Evidential Knowledge Distillation

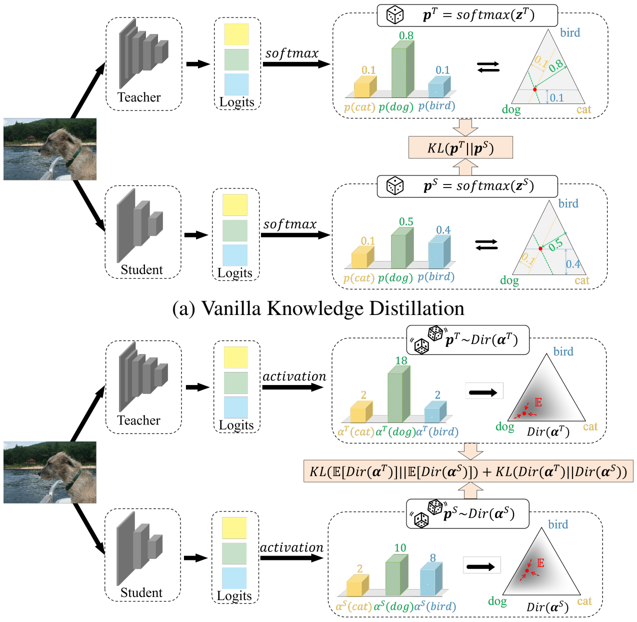
Liangyu Xiang[1, 2] Junyu Gao[1, 2 †] Changsheng Xu[1, 2, 3]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3]Peng Cheng Laboratory, ShenZhen, China
xiangliangyu2023@ia.ac.cn; {junyu.gao, csxu}@nlpr.ia.ac.cn

Abstract

Existing logit-based knowledge distillation methods typically employ singularly deterministic categorical distributions, which eliminates the inherent uncertainty in network predictions and thereby limiting the effective transfer of knowledge. To address this limitation, we introduce distribution-based probabilistic modeling as a more comprehensive representation of network knowledge. Specifically, we regard the categorical distribution as a random variable and leverage deep neural networks to predict its distribution, representing it as an evidential second-order distribution. Based on the second-oder modeling, we propose Evidential Knowledge Distillation (EKD) which 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. Additionally, we theoretically show that EKD's distillation objective provides an upper bound on the student's expected risk when treating the teacher's predictions as ground truth. Extensive experiments on several standard benchmarks across various teacher-student network pairs highlight the effectiveness and superior performance of EKD. Our code is available at https://github.com/lyxiang-casia/EKD.

1. Introduction

As deep neural networks (DNNs) continue to advance, their performance has rapidly improved, yet their increasing network size and training costs pose challenges, especially in resource-constrained and real-time processing scenarios [62, 89]. Consequently, logit-based knowledge distillation (KD) [30, 88], which aims to compress the predictive capabilities of complex models into lightweight ones, has gained increasing popularity. KD transfers knowledge by guiding the predictions of a smaller network to align with those of a larger network during the training.



(a) Vanilla Knowledge Distillation

Teacher
Logits
activation
$p^T \sim Dir(\alpha^T)$
2
18
2
$\alpha^T(cat)\alpha^T(dog)\alpha^T(bird)$
bird
dog
cat
$Dir(\alpha^T)$
$\mathbb{E}$
$KL(\mathbb{E}[Dir(\alpha^T)]||\mathbb{E}[Dir(\alpha^S)]) + KL(Dir(\alpha^T)||Dir(\alpha^S))$
Student
Logits
activation
$p^S \sim Dir(\alpha^S)$
2
10
8
$\alpha^S(cat)\alpha^S(dog)\alpha^S(bird)$
bird
dog
cat
$Dir(\alpha^S)$
$\mathbb{E}$

(b) Evidential Knowledge Distillation

Figure 1. Illustrations of vanilla KD and our EKD. (a) Vanilla KD deterministically converts the logits into categorical distributions through the softmax function. Each categorical distribution corresponds to a point on the simplex, with each categorical probability representing the barycentric coordinates of that point. Vanilla KD is achieved by minimizing the differences between the teacher's and student's barycentric coordinates. (b) EKD treats the categorical distribution as a random variable following a Dirichlet distribution, parameterized by the logits. The heatmap depicts the probability density of the Dirichlet distribution across the simplex. Moreover, EKD includes distillation of both the expectation of the second-order distribution, and the second-order distribution itself.

Current knowledge distillation methods largely adhere to the vanilla paradigm [30] where only the categorical distri-

[†]Corresponding author.

butions are aligned. As illustrated in Figure 1a, the logits of a complex network (teacher network) and a lightweight network (student network) are firstly converted into categorical probabilities using a softmax function. Knowledge is then transferred through the process of minimizing the KL divergence between the two categorical distributions. Enhancements to this distillation process, such as decoupling the categorical distributions [88], smoothing the distribution with multiple levels [45], or adopting sample-wise distribution smoothness [69], have demonstrated notable improvements in the student network's predictive performance. Despite these advancements, all such methods are constrained by a fundamental limitation in their probabilistic modeling, which assumes that categorical probabilities are singularly deterministic [12, 14]. Specifically, vanilla KD implicitly assumes that for any given sample, its probability belonging to a particular class is deterministic and can be approximated by a deep neural network. In the case of a three-class classification problem, as shown in Figure 1a, this assumption is akin to identifying a single point on a 2D simplex that best represents the entire simplex. However, in practice, networks trained on finite data with limited capability exhibit inherent uncertainty in their predictions for specific samples [39, 47]. For example, networks with different initial weights may yield different predictions for the same test samples. Consequently, this singularly deterministic probabilistic modeling creates a bottleneck in the effective representation of diverse and uncertain knowledge.

To overcome the limitations of singularly deterministic probabilistic modeling, we employ a distribution-based probabilistic modeling, treating the categorical distribution as a random variable governed by a second-order distribution. Following Subjective Logic [35] and Dempster–Shafer Theory of Evidence [81], the Dirichlet distribution, a conjugate prior of the categorical likelihood, instantiates this second-order distribution, facilitating the success of evidential deep learning across multiple domains [5, 16, 25]. However, recent studies employing distribution-based probabilistic modeling are largely confined to macro-level training within a single network, where the second-order distribution is integrated into a single categorical distribution during the network's training process [8, 11, 64]. The second-order distribution inherently provides micro-level information across all possible categorical distributions, offering advantages in cross-network information propagation. Therefore, the second-order distribution serves as a bridge to enable knowledge transfer at both the macro and micro levels, supported by PAC-Bayesain theory [19]. As illustrated in Figure 1b, the logits output by each network are converted into a set of Dirichlet parameters, forming respective second-order distributions. Essentially, the second-order distribution assigns varying weights to all possible first-order categorical distributions, where the non-uniformity of these weights reflects the network's capacity for classification. Building on second-order predictions, we propose the Evidential Knowledge Distillation (EKD), integrating both macro and micro perspectives. At the macro level, we reduce each second-order distribution to their first-order categorical distributions via an expectation operation in order to ensure that the centroids of their Dirichlet distributions coincide. By representing the entire simplex with a single point, the student is encouraged to focus more on the overall density distribution of the teacher at this stage, which leads to the proportions of student's outputs across categories aligning with those of the teacher. At the micro level, we replace the first-order distribution with the second-order distribution to enable the transfer of more granular classification information. As a complement to proportional alignment, this step refines the magnitude in the student network's outputs. Ultimately, the student model benefits from both the teacher's overall characteristics and its detailed classification structure.

To summarize, Our main contribution are as follows:

- We reformulate network predictions in knowledge distillation through distribution-based probabilistic modeling. Building on this reformulation, we propose EKD, which distills the teacher's second-order distribution into the student model across both macro and micro levels.
- With the help of PAC-Bayesian theory [19], we prove that the distillation objective in EKD serves as an upper bound on the expected risk of the student network when treating the teacher network's outputs as ground truth.
- Extensive experiments are conducted on several standard benchmarks, encompassing a wide range of teacher and student models. Without bells and whistles, EKD achieves performance improvement of up to 3.61% over KD across various teacher-student network combinations.

2. Related Work

Knowledge Distillation Knowledge distillation enhances a simpler student network using insights from a complex teacher network. Existing methods fall into three categories: logit-based [2, 30, 44, 45, 54, 85, 88], feature-based [33, 56, 57, 70, 73], and hybrid approaches [22, 31, 34, 58, 80, 82]. Unlike feature-based methods, which demand architecture-specific designs, logit-based methods offer wider applicability by relying solely on network outputs [4, 53, 84]. Introduced by Hinton and Geoffrey [30], logit-based distillation converts teacher and student logits into categorical distributions aligned via KL divergence. Subsequent efforts have refined this by adjusting the temperature parameter dynamically, based on training stages [45], teacher-student gaps [21], and output variance [69], or by decomposing distributions [88] and leveraging intermediate networks [31, 54].

The aforementioned traditional distillation methods share a common limitation in their reliance on point estimates for model parameters or categorical distributions, whereas Bayesian distillation [13, 37, 50, 74, 77] employs distributional estimation, which enables the student network to inherit the teacher's strengths in uncertainty estimation. Initial explorations [24, 37, 46, 51] adopt teacher ensembles to approximate the distribution of teacher model parameters and utilize the vanilla KD [30] for distillation. Subsequent extensions continue to advance by improving the student network architecture [20, 50, 61, 72, 76], optimization objectives [74], and sampling methods [13]. Additionally, some approaches introduce the utilization of quantized uncertainty during training to enable more fine-grained knowledge transfer [28, 66, 87]. While these methods introduce uncertainty into the model parameters and the resultant predictive distributions, they remain limited by their reliance on a finite number of ensembles to approximate the distributions. Consequently, we propose to adopt a continuous probabilistic modeling framework for both the teacher and the student, while leveraging more advanced distillation techniques. Unlike Bayesian distillation, which performs uncertainty-related distillation, we implicitly express the uncertainty of the predictive distribution through second-order probabilistic modeling and focus on enhancing the performance of the student network.

Evidential Deep Learning Based on Subjective Logic (SL) [35] and Dempster-Shafer Theory (DST) [65], Sensoy et al. [64] initially propose a framework in which evidence collected by a deep neural network for each class is used to parameterize a second-order Dirichlet distribution for network predictions. This concept of evidence collection allows evidential networks to express "I don't know" , showing advantages in uncertainty estimation. Since then, numerous studies have advanced and expanded upon Evidential Deep Learning (EDL) from both theoretical and application-oriented perspectives [17]. Theoretical investigations have primarily focused on refining the evidence collection process [11, 15, 18, 40, 55], delving into different training strategies [5, 26, 43] and extending evidential theory to regression networks [1, 49, 52]. Moreover, EDL has been widely applied in computer vision tasks, including action localization [5–7, 16], stereo matching [75], anomaly detection [68], and multi-view classification [25].

3. Method

In this section, we first present the fundamental framework of our Evidential Knowledge Distillation (EKD), which includes both macroscopic first-order and microscopic second-order distribution distillation. Theoretically, we utilize PAC-Bayesian theory to demonstrate that the optimization objective of EKD effectively represents an upper bound on the student's expected risk. At the end of this section, we illustrate the complementary effect of the two distillation strategies with a concrete toy case.

3.1. Preliminaries

Evidential Deep Learning. In accordance with the formulation of evidential deep learning [8, 9, 64], for a classification task involving K classes, the categorical probability vector $\boldsymbol{p}$ is treated as a random variable governed by a conjugate prior Dirichlet distribution. To define this distribution's parameters, an evidential activation function σ which we designate as the exponential function (exp), is employed to transform the network's output logits $\boldsymbol{z}$ into a non-negative evidence vector $\boldsymbol{e} = \sigma(\boldsymbol{z})$. This evidence vector is subsequently combined with a prior weight λ to yield the the K-dimensional Dirichlet parameters $\boldsymbol{\alpha}$ which fully specify a Dirichlet distribution $Dir(\boldsymbol{\alpha})$ [8]. In terms of optimization, the widely adopted cross-entropy loss is computed by integrating over the categorical probabilities, guided by second-order the Dirichlet distribution.

$$\begin{aligned}\mathcal{L}_{EDL-ce} &= \int \left[\sum_{i=1}^{K} -y_i \log(p_i)\right] Dir(\boldsymbol{p}|\boldsymbol{\alpha}) d\boldsymbol{p} \\ &= \sum_{i=1}^{K} y_i \left(\psi(\alpha_0) - \psi(\alpha_i)\right) \end{aligned} \tag{1}$$

Here, y_i denotes the ground truth, $\alpha_0 = \sum_{i=1}^{K}(\alpha_i)$ represents the Drichlet strength, and ψ refers to the digamma function. To support the acquisition of evidence, we employ Eq. (1), rather than the conventional cross-entropy loss, across all network training stages, including both the training of the teacher networks and student networks during the knowledge distillation process.

Knowledge Distillation. Existing logit-based knowledge distillation methods rely on categorical distribution predictions derived from the softmax function. Specifically, for a given input sample x, the network produces logits $\boldsymbol{z}$, which are then transformed into a categorical distribution via the softmax operation, denoted as $\boldsymbol{p} = \text{softmax}(\boldsymbol{z})$. Knowledge transfer is accomplished by minimizing the Kullback-Leibler divergence(KL) divergence between the teacher's and student's predicted distributions, $\boldsymbol{p}^T$ and $\boldsymbol{p}^S$:

$$\mathcal{L}_{KD} = KL(\boldsymbol{p}^T || \boldsymbol{p}^S) \tag{2}$$

The teacher's predictions serve as soft labels, replacing the traditional one-hot labels, capturing inter-class similarities.

3.2. Evidential Knowledge Distillation

Previous knowledge distillation methods have generally followed a unified paradigm, where the network's output logits are transformed into categorical probabilities via the softmax function, and the student network is guided to align its categorical distribution with that of the teacher network. While these approaches have yielded notable success, a fundamental limitation persists: they treat categorical probabilities as singularly deterministic values. This assumption overlooks the uncertainty in network predictions, thus constraining the sharing of more granular information. As a more robust alternative, we predict the distribution of categorical distributions using neural networks, specifically modeling it as a second-order Dirichlet distribution. More importantly, we align both the first-order distribution, obtained by integrating the second-order distribution, and the second-order distribution itself.

Firstly, we aim to extract the overall characteristics of the second-order distribution to capture the general distribution trend of probability density across the simplex, as well as to achieve alignment between networks in terms of final predicted classes. Specifically, given the second-order distributions predicted by the teacher and student, $Dir(\boldsymbol{\alpha}^T)$ and $Dir(\boldsymbol{\alpha}^S)$, we compute their expectations where all possible categorical distributions are summed under the weights specified by the second-order distribution:

$$\hat{\boldsymbol{p}}^T = \mathbb{E}_{\boldsymbol{p}^T \sim Dir(\boldsymbol{\alpha}^T)}[\boldsymbol{p}^T], \hat{\boldsymbol{p}}^S = \mathbb{E}_{Dir(\boldsymbol{p}^S \sim \boldsymbol{\alpha}^S)}[\boldsymbol{p}^S] \tag{3}$$

The expectation above, serving as the final predictive outcome for both teacher and student, is equivalent to the centroid of the second-order distribution, enabling the transfer of comprehensive probability density information via first-order distillation at a macro level.

$$\mathcal{L}_{1st} = KL(\hat{\boldsymbol{p}}^T || \hat{\boldsymbol{p}}^S) = \sum_{i=1}^{K} \frac{\alpha_i^T}{\alpha_0^T} log(\frac{\alpha_i^T \alpha_0^S}{\alpha_0^T \alpha_i^S}) \tag{4}$$

Formally, optimizing Eq. (4) involves adjusting the proportion of each class's Dirichlet parameter relative to the total, which significantly enhances the alignment between the student's and teacher's predicted classes. In terms of formulation, Eq. (4) is very similar to vanilla KD, yet it is emphasized that vanilla KD can be regarded as a special case of first-order distillation when $\sigma = exp(\cdot)$ and $\lambda = 0$. Proof is provided in the **Supplementary Material**.

The first-order distillation objective above only provides macroscopic guidance for the student's second-order distribution, failing to accurately and comprehensively represent the weighting of each categorical distribution. In other words, given a fixed teacher distribution, the student Dirichlet distribution that minimizes Eq. (4) is not unique. Therefore, we propose a microscopic second-order distillation as a complement to first-order distillation, aiming to provide fine-grained teacher supervision signals across the entire simplex for the student distribution. Specifically, we achieve this by applying the KL divergence directly between the two Dirichlet distributions:

$$\begin{aligned}\mathcal{L}_{2nd} &= KL(Dir(\boldsymbol{p}^T|\boldsymbol{\alpha}^T)||Dir(\boldsymbol{p}^S|\boldsymbol{\alpha}^S)) \\ &= log\frac{\Gamma(\alpha_0^T)}{\Gamma(\alpha_0^S)} - \sum_{i=1}^{K} log\frac{\Gamma(\alpha_i^T)}{\Gamma(\alpha_i^S)} \\ &+ \sum_{i=1}^{K}(\alpha_i^T - \alpha_i^S)(\psi(\alpha_i^T) - \psi(\alpha_0^T))\end{aligned} \tag{5}$$

Here, $\Gamma(\cdot)$ denotes the Gamma function. The detailed derivation for Eq. (5) is provided in **Supplementary Material**. Since Eq. (5) considers the whole $(K-1)$ dimensional simplex, overflow may occur [50]; thus, we apply the $\log(1+x)$ function to smooth the network outputs. While the above Eq. (4) aims to optimize the proportional relationships between classes, Eq. (5) here places greater emphasis on aligning the magnitudes within each class. Here, the alignment within the class means that the gradients of a loss function with respect to the output for a particular class do not conflict with the gradients of the output for other classes. This meticulous numerical alignment supplies each point on the $(K-1)$-dimensional simplex of the student's second-order distribution with the corresponding teacher label, ensuring a precise mapping of knowledge across the entire distribution space. By doing so, it strives to faithfully replicate the teacher's second-order predictive nuances at a granular, micro-level, capturing the fine-grained patterns and uncertainties inherent in the teacher model's outputs.

By combining these two complementary optimization objectives, one optimizing proportional relationships at the macro level and the other aligning numerical values at the micro level, we arrive at the final EKD distillation:

$$\mathcal{L}_{EKD} = \mathcal{L}_{1st} + \gamma\mathcal{L}_{2nd}, \tag{6}$$

where γ represents the coefficient that adjusts the balance between the first-order and second-order distillation.

3.3. Theoretical Analysis

In this section, we provide theoretical support for Eq. (6) using PAC-Bayesian Theory [19]. While PAC-Bayesian theory has been explored in the context of the Dirichlet distribution [26], its application remains confined to single-network optimization, neglecting the relationship among multiple networks. Therefore, we aim to estimate the alignment of different networks across the entire data distribution based on their alignment on training samples.

Let x denote a certain sample and y denote its corresponding label. Under the PAC setting, samples are drawn from a fixed but unknown distribution $\mathcal{D}$. A classifier h maps x to y. Moreover, we define a classifier h correspond to a point on the $(K-1)$-dimensional simplex, i.e., a categorical distribution $\boldsymbol{p}$. Essentially, the Dirichlet distribution made by each network serves as the posterior distribution of a set of classifiers. By weighting the set of classifiers with its posterior distribution, we obtain the final majority vote classifier (or Bayesain classifier):

$$\boldsymbol{B}^T = \mathbb{E}_{Dir(\boldsymbol{\alpha}^T)}[\boldsymbol{p}^T], \tag{7}$$

$$\boldsymbol{B}^S = \mathbb{E}_{Dir(\boldsymbol{\alpha}^S)}[\boldsymbol{p}^S], \tag{8}$$

where $\boldsymbol{B}^T$ and $\boldsymbol{B}^S$ denote the final classification results of the teacher and student classifiers, respectively.

In knowledge distillation, the student treats the teacher's predictions as soft labels. Below, we define the student's misclassification risk using cross-entropy, which is equivalent to KL divergence from an optimization perspective.

$$R^S = \mathbb{E}_{x\sim\mathcal{D}}\big[-\boldsymbol{B}^T \log \boldsymbol{B}^S\big], \tag{9}$$

$$\hat{R}^S = \frac{1}{N}\sum_{i=1}^{N}\big[-\boldsymbol{B}_i^T \log \boldsymbol{B}_i^S\big]. \tag{10}$$

where R^S represents the expected risk of the student over the entire data distribution $\mathcal{D}$ and $\hat{R}^S$ denotes the empirical risk on the training set $\mathcal{D}_t$ containing N training samples.

Theorem 1. *For any $\delta \in (0,1]$ and $\gamma \in \mathbb{R}^+$, R^S is bounded above by the sum of $\hat{R}^S$, $\sum_1^N KL(Dir(\boldsymbol{\alpha}_i^T)||Dir(\boldsymbol{\alpha}_i^S))$, and a constant related to δ, denoted as $C(\delta)$:*

$$P_{\mathcal{D}_t\sim\mathcal{D}^N}(R^S \leq \hat{R}^S + \frac{\gamma}{N}\sum_{i=1}^{N} KL(Dir(\boldsymbol{\alpha}_i^T)||Dir(\boldsymbol{\alpha}_i^S)) + C(\delta)) \geq 1-\delta. \tag{11}$$

Detailed proof and definition of $C(\delta)$ can be found in **Supplementary Material**. Theorem 1 indicates that the upper bound on the student's excepted risk over the unknown data distribution $\mathcal{D}$ consists of two components: the alignment of the first-order distribution with the true labels (teacher labels) and the similarity between the predicted second-order distribution and the true second-order distribution (teacher distribution). Clearly, this upper bound indicates that the empirical risk optimization represented by Eq. (2) does not guarantee a reduction in the student's expected risk, suggesting that the transfer of the teacher's generalization performance is both inefficient and incomplete. In contrast, EKD directly adopts this upper bound as optimization objectives, refining the formulation of empirical risk while incorporating second-order distribution alignment between the student and teacher, thereby striving to minimize the student's expected risk.

3.4. Toy Case

Figure 2 demonstrates the complementary roles of first-order and second-order distillation through a three-class example. Initially, the student Dirichlet parameters differ significantly from those of the teacher, both in relative proportion and in absolute values. The left side of Figure 2 illustrates the essence of first-order distillation, which minimizes the distance between the expectations of two Dirichlet distributions over the $(K-1)$ dimensional simplex measured by KL divergence. However, the expectation of the Dirichlet distribution depends solely on the proportion of each class's Dirichlet parameter relative to the overall sum, which can lead to discrepancies in the specific values of the student's output compared to the teacher's. Although this discrepancy does not affect the final class prediction, it causes the student's second-order distribution to differ significantly from the teacher's. To address this, we use second-order distillation as a complement to first-order distillation, aiming to align the Dirichlet parameter values. In contrast to first-order distillation, second-order distribution alignment encourages consistency of parameters within the same class across different networks, disregarding the relative magnitudes among different classes. As a result, we combine these complementary distillation objectives to achieve a comprehensive alignment of the second-order distribution. The quantitative effects of the two types of distillation will be discussed in detail in Section 4.2

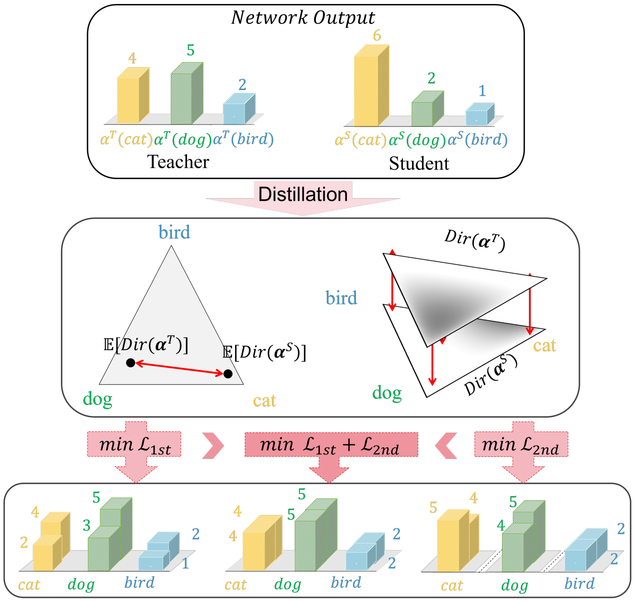

Network Output
4
5
2
$\alpha^T(cat)\alpha^T(dog)\alpha^T(bird)$
Teacher
6
2
1
$\alpha^S(cat)\alpha^S(dog)\alpha^S(bird)$
Student
Distillation
bird
$\mathbb{E}[Dir(\boldsymbol{\alpha}^T)]$
$\mathbb{E}[Dir(\boldsymbol{\alpha}^S)]$
dog
cat
$Dir(\boldsymbol{\alpha}^T)$
bird
cat
dog
$Dir(\boldsymbol{\alpha}^S)$
$min\ \mathcal{L}_{1st}$
$min\ \mathcal{L}_{1st} + \mathcal{L}_{2nd}$
$min\ \mathcal{L}_{2nd}$
4
2
5
3
2
1
cat
dog
bird
4
4
5
5
2
2
cat
dog
bird
5
4
5
4
2
2
cat
dog
bird

Figure 2. A toy case illustrating the complementary roles of the two distillation losses in EKD. Given the Dirichlet parameters output by teacher and student networks, the left panel shows the results when only first-order distillation is used to optimize the proportion of each category. The right panel presents the outcome when only second-order distillation is applied for intra-class numerical alignment. The center panel presents the results of combining both objectives, which yields the optimal performance.

4. Experiments

Datasets. Our experiments are primarily conducted on two standard datasets including CIFAR100 [38] and ImageNet [60]. The CIFAR-100 dataset [38] consists of 60k 32x32 color images divided into 100 classes, with each class containing 600 images. It is split into a training set of 50k images and a validation set of 10k images. ImageNet [60] is a large-scale visual database comprising 1,000 categories that cover a diverse range of objects and scenes. It includes approximately 1.28 million labeled images for training and an additional 50,000 images for validation.

Implementation Details. We adopt the experimental settings established in prior studies [10, 69, 88]. For the

Table 1. Results on the CIFAR-100 validation set. The teacher and student networks share the same architecture. The best and second-best results among logits-based methods are highlighted in **bold** and underlined, respectively. Δ denotes the improvement of the distilled student model compared to the independently trained student model. Note that "Softmax" indicates that the model uses the vanilla KD probabilistic model, while "EDL" indicates that the model are trained by Eq. (1). "*" indicates our reproduced results.

	Experiment Group	1	2	3	4	5	6	7
Teacher	Architecture	ResNet32×4	VGG13	WRN-40-2	WRN-40-2	ResNet56	ResNet110	ResNet110
	Softmax	79.42	74.64	75.61	75.61	72.34	74.31	74.31
	EDL	79.53	74.96	75.41	75.41	73.09	74.67	74.67
Student	Architecture	ResNet8×4	VGG8	WRN-40-1	WRN-16-2	ResNet20	ResNet32	ResNet20
	Softmax	72.50	70.36	71.98	73.26	69.06	71.14	69.06
	EDL	72.77	70.67	71.84	73.36	69.22	70.9	69.22
Feature	RKD[CVPR'19] [56]	71.90	71.48	72.22	73.35	69.61	71.82	69.25
	CRD[ICLR'20] [70]	75.51	73.94	74.14	75.48	71.16	73.48	71.46
	ReviewKD[CVPR'21] [10]	75.63	74.84	75.09	76.12	71.89	73.89	71.34
	SimKD[CVPR'22] [3]	78.08	74.89	74.53	75.53	71.05	73.92	71.06
	CAT-KD[CVPR'23] [23]	76.91	74.65	74.82	75.60	71.62	73.62	71.37
	TopKD[ICML'24] [36]	75.40	74.01	74.43	75.75	71.58	73.77	71.47
Logit	KD[NeurIPS'14] [30]	73.33	72.98	73.54	74.92	70.66	73.08	70.67
	Δ	0.83	2.62	1.56	1.66	1.6	1.94	1.61
	DKD[CVPR'22] [88]	76.32	74.68	**74.81**	**76.24**	**71.97**	74.11	71.06
	Δ	3.82	4.32	2.83	2.98	2.91	2.97	2.0
	CTKD[AAAI'23] [45]	73.39	73.52	73.93	75.45	71.19	73.52	70.99
	Δ	0.89	3.16	1.95	2.19	2.13	2.38	1.93
	Logit_Stand[CVPR'24] [69]	76.62	74.36	74.37	76.11	71.43	**74.17**	71.48
	Δ	4.12	4.00	2.39	2.85	2.37	3.03	2.42
	SDD[CVPR'24] [48]	75.92*	73.90*	74.26*	75.98*	70.72*	73.90*	71.43*
	Δ	3.42*	3.54*	2.28*	2.72*	1.66*	2.76*	2.37*
	TeKAP[ICLR'25] [31]	74.79	73.8*	73.80	75.21	71.71	73.50*	71.00*
	Δ	2.29	3.44*	1.82	1.95	2.65	2.36*	1.94*
	EKD(Ours)	**77.21**	**74.71**	74.43	76.15	71.48	73.68	**71.51**
	Δ	4.44	4.04	2.59	2.79	2.26	2.78	2.29

Table 3. The top-1 and top-5 accuracy (%) on the ImageNet validation set [60]. The best and second best results are emphasized in **bold** and underlined. "*" indicates our reproduced results.

Teacher	Architecture	ResNet34		ResNet50	
	Accuracy	top-1	top-5	top-1	top-5
	Softmax	73.31	91.42	76.16	92.86
	EDL	73.83	91.69	76.10	92.98
Student	Architecture	ResNet18		MNV1	
	Accuracy	top-1	top-5	top-1	top-5
	Softmax	69.75	89.07	68.87	88.76
	EDL	70.28	89.58	69.92	89.46
Feature	CRD [70]	71.17	90.13	71.37	90.41
	ReviewKD [10]	71.61	90.51	72.56	91.00
	SimKD [3]	71.59	90.48	72.25	90.86
	CAT-KD [23]	71.26	90.45	72.24	91.13
Logit	KD [30]	71.03	90.05	70.50	89.80
	DKD [88]	71.70	90.41	72.05	91.05
	CTKD [45]	71.38	90.27	71.16	90.11
	Logit_Stand [69]	71.42	90.29	72.18	90.80
	SDD [48]	71.44	90.05	72.15*	90.79*
	EKD(Ours)	**71.73**	**90.69**	**72.54**	**91.20**

experiments on CIFAR-100, we use the SGD optimizer with 240 epochs. The learning rate is set to 0.01 for MobileNets [32, 63] and ShuffleNets [86] and 0.05 for other networks, including ResNets [27], WRNs [83] and VGGs [67], with a decay factor of 0.1 applied at epochs 150, 180, and 210. Regarding hyperparameter configuration, we set γ in Eq. (6) to 1 by default, requiring no additional hyperparameter tuning. All results represent the mean of five independent runs. Additional experimental details are provided in the **Supplementary Material**.

Baselines. To benchmark EKD, we incorporate five state-of-the-art logit-based distillation methods: DKD (CVPR'22) [88], CTKD (AAAI'23) [45], Logit_Stand (CVPR'24) [69], SDD (CVPR'24) [48] and TeKAP (ICLR'25) [31]. Additionally, a selection of feature-based distillation methods [3, 10, 23, 56, 59, 70] is included as references to facilitate a broader contextual analysis.

4.1. Main Results

Performance comparison on CIFAR100. We conduct experiments comparing EKD with other distillation methods under various teacher-student pair settings. Table 1 and Table 2 present the accuracy results for settings where the teacher and student have identical and distinct architectures. As presented in Table 1, EKD demonstrates favourable performance improvements over KD, ranging from 0.66% to 3.61%. In particular, EKD demonstrates its strongest performance in Groups 1, 2, and 7, while achieving comparable results to DKD [88] in Groups 3–6. The underlying reason for this phenomenon lies in the performance gap between the teacher and student. In groups 1, 2, and 7, the average performance gap between the teacher and student is 5.51%, whereas it is 3.31% in groups 3–6. Furthermore, we observe that pairs with a larger performance gap tend to have a higher distillation loss, indicating a greater opti-

Table 2. Results on the CIFAR-100 validation set. The teacher and student networks have different architectures. The best and second-best results among logits-based methods are highlighted in **bold** and underlined, respectively. Δ denotes the improvement of the distilled student model compared to the independently trained student model. Note that "Softmax" indicates that the model uses the vanilla KD probabilistic model, while "EDL" indicates that the model are trained by Eq. (1). "*" indicates our reproduced results.

	Experiment Group	1	2	3	4	5	6	7
Teacher	Architecture	ResNet32×4	ResNet32×4	ResNet32×4	WRN-40-2	WRN-40-2	VGG13	ResNet50
	Softmax	79.42	79.42	79.42	75.61	75.61	74.64	79.34
	EDL	79.53	79.53	79.53	75.41	75.41	75.11	79.69
Student	Architecture	SHN-V2	WRN-16-2	WRN-40-2	ResNet8×4	MN-V2	MN-V2	MN-V2
	Softmax	71.82	73.26	75.61	72.50	64.60	64.60	64.60
	EDL	72.73	73.36	75.41	72.77	65.39	65.39	65.39
Feature	RKD[CVPR'19] [56]	73.21	74.86	77.82	75.26	69.27	64.52	64.43
	CRD[ICLR'20] [70]	75.65	75.65	78.15	75.24	70.28	69.73	69.11
	ReviewKD[CVPR'21] [10]	77.78	76.11	78.96	74.34	71.28	70.37	69.89
	SimKD[CVPR'22] [3]	78.39	77.17	79.29	75.29	70.10	69.44	69.97
	CAT-KD[CVPR'23] [23]	78.41	76.97	78.59	75.38	70.24	69.13	71.36
	TopKD[ICML'24] [36]	76.33	74.34*	76.18*	75.83*	69.54*	69.32*	70.11*
Logit	KD[NeurIPS'14] [30]	74.45	74.90	77.70	73.97	68.36	67.37	67.35
	Δ	2.63	1.64	2.09	1.47	3.76	2.77	2.75
	DKD[CVPR'22] [88]	77.07	75.70	78.46	75.56	69.28	69.71	70.35
	Δ	5.25	2.44	2.85	3.06	4.68	5.11	5.75
	CTKD[AAAI'23] [45]	75.37	74.57	77.66	74.61	68.34	68.50	68.67
	Δ	3.55	1.31	2.05	2.11	3.74	3.90	4.07
	Logit_Stand[CVPR'24] [69]	75.56	75.26	77.92	**77.11**	69.23	68.61	69.02
	Δ	3.74	2.00	2.31	4.61	4.63	4.01	4.42
	SDD[CVPR'24] [48]	76.88*	75.92*	78.54*	74.83*	**70.14***	69.06*	69.31*
	Δ	5.06*	2.66*	2.93*	2.33*	5.54*	4.46*	4.71*
	TeKAP[ICLR'25] [31]	75.43	75.26*	76.93*	74.99*	68.65*	68.27*	68.75*
	Δ	3.61	2.00*	1.32*	2.49*	4.05*	3.67*	4.15*
	EKD(Ours)	**77.65**	**76.51**	**78.91**	76.75	69.20	**69.94**	**70.38**
	Δ	4.92	3.15	3.50	3.98	3.81	4.55	4.99

Table 4. Ablation study for each components in EKD. The "MSE" refers to aligning the magnitude of model outputs with mean square error. The "Accuracy" denotes the Top-1 accuracy (%) on the CIFAR100 validation set. ResNet32×4 and ResNet8×4 are adopted as teacher and student.

Index	Distillation Objective				Accuracy
	$\mathcal{L}_{KD}$	$\mathcal{L}_{1st}$	$\mathcal{L}_{2nd}$	MSE	
1	✓				74.45
2		✓			73.80
3			✓		76.42
4				✓	76.10
5		✓	✓		**77.21**
6		✓		✓	76.83
7	✓		✓		76.34
8	✓			✓	76.13

mization space during distillation. Conversely, for teacher-student pairs with closer performance, their distillation loss is generally smaller, leading EKD to encounter local minima during optimization. For the distillation of distinct networks shown in Table 2, EKD achieves the highest performance in half of the experimental groups. When using MobileNet [63] as the student network, the student trained solely with EDL achieves a 0.79% higher performance than the one trained with softmax alone. However, the result for EKD is only comparable to that of other methods. We attribute this phenomenon to the inherent performance ceiling of the student network due to its limited capacity.

Performance comparison on ImageNet. Table 3 presents the performance of EKD on the large-scale ImageNet dataset. EKD markedly outperforms other logit-based methods in the ResNet50-MNV1 experiment and achieves results comparable to top methods in the ResNet34-ResNet18 experiment. The modest gains in the latter can also be attributed to the narrow performance gap between the teacher and student.

4.2. Ablation Studies

To elucidate the superiority of the two alignment strategies implemented in EKD and their alignment with theoretical underpinnings, the ablation study presented in Table 4 evaluates various distillation objectives on student model performance. In addressing alignment with respect to the proportional distribution across output classes, we investigate two methodologies: the conventional vanilla KD approach presented in Eq. (2), which aligns softmax-based categorical distributions, and $\mathcal{L}_{1st}$ in Eq. (4). Comparative analyses between Index 1 and Index 2, Index 5 and Index 7, and Index 6 and Index 8 consistently demonstrate that $\mathcal{L}_{1st}$ surpasses vanilla KD. This finding underscores that EKD's first-order alignment serves as a robust generalization of vanilla KD, effectively incorporating the alignment of categorical distributions by leveraging the centroids of second-order Dirichlet distributions.

In addition to the second-order distillation, $\mathcal{L}_{2nd}$, shown

Table 5. The Top-1 accuracy (%) of combining EKD with feature distillation-based methods on CIFAR-100.

Teacher	Architecture	VGG13	ResNet32x4
	Accuracy	74.96	79.53
Student	Architecture	VGG8	SHN-V2
	Accuracy	70.67	72.73
Logit	EKD	74.71	77.65
Feature	RKD [56]	71.48	73.21
	CAT-KD [23]	74.65	78.41
Logit+Feature	EKD+RKD [56]	**74.99**↑	**78.08**↑
	EKD+CAT-KD [23]	**74.91**↑	**78.46**↑

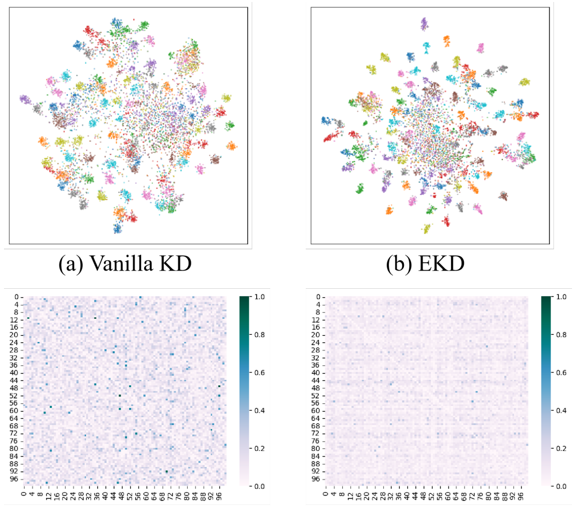
in Eq. (5), another intuitive approach to aligning model output magnitudes is through mean squared error (MSE). Consequently, we compare $\mathcal{L}_{2nd}$ and MSE in Table 4. Comparisons across Index 3 and Index 4, Index 5 and Index 6, and Index 7 and Index 8 reveal that $\mathcal{L}_{2nd}$ consistently outperforms MSE. These empirical results corroborate the theoretical conclusions delineated in Section 3.3. It is suggested that $\mathcal{L}_{2nd}$, being part of the upper bound on the student's expected risk, directly optimizes the student's generalization performance, whereas the MSE objective does so indirectly.

4.3. Further Remarks

Visualizations. Figure 3 presents the t-SNE visualizations of the laerned feature representations and, in parallel, the differences in logits between the teacher and student networks, where the teacher is implemented as ResNet32×4 and the student as ResNet8×4. As shown in the first row, EKD achieves a clear improvement over vanilla KD baseline with respect to class separability, leading to more distinctly clustered feature distributions. Furthermore, the logits produced under EKD exhibit a noticeably stronger alignment with those of the teacher network, indicating a closer match in predictive behavior compared to the results obtained with vanilla KD.

Compatibility with feature-based methods. Table 5 presents the results of integrating EKD with several representative feature-based distillation methods [23, 56]. For feature-based approaches, the introduction of EKD enhances their performance, demonstrating strong compatibility, particularly with methods that originally exhibit lower performance [56]. Conversely, these methods can further compensate for the limitation of the EKD approach in expressing knowledge through logits.

Distilling ViT. Transformer-based models are gaining increasing popularity, and logit-based distillation methods can be easily applied to these models. Therefore, we test the distillation effectiveness of EKD on the ViT model, as presented in Table 6. Following the configuration in [41, 42], we use ResNet56 as the teacher model and four ViT variants, DeiT-Ti [71], PiT-Ti [29], PVT-Ti [78], and PVTv2 [79], as the student models. EKD achieves an average performance improvement of 3.75% over KD [30] and an average improvement of 0.25% over AutoKD [42].



(a) Vanilla KD (b) EKD

(c) Vanilla KD (d) EKD

Figure 3. Visualization of features and logits output. The first row shows the t-SNE of features learned by vanilla KD and EKD. The second row illustrates the difference of correlation matrices of logits between the student and teacher.

Table 6. Top-1 accuracy of various ViT student models on CIFAR100. Teacher model is ResNet56.

Architecture	Size	EDL	Softmax	KD [30]	AutoKD [42]	EKD(Ours)
DeiT-Ti [71]	5M	64.77	65.08	73.25	78.55	**78.64**
PiT-Ti [29]	5M	73.48	73.58	75.47	78.76	**79.33**
PVT-Ti [78]	13M	69.02	69.22	73.60	78.43	**78.74**
PVTv2 [79]	3M	76.68	77.44	78.81	79.37	**79.80**

5. Conclusion

This paper reinterprets vanilla knowledge distillation (KD) through a distribution-based probabilistic framework, employing evidential second-order distributions to effectively capture predictive uncertainty and provide a comprehensive knowledge representation. This pioneering methodology underpins Evidential Knowledge Distillation (EKD), facilitating knowledge transfer across macro and micro levels. At the macro level, aligning the expectation of the second-order distribution, which reflects its global characteristics, enhances optimization of inter-class proportions in the student's output. At the micro level, aligning the second-order distribution aligns the magnitudes of the student's output. These complementary mechanisms ensure a robust distillation process. Through PAC-Bayesian theory, EKD optimizes the upper bound of the student's expected risk. Extensive experiments underscore EKD's pronounced advantage over prevailing logit-based distillation techniques.

Acknowledgements

This work was supported in part by the National Key Research and Development Plan of China under Grant 2023YFC3310700, in part by the National Natural Science Foundation of China under Grants 62036012, U21B2044, 62236008, 62472422, and U2333215, and in part by Beijing Natural Science Foundation under Grant 4242051.

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *NIPS*, 33:14927–14937, 2020. 3

[2] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *AAAI*, pages 3430–3437, 2020. 2

[3] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *CVPR*, pages 11933–11942, 2022. 6, 7

[4] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *NIPS*, 30, 2017. 2

[5] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *ECCV*, pages 192–208. Springer, 2022. 2, 3

[6] Mengyuan Chen, Junyu Gao, and Changsheng Xu. Cascade evidential learning for open-world weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14741–14750, 2023.

[7] Mengyuan Chen, Junyu Gao, and Changsheng Xu. Uncertainty-aware dual-evidential learning for weakly-supervised temporal action localization. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):15896–15911, 2023. 3

[8] Mengyuan Chen, Junyu Gao, and Changsheng Xu. R-edl: Relaxing nonessential settings of evidential deep learning. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[9] Mengyuan Chen, Junyu Gao, and Changsheng Xu. Revisiting essential and nonessential settings of evidential deep learning. *arXiv preprint arXiv:2410.00393*, 2024. 3

[10] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, pages 5008–5017, 2021. 5, 6, 7

[11] Danruo Deng, Guangyong Chen, Yang Yu, Furui Liu, and Pheng-Ann Heng. Uncertainty estimation by fisher information-based evidential deep learning. In *ICML*, pages 7596–7616. PMLR, 2023. 2, 3

[12] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000. 2

[13] Luyang Fang, Yongkai Chen, Wenxuan Zhong, and Ping Ma. Bayesian knowledge distillation: A bayesian perspective of distillation with uncertainty quantification. In *Forty-first ICML*, 2024. 3

[14] Yarin Gal et al. Uncertainty in deep learning. 2016. 2

[15] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18827–18836, 2023. 3

[16] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Vectorized evidential learning for weakly-supervised temporal action localization. *PAMI*, 2023. 2, 3

[17] Junyu Gao, Mengyuan Chen, Liangyu Xiang, and Changsheng Xu. A comprehensive survey on evidential deep learning and its applications. *arXiv preprint arXiv:2409.04720*, 2024. 3

[18] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Learning probabilistic presence-absence evidence for weakly-supervised audio-visual event perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[19] Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. Pac-bayesian learning of linear classifiers. In *ICML*, pages 353–360, 2009. 2, 4

[20] Hongji Guo, Hanjing Wang, and Qiang Ji. Bayesian evidential deep learning for online action detection. In *ECCV*, pages 283–301. Springer, 2025. 3

[21] Jia Guo. Reducing the teacher-student gap via adaptive temperatures. 2022. 2

[22] Yong Guo, Shulian Zhang, Haolin Pan, Jing Liu, Yulun Zhang, and Jian Chen. Gap preserving distillation by building bidirectional mappings with a dynamic teacher. *arXiv preprint arXiv:2410.04140*, 2024. 2

[23] Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation. In *CVPR*, pages 11868–11877, 2023. 6, 7, 8

[24] Corina Gurau, Alex Bewley, and Ingmar Posner. Dropout distillation for efficiently estimating model confidence. *arXiv preprint arXiv:1809.10562*, 2018. 3

[25] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *ICLR*, 2020. 2, 3

[26] Manuel Haussmann, Sebastian Gerwinn, and Melih Kandemir. Bayesian evidential deep learning with pac regularization. *arXiv preprint arXiv:1906.00816*, 2019. 3, 4

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[28] Mina Hemmatian, Ali Shahzadi, and Saeed Mozaffari. Uncertainty-based knowledge distillation for bayesian deep neural network compression. *International Journal of Approximate Reasoning*, 175:109301, 2024. 3

[29] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *CVPR*, pages 11936–11945, 2021. 8

[30] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 3, 6, 7, 8

[31] Md Imtiaz Hossain, Sharmen Akhter, Choong Seon Hong, and Eui-Nam Huh. Single teacher, multiple perspectives: Teacher knowledge augmentation for enhanced knowledge distillation. In *ICLR*, 2025. 2, 6, 7

[32] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 6

[33] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *NIPS*, 35:33716–33727, 2022. 2

[34] Yaomin Huang, Zaomin Yan, Chaomin Shen, Faming Fang, and Guixu Zhang. Harmonizing knowledge transfer in neural network with unified distillation. In *ECCV*, pages 58–74. Springer, 2024. 2

[35] Audun Jøsang. *Subjective logic*. Springer, 2016. 2, 3

[36] Jungeun Kim, Junwon You, Dongjin Lee, Ha Young Kim, and Jae-Hun Jung. Do topological characteristics help in knowledge distillation? In *ICML*, 2024. 6, 7

[37] Anoop Korattikara Balan, Vivek Rathod, Kevin P Murphy, and Max Welling. Bayesian dark knowledge. *NIPS*, 28, 2015. 3

[38] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[39] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *NIPS*, 30, 2017. 2

[40] Changbin Li, Kangshuo Li, Yuzhe Ou, Lance M Kaplan, Audun Jøsang, Jin-Hee Cho, Dong Hyun Jeong, and Feng Chen. Hyper evidential deep learning to quantify composite classification uncertainty. *arXiv preprint arXiv:2404.10980*, 2024. 3

[41] Kehan Li, Runyi Yu, Zhennan Wang, Li Yuan, Guoli Song, and Jie Chen. Locality guidance for improving vision transformers on tiny datasets. In *ECCV*, pages 110–127. Springer, 2022. 8

[42] Lujun Li, Peijie Dong, Zimian Wei, and Ya Yang. Automated knowledge distillation via monte carlo tree search. In *CVPR*, pages 17413–17424, 2023. 8

[43] Xue Li, Wei Shen, and Denis Charles. Tedl: A two-stage evidential deep learning method for classification uncertainty quantification. *arXiv preprint arXiv:2209.05522*, 2022. 3

[44] Zheng Li, Ying Huang, Defang Chen, Tianren Luo, Ning Cai, and Zhigeng Pan. Online knowledge distillation via multi-branch diversity enhancement. In *Proceedings of the Asian conference on computer vision*, 2020. 2

[45] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *AAAI*, pages 1504–1512, 2023. 2, 6, 7

[46] Jakob Lindqvist, Amanda Olmin, Fredrik Lindsten, and Lennart Svensson. A general framework for ensemble distribution distillation. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2020. 3

[47] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *ICML*, pages 6448–6458. PMLR, 2020. 2

[48] Shicai Wei Chunbo Luo Yang Luo. Scale decoupled distillation. *arXiv preprint arXiv:2403.13512*, 2024. 6, 7

[49] Huan Ma, Zongbo Han, Changqing Zhang, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. *NIPS*, 34:6881–6893, 2021. 3

[50] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019. 3, 4

[51] Zelda E Mariet, Rodolphe Jenatton, Florian Wenzel, and Dustin Tran. Distilling ensembles improves uncertainty estimates. In *Third symposium on advances in approximate bayesian inference*, 2021. 3

[52] Nis Meinert and Alexander Lavin. Multivariate deep evidential regression. *arXiv preprint arXiv:2104.06135*, 2021. 3

[53] Zhong Meng, Jinyu Li, Yong Zhao, and Yifan Gong. Conditional teacher-student learning. In *ICASSP*, pages 6445–6449. IEEE, 2019. 2

[54] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, pages 5191–5198, 2020. 2

[55] Deep Shankar Pandey and Qi Yu. Learn to accumulate evidence from all training samples: theory and practice. In *ICML*, pages 26963–26989. PMLR, 2023. 3

[56] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pages 3967–3976, 2019. 2, 6, 7, 8

[57] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *CVPR*, pages 5007–5016, 2019. 2

[58] Bo Peng, Zhen Fang, Guangquan Zhang, and Jie Lu. Knowledge distillation with auxiliary variable. In *ICML*, 2024. 2

[59] 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. 6

[60] 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. *IJCV*, 115:211–252, 2015. 5, 6

[61] Max Ryabinin, Andrey Malinin, and Mark Gales. Scaling ensemble distribution distillation to many classes with proxy targets. *NIPS*, 34:6023–6035, 2021. 3

[62] Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. From words to watts: Benchmarking the energy costs of large language model inference. In *HPEC*, pages 1–9. IEEE, 2023. 1

[63] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 6, 7

[64] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *NIPS*, 31, 2018. 2, 3

[65] Glenn Shafer. Dempster-shafer theory. *Encyclopedia of artificial intelligence*, 1:330–331, 1992. 3

[66] Yichen Shen, Zhilu Zhang, Mert R Sabuncu, and Lin Sun. Real-time uncertainty estimation in computer vision via uncertainty-aware distribution distillation. In *WACV*, pages 707–716, 2021. 3

[67] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[68] Che Sun, Yunde Jia, and Yuwei Wu. Evidential reasoning for video anomaly detection. In *ACMMM*, pages 2106–2114, 2022. 3

[69] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *CVPR*, pages 15731–15740, 2024. 2, 5, 6, 7

[70] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 2, 6, 7

[71] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357. PMLR, 2021. 8

[72] Linh Tran, Bastiaan S Veeling, Kevin Roth, Jakub Swiatkowski, Joshua V Dillon, Jasper Snoek, Stephan Mandt, Tim Salimans, Sebastian Nowozin, and Rodolphe Jenatton. Hydra: Preserving ensemble diversity for model distillation. *arXiv preprint arXiv:2001.04694*, 2020. 3

[73] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *CVPR*, pages 1365–1374, 2019. 2

[74] Meet Vadera, Brian Jalaian, and Benjamin Marlin. Generalized bayesian posterior expectation distillation for deep neural networks. In *Conference on Uncertainty in Artificial Intelligence*, pages 719–728. PMLR, 2020. 3

[75] Chen Wang, Xiang Wang, Jiawei Zhang, Liang Zhang, Xiao Bai, Xin Ning, Jun Zhou, and Edwin Hancock. Uncertainty estimation for stereo matching based on evidential deep learning. *pattern recognition*, 124:108498, 2022. 3

[76] Hanjing Wang and Qiang Ji. Beyond dirichlet-based models: when bayesian neural networks meet evidential deep learning. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024. 3

[77] Kuan-Chieh Wang, Paul Vicol, James Lucas, Li Gu, Roger Grosse, and Richard Zemel. Adversarial distillation of bayesian neural network posteriors. In *ICML*, pages 5190–5199. PMLR, 2018. 3

[78] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *CVPR*, pages 568–578, 2021. 8

[79] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 8

[80] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *ECCV*, pages 588–604. Springer, 2020. 2

[81] Ronald R Yager and Liping Liu. *Classic works of the Dempster-Shafer theory of belief functions*. Springer, 2008. 2

[82] Jing Yang, Brais Martinez, Adrian Bulat, Georgios Tzimiropoulos, et al. Knowledge distillation via softmax regression representation learning. ICLR (ICLR), 2021. 2

[83] Sergey Zagoruyko. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 6

[84] Feng Zhang, Xiatian Zhu, and Mao Ye. Fast human pose estimation. In *CVPR*, pages 3517–3526, 2019. 2

[85] 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. 2

[86] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, pages 6848–6856, 2018. 6

[87] Yuan Zhang, Weihua Chen, Yichen Lu, Tao Huang, Xiuyu Sun, and Jian Cao. Avatar knowledge distillation: self-ensemble teacher paradigm with uncertainty. In *ACMMM*, pages 5272–5280, 2023. 3

[88] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, pages 11953–11962, 2022. 1, 2, 5, 6, 7

[89] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *CVPR*, pages 16965–16974, 2024. 1