Comparative Knowledge Distillation

WACV 2025

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

- Problem/Objective
  - Knowledge distillation
- Contribution/Key Idea
  - Novel KD framework
  - SOTA

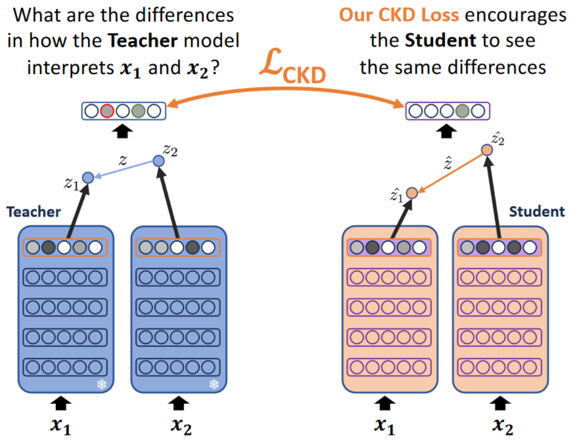
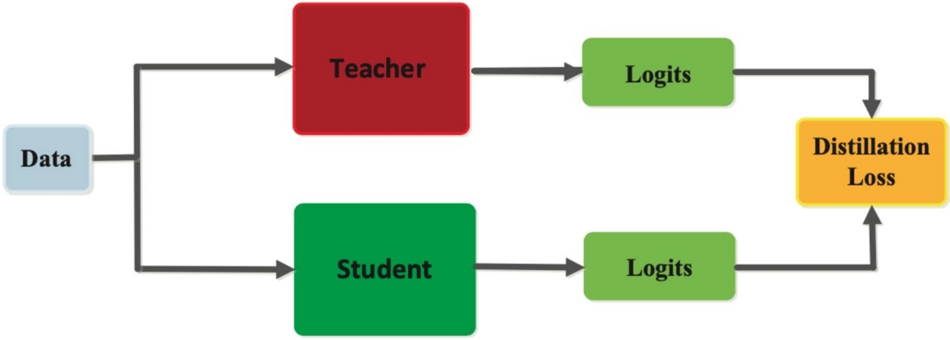


Figure 1. Comparative Knowledge Distillation (CKD): a novel training paradigm that encourages student and teacher representations of the *differences between sample representations*. Critically, since teacher representations can be cached and recombined into many possible comparisons, CKD offers an additional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

김범준

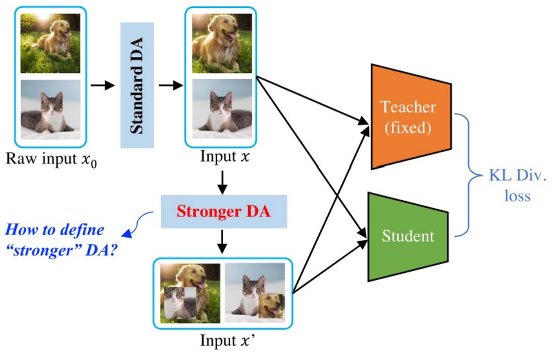
- **Knowledge Distillation (KD)**



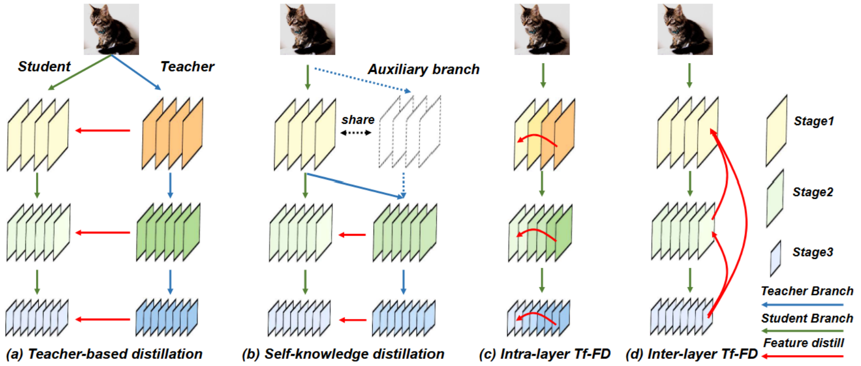
- 최근 Foundation model의 발전 등 Teacher model로 사용할 model들은 갈수록 large, heavy 해짐
- 각 data에 대한 inference의 cost ↑
  - Efficient 한 KD framework 필요
  - ***fewer teacher call*** 방법에 대한 고민

This naturally raises the question: how can we perform effective knowledge distillation while using *fewer teacher calls*?

- **Teacher model's fewer calls**



Data-Augmentation [1]

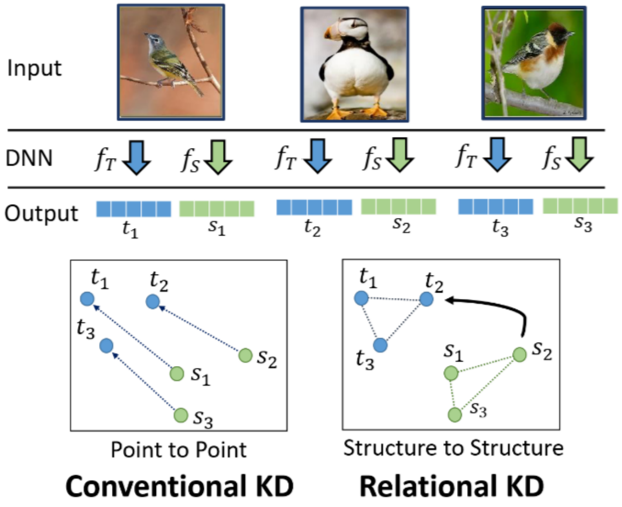


Teacher-Free Distillation [2]

- Data 수↓ 경우 **fewer calls(목표)**, but 이런 경우는 오히려 data augmentation
  - Multi-Teacher KD 방법
- Fewer inference call을 유지하는 방법에 대해선 연구된적 x
  - self-distillation, teacher-free distillation 방법 o but, 효과↓
  - Strong Teacher를 유지하지만, fewer call 방법 연구

[1] Wang, Huan, et al. "What makes a" good" data augmentation in knowledge distillation-a statistical perspective." Advances in Neural Information Processing Systems 35 (2022): 13456-13469.
[2] Li, Lujun. "Self-regulated feature learning via teacher-free feature distillation." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **기존의 KD, relation KD [1]**



$$L = \mathrm{Loss}(\hat{z}_i,\ z_i)$$

$$L_{\mathrm{RKD}} = L(\psi(\hat{z}_i,\ \hat{z}_j),\ \psi(z_i,\ z_j))$$

- ψ → 유클리드 거리, 각도 등 관계 측정 함수 (to low-dimensional description)
- 모든 쌍 i, j 에 대해 학습 신호 O(n^2) 만들 수 있고, teacher inference는 n번
  - 단점 : ψ 가 벡터 → 스칼라 → 풍부한 정보 loss

[1] Park, Wonpyo, et al. "Relational knowledge distillation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

- **CKD (Comparative Knowledge Distillation)**

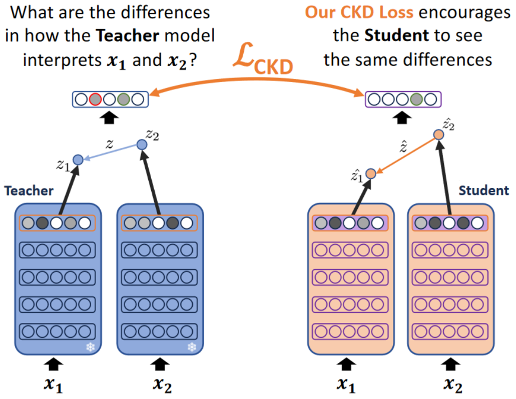


Figure 1. Comparative Knowledge Distillation (CKD): a novel training paradigm that encourages student and teacher representations of the *differences between sample representations*. Critically, since teacher representations can be cached and recombined into many possible comparisons, CKD offers an additional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

$$L_{\mathrm{CKD}}(\hat{z}_i,\ \hat{z}_j,\ z_i,\ z_j) = \mathrm{MSE}(\hat{z}_i - \hat{z}_j,\ z_i - z_j)$$

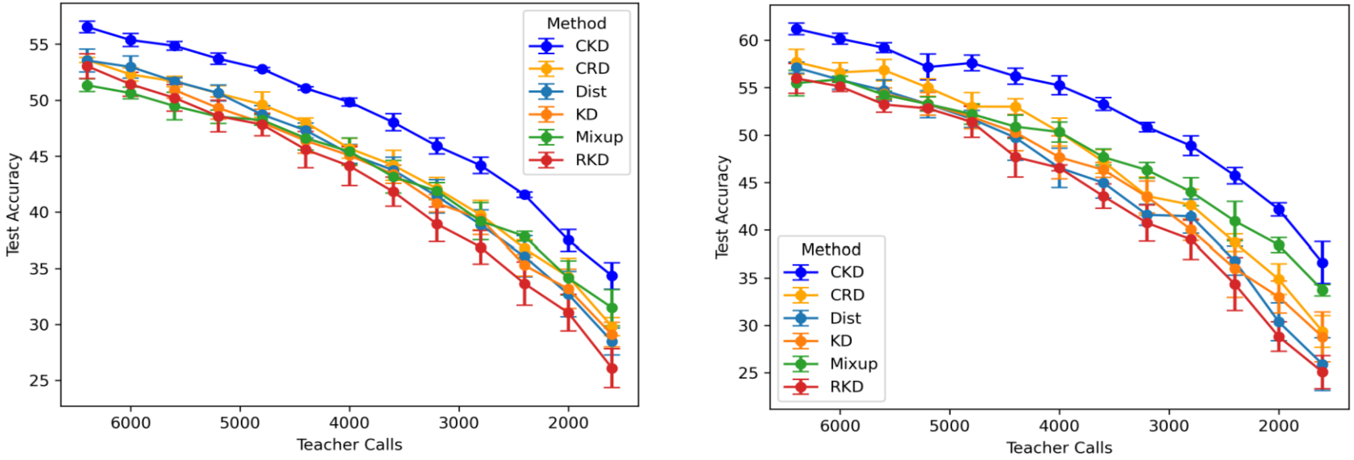
- **CKD**
- 모든 쌍 i, j 에 대해 학습 동일 (low teacher inference 사용)
  - But) 벡터 간의 high-dimensional 차이 직접 학습

$$L = \sum_{i,j} [\mathrm{MSE}(\hat{y}_i - \hat{y}_j,\ y_i - y_j) + \mathrm{MSE}(\hat{z}_i - \hat{z}_j,\ z_i - z_j)]$$

- 기존 방법들과의 공정한 비교를 위해 GT 역시 이러한 방법 사용
  - teacher 모델이 어떻게 샘플간의 차이를 학습하는지를 student가 학습
  - CKD는 학습을 regularize

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.

- **Experiments**
  - CIFAR-100 & Stanford Cars 에서 랜덤하게 n개의 데이터셋만 추출하여 실험을 진행

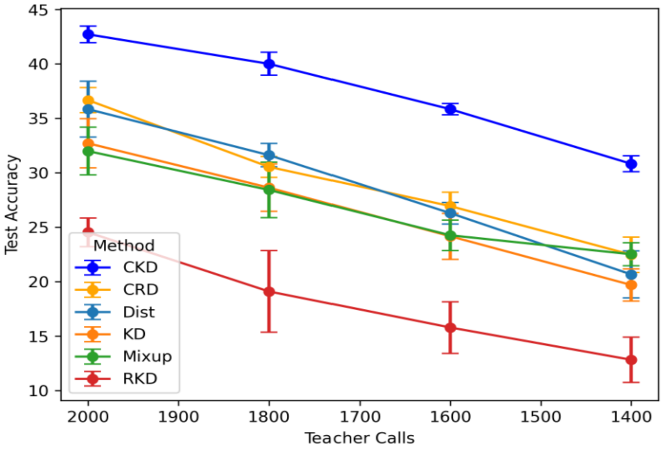


(a) VGG13 → VGG8

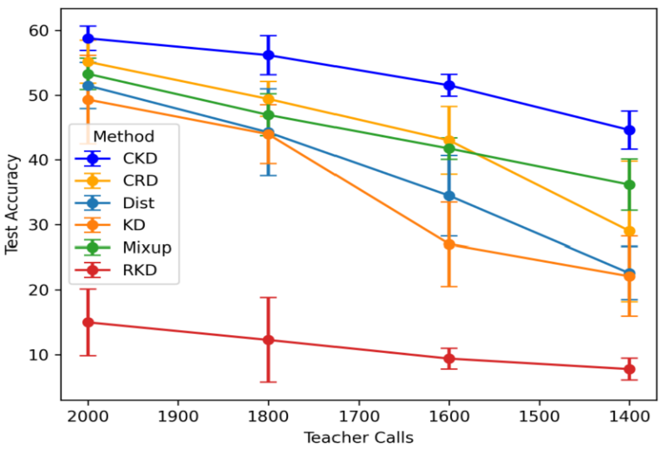
(b) WRN-40-2 → WRN-16-2

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.

- **Experiments**
  - CIFAR-100 & Stanford Cars 에서 랜덤하게 n개의 데이터셋만 추출하여 실험을 진행



(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.

- **Experiments**
  - 특정 성능에 도달하기 위해 필요한 teacher call 수

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%**	–

- **Experiments**
  - 줄어든 teacher call 수 고정일 때 성능 비교

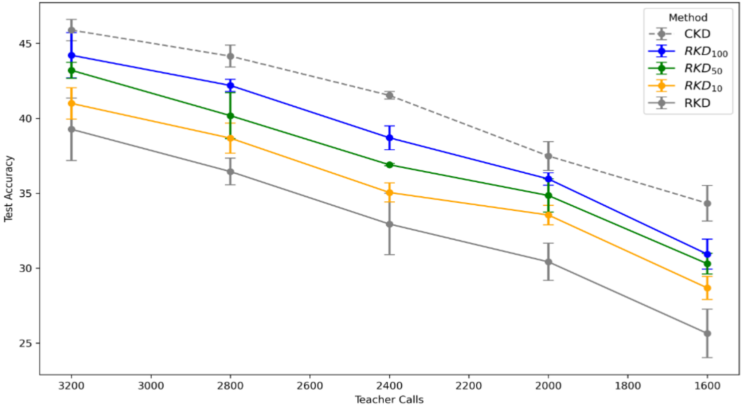
Table 2. Given white-box access to intermediate teacher outputs, CKD seamlessly integrates with KD losses designed to learn from intermediate representations, improving their performances in the RTI-KD setting (and even improves over adding CRD loss).

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	$\mathbf{41.19_{0.54}}$	$\mathbf{36.97_{0.59}}$	$\mathbf{29.73_{1.32}}$

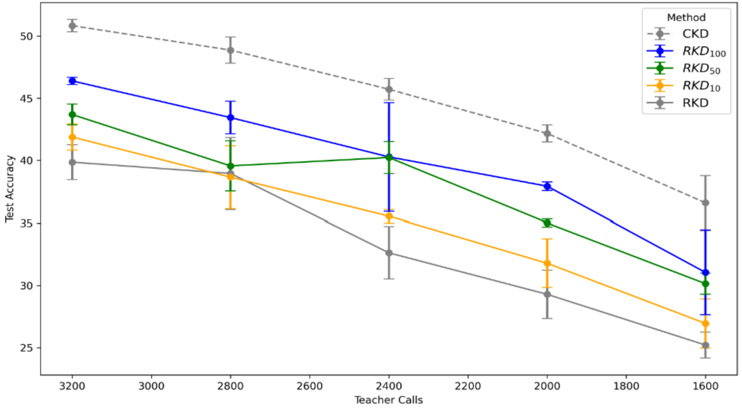
[Comparative Knowledge Distillation](#)

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

- **Experiments**
  - High-Dimensional KD에 대한 유효성



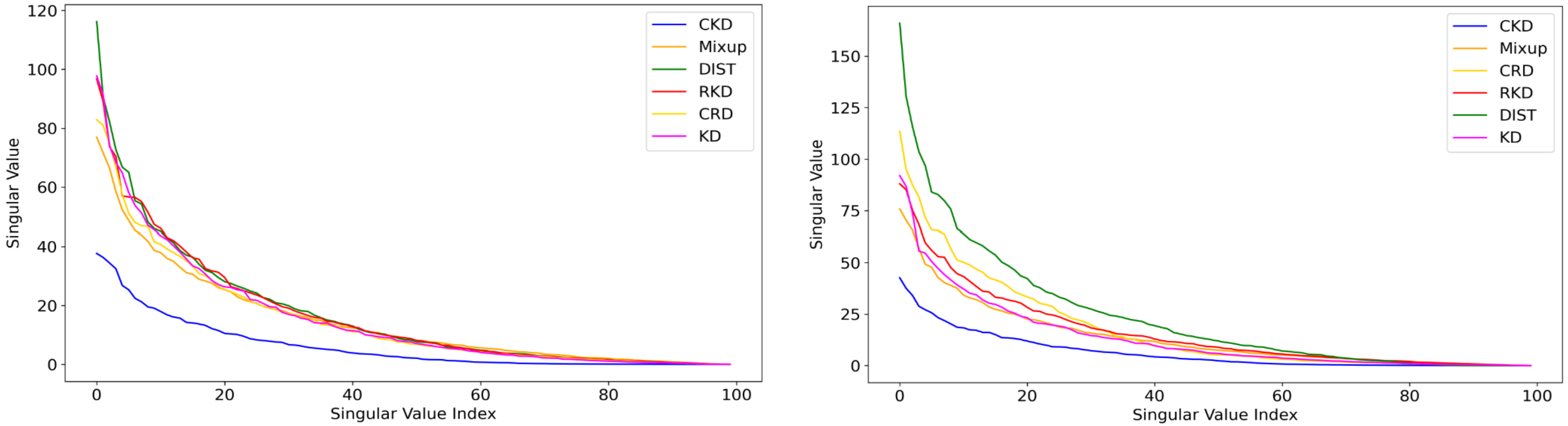
(a) VGG models



(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.

- **Experiments**
  - CKD가 regularized가 잘되었는지, SVD (singular Value Decomposition) 비교



(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].

- **Conclusion**
  - RTI-KD (Reduced Teacher Inference - Knowledge distillation) 분야의 가능성을 보여줌
  - Black Box KD 뿐만 아니라 네트워크 중간 단의 White Box에도 적용 가능함
  - Logit space의 regularization에 기여