Relational Knowledge Distillation

CVPR 2019

Wonpyo Park[*]
POSTECH

Dongju Kim
POSTECH

Yan Lu
Microsoft Research

Minsu Cho
POSTECH

Wonbeom Jang

Previous methods

- Distilling the Knowledge in a Neural Network Hinton et al. In NIPS, 2014.

$$\sum_{x_i \in \mathcal{X}} \mathrm{KL}\Big(\mathrm{softmax}(\frac{f_T(x_i)}{\tau}), \mathrm{softmax}(\frac{f_S(x_i)}{\tau})\Big)$$

- FitNets: Hints for Thin Deep Nets Romero et al. In ICLR, 2015

$$\sum_{x_i \in \mathcal{X}} \big\| f_T(x_i) - \beta\big(f_S(x_i)\big) \big\|_2^2$$

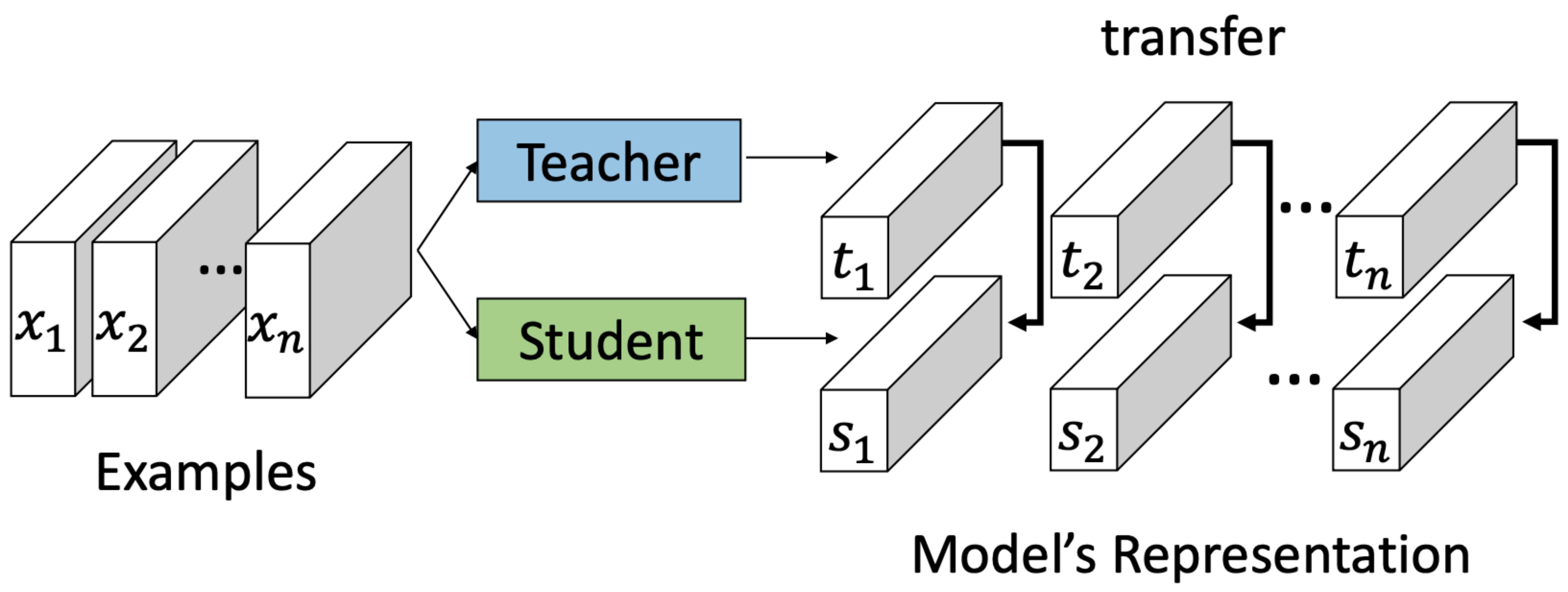
Previous methods

- Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer

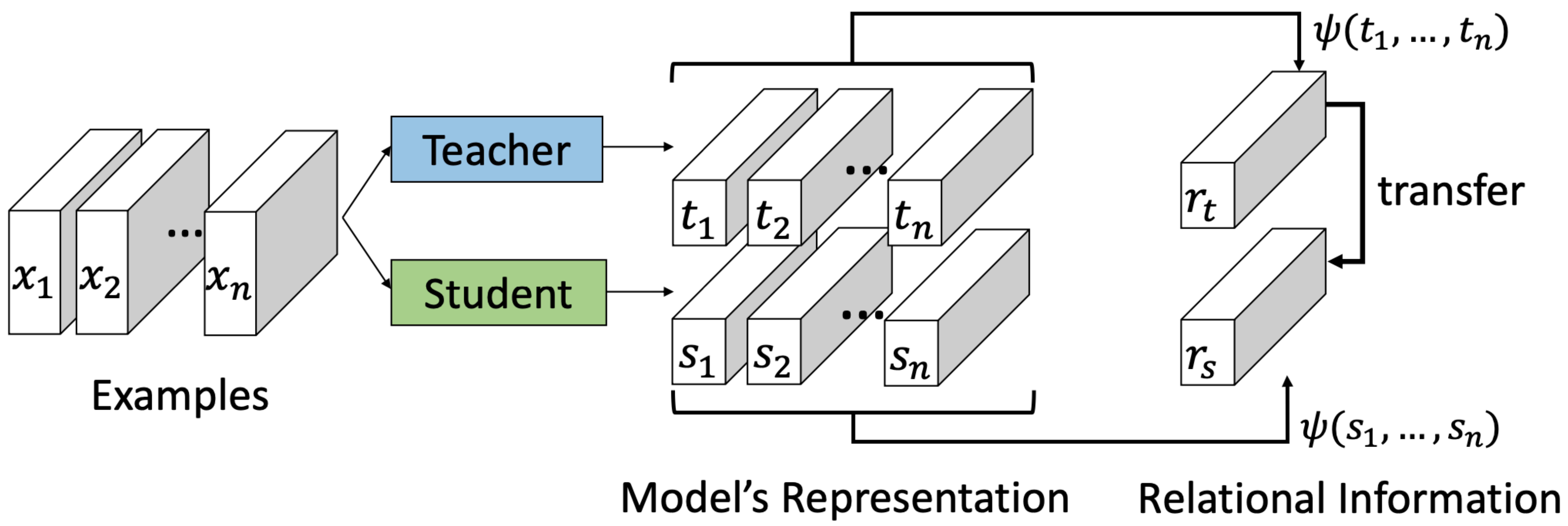
$$\sum_{x_i \in \mathcal{X}} \left\| \frac{Q_T^i}{\|Q_T^i\|} - \frac{Q_S^i}{\|Q_S^i\|} \right\|_2$$

- Label Refinery: Improving ImageNet Classification through Label Progression (Bagherinezhad et al. In arXiv, 2018.)

Previous methods



Individual Knowledge Distillation



Relational Knowledge Distillation

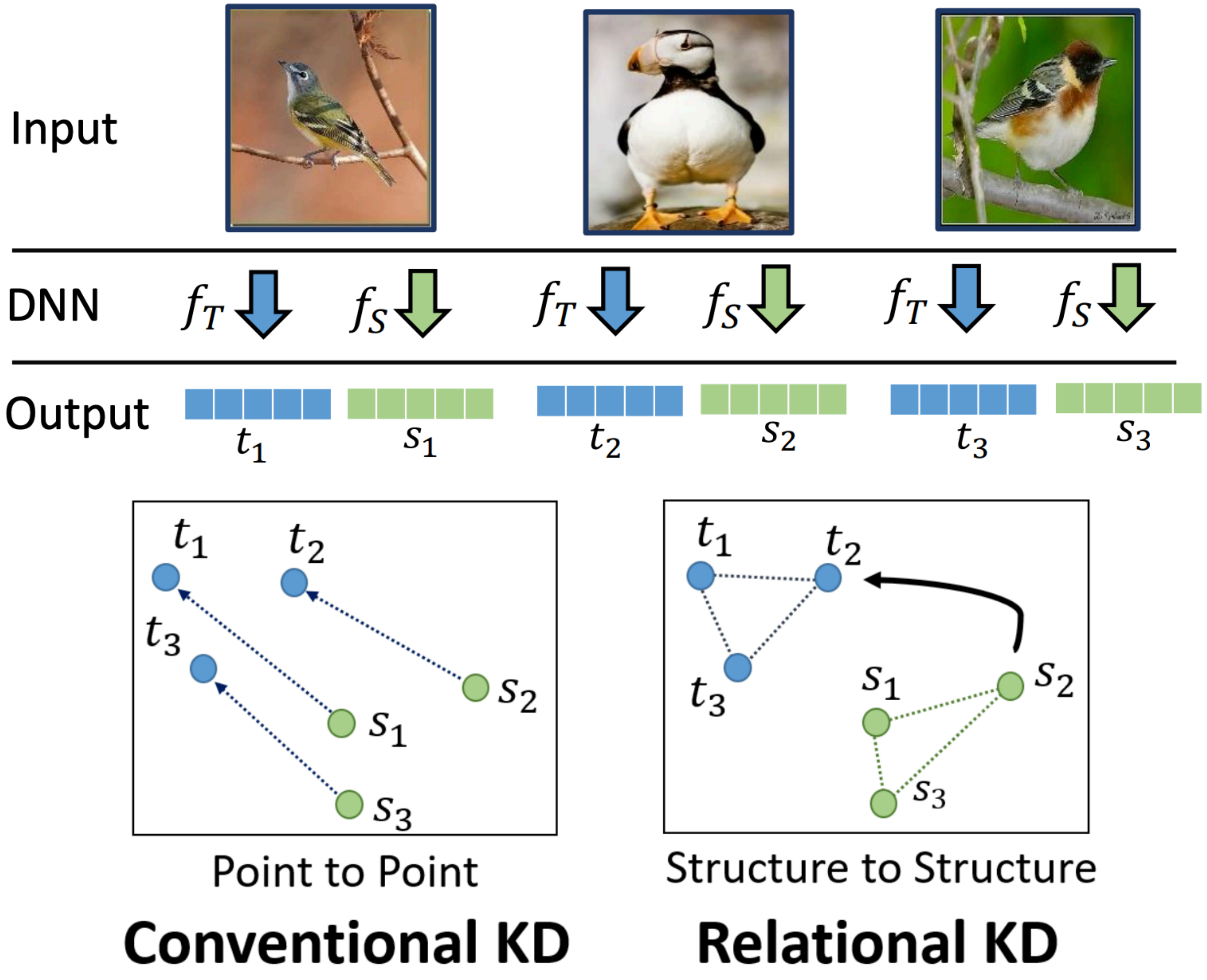
$$\mathcal{L}_{\mathrm{IKD}} = \sum_{x_i \in \mathcal{X}} l\big(f_T(x_i), f_S(x_i)\big), \qquad (1)$$

$$\mathcal{L}_{\mathrm{RKD}} = \sum_{(x_1,..,x_n) \in \mathcal{X}^N} l\big(\psi(t_1,..,t_n), \psi(s_1,..,s_n)\big), \qquad (4)$$

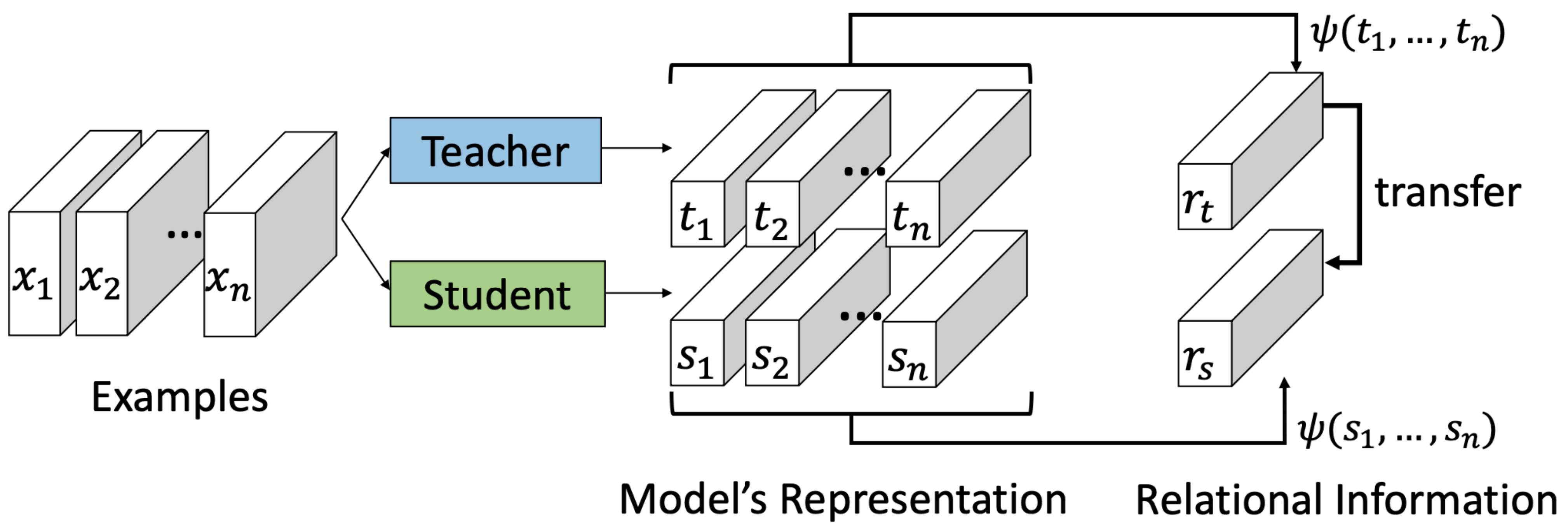
transfers individual outputs of the teacher directly to the student

distance-wise and angle-wise distillation losses that penalize structural differences in relations.

Previous methods



Relational knowledge distillation



Relational Knowledge Distillation

$$\mathcal{L}_{\text{RKD}} = \sum_{(x_1,..,x_n)\in\mathcal{X}^N} l\big(\psi(t_1,..,t_n), \psi(s_1,..,s_n)\big), \quad (4)$$

$$\mathcal{L}_{\text{task}} + \lambda_{\text{KD}} \cdot \mathcal{L}_{\text{KD}}, \quad (11)$$

$t_i = f_T(x_i)$

$s_i = f_S(x_i)$

$(x_1, x_2, \ldots, x_n)$ is a n-tuple drawn from $\mathcal{X}^N$

ψ is a relational potential function that measures a relational energy of the given n-tuple

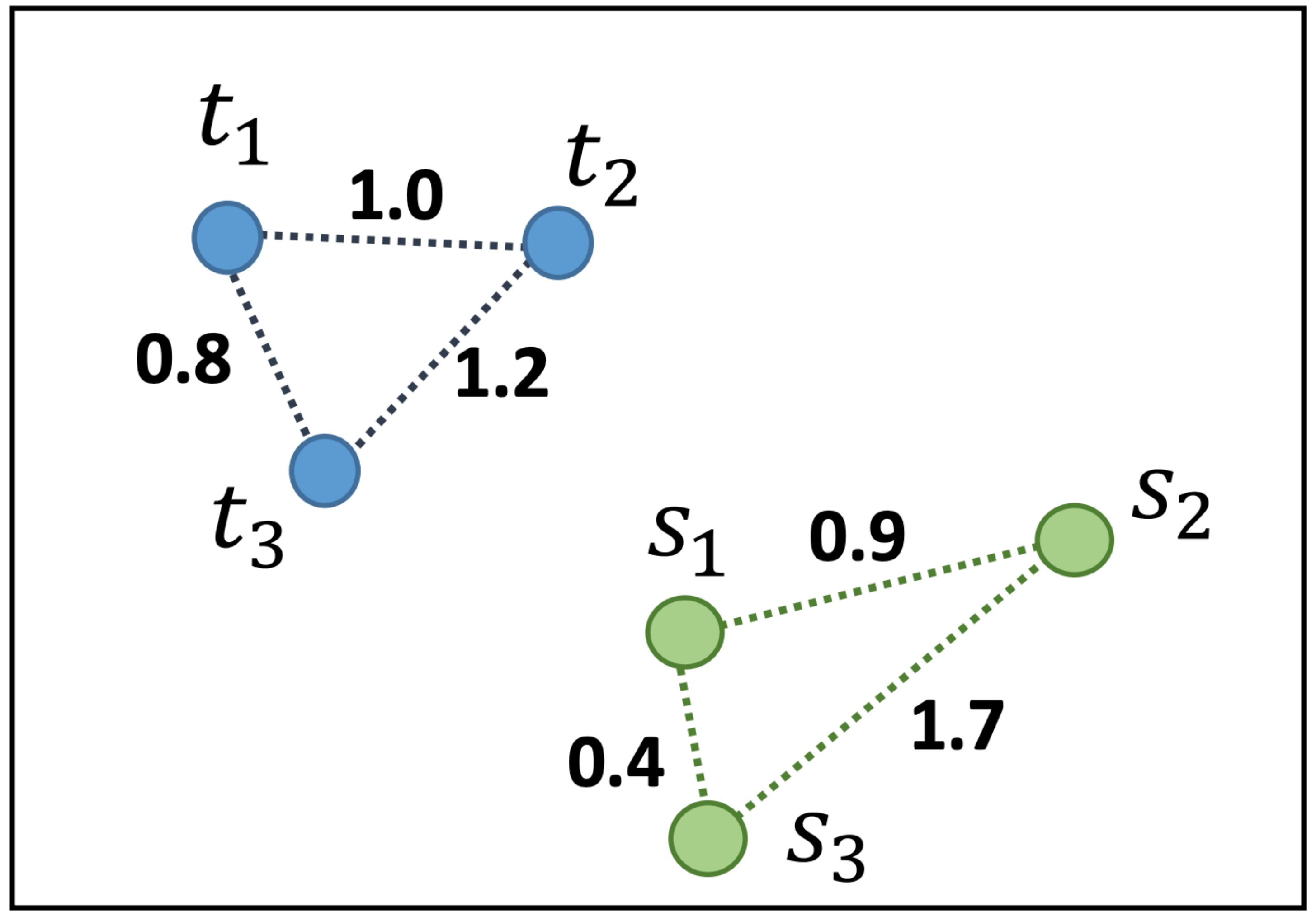
Distance-wise distillation loss

$$\psi_{\mathrm{D}}(t_i, t_j) = \frac{1}{\mu} \left\| t_i - t_j \right\|_2 , \tag{5}$$

$$\mu = \frac{1}{|\mathcal{X}^2|} \sum_{(x_i, x_j) \in \mathcal{X}^2} \left\| t_i - t_j \right\|_2 . \tag{6}$$

$$\mathcal{L}_{\text{RKD-D}} = \sum_{(x_i, x_j) \in \mathcal{X}^2} l_\delta\big(\psi_{\mathrm{D}}(t_i, t_j), \psi_{\mathrm{D}}(s_i, s_j)\big), \tag{7}$$

$$l_\delta(x, y) = \begin{cases} \frac{1}{2}(x-y)^2 & \text{for } |x-y| \le 1, \\ |x-y| - \frac{1}{2}, & \text{otherwise.} \end{cases} \tag{8}$$



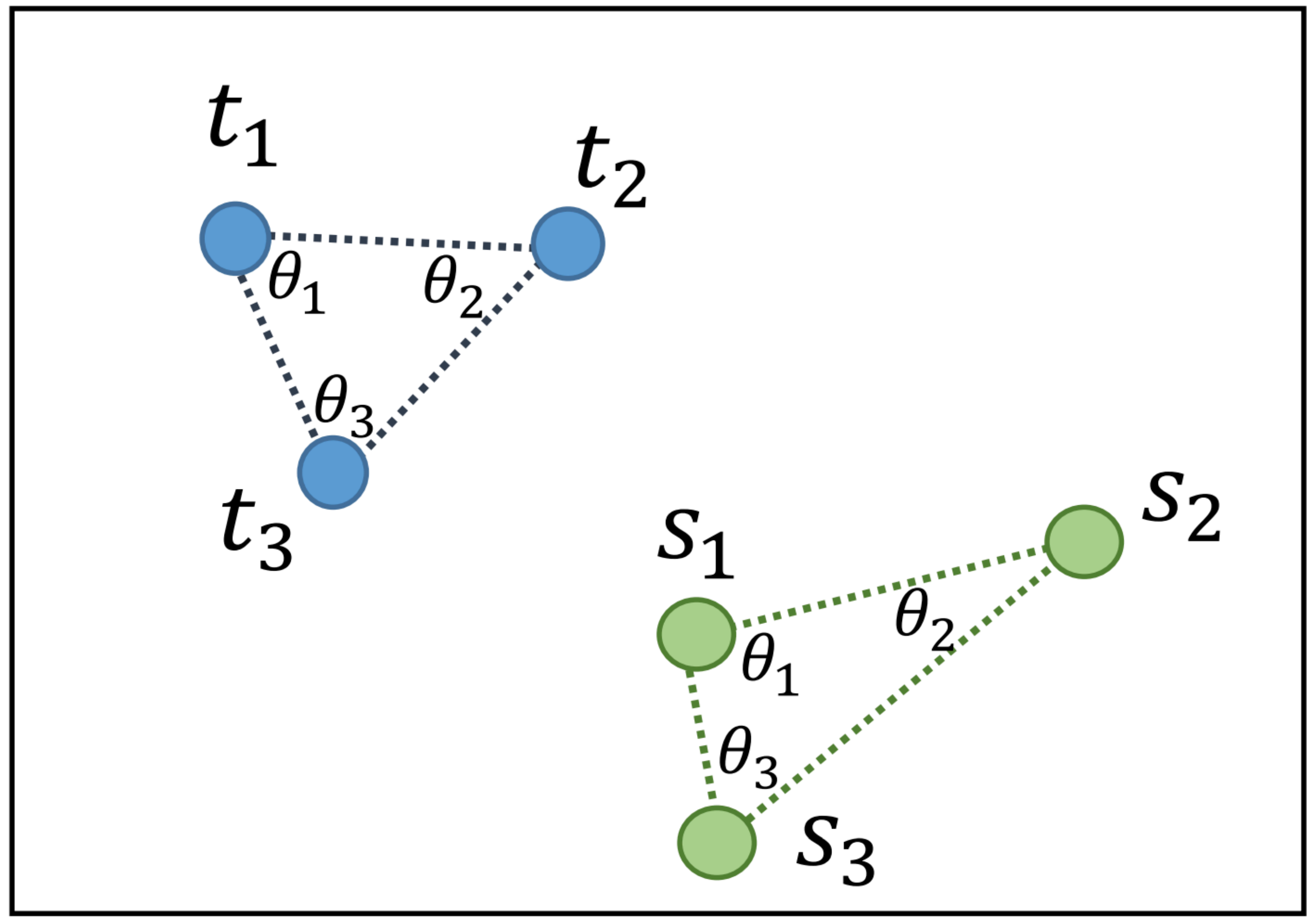
Embedding Space

Angle-wise distillation loss

$$\psi_{\mathrm{A}}(t_i, t_j, t_k) = \cos \angle t_i t_j t_k = \langle \mathbf{e}^{ij}, \mathbf{e}^{kj} \rangle \qquad (9)$$

$$\text{where} \quad \mathbf{e}^{ij} = \frac{t_i - t_j}{\|t_i - t_j\|_2}, \mathbf{e}^{kj} = \frac{t_k - t_j}{\|t_k - t_j\|_2}.$$

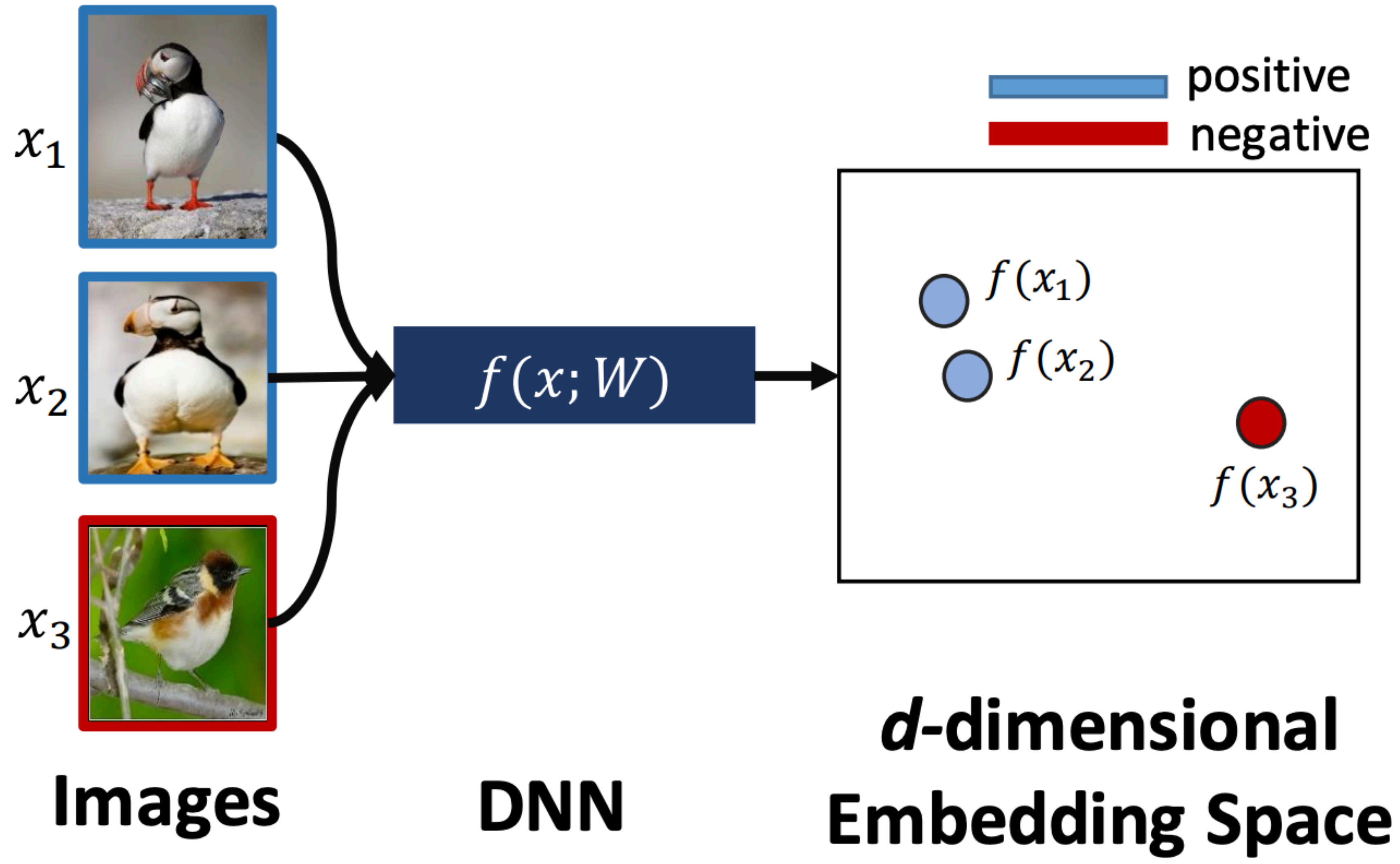
$$\mathcal{L}_{\text{RKD-A}} = \sum_{(x_i, x_j, x_k) \in \mathcal{X}^3} l_\delta \big(\psi_{\mathrm{A}}(t_i, t_j, t_k), \psi_{\mathrm{A}}(s_i, s_j, s_k)\big), \qquad (10)$$



Embedding Space

Experiments

Metric learning



$$\mathcal{L}_{\text{triplet}} = \left[\|f(x_a) - f(x_p)\|_2^2 - \|f(x_a) - f(x_n)\|_2^2 + m \right]_+ . \tag{12}$$

- It aims to train an embedding model.
- In embedding space, distances between projected examples correspond to their semantic similarity.

Experiments

Metric learning

(a) Results on CUB-200-2011 [40]

	Baseline (Triplet [31])	FitNet [27]	Attention [47]	DarkRank [7]	Ours RKD-D	Ours RKD-A	Ours RKD-DA
$\ell 2$ normalization	O	O	O	O	O / X	O / X	O / X
ResNet18-16	37.71	42.74	37.68	46.84	46.34 / 48.09	45.59 / **48.60**	45.76 / 48.14
ResNet18-32	44.62	48.60	45.37	53.53	52.68 / **55.72**	53.43 / 55.15	53.58 / 54.88
ResNet18-64	51.55	51.92	50.81	56.30	56.92 / 58.27	56.77 / 58.44	57.01 / **58.68**
ResNet18-128	53.92	54.52	55.03	57.17	58.31 / 60.31	58.41 / **60.92**	59.69 / 60.67
ResNet50-512	61.24						

(b) Results on Cars 196 [14]

	Baseline (Triplet [31])	FitNet [27]	Attention [47]	DarkRank [7]	Ours RKD-D	Ours RKD-A	Ours RKD-DA
$\ell 2$ normalization	O	O	O	O	O / X	O / X	O / X
ResNet18-16	45.39	57.46	46.44	64.00	63.23 / 66.02	61.39 / **66.25**	61.78 / 66.04
ResNet18-32	56.01	65.81	59.40	72.41	73.50 / **76.15**	73.23 / 75.89	73.12 / 74.80
ResNet18-64	64.53	70.67	67.24	76.20	78.64 / **80.57**	77.92 / 80.32	78.48 / 80.17
ResNet18-128	68.79	73.10	71.95	77.00	79.72 / 81.70	80.54 / 82.27	80.00 / **82.50**
ResNet50-512	77.17						

Experiments

Metric learning

Table 2: Recall@1 of self-distilled models. Student and teacher models have the same architecture. The model at Gen(n) is guided by the model at Gen(n-1).

	CUB [40]	Cars [14]	SOP [21]
ResNet50-512-Triplet	61.24	77.17	76.58
ResNet50-512@Gen1	**65.68**	**85.65**	**77.61**
ResNet50-512@Gen2	65.11	85.61	77.36
ResNet50-512@Gen3	64.26	85.23	76.96

Experiments

Metric learning

RKD performing better without l2 normalization.

L2 norm forces out- put points of an embedding model to lie on the surface of unit-hypersphere, and thus a student model without l2 norm is able to fully utilize the embedding space.

Ours		
RKD-D	RKD-A	RKD-DA
O / X	O / X	O / X
46.34 / 48.09	45.59 / **48.60**	45.76 / 48.14
52.68 / **55.72**	53.43 / 55.15	53.58 / 54.88
56.92 / 58.27	56.77 / 58.44	57.01 / **58.68**
58.31 / 60.31	58.41 / **60.92**	59.69 / 60.67

Ours		
RKD-D	RKD-A	RKD-DA
O / X	O / X	O / X
63.23 / 66.02	61.39 / **66.25**	61.78 / 66.04
73.50 / **76.15**	73.23 / 75.89	73.12 / 74.80
78.64 / **80.57**	77.92 / 80.32	78.48 / 80.17
79.72 / 81.70	80.54 / 82.27	80.00 / **82.50**

Experiments

Metric learning

Students excelling teachers.

Continuous target la- bels of RKD (*e.g.*, distance or angle) may also carry useful information, which cannot properly be encoded in binary (positive/negative) ground-truth labels used in conventional losses, *i.e.*, the triplet loss.

	Baseline (Triplet [31])	FitNet [27]	Attention [47]	DarkRank [7]	Ours RKD-D	Ours RKD-A	Ours RKD-DA
$\ell 2$ normalization	O	O	O	O	O / X	O / X	O / X
ResNet18-16	45.39	57.46	46.44	64.00	63.23 / 66.02	61.39 / **66.25**	61.78 / 66.04
ResNet18-32	56.01	65.81	59.40	72.41	73.50 / **76.15**	73.23 / 75.89	73.12 / 74.80
ResNet18-64	64.53	70.67	67.24	76.20	78.64 / **80.57**	77.92 / 80.32	78.48 / 80.17
ResNet18-128	68.79	73.10	71.95	77.00	79.72 / 81.70	80.54 / 82.27	80.00 / **82.50**
ResNet50-512	77.17						

Experiments

Metric learning

RKD as a training domain adaptation.

These results reveal an interesting effect of RKD that it strongly adapts models on the training domain at the cost of sacrificing generalization to other domains.

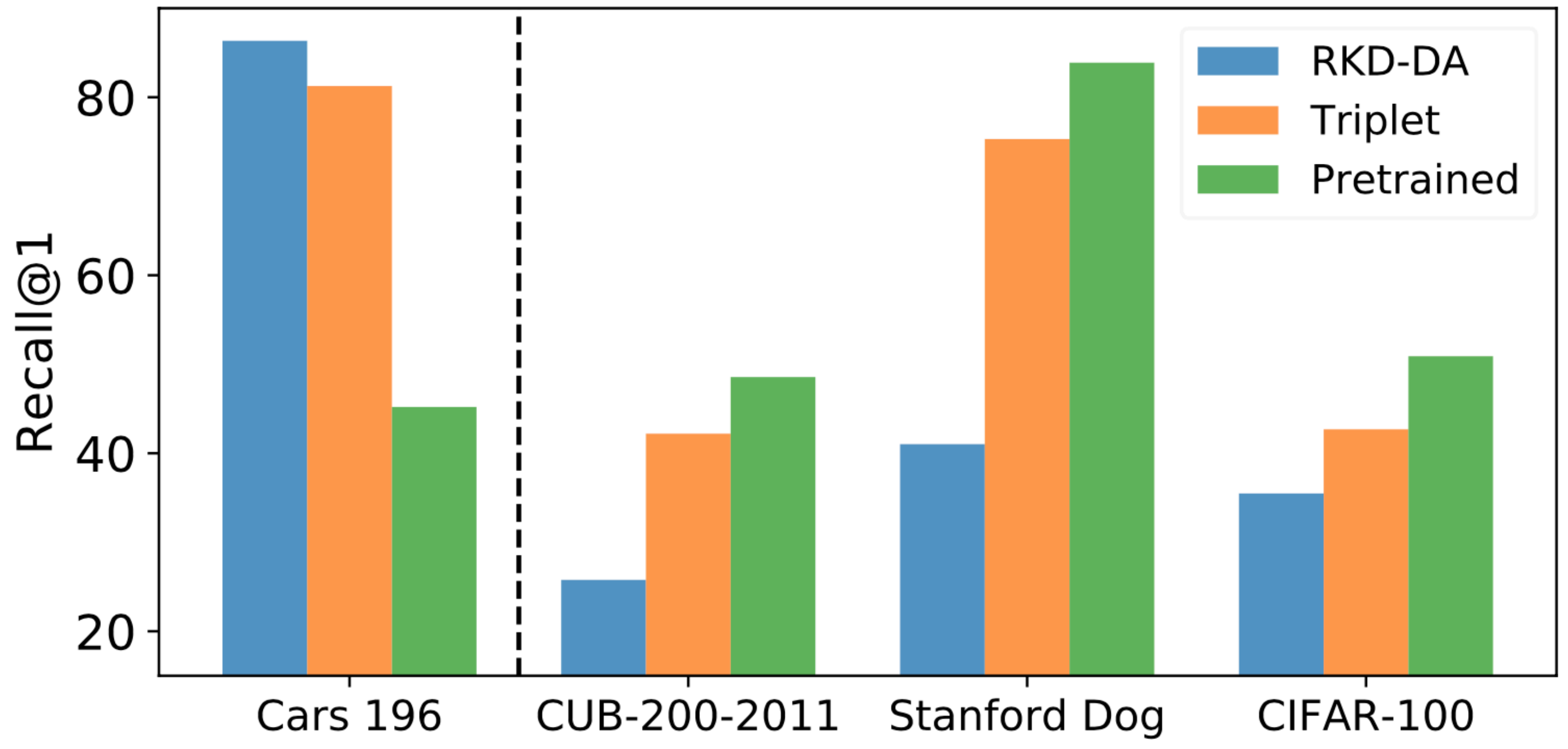


Figure 3: Recall@1 on the test split of Cars 196, CUB-200-2011, Stanford Dog and CIFAR-100. Both Triplet (teacher) and RKD-DA (student) are trained on Cars 196. The left side of the dashed line shows results on the training domain, while the right side presents results on other domains.

Experiments

Metric learning

	K	CUB-200-2011 [40] 1	2	4	8	Cars 196 [14] 1	2	4	8	Stanford Online Products [21] 1	10	100	1000
GoogLeNet [35]	LiftedStruct [21]-128	47.2	58.9	70.2	80.2	49.0	60.3	72.1	81.5	62.1	79.8	91.3	97.4
	N-pairs [34]-64	51.0	63.3	74.3	83.2	71.1	79.7	86.5	91.6	67.7	83.8	93.0	97.8
	Angular [41]-512	54.7	66.3	76.0	83.9	71.4	81.4	87.5	92.1	70.9	85.0	93.5	98.0
	A-BIER [22]-512	57.5	68.7	78.3	86.2	82.0	89.0	93.2	96.1	74.2	86.9	94.0	97.8
	ABE8 [13]-512	60.6	71.5	79.8	87.4	**85.2**	**90.5**	94.0	96.1	**76.3**	**88.4**	94.8	98.2
	RKD-DA-128	60.8	72.1	81.2	**89.2**	81.7	88.5	93.3	96.3	74.5	88.1	95.2	98.6
	RKD-DA-512	**61.4**	**73.0**	**81.9**	89.0	82.3	89.8	**94.2**	**96.6**	75.1	88.3	**95.2**	**98.7**
ResNet50 [10]	Margin [42]-128	63.6	74.4	83.1	90.0	79.6	86.5	91.9	95.1	72.7	86.2	93.8	98.0
	RKD-DA-128	**64.9**	**76.7**	**85.3**	**91.0**	**84.9**	**91.3**	**94.8**	**97.2**	**77.5**	**90.3**	**96.4**	**99.0**

Experiments

Image classification

Table 4: Accuracy (%) on CIFAR-100 and Tiny ImageNet.

	CIFAR-100 [15]	Tiny ImageNet [46]
Baseline	71.26	54.45
RKD-D	72.27	54.97
RKD-DA	72.97	56.36
HKD [11]	74.26	57.65
HKD+RKD-DA	**74.66**	**58.15**
FitNet [27]	70.81	55.59
FitNet+RKD-DA	72.98	55.54
Attention [47]	72.68	55.51
Attention+RKD-DA	73.53	56.55
Teacher	77.76	61.55

Experiments

Few-shot Learning

As the prototypical networks build on shallow net- works that consist of only 4 convolutional layers, we use the same architecture for the student model and the teacher, *i.e.*, self-distillation, rather than using a smaller student net- work.

Table 5: Accuracy (%) on Omniglot [16].

	5-Way Acc.		20-Way Acc.	
	1-Shot	5-Shot	1-Shot	5-Shot
RKD-D	98.58	**99.65**	95.45	**98.72**
RKD-DA	**98.64**	99.64	**95.52**	98.67
Teacher	98.55	99.56	95.11	98.68

Table 6: Accuracy (%) on *mini*ImageNet [39].

	1-Shot 5-Way	5-Shot 5-Way
RKD-D	49.66 ± 0.84	67.07 ± 0.67
RKD-DA	50.02 ± 0.83	**68.16** ± 0.67
FitNet	**50.38** ± 0.81	68.08 ± 0.65
Attention	34.67 ± 0.65	46.21 ± 0.70
Teacher	49.1 ± 0.82	66.87 ± 0.66