计算机科学
COMPUTER SCIENCE

# 基于对比学习的超多类深度图像聚类模型

胡深, 钱宇华, 王婕婷, 李飞江, 吕维

引用本文

胡深, 钱宇华, 王婕婷, 李飞江, 吕维. 基于对比学习的超多类深度图像聚类模型[J]. 计算机科学, 2023, 50(9): 192-201.

HU Shen, QIAN Yuhua, WANG Jieting, LI Feijiang, LYU Wei. Super Multi-class Deep Image Clustering Model Based on Contrastive Learning [J]. Computer Science, 2023, 50(9): 192-201.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900133

# 基于对比学习的超多类深度图像聚类模型

胡 深[1,3] 钱宇华[1,2,3] 王婕婷[1,3] 李飞江[1,3] 吕 维[1,3]
1 山西大学计算机与信息技术学院 太原 030006
2 山西大学计算智能与中文信息处理教育部重点实验室 太原 030006
3 山西大学大数据科学与产业研究院 太原 030006
(hushenshen@aliyun.com)

**摘 要** 图像聚类通过表征学习对图像数据降维并提取有效特征而后进行聚类分析。当图像数据存在超多类别时,数据分布的复杂性和类簇的密集性严重影响了现有方法的实用性。为此,提出了基于对比学习的超多类深度图像聚类模型,主要分为3个阶段:首先,改进对比学习方法训练特征模型以使类簇分布均匀;其次,基于语义相似性原则多视角挖掘实例语义最近邻信息;最后,将实例及其最近邻作为自监督信息训练聚类模型。根据实验类型的不同,设计了消融实验和对比实验。在消融实验中,证明了所提方法使类簇均匀分布在映射空间,并可靠挖掘语义最近邻信息。在对比实验中,将其与先进算法在7个基准数据集上进行了比较,在ImageNet-200类数据集上,其准确率比目前先进方法提升了10.6%;在ImageNet-1000类数据集上,其准确率比目前先进算法提升了9.2%。

**关键词**:超多类聚类;对比学习;特征模型;语义相似性;图像聚类
**中图法分类号** TP391

## Super Multi-class Deep Image Clustering Model Based on Contrastive Learning

HU Shen[1,3],QIAN Yuhua[1,2,3],WANG Jieting[1,3],LI Feijiang[1,3] and LYU Wei[1,3]
1 School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2 Shanxi University Key Laboratory of Computational Intelligence and Chinese Information Processing,Ministry of Education,Taiyuan 030006,China
3 Institute of Big Data Science and Industry,Shanxi University,Taiyuan 030006,China

**Abstract** Image clustering reduces the dimensionality of image data,extracts effective features through representation learning,and performs cluster analysis. When there are many categories of image data,the complexity of data distribution and the density of clusters seriously affect the practicability of existing methods. To this end,this paper proposes a super-multi-class deep image clustering model based on contrastive learning,which is mainly divided into three stages:firstly,improving the contrastive learning method to train the feature model to make the cluster distribution uniform;secondly,based on the principle of semantic similarity,the perspective mines instance semantic nearest neighbor information;and finally,the instance and its nearest neighbors are used as self-supervised information to train a clustering model. According to the different types of experiments,ablation experiments and contrast experiments are designed in this paper. The ablation experiments prove that the proposed method could make the clusters evenly distributed in the mapping space and mine the semantic nearest neighbor information reliably. In the comparative experiments,it's compared with the advanced algorithms on 7 benchmark datasets. On the ImageNet-200 class dataset,it's accuracy is 10.6% higher than the advanced method. It's accuracy rate on the ImageNet-1000 class dataset is higher than that of the advanced algorithm,which improves by 9.2%.

**Keywords** Super multi-class clustering,Contrastive learning,Feature model,Semantic similarity,Image clustering

到稿日期:2022-08-02 返修日期:2022-10-10
基金项目:国家自然科学基金重点项目(62136005);国家重点研发计划(2021ZD0112400);国家自然科学基金青年科学基金(62106132);山西省三晋学者项目资助;山西省基础研究计划(20210302124271,202103021223026)
This work was supported by the Key Program of the National Natural Science Foundation of China(62136005), National Key Research and Development Program of China(2021ZD0112400),Young Scientists Fund of the National Natural Science Foundation of China(62106132),Program for the San Jin Young Scholars of Shanxi and Shanxi Provincial Research Foundation for Basic Research, China (20210302124271, 202103021223026).
通信作者:钱宇华(jinchengqyh@126.com)

## 1　引言

图像聚类通过算法模型对原始图像提取特征，然后采用聚类算法对特征进行聚类，将相似图像归为同一类。图像聚类在无标注情况下挖掘图像数据潜在的结构信息，多应用在医疗影像[1]、图像识别[2-3]、遥感监测[4]、参数辨别[5]等领域。

现有图像聚类方法在类别较少的数据集上的性能已经逼近有监督分类结果，但在超多类数据集上的性能仍有提升空间。先进图像聚类方法 RUC[6] 和 SPICE[7] 等在 STL-10 类数据集[8] 上的准确率达到了 86%以上，在 CIFAR-10 类数据集[9] 上的准确率达到了 90%以上，但在 CIFAR-20 类数据集[9] 上的准确率目前最高只有 58%。IDFD[10]，ProPos[11] 和 ConCURL[12] 等在 ImageNet-10 类数据集[13] 上的聚类准确率普遍达到了 95%以上，但在 ImageNet-1000 类数据集[14] 上目前先进方法的准确率只有 39.9%。

超多类图像聚类方法仍面临以下两个方面的挑战：(1)超多类数据类型和分布的复杂化导致实例间关系的不确定性增强[15]，现有聚类方法特征模型不利于超多类簇密集问题；(2)构造的自监督信息存在噪声数据，导致学习结果具有随机一致性[16]。

目前，深度图像聚类的相关工作大致可分为基于传统聚类方法和基于自监督方法两种。基于传统聚类方法使用神经网络提取特征之后，直接使用 $K$-Means、谱聚类等传统算法进行聚类分析，但这类方法依赖算法模型的初始化参数，无法有效提取高维数据的语义特征，难以获得优质聚类效果，在实际场景中也难以应用。自监督方法为提升聚类方法的稳定性[17] 和泛化性[18]，通过构造图像伪标签、正样本对、最近邻等自监督信息训练模型来提取图像高级语义信息，以对实例产生具有判别性的表示，便于衡量实例之间的相似度，提高了聚类算法的性能。

本文沿用自监督学习思想，针对上述超多类图像聚类方法中存在的不足，提出了基于对比学习的超多类深度图像聚类模型。该模型主要分为 3 个阶段：首先，使用新的正向对齐策略构造对比损失，同时提升对比学习的对齐性(见 3.1.1 节)和均匀性(见 3.1.2 节)性质以训练特征模型；其次，基于语义相似性原则，在原型、强增强和弱增强 3 个视图的特征空间中，逐步对每一实例挖掘语义最近邻信息(见 3.2 节)；最后，将实例和其最近邻作为可靠自监督信息，结合特定损失函数训练聚类模型(见 3.3 节)。本文的贡献总结如下：

(1)针对超多类数据分布复杂和类簇密集问题，改进对比学习方法训练特征模型，使类簇均匀分布在映射空间。

(2)针对自监督信息构建存在噪声的问题，改进构建策略，去除类簇边界语义不一致数据，以正确指导聚类模型训练。

(3)关注超多类聚类问题，且在超多类数据集 ImageNet-200 类上的准确率比目前先进聚类算法提高了 10.6%，在 ImageNet-1000 类数据集上的准确率比目前先进方法提升了 9.2%。

本文第 2 章主要介绍图像聚类和对比学习相关研究的现状；第 3 章阐述基于对比学习的超多类深度图像聚类模型；第 4 章将实验分为消融实验、对比实验和扩充实验并对实验结果进行分析；最后总结全文并展望未来。

## 2　相关工作

自监督学习为深度图像聚类联合优化特征模型和聚类模型提供了新的思路，日益成为研究热点。2.1 节介绍相关传统聚类算法以及基于传统聚类算法的聚类工作；2.2 节介绍基于自监督学习的深度图像聚类相关工作；2.3 节介绍使用对比学习优化特征模型的相关工作。另外现实世界数据往往含有多个类别的语义信息，如果只考虑单一语义标签对其进行学习，就很难获得很好的分类/聚类效果。Zhang 等提出 ML-GAT[19] 用于提升模型的多标签图像分类性能。

### 2.1　基于传统聚类算法的相关工作

大规模数据给聚类分析带来了挑战，当数据类别较多、数据量较大时，聚类算法的性能和聚类时间都会受到不同程度的影响，SA-ART[20]，VA-ART[21]，GHF-ART[22] 和 ACH-ART[23] 等工作基于自适应共振理论做了大量研究以解决大规模数据聚类问题。Lei 等基于自适应理论 ART[24] 在 SA-ART[20] 中提出聚类空间匹配和显著特征加权以对 Fuzzy ART[25] 进行扩展，引入了在共享特征空间中测量相似性、自调节学习率和集群权重恢复等策略稳定地加强对大规模数据的学习，SA-ART 具有良好的可扩展性，可以纳入新的理论和方法。SA-ART 使用聚类空间匹配策略，能够对集群的突出特征进行建模，具有很好的鲁棒性，在大规模超多类数据集上取得了很好的效果，但 SA-ART 对噪声数据具有较强的鲁棒性，泛化能力略低。

图像等高维数据也给传统聚类算法带来了挑战，因其在高维特征空间中难以有效衡量数据之间的相似性，从而无法进行有效聚类。部分学者研究使用深度神经网络提取实例的高级语义特征，然后在特征空间中使用传统聚类算法进行聚类。例如：结合 $K$-Means 算法的 DEN[26] 和 DEC[27] 以及 DCN[28]、基于谱聚类算法高维空间表达能力的 DSC[29]、联合子空间聚类算法的 DASC[30] 和结合相对熵的 DEPICT[31] 等。此类深度聚类方法大多将自动编码器与传统聚类算法相结合，自动编码器过度强调低级特征导致早期聚类方法缺乏对象层次的高级语义信息而表现较差。

### 2.2　基于自监督学习的深度图像聚类相关工作

学者研究在无监督情况下构造自监督信息训练聚类模型以提升性能。早期，JULE[32] 和 Deep Cluster[33] 等方法交替进行聚类。例如 Deep Cluster[33] 将特征表示后的实例使用 $K$-Means 生成伪标签形成自监督信息，伪标签监督分类网络训练优化特征模型，优化后的特征模型指导 $K$-Means 更新伪标签，如此交替循环优化，最终得到聚类结果。但是以上交替聚类算法在特征模型和聚类模型交替时会产生误差，例如伪标签存在较多错误标注等问题，这种误差在迭代优化过程中会逐步累积，最终导致聚类性能恶化。接下来 DAC[13]，DCCM[34]，IIC[35]，PICA[36]，SCAN[37]，CC[38]，IDFD[10] 等方法表明可以通过判别性的表示来提高图像聚类的性能。它们利用自监督学习中的对比学习思想，直接学习将图像映射成标签特征，在训练时将其作为表示特征，在测试时作为 one-hot 编码的聚类索引。

Li 等在 2021 年提出了 CC(Contrastive Clustering)[38]，通过构造正负样本对作为自监督信息，最大化正样本对的相似性，同时最小化负样本对的相似性，分别在行和列空间进行

实例和集群级的对比学习，优化表示模型，最终对实例得到one-hot编码表示，直接得到聚类结果。但CC使用实例级的对比损失不关注集群的统一性，会将同类簇内部分实例作为负样本，产生类碰撞问题。

Van Gansbeke等在2020年提出的SCAN(Learning to Classify Images Without Labels)[37]使用对比学习训练特征模型对实例进行表示，在此基础上，基于语义相似性挖掘实例最近邻信息，利用相似样本具有相同的标签来指导优化聚类模型。然而，由于表示特征并不完美，当实例位于不同类簇边界线时，相似实例并不总是具有相同的语义，只使用实例相似性而忽略集群之间的语义差异来挖掘最近邻会限制聚类的性能[7]。

### 2.3 对比学习

对比学习的目标是：通过设计前置任务，构造正负样本对训练特征模型，使得相似实例在映射空间中比较接近，而相异实例在映射空间中距离较远。如何设计正负样本对、如何构造遵循上述指导原则的对比学习结构，以及如何防止模型坍塌是对比学习中比较重要的关键点。对比学习又可以依据是否使用对抗信息分为两类。

#### 2.3.1 使用对抗信息

使用对抗信息的代表性工作有MoCo[39]和SimCLR[40]等。SimCLR的正样本来源于同一实例 $\boldsymbol{x}$ 的两种增强视图 $\boldsymbol{u}_x^a, \boldsymbol{u}_x^b$。负样本来源于同一Batch，对 $\boldsymbol{u}_x^a$ 和Batch内任一样本 $\boldsymbol{x}_i^-$ 的增强视图 $\boldsymbol{u}_{xi^-}^a$，$\boldsymbol{u}_{xi^-}^b$ 均构造负样本对 $(\boldsymbol{u}_x^a, \boldsymbol{u}_{xi^-}^a)$ 和 $(\boldsymbol{u}_x^a, \boldsymbol{u}_{xi^-}^b)$。样本对 $\boldsymbol{u}^a, \boldsymbol{u}^b$ 之间的相似性采用余弦相似度进行度量，如式(1)所示。

$$\mathrm{sim}(\boldsymbol{f}_{u^a}, \boldsymbol{f}_{u^b}) = \boldsymbol{f}_{u^a}^{\mathrm{T}} \boldsymbol{f}_{u^b} / (\| \boldsymbol{f}_{u^a} \|_2 \| \boldsymbol{f}_{u^b} \|_2) \tag{1}$$

其中，$f$ 为特征模型输出。

将 $\mathrm{sim}(\boldsymbol{f}_{u^a}, \boldsymbol{f}_{u^b})$ 简写为 $s(u^a, u^b)$，SimCLR使用的对比损失InfoNCE可写为式(2)：

$$L_{\mathrm{InfoNCE}} = -\log \frac{\exp(s(\boldsymbol{u}^a, \boldsymbol{u}^b)/t)}{\exp(s(\boldsymbol{u}^a, \boldsymbol{u}^b)/t) + \sum_{i=0}^{D} \exp(s(\boldsymbol{u}^a, \boldsymbol{u}_i^-)/t)} \tag{2}$$

使用对比损失训练特征模型，目的是将实例映射到一个空间上，并在这个空间内拉近正样本对的距离，推远负样本对的距离，迫使特征模型忽略表面因素，学习实例的内在一致结构信息，即学会某些类型的不变性，如遮挡不变性、旋转不变性、颜色不变性等。

如果不使用负样本对，只关注正样本对之间的相似性，特征模型很容易退化到一个崩溃解，即特征模型塌缩到对所有输入都输出同样的特征。但使用负样本对计算开销较大且需要使用较大Batch size加载足量负样本，对内存要求较高。并且由于将增强视图外所有实例均视为负例，不可避免会将同类簇内部分实例作为负样本对参与训练，产生类碰撞问题。

#### 2.3.2 不使用对抗信息

不使用对抗信息的代表性工作有BYOL(Bootstrap your own latent: A new approach to self-supervised Learning)[41]和SimSiam(Exploring Simple Siamese Representation Learning)[42]等。BYOL在不使用负样本对的情况下，为了防止模型塌缩得到崩溃解，使用了两个参数不同的网络 $z_1$ 和 $z_2$，并对其中一个网络 $z_2$ 停止梯度传播，以避免两个网络相似，并加入预测器 $g$ 来缩小两个网络之间的差距。当温度系数 $\tau = 0.5$ 时，BYOL使用的损失可以写为式(3)：

$$\begin{aligned} L_{\text{non-contr}} &= -2g(z_1(\boldsymbol{x}))^{\mathrm{T}} z_2(\boldsymbol{x}_+) \\ &= \| g(z_1(\boldsymbol{x})) - z_2(\boldsymbol{x}_+) \|_2^2 + const \end{aligned} \tag{3}$$

其中，$(\boldsymbol{x}, \boldsymbol{x}_+)$ 为正样本对，$g(z_1(\cdot))$ 和 $z_2(\cdot)$ 的输出是L2正则标准化后的。

不使用负样本对的学习方法经常受到不稳定训练的影响，并且高度依赖批量统计学和超参数调整等更多手段来避免表示崩溃。

## 3 基于对比学习的超多类深度图像聚类模型

本模型使用自监督学习思想训练聚类模型以对实例产生one-hot表示，基于SCAN[37]方法进行改进。SCAN方法首先训练特征模型对数据进行表示，之后基于特征相似度计算近邻，最后使用损失训练聚类网络以对实例产生one-hot表示。本模型和SCAN方法有以下两个主要不同点：(1)特征模型训练方法不同，本文模型从对齐性和均匀性出发建立了新的对比损失训练特征模型；(2)近邻寻找策略不同，本文模型从多视图角度并考虑语义差异寻找最近邻作为自监督信息。本文将聚类模型分为3个阶段，整体架构图如图1所示。

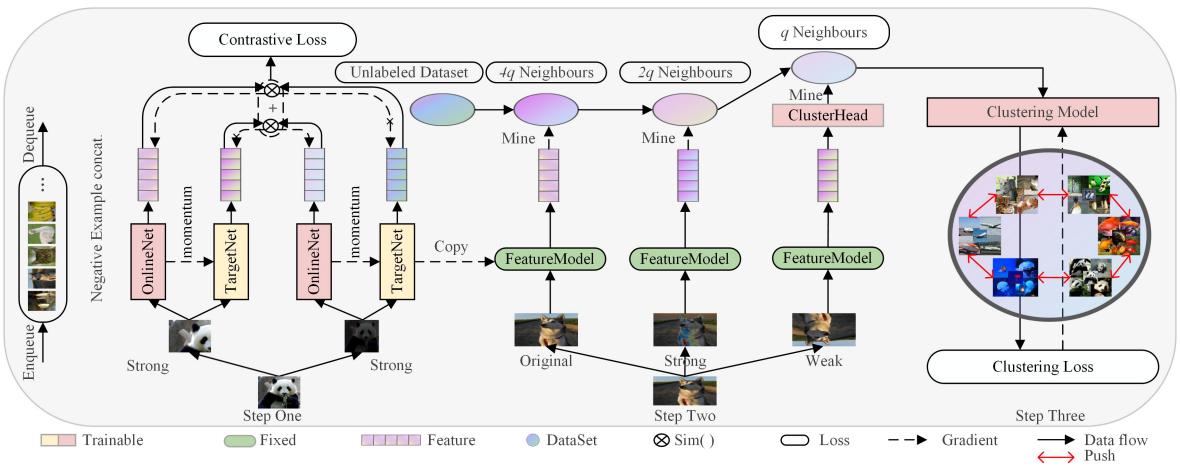



图1 整体架构图

Fig. 1 Overall architecture

第一阶段，改进对比学习方法，训练特征模型；第二阶段，为每一实例挖掘语义最近邻信息；第三阶段，将实例和语义最近邻作为自监督信息优化聚类模型。

### 3.1　适用于超多类深度图像聚类的特征模型

本小节介绍使用对比学习思想训练特征模型的算法。对常用对比损失 InfoNCE(见式(2))进行分析，当负样本的数量趋近于无穷大时，此损失可近似解耦为两项[43]，如式(4)所示：

$$\begin{aligned}L_{\text{contrast}}&=-\log\frac{\exp(s(\boldsymbol{u}^a,\boldsymbol{u}^b)/t)}{\exp(s(\boldsymbol{u}^a,\boldsymbol{u}^b)/t)+\sum_{i=0}^{D}\exp(s(\boldsymbol{u}^a,\boldsymbol{u}_i^-)/t)}\\&=-s(\boldsymbol{u}^a,\boldsymbol{u}^b)/t+\log(\exp(s(\boldsymbol{u}^a,\boldsymbol{u}^b)/t)+\\&\quad\sum_{i=0}^{D}\exp(s(\boldsymbol{u}^a,\boldsymbol{u}_i^-)/t))\end{aligned}\tag{4}$$

因式(4)中$\sum_{i=0}^{D}\exp(s(\boldsymbol{u}^a,\boldsymbol{u}_i^-)/t)$有下界且总为正值，此时损失函数倾向于较小值$-s(\boldsymbol{u}^a,\boldsymbol{u}^b)/t$，假设编码器是完全对齐的，即$P[(\boldsymbol{u}_{\boldsymbol{u}^a}=f_{\boldsymbol{u}^b})]=1$，则式(4)可推导为式(5)：

$$\begin{aligned}L_{\text{contrast}}&=-\log\frac{\exp(s(\boldsymbol{u}^a,\boldsymbol{u}^b)/t)}{\exp(s(\boldsymbol{u}^a,\boldsymbol{u}^b)/t)+\sum_{i=0}^{D}\exp(s(\boldsymbol{u}^a,\boldsymbol{u}_i^-)/t)}\\&=\underbrace{-s(\boldsymbol{u}^a,\boldsymbol{u}^b)/t}_{\text{positive alignment}}+\\&\quad\underbrace{\log(\exp(1/t)+\sum_{i=0}^{D}\exp(s(\boldsymbol{u}^a,\boldsymbol{u}_i^-)/t))}_{\text{uniformity}}\end{aligned}\tag{5}$$

式(5)将对比损失分解为两个部分：第一部分为对齐性，即相似的正例映射到特征空间后，距离应该较近；第二部分为均匀性，即类簇应在映射空间上均匀分布，特征内保留尽可能多的独有信息。这两部分都是需要的，假如只有对齐性没有均匀性，模型很容易坍缩得到退化解，在 2.2.1 节中也有描述。假如只有均匀性没有对齐性，同类实例两两之间都被错误作为负样本对参与训练，致使模型退化，同类样本之间被推开，并且没有正样本对之间的对齐监督信息，就失去了对比学习的意义。

本节研究同时优化对齐性(见 3.1.1 节)和均匀性(见 3.1.2 节)以训练特征模型，使类簇更均匀分布在映射空间，拉开不同集群之间的距离，并使同类簇内实例距离更接近，从而有效指导超多类聚类。

#### 3.1.1　对齐性

本文使用新正向对齐策略，将 Cutout[44] 和 Rand Augment[45] 的 4 个变换以不同方式组合起来生成两个强增强策略 $t_a$ 和 $t_b$，分别作用在实例$\boldsymbol{x}$上，得到两个增强视图$\boldsymbol{u}^a$，$\boldsymbol{u}^b$；并分别经过在线网络 $v_1$ 提取特征并变换到映射空间得到$\boldsymbol{l}_a$，$\boldsymbol{l}_b$，经过目标网络 $v_2$ 得到$\boldsymbol{h}_a$，$\boldsymbol{h}_b$。在线网络 $v_1$ 由一个骨干网络 ResNet-50 和两个多层感知机结构组成，目标网络 $v_2$ 由一个骨干网络 ResNet-50 和一个多层感知机结构组成，构建($\boldsymbol{l}_a$，$\boldsymbol{h}_b$)和($\boldsymbol{l}_b$，$\boldsymbol{h}_a$)为两对正样本对，建立对称对齐损失如式(6)所示，其中 $sim(\cdot)$ 为式(1)样本对之间的相似度度量方式。

$$L_{\text{con}}=\frac{1}{2}sim(\boldsymbol{h}_a,\boldsymbol{h}_b)+\frac{1}{2}sim(\boldsymbol{h}_b,\boldsymbol{h}_a)\tag{6}$$

#### 3.1.2　均匀性

负样本来自于一个可维护的队列，负样本队列内任一实例$\boldsymbol{x}_i^-$的目标网络表示$\boldsymbol{h}_i^-\in\{\boldsymbol{h}^-\}$均和实例$\boldsymbol{x}$的在线网络表示$\boldsymbol{l}_a$构为负样本对($\boldsymbol{l}_a$，$\boldsymbol{h}_i^-$)。对$\boldsymbol{l}_b$同理构造负样本对($\boldsymbol{l}_b$，$\boldsymbol{h}_i^-$)。在任意一个 Epoch 的任意一次迭代中，只使用反向传播更新在线网络 $v_1$ 的参数，然后通过式(7)的移动平均方法更新目标网络 $v_2$ 的参数。同时，队列尾部删除一个 Batch 大小的负样本列，头部更新一个 Batch 大小的负样本列，完成这次迭代的负样本队列的更新。负样本队列内的值是负样本通过目标网络 $v_2$ 输出的特征值，它代表所有负样本的视觉表示，队列足够大，就可以构造足够多的负样本对。

$$v_2=mv_2+(1-m)v_1\tag{7}$$

$$m\triangleq 1-(1-m_{\text{base}})\cdot(\cos(\pi n/N)+1)/2\tag{8}$$

式(7)中 $m$ 取值从 $m_{\text{base}}=0.996$ 开始，按照式(8)更新参数 $m$ 数值，$n$ 是当前训练步数，$N$ 是最大训练步数。因为权重 $m$ 很接近 1，所以目标网络的更新极其缓慢，目标网络的更新相当于是参考了很多的 Batch，也就是参考了很多负样本，在 3.1.1 节式(6)中引入负样本对来提升均匀性。

#### 3.1.3　构造对比损失

在全部训练数据集 $D$ 上，对实例 $x\in D$ 建立新对比损失，如式(9)所示，其中($\boldsymbol{l}_a$，$\boldsymbol{h}_b$)和($\boldsymbol{l}_b$，$\boldsymbol{h}_a$)为使用 3.1.1 节方法构造的两对正样本对。

$$L_{\text{contrast}}=\frac{1}{2}\mathscr{L}(\boldsymbol{l}_a,\boldsymbol{h}_b,\{\boldsymbol{h}^-\})+\frac{1}{2}\mathscr{L}(\boldsymbol{l}_b,\boldsymbol{h}_a,\{\boldsymbol{h}^-\})\tag{9}$$

其中，$\{\boldsymbol{h}^-\}$为 3.1.2 节所述负样本队列经过目标网络 $v_2$ 输出特征的集合，通过 $\mathscr{L}(\cdot)$ 引入负样本队列 $\{\boldsymbol{h}^-\}$。如式(10)所示：

$$L(\boldsymbol{l},\boldsymbol{h}^+,\{\boldsymbol{h}^-\})=-\log\frac{\exp(s(\boldsymbol{l},\boldsymbol{h}^+)/t)}{\exp(s(\boldsymbol{l},\boldsymbol{h}^+)/t)+\sum\limits_{\{\boldsymbol{h}^-\}}\exp(s(\boldsymbol{l},\boldsymbol{h}_i^-)/t)}\tag{10}$$

其中，超参数温度系数 $\tau$ 设置为 0.2，温度系数决定了对比损失对未被推远负样本的关注程度。为更直观地了解温度系数的作用，分析温度系数 $\tau$ 趋向于 0 和无穷大[46]的两种极端情况。温度系数趋近于 0 时，式(10)可推导为式(11)：

$$\begin{aligned}&\lim_{\tau\to0^+}-\log\frac{\exp(s(\boldsymbol{l},\boldsymbol{h}^+)/\tau)}{\exp(s(\boldsymbol{l},\boldsymbol{h}^+)/\tau)+\sum\limits_{\{\boldsymbol{h}^-\}}\exp(s(\boldsymbol{l},\boldsymbol{h}_i^-)/\tau)}\\&=\lim_{\tau\to0^+}+\log[1+\sum_{\{\boldsymbol{h}^-\}}\exp((s(\boldsymbol{l},\boldsymbol{h}_i^-)-s(\boldsymbol{l},\boldsymbol{h}^+))/\tau)]\\&=\lim_{\tau\to0^+}+\log[1+\sum_{s(\boldsymbol{l},\boldsymbol{h})\geqslant s(\boldsymbol{l},\boldsymbol{h}^+)}\exp((s(\boldsymbol{l},\boldsymbol{h})-s(\boldsymbol{l},\boldsymbol{h}^+))/\tau)]\\&=\lim_{\tau\to0^+}\frac{1}{\tau}\max[s_{\max}-s(\boldsymbol{l},\boldsymbol{h}^+),0]\end{aligned}\tag{11}$$

此时式(10)退化为只关注距离本实例最近的未被推远负样本的损失函数。而当温度系数趋向于无穷大时，式(10)可推导为式(12)，此时损失对所有负样本的权重都相同，都为 $1/N\tau$，即损失函数失去了对未被推开负样本的关注特性，即当温度系数趋向于无穷时，该损失便变成了简单损失 $\mathscr{L}_{\text{simple}}$，如式(13)所示：

$$\begin{aligned}&\lim_{\tau\to+\infty}-\log\frac{\exp(s(\boldsymbol{l},\boldsymbol{h}^+)/\tau)}{\exp(s(\boldsymbol{l},\boldsymbol{h}^+)/\tau)+\sum\limits_{\{\boldsymbol{h}^-\}}\exp(s(\boldsymbol{l},\boldsymbol{h}_i^-)/\tau)}\\&=\lim_{\tau\to+\infty}-\frac{1}{\tau}s(\boldsymbol{l},\boldsymbol{h}^+)+\log\sum\exp(s(\boldsymbol{l},\boldsymbol{h})/\tau)\\&=\lim_{\tau\to+\infty}-\frac{1}{\tau}s(\boldsymbol{l},\boldsymbol{h}^+)+\frac{1}{N}\sum\exp(s(\boldsymbol{l},\boldsymbol{h})/\tau)-1+\log N\end{aligned}$$

$$= \lim_{\tau \to +\infty} -\frac{N-1}{N\tau} s(\boldsymbol{l}, \boldsymbol{h}^+) + \frac{1}{N\tau} \sum_{\{\boldsymbol{h}^-\}} s(\boldsymbol{l}, \boldsymbol{h}_i^-) + \log N \quad (12)$$

$$\mathcal{T}_{\text{simple}}(\boldsymbol{l}, \boldsymbol{h}^+, \{\boldsymbol{h}^-\}) = -s(\boldsymbol{l}, \boldsymbol{h}^+) + \lambda \sum_{\{\boldsymbol{h}^-\}} s(\boldsymbol{l}, \boldsymbol{h}_i^-) \quad (13)$$

最后，使用构建的正负样本对和新对比损失(见式(9))训练在线网络和目标网络，优化目标是使正样本对$(\boldsymbol{l}_a, \boldsymbol{h}_b)$和$(\boldsymbol{l}_b, \boldsymbol{h}_a)$内的$\boldsymbol{l}$和$\boldsymbol{h}$的相似程度尽可能高，使$\boldsymbol{l}$和$\{\boldsymbol{h}^-\}$内所有负样本的相似程度尽可能低。

### 3.2 语义最近邻

SCAN[37]量化了挖掘到的最近邻属于同一语义集群的程度：对于挖掘不同数目的近邻，在4个数据集CIFAR-10[9]，CIFAR100-20[9]，STL-10[8]和ImageNet[14]中都有较高的语义相似性。本文语义最近邻寻找策略同时考虑实例的相似性和语义差异性，从多视图角度挖掘语义最近邻，多视图数据的每一个视图都包含其他视图不具备的信息或知识，本文利用不同视图间的差异性和互补性信息更加全面地跨视图挖掘语义最近邻。

将本文3.1节所述在线网络$v_1$的特征模型$F$的权重记为$\theta_r$。首先，对每个实例$x_i \in \mathcal{D}$，得到特征$\boldsymbol{f}_{x_i} = \Phi(\boldsymbol{x}_i; \theta_r)$，在实例特征空间中使用向量检索库Faiss比较多维空间中数据之间的相似性来搜索与实例$\boldsymbol{x}_i$最相近的$4q$个近邻，记为集合$N_{4q}x_i$。其次，对实例$\boldsymbol{x}_i$作强变换$\alpha$得到$\boldsymbol{\alpha}_{x_i} = \alpha(\boldsymbol{x}_i)$，同样得到特征$\boldsymbol{f}_{\boldsymbol{\alpha}_{x_i}} = \Phi(\boldsymbol{\alpha}_{x_i}; \theta_r)$。对于强变换$\alpha$，本文将Rand Augment[45]中随机选择的4个变换和Cutout[44]组合起来，对图像进行强增强。在强变换映射空间中在集合$N_{4q}x_i$上使用向量检索库Faiss比较多维空间中数据之间的相似性来搜索与实例$\boldsymbol{x}_i$最相近的$2q$个近邻，得到近邻集合$N_{2q}x_i$。最后，对实例$\boldsymbol{x}_i$作弱变换$\beta$得到$\boldsymbol{\beta}_{x_i} = \beta(\boldsymbol{x}_i)$，得到特征$\boldsymbol{f}_{\boldsymbol{\beta}_{x_i}} = \Phi(\boldsymbol{\beta}_{x_i}; \theta_r)$。对于弱变换$\beta$，本文采用FixMatch[47]中的标准翻转和移位增强策略，将$\boldsymbol{f}_{\boldsymbol{\beta}_x}$输入到一个多层感知机聚类头$\theta_{\text{clu}}$来预测实例属于某一类的概率$p_{\beta_x} = \Phi(\boldsymbol{f}_{\boldsymbol{\beta}_x}; \theta_{\text{clu}})$，权重$\theta_{\text{clu}}$部分来自预训练模型，在集合$N_{2q}x_i$上去除$q$个$p_{\beta_x}$较小的近邻，最终得到实例$\boldsymbol{x}_i$的$q$个最近邻，在数据集$D$中将实例$\boldsymbol{x}_i$的$q$个最近邻定义为集合$\mathcal{N}x_i$。

### 3.3 聚类算法

本文模型的伪代码如算法1所示。

**算法1** Super Multi-Class Deep Image Clustering Model

Input: Dataset $\mathcal{D}$, Clusters $\mathbf{C}$, Nets $\Phi_v$, $\Phi_c$ and $\Phi_\eta$, Nei $\mathcal{N}_\mathcal{D}$

Draw augmentation functions $t_1, t_2, \alpha, \beta$

For $\mathbf{x}_i \in D$ do ▶Feature Model Step 1, Sec. 3.1

$\mathbf{u}^a, \mathbf{u}^b = t_1(\mathbf{x}_i), t_2(\mathbf{x}_i)$

$\mathbf{l}_a, \mathbf{l}_b = \Phi_{v_1}(\mathbf{u}^a), \Phi_{v_1}(\mathbf{u}^b)$

$\mathbf{h}^a, \mathbf{h}^b = \Phi_{v_2}(\mathbf{u}^a), \Phi_{v_2}(\mathbf{u}^b)$

For $\mathbf{x}_i^-$ in Negative queen do

$\mathbf{h}_i^- = \Phi_{v_2}(\mathbf{x}_i^-)$

End for

Loss $= \mathcal{L}(\mathbf{l}_a, \mathbf{h}_b, \{\mathbf{h}^-\}) + \mathcal{L}(\mathbf{l}_b, \mathbf{h}_a, \{\mathbf{h}^-\})$ in Eq. 9

Update $v_1$, Update $v_2 = mv_2 + (1-m)v_1$ in Eq. 7

Update Negative queen in each batch

End for

Update $\Phi_c$ from $\Phi_v$

For $\mathbf{x}_i \in \mathcal{D}$ do ▶Mine Neighbors Step 2, Sec. 3.2

$\mathrm{N}_{4q}\mathbf{x}_i = $ 4q neighboring samples of $\Phi_{v_1}(\mathbf{x}_i)$ in $\mathcal{D}$

$\boldsymbol{\alpha}_{\mathbf{x}_i} = \alpha(\mathbf{x}_i)$

$\mathrm{N}_{2q}\mathbf{x}_i = $ 2q neighboring samples of $\Phi_{v_1}(\mathbf{a}_{\mathbf{x}_i})$ in $\mathrm{N}_{4q}\mathbf{x}_i$

$\boldsymbol{\beta}_{\mathbf{x}_i} = \beta(\mathbf{x}_i)$

$\mathcal{N}\mathbf{x}_i = $ q max $p_{\beta_x} = \Phi_c(\Phi_{v_1}(\boldsymbol{\beta}_x))$ samples of $\mathrm{N}_{2q}\mathbf{x}_i$

$\mathcal{N}_\mathcal{D} \leftarrow \mathcal{N}_\mathcal{D} \bigcup \mathcal{N}\mathbf{x}_i$

End for

while Loss decreases do ▶Clustering Step 3, Sec. 3.3

Update $\Phi_\eta$ with Loss, i.e. $\Lambda(\Phi_\eta(\mathcal{D}), \mathcal{N}_\mathrm{D}, \mathscr{C})$ in Eq. 14

End while

Return: $\Phi_\eta(\mathcal{D})$ ▶$\mathcal{D}$ is divided over $\mathscr{C}$ clusters

聚类模型由骨干网络ResNet-50和聚类头组成，骨干网络的权重来自本文3.1节在线网络$v_1$的特征模型$F$，聚类头$\Phi_\eta$是权重为$\eta$的神经网络。经由第一步特征模型训练和第二步寻找最近邻生成可靠自监督信息，第三步聚类模型训练阶段损失的目的是将3.2节对实例寻找到的最近邻$\mathcal{N}_X$作为自监督信息来指导训练神经网络$\Phi_\eta$，以将实例映射成标签特征，在训练时作为表示特征，在测试时作为one-hot编码的聚类索引。损失函数使用SCAN-loss[37]，如式(14)所示：

$$\Lambda = -\frac{1}{|\mathcal{D}|} \sum_{X \in \mathcal{D}} \sum_{V \in \mathcal{N}_X} \log \langle \Phi_\eta(\boldsymbol{X}), \Phi_\eta(\boldsymbol{V}) \rangle + \lambda \sum_{c \in C} \Phi_\eta'^c \log \Phi_\eta'^c \quad (14)$$

其中，$\langle \cdot \rangle$表示点积计算，样本$\boldsymbol{x}_i$被分配到集群$c$的概率表示为$\Phi_\eta^c(\boldsymbol{x}_i), \Phi_\eta(\boldsymbol{x}_i) \in [0,1]^c$，$\mathscr{C} = \{1, \cdots, C\}$。

为了避免聚类头$\Phi_\eta$将数据集$\mathcal{D}$内所有样本都分配到一个集群中，使用式(14)中的第二项$\lambda \sum_{c \in C} \Phi_\eta'^c \log \Phi_\eta'^c$将聚类模型的预测均匀地分散到各类簇中，其中$\Phi_\eta'^c$的计算方式如式(15)所示：

$$\Phi_\eta'^c = \frac{1}{|\mathcal{D}|} \sum_{X \in \mathcal{D}} \Phi_\eta^c(\boldsymbol{X}) \quad (15)$$

当聚类头$\Phi_\eta$将$\boldsymbol{x}_i$和其近邻$Nx_i$都分配到同一个类簇并且向量表示是one-hot时，点积最大，也就是损失最小，从而训练聚类头$\Phi_\eta$对输入实例产生一致且有判别度的one-hot向量表示，进而对one-hot向量使用argmax函数对实例得出类别预测值，对所有数据得出聚类结果。

## 4 实验结果和分析

### 4.1 基准数据集和实验设置

本节在7个常用的图像基准数据集上评估模型的聚类性能，包括STL-10[8]，ImageNet-10[13]，ImageNet-Dogs[13]，ImageNet-50[37]，ImageNet-100[37]，ImageNet-200[37]和ImageNet-1000[14]。表1列出了每个数据集的关键参数，包括类别数目、图像尺寸和实例数量等。对于ImageNet数据集采用最近自监督、分类和聚类工作广泛使用的尺寸224×224对原始图像进行调整，其他数据集使用原尺寸。骨干网络统一使用ResNet-50。语义最近邻数目$q$的设置为：ImageNet数据集每个实例采用50个最近邻，STL-10数据集每个实例采用20个最近邻。特征模型训练阶段使用256 Batch size训练300个Epoch，聚类模型训练阶段使用512 Batch size训练100个Epoch。

表 1　基准数据集参数

Table 1　Dataset parameters overview

| Dataset | # Classes | Image size | # Split | # Samples |
|---|---|---|---|---|
| STL-10 | 10 | 96×96 | Train+Test | 13 000 |
| ImageNet-10 | 10 | 224×224 | Train+Test | 13 500 |
| ImageNet-Dogs-15 | 15 | 224×224 | Train+Test | 20 250 |
| ImageNet-50 | 50 | 224×224 | Train+Test | 66 774 |
| ImageNet-100 | 100 | 224×224 | Train+Test | 133 545 |
| ImageNet-200 | 200 | 224×224 | Train+Test | 266 558 |
| ImageNet-1000 | 1 000 | 224×224 | Train+Test | 1 331 167 |

本文实验使用 3 个主流的深度聚类评价指标，包括聚类准确率(ACC)、归一化互信息(NMI)和调整兰德系数(ARI)。ACC 指标是聚类正确的样本数量与总样本数量的比值，取值范围为[0,1]。NMI 即归一化互信息，用来衡量两个数据集分布的吻合程度，聚类结果和数据已知划分越相似，互信息就越大，取值范围为[0,1]。ARI 指标反映聚类结果和数据已知划分的重叠程度，取值范围为[−1,1]。以上 3 个指标都是值越大代表聚类效果越好。

### 4.2　消融实验

本节主要分析提出的特征模型训练策略和语义最近邻寻找策略两个模块的作用。需要注意寻找语义近邻的性能和特征模型的表示性能是正相关的。本文在少类数据集 STL-10 上进行消融实验，结果如表 2 所列。可以看出，在 SCAN 方法上将近邻寻找策略更换为本文近邻寻找策略，3 个指标的提升度为：ACC(2.6%)，NMI(2.9%)，ARI(3.6%)。将 SCAN 特征模型训练策略更换为本文训练策略，3 个指标的提升度为：ACC(5.6%)，NMI(7.9%)，ARI(10.1%)。最后，同时采用特征模型训练策略和语义最近邻寻找策略，3 个指标的提升度为：ACC(9.9%)，NMI(12%)，ARI(16%)。

表 2　在 STL-10 和 ImageNet-200 数据集上的消融实验

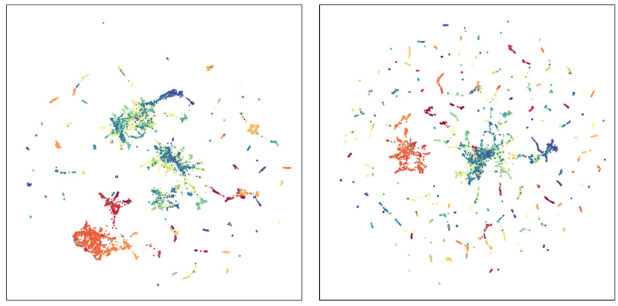
Table 2　Ablation experiments on STL-10 and ImageNet-200 datasets

| Method | Our Feature Model | Our Mine Neighbors | STL-10 | | | ImageNet-200 | | |
|---|---|---|---|---|---|---|---|---|
| | | | *ACC* | *NMI* | *ARI* | *ACC* | *NMI* | *ARI* |
| Ours | | | 80.9 | 69.8 | 64.6 | 56.3 | 75.7 | 44.1 |
| | | √ | 83.5 | 72.7 | 68.2 | 59.8 | 76.8 | 46.8 |
| | √ | | 86.5 | 76.9 | 74.7 | 64.4 | 79.1 | 50.1 |
| | √ | √ | **90.8** | **81.8** | **80.6** | **68.7** | **80.8** | **54.2** |

本文主要关注超多类图像聚类问题，进而在超多类数据集 ImageNet-200 上进行实验，衡量本文所提算法的有效性，实验结果如表 2 所列。可以看出，在 SCAN 方法上，将近邻寻找策略更换为本文策略，3 个指标的提升度为：ACC(3.5%)，NMI(1.1%)，ARI(2.7%)。将特征模型训练策略更换为本文训练策略，3 个指标的提升度为：ACC(8.1%)，NMI(3.4%)，ARI(6%)。最后，同时采用特征模型训练策略和语义最近邻寻找策略，3 个指标的提升度为：ACC(12.4%)，NMI(5.1%)，ARI(10.1%)。实验证明了本文方法在超多类数据集上的适用性。

#### 4.2.1　特性模型训练策略对比分析

下文从实例方面对本文所提特征模型训练策略和 SCAN 方法进行对比分析。分别使用 SCAN 和本文特征模型对 ImageNet-200 数据集进行表示，并使用 PaCMAP[48] 方法进行降维可视化。如图 2(a)所示，SCAN 方法没有很好地使 200 个类簇均匀分布，大部分类簇较密集。相较于 SCAN 方法，本文特征模型使 200 个类簇更均匀分布在映射空间中，不同类簇之间距离被拉远，并且同一类簇内实例分布更紧凑，如图 2(b)所示。证明了本文从对齐性和均匀性出发训练特征模型更有利于聚类分析。



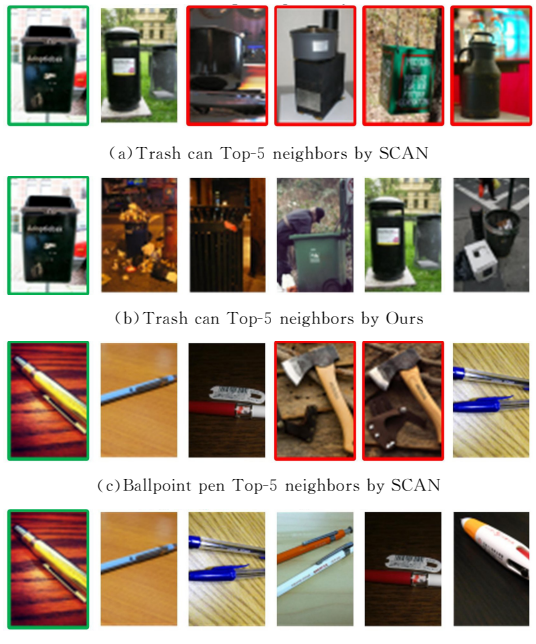
(a)SCANs feature model on ImageNet-200　(b)Ours feature model on ImageNet-200

图 2　SCAN 和本文特征模型对 ImageNet-200 数据集特征表示可视化

Fig. 2　Visualizstion of feature representation of ImageNet-200 dataset on SCAN and our feature model

#### 4.2.2　近邻寻找策略对比分析

从实例方面对比分析近邻寻找策略，分别使用 SCAN 和本文策略，对同一实例寻找 5 个最近邻，结果如图 3 所示，首列图像为原型图像，第一行和第三行为 SCAN 结果，第二行和第四行为本文结果。



(a)Trash can Top-5 neighbors by SCAN

(b)Trash can Top-5 neighbors by Ours

(c)Ballpoint pen Top-5 neighbors by SCAN

(d)Ballpoint pen Top-5 neighbors by Ours

图 3　语义最近邻寻找结果对比(电子版为彩图)

Fig. 3　Seach results comparison of semantic nearest neighbors

如图 3 所示，SCAN 方法会把语义边缘样本做为邻居，第一行 SCAN 对原型图像(垃圾桶)寻找的 5 个最近邻，只有第一个近邻和原型图像属于同一类，其余 4 个近邻(红色框)皆和原型不属于同一类。第三行 SCAN 对原型图像(圆珠笔)寻找 5 个最近邻，其中有两个近邻 3 和 4(红色框)和原型不属于同一类。相较于 SCAN 方法，本文提出的近邻寻找策略去除了类簇边缘语义不一致的噪声数据，如图 3 第二行和

第四行所示，均寻找到了和原型属于同一类的语义近邻，相较于单视图角度直接使用特征相似度寻找近邻的方法，本文从多视图角度并考虑语义差异性寻找最近邻能更好地挖掘到语义一致最近邻信息。

### 4.3 聚类方法性能对比

由于现有聚类方法都是在类别数量较少的数据集上进行实验，本文按照类别数目分多部分进行讨论，对于预训练使用了测试集数据的方法，不参与比较。

#### 4.3.1 类别数较少时的聚类性能对比

在 STL-10，ImageNet-10，ImageNet-Dogs 这 3 个图像基准数据集上进行测试，结果如表 3 所列，可以看出，本文模型在以上 3 个数据集上的 3 个流行评价指标 ACC，NMI，ARI 的性能均是最优的。其中，模型在 ImageNet-10 和 STL-10 上的准确率接近于有监督性能，均达到了 90%以上。

表 3 在 STL-10，ImageNet-10 和 ImageNet-Dogs 上和先进方法比较

Table 3 Comparison with advanced methods on STL-10, ImageNet-10 and ImageNet-Dogs

| Method | ImageNet-10 | | | STL-10 | | | ImageNet-Dogs | | |
|---|---|---|---|---|---|---|---|---|---|
| | *ACC* | *NMI* | *ARI* | *ACC* | *NMI* | *ARI* | *ACC* | *NMI* | *ARI* |
| *K*-Means | 24.1 | 11.9 | 5.7 | 19.2 | 12.5 | 6.1 | 10.5 | 5.5 | 2.0 |
| SC[49] | 27.4 | 15.1 | 7.6 | 15.9 | 9.8 | 4.8 | 11.1 | 3.8 | 1.3 |
| AE[50] | 31.7 | 21.0 | 15.2 | 30.3 | 25.0 | 16.1 | 18.5 | 10.4 | 7.3 |
| VAE[51] | 33.4 | 19.3 | 16.8 | 28.2 | 20.0 | 14.6 | 17.9 | 10.7 | 7.9 |
| JULE[32] | 30.0 | 17.5 | 13.8 | 27.7 | 18.2 | 16.4 | 13.8 | 5.4 | 2.8 |
| DEC[27] | 38.1 | 28.2 | 20.3 | 35.9 | 27.6 | 18.6 | 19.5 | 12.2 | 7.9 |
| DAC[13] | 52.7 | 39.4 | 30.2 | 47.0 | 36.6 | 25.7 | 27.5 | 21.9 | 11.1 |
| IIC[35] | — | — | — | 49.9 | 43.1 | 29.5 | — | — | — |
| DCCM[34] | 71.0 | 60.8 | 55.5 | 48.2 | 37.6 | 26.2 | 38.3 | 32.1 | 18.2 |
| PICA[36] | 85.0 | 78.2 | 73.3 | — | — | — | 32.4 | 33.6 | 17.9 |
| SCAN[37] | — | — | — | 80.9 | 69.8 | 64.6 | — | — | — |
| CC[38] | 89.3 | 85.9 | 82.2 | 85.0 | 76.4 | 72.6 | 42.9 | 44.5 | 27.4 |
| GCC[52] | 90.1 | 84.2 | 82.2 | 78.8 | 68.4 | 63.1 | 52.6 | 49.0 | 36.2 |
| PCL[53] | 90.7 | 84.1 | 82.2 | 41.0 | 71.8 | 67.0 | 41.2 | 44.0 | 29.9 |
| MiCE[54] | — | — | — | 75.2 | 63.5 | 57.5 | 43.9 | 42.3 | 28.6 |
| IDFD[10] | 95.4 | 89.8 | 90.1 | 75.6 | 64.3 | 57.5 | 59.1 | 54.6 | 41.3 |
| BYOL[41] | 93.9 | 86.6 | 87.2 | 82.5 | 71.3 | 65.7 | 69.4 | 63.5 | 54.8 |
| ProPos[11] | 95.6 | 89.6 | 90.6 | 86.7 | 75.8 | 73.7 | 74.5 | 69.2 | 62.7 |
| Ours | **95.8** | **90.1** | **90.7** | **90.8** | **81.8** | **80.6** | **75.3** | **72.7** | **62.8** |

本节参与比较的方法分为 3 部分，对应于本文相关工作对深度聚类的划分：第一部分主要是传统聚类算法（如 *K*-means）和一些将深度神经网络与传统聚类算法相结合的方法（如 DEC[27]），以及交替聚类方法（如 JULE[32] 等）；第二部分主要是利用自监督学习中的对比学习思想，直接学习将图像映射成标签特征的算法（如 CC[38]，SCAN[37]，IDFD[10] 等）；第三部分是相较于在自监督学习中使用对抗信息的第二部分工作，为避免产生类碰撞问题而不使用对抗信息的方法（如 ProPos[11] 等）。对于传统聚类算法 *K*-Means，本文与 SCAN[37]，SPICE[7] 和 ProPos[11] 等已发表论文保持一致，使用在 ImageNet 上的对比学习预训练模型 SimCLR[40] 进行实验，聚类参数集群数量均设置为相对应数据集的真实类簇数目。表 3 中第一部分除 *K*-Means 之外，非传统聚类算法的性能在 SCAN[37]，SPICE[7] 和 ProPos[11] 等已发表论文中的表现均一致，本文也采用和以上论文一致的实验结果；表 3 中第二部分基于对比学习的深度聚类算法结果均来源于对应论文中公开公布的结果；表 3 中第三部分 BYOL[43] 和 ProPos[11] 算法的结果来源于 ProPos[11] 论文中公开公布的实验结果。

2017 年之前的深度聚类算法大多只是简单地将神经网络和传统聚类算法相结合，由于其依赖网络初始特征表示，会容易锁定到低层次的特征，如纹理、颜色、边缘、棱角等。因此，这些方法在对复杂图像进行聚类时，由于缺乏对象层次的高级语义信息而表现较差。与工作 DAC 和 IIC 相比，本模型不需要对任何特定数据集进行微调。本模型优于使用对比学习方法进行图像聚类的工作，并且优于为解决类碰撞问题而不使用对抗信息训练的方法，如 BYOL 和 ProPos。在此之前，ProPos 方法优于 CC，SCAN，IDFD 等使用对抗信息训练的方法。实验结果证明了本文模型在深度图像聚类中捕捉语义信息的优越性。

#### 4.3.2 超多类聚类性能对比

本小节实验数据集采用了 ImageNet-50，ImageNet-100，ImageNet-200 和 ImageNet-1000 这 4 个数据集。将本文方法与先进方法进行比较，结果如表 4 所列，其中 Noun Prior[55] 方法使用了额外数据集进行训练，对于传统聚类算法 *K*-Means，本文与 SCAN[37] 保持一致，使用了在 ImageNet 数据集上的对比学习预训练模型 MOCO[56] 进行实验，聚类参数集群数量均设置为对应数据集的真实类簇数目，表 4 中其他算法的实验数据均来源于对应论文中公开公布的结果。在 ImageNet-50 数据集上，本文模型的 ACC 指标仅次于 Noun Prior。在 ImageNet-100 数据集上，本文模型的 ACC 和 NMI 指标均超过了 Noun Prior，性能最好。在 ImageNet-200 和 ImageNet-1000 数据集上，本文模型的 ACC，NMI，ARI 指标均超过了使用了额外数据集训练的 Noun Prior 方法，3 个指标性能均为最佳。值得注意的是，在 200 类数据集上，本文模型的 ACC 指标比使用了额外数据集训练的方法 Noun Prior 提高了 8.9%。在不使用额外数据训练的情况下，在 200 类数据集上本文方法的 ACC 指标比先进算法提高了 10.6%，在 1000 类数据集上 ACC 指标比先进算法提高了 9.2%。综上所述，本模型不仅超越了使用额外数据集训练的方法 Noun Prior，而且在超多类数据集上准确率取得了较大提升。

表 4　在多类数据集上和先进方法比较

Table 4　Comparison with advanced methods on multi-class datasets

| Method | Extra Train Data | 50 Classes | | | 100 Classes | | | 200 Classes | | | 1 000 Classes | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *ACC* | *NMI* | *ARI* | *ACC* | *NMI* | *ARI* | *ACC* | *NMI* | *ARI* | *ACC* | *NMI* | *ARI* |
| *K*-Means | | 65.9 | 77.5 | 57.9 | 59.7 | 76.1 | 50.8 | 52.5 | 75.5 | 43.2 | — | — | — |
| Scan2(2020) | | 75.1 | 80.5 | 63.5 | 66.2 | 78.7 | 54.4 | 56.3 | 75.7 | 44.1 | 39.9 | 72.0 | 27.5 |
| Scan3(2020) | | 76.8 | 82.2 | 66.1 | 68.9 | 80.8 | 57.6 | 58.1 | 77.2 | 47.0 | — | — | — |
| Noun Prior(2021)[55] | √ | **82.7** | **84.7** | **74.4** | 73.1 | 80.5 | **62.8** | 59.8 | 74.7 | 48.6 | — | — | — |
| Ours | | 79.1 | 81.1 | 61.5 | **75.0** | **81.5** | 53.1 | **68.7** | **80.8** | **54.2** | **49.1** | **75.8** | **29.3** |

4.3.3　ImageNet-1000 数据集聚类性能对比

在 ImageNet-1000 数据集上将先进聚类方法 SA-ART[20]，DSC[29]，SCAN[37]和本文模型进行对比，使用了 3 个评价指标：加权平均精度 $P$(见式(16))、加权平均召回率 $R$(式(17))和加权平均 $F1$ Score(见式(18))。

$$P=\sum_c \frac{L_c}{\sum_c L_c} p_c \tag{16}$$

$$R=\sum_c \frac{L_c}{\sum_c L_c} r_c \tag{17}$$

$$F=\sum_c \frac{L_c}{\sum_c L_c} \frac{2p_c r_c}{p_c+r_c} \tag{18}$$

$$p_c=\frac{L_{c,t}}{L_c} \tag{19}$$

$$r_c=\frac{L_{c,t}}{\sum_c L_{c,t}} \tag{20}$$

其中，$L_c$ 是集群 $c$ 的大小，$p_c$ 和 $r_c$ 分别是 $c$ 的精度和召回率，$t$ 是 $c$ 中大多数样本的真实类别，$L_{c,t}$ 是 $c$ 中 $t$ 类样本的数量。

对算法 SA-ART 的实验设置与本文对 *K*-Means 算法的实验设置保持一致，同使用在 ImageNet 上的对比学习预训练模型 MOCO[56]进行实验。对算法 SA-ART 的参数 $J$(聚类集群数量)使用递增数值分别进行了实验，验证了 SA-ART 算法的优越性。SA-ART 工作提供了 DSC 方法的聚类结果，本文直接将其引用到表 5 中。其中，在使用 MOCO[56]预训练模型对 SA-ART 进行的实验中，随着参数 $J$(聚类集群数量)的增加，查准率 $P$ 也随之增加；当参数 $J$ 和原文保持一致时(349 类)，$P$ 指标低于原文结果，$R$ 指标有较大提升；当设置参数 $J$ 与数据集真实类簇数量一致时(1000 类)，查准率 $P$ 与原文结果相近，在召回率 $R$ 指标方面较原文提高了 9%左右，证明了 SA-ART 算法在保证较高查准率 $P$ 的情况下，仍能稳步提升召回率 $R$ 性能，进而证明了 SA-ART 在处理大规模数据集时的优越性。

表 5　在 ImageNet-1000 数据集上和先进方法比较

Table 5　Comparison with advanced methods on ImageNet-1000 dataset

| Dataset | Method | Clusters($J$) | $P$ | $R$ | $F$ |
|---|---|---|---|---|---|
| ImageNet-1000 | SCAN | 1 000 | 0.389 | 0.399 | 0.394 |
| | DSC-Nets | 400 | 0.413±0.039 | 0.193±0.028 | 0.249±0.029 |
| | SA-ART | 349 | 0.449±0.033 | 0.129±0.036 | 0.252±0.037 |
| | SA-ART$_{moco}$ | 349 | 0.309 | 0.446 | 0.288 |
| | SA-ART$_{moco}$ | 616 | 0.383 | 0.338 | 0.287 |
| | SA-ART$_{moco}$ | 1 000 | 0.453 | 0.259 | 0.267 |
| | Ours | 1 000 | **0.520** | **0.492** | **0.506** |

使用上述 3 个指标衡量 SCAN 方法和本文模型，结果如表 5 所列。在加权平均精度 $P$ 指标方面，2019 年的工作 SA-ART 超越了 2020 年使用大规模数据集进行预训练的深度聚类工作 SCAN，说明了 SA-ART 在省时间、省算力、省步骤的同时还可以获得很高的精度，证明了 SA-ART 处理大规模稀疏数据集的优越性。由于 SA-ART 对集群的局部突出分布进行建模，所以查准率 $P$ 指标较高，在召回率 $R$ 方面的表现略低但合理，本文模型精度 $P$ 略高于 SA-ART 方法，在 $P$，$R$，$F$ 这 3 个指标上较基础方法 SCAN 有较大提升，说明了本文所提方法的有效性。

**结束语**　本文提出了一个适用于超多类图像数据的深度图像聚类模型，可满足实际应用过程中不同场景下的图像聚类无监督自适应。本文提出的算法模型主要有以下步骤：首先，提升对比学习方法的对齐性和均匀性优化特征模型，使类簇均匀分布在映射空间；其次，基于语义相似性原则多视角挖掘近邻关系，进一步改进现有的自监督信息构建策略，可以有效去除语义不一致噪声；最后，基于最大化分离异类特征的损失函数高效利用自监督信息，以正确引导超多类聚类模型训练。本文提出的超多类聚类模型可有效缓解类簇密集问题，提高聚类算法在超多类图像数据上的适用性，并在超多类数据集上大大缩小无监督和有监督分类之间的差距。

未来研究主要聚焦于以下问题：使用对比学习方法时避免引入类碰撞问题，防止模型退化；提出适用于超多类数据的相似性度量方法以提升超多类聚类性能。

## 参 考 文 献


[1] REN X T, ZHAO J J, QIANG Y, et al. Lung Cancer Subtype Recognition with Unsupervised Learning Combining Paired Learning and Image Clustering[J]. Computer Science, 2020, 47(10): 200-206.

[2] ZHU Z Q, GENG H J, QIAN Y H. Line-Segment Clustering Algorithm for Chemical Structure[J]. Computer Science, 2022, 49(5): 113-119.

[3] LU K, DING Z, GE S. Sparse-Representation-Based Graph Embedding for Traffic Sign Recognition[J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(4): 1515-1524.

[4] LIU X, WANG G Y, LUO X B. Multi-granularity Clustering of

Remote Sensing Image Based on Gaussian Cloud Transformation [J]. Computer Science, 2017, 44(9): 23-27, 52.

[5] HUANG M, LI T, WEN X, et al. Induction Motor Parameter Identification Incorporating Image Recognition and Cluster Analysis[J]. Journal of Chongqing University of Technology(Natural Science), 2022, 36(9): 195-201.

[6] PARK S, HAN S, KIM S, et al. Improving Unsupervised Image Clustering With Robust Learning [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE Press, 2021: 15750-15758.

[7] NIU C, SHAN H, WANG G. SPICE: Semantic Pseudo-labeling for Image Clustering[J]. arXiv: 2103.09382, 2021.

[8] COATES A, LEE H, NG A Y. An Analysis of Single-Layer Networks in Unsupervised Feature Learning[C]// Proceedings of the fourteenth International Conference on Artificial Intelligence and Statistics. Lauderdale: JMLR Press, 2011: 215-223.

[9] KRIZHEVSKY A. Learning Multiple Layers of Features from Tiny Images[R]. University of Toronto, Technical Report, 2009.

[10] TAO Y, TAKAGI K, NAKATA K. Clustering-friendly Representation Learning via Instance Discrimination and Feature Decorrelation[C]// Proceedings of the International Conference on Learning Representations. ICLR Press, 2021.

[11] HUANG Z, CHEN J, ZHANG J, et al. Learning Representation for Clustering via Prototype Scattering and Positive Sampling [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7509-7524.

[12] DESHMUKH A A, REGATTI J R, MANAVOGLU E, et al. Representation Learning for Clustering via Building Consensus [J]. Springer Machine Learning Journal, 2022, 111(12): 4601-4638.

[13] CHANG J, WANG L, MENG G, et al. Deep Adaptive Image Clustering[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Venice: IEEE Press, 2017: 5880-5888.

[14] DENG J, DONG W, SOCHER R, et al. ImageNet: A Large-Scale Hierarchical Image Database[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Miami: IEEE Press, 2009: 248-255.

[15] LI F, QIAN Y, WANG J, et al. Clustering method based on sample's stability[J]. SCIENTIA SINICA Informationis, 2020, 50(8): 1239-1254.

[16] WANG J, QIAN Y, LI F, et al. Support Vector Machine with Eliminating the Random Consistency[J]. Journal of Computer Research and Development, 2020, 57(8): 1581-1593.

[17] LI F, QIAN Y, WANG J, et al. Clustering ensemble based on sample's stability[J]. Artificial Intelligence, 2019, 273: 37-55.

[18] WANG J, QIAN Y, LI F, et al. Generalization Performance of Pure Accuracy and Its Application in Selective Ensemble Learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1798-1816.

[19] ZHANG H, ZHANG J, HUANG J. Multi-label Image Classification Model Based on Graph Attention Network[J]. Journal of Chongqing Technology and Business University(Natural Science Edition), 2022, 39(1): 34-41.

[20] MENG L, TAN A H, MIAO C. Salience-aware adaptive resonance theory for large-scale sparse data clustering[J]. Neural Networks, 2019, 120: 143-157.

[21] MENG L, TAN A H, WUNSCH D C. Vigilance adaptation in adaptive resonance theory[C]// Proceedings of the the 2013 International Joint Conference on Neural Networks. IEEE Press, 2013: 1-7.

[22] MENG L, TAN A H, XU D. Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-Clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(9): 2293-2306.

[23] MENG L, TAN A H, WUNSCH D C. Adaptive Scaling of Cluster Boundaries for Large-Scale Social Media Data Clustering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 27(12): 2656-2669.

[24] CARPENTER G A, GROSSBERG S. A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine[J]. Computer Vision, Graphics, and Image Processing, 1987, 37(1): 54-115.

[25] CARPENTER G A, GROSSBERG S, ROSEN D B. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system[J]. Neural Networks, 1991, 4(6): 759-771.

[26] HUANG P, HUANG Y, WANG W, et al. Deep Embedding Network for Clustering[C]// Proceedings of the 22nd International Conference on Pattern Recognition. Washington: IEEE Press, 2014: 1532-1537.

[27] XIE J, GIRSHICK R, FARHADI A. Unsupervised Deep Embedding for Clustering Analysis[C]// Proceedings of the International Conference On Machine Learning. New York: ACM Press, 2016: 478-487.

[28] YANG B, FU X, SIDIROPOULOS N D, et al. Towards *K*-means-friendly Spaces: Simultaneous Deep Learning and Clustering[C]// Proceedings of the International Conference On Machine Learning. New York: ACM Press, 2017: 3861-3870.

[29] JI P, ZHANG T, LI H, et al. Deep Subspace Clustering Networks[C]// Proceedings of the Annual Conference on Neural Information Processing Systems. MIT Press, 2017.

[30] ZHOU P, HOU Y, FENG J. Deep Adversarial Subspace Clustering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Press, 2018: 1596-1604.

[31] DIZAJI K G, HERANDI A, DENG C, et al. Deep Clustering via Joint Convolutional Autoencoder Embedding and Relative Entropy Minimization[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Venice: IEEE Press, 2017: 5747-5756.

[32] YANG J, PARIKH D, BATRA D. Joint Unsupervised Learning of Deep Representations and Image Clusters[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Press, 2016: 5147-5156.

[33] CARON M, BOJANOWSKI P, JOULIN A, et al. Deep Clustering for Unsupervised Learning of Visual Features[C]// Pro-

ceedings of the European conference on computer vision. Cham: Springer International Publishing, 2018: 139-156.

[34] WU J, LONG K, WANG F, et al. Deep Comprehensive Correlation Mining for Image Clustering [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE Press, 2019: 8149-8158.

[35] JI X, VEDALDI A, HENRIQUES J. Invariant Information Clustering for Unsupervised Image Classification and Segmentation [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE Press, 2019: 9864-9873.

[36] HUANG J, GONG S, ZHU X. Deep Semantic Clustering by Partition Confidence Maximisation [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE Press, 2020: 8846-8855.

[37] VAN GANSBEKE W, VANDENHENDE S, GEORGOULIS S, et al. SCAN: Learning to Classify Images Without Labels[C] // Proceedings of the European conference on computer vision. Cham: Springer International Publishing, 2020: 268-285.

[38] LI Y, HU P, LIU Z, et al. Contrastive Clustering [C] // Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2021: 8547-8555.

[39] HE K, FAN H, WU Y, et al. Momentum Contrast for Unsupervised Visual Representation Learning[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE Press, 2020: 9729-9738.

[40] CHEN T, KORNBLITH S, NOROUZI M, et al. A Simple Framework for Contrastive Learning of Visual Representations [C] // Proceedings of the International Conference on Machine Learning. New York: ACM Press, 2020: 1597-1607.

[41] GRILL J B, STRUB F, ALTCHÉ F, et al. Bootstrap your own latent: A new approach to self-supervised Learning [C] // Proceedings of the Annual Conference on Neural Information Processing Systems. MIT Press, 2020: 21271-21284.

[42] CHEN X, HE K. Exploring Simple Siamese Representation Learning[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE Press, 2021: 15750-15758.

[43] WANG T, ISOLA P. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere [C] // Proceedings of the International Conference On Machine Learning. New York: ACM Press, 2020: 9929-9939.

[44] DEVRIES T, TAYLOR G W. Improved Regularization of Convolutional Neural Networks with Cutout [J]. arXiv: 1708.04552, 2017.

[45] CUBUK E D, ZOPH B, SHLENS J, et al. Randaugment: Practical automated data augmentation with a reduced search space [C] // Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. IEEE Press, 2020: 3008-3017.

[46] WANG F, LIU H. Understanding the Behaviour of Contrastive Loss[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE Press, 2021: 2495-2504.

[47] SOHN K, BERTHELOT D, LI C L, et al. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence[J]. Advances in neural information processing systems, 2020, 33: 596-608.

[48] WANG Y, HUANG H, RUDIN C, et al. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization[J]. Journal of Machine Learning Research, 2021, 22(201): 1-73.

[49] ZELNIK-MANOR L, PERONA P. Self-Tuning Spectral Clustering[C] // Proceedings of the Annual Conference on Neural Information Processing Systems. MIT Press, 2004.

[50] BENGIO Y, LAMBLIN P, POPOVICI D, et al. Greedy Layer-Wise Training of Deep Networks[C] // Proceedings of the Annual Conference on Neural Information Processing Systems. MIT Press, 2006.

[51] KINGMA D P, WELLING M. Auto-Encoding Variational Bayes [C] // Proceedings of the International Conference on Learning Representations. ICLR Press, 2014.

[52] ZHONG H, WU J, CHEN C, et al. Graph Contrastive Clustering [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE Press, 2021: 9204-9213.

[53] LI J, ZHOU P, XIONG C, et al. Prototypical Contrastive Learning of Unsupervised Representations[C] // Proceedings of the International Conference on Learning Representations. ICLR Press, 2021.

[54] TSAI T W, LI C, ZHU J. MiCE: Mixture of Contrastive Experts for Unsupervised Image Clustering[C] // Proceedings of the International Conference on Learning Representations. ICLR Press, 2021.

[55] COHEN N, HOSHEN Y. The Single-Noun Prior for Image Clustering[J]. arXiv: 2104.03952, 2021.

[56] CHEN X, FAN H, GIRSHICK R, et al. Improved Baselines with Momentum Contrastive Learning[J]. arXiv: 2003.04297, 2020.

**HU Shen**, born in 1997, postgraduate, is a student member of China Computer Federation. His main research interests include self-supervised learning and super multi-class image clustering.

**QIAN Yuhua**, born in 1976, Ph.D, professor, is a member of China Computer Federation. His main research interests include artificial intelligence, big data, machine learning and data mining.

（责任编辑：何杨）