A Survey on Knowledge Distillation Tiny ML Reasoning Models Compression and Small-Scale AI

www.surveyx.cn

Abstract

In the context of AI deployment on resource-constrained devices, the survey explores the interplay of knowledge distillation, Tiny ML, reasoning models, and model compression techniques. These methodologies focus on simplifying neural architectures and optimizing performance while maintaining accuracy. Key advancements include data-free knowledge distillation, which enhances model efficiency without requiring original datasets, and dynamic distillation methods that adapt to specific model and data characteristics. Ensemble and multi-teacher strategies are shown to improve student model performance by leveraging diverse teacher models. The study also highlights innovative loss functions and the integration of symbolic knowledge for enhanced interpretability. Pruning and quantization methods are emphasized as critical for reducing model complexity, while hybrid techniques combine these strategies to achieve greater efficiency. Applications span diverse fields, including graph processing, language understanding, image classification, and medical diagnostics, showcasing the potential of these techniques to optimize AI models for real-world deployment. The survey concludes by identifying future research directions, such as integrating emerging technologies and refining data utilization strategies, to further enhance model simplification and efficiency. These advancements promise to make AI more accessible and effective across various domains, particularly in environments with limited computational resources.

1 Introduction

1.1 Theme of Complexity and Resource Optimization

The reduction of complexity and optimization of resources are critical for deploying AI models on devices with limited computational and storage capabilities, particularly in Internet of Things (IoT) environments. Knowledge Distillation (KD) plays a vital role in this process, facilitating the compression of large models into more compact forms while aiming to preserve performance. This approach is especially pertinent where traditional distillation methods are hindered by the requirement for complete datasets and single-stage operations. Additionally, the inefficacy of conventional federated learning techniques in managing non-i.i.d. data distributions underscores the need for optimized AI solutions capable of functioning effectively in varied contexts [1].

In computer vision, deploying deep neural networks (DNNs) on resource-constrained systems necessitates innovative strategies to minimize computational and storage demands. Enhancements in convolutional neural networks (CNNs) for object detection, such as improved accuracy and speed, are crucial [2]. Moreover, ensemble knowledge distillation has demonstrated its ability to enhance generalization and classification performance in compact CNNs, despite existing methods' limitations regarding computational efficiency [3]. The substantial memory and computation requirements of CNNs remain significant barriers to their deployment on resource-limited systems [4].

Generated by SurveyX.cn.

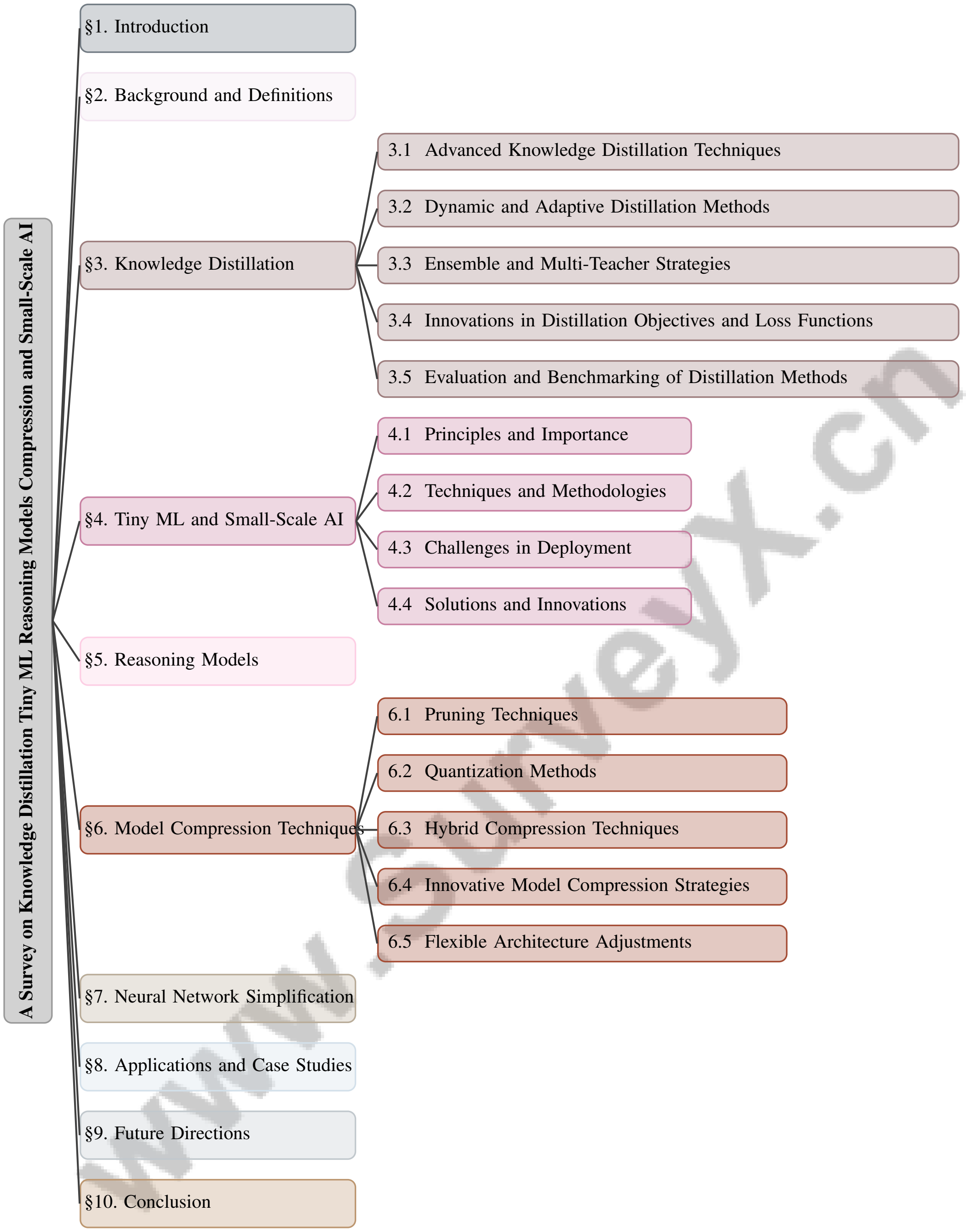

A Survey on Knowledge Distillation Tiny ML Reasoning Models Compression and Small-Scale AI
§1. Introduction
§2. Background and Definitions
§3. Knowledge Distillation
3.1 Advanced Knowledge Distillation Techniques
3.2 Dynamic and Adaptive Distillation Methods
3.3 Ensemble and Multi-Teacher Strategies
3.4 Innovations in Distillation Objectives and Loss Functions
3.5 Evaluation and Benchmarking of Distillation Methods
§4. Tiny ML and Small-Scale AI
4.1 Principles and Importance
4.2 Techniques and Methodologies
4.3 Challenges in Deployment
4.4 Solutions and Innovations
§5. Reasoning Models
§6. Model Compression Techniques
6.1 Pruning Techniques
6.2 Quantization Methods
6.3 Hybrid Compression Techniques
6.4 Innovative Model Compression Strategies
6.5 Flexible Architecture Adjustments
§7. Neural Network Simplification
§8. Applications and Case Studies
§9. Future Directions
§10. Conclusion

Figure 1: chapter structure

The proliferation of mobile robots with constrained computing resources necessitates the execution of complex machine learning models directly on these platforms, further emphasizing the importance of model compression and resource optimization [5]. In natural language processing, deploying large language models (LLMs) in sectors like education requires reducing model size and computational needs to address challenges related to resources, latency, and energy efficiency. Integrating minimum Bayes risk (MBR) decoding into knowledge distillation processes enhances the transfer of advanced capabilities from larger models to more efficient counterparts [6].

These collective efforts underscore the imperative to minimize complexity and optimize resource utilization in AI models through advanced techniques such as model compression, knowledge distillation, and efficient architecture design. These strategies not only improve training efficiency and reduce memory footprint but also facilitate the effective deployment of sophisticated models on devices with limited computational capabilities, addressing the growing demand for AI applications in resource-constrained environments [7, 8, 9, 10, 11].

1.2 Importance in Resource-Constrained Deployments

Deploying AI models on resource-constrained devices is crucial due to their inherent limitations in computational power, storage capacity, and energy efficiency. The impracticality of large and complex models in real-world applications highlights the need for efficient AI deployment in such settings [5]. This is particularly relevant in edge computing environments, where the high computational and storage demands of deep neural networks (DNNs) pose significant challenges [12]. In applications like CNN-based object detection, maintaining high accuracy while minimizing processing time is essential for practical deployment [2].

Traditional federated learning methods struggle with non-i.i.d. data distributions, a common issue in resource-constrained environments, necessitating more efficient AI deployment strategies [1]. Moreover, the deployment of large language models (LLMs) in fields such as education faces hurdles due to their computational demands and latency issues [13]. The complexity involved in transferring knowledge from sophisticated LLM teachers to simpler student models, particularly in translation tasks, underscores the need for efficient distillation methods that can maintain performance [6].

Knowledge distillation encounters challenges, particularly in aligning feature distributions between teacher and student networks, which can result in suboptimal performance in student models [4]. Existing techniques often struggle to capture nuanced differences between teacher and student outputs, leading to performance degradation [14]. Furthermore, the slow inference speed of large models in applications such as web-based question answering systems accentuates the necessity for compression techniques that can reduce latency and enhance throughput [15].

The effective deployment of over-parameterized neural networks on resource-constrained devices demands robust knowledge transfer methods to ensure that smaller models can perform adequately without the extensive resources required by their larger counterparts [16]. Benchmarking results illustrate the challenge of compressing DNNs without significant performance loss, marking this as a critical area for research [3]. Addressing these multifaceted challenges is essential for the practical deployment of AI across various applications in resource-limited environments.

1.3 Structure of the Survey

This survey is structured to provide a comprehensive exploration of methodologies and techniques aimed at simplifying machine learning models for deployment on resource-constrained devices. It begins with an introduction that discusses the overarching theme of complexity reduction and resource optimization in AI models, emphasizing their significance in environments with limited computational resources. Following this, a detailed background section defines core concepts such as Knowledge Distillation, Tiny ML, Reasoning Models, Model Compression, Neural Network Simplification, and Small-Scale AI, establishing a foundation for subsequent discussions.

The survey delves into specific areas of interest, starting with Knowledge Distillation, where various methodologies and advancements are examined. This includes discussions on advanced distillation techniques, dynamic and adaptive methods, ensemble strategies, and innovations in distillation objectives and loss functions. The section concludes with an analysis of how these methods are evaluated and benchmarked for effectiveness.

Subsequently, the principles and applications of Tiny ML and Small-Scale AI are explored, addressing the challenges and solutions associated with deploying efficient models on constrained devices. The survey thoroughly examines Reasoning Models, highlighting their significance in streamlining neural architectures without compromising performance. It also investigates enhancement techniques, such as knowledge distillation and architectural innovations, alongside integration strategies that facilitate the deployment of smaller, more efficient models while preserving reasoning capabilities. Additionally, the effectiveness of alternative reasoning schemes, like SOCRATIC COT, in improving the performance of distilled models in complex problem-solving scenarios is discussed [11, 17, 18, 19].

AI-generated, for reference only.

Model Compression Techniques are detailed, covering methods such as pruning, quantization, and hybrid approaches, along with innovative strategies and flexible architecture adjustments. This discussion is enriched by an in-depth analysis of Neural Network Simplification, focusing on advanced strategies for architecture search and parameter reduction, including knowledge distillation techniques that optimize the performance of smaller models, as well as a novel architecture slimming method that automates the determination of compressed architectures while preserving model capacity. This includes examining loss functions specific to cross-encoder architectures and integrating principal component analysis to maximize parameter variance, ultimately enhancing efficiency in document ranking tasks and overall model performance [11, 8].

The survey also features a section on Applications and Case Studies, providing real-world examples and highlighting the benefits and challenges of implementing these techniques. The paper concludes with a discussion on Future Directions, considering potential research avenues and the integration of emerging technologies, followed by a concise conclusion that reiterates the significance of the discussed topics in advancing AI deployment on resource-constrained devices.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Background and Definitions

The exploration of Knowledge Distillation (KD), Tiny ML, Reasoning Models, Model Compression, Neural Network Simplification, and Small-Scale AI is critical for optimizing machine learning models, particularly in resource-constrained environments. Techniques such as reverse Kullback-Leibler divergence in the MiniLLM approach exemplify KD's capability to transform large language models (LLMs) into smaller, efficient versions, enhancing performance and scalability across various model sizes [5, 20]. These methodologies aim to reduce model complexity while maintaining or improving performance.

Knowledge Distillation involves training a smaller 'student' model to emulate a larger 'teacher' model's outputs, thereby retaining performance with reduced computational demands [5]. This approach is beneficial in data-scarce scenarios like medical image classification, addressing model compression challenges through response-based, feature-based, and relation-based knowledge transfer, which enhance interpretability across domains such as banking [16, 12].

Tiny ML focuses on developing compact machine learning models for low-power devices, crucial for AI applications within the Internet of Things (IoT) [2]. Techniques such as model compression and neural network simplification optimize resource usage.

Reasoning Models enhance AI decision-making by integrating cognitive abilities and specialized knowledge, streamlining neural architectures without compromising performance. Techniques like model compression—including quantization, pruning, and KD—shrink model size and inference time while maintaining accuracy, enabling large-scale models' deployment in resource-limited settings [7, 10, 11, 8]. Symbolic knowledge distillation, a subset of reasoning models, converts implicit knowledge from LLMs into explicit forms, enhancing interpretability and efficiency.

Model Compression includes techniques such as pruning, quantization, and low-precision networks, essential for deploying AI on devices with limited computational and storage capabilities [3]. Quantization-aware training (QAT) is notable for reducing model precision while maintaining accuracy, despite often requiring complex training procedures.

Neural Network Simplification employs strategies like architecture search and parameter reduction to develop efficient models, crucial for resource-constrained environments [2]. These techniques are categorized into parameter pruning and quantization, low-rank factorization, transferred/compact convolutional filters, and knowledge distillation.

Small-Scale AI adapts AI technologies for effective operation on limited-resource devices. Although pre-trained language models (PLMs) excel in natural language processing tasks, their size often makes them impractical for low-capacity deployment. Thus, the field emphasizes model compression techniques—such as quantization, pruning, and KD—to improve efficiency in constrained environments. A research gap remains in the compression of decoder-based models, essential for edge device

AI-generated, for reference only.

applicability [9, 21, 22, 23]. Efficient training methods are crucial, as traditional KD can increase computational costs due to reliance on teacher networks.

Integrating these concepts highlights the need for AI models that are both efficient and effective in diverse constrained environments. Techniques like Memory Access Prediction (MAP) models optimize resource usage by enhancing data prefetching and minimizing memory latency. Recent advancements, such as Pattern-Clustered Knowledge Distillation (PaCKD), significantly improve model efficiency, achieving a 552-fold reduction in model size with minimal accuracy loss, aligning with the survey's emphasis on optimizing resource utilization and addressing the challenges of deploying large models in real-world applications [7, 24, 10, 25, 23]. Additionally, the interplay between data augmentation and model compression in supervised classification settings offers further opportunities for advancing model efficiency.

3 Knowledge Distillation

In recent years, knowledge distillation has garnered significant attention as a powerful technique for transferring knowledge from complex teacher models to simpler student models, thereby enhancing their performance while maintaining computational efficiency. This section delves into the various advanced techniques that have emerged within the domain of knowledge distillation, highlighting their innovative approaches and contributions to the field. The subsequent subsection will explore these advancements in greater detail, specifically focusing on advanced knowledge distillation techniques that have been developed to overcome the limitations of traditional methods and improve the overall efficiency of the distillation process.

3.1 Advanced Knowledge Distillation Techniques

Advanced knowledge distillation techniques have evolved to address the limitations of traditional methods by enhancing the efficiency and adaptability of the distillation process. A notable advancement in this domain is the use of multiple teacher models with diverse architectures, as exemplified by the Multi-Teacher Knowledge Distillation (MTKD) method. MTKD leverages the complementary strengths of different teachers to provide a more comprehensive and effective knowledge transfer, surpassing approaches that rely on a single teacher model [26].

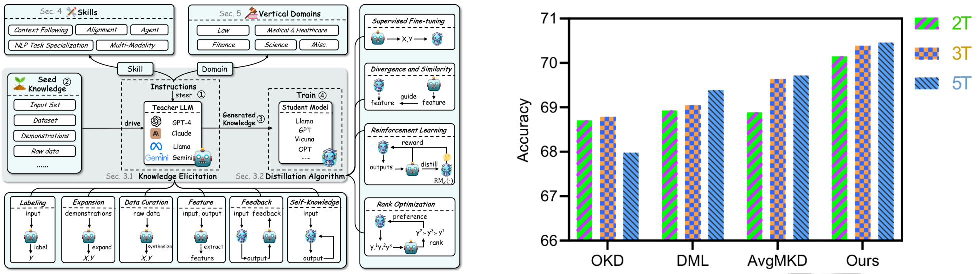
Data-Free Knowledge Distillation (DFKD) represents a significant leap forward, enabling the training of student models without access to the original dataset. This is particularly beneficial in scenarios where data privacy is a concern. Techniques such as Data-Free Knowledge Distillation with Soft Targeted Transfer Set Synthesis (DFKD-STTS) synthesize pseudo-training samples by modeling the intermediate feature space of a teacher model, effectively circumventing data availability constraints [27].

Self-supervised learning has also been integrated into the distillation process to enrich the learning experience. The Batch Knowledge Ensembling (BAKE) method, for instance, achieves better soft targets for training with minimal overhead by ensembling knowledge within batches, thereby enhancing the robustness of the student model [28]. Moreover, SimKD proposes reusing the teacher's classifier for student inference and aligning features through a simple L2 loss, aiming to eliminate the performance gap without complex representations [29].

Innovative strategies like the Random Intermediate Layer Knowledge Distillation (RAIL-KD) introduce randomness in selecting teacher layers for distillation, which significantly reduces the need for hyperparameter tuning and computational costs [30]. Additionally, the TMKD approach combines early calibration through multi-teacher knowledge distillation during the training phase, contrasting with traditional late calibration methods [15].

The exploration of selective knowledge distillation methods, which involve choosing training samples based on their properties, further improves the efficiency of the distillation process [27]. Furthermore, the introduction of distinct distillation schemes focusing on hard samples, such as HGMD-weight and HGMD-mixup, has been shown to enhance the learning process by concentrating on challenging instances [31].

These advanced techniques collectively illustrate the transformative potential of knowledge distillation in improving model efficiency and adaptability. By employing advanced techniques such as self-supervised learning, data-free methods, and collaborative frameworks, knowledge distillation emerges as a crucial approach for enhancing the performance of machine learning models, particularly when deploying them on resource-constrained devices. For instance, self-supervised learning enables models to learn feature representations without the need for labeled data, while strategies like Distill-on-the-Go (DoGo) facilitate the online distillation process between models, leading to improved representation quality, especially in smaller networks. Furthermore, by exploring cross-stage connection paths in distillation, researchers have developed efficient mechanisms that significantly boost student network performance across various tasks, all while minimizing computational overhead. This adaptability of knowledge distillation not only optimizes model efficiency but also enhances generalization, making it a vital tool in the field of machine learning. [32, 33, 34]



(a) Skill-based Knowledge Distillation for Multi-modal NLP[35]

(b) Comparison of Text Classification Accuracy across Different Models and Data Sets[36]

Figure 2: Examples of Advanced Knowledge Distillation Techniques

As shown in Figure 2, The concept of knowledge distillation has gained significant traction in the field of artificial intelligence, particularly in the realm of Natural Language Processing (NLP). This process involves transferring knowledge from a large, complex model (teacher) to a smaller, more efficient model (student), aiming to retain high performance while reducing computational costs. Advanced techniques in knowledge distillation are pushing the boundaries of this field, as illustrated by two compelling examples. The first example, "Skill-based Knowledge Distillation for Multimodal NLP," presents a flowchart detailing the stages of distilling knowledge for multimodal NLP tasks. This process begins with "Seed Knowledge," which encompasses various types of foundational data such as input sets, datasets, demonstrations, and raw data, all crucial for guiding the knowledge elicitation process. The second example, "Comparison of Text Classification Accuracy across Different Models and Data Sets," features a bar chart that vividly contrasts the accuracy of text classification models applied to various datasets. This comparison highlights the performance of models such as OKD, DML, AvgMKD, and a novel approach labeled "Ours," with accuracy metrics spanning from 66 to 71. These visual aids underscore the advancements in knowledge distillation techniques, offering insights into their application and efficacy in optimizing NLP models. [**?**]xu2024surveyknowledgedistillationlarge,liu2020adaptive)

3.2 Dynamic and Adaptive Distillation Methods

Dynamic and adaptive distillation methods, such as spot-adaptive knowledge distillation and dynamic importance sampling, have emerged as critical strategies for optimizing knowledge transfer in machine learning. These approaches enhance flexibility by allowing the selection of distillation spots to vary based on training samples and epochs, as well as reducing computational costs through efficient sampling techniques. Furthermore, dynamic knowledge distillation frameworks facilitate the transfer of knowledge from proprietary black-box models to specific local datasets, thereby addressing challenges in real-world applications like healthcare. Collectively, these methods significantly improve the effectiveness of knowledge transfer processes across diverse machine learning contexts, including federated learning and model compression [37, 38, 39, 40, 41]. These methods are designed to optimize the distillation process by adapting to the specific characteristics of both the teacher and student models, as well as the data being processed.

As illustrated in Figure 3, the hierarchical structure of dynamic and adaptive distillation methods highlights key approaches such as spot-adaptive knowledge distillation, ideal joint classifier knowledge distillation, and adaptive multi-teacher multi-level knowledge distillation. Each method's core

AI-generated, for reference only.

strategies and innovations are depicted, emphasizing their roles in optimizing knowledge transfer processes in machine learning.

The Ideal Joint Classifier Knowledge Distillation (IJCKD) exemplifies this approach by focusing on minimizing the error of the student network in relation to the teacher's error while also addressing discrepancies between their classifiers. This dual focus facilitates improved knowledge transfer, ensuring that the student model not only learns from the teacher's outputs but also aligns its classification boundaries more closely with those of the teacher [42].

Another innovative approach is the Adaptive Multi-Teacher Multi-Level Knowledge Distillation (AMTML-KD), which integrates high-level soft-target knowledge with intermediate-level hints from multiple teachers. By employing a multi-group hint strategy, AMTML-KD dynamically selects the most relevant information from each teacher, thereby enhancing the student's learning process and improving its performance across different tasks [36].

Feature-based knowledge distillation methods further contribute to the adaptability of the distillation process by focusing on the alignment of intermediate features between the teacher and student models. By evaluating the ability of these methods to improve student performance through feature alignment, researchers have demonstrated the potential of dynamic distillation strategies to optimize the learning process in diverse settings [16].

The implementation of dynamic and adaptive distillation methods, such as spot-adaptive knowledge distillation (SAKD) and dynamic importance sampling, emphasizes the necessity of customizing the knowledge transfer process to align with the unique requirements of both the models and the specific tasks. SAKD enhances distillation by adaptively selecting the optimal layers in the teacher network for each training sample and iteration, while dynamic importance sampling reduces computational costs by selectively focusing on the most relevant classes during training, thereby improving efficiency without compromising performance [41, 38]. By leveraging the strengths of multiple teachers and focusing on both output and feature alignment, these methods enhance the overall efficiency and effectiveness of knowledge distillation, making them invaluable tools for deploying AI models on resource-constrained devices.

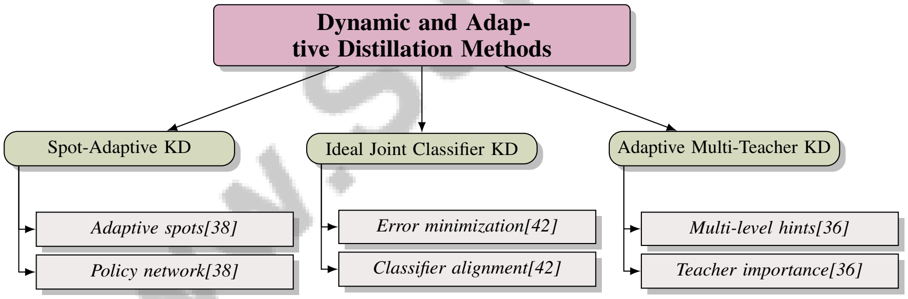


Figure 3: This figure illustrates the hierarchical structure of dynamic and adaptive distillation methods, highlighting key approaches such as spot-adaptive knowledge distillation, ideal joint classifier knowledge distillation, and adaptive multi-teacher multi-level knowledge distillation. Each method's core strategies and innovations are depicted, emphasizing their roles in optimizing knowledge transfer processes in machine learning.

3.3 Ensemble and Multi-Teacher Strategies

Ensemble and multi-teacher strategies in knowledge distillation have gained prominence as effective methods for enhancing the capabilities of student models by harnessing the unique strengths of multiple teacher models. These approaches not only facilitate the transfer of knowledge from various teachers to a single student model but also incorporate adaptive techniques to prioritize the importance of each teacher's contributions based on their prediction accuracy. For instance, recent advancements in ensemble knowledge distillation emphasize the use of both labeled and unlabeled data to optimize knowledge transfer, while multi-teacher frameworks allow for the integration of instance-level importance weights and intermediate-level hints, thereby enriching the learning experience and improving overall model performance. Such methodologies have been shown to significantly boost classification accuracy and model generalization, particularly in scenarios with limited training data.

[43, 44, 36, 45, 46]. These strategies aim to provide richer and more comprehensive supervision, which is crucial for improving the performance of student networks, particularly in complex tasks.

Multi-Teacher Knowledge Distillation (MTKD) is a prominent strategy that utilizes knowledge from multiple teacher models to enhance various components of a target model. For instance, MTKD has been effectively applied to improve the image encoder, sequential encoder, and decoder of the TIMT model by integrating insights from three distinct teacher models [47]. This approach not only enriches the learning experience of the student model but also addresses the limitations of relying on a single teacher, which may not capture the full spectrum of knowledge necessary for optimal performance.

Despite the advantages, multi-teacher distillation methods often face challenges related to computational complexity. The integration of knowledge from multiple teachers can be computationally expensive, necessitating innovative solutions to balance the trade-off between performance gains and resource consumption [48]. To mitigate these challenges, recent research has focused on developing efficient frameworks that minimize computational overhead while maximizing the benefits of multi-teacher distillation.

Relation-based knowledge distillation approaches further contribute to the ensemble and multi-teacher strategies by capturing high-order relationships between data instances. These approaches emphasize the importance of understanding the intricate mechanisms behind knowledge transfer, moving beyond intuition to establish a more rigorous theoretical framework [16]. By analyzing the relational dynamics between teacher and student models, relation-based KD provides unique contributions to the overall efficacy of knowledge transfer.

The primary challenge in ensemble and multi-teacher strategies lies in understanding the complex mechanisms underlying knowledge transfer. Existing methods often rely on intuition, highlighting the need for a more structured theoretical framework to guide the design and implementation of these strategies [42]. Addressing this challenge is essential for advancing the field and realizing the full potential of ensemble and multi-teacher strategies in knowledge distillation.

3.4 Innovations in Distillation Objectives and Loss Functions

Recent advancements in distillation objectives and loss functions have significantly enhanced the efficiency and adaptability of knowledge transfer, facilitating the deployment of AI models in resource-constrained environments. A notable innovation is the integration of Knowledge Distillation (KD) with Mixture of Experts (MoE) models, which enables the development of modular architectures capable of dynamically allocating tasks to specialized experts [49]. This approach leverages the strengths of both KD and MoE to optimize task performance and model efficiency.

In the realm of object detection, new loss functions such as weighted cross-entropy loss and teacher bounded loss have been introduced to address the complexities inherent in this domain [2]. These loss functions are specifically designed to handle the diverse challenges of object detection tasks, ensuring more accurate and robust model performance.

The application of triplet loss in knowledge distillation represents another significant innovation, allowing student models to learn effectively from similar outputs while distinguishing between dissimilar outputs [14]. This approach enhances the discriminative power of student models, contributing to improved performance across various tasks.

FedSiKD exemplifies the use of knowledge distillation to address non-i.i.d. data issues by effectively transferring knowledge between teacher and student models [1]. This method underscores the importance of tailored distillation strategies in federated learning environments, where data heterogeneity poses significant challenges.

The Similarity Transfer Knowledge Distillation (STKD) method introduces a novel approach by transferring similarity correlations between instances using virtual samples created through the mixup technique [50]. This departure from traditional instance-based methods enables more comprehensive knowledge transfer, enhancing model generalization capabilities.

Innovations in the use of multiple candidate sequences from Minimum Bayes Risk (MBR) decoding for knowledge distillation offer a new perspective on leveraging diverse outputs for enhanced learning

AI-generated, for reference only.

[6]. This approach contrasts with traditional methods that focus on single best outputs, providing a richer set of training signals for student models.

The introduction of reverse Kullback-Leibler Divergence (KLD) as an objective in the M INI LLM method allows for more accurate and reliable student model outputs, highlighting the potential of reverse divergence measures in distillation processes [20]. Additionally, derivative matching and online distillation techniques have been developed to further refine the performance of simpler models [5].

These recent innovations in knowledge distillation highlight a significant shift in objectives and loss functions, addressing challenges such as distribution shifts and the representation gap between teacher and student models. By introducing frameworks like DiffKD, which utilizes diffusion models for feature denoising, and DistPro, which employs meta-optimization to discover optimal distillation pathways, these advancements pave the way for more efficient and robust knowledge transfer processes. Such improvements are crucial for the deployment of AI models in diverse and resource-constrained environments, particularly in contexts like federated learning where communication efficiency and model adaptability are paramount. [51, 52, 53, 39, 54]

3.5 Evaluation and Benchmarking of Distillation Methods

Benchmark	Size	Domain	Task Format	Metric
KD-Edge[55]	1,000,000	Computer Vision	Knowledge Distillation	Accuracy, KD Time
CNN-COMP[56]	70,000	Image Recognition	Image Classification	Accuracy, Model Size
MT-Compress[57]	3,000,000	Machine Translation	Translation	BLEU, chrF2
ComKD[58]	1,000,000	Natural Language Inference	Classification	Accuracy, F1-score
DM-KD[59]	200,000	Image Classification	Knowledge Distillation	Top-1 Accuracy
KD[60]	1,000,000	Image Classification	Knowledge Distillation	Top-1 Error
KD-TF[21]	600	Natural Language Understanding	Classification	Accuracy, F1-score
KD-Bench[39]	60,000	Image Classification	Image Classification	Accuracy Gain, Forgetting

Table 1: This table presents a comprehensive overview of various benchmarks used in the evaluation of knowledge distillation methods across different domains and tasks. It details the size of each benchmark, the specific domain it pertains to, the task format involved, and the metrics employed for performance assessment. The inclusion of diverse benchmarks underscores the multifaceted nature of knowledge distillation evaluation, spanning computer vision, image recognition, machine translation, and natural language processing.

The evaluation and benchmarking of knowledge distillation methods are crucial for determining their effectiveness in improving model performance and optimizing resource efficiency. Table 1 provides a detailed overview of the benchmarks utilized in the evaluation and benchmarking of knowledge distillation methods, highlighting their size, domain, task format, and the metrics used for performance assessment. The assessment of student models distilled through various techniques utilizes a range of metrics and datasets, enabling a comprehensive evaluation of performance improvements. This analysis not only highlights enhancements in accuracy but also sheds light on the interpretability of models, as knowledge distillation facilitates the transfer of class-similarity information from teacher to student networks. Additionally, the impact of data augmentation strategies, such as Mixed Sample Regularization, on generalization capabilities during the distillation process is examined, revealing potential trade-offs in learning example-specific features. Overall, the findings provide valuable insights into the multifaceted benefits of knowledge distillation across different model architectures and tasks. [34, 61, 62, 24, 63]

In the context of classification tasks, methods like SimKD have been evaluated on datasets such as CIFAR-100 and ImageNet, where they are compared against state-of-the-art approaches, including vanilla KD and FitNet, using metrics like top-1 test accuracy [29]. Similarly, BAKE has demonstrated consistent improvements in classification performance across various architectures with minimal computational overhead, outperforming state-of-the-art single-network methods on all benchmarks [28].

For tasks involving natural language processing, RAIL-KD was assessed on eight tasks from the GLUE benchmark, where it was compared against baseline methods such as vanilla KD, PKD, and ALP-KD [30]. The performance of the proposed method in selective knowledge distillation was evaluated by measuring BLEU scores against baseline models, employing a maximum training step of 300K and analyzing the effects of different sample selection strategies [27].

Furthermore, innovative approaches like M INI LLM have been evaluated using various metrics such as Rouge-L scores, GPT-4 feedback, and human evaluations, comparing their performance against standard KD baselines [20]. These evaluations highlight the potential of advanced distillation techniques to enhance model performance while maintaining efficiency.

Overall, rigorous evaluation and benchmarking of knowledge distillation methods provide valuable insights into their performance and resource demands. By utilizing a diverse array of metrics and datasets, researchers can gain deeper insights into the strengths and weaknesses of various knowledge distillation techniques, such as the novel residual learning frameworks and mixed sample augmentation strategies, thereby informing and driving future innovations in the field of model training and interpretability. [62, 63, 61, 24]

4 Tiny ML and Small-Scale AI

Tiny ML and Small-Scale AI have significantly reshaped the machine learning landscape by focusing on creating efficient models that operate effectively within limited computational resources. This shift is essential for deploying AI technologies in diverse environments, requiring compact, high-performing models. Knowledge distillation plays a crucial role in enhancing smaller models using advanced large language models (LLMs), bridging foundational and specialized models. Innovative strategies, such as interpretable knowledge distillation and teaching committees, optimize performance while addressing computational limitations and privacy concerns in edge device deployment [64, 65, 20].

Figure 4 illustrates the hierarchical structure of Tiny ML and Small-Scale AI, highlighting key principles, methodologies, challenges, and innovations. This figure categorizes the essential aspects of efficient model deployment on resource-constrained devices, emphasizing knowledge distillation techniques, adaptive learning, and innovative frameworks to enhance AI applications across diverse environments. By integrating this visual representation, we can better understand the intricate relationships and strategies that underpin the successful implementation of these technologies.

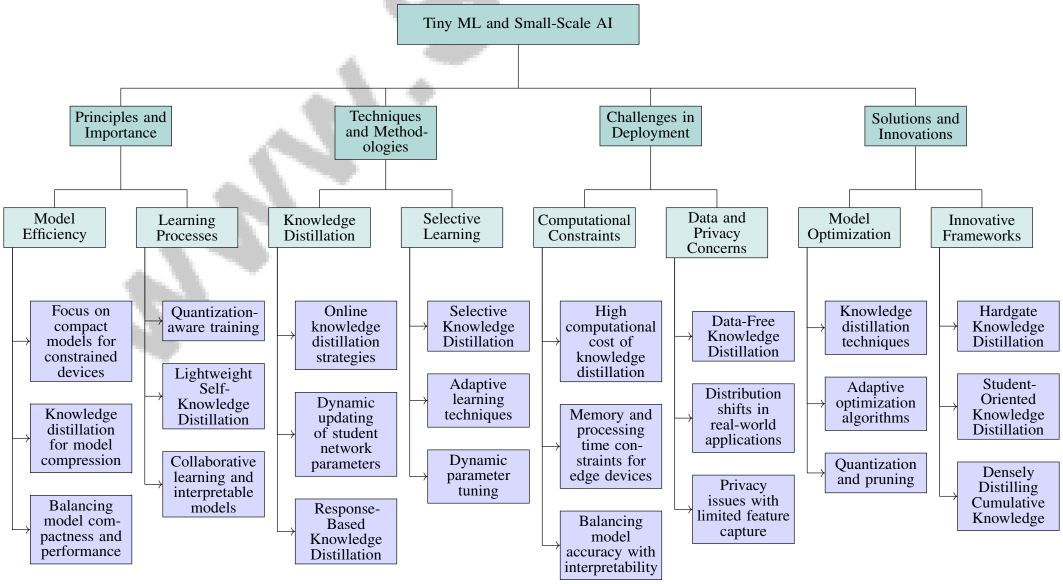

Tiny ML and Small-Scale AI
Principles and Importance
Techniques and Methodologies
Challenges in Deployment
Solutions and Innovations
Model Efficiency
Learning Processes
Knowledge Distillation
Selective Learning
Computational Constraints
Data and Privacy Concerns
Model Optimization
Innovative Frameworks
Focus on compact models for constrained devices
Knowledge distillation for model compression
Balancing model compactness and performance
Quantization-aware training
Lightweight Self-Knowledge Distillation
Collaborative learning and interpretable models
Online knowledge distillation strategies
Dynamic updating of student network parameters
Response-Based Knowledge Distillation
Selective Knowledge Distillation
Adaptive learning techniques
Dynamic parameter tuning
High computational cost of knowledge distillation
Memory and processing time constraints for edge devices
Balancing model accuracy with interpretability
Data-Free Knowledge Distillation
Distribution shifts in real-world applications
Privacy issues with limited feature capture
Knowledge distillation techniques
Adaptive optimization algorithms
Quantization and pruning
Hardgate Knowledge Distillation
Student-Oriented Knowledge Distillation
Densely Distilling Cumulative Knowledge

Figure 4: This figure illustrates the hierarchical structure of Tiny ML and Small-Scale AI, highlighting key principles, methodologies, challenges, and innovations. It categorizes the essential aspects of efficient model deployment on resource-constrained devices, emphasizing knowledge distillation techniques, adaptive learning, and innovative frameworks to enhance AI applications across diverse environments.

AI-generated, for reference only.

4.1 Principles and Importance

Tiny ML and Small-Scale AI focus on developing highly efficient, compact models for devices with constrained computational resources, such as IoT microcontrollers. Techniques like Knowledge Distillation compress large language models into smaller versions, maintaining performance and enabling complex tasks in resource-limited settings. Model compression methods, including multi-segment activation and modular knowledge distillation, enhance these compact models' capabilities, expanding their applications in IoT and edge computing [12, 20, 66, 67, 68]. The core principles of Tiny ML emphasize balancing model compactness and performance, ensuring high efficiency under constrained conditions, critical for AI applications with limited power and data availability.

A key aspect of Tiny ML is integrating quantization-aware training with knowledge distillation, as demonstrated by frameworks like CAKD, which focus on influential components of the distillation process to facilitate effective knowledge transfer [69]. This highlights the importance of efficient learning processes, enabling superior performance even in data-scarce scenarios [70].

Methods like Lightweight Self-Knowledge Distillation (LightSKD) enhance model robustness and performance by incorporating diverse feature representations from various layers [71]. This aligns with Small-Scale AI principles, emphasizing adaptability to different computational capabilities for resource-constrained device deployment.

Collaborative learning and simpler, interpretable models are central to Tiny ML and Small-Scale AI. Knowledge distillation techniques facilitate deploying these models in complex environments, such as banking, by utilizing simpler architectures without compromising interpretability or performance [72]. Furthermore, methods like GTN advance Neural Architecture Search (NAS), allowing for the configuration of new student models with enhanced accuracy [73].

Structured knowledge from self-supervised learning is crucial for enabling models to generalize better and withstand noise during training, essential for Tiny ML applications [74]. Leveraging diverse datasets ensures models are trained across varied scenarios, enhancing their real-world applicability.

The proposed method emphasizes the synergy between a high-capacity teacher model and a lightweight student model to improve efficiency in object detection, demonstrating practical Tiny ML applications [2]. Effective knowledge distillation relies on the student's ability to approximate the teacher's predictive distribution, often hindered by optimization challenges [33]. Advancements in KD have markedly improved smaller models' performance through effective knowledge transfer from larger teacher models, enhancing model efficiency [16]. Notably, MBR-n exhibits significant improvements in data efficiency and translation quality, highlighting its potential as a superior knowledge distillation choice [6].

4.2 Techniques and Methodologies

The evolution of Tiny ML and Small-Scale AI is driven by innovative techniques aimed at creating efficient models for resource-constrained devices. Versatile frameworks like TCS are applicable across various architectures, making them promising for both traditional knowledge distillation (KD) and few-shot learning scenarios [75].

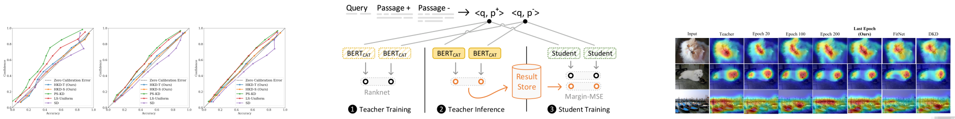
Online knowledge distillation strategies, involving multiple student branches learning from ground truth labels and weighted ensemble predictions, enhance adaptability and efficiency, culminating in a group leader model optimized for deployment [46].

Dynamic updating of student network parameters addresses the retention and integration of new knowledge while maintaining performance on previously encountered examples, effectively combating deployment challenges in dynamic environments [76]. Additionally, dynamic temperature adjustments in Meta Knowledge Distillation (MKD) optimize distillation performance by tailoring the learning process to each model's needs [77].

Response-Based Knowledge Distillation (RBKD) enhances efficiency by allowing the student model to learn about multiple classes from a single input, maximizing training utility [78]. Methods like PS-KD focus on hard examples during training, enabling models to learn from their mistakes and improve generalization [79].

Selective Knowledge Distillation prioritizes training samples based on cross-entropy scores, optimizing the process by focusing on the most informative examples [27]. Utilizing datasets like CIFAR-100 provides a robust foundation for evaluating these methodologies [24].

These techniques underscore the innovative strategies employed in Tiny ML and Small-Scale AI to achieve effective model deployment on resource-limited devices. By integrating adaptive learning techniques, dynamic parameter tuning, and selective knowledge transfer methods, these approaches significantly enhance AI models' performance and efficiency, particularly in constrained environments such as federated learning and healthcare [37, 39, 42, 80].



(a) Comparison of Confidence and Accuracy for Different Methods in Image Classification[81]

(b) Teacher-Student Training Framework for Answering Questions in Text[82]

(c) Comparison of Image Restoration Techniques on Different Scenes[44]

Figure 5: Examples of Techniques and Methodologies

As shown in Figure 5, Tiny ML and Small-Scale AI are rapidly evolving, offering innovative techniques and methodologies that cater to the constraints of limited computational resources while maintaining impressive performance. The provided examples illustrate a range of applications, from image classification to text-based question answering and image restoration. The first example highlights a comparative analysis of confidence and accuracy across various image classification methods, showcasing the efficacy of techniques like Zero Calibration Error and HKD variants. The second example employs a teacher-student training framework to enhance the accuracy of answering text-based queries, leveraging models like BERTCAT to optimize the learning process. Lastly, the image restoration techniques comparison demonstrates the progression of model performance over training epochs, emphasizing the refinement of outputs achieved through methods like FitNet and DKD. Collectively, these examples underscore the potential of Tiny ML and Small-Scale AI in delivering robust solutions across diverse applications, even under resource constraints [81, 82, 44].

4.3 Challenges in Deployment

Deploying Tiny ML and Small-Scale AI solutions presents several challenges, particularly regarding computational constraints and model efficiency. The computational cost of knowledge distillation processes involving multiple teacher networks is high, making it impractical for resource-constrained environments due to the memory overhead required to generate soft targets from multiple networks [28]. Additionally, significant divergence in data distributions among clients hampers federated learning deployment, necessitating strategies to handle non-i.i.d. data effectively [1].

In educational settings, the deployment of large language models (LLMs) is hindered by their substantial size and computational demands, limiting practical applications in schools with restricted processing power [13]. This issue is compounded by the need to balance model accuracy with interpretability in practical applications, such as retail banking, where complex models often compromise transparency [72].

Designing lightweight deep neural networks (DNNs) for edge devices must address memory and processing time constraints to ensure efficient operation [12]. Furthermore, employing cheap convolutions reduces computational costs and improves performance but requires careful consideration to avoid extensive redesign [83].

Data privacy concerns present another significant challenge, especially where access to original training data is restricted. Data-Free Knowledge Distillation techniques enable model compression without the original datasets, addressing privacy issues. However, reliance on computational resources and limited feature capture in existing Self-Knowledge Distillation methods hinder effectiveness, posing barriers to efficient deployment [71].

Moreover, distribution shifts in real-world applications complicate AI model deployment, highlighting the necessity for robust knowledge distillation techniques to maintain performance under varying conditions [54]. Careful selection of student models is crucial to avoid redundancy in knowledge, as demonstrated in methods requiring multiple student networks [84]. Challenges persist in achieving reliable performance in few-shot network compression, with high estimation errors during inference leading to overfitting on limited training instances [39].

The multifaceted challenges associated with deploying Tiny ML and Small-Scale AI solutions highlight the intricate nature of implementing these technologies in resource-constrained environments. This complexity necessitates continuous research and innovation, particularly in knowledge distillation (KD) for pre-trained models and optimizing model architectures. Recent studies show that enhancing KD techniques can significantly improve distributed training efficiency and federated learning, reducing communication demands and facilitating effective AI applications in collaborative settings. Moreover, advancements in KD methodologies, such as MiniLLM, illustrate the potential for smaller models to replicate larger language models' performance while maintaining high quality and scalability, underscoring the importance of ongoing exploration to refine AI technologies for practical use in limited-resource scenarios [39, 20].

4.4 Solutions and Innovations

Deploying Tiny ML and Small-Scale AI solutions on resource-constrained devices necessitates innovative strategies to overcome challenges such as computational limitations and data privacy concerns. Recent advancements in model optimization have led to innovative solutions aimed at enhancing the efficiency and adaptability of large-scale language models. Techniques like knowledge distillation, which trains smaller student models to emulate larger teacher models' performance, have been refined to reduce online inference latency without sacrificing accuracy. Tailored loss functions and strategies, such as pairwise training samples for cross-encoder architectures, are critical for improving document ranking tasks. Online knowledge distillation methods promoting diversity among multiple student models have shown promise in enhancing performance. Additionally, integrating adaptive optimization algorithms, parallel computing, and model compression techniques—such as quantization and pruning—has been systematically reviewed, highlighting their effectiveness in accelerating convergence and reducing memory footprints while maintaining predictive accuracy. These advancements ensure effective model deployment and performance optimization in real-world applications [7, 11, 46].

Notable innovations include the Hardgate Knowledge Distillation (HKD) approach, which enhances model calibration and generalization without increasing learnable parameters compared to baseline methods, maintaining model performance within resource constraints [81]. Similarly, the Student-Oriented Knowledge Distillation (SoKD) framework adapts teacher knowledge to meet the specific needs of the student network, improving knowledge transfer efficiency by focusing on distinctive areas of interest [44].

Exploring teacher-free methods in Natural Language Processing (NLP) offers promising solutions, as evidenced by benchmarks suggesting these approaches can match traditional Knowledge Distillation (KD) performance while reducing computational overhead [21]. This reduction in demand is critical for deploying AI models in resource-limited environments.

Densely Distilling Cumulative Knowledge (DKD) demonstrates versatility in offline scenarios, such as model compression, addressing deployment challenges in Tiny ML and Small-Scale AI. By efficiently compressing models without sacrificing performance, DKD facilitates AI solution deployment in constrained environments [85].

Differentiable Feature Aggregation (DFA) offers a cost-effective alternative to multi-teacher distillation, mimicking its benefits while significantly reducing computational costs, making it particularly suitable for resource-constrained environments [48].

Parameter-Efficient Student-Friendly Knowledge Distillation (PESF-KD) reduces computational costs and enhances knowledge transfer efficiency, ideal for deployment in limited-resource environments. By focusing on parameter efficiency, PESF-KD ensures AI models can be deployed without extensive computational requirements [86].

These solutions and innovations illustrate ongoing efforts to optimize AI model deployment in resource-constrained environments. By integrating advanced distillation techniques, such as interpretable knowledge distillation, mixed distillation frameworks, and innovative data-free approaches, researchers significantly enhance smaller language models' (LLMs) performance for Tiny ML and Small-Scale AI applications. These methods improve reasoning capabilities and training efficiency while addressing critical challenges related to computational demands and data privacy. For instance, the interpretable knowledge distillation method allows smaller models to learn from larger models' strategies without direct parameter manipulation, while the mixed distillation framework combines various prompting techniques to boost reasoning accuracy. Additionally, data-free knowledge distillation enables effective training without accessing original datasets, mitigating privacy concerns.

Figure 6 illustrates the diverse strategies and innovations in AI deployment, focusing on knowledge distillation techniques, data-free and efficient methods, and optimization and compression strategies. Each category highlights key advancements in enhancing AI model performance and deployment efficiency in resource-constrained environments. Collectively, these advancements pave the way for more practical and widespread applications of AI technologies across diverse domains [87, 88, 64].

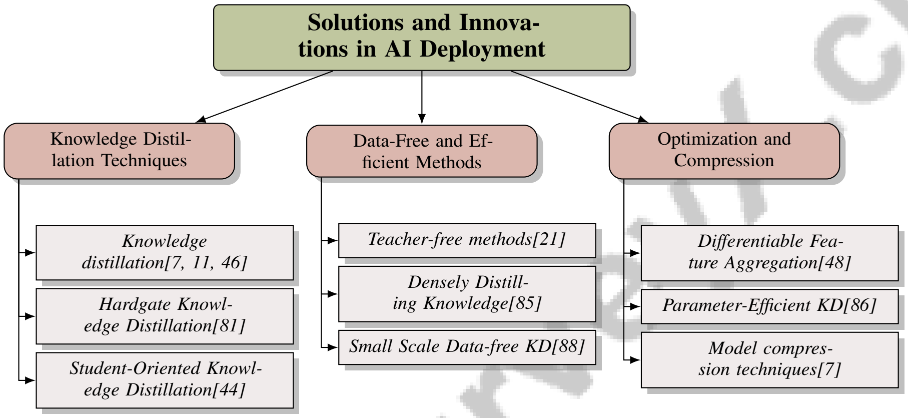

Solutions and Innovations in AI Deployment
Knowledge Distillation Techniques
Knowledge distillation[7, 11, 46]
Hardgate Knowledge Distillation[81]
Student-Oriented Knowledge Distillation[44]
Data-Free and Efficient Methods
Teacher-free methods[21]
Densely Distilling Knowledge[85]
Small Scale Data-free KD[88]
Optimization and Compression
Differentiable Feature Aggregation[48]
Parameter-Efficient KD[86]
Model compression techniques[7]

Figure 6: This figure illustrates the diverse strategies and innovations in AI deployment, focusing on knowledge distillation techniques, data-free and efficient methods, and optimization and compression strategies. Each category highlights key advancements in enhancing AI model performance and deployment efficiency in resource-constrained environments.

5 Reasoning Models

5.1 Enhancing Reasoning with Multi-Prompting Techniques

Multi-prompting techniques have become instrumental in enhancing AI systems' interpretability and performance, particularly in complex decision-making tasks. These techniques utilize multiple prompts to explore diverse perspectives, improving reasoning capabilities and generating comprehensive responses. This approach is particularly effective in customer service, where interpretability and contextual relevance are crucial [64]. By integrating various cognitive abilities, models synthesize information from multiple sources, developing well-rounded solutions to complex problems. The SOCRATIC COT method exemplifies this by showing significant performance improvements on reasoning datasets like GSM8K and SVAMP, demonstrating that smaller models can outperform larger ones in specific scenarios through tailored distillation strategies [11, 18].

Beyond reasoning enhancements, multi-prompting improves model interpretability. Studies show that knowledge distillation can transfer class-similarity information from larger to smaller models, increasing identifiable concept detectors and facilitating reliable applications across fields [87, 63, 18, 64]. By providing clear prompts, these techniques enhance user understanding of the model's decision-making process, fostering trust and transparency—essential in domains like customer service.

The adoption of multi-prompting techniques marks a significant advancement in reasoning models, enhancing AI performance through effective problem decomposition and strategic guidance while improving interpretability. This development addresses security and privacy concerns associated with larger models and plays an increasingly vital role in evolving sophisticated reasoning models capable of tackling complex challenges across diverse applications.

5.2 Integration of Cognitive Abilities and Specialized Knowledge

Integrating cognitive abilities and specialized knowledge in reasoning models is crucial for advancing AI systems, enabling them to perform complex tasks with enhanced precision and adaptability. This integration fosters models that efficiently process information and exhibit a comprehensive understanding of domain-specific knowledge, improving decision-making and interpretability [63, 24]. Cognitive abilities are enhanced through multi-prompting, allowing models to explore diverse perspectives and synthesize information for improved accuracy and reliability [64]. This approach is effective in applications requiring nuanced understanding, such as customer service and medical diagnostics.

Specialized knowledge is integrated through domain-specific datasets and customized training methodologies, such as knowledge distillation and modular architectures, refining knowledge transfer between larger teacher models and smaller student models [24, 89, 44, 18, 49]. By embedding expert knowledge into the architecture, these systems deliver more accurate outputs in specialized fields, such as interpreting complex medical data.

Moreover, integrating cognitive abilities and specialized knowledge enhances model interpretability. Knowledge distillation models provide systematic insights into decision-making processes, fostering trust and transparency, especially in critical sectors like healthcare and finance [72, 24, 64, 63, 65]. The ability to explain and justify decisions based on cognitive reasoning and specialized knowledge ensures users can rely on these models for informed decision-making.

5.3 Variance as a Metric for Information Retention

Variance has emerged as a significant metric for assessing information retention in models, focusing on preserving critical information while reducing complexity. Analyzing variance within model parameters helps identify essential components, enabling effective model slimming without compromising performance [8]. Variance-based methods excel in filter pruning, eliminating redundant filters from convolutional layers in neural networks. By quantifying variance, these methods prioritize retaining filters that significantly contribute to predictive capabilities, ensuring essential functions are preserved while reducing model size.

This optimization is crucial for deployment on devices with limited computational resources, as connected devices necessitate compressed models that maintain performance despite processing and energy constraints. Techniques like weight quantization, parameter pruning, and knowledge distillation enhance model efficiency, ensuring large language and vision models can be effectively utilized in practical applications [10, 9]. The effectiveness of variance lies in capturing the distribution of information across model parameters, facilitating a deeper understanding of components essential for accuracy while identifying those that can be pruned without performance loss. This analysis highlights the importance of cross-level connections and the role of pruned models in knowledge distillation, optimizing knowledge transfer from teacher to student networks [24, 90].

5.4 SOCRATIC COT Method for Problem Decomposition

The SOCRATIC COT method enhances reasoning capabilities in machine learning models by decomposing complex problems into manageable components. This method enables smaller models to engage in effective reasoning processes through structured guidance, facilitating the learning of intricate tasks that would otherwise be challenging [18]. Its efficacy lies in systematically breaking down complex problems, allowing models to tackle each component individually and build a comprehensive understanding incrementally. Utilizing the SOCRATIC COT method allows reasoning models to adopt a structured framework that facilitates learning and ensures comprehensive analysis of each problem component.

Research indicates this approach significantly improves the performance of smaller distilled models, achieving over a 70

AI-generated, for reference only.

5.5 Symbolic Knowledge Distillation for Interpretability

Symbolic Knowledge Distillation (SKD) enhances the interpretability of large language models (LLMs) by transforming implicit knowledge into explicit symbolic forms, fostering a transparent understanding of model decision-making processes. This approach is crucial for applications requiring high trust and accountability [17]. SKD mitigates overfitting and underfitting challenges in teacher models, employing techniques like cross-fitting and loss correction to enhance learning, ensuring robust and reliable student models [89].

The adaptability of SKD is augmented by methods like Spot Adaptive Knowledge Distillation (SAKD), which selectively identifies optimal distillation points during training, enhancing learning efficiency and student model performance by focusing on the most informative aspects of the teacher's knowledge [38]. By strategically selecting distillation points, SAKD ensures distilled knowledge remains relevant and interpretable. Integrating ensemble methods in SKD offers a unique perspective on optimizing interpretability in deep learning models. Unlike traditional feature selection methods, ensemble techniques provide a holistic view of model behavior, enhancing the interpretability of distilled knowledge [91].

5.6 Self-Learning and Generalization through PS-KD

Progressive Self-Knowledge Distillation (PS-KD) represents a significant advancement in developing efficient AI models for resource-constrained environments. PS-KD leverages self-learning, allowing models to iteratively refine their knowledge with minimal external supervision, enhancing their ability to generalize from limited data—ideal for applications with scarce data or privacy restrictions [79]. The core principle involves progressively enhancing the student model's knowledge by distilling information from itself, reducing dependency on large teacher models and fostering a deeper understanding of the data through repeated exposure and self-refinement.

By focusing on challenging examples during training, PS-KD improves generalization capabilities. Additionally, PS-KD addresses the challenge of maintaining performance while minimizing computational complexity. By utilizing a self-learning framework, PS-KD enables models to achieve high accuracy with reduced resource requirements, aligning with the goals of Tiny ML and Small-Scale AI for deploying efficient models on devices with limited computational power [79]. This approach is beneficial in scenarios where traditional knowledge distillation methods are impractical due to computational constraints.

The effectiveness of PS-KD is enhanced by its ability to adaptively adjust the learning process based on model performance. By continuously evaluating and refining predictions, the student model can dynamically enhance its understanding of tasks, leading to more robust and reliable outputs. This adaptability is crucial for deploying AI models in dynamic environments where conditions and data distributions may vary [79].

5.7 Modularity and Knowledge Retention across Domains

Modularity in machine learning models enhances knowledge retention across domains, facilitating systems that efficiently adapt to diverse tasks and environments. This approach decomposes complex models into smaller, manageable components or modules, each specializing in specific task aspects. Such specialization promotes effective knowledge transfer and retention, enabling each module to independently learn and refine its capabilities while maintaining overall system coherence [64].

A key benefit of modularity is its support for transfer learning across different domains. By leveraging pre-trained modules that encapsulate domain-specific knowledge, models can rapidly adapt to new tasks with minimal retraining, thus preserving knowledge acquired from previous tasks. This strategy enhances knowledge transfer efficiency and reduces the computational resources required for training, aligning with Tiny ML and Small-Scale AI objectives for deploying efficient models on resource-constrained devices [64].

Furthermore, the modular architecture allows for integrating specialized knowledge from multiple domains, facilitating the development of systems capable of handling complex, multi-faceted tasks. By combining modules with distinct expertise, models can leverage a broader knowledge base, improving decision-making capabilities and adaptability to varying conditions [64]. This approach is

AI-generated, for reference only.

especially beneficial in applications like customer service and medical diagnostics, where models must process diverse information and deliver accurate, contextually relevant responses.

Knowledge retention across domains is supported by techniques like multi-prompting, enabling models to explore various perspectives and generate comprehensive solutions to complex problems. By incorporating insights from different domains, models enhance reasoning capabilities and improve performance across a wide range of tasks [64]. This cross-domain knowledge retention is crucial for developing versatile and robust AI systems capable of operating effectively in dynamic and resource-constrained environments.

5.8 Leveraging Inter-Instance Similarities with STKD

The Similarity Transfer Knowledge Distillation (STKD) method advances the knowledge distillation process by leveraging inter-instance similarities to enhance student model generalization capabilities. STKD transcends traditional instance-based methods by utilizing virtual samples created through the mixup technique, enabling a richer knowledge transfer [50]. By employing virtual samples, STKD captures relationships between different data instances, facilitating a nuanced transfer of knowledge from teacher to student models. This method addresses limitations of conventional knowledge distillation techniques, which often focus solely on individual instances without considering broader inter-instance relationships.

STKD's ability to leverage inter-instance similarities is beneficial in scenarios where data diversity and variability are critical for model performance. By incorporating these similarities into the distillation process, STKD ensures student models gain a holistic understanding of the data, leading to improved accuracy and robustness. This approach aligns with broader goals of model simplification and efficiency, enabling the development of compact models that maintain high performance in resource-constrained environments [50].

The STKD method showcases the potential of utilizing inter-instance similarities to enhance knowledge distillation, paving the way for efficient AI models capable of performing effectively across diverse applications. By leveraging similarity correlations between instances and employing innovative techniques like the mixup method to create virtual samples, STKD not only improves student model accuracy but also outperforms traditional knowledge distillation approaches. This advancement suggests a promising direction for building compact neural networks that address computational challenges in domains such as classification and object detection tasks [92, 34, 38, 50]. As research in this area evolves, integrating inter-instance similarities is likely to play an increasingly important role in advancing knowledge distillation techniques.

6 Model Compression Techniques

Category	Feature	Method
Pruning Techniques	Feature Retention Attention-Based	PCA-KD[93] CD[94]
Quantization Methods	Model Optimization Strategies	MC[95]
Innovative Model Compression Strategies	Knowledge Transfer and Integration	NesyCD[96], STMSF[97], SKD[98], PS-KD[79], KD[13]
	Memory and Architecture Optimization	LMA[67], 3AS[8]
	Structured Learning Processes	RCO[99]
	Data Augmentation and Generation	MKD[100], IN[101]
Flexible Architecture Adjustments	Gradient Flow Enhancement	KD-HDNN[102]

Table 2: This table presents a comprehensive overview of various model compression techniques, categorizing them into pruning techniques, quantization methods, innovative model compression strategies, and flexible architecture adjustments. Each category is further detailed with specific methods and features, highlighting the diverse approaches employed to optimize neural networks for resource-constrained environments.

Model compression techniques are pivotal for optimizing neural networks in environments with limited resources, focusing on reducing model size and complexity while maintaining performance. Table 3 provides a detailed classification of model compression techniques, illustrating the range of methodologies applied to enhance neural network efficiency and performance in environments with limited computational resources. Pruning techniques are foundational in this domain, targeting

AI-generated, for reference only.

the elimination of redundant parameters to enhance efficiency. This section delves into pruning methodologies, their applications, and their impact on performance.

6.1 Pruning Techniques

Pruning techniques are essential for model compression, systematically removing redundant parameters to decrease complexity and resource usage. These methods optimize models for low-compute devices, potentially enhancing performance through improved efficiency. Techniques like weight quantization and network pruning significantly reduce model size and computational costs, often maintaining or even improving accuracy. Recent studies demonstrate the effectiveness of automated architecture slimming and the synergy of pruning with knowledge distillation to boost model transferability and performance [10, 8, 9, 90].

A notable approach uses Principal Component Analysis (PCA) to guide pruning, as evidenced by PCA-based knowledge distillation methods that create ultra-compact models, enhancing execution speed without sacrificing task quality [93]. The 3AS method automates architecture determination by maximizing filter variance, ensuring retention of informative features for efficient model performance [8].

Channel pruning techniques employing channel-wise attention mechanisms effectively reduce computational costs while preserving accuracy by selectively pruning channels based on their contribution to overall model performance [94]. Additionally, integrating knowledge distillation with pruning produces simpler student models that retain the performance of complex teacher models, highlighting the importance of distilling essential knowledge into compact forms [95].

Evaluating pruning techniques involves metrics assessing both accuracy and computational efficiency, ensuring pruned models meet real-world deployment demands [10]. This metrics focus aids in understanding pruning trade-offs and optimizing models for specific applications.

6.2 Quantization Methods

Quantization methods are crucial for model compression, reducing parameter precision to lower memory usage and computational demands while maintaining accuracy. These methods enhance efficiency and speed for machine learning models on resource-constrained devices through techniques like model compression, knowledge distillation, and adaptive optimization algorithms [7, 10, 11, 103].

Quantization typically involves converting floating-point weights and activations into lower bit-width representations, such as 8-bit integers, significantly reducing memory footprint and accelerating inference. This enhances operational efficiency, making deep learning models more suitable for deployment in limited-resource environments, while balancing model size, accuracy, and inference time across applications like image classification and natural language processing [104, 10, 9, 56].

Post-training quantization enables converting pre-trained models into quantized versions without additional training data, facilitating deployment on resource-constrained devices. Recent advancements like Self-Supervised Quantization-Aware Knowledge Distillation (SQAKD) improve this process by enabling effective quantization without labeled data, thus enhancing performance while simplifying training procedures [105, 10, 106, 9]. This method is particularly advantageous when retraining is impractical due to data scarcity or computational limitations.

Quantization-aware training (QAT) integrates quantization effects during model training, facilitating low-bit deep learning models that maintain competitive performance. QAT allows models to learn to compensate for reduced precision, resulting in better performance compared to post-training quantization, especially in accuracy-critical tasks [9, 105, 106, 11, 107].

Combining quantization with other model compression techniques, such as pruning, enhances neural network efficiency. This synergy is vital for deploying AI models in resource-constrained environments, where memory and computational resources are limited [95].

6.3 Hybrid Compression Techniques

Hybrid compression techniques combine multiple strategies, such as pruning and quantization, to achieve greater efficiency and performance. These techniques are essential for optimizing neural networks for deployment on resource-constrained devices, addressing both size and computational demands [104, 10, 9, 23].

Integrating pruning and quantization reduces model size by eliminating redundant parameters and decreasing the precision of remaining parameters, further lowering memory footprint and computational requirements [10]. A comprehensive evaluation framework has been developed to assess hybrid methods across various architectures, providing a holistic view of performance gains achieved through these techniques [10].

The synergy between pruning and quantization enables the creation of models that are smaller, faster, and capable of maintaining accuracy. This is particularly crucial for edge computing and Internet of Things (IoT) applications, where computational resources are limited. Innovative approaches, such as lightweight DNN design through knowledge distillation and early halting techniques, enhance resource utilization and training speed, ensuring adequate accuracy while minimizing resource consumption. Model compression techniques like quantization and pruning are vital for reducing model size and inference delays, supporting operational constraints in edge computing and IoT scenarios [12, 7].

6.4 Innovative Model Compression Strategies

Innovative model compression strategies are critical for enhancing neural network efficiency, facilitating deployment on resource-constrained devices without sacrificing performance. These strategies utilize advanced techniques such as network pruning, knowledge distillation, and architecture slimming to achieve substantial reductions in model size and computational demands while maintaining or improving accuracy [8, 23].

One notable strategy is the Light Multi-Segment Activation (LMA) model, which enhances accuracy while reducing memory costs compared to other multi-segment activations [67]. Route Constrained Optimization (RCO) improves small student networks' performance by leveraging a structured learning sequence derived from the teacher's training trajectory, optimizing knowledge transfer [99].

The Neural-Symbolic Collaborative Distillation (NesyCD) method integrates symbolic knowledge bases with neural networks, enhancing knowledge transfer and performance in complex reasoning tasks [96]. Mixup, a data augmentation technique, effectively enhances student model performance in knowledge distillation without excessive smoothness [100].

The novel architecture slimming method automates the determination of compressed architectures, enhancing model compression efficiency without extensive human input [8]. Peer-to-peer learning facilitates simultaneous training of teacher and student networks, fostering knowledge transfer and collaborative learning among students [97]. Few-shot learning techniques utilizing optimized pseudo training examples enable efficient model compression in data-scarce scenarios [101].

Stage-wise Knowledge Distillation (SKD) presents opportunities for future research by integrating with other compression techniques and applying it to computer vision tasks such as object detection and pose estimation [98]. Progressive Self-Knowledge Distillation (PS-KD) enhances generalization performance across tasks, iteratively refining knowledge within models for continuous improvement [79].

Lastly, approaches that reduce model size and computational requirements while maintaining accuracy, such as automatic scoring systems, underscore the potential for deploying efficient models on less powerful hardware [13]. These innovations highlight the importance of model compression in the practical application of AI technologies across diverse and resource-limited environments.

6.5 Flexible Architecture Adjustments

Flexible architecture adjustments are crucial for effective model compression, enabling efficient deployment of neural networks on resource-constrained devices without sacrificing performance. One innovative approach is the use of Highway Deep Neural Networks (HDNNs), which incorporate gate functions to facilitate smoother gradient flow and efficient learning [102]. These mechanisms maintain model performance while allowing significant reductions in complexity and size.

Evaluating models based on performance metrics across multiple trials after applying various compression techniques ensures the identification of the most effective strategies, optimizing model efficiency and accuracy [3]. This systematic testing of configurations allows researchers to fine-tune architectures for a balance between computational efficiency and predictive performance.

Such flexible adjustments are particularly beneficial in scenarios requiring models to adapt to diverse operational environments with varying resource constraints. By leveraging architectural innovations like HDNNs and rigorous evaluation methodologies, it is possible to develop compact, efficient, and robust models adaptable to a wide range of applications. Techniques such as knowledge distillation and model compression address limitations in processing power and storage capacity, enabling the development of smaller models that maintain performance levels comparable to larger counterparts. Optimizing inference and training processes through methods like elastic heterogeneous computing and adaptive optimization algorithms enhances AI application efficiency, even on devices with restricted resources [7, 11, 108].

Feature	Pruning Techniques	Quantization Methods	Hybrid Compression Techniques
Optimization Focus	Parameter Elimination	Precision Reduction	Combined Strategies
Efficiency Enhancement	Improved Efficiency	Lower Memory Usage	Greater Efficiency
Deployment Suitability	Low-compute Devices	Resource-constrained Devices	Edge Computing

Table 3: This table presents a comparative analysis of various model compression techniques, including pruning, quantization, and hybrid strategies. Each method is evaluated based on its optimization focus, efficiency enhancement, and suitability for deployment in resource-constrained environments. The comparison highlights the strengths of each technique in improving neural network performance while reducing computational demands.

7 Neural Network Simplification

7.1 Neural Network Simplification

Neural network simplification is essential for model compression, focusing on reducing architecture size and complexity to facilitate deployment on resource-constrained devices. Techniques such as pruning, quantization, and knowledge distillation are pivotal in lowering computational demands and model sizes. Recent research underscores the importance of data augmentation in optimizing these compression methods, suggesting that tailored augmentation strategies can enhance performance across different model sizes, thereby addressing the need for efficient models in a connected world [104, 10].

A central strategy in this field is architecture search, which seeks to identify efficient model architectures that achieve high performance with minimal resources. Neural Architecture Search (NAS) automates this process, enabling the discovery of optimized structures tailored to specific tasks and environments [8]. By systematically exploring a range of architectural configurations, NAS identifies compact models suitable for devices with limited capabilities.

Parameter reduction is equally critical, employing techniques like pruning and quantization to decrease model size and computational demands. Pruning removes redundant parameters, while quantization reduces the precision of weights and activations, both contributing to enhanced model efficiency [95]. These methods are vital for deploying AI models in environments with strict memory and power constraints.

The integration of architecture search and parameter reduction is exemplified by hybrid compression techniques, which combine multiple strategies to enhance efficiency. By merging methods such as pruning and quantization, these techniques offer comprehensive solutions for reducing model complexity while maintaining accuracy [10]. This approach aligns with the goal of deploying AI on resource-limited devices, enabling the development of models that are both compact and powerful.

Furthermore, using variance as a metric for information retention in model parameters helps identify critical components of a neural network. Analyzing parameter variance allows researchers to prioritize essential information retention, ensuring that simplified models maintain their predictive capabilities [8]. This methodology supports the efficient slimming of neural architectures, enhancing their deployment suitability in resource-limited settings.

8 Applications and Case Studies

Efficient AI models are pivotal across domains such as graph processing, language understanding, image classification, and medical diagnostics. This section explores practical applications of these models, focusing on methodologies like knowledge distillation and model compression that optimize performance in resource-limited environments. The subsequent subsections delve into specific applications, beginning with graph applications and memory access prediction, where efficient AI models notably enhance operational capabilities.

8.1 Graph Applications and Memory Access Prediction

In graph applications and memory access prediction, efficient AI models are crucial, especially in resource-constrained settings. Challenges in deploying machine learning models for graph processing involve managing large datasets and optimizing computational resources. Techniques such as knowledge distillation and model compression reduce model complexity while preserving accuracy [50]. Simplified neural networks, achieved through pruning and quantization, enable efficient graph data processing, retaining essential features for accurate analysis [95].

Memory access prediction benefits from model efficiency, optimizing memory usage and reducing latency, critical for environments with limited resources [3]. Advanced techniques like Similarity Transfer Knowledge Distillation (STKD) enhance model capabilities by leveraging inter-instance similarities, improving generalization and accuracy [50]. Strategies such as "prune, then distill" facilitate real-time applications by transferring knowledge from pruned teacher models to smaller student models [34, 11, 109, 90].

8.2 Language Processing and Understanding

Efficient AI models advance natural language processing (NLP) in resource-limited scenarios. Neural-symbolic collaborative distillation (NesyCD) combines symbolic knowledge with neural networks, enhancing performance on complex reasoning tasks [96]. Experiments with datasets such as BBH and GSM8K demonstrate NesyCD's efficacy over standard methods like CoT distillation [96].

The integration of symbolic reasoning with neural networks enables nuanced language models for applications like automated customer service. Interpretable knowledge distillation improves goal-oriented dialogues, ensuring secure self-hosting and high-quality service [64, 109, 11, 49]. Techniques like NesyCD transfer capabilities from large proprietary models to smaller, open-source counterparts, enhancing performance and reducing computational demands in constrained environments [35, 20].

8.3 Image Classification and Object Detection

In computer vision tasks like image classification and object detection, efficient AI models are vital for real-time performance on resource-constrained devices. Model compression techniques such as knowledge distillation, pruning, and quantization reduce computational burdens while maintaining accuracy, facilitating compact object detection networks [110, 2, 9, 56, 111].

Knowledge distillation transfers knowledge from large models to smaller ones, ensuring high accuracy under limited computational resources. PCA-based knowledge distillation creates ultra-compact models excelling in tasks like stylization [93]. Ensemble and multi-teacher strategies enhance student models in object detection by providing diverse training signals, improving generalization and accuracy [47]. Pruning and quantization reduce model complexity, resulting in compact yet powerful models [95].

8.4 Medical Imaging and Diagnostic Applications

Advanced AI models like the Hybrid Data-Efficient Knowledge Distillation Network (HDKD) and interpretable embedding techniques transform medical imaging and diagnostics by improving accuracy and speed. These models leverage knowledge from convolutional neural networks to enhance vision transformers, especially in data-limited scenarios, providing interpretable insights into embedding procedures [70, 112].

Efficient processing of large datasets is essential in medical imaging for accurate diagnoses and treatment planning. HDKD enhances model performance on limited data by distilling knowledge from CNNs, improving ViTs' generalization while minimizing computational overhead [70, 112]. Quantization methods further enhance AI model efficiency by reducing parameter precision, allowing faster processing times essential for real-time diagnostics [95]. Pruning techniques reduce model size by eliminating redundant parameters, ensuring critical features are retained for accurate diagnostic outcomes [95].

8.5 Educational Technology and Large Language Models

Model compression and knowledge distillation techniques are crucial for developing large language models (LLMs) in educational technology. These techniques enable efficient LLMs to operate effectively on resource-constrained devices, facilitating their use in diverse educational settings [13]. In educational technology, LLMs' ability to process and generate natural language text accurately is vital for applications like automated tutoring and personalized learning experiences. Knowledge distillation allows smaller models to learn from larger, complex teacher models, maintaining high performance while being optimized for deployment on devices with limited computational capabilities [13].

Model compression techniques such as pruning and quantization enhance LLM efficiency by reducing size and computational demands, ensuring accessibility and effectiveness across various settings [95]. Innovative loss functions and ensemble strategies in knowledge distillation enhance the training process for student models by providing diverse training signals, leveraging complementary knowledge from multiple teacher models [43, 29, 42, 45, 46].

8.6 General Applications Across Diverse Datasets

Efficient AI models' development and deployment across diverse datasets are critical for advancing machine learning applications on resource-constrained devices. The KroneckerBERT framework exemplifies innovative model compression techniques, achieving a remarkable compression factor of 19 on the BERTBASE model, surpassing existing state-of-the-art methods [113]. This significant reduction in model size demonstrates the framework's capability to maintain high performance while being optimized for deployment in limited-resource environments.

The application of knowledge distillation and model compression techniques across various datasets highlights the versatility of these methods. Advanced distillation strategies enhance model performance metrics, such as accuracy, while significantly reducing model size [3]. These improvements are crucial for deploying AI models in diverse applications, ranging from natural language processing to image classification. By employing advanced model compression techniques, such as Kronecker decomposition and knowledge distillation, researchers can create deep learning models that are smaller, more energy-efficient, and maintain high performance across a range of applications [10, 5, 113].

9 Future Directions

9.1 Integration with Emerging Technologies

Integrating emerging technologies with model simplification techniques offers significant potential for enhancing AI model efficiency across various domains. Research should focus on optimizing distillation processes and exploring diverse datasets with augmentation strategies to enhance knowledge distillation fidelity [33]. Curriculum learning approaches can mitigate the capacity curse by providing structured pathways that optimize model performance [6]. Self-supervised learning advancements offer integration opportunities with knowledge distillation, potentially improving efficiency and enabling cross-modal knowledge transfer [16]. Refining sample selection criteria in selective knowledge distillation could enhance adaptability and performance in neural machine translation architectures [27]. Optimizing data sampling strategies to ensure higher sample similarity in mini-batches and applying methods like BAKE to tasks beyond classification, such as segmentation and detection, are valuable research directions [28]. Expanding Random Intermediate Layer Knowledge Distillation (RAIL-KD) to larger models and a broader range of natural language understanding tasks could further extend model simplification capabilities [30]. Integrating advanced data augmentation strategies with knowledge distillation methodologies can significantly enhance model simplification techniques, improving AI model efficiency and performance on resource-limited edge devices by allowing tailored data augmentation policies that enhance interpretability through class-similarity information transfer [104, 63].

9.2 Enhanced Data Utilization and Augmentation

Enhancing data utilization and augmentation is vital for maximizing AI model performance in resource-constrained environments. Adaptive optimization algorithms, parallel computing, and mixed precision training can accelerate convergence and reduce memory usage. Combining model compression methods like quantization, pruning, and knowledge distillation with tailored data augmentation strategies optimizes model size and inference speed while preserving accuracy. Understanding the relationship between model architecture and data augmentation enables developing effective policies that leverage various model sizes to improve performance in challenging computational contexts [104, 7]. Future research could further enhance these processes, improving knowledge distillation and model compression techniques. Refining Kronecker decomposition, as demonstrated in the KroneckerBERT framework, offers substantial compression while maintaining performance, and extending this technique to other Transformer-based models could yield compact, powerful models for limited-resource devices [113]. Optimizing proposal distributions beyond knowledge distillation could enhance large-scale distillation processes, broadening machine learning applications [41]. Exploring tighter bounds on mutual information and mitigating bias transfer during distillation is essential for improving fidelity and fairness in student models [114]. Integrating environmental labels to guide model learning may address distribution shifts, enhancing the robustness of knowledge distillation methods [54]. Developing standardized formats for deep neural network models and exploring additional metadata collection methods could enhance reconstruction accuracy in data-free knowledge distillation [115]. Employing features from within stages and investigating further loss functions could refine the distillation framework, improving performance across diverse applications [34]. Combining different distillation techniques and evaluating their effectiveness across various use cases, such as retail banking, could optimize model performance in practical applications [72]. Novel distillation techniques and improved interpretability of distilled models are critical for addressing real-time deployment challenges in practical settings [111]. Addressing hard sampling issues and exploring combinations of proposed methods with other knowledge distillation techniques could enhance performance, particularly with limited data availability [14]. Optimizing frameworks by integrating additional techniques and testing their applicability across different network architectures and tasks could lead to more efficient and adaptable models [116]. These research directions underscore the potential for significant advancements in data utilization and augmentation, driving the development of more efficient and robust AI models. Leveraging advancements in knowledge distillation and data augmentation techniques can improve AI technology implementation across various applications, particularly in resource-constrained environments. Innovations such as tailored data generation for teacher-student models and optimized loss functions for document ranking facilitate more efficient model training and inference, enhancing AI solutions' accessibility and effectiveness in challenging settings [117, 118, 11, 109].

10 Conclusion

Exploring knowledge distillation and model compression techniques underscores their essential contribution to reducing machine learning model complexity and resource demands, enabling their deployment on devices with limited resources. The DFKD-T3 framework exemplifies significant progress in distillation performance across NLP tasks, outperforming existing methods and demonstrating strong scalability, which underscores the effectiveness of data-free approaches in improving knowledge transfer. The TMKD method illustrates the benefits of multi-teacher strategies, achieving results akin to original teacher models while enhancing accuracy and inference speed, crucial for optimizing large language models in educational technology and expanding access to advanced capabilities. Symbolic knowledge distillation further enhances the transparency and functionality of compact AI models, aligning with the objective of model simplification and promoting efficient AI solutions across various sectors. Innovations, such as those by Chen et al., reveal substantial improvements in the accuracy-speed trade-offs, highlighting the transformative potential of knowledge distillation techniques for future AI deployments. The EarlyLight approach exemplifies how training

AI-generated, for reference only.

time and resource consumption can be minimized while maintaining high accuracy, vital for real-time applications. These advancements in knowledge distillation and model compression are pivotal in shaping AI deployment's future, enabling substantial model compression while retaining competitive performance. By addressing the complexity and resource challenges of machine learning models, these techniques hold significant promise for advancing more efficient and accessible AI solutions across diverse domains.

References

[1] Yousef Alsenani, Rahul Mishra, Khaled R. Ahmed, and Atta Ur Rahman. Fedsikd: Clients similarity and knowledge distillation: Addressing non-i.i.d. and constraints in federated learning, 2024.

[2] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017.

[3] Het Shah, Avishree Khare, Neelay Shah, and Khizir Siddiqui. Kd-lib: A pytorch library for knowledge distillation, pruning and quantization, 2020.

[4] Suhas Lohit and Michael Jones. Model compression using optimal transport, 2020.

[5] George Papamakarios. Distilling model knowledge, 2015.

[6] Jun Wang, Eleftheria Briakou, Hamid Dadkhahi, Rishabh Agarwal, Colin Cherry, and Trevor Cohn. Don't throw away data: Better sequence knowledge distillation, 2024.

[7] Taiyuan Mei, Yun Zi, Xiaohan Cheng, Zijun Gao, Qi Wang, and Haowei Yang. Efficiency optimization of large-scale language models based on deep learning in natural language processing tasks, 2024.

[8] Dongqi Wang, Shengyu Zhang, Zhipeng Di, Xin Lin, Weihua Zhou, and Fei Wu. A novel architecture slimming method for network pruning and knowledge distillation, 2022.

[9] Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhijun Tu, Kai Han, Hailin Hu, and Dacheng Tao. A survey on transformer compression, 2024.

[10] Aayush Saxena, Arit Kumar Bishwas, Ayush Ashok Mishra, and Ryan Armstrong. Comprehensive study on performance evaluation and optimization of model compression: Bridging traditional deep learning and large language models, 2024.

[11] Xubo Qin, Xiyuan Liu, Xiongfeng Zheng, Jie Liu, and Yutao Zhu. An empirical study of uniform-architecture knowledge distillation in document ranking, 2023.

[12] Rahul Mishra and Hari Prabhat Gupta. Designing and training of lightweight neural networks on edge devices using early halting in knowledge distillation, 2022.

[13] Ehsan Latif, Luyang Fang, Ping Ma, and Xiaoming Zhai. Knowledge distillation of llm for automatic scoring of science education assessments, 2024.

[14] Hideki Oki, Motoshi Abe, Junichi Miyao, and Takio Kurita. Triplet loss for knowledge distillation, 2020.

[15] Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. Model compression with two-stage multi-teacher knowledge distillation for web question answering system, 2019.

[16] Chuanguang Yang, Xinqiang Yu, Zhulin An, and Yongjun Xu. Categories of response-based, feature-based, and relation-based knowledge distillation, 2023.

[17] Kamal Acharya, Alvaro Velasquez, and Houbing Herbert Song. A survey on symbolic knowledge distillation of large language models, 2024.

[18] Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models, 2023.

[19] Jiacheng Liu, Peng Tang, Wenfeng Wang, Yuhang Ren, Xiaofeng Hou, Pheng-Ann Heng, Minyi Guo, and Chao Li. A survey on inference optimization techniques for mixture of experts models, 2025.

[20] Minillm: Knowledge distillation.

AI-generated, for reference only.

[21] Ivan Kobyzev, Aref Jafari, Mehdi Rezagholizadeh, Tianda Li, Alan Do-Omri, Peng Lu, Pascal Poupart, and Ali Ghodsi. Do we need label regularization to fine-tune pre-trained language models?, 2023.

[22] Tianda Li, Yassir El Mesbahi, Ivan Kobyzev, Ahmad Rashid, Atif Mahmud, Nithin Anchuri, Habib Hajimolahoseini, Yang Liu, and Mehdi Rezagholizadeh. A short study on compressing decoder-based language models, 2021.

[23] Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models, 2024.

[24] Apoorva Verma, Pranjal Gulati, and Sarthak Gupta. [re] distilling knowledge via knowledge review, 2022.

[25] Neelesh Gupta, Pengmiao Zhang, Rajgopal Kannan, and Viktor Prasanna. Packd: Pattern-clustered knowledge distillation for compressing memory access prediction models, 2024.

[26] Yuxuan Jiang, Chen Feng, Fan Zhang, and David Bull. Mtkd: Multi-teacher knowledge distillation for image super-resolution, 2024.

[27] Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. Selective knowledge distillation for neural machine translation, 2021.

[28] Yixiao Ge, Xiao Zhang, Ching Lam Choi, Ka Chun Cheung, Peipei Zhao, Feng Zhu, Xiaogang Wang, Rui Zhao, and Hongsheng Li. Self-distillation with batch knowledge ensembling improves imagenet classification, 2021.

[29] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier, 2022.

[30] Md Akmal Haidar, Nithin Anchuri, Mehdi Rezagholizadeh, Abbas Ghaddar, Philippe Langlais, and Pascal Poupart. Rail-kd: Random intermediate layer mapping for knowledge distillation, 2021.

[31] Lirong Wu, Yunfan Liu, Haitao Lin, Yufei Huang, and Stan Z. Li. Teach harder, learn poorer: Rethinking hard sample distillation for gnn-to-mlp knowledge distillation, 2024.

[32] Prashant Bhat, Elahe Arani, and Bahram Zonooz. Distill on the go: Online knowledge distillation in self-supervised learning, 2021.

[33] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work?, 2021.

[34] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5008–5017, 2021.

[35] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models, 2024.

[36] Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020.

[37] Yiqin Yu, Xu Min, Shiwan Zhao, Jing Mei, Fei Wang, Dongsheng Li, Kenney Ng, and Shaochun Li. Dynamic knowledge distillation for black-box hypothesis transfer learning, 2020.

[38] Jie Song, Ying Chen, Jingwen Ye, and Mingli Song. Spot-adaptive knowledge distillation, 2022.

[39] Norah Alballa and Marco Canini. Practical insights into knowledge distillation for pre-trained models, 2024.

AI-generated, for reference only.

[40] Songling Zhu, Ronghua Shang, Bo Yuan, Weitong Zhang, Yangyang Li, and Licheng Jiao. Dynamickd: An effective knowledge distillation via dynamic entropy correction-based distillation for gap optimizing, 2023.

[41] Minghan Li, Tanli Zuo, Ruicheng Li, Martha White, and Weishi Zheng. Accelerating large scale knowledge distillation via dynamic importance sampling, 2018.

[42] Huayu Li, Xiwen Chen, Gregory Ditzler, Janet Roveda, and Ao Li. Knowledge distillation under ideal joint classifier assumption, 2024.

[43] Umar Asif, Jianbin Tang, and Stefan Harrer. Ensemble knowledge distillation for learning improved and efficient networks, 2020.

[44] Chaomin Shen, Yaomin Huang, Haokun Zhu, Jinsong Fan, and Guixu Zhang. Student-oriented teacher knowledge refinement for knowledge distillation, 2024.

[45] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Unified and effective ensemble knowledge distillation, 2022.

[46] Zheng Li, Ying Huang, Defang Chen, Tianren Luo, Ning Cai, and Zhigeng Pan. Online knowledge distillation via multi-branch diversity enhancement, 2020.

[47] Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. Multi-teacher knowledge distillation for text image machine translation, 2023.

[48] Yushuo Guan, Pengyu Zhao, Bingxuan Wang, Yuanxing Zhang, Cong Yao, Kaigui Bian, and Jian Tang. Differentiable feature aggregation search for knowledge distillation, 2020.

[49] Mohammed Al-Maamari, Mehdi Ben Amor, and Michael Granitzer. Mixture of modular experts: Distilling knowledge from a multilingual teacher into specialized modular language models, 2024.

[50] Haoran Zhao, Kun Gong, Xin Sun, Junyu Dong, and Hui Yu. Similarity transfer for knowledge distillation, 2021.

[51] Xueqing Deng, Dawei Sun, Shawn Newsam, and Peng Wang. Distpro: Searching a fast knowledge distillation process via meta optimization, 2022.

[52] Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge diffusion for distillation, 2023.

[53] Yaomin Huang, Zaomin Yan, Chaomin Shen, Faming Fang, and Guixu Zhang. Harmonizing knowledge transfer in neural network with unified distillation, 2024.

[54] Songming Zhang, Ziyu Lyu, and Xiaofeng Chen. Revisiting knowledge distillation under distribution shift, 2024.

[55] John Violos, Symeon Papadopoulos, and Ioannis Kompatsiaris. Towards optimal trade-offs in knowledge distillation for cnns and vision transformers at the edge, 2024.

[56] Maniratnam Mandal and Imran Khan. Analyzing compression techniques for computer vision, 2023.

[57] Harshita Diddee, Sandipan Dandapat, Monojit Choudhury, Tanuja Ganu, and Kalika Bali. Too brittle to touch: Comparing the stability of quantization and distillation towards developing lightweight low-resource mt models, 2022.

[58] Tianda Li, Ahmad Rashid, Aref Jafari, Pranav Sharma, Ali Ghodsi, and Mehdi Rezagholizadeh. How to select one among all? an extensive empirical study towards the robustness of knowledge distillation in natural language understanding, 2021.

[59] Zheng Li, Yuxuan Li, Penghai Zhao, Renjie Song, Xiang Li, and Jian Yang. Is synthetic data from diffusion models ready for knowledge distillation?, 2023.

[60] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation, 2019.

[61] Deepan Das, Haley Massa, Abhimanyu Kulkarni, and Theodoros Rekatsinas. An empirical analysis of the impact of data augmentation on knowledge distillation, 2020.

[62] Yiqing Shen, Liwu Xu, Yuzhe Yang, Yaqian Li, and Yandong Guo. Mixed sample augmentation for online distillation, 2023.

[63] Hyeongrok Han, Siwon Kim, Hyun-Soo Choi, and Sungroh Yoon. On the impact of knowledge distillation for model interpretability, 2023.

[64] Tong Wang, K. Sudhir, and Dat Hong. Using advanced llms to enhance smaller llms: An interpretable knowledge distillation approach, 2024.

[65] Zichang Liu, Qingyun Liu, Yuening Li, Liang Liu, Anshumali Shrivastava, Shuchao Bi, Lichan Hong, Ed H. Chi, and Zhe Zhao. Wisdom of committee: Distilling from foundation model to specialized application model, 2024.

[66] Ka Man Lo, Yiming Liang, Wenyu Du, Yuantao Fan, Zili Wang, Wenhao Huang, Lei Ma, and Jie Fu. m2mkd: Module-to-module knowledge distillation for modular transformers, 2024.

[67] Zhenhui Xu, Guolin Ke, Jia Zhang, Jiang Bian, and Tie-Yan Liu. Light multi-segment activation for model compression, 2019.

[68] Ke Zhang, Hanbo Ying, Hong-Ning Dai, Lin Li, Yuangyuang Peng, Keyi Guo, and Hongfang Yu. Compacting deep neural networks for internet of things: Methods and applications, 2021.

[69] Zao Zhang, Huaming Chen, Pei Ning, Nan Yang, and Dong Yuan. Cakd: A correlation-aware knowledge distillation framework based on decoupling kullback-leibler divergence, 2024.

[70] Omar S. EL-Assiouti, Ghada Hamed, Dina Khattab, and Hala M. Ebied. Hdkd: Hybrid data-efficient knowledge distillation network for medical image classification, 2024.

[71] Xucong Wang, Pengchao Han, and Lei Guo. Lightweight self-knowledge distillation with multi-source information fusion, 2023.

[72] Maxime Biehler, Mohamed Guermazi, and Célim Starck. Using knowledge distillation to improve interpretable models in a retail banking context, 2022.

[73] Kuluhan Binici, Weiming Wu, and Tulika Mitra. Generalizing teacher networks for effective knowledge distillation across student architectures, 2025.

[74] Ye Lin, Yanyang Li, Ziyang Wang, Bei Li, Quan Du, Tong Xiao, and Jingbo Zhu. Weight distillation: Transferring the knowledge in neural network parameters, 2021.

[75] Junjie Zhou, Ke Zhu, and Jianxin Wu. All you need in knowledge distillation is a tailored coordinate system, 2025.

[76] Gaurav Patel, Konda Reddy Mopuri, and Qiang Qiu. Learning to retain while acquiring: Combating distribution-shift in adversarial data-free knowledge distillation, 2023.

[77] Jihao Liu, Boxiao Liu, Hongsheng Li, and Yu Liu. Meta knowledge distillation, 2022.

[78] Vibhas Vats and David Crandall. Controlling the quality of distillation in response-based network compression, 2021.

[79] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation with progressive refinement of targets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6567–6576, 2021.

[80] Chenglong Wang, Yi Lu, Yongyu Mu, Yimin Hu, Tong Xiao, and Jingbo Zhu. Improved knowledge distillation for pre-trained language models via knowledge selection, 2023.

[81] Dongkyu Lee, Zhiliang Tian, Yingxiu Zhao, Ka Chun Cheung, and Nevin L. Zhang. Hard gate knowledge distillation – leverage calibration for robust and reliable language model, 2022.

[82] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation, 2021.

AI-generated, for reference only.

[83] Jiao Xie, Shaohui Lin, Yichen Zhang, and Linkai Luo. Training convolutional neural networks with cheap convolutions and online distillation, 2019.

[84] Jihyeon Seo, Kyusam Oh, Chanho Min, Yongkeun Yun, and Sungwoo Cho. Deep collective knowledge distillation, 2023.

[85] Zenglin Shi, Pei Liu, Tong Su, Yunpeng Wu, Kuien Liu, Yu Song, and Meng Wang. Densely distilling cumulative knowledge for continual learning, 2024.

[86] Jun Rao, Xv Meng, Liang Ding, Shuhan Qi, and Dacheng Tao. Parameter-efficient and student-friendly knowledge distillation, 2022.

[87] Chenglin Li, Qianglong Chen, Liangyue Li, Caiyu Wang, Yicheng Li, Zulong Chen, and Yin Zhang. Mixed distillation helps smaller language model better reasoning, 2024.

[88] He Liu, Yikai Wang, Huaping Liu, Fuchun Sun, and Anbang Yao. Small scale data-free knowledge distillation, 2024.

[89] Tri Dao, Govinda M Kamath, Vasilis Syrgkanis, and Lester Mackey. Knowledge distillation as semiparametric inference, 2021.

[90] Jinhyuk Park and Albert No. Prune your model before distill it, 2022.

[91] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

[92] Sajjad Abbasi, Mohsen Hajabdollahi, Nader Karimi, and Shadrokh Samavi. Modeling teacher-student techniques in deep neural networks for knowledge distillation, 2019.

[93] Tai-Yin Chiu and Danna Gurari. Pca-based knowledge distillation towards lightweight and content-style balanced photorealistic style transfer models, 2022.

[94] Zaida Zhou, Chaoran Zhuge, Xinwei Guan, and Wen Liu. Channel distillation: Channel-wise attention for knowledge distillation, 2020.

[95] Timotheos Souroulla, Alberto Hata, Ahmad Terra, Özer Özkahraman, and Rafia Inam. Model compression for resource-constrained mobile robots, 2022.

[96] Huanxuan Liao, Shizhu He, Yao Xu, Yuanzhe Zhang, Kang Liu, and Jun Zhao. Neural-symbolic collaborative distillation: Advancing small language models for complex reasoning tasks, 2025.

[97] Usma Niyaz and Deepti R. Bathula. Augmenting knowledge distillation with peer-to-peer mutual learning for model compression, 2021.

[98] Akshay Kulkarni, Navid Panchi, Sharath Chandra Raparthy, and Shital Chiddarwar. Data efficient stagewise knowledge distillation, 2020.

[99] Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. Knowledge distillation via route constrained optimization, 2019.

[100] Hongjun Choi, Eun Som Jeon, Ankita Shukla, and Pavan Turaga. Understanding the role of mixup in knowledge distillation: An empirical study, 2022.

[101] Akisato Kimura, Zoubin Ghahramani, Koh Takeuchi, Tomoharu Iwata, and Naonori Ueda. Few-shot learning of neural networks from scratch by pseudo example optimization, 2018.

[102] Liang Lu, Michelle Guo, and Steve Renals. Knowledge distillation for small-footprint highway networks, 2016.

[103] Junmo Kang, Wei Xu, and Alan Ritter. Distill or annotate? cost-efficient fine-tuning of compact models, 2023.

[104] Muzhou Yu, Linfeng Zhang, and Kaisheng Ma. Revisiting data augmentation in model compression: An empirical and comprehensive study, 2023.

AI-generated, for reference only.

[105] Jangho Kim, Yash Bhalgat, Jinwon Lee, Chirag Patel, and Nojun Kwak. Qkd: Quantization-aware knowledge distillation, 2019.

[106] Kaiqi Zhao and Ming Zhao. Self-supervised quantization-aware knowledge distillation, 2024.

[107] Xinpeng Wang, Leonie Weissweiler, Hinrich Schütze, and Barbara Plank. How to distill your bert: An empirical study on the impact of weight initialisation and distillation objectives, 2023.

[108] Ji Liu, Daxiang Dong, Xi Wang, An Qin, Xingjian Li, Patrick Valduriez, Dejing Dou, and Dianhai Yu. Large-scale knowledge distillation with elastic heterogeneous computing resources, 2022.

[109] Jingxuan Wei, Linzhuang Sun, Yichong Leng, Xu Tan, Bihui Yu, and Ruifeng Guo. Sentence-level or token-level? a comprehensive study on knowledge distillation, 2024.

[110] Cheng Han, Qifan Wang, Sohail A. Dianat, Majid Rabbani, Raghuveer M. Rao, Yi Fang, Qiang Guan, Lifu Huang, and Dongfang Liu. Amd: Automatic multi-step distillation of large-scale vision models, 2024.

[111] Gousia Habib, Tausifa jan Saleem, Sheikh Musa Kaleem, Tufail Rouf, and Brejesh Lall. A comprehensive review of knowledge distillation in computer vision, 2024.

[112] Seunghyun Lee and Byung Cheol Song. Interpretable embedding procedure knowledge transfer via stacked principal component analysis and graph neural network, 2021.

[113] Marzieh S. Tahaei, Ella Charlaix, Vahid Partovi Nia, Ali Ghodsi, and Mehdi Rezagholizadeh. Kroneckerbert: Learning kronecker decomposition for pre-trained language models via knowledge distillation, 2021.

[114] Aman Shrivastava, Yanjun Qi, and Vicente Ordonez. Estimating and maximizing mutual information for knowledge distillation, 2023.

[115] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks, 2017.

[116] Junjie Liu, Dongchao Wen, Hongxing Gao, Wei Tao, Tse-Wei Chen, Kinya Osa, and Masami Kato. Knowledge representing: Efficient, sparse representation of prior knowledge for knowledge distillation, 2019.

[117] Ziqi Wang, Chi Han, Wenxuan Bao, and Heng Ji. Understanding the effect of data augmentation on knowledge distillation, 2023.

[118] Jie Fu, Xue Geng, Zhijian Duan, Bohan Zhuang, Xingdi Yuan, Adam Trischler, Jie Lin, Chris Pal, and Hao Dong. Role-wise data augmentation for knowledge distillation, 2020.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.