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), where only the most informative features are selectively distilled, enhancing student performance. Our approach is grounded in the Information Bottleneck (IB) principle, where focusing on features with higher mutual information with the input leads to more effective distillation. SFKD integrates with existing KD variants and enhances the transfer of "dark knowledge". It consistently improves image classification accuracy across diverse models, including ConvNeXt and ViT, achieving gains of 5.44% on CIFAR-100 and 3.57% on ImageNet-1K. When combined with current KD methods, SFKD outperforms state-of-the-art results by 1.47%.

1 Introduction

Knowledge distillation (KD) (Bucila et al., 2006; Hinton et al., 2015) captures implicit information from a teacher network to guide the training of a student network, making it a powerful approach for model compression and enhancing transfer learning. This implicit information, often termed "dark knowledge", forms the core of the distillation process. Most existing KD methods transfer the full spectrum of knowledge cues from teacher to student (Chen et al., 2021b; Komodakis & Zagoruyko, 2017; Li et al., 2022b;c; Olvera-López et al., 2010; Romero et al., 2014; Tian et al., 2019). This includes soft predictions (Hinton et al., 2015), intermediate representations (Romero et al., 2014), and attention maps (Komodakis & Zagoruyko, 2017), all distilled throughout the entire training process.

A significant challenge, however, arises from the intricate nature of the representations learned by high-capacity teacher models, which often generate high-dimensional predictive outputs, capture fine-grained details, and encode specific training instances. When transferred indiscriminately, such representations can overwhelm the student model, leading to overfitted features that hinder generalization (Ojha et al., 2023). Consequently, current approaches may suffer from knowledge redundancy, where the student passively absorbs all information without distinguishing critical knowledge from irrelevant details. This challenge parallels information redundancy problems studied in feature selection Liu et al. (2011).

Classification: Traditional KD methods primarily focus on transferring knowledge from the teacher model's probability distributions (Gou et al., 2021). However, these distributions often contain incorrect or noisy predictions, which can mislead the student model. The student may struggle to distinguish between relevant and irrelevant features, sometimes treating irrelevant information as significant (Li et al., 2022a; Song et al., 2022). This interference can degrade the model's classification performance, especially when the teacher conveys information that contradicts the true class labels. To address this, we propose a salient feature-informed classification approach. By transferring only the top-K highest confidence predictions from the teacher's softmax output, we minimize the impact of low-confidence, biased information. This selective transfer aims to capture the most valuable knowledge, enhancing the overall effectiveness of the distillation process.

Salient Feature Extraction: Existing feature-based KD methods enhance student models by leveraging both the teacher's outputs and intermediate representations as hints. We build on this by applying salient

feature masking across both intermediate feature maps and attention mechanisms, selectively distilling only the most informative elements. For attention-based architectures, we refine traditional attention transfer (AT) (Komodakis & Zagoruyko, 2017) by employing top-K selection, filtering out less significant components. This targeted distillation guides the student to focus on critical details, balancing capacity and efficiency while boosting interpretability and performance.

We introduce **S**alient **F**eature masking for **K**nowledge **D**istillation (SFKD), a method that selectively distills only the most relevant logits, feature maps, and attention maps from the teacher model. Saliency is determined by filtering out the lowest elements using a top-K selection rule (see Fig. 1). This filtering reduces bias during knowledge transfer, enhancing the quality of high-order feature representations. Using Information Bottleneck (IB) theory (Tishby & Zaslavsky, 2015a), we analyze the mutual information in the context of salient feature masking and demonstrate that selective distillation preserves higher mutual information between inputs and features, resulting in improved performance. Extensive experiments show that SFKD achieves performance nearly equivalent to the teacher model. Our contributions are summarized as follows:

1. we propose SFKD as applied to the most salient logits, feature representations, and attention maps;
2. we provide a theoretical foundation using the Information Bottleneck theory, demonstrating that higher mutual information exists between the input and filtered activation maps when less relevant values are discarded;
3. SFKD can be applied alone, or integrated with existing KD methods. In all cases, it consistently improves classification accuracy;
4. drawing from our observations and analyses, we develop a novel, simple yet effective algorithm to identify the top-K salinet feature for enhancing distillation performance.

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, 2015a), 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.

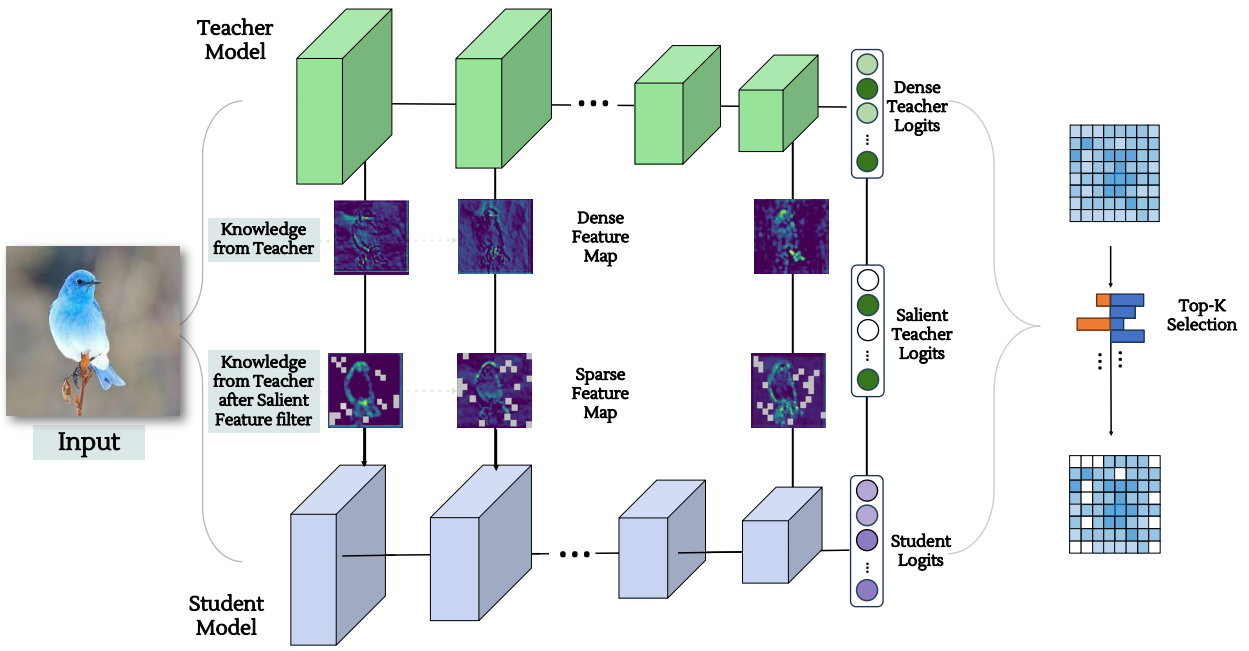


Figure 1: The concept of the proposed SFKD. SFKD distinctively concentrates on: 1) distilling critical classification knowledge, 2) transferring essential information from intermediate layers, and 3) refining attention mechanisms for knowledge distillation.

3 Methods

3.1 Formulation

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* (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:

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

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 (using additional parameters r) to match the dimensions of the teacher's features, and their similarity is then optimized by minimizing the mean squared error:

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

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:

$$L_{AT} = L_{CLS} + \frac{\beta}{2} \sum_{j \in I} \left\| \frac{Q_s^j}{\|Q_s^j\|_2} - \frac{Q_{t^K}^j}{\|Q_{t^K}^j\|_2} \right\|_p \tag{3}$$

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).

3.2 Salient Feature Masking for Selective Knowledge Sharing

In SFKD, during knowledge distillation, only the most salient values of a given feature map (attention map, logits vector) $\mathbf{F} \in \mathbb{R}^N$ from the teacher model are retained, while the remaining values are set to zero. For logits, N is the number of classes; for feature maps, it is the product of their dimensions. Let $I_{\text{top-K}}$ be the set of indices of the top-K values. The masking operator $\mathbf{M} \in \mathbb{R}^N$ can be expressed as:

$$M_i = \begin{cases} 1, & \text{if } i \in I_{\text{top-K}} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where i represents an index in $\mathbf{F}$, and M_i is the corresponding mask value at index i. The masked version of $\mathbf{F}$, denoted as $\mathbf{F}_K$, is obtained by element-wise multiplication of $\mathbf{F}$ and $\mathbf{M}$:

$$\mathbf{F}_K = \mathbf{F} \odot \mathbf{M}. \tag{5}$$

This operation zeroes out all values in $\mathbf{F}$ except for the top-K values.

3.3 Information-theoretic Analysis

Connecting Information Bottleneck with Knowledge Distillation. Let X and Y denote the input variable and the target output of a learning model, respectively. The intermediate representation F is established through an encoder $P(F|X)$ and a decoder $P(Y|F)$. To optimize the structure of a deep neural network, we leverage the Information Bottleneck (IB) principle Tishby & Zaslavsky (2015a); Shwartz-Ziv & Tishby (2017), which provides a theoretical foundation for balancing information compression and preservation. Specifically, the optimization objective can be formulated as:

$$\min_F \; I(X;F) - \zeta I(F;Y), \tag{6}$$

where $I(X;F)$ and $I(F;Y)$ denote the mutual information between X and F, and between Y and F, respectively. The parameter ζ controls the trade-off between compressing the input information and preserving the relevant information for the output.

The IB framework and interpretations have been applied to knowledge transfer in teacher-student networks. The key insight is that preserving high mutual information between the teacher and student networks is essential for effective knowledge distillation. Let F_t and F_s denote the representations of the teacher model and student model, respectively. Under the knowledge distillation framework, the optimization objective of the student model Ahn et al. (2019) is:

$$\min_s \; \{I(X;F_s) - \zeta I(Y;F_s) - \xi I(F_t;F_s)\}. \tag{7}$$

Here, $\xi > 0$ is a balancing parameter, and $I(F_t;F_s)$ denotes the mutual information between the teacher and student representations. The IB principle can be corroborated by an information plane depicting the trajectory of the points $I((X;F), I(F;Y))$ along the neural network training epochs. For the teacher-student knowledge distillation, we plot the trajectories of $(I(X;F_t), I(F_t;Y))$ and $(I(X;F_s), I(F_s;Y))$ on the information plane (see Figs. 2b -2d).

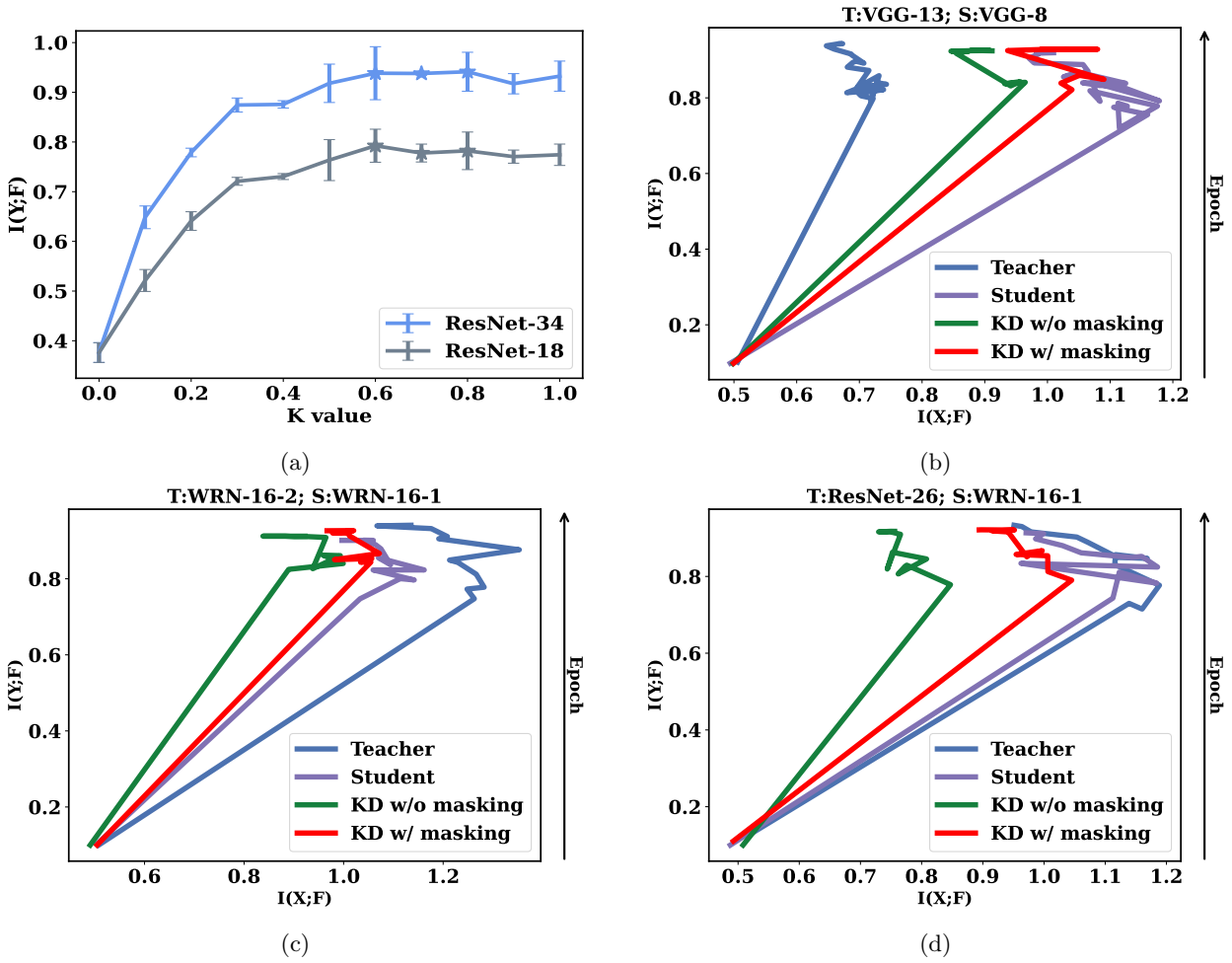


Figure 2: (a) An example of the optimal K selection for salient feature masking. The horizontal axis shows the degree of top-K sparsity ($K = 1$ meaning no masking). (b)–(d) The information plane for different teacher-student networks: the mutual information trajectories with respect to the training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

Optimal K Selection. As an example of the selecting process, mutual information trajectories for ResNet-18 and ResNet-34 are depicted in Fig. 2a under different top-K sparsity conditions. These observations have led to the following key insights: 1) The progression of all observed curves is consistent and aligns with the IB principle as delineated in (Tishby & Zaslavsky, 2015b). Notably, the trajectory of the larger model (ResNet-34) consistently lies above that of the smaller model (ResNet-18), effectively serving as an upper bound for the latter. 2) We observe a clear increase in mutual information $I(X;F)$ as sparsity decreases toward density, peaking before oscillating around a stable value. Interestingly, at complete density $K^{1.0}$, the mutual information is lower than at intermediate top-K values, indicating that top-K sparsity allows the teacher to continue distilling useful information for the student. The point where top-K sparsity stabilizes generally falls between $K^{0.5}$ and $K^{0.8}$.

The theoretical foundation and merit of the top-K masking approach can be argued through the lens of entropy H and Kullback–Leibler (KL) divergence D_{KL}. Specifically, we show that masking transformation reduces uncertainty in distilled data by filtering out less informative knowledge. This is formally established in the following proposition.

Proposition 1 (Masking concentrates distilled knowledge)**.** Let $F_t = \{p_1, p_2, \ldots, p_n\}$ and $F_{t^K} = \{q_1, q_2, \ldots, q_n\}$ be two discrete probability distributions over n states, which represent the logit outputs of the teacher model and its counterpart after top-K masking, respectively. Furthermore, let $U = \left\{\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right\}$

denote the uniform distribution over these n states. Then, we have

$$D_{\mathrm{KL}}(F_t||U) \leq D_{\mathrm{KL}}(F_{t^K}||U), \tag{8}$$

meaning F_t is more dispersed than F_{t^K}. In other words, top-K masking helps distill the concentrated knowledge.

Proof: Let $H(F_t)$ and $H(F_{t^K})$ denote the binary entropy of the two distributions F_t and F_{t^K}, respectively. Note that the uniform distribution U on a finite set has the maximum entropy ($\log n$) among all discrete distributions on that set Cover & Thomas (2006). In addition, the KL divergence from any distribution F to U is given as: $D_{\mathrm{KL}}(F||U) = \log n - H(F)$. For F_t and F_{t^K}, we get:

$$D_{\mathrm{KL}}(F_t||U) = \log n - H(F_t), \tag{9}$$

$$D_{\mathrm{KL}}(F_{t^K}||U) = \log n - H(F_{t^K}). \tag{10}$$

Since post-processing reduces entropy (Cover & Thomas, 2006, pg. 44), we have $H(F_{t^K}) = H(M_K(F_t)) \leq H(F_t)$, where $M_K(F_t)$ is the top-K masking operator. Thus, it follows that $D_{\mathrm{KL}}(F_t||U) \leq D_{\mathrm{KL}}(F_{t^K}||U)$: F_t is closer to the uniform distribution U than F_{t^K} in the sense of KL divergence.

IB-based Analysis of SFKD. We establish a theoretical foundation for our proposed top-K masking mechanism in SFKD by linking KD with the IB theory. Our analysis shows that during deep network training, the mutual information $I(Y;F)$ with the labels consistently increases, while $I(X;F)$ with the input data initially rises but then declines. While maximizing $I(Y;F)$ aids the teacher model, it is less crucial for KD since the ground truth labels are already incorporated into the distillation objective. Instead, preserving $I(X;F)$ captures "dark knowledge", which is vital for effective KD. For instance, an image labeled "camel, horse, car" may still encode features relevant to "people", offering valuable information for the student model.

Existing KD methods often utilize high temperatures to soften network predictions, exposing such information from the teacher model. However, IB theory demonstrates that a full feature knowledge transferred from teacher to student tends to be overconfident, making it challenging to recover suppressed knowledge by only scaling the temperature. In contrast, selective features, despite sparse information due to non-optimal $I(Y;F)$, may possess a larger $I(X;F)$ and improve KD performance. This insight corroborates our observation that selected features serve as valuable and more precise information.

Mutual Information Estimation. Estimating mutual information is challenging, as it often requires inferring complex probability densities from limited data, complicating accurate dependence measurement, especially for continuous or high-dimensional data. This limitation introduces potential bias, particularly when true underlying distributions are unknown. In pursuit of approximating the reconstruction loss, we integrate a decoder with the last convolutional layer of a pre-trained and static network model, aiming to reconstruct a pseudo input image, denoted as $\overline{X}$. Subsequently, this decoder undergoes training until convergence using the Adam optimizer, coupled with a binary cross-entropy (BCE) loss between $\overline{X}$ and the original input X. This method of calculating reconstruction loss serves as a proxy for estimating $I(X;F)$, adhering to the methodology outlined in Ref. (Wang et al., 2021). More detailed settings are provided in the supplementary material (Sec. A).

Our motivation for SFKD aligns with the observed phenomenon, as illustrated in Figs. 2b -2d. Specifically, we find that: 1) Compared with the independent student case, masked or unmasked KD always facilitates faster learning of the student with steeper initial trajectories, which yields higher values of $I(Y;F)$. This phenomenon is consistently observed for various teacher-student networks. 2) Masked KD (SFKD) has higher values of $I(X;F)$ than the unmasked KD throughout the training epochs. This implies masked KD tends to compress less information from the input X, which can improve the student's performance. 3) As shown in Fig. 2b, when the teacher and student have the same type of network, mutual information $I(X;F)$ decreases with the depth of layers, which is justified by the data processing inequality (DPI) (Cover & Thomas, 2006, pg. 34).

4 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.

Table 1: Classification accuracy of different pairs of teacher (T) and student model (S) with various knowledge distillation methods on the CIFAR-100 dataset. The table is divided into two sections: the ***left*** section presents results for homogeneous T/S model pairs, averaged over 5 runs, while the ***right*** section details the classification accuracy for heterogeneous T/S model pairs, with averages taken over 3 runs.

T/S Pair	WRN40-2/ WRN16-2(%)		WRN40-2/ WRN40-1(%)		ResNet-32x4/ ResNet-8x4 (%)		VGG-13/ VGG-8 (%)		VGG-13/ MobileNetV2 (%)		ResNet-32x4/ ShuffleNetV1 (%)		ResNet-32x4/ ShuffleNetV2 (%)		WRN-40-2/ ShuffleNetV1 (%)	
Teacher	75.61		75.61		79.42		74.64		74.64		79.42		79.42		75.61	
Student	73.26		71.98		73.09		70.36		64.60		70.50		71.82		70.50	
Method	baseline	SFKD	baseline	SFKD	baseline	SFKD	baseline	SFKD	baseline	SFKD	baseline	SFKD	baseline	SFKD	baseline	SFKD
KD (Hinton et al., 2015)	74.92	75.39	73.54	74.05	73.33	74.41	72.98	73.58	67.37	68.30	74.07	74.73	74.45	75.50	74.83	76.16
FitNet (Romero et al., 2014)	73.58	73.80	72.24	72.60	73.50	74.43	71.02	72.35	64.14	65.29	73.59	74.44	73.54	75.30	73.73	74.99
AT (Komodakis & Zagoruyko, 2017)	74.08	74.58	72.77	73.42	73.44	73.71	71.43	72.54	59.40	60.82	71.73	73.42	72.73	73.62	73.32	74.31
SP (Tung & Mori, 2019)	73.83	74.88	72.43	73.51	72.94	73.21	72.68	73.23	66.30	67.05	73.48	76.05	74.56	76.20	74.52	76.11
CC (Peng et al., 2019)	73.56	73.73	72.21	72.42	72.97	73.17	70.71	71.97	64.86	65.58	71.14	71.84	71.29	73.04	71.38	72.07
VID (Ahn et al., 2019)	74.11	74.63	73.30	73.62	73.09	73.39	71.23	71.94	65.56	65.72	73.38	74.19	73.40	74.93	73.61	75.05
RKD (Park et al., 2019)	73.35	73.76	72.22	72.56	71.90	72.59	71.48	71.66	64.52	65.58	72.28	73.15	73.21	74.13	72.21	73.53
PKT (Passalis & Tefas, 2018)	74.54	75.26	73.45	73.83	73.64	74.36	72.88	73.19	67.13	68.03	74.10	75.02	74.69	75.84	73.89	75.29
CRD (Tian et al., 2019)	75.48	75.76	74.14	74.43	75.51	75.80	73.94	74.08	69.73	69.84	75.11	75.84	75.65	76.33	76.05	76.36
SimKD (Chen et al., 2022)	76.23	76.53	75.56	75.87	78.08	78.53	74.93	75.23	68.95	70.38	77.18	77.82	78.39	78.48	77.09	77.64

4.1 Experiment Settings

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. / 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 (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+Ours	**80.62**$_{+1.88}$	**78.90**$_{+1.64}$	**79.18**$_{+1.39}$	**78.87**$_{+2.43}$

Datasets and Baselines. In this study, we utilize the CIFAR-10/100 (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. Further details can be found in suppl. mat.

4.2 Results

Results on CIFAR-100. Tab. 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 Tab. 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]. Moreover, as illustrated in Tab. 4, our SFKD outperforms most of the previous SOTA methods on CIFAR-100 dataset. 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 Tab. 6. This analysis reveals how each technique affects the overall performance and how their interactions contribute to the final results.

Results on ImageNet. We report Top-1 and Top-5 accuracies (%) for our proposed method, referred to as "Ours" in Tab. 3. The results of baseline methods taken from the original papers report (Chen et al., 2021b; Tian et al., 2019). Experiments' outcome demonstrates the scalability of our method accross a broader range of dataset sizes.

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
LS (MLKD+LS) (Sun et al., 2024)	72.08	90.74	73.22	91.59
AT (Komodakis & Zagoruyko, 2017)	70.69	90.01	69.56	89.33
AT+Ours	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+Ours	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+Ours	**72.10**$_{+0.4}$	**90.70**$_{+0.29}$	**72.95**$_{+0.9}$	**91.30**$_{+0.25}$

Results on CUB200. CUB200 (Welinder et al., 2010) is utilized for assessing fine-grained classification tasks, which consist of 200 different bird species. The results are presented in Tab. 5.

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

It would be advantageous to apply our method when data comes not only from a single teacher but multiple teachers. Unlike conventional multi-teacher KD methods (Du et al., 2020; You et al., 2017; Fukuda et al., 2017; Wu et al., 2019) that might dilute the specificity of knowledge due to averaging or varying confidence levels among teachers, our SFKD method stands out by selectively channeling the most pertinent insights from the teacher(s). This selective sharing mechanism makes SFKD especially suited for distillation contexts involving multiple teachers, as evidenced by the results presented in Tab. 7. SFKD's superior accuracy further validates its utility in enhancing the quality and efficiency of knowledge distillation. We apply our selective knowledge sharing technique to "AEKD" (Du et al., 2020) and its enhanced version "AEKD-F", which respectively aggregate teacher predictions using an adaptive weighting strategy and incorporate

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

Table 4: Comparison with recent SOTA methods on the CIFAR-100 dataset. **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
LS (Sun et al., 2024) (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 5: 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**$_{+4.78}$

intermediate features. As shown in Tab. 7, SFKD always achieves the best performance. Tri-ResNet-32x4, which includes three ResNet-32x4 models, serves as a teacher for both students VGG-8 and ShuffleNetV2.

4.4 Application 2: Salient Feature Masking in Data-Free Knowledge Distillation

Leveraging SFKD in the domain of Data-Free Knowledge Distillation (DFKD) (Chen et al., 2019; Yin et al., 2020) offers a strategic enhancement, especially when navigating the challenges of training without access to original datasets. Our methodology, centered around the selective transmission of highly informative features, is particularly beneficial in these settings. SFKD's selective nature preserves the integrity of the knowledge transferred, filtering out noise and enhancing the quality of synthetic data. Such a targeted method of knowledge transfer is instrumental in ensuring that the student model is primed to generalize effectively to new, unseen data, a critical advantage in practical applications where the model encounters data variants not represented in the synthetic training set.

Table 6: 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 + ours			✓			
SP + KD		✓				
SP + (KD+ours)					✓	
(SP + ours) +KD				✓		
(SP + ours) + (KD+ours)						✓
	74.52	75.56	76.11	76.76	76.63	76.68

Table 7: 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 8: 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%**

We synthesized 100K images using the DeepInversion (DI) technique (Yin et al., 2020) for DFKD, generating synthetic data from VGG-11 and ResNet-34 models trained on CIFAR-10. Tab. 8 shows that SFKD improves all teacher-student pair combinations in DFKD. Our findings highlight the benefits of SFKD in improving model training strategies and synthetic data utilization, enhancing model interpretability and effectiveness.

4.5 Visualization

Visualization of Class Activation Maps. A Class Activation Map (CAM) (Zhou et al., 2016) serves as a visualization technique that displays the activation regions for specific categories within neural networks, thereby aiding in understanding how these networks classify images across different categories. In Fig. 3, the first row presents the input images, while the second, third, and fourth rows show the class activation maps of the teacher model, baseline AT (K^1), and SFKD ($K^\star$) respectively. $\star$ denotes the optimal K value to precise. In line 3, it is observed that the model with baseline AT does not focus exclusively on the car,

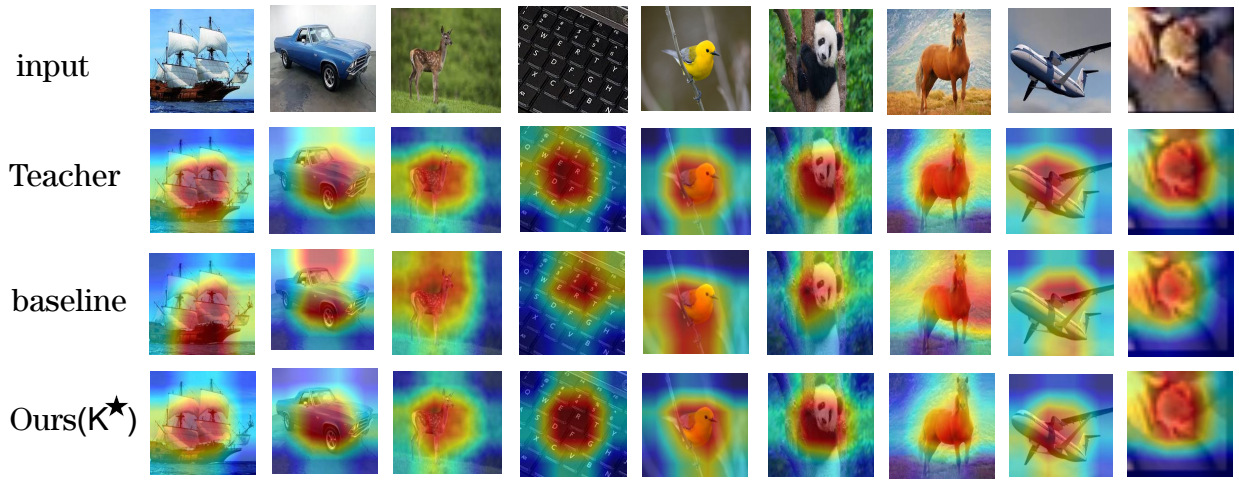
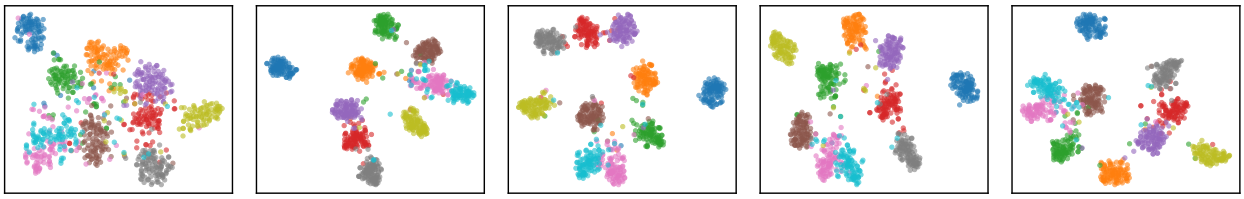


Figure 3: Class activation map of the distilled student model deployed with our method and baseline AT, the teacher model. The deeper the color, the more salient the corresponding feature of the image. The top row presents the input images, while the second, third, and fourth rows display the class activation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



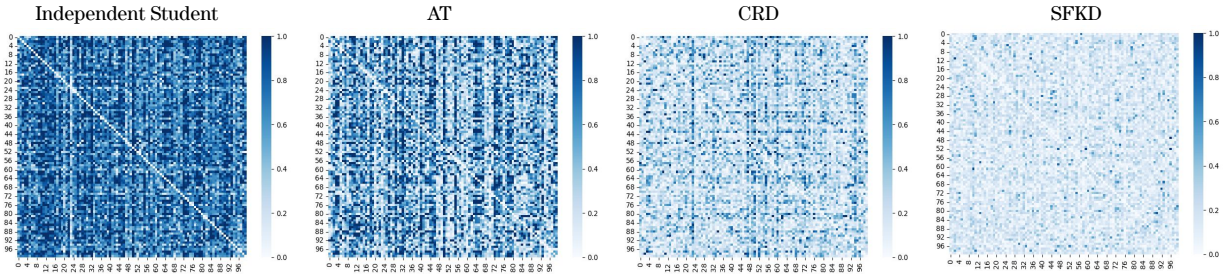
(a) Student (b) SFKD ($K^{0.3}$) (c) SFKD ($K^{0.5}$) (d) SFKD ($K^{0.7}$) (e) baseline CRD

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.

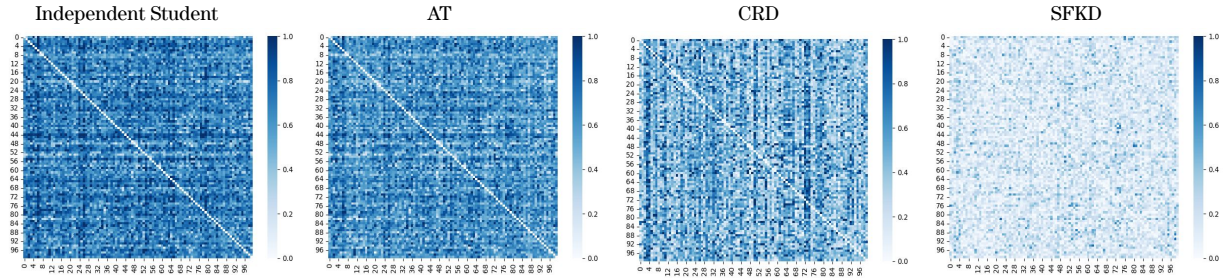
bird, and horse categories, but rather on outer areas unrelated to the target. Conversely, in line 4 of Fig. 3, the student models trained with AT+SFKD are not distracted; their attention is fully concentrated on the car, bird and horse area in the image almost acting like a teacher model, demonstrating that our proposed methods enhance the focus on deep features. Since our SFKD is method-agnostic, we can choose any of the previous KD techniques as a baseline. In this case, ResNet-32x4 as a teacher & ResNet-8x4 as a student, and AT as a baseline was used.

Visualization with t-SNE. t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten & Hinton, 2008) provides a clear view of the image processing workflow in student models. A tight clustering within classes and a distinct separation between classes indicate the model's strong classification capabilities. We present visualizations from an independent student model, baseline CRD, and our approach SFKD with varying K. Fig. 4 shows experiments on CIFAR-100 using ResNet-32x4 → ResNet-8x4, comparing feature distributions of student networks. It is observed that the features of the student trained with SFKD ($K^{0.3}$, $K^{0.5}$), as shown in Fig. 4(b, c), are more distinguishable among different classes compared to those of the student trained from scratch. This indicates that precise selective teacher knowledge can effectively enhance the discrimination ability of the student network. Moreover, compared to the baseline method, our SFKD approach results in more compact feature clusters within the same classes, as exemplified by the blue, green, light green, and orange clusters in the figure. We calculate the L1 error between the teacher's and student's

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 (Komodakis & Zagoruyko, 2017), CRD (Tian et al., 2019), 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. We also provide an additional evaluation of our method in the suppl. mat.



(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.

5 Conclusion and Discussion

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.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*, 2006.

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.

Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *International Conference on Computer Vision*, pp. 3513–3521, 2019.

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.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ, 2 edition, 2006. ISBN 978-0471241959.

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.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

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.

Chenxin Li, Mingbao Lin, Zhiyuan Ding, Nie Lin, Yihong Zhuang, Yue Huang, Xinghao Ding, and Liujuan Cao. Knowledge condensation distillation. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2022a.

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.

S. Li, M. Lin, Y. Wang, C. Fei, L. Shao, and R. Ji. Learning efficient gans for image translation via differentiable masks and co-attention distillation. *IEEE Transactions on Multimedia (TMM)*, 2022b.

S. Li, M. Lin, Y. Wang, Y. Wu, Y. Tian, L. Shao, and R. Ji. Distilling a powerful student model via online knowledge distillation. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2022c.

Si Liu, Hairong Liu, Longin Jan Latecki, Shuicheng Yan, Changsheng Xu, and Hanqing Lu. Size adaptive selection of most informative features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pp. 392–397, 2011.

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.

J. A. Olvera-López, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, and J. Kittler. A review of instance selection methods. *Artificial Intelligence Review*, 34(2):133–143, 2010.

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.

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.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11): 8135–8153, 2022.

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, 2024.

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)*, 2015a.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *ITW*, 2015b.

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.

Yulin Wang, Zanlin Ni, Shiji Song, Le Yang, and Gao Huang. Revisiting locally supervised learning: an alternative to end-to-end training. In *ICLR*, 2021.

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.

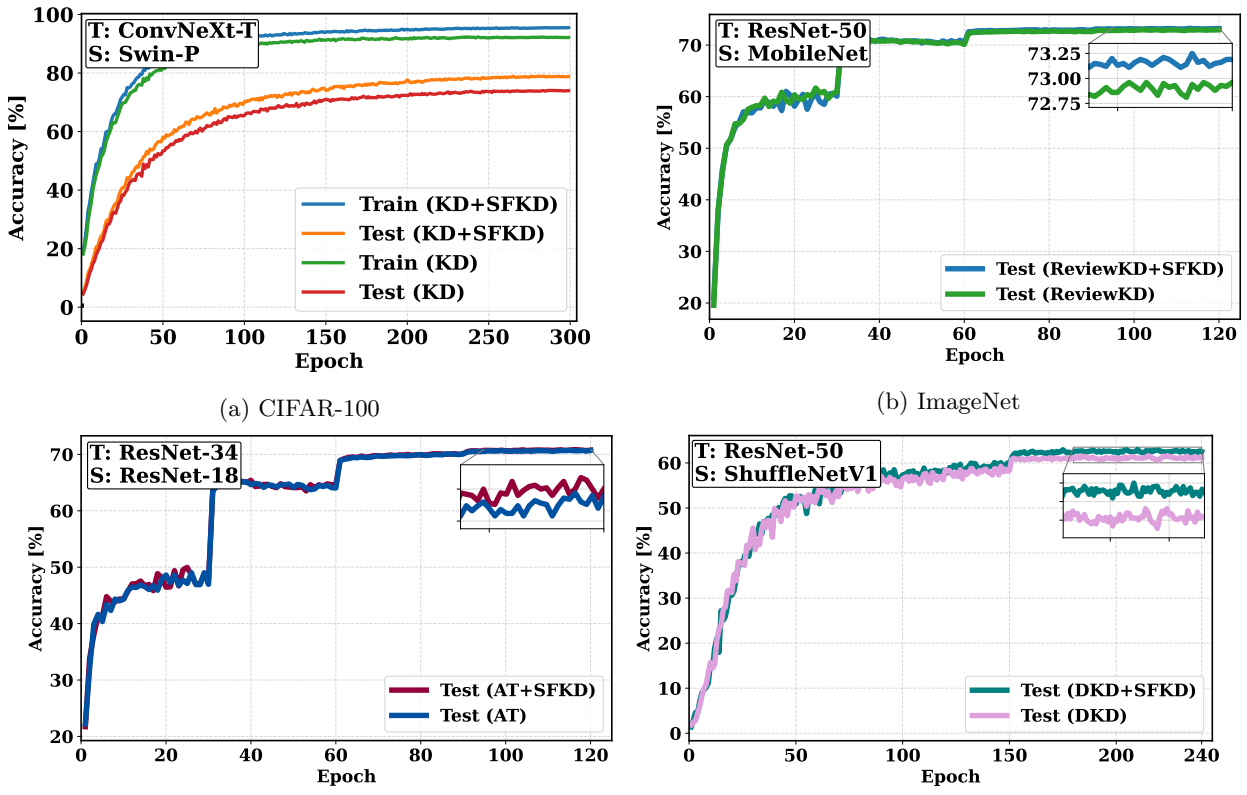
Appendix

This supplementary document details mutual information estimation for $I(X;F)$ and $I(Y;F)$ (Sec. A), complete training setup with hyperparameters for CIFAR-100, CUB200, and ImageNet (Sec. B), and additional experiments including ablations, CIFAR-10 evaluation, data-free knowledge distillation results, and representation visualizations (Sec. 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 Training Details.

The hyperparameters for the baselines (when K^1) are aligned with those in the original papers, as indicated in Tab. 9. The experimental setup for CIFAR-100 and CUB200 is identical to that used in CRD (Tian et al., 2019); the training lasts 240 epochs, except for MLKD being 480 as in (Jin et al., 2023; Sun et al., 2024), 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 5×10^{-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. Throughout this paper, the temperature τ for the KD loss is consistently set at 4. 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.

Table 9: Hyperparameter settings of baseline distillation methods. β is the weight balance of distillation loss in the baselines.

Methods	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)	RKD (Park et al., 2019)	PKT (Passalis & Tefas, 2018)	CRD (Tian et al., 2019)	SimKD (Chen et al., 2022)
β	0	100	1000	3000	0.02	1	1	30000	0.9	1

C More Ablation Studies and Results

More Experiments. Our approach effectively supports ViT distillation due to its model-agnostic nature, with experiments conducted on CIFAR100 under the same conditions as (Hao et al., 2024). As shown in Tab. **2** of our main manuscript, our method consistently enhances KD performance across various *ViT-based, CNN-based*, and *MLP-based* (Mixer-B/16) models. Fig. 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 Fig. 6b. Validation on CUB200 dataset (Welinder et al., 2010), DKD with our approach in Fig. 6d shows that it enhances the accuracy.

Results on CIFAR-10. For CIFAR-10, we run three scenarios with varying network architectures for both student and teacher networks: the first two experiments used WRNs, the next two used ResNets, and in the final experiment, the teacher and student networks had different architectures.

Table 10: Top-1 accuracy (%) of various knowledge distillation methods on CIFAR-10.

Distillation type				Logits		Attention		Features	
Teacher	Student	Teacher	Student	KD (Hinton et al., 2015)	**Ours**	AT (Komodakis & Zagoruyko, 2017)	**Ours**	SP (Tung & Mori, 2019)	**Ours**
WRN-16-2	WRN-16-1	93.94	91.26	92.06	**92.95**	91.72	**92.68**	92.48	**92.83**
WRN-40-2	WRN-16-2	95.07	93.94	94.00	**94.54**	94.11	**94.84**	94.48	**94.66**
ResNet-26	ResNet-8	93.40	87.74	88.75	**89.09**	88.15	**89.0**	88.94	**89.19**
ResNet-26	ResNet-14	93.40	91.51	92.57	**92.84**	92.11	**92.84**	92.55	**92.61**
ResNet-26	WRN-16-1	93.40	91.26	92.43	**92.89**	91.32	**92.9**	92.47	**92.74**

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 Tab. 11 for student and teacher models that share the same architecture, over five runs, and dissimilar architectural designs, over three runs. The variance of the baseline has been omitted due to space constraints. Additional ablation study by combining distillation methods with KD, demonstrating the compatibility of our objective is provided in Tab. 13.

Table 11: Classification accuracy of different pairs of the teacher model (T) and the student model (S) with various knowledge distillation methods on the CIFAR-100 dataset. The table is divided into two sections: the ***left*** section presents results for homogeneous T/S model pairs, averaged over 5 runs, while the ***right*** section details the classification accuracy for heterogeneous T/S model pairs, with averages taken over 3 runs.

T/S Pair	WRN40-2/ WRN16-2(%)		WRN40-2/ WRN40-1(%)		ResNet-32x4/ ResNet-8x4 (%)		VGG-13/ VGG-8 (%)		VGG-13/ MobileNetV2 (%)		ResNet-32x4/ ShuffleNetV1 (%)		ResNet-32x4/ ShuffleNetV2 (%)		WRN-40-2/ ShuffleNetV1 (%)	
Teacher	75.61		75.61		79.42		74.64		74.64		79.42		79.42		75.61	
Student	73.26		71.98		73.09		70.36		64.60		70.50		71.82		70.50	
Method	baseline	SFKD	baseline	SFKD	baseline	SFKD	baseline	SFKD	baseline	SFKD	baseline	SFKD	baseline	SFKD	baseline	SFKD
KD (Hinton et al., 2015)	74.92	$75.39_{\pm 0.29}$	73.54	$74.05_{\pm 0.22}$	73.33	$74.41_{\pm 0.12}$	72.98	$73.58_{\pm 0.23}$	67.37	$68.30_{\pm 0.17}$	74.07	$74.73_{\pm 0.17}$	74.45	$75.50_{\pm 0.08}$	74.83	$76.16_{\pm 0.15}$
FitNet (Romero et al., 2014)	73.58	$73.80_{\pm 0.13}$	72.24	$72.60_{\pm 0.27}$	73.50	$74.43_{\pm 0.26}$	71.02	$72.35_{\pm 0.26}$	64.14	$65.29_{\pm 0.13}$	73.59	$74.44_{\pm 0.13}$	73.54	$75.30_{\pm 0.17}$	73.73	$74.99_{\pm 0.34}$
AT (Komodakis & Zagoruyko, 2017)	74.08	$74.58_{\pm 0.30}$	72.77	$73.42_{\pm 0.18}$	73.44	$73.71_{\pm 0.16}$	71.43	$72.54_{\pm 0.32}$	59.40	$60.82_{\pm 0.31}$	71.73	$73.42_{\pm 0.13}$	72.73	$73.62_{\pm 0.27}$	73.32	$74.31_{\pm 0.19}$
SP (Tung & Mori, 2019)	73.83	$74.88_{\pm 0.30}$	72.43	$73.51_{\pm 0.35}$	72.94	$73.21_{\pm 0.07}$	72.68	$73.23_{\pm 0.19}$	66.30	$67.05_{\pm 0.29}$	73.48	$76.05_{\pm 0.32}$	74.56	$76.20_{\pm 0.29}$	74.52	$76.11_{\pm 0.13}$
CC (Peng et al., 2019)	73.56	$73.73_{\pm 0.22}$	72.21	$72.42_{\pm 0.15}$	72.97	$73.17_{\pm 0.12}$	70.71	$71.97_{\pm 0.30}$	64.86	$65.58_{\pm 0.14}$	71.14	$71.84_{\pm 0.16}$	71.29	$73.04_{\pm 0.36}$	71.38	$72.07_{\pm 0.22}$
VID (Ahn et al., 2019)	74.11	$74.63_{\pm 0.29}$	73.30	$73.62_{\pm 0.18}$	73.09	$73.39_{\pm 0.16}$	71.23	$71.94_{\pm 0.22}$	65.56	$65.72_{\pm 0.42}$	73.38	$74.19_{\pm 0.28}$	73.40	$74.93_{\pm 0.07}$	73.61	$75.05_{\pm 0.21}$
RKD (Park et al., 2019)	73.35	$73.76_{\pm 0.20}$	72.22	$72.56_{\pm 0.24}$	71.90	$72.59_{\pm 0.28}$	71.48	$71.66_{\pm 0.23}$	64.52	$65.58_{\pm 0.21}$	72.28	$73.15_{\pm 0.19}$	73.21	$74.13_{\pm 0.38}$	72.21	$73.53_{\pm 0.39}$
PKT (Passalis & Tefas, 2018)	74.54	$75.26_{\pm 0.14}$	73.45	$73.83_{\pm 0.20}$	73.64	$74.36_{\pm 0.17}$	72.88	$73.19_{\pm 0.21}$	67.13	$68.03_{\pm 0.20}$	74.10	$75.02_{\pm 0.03}$	74.69	$75.84_{\pm 0.35}$	73.89	$75.29_{\pm 0.29}$
CRD (Tian et al., 2019)	75.48	$75.76_{\pm 0.18}$	74.14	$74.43_{\pm 0.29}$	75.51	$75.80_{\pm 0.19}$	73.94	$74.08_{\pm 0.07}$	69.73	$69.84_{\pm 0.27}$	75.11	$75.84_{\pm 0.27}$	75.65	$76.33_{\pm 0.26}$	76.05	$76.36_{\pm 0.39}$
SimKD (Chen et al., 2022)	76.23	$76.53_{\pm 0.31}$	75.56	$75.87_{\pm 0.21}$	78.08	$78.53_{\pm 0.24}$	74.93	$75.23_{\pm 0.08}$	68.95	$70.38_{\pm 0.31}$	77.18	$77.82_{\pm 0.15}$	78.39	$78.48_{\pm 0.13}$	77.09	$77.64_{\pm 0.10}$

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. Tab. 12 presents a comparative analysis between our synthesized images and those generated by WGAN-GP, a baseline GAN-based model trained on original data.

Table 12: 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 Fig. 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 Fig. 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.

Table 13: Test accuracy (%) of student networks on CIFAR-100 by combining distillation methods with KD, demonstrating the compatibility of our objective.

T/S Pair	WRN40-2/ WRN16-2 (%)		WRN40-2/ WRN40-1 (%)		ResNet-32x4/ ResNet-8x4 (%)		VGG-13/ VGG-8 (%)	
Teacher	75.61		75.61		79.42		74.64	
Student	73.26		71.98		73.09		70.36	
Method	baseline	SFKD	baseline	SFKD	baseline	SFKD	baseline	SFKD
KD	74.92	75.39	73.54	74.05	73.33	74.41	72.98	73.58
KD + FitNet	75.75	75.96	74.12	74.48	74.31	75.26	73.54	73.70
KD + AT	75.28	75.73	74.45	74.80	74.26	75.43	73.62	73.70
KD + SP	75.34	75.56	73.15	74.15	74.74	75.13	73.44	73.64
KD + VID	74.79	75.32	74.20	74.60	74.82	75.36	73.96	74.18
KD + RKD	75.40	75.77	73.87	74.06	74.47	75.05	73.72	73.94
KD + PKT	76.01	76.09	74.40	74.61	74.17	74.66	73.37	73.64

T/S Pair	VGG-13/ MobileNetV2 (%)		ResNet-32x4/ ShuffleNetV1 (%)		ResNet-32x4/ ShuffleNetV2 (%)		WRN-40-2/ ShuffleNetV1 (%)	
Teacher	74.64		79.42		79.42		75.61	
Student	64.60		70.50		71.82		70.50	
KD	67.37	68.30	74.07	74.73	74.45	75.50	74.83	76.16
KD + FitNet	68.58	68.86	74.82	75.37	75.11	76.14	75.55	76.06
KD + AT	69.34	69.62	74.76	76.40	75.30	76.70	75.61	76.61
KD + SP	66.89	68.19	73.80	75.88	75.15	76.41	75.56	76.76
KD + VID	66.91	68.49	74.28	75.32	75.78	76.15	75.36	76.82
KD + RKD	68.50	69.03	74.20	75.19	75.74	75.98	75.45	76.40
KD + PKT	67.89	68.89	74.06	75.18	75.18	76.21	75.51	76.35

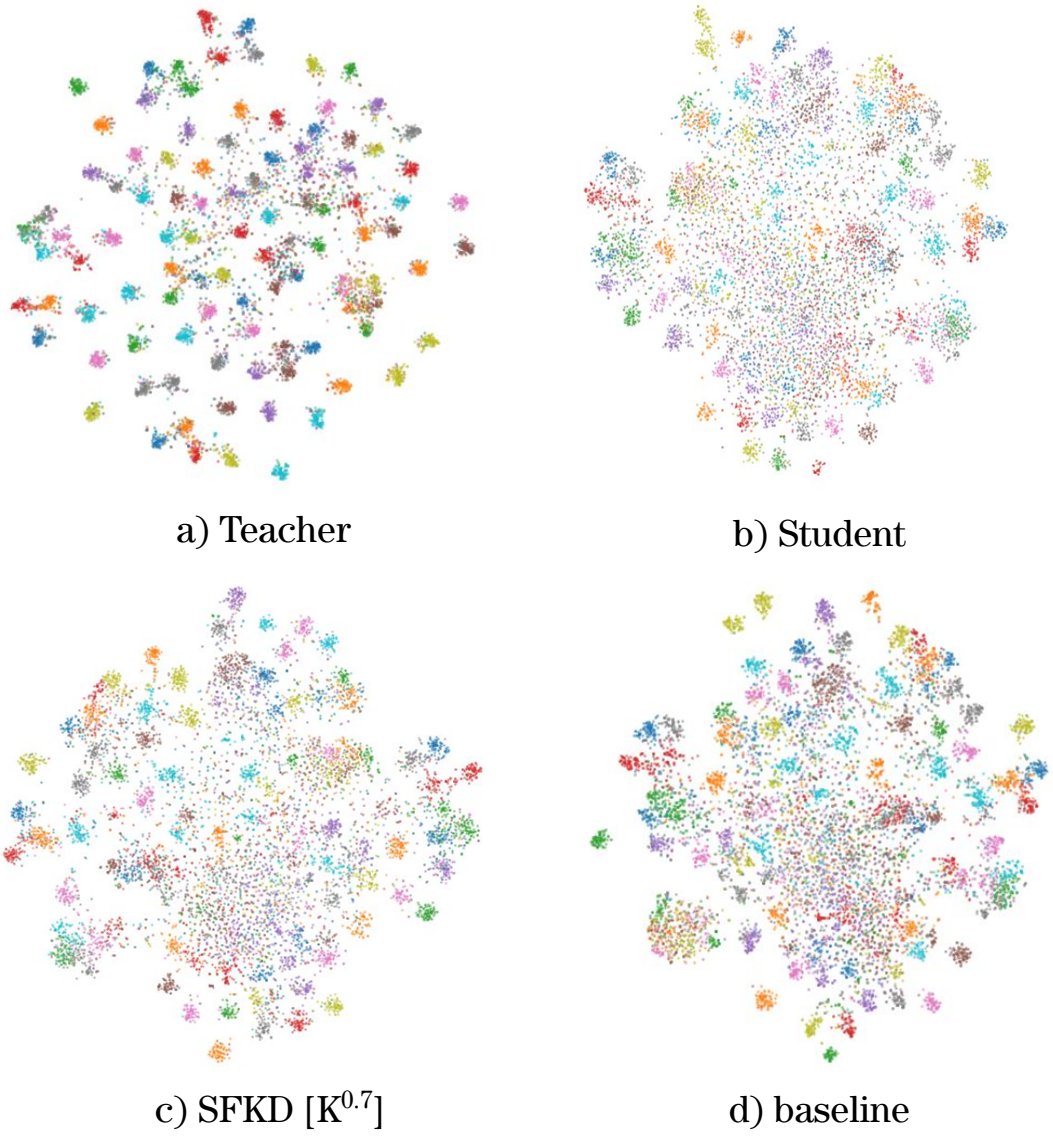


Figure 7: t-SNE clustering: demonstrating model accuracy on CIFAR-100. Points with the same color indicate they are from the same category, highlighting the model's proficiency in distinguishing between classes. A model that groups data points closely within the same class while keeping them widely separated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