Online Deep Clustering for Unsupervised Representation Learning



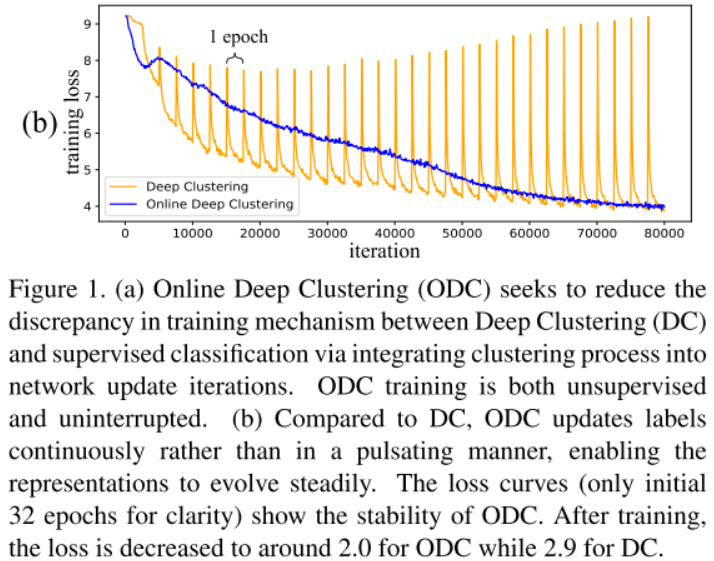
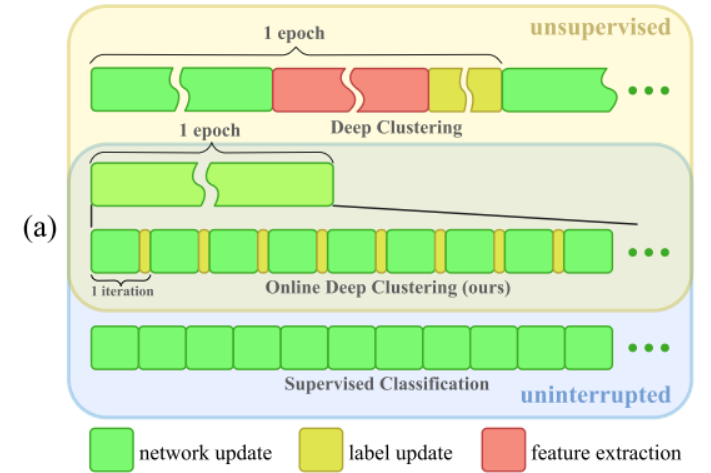
1. 摘要

联合聚类和特征学习方法在无监督的表征学习中表现出卓越的性能。然而，在特征聚类和网络参数更新之间交替进行的训练计划导致了视觉表征的不稳定学习。为了克服这一挑战，我们提出了在线深度聚类（ODC），它同时执行聚类和网络更新，而不是交替进行。我们的主要观点是，聚类中心点应该稳定地发展，以保持分类器的稳定更新。具体来说，我们设计并维护两个动态内存模块，即存储样本标签和特征的样本内存，以及用于中心点进化的中心点内存。我们将突然的全局聚类分解为稳定的内存更新和批量化的标签重新分配。这个过程被整合到网络更新迭代中。这样一来，标签和网络就会肩并肩地进化，而不是交替进行。大量的实验表明，ODC稳定了训练过程并有效地提高了性能。

1. 问题重述

与仅仅捕捉到图像内不变性的对比性表示学习[15, 16, 17]相比，基于聚类的方法能够探索图像间的相似性。与通常在固定特征上进行的传统聚类不同[18, 19]，这些工作联合优化聚类和特征学习。

Caron等人提出的深度聚类[13]（DC）是第一个尝试扩大基于聚类的表示学习。DC在深度特征聚类和CNN参数更新之间交替进行。特别是，在每个epoch的开始，它对整个数据集进行离线聚类算法，以获得伪标签作为下一个epoch的监督。离线聚类不可避免地在不同的epoch中对所分配的标签进行置换，也就是说，即使一些聚类没有变化，它们在聚类后的指数也会被随机置换。因此，分类器中的参数不能从上一个历时中继承下来，它们必须在每个历时前随机初始化。这种机制引入了训练的不稳定性，并使表征暴露在表征损坏的高风险中。如图1（a）所示，DC的网络更新被每个历时的特征提取和聚类所打断。这与传统的监督分类不同，监督分类是以不间断的方式使用固定标签进行的，其中一个迭代包括网络的前向和后向传播。



在这项工作中，我们试图设计出一种具有高稳定性的联合聚类和特征学习范式。为了减少聚类和监督学习之间训练机制的差异，我们将聚类过程分解为小批量的标签更新，并将这个更新过程整合到网络更新的迭代中。基于这种直觉，我们提出了在线深度聚类（ODC），用于联合聚类和特征学习。具体来说，一个ODC迭代包括前向和后向传播、标签重新分配和中心点更新。对于标签的更新，ODC重用了前向传播中的特征，从而避免了额外的特征提取。为了促进在线标签重新分配和中心点更新，我们设计并维护了两个动态内存模块，即用于存储样本标签和特征的样本内存，以及用于中心点进化的中心点内存。通过这种方式，ODC以一种不间断的方式进行训练，类似于监督分类，同时不需要人工标注。在训练过程中，标签和网络参数是肩并肩地进化的，而不是交替进行的。由于标签在每次迭代中都是连续和即时更新的，CNN中的分类器也会更稳定地进化，从而产生更稳定的损失曲线，如图1（b）所示。

虽然ODC在各种基准上取得了令人信服的无监督表示学习性能，但它可以自然地用于微调使用其他无监督学习方法训练的模型。广泛的实验表明，ODC的稳定性有助于它作为无监督的微调工具比DC表现得更好。我们总结了我们的贡献如下：

1）我们提出了ODC，它以无监督的方式学习图像表征，具有很高的稳定性。2) ODC也是一个统一的无监督的微调方案，进一步改进了以前的自监督表征学习方法。3）在不同的基准上观察到令人鼓舞的表现，表明联合聚类和特征学习的巨大潜力。

1. 相关工作

无监督的表征学习。许多无监督的视觉表征学习算法是基于生成模型的，它通常使用一个潜在的表征瓶颈来重建输入图像。现有的基于生成的模型包括自动编码器[20, 21]、受限波尔兹曼机[22, 23, 24]、可变自动编码器[25]和生成对抗网络[26]，其中一些在生成图像或视频方面表现出强大的能力[27, 28, 29, 30, 31, 32]。通过学习生成实例，这些模型可以学习有意义的潜在表征，可以用于下游的任务[5, 33, 34]。另一种流行的无监督的表征形式

学习的另一种流行形式是自我监督学习，其中一个辅助任务被设计为从原始数据中得出代理标签。通过鼓励CNN从数据中预测代理标签来学习表征。已经探索了各种辅助任务，例如，预测图像中的相对补丁位置[1]，解决拼图[4]，灰度图像着色[3，35]，缺失像素的绘画[2]，跨通道预测[36]，计数视觉基元[37]，预测图像旋转[8]，以及多视角对比学习[38]。对于视频，自发的监督信号来自于时间连续性[39, 40, 41, 42, 43, 44, 45]或运动一致性[46, 47, 48, 49, 9]。

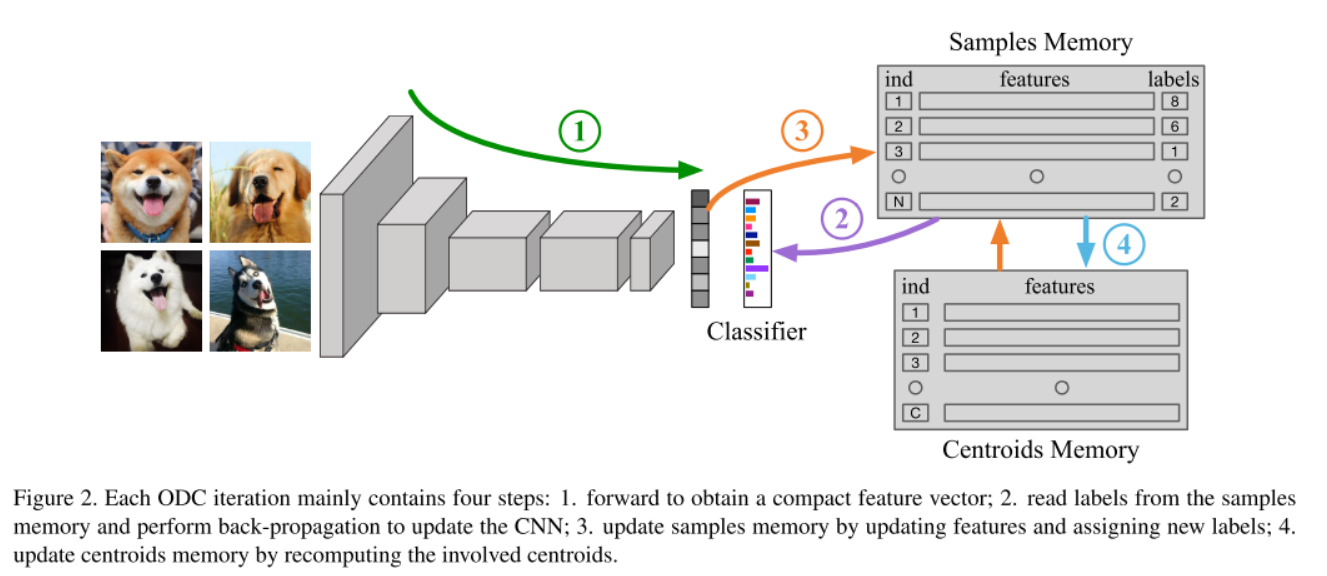
联合聚类和特征学习。基于聚类的无监督表示学习是最近特别感兴趣的。人们提出了各种方法来联合优化特征学习和图像聚类。值得注意的是，这些方法在小数据集的无监督特征学习中显示出巨大的潜力[11, 12, 50, 51]。为了扩展到像ImageNet这样的大数据集[52]，Caron等人[13]提出了DeepCluster来聚类特征，并在每个历时中用随后分配的伪标签来更新CNN。在随后的研究中，Caron等人[14]提出DeeperCluster来利用自我监督和聚类，并在非策划数据上验证其方法的表示学习能力。尽管深度聚类方法能够从大规模的无标签数据中学习良好的表征，但特征聚类和CNN参数更新的交替进行导致了训练的不稳定性。

自监督学习的改进。一些工作旨在从不同的角度改进以前的自监督学习方法。例如，Larsson等人[6]首次深入分析了作为辅助任务的着色，并对提高其有效性提出了一些见解。Mundhenk等人[53]探索了一套方法来避免一些琐碎的捷径，如基于上下文的自监督学习的色差。Noroozi等人[54]使用一种基于聚类的知识转移方法，在预训练时允许更深的网络，提高了自监督模型的性能。Wang等人[55]和Doersch等人[56]利用不同辅助任务中包含的多个线索来改进自监督模型。最近，一些工作[57, 58]广泛研究了现有自监督方法的架构和扩展能力。作为对这些工作的补充，ODC作为一个灵活统一的无监督微调方案来提升一般的自监督学习方法，尽管它可以单独用于从头开始进行无监督的表示学习

1. 详细阐述
   1. Online Deep Clustering

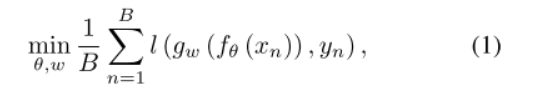
我们首先讨论DC的基本思想[13]，然后详细介绍所提出的ODC。为了学习表征，DC在离线特征聚类和带有伪标签的网络反向传播之间交替进行。离线聚类过程需要对整个训练集进行深度特征提取，然后采用全局聚类算法，例如K-Means聚类。全局聚类对伪标签进行了极大的改变，要求网络在随后的历时中迅速适应新标签。

**框架概述**。与DC不同，ODC不需要一个额外的特征提取过程。此外，标签与网络参数的更新同步进行。这是由新引入的样本和中心点存储器实现的。如图2所示，样本存储器存储整个数据集的特征和伪标签；而中心点存储器存储类中心点的特征，即一个类中所有样本的平均特征。这里的 "类 "代表了一个在训练中不断演变的临时集群。标签和网络参数在ODC的不间断迭代中同时更新。为了避免ODC陷入琐碎的解决方案，引入了包括损失再加权和处理小聚类的具体技术。

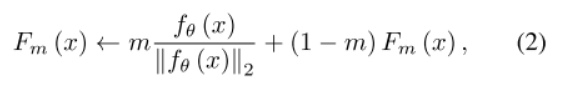


**An ODC Iteration**。假设我们有一个随机初始化的网络fθ（∗）和一个线性分类器gw（∗），目标是训练骨干参数θ，以产生高辨识度的表示。为了准备ODC，样本和中心点记忆通过一个全局聚类过程初始化，例如K-Means。接下来，我们可以不间断地迭代执行ODC。

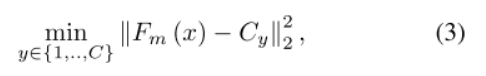
一个ODC迭代包含四个步骤。首先，给定一批输入图像{x}，网络将图像映射成紧凑的特征向量F=fθ（x）。其次，我们从样本存储器中读取这批图像的伪标签。利用这些伪标签，我们用随机梯度下降法更新网络，以解决以下问题：



其中yn是来自样本存储器的当前伪标签，B表示每个小批的大小。第三，L2归一化后的fθ（x）被重新用于更新样本存储器：



其中Fm (x)是样本存储器中x的特征，m∈(0, 1)是动量系数。同时，通过寻找最近的中心点，为每个涉及的样本分配一个新的标签。



其中Cy表示类y的中心点特征。 最后，所涉及的中心点，包括新成员加入的中心点和老成员离开的中心点，都被记录下来。每隔k次迭代，通过对属于其相应中心点的所有样本的特征进行平均化来更新这些中心点。

* 1. Handling Clustering Distribution in ODC

**损失再加权**。为了避免训练中崩溃成几个巨大的集群，DC在每个历时前都采用了统一的采样。然而，对于ODC来说，每个迭代中集群的样本数量都会发生变化。使用统一抽样需要在每次迭代中对整个数据集重新抽样，这个过程被认为是多余的，而且成本很高。我们提出了另一种方法，即根据每个类别中的样本数量对损失进行重新加权。为了验证它们的等价性，我们实施了一个带有损失重新加权的DC模型，并根据经验发现，当权重符合下列条件时，性能保持不变：



Nc代表着类别c的样本数目。我们在ODC中采用了相同的损失再加权公式。通过损失再加权，小群组中的样本对反向传播的贡献更大，从而将决策边界推得更远，以接受更多的潜在样本。

**处理小集群**。损失重新加权有助于防止巨大集群的形成。然而，我们仍然面临着一些小聚类崩溃成空聚类的风险。为了克服这个问题，我们建议在极小的聚类崩溃之前提前处理并消除它们。我们把正常的聚类称为Cn，其大小大于阈值，小聚类称为Cs，对于c∈Cs，我们首先把c中的样本分配给Cn中最近的中心点，使c为空。接下来，我们通过K-Means将最大的聚类cmax∈Cn分成两个子聚类，并随机选择其中一个子聚类作为新的c。尽管这个过程突然改变了一些聚类，但它只影响了参与这个过程的一小部分样本。

**降低维度**。一些骨干网络将图像映射为高维向量，例如，AlexNet产生4,096维的特征，ResNet-50产生2,048维的特征，导致后续聚类的空间和时间都很复杂。DC对整个数据集的特征进行PCA以减少维度。然而，对于ODC来说，不同样本的特征具有不同的时间戳，导致样本之间的统计数据不兼容。因此，PCA就不再适用了。在每次迭代中执行PCA也是很昂贵的。因此，我们增加了一个{fc-bn-relu- dropout-fc-relu}的非线性头层，将高维的特征减少到256维。它在ODC迭代过程中被联合调整。头层在下游任务中被移除。

* 1. ODC for Unsupervised Fine-tuning

与自监督学习方法相比，自监督学习方法倾向于捕捉图像内部语义。基于聚类的方法更注重图像间的信息。因此，DC和ODC是对以前自监督学习方法的自然补充。由于DC和ODC并不局限于一个专门设计的目标，如旋转角度或颜色预测，它们很容易成为一个无监督的微调方案，以提高现有自监督方法的性能。在本文中，我们研究了DC和ODC作为微调过程的有效性，其初始化来自不同的自监督学习方法。

* 1. Implementation Details

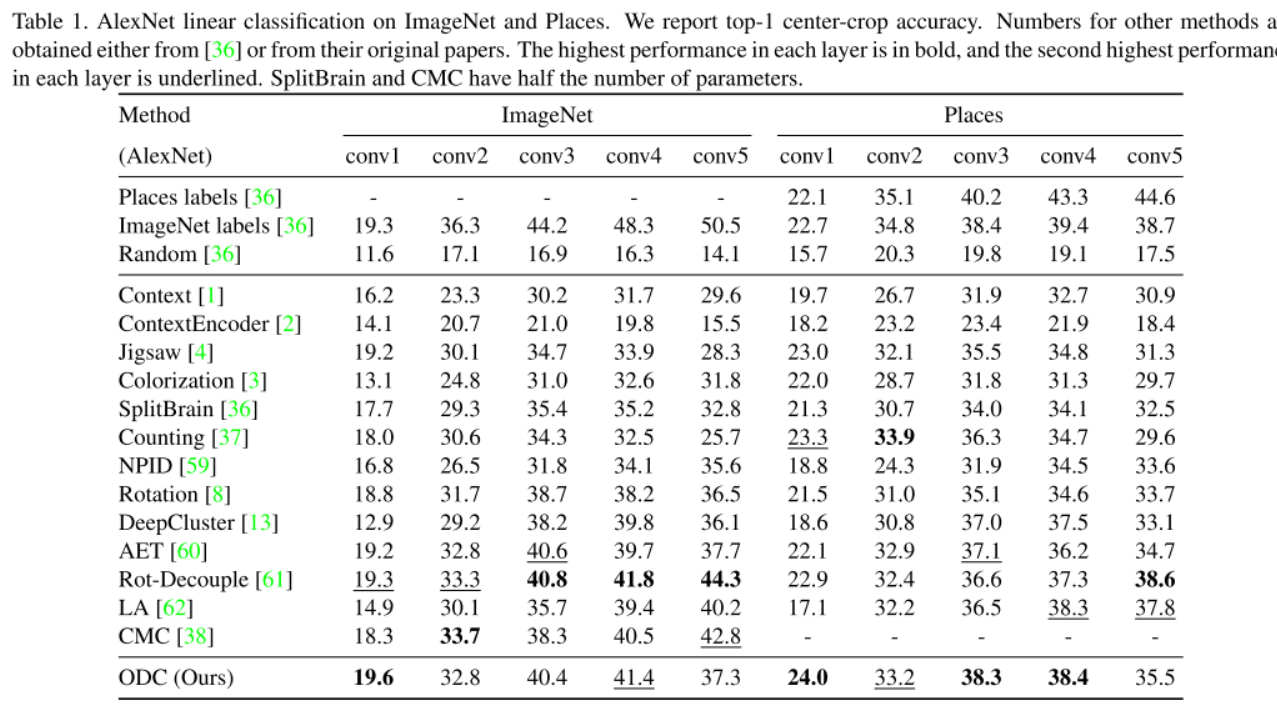
**Data Pre-processing**。我们使用包含128万张无标签图像的ImageNet进行训练。图像首先被随机裁剪为224x224的分辨率，并进行了包括随机翻转和旋转（±2◦）的增强。DC对图像采用Sobel滤波器，以避免利用颜色作为捷径。这样的预处理步骤要求下游任务也包括Sobel层，这有可能限制其应用。我们发现，强色彩抖动在避免捷径方面显示出与Sobel滤波器相同的效果，同时它允许正常的RGB图像作为输入。具体来说，我们采用PyTorch风格的颜色抖动变换，亮度系数（0.6，1.4），对比度系数（0.6，1.4），饱和度系数（0，2）和色调系数（-0.5，0.5）。此外，我们以0.2的概率随机地将图像转换为灰度。应用在训练样本上的随机颜色抖动和灰度使以颜色衡量的相似度随机化。这就阻止了网络从颜色中利用琐碎的信息。

**ODC的训练**。我们使用ResNet-50作为backbone。考虑到大多数早期的工作都使用AlexNet，我们也在AlexNet上进行了实验以进行比较。遵循[13]，我们使用没有局部反应规范化的AlexNet架构，并增加了批量规范化层。AlexNet和ResNet-50的ODC模型是从头开始训练的。批量大小为512，分配给8个GPU。AlexNet的学习率为0.04，ResNet-50的学习率为0.06，持续了400个历时，并在接下来的40个历时中衰减了0.1。在DC之后，集群的数量被设定为10,000个，这比ImageNet注释的类的数量大10倍。动量系数m被设定为0.5。识别小聚类的阈值被设定为20。只要不超过一个聚类中的平均样本数，改变这个阈值不会对结果产生很大影响。中心点存储器每10次迭代更新一次。中心点的更新频率构成了学习效果和效率之间的权衡。在我们的实验中，我们观察到，只要频率被限制在一个合理的范围内，ODC的性能对它不敏感。

1. 实验
   1. Evaluation on Unsupervised Representation

在对ODC模型进行预训练后，我们评估了无监督特征在标准下游任务中的质量，包括ImageNet分类、Places205[63]分类、VOC2007[64]SVM分类和VOC2007 Low-shot分类。我们提供每个基准的细节，并展示我们的竞争结果如下。深度聚类的重新实现。由于DC的原始论文不包括ResNet-50，我们用ResNet-50实现了DC模型。DC模型采用了与ODC相同的数据增强，只是DC在图像上应用了一个Sobel滤波器。为了公平比较，DC的训练超参数与ODC相同，只是我们根据经验发现lr = 0.1更适合DC。ImageNet分类。按照Zhang等人[36]的设置，我们保持包括所有卷积层和批量归一化层在内的骨干层冻结，并对来自不同深度的卷积层的特征进行1000路线性分类器训练。这些特征通过平均池化被映射到大约9000个维度。我们使用动量为0.9的SGD和批次大小为256的SGD来训练所有模型，总共训练了100个epochs。学习率初始化为0.01，每30个历时后衰减10倍。其他的超参数是按照Goyal等人[58]设置的。我们报告了在ImageNet的官方验证分割上的前1名中心裁剪准确率。

对于AlexNet，如表1所示，ODC在所有conv层都比DC有持续的改进，其中conv1层的改进最大（6.7%）。conv1层的表现超过了ImageNet的预训练模型。关于表现最好的层，ODC在conv4层达到了41.4%，超过了最新的LA[62]，仅次于Rot-Decoupling[61]。尽管 ODC 在其最佳表现层上没有超过 Rot-Decoupling，但它为基于旋转的方法提供了一个补充性的视角。ODC在更深的架构中也有很好的扩展性。对于ResNet-50，如表2所示，ODC在conv5层达到了57.6%的中心裁剪精度，比重新实现的DC的最佳表现层高出5.4%。与同时进行的最先进的方法LA[62]相比，我们的方法产生了竞争性的结果。虽然conv5的结果略低于LA，但ODC从conv1到conv4层都以较大的优势胜过LA。我们观察到，从较浅的层到较深的层，性能都在不断提高，这表明ODC充分利用了所有的剩余层。



ODC在更深的架构中也有很好的扩展性。对于ResNet-50，如表2所示，ODC在conv5层达到了57.6%的中心裁剪精度，比重新实现的DC的最佳表现层高出5.4%。与同时进行的最先进的方法LA[62]相比，我们的方法产生了竞争性的结果。虽然conv5的结果略低于LA，但ODC从conv1到conv4层都以很大的幅度超过了LA。我们观察到，从较浅的层到较深的层，性能都在不断提高，这表明ODC充分利用了所有的剩余层。Places205分类。继Zhang等人[36]之后，为了测试在其他领域的泛化能力，我们还将学到的模型转移到Places205数据集，该数据集包含205个场景类别的245万张图像。与ImageNet上的实验类似，我们在Places205的训练分割上，在每个冻结卷积层之上训练一个205路线性分类器，并在标准验证分割上报告top-1中心裁剪的准确性。评估设置和超参数与ImageNet分类任务中的参数相同。

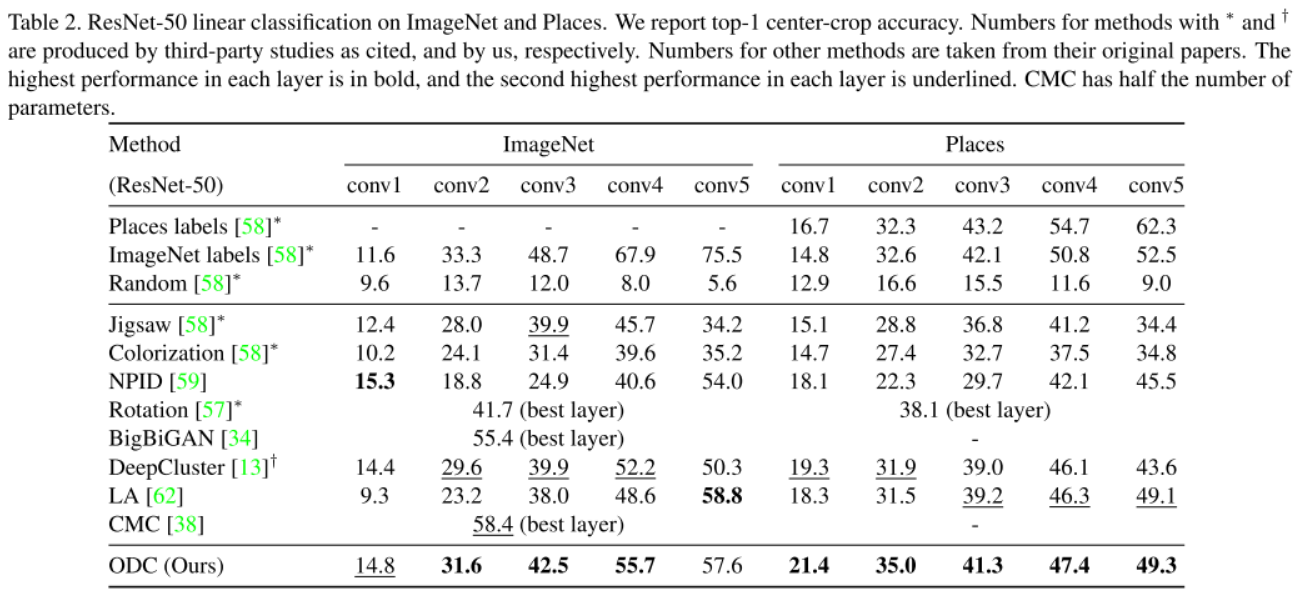
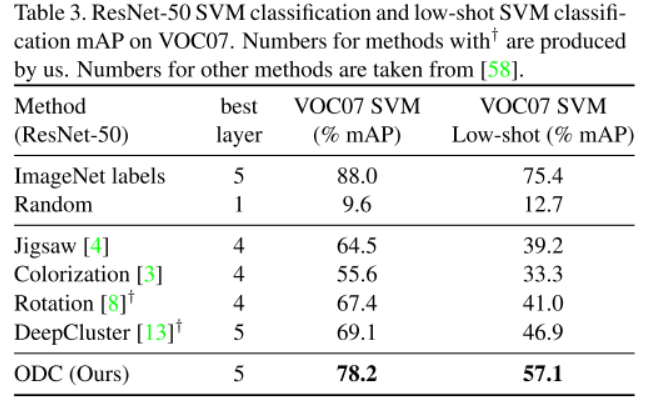
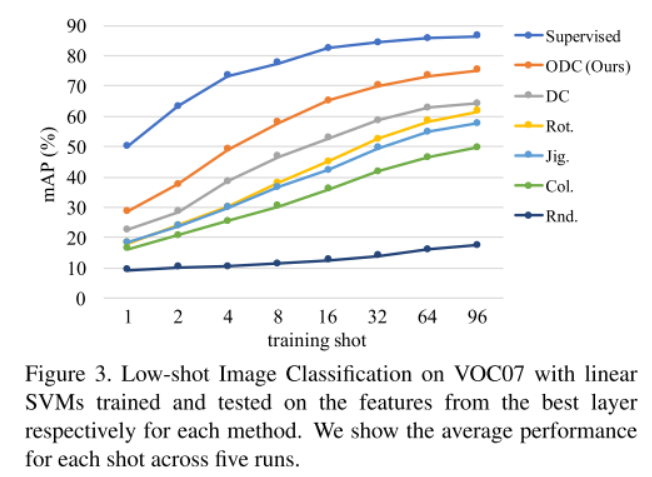


表1中的结果显示，以AlexNet为骨干的ODC在各层的表现也都优于DC。ODC在conv1、conv3和conv4层都超过了以前的所有作品。与ImageNet分类任务中的观察结果类似，当ODC与ResNet-50转移到Places205时，它在更深的架构上有很好的扩展性。如表2所示，在所有层中，ODC都超过了以前的所有作品，在conv2层中观察到与亚军的最大差距（3.1%）。关于表现最好的层，ODC在conv5层达到了49.3%的中心裁剪精度，在各自的最佳层中超过了重新实现的DC 3.2%。我们观察到ODC在conv1和conv2层的表现优于使用Places标签或ImageNet标签的监督模型。我们的方法在Places205分类任务中的转移性能表明，通过ODC学习的表征可以很好地推广到ImageNet的不同领域。VOC2007 SVM分类。为了进一步评估所学特征的泛化，我们在VOC2007转移学习任务中进行了实验，该任务类似于具有较小数据集的实际应用。按照文献[58]，我们在VOC2007的 "trainval "部分对从冷冻骨干中提取的特征进行线性SVM训练，并在测试部分进行评估。我们沿用了[58]中的测试设置和超参数，并报告了ResNet-50的不同方法的最佳表现层。表3中的结果显示，在VOC2007的SVM分类任务上，ODC以很大的优势超过了以前的方法。通过ODC，我们实现了78.2%的mAP性能，比DC高9.1%。然而，我们也注意到，我们的ODC与用Ima GeNet标签预训练的监督模型之间仍有9.8%的显著性能差距，这为进一步探索留下了余地。

Low-Shot：VOC2007 SVM分类。继[58]之后，我们还将学到的表征转移到VOC2007 SVM分类的Low-Shot中，以测试每个类别的训练实例很少时的特征质量。我们改变了每个类别中阳性样本的数量，并使用VOC2007 SVM分类的相同设置在冻结的ResNet-50骨干上训练线性SVMs。在训练中使用VOC2007的标准 "trainval "分割，在测试中使用test分割。我们在图3中报告了不同低射值的五个独立样本的平均精度（mAP）。表3中显示的最终mAP结果是所有低射值和所有独立运行的平均数。图3中显示的是每次拍摄的结果。ODC在每次拍摄中都比DC有持续的改进，当每类允许有更多的正面例子时，性能差距进一步加大。我们还观察到，随着训练次数的增加，ODC和用ImageNet标签预训练的监督模型之间的性能差距也逐渐缩小了。表3显示，ODC在VOC2007上的低镜头SVM分类中取得了57.1%的mAP性能，比我们的同行DC高出10.2%。ODC在该基准中的低照度结果表明，通过ODC学习的特征在低照度分类中具有良好的通用性。





1. 评析