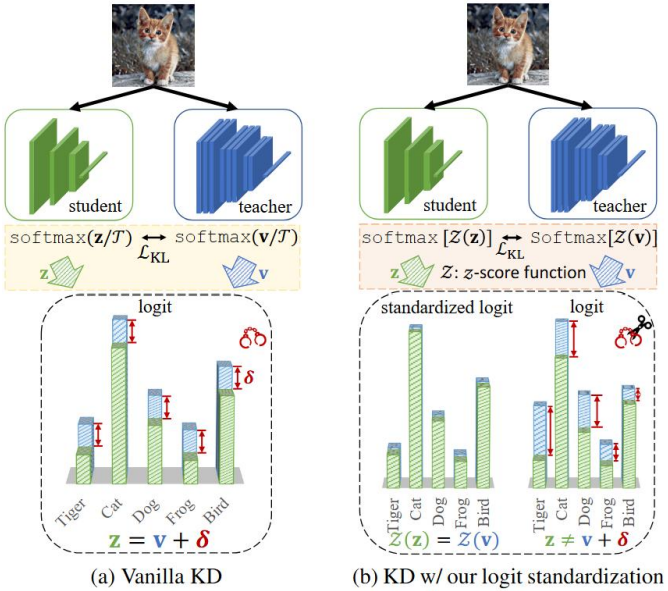
VEHICLE 250904

Knowledge Distillation with Refined Logits

Wujie Sun[1,2,3] Defang Chen[4†] Siwei Lyu[4] Genlang Chen[5] Chun Chen[1,3] Can Wang[1,3]
[1]State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2]School of Software Technology, Zhejiang University
[3]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[4]University at Buffalo, State University of New York [5]NingboTech University
sunwujie@zju.edu.cn, {defangch, siweilyu}@buffalo.edu, {cgl, chenc, wcan}@zju.edu.cn

- Problem/Objective
  - 정통 KD (Logit distillation task)
  - Image Classification
- Contribution/Key Idea
  - Prove that conventional distillation is wrong
  - class correlations을 고려한 distillation 방법 제시



Logit Standardization in KD
(CVPR 2024 Highlight)

- **기존 연구 분석 for KD**

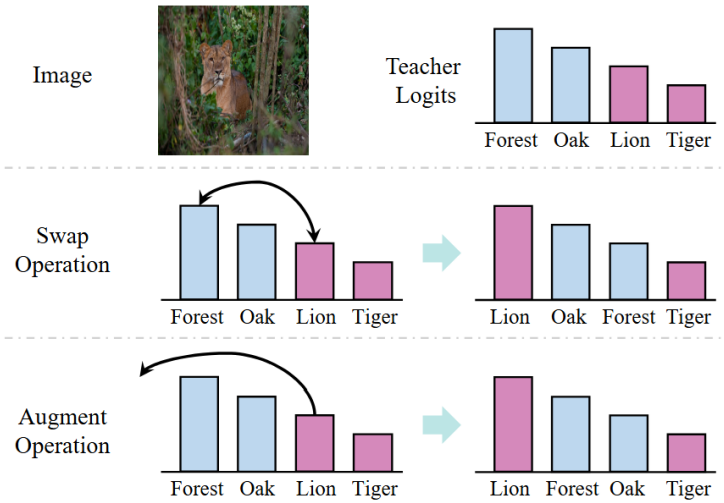
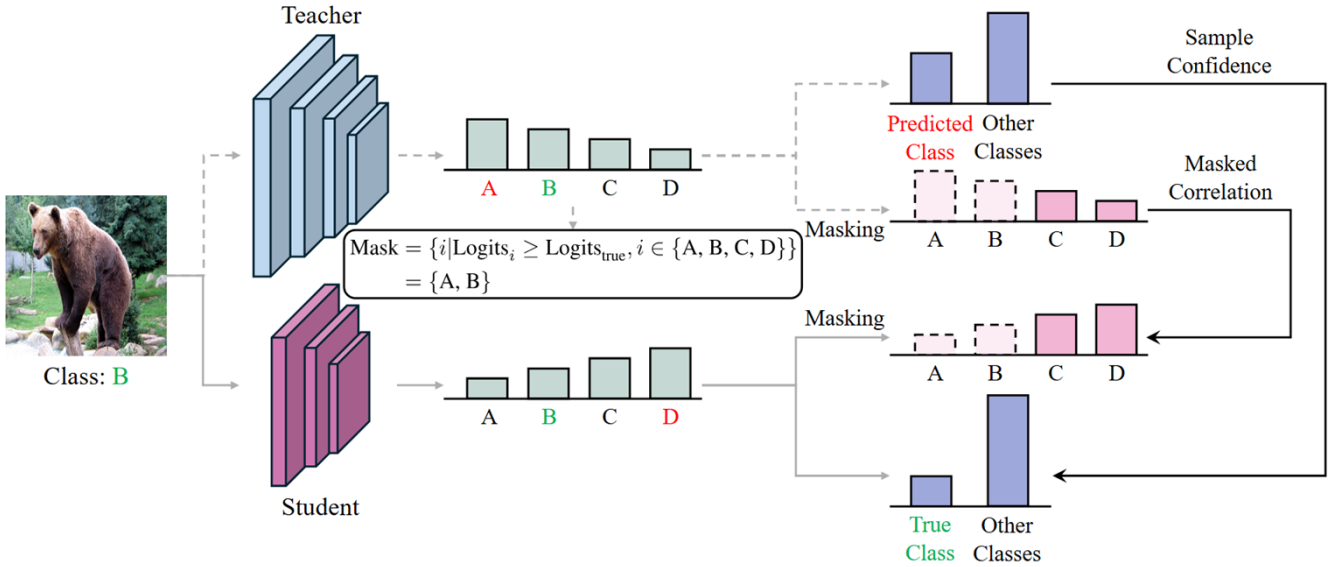


Figure 1. A toy example of existing correction-based distillation approaches. **Classes represented by the same color are highly correlated and should be ranked closely.** The image displayed is a "lion", yet the teacher model incorrectly classifies it as the "forest". Both the swap and augment operations disrupt the close correlation between "lion" and "tiger". A more detailed example of class correlation is provided in the Appendix.

- 기존 연구 1: Swap Operation (Forest ⟷ Lion)
  - 정답 클래스와 ⟷ 예측된 클래스 점수를 강제로 바꿔 치기(Swap)
- 기존 연구 2: Augment Operation (Lion ↑)
  - 정답 클래스의 확률을 인위적으로 키워줌

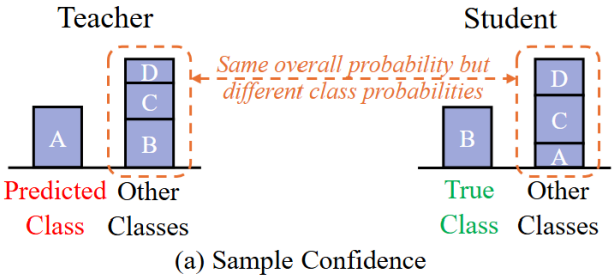
- 두 연구 모두 Class correlation이 깨진다는 한계점
  - 비슷한 수준의 prediction 값을 가져야 함
  - ex) Lion & Tiger pair

- **Method**

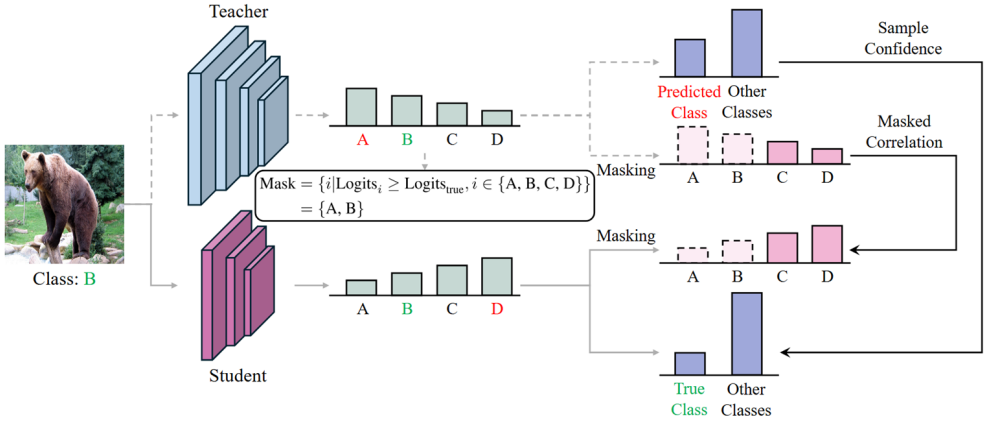


- 본 연구에서는 Class correlation을 maintain하면서 distillation 성능을 향상시킬 수 있는 2가지 방법을 제시
  - Sample Confidence (SC)
  - Masked Correlation (MC)

- **Method - 1st**



(a) Sample Confidence



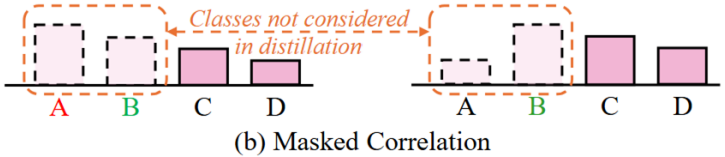
- Sample Confidence (SC)
  - Teacher가 가장 높게 예측한 class의 “확률”만 align **(Class x)**
  - “Confidence” 분포만 align

$$b^{\mathrm{T}} = \{\hat{p}^{\mathrm{T}}_{\mathrm{max}}, 1 - \hat{p}^{\mathrm{T}}_{\mathrm{max}}\},$$

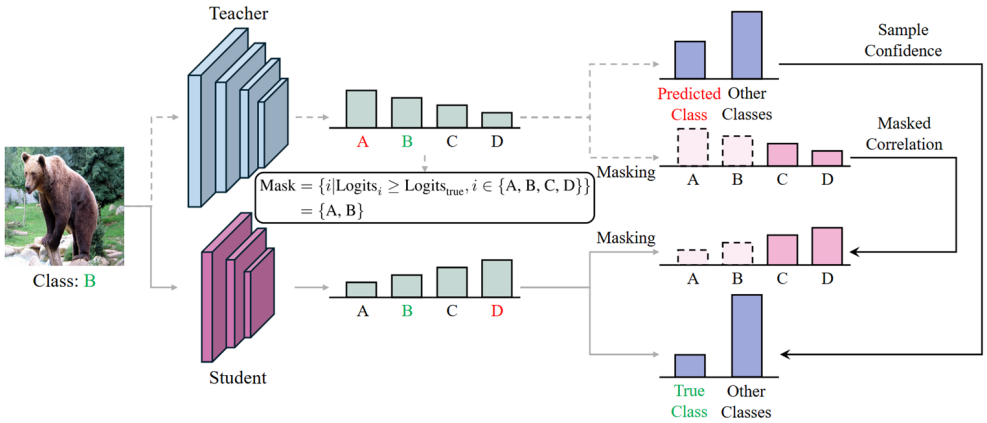
$$b^{\mathrm{S}} = \{\hat{p}^{\mathrm{S}}_{\mathrm{true}}, 1 - \hat{p}^{\mathrm{S}}_{\mathrm{true}}\}.$$

$$L_{\mathrm{SCD}} = \tau^2 \mathrm{KL}\left(b^{\mathrm{T}}, b^{\mathrm{S}}\right)$$

- **Method - 2nd**



(b) Masked Correlation



- Masked Correlation (MC)
  - Teacher가 정답보다 높게 예측한 class들은 마스킹 처리 (제거)
  - 남은 class들 정규화하여 distill

$$M_{\text{ge}} = \{i | z_i^{\text{T}} \geq z_{\text{true}}^{\text{T}}, 1 \leq i \leq C\}$$

$$\tilde{p}_i = \frac{\exp\left(z_i/\tau\right)}{\sum_{c=1, c \notin M_{\text{ge}}}^{C} \exp\left(z_c/\tau\right)}$$

$$L_{\text{MCD}} = \tau^2 \text{KL}\left(\tilde{p}^{\text{T}}, \tilde{p}^{\text{S}}\right)$$

Method - total

- Sample Confidence (SC)
  - 잘못된 class의 이름은 따라가지 않고 confidence만 따라감
  - CE와 유사한 binary correlation but flexibility

- Masked Correlation (MC)
  - Teacher가 맞으면 → 대부분의 class correlation 보존
  - Teacher가 틀리면 → mis-info 줄이고, correlation의 일부만 보존

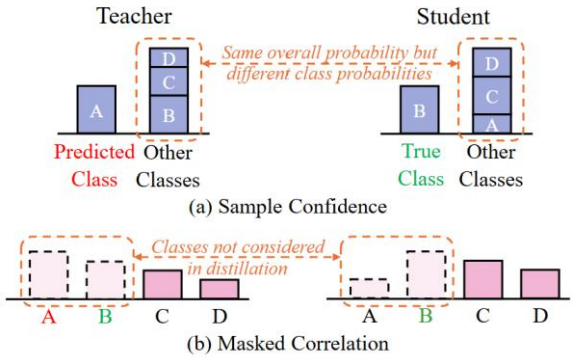


Figure. RLD framework

$$L_{\mathrm{RLD}} = L_{\mathrm{CE}} + \alpha L_{\mathrm{SCD}} + \beta L_{\mathrm{MCD}},$$

Method - total

- Decoupled Knowledge Distillation (DKD) [1] , CVPR 2022와 비교
  - 정답 / not 정답 분포를 decouple해서 각각 distill 하는 전략
    - Teacher 정확 :
      - RLD = DKD
    - Teacher 틀릴 때 :
      - DKD는 잘못된 logit 포함해서 align
      - RLD는 SC + MC로 강하게 틀린 class 제거,
      - 나머지로만 align

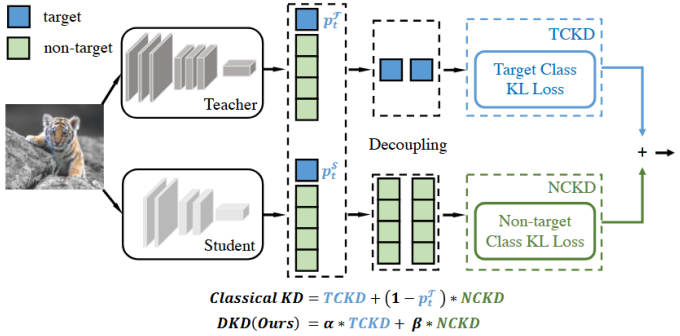


Figure. DKD framework

Relevance to DKD. Although RLD and DKD [50] consider logit distillation from distinct perspectives, they become equivalent when the teacher model consistently makes accurate predictions. Besides, DKD does not explicitly explain why transferring non-target class knowledge (i.e., the probability distribution when the true class is masked, referred to as NCKD) can significantly enhance model performance. Beyond the idea that the class relationships embedded in this knowledge facilitate training, RLD offers a new

[1] Zhao, Borui, et al. "Decoupled knowledge distillation." Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2022.

- **Experiment - CIFAR-100 (Homogeneous)**

Type	Teacher	ResNet32×4	VGG13	WRN-40-2	ResNet56	ResNet110	ResNet110
		79.42	74.64	75.61	72.34	74.31	74.31
	Student	ResNet8×4	VGG8	WRN-40-1	ResNet20	ResNet32	ResNet20
		72.50	70.36	71.98	69.06	71.14	69.06
Feature	FitNet	73.50	71.02	72.24	69.21	71.06	68.99
	AT	73.44	71.43	72.77	70.55	72.31	70.65
	RKD	71.90	71.48	72.22	69.61	71.82	69.25
	CRD	75.51	73.94	74.14	71.16	73.48	71.46
	OFD	74.95	73.95	74.33	70.98	73.23	71.29
	ReviewKD	75.63	74.84	*75.09*	71.89	73.89	71.34
	SimKD	*78.08*	74.89	74.53	71.05	73.92	71.06
	CAT-KD	76.91	74.65	74.82	71.62	73.62	71.37
Logit	KD	73.33	72.98	73.54	70.66	73.08	70.67
	CTKD	73.39	73.52	73.93	71.19	73.52	70.99
	DKD	<u>76.32</u>	<u>74.68</u>	<u>74.81</u>	<u>71.97</u>	**74.11**	71.06
	LA	73.46	73.51	73.75	71.24	73.39	70.86
	RC	74.68	73.37	74.07	71.63	73.44	<u>71.41</u>
	LR	76.06	74.66	74.42	70.74	73.52	70.61
	RLD (ours)	**76.64**	**74.93**	**74.88**	**72.00**	<u>74.02</u>	**71.67**

Table 1. Top-1 accuracy (%) on the CIFAR-100 validation set when the teacher and student models are homogeneous. The best and second best results of logit distillation are highlighted in **bold** and <u>underlined</u> text, respectively. For the case where the best result of feature distillation is better than the best result of logit distillation, we highlight it with *italic* text. The reported results are the mean of three trials.

- **Experiment - CIFAR-100 (Heterogeneous)**

Type	Teacher	ResNet32×4	ResNet32×4	WRN-40-2	WRN-40-2	VGG13	ResNet50
		79.42	79.42	75.61	75.61	74.64	79.34
	Student	SHN-V2	WRN-40-2	ResNet8×4	MN-V2	MN-V2	MN-V2
		71.82	75.61	72.50	64.60	64.60	64.60
Feature	FitNet	73.54	77.69	74.61	68.64	64.16	63.16
	AT	72.73	77.43	74.11	60.78	59.40	58.58
	RKD	73.21	77.82	75.26	69.27	64.52	64.43
	CRD	75.65	78.15	75.24	70.28	69.73	69.11
	OFD	76.82	79.25	74.36	69.92	69.48	69.04
	ReviewKD	77.78	78.96	74.34	*71.28*	*70.37*	69.89
	SimKD	78.39	*79.29*	75.29	70.10	69.44	69.97
	CAT-KD	*78.41*	78.59	75.38	70.24	69.13	*71.36*
Logit	KD	74.45	77.70	73.97	68.36	67.37	67.35
	CTKD	75.37	77.66	74.61	68.34	68.50	68.67
	DKD	77.07	78.46	75.56	69.28	69.71	70.35
	LA	75.14	77.39	73.88	68.57	68.09	68.85
	RC	75.61	77.58	75.22	68.72	68.66	68.98
	LR	76.27	78.73	75.26	69.02	69.78	70.38
	RLD (ours)	**77.56**	**78.91**	**76.12**	**69.75**	**69.97**	**70.76**

Table 2. Top-1 accuracy (%) on the CIFAR-100 validation set when the teacher and student models are heterogeneous. The same convention is used as in Table 1.

- **Experiment - ImageNet**

Teacher/Student	Res34/Res18		Res50/MN-V1	
Accuracy	Top-1	Top-5	Top-1	Top-5
Teacher	73.31	91.42	76.16	92.86
Student	69.75	89.07	68.87	88.76
AT	70.69	90.01	69.56	89.33
OFD	70.81	89.98	71.25	90.34
CRD	71.17	90.13	71.37	90.41
ReviewKD	71.61	<u>90.51</u>	<u>72.56</u>	91.00
SimKD	71.59	90.48	72.25	90.86
CAT-KD	71.26	90.45	72.24	<u>91.13</u>
KD	71.03	90.05	70.50	89.80
CTKD	71.38	90.27	71.16	90.11
DKD	<u>71.70</u>	90.41	72.05	91.05
LA	71.17	90.16	70.98	90.13
RC	71.59	90.21	71.86	90.54
LR	70.29	89.98	71.76	90.93
RLD (ours)	**71.91**	**90.59**	**72.75**	**91.18**

Table 3. Top-1 and top-5 accuracy (%) on the ImageNet validation set. The best and second best results are highlighted in **bold** and <u>underlined</u> text, respectively. The reported results are the mean of three trials.

- **Experiment**

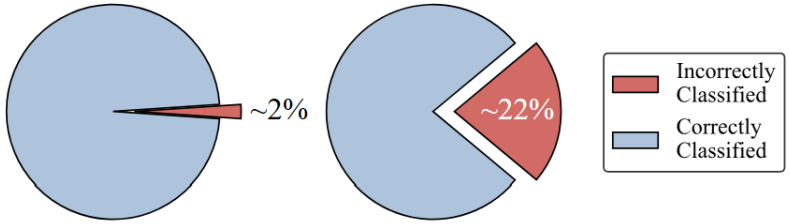


Figure 4. Proportion of predictions from teacher models on the training set. (a) ResNet56. (b) ResNet50.

Teacher	ResNet56 72.34	ResNet110 74.31	VGG13 74.64
Student	WideResNet-40-2 75.61		
KD	76.72	77.37	76.86
DKD	77.34	77.70	77.45
RLD (ours)	**78.03**	**78.28**	**77.88**
Δ	+0.69	+0.58	+0.43

Table 4. Top-1 accuracy (%) on the CIFAR-100 validation set when distilling with inferior teachers. Optimal results are highlighted in **bold**. The reported results are the mean of three trials.

↓

정답률 차이

↓

Inferior Teacher

- **Experiment**

Teacher	WRN-40-2 75.61	VGG13 74.64	ResNet50 79.34
Student	MobileNet-V2 64.60		
KD	69.23	68.61	69.02
CTKD	69.53	68.98	69.36
DKD	70.01	69.98	70.45
RLD (ours)	**70.35**	**70.63**	**71.06**

Table 5. Top-1 accuracy (%) on the CIFAR-100 validation set when training with logit standardization technique LSKD [35]. The optimal results are highlighted in **bold** text. The reported results are the mean of three trials.

↓

LSKD를 더했을때

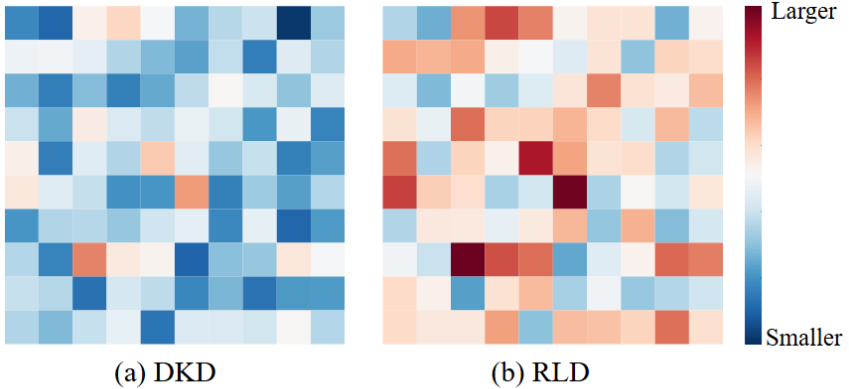


Figure 5. Visualized teacher-student logit discrepancy learned by DKD and RLD on the CIFAR-100 validation set. For better visualization, 100 classes are reshaped into a 10×10 matrix. The teacher is ResNet32×4, and the student is ResNet8×4.

↓

teacher와 student의 logit 차이

- **Experiment**

L_{CE}	L_{SCD}	L_{MCD} M_g	L_{MCD} M_{ge}	Accuracy
✓				72.50
✓	✓			73.55
✓		✓		75.50
✓			✓	75.64
✓	✓	✓		75.53
✓	✓		✓	**76.64**

Table 6. Ablation study on the importance of each component in RLD. Top-1 accuracy (%) on the CIFAR-100 validation set is reported. The teacher is ResNet32×4, and the student is ResNet8×4. The reported results are the mean of three trials.

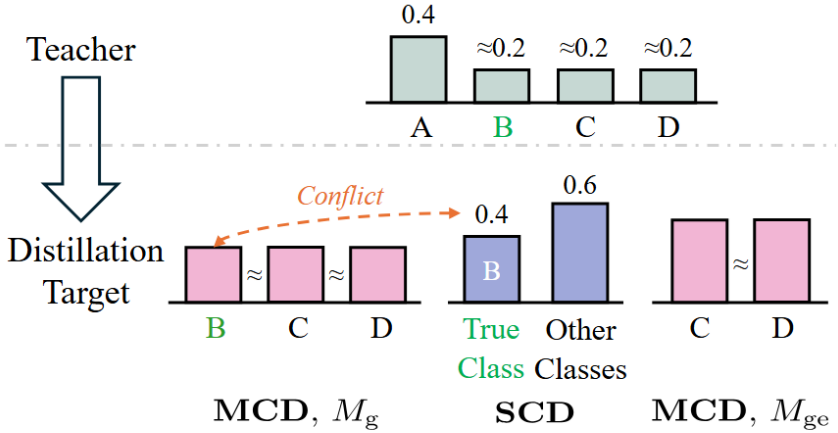


Figure 6. A toy example illustrates how the M_g masking strategy can lead to loss conflict. In this example, the probabilities of classes B, C, and D in the teacher distribution are close, and the value corresponding to B is slightly larger than those of C and D.

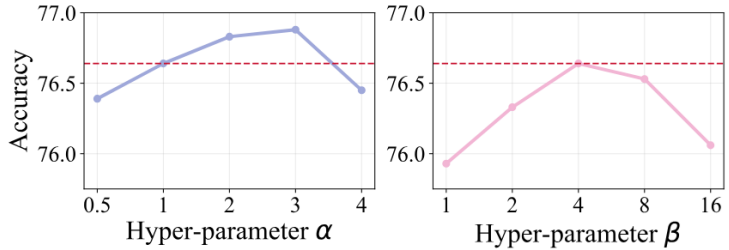
↓

equal의 포함 여부

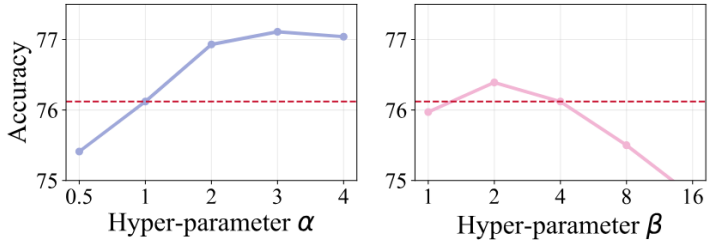
↓

왜곡 현상
B+C+D = 1.2 (?)

- **Experiment**



(a) Teacher: ResNet32x4, Student: ResNet8x4



(b) Teacher: WRN-40-2, Student: ResNet8x4

Figure 7. Impact of the hyper-parameters α (L_{SCD}) and β (L_{MCD}) on the CIFAR-100 validation set. By default, for both distillation pairs, $\alpha = 1$ and $\beta = 4$ (corresponding to the accuracies reported in Section 5.2, marked with dashed lines). The reported results are the mean of three trials.

→

DKD와의 차이점 부각 α↑ 괜찮다.
MC가 마스킹을 잘해줘서

- **Experiment – Appendix**

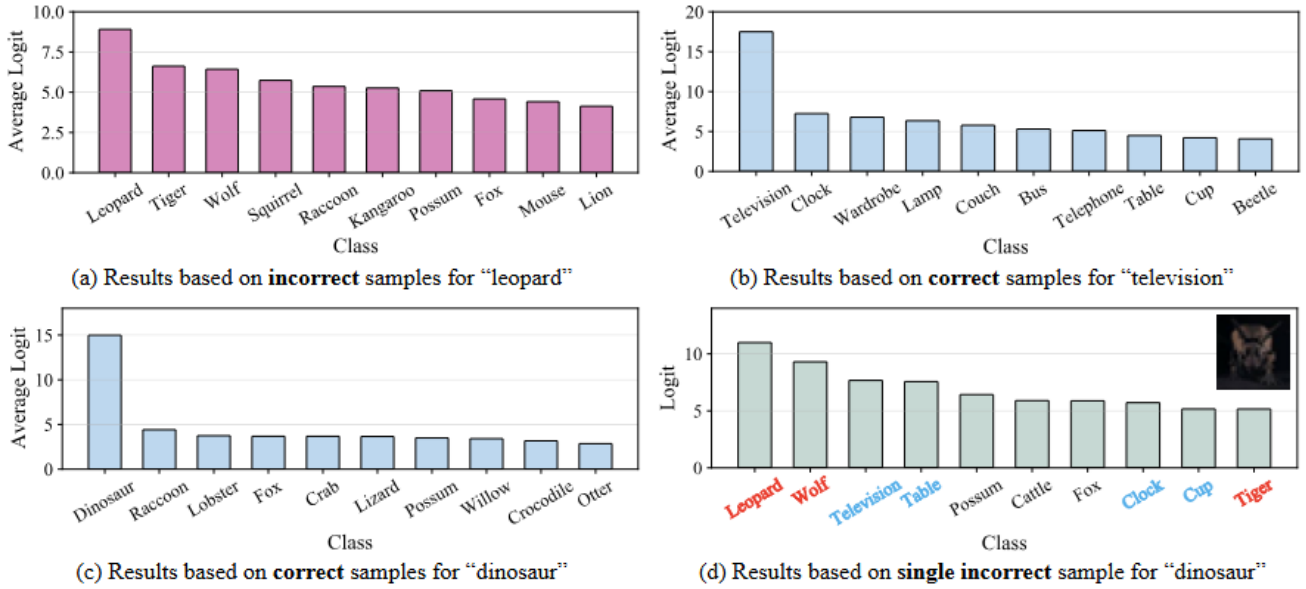


Figure 8. Prediction results for different classes on the CIFAR-100 validation set. The model for calculation is ResNet110. Bold colored classes in (d) denote that they are among the top 10 classes most similar to either "leopard" (as inferred from (a)) or "television" (as inferred from (b)), yet they do not appear in the top 10 classes most similar to "dinosaur" (as inferred from (c)).

- **Experiment – Appendix**

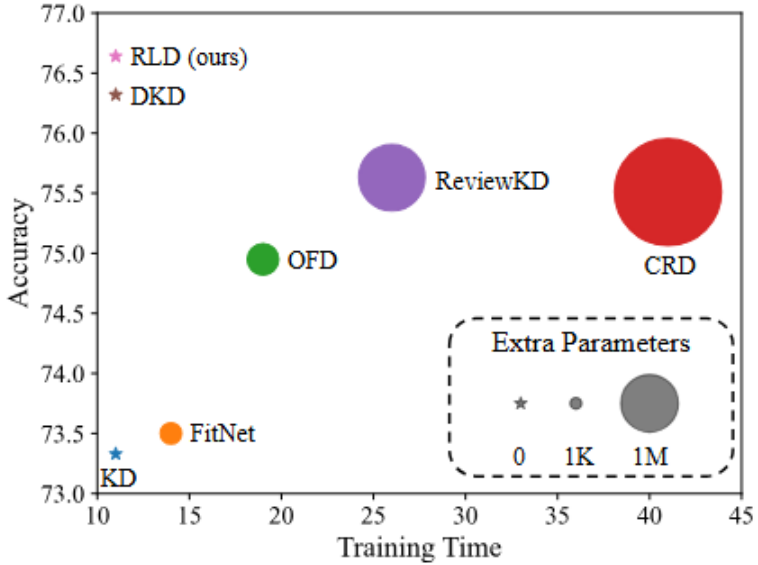


Figure 9. Batch training time (ms) *vs.* top-1 validation accuracy (%) on the CIFAR-100 dataset. The teacher is ResNet32×4, and the student is ResNet8×4. Larger circle denotes more extra parameters.