33	
wL_{KD}	**75.14** +1.81	**74.76** +1.43
$(H_{\text{ub}} - w)L_{\text{KD}}$	72.73 −0.60	68.90 −4.43

sample in KD, as high-entropy outputs shall correspond to greater information contents that are crucial for learning. Our preliminary analysis reveals that higher entropy samples[2] (i) correlate with larger teacher-student accuracy gaps (Fig. 2a) and (ii) often lie near class boundaries in t-SNE visualizations [41] (Fig. 2b), suggesting these informative samples not only offer valuable learning opportunities but are also pivotal in defining decision boundaries. Thus, entropy can serve as a reliable metric for identifying the most valuable samples in KD.

Leveraging this insight, adapting the focus of KD to valuable samples should fill the need for enhanced distillation. To validate this, we conducted a preliminary experiment that compared the performance of reweighting with teacher ($H^{\mathcal{T}}$) and student ($H^{\mathcal{S}}$) entropy, along with their linear-inverted variants ($H_{\text{ub}} - w$), where high-entropy samples received lower weight. Here, H_{ub} denotes the upper bound of entropy (*i.e.*, $\log 100$ for CIFAR-100). As shown in Tab. 1, reweighting with either $H^{\mathcal{T}}$ or $H^{\mathcal{S}}$ significantly improved KD accuracy, while inverted reweighting led to decreased performance. This supports our hypothesis that focusing on samples with valuable knowledge enhances student learning, with $H^{\mathcal{T}}$ proving more effective due to the teacher's more reliable assessment of sample value.

Exploring deeper, we observed that $H^{\mathcal{S}}$ exhibited increasing variability across training epochs for the top 10% $H^{\mathcal{T}}$ samples (Fig. 3). This suggests that while the teacher finds these samples valuable, potentially due to the inherent differences in architecture or capacity, the student's assessment fluctuates during training. Some samples remain consistently challenging (high $H^{\mathcal{S}}$), while others become progressively simpler (low $H^{\mathcal{S}}$) over time. This misalignment reveals the limitation of reweighting solely with $H^{\mathcal{T}}$, as it remains constant throughout the epochs and thus fails to capture the student's evolving learning process.

[2]Results for high teacher entropy samples are shown here; high student entropy plots are in the Appendix B.1.

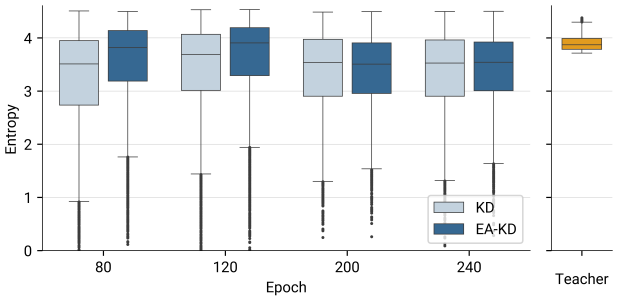


Figure 3. **Box Plot of $H^{\mathcal{S}}$ *vs*. $H^{\mathcal{T}}$ for Top 10% High $H^{\mathcal{T}}$ Samples.** $H^{\mathcal{S}}$ shows notable variation, reflecting the student's learning dynamics and differing perspective from the teacher. EA-KD enhances mimicry with more stable and aligned entropy.

In light of the above analysis, we introduce Entropy-based Adaptive Knowledge Distillation (EA-KD), a simple yet effective plug-and-play KD method that enhances distillation by reweighting the loss toward high-value samples. Leveraging both teacher and student entropies, EA-KD emphasizes samples that encapsulate critical knowledge as identified by the teacher, while dynamically adapting to the student's evolving needs. This method directly addresses a fundamental limitation in standard KD frameworks, which often overlook the unique learning value of each sample and tend to bias towards simpler knowledge. Furthermore, EA-KD can be seamlessly integrated into most KD frameworks, enhancing their performance with negligible computational cost (Fig. 1). Extensive experiments on image classification, object detection, and LLM distillation demonstrate its efficacy and versatility across diverse KD frameworks. Our main contributions are summarized as follows:

- We reveal that high-entropy samples carry critical knowledge in KD and propose an entropy-based reweighting factor that integrates both teacher and student entropy to provide a dynamic and tailored learning focus.
- We introduce EA-KD, a plug-and-play KD method that adaptively reweights the distillation loss to prioritize valuable samples, enabling more effective and efficient knowledge transfer.
- We demonstrate that EA-KD consistently improves performance across logit- and feature-based KD methods, achieving SOTA results on both CV and LLM tasks with minimal computational overhead.

2. Related Work

Logit and Feature Distillation. Logit distillation [11, 14, 47] aligns the softened output logits of the teacher and student, valued for its simplicity and broad applicability. On the other hand, feature distillation minimizes divergence in intermediate feature representations, offering enhanced learning but often with higher computational costs [2, 26, 33]. Both pathways have achieved SOTA performance across tasks and domains. However, most of them typically adopt a static distillation scheme, such as treating all samples uniformly. Adaptive distillation addresses this limitation by introducing more dynamic knowledge transfer processes [17, 20, 28, 38, 46, 48, 49]

Adaptive Distillation. These methods improve knowledge transfer by dynamically adjusting knowledge at different levels. For sample-level, RW-KD [28] employs meta-learning to optimize weights for each sample, which introduces high computational overhead. PAD [46] showed that traditional hard-mining weighting is unsuitable for KD and instead prioritizes samples with low uncertainty and small teacher-student gaps. This shares similarities with EA-KD, as we will show that high-entropy samples often exhibit lower KLD. However, PAD relies on an auxiliary estimation branch, whereas EA-KD offers a more interpretable and efficient approach by directly utilizing entropy. For logit-level, CTKD [20] and LS [38] dynamically adjust the temperature parameter T to refine knowledge transfer, but remain limited to logit-based KD. Instead, EA-KD's sample-wise reweighting ensures broader compatibility across KD frameworks. Importantly, EA-KD and these methods serve distinct yet complementary roles, combining them could further improve performance, as we will show.

Entropy in KD. Cheng et al. introduced an entropy-based metric to quantify knowledge retention in KD [3]. Inspired by this, we utilize entropy to identify valuable samples in KD. AKD [17] assign higher weights to low-entropy teacher predictions in multi-teacher settings. However, such weighting can degrade performance in the more common single-teacher settings (Tab. 1). DynamicKD [49] refines logit-level knowledge through entropy correction similar to CTKD and LS, and is also constrained to logit-based KD. TTM [48] removes the student temperature, revealing an inherent Rényi entropy regularization, while WTTM further emphasizes the uncertain samples. In contrast, EA-KD actively leverages both teacher and student entropy for dynamic weighting, yielding stronger performance[3].

3. Methodology

3.1. Preliminaries

Information Theory. Entropy quantifies the uncertainty or information content of a random variable [36]. For a given sample x_n, the entropy H_n is computed as follows:

$$H_n = -\sum_{i=1}^{C} \sigma(z_{n,i}) \log(\sigma(z_{n,i}))\,, \tag{1}$$

where $\sigma(\cdot)$ denotes the softmax function, $z_{n,i}$ is the logit for class i of sample x_n, and C is the number of classes.

Knowledge Distillation. The goal of vanilla KD is to transfer the knowledge encapsulated in the teacher's softened probability outputs to the student [11]. In classification tasks, the probabilities p are softened using the temperature-scaled softmax function:

$$p_i(T) = \sigma(z, T)_i = \frac{\exp(\frac{z_i}{T})}{\sum_{k=1}^{C} \exp(\frac{z_k}{T})}\,, \tag{2}$$

where $p_i(T)$ denotes the softened probability for class i, and $\sigma(z_i, T)$ is the temperature-scaled softmax function. The temperature T controls the smoothness of the distribution, revealing the subtle inter-class relationships.

The core of KD is to minimize the Kullback-Leibler divergence (KLD) between the teacher's and student's softened probabilities, the KD loss is defined as:

$$\begin{aligned} L_{\text{KD}} &= \text{KLD}(p^{\mathcal{T}}(T) \| p^{\mathcal{S}}(T)) \cdot T^2 \\ &= \sum_{i=1}^{C} p_i^{\mathcal{T}}(T) \log\left(\frac{p_i^{\mathcal{T}}(T)}{p_i^{\mathcal{S}}(T)}\right) \cdot T^2\,, \end{aligned} \tag{3}$$

where $p^{\mathcal{T}}$ and $p^{\mathcal{S}}$ denote the teacher's and student's softened outputs, respectively. For simplicity of theoretical analysis, we set $T = 1$ in this section. L_{KD} then simplifies to:

$$L_{\text{KD}} = \sum_{i=1}^{C} p_i^{\mathcal{T}} \log\left(\frac{p_i^{\mathcal{T}}}{p_i^{\mathcal{S}}}\right). \tag{4}$$

3.2. EA-KD

Limitations in Standard KDs. Most logit- and feature-based KD methods [2, 11, 14, 26, 33, 47] treat all samples uniformly, overlooking their unique learning value. This oversight can cause the model to over-prioritize simpler samples at the expense of more valuable, high-entropy ones. Taking KLD—the main loss function for logit-based methods—as an example, consider a student initialized with a uniform distribution $p^{\mathcal{S}}$ where $p_i^{\mathcal{S}} = \frac{1}{C}\,\forall\, i$. For a low-entropy sample with a teacher output $p_{\text{low}}^{\mathcal{T}}$ where $p_{\text{low},j}^{\mathcal{T}} \approx 1$ and $p_{\text{low},i}^{\mathcal{T}} \approx 0\,(i \neq j)$, the KLD becomes:

$$\begin{aligned} \text{KLD}(p_{\text{low}}^{\mathcal{T}} \| p^{\mathcal{S}}) &\approx p_{\text{low},j}^{\mathcal{T}} \cdot \log\left(\frac{p_{\text{low},j}^{\mathcal{T}}}{p_j^{\mathcal{S}}}\right) \\ &= \log\left(\frac{1}{p_j^{\mathcal{S}}}\right) \\ &= \log(C). \end{aligned} \tag{5}$$

[3] See Appendix A.1 for a more detailed comparison.

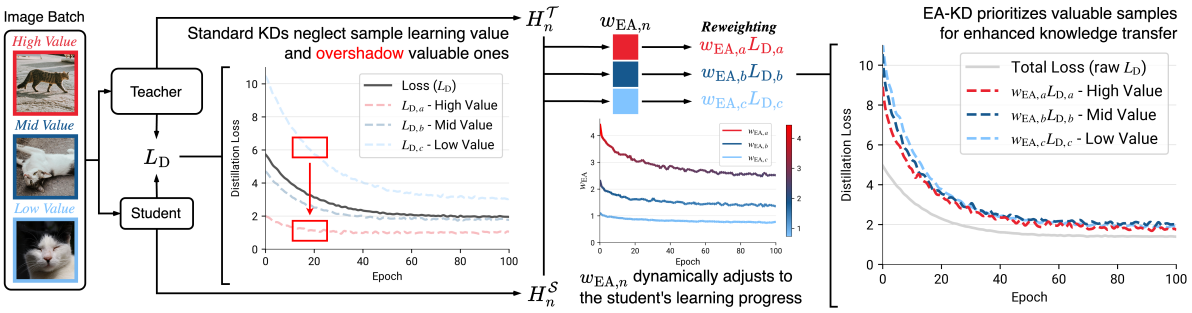


Figure 4. **Illustration of the Uniform Distillation Scheme in Standard KD (Left) and Entropy-based Reweighting in EA-KD (Right).** Standard KD not only overlook the varying learning value of individual samples but also bias toward low-entropy (easier) ones. In contrast, EA-KD effectively guides learning toward valuable samples based on both teacher and student assessments.

For a high-entropy distribution $p^{\mathcal{T}}_{\text{high}} \approx \frac{1}{C}\,\forall\, i$, the KLD is:

$$\begin{aligned}\text{KLD}(p^{\mathcal{T}}_{\text{high}} \| p^{\mathcal{S}}) &\approx \sum_{i=1}^{C} \frac{1}{C} \cdot \log\left(\frac{1/C}{1/C}\right) \\ &= 0.\end{aligned} \tag{6}$$

Thus, we obtain:

$$\text{KLD}(p^{\mathcal{T}}_{\text{low}} \| p^{\mathcal{S}}) > \text{KLD}(p^{\mathcal{T}}_{\text{high}} \| p^{\mathcal{S}}). \tag{7}$$

This inequality highlights KLD's inherent bias toward low-entropy samples, which dominate the loss over the valuable, high-entropy ones. Similarly, the MSE loss in feature-based KD also biases learning toward high-magnitude activations, which typically correspond to low-entropy samples. Such imbalance shifts the learning focus toward simpler samples, hindering the transfer of knowledge crucial for learning, especially during early training stages (refer to the overshadowing effect in the left part of Fig. 4).

Entropy-based Reweighting. As discussed in Sec. 1, entropy can serve as a measure of the sample learning value in KD. In practice, we soften the entropy with an alternative temperature T' to better reflect sample value:

$$H_n = -\sum_{i=1}^{C} p_{n,i}(T') \log(p_{n,i}(T'))\,, \tag{8}$$

where $p_{n,i}(T')$ is the temperature-scaled probability of class i for sample x_n.

To dynamically emphasize valuable samples, EA-KD's reweighting factor w_{EA} is formulated using two components: a base term w_{base} and an interaction term w_{interact}. The base term captures the inherent value of each sample based on the teacher's output entropy, defined as:

$$w_{\text{base},n} = H^{\mathcal{T}}_n\,, \quad H^{\mathcal{T}}_n \in [0, H_{\text{ub}}]\,, \tag{9}$$

where $H^{\mathcal{T}}_n$ denotes the entropy of the teacher's prediction for sample x_n, and $H_{\text{ub}} = \log(C)$ is the upper bound of entropy for C classes. The interaction term w_{interact}, on the other hand, captures the interplay between the teacher and student perspectives by taking the normalized product of their entropies:

$$w_{\text{interact},n} = \frac{H^{\mathcal{T}}_n \cdot H^{\mathcal{S}}_n}{H_{\text{ub}}}, \quad w_{\text{interact},n} \in [0, H_{\text{ub}}]\,. \tag{10}$$

where $H^{\mathcal{S}}_n$ is the student's entropy. Finally, the EA-KD reweighting factor $w_{\text{EA},n}$ is defined as the average of the base and interaction terms:

$$w_{\text{EA},n} = \frac{w_{\text{base},n} + w_{\text{interact},n}}{2}, \quad w_{\text{EA},n} \in [0, H_{\text{ub}}]\,. \tag{11}$$

By integrating both the teacher's evaluation and the student's evolving understanding, w_{EA} effectively prioritizes valuable samples while dynamically adjusting the focus throughout training.

Reformulation. To better illustrate the influence of the student's perspective in $w_{\text{EA},n}$, we reformulate Eq. (11) as:

$$\begin{aligned} w_{\text{EA},n} &= \frac{H^{\mathcal{T}}_n + \frac{H^{\mathcal{T}}_n \cdot H^{\mathcal{S}}_n}{H_{\text{ub}}}}{2} \\ &= \frac{1}{2} H^{\mathcal{T}}_n \left(1 + \frac{H^{\mathcal{S}}_n}{H_{\text{ub}}}\right). \end{aligned} \tag{12}$$

This reformulation expresses $w_{\text{EA},n}$ as the product of $H^{\mathcal{T}}_n$ and a scaling factor that depends on the $H^{\mathcal{S}}_n$. Thus, $w_{\text{EA},n}$ can be regarded as the teacher's assessment of sample value, adaptively adjusted based on the student's perspective. Depending on the entropy values, w_{EA} behaves as follows:

$$w_{\text{EA},n} = \begin{cases} H_{\text{ub}} & \text{if } H^{\mathcal{T}}_n \to H_{\text{ub}} \wedge H^{\mathcal{S}}_n \to H_{\text{ub}} \quad \text{(13a)} \\ \dfrac{H_{\text{ub}}}{2} & \text{if } H^{\mathcal{T}}_n \to H_{\text{ub}} \wedge H^{\mathcal{S}}_n \to 0 \quad \text{(13b)} \\ 0 & \text{if } H^{\mathcal{T}}_n \to 0 \quad \forall H^{\mathcal{S}}_n \quad \text{(13c)} \\ w_{\text{EA},n} & \text{otherwise.} \end{cases}$$

When the teacher considers a sample highly valuable for learning ($H^{\mathcal{T}}_n \to H_{\text{ub}}$), the scaling factor adjusts $w_{\text{EA},n}$

Table 2. **Results on CIFAR-100.** Accuracy (%) of EA-methods *vs*. baselines across teacher-student pairs, with relative improvements highlighted. Avg. Δ shows average improvement across methods and model pairs. Best results are **bolded**, and second-best are underlined.

Type		Teacher / Student	ResNet32×4 / ResNet8×4	WRN-28-4 / WRN-16-2	WRN-40-2 / WRN-40-1	VGG13 / VGG8	VGG13 / MN-V2	ResNet50 / MN-V2	ResNet32×4 / SN-V2	Avg. Δ
			79.42 / 72.50	78.60 / 73.26	75.61 / 71.98	74.64 / 70.36	74.64 / 64.60	79.34 / 64.60	79.42 / 71.82	
Logit	KD [11]		73.33	75.04	73.54	72.98	67.37	67.35	74.45	
	EA-KD		75.46 +2.13	75.79 +0.75	74.38 +0.84	74.08 +1.10	69.17 +1.80	69.67 +2.32	75.91 +1.46	+1.48
	CTKD [20]		73.91	75.29	73.93	73.52	68.46	68.47	75.31	
	EA-CTKD		75.18 +1.27	75.72 +0.43	74.03 +0.10	73.79 +0.27	69.19 +0.73	69.38 +0.91	76.02 +0.71	+0.63
	DKD [47]		76.32	76.45	74.81	74.68	69.71	70.35	77.07	
	EA-DKD		76.80 +0.48	76.74 +0.29	74.98 +0.17	75.07 +0.39	70.39 +0.68	70.98 +0.63	77.72 +0.65	+0.47
	MLD [14]		77.08	76.83	75.35	75.18	70.57	71.04	78.44	
	EA-MLD		77.65 +0.57	77.47 +0.64	75.77 +0.42	75.28 +0.10	70.72 +0.15	71.43 +0.39	78.85 +0.41	+0.38
	MLD+LS [38]		78.28	77.20	75.56	75.22	70.94	71.19	78.76	
	EA-MLD+LS		**78.38** +0.10	**77.60** +0.39	**75.78** +0.22	**75.38** +0.16	70.67 −0.27	71.36 +0.17	**79.13** +0.37	+0.16
Feature	ReviewKD [2]		75.63	76.39	74.45	74.45	70.37	69.89	77.78	
	EA-ReviewKD		76.10 +0.47	76.95 +0.56	75.43 +0.98	74.56 +0.11	70.55 +0.18	69.80 −0.09	78.22 +0.44	+0.38
	FCFD [26]		76.62	77.00	75.46	75.22	70.65	71.00	78.18	
	EA-FCFD		77.50 +0.88	77.15 +0.15	75.30 −0.16	75.36 +0.14	**71.02** +0.37	**71.97** +0.97	78.75 +0.56	+0.42
	Avg. Δ		+0.84	+0.46	+0.37	+0.33	+0.52	+0.76	+0.66	+0.56

based on the student's view. If the student aligns (Eq. (13a)), the weight is maximized to emphasize this sample. Conversely, if the student is confident (Eq. (13b)), the weight reduces to half for moderate focus. However, when the teacher considers the sample simple (Eq. (13c)), the weight remains low regardless of H_n^S, as the teacher considers it lacks valuable knowledge.

Loss Integration. As shown in Fig. 4, the flexible nature of w_{EA} allows it to be plug-and-play into most distillation frameworks by reweighting the contribution of each sample based on its learning value. For instance, when integrated with vanilla KD, the loss is defined as:

$$L_{\text{EA-KD}} = \sum_{n=1}^{N} w_{\text{EA},n} \cdot L_{\text{KD},n}, \tag{14}$$

where $L_{\text{KD},n}$ is the distillation loss for sample x_n. As a result, EA-KD enhances standard KD methods by facilitating a more nuanced and adaptive transfer of knowledge, ensuring that informative samples receive increased focus throughout training.

4. Experiments

4.1. Experimental Setup

Datasets. We used CIFAR-100 [16] (50k training, 10k validation images; 100 classes), Tiny-ImageNet [18] (100k training, 5k validation images; 200 classes), and ImageNet [6] (1.28M training, 50k validation images; 1,000 classes) to evaluate our method for image classification. For object detection, we adopted MS-COCO [25] (118k training, 5k validation images; 80 classes). Furthermore, for LLM distillation, we employed five instruction-following datasets following [9]: Dolly [5], SInst [42], Vicuna [4], S-NI [43], and UnNI [12].

KD Frameworks. We evaluated EA-KD across representative SOTA logit-based (KD [11], CTKD [20], DKD [47], MLD [14], MLD+LS [38]) and feature-based methods (ReviewKD [2], FCFD [26]), reweighting their distillation loss while preserving each framework's original structure. For fair comparisons, no hyperparameter was tuned for the EA-variants, except for the KD weight of MLD+LS [38] was reduced from 9.0 to 4.0 to avoid over-penalizing.

Implementation Details. EA-KD and other variants were evaluated on various CNNs (VGG [37], ResNet [10], WideResNet [44], MobileNet [34], ShuffleNet [29]), transformer teachers (ViT [7], DeiT [40], Swin [27]), and GPT2 [32] for LLM distillation. We used the training settings of [39] for vanilla KD and the respective settings of each framework for classification, [2] for detection, and [45] for LLMs. We set $T' = 3$ based on our ablation study, except $T' = 2$ for ImageNet and transformer teachers. All results averaged over five runs, and analyses were conducted using ResNet32×4-ResNet8×4 pair on CIFAR-100.

4.2. Results

CIFAR-100. Tab. 2 presents results of various EA-methods and their baselines on CIFAR-100. EA-KD consistently improves both logit- and feature-based KD frameworks across most teacher-student pairs, with an average gain of 0.56%. The logit-based EA-MLD+LS achieved SOTA results in most pairings, while EA-FCFD excelled in two heteroge-

Table 3. **Results on Tiny-ImageNet.** Accuracy (%) of the ResNet32×4 teacher and ResNet8×4 student.

Teacher	Student	KD	EA-KD	MLD	EA-MLD	MLD+LS	EA-MLD+LS	FCFD	EA-FCFD
64.41	55.25	56.00	59.39 +3.39	61.91	**62.65** +0.74	61.36	62.41 +1.05	60.12	60.51 +0.39

Table 4. **Results on ImageNet.** Accuracy (%) of the ResNet34 teacher and ResNet18 student, averaged over three runs.

Teacher	Student	KD [11]	EA-KD	KD+LS [38]	DKD [47]	EA-DKD	DKD+LS [38]	EA-DKD+LS	PAD [46]
73.31	69.75	71.03	**71.79** +0.76	71.42	71.70	71.96 +0.26	71.88	**71.99** +0.11	71.71

Table 5. **Results on Tiny-ImageNet with Transformers.** Accuracy (%) of transformer-based teachers and a ResNet8×4 student, averaged over three runs.

Teacher		Student	KD	EA-KD
ViT-B	71.12	ResNet8×4	54.70	**57.53** +2.83
DeiT-B	85.55	55.25	56.30	**58.36** +2.06
Swin-B	86.30		56.15	**59.58** +3.43

Table 6. **Results on MS-COCO.** AP (overall), AP_{50}, and AP_{75} are reported using Faster R-CNN with FPN.

	R-101 & R-18			R-50 & MV2		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Teacher	42.04	62.48	45.88	40.22	61.02	43.81
Student	33.26	53.61	35.26	29.47	48.87	30.90
KD	33.97	54.66	36.62	30.13	50.28	31.35
EA-KD	**34.78**	**56.14**	**37.19**	**31.81**	**53.18**	**33.18**
Δ	+0.81	+1.48	+0.57	+1.68	+2.90	+1.83

neous pairs. Additionally, EA-MLD and EA-FCFD secured most second-best results, outperforming the previous SOTA, MLD+LS [38]. These findings highlight EA-KD's broad applicability and effectiveness in enhancing knowledge transfer across diverse types of KD frameworks.

Tiny-ImageNet and ImageNet. Tab. 3 and Tab. 4 show that EA-methods consistently outperform their baselines on both Tiny-Imagenet and ImageNet. On Tiny-ImageNet, EA-KD improves KD by 3.39% and achieves performance comparable to the novel FCFD [26]. On ImageNet, EA-KD surpasses other adaptive KDs, including logit-level adaptive LS [38] and sample-level adaptive approach PAD [46]. These results highlight EA-methods' scalability on larger, more diverse datasets. Furthermore, Tab. 5 underscores EA-KD's advantage in distilling valuable knowledge from vision transformer-based teachers to CNN students. Unlike standard KD, which plateaued when using DeiT-B as the teacher, EA-KD consistently improved student performance, mitigating the capacity gap and utilizing the guidance of stronger teachers more effectively.

MS-COCO. Extending to object detection, Tab. 6 presents the performance of EA-KD and KD on the MS-COCO dataset. EA-KD consistently outperforms KD across AP metrics for both model pairings, with particularly notable gains in AP_{50}. This highlights EA-KD's ability to effectively guide the student with valuable knowledge in a more complex visual task, thereby enhancing its performance.

Table 7. **LLM Distillation Results.** Rouge-L [24] scores averaged over five seeds on each dataset are reported. SFT denotes a student supervised fine-tuned on the dataset.

Method	Dataset					Avg.
	Dolly	SInst	Vicuna	S-NI	UnNI	
Teacher GPT2-XL	27.19	14.64	16.30	27.55	31.42	23.42
SFT GPT2-S	22.94	10.11	15.17	16.21	18.68	16.62
KD	24.54	10.43	15.66	17.24	20.28	17.63
RKLD	24.38	**10.73**	15.71	17.31	20.96	17.82
JSD	23.86	10.20	15.50	16.20	19.17	16.98
EA-KD	**24.95**	10.59	**16.41**	**18.27**	**21.46**	**18.34**
Δ to KD	+0.41	+0.16	+0.75	+1.03	+1.18	+0.71

LLM Distillation. To explore the full potential of EA-KD, we extended our experiments to LLM distillation, applying the reweighting at the sequence level of the model outputs. As shown in Tab. 7, EA-KD consistently outperforms standard KD across datasets. Moreover, EA-KD also surpasses two recently emerging methods in this field, Reverse KLD (RKLD) and Jensen-Shannon divergence (JSD), which aim to address KLD's limitation of forcing the student to cover all modes of the complex LLM teacher distribution [9]. In this scenario, EA-KD's dynamic prioritization of the most critical knowledge performed effectively in such complexity. This highlights EA-KD's versatility beyond visual tasks, showcasing its ability to target valuable knowledge across diverse domains.

4.3. Empirical Analysis

In this section, we conduct ablation studies to evaluate two key components of EA-KD: T' and the reweighting factor. We then compare EA-KD and vanilla KD from multiple perspectives, showing how EA-KD improves teacher-student alignment and class separability. Additionally, we discuss the synergy between EA and DKD, demonstrating enhanced robustness in both the loss surface and hyperparameter stability. Finally, we highlight EA-methods computational efficiency across diverse KD frameworks.

Table 8. **Impact of T'.** Setting T' to 3 yields the best results across architectures, demonstrating its robustness and consistency.

Teacher	Student	T'		
		2	3	4
ResNet32×4	ResNet8×4	73.90	**75.46**	75.22
WRN-28-4	WRN-16-2	74.89	**75.79**	75.62
ResNet32×4	SN-V2	75.51	**75.91**	75.72

Table 9. **Impact of Reweighting Factors.** The proposed w_{EA} outperforms the single-sided w_{base}, while w_{interact}, lacking teacher base guidance, shows limited improvement.

Method	Acc.	Reweighting Factor		
		w_{base}	w_{interact}	w_{EA}
KD [11]	73.33	75.14 ↑	74.76 ↑	**75.46** ↑
MLD [14]	77.08	77.47 ↑	77.45 ↑	**77.65** ↑
MLD+LS [38]	78.28	78.30 ↑	78.20 ↓	**78.38** ↑
FCFD [26]	76.62	77.50 ↑	77.42 ↑	**77.44** ↑

Ablation Study. Two experiment were performed to evaluate the sensitivity of T' and the effect of each reweighting components in w_{EA}. (i) Tab. 8 shows that a T' of 3 consistently delivers optimal performance across model combinations, highlighting its robustness in reflecting sample value. (ii) As shown in Tab. 9, w_{base} enhances KD by emphasizing valuable samples under the teacher's guidance. While w_{interact} integrates student dynamics, it may introduce noise and dilute the teacher's guidance, reducing effectiveness. However, when integrated into w_{EA}, $H^{\mathcal{S}}$ acts as a scaling factor for $H^{\mathcal{T}}$ (Eq. (12)). This ensures early training, where $H^{\mathcal{S}}$ is mostly high and the scaling near 1, is mainly guided by $H^{\mathcal{T}}$. As training progresses, $H^{\mathcal{S}}$ adaptively adjusts and tailors the weighting, leading to superior performance.

Distillation Loss and Sample Value. In Fig. 5, we analyze the distribution of distillation loss across samples grouped by low- to high-entropy quartiles (Q1–Q4) for KD and EA-KD on CIFAR-100 and Tiny-ImageNet. Notably, KD loss is dominated by low-value samples throughout training on both datasets. The overshadowing effect in KD's uniform distillation scheme hinders the transfer of critical knowledge from high-value samples, as discussed in Sec. 3, ultimately leading to more significant accuracy gaps (Fig. 2a). In contrast, EA-KD emphasizes the focus on valuable samples (Fig. 5), bringing holistic performance improvements across entropy segments (Fig. 2a).

Student's Unique Perspective. Fig. 6 compares the t-SNE visualizations [41] of KD and EA-KD with the high $H^{\mathcal{S}}$ and w_{EA} samples highlighted[4]. Similar to Fig. 2b, high $H^{\mathcal{S}}$ samples also lie near decision boundaries in the KD-student. This suggests that, despite having a different view

[4]See Appendix B.1 for further t-SNE visualizations showing the student's evolving sample focus over training epochs.

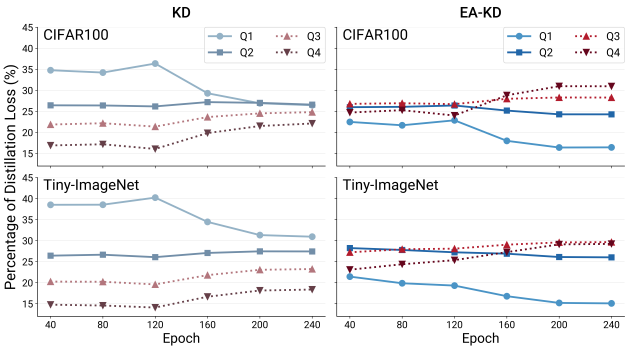


Figure 5. **Loss Distribution *vs*. Entropy Quartiles over Epochs.** Low-value samples (Q1, Q2; blue) dominate the KD loss, whereas EA-KD places more focus on high-value ones (Q3, Q4; red).

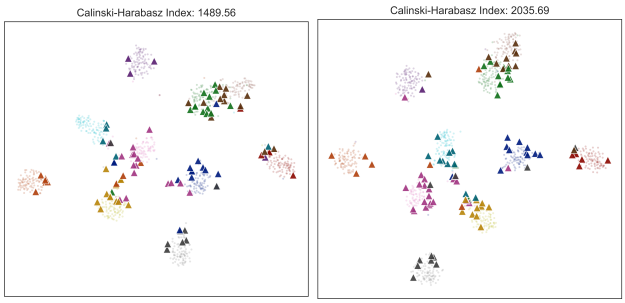


Figure 6. **t-SNE of Students.** Top 10% high $H^{\mathcal{S}}$ samples for KD (left) and w_{EA} samples for EA-KD (right) are highlighted with triangles. High $H^{\mathcal{S}}$ samples cluster near decision boundaries, underscoring their learning value for the student.

from the teacher (Fig. 3), the student's evolving capacity enables $H^{\mathcal{S}}$ to capture the samples crucial for its own learning progress. In EA-KD, integrating both $H^{\mathcal{T}}$ and $H^{\mathcal{S}}$ in w_{EA} captures both the teacher's informed assessment and the student's dynamic learning needs. As a result, EA-KD achieves superior class separability as reflected by a higher Calinski-Harabasz (CH) index.

Consistency in Entropy Levels. As shown in Fig. 3, EA-KD mitigates the increasing entropy variation across epochs observed in KD, enabling the student to maintain more stable $H^{\mathcal{S}}$ and align more closely to $H^{\mathcal{T}}$. This stability creates a feedback loop in EA-KD, where a steady $H^{\mathcal{S}}$ leads to a stable w_{EA}, promoting a more focused learning process. As a result, the student more effectively mimics the teacher's response and achieves improved performance.

Loss Landscape and Synergy in EA-DKD. Visualizing the loss landscape [19] offers insights into the student's robustness against noise and generalization. As illustrated in Fig. 7, DKD shows narrow and fluctuating contours with significantly higher variance across epochs (*e.g.* 366.26 at epoch 240), indicating instability in the loss surface and weaker generalization. In contrast, EA-DKD consistently produces a smoother surface with more stable contours (*e.g.*

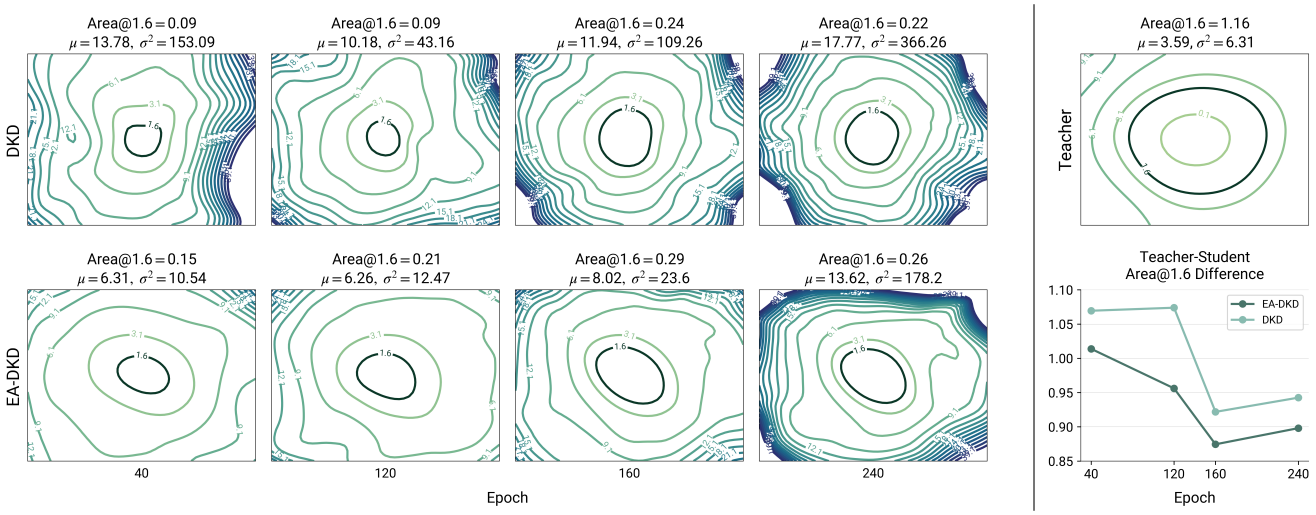


Figure 7. **Loss Surface and Differences in Area@1.6 for DKD and EA-DKD Students Across Epochs.** The mean and variance for the surface, along with the contour areas at level 1.6 (Area@1.6), are provided for each subplot. The line plot (lower right) tracks the differences in Area@1.6 between the teacher and students over epochs. EA-DKD consistently shows larger contours with less fluctuation, signifying smoother learning surfaces and a more robust generalization process compared to DKD.

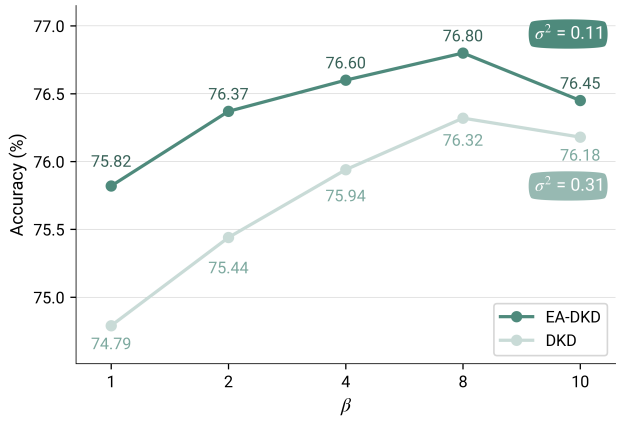


Figure 8. **Comparison of DKD and EA-DKD Performance Across Varying β Values.** EA-DKD consistently outperforms DKD and significantly reduces performance variance over β (0.10 *vs*. 0.31), highlighting EA-DKD's enhanced robustness.

variance of 178.20 at epoch 240) and broader low-loss regions (Area@1.6) compared to DKD, suggesting improved generalizability. In addition, the Area@1.6 comparison (Fig. 7, lower right) underscores EA-DKD's closer alignment with the teacher's surface throughout training.

The enhanced generalization in EA-DKD also extends to hyperparameter robustness. DKD was introduced to decouple target class KD (TCKD) and non-target class KD (NCKD) in vanilla KD, balancing them with hyperparameters α and β. While α remains stable around 1.0, adjusting β from 1.0 to 10.0 leads to fluctuating performance [47], with a variance of 0.31 (see Fig. 8). Notably, EA-DKD effectively reduces this fluctuation to a variance of 0.10 and consistently improves the performance. This enhanced robustness may be attributed to the synergy between DKD and EA in handling different aspects of knowledge transfer: DKD prevents NCKD from being overshadowed by TCKD at the class level, while EA ensures valuable samples—often rich in NCKD—are not dominated by simpler ones at the sample level. Additionally, the dynamic focus in EA compensates for the static nature of β. Together, EA-DKD facilitates a nuanced and balanced knowledge transfer process, both class-wise and sample-wise.

Computational Efficiency. Fig. 1 illustrates the efficiency of EA-methods, showing significant performance improvements across various SOTA KD frameworks with negligible cost. This remarkable efficiency, combined with its streamlined integration into both logit- and feature-based KD methods, underscores the potential of our method as a versatile and practical enhancement for KD.

5. Conclusion

In this paper, we revisited existing KD methods from a novel perspective and revealed a key limitation in their inherent uniform distillation strategy, which often hinders the transfer of high-entropy samples that carry critical knowledge. To address this, we proposed EA-KD, a plug-and-play KD approach that dynamically reweights the distillation loss, directing the learning focus toward valuable samples. EA-KD consistently enhances representative KD baselines across image classification, object detection, and LLM distillation, all with negligible cost. We believe EA-KD showcases a great paradigm of the meticulous handling of knowledge transfer, adapting KD to the varying learning value of samples while accounting for the student's evolving learning dynamics throughout the distillation process.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[2] 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*, pages 5008–5017, 2021. 1, 3, 5

[3] Xu Cheng, Zhefan Rao, Yilan Chen, and Quanshi Zhang. Explaining knowledge distillation by quantifying the knowledge. *CoRR*, abs/2003.03622, 2020. 3

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 5

[5] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. 5

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5

[8] Zhan Gao, Qika Lin, Huaxuan Wen, Bin Pu, Mengling Feng, and Kenli Li. Incorporating large vision model distillation and fuzzy perception for improving disease diagnosis. *IEEE Transactions on Fuzzy Systems*, 2025. 1

[9] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 5, 6

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 3, 5, 6, 7

[12] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*, 2022. 5

[13] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019. 1

[14] Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24276–24285, 2023. 1, 3, 5, 7

[15] Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 24872–24895, 2024. 1

[16] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 5

[17] Kisoo Kwon, Hwidong Na, Hoshik Lee, and Nam Soo Kim. Adaptive knowledge distillation based on entropy. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7409–7413, 2020. 1, 3

[18] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[19] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2018. 7

[20] 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*, pages 1504–1512, 2023. 1, 3, 5

[21] Pengchen Liang, Haishan Huang, Bin Pu, Jianguo Chen, Xiang Hua, Jing Zhang, Weibo Ma, Zhuangzhuang Chen, Yiwei Li, and Qing Chang. Task-specific knowledge distillation from the vision foundation model for enhanced medical image segmentation. *arXiv preprint arXiv:2503.06976*, 2025. 1

[22] Pengchen Liang, Leijun Shi, Huiping Yao, Bin Pu, Jianguo Chen, Lei Zhao, Haishan Huang, Zhuangzhuang Chen, Zhaozhao Xu, Lite Xu, et al. Rapid bone scintigraphy enhancement via semantic prior distillation from segment anything model. *arXiv preprint arXiv:2503.02321*, 2025.

[23] Pengchen Liang, Leijun Shi, Huiping Yao, Bin Pu, Jianguo Chen, Lei Zhao, Haishan Huang, Zhuangzhuang Chen, Zhaozhao Xu, Lite Xu, et al. Semantic prior distillation with vision foundation model for enhanced rapid bone scintigraphy image restoration. *arXiv e-prints*, pages arXiv–2503, 2025. 1

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[25] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[26] Dongyang Liu, Meina Kan, Shiguang Shan, and CHEN Xilin. Function-consistent feature distillation. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 3, 5, 6, 7

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[28] Peng Lu, Abbas Ghaddar, Ahmad Rashid, Mehdi Rezagholizadeh, Ali Ghodsi, and Philippe Langlais. RW-KD: Sample-wise loss terms re-weighting for knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3145–3152, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 3

[29] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 5

[30] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5191–5198, 2020. 1

[31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[32] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 5

[33] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 1, 3

[34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 5

[35] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 1

[36] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 1, 3

[37] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015. 5

[38] 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*, pages 15731–15740, 2024. 3, 5, 6, 7

[39] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 5

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 1, 5

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 2, 7

[42] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 5

[43] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*, 2022. 5

[44] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Procedings of the British Machine Vision Conference 2016*. British Machine Vision Association, 2016. 5

[45] Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. Dual-space knowledge distillation for large language models. *arXiv preprint arXiv:2406.17328*, 2024. 5

[46] Youcai Zhang, Zhonghao Lan, Yuchen Dai, Fangao Zeng, Yan Bai, Jie Chang, and Yichen Wei. Prime-aware adaptive distillation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 658–674. Springer, 2020. 3, 6

[47] 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*, pages 11953–11962, 2022. 1, 3, 5, 6, 8

[48] Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching. *arXiv preprint arXiv:2402.11148*, 2024. 3, 1

[49] Songling Zhu, Ronghua Shang, Bo Yuan, Weitong Zhang, Wenjie Li, Yangyang Li, and Licheng Jiao. Dynamickd: An effective knowledge distillation via dynamic entropy correction-based distillation for gap optimizing. *Pattern Recognition*, 153:110545, 2024. 1, 3