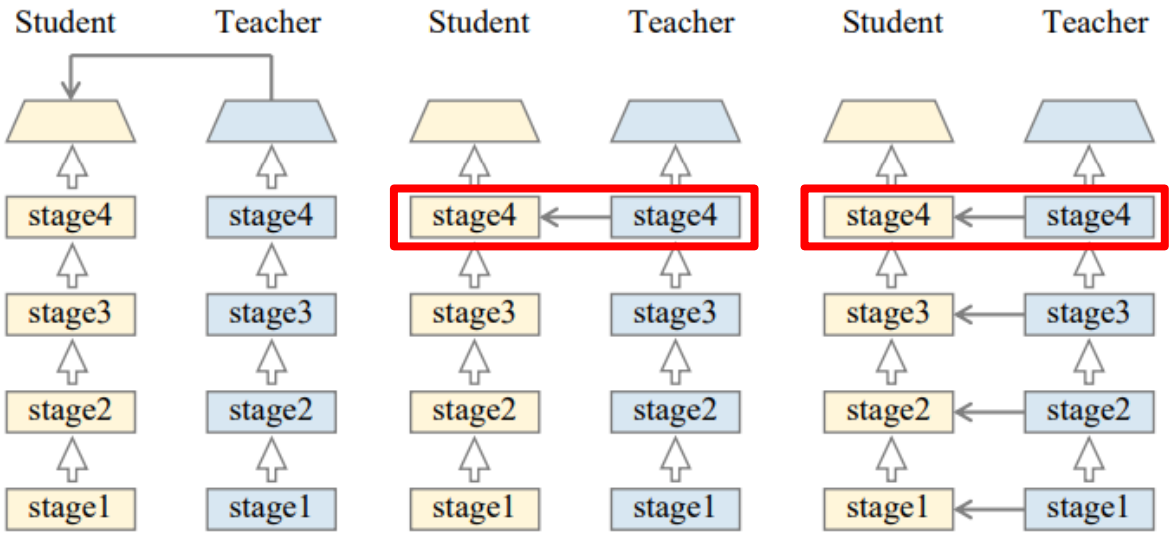
Distilling Knowledge via Knowledge Review

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]
{pgchen, leojia}@cse.cuhk.edu.hk liushuhust@gmail.com hengshuang.zhao@eng.ox.ac.uk

Presenter: Seonghak KIM

Introduction

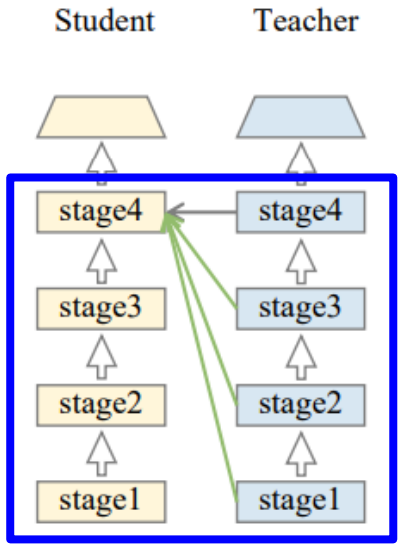
- **Knowledge Distillation**
  - **Training small networks under the supervision of a larger networks**
  - **Type**
    - Logit-based distillation
    - Feature-based distillation with intermediate layers
  - **(–) only use the same level information to guide the student**



[Previous knowledge distillation frameworks. They only transfer knowledge within the same levels]

Introduction

- **Knowledge Review**
  - **Review mechanism**
    - Multi-level information of the teacher to guide one-level learning of the student
  - **Residual learning**
  - **Attention based fusion (ABF)**
  - **Hierarchical context loss (HCL)**



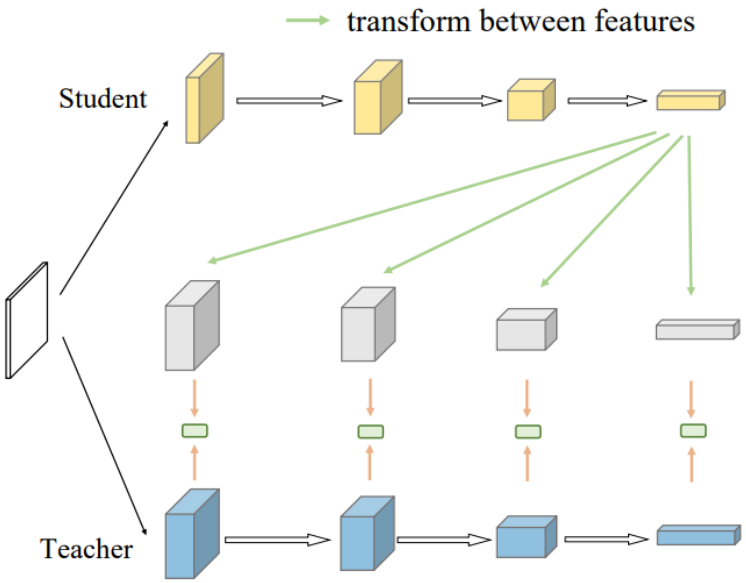
[Knowledge review mechanism. Multiple layers of the teacher is used to supervise one layer in the student]

Method

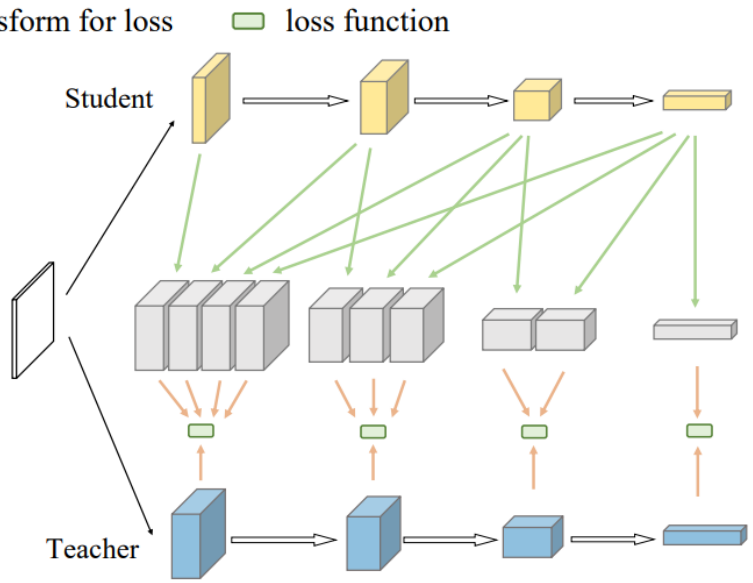
Review mechanism

Notation

- Given input image $\mathbf{X}$ and student network $\mathcal{S}$, the output logit of the student is $\mathbf{Y_s} = \mathcal{S}(\mathbf{X})$.
- When $\mathcal{S} \rightarrow (\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_n, \mathcal{S}_c)$, $\mathbf{Y_s} = \mathcal{S}_c \circ \mathcal{S}_n \circ \cdots \circ \mathcal{S}_1(\mathbf{X})$ [cf., $g \circ f(x) = g(f(x))$]
- Intermediate features are $\left(\boldsymbol{F}_s^1, , \dots, \boldsymbol{F}_s^i \dots, \boldsymbol{F}_s^n\right)$ where $\boldsymbol{F}_s^i = \mathcal{S}_i \circ \cdots \circ \mathcal{S}_1(\mathbf{X})$



[Single-layer knowledge distillation with the review mechanism]



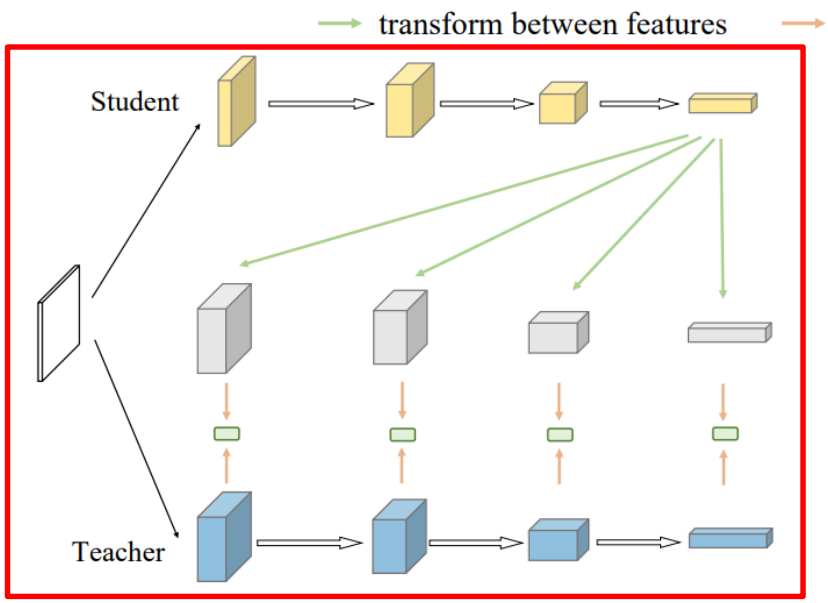
[Multiple-layers knowledge distillation with the review mechanism]

Method

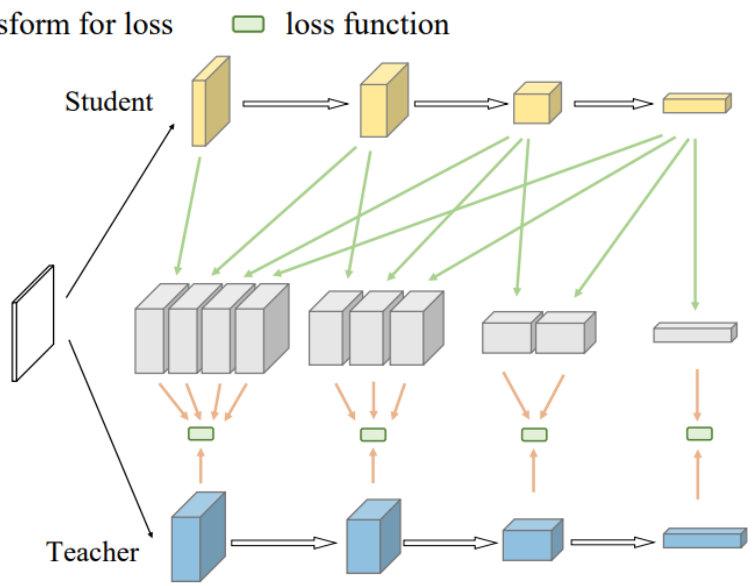
Review mechanism

Single-layer Knowledge distillation

- $\mathcal{L}_{\text{SKD}} = \mathcal{D}\left(\mathcal{M}_s^{\text{i}}(\boldsymbol{F}_s^{\text{i}}), \mathcal{M}_t^{\text{i}}(\boldsymbol{F}_t^{\text{i}})\right)$ ➔ $\mathcal{L}_{\text{SKD\_R}} = \sum_{\text{j}=1}^{\text{i}} \mathcal{D}\left(\mathcal{M}_s^{\text{i,j}}(\boldsymbol{F}_s^{\text{i}}), \mathcal{M}_t^{\text{j,i}}\left(\boldsymbol{F}_t^{\text{j}}\right)\right)$
  - $\mathcal{M}$ is transformation, which is simply composed of convolution and nearest interpolation layers for matching the size.
  - $\mathcal{D}$ is distance function.



[Single-layer knowledge distillation with the review mechanism]



[Multiple-layers knowledge distillation with the review mechanism]

Method

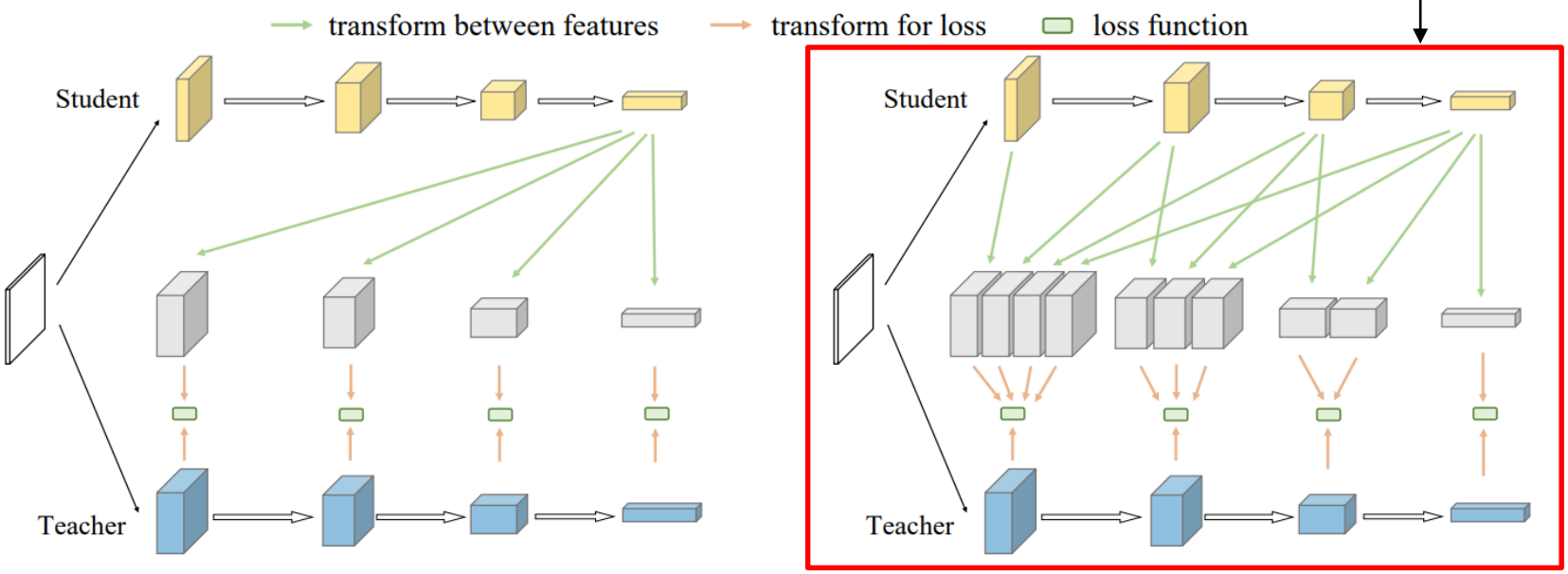
Review mechanism

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{MKD\_R}}$$

Multiple-layers Knowledge distillation

- $\mathcal{L}_{\text{MKD}} = \sum_{i \in I} \mathcal{D}\left(\mathcal{M}_s^i(F_s^i), \mathcal{M}_t^i(F_t^i)\right)$ ➔ $\mathcal{L}_{\text{MKD\_R}} = \sum_{i \in I}\left(\sum_{j=1}^{i} \mathcal{D}\left(\mathcal{M}_s^{i,j}(F_s^i), \mathcal{M}_t^{j,i}\left(F_t^j\right)\right)\right)$
- $(-)$ cumbersome learning process [e.g., network with n stages $\rightarrow$ calculations of $\frac{1}{2}n(n+1)$]

Shallower features of the teacher are used only.

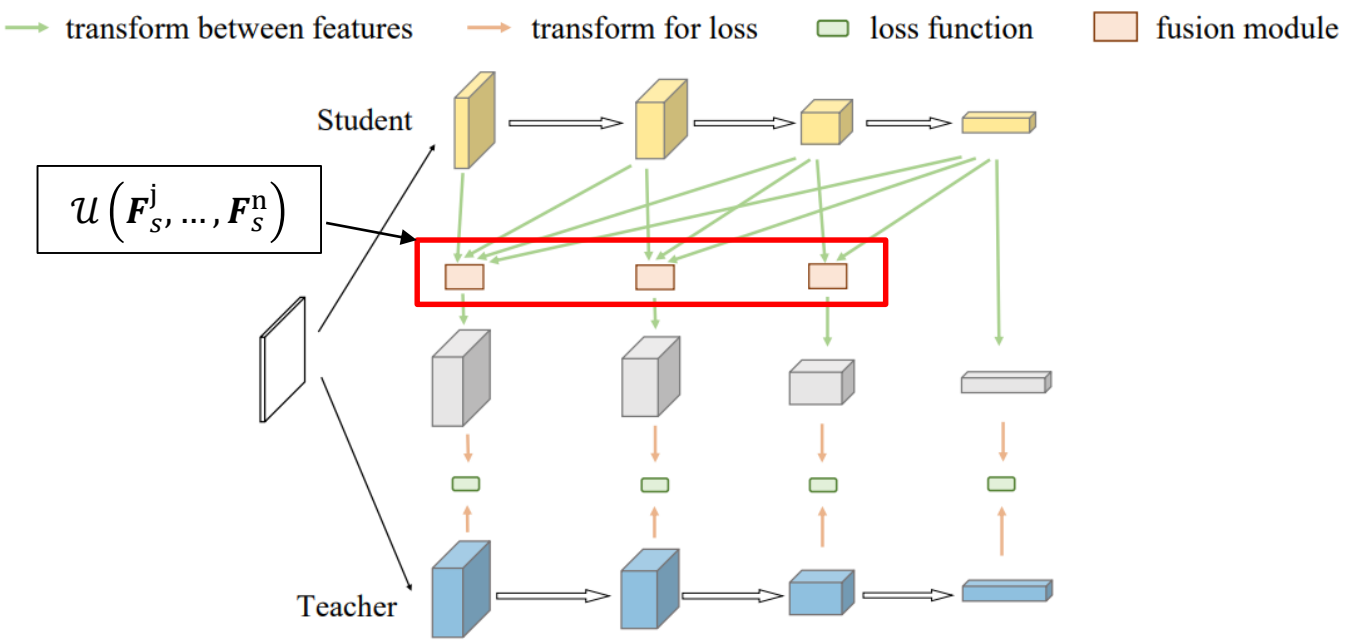


[Single-layer knowledge distillation with the review mechanism]

[Multiple-layers knowledge distillation with the review mechanism]

Method

- **Residual Learning**
  - **Fusion module $(\mathcal{U})$**
    - $\mathcal{L}_{\mathrm{MKD\_R}} = \sum_{i\in I}\left(\sum_{j=1}^{i}\mathcal{D}\left(\mathcal{M}_s^{i,j}(F_s^i), \mathcal{M}_t^{j,i}\left(F_t^j\right)\right)\right)$ ➔ $\mathcal{L}_{\mathrm{MKD\_R}} = \sum_{j=1}^{n}\sum_{i=j}^{n}\mathcal{D}\left(F_s^i, F_t^j\right)$
    - $\sum_{i=j}^{n}\mathcal{D}\left(F_s^i, F_t^j\right) \approx \mathcal{D}\left(\mathcal{U}\left(F_s^j, \dots, F_s^n\right), F_t^j\right)$



[Optimized architecture with fusion modules]

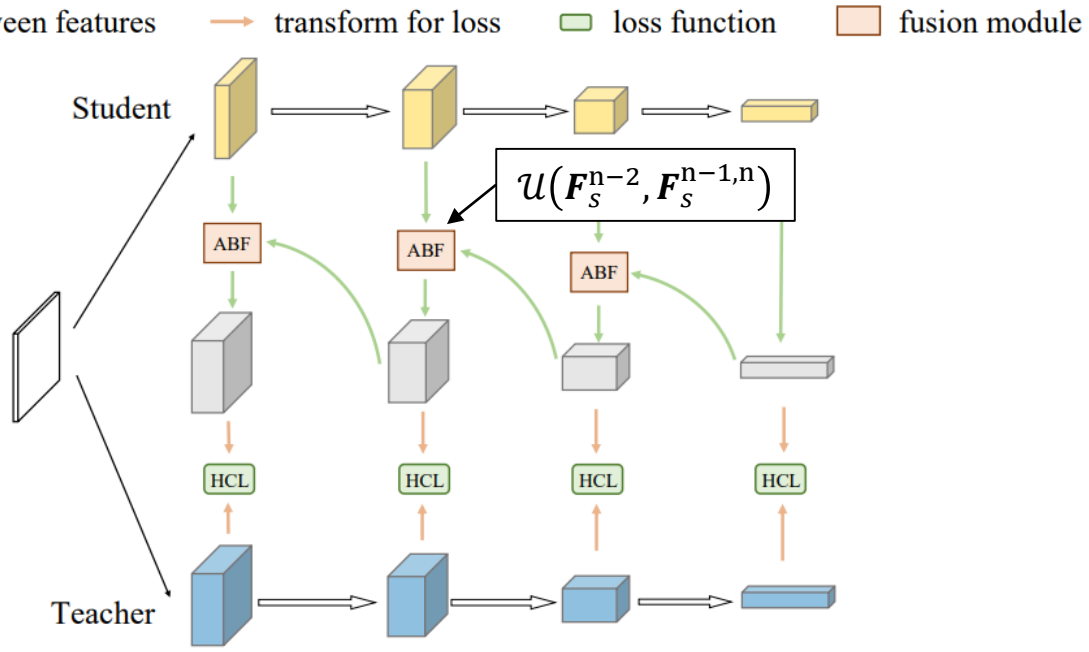
Method

Residual Learning

cf. Residual learning
$F_4^{\mathcal{S}} + F_3^{\mathcal{S}} \rightarrow F_3^{\mathcal{T}} \Longrightarrow F_4^{\mathcal{S}} \rightarrow F_3^{\mathcal{S}} - F_3^{\mathcal{T}}$

Recursive operation

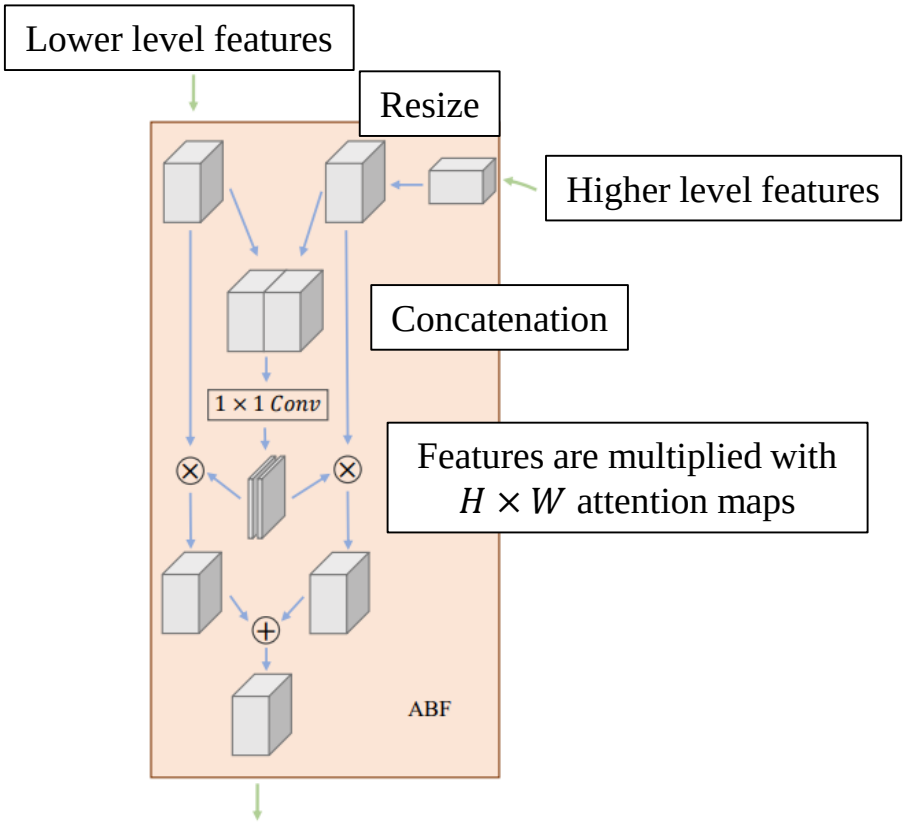
- $\mathcal{U}\left(\boldsymbol{F}_s^{\mathrm{j}}, \boldsymbol{F}_s^{\mathrm{j+1}}, \ldots, \boldsymbol{F}_s^{\mathrm{n}}\right)$ ➔ combination of $\boldsymbol{F}_s^{\mathrm{j}}$ and $\mathcal{U}\left(\boldsymbol{F}_s^{\mathrm{j+1}}, \ldots, \boldsymbol{F}_s^{\mathrm{n}}\right)$
- $\therefore \mathcal{L}_{\mathrm{MKD\_R}} = \mathcal{D}(\boldsymbol{F}_s^{\mathrm{n}}, \boldsymbol{F}_t^{\mathrm{n}}) + \sum_{\mathrm{j=n-1}}^{1} \mathcal{D}\left(\mathcal{U}\left(\boldsymbol{F}_s^{\mathrm{j}}, \boldsymbol{F}_s^{\mathrm{j+1,n}}\right), \boldsymbol{F}_t^{\mathrm{j}}\right)$
  - $\boldsymbol{F}_s^{\mathrm{j+1,n}}$ denotes a fusion of features from $\boldsymbol{F}_s^{\mathrm{j+1}}$ to $\boldsymbol{F}_s^{\mathrm{n}}$



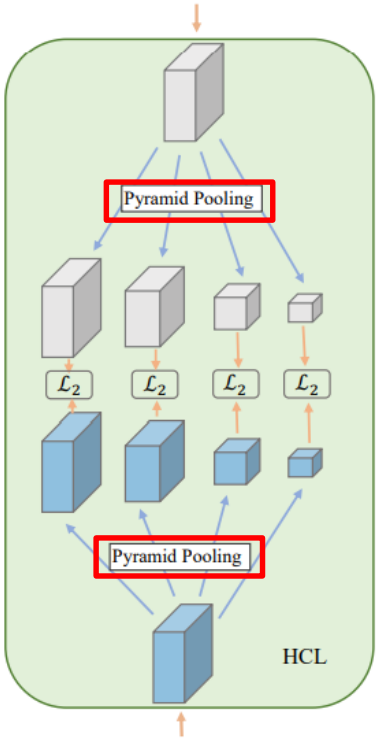
[Final architecture]

Method

- **ABF and HCL**
  - **ABF (attention based fusion)**
  - **HCL (hierarchical context loss)**



[Architecture of ABF]



[Architecture of HCL]

Experiments

Classification

Results on CIFAR 100 (same architectures)

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**

[Results on CIFAR-100 with the teacher and student having same architectures.]

Experiments

Classification

Results on CIFAR 100 (different architectures)

Distillation Mechanism	Teacher Acc	ResNet32x4 79.42	WRN40-2 75.61	VGG13 74.64	ResNet50 79.34	ResNet32x4 79.42
	Student Acc	ShuffleNetV1 70.50	ShuffleNetV1 70.50	MobileNetV2 64.6	MobileNetV2 64.6	ShuffleNetV2 71.82
Logits	KD [9]	74.07	74.83	67.37	67.35	74.45
Single Layer	FitNet [25]	73.59	73.73	64.14	63.16	73.54
Single Layer	PKT [23]	74.10	73.89	67.13	66.52	74.69
Single Layer	RKD [22]	72.28	72.21	64.52	64.43	73.21
Single Layer	CRD [28]	75.11	76.05	69.73	69.11	75.65
Multiple Layers	AT [38]	71.73	73.32	59.40	58.58	72.73
Multiple Layers	VID [1]	73.38	73.61	65.56	67.57	73.40
Multiple Layers	OFD [8]	75.98	75.85	69.48	69.04	76.82
Review	Ours	**77.45**	**77.14**	**70.37**	**69.89**	**77.78**

[Results on CIFAR-100 with the teacher and student having different architectures.]

Experiments

- **Classification**
  - **Results on ImageNet**
    - (a) – different architectures (student: MobileNet, teacher: ResNet50)
    - (b) – same architectures (student: ResNet18, teacher: ResNet34)

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**

[Results on ImageNet]

Experiments

Object Detection

Results on 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**
Teacher	Faster R-CNN w/ R50-FPN	40.22	61.02	43.81	51.98	43.53	24.16
Student	Faster R-CNN w/ MV2-FPN	29.47	48.87	30.90	38.86	30.77	16.33
	w/ KD [9]	30.13 (+0.66)	50.28	31.35	39.56	31.91	16.69
	w/ FitNet [25]	30.20 (+0.73)	49.80	31.69	39.69	31.64	16.39
	w/ FGFI [31]	31.16 (+1.69)	50.68	32.92	42.12	32.63	16.73
	w/ Our Method	**33.71 (+4.24)**	**53.15**	**36.13**	**46.47**	**35.81**	**16.77**
Teacher	RetinaNet101	40.40	60.25	43.19	52.18	44.34	24.03
Student	RetinaNet50	36.15	56.03	38.73	46.95	40.25	21.37
	w/ KD [9]	36.76 (+0.61)	56.60	39.40	48.17	40.56	21.87
	w/ FitNet [25]	36.30 (+0.15)	55.95	38.95	47.14	40.32	20.10
	w/ FGFI [31]	37.29 (+1.14)	57.13	40.04	49.71	41.47	21.01
	w/ Our Method	**38.48 (+2.33)**	**58.22**	**41.46**	**51.15**	**42.72**	**22.67**

[Results on object detection]

Experiments

- **Analysis**
  - **Ablation study**
    - Student: Wide Residual Network (WRN16-2)
    - Teacher: WRN40-2
    - CIFAR100 dataset

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

[RM: Review mechanism.
RLF: Residual learning framework.
ABF: Attention based fusion module.
HCL: Hierarchical context loss function.]

Conclusion

- **Contributions**
  - **The *review mechanism*, which uses multiple layers in the teacher was proposed to supervise one layer in the student.**
  - **Significant improvement on all classification, detection, and segmentation**

Thank you.