electronics

Article

Bi-Level Orthogonal Multi-Teacher Distillation

Shuyue Gong and Weigang Wen *

School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China; 21222005@bjtu.edu.cn

* Correspondence: wgwen@bjtu.edu.cn

Abstract: Multi-teacher knowledge distillation is a powerful technique that leverages diverse information sources from multiple pre-trained teachers to enhance student model performance. However, existing methods often overlook the challenge of effectively transferring knowledge to weaker student models. To address this limitation, we propose BOMD (Bi-level Optimization for Multi-teacher Distillation), a novel approach that combines bi-level optimization with multiple orthogonal projections. Our method employs orthogonal projections to align teacher feature representations with the student's feature space while preserving structural properties. This alignment is further reinforced through a dedicated feature alignment loss. Additionally, we utilize bi-level optimization to learn optimal weighting factors for combining knowledge from heterogeneous teachers, treating the weights as upper-level variables and the student's parameters as lower-level variables. Extensive experiments on multiple benchmark datasets demonstrate the effectiveness and flexibility of BOMD. Our method achieves state-of-the-art performance on the CIFAR-100 benchmark for multi-teacher knowledge distillation across diverse scenarios, consistently outperforming existing approaches. BOMD shows significant improvements for both homogeneous and heterogeneous teacher ensembles, even when distilling to compact student models.

Keywords: knowledge distillation; deep learning; convolutional neural networks; teacher-student model; optimization; multi-model learning; soft labeling; supervised learning

1. Introduction

In recent years, deep learning models have achieved remarkable success across various domains, pushing the boundaries of artificial intelligence. However, the increasing complexity and size of these models have led to significant computational challenges [1–7] . State-of-the-art deep neural networks often contain millions or even billions of parameters, requiring substantial computational resources for training and inference. This trend has created a pressing need for efficient model compression techniques that can reduce model size and computational requirements while maintaining performance. Among the various approaches to model compression, Knowledge Distillation (KD) [8] has emerged as a promising method. This technique aims to transfer the knowledge from a large, complex teacher model to a smaller, more efficient student model. In recent years, there has been a growing interest in multi-teacher methods of knowledge distillation, where multiple teacher models are employed to provide diverse sources of information to enhance the learning of the student model. In this review, we discuss the advancements and effectiveness of multi-teacher methods in knowledge distillation [9,10].

One of the key motivations behind using multiple teachers is to leverage the diverse knowledge sources they possess. Each teacher model may have been trained on different datasets, architectures, or with various regularization techniques, leading to different areas of expertise. By combining the knowledge from multiple teachers, the student model can benefit from a more comprehensive understanding of the data distribution and improve its generalization ability. Existing multi-teacher methods [11] have explored different strategies to effectively utilize the knowledge from multiple teachers. Weighted ensemble approaches

check for updates

Citation: Gong, S.; Wen, W. Bi-Level Orthogonal Multi-Teacher Distillation. *Electronics* **2024**, *13*, 3345. https://doi.org/10.3390/electronics13163345

Received: 12 July 2024
Revised: 2 August 2024
Accepted: 15 August 2024
Published: 22 August 2024

assign different weights to each teacher's predictions, either learned or based on heuristics, to optimize the knowledge transfer process. These methods [12] aim to find the optimal combination of teacher models, considering their expertise and reliability. Ensemble techniques, such as model averaging or knowledge fusion, have also been employed to aggregate the knowledge from multiple teachers. Another aspect that has been explored in multi-teacher methods is the introduction of diversity in the training process. Teachers can be trained with different objectives, such as incorporating auxiliary tasks or applying different regularization techniques, to provide complementary knowledge to the student model. By encouraging diversity among the teachers, the student model can provide a more robust representation and capture a broader range of information. Furthermore, recent advancements [13] have focused on addressing the limitations of existing multi-teacher methods. Some approaches have introduced techniques to handle the imbalance between teachers or to adaptively select the most informative teachers for each training instance. Others have proposed methods to incorporate self-distillation, where the student model distills knowledge from other teachers or even itself, creating a self-supervised learning loop that further enhances the knowledge transfer process. However, challenges still exist in multi-teacher knowledge distillation. Determining the optimal number of teachers, designing effective weighting strategies, and managing the computational complexity associated with multiple teachers are areas that require further investigation.

To address this challenge, we propose Bi-Level Orthogonal Multi-Teacher Distillation (BOMD), a novel approach that utilizes bi-level optimization and multiple orthogonal projections. To effectively leverage the knowledge from an ensemble of diverse teacher models during distillation, we propose a novel methodology that combines two essential techniques. First, we employ multiple orthogonal projections to align the direct feature representations from different teacher models with the student model's feature space. Orthogonal projections preserve structural properties and relationships between features, enhancing the retention of relevant knowledge during transfer. We introduce orthogonal projection matrices, one for each teacher, which project the teacher's features into the student's feature space while satisfying orthogonality conditions to maintain angle and norm preservation. A feature alignment loss is minimized to encourage the projected teacher features to align with the student's representations. Second, we optimize the weighting factors that combine the knowledge from multiple teachers using a bi-level optimization approach. Unlike heuristic weighting strategies, bi-level optimization directly learns an optimal weighting strategy tailored to the specific characteristics of the teacher ensemble and student model. This approach treats the weighting factors as upper-level variables and the student's parameters as lower-level variables in a nested optimization problem, allowing for effective knowledge transfer even with heterogeneous architectures and diverse complementary knowledge sources.

Through extensive experiments on multiple benchmark datasets, we validate the effectiveness and flexibility of our approach. Our proposed BOMD method achieves state-of-the-art performance on the CIFAR-100 benchmark for multi-teacher knowledge distillation across diverse teacher–student scenarios. In experiments involving homogeneous teacher ensembles, BOMD consistently outperforms existing approaches, with gains ranging from 0.63% when distilling VGG13 to VGG8, up to an impressive 2.79% boost transferring knowledge from WRN-40-2 to MobileNetV2. When distilling from heterogeneous teacher ensembles of three architectures, BOMD surpasses the following best method by 0.89–1.35% across different student models. Even in the challenging case of five diverse teacher architectures, BOMD maintains its edge, outperforming alternative techniques by 0.26–1.16%. These state-of-the-art results demonstrate the effectiveness of our orthogonal projection strategy and bi-level optimization in optimally blending teacher representations for enhanced knowledge transfer. Our contributions are as follows:

- Our work introduces a novel BOMD approach that combines orthogonal projections and bi-level optimization for effective knowledge transfer from an ensemble of diverse teacher models.

- A key component of our BOMD method is the use of bi-level optimization to learn optimal weighting factors for combining knowledge from multiple teachers. Unlike heuristic weighting strategies, our approach treats the weighting factors as upper-level variables and the student's parameters as lower-level variables in a nested optimization problem.
- Through extensive experiments on benchmark datasets, we validate the effectiveness and flexibility of our BOMD approach. Our method achieves state-of-the-art performance on the CIFAR-100 benchmark for multi-teacher knowledge distillation, consistently outperforming existing approaches across diverse teacher–student scenarios, including homogeneous and heterogeneous teacher ensembles.

Our paper is structured to provide a comprehensive exploration of our method. Following the related work sections, we begin by presenting an overview of our approach, illustrated in Figure 1. The core of our paper is Section 3, which delves into the technical details of BOMD. We start by discussing multi-teacher feature-based and logit-based distillation techniques, providing mathematical formulations for each. We then introduce our novel multiple orthogonal projections strategy, which aims to align teacher features with the student's feature space while preserving structural properties. The section concludes with an explanation of our bi-level optimization approach for determining optimal weighting factors for each teacher model. Then, we present a comprehensive evaluation in the Experiment section and summarize the approach and outlook in the conclusion Section 6.

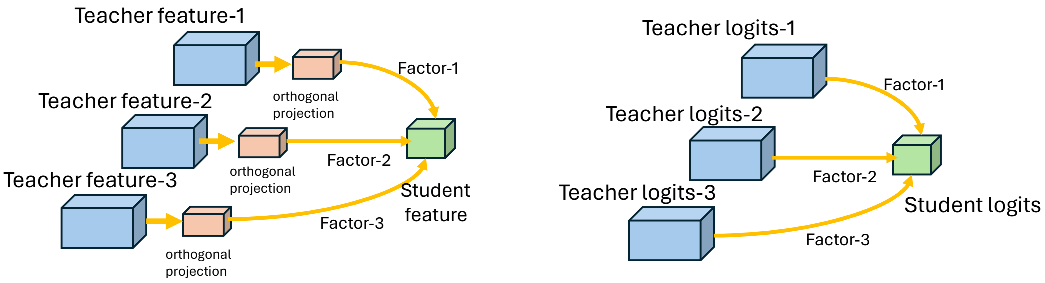

Figure 1. A schematic overview of our BOMD. During the training phase, BOMD utilizes the feature distillation (**left**) and logit distillation (**right**) between teacher–student models.

2. Related Work

2.1. Knowledge Distillation

Knowledge distillation is a powerful model compression technique that transfers knowledge from a large, complex teacher model to a smaller, more efficient student model, enabling the deployment of sophisticated AI systems in resource-constrained environments [14–19]. This approach encompasses various technology categories, each targeting different aspects of knowledge transfer. Response-based distillation focuses on the teacher's final output, utilizing soft targets that contain richer information than hard labels to guide the student's learning [8,20]. Feature-based distillation aims to transfer intermediate representations, allowing the student to learn more nuanced internal knowledge from the teacher [21–23]. Relation-based distillation preserves relationships between data instances or model components, capturing structural knowledge that is crucial for maintaining performance.

Single-teacher knowledge distillation has shown remarkable success across numerous domains, with several innovative approaches emerging to enhance its effectiveness. CRD (Contrastive Representation Distillation) introduced a contrastive objective for knowledge distillation, significantly improving image classification accuracy by leveraging contrastive learning to capture fine-grained structural knowledge from the teacher, enabling the student to learn more discriminative features [24]. FitNets pioneered the concept of hint training to guide the learning of intermediate representations, particularly beneficial for deeper architectures where direct knowledge transfer can be challenging [18].

These advancements in knowledge distillation have not only improved the performance of compact models but have also opened up new possibilities for efficient AI deployment. As research in this field continues to evolve, we can expect further innovations that push the boundaries of model compression and efficiency, potentially revolutionizing the way we design and deploy AI systems in resource-constrained scenarios.

2.2. Multi-Teacher Knowledge Distillation

Multi-teacher knowledge distillation emerged as a promising technique to enhance student performance by leveraging collective knowledge from multiple pre-trained teachers, offering several potential advantages over single-teacher approaches [9]. This method provides diverse knowledge sources, increased robustness to individual teacher biases, and the ability to specialize in different aspects of the task, showing particular promise in scenarios where different teacher models excel in complementary aspects of the problem domain.

Early multi-teacher methods employed simple averaging or weighted averaging strategies to combine logits or feature representations from teachers, but these approaches treated teachers as equal contributors, failing to exploit each teacher's unique strengths and specializations fully [9]. This limitation led to suboptimal knowledge transfer, especially when dealing with heterogeneous teacher ensembles with varying architectures or training paradigms, prompting the development of more advanced multi-teacher methods focused on adaptive weighting strategies and selective knowledge transfer.

EBKD introduced an attention-based weighting scheme to emphasize knowledgeable teachers for each input sample, showing significant improvements over fixed weighting strategies, particularly on challenging datasets like CIFAR-100 [12]. The attention mechanism in EBKD allows the student to dynamically focus on the most relevant teacher knowledge for each specific input, leading to more effective and targeted learning. OKD-Dip proposed an online knowledge distillation framework that dynamically adjusts the importance of each teacher based on their performance on mini-batches during training, enabling the student to adapt to changing teacher contributions throughout the learning process and potentially capturing temporal dynamics in teacher expertise [10].

Recent works have also explored multi-teacher distillation under specific constraints, with CA-MKD addressing the challenge of noisy or adversarial teachers through robust optimization techniques, demonstrating resilience to corrupted knowledge sources, an important consideration in real-world applications where teacher quality may vary [11]. Another line of research explored complementary knowledge distillation, encouraging the student to learn distinctive information from each teacher not captured by others, with Yuan et al. proposing a reinforced multi-teacher selection strategy that adaptively chooses the most informative subset of teachers for each training instance [25]. This approach showed promise in scenarios with many diverse teachers, effectively navigating the trade-off between leveraging multiple knowledge sources and avoiding redundancy or conflicting information.

As research in multi-teacher knowledge distillation continues to evolve, we can expect further innovations that address the challenges of effectively combining and prioritizing diverse knowledge sources, potentially leading to more robust and versatile student models capable of excelling across a wide range of tasks and domains.

2.3. Difference of Our Method vs. Existing Methods

The field of multi-teacher knowledge distillation has seen significant advancements, with various methods addressing different aspects of the challenge. CA-MKD focuses on robustness to noisy teachers, demonstrating resilience in scenarios where teacher quality may vary [11]. In contrast, our proposed BOMD method is designed as a general framework applicable to various teacher–student scenarios, consistently outperforming existing approaches by substantial margins on the CIFAR-100 benchmark (Tables 1–3), even in scenarios without explicitly noisy teachers. This superior performance suggests that BOMD's orthogonal projection strategy and bi-level optimization provide benefits beyond robustness to noise, potentially capturing more nuanced and complementary information from the teacher ensemble.

Table 1. Distillation results (Top-1 accuracy) of multi-teacher KD methods on CIFAR-100. $ARI = \frac{1}{M}\sum_{i=1}^{M}\frac{Acc^{i}_{MMKD}-Acc^{i}_{BKD}}{Acc^{i}_{BKD}-Acc^{i}_{Stu}} \times 100\%$.

Teacher	VGG13 75.17 ± 0.18	ResNet32x4 79.31 ± 0.14	ResNet32x4 79.31 ± 0.14	WRN-40-2 76.62 ± 0.26	WRN-40-2 76.62 ± 0.26	ResNet20x4 78.632 ± 0.24	ARI (%)
Ensemble	77.07	81.16	81.16	79.62	79.62	80.81	
Student	VGG8 70.74 ± 0.40	MobileNetV2 65.64 ± 0.19	VGG8 70.74 ± 0.40	MobileNetV2 65.64 ± 0.19	WRN-40-1 71.93 ± 0.22	ShuffleNetV1 71.70 ± 0.43	/
AVER [9]	73.98 ± 0.13	68.42 ± 0.06	73.23 ± 0.35	69.67 ± 0.01	74.56 ± 0.13	75.73 ± 0.02	49.97%
AEKD-logits [13]	73.82 ± 0.09	68.39 ± 0.13	73.22 ± 0.29	69.56 ± 0.34	74.18 ± 0.25	75.93 ± 0.32	54.87%
FitNet-MKD [21]	74.05 ± 0.07	68.46 ± 0.49	73.24 ± 0.24	69.29 ± 0.42	74.95 ± 0.30	75.98 ± 0.06	46.97%
AEKD-feature [13]	73.99 ± 0.15	68.18 ± 0.06	73.38 ± 0.16	69.44 ± 0.25	74.96 ± 0.18	76.86 ± 0.03	43.16%
CA-MKD [11]	74.27 ± 0.16	69.19 ± 0.04	75.08 ± 0.07	70.87 ± 0.14	75.27 ± 0.21	77.19 ± 0.49	11.98%
BOMD	74.90 ± 0.07	69.88 ± 0.04	75.86 ± 0.18	71.56 ± 0.03	75.78 ± 0.11	77.98 ± 0.35	/

Table 2. Distillation accuracy of BOMD.

Teacher	ResNet32x4 79.31 ± 0.14	WRN-40-2 76.62 ± 0.26	WRN-40-2 76.62 ± 0.26
Student	MobileNetV2 65.64 ± 0.19	MobileNetV2 65.64 ± 0.19	WRN-40-1 71.93 ± 0.22
KD [8]	67.57 ± 0.10	69.31 ± 0.20	74.22 ± 0.09
AT [23]	67.38 ± 0.21	69.18 ± 0.37	74.83 ± 0.15
VID [26]	67.78 ± 0.13	68.57 ± 0.11	74.37 ± 0.22
CRD [24]	69.04 ± 0.16	70.14 ± 0.06	74.82 ± 0.06
SRRL [27]	68.77 ± 0.06	69.44 ± 0.13	74.60 ± 0.04
SemCKD [28]	68.86 ± 0.26	69.61 ± 0.05	74.41 ± 0.16
BOMD	69.89 ± 0.12	71.45 ± 0.12	75.76 ± 0.15

Table 3. Distillation accuracy of multi-teacher KD methods.

Dataset	**Stanford Dogs**		**Tiny-ImageNet**	
Teacher	ResNet101 68.39 ± 1.44	ResNet34x4 66.07 ± 0.51	ResNet32x4 53.38 ± 0.11	VGG13 49.17 ± 0.33
Student	ShuffleNetV2x0.5 59.36 ± 0.73	ShuffleNetV2x0.5 59.36 ± 0.73	MobileNetV2 39.46 ± 0.38	MobileNetV2 39.46 ± 0.38
AVER [9]	65.13 ± 0.13	63.46 ± 0.21	41.78 ± 0.15	41.87 ± 0.11
EBKD [12]	64.28 ± 0.13	64.19 ± 0.11	41.24 ± 0.11	41.46 ± 0.24
CA-MKD [11]	64.09 ± 0.35	64.28 ± 0.20	43.90 ± 0.09	42.65 ± 0.05
AEKD-feature [13]	64.91 ± 0.21	62.13 ± 0.29	42.03 ± 0.12	41.56 ± 0.14
AEKD-logits [13]	65.18 ± 0.24	63.97 ± 0.14	41.46 ± 0.28	41.19 ± 0.23
BOMD	65.54 ± 0.12	64.67 ± 0.18	44.21 ± 0.04	44.35 ± 0.12

Despite these promising developments, several challenges remain in the field of multi-teacher knowledge distillation, with a key challenge being the efficient exploration of the vast search space of teacher ensembles and knowledge transfer strategies. While methods like EBKD and OKDDip have made progress in adaptive weighting, they still rely on predefined architectures and transfer mechanisms, potentially hindering their ability to fully exploit the potential of diverse teacher ensembles, especially when dealing with significantly different model architectures or knowledge representations [10,12].

Our BOMD approach takes a different stance on leveraging diverse teacher knowledge by aligning teacher representations in the student's feature space while preserving structural properties through orthogonal projections. This novel approach effectively blends diverse teacher knowledge, allowing the student to benefit from the full ensemble while maintaining computational efficiency. The use of orthogonal projections is particularly innovative, as it enables the preservation of important geometric relationships in the feature space during the knowledge transfer process.

Unlike previous methods that rely on heuristics or simple attention mechanisms, BOMD's bi-level optimization allows for a more principled and flexible approach to weight assignment, tailored to the specific teacher ensemble and student model. This bi-level formulation treats the weighting factors as upper-level variables and the student's parameters as lower-level variables, enabling a more nuanced optimization process that can adapt to the intricate relationships between teachers and students. By jointly optimizing these factors, BOMD can potentially discover more effective knowledge transfer strategies that are tailored to the specific characteristics of each teacher–student combination.

As research in this field continues to evolve, we anticipate that approaches like BOMD will pave the way for more sophisticated and adaptive multi-teacher distillation methods, potentially revolutionizing how we leverage collective knowledge from diverse model ensembles to train highly efficient and performant student models.

3. Bi-Level Orthogonal Multi-Teacher Distillation

A common approach in multi-teacher knowledge distillation is to optimize the student model by minimizing the divergence between its representations (features or logits) and those of an ensemble of teacher models (see Figure 1). Two widely used divergence measures in this context are the L2 distance for feature-based distillation and the Kullback–Leibler (KL) divergence for logit-based distillation.

3.1. Multi-Teacher Feature-Based Distillation

Consider a student model S and a collection of N teacher models represented as $\mathcal{T} = T_1, T_2, \ldots, T_N$. For any input x, let $S(x)$ denote the feature representation extracted by the student model, and $T_i(x)$ represent the feature representation from the i-th teacher model, where i ranges from 1 to N.

Our objective is to minimize the L2 distance between the student's feature representation and the average of the teachers' representations. This can be expressed mathematically as

$$\mathcal{L}_{\text{feat}}(x) = \left\| S(x) - \frac{1}{N} \sum_{i=1}^{N} T_i(x) \right\|_2^2 \tag{1}$$

3.2. Multi-Teacher Logit-Based Distillation

In logit-based distillation, the Kullback–Leibler (KL) divergence is employed to measure the discrepancy between the softmax distributions of the student and teacher models. Let $p_S(x)$ and $p_i(x)$ denote the softmax distributions of the student and the i-th teacher model, respectively, for the input sample x. The KL divergence between the student and the mean of the teacher distributions is minimized, as formulated in the following objective function:

$$\mathcal{L}_{\text{logit}}(x) = \text{KL}\left(p_S(x), \middle\|, \frac{1}{N} \sum_{i=1}^{N} p_i(x) \right) \tag{2}$$

The KL divergence between two probability distributions p and q is defined as

$$\text{KL}(p, \|, q) = \sum_k p(k) \log \frac{p(k)}{q(k)} \tag{3}$$

where k iterates over all possible classes or output categories.

By minimizing $\mathcal{L}_{\text{logit}}(x)$, the student model is encouraged to mimic the ensemble behavior of the teacher models, effectively distilling their collective knowledge into its softmax output distributions. These feature-based and logit-based distillation objectives can be combined with the standard cross-entropy loss for supervised learning, forming a multi-task optimization objective that leverages both the ground-truth labels and the knowledge from the teacher ensemble.

3.3. Multiple Orthogonal Projections

Effectively retaining relevant features from multiple teacher models is crucial for successful knowledge transfer during the distillation process. We propose using multiple orthogonal projections to align the direct feature representations from different teacher models with the student model's feature space. Orthogonal projections can enhance feature retention by preserving structural properties and relationships between features.

Let $\mathcal{T} = T_1, T_2, \ldots, T_M$ denote the set of M teacher models, and S represent the student model. For each teacher model T_i, we extract its direct feature representation $\mathbf{F}_{T_i} \in \mathbb{R}^{d_T \times n}$, where d_T is the dimensionality of the teacher's feature space, and n is the number of instances. Similarly, we have the student model's feature representations $\mathbf{F}_S \in \mathbb{R}^{d_S \times n}$, where d_S is the dimensionality of the student's feature space. Our objective is to align the teacher features $\mathbf{F}_{T_i}$ with the student's feature space while preserving their structural properties. To achieve this, we introduce M orthogonal projection matrices $\mathbf{P}_I \in \mathbb{R}^{d_S \times d_T}$, one for each teacher model, which satisfies the following orthogonality condition:

$$\mathbf{P}_i^\top \mathbf{P}_i = \mathbf{I}_{d_T} \tag{4}$$

where $\mathbf{I}_{d_T}$ is the $d_T \times d_T$ identity matrix, ensuring that the projections preserve the norms and angles between feature vectors. The projected teacher features $\mathbf{F}_{T_i}^P$ are obtained by applying the corresponding orthogonal projection matrix $\mathbf{P}_i$:

$$\mathbf{F}_{T_i}^P = \mathbf{P}_i \mathbf{F}_{T_i} \tag{5}$$

The projected teacher features $\mathbf{F}_{T_i}^P$ now reside in the same feature space as the student model, facilitating the knowledge distillation process. To optimize the orthogonal projection matrices $\mathbf{P}_i$, we introduce a feature alignment loss $\mathcal{L}align$ that aims to minimize the distance between the projected teacher features and the student's features:

$$\mathcal{L}align = \sum_{i=1}^{M} \left\| \mathbf{F}_{T_i}^P - \mathbf{F}_S \right\|_F^2 \tag{6}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

3.4. Benefits and Limitations

The use of multiple orthogonal projections in our approach offers several key benefits for multi-teacher knowledge distillation. Firstly, it allows for the preservation of important geometric relationships in the feature space during the knowledge transfer process, ensuring that critical structural information from each teacher is retained. This is particularly valuable when dealing with diverse teacher models that may capture different aspects of the problem domain. Secondly, the orthogonal nature of the projections minimizes interference between different teachers' knowledge, allowing the student to effectively learn from multiple sources without conflicting information. This can lead to more robust and comprehensive knowledge transfer, especially in scenarios where teachers have complementary strengths. Additionally, the flexibility of our approach in adapting to different teacher–student combinations through the optimization of projection matrices enables more efficient and targeted knowledge distillation. This adaptability is crucial for handling heterogeneous teacher ensembles and varying student architectures, potentially leading to improved generalization and performance across a wide range of tasks. Furthermore, by aligning teacher representations in the student's feature space, our method facilitates more direct and interpretable knowledge transfer, which can be particularly beneficial for analyzing and understanding the distillation process. Overall, these benefits contribute to a more effective and versatile multi-teacher knowledge distillation framework, capable of leveraging diverse teacher ensembles to train highly efficient and performant student models.

3.5. Bi-Level Optimization for Weighting Factors

Traditional multi-teacher distillation approaches often rely on heuristic weighting strategies, such as equal weighting or similarity-based weighting, to combine knowledge from different teachers. However, these heuristics may not be optimal, especially when teachers and students have heterogeneous architectures or possess diverse complementary knowledge. To address this limitation, we propose optimizing the weighting factors matrix $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_M]$ using bi-level optimization, where $\alpha_i \in [0,1]$ and $\sum_{i=1}^{M} \alpha_i = 1$ quantifies the relative importance of each teacher model.

The student model S is trained to minimize a weighted combination of the knowledge distillation losses from the individual teachers:

$$\mathcal{L}_{KD}(w, \alpha) = \sum_{i=1}^{M} \alpha_i \mathcal{L}_{KD}(S(w), T_i) \tag{7}$$

where w represents the trainable parameters of the student model S, and $\mathcal{L}_{KD}(\cdot,\cdot)$ is the knowledge distillation loss function, which can be defined as the cross-entropy between the student's logits and the teacher's logits, or any other suitable loss measure.

To optimize the weight vector α, we employ a bilevel optimization approach. In this framework, the weight vector α is treated as the upper-level variable, and the student model's parameters w are the lower-level variables. The bilevel optimization problem can be formulated as

$$\min_{\alpha} \quad \mathcal{L}_{val}(w^*(\alpha), \alpha) \tag{8}$$

$$\text{s.t.} \quad w^*(\alpha) = \text{argmin}_w \ \ \mathcal{L}_{train}(w, \alpha) \tag{9}$$

where $\mathcal{L}val(\cdot,\cdot)$ is the validation loss function, which evaluates the performance of the student model with parameters $w^*(\alpha)$ optimized for the given weight vector α. The nested formulation in Equation (9) implies that for any fixed value of α, the student model's parameters w are optimized to minimize the weighted combination of knowledge distillation losses from the teachers. Subsequently, in the outer optimization problem (Equation (8)), the weight vector α is adjusted to minimize the validation loss of the student model with the optimized parameters $w^*(\alpha)$.

The computational complexity of the bi-level optimization approach for multi-teacher knowledge distillation scales significantly with both the number of teachers and the size of the student model. The outer optimization problem (Equation (8)) optimizes M variables, one for each teacher, and the inner optimization problem (Equation (9)) computes a weighted sum of M individual knowledge distillation losses. As M increases, the dimensionality of the search space for α grows linearly, potentially requiring more iterations to converge, and the time complexity of computing the combined loss grows linearly with M. The inner optimization problem involves training the student model, which scales with the model's size (number of parameters), and larger student models require more computation per forward and backward pass, impacting the time for each iteration. The memory requirements also increase with the student model size, potentially limiting the batch size and affecting convergence speed. For each update of α, the student model needs to be retrained or fine-tuned, leading to a multiplicative effect on computational cost, and the gradient computation for the outer problem may require approximations (e.g., implicit differentiation), which can be computationally expensive for large models and many teachers.

The bi-level nature of the problem can lead to instabilities and slow convergence, potentially requiring more iterations as the problem scale increases, and the interdependence between α and w may necessitate careful balancing of inner and outer optimization steps, further impacting scalability. In practice, the approach might become computationally prohibitive for very large student models or a high number of teachers, and approximate methods, such as truncated back-propagation or meta-learning techniques, might be necessary to make the approach feasible for large-scale problems. Additionally, techniques like progressive training or teacher pruning could be explored to manage the computational complexity as the number of teachers increases.

This bi-level optimization problem can be approached using gradient-based methods similar to hyperparameter optimization. However, optimizing α is more challenging than scalar-valued hyperparameters due to its higher dimensionality and the nested optimization problem's complexity. Bi-level optimization offers several advantages: (1) Adaptability: By directly optimizing the weighting factors, the method can adapt to the specific characteristics of the teacher ensemble and the student model, leading to more effective knowledge transfer. (2) Heterogeneity Handling: Unlike heuristic weighting strategies that assume homogeneity among teachers and students, bi-level optimization can handle heterogeneous architectures and diverse knowledge sources, allowing for an optimal combination of complementary information. (3) Flexibility: The bi-level optimization formulation is flexible and can accommodate various distillation loss functions, feature representations, and architectural configurations, making it applicable to a wide range of multi-teacher distillation scenarios. (4) Performance Improvement: By optimally aligning teacher–student features and effectively distilling collective knowledge from the teacher ensemble, bi-level optimization has the potential to significantly improve the student model's performance compared to heuristic weighting strategies. Through this bi-level optimization approach, we aim to unlock the full potential of multi-teacher distillation by learning an optimal weighting strategy that maximizes knowledge transfer while accounting for the heterogeneity and complementarity of the teacher ensemble.

4. Experiments

To evaluate the effectiveness of our proposed BOMD method, we conducted extensive experiments on multiple benchmark datasets. We compared our approach with state-of-the-art multi-teacher knowledge distillation methods across various teacher–student scenarios.

4.1. Datasets and Implementation Details

Our study employed the CIFAR-100 benchmark dataset to evaluate the effectiveness of the BOMD method. CIFAR-100 is a widely used image classification dataset consisting of 60,000 32×32 color images across 100 classes. We adhered to the standard split of 50,000 training images and 10,000 test images for our experiments. To comprehensively assess BOMD, we utilized a variety of model architectures for both teacher and student networks. Our teacher models included VGG13, ResNet32, and Wide ResNet (WRN-40-2), while student models comprised VGG8, ResNet8, and MobileNetV2. This diverse set of architectures allowed us to test BOMD's performance in both homogeneous and heterogeneous teacher ensemble scenarios. The experimental procedure involved first training the teacher models independently on the CIFAR-100 training set. These pre-trained teacher models were then used to distill knowledge to the student model using our BOMD approach. The student model's training process incorporated a combination of cross-entropy loss with ground truth labels, knowledge distillation loss from teacher logits (using KL divergence), and our novel feature alignment loss utilizing orthogonal projections.

4.2. Settings and Hyperparameters

For our experiments, we employed a batch size of 128 and trained the models for 200 epochs. We used the Adam optimizer with an initial learning rate of 0.001, which was decreased by a factor of 0.1 at epochs 80 and 160. These hyperparameters were consistently applied across all experiments to ensure fair comparisons. The core of our BOMD method consists of two key components: multiple orthogonal projections and bi-level optimization. For the orthogonal projections, we introduced projection matrices for each teacher to align their features with the student's feature space. These projections were optimized by minimizing a feature alignment loss. The bi-level optimization component was used to determine optimal weighting factors for combining knowledge from multiple teachers. This was formulated as a nested optimization problem, with the upper-level optimizing the weighting factors and the lower-level optimizing the student model parameters. To solve the bi-level optimization problem, we employed a gradient-based approach using the Adam optimizer. The upper-level optimization used a learning rate of 0.0005, while the lower-level optimization used the same learning rate as the overall student model training (0.001). We found that five inner optimization steps for every outer optimization step provided a good balance between computational efficiency and optimization quality. We evaluated the performance of BOMD and baseline methods using top-1 classification accuracy on the CIFAR-100 test set. Our baseline comparisons included established multi-teacher knowledge distillation methods such as EBKD, OKDDip, and CA-MKD. For each method, including BOMD, we conducted experiments with varying numbers of teachers (from 1 to 5) and different student model architectures to provide a comprehensive performance analysis.

4.3. Experimental Framework and Devices

All experiments were implemented using PyTorch 1.8.0 and conducted on a system with 4 NVIDIA Tesla V100 GPUs. To ensure reproducibility, we have made our code publicly available on GitHub, including detailed instructions for replicating our experimental setup and results. This comprehensive experimental framework allows for a thorough evaluation of BOMD's performance across various scenarios, providing insights into its effectiveness and versatility in multi-teacher knowledge distillation tasks.

5. Experiment Results

5.1. Distillation Performance of Multi-Teacher KD Methods on CIFAR-100

The results in Table 1 clearly demonstrate the effectiveness of our BOMD approach across diverse teacher–student scenarios on CIFAR-100. Consistently outperforming existing multi-teacher knowledge distillation methods, BOMD achieves state-of-the-art performance in transferring knowledge from an ensemble of homogeneous teachers to students of varying architectures. What stands out is the significant performance gain of BOMD over the following best method, spanning increases ranging from 0.63% for distilling VGG13 to VGG8 up to an impressive 2.79% boost when transferring knowledge from WRN-40-2 to the lightweight MobileNetV2 student. This substantial improvement can be attributed to our novel orthogonal projection strategy that aligns teacher representations in the student's feature space while preserving structural properties, coupled with the bi-level optimization that learns optimal teacher weightings tailored to the specific teacher ensemble and student model. Furthermore, the Average Relative Improvement (ARI) metric, which measures the relative accuracy gain compared to the student's initial performance, reveals that BOMD consistently surpasses other methods across all teacher–student combinations. This underscores the robustness and broad applicability of our approach, regardless of the specific architectures involved.

5.2. Compared to Single-Teacher Methods

Extending our analysis to scenarios where knowledge is distilled from a single teacher, Table 2 showcases the competitive performance of BOMD against state-of-the-art single-teacher knowledge distillation techniques. Across diverse teacher–student pairs on CIFAR-100, our method consistently outperforms well-established approaches such as KD, FitNets, Attention Transfer (AT), Variational Information Distillation (VID), Contrastive Representation Distillation (CRD), SemCKD, and SRRL.

Notably, when distilling knowledge from the powerful ResNet32x4 teacher to the lightweight MobileNetV2 student, BOMD achieves a remarkable 69.89% accuracy, surpassing the next best method, CRD, by a substantial 0.85% margin. Similarly, for the WRN-40-2 to MobileNetV2 transfer, our approach outperforms the runner-up CRD by 1.31%, reaching an impressive 71.45% accuracy.

These results demonstrate that, in addition to excelling in multi-teacher scenarios, our BOMD framework is highly competitive and often superior to dedicated single-teacher distillation techniques tailored for specific teacher–student pairs. The effectiveness of our orthogonal projection and bi-level optimization strategies shines through, enabling efficient and robust knowledge transfer even in traditional single-teacher distillation settings.

5.3. Distillation Performance on Large-Scale Datasets

Extending our evaluation to large-scale datasets, Table 3 represents the results of our BOMD method on the Stanford Dogs and Tiny-ImageNet benchmarks. Once again, our approach consistently outperforms existing multi-teacher knowledge distillation methods across diverse teacher ensembles and student architectures.

On the Stanford Dogs dataset, BOMD achieves top accuracies of 65.54% and 64.67% when distilling knowledge from ResNet101 and ResNet34x4 teacher ensembles, respectively, to the lightweight ShuffleNetV2x0.5 student. These results represent substantial improvements over the following best methods, with gains ranging from 0.41% to 2.54%.

5.4. Results on CIFAR-100 with Three Teachers

In Table 4, we present the top-1 test accuracy results of our Bi-level Optimization for Multi-Teacher Distillation (BOMD) method and other state-of-the-art multi-teacher knowledge distillation approaches on the CIFAR-100 dataset. The experiments involve three teacher models with different architectures, serving as a challenging and diverse teacher ensemble for knowledge distillation.

Table 4. Distillation accuracy of MKD methods.

Teacher	ResNet56	73.47	ResNet8	59.32	VGG11	71.52
	ResNet20x4	78.39	WRN-40-2	76.51	VGG13	75.19
	VGG13	75.19	ResNet20x4	78.39	ResNet32x4	79.31
Student	VGG8	70.74 ± 0.40	ResNet8x4	72.79 ± 0.14	VGG8	70.74 ± 0.40
FitNet-MKD [21]	75.06 ± 0.13		75.21 ± 0.12		73.43 ± 0.08	
AVER [9]	75.11 ± 0.57		75.16 ± 0.11		73.59 ± 0.06	
EBKD [12]	74.18 ± 0.22		75.44 ± 0.29		73.45 ± 0.08	
AEKD-feature [13]	74.69 ± 0.57		73.98 ± 0.18		73.40 ± 0.06	
AEKD-logits [13]	75.17 ± 0.30		73.93 ± 0.17		74.15 ± 0.08	
CA-MKD [11]	75.53 ± 0.14		75.27 ± 0.18		74.63 ± 0.17	
BOMD	76.42 ± 0.15		76.49 ± 0.14		75.98 ± 0.14	

Our BOMD method achieves the highest top-1 accuracy across all three teacher ensemble configurations, outperforming the existing approaches by a considerable margin. Specifically, when distilling knowledge from ResNet56, ResNet20x4, and VGG13 teachers into a VGG8 student, BOMD attains a remarkable 76.42% accuracy, surpassing the second-best CA-MKD method by 0.89%. This superior performance highlights the effectiveness of our bi-level optimization strategy in learning optimal weighting factors, enabling efficient knowledge transfer from the heterogeneous teacher ensemble.

Furthermore, our approach consistently demonstrates its strength in the other two teacher ensemble setups, achieving 76.49% accuracy with the ResNet8, WRN-40-2, and ResNet20x4 teachers, and 75.98% accuracy with the VGG11, VGG13, and ResNet32x4 teachers. These results underscore the adaptability and robustness of BOMD in handling diverse teacher architectures, leveraging their complementary knowledge to enhance the student's performance significantly.

5.5. Results on CIFAR-100 with Five Teachers

Extending our evaluation to a more challenging scenario with five teacher models, Table 5 showcases the top-1 test accuracy of our BOMD method and competing approaches on the CIFAR-100 dataset. The increased number of teachers with varying architectures further amplifies the complexity of effectively distilling knowledge from the ensemble.

Table 5. Distillation accuracy of our methods.

Teacher	ResNet8	59.32	VGG11	71.52	ResNet8	59.32
	VGG11	71.52	ResNet56	73.47	VGG11	71.52
	ResNet56	73.47	VGG13	75.19	VGG13	75.19
	VGG13	75.19	ResNet20x4	78.39	WRN-40-2	76.51
	ResNet32x4	79.31	ResNet32x4	79.31	ResNet20x4	78.39
Student	VGG8	70.74 ± 0.40	VGG8	70.74 ± 0.40	MobileNetV2	65.64 ± 0.19
AEKD-feature [13]	74.02 ± 0.08		75.06 ± 0.03		69.41 ± 0.21	
AVER [9]	74.47 ± 0.47		74.48 ± 0.12		69.41 ± 0.04	
AEKD-logits [13]	73.53 ± 0.10		74.90 ± 0.17		69.28 ± 0.21	
EBKD [12]	74.37 ± 0.07		73.94 ± 0.29		69.26 ± 0.64	
CA-MKD [11]	74.64 ± 0.23		75.02 ± 0.21		70.30 ± 0.51	
BOMD	75.56 ± 0.34		75.32 ± 0.13		71.46 ± 0.26	

Once again, our proposed BOMD method emerges as the top-performing technique across all three teacher ensemble configurations. When distilling knowledge from ResNet8, VGG11, ResNet56, VGG13, and ResNet32x4 teachers into a VGG8 student, BOMD achieves a remarkable 75.56% accuracy, outperforming the second-best CA-MKD method by 0.92%. This remarkable improvement demonstrates the power of our bi-level optimization framework in optimally combining diverse knowledge sources, even in the presence of a large and heterogeneous teacher ensemble.

Moreover, our method maintains its superiority in the other two setups, attaining 75.32% accuracy with the VGG11, ResNet56, VGG13, ResNet20x4, and ResNet32x4 teachers, and 71.46% accuracy with the ResNet8, VGG11, VGG13, WRN-40-2, and ResNet20x4 teachers distilled into a MobileNetV2 student. These results further validate the effectiveness and adaptability of our approach, showcasing its ability to handle various student-teacher architectural configurations while maximizing knowledge transfer.

By leveraging the bi-level optimization technique to learn optimal weighting factors, our BOMD method effectively aligns and combines the heterogeneous teacher representations, enabling the student model to benefit from the collective knowledge of the diverse teacher ensemble. This unique capability positions BOMD as a powerful tool for multi-teacher knowledge distillation, facilitating improved performance and efficient knowledge transfer across a wide range of scenarios.

5.6. Advantages over Other Method

Our proposed BOMD approach demonstrates significant improvements over existing multi-teacher knowledge distillation methods across a variety of scenarios. The consistent performance gains, ranging from 0.26% to 2.79% on CIFAR-100 benchmarks, highlight the effectiveness of combining orthogonal projections with bi-level optimization for knowledge transfer. The superiority of BOMD over methods like CA-MKD [11], which focuses primarily on robustness to noisy teachers, suggests that our approach captures more nuanced and complementary information from the teacher ensemble. BOMD's bi-level optimization strategy for learning optimal weighting factors provides a more principled and flexible approach compared to heuristic weighting schemes used in previous work like EBKD and OKDDip. This adaptability likely contributes to BOMD's strong performance across diverse teacher–student scenarios, including both homogeneous and heterogeneous teacher ensembles. The impressive gains achieved when distilling to compact student models, such as MobileNetV2, are particularly noteworthy, suggesting that BOMD is especially effective at transferring knowledge to resource-constrained models, which is crucial for deploying efficient AI systems in real-world applications with limited computational resources.

5.7. Analysis of Our Method

The success of BOMD can be attributed to several key factors. First, the use of orthogonal projections to align teacher features with the student's feature space while preserving structural properties appears to be a powerful mechanism for effective knowledge transfer. This approach likely allows the student to benefit from the diverse expertise of multiple teachers without losing important relational information within the feature representations. Second, by treating the weighting factors as upper-level variables and the student's parameters as lower-level variables, BOMD can adapt the knowledge transfer process to the specific characteristics of both the teacher ensemble and the student model. This bi-level optimization framework enables a more nuanced and tailored approach to knowledge distillation, potentially capturing complex interactions between teachers and students that simpler methods might miss. The consistent performance improvements across various teacher–student combinations suggest that BOMD's approach to feature alignment and knowledge integration is robust and generalizable.

5.8. Limitations of Our Method

While BOMD shows promising results, it is important to acknowledge potential limitations of our approach. The computational complexity of bi-level optimization may become prohibitive for very large teacher ensembles or student models. This scalability issue could limit BOMD's applicability in scenarios involving numerous teachers or extremely large model architectures. Future work could explore approximation techniques or progressive training strategies to improve scalability.

6. Conclusions

In this paper, we introduced BOMD, a novel approach for multi-teacher knowledge distillation that combines orthogonal projections with bi-level optimization. Our method effectively addresses the challenge of knowledge transfer from diverse teacher ensembles to compact student models. Through extensive experiments on the CIFAR-100 benchmark, we demonstrated that BOMD consistently outperforms existing multi-teacher distillation methods across various scenarios, including both homogeneous and heterogeneous teacher ensembles.

The key strengths of BOMD lie in its ability to achieve the following:

- Align teacher features with the student's feature space through orthogonal projections, preserving structural properties during knowledge transfer.
- Optimize weighting factors for combining teacher knowledge using a principled bi-level optimization approach.
- Achieve significant performance improvements even when distilling to very compact student models.

Limitations and Future Work

Future work could explore the application of BOMD to other domains beyond image classification, such as natural language processing or speech recognition. Additionally, investigating the scalability of our approach to even larger and more diverse teacher ensembles could provide further insights into its effectiveness and limitations.

In conclusion, BOMD represents a significant step forward in multi-teacher knowledge distillation, offering a flexible and effective framework for leveraging diverse knowledge sources to train high-performing compact models. As the field of AI continues to evolve towards more efficient and deployable solutions, techniques like BOMD will play a crucial role in bridging the gap between state-of-the-art performance and practical deployment constraints.

Author Contributions: Methodology, writing, experiment(s), S.G.; writing—review and editing, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dong, P.; Niu, X.; Li, L.; Xie, L.; Zou, W.; Ye, T.; Wei, Z.; Pan, H. Prior-Guided One-shot Neural Architecture Search. *arXiv* **2022**, arXiv:2206.13329. [CrossRef]
2. Dong, P.; Li, L.; Wei, Z.; Niu, X.; Tian, Z.; Pan, H. EMQ: Evolving Training-free Proxies for Automated Mixed Precision Quantization. In Proceedings of the International Conference on Computer Vision (ICCV), Paris, France, 4–6 October 2023.
3. Zhu, C.; Li, L.; Wu, Y.; Sun, Z. Saswot: Real-time semantic segmentation architecture search without training. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 7722–7730.
4. Wei, Z.; Dong, P.; Hui, Z.; Li, A.; Li, L.; Lu, M.; Pan, H.; Li, D. Auto-prox: Training-free vision transformer architecture search via automatic proxy discovery. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 15814–15822.
5. Wei, Z.; Pan, H.; Li, L.; Dong, P.; Tian, Z.; Niu, X.; Li, D. TVT: Training-Free Vision Transformer Search on Tiny Datasets. *arXiv* **2023**, arXiv:2311.14337. [CrossRef]
6. Lu, L.; Chen, Z.; Lu, X.; Rao, Y.; Li, L.; Pang, S. UniADS: Universal Architecture-Distiller Search for Distillation Gap. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024.
7. Dong, P.; Li, L.; Pan, X.; Wei, Z.; Liu, X.; Wang, Q.; Chu, X. ParZC: Parametric Zero-Cost Proxies for Efficient NAS. *arXiv* **2024**, arXiv:2402.02105. [CrossRef]
8. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531. [CrossRef]
9. Fukuda, T.; Suzuki, M.; Kurata, G.; Thomas, S.; Cui, J.; Ramabhadran, B. Efficient Knowledge Distillation from an Ensemble of Teachers. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 3697–3701.
10. Chen, D.; Mei, J.P.; Wang, C.; Feng, Y.; Chen, C. Online knowledge distillation with diverse peers. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 3430–3437.
11. Zhang, H.; Chen, D.; Wang, C. Confidence-aware multi-teacher knowledge distillation. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 4498–4502.
12. Kwon, K.; Na, H.; Lee, H.; Kim, N.S. Adaptive knowledge distillation based on entropy. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 7409–7413.
13. Du, S.; You, S.; Li, X.; Wu, J.; Wang, F.; Qian, C.; Zhang, C. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12345–12355.
14. Li, L. Self-regulated feature learning via teacher-free feature distillation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 347–363.
15. Dong, P.; Li, L.; Wei, Z. Diswot: Student architecture search for distillation without training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 11898–11908.
16. Liu, X.; Li, L.; Li, C.; Yao, A. Norm: Knowledge distillation via n-to-one representation matching. *arXiv* **2023**, arXiv:2305.13803. [CrossRef]
17. Li, L.; Liang, S.N.; Yang, Y.; Jin, Z. Teacher-free distillation via regularizing intermediate representation. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
18. Li, L.; Dong, P.; Wei, Z.; Yang, Y. Automated knowledge distillation via monte carlo tree search. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 17413–17424.
19. Li, L.; Dong, P.; Li, A.; Wei, Z.; Yang, Y. Kd-zero: Evolving knowledge distiller for any teacher–student pairs. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 69490–69504.
20. Buciluǎ, C.; Caruana, R.; Niculescu-Mizil, A. Model compression. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 535–541.
21. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550. [CrossRef]

22. Yim, J.; Joo, D.; Bae, J.; Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4133–4141.
23. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928. [CrossRef]
24. Tian, Y.; Krishnan, D.; Isola, P. Contrastive Representation Distillation. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 30 April 2020.
25. Yuan, F.; Shou, L.; Pei, J.; Lin, W.; Gong, M.; Fu, Y.; Jiang, D. Reinforced multi-teacher selection for knowledge distillation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 14284–14291.
26. Ahn, S.; Hu, S.X.; Damianou, A.; Lawrence, N.D.; Dai, Z. Variational information distillation for knowledge transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9163–9171.
27. Yang, J.; Martinez, B.; Bulat, A.; Tzimiropoulos, G. Knowledge distillation via softmax regression representation learning. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2021.
28. Chen, D.; Mei, J.; Zhang, Y.; Wang, C.; Wang, Z.; Feng, Y.; Chen, C. Cross-Layer Distillation with Semantic Calibration. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; pp. 7028–7036.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.