西安工程大学

# 学术学位硕士研究生学位论文

# 选 题 报 告

| | |
|---|---|
| 学位论文题目 | 基于多级特征先验引导的知识蒸馏方法研究 |
| 研究生姓名 | 曹琦智 |
| 学　　号 | 220411030 |
| 所在学院 | 电子信息学院 |
| 学科、专业 | 控制科学与工程 |
| 导师姓名 | 张凯兵 |
| 开题时间 | 2023 年 11 月 29 日 |

西安工程大学研究生院

# 填 写 要 求

一、论文题目是论文中心思想的高度概括，要求准确、规范、用词科学、简洁，一般不能超过 25 个汉字。

二、参考文献（参考文献应为近五年的 50 篇以上，其中外文资料不少于 1/3）。

三、硕士学位论文选题报告通过后，此表一式两份，均由学院保存（其中一份在研究生获得硕士学位后装入研究生个人学位论文档案袋存档）。

四、本表个别栏目填写空间不足时，可续页。

五、选题报告为 A4 纸双面打印，字号：宋体、小四，1.3 倍行距，于左侧装订成册。

# 一、 选题依据

**1.选题类型**

本课题类型为基础研究。

**2.选题来源**

本课题来源于国家自然科学基金研究计划：基于分治策略与增量字典学习的图像超分辨重建方法研究（项目编号: 61971339）。

**3.研究意义**

随着深度学习方法的兴起，以深度神经网络为代表的深度学习模型在计算机视觉任务中取得了巨大的成功，使得计算机系统能够准确、快速地识别检测识别目标物体。这在自动驾驶汽车、监控系统、医疗影像分析和军事应用等领域产生了深远的影响，对提高生产效率和改善生活质量有着重要意义。然而，当前主流的图像分类和目标检测方法多是基于大型深度神经网络（DNNs）或其集合，这些模型往往拥有数百万甚至上亿的参数。例如，VGGNet 等深度神经网络已经在各类视觉任务中展现了良好的性能，但这种复杂模型带来的计算开销巨大，使其不适用部署到计算资源有限的移动终端和嵌入式设备中，大大限制了图像分类、目标检测等技术的实际应用。

为了解决这一问题，众多模型轻量化方法被提出，其中知识蒸馏[1]（Knowledge Distillation, KD）作为当下备受关注的模型轻量化方法之一，已经在自然语言处理、语义识别和计算机视觉等领域取得良好的效果。知识蒸馏利用一个带有较多网络参数规模较大、性能较好的模型作为教师模型，把它的知识迁移到一个小型神经网络（学生模型），由此获得一个性能与教师性能能够相媲美甚至性能更好的学生网络模型，并且其规模和参数量远小于教师模型，使其在硬件资源有限的轻量型设备上也能够实现部署并取得优秀的应用效果。

然而，在现有多数知识蒸馏方法中：

（1） 只有对应层的中间特性知识被蒸馏，跨层级特性很少被考虑，导致特性知识难以被充分地从教师模型迁移至学生模型；

（2） 在学生模型的训练过程中，教师模型以往只被作为监督信号来辅助学生模型的训练，忽视了它的先验引导作用；

（3） 训练过程中学生模型和一个甚至多个教师模型的同时加载需要更多的存储和计算资源，很大程度上加重了模型训练负担。

为了弥补现存知识蒸馏方法中的以上不足，我们提出基于多级特征教师先验引导的知识蒸馏方法研究，力求在压缩模型参数量的同时获得性能良好甚至更优的神经网络模型，从而使其能够在移动终端、嵌入式平台上等存储和计算资源有限的轻量型设备上部署，并探索硬件要求更低的训练方式，提升模型训练速度，对深度学习模型的应用和发展具有重要的现实研究意义。

## 4. 国内外的研究现状

深度学习由于对目标多样性变化具有很好的鲁棒性，近年来得到广泛的关注并取得快速的发展。然而，性能越好的深度学习模型往往需要更多的存储和计算资源，使得一些模型在物联网、移动嵌入式等低资源设备的应用上受到限制[2]。因此高效的模型轻量化方法研究受到了高度关注，其目的是使具有高性能的模型能够满足低资源设备的低功耗和实时性等要求，同时尽可能地不降低模型的性能。

近些年，先后出现了参数剪枝[3]、参数量化[4]、自动化搜索[5]等多种模型轻量化和模型压缩方法。参数剪枝通过去除冗余和不重要的参数来减小模型的大小。剪枝可以分为结构化剪枝和非结构化剪枝，前者是指按照一定规则剪掉整个卷积核或通道，后者是指任意剪掉单个参数[6,7]。参数量化方法将模型中的浮点数参数转换为较低精度的表示，从而减少模型的存储和计算需求[8,9]。自动化搜索方法使用启发式算法或强化学习来自动地搜索适合轻量化要求的网络结构和超参数[10]。这些方法在不同场景和任务中都有一定的适用性，为实现在资源受限设备上部署高效神经网络模型提供了多样化的解决方案[11,12]。

除了以上方法之外，知识蒸馏（Knowledge Distillation，KD）作为一种有前景的模型轻量化方法，近年来在深度学习领域引起了广泛的关注和研究，它与较早提出的并被广泛应用的一种机器学习方法的思想较为相似，即迁移学习[13]。知识蒸馏与迁移学习都涉及到知识的迁移，然而它们在数据域、网络结构、学习方式和目标任务方面均不同。知识蒸馏在 2015 被 Hinton[14]等人提出，其核心思想是从一个大型复杂的神经网络（教师网络）提取知识，并将其迁移到一个紧凑的模型（学生网络）上，以获得性能良好且更易于部署的轻量模型[15]。

在知识蒸馏的过程中，除了传统的硬标签（One-hot）来指导学生网络的训练外，还引入了教师网络在训练过程中的软标签（Soft Targets）来辅助学生网络的训练。通过最小化学生网络的输出与教师网络的输出之间的 KL 散度（Kullback-Leibler Divergence）损失，学生网络可以学习到教师网络中包含的更多细节信息，从而在不损失太多模型性能的情况下实现模型轻量化。除了原始的知识蒸馏方法外，研究者们还提出了一系列的改进和拓展技术。例如，FitNets 方法将知识蒸馏扩展到多个教师网络之间的知识传递[16]；Attention Transfer 通过对教师网络中的注意力机制进行蒸馏，提高了学生网络在注意力分布上的表现[17]；这些技术的引入进一步丰富了知识蒸馏的研究内容，提高了模型性能和泛化能力。

随着计算机视觉和自然语言处理[18]等领域的不断发展，知识蒸馏将在更广泛的应用场景中发挥更重要的作用。未来的研究方向可能包括更加灵活和高效的蒸馏策略，以及探索知识蒸馏在跨模态学习、多任务学习等更复杂场景中的应用，为模型的轻量化和泛化能力提升提供更多可能性。

## 二、文献综述（综述中引用的文献应按文中标注出现的顺序附后）

**对文献进行归纳总结、分类评价。**

深度学习由于对目标多样性变化具有很好的鲁棒性，近年来得到广泛的关注并取得快速的发展。然而，性能越好的深度学习模型往往需要更多的存储和计算资源，这使一些模型在物联网、移动嵌入式等低资源设备的应用上受到限制[19]。因此，研究人员开始对高效的深度学习模型展开研究，剪枝[20]、量化[21]、知识蒸馏[22]等模型轻量化方法被相继提出。 其中，知识蒸馏作为一种新兴的方法，目前已成为深度学习领域的一个研究热点和重点。在知识蒸馏网络中，这个大模型我们通常称之为 Teacher（教师模型），小模型称之为 Student（学生模型），来自教师模型输出的监督信息称之为 Knowledge（知识），而学生学习迁移来自教师的监督信息的过程称之为 Distillation（蒸馏），其核心思想如图 1-1 所示。

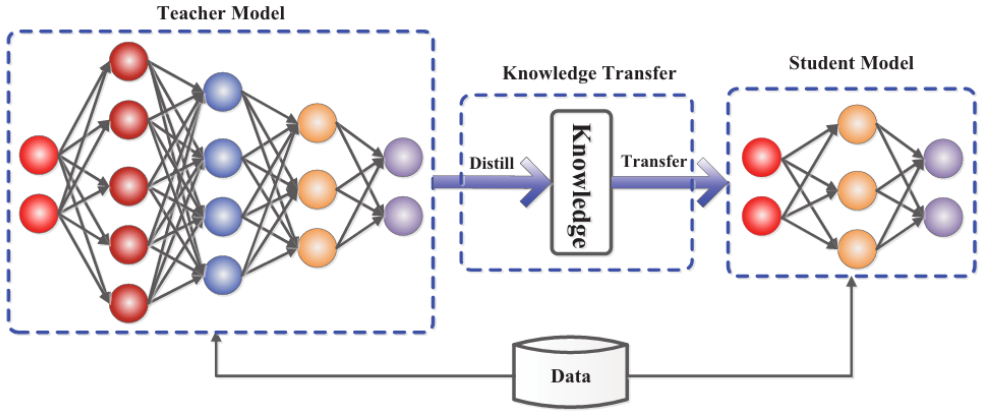



图 1-1 知识蒸馏

### 1 知识蒸馏中的知识类型

知识蒸馏利用性能更好的大模型的监督信息来辅助小模型的训练，在压缩模型大小的同时保持较好的模型性能。原始知识蒸馏(Vanilla Knowledge Distillation)仅仅是从教师模型输出的软目标中学习出轻量级的学生模型。然而当教师模型变得更深时，仅仅学习软目标是不够的。因此，我们不仅需要获取教师模型输出的知识，还需要学习隐含在教师模型中的其它知识。根据所蒸馏知识的不同，现有的方法又可以分为：（1）基于响应的知识蒸馏[23,24]（2）基于特征的知识蒸馏[25,26]（3）基于关系的知识蒸馏[27,28]。

### 1.1 基于响应的知识蒸馏

基于响应的知识蒸馏方法的核心思想是以教师模型的输出结果作为先验知识，结合样本的真实类别标签，共同指导学生网络的训练。这类方法最早是 2015 年由 Hinton[14]等人在 NIPS 大会上提出的，通过软化教师网络的输出来指导学生网络，并将学生网络的优化目标分为两部分：1）硬标签（Hard Target）：学生网络输出的类别

概率与样本真实类别标签（One-hot）之间的交叉熵；2）软标签（Soft Target）：学生与教师网络输出结果之间的 KL 散度损失，软目标为经过带温度参数 $T$ 的 softmax 结果，如公式（1-1）所示：

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{1-1}$$

其中，$T$为温度参数，$z_i$为神经网络的概率分布，$q_i$为输出的软目标。联合两个优化目标进行训练的学生网络能够模仿教师网络输出的概率分布，并获得与教师网络相近甚至更好的拟合能力。

该类方法训练框架如下图 1-2 所示，首先训练出一个性能较好的教师模型，有一个好的教师模型是知识蒸馏的前提；然后用教师模型的输出作为 Soft Target 来监督学生模型的训练。

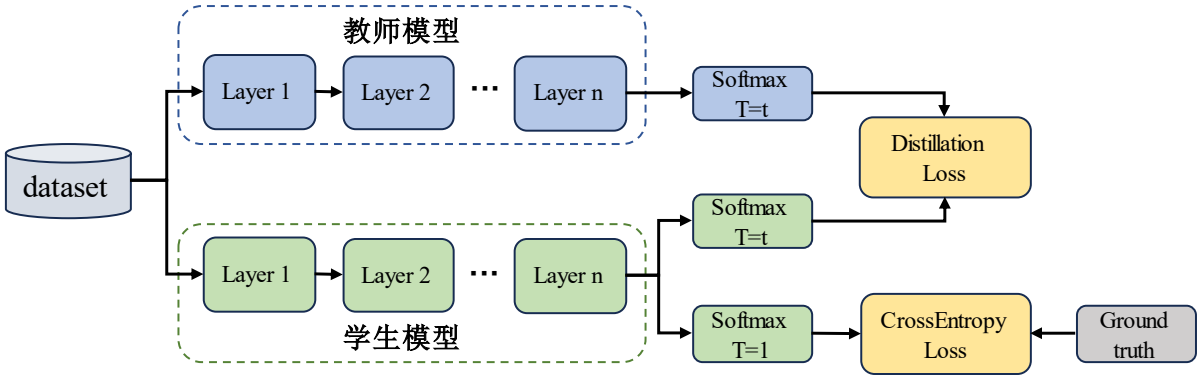



图 1-2 Hinton 提出的基于响应的蒸馏

Hinton 方法中损失函数表示如下：

$$L_{KD} = \alpha \text{CrossEntropy}(y_s, y) + (1-\alpha)\text{H}(y_s, y_t) \tag{1-2}$$

其中$\text{CrossEntropy}(*)$为交叉熵函数, $\text{H}(*)$ 为 KL 散度损失，$y_s$ 表示学生模型的预测结果，$y_t$表示教师模型的预测结果，$y$是硬标签。第一部分交叉熵函数是学生模型预测结果与硬标签之间的交叉熵损失，第二部分是学生模型与教师软标签之间的 KL 散度损失[29,30]。

响应知识通常指的是教师模型的最后一层输出，主要包括逻辑单元和软标签的知识。响应知识蒸馏的主要思想是促使学生能够学习到教师模型的最终输出，以达到和教师模型相近甚至更好的性能。原始知识蒸馏是针对分类任务来提出的，迁移的知识仅包含分类软标签知识，然而其它任务(如目标检测)网络最后一层输出中还可能包含有目标定位的信息，即不同任务教师模型的最后一层输出是不尽相同的。

**1.2 基于特征的知识蒸馏**

在深度学习中，一般将神经网络隐藏层的输出看作模型的特征。基于特征的知识蒸馏方法利用中间层的特征作为知识，指导学生网络进行训练[31]。Gotmare 等人[32]的研究表明：教师的软标签主要指导学生浅层网络的训练，而在学生网络的特征提取层的指导较少。换句话说，如果网络较深的话，单单学习教师的响应输出是不够的。复

杂教师和简单学生模型在中间的隐含层之间存在显著的容量差异，这导致它们不同的特征表达能力，教师的中间特征状态知识可以用于解决教师和学生模型在容量之间存在的“代沟”(Gap)问题[33]，其主要思想是从教师中间的网络层中提取特征来充当学生模型中间层输出的提示(Hint)。这一过程称为基于中间特征的知识蒸馏，如图 1-3 所示。一个使用教师模型中间特征知识的代表性方法是 FitNets[16]，其主要思想是促使学生的隐含层与教师隐含层的输出趋近于一致。中间特性蒸馏的损失可以被定义为：

$$L_{Feat}(f_t(x), f_s(x)) = \mathcal{L}_F\left(\Phi_t(f_t(x)), \Phi_s(f_s(x))\right) \quad (1\text{-}3)$$

其中，$f_t(x)$和$f_s(x)$分别是教师网络和学生网络的中间层特征，$\Phi_t(f_t(x))$和$\Phi_s(f_s(x))$将两个特征图转换为同一形状，$\mathcal{L}_F$通常为$\mathcal{L}_2$损失函数。

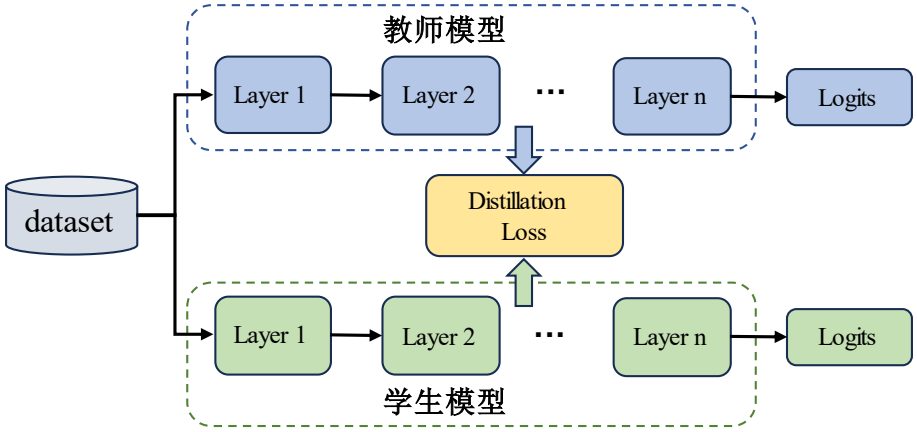



图 1-3 基于特征的知识蒸馏

基于中间特征的知识蒸馏将教师模型的特征提取能力迁移到学生模型中。在网络层的蒸馏节点上可以隔层、逐层或逐块地将教师的中间特征知识转移到学生模型中，或者仅蒸馏教师模型较高的隐含层和最后一个卷积层的特征知识。在网络层的迁移手段上，可以借助于构造的网络块使学生模型通过模仿学习来获得教师模型的中间层特征，如可学习的投影矩阵和定义的注意力映射图[17]。不同于将教师模型的中间特征知识迁移或投射到学生模型，一些工作通过共享网络的网络层直接利用教师的中间特征，基于中间特征的知识蒸馏方法实质上是要最小化教师与学生之间中间特征的映射距离[34]。

### 1.3 基于关系的知识蒸馏

基于响应和特征的知识蒸馏方法都使用教师模型中特定层的输出，而基于关系的知识蒸馏进一步探索了不同层或者数据样本之间的关系，其思想如图 1-4 所示。Yim 等人[35]通过计算表示方向的内积生成一个 FSP 矩阵，来刻画层与层之间的特征关系，然后最小化教师网络 FSP 矩阵和学生网络 FSP 矩阵之间的 L2 损失，以达到知识迁移的目的。FSP 矩阵是测量网络间的关系特征，而后续工作更强调样本的关系知识。例如，Park 等人[36]提出了基于样本的角度关系和距离关系蒸馏。三个样本角度关系可表示如下：

$$\begin{cases} \psi_A(t_i,t_j,t_k)=\cos\angle t_i t_j t_k=\left\langle e^{ij},e^{kj}\right\rangle \\ e^{ij}=\dfrac{t_i-t_j}{\left\|t_i-t_j\right\|_2},e^{kj}=\dfrac{t_k-t_j}{\left\|t_k-t_j\right\|_2} \end{cases} \tag{1-4}$$

其中，$t_i,t_j,t_k$ 表示不同样本，余弦函数被用来衡量样本的距离关系，定义为：

$$\begin{cases} \psi_A(t_i,t_j)=\dfrac{1}{\mu}\left\|t_i-t_j\right\|_2 \\ \mu=\dfrac{1}{\left|\chi^2\right|}\sum\left\|t_i-t_j\right\|_2 \end{cases} \tag{1-5}$$

其中，$\chi^2$ 表示一对具有区分性样本的集合，$\mu$ 是距离的归一化因子，其值为一个训练批次中样本对的平均距离。

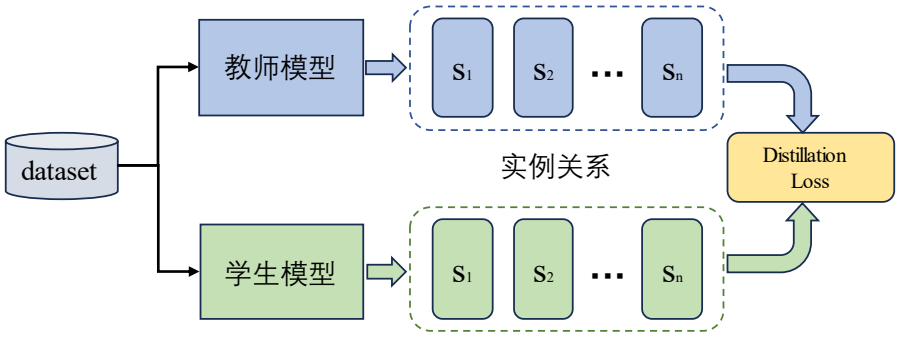



图 1-4 基于关系的知识蒸馏

Lee 等人[37]利用特征图之间的相关性提取知识，通过引入奇异值分解（SVD）来有效地去除特征映射中的空间冗余，并在降低特征维数的过程中获得有意义的隐含特征信息。Tung 等人[38]提出了一个新的蒸馏损失——相似性保留损失，即一对输入送到教师网络中产生的特征向量相似，那么送到学生网络中产生的特征向量也应该相似，反之不相似的话同样在学生网络中也不应该相似。将教师模型计算出的样本对之间的相似性信息作为知识迁移给学生网络，指导学生网络的训练。针对多个输入样本，教师网络和学生网络分别得到各自的样本相似性矩阵，通过两个相似性矩阵的均方误差最小化，达到知识蒸馏的目的[39]。除样本间相似性知识之外，关系蒸馏还可以利用相互关系知识[40]和相关性知识[41]。基于样本的关系知识蒸馏不仅传递了单个样本的信息，而且传输多个样本间的关系知识。另外样本的关系知识还能借助于辅助技术，如通过图描述数据内部关系来实现样本关系的知识迁移[42]。

**2 知识蒸馏中的蒸馏框架**

在知识蒸馏中，蒸馏方案也是一个重要的部分。根据教师模型的数量，知识蒸馏方案可以分为三类：单教师蒸馏[43]、多教师蒸馏[44]和自蒸馏[45]。没有教师模型参与的蒸馏方式被称为自蒸馏[46,47]；单教师蒸馏中只有一个教师模型来监督学生模型的训练，是最常用的知识蒸馏方法；还有一种方案是由多个教师模型参与的多教师蒸馏[48,49]，包括多个学生网络相互学习的方式。

## 2.1 单教师蒸馏

单教师蒸馏是知识蒸馏中最常见的一种方式之一。它使用性能良好的大规模神经网络作为教师模型，然后将其知识迁移至一个小规模的学生模型。在训练过程中，学生模型通常在教师模型的监督下学习。例如，在分类任务中，学生模型不仅简单地学习数据集中的 Ground Truth 标签，还会学习教师模型的预测概率或中间特征[50,51]。通过这种方式，学生模型可以从教师模型的‘经验’中受益，从而提高性能并在更小的模型尺寸下进行有效的推理。

单教师蒸馏主要分为两种情况。一种被称为“离线蒸馏”，它使用已经完成预训练的教师模型来监督学生模型的训练。已经完成预训练的教师模型可以提供高质量的软标签，以提高学生模型的学习质量。该训练框架如图 2-1 所示。另一种情况是教师模型和学生模型同时进行训练，以使学生模型的输出结果与教师模型趋于一致，称为“在线蒸馏”[52]。该框架如图 2-2 所示。由于教师模型具有更强的学习能力，它可以输出更好的预测分布。

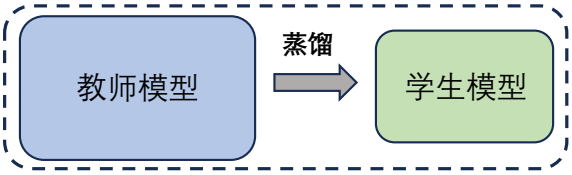



图 2-1 离线蒸馏

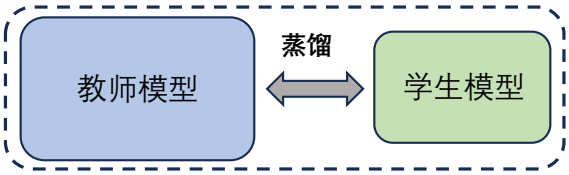



图 2-2 在线蒸馏

值得注意的是，越大或越深的神经网络不一定总是更适合作为教师模型。这就好像学生模型的学习能力是有限的，当教师模型和学生模型之间的‘gap’太大时，会恶化蒸馏效果。此外，单教师蒸馏对教师模型有很强的依赖性，教师模型的性能直接影响学生模型的学习效果。

为了减轻教师和学生之间 gap 带来的消极影响，Mirzadeh 等人[53]提出了一种教师助教知识蒸馏方法（Teacher Assistant Knowledge Distillation，TAKD），来弥补教师网络和学生网络之间的差距。Ni 等人[54]提出了一种增强的知识蒸馏方法来取代传统的蒸馏方法，即两阶段知识蒸馏。Yan 等人[55]提出了一种递归知识蒸馏的方法，来减小教师网络和学生网络之间的差距，进一步提高轻量化模型的泛化能力。

## 2.2 多教师蒸馏

在多教师蒸馏过程中，多个神经网络被用作教师模型来监督学生模型的训练。在训练过程中，学生网络可以根据每个教师网络输出的质量级别进行选择性学习，或者将多个教师模型的输出集合作为软标签用于监督学习，从而解决了学生模型对单个教师模型的依赖性。

类似于单教师蒸馏，多教师蒸馏也可以分为两种情况。一种是对教师模型进行预训练，然后使用已经完成预训练的多个教师模型来监督学生模型的训练。例如，Dvornik 等人[56]提出了将多个教师网络的预测结果进行整合和平均，并通过比较学生

网络的预测概率与多个教师网络的平均预测概率的 KL 散度来将它们传递给学生网络，被称为“集成蒸馏”。该框架如 2-3 所示。另一种情况是多个模型同时进行训练，并相互学习[57]，将它们的输出集成为一种常见的软标签进行学习。这样每个网络模型都可以从多个教师模型中学习，从而有效地提高学习效果，被称为“协作蒸馏”，该框架如图 2-4 所示。

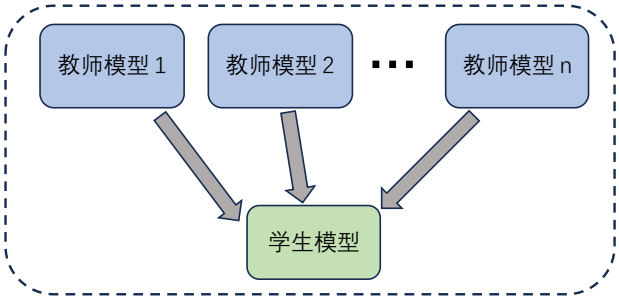



图 2-3 集成蒸馏

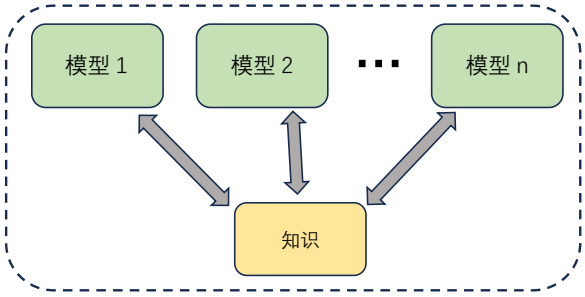



图 2-4 协作蒸馏

与单教师蒸馏相比，多教师蒸馏需要完成多个教师模型的预训练，准备过程更为复杂。此外，在训练学生模型时，多个教师网络的存在将占用大量存储空间和计算资源，需要更长的训练时间，并对硬件设备有更高的要求，而且还需要考虑选择多个教师模型或权重分配的问题[58,59]。Zhu 等人[60]提出了一种更有效的方法来对不同教师网络的知识进行加权，并设计了一个门控网络来获取不同教师网络的权重，显示出比基于单个教师的网络更为出色的结果。Shen 等人[61]集成多个教师网络的预测结果对其求平均，然后迁移到学生网络，通过 KL 散度度量学生网络的预测概率与多个教师网络的平均预测概率之间的相似性。

### 2.3 自蒸馏

自蒸馏方案是仅通过神经网络模型自身进行知识的更新。事实上，自蒸馏可以被看作是使用网络的一部分作为教师模型来指导另一部分[62,63]，将深层的知识蒸馏给浅层。或者使用前一周期的输出来指导后一周期[64]。


图 2-5 自蒸馏

对比单教师蒸馏和多教师蒸馏，自蒸馏不涉及其他神经网络，也不需要预训练教师模型，因此准备工作简单得多。此外，在训练过程中只涉及一个神经网络，所以参数数量和计算资源占用较少，训练速度也大大提高。在硬件设备资源有限的情况下，自蒸馏是一个很好的选择。最重要的是，自蒸馏方法无需考虑教师模型和学生模型之间的差距是否合适。

自蒸馏方法在图像分类任务和自然语言处理任务中得到了广泛应用。在图像分类

任务中，通过自蒸馏可以使浅层网络学习到更多知识，并在减少参数和计算资源的情况下提高模型性能。在自然语言处理任务中，自蒸馏可以提高语言模型的性能，通过前一时期的预测结果指导后一时期的训练。此外，研究人员还探索了将自蒸馏方法与其他知识蒸馏方法相结合的可能性，以进一步提高模型性能并减少训练过程中的存储和计算开销。然而，自蒸馏也有其局限性。由于缺乏其他神经网络和自身较弱的学习能力，无法保证优质的软标签。缺少外部知识将会恶化蒸馏效果。为了克服自蒸馏的一些限制，研究人员提出了各种改进的自蒸馏技术，例如引入注意力机制来增强自蒸馏效果，或者设计更复杂的网络结构来提高自蒸馏的性能。

**3 知识蒸馏与其他算法的融合**

在知识蒸馏的探索过程中，除了知识类型和蒸馏方案，还整合了其他深度学习技术和知识蒸馏方法，并取得了非常好的结果。许多深度学习技术在知识蒸馏中发挥着重要作用，例如生成对抗网络（GANs）、对比学习、图神经网络、注意力机制等。

生成对抗网络（GANs）：已将 GANs 纳入知识蒸馏中，以生成真实样本，提高学生模型的鲁棒性，基于 GAN 的蒸馏有助于学习更多样化的表示并减少过拟合[65]。

对比学习：将对比学习技术，例如对比预测编码和 InfoNCE 损失，应用于知识蒸馏[66]。这些方法通过比较正样本和负样本，鼓励学生模型学习更具信息量和可区分性的表示。

图神经网络（GNNs）：知识蒸馏技术已扩展到基于图的数据，并使用图神经网络进行处理， 图知识蒸馏旨在将图相关的知识，如节点嵌入或图结构，从教师模型传递给学生模型[67]。

注意力机制：将注意力机制，例如自注意力或基于 Transformer 的注意力，融入知识蒸馏中，以捕捉数据中的重要信息和关系。基于注意力的蒸馏帮助学生模型在训练过程中关注重要元素[68]。

联邦学习：将联邦学习，一种分布式学习方法，与知识蒸馏相结合，实现多设备或客户端之间的协作训练[69]。这样可以在不集中存储数据的情况下实现知识共享。

除了以上方法外，还有强化学习、跨模态迁移等深度学习技术在知识蒸馏领域也有相关的应用。这些技术融合方法旨在充分发挥各种深度学习技术的优势，大大提升了知识蒸馏过程中的知识迁移效果，使知识蒸馏能够更好的满足各个领域不同任务的模型轻量化需要。

**综上所述，本课题将致力于在图像分类和目标检测任务场景中探索知识蒸馏技术，针对以往蒸馏方法中教师多尺度特征知识利用不充分、学生模型训练缺少先验知识引导和模型训练过程中硬件设备负担较重的问题，提出有效方案以实现更加快速、有效地压缩和简化深度神经网络模型，为图像分类和目标检测等技术在实际应用中的推广和发展提供新的可能性。**

**参考文献**


[1] Gou J, Yu B, Maybank S, et al. Knowledge Distillation: A Survey[J]. International Journal of Computer Vision, 2020, 129: 1789 - 1819.

[2] 贺才郡,李开成,董宇飞,等.基于知识蒸馏与 RP-MobileNetV3 的电能质量复合扰动识别[J].电力系统保护与控制, 2023, 51(14): 75-84.

[3] Lin M, Ji R, Wang Y, et al. HRank: Filter Pruning Using High-Rank Feature Map[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 1526-1535.

[4] Yamamoto K. Learnable Companding Quantization for Accurate Low-bit Neural Networks[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 5027-5036.

[5] Liu X, Zhao J, Li J, et al. Federated Neural Architecture Search for Medical Data Security[J]. IEEE Transactions on Industrial Informatics, 2022, 18: 5628-5636.

[6] Zhang C, Ma Y, Wu J, et al. HCov: A Target Attention-based Filter Pruning with Retaining High-Covariance Feature Map[C]//International Joint Conference on Neural Networks (IJCNN). 2021: 1-8.

[7] Chen Y, Shuai M, Lou S, et al. FPAR: Filter Pruning Via Attention and Rank Enhancement[C]//IEEE International Conference on Multimedia and Expo (ICME). 2022: 1-6.

[8] Long X, Zeng X, Liu Y, et al. Low Bit Neural Networks with Channel Sparsity and Sharing[C]//International Conference on Image, Vision and Computing (ICIVC). 2022: 889-894.

[9] Liu D, Chen X, Ma C, et al. Hyperspherical Quantization: Toward Smaller and More Accurate Models[C]//IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2023: 5251-5261.

[10] Das P, Singh M, Roy D, et al. A Secure Softwarized Blockchain-based Federated Health Alliance for Next Generation IoT Networks[J]. IEEE Globecom Workshops (GC Wkshps), 2021, 12: 1-6.

[11] 秦荣荣,高晓蓉,罗林,等.基于注意力反向知识蒸馏的车轮踏面异常检测[J].激光与光电子学进展, 2023, 60(24): 44-52.

[12] 苗德邻,刘磊,莫涌超,等.基于知识蒸馏的夜间低照度图像增强及目标检测[J].应用光学, 2023, 44(5):1037-1044.

[13] Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359

[14] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network[J]. Computer Science, 2015, 14(7):38-39.

[15] Wang L, Yoon K. Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks[J]. IEEE Transactions on Pattern Analysis and

Machine Intelligence, 2020, 44: 3048-3068.
[16] Romero A, Ballas N, Kahou S E, et al. FitNets: Hints for Thin Deep Nets[C]//International Conference on Learning Representations(ICLR). 2015: 36-42.
[17] Zagoruyko S, Komodakis N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer[C]//International Conference on Learning Representations(ICLR). 2017: 24-26.
[18] Chen Y, Liu Y, Dong L, et al. AdaPrompt: Adaptive Model Training for Prompt-based NLP[C]//Conference on Empirical Methods in Natural Language Processing. 2022:34-41.
[19] 黄震华,杨顺志,林威,等.知识蒸馏研究综述[J].计算机学报, 2022(003):045.
[20] Chen Y, Wen X, Zhang Y, et al. FPC: Filter pruning via the contribution of output feature map for deep convolutional neural networks acceleration[J]. Knowledge-based systems, 2022, 28:238.
[21] Dai S, Venkatesan R, Ren H, et al. VS-Quant: Per-vector Scaled Quantization for Accurate Low-Precision Neural Network Inference[J]. 2021, arXiv:2102.04503.
[22] Wang L, Yoon K. Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44: 3048-3068.
[23] Zhao B, Cui Q, Song R, et al. Decoupled Knowledge Distillation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 11943-11952.
[24] Li X, Fan D, Yang F, et al. Probabilistic Model Distillation for Semantic Correspondence[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 7501-7510.
[25] Song J, Chen Y, Ye J, et al. Spot-Adaptive Knowledge Distillation[J]. IEEE Transactions on Image Processing, 2022, 31: 3359-3370.
[26] Chen P, Liu S, Zhao H, et al. Distilling Knowledge via Knowledge Review[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 5006-5015.
[27] Wang X, Wang Z, Hu W. Serial Contrastive Knowledge Distillation for Continual Few-shot Relation Extraction[J]. ArXiv: 2305.06616, 2023.
[28] Zhu J, Tang S, Chen D, et al. Complementary Relation Contrastive Distillation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 9256-9265.
[29] Qiu Z, Ma X, Yang K, et al. Better Teacher Better Student: Dynamic Prior Knowledge for Knowledge Distillation[J]. ArXiv: 2206. 06067, 2023.
[30] 韩宇. 深度神经网络知识蒸馏综述[J]. 计算机科学与应用, 2020, 10（9）:6-14.
[31] Wang M, Liu R, Abe N, et al. Discover the effective strategy for face recognition model compression by improved knowledge distillation[C]//2018 25th IEEE International

Conference on Image Processing (ICIP). 2018: 2416-2420.
[32] Romero A, Ballas N, Kahou S E, et al. FITNETS: HINTS FOR THIN DEEP NETS[J]. arXiv preprint arXiv: 1412.6550, 2014.
[33] Mirzadeh S, Farajtabar M, Li A, et al. Improved Knowledge Distillation via Teacher Assistant[C]//AAAI Conference on Artificial Intelligence. 2019: 89-97.
[34] Passalis N, Tefas A. Learning deep representations with probabilistic knowledge transfer[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 268-284
[35] Yim J, Joo D, Bae J, et al. A gift from knowledge distillation: fast optimization, network minimization and transfer learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4133-4141.
[36] Park W, Kim D, Lu Y, et al. Relational knowledge distillation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 3967-3976.
[37] Lee S H, Kim D H, Song B C. Self-supervised knowledge distillation using singular value decomposition[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 335-350.
[38] Tung F, Mori G. Similarity-preserving knowledge distillation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1365-1374.
[39] 王新生,朱小飞,黄贤英.双通道知识蒸馏的节点分类方法[J].小型微型计算机系统, 2023, 44(10): 2284-2290.
[40] Ni Z, Yang F, Wen S, et al. Dual Relation Knowledge Distillation for Object Detection [J]. ArXiv: 2302.05637, 2023.
[41] Peng B, Jin X, Liu J, et al. Correlation congruence for knowledge distillation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 5006-5015.
[42] Lee S, Song B C. Graph-based knowledge distillation by multi-head attention network [C]//Proceedings of the 30th British Machine Vision Conference. 2019: 141.
[43] Johnson R, Zhang T. Guided Learning of Nonconvex Models through Successive Functional Gradient Optimization[C]//International Conference on Machine Learning, 2020: 572-585.
[44] 杜潇鉴,吕卫东,孙钰华.基于多教师知识蒸馏的新闻文本分类方法[J].计算机科学与应用, 2023, 13(8):11.
[45] 陈建炜,杨帆,赖永炫.一种基于信息熵迁移的文本检测模型自蒸馏方法[J].自动化学报, 2023, 49(11):1-12.
[46] Ge Y, Choi C L, Zhang X, et al. Self-distillation with Batch Knowledge Ensembling Improves ImageNet Classification[J]. ArXiv: 2104. 13298, 2021.
[47] 马金林,刘宇灏,马自萍,等.解耦同类自知识蒸馏的轻量化唇语识别方法[J].北京航空航天大学学报, 2023, 48: 476-483.

[48] Yuan F, Shou L, Pei J, et al. Reinforced Multi-Teacher Selection for Knowledge Distillation[C]//AAAI Conference on Artificial Intelligence. 2020: 34-46.
[49] Liu Y, Zhang W, Wang J, et al. Adaptive multi-teacher multi-level knowledge distillation [J]. Neurocomputing, 2020, 415: 106-113.
[50] Guo Z, Yan H, Li H, et al. Class Attention Transfer Based Knowledge Distillation [C]//In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 11868-11877.
[51] Zhang Y, Li S, Yang X S. Knowledge Distillation with Active Exploration and Self-Attention Based Inter-Class Variation Transfer for Image Segmentation[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023: 5076-5088.
[52] Li S, Lin M, Wang Y, et al. Distilling a Powerful Student Model via Online Knowledge Distillation[J]. IEEE transactions on neural networks and learning systems, 2021, 16: 372-379.
[53] Mirzadeh S I, Farajtabar M, Li A, et al. Improved knowledge distillation via teacher assistant[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 5191-5198.
[54] Ni H, Shen J, Yuan C. Enhanced Knowledge Distillation for Face Recognition[C]//2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking （ISPA/BDCloud/SocialCom/SustainCom. IEEE, 2019: 1441-1444.
[55] Yan M, Zhao M, Xu Z, et al. VarGFaceNet: An Efficient Variable Group Convolutional Neural Network for Lightweight Face Recognition[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop （ICCVW）. IEEE, 2019: 2647-2654.
[56] Dvornik N, Schmid C, Mairal J. Diversity With Cooperation: Ensemble Methods for Few-Shot Classification[C]//IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 3722-3730.
[57] Zou P, Teng Y, Niu T. Multi scale Feature Extraction and Fusion for Online Knowledge Distillation[C]//International Conference on Artificial Neural Networks. 2022: 698-706.
[58] 陈诗琪,王威,占荣辉,等.特征图知识蒸馏引导的轻量化任意方向 SAR 舰船目标检测器[J].雷达学报, 2023, 12(1):14.
[59] 张翼,朱永利.结合知识蒸馏和图神经网络的局部放电增量识别方法[J].电工技术学报, 2023, 38(5):11.
[60] Zhu J, Liu J, Li W, et al. Ensembled CTR Prediction via Knowledge Distillation [C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 377-485.
[61] Shen Z, Savvides M. MEAL V2: Boosting Vanilla ResNet-50 to 80%+ Top-1 Accuracy on ImageNet without Tricks[J]. arXiv preprint arXiv:2009.08453, 2020.

[62] Zhang L, Song J, Gao A, et al. Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation[C]//IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 3712-3721.

[63] Zhang L, Bao C, Ma K. Self-Distillation: Towards Efficient and Compact Neural Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44: 4388-4403.

[64] Shen Y, Xu L, Yang Y, et al. Self-Distillation from the Last Mini-Batch for Consistency Regularization[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 11933-11942.

[65] Huang B, Chen M, Wang Y, et al. Boosting Accuracy and Robustness of Student Models via Adaptive Adversarial Distillation[C]//In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 24668-24677.

[66] Xu Q, Chen Z, Ragab M, et al. Contrastive adversarial knowledge distillation for deep model compression in time-series regression tasks[J]. Neurocomputing, 2021, 485: 242-251.

[67] Ghosh P, Saini N, Davis L S, et al. Learning Graphs for Knowledge Transfer with Limited Labels[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 11146-11156.

[68] Diao C, Loynd R. Relational Attention: Generalizing Transformers for Graph-Structured Tasks[J]. ArXiv: 2210.05062, 2022.

[69] Sui D, Chen Y, Zhao J, et al. FedED: Federated Learning via Ensemble Distillation for Medical Relation Extraction[C]//Conference on Empirical Methods in Natural Language Processing. 2020: 564-577.

## 三、研究内容和方法

### 1.研究的基本内容

本课题旨在利用知识蒸馏的基本方法，研究基于多级特征传递和教师先验引导的理论、方法。在知识蒸馏网络中，根据知识的不同可以分为基于响应的知识蒸馏、基于特征的知识蒸馏和基于关系的知识蒸馏，我们通过不同的途径来充分挖掘和利用教师模型中的知识。首先，我们提出多级特性渐进式传递蒸馏，将学生模型的中间特征从后向和前向两条路径进行渐进式融合传递，保证学生特征能够充分受到教师模型的特征知识监督；其次，除了对学生模型特征的监督外，利用教师模型对学生特征进行先验引导，将教师的特性知识作为先验引导与学生模型的特性知识在通道上和空间上进行融合，把融合后的特征作为学生的特征知识，令其与教师特性趋向于一致，很好地保证了教师在学生整个学习过程中的参与，充分发挥教师在学生训练过程中的指导作用；最后，我们利用自蒸馏思想，将模型的所有中间层特征信息进行提取，作为监督来指导模型自身的中间层特征，而在输出层将纠正后的模型预测结果作为监督信息，大大提升模型训练速度并节省了硬件资源。图 3-1 为本课题研究内容的基本框架：

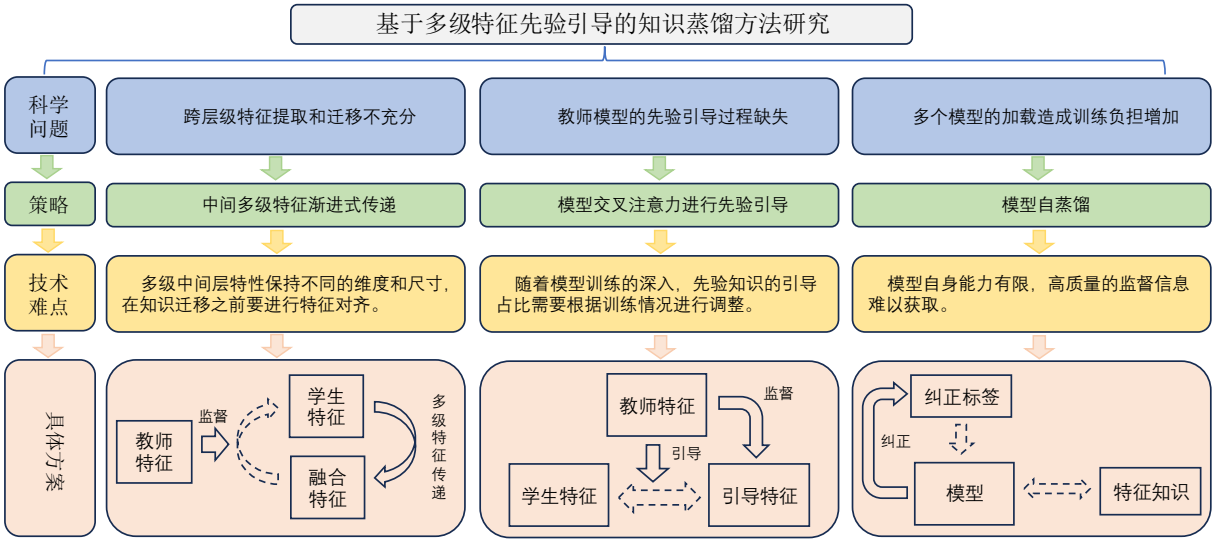



图 3-1 研究内容基本框架

**（1）基于多级特征渐进式传递的知识蒸馏方法研究**

基于特征的知识蒸馏理论研究是实现本部分的研究基础，高质量的特征能够同时满足多种下游任务。因此，如何将教师模型的特征知识充分地蒸馏至学生模型是本部分的一项重要研究内容。

考虑到现有知识蒸馏方法受多级中间层特性在维度和尺寸上不统一的限制，只有对应层的中间特性被蒸馏，跨层级特征很少被考虑，导致特性知识难以被充分地从教师模型迁移至学生模型。本课题拟利用以这个双向渐进式多级传递过程对学生特征进行逐级的递进融合，保证融合后的特征包含来自每个中间层特征的信息，从而实现特征知识高效且充分的迁移。

**（2）基于模型交叉注意力先验引导的知识蒸馏识别方法研究**

注意力机制模仿人类大脑的信息处理方式，允许模型根据输入数据的不同部分自动分配不同的关注度，从而更好地捕捉数据之间的关联性。因此，如何合理利用教师模型的注意力引导学生模型训练是本部分的重要研究内容。

考虑到大多数知识蒸馏方法在学生模型的训练过程中，教师模型只被作为监督信号来辅助学生模型的训练，忽视了它的先验引导作用，没有充分挖掘教师模型的使用价值。本课题拟提出模型交叉注意力作为先验引导，通过动态地计算和调整权重，学生模型可以模仿教师模型在输入数据上的关注模式，学会在不同部分的输入数据上分配注意力，而不是仅仅将教师特征作为最终的监督信息，从而更好地捕捉和迁移教师模型的知识。

**（3）基于多级特征传递和预测纠正的自蒸馏方法研究**

在知识蒸馏中知识的选择和蒸馏过程对学生网络的性能有着直接的影响，而其他蒸馏方式相比自蒸馏方法中高质量知识的引导相对缺失。因此，如何选择蒸馏的知识类型和高质量知识获取是本部分的重难点内容。

考虑到以往方法训练过程中，学生模型和一个甚至多个教师模型的同时加载需要更多的存储和计算资源，造成较重的模型训练负担并大大延长了模型训练所需的时间。本课题拟提出基于多级特征传递和预测纠正的自蒸馏方法，通过多级特征渐进式传递的方式得到包含模型的所有中间层特征信息的高质量特征，然后用其来辅助模型自身的训练；此外，将纠正后的模型预测作为软标签与 Ground Truth 共同监督模型的训练。

**2.拟采取的技术路线、研究手段和研究方法**

在知识蒸馏网络中，对于一个训练良好的教师网络，其输出不仅包含了输入图像对应的标签，还包含了除标签以外其他类别的信息。学生网络一般选用参数量少的轻量级网络，教师网络的输出由于附加类别信息的存在使得学生网络可以更容易地学习到更多的特征。除了从教师模型学习更丰富的信息之外，学模型向自身的内部探索也同样重要，以该思想为基础的自蒸馏方法不仅省去教师模型预训练时间，而且对训练设备的要求也更低。根据以上研究基础，本课题提出基于多级特征传递和教师先验引导的知识蒸馏方法研究，拟采取的技术路线具体为：

**（1）基于多级特征渐进式传递的知识蒸馏方法研究**

在基于特性的知识蒸馏过程中，大多数方法仅仅迁移同层级的特性知识，它们丢失了不同层级特性中包含的重要信息。回顾机制的提出一定程度上解决了这个问题，它将不同层的中间特征通过回顾的方式融合起来，实现了跨层特征的迁移并取得一定效果。但回顾机制仅仅使用了教师的浅层特征去指导学的深层特征，而教师深层特征中也包含着丰富的知识，如何把这部分信息有效利用起来成为一个重要的问题。

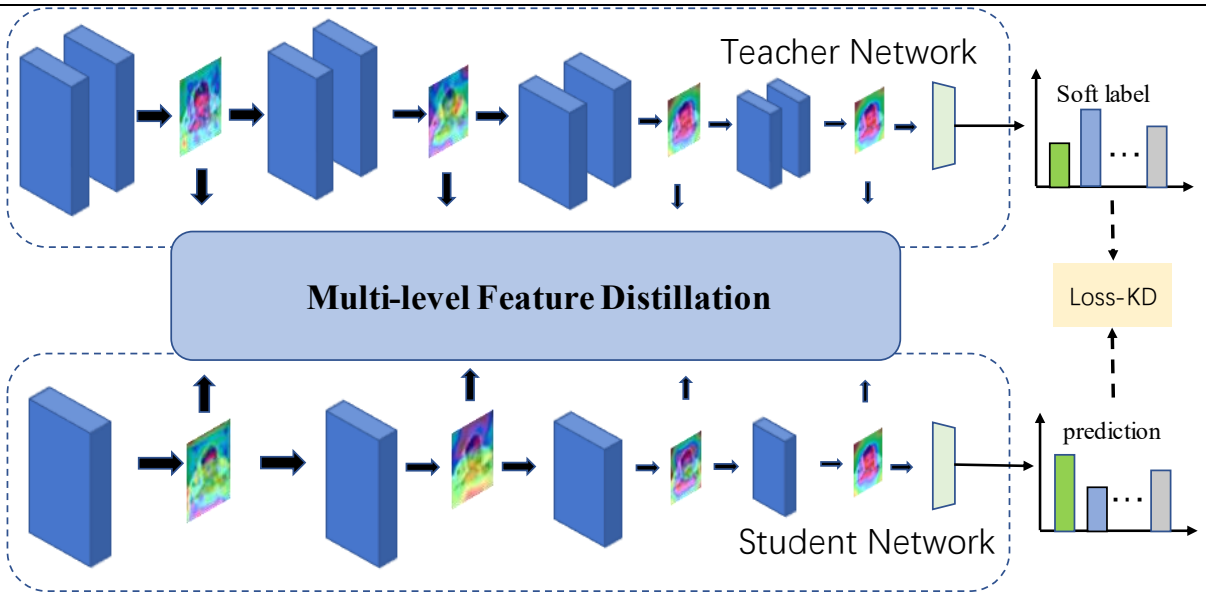
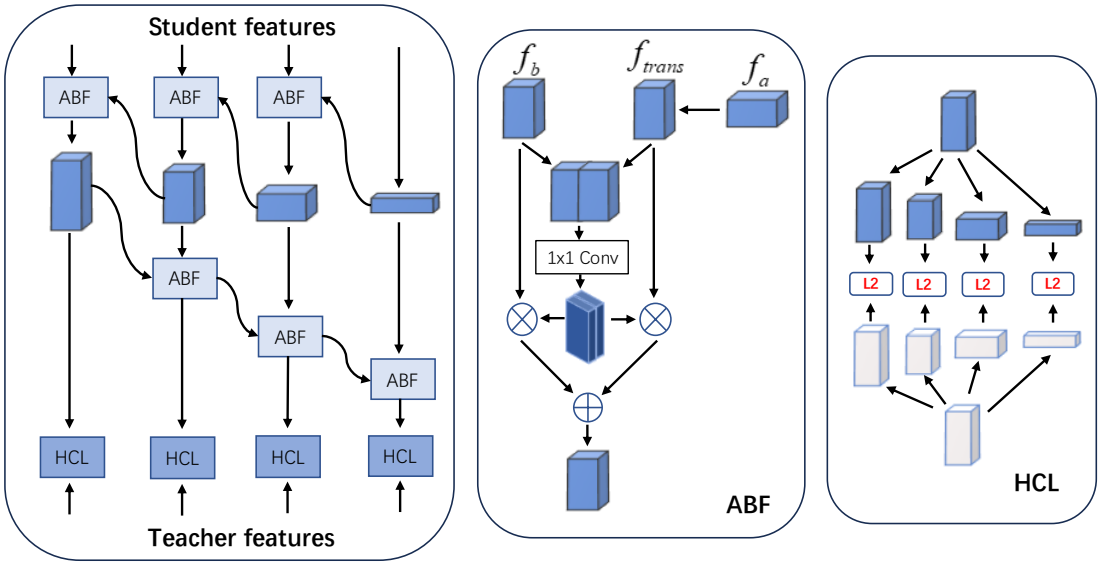



图 3-1 基于多级特征渐进式传递的知识蒸馏

因此本课题拟采用多级特征渐进式传递的方式进行知识蒸馏。首先，将学生深层特性向浅层特性进行反向融合得到一阶融合特征；然后将得到的融合特征再进行由浅层到深层的前向融合得到二阶融合特征。学生模型的中间特性经过两次渐进式融合后，最后的二阶融合特征包含来自每层中间特性的信息，有效提升学生模型的特性学习效率。网络架构如图 3-1 所示，其中多级特征蒸馏（Multi-level Feature Distillation）包含三部分：基于注意力的特征融合模块、多级特征渐进式融合、多级上下文损失。




(a) Multi-level feature progressive transfer (b) Attention Based Fusion (c) Hierarchical Context Loss

图 3-2 Multi-level Feature Distillation

**1）基于注意力的特征融合模块（Attention Based Fusion，ABF）**。首先，我们通过卷积和插值使来自不同层的特征能够对齐，这个过程可以表示为：

$$F_{trans} = G(F_a) \tag{3-1}$$

我们用$G(\bullet)$代表卷积和插值操作，$F_a$为来自上层或下层的特性表征。经过变形后的特性表征$F_a$与本层的特性表征$F_b$在通道和尺寸上保持一致。然后，一个 1×1 的卷积被使用来获取特征$F_a$和$F_b$分别对应的特征注意力图，可表示如下：

$$F_{attn} = Conv(F) \tag{3-2}$$

其中，$F_{attn}$表示注意力图，$Conv(\bullet)$表示 1×1 的卷积操作。

最后将来自不同层级的特征和他们对应的注意力图进行点乘后相加，得到融合特征。这个融合模块可表示如下：

$$F_{\text{ABF}}(F_b, F_a) = Conv(G(F_a)) * G(F_a) + Conv(F_b) * F_b \tag{3-3}$$

**2）多级特征渐进式融合（Multi-level Feature Progressive Transfer，MFPT）**。对于学生模型的中间特征，我们通过 ABF 模块进行多级特征融合。为了方便，我们令$f_i \circ f_{i+1} = F_{\text{ABF}}(f_i, f_{i+1})$，则反向渐进式融合可表示如下：

$$\dot{f}_i = f_i \circ f_{i+1} \cdots \circ f_n \tag{3-4}$$

其中，$\dot{f}_i$表示第 i 层一阶融合特征。通过等式（3-4），我们可以得到学生模型每个中间层的一阶融合特征$\dot{F}_S$，它包含了来自比当前第 i 层更深的特征信息。

类似的，我们可令$f_i \bullet f_{i+1} = F_{\text{ABF}}(f_{i+1}, f_i)$，则前向渐进式融合可表示为：

$$\ddot{f}_i = \dot{f}_1 \bullet \dot{f}_2 \cdots \bullet \dot{f}_i \tag{3-5}$$

其中，$\ddot{f}_i$表示第 i 层二阶融合特征，它包含学生模型所有中间层特征的信息。

**3）多级上下文损失（Hierarchical Context Loss，HCL）**。ABF 模块将学生模型在不同尺度和维度上的特征进行融合。然而，直接进行特征蒸馏将导致特征信息提取不充分。这个问题由 HCL 模块有效地解决，该模块利用金字塔池化来处理教师和学生模型的特征，这个过程可表示如下：

$$F^P = P(F) \tag{3-6}$$

其中，$F^P$为经过金字塔池化处理的特征。此时可得到多个尺度上的教师和学生模型特征，并用$L_2$距离来计算损失函数，可表示如下：

$$L_{RP} = D(F_S^P, F_T^P) \tag{3-7}$$

其中，$D(\bullet)$为$L_2$损失，$F_S^P$和$F_T^P$分别为经过金字塔池化后的学生和教师特征。

我们拟提出的多级特征渐进式传递蒸馏借鉴了残差网络的堆叠思想，确保经过多轮融合后，每一级输出特征包含了来自学生所有中间层的特征信息，弥补了以往方法中跨层教师特征利用不充分的问题，极大地提升了特征知识蒸馏的效果。

**（2）基于模型交叉注意力先验引导的知识蒸馏识别方法研究**

在现有知识蒸馏方法中，教师知识在学生模型训练过程中的监督作用一直被作为唯一的关注重点，纯粹地将教师的知识作为目标进行学习，只关注了输出上的知识学习，而忽视了教师知识在学生模型学习过程中的引导作用。

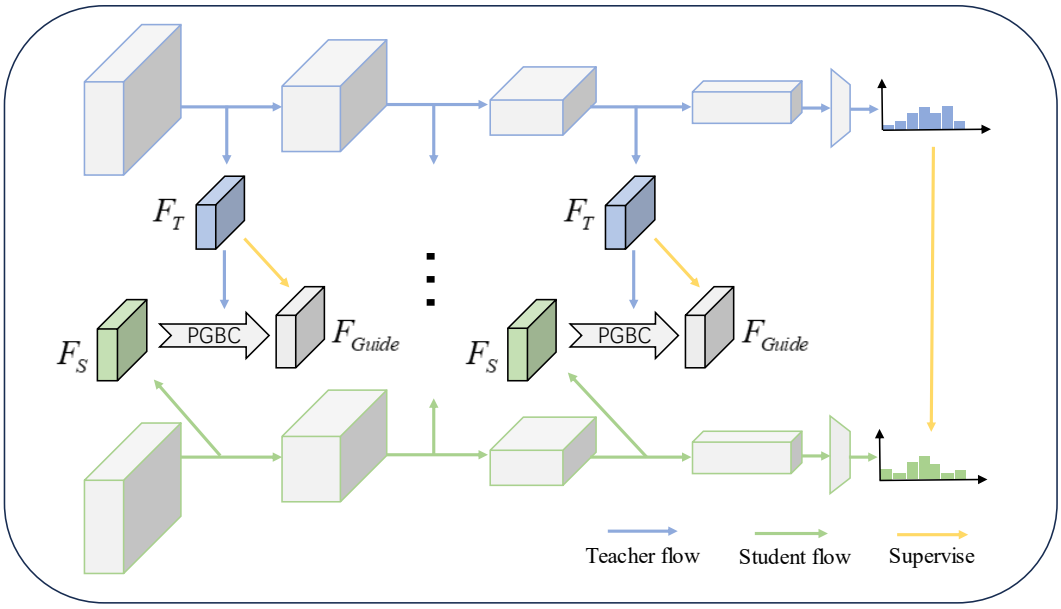



图 3-3 基于模型交叉注意力先验引导的知识蒸馏

针对这一问题，我们借助自注意力的提取形式，将教师与学生的中间层特性进行交叉传递，并在通道和空间上分别生成通道交叉注意力图和空间交叉注意力图，以此作为教师的特征知识先验引导；然后根据生成的通道和空间注意力图来引导学生输出特征，此时再将教师模型的输出特征作为监督信号，令学生输出特征与其趋向于一致，保证教师在学生整个学习过程中的参与来提升知识蒸馏效果。该蒸馏过程的总体框架如图 3-3 所示。

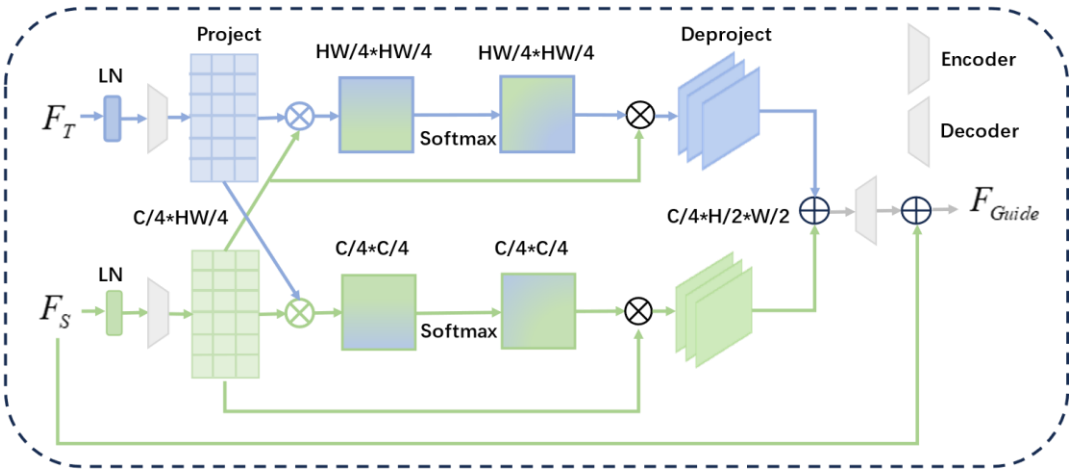



图 3-4 Priori Guidance Based on Cross-attention

在我们提出的方法中，不再纯粹地将教师的特征知识作为监督来辅助学生模型的训练，而是利用教师特征作为先验知识引导学生模型生成先验引导特征，先验引导特征的生成由基于交叉注意力的先验引导模块（Priori Guidance Based on Cross-attention,

PGBC）来实现，该模块具体细节如图 3-4 所示。

首先，为了不破坏样本层的特征像素关系，我们对教师特征和学生特征进行层归一化处理，随后经过层归一化的特征被送入编码器在通道数和尺寸上进行缩减，可以大大减少后续操作的参数量和计算量并节省存储空间。然后对经过编码的特征进行特征映射，以上过程可表示为：

$$F^{Pro} = Project(Encoder(LN(F))) \tag{3-8}$$

其中，$LN(\bullet)$ 表示层归一化，$Encoder(\bullet)$ 表示特征编码，$Project(\bullet)$ 表示特征映射，$F^{Pro}$ 表示映射后的编码特征。此时输入的教师特征和学生特征经过编码在维度上由 C*H*W 变换为 C/4*H/2*W/2，再经过映射变换为 C/4*HW/4。根据等式 3-8 可得到变换后的教师特征 $F_T^{Pro}$ 和学生特征 $F_S^{Pro}$，然后将变换特征 $F_T^{Pro}$ 和 $F_S^{Pro}$ 通过矩阵相乘的方式获取空间和通道上的交叉注意力图，该过程可表示如下：

$$Map\_c = F_T^{Pro} \times (F_S^{Pro})^T, Map\_s = (F_T^{Pro})^T \times F_S^{Pro} \tag{3-9}$$

其中，$Map\_c$ 和 $Map\_s$ 分别代表生成的通道和空间注意力图。与以往的注意图相比，该注意力图包含了教师特征中的先验信息，可以更好地利用教师知识来引导学生特征。由生成的空间和通道注意力图可得到学生模型的空间和通道引导特征，此过程可表示如下：

$$F_C^G = Map\_c \times F_S^{Pro}, F_S^G = F_S^{Pro} \times Map\_s \tag{3-10}$$

其中，$F_C^G$ 和 $F_S^G$ 分别为通道引导特征和空间引导特征。然后将得到的空间和通道引导特征分别进行逆映射并进行点和相加，该过程如下：

$$F^{De} = Deproject(F_c^G) + Deproject(F_s^G) \tag{3-11}$$

其中，$Deproject(\bullet)$ 表示特征逆映射操作，$F^{De}$ 表示逆映射后的空间和通道引导特征之和，此时逆映射后的特征 $F^{De}$ 维度由 C/4*HW/4 变换至 C/4*H/2*W/2。再将特征 $F^{De}$ 送入解码器中，解码后的特征维度变换至 C*H*W。解码后的输出特诊与学生原始特征相加得到最终的引导特征，以上过程可表示如下：

$$F_{Guide} = Decoder(F^{De}) + F_S \tag{3-12}$$

最终，我们用教师特征来监督学生的先验引导特征得到特征知识蒸馏损失，这个过程可表示如下：

$$L_{feat} = D(F_T, F_{Guide}) \tag{3-13}$$

其中，$D(\bullet)$ 表示 $L_2$ 距离损失函数。

我们拟提出的模型交叉注意力先验引导的知识蒸馏方法将教师模型特征知识的作用从输出监督扩展到先验引导，增强了教师在学生模型训练过程中的参与程度，使学生模型以更加合理且高效的形式得到训练。

**（3）基于多级特征传递和预测纠正的自蒸馏方法研究**

单教师和多教师知识蒸馏方法在训练学生模型过程中需要加载一个或多个教师模型，这就造成参数和计算量的大幅增加，给模型训练的硬件设备带来了更大的负担并且训练时间也会被大大延长。

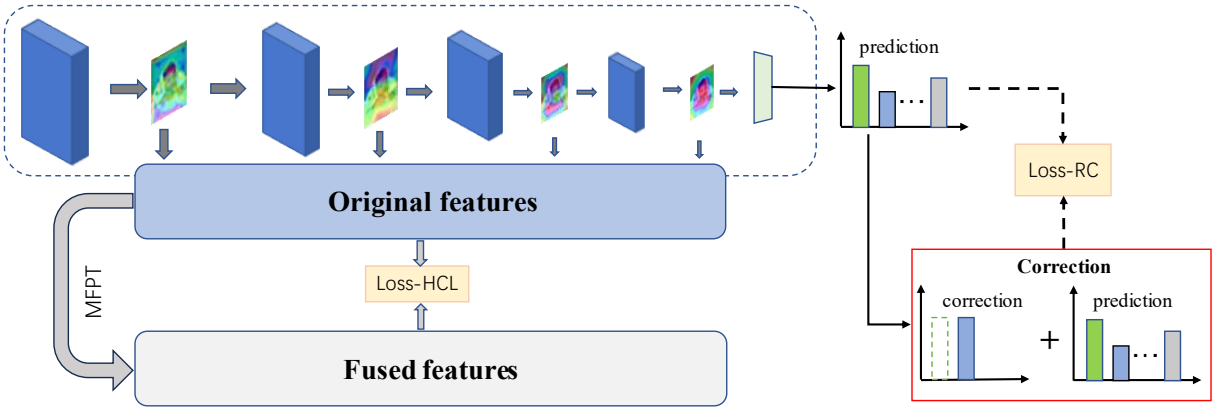
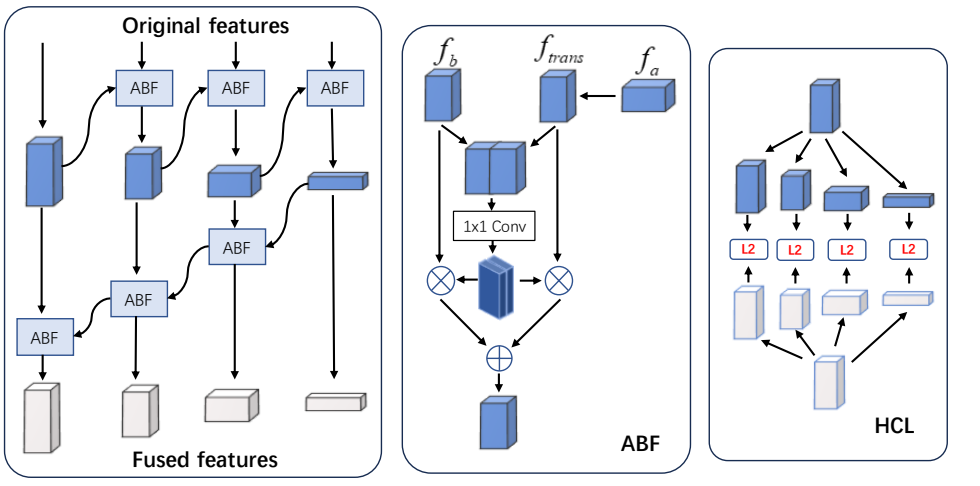



图 3-5 基于多级特征传递和预测纠正的自蒸馏

自蒸馏方法很好的弥补了这个问题，与单教师蒸馏和多教师蒸馏相比，它只有一个神经网络的参与，参数量和计算资源占用较少，所以训练速度也有相当大的提升。在硬件设备资源有限的情况下，自蒸馏是一种很好的选择。但自蒸馏方法，也有自身的一些限制。由于没有其他神经网络的参与，且其自身能力较弱，优秀的监督信息无法被保障，外部知识的缺失影响了蒸馏效果。为了解决这个问题，我们提出基于多级特征传递和预测纠正的自蒸馏方法，该蒸馏方法的总体框架如图 3-5 所示。

在我们的方法中，我们通过多级特征传递得到包含学生模型每一层特征信息的融合特征，将它作为模型特征层的监督信息。此外，我们提出一个预测纠正机制，对模型的预测结果进行纠正，并将纠正后的结果作为输出层的监督信息。总结以上，我们的方法可以分为两部分：多级特征监督和预测纠正机制。

**1）多级特征监督**




(a) Multi-level Feature Progressive Transfer　(b) Attention Based Fusion　(c) Hierarchical Context Loss

图 3-6 多级特征监督

首先，我们将模型的原始特征 $F$ 通过 ABF 模块进行多级特征融合来得到多级融合特征，这里的 ABF 模块与阶段一中提到的 ABF 模块相同，即等式（3-3）。为了方便，我们可令 $f_i \bullet f_{i+1} = F_{\text{ABF}}(f_{i+1}, f_i)$，则前向渐进式融合可表示为：

$$\dot{f}_i = f_1 \bullet f_2 \cdots \bullet f_i \tag{3-14}$$

其中，$f_i$ 表示第 i 层原始特征，$\dot{f}_i$ 表示第 i 层的一阶融合特征。通过等式（3-14），我们可以得到学生模型每个中间层的一阶融合特征 $\dot{F}$，此时的一阶融合特征包含了来自比当前第 i 层更浅的特征。

类似的，我们令 $f_i \circ f_{i+1} = F_{\text{ABF}}(f_i, f_{i+1})$，则反向渐进式融合可表示如下：

$$\ddot{f}_i = \dot{f}_i \circ \dot{f}_{i+1} \cdots \circ \dot{f}_n \tag{3-15}$$

其中，$\ddot{f}_i$ 表示第 i 层的二阶融合特征。通过等式（3-15），我们可以得到学生模型每个中间层二阶融合特征 $\ddot{F}$，此时的二阶融合特征包含来自模型所有中间层特征的信息。

ABF 模块将学生模型在不同尺度和维度上的特征进行融合。然而，直接进行特征蒸馏将导致特征信息提取不充分。这个问题由 HCL 模块有效地解决，。该模块利用金字塔池化来处理模型的原始特征和二阶融合特征，这个过程可表示如下：

$$F^P = P(F) \tag{3-16}$$

其中，$F^P$ 为经过金字塔池化处理的特征。此时可得到多个尺度上的原始特征和二阶融合特征，并用 $L_2$ 距离来计算损失函数，可表示如下：

$$L_{HCL} = D(F^P, \ddot{F}^P) \tag{3-17}$$

其中，$D(\bullet)$ 为 $L_2$ 损失。$F^P, \ddot{F}^P$ 分别为经过金字塔池化后的原始特征和二阶融合特征。

**2）预测纠正机制**

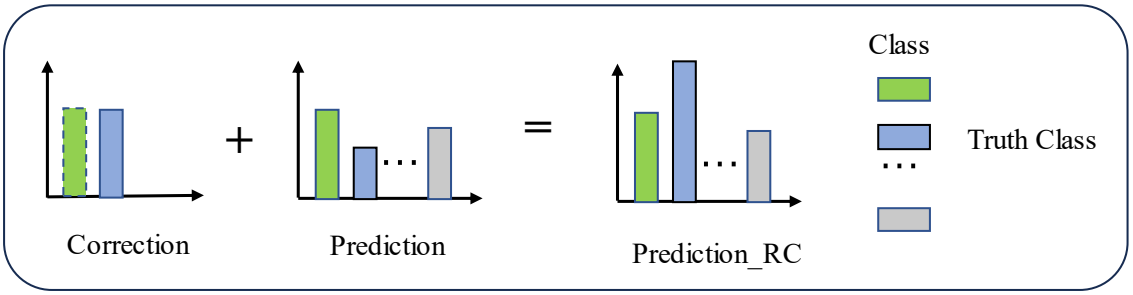



图 3-7 预测纠正

在对预测输出的处理过程中，通过在 Softmax 函数的基础上定义一个温度系数 $\tau$，

将网络的输出转换为概率预测分数，该过程可表示如下：

$$p_i^{\tau} = \frac{\exp(p_i / \tau)}{\sum_j \exp(p_j / \tau)} \tag{3-18}$$

其中， $p_i$ 是对第 i 个样本的类概率预测， $\tau$ 是温度系数。我们取学生模型对第 i 个样本的预测为 $P = \{p_1, p_2, \cdots, p_{cls}, \cdots, p_n\}$。然后，将预测分布中的最大预测值添加至正确类别的预测值，这个过程可表示如下：

$$p_{rc} = p_{cls} + max(P) \tag{3-19}$$

其中 $p_{cls}$ 是正确类别的预测概率值。我们可以得到 $P_{rc} = \{p_1, p_2, \cdots, p_{rc}, \cdots, p_n\}$，然后，令学生模型在纠正后的预测标签监督下进行训练，知识纠正损失可表示如下：

$$L_{RC} = H(P_{rc}^{\tau}, P^{\tau}) \tag{3-20}$$

其中， $H(\bullet)$ 是 KL 散度损失函数。

我们拟提出的多级特征传递和预测纠正的自蒸馏方法通过多级特征监督和预测纠正机制来获取更优质的监督信息，从特征层和输出层两个方面对模型进行自我优化，保证了模型在无外部教师模型监督的情况下得到快速、高效的训练。

**3.创新点**

**（1）针对跨层特征提取和利用不充分问题，本课题提出基于多级特征渐进式传递的知识蒸馏方法。**我们的方法以残差网络堆叠思想为基础，对学生特征进行双向渐进式的融合传递，保证学生模型的每一层中间特征能够得到教师模型的监督，以充分利用教师模型的跨层特征知识，更好地将知识从教师网络转移到学生网络。

**（2）针对教师模型先验引导过程缺失问题，本课题提出基于模型交叉注意力先验引导的知识蒸馏方法。**我们的方法将模型交叉注意力作为先验引导，学生模型可以学会在不同部分的输入数据上分配注意力，而不是仅仅将教师特征作为最终的监督信息，从而更好地捕捉和迁移教师模型的知识。

**（3）针对多个模型加载造成训练负担增加的问题，本课题提出基于多级特征传递和预测纠正的自蒸馏方法。**我们的方法通过多级特征渐进式传递和预测纠正机制分别得到高质量的特征监督信息和输出监督信息，在没有其他模型参与的情况下实现了模型的自我优化，减轻模型训练的硬件压力并提升了训练速度。

**4.在研究过程中可能遇到的困难和问题**

**（1）教师网络和学生网络的特征对齐问题。**

由于教师网络和学生网络的网络架构不同，提取到的特征维数也不同，不能直接进行匹配，如何选择最优的特征回归损失是解决问题的关键。因此，寻找一种有效的特征融合和映射方法，从而减小教师网络和学生网络特征之间的差距。

**（2）训练过程中的先验知识占比调整问题。**

随着学生模型训练过程的深入，学生模型的能力也在不断增强，教师模型的先验引导知识学生模型训练过程中需要进行不断地调整。因此，有效的学生知识衡量手段需要被提出来调整先验知识在训练中的引导占比。

**（3）高质量的监督信息获取问题。**

自蒸馏方法与单教师蒸馏和多教师蒸馏相比，它只有一个神经网络的参与，参数量和计算资源占用较少。但也正是由于没有其他神经网络的参与，且其自身能力较弱，优秀的监督信息无法被保障，外部知识的缺失会影响了蒸馏效果。

**5.预期研究的成果**

（1） 预期本课题结束时，对本课题所涉及的关键技术问题提出解决方案。

（2） 完成三种具有一定创新性的神经网络模型轻量化方法，并完成相应的算法仿真和实验结果分析，给出仿真分析报告一份。

（3） 完成以《基于多级特征先验引导的知识蒸馏方法研究》为题的硕士研究生学位论文。

（4） 预期在国内核心刊物上发表论文 1-2 篇，在国外刊物上发表论文 1-2 篇。

（5） 预期申请发明型专利 1-2 项。

## 四、研究基础

**1.学术条件**

（1）国内外核心期刊文献数据库，已阅读 40+篇知识蒸馏网络、图像分类、目标检测等相关中外文献，为后续研究积累了基础知识。

（2）Cifar-10、Cifar-100、ImageNet、Coco2017 等 5 个开源公共数据库资源。

（3）对知识蒸馏领域已公布代码的代表性方法进行了复现，掌握了 8 种现有知识蒸馏方法在 Cifar-100 数据集上的实验数据，为后续的实验结果对比做了准备。

（4）部分研究内容，已取得进展，向国际期刊《IEEE Signal Processing Letters》（中科院 JCR 工程技术领域二区，影响因子 3.9）投稿题为《Be An Excellent Student: Review, Preview, and Correction》论文一篇（第一作者），已发表。

**2.设施条件**

（1） 硬件方面：高性能计算机系统、多个 2080/3080/3090/4090 深度学习工作站、Jeston nano/TX2 开发板等设备。

（2）软件方面：Matlab、Visual C++、Pycharm、Anaconda 等专业软件。

**3.经费预算**

（金额单位：元 ）

| 预算科目 | 预算经费 | 备注（计算依据与说明） |
|---|---|---|
| 文献查询 | 500 | 书籍、数据库购买 |
| 论文发表 | 5500 | 论文版面费 |
| 实验材料消耗 | 200 | 资料打印 |
| 小型设备购置 | 300 | 存储设备等硬件 |
| 学术会议 | 3000 | 学习调研差旅 |
| 专利申请 | 500 | 专利下载及咨询 |
| 合计 | 10000 | \ |

注：经费来源为研究生培养经费与相关项目经费。

## 五、论文工作计划

| 起止时间 | 工作内容及完成指标 |
| --- | --- |
| 2023.09-2023.12 | 完成基于多级特征先验引导的知识蒸馏算法的初步建立。 |
| 2024.01-2024.03 | 设计并完成基于多级特征渐进式传递的知识蒸馏方法，撰写论文。 |
| 2024.04 -2024.06 | 完成基于模型交叉注意力先验引导的知识蒸馏方法，撰写论文。 |
| 2024.07-2024.09 | 完成基于多级特征传递和预测纠正的自蒸馏方法，撰写论文。 |
| 2024.10-2024.12 | 利用实验仿真结果找出算法的不足之处，对算法进一步改进和优化。 |
| 2025.01-2025.02 | 对所有实验结果和方法进行归纳整理。 |
| 2025.03-2025.05 | 撰写学位论文，准备毕业答辩。 |

## 六、导师评审意见

**导师意见（导师就研究生对国内外研究现状的了解情况、研究方法、研究手段及论文工作计划的评价）：**

**导师签名：** **年 月 日**

## 七、评审小组和学院意见

| | 姓 名 | 职 称 | 专 业 | 成 绩 | 结 论 | 签 字 |
|---|---|---|---|---|---|---|
| 组长 | 景军峰 | 教授 | 控制科学与工程 | | | |
| 成员 | 钱慧芳 | 副教授 | 控制科学与工程 | | | |
| | 宋玉琴 | 副教授 | 控制科学与工程 | | | |
| | 张榆红 | 博士 | 控制科学与工程 | | | |
| | 高原 | 高工 | 控制科学与工程 | | | |
| 总成绩 | | | | | | |

评审小组审查结论（论文选题的意义、论文的难度与工作量是否适当、研究方案的可行性、是否同意论文开题）：

组长签名：　　年　月　日

学院教授委员会意见：

学院教授委员会主任签名：　　年　月　日