南京航空航天大學

Nanjing University of Aeronautics and Astronautics

ParNeC 模式识别与神经计算研究组

PAttern Recognition and NEural Computing

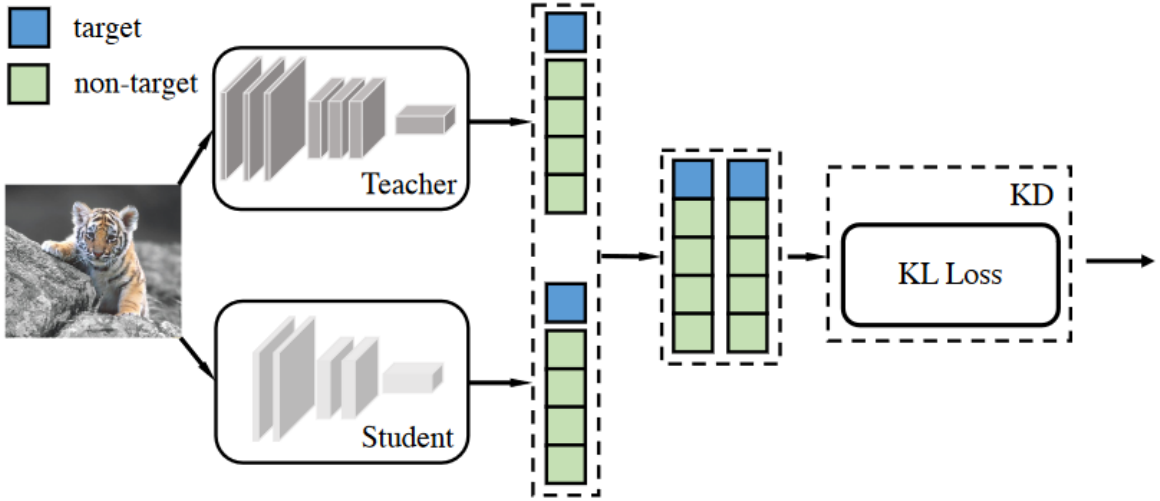
Decoupled Knowledge Distillation

Borui Zhao[1] Quan Cui[2] Renjie Song[1] Yiyu Qiu[1,3] Jiajun Liang[1]

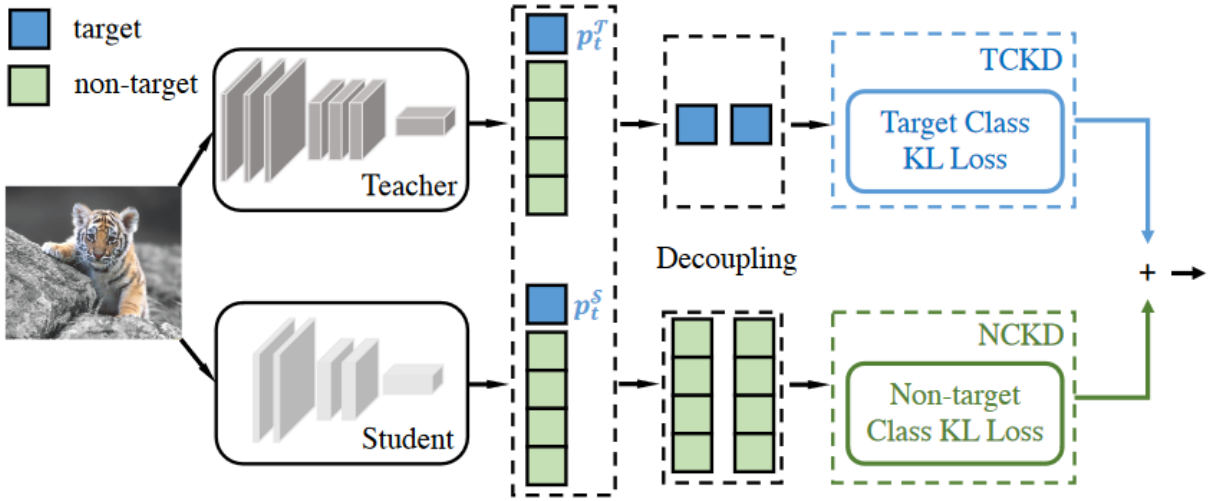
[1]MEGVII Technology [2]Waseda University [3]Tsinghua University

CVPR 2022

Decoupled Knowledge Distillation



(a) Classical Knowledge Distillation (KD).



$$\boldsymbol{Classical\ KD} = \boldsymbol{TCKD} + (\mathbf{1} - p_t^{\mathcal{T}}) * \boldsymbol{NCKD}$$

$$\boldsymbol{DKD(Ours)} = \boldsymbol{\alpha} * \boldsymbol{TCKD} + \boldsymbol{\beta} * \boldsymbol{NCKD}$$

(b) Decoupled Knowledge Distillation (DKD).

$$p_t = \frac{\exp(z_t)}{\sum_{j=1}^{C} \exp(z_j)}$$

$$p_{\backslash t} = \frac{\sum_{k=1,k\neq t}^{C} \exp(z_k)}{\sum_{j=1}^{C} \exp(z_j)}$$

$$\hat{p}_i = \frac{\exp(z_i)}{\sum_{j=1,j\neq t}^{C} \exp(z_j)}$$

$$\mathrm{KD} = \mathrm{KL}(\mathbf{p}^{\mathcal{T}} || \mathbf{p}^{\mathcal{S}})$$

$$= p_t^{\mathcal{T}} \log(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}) + \sum_{i=1,i\neq t}^{C} p_i^{\mathcal{T}} \log(\frac{p_i^{\mathcal{T}}}{p_i^{\mathcal{S}}})$$

$$= p_t^{\mathcal{T}} \log(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}) + p_{\backslash t}^{\mathcal{T}} \sum_{i=1,i\neq t}^{C} \hat{p}_i^{\mathcal{T}} (\log(\frac{\hat{p}_i^{\mathcal{T}}}{\hat{p}_i^{\mathcal{S}}}) + \log(\frac{p_{\backslash t}^{\mathcal{T}}}{p_{\backslash t}^{\mathcal{S}}}))$$

$$= \underbrace{p_t^{\mathcal{T}} \log(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}) + p_{\backslash t}^{\mathcal{T}} \log(\frac{p_{\backslash t}^{\mathcal{T}}}{p_{\backslash t}^{\mathcal{S}}})}_{\mathrm{KL}(\mathbf{b}^{\mathcal{T}} || \mathbf{b}^{\mathcal{S}})} + \underbrace{p_{\backslash t}^{\mathcal{T}} \sum_{i=1,i\neq t}^{C} \hat{p}_i^{\mathcal{T}} \log(\frac{\hat{p}_i^{\mathcal{T}}}{\hat{p}_i^{\mathcal{S}}})}_{\mathrm{KL}(\hat{\mathbf{p}}^{\mathcal{T}} || \hat{\mathbf{p}}^{\mathcal{S}})}$$

Decoupled Knowledge Distillation

$$
\begin{aligned}
\mathrm{KD} &= \mathrm{KL}(\mathbf{p}^{\mathcal{T}} || \mathbf{p}^{\mathcal{S}}) \\
&= p_t^{\mathcal{T}} \log(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}) + \sum_{i=1, i \neq t}^{C} p_i^{\mathcal{T}} \log(\frac{p_i^{\mathcal{T}}}{p_i^{\mathcal{S}}}) \\
&= p_t^{\mathcal{T}} \log(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}) + p_{\backslash t}^{\mathcal{T}} \sum_{i=1, i \neq t}^{C} \hat{p}_i^{\mathcal{T}} (\log(\frac{\hat{p}_i^{\mathcal{T}}}{\hat{p}_i^{\mathcal{S}}}) + \log(\frac{p_{\backslash t}^{\mathcal{T}}}{p_{\backslash t}^{\mathcal{S}}})) \\
&= \underbrace{p_t^{\mathcal{T}} \log(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}) + p_{\backslash t}^{\mathcal{T}} \log(\frac{p_{\backslash t}^{\mathcal{T}}}{p_{\backslash t}^{\mathcal{S}}})}_{\mathrm{KL}(\mathbf{b}^{\mathcal{T}} || \mathbf{b}^{\mathcal{S}})} + \underbrace{p_{\backslash t}^{\mathcal{T}} \sum_{i=1, i \neq t}^{C} \hat{p}_i^{\mathcal{T}} \log(\frac{\hat{p}_i^{\mathcal{T}}}{\hat{p}_i^{\mathcal{S}}})}_{\mathrm{KL}(\hat{\mathbf{p}}^{\mathcal{T}} || \hat{\mathbf{p}}^{\mathcal{S}})}
\end{aligned}
$$

Target Class Knowledge Distillation (TCKD)

Non-Target Class Knowledge Distillation(NCKD)

$$\mathrm{KD} = \mathrm{TCKD} + (1 - p_t^{\mathcal{T}})\mathrm{NCKD}$$

Effects of TCKD and NCKD

student	TCKD	NCKD	top-1	Δ
ResNet32×4 as the teacher				
ResNet8×4			72.50	-
	✓	✓	73.63	+1.13
	✓		68.63	-3.87
		✓	74.26	+1.76
ShuffleNet-V1			70.50	-
	✓	✓	74.29	+3.79
	✓		70.52	+0.02
		✓	74.91	+4.41
WRN-40-2 as the teacher				
WRN-16-2			73.26	-
	✓	✓	74.96	+1.70
	✓		70.96	-2.30
		✓	74.76	+1.50
ShuffleNet-V1			70.50	-
	✓	✓	74.92	+4.42
	✓		70.62	+0.12
		✓	75.12	+4.62

1. Applying Strong Augmentation

student	TCKD	top-1	Δ
ResNet8×4		73.82	-
	✓	75.33	+1.51
ShuffleNet-V1		77.13	-
	✓	77.98	+0.85

2. Noisy Labels

noisy ratio	TCKD	top-1	Δ
0.1		70.99	-
	✓	70.96	-0.03
0.2		67.55	-
	✓	68.03	+0.48
0.3		64.62	-
	✓	65.26	+0.64

3. Challenging Datasets(e.g., ImageNet)

TCKD	top-1	Δ
	70.71	-
✓	71.03	+0.32

Effects of TCKD and NCKD

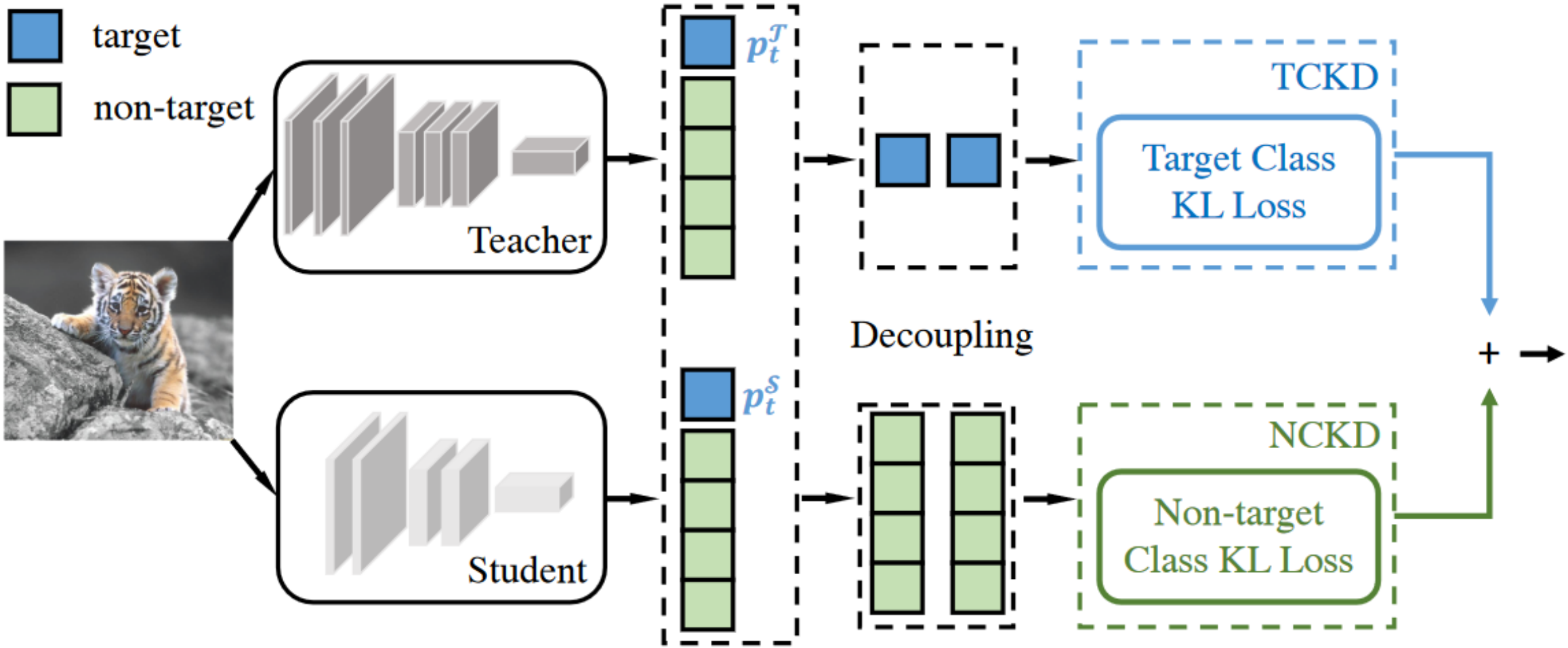
$$\mathrm{KD} = \mathrm{TCKD} + (1 - p_t^{\mathcal{T}})\mathrm{NCKD}$$

The loss weight of well-predicted samples are suppressed by the high confidence of the teacher.

To verify this, authors rank the training samples according to p_t^T , and evenly split them into two sub-sets. For clarity, one sub-set includes samples with top-50% p_t^T while remaining samples are in the other sub-set. Then they train student networks with NCKD on each subset to compare the performance gain (while the cross-entropy loss is still on the whole set).

0-50%	50-100%	top-1
✓	✓	74.26
✓		74.23
	✓	73.96

Decoupled Knowledge Distillation



$$Classical\ KD = TCKD + \left(1 - p_t^T\right) * NCKD$$

$$DKD(Ours) = \alpha * TCKD + \beta * NCKD$$

Multi-Label Knowledge Distillation

$$\text{if } i \in \boldsymbol{y},\ p_i = \frac{\sigma(z_i)}{\sum_{k=1}^{C}\sigma(z_i)} = \frac{\sigma(z_i)}{\sum_{k\in\boldsymbol{y}}\sigma(z_k)}\frac{\sum_{k\in\boldsymbol{y}}\sigma(z_k)}{\sum_{k=1}^{C}\sigma(z_k)} = \hat{p}_{t,i}\ p_t$$

$$\text{if } i \notin \boldsymbol{y},\ p_i = \frac{\sigma(z_i)}{\sum_{k=1}^{C}\sigma(z_i)} = \frac{\sigma(z_i)}{\sum_{k\notin\boldsymbol{y}}\sigma(z_k)}\frac{\sum_{k\notin\boldsymbol{y}}\sigma(z_k)}{\sum_{k=1}^{C}\sigma(z_k)} = \hat{p}_{\setminus t,i}\ p_{\setminus t}$$

$$\begin{aligned} D_{KL}(\boldsymbol{p}^T||\boldsymbol{p}^S) &= \sum_{i=1}^{N} p_i^T \log\frac{p_i^T}{p_i^S} \\ &= \sum_{i\in\boldsymbol{y}} p_i^T \log\frac{p_i^T}{p_i^S} + \sum_{i\notin\boldsymbol{y}} p_i^T \log\frac{p_i^T}{p_i^S} \\ &= p_t^T \log\frac{p_t^T}{p_t^S} + p_{\setminus t}^T \log\frac{p_{\setminus t}^T}{p_{\setminus t}^S} + p_t^T \sum_{i\in\boldsymbol{y}} \hat{p}_{t,i}^T \log\frac{\hat{p}_{t,i}^T}{\hat{p}_{t,i}^S} + p_{\setminus t}^T \sum_{i\notin\boldsymbol{y}} \hat{p}_{\setminus t,i}^T \log\frac{\hat{p}_{\setminus t,i}^T}{\hat{p}_{\setminus t,i}^S} \end{aligned}$$

Hyperparameter Selection

ResNet32×4 and ResNet8×4 are set as the teacher and the student, respectively.
Firstly, they prove that decoupling $(1 - p_t^T)$ and NCKD can bring reasonable performance gain (73.63% vs. 74.79%) in the first table.
Then, they demonstrate that decoupling weights of NCKD and TCKD could contribute to further improvements (74.79% vs. 76.32%).
Moreover, the second table indicates that TCKD is indispensable, and the improvements from TCKD are stable with different α around 1.0.

β	$1 - p_t^{\mathcal{T}}$	1.0	2.0	4.0	8.0	10.0
top-1	73.63	74.79	75.44	75.94	**76.32**	76.18

α	0.0	0.2	0.5	1.0	2.0	4.0
top-1	75.30	75.64	76.12	**76.32**	76.11	75.42

Experiments

results on the CIFAR-100 validation with teachers and students in the same architectures

distillation manner		ResNet56	ResNet110	ResNet32×4	WRN-40-2	WRN-40-2	VGG13
	teacher	72.34	74.31	79.42	75.61	75.61	74.64
	student	ResNet20	ResNet32	ResNet8×4	WRN-16-2	WRN-40-1	VGG8
		69.06	71.14	72.50	73.26	71.98	70.36
features	FitNet [28]	69.21	71.06	73.50	73.58	72.24	71.02
	RKD [23]	69.61	71.82	71.90	73.35	72.22	71.48
	CRD [33]	71.16	73.48	75.51	75.48	74.14	73.94
	OFD [10]	70.98	73.23	74.95	75.24	74.33	73.95
	ReviewKD [1]	71.89	73.89	75.63	76.12	**75.09**	**74.84**
logits	KD [12]	70.66	73.08	73.33	74.92	73.54	72.98
	DKD	**71.97**	**74.11**	**76.32**	**76.24**	74.81	74.68
	Δ	+1.31	+1.03	+2.99	+1.32	+1.27	+1.70

Table 6. **Results on the CIFAR-100 validation.** Teachers and students are in the **same** architectures. And Δ represents the performance improvement over the classical KD. All results are the average over 5 trials.

Experiments

results on the CIFAR-100 validation with teachers and students in different architectures

distillation manner	teacher	ResNet32×4	WRN-40-2	VGG13	ResNet50	ResNet32×4
		79.42	75.61	74.64	79.34	79.42
	student	ShuffleNet-V1	ShuffleNet-V1	MobileNet-V2	MobileNet-V2	ShuffleNet-V2
		70.50	70.50	64.60	64.60	71.82
features	FitNet [28]	73.59	73.73	64.14	63.16	73.54
	RKD [23]	72.28	72.21	64.52	64.43	73.21
	CRD [33]	75.11	76.05	69.73	69.11	75.65
	OFD [10]	75.98	75.85	69.48	69.04	76.82
	ReviewKD [1]	**77.45**	**77.14**	**70.37**	69.89	**77.78**
logits	KD [12]	74.07	74.83	67.37	67.35	74.45
	DKD	76.45	76.70	69.71	**70.35**	77.07
	Δ	+2.38	+1.87	+2.34	+3.00	+2.62

Table 7. **Results on the CIFAR-100 validation.** Teachers and students are in **different** architectures. And Δ represents the performance improvement over the classical KD. All results are the average over 5 trials.

Experiments

distillation manner			features				logits		
	teacher	student	AT [43]	OFD [10]	CRD [33]	ReviewKD [1]	KD [12]	KD*	**DKD**
top-1	73.31	69.75	70.69	70.81	71.17	71.61	70.66	71.03	**71.70**
top-5	91.42	89.07	90.01	89.98	90.13	**90.51**	89.88	90.05	90.41

Table 8. **Top-1 and top-5 accuracy (%) on the ImageNet validation.** We set **ResNet-34** as the teacher and **ResNet-18** as the student. KD* represents the result of our implementation. All results are the average over 3 trials.

distillation manner			features				logits		
	teacher	student	AT [43]	OFD [10]	CRD [33]	ReviewKD [1]	KD [12]	KD*	**DKD**
top-1	76.16	68.87	69.56	71.25	71.37	**72.56**	68.58	70.50	72.05
top-5	92.86	88.76	89.33	90.34	90.41	91.00	88.98	89.80	**91.05**

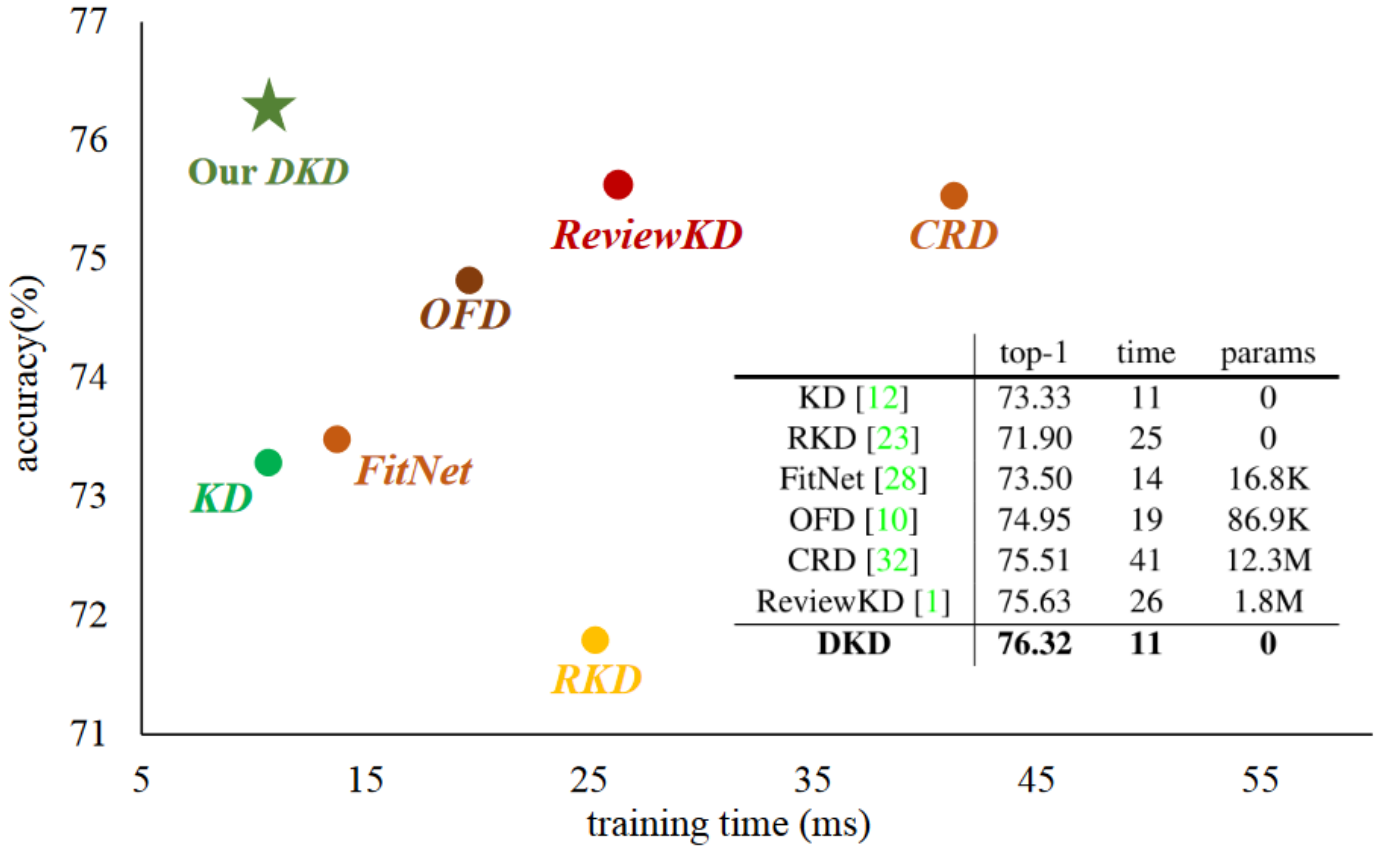
Table 9. **Top-1 and top-5 accuracy (%) on the ImageNet validation.** We set **ResNet-50** as the teacher and **MobileNet-V2** as the student. KD* represents the result of our implementation. All results are the average over 3 trials.

Experiments

	R-101 & R-18			R-101 & R-50			R-50 & MV2		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
teacher	42.04	62.48	45.88	42.04	62.48	45.88	40.22	61.02	43.81
student	33.26	53.61	35.26	37.93	58.84	41.05	29.47	48.87	30.90
KD [12]	33.97	54.66	36.62	38.35	59.41	41.71	30.13	50.28	31.35
FitNet [28]	34.13	54.16	36.71	38.76	59.62	41.80	30.20	49.80	31.69
FGFI [38]	35.44	55.51	38.17	39.44	60.27	43.04	31.16	50.68	32.92
ReviewKD [1]	36.75	56.72	34.00	40.36	60.97	44.08	33.71	53.15	36.13
DKD	35.05	56.60	37.54	39.25	60.90	42.73	32.34	53.77	34.01
DKD+ReviewKD	**37.01**	**57.53**	**39.85**	**40.65**	**61.51**	**44.44**	**34.35**	**54.89**	**36.61**

Table 10. **Results on MS-COCO based on Faster-RCNN [27]-FPN [19]**: AP evaluated on `val2017`. Teacher-student pairs are ResNet-101 (R-101) & ResNet-18 (R-18), ResNet-101 & ResNet-50 (R-50) and ResNet-50 & MobileNet-V2 (MV2) respectively. All results are the average over 3 trials. More details are attached in supplement.

Experiments-Training Efficiency



	top-1	time	params
KD [12]	73.33	11	0
RKD [23]	71.90	25	0
FitNet [28]	73.50	14	16.8K
OFD [10]	74.95	19	86.9K
CRD [32]	75.51	41	12.3M
ReviewKD [1]	75.63	26	1.8M
DKD	**76.32**	**11**	**0**

Figure 2. Training time (per batch) *vs*. accuracy on CIFAR-100. We set ResNet32×4 as the teacher and ResNet8×4 as the student. The table shows the number of extra parameters for each method.

Experiments-Big Teachers

teacher	W-28-2 75.45	W-40-2 75.61	W-16-4 77.51	W-28-4 78.60
KD	75.37	74.92	75.79	75.04
DKD	75.92	76.24	76.00	76.45

Table 11. Results on CIFAR-100. We set WRN-16-2 as the student and WRN series networks as teachers.

teacher	VGG13 74.64	WRN-16-4 77.51	ResNet50 79.34
KD	74.93	75.79	75.36
DKD	75.45	76.00	76.60

Table 12. Results on CIFAR-100. We set WRN-16-2 as the student and networks from different series as teachers.

Experiments-Feature Transferability

	baseline	KD	FitNet	CRD	ReviewKD	**DKD**
STL-10	69.7	70.9	70.3	71.6	72.4	**72.9**
TI	33.7	33.9	33.5	35.6	36.6	**37.1**

Table 13. **Comparison with previous methods on transferring features** from CIFAR-100 to STL-10 and Tiny-ImageNet (TI).

Experiments-Visualizations

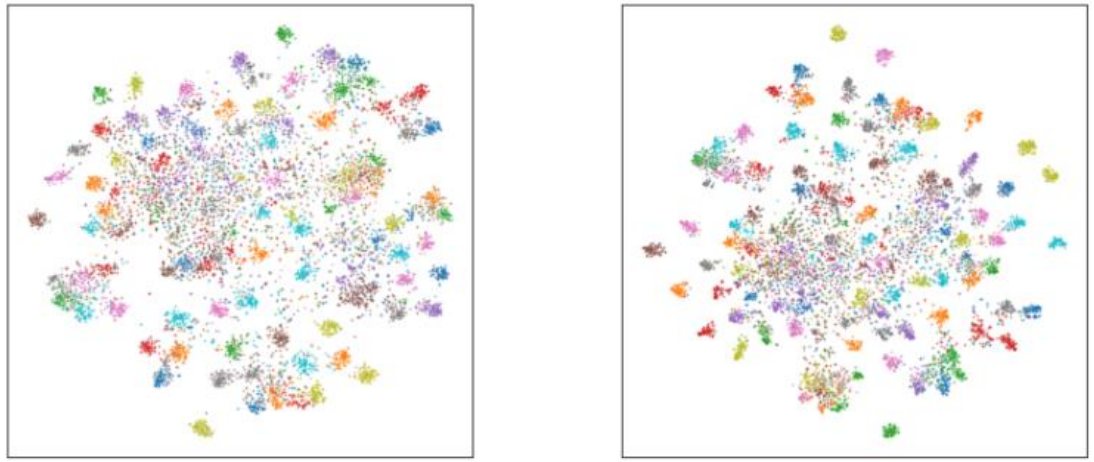


Figure 3. t-SNE of features learned by KD (left) and DKD (right).

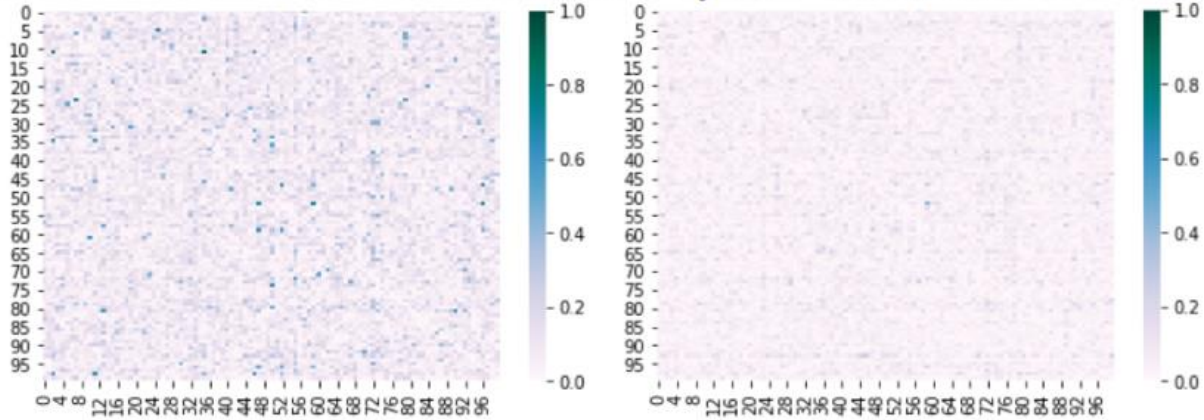


Figure 4. Difference of correlation matrices of student and teacher logits. Obviously, DKD (right) leads to a smaller difference (more similar prediction) than KD (left).