Distilling Knowledge via Knowledge Review

Pengguang Chen[1] Shu Liu[2] Hengshuang Zhao[3] Jiaya Jia[1,2]

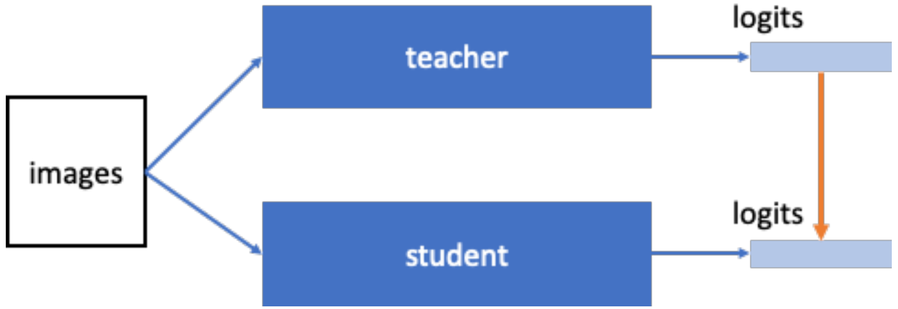
The Chinese University of Hong Kong[1] SmartMore[2] University of Oxford[3]

Review

Knowledge distillation

1.Traditional teacher-student distillation[1]

$$D_{\mathrm{KL}}(\boldsymbol{p}||\boldsymbol{q}) = \mathcal{H}(\boldsymbol{p},\boldsymbol{q}) - \mathcal{H}(\boldsymbol{p})$$
$$= -\sum_i p_i \log q_i - (-\sum_i p_i \log p_i).$$

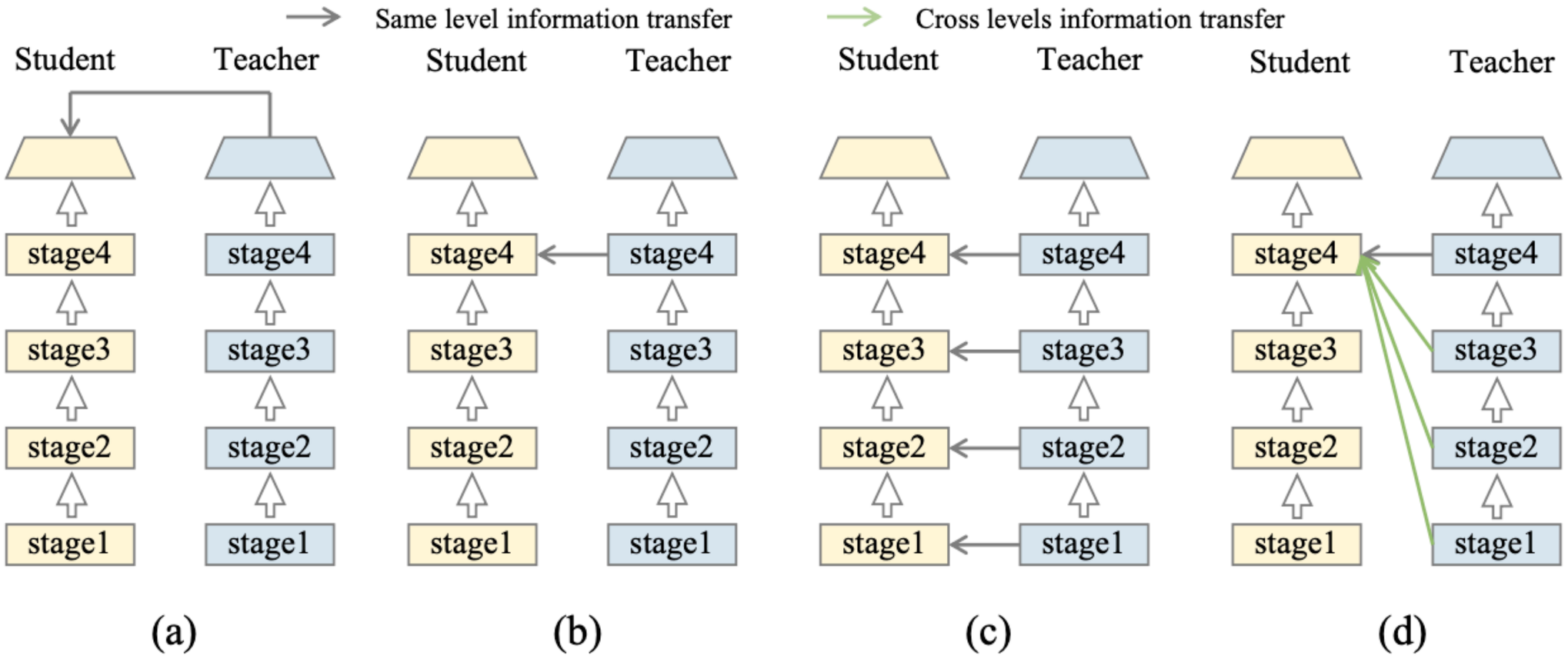


2. Fitnets (Hints)[2]

$$\mathcal{L}_{hint} = \frac{1}{2}\|\boldsymbol{Z}^t - r(\boldsymbol{Z}^s)\|^2$$

$$\mathbf{W}^*_{\mathbf{Guided}} = \underset{\mathbf{W}_{\mathbf{Guided}}}{\operatorname{argmin}}\ \mathcal{L}_{HT}(\mathbf{W}_{\mathbf{Guided}}, \mathbf{W}_r)$$

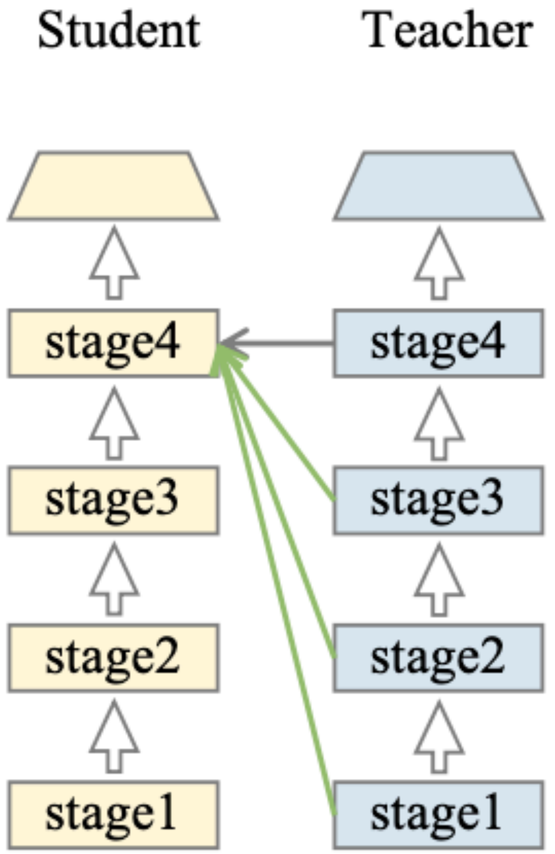
Motivation


Same level information transfer
Cross levels information transfer
Student
Teacher
Student
Teacher
Student
Teacher
Student
Teacher
stage4
stage3
stage2
stage1
(a)
(b)
(c)
(d)

Method

Knowledge Review

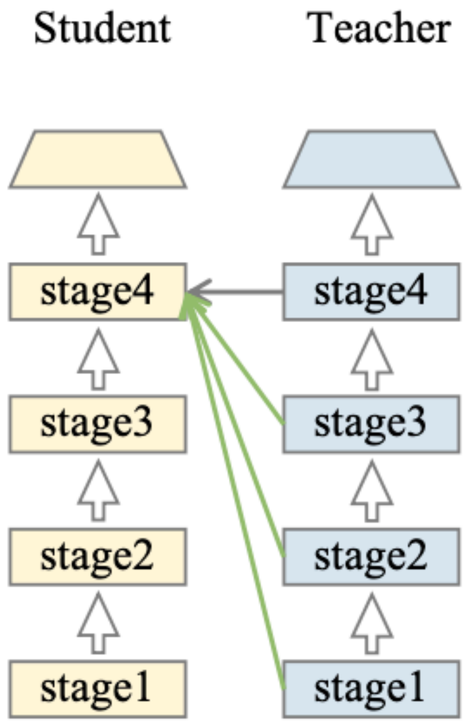
- Definition: use previous (shallower) features (of the teacher) to guide the current (deeper) feature (of the student).
- How to extract useful information from multi-level information from the teacher and how to transfer them to the student



(d)

Method

Review Mechanism

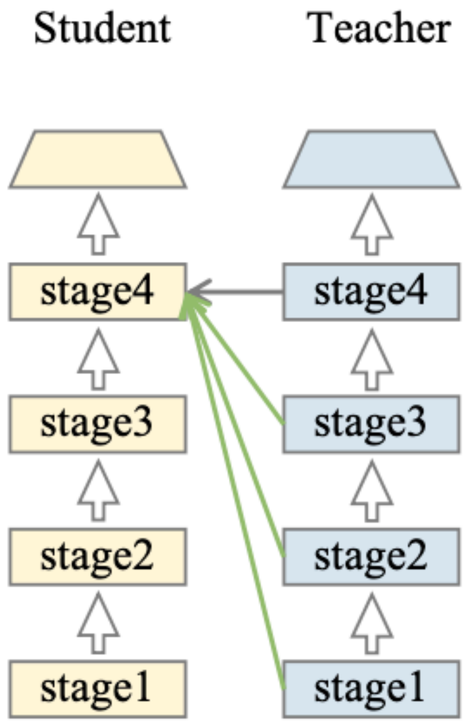


1.Symbols

- input image X
- teacher network Γ
- student network $\mathcal{S}$ divided into $(\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_n, \mathcal{S}_c)$
- $Y_s = \mathcal{S}(X)$ is the logit of the student
- $Y_s = \mathcal{S}_c \circ \mathcal{S}_n \circ \cdots \circ \mathcal{S}_1(X)$
- Intermediate features $(\mathbf{F}_s^1, \cdots, \mathbf{F}_s^n)$, $\mathbf{F}_s^i = \mathcal{S}_i \circ \cdots \circ \mathcal{S}_1(X)$

Method

Review Mechanism



2. single-layer knowledge distillation

$$\mathcal{L}_{SKD} = \mathcal{D}\left(\mathcal{M}_s^i(\mathbf{F}_s^i), \mathcal{M}_t^i(\mathbf{F}_t^i)\right)$$

3. multiple-layers knowledge distillation

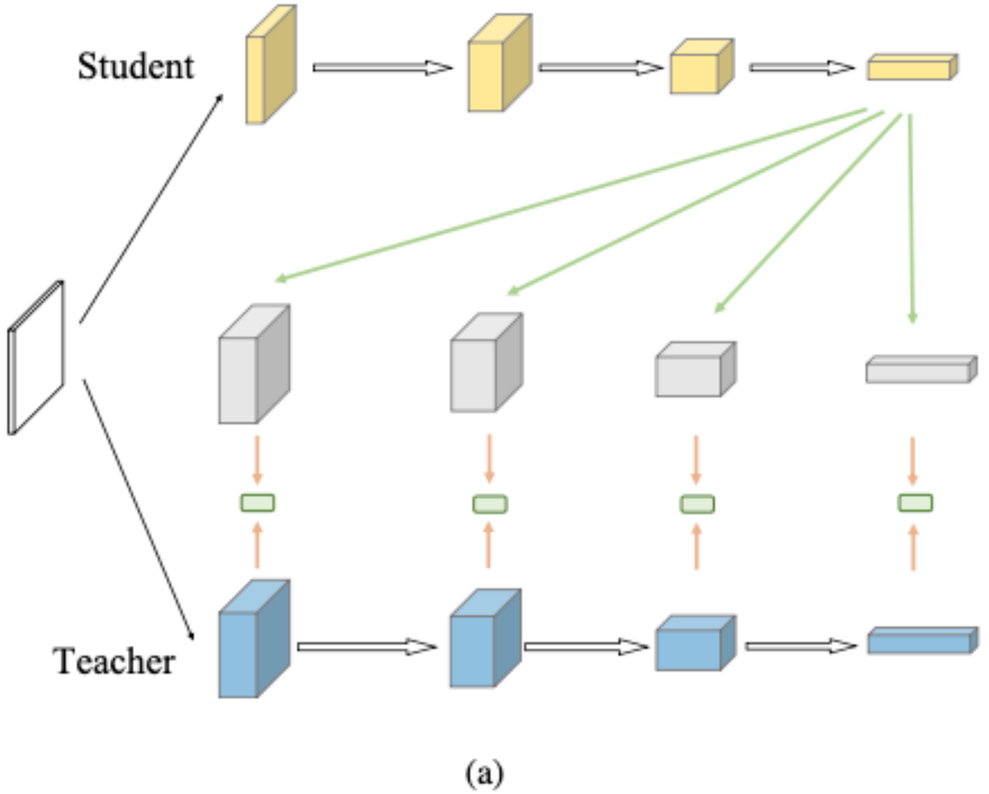
$$\mathcal{L}_{MKD} = \sum_{i \in \mathbf{I}} \mathcal{D}\left(\mathcal{M}_s^i(\mathbf{F}_s^i), \mathcal{M}_t^i(\mathbf{F}_t^i)\right)$$

4. single-layer knowledge distillation with review mechanism

$$\mathcal{L}_{SKD\_R} = \sum_{j=1}^{i} \mathcal{D}\left(\mathcal{M}_s^{i,j}(\mathbf{F}_s^i), \mathcal{M}_t^{j,i}(\mathbf{F}_t^j)\right)$$

Method

Review Mechanism



2. single-layer knowledge distillation

$$\mathcal{L}_{SKD} = \mathcal{D}\left(\mathcal{M}_s^i(\mathbf{F}_s^i), \mathcal{M}_t^i(\mathbf{F}_t^i)\right)$$

3. multiple-layers knowledge distillation

$$\mathcal{L}_{MKD} = \sum_{i \in \mathbf{I}} \mathcal{D}\left(\mathcal{M}_s^i(\mathbf{F}_s^i), \mathcal{M}_t^i(\mathbf{F}_t^i)\right)$$

4. single-layer knowledge distillation with review mechanism

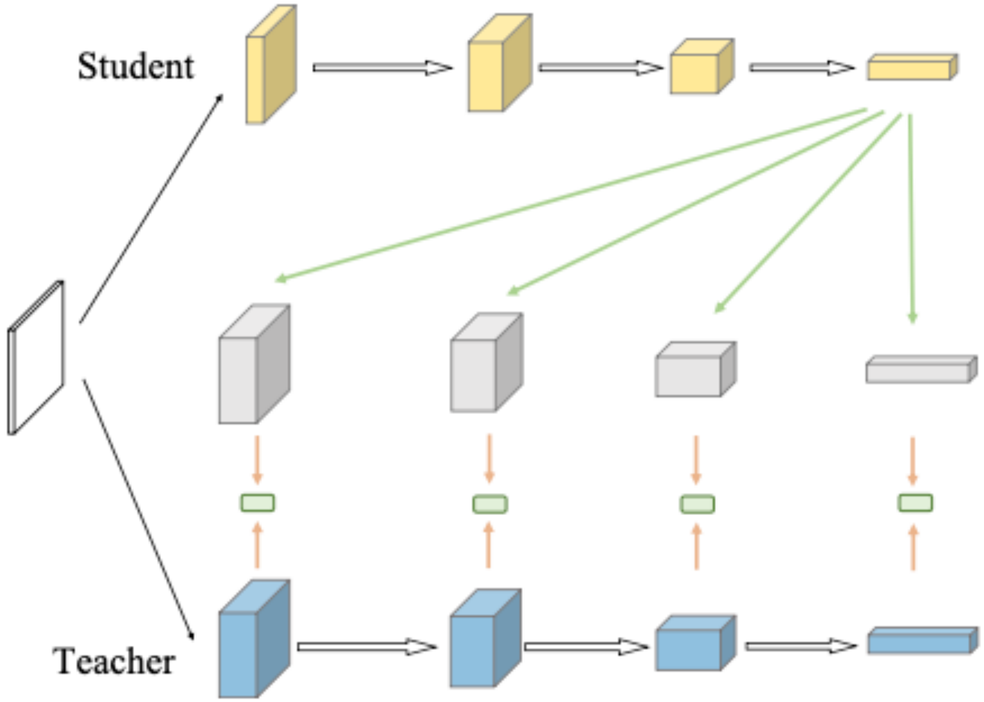
$$\mathcal{L}_{SKD\_R} = \sum_{j=1}^{i} \mathcal{D}\left(\mathcal{M}_s^{i,j}(\mathbf{F}_s^i), \mathcal{M}_t^{j,i}(\mathbf{F}_t^j)\right)$$

Method

Review Mechanism

5. multiple-layers knowledge distillation with review mechanism

$$\mathcal{L}_{MKD\_R} = \sum_{i \in \mathbf{I}} \left(\sum_{j=1}^{i} \mathcal{D}\left(\mathcal{M}_s^{i,j}(\mathbf{F}_s^i), \mathcal{M}_t^{j,i}(\mathbf{F}_t^j) \right) \right)$$



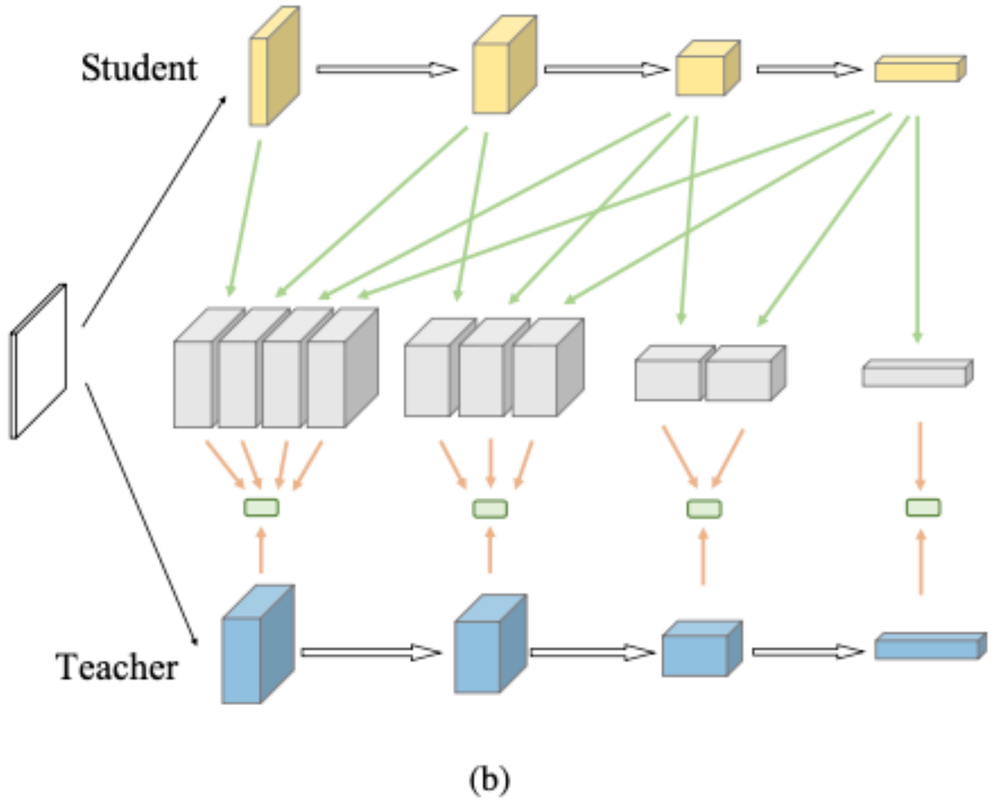
(a)

Method

Review Mechanism

5. multiple-layers knowledge distillation with review mechanism

$$\mathcal{L}_{MKD\_R} = \sum_{i \in \mathbf{I}} \left(\sum_{j=1}^{i} \mathcal{D}\left(\mathcal{M}_s^{i,j}(\mathbf{F}_s^i), \mathcal{M}_t^{j,i}(\mathbf{F}_t^j) \right) \right)$$



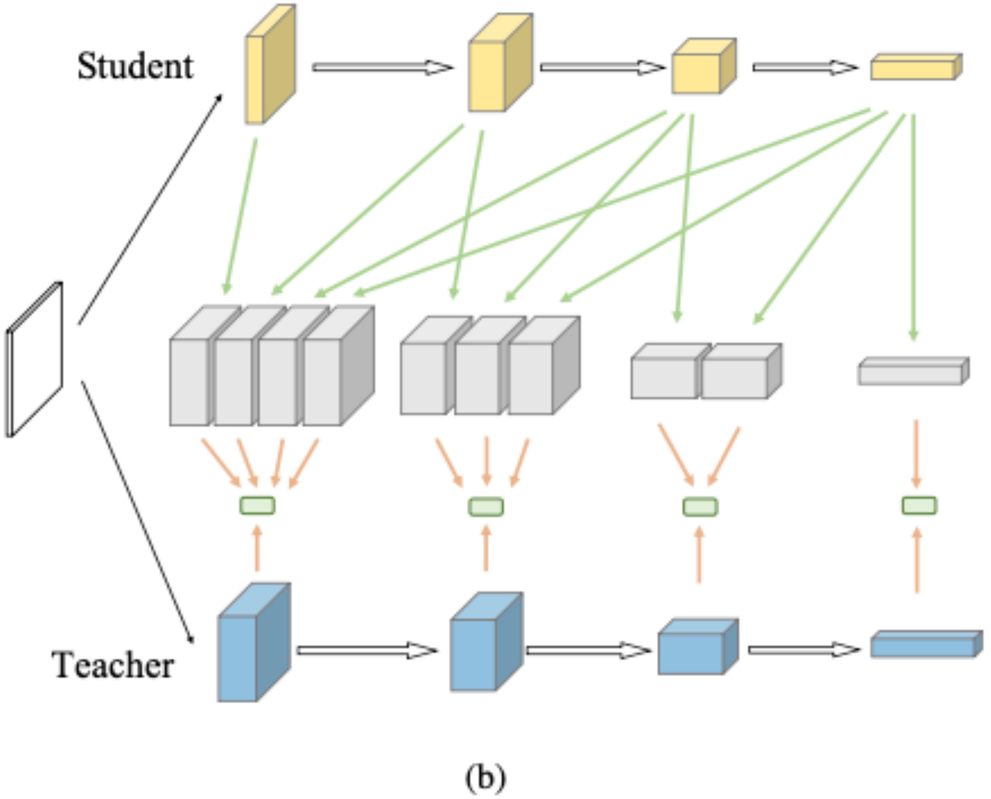
(b)

Method

Review Mechanism

5. multiple-layers knowledge distillation with review mechanism

$$\mathcal{L}_{MKD\_R} = \sum_{i \in \mathbf{I}} \left(\sum_{j=1}^{i} \mathcal{D}\left(\mathcal{M}_s^{i,j}(\mathbf{F}_s^i), \mathcal{M}_t^{j,i}(\mathbf{F}_t^j) \right) \right)$$



(b)

$$\mathcal{L}_{MKD\_R} = \sum_{i=1}^{n} \left(\sum_{j=1}^{i} \mathcal{D}\left(\mathbf{F}_s^i, \mathbf{F}_t^j \right) \right)$$

$$\mathcal{L}_{MKD\_R} = \sum_{j=1}^{n} \left(\sum_{i=j}^{n} \mathcal{D}\left(\mathbf{F}_s^i, \mathbf{F}_t^j \right) \right)$$

Method

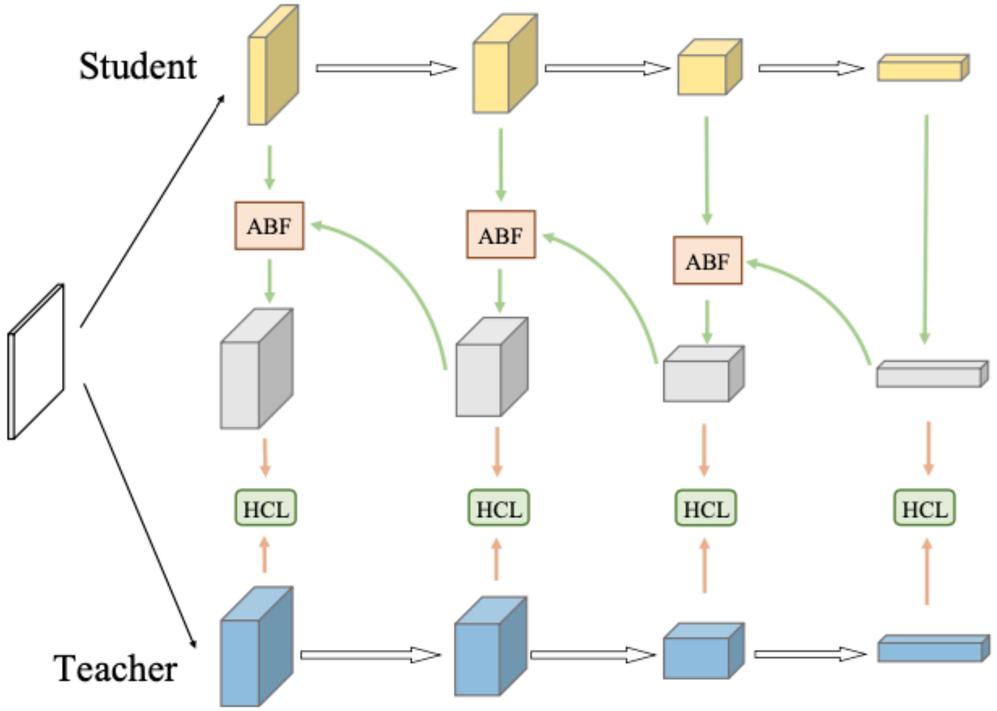
Residual Learning Framework

Fusion of features

$$\mathcal{L}_{MKD\_R} = \sum_{j=1}^{n} \left(\sum_{i=j}^{n} \mathcal{D}\left(\mathbf{F}_s^i, \mathbf{F}_t^j\right) \right)$$

$$\sum_{i=j}^{n} \mathcal{D}\left(\mathbf{F}_s^i, \mathbf{F}_t^j\right) \approx \mathcal{D}\left(\mathcal{U}(\mathbf{F}_s^j, \cdots, \mathbf{F}_s^n), \mathbf{F}_t^j\right)$$

$$\mathcal{L}_{MKD\_R} = \mathcal{D}(\mathbf{F}_s^n, \mathbf{F}_t^n) + \sum_{j=n-1}^{1} \mathcal{D}\left(\mathcal{U}(\mathbf{F}_s^j, \mathbf{F}_s^{j+1,n}), \mathbf{F}_t^j\right)$$



(d)

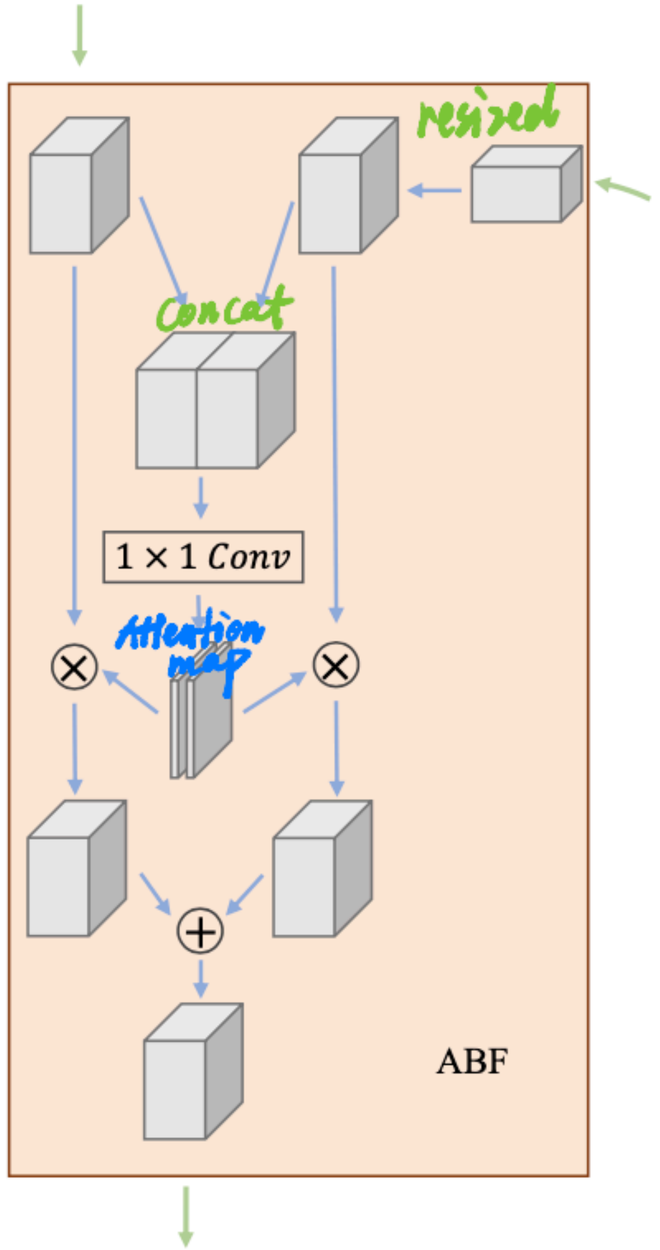
Method

Attention based fusion (ABF)

1. The higher level features are first resized to the same shape as the lower level features.

2. Then two features from different levels are concatenated together to generate two H × W attention maps.

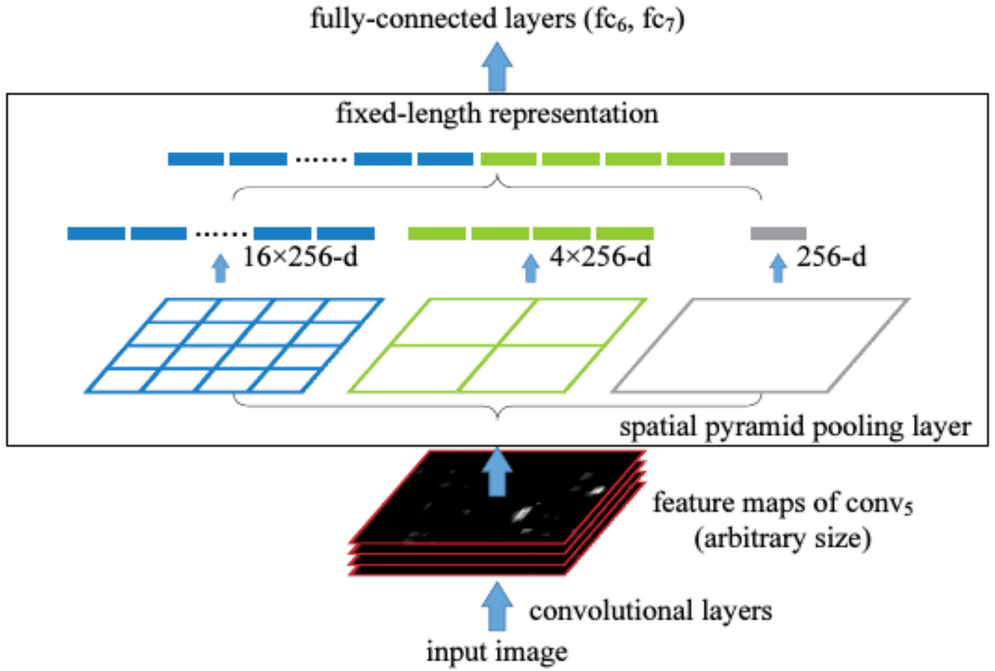
3. These maps are multiplied with two features, respectively. Finally, the two features are added.

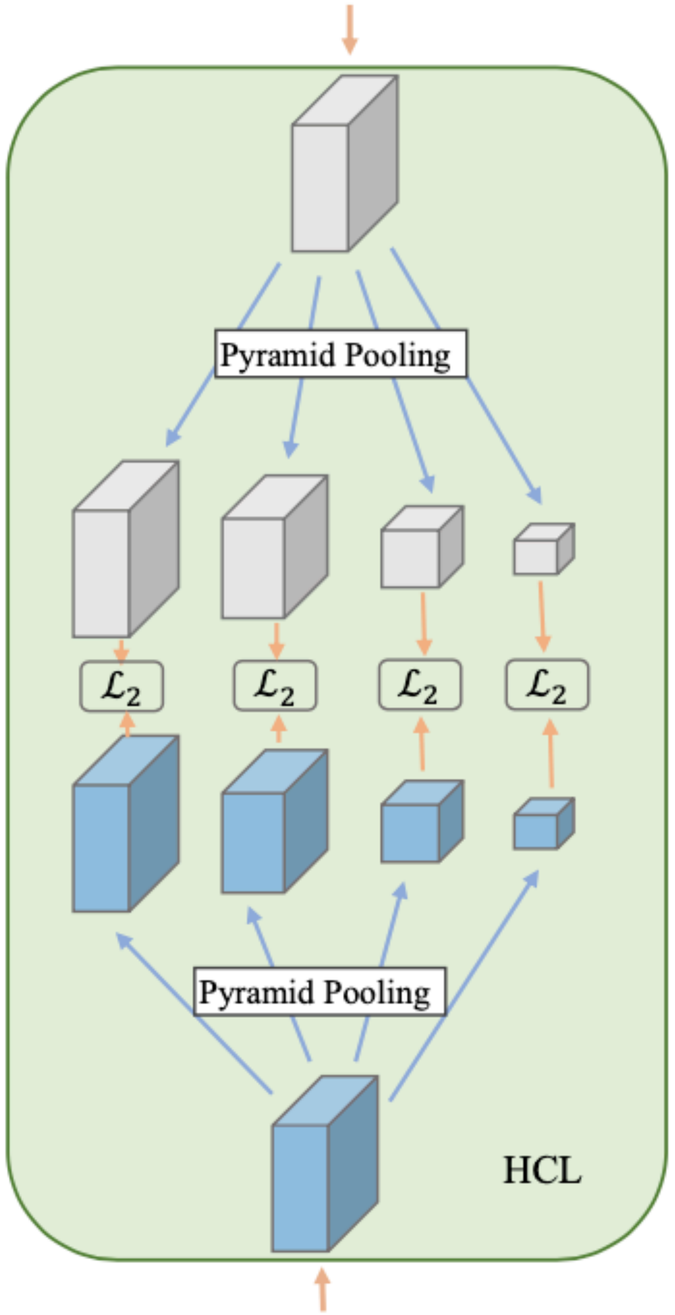


Method

Hierarchical context loss (HCL) function

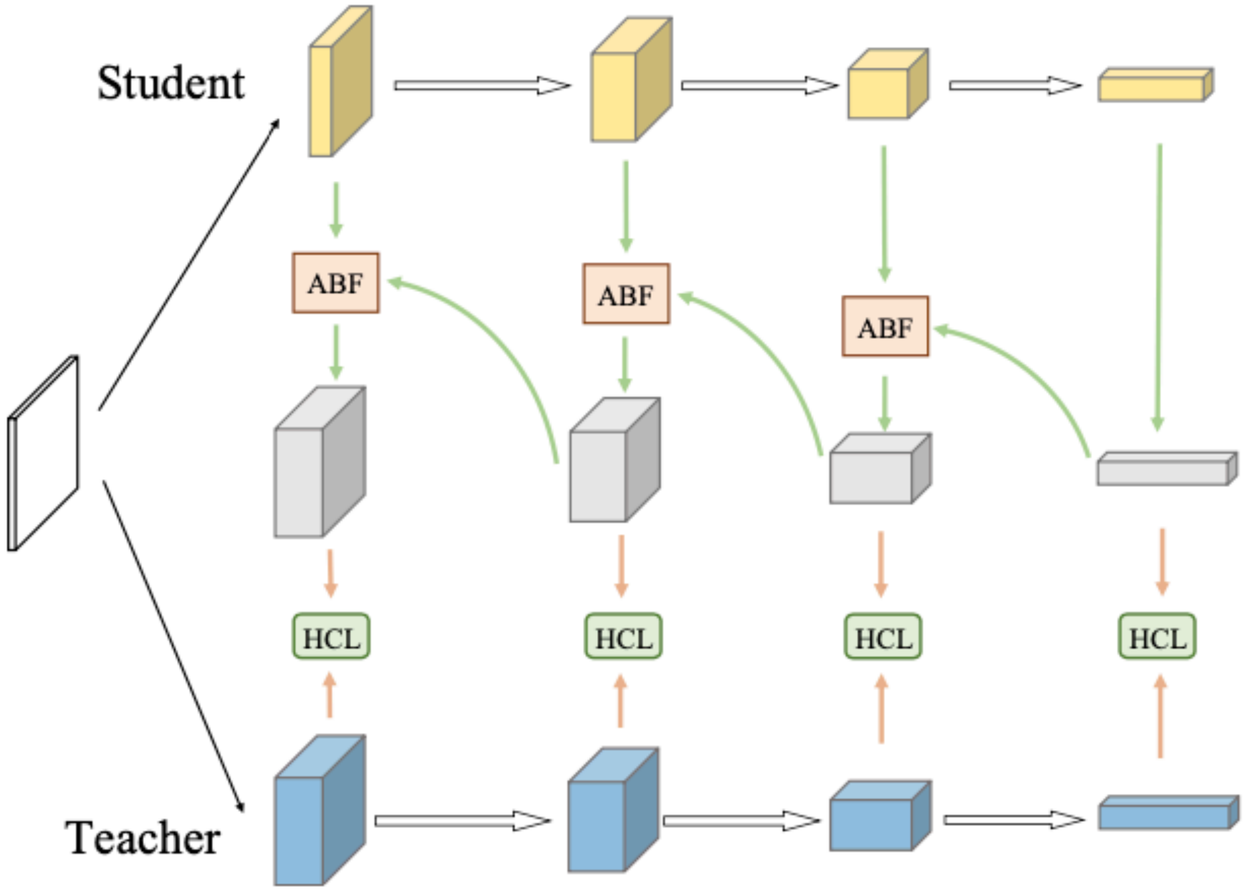
- L2 distance is only effective to transfer information between features from the same level.
- Spatial pyramid pooling





Method

ReviewKD



(d)

Experiments

Classification

Distillation Mechanism	Teacher Acc	ResNet56 72.34	ResNet110 74.31	ResNet32x4 79.42	WRN40-2 75.61	WRN40-2 75.61	VGG13 74.64
	Student Acc	ResNet20 69.06	ResNet32 71.14	ResNet8x4 72.50	WRN16-2 73.26	WRN40-1 71.98	VGG8 70.36
Logits	KD [9]	70.66	73.08	73.33	74.92	73.54	72.98
Single Layer	FitNet [25]	69.21	71.06	73.50	73.58	72.24	71.02
Single Layer	PKT [23]	70.34	72.61	73.64	74.54	73.54	72.88
Single Layer	RKD [22]	69.61	71.82	71.90	73.35	72.22	71.48
Single Layer	CRD [28]	71.16	73.48	75.51	75.48	74.14	73.94
Multiple Layers	AT [38]	70.55	72.31	73.44	74.08	72.77	71.43
Multiple Layers	VID [1]	70.38	72.61	73.09	74.11	73.30	71.23
Multiple Layers	OFD [8]	70.98	73.23	74.95	75.24	74.33	73.95
Review	Ours	**71.89**	**73.89**	**75.63**	**76.12**	**75.09**	**74.84**

Table 1. Results on CIFAR-100. The teacher and student have architectures of the same style.

Experiments

Classification

Setting		Teacher	Student	KD [9]	AT [38]	OFD [8]	CRD [28]	Ours
(a)	Top-1	76.16	68.87	68.58	69.56	71.25	71.37	**72.56**
	Top-5	92.86	88.76	88.98	89.33	90.34	90.41	**91.00**
(b)	Top-1	73.31	69.75	70.66	70.69	70.81	71.17	**71.61**
	Top-5	91.42	89.07	89.88	90.01	89.98	90.13	**90.51**

Table 3. Results on ImageNet. (a) MobileNet as student, ResNet50 as teacher. (b) ResNet18 as student, ResNet34 as teacher.

Experiments

Object Detection

student: Mask R-CNN[3] teacher: from Detectron2[4]

dataset: COCO2017

	Method	mAP	AP50	AP75	APl	APm	APs
Teacher	Faster R-CNN w/ R101-FPN	42.04	62.48	45.88	54.60	45.55	25.22
Student	Faster R-CNN w/ R18-FPN	33.26	53.61	35.26	43.16	35.68	18.96
	w/ KD [9]	33.97 (+0.61)	54.66	36.62	44.14	36.67	18.71
	w/ FitNet [25]	34.13 (+0.87)	54.16	36.71	44.69	36.50	18.88
	w/ FGFI [31]	35.44 (+2.18)	55.51	38.17	47.34	38.29	19.04
	w/ Our Method	**36.75 (+3.49)**	**56.72**	**34.00**	**49.58**	**39.51**	**19.42**
Teacher	Faster R-CNN w/ R101-FPN	42.04	62.48	45.88	54.60	45.55	25.22
Student	Faster R-CNN w/ R50-FPN	37.93	58.84	41.05	49.10	41.14	22.44
	w/ KD [9]	38.35 (+0.42)	59.41	41.71	49.48	41.80	22.73
	w/ FitNet [25]	38.76 (+0.83)	59.62	41.80	50.70	42.20	22.32
	w/ FGFI [31]	39.44 (+1.51)	60.27	43.04	51.97	42.51	22.89
	w/ Our Method	**40.36 (+2.43)**	**60.97**	**44.08**	**52.87**	**43.81**	**23.60**

Experiments

Instance Segmentation

student: Mask R-CNN[3] teacher: from Detectron2[4]

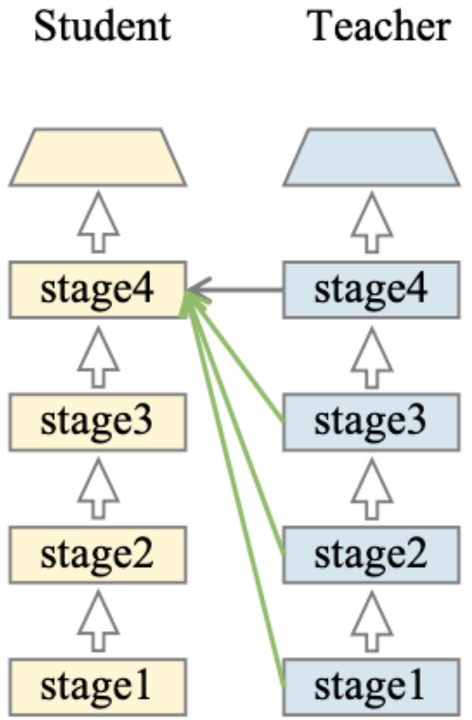
dataset: COCO2017

	Method	mAP	AP50	AP75	APl	APm	APs
Teacher	Mask R-CNN w/ R101-FPN	38.63	60.45	41.28	55.29	41.33	19.48
Student	Mask R-CNN w/ R18-FPN	31.25	51.07	33.10	45.53	32.80	14.18
	+ Our Method	**33.62 (+2.37)**	**53.91**	**35.96**	**50.30**	**35.31**	**15.03**
Teacher	Mask R-CNN w/ R101-FPN	38.63	60.45	41.28	55.29	41.33	19.48
Student	Mask R-CNN w/ R50-FPN	35.24	56.32	37.49	50.34	37.71	17.16
	+ Our Method	**36.98 (+1.74)**	**58.13**	**39.60**	**53.19**	**39.57**	**17.54**
Teacher	Mask R-CNN w/ R50-FPN	37.17	58.60	39.88	53.30	39.49	18.63
Student	Mask R-CNN w/ MV2-FPN	28.37	47.19	29.95	41.70	29.01	12.09
	+ Our Method	**31.56 (+3.19)**	**50.70**	**33.44**	**47.39**	**32.44**	**12.76**

Experiments

Ablation Study

- ResNet20 as the student and ResNet56 as the teacher on CIFAR100
- The student's baseline result is 69.1
- Red - lower than baseline
- Blue - higher than baseline



		Teacher Stage 1	2	3	4
Student Stage	1	69.5	69.0	68.2	66.3
	2	69.6	69.6	61.4	61.1
	3	69.2	69.8	71.0	50.4
	4	69.2	69.3	70.3	70.3

Experiments

Ablation Study

student: WRN16-2

teacher: WRN40-2

dataset: CIFAR-100

teacher - 75.61

RM	RLF	ABF	HCL	Accuracy (Variance)
				74.3 (5e-2)
✔				75.2 (6e-2)
✔	✔			75.6 (6e-2)
✔	✔	✔		76.0 (6e-2)
✔	✔		✔	75.8 (5e-2)
✔	✔	✔	✔	76.2 (4e-2)

Table 7. RM: The proposed review mechanism (Section 3.1). RLF: Residual learning frame work (Section 3.2). ABF: Attentation based fusion module (Section 3.3). HCL: Hierarchical context loss function (Section 3.3).

Reference

[1]Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).
[2]Romero, Adriana, et al. "Fitnets: Hints for thin deep nets." *arXiv preprint arXiv:1412.6550* (2014).
[3] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.
[4] Wu, Yuxin, et al. "Detectron2." (2019).