Streamlined Knowledge Distillation

Hyeon-Jin Jeong[1] Han-Jin Lee[2] Seok-Hwan Choi[†]
[1,2]Department of Computer Science, Yonsei University
[†]Department of Software, Yonsei University
{wjdehdtod, han-.-jin, sh.choi}@yonsei.ac.kr

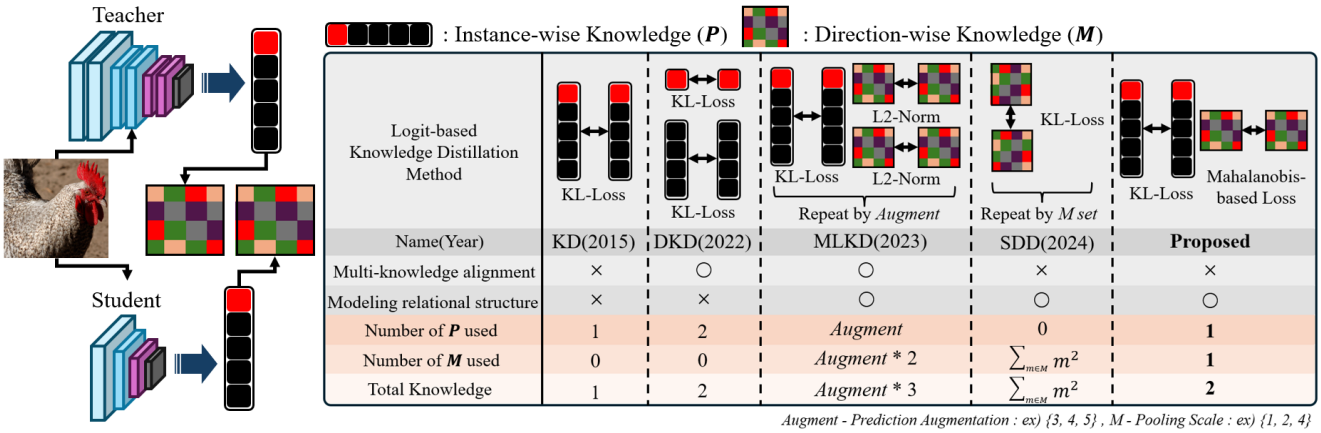
Abstract

Logit-based Knowledge Distillation (KD) has emerged as a lightweight alternative to feature-based KD. Recent logit-based methods often rely on multi-knowledge alignment and relational modeling. These methods are often inefficient due to redundant objectives, suboptimal transformations, and poorly designed loss functions. Motivated by these issues, we propose Streamlined Knowledge Distillation (SKD), a simple yet effective logit-based method that transfers only two essential forms of knowledge without requiring additional alignment or relational modeling. Specifically, SKD transfers instance-wise knowledge via Kullback-Leibler divergence and direction-wise knowledge by aligning the Gramian matrix of normalized logits. For the latter, we introduce a Mahalanobis distance-based direction-wise loss stabilized through Tikhonov regularization and Cholesky decomposition. This direction-wise loss accounts for variance and correlation in the output space and, as we formally show, is equivalent to the L2-norm in a covariance-whitened space. Extensive experiments demonstrate that SKD consistently outperforms existing logit-based methods and even surpasses feature-based methods, despite its simpler design. Code is available at https://github.com/HyunJunSik/StreamLined.

1. Introduction

The capacity of deep learning models has substantially increased, contributing to excellent performance in computer vision tasks such as image classification, object detection, and segmentation [39]. However, these large models face deployment challenges in resource-constrained environments such as embedded systems and mobile devices [12]. Recent studies have explored methods to address such constraints in real-world deployment [20, 34]. As part of these efforts, Knowledge Distillation (KD) [16] has been widely adopted as a practical solution. KD transfers knowledge from a large model (teacher) to a small model (student) by distilling knowledge. This allows the student model to achieve high performance while being deployable in resource-constrained environments. KD methods are commonly categorized as feature-based and logit-based [9]. Feature-based methods focus on aligning intermediate feature representations between the teacher and the student. This distillation process encourages the student to acquire richer knowledge by mimicking the intermediate feature representations of the teacher. However, feature-based methods often face practical limitations in real-world applications, including high computational overheads and potential security vulnerabilities. Specifically, when the teacher and student are heterogeneous, extra projection modules are required to match their feature representations, which increases training time and cost [3, 26]. Also, since feature-based methods use intermediate features, they are more exposed to security risks. For example, if a backdoor is embedded in the teacher model, it may be transferred to the student through feature-based methods [4]. Due to these problems, logit-based KD methods have started to gain more attention.

Logit-based KD methods guide the student to mimic the output predictions of the teacher. They provide an easy and effective distillation process and do not require access to the teacher's intermediate features, making them easier to use. However, since their performance is usually lower than feature-based KD methods, recent studies have tried to close this gap through multi-knowledge alignment or relational structure modeling [16, 17, 36, 42]. In Figure 1, we show the evolution of logit-based KD methods from simple to complex design. Since KD [16] was first proposed, DKD [42] has introduced multi-knowledge alignment by decoupling instance-wise knowledge. Building on this idea, MLKD [17] further augmented both instance-wise and direction-wise knowledge to enable multi-knowledge alignment and capture relational structure. More recently, SDD [36] proposed a pooling-based strategy to construct direction-wise knowledge from output logits at multiple scales. Collectively, these recent methods achieve multi-knowledge alignment and relational structure modeling by employing multiple forms of instance-wise knowledge ($\boldsymbol{P}$



Logit-based Knowledge Distillation Method	KL-Loss	KL-Loss	KL-Loss, L2-Norm	KL-Loss	KL-Loss, Mahalanobis-based Loss
Name(Year)	KD(2015)	DKD(2022)	MLKD(2023)	SDD(2024)	**Proposed**
Multi-knowledge alignment	×	○	○	×	×
Modeling relational structure	×	×	○	○	○
Number of $\boldsymbol{P}$ used	1	2	*Augment*	0	**1**
Number of $\boldsymbol{M}$ 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**

Figure 1. Evolution of logit-based distillation methods. Recent works introduce multi-knowledge alignment, leading to increased distillation complexity.

in Figure 1) and direction-wise knowledge ($\boldsymbol{M}$ in Figure 1). Although these efforts help narrow the performance gap with feature-based methods, three key limitations remain. First, multi-knowledge alignment strategies often lead to redundant and overlapping objectives, which can hinder training efficiency and introduce unnecessary complexity. Second, the quality of the modeled relational structure is frequently degraded by suboptimal transformations in output space. Third, ill-suited loss functions in the output space (e.g., L2-norm) fail to preserve relational structures, since they ignore uneven variance across directional relationships [14, 37].

Motivated by these observations, we propose Streamlined Knowledge Distillation (SKD), a simple yet effective logit-based KD method. Our method streamlines the distillation process by focusing solely on two essential forms of knowledge: instance-wise semantics and direction-wise relational structures. As shown in the rightmost part of Figure 1, our method avoids using multi-knowledge alignment and instead adopts a single instance-wise knowledge based on the Kullback-Leibler divergence loss [16]. SKD further transfers direction-wise knowledge by aligning the Gramian matrix of normalized output logits, which encodes pairwise directional relationships among samples. We then introduce a novel direction-wise loss function based on Mahalanobis distance, which normalizes uneven variances across relational structures, and stabilize this loss with Tikhonov regularization and Cholesky decomposition. Importantly, we provide a formal mathematical proof showing that the proposed direction-wise loss is exactly equivalent to the L2-norm in a covariance-whitened space, establishing it as a principled whitening formulation. Ultimately, our approach enables a simplified yet principled design of logit-based KD by transferring two complementary forms of knowledge in a unified framework. The main contributions of this paper are summarized as follows:

- We propose a simple logit-based distillation framework that transfers only two essential forms of knowledge–instance-wise and direction-wise–thereby removing the need for extra alignment or relational structure modeling.
- To realize direction-wise knowledge transfer, we build upon direction-aware representations and design a variance-aware matching loss that encourages the student to preserve pairwise directional relationships in the teacher's output space.
- We conduct extensive experiments on benchmarks under both homogeneous and heterogeneous architectures. From the results, SKD consistently outperforms state-of-the-art methods–even surpassing the teacher–and achieves strong scalability and training efficiency.

2. Related Works

Feature-based KD was first introduced by FitNet [29], which transfers knowledge from intermediate feature layers. Subsequent works have explored richer representations by using end-of-group features [1, 5] or transforming feature maps through attention mechanisms [33] and contrastive learning [31]. Other works focusing on the distance function perspective have attempted to overcome the limitations of the L2-norm by designing more structured alignment objectives [15, 25]. Despite their success, these methods often suffer from high computational overhead and require projection modules between teacher and student, which limits their practicality.

To address these issues, a logit-based KD method was first proposed by Hinton et al. [16]. By operating solely in the output space, this approach provides greater scalability and simplicity than feature-based KD methods. Since then, several methods have been proposed to enhance its performance, either by decoupling the output logits [30, 42]

or by adopting a collaborative training strategy [11]. More recent approaches further enhance distillation quality by introducing multi-knowledge alignment [36, 41] to transfer fine-grained semantics or by modeling directional relationships among samples to preserve relational structure [17]. Although these advanced methods improve distillation performance, they often require additional objectives and complex modeling strategies, which increase optimization overhead and may overload the student with excessive knowledge. To mitigate such overhead, we streamline the distillation process by transferring only two complementary forms of knowledge directly from the output space, which provides rich yet efficient supervision for the student.

3. Method

3.1. Preliminaries

We consider a conventional KD framework, where a pretrained teacher model T guides the training of a student model S on a supervised classification task. Let $\mathbf{x} \in \mathbb{R}^d$ be an input sample with label $y \in \{1, \dots, C\}$, and let $z_t = T(\mathbf{x})$, $z_s = S(\mathbf{x})$ denote the logits produced by the teacher and student, respectively. Here, $z_t, z_s \in \mathbb{R}^{B\times C}$ represent the output logits over C classes for a batch of B samples. The conventional KD aims to align the output distributions of the individual samples between T and S via soft targets. In this work, we focus on logit-based KD, a specific form of the conventional KD that transfers knowledge solely through the teacher's output logits. We improve logit-based KD by distilling two complementary forms of knowledge: instance-wise semantics and direction-wise relational structure. Detailed explanations are provided in Section 3.2 and Section 3.3, respectively.

3.2. Instance-wise Knowledge

Different from recent logit-based KD methods that introduce multiple forms of instance-wise knowledge, such as using two separate $\boldsymbol{P}$ distributions [42] or augmented logits [17], we adopt the original formulation of KD [16], which supervises the student using a single soft target from the teacher. Specifically, the student is trained to mimic the teacher's softened class probabilities by minimizing the Kullback-Leibler (KL) divergence between their logits:

$$L_{\text{INS}} = \text{KL}\left(\text{softmax}(z_{(t)}/\tau), \text{softmax}(z_{(s)}/\tau)\right), \quad (1)$$

where τ is the temperature factor that softens the output distributions. This loss encourages the student to align with the teacher's instance-wise output distribution, effectively capturing inter-class similarity information [16]. However, such a distillation process focuses only on individual sample outputs and ignores structural relationships between samples [27, 33]. To address this, we complement instance-wise knowledge with direction-wise knowledge.

3.3. Direction-wise Knowledge

To capture pairwise directional relationships in the output space, we propose a novel direction-wise knowledge based on Gramian matrix. The Gramian matrix, originally introduced to represent feature correlations between channels in convolutional neural networks [8], effectively captures directional relationships among representations. Given this advantage, it has been widely used in feature-based KD methods to model intermediate feature representations with L2-norm [6, 33, 38, 43].

Motivated by these strengths, a recent logit-based KD method [17] has applied the Gramian matrix to the output space, demonstrating its applicability to the teacher's output space. However, direct application to the output space poses challenges due to scale instability and ill-suited L2-norm objective. Unlike intermediate features, which are routinely stabilized by normalization layers (e.g., BN), the output space often shows large-scale variations across samples. Also, it may introduce nonlinear distortions in its directional structure when common transformations such as softmax or temperature scaling are applied [10]. In the case of the L2-norm, it treats all relationships in the Gramian matrix equally [13]. However, this uniform weighting tends to penalize unstable directions with a large variance more strongly, which weakens the preservation of meaningful directional relationships [24].

To address these issues, we aim to preserve the underlying directional relationships through a suitable transformation and a carefully designed loss function. Specifically, our method (i) generates a Gramian matrix by applying Euclidean Batch Normalization (EBN) to the output logits, thereby retaining pure directional information, and (ii) employs a Mahalanobis distance-based loss function to preserve meaningful directional relationships.

3.3.1. Generate Gramian Matrix via Euclidean Batch Normalization

EBN transforms each logit vector $z_i \in \mathbb{R}^C$ into a unit vector, preserving only directional structure and removing scale. This enables the Gramian matrix to capture pure pairwise directional relationships such as cosine similarities. Let the logits be $z \in \mathbb{R}^{B\times C}$, where each row represents a sample. We normalize each row as follows:

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

The normalized logit $\hat{z} \in \mathbb{R}^{B\times C}$ contains unit-norm row vectors for each sample. We then compute the Gramian matrix $G \in \mathbb{R}^{B\times B}$ as follows:

$$G = \hat{z} \cdot \hat{z}^\top \quad G_{ij} = \langle \hat{z}_i, \hat{z}_j \rangle = \sum_{k=1}^{C} \hat{z}_{ik}\hat{z}_{jk}, \quad (3)$$

which encodes pairwise directional relationships–captured as cosine similarities–between normalized logits, with each entry $G_{ij} = \langle \hat{z}_i, \hat{z}_j \rangle$ reflecting the similarity between samples i and j.

3.3.2. Design Direction-wise Loss

Although EBN preserves pairwise directional relationships, the strength of these relations varies across sample pairs, and some parts are much more unstable than others. Under the standard L2-norm, all relations are treated equally, allowing high-variance ones to dominate training and distort the overall structure. To address this, we design a direction-wise loss based on the Mahalanobis distance, which effectively normalizes uneven variances and accounts for correlations among directional relationships. Conceptually, this loss aligns directional relationships in a covariance-whitened space, functioning as a whitened L2-norm.

We begin by defining the Gramian difference between the student and teacher as

$$D = G_{(s)} - G_{(t)} \in \mathbb{R}^{B \times B}, \tag{4}$$

and interpret each row $D_{i,:} \in \mathbb{R}^B$ as a directional relationship vector over the batch. Based on these row-wise directional relationships, we estimate the empirical batch covariance:

$$\Sigma = \mathrm{Cov}\left(\{D_{i,:}\}_{i=1}^{B}\right) \in \mathbb{R}^{B \times B}. \tag{5}$$

Given D and Σ, we define the direction-wise loss as

$$L_{\text{DIR}} = \frac{1}{B} \sum_{i=1}^{B} \sqrt{D_{i,:}^{\top} \, \Sigma^{-1} \, D_{i,:}}. \tag{6}$$

This covariance-aware form balances contributions across directional relationships and prevents unstable, high-variance components from dominating the objective. Despite its conceptual appeal, this formulation introduces two practical challenges.

First, Σ is not guaranteed to be positive definite, which can lead to instability during matrix inversion due to near-singularity [18]. Second, computing Σ^{-1} incurs a cubic complexity of $\mathcal{O}(d^3)$, where $d = B$ is the dimensionality of the Gramian difference, potentially creating a computational bottleneck. These issues compromise numerical robustness and hinder practical deployment. To mitigate them, we apply two stabilization strategies: Tikhonov regularization ensures the positive definiteness, and Cholesky decomposition is used to accelerate and stabilize the inversion process. They are described below.

3.3.3. Stabilizing Direction-wise Loss

To guarantee the positive definiteness and numerical invertibility of the empirical covariance matrix Σ, we apply Tikhonov regularization by adding a scaled identity term λI, yielding a stabilized covariance $\Sigma' = \Sigma + \lambda I$ with $\lambda > 0$. We then apply Cholesky decomposition to factorize the regularized matrix as:

$$\Sigma' = LL^{\top}, \quad L = \mathrm{Cholesky}(\Sigma') \tag{7}$$

Here, L denotes a unique lower triangular matrix with positive diagonal entries from the Cholesky decomposition of Σ' (i.e., $\Sigma' = LL^{\top}$), which implies $(LL^{\top})^{-1} = L^{-\top}L^{-1}$. Consequently, the final direction-wise loss is defined as

$$L_{\text{DIR}} = \frac{1}{B} \sum_{i=1}^{B} \sqrt{D_{i,:}^{\top} \, \Sigma'^{-1} \, D_{i,:}}. \tag{8}$$

This stabilized formulation preserves semantic directionality while ensuring numerical robustness. We then formally show that L_{DIR} is exactly the L2-norm in the covariance-whitened space defined by Σ'.

Proposition (Equivalence to a Whitened L2-norm form) We show that the proposed L_{DIR} can be expressed as an L2-norm in a covariance-whitened space. In Section 3.3.3, the regularized covariance is defined as $\Sigma' = \Sigma + \lambda I$. Then, Σ' is represented by the Cholesky decomposition $\Sigma' = LL^{\top}$, where L is a unique lower triangular matrix. Given these notations, L_{DIR} can be substituted as follows:

$$L_{\text{DIR}} = \frac{1}{B} \sum_{i=1}^{B} \sqrt{D_{i,:}^{\top} (LL^{\top})^{-1} D_{i,:}}. \tag{9}$$

Since $(LL^{\top})^{-1} = (L^{\top})^{-1}L^{-1}$, the term inside the square root can be rewritten as a standard L2-norm:

$$\frac{1}{B} \sum_{i=1}^{B} \sqrt{(L^{-1}D_{i,:})^{\top}(L^{-1}D_{i,:})} = \frac{1}{B} \sum_{i=1}^{B} \sqrt{||L^{-1}D_{i,:}||_2^2}. \tag{10}$$

Hence, the proposed direction-wise loss is expressed as

$$L_{\text{DIR}} = \frac{1}{B} \sum_{i=1}^{B} ||L^{-1}D_{i,:}||_2. \tag{11}$$

This equivalence shows that L_{DIR} is variance- and correlation-aware yet retains the simplicity of the L2-norm objective.

3.4. Streamlined Knowledge Distillation

Having defined both instance-wise and direction-wise knowledge, we now present the final formulation of our proposed SKD. Figure 2 provides a schematic overview of the SKD process. The total loss is defined as:

$$L_{\text{SKD}} = L_{\text{INS}} + L_{\text{DIR}}. \tag{12}$$

By integrating these two complementary components, SKD enables the student model to capture fine-grained class-wise

semantics and pairwise relational structures embedded in the teacher's output space. This allows for more comprehensive and efficient knowledge transfer and eliminates the need for extra alignment or relational structure modeling. The full PyTorch-style pseudocode of our SKD is provided in Algorithm 1.

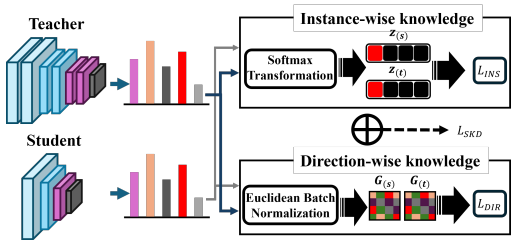


Figure 2. Schematic overview of our Streamlined Knowledge Distillation.

Algorithm 1 Pseudocode of SKD in a PyTorch-like style.

```
# l_stu: output logits of student
# l_tea: output logits of teacher
# lamb: tikhonov regularization factor
# tau: temperature
pred_stu = F.softmax(l_stu / tau)
pred_tea = F.softmax(l_tea / tau)
# instance-wise loss
l_instance = F.kl_div(pred_stu, pred_tea)
norm_stu = F.normalize(l_stu)
norm_tea = F.normalize(l_tea)
G_stu = torch.mm(norm_stu, norm_stu.T)
G_tea = torch.mm(norm_tea, norm_tea.T)
diff = G_stu - G_tea
cov_mat = torch.cov(diff.T) +
          lamb * torch.eye(diff.shape[0])
L = torch.linalg.cholesky(cov_mat)
cov_mat_inv = torch.cholesky_inverse(L)
dist = torch.einsum('bi,ij,bj->b',
           diff, cov_mat_inv, diff)
# direction-wise loss
l_direction = torch.sqrt(dist).mean()
total_loss = l_instance + l_direction
```

4. Experiments

4.1. Experimental settings

We conducted experiments on the CIFAR-100 [19] and ImageNet [7] datasets to evaluate our method for image classification. Also, we utilized the MS-COCO dataset [21] for object detection. To assess the effectiveness of our method, we consider two types of teacher-student architecture pairs: (1) homogeneous pairs using the same architectures (e.g., VGG16 $\rightarrow$ VGG8), and (2) heterogeneous pairs involving different architectures (e.g., ResNet50 $\rightarrow$ MobileNetV1). We included a range of architectures covering different capacity scales, such as ResNet, WideResNet (WRN), ShuffleNetV1/V2, MobileNetV1, and VGG.

Implementation Details For CIFAR-100, we follow the experimental setup of CRD [31]. Models are trained using SGD for 240 epochs with a batch size of 64. The initial learning rate is set to 0.1 and decayed by a factor of 0.1 at epochs 150, 180, and 210. Momentum and weight decay are set to 0.9 and 5×10^{-4}, respectively. For ImageNet, models are trained for 100 epochs with a batch size of 512. Initial learning rate is set to 0.2, decayed by a factor of 0.1 every 30 epochs. We use momentum 0.9 and weight decay 1×10^{-4} for all ImageNet experiments. For MS-COCO, we set 180,000 epochs for training with a batch size of 8, learning rate of 0.01. All experiments are conducted using Intel Gaudi (Habana HL-200) and NVIDIA A6000 accelerators. Each experiment is repeated five times, and we report the average results across runs.

Experimental Results We evaluate our method on CIFAR-100, ImageNet and MS-COCO under both homogeneous and heterogeneous architecture settings. On CIFAR-100, Table 1 shows that SKD achieves notable improvements over prior logit-based KD methods. Also, SKD even surpasses feature-based KD methods despite its simplicity. Under heterogeneous settings (Table 2), SKD consistently boosts the performance of student models, outperforming previous logit-based and feature-based methods. To validate scalability, we further conducted experiments on ImageNet. As shown in Table 3, SKD maintains robust performance across both homogeneous (e.g., ResNet34 $\rightarrow$ ResNet18) and heterogeneous (e.g., ResNet50 $\rightarrow$ MobileNetV1) settings. Also, we evaluate the performance of SKD on object detection using Faster R-CNN [28]-FPN [22] and report AP, AP_{50}, and AP_{75} as metrics. As shown in Table 4, SKD provides consistent AP gains and remains competitive or superior to prior KD methods. Collectively, these results demonstrate that SKD improves student models under resource constraints and generalizes well to large-scale datasets and tasks.

4.2. Analysis

While the previous experiments demonstrate the effectiveness of our method, we perform additional analyses to better understand its behavior under different conditions and design choices.

Ablation Study We investigate the effectiveness of each component in our method through an ablation study. The baseline configuration combines instance-wise knowledge with direction-wise knowledge computed from softmax-transformed logits and a simple L2-norm. We then

Table 1. Comparison of Top-1 accuracy on CIFAR-100. We report the individual performance of the teacher and student models under homogeneous architecture settings.

Method							
	Teacher	ResNet56 72.40	ResNet110 74.31	ResNet32x4 79.42	WRN-40-2 75.59	WRN-40-2 75.59	VGG13 74.64
	Student	ResNet20 65.79	ResNet32 67.92	ResNet8x4 72.48	WRN-16-2 71.12	WRN-40-1 69.54	VGG8 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**

Table 2. Comparison of Top-1 accuracy on CIFAR-100. We report the individual performance of the teacher and student models under heterogeneous architecture settings.

Method						
	Teacher	ResNet32x4 79.42	WRN-40-2 75.59	VGG13 74.64	ResNet50 79.33	ResNet32x4 79.42
	Student	ShuffleNetV1 69.43	ShuffleNetV1 69.43	MobileNetV1 60.09	MobileNetV1 60.09	ShuffleNetV2 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**

progressively apply EBN, replace the L2-norm with a Mahalanobis-based distance, and incorporate stabilization tricks (Tikhonov regularization and Cholesky decomposition). Each step leads to steady performance gains and this confirms the effectiveness of our design components. As shown in Table 5, every component contributes meaningfully to overall accuracy. This result underscores the importance of preserving directional relationships and ensuring optimal distillation.

Performance Comparison with Teacher Model We further analyze the performance gap between the teacher and student models under our SKD method. As shown in Table 6, we report the gap between their Top-1 accuracies, where a negative value indicates that the student outperforms the teacher. Across the results, the student matches or exceeds the teacher in most cases, indicating that SKD retains transfer efficiency despite reduced capacity. We conjecture that this phenomenon stems from the effective guidance provided by our direction-wise knowledge, which encourages the student to learn a more structured representation space.

Complementarity with Feature-based Methods Our SKD method can function not only as a logit-based standalone distillation strategy, but also as a complementary component to enhance existing feature-based methods. Since it does not require any additional projection layers or auxiliary modules, it can be seamlessly integrated into various feature-based methods. When combined with FitNet and RKD, our SKD consistently improves them, with absolute gains from 0.11% to 3.55% (see Table 7). These results suggest that our direction-aware logit distillation complements intermediate feature supervision by reinforcing struc-

Table 3. Comparison of Top-1 accuracy on ImageNet with both homogeneous (e.g., R34 → R18) and heterogeneous (e.g., R50 → MV1) teacher-student architectures.

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**

Table 4. We conduct object detection with Faster-RCNN [28]-FPN [22]. Teacher → Student backbone pairs under homogeneous and heterogeneous settings. Evaluation metrics are AP, AP_{50}, and AP_{75}.

		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Method	Teacher	ResNet101			ResNet101			ResNet50		
		42.04	62.48	45.88	42.04	62.48	45.88	40.22	61.02	43.81
	Student	ResNet18			ResNet50			MobileNetV2		
		33.26	53.61	35.26	37.93	58.84	41.05	29.47	48.87	30.90
Feature	FitNet [29]	34.13	54.16	36.71	38.76	59.62	41.80	30.20	49.80	31.69
	FGFI [35]	35.44	55.51	38.17	39.44	60.27	43.04	31.16	50.68	32.92
	ReviewKD [5]	36.75	56.72	34.00	40.36	60.97	44.08	33.71	53.15	36.13
Logit	KD [16]	33.97	54.66	36.62	38.35	59.41	41.71	30.13	50.28	31.35
	DKD [42]	35.05	56.60	37.54	39.25	60.90	42.73	32.34	53.77	34.01
	MLKD [17]	36.03	**57.28**	38.51	40.15	**61.67**	44.57	33.83	54.01	35.22
	SKD (Ours)	**36.74**	57.03	**39.76**	**40.62**	61.61	**44.61**	**34.25**	**54.51**	**36.29**

Table 5. Results of the ablation study on CIFAR-100, with ResNet32x4 as the teacher and ResNet8x4 as the student. Top-1 accuracy is reported as the evaluation metric. We defined Euclidean Batch Normalization as EBN, Mahalanobis as Mal.

Baseline	EBN (X: Softmax)	Mal Loss (X: L2)	Stability Tricks (X: None)	Top-1 accuracy
O	X	X	X	76.61
O	O	X	X	77.28
O	O	O	X	77.76
O	O	O	O	**78.33**

Table 6. Experiments are implemented on CIFAR-100, with teacher and student in a homogeneous architecture. Top-1 accuracy as the evaluation metric. Note that when the student model outperforms the teacher model, the gap is negative. The settings are the same as Table 1.

Teacher	72.40	74.31	79.42	75.59	75.59	74.64
Student(Ours)	72.50	74.84	78.33	76.60	76.04	75.75
Gap	-0.10	-0.53	1.09	-1.01	-0.45	-1.11

tural alignment in the output space.

Applicability to Vision Transformers Recently, Vision Transformers (ViTs) have attracted considerable attention due to their competitive performance in classification benchmarks [40], but their high computational cost poses challenges for practical use. Then, DeiT [32] and Swin [23] emerged as the solution to these problems. Specifically,

Table 7. CIFAR-100 results (Top-1 accuracy) of combining our method with feature-based distillation methods. Experiments follow the same setting as Table 1.

FitNet [29]	69.64	72.35	75.26	74.46	73.28	70.09
+ Ours	**71.29**	**73.46**	**77.01**	**75.60**	**74.68**	**72.24**
RKD [25]	70.51	73.31	74.55	74.08	73.52	71.61
+ Ours	**71.42**	**73.71**	**75.78**	**74.45**	**73.63**	**75.16**

DeiT mitigates this issue through soft and hard distillation and Swin improves efficiency and multi-scale features. However, these methods still struggle to exploit relational structures in the output space. Given this limitation, we apply SKD to ViT-Tiny with RegNetY-16GF as the teacher to assess its ability to model relational structures more effectively. As shown in Table 8, our method achieves superior accuracy compared to the baselines, and it proves effective even on non-convolutional architectures such as ViTs.

Visualization We present visualizations from two perspectives to show that SKD transfers knowledge effectively. (1) To quantify how well relational structure is preserved relative to the L2-norm, we visualize teacher-student Gramian differences on normalized logits. As shown in Figure 4(a-c), SKD has lighter off-diagonal regions than the student and L2-norm. This implies higher representation similarity and stronger preservation of pairwise directional relationships. We also report linear Centered Kernel Alignment (CKA) score with the teacher computed from Gramian

Table 8. Top-1 accuracy on ImageNet for ViT-Tiny student distilled from RegNetY-16GF teacher. We compare Soft KD, Hard KD, DeiT, and our SKD.

Teacher	Student	Soft KD [32]	Hard KD [32]	DeiT [32]	Swin [23]	Ours
RegNetY-16GF (82.90)	ViT-T (66.63)	67.01	67.27	68.03	67.88	**68.68**

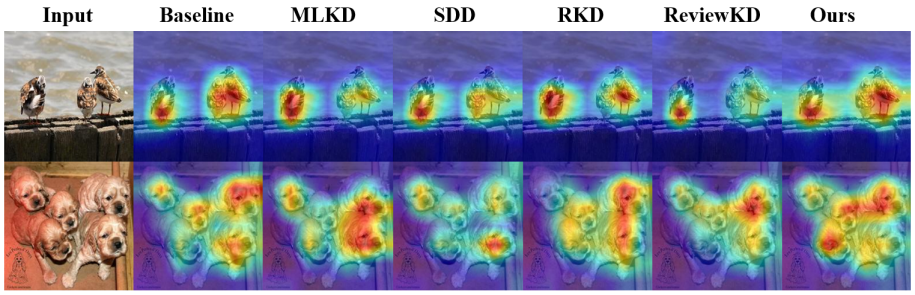
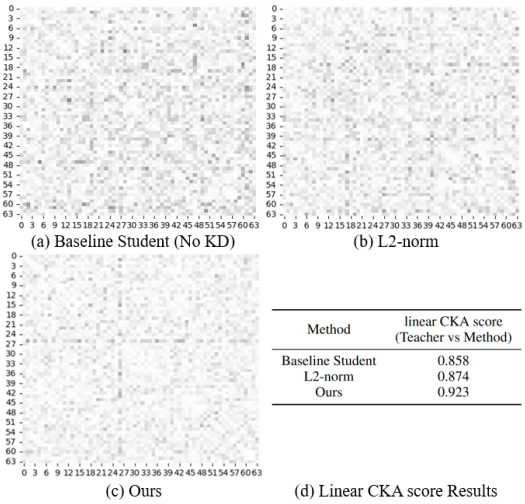


Figure 3. Grad-CAM++ heatmaps of MobileNetV1 trained with different KD methods.

matrices in Figure 4(d). SKD attains the highest linear CKA score (over the baseline student and L2-norm), confirming the advantage of our method in preserving relational structure. (2) To assess semantic focus, we visualize discriminative heatmaps using Grad-CAM++ [2] in Figure 3 (teacher: ResNet50, student: MobileNetV1 on ImageNet). Heatmaps indicate that SKD enables the student to attend to more fine-grained and important regions than prior feature-based and logit-based approaches. These visualizations confirm that SKD enables effective transfer of fine-grained semantics and relational structures, allowing the student to utilize its limited capacity more efficiently.



(a) Baseline Student (No KD) (b) L2-norm (c) Ours

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

(d) Linear CKA score Results

Figure 4. (a-c) Heatmaps of the absolute teacher-student Gramian difference, $|G_{(s)} - G_{(t)}|$, computed from Gramian matrix of normalized logits: (a) Baseline Student, (b) L2-norm, (c) Ours. Lighter colors indicate closer to the teacher. (d): Linear CKA scores with the teacher computed from Gramian matrices.

Training Efficiency We compare the per-batch training time of various KD methods under identical conditions. Specifically, we use ResNet56 as the teacher and ResNet20 as the student, and report the average training time over fifty runs. As shown in Figure 5, our method achieves the shortest training time among representative logit-based and feature-based approaches. This efficiency stems from our streamlined design, which avoids complex components and transfers only two forms of knowledge.

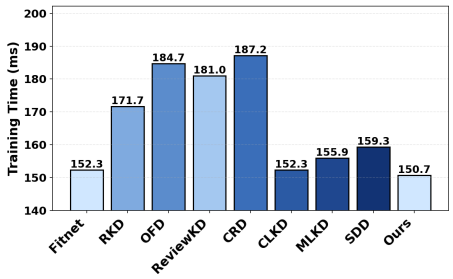


Figure 5. Average per-batch training time (ms) on CIFAR-100.

5. Conclusion

In this paper, we proposed Streamlined Knowledge Distillation (SKD), a simple yet effective logit-based distillation. By transferring only two forms of knowledge–instance-wise semantics and direction-wise relational structure–with stabilized learning objectives, SKD simplifies multi-knowledge distillation and removes the need for additional alignment or relational structure modeling. Extensive experiments and comprehensive analyses show that SKD consistently improves student performance across diverse scenarios, confirming its effectiveness and robustness. Nonetheless, the covariance Σ' in L_{DIR} can be sensitive under large batch sizes or noisy labels, which may affect stability. Future work will improve covariance estimation and extend SKD to sequential and multi-modal domains.

Acknowledgements. This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korean Government of the Ministry of Science and Information and Communication Technology (MSIT), South Korea, Digital Columbus Project (RS-2025-02304331) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25472714).

References

[1] Ziqi Bing, Linyang Li, and Junhao Liang. Optimizing knowledge distillation in transformers: Enabling multi-head attention without alignment barriers. *arXiv preprint arXiv:2502.07436*, 2025. 2

[2] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 8

[3] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7028–7036, 2021. 1

[4] Jia Chen, Xuan Zhao, Huan Zheng, Xiaoyang Li, Shiming Xiang, and Hongkai Guo. Robust knowledge distillation based on feature variance against backdoored teacher model. *Applied Soft Computing*, 163:111907, 2024. 1

[5] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021. 2, 6, 7

[6] Hengliang Cheng, Lingli Yang, and Zhaoyu Liu. Relation-based knowledge distillation for anomaly detection. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part I*, pages 105–116. Springer, 2021. 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 3

[9] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021. 1

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 3

[11] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11020–11029, 2020. 3

[12] Gousia Habib, Sheikh Musa Kaleem, Tufail Rouf, Brejesh Lall, et al. A comprehensive review of knowledge distillation in computer vision. *arXiv preprint arXiv:2404.00936*, 2024. 1

[13] Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. The elements of statistical learning, 2009. 3

[14] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hwanjun Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019. 2, 6

[15] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3779–3787, 2019. 2

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 3, 6, 7

[17] Yukun Jin, Jindong Wang, and Dahua Lin. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24276–24285, 2023. 1, 3, 6, 7

[18] Andrew D Ker. Stability of the mahalanobis distance: A technical note. *Technical Report CS-RR-10-20*, 2010. 4

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[20] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3041–3050, 2023. 1

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 5, 7

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7, 8

[24] William Menke. Review of the generalized least squares method. *Surveys in Geophysics*, 36(1):1–25, 2015. 3

[25] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 2, 6, 7

[26] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Heterogeneous knowledge distillation using information flow modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2339–2348, 2020. 1

[27] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5007–5016, 2019. 3

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 5, 7

[29] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2, 6, 7

[30] Wujie Sun, Defang Chen, Siwei Lyu, Genlang Chen, Chun Chen, and Can Wang. Knowledge distillation with refined logits. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1110–1119, 2025. 2, 6, 7

[31] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 2, 5, 6

[32] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 7, 8

[33] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1365–1374, 2019. 2, 3

[34] Tao Wang. Learning to detect and segment for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7051–7060, 2023. 1

[35] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4933–4942, 2019. 7

[36] Shicai Wei, Chunbo Luo, and Yang Luo. Scaled decoupled distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15975–15983, 2024. 1, 3, 6, 7

[37] Yuchen Xu, Cheng Cao, Feng Yuan, Rui Mi, Dong Wang, Ying Liu, and Meng Su. Data-efficient knowledge distillation with teacher assistant-based dynamic objective alignment. In *International Conference on Computational Science*, pages 181–195. Springer Nature Switzerland, 2024. 2

[38] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017. 3

[39] Mengwen Yuan, Chengjun Zhang, Ziming Wang, Huixiang Liu, Gang Pan, and Huajin Tang. Trainable spiking-yolo for low-latency and high-performance object detection. *Neural Networks*, 172:106092, 2024. 1

[40] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 7

[41] Shuang Zhang, Haiping Liu, John E Hopcroft, and Kaiming He. Class-aware information for logit-based knowledge distillation. *arXiv preprint arXiv:2211.14773*, 2022. 3, 6

[42] Bowen Zhao, Qibin Cui, Rui Song, Yikang Qiu, and Jian Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962, 2022. 1, 2, 3, 6, 7

[43] Zhenyu Zhou, Yifan Shen, Shishuo Shao, Ling Gong, and Sheng Lin. Rethinking centered kernel alignment in knowledge distillation. *arXiv preprint arXiv:2401.11824*, 2024. 3