控制与决策
Control and Decision

基于自监督对抗学习的多尺度知识蒸馏方法

张建, 梁兴柱, 张康, 林玉娥, 夏晨星

引用本文:
张建, 梁兴柱, 张康, 等. 基于自监督对抗学习的多尺度知识蒸馏方法[J]. 控制与决策, 2025, 40(3): 880-888.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2024.0290

您可能感兴趣的其他文章

Articles you may be interested in

输入约束不确定系统的点对点迭代学习控制与优化

Point-to-point iterative learning control and optimization for uncertain systems with constrained input

控制与决策. 2021, 36(6): 1435-1441 https://doi.org/10.13195/j.kzyjc.2019.0908

基于条件生成对抗网络的不平衡学习研究

Research on imbalanced learning based on conditional generative adversarial networks

控制与决策. 2021, 36(3): 619-628 https://doi.org/10.13195/j.kzyjc.2019.0522

基于深度强化学习与迭代贪婪的流水车间调度优化

Scheduling optimization for flow-shop based on deep reinforcement learning and iterative greedy method

控制与决策. 2021, 36(11): 2609-2617 https://doi.org/10.13195/j.kzyjc.2020.0608

MADDPG算法经验优先抽取机制

Multi-agent deep deterministic policy gradient algorithm via prioritized experience selected method

控制与决策. 2021, 36(1): 68-74 https://doi.org/10.13195/j.kzyjc.2019.0834

结合注意力机制的循环神经网络复述识别模型

Recurrent neural networks based paraphrase identification model combined with attention mechanism

控制与决策. 2021, 36(1): 152-158 https://doi.org/10.13195/j.kzyjc.2019.0638

基于自监督对抗学习的多尺度知识蒸馏方法

张 建[1]，梁兴柱[1†]，张 康[2]，林玉娥[1]，夏晨星[1]

(1. 安徽理工大学 计算机科学与工程学院，安徽 淮南 232001；2. 西安电子科技大学 微电子学院，西安 710126)

摘 要: 针对离线知识蒸馏中因教师与学生之间规模差距过大, 知识难以有效传递，导致学生性能不佳的问题, 提出一种基于自监督对抗学习的多尺度知识蒸馏方法 (SAMKD), 并利用自监督和对抗学习进一步开发中间多尺度特征与网络末端输出特征 logits 的潜力. 首先, 引入多角度几何变换图像监督网络学习; 然后, 设计多分支辅助网络提取主干网络的多尺度特征, 进而获得更多监督信息; 最后, 利用对抗学习的二元博弈思想进行多阶段对抗训练, 将多层次的知识通过蒸馏方法充分传递. 在 3 个具有挑战性的公开数据集 CIFAR-10、CIFAR-100 和 Tiny-ImageNet 上进行广泛评估, 实验结果表明所提出方法相较其他先进知识蒸馏方法具有强大的竞争力.

关键词: 知识蒸馏；自监督学习；对抗学习；多尺度特征

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.0290

引用格式: 张建, 梁兴柱, 张康, 等. 基于自监督对抗学习的多尺度知识蒸馏方法 [J]. 控制与决策, 2025, 40(3): 880-888.

Multi-scale knowledge distillation method based on self-supervised adversarial learning

ZHANG Jian[1]，*LIANG Xing-zhu*[1†]，*ZHANG Kang*[2]，*LIN Yu-e*[1]，*XIA Chen-xing*[1]

(1. School of Computer Science and Engineering，Anhui University of Science and Technology，Huainan 232001，China；2. School of Microelectronics，Xidian University，Xi'an 710126，China)

Abstract: To address the challenge of ineffective knowledge transfer and poor student performance in offline knowledge distillation due to the significant scale gap between teachers and students, a multi-scale knowledge distillation method based on self-supervised adversarial learning (SAMKD) is proposed. This method leverages self-supervision and adversarial learning to further develop the potential of intermediate multi-scale features and network logits. Firstly, the paper introduces supervised network learning using multi-angle geometrically transformed images. Then, it designs a multi-branch auxiliary network to extract multi-scale features from the backbone network, thereby enhancing supervisory information. Finally, it employs a binary adversarial training approach inspired by adversarial learning for multi-stage adversarial training, effectively facilitating comprehensive knowledge transfer across multiple levels through distillation. Extensive evaluations on three challenging public datasets, CIFAR-10, CIFAR-100, and Tiny-ImageNet, demonstrate that the proposed method exhibits robust competitiveness and outperforms other state-of-the-art knowledge distillation methods.

Keywords: knowledge distillation；self-supervised learning；adversarial learning；multi-scale features

0 引 言

近年来, 深度神经网络得到了飞速发展, 这也成功地促进了计算机视觉领域的进步, 在图像分类[1]、目标检测[2] 和对抗学习[3] 等领域中得到广泛应用. 然而, 深度神经网络模型被设计得深而复杂, 产生高计算和存储成本, 进而导致这些高性能的大规模模型很难应用在计算资源有限的移动或嵌入式设备上.

为了解决模型过大难以压缩的问题, 许多研究工作旨在构建更小规模但保持精度的深度网络模型. 当前, 主流的模型压缩方法分别为剪枝[4]、量化[5]、轻量级网络架构设计[6] 和知识蒸馏[7](knowledge distillation, KD). 其中, 知识蒸馏提供了一种直接高

收稿日期: 2024-03-20；**录用日期:** 2024-07-12.

基金项目: 安徽理工大学医学专项培育项目 (YZ2023H2C005)；国家自然科学基金项目 (62102003)；安徽理工大学环境友好材料与职业健康研究院研发专项基金项目 (ALW2021YF04)；安徽理工大学研究生创新基金项目 (2023cx2139).

责任编委: 曹进德.

†通信作者. E-mail: xzliang@aust.edu.cn.

效且极具潜力的范例[8], 其主要思想是将性能更优秀的大规模教师模型中的知识转移到小规模的学生模型中, 使学生模型逼近甚至超越教师模型性能[9].

知识蒸馏方法可大致分为两类, 一类是基于 logits 的知识蒸馏方法, 另一种是基于特征的知识蒸馏方法. 前者主要通过最小化教师与学生的预测逻辑 logits 之间的 KL 散度 (KL-Divergence) 来传递知识[10-13]. 然而, logits 蒸馏方法并未发挥完整的教师模型作用, 未充分利用中间层的教师特征. 另外, 一些研究方法尝试引导学生模型模仿中间层多样化的深层特征, 如中间表示、注意力图等. 这些方法主要集中在从中间层的深层特征中提取知识, 称为基于特征的知识蒸馏[14-17]. 虽然特征蒸馏方法取得了不错的性能, 但在多尺度特征的利用上还不够充分, 导致损失了较多的特征信息.

现有的方法侧重于教师模型以输出 logits 或中间特征的形式传递给学生模型, 忽略了教师与学生之间差距过大导致学生性能不佳的问题. 特别是教师与学生之间的规模差距很大时[18-19], 学生模仿教师能力有限, 因此学生不一定能达到更强的性能.

为此, 本文提出一种基于自监督对抗学习的多尺度知识蒸馏方法 (multi-scale knowledge distillation based on self-supervised adversarial learning, SAMKD), 缩小教师模型与学生模型之间差距, 提高蒸馏性能. 具体而言, 首先进行多角度监督图像处理, 对输入图像进行不同角度的旋转, 并适配相应的扩展标签, 利用多角度几何变换图像监督网络模型学习; 其次, 设计多分支辅助网络提取主干网络的多尺度特征, 形成多分支辅助网络监督 (multi-branch auxiliary network supervision, MANS); 最后, 利用对抗网络二元博弈的思想进行多阶段对抗训练 (multi-stage adversarial training, MAT), 教师与学生角色互换对抗训练, 强有力地将多层次的知识传递给学生模型. 本文提出的 SAMKD在 3 个公开标准数据集上进行广泛评估, 实验结果展示了 SAMKD 的出色性能.

1　方　法

1.1　问题定义

传统知识蒸馏的关键思想是通过训练一个小容量的学生模型来模仿教师的输出, 如图 1 所示. 为了实现这一目标, 通常利用交叉熵损失和 KL 散度来优化学生训练. 给定数据集$D=\{x,y\}_M^N$, 其中包含 M 个类别的 N 个样本, x 为输入样本, y 为相应的真实标签. 输入图像由神经网络进行特征提取,最终经过线性分类器得到 logits 输出 z 如下所示:

$$z=\{z_1,z_2,\dots,z_i,\dots,z_M\}. \tag{1}$$

其中: z_i为第 i 个类别的 logit 输出, M 为类别数. 在训练过程中, 最后使用 softmax 函数将 logits 输出转换为 0 至 1 之间的预测概率分布. 因此定义最终预测概率分布 p 如下所示:

$$p=\{p_1,p_2,\dots,p_i,\dots,p_M\}, \tag{2}$$

$$p_i^T=\exp(z_i/T)\Big/\sum_{j=1}^{M}\exp(z_j/T). \tag{3}$$

其中: p_i^T为第 i 个类别的输出概率; T 为温度系数, 控制输出概率的软化程度. 蒸馏过程中, 学生模型由教师预测的软化概率分布进行指导, 并由自身真实标签约束, 最终损失函数由 KL 散度与交叉熵损失组成. 学生网络通过最小化如下方式进行训练:

$$\mathcal{L}_{\mathrm{KD}}^{s}=\mathcal{L}_{\mathrm{CE}}(p_s^T,y)+T^2\mathcal{L}_{\mathrm{KL}}(p_s^T,p_t^T). \tag{4}$$

其中: s 和 t 分别为学生和老师, 交叉熵中蒸馏温度 T 为 1. KL 散度用来对齐教师与学生之间软化的概率输出, 温度$T>1$. 由于软化分布产生的梯度尺度为 $1/T^2$, 需要乘T^2补偿由$\mathcal{L}_{\mathrm{KL}}$中的温度软化引起相关值减少的权重.

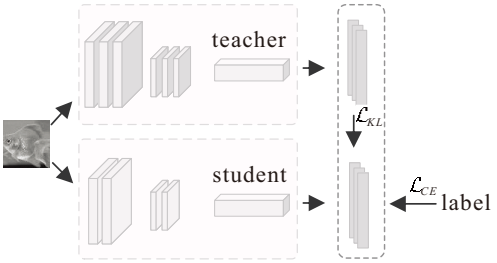


图1　经典知识蒸馏

1.2　模型整体架构

基于自监督对抗学习的多尺度知识蒸馏方法框架如图 2 所示. 首先通过双重自监督学习预训练小容量教师网络, 采用双重自监督方法辅助教师网络的优化, 多角度自监督标签和原始标签共同利用交叉熵损失约束教师网络; 然后, 进入第 1 阶段对抗训练过程, 小容量教师网络指导大容量学生网络学习, 利用 KL 散度进行教师与学生之间的知识传递, 大容量学生网络需要交叉熵损失来保证预测的准确性, 这种容量学生网络称为对抗教师网络; 最后, 对抗教师网络会作为最终教师指导目标学生网络模型学习, 称为第 2 阶段对抗训练过程. 两阶段训练过程中, 对抗博弈思想实现教师与学生之间的角色互换, 通过双重自监督构建辅助任务, 为网络学习提供更多有意义的额外特征.

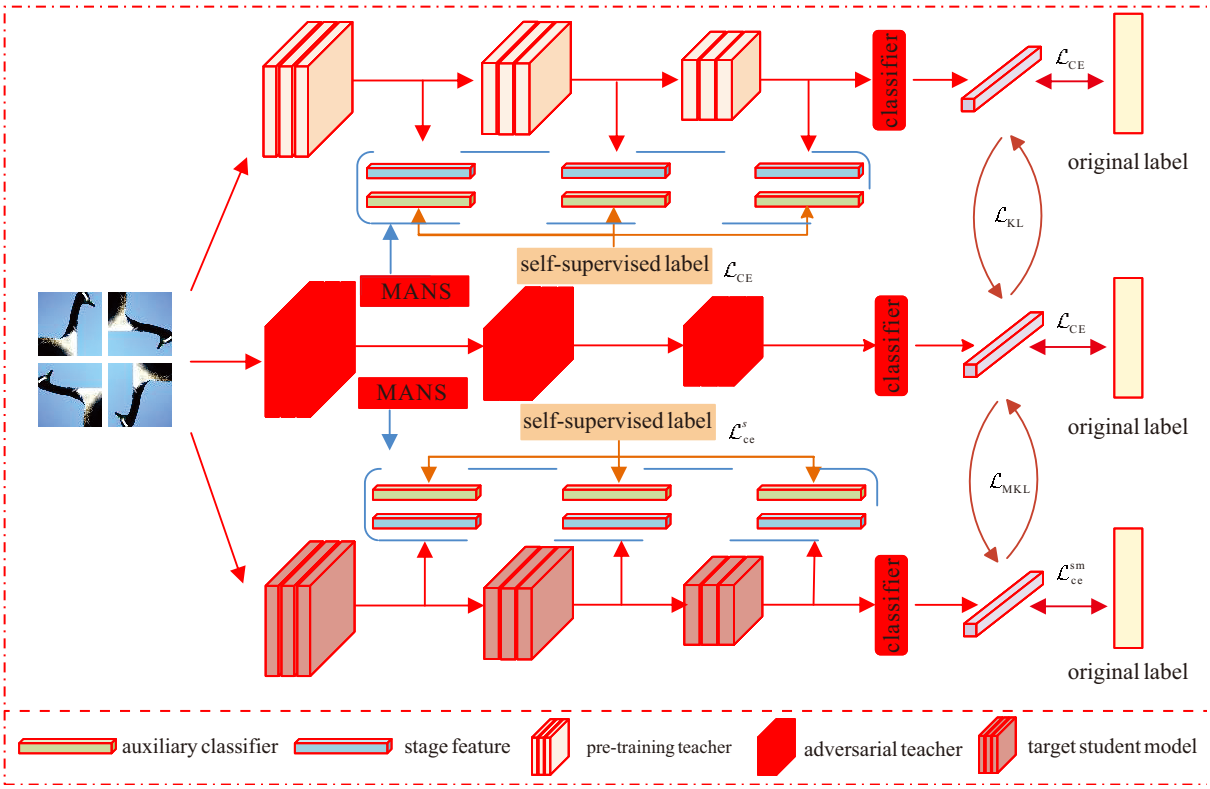


图2 SAMKD 模型框架

1.3 双重自监督学习

1.3.1 辅助分支网络自监督

传统分类骨干神经网络通常分为两部分, 一个是特征提取部分, 另一个是最终的分类器部分. 其中特征提取部分往往由浅入深被分为多个特征提取块, 这些特征提取块会依次产生不同尺度的特征图. 图像在经过特征提取部分后会得到一个语义丰富的精细特征图, 随后进入分类器得到最终的预测结果. 标准的分类网络由于其固定的卷积核大小和步幅, 图像感受野较小, 得到的特征信息有限, 不利于知识蒸馏过程中知识的传递. 为了使单个模型从多个网络中学习更多的特征知识, 受 TSKD[20] 多出口网络的启发, 设计了多分支的辅助分支网络, 其结构如图 2 中 MANS 所示. 该网络在传统的神经网络各个特征提取阶段设计辅助分支并逐级提取特征, 因此, 主干网络受到每一个辅助分支网络额外监督指导.

1.3.2 多角度旋转图像自监督

经典的自监督学习工作一般通过设计辅助任务来进一步挖掘数据自身的表征特性作为监督信息, 如几何变换与多任务学习等. 受其中几何变换[21]的启发, 提出一种简单且高效实用的自监督任务, 即多角度旋转图像自监督. 将几何变换定义为 90°倍数旋转, 分别为 0°、90°、180°和 270°. 在对数据集采用标准的数据预处理和增强技术后, 通过这种简单的几何变换, 将单个图像扩展成 4 倍作为输入, 从而使网络学习到更多有意义的特征. 具体而言, 给定一个训练模型$F(\cdot)$, 由原始主干网络与辅助分支网络构成; 一组 K 个离散几何变换$V=\{Q(\cdot|y)\}_{y}^{K}$, 原始图像 x 经过 90° 倍数旋转变换后变为图像 $X^{v}=\{Q(x|y)\}_{y=1}^{K}$. 经过模型$F(\cdot)$的特征提取分别得到主干网络输出 logits 概率分布F_{core}和辅助分支网络输出 logits 概率分布F_{aux}为

$$F_{\text{core}}, F_{\text{aux}} = F(Q(x|y))_{y=1}^{K}. \tag{5}$$

其中$Q(\cdot|y)$为将带有标签 y 的几何变换应用于图像 x 的运算操作, 产生几何变换后的图像.

1.4 多阶段对抗训练

1.4.1 第 1 阶段对抗训练

第 1 阶段对抗训练是转换教师网络与学生网络角色, 从而获得对抗教师网络. 本文采用逆向对抗思想, 目的是训练出拟合小容量教师网络的过渡大容量学生网络. 这种过渡大容量学生网络称为对抗教师网络. 具体而言, 结构简单、参数少的网络作为教师$F^{t}(\cdot)$, 相对应的结构复杂、参数多的网络作为学生$F^{s}(\cdot)$. 将多角度旋转图像X^{v}作为输入, 教师与学生模型分别对其进行特征提取, 最终分别得到各自的主干网络输出 logits 概率分布F_{core}与辅助分支网络输出 logits 概率分布F_{aux}. 该过程可描述如下:

$$F^{t}_{\text{core}}, F^{t}_{\text{aux}} = F^{t}(Q(x|y))_{y=1}^{K}, \tag{6}$$

$$F^{s}_{\text{core}}, F^{s}_{\text{aux}} = F^{s}(Q(x|y))_{y=1}^{K}. \tag{7}$$

其中:F^{t}_{core}, F^{t}_{aux}分别为教师模型主干网络输出

logits 概率分布和辅助分支网络输出 logits 概率分布; F^s_{core}, F^s_{aux} 分别为学生模型主干网络输出 logits 概率分布和辅助分支网络输出 logits 概率分布.

通过最小化预训练好的教师模型与学生模型之间的 KL 散度来得到对抗教师模型, 使得对抗教师模型在保持精度的情况下更好地吸收目标学生模型的知识. 另外, 将 F^s_{core} 进行 1/4 切片变换成原始 0°旋转图像的输出 logits 概率分布 F^*. F^* 对应真实图像标签为 y^*, 两者的交叉熵作为约束学生模型训练的另一部分. 对抗教师的训练过程定义如下:

$$F^* = \zeta(F^s_{\text{core}}), \tag{8}$$

$$\text{Min}\Big(\sum_{i=1}^{N}\mathcal{L}_{\text{CE}}(F^*_i, y^*_i) + \sum_{i=1}^{N}\mathcal{L}_{\text{KL}}((\hat{F}^s_{\text{core}})_i, (\hat{F}^t_{\text{core}})_i)\Big). \tag{9}$$

其中: ζ 为 1/4 张量切片变换操作, y^* 为 1/4 切片原始对应的真实标签,最小化 F^* 与 y^* 交叉熵损失来保证学生模型的准确率; $\hat{F}^s_{\text{core}}$ 与 $\hat{F}^t_{\text{core}}$ 分别为学生模型主干网络软化输出和教师模型主干网络软化输出, 同时最小化 $\hat{F}^s_{\text{core}}$ 与 $\hat{F}^t_{\text{core}}$ 的 KL 散度来促进学生模型 $F^s(\cdot)$ 与教师模型 $F^t(\cdot)$ 之间的知识传递.

1.4.2　第 2 阶段对抗训练

目前知识蒸馏中学生性能不佳, 其中一个关键原因是教师与学生之间规模差距过大, 这会导致无法有效地将教师知识迁移至学生模型. 为此, 通过第 1 阶段对抗训练得到更适合学生学习的对抗教师网络. 第 2 阶段对抗训练是利用第 1 阶段对抗训练产生的对抗教师网络作为最终指导目标学生的教师. 对抗教师网络的表现与正常训练的教师几乎相同, 但学生在训练过程中更容易提取知识. 通过第 1 阶段初步的对抗训练, 对抗教师网络包含了大量从目标学生学到的语义信息, 跨越了教师与学生之间的代沟. 具体而言, 首先, 冻结对抗教师网络参数, 与目标学生网络训练时分别得到教师与学生主干网络对多角度监督图像的软化概率分布 $\hat{F}^t_{\text{core}}$ 与 $\hat{F}^s_{\text{core}}$, 并利用 KL 散度计算它们的改进相似度 $\mathcal{L}_{\text{MKL}}$. 另外, 目标学生网络的各个分支辅助网络对主干网络产生的多尺度特征进行进一步特征提取, 得到多角度监督图像的输出 logits 概率分布 F^s_{aux}. 给定多角度监督图像真实标签 c, 利用交叉损失 $\mathcal{L}^s_{\text{ce}}$ 约束辅助分支网络的优化方向. 最后, 目标学生网络主干网络需要预测原始未旋转图像的类别, 并通过计算该 logits 输出概率分布 F^* 与原始真实标签 y^* 的交叉熵损失 $\mathcal{L}^{\text{sm}}_{\text{ce}}$ 来进一步训练学生模型. 上述训练过程可描述如下:

$$\mathcal{L}_{\text{MKL}} = \frac{1}{N}\sum_{i=1}^{N}\text{KL}((\hat{F}^s_{\text{core}})_i, (\hat{F}^t_{\text{core}})_i), \tag{10}$$

$$\mathcal{L}^s_{\text{ce}} = -\sum_{i=1}^{N} c_i \log((F^s_{\text{aux}})_i), \tag{11}$$

$$\mathcal{L}^{\text{sm}}_{\text{ce}} = -\sum_{i=1}^{N} y^*_i \log(F^*_i). \tag{12}$$

其中: N 为数据集样本数, i 为第 i 个图片样本.

1.5　损失函数优化

交叉熵损失与 KL 散度广泛用于 KD 分类任务. 交叉熵约束网络预测准确率, KL 散度作为蒸馏损失, 用作对齐教师模型向学生模型知识迁移过程中的预测分布. 本文根据 SAMKD 框架以及训练策略, 对辅助分支网络、主干网络和多阶段对抗训练损失函数作出相对有效的调整组合和优化. 模型整个过程中两个对抗训练阶段的损失函数记为 $\mathcal{L}^{s1}_{\text{adv}}$ 和 $\mathcal{L}^{s2}_{\text{adv}}$. 第 1 阶段对抗训练目标是获得适合教学目标学生网络的对抗教师网络, 该对抗教师网络包含了丰富且易于目标学生网络学习的知识. 因此, 第 1 阶段对抗训练机制损失函数 $\mathcal{L}^{s1}_{\text{adv}}$ 表示如下:

$$\mathcal{L}^{s1}_{\text{adv}} = \sum_{i=1}^{N}\mathcal{L}_{\text{CE}}(F^*_i, y^*_i) + \sum_{i=1}^{N}\mathcal{L}_{\text{KL}}((\hat{F}^s_{\text{core}})_i, (\hat{F}^t_{\text{core}})_i). \tag{13}$$

第 2 阶段对抗训练是利用优秀的对抗教师网络指导目标学生网络学习来最终构建速度快、计算量小同时不失精度的轻量级压缩学生网络模型. 此对抗训练机制损失函数 $\mathcal{L}^{s2}_{\text{adv}}$ 定义为

$$\mathcal{L}^{s2}_{\text{adv}} = \mathcal{L}^{\text{sm}}_{\text{ce}} + \alpha\mathcal{L}_{\text{MKL}} + \beta\mathcal{L}^s_{\text{ce}}, \tag{14}$$

其中 α 与 β 为权重系数, 用来平衡调控模型各部分的影响. 结合上述两个阶段损失, 模型最终总损失为

$$\mathcal{L}_{\text{total}} = \mathcal{L}^{s1}_{\text{adv}} + \mathcal{L}^{s2}_{\text{adv}}. \tag{15}$$

1.6　整体算法流程

相较于其他知识蒸馏方法, SAMKD 方法的改进主要在于采用了双重自监督学习以及多阶段对抗训练, 具体算法流程如下.

step1: 标准的数据预处理和增强技术, 包括边缘填充、随机裁剪和翻转等操作.

step2: 多角度图像旋转处理

train_batch = torch. rot90(batch, *K*, (*H*, *W*)).

step3: 第 1 阶段对抗训练:

with torch.no_grad( ):

output_t, aux_logits = model_t(train_batch)

output_s,aux_logits = model(train_batch)

loss1 = CEloss + KLloss

step4: 第 2 阶段对抗训练:

```
with torch.no_grad( ):
output_a, aux_logits = a_model(train_batch)
output_s,aux_logits = s_model(train_batch)
loss2 =SM_CEloss + MKLloss + S_CEloss
```

step5: 直到对抗训练模型收敛.

2 实验结果与分析

2.1 数据集与网络架构

本节介绍了实验中 3 个公开数据集 CIFAR-10、CIFAR-100、Tiny-ImageNet 以及相关网络架构.

CIFAR-10[22]: 包含 60 000 张彩色图像, 分为 50 000 张训练彩色图像和 10 000 张测试彩色图像, 每张图像的尺寸为 32 × 32 像素, 分为 10 个类别. CIFAR-100[22]: 包含 60 000 张彩色图像, 分为 50 000 张训练彩色图像和 10 000 张测试彩色图像, 每张图像的尺寸为 32 × 32 像素, 分为 100 个类别. Tiny-ImageNet[23]: 包含 100 000 张训练彩色图像, 10 000 张测试彩色图像, 10 000 张验证彩色图像, 每张图像的尺寸为 64 × 64 像素, 分为 200 个类别.

为了验证本文所提方法在各种网络架构上的通用性, 采用 3 个主流网络 ResNet[24]、WRN[25] 和 VGG[26] 系列的网络架构. 另外, 由于 CIFAR-10 是公开标准小数据集, 包含特征较少, 对抗教师网络的训练使用标准的骨干网络, 以此来获得较好的性能. 在 CIFAR-100 与 Tiny-ImageNet 上, 将使用蒸馏领域公认通用的骨干网络进行实验.

2.2 实验设置与环境配置

在整个实验过程中, 对这 3 个数据集采用标准的数据预处理和增强技术, 包括边缘填充、随机裁剪和翻转等操作. 另外, 所有图像均按蒸馏领域公认的通道均值和标准差进行归一化. 对于 3 个数据集, 将批量大小设置为 128, 使用动量为 0.9 与权重衰减为 5e-4 的 SGD 作为优化器, 蒸馏温度 T 设置为 4. 对于 CIFAR-10 数据集, 初始学习率设置为 0.1, 共 160 轮. 学习率在 80 和 100 个轮次衰减 0.1. 在 CIFAR-100 数据集中, 学习率初始化为 0.05, 学习率在第 150、180 和 210 个轮次衰减 0.1, 共 240个轮次. 对于 Tiny-ImageNet 数据集, 初始学习率设为 0.1, 在共 200 个轮次训练中, 学习率在第 60、120 和 160 轮次处衰减 0.2. 高蒸馏损失值对于知识传递非常重要, 因此 α需要相对较大权重值. 该系数的确定参考了 KD 与 DKD 方案. 通过相关实验分析, 取 $\alpha = 1.5, \beta = 1$. 尽管该权重比例并不是最佳的, 但它们已经提供了局部最优和先进的结果. 将所提出方法在 PyTorch 环境中实现, 所有实验均在 Linux 操作系统下的 NVIDIA RTX 3090 上进行. 实验结果以 3 次实验的平均值形式报告.

2.3 实验结果分析

首先, 将 SAMKD 与经典 KD[7] 比较, 初步验证 SAMKD 的有效性以及带来的巨大性能提升. 然后, 与其他具有代表性的离线知识蒸馏方法进行比较以进一步体现 SAMKD 优势, 这些先进的方法包括 FitNet[14]、AT[15]、FT[27]、AB[28]、PKT[29]、CC[30]、CRD[16]、RKD[31]、VID[17]、DKD[11]、SP[32]、SRRL[10]、NKD[12]、NormKD[13]、NORM[33]、CTKD[34].

2.3.1 与经典蒸馏对比

本文提出的 SAMKD 与经典 KD 相比较, 在多种网络架构组合中性能有显著提升. 表 1 ~ 表 3 展示了基于 9 个不同的师生对在 3 个数据集上的 top-1 分类精度, 并且在某些情况下改进较为显著. 同时这 3 个表格数据还显示了经典 KD 与 SAMKD 带来的提升. 可以看出, 最初新兴模型压缩方法相较经典 KD 毫无疑问展示了它的实用性、效率以及相当大的潜力,它形成了一条非常通用的模型压缩路线, 几乎适用于所有的网络架构. 本文提出的 SAMKD 在各种网络架构的师生对中均优于经典 KD, 验证了 SAMKD 的有效性. 自监督与对抗学习策略结合知识蒸馏表现优越.

SAMKD 在 CIFAR-10 上相比较经典 KD 整体上平均提高约 1.5 %. 在更具挑战的数据集 CIFAR-100 和 Tiny-ImageNet 上, SAMKD 同样表现优秀. 在 CIFAR-100 上, STAKD 整体平均提升约 1.65 %, 同样地, 在更具挑战性的 Tiny-ImageNet 上, 仍然保持整体 2.14 % 的提升.

2.3.2 与代表性蒸馏方法对比

表 4 为所提出的 SAMKD 与其他代表性先进方法在 CIFAR-100 上的实验结果. 从不同架构的教师

表1 CIFAR-100 验证的 Top-1 准确率 %

Teacher	ResNet-110	ResNet-32x4	ResNet-110	ResNet-56	VGG-13	WRN-40-2	WRN-40-2
	74.31	79.42	74.31	72.34	74.64	75.61	75.61
Adv_Teacher	ResNet-110	ResNet-32x4	ResNet-110	ResNet-56	VGG-13	WRN-40-2	WRN-40-2
	74.45	79.17	74.15	72.73	74.26	76.19	76.34
Student	ResNet-20	ResNet-8x4	ResNet-32	ResNet-20	VGG-8	WRN-16-2	WRN-40-1
	69.06	72.50	71.14	69.06	70.36	73.26	71.98
KD	70.67	74.42	73.08	70.66	72.98	74.92	73.54
SAMKD	**72.43**	**76.10**	**74.45**	**72.30**	**74.49**	**76.54**	**75.49**

表2　CIFAR-10 验证的 Top-1 准确率　%

Teacher	ResNet-18	ResNet-18	Vgg-13
	95.13	95.13	94.09
Adv_Teacher	ResNet-18	ResNet-18	Vgg-13
	95.42	95.09	93.68
Student	ResNet-20	ResNet-32	Vgg-8
	92.19	93.04	91.68
KD	92.49	93.31	92.63
SAMKD	**94.12**	**95.30**	**93.48**

表3　Tiny-ImageNet 验证的 Top-1 准确率　%

Teacher	ResNet-56	VGG-13
	57.93	59.26
Adv_Teacher	ResNet-56	VGG-13
	57.39	59.29
Student	ResNet-20	VGG-8
	52.48	54.30
KD	52.83	57.70
SAMKD	**54.66**	**60.14**

和学生组合来全面评估这些方法可以看到, SAMKD 整体性能优于其他方法. 与一些先进的知识蒸馏方法相比其性能同样具有竞争力. 例如, SAMKD 在 ResNet-110/ResNet-32 师生架构上的表现优于一些最先进的方法, 比 DKD 和 NormKD 分别提高了 0.34 % 与 0.54 %. 另外, 在 VGG-13/VGG-8 和 WRN-40-2/WRN-40-1 的架构组合中, 本文提出的 SAMKD 仍然具有一定优势. 同样地, 在 ResNet-56/ResNet-20 模型上, SAMKD 展示了最佳性能, 超过 NORM 和 CTKD 等方法.

此外, 为了更准确和直观地将训练优化过程表示出来, 在 CIFAR-100 上使用 ResNet-110/ResNet-32 模型以展示 Top-1 精确度可视化的效果. 从图 3 可以看到 3 种方法的准确率优化趋势. 在 0 ～30 轮, KD 与 DKD 准确率差距相差不大, 不过 SAMKD 仍保持一定的优势, 这表明即使在训练轮次较少的情况下, SAMKD 仍有较强的接受知识的能力. 30 ～150 轮之间, SAMKD 已经明显优于其他方法. 在 150 ～240 轮次时, 特别是在学习率衰减关键轮次之后, SAMKD 仍保持相当大的竞争力. 相比较其他方法, SAMKD 整体波动平缓且优势明显, 意味着在知识传递过程中能够学习更丰富的知识并优化更稳定的目标学生模型, 最终收敛达到最佳性能.

表4　不同蒸馏方法在 CIFAR-100 上 Top-1 准确率　%

distillation manner	teacher student	ResNet-56 72.34 ResNet-20 69.06	ResNet-110 74.31 ResNet-32 71.14	ResNet-32x4 79.42 ResNet-8x4 72.50	WRN-40-2 75.61 WRN-16-2 73.26	WRN-40-2 75.61 WRN-40-1 71.98	VGG-13 74.64 VGG-8 70.36
FITNET	ICLR 2015	69.21	71.06	73.50	73.58	72.24	71.02
AT	ICLR 2017	70.55	72.31	73.44	74.08	72.77	71.43
FT	NeurIPS 2018	69.84	72.37	72.86	73.25	71.59	70.58
AB	AAAI 2019	69.47	70.98	73.17	72.50	72.38	70.94
PKT	ECCV 2018	70.34	72.61	73.64	74.54	73.45	72.88
CC	ICCV 2019	69.63	71.48	72.97	73.56	72.21	70.71
CRD	ICLR 2020	71.16	73.48	75.51	75.48	74.14	73.94
RKD	CVPR 2019	69.61	71.82	71.90	73.35	72.22	71.48
VID	CVPR 2019	70.38	72.63	73.09	74.11	73.30	71.23
SP	ICCV 2019	69.67	72.69	72.94	73.83	72.43	72.68
SRRL	ICLR 2021	71.57	73.48	75.39	75.69	74.64	74.04
DKD	CVPR 2022	71.97	74.11	76.32	76.24	74.81	74.41
NKD	ICCV 2023	70.14	73.37	75.34	75.73	75.21	74.46
NormKD	arXiv 2023	71.40	73.91	76.57	76.40	74.84	74.45
NORM	ICLR 2023	71.61	73.95	**76.98**	76.26	75.42	74.46
CTKD	AAAI 2023	71.19	73.52	76.16	75.45	73.93	73.52
SAMKD	—	**72.30**	**74.45**	76.10	**76.54**	**75.49**	**74.49**

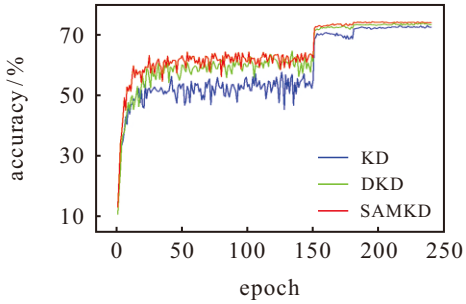


图3　不同蒸馏方法收敛对比

2.3.3　时间成本与性能分析

为了进一步体现所提方法的时间成本消耗, 在 CIFAR-100 数据集上选择 4 种经典蒸馏方法作为对比对象, 对比结果如表 5 所示. 从表 5 中可以看出, 为了提高最终目标学生模型性能, SAMKD 主要牺牲了一些时间成本. 教师与学生之间规模差距过大时, 学生模仿教师能力有限. 因此, 哪怕花费大量时间成本去训练一个超大规模的教师, 学生也不一定

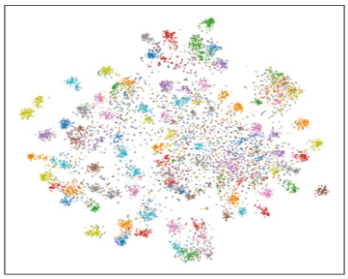
能达到更强的性能, 这就产生了时间与性能都无法兼顾的问题. 因此, 在时间成本与性能之间做出平衡, 花费一些时间成本去达到更优秀的性能是值得的.

表5 时间成本与性能分析

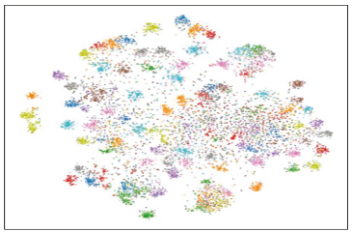
method	ResNet-32	KD	AT	VID	DKD	**SAMKD**
accuracy	71.14	73.08	72.31	72.63	74.11	**74.45**
time(hours)	**0.60**	2.39	2.55	2.87	2.26	9.58

2.3.4 定性分析

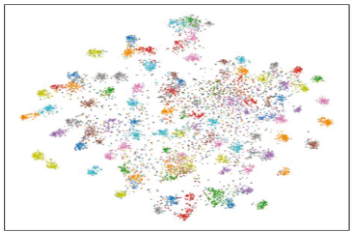
本文从两个角度呈现可视化模型的高维特征, 以直观地探究高维特征分布及其带来的影响. 在CIFAR-100 上展示 T-SNE[35] 与相关矩阵的可视化, 设置 ResNet-110 作为教师, ResNet-32 作为学生. T-SNE 技术将高维特征转化为二维特征, 形成一系列聚集的团簇, 每一个团簇代表一个类别. 如图 4 所示, SAMKD 特征表示比其他蒸馏方法更具可分离性,



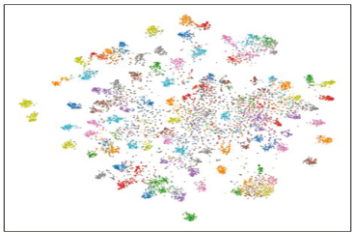
(a) KD



(b) AT



(c) DKD

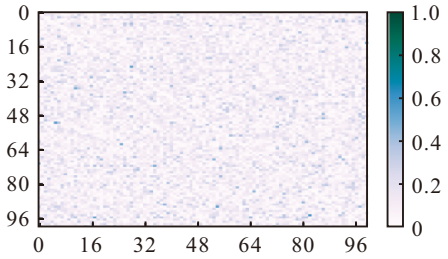


(d) SAMKD

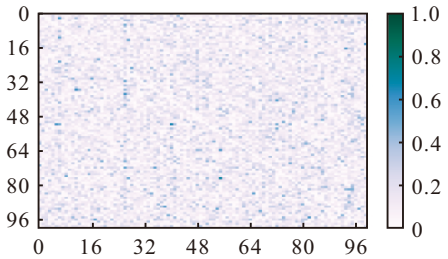
图4 T-SNE 分类散点图

表明 SAMKD 更有利于深度特征的区分.

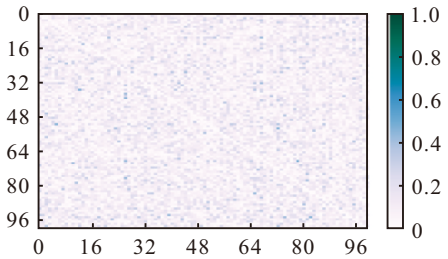
本文还可视化了学生与教师 logits 相关矩阵的差异, 颜色越深意味着学生与老师差异越大, 见图 5. 与其他方法相比, SAMKD 帮助学生输出与老师更相似的 logits, 代表着学生通过蒸馏学习到了更多来自教师的知识, 即获得更好的蒸馏性能. SAMKD 的有效性在这些可视化结果中得到了最直观地展示, 在复杂情况下能获得较高的性能.



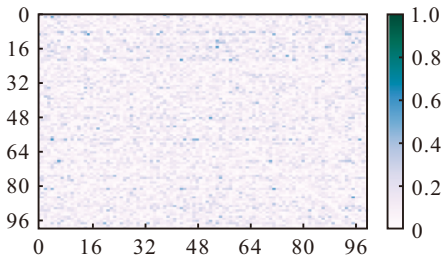
(a) KD



(b) AT



(c) DKD



(d) SAMKD

图5 学生与教师相关矩阵

2.4 消融实验

为了验证所提出的自监督学习与对抗学习在知识蒸馏的有效性, 以原始骨干模型为基线, 对蒸馏散度对齐 MKL、多阶段对抗训练机制 MAT、自监督辅助分支网络 MANS 进行消融实验,结果如表 6 所示. 所有消融实验均在CIFAR-100 数据集上进行, ResNet-110 为教师网络, ResNet-32 为学生网络, 即

最终模型压缩目标网络.

如表 6 所示, 为了更好地体现知识蒸馏方法所带来的性能提升, 将情况 1 列为消融实验的基线, 无蒸馏操作. 可以看出, 情况 1 性能最低, 准确率仅达到 71.14 %, 与其他对照组相比差距较大. 其他对照情况则是在具有蒸馏操作前提下进行的, 即存在 MKL. 情况 2 代表应用蒸馏损失, 相比较情况 1 准确率提升近 2 %. 这种飞跃式性能提升体现了新兴模型压缩方法知识蒸馏的实用性与有效性. 另外, 从情况 3 可以看出, 多阶段对抗训练机制同样发挥了重要作用, 相比较情况 2 性能提升约 0.78 %. 与此同时, 多分支辅助网络监督同样不可或缺, 情况 4 已经验证其有效性. 情况 5 代表本文提出的完整 SAMKD, 当这些关键组成部分和必要策略结合在一起时, 模型性能达到最佳.

表6　消融实验分析表

情况	MKL	MAT	MANS	accuracy
情况1	—	—	—	71.14
情况2	√	—	—	73.08
情况3	√	√	—	73.86
情况4	√	—	√	74.15
情况5	√	√	√	**74.45**

对于α在第 2 阶段对抗训练损失函数的权重占比, 选取不同的α值进行实验, 如图 6 所示. 首先, SAMKD 性能整体表现依然较好, 但仍然有一些较小的波动. 本文发现当$\alpha \in [1.2, 1.6]$时, 模型效果较优. 与此同时, 当α大于 1.6 时, 性能会下降. 推测过大百分比的蒸馏 KL 散度可能会影响整体模型任务损失. 此外, 观察到$\alpha = 1.5$时性能最优. 因此本文选择$\alpha = 1.5$作为最佳选择. 当固定α值, 调整β参数时, 超参数权重为 1 时, 已经可以达到一个相对整体最优, 因此本文选择其值为 1. 该权重比例尽管不能确定为全局最佳, 但根据实验结果已经提供了最先进的结果, 达到了性能相对优秀的局部最优.

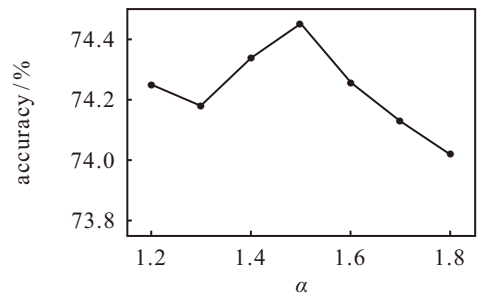


图6　超参数敏感性分析图

3　结　论

本文提出了一种基于自监督对抗学习的多尺度知识蒸馏方法 SAMKD 用于模型压缩. SAMKD 有力缓解了因教师与学生之间规模差距过大导致学生性能不佳的问题, 成功构建了高效的轻量级压缩网络模型. 首先将自监督学习引入, 通过其中简洁的多角度旋转监督图像和辅助分支网络构建, 充分利用多尺度特征, 形成双重自监督来辅助模型优化. 然后, 与对抗学习结合, 获得契合目标学生网络的对抗教师网络, 拉近了教师与学生之间的差距. 在 3 个公开数据集上的实验结果表明, 与其他方法相比, 本文的 SAMKD 取得了较好的性能. 在未来的工作中, 将进一步研究如何提升知识蒸馏的压缩模型的效率, 同时探究与更多领域结合, 以实现更广泛的应用.

参考文献 (References)

[1] Tang Y, Chen Y. Self-knowledge distillation based on dynamic mixed attention[J]. Control and Decision, DOI: 10.13195/j.kzyjc.2024.0036.

[2] Shi Y H, Wang N Y, Guo X J. YOLOV: Making still image object detectors great at video object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(2): 2254-2262.

[3] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]. International Conference on Neural Information Processing Systems. New York, 2014: 2672-2680.

[4] Cai L H, An Z L, Yang C G, et al. Prior gradient mask guided pruning-aware fine-tuning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 140-148.

[5] Gholami A, Kim S, Dong Z, et al. A survey of quantization methods for efficient neural network inference[C]. Low-Power Computer Vision. Boca Raton: Chapman and Hall/CRC, 2022: 291-326.

[6] Cheng Q, Li J, Gao X L, et al. Lightweight method of deep neural network based on deep sparse low rank decomposition[J]. Control and Decision, 2023, 38(3): 751-758.

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

[8] 潘瑞东, 孔维健, 齐洁. 基于预训练模型与知识蒸馏的法律判决预测算法[J]. 控制与决策, 2022, 37(1): 67-76. (Pan R D, Kong W J, Qi J. Legal judgment prediction based on pre-training model and knowledge distillation[J]. Control and Decision, 2022, 37(1): 67-76.)

[9] Liang X Z, Bi F L, Liu W, et al. Trained teacher: Who is good at teaching[J]. Displays, 2023, 80: 102543.

[10] Yang J, Martinez B, Bulat A, et al. Knowledge distillation via softmax regression representation learning[Z]. International Conference on Learning Representations, 2021.

[11] Zhao B R, Cui Q, Song R J, et al. Decoupled knowledge distillation[C]. IEEE/CVF Conference on Computer

Vision and Pattern Recognition. New Orleans, 2022: 11943-11952.

[12] Yang Z D, Zeng A L, Li Z, et al. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels[C]. IEEE/CVF International Conference on Computer Vision. Paris, 2023: 17139-17148.

[13] Chi Z, Zheng T, Li H, et al. Norm KD: Normalized logits for knowledge distillation[J/OL]. 2023, arXiv: 2308.00520.

[14] Adriana R, Nicolas B, Ebrahimi K S, et al. Fitnets: Hints for thin deep nets[M]. San Diego, 2015.

[15] 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. Hilton Vancouver Metrotown, 2017: 1-10.

[16] Tian Y, Krishnan D, Isola P. Contrastive representation distillation[J/OL]. 2019, arXiv: 1910.10699.

[17] Ahn S, Hu S X, Damianou A, et al. Variational information distillation for knowledge transfer[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 9163-9171.

[18] Mirzadeh S I, Farajtabar M, Li A, et al. Improved knowledge distillation via teacher assistant[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 5191-5198.

[19] Yang C L, Xie L X, Su C, et al. Snapshot distillation: Teacher-student optimization in one generation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 2859-2868.

[20] Xu C Y, Gao W J, Li T, et al. Teacher-student collaborative knowledge distillation for image classification[J]. Applied Intelligence, 2023, 53(2): 1997-2009.

[21] Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations[J/OL]. 2018, arXiv: 1803.07728.

[22] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[J]. Handbook of Systemic Autoimmune Diseases, 2009, 1(4): 1-60.

[23] Le Y, Yang X. Tiny imagenet visual recognition challenge[J]. CS 231N, 2015, 7(7): 3.

[24] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[25] Zagoruyko S, Komodakis N. Wide residual networks[J/OL]. 2016, arXiv: 1605.07146.

[26] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]. IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015: 1-9.

[27] Kim J, Park S U, Kwak N. Paraphrasing complex network: Network compression via factor transfer[C]. Advances in Neural Information Processing Systems. Montreal, 2018: 2765-2774.

[28] Heo B, Lee M, Yun S, et al. Knowledge transfer via distillation of activation boundaries formed by hidden neurons[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 3779-3787.

[29] Passalis N, Tefas A. Learning deep representations with probabilistic knowledge transfer[C]. European Conference on Computer Vision. Cham: Springer, 2018: 283-299.

[30] Peng B Y, Jin X, Li D S, et al. Correlation congruence for knowledge distillation[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2019: 5006-5015.

[31] Park W, Kim D, Lu Y, et al. Relational knowledge distillation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 3967-3976.

[32] Tung F, Mori G. Similarity-preserving knowledge distillation[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2019: 1365-1374.

[33] Liu X, Li L, Li C, et al. NORM: Knowledge distillation via *N*-to-One representation matching[J/OL]. 2024, arXiv: 2402.11148.

[34] Li Z, Li X, Yang L F, et al. Curriculum temperature for knowledge distillation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(2): 1504-1512.

[35] Belkina A C, Ciccolella C O, Anno R, et al. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets[J]. Nature Communications, 2019, 10: 5415.

作者简介

张建 (2000−), 男, 硕士生, 主要研究方向为知识蒸馏、图像处理、计算机视觉, E-mail: 2022201366@aust.edu.cn;

梁兴柱 (1979−), 男, 副教授, 硕士, 主要研究方向为模式识别、计算机视觉, E-mail: xzliang@aust.edu.cn;

张康 (2003−), 男, 本科生, 主要研究方向为半导体器件、集成电路设计, E-mail: 2103617820@qq.com;

林玉娥 (1979−), 女, 副教授, 博士, 主要研究方向为图像处理、计算机视觉, E-mail: yelin@aust.edu.cn;

夏晨星 (1991−), 男, 副教授, 博士, 主要研究方向为图像处理、计算机视觉, E-mail: cxxia@aust.edu.cn.