REDistill: Robust Estimator Distillation for Balancing Robustness and Efficiency

Ondrej Tybl[1] and Lukas Neumann[1]

CMP Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague
{ondrej.tybl,lukas.neumann}@cvut.cz

Abstract. Knowledge Distillation (KD) transfers knowledge from a large teacher model to a smaller student by aligning their predictive distributions. However, conventional KD formulations - typically based on Kullback–Leibler divergence - assume that the teacher provides reliable targets. In practice, teacher predictions are often noisy or overconfident, and existing correction-based approaches rely on ad-hoc heuristics and extensive hyper-parameter tuning, which hinders generalization. We introduce REDistill (Robust Estimator Distillation), a simple yet principled framework grounded in robust statistics. REDistill replaces the standard KD objective with a power divergence loss, a generalization of KL divergence that adaptively downweights unreliable teacher output. This formulation provides a unified and interpretable treatment of teacher noise, requires only logits, integrates seamlessly into existing KD pipelines, and incurs negligible computational overhead. Extensive experiments on CIFAR-100 and ImageNet-1k demonstrate that REDistill consistently improves student accuracy in diverse teacher-student architectures. Remarkably, it achieves these gains without model-specific hyper-parameter tuning, underscoring its robustness and strong generalization to unseen teacher–student pairs.

Keywords: Knowledge Distillation · Robustness · Power Divergence

1 Introduction

Rapid development of deep neural networks has been coupled with a substantial increase in model size and therefore computational cost, posing challenges for efficient training and deployment of such models. Although larger models generally exhibit higher accuracy, considerable research effort has been devoted to reduce model complexity without sacrificing accuracy. Knowledge Distillation (KD) has recently gained significant research attention [8, 28, 50, 51], because compared to other strategies such as pruning or quantization, knowledge distillation has fewer architectural constraints and therefore offers greater flexibility and wider applicability.

In Knowledge Distillation, the knowledge accumulated in a pre-trained high-capacity *teacher* model is transferred to a more efficient, smaller *student* network.

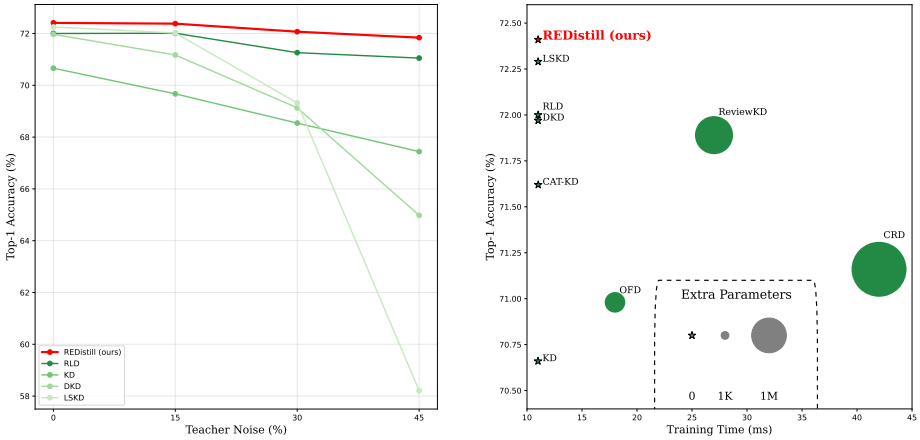

Fig. 1: Our method achieves a strong balance between robustness and efficiency, consistently delivering state-of-the-art performance even when the teacher's predictions are noisy, without introducing additional computational overhead. As teacher network noise increases (i.e., more teacher mistakes), our method's performance degrades the least (left). Since our method relies solely on the teacher's logits and introduces no additional trainable parameters, the training time slowdown is negligible compared to other methods (right). Results reported on CIFAR-100 for the ResNet56/ResNet20 student/teacher pair.

The main challenge, however, is that even the *teacher's* predictions are not perfect, in other words, the **teacher is noisy**, and as a result, blindly mimicking teacher outputs only leads to error accumulation and to a significant degradation of the *student* model accuracy.

Existing correction-based distillation methods [8, 27, 56] adjust the teacher's logits using ground-truth information to address this issue. Specifically, they either swap the logits between the predicted top class and the ground-truth class (the swap operation) [56], or amplify the probability assigned to the ground-truth class (the augment operation) [8, 27] or more recently, [51] masks certain classes during distillation. However, these ad-hoc modifications and heuristics distort the underlying correlations among classes and their effectiveness is highly sensitive to hyper-parameter tuning, whose values need to be *empirically determined for each teacher/student pair* and for *each dataset*.

In this paper, we introduce a robust principled framework based on the power divergence family from robust statistics. By replacing the standard KD objective with its power divergence counterpart, we adaptively down-weight unreliable teacher outputs while preserving essential logit relationships. As a result, our method tackles mistakes of the noisy teacher network in a unified and an interpretable manner. Thanks to its fundamental grounding, the method is less sensitive to specific hyper-parameter values; even more so, we are able to determine the most important hyper-parameter value theoretically (and we empiri-

cally show that this choice is indeed optimal). Our method, without bells and whistles, not only outperforms previous KD methods, but it does so **without model-specific hyper-parameter tuning for each teacher/student pair**, which makes our method more universally applicable.

Our main contributions are summarized as follows:

- We propose a new principled method for Knowledge Distillation, which compensates for unreliable teacher predictions. The method replaces the standard KD objective with its power divergence counterpart, which adaptively down-weights unreliable teacher outputs while preserving essential logit relationships.
- The method does not rely on model-specific hyper-parameter tuning and heuristics, which makes it more generalizable. Experiments in the model-agnostic setting suggest that the method will also likely work on unseen teacher/student pairs, because without a single modification and using hyper-parameter value which was first obtained theoretically, the method outperformed previous methods on 14 different teacher/student pairs.

The remainder of the paper is organized as follows. In Section 2, we review related work on knowledge distillation and robust loss design. Section 3 introduces the necessary background and notation, and Section 4 presents the proposed method in detail. Experimental results are provided in Section 5, followed by conclusions in Section 6.

2 Related Work

Knowledge Distillation (KD) was first introduced by Hinton et al. [20], who proposed transferring knowledge from a high-capacity *teacher* network to a smaller *student* by minimizing the Kullback–Leibler (KL) divergence between their softened logit distributions. Zhao et al. [67] later refined this formulation by decoupling target and non-target logits, allowing the student to learn class confidence and inter-class relationships more effectively.

Subsequent KD research can be broadly categorized into four families: (1) Logit-based distillation, which aligns the teacher's output distribution with the student's predictions [8,9,20,24,27,28,37,41,49–51,55,56,63,66,67]; (2) Feature-based distillation, where intermediate representations or attention maps are transfered [1, 12, 16, 18, 23, 29, 30, 32, 53]; (3) Relation-based distillation, which captures structural dependencies between samples or classes [33, 36, 39, 40, 64]; and (4) Multi-teacher distillation, where the student aggregates knowledge from multiple teachers [48,58,62]. Feature- and relation-based methods often face challenges due to architectural heterogeneity and the complexity of aligning feature spaces, while multi-teacher approaches increase computational cost and parameterization. Consequently, *logit-based distillation* remains the most practical and widely used paradigm because of its simplicity, generality, and effectiveness.

Within the logit-based family, recent work has focused on refining how the student interacts with the teacher's softened outputs. Temperature-based methods modify the distillation temperature to control confidence: Li et al. [28] adapt

temperature adversarially in a sample-wise manner, Sun et al. [50] dynamically optimize per-sample temperatures, and Jin et al. [24] exploit multi-level logits for richer supervision. Other approaches attempt to mitigate noisy or overconfident teacher predictions through heuristic corrections [8, 27, 51, 56]. Despite their empirical gains, these strategies rely heavily on ad-hoc adjustments and hyper-parameter tuning tailored to specific models or datasets. None provide a principled mechanism grounded in statistical theory to handle unreliable teacher outputs.

An orthogonal line of research in *robust statistics* seeks estimators that remain reliable under noisy or contaminated data [22]. Rather than correcting logits heuristically, robustness can be achieved by modifying the underlying loss function so that outliers have a diminished influence. Classical examples include M-estimators [22], generalized divergence functions [4], and the *power divergence* family [11, 14, 15, 35, 42], which generalizes the KL divergence and introduces a tunable robustness parameter to balance efficiency and noise tolerance. Such divergences have proven effective in diverse settings, including robust variational autoencoders [2], outlier-resistant regression [19], stabilized GAN training [7], and language model optimization [44]. However, their integration into deep learning frameworks remains limited and in KD completely unexplored.

Our power divergence is closely related to other robust loss functions, but differs in what it makes robust. The Huber loss and our power divergence both belong to the family of *M-estimators* [22], which attain robustness by bounding the influence of outliers on the estimate; the Huber loss is the canonical robust M-estimator for *location* estimation, whereas the power divergence plays the analogous role for *distributional* estimation, which is precisely the regime of knowledge distillation. The focal loss [31] also modifies cross-entropy, but reweights samples by the *model's* prediction confidence $(1-p_t)^\gamma$, focusing training on hard examples; as its authors note, this "performs the opposite role of a robust loss". In contrast, the power divergence reweights by the *likelihood ratio* p/q_θ, bounding the influence of samples where teacher and student disagree strongly and thereby targeting robustness to teacher errors directly. The most closely related KD method is ABKD [55], which is motivated by a similar robustness–efficiency trade-off and also draws on a power-divergence family. However, ABKD selects its loss configuration by grid search, yielding dataset- and architecture-specific choices, whereas we derive a single robustness parameter ($\lambda = 2/3$) from statistical theory and keep it fixed across all settings (see Section 4).

Our work bridges knowledge distillation and robust statistics. To the best of our knowledge, this is the first work to provide a unified and statistically grounded framework for *robust knowledge distillation*.

3 Background

We start by recalling the standard notation and concepts of knowledge distillation.

Suppose that we have a student model f_S with a parameter vector θ and our aim is to train it for a K-fold classification. In the vanilla setting, we minimize the average *Kullback–Leibler (KL)* divergence loss between the student prediction and one-hot encoded target

$$\mathcal{L}_{\text{vanilla}}(\theta, x, y) \coloneqq \text{KL}(y, q_\theta(\cdot|x)) = \sum_{k=1}^{K} y_k \log \frac{y_k}{q_\theta(k \,|\, x)} \tag{1}$$

over the training dataset consisting of labeled images $(x, y) \in \mathbb{R}^{3\times H\times W} \otimes \mathbb{R}^K$, where

$$q_\theta(k \,|\, x) = \frac{\exp\left(f_S(x, \theta)_k\right)}{\sum_{d=1}^{K} \exp\left(f_S(x, \theta)_d\right)}, \quad k = 1, \ldots, K \tag{2}$$

are the student output probabilities.

In knowledge distillation ([20]), we work with an enhanced information: a trained teacher model f_T with output probabilities

$$p(k \,|\, x) = \frac{\exp\left(f_T(x)_k\right)}{\sum_{d=1}^{K} \exp\left(f_T(x)_d\right)}, \quad k = 1, \ldots, K \tag{3}$$

is given and we train the student by also matching its output to the teacher using a loss in the form of a linear combination

$$\mathcal{L}_{\text{KD}}\left(\theta, x, y\right) \coloneqq c_1 \,\text{KL}\left(y, q_\theta\left(\cdot|x\right)\right) + c_2 \,\text{KL}\left(p\left(\cdot|x\right), q_\theta\left(\cdot|x\right)\right) \tag{4}$$

for hyper-parameters $c_1, c_2 > 0$ that balance the influence of ground-truth and teacher predictions. Various modifications to Equation (4) have been introduced to further improve the speed and efficiency of knowledge distillation; one of the examples is the Decoupled Knowledge Distillation ([67]) where they treat the probabilities corresponding to the ground-truth class separately. As a result in practice, we replace $\text{KL}(p(\cdot|x), q_\theta(\cdot|x))$ from Equation (4) with the following formulation

$$\alpha \,\text{KL}(p_{\text{target}}(\cdot|x), q_{\theta,\text{target}}(\cdot|x)) + \beta \,\text{KL}(p_{\text{nontarget}}(\cdot|x), q_{\theta,\text{nontarget}}(\cdot|x)), \tag{5}$$

where α, β balance the influence of target and non-target classes.

4 Method

Knowledge distillation typically relies on minimizing the Kullback–Leibler (KL) divergence between teacher and student predictions. Although effective in noise-free regimes (corresponding to a perfect teacher), KL-based training is known to be highly sensitive to distributional misspecification and outliers—issues that arise naturally when the teacher itself is imperfect or overconfident [4]. In this work, we revisit the distillation objective from the perspective of robust statis-

tical estimation. By formulating distillation as a problem of estimating a target distribution under model uncertainty, we connect recent insights from robust divergence estimation in statistics to deep learning objectives. Through this analysis, we identify a principled replacement for the KL divergence: the power divergence, a family that balances efficiency and robustness in a theoretically grounded way.

4.1 Power-Divergence

The KL divergence between a data distribution p and a model distribution q_θ can be written as the expected excess "surprise" of using q_θ to approximate p where "surprise" is measured by the logarithm of the likelihood ratio p/q_θ:

$$\mathrm{KL}(p, q_\theta) = \sum_{k=1}^{K} p_k \log \frac{p_k}{q_{\theta,k}} = \mathbb{E}_p\left[\log\left(\frac{p}{q_\theta}\right)\right]. \tag{6}$$

Minimizing this divergence corresponds to maximum likelihood estimation, which yields estimators that are unbiased and efficient [13]. However, when the data or teacher predictions are contaminated by noise and/or mistakes, such estimators become unstable.

This motivates the search for *robust alternatives* – estimators that maintain desirable statistical properties even under model misspecification [22]. We do so by modifying the logarithmic transformation in Equation (6). Ideally, this generalized divergences reduces the influence of large likelihood ratios p/q_θ in presence of mistakes in p, thereby limiting sensitivity to outliers [4, 6, 14, 38, 42].

First, we define γ-logarithm (also relaxed logarithm or equivalently, the Box–Cox transform of order $1-\gamma$ [5, 14, 57]) as

$$\log_\gamma(x) \coloneqq \begin{cases} \log(x), & \text{if } \gamma = 1, \\ \dfrac{x^{1-\gamma} - 1}{1-\gamma}, & \text{if } \gamma \neq 1, \\ \text{undefined}, & \text{if } x \leq 0. \end{cases} \tag{7}$$

We have that γ-logarithm continuously approximates the natural logarithm as $\gamma \to 1$ and the tunable parameter γ we can control the convexity properties, see Figure 2. Substituting the $(1-\lambda)$-logarithm for the natural logarithm in Equation (6) then yields the *power divergence* of order λ

$$\begin{aligned} \mathrm{D}_\lambda(p, q_\theta) &\coloneqq \frac{1}{1+\lambda}\,\mathbb{E}_p\left[\log_{1-\lambda}\left(\frac{p}{q_\theta}\right)\right] \\ &= \frac{1}{\lambda(\lambda+1)} \sum_{k=1}^{K} p(k \mid x)\left[\left(\frac{p(k \mid x)}{q_\theta(k \mid x)}\right)^{\lambda} - 1\right]. \end{aligned} \tag{8}$$

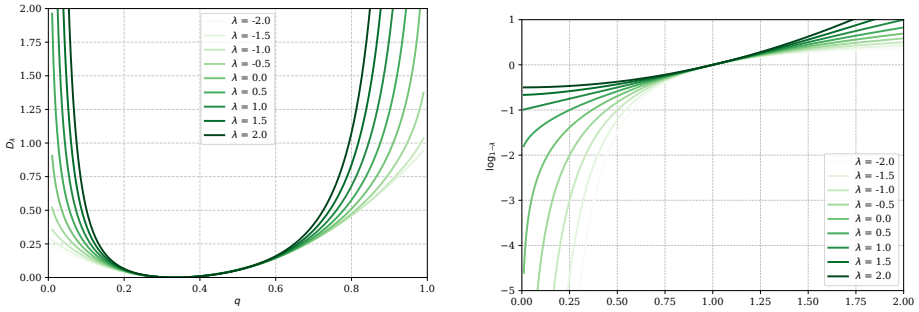

Fig. 2: The divergence D_λ corresponds to the KL divergence for $\lambda = 0$. For other values, logarithm as a measure of surprise in the divergence computation is replaced by its smooth relaxation (known as $(1-\lambda)$ -logarithm, see Eq. (7)). (a) shows graph of $(1-\lambda)$ -logarithm, (b) shows $\mathrm{D}_\lambda(P\|Q)$ between $P = (1/3, 2/3)$ and $Q = (q, 1-q)$ as a function of q for different λ values.

This defines a continuous family of *power divergence* [3, 14] that includes the KL divergence as a special case $\lambda = 0$, i.e., $\mathrm{D}_0 = \mathrm{KL}$. The parameter λ governs the trade-off between statistical efficiency and robustness: smaller values emphasize accuracy under correct model specification, while larger values attenuate the impact of outliers. The divergence is convex, nonnegative, and vanishes if and only if $p = q_\theta$. Figure 2 visualizes the behavior of D_λ as λ varies.

4.2 Robust Estimator Distillation (REDistill)

We formulate a robust distillation objective that replaces the KL divergence in the standard formulation with the power divergence of order $\lambda = 2/3$; the choice of λ based on the balance between efficiency and robustness is addressed below. The resulting loss, which we call **Robust Estimator Distillation (REDistill)**, is defined as

$$\begin{aligned}\mathcal{L}_{\mathrm{REDistill}}(\theta, x, y) \coloneqq \mathrm{KL}(y,\, q_\theta(\cdot \mid x)) \\ + \alpha\, \mathrm{D}_{2/3}(p_{\mathrm{target}}(\cdot \mid x)\,,\, q_{\theta,\mathrm{target}}(\cdot \mid x)) \\ + \beta\, \mathrm{D}_{2/3}(p_{\mathrm{nontarget}}(\cdot \mid x)\,,\, q_{\theta,\mathrm{nontarget}}(\cdot \mid x))\,.\end{aligned} \tag{9}$$

This formulation inherits the efficiency of KL-based distillation but provides improved robustness to noisy or miscalibrated teacher predictions [51].

To justify our choice of λ, we first formally characterize robustness, we analyze the sensitivity of the estimator induced by D_λ using the *influence function* [22,25] from statistical literature. Given a loss function $\mathcal{L}$ and a training set $\{(x_n, y_n)\}_{n=1}^N$, parameters are estimated as:

$$\hat{\theta} \coloneqq \arg\min_\theta \sum_{n=1}^{N} \mathcal{L}(\theta, x_n, y_n). \tag{10}$$

When an outlier (x, y) with small weight $\epsilon > 0$ is added, the estimator becomes:

$$\hat{\theta}_{\epsilon,(x,y)} \coloneqq \arg\min_{\theta} \sum_{n=1}^{N} \mathcal{L}(\theta, x_n, y_n) + \epsilon \mathcal{L}(\theta, x, y). \tag{11}$$

The influence function quantifies the resulting perturbation:

$$\mathrm{IF}(\mathcal{L}, (x, y)) \coloneqq \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \big(\hat{\theta}_{\epsilon,(x,y)} - \hat{\theta}\big), \tag{12}$$

and robustness can be measured by

$$\sup_{(x,y)} \| \mathrm{IF}(\mathcal{L}, (x, y))\|, \tag{13}$$

where a smaller value implies greater resistance to outliers.

Table 1: Student network top-1 accuracy (%) on the CIFAR-100 validation set, using the *model-agnostic* hyper-parameters protocol (see Sec. 5.1). The reported results are averages of four trials.

teacher	ResNet32	ResNet32	ResNet32	WRN-40-2	WRN-40-2	VGG13	ResNet50
	79.42	79.42	79.42	75.61	75.61	74.64	79.34
student	SHN-V2	WRN-16-2	WRN-40-2	ResNet8	MN-V2	MN-V2	MN-V2
	71.82	73.26	75.61	72.50	64.60	64.60	64.60
KD [20]	75.51	72.93	77.15	75.03	64.66	64.23	64.40
DKD [67]	77.01	76.18	78.81	75.50	69.27	69.70	70.60
LSKD [50]	77.11	76.25	78.86	76.55	69.57	69.82	70.61
RLD [51]	76.74	75.97	78.68	75.12	68.91	69.72	70.22
REDistill (ours)	**77.52**	**76.31**	**79.01**	**77.21**	**69.98**	**70.12**	**70.92**
teacher	ResNet32	VGG13	WRN-40-2	WRN-40-2	ResNet56	ResNet110	ResNet110
	79.42	74.64	75.61	75.61	72.34	74.31	74.31
student	ResNet8	VGG8	WRN-40-1	WRN-16-2	ResNet20	ResNet32	ResNet20
	72.50	70.36	71.98	73.26	69.06	71.14	69.06
KD [20]	73.94	70.69	71.40	73.83	71.01	72.17	69.95
DKD [67]	76.17	74.58	74.57	75.56	71.71	73.53	71.64
LSKD [50]	76.67	74.61	74.88	75.55	71.51	73.94	**71.67**
RLD [51]	76.12	74.40	74.69	75.62	70.85	73.09	70.18
REDistill (ours)	**77.05**	**74.78**	**74.92**	**76.82**	**71.95**	**74.18**	71.60

When training via minimization of D_λ, the per-sample loss is

$$\mathcal{L}_\lambda(\theta, x, y) = \mathrm{D}_\lambda \left(y, q_\theta(\cdot|x)\right), \tag{14}$$

and the corresponding influence function evaluates to (see Supplementary material for details)

$$\mathrm{IF}(\mathcal{L}_\lambda, (x, y)) = \frac{1}{1+\lambda}\, q_\theta\left(\cdot|x\right). \tag{15}$$

Equation (15) demonstrates that larger positive λ values reduce the influence of individual samples, thereby enhancing robustness.

Following Equation (15) and the analysis of [42], estimators derived from D_λ remain asymptotically unbiased, with a tunable trade-off: 1) larger λ increases robustness to outliers; 2) smaller $|\lambda|$ improves statistical efficiency. Following theoretical studies [14] we identify $\lambda = 2/3$ as a near-optimal balance between these competing objectives (see Supplementary material), which we adopt in *all our experiments*.

Unlike previous regularization methods which are mostly heuristic, the robustness in REDistill arises directly from a principled reformulation grounded in robust statistical estimation, without introducing additional hyper-parameters or computational overhead.

Compatibility with Temperature Scaling. In many practical applications, it is desirable to pre-process model logits by a temperature scaling, that is, we want to soften (possibly) both the teacher p and model logits q_θ by a temperature $\tau > 0$ so that it becomes

$$p^\tau(k \,|\, x) = \frac{\exp\left(f_T(x)_k/\tau\right)}{\sum_{d=1}^{K} \exp\left(f_T(x)_d/\tau\right)}, \quad q^\tau(k \,|\, x) = \frac{\exp\left(f_S(x)_k/\tau\right)}{\sum_{d=1}^{K} \exp\left(f_S(x)_d/\tau\right)} \tag{16}$$

However, even at moderate temperatures τ, softening the logits leads to a significant impact on the magnitude of the loss gradient [20] and therefore slows down the training. We propose that in cases when temperature is used, the power-divergence is scaled with a factor τ^2, i.e. it becomes

$$\tau^2 \, \mathrm{D}_{2/3}\left(p^\tau, q^\tau\right). \tag{17}$$

See Supplementary material for details.

5 Experiments

5.1 Experimental Setup

In line with previous work, we conduct experiments on two standard image classification benchmarks: CIFAR-100 [26] and ImageNet-1k [45], with input resolutions of 32×32 and 224×224, respectively. The teacher–student model pairs include variants of ResNet [17], WideResNet (WRN) [61], VGG [47], ShuffleNet (SHN-V2) [34, 65] and MobileNetV2 (MN-V2) [21, 46].

All the compared KD losses for different methods depend on various hyper-parameter settings (i.e. logit temperature and c_1, c_2 in Equation (4) for KD [20], α, β in Equation (5) in DKD [67]). First, we compare all methods in **model-agnostic evaluation protocol**, where all *hyper-parameter values are fixed for a given method* for all evaluated models. We argue that this setting is more realistic, because it is impractical to assume that one has to find optimal hyper-parameter setting for each teacher-student pair on every dataset. It also makes the comparison between methods more fair, because there is no possibility of hyper-parameter tuning on the validation set to artificially boost accuracy of

given method. In contrast, a model-specific protocol is useful when the goal is to determine the best achievable performance for each method.

Table 2: Student network top-1 accuracy (%) on the CIFAR-100 validation set, using the *model-specific* hyper-parameters protocol (see Sec. 5.1). The reported results are averages of four trials.

teacher	ResNet32	ResNet32	ResNet32	WRN-40-2	WRN-40-2	VGG13	ResNet50
	79.42	79.42	79.42	75.61	75.61	74.64	79.34
student	SHN-V2	WRN-16-2	WRN-40-2	ResNet8	MN-V2	MN-V2	MN-V2
	71.82	73.26	75.61	72.50	64.60	64.60	64.60
KD [20]	74.45	74.90	77.70	73.97	68.36	67.37	67.35
CTKD [28]	75.37	74.57	77.66	74.61	68.34	68.50	68.67
DKD [67]	77.07	75.70	78.46	75.56	69.28	69.71	70.35
LSKD [50]	77.37	76.19	*78.95*	*76.75*	*70.01*	*69.98*	70.45
LA [56]	75.14	–	77.39	73.88	68.57	68.09	68.85
RC [8]	75.61	–	77.58	75.22	68.72	68.66	68.98
LR [27]	76.27	–	78.73	75.26	69.02	69.78	70.38
RLD [51]†	*77.56*	–	78.91	76.12	69.75	69.97	*70.76*
EA-KD [49]	75.91	–	–	–	–	69.17	69.67
REDistill (ours)	**77.84**	**76.37**	**79.23**	**77.46**	**70.48**	**70.58**	**71.03**

teacher	ResNet32	VGG13	WRN-40-2	WRN-40-2	ResNet56	ResNet110	ResNet110
	79.42	74.64	75.61	75.61	72.34	74.31	74.31
student	ResNet8	VGG8	WRN-40-1	WRN-16-2	ResNet20	ResNet32	ResNet20
	72.50	70.36	71.98	73.26	69.06	71.14	69.06
KD [20]	73.33	72.98	73.54	74.92	70.66	73.08	70.67
CTKD [28]	73.39	73.52	73.93	75.45	71.19	73.52	70.99
DKD [67]	76.32	74.68	74.81	76.24	71.97	74.11	71.06
LSKD [50]	*77.01*	74.81	*74.89*	*76.39*	*72.32*	*74.29*	*71.85*
LA [56]	73.46	73.51	73.75	–	71.24	73.39	70.86
RC [8]	74.68	73.37	74.07	–	71.63	73.44	71.41
LR [27]	76.06	74.66	74.42	–	70.74	73.52	70.61
RLD [51]	76.64	**74.93**	74.88	–	72.00	74.02	71.67
ABKD [55]	75.62	74.22	74.41	76.11	71.79	74.14	71.72
EA-KD [49]	75.46	74.08	74.38	–	–	–	–
REDistill (ours)	**77.31**	*74.88*	**74.97**	**76.99**	**72.41**	**74.40**	**71.89**

Second, in line with previous work [50,51,67], we also include **model-specific evaluation protocol**, where each teacher-student pair has its own unique hyper-parameter values. While model-specific hyper-parameters yield slightly better results and are important for assessment of the best possible student accuracy, training for a single teacher-student pair means training for each possible hyper-parameter combination separately, which leads to extensive computation demands as the optimal hyper-parameters are unknown and need to be searched for.

Our comparisons focus on logit-based knowledge distillation approaches KD [20], CTKD [28], DKD [67], LSKD [50], LA [56], RC [8], LR [27], RLD [51], ED-KD [49] and ABKD [55]. For feature-based methods, we include FitNet [43], AT [60], RKD [39], CRD [53], OFD [18], ReviewKD [12], SimKD [10], CAT-KD [16]. These methods are compared separately from logit-based methods as they are more computationally demanding [10, 51] and less generalizable across heterogeneous architectures [54].

All results were obtained following an extensive review of prior work and publicly available source codes to ensure fair and consistent comparisons. Fur-

Table 3: Student network accuracy (%) on the ImageNet-1k validation set. The reported results are averages of three trials.

	Res34/Res18		Res50/MN-V1	
	Top-1	Top-5	Top-1	Top-5
Teacher	73.31	91.42	76.16	92.86
Student	69.75	89.07	68.87	88.76
KD [20]	71.03	90.05	70.50	89.80
CTKD [28]	71.38	90.27	71.16	90.11
DKD [67]	71.70	90.41	72.05	91.05
LSKD [50]	71.88	90.58	72.85	91.23
LA [56]	71.17	90.16	70.98	90.13
RC [8]	71.59	90.21	71.86	90.54
LR [50]	70.29	89.98	71.76	90.93
RLD [51]	71.91	90.59	72.75	91.18
REDistill (ours)	**72.00**	**90.71**	**72.98**	**91.48**

thermore, we note that AID and SKD [41, 59] are not included (same as [51]), as both involve teacher fine-tuning either before distillation (AID) or during distillation through an attention head (SKD), making their teacher performance incomparable with other methods. For *model-agnostic evaluation*, we reproduce all results using the authors' official code, using the default/best reported hyperparameters from official implementations. For our method, we used $\lambda = 2/3$ and used the α and β values from DKD [50]. For *model-specific evaluation*, the results are taken directly as reported in [50, 51].

In both cases, we follow the same training setup as in [50,51,67]. For CIFAR-100, we train with SGD [52] for 240 epochs with a batch size of 64, weight decay of 5×10^{-4}, and momentum 0.9. The initial learning rate is 0.01 for SHN-V2 and MN-V2 and 0.05 for the other architectures; it is reduced by a factor of 10 at epochs 150, 180, and 210. For ImageNet, we train with SGD [52] using a batch size of 512, 100 epochs, weight decay 1×10^{-4}, momentum 0.9, and initial learning rate 0.2, decayed by 10 every 30 epochs. Unlike [51], we do not use any data augmentation. All experiments were conducted on single Tesla V100-SXM2-32GB GPU; one training on CIFAR-100 required up to 2 hours, while ImageNet training took $\sim$ 2 days.

5.2 Results

For CIFAR-100 and ImageNet-1k, we present results for both *model-agnostic* (Table 1) and *model-specific* (Tables 2 and 3) protocols. *Our method consistently surpasses all previous knowledge distillation methods*, with the exception of the VGG13/VGG8 pair where it places second. On ImageNet-1k (Table 3), our method achieves **state-of-the-art accuracy** with a substantial margin over the best existing baselines.

These results confirm our central hypothesis: incorporating a robust training loss into knowledge distillation leads to stronger student models, and the gains

Table 4: Combining REDistill with existing methods. Student network top-1 accuracy (%) on the CIFAR-100 validation set.

teacher	ResNet32	ResNet32	ResNet32	WRN-40-2	WRN-40-2	VGG13	ResNet50
	79.42	79.42	79.42	75.61	75.61	74.64	79.34
student	SHN-V2	WRN-16-2	WRN-40-2	ResNet8	MN-V2	MN-V2	MN-V2
	71.82	73.26	75.61	72.50	64.60	64.60	64.60
KD [20]	74.45	74.90	77.70	73.97	68.36	67.37	67.35
KD+REDistill	**74.99**	**75.62**	**77.85**	**74.20**	**68.69**	**67.52**	**68.02**
DKD [67]	77.07	75.70	78.46	75.56	69.28	69.71	70.35
DKD+REDistill	**77.84**	**76.37**	**79.23**	**77.46**	**70.48**	**70.58**	**71.03**
MLKD [24]	78.44	76.52	79.26	77.33	70.78	70.57	71.04
MLKD+REDistill	**78.82**	**77.50**	**79.60**	**77.63**	**71.63**	**70.95**	**71.20**
teacher	ResNet32	VGG13	WRN-40-2	WRN-40-2	ResNet56	ResNet110	ResNet110
	79.42	74.64	75.61	75.61	72.34	74.31	74.31
student	ResNet8	VGG8	WRN-40-1	WRN-16-2	ResNet20	ResNet32	ResNet20
	72.50	70.36	71.98	73.26	69.06	71.14	69.06
KD [20]	73.33	72.98	73.54	74.92	70.66	73.08	70.67
KD+REDistill	**74.12**	**73.41**	**74.62**	**75.25**	**71.01**	**73.25**	**70.89**
DKD [67]	76.32	74.68	74.81	76.24	71.97	74.11	71.06
DKD+REDistill	**77.31**	**74.88**	**74.97**	**76.99**	**72.41**	**74.40**	**71.89**
MLKD [24]	77.08	75.18	75.35	76.63	72.19	74.11	71.89
MLKD+REDistill	**78.31**	**75.24**	**75.53**	**76.91**	**72.40**	**74.29**	**72.23**

are not attributable to hyper-parameter tuning. Moreover, comparing results from the *model-agnostic* and *model-specific* settings reveals that several high-performing methods in Table 2 rely on extensive hyper-parameter tuning and/or data augmentation – without it, their performance drops significantly. For example, RLD underperforms even the standard KD baseline for ResNet56/ResNet20 and yields only marginal gains ($\sim 0.2\%$) for ResNet110/ResNet20 and WRN-40-2/ResNet8x4 – an order of magnitude smaller than under the *model-specific* protocol. Similarly, LSKD does not consistently outperform DKD. In contrast, our method maintains strong and stable improvements in both evaluation settings, demonstrating its robustness and generalization across diverse teacher–student architectures.

Combining methods. Next, we demonstrate the extensive applicability of our method by incorporating it into existing distillation pipelines within *model-specific* protocol to demonstrate the results improvements. We integrate it with three representative approaches – KD [20], DKD [67], and MLKD [24]. For fair comparison, we don't alter the previous author's training pipelines to avoid introducing bias due to suboptimal hyperparameter choices. Specifically, for MLKD and REDistill+MKLD we follow exactly the training pipeline of the authors [24], and this includes data augmentation and training for 480 epochs.

As seen in Table 4, adding our loss leads to further improvement of the student model in every case, indicating that the robustness introduced by our loss complements rather than competes with prior formulations.

Furthermore, we compare our best result achieved by combining our method with MLKD with feature-based distillation methods (see Table 5). Our method, despite being significantly simpler and requiring no additional feature matching or architectural modifications, *outperforms a whole family of different and more*

Table 5: Comparison with feature-based KD methods. Student network top-1 accuracy (%) on the CIFAR-100 validation set.

teacher	ResNet32	ResNet32	ResNet32	WRN-40-2	WRN-40-2	VGG13	ResNet50
	79.42	79.42	79.42	75.61	75.61	74.64	79.34
student	SHN-V2	WRN-16-2	WRN-40-2	ResNet8	MN-V2	MN-V2	MN-V2
	71.82	73.26	75.61	72.50	64.60	64.60	64.60
FitNet [43]	73.54	74.70	77.69	74.61	68.64	64.16	63.16
AT [60]	72.73	73.91	77.43	74.11	60.78	59.40	58.58
RKD [39]	73.21	74.86	77.82	75.26	69.27	64.52	64.43
CRD [53]	75.65	75.65	78.15	75.24	70.28	69.73	69.11
OFD [18]	76.82	76.17	79.25	74.36	69.92	69.48	69.04
ReviewKD [12]	77.78	76.11	78.96	74.34	71.28	70.37	69.89
SimKD [10]	78.39	77.17	79.29	75.29	70.10	69.44	69.97
CAT-KD [16]	78.41	76.97	78.59	75.38	70.24	69.13	**71.36**
REDistill+MLKD	**78.82**	**77.50**	**79.60**	**77.63**	**71.63**	**70.95**	71.20
teacher	ResNet32	VGG13	WRN-40-2	WRN-40-2	ResNet56	ResNet110	ResNet110
	79.42	74.64	75.61	75.61	72.34	74.31	74.31
student	ResNet8	VGG8	WRN-40-1	WRN-16-2	ResNet20	ResNet32	ResNet20
	72.50	70.36	71.98	73.26	69.06	71.14	69.06
FitNet [43]	73.50	71.02	72.24	73.58	69.21	71.06	68.99
AT [60]	73.44	71.43	72.77	74.08	70.55	72.31	70.65
RKD [39]	71.90	71.48	72.22	73.35	69.61	71.82	69.25
CRD [53]	75.51	73.94	74.14	75.48	71.16	73.48	71.46
OFD [18]	74.95	73.95	74.33	75.24	70.98	73.23	71.29
ReviewKD [12]	75.63	74.84	75.09	76.12	71.89	73.89	71.34
SimKD [10]	78.08	74.89	74.53	75.53	71.05	73.92	71.06
CAT-KD [16]	76.91	74.65	74.82	75.60	71.62	73.62	71.37
REDistill+MLKD	**78.31**	**75.24**	**75.53**	**76.91**	**72.40**	**74.29**	**72.23**

complex feature-based methods. This highlights that **our contribution is orthogonal to existing strategies**: it enhances them by providing a more stable and reliable training signal. We note that since feature-based methods rely on fundamentally different training pipelines and hyper-parameters, direct mapping of training parameters to logit-based methods is not possible.

5.3 Ablations

Robustness parameter. We investigate the impact of λ on the performance of our method (see Equation (8)). Recall that λ is not treated as a hyper-parameter as it was fixed at $2/3$ based on theoretical evidence. To also support this choice empirically, we evaluate top-1 accuracy on the CIFAR-100 and ImageNet-1k validation set for various values of λ, ranging from 0 to 2, in Table 6. The results show that values around $\lambda = 2/3$ yield the highest top-1 accuracy. This supports our theoretical argument that $\lambda = 2/3$ provides an optimal balance between the teacher's and student's knowledge, resulting in a more robust distillation process. Furthermore, the choice of λ appears somewhat flexible, with other values in this region also yielding strong performance. However, when λ exceeds 1, the performance starts to degrade. Specifically, $\lambda = 3/2$ and $\lambda = 2$ lead to lower accuracies.

Inaccurate teacher. In Table 7, we present empirical evidence showing that our method remains robust even when the teacher provides inaccurate predictions –

a central motivation for our approach. In this ablation study, we systematically introduce noise into the teacher's outputs. Specifically, for each training example, we randomly shuffle a portion of the teacher's logits. We examine multiple levels of noise: no shuffling (original setting), 15%, 30%, and 45% of the logits shuffled. To isolate the effect of teacher corruption, we set the true-label cross-entropy term to zero, ensuring that the model relies entirely on the noisy teacher signal. This setup allows us to directly assess how resilient our method is to inaccuracies in the teacher's predictions.

Across all noise levels and teacher–student pairs, our method consistently achieves the best performance. Notably, as the noise level increases, KD, DKD, and LSKD suffer severe performance degradation, whereas REDistill (ours) and RLD exhibit only moderate declines. These results offer direct empirical evidence that our method (and RLD) is robust to noisy teacher predictions. Moreover, compared to RLD, our method achieves *consistently higher overall accuracy while maintaining the same level of robustness*.

Table 6: Student top-1 accuracy (%) for different λ (the leftmost column) in model-agnostic setting; $\lambda = 0$ corresponds to DKD [67].

dataset	CIFAR-100							ImageNet-1k-100	
teacher	ResNet32	ResNet32	ResNet32	WRN-40-2	WRN-40-2	VGG13	Res50	Res34	Res50
student	SHN-V2	WRN-16-2	WRN-40-2	ResNet8	MN-V2	MN-V2	MN-V2	Res18	MN-V1
0	77.01	76.18	78.81	75.50	69.27	69.70	70.60	71.70	72.05
1/3	77.12	76.20	78.82	76.91	69.20	70.01	70.55	71.72	72.00
2/3	**77.52**	**76.31**	**79.01**	**77.21**	**69.98**	**70.12**	**70.92**	**72.00**	**72.98**
1	77.50	76.20	78.78	77.15	69.90	69.68	70.51	70.80	72.72
3/2	77.01	76.01	78.12	76.21	69.51	69.11	70.42	70.12	72.19
2	76.31	75.80	78.01	75.80	69.08	69.12	70.48	70.13	71.33

Table 7: Student top-1 accuracy (%) ablation on CIFAR-100 when noise is added by corrupting a random proportion of teacher logits. True-label cross entropy term weight is set to zero to fully evaluate impact of the teacher signal only. Results averaged over three runs.

teacher / student	ResNet32/ SHN-V2				WRN-40-2 / ResNet8				ResNet50 / MN-V2			
Teacher noise	0%	15%	30%	45%	0%	15%	30%	45%	0%	15%	30%	45%
KD [20]	76.12	75.13	74.00	72.90	74.85	73.28	69.59	65.67	63.71	63.15	62.74	59.44
DKD [62]	76.82	76.02	73.98	69.83	75.19	74.13	72.44	68.63	70.25	68.00	61.88	54.45
LSKD [47]	76.67	76.40	73.70	62.59	76.42	75.83	74.84	70.95	70.23	67.50	53.06	41.91
RLD [48]	76.58	76.59	75.84	75.63	74.07	74.31	74.33	74.71	70.12	70.28	69.35	68.53
REDistill (ours)	**77.32**	**77.29**	**76.98**	**76.75**	**76.43**	**76.29**	**76.00**	**75.61**	**70.76**	**70.50**	**70.12**	**69.24**

6 Conclusion

In this work, we revisited the robustness of knowledge distillation (KD) through the lens of robust statistical estimation. We identify a key weakness in conven-

tional KD – the assumption that teacher soft targets are always reliable. When teachers are imperfect or overconfident, the standard KL divergence becomes overly sensitive to noise, leading to unstable or suboptimal student training. To address this, we propose Robust Estimator Distillation (REDistill), which replaces the KL divergence with a power divergence. This yields a principled loss that reduces unreliable teacher output while preserving informative structure – *without heuristics, extra hyper-parameters, architectural changes, or additional computational cost*. Experiments on CIFAR-100 and ImageNet showed that REDistill consistently improved student performance and achieves *state-of-the-art results* while (i) being effective and outperforming logit-based baselines; (ii) not being sensitive to hyper-parameter tuning and/or data augmentation – an important property for generalization to new teacher-student pairs; and (iii) *easy to plug in*, providing consistent improvements when combined with other distillation objectives, showing that our contribution is orthogonal to existing strategies.

Acknowledgement

The research was supported by Czech Science Foundation Grant No. 24-10738M. The access to the computational infrastructure of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics" and of the Ministry of Education, Youth and Sports of the Czech Republic project e-INFRA CZ (ID:90254) is also gratefully acknowledged.

References

1. Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9163–9171 (2019)
2. Akrami, H., Joshi, A.A., Li, J., Aydöre, S., Leahy, R.M.: A robust variational autoencoder using beta divergence. Knowledge-based systems **238**, 107886 (2022)
3. Amari, S.i.: Differential-geometrical methods in statistics, vol. 28. Springer Science & Business Media (2012)
4. Basu, A., Harris, I.R., Hjort, N.L., Jones, M.: Robust and efficient estimation by minimising a density power divergence. Biometrika **85**(3), 549–559 (1998)
5. Box, G.E., Cox, D.R.: An analysis of transformations. Journal of the Royal Statistical Society Series B: Statistical Methodology **26**(2), 211–243 (1964)
6. Broniatowski, M., Keziou, A.: Parametric estimation and tests through divergences and the duality technique. Journal of Multivariate Analysis **100**(1), 16–36 (2009)
7. Cai, L., Chen, Y., Cai, N., Cheng, W., Wang, H.: Utilizing amari-alpha divergence to stabilize the training of generative adversarial networks. Entropy **22**(4), 410 (2020)
8. Cao, Q., Zhang, K., He, X., Shen, J.: Be an excellent student: Review, preview, and correction. IEEE Signal Processing Letters **30**, 1722–1726 (2023)
9. Chen, D., Mei, J.P., Wang, C., Feng, Y., Chen, C.: Online knowledge distillation with diverse peers. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 3430–3437 (2020)

10. Chen, D., Mei, J.P., Zhang, H., Wang, C., Feng, Y., Chen, C.: Knowledge distillation with the reused teacher classifier. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11933–11942 (2022)
11. Chen, H.S., Lai, K., Ying, Z.: Goodness-of-fit tests and minimum power divergence estimators for survival data. Statistica Sinica pp. 231–248 (2004)
12. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5008–5017 (2021)
13. Cramér, H.: Mathematical methods of statistics, vol. 9. Princeton university press (1999)
14. Cressie, N., Read, T.R.: Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society Series B: Statistical Methodology **46**(3), 440–464 (1984)
15. Frýdlová, I., Vajda, I., Kůs, V.: Modified power divergence estimators in normal models–simulation and comparative study. Kybernetika **48**(4), 795–808 (2012)
16. Guo, Z., Yan, H., Li, H., Lin, X.: Class attention transfer based knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11868–11877 (2023)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
18. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1921–1930 (2019)
19. Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernández-Lobato, D., Turner, R.: Black-box alpha divergence minimization. In: International conference on machine learning. pp. 1511–1520. PMLR (2016)
20. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
21. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
22. Huber, P.J.: Robust statistics. In: International encyclopedia of statistical science, pp. 1248–1251. Springer (2011)
23. Ji, M., Heo, B., Park, S.: Show, attend and distill: Knowledge distillation via attention-based feature matching. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 7945–7952 (2021)
24. Jin, Y., Wang, J., Lin, D.: Multi-level logit distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24276–24285 (2023)
25. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International conference on machine learning. pp. 1885–1894. PMLR (2017)
26. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009), technical Report
27. Lan, W., Cheung, Y.m., Xu, Q., Liu, B., Hu, Z., Li, M., Chen, Z.: Improve knowledge distillation via label revision and data selection. IEEE Transactions on Cognitive and Developmental Systems (2025)
28. Li, Z., Li, X., Yang, L., Zhao, B., Song, R., Luo, L., Li, J., Yang, J.: Curriculum temperature for knowledge distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1504–1512 (2023)

29. Li, Z., Ye, J., Song, M., Huang, Y., Pan, Z.: Online knowledge distillation for efficient pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11740–11750 (2021)
30. Lin, S., Xie, H., Wang, B., Yu, K., Chang, X., Liang, X., Wang, G.: Knowledge distillation via the target-aware transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10915–10924 (2022)
31. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
32. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2604–2613 (2019)
33. Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y., Duan, Y.: Knowledge distillation via instance relationship graph. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7096–7104 (2019)
34. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV). pp. 116–131 (2018)
35. Maji, A., Ghosh, A., Basu, A., Pardo, L.: Robust statistical inference based on the c-divergence family. Annals of the Institute of Statistical Mathematics **71**(5), 1289–1322 (2019)
36. Miles, R., Elezi, I., Deng, J.: Vkd: Improving knowledge distillation using orthogonal projections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15720–15730 (2024)
37. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 5191–5198 (2020)
38. Pardo, L.: Statistical inference based on divergence measures. Chapman and Hall/CRC (2018)
39. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3967–3976 (2019)
40. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5007–5016 (2019)
41. Qian, C., Le, T., Harandi, M.: A good teacher adapts their knowledge for distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1239–1248 (2025)
42. Read, T.R., Cressie, N.A.: Goodness-of-fit statistics for discrete multivariate data. Springer Science & Business Media (2012)
43. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: International Conference on Learning Representations (2015)
44. Roulet, V., Liu, T., Vieillard, N., Sander, M.E., Blondel, M.: Loss functions and operators generated by f-divergences. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=V1YfPJDliw
45. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)

46. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
47. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
48. Son, W., Na, J., Choi, J., Hwang, W.: Densely guided knowledge distillation using multiple teacher assistants. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9395–9404 (2021)
49. Su, C.P., Tseng, C.H., Pu, B., Zhao, L., Yang, J., Chen, Z., Lee, S.J.: Ea-kd: Entropy-based adaptive knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 731–740 (2025)
50. Sun, S., Ren, W., Li, J., Wang, R., Cao, X.: Logit standardization in knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15731–15740 (2024)
51. Sun, W., Chen, D., Lyu, S., Chen, G., Chen, C., Wang, C.: Knowledge distillation with refined logits. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1110–1119 (2025)
52. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: International conference on machine learning. pp. 1139–1147. pmlr (2013)
53. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019)
54. Wang, C., Chen, D., Mei, J.P., Zhang, Y., Feng, Y., Chen, C.: Semckd: Semantic calibration for cross-layer knowledge distillation. IEEE Transactions on Knowledge and Data Engineering **35**(6), 6305–6319 (2022)
55. Wang, G., Yang, Z., Wang, Z., Wang, S., Xu, Q., Huang, Q.: ABKD: Pursuing a proper allocation of the probability mass in knowledge distillation via α-β-divergence. arXiv preprint arXiv:2505.04560 (2025)
56. Wen, T., Lai, S., Qian, X.: Preparing lessons: Improve knowledge distillation with better supervision. Neurocomputing **454**, 25–33 (2021)
57. Yamano, T.: Some properties of q-logarithm and q-exponential functions in tsallis statistics. Physica A: Statistical Mechanics and its Applications **305**(3-4), 486–496 (2002)
58. Yuan, F., Shou, L., Pei, J., Lin, W., Gong, M., Fu, Y., Jiang, D.: Reinforced multi-teacher selection for knowledge distillation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 14284–14291 (2021)
59. Yuan, M., Lang, B., Quan, F.: Student-friendly knowledge distillation. Knowledge-Based Systems **296**, 111915 (2024)
60. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
61. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
62. Zhang, H., Chen, D., Wang, C.: Confidence-aware multi-teacher knowledge distillation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4498–4502. IEEE (2022)
63. Zhang, R., Shen, J., Liu, T., Liu, J., Bendersky, M., Najork, M., Zhang, C.: Do not blindly imitate the teacher: Using perturbed loss for knowledge distillation. arXiv preprint arXiv:2305.05010 (2023)

64. Zhang, W., Xie, F., Cai, W., Ma, C.: Vrm: Knowledge distillation via virtual relation matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2707–2717 (2025)
65. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6848–6856 (2018)
66. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4320–4328 (2018)
67. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 11953–11962 (2022)