Direction Sensitivity–based Knowledge Distillation: Optimization-Aware Low-Rank Knowledge Transfer

Yongkai Liao[1], **Xinxing Chen**[1,2*], **Zhongzheng Fu**[1], **Haoyuan Wang**[1], **Jian Huang**[1*]

[1]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, 430074, China

[2]Shenzhen Huazhong University of Science and Technology Research Institute, Shenzhen, 518052, China

{lyk, cxx, fuzhongzheng, why427} @hust.edu.cn, huang_jan@mail.hust.edu.cn

Abstract

Knowledge distillation (KD) aims to enhance the performance of lightweight student networks through the guidance of teacher models. However, the existing methods have deficiencies in two key aspects: First, these methods rely heavily on static representation alignment, failing to account for optimization sensitivity in different directions within the distillation subspace; second, they lack a fine-grained mechanism to align critical directional features. To address these issues, we propose Direction Sensitivity–based Knowledge Distillation method (DSKD), which can quantitatively measure the sensitivity of each direction to the loss function at different training stages and dynamically select the optimization direction accordingly. Meanwhile, we designed a directional sensitivities weighted distillation loss. By aligning the parameter matrices of the teacher and student models in the key directions, we can more effectively transfer knowledge and improve the distillation effect. We combined DSKD with multiple advanced distillation strategies and conducted an empirical evaluation in the GLUE benchmark and CIFAR-100. The results showed that this method could significantly improve the performance of existing distillation techniques.

Code — https://github.com/theChoseno/DSKD

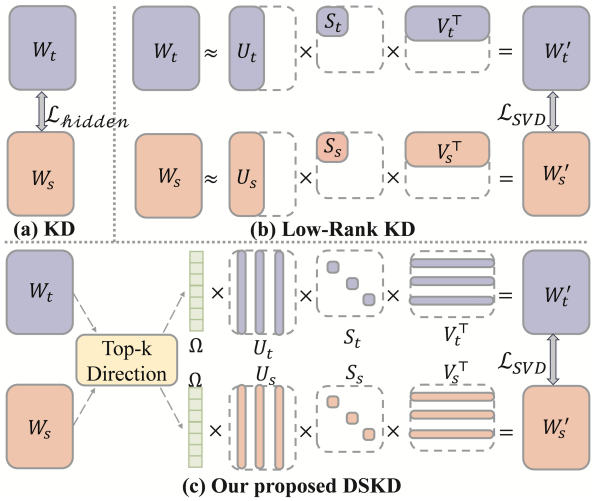


Figure 1: The differences between (a) vanilla knowledge distillation (b) low rank knowledge distillation and (c) direction sensitivity-based knowledge distillation (DSKD). Our method uses key directions to construct low-rank distillation losses.

Introduction

Knowledge Distillation (KD) has emerged as a cornerstone paradigm for model compression, enabling efficient deployment of deep learning systems in resource-constrained scenarios across computer vision and natural language processing (Hinton, Vinyals, and Dean 2015). By transferring knowledge from a high-capacity teacher to a compact student model, KD effectively balances performance retention with reduced inference latency and memory footprint. The proliferation of powerful Transformer-based pre-trained language models (PLMs) such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) has intensified the urgency for effective compression techniques, as their computational demands often preclude deployment on edge devices. Consequently, designing KD strategies that preserve the nuanced representational capacity of these architectures while accommodating stringent resource limitations has become a critical research frontier (Gou et al. 2021), necessitating innovations beyond conventional distillation frameworks.

Among the many KD strategies, low-rank distillation (Figure.1(b)) has recently attracted attention due to its ability to compress model representations while preserving critical information. These approaches typically rely on Singular Value Decomposition (SVD) to extract the dominant directions—those corresponding to the largest singular values—from weight or feature matrices as knowledge carriers (Zhang et al. 2025; Lee, Kim, and Song 2018). While effective in reducing dimensionality, this strategy is inherently static: it ranks directions solely by singular value magnitude, under the assumption that the largest ones always capture the most informative structure. However, this assumption has limitations. Recent studies suggest that the influence of a direction on model performance may not always align

*Corresponding authors.

with its singular value rank—especially after fine-tuning, where directions with small singular values can exert significant impact (Staats, Thamm, and Rosenow 2024; Oymak et al. 2019). Moreover, SVD-based approximation assumes uniform importance across parameters, often leading to degraded performance in downstream tasks (Hsu et al. 2022). Attempts to improve low-rank estimation by re-weighting parameter importance (Hua et al. 2023, 2022) mark a step forward, but still fall short of modeling how these directions influence loss function during training. In light of these findings, we argue that direction selection in KD should be guided not merely by static structure (e.g., singular value magnitude), but by its optimization sensitivity—that is, how strongly a direction responds to the loss function during training. While sensitivity-based metrics have been used in pruning and robustness (Liu et al. 2024b,a), they focus on parameter-level importance or post-hoc analysis—aimed at removing weights, not preserving critical directions in knowledge transfer. This highlights a key gap: current distillation methods lack a mechanism to identify singular directions that are functionally important during optimization. Relying on static criteria (e.g., singular value magnitude) or heuristic reweighting, they fail to capture how directional influence evolves in training. As a result, informative low-energy directions may be ignored, while dominant but unstable ones may mislead the student. No existing method dynamically selects optimization-relevant subspaces—those most sensitive to the loss landscape—to guide knowledge transfer.

To address the above problem, this paper proposes Direction Sensitivity–based Knowledge Distillation (DSKD), a framework that selects low-rank subspaces based on their optimization sensitivity rather than static singular values (Figure.1(c)). We introduce a training-phase-aware metric that fuses gradient and curvature signals to identify the most influential directions during training, and design a direction-weighted distillation loss to align teacher and student models within this adaptive subspace. Our approach shifts the paradigm of low-rank distillation from structure-centric to optimization-aware. By dynamically integrating first-order and second-order sensitivity signals to adaptively update the distillation subspace, DSKD establishes a functional alignment mechanism that evolves with the optimization trajectory, moving beyond the limitations of static SVD. Compared to traditional structure-only methods, our approach provides a more interpretable and targeted mechanism for knowledge transfer by considering the stage-specific direction importance. Our work makes three key contributions:

- We introduce a novel perspective on direction selection in knowledge distillation by considering optimization sensitivity rather than static singular value magnitudes. Our method jointly leverages first-order and second-order information to quantify directional importance, and employs a training-phase-aware fusion strategy that dynamically adapts direction selection throughout the optimization process.
- We introduce a direction-weighted distillation loss that emphasizes optimization-sensitive directions via top-k selection for low-rank reconstruction, enabling more effective knowledge transfer than structure-based methods.
- Extensive experiments on GLUE and CIFAR-100 show consistent improvements over state-of-the-art baselines, demonstrating the effectiveness of our approach in enhancing distillation across downstream tasks.

Related Work

Knowledge Distillation

KD transfers knowledge from teacher to student via structured alignment (Hinton, Vinyals, and Dean 2015). Methods include logit-based KD (Jing Yang 2021; Li et al. 2022; Li and Zhe 2022; Zhou et al. 2023; Li et al. 2023; Gong et al. 2023; Sun et al. 2024; Gao et al. 2025), using softened outputs, and feature-based KD (Park et al. 2019; Tung and Mori 2019; Chen et al. 2022; Guo et al. 2023; Yang et al. 2024; Li et al. 2025; Dai et al. 2025), aligning intermediate representations. However, direct feature matching between disparate Transformers often causes optimization instability due to student capacity limits (Liang et al. 2023). Attention (Zagoruyko and Komodakis 2017; Jiao et al. 2020) and relational distillation (Tian, Krishnan, and Isola 2020) partially help but fail to balance fidelity and learnability. DSKD identifies discriminative teacher subspaces via loss sensitivity analysis, guiding the student toward critical knowledge without architectural overhead—enabling more effective and targeted distillation.

SVD-based Distillation Methods

SVD is a widely used technique for model compression (Denton et al. 2014; Zhang et al. 2015; Yang et al. 2020) and has been extended to KD by selecting top singular directions as knowledge carriers (Lee, Kim, and Song 2018; Zhang et al. 2025). Extensions include cross-model singular vector alignment and tensor-based student enhancement (Zhan et al. 2024). However, these methods typically assume that large singular values directly indicate direction importance, overlooking that some low-energy directions can become critical during fine-tuning (Staats, Thamm, and Rosenow 2024). To address this, re-weighting approaches have been proposed (Hua et al. 2023, 2022), but they often lack explicit modeling of training dynamics. Our work extends this line by incorporating both gradient and Hessian information to adaptively identify optimization-relevant directions.

Low-Rank Approximation and Modeling

Low-rank modeling is fundamental to efficient neural architectures, with methods such as Linformer (Wang et al. 2020), automatic rank selection (Gao et al. 2024; Horváth et al. 2024; Ryu et al. 2024) and curvature-aware approximation (Li et al. 2024). These studies show that effective low-rank strategies preserve performance when guided by appropriate selection criteria. Yet, few works integrate optimization sensitivity—captured via gradients and curvature—into directional selection. Our method fills this gap by quantifying directional responses to the loss function, enabling principled low-rank distillation aligned with the training dynamics.

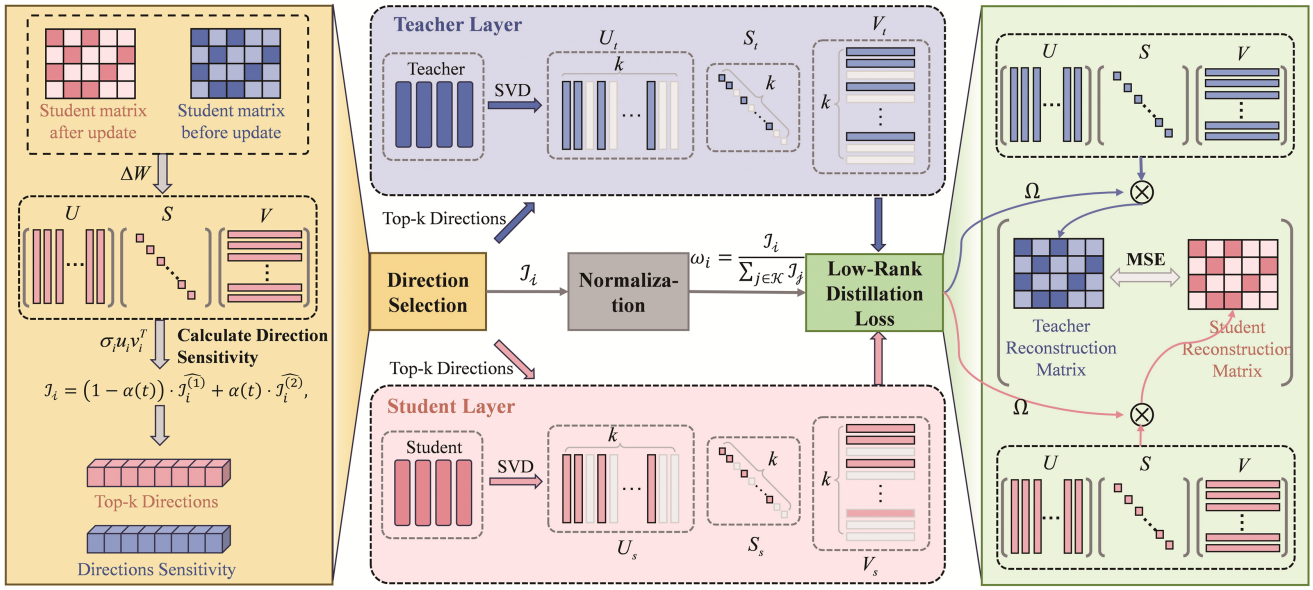


Figure 2: An overview of our DSKD. We apply SVD on the parameter update part of the student model and use the direction sensitivity metric to select the k most sensitive singular directions. Using normalized sensitivity weights, we reconstruct low-rank approximations of both teacher and student models along these directions. The alignment in the sensitive subspace is then measured via MSE loss between the reconstructed matrices. Our approach preserves the most influential update directions and enables fine-grained, direction-aware knowledge transfer.

Method

This section provides a detailed introduction of DSKD. The overall framework is illustrated in Figure.2

SVD Decomposition of Matrix

Let $\mathbf{W} \in \mathbb{R}^{m\times n}$ denote a parameter matrix in the teacher model, with SVD:

$$\mathbf{W} = \mathbf{U}\mathbf{S}\mathbf{V}^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \tag{1}$$

where $r = \text{rank}(\mathbf{W})$, σ_i is the i-th singular value, and $\mathbf{u}_i$, $\mathbf{v}_i$ are the corresponding left and right singular vectors.

Conventional low-rank distillation methods (Zhang et al. 2025; Lee, Kim, and Song 2018) select the top-k singular directions based on magnitude (σ_i), assuming larger values indicate greater importance. However, this static criterion ignores the dynamic role of directions during optimization. A direction with small σ_i may still be critical if it induces large changes in the loss function. To address this, we propose a training-phase-aware direction selection strategy grounded in optimization sensitivity.

Loss Sensitivity under Singular Direction Perturbation

We quantify the impact of each singular direction $(\mathbf{u}_i, \mathbf{v}_i)$ by analyzing the second-order Taylor expansion of the loss $\mathcal{L}(\mathbf{W})$ under perturbation $\Delta\mathbf{W} = \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$:

$$\mathcal{L}(\mathbf{W}+\Delta\mathbf{W}) = \mathcal{L}(\mathbf{W}) + \langle \nabla\mathcal{L}(\mathbf{W}), \Delta\mathbf{W}\rangle \tag{2}$$

$$+ \frac{1}{2}\langle \Delta\mathbf{W}, \mathcal{H}(\mathbf{W})[\Delta\mathbf{W}]\rangle + o(\|\Delta\mathbf{W}\|_F^2), \tag{3}$$

where $\nabla\mathcal{L}(\mathbf{W})$ is the gradient and $\mathcal{H}(\mathbf{W})$ is the Hessian operator, $\langle\cdot,\cdot\rangle$ denotes the Frobenius inner product.

The first-order term evaluates gradient alignment:

$$\langle \nabla\mathcal{L}, \Delta\mathbf{W}\rangle = \sigma_i \cdot \mathbf{u}_i^\top \nabla\mathcal{L}(\mathbf{W})\mathbf{v}_i. \tag{4}$$

The second-order term captures curvature response:

$$\langle \Delta\mathbf{W}, \mathcal{H}[\Delta\mathbf{W}]\rangle = \sigma_i^2 \cdot (\mathbf{v}_i \otimes \mathbf{u}_i)^\top \mathcal{H}(\mathbf{W})(\mathbf{v}_i \otimes \mathbf{u}_i), \tag{5}$$

where $\otimes$ denotes the Kronecker product.

Dynamic Direction Selection

In the early training stage, the parameter $\mathbf{W}$ is far from the local optimum, and the gradient dominates the update direction. A first-order Taylor approximation yields:

$$|\Delta\mathcal{L}_i| = \sigma_i \cdot \left|\mathbf{u}_i^\top \mathbf{G}\mathbf{v}_i\right| + o(\sigma_i^2), \tag{6}$$

where $\mathbf{G} = \nabla\mathcal{L}(\mathbf{W})$, the second-order term is negligible when σ_i is small. Therefore, we define the first-order metric as:

$$\mathcal{I}_i^{(1)} = \sigma_i \cdot |\mathbf{u}_i^\top \mathbf{G}\mathbf{v}_i| \tag{7}$$

In the later training stage, assuming the model approaches a local optimum $\mathbf{W}^*$ with $\nabla\mathcal{L}(\mathbf{W}^*) = 0$, the second-order

loss change under a rank-one perturbation $\sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ is:

$$\mathcal{L}(\mathbf{W}^* + \sigma_i \mathbf{u}_i \mathbf{v}_i^\top) - \mathcal{L}(\mathbf{W}^*) \tag{8}$$

$$= \frac{1}{2}\sigma_i^2 \cdot (\mathbf{v}_i \otimes \mathbf{u}_i)^\top \mathcal{H}(\mathbf{W}^*)(\mathbf{v}_i \otimes \mathbf{u}_i). \tag{9}$$

To avoid Hessian computation, we approximate it via the empirical Fisher matrix $\mathbf{F} \approx \mathbb{E}[\mathbf{G}\mathbf{G}^\top]$ (Kunstner, Balles, and Hennig 2019; Li et al. 2024), noting that $(\mathbf{v}_i \otimes \mathbf{u}_i)^\top \mathbf{F}(\mathbf{v}_i \otimes \mathbf{u}_i) \approx (\mathbf{u}_i^\top \mathbf{G}\mathbf{v}_i)^2$.

We can get:

$$(\mathbf{v}_i \otimes \mathbf{u}_i)^\top \mathcal{H}(\mathbf{W}^*)(\mathbf{v}_i \otimes \mathbf{u}_i) \tag{10}$$

$$\approx (\mathbf{v}_i \otimes \mathbf{u}_i)^\top \mathbf{F}(\mathbf{v}_i \otimes \mathbf{u}_i) \tag{11}$$

$$\approx (\mathbf{u}_i^\top \mathbf{G}\mathbf{v}_i)^2 \tag{12}$$

$$\leq |(\mathbf{u_i}^\top \mathbf{G}\mathbf{G}^\top \mathbf{u_i})(\mathbf{v_i}^\top \mathbf{G}^\top \mathbf{G}\mathbf{v_i})|. \tag{13}$$

We propose a conservative second-order term sensitivity metric:

$$\mathcal{I}_i^{(2)} = \frac{1}{2}\sigma_i^2 \cdot |(\mathbf{u_i}^\top \mathbf{G}\mathbf{G}^\top \mathbf{u_i})(\mathbf{v_i}^\top \mathbf{G}^\top \mathbf{G}\mathbf{v_i})|, \tag{14}$$

while the Fisher quadratic form $(\mathbf{u}_i^\top \mathbf{G}\mathbf{v}_i)^2$ provides a direct approximation of the Hessian-induced curvature, it is highly sensitive to mini-batch noise. Our proposed upper bound serves as a robust, high-recall alternative. By amplifying directions with high gradient energy in both input ($\mathbf{v}_i$) and output ($\mathbf{u}_i$) spaces, it identifies structurally critical pathways that are likely to exhibit significant second-order effects, even when the instantaneous joint response $(\mathbf{u}_i^\top \mathbf{G}\mathbf{v}_i)$ is moderate. This makes it a reliable criterion for identifying optimization-relevant directions in the late training phase.

To adaptively balance these across training, we introduce a time-varying coefficient $\alpha(t) \in [0, 1]$ and define the composite sensitivity score:

$$\mathcal{I}_i = (1 - \alpha(t)) \cdot \hat{\mathcal{I}}_i^{(1)} + \alpha(t) \cdot \hat{\mathcal{I}}_i^{(2)}, \tag{15}$$

where $\hat{\mathcal{I}}_i = \frac{\mathcal{I}_i}{\sum \mathcal{I}_i}$ denotes normalization across directions to ensure numerical stability and fair comparison. $\alpha(t)$ increases from 0 to 1 over training steps.

We select the top-k most sensitive directions:

$$\mathcal{K} = \text{argmax}_i^{(k)} \mathcal{I}_i, \tag{16}$$

forming an optimization-aware subspace for distillation.

To balance efficiency and adaptivity, we update the direction set $\mathcal{K}$ periodically every T_s steps which ensures that sensitivity rankings remain stable over short intervals. Periodic updates reduce computational overhead from repeated gradient, while still capturing dynamic changes in directional importance.

Weighted Low-Rank Distillation Loss

Let $\mathbf{W}_s$ and $\mathbf{W}_t$ denote corresponding parameter matrices in the student and teacher models. We compute their SVDs and extract the top-k directions in $\mathcal{K}$ to form low-rank approximations:

$$\mathbf{U}_t^{(k)} = [\mathbf{u}_t^{(i_1)}, \ldots, \mathbf{u}_t^{(i_k)}] \in \mathbb{R}^{m \times k}, \tag{17}$$

$$\mathbf{S}_t^{(k)} = \text{diag}(\sigma_t^{(i_1)}, \ldots, \sigma_t^{(i_k)}) \in \mathbb{R}^{k \times k}, \tag{18}$$

$$\mathbf{V}_t^{(k)} = [\mathbf{v}_t^{(i_1)}, \ldots, \mathbf{v}_t^{(i_k)}] \in \mathbb{R}^{n \times k}, \tag{19}$$

with analogous definitions for $\mathbf{U}_s^{(k)}, \mathbf{S}_s^{(k)}, \mathbf{V}_s^{(k)}$.

To emphasize the importance of sensitive directions, we introduce a diagonal weight matrix $\mathbf{\Omega} = \text{diag}(\omega_{i_1}, \ldots, \omega_{i_k})$, where:

$$\omega_i = \frac{\mathcal{I}_i}{\sum_{j \in \mathcal{K}} \mathcal{I}_j}, \quad \forall i \in \mathcal{K}. \tag{20}$$

This ensures $\sum \omega_i = 1$, promoting numerical stability.

We define the weighted low-rank representations:

$$\widetilde{\mathbf{W}}_t^{(k)} = \mathbf{\Omega}\mathbf{U}_t^{(k)}\mathbf{S}_t^{(k)}(\mathbf{V}_t^{(k)})^\top, \tag{21}$$

$$\widetilde{\mathbf{W}}_s^{(k)} = \mathbf{\Omega}\mathbf{U}_s^{(k)}\mathbf{S}_s^{(k)}(\mathbf{V}_s^{(k)})^\top, \tag{22}$$

note that weighting is applied at the reconstruction level, rescaling the contribution of each direction according to its optimization sensitivity.

The distillation loss is defined as:

$$\mathcal{L}_{\text{SVD}} = \left\| \widetilde{\mathbf{W}}_s^{(k)} - \widetilde{\mathbf{W}}_t^{(k)} \right\|_F^2. \tag{23}$$

Finally, the overall objective combines task-specific loss $\mathcal{L}_{\text{task}}$ (e.g., cross-entropy) and distillation loss:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \cdot \mathcal{L}_{\text{SVD}}, \tag{24}$$

where $\lambda > 0$ balances the two components. This joint objective enables the student to inherit structurally and functionally critical knowledge from the teacher while preserving task performance.

Experiments

We conduct a comprehensive evaluation of our method by integrating it into state-of-the-art knowledge distillation frameworks, aiming to enhance their performance. Experiments span two domains—natural language processing and computer vision—to demonstrate broad effectiveness. Code is provided in the appendix.

Datasets

GLUE The GLUE benchmark (Wang et al. 2019) includes sentiment analysis (SST-2), linguistic acceptability (CoLA), paraphrasing (MRPC, QQP), natural language inference (RTE, MNLI, QNLI), and semantic similarity (STS-B). We report accuracy for MNLI (averaged over matched and mismatched splits), SST-2, QNLI, QQP, and RTE; F1 for MRPC; Matthews correlation for CoLA; and Spearman correlation for STS-B (Wu et al. 2023; Zhou et al. 2023).

CIFAR-100 This image classification dataset (Krizhevsky, Hinton et al. 2009) contains 60,000 images from 100 classes, with 50,000 training and 10,000 test samples. We evaluate using top-1 and top-5 accuracy.

Baseline Models

For the GLUE benchmark, we integrate our module into recent distillation methods: BERT-of-Theseus (Xu et al. 2020), LGTM (Ren et al. 2023), DBKD (Zhou et al. 2023), and AD-KD (Wu et al. 2023). For comparison with tensor-based approaches, we also include OPDF (Zhan et al. 2024). Note that DBKD does not report results on STS-B, as it focuses on logit estimation from decision distributions.

Methods	RTE Acc.	MRPC F1/Acc.	STS-B Corr.	CoLA Mcc.	SST-2 F1/Acc.	QNLI Acc.	QQP F1/Acc.	MNLI Acc.	Avg.	Train params (M)	Inference params (M)
BERT-base											
	70.5	88.3/82.8	86.4	54.7	91.7	91.7	88.0/91.0	84.2/83.3	81.7	110	110
BERT-of-Theseus											
None	66.1	87.6/81.1	85.7	37.2	90.1	88.0	86.7/89.9	81.7	77.7	66	66
+SVD	65.7	87.2/80.7	85.0	40.0	90.6	88.2	85.8/89.1	80.5	77.7	66	66
+SVD (OP)	65.5	87.0/80.6	85.2	39.1	91.0	87.9	86.2/89.7	79.6	77.5	90	66
+OPDF	66.3	87.1/80.9	**86.1**	41.2	91.4	89.0	86.8/90.2	81.6	78.5	160	66
+DSKD (Ours)	**67.6**	**87.7/81.6**	85.6	**43.0**	**91.5**	**89.2**	**87.2/90.5**	**82.8**	**79.2**	66	66
LGTM											
None	64.8	85.5/78.4	85.0	**49.9**	91.7	89.2	88.0/91.1	82.5	79.3	67	67
+SVD	64.9	84.1/75.9	84.8	48.8	90.6	88.3	86.9/89.0	82.6	78.5	67	67
+SVD (OP)	65.2	85.8/80.9	83.3	42.5	91.2	88.4	87.0/89.5	81.5	77.9	91	67
+OPDF	65.5	**86.5/81.2**	**85.4**	48.6	91.5	**89.3**	87.5/89.9	82.9	79.5	163	67
+DSKD (Ours)	**66.2**	85.9/78.9	85.2	49.4	**92.2**	89.1	**88.6/91.7**	**83.6**	**79.8**	67	67
DBKD											
None	63.3	82.6/74.8	-	25.6	88.9	86.7	**86.3/89.8**	76.5	72.5	53	53
+SVD	63.6	83.1/75.0	-	28.1	88.7	87.1	85.0/87.9	76.8	72.8	53	53
+SVD (OP)	64.2	84.1/76.5	-	26.9	89.1	86.5	85.1/88.1	75.5	72.7	69	53
+OPDF	**65.1**	**84.8/77.0**	-	27.7	**89.8**	**87.7**	86.1/89.2	77.0	73.7	83	53
+DSKD (Ours)	64.7	83.7/75.8	-	**28.8**	89.7	87.5	86.2/89.7	**77.6**	**73.7**	53	53
AD-KD											
None	67.5	84.7/79.0	88.4	49.7	91.1	90.8	86.0/89.5	81.0	79.8	67	67
+SVD	67.0	84.5/78.6	88.1	49.8	90.5	89.5	86.5/89.9	81.2	79.5	67	67
+SVD (OP)	68.1	84.7/78.9	88.4	50.1	89.9	90.0	86.1/89.6	81.2	79.7	91	67
+OPDF	**68.5**	85.0/79.1	88.5	50.9	91.3	**91.0**	86.9/90.1	**82.0**	80.3	182	67
+DSKD (Ours)	66.9	**85.7/80.1**	**88.9**	**51.5**	**91.5**	90.5	**87.1/90.7**	81.9	**80.4**	67	67

Table 1: Overall results on the GLUE benchmark (in percent). Results are averaged over 3 runs. The best results are shown in bold.

For the CIFAR-100 dataset, We evaluate on ReviewKD (Chen et al. 2021), DKD (Zhao et al. 2022), and Exp-KD (Sun et al. 2025), integrating our method to validate its generalizability across vision-based distillation frameworks.

Implementation Details

We implement our framework in PyTorch using two NVIDIA RTX A6000 GPUs (48GB each). For the GLUE benchmark, we use BERT-base-uncased as the base model. We report results using the checkpoint with the best validation performance.We compare our method (DSKD) against standard SVD-based distillation and two over-parameterized variants: SVD(OP) and OPDF. Notably, SVD(OP) and OPDF perform pre-training over-parameterization, whereas our method and standard SVD apply decomposition dynamically during training.

For the CIFAR-100 dataset, we evaluate four representative CNN-based distillation frameworks. For convolutional layers, we consider both channel-wise (Zhang et al. 2015) and spatial (Jaderberg, Vedaldi, and Zisserman 2014) SVD strategies. Since OPDF was not originally designed for convolutional layers, we exclude it from comparison and focus on SVD and DSKD.

Hyperparameters: the SVD loss weight $\lambda \in \{0.1, 0.5, 1, 2, 10\}$, and the number of directions $k \in \{16, 32, 64, 128, 256\}$. On CIFAR-100, due to the small intrinsic rank of convolutional kernels, we set k to full rank, the time step for choosing the direction $T_s \in \{1, 10, 50, 100\}$. The training-phase-aware weighting factor is scheduled linearly: $\alpha(t) = t/T$, where t is the current step and T is the total.

Main Experimental Results

For the GLUE benchmark, as shown in Table.1, DSKD consistently outperforms baseline methods across most GLUE tasks. Notably, DSKD surpasses SVD-based over-parameterization methods (e.g., SVD (OP), OPDF) without increasing parameter count, demonstrating superior parameter efficiency. Compared to standard SVD, the performance gain validates the importance of optimization-sensitive direction selection in distillation. These results confirm that DSKD effectively enhances existing distillation frameworks for natural language understanding. Training complexity analysis is provided in the appendix.

For the CIFAR-100 dataset, as shown in Table.2, DSKD achieves state-of-the-art performance on most evaluated architectures. SVD with direction sensitivity consistently outperforms its non-sensitive counterpart, verifying the utility

Methods		ResNet56 ResNet20		ResNet110 ResNet32		ResNet32×4 ResNet8×4		WRN-40-2 WRN-16-2	
		top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
Teacher		72.4	92.5	74.3	92.9	79.4	94.6	75.6	93.6
ReviewKD									
None		**69.8**	**92.1**	73.1	92.5	75.3	93.7	73.2	92.1
spatial-wise	+SVD	68.4	91.4	72.8	92.3	74.7	92.4	73.5	92.3
	+DSKD (Ours)	68.5	91.5	73.1	92.5	74.8	92.5	73.3	92.2
channel-wise	+SVD	69.5	91.8	73.5	92.7	74.8	92.5	73.8	92.5
	+DSKD (Ours)	69.3	91.8	**73.6**	**92.8**	**75.9**	**94.0**	**74.1**	**92.8**
DKD									
None		71.1	92.3	73.3	93.0	75.5	93.7	73.2	92.1
spatial-wise	+SVD	70.1	91.8	72.1	91.9	74.5	92.8	73.4	92.2
	+DSKD (Ours)	70.6	92.1	71.8	91.2	74.8	92.9	72.8	91.8
channel-wise	+SVD	70.3	91.9	72.5	92.1	75.6	93.5	**73.6**	**92.3**
	+DSKD (Ours)	**71.5**	**92.5**	**73.9**	**93.5**	**75.7**	**93.8**	73.4	92.2
Exp-KD									
None		67.8	91.2	68.4	90.5	68.2	90.7	70.1	91.3
spatial-wise	+SVD	67.1	90.6	68.1	90.2	67.8	90.1	69.9	91.0
	+DSKD (Ours)	67.5	91.0	68.5	90.7	68.2	90.5	70.3	91.4
channel-wise	+SVD	68.0	91.3	68.7	90.8	68.4	90.6	70.9	91.6
	+DSKD (Ours)	**68.3**	**91.5**	**69.1**	**91.9**	**68.6**	**90.8**	**70.9**	**91.9**

Table 2: Overall results on CIFAR-100 datasets (in percent). Results are averaged over 3 runs. The best results are shown in bold.

λ	RTE Acc.	MRPC F1/Acc.	STS-B Corr.	CoLA Mcc.	SST-2 F1/Acc.	QNLI Acc.
0	66.1	87.6/81.1	85.7	37.2	90.1	88.0
0.1	66.2	87.6/81.1	85.7	40.7	91.5	89.1
0.5	66.9	87.5/81.4	**85.7**	**43.0**	91.2	**89.2**
1	**67.6**	**87.7/81.6**	85.5	39.0	**91.5**	88.2
2	64.7	87.1/80.6	85.4	37.1	90.8	88.1
10	64.7	86.6/79.7	85.4	34.0	88.9	87.5

Table 3: Ablation study of different λ values on GLUE tasks (in percent).

of direction selection. We further observe that channel-wise decomposition significantly exceeds spatial decomposition in accuracy. We attribute this to the stronger semantic alignment between singular directions and feature channels, enabling sensitivity metrics to identify task-critical subspaces. In contrast, spatial decomposition lacks such interpretability, limiting the effectiveness of directional analysis.

In summary, DSKD improves upon existing distillation methods by adaptively selecting optimization-sensitive directions. It provides a general and effective framework for incorporating direction sensitivity into both transformer and CNN-based models.

Ablation Study

Impact of λ. Table.3 shows the effect of the distillation weight λ, which balances the SVD loss and task-specific objective. Performance is sensitive to λ and varies across tasks. Optimal values yield significant gains—up to +1.5% on RTE and +5.8% on CoLA—while $\lambda = 0$ results in performance drops, confirming the effectiveness of DSKD. Overly large λ degrades performance across most tasks, suggesting that excessive distillation pressure can impede student learning by enforcing suboptimal teacher structures. Overall, $\lambda \in [0.1, 1.0]$ provides a robust trade-off between knowledge transfer and model adaptability.

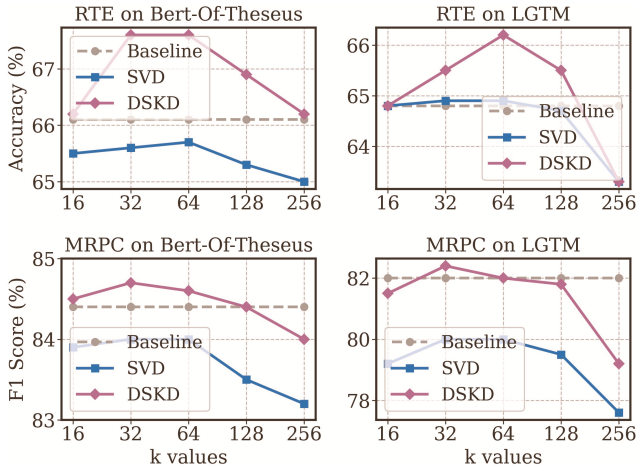


Figure 3: Results on RTE and MRPC with different k (in percent). The results on MPRC are the average of the accuracy rate and the f1 score.

Impact of k. DSKD uses a top-k strategy to focus on loss-sensitive directions. We analyze k's impact using Bert-

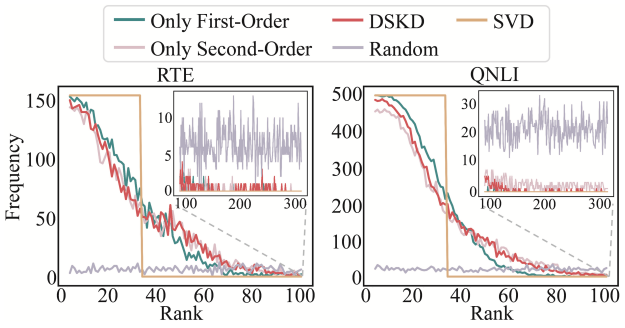


Figure 4: The frequency at which the direction is selected with different selection strategies.

of-Theseus on RTE and MRPC (Figure.3). Both SVD and DSKD show a rise-then-fall trend in performance as k increases. Too small k risks losing critical knowledge; too large k introduces noise and harms distillation. This highlights the importance of balancing information retention and noise suppression. In our experiment, k is set to 32.

Impact of T_s. We analyze the sensitivity to T_s on RTE ($\sim$2.5K examples) and QNLI ($\sim$108K examples). Performance on RTE varies significantly with T_s, whereas on QNLI the effect is much smaller. We attribute this to data efficiency: in low-data regimes, a small T_s may hinder the student's ability to capture the teacher's structural knowledge, while ample data in QNLI mitigate this sensitivity by providing stronger optimization signals.

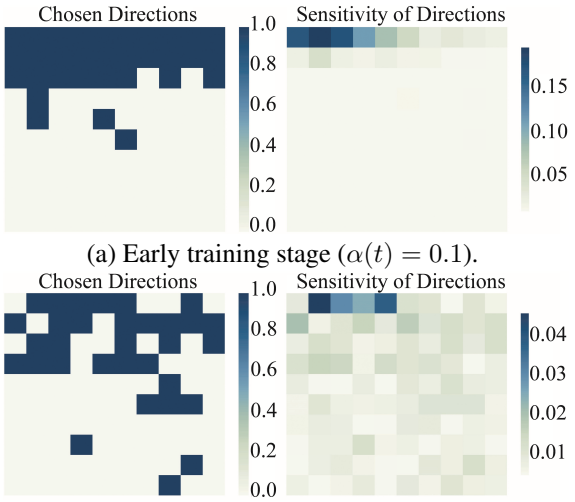
T_s	RTE	QNLI
None	66.1	88.0
1	$64.7_{-1.4}$	$89.0_{+1.0}$
10	$66.2_{+0.1}$	$89.2_{+1.2}$
50	$66.9_{+0.8}$	$89.0_{+1.0}$
100	$\mathbf{67.6_{+1.5}}$	$\mathbf{89.2_{+1.2}}$

Table 4: Results on RTE and QNLI with different T_s (in percent).

Impact of direction selection strategies. We evaluate different direction selection strategies on RTE and QNLI in Table.5. Using only the second-order component performs worse than the first-order alone, indicating that gradient signals provide stronger initial guidance, while combining both yields the best performance, showing that curvature information complements gradient dynamics for more stable alignment. To analyze the selection pattern, we visualize the frequency of the selected direction in a certain layer of the weight matrix (Figure.4). On the smaller RTE set, the influence of first- and second-order components is similar; on QNLI, second-order-only selection overemphasizes small singular-value directions, whereas first-order-only focus on top singular directions may miss fine-grained semantics. Our method dynamically balances the two strategies, enhancing the effect of distillation.

Methods	RTE	QNLI
None	66.1	88.0
random	$63.5_{-2.6}$	$87.0_{-1.0}$
SVD	$65.7_{-0.4}$	$88.2_{+0.2}$
only First-order response	$67.2_{+1.1}$	$88.9_{+0.9}$
only Second-order response	$66.8_{+0.7}$	$88.2_{+0.2}$
DSKD	$\mathbf{67.6_{+1.5}}$	$\mathbf{89.2_{+1.2}}$

Table 5: Results on RTE and QNLI with different direction selection strategies (in percent).



(a) Early training stage ($\alpha(t) = 0.1$).

(b) Late training stage ($\alpha(t) = 0.9$).

Figure 5: Visualization of direction selection at different training stages.

Visualized Analysis on Direction Sensitivity

We visualized the direction sensitivity of the parameter matrix training of a certain layer of the student model in the early ($\alpha(t) = 0.1$) and late ($\alpha(t) = 0.9$) stages (Figure.5) (we presented the first 100 directions and expanded them into matrices. Each grid point represents a direction. The darkest region in the left figure indicates the selected direction; in the right figure, darker areas represent higher sensitivity.). Initially, sensitivity concentrates on large-singular-value directions, it spreads to smaller-singular-value directions and becomes more uniform. This adaptive shift promotes efficient knowledge transfer and mitigates overfitting.

Conclusion

We propose a novel knowledge distillation framework that selects singular directions based on optimization sensitivity rather than static magnitude. Our method introduces a gradient- and Hessian-aware metric to capture the first- and second-order impact of directions on the loss landscape, with a training-phase-aware fusion mechanism for adaptive importance scoring. A direction-weighted distillation loss aligns the student and teacher in the top-k sensitive subspaces. Experiments on GLUE and CIFAR-100 show consistent gains over state-of-the-art baselines.

Acknowledgments

This work was supported by Hubei Science and Technology Major Project (2024BAA007), National Natural Science Foundation of China (U24A20280, 62333007), Hubei Provincial Technology Innovation Program (2025DJA047), and Guangdong Basic and Applied Basic Research Foundation (2025A1515012194).

References

Chen, D.; Mei, J.-P.; Zhang, H.; Wang, C.; Feng, Y.; and Chen, C. 2022. Knowledge Distillation with the Reused Teacher Classifier. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11923–11932. IEEE.

Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021. Distilling Knowledge via Knowledge Review. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5006–5015. IEEE Computer Society.

Dai, T.; Lin, Y.; Guo, H.; Wang, J.; and Zhu, Z. 2025. DCSF-KD: Dynamic Channel-wise Spatial Feature Knowledge Distillation for Object Detection. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, 2627–2635.

Denton, E.; Zaremba, W.; Bruna, J.; LeCun, Y.; and Fergus, R. 2014. Exploiting linear structure within convolutional networks for efficient evaluation. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 1269–1277.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Gao, S.; Hua, T.; Hsu, Y.-C.; Shen, Y.; and Jin, H. 2024. Adaptive rank selections for low-rank approximation of language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 227–241.

Gao, Z.; Han, S.; Zhang, X.; Xu, K.; Zhou, D.; Mao, X.; Dou, Y.; and Wang, H. 2025. Maintaining fairness in logit-based knowledge distillation for class-incremental learning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, 16763–16771.

Gong, L.; Lin, S.; Zhang, B.; Shen, Y.; Li, K.; Qiao, R.; Ren, B.; Li, M.; Yu, Z.; and Ma, L. 2023. Adaptive hierarchy-branch fusion for online knowledge distillation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, 7731–7739.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129(6): 1789–1819.

Guo, Z.; Yan, H.; Li, H.; and Lin, X. 2023. Class Attention Transfer Based Knowledge Distillation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11868–11877. IEEE.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

Horváth, S.; Laskaridis, S.; Rajput, S.; and Wang, H. 2024. Maestro: uncovering low-rank structures via trainable decomposition. In *Proceedings of the 41st International Conference on Machine Learning*, 18859–18881.

Hsu, Y.-C.; Hua, T.; Chang, S.; Lou, Q.; Shen, Y.; and Jin, H. 2022. Language model compression with weighted low-rank factorization. In *International Conference on Learning Representations*.

Hua, T.; Hsu, Y.-C.; Wang, F.; Lou, Q.; Shen, Y.; and Jin, H. 2022. Numerical Optimizations for Weighted Low-rank Estimation on Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1404–1416.

Hua, T.; Li, X.; Gao, S.; Hsu, Y.-C.; Shen, Y.; and Jin, H. 2023. Dynamic Low-rank Estimation for Transformer-based Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9275–9287. Singapore: Association for Computational Linguistics.

Jaderberg, M.; Vedaldi, A.; and Zisserman, A. 2014. Speeding up Convolutional Neural Networks with Low Rank Expansions. In *Proceedings of the British Machine Vision Conference 2014*, 88–1. British Machine Vision Association.

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4163–4174.

Jing Yang, A. B. G. T., Brais Martinez. 2021. Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kunstner, F.; Balles, L.; and Hennig, P. 2019. Limitations of the empirical fisher approximation for natural gradient descent. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 4156–4167.

Lee, S. H.; Kim, D. H.; and Song, B. C. 2018. Self-supervised knowledge distillation using singular value decomposition. In *Proceedings of the 15th European conference on computer vision (ECCV)*, 339–354. Springer Verlag.

Li, C.; Lin, M.; Ding, Z.; Lin, N.; Zhuang, Y.; Huang, Y.; Ding, X.; and Cao, L. 2022. Knowledge condensation distillation. In *Proceedings of the 17th European conference on computer vision (ECCV)*, 19–35. Springer.

Li, J.; Lai, Y.; Wang, R.; Shui, C.; Sahoo, S.; Ling, C. X.; Yang, S.; Wang, B.; Gagné, C.; and Zhou, F. 2024. Hessian aware low-rank perturbation for order-robust continual learning. *IEEE Transactions on Knowledge and Data Engineering*, 36(11): 6385–6396.

Li, L.; and Zhe, J. 2022. Shadow knowledge distillation: bridging offline and online knowledge transfer. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 635–649.

Li, T.; Liu, L.; Liu, K.; Wang, X.; Zhou, B.; Yang, H.; and Lu, K. 2025. Adaptive Dual Guidance Knowledge Distillation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, 18457–18465.

Li, Z.; Li, X.; Yang, L.; Zhao, B.; Song, R.; Luo, L.; Li, J.; and Yang, J. 2023. Curriculum temperature for knowledge distillation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, 1504–1512.

Liang, C.; Zuo, S.; Zhang, Q.; He, P.; Chen, W.; and Zhao, T. 2023. Less is more: task-aware layer-wise distillation for language model compression. In *Proceedings of the 40th International Conference on Machine Learning*, 20852–20867.

Liu, K.; Zhang, Y.; Cheng, N.; Li, Z.; Wang, S.; and Xiao, J. 2024a. GRASP: Replace Redundant Layers with Adaptive Singular Parameters for Efficient Model Compression. arXiv:2501.00339.

Liu, K.; Zhang, Y.; Cheng, N.; Li, Z.; Wang, S.; and Xiao, J. 2024b. Rethinking Layer Removal: Preserving Critical Components with Task-Aware Singular Value Decomposition. arXiv:2501.00339.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Oymak, S.; Fabian, Z.; Li, M.; and Soltanolkotabi, M. 2019. Generalization Guarantees for Neural Networks via Harnessing the Low-Rank Structure of the Jacobian. arXiv:1906.05392.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational Knowledge Distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3962–3971. IEEE.

Ren, Y.; Zhong, Z.; Shi, X.; Zhu, Y.; Yuan, C.; and Li, M. 2023. Tailoring Instructions to Student's Learning Levels Boosts Knowledge Distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1990–2006.

Ryu, J. J.; Xu, X.; Erol, H. M.; Bu, Y.; Zheng, L.; and Wornell, G. W. 2024. Operator SVD with neural networks via nested low-rank approximation. In *Proceedings of the 41st International Conference on Machine Learning*, 42870–42905.

Staats, M.; Thamm, M.; and Rosenow, B. 2024. Small Singular Values Matter: A Random Matrix Analysis of Transformer Models. arXiv:2410.17770.

Sun, S.; Ren, W.; Li, J.; Wang, R.; and Cao, X. 2024. Logit Standardization in Knowledge Distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15731–15740. IEEE.

Sun, T.; Chen, H.; Hu, G.; and Zhao, C. 2025. Explainability-based knowledge distillation. *Pattern Recognition*, 159: 111095.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Representation Distillation. In *International Conference on Learning Representations*.

Tung, F.; and Mori, G. 2019. Similarity-Preserving Knowledge Distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1365–1374. IEEE.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*.

Wang, S.; Li, B. Z.; Khabsa, M.; Fang, H.; and Ma, H. 2020. Linformer: Self-Attention with Linear Complexity. arXiv:2006.04768.

Wu, S.; Chen, H.; Quan, X.; Wang, Q.; and Wang, R. 2023. AD-KD: Attribution-Driven Knowledge Distillation for Language Model Compression. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 8449–8465.

Xu, C.; Zhou, W.; Ge, T.; Wei, F.; and Zhou, M. 2020. BERT-of-Theseus: Compressing BERT by Progressive Module Replacing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7859–7869.

Yang, H.; Tang, M.; Wen, W.; Yan, F.; Hu, D.; Li, A.; Li, H.; and Chen, Y. 2020. Learning Low-rank Deep Neural Networks via Singular Vector Orthogonality Regularization and Singular Value Sparsification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2899–2908. IEEE.

Yang, S.; Yang, J.; Zhou, M.; Huang, Z.; Zheng, W.-S.; Yang, X.; and Ren, J. 2024. Learning from human educational wisdom: A student-centered knowledge distillation method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6): 4188–4205.

Zagoruyko, S.; and Komodakis, N. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *International Conference on Learning Representations*.

Zhan, Y.-L.; Lu, Z.-Y.; Sun, H.; and Gao, Z.-F. 2024. Over-parameterized student model via tensor decomposition boosted knowledge distillation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 69445–69470.

Zhang, J.; Gao, Y.; Zhou, M.; Liu, R.; Cheng, X.; Nikoli, S. V.; and Chen, S. 2025. SVD-KD: SVD-based hidden layer feature extraction for Knowledge distillation. *Pattern Recognition*, 167: 111721.

Zhang, X.; Zou, J.; He, K.; and Sun, J. 2015. Accelerating very deep convolutional networks for classification and detection. *IEEE transactions on pattern analysis and machine intelligence*, 38(10): 1943–1955.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled Knowledge Distillation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11943–11952. IEEE Computer Society.

Zhou, Q.; Yang, Z.; Li, P.; and Liu, Y. 2023. Bridging the Gap between Decision and Logits in Decision-based Knowledge Distillation for Pre-trained Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 13234–13248.