图像聚类方法 SCAN 的复现与改进

摘要

图像聚类是一个重要且具有挑战的计算机视觉任务，已经有许多无监督的深度聚类方法被提出。本文主要针对 2020 年提出的一种无标签图像分类方法 SCAN 进行实验复现，并利用结构学习的思想对损失函数进行改进，最终复现以及改进的实验结果在三个小规模数据集上都能取得较好的实验结果。

关键词:；图像聚类；无监督；结构学习

1 引言

图像分类是计算机视觉领域的重要分支，目标是从预定义的类别集合中分配语义标签给相应的图像，例如，一张图像可以对应猫、狗、汽车、飞机等具体的小类，也可以抽象到动物、交通工具等大类。

传统的图像分类方法通常需要在带有人工标注语义标签的大规模图像数据集上进行有监督的训练，使得模型学习到对于预定类更加具判别性的特征表示，从而实现良好的分类效果。而在真实语义标签缺失，甚至语义类以及类的数目都不是先验已知的情况下，理想的目标是将图像分组到不同的簇中，在同一簇中的图像属于同一个或者相似的语义类，而不同簇中的图像在语义上是不相同的，一种可能且有效的解决方法是使用无监督的图像分类方法，即图像聚类，根据图像之间的相似性来将它们分组。与传统的有监督学习方法不同，图像聚类方法不需要人工标注的语义标签，而是通过图像自身的特征进行分类。这种方法可以自动地提取图像中的内在结构和模式，学习到图像的高质量表征并定义类别集合的伪标签，从而实现对图像的分类。

但是图像聚类也存在着一些挑战。首先，由于缺少语义标签作为监督信息，模型需要具备从无标签的数据中学习有用的特征的能力。其次，由于没有明确的类别标签，模型需要自行确定图像的类别。目前已经有许多无监督的图像分类方法提出，比如变分自编码器、生成对抗网络以及掩码自编码器等，先从无标签的数据中学习有用的特征表示，并通过重构输入数据或生成类似数据来优化表示的质量，然后使用聚类算法对学习到的特征表示进行分类，从而进一步实现对图像的分类。

本文主要针对 2020 年提出的一种图像聚类方法 SCAN [19] 进行实验分析与改进，这是一种将特征学习与聚类解耦的两阶段方法，在结合表征学习和端到端学习两者优势的同时，克服了两者存在的不足，当时在几个数据集上的实验结果大大优于先前的方法。

2 相关工作

2.1 深度聚类

深度聚类的早期工作通常简单地将特征学习与浅层聚类相结合，例如，一些方法将传统的聚类算法（如 k-means [18,22,24]、子空间聚类 [12,23] 以及谱聚类 [17]）与堆叠自编码器或者卷积神经网络相结合。但是上述方法更适用于具有特定分布的数据，而不是分布更具广泛性的数据。因此，这些方法在许多实际应用中对于数据的处理存在挑战性，并且通常需要传统聚类算法进行初始化，或者进一步的处理，才能获得最终的分类结果，对聚类的性能表现有一定限制。

为了实现表征学习与聚类的同时优化，一些研究将图像表征直接映射为标签。例如，一些方法通过最大化原始图像与增强之间标签的互信息 [7,15,25]，或者最大化样本及其近邻样本之间聚类指示的似然估计 [6,7]，以端到端的形式将图像用分类模型直接映射为聚类指示概率并得到标签。但是这类方法对初始化参数敏感或者受低质量初始嵌入特征的影响，聚类性能往往不佳。

为了克服上述不足，一些研究通过以两阶段的训练方式来进行聚类。在第一个阶段，这些方法通常借助能够学习到高质量图像特征的借口任务来进行特征学习 [3,11]，然后通过最大化紧邻样本之间聚类概率指示相似度的方式来训练分类模型 [2,19,21,26]。在下一阶段，通常选择聚类指示概率逼近 one-hot 向量的置信样本，并用 one-hot 形式的伪标签对这些样本进行标记，最后，利用标记过的样本 [19,21] 来对模型进行进一步的训练。

2.2 表征学习

图像表征学习的早期工作使用手工制作的特征，比如尺度不变特征变换 [16] 和定向梯度直方图 [5]。之后的一些工作采用深度神经网络，例如自编码器 [20] 和生成对抗学习 [8]，来提取图像的有效特征。最近，随着自监督学习的兴起，表征学习得到了很大的发展。自监督学习旨在构建一系列借口任务，并利用图像的不同先验来探索数据样本的内在分布，已经成为一种越来越流行的深度图像表征学习方法。相关工作主要集中在借口任务上，基于图像变换前后标签不变性的特征，通过对输入图像应用变换来学习良好的特征表示。对比学习作为图像处理学习方法的主要组成部分之一，通常优化对比损失（比如最大化不同视图之间的互信息 [1] 或者动量对比学习 [11]）来拉近邻居和推远非邻居，还有一些对比学习方法通过结合数据增强策略（如 SimCLR [3]、BYOL [9]、PCL [14] 和 HCSC [10]），在图像表征方面取得了很大成功。

3 本文方法

3.1 本文方法概述

SCAN [19] 是一种两阶段的图像聚类方法，即基于最近邻的语义聚类，它利用了表征学习和端到端学习方法的优点，同时解决了两者的不足。在第一阶段，通过借口任务学习图像的特征表示，并依据特征的相似性挖掘每个图像的最近邻；在第二个阶段，将获得的语义上

有意义的最近邻作为先验集成到可学习的聚类方法中。下面将从语义聚类表征学习、语义聚类损失、自标记微调三个部分来详细介绍实验方法。

3.2 语义聚类表征学习

在有监督学习的设置中，每个样本可以通过真实语义标签与正确的类别相关联，具体来说，图像集合 $\mathcal{D} = \{X_1, \ldots, X_{|D|}\}$ 和语义类别 $\mathcal{C}$ 通常可以通过最小化交叉熵损失来学习。然而，在缺少真实语义标签时，就需要定义一个先验来获得样本可能属于同一类或是其他类的聚类概率指示。

端到端学习的方法利用卷积神经网络的架构作为先验，或者依据图像与其增强之间的一致性来定义聚类簇。在这些情形下，聚类学习已知是对网络初始化敏感的，而且在训练的开始阶段，网络还无法从图像中提取到高质量的特征信息，因此很容易学习到对于语义聚类次优的低质量图像特征。为了克服这一不足，SCAN 使用表征学习来获得对于语义聚类更好的先验。

在图像表征学习中，通过借口任务 τ 以自监督的方式学习嵌入函数 Φ_θ，将图像映射到特征表示，其中 Φ_θ 由权重为 θ 的神经网络参数化。某些借口任务基于特定的图像变换，会导致学习到的特征表示与所采用的变换协变，即不同的仿射变换可能导致 Φ_θ 产生不同的输出预测。因此，SCAN 为借口任务 τ 施加最小化图像 X_i 及其增强 $T[X_i]$ 的特征表示之间的距离，从而使得特征表示在图像变换下尽量不变。此外，在满足约束的情况，依据表征学习检索图像的最近邻可以发现特征相似的图像通常在语义上相似，因此，认为表征学习中的借口任务可用于获得语义上有意义的高质量特征。

3.3 语义聚类损失

依据图像最近邻通常属于同一语义类这一发现，SCAN 将由借口任务　获得的最近邻作为语义聚类的先验。通过表征学习，在无标签数据集 $\mathcal{D}$ 上训练模型 Φ_θ 来解决借口任务 τ，然后对于每一样本 X_i，在特征空间 Φ_θ 中检索其对应的 K 个最近邻，记为 $\mathcal{N}_{X_i}$。在聚类阶段，训练一个聚类函数 Φ_η，将 X_i 及其所有近邻 $\mathcal{N}_{X_i}$ 分配到同一类别，其中 Φ_η 为一权重 η 为的神经网络。Φ_η 最终的输出经过 softmax 处理得到类别集 $\mathcal{C}$ 的一个软分配。样本 X_i 被分配到类别 c 的概率表示为 $\Phi_\eta^c(X_i)$，最终的目标函数为式（1）所示：

$$\begin{aligned} \Lambda = -\frac{1}{|\mathcal{D}|} \sum_{X \in \mathcal{D}} \sum_{k \in \mathcal{N}_X} \log \langle \Phi_\eta(X), \Phi_\eta(k) \rangle + \lambda \sum_{c \in \mathcal{C}} \Phi_\eta'^c \log \Phi_\eta'^c, \\ \text{with } \Phi_\eta'^c = \frac{1}{|\mathcal{D}|} \sum_{X \in \mathcal{D}} \Phi_\eta^c(X). \end{aligned} \tag{1}$$

其中 $\langle \cdot \rangle$ 表示向量点乘。目标函数中第一个损失项用于约束 Φ_η 对 X_i 及其 K 个近邻 N_{X_i} 做出一致性预测，而第二个损失项，即熵项，用于样本的均匀分配，避免 Φ_η 将所有样本分配为同一类。

3.4 自标记微调

实际上，图像最近邻属于同一语义类这一假设只考虑到通常情况，在 Cifar10 数据集中，即使是前 5 个最近邻对，也只有 75% 左右真正属于同一类别，因而在 $K \geq 1$ 的情况下，不

可避免地会引入一些不合格的正对，对聚类的准确性产生影响。注意到在样本针对某一类别有着高置信度的预测时，即 $p_{max} \approx 1$，样本的真实类别往往与之对应，因此，SCAN 提出一种自标记方法来利用这些高置信度样本，并对引入的噪声最近邻造成的错误进行纠正。具体来说，在训练过程，通过预定义的置信度阈值来筛选出具有高置信度的样本，将这些样本视为预测类别对应的“原型”，并由此生成伪标签分配给置信度不高的其他样本，实现模型的自纠正，提高聚类分配的准确性。

4 复现细节

4.1 与已有开源代码对比

SCAN 方法的代码开源，网址为github.com/wvangansbeke/Unsupervised-Classification，实验主要依赖 3 个 python 文件，需要按顺序依次运行，其中 *simclr.py* 通过表征学习中的借口任务获取所有图像的 K 近邻集合，*scan.py* 使用近邻作为先验进行聚类，*selflabel.py* 进行自标记微调，提高聚类的准确性。

与已有开源代码对比，由于是对聚类的损失函数进行改进，因此主要在 *losses* 文件夹的 *losses.py* 中添加新的损失函数 *StruturedLearning*，以及在 *utils* 文件夹中的 *train_utils.py* 添加新的训练函数 *SSL_train*，然后在 *scan.py* 中调用。

4.2 数据集介绍

实验分别在 Cifar10 [13]、Cifar100-20 [13] 以及 STL10 [4] 这 3 个不同的基准数据集上进行评测，基准数据集的主要信息如表 1 所示。

其中 Cifar10 数据集是一个由 Hinton 的学生整理的一个用于识别普适物体的小型数据集，一共包含 10 个类别的 RGB 彩色图片，分别为飞机、汽车、鸟类、猫、鹿、狗、蛙类、马、船以及卡车，每个类别有 6000 张图像；Cifar100-20 数据集和 Cifar10 数据集类似, 也是用于机器视觉领域图像分类的一个图像数据集, 共有 100 个具体的类别, 对应图像的细粒度标签, 每个类别有 600 张图像。这 100 个类可以分为 20 个超类, 对应图像的粗粒度标签；STL10 数据集由 Adam 在 2011 年发布，是一个用于开发无监督特征学习、深度学习、自监督学习算法的图像识别数据集。其灵感来自 Cifar10 数据集，每个类别的标注图像数量相比 Cifar10 中的要少，但提供了大量的无标注图像来做无监督预训练，其主要的挑战是利用无标签图像来构建先验知识。

表 1. 基准数据集的主要信息

数据集	图像大小	训练集	测试集	类别数目
STL10	96×96	$5,000$	$8,000$	10
Cifar10	32×32	$50,000$	$10,000$	10
Cifar100-20	32×32	$50,000$	$10,000$	20

4.3 创新点

对 SCAN 的改进主要针对语义聚类损失部分，重新设计聚类的损失函数，使用结构重建的方法来更好地保留图像的邻域结构。

定义样本 x_i 及其邻域集合 $\mathcal{N}(x_i)$，对于任意的 $x' \in \mathcal{N}(x_i)$，满足 $\|x' - x_i\| \dot{\leq} r$，其中 r 为一给定的参数。定义 $\mathcal{T}$ 为用于数据增强的一些变换集合，$\mathcal{B}(x_i)$ 为 x_i 增强的邻域集合，即对于任意 $x_{\mathcal{B}} \in \mathcal{B}(x_i)$ 的以及任意的变换 $T \in \mathcal{T}$，有 $\|x_{\mathcal{B}} - T(x_i)\| \dot{\leq} r_{\mathcal{B}}$ 满足。综上定义可以推导得到，对于任意的变换 $T \in \mathcal{T}$，有

$$\begin{aligned} \|T(x_i) - T(x')\| &= \|(T(x_i) - x_i) + (x_i - x') + (x' - T(x'))\| \\ &\leq \|T(x_i) - x_i\| + \|x_i - x'\| + \|x' - T(x')\| \\ &\leq r + 2r_{\mathcal{B}}. \end{aligned} \tag{2}$$

不失一般性，我们假定 $r_{\mathcal{B}} << r$，那么可以认为 $T(x_i)$ 也在 $T(x')$ 的邻域中，即邻域结构在某些数据增强下具有保留性。

映射函数用 $\Phi(x_i; \varphi)$ 表示，输出为 d 维特征表示 $\mathbf{z}_i$，模型用 $f(x_i; \theta)$ 表示，输出为软分配的 c 维向量 $\mathbf{q}_i$。从上述假定出发，设计的损失函数如下：

$$\mathcal{L}_{sl}(f(\mathcal{D}; \theta)) = \sum_{i=1}^{n} \mathcal{L}_{cl}(f(x_i; \theta)) + \eta_{cc}\mathcal{L}_{cc}(f(\mathcal{D}; \theta)) + \lambda_b \mathcal{L}_b(f(\mathcal{D}; \theta)). \tag{3}$$

下面对各损失项，即一致性损失、对比损失以及熵损失，依次进行介绍。

4.3.1 一致性损失

从 $\mathcal{T}$ 中随机选取两个变换 $T^{(a)}$ 和 $T^{(b)}$，对 x_i 应用上述变换得到 $x_i^{(a)} = T^{(a)}(x_i)$ 和 $x_i^{(b)} = T^{(b)}(x_i)$，对应的输出分别为 $\mathbf{q}_i^{(a)} = f^{(a)}(x_i; \theta)$ 和 $\mathbf{q}_i^{(b)} = f^{(b)}(x_i; \theta)$。一致性损失计算如下：

$$\mathcal{L}_{cl}\big(f(x_i; \theta)\big) = -\sum_{j=rn(\mathcal{N}(x_i))\big)} w_{ij} \log\left(\left(\mathbf{q}_i^{(a)}\right)^T \mathbf{q}_j^{(b)}\right) \tag{4}$$

其中 $rn(\cdot)$ 表示从给定集合中随机选取，相似度矩阵 $\mathbf{W}$ 计算如下：

$$w_{ij} = \begin{cases} \exp\left(\frac{\mathbf{z}_i^T \mathbf{z}_j}{\tau_c}\right) & j \in \mathcal{N}(x_i) \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

其中 τ_c 为用于调整相似度的温度系数。

4.3.2 对比损失

假定一个批次中样本数为 B，$\mathbf{Q}^{(a)} \in \mathbb{R}^{B \times c}$ 为所有样本在数据增强 $T^{(a)}$ 下得到的聚类分配概率矩阵，而 $\widetilde{\mathbf{Q}}^{(b)} \in \mathbb{R}^{B \times c}$ 为所有样本的随机近邻在数据增强 $T^{(b)}$ 下得到的聚类分配概率矩阵，令 $\mathbf{P}^{(a)} = (\mathbf{Q}^{(a)})^T$ 以及 $\widetilde{\mathbf{P}}^{(b)} = (\widetilde{\mathbf{Q}}^{(b)})^T$，那么 $\mathbf{p}_l^{(a)}$ 和 $\widetilde{\mathbf{q}}_l^{(b)}$ 对应第 l 个类别的概率分配。具体来说，对比损失计算如下：

$$l_{cc}^{(a,b)} = -\sum_{i=1}^{m}\sum_{l=1}^{c} \log \frac{exp((\mathbf{p}_l^{(a)})^T \tilde{\mathbf{p}}_l^{(b)} / \tau_{cc})}{\sum_{j=1, j\neq l}^{c} \gamma_{ij} exp((\mathbf{p}_l^{(a)})^T \tilde{\mathbf{p}}_j^{(b)} / \tau_{cc})}. \tag{6}$$

其中 m 为批次数，τ_{cc} 为一温度系数，γ_{ij} 为 $\mathbf{p}_l^{(a)}$ 和 $\widetilde{\mathbf{q}}_l^{(b)}$ 两者的权重，用于推远非近邻。最终的对比损失如下：

$$\mathcal{L}_{cc}(f(\mathcal{D};\theta)) = \frac{1}{2mc}(l_{cc}^{(a,b)} + l_{cc}^{(b,a)}). \tag{7}$$

4.3.3 熵损失

使用负熵损失，计算如下：

$$\mathcal{L}_b(f(\mathcal{D};\theta)) = \sum_{l=1}^{c} \bar{q}_l \log \bar{q}_l, \quad \text{with } \bar{q}_l = \frac{\sum_{i=1}^{n} q_{il}}{n}. \tag{8}$$

5 实验结果分析

实验使用准确度、标准化互信息以及调整兰德指数三类指标进行评估，分别用 ACC、NMI 以及 ARI 表示。关于参数设置，三个数据集的训练批次为 100，批次大小为 128，温度系数 τ_c=8，τ_{cc}=1，权重参数 η_{cc}=1，λ_b=5，优化器使用 Adam，学习率和权重衰减均为 10^{-4}。

由于自标记微调部分相当于是对聚类阶段进行优化，因此原始的实验结果分为两组，一组在聚类步骤之后，一组在自标记步骤之后，分别用 SCAN* 以及 SCAN† 表示，这里我用 SCAN^*_{re} 和 $\text{SCAN}^\dagger_{re}$ 分别表示复现得到的聚类和自标记实验结果，具体如表 2 所示。

表 2. 原论文与复现实验结果

数据集	STL10			Cifar10			Cifar100-20		
评估指标	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
SCAN* (Avg±Std)	$75.5_{\pm 2.0}$	$65.4_{\pm 1.2}$	$59.0_{\pm 1.6}$	$81.8_{\pm 0.3}$	$71.2_{\pm 0.4}$	$66.5_{\pm 0.4}$	$42.2_{\pm 3.0}$	$44.1_{\pm 1.0}$	$26.7_{\pm 1.3}$
SCAN† (Avg±Std)	$76.7_{\pm 1.9}$	$68.0_{\pm 1.2}$	$61.6_{\pm 1.8}$	$87.6_{\pm 0.4}$	$78.7_{\pm 0.5}$	$75.8_{\pm 0.7}$	$45.9_{\pm 2.7}$	$46.8_{\pm 1.3}$	$30.1_{\pm 2.1}$
SCAN† (Best) [19]	80.9	69.8	64.6	88.3	79.7	77.2	50.7	48.6	33.3
SCAN^*_{re}	74.6	63.7	57.2	80.4	68.2	63.7	39.7	41.2	25.1
$\text{SCAN}^\dagger_{re}$	77.8	65.2	60.1	85.1	74.9	71.9	46.7	47.6	31.0

可以看到复现的实验结果和原论文有着一定差距，但基本上在给定的标准差范围内；此外，自标记微调能够有效提高聚类的准确性，平均能提升四五个点左右。

本文对 SCAN 做出的改进是针对聚类阶段，因此改进后应该对聚类和自标记两组结果都有提升效果，用 SCAN^*_{new} 和 $\text{SCAN}^\dagger_{new}$ 分别表示复现得到的聚类和自标记实验结果，如表 3 所示，得到的结果与预期相符。

此外，还针对权重系数 η_{cc}，λ_b 在 STL10 数据集上作了参数敏感性分析，如图 1 所示，在 λ_b 值较大时，聚类结果会表现得更好；在 λ_b 值较小时，选择较大的 η_{cc} 值有助于提高准确度；最优的组合是 η_{cc}=1，λ_b=5。

表 3. 改进后的实验结果

数据集	STL10			Cifar10			Cifar100-20		
评估指标	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
SCAN^{*}_{new}	75.6	66.2	60.3	82.6	71.5	67.4	41.7	41.7	25.2
$\text{SCAN}^{\dagger}_{new}$	78.9	68.1	64.8	86.5	77.4	74.7	48.9	49.2	32.7

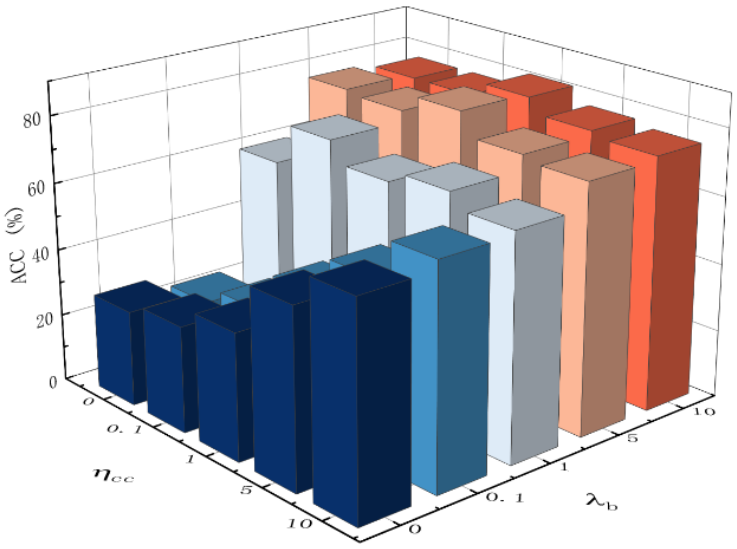

图 1. η_{cc}，λ_b 在 STL10 数据集上的敏感性分析

图 2、3、4 所示依次为 STL10、Cifar10 以及 Cifar100-20 在改进后的 SCAN 算法下得到的混淆矩阵的图示结果，竖轴表示真实类别，横轴代表预测类别。可以看到，大部分的错误存在于较难区分的类别之间，比如说猫、狗，而对于 Cifar100-20，由于超类存在一定的模糊性，结果并不算好，因此这也是一个可以改进的点。

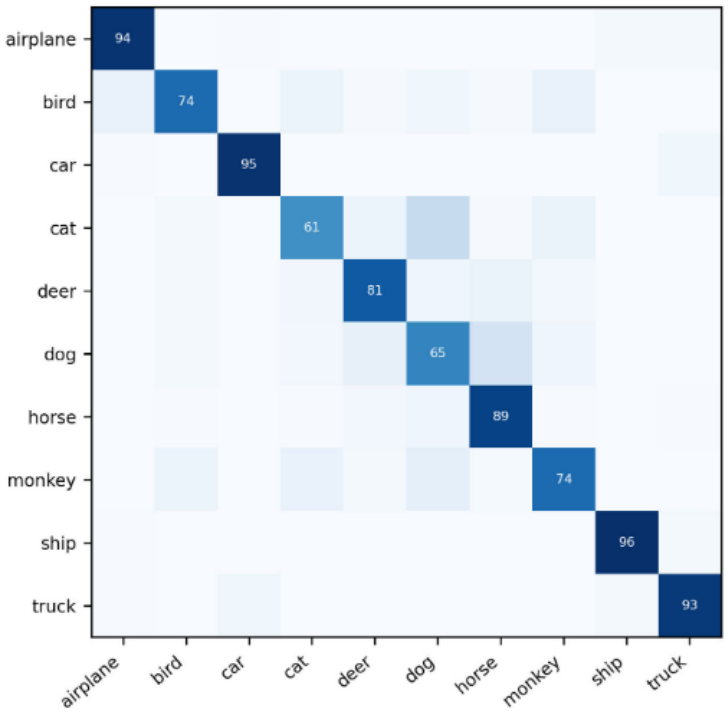

图 2. STL10 的混淆矩阵

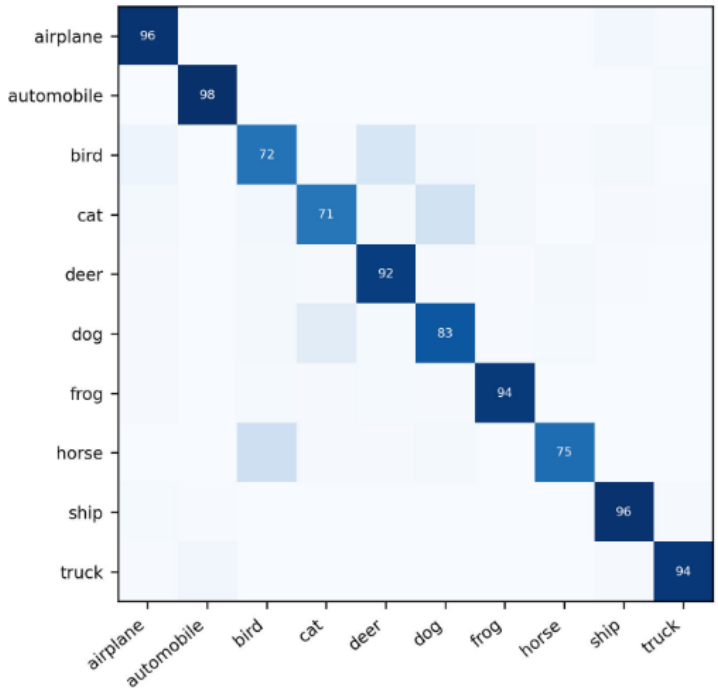

图 3. Cifar10 的混淆矩阵

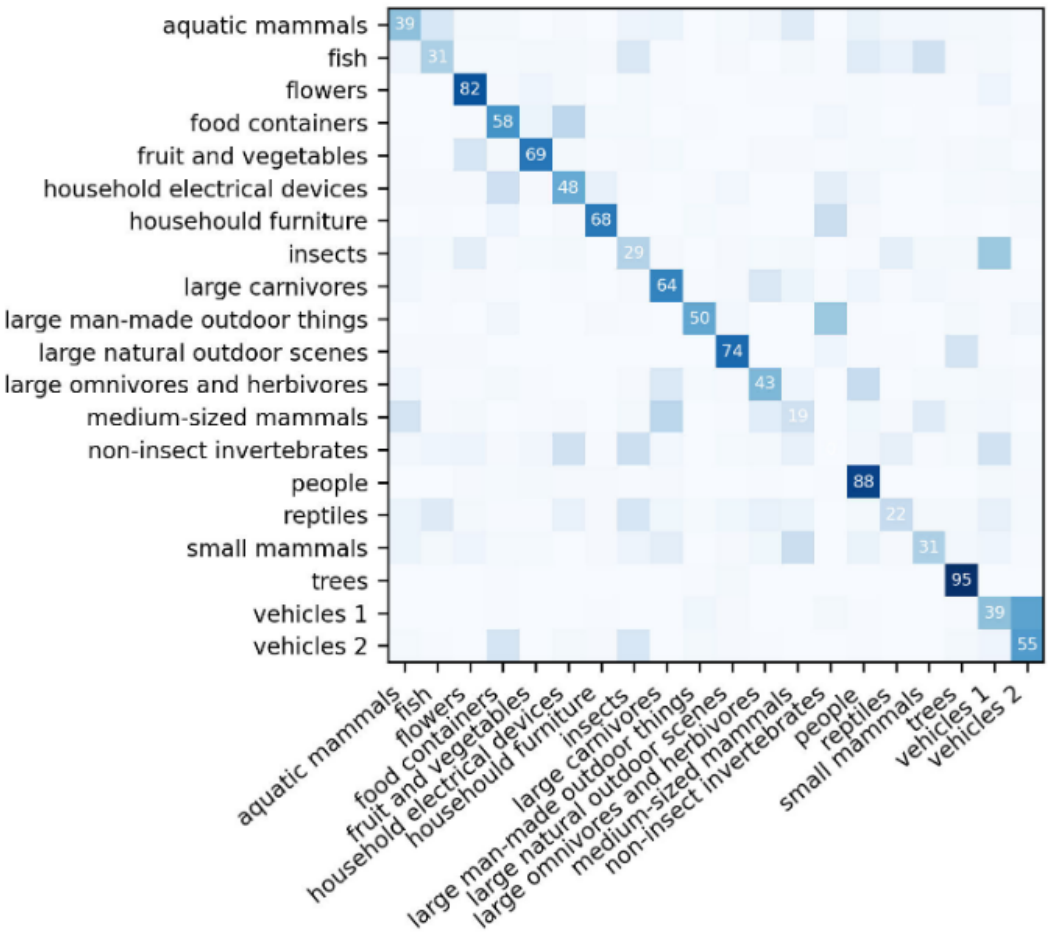

图 4. Cifar100-20 的混淆矩阵

6 总结与展望

本文针对 2020 年提出的图像聚类方法 SCAN 进行复现，并使用结构学习的思想优化损失函数来更好地保留图像的邻域结构，在三个小规模基准数据集上的实验结果表明，复现工作较为成功且在其上的改进能够进一步提高聚类结果的准确度。在未来的工作中，可以尝试将更多的图学习方法结合到图像聚类任务中，实现进一步的提升，此外，将原论文的做法以及结构学习的思想运用到半监督方法中也值得研究。

参考文献

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

[2] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 5879–5887, 2017.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[6] Zhiyuan Dang, Cheng Deng, Xu Yang, Kun Wei, and Heng Huang. Nearest neighbor matching for deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13693–13702, 2021.

[7] Kien Do, Truyen Tran, and Svetha Venkatesh. Clustering by maximizing mutual information across views. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9928–9938, 2021.

[8] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. *Advances in neural information processing systems*, 32, 2019.

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[10] Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, and Yi Xu. Hcsc: Hierarchical contrastive selective coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9706–9715, 2022.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[12] Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. Deep subspace clustering networks. *Advances in neural information processing systems*, 30, 2017.

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[14] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[15] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8547–8555, 2021.

[16] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[17] Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. *arXiv preprint arXiv:1801.01587*, 2018.

[18] Kai Tian, Shuigeng Zhou, and Jihong Guan. Deepcluster: A general clustering framework based on deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part II 17*, pages 809–825. Springer, 2017.

[19] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pages 268–285. Springer, 2020.

[20] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[21] Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8150–8159, 2019.

[22] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[23] Masataka Yamaguchi, Go Irie, Takahito Kawanishi, and Kunio Kashino. Subspace structure-aware spectral clustering for robust subspace clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9875–9884, 2019.

[24] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning*, pages 3861–3870. PMLR, 2017.

[25] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4066–4075, 2019.

[26] Huasong Zhong, Jianlong Wu, Chong Chen, Jianqiang Huang, Minghua Deng, Liqiang Nie, Zhouchen Lin, and Xian-Sheng Hua. Graph contrastive clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9224–9233, 2021.