Computer Science and Application 计算机科学与应用, 2026, 16(2), 65-75
Published Online February 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.162039

Hans汉斯

基于解耦对比注意力的图像聚类算法研究

周佳成，胡文军

湖州师范学院信息工程学院，浙江 湖州

收稿日期：2025年12月26日；录用日期：2026年1月23日；发布日期：2026年1月30日

摘　要

针对深度聚类中图像特征判别力不足、复杂背景干扰以及传统对比损失函数存在的正负耦合效应导致优化效率低下等问题，提出一种鲁棒且高效的图像聚类方法——解耦对比注意力聚类网络(DCACN)。首先，采用多裁剪数据增强策略，融合全局语义视图与局部细节视图来捕捉图像的多尺度不变性特征。然后，在ResNet-34骨干网络中无缝集成卷积块注意力模块，进而利用通道和空间注意力机制引导网络聚焦前景主体并抑制背景噪声。最后，引入解耦对比学习模块，消除正负样本间的梯度耦合来提升训练效率。在ImageNet-10、CIFAR-10和STL-10数据集上的实验结果表明，DCACN的聚类准确率分别达到了0.935、0.918和0.876，显著优于其他主流算法。

关键词

深度聚类，对比学习，注意力机制，图像聚类

Research on Image Clustering Algorithm Based on Decoupled Contrastive Attention

Jiacheng Zhou, Wenjun Hu

School of Information Engineering, Huzhou University, Huzhou Zhejiang

Received: December 26, 2025; accepted: January 23, 2026; published: January 30, 2026

Abstract

To address issues such as insufficient discriminative power of image features in deep clustering, interference from complex backgrounds, and the inefficiency of optimization caused by positive-negative coupling effects in traditional contrastive loss functions, a robust and efficient image clustering method—Decoupled Contrastive Attention Clustering Network (DCACN)—is proposed. First, a multi-crop data augmentation strategy is employed to integrate global semantic views and local detail views, capturing multi-scale invariant features of images. Then, the Convolutional Block

文章引用: 周佳成, 胡文军. 基于解耦对比注意力的图像聚类算法研究[J]. 计算机科学与应用, 2026, 16(2): 65-75.
DOI: 10.12677/csa.2026.162039

Attention Module is seamlessly integrated into the ResNet-34 backbone network, utilizing channel and spatial attention mechanisms to guide the network in focusing on foreground subjects while suppressing background noise. Finally, a decoupled contrastive learning module is introduced to eliminate gradient coupling between positive and negative samples, thereby improving training efficiency. Experimental results on ImageNet-10, CIFAR-10, and STL-10 datasets demonstrate that DCACN achieves clustering accuracies of 0.935, 0.918, and 0.876, significantly outperforming other mainstream algorithms.

Keywords

Deep Clustering, Contrastive Learning, Attention Mechanism, Image Clustering

1. 引言

近年来，深度聚类已成为计算机视觉和机器学习领域的一个研究热点，其目的旨在利用神经网络强大的非线性映射能力提取特征，从而在无监督场景下实现对高维图像数据的自动分类。在诸如自动驾驶、图像检索和医学影像分析等实际应用场景中，获取高质量的判别性语义特征对于提升聚类性能具有重要意义。

目前，深度聚类领域已取得了显著的研究成果，特别是基于对比学习的方法通过拉近正样本对、推远负样本对的空间距离，极大地提升了表征的质量。然而，传统方法在处理复杂图像聚类任务时仍面临挑战。首先，现有的数据增强策略往往仅依赖简单的随机裁剪，容易导致模型忽视图像的全局语义结构或在面对剧烈尺度变化时表现不佳。其次，标准的卷积神经网络往往平等地对待所有像素，缺乏区分前景主体与背景噪声的能力，导致模型容易学习到错误的聚类模式。此外，广泛使用的对比损失函数存在正负耦合效应，限制了模型的学习效率和优化稳定性。

为应对上述问题，本文提出了一种解耦对比注意力聚类网络(Decoupled Contrastive Attention Clustering Network, DCACN)。该模型旨在通过多层次的改进构建一个鲁棒的聚类框架：在输入端，融合了多裁剪思想设计混合增强策略，以捕获图像的多尺度不变性特征；在特征提取阶段，引入混合注意力机制引导网络聚焦关键区域并抑制背景干扰；在损失函数设计上，通过引入解耦对比学习消除正负样本间的梯度耦合。主要贡献如下：

(1) 提出一种基于多裁剪的混合数据增强策略，通过构建包含全局语义和局部细节的多样化视图集合，增强模型对多尺度特征的捕获能力；

(2) 设计并集成了融合注意力机制的特征编码器，利用通道和空间双重维度自适应聚焦图像前景，有效抑制背景噪声对聚类结构的干扰；

(3) 引入了解耦对比损失优化模块，分别在实例级和聚类级解除正负样本对之间的梯度耦合，在提升模型训练效率的同时，进一步增强聚类分配的准确性。

2. 相关工作

基于自编码器的深度聚类，其最初形态是一种简单直接的两阶段或分离式方法，这类方法往往将特征学习与聚类过程独立开。具体地，先利用全部无标签图像数据，以最小化重构误差为目标，独立地训

练一个深度自编码器网络，如栈式自编码器(Stacked Autoencoder, SAE) [1] [2]，当网络训练收敛后，将编码器作为一个固定的、非线性的特征提取器[3]；然后，将所有图像数据输入到这个预训练好的编码器中，提取出每个样本对应的低维潜在特征；最后，将这些特征应用于诸如 K-Means [4]、高斯混合模型等传统聚类算法并最终得到聚类结果。这种两阶段方法的优势在于其概念清晰、实现简单，并且相比于在原始像素空间上直接聚类，性能通常有显著提升，证明了深度表示对于聚类的有效性。然而，其根本性的缺陷在于特征学习与聚类任务的完全解耦[5]。自编码器的训练目标是尽可能精确地重构输入，这意味着它学习到的特征是为了表示数据，而非为了区分数据[6]。因此，这种方式下学习到的特征对于后续的聚类任务来说是次优的，无法保证最大化类间距离和最小化类内距离，从而限制了聚类性能的上限。

对比学习作为自监督学习的一种主流范式，在无监督特征表示领域取得了突破性进展。该方法的核心思想是通过拉近同一图像的不同增强视图[7]，并推远不同图像的视图，从而学习到具有判别力的特征。典型的对比聚类方法同时在实例级和聚类级进行对比学习，显著提升了聚类精度。然而，传统的对比学习方法面临两个主要瓶颈：一是正负样本对之间存在的“梯度耦合”效应[8]，即正样本的拉近受到负样本排斥强度的隐式抑制，降低了优化效率；二是单一尺度的视图生成策略难以兼顾图像的全局语义与局部细节。

注意力机制通过模拟人类视觉系统对关键信息的捕捉能力，使网络能够自适应地关注特征图中的重要区域。在计算机视觉领域，卷积块注意力模块是一种极具代表性的轻量化模块，它通过串联通道注意力模块和空间注意力模块，分别在关注什么和关注哪里两个维度上对特征进行重标定[9]。在深度聚类任务中引入注意力机制，可以引导模型从复杂的背景中剥离出具有判别性的前景主体特征，从而抑制噪声干扰并增强聚类分配的准确性。这一思想若与残差网络深度融合，便构建出更具鲁棒性的特征编码器[10]。

3. 解耦对比注意力聚类网络

3.1. 网络框架概述

本文提出的解耦对比注意力聚类网络是一个端到端的深度无监督学习框架，如图 1 所示，其旨在通过联合优化特征表示学习和聚类分配，解决传统聚类方法在高维图像数据上特征判别力不足及聚类结构不清晰的问题。图 1 网络框架采用了一种双分支、多阶段的协同优化策略，主要由三个核心模块组成，具体包括多裁剪的数据增强模块、融合注意力机制的特征提取器、双分支对比投影头的解耦对比损失优化模块。

网络的输入端采用了基于多裁剪的数据增强策略。对于每一个原始输入图像，该模块并非简单地生成单一增强视图，而是构建了一组包含不同尺度和视角的视图集合。如图中左侧所示，原始图像经过随机裁剪、翻转、颜色抖动等操作，生成了两个包含全局语义信息的高分辨率视图以及多个关注局部细节的低分辨率视图。所有生成的增强视图共享同一个特征提取器。该模块以深度残差网络为主干架构，并在其基础残差块中无缝集成了卷积块注意力模块。特征提取器包含四个阶段，分别堆叠了 3、4、6、3 个融合了 CBAM 的残差块。特征提取器输出的高维特征向量被送入两个并行的非线性映射模块，实例级对比头和聚类级对比头。实例级对比头由多层感知机构成[11]，包含全连接层和 ReLU 激活函数。其作用是将特征映射到一个低维的实例特征空间，用于计算样本个体之间的相似度，旨在拉近同一图像不同视图的特征距离，推远不同图像的特征距离，从而学习到细粒度的实例判别性特征。聚类级对比头同样由 MLP 构成，但在末端接有一个 Softmax 归一化层。该分支将特征映射到一个维度等于预设聚类数 K 的概率空间，输出每个样本属于各个簇的软分配概率。网络的优化目标由解耦对比损失优化模块定义，该模块摒弃了传统的耦合式对比损失，转而在实例级和聚类级分别计算解耦对比损失。解耦损失通过移除分母中的正样本项，消除了正负样本之间的梯度耦合效应，提升了学习效率[4]。

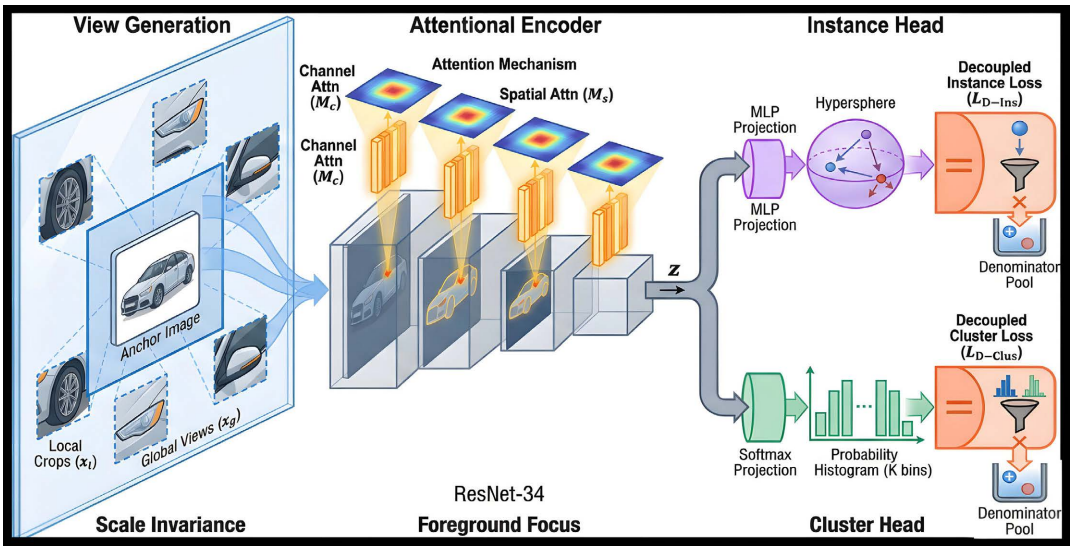

Figure 1. Decoupling contrastive attention-based clustering network framework
图 1. 解耦对比注意力聚类网络框架

3.2. 多裁剪数据增强策略

在深度聚类任务中，数据增强作为自监督学习的核心驱动力，其质量直接决定了模型所能学习到的特征不变性和泛化能力。传统的对比学习方法通常采用简单的“双视图”策略，即对每张图像仅生成两个随机增强视图。然而，这种策略在处理复杂图像数据时往往存在局限性。一方面，单一尺度的视图难以同时捕获图像的全局语义和局部细节；另一方面，有限的样本对数量限制了模型对特征空间探索的充分性。为了克服这些不足，基于 SwAV [12]中的多裁剪思想，设计一种包含全局视图与局部视图的混合增强策略。该策略的核心在于构建一个多尺度、多视角的正样本集合。具体而言，给定输入图像，生成两类具有不同分辨率和覆盖范围的视图：全局视图和局部视图。全局视图包含两个高分辨率的视图，旨在它们保留了图像的完整语义结构和上下文信息，其裁剪区域覆盖了原始图像的大部分，这些视图经过标准的随机翻转、颜色抖动等弱增强处理，确保模型能够学习到图像的宏观特征和主体结构[13]。局部视图包含多个低分辨率的局部视图，旨在迫使模型关注图像的细粒度特征和局部纹理，其裁剪区域仅覆盖原始图像的一小部分，这些视图通常应用于增强操作，如高斯模糊、灰度化等，以增加任务的难度。

3.3. 融合注意力机制的特征编码器

特征编码器是深度聚类网络的核心组件，其提取特征的质量直接决定了后续对比学习和聚类分配的性能上限。传统的卷积神经网络在提取特征时，通常将特征图中的每个通道和空间位置视为同等重要。然而，在无监督聚类任务中，图像的背景噪声、无关物体或纹理细节往往会对聚类结果产生干扰[8] [14]。为使网络能够自适应地聚焦于具有判别力的关键区域和特征通道[15]，这里在特征编码器中融合卷积块注意力模块(Convolutional Block Attention Module, CBAM)。选用的特征编码器主干网络为 ResNet-34 [16]，其基本构建单元是残差块。为了在不显著增加计算开销的前提下增强特征表达能力，将 CBAM 无缝嵌入到每个残差块的卷积层之后，具体如图 2 所示，其中 CBAM 包含两个序贯连接的子模块：通道注意力模块 CAM 和空间注意力模块 SAM。

通道注意力模块旨在关注“什么”是重要的特征。在输入特征图 $F \in \mathbb{R}^{C\times H\times W}$ 上，CAM 首先分别进行全局最大池化和全局平均池化，以聚合特征图的空间信息，得到两个不同的通道描述符 $F^{C}_{\max} \in \mathbb{R}^{C\times 1\times 1}$

$F_{\text{avg}}^{C} \in \mathbb{R}^{C\times1\times1}$。这两种池化操作分别捕获了特征图中最显著的特征响应和整体的背景信息，具有互补性。随后，这两个描述符被送入一个共享的多层感知机。该 MLP 包含两层全连接层，中间层维度缩减为 C/r，并使用 ReLU 激活函数[16]。MLP 输出的两个特征向量逐元素相加后，经过 Sigmoid 激活函数，生成最终的通道注意力图 $M_c \in \mathbb{R}^{C\times1\times1}$：

$$M_c(F) = o\left(\text{MLP}\left(\text{AvgPool}(F)\right) + \text{MLP}\left(\text{MaxPool}(F)\right)\right) \tag{1}$$

最后，将通道注意力图 M_c 与原始特征图 F 进行广播乘法，得到经过通道注意力细化的特征图 F'

$$F' = M_c(F) \otimes F \tag{2}$$

通过这一过程，网络能够自动识别并增强对聚类任务有贡献的关键特征通道，同时抑制无关或冗余的通道。

空间注意力模块旨在关注哪里是重要的区域。它以经过通道细化的特征图 F 作为输入。首先，沿通道轴分别进行最大池化和平均池化，生成两个二维的空间特征描述图 $F_{\max}^{s} \in \mathbb{R}^{C\times1\times1}$ 和 $F_{\max}^{c} \in \mathbb{R}^{C\times1\times1}$，这一步骤有效地压缩了通道信息，突出了空间维度的显著区域。接着，将这两个描述图在通道维度上进行拼接，并通过一个 7×7 的卷积层进行融合。最后，经过 Sigmoid 激活函数，生成空间注意力图 $M_s \in \mathbb{R}^{C\times1\times1}$

$$M_s(F') = \sigma\left(\text{MLP}\left(\text{AvgPool}(F)\right) + \text{MLP}\left(\text{MaxPool}(F)\right)\right) \tag{3}$$

最终，将空间注意力图 M_s 与特征图 F' 进行逐元素相乘，得到最终的精炼特征图 F'

$$F'' = M_s(F') \otimes F' \tag{4}$$

空间注意力机制使得网络能够定位图像中的显著目标区域，并忽略背景区域的干扰。这对于无监督聚类尤为重要，因为它确保了聚类是基于物体本身的语义特征，而非背景环境。

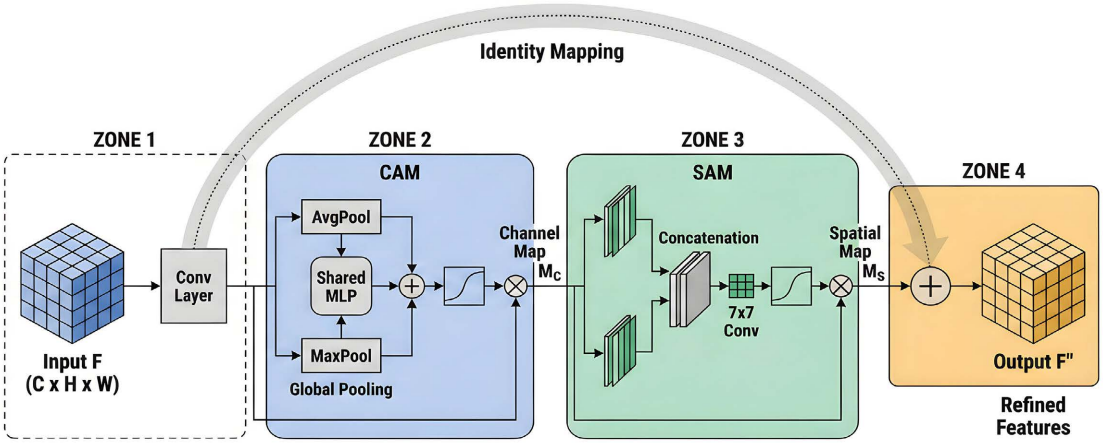

Figure 2. Residual block structure with combined CBAM
图 2. 融合 CBAM 的残差块结构

3.4. 解耦对比学习模块

在获得图像的高质量特征表示后，如何设计有效的损失函数来驱动网络学习是关键。传统的 InfoNCE [17]损失函数虽然在对比学习中取得了巨大成功，但其正样本对的相似度计算同时出现在分子和分母中，导致了正负耦合现象。研究表明，这种耦合效应会降低模型对正样本对的优化效率，因为正样本的拉近会隐式地受到负样本排斥强度的抑制。为了解决这一问题，在实例级和聚类级两个层面均引入了解耦对

比损失，旨在解除正负样本间的梯度耦合，提升模型的学习效率和最终聚类性能[18]。

3.4.1. 实例级对比损失

实例级对比学习的目标是学习具有个体判别性的特征表示[19]，即在特征空间中拉近同一图像不同视图的距离，推远不同图像视图之间的距离，常用信息噪声对比估计(Information Noise Contrastive Estimation, InfoNCE)作为损失。首先编码器输出的特征经过一个非线性的实例投影头 $g_{inst}(\cdot)$ 映射到实例特征空间，得到 $h=g_{inst}(z)$。对于批次中的任意一个图像 x_i，假设其生成的两个视图特征分别为 h_i^1， h_i^2。传统的 InfoNCE 损失形式为：

$$L_{NCE}=-\log\frac{\exp\left(\sin\left(h_i^1,h_i^2\right)/\tau\right)}{\sum_{k=1}^{2N}\mathbb{I}_{k\neq i}\exp\left(\sin\left(h_i^1,h_k\right)/\tau\right)+\exp\left(\sin\left(h_i^1,h_i^2\right)/\tau\right)} \tag{5}$$

其中 τ 是温度参数，$\mathbb{I}(\cdot)$ 代表指示函数，用于逻辑判断，$\sin(u,v)=\frac{u^{\mathrm{T}}v}{\|u\|\|v\|}$ 表示余弦相似度。可以看到，分子中的正样本相似度项也出现在分母中。这里采用式(6)的解耦实例对比损失，其显式地移除了分母中的正样本项：

$$L_{inst}=-\log\frac{\exp\left(\sin\left(h_i^1,h_i^2\right)/\tau\right)}{\sum_{k=1,k\neq i}^{2N}\exp\left(\sin\left(h_i^1,h_k\right)/\tau\right)} \tag{6}$$

3.4.2. 聚类级对比损失

聚类级对比学习的目标是挖掘数据的语义簇结构，即同一图像的不同视图应该被分配到同一个聚类簇中，而不同语义类别的图像应具有不同的聚类分配分布[17]。特征 z 经过另一个并行的聚类投影头 $g_{clus}(\cdot)$，映射到维度为 K 的向量空间，并经过 Softmax 归一化得到软聚类分配概率 $p=\mathrm{Softmax}\left(g_{clus}(z)\right)\in\mathbb{R}^K$。这里本文将 p_i 视为样本 i 在聚类空间中的特征表示。

类似于实例级，本文定义解耦聚类对比损失。对于同一图像的两个视图，其聚类分配概率 p_i^1， p_i^2 应尽可能相似。相似度度量依然采用余弦相似度。损失函数定义如下：

$$L_{clus}=-\log\frac{\exp\left(\sin\left(p_i^1,p_i^2\right)/\tau_c\right)}{\sum_{k=1,k\neq i}^{2N}\exp\left(\sin\left(p_i^1,p_k\right)/\tau_c\right)} \tag{7}$$

其中，τ_c 是聚类级的温度参数。最终，DCACN 的总优化目标为实例级损失和聚类级损失的加权和：

$$L_{total}=L_{inst}+\lambda L_{clus} \tag{8}$$

其中，λ 是平衡两个损失项权重的超参数。通过这种双层次的解耦对比学习，网络不仅能够学习到细粒度的实例特征，还能自适应地形成语义一致的聚类结构。在训练过程中，这种联合优化策略使得特征表示学习和聚类分配相互促进。

4. 实验结果与分析

4.1. 数据集及评价指标

为了全面、客观地评估所提出的解耦对比注意力聚类网络的性能，这里使用 3 个广泛使用的标准图像聚类基准数据集进行实验验证。这些数据集涵盖了从简单的灰度数字图像到复杂的彩色自然物体图像，具

有不同的类别数量、图像分辨率和背景复杂度，能够充分检验模型在不同场景下的泛化能力和鲁棒性。本实验选用的四个数据集分别是：ImageNet-10、CIFAR-10 和 STL-10。表 1 给出了这些数据集的具体信息。

Table 1. Statistical information of experimental datasets
表 1. 实验数据集统计信息

数据集	类别数	样本总数	图像尺寸	通道数	内容描述
ImageNet-10 [20]	10	13,000	-	3	ImageNet 自然物体子集
CIFAR-10 [21]	10	60,000	32 × 32	3	自然物体
STL-10 [22]	10	13,000	96 × 96	3	高分辨率自然物体

为了定量地衡量深度聚类算法的性能，本章采用了三个被广泛认可和使用的标准评价指标：聚类准确率(Clustering Accuracy, ACC)、标准互信息(Normalized Mutual Information, NMI)和调整兰德系数(Adjusted Rand Index, ARI)。

(1) ACC 是最直观的性能指标，它衡量的是聚类算法正确预测样本类别的比例。由于无监督聚类算法输出的簇标签与数据集的真实类别标签之间通常不存在直接的一一对应关系，因此在计算 ACC 之前，必须先利用匈牙利算法找到预测标签与真实标签之间的最佳匹配映射。给定数据集包含 N 个样本，其真实标签集合为 $y=\{y_1,y_2,\cdots,y_N\}$，聚类算法输出的预测标签集合为 $c=\{c_1,c_2,\cdots,c_N\}$。ACC 定义如下：

$$\mathrm{ACC}=\max_{m\in\mathcal{M}}\frac{\sum_{i=1}^{N}\mathbb{I}\left(y_i=m\left(c_i\right)\right)}{N} \tag{9}$$

其中，y_i 是样本 i 的真实标签，c_i 是样本 i 的聚类预测标签，$\mathcal{M}$ 是所有可能的从预测标签空间到真实标签空间的一一映射函数的集合，$m(\cdot)$ 是其中一个映射函数。$\mathbb{I}(\cdot)$ 是指示函数，当条件为真时取值为 1，否则为 0。

(2) NMI 是一个基于信息论的度量指标，用于衡量两个随机变量(即聚类结果和真实标签)之间的共享信息量。NMI 的优势在于它不受簇的排列顺序影响，且对簇的大小不敏感。NMI 的计算公式为：

$$\mathrm{NMI}\left(y,c\right)=\frac{2\cdot I\left(y;c\right)}{H\left(y\right)+H\left(c\right)} \tag{10}$$

其中，$I\left(y;c\right)$ 表示真实标签 y 与预测标签 c 之间的互信息，量化了已知聚类结果 c 后对真实标签 y 的不确定性的减少程度。$H\left(y\right)$ 和 $H\left(c\right)$ 分别是真实标签 y 和标签 c 的熵。NMI 通过将互信息用各自的熵进行归一化，使得其值域固定在[0, 1]之间，NMI = 1 表示聚类结果与真实标签完全一致，NMI = 0 表示两者完全独立。

(3) ARI 是兰德系数的改进版本。ARI 衡量的是成对样本在聚类结果和真实标签中是否被一致地划分。然而，ARI 在随机划分情况下的期望值不为 0，这使得它在不同簇数的情况下难以进行公平比较。ARI 通过引入期望值校正，解决了这一问题，使得随机聚类的 ARI 期望值为 0，而完美聚类的 ARI 为 1。ARI 的计算基于列联表，定义如下

$$\mathrm{ARI}=\frac{\sum_{ij}\binom{n_{ij}}{2}-\left[\sum_{i}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}\right]\Big/\binom{N}{2}}{\frac{1}{2}\left[\sum_{i}\binom{a_i}{2}+\sum_{j}\binom{b_j}{2}\right]-\left[\sum_{i}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}\right]\Big/\binom{N}{2}} \tag{11}$$

其中，n_{ij} 是同时属于真实类别 i 和预测簇 j 的样本数量，a_i 是属于真实类别 i 的样本总数，b_j 是属于预测簇 j 的样本总数，N 是样本总数。ARI 综合考虑了正确聚类对和错误聚类对，是一个非常鲁棒的聚类评

价指标，尤其在簇的大小不均衡或簇的数量较多时，ARI 往往比 ACC 更能反映聚类结构的真实质量。

4.2. 实验结果对比与性能分析

(1) DCACN 稳定性实验

为了进一步探究 DCACN 模型的训练稳定性和收敛速度，本文记录了模型在训练过程中三个评价指标随迭代轮数的变化情况。图 3 展示了三个数据集在 1000 个 Epoch 训练周期内的性能曲线，其中红色曲线表示 ACC，蓝色曲线表示 NMI，绿色曲线表示 ARI。从图 3 可以看出，DCACN 的训练表现兼具效率与稳定性：训练初期(前 200 个 Epoch)，三个数据集的 ACC、NMI、ARI 均快速攀升，CIFAR-10 的 ACC 在 150 个 Epoch 便突破 0.8，体现了解耦对比损失对优化效率的提升；训练中后期(400 个 Epoch 后)，指标曲线趋于平稳，波动幅度收窄，验证了多裁剪增强与注意力编码的鲁棒性；且模型在样本规模、分辨率不同的数据集上，最终指标均收敛至 0.8 以上，泛化性良好。这些结果印证了 DCACN 核心模块的设计合理性。

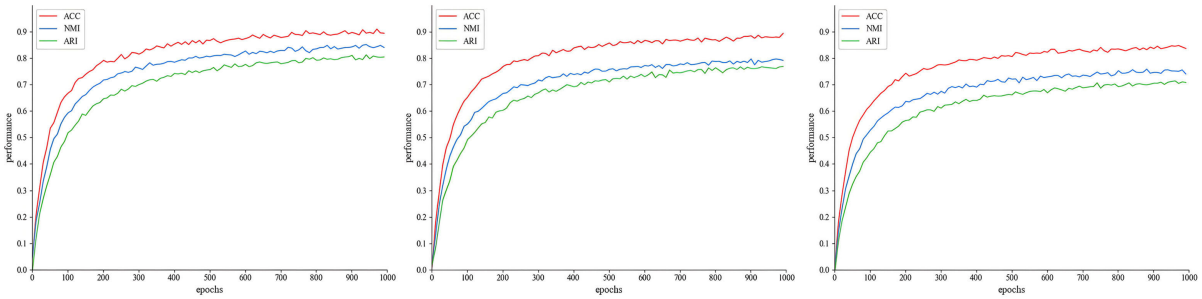

Figure 3. Cluster results visualization of three image datasets
图 3. 三种图像数据集的聚类结果展示图

(2) DCACN 特征表示能力实验

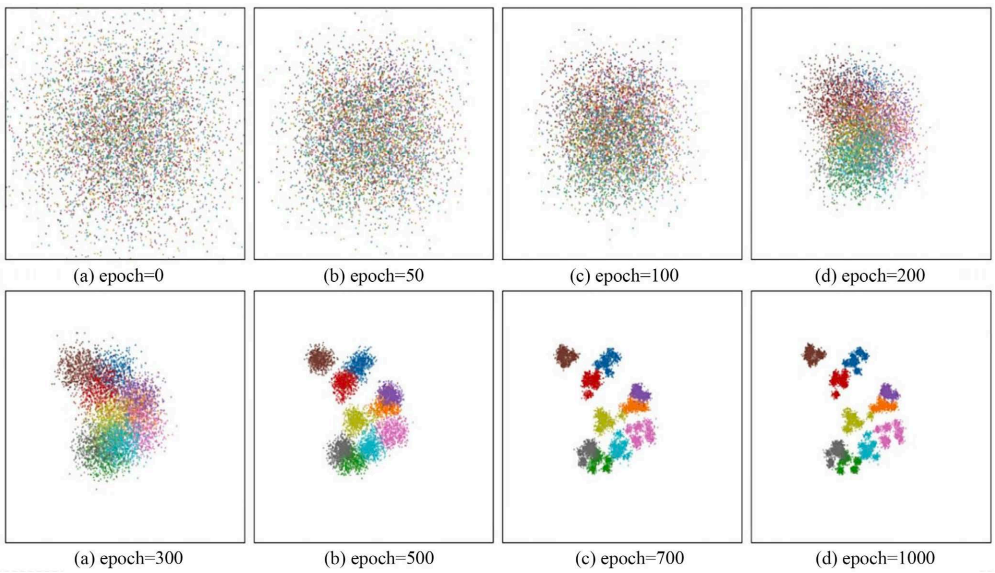

Figure 4. Visualization of the evolution of t-SNE features on the ImageNet-10 dataset
图 4. 在 ImageNet-10 数据集上的 t-SNE 特征可视化演变

为了更直观地展示 DCACN 学习到的特征表示的判别能力，本文利用 t-SNE 算法将 ImageNet-10 数

据集的高维特征映射到二维平面进行可视化。图 4 展示了随着训练过程的推进，特征空间分布的演变过程。在训练初始阶段，所有类别的样本点混杂在一起，特征分布呈现出无序的混沌状态，表明此时网络尚未提取到有意义的语义特征。随着训练的进行，可以看到不同颜色的样本点开始逐渐聚拢，原本混杂的分布开始出现分裂，形成了初步的簇结构轮廓。在训练后期，簇与簇之间的边界变得日益清晰，类内样本高度紧凑，类间距离显著增大。在 Epoch1000 时，10 个类别的样本点已经形成了 10 个界限分明、独立的簇团。

(2) 其他算法聚类对比实验

为了验证 DCACN 模型的有效性，利用 ImageNet-10、CIFAR-10 和 STL-10 数据集开展无监督聚类实验，并与近年来在该领域表现优异的多种深度聚类算法进行对比，表 2 是具体的实验结果。

Table 2. Performance comparison of different clustering algorithms on ImageNet-10, CIFAR-10, and STL-10 datasets
表 2. 不同聚类算法在 ImageNet-10，CIFAR-10 和 STL-10 数据集上的性能对比

方法	ImageNet-10			CIFAR-10			STL-10		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
K-Means	0.241	0.119	0.057	0.229	0.087	0.049	0.192	0.125	0.061
DEC	0.381	0.282	0.203	0.301	0.257	0.161	0.359	0.276	0.186
DCEC [20]	0.401	0.302	0.287	0.352	0.342	0.239	0.389	0.347	0.255
IIC [18]	-	-	-	0.617	0.551	0.411	0.596	0.496	0.397
PICA [23]	0.870	0.802	0.761	0.696	0.591	0.512	0.713	0.611	0.531
CC [24]	0.893	0.859	0.822	0.790	0.705	0.637	0.850	0.764	0.726
SCCMD [7]	-	-	-	0.903	0.814	0.805	0.793	0.756	0.733
DCACN (Ours)	0.935	0.887	0.864	0.918	0.832	0.821	0.876	0.789	0.758

可以看出 DCACN 在各项指标上均取得了显著领先。相比于经典的对比聚类方法 CC，DCACN 在 ACC 上提升了约 12.8%，在 ARI 上提升了 18.4%。这表明引入的多裁剪策略和注意力机制有效弥补了低分辨率图像中局部细节丢失的问题，使得模型能更准确地捕捉物体的细微语义差异。即使与最新的 SCCMD 相比，DCACN 在 ACC 上也保持了 1.5%的优势，验证了解耦对比损失在优化效率上的贡献。STL-10 具有较高的图像分辨率但有标签训练样本较少。DCACN 在该数据集上同样表现出色，ACC 达到了 0.876，超过了 CC 和 SCCMD。这一结果尤为重要，说明 DCACN 在有限数据下具有更强的特征泛化能力。CBAM 注意力模块在此处发挥了关键作用，它帮助模型在缺乏大量样本覆盖的情况下，依然能够精准定位图像主体，减少背景噪声对聚类原型的干扰。ImageNet-10 因其复杂的背景和多样的物体姿态而最具挑战性。DCACN 在此数据集上取得了 0.935 的极高准确率，大幅超越了 PICA 和 CC。这证明了 DCACN 具有抗干扰能力和语义理解能力。多裁剪策略生成的全局–局部视图对，使得模型能够学习到具有高度尺度不变性和视角不变性的特征表示，从而在面对真实世界复杂场景时依然能保持稳健的聚类性能。

5. 结束语

本文针对深度聚类中特征判别力不足及优化效率低的问题，提出了一种端到端的解耦对比注意力聚类网络。通过设计了基于多裁剪的混合数据增强策略，构建包含全局与局部视图的多样化样本对，增强了模型对多尺度特征的学习能力。在特征编码器中融入 CBAM 注意力机制，实现了通道与空间维度的自

适应特征重标定，有效抑制了背景噪声并突出了关键语义特征。最后引入了解耦对比学习模块，在实例级和聚类级分别设计了解耦损失函数，消除了正负样本间的梯度耦合，显著提升了优化效率和聚类精度。在 ImageNet-10、CIFAR-10 和 STL-10 三个数据集上的实验结果表明，DCACN 在各项聚类指标上均优于现有的主流方法，验证了该框架的有效性与鲁棒性。

未来研究可从以下三个方面推进：其一，探索模型轻量化与加速策略。研究通过知识蒸馏、模型剪枝或低秩近似等技术，在保留 CBAM 特征重标定能力的同时降低编码器参数量，并优化多裁剪视图的融合效率。其二，引入更细粒度的判别性机制。尝试结合度量学习或难样本挖掘技术，进一步优化解耦对比损失函数，以提升模型在处理背景复杂、类间差异微小的图像数据时的聚类稳定性。其三，研究动态聚类与开放场景扩展。探索将非参数聚类方法或类别发现机制引入 DCACN 框架，使模型能够自适应地推断潜在的聚类中心数目，从而提升在真实、动态变化的数据环境中的通用性与泛化能力。

参考文献

[1] Vincent, P., Larochelle, H., Bengio, Y., *et al*. (2023) Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, **11**, 3371-3408.

[2] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[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, 18-22 June 2018, 7132-7141. https://doi.org/10.1109/cvpr.2018.00745

[4] Xie, J., Girshick, R. and Farhadi, A. (2024) Unsupervised Deep Embedding for Clustering Analysis. *International Conference on Machine Learning*. *PMLR*, Vienna, 21-27 July 2024, 478-487.

[5] Hinton, G.E. and Salakhutdinov, R.R. (2006) Reducing the Dimensionality of Data with Neural Networks. *Science*, **313**, 504-507. https://doi.org/10.1126/science.1127647

[6] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) Rethinking the Inception Architecture for Computer Vision. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 2818-2826. https://doi.org/10.1109/cvpr.2016.308

[7] Li, Y., Hu, P., Liu, Z., Peng, D., Zhou, J.T. and Peng, X. (2021) Contrastive Clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 8547-8555. https://doi.org/10.1609/aaai.v35i10.17037

[8] Woo, S., Park, J., Lee, J. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. In: Ferrari, V., *et al*., Eds., *Proceedings of the European Conference on Computer Vision*, Springer International Publishing, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[9] 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

[10] MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281-297.

[11] Reynolds, D.A. (2009) Gaussian Mixture Models. In: Li, S.Z. and Jain, A., Eds., *Encyclopedia of Biometrics*, Springer US, 659-663. https://doi.org/10.1007/978-0-387-73003-5_196

[12] Caron, M., Misra, I., Mairal, J., *et al*. (2020) Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Advances in Neural Information Processing Systems* 33: *Annual Conference on Neural Information Processing Systems* 2020, *NeurIPS* 2020, 6-12 December 2020, 9912-9924.

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2021) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Proceedings of the* 9*th International Conference on Learning Representations*, 3-7 May 2021, 1-15.

[14] Guo, X., Gao, L., Liu, X. and Yin, J. (2017) Improved Deep Embedded Clustering with Local Structure Preservation. *Proceedings of the* 26*th International Joint Conference on Artificial Intelligence*, Melbourne, 19-25 August 2017, 1753-1759. https://doi.org/10.24963/ijcai.2017/243

[15] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W. and Hu, Q. (2020) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 11534-11542. https://doi.org/10.1109/cvpr42600.2020.01155

[16] Dizaji, K.G., Herandi, A., Deng, C., Cai, W. and Huang, H. (2017) Deep Clustering via Joint Convolutional Autoencoder Embedding and Relative Entropy Minimization. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 5736-5745. https://doi.org/10.1109/iccv.2017.612

[17] Kingma, D.P. and Welling, M. (2014) Auto-Encoding Variational Bayes. *Proceedings of the International Conference on Learning Representations*, Banff, 14-16 April 2014, 1-14.

[18] Guo, X., Liu, X., Zhu, E. and Yin, J. (2017) Deep Clustering with Convolutional Autoencoders. In: Liu, D.R., *et al*., Eds., *Neural Information Processing*, Springer International Publishing, 373-382. https://doi.org/10.1007/978-3-319-70096-0_39

[19] Blum, A. and Mitchell, T. (1998) Combining Labeled and Unlabeled Data with Co-Training. *Proceedings of the* 11*th Annual Conference on Computational Learning Theory*, Madisson, 24-26 July 1998, 92-100. https://doi.org/10.1145/279943.279962

[20] Jiang, Z., Zheng, Y., Tan, H., Tang, B. and Zhou, H. (2017) Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering. *Proceedings of the* 26*th International Joint Conference on Artificial Intelligence*, Melbourne, 19-25 August 2017, 1965-1972. https://doi.org/10.24963/ijcai.2017/273

[21] Almahairi, A., Ballas, N., Cooijmans, T., *et al*. (2016) Dynamic Capacity Networks. *Proceedings of the* 33*rd International Conference on Machine Learning*, New York, 20-22 June 2016, 1228-1237.

[22] Zagoruyko, S. and Komodakis, N. (2016) Wide Residual Networks. In: *Proceedings of the British Machine Vision Conference* 2016, BMVA Press, 87.1-87.12. https://doi.org/10.5244/c.30.87

[23] Ji, X., Vedaldi, A. and Henriques, J. (2019) Invariant Information Clustering for Unsupervised Image Classification and Segmentation. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27-28 October 2019, 9865-9874. https://doi.org/10.1109/iccv.2019.00996

[24] Huang, J., Gong, S. and Zhu, X. (2020) Deep Semantic Clustering by Partition Confidence Maximisation. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 8849-8858. https://doi.org/10.1109/cvpr42600.2020.00887