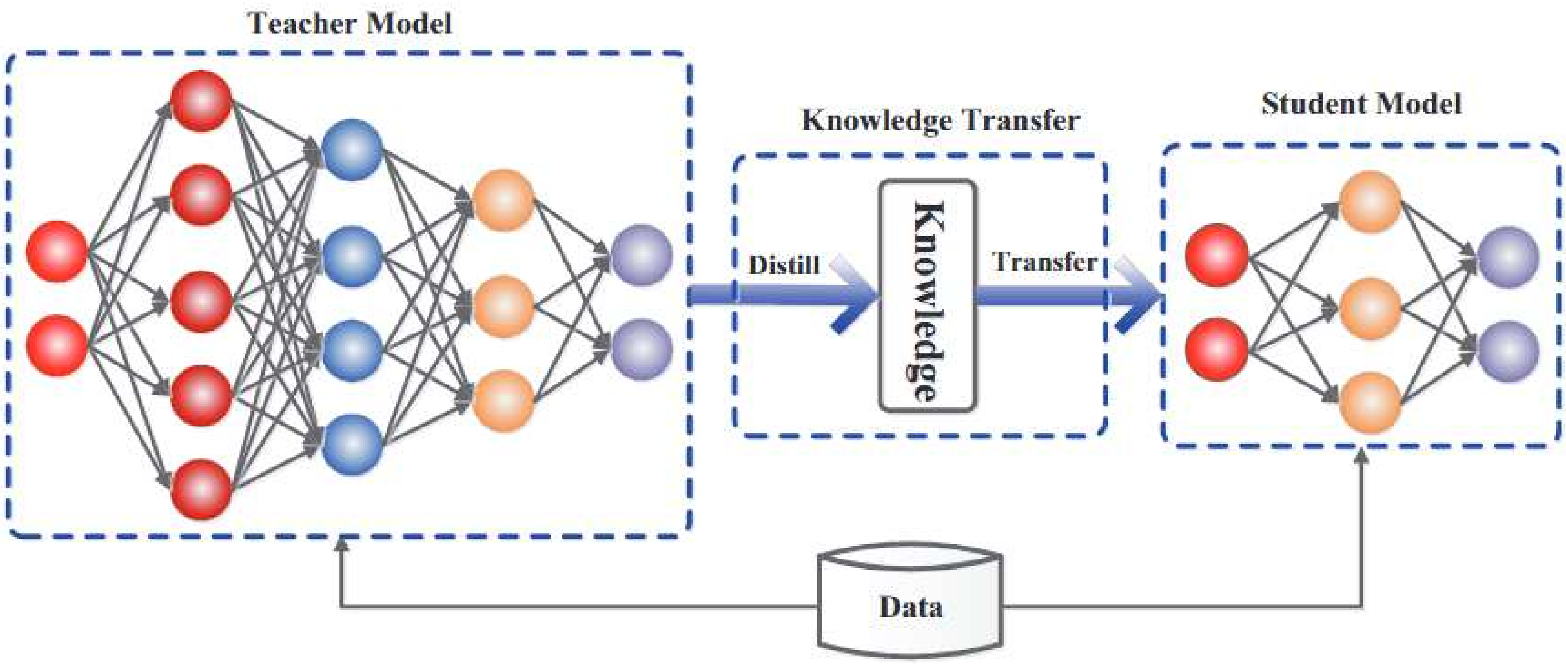
南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

TC3KD: Knowledge distillation via teacher-student cooperative curriculum customization

Chaofei Wang [a], Ke Yang [b], Shaowei Zhang [c], Gao Huang [a], Shiji Song [a,*]

[a] *Department of Automation, Tsinghua University, Beijing, China*
[b] *College of Chemistry, Beijing University of Chemical Technology, Beijing, China*
[c] *School of Microelectronics, Tianjin University, Tianjin, China*

Knowledge Distillation



Knowledge and Distillation

[1] Gou J, Yu B, Maybank S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

Knowledge

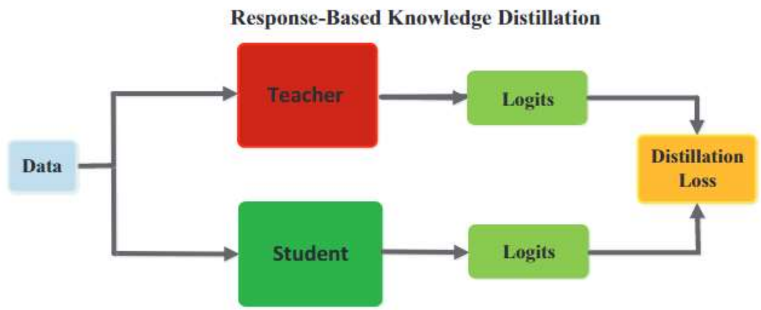


Fig. 4 The generic response-based knowledge distillation.

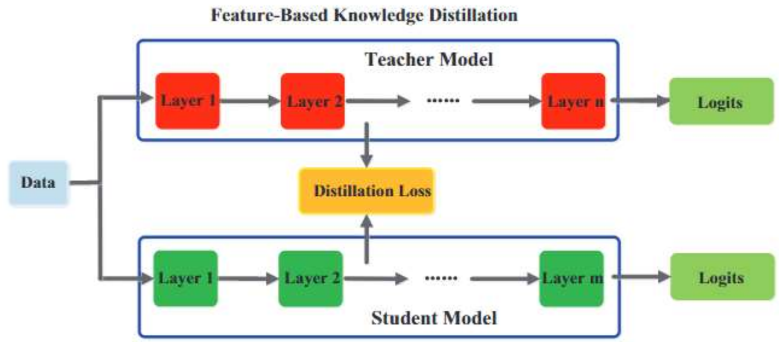


Fig. 6 The generic feature-based knowledge distillation.

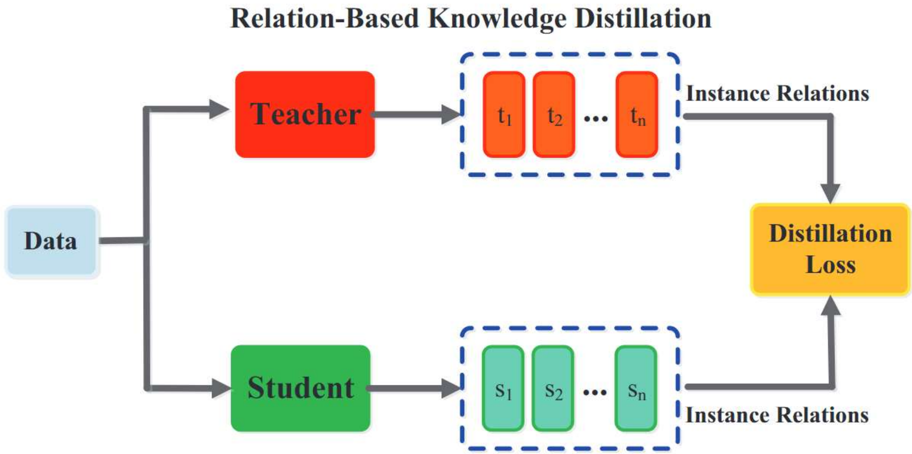


Fig. 7 The generic instance relation-based knowledge distillation.

[1] Gou J, Yu B, Maybank S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

Distillation

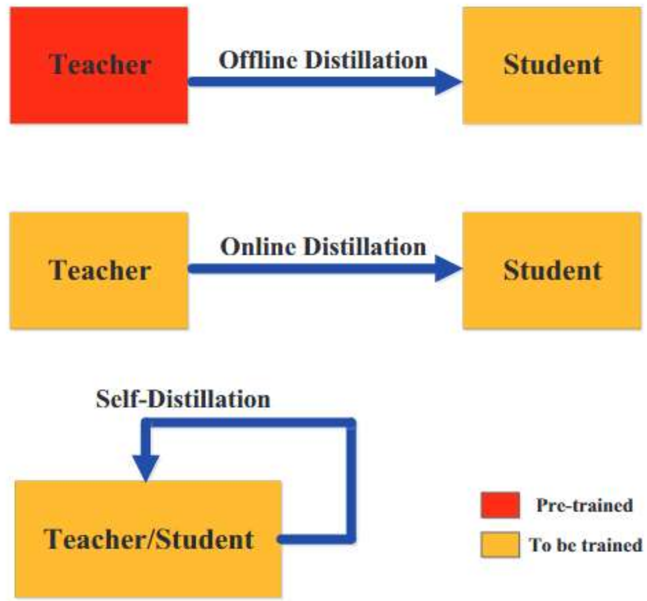


Fig. 8 Different distillations. The red color for "pre-trained" means networks are learned before distillation and the yellow color for "to be trained" means networks are learned during distillation

[1] Gou J, Yu B, Maybank S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Knowledge Distillation

cross entropy loss:

$$L(B,\theta) = -\frac{1}{|B|}\sum_{(x_i,y_i)\in B} y_i^T \cdot \log f(x_i;\theta)$$

total loss:

$$L^s(B,\theta^s) = -\frac{1}{|B|}\sum_{(x_i,y_i)\in B}\left\{\lambda y_i^T \cdot \log f(x_i;\theta^s) + (1-\lambda)\mathrm{KL}\left[f^\tau(x_i;\theta^t)\,||\,f^\tau(x_i;\theta^s)\right]\right\},$$

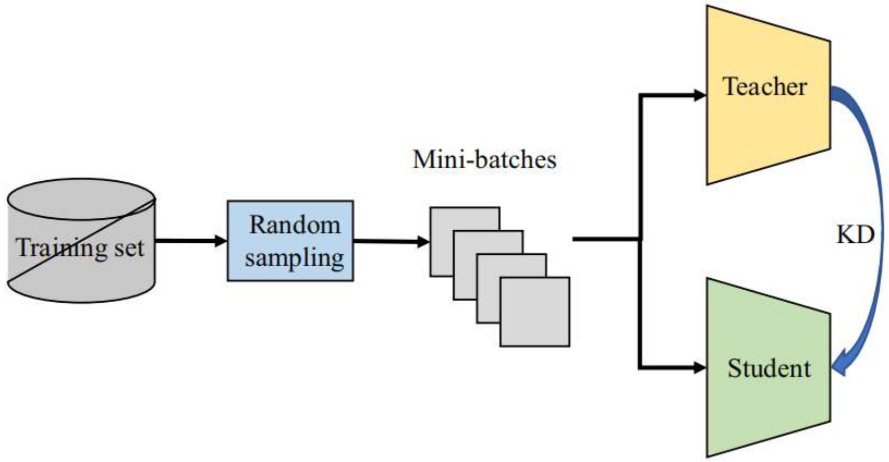


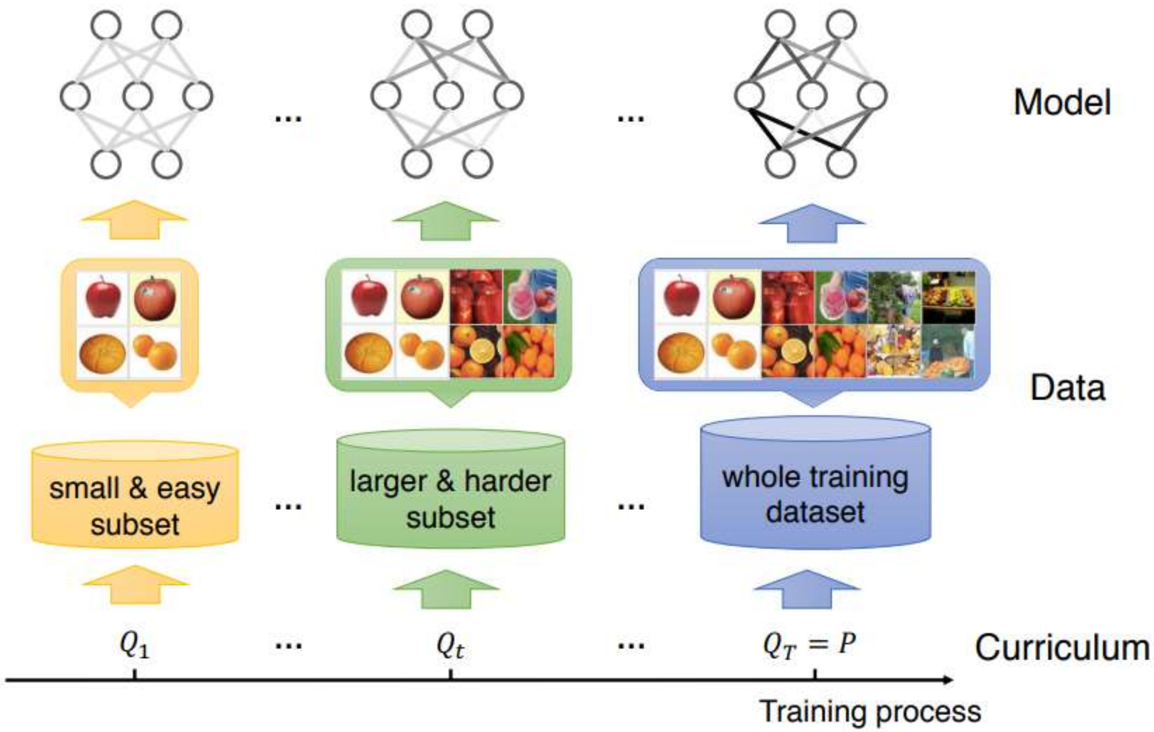
Fig. 2. General framework of knowledge distillation.

[2]Wang, Chaofei, et al. "TC3KD: Knowledge distillation via teacher-student cooperative curriculum customization." Neurocomputing 508 (2022): 284-292.

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Curriculum Learning

ranking function:
1、Difficulty Measurer
2、Training Scheduler



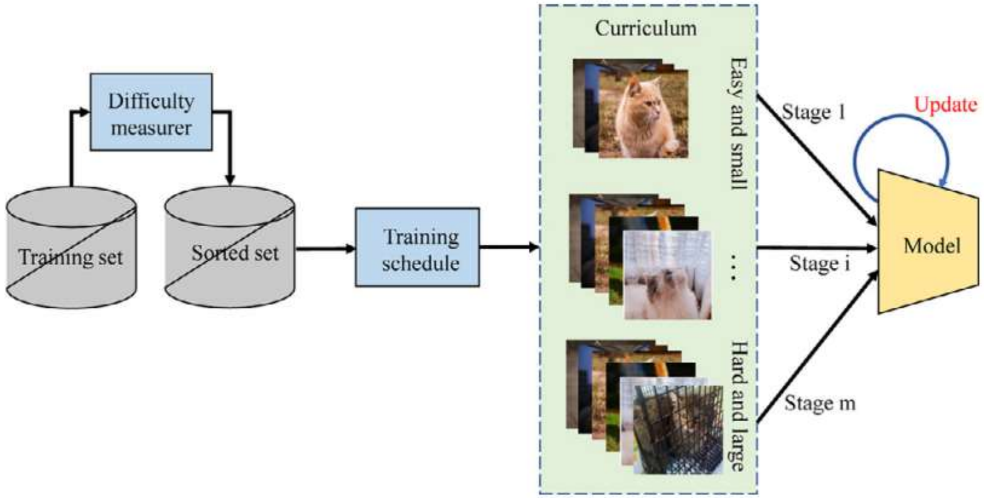


Fig. 3. Common paradigm of curriculum learning.

[3]Wang, Xin, Yudong Chen, and Wenwu Zhu. "A survey on curriculum learning." IEEE Transactions on Pattern Analysis and Machine Intelligence 44.9 (2021): 4555-4576.

Curriculum Learning

1、Difficulty Measurer： $g(x_i, y_i) > g(x_j, y_j)$

2、Training Scheduler（Baby Step算法）

$$h : D' \rightarrow \{S_1, S_2, S_3 \cdots, S_m\}$$

$$S_1 \subseteq S_2 \subseteq S_3 \subseteq \cdots \subseteq S_m = D'$$

Algorithm:1 General curriculum learning method

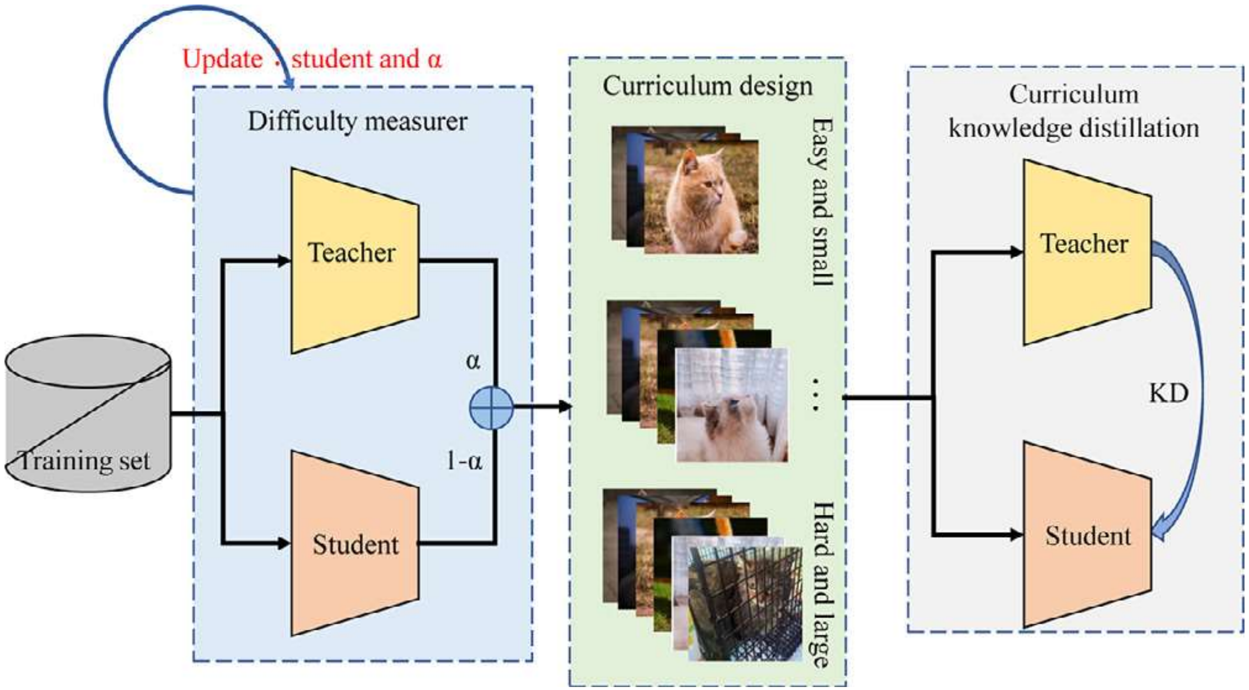
Input: M: initialized model; D: training dataset; g: difficulty measurer; h: training scheduler; m the number of subsets;
Output: M^*: optimal model;

```
D' = g(D);
{S_1, S_2, S_3 ..., S_m} = h(D');
for i = 1, 2, ..., m do
    while the model M does not converge do
        train M with S_i;
    end while
end for
```

[2]Wang, Chaofei, et al. "TC3KD: Knowledge distillation via teacher-student cooperative curriculum customization." Neurocomputing 508 (2022): 284-292.

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

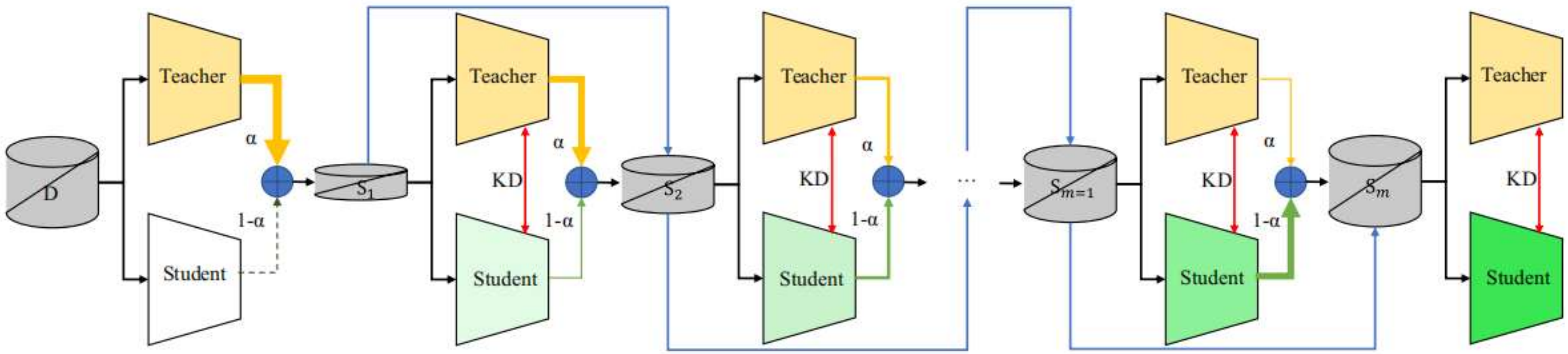
TC3KD: Knowledge distillation via teacher-student cooperative curriculum customization



贡献点:
1、课程学习引入到知识蒸馏中，并利用师生共同进行难度度量
2、在师生共同进行难度度量时，作者提出一个动态的权重设置，来平衡教师和学生在不同训练阶段的权重比例
3、为了提高蒸馏性能、降低训练成本，作者提出一种“在平衡中取舍”的训练调度方法

[2]Wang, Chaofei, et al. "TC3KD: Knowledge distillation via teacher-student cooperative curriculum customization." Neurocomputing 508 (2022): 284-292.

Difficulty Measurer



$$g((x_i, y_i)) = \alpha(-y_i^T \cdot logf(x_i; \theta^t)) + (1-\alpha)(-y_i^T \cdot logf(x_i; \theta^s)),$$

$$\alpha = 1 - \frac{k-1}{m}, k = \{1, \cdots, m\},$$

[2]Wang, Chaofei, et al. "TC3KD: Knowledge distillation via teacher-student cooperative curriculum customization." Neurocomputing 508 (2022): 284-292.

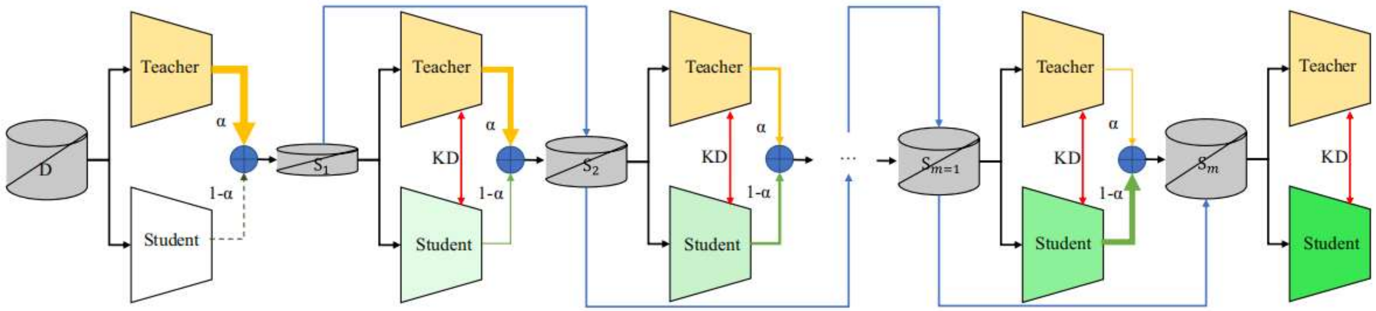
Training Scheduler

$$h : D' \rightarrow \{S_1, S_2, S_3 \cdots, S_m\}, \quad S_1 \subseteq S_2 \subseteq S_3 \subseteq \cdots \subseteq S_m = D',$$

instead of sorting of all samples. Specifically, 1) in the first stage, we rank the whole original training set D from easy to hard with the difficulty measurer, and fetch the $\frac{N}{cm}$ (c denotes the number of classes) simplest samples from each class to form S_1. Then, we do distillation on S_1 to get a snapshot student, and update the difficulty measurer. 2) In the second stage, we remove S_1 from the training set, and rank the residual training set $D - S_1$ from easy to hard with the updated difficulty measure. Then, we fetch the $\frac{N}{cm}$ simplest samples from each class, which are merged with S_1 to form S_2. We do distillation on S_2 to get an improved snapshot student, and update the difficulty measurer. 3) We repeat this process until S_m is equal to D. Finally, we conduct distillation on S_m to get the optimal student. The integrated algorithm of TC^{3}KD is shown in Algorithm 2.

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

TC3KD



Algorithm 2: Algorithm of knowledge distillation via teacher-student cooperative curriculum customization

Input: $f(\theta^t)$: teacher network, $f(\theta^s)$: student network, $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$: training set, m: the number of training stages, $S_0 = \phi$: initial training subset;
Output: $f^*(\theta^s)$: optimal student network;
for all $k = 1, 2, \ldots, m$ **do**
  $D_k = D - S_{k-1}$;
  calculate $\alpha = 1 - \frac{k-1}{m}$;
  calculate the difficulty of samples from D_k by Equ. 3, and get D'_k = ascending sort (D_k);
  select top $\frac{N}{cm}$ samples of each class in D'_k to get S_{top};
  $S_k = S_{k-1} \cup S_{top}$;
  while $f(\theta^s)$ does not converge **do**
    Train $f(\theta^s)$ on S_k with Equ. 2;
  end while
end for

$$g((x_i, y_i)) = \alpha(-y_i^T \cdot log f(x_i; \theta^t)) + (1 - \alpha)(-y_i^T \cdot log f(x_i; \theta^s)),$$

$$L^s(B, \theta^s) = -\frac{1}{|B|} \sum_{(x_i, y_i) \in B} \{\lambda y_i^T \cdot log f(x_i; \theta^s) + (1 - \lambda) \mathrm{KL}[f^\tau(x_i; \theta^t) || f^\tau(x_i; \theta^s)]\},$$

Experiments

1、Comparison with the mainstream methods

Table 1
Comparison results between the mainstream methods and TC^3KD on three datasets. Different network structures and teacher-student pairs are adopted. Top 1 accuracy (%) is averagely evaluated in three independent experiments. For the baseline methods, we reproduce the results following their published code. The best results are **bold**.

Dataset	Network structure	Teacher	Student	KD [18]	AT [52]	FitNets [39]	CCKD [37]	SLKD [56]	Ours
CIFAR-100	WRN-40-2/WRN-40-1	76.53	71.95	72.68	72.94	72.94	72.22	72.89	**73.55**
	ResNet-110/ResNet-20	73.41	68.91	70.67	70.91	70.67	70.88	71.10	**71.95**
	WRN-40-2/MobileNetV2	76.53	64.49	68.03	68.37	68.19	68.22	68.53	**69.11**
	ResNet-110/MobileNetV2	73.41	64.49	68.63	68.84	68.62	68.61	69.27	**70.08**
CINIC-10	ResNet-110/ResNet-20	86.45	82.43	82.58	82.84	82.90	82.98	83.16	**83.69**
ImageNet	ResNet-34/ResNet-18	73.31	69.75	70.66	70.70	69.89	69.96	70.54	**71.13**
	ResNet-50/MobileNetV2	75.54	64.23	66.72	66.85	66.21	66.71	66.31	**67.24**

2、Combination with the mainstream methods

Table 2
Results of combination Teacher-student Cooperative Curriculum Customization (TC^3) with the mainstream methods on CIFAR-100. Different network structures and teacher-student pairs are adopted. Top 1 accuracy (%) is averagely evaluated in three independent experiments. The superscript numbers represent the variations of results after adding TC^3, $\uparrow$ for increase, $\downarrow$ for decrease. The improved results with TC^3 are **bold**.

Network structure	Teacher	Student	AT [52]	AT + TC^3	FitNets [39]	FitNets + TC^3	CCKD [37]	CCKD + TC^3
WRN-40-2/WRN-40-1	76.53	71.95	72.94	**73.69**$^{\uparrow 0.75}$	72.94	**73.29**$^{\uparrow 0.35}$	72.22	71.76$^{\downarrow 0.46}$
ResNet-110/ResNet-20	73.41	68.91	70.91	**71.69**$^{\uparrow 0.78}$	70.67	**71.50**$^{\uparrow 0.83}$	70.88	70.65$^{\downarrow 0.23}$
WRN-40-2/MobileNetV2	76.53	64.49	68.37	**70.03**$^{\uparrow 1.66}$	68.19	**69.52**$^{\uparrow 1.33}$	68.22	**68.25**$^{\uparrow 0.03}$
ResNet-110/MobileNetV2	73.41	64.49	68.84	**69.28**$^{\uparrow 0.44}$	68.62	**69.58**$^{\uparrow 0.96}$	68.61	68.06$^{\downarrow 0.55}$
Average	74.97	67.46	70.26	**71.17**$^{\uparrow 0.91}$	70.10	**70.97**$^{\uparrow 0.87}$	69.98	69.68$^{\downarrow 0.30}$

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Ablation study

1、Different difficulty measurers.

Table 3
Comparison results of different difficulty measurers on CIFAR-100. "DM" denotes difficulty measurer. "Fixed teacher" represents the pre-trained teacher. "Fixed student" represents the trained student by standard KD [18]. Top 1 accuracy (%) is averagely evaluated in three independent experiments. The best results are **bold**.

Structure	Teacher Student	WRN-40-2 WRN-40-1	ResNet-110 ResNet-20
Accuracy	Teacher	76.53	73.41
	Student	71.95	68.91
	KD [18]	72.68	70.67
DM	Fixed teacher	72.33	69.76
	Fixed student	72.77	70.91
	SLKD [56]	72.89	71.10
	TC[3]	**73.55**	**71.95**

2、Different weight settings.

equ.5 $$\alpha = \frac{k}{m}, k = \{1, \cdots, m\},$$

$$\alpha = 0.5$$

equ.4 $$\alpha = 1 - \frac{k-1}{m}, k = \{1, \cdots, m\},$$

Comparison results of different weight settings on CIFAR-100. "WS" denotes weight setting. "Decreasing α" follows Equ. 4. "Fixed $\alpha = 0.5$" represents equal distribution. "Increasing α" follows Equ. 5. Top 1 accuracy (%) is averagely evaluated in three independent experiments. The best results are **bold**.

Structure	Teacher Student	WRN-40-2 WRN-40-1	ResNet-110 ResNet-20
Accuracy	Teacher	76.53	73.41
	Student	71.95	68.91
	KD [18]	72.68	70.67
WS	Increasing α	69.79	67.52
	Fixed $\alpha = 0.5$	72.82	69.94
	Decreasing α	**73.55**	**71.95**

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Ablation study

3、Different training schedulers.

Table 5
Comparison results between different training schedulers on CIFAR-100. "TS-1" represents "fetch and remove in balance". "TS-2" represents "fetch and remove without balance". "TS-3" represents "fetch but do not remove in balance". "TS-4" represents "fetch but do not remove without balance". Top 1 accuracy (%) is averagely evaluated in three independent experiments. Computational cost (minutes) is estimated on TITAN Xp. The best results are **bold**.

Network structure	Accuracy of baseline			KD accuracy				Computational cost			
	T	S	KD [18]	TS-1	TS-2	TS-3	TS-4	TS-1	TS-2	TS-3	TS-4
WRN-40-2/WRN-40-1	76.53	71.95	72.68	**73.55**	72.52	73.43	72.61	**7.2**	**7.2**	13.8	13.8
ResNet-110/ResNet-20	73.41	68.91	70.67	**71.95**	70.21	71.58	70.36	**8.5**	**8.5**	16.3	16.3

4、Different stage partitions.

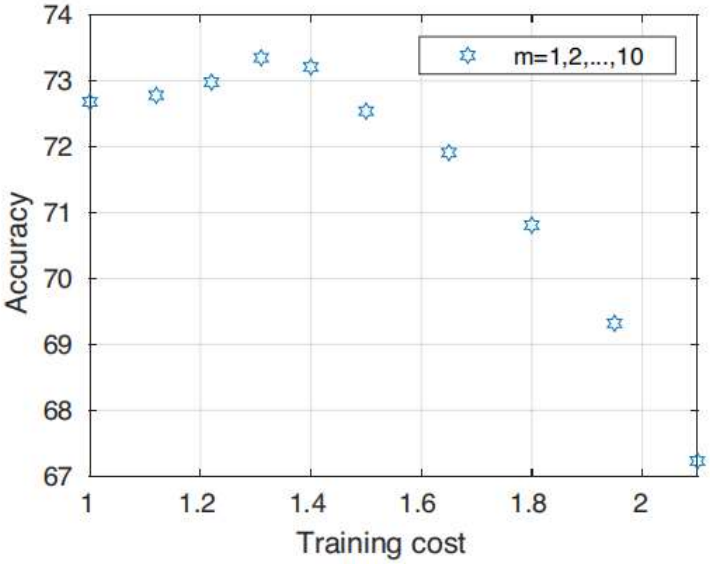


Fig. 5. Ablation study of the number of stages m. WRN-40-2/ WRN-40-1 is adopted as teacher-student pair. Total epochs are fixed 200. We try different m from 1 to 10.

Algorithm 2: Algorithm of knowledge distillation via teacher-student cooperative curriculum customization

Input: $f(\theta^t)$: teacher network, $f(\theta^s)$: student network, $D=\{(x_1,y_1),\ldots,(x_N,y_N)\}$: training set, m: the number of training stages, $S_0=\phi$: initial training subset;
Output: $f^*(\theta^s)$: optimal student network;
for all $k=1,2,\ldots,m$ **do**
  $D_k = D - S_{k-1}$;
  calculate $\alpha = 1 - \frac{k-1}{m}$;
  calculate the difficulty of samples from D_k by Equ. 3, and get D'_k = ascending sort (D_k);
  select top $\frac{N}{cm}$ samples of each class in D'_k to get S_{top};
  $S_k = S_{k-1} \cup S_{top}$;
  while $f(\theta^s)$ does not converge **do**
    Train $f(\theta^s)$ on S_k with Equ. 2;
  end while
end for

Conclusion and limitation

1、不适用在语义分割任务上

2、方法的局限性：对基于关系或基于图的知识蒸馏方法不友好

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

$$L^{s}(B,\theta^{s}) = -\frac{1}{|B|}\sum_{(x_i,y_i)\in B}\{\lambda y_i^{T}\cdot logf(x_i;\theta^{s})$$
$$+(1-\lambda)\mathrm{KL}\left[f^{\tau}(x_i;\theta^{t})||f^{\tau}(x_i;\theta^{s})\right]\},$$

while the student network has not converged **do**
  $L_{task} = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W} CE\left(\sigma\left(\mathbf{Z}_{h,w}\right), \mathbf{y}_{h,w}\right)$
  $L_{kd} = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W} KL\left(\sigma\left(\frac{\mathbf{Z}^{s}_{h,w}}{T}\right)\|\sigma\left(\frac{\mathbf{Z}^{t}_{h,w}}{T}\right)\right)$
  if $iter \leq iter_{warm-up}$ **then**
    $D_{kl} = f\left(x \mid \theta_t\right)\log\left(\frac{f(x|\theta_t)}{f_{aux}(x|\theta_t)}\right)$
    $\mathrm{TERD_e} = \exp\left\{-D_{kl}\right\}$
    $L_{overall} = \mathrm{TERD_e}\cdot L_{task} + L_{kd}$
  else
    $\boxed{\mathrm{TSRD_e} = \left(f(x \mid \theta_s) \leq t\right) \oplus \left(f(x \mid \theta_t) \leq t\right)}$
    $L_{overall} = \mathrm{TSRD_e}\cdot L_{task} + L_{kd}$
  end if
  $L_{overall}.backward()$
end while
return θ_s

TSRD=a*TSRD

a	**0.8**	**0.9**	**1.0**	**1.1**	**1.2**
mIoU	77.18	76.45	76.69	77.16	

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

结　　束