Comparative Knowledge Distillation

Alex Tianyi Xu* Alex Wilf* Paul Pu Liang Alexander Obolenskiy
Daniel Fried Louis-Philippe Morency

Carnegie Mellon University, Pittsburgh, PA 15213, USA
{alextiax,awilf,pliang,aobolens,dfried,morency}@cs.cmu.edu

Abstract

*In the era of large-scale pretrained models, Knowledge Distillation (KD) serves an important role in transferring the wisdom of computationally-heavy teacher models to lightweight, efficient student models while preserving performance. Yet KD settings often assume readily available access to teacher models capable of performing many inferences—a notion increasingly at odds with the realities of costly large-scale models. Addressing this gap, we study an important question: how KD algorithms fare as the number of teacher inferences decreases, a setting we term Reduced-Teacher-Inference Knowledge Distillation (RTI-KD). We observe that the performance of prevalent KD techniques and state-of-the-art data augmentation strategies suffers considerably as the number of teacher inferences is reduced. One class of approaches, termed "relational" knowledge distillation underperforms the rest, yet we hypothesize that they hold promise for reduced dependency on teacher models because they can augment the effective dataset size without additional teacher calls. We find that a simple change — performing high-dimensional comparisons instead of low-dimensional relations, which we term **Comparative Knowledge Distillation** — vaults performance well over existing KD approaches. We perform empirical evaluation across varied experimental settings and rigorous analysis to understand the learning outcomes of our method. All code is made publicly available.*

1. Introduction

The growing demand for smaller models that retain the capabilities of large pretrained ones has spurred interest in efficient compression techniques [16]. Although Knowledge Distillation (KD) [9] stands out as a promising solution approach [16], KD settings usually assume access to many teacher outputs, which are often costly to obtain from today's large models [26]. This naturally raises the question: how can we perform effective knowledge distillation while using *fewer teacher calls*?

KD is commonly performed by learning to imitate the teacher's representation of an input sample [9], and when few samples are available, data augmentation techniques can be used to generate new samples to ask *additional* questions of the teacher [2]. Yet to the best of our knowledge, no works have investigated how well these methods fare as the number of teacher calls they have access to is reduced, a setting we term Reduced-Teacher-Inference Knowledge Distillation (RTI-KD).

In this paper, we study the RTI-KD setting and find that existing KD approaches struggle to effectively distill knowledge as we reduce teacher calls. We investigate different commonly used KD strategies and hypothesize that one method called "relational" KD [21] shows promise in this setting even though it underperforms other methods. While relational KD has the interesting property that it can create additional learning signals by recombining existing teacher calls, it loses information in its training procedure because it encourages student models to match teacher models' low-dimensional "relations": low-dimensional metrics such as Euclidean distance that define how a model interprets sample representations differently.

We propose a new method called **Comparative Knowledge Distillation** (CKD) which builds on relational KD by learning from pairs of teacher representations, and extends beyond it by encouraging students to learn how teachers view *comparisons*: high-dimensional vector differences between pairs of samples. By training to predict these higher-dimensional comparisons *directly* we can boost performance to well above any other recently proposed algorithm in this setting. Across different image classification architectures, number of teacher calls, and the depth of access to the teacher model (intermediate outputs vs. logits-only), CKD consistently *outperforms baselines for given teacher calls* and requires *fewer teacher inferences to achieve the same performance*, often reducing teacher

[*] Equal contribution

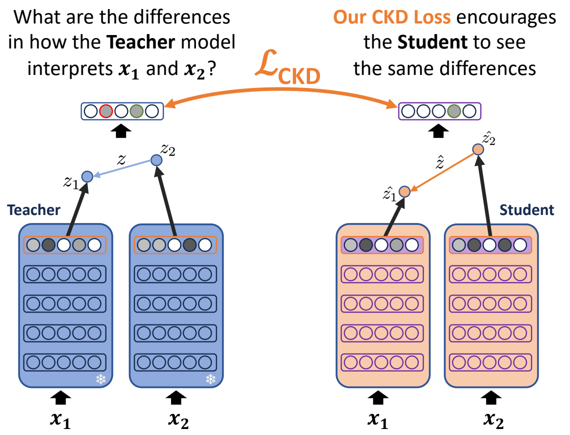


Figure 1. Comparative Knowledge Distillation (CKD): a novel training paradigm that encourages student and teacher representations of the *differences between sample representations*. Critically, since teacher representations can be cached and recombined into many possible comparisons, CKD offers an additional learning signal *without requiring additional teacher calls*, building on relational methods by introducing a high-dimensional loss term.

calls required by the next-best KD technique by a surprising **15%** and for some performance targets, by up to **23%**. We provide extensive analysis on the differences between relational and comparative knowledge distillation and some insight into what CKD learns that makes its representation space so performant in the RTI-KD setting. Our code is publicly available.*

2. Related Work

There are four closely related areas in Knowledge Distillation to our work: KD-Specific loss functions, Data Augmentation Strategies for KD, Relational KD approaches, Contrastive Learning.

KD-Specific Loss Functions Starting with [9]'s KL divergence loss between teacher and student losses, many papers built different loss functions specific to KD [1, 11, 22, 23]. Many papers have also applied KD to intermediate layer representations when given "white box" access to the teacher model's intermediate representations [8, 15, 18, 24, 25, 27, 35, 42]. CKD is *complementary* to these approaches, as these loss functions can be applied to our comparative representations and to representations of single samples. We explain this further in Section 4.4.

Data Augmentation Data augmentations such as flipping, cropping, rotating, and cutout have set the state of the art on some KD tasks [5, 6, 37, 38] and aggregating these strategies has shown promise as well [3]. Augmentation strategies based on Mixup [41] have been particularly performant [17, 33] and synthetic data generation techniques have enabled KD in extremely low-resource settings [20, 32, 34]. Data augmentation strategies can be very effective at augmenting the amount of data that can be used to query the teacher, but in the RTI-KD setting teacher calls are limited due to the cost of teacher queries. By contrast, CKD adds additional learning signals *without additional teacher calls*.

Relation-Based KD In relation-based KD losses, a student's learning signal is derived from a distance metric applied to both the student and the teacher's representations of a pair or group of samples. Many methods implement variants of this approach [4, 19, 21, 23, 36], some applying these methods across or within representation channels [7, 10] or within prediction classes [10]. Relation-Based KD losses are similar to CKD in that they compare student and teacher representations of different samples, but different in that they collapse the representation space into a single number: the Euclidean distance or angle between vectors [21]. To the best of our knowledge, no existing KD approaches have considered learning from high-dimensional comparisons between samples.

Contrastive Learning Contrastive Learning approaches for KD such as CRD [29] and ReKD [44] represent a different but related approach to ours. Contrastive learning methods encourage the student's representation of one sample to be similar or different to the teacher's representation of another, depending on whether the two samples are considered a "positive" or "negative" pair by a pseudo-labeling function that may require ground truth labels [29]. This is similar to our method in that representations from multiple samples are involved, but different in the objective we optimize. CKD encourages students to match a teacher's *comparison* between two samples by having the student consider both samples itself, and requires no pseudo-labeling (i.e., positive and negative pairs).

Teacher-less Distillation Self-distillation [40, 43] and teacher-free distillation [14, 39] approaches have been proposed to address the heavy computation cost of teacher models in the canonical knowledge distillation framework by eliminating the use of a strong teacher altogether. In contrast, CKD focuses on the RTI-KD setting where a performant (albeit black-box) teacher is readily available.

3. Comparative Knowledge Distillation

3.1. Notation and Method Intuition

In this paper, we investigate the Reduced-Teacher-Inference Knowledge Distillation setting (RTI-KD), studying how well KD methods preserve student performance as

[*] https://github.com/alextxu/ckd

the number of teacher calls n is reduced. KD losses are commonly determined by comparing a student's representation $\hat{z}_i$ of sample x_i to the teacher's representation z_i, where z are encoded representations such as logit vectors or in the case of "white-box" access, hidden layer outputs.

By contrast, Relational KD (RKD) [21] learns from how students and teachers view the "relation" between two datapoints i, j, formalized as

$$\mathcal{L}_{RKD} = \mathcal{L}(\psi(\hat{z}_i, \hat{z}_j), \psi(z_i, z_j)) \quad (1)$$

where ψ is a function that describes a low-dimensional "relation" between two vectors such as the Euclidean distance. This has the fascinating property that RKD has access to n^2 learning signals $\mathcal{L}_{RKD}((\hat{z}_i, \hat{z}_j), (z_i, z_j))$ from all pairs of i, j *without requiring additional teacher calls* because representations z_i, z_j can be precomputed and cached. We hypothesize this property may be important for effective KD with reduced teacher calls because students may get valuable learning signals from the nuances of how the teacher interprets the similarities and differences between the many pairs of sample representations in a dataset.

Yet although RKD learns from these many pairs, its relational function ψ reduces the dimensionality of the learning signal to a single value such as Euclidean distance, removing critical information from the learning process. It is perhaps unsurprising, then that RKD consistently underperforms recently proposed methods [10, 29].

3.2. From Relational to Comparative: the CKD Loss Function

To address this shortcoming we propose a method, termed **Comparative Knowledge Distillation**, which also draws pairs of datapoints from the n^2 possible combinations, but attempts to match **high-dimensional comparisons** between teacher and student representations *directly*, without the dimensionality reduction involved in calculating relations such as the angle between vectors or the Euclidean distance.

Our method is illustrated in Figure 1, and our loss can be formulated for datapoints i, j, student representations $\hat{z}$ and teacher representations z as:

$$\mathcal{L}_{CKD}(\hat{z}_i, \hat{z}_j, z_i, z_j) = \mathcal{L}_{MSE}(\hat{z}_i - \hat{z}_j, z_i - z_j) \quad (2)$$

For a fair comparison with prior work [9, 10, 21] we also apply our loss to the ground truth labels y. The complete loss function is then

$$\mathcal{L} = \sum_{i,j} \mathcal{L}_{CKD}(\hat{y}_i, \hat{y}_j, y_i, y_j) + \mathcal{L}_{CKD}(\hat{z}_i, \hat{z}_j, z_i, z_j) \quad (3)$$

Our intuition is that this method will help to regularize the learning process in the presence of reduced teacher calls by encouraging students to match how the teacher interprets similarities and differences between many pairs of datapoints in a rich high-dimensional space.

4. Experimental Setup

4.1. Methodology

Datasets We conduct our experiments on the CIFAR-100 [13] and Stanford Cars [12] datasets which have been commonly used in KD experiments [21, 25, 29]. We investigate reduced teacher call settings by constraining the dataset to randomly chosen subsets (n) in the range $[6400, 1600]$ by decrements of 400 for CIFAR, and $[2000, 1400]$ by decrements of 200 for Cars. We chose these ranges to test our performance on as few teacher calls as possible while achieving reasonable levels of variance during evaluation. This allows us to establish a meaningful comparison with the baselines. We explore a narrower range of teacher calls for Cars because runs take much longer due to the higher resolution of the images. We split the data 80-20% for train and validation and evaluate on the CIFAR-100 and Cars test sets, maintaining the same train, validation and test splits for all experiments.

Teacher-Student Combinations We explore the same teacher-student model combinations as prior work [29]: VGG13 to VGG8, WRN-40-2 to WRN-16-2, ResNet110 to ResNet32.

Data Preprocessing When passing samples through any model (teacher or student) on CIFAR-100 [13], we perform a random cropping of 32x32 with a padding of 4, followed by a random horizontal flip as in [29]. For Stanford Cars [12], we resize the images to 64x64 and take a random cropping of 56x56, followed by a random horizontal flip as in RKD [21].

Training Details We run each student model over five trials and report the mean and standard deviation of our results. We train to convergence using early stopping on the validation loss instead of fixed epochs so that each algorithm runs to convergence before evaluation. We use trained teacher models from [29] and searched over hyperparameters centered around the defaults found in [29]. Additional training details can be found in Appendix A.

4.2. Baselines

We report results on the following baselines, selected because of their strong performance on KD tasks and their data augmentation properties in low-resource settings.

1. **Knowledge Distillation (KD)** [9]: this is the standard KD loss, employing KL divergence loss between the teacher and student logits.

2. **Contrastive Representation Distillation (CRD)** [29] is a contrastive learning method that uses the label to group "positives" and "negatives" in each batch and

encourage the student's representations to be similar to the teacher's for positives and dissimilar for negatives.

3. **Mixup**: As Mixup applied to KD requires additional teacher calls on the mixed up inputs [17], we implement the "Fixed Teacher" [2] version, in which the teacher's output logits from the original datapoints are recombined and used for supervision.

4. **Relational Knowledge Distillation (RKD)** [21] is a "relational" KD approach based on learning a low-dimensional value such as angle-between or euclidean distance that describes how the teacher interprets two samples differently. By contrast, our proposed CKD encourages students to match the teacher's *high-dimensional comparisons* between samples.

5. **Distillation from a Stronger Teacher (DIST)** [10] is an approach that improves over the standard KD loss by encouraging the student to match the intra-class probabilities across samples. Explicitly, for two matrices representing the teacher and student's logits of a batch, KD [9] attempts to match the rows (per-sample logits), whereas DIST also includes a loss based on the columns (encouraging representations across samples in a batch to be similar for each dimension).

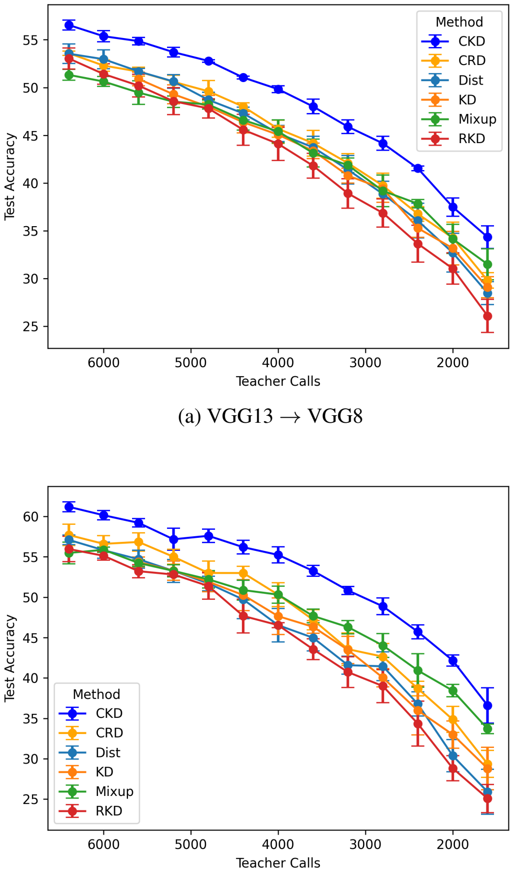
4.3. Evaluation

We evaluate our method using two metrics: **Test Accuracy** and **Teacher Calls Required for Target Accuracy**. The former is calculated simply as top-1 accuracy on the test set. For the latter, we are interested in, for a given target accuracy, *how many teacher calls a method requires to achieve that target accuracy*. Because we test discrete values of teacher calls that may not perfectly match a performance target, we linearly interpolate between the two closest performance values to arrive at the estimate for the number of teacher calls required to achieve a given target performance. For example, if a method required 4000 teacher calls to achieve 40% accuracy and 4400 teacher calls to achieve 50%, we would estimate the number of teacher calls required to achieve a target accuracy of 45% to be 4200.

4.4. Extension to White-Box Setting

One common KD setting is "white-box," in which not only are the teacher-produced logits available for training, but so too are the teacher model's intermediate layer outputs for those samples. Some KD loss functions are designed specifically for intermediate outputs [1, 25]. Our approach is *complementary* to these; we can simply replace the teacher and student representations of a *single sample* with the teacher and student's representations of the *comparison*: the difference between two samples' representations. Concretely, for student networks with k_s layers and teacher networks with k_t layers, white-box KD approaches train the student with loss from a *sequence* of latent output representations containing representations from each layer of the student and teacher:

$$\mathcal{L}_{WB}(\hat{z}_i^1, ..., \hat{z}_i^{k_s}, z_i^1, ..., z_i^{k_t}) \tag{4}$$



(a) VGG13 → VGG8

(b) WRN-40-2 → WRN-16-2

Figure 2. Experimental results on CIFAR-100 represented visually for VGG and WRN models. CKD consistently outperforms baselines as teacher calls are reduced for different teacher-student distillation settings common in the literature [29]. → indicates distilling a teacher into a student model. Points and error bars are the mean and standard deviation of runs over five trials.

CKD is compatible with these methods because we can replace these latent vectors with *high-dimensional comparisons*: the vector difference between representations of samples i, j at the same layer of the network. This allows CKD to then integrate with the white-box losses without modification.

$$\begin{aligned}\mathcal{L}_{WB+CKD} = \mathcal{L}_{WB}(\hat{z}_i^1 - \hat{z}_j^1, ..., \hat{z}_i^{k_s} - \hat{z}_j^{k_s}, \\ z_i^1 - z_j^1, ..., z_i^{k_t} - z_j^{k_t})\end{aligned} \tag{5}$$

In our experiments, we investigate whether this comparative version can improve performance when used with two widely used intermediate layer losses, FitNets [25] and Variational Information Distillation (VID) [1]. We also integrate the high performing KD method CRD [29] with these white box losses, but it is important to note that because CRD is not a method that alters the inputs, it must then combine with white-box losses instead of integrating with them as CKD can.

5. Results and Discussion

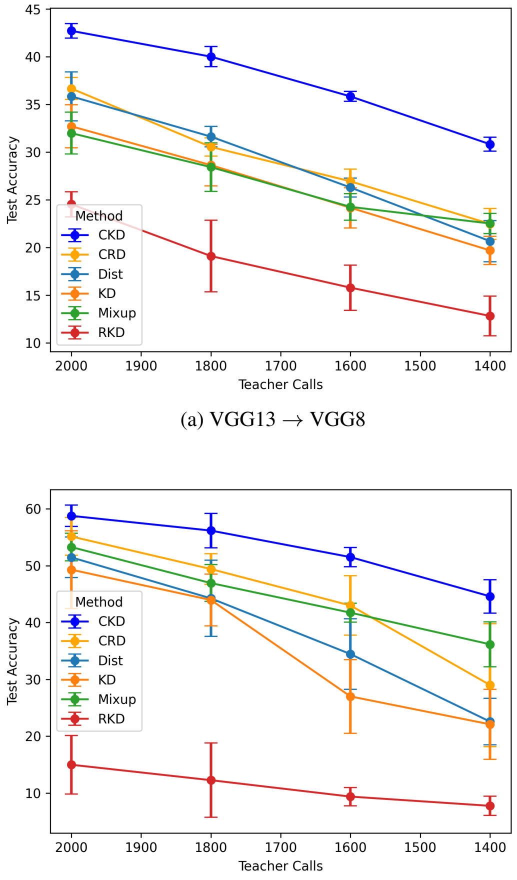
5.1. RTI-KD Results

Our results are depicted in Table 1 and Figures 2 and 3. We find that across a variety of student-teacher combinations including Wide ResNet (WRN), VGG, and ResNet models, our approach consistently outperforms KD baselines as the number of available teacher calls is reduced and significantly reduces the number of teacher calls necessary to achieve the same performance targets.

For WRN models to reach a target accuracy of 55% on CIFAR-100, RKD requires 5981 teacher calls, CRD requires 5199, and CKD requires only 3969, a reduction of **33.64%** compared to RKD and **23.66%** compared to CRD, the next best performing approach. For WRN, VGG, and ResNet models on CIFAR-100, averaged across all performance targets in Table 1, CKD reduces the number of teacher calls required compared with RKD by **36.63%, 30.44%, 30.56%**. Compared with the next-best-performing method, CRD, this reduces teacher calls by **25.9%, 21.34% and 16.83%**.

Considering performance at different teacher call values instead of reduction required to reach the same performance, we find that CKD outperforms baselines substantially. For WRN models, CKD outperforms the baselines Dist, RKD, Mixup, KD, and CRD by **7.26%, 8.46%, 4.68%, 7.07%, and 5.03%** absolute test accuracy averaged across teacher calls; on VGG models too, CKD outperforms these approaches by **4.18%, 5.88%, 4.44%, 4.82% and 3.64%**, and for ResNet models by **7.82%, 7.81%, 4.00%, 7.27% and 4.35%**. These results are visualized in Figure 2. This trend holds for even larger number of teacher calls up to 16000, before finally converging with top baseline methods, as shown in Appendix B.

On the Stanford Cars dataset as well, our method significantly outperforms prior work. Compared with the next best method, CRD, on VGG models CKD performs **8.19%** better on average across teacher calls, by **2.78%** on ResNet models, and by **8.63%** on WRN models. This result is visualized in Figure 3.



(a) VGG13 → VGG8

(b) WRN-40-2 → WRN-16-2

Figure 3. Experimental results on Stanford Cars represented visually. Similarly to Figure 2, points and error bars are mean and standard deviations over five trials.

5.2. Extension to White-Box Access

We find that CKD also integrates with different intermediate layer loss functions well, often improving two commonly used intermediate layer loss functions by substantial margins. Our results are depicted in Table 2. In the WRN distillation setting averaged across low resource teacher calls n ranging from 3200 to 1600, adding CKD to FitNets white-box loss led to an improvement of **8.84%** absolute top-1 accuracy improvement, and an improvement of **8.50%** over adding CRD to FitNets. Results of adding CKD to VID were similar although not quite as pronounced, leading to average improvements of **4.27%** and **4.55%** over VID and VID+CRD, respectively. With VGG models the results were even less pronounced, on VID leading to an average improvement of **0.62%** over VID and **1.68%** over VID+CRD. CKD modifications to the VID loss may be less performant because of VID's use of mutual information to

Table 1. We calculate how many teacher calls (in thousands) are needed to achieve desired test accuracy thresholds on CIFAR-100 for different teacher → student distillations. We find that CKD can achieve the same performance while reducing the number of teacher calls required to do so. Δ computes the percent reduction in teacher calls from the next closest baseline for that target accuracy.

Target Acc	55	50	45	40	35	30
WRN-40-2→WRN-16-2						
KD [9]	–	4.37	3.45	2.80	2.29	1.72
RKD [21]	5.98	4.63	3.83	3.13	2.46	2.02
Dist [10]	5.88	4.59	3.82	2.69	2.27	1.97
Mixup [41]	5.90	3.95	3.03	2.30	1.71	–
CRD [29]	5.20	3.96	3.35	2.56	2.08	1.65
CKD	3.97	3.11	2.34	1.84	–	–
Δ	**↓ 23.66%**	**↓ 21.45%**	**↓ 22.97%**	**↓ 19.72%**	–	–
ResNet110→ResNet32						
KD [9]	–	5.07	3.89	3.14	2.50	1.97
RKD [21]	–	5.00	3.76	3.34	2.70	2.19
Dist [10]	6.34	5.13	3.95	3.16	2.63	2.19
Mixup [41]	–	4.58	3.33	2.38	1.87	–
CRD [29]	5.50	3.97	3.39	2.76	2.22	1.85
CKD	4.61	3.40	2.62	2.18	1.86	1.66
Δ	**↓ 16.13%**	**↓ 14.46%**	**↓ 21.29%**	**↓ 8.40%**	**↓ 0.48%**	**↓ 10.63%**
VGG13→VGG8						
KD [9]	–	5.44	3.99	3.10	2.37	1.69
RKD [21]	–	5.57	4.32	3.37	2.59	1.92
Dist [10]	–	5.10	3.97	3.03	2.29	1.75
Mixup [41]	–	5.86	3.93	2.91	2.04	–
CRD [29]	–	5.10	3.90	2.93	2.20	1.62
CKD	5.92	4.22	3.08	2.23	1.69	–
Δ	–	**↓ 17.11%**	**↓ 21.06%**	**↓ 23.45%**	**↓ 17.52%**	–

Table 2. Given white-box access to intermediate teacher outputs, CKD seamlessly integrates with KD losses designed to learn from intermediate representations, improving their performances in the RTI-KD setting (and even improves over adding CRD loss).

Teacher Calls	3200	2400	1600
WRN-40-2→WRN-16-2			
FitNets [25]	$39.44_{4.50}$	$30.90_{3.67}$	$24.08_{0.74}$
+CRD [29]	$41.59_{1.18}$	$32.62_{2.55}$	$21.20_{1.59}$
+CKD	$\mathbf{47.78_{0.96}}$	$\mathbf{41.43_{2.29}}$	$\mathbf{31.72_{2.46}}$
VID [1]	$42.70_{0.97}$	$37.40_{1.33}$	$28.82_{1.34}$
+CRD [29]	$45.29_{1.19}$	$36.28_{1.00}$	$26.53_{2.47}$
+CKD	$\mathbf{47.23_{1.27}}$	$\mathbf{41.52_{1.69}}$	$\mathbf{32.99_{1.06}}$
VGG13→VGG8			
FitNets [25]	$39.27_{1.44}$	$33.98_{1.41}$	$27.12_{1.85}$
+CRD [29]	$36.89_{0.83}$	$32.24_{1.12}$	$24.96_{1.72}$
+CKD	$\mathbf{40.91_{0.97}}$	$\mathbf{36.18_{0.91}}$	$\mathbf{30.15_{1.53}}$
VID [1]	$40.87_{1.09}$	$35.87_{1.02}$	$29.29_{1.28}$
+CRD [29]	$39.88_{1.18}$	$34.74_{1.29}$	$28.23_{1.01}$
+CKD	$\mathbf{41.19_{0.54}}$	$\mathbf{36.97_{0.59}}$	$\mathbf{29.73_{1.32}}$

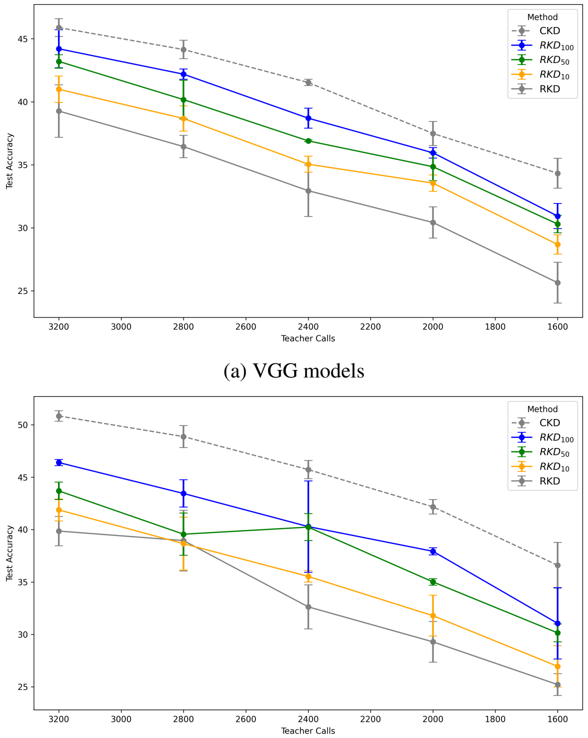
encourage student and teacher representations together – it may be the case that our high-dimensional comparison vectors require a particular normalization to integrate well with VID. Using FitNets, however, average improvement on the WRN and VGG models was more substantial: **2.29%** and **4.38%** over FitNets and FitNets+CRD respectively. These results indicate that CKD is not only capable of replacing single-sample representations in white-box distillation, but that it can often outperform single sample losses as well.

6. Analysis

6.1. Understanding High-Dimensional Comparison

In this section, we investigate how important it is that these students learn from *high-dimensional* comparative signals. To explore this, we look to the relationship between CKD and its closest counterpart, RKD.

As described in Section 2, RKD calculates low-dimensional metrics they describe as a "relation" between two samples, such as Euclidean distance. We investigate how performance changes with different dimensionalities in between single-dimension Euclidean distance and full dimension (for CIFAR-100 logits, this is 100-dimensional). To study these in-between levels of dimensionality, we apply average pooling (θ_m) with a pool size and stride length equal to $d = D/m$, effectively downsampling the differ-



(a) VGG models

(b) WRN models

Figure 4. Dimensionality is important in transferring information from teacher to student in the RTI-KD setting. Higher dimensional versions of RKD, RKD$_{100}$, RKD$_{50}$, and RKD$_{10}$ lead to increased performance over the original RKD algorithm. Additionally, the gap between RKD$_{100}$ and CKD illustrates that it is also important to apply comparative loss to the ground truth labels as well as teacher representations.

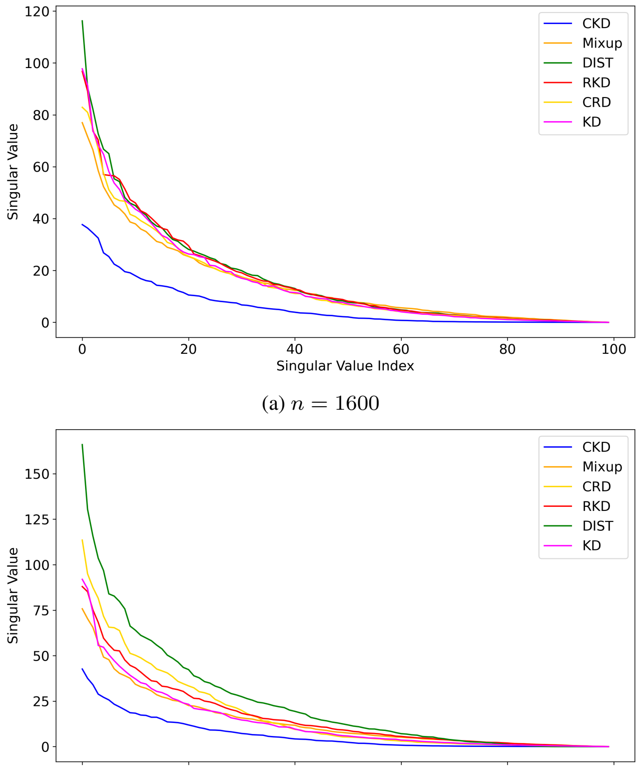
ence vector to dimensionality d to before calculating the MSE loss.

Formally, this loss function is described in Equation 6 below. As in Section 3, z_i and z_j are the teacher's representations of samples i, j and $\hat{z}_i, \hat{z}_j$ are the student's representations. θ_d represents the pooling function that downsamples the D-dimensional vector into an d dimensional vector.

$$\mathcal{L}_{RKD_d}(\hat{z}_i, \hat{z}_j, z_i, z_j) = \mathcal{L}_{mse}\left(\theta_d(\hat{z}_i - \hat{z}_j), \theta_d(z_i - z_j)\right) \tag{6}$$

We investigate how this loss performs on VGG and WRN models on the CIFAR-100 dataset in Figure 4. At $d = 100$, the only difference between this loss and CKD is that CKD also applies its comparative loss to ground truth labels y, as described in Equation 3, whereas $\mathcal{L}_{RKD_d}$ uses standard cross-entropy loss between single samples, as in RKD [21].

We find that dimensionality is indeed important; the 100-dimensional comparison yields substantially better performance than the 50, 10, or 1-dimensional versions, as shown in Figure 4. On average across teacher calls on VGG models, RKD performs at 32.95% accuracy, RKD$_{10}$ at 35.39%, RKD$_{50}$ at 37.08%, and RKD$_{100}$ at 38.4%. The trend also holds for WRN models: RKD performs at 33.19%, RKD$_{10}$ at 34.96%, RKD$_{50}$ at 37.73%, and RKD$_{100}$ at 39.82%. This supports our intuition that rich high-dimensional information is important for transferring learning from teacher to student in the RTI-KD setting. Additionally, as RKD$_{100}$ and CKD differ only in their application of comparative loss to the ground truth labels, CKD's improvements over RKD$_{100}$ articulate the importance as well of using comparative losses with both the labels and the teacher representations.



(a) $n = 1600$

(b) $n = 2000$

Figure 5. CKD acts as a regularizer, flattening student models' representation spaces: a property that is closely tied to generalization [28, 30].

6.2. Analyzing Representations Learned by CKD

Our intuition for why CKD performs well is twofold: first, because CKD provides additional learning signals that may help regularize the student in the low-data regime, and second, because it provides rich high-dimensional comparisons that may allow the student to better match the teacher's representations. We evaluate these intuitions in the two experiments that follow, and find that CKD does indeed act as a regularizer, and encourages the student to better match the teacher's representations than baseline approaches.

CKD Flattens the Representation Space Our first intuition is that CKD may act as a regularizer, introducing an additional learning signal that helps shape the optimization space in ways that are favorable to generalizable learning of the teacher model under low-resource conditions. To investigate this, we analyze the *flatness* of logit representations space, which has been linked to generalization by established theory [28, 30]. We do this by performing the analysis from [31] which analyzes the flatness of the representations by performing Singular Value Decomposition (SVD) on the student representations, where a lower curve indicates flatter representations. We perform this experiment across two low-resource settings of n on the saved WRN student models' logit representations. Our results are visualized in Figure 5 – CKD's SVD curve is substantially below others, indicating that CKD may act as a regularizer, promoting generalization in the challenging low-resource RTI-KD setting. Notably, the CKD curve is significantly lower than even Mixup [41], a data augmentation strategy specifically designed for regularization. The link between regularization and generalization has been explored in many works [41], and it is promising that CKD exhibits strong regularization effects on the training process.

Table 3. Training with CKD leads to an improvement in matching the student's correlation across class logits to the teacher's, a property CRD [29] found important for KD representation learning. This table depicts the average absolute difference of student and teacher's correlation matrices; lower is better. Surprisingly, CKD outperforms even CRD, which *explicitly* optimizes this objective.

Teacher Student	ResNet110 ResNet32	VGG13 VGG8	WRN-40-2 WRN-16-2
Mixup [41]	0.109	0.109	0.101
DIST [10]	0.103	0.097	0.102
RKD [21]	0.102	0.099	0.098
CRD [29]	0.096	0.096	0.093
CKD	**0.081**	**0.089**	**0.083**

Student-Teacher Logit Correlations Tian et al. [29] showed that capturing the inter-class correlations between teacher logits is important to successful KD outcomes in students. We reproduce their experiment [29] to analyze how well CKD encourages this desirable property in students: the details are described below.

Across 100 randomly chosen samples from the CIFAR-100 test set, we first calculate the correlation matrices between class logits for both the teacher and the student. This is done by centering the data by mean, computing the outer product of the resulting vectors to arrive at the covariance matrix, then normalizing by standard deviation to yield the correlation matrix. Then, we report the average absolute difference between the student (trained in different ways) and the teacher's correlation matrices. Lower is better, because a value of 0 would indicate perfect imitation of the teacher's inter-class logit correlations.

In Table 3 we report the numerical results of this correlation analysis. We find that CKD outperforms baselines including CRD [29], whose objective *explicitly* attempts to capture inter-class correlations. This analysis, along with the main results, indicates that CKD's comparative loss function is providing strong KD learning outcomes that enable the student to match the teacher's representations. We hypothesize that this is largely due to the richness of the *high-dimensional comparisons*, as RKD (without high-dimensional comparison) augments the dataset size but does not achieve nearly so precise a correlation with teacher outputs as CRD or CKD.

7. Conclusion

In this paper we introduced **Comparative Knowledge Distillation (CKD)**, a novel learning paradigm that we show is useful in performing Knowledge Distillation as teacher inferences are reduced (RTI-KD). CKD does this by recombining existing teacher calls into high-dimensional *comparisons*, encouraging student models to mimic teacher's difference in representation between samples directly in a high-dimensional space. Empirical evaluations reveal CKD's superiority over state-of-the-art KD techniques across various settings. Moreover, CKD is complementary to KD loss functions designed specifically for intermediate representations; by modifying single sample representations to be "comparative" sample representations before feeding them into white-box loss functions, with no additional changes to the loss functions we find that CKD often outperforms those loss functions. In our analysis, we find that CKD captures critical inter-class correlations and acts as a regularizer on the logit space, enhancing generalization in the low-resource setting.

Finally, this study attempts to lay a foundation for efficient KD research in the era of large-scale pretrained models, where teacher inferences can often be costly. One important limitation of this line of research is a deeper understanding of when and how biases in teacher models can be inherited by student models. It will be important for future work to conduct a principled investigation of bias transfer in knowledge distillation to advise future directions in this space. Future work may also find it fruitful to further investigate the reduced teacher inference setting and additional applications of comparative training (for example, to large language models or to different vision tasks).

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9163–9171, 2019.

[2] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10925–10934, 2022.

[3] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[4] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7842–7851, 2021.

[5] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[6] Jie Fu, Xue Geng, Zhijian Duan, Bohan Zhuang, Xingdi Yuan, Adam Trischler, Jie Lin, Chris Pal, and Hao Dong. Role-wise data augmentation for knowledge distillation. *arXiv preprint arXiv:2004.08861*, 2020.

[7] Jianping Gou, Xiangshuo Xiong, Baosheng Yu, Yibing Zhan, and Zhang Yi. Channel correlation-based selective knowledge distillation. *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[8] Md Akmal Haidar, Nithin Anchuri, Mehdi Rezagholizadeh, Abbas Ghaddar, Philippe Langlais, and Pascal Poupart. Rail-kd: Random intermediate layer mapping for knowledge distillation. *arXiv preprint arXiv:2109.10164*, 2021.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[10] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727, 2022.

[11] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[12] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

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

[14] Lujun Li. Self-regulated feature learning via teacher-free feature distillation. In *European Conference on Computer Vision*, pages 347–363. Springer, 2022.

[15] Linfeng Li, Weixing Su, Fang Liu, Maowei He, and Xiaodan Liang. Knowledge fusion distillation: Improving distillation with multi-scale attention mechanisms. *Neural Processing Letters*, pages 1–16, 2023.

[16] Zhuo Li, Hengyi Li, and Lin Meng. Model compression for deep neural networks: A survey. *Computers*, 12(3):60, 2023.

[17] Kevin J Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, and Lawrence Carin. Mixkd: Towards efficient distillation of large-scale language models. *arXiv preprint arXiv:2011.00593*, 2020.

[18] Xiaolong Liu, Lujun Li, Chao Li, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching, 2023.

[19] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7096–7104, 2019.

[20] Dang Nguyen, Sunil Gupta, Kien Do, and Svetha Venkatesh. Black-box few-shot knowledge distillation. In *European Conference on Computer Vision*, pages 196–211. Springer, 2022.

[21] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

[22] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–284, 2018.

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

[24] Cuong Pham, Van-Anh Nguyen, Trung Le, Dinh Phung, Gustavo Carneiro, and Thanh-Toan Do. Frequency attention for knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2277–2286, 2024.

[25] 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.

[26] 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 *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9. IEEE, 2023.

[27] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021.

[28] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

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

[30] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015.

[31] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Na-

jafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019.

[32] Dongdong Wang, Yandong Li, Liqiang Wang, and Boqing Gong. Neural networks are more productive teachers than human raters: Active mixup for data-efficient knowledge distillation from a blackbox model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1498–1507, 2020.

[33] Huan Wang, Suhas Lohit, Michael N Jones, and Yun Fu. What makes a" good" data augmentation in knowledge distillation-a statistical perspective. *Advances in Neural Information Processing Systems*, 35:13456–13469, 2022.

[34] Zi Wang. Zero-shot knowledge distillation from a decision-based black-box model. In *International Conference on Machine Learning*, pages 10675–10685. PMLR, 2021.

[35] Yimeng Wu, Mehdi Rezagholizadeh, Abbas Ghaddar, Md Akmal Haidar, and Ali Ghodsi. Universal-kd: Attention-based output-grounded intermediate layer knowledge distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7649–7661, 2021.

[36] Xiaomeng Xin, Heping Song, and Jianping Gou. A new similarity-based relational knowledge distillation method. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3535–3539. IEEE, 2024.

[37] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *European Conference on Computer Vision*, pages 588–604. Springer, 2020.

[38] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Hierarchical self-supervised augmented knowledge distillation. *arXiv preprint arXiv:2107.13715*, 2021.

[39] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3903–3911, 2020.

[40] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13876–13885, 2020.

[41] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[42] Linfeng Zhang, Xin Chen, Junbo Zhang, Runpei Dong, and Kaisheng Ma. Contrastive deep supervision. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022.

[43] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3712–3721, 2019.

[44] Kai Zheng, Yuanjiang Wang, and Ye Yuan. Boosting contrastive learning with relation knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3508–3516, 2022.