Multi-level Logit Distillation

Ying Jin[1] Jiaqi Wang[2] Dahua Lin[1,2]
[1]The Chinese University of Hong Kong
[2]Shanghai AI Laboratory
THU-PM-351

Multi-Level Logit Distillation

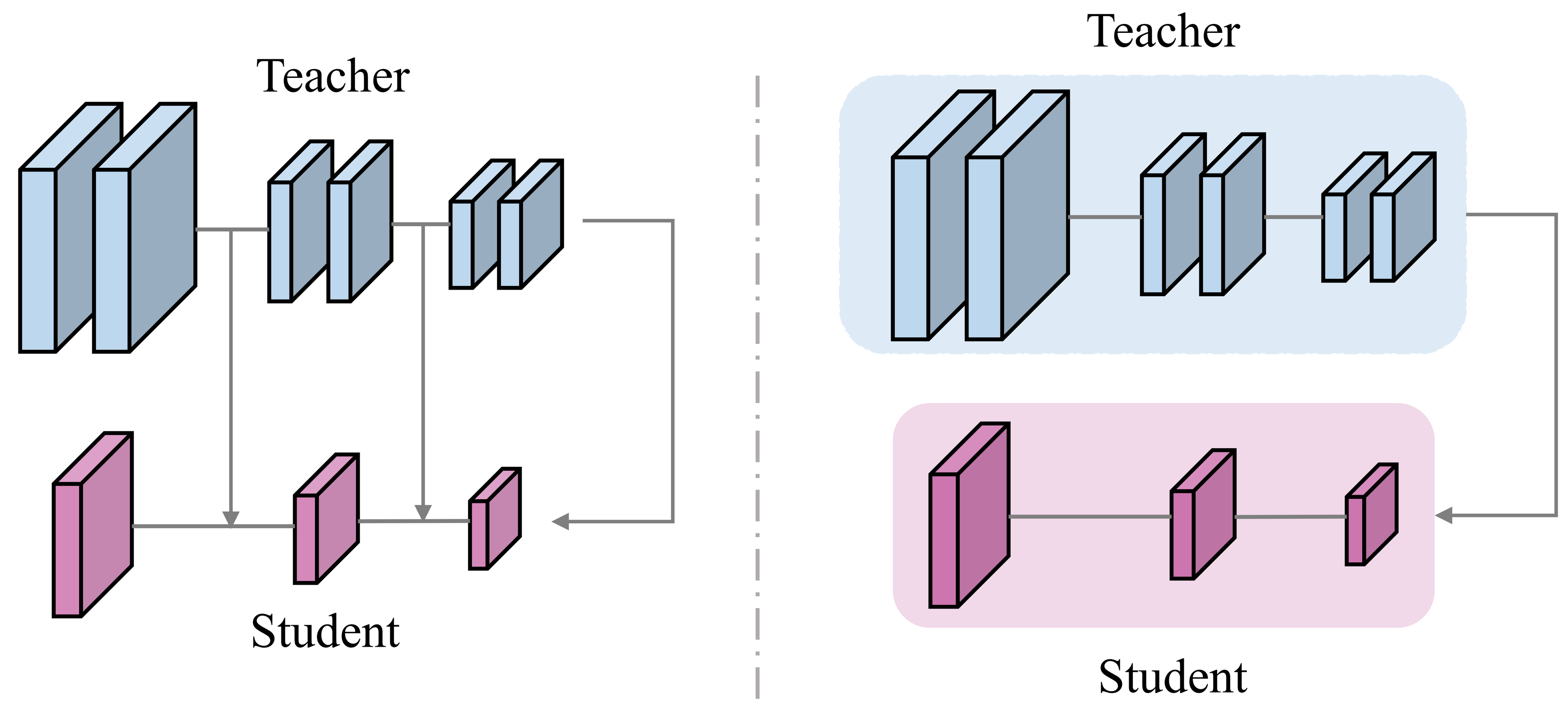
Ying Jin, Jiaqi Wang, and Dahua Lin
The Chinese University of Hong Kong, Shanghai Artificial Intelligence Laboratory

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Feature Distillation V.S. Logit Distillation

Feature Distillation
- Convey knowledge in both feature and logit level

Logit Distillation
- Convey knowledge in merely logit level
- Intermediate layers in the teacher model is invisible

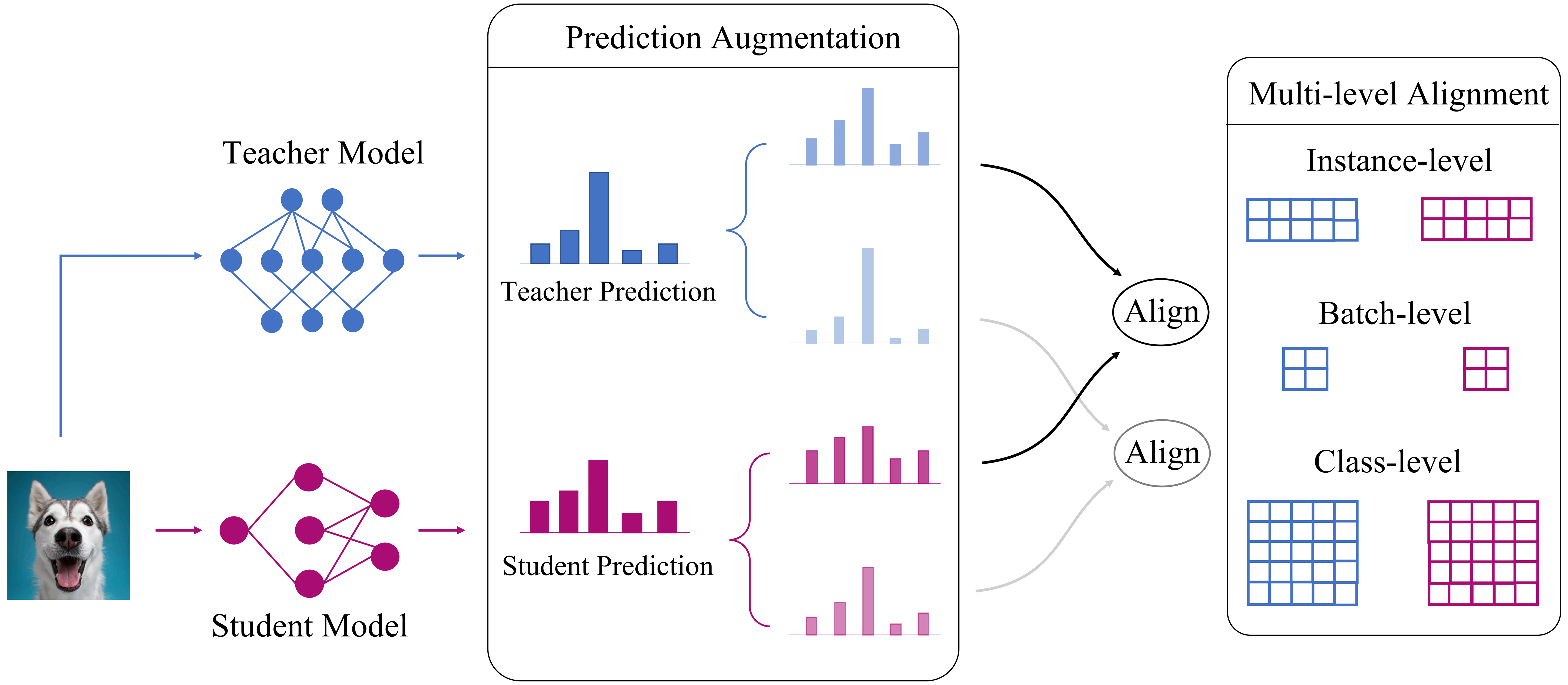


Feature Distillation Logit Distillation

Our Motivation

Multi-level Logit Distillation
- Multi-level Alignment: instance, class, batch level
- Merely on logit outputs
- Prediction Augmentation: further enhance diversity

Method Framework



Efficacy and Results:

Our Method
- Surpass previous logit distillation methods
- Comparable with feature distillation methods

Method	Teacher / Student	ResNet56 72.34 / ResNet20 69.06	ResNet110 74.31 / ResNet32 71.14	ResNet32×4 79.42 / ResNet8×4 72.50	WRN-40-2 75.61 / WRN-16-2 73.26	WRN-40-2 75.61 / WRN-40-1 71.98	VGG13 74.64 / VGG8 70.36	Avg
Feature	FitNet [22]	69.21	71.06	73.50	73.58	72.24	71.02	71.77
	RKD [19]	69.61	71.82	71.90	73.35	72.22	71.48	71.73
	CRD [27]	71.16	73.48	75.51	75.48	74.14	73.94	73.95
	OFD [8]	70.98	73.23	74.95	75.24	74.33	73.95	73.78
	ReviewKD [1]	71.89	73.89	75.63	76.12	75.09	74.84	74.58
Logit	KD [10]	70.66	73.08	73.33	74.92	73.54	72.98	73.09
	DML [35]	69.52	72.03	72.12	73.58	72.68	71.79	71.95
	TAKD [18]	70.83	73.37	73.81	75.12	73.78	73.23	73.36
	Ours	**72.19**	**74.11**	**77.08**	**76.63**	**75.35**	**75.18**	**75.09**

Knowledge Distillation

Knowledge Distillation

- Convey knowledge from a big teacher model to a lightweight student model

Feature Distillation

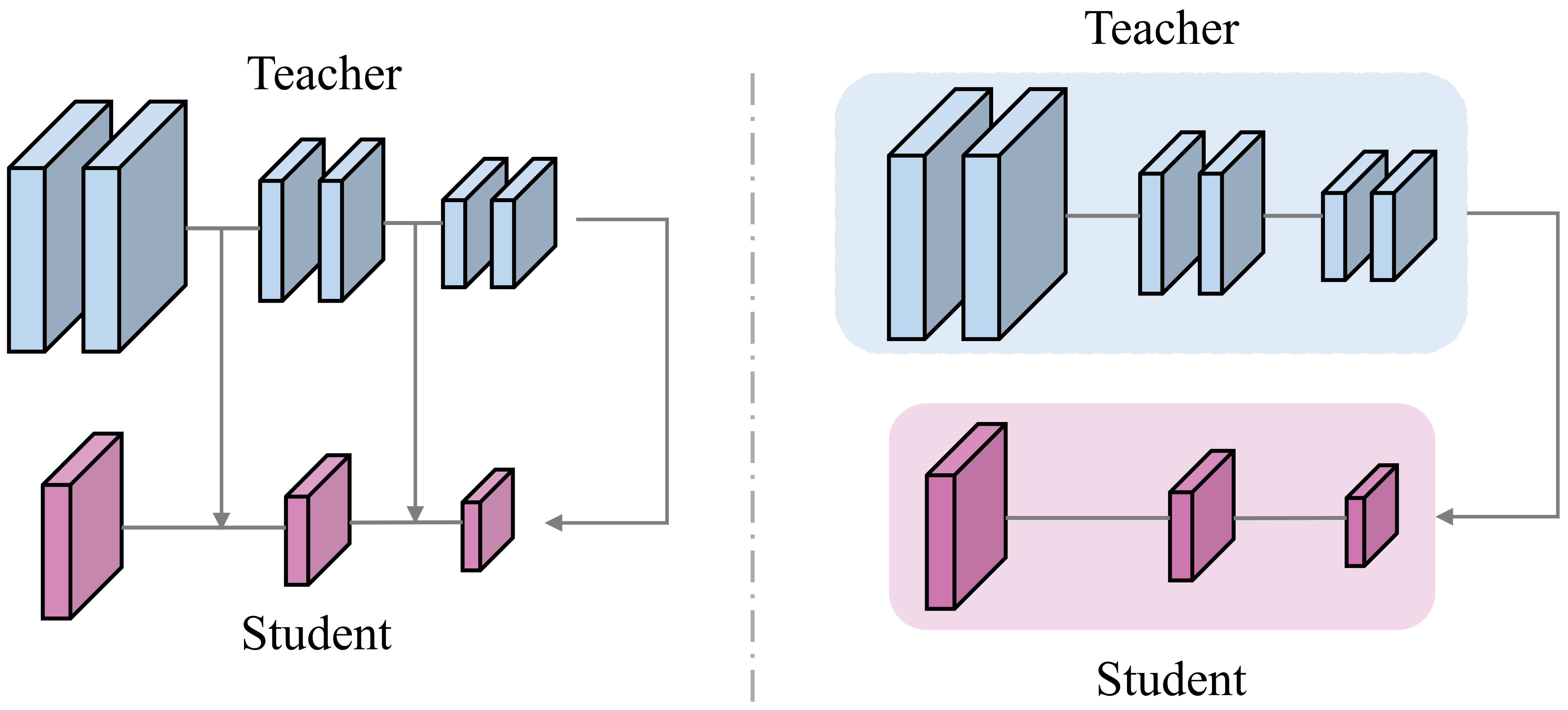
- Convey knowledge in both feature and logit level

Logit Distillation

- Convey knowledge in merely logit level
- Intermediate layers in the teacher model are invisible

We focus on Logit Distillation

- Performance is always inferior to feature distillation



Feature Distillation

Logit Distillation

Multi-level Logit Distillation: Preliminaries

Logit output

$$p_j = \frac{e^{z_j/T}}{\sum_{c=1}^{C} e^{z_c/T}},$$

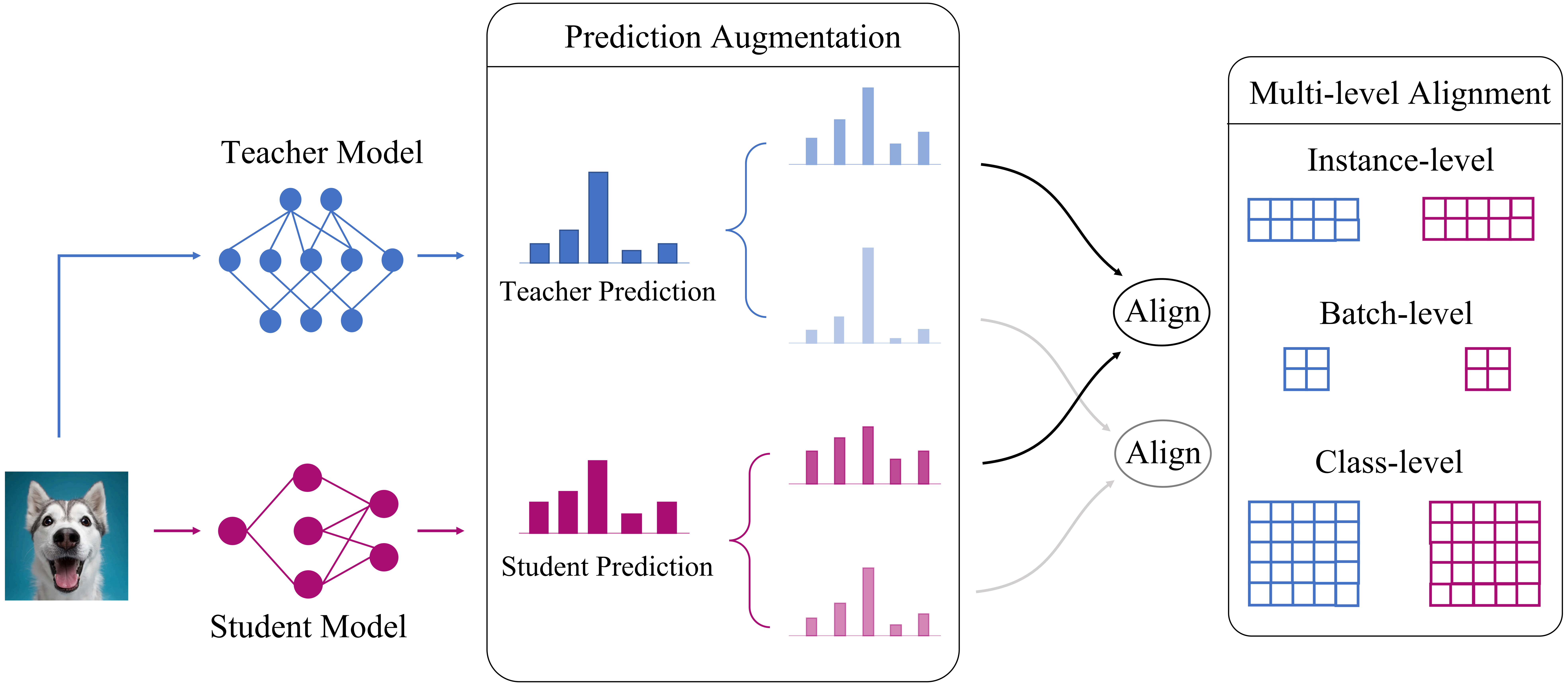
Knowledge distillation loss

$$L_{KD} = KL(p^{tea}||p^{stu}) = \sum_{j=1}^{C} p_j^{tea} log(\frac{p_j^{tea}}{p_j^{stu}}),$$

Multi-level Logit Distillation

Multi-level Logit Distillation

- Multi-level Alignment: instance, class, and batch level alignment
- Merely on logit outputs
- Prediction Augmentation: further enhance diversity



Multi-level Logit Distillation

Prediction augmentation

- Gain richer knowledge from predictions
- Temperature scaling

$$p_{i,j,k} = \frac{e^{z_{i,j}/T_k}}{\sum_{c=1}^{C} e^{z_{i,c}/T_k}},$$

Instance-level alignment

- Inherit the original mechanism in KD
- Minimize the KL divergence between augmented predictions

$$\begin{aligned} L_{ins} &= \sum_{i=1}^{N}\sum_{k=1}^{K} KL(p_{i,k}^{tea} \| p_{i,k}^{stu}) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{C} p_{i,j,k}^{tea} log(\frac{p_{i,j,k}^{tea}}{p_{i,j,k}^{stu}}), \end{aligned}$$

Multi-level Logit Distillation

Batch-level alignment

- Conduct batch-level alignment by input correlation , the relation between two inputs
- Modeled via features in previous works
- We take logit predictions to quantify it by Gram Matrix

$$G^k = p_k p_k^T, G_{ab}^k = \sum_{j=1}^{C} p_{a,j,k} \cdot p_{b,j,k},$$

$$L_{batch} = \frac{1}{B} \sum_{k=1}^{K} ||G_{tea}^k - G_{stu}^k||_2^2,$$

Multi-level Logit Distillation

Class-level alignment

- Model predictions can depict the relationship between categories
- Enforce the student model to absorb this part of knowledge

$$M^k = p_k^T p_k, M_{ab}^k = \sum_{i=1}^{N} p_{i,a,k} \cdot p_{i,b,k},$$

$$L_{class} = \frac{1}{C} \sum_{k=1}^{K} ||M_{tea}^k - M_{stu}^k||_2^2$$

Multi-level alignment

$$L_{total} = L_{ins} + L_{batch} + L_{class}.$$

Multi-level Logit Distillation

Algorithm 1: Pseudo-code in a PyTorch-like style.

```
# z_stu, z_tea: student, teacher logit outputs
# T = [T_1, T_2, ..., T_K]:
#      one set of K different temperatures
# l_ins, l_batch, l_class:
#      three parts of alignment loss
# l_total: total loss
l_total = 0
for t in T do
    p_stu = F.softmax(z_stu / t) # B x C
    p_tea = F.softmax(z_tea / t) # B x C
    l_ins = F.kl_div(p_tea, p_stu)
    G_stu = torch.mm(p_stu, p_stu.t()) # B x B
    G_tea = torch.mm(p_tea, p_tea.t()) # B x B
    l_batch = ((G_stu - G_tea) ** 2).sum() / B
    M_stu = torch.mm(p_stu.t(), p_stu) # C x C
    M_tea = torch.mm(p_tea.t(), p_tea) # C x C
    l_class = ((M_stu - M_tea) ** 2).sum() / C
    l_total += (l_ins + l_batch + l_class)
end
```

Multi-level Logit Distillation

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Table 2. **Results on CIFAR-100, Homogenous Architecture.** Top-1 accuracy is adopted as the evaluation metric. The teacher model and student model are in homogenous architecture and their original performance is reported respectively.

Method	Teacher	ResNet56	ResNet110	ResNet32×4	WRN-40-2	WRN-40-2	VGG13	Avg
		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	
Feature	FitNet [22]	69.21	71.06	73.50	73.58	72.24	71.02	71.77
	RKD [19]	69.61	71.82	71.90	73.35	72.22	71.48	71.73
	CRD [27]	71.16	73.48	75.51	75.48	74.14	73.94	73.95
	OFD [8]	70.98	73.23	74.95	75.24	74.33	73.95	73.78
	ReviewKD [1]	71.89	73.89	75.63	76.12	75.09	74.84	74.58
Logit	KD [10]	70.66	73.08	73.33	74.92	73.54	72.98	73.09
	DML [35]	69.52	72.03	72.12	73.58	72.68	71.79	71.95
	TAKD [18]	70.83	73.37	73.81	75.12	73.78	73.23	73.36
	Ours	**72.19**	**74.11**	**77.08**	**76.63**	**75.35**	**75.18**	**75.09**

Multi-level Logit Distillation

Table 3. **Results on CIFAR-100, Heterogeneous Architecture.** Top-1 accuracy is adopted as the evaluation metric. The teacher model and student model are in heterogeneous architecture and their original performance is reported respectively.

Method	Teacher / Student	ResNet32×4 79.42 ShuffleNet-V1 70.50	WRN-40-2 75.61 ShuffleNet-V1 70.50	VGG13 74.64 MobileNet-V2 64.60	ResNet50 79.34 MobileNet-V2 64.60	ResNet32×4 79.42 ShuffleNet-V2 71.82	Avg
Feature	FitNet [22]	73.59	73.73	64.14	63.16	73.54	69.63
	RKD [19]	72.28	72.21	64.52	64.43	73.21	69.33
	CRD [27]	75.11	76.05	69.73	69.11	75.65	73.13
	OFD [8]	75.98	75.85	69.48	69.04	76.82	73.43
	ReviewKD [1]	**77.45**	77.14	70.37	69.89	77.78	74.53
Logit	KD [10]	74.07	74.83	67.37	67.35	74.45	71.60
	DML [35]	72.89	72.76	65.63	65.71	73.45	70.09
	TAKD [18]	74.53	75.34	67.91	68.02	74.82	72.12
	Ours	77.18	**77.44**	**70.57**	**71.04**	**78.44**	**74.93**

Multi-level Logit Distillation

Table 4. **Results on ImageNet.** Top-1 and Top-5 accuracy is adopted as the evaluation metric. The original accuracies of the teacher and student model are also reported.

		Top-1	Top-5	Top-1	Top-5
Method	Teacher	ResNet34		ResNet50	
		73.31	91.42	76.16	92.86
	Student	ResNet18		MobileNet-V2	
		69.75	89.07	68.87	88.76
Feature	AT [33]	70.69	90.01	69.56	89.33
	OFD [8]	70.81	89.98	71.25	90.34
	CRD [27]	71.17	90.13	71.37	90.41
	ReviewKD [1]	71.61	90.51	72.56	91.00
Logit	KD [10]	70.66	89.88	68.58	88.98
	DML [35]	70.82	90.02	71.35	90.31
	TAKD [18]	70.78	90.16	70.82	90.01
	DKD [36]	71.70	90.41	72.05	91.05
	Ours	**71.90**	**90.55**	**73.01**	**91.42**

Multi-level Logit Distillation

Instance-level Alignment	Batch-level Alignment	Class-level Alignment	Prediction Augmentation	Acc
✓	✗	✗	✗	73.33
✓	✓	✗	✗	74.58
✓	✓	✓	✗	76.26
✓	✓	✓	✓	**77.08**

Teacher	72.34	74.31	79.42	75.61	75.61	74.64	75.32 (Avg)
Student (Ours)	72.19	74.11	77.08	76.63	75.35	75.18	75.09 (Avg)
Gap	0.15	0.20	2.34	-1.02	0.26	- 0.54	0.23 (Avg)

Thank you