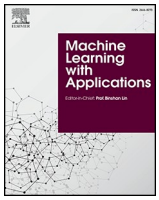
Machine Learning with Applications 18 (2024) 100605

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa



A survey on knowledge distillation: Recent advancements

Amir Moslemi [a,b,*], Anna Briskina [a], Zubeka Dang [a], Jason Li [a]

[a] School of Software Design & Data Science, Seneca Polytechnic, Toronto, Ontario, Canada
[b] Department of Physics, Toronto Metropolitan University, Toronto, Ontario, Canada

ARTICLE INFO

Keywords:
Deep learning
Knowledge distillation
Model compression
Self-distillation
Adversarial distillation

ABSTRACT

Deep learning has achieved notable success across academia, medicine, and industry. Its ability to identify complex patterns in large-scale data and to manage millions of parameters has made it highly advantageous. However, deploying deep learning models presents a significant challenge due to their high computational demands. Knowledge distillation (KD) has emerged as a key technique for model compression and efficient knowledge transfer, enabling the deployment of deep learning models on resource-limited devices without compromising performance. This survey examines recent advancements in KD, highlighting key innovations in architectures, training paradigms, and application domains. We categorize contemporary KD methods into traditional approaches, such as response-based, feature-based, and relation-based knowledge distillation, and novel advanced paradigms, including self-distillation, cross-modal distillation, and adversarial distillation strategies. Additionally, we discuss emerging challenges, particularly in the context of distillation under limited data scenarios, privacy-preserving KD, and the interplay with other model compression techniques like quantization. Our survey also explores applications across computer vision, natural language processing, and multimodal tasks, where KD has driven performance improvements and enhanced model compression. This review aims to provide researchers and practitioners with a comprehensive understanding of the state-of-the-art in knowledge distillation, bridging foundational concepts with the latest methodologies and practical implications.

1. Introduction

Knowledge distillation (KD), a technique for model compression and knowledge transfer, has gained significant attention in the field of machine learning since its formal introduction by Hinton et al. (2015). This approach builds upon the earlier work of Caruana et al. (2006), who demonstrated that the collective knowledge of an ensemble of models could be condensed into a single, smaller model. This groundbreaking idea laid the foundation for addressing the challenge of deploying deep models on resource-constrained devices such as mobile phones and embedded systems, which often have limited memory and processing power. Hinton et al. (2015) expanded on this concept, showing that information from a large, complex model (the teacher) could be effectively transferred to a smaller, more compact model (the student) through a process they termed "knowledge distillation" (shown as Fig. 1). This technique has since become a cornerstone in the ongoing effort to create more efficient and deployable machine learning models. In the years following its introduction, knowledge distillation has been extensively studied, developed, and refined, finding wide-ranging applications across various domains of artificial intelligence, particularly in scenarios where model efficiency is crucial.

Knowledge distillation offers solutions to a range of industry challenges. Notable applications include reducing the memory requirements of models, maintaining data privacy during training, and decreasing energy consumption. For example, reducing the memory requirements of a model without compromising performance is particularly beneficial for embedded systems in self-driving cars (Li et al., 2022c). In situations where privacy is a concern, data-free knowledge distillation addresses these issues by generating data, thereby eliminating the need to collect and store sensitive information (Wang et al., 2024). Additionally, knowledge distillation can help lower the energy consumption of a model by transferring the performance of a larger model to a smaller one, with minimal performance loss (Gowda et al., 2024).

Knowledge distillation can be classified into several schemes and algorithms, each focusing on different strategies for transferring knowledge from the teacher model to the student model. **Response-**

* Corresponding author.
E-mail addresses: amir.moslemi@ryerson.ca (A. Moslemi), anna.briskina@senecapolytechnic.ca (A. Briskina), zubeka-dane.dang@senecapolytechnic.ca (Z. Dang), jason.li@senecapolytechnic.ca (J. Li).

https://doi.org/10.1016/j.mlwa.2024.100605
Received 24 October 2024; Received in revised form 8 November 2024; Accepted 9 November 2024
Available online 10 November 2024

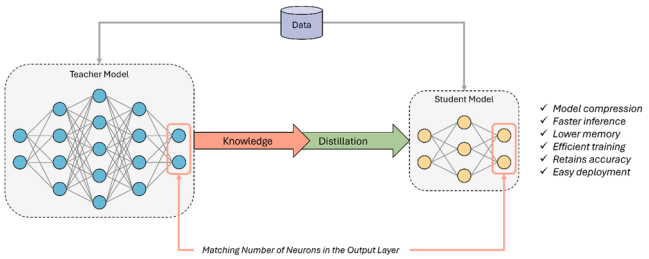


Fig. 1. Overview of the Teacher–Student Knowledge Distillation Framework.

based distillation is the traditional form of KD, where knowledge is transferred by training the student model to mimic the output logits of the teacher model, often using soft targets (Song et al., 2023). This method has been widely utilized due to its simplicity and effectiveness in improving model performance across various tasks. In contrast, **feature-based distillation** does not focus solely on the outputs but rather transfers intermediate features from the teacher to the student. This method aims to align the feature representations and activation maps of the two models, thereby capturing richer knowledge from the teacher network (Ji et al., 2021a). Feature-based distillation has gained popularity for its ability to convey more nuanced information from the teacher to the student. **Relation-based distillation** transfers relational knowledge by teaching the student to understand interdependencies or similarities between instances learned by the teacher. It often finds its application in tasks where understanding the relationships between data points is crucial, making it particularly suitable for graph-based, visual, and language models. Table 1 provides an overview of the various types of knowledge distillation and their applications across different domains, showcasing the distinct methods, knowledge types, and network architectures employed in response-based, feature-based, and relation-based distillation techniques.

Offline distillation involves training the teacher model first and then transferring its knowledge to the student in a separate process. It is commonly used when a strong pre-trained teacher is available, allowing efficient model compression (Srinivasagan et al., 2023; Yin et al., 2022). **Online distillation**, in contrast, simultaneously trains both teacher and student models, facilitating real-time knowledge transfer as both models evolve together. This approach is suitable for applications requiring continual learning (Li et al., 2022a; Chen et al., 2020). Another approach is **self-distillation,** which differs from traditional KD by involving a single model that acts as both the teacher and the student. In this scheme, the model refines its own knowledge through recursive learning processes, enhancing performance without needing a larger teacher network (Zhang et al., 2019; Zhang et al., 2021). This technique simplifies the training process while still achieving performance gains. Table 2 shows an overview of the different knowledge distillation approaches and their applications across various domains, illustrating the diverse methodologies, network architectures, and datasets used in offline, online, and self-distillation approaches.

Various knowledge distillation techniques employ specialized mechanisms and architectural components to effectively capture and transfer complex relationships and contextual information from teacher models to student models. For instance, **cross-modal distillation** involves transferring knowledge between models trained on different modalities (e.g., images to text), enabling the student to understand concepts from diverse data types. In graph neural networks (GNNs), KD is employed to transfer the structural knowledge of graph data. In **graph-based distillation**, the student model learns not only from the node features but also from the intricate relationships between nodes, such as edges and connectivity patterns (Dong et al., 2021; Yang et al., 2022). By capturing the pairwise interactions and the overall graph topology, the student model can effectively mimic the teacher model's performance in tasks like node classification, link prediction, and graph classification. This approach is essential for graph-structured data, where relational information often holds the key to understanding complex structures and patterns. Building on feature-based distillation, **attention-based distillation** guides the student model to focus on specific regions or aspects of the input data deemed important by the teacher. This method leverages attention mechanisms to refine the learning process, making it particularly useful for tasks that require interpretability and attention to detail (Ji et al., 2021a).

Certain knowledge distillation methods are designed to function independently of the original training data, enabling distillation in scenarios where data access is limited or restricted. For example, **data-free distillation** conducts distillation without access to the original training data. This distillation technique is especially useful when access to the original training data is restricted due to privacy or security concerns (Zhu et al., 2021; Chawla et al., 2021). It utilizes synthetic data or generative models to create pseudo-data, allowing the student to learn from the teacher without relying on the original dataset. Alternatively, **adversarial distillation** introduces an adversarial training framework into the distillation process, where the student model is trained to not only mimic the teacher but also to compete against a discriminator model that differentiates between the teacher and student outputs (Goodfellow et al., 2014). This adversarial learning process improves the robustness and generalization of the student model, making it suitable for tasks where security and robustness are critical. Recently, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have become increasingly popular in deep learning and are now widely applied in adversarial distillation (Ye & Bors, 2021; Zhai et al., 2021). In this context, GANs generate challenging examples that help the student model learn to generalize better, effectively transferring more nuanced knowledge from the teacher. By leveraging the adversarial framework, the student model can acquire knowledge that enhances

Table 1
Overview of knowledge distillation types and their applications.

Authors	Knowledge schemes	Methods	Knowledge types	Teacher networks	Student networks	Dataset
Song et al. (2023)	Response-Based Distillation;	EffDstl	Soft Targets	ResNet34	ResNet18	ImageNet
Ahmad et al. (2024)	Response-Based Distillation; Multi-teacher cross-modal distillation;	MTCM-KD	Soft Labels	CDSFL	T1CE modality	BraTS-2021
Kim et al. (2023)	Response-Based Distillation;	RCKD	Soft Labels	MoNuSeg2018	CSAT	TCGA
Ji et al. (2021a)	Feature-Based Distillation; Attention-Based Distillation	AFD	Feature Links; Activation Map	ResNet	ResNet	CIFAR-100; tinyImageNet; ImageNet
Ji et al. (2021b)	Feature-Based Distillation; Self-Distillation	FRSKD	Feature Maps; Soft Labels	ResNet	ResNet	CIFAR-100; tinyImageNet; ImageNet
Sepahvand et al. (2022)	Feature-Based Distillation	Distillation Module	Feature Maps	ResNet-50; VGG-19	LeNet-5; AlexNet	CIFAR-10; Pokemon
Dong et al. (2021)	Relation-Based Distillation	ERDIL	Relation Graph	ResNet	ResNet	CIFAR100; miniImageNet; CUB200
Yang et al. (2022)	Relation-Based Distillation	CIRKD	Pixel-to-Pixel; Pixel-to-Region	DeepLabV3; ResNet-101	DeepLabV3; PSPNet; ResNet18; MobileNetV2	Cityscapes; CamVid; Pascal VOC

Table 2
Overview of knowledge distillation approaches and their applications.

Authors	Knowledge schemes	Methods	Knowledge types	Teacher networks	Student networks	Dataset
Srinivasagan et al. (2023)	Offline Distillation	ITS	Soft Labels	ViTSTR	ViTSTR	MJ; ST
Yin et al. (2022)	Offline Distillation	CVRKD-IQA	High-Quality and Low-Quality Distribution	FR-teacher	NAR-student	Kaddid10K; LIVE; TID2013; KonIQ-10K
Schmid et al. (2022)	Offline Distillation	PaSST	Soft Labels	Transformer	MobileNetV3-Large	AudioSet
Li et al. (2022a)	Online Distillation	OKDHP	Pixel-wise; Feature Maps	HG; FAU	HG; FAU	MPII; COCO
Chen et al. (2020)	Online Distillation	OKDDip	Feature Maps; Soft Labels	ResNet; VGG; DensNet; WRN	ResNet; VGG; DensNet; WRN	CIFAR; ImageNet
Li et al. (2020)	Online Distillation	FFM	Feature Fusion Module	ResNet; Xception; ShuffleNet	ResNet; Xception; ShuffleNet	CIFAR; CINIC-10
Kim et al. (2021)	Self-Distillation	PS-KD	Soft Labels	ResNet; DenseNet; PyramidNet	ResNet; DenseNet; PyramidNet	CIFAR-100
Ge et al. (2021)	Self-Distillation	BAKE	Soft Labels	RestNet; MobileNet; EfficientNet	RestNet; MobileNet; EfficientNet	ImageNet-1K; CIFAR-100; TinyImageNet; CUB-200-2011; Stanford Dogs; MIT67
Yue et al. (2022)	Self-Distillation	SSAD	Soft Labels	SSJD; PCN; SSL	SSJD; PCN; SSL	Indian Pines; University of Pavia; Houston
Ji et al. (2021b)	Feature-Based Distillation; Self-Distillation	FRSKD	Feature Maps; Soft Labels	ResNet	ResNet	CIFAR-100; tinyImageNet; ImageNet

its ability to handle complex, real-world scenarios, particularly in domains like image synthesis, anomaly detection, and robust classification (Higuchi et al., 2022; Lee et al., 2022; He et al., 2022).

Several distillation techniques focus on optimizing the KD process to enhance efficiency and performance. For instance, **quantized distillation** aims to reduce the memory and computational footprint of the student model by employing techniques like quantization during the distillation process (Kim et al., 2019). This involves compressing the model's parameters into lower precision representations without significantly sacrificing performance. Additionally, **NAS-based distillation** integrates Neural Architecture Search (NAS) techniques with KD to automatically find the most optimized student network architecture (Lee et al, 2023). This approach strikes a balance between performance and computational efficiency and is increasingly popular for designing student networks tailored to specific tasks and constraints.

Additionally, ensemble-based knowledge distillation approaches combine insights from multiple teacher models, aiming to improve the accuracy and robustness of the resulting student model. For example, **lifelong distillation** focuses on scenarios where models need to continuously learn from new data while retaining previously acquired knowledge, employing distillation to mitigate catastrophic forgetting (Ye & Bors, 2021; Zhai & Mori, 2021). This approach is particularly relevant in dynamic environments where models need to adapt over time, such as real-time monitoring systems and evolving data streams. Lastly, **multi-teacher distillation** employs multiple teacher models, often from different modalities, to provide a comprehensive knowledge base for training the student model (Ahmad et al., 2024). By combining diverse forms of knowledge, this approach improves the student model's generalization capabilities.

This paper provides a comprehensive overview of how knowledge distillation has evolved and been applied, exploring the diverse approaches that have emerged in this rapidly advancing field and emphasizing the importance of KD in enabling advanced artificial intelligence capabilities on resource-constrained devices. The methods presented in this paper demonstrate the versatility and applicability of knowledge distillation in a range of real-world contexts. Table 3 offers a detailed summary of KD methods employed across various domains.

The structure of this paper is illustrated in Fig. 2. This figure categorizes various types of knowledge distillation (KD) techniques based on their approaches, the types of knowledge they leverage, and their respective applications (Figs. 3–16).

2. Types of knowledge distillation

This section explores the primary types of knowledge distillation, including response-based, feature-based, and relation-based distillation. These methods vary in the way they extract and transfer knowledge from the teacher model, offering diverse strategies to improve student model performance while maintaining efficiency.

2.1. Response-based knowledge distillation

Response-based knowledge distillation is a technique where a student model learns to mimic a teacher model's soft output probabilities, or "soft targets." These soft targets provide a nuanced view of data, encoding crucial class relationships for generalization. This approach helps the student model understand the teacher's decision-making process, improving performance and avoiding overfitting (Gou et al., 2021).

The loss function of response-based KD focuses on matching the output of the teacher and student models. This is done by minimizing the Kullback-Leibler (KL) divergence between the softened output probabilities of the teacher and the student networks (Hilton et al., 2015). The softened output of teacher and student is formulated as:

$$p_t = \sigma\left(\frac{z_t}{T}\right) = \frac{\exp\left(\frac{z_t}{T}\right)}{\sum_{i=1}^{N}\exp\left(\frac{z_t^i}{T}\right)} \tag{1.1.1}$$

$$p_s = \sigma\left(\frac{z_s}{T}\right) = \frac{\exp\left(\frac{z_s}{T}\right)}{\sum_{i=1}^{N}\exp\left(\frac{z_s^i}{T}\right)} \tag{1.1.2}$$

Where:

N is the number of classes.

z_t and z_s represent the logits (pre-softmax outputs) of the teacher and student networks, respectively.

T is the temperature parameter used to soften the logits and normally set to 1 or higher to soften the probability.

$\sigma(x)$ is the softmax function.

The loss function of Response-Based KD is combined of KL divergence distillation loss and the Cross-Entropy (CE) task loss. The KL divergence loss minimizes the difference between the soft logits of the teacher and the student, while the CE loss minimizes the error between

Table 3
Comprehensive overview of specialized knowledge distillation techniques.

Authors	Knowledge schemes	Methods	Knowledge types	Teacher networks	Student networks	Dataset
Ahmad et al. (2024)	Response-Based Distillation; Multi-teacher cross-modal distillation;	MTCM-KD	Soft Labels	CDSFL	T1CE modality	BraTS-2021
Ji et al. (2021a)	Feature-Based Distillation; Attention-Based Distillation	AFD	Feature Links; Activation Map	ResNet	ResNet	CIFAR-100; tinyImageNet; ImageNet
Lee et al. (2022)	Adversarial Distillation	GCNs	Feature Matrix; Similarity Matrix:	ResNet; MobileNet; Wide ResNet	ResNet; MobileNet; Wide ResNet	CIFAR
Higuchi et al. (2022)	Adversarial Distillation	adv-CNN	Soft Labels	ResNet	ResNet	CIFAR-10
He et al. (2022)	Adversarial Distillation	GraphAKD	Discriminators	GCNII; GAMLP	GCN; Cluster-GCN	Cora; CiteSeer; PubMed; Flickr; Arxiv; Reddit; Yelp; Products
Wu et al. (2021)	Multi-teacher Distillation	MT-BERT	Soft Labels	BERT; RoBERTa; UniLM	UniLM	SST2; RTE; MIND
Pham et al. (2022)	Multi-teacher Distillation	CMT-KD	Feature Maps	AlexNet; ResNet18	AlexNet; ResNet18	CIFAR-100; ImageNet
Liu et al. (2020)	Multi-teacher Distillation	AMTML-KD	Soft Labels; Intermediate Features	ResNet; VGG; DensNet	Stu1; Stu2; Stu3	CIFAR; TinyImageNet
Xue et al. (2022)	Cross-Modal Distillation	MFH; MVD	Modality Features	ResNet	ResNet	Gaussian; AVMNIST; RAVDESS; VGGSound; NYU Depth V2; MM-IMDB
Sarkar and Etemad (2022)	Cross-Modal Distillation	XKD	Attention Maps; Modality Features	ViT	ViT	UCF101; HMDB51; Kinetics400; Kinetics-Sound; AudioSet; ESC50; FSD50K
Xia et al. (2023)	Cross-Modal Distillation	MNF; CSC	Modality Features	ResNet-50	ResNet-50	UCF51; ActivityNet
Yang et al. (2020)	Graph-Based Distillation	LSP	Feature Maps	GCN	GCN	PPI; ModelNet40
Liu et al. (2022a)	Graph-Based Distillation	HIRE	Soft Labels; Semantic Relations; Intermediate Features	GCN; GAT; RGCN; HAN; HGT; HGConv	GCN; GAT; RGCN; HAN; HGT; HGConv	ACM; IMDB; DBLP
Wang et al. (2022a)	Graph-Based Distillation	CKD	Regional Knowledge; Global Knowledge	HIN	HIN	Pubmed; DBLP; ACM; Freebas
Ji et al. (2021a)	Attention-Based Distillation	AFD	Feature Maps	ResNet; Wide ResNet	ResNet; Wide ResNet	CIFAR-100; tinyImageNet; ImageNet
Passban et al. (2020)	Attention-Based Distillation	ALP-KD	Soft Labels	BERT	BERT	CoLA; MNLI; MRPC; QNLI; QQP; RTE; SST-2; STS-B
Wu et al. (2021)	Attention-Based Distillation	Universal-KD	Intermediate Features	BERT	BERT	GLUE
Zhu et al. (2021)	Data-Free Distillation	FEDGEN	Feature Maps	FEDGEN	FEDGEN	MNIST; EMNIST; CELEBA
Chawla et al. (2021)	Data-Free Distillation	DIODE	Soft Labels	Yolo-V3	Yolo-V3	MSCOCO
Fang et al. (2022)	Data-Free Distillation	FastDFKD	Feature Maps	ResNet; VGG; WRN	ResNet; VGG; WRN	CIFAR; NYUv2; ImageNet
Kim et al. (2019)	Quantized Distillation	QKD	Feature Maps	ResNet; MobileNet	ResNet; MobilenetV3	CIFAR; ImageNet
Zhao and Zhao (2024)	Quantized Distillation	SQAKD	Feature Maps	ResNet; VGG	MobileNet; ShuffleNet; SqueezeNet	CIFAR-10; CIFAR-100; TinyImageNet
Boo et al. (2021)	Quantized Distillation	SPEQ	Intermediate Features	ResNet; VGG; MobileNet	ResNet; VGG; MobileNet	CIFAR10; CIFAR100; ImageNet
Ye and Bors (2021)	Lifelong Distillation; Adversarial Distillation	LT-GANs	Discriminator	LT-GANs	LT-GANs	CIFAR; ImageNet; MNIST; SVHN; Fashion; Omniglot
Zhai et al. (2021)	Lifelong Distillation	Hyper-LifelongGAN	Discriminator	Hyper-LifelongGAN	Hyper-LifelongGAN	Image
Lee et al. (2023)	NAS-Based Distillation	DaSS	Meta-Knowledge; Search Strategy	ResNet	ResNet	TinyImageNe
Trivedi et al. (2023)	NAS-Based Distillation	KD-NAS	Search Strategy	XLM-Roberta	KD-NAS	CC100
Liu et al. (2022b)	Cross-Architecture Distillation	ViT; CNNs	Feature Maps; PCA	ViT; CNNs	ViT; CNNs	CIFAR; ImageNet
Hao et al. (2024)	Cross-architecture distillation	OFA-KD	Intermediate Features	ResNet	Swin	CIFAR; ImageNet
Dong et al. (2023)	NAS-Based Distillation	DisWOT	Feature Semantics	ResNet	ResNet	CIFAR; ImageNet
Fang et al. (2021a)	Cross-Modal Distillation; VL Distillation	DistillVLM	Soft Labels; Intermediate Features	VL	DistillVLM	COCO Captioning; VQA
Fang et al. (2021b)	Self-Distillation	SEED	Feature Maps	EfficientNet; MobileNet	EfficientNet; MobileNet	ImageNet
Fang and Yang (2023)	Cross-Modal Distillation; VL Distillation	VL	Soft Labels; Intermediate Features	VL	DistillVLM	COCO Captioning; VQA

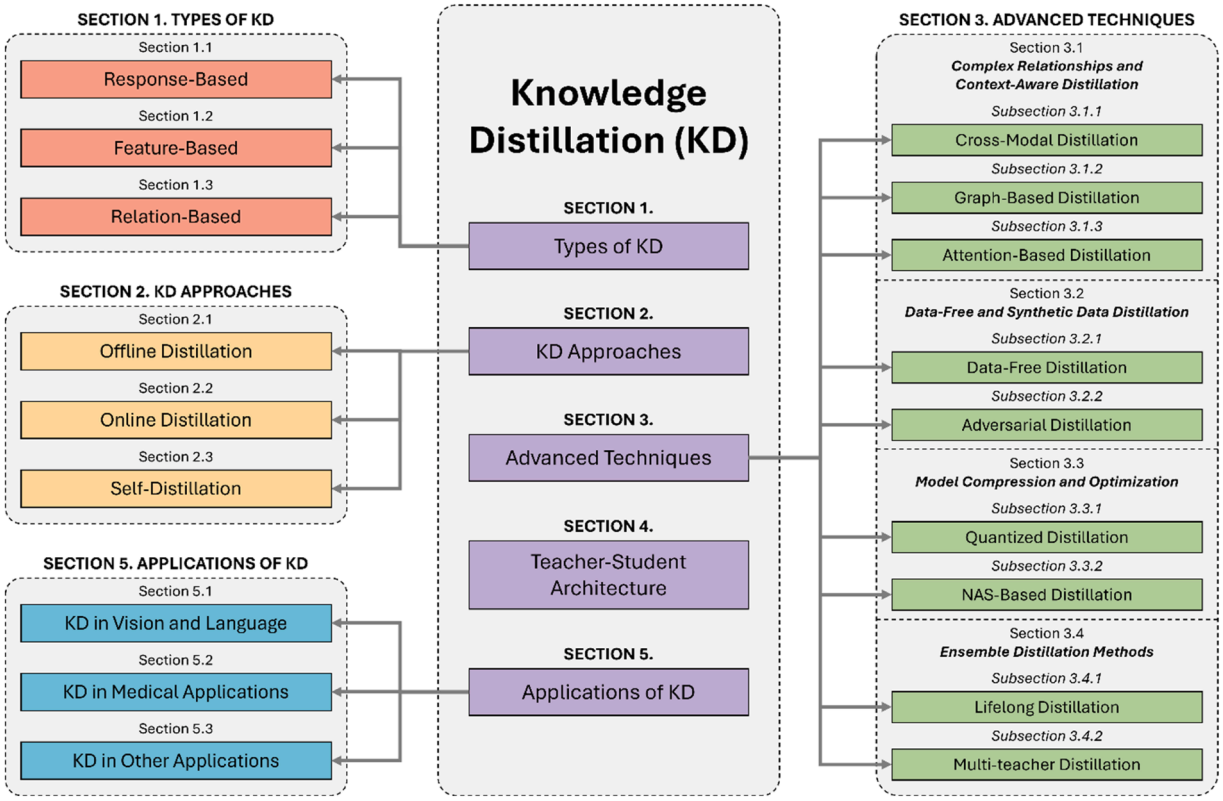


Fig. 2. The schematic structure of the paper and the relationship between its sections. The body of this survey mainly contains the fundamentals of knowledge distillation approaches, advanced techniques, and applications of knowledge distillation. Subsections of each main section are listed in this figure.

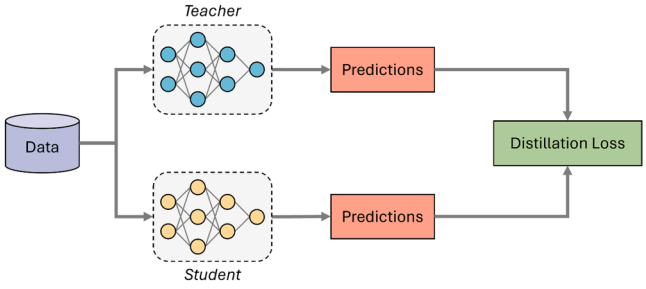


Fig. 3. A schematic framework of response-based knowledge distillation.

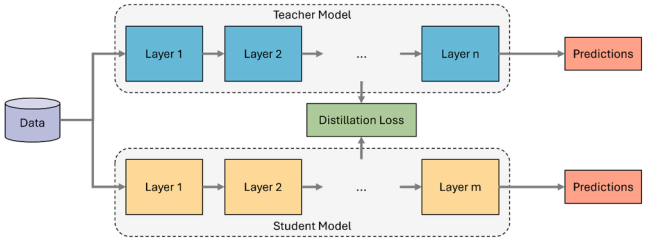


Fig. 4. A schematic framework of feature-based knowledge distillation.

the student's soft logits and the ground truth labels y.

The KL divergence loss is formulated as:

$$L_{KL} = T^2 \sum_{i}^{N} p_t^i(T) \log \frac{p_t^i(T)}{p_s^i(T)} \tag{1.1.3}$$

Where:

p_t^i and p_s^i is the teacher and the student model's predicted probability

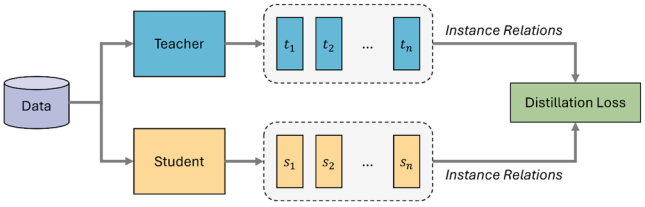


Fig. 5. A schematic framework of relation-based knowledge distillation.

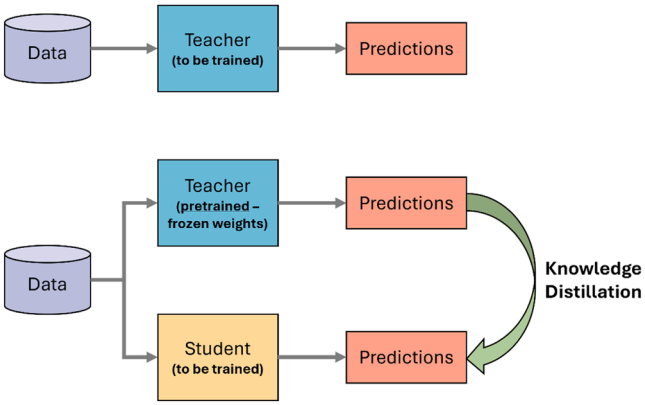


Fig. 6. A schematic framework of offline knowledge distillation.

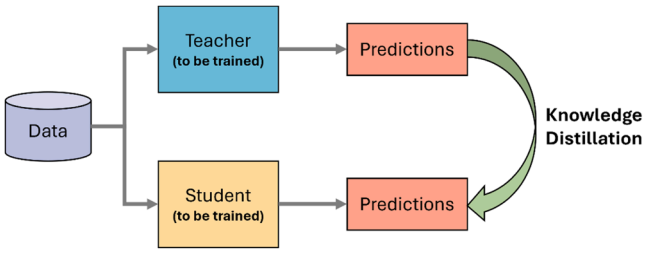


Fig. 7. A schematic framework of online knowledge distillation.

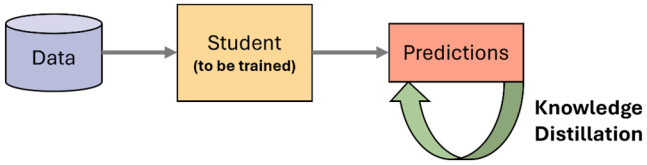


Fig. 8. A schematic framework of self-distillation.

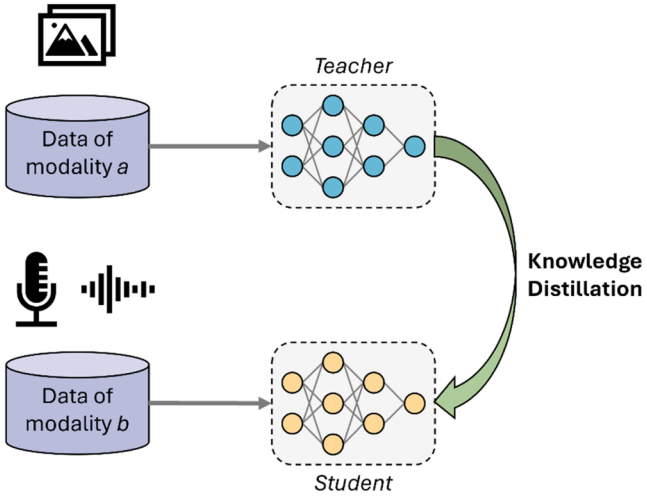


Fig. 9. A schematic framework of cross-modal knowledge distillation.

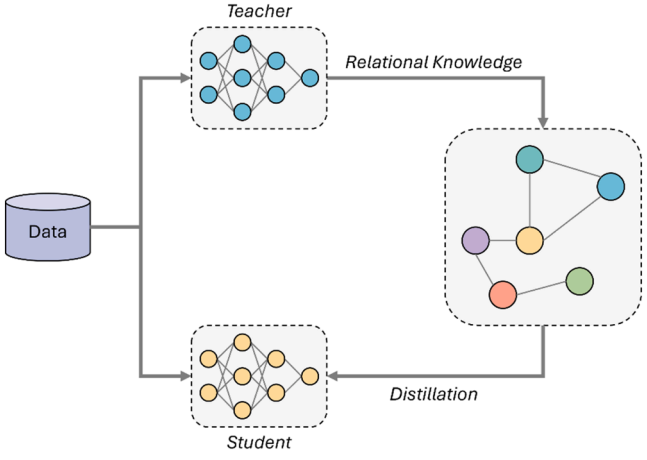


Fig. 10. A schematic framework of graph-based knowledge distillation.

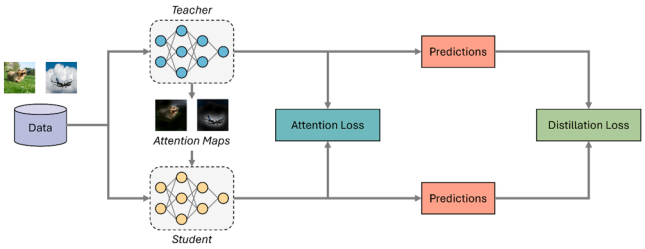


Fig. 11. A schematic framework of attention-based knowledge distillation.

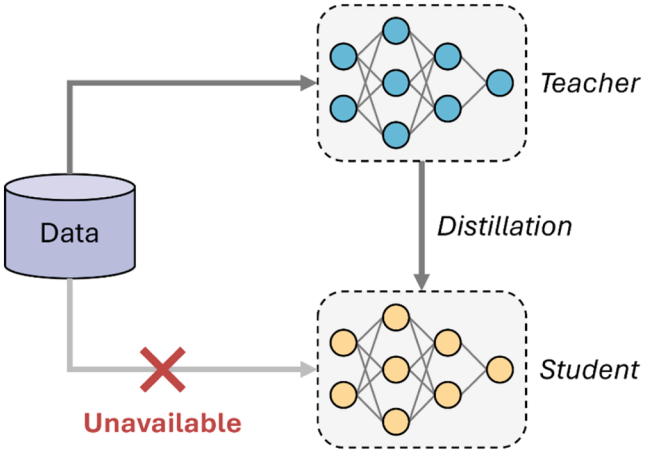


Fig. 12. A schematic framework of data-free knowledge distillation.

for class i , respectively.

The CE task loss is formulated as:

$$L_{CE} = CE(y, \sigma(z_s)) = -\sum_{i=1}^{N} y_i \log\left(p_s^i\right) \tag{1.1.4}$$

Where y_i is the truth label for class .

By combining equations (1.1.3) and (1.1.4), the overall loss function for Response-Based KD becomes:

$$L_{ResponseKD} = \alpha L_{CE} + (1-\alpha) L_{KL} \tag{1.1.5}$$

Where α is a hyperparameter that balances the contributions of the KL distillation loss and the CE task loss.

Song et al. (2023) analyzed knowledge transfer from weak and strong teachers, concluding that strong teachers generally lead to more effective distillation. However, they also demonstrated the value of weak teachers through their Efficient Distillation (EffDstl) framework, which minimizes reliance on computationally heavy models. Ahmad et al. (2024) introduced MTCM-KD, a multi-teacher cross-modal distillation approach for unimodal segmentation. This technique integrates response-based KD from multiple teachers trained on different modalities, achieving improved segmentation accuracy through a cooperative deep supervision fusion learning framework. Kim et al. (2023) proposed Response-Based Cross-Task Knowledge Distillation (RCKD), applying response-based KD across different tasks. This method uses soft targets from a teacher model trained on one task to guide a student model on a related but distinct task, effectively capturing shared patterns and improving performance in scenarios with limited data.

In summary, response-based KD is widely adopted for its simplicity and direct method of transferring knowledge, making it effective across various applications. It uses soft target outputs from the teacher model, which can help avoid overfitting in the student model. However, this approach can face performance limitations when there is a large capacity gap between the teacher and student models, which can make it difficult for the student to fully replicate the teacher's output patterns,

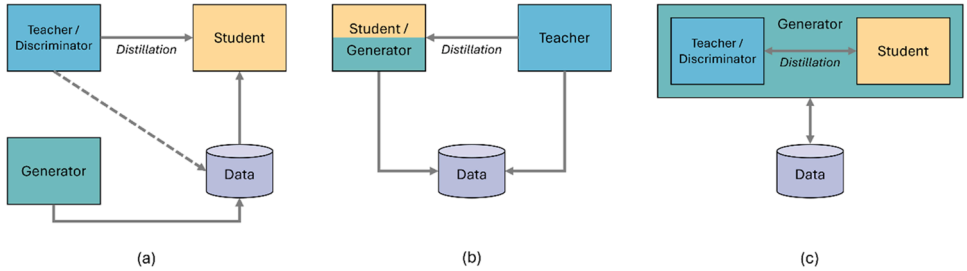


Fig. 13. Types of primary adversarial distillation techniques. (a) The generator in the GAN framework creates training data to enhance the effectiveness of knowledge distillation, with the teacher model serving as the discriminator. (b) The GAN discriminator ensures that the student model, which also acts as a generator, closely imitates the teacher model. (c) Both teacher and student models operate as a generator, while the discriminator supports online knowledge distillation by refining the process.

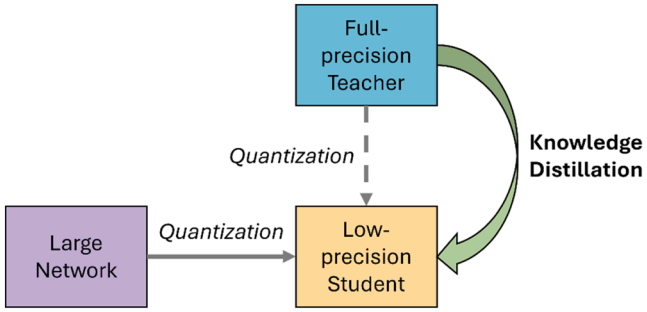


Fig. 14. A schematic framework of quantized knowledge distillation.

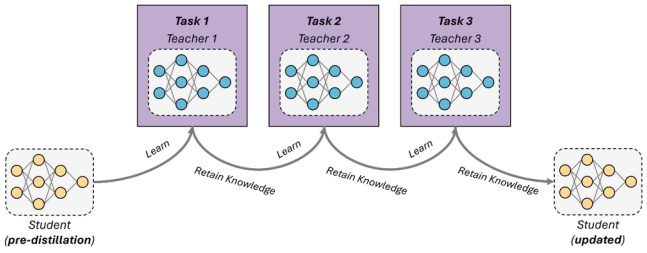


Fig. 15. A schematic framework of lifelong knowledge distillation.

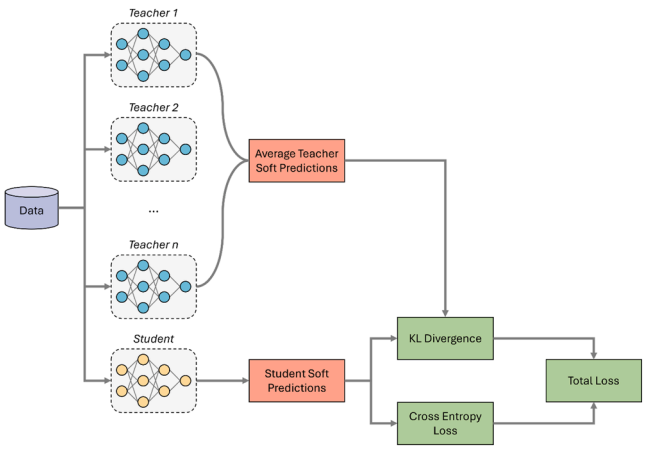


Fig. 16. A schematic framework of multi-teacher knowledge distillation.

leading to suboptimal learning outcomes (Mirzadeh et al., 2020).

2.2. Feature-based knowledge distillation

Feature-based knowledge distillation transfers internal representations from a teacher model to a student model, enabling the student to capture intricate structures and relationships encoded in the teacher's feature maps (Yang et al., 2023a). This approach provides richer knowledge transfer than simply mimicking output probabilities.

The loss function of feature-based KD focuses on aligning the intermediate feature representations between the teacher and student networks, and normally can be formulated as:

$$L_{Feature} = \frac{1}{N}\sum_{i=1}^{N} \| F_T^i - F_S^i \|_2^2 \tag{1.2.1}$$

Where:

$L_{Feature}$ is the feature-based KD loss function.

N is the number of feature layers or map points considered for distillation.

F_T^i and F_S^i represent the feature maps of the teacher and student networks, respectively.

$\| \cdot \|_2^2$ denotes the squared Euclidean (L_2) norm between the teacher's and student's feature maps.

The total KD loss is a combination of the feature-based KD loss and the task loss (typically CE loss), expressed as:

$$L_{KD} = \alpha L_{CE} + (1-\alpha) L_{Feature} \tag{1.2.2}$$

Where α is a hyperparameter that balances the contributions of the feature-based KD and the CE task loss.

Ji et al. (2021a) proposed an attention-based feature alignment mechanism in "Show, Attend and Distill: Knowledge Distillation via Attention-Based Feature Matching." This method uses an attention mechanism to align intermediate feature representations, ensuring the student focuses on critical parts of the feature space. In another paper, Ji et al. (2021b) introduced Feature Refinement via Self-Knowledge Distillation (FRSKD), a self-teaching approach where a model refines its own features through iterative self-knowledge distillation. This method incorporates top-down and bottom-up paths in an auxiliary self-teacher network, generating refined feature maps and soft labels as pseudo-labels. Sepahvand et al. (2022) presented a technique for knowledge distillation in resource-constrained environments, decomposing the teacher's deep feature representations into manageable components. This approach is particularly suitable for intelligent mobile applications.

Overall, feature-based KD captures detailed internal representations from the teacher model, leading to a richer transfer of knowledge and improved performance in complex tasks. However, despite its effectiveness, feature-based KD faces challenges, such as difficulty in aligning feature maps between teacher and student models with different

architectures or capacities (Yang et al., 2023a); potential transfer of redundant or irrelevant information (Heo et al., 2019); and sensitivity to the selection of layers for knowledge distillation (Yim et al., 2017).

2.3. Relation-based knowledge distillation

Relation-based knowledge distillation is an advanced technique that preserves structural relationships and dependencies within data learned by a teacher model (Park et al. 2019). Unlike earlier KD methods that focus on output probabilities or feature representations, relation-based KD captures and transfers relational information between different data points or features, such as maintaining pairwise similarities in the teacher's feature space (Tung and Mori, 2019), angular relationships between data points (Park et al. 2019), or the correlation between instances (Peng et al., 2019).

The relation-based knowledge distillation loss function focuses on preserving the relational structure of the data learned by the teacher to transfer the relational knowledge between pairs or groups of data points from the teacher network to the student network. After establishing a relational measure between data points (e.g., pairwise distance or cosine similarity), the correlation between the teacher's and student's feature maps is calculated.

The pairwise distance between two different data points can be computed as:

$$R_T(i,j) = \|F_T^i - F_T^j\|_2 \tag{1.3.1}$$

Additionally, the cosine similarity to measure the similarity between feature representations is defined as:

$$R_T(i,j) = \frac{F_T^i \cdot F_T^j}{\|F_T^i\| \cdot \|F_T^j\|} \tag{1.3.2}$$

Here F_T^i and F_T^j are the feature representations of data points i and j in the teacher network Similarly, F_S^i and F_S^j represent the corresponding feature representations of data points i and j in the student network.

The relation-based knowledge distillation loss can be formulated as:

$$L_{Relation} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \| R_T(i,j) - R_S(i,j) \|_2^2 \tag{1.3.3}$$

Where:

N is the number of data points.

$R_T(i,j)$ and $R_S(i,j)$ are the distances or the similarities for the teacher and student networks for data points i and j, respectively.

$\| \cdot \|_2^2$ denotes the squared Euclidean (L_2) norm between the teacher's and student's relational matrices.

The total KD loss is a combination of the relation-based KD loss and the task loss (typical CE loss), expressed as:

$$L_{KD} = \alpha L_{CE} + (1-\alpha) L_{Relation} \tag{1.3.4}$$

Where α is a hyperparameter that balances the contributions of the relation-based KD and the CE task loss.

Recent applications have demonstrated the versatility and effectiveness of relation-based KD in various domains. Dong et al. (2021) extended the approach to few-shot class-incremental learning (FSCIL) with their Exemplar Relation Distillation Incremental Learning (ERDIL) framework. ERDIL leverages an Exemplar Relation Graph (ERG) to capture relational knowledge between classes and transfers this knowledge through an Exemplar Relation Loss function. This innovative approach has shown impressive results, outperforming state-of-the-art methods on benchmark datasets such as CIFAR100 and Mini ImageNet. In the field of semantic segmentation, Yang et al. (2022) proposed Cross-Image Relational KD (CIRKD), a method that transfers structured pixel-to-pixel and pixel-to-region relations across entire images. CIRKD employs a memory bank to store rich embeddings, allowing the student model to capture global pixel dependencies across different images. This approach has demonstrated superior performance compared to existing KD techniques on datasets like Cityscapes and Pascal VOC, highlighting the importance of preserving complex relational information in image segmentation tasks.

In essence, relation-based KD effectively captures structural and relational information within data, making it highly suited for graph-based models and tasks that require understanding inter-data relationships. However, this technique can be computationally intensive and may require careful design of relational loss functions to avoid losing critical structural information during transfer.

A practical example provided by Zhao et al. (2023) demonstrates how response-based, feature-based, and relation-based KD techniques can be effectively utilized in addressing industry challenges. The study uses a Semi-Supervised Knowledge Distillation (SSKD) framework to improve the performance of deep learning models for vision-based robot guidance in smart manufacturing. The framework addresses the challenge of obtaining sufficient labeled data and the need for efficient models in real-world factory settings. It uses KD methods to transfer knowledge from a larger, pre-trained teacher model (YOLOv5m) to a smaller, more efficient student model (YOLOv5s). The study employs a combination of response-based, feature-based, and relation-based KD techniques to train the student model. This process involves using the teacher model's outputs, intermediate feature maps, and relational information between detected objects to guide the student model's learning. By using KD, the framework reduces inference time from 185 ms to 45 ms while maintaining high accuracy, achieving recall and precision values exceeding 99.5% and 92.6%, respectively. This results in a significant improvement in efficiency and generalizability across different working environments, ultimately achieving a 200% improvement in labor efficiency.

3. Knowledge distillation approaches

Knowledge distillation can be implemented in several ways, each tailored to different training scenarios and model configurations. The primary approaches include offline distillation, online distillation, and self-distillation. This section will explore these approaches in detail, highlighting their unique characteristics, benefits, and new applications across various domains.

3.1. Offline distillation

Offline knowledge distillation is a technique where knowledge is transferred from a pre-trained teacher model to a smaller student model during a separate training phase. As described by Gou et al. (2021), this method allows for the compression of large, complex models into more efficient versions suitable for deployment on resource-constrained devices while maintaining comparable performance.

The offline knowledge distillation process typically involves two main steps. First, a teacher model is pretrained on the dataset. Next, the outputs from this teacher model are used as soft labels to guide the training of a student model. In this stage, the student model learns by minimizing the loss between the true labels and the teacher's soft labels (Gou et al., 2024).

Despite its efficiency, offline KD faces several challenges. The static nature of the teacher model can limit the student's adaptability to new data patterns, and biases in the teacher may be transferred to the student (Zhang et al., 2018). The method's reliance on large, computationally expensive teacher models can be impractical in resource-limited environments. Additionally, the student's performance is often tied to the teacher's quality, potentially inheriting its limitations (Lan et al., 2018). Offline KD may also struggle to transfer fine-grained knowledge from the teacher's intermediate layers (Chen et al., 2017).

Recent studies demonstrate the versatility and effectiveness of offline KD across various domains. Srinivasagan et al. (2023) applied offline distillation to compress an image-to-speech system for low-resource

devices, achieving significant reductions in model size and inference time while maintaining audio quality. In the field of image quality assessment, Yin et al. (2022) used offline KD to transfer knowledge from a full-reference teacher model to a non-aligned reference student model, enhancing performance while maintaining efficiency. Schmid et al. (2022) employed offline KD to transfer knowledge from a high-performing Transformer to a more efficient Convolutional Neural Network (CNN) for large-scale audio tagging, achieving state-of-the-art results on the AudioSet dataset.

Ultimately, offline KD enables efficient compression of large models for deployment on resource-limited devices, though it can limit the student's adaptability to new data and may inherit biases from the static teacher model.

3.2. Online distillation

Online knowledge distillation is an innovative approach that involves the simultaneous training of multiple model branches in a mutual learning environment. Unlike offline distillation, which requires a pre-trained teacher model, online KD allows for continuous feedback and knowledge sharing between peer models during the learning process. This method aims to enhance overall model performance by improving diversity and preventing homogenization among student branches (Chen et al., 2020).

The standard process of online knowledge distillation involves training multiple student models concurrently. The outputs of these models are aggregated and compared with the true labels as well as with each other. This approach allows the student models to learn simultaneously and strengthens the robustness of the learning process (Chen et al., 2020).

Nevertheless, online KD faces certain challenges, such as the increased computational complexity, as simultaneous training of multiple models can be resource-intensive, particularly for large-scale models or datasets (Anil et al., 2018). Managing diversity among student branches remains challenging, as they may still converge toward similar representations without careful design (Chen et al., 2020). Additionally, the evolving nature of the teacher model during training can lead to suboptimal knowledge transfer if it is not adequately robust early in the process (Lan et al., 2018).

Recent research has demonstrated the versatility and effectiveness of online KD in various domains. Li et al. (2022a) proposed a novel approach for pixel-level human pose estimation using a Feature Aggregation Unit (FAU). This method generates and aggregates heatmaps from diverse student branches, resulting in a more robust and accurate pose estimation model. The diversity introduced by different student architectures enhances overall accuracy by providing varied perspectives on the task. Chen et al. (2020) introduced the OKDDip framework, which employs a two-stage process to reduce the risk of homogenization in student models. This method uses auxiliary student branches and a group leader to aggregate learned knowledge, incorporating an attention mechanism to assign weights based on branch performance. Li et al. (2020) further improved upon this framework by introducing a feature fusion module and a classifier diversification loss function, enhancing the diversity and robustness of the knowledge-sharing process.

Reflecting the real-world utility of this approach, Lin et al. (2023) use online KD, specifically deep mutual learning (DML), to improve the accuracy and stability of carbon emission forecasting in the electric power industry. The study employs two student models — Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) — that learn from each other during training. This cooperative learning process helps to reduce overfitting and enhances the models' ability to capture complex patterns within the time series data. By improving forecasting accuracy, the study aims to help power industry decision-makers adjust power generation policies, ultimately leading to the achievement of carbon reduction targets and the goals of carbon peaking and neutrality.

Overall, online KD enhances model diversity and robustness by allowing continuous feedback among multiple student branches, though its simultaneous training approach can be computationally intensive and requires careful management to maintain representation diversity.

3.3. Self-Distillation

Self-distillation is a variant of knowledge distillation where a model acts as both the teacher and the student, using its own knowledge to guide its learning process. Unlike traditional distillation, where a large teacher model transfers knowledge to a smaller student model, self-distillation encourages the model to refine its own predictions iteratively during training (Zhang et al., 2019). While self-distillation can lead to improved model performance and generalization, it has several limitations. One of the primary challenges is the risk of reinforcing incorrect or suboptimal predictions, as the model relies heavily on its own outputs without external supervision. This can lead to overfitting, especially if the initial predictions are flawed (Chen & Chu, 2023). Additionally, self-distillation requires sophisticated training schedules and careful tuning of hyperparameters to ensure that the model refines its knowledge in a meaningful way, which can add complexity to the training process (Pham et al., 2022a).

The typical process of self-distillation involves dividing the model into segments, such as each residual block in a ResNet model. Each segment's feature maps are then connected to a fully connected layer and a classifier. These branched out segments are further connected to each other to enable the transfer of complex knowledge from deeper blocks to shallower ones (Zhang et al., 2019). In other words, the deepest block acts as a teacher model for the shallower segments. It transfers knowledge by minimizing the loss between its branch and those of the shallower layers, as well as the ground truth. Once training is complete, the branches are discarded.

Recent advancements in self-distillation techniques include the work of Zhang et al. (2021), who introduced a method of self-distillation that progressively refines the network's internal representations. In this approach, intermediate layers act as "teachers" for deeper layers, allowing the network to become more efficient and compact without external supervision. Their method successfully reduces model size while maintaining or even improving performance, making it particularly useful for deployment in resource-constrained environments. This approach has demonstrated strong results in model compression and efficiency, further supporting the potential of self-distillation in practical applications.

Kim et al. (2021) introduced progressive self-knowledge distillation (PS-KD), which continuously updates predictions during training to find flatter minima in the loss landscape. While improving generalization and robustness, PS-KD requires precise control of training dynamics and careful hyperparameter tuning. Ge et al. (2021) proposed Batch Knowledge Ensembling (BAKE), which aggregates predictions from different mini batches during training. This approach creates more diverse learning targets, improving generalization and mitigating the risk of reinforcing incorrect predictions. BAKE demonstrated significant improvements in ImageNet classification accuracy. Yue et al. (2022) combined self-supervised learning with adaptive distillation for hyperspectral image classification. This method uses self-supervised pre--training to extract features from unlabeled data, followed by adaptive distillation that dynamically adjusts the model's self-learning process. This approach effectively handles complex data structures and improves classification accuracy in hyperspectral imaging.

The main superiority of self distillation over teacher-student approach is that no extra teacher is required. Conversely, teacher-student approach is working based on training an overparameterized teacher model at first to teach shallow model (student). Designing and training the teacher model are the main challenges for traditional KD techniques. Additionally, training teachers (deep networks) takes long time, since teacher model is an over-parameterized network. Zhang et al. (2019) experimentally showed that self distillation technique

outperformed deep networks such as VGG19, ResNet18 and ResNet50 on CIFAR dataset. The reason why shallow network with self distillation could outperform all deep supervision is that an extra bottleneck was added to identify classifier-specific features.

Dual Teachers for Self-Knowledge Distillation (DTSKD) was proposed to enhance the performance of self distillation approach (Li et al. 2024a). In DTSKD, student is trained by two teachers based on self supervision staretgy from two inherently different domains including the past learning history and the current network structure. In specific, historical teacher and structural teacher are two teachers of DTSKD to train student. Historical teacher provides knowledge from the previous epoch and structural teacher distils the knowledge form the current iteration. Yang et al. (2023b) proposed Universal Self-Knowledge Distillation (USKD). In USKD, KD loss is decomposed to a Normalized KD (NKD) loss and soft labels are constructed for both target class and non-target classes. USKD constructs soft labels for both target and non-target classes without having a teacher by smoothing the target logit of the student as the soft target label. Lee et al. (2023b) proposed self- distillation drop-out (SD-Dropout) to reduce the number of trainable parameters. In SD-Dropout, distributions of multiple models are distilled using a dropout sampling. SD-Dropout is simple yet effective approach to train a shallow model based on self supervise learning. Inconsistency between deep and shallow classifiers in self-KD during distilling knowledge is one of the concerns. In this regard, Liang et al. (2024) proposed Neighbor Self-Knowledge Distillation (NSKD) to circumvent mismatch problem between deep and shallow classifiers using auxiliary classifiers which are added to the shallow parts of the network. These auxiliary classifiers provide distillations of multiple neighboring which leads to reduce the mismatch.

To demonstrate a practical application of self-distillation, the study by Li et al. (2024b) proposes a deep knowledge distillation model for traffic prediction that leverages self-distillation and mutual learning techniques to improve the accuracy of traffic flow forecasting. The model employs two graph neural networks with encoder-decoder structures. Self-distillation is employed within each network, enhancing feature sensitivity by transferring knowledge from the deeper structure to the shallower structure. Mutual learning enables the two networks to learn collaboratively, improving feature learning and generalization by minimizing the divergence between their predictions. Experiments on real-world datasets demonstrate the model's effectiveness, particularly for long-term predictions (beyond 30 minutes). The deep knowledge distillation model achieved lower Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE) compared to the baseline DCRNN model, with improvements of 0.19, 0.7%, and 0.49 respectively on the METR-LA dataset. On the PEMS-BAY dataset, the model achieved even better results, reducing MAE by 0.18, MAPE by 0.4%, and RMSE by 0.41. These results indicate the effectiveness of the hybrid KD approach in capturing spatiotemporal dependencies and enhancing the feature extraction capabilities of graph neural networks for improved traffic flow prediction.

In summary, self-distillation allows a model to refine its predictions by learning from its own knowledge, which can lead to enhanced performance without needing a larger teacher model. However, this technique may inadvertently reinforce its own inaccurate predictions and requires precise training schedules and fine-tuning, which can add complexity to the process.

4. Advanced Techniques in Knowledge Distillation

Advanced techniques in knowledge distillation expand on traditional methods to tackle more complex tasks. These methods can be grouped into four main categories: Complex Relationships and Context-Aware Distillation, Data-Free and Synthetic Data Distillation, Model Compression and Architecture Optimization, and Ensemble Distillation Methods. This section provides an overview of these techniques and their various applications across different domains.

4.1. Complex relationships and context-aware distillation

This section explores advanced KD methods leverage specific mechanisms or architectural components to enhance knowledge transfer. Cross-modal distillation transfers knowledge between different data modalities (e.g., from images to text). Graph-based distillation employs graph structures to capture and convey relational information. Finally, attention-based distillation utilizes attention mechanisms to align the focus of the student model with that of the teacher. By exploiting these mechanisms, the student model can more effectively replicate the teacher's performance.

4.1.1. Cross-modal distillation

Cross-modal knowledge distillation involves transferring knowledge from a teacher network trained on one modality, such as images, to a student network trained on a different modality, such as audio. This approach allows the student model to leverage rich information from another modality, often improving performance in tasks where data from one modality is scarce or less informative (Gupta et al., 2016). However, cross-modal KD introduces unique challenges, such as aligning disparate feature spaces and dealing with modality-specific noise or irrelevant features that can hinder the effectiveness of the knowledge transfer (Wang et al., 2023; Huo et al., 2024).

A common cross-modal distillation process incorporates several main stages. First, a high-capacity teacher model is trained on the source modality with abundant labeled data, learning rich feature representations. Second, a student model designed for the target modality is initialized, often facing limited labeled data (Gupta et al., 2016). Third, mechanisms are established to align the feature spaces of the teacher and student models, such as shared embedding spaces or adaptation layers that bridge the modality gap. Fourth, the cross-modal distillation loss function is designed to transfer knowledge between two different modalities, such as from a teacher model trained on one modality (e.g., vision) to a student model working on another modality (e.g., audio or text) (Gupta et al., 2016; Huo et al., 2024). This loss function typically combines a task-specific loss (e.g., CE loss) with a distillation loss (e.g., KL Divergence) that helps align the student's predictions or feature representations with those of the teacher. Finally, the student model is trained using both the distillation loss and any available supervised loss on the target modality, enabling it to leverage the teacher's knowledge despite the modality differences.

Xue et al. (2022) introduced the Modality Focusing Hypothesis (MFH) and Modality Venn Diagram (MVD) to understand and improve cross-modal distillation. They proposed adjusting the teacher network to focus on modality-general features, significantly improving the student model's performance by preserving relevant information during cross-modal transfer. Sarkar and Etemad (2022) developed XKD, a self-supervised framework for learning transferable features across video modalities. Their method combines Masked Data Modeling with Knowledge Distillation and Domain Alignment, using attention maps and maximum mean discrepancy loss to align features between audio and visual streams. This approach improves the generalization and transferability of learned representations in various video-related tasks. Xia et al. (2023) addressed challenges in cross-modal knowledge transfer for unconstrained videos. They proposed the Modality Noise Filter (MNF) to remove irrelevant modality-specific noise and the Contrastive Semantic Calibration (CSC) module to align visual and audio features based on semantic relevance. These innovations ensure more effective knowledge transfer in noisy, unconstrained video environments.

To demonstrate this method in action, the study by Yang and Xu (2021) proposes a cross-modality knowledge distillation (CMKD) method for multi-modal aerial view object classification, aiming to improve the performance of object classification models using both synthetic aperture radar (SAR) and electro-optical (EO) images. The study uses two different network structures, CMKD-s and CMKD-m, to

transfer knowledge between SAR and EO modalities. CMKD-s employs online KD to transfer information between the two sensors, enhancing the robustness of the aerial view object classification model. CMKD-m further improves upon this by introducing a semi-supervised enhanced training approach, enabling mutual knowledge transfer between the models. Both CMKD-s and CMKD-m were evaluated on the NTIRE2021 SAR-EO challenge dataset, achieving higher accuracy compared to a baseline method without knowledge transfer. The study demonstrates that leveraging KD techniques across different imaging modalities can significantly improve the accuracy and robustness of object classification models, particularly in challenging scenarios where a single sensor may not capture sufficient information.

To recapitulate, cross-modal KD enriches the student model's versatility across modalities, though alignment of feature spaces can be complex and may require handling modality-specific noise.

4.1.2. Graph-based distillation

Graph-based knowledge distillation focuses on transferring knowledge from teacher models to student models specifically designed for graph-structured data. The graph structure allows for the preservation of high-order dependencies and structural information that might be lost in conventional distillation techniques (Zhang & Peng, 2018).

A typical graph-based KD architecture consists of a teacher and a student graph neural network (GNN). The teacher model is responsible for capturing detailed structural patterns and relational information from the graph data. The student model aims to replicate the teacher's performance by learning from its representations. According to Liu et al. (2023), this is achieved by aligning the embeddings and outputs of the student GNN with those of the teacher through specialized distillation loss functions that consider the graph topology and node relationships. This architecture enables the student model to effectively inherit the teacher's ability to understand complex graph structures while being more efficient in computation.

The graph-based distillation loss (Liu et al., 2023) encourages the student network to learn the relational structure (graph) between data points as captured by the teacher. A common approach is to minimize the squared difference (Frobenius norm) between the adjacency matrices of the teacher and student graphs:

$$L_{Graph} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \| A_T(i,j) - A_S(i,j) \|_2^2 \tag{3.1.2.1}$$

Where:

N is the number of data points.

$A_T(i,j)$ and $A_S(i,j)$ represent the strength of the relationship between the i -th and j -th data points in the teacher's and student's graphs, respectively.

$\| \cdot \|_2^2$ denotes the squared Euclidean (L_2) norm between the entries of the teacher's and student's adjacency matrices.

The total KD loss is a combination of the graph-based KD loss and the task loss (typical CE loss), expressed as:

$$L_{KD} = \alpha L_{CE} + (1-\alpha) L_{Graph} \tag{3.1.2.2}$$

Where α is a hyperparameter that balances the contributions of the graph-based KD and the CE task loss.

Yang et al. (2020) introduced the Local Structure Preserving (LSP) module for Graph Convolutional Networks (GCNs). This module transfers topological knowledge from teacher to student GCNs by preserving local graph structures, ensuring the student retains the input graph's topological relationships. The method adapts to dynamic graphs and shows strong performance in node classification and 3D object recognition tasks. Liu et al. (2022a) developed HIRE, a framework for distilling knowledge from heterogeneous graphs. HIRE combines Node-level Knowledge Distillation (NKD) and Relation-level Knowledge Distillation (RKD) to transfer soft labels and capture high-order semantic relations between different node types. This approach enhances the performance of heterogeneous graph neural networks (HGNNs) across multiple datasets. Wang et al. (2022a) proposed Collaborative Knowledge Distillation (CKD) for embedding nodes in heterogeneous information networks (HINs). CKD uses a three-stage approach: Semantic Context Subgraph Sampling, Heterogeneous Knowledge Modeling, and Collaborative Knowledge Distillation. This method effectively handles large datasets and complex HINs, achieving superior performance in node classification and link prediction tasks.

Overall, graph-based methods excel in tasks that require an understanding of structural dependencies but, similar to relation-based KD, this technique may increase computational demands due to the complexity of graph data.

4.1.3. Attention-Based Distillation

Attention-based distillation leverages attention mechanisms to enhance the process of transferring knowledge from a teacher model to a student model by focusing on the most relevant features or representations. Unlike traditional methods that manually select feature alignments between teacher and student models, attention-based distillation automatically identifies and emphasizes important regions of the model's outputs or intermediate layers (Zagoruyko & Komodakis, 2016). This approach improves the efficiency and effectiveness of knowledge transfer by enabling the student model to focus on the most informative aspects of the teacher's learned representations.

Attention-based distillation employs activation-based attention transfer or gradient-based attention transfer methods (Zagoruyko & Komodakis, 2016). In activation-based attention transfer, the attention maps are computed using the activations of a teacher network. The goal is to train the student network not only to make accurate predictions but also to produce attention maps that closely resemble those of the teacher. This helps the student network mimic the teacher's focus on specific spatial regions of the input. While in gradient-based attention transfer, attention is encoded by the sensitivity of the model's output prediction with respect to the input spatial locations. This approach evaluates how changes in specific input locations (e.g., pixels) affect the output, implying that the network is "paying attention" to those locations.

The activation-based attention transfer loss can be formulated as:

$$L_{activation} = \frac{1}{N} \sum_{i=1}^{N} \| f(A_T^i) - f(A_S^i) \|_2^2 \tag{3.1.3.1}$$

Where:

N is the number of spatial locations in the attention map.

A_T^i and A_S^i are the attention values at location i for the teacher and student networks, respectively.

$f(x)$ is the attention mapping function.

$\| \cdot \|_2^2$ denotes the squared Euclidean (L_2) norm between the teacher's and student's attention maps.

The gradient-based attention transfer loss can be formulated as:

$$L_{gradient} = \frac{1}{N} \sum_{i=1}^{N} \| g_T(x_i) - g_S(x_i) \|_2^2 \tag{3.1.3.2}$$

Where:

$g_T(x_i)$ and $g_S(x_i)$ the gradient-based attention values at input location x_i for the teacher and student networks, respectively.

Guo et al. (2023) instituted a class attention transfer method where they proposed a novel attention-based distillation framework that focuses on class-specific features. In their architecture, attention maps are generated for each class, and the student model is trained to mimic these class-wise attention maps from the teacher model. This class-wise attention alignment allows the student to capture more discriminative features relevant to each class, leading to improved performance in classification tasks. By concentrating on the class-specific attention, the student model benefits from a more targeted knowledge transfer, which

enhances its ability to generalize and recognize intricate patterns associated with each category.

Ji et al. (2021a) introduced Attention-based Feature Distillation (AFD), which uses an attention mechanism to identify similarities between teacher and student networks. AFD automatically controls the intensity of the distillation process based on feature similarity, eliminating the need for manually selected links and ensuring the student focuses on the most relevant teacher features. Passban et al. (2020) proposed ALP-KD, an attention-based layer projection technique addressing the skip and search problems in intermediate layer distillation. ALP-KD uses attention to project information from all teacher layers to the student, weighted by relevance, ensuring the student learns from all teacher layers even when there's a mismatch in layer depth. Wu et al. (2021b) developed Universal-KD, an attention-based output-grounded intermediate layer distillation method. It uses attention to assign weights to pseudo classifiers attached to intermediate layers, providing interpretable output space probabilities. This approach addresses the capacity gap between different-sized models and supports cross-architecture distillation.

To showcase how this method functions in practical scenarios, a study by López-Cifuentes et al. (2023) utilizes attention-based KD to enhance the performance of a smaller student network in scene recognition by transferring knowledge from a larger teacher network. This is achieved by matching the activation maps, which represent attention to relevant image regions, between the teacher and student networks. Instead of relying on pixel-wise comparisons using the $\ell 2$ norm, this study introduces a novel DCT-based metric to compare the activation maps in the frequency domain. This method aims to better capture global image cues and spatial relationships that are crucial for scene recognition, enabling the student network to learn more descriptive features and achieve higher performance. By minimizing the differences between the transformed activation maps, the student network learns to focus on the same relevant image areas as the teacher network, improving its ability to recognize complex scenes.

In summary, attention-based KD enhances transfer efficiency and interpretability but requires careful tuning to ensure attention is focused on the most informative features.

4.2. Data-Free and Synthetic Data Distillation

This section focuses on distillation methods that operate without access to the original training data. Data-Free Distillation and Adversarial Distillation generate synthetic data, often using techniques like GANs or model inversion, to enable the student model to learn from the teacher. By creating artificial data that approximates the original data distribution, these methods facilitate effective knowledge transfer while addressing privacy or data availability concerns.

4.2.1. Data-Free Distillation

Data-free knowledge distillation (DFKD) is a technique that enables the transfer of knowledge from a teacher model to a student model without requiring access to the original training data (Chen et al., 2019b). This approach is especially valuable in scenarios where the data is proprietary, sensitive, or unavailable due to privacy concerns (Lopes, 2017). Data-free distillation techniques typically rely on synthetic data generation, where the student model learns from examples created by the teacher model or a generator, rather than directly from the original dataset. However, the absence of real data introduces unique challenges, such as generating high-quality, informative synthetic samples that effectively capture the underlying data distribution and ensuring efficient training without the use of ground-truth data (Chen et al., 2019b).

A typical data-free distillation process involves generating synthetic data using information extracted from the teacher model to train the student model. In this approach, 'meta-data' is transferred from the teacher's batch normalization layers to reconstruct the data distribution. Specifically, synthetic samples are generated by optimizing random noise inputs so that their activations match those recorded from the teacher model (Lopes et al., 2017). The student model is then trained on these synthetic samples using a combination of loss functions: cross-entropy loss measures how well the student predicts target labels; activation-based distillation loss aligns internal representations between the student and teacher networks; and information entropy loss minimizes uncertainty in the student's predictions by pushing the softmax outputs toward one-hot vectors (Lopes et al., 2017). This framework allows the student to learn effectively from the teacher without accessing the original training data.

The data-free distillation the loss function combines three components: cross-entropy loss, activation-based distillation loss, and information entropy loss. While the cross-entropy loss measures how well the student model can predict the target labels, and the activation-based distillation loss helps align the internal representations (activations) between the student and teacher networks, the information entropy loss is used to minimize uncertainty in its predictions by pushing the softmax outputs toward one-hot vectors. The information entropy loss is given by:

$$L_{information} = -\frac{1}{N}\sum_{i=1}^{N} p_S^i \log\left(p_S^i\right) \tag{3.2.1.1}$$

By combining equations (1.1.4), (3.1.3.1), and (3.2.1.1), the total loss for data-free distillation can be expressed as:

$$L_{KD} = \alpha L_{CE} + \beta L_{activation} + \gamma L_{information} \tag{3.2.1.2}$$

Where α , β , γ are hyperparameters that balance the importance of the cross-entropy loss, activation-based distillation loss, and information entropy loss, respectively.

Zhu et al. (2021) introduced FEDGEN for heterogeneous federated learning. This method uses a lightweight generator to aggregate user information without external data, improving model generalization and reducing communication rounds in decentralized environments where data privacy and heterogeneity are critical. Chawla et al. (2021) developed DIODE (DeepInversion for Object Detection) for data-free distillation in object detection tasks. DIODE synthesizes high-quality images by inverting a pre-trained object detection network, employing differentiable augmentations and a novel bounding box sampling strategy. This approach outperforms proxy datasets in distillation efficacy when real data is inaccessible. Fang et al. (2022) proposed FastDFKD, accelerating data-free distillation by reusing common features from training data. A meta-synthesizer learns and reuses shared features across multiple data points, achieving 10 × to 100 × speedup in data synthesis while maintaining competitive performance on various tasks. This method is particularly beneficial for large-scale tasks where traditional DFKD methods are computationally prohibitive.

To illustrate a practical application of data-free distillation, Xiang et al. (2024) proposed a novel technique called Data-Free Knowledge Distillation for Diffusion Models (DKDM) that significantly accelerates diffusion models without requiring access to the original training data. Diffusion models excel at generating realistic images, videos, and audio but suffer from slow inference speeds. DKDM addresses this limitation by distilling the knowledge from a large, pre-trained "teacher" diffusion model into a faster "student" model with a smaller architecture. The key innovation lies in DKDM's ability to achieve this distillation without using the source data. The authors achieve this by synthesizing denoising data from the teacher model and employing a dynamic iterative distillation method to prevent the data generation process from becoming the main bottleneck. Experiments demonstrate that DKDM successfully compresses diffusion models while preserving comparable performance, enabling up to 2x faster models. The paper concludes that DKDM is a significant advancement in the field, allowing for efficient model compression and facilitating wider adoption of diffusion models across diverse applications.

In summary, the process of data-free knowledge distillation involves

utilizing activation statistics to bypass the need for original training data. These statistics capture the specific neurons activated during the training of the teacher model (Lopes et al., 2017). By leveraging these activation statistics and labels from the teacher model, it is possible to recreate a replica of the original training data, which can then be used to train the student model.

4.2.2. Adversarial distillation

Adversarial knowledge distillation (AKD) is a technique that combines the principles of adversarial training and knowledge distillation to enhance the performance and robustness of student models. In AKD, the student model is trained not only to mimic the teacher model's outputs but also to resist adversarial attacks, ensuring that the student model retains essential knowledge while improving its generalization capabilities (Goodfellow et al., 2014).

AKD employs three primary strategies to enhance the knowledge transfer process between teacher and student models. First, it utilizes an adversarial generator to create synthetic data, which either serves as the main training dataset or augments existing data. Second, AKD incorporates one or more discriminators that distinguish between outputs from the student and teacher models, using either logits or features, and leveraging unlabeled data to facilitate knowledge transfer. Finally, AKD implements a joint optimization approach where both the teacher and student models are simultaneously refined in each training iteration. These methods collectively harness adversarial techniques to boost the efficiency of knowledge distillation, ultimately leading to improved performance in the student model while maintaining a compact architecture (Gou et al., 2021).

The AKD loss function typically consists of two components: the discriminator loss and the adversarial loss. These losses are optimized separately, with the discriminator loss training the discriminator to differentiate between the teacher and student outputs, and the adversarial loss training the student to fool the discriminator by mimicking the teacher's outputs.

Since the discriminator is trained to maximize its ability to differentiate between the teacher's output and the student's output, the discriminator's loss can be formulated as:

$$L_{Discriminator} = -E_{x\sim P_{data}}[\log(D(T(x)))] - E_{x\sim P_{data}}[\log(1-D(S(x)))] \quad (3.2.2.1)$$

Where:

E_x are the expectation value of the input data points.

P_{data} refers to the data distribution from which the training data or inputs are sampled.

$x \sim P_{data}$ means that x is a random variable sampled from the probability distribution P_{data}, which is typically the distribution.

$D(T(x))$ is the probability of the discriminator's output for the teacher's output.

$D(S(x))$ is the probability of the discriminator's output for the student's output.

While the discriminator tries to maximize its loss, student network is trained to minimize its adversarial loss by generating outputs that can "fool" the discriminator into classifying them as if they were produced by the teacher. The adversarial loss function can be formulated as:

$$L_{Adversarial} = -E_{x\sim P_{data}}[\log(D(S(x)))] \quad (3.2.2.2)$$

One of the main challenges of AKD lies in balancing the adversarial and distillation losses, as improper tuning can lead to either reduced accuracy or vulnerability to adversarial examples (Wang et al., 2018).

To enhance the transfer of complex knowledge structures between teacher and student models, Lee et al. (2022) introduced a method using Graph Convolutional Networks (GCNs) to preserve similarity relationships between data samples during distillation. By combining traditional distillation loss with adversarial loss, this approach improves accuracy and maintains relational integrity in the student model. Higuchi et al. (2022) proposed a framework that focuses on both final outputs and intermediate representations in CNNs. This method combines adversarial training for robustness with knowledge distillation, aligning intermediate feature representations between teacher and student models to enhance resilience against adversarial attacks. He et al. (2022) developed a technique for compressing Graph Neural Networks (GNNs) using adversarial knowledge distillation. Their student-teacher framework challenges the student GNN to match the teacher's outputs while preserving essential graph structural information, resulting in efficient, smaller GNNs that maintain performance on benchmark datasets.

For a concrete example of this approach in action, Bai et al. (2022) use adversarial knowledge distillation to improve the performance of BioBERT, a pre-trained language model, on a biomedical factoid question-answering task. The goal is to enhance the model's ability to identify the exact answer to a question within a given passage of text. This is achieved by using a teacher-student framework. The teacher model is also based on BioBERT but includes an additional module designed to capture question-answer interaction knowledge. This knowledge is then distilled to the student model, which is a basic BioBERT model. To further enhance the student model's robustness and prevent overfitting, adversarial training is incorporated into the distillation process. This involves constructing perturbed training examples by adding noise to the original training data. By forcing the student model to mimic the teacher model's predictions on both original and perturbed examples, the student model can learn the knowledge of question-answer interaction and improve its performance on the question-answering task.

In essence, adversarial knowledge distillation improves generalization and security but requires careful balancing of adversarial and distillation losses, which can be challenging to achieve without impacting accuracy.

4.3. Model compression and architecture optimization

This section encompasses distillation techniques aimed at producing efficient student models with reduced size and computational requirements. Quantized distillation combines quantization with distillation to create smaller, faster models suitable for deployment on resource-constrained devices. NAS-based distillation integrates Neural Architecture Search (NAS) to automatically find the optimal student model architecture, balancing performance and efficiency during the distillation process.

4.3.1. Quantized Distillation

Quantized knowledge distillation (QKD) focuses on improving the performance of quantized neural networks (QNNs) by leveraging knowledge distillation techniques (Kim et al., 2019). QNNs, which reduce the precision of weights and activations to lower-bit representations, are designed for deployment on resource-constrained devices like mobile phones and embedded systems. Although quantization significantly reduces memory usage and increases computational efficiency, it often leads to performance degradation compared to full-precision models. Quantized distillation addresses this problem by using knowledge from a full-precision teacher model to guide the training of a quantized student model, helping the student overcome the accuracy loss that typically accompanies quantization (Kim et al., 2019).

The total loss function for QKD combines the CE task loss, the KL divergence distillation loss, and the quantization-aware loss. The quantization-aware loss minimizes the difference between the quantized and full-precision student activations or weights by penalizing the quantization error between the quantized student weights W_q and the full-precision student weights W_f:

$$L_{Quantized} = \| W_f - W_q \|_2^2 \quad (3.3.1.1)$$

Where W_f and W_q represent the full-precision weights and the

quantized weights, respectively.

The total loss for quantized distillation can be expressed as:

$$L_{KD} = \alpha L_{CE} + \beta L_{KL} + \gamma L_{Quantized} \quad (3.3.1.2)$$

Where α, β, γ are hyperparameters that balance the importance of the CE task loss, the KL divergence distillation loss, and the quantization-aware loss, respectively.

A typical quantized KD architecture involves a full-precision teacher model and a quantized student model, where the teacher provides soft targets and activation maps to guide the student in learning to replicate the teacher's outputs despite the reduced precision of its own weights and activations. Kim et al. (2019) proposed QKD, a three-phase method for improving quantized neural networks. It involves "self-studying" for good initialization, "co-studying" to make the teacher more quantization-friendly, and "tutoring" to transfer knowledge efficiently. The architecture incorporates quantization operations into the forward and backward passes of the student model, allowing it to learn quantization effects during training. The loss function combines the standard classification loss with a distillation loss that measures the discrepancy between the teacher's and student's outputs. By integrating quantization processes into the knowledge distillation framework, the student model can achieve higher accuracy compared to traditional quantization methods, effectively narrowing the performance gap between quantized and full-precision models. This comprehensive approach significantly improves the performance of quantized models, making them comparable to full-precision models while maintaining reduced computational demands.

Zhao and Zhao (2024) introduced a self-supervised quantization-aware knowledge distillation (SQAKD) method. This approach combines self-supervised learning with SQAKD, using proxy tasks and unlabeled data to train models. It improves the efficiency and accuracy of quantized models while reducing reliance on labeled datasets, making it valuable when labeled data is scarce. Boo et al. (2021) developed the Stochastic Precision Ensemble (SPEQ) technique. SPE assigns random precision levels during training, creating an ensemble of models with varying precisions. Knowledge from this ensemble is distilled into a single quantized model, enabling it to learn from diverse precision scenarios. This self-knowledge distillation approach mitigates accuracy loss in quantized models, resulting in robust and efficient models for edge devices.

Ultimately, QKD provides an efficient model for low-resource environments, though precision reduction can introduce noise, impacting performance if not managed carefully.

4.3.2. NAS-Based Distillation

Neural Architecture Search (NAS) is a powerful method for automatically discovering optimal neural network architectures tailored to across a large variety of artificial intelligence tasks (Liu et al., 2019a). When combined with knowledge distillation, NAS-based distillation aims to find the most effective student architectures for learning from a pre-trained teacher model. The goal is to improve student model performance while maintaining efficiency, making NAS-based distillation particularly useful for creating smaller, faster models suited for deployment on resource-constrained devices.

A typical NAS-based distillation process involves an iterative search that explores a vast space of potential student models, guided by knowledge distillation from a robust teacher model (Nath et al., 2024). Initially, a weight-sharing super-network is constructed, encompassing all candidate architectures within a predefined search space. This over-parameterized network includes distinct paths for each possible architecture, with shared weights optimized through gradient descent during training. Throughout the search phase, each candidate architecture (or path) is evaluated based on its performance when distilled from the teacher model. The student models not only learn from the teacher's soft labels but may also align their intermediate feature representations with those of the teacher, enhancing both accuracy and robustness. Once the optimal architecture is identified, balancing performance metrics like accuracy, model size, and computational efficiency, it is trained from scratch using conventional methods and the distilled knowledge from the teacher.

However, traditional NAS methods are often computationally expensive for optimizing the neural architecture for the target task and can struggle to generalize small datasets (Elsken et al., 2020). Moreover, a significant drawback is that most NAS approaches concentrate mainly on optimizing accuracy, model size, or computational efficiency (FLOPs), while paying insufficient attention to the robustness of the searched architectures against adversarial attacks. This lack of focus on adversarial robustness is crucial because it affects the security and resilience of machine learning systems in real-world applications (Nath et al., 2024)

The total loss function for NAS-based distillation combines the CE task loss, the KL divergence distillation loss, and the architecture search loss (Liu et al., 2019b). The architecture search loss optimizes the architecture of the student network, which normally is formulated as:

$$L_{Architecture} = E_{\alpha \sim A}[L_{val}(w(\alpha), \alpha)] \quad (3.3.2.1)$$

Where:

$E_{\alpha \sim A}$ is the expectation over architectures α sampled from the architecture search space A.

$w(\alpha)$ is the weight of the student network given an architecture α, which is optimized with respect to the task-specific loss on the training set.

L_{val} is the validation loss, used to guide architecture optimization based on how well the architecture performs on the validation set.

The total loss for NAS-based distillation can be expressed as:

$$L_{KD} = \alpha L_{CE} + \beta L_{KL} + \gamma L_{Architecture} \quad (3.3.2.2)$$

Where α, β, γ are hyperparameters that balance the importance of the CE task loss, the KL divergence distillation loss, and the architecture search loss, respectively.

Lee et al. (2023) introduced Distillation-aware Student Search (DaSS), a method that predicts student architecture performance without extensive training on target tasks. DaSS incorporates a distillation-aware task encoding that adapts to the teacher model's accuracy, enabling efficient and scalable architecture search. By leveraging meta-prediction and gradient-based adaptation, DaSS generalizes across different datasets and teacher models, reducing computational costs while improving the quality of discovered student models. Trivedi et al. (2023) proposed KD-NAS, a system combining NAS with knowledge distillation to optimize student architectures for multilingual language models. KD-NAS uses a NAS controller to predict rewards based on distillation loss and inference latency, selecting top candidate architectures for knowledge transfer. It also introduces a multi-layer hidden state distillation method, allowing students to learn from multiple teacher layers without requiring pre-training. KD-NAS demonstrates significant improvements in speed and efficiency while maintaining strong performance in multilingual tasks.

In essence, NAS for knowledge distillation involves searching within a predefined space of student architectures (Li et al., 2020). During this process, a pre-trained teacher model transfers its knowledge to a candidate student model. After the search is completed, the student model that best meets the desired objectives, such as performance and computational efficiency, is selected.

Ultimately, NAS-based distillation is powerful for designing efficient models but can be computationally expensive due to the architecture search process and may overlook robustness to adversarial attacks, which is essential for secure applications.

4.4. Ensemble Distillation Methods

This section provides an overview of several ensemble distillation methods, including lifelong distillation and multi-teacher distillation.

Lifelong distillation applies knowledge distillation within a continual learning framework, enabling the student model to incrementally learn new tasks over time without forgetting previous knowledge. Multi-teacher distillation involves distilling knowledge from multiple teacher models into a single student model, aggregating expertise from various sources to improve generalization and performance.

4.4.1. Lifelong Distillation

Lifelong distillation focuses on enabling machine learning models to learn continuously from a sequence of tasks while retaining knowledge from previously learned tasks (Wang et al., 2022b). This approach addresses the challenge of "catastrophic forgetting," where a model forgets previously learned information when trained on new tasks. Lifelong knowledge distillation allows models to retain essential knowledge through teacher-student frameworks, where the teacher model guides the student to retain old knowledge while learning new tasks (Wang et al., 2022b). This method is particularly useful for applications that require continuous learning and adaptation, such as autonomous systems, generative models, and image processing tasks.

Lifelong distillation typically involves a dynamic teacher-student framework, in which the teacher model retains knowledge from previous tasks, while the student model focuses on learning new tasks while preserving previously acquired knowledge. In this architecture, the teacher model can be an earlier snapshot of the student, or an ensemble of models trained on previous tasks. The student model learns from new data and receives guidance from the teacher through knowledge distillation techniques that minimize discrepancies between their outputs (Hong et al., 2020). This setup ensures that the student model not only acquires new knowledge but also retains proficiency in previously learned tasks. However, a drawback of lifelong KD is that the "previous knowledge" might not be easily applicable or reusable for future tasks due to differences in task domains (Hong et al., 2020). Therefore, carefully evaluating which previous knowledge can be reused for new target domains is essential.

The total loss function for Lifelong distillation combines the CE task loss, the KL divergence distillation loss, and the knowledge retain loss. The knowledge retention loss prevents catastrophic forgetting of previous tasks by regularizing the model's parameters, such as Elastic Weight Consolidation (EWC)(Kirkpatrick et al., 2017), which adds a penalty term to restrict significant changes in important weights:

$$L_{Retain} = \lambda \sum_{i} \Omega_i (\theta_i - \theta_{i-1})^2 \quad (3.4.1.1)$$

Where:

θ_i are the current model parameters.

θ_{i-1} are the parameters learned from previous task.

Ω_i are importance weights (indicating how crucial each parameter is for the previous task).

λ is a hyperparameter that controls the strength of this regularization.

The total loss for lifelong distillation can be expressed as:

$$L_{KD} = \alpha L_{CE} + \beta L_{KL} + \gamma L_{Retain} \quad (3.4.1.2)$$

Where α, β, γ are hyperparameters that balance the importance of the CE task loss, the KL divergence distillation loss, and the knowledge retain loss, respectively.

Soltoggio et al. (2024) introduced the Shared Experience Lifelong Learning (ShELL) framework for distributed AI systems. ShELL enables AI agents to learn and share knowledge continuously at the network edge, integrating knowledge-sharing mechanisms across agents. This approach allows for collaboration in decentralized environments and augments advanced AI models with lifelong learning capabilities. Ye and Bors (2021) proposed Lifelong Twin GANs (LT-GANs), a dual-GAN system for retaining knowledge across tasks. LT-GANs use a Teacher-Assistant mechanism where roles alternate after learning new tasks, ensuring previously learned tasks are not forgotten. This method demonstrates effectiveness in image generation tasks, balancing new learning with knowledge retention. Zhai et al. (2021) developed Hyper-LifelongGAN, a scalable architecture for lifelong learning in image-conditioned generation tasks. It decomposes filters into dynamic base filters and a shared weight matrix, using task-specific coefficients to adjust for each task. This approach leverages knowledge distillation to retain previous task knowledge, addressing catastrophic forgetting and scalability issues efficiently.

As a practical application, the study by Li et al. (2021) on continual multi-task image restoration uses lifelong knowledge distillation on both the generators and discriminators in its CycleGAN framework. The goal is to alleviate the catastrophic forgetting problem, where a model forgets previously learned tasks when learning new ones. The distillation process involves constraining the current task's generator and discriminator outputs to be similar to those of the previous task's models. This ensures that knowledge acquired from past tasks is retained while learning a new task. Specifically, the generators for the current task are constrained to produce outputs similar to the previous task's generators, ensuring the transfer of learned restoration techniques. Similarly, the discriminators for the current task are trained to have consistent outputs with those of the previous task, promoting the preservation of learned image quality evaluation. This approach enables the model to continually learn and adapt to new image restoration tasks, such as low-light enhancement, deblurring, and denoising, without losing proficiency in previously acquired skills.

To recapitulate, lifelong KD enables continuous learning by allowing models to acquire new information while retaining previous knowledge. However, a challenge is that prior knowledge may not always be applicable to future tasks due to differences in task domains. Additionally, balancing new and old knowledge without excessive retention loss is crucial for optimal performance across tasks, as excessive focus on preserving past knowledge can hinder the learning of new tasks.

4.4.2. Multi-teacher Distillation

Multi-teacher knowledge distillation involves leveraging the knowledge of multiple teacher models to guide the learning of a single student model (You et al., 2017). This approach can enhance the performance of the student by combining diverse knowledge sources, but it also introduces challenges related to managing the variability among the teachers and ensuring effective knowledge aggregation. One major advantage of multi-teacher distillation is that it can improve generalization by exposing the student to a wider range of information, but aligning the teachers' outputs and managing their contributions can be computationally expensive and complex (You et al., 2017).

A typical multi-teacher distillation framework utilizes softened outputs, intermediate feature-based distillation, or a combination of both to guide the student model. In the approach proposed by Pham et al. (2022b), multiple quantized teacher models with varying bit-widths collaborate to transfer knowledge. The contributions of each teacher are balanced according to its individual capacity. Each teacher's intermediate layer outputs are fused into a shared feature map, weighted by an importance factor that reflects the teacher's influence on the collective knowledge. This collaborative learning approach encourages teachers with varying bit-widths to contribute optimally to the shared knowledge, forming a robust representation that the student can mimic. The student learns from both the ensemble logits of the teachers (via softened outputs and cross-entropy loss) and the shared feature maps at selected layers using distance-based losses, such as attention loss or FitNet hint loss, to align the student's intermediate representations with those of the teachers.

With single-teacher distillation, the student network learns from the softened output of the teacher network. When there are multiple teachers, the student network learns from the average of the softened outputs of the multiple teacher networks. The loss function for multi-teacher knowledge distillation can be formulated as:

$$L_{MTKD} = \alpha L_{CE} + (1-\alpha)\cdot\frac{1}{M}\sum_{i=1}^{M} L_{KL} \quad (3.4.2.1)$$

Where:

L_{CE} is the cross-entropy loss, or the standard task loss, which minimizes the error between the student's soft logits and the ground truth labels.

L_{KL} represents Kullback-Leibler divergence, or the distillation loss, minimizes the difference between the soft logits of the i-th teacher and the student.

M is the number of teachers.

α is a balancing hyperparameter that controls the trade-off between the cross-entropy loss and Kullback-Leibler divergence.

Wu et al. (2021a) proposed a framework for distilling knowledge from multiple pre-trained language models (MT-BERT) by aligning their hidden states. They introduced multi-teacher hidden loss and multi-teacher distillation loss, which align intermediate representations and weigh teacher contributions based on confidence, effectively handling inconsistencies in pre-trained language model feature spaces. Liu et al. (2020) presented an adaptive framework that integrates knowledge from multiple teacher networks by assigning weights based on teacher confidence. They introduced latent representations and a multi-group hint strategy to capture hidden representations and facilitate intermediate-level knowledge transfer, ensuring more effective learning from reliable teachers.

To illustrate a real-world application, the study by Tang and Liu (2024) uses a multi-teacher KD approach to detect financial fraud. The goal is to improve detection accuracy, inference speed, and generalization ability when dealing with imbalanced datasets and industry-specific challenges. The process starts by training multiple teacher models, each specialized in detecting fraud within a specific industry. These models are then used to transfer their knowledge to a single, more compact student model. This distillation process leverages both hard targets (true labels) and soft targets (probability distributions from the teachers) to guide the student's learning. By incorporating insights from multiple expert models, the student model becomes capable of detecting fraud across different industries while maintaining a smaller size for faster inference. This distributed approach addresses the limitations of traditional methods that struggle with industry-specific variations and large, complex models.

In summary, multi-teacher distillation improves model robustness by integrating diverse knowledge but can be computationally intensive due to the need to manage multiple teacher contributions. The process of multi-teacher distillation typically involves pretraining multiple teacher models and aggregating their outputs during the training of the student model (Amirkhani et al., 2021). These aggregated outputs serve as soft labels, which are used alongside the true labels to train the student model.

5. Teacher–student architecture variations

Knowledge distillation involves transferring knowledge from a large, complex teacher model to a smaller, more efficient student model (Hilton et al., 2015). This technique traditionally assumes that the teacher and student share similar architectures. However, in real-world applications, models with different architectures may need to be deployed due to varying computational constraints or specific application requirements.

Liu et al. (2022b) propose a method for cross-architecture knowledge distillation (KD) using two projectors, partially cross-attention (PCA) and group-wise linear (GL), to align teacher-student feature representations. This improves student model performance, even with simpler architectures, outperforming traditional KD across multiple benchmarks. Hao et al. (2024) introduce the One-for-All (OFA-KD) framework to transfer knowledge between heterogeneous architectures by projecting intermediate features into a unified latent space. OFA-KD enhances knowledge transfer with adaptive target enhancement, significantly improving student model performance, especially with dissimilar teacher-student architectures. Dong et al. (2023) present DisWOT, a zero-training method for student architecture search in KD. DisWOT uses similarity metrics and evolutionary algorithms to predict optimal student architectures without training, reducing computational costs while achieving competitive performance.

In summary, recent advancements in knowledge distillation techniques have broadened the scope of model compatibility, enabling effective knowledge transfer between heterogeneous architectures. These developments indicate a significant progression in the applicability of knowledge distillation, making it a viable approach for diverse deployment scenarios.

6. Applications

Knowledge distillation techniques have found extensive applications across multiple domains, enabling the deployment of efficient and compact models in areas where computational and memory resources are limited. From computer vision and natural language processing to the medical field and autonomous systems, KD plays a critical role in achieving high-performance models suitable for real-world constraints. This section explores the use of KD in various applications, discussing how different techniques address specific needs in each field.

To aid in understanding the trade-offs among KD techniques, Table 4 provides a comparative overview based on key metrics such as accuracy boost, resource efficiency, computational complexity, and ideal use cases. This table serves as a quick reference for selecting the appropriate KD method depending on the target application's requirements and constraints, helping to weigh the strengths and limitations of each approach.

6.1. KD in vision and language

Knowledge distillation has been applied across various domains, including vision and language (VL), where large models are compressed into smaller, efficient ones while maintaining high performance. VL models integrate visual and textual data for tasks like image captioning and visual question answering but are often computationally intensive, making KD crucial for their deployment in resource-limited settings.

Fang et al. (2021a) propose DistillVLM to address misalignment in visual tokens between teacher and student models. By aligning visual and linguistic features, they achieve significant compression without sacrificing accuracy, making the model suitable for resource-constrained tasks. In another work, Fang et al. (2021b) introduce SEED, a self-supervised distillation method that improves visual representations by guiding the student model during pre-training. SEED enhances performance without labeled data, proving effective in various tasks with limited data. Fang and Yang (2023) discuss cross-modal KD in VL, highlighting techniques like shared embeddings and contrastive learning to improve tasks such as visual question answering and language-driven image generation. Liu et al. (2022c) apply cross-domain KD for deepfake detection, focusing on localized inconsistencies in spatial and frequency domains. Their approach improves detection accuracy across various deepfake styles. Yun et al. (2021) defend KD for task incremental learning in 3D object detection, using KD with a Bayesian approach to prevent catastrophic forgetting, ensuring high accuracy across multiple detection tasks.

In the work on enhancing Alexa's natural language understanding (NLU) system, FitzGerald et al. (2022) at Amazon explore KD to train multi-billion-parameter encoders for language tasks. The Alexa Teacher Model demonstrates the potential of KD in efficiently handling large-scale, real-world NLU applications by pretraining and distilling high-capacity teacher models into smaller, deployable models. This approach enables Alexa to provide fast and accurate responses in resource-constrained voice-activated devices.

Table 4
Comparative analysis of knowledge distillation techniques based on different metrics.

Technique	Accuracy Boost	Resource Efficiency	Computational Complexity	Ideal Use Cases	Key Trade-Offs
Response-Based KD	Moderate	High	Low	General-purpose, NLP tasks	Limited by capacity gap between teacher and student
Feature-Based KD	High	Moderate	Moderate	Vision, complex visual tasks	May struggle with feature alignment across architectures
Relation-Based KD	High	Low	High	Graph-based applications	High computational cost for graph-structured data
Offline Distillation	Moderate-High	High	Moderate	Pre-trained models, resource-limited devices	Inflexible to new data patterns
Online Distillation	High	Moderate	High	Real-time learning, continual updates	High computational cost due to simultaneous training
Self-Distillation	Moderate	High	Low	Mobile deployment	Risk of reinforcing incorrect predictions
Cross-Modal Distillation	Moderate	Moderate	High	Multi-modal tasks (e.g., image-text, audio-video)	Requires advanced alignment for disparate feature spaces
Graph-Based Distillation	High	Low	High	Node classification, social network analysis	Requires complex loss functions for graph integrity
Attention-Based Distillation	Moderate-High	Moderate	High	Vision, NLP with interpretability needs	High computational demands for attention mechanism
Data-Free KD	Moderate	Moderate	High	Privacy-sensitive environments	Quality of synthetic data impacts effectiveness
Adversarial KD	Moderate-High	Low	High	Security-sensitive applications	Balancing adversarial and distillation loss
Quantized KD	Low-Moderate	Very High	Moderate	IoT, edge devices	Accuracy may decrease with reduced bit precision
NAS-Based Distillation	High	High	Very High	Edge devices, compact model deployment	Initial search is computationally demanding
Lifelong Distillation	Moderate	Moderate	High	Dynamic applications (e.g., language learning, autonomous systems)	Must carefully select reusable prior knowledge for new domains
Multi-Teacher Distillation	High	Moderate	High	Multi-modal, diverse information tasks	Alignment of multiple teachers can be resource-intensive

McDonald et al. (2024) document the use of KD by Mistral AI to reduce hallucination rates in LLMs, a common challenge in generative models. By distilling knowledge from teacher to student models, they improve the accuracy and reliability of AI-generated content across benchmarks like the MMLU. This distillation approach results in smaller, more efficient LLMs that are better suited for deployment in environments requiring controlled output reliability.

In 3D computer vision, Chan et al. (2022) leverage KD to enhance the performance of geometry-aware GANs for applications requiring realistic 3D image synthesis. Their work on Efficient Geometry-Aware 3D GANs focuses on refining spatial and feature representation, reducing the computational cost of rendering realistic 3D visuals. KD facilitates model compression, allowing high-quality 3D generation to be deployed on resource-limited devices, making it suitable for applications like AR/VR.

6.2. KD in medical applications

Knowledge distillation has become crucial in the medical field, where efficient and deployable models are essential for tasks like disease detection, medical image segmentation, and patient monitoring. KD allows for the creation of lightweight models that maintain the performance of larger ones, making it ideal for resource-constrained medical environments.

In medical imaging, Generative Adversarial Networks (GANs) play a transformative role in handling various tasks, including segmentation, reconstruction, detection, classification, augmentation, registration, and image synthesis (Islam et al., 2024). With the capacity to generate high-quality synthetic images even with limited datasets, GANs have notably improved diagnostic precision and image quality enhancement. According to Alamir and Alghamdi (2023), GANs contribute to a new level of diagnostic capability by producing realistic medical images that aid in training and validating other diagnostic models. GANs are especially impactful in the segmentation of critical organs; Makhlouf et al. (2023) highlight that brain, chest, breast, and lung images are among the most frequently segmented areas, reflecting the high demand for accurate organ-specific image processing in diagnostics.

Zou et al. (2021) introduce Coco DistillNet for pathological gastric cancer segmentation. This approach improves segmentation accuracy by transferring knowledge from a teacher model to a lightweight student model through cross-layer correlation distillation. It ensures better visual feature understanding, making it suitable for clinical settings with limited computational resources. Li et al. (2022b) combine self-supervised learning with self-knowledge distillation for COVID-19 detection from chest X-rays. This method helps overcome the scarcity of labeled data by allowing the student model to refine its representations, achieving high detection accuracy with minimal labeled data. Gordienko et al. (2023) propose an ensemble KD framework for edge intelligence in medical applications. Multiple teacher models distill knowledge into a compact student model, enhancing its performance and generalization. This framework is particularly useful for real-time, low-latency tasks on edge devices like portable diagnostic tools.

6.3. KD in Other Applications

Knowledge distillation extends beyond traditional fields like computer vision and NLP, proving useful in areas such as autonomous systems, neural architecture search (NAS), and industrial anomaly detection. KD helps reduce model complexity, enhance performance, and enable deployment in resource-limited environments.

Kargin and Petrenko (2023) introduce KD for autonomous unmanned systems like UAVs and AGVs. Their method improves decision-making by distilling knowledge about environmental factors and event streams, allowing lightweight models to function efficiently in dynamic environments with real-time constraints. Trofimov et al. (2023) combine multi-fidelity NAS with KD to accelerate the search for optimal architectures. By leveraging teacher models for guidance, KD reduces computational costs and improves the efficiency of NAS, discovering high-performing architectures faster. Rakhmonov et al. (2023) propose an Extensive Knowledge Distillation (EKD) model for real-time anomaly detection in industrial settings. EKD transfers knowledge from complex teacher models to lightweight students,

enabling accurate, real-time fault detection, enhancing safety and operational efficiency.

Mora et al. (2022) explore the integration of KD into federated learning (FL) systems. In federated learning, data privacy and communication efficiency are key challenges. The authors present KD as a solution that allows local models to distill their knowledge into a central student model without sharing raw data. This not only improves the generalization of the global model but also reduces communication overhead in FL. Their framework enhances model performance in federated settings, particularly in environments with non-IID data and limited communication resources, making it a valuable technique for distributed machine learning tasks.

7. Conclusion

Knowledge distillation has revolutionized model compression and knowledge transfer in machine learning since its introduction by Hinton et al. (2015). This paper has provided a comprehensive overview of its evolution, exploring various types of knowledge transfer (response--based, feature-based, and relation-based), distillation schemes (offline, online, and self-distillation), and advanced algorithms. The advancements in knowledge distillation have significantly improved the deployment of efficient models on resource-constrained devices while maintaining high performance, enabling sophisticated AI capabilities in diverse real-world scenarios. As the field continues to evolve, future research may focus on improving existing approaches and exploring novel applications. Knowledge distillation remains crucial in bridging the gap between complex state-of-the-art AI models and practical deployment constraints, promising continued innovation in the field of artificial intelligence.

Declaration of competing interest

There is not conflict of interest.

Funding

This work was supported by the Applied Research and Technology Partnership grants (ARTPs) funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant CCB21-2021-00266 and by BFI Energy Group Inc.

Data availability

No data was used for the research described in the article.

References

Ahmad, S., Ullah, Z., & Gwak, J. (2024). *Multi-Teacher Cross-Modal Distillation with Cooperative Deep Supervision Fusion Learning for Unimodal Segmentation, 297*. Knowledge-Based Systems, Article 111854.

Amirkhani, A., Khosravian, A., Masih-Tehrani, M., & Kashiani, H. (2021). Robust semantic segmentation with multi-teacher knowledge distillation. *IEEE Access, 9*, 119049–119066. https://doi.org/10.1109/ACCESS.2021.3107841

Anil, R., Pereyra, G., Passos, A., Ormandi, R., Dahl, G. E., & Hinton, G. E. (2018). Large scale distributed neural network training through online distillation. *arXiv preprint. arXiv:1804.03235*.

Bai, J., Yin, C., Zhang, J., Wang, Y., Dong, Y., Rong, W., & Xiong, Z. (2022). Adversarial knowledge distillation based biomedical factoid question answering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1*. https://doi.org/10.1109/tcbb.2022.3161032

Boo, Y., Shin, S., Choi, J., & Sung, W. (2021). Stochastic precision ensemble: Self-knowledge distillation for quantized deep neural networks. In , *35. Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 6794–6802).

Caruana, R., Buciluǎ, C., & Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 535–541).

Chan, E. R., Lin, C. Z., Chan, M. A., Nagano, K., Pan, B., De Mello, S., … Wetzstein, G. (2022). Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16123–16133).

Chawla, A., Yin, H., Molchanov, P., & Alvarez, J. (2021). Data-free knowledge distillation for object detection | IEEE conference publication | IEEE Xplore. *Data-Free Knowledge Distillation for Object Detection*. https://ieeexplore.ieee.org/document/9423376.

Chen, D., Mei, J. P., Wang, C., Feng, Y., & Chen, C. (2020). Online knowledge distillation with diverse peers. *arXiv.org*. https://arxiv.org/abs/1912.00350.

Chen, G., Choi, W., Yu, X., Han, T., & Chandraker, M. (2017). Learning efficient object detection models with knowledge distillation. *Advances in Neural Information Processing Systems, 30*.

Chen, W. C., & Chu, W. T. (2023). Sssd: Self-supervised self distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2770–2777).

Dong, P., Li, L., & Wei, Z. (2023). Diswot: Student architecture search for distillation without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11898–11908).

Dong, S., Hong, X., Tao, X., Chang, X., Wei, X., & Gong, Y. (2021). Few-shot class-incremental learning via relation knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. https://ojs.aaai.org/index.php/AAAI/article/view/16213.

Elsken, T., Staffler, B., Metzen, J. H., & Hutter, F. (2020). Meta-learning of neural architectures for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12365–12375).

Fang, G., Mo, K., Wang, X., Song, J., Bei, S., Zhang, H., & Song, M. (2022). Up to 100 x faster data-free knowledge distillation. *arXiv.org*. https://arxiv.org/abs/2112.06253.

Fang, Z., Wang, J., Hu, X., Wang, L., Yang, Y., & Liu, Z. (2021a). Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1428–1438).

Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y., & Liu, Z. (2021b). Seed: Self-supervised distillation for visual representation. *arXiv preprint. arXiv:2101.04731*.

Fang, Z., & Yang, Y. (2023). Knowledge distillation across vision and language. *Advancements in Knowledge Distillation: Towards New Horizons of Intelligent Systems* (pp. 65–94). Cham: Springer International Publishing.

FitzGerald, J., Ananthakrishnan, S., Arkoudas, K., Bernardi, D., Bhagia, A., Delli Bovi, C., … Natarajan, P. (2022). Alexa teacher model: Pretraining and distilling multi-billion-parameter encoders for natural language understanding systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 2893–2902).

Ge, Y., Zhang, X., Choi, C. L., Cheung, K. C., Zhao, P., Zhu, F., … Li, H. (2021). Self-distillation with batch knowledge ensembling improves imagenet classification. *arXiv preprint*. arXiv:2104.13298.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Gordienko, Y., Shulha, M., Kochura, Y., Rokovyi, O., Alienin, O., Taran, V., & Stirenko, S. (2023). Ensemble knowledge distillation for edge intelligence in medical applications. *Advancements in Knowledge Distillation: Towards New Horizons of Intelligent Systems* (pp. 135–168). Cham: Springer International Publishing.

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision, 129*(6), 1789–1819.

Gou, J., Chen, Y., Yu, B., Liu, J., Du, L., Wan, S., & Yi, Z. (2024). Reciprocal teacher-student learning via forward and feedback knowledge distillation. *IEEE Transactions on Multimedia, 26*, 7901–7916. https://doi.org/10.1109/TMM.2024.3372833. IEEE Transactions on Multimedia.

Gowda, S. N., Hao, X., Li, G., Gowda, S. N., Jin, X., & Sevilla-Lara, L. (2024). Watt for what: Rethinking deep learning's energy-performance relationship (No. arXiv: 2310.06522). *arXiv*. http://arxiv.org/abs/2310.06522.

Guo, Ziyao, Yan, Haonan, Li, Hui, & Lin, Xiaodong (2023). Class attention transfer based knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11868–11877).

Gupta, S., Hoffman, J., & Malik, J. (2016). Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2827–2836).

Hao, Z., Guo, J., Han, K., Tang, Y., Hu, H., Wang, Y., & Xu, C. (2024). One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *Advances in Neural Information Processing Systems*, 36.

He, H., Wang, J., Zhang, Z., & Wu, F. (2022). Compressing deep graph neural networks via adversarial knowledge distillation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 534–544).

Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., & Choi, J. Y. (2019). A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1921–1930).

Higuchi, H., Suzuki, S., & Shouno, H. (2022). Adversarial training with knowledge distillation considering intermediate representations in CNNs. In *International Conference on Neural Information Processing* (pp. 683–691). Singapore: Springer Nature Singapore.

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

Hong, X., Guan, S. U., Man, K. L., & Wong, P. W. (2020). Lifelong machine learning architecture for classification. *Symmetry, 12*(5), 852.

Huo, F., Xu, W., Guo, J., Wang, H., & Guo, S. (2024). C2KD: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16006–16015).

Islam, S., A, A., Aziz, M. T., Nabil, H. R., Jim, J. R., Kabir, M. M., Mridha, M. F., … Shin, J (2024). Generative adversarial networks (GANs) in medical imaging: advancements, applications and challenges. *IEEE Access*.

Ji, M., Heo, B., & Park, S. (2021a). *Show, Attend and Distill: Knowledge distillation via attention-based feature matching*. Cornell University. *arXiv*.

Ji, M., Shin, S., Hwang, S., Park, G., & Moon, I. C. (2021b). Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10664–10673).

Kargin, A., & Petrenko, T. (2023). Knowledge distillation for autonomous intelligent unmanned system. *Advancements in Knowledge Distillation: Towards New Horizons of Intelligent Systems* (pp. 193–230). Cham: Springer International Publishing.

Kim, H., Kwak, T. Y., Chang, H., Kim, S. W., & Kim, I. (2023). RCKD: response-based cross-task knowledge distillation for pathological image analysis. *Bioengineering (Basel), 10*(11), 1279-. https://doi.org/10.3390/bioengineering10111279

Kim, J., Bhalgat, Y., Lee, J., Patel, C., & Kwak, N. (2019). Qkd: Quantization-aware knowledge distillation. *arXiv preprint.* arXiv:1911.12491.

Kim, K., Ji, B., Yoon, D., & Hwang, S. (2021). Self-knowledge distillation with progressive refinement of targets. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 6547–6556). https://doi.org/10.1109/ICCV48922.2021.00650

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., … Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences, 114*(13), 3521–3526.

Lan, X., Zhu, X., & Gong, S. (2018). Knowledge distillation by on-the-fly native ensemble. *Advances in Neural Information Processing Systems, 31*.

Lee, H., An, S., Kim, M., & Hwang, S. J. (2023). Meta-prediction model for distillation-aware NAS on unseen datasets. *arXiv.org*. https://arxiv.org/abs/2305.16948.

Lee, H., Park, Y., Seo, H., & Kang, M. (2023b). Self-knowledge distillation via dropout. *Computer Vision and Image Understanding, 233*, Article 103720.

Lee, S., Kim, S., Kim, S. S., & Seo, K. (2022). Similarity-based adversarial knowledge distillation using graph convolutional neural network. *Electronics Letters, 58*(16), 606–608.

Li, C., Peng, J., Yuan, L., Wang, G., Liang, X., Lin, L., & Chang, X. (2020). Block-wisely supervised neural architecture search with knowledge distillation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1986–1995). https://doi.org/10.1109/CVPR42600.2020.00206

Li, Y., Nie, X., Diao, W., & Zheng, S. (2021). Lifelong CycleGAN for continual multi-task image restoration. *Pattern Recognition Letters, 153*, 183–189. https://doi.org/10.1016/j.patrec.2021.12.010

Li, Z., Ye, J., Song, M., Huang, Y., & Pan, Z. (2022a). Online knowledge distillation for efficient pose estimation. *arXiv.org*. https://arxiv.org/abs/2108.02092.

Li, G., Togo, R., Ogawa, T., & Haseyama, M. (2022b). Self-knowledge distillation based self-supervised learning for covid-19 detection from chest x-ray images. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1371–1375). IEEE.

Li, W., Wang, J., Ren, T., Li, F., Zhang, J., & Wu, Z. (2022c). Learning accurate, speedy, lightweight CNNs via instance-specific multi-teacher knowledge distillation for distracted driver posture identification. *IEEE Transactions on Intelligent Transportation Systems, 23*(10), 17922–17935. https://doi.org/10.1109/TITS.2022.3161986. IEEE Transactions on Intelligent Transportation Systems.

Li, Z., Li, X., Yang, L., Song, R., Yang, J., & Pan, Z. (2024a). Dual teachers for self-knowledge distillation. *Pattern Recognition, 151*, Article 110422.

Li, Y., Li, P., Yan, D., Liu, Y., & Liu, Z. (2024b). Deep knowledge distillation: A self-mutual learning framework for traffic prediction. *Expert Systems With Applications, 252*, Article 124138. https://doi.org/10.1016/j.eswa.2024.124138

Liang, P., Zhang, W., Wang, J., & Guo, Y. (2024). Neighbor self-knowledge distillation. *Information Sciences, 654*, Article 119859.

Lin, R., Lv, X., Hu, H., Ling, L., Yu, Z., & Zhang, D. (2023). Dual-stage ensemble approach using online knowledge distillation for forecasting carbon emissions in the electric power industry. *Data Science and Management, 6*(4), 227–238. https://doi.org/10.1016/j.dsm.2023.09.001

Liu, C., Chen, L. C., Schroff, F., Adam, H., Hua, W., Yuille, A. L., & Fei-Fei, L. (2019a). Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 82–92).

Liu, J., Zheng, T., & Hao, Q. (2022a). *HIRE: Distilling High-Order Relational Knowledge from Heterogeneous Graph Neural Networks*. Cornell University. https://doi.org/10.48550/arxiv.2207.11887. *arXiv*.

Liu, J., Zheng, T., Zhang, G., & Hao, Q. (2023). Graph-based knowledge distillation: A survey and experimental evaluation. *arXiv preprint. arXiv:2302.14643*.

Liu, H., Simonyan, K., & Yang, Y. (2019b). Darts: Differentiable architecture search. *arXiv preprint. arXiv:1806.09055*.

Liu, Y., Zhang, W., & Wang, J. (2020). Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing, 415*, 106–113. https://doi.org/10.1016/j.neucom.2020.07.048

Liu, Y., Cao, J., Li, B., Hu, W., Ding, J., & Li, L. (2022b). Cross-architecture knowledge distillation. In *Proceedings of the Asian conference on computer vision* (pp. 3396–3411).

Liu, Z., Wang, H., & Wang, S. (2022c). Cross-domain local characteristic enhanced deepfake video detection. In *Proceedings of the Asian Conference on Computer Vision* (pp. 3412–3429).

Lopes, R. G., Fenu, S., & Starner, T. (2017). Data-free knowledge distillation for deep neural networks. *arXiv preprint. arXiv:1710.07535*.

López-Cifuentes, A., Escudero-Viñolo, M., Bescós, J., & Miguel, J. C. S. (2023). Attention-based knowledge distillation in scene recognition: The impact of a DCT-Driven loss. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(9), 4769–4783. https://doi.org/10.1109/tcsvt.2023.3250031

McDonald, D., Papadopoulos, R., & Benningfield, L. (2024). Reducing llm hallucination using knowledge distillation: A case study with mistral large and mmlu benchmark. *Authorea Preprints*.

Makhlouf, A., Maayah, M., Abughanam, N., et al. (2023). The use of generative adversarial networks in medical image augmentation. *Neural Comput & Applic, 35*, 24055–24068. https://doi.org/10.1007/s00521-023-09100-z

Mirzadeh, S. I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., & Ghasemzadeh, H. (2020, April). Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 04, pp. 5191-5198). https://doi.org/10.1609/aaai.v34i04.5963.

Mora, A., Tenison, I., Bellavista, P., & Rish, I. (2022). Knowledge distillation for federated learning: a practical guide. *arXiv preprint. arXiv:2211.04742*.

Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3967–3976).

Passban, P., Wu, Y., Rezagholizadeh, M., & Liu, Q. (2020). *ALP-KD: Attention-Based Layer Projection for Knowledge Distillation*. Cornell University. https://doi.org/10.48550/arxiv.2012.14022. *arXiv*.

Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., & Zhang, Z. (2019). Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 5007–5016).

Pham, C., Hoang, T., & Do, T. (2022b). Collaborative multi-teacher knowledge distillation for learning low bit-width deep neural networks. *arXiv.org*. https://arxiv.org/abs/2210.16103.

Pham, M., Cho, M., Joshi, A., & Hegde, C. (2022a). Revisiting self-distillation. *arXiv preprint. arXiv:2206.08491*.

Rakhmonov, A. A. U., Subramanian, B., Olimov, B., & Kim, J. (2023). Extensive knowledge distillation model: An end-to-end effective anomaly detection model for real-time industrial applications. *IEEE Access*.

Sarkar, P., & Etemad, A. (2022). *XKD: cross-modal knowledge distillation with domain alignment for video representation learning*. Cornell University. https://doi.org/10.48550/arxiv.2211.13929. arXiv.

Schmid, F., Koutini, K., & Widmer, G. (2022). *Efficient large-scale audio tagging via transformer-to-CNN knowledge distillation*. Cornell University. https://doi.org/10.48550/arxiv.2211.04772. *arXiv*.

Sepahvand, M., Abdali-Mohammadi, F., & Taherkordi, A. (2022). Teacher–student knowledge distillation based on decomposed deep feature representation for intelligent mobile applications. *Expert Systems with Applications, 202*, Article 117474.

Soltoggio, A., Ben-Iwhiwhu, E., Braverman, V. et al. A collective AI via lifelong learning and sharing at the edge. Nature Machine Intelligence 6, 251–264 (2024).

Song, L., Gong, X., Zhou, H., Chen, J., Zhang, Q., Doermann, D., & Yuan, J. (2023). Exploring the knowledge transferred by response-based teacher-student distillation. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 2704–2713).

Srinivasagan, G., Deisher, M., & Georges, M. (2023). *Compression of End-to-End Non-Autoregressive Image-to-Speech System for Low-Resourced Devices*. Cornell University. https://doi.org/10.48550/arxiv.2312.00174. arXiv.

Tang, Y., & Liu, Z. (2024). A Distributed knowledge distillation framework for financial fraud detection based on transformer. *IEEE Access, 12*, 62899–62911. https://doi.org/10.1109/ACCESS.2024.3387841

Trivedi, A., Udagawa, T., Merler, M., Panda, R., El-Kurdi, Y., & Bhattacharjee, B. (2023). Neural architecture search for effective teacher-student knowledge transfer in 0language models. *arXiv.org*. https://arxiv.org/abs/2303.09639.

Trofimov, I., Klyuchnikov, N., Salnikov, M., Filippov, A., & Burnaev, E. (2023). Multi-fidelity neural architecture search with knowledge distillation. *IEEE Access, 11*, 59217–59225.

Tung, F., & Mori, G. (2019). Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1365–1374).

Wang, C., Zhou, S., Yu, K., Chen, D., Li, B., Feng, Y., & Chen, C. (2022a). Collaborative knowledge distillation for heterogeneous information network embedding. In *Proceedings of the ACM Web Conference 2022*. https://doi.org/10.1145/3485447.3512209

Wang, W., Liu, F., Liao, W., & Xiao, L. (2023). Cross-modal graph knowledge representation and distillation learning for land cover classification. *IEEE Transactions on Geoscience and Remote Sensing, 61*, 1. https://doi.org/10.1109/TGRS.2023.3307604

Wang, X., Zhang, R., Sun, Y., & Qi, J. (2018). Kdgan: Knowledge distillation with generative adversarial networks. *Advances in neural information processing systems, 31*.

Wang, Y. H., Lin, C. Y., Thaipisutikul, T., & Shih, T. K. (2022b). Single-head lifelong learning based on distilling knowledge. *IEEE Access, 10*, 35469–35478.

Wang, N., Deng, Y., Feng, W., Yin, J., & Ng, S. K. (2024). Data-free federated class incremental learning with diffusion-based generative memory (no. arXiv: 2405.17457). *arXiv*. https://doi.org/10.48550/arXiv.2405.17457

Wu, C., Wu, F., & Huang, Y. (2021a). One teacher is enough? pre-trained language model distillation from multiple teachers. *arXiv.org*. https://arxiv.org/abs/2106.01023.

Wu, Y., Rezagholizadeh, M., Ghaddar, A., Haidar, M. A., & Ghodsi, A. (2021b). Universal-KD: attention-based output-grounded intermediate layer knowledge distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2021.emnlp-main.603

Xia, W., Li, X., Deng, A., Xiong, H., Dou, D., & Hu, D. (2023). *Robust Cross-Modal knowledge distillation for unconstrained videos*. Cornell University. https://doi.org/10.48550/arxiv.2304.07775. arXiv.

Xiang, Q., Zhang, M., Shang, Y., Wu, J., Yan, Y., & Nie, L. (2024). DKDM: Data-Free Knowledge Distillation for Diffusion Models with Any Architecture. *arXiv.org*. https://arxiv.org/abs/2409.03550.

Xue, Z., Gao, Z., Ren, S., & Zhao, H. (2022). The modality focusing hypothesis: towards understanding crossmodal knowledge distillation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2206.06487.

Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., & Zhang, Q. (2022). Cross-image relational knowledge distillation for semantic segmentation. *arXiv.org*. https://arxiv.org/abs/2204.06986.

Yang, C., Yu, X., An, Z., & Xu, Y. (2023a). Categories of response-based, feature-based, and relation-based knowledge distillation. *Advancements in Knowledge Distillation: Towards New Horizons of Intelligent Systems* (pp. 1–32). Cham: Springer International Publishing.

Yang, Z., Zeng, A., Li, Z., Zhang, T., Yuan, C., & Li, Y. (2023b). From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 17185–17194).

Yang, L., & Xu, K. (2021). Cross modality knowledge distillation for multi-modal aerial view object classification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 382–387). https://doi.org/10.1109/cvprw53098.2021.00048

Yang, Y., Qiu, J., Song, M., Tao, D., & Wang, X. (2020). *Distilling Knowledge from Graph Convolutional Networks*. Cornell University. https://doi.org/10.48550/arxiv.2003.10477. *arXiv*.

Yim, J., Joo, D., Bae, J., & Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4133–4141).

Yin, G., Wang, W., Yuan, Z., Han, C., Ji, W., Sun, S., & Wang, C. (2022). *Content-Variant Reference Image Quality Assessment Via Knowledge Distillation*. Cornell University. https://doi.org/10.48550/arxiv.2202.13123. *arXiv*.

Ye, F., & Bors, A. G. (2021). Lifelong twin generative adversarial networks. In *2021 IEEE International Conference on Image Processing (ICIP)* (pp. 1289–1293). IEEE. https://doi.org/10.1109/ICIP42928.2021.9506116.

You, S., Xu, C., Xu, C., & Tao, D. (2017). Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1285–1294). https://doi.org/10.1145/3097983.3098135

Yue, J., Fang, L., Rahmani, H., & Ghamisi, P. (2022). Self-supervised learning with adaptive distillation for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–13. https://doi.org/10.1109/TGRS.2021.3057768

Yun, P., Liu, Y., & Liu, M. (2021). In defense of knowledge distillation for task incremental learning and its application in 3D object detection. *IEEE Robotics and Automation Letters, 6*(2), 2012–2019.

Zagoruyko, S., & Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint. arXiv:1612.03928*.

Zhai, M., Chen, L., & Mori, G. (2021). Hyper-lifelonggan: Scalable lifelong learning for image conditioned generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2246–2255).

Zhang, C., & Peng, Y. (2018). Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. *arXiv preprint. arXiv:1804.10069*.

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., & Ma, K. (2019). Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3713–3722).

Zhang, L., Bao, C., & Ma, K. (2021). Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(8), 4388–4403.

Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. (2018). Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4320–4328).

Zhao, Z., Lyu, J., Chu, Y., Liu, K., Cao, D., Wu, C., Qin, L., & Qin, S. (2023). Toward generalizable robot vision guidance in real-world operational manufacturing factories: A semi-supervised knowledge distillation approach. *Robotics and Computer-Integrated Manufacturing, 86*, Article 102639. https://doi.org/10.1016/j.rcim.2023.102639

Zhao, K., & Zhao, M. (2024). Self-supervised quantization-aware knowledge distillation. *arXiv preprint. arXiv:2403.11106*.

Zhu, Z., Hong, J., & Zhou, J. (2021). Data-free knowledge distillation for heterogeneous federated learning. *arXiv.org*. https://arxiv.org/abs/2105.10056.

Zou, W., Qi, X., Wu, Z., Wang, Z., Sun, M., & Shan, C. (2021). Coco distillnet: a cross-layer correlation distillation network for pathological gastric cancer segmentation. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1227–1234). IEEE.