CVPR JUNE 3-7, 2026 DENVER COLORADO

ADD Agency for Defense Development

Distilling Balanced Knowledge from a Biased Teacher

Seonghak Kim

Defense AI R&D Institute

Agency for Defense Development

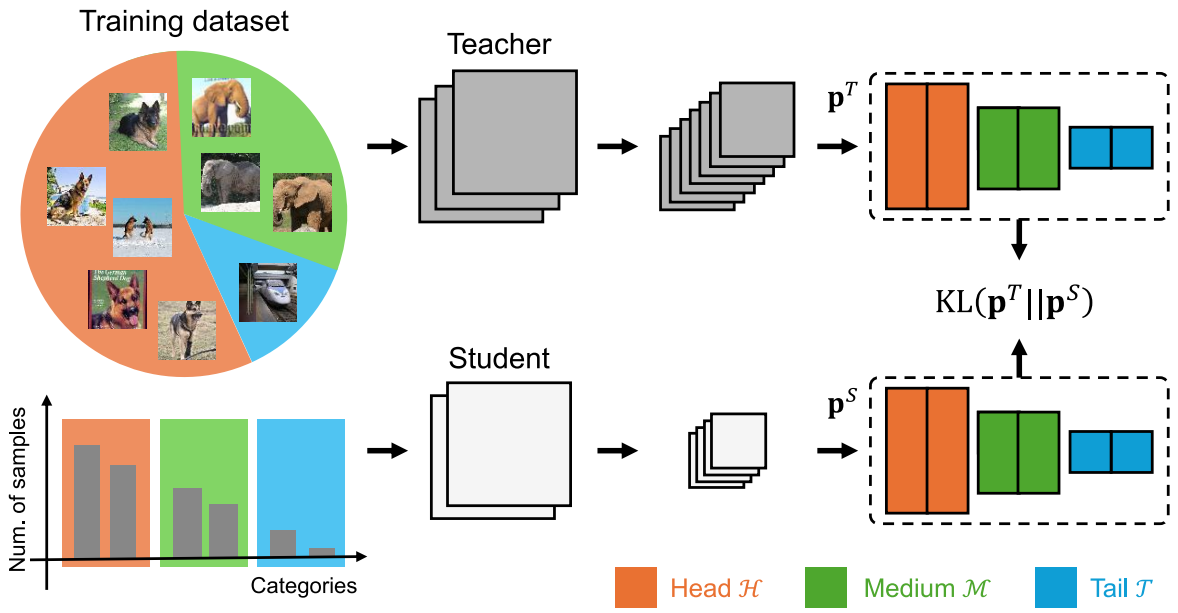
Motivation

- **Knowledge distillation (KD)**
  - Model compression: large teacher (T) → compact student (S)
  - Conventional KD assumes **balanced datasets**.

Can a teacher trained on imbalanced data still offer trustworthy supervision?

- **Long-tailed dataset**
  - Common in real-world data
  - Head class biased teacher
  - Poor performance on tail classes (∵ insufficient exposure)

➔ ***Failure of standard KD***



Method

- **Notation and definition**
  - For a classification task with C classes, predictive probability vector, $\mathbf{p} = [p_1, p_2, \dots, p_C] \in \mathbb{R}^C$

$$p_i = \sigma(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)}$$

  - Under long-tailed distributions, $C \to \mathcal{G} \in \{\mathcal{H}, \mathcal{M}, \mathcal{T}\}$

$$\begin{aligned} \mathrm{KD} &= \mathrm{KL}(\mathbf{p}^T \parallel \mathbf{p}^S) \\ &= \sum_{\mathcal{G}} \sum_{i \in \mathcal{G}} p_i^T \log\left(\frac{p_i^T}{p_i^S}\right) \end{aligned}$$

  - **Cross-group probability,** $\mathbf{p}_{\mathcal{G}} = [p_{\mathcal{H}}, p_{\mathcal{M}}, p_{\mathcal{T}}] \in \mathbb{R}^3$

$$p_{\mathcal{G}} = \frac{\sum_{i \in \mathcal{G}} \exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)}$$

  - **Within-group probability,** $\tilde{\mathbf{p}}_{\mathcal{G}} = \left[\tilde{p}_{\mathcal{G}_1}, \tilde{p}_{\mathcal{G}_2}, \dots, \tilde{p}_{\mathcal{G}_i}\right]_{i \in \mathcal{G}} \in \mathbb{R}^{|\mathcal{G}|}$

$$\tilde{p}_{\mathcal{G}_i} = \frac{\exp(z_i)}{\sum_{j \in \mathcal{G}} \exp(z_j)}$$

Method

ADD

- **Revisiting KL divergence**
  - Using cross- and within-group probability,

$$p_{\mathcal{G}} = \frac{\sum_{i \in \mathcal{G}} \exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)}, \tilde{p}_{\mathcal{G}_i} = \frac{\exp(z_i)}{\sum_{j \in \mathcal{G}} \exp(z_j)}$$

$$\begin{aligned}
\mathrm{KD} &= \mathrm{KL}(\mathbf{p}^T \parallel \mathbf{p}^S) \\
&= \sum_{\mathcal{G}} \sum_{i \in \mathcal{G}} p_i^T \log\left(\frac{p_i^T}{p_i^S}\right) \\
&= \sum_{\mathcal{G}} \sum_{i \in \mathcal{G}} p_i^T \log\left(\frac{p_{\mathcal{G}}^T}{p_{\mathcal{G}}^S}\right) + \sum_{\mathcal{G}} \sum_{i \in \mathcal{G}} p_i^T \log\left(\frac{\tilde{p}_{\mathcal{G}_i}^T}{\tilde{p}_{\mathcal{G}_i}^S}\right) \\
&= \sum_{\mathcal{G}} p_{\mathcal{G}}^T \log\left(\frac{p_{\mathcal{G}}^T}{p_{\mathcal{G}}^S}\right) + \sum_{\mathcal{G}} p_{\mathcal{G}}^T \sum_{i \in \mathcal{G}} \tilde{p}_{\mathcal{G}_i}^T \log\left(\frac{\tilde{p}_{\mathcal{G}_i}^T}{\tilde{p}_{\mathcal{G}_i}^S}\right) \\
&= \mathrm{KL}\left(\mathbf{p}_{\mathcal{G}}^T \parallel \mathbf{p}_{\mathcal{G}}^S\right) + \sum_{\mathcal{G}} p_{\mathcal{G}}^T \cdot \mathrm{KL}\left(\tilde{\mathbf{p}}_{\mathcal{G}}^T \parallel \tilde{\mathbf{p}}_{\mathcal{G}}^S\right)
\end{aligned}$$

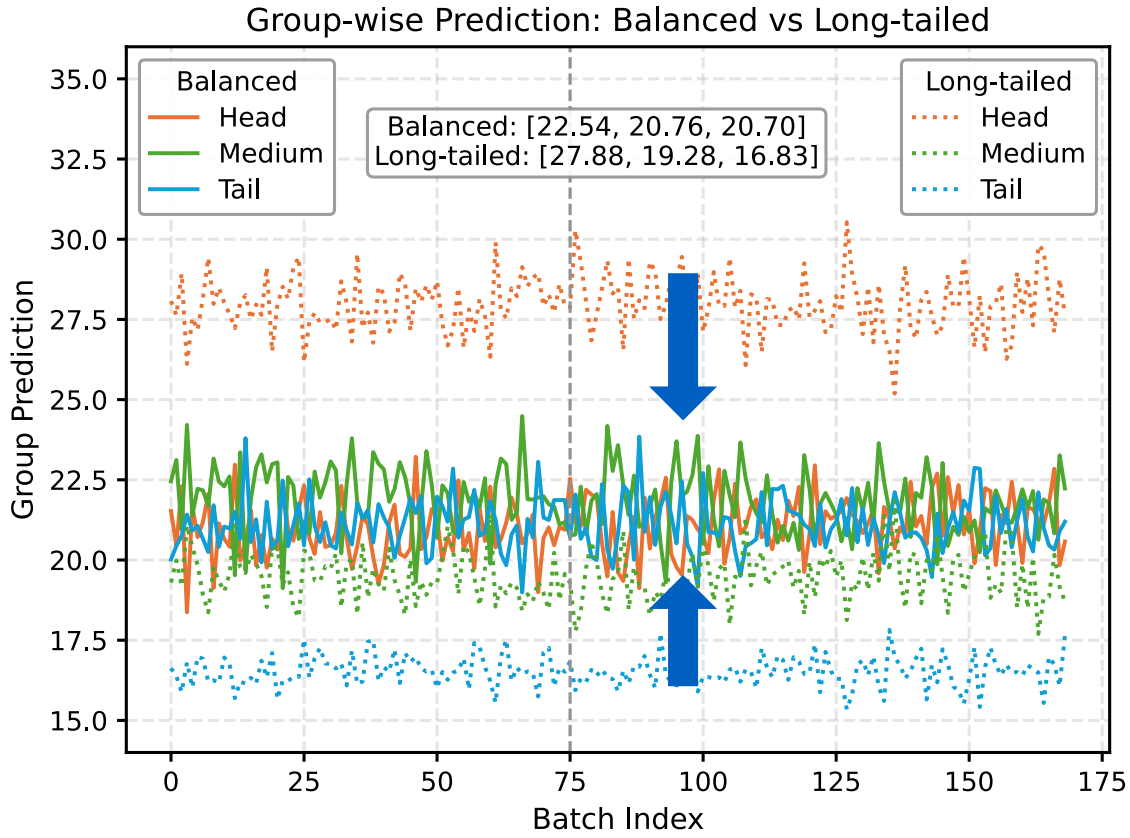
Cross-group loss ***Weighted sum of within-group losses***

Method

- **Rebalanced cross-group loss**
  - Teachers' cross-group predictions: **uniform** on balanced data *vs.* **head-biased** under long-tailed data
  - **Rebalancing before distillation** w/ **scaling factors** for each group

$$s_{\mathcal{H}} = \frac{p^B_{\text{avg}}}{p^B_{\mathcal{H}}}, s_{\mathcal{M}} = \frac{p^B_{\text{avg}}}{p^B_{\mathcal{M}}}, s_{\mathcal{T}} = \frac{p^B_{\text{avg}}}{p^B_{\mathcal{T}}}$$

$$\mathbf{p}_{\text{batch}} = \left[p^B_{\mathcal{H}}, p^B_{\mathcal{M}}, p^B_{\mathcal{T}}\right]$$
$$p^B_{\text{avg}} = \text{Mean}\left(p^B_{\mathcal{H}}, p^B_{\mathcal{M}}, p^B_{\mathcal{T}}\right)$$



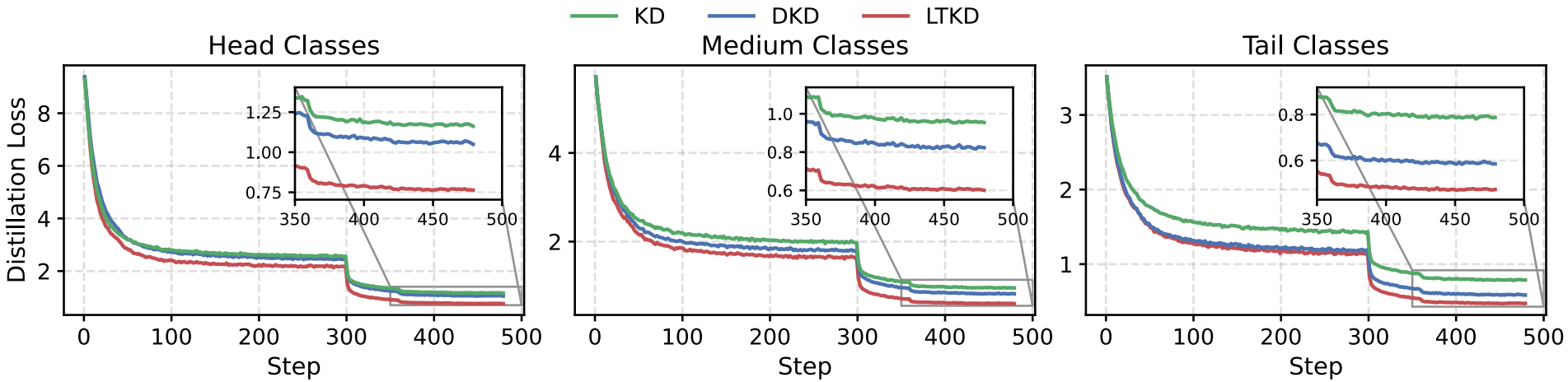
- After normalization, $\hat{\mathbf{p}}^T_{\mathcal{G}} = \left[\frac{s_{\mathcal{H}} p^T_{\mathcal{H}}}{\sum_{\mathcal{G}} s_{\mathcal{G}} p^T_{\mathcal{G}}}, \frac{s_{\mathcal{M}} p^T_{\mathcal{M}}}{\sum_{\mathcal{G}} s_{\mathcal{G}} p^T_{\mathcal{G}}}, \frac{s_{\mathcal{T}} p^T_{\mathcal{T}}}{\sum_{\mathcal{G}} s_{\mathcal{G}} p^T_{\mathcal{G}}}\right]$

Method

- **Reweighted within-group loss**

$$\sum_{\mathcal{G}} p_{\mathcal{G}}^{T} \cdot \mathrm{KL}\left(\tilde{\mathbf{p}}_{\mathcal{G}}^{T} \,\|\, \tilde{\mathbf{p}}_{\mathcal{G}}^{S}\right) = p_{\mathcal{H}}^{T}\mathrm{KL}\left(\tilde{\mathbf{p}}_{\mathcal{H}}^{T} \,\|\, \tilde{\mathbf{p}}_{\mathcal{H}}^{S}\right) + p_{\mathcal{M}}^{T}\mathrm{KL}\left(\tilde{\mathbf{p}}_{\mathcal{M}}^{T} \,\|\, \tilde{\mathbf{p}}_{\mathcal{M}}^{S}\right) + p_{\mathcal{T}}^{T}\mathrm{KL}\left(\tilde{\mathbf{p}}_{\mathcal{T}}^{T} \,\|\, \tilde{\mathbf{p}}_{\mathcal{T}}^{S}\right)$$

  - Weaker supervision for tail classes $\left(p_{\mathcal{H}}^{T} > p_{\mathcal{M}}^{T} > p_{\mathcal{T}}^{T}\right)$ → suboptimal convergence

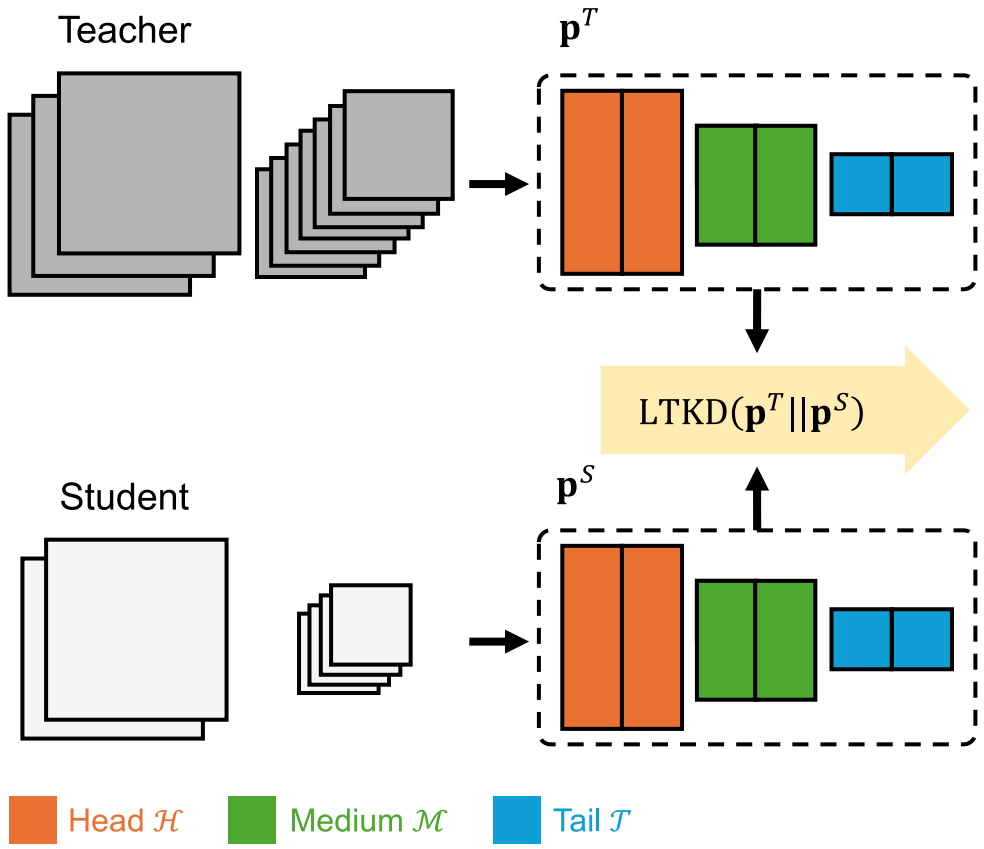


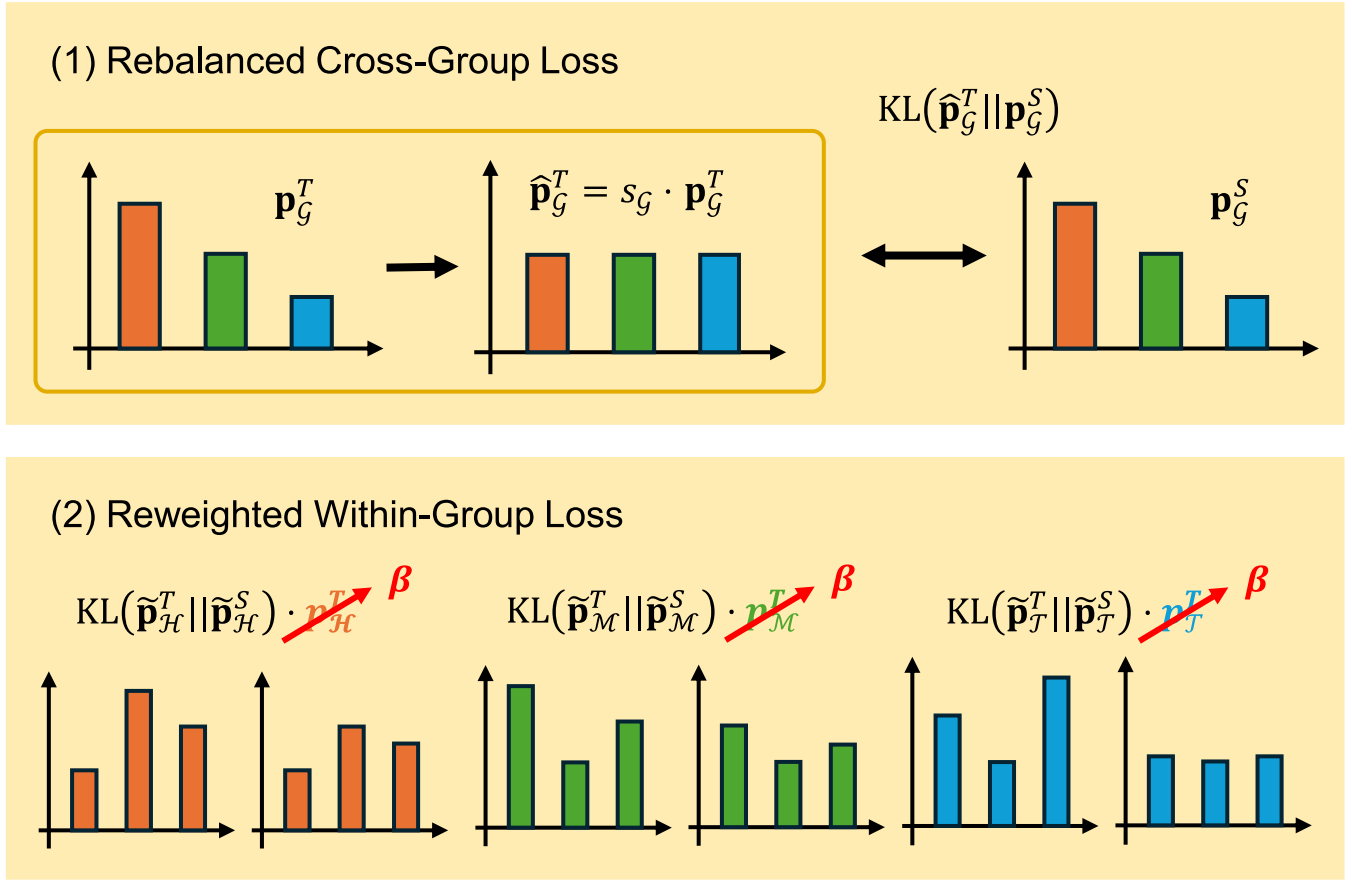
  - **Equal importance** to all groups
    by replacing teacher-derived weights $p_{\mathcal{G}}^{T}$ → **uniform constant $\boldsymbol{\beta}$**

Method

- **Long-tailed knowledge distillation (LTKD)**

$$\alpha \cdot \mathrm{KL}\left(\widehat{\mathbf{p}}_{\mathcal{G}}^{T} \parallel \mathbf{p}_{\mathcal{G}}^{S}\right) + \beta \cdot \sum_{\mathcal{G}} \mathrm{KL}\left(\widetilde{\mathbf{p}}_{\mathcal{G}}^{T} \parallel \widetilde{\mathbf{p}}_{\mathcal{G}}^{S}\right)$$





Main results

- **CIFAR-100-LT**

T-S Pairs	ResNet32×4 – ResNet8×4						VGG13 – VGG8					
γ	10		20		100		10		20		100	
Group	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All
Teacher	50.72	64.95	39.19	58.82	15.28	45.35	45.67	60.77	36.43	55.10	14.01	43.11
Student	47.32	60.59	36.99	55.44	13.38	42.48	43.67	57.43	33.89	52.29	13.13	40.70
DKD [52]	49.86	64.55	37.87	58.78	13.25	46.11	48.00	61.84	37.65	56.68	14.42	44.22
ReviewKD [6]	52.08	64.71	40.12	59.17	15.09	45.91	47.75	61.43	37.69	56.51	14.76	44.19
DIST [17]	50.28	63.74	38.69	58.28	13.86	45.21	45.57	60.53	34.36	54.68	12.46	42.12
CAT-KD [12]	49.83	64.74	37.67	58.73	12.83	45.33	48.53	62.01	37.95	56.78	14.22	44.33
LTKD	**58.66**	**66.76**	**49.70**	**62.54**	**27.21**	**51.08**	**53.95**	**63.04**	**45.77**	**58.86**	**23.30**	**47.66**
Δ	**+6.58**	**+2.02**	**+9.58**	**+3.37**	**+12.12**	**+4.97**	**+5.42**	**+1.03**	**+7.82**	**+2.08**	**+8.54**	**+3.33**

T-S Pairs	WRN–40–2 – ShuffleNetV1						ResNet50 – MobileNetV2					
γ	10		20		100		10		20		100	
Group	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All
Teacher	49.77	63.05	39.88	58.27	14.88	44.78	49.74	63.51	37.74	56.70	14.42	42.26
Student	40.04	54.06	29.91	48.22	10.74	36.21	30.47	44.58	22.32	39.25	7.04	27.56
DKD [52]	50.86	63.65	39.94	58.28	15.04	45.24	43.29	57.20	33.23	52.20	12.45	39.21
ReviewKD [6]	51.24	63.90	40.44	58.63	15.81	45.40	33.68	47.75	24.80	42.08	9.75	31.86
DIST [17]	48.40	62.47	37.48	56.92	12.23	41.95	37.86	52.36	27.11	46.50	9.81	34.96
CAT-KD [12]	51.02	63.68	40.23	58.26	14.68	44.84	43.18	57.23	33.17	51.90	11.61	38.45
LTKD	**57.40**	**65.42**	**48.42**	**60.94**	**23.99**	**48.60**	**48.43**	**57.79**	**40.82**	**53.70**	**21.04**	**42.45**
Δ	**+6.16**	**+1.52**	**+7.98**	**+2.31**	**+8.18**	**+3.20**	**+5.14**	**+0.56**	**+7.59**	**+1.50**	**+8.59**	**+3.24**

Main results

- **TinyImageNet-LT**

T-S Pairs	ResNet32×4 – ResNet8×4						VGG13 – VGG8					
γ	10		20		100		10		20		100	
Group	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All
Teacher	38.47	52.64	28.74	47.49	9.53	35.37	32.53	45.23	21.85	39.75	6.29	29.71
Student	29.72	44.60	21.46	40.25	4.73	30.62	31.12	43.76	22.19	39.00	6.88	29.96
KD [16]	27.34	45.38	18.11	41.35	3.38	31.42	31.92	47.16	20.60	41.49	3.99	30.95
DKD [52]	34.70	48.93	26.58	44.84	9.09	34.61	33.20	48.01	22.65	42.44	5.88	31.82
ReviewKD [6]	32.85	49.13	23.43	44.62	5.39	33.51	34.39	47.66	24.84	42.36	7.61	32.18
DIST [17]	34.81	50.14	25.71	45.52	7.30	33.98	33.48	47.22	23.19	41.46	5.98	31.01
LTKD	**40.66**	**51.33**	**31.33**	**47.05**	**10.48**	**36.21**	**38.90**	**49.43**	**29.30**	**44.22**	**9.73**	**33.78**
Δ	**+5.85**	**+1.19**	**+4.75**	**+1.53**	**+1.39**	**+1.60**	**+4.51**	**+1.42**	**+4.46**	**+1.78**	**+2.12**	**+1.60**

T-S Pairs	ResNet32×4 – ShuffleNetV1						VGG13 – MobileNetV2					
γ	10		20		100		10		20		100	
Group	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All
Teacher	38.47	52.64	28.74	47.49	9.53	35.37	32.53	45.23	21.85	39.75	6.29	29.71
Student	24.43	37.10	16.80	31.85	4.71	22.77	26.98	39.73	17.38	33.14	3.97	22.99
KD [16]	34.67	49.05	24.12	42.81	5.62	30.74	31.24	45.60	19.53	39.50	3.27	28.44
DKD [52]	36.83	50.22	26.64	44.38	8.34	33.23	33.07	46.79	22.06	40.87	5.70	29.97
ReviewKD [6]	36.30	49.27	27.09	44.72	8.49	33.12	32.37	44.99	22.77	39.36	7.02	29.20
DIST [17]	36.49	50.07	25.74	43.59	7.23	31.19	32.92	46.09	21.77	40.05	5.52	29.08
LTKD	**42.12**	**51.64**	**33.06**	**46.41**	**12.85**	**35.09**	**39.04**	**48.71**	**28.28**	**43.22**	**9.52**	**32.30**
Δ	**+5.29**	**+1.42**	**+5.97**	**+1.69**	**+4.36**	**+1.86**	**+5.97**	**+1.92**	**+5.51**	**+2.35**	**+2.50**	**+2.33**

Main results

- **ImageNet-LT**
  - **Consistent improvements** on the **large-scale, imbalanced datasets** as well as general benchmarks such as CIFAR-100-LT and TinyImageNet-LT

T-S Pairs	ResNet34 – ResNet18						ResNet50 – MobileNetV1					
γ	5		10		20		5		10		20	
Group	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All
Teacher	57.75	67.61	50.61	63.85	43.00	59.91	59.62	69.06	52.69	65.57	44.82	61.46
Student	54.48	64.73	47.45	61.18	39.98	57.25	55.42	65.46	49.49	62.55	41.68	58.94
KD [16]	56.14	66.27	49.35	63.03	41.72	59.15	56.58	66.51	50.42	63.70	42.45	59.91
DKD [52]	56.83	66.84	49.95	63.50	42.65	59.82	58.54	68.09	52.39	65.04	45.04	61.46
ReviewKD [6]	57.27	66.96	50.80	63.72	43.48	60.15	58.59	68.10	53.06	65.31	45.35	61.84
DIST [17]	56.79	66.66	50.28	63.56	42.78	59.94	57.29	66.94	50.89	64.09	43.70	60.71
CAT-KD [12]	55.92	66.14	49.58	63.20	42.20	59.67	57.08	66.96	51.00	64.07	43.55	60.54
LTKD	**58.33**	**67.23**	**52.55**	**64.29**	**45.88**	**60.80**	**60.17**	**68.48**	**54.40**	**65.52**	**48.55**	**62.22**
Δ	**+1.06**	**+0.27**	**+1.75**	**+0.57**	**+2.40**	**+0.65**	**+1.58**	**+0.38**	**+1.34**	**+0.21**	**+3.20**	**+0.38**

Ablation study

- **Cross-group loss:** Non-rebalanced $\left(\mathbf{p}_{\mathcal{G}}^{\mathcal{T}}\right)$ *vs.* rebalanced $\left(\widehat{\mathbf{p}}_{\mathcal{G}}^{\mathcal{T}}\right)$

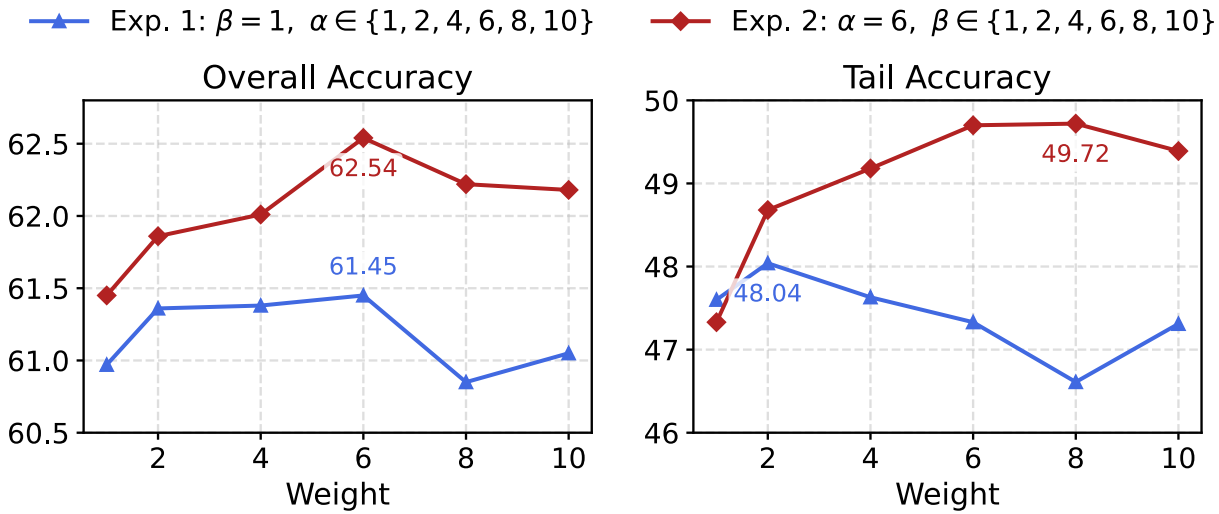
Models	Teacher	ResNet32×4				VGG13				WRN–40–2			
	Student	ResNet8×4		ShuffleNetV1		VGG8		MobileNetV2		WRN–40–1		ShuffleNetV1	
Loss	Cross / Within	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All
Biased	✓ / ✗	38.30	55.77	30.01	**48.82**	35.55	53.30	23.42	40.29	34.64	53.97	30.23	49.31
Ours	✓ / ✗	**40.51**	**56.55**	**30.68**	48.81	**37.26**	**53.70**	**23.97**	**40.30**	**35.23**	**54.35**	**32.44**	**49.91**
	△	**+2.21**	**+0.78**	**+0.67**	-0.01	**+1.71**	**+0.40**	**+0.55**	**+0.01**	**+0.59**	**+0.38**	**+2.21**	**+0.60**

- **Complementary cross- and within-group components**

Models	Teacher	ResNet32×4				VGG13				WRN–40–2			
	Student	ResNet8×4		ShuffleNetV1		VGG8		MobileNetV2		WRN–40–1		ShuffleNetV1	
Loss	Cross / Within	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All	$\mathcal{T}$	All
Baseline	✗ / ✗	36.81	57.41	36.13	55.03	37.29	56.13	29.05	47.69	35.63	56.71	38.30	56.61
Ours	✓ / ✗	40.51	56.55	30.68	48.81	37.26	53.70	23.97	40.30	35.23	54.35	32.44	49.91
	✗ / ✓	42.34	59.78	39.10	56.84	38.54	56.83	31.99	49.20	39.67	58.11	40.45	57.65
	✓ / ✓	**49.70**	**62.54**	**45.94**	**59.62**	**45.77**	**58.86**	**38.33**	**52.03**	**45.74**	**59.91**	**48.42**	**60.94**

Ablation study

- **Hyperparameters**
  - Consistently superior to ReviewKD (Overall: 59.17%, Tail: 40.12%) across all settings
    → **High robustness**



- **Number of groups**
  - $n(\mathcal{G})$ ↑: fine-grained bias correction → performance ↑

Continuous reweighting

$n(\mathcal{G})$	3	4	5	10	20	25	50	100
R32×4−R8×4	51.08	51.08	51.10	**51.14**	50.99	50.34	50.06	50.41
VGG13−VGG8	47.66	47.85	48.06	48.26	**48.69**	48.58	48.19	47.82
WRN402−SV1	48.60	48.98	49.03	49.27	49.54	**49.90**	49.64	47.95
R50−MV2	42.45	43.01	42.99	43.40	**43.43**	43.42	42.51	41.18

CVPR JUNE 3-7, 2026 DENVER COLORADO

ADD Agency for Defense Development

Thank you!

Seonghak Kim

Defense AI R&D Institute

Agency for Defense Development