Efficient Knowledge Distillation via Salient Feature Masking

Anonymous authors
Paper under double-blind review

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 across architectures (ConvNeXt, ViT) and scales (CIFAR-100: +5.44 pp; CUB: +6.39 pp; ImageNet-1K: +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.

1 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 feature and attention maps (Romero et al., 2014; Komodakis & Zagoruyko, 2017; Tian et al., 2019; Chen et al., 2021b). However, such indiscriminate transfer overwhelms the student model with peripheral or even misleading activations, thereby misguiding its limited capacity and hindering generalization (Ojha et al., 2023).

We reinterpret distillation through the lens of the *Information Bottleneck* (IB) principle (Saxe et al., 2018). 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 justification and guarantee.** By re-casting the mask selection as an Information-Bottleneck optimization, we prove that the retained activations maximize mutual

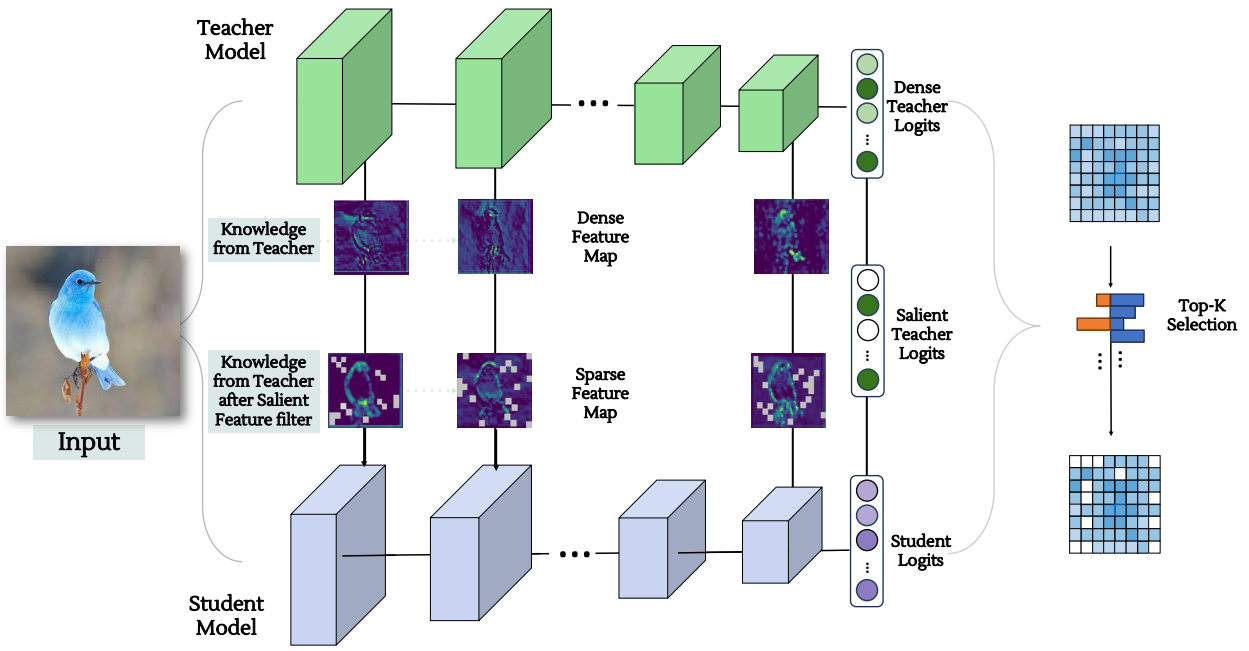


Figure 1: The concept of the proposed SFKD. SFKD distinctively concentrates on: 1) distilling critical classification knowledge, 2) transferring essential information from intermediate layers, and 3) refining attention mechanisms for knowledge distillation.

information $I(X; F_K)$ under a budget on student capacity, and we bound the information lost when discarding the remaining entries.

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.

2 Related Work

Knowledge Distillation (KD) variants generally fall into three categories based on the type of knowledge transferred: logits (Hinton et al., 2015; Furlanello et al., 2018; Mirzadeh et al., 2019; Zhao et al., 2022; Jin et al., 2023), features (Chen et al., 2021b; Heo et al., 2019; Park et al., 2019; Peng et al., 2019; Romero et al., 2014; Tung & Mori, 2019; Tian et al., 2019; Liu et al., 2023), and attention (Komodakis & Zagoruyko, 2017; Guo et al., 2023). Vanilla KD (Hinton et al., 2015) 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 (Romero et al., 2014) aligns feature maps between specific teacher-student layers. Attention-based methods (Komodakis & Zagoruyko, 2017) use attention maps derived from feature representations for comprehensive knowledge transfer across layers. Subsequent studies explore applications of KD in semantic segmentation (Liu et al., 2019; Yang et al., 2022), object detection (Li et al., 2024; Zhang et al., 2024), and student architecture search (Dong et al., 2023).

Information Bottleneck (IB) is a principle introduced by (Tishby et al., 2000), 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 (Tishby & Zaslavsky, 2015), proposing that deep neural networks (DNNs) implicitly optimize this trade-off during training. (Shwartz-Ziv & Tishby, 2017) 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 & Latham, 2020) 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., 2022) 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 work uniquely applies the IB principle to interpret the KD process. While (Goldfeld et al., 2019) 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.

3 Methods

Let $\mathbf{a} \in \mathbb{R}^N$ denote a one–dimensional teacher activation (e.g., the class–logit vector, a flattened feature map, or a flattened attention tensor). Our goal is to retain only the K most informative elements (as measured via mutual information) and suppress the rest. We define the top-K index set as:

$$I_{Top-K} = \{\, i \in \{1, \ldots, N\} \mid \mathbf{a}_i \text{ is among the } K \text{ largest elements of } \mathbf{a} \,\} \tag{1}$$

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

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

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

$$\mathbf{a}_K := \mathbf{M} \odot \mathbf{a}. \tag{3}$$

This operation leaves the K salient entries untouched ($a_{K,i} = a_i$ for $i \in I_{\text{Top-}K}$) and zeros out all others ($a_{K,i} = 0$ otherwise). Whenever the activation vector is denoted $\mathbf{F}$ we write $\mathbf{F}_K = \mathbf{M} \odot \mathbf{F}$ for brevity.

Our findings have broad applicability, covering a wide range of distillation techniques, as illustrated in Figure 1. We focus on standard methods representing three main families of distillation approaches: *output-based* (Hinton et al., 2015), *feature-based* (Romero et al., 2014), and *attention-based* (Komodakis & Zagoruyko, 2017). 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: The salient feature masking operates on the logit space by retaining only the top-K logits from the teacher's distribution based on their magnitude. The knowledge transfer is then performed through KL-divergence minimization between the masked teacher distribution and student predictions:

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

where z_{t^K} denotes the teacher's logits after top-K masking; τ is a scaling temperature; and the overall loss function is $\gamma L_{CLS} + \alpha L_{KL}$ with balancing parameters γ and α.

(2) Feature-based: The student's intermediate features $F_s^{(l)}$ are trained to mimic only the K relevant features of the teacher's $F_{t^K}^{(l)}$ for a given image X at layer 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 linear projection layer to align the number of channels in F_s with those in F_t). Their similarity is then optimized by minimizing the mean squared error:

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

The total loss is $\gamma L_{CLS}+\beta L_{Hint}$, where γ and β are balancing parameters. '*Hint*' represents all feature-based KD methods.

(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 $Q_s^j = \text{vec}(\phi(A_s^j))$ and $Q_{t^K}^j = \text{vec}(\phi(\text{Top}_K(A_t^j))$ are respectively the j-th pair of student and top-K elements in the teacher's attention maps in vectorized form. A mapping function ϕ maps a 3D activation tensor $A \in R^{C \times H \times W}$ to a spatial attention map. $\beta > 0$ is a balancing parameter and $\|x\|_p$ is the ℓ_p norm of vector x (typically ℓ_2 norm).

4 An Information-Theoretic Perspective On SFKD

In this section, we use the well-established Information Bottleneck (IB) 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.

4.1 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 (Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017) 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 networks, the mutual information $I(X;F)$ is technically infinite. Following common practice in IB analysis of deep neural networks, we use practical MI estimators that rely on well-established lower bounds, allowing us to qualitatively analyze the information flow during training (Shwartz-Ziv & Tishby, 2017; Ahn et al., 2019).

Viewing this through the IB lens, the goal of knowledge distillation should be to create a "bottleneck" that filters the teacher's knowledge before it is transferred to the student. This ensures the student focuses its limited capacity on the most salient information.

The IB principle inspires several hypotheses about how SFKD should affect the student's learning dynamics, which we can visualize on the "information plane" ($I(X;F)$ vs. $I(F;Y)$).

- **H1** *Less input compression* – By focusing only on salient features, SFKD allows the student to retain more information about the original input, leading to a higher $I(X;F_s)$.
- **H2** *Faster label informativeness* – By receiving a cleaner, more concentrated learning signal, the student model trained with SFKD should learn the relationship with the output labels more quickly, resulting in a steeper initial increase in $I(F_s;Y)$.
- **H3** *Intermediate K is best* – Transferring too little information (a very small K) or too much (no masking) is suboptimal. An ideal "elbow point" for K should exist, where the student's performance is maximized.

Our empirical results, visualized in Figure 2, strongly support these hypotheses. We observe that SFKD consistently guides the student to a better position on the information plane (H1) and accelerates the learning of label-relevant information (H2) compared to standard KD. Furthermore, our ablation on the value of K confirms that an intermediate level of masking is indeed optimal (H3), validating the core idea of creating an information bottleneck.

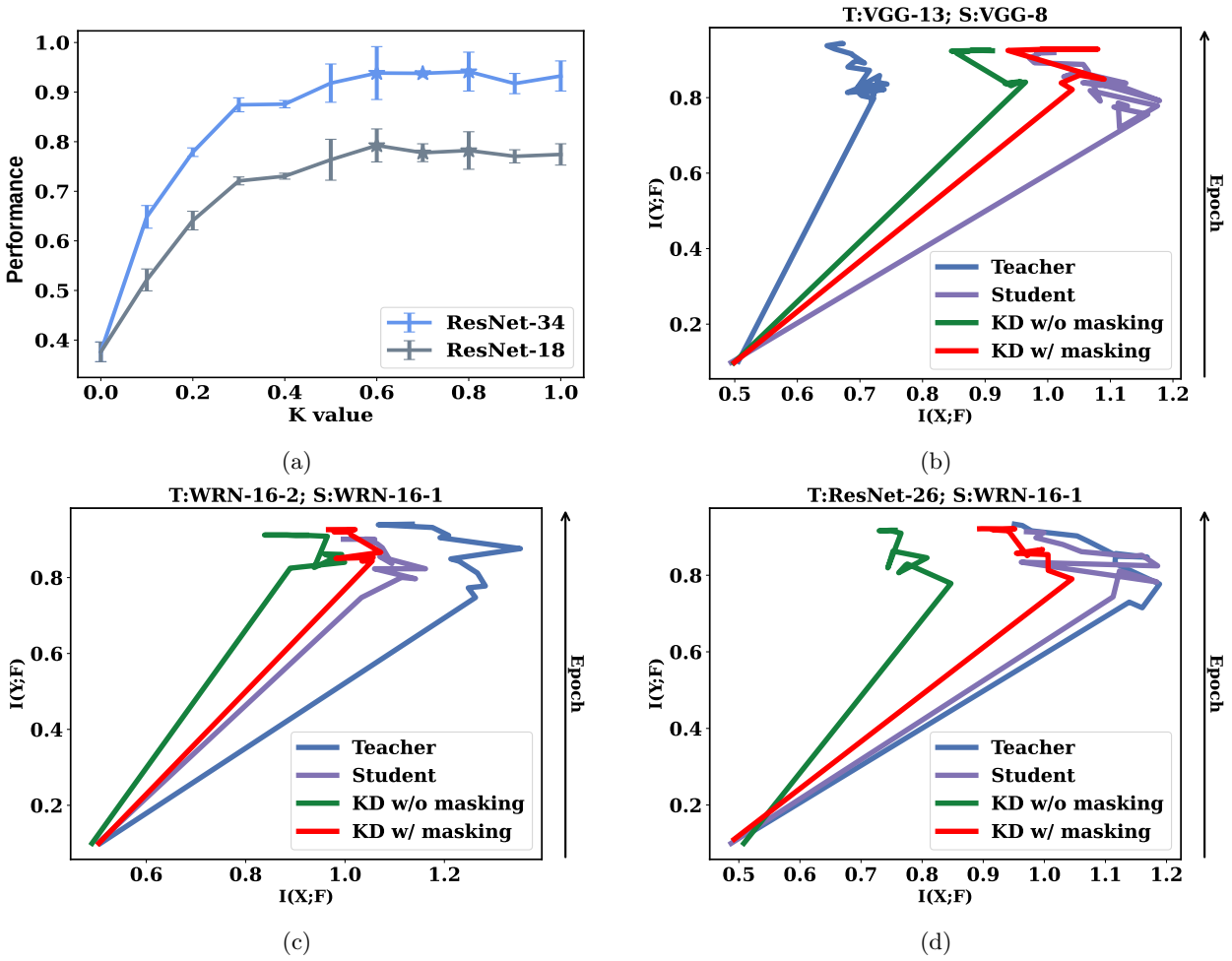


Figure 2: (a) An example of the optimal K selection for salient feature masking. The horizontal axis shows the degree of Top-K sparsity ($K = 1$ meaning no masking). (b)–(d) The information plane for different teacher-student networks: the mutual information trajectories with respect to the training epochs.

5 Experiments

We demonstrate that our SFKD approach is method-agnostic by testing it across various existing distillation methods. Additionally, 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.

Results on CIFAR-100. Table 1 demonstrates the performance and robustness of SFKD when applied to similar vs. dissimilar model architectures. 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 2. This broader testing scope further validates the versatility of our method across diverse neural network designs to distill knowledge effectively across any architectures, such as from *ViTs* (Swin-T) to *CNN-based student* (ResNet-18) and from *ConvNeXt* to *Swin-P*[1].

Results on ImageNet. Table 3 reports Top-1 and Top-5 accuracies for SFKD ("Ours") compared to baseline methods (Chen et al., 2021b; Tian et al., 2019). Across teacher–student pairs of varying capacity

[1]Swin-Pico referred to as Swin-P

gaps, SFKD yields consistent gains in both metrics. These findings demonstrate the scalability of our approach to large-scale datasets, where maintaining signal quality during transfer is particularly challenging.

Results on CUB200. Table 4 evaluates SFKD on the fine-grained CUB200 bird classification task (Welinder et al., 2010), which requires discriminating between subtle inter-class variations. SFKD achieves substantial accuracy improvements across all tested configurations, indicating that its selective feature transfer enhances the capture of fine-grained discriminative cues.

Table 1: Comprehensive performance comparison on CIFAR-100. **Bold** indicates the best, underbar is the second-best value.

Method	Same architecture style				Different architecture style		
T/S Pair	WRN-40-2	WRN-40-2	ResNet-32x4	VGG-13	VGG-13	ResNet-32x4	WRN-40-2
	WRN-16-2	WRN-40-1	ResNet-8x4	VGG-8	MobileNetV2	ShuffleNetV2	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 (Guo et al., 2023) (CVPR'23)	75.60	74.82	76.91	74.65	69.13	78.41	77.35
ReviewKD (Chen et al., 2021b) (CVPR'21)	76.12	75.09	75.63	74.84	70.37	77.78	77.14
DIST (Huang et al., 2022) (NeurIPS'22)	N/A	74.73	76.31	N/A	N/A	77.35	N/A
KD-Zero (Li et al., 2023a) (NeurIPS'23)	76.42	N/A	77.85	75.26	70.42	77.45	77.52
Auto-KD (Li et al., 2023b) (ICCV'23)	76.86	N/A	77.61	75.36	70.58	77.52	77.46
RLD (Sun et al., 2024b) (ICCV'25)	76.02	74.88	76.64	74.93	69.97	77.56	N/A
LS (MLKD+LS) (Sun et al., 2024a) (CVPR'24)	76.95	75.56	78.28	75.22	**70.94**	78.76	N/A
DKD (Zhao et al., 2022) (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 (Chen et al., 2022) (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 (Jin et al., 2023) (CVPR'23)	76.63	75.35	77.08	75.18	70.57	78.44	77.44
MLKD + SFKD	**77.01**	75.72	78.06	**75.60**	70.58	**79.16**	77.50

Table 2: SFKD with heterogeneous architectures on CIFAR-100: *ViT-based teachers* distilled to both *CNN-based* and *ViT-based students*.

ViT-based Teachers	T.	Swin-T	ViT-S	Mixer-B/16	ConvNeXt-T
	S.	ResNet-18	ResNet-18	ResNet-18	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 (Huang et al., 2022)	77.75	76.49	76.36	76.41
	KD (Hinton et al., 2015)	78.74	77.26	77.79	76.44
	KD + SFKD	**80.62**$_{+1.88}$	**78.90**$_{+1.64}$	**79.18**$_{+1.39}$	**78.87**$_{+2.43}$

6 Discussion

Across CIFAR-100, ImageNet, and CUB200, SFKD consistently outperforms strong KD baselines, with relative gains up to +6.39 percentage points in accuracy on the CUB200 dataset. The breadth of tested teacher–student combinations—from homogeneous CNN–CNN settings to heterogeneous ViT–CNN and ConvNeXt–ViT transfers—demonstrates that SFKD's masking mechanism generalizes well to diverse architectural paradigms.

Furthermore, these performance improvements align with the Information Bottleneck perspective underpinning SFKD: by filtering out low-informative activations, the method reduces transfer noise and compels the

Table 3: Top-1 and Top-5 accuracy (%) on ImageNet validation.

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 (Chen et al., 2021b)	71.61	90.51	72.56	91.00
SimKD (Chen et al., 2022)	71.59	90.48	72.25	90.86
CAT-KD (Guo et al., 2023)	71.26	90.45	72.24	91.13
AT (Komodakis & Zagoruyko, 2017)	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 (Hinton et al., 2015)	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 (Zhao et al., 2022)	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 4: 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.

Teacher	ResNet-32x4	ResNet-32x4	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 (Tung & Mori, 2019)	48.49	61.83	44.28	54.78	55.31
CRD (Tian et al., 2019)	57.45	62.28	56.45	66.10	57.45
SemCKD (Chen et al., 2021a)	56.89	63.78	68.23	66.54	57.20
ReviewKD (Chen et al., 2021b)	-	64.12	58.66	67.10	-
KD (Hinton et al., 2015)	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 (Zhao et al., 2022)	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**$_{+4.78}$

student to focus on the most predictive components. This not only yields higher accuracy but also enhances representation quality, as corroborated by t-SNE and class activation map analyses (Section 7.3). The results further suggest that SFKD is particularly beneficial in settings with limited student capacity or noisy supervision signals, such as data-free or multi-teacher distillation.

The empirical evidence provided supports SFKD as a lightweight, 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.

7 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.

Table 5: SFKD with Multi-Teacher Knowledge Distillation. The student models ShuffleNetV2 & VGG-8 were trained under the configuration of pre-trained Tri-ResNet-32x4.

Teacher Networks	Student Network	S.	AEKD	SFKD + AEKD	SFKD + AEKD-F	Ensemble
Tri-ResNet-32x4	ShuffleNetV2	71.82%	75.87%	**76.17%**	**77.16**%	81.31%
Tri-ResNet-32x4	VGG-8	70.36%	73.11%	**73.36%**	**73.80%**	81.31%

Table 6: Results of DFKD to various students on CIFAR-10.

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 $\sim \mathcal{N}(0, 1)$		13.55%	13.45%	13.61%
DeepDream	No	36.59%	39.67%	29.98%
DeepInversion (DI)	No	84.16%	83.82%	91.43%
DI + SFKD ($K^{0.7}$)	No	**85.24%**	**84.86%**	**91.82%**

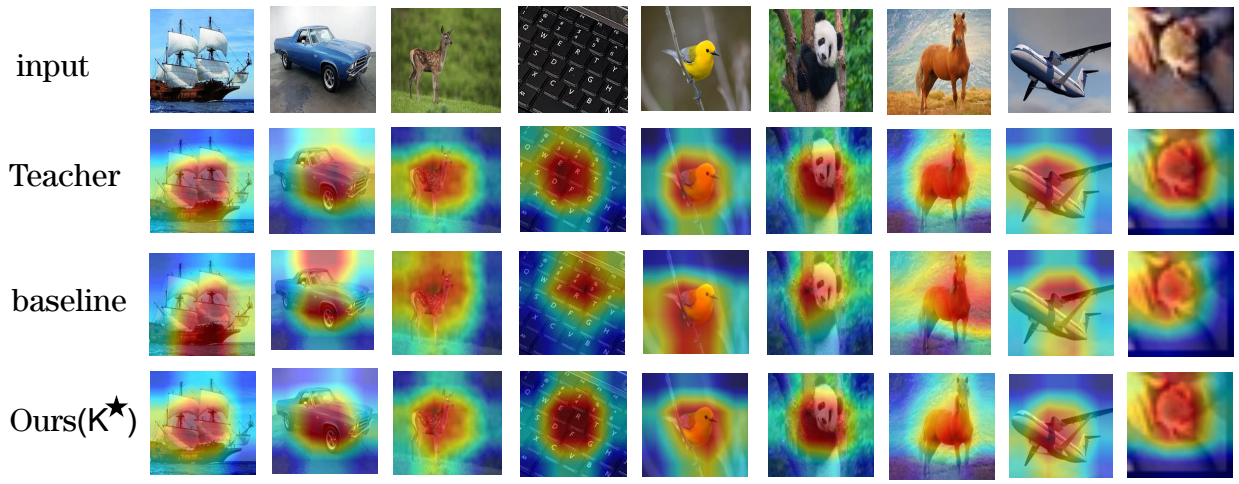
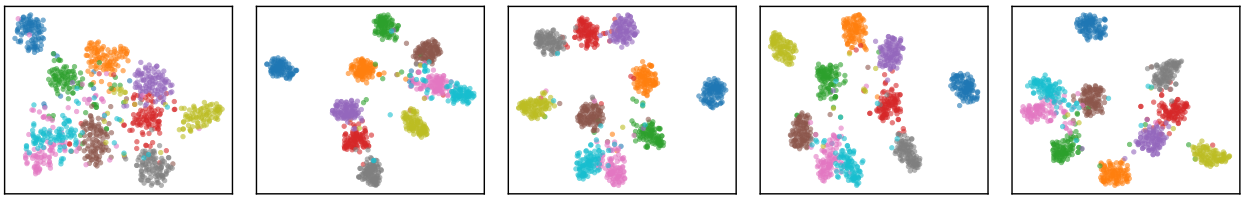


Figure 3: Class activation map of the distilled student model deployed with our method and baseline AT, the teacher model. The deeper the color, the more salient the corresponding feature of the image. The top row presents the input images, while the second, third, and fourth rows display the class activation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

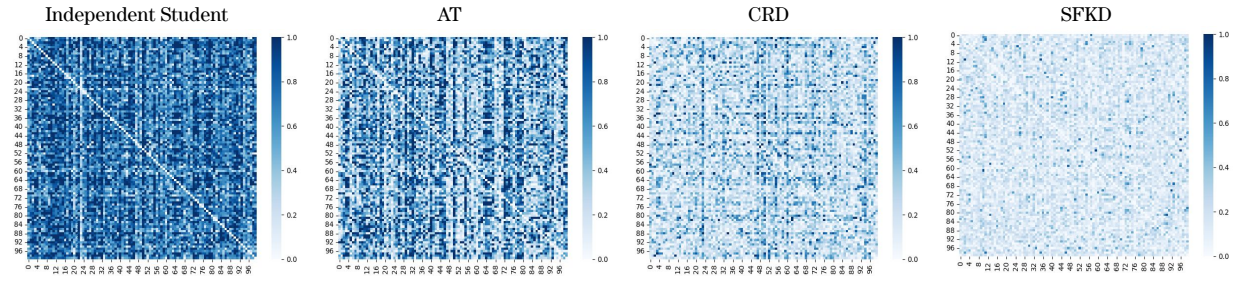
7.1 Application 1: Selective Knowledge Sharing in Multi-Teacher Knowledge Distillation

In multi-teacher distillation, conventional methods [(Du et al., 2020; You et al., 2017; Fukuda et al., 2017; Wu et al., 2019)] 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 5). We demonstrate this by applying SFKD to the AEKD framework (Du et al., 2020), using a Tri-ResNet-32x4 ensemble to teach both VGG-8 and ShuffleNetV2 students. Across all configurations, SFKD consistently achieves the best performance by effectively channeling the most pertinent insights from the multiple experts.

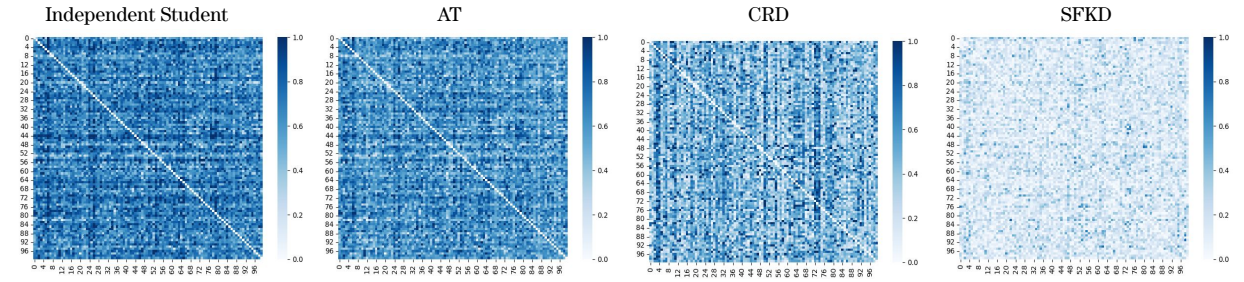


(a) Student (b) SFKD ($K^{0.3}$) (c) SFKD ($K^{0.5}$) (d) SFKD ($K^{0.7}$) (e) baseline CRD

Figure 4: t-SNE clustering: demonstrating model accuracy on CIFAR-100. 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 suggests better model classification accuracy.



(a) Teacher: ResNet-32×4, Student: ResNet-8×4



(b) Teacher: VGG-13, Student: MobileNetV2

Figure 5: Contrast in correlation matrices of teacher and student classifier weights on CIFAR-100. The correlation matrices are computed using normalized weights.

7.2 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) (Yin et al., 2020) from CIFAR-10-trained VGG-11 and ResNet-34 teachers. As shown in Table 6, applying SFKD during distillation consistently improves the student's accuracy across all tested teacher-student pairs, highlighting its effectiveness in improving synthetic data utilization.

7.3 Visualization

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

Focused Attention with Class Activation Maps (CAMs). We use CAMs (Zhou et al., 2016) to visualize where the model is "looking". Figure 3 contrasts the student model's attention when trained with a baseline method (AT) versus 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.

Improved Feature Separability with t-SNE. To assess feature quality, we use t-SNE (van der Maaten & Hinton, 2008) to project the feature distributions of student networks trained on CIFAR-100 (ResNet-32x4 $\rightarrow$ ResNet-8x4). As shown in Figure 4, 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.

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 matricesand illustrate this variance using a heatmap (Figure 5). Four methods were examined: the independent student without any distillation, alongside students trained with AT (Komodakis & Zagoruyko, 2017), CRD (Tian et al., 2019), and our approach, SFKD ($\mathrm{K}^{0.3}$). 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.

8 Conclusion

In this work, 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 pre-trained network's logits by preserving high probability class values. This effective technique is easily applicable to large networks in real-world scenarios, which requires no retraining or modification of the original model. Leveraging the information bottleneck principle, we provide theoretical analysis and interoperability of SFKD's effectiveness, which explores insights into the teacher model's decision-making process. Our work 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.

References

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9163–9171, 2019.

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*, pp. 11933–11942, 2022.

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*, volume 35, pp. 7028–7036, 2021a.

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)*, pp. 5008–5017, 2021b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009.

Peijie Dong, Lujun Li, and Zimian Wei. Diswot: Student architecture search for distillation without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11898–11908, 2023.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. In *Advances in Neural Information Processing Systems*, 2020.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pp. 3697–3701, 2017.

Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Animashree Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, 2018.

Ziv Goldfeld, Ewout van den Berg, Kristjan H. Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In *ICML*, 2019.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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*, pp. 11868–11877, 2023.

Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *In ICLR*, 2019.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727, 2022.

Y. Jin, J. Wang, and D. Lin. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24276–24285, 2023.

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.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Lujun Li, Peijie Dong, Anggeng Li, Zimian Wei, and Ya Yang. Kd-zero: Evolving knowledge distiller for any teacher-student pairs. *Advances in Neural Information Processing Systems*, 36:69490–69504, 2023a.

Lujun Li, Peijie Dong, Zimian Wei, and Ya Yang. Automated knowledge distillation via monte carlo tree search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17413–17424, 2023b.

Lujun Li, Yufan Bao, Peijie Dong, Chuanguang Yang, Anggeng Li, Wenhan Luo, Qifeng Liu, Wei Xue, and Yike Guo. Detkds: Knowledge distillation search for object detectors. In *Forty-first International Conference on Machine Learning*, 2024.

Xiaolong Liu, Lujun Li, Chao Li, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching. *arXiv preprint arXiv:2305.13803*, 2023.

Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2604–2613, 2019.

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*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131, 2018.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Utkarsh Ojha, Yuheng Li, Anirudh Sundara Rajan, Yingyu Liang, and Yong Jae Lee. What knowledge gets distilled in knowledge distillation? *Advances in Neural Information Processing Systems*, 36:11037–11048, 2023.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 268–284, 2018.

Baoyun Peng, Xiao Jin, Jiaheng Liu, Shunfeng Zhou, Yichao Wu, Yu Liu, Dongsheng Li, and Zhaoning Zhang. Correlation congruence for knowledge distillation. *arXiv preprint arXiv:1904.01802*, 2019.

Roman Pogodin and Peter E. Latham. Kernelized information bottleneck leads to biologically plausible 3-factor hebbian learning in deep networks. In *NeurIPS*, 2020.

A. Romero, N. Ballas, S.E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2234–2242, 2016.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Andrew M. Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D. Tracey, and David D. Cox. On the information bottleneck theory of deep learning. In *ICLR*, 2018.

R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15731–15740, 2024a.

Wujie Sun, Defang Chen, Siwei Lyu, Genlang Chen, Chun Chen, and Can Wang. Knowledge distillation with refined logits. *arXiv preprint arXiv:2408.07703*, 2024b.

Y. Tian, D. Krishnan, and P. Isola. Contrastive representation distillation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *Proceedings of the Information Theory Workshop (ITW)*, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

F. Tung and G. Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1365–1374, 2019.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. Efficient knowledge distillation from model checkpoints. *Advances in Neural Information Processing Systems*, 35:607–619, 2022.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

Meng-Chieh Wu, Ching-Te Chiu, and Kun-Hsuan 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)*, pp. 2202–2206. IEEE, 2019.

Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12319–12328, 2022.

Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8715–8724, 2020.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1285–1294, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Haonan Zhang, Longjun Liu, Yuqi Huang, Zhao Yang, Xinyu Lei, and Bihan 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)*, pp. 15331–15341. IEEE, 2024.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, 2018.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. *arXiv preprint arXiv:2203.08679*, 2022.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

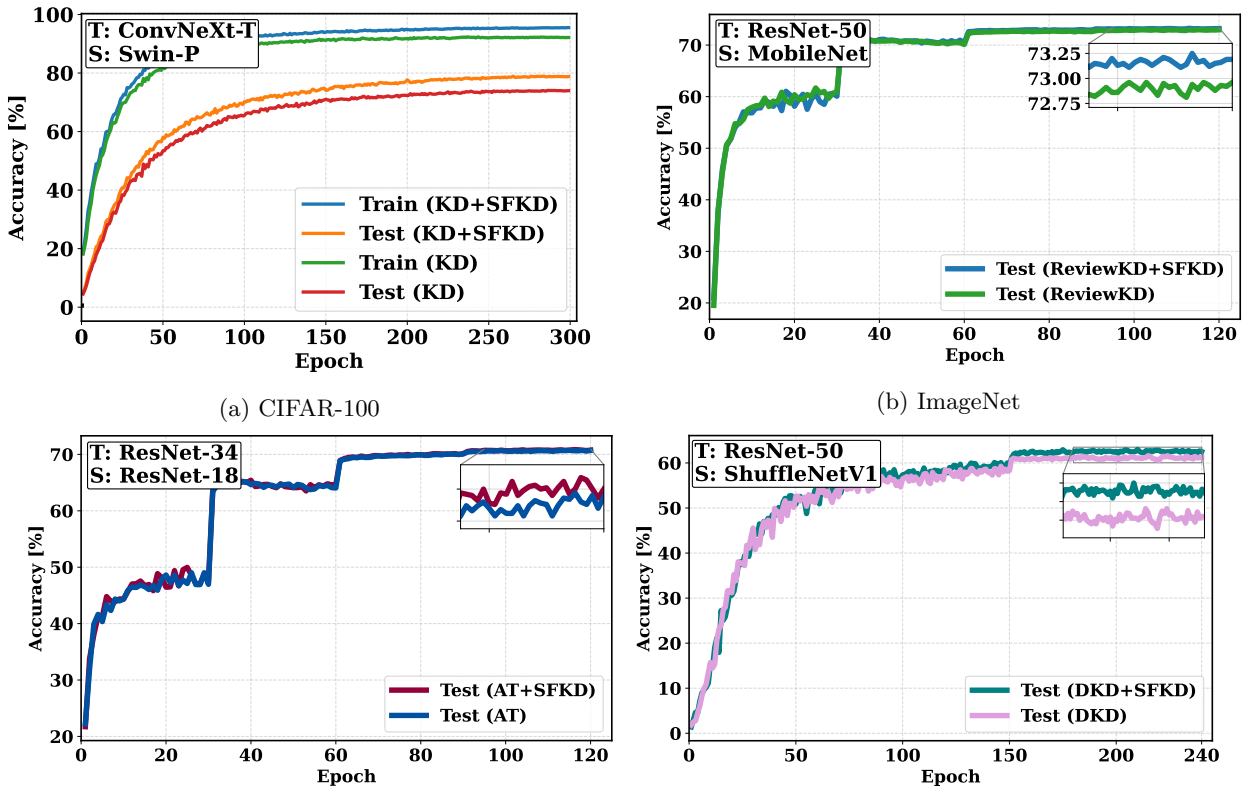
Appendix

This supplementary document details mutual information estimation for $I(X;F)$ and $I(Y;F)$ (Section A), complete training setup with hyperparameters for CIFAR-100, CUB200, and ImageNet (Section B), and additional experiments including ablations, data-free knowledge distillation results, and representation visualizations (Section C).

A Mutual Information Estimation.

Estimating *I(X;F).* Let $R(X|F)$ denote the expected error for reconstructing X from F. It is well known that $R(X|F)$ follows $I(X;F) = H(X) - H(X|F) \geq H(X) - R(X|F)$, where $H(X)$ is the Shannon entropy of X, which is a constant (Hjelm et al., 2019). Therefore, we estimate $I(X;F)$ by training a decoder parameterized by w to obtain the minimal reconstruction loss, namely $I(X;F) \approx \max_w [H(X) - R_w(X|F)]$. In practice, we use the binary cross-entropy loss for $R_w(X|F)$.

Estimating *I(Y;F).* Since $I(Y;F) = H(Y) - H(Y|F) = H(Y) - \mathbb{E}_{(F,Y)}[-\log p(Y|F)]$, a straightforward approach is to train an auxiliary classifier $q_\psi(Y|F)$ with parameters ψ to approximate $p(Y|F)$, such that we have $I(Y;F) \approx \max_\psi \{H(Y) - \mathbb{E}_F[\sum_Y -p(Y|F)\log q_\psi(Y|F)]\}$. Finally, we estimate the expectation over F using its sample mean $I(Y;F) \approx \max_\psi \{H(Y) - \frac{1}{N}[\sum_{i=1}^N -\log q_\psi(Y_i|F_i)]\}$, where $\{(X_i, F_i, Y_i)\}_{i=1}^N$ are the samples. Consequently, $q_\psi(Y|F)$ can be trained in a regular classification fashion with the cross-entropy loss.



(a) CIFAR-100

(b) ImageNet

(c) ImageNet

(d) CUB200

Figure 6: (a) ConvNeXt-T distilled to ViT-based student, evaluated on CIFAR-100. (b)–(c) ReviewKD and AT with our method, evaluated on ImageNet. (d) DKD with our approach, evaluated on CUB200.

B Experimental Settings

Datasets and Baselines. In this study, we utilize the CIFAR-100 (Krizhevsky & Hinton, 2009), CUB200 (Welinder et al., 2010) datasets and ImageNet (Deng et al., 2009). To demonstrate SFKD's versatility, we evaluate it with multiple KD methods: vanilla KD (Hinton et al., 2015), FitNet (Romero et al., 2014), AT (Komodakis & Zagoruyko, 2017), SP (Tung & Mori, 2019), CC (Peng et al., 2019), VID (Ahn et al., 2019), CRD (Tian et al., 2019), RKD (Park et al., 2019), PKT (Passalis & Tefas, 2018), DKD (Zhao et al., 2022) and Simple Knowledge Distillation (SimKD) (Chen et al., 2022). SFKD was implemented both as a standalone approach and in conjunction with the KD loss (except vanilla KD and SimKD) to demonstrate its efficacy and versatility. Experiments were performed using renowned backbone networks such as VGG (Simonyan & Zisserman, 2014), ResNet (He et al., 2016), Wide Residual Networks (WRN) (Zagoruyko & Komodakis, 2016), MobileNet (Sandler et al., 2018), ShuffleNet (Ma et al., 2018; Zhang et al., 2018) and more advanced networks, including ConvNeXt (Liu et al., 2022) and Vision Transformers (ViTs): ViT (Dosovitskiy, 2020) and Swin (Liu et al., 2021), across a range of teacher-student model pairings. To ensure a fair comparison with baseline methods, all training settings, including learning rate, batch size, and temperature, were standardized according to the baseline configurations.

Training details: We follow the conventional experimental settings of previous works (Tian et al., 2019; Zhao et al., 2022; Sun et al., 2024a;b) for CIFAR-100 and CUB200, training models for 240 epochs, except for MLKD being 480 as in (Jin et al., 2023; Sun et al., 2024a), with the learning rate being reduced by a factor of 10 at the 150th, 180th, and 210th epochs. The initial learning rate for architectures in the MobileNet/ShuffleNet series is 0.01, while it is 0.05 for all other architectures. A batch size 64 is used, alongside a weight decay of $5{\times}10^{-4}$ and stochastic gradient descent (SGD) optimizer. All results are presented as averages from 5 trials for homogeneous (Teacher/Student) T/S pairs and 3 trials for heterogeneous T/S pairs. We run ViT-based knowledge distillation processes for 300 epochs following the training scheme in (Hao et al., 2024). In IB analysis, we train the decoder to convergence with the Adam optimizer, with learning rate set to 0.05. All models on CIFAR-100 of the paper were run on NVIDIA GeForce GTX 1080 Ti GPUs (6 GPUs). Note: the default setting is for a single-GPU training. For ImageNet, the initial learning rate is set to 0.1 and then divided by 10 at 30th, 60th, 90th of the total 120 training epochs. We conducted experiments on ImageNet using 24 NVIDIA A100 GPUs.

C More Ablation Studies and Results

More Experiments. Our approach effectively supports ViT distillation due to its model-agnostic nature, with experiments conducted on CIFAR-100 under the same conditions as (Hao et al., 2024). As shown in Table **2** of our main manuscript, our method consistently enhances KD performance across various *ViT-based, CNN-based*, and *MLP-based* (Mixer-B/16) models. Figure 6a illustrates both training and testing accuracy measurements comparing standard KD against KD enhanced with our SFKD method when distilling from ConvNeXt-T to *ViT-based student*. Additional validation on ImageNet (Deng et al., 2009) with ReviewKD (Chen et al., 2021b) was conducted on ResNet-50 to MobileNet, achieving Top-1 accuracy 73.25% with SFKD, as shown in Figure 6b. Validation on CUB200 dataset (Welinder et al., 2010), DKD with our approach in Figure 6d shows that it enhances the accuracy.

Standard Deviation for CIFAR-100 Benchmark Results. Sensitivity analyses involving a broader range of K values variability measured by standard deviation across multiple trials on the CIFAR-100 benchmark is provided in Table 7 for student and teacher models that share the same architecture, over five runs, and dissimilar architectural designs, over three runs.

Ablation study. To provide a deeper understanding of our approach, we conducted an ablation study exploring the impact of each technique both individually and in combination with KD. The results of this analysis can be found in Table 8. This analysis reveals how each technique affects the overall performance and how their interactions contribute to the final results.

Table 7: Comprehensive performance comparison on CIFAR-100. The table shows the classification accuracy (%) for various distillation methods and their performance when enhanced with SFKD. Teacher/student pairs are abbreviated for space (e.g., R32x4/R8x4 is ResNet-32x4/ResNet-8x4; MNV2 is MobileNetV2; SNV2 is ShuffleNetV2).

T/S Pair	**WRN-40-2/ WRN-40-1**	**R32x4/ R8x4**	**VGG13/ VGG8**	**VGG13/ MNV2**	**R32x4/ SNV2**
Teacher Acc.	75.61	79.42	74.64	74.64	79.42
Student Acc.	71.98	73.09	70.36	64.60	71.82
*Performance Format: Baseline Method / **Baseline + SFKD***					
KD (Hinton et al., 2015)	73.54 / **74.05**$_{\pm 0.22}$	73.33 / **74.41**$_{\pm 0.12}$	72.98 / **73.58**$_{\pm 0.23}$	67.37 / **68.30**$_{\pm 0.17}$	74.45 / **75.50**$_{\pm 0.08}$
FitNet (Romero et al., 2014)	72.24 / **72.60**$_{\pm 0.27}$	73.50 / **74.43**$_{\pm 0.26}$	71.02 / **72.35**$_{\pm 0.26}$	64.14 / **65.29**$_{\pm 0.13}$	73.54 / **75.30**$_{\pm 0.17}$
AT (Komodakis & Zagoruyko, 2017)	72.77 / **73.42**$_{\pm 0.18}$	73.44 / **73.71**$_{\pm 0.16}$	71.43 / **72.54**$_{\pm 0.32}$	59.40 / **60.82**$_{\pm 0.31}$	72.73 / **73.62**$_{\pm 0.27}$
SP (Tung & Mori, 2019)	72.43 / **73.51**$_{\pm 0.35}$	72.94 / **73.21**$_{\pm 0.07}$	72.68 / **73.23**$_{\pm 0.19}$	66.30 / **67.05**$_{\pm 0.29}$	74.56 / **76.20**$_{\pm 0.29}$
CC (Peng et al., 2019)	72.21 / **72.42**$_{\pm 0.15}$	72.97 / **73.17**$_{\pm 0.12}$	70.71 / **71.97**$_{\pm 0.30}$	64.86 / **65.58**$_{\pm 0.14}$	71.29 / **73.04**$_{\pm 0.36}$
VID (Ahn et al., 2019)	73.30 / **73.62**$_{\pm 0.18}$	73.09 / **73.39**$_{\pm 0.16}$	71.23 / **71.94**$_{\pm 0.22}$	65.56 / **65.72**$_{\pm 0.42}$	73.40 / **74.93**$_{\pm 0.07}$
RKD (Park et al., 2019)	72.22 / **72.56**$_{\pm 0.24}$	71.90 / **72.59**$_{\pm 0.28}$	71.48 / **71.66**$_{\pm 0.23}$	64.52 / **65.58**$_{\pm 0.21}$	73.21 / **74.13**$_{\pm 0.38}$
PKT (Passalis & Tefas, 2018)	73.45 / **73.83**$_{\pm 0.20}$	73.64 / **74.36**$_{\pm 0.17}$	72.88 / **73.19**$_{\pm 0.21}$	67.13 / **68.03**$_{\pm 0.20}$	74.69 / **75.84**$_{\pm 0.35}$
CRD (Tian et al., 2019)	74.14 / **74.43**$_{\pm 0.29}$	75.51 / **75.80**$_{\pm 0.19}$	73.94 / **74.08**$_{\pm 0.07}$	69.73 / **69.84**$_{\pm 0.27}$	75.65 / **76.33**$_{\pm 0.26}$
SimKD (Chen et al., 2022)	75.56 / **75.87**$_{\pm 0.21}$	78.08 / **78.53**$_{\pm 0.24}$	74.93 / **75.23**$_{\pm 0.08}$	68.95 / **70.38**$_{\pm 0.31}$	78.39 / **78.48**$_{\pm 0.13}$

Table 8: Individual and joint contributions to performance are illustrated through feature-based and logit-based combinations. The baseline method represents a feature-based approach, while KD indicates a logit-based method.

Method/T-S pair	WRN-40-2/ShuffleNetV1					
SP (Tung & Mori, 2019)	✓					
SP + SFKD			✓			
SP + KD		✓				
SP + (KD+SFKD)					✓	
(SP + SFKD) +KD				✓		
(SP + SFKD) + (KD+SFKD)						✓
	74.52	75.56	76.11	76.76	76.63	76.68

C.1 Data-Free Knowledge Distillation.

We extend our investigation to the domain of Data-Free Knowledge Distillation (DFKD) specifically to evaluate our method's robustness when dealing with potentially degraded and/or suboptimal feature maps. While our method is designed to leverage high-quality feature maps from well-trained teacher models, we recognize that such optimal conditions may not always be available in real-world applications. Through DFKD experiments, we deliberately test our approach in scenarios where feature map quality is inherently compromised due to the synthetic nature of the training data.

Using DeepInversion (DI) (Yin et al., 2020), we synthesize 100K CIFAR-10 images from teacher models VGG-11 and ResNet-34. To comprehensively assess how our method performs with these potentially degraded feature maps, we employ multiple evaluation metrics: (a) single-value measures including Inception Score (IS) (Salimans et al., 2016) and Frechet Inception Distance (FID) (Heusel et al., 2017), and (b) two-value measures such as Precision and Recall (P&R) (Sajjadi et al., 2018). These metrics help quantify both the quality degradation in synthetic data and our method's resilience to such degradation. Table 9 presents a comparative analysis between our synthesized images and those generated by WGAN-GP, a baseline GAN-based model trained on original data.

Table 9: Metric result of synthesized images. A higher score of IS, Precision and Recall is better, whereas a lower score of FID is better.

CIFAR-10				
Inverted Model	**IS ↑**	**FID ↓**	**Precision ↑**	**Recall ↑**
VGG-11	2.91	176.76	0.3824	0.0022
ResNet-34	4.21	99.79	0.5824	0.1928
WGAN-GP (Gulrajani et al., 2017)	7.86	29.30	0.7040	0.4353

C.2 More Visualizations

In Figure 7, we present visualizations comparing feature representations from models trained with our proposed distillation method (SFKD), alongside those from a teacher model, a student model trained without distillation, and CRD (Tian et al., 2019). The visual evidence in Figure 7 demonstrates that combining CRD with SFKD results in more distinct and separable features compared to the original representations, suggesting that SFKD enhances the distinguishability of deep features within the student model.

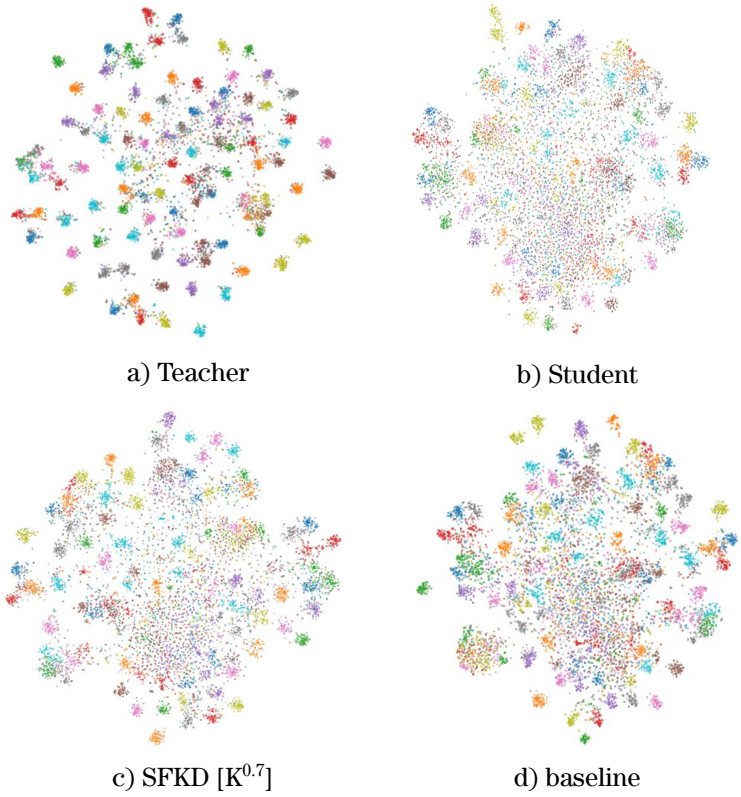


Figure 7: t-SNE clustering: demonstrating model accuracy on CIFAR-100. Points with the same color indicate they are from the same category, highlighting the model's proficiency in distinguishing between classes. A model that groups data points closely within the same class while keeping them widely separated from points of other classes demonstrates effective classification performance.