Modeling and Simulation 建模与仿真, 2025, 14(2), 279-290
Published Online February 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.142151

Hans汉斯

基于金字塔池化以及掩码生成的特征知识蒸馏

陆新才[1]，孙占全[1]，王　贺[2]，李庆蓬[1]

[1]上海理工大学光电信息与计算机工程学院，上海
[2]河南大学经济学院，河南　郑州

收稿日期：2025年1月24日；录用日期：2025年2月17日；发布日期：2025年2月27日

摘　要

知识蒸馏(KD)的目标是将知识从大型教师网络传递到轻量级的学生网络中去。主流的KD方法可以被分为Logit蒸馏和特征蒸馏。基于特征的知识蒸馏是KD的重要组成部分，它利用中间层来监督学生网络的训练过程。然而，中间层的潜在不匹配可能会在训练过程中适得其反，并且目前的学生模型往往直接通过模仿老师的特征来学习。针对这一问题，本文提出了一种新的知识蒸馏框架，称为解耦空间金字塔池知识蒸馏，以区分特征图中区域的重要性。同时，本文还提出了一种掩码生成特征蒸馏模块，指导学生模型通过一个块生成而不是模仿教师的完整特征。与之前复杂的蒸馏方法相比，本文提出的方法在CIFAR-100和Tiny-ImageNet数据集上取得了更高的知识蒸馏模型分类结果。

关键词

模型压缩，知识蒸馏，特征蒸馏

Feature Knowledge Distillation Based on Pyramid Pooling and Mask Generation

Xincai Lu[1], Zhanquan Sun[1], He Wang[2], Qingpeng Li[1]

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai
[2]School of Economics, Henan University, Zhengzhou Henan

Received: Jan. 24th, 2025; accepted: Feb. 17th, 2025; published: Feb. 27th, 2025

Abstract

The goal of Knowledge Distillation (KD) is to transfer knowledge from a large teacher network to a

文章引用: 陆新才, 孙占全, 王贺, 李庆蓬. 基于金字塔池化以及掩码生成的特征知识蒸馏[J]. 建模与仿真, 2025, 14(2): 279-290. DOI: 10.12677/mos.2025.142151

lightweight student network. Mainstream KD methods can be divided into logit distillation and feature distillation. Feature-based knowledge distillation is a critical component of KD, utilizing intermediate layers to supervise the training process of the student network. However, potential mismatches in intermediate layers may backfire during training, and current student models often learn directly by imitating the teacher's features. To address this issue, this paper proposes a novel distillation framework called Decoupled Spatial Pyramid Pooling Knowledge Distillation, which distinguishes the importance of regions in feature maps. This paper also introduces a mask-based feature distillation module, which guides the student model to generate features from a block rather than mimicking the complete features of the teacher model. Compared to previous complex distillation methods, the proposed approach achieves superior classification results on the CIFAR-100 and Tiny-ImageNet datasets.

Keywords

Model Compression, Knowledge Distillation, Feature Distillation

1. 引言

在过去的几十年中，深度神经网络推动了计算机视觉领域的蓬勃发展，在比如图像分类[1]-[3]、目标检测[4] [5]、和分割[6] [7]方面获得了巨大的性能提升。但是，由于强大的网络性能常常依赖于庞大的模型容量，它们严重依赖于算力和存储资源，在某些特定环境比如移动设备上无法完成部署。为了解决这一问题，许多方法被提出用来压缩模型的大小。知识蒸馏(Knowledge Distillation)正是其中的一种方法。具体来说，知识蒸馏框架主要包括一个大型模型(教师)和一个小型模型(学生)，通过将知识从教师转移到学生的方法，在不增加额外成本的前提下提高了小型模型(学生)的网络性能。

主流的知识蒸馏方法主要分为软目标蒸馏和特征蒸馏。软目标蒸馏仅仅在软目标层面，通过最小化教师和学生之间的KL 散度(Kullback-Leibler Divergence) [8] [9]来转移知识。为了更好地利用教师的知识，最近的研究更加专注于教师模型的特征层，通过匹配教师和学生之间的特征分布来进行蒸馏。这种方法被称为特征蒸馏[10]。以前基于特征的蒸馏方法通常会让学生尽可能模仿老师的输出，因为老师的特征具有更强的表现力。然而，本文认为没有必要直接模仿老师来提高学生特征的表征能力。用于蒸馏的特征一般是通过深度网络得到的高阶语义信息。特征像素已经在一定程度上包含了相邻像素的信息。

面对这种情况，本文重点提出了一种掩码生成式解耦特征蒸馏算法(MDKD)，这是一种简单高效的基于特征的蒸馏方法。具体来说，首先，本文方法屏蔽了学生特征的随机像素，然后通过一个简单的模块使用屏蔽特征生成教师的完整特征。由于每次迭代都使用随机像素，因此在整个训练过程中将使用所有像素，这意味着特征将更加鲁棒，并且其表示能力将得到提高。其次，本文还提出了一种解耦空间金字塔池化知识蒸馏(DSPP)，该方法中应用了空间金字塔池化[11]架构来自动捕获知识，它可以有效地捕获特征图不同尺度的信息知识。然后，基于特征图中较低激活区域在 KD 中发挥更重要作用的观察结果，即较低激活区域包含更多信息知识线索，设计了一个解耦模块来分析学生和教师网络之间的区域级语义损失。通过利用空间金字塔池化和解耦的区域级损失分配，可以通过更复杂的监督有效优化学生网络。大量实验数据证明，在主流基准测试中我们的方法在同构及异构网络知识蒸馏配置中优于现有的蒸馏技术。

2. 相关知识

知识蒸馏(Knowledge Distillation, KD)的概念最早由 Hinton 等人提出。涉及一个大型的教师模型和一个轻量级的学生模型。该框架的目标是将教师模型中的知识提炼并转移至学生模型中。具体操作上，通过最小化教师和学生模型预测之间的差异，迫使学生模型模仿教师的输出。知识蒸馏旨在通过从较大教师网络中提取的暗知识，提升较小学生网络的性能。

根据现有的知识，主要把知识蒸馏方法分为三类：基于软目标的蒸馏[8] [9] [12]-[15]、基于中间特征的蒸馏[11] [18]-[23]和基于关系的蒸馏[24]。软目标蒸馏方法侧重于模型输出的 Logit，而特征蒸馏方法则侧重于模型内部特征的提取和转移，现有的基于关系的知识蒸馏方法则对不同层与数据样本之间以及不同样本之间的关系进行了探究。这些方法各有其特点和应用场景，为模型压缩和知识蒸馏提供了多样的技术选择。

软目标蒸馏方法通过输出的 Logit 实现知识的提取。先前的关于软目标蒸馏的研究主要集中在开发有效优化方法上，例如，DKD [8]通过用一个常数值替换与教师置信度负相关的系数，从而将损失从整体软目标蒸馏中解耦出来，提高了对预测良好样本的蒸馏效果。DML [13]通过同时训练两个小型学生网络，利用分类损失函数和模仿对方网络的损失函数，实现学生网络的类别后验对齐。TAKD [12]则引入了一个名为“教师助理”的中型网络，旨在缩小教师和学生之间的差距。MLKD [14]通过实例、批次和类级别的多级预测对齐，优化了知识转移过程。CTKD [15]通过动态调整学习阶段的任务难度，实现了渐进式学习。CTKD 以易到难的课程形式，逐步提升蒸馏损失的温度，从而增加学习任务的难度。此外，还有一些研究[16] [17]侧重于对经典知识蒸馏方法进行解释。基于软目标的知识蒸馏算法简单，可以与其他知识蒸馏方法结合，但单独使用的话，模型的性能提升有限。除此之外，这种知识蒸馏方法依赖于 softmax 函数，只能用于分类相关的任务，并与类别的数量有关。最后，这些方法不能应用于无标签任务。

为了进一步促进知识蒸馏，提出了特征蒸馏这一新的研究方向，该方法对中间特征而不是软目标输出进行蒸馏。具体而言，FitNets [10]扩展了 Hinton 提出的知识蒸馏(KD) [8]方法，通过利用教师模型特征提取器的中间层输出作为提示，结合 KD 对更深且更窄的学生模型进行知识传递。AT [21]通过提取复杂模型生成的注意力图来指导简单模型，使其生成的注意力图与复杂模型相似。与 FitNet 不同，AT 采用了多个层来进行知识转移，来捕获多层次信息，从而更好地提高学生的性能。MGD [23]通过随机屏蔽学生模型特征的像素并重构教师模型的完整特征来增强学生模型的表征能力。CRD [20]通过对比性学习来训练学生模型，使其能在教师的数据表述中捕捉到更多信息的新目标。其他研究[25]则通过提取输入的相关性来传递教师的知识。此外，还有一些研究通过同时对中间特种和软目标输出一同蒸馏，SAKD [26]在整个蒸馏期间的每次训练迭代中自适应地确定每个样本的教师网络中的中间特征和软目标层的蒸馏点。CAT-KD [27]引入了类别注意力转移(CAT)和类激活映射(CAM)转移，通过转移它们来增强学生模型的性能。研究者们提出了很多有效的基于特征知识的蒸馏方法，但是如何选取教师模型的暗示依然是一个值得探究的问题。比较常用的选取策略是跨层选取，例如 12 层模型蒸馏到 4 层模型，选取策略为每三层的输出选取一层作为暗示。其他有效的选取策略有待进一步挖掘。

与之前只关注单个样本的输出结果不同，RKD [24]将输出样本之间的结构关系迁移给学生。通过将同批次的两个样本之间的距离关系以及三个样本之间的角度关系作为知识传递给学生，让学生学习教师模型的结构化信息。

3. 方法

3.1. 知识蒸馏

对于包含 C 类的样本，其分类概率可以表示为 $P=\left[p_1,p_2,\cdots,p_i,\cdots,p_C\right]\in\mathbb{R}^{1\times C}$，其中 p_i 表示第 i 类的

概率，C 为该数据集类别数。P 中的每个元素都可以通过 softmax 函数获得：

$$p_i = \frac{\exp\left(z_i/T\right)}{\sum_{j=1}^{C}\exp\left(z_j/T\right)}. \tag{1}$$

其中 z_i 表示第 i 类的 logit 值，T 表示用于温度缩放的超参数。在知识蒸馏中，T 通常大于 1.0，这有助于帮助学生更好的学习教师网络。知识蒸馏的过程是通过温度变化控制教师网络输出，最小化教师模型和学生模型输出之间的 KL 散度来实现蒸馏。

$$L_{KD} = KL\left(p^{tea}\middle\|p^{stu}\right) = \sum_{j=1}^{C} p^{tea}\log\left(\frac{p_j^{tea}}{p_j^{stu}}\right). \tag{2}$$

对于不同的任务，模型的架构差异很大。此外，大多数蒸馏方法都是为特定任务而设计的。然而，基于特征的蒸馏可以应用于分类和密集预测。特征蒸馏的基本方法可以表述为：

$$L_{fea} = \sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(F_{k,i,j}^{T} - f_{align}\left(F_{k,i,j}^{S}\right)\right)^2 \tag{3}$$

其中 F^T 和 F^S 分别表示教师和学生的特征，f_{align} 是学生的特征与教师的特征的对齐的适应层。C,H,W 表示特征图的形状。

在本节中，主要讨论了基于特征的知识蒸馏的机制，整体结构如图 1 所示。以往的特征蒸馏方法可以指导学生直接模仿老师的特征，然而，本文提出了掩码生成蒸馏，指导学生生成老师的特征而不是模仿他。此外，本文提出了一种新的解耦金字塔池化蒸馏方法定义知识，以及利用解耦特征图来优化知识蒸馏(KD)训练过程，从而更充分地利用中间层知识。

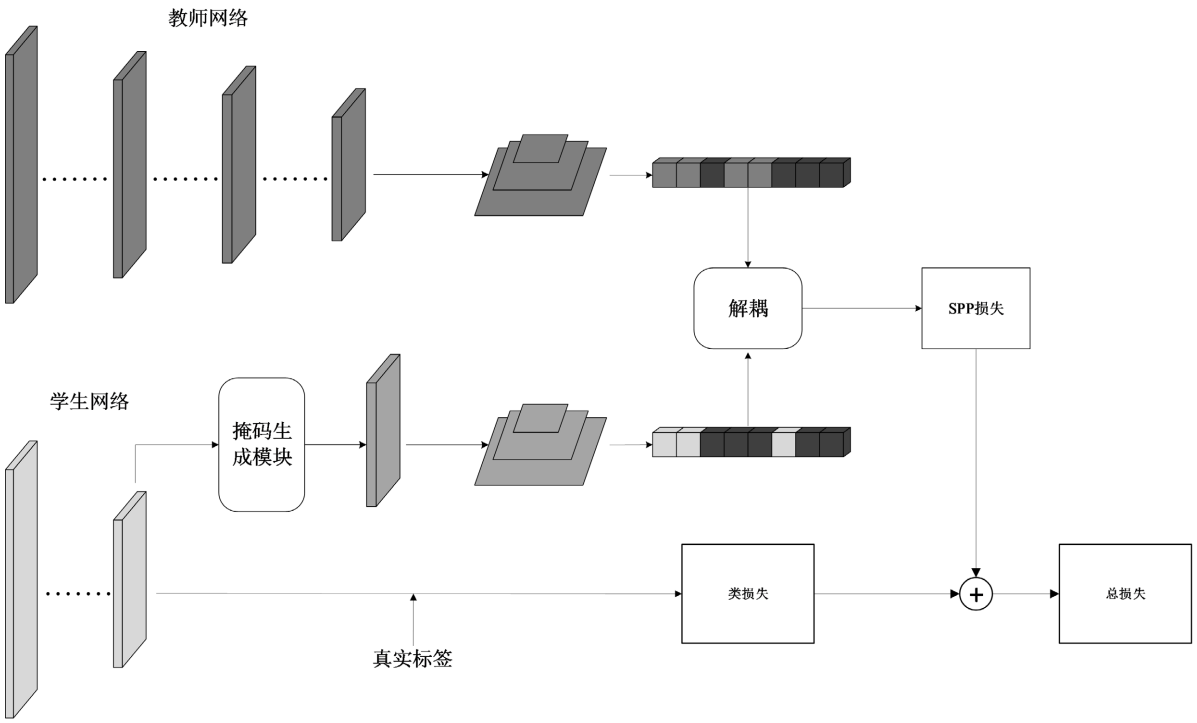

Figure 1. General structure diagram
图 1. 总体结构图

3.2. 掩码特征生成

对于 CNN 模型来说，深层次的特征具有更大的感受野和更好的原始输入图像表示。也就是说，特征像素已经一定程度上包含了相邻像素的信息。因此，使用部分像素来恢复完整的特征图的方式是可以的，本文的方法旨在通过学生的屏蔽特征生成教师的特征，这可以帮助学生获得更好的表示。图 2 为本文提出的掩码特征生成结构图。

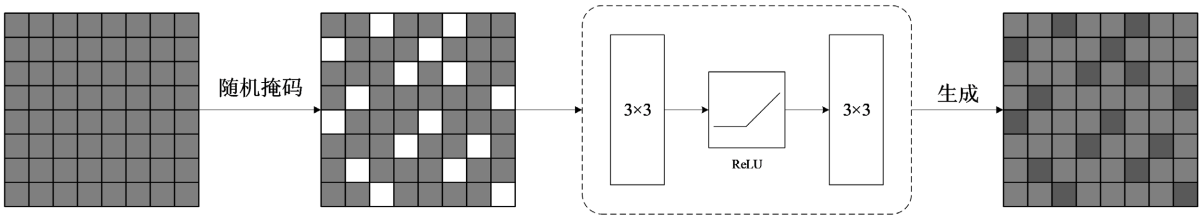

Figure 2. Mask feature generation structure diagram
图 2. 掩码特征生成结构图

本文分别将教师和学生的第 l 个特征图表示为 $T^l \in R^{C\times H\times W}$ 和 $S^l \in R^{C\times H\times W}\left(l=1,\cdots,L\right)$。首先，设置第 l 个随机掩码来覆盖学生的第 l 个特征，其表示为：

$$M_{i,j}^{l}=\begin{cases}0, & R_{i,j}^{l}<\lambda\\ 1, & R_{i,j}^{l}\geq\lambda\end{cases} \tag{4}$$

其中 $R_{i,j}^{l}$ 是(0, 1)中的随机数，i，j 分别是特征图的水平和垂直坐标。λ 是一个超参数，表示掩码的比例。第 l 个特征图将会被第 l 个随机掩码所覆盖。

然后我们使用相应的掩码来覆盖学生的特征图，并尝试用左侧像素生成教师的特征图，具体公式如下：

$$T^{l}\leftarrow\vartheta\left(f_{align}\left(S^{l}\right)\cdot M^{l}\right) \tag{5}$$

$$\vartheta\left(F\right)=W_{l2}\left(\mathrm{ReLU}\left(W_{l1}\left(F\right)\right)\right) \tag{6}$$

其中，ϑ 表示为包括两个卷积层 Wl_1，Wl_2 和一个激活层 ReLU 的投影层。在本文中，适应层被设置为 1 × 1 的卷积层，Wl_1，Wl_2 采用 3 × 3 的卷积层。

根据以上方法，本文的掩码特征生成蒸馏的损失可以被表示为：

$$L_{MFG}\left(S,T\right)=\sum_{l=1}^{L}\sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(T_{k,i,j}^{l}-\vartheta\left(f_{align}\left(S_{k,i,j}^{l}\right)\cdot M_{i,j}^{l}\right)\right)^{2} \tag{7}$$

其中 L 为蒸馏层的数目，C，H，W 表示特征图的形状大小。S 和 T 分别表示学生和教师的特征。

3.3. 解耦金字塔池知识蒸馏

本文提出的 DSPP 架构如图 3 所示，学生网络和教师网络通过解耦的空间金字塔池相互交互。从文献[8]中能够得知，logit 实际上是类别的概率分布，其对于学生网络过于抽象而使得学生网络无法获得全面的信息性知识。此外，由于模型大小的不同，找到适当匹配的教师和学生模型的提示层是非常困难的。由于以上原因，本文选择使用最后一个提示层来解决 logit 高度抽象问题以及匹配提示层的复杂问题。在最后一层提示层中，本文使用了转换的操作对齐了学生和教师的最后一层提示层。受到文献[28]的启发，

本文引入了一种新的方法来定义提示层中的知识，应用空间金字塔池来捕获不同尺度的提示层中的知识，从而解决了最后一层提示层可能存在的知识过于集中而不全面的问题。此外，本文还提出了一个解耦模块来提高了空间金字塔中较低激活区域的重要性。

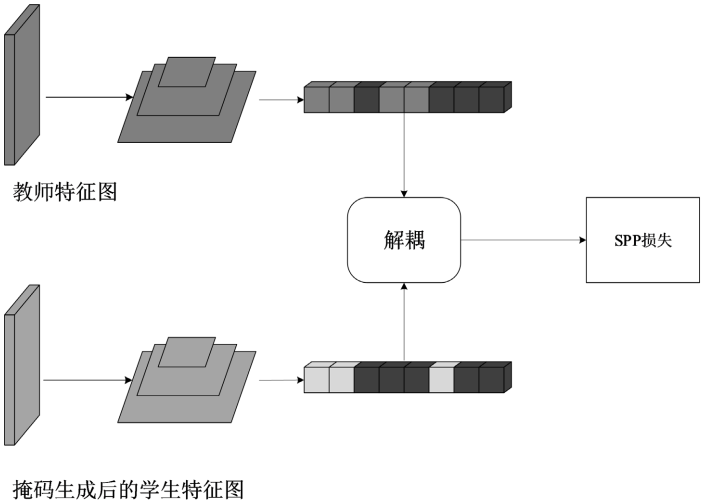

Figure 3. Decoupled pyramid pool knowledge distillation structure diagram
图 3. 解耦金字塔池知识蒸馏结构图

3.3.1. 空间金字塔池

空间金字塔池化技术最初在文献[11]中被引入至视觉识别领域，该创新策略显著地解除了卷积神经网络对于固定输入尺寸的依赖。鉴于教师模型与学生模型在架构设计上存在的差异，本文采用了空间金字塔池化方法，以应对两者在最后一个特征层(提示层)上形状不匹配的问题。进一步地，空间金字塔池化通过其分层结构，为最后一个特征层引入了多尺度的感受野，这一特性赋予了本文模型从该特征层中同时捕获全局与局部知识的能力。具体而言，空间金字塔池化的实施过程可形式化描述如下：

$$f_{pyramid} = Pooling\left(L, W\left(L\right)/k\right), k = 1,2,\cdots,n, L \in \mathbb{R}^{b\times c\times h\times w} \tag{8}$$

其中，L 表示输入提示层，$W(\cdot)$表示计算 L 的长度函数，k 为金字塔层的层数，k 共有 n 层。函数 $Pooling(\cdot)$ 中总包含两个参数，输入特征图以及池化核的大小。

众所周知，学生模型的 logit 来自全连接层，与教师模型是高度相似，即它们在一个数据集上的预测是相同的。然而，对于图像分类任务，全连接层缺少了输入图像的二维或三维空间信息。如前所述，很难找到教师和学生模型的提示层的适当匹配，并且这可能会降低 KD 的可解释性，这是本文选择了最后一个提示层的动机。此外，全连接层直接从最后一个提示层计算得出，该层在理论上最接近所有提示层中的 logit。

3.3.2. 解耦模块

为了更多的关注较低的激活区域，本文提出了一个解耦模块来处理空间金字塔池化的扁平化特征。在解耦模块中，根据特征中的每个元素的值将扁平化的特征解耦为两个组件。如图 4 所示，学生特征 V_s 通过双向箭头与教师特征 V_t 进行元素匹配。红色箭头指向 V_t 中的 n 个最大元素，其另一端指向 V_s 中相应的位置。相反，蓝色箭头指向 V_t 中的最后一个尾部$(N-n)$个元素，其中 N 表示 V_s 或 V_t 的长度，SPP 的损失可以计算为：

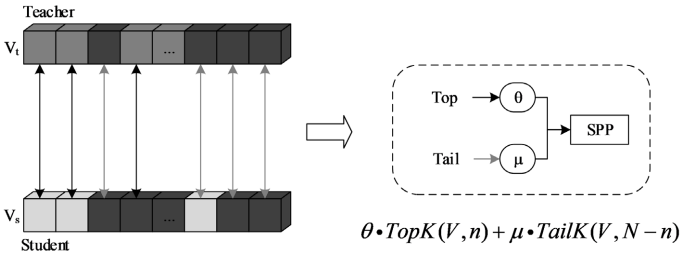

Figure 4. Decoupling module structure diagram
图 4. 解耦模块结构图

$$\begin{aligned} &top(n) = \arg\max\left(V_t\right)[0:n], n = 0,1,\cdots,N; \\ &tail(n) = \arg\max\left(V_t\right)[0:m], m = N-n; \\ &L_{DSPP} = \theta L_2\left(V_t\left[top(n)\right], V_s\left[top(n)\right]\right) + \mu L_2\left(V_t\left[tail(m)\right], V_s\left[tail(m)\right]\right), V_t, V_s \in \mathbb{R}^N \end{aligned} \tag{9}$$

其中 $top(\cdot)$表示 V_t 中 top-$(\cdot)$元素的索引，$tail(\cdot)$表示 V_t 中 $tail$-$(\cdot)$元素的索引，函数 $L2(\cdot)$表示 $L2$ 范数距离。θ 和 μ 是控制解耦权重的超参数。为了提高较低激活区域的重要性，本文让 μ 大于 θ。本文方法更加关注较低激活区域的原因是，具有大量参数的强大教师模型可能具有更复杂的机制来查找反应较低激活区域的输入的更多细节。较低的激活区域有助于提高学生模型的准确性和泛化性。

综上所述，本文将 DSPP 应用到知识蒸馏任务中去，并将其与本文提出的掩码特征生成蒸馏相结合，其公式如下：

$$L_{total} = L_{CE} + \alpha L_{\mathrm{MFG}} + \beta L_{DSPP}. \tag{10}$$

其中 L_{CE} 代表目标与仅来自学生模型预测之间的广泛使用的交叉熵损失，α，β 分别用于平衡 L_{MFG} 和 L_{DSPP} 的权重。

通过将损失的三个部分整合到一起，本文的方法，不仅通过学生的屏蔽特征生成教师的特征，帮助学生获得更好的表示，也应用解耦的语义损失分配来提高在 KD 中发挥更重要作用的较低激活区域的权重，旨在减轻学生网络的训练难度。通过该方法，可以让学生更加学习更加丰富的教师知识，对于学生网络的性能提升起着重要的作用。

4. 实验

4.1. 数据集与设置

在本文的实验中，我们对图像分类的性能进行了评估。数据集方面，本文选择了两个广泛研究的数据集：1) CIFAR-100 [29]，这是一个著名的图像分类数据集，包含 100 个类别的 32 × 32 像素的图片，其中 50000 张图片作为训练集，10000 张作为验证集。2) Tiny-ImageNet [30]，这是一个大规模的分类数据集，是图像分类领域最重要的基准数据集之一，包含 200 个类别的 100000 张图像(每个类别 500 张)，缩小为 64 × 64 彩色图像。每个类别有 500 张训练图像、50 张验证图像和 50 张测试图像。

设置方面，本文的实验着重在知识蒸馏上，具体包括了两种不同的设置：1) 同构架构，即教师模型与学生模型采用相同的模型架构，仅仅是模型层数不相同，例如 ResNet56 和 ResNet20。2) 异构架构，即教师模型的模型架构是与学生模型完全不相同的，例如 ResNet 32 × 4 和 ShuffleNetV2。本文的实验包括了多种神经网络架构，如 ResNet [1]，ShuffleNet [31]，vgg [32]，WRN [33]，MobileNet [34]。

实验配置方面，在 CIFAR-100 数据集实验中，本文将 batch 大小设置为 64，基础学习率为 0.05。在 ImageNet 数据集实验中，本文将 batch 大小设置为 128，基础学习率为 0.01。本文使用 1 块 Nvidia RTX

3090 作为训练显卡。

4.2. 实验结果

在本文的实验中，我们评估了该方法的性能，同时与目前主流知识蒸馏方法进行了比较，包括了主流的软目标蒸馏方法以及特征蒸馏方法。表 1 给出了基于五种同构网络模型组合和八种 KD 方法的 CIFAR-100 上的 Top-1 测试精度，与我们提出的 MD 知识蒸馏进行了比较。其他方法的部分结果引用自文献[28]。根据表 1，表明本文方法在 KD 的参与下始终比最先进的蒸馏方法获得更高的精度。七种异构网络模型组合的结果如表 2 所示。显然，当时教师-学生模型组合从同构切换到异构时，在多个中间层上构建的方法往往比提取最后基层或 logit 的方法表现更差。一些方法甚至可能在学生网络的训练过程中起到相反的负面作用。例如，AT 和 FitNet 的表现甚至比普通学生还要差。这就如之前章节所述，可能是由于提示层的不匹配而造成得这种现象。

具体来说，在本研究中，我们对 CIFAR-100 和 Tiny-ImageNet 数据集进行了深入的实验分析，探讨了同构与异构架构下的知识蒸馏效果。在 CIFAR-100 数据集中，同构架构实验中，教师模型 ResNet110 的 Top-1 准确率为 74.31%，而采用相同架构但层数减少的学生模型 ResNet32 的 Top-1 准确率为 71.14%。通过与现有的 ReviewKD (特征蒸馏)方法的比较，我们发现，ReviewKD 方法将准确率提升至 73.89%，而本研究提出的方法能将学生模型的准确率提升至 74.13%。在异构架构实验中，教师模型 ResNet 32 × 4 的 Top-1 准确率为 79.42%，学生模型 ShuffleNetV2 的原始准确率为 71.82%。相较于 ReviewKD (特征蒸馏)和 DKD (Logit 蒸馏)方法，其分别将准确率提升至 77.78%和 77.07%，本研究方法能显著提升至 78.13%。

为了评估本文方法的泛化性能，本文同样在 Tiny-ImageNet 上对三种经典的师生架构进行了一系列的实验，如表 3 所示。结果表明本文方法优于其他方法，包括 CRD 和 SAKD 的组合，这进一步证明了本文方法的有效性。Tiny-ImageNet 中的图像比 CIFAR-100 中的图像大两倍，特征图同样大两倍，可以提供更多的信息。因此，本文方法在 Tiny-ImageNet 上的性能优于 CIFAR-100。

Table 1. Homogeneous architecture CIFAR-100 results
表 1. 同构架构 CIFAR-100 结果

方法						
	教师网络	ResNet56	ResNet110	ResNet32×4	WRN-40-2	VGG13
		72.34	74.31	79.42	75.61	75.61
	学生网络	ResNet20	ResNet32	ResNet8×4	WRN-16-2	VGG8
		69.06	71.14	72.50	73.26	70.36
软目标	KD	70.66	73.08	74.92	73.54	72.98
	DML	69.52	72.03	73.58	72.68	71.79
	TAKD	70.83	73.37	75.06	74.33	73.23
特征	FitNet	69.21	71.06	73.50	72.24	71.02
	RKD	69.61	71.82	73.35	72.22	71.48
	CRD	71.16	73.48	75.51	74.14	73.94
	OFD	70.98	73.23	75.48	74.33	73.95
	ReviewKD	71.89	73.89	75.63	75.09	74.84
	Ours	71.35	74.13	76.03	76.87	74.96

Table 2. Heterogeneous architecture CIFAR-100 results
表 2. 异构架构 CIFAR-100 结果

方法						
	教师网络	ResNet 32 × 4	ResNet 32 × 4	ResNet 32 × 4	WRN-40-2	WRN-40-2
		79.42	79.42	79.42	75.61	75.61
	学生网络	ShuffleNet-V2	WRN-16-2	WRN-40-2	ResNet 8 × 4	MobileNet-V2
		71.82	73.26	75.61	72.5	64.6
软目标	KD	74.45	74.9	77.7	73.97	68.36
	CTKD	75.37	74.57	77.66	74.61	68.34
	DKD	77.07	75.7	78.46	75.56	69.28
特征	FitNet	73.54	74.7	77.69	74.61	68.64
	AT	72.73	73.91	77.43	74.11	60.78
	RKD	73.21	74.86	77.82	75.26	69.27
	CRD	75.65	75.65	78.15	75.24	70.28
	OFD	76.82	76.17	79.25	74.36	69.92
	ReviewKD	77.78	76.11	78.96	74.34	71.28
	SimKD	78.39	77.17	79.29	75.29	70.1
	Ours	78.13	76.37	78.77	76.83	70.88

对于 Tiny-ImageNet 数据集，同构架构中教师模型 ResNet34 的 Top-1 准确率为 73.31%，学生模型 ResNet18 的准确率原为 69.75%，而本研究方法提升至 71.93%，相比 KD 方法的 70.66%表现出显著优势。在异构架构中，使用 ResNet50 作为教师模型，MobileNetV2 作为学生模型，其原始 Top-1 准确率为 68.87%，通过本研究方法提升至 72.64%，而 DKD 方法的效果为 72.05%。

Table 3. Tiny-ImageNet results
表 3. Tiny-ImageNet 结果

		Top-1	Top-5	Top-1	Top-5
方法	教师网络	ResNet34		ResNet50	
		73.31	91.42	76.16	92.86
	学生网络	ResNet18		MobileNet-V2	
		69.75	89.07	68.87	88.76
软目标	KD	70.66	89.88	68.58	88.98
	DML	70.82	90.02	71.35	90.31
	TAKD	70.78	90.16	70.82	90.01
	DKD	71.7	90.41	72.05	91.05
特征	AT	70.69	90.01	69.56	89.33
	OFD	70.81	89.98	71.25	90.34
	CRD	71.17	90.13	71.37	90.41
	ReviewKD	71.61	90.51	72.56	91
	Ours	71.93	90.38	72.64	90.82

4.3. 消融实验

该小节研究了本文方法中每个组成部分的贡献，包括掩码特征生成蒸馏(MFD)和解耦金字塔池知识蒸馏(DSPP)，如表 4 所示。实验在 CIFAR-100 上进行，以 ResNet 32 × 4，ResNet 8 × 4 和 ShuffleNet-V2 分别作为同构和异构的学生网络，以 Top-1 准确率作为评价指标。当采用所有结构时，该方法的表现超过了所有其他蒸馏方法，证明了我们方法的每个部分都是不可或缺的。

Table 4. Ablation experiment
表 4. 消融实验

MFD	DSPP	ResNet 8 × 4	ShuffleNet-V2
√		75.85	77.91
√	√	76.03	78.13

5. 结论

本文提出了一种新型知识蒸馏方法，主要包括掩码特征生成蒸馏(MFD)和解耦空间金字塔池化知识蒸馏(DSPP)。这些方法通过改进学生模型与教师模型的特征对齐方式，实现了知识蒸馏的性能提升。具体来说，MFD 创新性地引入了特征生成策略，与传统的模仿教师特征的方式不同，学生模型通过屏蔽随机像素并利用简单的生成模块直接生成教师的完整特征。由于训练过程中屏蔽像素的随机性，模型得以利用全部像素，从而提升了特征的鲁棒性与表示能力。另一方面，DSPP 通过解耦的空间金字塔池化操作，减少了对中间层的依赖，并有效捕获了多尺度知识。针对特征图低激活区域包含更多知识线索的观察结果，本文设计了一个解耦模块，用于分析教师和学生网络之间的区域级语义损失。结合空间金字塔池化与解耦区域损失分配的策略，DSPP 实现了对学生网络的高效优化。在 CIFAR-100 和 Tiny-ImageNet 数据集上的实验结果表明，无论是同构还是异构网络配置，本文方法均显著优于现有蒸馏技术，证明了其在图像分类任务中的优越性。这种结合生成式掩码与解耦池化的蒸馏策略，为知识蒸馏领域提供了一种全新的解决思路。

参考文献

[1] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[2] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90. https://doi.org/10.1145/3065386

[3] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7132-7141. https://doi.org/10.1109/cvpr.2018.00745

[4] Li, Q., Jin, S. and Yan, J. (2017) Mimicking Very Efficient Network for Object Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 7341-7349. https://doi.org/10.1109/cvpr.2017.776

[5] Lin, T., Dollar, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017) Feature Pyramid Networks for Object Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 936-944. https://doi.org/10.1109/cvpr.2017.106

[6] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/cvpr.2015.7298965

[7] Ren, S., He, K., Girshick, R. and Sun, J. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[8] Geoffrey, H., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network. arXiv: 1503.02531. https://doi.org/10.48550/arXiv.1503.02531

[9] Zhao, B., Cui, Q., Song, R., Qiu, Y. and Liang, J. (2022) Decoupled Knowledge Distillation. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 11943-11952. https://doi.org/10.1109/cvpr52688.2022.01165

[10] Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C. and Bengio, Y. (2015) FitNets: Hints for Thin Deep Nets. arXiv: 1412.6550. https://doi.org/10.48550/arXiv.1412.6550

[11] He, K., Zhang, X., Ren, S. and Sun, J. (2015) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 1904-1916. https://doi.org/10.1109/tpami.2015.2389824

[12] Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A. and Ghasemzadeh, H. (2020) Improved Knowledge Distillation via Teacher Assistant. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 5191-5198. https://doi.org/10.1609/aaai.v34i04.5963s

[13] Zhang, Y., Xiang, T., Hospedales, T.M. and Lu, H. (2018) Deep Mutual Learning. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4320-4328. https://doi.org/10.1109/cvpr.2018.00454

[14] Jin, Y., Wang, J. and Lin, D. (2023) Multi-Level Logit Distillation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 24276-24285. https://doi.org/10.1109/cvpr52729.2023.02325

[15] Li, Z., Li, X., Yang, L., Zhao, B., Song, R., Luo, L., *et al*. (2023) Curriculum Temperature for Knowledge Distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 1504-1512. https://doi.org/10.1609/aaai.v37i2.25236

[16] Phuong, M. and Lampert, C.H. (2019) Towards Understanding Knowledge Distillation. International Conference on Machine Learning. arXiv: 2105.13093. https://doi.org/10.48550/arXiv.2105.13093

[17] Cheng, X., Rao, Z., Chen, Y. and Zhang, Q. (2020) Explaining Knowledge Distillation by Quantifying the Knowledge. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 12922-12932. https://doi.org/10.1109/cvpr42600.2020.01294

[18] Chen, P., Liu, S., Zhao, H. and Jia, J. (2021) Distilling Knowledge via Knowledge Review. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 5006-5015. https://doi.org/10.1109/cvpr46437.2021.00497

[19] Chen, D., Mei, J., Zhang, H., Wang, C., Feng, Y. and Chen, C. (2022) Knowledge Distillation with the Reused Teacher Classifier. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 11923-11932. https://doi.org/10.1109/cvpr52688.2022.01163

[20] Tian, Y., Krishnan, D. and Isola P. (2019) Contrastive Representation Distillation. arXiv: 1910.10699. https://doi.org/10.48550/arXiv.1910.10699

[21] Zagoruyko, S. and Komodakis, N. (2016) Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. arXiv: 1612.03928. https://doi.org/10.48550/arXiv.1612.03928

[22] Gou, J., Yu, B., Maybank, S.J. and Tao, D. (2021) Knowledge Distillation: A Survey. *International Journal of Computer Vision*, **129**, 1789-1819. https://doi.org/10.1007/s11263-021-01453-z

[23] Yang, Z., Li, Z., Shao, M., Shi, D., Yuan, Z. and Yuan, C. (2022) Masked Generative Distillation. *Computer Vision—ECCV* 2022, Tel Aviv, 23-27 October 2022, 53-69. https://doi.org/10.1007/978-3-031-20083-0_4

[24] Park, W., Kim, D., Lu, Y. and Cho, M. (2019) Relational Knowledge Distillation. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 3962-3971. https://doi.org/10.1109/cvpr.2019.00409

[25] Tung, F. and Mori, G. (2019) Similarity-Preserving Knowledge Distillation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, 27 October-2 November 2019, 1365-1374. https://doi.org/10.1109/iccv.2019.00145

[26] Song, J., Chen, Y., Ye, J. and Song, M. (2022) Spot-Adaptive Knowledge Distillation. *IEEE Transactions on Image Processing*, **31**, 3359-3370. https://doi.org/10.1109/tip.2022.3170728

[27] Guo, Z., Yan, H., Li, H. and Lin, X. (2023) Class Attention Transfer Based Knowledge Distillation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 11868-11877. https://doi.org/10.1109/cvpr52729.2023.01142

[28] Gao, L. and Gao, H. (2023) Feature Decoupled Knowledge Distillation via Spatial Pyramid Pooling. *Computer Vision—ACCV* 2022, Macao, 4-8 December 2022, 732-745. https://doi.org/10.1007/978-3-031-26351-4_44

[29] Krizhevsky, A. and Hinton, G. (2009) Learning Multiple Layers of Features from Tiny Images. Technical Report, University of Toronto, Toronto.

[30] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., *et al*. (2015) ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, **115**, 211-252. https://doi.org/10.1007/s11263-015-0816-y

[31] Zhang, X., Zhou, X., Lin, M. and Sun, J. (2018) Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6848-6856. https://doi.org/10.1109/cvpr.2018.00716

[32] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv: 1409.1556. https://doi.org/10.48550/arXiv.1409.1556

[33] Zagoruyko, S. and Komodakis, N. (2016) Wide Residual Networks. In: Wilson, R.C., Hancock, E.R. and Smith, W.A.P., Eds., *Proceedings of the British Machine Vision Conference* 2016, BMVA Press. https://doi.org/10.5244/c.30.87

[34] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., *et al*. (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv: 1704.04861. https://doi.org/10.48550/arXiv.1704.04861