APL Machine Learning

RESEARCH ARTICLE | FEBRUARY 06 2026

Efficient knowledge distillation via salient feature masking

Assel Kembay ; Skye Gunasekaran ; Rui-Jie Zhu ; Yu Zhang ; Jason K. Eshraghian

Check for updates

APL Mach. Learn. 4, 016104 (2026)
https://doi.org/10.1063/5.0312051

View Online Export Citation

Articles You May Be Interested In

Learning to see high-density random images long-term transmitted in multimode fiber
AIP Advances (April 2024)

An auxiliary detection model for gastrointestinal nursing based on deep learning
AIP Advances (December 2025)

Comparison of semantic segmentation capabilities of pre-trained models for co-salient object detection
AIP Conf. Proc. (July 2023)

08 May 2026 01:37:23

AIP Publishing

Efficient knowledge distillation via salient feature masking

Cite as: APL Mach. Learn. 4, 016104 (2026); doi: 10.1063/5.0312051
Submitted: 13 November 2025 • Accepted: 22 January 2026 •
Published Online: 6 February 2026

View Online Export Citation CrossMark

Assel Kembay, Skye Gunasekaran, Rui-Jie Zhu, Yu Zhang, and Jason K. Eshraghian[a]

AFFILIATIONS

Department of Electrical and Computer Engineering, University of California, Santa Cruz, California 95064, USA

[a]Author to whom correspondence should be addressed: jsn@ucsc.edu

ABSTRACT

Traditional Knowledge Distillation (KD) transfers all outputs from a teacher model to a student model, often introducing knowledge redundancy. This redundancy dilutes critical information, leading to degraded student model performance. To address this, we propose Salient Feature Masking for Knowledge Distillation (SFKD), a lightweight enhancement that masks out less informative components and selectively distills only the top-K activations. SFKD is a drop-in modification applicable to both logit-based and feature-based KD, incurs negligible overhead, and sharpens the student's learning signal. Empirically, SFKD yields consistent gains over strong KD baselines across architectures (ConvNeXt, ViT) and datasets (CIFAR-100: up to +2.43 pp; CUB-200: up to +6.39 pp; ImageNet-1K: up to +3.57 pp). We also provide intuition from the information bottleneck perspective to motivate why filtering out less salient teacher signals benefits the student. Overall, SFKD is a simple, empirically validated method for training student models that are both leaner and more accurate.

© 2026 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1063/5.0312051

08 May 2026 01:37:23

I. INTRODUCTION

While deep neural networks continue to grow in depth, width, and computational demands, the devices that ultimately rely on these algorithms—mobile phones, autonomous drones, and battery-constrained sensors—operate under tight budgets with respect to memory, energy, and latency. *Knowledge distillation (KD)* addresses this gap by transferring the behavior of a high-capacity *teacher* network to a compact *student*. Conventional pipelines, however, relay the full spectrum of teacher signals: the entire logit vector, intermediate features, and attention maps.[1–4] However, such indiscriminate transfer overwhelms the student model with peripheral or even misleading activations, thereby misguiding its limited capacity and hindering generalization.[5]

We reinterpret distillation through the lens of the *Information Bottleneck (IB)* principle.[6] Each teacher activation constitutes a noisy channel between the input–label pair (X, Y) and a representation F. The IB objective seeks the most concise F that maximizes $I(F;\ Y)$ while suppressing redundant information $I(X;\ F)$. From this perspective, only a subset of the teacher's knowledge is worth transmitting.

Guided by the IB principle, we derive **S**alient **F**eature masking for **K**nowledge **D**istillation (**SFKD**), a unified top-K masking rule that filters teacher signals before they reach the student. Viewing each teacher activation as a noisy communication channel, SFKD ranks logit entries, feature map channels, and attention coefficients by a lightweight mutual information proxy and retains only the K most informative elements. By discarding poor cues, the method suppresses transfer bias and compels the student to focus on the evidence most predictive of Y, thereby improving accuracy, robustness, and interpretability at negligible computational cost. Our contributions are as follows:

1. **Unified saliency mask for distillation.** We introduce SFKD, a single top-K masking rule that selects the most informative logits, feature-map values, and attention coefficients, and distills only these signals from teacher to student.
2. **Information-bottleneck motivation and empirical validation.** Motivated by the information bottleneck principle, we propose a simple top-K masking strategy that retains only the most salient activations during knowledge distillation. Through extensive information-plane analysis, we empirically demonstrate that this selective transfer approach helps students achieve better trade-offs between compression [lower I(X; F)] and task performance [higher I(F; Y)] compared to transferring all teacher knowledge indiscriminately.

© Author(s) 2026

3. **SFKD drops straight into existing KD pipelines.** We find that SFKD consistently raises top-1 accuracy across CIFAR-100, CUB200, and ImageNet-1K setups, while adding negligible computational overhead.

II. RELATED WORK

Knowledge Distillation (KD) variants generally fall into three categories based on the type of knowledge transferred: logits,[7–11] features,[1,3,4,12–16] and attention.[2,17] Vanilla KD[7] transfers class predictions from the teacher's output layer to guide the student's training. In contrast, feature distillation extracts knowledge from intermediate layers; for example, FitNet[1] aligns feature maps between specific teacher–student layers. Attention-based methods[2] use attention maps derived from feature representations for comprehensive knowledge transfer across layers. Subsequent studies explore applications of KD in semantic segmentation,[18,19] object detection,[20,21] and student architecture searches.[22]

Selective and Rank-based Knowledge Distillation. While traditional KD transfers all teacher outputs indiscriminately, recent studies explore selective transfer strategies. Decoupled Knowledge Distillation (DKD)[10] decomposes logit-based KD into target class knowledge and non-target class knowledge, selectively weighting these components. Rank-based distillation methods[4,23,24] emphasize preserving the relative ordering of predictions rather than absolute values. However, these approaches still transfer information from all classes or features. In contrast, SFKD explicitly filters out low-magnitude activations, creating a true information bottleneck.

Information Bottleneck (IB) is a principle introduced by Tishby, Pereira, and Bialek,[25] which aims to extract the most relevant information from an input. The IB method defines a trade-off between compressing the input representation and preserving information about the target variable. The IB framework was extended to deep learning,[26] proposing that deep neural networks (DNNs) implicitly optimize this trade-off during training. Shwartz-Ziv and Tishby[27] applied the IB principle to analyze the training dynamics of DNNs, showing that the learning process can be viewed as a progression from fitting the data to compressing irrelevant information, thereby enhancing generalization.

Pogodin and Latham[28] further advanced this field by proposing learning rules based on the IB principle, achieving performance comparable to backpropagation in image classification tasks. More recently, Wang *et al.*[29] found that an intermediate model, often at an optimal training checkpoint, can serve as a more effective teacher than a fully converged model, despite its lower accuracy. In contrast, our study uniquely applies the IB principle to interpret the KD process. While Goldfeld *et al.*[30] analyzed mutual information compression in representation learning, we are the first to use the IB framework to specifically examine information flow during distillation, offering novel insights into the underlying dynamics of the process.

III. METHODS

Let $\mathbf{a} \in \mathbb{R}^N$ denote a teacher activation vector (e.g., the class logit vector, a flattened feature map, or a flattened attention tensor). Our objective is to retain only the K most informative elements, where element magnitude serves as a proxy for information content, following the intuition that larger activation values contribute more significantly to the final prediction. We define the top-K index set as

$$I_{\text{Top}-K} = \{i \in \{1, \ldots, N\} \mid \mathbf{a}_i \text{ ranks among the } K \text{ largest in } \mathbf{a}\}. \quad (1)$$

From $I_{\text{Top-}K}$, we construct the binary mask $\mathbf{M} \in \{0,1\}^N$ with components

$$M_i = \begin{cases} 1, & \text{if } i \in I_{\text{Top}-\text{K}}, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Applying the mask to $\mathbf{a}$ via the element-wise (Hadamard) product $\odot$ yields the top-K masked activation,

$$\mathbf{a}_K := \mathbf{M} \odot \mathbf{a}. \quad (3)$$

This operation preserves the K most salient entries ($a_{K,i} = a_i$ for $i \in I_{\text{Top-}K}$) while zeroing out all others ($a_{K,i} = 0$ otherwise), effectively creating an information bottleneck that filters low-importance signals before knowledge transfer. For notational consistency, when the activation vector is denoted $\mathbf{F}$, we write $\mathbf{F}_K = \mathbf{M} \odot \mathbf{F}$.

Our findings have broad applicability, covering a wide range of distillation techniques, as illustrated in Fig. 1. We focus on standard methods representing three main families of distillation approaches: *output-based*,[7] *feature-based*,[1] and *attention-based*.[2] The objectives of these methods are combined with the cross-entropy loss $L_{CLS}(z_s, y) := -\sum_{j=1}^{c} y_j \log \sigma_j(z_s)$, where y is the ground-truth one-hot label vector, z_s is the student's logit output, $\sigma_j(z) = e^{z_j}/\sum_i e^{z_i}$ is the softmax function, and c is the number of classes.

(1) **Output-based:** Salient feature masking operates on the logit space by applying the top-K mask to the teacher's logit vector, retaining only the K highest-magnitude logits and zeroing out the rest. The knowledge transfer is then performed through KL-divergence minimization between the masked teacher distribution and student predictions,

$$L_{KL}(z_s, z_{t^K}) := -\tau^2 \sum_{j=1}^{c} \sigma_j\left(\frac{z_{t^K}}{\tau}\right) \log \sigma_j\left(\frac{z_s}{\tau}\right), \quad (4)$$

where $z_{t^K} = \mathbf{M} \odot z_t$ denotes the teacher's logits after top-K masking; τ is a temperature scaling parameter that controls the softness of the probability distributions; and the overall loss function is $\gamma L_{CLS} + \alpha L_{KL}$ with balancing hyperparameters γ and α.

(2) **Feature-based:** The student's intermediate features $F_s^{(l)}$ are trained to mimic only these K salient components of the teacher's masked features $F_{t^K}^{(l)}$. The student's features are first projected via a transformation function r to match the spatial dimensions or number of channels of the teacher's features (e.g., a 1×1 convolutional layer or linear projection to align channel dimensions between F_s and F_t). Their similarity is then optimized by minimizing the mean squared error,

$$L_{\text{Hint}}\left(F_s^{(l)}, F_{t^K}^{(l)}\right) = \frac{1}{2}\left\| F_{t^K}^{(l)} - r\left(F_s^{(l)}\right)\right\|_2^2, \quad (5)$$

where $F_{t^K}^{(l)} = \mathbf{M}^{(l)} \odot F_t^{(l)}$ represents the masked teacher features. The total loss is $\gamma L_{CLS} + \beta L_{\text{Hint}}$, where γ and β are

08 May 2026 01:37:23

© Author(s) 2026

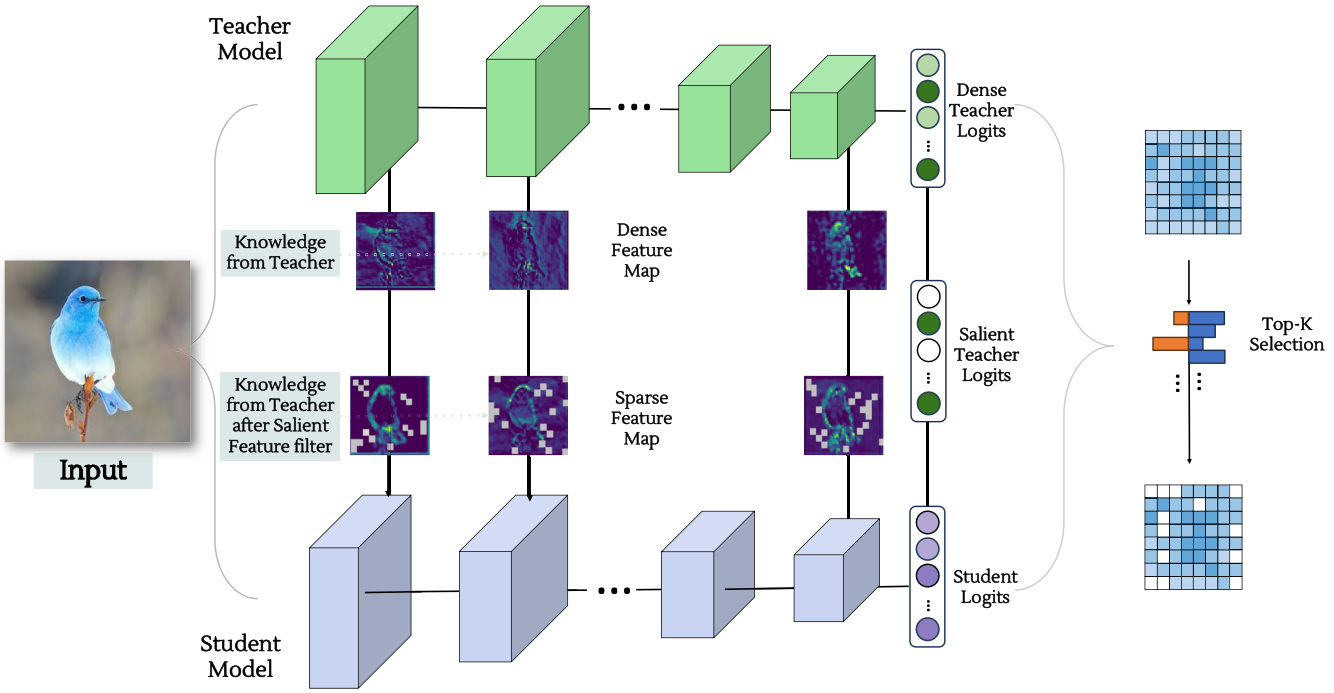

FIG. 1. Concept of the proposed SFKD. SFKD distinctively concentrates on (1) distilling critical classification knowledge, (2) transferring essential information from intermediate layers, and (3) refining attention mechanisms for knowledge distillation.

balancing hyperparameters. The term "Hint" represents all feature-based KD methods following Romero *et al.*[1]

(3) **Attention-based:** Let I be the set of indices representing the teacher–student activation layer pairs where attention maps are transferred. The total attention transfer loss is then defined as:

$$L_{AT} = L_{CLS} + \frac{\beta}{2}\sum_{j\in I}\left\| \frac{Q_s^j}{\|Q_s^j\|_2} - \frac{Q_{t^K}^j}{\|Q_{t^K}^j\|_2} \right\|_p, \tag{6}$$

where $Q_s^j = \mathrm{vec}(\phi(A_s^j))$ and $Q_{t^K}^j = \mathrm{vec}(\phi(\mathbf{M}^j \odot A_t^j))$ are, respectively, the j-th pair of student and top-K masked teacher attention maps in vectorized form. The mapping function $\phi : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{H\times W}$ transforms a 3D activation tensor A into a spatial attention map by aggregating across channels (e.g., via summation or ℓ_2 norm). $\beta > 0$ is a balancing hyperparameter, and $\|\cdot\|_p$ denotes the ℓ_p norm (typically $p = 2$).

IV. AN INFORMATION-THEORETIC PERSPECTIVE ON SFKD

In this section, we use the well-established Information Bottleneck theory as a conceptual lens to motivate and analyze SFKD. This perspective provides a clear intuition for why selectively distilling information, rather than transferring the teacher's entire knowledge base, can lead to more efficient and effective student models.

A. The IB principle

Let X and Y denote the input and label random variables, and let F be an intermediate representation generated by a parameterized encoder $p_\phi(F|X)$. The classical IB objective[26,27] seeks a trade-off between compressing the input and preserving predictive information about the label,

$$\min_{\phi}\ I(X;F) - \zeta\, I(F;Y), \tag{7}$$

where $\zeta > 0$ controls the trade-off. We acknowledge that for deterministic neural networks, the mutual information $I(X;\ F)$ is formally infinite. Following established practices in IB analysis of deep learning,[27,31,32] we employ practical MI estimators based on well-established lower bounds. These estimators provide reliable *relative* measures for comparing information flow across different training configurations, which is sufficient for our qualitative analysis of distillation dynamics.

Viewing knowledge distillation through the IB lens, we posit that the teacher's complete output acts as a noisy communication channel. Not all teacher signals are equally informative—some may even mislead the student by emphasizing spurious correlations or overfitted patterns. SFKD creates an information bottleneck that filters the teacher's knowledge before transfer, ensuring the student focuses its limited capacity on the most predictive and generalizable information.

The IB principle inspires several testable hypotheses about how SFKD should affect the student's learning dynamics, which we

08 May 2026 01:37:23

© Author(s) 2026

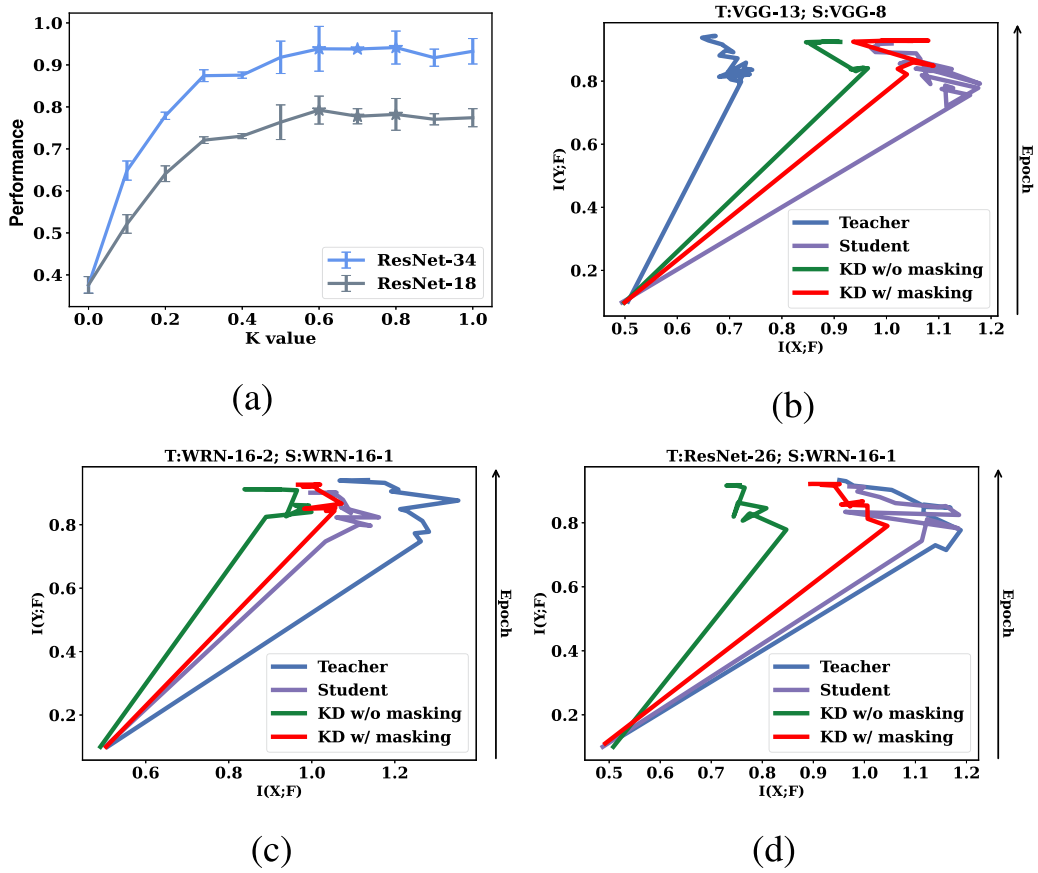

FIG. 2. (a) Optimal K selection via reconstruction-based analysis. The horizontal axis shows the masking ratio ($K = 1.0$ means no masking, and $K = 0.6$ means retaining the top 60%). Lower BCE reconstruction loss indicates better information retention. [(b)–(d)] Information plane trajectories for different teacher–student pairs on CIFAR-10. Each point represents a measurement epoch (every 20 epochs over 240 total). Arrows indicate temporal progression. MI estimates are computed on the test set: $I(X; F)$ via decoder reconstruction quality and $I(Y; F)$ via classification accuracy. SFKD (red) achieves superior trajectories compared to full KD (green), validating hypotheses H1–H3.

can visualize on the "information plane" [$I(X; F)$ vs $I(F; Y)$], as follows:

- **H1: Enhanced compression-prediction trade-off**—By filtering out noisy or redundant teacher signals, SFKD should guide students toward a better position on the information plane, achieving higher $I(F_s; Y)$ (label informativeness) for a given level of $I(X; F_s)$ (input retention).
- **H2: Accelerated learning dynamics**—By receiving a cleaner, more concentrated learning signal, students trained with SFKD should exhibit faster convergence in $I(F_s; Y)$ during early training epochs compared to full KD, as they are not distracted by less informative teacher outputs.
- **H3: Optimal masking level exists**—Transferring too little information (very small K) starves the student, while transferring everything ($K = N$, no masking) introduces noise. An intermediate masking level should maximize student performance, reflecting the optimal information bottleneck capacity.

Our empirical results (Fig. 2) strongly support these hypotheses. SFKD consistently achieves superior information plane trajectories (H1), accelerates label information acquisition (H2), and exhibits clear optimal K values (H3), validating the IB-motivated design of selective knowledge transfer.

V. EXPERIMENTS

We demonstrate that our SFKD approach is method-agnostic by testing it across various existing distillation methods. In addition, we show its two applications: (i) selective knowledge sharing in multi-teacher knowledge distillation and (ii) salient feature masking in data-free knowledge distillation. Furthermore, we perform ablation studies, implementing both as a standalone approach and in conjunction with the KD loss.

A. Results on CIFAR-100

Table I demonstrates the consistent effectiveness of SFKD across diverse teacher–student configurations: both homogeneous

08 May 2026 01:37:23

© Author(s) 2026

TABLE I. Comprehensive performance comparison on CIFAR-100. The **bold** values indicate the best, and the underlined values are the second-best value.

Method	Same architecture style				Different architecture style		
T/S pair	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	ResNet-32 × 4 ResNet-8 × 4	VGG-13 VGG-8	VGG-13 MobileNetV2	ResNet-32 × 4 ShuffleNetV2	WRN-40-2 ShuffleNetV1
Teacher	75.61	75.61	79.42	74.64	74.64	79.42	75.61
Student	73.26	71.98	73.09	70.36	64.60	71.82	70.50
CAT-KD[17] (CVPR'23)	75.60	74.82	76.91	74.65	69.13	78.41	77.35
ReviewKD[4] (CVPR'21)	76.12	75.09	75.63	74.84	70.37	77.78	77.14
DIST[34] (NeurIPS'22)	N/A	74.73	76.31	N/A	N/A	77.35	N/A
KD-zero[35] (NeurIPS'23)	76.42	N/A	77.85	75.26	70.42	77.45	<u>77.52</u>
Auto-KD[36] (ICCV'23)	76.86	N/A	77.61	<u>75.36</u>	<u>70.58</u>	77.52	77.46
RLD[24] (ICCV'25)	76.02	74.88	76.64	74.93	69.97	77.56	N/A
LS (MLKD+LS)[23] (CVPR'24)	<u>76.95</u>	75.56	<u>78.28</u>	75.22	**70.94**	<u>78.76</u>	N/A
DKD[10] (CVPR'22)	76.24	74.81	76.32	74.68	69.71	77.07	76.70
DKD + SFKD	76.51	74.96	76.68	74.82	69.94	77.34	76.95
SimKD[37] (CVPR'22)	76.23	75.56	78.08	74.93	68.95	78.39	N/A
SimKD + SFKD	76.53	**75.87**	**78.53**	75.23	70.38	78.48	**77.64**
MLKD[11] (CVPR'23)	76.63	75.35	77.08	75.18	70.57	78.44	77.44
MLKD + SFKD	**77.01**	<u>75.72</u>	78.06	**75.60**	<u>70.58</u>	**79.16**	77.50

pairs (same architecture family, e.g., WRN-40-2 → WRN-16-2) and heterogeneous pairs (different architectures, e.g., ResNet-32 × 4 → ShuffleNetV2). SFKD yields improvements of +0.3 to +1.7 percentage points across all tested baselines (DKD, SimKD, MLKD), demonstrating its method-agnostic nature. We extend our evaluation to more advanced network architectures, including ConvNeXt and Vision Transformers: *ViT-based teachers* distilled to both *CNN-based* and *ViT-based students*, with results presented in Table II. This broader testing scope validates SFKD's versatility across fundamentally different architectural paradigms—from attention-based transformers (Swin-T) to CNNs (ResNet-18), and between modern architectures (ConvNeXt → Swin-P[33]), with gains up to +2.43 pp.

B. Results on ImageNet

Table III reports Top-1 and Top-5 accuracies on the large-scale ImageNet-1K dataset (1000 classes). SFKD consistently improves upon strong baselines across different capacity gaps: ResNet-34 → ResNet-18 (+1.16 pp top-1) and ResNet-50 → MobileNet (+3.57 pp top-1). The consistent gains in both top-1 and top-5 metrics demonstrate SFKD's scalability to large-scale, high-diversity datasets where filtering noisy teacher signals becomes increasingly critical.

C. Results on CUB200

Table IV evaluates SFKD on the fine-grained CUB-200-2011 bird classification task[38] (11 788 images, 200 bird species), which requires discriminating between subtle inter-class variations (e.g., different warbler species). SFKD achieves substantial improvements across all configurations, with the largest gain of +6.39 pp (VGG-13 → MobileNetV2). This suggests that selective transfer is particularly beneficial for fine-grained tasks where salient discriminative features (e.g., beak shape and plumage patterns) must be emphasized over background or irrelevant texture information.

TABLE II. SFKD with heterogeneous architectures on CIFAR-100: *ViT-based teachers* distilled to both *CNN-based* and *ViT-based students*. Boldface denotes results obtained by augmenting baseline knowledge distillation method with SFKD.

ViT-based Teachers	T. S.	Swin-T ResNet-18	ViT-S ResNet-18	Mixer-B/16 ResNet-18	ConvNeXt-T Swin-P
	Teacher acc	89.26	92.43	87.62	88.41
	Student acc	74.01	74.01	74.01	72.63
Logit-based	DIST[34]	77.75	76.49	76.36	76.41
	KD[7]	78.74	77.26	77.79	76.44
	KD + SFKD	**80.62**$_{+1.88}$	**78.90**$_{+1.64}$	**79.18**$_{+1.39}$	**78.87**$_{+2.43}$

08 May 2026 01:37:23

© Author(s) 2026

TABLE III. Top-1 and Top-5 accuracy (%) on ImageNet validation. Boldface denotes the best result in each comparison between baseline methods and their SFKD-enhanced variants.

Teacher/Student	ResNet-34/ResNet-18		ResNet-50/MobileNet	
Accuracy	Top-1	Top-5	Top-1	Top-5
Teacher	73.31	91.42	76.16	92.86
Student	69.75	89.07	68.87	88.76
ReviewKD[4]	71.61	90.51	72.56	91.00
SimKD[37]	71.59	90.48	72.25	90.86
CAT-KD[17]	71.26	90.45	72.24	91.13
AT[2]	70.69	**90.01**	69.56	89.33
AT+SFKD	**70.84**$_{+0.15}$	89.91	**70.88**$_{+1.32}$	**90.00**$_{+0.67}$
KD[7]	70.66	89.88	68.58	88.98
KD+SFKD	**71.82**$_{+1.16}$	**90.41**$_{+0.53}$	**72.15**$_{+3.57}$	**90.52**$_{+1.54}$
DKD[10]	71.70	90.41	72.05	91.05
DKD+SFKD	**72.10**$_{+0.4}$	**90.70**$_{+0.29}$	**72.95**$_{+0.9}$	**91.30**$_{+0.25}$

TABLE IV. Performance on the CUB200 dataset was evaluated across three teacher–student configurations: (1) identical structure but different sizes, (2) different architectures with equivalent depth, and (3) completely different networks in both architecture and depth. Boldface denotes the best result in each comparison between baseline methods and their SFKD-enhanced variants.

Teacher	ResNet-32 × 4	ResNet-32 × 4	VGG-13	VGG-13	ResNet-50
Acc	66.17	66.17	70.19	70.19	60.01
Student	MobileNetV2	ShuffleNetV1	MobileNetV2	VGG-8	ShuffleNetV1
Acc	40.23	37.28	40.23	46.32	37.28
SP[15]	48.49	61.83	44.28	54.78	55.31
CRD[3]	57.45	62.28	56.45	66.10	57.45
SemCKD[39]	56.89	63.78	68.23	66.54	57.20
ReviewKD[4]	· · ·	64.12	58.66	67.10	· · ·
KD[7]	56.09	61.68	53.98	64.18	57.21
KD+SFKD	**61.68**$_{+5.59}$	**65.67**$_{+3.99}$	**60.37**$_{+6.39}$	**65.64**$_{+1.46}$	**61.01**$_{+3.8}$
DKD[10]	59.94	64.51	58.45	67.20	59.21
DKD+SFKD	**62.15**$_{+2.21}$	**67.09**$_{+2.58}$	**61.49**$_{+3.04}$	**68.88**$_{+1.68}$	**63.99**$_{+4.78}$

VI. DISCUSSION

Across three diverse benchmarks—CIFAR-100 (100 classes), ImageNet-1K (1000 classes), and CUB200 (200 fine-grained categories)—SFKD demonstrates consistent and substantial improvements over strong baselines, with gains ranging from +0.3 pp to +6.39 pp. The magnitude of improvement appears correlated with task difficulty and capacity gap: larger gains on fine-grained CUB200 (+6.39 pp) vs standard CIFAR-100 (+0.3–1.7 pp), suggesting that selective transfer is most beneficial when discriminative information is subtle or when the teacher–student capacity mismatch is large.

The breadth of tested configurations—15+ teacher–student pairs spanning homogeneous (CNN–CNN), heterogeneous (ViT-CNN), and modern architectures (ConvNeXt-Swin)—establishes SFKD's architectural agnosticism. Critically, SFKD enhances *all* tested baseline methods (KD, DKD, SimKD, MLKD, AT, etc.) without requiring method-specific modifications, confirming its drop-in compatibility.

These performance improvements are consistent with our IB-theoretic analysis: by selectively retaining only the top-K most informative teacher activations, SFKD reduces the signal-to-noise ratio in knowledge transfer, enabling students to converge faster to better information plane positions [higher $I(F;\ Y)$ for given $I(X;\ F)$]. The visualization analyses (Sec. VII C) further corroborate this: SFKD-trained students exhibit more compact, discriminative feature clusters (t-SNE) and more focused attention to target objects (CAMs), indicating improved representation quality beyond mere accuracy gains.

08 May 2026 01:37:23

© Author(s) 2026

The empirical evidence provided supports SFKD as a lightweight (negligible computational overhead), theoretically grounded enhancement to a wide range of KD frameworks, offering both practical performance gains and conceptual clarity on the role of selective knowledge transfer.

A. Limitations and future work

While SFKD demonstrates consistent improvements, several aspects warrant further investigation. First, we use magnitude-based top-K selection as a computationally efficient proxy for informativeness; future work could compare this against random masking controls to isolate sparsity effects and explore soft weighting schemes (e.g., temperature-annealed top-K weights). Second, we focus on accuracy metrics; future studies should evaluate calibration quality (ECE, NLL, and Brier scores) to verify that masking does not degrade uncertainty estimates. We believe these directions will further strengthen the theoretical foundations and practical applicability of selective knowledge transfer.

VII. ADVANCED APPLICATION OF SFKD

Beyond the standard single-teacher distillation framework, we demonstrate that SFKD's core principle provides significant advantages in more complex scenarios. In this section, we explore two such advanced applications: (1) enhancing knowledge transfer from an ensemble of models in multi-teacher knowledge distillation and (2) improving student performance in data-free knowledge distillation. Finally, we provide a series of visualizations that offer qualitative insights into *how* SFKD achieves its performance gains by improving the student's feature representations and focus.

A. Application 1: Selective knowledge sharing in multi-teacher knowledge distillation

In multi-teacher distillation, conventional methods[40–43] that average teacher outputs risk diluting specialized knowledge. SFKD avoids this pitfall by selectively distilling only the most salient signals from the teacher ensemble. This makes it uniquely suited for multi-teacher contexts, a claim supported by its superior accuracy in our experiments (Table V). We demonstrate this by applying SFKD to the AEKD framework,[40] using a Tri-ResNet-32 × 4 ensemble to teach both VGG-8 and ShuffleNetV2 students. Across all configurations, SFKD consistently improves the AEKD baseline by effectively channeling the most pertinent insights from the multiple experts.

B. Application 2: Salient feature masking in data-free knowledge distillation

Data-Free Knowledge Distillation (DFKD) relies on synthetic data, making it critical to filter out noise and artifacts. SFKD's methodology is particularly advantageous here, as it preserves the integrity of the distilled knowledge by focusing only on highly informative features. This targeted knowledge transfer helps the student model generalize better to real-world data. To validate this, we synthesized 100K images via DeepInversion (DI)[45] from CIFAR-10-trained VGG-11 and ResNet-34 teachers. We also evaluated with NaturalInversion (NI)[46] 50K synthetic images from the VGG-11 model. As shown in Table VI, applying SFKD during distillation consistently improves the student's accuracy across tested configurations, highlighting its effectiveness in improving synthetic data utilization.

TABLE V. SFKD with multi-teacher knowledge distillation. The student models ShuffleNetV2 & VGG-8 were trained under the configuration of pre-trained Tri-ResNet-32 × 4 on CIFAR-100. **Boldface** denotes results obtained by augmenting multi-teacher distillation methods with SFKD.

Teacher networks	Student network	S. (%)	CA-MKD[44] (%)	AEKD[40] (%)	SFKD+AEKD (%)	SFKD +AEKD-F (%)	Ensemble (%)
Tri-ResNet-32 × 4	ShuffleNetV2	71.82	77.41	75.87	**76.17**	**77.16**	81.31
Tri-ResNet-32 × 4	VGG-8	70.36	75.26	73.11	**73.36**	**73.80**	81.31

TABLE VI. Results of DFKD for various students on CIFAR-10. Boldface denotes results obtained by augmenting data-free knowledge distillation methods with SFKD.

Teacher student	Required data	VGG-11 VGG-11	VGG-11 ResNet-18 (%)	ResNet-34 ResNet-18 (%)
Student accuracy	Yes	92.25%	95.20	95.20
Noise ~ $\mathcal{N}(0,1)$		13.55%	13.45	13.61
DeepDream	No	36.59%	39.67	29.98
SpaceshipNet[47]	No	N/A	92.27	95.39
DFKD-MSARC [48]	No	N/A	92.34	94.91
NaturalInversion (NI)[46]	No	89.79%	90.10	93.72
NI + SFKD ($K^{0.7}$)	No	**90.46%**	**90.85**	···
DeepInversion (DI)[45]	No	84.16%	83.82	91.43
DI + SFKD ($K^{0.7}$)	No	**85.24%**	**84.86**	**91.82**

08 May 2026 01:37:23

© Author(s) 2026

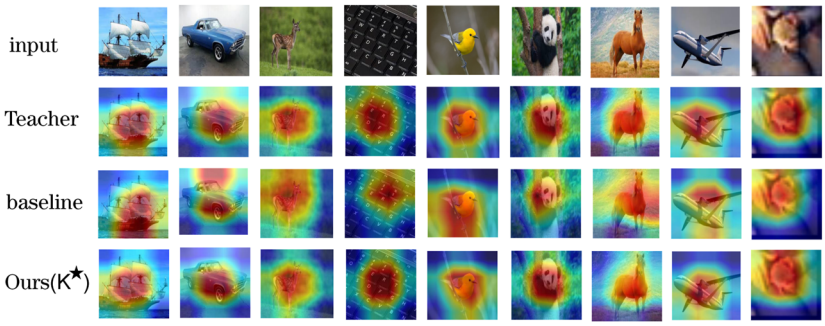

FIG. 3. Class activation map of the distilled student model deployed with our method and baseline AT. The top row presents the input images, while the second, third, and fourth rows display the class activation maps of the teacher model, baseline AT (K^1), and SFKD ($K^\star$), respectively. The deeper the color, the more salient the corresponding feature of the image.

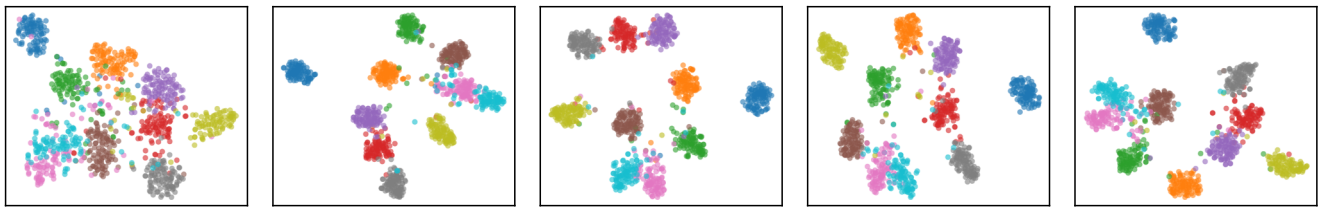

FIG. 4. t-SNE clustering: demonstrating model accuracy on CIFAR-100. (a) Student trained from scratch (ResNet-8 × 4). [(b)–(d)] Student trained with CRD+SFKD. (e) Student trained with CRD only. 10 out of 100 classes were randomly sampled, as indicated by their respective colors. A high density of same-class dots and large separation among classes suggest better model classification accuracy. (a) Student, (b) SFKD ($K^{0.3}$), (c) SFKD ($K^{0.5}$), (d) SFKD ($K^{0.7}$), and (e) baseline CRD.

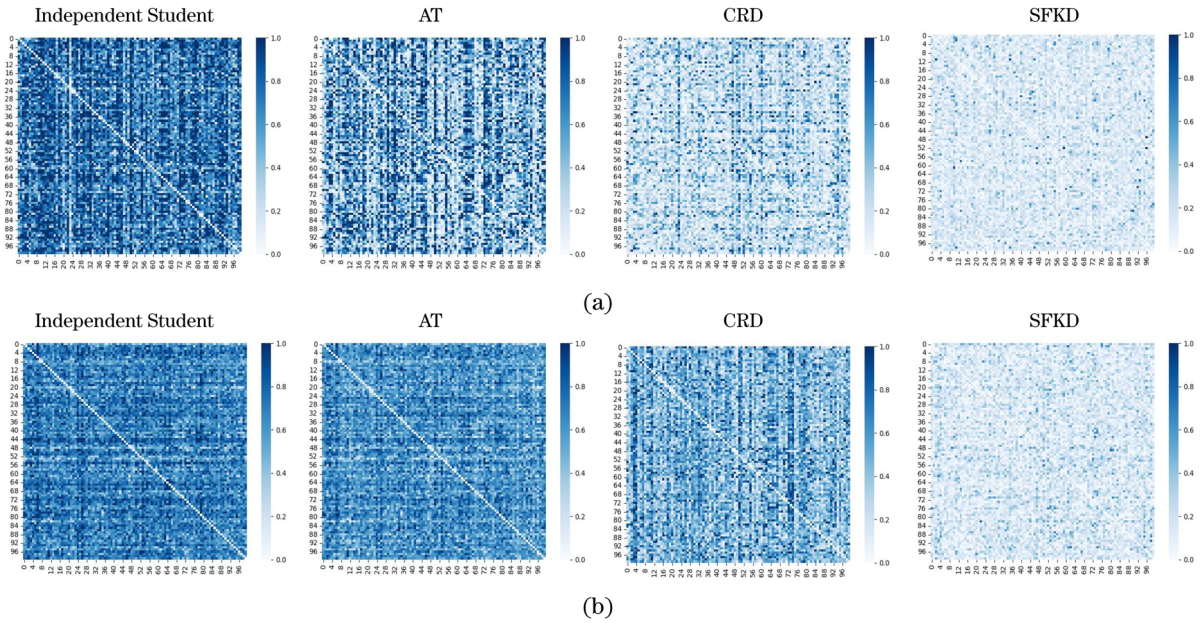

FIG. 5. Contrast in correlation matrices of teacher and student classifier weights on CIFAR-100. The correlation matrices are computed using normalized weights. (a) Teacher: ResNet-32 × 4, Student: ResNet-8 × 4. (b) Teacher: VGG-13, Student: MobileNetV2. Each subplot shows the L1 error heatmap between teacher and student correlation matrices for independent student, AT, CRD, and SFKD methods. (a) Teacher: ResNet-32 × 4, Student: ResNet-8× 4. (b) Teacher: VGG-13, Student: MobileNetV2.

08 May 2026 01:37:23

© Author(s) 2026

C. Visualization

To provide insight into *how* SFKD improves student models, we visualize and analyze the learned representations.

1. Focused attention with class activation maps (CAMs)

We use CAMs[49] to visualize where the model is "looking." Figure 3 contrasts the student model's attention when trained with a baseline method (AT) vs our AT+SFKD. The baseline model's focus often spreads to irrelevant background areas. In contrast, the SFKD-trained student concentrates its attention squarely on the target objects (car, bird, horse), closely mimicking the teacher's focus and demonstrating an improved ability to learn salient features.

2. Improved feature separability with t-SNE

To assess feature quality, we use t-SNE[50] to project the feature distributions of student networks trained on CIFAR-100 (ResNet-32 × 4 → ResNet-8 × 4). As shown in Fig. 4(a), a student trained from scratch or with a baseline (CRD) exhibits significant class overlap. The student trained with SFKD, however, produces feature clusters that are far more compact and clearly separated, indicating a more discriminative and effective representation.

3. Classifier pattern matching

We further quantify the student's ability to learn the teacher's internal logic by measuring the L1 error between their classifier weight correlation matrices and illustrate this variance using a heatmap (Fig. 5). Four methods were examined: the independent student without any distillation, alongside students trained with AT,[2] CRD,[3] and our approach, SFKD ($K^{0.3}$). The findings demonstrate that SFKD records the minimal difference across both sets of teacher–student pairs, showcasing SFKD's superior ability to replicate the teacher's correlation patterns.

VIII. CONCLUSION

In this study, we introduce salient feature masking for knowledge distillation, a simple but effective method that selectively distills the most pertinent features to enhance student performance. Compatible with existing KD variants, logit-based SFKD allows direct manipulation of a pretrained network's logits by preserving high probability class values. This effective technique is easily applicable to large networks in real-world scenarios, which require no retraining or modification of the original model. Leveraging the information bottleneck principle, we provide a theoretical analysis and an interoperability of SFKD's effectiveness, which explores insights into the teacher model's decision-making process. Our study opens up a few interesting research directions. First, it is intriguing to explore the characteristics of information flow during the distillation process. Second, finding the optimal K value effectively without extensive tuning is important for the top-K salient feature distillation regarding heterogeneous teacher–student networks.

SUPPLEMENTARY MATERIAL

See the supplementary material for additional experimental results, including comprehensive ablation studies, sensitivity analyses across different K values, mutual information estimation details, and extended visualizations of learned representations.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Assel Kembay: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Methodology (equal); Software (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Skye Gunasekaran**: Formal analysis (equal); Validation (equal); Writing – review & editing (equal). **Rui-Jie Zhu**: Formal analysis (equal); Investigation (equal); Validation (equal). **Yu Zhang**: Formal analysis (equal); Methodology (equal); Supervision (supporting); Validation (equal); Writing – review & editing (equal). **Jason K. Eshraghian**: Conceptualization (equal); Funding acquisition (equal); Investigation (equal); Project administration (equal); Resources (equal); Supervision (equal); Validation (equal); Writing – review & editing (equal).

DATA AVAILABILITY

Our code is available at https://github.com/akembay/SFKD.

REFERENCES

[1] A. Romero, N. Ballas, S. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," arXiv:1412.6550 (2014).
[2] N. Komodakis and S. Zagoruyko, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in Proceedings of the International Conference on Learning Representations (ICLR), 2017.
[3] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in Proceedings of the International Conference on Learning Representations (ICLR), 2019.
[4] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2021), pp. 5008–5017.
[5] U. Ojha, Y. Li, A. Sundara Rajan, Y. Liang, and Y. J. Lee, "What knowledge gets distilled in knowledge distillation?," 36 (Advances in Neural Information Processing Systems, 2023), pp. 11037–11048.
[6] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," in ICLR, 2018.
[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531 (2015).
[8] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in International Conference on Machine Learning, 2018.
[9] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in Proceedings of the AAAI Conference on Artificial Intelligence, 2019.
[10] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR, 2022), pp. 11953-11962.

08 May 2026 01:37:23

© Author(s) 2026

[11] Y. Jin, J. Wang, and D. Lin, "Multi-level logit distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2023), pp. 24276–24285.
[12] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2019), pp. 1921–1930.
[13] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2019), pp. 3967–3976.
[14] B. Peng, X. Jin, J. Liu, S. Zhou, Y. Wu, Y. Liu, D. Li, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5007-5016, 2019.
[15] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 2019), pp. 1365–1374.
[16] X. Liu, L. Li, C. Li, and A. Yao, *NORM: Knowledge Distillation via N-to-One Representation Matching* (ICLR, 2023).
[17] Z. Guo, H. Yan, H. Li, and X. Lin, "Class attention transfer based knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2023), pp. 11868–11877.
[18] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2019), pp. 2604–2613.
[19] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2022), pp. 12319–12328.
[20] L. Li, Y. Bao, P. Dong, C. Yang, A. Li, W. Luo, Q. Liu, W. Xue, and Y. Guo, "DetKDS: Knowledge distillation search for object detectors," in Forty-First International Conference on Machine Learning, 2024.
[21] H. Zhang, L. Liu, Y. Huang, Z. Yang, X. Lei, and B. Wen, "CaKDP: Category-aware knowledge distillation and pruning framework for lightweight 3D object detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2024), pp. 15331–15341.
[22] P. Dong, L. Li, and Z. Wei, "DisWOT: Student architecture search for distillation without training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2023), pp. 11898–11908.
[23] S. Sun, W. Ren, J. Li, R. Wang, and X. Cao, "Logit standardization in knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024), pp. 15731–15740.
[24] W. Sun, D. Chen, S. Lyu, G. Chen, C. Chen, and C. Wang, "Knowledge distillation with refined logits," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2025), pp. 1110–1119.
[25] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," arXiv:physics/0004057 (2000).
[26] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in Proceedings of the Information Theory Workshop (ITW), 2015.
[27] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," Entropy **21**(219), 3390 (2019).
[28] R. Pogodin and Pe. E. Latham, "Kernelized information bottleneck leads to biologically plausible 3-factor Hebbian learning in deep networks," arXiv:2006.07123v2.
[29] C. Wang, Q. Yang, R. Huang, S. Song, and G. Huang, "Efficient knowledge distillation from model checkpoints," in *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems* (2022), pp. 607–619.
[30] Z. Goldfeld, E. van den Berg, K. H. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy, "Estimating information flow in deep neural networks," in International Conference on Learning Representations (ICML), 2019.
[31] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, 2019, pp. 9163–9171.
[32] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," arXiv:1703.00810 (2017).
[33] Swin-Pico referred to as Swin-P.
[34] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," Advances in Neural Information Processing Systems **35**, 33716–33727 (2022).
[35] L. Li, P. Dong, A. Li, Z. Wei, and Y. Yang, "Kd-zero: Evolving knowledge distiller for any teacher-student pairs," 36 (Advances in Neural Information Processing Systems, 2023), pp. 69490–69504.
[36] L. Li, P. Dong, Z. Wei, and Y. Yang, "Automated knowledge distillation via monte carlo tree search," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2023), pp. 17413–17424.
[37] D. Chen, J.-P. Mei, H. Zhang, C. Wang, Y. Feng, and C. Chen, "Knowledge distillation with the reused teacher classifier," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2022), pp. 11933–11942.
[38] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona (2010). "Caltech-UCSD birds 200," Caltech Vision Lab, http://www.vision.caltech.edu/visipedia/CUB-200.html.
[39] D. Chen, J.-P. Mei, Y. Zhang, C. Wang, Z. Wang, Y. Feng, and C. Chen, "Cross-layer distillation with semantic calibration," in *Proceedings of the AAAI Conference on artificial Intelligence* (AAAI, 2021), Vol. 35, pp. 7028–7036.
[40] S. Du, S. You, X. Li, J. Wu, F. Wang, C. Qian, and C. Zhang, "Agree to disagree: Adaptive ensemble knowledge distillation in gradient space," in *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (2020), pp. 12345–123.
[41] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2017), pp. 1285–1294.
[42] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," Interspeech **2017**, 3697–3701.
[43] M.-C. Wu, C.-T. Chiu, and K.-H. Wu, "Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2019), pp. 2202–2206.
[44] H. Zhang, D. Chen, and C. Wang, "Confidence-aware multi-teacher knowledge distillation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 4498–4502.
[45] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "Dreaming to distill: Data-free knowledge transfer via deep inversion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2020), pp. 8715–8724.
[46] Y. Kim, D. Park, D. Kim, and S. Kim, "Naturalinversion: Data-free image synthesis improving real-world consistency," in *Proceedings of the AAAI Conference on artificial Intelligence* (AAAI, 2022), Vol. 36, pp. 1201–1209.
[47] S. Yu, J. Chen, H. Han, and S. Jiang, "Data-free knowledge distillation via feature exchange and activation region constraint," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2023), pp. 24266–24275.
[48] P. Liang, J. Chen, Y. Wu, B. Pu, H. Huang, Q. Chang, and G. Ran, "Data free knowledge distillation with feature synthesis and spatial consistency for image analysis," Sci. Rep. **14**, 27557 (2024).
[49] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 2921–2929.
[50] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res. **9**, 2579–2605 (2008), available at http://jmlr.org/papers/v9/vandermaaten08a.html.
[51] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," Advances in Neural Information Processing Systems **35**, 33716–33727 (2022).

08 May 2026 01:37:23

© Author(s) 2026