https://crad.ict.ac.cn
63(1): 90−122, 2026

面向视觉算法的知识蒸馏研究综述

潘海为　于丰铭　张可佳　兰海燕　孟庆宇　李　哲
（哈尔滨工程大学计算机科学与技术学院　哈尔滨　150001）
（panhaiwei@hrbeu.edu.cn）

Knowledge Distillation in Visual Algorithms: A Survey

Pan Haiwei, Yu Fengming, Zhang Kejia, Lan Haiyan, Meng Qingyu, and Li Zhe
（*College of Computer Science and Technology*, *Harbin Engineering University*, *Harbin* 150001）

Abstract　Knowledge distillation, as a key technique in deep learning, achieves model compression and acceleration by transferring knowledge from a large teacher model to a smaller student model. Under the premise of maintaining performance, this technology significantly reduces the requirements of computational resources and storage, and facilitates the deployment of high-performance models on resource-constrained edge devices. Firstly, we provide a systematic review of the recent research in knowledge distillation and categorize it from two perspectives: the type of knowledge and teacher-student model architectures. We comprehensively summarize the distillation methods based on three typical types of knowledge: output feature knowledge, intermediate feature knowledge, and relational feature knowledge, as well as distillation methods based on CNN to CNN architecture, CNN to ViT (vision Transformer) architecture, ViT to CNN architecture, and ViT to ViT architecture. Next, we explore various learning paradigms such as offline distillation, online distillation, self-distillation, data-free distillation, multi-teacher distillation, and assistant distillation. Then, we summarize distillation optimization methods based on the distillation process, knowledge structure, temperature coefficient, and loss functions and analyze improvements in distillation brought by adversarial techniques, automated machine learning, reinforcement learning, and diffusion models, and conclude with the implementation of distillation technology in common applications. Despite significant advancements in knowledge distillation, numerous challenges remain in both practical applications and theoretical research. Finally, we provide an in-depth analysis of these issues and offer insights into future development directions.

Key words　knowledge distillation; model compression; deep learning; convolutional neural network; vision Transformer

摘　要　知识蒸馏作为深度学习中的关键技术，通过将大型教师模型的知识传递给较小的学生模型，实现了模型的压缩与加速。在保证性能的前提下，显著减少了计算资源和存储需求，促进了高性能模型在资源受限的边缘设备上的部署。围绕知识蒸馏的最新研究进展进行了系统性的综述，从知识类型和师生模型架构2个角度对知识蒸馏进行分类，详细汇总了输出特征知识、中间特征知识、关系特征知识3种典型知识类型的蒸馏方法，以及卷积架构到卷积架构、卷积架构到ViT（vision Transformer）架构、ViT架构到卷积架构和ViT架构到ViT架构的蒸馏方法；探讨了离线蒸馏、在线蒸馏、自蒸馏、无数据蒸馏、多教师蒸馏和助理蒸馏的学习方式；归纳了基于蒸馏过程、知识结构、温度系数及损失函数的蒸馏优化方法，分析

收稿日期：2024-08-29；**修回日期**：2025-04-29
基金项目：国家自然科学基金项目（62072135）；国家工业和信息化部项目（CBZ3N21-2）
This work was supported by the National Natural Science Foundation of China (62072135) and the Project of Ministry of Industry and Information Technology (CBZ3N21-2).
通信作者：张可佳（kejiazhang@hrbeu.edu.cn）

了对抗性技术、自动机器学习、强化学习和扩散模型对蒸馏的改进，并总结了蒸馏技术在常见应用中的实现。尽管知识蒸馏取得了显著进展，但在实际应用和理论研究中仍面临诸多挑战。最后，对这些问题进行了深入分析，并对未来发展方向提出了见解。

关键词 知识蒸馏；模型压缩；深度学习；卷积神经网络；视觉 Transformer

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440694 **CSTR:** 32373.14.issn1000-1239.202440694

近年来，随着深度学习技术的迅猛发展，神经网络模型在计算机视觉、自然语言处理等诸多领域取得了显著的突破。1998 年，LeNet-5[1] 作为最早的卷积神经网络（convolutional neural network, CNN）之一，奠定了现代 CNN 的基础框架。2012 年，AlexNet[2] 通过学习到的特征超越手工设计特征，掀起深度学习热潮，并推动了如 VGGNet[3] 和 Inception 系列[4-6] 网络的出现。随着网络深度增加，梯度消失问题成为瓶颈，ResNet[7] 和 DenseNet[8] 提出残差连接与密集连接，有效解决该问题并提升网络深度上限。与此同时，深度网络带来的高计算开销限制了其在资源受限环境中的部署，促使 MobileNet 系列[9-11] 和 EfficientNet 系列[12-13] 网络聚焦于轻量化和高效性设计。

随着 Transformer[14] 在自然语言处理（natural language processing, NLP）领域的成功，Dosovitskiy 等人[15] 提出 ViT（vision Transformer），将其架构引入视觉任务，引领视觉 Transformer 的研究。随后，Swin Transformer[16] 通过滑动窗口机制实现高分辨率图像的高效处理；DeiT[17] 通过蒸馏技术降低了 ViT 对大规模数据的依赖，PVT[18] 结合金字塔结构提升多尺度建模能力。此外，CLIP[19] 和 DALL-E[20] 等模型结合视觉与语言，将 ViT 应用于多模态任务中，展现了强大的跨模态理解与生成能力。

然而，这些性能卓越的模型往往具有庞大的参数规模和计算复杂度，导致在资源受限的环境中难以部署。为了解决这一问题，模型压缩技术应运而生。模型压缩方法主要包括权重剪枝、量化、低秩分解以及知识蒸馏等。其中知识蒸馏作为一种有效的模型压缩方法，受到了广泛关注，目前已经成为深度学习领域的一大研究热点。

知识蒸馏最早由 Hinton 等人[21] 提出，其基本思想是通过训练一个较小的学生模型，来逼近一个较大且性能较好的教师模型。具体而言，教师模型通常预先训练完成，并在训练学生模型时提供额外的监督信号，帮助学生模型更好地学习。这样的机制不仅可以显著减少模型的参数数量和计算开销，还能够在保持精度的情况下提升模型的推理速度。

知识蒸馏方法发展至今，已经衍生出多种变体和应用。从最初的经典方法到对教师模型输出分布的探索，再到多教师模型和多任务学习的结合，研究者们在这一领域不断创新，提出了多种改进策略。例如，基于中间特征层的蒸馏方法通过教师模型的中间层特征来指导学生模型的训练；自蒸馏方法则利用同一模型在不同训练阶段的信息进行蒸馏。此外，还有结合对抗训练和强化学习等新兴技术的知识蒸馏方法，这些策略不仅在理论上深化了我们对知识蒸馏的理解，也在实践中推动了模型压缩技术的广泛应用。

知识蒸馏的应用领域也非常广泛。除了在计算机视觉和自然语言处理等传统深度学习领域的应用之外，知识蒸馏还在自动驾驶、医疗影像分析、语音识别等领域展现出了强大的潜力。这些应用不仅验证了知识蒸馏的有效性，也为未来的研究提供了丰富的实践案例。

本综述旨在系统地梳理和总结知识蒸馏在视觉领域的研究进展。首先，我们将回顾知识蒸馏的基本原理和经典方法；然后从知识蒸馏的分类、学习方式、优化方法以及应用的角度介绍近年来出现的各种改进和扩展；最后我们将讨论当前研究中存在的挑战和未来的发展方向，为该领域的进一步研究提供参考。

先前 2 项比较全面的外文综述[22-23] 和 3 项不错的中文综述[24-26] 已经对知识蒸馏有了一个较为系统的介绍，但它们都忽略了对于知识蒸馏学习范式优化的梳理以及关于视觉 Transformer 架构模型蒸馏的总结，通过这篇综述希望能够为读者提供一个全面而清晰的知识蒸馏研究脉络，帮助相关领域的研究人员和工程师更好地理解和应用这一重要的模型压缩技术。

本文的章节将从知识蒸馏的背景、知识蒸馏的分类、知识蒸馏的学习方式、知识蒸馏的优化和知识蒸馏的应用详细介绍，如图 1 所示。

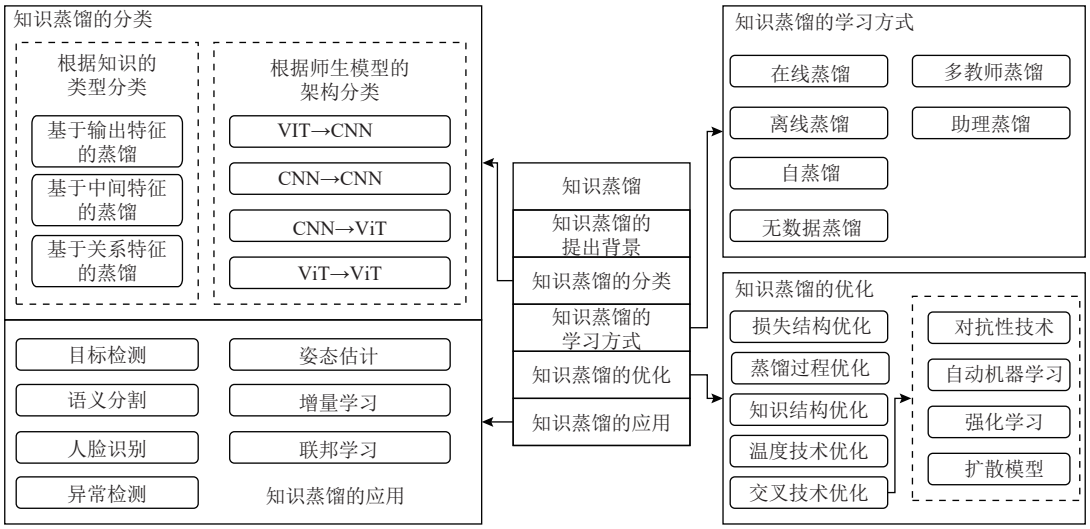


Fig. 1 Overview of the general structure in our paper

图 1 本文总体结构概览

1 知识蒸馏的背景

随着深度学习技术的快速发展，模型的规模和复杂度不断攀升。从早期的简单卷积网络（如 LeNet-5[1]）到近年来的深层宽层网络，再到以 Transformer 为基础的大型预训练模型，这些创新架构在训练时通常能够取得卓越的性能，推动了计算机视觉、自然语言处理等领域的飞跃。然而，这些高性能模型往往面临着庞大的参数量和计算需求，这使得它们在计算资源和存储空间受限的环境中部署变得极为困难，亟需解决效率和资源消耗的问题。

为了解决这一挑战，研究者提出了多种模型压缩技术，旨在减少大规模模型对资源的占用及推理过程中的计算消耗。常见的压缩策略包括权重剪枝、量化、低秩分解以及知识蒸馏等，其中知识蒸馏因其在保持模型性能的同时显著降低计算负担而受到广泛关注，成为近年来深度学习研究的重点之一。

相比于其他模型压缩方法，知识蒸馏具有独特的优势。权重剪枝通过移除冗余权重来减少模型参数，但往往需要精细的剪枝策略以平衡性能与压缩率；量化通过将权重和激活值映射到较低的比特表示以降低存储和计算成本，但可能会带来一定的精度损失；低秩分解通过将权重矩阵分解为低秩矩阵来减少计算复杂度，但实现过程较为复杂；知识蒸馏则通过让学生模型模仿教师模型的行为，直接优化学生模型的输出，使其在较小规模下仍能保持高性能，并且知识蒸馏不仅能够有效地压缩模型，还能保持甚至提高学生模型的泛化能力。知识蒸馏在许多实际应用中表现出色，成为模型压缩领域的重要方法。

1.1 符号系统

数据集输入表示为$\{\boldsymbol{x}_i\}_{i=1}^N=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$，对应的标签集合表示为$\{\boldsymbol{y}_i\}_{i=1}^N=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_N\}$。针对 2 种不同的网络架构分别给出形式化描述：

1）卷积架构。对于具有l层的卷积网络，如图 2 所示，$f(\cdot)$表示整个网络架构，$\boldsymbol{F}$表示中间层特征。

$$\{\boldsymbol{F}_i\}_{i=1}^l=\{\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_l\}\text{。} \tag{1}$$

$\boldsymbol{z}=f(\boldsymbol{x})$表示 Logits 输出，Logits 输出经过 softmax 层激活后为概率分布输出，表示为$\boldsymbol{p}=softmax_\tau(\boldsymbol{z})$，其中：

$$softmax_\tau(\boldsymbol{z})=\frac{\exp(\boldsymbol{z}/\tau)}{\sum_j\exp(\boldsymbol{z}_j/\tau)}, \tag{2}$$

τ表示温度系数。

对于教师模型，具有l_{T}层的网络表示为$f^{\mathrm{T}}(\cdot)$，$\{\boldsymbol{F}_i^{\mathrm{T}}\}_{i=1}^{l_{\mathrm{T}}}=\{\boldsymbol{F}_1^{\mathrm{T}},\boldsymbol{F}_2^{\mathrm{T}},\cdots,\boldsymbol{F}_{l_{\mathrm{T}}}^{\mathrm{T}}\}$表示教师模型的特征层，$\boldsymbol{z}^{\mathrm{T}}=f^{\mathrm{T}}(\boldsymbol{x})$，$\boldsymbol{p}^{\mathrm{T}}=softmax_\tau(\boldsymbol{z}^{\mathrm{T}})$分别表示教师模型的 Logits 输出和概率分布输出。对于学生模型同理。

2）视觉 Transformer 架构。对于具有L个 Transformer 块的 ViT[15]，如图 3 所示。

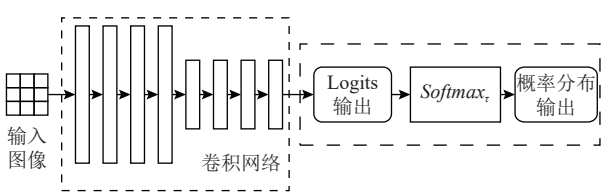


Fig. 2 Convolutional network model architecture

图 2 卷积网络模型架构

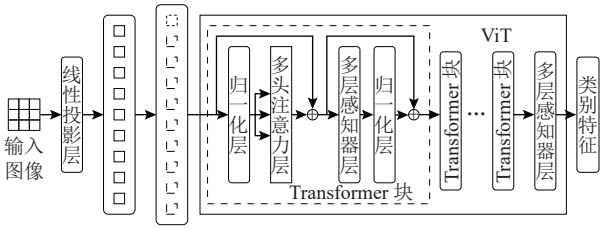


Fig. 3 ViT model architecture

图 3 ViT 模型架构

输入图像$\boldsymbol{x}\in\mathbb{R}^{H\times W\times C}$被切分为$N$个图像块并拉伸为 1 维向量$\{\boldsymbol{x}_{\mathrm{p}}^{j}\in\mathbb{R}^{P^2C}|j\in N,1\leqslant j\leqslant N\}$，$P$表示图像块大小，$H$和$W$分别表示图像的高和宽，$C$表示图像通道数，$N=HW/P^2$，每一个图像块通过$D$维向量模型编码：

$$\boldsymbol{e}_j=\boldsymbol{x}_{\mathrm{p}}^{j}\cdot\boldsymbol{E},\ \ \boldsymbol{E}\in\mathbb{R}^{(P^2C)\times D}\ 。\tag{3}$$

编码后的图像块插入一个可学习的分类 token，并对整个序列添加位置编码：

$$Input=\boldsymbol{v}_0=[\boldsymbol{x}_{\mathrm{cls}};\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_N]+\boldsymbol{E}_{\mathrm{pos}},\ \ \boldsymbol{E}_{\mathrm{pos}}\in\mathbb{R}^{(N+1)\times D}\ 。\tag{4}$$

将编码后的序列输入到 Transformer 块中，Transformer 块由多头自注意力（mutil-head attention, MHA）层、多层感知器（multilayer perceptron, MLP）层和层归一化（layer normalization, LN）层组成，每一个 Transformer 块使用上一个块的输出$\boldsymbol{v}_{l-1}$作为输入，得到当前块的特征：

$$\boldsymbol{v}_l=TransformerBlock(\boldsymbol{v}_{l-1}),\ \ l\in\{1,2,\cdots,L\}\ 。\tag{5}$$

最后一个块的输出$\boldsymbol{v}_L$称之为输出特征，$\boldsymbol{q}$表示由分类 token$\boldsymbol{v}_L^0$计算得到的类别特征。块内特征可以表示为

$$\begin{aligned}&\boldsymbol{v}'_l=MHA(LN(\boldsymbol{v}_{l-1}))+\boldsymbol{v}_{l-1},\ \ l=1,2,\cdots,L,\\&\boldsymbol{v}_l=LN(MLP(\boldsymbol{v}'_l))+\boldsymbol{v}'_l,\ \ l=1,2,\cdots,L,\\&\boldsymbol{q}=LN(\boldsymbol{v}_L^0)\ 。\end{aligned}\tag{6}$$

对于具有L_{T}个 Transformer 块的教师模型，$\{\boldsymbol{v}^{\mathrm{T}}\}_{l=1}^{L_{\mathrm{T}}}=\{\boldsymbol{v}_1^{\mathrm{T}},\boldsymbol{v}_2^{\mathrm{T}},\cdots,\boldsymbol{v}_{L_{\mathrm{T}}}^{\mathrm{T}}\}$表示 Transformer 块的特征，$\boldsymbol{v}_{L_{\mathrm{T}}}^{\mathrm{T}}$和$\boldsymbol{q}^{\mathrm{T}}$分别表示教师模型的输出特征和类别特征。对于学生模型同理。

针对下文涉及的距离度量，表 1 给出其符号缩写、英文全称和中文名称。

1.2 知识蒸馏的提出

蒸馏这一概念在 2015 年首次由 Hinton 等人[21]提出，核心思想是通过训练一个较小的学生模型来逼近一个较大且性能较好的教师模型。这一类似的思想最早由 Bucila 等人[28]在 2006 年提出，他们的工作令人信服地证明了较大的模型获得的知识可以转移到较小的模型中。

Table 1 Reference Table of Distance Symbols

表 1 距离符号对照表

符号缩写	英文全称	中文名称
L_1	L_1-Norm	曼哈顿距离
L_2	L_2-Norm	欧氏距离
L_∞	L_∞-Norm	切比雪夫距离
CE	cross entropy	交叉熵
KL	Kullback-Leibler divergence	KL 散度
MSE	mean-square error	均方误差
MMD	maximum mean discrepancy	最大均值差异
EMD	earth mover's distance	推土机距离
InfoNCE[27]	information noise-contrastive estimation	

知识蒸馏系统通常由 3 个关键部分组成：知识、蒸馏算法和师生模型架构。图 4 展示了一个知识蒸馏的基本框架。

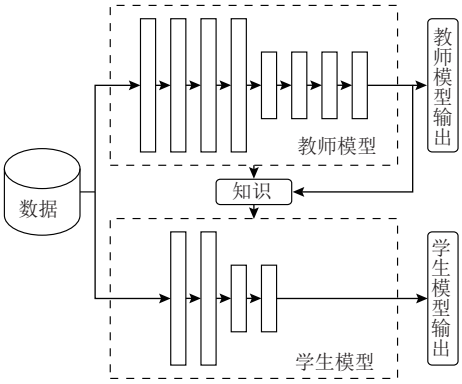


Fig. 4 Basic architecture of knowledge distillation

图 4 知识蒸馏的基本架构

Hinton 等人[21]提出通过最小化教师模型产生的概率分布输出（$\boldsymbol{p}^{\mathrm{T}}$）和学生模型产生的概率分布输出（$\boldsymbol{p}^{\mathrm{S}}$）之间的差异来进行知识的转移，然而训练有素的教师模型在 softmax 函数上产生的概率分布输出具有较高的真实类置信度，而其他类的概率都接近于 0。因此引入了 softmax 函数的温度概念来软化标签，称之为温度系数，软化后的标签称为软标签，表示为

$$\boldsymbol{p}^{\mathrm{T}}=\frac{\exp(z^{\mathrm{T}}/\tau)}{\sum_j\exp(z_j^{\mathrm{T}}/\tau)},\ \ \boldsymbol{p}^{\mathrm{S}}=\frac{\exp(z^{\mathrm{S}}/\tau)}{\sum_j\exp(z_j^{\mathrm{S}}/\tau)},\tag{7}$$

其中τ表示温度系数，随着τ的不断增大，概率分布输出分布变得越来越平稳，最终趋于均匀分布，当$\tau=1$时，退化为普通的 softmax 函数。更平稳的概率分布输出分布能更好地表达教师模型发现的与真实类相近的类别信息，这部分知识被称为暗知识，众多学者认为暗知识是使知识蒸馏有效的因素之一。

除了教师模型软标签提供的暗知识,数据的真实标签也能提供有效的信息,一个典型的蒸馏损失可以形式化表述为

$$L_{\mathrm{KD}}=\alpha H(\boldsymbol{p}^{\mathrm{S}};\boldsymbol{y})+\beta D_{\mathrm{KL}}(\boldsymbol{p}^{\mathrm{T}}\|\boldsymbol{p}^{\mathrm{S}})=-\alpha\sum_i \boldsymbol{y}_i\log(\boldsymbol{p}_i^{\mathrm{S}})+\beta\sum_i \boldsymbol{p}_i^{\mathrm{T}}\log\frac{\boldsymbol{p}_i^{\mathrm{T}}}{\boldsymbol{q}_i^{\mathrm{S}}}。\tag{8}$$

知识蒸馏在深度学习领域有多种应用,主要通过将大型、复杂的教师模型的知识传递给较小、较简单的学生模型,达到在推理阶段加速并减少存储需求的目的;其次知识蒸馏还有进行知识迁移和隐私保护等应用,有以下 3 个方面:

1)模型压缩。知识蒸馏最主要的用途是模型压缩。通过将大型教师模型的知识传递给较小的学生模型,能够在保持模型性能的同时,显著减少模型的参数量和计算复杂度,使其适用于资源受限的环境。

2)知识迁移。知识蒸馏可以利用其训练时传递知识的特性进行跨模态或跨架构知识的迁移,这进一步扩大了蒸馏的灵活性和适用性,能够使学生模型获得更多样化和丰富的知识,从而提升其在不同任务中的性能。

3)隐私保护。知识蒸馏可以在不直接暴露原始数据的情况下传递知识,这在涉及隐私数据的应用中(如医疗数据、金融数据等)具有重要意义。通过知识蒸馏,敏感数据可以用于训练教师模型,而学生模型则可以在不接触这些敏感数据的情况下获取知识。

2 知识蒸馏的分类

本文从知识类型和师生模型架构 2 个角度对知识蒸馏方法进行了分类,如图 1 中所示的知识蒸馏的分类。知识类型的分类遵循先前的工作[22-26],通常从输出特征、中间特征和关系特征 3 个层面进行汇总。近年来,随着 Transformer 模型的出现,师生模型的架构不再局限于卷积模型。因此,本文还从师生模型架构的角度对蒸馏方法进行了分类和汇总。

2.1 知识类型

本节从学生模型获取知识的来源对知识蒸馏方法进行分类,通常包括基于输出特征的蒸馏、基于中间特征的蒸馏和基于关系特征的蒸馏。这些方法分别利用教师模型的输出特征、中间层特征,以及样本间或特征图间的关系作为知识来指导学生模型训练。

2.1.1 基于输出特征的蒸馏

基于输出特征的蒸馏通常是指利用教师模型最后一层神经元的输出指导学生模型训练,其主要思想是让学生模型来模仿训练有素的教师模型的输出来获得高可信的结果。基于输出特征的知识蒸馏简单有效,有着广泛的应用。根据输出特征知识的位置,将经过 softmax 层激活的知识称之为概率分布输出;将未经过 softmax 层激活的知识称之为 Logits 输出。

1)概率分布输出。概率分布输出指神经网络经过 softmax 层之后的输出,如图 5 所示,基于概率分布输出的蒸馏损失可以描述为

$$L_{\mathrm{KD}}=L_{\mathrm{Dis}}(\boldsymbol{p}^{\mathrm{T}},\boldsymbol{p}^{\mathrm{S}}),\tag{9}$$

其中 L_{Dis} 表示距离度量函数。

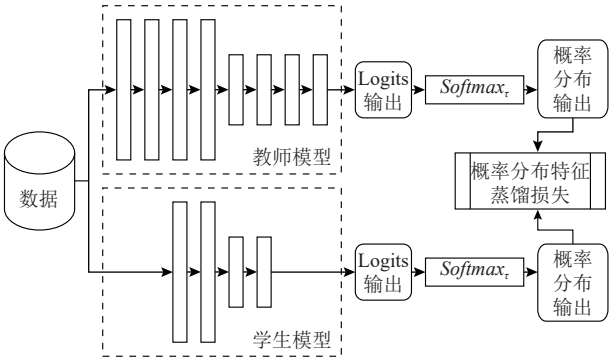


Fig. 5 Illustration of distillation method based on probability distribution output

图 5 基于概率分布输出的蒸馏方法示意图

2015 年,Hinton 等人[21]提出了利用软目标中蕴含的来自教师模型的暗知识来指导学生模型学习,并利用温度系数来加强概率较低的标签概率对模型的影响。Sau 等人[29]对教师模型输出的知识添加噪声来模拟同一领域的不同教师模型,证明了添加噪声相当于正则化项的作用,实现了更好的深度模型的压缩。Zhang 等人[30]提出了互学习的概念,让多个模型之间互相学习其概率分布输出,证明了具有高熵的概率模型能训练出更鲁棒的结构。Furlanello 等人[31]提出了序列化建模方案,令前一个模型指导后一个模型的训练,最终将所有学生模型集成得到最终的学生模型。Yang 等人[32]认为视觉上相似的物体同样包含着有效的信息,引入了类别相似性知识,计算主类与前 k 个类别的置信度差异,加强次要类别的置信度来得到一个更宽容的教师模型。Lee 等人[33]认为线性可分的任务经过某些数据增强手段之后可能不再线性可分,提出学习一个原始标签与增强标签的联合分布任务来进行蒸馏学习。Yang 等人[34]专注于教师模型分类器的学习,在倒数第 2 层匹配师生模型特征后通过教师模型的 softmax 函数计算概率分布输出,计算交叉熵来优化整体的蒸馏损失。

2)Logits 输出。Logits 输出指神经网络最后 1 层

和 softmax 层之前的输出，如图 6 所示，基于 Logits 输出的蒸馏损失可以描述为

$$L_{KD} = L_{Dis}(z^T, z^S), \tag{10}$$

其中 L_{Dis} 表示距离度量函数。

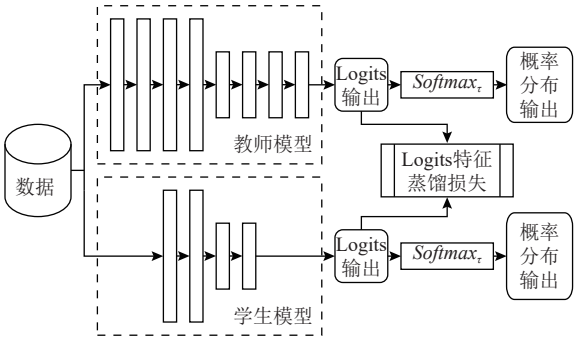


Fig. 6 Illustration of distillation method based on Logits output

图 6 基于 Logits 输出的蒸馏方法示意图

在 2014 年，Ba 等人[35]提出了与 Hinton 等人[21]类似的观点，他们认为模型压缩的成功表明了较小的网络也能够达到与复杂网络相似的精度，使用简单的网络模仿已经训练好的复杂网络的 Logits 输出，并在隐藏层的学习中加入了矩阵分解来加快模型的学习。Zhou 等人[36]实验发现对 Logits 输出进行蒸馏要优于对概率分布输出进行蒸馏，提出一个共享低级别参数的师生模型架构，并通过梯度阻断模块来避免学生网络的梯度对教师模型的负优化。Son 等人[37]针对 Mirzadeh 等人[38]提出的助理模型蒸馏来弥合师生模型能力差距的方案，助理模型会对错误的预测进行累计，造成性能的下降，Son 等人借鉴 DenseNet[8]的思想将上级教师模型的输出知识始终作为下级的指导信息，并借鉴 Dropout 的思想提出随机丢弃来缓解过拟合。Li 等人[39]认为师生模型之间的差距可以作为知识来指导学生模型训练，利用神经架构搜索（neural architecture search, NAS）得到残差学生模型架构，提出能量熵来设定停止条件，在推理阶段提出自适应推理，对不同难度的样本使用不同数量的残差学生模型来加快简单样本的推理速度。

基于输出特征的蒸馏方法汇总于表 2。

Table 2 Summary of Knowledge Distillation Methods Based on Output Features

表 2 基于输出特征的知识蒸馏方法汇总

方法来源	知识的来源	距离度量	关键技术
文献 [35]	Logits 输出	L_2	提出模仿复杂模型的 Logits 输出，获得与复杂模型精度相当的轻量模型
文献 [21]	概率分布输出	KL	通过温度系数软化标签强调暗知识的作用
文献 [29]	概率分布输出	L_2	提出对教师模型概率分布输出添加噪声，模拟同专业的不同教师模型，缓解知识多样性问题
文献 [30]	概率分布输出	KL	提出让多个模型互相学习来获得更有效更紧凑的模型
文献 [36]	Logits 输出	L_2	通过共享师生模型架构的低级别参数来降低推理所需要的时间
文献 [31]	概率分布输出	CE	提出了一种新的学习范式，通过序列化建模获得高性能的学生模型
文献 [32]	概率分布输出	最高分差	考虑了次要类别概率分布输出的知识，缓解严格教师模型带来的性能损失
文献 [33]	概率分布输出	CE	提出标签与增强标签联合优化，处理样本经过数据增强后不一定线性可分的问题
文献 [34]	概率分布输出 倒数第 2 层输出	CE L_2	提出将表示学习和分类解耦，利用教师分类器辅助学生模型训练，改善学生模型的特征表示
文献 [37]	Logits 输出概率分布输出	CE KL	始终将上级教师模型的知识作为多级下级学生模型的指导信息，缓解教师模型错误累计问题
文献 [39]	师生 Logits 输出差	KL	利用师生模型 Logits 输出的差异作为知识来缩减师生模型之间的能力差距

注：L_2，KL，CE 分别表示欧氏距离、KL 散度和交叉熵。

2.1.2 基于中间特征的蒸馏

基于中间特征的蒸馏通常指利用神经网络中间层特征图中的知识指导学生模型训练，如图 7 所示，通过设计特定的投射层来对齐师生模型特征图的语义信息，基于中间特征的蒸馏损失可以描述为

$$L_{KD} = L_{Dis}(\boldsymbol{F}^T, r(\boldsymbol{F}^S)), \tag{11}$$

其中 $\boldsymbol{F}^T$，$\boldsymbol{F}^S$ 分别表示教师模型和学生模型的中间层特征图，$r(\cdot)$ 表示投射层。

Romero 等人[40]在 2015 年提出了迁移中间层的知识，即训练一个比教师模型更深且更细的网络对模型进行压缩，选择教师模型的隐藏层作为提示层，学生模型的某个隐藏层作为引导层，引导层从提示层进行学习。

如何选择中间层特征传递的知识是基于中间特征蒸馏的一个焦点，Zagoruyko 等人[41]提出了使用基于组的注意力图转移特征知识，给出了基于绝对值和、绝对值的 p 次幂的和以及绝对值的 p 次幂的最大值 3 种计算注意力图的方式。Lin 等人[42]针对特

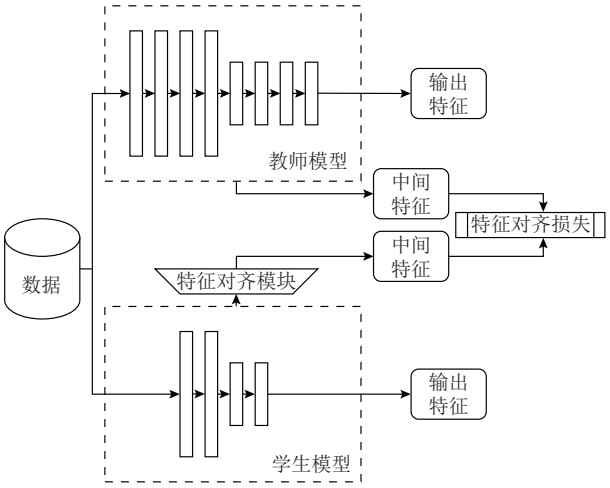


Fig. 7 Illustration of distillation method based on intermediate features

图 7 基于中间特征的蒸馏方法示意图

征图语义不匹配的问题提出目标感知对齐模块, 利用教师模型特征来对学生模型语义特征进行重建, 并且针对大特征图提出分组蒸馏和锚点蒸馏来提取局部特征和全局特征。Lao 等人[43]认为基于中间特征的蒸馏方法的重要知识很容易被简单知识所压制, 提出使用掩码图像机制来传递更有效的知识, 强制学生模型学习教师模型的输出来恢复缺失特征。Guo 等人[44]给出了更具解释性的注意力图蒸馏方案, 调整类注意力图的生成位置, 提出根据类别生成注意力图, 给出了更具解释性的可视化知识转移结果。Zhang 等人[45]将频率域的知识引入蒸馏学习中, 通过 Prompt 方法[46]和离散小波变换来微调教师模型以引入频率域知识, 并利用连续 2 层之间的相关性矩阵作为频率损失的权重。

教师模型的激活边界被认为是训练学生模型的关键信息, Huang 等人[47]通过最小化师生模型特征图分布之间的最大均值差异来匹配神经元的激活分布, 并且在计算过程中提出使用核函数进行映射来辅助计算。Heo 等人[48]给出了线性层和卷积层的激活信息的传递方案, 并给出一个替代函数将离散的激活边界近似为连续函数进行优化。

Shen 等人[49]认为重用预训练的模型能够大大降低训练的成本, 从多个训练有素的教师模型中提取并合并知识, 获得一个能够综合处理分类的轻量级学生模型。Wei 等人[50]将量化方法与蒸馏方法结合, 对学生模型特征图维度对齐后对师生模型进行均匀量化, 对量化后的特征图进行蒸馏。Ahn 等人[51]从信息论的角度对知识蒸馏进行了解释, 将知识的转移表述为师生模型之间的互信息最大化, 通过最大化变分下界来近似分布帮助互信息的计算。Heo 等人[52]对先前具有代表性的 6 项工作(FitNets[40], AT[41], FSP[53], Jacobian[54], FT[55], AB[48])进行对比总结, 对 ReLU 进行修改, 提出 margin-ReLU 来保存教师模型信息, 并提出一种新的距离度量用于师生模型的特征蒸馏。Yang 等人[56]认为 CRD[57]基于对比学习的蒸馏方案会损害模型对原任务的表征能力, 将联合标签增强的自蒸馏方案[33]迁移到传统蒸馏中。教师模型的强大性能不仅仅归功于骨干网络的特征提取能力, 最后的判别器也同等重要。Yang 等人[34]将师生模型输出均输入到教师模型分类器中来计算两者的交叉熵, 并引入一项简化的 FitNets[40]提示损失来进行师生模型中间层的匹配。Chen 等人[58]通过一个简单的投射层对齐师生模型的倒数第 2 层特征, 直接使用教师模型的分类器进行预测输出。

基于中间特征的蒸馏方法汇总如表 3 所示。

2.1.3 基于关系特征的蒸馏

基于关系特征的知识通常指不同层之间或不同输入样本特征之间的关系知识, 如图 8 所示, 基于关系的蒸馏可以描述为

$$L_{\text{L-KD}} = L_{\text{Dis}}\left(\psi_{\text{T}}(\boldsymbol{F}_i^{\text{T}}, \boldsymbol{F}_j^{\text{T}}), \psi_{\text{S}}(\boldsymbol{F}_i^{\text{S}}, \boldsymbol{F}_j^{\text{S}})\right), \tag{12}$$

$$L_{\text{I-KD}} = L_{\text{Dis}}\left(\varphi_{\text{T}}(\boldsymbol{z}_i^{\text{T}}, \boldsymbol{z}_j^{\text{T}}), \varphi_{\text{S}}(\boldsymbol{z}_i^{\text{S}}, \boldsymbol{z}_j^{\text{S}})\right), \tag{13}$$

其中 $L_{\text{L-KD}}$, $L_{\text{I-KD}}$分别表示不同层之间的关系和不同样本特征之间的关系, $\psi(\cdot)$, $\varphi(\cdot)$分别表示层间关系抽取模块和样本特征关系抽取模块, $\boldsymbol{F}^{\text{T}}$, $\boldsymbol{F}^{\text{S}}$表示师生模型的特征图, $\boldsymbol{z}^{\text{T}}$, $\boldsymbol{z}^{\text{S}}$表示师生模型样本特征。

对于层间关系的研究, Yim 等人[53]将教师模型处理问题的流程抽象为求解程序的流程(flow of solution procedure, FSP)矩阵, 学生模型通过学习教师模型的 FSP 矩阵来获得解决问题的流程。You 等人[61]认为不同样本的中间层特征之间的相对差异可以作为教师模型对学生模型更明确更适合的指导知识。Zhang 等人[62]使用每个教师模型的 Logits 输出特征和中间特征作为节点来构成 2 个图, 通过图结构对不同教师模型的重要性和关系进行建模。直接对师生模型特征层进行对齐存在语义不匹配的问题, Chen 等人[63]提出使用相似度矩阵和注意力机制来进行中间层的关联, 从而获得语义对齐的知识来进行蒸馏学习。Liu 等人[64]同样计算特征图通道之间的相似性矩阵用于蒸馏学习, 并针对大特征图给出了分组计算聚合方案。Jin 等人[65]针对概率分布输出蒸馏, 提出多层次对齐方案, 分别从样本实例角度、批次角度和类别角度对师生模型的输出特征进行对齐。

Table 3 Summary of Distillation Methods Based on Intermediate Features

表 3 基于中间特征的蒸馏方法汇总

方法来源	知识的来源	知识的转换	距离度量	关键技术
文献 [40]	中间层特征图	线性层	L_2	提出从教师模型中间层抽取知识来改善训练过程和学生模型性能
文献 [41]	注意力图	归一化层	L_2	利用网络层块注意力图转移特征知识模仿人类视觉系统来提高网络性能
文献 [47]	神经元激活模式	核方法	MMD	提出转移教师模型神经元激活模式来传递知识以提高蒸馏知识的有效性
文献 [50]	中间层特征图	线性层	L_2	通过结合量化和蒸馏 2 种方法来专注于非常微小网络的训练
文献 [51]	中间层特征图		互信息	提出知识转移信息论框架，将知识转移表述为师生模型的互信息最大化
文献 [52]	中间层特征图	margin-ReLU	dp	提出新的激活函数来保存教师模型信息，引入新距离度量用于特征蒸馏
文献 [59]	概率分布输出中间层特征图	线性层	CE，L_2	鼓励同一卷积层的不同卷积核生成多样性特征图，缓解学生模型性能低下的问题
文献 [56]	联合分布特征中间层特征图概率分布输出	线性层	CE，KL	将联合标签增强自蒸馏方法迁移到传统蒸馏，缓解自监督对比学习对原始类表征的损害
文献 [34]	倒数第 2 层特征概率分布输出	线性层	L_2，CE	提出将表示学习和分类解耦，利用教师分类器辅助学生模型训练，改善学生模型的特征表示
文献 [60]	金字塔网络特征图	注意力聚合	L_2	提出跨阶段连接路径，学生模型高阶层学习教师模型低层特征以提高模型的性能表现
文献 [58]	倒数第 2 层特征教师模型分类器	线性层	L_2	提出对齐师生模型倒数第 2 层特征并直接复用教师模型分类器来降低蒸馏算法的复杂性
文献 [42]	中间层特征图	相似加权线性层	L_2	利用教师模型对学生模型特征重建，解决模型架构差异带来的语义偏差
文献 [43]	金字塔网络特征图	随机掩码、线性层	L_2	利用掩码图像机制进行知识传递，提高教师模型知识的完整性和有效性
文献 [44]	类别注意力图	归一化层	L_2	提出类别注意力图来进行知识的迁移，缓解蒸馏可解释性差的问题
文献 [45]	中间层频率特征	线性层	L_1	通过离散小波变换获得频率域特征，缓解空间域连续下采样的信息丢失

注：L_1和L_2表示曼哈顿距离和欧氏距离，KL 和 CE 表示 KL 散度和交叉熵，MMD 表示最大均值差异，dp[52] 表示修正的欧氏距离。

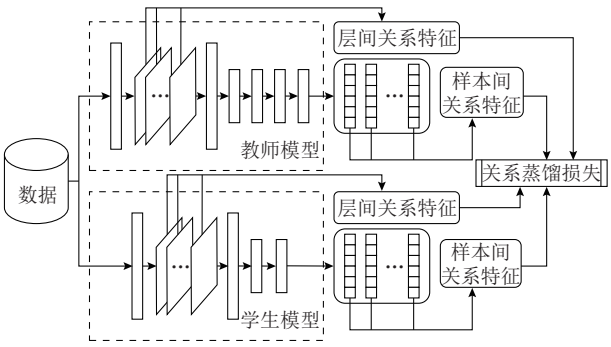


Fig. 8 Illustration of distillation method based on relational features

图 8 基于关系特征的蒸馏方法示意图

对于样本间关系的研究，Chen 等人[66] 提出了样本交叉相似性，利用L_2范数给出一个样本相似性的打分函数，根据相似性分数来选择不同的蒸馏策略。Liu 等人[67] 引入了图结构来传递样本间的实例关系，将实例特征和实例关系分别作为关系图的点和边，有效地捕获了整个网络中的知识。Peng 等人[68] 不仅考虑了样本内容的一致性，还考虑了样本关系的一致性，引入核方法来捕获样本之间的相似性。Tung 等人[69] 认为相似的样本会引发神经网络中相似的激活模式，提出利用样本相似性向学生模型传递教师模型的激活模式。Tian 等人[57] 引入了对比损失来表达样本之间的依赖关系，通过最大化教师模型和学生模型特征之间的互信息下界来优化学生模型。Xu 等人[70] 引入了词袋模型来聚合相似的数据，通过样本内蒸馏损失和样本间蒸馏损失来最小化师生模型距离并传递词袋知识。

基于关系特征的蒸馏方法汇总如表 4 所示。

2.1.4 小 结

本节介绍了 3 种知识类型。基于输出特征的蒸馏的优势在于结构简单、易于学习且不受模型架构的限制，但这种简单的结构难以为学生模型提供充足的知识。如何从输出中挖掘更有价值的信息是未来研究的一大趋势。一类方法通过调整温度系数来为学生模型带来更多的知识，这将在 4.3 节中详细介绍。基于中间特征的蒸馏相比于基于输出特征的蒸馏，知识类型更加丰富，能从中间层提取更复杂的知识传递给学生模型。然而，对于架构不同的师生模型，从中间层提取的知识可能存在维度和语义不匹配的问题。现有方法大多通过引入额外的投射层来进行语义对齐，但这会带来额外的计算参数。如何在不引入额外开销的情况下完美对齐师生模型的语义维度特征，是一个值得深入研究的问题。此外，从中间层提取知识的位置以及学生模型接收知识的位置也值得进一步研究。基于关系特征的蒸馏通过特

Table 4 Summary of Distillation Methods Based on Relational Features

表 4 基于关系特征的蒸馏方法汇总

方法来源	知识的来源	相似性知识	距离度量	关键技术
文献 [61]	中间层特征图	L_2	L_∞	利用不同样本的中间层特征之间的相对差异作为知识，提高知识的多样性
文献 [53]	中间层特征图	哈达玛乘积	L_2	提出将教师模型求解问题流程抽象为关系矩阵来作为知识提高学生模型的性能
文献 [62]	Logits 输出 中间层特征图	Logits 图表示图	EMD MMD	利用教师模型 Logits 输出特征和中间特征构造图结构，解决特征之间互补性的缺失问题
文献 [66]	样本特征向量	L_2	KL	提出一种新的知识类型，给出样本相似性打分函数并根据相似性来进行蒸馏
文献 [67]	不同层样本 特征向量	实例关系图	L_2	利用实例特征构造图关系知识，解决由于师生模型架构不同 导致的性能下降和类间差异性低的问题
文献 [68]	样本特征向量	核方法	L_2	考虑样本内容一致性和样本关系一致性，解除师生模型的实例强约束限制
文献 [69]	中间层特征图	转置相乘	L_2	利用样本相似性向学生模型传递教师模型的激活模式，推动知识获取的发展
文献 [57]	样本特征向量	互信息	最大互信息	引入对比损失表达样本关系，处理先前方法忽略教师模型结构信息的问题
文献 [63]	中间层特征图	转置相乘	MSE	利用相似度矩阵和注意力机制捕获中间层语义关联，解决手动关联不匹配问题
文献 [64]	中间层特征图	转置相乘	L_2	提出通道间相似性知识蒸馏，解决教师模型特征多样性不足的问题
文献 [70]	样本特征向量	余弦相似度	InfoNCE	提出词袋聚合蒸馏传递关系知识，解决基于对比学习方法优化困难的问题
文献 [65]	概率分布输出 中间层特征图	转置相乘	KL L_2	从样本实例、批次和类别角度构造关系知识，解决基于概率 分布输出蒸馏方法性能欠佳的问题

注：L_2，L_∞分别表示欧氏距离和切比雪夫距离，EMD，MMD，MSE，KL 分别表示推土机距离、最大均值差异、均方误差、KL 散度， InfoNCE[27] 表示一项对比损失。

征图或样本特征之间的关系向学生模型传递知识。现有方法大多通过样本之间或层之间的相似性来提取关系知识，然而，如何获得更复杂且包含更多语义知识的关系知识来帮助学生模型学习，仍然是一个具有挑战性的问题。图结构在基于关系特征的蒸馏中是一种常用的数据结构，刘静等人[71]的一项关于图知识的蒸馏综述在这一部分具有一定参考价值。

2.2 师生模型架构

随着 Transformer 架构模型的出现，蒸馏方法不再局限于卷积架构模型。本节从教师模型和学生模型的架构角度进行分类，CNN 和 ViT[15] 是 2 种主流的视觉网络框架，它们分别应用于教师模型和学生模型，共有 4 种组合：CNN2CNN，CNN2ViT，ViT2CNN，ViT2ViT。具体而言，这 4 种组合分别表示：卷积架构的教师模型向卷积架构的学生模型进行蒸馏、卷积架构的教师模型向 ViT 架构的学生模型进行蒸馏、ViT 架构的教师模型向卷积架构的学生模型进行蒸馏，以及 ViT 架构的教师模型向 ViT 架构的学生模型进行蒸馏。

2.2.1 CNN2CNN

CNN2CNN 是指卷积架构的教师模型向卷积架构的学生模型进行蒸馏，通常为一个较深且性能较好的卷积网络教师模型来指导一个较浅且较为轻量的卷积网络学生模型训练，卷积网络经过近年来的爆炸式发展，已经广泛应用于各个领域并取得显著的成果，因此在知识蒸馏中卷积架构到卷积架构蒸馏的研究占到了绝大多数，2.1 节总结的基于输出特征的蒸馏[21,32,36]、基于中间特征的蒸馏[40-41,48] 以及基于关系特征的蒸馏[53,69,72] 大多数均以卷积架构师生模型为基础。

2.2.2 CNN2ViT

CNN2ViT 是指卷积架构的教师模型向 ViT 架构的学生模型进行蒸馏，通常利用卷积网络前期的归纳偏差来帮助 ViT 架构的网络模型训练，如图 9 所示，蒸馏损失可以描述为

$$L_{\mathrm{KD}}=\alpha L_{\mathrm{Dis}}(\rho(\boldsymbol{F}^{\mathrm{T}}),\boldsymbol{v}^{\mathrm{S}})+\beta L_{\mathrm{Dis}}(\boldsymbol{z}^{\mathrm{T}},\boldsymbol{q}^{\mathrm{S}}),\tag{14}$$

其中$\boldsymbol{F}^{\mathrm{T}}$表示卷积架构教师模型的中间层特征，$\boldsymbol{v}^{\mathrm{S}}$表示 ViT 架构学生模型的 Transformer 块特征，$\rho(\cdot)$表示特征转换层，用于对齐教师模型的卷积特征和学生模型的 Transformer 块特征。$\boldsymbol{z}^{\mathrm{T}}$和$\boldsymbol{q}^{\mathrm{S}}$分别表示卷积架构教师模型的 Logits 输出特征和 ViT 架构学生模型的分类标记输出特征，两者的特征维度均为类别数量。α和β为超参数，用于平衡两者的权重。

ViT[15] 在没有大规模数据的情况下不能很好地泛化，Touvron 等人[17] 在尾部额外添加一个蒸馏 token 来接收来自 CNN 教师模型的知识。Ren 等人[73] 对 Touvron 等人提出的 DeiT[17] 做出改进，认为不同教师模型具有不同的归纳偏差，提出引入多个 token 分别接收卷积网络的知识和内卷积网络[74] 的知识。Zhao 等人[75] 认为 CNN 后期的局部特征直接被使用会损害 ViT 的优化，将 CNN 知识通过池化层分为局部知识和全局知识，在训练前期通过局部知识充分利用

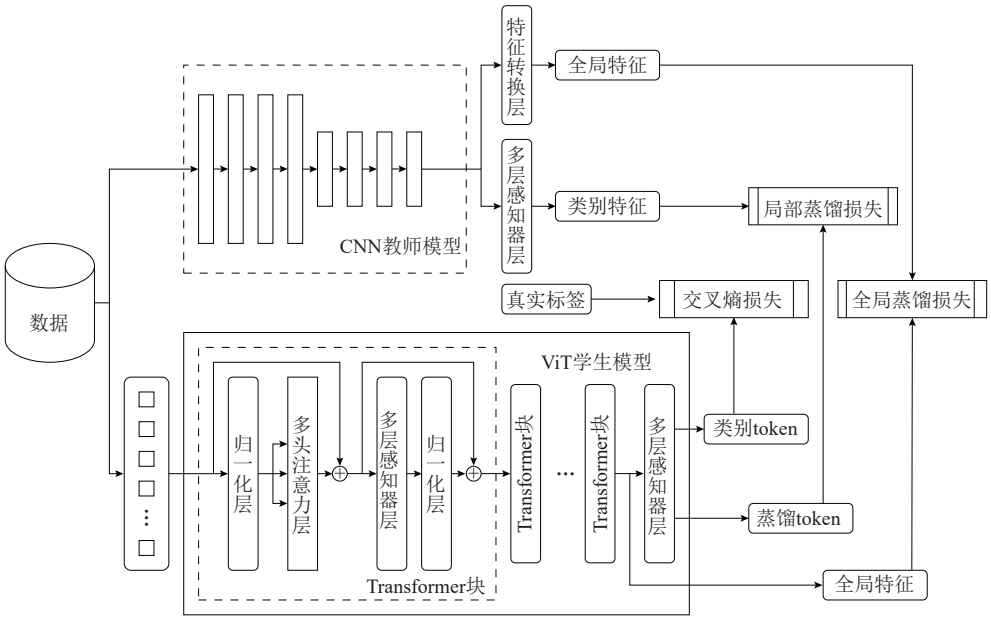


Fig. 9 Illustration of CNN2ViT method

图 9 CNN2ViT 方法示意图

归纳偏差的益处，在后期利用全局特征继续向 ViT 输送知识。

2.2.3 ViT2CNN

ViT2CNN 是指 ViT 架构的教师模型向卷积架构的学生模型进行蒸馏，如图 10 所示，蒸馏损失可以描述为

$$L_{\mathrm{KD}}=\alpha L_{\mathrm{Dis}}(\boldsymbol{v}^{\mathrm{T}},\phi(\boldsymbol{F}^{\mathrm{S}}))+\beta L_{\mathrm{Dis}}(\boldsymbol{q}^{\mathrm{T}},\boldsymbol{z}^{\mathrm{S}}),\tag{15}$$

其中 $\boldsymbol{v}^{\mathrm{T}}$ 表示 ViT 架构教师模型的 Transformer 块特征，$\boldsymbol{F}^{\mathrm{S}}$ 表示卷积架构学生模型的中间层特征，$\phi(\cdot)$ 表示特征转换层，用于对齐教师模型的 Transformer 块特征和学生模型的卷积特征。$\boldsymbol{q}^{\mathrm{T}}$ 和 $\boldsymbol{z}^{\mathrm{S}}$ 分别表示 ViT 架构教师模型的分类标记输出特征和卷积架构学生模型的 Logits 输出特征，两者的特征维度均为类别数量。α 和 β 为超参数，用于平衡两者的权重。

Liu 等人[76] 致力于将 Transformer 的知识传递给 CNN 模型，主要解决异构模型的特征语义不匹配问题，提出部分交叉注意力投射层和分组线性投射层来映射学生模型特征使之与教师模型特征更相似，并引入多种数据增强手段来提升学生模型的鲁棒性。

Zhao 等人[77] 通过将像素级特征映射到具有统一感受野的局部特征中来解决跨架构师生模型的特征对齐问题；Hao 等人[78] 将学生模型中间特征映射到 Logits 空间，来解决异构模型的特征对齐问题；Li 等人[79] 通过引入助理模型来进行跨架构的蒸馏，弥合异构模型之间的特征差异；Ni 等人[80] 引入多个具有不同归纳偏差的学生模型进行相互学习，来缓解跨

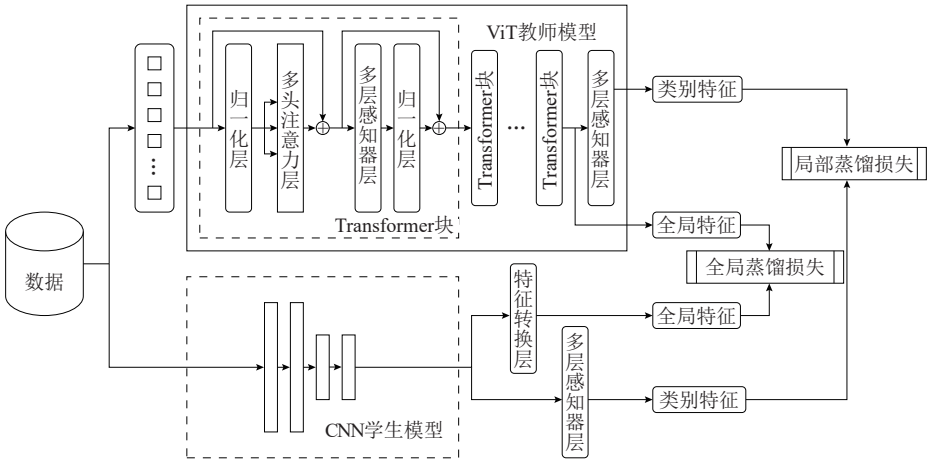


Fig. 10 Illustration of ViT2CNN method

图 10 ViT2CNN 方法示意图

架构蒸馏的性能问题; Zheng 等人[81]通过一个特征适应模块来处理教师特征, 让其与异构学生模型特征对齐。

2.2.4 ViT2ViT

ViT2ViT 是指 ViT 架构的教师模型向 ViT 架构的学生模型进行蒸馏, 将知识从较大的 ViT 架构教师模型转移到较小的 ViT 架构学生模型中, 如图 11 所示, 蒸馏损失可以描述为

$$L_{\mathrm{KD}}=\alpha L_{\mathrm{Dis}}(\boldsymbol{v}^{\mathrm{T}},r(\boldsymbol{v}^{\mathrm{S}}))+\beta L_{\mathrm{Dis}}(\boldsymbol{q}^{\mathrm{T}},\boldsymbol{q}^{\mathrm{S}}),\tag{16}$$

其中 $\boldsymbol{v}^{\mathrm{T}}$ 和 $\boldsymbol{v}^{\mathrm{S}}$ 分别表示 ViT 架构的教师模型和学生模型的 Transformer 块特征, $r(\cdot)$ 表示特征映射层。$\boldsymbol{q}^{\mathrm{T}}$ 和 $\boldsymbol{q}^{\mathrm{S}}$ 分别表示 ViT 架构教师模型和学生模型的分类标记输出特征。α 和 β 为超参数, 用于平衡两者的权重。

Wu 等人[82]将教师模型的推理与学生模型的蒸馏分离, 提出稀疏的特征编码和数据增强编码来存储教师模型的推理结果以进行快速蒸馏, 从预训练的知识中获得性能提升。Zhang 等人[83]发现直接将权重共享技术应用到 ViT 中会带来训练上的不稳定和性能下降的问题, 提出使用权重复用和权重蒸馏, 前者通过在多头注意力前后插入线性层和在多层感知器层中插入转换模块来增加参数的多样性, 后者利用概率分布输出、自注意力 QKV 矩阵和多层感知器层的关系矩阵进行蒸馏来保证性能具有竞争性。Yang 等人[84]认为不同的 ViT 层具有不同的特征分布, 提出对浅层使用基于模仿的线性特征蒸馏, 对于深层特征使用基于掩码的生成式蒸馏。Zhao 等人[85]认为对多头注意力的 1 对 1 蒸馏忽略了语义的一致性, 对于不同数量的注意力头不能完美匹配, 提出了注意力头的对齐蒸馏, 利用教师模型注意力头与学生模型注意力头之间的余弦相似度, 每一个教师模型注意力头指导所有的学生模型注意力头, 并通过相似性来控制权重。

掩码图像建模(masked image modeling, MIM)具有强大的特征表征能力, Ren 等人[86]通过蒸馏技术将 MIM 的知识转移到较小的学生模型中, 研究发现对自注意力模块的 QKV 矩阵的 Q-K, V-V 关系矩阵进行蒸馏效果最好。Lin 等人[87]发现 ViT 在小样本的设定下容易过拟合且缺乏类似 CNN 模型归纳偏置的机制, 有严重的性能下降问题, 利用分类 token 的交叉熵和 MIM 损失进行自监督训练, 并通过余弦相似度进行跨类内图像块蒸馏学习。Bai 等人[88]提出了直接从 MAE[89]预训练模型中提取知识, 将掩码率提升到 95% 并利用曼哈顿距离匹配中间特征来训练学生模型。Zhao 等人[90]认为 Bai 等人提出的 DMAE[88]对教师模型的高掩码率会丢失重要信息, 提出让教师模型拥有更多的可见块, 学生模型的可见块是教师模型的子集, 通过直接对齐和生成式对齐来处理师生模型中均可见的块和仅在教师模型中可见的块。

2.2.5 小 结

涉及 ViT 架构的蒸馏方法总结如表 5 所示, InfoNCE[27]表示一项对比损失。基于 Transformer 的视觉模型由于其强大的性能, 在多个领域都有重要应用。对于异构架构的蒸馏, 大多数工作集中在利用卷积网络的归纳偏差来辅助 ViT 的训练, 但很少

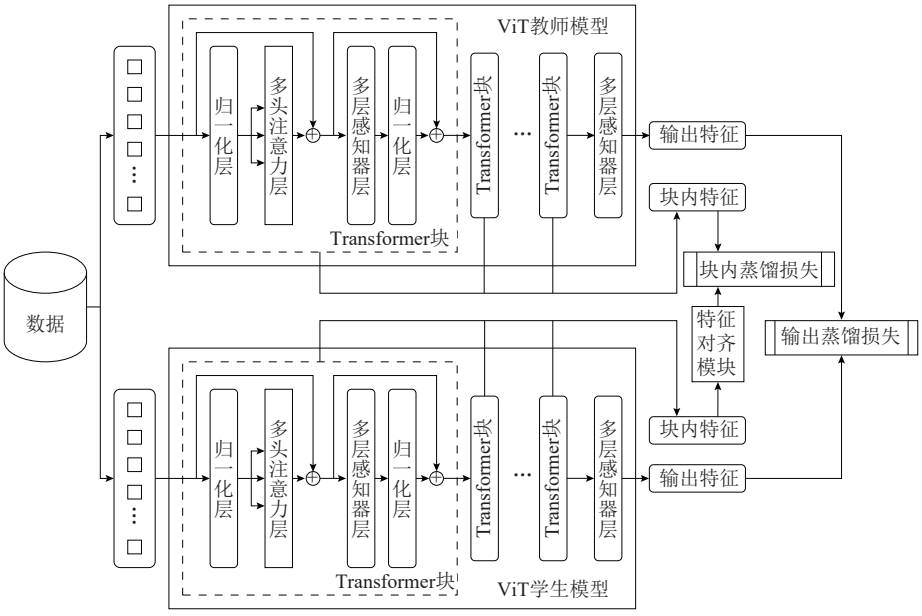


Fig. 11 Illustration of ViT2ViT method

图 11 ViT2ViT 方法示意图

Table 5 Summary of ViT Architecture Relative Distillation Methods

表 5 ViT 架构相关蒸馏方法汇总

方法来源	知识的来源	知识的转换	架构	距离度量	关键技术
文献 [17]	CNN 概率分布输出	蒸馏 token	CNN→ViT	KL	引入一个蒸馏 token 接收卷积教师模型的知识，解决 ViT 在数据不足时的泛化性问题
文献 [76]	QKV 矩阵	线性层	ViT→CNN	L_2	提出交叉注意力分组线性映射对齐师生模型特征，缓解跨架构蒸馏的语义对齐问题
文献 [73]	CNN 概率分布输出、INN 概率分布输出	蒸馏 token	CNN→ViT	KL	引入多个蒸馏 token 接收卷积网络和内卷网络的知识，探究了归纳偏差的作用
文献 [82]	分类 token	稀疏编码	ViT→ViT	CE	将教师模型推理与蒸馏分离，缓解资源有限设备上的局限性
文献 [83]	分类 token、QKV 矩阵、MLP 输出	自注意力矩阵	ViT→ViT	CE	提出权重复用和权重蒸馏，缓解在资源有限设备上的适用性
文献 [84]	MLP 输出	线性层、掩码	ViT→ViT	L_2	提出了一种基于特征的 ViT 蒸馏方法，缓解 CNN 架构的蒸馏方法无法应用到 ViT 架构上的问题
文献 [75]	CNN Logits 输出	全局池化层	CNN→ViT	KL	提出累计空间知识蒸馏，缓解跨架构语义对齐和 CNN 知识抑制模型后期训练的问题
文献 [86]	QKV 矩阵	自注意力矩阵	ViT→ViT	KL	将 MIM 的知识转移到较小的模型中，解决小模型只能从预训练方法中获益的问题
文献 [87]	分类 token、图像掩码块	无	ViT→ViT	CE	通过分类 token 和 MIM 损失进行自监督训练，缓解小样本下 ViT 的性能损失问题
文献 [88]	MLP 输出	线性层	ViT→ViT	L_1	利用 MAE 预训练模型提取知识，缓解蒸馏计算效率的问题
文献 [90]	MLP 输出	线性层 MHA	ViT→ViT	L_1	提出非对称掩码蒸馏，缓解教师模型高掩码率信息丢失问题
文献 [77]	MHA 特征图	自注意力矩阵	ViT→CNN	MSE	将像素级特征映射到具有统一感受野的局部特征中，解决师生模型特征的对齐问题
文献 [85]	MHA 特征图	余弦相似度	ViT→ViT	L_1	提出多头注意力对齐策略，解决师生模型输出头对齐问题
文献 [78]	MLP 输出 中间层特征图	特征投影层	ViT→CNN CNN→ViT	KL	提出通过将特征映射到 Logits 空间，解决异构模型的特征对齐问题
文献 [79]	MLP 输出中间层特征图	助理模型	ViT→CNN CNN→ViT	InfoNCE KL	引入一个辅助模型来促进异构教师模型和学生模型之间的特征传递
文献 [80]	MHA 特征、分类 token	线性层	ViT→CNN	KL CE	通过多个不同归纳偏差的学生模型相互学习，来缓解跨架构蒸馏的性能问题
文献 [81]	MLP 输出	特征适应模块	ViT→CNN	L_2	提出特征适应块处理教师模型特征，缓解异构特征的不匹配

注：KL，CE，MSE，L_1，L_2分别表示 KL 散度、交叉熵、均方误差、曼哈顿距离和欧氏距离。

关注将 ViT 中丰富的知识转移到更普适的卷积网络中。将 ViT 强大的全局建模能力转移到卷积网络中，是一个值得深入研究的课题。

3 知识蒸馏的学习方式

知识蒸馏的学习方式通常可以根据训练过程的不同分为离线蒸馏、在线蒸馏、自蒸馏、无数据蒸馏、多教师蒸馏和助理蒸馏。其中离线蒸馏是最常见的学习方式，通过预先训练好的教师模型来训练学生模型；在线蒸馏允许师生模型同时更新；在自蒸馏方法中，模型本身同时作为教师模型和学生模型，降低了计算和存储的开销；无数据蒸馏可以在不访问原始数据的情况下完成对学生模型的训练；多教师蒸馏打破了单一教师模型的限制，通过多个教师模型共同训练学生模型；助理蒸馏通过引入一个能力介于师生模型之间的助理模型来帮助学生模型更好地优化。各种学习方式之间并非相互独立，某些方法可以同时归属于多种学习方式。以上学习方式示意图如图 12 所示。

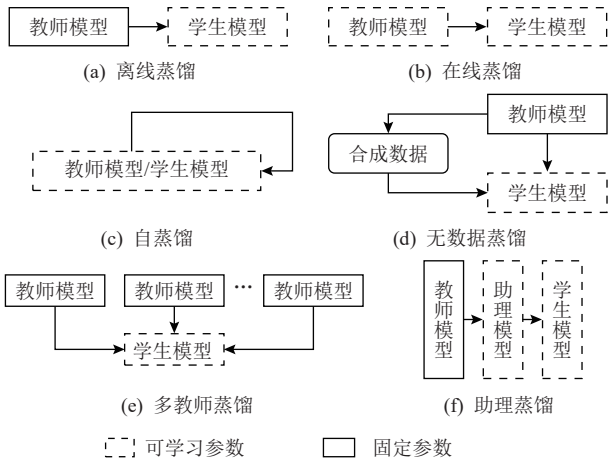


Fig. 12 Style of learning of knowledge distillation

图 12 知识蒸馏的学习方式

3.1 离线蒸馏

离线蒸馏中，教师模型和学生模型的训练过程是分开的，首先在蒸馏学生模型之前先在数据集上训练教师模型，然后利用训练好的教师模型中提取

的知识来指导较小的学生模型训练。离线蒸馏的方式简单易行，师生模型的训练可以分开进行，并且在蒸馏时仅需关注学生模型的学习，如图 12 (a)所示；但学生模型只能从教师模型学习到固定的知识，无法适应动态的变化，并且学生模型的能力极大地依赖于教师模型提供的知识。为应对不同任务场景的需求，离线蒸馏逐步发展出标准蒸馏和跨任务及跨模态蒸馏 2 个研究方向，分别聚焦于单一任务内的知识高效传递与跨任务跨模态知识迁移的优化。

1)标准蒸馏。大多数蒸馏方法采用离线方式进行，其第 1 阶段通常不被纳入知识蒸馏的讨论范围。离线蒸馏的重点在于知识的有效转移，例如，Hinton 等人[21]通过软化预测标签并引入温度系数，增强置信度较低标签的影响，以转移输出特征知识；Romero 等人[40]通过线性层对齐教师模型与学生模型的特征层，实现中间特征知识的迁移；Yim 等人[53]则提取教师模型的关系矩阵，反映问题求解流程中的关系特征，从而完成关系特征知识的转移。

2)跨任务及跨模态蒸馏。一些工作利用离线蒸馏进行跨任务和跨模态学习，通过其他模态任务的数据来提升下游任务性能。Ye 等人[91]提出分步蒸馏，先通过对比蒸馏增强特征提取器的判别性，再利用相似性蒸馏优化分类器，重用跨任务模型知识；Lê 等人[92]探讨有限数据下目标检测与语义分割的跨任务学习，通过蒸馏技术缓解数据不足问题；Li 等人[93]通过将教师模型的知识应用于多种任务场景，有效缓解了不同标签空间的约束问题，提升了学生模型的泛化能力；Auty 等人[94]利用反向投影，将在不同任务中训练的教师模型知识迁移至目标任务的学生模型中，从而实现知识的有效传递；Yuan 等人[95]将图像语义理解任务模型中的知识蒸馏到文本图像生成任务中，确保了生成图像与文本语义的一致性；Bayramli 等人[96]利用注意力块实现语义分割教师模型向场景流学生模型的知识迁移；Wang 等人[97]提出双向模态重建模块进行跨模态蒸馏，利用标注丰富的模态补充标注稀缺模态的数据。

3.2 在线蒸馏

在没有大容量高性能的教师模型或教师模型与学生模型能力差距很大时，学生模型的性能会受到影响，为了进一步提升学生模型的性能，在线蒸馏被提出。在线蒸馏中教师模型和学生模型同时更新，如图 12(b)所示，在线蒸馏的典型模式有互学习、共享学习和协同学习[23,25]。

1)互学习。Zhang 等人[30]在 2017 年提出了互学习的概念，让多个模型之间互相学习其概率分布输出来提高模型性能。为了缓解师生模型能力差距的影响，Qian 等人[98]通过自适应切换阈值来让师生模型角色互换和互相学习。Yang 等人[99]提出一个对比互学习框架，在一组模型之间进行交互和对比分布的传递，以增强模型的特征表示能力。Jacob 等人[100]提出自适应特征蒸馏和在线任务加权方案为每个任务选择训练层来缓解多任务学习中准确率受损的问题。Rao 等人[101]通过在教师模型中插入可学习的适配器模块进行更新，来缓解师生模型预测分布的尖锐度差距。

2)共享学习。Anil 等人[102]2018 年提出了共蒸馏，每个子网在数据的不相交的子集上训练，迁移其他模型的知识来互相学习。Zhou 等人[36]提出一个共享低级别参数的师生模型架构的蒸馏框架。

3)协同学习。Song 等人[103]提出了协同学习，在相同数据上训练多个分类器，来提高模型的泛化能力。Chen 等人[104]通过多个辅助网络和 1 个主网络进行 2 级蒸馏来提高模型多样性；而 Gong 等人[105]提出自适应层级分支融合在线蒸馏框架来提高模型的多样性，缓解多学生模型集成的同质化问题。Wu 等人[106]构建一个多分支网络，将分支随机增强后与另一个分支分类器集成构成教师模型相互学习。Li 等人[107]将学生模型划分为领导学生模型和普通学生模型，领导学生模型指导普通学生模型训练，并且普通学生模型的知识能够整合提炼给领导学生模型来提高模型的整体能力。

3.3 自蒸馏

自蒸馏不同于传统的知识蒸馏，其教师模型和学生模型都是同一个模型，如图 12(c)所示，自蒸馏通过利用模型自身的预测结果或中间特征来提高其模型性能。

Zhang 等人[108]在 2019 年提出自蒸馏的概念，通过特征相应的 KL 散度和中间层特征的对齐损失来将模型深层的知识转移到浅层。Lan 等人[109]有相似的思想，将训练中模型发现的知识提炼回自身来规范后续学习过程。与之相反，Yang 等人[110]提出将网络训练前期的知识转移到网络训练后期，来实现同一网络的监督训练。Mobahi 等人[111]对自蒸馏进行理论分析，发现自蒸馏迭代逐渐限制了用于表示解决方案的基函数的数量。自蒸馏方法存在即时信息丢失、额外内存消耗和并行化程度低的问题，Shen 等人[112]提出了利用上一批次输出结果进行蒸馏，来保持模型的一致性。Gou 等人[113]提出学生模型可以通

过自我反思，进行更有效的训练，通过自蒸馏将前几层的知识蒸馏到后几层中。余鹰等人[114]提出一种多层次知识自蒸馏联合多步训练方法，将特征学习和知识自蒸馏分步进行，解决了单步训练网络优化难的问题。

3.4 无数据蒸馏

Lopes 等人[115]在 2017 年提出无数据蒸馏的概念，如图 12（d）所示，现有蒸馏方法需要访问原始训练数据，随着数据集规模的增大其分布可能变得越来越复杂，并且许多数据集涉及隐私或安全问题（生物特征、敏感数据）使得数据共享和数据采集变得困难，提出通过利用教师模型的层级或信息重建训练学生模型的数据。与之思想类似的工作，Nayak 等人[116]提出通过狄利克雷分布对 softmax 函数进行建模来构建样本数据。针对无数据蒸馏训练的稳定性和效率问题，Fang 等人[117]通过重用训练数据中的共同特征来提高数据合成的效率。Yu 等人[118]设计了通道级特征交换和多尺度空间激活来解决合成样本的多样性和效率问题。Li 等人[119]提出了一种动态生成由易到难伪样本的方法来解决无数据蒸馏训练不充分及不稳定的问题。Wang 等人[120]探究了师生模型之间的性能差距与合成样本质量之间的关系，提出了一种差距敏感的样本生成方法。张晶等人[121]建立教师模型与学生模型生成器并同时优化，实现网络任务与优化目标一致，提升学生模型判别性能，并利用教师模型潜在分布先验信息优化生成器，保证学生模型识别准确率并提升泛化性。

Chen 等人[122]在 2019 年提出利用生成对抗网络合成数据样本来训练学生模型，推动了接下来一系列基于生成对抗网络（generative adversarial nets, GAN）[123]工作的发展[124-127]。Do 等人[128]提出动量对抗蒸馏方法来缓解生成器合成数据偏移的问题。Zhao 等人[129]提出通过双判别器对抗蒸馏来获得教师模型的内在统计信息和与学生模型的最大差异，从而更好地生成样本。Shao 等人[130]认为现有的合成方法无法区分不同类别的样本分布，引入条件 GAN 来合成特定类别的多样性样本。Patel 等人[131]认为伪标签在多个生成器更新下为非平稳分布，准确性受到影响，在知识保持和任务获取之间设置一个对齐因子来缓解该问题。

3.5 多教师蒸馏

多教师蒸馏的概念在 2017 年由 You 等人[61]提出，旨在通过多个教师模型为学生模型提供更全面的指导，如图 12（e）所示，通过平均教师模型的软标签并通过一种投票策略筛选多个教师模型的中间特征，复合多个教师模型的知识来指导学生模型训练。他们的工作引发了后续一系列工作对于多教师蒸馏方法的研究[31,38,102,132-135]。

Son 等人[37]借鉴 DenseNet[8]的思想将上级教师模型的输出知识始终作为下级教师模型的指导信息来缓解教师模型预测的错误累计，并借鉴 Dropout 的思想提出随机丢弃来缓解过拟合。Shi 等人[136]利用学生模型学习效果的反馈来关注学生模型对不同教师模型知识的偏好，提高知识转移的效率。Pham 等人[137]融合量化的方法，鼓励量化教师模型之间的协作学习以及量化教师模型与量化学生模型之间的相互学习。针对教师模型可能出现的低质量预测，Zhang 等人[138]利用真实标签为每个教师模型分配预测置信度，提高教师模型软标签的可靠性。Cheng 等人[139]将多教师蒸馏损失解耦，通过预测熵为每个教师模型分配权重，缓解低质量预测带来的负面影响。

近期部分工作聚焦于跨模态和跨任务的知识转移，通过不同来源的教师模型协作提升学生模型性能。Klingner 等人[140]将基于激光雷达的 3D 目标检测教师模型的跨模态知识蒸馏用于多相机 3D 目标检测学生模型，并通过预训练的实例分割教师模型进行跨任务蒸馏。Zhao 等人[141]设计 4 个蒸馏模块，从激光雷达点云和多视角相机检测器到雷达点云和多视角相机检测器，实现高效跨模态蒸馏。

3.6 助理蒸馏

由于教师模型和学生模型能力差距过大，会影响知识蒸馏的性能，其中一种解决方案是引入一个能力介于教师模型和学生模型之间的助理模型来弥合性能鸿沟[38]。该思想由 Wang 等人[142]提出，利用助理模型作为 GAN 的判别器，来缩减教师模型和学生模型产生特征之间的差异。我们将此类方法称之为助理蒸馏，如图 12（f）所示。

Passalis 等人[143]训练了一个助理模型来建立师生模型各层之间 1 对 1 的匹配关系；Ganta 等人[144]实现了多个助理模型的集成，实现了更好的蒸馏性能；Gao 等人[145]引入一个助理模型来学习教师模型和学生模型之间的残差来弥合其学习能力的差异；Son 等人[37]引入多级助理模型，并利用高层助理模型和教师模型来同时指导低层助理模型和学生模型训练，缓解错误累计的问题；Mirzadeh 等人[38]在训练助理模型时只考虑了输出特征；Morikawa 等人[146]结合了输出特征和中间层特征来训练助理模型；He 等人[147]引入了在线学习，让助理模型和学生模型相互学习，

提升了学生模型的表现；Zhao 等人[148]微调教师模型来构造自适应助理模型，提供多样化知识并缩小师生模型之间的能力差距；Li 等人[79]通过助理模型来进行跨架构的蒸馏，弥合异构模型之间的特征差异；Huo 等人[149]提出了一种动态知识选择策略，从非目标类别中提取有效的跨模态知识，并通过助理模型逐步传递这些知识，增强跨模态学习效果。

3.7 小 结

本节介绍了 6 种常见的学习方式，针对各种学习方式的蒸馏方法汇总于表 6 所示。离线蒸馏方法简单易行，但仅能学习到教师模型的固定知识，灵活性不强；在线蒸馏中，教师模型也参与参数更新，尽管能带来更大的性能提升，但引入了额外的计算开销；自蒸馏方法中，教师模型和学生模型为同一模型，计算和存储开销较小，但知识结构相对单一；无数据蒸馏能够很好地保护数据隐私，但由于缺乏数据集支持，在复杂场景中性能较差；多教师蒸馏由多个教师模型提供知识，知识结构较为丰富，但特征融合较难且会引入额外的开销；助理蒸馏能够在一定程度上弥合师生模型之间的性能差异，但助理模型的训练也引入了额外的计算开销。每种方式各有优劣，在实际应用中应根据具体场景选择适当的方法。

4 知识蒸馏的优化

第 2 节和第 3 节归纳的方法大多是基于 Hinton 等人[21]提出的经典蒸馏范式。近年来，随着知识蒸馏的发展，出现了一些研究打破了传统固定模式，对学习范式进行了改进和优化。本文将这类方法总结为知识蒸馏的优化，如图 13 所示，分为 4 个类别：蒸馏过程的优化、知识结构的优化、温度系数的优化和损失结构的优化。蒸馏过程的优化主要针对师生模型特征的对齐和训练过程的调整；知识结构的优化汇总了对知识的调整和筛选的方法，以高效地向学生模型传递知识；温度系数的优化打破了先前固定系数的模式，总结了动态温度的设置方法；损失结构的优化汇总了对损失函数的解耦方法，解决了损失内部的冲突博弈。此外，本节还总结了利用其他技术对知识蒸馏方法进行优化的研究，包括对抗性技术、自动机器学习、强化学习和扩散模型。

4.1 蒸馏过程优化

知识蒸馏训练过程的优化主要针对于特征对齐模块的改进以及蒸馏学习过程的调整，来获得更好的学生模型。Cho 等人[150]发现蒸馏损失后期会损害模型的训练，提出停止方案以阻止后期蒸馏损失对模型性能的削弱。Mirzadeh 等人[38]发现师生能力差距过大会导致学生模型的性能下降，提出引入一个介于师生模型之间的助理模型来帮助学生模型学习，弥合性能的差距。Yuan 等人[151]证明了用较弱的学生模型指导教师模型或训练不佳的教师模型指导性能更佳的学生模型同样能够带来性能的提升。Stanton 等人[152]探究影响蒸馏性能的原因，对学生模型容量、网络架构、数据集规模、复杂性、数据域以及优化方法进行了大量实验，认为优化方法限制了模型的收敛。He 等人[153]认为在师生模型之间添加可学习模块会损害表示学习的性能，提出了一种非参数的对齐方式，通过奇异值分解来完成师生模型的对齐。Shen 等人[154]认为针对蒸馏训练需要额外对教师模型进行推理，提出 FKD，在教师模型推理阶段保存数据增强方法及输出结果，在训练学生模型时直接复用来加快蒸馏的训练。Song 等人[155]更关注“从哪里蒸馏”，提出点自适应的蒸馏模块，自适应地确定每个样本在教师模型中的蒸馏位置。Zhang 等人[156]提出一个在线蒸馏框架，旨在寻找更平坦的最优解，提出使用混合权重模型配合间歇融合策略来帮助模型更好地收敛。Wang 等人[157]将学生模型中间特征转移到教师模型进行推理，与教师模型软标签进行蒸馏，学生模型最后几层只通过真实标签优化，通过这种方式缓解真实标签与教师模型的错误分类之间的冲突。

4.2 知识结构优化

知识蒸馏的知识结构优化指的是对教师模型提供的知识进行再抽取来减少信息的冗余或进行再调整来提升知识的传递效率，大大提高教师模型知识的有效性，从而能够训练得到性能更好的学生模型。

Xu 等人[158]认为学生模型并没有获得教师模型全面的知识，提出通过 top$K$ 选择性转移策略来更好地利用样本变换带来的暗知识。Li 等人[159]认为对所有的知识进行蒸馏是冗余的，将知识点分为强、中、弱 3 个档次，强价值知识直接蒸馏，中价值知识利用弱价值知识增强后再蒸馏。Huang 等人[160]认为学生模型应当关注教师模型的相对排名而不是准确的概率输出，引入类内相似性和类间相似性，并用皮尔森系数取代 KL 散度作为新的距离度量。Shen 等人[112]针对先前自蒸馏方法即时信息丢失、额外内存消耗和并行化程度低的问题，提出了利用上一批次输出结果进行蒸馏的方法来保持模型的一致性。Choi 等人[161]认为教师模型可能将错误信息传递给学生模型，

Table 6 Summary of Learning Methods of Knowledge Distillation

表 6 知识蒸馏的学习方式汇总

方法来源	学习方法	关键技术
文献 [115]	无数据学习	提出利用网络的元数据蒸馏，无需访问原始数据，解决数据隐私问题
文献 [30]	互学习	提出让多个模型互相学习其输出，在没有强大教师模型也能提高模型性能
文献 [102]	共享学习	使每个模型在不相交的数据子集上训练并迁移其他模型知识来降低复杂性
文献 [103]	协同学习	提出协同学习，在相同数据上训练多个分类器，提高模型的泛化能力
文献 [36]	共享学习	通过共享师生模型架构的低级别参数来降低推理所需要的时间
文献 [122]	无数据学习	提出一种利用 GAN 合成数据的方法，缓解蒸馏中的隐私问题
文献 [109]	自学习	提出将模型中发现的知识提炼回自身来避免训练计算开销大的教师模型
文献 [108]	自学习	提出自蒸馏的概念来缓解网络规模增加带来的计算存储成本
文献 [116]	无数据学习	将 softmax 函数建模为狄利克雷分布抽取教师模型数据分布，缓解数据隐私问题
文献 [110]	自学习	将模型训练早期的知识转移到模型训练后期来降低训练的计算资源消耗
文献 [104]	协同学习	提出多个辅助网络与主网络进行 2 级蒸馏，提高多样性来缓解学生模型同质化问题
文献 [111]	自学习	对自蒸馏进行理论分析，自蒸馏迭代逐渐限制表示解决方案的基函数数量
文献 [38]	助理学习	提出通过助理模型缓解师生模型的能力差距问题
文献 [143]	助理学习	通过助理模型建立师生模型层之间 1 对 1 的匹配关系，缓解师生模型的能力差距问题
文献 [106]	协作学习	构建分支网络，分支增强后与另一分支分类器集成后互相学习来构建高水平教师模型
文献 [144]	助理学习	集成了多个助理模型来缩减师生模型性能差异，实现了更好的蒸馏性能
文献 [145]	助理学习	提出通过助理模型学习师生模型之间的残差来弥合两者学习能力的差异
文献 [37]	助理学习	利用高层助理模型和教师模型指导低层助理模型和学生模型缓解错误累计问题
文献 [128]	无数据学习	提出动量对抗蒸馏来缓解合成数据的分布偏移问题
文献 [98]	互学习	通过自适应切换阈值互换师生模型角色，缓解师生模型能力差距带来的性能影响
文献 [129]	无数据学习	提出通过双判别器对抗蒸馏来解决数据不可用的问题
文献 [112]	自学习	通过利用上一次迭代结果更新自蒸馏模型来保持模型的一致性
文献 [130]	无数据学习	利用条件生成对抗网络改进合成样本的歧义性和多样性不足的问题
文献 [117]	无数据学习	提出利用重用共享的共同特征来解决数据合成效率低下的问题
文献 [146]	助理学习	同时考虑了输出和中间特征来训练助理模型，更好地缓解了师生模型之间的性能差异
文献 [101]	互学习	在教师模型中添加可学习适配器模块来缓解师生模型预测分布尖锐度差异
文献 [99]	互学习	提出对比互学习框架来从其他网络学习对比知识以增强特征表示能力
文献 [131]	无数据学习	提出一个元学习启发框架来解决生成的伪样本分布不稳定的问题
文献 [100]	互学习	提出自适应特征蒸馏和在线任务加权来缓解多任务学习中准确率受损问题
文献 [105]	协作学习	提出自适应分支融合框架提高模型多样性来缓解学生模型集成的同质化
文献 [113]	自学习	将模型前几层知识蒸馏到后几层，通过自我反思进行更有效的学习
文献 [118]	无数据学习	提出信道特征交换和多尺度区域一致性，处理数据合成在多样性和效率方面的局限性
文献 [119]	无数据学习	通过逐渐生成由易到难的伪样本，模拟人类学习方式解决学习不充分不稳定的问题
文献 [147]	助理学习	令助理模型和学生模型相互学习，通过离线在线联合蒸馏的方法缓解师生模型差异
文献 [148]	助理学习	构造自适应助理模型，提供多样化知识并缩小师生模型之间的能力差距
文献 [120]	无数据学习	利用师生模型的差异生成样本来解决无数据蒸馏中样本质量不理想的问题
文献 [79]	助理学习	通过助理模型来进行跨架构的蒸馏，弥合异构模型之间的巨大特征差异

提出分组训练的框架，学生模型组先反馈出一组易于出错的样本，经过数据增强后交给教师模型加强学习，利用教师模型和性能最好的学生模型一起指导学生模型组中的所有学生模型训练。Shen 提出 FerKD[162]，该方法是在他们的另一项工作 FKD[154] 的基础上提出自适应标签校正，丢弃非常简单和非常困难的样本，重点来优化中等难度的样本。Miles 等人[163] 专注于高效的知识转移，提出在 Stiefel 流形上

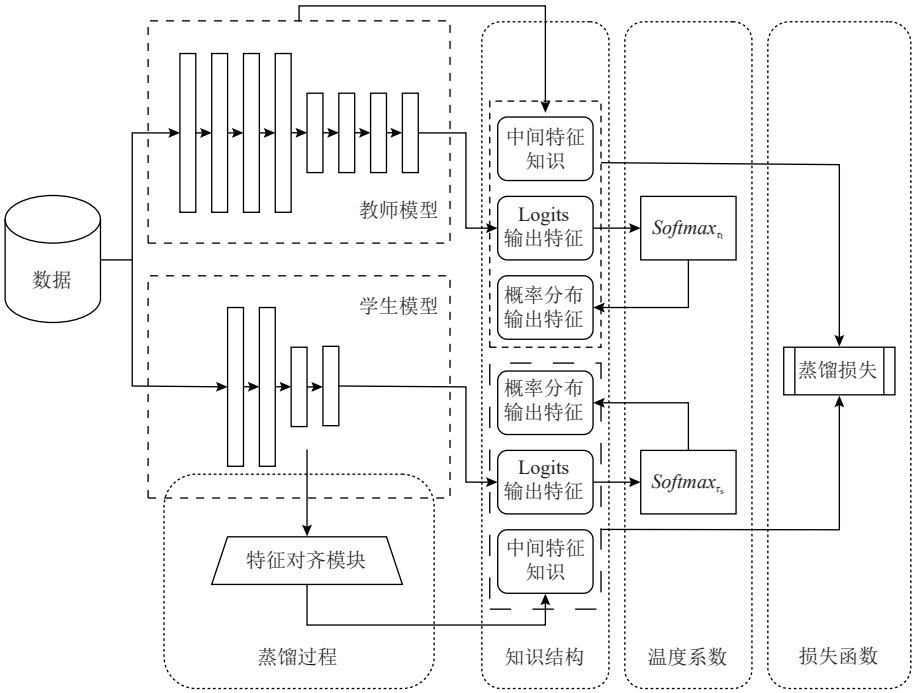


Fig. 13 Illustration of optimized location of knowledge distillation

图 13 知识蒸馏的优化位置示意图

构建一个正交投影层，最大限度地将蒸馏知识传递给学生模型。他们还针对蒸馏中的投影层、归一化层和距离函数进行了更深层次的实验探究[164]。Sun等人[165]提出一种基于掩码的方法屏蔽教师模型的错误分类，来减少教师模型的错误输出对学生模型的负面影响。

4.3 温度系数优化

知识蒸馏的温度系数从被提出时始终作为固定超参数，然而针对不同的样本使用相同的温度系数来软化目标，得到的结果往往是次优的，因此针对不同的特征自适应地设置不同的温度系数，能够获得更好的软标签来指导学生模型训练，进一步提升学生模型的性能。

Xu等人[166]从one-hot标签角度研究了标签分布与真实分布之间的差距，发现网络中倒数第2层特征的L_2范数可以作为标签中的噪声强度来代替温度系数。Jafari等人[167]通过一个退火函数来生成动态温度，并用均方误差取代KL散度作为距离度量。Li等人[168]认为训练学生模型应当遵循由易到难的原则，提出使用温度系数来调整任务的难度，通过GAN的思想来学习调整温度系数。Chi等人[169]将样本的Logits输出视为高斯分布，利用其标准差作为温度系数。Wei等人[170]提出了样本Logits输出平滑度的概念，利用log-sum-exp函数评估Logits输出的尖锐度，动态调整温度系数，对尖锐度高的教师模型设置较高温度，对尖锐度低的学生模型设置较低温度，以平衡师生模型输出一致性。Sun等人[171]从信息论的最大熵理论利用拉格朗日乘子法推导温度系数，根据推导后的结果提出对Logits输出特征进行标准化，即利用均值对Logits输出进行平移后将标准差作为温度系数。

4.4 损失结构优化

知识蒸馏的损失函数通常包括任务损失和蒸馏损失，有研究发现这2项损失之间以及蒸馏损失内部存在耦合和冲突，会影响蒸馏学习的性能。因此近期一些工作[172-175]专注于打破这种关系来进一步提升蒸馏的性能。

Zhao等人[172]提出将损失解耦为针对目标类的二元蒸馏和针对非目标类的多类别蒸馏；Yang等人[173]发现Zhao等人[172]提出的DKD非目标类别师生模型的概率分布输出和不同，这阻碍了分布的相似，提出对分布进行归一化；Zhao等人[175]认为任务损失与蒸馏损失发生冲突，退化为多任务学习，提出调整优化器动量值让蒸馏损失占据主导地位来打破冲突的博弈；Park等人[174]将DKD[172]中对非目标类暗知识的动态自适应调整学习同样应用到目标类上，并且对任务进行序列化训练，来降低2种不同分布优化任务的冲突。

4.5 交叉技术优化

交叉技术优化旨在利用其他领域的技术，例如

对抗性技术、自动机器学习、强化学习、扩散模型等，通过这些技术的优势来改善优化知识蒸馏学习的过程，来获得性能更好的学生模型。

1)对抗性技术。对抗学习在2014年Goodfellow等人[123]提出的GAN取得巨大成功后受到广泛关注，GAN中判别器估计样本来自训练数据分布的概率，生成器试图利用生成的数据来欺骗判别器，两者相互博弈来提高生成数据的质量。而该思想同样可以应用到知识蒸馏中来提升学生模型的性能。Shen等人[176]为了从不同教师模型中提取多样化知识，定义了一个分块训练损失来指导学生模型恢复教师模型中的知识。Li等人[177]在对抗训练框架下引入任务驱动的注意力机制，将知识设计成紧凑的形式传递给学生模型。Ma等人[178]提出了一项利用对抗性思想防止闭源模型通过蒸馏的方式被窃取，通过最大化教师模型与预训练模型的KL散度来尽可能地获得释放错误信号的能力。

在蒸馏中决策边界一直被认为是重要的，Heo等人[179]利用对抗攻击发现支持决策边界的样本来向学生模型传递边界信息。Nguyen-Duc等人[180]构造了一个给定输入球约束内所有对抗样本的集合，来挖掘决策边界。

鲁棒性蒸馏是指通过对抗性训练或其他技术得到一个鲁棒性的教师模型，然后通过蒸馏技术将鲁棒性的知识传递给学生模型的方法。Bai等人[181]发现学生模型很难从教师模型中继承对抗鲁棒性，通过一种重加权的策略有选择地从教师模型学习。Huang等人[182]设计了自适应对抗性蒸馏方案，在内部优化中自适应搜索最优匹配点。Yue等人[183]考虑了鲁棒性的类别平衡性，提出公平对抗鲁棒性蒸馏，增加困难类的权重来增强学生模型的鲁棒公平性。Wang等人[184]认为提供鲁棒性知识的教师模型并不总是能正确预测，通过多个实例之间的结构化关系代替1对1的模仿学习。

2)自动机器学习。自动机器学习(automated machine learning, AutoML)是指通过自动化的方式来完成机器学习模型的设计、训练和优化过程。AutoML的目标是使得机器学习的应用变得更加简单和高效，即使对于没有专业知识的用户，也能使用机器学习技术解决实际问题，其中NAS是其中最流行的一项技术之一，在知识蒸馏中学生模型的性能不仅仅取决于教师模型提供的知识，还取决于学生模型架构的设计，因此有的工作将NAS用于改进知识蒸馏的师生模型架构。

Li等人[185]提出一种分块搜索方案来评估每一块的候选框架，其搜索结果超过了教师模型的性能。Liu等人[186]提出一种基于强化学习的NAS搜索算法来寻找给定教师模型的最佳的学生模型。Kang等人[187]通过提出一种预言机损失来帮助NAS搜索，缓解师生模型能力差距带来的性能下降问题。Wei等人[188]提出了一种在小规模数据集上对ViT架构进行搜索的方案，找到最佳的学生模型架构进行蒸馏学习。Dong等人[189]提出一种基于NAS的学生模型搜索策略，通过衡量类别样本之间的相似性来获得最佳学生模型。Lu等人[190]提出一种基于NSGA-II遗传算法的搜索框架，并提出连续减半的优化加速策略。有的工作通过总结先前工作构造搜索空间，利用搜索算法来获得最佳的蒸馏设置。Li等人[191]通过蒙特卡洛搜索最佳的蒸馏策略，提出离线推理、随机掩码的稀疏蒸馏，提前停止策略来加速蒸馏。

3)强化学习。强化学习(reinforcement learning, RL)是机器学习的一个重要领域，强化学习可以被简单理解为智能体在环境中采取行动并得到奖励来扩充自己的知识储备，使得自己在环境中有更好的表现，得到更多的回报。Liu等人[186]设计了基于知识蒸馏的奖励函数，通过强化学习的方法搜索最佳的师生模型架构。

4)扩散模型。扩散模型(diffusion models)是生成模型的一类，它通过逐步去噪的过程生成新数据。受非平衡态热力学的启发，在2015年扩散模型的概念[192]被提出，并在2020年由Ho等人[193]提出的DDPM推动其进一步发展，扩散模型的基本思想是将数据逐步转化为噪声，然后通过逆过程逐步去噪生成新数据。这一过程可以看作是一种马尔可夫链，它由前向过程(添加噪声)和逆向过程(去除噪声)组成。扩散模型已经在图像生成等领域取得了显著的成功。Huang等人[194]认为蒸馏特征中的噪声是影响蒸馏性能的原因之一，受扩散模型逆过程的启发，提出通过扩散模型逐步地从图像或特征中去除噪声。

4.6 小 结

针对知识蒸馏的优化方法汇总如表7所示。这类研究是近年来的热点，通过从训练优化的角度改进知识蒸馏的学习范式，取得了显著的性能提升。然而，目前的方法对于师生模型架构的设计仍然依赖于人工设计，尽管通过搜索方法可以找到最佳的蒸馏设置，但这会引入额外的计算负担。因此，如何在低开销的情况下获得针对特定任务场景的最佳师生模型架构设计仍值得深入研究。此外，当前知识

Table 7 Summary of Optimization Methods for Knowledge Distillation

表 7 知识蒸馏的优化方法汇总

方法来源	优化位置	优化方法	关键技术
文献 [150]	训练过程	提前停止	提出提前停止优化方法，缓解蒸馏损失后期对模型的损害
文献 [176]	损失函数	对抗性技术	提出利用分块训练损失来从教师模型中提取多样化知识，缓解知识单一的问题
文献 [179]	训练过程	对抗性技术	利用对抗攻击发现决策边界样本，传递准确的边界信息来提升性能
文献 [151]	训练过程	正则化约束	性能较弱的模型能够提供正则化约束给性能较强的模型来提升性能
文献 [38]	训练过程	辅助模型	引入一个助理模型来弥合师生模型的性能差距，缓解师生模型能力差距带来的性能下降
文献 [158]	知识结构	特征筛选	提出选择性转移策略，利用样本变换带来的暗知识以提高模型性能
文献 [166]	温度系数	动态温度	利用倒数第 2 层特征的L_2范数作为温度系数来缓解标签噪声的影响
文献 [177]	训练过程	对抗性技术	设计任务驱动的注意力机制，构造紧凑的知识结构以减少信息丢失
文献 [186]	模型架构	强化学习	设计奖励函数寻找教师模型的最佳学生模型，缓解师生模型能力差距带来的性能问题
文献 [187]	模型架构	自动机器学习	提出一种基于预言机的搜索算法来缓解师生模型能力差距带来的性能问题
文献 [185]	模型架构	自动机器学习	提出一种分块的搜索评估方案以保证候选框架得到公平的训练
文献 [167]	温度系数	动态温度	通过退火函数生成动态温度来缓解师生模型能力差距过大带来的性能下降
文献 [152]	训练过程		综合探究影响蒸馏性能的原因，优化不佳是主要原因
文献 [178]	损失函数	对抗性技术	提出自我破坏知识蒸馏来缓解蒸馏知识产权暴露的风险
文献 [172]	损失函数	解耦	提出将损失函数解耦来缓解经典蒸馏损失本身内部的冲突博弈
文献 [153]	训练过程	非参数化对齐	提出通过奇异值分解对齐师生模型来缓解投射层性能效率不佳的问题
文献 [154]	训练过程	推理蒸馏分离	通过保存教师模型推理的数据增强及输出来缓解蒸馏过程教师模型推理的开销
文献 [159]	知识结构	特征筛选	提出知识凝聚蒸馏，对知识进行分级来处理蒸馏知识冗余的问题
文献 [160]	知识结构	皮尔森系数	利用师生模型预测输出类别的相对关系来处理师生模型预测的差异问题
文献 [112]	知识结构	批次结果重用	提出小批量自蒸馏算法，利用上一批次的信息缓解难以并行化的问题
文献 [155]	训练过程	特征路由	提出点自适应位置蒸馏来处理师生模型蒸馏对齐位置不确定的问题
文献 [188]	模型架构	自动机器学习	提出小规模数据集 ViT 搜索方案来缓解师生模型能力差距带来的性能问题
文献 [191]	训练设置	自动机器学习	构造蒙特卡洛搜索空间探寻最佳设置来缓解手工设计蒸馏方案的压力
文献 [189]	模型架构	自动机器学习	通过衡量样本间相似性寻找最佳师生模型架构来解决师生模型能力差距带来的性能下降
文献 [180]	训练过程	对抗性技术	通过构造给定输入球约束内对抗样本集合挖掘决策边界来提升模型性能
文献 [181]	训练过程	对抗性技术	提出一种重加权策略的引导对抗对比蒸馏来帮助学生模型继承对抗鲁棒性
文献 [182]	训练过程	对抗性技术	设计了自适应对抗性蒸馏方案来缓解先前模型过度修正的问题
文献 [184]	训练过程	对抗性技术	利用多个实例之间的结构化关系来缓解教师模型不稳定预测的影响
文献 [173]	损失函数	解耦	通过对解耦后的输出归一化来处理师生模型输出不一致的问题
文献 [175]	损失函数	解耦	调整优化器动量值让蒸馏损失占据主导，打破不同损失的冲突博弈
文献 [162]	知识结构	数据筛选	通过困难样本挖掘和校准来缓解教师模型软标签置信度可靠性问题
文献 [156]	训练过程	正则化约束	提出混合权重和间歇融合寻找更平坦的最优解来提高模型泛化性
文献 [161]	知识结构	困难样本挖掘	通过筛选易错样本加强学习来缓解教师模型传播错误知识的风险
文献 [168]	温度系数	对抗温度	利用温度系数调整任务难度来更合理地为学生模型设置学习进程
文献 [169]	温度系数	标准化 Logits 温度	将 Logits 输出标准差作为温度系数来缓解样本知识传递不充分的问题
文献 [163]	知识结构	正交投影层	提出在 Stiefel 流形上构建正交投影层来提高知识传递效率
文献 [170]	温度系数	师生模型差异温度	提出通过平衡输出特征的平滑度来解决不同难度样本的知识转移问题
文献 [171]	温度系数	标准化 Logits 温度	提出对 Logits 输出进行标准化来解除师生模型输出在范围和方差上的耦合
文献 [183]	训练过程	对抗性技术	提出公平的鲁棒性蒸馏框架来解决鲁棒性的类别差异性
文献 [194]	训练过程	扩散模型	提出扩散模型去噪模块来解决蒸馏特征中的噪声问题
文献 [190]	模型架构	自动机器学习	通过遗传算法寻找最佳师生模型架构来缓解师生模型能力差距带来的性能问题
文献 [174]	损失函数	解耦、序列化优化	将真实标签和教师模型标签的优化任务序列化进行，缓解不同分布优化任务的冲突
文献 [157]	训练过程	优化路径调整	通过将学生模型与教师模型交叉蒸馏来缓解真实标签与教师模型错误分类之间的冲突问题
文献 [165]	知识结构	错误标签掩码	利用掩码屏蔽教师模型的错误分类来减少教师模型的错误输出对学生模型的负面影响

筛选的工作主要集中在教师模型上，而对学生模型难以区分的困难样本进行挖掘将是一项有意义的研究方向。并且目前的研究大多基于实验观察结果，缺乏理论上的完整性。

5 知识蒸馏的应用

知识蒸馏有着广泛的应用，通常是利用其模型压缩的能力和知识转移的特性，本节总结了 5 个较为热门的下游任务，例如目标检测、语义分割、人脸识别、异常检测和姿态估计等；不仅仅局限于下游任务，知识蒸馏中知识转移和隐私保护的特性还能够直接应用于某一个领域，例如增量学习和联邦学习等。

5.1 目标检测

目标检测旨在从图像或视频中识别并定位特定类别的目标物体。与图像分类不同，目标检测不仅需要识别图像中的目标类别，还需要确定每个目标在图像中的具体位置。通常通过在图像中绘制边界框来实现，每个边界框标记一个目标物体的空间位置。根据目标检测的维度不同，可以将目标检测分为 2D 目标检测和 3D 目标检测。

大量的假阴性样本和较低定位精度样本影响性能，Guo 等人[195]采用一致性正则化来处理假阴性样本和低精度定位带来的噪声问题。Zhou 等人[196]引导模型从简单样本到困难样本逐步蒸馏显著性知识，并通过预测边界周围纹理来细化噪声标签的边界。

视觉语言模型（vision language model, VLM）中包含着丰富的知识，Wang 等人[197]提出一个物体感知的蒸馏金字塔框架，从预训练的 VLM 提取知识，并引入全局和分块蒸馏进行知识传递。Li 等人[198]将 VLM 获得的知识提取到 DERT 的检测器中，通过语义蒸馏和关系蒸馏来传递语义知识和利用物体之间的隐式关系。Zhou 等人[199]提出了一个通用的跨模态知识蒸馏框架来提高单模态检测器的性能，将不同模态的信息蒸馏到同一个模态来合成。

Wu 等人[200]认为提取物体特征的域不变表示具有重要意义，提出循环解耦自蒸馏用于从视觉特征中循环提取域的不变特征表示。针对以上方法无法从异构教师模型中提取知识的问题，Lao 等人[201]引入额外可变形的交叉注意力解码头，让不同架构的师生模型互相学习。Zeng 等人[202]研究了如何从一个不完美的教师模型中提取知识。

在 3D 目标检测中 Klingner 等人[140]介绍了一种跨任务、跨模态、跨阶段的多相机 3D 目标检测综合知识蒸馏框架。Liu 等人[203]提出一种跨模态蒸馏方法，通过 Logits 输出特征蒸馏来自上级激光雷达点云 LiDAR（light detection and ranging）模型的立体探测器，来缩小基于立体和基于 LiDAR 方法之间的差距。来自估计的深度先验不可避免的误差可能会导致不一致的语义信息和 3D 定位，Wu 等人[204]提出了具有 3D 感知位置编码的注意力深度知识蒸馏框架来约束学生模型学习。

基于多视角鸟瞰图 BEV（bird's eye view）的 3D 目标检测与基于 LiDAR 的方法有明显的性能差距，Wang 等人[205]提出基于多视角 BEV 的学生模型，利用蒸馏训练好的 LiDAR 教师模型的特征来弥补性能差距。Zhang 等人[206]提出一种新视图引导蒸馏目标，可以稳定量化感知训练，同时利用图像特征和 BEV 特征来增强模型性能。Zhao 等人[207]提出了一种用于 BEV 目标检测的多模态蒸馏方案，将教师模型的 LiDAR 特征知识转移到多模态学生模型中。

现有方法只有在师生模型共享相似的中间特征表示时才会显示出有效性，点云的内在稀疏性和不规则性导致了表示差异。Li 等人[208]提出了一种表征差异蒸馏，减少较小的学生模型与过度参数化的教师模型之间的性能差异。Gambashidze 等人[209]提出通过 X 射线蒸馏解决 LiDAR 目标检测中的稀疏性和遮挡性。Huang 等人[210]将动力学与雨环境理论统一起来，提出晴雨知识蒸馏，增强雨天条件下的 LiDAR 目标检测。

5.2 语义分割

语义分割旨在将图像中的每一个像素分类到某一个特定的类别中。与图像分类和目标检测不同，语义分割不仅需要识别图像中的物体，还需要准确地定位每一个物体的边界。它在自动驾驶、医学影像分析、遥感影像处理等领域有广泛的应用。近期利用知识蒸馏的语义分割方法主要通过蒸馏算法来对特征和结构信息进行补充，从而构建更好的特征表示来增强模型的泛化性和准确性。

Ji 等人[211]充分利用低层的结构纹理知识，通过迭代拉普拉斯金字塔和方向滤波来分解低级特征，挖掘纹理知识。Gao 等人[212]提出一种跨域相关蒸馏的域适应框架处理光照影响，提取内容和风格知识特征，计算 2 幅图像之间的固有差异和光照差异。Yang 等人[213]侧重在整个图像之间传递结构化的像素和像素到区域的关系。Peng 等人[214]利用自注意力模块建立层次化稠密特征，引入相关性蒸馏来促进细粒度的分割。Liu 等人[215]提出通过多个教师模型解码

器生成可靠软标签，提出解耦特征集成蒸馏机制，利用重要的特征图进行编码器学习。

在3D点云语义分割方面，由于点云稀疏性和随机性导致直接蒸馏效果欠佳，Hou等人[216]提出从点级和体素级转移隐藏的知识，利用逐点和体素蒸馏来补充稀疏的监督信号。Chen等人[217]尝试对点云进行完全无监督的语义分割，将多视图的视觉特征反向投影到3D空间聚合为统一的点特征，将点特征聚合为超体素进行语义的挖掘。

5.3 人脸识别

人脸识别任务是一种生物识别技术，旨在通过计算机视觉和机器学习算法，自动检测和识别图像或视频中的人脸。近期利用知识蒸馏的人脸识别工作主要利用蒸馏技术知识迁移的能力来获得更好的人脸特征表示。

Xu等人[218]提出对人脸的贝叶斯集成平均模型进行蒸馏压缩来提高模型性能。Li等人[219]试图去除学生模型训练时的监督信号，通过反向蒸馏约束教师模型的搜索空间来缩小师生模型的内在差距。学生模型的类内相似度与教师模型有较大差距，Yu等人[220]提出估计师生模型的概率分布，并引入相似性分布一致性损失来提高学生模型的类内紧凑型。轻量级的学生模型由于模型容量较低难以拟合教师模型的Logits输出特征，Zhao等人[221]通过探测Logits输出特征来提取与面孔身份相关的主要知识，并抛弃其他不相关知识。Zhang等人[222]为减轻训练数据集中人脸姿态分布不均匀的影响，将教师模型人脸的角度知识提取到学生模型中，使之能够处理较大姿态变化的人脸。为解决面部标注数据有限的问题，Hu等人[223]提出一个自监督人脸表示学习框架，能够从人脸视频中爬取通用的、多样的人脸表示。

5.4 异常检测

异常检测是一种识别数据中异常模式或异常行为的技术。异常通常指那些显著偏离正常模式的数据点，这些异常数据可能表示某种问题或罕见事件。

反向蒸馏是Deng等人[224]在2022年为解决异常检测问题提出的一项技术，让学生模型将教师模型对于单个类别的编码作为输入来恢复教师模型的多尺度表示。Tien等人[225]通过反向蒸馏的思想提出一个新的异常检测框架，在低延迟的情况下表现良好。

Gu等人[226]通过存储常态信息来强化学生模型生成特征的常态性，作为其进行异常检测的先验知识。Binh等人[227]提出基于注意力机制的深度伪造检测蒸馏模型，应用知识蒸馏领域中的频域学习和最优传输理论，改善低质量压缩深度伪造图像的检测。Ma等人[228]提出一种多教师的蒸馏框架进行异常检测，来融合多种不同神经网络的优势。Ristea等人[229]提出了一种自蒸馏掩码自动编码器的视频异常检测框架，利用师生模型解码器之间的差异来提升检测性能，并通过合成异常事件来增强训练。刑鹏等人[230]提出一种基于特征约束的蒸馏学习方法，充分利用教师模型的特征来指导学生模型高效地识别异常图像。

5.5 姿态估计

姿态估计旨在从图像或视频中检测并估计人体、动物或物体的姿态。姿态估计的主要目标是识别图像或视频中的特定关键点(如关节、骨骼点等)的位置和姿态信息。

通过分类的预训练模型来生成视频特征用于动作定位，不同优化任务会使时间定位结果出现不一致的问题，Ju等人[231]提出从预训练的VLM提取自由动作知识，这类知识与基于分类的模型知识互补，通过蒸馏将两者融合。Wu等人[232]提出一种基于VLM的端到端的零样本人-物交互检测框架，结合2阶段二分匹配算法交互评分模块，以动作无关的方式实现对人-物对的交互区分。

依赖于跟踪过去位置的方法在拥挤环境可能会累计错误，导致跟踪漂移。Monti等人[233]引入时空注意力模块恢复不同时间步之间的依赖关系，迁移至学生模型使其能用较短的序列进行预测。Ma等人[234]提出一个多目标条件变分推荐预测模型来平衡准确性采样和多样性采样。Yang等人[235]提出利用多人线性边缘特征提取来处理复杂姿态和物体遮挡问题，生成一个包含人-物边界和遮挡信息的重叠边缘，并在该边缘上进行特征蒸馏。

Ye等人[236]将基于热力图的模型作为教师模型、基于回归的模型作为学生模型，通过蒸馏集成两者优势，生成兼具高效性和优越性的学生模型。为了使Transformer架构的模型能有效部署到边缘设备上，Chen等人[237]提出一种改进的多重循环Transformer模块架构的自蒸馏算法，来改进小型模型欠拟合的问题。

5.6 增量学习

增量学习又称为持续学习或终身学习，是指模型能够随着时间的推移逐步学习新知识，其中有一项主要的挑战是灾难性遗忘问题，模型在学习新知识的过程中，可能出现对已学知识的遗忘现象。将知识蒸馏技术应用于增量学习是一个有效的方法，利用蒸馏技术知识迁移的能力，可以缓解增量学习

的灾难性遗忘问题。

Lu 等人[238]提出增强几何蒸馏，在记忆不同的视图上以成对和交叉的方式进行蒸馏，并构建一个通用无数据增量框架来避免隐私泄露。Shang 等人[239]在解码器中添加新的类标记来进行新类的学习，并通过对旧类知识的蒸馏来减少灾难性遗忘。Phan 等人[240]提出一种基于类相似性的知识蒸馏方法，将先前模型知识提取到与新类相似的旧类上进行联合学习。Gu 等人[241]对旧类和新类分别设计了损失函数，并对旧类进行重规范化来淡化对旧知识的丢失。Li 等人[242]提出通过重要性采样的对比连续学习来保存知识，提出一种重放缓冲区的策略来恢复先前的数据分布。

增量学习的另一个问题是对新样本的欠拟合和对少数旧样本的过拟合，Zhao 等人[243]提出双教师模型结构，一个模型在基类数据上训练有丰富通用知识的模型来缓解新类的过拟合，另一个模型用来获取新类的知识。Yan 等人[244]训练了堆叠的子教师模型，集成多级监督和反向自蒸馏，从各个教师模型中收集其优势。对无限数据流进行训练时只能考虑短期信息，Ye 等人[245]提出使用变分自动编码器开发一个短时记忆模块和一个参数化可扩展性记忆模块来构成无监督持续学习框架。

增量学习中还存在着训练的不稳定的问题，Wen 等人[246]提出多教师蒸馏，通过权重排列、特征扰动和多样性正则化技术确保教师模型知识的多样性和有效性。Yang 等人[247]发现背景类和新类之间存在混淆影响增量学习的性能，将旧模型的输出在真实标签的引导下进行扩展移植，使其与新的模型输出形成相匹配的类对应关系。

5.7 联邦学习

联邦学习是一种分布式机器学习技术，通过在多个设备（如智能手机、物联网设备等）上本地训练模型，并在不共享原始数据的前提下将模型参数或梯度汇总到中央服务器，从而构建全局模型。这种方法能够在保护数据隐私的同时，实现高效的机器学习模型训练。

数据异构性是联邦学习的一大挑战，Zhang 等人[248]提出一种无数据知识蒸馏来微调服务器中的全局模型，引入一个生成器来探索局部模型的输入空间并将其知识迁移到全局模型中。Shen 等人[249]提出一种基于信道分配循环蒸馏引导的通道解耦框架。

Kim 等人[250]使用联邦学习来提高单分类器的性能，提出基于原型表示蒸馏的无监督联邦学习来增强全局模型的表达能力。为处理具有极端标签噪声的客户端，Lu 等人[251]提出负蒸馏，令标签噪声上训练的模型充当“坏老师”来降低提供错误信息的风险。Wang 等人[252]将每个客户端中的本地数据视为一个特定的领域，提出一种领域知识感知的联邦蒸馏方法。乐俊青等人[253]设计了一种基于双重知识蒸馏的训练算法，确保模型学习到每个样本的知识，使模型保持较高的准确度，并通过自适应的差分隐私策略来实现客户端级的强隐私保护。

5.8 小 结

知识蒸馏主要用于模型压缩，但由于其在知识转移中的独特优势，被广泛应用于其他领域，特别是在跨模态知识转移任务中。本节总结了知识蒸馏在 5 种下游任务（目标检测、语义分割、人脸识别、异常检测和姿态估计）和 2 种特定领域（增量学习和联邦学习）中的应用。在实际应用中，应根据具体场景和需求选择合适的蒸馏方法，以最大限度地提升模型性能和效率。通过知识蒸馏，不仅能有效压缩模型，还能实现特征融合和知识共享，拓展其在多种复杂任务中的应用潜力。

6 总结与展望

知识蒸馏作为深度学习模型压缩和加速的重要技术，近年来在学术界和工业界都得到了广泛关注和深入研究。

最初，知识蒸馏主要集中在对教师模型知识来源的探索。Hinton 等人[21]提出的蒸馏算法通过利用教师模型的 Logits 输出特征进行知识传递，奠定了蒸馏的基础。随着研究的深入，发现中间层特征和关系特征同样具有重要的价值，进一步提升了蒸馏效果。在大模型出现后，Transformer 架构的系列模型也为知识蒸馏提供了新的视角，探索了如何在不同类型的模型之间进行有效的知识迁移。

随后，知识蒸馏的研究逐步转向了不同学习方式的探索。早期的蒸馏方法主要基于离线蒸馏，通过预先训练好的教师模型传递知识给学生模型。随着需求的多样化，在线蒸馏、自蒸馏以及无数据蒸馏等方法应运而生，允许在没有教师模型或原始数据的情况下进行蒸馏。此外，多教师蒸馏和助理蒸馏等方法也获得广泛关注，通过引入多个教师模型或助理模型的知识，进一步提高了蒸馏的效果和鲁棒性。

近几年的研究更加关注优化方式的探讨，旨在进一步提升蒸馏的效率和效果。研究者们通过知识

结构的筛选来确保传递给学生模型的知识更具代表性和有效性,避免冗余和错误信息的传递;优化蒸馏过程的交互方式和改进损失函数以更好地对接教师与学生模型的特征输出,也是近期研究的重点。温度系数的优化作为一个重要方向,通过调节软标签的平滑程度,进一步平衡教师与学生模型之间的知识传递效果。

尽管知识蒸馏在多个领域取得了显著的进展,但在实际应用和理论研究中仍面临诸多挑战:

1)蒸馏知识的结构。教师模型中的知识可以呈现多种形式,包括输出特征知识、中间层特征知识、样本或通道关系知识以及 ViT 架构模型中的知识等。如何选择合适的知识类型依然是一个值得研究的问题。针对这些多样化的知识类型,选择合适的提取位置也是一个重要的研究方向。此外,从教师模型中抽取的知识可能会包含噪声或不确定性,这些因素可能对学生模型产生不利影响。因此,研究如何处理和过滤教师模型知识中的不利因素是一个具有挑战性的课题。另一方面,如何从教师模型中挖掘出令学生模型难以区分的困难特征知识以加强学习,也是一个值得深入探索的方向。

2)师生模型架构的探索。教师模型的选择对知识蒸馏的效果至关重要。在实际应用中,教师模型的选择通常依赖于任务的性质、数据的特点和模型的计算资源。如何选择合适的教师模型,以确保其能够有效地指导学生模型进行训练,仍然是一个开放性的问题。

已有研究[38]证明,差距过大的师生模型架构会对学生模型产生不利影响。然而,目前对师生模型架构设计的探索仍主要依赖人工设计。尽管一些工作[189-191]通过搜索方法找到了最佳的师生模型架构设置,但这会引入大量额外的计算开销。在较少资源消耗的前提下找到最佳的师生模型架构仍是一个重要的挑战。

目前,大多数知识蒸馏方法主要针对特定类型的模型或特定任务展开研究。然而,对于跨架构的蒸馏,尤其是在 ViT 架构与 CNN 架构之间进行蒸馏时存在显著的挑战。这是因为不同架构所提取的特征通常存在语义不匹配的问题,这使得直接将教师模型的知识传递给学生模型变得困难。因此,设计合适的语义对齐模块,以便更有效地将教师模型的知识传递给学生模型,成为一个亟待解决的研究课题。此外,如何设计一种通用的蒸馏方法,能够跨越不同的网络架构和任务类型,适应各种应用场景,依然是知识蒸馏领域面临的重大挑战。

3)蒸馏学习的优化。尽管知识蒸馏能够有效压缩模型并提高学生模型的泛化能力,但在实际应用中,学生模型往往难以完美复制教师模型的能力。如何控制蒸馏过程中的性能损失,确保学生模型的效果尽可能接近教师模型,仍然是一个长期存在的挑战。

现有的蒸馏学习范式主要基于 Hinton 等人[21]在 2015 年提出的蒸馏学习算法。但有研究[152]指出,这一范式可能限制了学生模型性能的提升。尽管当前的研究通过优化损失函数、温度系数等方面进行了改进,仍未形成被广泛接受的新范式,且大多数改进基于实验观察,缺乏理论上的系统性。因此,构建一种新的优化学习体系,以突破现有瓶颈,仍是亟待解决的核心问题。

此外,知识蒸馏过程通常需要较长的训练时间和大量的计算资源,尤其是当教师模型规模庞大时,这一问题更为突出。与此同时,随着蒸馏方法应用的扩展,模型的可解释性和透明度逐渐成为关注重点。如何设计蒸馏算法,使得学生模型不仅能接近教师模型的性能,还能够提供清晰的决策过程解释,是当前研究中的一个重要难题。

本节从知识的来源、蒸馏的方式和蒸馏的优化总结了目前知识蒸馏的发展,并从知识蒸馏的关键组成:知识、师生模型架构和蒸馏算法 3 个方面讨论了目前面临的挑战,未来的研究需要在这些方向上继续探索,以进一步提升蒸馏效果和学生模型的性能。自 2015 年知识蒸馏提出以来,该技术在学术界和工业界受到了广泛关注。希望本综述能为相关研究人员和工程师提供一些启发和参考,继续推动知识蒸馏的进一步发展。

作者贡献声明: 潘海为、于丰铭完成相关文献资料的收集和分析,是论文的主要撰写人;张可佳、兰海燕负责论文整体结构的设计,并对论文进行修改;孟庆宇、李哲参与文献资料的分析、整理。

参 考 文 献

[1] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278−2324

[2] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Proc of the 26th Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT, 2012: 1106−1114

[3] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint, arXiv: 1409.1556, 2014

[4] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions[C]//Proc of 2015 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1−9

[5] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proc of 2016 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2818−2826

[6] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[C]//Proc of the 31st AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 4278−4284

[7] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of 2016 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[8] Huang Gao, Liu Zhuang, van der M L, et al. Densely connected convolutional networks[C]//Proc of 2017 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 4700−4708

[9] Howard A G, Zhu Menglong, Chen Bo, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint, arXiv: 1704.04861, 2017

[10] Sandler M, Howard A, Zhu Menglong, et al. MobileNetV2: Inverted residuals and linear bottlenecks[C]//Proc of 2018 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4510−4520

[11] Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3[C]//Proc of 2019 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 1314−1324

[12] Tan Mingxing, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks[C]//Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 6105−6114

[13] Tan Mingxing, Le Q. EfficientNetV2: Smaller models and faster training[C]//Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2021: 10096−10106

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of the 31st Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2017: 5998−6008

[15] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv preprint, arXiv: 2010.11929, 2021

[16] Liu Ze, Lin Yutong, Cao Yue, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows[C]//Proc of 2021 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 10012−10022

[17] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 10347−10357

[18] Wang Wenhai, Xie Enze, Li Xiang, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//Proc of 2021 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 568−578

[19] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 8748−8763

[20] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 8821−8831

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

[22] Gou Jianping, Yu Baosheng, Maybank S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789−1819

[23] Wang Lin, Yoon K J. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3048−3068

[24] Huang Zhenhua, Yang Shunzhi, Lin Wei, et al. Knowledge distillation: A survey[J]. Chinese Journal of Computers, 2022, 45(3): 624−653 (in Chinese)
(黄震华, 杨顺志, 林威, 等. 知识蒸馏研究综述[J]. 计算机学报, 2022, 45(3): 624−653)

[25] Shao Renrong, Liu Yuang, Zhang Wei, et al. A survey of knowledge distillation in deep learning[J]. Chinese Journal of Computers, 2022, 45(8): 1638−1673 (in Chinese)
(邵仁荣, 刘宇昂, 张伟, 等. 深度学习中知识蒸馏研究综述[J]. 计算机学报, 2022, 45(8): 1638−1673)

[26] Si Zhaofeng, Qi Honggang. Survey on knowledge distillation and its application[J]. Journal of Image and Graphics, 2023, 28(9): 2817−2832 (in Chinese)
(司兆峰, 齐洪钢. 知识蒸馏方法研究与应用综述[J]. 中国图象图形学报, 2023, 28(9): 2817−2832)

[27] Oord A, Li Yazhe, Vinyals O. Representation learning with contrastive predictive coding[J]. arXiv preprint, arXiv: 1807.03748, 2019

[28] Bucila C, Caruana R, Niculescu-Mizil A. Model compression[C]//Proc of the 12th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2006: 535−541

[29] Sau B B, Balasubramanian V N. Deep model compression: Distilling knowledge from noisy teachers[J]. arXiv preprint, arXiv: 1610.09650, 2016

[30] Zhang Ying, Xiang Tao, Hospedales T M, et al. Deep mutual learning[C]//Proc of 2018 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4320−4328

[31] Furlanello T, Lipton Z, Tschannen M, et al. Born again neural networks[C]//Proc of the 35th Int Conf on Machine Learning. New York: ACM, 2018: 1607−1616

[32] Yang Chenglin, Xie Lingxi, Qiao Siyuan, et al. Training deep neural networks in generations: A more tolerant teacher educates better students[C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 5628−5635

[33] Lee H, Hwang S J, Shin J. Self-supervised label augmentation via input transformations[C]//Proc of the 37th Int Conf on Machine Learning. New York: ACM, 2020: 5714−5724

[34] Yang Jing, Martinez B, Bulat A, et al. Knowledge distillation via softmax regression representation learning[C/OL]//Proc of 9th Int Conf on Learning Representations, New York: Ithaca, 2020[2024-08-20]. https://openreview.net/forum?id=ZzwDy_wiWv

[35] Ba J, Caruana R. Do deep nets really need to be deep?[C]//Proc of the 28th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2014: 2654−2662

[36] Zhou Guorui, Fan Ying, Cui Runpeng, et al. Rocket launching: A universal and efficient framework for training well-performing light net[C]//Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 4580−4587

[37] Son W, Na J, Choi J, et al. Densely guided knowledge distillation using multiple teacher assistants[C]//Proc of 2021 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 9375−9384

[38] Mirzadeh S I, Farajtabar M, Li Ang, et al. Improved knowledge distillation via teacher assistant[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 5191−5198

[39] Li Xuewei, Li Songyuan, Omar B, et al. ResKD: Residual-guided knowledge distillation[J]. IEEE Transactions on Image Processing, 2021, 30: 4735−4746

[40] Romero A, Ballas N, Kahou S E, et al. FitNets: Hints for thin deep nets[J]. arXiv preprint, arXiv: 1412.6550, 2015

[41] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[J]. arXiv preprint, arXiv: 1612.03928, 2017

[42] Lin Sihao, Xie Hongwei, Wang Bing, et al. Knowledge distillation via the target-aware transformer[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10915−10924

[43] Lao Shanshan, Song Guanglu, Liu Boxiao, et al. Masked autoencoders are stronger knowledge distillers[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 6361−6370

[44] Guo Ziyao, Yan Haonan, Li Hui, et al. Class attention transfer based knowledge distillation[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 11868−11877

[45] Zhang Yuan, Huang Tao, Liu Jiaming, et al. FreeKD: Knowledge distillation via semantic frequency prompt[C]//Proc of 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 15931−15940

[46] Jia Menglin, Tang Luming, Chen B, et al. Visual prompt tuning[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 709−727

[47] Huang Zehao, Wang Naiyan. Like what you like: Knowledge distill via neuron selectivity transfer[J]. arXiv preprint, arXiv: 1707.01219, 2017

[48] Heo B, Lee M, Yun S, et al. Knowledge transfer via distillation of activation boundaries formed by hidden neurons[C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 3779−3787

[49] Shen Chengchao, Wang Xinchao, Song Jie, et al. Amalgamating knowledge towards comprehensive classification[C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 3068−3075

[50] Wei Yi, Pan Xinyu, Qin Hongwei, et al. Quantization mimic: Towards very tiny CNN for object detection[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 274−290

[51] Ahn S, Hu S X, Damianou A, et al. Variational information distillation for knowledge transfer[C]//Proc of 2019 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 9155−9163

[52] Heo B, Kim J, Yun S, et al. A comprehensive overhaul of feature distillation[C]//Proc of 2019 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 1921−1930

[53] Yim J, Joo D, Bae J, et al. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning[C]//Proc of 2017 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 7130−7138

[54] Srinivas S, Fleuret F. Knowledge transfer with Jacobian matching[C]//Proc of the 35th Int Conf on Machine Learning. New York: ACM, 2018: 4730−4738

[55] Kim J, Park S, Kwak N. Paraphrasing complex network: Network compression via factor transfer[C]//Proc of the 32nd Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2018: 2765−2774

[56] Yang Chuanguang, An Zhulin, Cai Linhang, et al. Hierarchical self-supervised augmented knowledge distillation[C]//Proc of the 30th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2021: 1217−1223

[57] Tian Yonglong, Krishnan D, Isola P. Contrastive representation distillation[J]. arXiv preprint, arXiv: 1910.10699, 2022

[58] Chen Defang, Mei Jian Ping, Zhang Hailin, et al. Knowledge distillation with the reused teacher classifier[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 11923−11932

[59] Wang Xiaobo, Fu Tianyu, Liao Shengcai, et al. Exclusivity-consistency regularized knowledge distillation for face recognition[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 325−342

[60] Chen Pengguang, Liu Shu, Zhao Hengshuang, et al. Distilling knowledge via knowledge review[C]//Proc of 2021 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 5006−5015

[61] You Shan, Xu Chang, Xu Chao, et al. Learning from multiple teacher networks[C]//Proc of the 23rd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2017: 1285−1294

[62] Zhang Chenrui, Peng Yuxin. Better and faster: Knowledge transfer from multiple self-supervised learning tasks via graph distillation for

video classification[C]//Proc of the 27th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2018: 1135−1141

[63] Chen Defang, Mei Jianping, Zhang Yuan, et al. Cross-layer distillation with semantic calibration[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 7028−7036

[64] Liu Li, Huang Qingle, Lin Sihao, et al. Exploring inter-channel correlation for diversity-preserved knowledge distillation[C]//Proc of 2021 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 8251−8260

[65] Jin Ying, Wang Jiaqi, Lin Dahua. Multi-level logit distillation[C]// Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 24276−24285

[66] Chen Yuntao, Wang Naiyan, Zhang Zhaoxiang. DarkRank: Accelerating deep metric learning via cross sample similarities transfer[C]//Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 2852−2859

[67] Liu Yufan, Cao Jiajiong, Li Bing, et al. Knowledge distillation via instance relationship graph[C]//Proc of 2019 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 7089−7097

[68] Peng Baoyun, Jin Xiao, Li Dongsheng, et al. Correlation congruence for knowledge distillation[C]//Proc of 2019 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 5006−5015

[69] Tung F, Mori G. Similarity-preserving knowledge distillation[C]// Proc of 2019 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 1365−1374

[70] Xu Haohang, Fang Jiemin, Zhang Xiaopeng, et al. Bag of instances aggregation boosts self-supervised distillation[J]. arXiv preprint, arXiv: 2107.01691, 2022

[71] Liu Jing, Zheng Tongya, Hao Qinfen. Survey on knowledge distillation with graph: Algorithms classification and application analysis[J]. Journal of Software, 2023, 35(2): 675−710 (in Chinese)
(刘静, 郑铜亚, 郝沁汾. 图知识蒸馏综述: 算法分类与应用分析[J]. 软件学报, 2023, 35(2): 675−710)

[72] Park W, Kim D, Lu Yan, et al. Relational knowledge distillation[C]//Proc of 2019 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3967−3976

[73] Ren Sucheng, Gao Zhengqi, Hua Tianyu, et al. Co-advise: Cross inductive bias distillation[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 16752−16761

[74] Li Duo, Hu Jie, Wang Changhu, et al. Involution: Inverting the inherence of convolution for visual recognition[C]//Proc of 2021 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 12316−12325

[75] Zhao Borui, Song Renjie, Liang Jiajun. Cumulative spatial knowledge distillation for vision transformers[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 6123−6132

[76] Liu Yufan, Cao Jiajiong, Li Bing, et al. Cross-architecture knowledge distillation[C]//Proc of the 16th Asian Conf on Computer Vision. Berlin: Springer, 2022: 179−195

[77] Zhao Weisong, Zhu Xiangyu, He Zhixiang, et al. Cross-architecture distillation for face recognition[C]//Proc of the 31st ACM Int Conf on Multimedia. New York: ACM, 2023: 8076−8085

[78] Hao Zhiwei, Guo Jianyuan, Han Kai, et al. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation [C]//Proc of the 37th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2023: 79570−79582

[79] Li Guopeng, Wang Qiang, Yan Ke, et al. Fuse before transfer: Knowledge fusion for heterogeneous distillation[C]// Proc of 2025 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2025: 3445-3454

[80] Ni Jianyuan, Tang Hao, Shang Yuzhang, et al. Adaptive cross-architecture mutual knowledge distillation[C]//Proc of the 18th Int Conf on Automatic Face and Gesture Recognition. Piscataway, NJ: IEEE, 2024: 1−5

[81] Zheng Zhimeng, Huang Tao, Li Gongsheng, et al. Promoting CNNs with cross-architecture knowledge distillation for efficient monocular depth estimation[J]. arXiv preprint, arXiv: 2404.16386, 2024

[82] Wu Kan, Zhang Jinnian, Peng Houwen, et al. TinyViT: Fast pretraining distillation for small vision transformers[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 68−85

[83] Zhang Jinnian, Peng Houwen, Wu Kan, et al. MiniViT: Compressing vision transformers with weight multiplexing[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 12135−12144

[84] Yang Zhendong, Li Zhe, Zeng Ailing, et al. ViTKD: Practical guidelines for ViT feature knowledge distillation[J]. arXiv preprint, arXiv: 2209.02432, 2022

[85] Zhao Tianyang, Singh K Y, Appalaraju S, et al. No head left behind–multi-head alignment distillation for transformers[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 7514−7524

[86] Ren Sucheng, Wei Fangyun, Zhang Zheng, et al. TinyMIM: An empirical study of distilling MIM pre-trained models[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 3687−3697

[87] Lin Han, Han Guangxing, Ma Jiawei, et al. Supervised masked knowledge distillation for few-shot transformers[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 19649−19659

[88] Bai Yutong, Wang Zeyu, Xiao Junfei, et al. Masked autoencoders enable efficient knowledge distillers[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 24256−24265

[89] He Kaiming, Chen Xinlei, Xie Saining, et al. Masked autoencoders are scalable vision learners[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 16000−16009

[90] Zhao Zhiyu, Huang Bingkun, Xing Sen, et al. Asymmetric masked

distillation for pre-training small foundation models[C]//Proc of 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 18516−18526

[91] Ye Hanjia, Lu Su, Zhan Dechuan. Distilling cross-task knowledge via relationship matching[C]//Proc of 2020 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 12396−12405

[92] Lê H Â, Pham M T. Self-training and multi-task learning for limited data: Evaluation study on object detection[C]//Proc of 2019 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 1003−1009

[93] Li Deng, Wu Aming, Han Yahong, et al. Prototype-guided cross-task knowledge distillation for large-scale models[J]. arXiv preprint, arXiv: 2212.13180, 2022

[94] Auty D, Miles R, Kolbeinsson B, et al. Learning to project for cross-task knowledge distillation[J]. arXiv preprint, arXiv: 2403.14494, 2024

[95] Yuan Mingkuan, Peng Yuxin. CKD: Cross-task knowledge distillation for text-to-image synthesis[J]. IEEE Transactions on Multimedia, 2020, 22(8): 1955−1968

[96] Bayramli B, Ding Yue, Lu Hongtao. Integrating semantic segmentation model for self-supervised scene flow estimation via cross task distillation[C]//Proc of 2024 Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2024: 1−8

[97] Wang Lin, Chae Y, Yoon S H, et al. EvDistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation[C]//Proc of 2021 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 608−619

[98] Qian Biao, Wang Yang, Yin Hongzhi, et al. Switchable online knowledge distillation[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 449−466

[99] Yang Chuanguang, An Zhulin, Zhou Helong, et al. Online knowledge distillation via mutual contrastive learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 10212−10227

[100] Jacob G M, Agarwal V, Stenger B. Online knowledge distillation for multi-task learning[C]//Proc of 2023 IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2023: 2358−2367

[101] Rao Jun, Meng Xv, Ding Liang, et al. Parameter-efficient and student-friendly knowledge distillation[J]. IEEE Transactions on Multimedia, 2024, 26: 4230−4241

[102] Anil R, Pereyra G, Passos A, et al. Large scale distributed neural network training through online distillation[J]. arXiv preprint, arXiv: 1804.03235, 2020

[103] Song Guocong, Chai Wei. Collaborative learning for deep neural networks[C]//Proc of the 32nd Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2018: 1837−1846

[104] Chen Defang, Mei Jianping, Wang Can, et al. Online knowledge distillation with diverse peers[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 3430−3437

[105] Gong Linrui, Lin Shaohui, Zhang Baochang, et al. Adaptive hierarchy-branch fusion for online knowledge distillation[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 7731−7739

[106] Wu Guile, Gong Shaogang. Peer collaborative learning for online knowledge distillation[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 10302−10310

[107] Li Shaojie, Lin Mingbao, Wang Yan, et al. Distilling a powerful student model via online knowledge distillation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(11): 8743−8752

[108] Zhang Linfeng, Song Jiebo, Gao Anni, et al. Be your own teacher: Improve the performance of convolutional neural networks via self distillation[C]//Proc of 2019 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 3712−3721

[109] Lan Xu, Zhu Xiatian, Gong Shaogang. Self-referenced deep learning[C]//Proc of the 14th Asian Conf on Computer Vision. Berlin: Springer, 2018: 284−300

[110] Yang Chenglin, Xie Lingxi, Su Chi, et al. Snapshot distillation: Teacher-student optimization in one generation[C]//Proc of 2019 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 2854−2863

[111] Mobahi H, Farajtabar M, Bartlett P L. Self-distillation amplifies regularization in Hilbert space[C]//Proc of the 34th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 3351−3361

[112] Shen Yiqing, Xu Liwu, Yang Yuzhe, et al. Self-distillation from the last mini-batch for consistency regularization[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 11933−11942

[113] Gou Jianping, Xiong Xiangshuo, Yu Baosheng, et al. Multi-target knowledge distillation via student self-reflection[J]. International Journal of Computer Vision, 2023, 131(7): 1857−1874

[114] Yu Ying, Wei Wei, Tang Hong, et al. Multi-stage training with multi-level knowledge self-distillation for fine-grained image recognition[J]. Journal of Computer Research and Development, 2023, 60(8): 1834−1845 (in Chinese)
(余鹰, 危伟, 汤洪, 等. 多层次知识自蒸馏联合多步骤训练的细粒度图像识别[J]. 计算机研究与发展, 2023, 60(8): 1834−1845)

[115] Lopes R G, Fenu S, Starner T. Data-free knowledge distillation for deep neural networks[J]. arXiv preprint, arXiv: 1710.07535, 2017

[116] Nayak G K, Mopuri K R, Shaj V, et al. Zero-shot knowledge distillation in deep networks[C]//Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 4743−4751

[117] Fang Gongfan, Mo Kanya, Wang Xinchao, et al. Up to 100× faster data-free knowledge distillation[C]//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 6597−6604

[118] Yu Shikang, Chen Jiachen, Han Hu, et al. Data-free knowledge distillation via feature exchange and activation region constraint[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 24266−24275

[119] Li Jingru, Zhou Sheng, Li Liangcheng, et al. Dynamic data-free knowledge distillation by easy-to-hard learning strategy[J].

Information Sciences, 2023, 642: 119202

[120] Wang Yang, Qian Biao, Liu Haipeng, et al. Unpacking the gap box against data-free knowledge distillation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(9): 6280−6291

[121] Zhang Jing, Ju Jialiang, Ren Yonggong. Double-generators network for data-free knowledge distillation[J]. Journal of Computer Research and Development, 2023, 60(7): 1615−1627 (in Chinese)
(张晶, 鞠佳良, 任永功. 基于双生成器网络的 Data-Free 知识蒸馏[J]. 计算机研究与发展, 2023, 60(7): 1615−1627)

[122] Chen Hanting, Wang Yunhe, Xu Chang, et al. Data-free learning of student networks[C]//Proc of 2019 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 3513−3521

[123] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Proc of the 28th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2014: 2672−2680

[124] Fang Gongfan, Song Jie, Shen Chengchao, et al. Data-free adversarial distillation[J]. arXiv preprint, arXiv: 1912.11006, 2020

[125] Luo Liangchen, Sandler M, Lin Zi, et al. Large-scale generative data-free distillation[J]. arXiv preprint, arXiv: 2012.05578, 2020

[126] Qu Xiaoyang, Wang Jianzong, Xiao Jing. Enhancing data-free adversarial distillation with activation regularization and virtual interpolation[C]//Proc of 2021 IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2021: 3340−3344

[127] Micaelli P, Storkey A J. Zero-shot knowledge transfer via adversarial belief matching[C]//Proc of the 33rd Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2019: 9547−9557

[128] Do K, Le T H, Nguyen D, et al. Momentum adversarial distillation: Handling large distribution shifts in data-free knowledge distillation[C]//Proc of the 36th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 10055−10067

[129] Zhao Haoran, Sun Xin, Dong Junyu, et al. Dual discriminator adversarial distillation for data-free model compression[J]. International Journal of Machine Learning and Cybernetics, 2021, 13(5): 1213−1230

[130] Shao Renrong, Zhang Wei, Wang Jun. Conditional pseudo-supervised contrast for data-free knowledge distillation[J]. Pattern Recognition, 2023, 143: 109781

[131] Patel G, Mopuri K R, Qiu Qiang. Learning to retain while acquiring: Combating distribution-shift in adversarial data-free knowledge distillation[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 7786−7794

[132] Fukuda T, Suzuki M, Kurata G, et al. Efficient knowledge distillation from an ensemble of teachers[C]//Proc of the 18th Annual Conf of the Int Speech Communication Association. Grenoble, FR: ISCA, 2017: 3697−3701

[133] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[C]//Proc of the 31st Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2017: 1195−1204

[134] Xiang Liuyu, Ding Guiguang, Han Jungong. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 247−263

[135] Nguyen L T, Lee K, Shim B. Stochasticity and skip connection improve knowledge transfer[C]//Proc of the 28th European Signal Processing Conf. Piscataway, NJ: IEEE, 2021: 1537−1541

[136] Shi Luyao, Jiang Ning, Tang Jialiang, et al. Correlation guided multi-teacher knowledge distillation[C]//Proc of the 30th Int Conf on Neural Information Processing. Berlin: Springer, 2023: 562−574

[137] Pham C, Hoang T, Do T T. Collaborative multi-teacher knowledge distillation for learning low bit-width deep neural networks[C]//Proc of 2023 IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2023: 6424−6432

[138] Zhang Hailin, Chen Defang, Wang Can. Confidence-aware multi-teacher knowledge distillation[C]//Proc of 2022 IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2022: 4498−4502

[139] Cheng Xin, Zhang Zhiqiang, Weng Wei, et al. DE-MKD: Decoupled multi-teacher knowledge distillation based on entropy[J]. Mathematics, 2024, 12(11): 1672

[140] Klingner M, Borse S, Kumar V R, et al. X3KD: Knowledge distillation across modalities, tasks and stages for multi-camera 3D object detection[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 13343−13353

[141] Zhao Lingjun, Song Jingyu, Skinner Katherine A. CRKD: Enhanced camera-radar object detection with cross-modality knowledge distillation[C]//Proc of 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 15470−15480

[142] Wang Yunhe, Xu Chang, Xu Chao, et al. Adversarial learning of portable student networks[C]//Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 4260−4267

[143] Passalis N, Tzelepi M, Tefas A. Heterogeneous knowledge distillation using information flow modeling[C]//Proc of 2020 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 2339−2348

[144] Ganta D P, Gupta H D, Sheng V S. Knowledge distillation via weighted ensemble of teaching assistants[C]//Proc of 2021 IEEE Int Conf on Big Knowledge. Piscataway, NJ: IEEE, 2021: 30−37

[145] Gao Mengya, Wang Yujun, Wan Liang. Residual error based knowledge distillation[J]. Neurocomputing, 2021, 433: 154−161

[146] Morikawa T, Kameyama K. Multi-stage model compression using teacher assistant and distillation with hint-based training[C]//Proc of 2022 IEEE Int Conf on Pervasive Computing and Communications Workshops and other Affiliated Events. Piscataway, NJ: IEEE, 2022: 484−490

[147] He Limuxuan, Hua Bei. Semi-supervised knowledge distillation via teaching assistant[J]. Highlights in Science, Engineering and Technology, 2023, 72: 429−436

[148] Zhao Lei, Ng W W Y, Zhang Jianjun. Distillation of ensemble adaptive teacher assistant knowledge[J]. International Journal of Machine Learning and Cybernetics, 2025, 16(10): 8063-8078

[149] Huo Fushuo, Xu Wenchao, Guo Jingcai, et al. C2KD: Bridging the

modality gap for cross-modal knowledge distillation[C]//Proc of 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 16006−16015

[150] Cho J H, Hariharan B. On the efficacy of knowledge distillation[C]//Proc of 2019 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4793−4801

[151] Yuan Li, Tay F E, Li Guilin, et al. Revisiting knowledge distillation via label smoothing regularization[C]//Proc of 2020 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 3902−3910

[152] Stanton S, Izmailov P, Kirichenko P, et al. Does knowledge distillation really work?[C]//Proc of the 35th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 6906−6919

[153] He Ruifei, Sun Shuyang, Yang Jihan, et al. Knowledge distillation as efficient pre-training: Faster convergence, higher data-efficiency, and better transferability[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 9151−9161

[154] Shen Zhiqiang, Xing E. A fast knowledge distillation framework for visual recognition[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 673−690

[155] Song Jie, Chen Ying, Ye Jingwen, et al. Spot-adaptive knowledge distillation[J]. IEEE Transactions on Image Processing, 2022, 31: 3359−3370

[156] Zhang Tianli, Xue Mengqi, Zhang Jiangtao, et al. Generalization matters: Loss minima flattening via parameter hybridization for efficient online knowledge distillation[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 20176−20185

[157] Wang Jiabao, Chen Yuming, Zheng Zhaohui, et al. CrossKD: Cross-head knowledge distillation for object detection[C]//Proc of 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 16520−16530

[158] Xu Guodong, Liu Ziwei, Li Xiaoxiao, et al. Knowledge distillation meets self-supervision[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 588−604

[159] Li Chenxin, Lin Mingbao, Ding Zhiyuan, et al. Knowledge condensation distillation[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 19−35

[160] Huang Tao, You Shan, Wang Fei, et al. Knowledge distillation from a stronger teacher[C]//Proc of the 36th Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 33716−33727

[161] Choi J, Cho H, Cheung S, et al. ORC: Network group-based knowledge distillation using online role change[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 17335−17344

[162] Shen Zhiqiang. FerKD: Surgical label adaptation for efficient distillation[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 1666−1675

[163] Miles R, Elezi I, Deng Jiankang. VkD: Improving knowledge distillation using orthogonal projections[C]//Proc of 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 15720−15730

[164] Miles R, Mikolajczyk K. Understanding the role of the projector in knowledge distillation[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 4233−4241

[165] Sun Wujie, Chen Defang, Lyu Siwei, et al. Knowledge distillation with refined logits[J]. arXiv preprint, arXiv: 2408.07703, 2024

[166] Xu Kunran, Rui Lai, Li Yishi, et al. Feature normalized knowledge distillation for image classification[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 664−680

[167] Jafari A, Rezagholizadeh M, Sharma P, et al. Annealing knowledge distillation[C]//Proc of the 16th Conf of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2021: 2493−2504

[168] Li Zheng, Li Xiang, Yang Lingfeng, et al. Curriculum temperature for knowledge distillation[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 1504−1512

[169] Chi Zhihao, Zheng Tu, Li Hengjia, et al. NormKD: Normalized logits for knowledge distillation[J]. arXiv preprint, arXiv: 2308.00520, 2023

[170] Wei Yukang, Bai Yu. Dynamic temperature knowledge distillation[J]. arXiv preprint, arXiv: 2404.12711, 2024

[171] Sun Shangquan, Ren Wenqi, Li Jingzhi, et al. Logit standardization in knowledge distillation[C]//Proc of 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 15731−15740

[172] Zhao Borui, Cui Quan, Song Renjie, et al. Decoupled knowledge distillation[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 11943−11952

[173] Yang Zhendong, Zeng Ailing, Li Zhe, et al. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 17139−17148

[174] Park H, Lee J S. Adaptive explicit knowledge transfer for knowledge distillation[J]. arXiv preprint, arXiv: 2409.01679, 2024

[175] Zhao Borui, Cui Quan, Song Renjie, et al. DOT: A distillation-oriented trainer[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 6166−6175

[176] Shen Zhiqiang, He Zhankui, Xue Xiangyang. MEAL: Multi-model ensemble via adversarial learning[C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 4886−4893

[177] Li Peng, Shu Chang, Xie Yuan, et al. Hierarchical knowledge squeezed adversarial network compression[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11370−11377

[178] Ma Haoyu, Chen Tianlong, Hu T K, et al. Undistillable: Making a nasty teacher that cannot teach students[J]. arXiv preprint, arXiv: 2105.07381, 2021

[179] Heo B, Lee M, Yun S, et al. Knowledge distillation with adversarial samples supporting decision boundary[C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 3771−3778

[180] Nguyen-Duc T, Le T, Zhao He, et al. Adversarial local distribution regularization for knowledge distillation[C]//Proc of 2023 IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2023: 4670−4679

[181] Bai Tao, Zhao Jun, Wen Bihan. Guided adversarial contrastive distillation for robust students[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 9643−9655

[182] Huang Bo, Chen Mingyang, Wang Yi, et al. Boosting accuracy and robustness of student models via adaptive adversarial distillation[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 24668−24677

[183] Yue Xinli, Mou Ningping, Wang Qian, et al. Revisiting adversarial robustness distillation from the perspective of robust fairness[C]//Proc of the 37th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2023: 30390−30401

[184] Wang Yuzheng, Chen Zhaoyu, Yang Dingkang, et al. Adversarial contrastive distillation with adaptive denoising[C]//Proc of 2023 IEEE Int Conf on Acoustics, Speech and Signal Processing, Piscataway, NJ: IEEE, 2023: 1−5

[185] Li Changlin, Peng Jiefeng, Yuan Liuchun, et al. Block-wisely supervised neural architecture search with knowledge distillation[C]//Proc of 2020 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 1986−1995

[186] Liu Yu, Jia Xuhui, Tan Mingxing, et al. Search to distill: Pearls are everywhere but not the eyes[C]//Proc of 2020 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 7536−7545

[187] Kang M, Mun J, Han B. Towards oracle knowledge distillation with neural architecture search[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 4404−4411

[188] Wei Zimian, Pan Hengyue, Li Lujun, et al. TVT: Training-free vision transformer search on tiny datasets[J]. arXiv preprint, arXiv: 2311.14337, 2023

[189] Dong Peijie, Li Lujun, Wei Zimian. DisWOT: Student architecture search for distillation without training[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 11898−11908

[190] Lu Liming, Chen Zhenghan, Lu Xiaoyu, et al. UniADS: Universal architecture-distiller search for distillation gap[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 14167−14174

[191] Li Lujun, Dong Peijie, Wei Zimian, et al. Automated knowledge distillation via monte carlo tree search[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 17367−17378

[192] Sohl-Dickstein J, Weiss E A, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 2256−2265

[193] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C]//Proc of the 34th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 6840−6851

[194] Huang Tao, Zhang Yuan, Zheng Mingkai, et al. Knowledge diffusion for distillation[C]//Proc of the 37th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2023, 36: 65299−65316

[195] Guo Qiushan, Mu Yao, Chen Jianyu, et al. Scale-equivalent distillation for semi-supervised object detection[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 14502−14511

[196] Zhou Huajun, Qiao Bo, Yang Lingxiao, et al. Texture-guided saliency distilling for unsupervised salient object detection[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 7257−7267

[197] Wang Luting, Liu Yi, Du Penghui, et al. Object-aware distillation pyramid for open-vocabulary object detection[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 11186−11196

[198] Li Liangqi, Miao Jiaxu, Shi Dahu, et al. Distilling DETR with visual-linguistic knowledge for open-vocabulary object detection[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 6478−6487

[199] Zhou Shengchao, Liu Weizhou, Hu Chen, et al. UniDistill: A universal cross-modality knowledge distillation framework for 3D object detection in bird's-eye view[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 5116−5125

[200] Wu Aming, Deng Cheng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 837−846

[201] Lao Shanshan, Song Guanglu, Liu Boxiao, et al. UniKD: Universal knowledge distillation for mimicking homogeneous or heterogeneous object detectors[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 6362−6372

[202] Zeng Jia, Chen Li, Deng Hanming, et al. Distilling focal knowledge from imperfect expert for 3D object detection[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 992−1001

[203] Liu Zhe, Ye Xiaoqing, Tan Xiao, et al. StereoDistill: Pick the cream from LiDAR for distilling stereo-based 3D object detection[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 1790−1798

[204] Wu Zizhang, Wu Yunzhe, Pu Jian, et al. Attention-based depth distillation with 3D-aware positional encoding for monocular 3D object detection[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 2892−2900

[205] Wang Zeyu, Li Dingwen, Luo Chenxu, et al. DistillBEV: Boosting multi-camera 3D object detection with cross-modal knowledge distillation[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 8603−8612

[206] Zhang Yifan, Dong Zhen, Yang Huanrui, et al. QD-BEV: Quantization-aware view-guided distillation for multi-view 3D object detection[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 3802−3812

[207] Zhao Haimei, Zhang Qiming, Zhao Shanshan, et al. SimDistill:

Simulated multi-modal distillation for BEV 3D object detection[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 7460−7468

[208] Li Yanjing, Xu Sheng, Lin Mingbao, et al. Representation disparity-aware distillation for 3D object detection[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 6692−6701

[209] Gambashidze A, Dadukin A, Golyadkin M, et al. Weak-to-strong 3D object detection with X-ray distillation[C]//Proc of 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 15055−15064

[210] Huang Xun, Wu Hai, Li Xin, et al. Sunshine to rainstorm: Cross-weather knowledge distillation for robust 3D object detection[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 2409−2416

[211] Ji Deyi, Wang Haoran, Tao Mingyuan, et al. Structural and statistical texture knowledge distillation for semantic segmentation[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 16855−16864

[212] Gao Huan, Guo Jichang, Wang Guoli, et al. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 9903−9913

[213] Yang Chuanguang, Zhou Helong, An Zhulin, et al. Cross-image relational knowledge distillation for semantic segmentation[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 12309−12318

[214] Peng Bohao, Tian Zhuotao, Wu Xiaoyang, et al. Hierarchical dense correlation distillation for few-shot segmentation[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 23641−23651

[215] Liu Yuang, Zhang Wei, Wang Jun. Self-decoupling and ensemble distillation for efficient segmentation[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 1772−1780

[216] Hou Yuenan, Zhu Xinge, Ma Yuexin, et al. Point-to-voxel knowledge distillation for LiDAR semantic segmentation[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 8469−8478

[217] Chen Zisheng, Xu Hongbin, Chen Weitao, et al. PointDC: Unsupervised semantic segmentation of 3D point clouds via cross-modal distillation and super-voxel clustering[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 14244−14253

[218] Xu Jianqing, Li Shen, Deng Ailin, et al. Probabilistic knowledge distillation of face ensembles[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 3489−3498

[219] Li Jingzhi, Guo Zidong, Li Hui, et al. Rethinking feature-based knowledge distillation for face recognition[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 20156−20165

[220] Yu Zhipeng, Liu Jiaheng, Qin Haoyu, et al. ICD-face: Intra-class compactness distillation for face recognition[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 20985−20995

[221] Zhao Weisong, Zhu Xiangyu, Guo Kaiwen, et al. Grouped knowledge distillation for deep face recognition[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 3615−3623

[222] Zhang Zhenduo, Chen Yongru, Yang Wenming, et al. Pose-invariant face recognition via adaptive angular distillation[C]//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 3390−3398

[223] Hu Siyuan, Wang Zheng, Hu Peng, et al. PrefAce: Face-centric pretraining with self-structure aware distillation[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 12538−12546

[224] Deng Hanqiu, Li Xingyu. Anomaly detection via reverse distillation from one-class embedding[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 9727−9736

[225] Tien T D, Nguyen A T, Tran N H, et al. Revisiting reverse distillation for anomaly detection[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 24511−24520

[226] Gu Zhihao, Liu Liang, Chen Xu, et al. Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 16355−16363

[227] Binh L M, Woo S. ADD: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images[C]//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 122−130

[228] Ma Ye, Jiang Xu, Guan Nan, et al. Anomaly detection based on multi-teacher knowledge distillation[J]. Journal of Systems Architecture, 2023, 138: 102861

[229] Ristea N C, Croitoru F A, Ionescu R T, et al. Self-distilled masked auto-encoders are efficient video anomaly detectors[C]//Proc of 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 15984−15995

[230] Xing Peng, Jiang Xin, Pan Yonghua, et al. Feature constrained restricted distillation learning for visual anomaly detection[J]. Journal of Software, 2022, 34(9): 4378−4391 (in Chinese)
(邢鹏, 蒋鑫, 潘永华, 等. 基于特征约束蒸馏学习的视觉异常检测[J]. 软件学报, 2022, 34(9): 4378−4391)

[231] Ju Chen, Zheng Kunhao, Liu Jinxiang, et al. Distilling vision-language pre-training to collaborate with weakly-supervised temporal action localization[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 14751−14762

[232] Wu Mingrui, Gu Jiaxin, Shen Yunhang, et al. End-to-end zero-shot HOI detection via vision and language knowledge distillation[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 2839−2846

[233] Monti A, Porrello A, Calderara S, et al. How many observations are enough? Knowledge distillation for trajectory forecasting[C]//Proc of 2018 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 6543−6552

[234] Ma Hengbo, Li Jiachen, Hosseini R, et al. Multi-objective diverse human motion prediction with knowledge distillation[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 8151−8161

[235] Yang C, Kong K, Min S, et al. SEFD: Learning to distill complex pose and occlusion[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 14895−14906

[236] Ye Suhang, Zhang Yingyi, Hu Jie, et al. DistilPose: Tokenized pose regression with heatmap distillation[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 2163−2172

[237] Chen Sichen, Zhang Yingyi, Huang Siming, et al. SDPose: Tokenized pose estimation via circulation-guide self-distillation[C]// Proc of 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 1082−1090

[238] Lu Yichen, Wang Mei, Deng Weihong. Augmented geometric distillation for data-free incremental person ReID[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 7319−7328

[239] Shang Chao, Li Hongliang, Meng Fanman, et al. Incrementer: Transformer for class-incremental semantic segmentation with knowledge distillation focusing on old class[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 7214−7224

[240] Phan M H, Ta T A, Phung S L, et al. Class similarity weighted knowledge distillation for continual semantic segmentation[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 16845−16854

[241] Gu Peiyan, Zhang Chuyu, Xu Ruijie, et al. Class-relation knowledge distillation for novel class discovery[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 16428−16437

[242] Li Jiyong, Azizov D, Li Yang, et al. Contrastive continual learning with importance sampling and prototype-instance relation distillation[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 13554−13562

[243] Zhao Linglan, Lu Jing, Xu Yunlu, et al. Few-shot class-incremental learning via class-aware bilateral distillation[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 11838−11847

[244] Yan Hongwei, Wang Liyuan, Ma Kaisheng, et al. Orchestrate latent expertise: Advancing online continual learning with multi-level supervision and reverse self-distillation[C]//Proc of 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 23670−23680

[245] Ye Fei, Bors A G. Continual variational autoencoder via continual generative knowledge distillation[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 10918−10926

[246] Wen Haitao, Pan Lili, Dai Yu, et al. Class incremental learning with multi-teacher distillation[C]//Proc of 2024 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 28443−28452

[247] Yang Ze, Li Ruibo, Ling Evan, et al. Label-guided knowledge distillation for continual semantic segmentation on 2D images and 3D point clouds[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 18555−18566

[248] Zhang Lin, Shen Li, Ding Liang, et al. Fine-tuning global model via data-free knowledge distillation for non-IID federated learning[C]// Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10164−10173

[249] Shen Yiqing, Zhou Yuyin, Yu Lequan. CD2-pFed: Cyclic distillation-guided channel decoupling for model personalization in federated learning[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10031−10040

[250] Kim H, Kwak Y, Jung M, et al. ProtoFL: Unsupervised federated learning via prototypical distillation[C]//Proc of 2023 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 6447−6456

[251] Lu Yang, Chen Lin, Zhang Yonggang, et al. Federated learning with extremely noisy clients via negative distillation[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 14184−14192

[252] Wang Haozhao, Li Yichen, Xu Wenchao, et al. DaFKD: Domain-aware federated knowledge distillation[C]//Proc of 2023 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 20412−20421

[253] Le Junqing, Tan Zhouyong, Zhang Di, et al. Secure and efficient federated learning for continuous IoV data sharing[J]. Journal of Computer Research and Development, 2024, 61(9): 2199−2212 (in Chinese)
(乐俊青, 谭州勇, 张迪, 等. 面向车联网数据持续共享的安全高效联邦学习[J]. 计算机研究与发展, 2024, 61(9): 2199−2212)



Pan Haiwei, born in 1975. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include remote sensing image processing, medical image processing, data mining, machine learning, and deep learning.

潘海为, 1975 年生。博士, 教授, 博士生导师。CCF 高级会员。主要研究方向为遥感图像处理、医学图像处理、数据挖掘、机器学习、深度学习。



Yu Fengming, born in 1998. PhD candidate. Student member of CCF. His main research interests include knowledge distillation, model compression, and machine learning.

于丰铭, 1998 年生。博士研究生。CCF 学生会员。主要研究方向为知识蒸馏、模型压缩、机器学习。



Zhang Kejia, born in 1981. PhD, associate professor, PhD supervisor. Member of CCF. His main research interests include machine learning, smart healthcare, and big data.

张可佳, 1981 年生。博士，副教授，博士生导师。CCF 会员。主要研究方向为机器学习、智慧医疗、大数据。



Lan Haiyan, born in 1983. PhD, lecturer, master supervisor. Senior member of CCF. Her main research interests include signal processing and information security.

兰海燕, 1983 年生。博士，讲师，硕士生导师。CCF 高级会员。主要研究方向为信号处理、信息安全。



Meng Qingyu, born in 1998. Master. His main research interests include model compression, natural language processing, and machine learning.

孟庆宇, 1998 年生。硕士。主要研究方向为模型压缩、自然语言处理、机器学习。



Li Zhe, born in 1999. PhD candidate. Student member of CCF. His main research interests include computer vision, image generation, and multimodal image processing.

李　哲, 1999 年生。博士研究生。CCF 学生会员。主要研究方向为计算机视觉、图像生成、多模态图像处理。