一种层次化自监督知识蒸馏框架，用于边缘计算上的高效多模态学习

Tarique Dahri [1] , Zulfiqar Ali Memon [1] , Zhenyu Yu [2,*] , Mohd. Yamani Idna Idris [2] , Sheheryar Khan [3] , Sadiq Ahmad [4,†] , Maged Shoman [5] , Saddam Aziz [6] and Rizwan Qureshi [7,*] , *Senior Member, IEEE*

***Abstract*—我们引入了层级自监督知识蒸馏（Layered Self-Supervised Knowledge Distillation, LSSKD）框架，用于训练紧凑的深度学习模型。与传统方法依赖于预训练的教师网络不同，我们的方法在中间特征图上附加辅助分类器，生成多样化的自监督知识，并使得在不同网络阶段间进行一对一的传递。我们的方法在 CIFAR-100 上相较于最先进的 PS-KD 方法平均提升了 4.54 %，相较于 SSKD 提高了 1.14 %，而在 ImageNet 上相比 HASSKD 提升了 0.32 %。在 Tiny ImageNet 和 CIFAR-100 的少样本学习场景下的实验也达到了最先进的结果。这些发现证明了我们的方法在提升模型泛化能力和性能上有效，而不需要大型超参数化的教师网络。重要的是，在推理阶段，所有辅助分类器都可以被移除，不会产生额外的计算成本。这使得我们的模型适合在低计算设备上部署小型语言模型。由于其轻量化设计和适应性，我们的框架特别适用于需要高效和快速推理的多模态传感和信息物理环境。LSSKD 促进了能够在弱监督下从有限的感官数据学习的智能体的发展。**

***Index Terms*—Self-Supervised Learning, Knowledge Distillation, Edge Computing, Multi-modal Learning, Lightweight Deep Models**

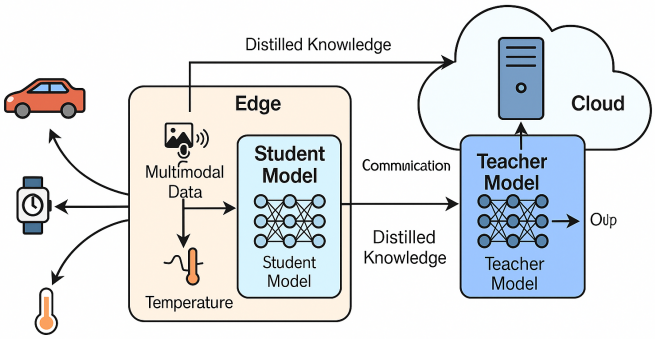

Fig. 1. 本文提出的分层自监督知识蒸馏（LSSKD）框架在物联网场景中的系统概览。资源密集型的教师模型部署在云端，而轻量级的学生模型运行在边缘设备上，以实现高效和低延迟的推理。

[1] Fast School of Computing, National University of Computer and Emerging Sciences, Karachi, Pakistan;
[2] Universiti Malaya, 50603 Kuala Lumpur, Malaysia;
[3] School of Professional Education and Executive Development, The Hong Kong Polytechnic University, Hong Kong;
[4] COMSATS University Islamabad, Wah Campus, 47040, Wah Cantt, Pakistan;
[5] Intelligent Transportation Systems University of Tennessee-Oak Ridge Innovation Institute' s Energy Storage and Transportation Convergent Research Initiative;
[6] Independent Researcher, USA;
[7] Center for research in Computer Vision, University of Central Florida; Orlando, Florida, USA;
† These authors also contributed equally to this work.
* Correspondence: Zhenyu Yu (yuzhenyuyxl@foxmail.com); Rizwan Qureshi (engr.rizwanqureshi786@gmail.com)

I. 引言

深度学习彻底改变了计算机视觉，在图像分类 [1], [2], [3] 、目标检测 [4] 、人脸识别 [5] 、姿态估计 [6] 、活动识别 [7] 和语义分割 [8], [9] 等任务中取得了显著的成功。然而，这些进步通常伴随着计算复杂度 [10] 、内存需求 [11] 和能量消耗 [12] 的增加，阻碍了大型模型在资源受限环境（如边缘设备 [13], [14] ）中的部署。所提出的框架将计算分离在云端和边缘之间，使设备上的推理轻量化，并在远程实现完全训练能力（见图 1 ）。

为了解决这些限制，模型压缩策略例如剪枝 [15] 、量化 [16] 和知识蒸馏（KD） [17] 已被广泛探索。其中，自知识蒸馏（Self-KD） [18] 因为让模型能够从自身预测中学习而不依赖于大型教师网络而受到欢迎。这一范式在保持高性能的同时显著降低了训练成本，特别是当与正则化技术如 $\mathcal{L}_1$ / $\mathcal{L}_2$ 权重衰减 [19] 、dropout [20] 、批量归一化 [21] 和先进的数据增强 [22] 结合使用时。进阶自知识蒸馏（PS-KD） [18] 通过将之前训练轮次的预测作为软目标来形成一个时间自监督环，从而扩展了这一概念。然而，PS-KD 仅利用了最终层的输出，忽略了网络层次结构中有价值的中间表示。

传统的知识蒸馏方法 [23], [18], [24] 通常通过最小化 Kullback-Leibler (KL) 散度来对齐学生网络和教师网络的最终输出分布。虽然这一方法有效，但未能捕捉到嵌入在中间层中的层次化知识。近期的扩展尝试结合中间表示 [25] ，但通常依赖于大型预训练的教师网络，并且未利用多层次标签软化或跨阶段监督策略。

为了克服这些限制，我们提出了分层自监督知识蒸馏（LSSKD）框架。如图 2 所示，LSSKD 在每个瓶颈阶段后引入辅助分类器，生成自监督增强分布（SADs）以支持分层标签软化。最终分类器产生预测的类别分布，而统一的软标签整合了浅层和深层阶段的输出。此外，LSSKD 使用 KL 散度实现从深层到浅层阶段的知识流动，以改进早期层分类器。额外的 $\mathcal{L}_2$ 损失用于最小化辅助和最终特征图之间的差异，促进内部一致性。与以固定硬标签为依赖的先前方法，如 HCSKD [26] 和 HASSKD [27] ，不同的是，LSSKD 利用冲突和软化的标签分布来增强学习的鲁棒性和泛化能力（见图 3 ）。

我们的主要贡献总结如下：

- 层次标签软化：我们提出了一种多层次的标签软化策略，使用来自中间分类器的概率分布，以实现硬目标的逐步优化。
- 跨层蒸馏：该框架不仅在层之间进行知识蒸馏，还在训练迭代过程中通过辅助自监督和特征一致性正则化来实现知识蒸馏。
- 实证表现：LSSKD 相较于 PS-KD 和 SSKD 在 CIFAR-100 上分别平均超出 4.54 % 和 1.14 %，并

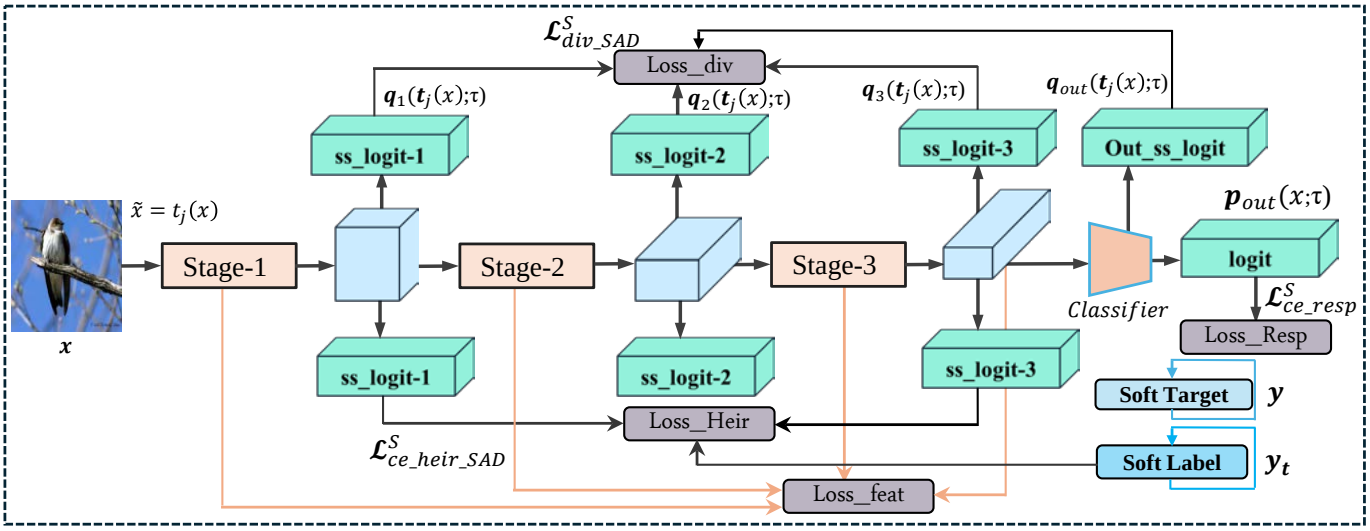

Fig. 2. 所提出的 LSSKD 机制概述。在每个瓶颈阶段之后的辅助分类器生成自监督的增强分布（SADs），而最终的线性分类器提供预测分布。自环表示通过多阶段预测逐步软化硬标签。

且平均超出 HASSKD 1.63 %。在 ImageNet 上，它实现了比 HSSAKD 高 0.32 % 的增益。

II. 相关工作

在知识蒸馏中，小的学生模型通常由大的教师模型或模型集监督。这个概念最早由 Buciluǎ 等人提出，后来 Geoffrey Hinton 对其进行了改进，他使用 softmax 温度来降低概率权重，从教师模型中提取更有用的信息。KD 在解决与大型深度学习模型相关的计算和存储需求方面至关重要。通过将知识从大型教师模型蒸馏到更高效的学生模型，KD 使得在资源受限系统上实现实时处理和部署成为可能。它压缩了知识，有助于计算资源的高效利用和在低资源设备上的部署。多模态感官系统的兴起，例如自主机器人或生物医学传感器，需要能够跨模式泛化并在数据稀缺的情况下有效运行的模型。LSSKD，通过在层级阶段间蒸馏知识和促进高效表示学习，展示了在传感驱动的 AI 系统中部署的强大潜力，在这些系统中，注释数据有限且系统的鲁棒性至关重要。此外，这种方法与自然智能中观察到的层次感官处理的生物启发范式相符。

除了模型压缩，知识蒸馏在多个领域中找到了应用，包括特权学习 [28] 、互学习 [29] 、辅教 [30] 、持续学习 [31] 、文本图像检索 [32] 等 [33] 。最近，在大型语言模型（LLMs） [34] 中，知识蒸馏在将先进能力从主流专有 LLMs [35] （如 GPT-4 [36] ）转移到开源对等体如 LLaMA-2 [37] 和 Mistral [38] 中发挥了关键作用。

虽然知识蒸馏在各种数据集和任务中表现出了强大的性能 [39] ，但在应用于更深层的神经网络时，由于信息瓶颈 [40] 、细粒度细节的丢失 [41] 以及过拟合 [42] 等问题，面临难以达到类似性能的挑战。为了提高准确性，后续的工作 [43], [44], [45] 探索了基于特征的信息以捕捉隐藏在中间层中的表现细节。例如，Fitnet [46] 提议在选定的学生-教师层对之间转移特征图，注意力图 [25] 旨在模仿教师网络的注意力图，而特征相似性保持（FSP） [47] 模仿二阶统计（Gram 矩阵） [22] ，使学生模型能够学习不同特征之间的关系。注意力迁移 [48] 将教师模型中注意力边界的信息转移到学生模型。深度监督（DS） [49] 最初是为了解决收敛问题并通过在深度神经网络的浅层中加入多个辅助分类器 [50] 以促进学习真实标签 [51] 来提高分类性能而引入的。

A. 自监督表示学习

近来的知识蒸馏（KD）进展结合了表示学习，以利用从数据中自然获得的监督。对比表示蒸馏（CRD） [52] 提出了一种目标函数，用来测量教师和学生网络学习的中间表示之间的互信息。自监督知识蒸馏（SSKD） [53] 利用从辅助模型生成的自监督信号之间的相似性进行蒸馏，使用 SimCLR [54] 框架作为从数据中自然监督的预设任务。

分层增强自监督知识蒸馏(Hierarchical Augmented Self-supervised Knowledge Distillation, HASKD）通过引入自监督增强任务，将表征学习与知识蒸馏结合。该任务旨在学习分类任务与辅助自监督任务的统一分布。HASKD 采用基于硬标签的监督方法来指导原始任务和辅助任务。虽然知识蒸馏（KD）认为软标签在类别间的泛化能力通常优于硬目标，HASKD 通过使用过去的预测来平滑硬目标。根据类似策略，我们持续利用过去的预测来优化硬标签，通过中间辅助层来提升最终分类器的泛化能力。这种方法有效利用了中间辅助层，增强了最终分类器的泛化能力。尽管之前关于知识蒸馏的研究主要集中在视觉任务和自然语言处理上，较少关注感知丰富的系统或多模态环境。我们的方法为可穿戴健康传感器、多模态感知系统以及能量受限的机器人中的紧凑推理模型提供了一个有前景的基础。

在自监督知识蒸馏（SSKD）中，学生模型在训练过程中从自身的预测中学习，实现高效的知识转移、增强的泛化能力和降低的复杂度。例如，CSKD 通过对同一类样本间的预测分布进行正则化。Tf-KD，即无教师的知识蒸馏，使用自训练和手动设计的正则化方法。自训练涉及学生模型学习自己的预测，用模型预测替代深度知识。手动设计的正则化使用具有 100% 准确率的虚拟教师模型作为正则项。Tf-KD 在不需要更强的教师模型或额外计算成本的情况下，实现了与正常知识蒸馏相当的性能。PS-KD 使用过去最后一层的预测对硬目标进行持续优化。我们采用了类似策略，使用软目标作为正则化。然而，我们的方法在关键方面有所不同。首先，我们采用自监督学习技术直接从

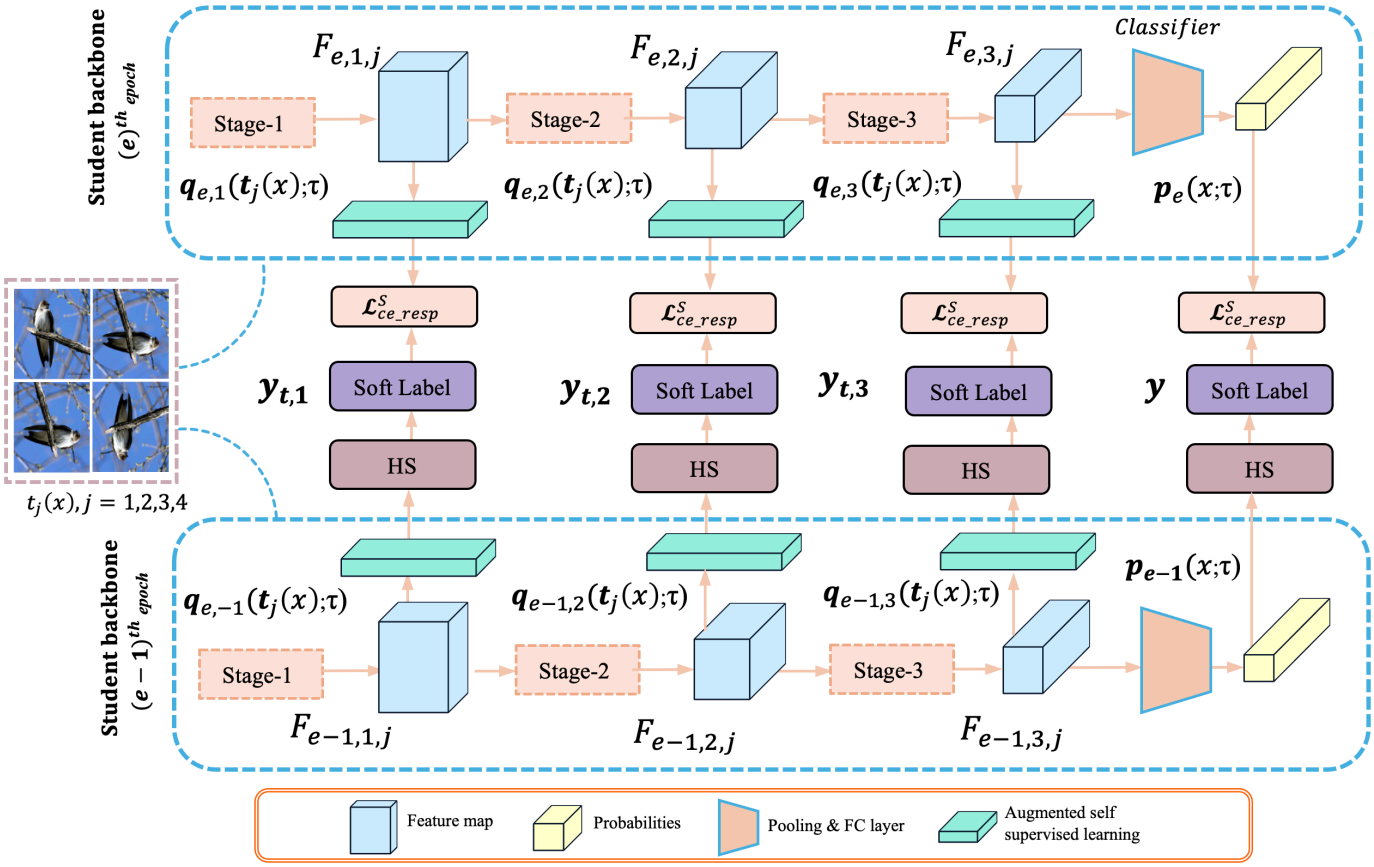

Fig. 3. 时间自我知识蒸馏：e 时刻的学生骨干模仿其 $(e-1)$ 轮时刻版本的预测。

数据中获得自然监督，在整个网络中捕获内在的语义和姿态信息以提高性能。此外，我们的方法利用中间层的预测来软化硬目标。在多个网络层次上考虑概率分布增强了小模型的泛化能力，从而提高了对未见数据的性能。

III. 方法

A. 概述

提出的架构如图 2 所示。我们使用预测类别分布和 SAD 解释学生网络的分级连续学习。

为了计算预测分布，令 $x \in X$ 和 $y \in \{1, \ldots, N\}$ 分别为训练样本和具有 N 类的目标空间。我们训练一个 CNN，记为 $f(\cdot)$，用于将输入样本 x 映射到预测类别概率 $p \in \mathbb{R}^N$。该 CNN 由特征提取器 $\Phi(\cdot)$ 和线性分类器 W 组成，前者编码特征嵌入向量，后者将其转换为类别 logit 概率分布：$f(x) = W^\top \Phi(x) \in \mathbb{R}^N$。为了生成类置信度较低的软概率，我们将带有 softmax 温度的预测类概率定义为 $p(x;\tau) = \sigma\left(\frac{g(z;w)}{\tau}\right) \in \mathbb{R}^N$ [23]，其中 τ 为温度超参数，$w \in \mathbb{R}^{d \times N}$。

对于自监督任务，我们利用原始标签和自监督增强标签（SAL）的联合分布。在使用数据集进行训练时，有 N 个标签经过 M 转换。因此，统一的标签空间变成了 $N \times M$，其中 $\times$ 表示元素级乘法。例如，我们使用旋转转换作为自监督来训练 CIFAR-100（100 个标签），这构造了图像的 $M = 4$ 个旋转变换（0°，90°，180°，270°），然后它学习所有可能组合的联合分布，即 400 个标签。

令 $x' = \{t_j\}_{j=1}^M$ 表示使用转换 t 的自监督增强训练样本。为了定义具有 softmax 温度的 SAD，我们采用以下方法：

其中 τ 是温度超参数，$w \in \mathbb{R}^{d \times N^* M}$。

现代卷积神经网络（CNN）使用阶段式构建模块来学习不同抽象层次的特征。为了从网络中寻求阶段式的表征监督，我们选择在每个阶段后附加一个辅助分支，由此产生 L 个分支 $\{c_l(\cdot)\}_{l=1}^L$，其中 $c_l(\cdot)$ 表示第 l 阶段后的辅助分支，受到 [50] 的启发。然而，如 [55] 实证验证，仅仅附加线性辅助分类器而不添加额外特征提取模块可能无法有效挖掘有意义的信息。因此，在每个辅助分支中包含特征提取模块、一个全局平均池化层和一个具有期望维度的线性分类器，以用于所用的辅助任务，例如，我们的自我监督增强任务的 $N \times M$。我们用 $p_e(x;\tau)$ 表示第 e 个时期的原始类分布，用 $q_{e,l}(x;\tau)$ 表示第 l 个阶段在第 e 个时期的 SAD，用 $q_{e-1,l}(t_j(x);\tau)$ 表示前一个时期的 SAD，用 $f_S(\cdot)$ 表示学生主干网络。

B. 连续自我知识蒸馏

在我们的方法中，我们利用过去几个时代的预测结果逐步软化后续时代的硬目标。具体而言，对于从最深分类器获得的预测类别分布，我们使用以下公式计算软目标：

$$y^s = (1-\alpha)y^* + \alpha p_S^{e-1}(x;\tau), \tag{1}$$

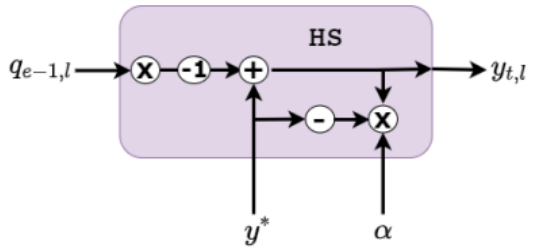

Fig. 4. 用于辅助分类器输出的混合自监督（HS）融合模块的架构。由前一时期的分布 $q_{e-1,l}$ 根据目标 y^* 进行调整，并通过因子 α 加权，以获得当前的预测 $y_{t,l}$ 。

，其中 $p_S^{e-1}(x;\tau)$ 表示来自前一个周期的预测类别概率分布，这有助于软化目标 y^s 。这里，y^s 表示 N 类别的独热编码标签。

对于从第 l 阶段浅层分类器获得的自监督增强分布，我们有以下用于将硬标签转换为软标签的公式：

其中 $q_S^{e-1,l}(t_j(x);\tau)$ 表示前一个周期中第 l 阶段辅助分类器的自监督增强分布。此处，$y_{*,t}$ 表示 N 类的转换后的单热编码，如图 4 所示。通过 M 转化，$y_{*,t}$ 的指示函数可以描述如下：

$$y_{*,t} = \frac{1}{K} 1_{\{\tilde{x}_i \in K\}} = \begin{cases} 1, & \text{when } \tilde{x}_i \in K \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

，其中 $\tilde{x}$ 代表转换后的样本，而 $K = N \times M$ 表示转换后样本的标签空间。

C. 训练学生网络

在我们的训练过程中，我们采用了一种端到端的方法，使学生模型能够从中间层和输出层中提取知识，增强其泛化能力。图 3 提供了我们方法的详细图解。

首先，我们基于中间层和最终层的输出计算两个损失。对于从最深的分类器获得的预测类别分布，记为 $p_S^e(x;\tau)$ ，我们使用交叉熵损失函数

$$\mathcal{L}_{S\text{ce resp}} = \mathcal{L}_{\text{ce}}\left(p_S^e(x;\tau), y_S\right), \tag{3}$$

来计算软目标以获得更具信息性的标签 y_S ，其中 $\mathcal{L}_{S\text{ce resp}}$ 代表学生模型的响应交叉熵损失。在这里，$\tau = 1$ ，并且 y_S 表示从先前预测 $p_S^{e-1}(x;\tau)$ 中推导出的软目标。

其次，我们将转换后的样本 $x^{\sim} = \{t_j\}_{j=1}^M$ 输入到学生主干 $f_S(\cdot)$ 中，以获取在每个阶段的转换后特征图。这些特征图随后被输入到 L 辅助分类器 $c_{S,l}(\cdot)$ 中，以计算在 $(e-1)$ 个周期时的自监督增强分布 $q_S^{e-1,l}(t_j(x);\tau)$ ，从而为 e 个周期软化硬目标。具有跨 M 转换的自监督软化标签的分层损失计算如下：

$$\mathcal{L}_{S,\text{ce heir SAD}} = \frac{1}{M}\sum_{j=1}^{M}\sum_{l=1}^{L}\mathcal{L}_{\text{ce}}\left(q_S^{e,l}(t_j(x);\tau), y_S\right), \tag{4}$$

，其中 $\mathcal{L}_{S,\text{ce heir SAD}}$ 表示 SAD 和软标签之间的分层交叉熵损失。标签 $y_{S,t,l}^{(2)}$ 表示原始监督标签（N ）和自监督增强标签（M ）的联合软化分布。

最后，我们结合这两个损失来获得标签监督（LS）损失，该损失模仿通过对硬目标平滑计算出的软标签：

$$\mathcal{L}_{LS}^S = \mathcal{L}_{\text{ce-resp}}^S + \mathcal{L}_{\text{ce-heir-SAD}}^S. \tag{5}$$

D. 基于特征和分类器的蒸馏

我们加入了两个额外的损失来利用网络后期阶段的知识。首先，为了将后期分类器的知识传递到早期分类器，我们引入了一个 KL 散度损失。此损失鼓励浅层分类器模仿由深层分类器生成的自监督增强分布：

$$\mathcal{L}_{\text{div-SAD}}^S = \frac{1}{M}\sum_{j=1}^{M}\sum_{l=1}^{L}\tau^2 DKL\left(q_S^l(t_j(x);\tau)\middle\|q_S^O(t_j(x);\tau)\right) \tag{6}$$

这里，$q_S^O(t_j(x);\tau)$ 代表由更深层的分类器生成的自监督增强分布。

其次，我们利用特征图中包含的更丰富的表示信息进行知识蒸馏。具体来说，我们计算瓶颈阶段的自监督增强特征图与最后一层的特征图之间的 $\mathcal{L}_2$ 损失函数：

$$\mathcal{L}_S^{\text{feat}} = \left\|F_t^l - F_t^o\right\|_2^2 \tag{7}$$

这里，F_t^l 代表 l -th 瓶颈阶段的平均池化层的特征图，F_t^o 代表最后一层的平均池化层的特征图，t 表示传递给 $f_S(\cdot)$ 主干的变换后数据样本。

我们的 LSSKD 框架的层次结构很好地与生物感知系统和机器人代理所观察到的阶段性信息处理相一致。此外，其自监督的转换机制可以通过设计特定模态的增强策略扩展到异构传感器模态（例如，雷达、视觉和音频），这使得 LSSKD 成为多模态传感器融合任务的一个有前景的候选者。

我们将这两个损失结合起来，并将其称为自监督（IS）损失，因为它们都是利用网络本身的知识进行监督的：

$$\mathcal{L}_S^{\text{IS}} = \mathcal{L}_S^{\text{div}} + \mathcal{L}_S^{\text{feat}} \tag{8}$$

最后，我们将标签监督损失和自身监督损失相加，以获得总损失，其中超参数 β 和 γ 为每个损失分配权重：

超参数 β 和 γ 根据其优化贡献为每个损失分配权重。这个过程使用软标签和硬标签来优化特征图和辅助分类器，同时确保早期阶段模仿后期阶段。在推理过程中，所有辅助分支都可以被移除，消除任何额外的计算负担。

IV. 实验结果

A. 实验设置

所有实验均在 NVIDIA RTX 4090 GPU 上进行。我们在三个基准数据集上评估了我们提出的 LSSKD 框架的有效性和泛化性：CIFAR-100 [56] 、ImageNet [57] 和 Tiny-ImageNet [58] 。使用了一组多样化的教师-学生架构，包括 PreAct ResNet-18 [55] 、ResNet [59] 、WRN [60] 、ResNeXt [61] 、VGG [62] 、MobileNetV2 [63] 和 ShuffleNet [64], [65] ，如表 II 所总结。在 ImageNet 上，我们使用 ResNet-18 作为教师和学生，以验证可扩展性，结果如表 IV-A 。

我们应用了标准的数据增强策略，遵循 [59] ，包括 4 像素填充、随机裁剪到 32×32 、水平翻转和归一化。所有模型均使用动量为 0.9 的随机梯度下降（SGD）进行训练，共进行 240 个周期。初始学习率设置为 0.05，并在第 150 和第 210 个周期时衰减 10 倍。批量大小为 64，权重衰减

设置为 5×10^{-5} 。LSSKD 引入了三个超参数：α 、β 和 γ ，它们分别控制软标签、辅助监督和特征级一致性的贡献。这些值是通过从训练数据中保留 10 % 验证集进行调优的。在所有实验中的最终使用值为 $\alpha = 0.8$ 、$\beta = 0.1$ 和 $\gamma = 0.1$ 。

我们提出的 LSSKD 在 CIFAR-100 上不同的学生-教师架构中实现了最先进的性能，而不需要任何预训练的教师模型。如表 II 所示，LSSKD 始终优于之前的 KD 方法，包括 SSKD [66] 、FitNet [46] 和 CRD [52] 。具体来说，在 CIFAR-100 上，LSSKD 相较于 SSKD 实现了平均 4.54 % 的提升。在 ImageNet 上，它在使用 ResNet-18 作为教师和学生模型时，实现了 0.33 % 的 top-1 增益（见表 IV-A ）。为了确保一致和公平的评估，我们在所有方法中应用了相同的训练流程和数据增强，包括随机裁剪、水平翻转以及具有相同学习率计划的 SGD 训练。这些结果验证了 LSSKD 在实现高准确率的同时保持训练和推理效率的优势。

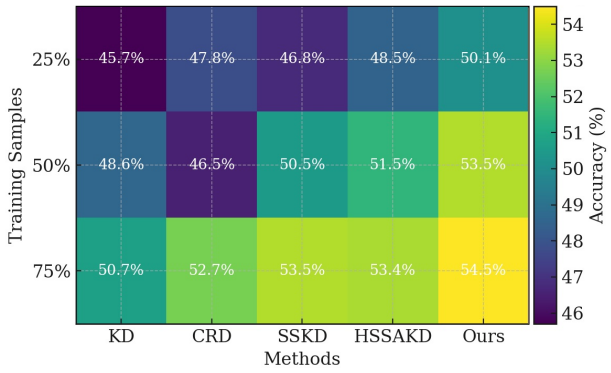

Fig. 5. 在小样本场景下使用不同比例的训练数据下，Tiny-ImageNet 上的 Top-1 准确率 (%) 比较。以 ResNet56-ResNet20 作为教师–学生对。

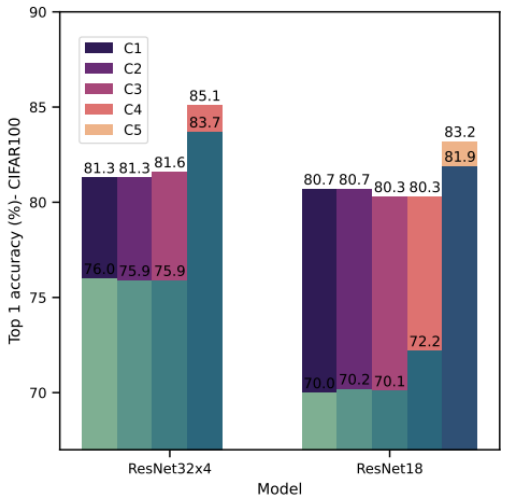

Fig. 6. 与 HASKD [26] 的阶段性比较。红色条表示 HASKD 的准确率，而蓝色条表示我们的方法。C4 (R32 × 4) 和 C5 (R18) 是最终分类器；其他是辅助分类器。

B. 少样本设定下的泛化

为了评估 LSSKD 的泛化能力，我们在使用 CIFAR-100 和 Tiny-ImageNet 的少样本学习场景下对其性能进行了评估。在每种情况下，我们保留训练样本的 25 % 、50 % 和 75 % ，同时保持完整的测试集不变。分层抽样策略确保不同比例之间的类别代表性平衡。如图 5 所示，在所有数据稀缺水平上，LSSKD 显著优于传统的 KD 方法，如 KD [23] 、CRD [52] 和 SSKD [66] 。这些结果展示了我们框架的稳健性和适应性，即使在数据不足的情况下，这也是实际应用中标注数据有限时的关键属性。

C. 与最新自蒸馏方法的比较

我们对 LSSKD 与近期的自蒸馏方法进行了全面比较，包括 LS [71] 、CSKD [72] 、TFKD [73] 和 PS-KD [18] 。所有评估均在 CIFAR-100 上进行，为了公平起见，使用一致的 ResNet34–ResNet18 架构。结果如表 III 所示。除了标准的自蒸馏外，我们还探讨了数据增强策略对性能的影响，特别是 Cutout 和 CutMix（COC） [22], [74] 。LSSKD 在所有基准中始终表现优异，相比 PS-KD 平均提高了 4.56 %，结合 COC 后额外提升了 2.58 %。值得注意的是，在像 MobileNetV2 和 ShuffleNetV2 这样的轻量级模型上，我们的方法在 top-1 准确率上提高了最多 5.6 %，展示了优异的正则化和可扩展性。

V. 消融研究与分析

A. 辅助分类器的影响

我们评估了中间辅助分类器在增强学生网络的表示能力方面的作用。受 HASKD [26] 启发，LSSKD 在瓶颈层跨阶段附加分类器，监督源自最深层分类器的软标签。如图 6 所示，我们的方法显著提高了每个阶段的分类准确性，相较于 HASKD。观察到的提升验证了两个关键假设：(1) 通过软化的标签进行分层监督比传统的硬目标更具信息量，以及 (2) 最深层的分类器提供更强的泛化能力，使其成为监督较浅模块的理想选择。在 ResNet-18 和 ResNet-32x4 上的实验证实了所有辅助分类器的显著性能提升。

B. 分层软化标签的有效性

不同于之前的方法如 PS-KD [18] ，仅依赖最终层的预测来软化硬目标，我们的 LSSKD 引入了使用中间分类器的层次软化策略。每一层的概率输出都有助于细化监督信号，形成多阶段软化标签。表 IV 证明这种层次策略在多个架构中持续提高学生表现。例如，我们的方法在 WRN-40-2 上提升了高达 8.79 % 的准确率，在 WRN-16-2 上提升了 6.39 % 的准确率，表明相比于基线硬标签训练，软化的层次监督增强了优化和泛化能力。

C. 跨网络架构的鲁棒性

为了评估 LSSKD 的普遍适用性，我们在包括 ResNet、WRN 和 ShuffleNet 家族的各种基线架构中评估了其性能。如表格 IV 所示，我们的方法无论网络深度或复杂性如何都能持续提高准确率。例如，LSSKD 在 ResNet20 上获得了 +3.05 % 的增益，在 ResNet56 上获得了 +3.96 %，在 WRN-40-2 上则高达 +8.79 %。这些结果证实了 LSSKD 在增强多样模型骨干方面的稳健性和多功能性，使其适用于轻量级和深度神经网络。

尽管 LSSKD 表现强劲，但若干方面仍需进一步探索：(1) 可扩展性。在更广泛领域的极大规模数据集上，LSSKD 的有效性尚未得到充分探索。(2) 任务泛化。未来的工作将研究其在图像分类之外任务的扩展，例如目标检测、机器

TABLE I
LSSKD 在 CIFAR-100 上的表现与基线方法的比较。教师网络：ResNet34；学生网络：ResNet18。在其他比较中，我们使用相同的 ResNet34-ResNet18 对以确保一致性。所有实验均在批量大小为 64 的条件下进行，学习率计划在第 150 个和第 210 个 epoch 时减小 10 倍。

Accuracy	Teacher	Student	KD [23]	AT [25]	CC [67]	SP [45]	RKD [44]	CRD [52]	SSKD [66]	HSSAKD [68]	MOKD [69]	Ours
Top-1	73.31	69.75	70.66	70.70	69.96	70.62	71.34	71.38	71.62	72.16	72.3	72.45
Top-5	91.42	89.07	89.88	90.0	89.17	89.80	90.37	90.49	90.67	90.85	90.9	91.15

TABLE II
我们的自蒸馏方法与各种教师-学生网络在 CIFAR-100 数据集上的蒸馏方法的 Top-1 准确率（%）对比。加粗的数字表示最佳准确率。已经加入了来自 [68] 的数据以提供全面的比较。

Distillation Mechanism	Teacher / Student	WRN-40-2 / WRN-16-2	WRN-40-2 / WRN-40-1	Resnet56 / Resnet20	ResNet32x4 / ResNet8x4	WRN-40-2 / ShuffleNetV1	ResNet32x4 / ShuffleNetV2
	Baseline	73.57	71.95	69.62	72.95	71.74	72.96
KD [23]		75.23	73.90	70.91	73.54	75.83	75.43
FitNet [46]		75.30	74.30	71.21	75.37	76.27	76.91
AT [25]		75.64	74.32	71.35	75.06	76.51	76.32
AB [48]		71.26	74.55	71.56	74.31	76.43	76.40
VID [43]		75.31	74.23	71.35	75.07	76.24	75.98
RKD [44]		75.33	73.90	71.67	74.17	75.74	75.42
SP [45]		74.35	72.91	71.45	75.44	76.40	76.43
CC [67]		75.30	74.46	71.44	74.40	75.63	75.74
CRD [52]		75.81	74.76	71.83	75.77	76.37	76.51
SSKD [66]		76.16	75.84	70.80	75.83	76.71	77.64
SemCKD [70]		75.02	73.54	71.54	75.89	77.39	78.26
Proposed Method		76.89	78.35	72.67	76.27	77.45	79.43

TABLE III
在 CIFAR-100 数据集上，不同学生网络中 SOTA 自蒸馏方法的 Top-1 准确率（%）比较。

Method	ResNet18	ResNet101	MobileNetV2
Baseline	75.87	79.25	68.38
LS [71]	79.06	80.16	–
CSKD [72]	78.7	79.53	–
TFKD [73]	77.36	77.36	70.88
$TFKD_{self}$	77.11	77.11	70.96
PS-KD [18]	79.18	80.57	69.89
Proposed MSAKD	83.16	82.24	73.95

TABLE IV
基准比较。以下标形式写的数字表示相对于基准方法的准确性提高，例如 72.67(+3.05) 表示在基准准确性 69.62 % 上提高了 3.05 % 。

Method	Baseline	Proposed MSAKD
ResNet20	69.62	72.67(+3.05)
ResNet56	71.83	78.79(+3.96)
WRN-16-2	76.77	83.16(+6.39)
WRN-40-2	76.89	85.68(+8.79)
ResNet18	73.57	76.89(+3.32)
ResNet50	75.81	82.63(+6.22)

翻译和模型压缩。(3) 模型简化。进一步的努力可以聚焦于通过剪枝更深层的网络层来降低模型复杂度，从而增强该方法在资源受限环境中的适用性。

VI. 结论

我们提出了 LSSKD 框架，该框架通过从中间层和最终层进行知识蒸馏来提高模型性能。我们的方法在 CIFAR-100 和 ImageNet 上实现了最先进的结果，而不依赖于大型预训练的教师网络。值得注意的是，所有辅助分类器在推理过程中都被移除，确保没有额外的计算成本。LSSKD 的高效性和紧凑性使其在成本效益高的设备上部署小型语言模型时尤为有利，从而能够以更少的参数和最低的资源需求进行有效训练。这支持在隐私敏感和低延迟场景中鲁棒且准确的模型。此外，通过将大型模型蒸馏成紧凑版本，LSSKD 减少了碳足迹并扩大了高性能人工智能在主要科技公司之外的使用。未来的工作中，我们计划探索其在目标检测、机器翻译和模型压缩等任务中的适用性。

References

[1] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 international interdisciplinary PhD workshop (IIPhDW)*. IEEE, 2018, pp. 117–122.

[2] Z. Yu and P. Wang, "Capan: Class-aware prototypical adversarial networks for unsupervised domain adaptation," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[3] P. Wang, Y. Yang, and Z. Yu, "Multi-batch nuclear-norm adversarial network for unsupervised domain adaptation," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[4] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[5] Y. Huang, J. Wu, X. Xu, and S. Ding, "Evaluation-oriented knowledge distillation for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 740–18 749.

[6] F. Zhang, X. Zhu, and M. Ye, "Fast human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3517–3526.

[7] Y. Liu, K. Wang, G. Li, and L. Lin, "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 5573–5588, 2021.

[8] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 41–65, 2018.

[9] Y. Luo, J. Wang, X. Yang, Z. Yu, and Z. Tan, "Pixel representation augmented through cross-attention for high-resolution remote sensing imagery segmentation," *Remote Sensing*, vol. 14, no. 21, p. 5415, 2022.

[10] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The computational limits of deep learning," *arXiv preprint arXiv:2007.05558*, 2020.

[11] S. Yu, H. Jiang, S. Huang, X. Peng, and A. Lu, "Compute-in-memory chips for deep learning: Recent trends and prospects," *IEEE circuits and systems magazine*, vol. 21, no. 3, pp. 31–56, 2021.

[12] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 09, 2020, pp. 13 693–13 696.

[13] J. Ji, Z. Shu, H. Li, K. X. Lai, M. Lu, G. Jiang, W. Wang, Y. Zheng, and X. Jiang, "Edge-computing based knowledge distillation and multi-task learning for partial discharge recognition," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[14] L. Yin, L. Wang, Z. Cai, S. Lu, R. Wang, A. AlSanad, S. A. AlQahtani, X. Chen, Z. Yin, X. Li *et al.*, "Dpal-bert: A faster and lighter question answering model." *CMES-Computer Modeling in Engineering & Sciences*, vol. 141, no. 1, 2024.

[15] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Advances in neural information processing systems*, vol. 28, 2015.

[16] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4820–4828.

[17] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 535–541.

[18] K. Kim, B. Ji, D. Yoon, and S. Hwang, "Self-knowledge distillation with progressive refinement of targets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6567–6576.

[19] S. J. Nowlan and G. E. Hinton, "Simplifying neural networks by soft weight sharing," in *The Mathematics of Generalization*. CRC Press, 2018, pp. 373–394.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[22] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

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

[24] Z. Yu, "Improved implicit diffusion model with knowledge distillation to estimate the spatial distribution density of carbon stock in remote sensing imagery," *arXiv preprint arXiv:2411.17973*, 2024.

[25] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[26] C. Yang, Z. An, L. Cai, and Y. Xu, "Hierarchical self-supervised augmented knowledge distillation," *arXiv preprint arXiv:2107.13715*, 2021.

[27] ——, "Knowledge distillation using hierarchical self-supervision augmented distribution," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[28] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," *arXiv preprint arXiv:1511.03643*, 2015.

[29] D. Chen, J.-P. Mei, C. Wang, Y. Feng, and C. Chen, "Online knowledge distillation with diverse peers," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 3430–3437.

[30] W. Son, J. Na, J. Choi, and W. Hwang, "Densely guided knowledge distillation using multiple teacher assistants," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9395–9404.

[31] S. Li, T. Su, X. Zhang, and Z. Wang, "Continual learning with knowledge distillation: A survey," *Authorea Preprints*, 2024.

[32] J. Rao, L. Ding, S. Qi, M. Fang, Y. Liu, L. Shen, and D. Tao, "Dynamic contrastive distillation for image-text retrieval," *IEEE Transactions on Multimedia*, vol. 25, pp. 8383–8395, 2023.

[33] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[34] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili *et al.*, "A survey on large language models: Applications, challenges, limitations, and practical usage," *Authorea Preprints*, 2023.

[35] X. Xu, M. Li, C. Tao, T. Shen, R. Cheng, J. Li, C. Xu, D. Tao, and T. Zhou, "A survey on knowledge distillation of large language models," *arXiv preprint arXiv:2402.13116*, 2024.

[36] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[37] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[38] S. Karamcheti, L. Orr, J. Bolton, T. Zhang, K. Goel, A. Narayan, R. Bommasani, D. Narayanan, T. Hashimoto, D. Jurafsky *et al.*, "Mistral–a journey towards reproducible language model training," 2021.

[39] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4794–4802.

[40] Y. Kim, W. Nam, H. Kim, J.-H. Kim, and G. Kim, "Curiosity-bottleneck: Exploration by distilling task-specific novelty," in *International conference on machine learning*. PMLR, 2019, pp. 3379–3388.

[41] Z. Hao, J. Guo, D. Jia, K. Han, Y. Tang, C. Zhang, H. Hu, and Y. Wang, "Learning efficient vision transformers via fine-grained manifold distillation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9164–9175, 2022.

[42] T. Chen, Z. Zhang, S. Liu, S. Chang, and Z. Wang, "Robust overfitting may be mitigated by properly learned smoothening," in *International Conference on Learning Representations*, 2020.

[43] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9163–9171.

[44] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[45] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1365–1374.

[46] R. Adriana, B. Nicolas, K. S. Ebrahimi, C. Antoine, G. Carlo, and B. Yoshua, "Fitnets: Hints for thin deep nets," *Proc. ICLR*, vol. 2, no. 3, p. 1, 2015.

[47] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4133–4141.

[48] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3779–3787.

[49] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial intelligence and statistics*. Pmlr, 2015, pp. 562–570.

[50] D. Chen, J.-P. Mei, H. Zhang, C. Wang, Y. Feng, and C. Chen, "Knowledge distillation with the reused teacher classifier," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 933–11 942.

[51] S. Luo, D. Chen, and C. Wang, "Knowledge distillation with deep supervision," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–8.

[52] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[53] W. Liu, S. Nie, J. Yin, R. Wang, D. Gao, and L. Jin, "Sskd: Self-supervised knowledge distillation for cross domain adaptive person re-identification," in *2021 7th IEEE International Conference on Network Intelligence and Digital Content (ICNIDC)*. IEEE, 2021, pp. 81–85.

[54] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 630–645.

[56] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[58] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, p. 3, 2015.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[60] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[61] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[63] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[64] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[65] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[66] G. Xu, Z. Liu, X. Li, and C. C. Loy, "Knowledge distillation meets self-supervision," in *European conference on computer vision*. Springer, 2020, pp. 588–604.

[67] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5007–5016.

[68] C. Yang, Z. An, L. Cai, and Y. Xu, "Knowledge distillation using hierarchical self-supervision augmented distribution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 2, pp. 2094–2108, 2024.

[69] K. Song, J. Xie, S. Zhang, and Z. Luo, "Multi-mode online knowledge distillation for self-supervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 848–11 857.

[70] D. Chen, J.-P. Mei, Y. Zhang, C. Wang, Z. Wang, Y. Feng, and C. Chen, "Cross-layer distillation with semantic calibration," in *AAAI Conference on Artificial Intelligence*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:227335337

[71] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[72] S. Yun, J. Park, K. Lee, and J. Shin, "Regularizing class-wise predictions via self-knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 876–13 885.

[73] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3903–3911.

[74] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.