Streamlined Knowledge Distillation

Hyeon-Jin Jeong[1], Han-Jin Lee[2], Seok-Hwan Choi[†]

CVPR JUNE 3-7, 2026 DENVER COLORADO

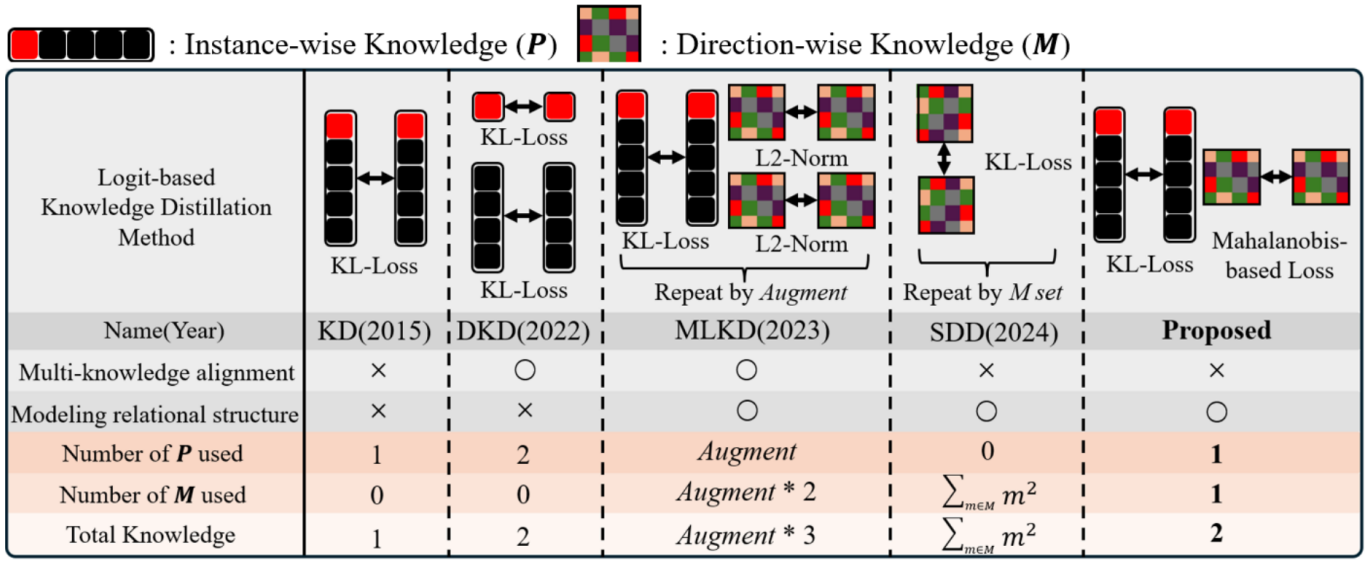
01. Background

- Knowledge Distillation (KD)
  - Feature-based KD
    - Aligns intermediate feature representations.
    - Pros : Rich knowledge transfer, high performance.
    - Cons : High computational overhead, requires extra projection modules.
  - Logit-based KD
    - Mimics output predictions directly.
    - Pros : Simple, scalable, easy to deploy.
    - Cons : Historically lower performance than feature-based methods.

The objective of recent logit-based KD has been to close performance gap with feature-based KD while keeping the scalability and simplicity of logit-level supervision.

02. Motivation

- Logit-based Approach Limitation
  - Recent logit-based KD often use multi-knowledge alignment and relational structure modeling.
  - But, there are three key limitations.
    - Redundant objectives) overlapping losses increases complexity.
    - Suboptimal transformations) output transformation can distort relational structure.
    - Ill-suited loss functions) L2-norm treats all relations equally and ignores variance.



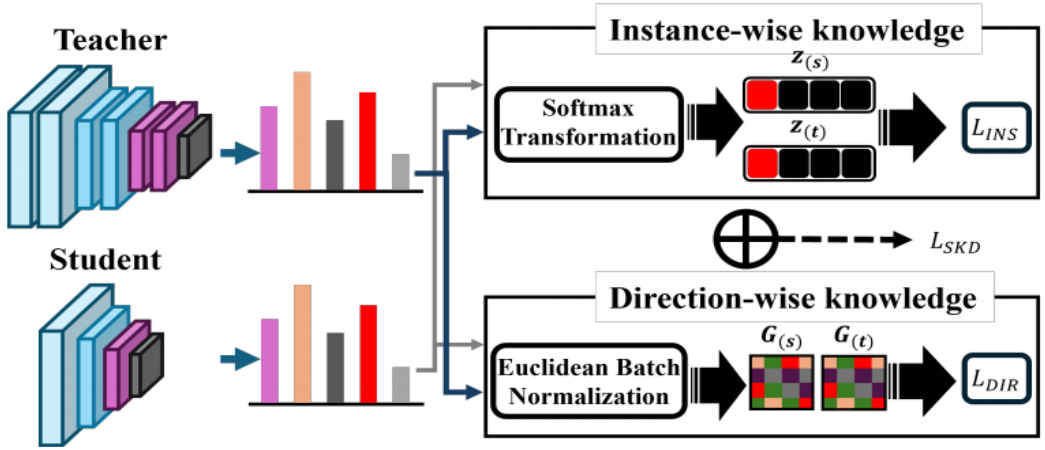
Logit-based Knowledge Distillation Method	KD(2015)	DKD(2022)	MLKD(2023)	SDD(2024)	**Proposed**
Name(Year)					
Multi-knowledge alignment	×	○	○	×	×
Modeling relational structure	×	×	○	○	○
Number of ***P*** used	1	2	*Augment*	0	**1**
Number of ***M*** used	0	0	*Augment* * 2	$\sum_{m \in M} m^2$	**1**
Total Knowledge	1	2	*Augment* * 3	$\sum_{m \in M} m^2$	**2**

Augment - Prediction Augmentation : ex) {3, 4, 5} , M - Pooling Scale : ex) {1, 2, 4}

➔ Motivated by these limitations, we streamlined the distillation process by transferring only two essential forms of knowledge.

03. Method

- Streamlined Knowledge Distillation (1 / 5) - Overview
  - Knowledge 1: Instance-wise Semantics
    - Inter-class similarity and fine-grained individual sample outputs.
  - Knowledge 2: Direction-wise Relational Structure
    - Pairwise directional relationships and output correlations across samples.



03. Method

- Streamlined Knowledge Distillation (2 / 5) - Instance-wise Semantic
  - Instance-wise Knowledge
    - To capture individual sample distributions and inter-class similarity, SKD utilizes the original KD.
    - We supervise the student using single soft target from teacher via KL divergence.

$$L_{INS} = KL\left(softmax\left(\frac{z_t}{\tau}\right), softmax\left(\frac{z_s}{\tau}\right)\right)$$

03. Method

- Streamlined Knowledge Distillation (3 / 5) - Direction-wise Relational Structure
  - Direction-wise Knowledge
    - To capture pairwise relationships, we use Gramian matrix on output logits.
    - Also, normalize the relational space via Mahalanobis distance-based loss, L_{DIR}.

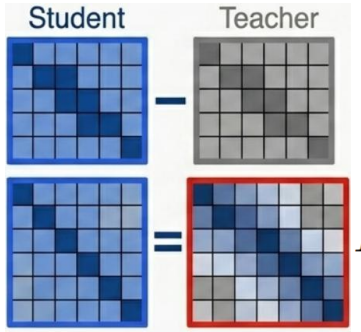
1. Isolate Direction (EBN)

- Euclidean Batch Normalization (EBN) transforms each logit into a unit vector.
- It removes scale, preserving pure directional structure.

$$\hat{z}_i = \frac{z_i}{\|z_i\|_2} = \frac{z_i}{\sqrt{\sum_{j=1}^{C} z_{ij}^2}}$$

2. Construct Gramian Matrix

- Build Gramian matrix from normalized logits.
- Then, get Gramian differences between teacher and student.



$$G = \hat{z} \cdot \hat{z}^\top$$

$$D = G_{(s)} - G_{(t)}$$

3. Measure Variance (Covariance)

- Estimate empirical batch covariance.
- Mahalanobis distance weight the Gramian difference by the inverse covariance.

$$\Sigma = \mathrm{Cov}\left(\{D_{i,:}\}_{i=1}^{B}\right)$$

$$L_{\mathrm{DIR}} = \frac{1}{B}\sum_{i=1}^{B}\sqrt{D_{i,:}^\top \Sigma^{-1} D_{i,:}}$$

03. Method

- Streamlined Knowledge Distillation (4 / 5) - Direction-wise Relational Structure
  - Computational bottlenecks of the inversion process of covariance matrix
    - Numerical Instability : covariance matrix is not guaranteed to be positive definite.
    - Computational Cost : computing the inverse matrix incurs a cubic complexity of $O(d^3)$.

➔ To overcome these challenges, we apply **stabilization tricks** on L_{DIR}.

  - Stabilizing the inversion process of covariance matrix
    - Tikhonov Regularization : add a small identity matrix term to ensure positive definiteness.

$$\Sigma' = \Sigma + \lambda I$$

    - Cholesky Decomposition : Factorize the regularized matrix into a unique lower triangular matrix.

$$\Sigma' = LL^\top, \quad L = \text{Cholesky}(\Sigma')$$

$$L_{\text{DIR}} = \frac{1}{B} \sum_{i=1}^{B} \sqrt{D_{i,:}^\top \, \Sigma'^{-1} \, D_{i,:}}$$

03. Method

- Streamlined Knowledge Distillation (5 / 5) - Direction-wise Relational Structure
  - Proposition) L_{DIR} is equivalent to Whitened L2-norm
    - $\sum'$ is represented by the Cholesky decomposition $LL^\top$.
    - Since, $(LL^\top)^{-1} = (L^\top)^{-1}L^{-1}$, the term inside the quare root can be rewritten as a standard L2-norm.

$$L_{\text{DIR}} = \frac{1}{B}\sum_{i=1}^{B}\sqrt{D_{i,:}^\top \Sigma'^{-1} D_{i,:}}$$

$$L_{\text{DIR}} = \frac{1}{B}\sum_{i=1}^{B}\sqrt{D_{i,:}^\top (LL^\top)^{-1} D_{i,:}}$$

$$\frac{1}{B}\sum_{i=1}^{B}\sqrt{(L^{-1}D_{i,:})^\top (L^{-1}D_{i,:})} = \frac{1}{B}\sum_{i=1}^{B}\sqrt{||L^{-1}D_{i,:}||_2^2}$$

$$\therefore L_{\text{DIR}} = \frac{1}{B}\sum_{i=1}^{B}||L^{-1}D_{i,:}||_2$$

➔ This means that L_{DIR} is still simple like L2-norm, but it is aware of variance and correlation.

04. Experiments

- Quantitative Results – Main Results under CIFAR-100/ImageNet

Method	Teacher	ResNet56	ResNet110	ResNet32x4	WRN-40-2	WRN-40-2	VGG13
		72.40	74.31	79.42	75.59	75.59	74.64
	Student	ResNet20	ResNet32	ResNet8x4	WRN-16-2	WRN-40-1	VGG8
		65.79	67.92	72.48	71.12	69.54	68.19
Feature	FitNet [29]	69.64	72.35	75.26	74.46	73.28	70.09
	RKD [25]	70.51	73.31	74.55	74.08	73.52	71.61
	CRD [31]	67.28	71.97	74.16	74.22	72.44	72.74
	OFD [14]	68.77	71.40	73.88	72.31	73.26	74.21
	ReviewKD [5]	69.54	71.12	75.71	73.92	74.50	72.29
Logit	KD [16]	71.74	74.01	74.75	76.04	74.52	74.08
	CLKD [41]	65.74	69.81	70.19	72.89	71.89	72.46
	DKD [42]	71.17	74.12	76.51	76.41	75.33	74.41
	MLKD [17]	72.21	74.24	75.59	**76.83**	74.78	74.25
	SDD [36]	69.42	72.78	74.40	75.01	72.53	72.29
	RLD [30]	72.00	74.02	76.64	76.06	74.88	74.93
	SKD (Ours)	**72.50**	**74.84**	**78.33**	76.60	**76.04**	**75.75**

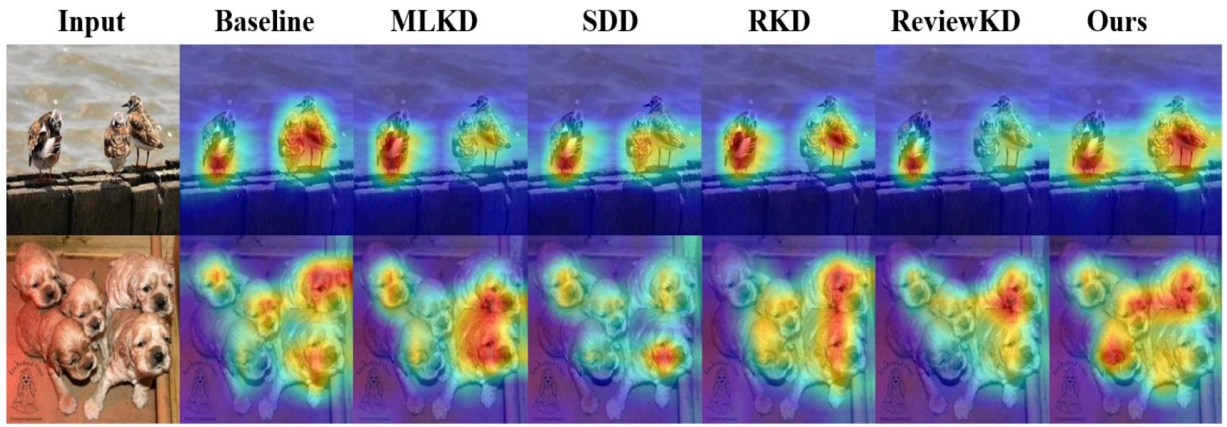
Method	Teacher	ResNet32x4	WRN-40-2	VGG13	ResNet50	ResNet32x4
		79.42	75.59	74.64	79.33	79.42
	Student	ShuffleNetV1	ShuffleNetV1	MobileNetV1	MobileNetV1	ShuffleNetV2
		69.43	69.43	60.09	60.09	73.5
Feature	FitNet [29]	74.09	73.49	64.98	64.06	75.42
	RKD [25]	75.76	75.88	63.25	63.17	77.13
	CRD [31]	75.5	76.15	66.55	66.28	74.66
	OFD [14]	77.87	77.39	64.82	67.65	77.74
	ReviewKD [5]	77.39	**77.68**	65.26	62.52	77.99
Logit	KD [16]	76.64	77.23	67.57	68.44	77.78
	CLKD [41]	73.86	74.27	62.19	61.53	75.56
	DKD [42]	76.75	75.94	67.58	67.51	78.43
	MLKD [17]	77.57	77.17	68.56	68.17	78.86
	SDD [36]	75.4	74.37	67.93	67.81	77.87
	RLD [30]	76.29	75.12	62.23	64.81	77.56
	SKD (Ours)	**77.91**	77.53	**68.60**	**68.48**	**79.02**

Teacher	Student	KD [16]	DKD [42]	MLKD [17]	SDD [36]	RLD [30]	AT [40]	ReviewKD [5]	RKD [25]	SKD (Ours)
R34 (73.30)	R18 (69.76)	69.11±0.21	70.88±0.16	70.97±0.14	70.30±0.19	70.52±0.17	70.54±0.15	70.51±0.18	69.94±0.20	**71.13±0.12**
R50 (76.14)	MV1 (66.51)	67.81±0.24	70.58±0.18	71.08±0.15	70.84±0.17	71.07±0.14	70.90±0.16	67.77±0.23	71.22±0.13	**71.53±0.11**

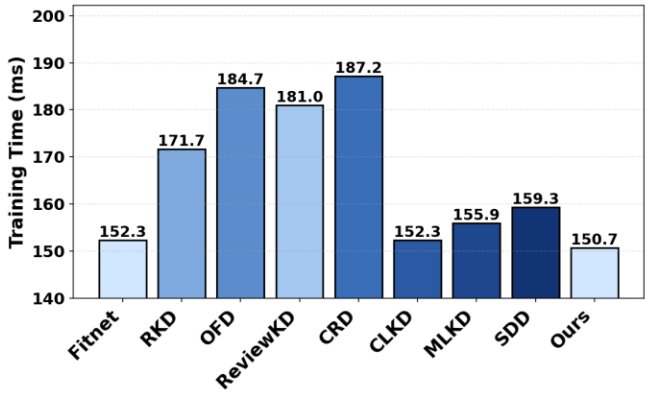
- Across CIFAR-100 and ImageNet, SKD consistently outperforms prior logit-based KD methods, showing that a streamlined two-knowledge design can achieve strong accuracy and scalability.

04. Experiments

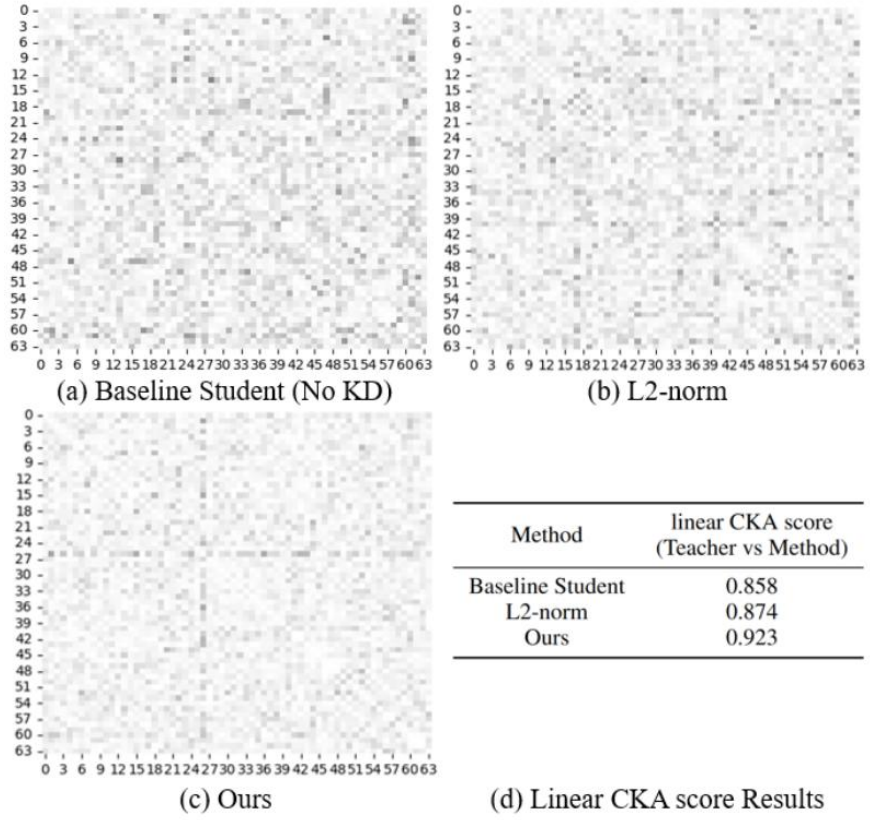
- Qualitative Results – Analysis of SKD



Grad-CAM++ Heatmap Visualization



Per-batch Training Time (ms)



Method	linear CKA score (Teacher vs Method)
Baseline Student	0.858
L2-norm	0.874
Ours	0.923

(d) Linear CKA score Results

Teacher-Student Gramian Matrix Difference Heatmap & CKA Scores

Hyeon-Jin Jeong

- Email : wjdehdtod@yonsei.ac.kr
- Github Link :