# Bootstrap Your Own Latent A New Approach to Self-Supervised Learning

# 自我引导潜在空间: 一种自我监督学习的新方法

Jean-Bastien Grill*1, Florian Strub*1, Florent Altché*, 1, Corentin Tallec*, 1, Pierre H. Richemond *, 1, 2 Elena Buchatskaya [1] , Carl Doersch [1] , Bernardo Avila Pires [1] , Zhaohan Daniel Guo [1]
Jean-Bastien Grill*1, Florian Strub*1, Florent Altché*, 1, Corentin Tallec*, 1, Pierre H. Richemond *, 1, 2 Elena Buchatskaya [1] , Carl Doersch [1] , Bernardo Avila Pires [1] , Zhaohan Daniel Guo [1]
Mohammad Gheshlaghi Azar [1] , Bilal Piot [1] , Koray Kavukcuoglu [1] , Rémi Munos [1] , Michal Valko [1]
Mohammad Gheshlaghi Azar [1] , Bilal Piot [1] , Koray Kavukcuoglu [1] , Rémi Munos [1] , Michal Valko [1]
[1] DeepMind [2] Imperial College
[1] DeepMind [2] 伦敦帝国学院
[jbgrill, fstrub, altche, corentint, richemond]@google.com
[jbgrill, fstrub, altche, corentint, richemond]@google.com

## Abstract

## 摘要

We introduce Bootstrap Your Own Latent (BYOL), a new approach to self-supervised image representation learning. BYOL relies on two neural networks, referred to as online and target networks, that interact and learn from each other. From an augmented view of an image, we train the online network to predict the target network representation of the same image under a different augmented view. At the same time, we update the target network with a slow-moving average of the online network. While state-of-the art methods rely on negative pairs, BYOL achieves a new state of the art without them. BYOL reaches 74.3% top-1 classification accuracy on ImageNet using a linear evaluation with a ResNet-50 architecture and 79.6% with a larger ResNet. We show that BYOL performs on par or better than the current state of the art on both transfer and semi-supervised benchmarks. Our implementation and pretrained models are given on GitHub. [3]

我们介绍了自我引导潜在空间 (BYOL)，这是一种新的自我监督图像表示学习方法。BYOL 依赖于两个神经网络，称为在线网络和目标网络，它们相互交互并学习。从图像的增强视图出发，我们训练在线网络以预测同一图像在不同增强视图下的目标网络表示。同时，我们使用在线网络的慢速移动平均来更新目标网络。尽管最先进的方法依赖于负对，BYOL 在没有负对的情况下实现了新的最先进水平。使用 ResNet-50 架构和 79.6% 更大的 ResNet，BYOL 在 ImageNet 上达到了 74.3% 的顶级分类准确率。我们展示了 BYOL 在迁移和半监督基准测试中表现与当前最先进水平相当或更好。我们的实现和预训练模型已在 GitHub 上提供。[3]

## 1 Introduction

## 1 引言

Learning good image representations is a key challenge in computer vision [1, 2, 3] as it allows for efficient training on downstream tasks [4, 5, 6, 7] . Many different training approaches have been proposed to learn such representations, usually relying on visual pretext tasks. Among them, state-of-the-art contrastive methods [8, 9, 10, 11, 12] are trained by reducing the distance between representations of different augmented views of the same image ('positive pairs'), and increasing the distance between representations of augmented views from different images ('negative pairs'). These methods need careful treatment of negative pairs [13] by either relying on large batch sizes [8, 12] , memory banks [9] or customized mining strategies [14, 15] to retrieve the negative pairs. In addition, their performance critically depends on the choice of image augmentations [34, 11, 8, 12] .

学习良好的图像表示是计算机视觉中的一个关键挑战 [1, 2, 3] ，因为它允许在下游任务上进行高效的训练 [4, 5, 6, 7] 。已经提出了许多不同的训练方法来学习这样的表示，通常依赖于视觉预文本任务。在这些方法中，最先进的对比方法 [8, 9, 10, 11, 12] 通过减少同一图像不同增强视图的表示之间的距离（“正对”）以及增加来自不同图像的增强视图的表示之间的距离（“负对”）进行训练。这些方法需要对负对

进行仔细处理 [13]，通常依赖于大批量大小 [8, 12] 、内存库 [9] 或定制的挖掘策略 [14, 15] 来检索负对。此外，它们的性能在很大程度上依赖于图像增强的选择 [34, 11, 8, 12] 。

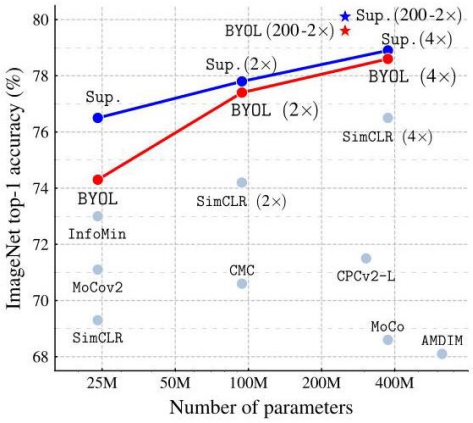



Figure 1: Performance of BYOL on ImageNet (linear evaluation) using ResNet-50 and our best architecture ResNet-200 (2×) , compared to other unsupervised and supervised (Sup.) baselines [8].

图 1: 使用 ResNet-50 和我们最佳架构 ResNet-200 (2×) 在 ImageNet 上的 BYOL 性能 (线性评估)，与其他无监督和监督 (Sup.) 基线 [8] 进行比较。

In this paper, we introduce Bootstrap Your Own Latent (BYOL), a new algorithm for self-supervised learning of image representations. BYOL achieves higher performance than state-of-the-art contrastive methods without using negative pairs. It iteratively bootstraps [4] the outputs of a network to serve as targets for an enhanced representation. Moreover, BYOL is more robust to the choice of image augmentations than contrastive methods; we suspect that not relying on negative pairs is one of the leading reasons for its improved robustness. While previous methods based on bootstrapping have used pseudo-labels [16], cluster indices [17] or a handful of labels [18, 19, 20], we propose to directly bootstrap the representations. In particular, BYOL uses two neural networks, referred to as online and target networks, that interact and learn from each other. Starting from an augmented view of an image, BYOL trains its online network to predict the target network's representation of another augmented view of the same image. While this objective admits collapsed solutions, e.g., outputting the same vector for all images, we empirically show that BYOL does not converge to such solutions. We hypothesize (Section 3.2) that the combination of (i) the addition of a predictor to the online network and (ii) the use of a moving average of online parameters, as the target network encourages encoding more and more information in the online projection and avoids collapsed solutions.

在本文中，我们介绍了一种新的自监督学习图像表示的算法——自引导潜在 (BYOL)。BYOL 在不使用负样本对的情况下，达到了比最先进的对比方法更高的性能。它通过迭代自引导 [4] 网络的输出，以作为增强表示的目标。此外，BYOL 对图像增强的选择比对比方法更具鲁棒性；我们怀疑不依赖负样本对是其改进鲁棒性的主要原因之一。虽然基于自引导的先前方法使用了伪标签 [16]、聚类索引 [17] 或少量标签 [18, 19, 20]，我们建议直接自引导表示。具体而言，BYOL 使用两个神经网络，称为在线网络和目标网络，它们相互交互并学习。从图像的增强视图开始，BYOL 训练其在线网络以预测目标网络对同一图像的另一个增强视图的表示。尽管这个目标允许收敛到崩溃的解决方案，例如对所有图像输出相同的向量，但我们通过实验证明，BYOL 不会收敛到这样的解决方案。我们假设 (第 3.2 节)(i) 向在线网络添加预测器和 (ii) 使用在线参数的移动平均，能够促使目标网络鼓励在在线投影中编码越来越多的信息，从而避免崩溃的解决方案。

We evaluate the representation learned by BYOL on ImageNet [21] and other vision benchmarks using ResNet architectures [22]. Under the linear evaluation protocol on ImageNet, consisting in training a linear classifier on top of the frozen representation, BYOL reaches 74.3% top-1 accuracy with a standard ResNet-50 and 79.6% top-1 accuracy with a larger ResNet (Figure 1). In the semi-supervised and transfer settings on ImageNet, we obtain results on par or superior to the current state of the art. Our contributions are: (i) We introduce BYOL, a self-supervised representation learning method (Section 3)

*Equal contribution; the order of first authors was randomly selected.
* 平等贡献；第一作者的顺序是随机选择的。
[3] https://github.com/deepmind/deepmind-research/tree/master/byol
[3] https://github.com/deepmind/deepmind-research/tree/master/byol

which achieves state-of-the-art results under the linear evaluation protocol on ImageNet without using negative pairs. (ii) We show that our learned representation outperforms the state of the art on semi-supervised and transfer benchmarks (Section 4). (iii) We show that BYOL is more resilient to changes in the batch size and in the set of image augmentations compared to its contrastive counterparts (Section 5). In particular, BYOL suffers a much smaller performance drop than SimCLR, a strong contrastive baseline, when only using random crops as image augmentations.

我们在 ImageNet [21] 和其他视觉基准上使用 ResNet 架构 [22] 评估 BYOL 学习到的表示。在 ImageNet 的线性评估协议下，训练一个线性分类器在冻结的表示之上，BYOL 在标准 ResNet-50 上达到 74.3% 的 top-1 准确率，在更大的 ResNet 上达到 79.6% 的 top-1 准确率 (图 1)。在 ImageNet 的半监督和迁移设置中，我们获得的结果与当前的最先进水平相当或更优。我们的贡献包括:(i) 我们引入了 BYOL，一种自监督表示学习方法 (第 3 节)，在不使用负对的情况下，在 ImageNet 的线性评估协议下实现了最先进的结果。(ii) 我们展示了我们学习到的表示在半监督和迁移基准上优于最先进水平 (第 4 节)。(iii) 我们表明，与其对比方法相比，BYOL 对批量大小和图像增强集的变化更具韧性 (第 5 节)。特别是，当仅使用随机裁剪作为图像增强时，BYOL 的性能下降远小于 SimCLR，这是一种强大的对比基线。

## 2 Related work

## 2 相关工作

Most unsupervised methods for representation learning can be categorized as either generative or discriminative [23, 8] . Generative approaches to representation learning build a distribution over data and latent embedding and use the learned embeddings as image representations. Many of these approaches rely either on auto-encoding of images [24,25,26] or on adversarial learning [27], jointly modelling data and representation [28, 29, 30, 31] . Generative methods typically operate directly in pixel space. This however is computationally expensive, and the high level of detail required for image generation may not be necessary for representation learning.

大多数无监督的表示学习方法可以分为生成式或判别式 [23, 8] 。生成式表示学习方法构建数据和潜在嵌入的分布，并使用学习到的嵌入作为图像表示。这些方法中的许多依赖于图像的自编码 [24,25,26] 或对抗学习 [27]，共同建模数据和表示 [28, 29, 30, 31] 。生成方法通常直接在像素空间中操作。然而，这在计算上是昂贵的，并且图像生成所需的高细节水平可能对表示学习并不是必要的。

Among discriminative methods, contrastive methods [9, 10, 32, 33, 34, 11, 35, 36] currently achieve state-of-the-art performance in self-supervised learning [37, 8, 38, 12]. Contrastive approaches avoid a costly generation step in pixel space by bringing representation of different views of the same image closer ('positive pairs'), and spreading representations of views from different images ('negative pairs') apart [39, 40]. Contrastive methods often require comparing each example with many other examples to work well [9, 8] prompting the question of whether using negative pairs is necessary.

在判别方法中，对比方法 [9, 10, 32, 33, 34, 11, 35, 36] 目前在自监督学习中实现了最先进的性能 [37, 8, 38, 12]。对比方法通过将同一图像的不同视图的表示拉近 ( “正对” )，并将来自不同图像的视图的表示分开 ( “负对” )，从而避免了在像素空间中代价高昂的生成步骤 [39, 40]。对比方法通常需要将每个示例与许多其他示例进行比较以获得良好的效果 [9, 8] ，这引发了使用负对是否必要的问题。

DeepCluster [17] partially answers this question. It uses bootstrapping on previous versions of its representation to produce targets for the next representation; it clusters data points using the prior representation, and uses the cluster index of each sample as a classification target for the new representation. While avoiding the use of negative pairs, this requires a costly clustering phase and specific precautions to avoid collapsing to trivial solutions.

DeepCluster [17] 部分回答了这个问题。它利用先前版本的表示进行自举，以生成下一个表示的目标；它使用先前的表示对数据点进行聚类，并将每个样本的聚类索引作为新表示的分类目标。虽然避免了使用负对，但这需要一个代价高昂的聚类阶段和特定的预防措施，以避免陷入平凡解。

Some self-supervised methods are not contrastive but rely on using auxiliary handcrafted prediction tasks to learn their representation. In particular, relative patch prediction [23, 40], colorizing gray-scale images [41, 42], image inpainting [43], image jigsaw puzzle [44], image super-resolution [45], and geometric transformations [46, 47] have been shown to be useful. Yet, even with suitable architectures [48], these methods are being outperformed by contrastive methods [37, 8, 12].

一些自监督方法不是对比性的，而是依赖于使用辅助手工预测任务来学习其表示。特别是，相对补丁预测 [23, 40]、为灰度图像上色 [41, 42]、图像修复 [43]、图像拼图 [44]、图像超分辨率 [45] 和几何变换 [46, 47] 已被证明是有用的。然而，即使有合适的架构 [48]，这些方法仍然被对比方法 [37, 8, 12] 超越。

Our approach has some similarities with Predictions of Bootstrapped Latents (PBL, [49]), a self-supervised representation learning technique for reinforcement learning (RL). PBL jointly trains the agent's history representation and an encoding of future observations. The observation encoding is used as a target to train the agent's representation, and the agent's representation as a target to train the observation encoding. Unlike PBL, BYOL uses a slow-moving average of its representation to provide its targets, and does not require a second network.

我们的方法与自助潜变量预测 (PBL, [49]) 有一些相似之处，PBL 是一种用于强化学习 (RL) 的自监督表示学习技术。PBL 联合训练代理的历史表示和未来观察的编码。观察编码被用作训练代理表示的目标，而代理的表示则作为训练观察编码的目标。与 PBL 不同，BYOL 使用其表示的慢动平均值来提供目标，并且不需要第二个网络。

The idea of using a slow-moving average target network to produce stable targets for the online network was inspired by deep RL [50, 51, 52, 53] . Target networks stabilize the bootstrapping updates provided by the Bellman equation, making them appealing to stabilize the bootstrap mechanism in BYOL. While most RL methods use fixed target networks, BYOL uses a weighted moving average of previous networks (as in [54]) in order to provide smoother changes in the target representation.

使用慢动平均目标网络为在线网络生成稳定目标的想法受到深度强化学习 [50, 51, 52, 53] 的启发。目标网络稳定了贝尔曼方程提供的自助更新，使其在稳定 BYOL 中的自助机制时变得更具吸引力。虽然大多数强化学习方法使用固定目标网络，但 BYOL 使用之前网络的加权移动平均 (如 [54] 所示) 以提供目标表示的更平滑变化。

In the semi-supervised setting [55, 56], an unsupervised loss is combined with a classification loss over a handful of labels to ground the training [19, 20, 57, 58, 59, 60, 61, 62]. Among these methods, mean teacher (MT) [20] also uses a slow-moving average network, called teacher, to produce targets for an online network, called student. An $\ell_2$ consistency loss between the softmax predictions of the teacher and the student is added to the classification loss. While [20] demonstrates the effectiveness of MT in the semi-supervised learning case, in Section 5 we show that a similar approach collapses when removing the classification loss. In contrast, BYOL introduces an additional predictor on top of the online network, which prevents collapse.

在半监督设置中 [55, 56]，无监督损失与少量标签的分类损失相结合，以为训练提供基础 [19, 20, 57, 58, 59, 60, 61, 62]。在这些方法中，平均教师 (MT) [20] 也使用一个称为教师的慢动平均网络，为一个称为学生的在线网络生成目标。教师和学生的 softmax 预测之间的 $\ell_2$ 一致性损失被添加到分类损失中。虽然 [20] 证明了 MT 在半监督学习中的有效性，但在第 5 节中我们展示了在去除分类损失时类似的方法会崩溃。相反，BYOL 在在线网络之上引入了一个额外的预测器，从而防止了崩溃。

Finally, in self-supervised learning, MoCo [9] uses a slow-moving average network (momentum encoder) to maintain consistent representations of negative pairs drawn from a memory bank. Instead, BYOL uses a moving average network to produce prediction targets as a means of stabilizing the bootstrap step. We show in Section 5 that this mere stabilizing effect can also improve existing contrastive methods.

最后，在自监督学习中，MoCo [9] 使用一个缓慢移动的平均网络 (动量编码器) 来保持从内存库中提取的负对的稳定表示。相反，BYOL 使用一个移动平均网络来生成预测目标，以稳定引导步骤。我们在第 5 节中展示，这种简单的稳定效应也可以改善现有的对比方法。

# 3 Method

# 3 方法

We start by motivating our method before explaining its details in Section 3.1. Many successful self-supervised learning approaches build upon the cross-view prediction framework introduced in [63]. Typically, these approaches learn representations by predicting different views (e.g., different random crops) of the same image from one another. Many such approaches cast the prediction problem directly in representation space: the representation of an augmented view of an image should be predictive of the representation of another augmented view of the same image. However, predicting directly in representation space can lead to collapsed representations: for instance, a representation that is constant across views is always fully predictive of itself. Contrastive methods circumvent this problem by reformulating the prediction problem into one of discrimination: from the representation of an augmented view, they

[4] Throughout this paper, the term bootstrap is used in its idiomatic sense rather than the statistical sense.
[4] 在本文中，术语自举是以其习惯用法而非统计意义使用的。

learn to discriminate between the representation of another augmented view of the same image, and the representations of augmented views of different images. In the vast majority of cases, this prevents the training from finding collapsed representations. Yet, this discriminative approach typically requires comparing each representation of an augmented view with many negative examples, to find ones sufficiently close to make the discrimination task challenging. In this work, we thus tasked ourselves to find out whether these negative examples are indispensable to prevent collapsing while preserving high performance.

我们首先阐述我们方法的动机，然后在第 3.1 节中解释其细节。许多成功的自监督学习方法建立在 [63] 中提出的跨视图预测框架之上。通常，这些方法通过预测同一图像的不同视图 (例如，不同的随机裁剪) 来学习表示。许多此类方法直接在表示空间中进行预测问题的表述: 增强视图的图像表示应该能够预测同一图像的另一个增强视图的表示。然而，直接在表示空间中进行预测可能导致表示的崩溃: 例如，一个在不同视图中保持不变的表示总是能够完全预测其自身。对比方法通过将预测问题重新表述为区分问题来规避这个问题: 从增强视图的表示中，它们学习区分同一图像的另一个增强视图的表示与不同图像的增强视图的表示。在绝大多数情况下，这防止了训练找到崩溃的表示。然而，这种区分方法通常需要将每个增强视图的表示与许多负例进行比较，以找到足够接近的负例，使得区分任务具有挑战性。因此，在本研究中，我们的任务是找出这些负例是否是防止崩溃而保持高性能的不可或缺的。

To prevent collapse, a straightforward solution is to use a fixed randomly initialized network to produce the targets for our predictions. While avoiding collapse, it empirically does not result in very good representations. Nonetheless, it is interesting to note that the representation obtained using this procedure can already be much better than the initial fixed representation. In our ablation study (Section 5), we apply this procedure by predicting a fixed randomly initialized network and achieve 18.8% top-1 accuracy (Table 5a) on the linear evaluation protocol on ImageNet, whereas the randomly initialized network only achieves 1.4% by itself. This experimental finding is the core motivation for BYOL: from a given representation, referred to as target, we can train a new, potentially enhanced representation, referred to as online, by predicting the target representation. From there, we can expect to build a sequence of representations of increasing quality by iterating this procedure, using subsequent online networks as new target networks for further training. In practice, BYOL generalizes this bootstrapping procedure by iteratively refining its representation, but using a slowly moving exponential average of the online network as the target network instead of fixed checkpoints.

为了防止崩溃，一个简单的解决方案是使用一个固定的随机初始化网络来生成我们预测的目标。虽然避免了崩溃，但经验上并没有产生非常好的表示。然而，有趣的是，使用该程序获得的表示已经比初始固定表示要好得多。在我们的消融研究中 (第 5 节)，我们通过预测一个固定的随机初始化网络来应用这一程序，并在 ImageNet 的线性评估协议上达到了 18.8% 的 top-1 准确率 (表 5a)，而随机初始化网络自身仅达到了 1.4% 。这一实验发现是 BYOL 的核心动机: 从给定的表示 (称为目标)，我们可以通过预测目标表示来训练一个新的、潜在增强的表示 (称为在线表示)。由此，我们可以期待通过迭代这一过程，使用后续的在线网络作为新的目标网络进行进一步训练，从而构建一系列质量不断提高的表示。在实践中，BYOL 通过迭代精炼其表示来推广这一自举程序，但使用在线网络的缓慢移动指数平均值作为目标网络，而不是固定的检查点。

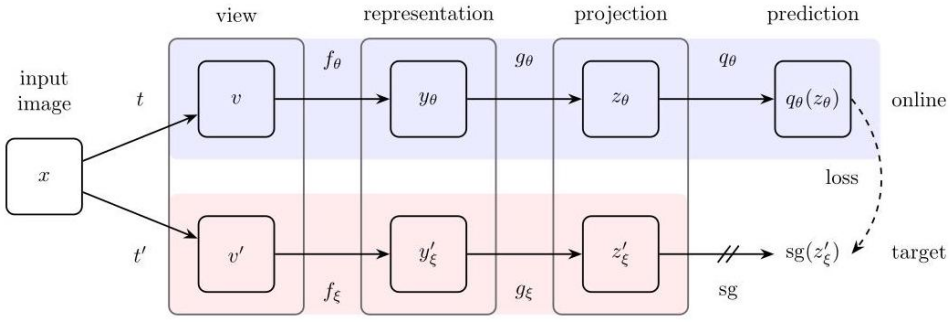



Figure 2: BYOL' s architecture. BYOL minimizes a similarity loss between $q_\theta(z_\theta)$ and $\mathrm{sg}\left(z'_\xi\right)$, where $\theta$ are the trained weights, $\xi$ are an exponential moving average of $\theta$ and sg means stop-gradient. At the end of training, everything but $f_\theta$ is discarded, and $y_\theta$ is used as the image representation.

图 2:BYOL 的架构。BYOL 最小化 $q_\theta(z_\theta)$ 和 $\mathrm{sg}\left(z'_\xi\right)$ 之间的相似性损失，其中 $\theta$ 是训练权重，$\xi$ 是 $\theta$ 的指数移动平均，而 sg 表示停止梯度。在训练结束时，除了 $f_\theta$ 之外的所有内容都被丢弃，$y_\theta$ 被用作图像表示。

## 3.1 Description of BYOL

## 3.1 BYOL 的描述

BYOL' s goal is to learn a representation $y_\theta$ which can then be used for downstream tasks. As described previously, BYOL uses two neural networks to learn: the online and target networks. The online network is defined by a set of weights $\theta$ and is comprised of three stages: an encoder $f_\theta$ , a projector $g_\theta$ and a predictor $q_\theta$ , as shown in Figure 2 and Figure 8. The target network has the same architecture as the online network, but uses a different set of weights $\xi$ . The target network provides the regression targets to train the online network, and its parameters $\xi$ are an exponential moving average of the online parameters $\theta$ [54]. More precisely, given a target decay rate $\tau \in [0,1]$ , after each training step we perform the following update,

BYOL 的目标是学习一个表示 $y_\theta$ ，该表示可以用于下游任务。如前所述，BYOL 使用两个神经网络进行学习: 在线网络和目标网络。在线网络由一组权重 $\theta$ 定义，并由三个阶段组成: 编码器 $f_\theta$ 、投影器 $g_\theta$ 和预测器 $q_\theta$ ，如图 2 和图 8 所示。目标网络与在线网络具有相同的架构，但使用不同的权重集 $\xi$ 。目标网络提供回归目标以训练在线网络，其参数 $\xi$ 是在线参数 $\theta$ 的指数移动平均 [54]。更准确地说，给定一个目标衰减率 $\tau \in [0,1]$ ，在每个训练步骤后，我们执行以下更新，

$$\xi \leftarrow \tau\xi + (1-\tau)\,\theta \tag{1}$$

Given a set of images $\mathcal{D}$ , an image $x \sim \mathcal{D}$ sampled uniformly from $\mathcal{D}$ , and two distributions of image augmentations $\mathcal{T}$ and $\mathcal{T}'$ , BYOL produces two augmented views $v \triangleq t(x)$ and $v' \triangleq t'(x)$ from $x$ by applying respectively image augmentations $t \sim \mathcal{T}$ and $t' \sim \mathcal{T}'$ . From the first augmented view $v$ , the online network outputs a representation $y_\theta \triangleq f_\theta(v)$ and a projection $z_\theta \triangleq g_\theta(y)$ . The target network outputs $y'_\xi \triangleq f_\xi(v')$ and the target projection $z'_\xi \triangleq g_\xi(y')$ from the second augmented view $v'$ . We then output a prediction $q_\theta(z_\theta)$ of $z'_\xi$ and $\ell_2$ -normalize both $q_\theta(z_\theta)$ and $z'_\xi$ to $\overline{q_\theta}(z_\theta) \triangleq q_\theta(z_\theta)/\left\|q_\theta(z_\theta)\right\|_2$ and $\bar{z}'_\xi \triangleq z'_\xi/\left\|z'_\xi\right\|_2$ . Note that this predictor is only applied to the online branch, making the architecture asymmetric between the online and target pipeline. Finally we define the following mean squared error between the normalized predictions and target projections, [5]

给定一组图像 $\mathcal{D}$ ，从 $\mathcal{D}$ 中均匀抽样的一幅图像 $x \sim \mathcal{D}$ ，以及两种图像增强的分布 $\mathcal{T}$ 和 $\mathcal{T}'$ ，BYOL 通过分别应用图像增强 $t \sim \mathcal{T}$ 和 $t' \sim \mathcal{T}'$ 从 $x$ 生成两个增强视图 $v \triangleq t(x)$ 和 $v' \triangleq t'(x)$ 。从第一个增强视图 $v$ ，在线网络输出一个表示 $y_\theta \triangleq f_\theta(v)$ 和一个投影 $z_\theta \triangleq g_\theta(y)$ 。目标网络从第二个增强视图 $v'$ 输出 $y'_\xi \triangleq f_\xi(v')$ 和目标投影 $z'_\xi \triangleq g_\xi(y')$ 。然后我们输出 $z'_\xi$ 和 $\ell_2$ 的预测 $q_\theta(z_\theta)$ ，并将 $q_\theta(z_\theta)$ 和 $z'_\xi$ 归一化到 $\overline{q_\theta}(z_\theta) \triangleq q_\theta(z_\theta)/\left\|q_\theta(z_\theta)\right\|_2$ 和 $\bar{z}'_\xi \triangleq z'_\xi/\left\|z'_\xi\right\|_2$ 。请注意，这个预测器仅应用于在线分支，使得在线和目标管道之间的架构不对称。最后，我们定义归一化预测和目标投影之间的均方误差 [5] 。

$$\mathcal{L}_{\theta,\xi} \triangleq \left\|\overline{q_\theta}(z_\theta) - \bar{z}'_\xi\right\|_2^2 = 2 - 2\cdot\frac{\left\langle q_\theta(z_\theta), z'_\xi\right\rangle}{\left\|q_\theta(z_\theta)\right\|_2 \cdot \left\|z'_\xi\right\|_2}. \tag{2}$$

We symmetrize the loss $\mathcal{L}_{\theta,\xi}$ in Eq. 2 by separately feeding $v'$ to the online network and $v$ to the target network to compute $\widetilde{\mathcal{L}}_{\theta,\xi}$ . At each training step, we perform a stochastic optimization step to minimize $\mathcal{L}^{\mathrm{BYOL}}_{\theta,\xi} = \mathcal{L}_{\theta,\xi} + \widetilde{\mathcal{L}}_{\theta,\xi}$ with respect to $\theta$ only, but not $\xi$ , as depicted by the stop-gradient in Figure 2. BYOL's dynamics are summarized as

我们通过分别将 $v'$ 输入在线网络，将 $v$ 输入目标网络来对损失 $\mathcal{L}_{\theta,\xi}$ 进行对称化，以计算 $\widetilde{\mathcal{L}}_{\theta,\xi}$ 。在每个训练步骤中，我们执行随机优化步骤，以最小化相对于 $\theta$ 的 $\mathcal{L}^{\mathrm{BYOL}}_{\theta,\xi} = \mathcal{L}_{\theta,\xi} + \widetilde{\mathcal{L}}_{\theta,\xi}$ ，而不是 $\xi$ ，如图 2 中的停止梯度所示。BYOL 的动态总结如下

$$\theta \leftarrow \text{optimizer}\left(\theta, \nabla_\theta \mathcal{L}^{\mathrm{BYOL}}_{\theta,\xi}, \eta\right) \quad \text{and} \quad \xi \leftarrow \tau\xi + (1-\tau)\,\theta,$$

where optimizer is an optimizer and $\eta$ is a learning rate. At the end of training, we only keep the encoder $f_\theta$ ; as in [9]. When comparing to other methods, we consider the number of inference-time weights only in the final representation $f_\theta$ . The architecture, hyper-parameters and training details are specified in Appendix A, the full training procedure is summarized in Appendix B, and python pseudo-code based on the libraries JAX [64] and Haiku [65] is provided in in Appendix J.

其中优化器是一个优化器，$\eta$ 是学习率。在训练结束时，我们只保留编码器 $f_\theta$ ；如 [9] 所示。在与其他方法比较时，我们仅考虑最终表示 $f_\theta$ 中的推理时权重数量。架构、超参数和训练细节在附录 A 中指定，完整的训练过程在附录 B 中总结，基于 JAX [64] 和 Haiku [65] 库的 Python 伪代码在附录 J 中提供。

## 3.2 Intuitions on BYOL's behavior

## 3.2 对 BYOL 行为的直觉

As BYOL does not use an explicit term to prevent collapse (such as negative examples [10]) while minimizing $\mathcal{L}_{\theta,\xi}^{\text{BYOL}}$ with respect to $\theta$ , it may seem that BYOL should converge to a minimum of this loss

由于 BYOL 在最小化相对于 $\theta$ 的 $\mathcal{L}_{\theta,\xi}^{\text{BYOL}}$ 时并没有使用显式项来防止崩溃 (例如负样本 [10])，因此似乎 BYOL 应该收敛到该损失的最小值

| Method | Top-1 | Top-5 |
|---|---|---|
| Local Agg. | 60.2 | - |
| PIRL [35] | 63.6 | - |
| CPC v2 [32] | 63.8 | 85.3 |
| CMC [11] | 66.2 | 87.0 |
| SimCLR [8] | 69.3 | 89.0 |
| MoCo v2 [37] | 71.1 | - |
| InfoMin Aug. [12] | 73.0 | 91.1 |
| BYOL (ours) | 74.3 | 91.6 |

| 方法 | Top-1 | Top-5 |
|---|---|---|
| 本地聚合 | 60.2 | - |
| PIRL [35] | 63.6 | - |
| CPC v2 [32] | 63.8 | 85.3 |
| CMC [11] | 66.2 | 87.0 |
| SimCLR [8] | 69.3 | 89.0 |
| MoCo v2 [37] | 71.1 | - |
| InfoMin Aug. [12] | 73.0 | 91.1 |
| BYOL(我们的) | 74.3 | 91.6 |

(a) ResNet-50 encoder.

(a) ResNet-50 编码器。

| Method | Architecture | Param. | Top-1 | Top-5 |
|---|---|---|---|---|
| SimCLR [8] | ResNet-50 (2×) | 94M | 74.2 | 92.0 |
| CMC [11] | ResNet-50 (2×) | 94M | 70.6 | 89.7 |
| BYOL (ours) | ResNet-50 (2×) | 94M | 77.4 | 93.6 |
| CPC v2 [32] | ResNet-161 | 305M | 71.5 | 90.1 |
| MoCo [9] | ResNet-50 (4×) | 375M | 68.6 | - |
| SimCLR [8] | ResNet-50 (4×) | 375M | 76.5 | 93.2 |
| BYOL (ours) | ResNet-50 (4×) | 375M | 78.6 | 94.2 |
| BYOL (ours) | ResNet-200 (2×) | 250M | 79.6 | 94.8 |

| 方法 | 架构 | 参数 | Top-1 | Top-5 |
|---|---|---|---|---|
| SimCLR [8] | ResNet-50 (2×) | 94M | 74.2 | 92.0 |
| CMC [11] | ResNet-50 (2×) | 94M | 70.6 | 89.7 |
| BYOL (我们的) | ResNet-50 (2×) | 94M | 77.4 | 93.6 |
| CPC v2 [32] | ResNet-161 | 305M | 71.5 | 90.1 |
| MoCo [9] | ResNet-50 (4×) | 375M | 68.6 | - |
| SimCLR [8] | ResNet-50 (4×) | 375M | 76.5 | 93.2 |
| BYOL (我们的) | ResNet-50 (4×) | 375M | 78.6 | 94.2 |
| BYOL (我们的) | ResNet-200 (2×) | 250M | 79.6 | 94.8 |

(b) Other ResNet encoder architectures.

[5] While we could directly predict the representation $y$ and not a projection $z$ , previous work [8] have empirically shown that using this projection improves performance.

[5] 虽然我们可以直接预测表示 $y$ 而不是投影 $z$ ，但之前的工作 [8] 已经实证表明，使用该投影可以提高性能。

(b) 其他 ResNet 编码器架构。

Table 1: Top-1 and top-5 accuracies (in %) under linear evaluation on ImageNet.

表 1: 在 ImageNet 上线性评估的 Top-1 和 Top-5 准确率 (以% 表示)。

with respect to $(\theta, \xi)$ (e.g., a collapsed constant representation). However BYOL' s target parameters $\xi$ updates are not in the direction of $\nabla_\xi \mathcal{L}^{\text{BYOL}}_{\theta,\xi}$ . More generally, we hypothesize that there is no loss $L_{\theta,\xi}$ such that BYOL' s dynamics is a gradient descent on $L$ jointly over $\theta, \xi$ . This is similar to GANs [66], where there is no loss that is jointly minimized w.r.t. both the discriminator and generator parameters. There is therefore no a priori reason why BYOL' s parameters would converge to a minimum of $\mathcal{L}^{\text{BVOL}}_{\theta,\xi}$ . While BYOL's dynamics still admit undesirable equilibria, we did not observe convergence to such equilibria in our experiments. In addition, when assuming BYOL's predictor to be optimal [6] i.e.,

关于 $(\theta, \xi)$ (例如，压缩常数表示)。然而，BYOL 的目标参数 $\xi$ 更新并不朝着 $\nabla_\xi \mathcal{L}^{\text{BYOL}}_{\theta,\xi}$ 的方向进行。更一般地，我们假设不存在损失 $L_{\theta,\xi}$ ，使得 BYOL 的动态是对 $L$ 的梯度下降，联合在 $\theta, \xi$ 上。这与 GANs [66] 类似，在 GANs 中，没有损失是相对于判别器和生成器参数共同最小化的。因此，没有先验理由说明 BYOL 的参数会收敛到 $\mathcal{L}^{\text{BVOL}}_{\theta,\xi}$ 的最小值。尽管 BYOL 的动态仍然允许不理想的平衡态，但我们在实验中并未观察到收敛到这种平衡态。此外，当假设 BYOL 的预测器是最优的 [6] 时，

$$q_\theta = q^\star \text{ with } q^\star \triangleq \arg\min_q \mathbb{E}\left[\left\|q\left(z_\theta\right) - z'_\xi\right\|_2^2\right] \text{ , where } q^\star\left(z_\theta\right) = \mathbb{E}\left[z'_\xi \mid z_\theta\right] \text{ , } \tag{3}$$

we hypothesize that the undesirable equilibria are unstable. Indeed, in this optimal predictor case, BYOL' s updates on $\theta$ follow in expectation the gradient of the expected conditional variance (see Appendix I for details), we note $z'_{\xi,i}$ the $i$ -th feature of $z'_\xi$ , then

我们假设不理想的平衡态是不稳定的。确实，在这个最优预测器的情况下，BYOL 在 $\theta$ 上的更新在期望上遵循期望条件方差的梯度 (详见附录 I)，我们注意到 $z'_{\xi,i}$ 是 $i$ -th 特征 $z'_\xi$ ，然后

$$\nabla_\theta \mathbb{E}\left[\left\|q^\star\left(z_\theta\right) - z'_\xi\right\|_2^2\right] = \nabla_\theta \mathbb{E}\left[\left\|\mathbb{E}\left[z'_\xi \mid z_\theta\right] - z'_\xi\right\|_2^2\right] = \nabla_\theta \mathbb{E}\left[\sum_i \operatorname{Var}\left(z'_{\xi,i} \mid z_\theta\right)\right], \tag{4}$$

Note that for any random variables $X, Y$ , and $Z, \operatorname{Var}\left(X \mid Y, Z\right) \leq \operatorname{Var}\left(X \mid Y\right)$ . Let $X$ be the target projection, $Y$ the current online projection, and $Z$ an additional variability on top of the online projection induced by stochasticities in the training dynamics: purely discarding information from the online projection cannot decrease the conditional variance.

注意，对于任何随机变量 $X, Y$ 和 $Z, \operatorname{Var}\left(X \mid Y, Z\right) \leq \operatorname{Var}\left(X \mid Y\right)$ 。设 $X$ 为目标投影，$Y$ 为当前在线投影，$Z$ 为在训练动态中由随机性引起的在线投影之上的额外变异性: 纯粹丢弃在线投影的信息无法降低条件方差。

In particular, BYOL avoids constant features in $z_\theta$ as, for any constant $c$ and random variables $z_\theta$ and $z'_\xi, \operatorname{Var}\left(z'_\xi \mid z_\theta\right) \leq \operatorname{Var}\left(z'_\xi \mid c\right)$ ; hence our hypothesis on these collapsed constant equilibria being unstable. Interestingly, if we were to minimize $\mathbb{E}\left[\sum_i \operatorname{Var}\left(z'_{\xi,i} \mid z_\theta\right)\right]$ with respect to $\xi$ , we would get a collapsed $z'_\xi$ as the variance is minimized for a constant $z'_\xi$ . Instead, BYOL makes $\xi$ closer to $\theta$ . incorporating sources of variability captured by the online projection into the target projection.

特别是，BYOL 避免了 $z_\theta$ 中的常量特征，因为对于任何常量 $c$ 和随机变量 $z_\theta$ 及 $z'_\xi, \operatorname{Var}\left(z'_\xi \mid z_\theta\right) \leq \operatorname{Var}\left(z'_\xi \mid c\right)$ ；因此我们对这些崩溃的常量平衡体的不稳定性的假设。有趣的是，如果我们要最小化 $\mathbb{E}\left[\sum_i \operatorname{Var}\left(z'_{\xi,i} \mid z_\theta\right)\right]$ 相对于 $\xi$ ，我们将得到一个崩溃的 $z'_\xi$ ，因为方差在常量 $z'_\xi$ 下被最小化。相反，BYOL 使得 $\xi$ 更接近于 $\theta$ ，将在线投影捕获的变异性源纳入目标投影中。

Furthermore, notice that performing a hard-copy of the online parameters $\theta$ into the target parameters $\xi$ would be enough to propagate new sources of variability. However, sudden changes in the target network might break the assumption of an optimal predictor, in which case BYOL's loss is not guaranteed to be close to the conditional variance. We hypothesize that the main role of BYOL's moving-averaged target network is to ensure the near-optimality of the predictor over training; Section 5 and Appendix J provide some empirical support of this interpretation.

此外，请注意，将在线参数 $\theta$ 硬拷贝到目标参数 $\xi$ 中将足以传播新的变异性源。然而，目标网络的突然变化可能会破坏最佳预测器的假设，在这种情况下，BYOL 的损失不保证接近条件方差。我们假设 BYOL 的移动平均目标网络的主要作用是确保预测器在训练过程中的近似最优性；第 5 节和附录 J 提供了对这一解释的一些实证支持。

| Method | Top-1 | | Top-5 | |
|---|---|---|---|---|
| | 1% | 10% | 1% | 10% |
| Supervised [77] | 25.4 | 56.4 | 48.4 | 80.4 |
| InstDisc | - | - | 39.2 | 77.4 |
| PIRL [35] | - | - | 57.2 | 83.8 |
| SimCLR [8] | 48.3 | 65.6 | 75.5 | 87.8 |
| BYOL (ours) | 53.2 | 68.8 | 78.4 | 89.0 |

| 方法 | Top-1 | | 前五名 | |
|---|---|---|---|---|
| | 1% | 10% | 1% | 10% |
| 监督学习 [77] | 25.4 | 56.4 | 48.4 | 80.4 |
| 实例对比 | - | - | 39.2 | 77.4 |
| PIRL [35] | - | - | 57.2 | 83.8 |
| SimCLR [8] | 48.3 | 65.6 | 75.5 | 87.8 |
| BYOL(我们的) | 53.2 | 68.8 | 78.4 | 89.0 |

(a) ResNet-50 encoder.
(a) ResNet-50 编码器。

| Method | Architecture | Param. | Top-1 | | Top-5 | |
|---|---|---|---|---|---|---|
| | | | 1% | 10% | 1% | 10% |
| CPC v2 [32] | ResNet-161 | 305M | - | - | 77.9 | 91.2 |
| SimCLR [8] | ResNet-50 (2×) | 94M | 58.5 | 71.7 | 83.0 | 91.2 |
| BYOL (ours) | ResNet-50 (2×) | 94M | 62.2 | 73.5 | 84.1 | 91.7 |
| SimCLR [8] | ResNet-50 (4×) | 375M | 63.0 | 74.4 | 85.8 | 92.6 |
| BYOL (ours) | ResNet-50 (4×) | 375M | 69.1 | 75.7 | 87.9 | 92.5 |
| BYOL (ours) | ResNet-200 (2×) | 250M | 71.2 | 77.7 | 89.5 | 93.7 |

| 方法 | 架构 | 参数 | 前 1 名 | | 前五名 | |
|---|---|---|---|---|---|---|
| | | | 1% | 10% | 1% | 10% |
| CPC v2 [32] | ResNet-161 | 305M | - | - | 77.9 | 91.2 |
| SimCLR [8] | ResNet-50 (2×) | 94M | 58.5 | 71.7 | 83.0 | 91.2 |
| BYOL (我们的) | ResNet-50 (2×) | 94M | 62.2 | 73.5 | 84.1 | 91.7 |
| SimCLR [8] | ResNet-50 (4×) | 375M | 63.0 | 74.4 | 85.8 | 92.6 |
| BYOL (我们的) | ResNet-50 (4×) | 375M | 69.1 | 75.7 | 87.9 | 92.5 |
| BYOL (我们的) | ResNet-200 (2×) | 250M | 71.2 | 77.7 | 89.5 | 93.7 |

(b) Other ResNet encoder architectures.
(b) 其他 ResNet 编码器架构。
Table 2: Semi-supervised training with a fraction of ImageNet labels.
表 2: 使用部分 ImageNet 标签的半监督训练。

| Method | Food101 | CIFAR10 | CIFAR100 | Birdsnap | SUN397 | Cars | Aircraft | VOC2007 | DTD | Pets | Caltech-101 | Flowers |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Linear evaluation: | | | | | | | | | | | | |
| BYOL (ours) | 75.3 | 91.3 | 78.4 | 57.2 | 62.2 | 67.8 | 60.6 | 82.5 | 75.5 | 90.4 | 94.2 | 96.1 |
| SimCLR (repro) | 72.8 | 90.5 | 74.4 | 42.4 | 60.6 | 49.3 | 49.8 | 81.4 | 75.7 | 84.6 | 89.3 | 92.6 |
| SimCLR [8] | 68.4 | 90.6 | 71.6 | 37.4 | 58.8 | 50.3 | 50.3 | 80.5 | 74.5 | 83.6 | 90.3 | 91.2 |
| Supervised-IN [8] | 72.3 | 93.6 | 78.3 | 53.7 | 61.9 | 66.7 | 61.0 | 82.8 | 74.9 | 91.5 | 94.5 | 94.7 |
| Fine-tuned: | | | | | | | | | | | | |
| BYOL (ours) | 88.5 | 97.8 | 86.1 | 76.3 | 63.7 | 91.6 | 88.1 | 85.4 | 76.2 | 91.7 | 93.8 | 97.0 |
| SimCLR (repro) | 87.5 | 97.4 | 85.3 | 75.0 | 63.9 | 91.4 | 87.6 | 84.5 | 75.4 | 89.4 | 91.7 | 96.6 |
| SimCLR [8] | 88.2 | 97.7 | 85.9 | 75.9 | 63.5 | 91.3 | 88.1 | 84.1 | 73.2 | 89.2 | 92.1 | 97.0 |
| Supervised-IN [8] | 88.3 | 97.5 | 86.4 | 75.8 | 64.3 | 92.1 | 86.0 | 85.0 | 74.6 | 92.1 | 93.3 | 97.6 |
| Random init [8] | 86.9 | 95.9 | 80.2 | 76.1 | 53.6 | 91.4 | 85.9 | 67.3 | 64.8 | 81.5 | 72.6 | 92.0 |

[6] For simplicity we also consider BYOL without normalization (which performs reasonably close to BYOL, see Appendix G.6) nor symmetrization.
[6] 为了简化，我们还考虑不进行归一化的 BYOL(其表现与 BYOL 相当接近，见附录 G.6) 或对称化。

| 方法 | Food101 | CIFAR10 | CIFAR100 | Birdsnap | SUN397 | 汽车 | 飞机 | VOC2007 | DTD | 宠物 | Caltech-101 | 花卉 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 线性评估: | | | | | | | | | | | | |
| BYOL(我们的方法) | 75.3 | 91.3 | 78.4 | 57.2 | 62.2 | 67.8 | 60.6 | 82.5 | 75.5 | 90.4 | 94.2 | 96.1 |
| SimCLR (重现) | 72.8 | 90.5 | 74.4 | 42.4 | 60.6 | 49.3 | 49.8 | 81.4 | 75.7 | 84.6 | 89.3 | 92.6 |
| SimCLR [8] | 68.4 | 90.6 | 71.6 | 37.4 | 58.8 | 50.3 | 50.3 | 80.5 | 74.5 | 83.6 | 90.3 | 91.2 |
| 监督式-IN [8] | 72.3 | 93.6 | 78.3 | 53.7 | 61.9 | 66.7 | 61.0 | 82.8 | 74.9 | 91.5 | 94.5 | 94.7 |
| 微调: | | | | | | | | | | | | |
| BYOL(我们的方法) | 88.5 | 97.8 | 86.1 | 76.3 | 63.7 | 91.6 | 88.1 | 85.4 | 76.2 | 91.7 | 93.8 | 97.0 |
| SimCLR (重现) | 87.5 | 97.4 | 85.3 | 75.0 | 63.9 | 91.4 | 87.6 | 84.5 | 75.4 | 89.4 | 91.7 | 96.6 |
| SimCLR [8] | 88.2 | 97.7 | 85.9 | 75.9 | 63.5 | 91.3 | 88.1 | 84.1 | 73.2 | 89.2 | 92.1 | 97.0 |
| 监督式-IN [8] | 88.3 | 97.5 | 86.4 | 75.8 | 64.3 | 92.1 | 86.0 | 85.0 | 74.6 | 92.1 | 93.3 | 97.6 |
| 随机初始化 [8] | 86.9 | 95.9 | 80.2 | 76.1 | 53.6 | 91.4 | 85.9 | 67.3 | 64.8 | 81.5 | 72.6 | 92.0 |

Table 3: Transfer learning results from ImageNet (IN) with the standard ResNet-50 architecture.
表 3: 使用标准 ResNet-50 架构从 ImageNet (IN) 的迁移学习结果。

# 4 Experimental evaluation

# 4 实验评估

We assess the performance of BYOL's representation after self-supervised pretraining on the training set of the ImageNet ILSVRC-2012 dataset [21]. We first evaluate it on ImageNet (IN) in both linear evaluation and semi-supervised setups. We then measure its transfer capabilities on other datasets and tasks, including classification, segmentation, object detection and depth estimation. For comparison, we also report scores for a representation trained using labels from the train ImageNet subset, referred to as Supervised-IN. In Appendix F, we assess the generality of BYOL by pretraining a representation on the Places365-Standard dataset [73] before reproducing this evaluation protocol.

我们评估了 BYOL 在 ImageNet ILSVRC-2012 数据集训练集上自监督预训练后的表示性能 [21]。我们首先在 ImageNet (IN) 上进行线性评估和半监督设置的评估。然后，我们测量其在其他数据集和任务上的迁移能力，包括分类、分割、目标检测和深度估计。为了进行比较，我们还报告了使用来自训练 ImageNet 子集的标签训练的表示的得分，称为监督-IN。在附录 F 中，我们通过在 Places365-Standard 数据集 [73] 上预训练一个表示来评估 BYOL 的通用性，然后再重现该评估协议。

Linear evaluation on ImageNet We first evaluate BYOL's representation by training a linear classifier on top of the frozen representation, following the procedure described in [48, 74, 41, 10, 8], and appendix D. 1; we report top-1 and top-5 accuracies in % on the test set in Table 1. With a standard ResNet-50 ( $\times 1$ ) BYOL obtains 74.3% top-1 accuracy (91.6% top-5 accuracy), which is a 1.3% (resp. 0.5% ) improvement over the previous self-supervised state of the art [12]. This tightens the gap with respect to the supervised baseline of [8], 76.5% , but is still significantly below the stronger supervised baseline of [75], 78.9%. With deeper and wider architectures, BYOL consistently outperforms the previous state of the art (Appendix D.2), and obtains a best performance of 79.6% top-1 accuracy, ranking higher than previous self-supervised approaches. On a ResNet-50 ($4\times$) BYOL achieves 78.6% , similar to the 78.9% of the best supervised baseline in [8] for the same architecture.

在 ImageNet 上的线性评估我们首先通过在冻结的表示上训练线性分类器来评估 BYOL 的表示，遵循 [48, 74, 41, 10, 8] 和附录 D.1 中描述的程序；我们在表 1 中报告测试集的 top-1 和 top-5 准确率 (以% 表示)。使用标准的 ResNet-50 ( $\times 1$ )，BYOL 获得了 74.3% 的 top-1 准确率 (91.6% 的 top-5 准确率)，这比之前的自监督最优结果 [12] 有了 1.3% (分别为 0.5% ) 的提升。这缩小了与 [8] 的监督基线 76.5% 之间的差距，但仍显著低于更强的监督基线 [75] 的 78.9%。在更深和更宽的架构下，BYOL 始终优于之前的最优结果 (附录 D.2)，并获得了 79.6% 的 top-1 准确率，排名高于之前的自监督方法。在 ResNet-50 ($4\times$) 上，BYOL 达到了 78.6% ，与 [8] 中相同架构的最佳监督基线的 78.9% 相似。

Semi-supervised training on ImageNet Next, we evaluate the performance obtained when fine-tuning BYOL's representation on a classification task with a small subset of ImageNet's train set, this time using label information. We follow the semi-supervised protocol of [74, 76, 8, 32] detailed in Appendix D. 1, and use the same fixed splits of respectively 1% and 10% of ImageNet labeled training data as in [8]. We report both top-1 and top-5 accuracies on the test set in Table 2. BYOL consistently outperforms previous approaches across a wide range of architectures. Additionally, as detailed in Appendix D.1, BYOL reaches 77.7% top-1 accuracy with ResNet-50 when fine-tuning over 100% of ImageNet labels.

半监督训练在 ImageNet 上接下来，我们评估在使用标签信息的小部分 ImageNet 训练集上微调 BYOL 表示时获得的性能。我们遵循附录 D.1 中详细说明的 [74, 76, 8, 32] 的半监督协议，并使用与 [8] 中相同的固定分割，即 ImageNet 标记训练数据的 1% 和 10% 。我们在表 2 中报告测试集的 top-1 和 top-5 准确率。BYOL 在各种架构中始终优于以前的方法。此外，如附录 D.1 中详细说明，BYOL 在对 ImageNet 标签的 100% 进行微调时，使用 ResNet-50 达到 77.7% 的 top-1 准确率。

Transfer to other classification tasks We evaluate our representation on other classification datasets to assess whether the features learned on ImageNet (IN) are generic and thus useful across image domains, or if they are ImageNet-specific. We perform linear evaluation and fine-tuning on the same set of classification tasks used in [8, 74] , and carefully follow their evaluation protocol, as detailed in Appendix E. Performance is reported using standard metrics for each benchmark, and results are provided on a held-out test set after hyperparameter selection on a validation set. We report results in Table 3, both for linear evaluation and fine-tuning. BYOL outperforms SimCLR on all benchmarks and the Supervised-IN baseline on 7 of the 12 benchmarks, providing only slightly worse performance on the 5 remaining benchmarks. BYOL's representation can transfer to small images, e.g., CIFAR [78], landscapes, e.g., SUN397 [79] or VOC2007 [80], and textures, e.g., DTD [81].

转移到其他分类任务我们在其他分类数据集上评估我们的表示，以评估在 ImageNet (IN) 上学习的特征是否是通用的，因此在图像领域中有用，或者它们是否是特定于 ImageNet 的。我们在与 [8, 74] 中使用的相同分类任务集上进行线性评估和微调，并仔细遵循他们的评估协议，如附录 E 中详细说明。性能使用每个基准的标准指标报告，并在验证集上进行超参数选择后，在保留的测试集上提供结果。我们在表 3 中报告线性评估和微调的结果。BYOL 在所有基准上都优于 SimCLR，并在 12 个基准中的 7 个上优于监督的 IN 基线，在其余 5 个基准上仅提供略差的性能。BYOL 的表示可以转移到小图像，例如 CIFAR [78]，风景，例如 SUN397 [79] 或 VOC2007 [80]，以及纹理，例如 DTD [81]。

Transfer to other vision tasks We evaluate our representation on different tasks relevant to computer vision practitioners, namely semantic segmentation, object detection and depth estimation. With this evaluation, we assess whether BYOL's representation generalizes beyond classification tasks.

转移到其他视觉任务我们在与计算机视觉从业者相关的不同任务上评估我们的表示，即语义分割、目标检测和深度估计。通过这一评估，我们判断 BYOL 的表示是否超越分类任务具有泛化能力。

We first evaluate BYOL on the VOC2012 semantic segmentation task as detailed in Appendix E.4, where the goal is to classify each pixel in the image [7]. We report the results in Table 4a. BYOL outperforms both the Supervised-IN baseline (+1.9 mIoU) and SimCLR (+1.1 mIoU).

我们首先在 VOC2012 语义分割任务上评估 BYOL，具体细节见附录 E.4，目标是对图像中的每个像素进行分类 [7]。我们在表 4a 中报告结果。BYOL 的表现优于 Supervised-IN 基线 (+1.9 mIoU) 和 SimCLR (+1.1 mIoU)。

Similarly, we evaluate on object detection by reproducing the setup in [9] using a Faster R-CNN architecture [82], as detailed in Appendix E.5. We fine-tune on trainval2007 and report results on test2007 using the standard $AP_{50}$ metric; BYOL is significantly better than the Supervised-IN baseline ($+3.1AP_{50}$) and SimCLR ($+2.3AP_{50}$) .

同样，我们通过使用 Faster R-CNN 架构 [82] 重现 [9] 中的设置来评估目标检测，具体细节见附录 E.5。我们在 trainval2007 上进行微调，并使用标准 $AP_{50}$ 指标在 test2007 上报告结果；BYOL 的表现显著优于 Supervised-IN 基线 ($+3.1AP_{50}$) 和 SimCLR ($+2.3AP_{50}$) 。

Finally, we evaluate on depth estimation on the NYU v2 dataset, where the depth map of a scene is estimated given a single RGB image. Depth prediction measures how well a network represents geometry, and how well that information can be localized to pixel accuracy [40]. The setup is based on [83] and detailed in Appendix E. 6. We evaluate on the commonly used test subset of 654 images and report results using several common metrics in Table 4b: relative (rel) error, root mean squared (rms) error, and the percent of pixels (pct) where the error, $\max(d_{gt}/d_p, d_p/d_{gt})$ , is below $1.25^n$ thresholds where $d_p$ is the predicted depth and $d_{gt}$ is the ground truth depth [40]. BYOL is better or on par with other methods for each metric. For instance, the challenging pct.<1.25 measure is respectively improved by +3.5 points and +1.3 points compared to supervised and SimCLR baselines.

最后，我们在 NYU v2 数据集上评估深度估计，其中给定单个 RGB 图像来估计场景的深度图。深度预测衡量网络对几何形状的表示能力，以及该信息在像素精度上的本地化能力 [40]。该设置基于 [83]，并在附录 E 中详细说明。我们在常用的 654 张图像的测试子集上进行评估，并在表 4b 中报告使用几种常见指标的结果: 相对 (rel) 误差、均方根 (rms) 误差，以及误差 $\max(d_{gt}/d_p, d_p/d_{gt})$ 低于 $1.25^n$ 阈值的像素百分比 (pct)，其中 $d_p$ 是预测的深度，$d_{gt}$ 是真实深度 [40]。在每个指标上，BYOL 的表现优于或与其他方法持平。例如，具有挑战性的 pct.<1.25 测量相较于监督和 SimCLR 基线分别提高了 +3.5 分和 +1.3 分。

| Method | $AP_{50}$ | mIoU |
|---|---|---|
| Supervised-IN [9] | 74.4 | 74.4 |
| MoCo [9] | 74.9 | 72.5 |
| SimCLR (repro) | 75.2 | 75.2 |
| BYOL (ours) | 77.5 | 76.3 |

| 方法 | $AP_{50}$ | mIoU |
|---|---|---|
| 监督式-IN [9] | 74.4 | 74.4 |
| MoCo [9] | 74.9 | 72.5 |
| SimCLR (重现) | 75.2 | 75.2 |
| BYOL (我们的) | 77.5 | 76.3 |

| Method | pct.< 1.25 | Higher better | | Lower better | |
|---|---|---|---|---|---|
| | | pct. $< 1.25^2$ | pct.$< 1.25^3$ | rms | rel |
| Supervised-IN [83] | 81.1 | 95.3 | 98.8 | 0.573 | 0.127 |
| SimCLR (repro) | 83.3 | 96.5 | 99.1 | 0.557 | 0.134 |
| BYOL (ours) | 84.6 | 96.7 | 99.1 | 0.541 | 0.129 |

| 方法 | pct.< 1.25 | 越高越好 | | 越低越好 | |
|---|---|---|---|---|---|
| | | pct. $< 1.25^2$ | pct.$< 1.25^3$ | rms | rel |
| 监督学习-IN [83] | 81.1 | 95.3 | 98.8 | 0.573 | 0.127 |
| SimCLR (重现) | 83.3 | 96.5 | 99.1 | 0.557 | 0.134 |
| BYOL (我们的) | 84.6 | 96.7 | 99.1 | 0.541 | 0.129 |

(a) Transfer results in semantic segmentation and object detection.

(b) Transfer results on NYU v2 depth estimation.

(a) 语义分割和物体检测的迁移结果 (b) NYU v2 深度估计的迁移结果。

Table 4: Results on transferring BYOL's representation to other vision tasks.

表 4: 将 BYOL 的表示迁移到其他视觉任务的结果。

# 5 Building intuitions with ablations

# 5 通过消融实验建立直觉

We present ablations on BYOL to give an intuition of its behavior and performance. For reproducibility, we run each configuration of parameters over three seeds, and report the average performance. We also report the half difference between the best and worst runs when it is larger than 0.25 . Although previous works perform ablations at 100 epochs [8, 12], we notice that relative improvements at 100 epochs do not always hold over longer training. For this reason, we run ablations over 300 epochs on 64 TPU v3 cores, which yields consistent results compared to our baseline training of 1000 epochs. For all the experiments in this section, we set the initial learning rate to 0.3 with batch size 4096, the weight decay to $10^{-6}$ as in SimCLR [8] and the base target decay rate $\tau_{\text{base}}$ to 0.99 . In this section we report results in top-1 accuracy on ImageNet under the linear evaluation protocol as in Appendix D.1.

我们对 BYOL 进行了消融实验，以直观地展示其行为和性能。为了确保可重复性，我们在三个随机种子上运行每种参数配置，并报告平均性能。当最佳和最差运行之间的差异大于 0.25 时，我们还报告它们之间的半差异。尽管之前的工作在 100 个周期内进行了消融实验 [8, 12]，但我们注意到在 100 个周期内的相对改进并不总是适用于更长的训练。因此，我们在 64 个 TPU v3 核心上进行了 300 个周期的消融实验，与我们 1000 个周期的基线训练相比，得到了一致的结果。在本节的所有实验中，我们将初始学习率设置为 0.3，批量大小为 4096，权重衰减设置为 $10^{-6}$ ，与 SimCLR [8] 一致，基础目标衰减率 $\tau_{\text{base}}$ 设置为 0.99。在本节中，我们报告在 ImageNet 上线性评估协议下的 top-1 准确率结果，如附录 D.1 所示。

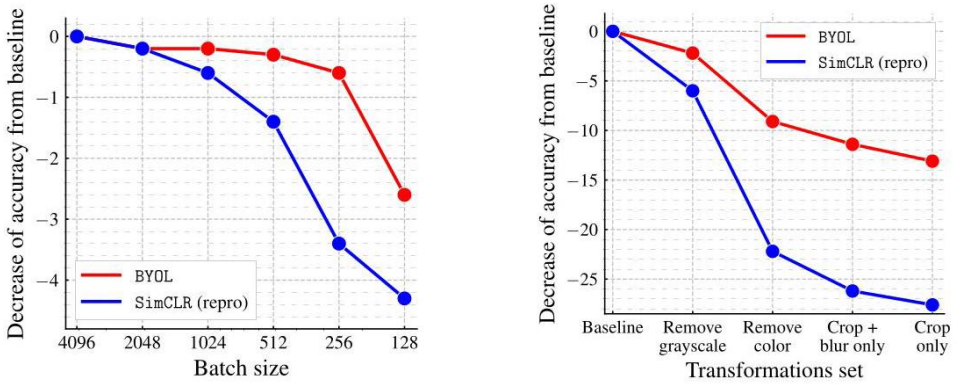
Batch size Among contrastive methods, the ones that draw negative examples from the minibatch suffer performance drops when their batch size is reduced. BYOL does not use negative examples and we expect it to be more robust to smaller batch sizes. To empirically verify this hypothesis, we train both BYOL and SimCLR using different batch sizes from 128 to 4096. To avoid re-tuning other hyperparameters, we average gradients over $N$ consecutive steps before updating the online network when reducing the batch size by a factor $N$ . The target network is updated once every $N$ steps, after the update of the online network; we accumulate the $N$ -steps in parallel in our runs. As shown in Figure 3a, the performance of SimCLR rapidly deteriorates with batch size, likely due to the decrease in the number of negative examples. In contrast, the performance of BYOL remains stable over a wide range of batch sizes from 256 to 4096, and only drops for smaller values due to batch normalization layers in the encoder. [7]

批量大小在对比方法中，从小批量中提取负样本的方法在减少批量大小时会遭遇性能下降。BYOL 不使用负样本，我们预计它对较小批量大小会更具鲁棒性。为了实证验证这一假设，我们使用从 128 到 4096 的不同批量大小训练 BYOL 和 SimCLR。为了避免重新调整其他超参数，我们在减少批量大小时，在更新在线网络之前对 $N$ 个连续步骤的梯度进行平均。目标网络在在线网络更新后每 $N$ 步骤更新一次；

我们在运行中并行累积 $N$ 步骤。如图 3a 所示，SimCLR 的性能随着批量大小的减少而迅速恶化，这可能是由于负样本数量的减少。相比之下，BYOL 的性能在 256 到 4096 的广泛批量大小范围内保持稳定，只有在较小值时由于编码器中的批量归一化层而下降。[7]

Image augmentations Contrastive methods are sensitive to the choice of image augmentations. For instance, SimCLR does not work well when removing color distortion from its image augmentations. As an explanation, SimCLR shows that crops of the same image mostly share their color histograms. At the same time, color histograms vary across images. Therefore, when a contrastive task only relies on random crops as image augmentations, it can be mostly solved by focusing on color histograms alone. As a result the representation is not incentivized to retain information beyond color histograms. To prevent that, SimCLR adds color distortion to its set of image augmentations. Instead, BYOL is incentivized to keep any information captured by the target representation into its online network, to

图像增强对比方法对图像增强的选择非常敏感。例如，当从其图像增强中去除颜色失真时，SimCLR 的效果不佳。作为解释，SimCLR 显示同一图像的裁剪大多共享其颜色直方图。同时，颜色直方图在不同图像之间是变化的。因此，当对比任务仅依赖随机裁剪作为图像增强时，它可以主要通过关注颜色直方图来解决。因此，表示不被激励保留超出颜色直方图的信息。为了防止这种情况，SimCLR 在其图像增强集中添加了颜色失真。相反，BYOL 被激励将目标表示捕获的任何信息保留到其在线网络中，以




(a) Impact of batch size

(b) Impact of progressively removing transformatic

Figure 3: Decrease in top-1 accuracy (in % points) of BYOL and our own reproduction of SimCLR at 300 epochs, under linear evaluation on ImageNet.

图 3: 在 ImageNet 上进行线性评估时，BYOL 和我们自己复现的 SimCLR 在 300 个周期下的 top-1 准确率 (以百分比点表示) 下降。

improve its predictions. Therefore, even if augmented views of a same image share the same color histogram, BYOL is still incentivized to retain additional features in its representation. For that reason, we believe that BYOL is more robust to the choice of image augmentations than contrastive methods.

改进其预测。因此，即使同一图像的增强视图共享相同的颜色直方图，BYOL 仍然被激励在其表示中保留额外的特征。因此，我们认为 BYOL 对图像增强的选择比对比方法更具鲁棒性。

Results presented in Figure 3b support this hypothesis: the performance of BYOL is much less affected than the performance of SimCLR when removing color distortions from the set of image augmentations (-9.1 accuracy points for BYOL, -22.2 accuracy points for SimCLR). When image augmentations are reduced to mere random crops, BYOL still displays good performance (59.4%, i.e. -13.1 points from 72.5% ), while SimCLR loses more than a third of its performance (40.3% , i.e. -27.6 points from 67.9%). We report additional ablations in Appendix G.3.

图 3b 中呈现的结果支持这一假设: 在从图像增强集合中去除颜色失真时，BYOL 的性能受到的影响远小于 SimCLR 的性能 (BYOL 减少 9.1 个准确率点，SimCLR 减少 22.2 个准确率点)。当图像增强减少到仅仅是随机裁剪时，BYOL 仍然表现良好 (59.4%，即比 72.5% 减少 13.1 个点)，而 SimCLR 的性能损失超过三分之一 (40.3% ，即从 67.9% 减少 27.6 个点)。我们在附录 G.3 中报告了更多的消融实验。

Bootstrapping BYOL uses the projected representation of a target network, whose weights are an exponential moving average of the weights of the online network, as target for its predictions. This way, the weights of the target network represent a delayed and more stable version of the weights of the

[7] The only dependency on batch size in our training pipeline sits within the batch normalization layers.

[7] 我们训练管道中对批量大小的唯一依赖存在于批量归一化层中。

online network. When the target decay rate is 1, the target network is never updated, and remains at a constant value corresponding to its initialization. When the target decay rate is 0, the target network is instantaneously updated to the online network at each step. There is a trade-off between updating the targets too often and updating them too slowly, as illustrated in Table 5a. Instantaneously updating the target network ($\tau = 0$) destabilizes training, yielding very poor performance while never updating the target ($\tau = 1$) makes the training stable but prevents iterative improvement, ending with low-quality final representation. All values of the decay rate between 0.9 and 0.999 yield performance above 68.4% top-1 accuracy at 300 epochs.

Bootstrapping BYOL 使用目标网络的投影表示，该网络的权重是在线网络权重的指数移动平均，作为其预测的目标。通过这种方式，目标网络的权重代表了在线网络权重的延迟和更稳定的版本。当目标衰减率为 1 时，目标网络从不更新，并保持在与其初始化相对应的常数值。当目标衰减率为 0 时，目标网络在每一步都瞬时更新为在线网络。更新目标过于频繁与更新过于缓慢之间存在权衡，如表 5a 所示。瞬时更新目标网络 ($\tau = 0$) 会使训练不稳定，导致性能非常差，而从不更新目标 ($\tau = 1$) 则使训练稳定，但阻止了迭代改进，最终导致低质量的最终表示。在 0.9 到 0.999 之间的所有衰减率值在 300 个周期内的性能均超过 68.4% 的 top-1 准确率。

| Target | $\tau_{\text{base}}$ | Top-1 |
|---|---|---|
| Constant random network | 1 | 18.8±0.7 |
| Moving average of online | 0.999 | 69.8 |
| Moving average of online | 0.99 | 72.5 |
| Moving average of online | 0.9 | 68.4 |
| Stop gradient of online | 0 | 0.3 |

| 目标 | $\tau_{\text{base}}$ | Top-1 |
|---|---|---|
| 恒定随机网络 | 1 | 18.8±0.7 |
| 在线移动平均 | 0.999 | 69.8 |
| 在线移动平均 | 0.99 | 72.5 |
| 在线移动平均 | 0.9 | 68.4 |
| 在线停止梯度 | 0 | 0.3 |

(a) Results for different target modes. † In the stop gradient of online, $\tau = \tau_{\text{base}} = 0$ is kept constant throughout training.

(a) 不同目标模式的结果。† 在在线的停止梯度中，$\tau = \tau_{\text{base}} = 0$ 在整个训练过程中保持不变。

| Method | Predictor | Target network | $\beta$ | Top-1 |
|---|---|---|---|---|
| BYOL | ✓ | ✓ | 0 | 72.5 |
| - | ✓ | ✓ | 1 | 70.9 |
| - | | ✓ | 1 | 70.7 |
| SimCLR | | | 1 | 69.4 |
| - | ✓ | | 1 | 69.1 |
| - | ✓ | | 0 | 0.3 |
| - | | ✓ | 0 | 0.2 |
| - | | | 0 | 0.1 |

| 方法 | 预测器 | 目标网络 | $\beta$ | Top-1 |
|---|---|---|---|---|
| BYOL | ✓ | ✓ | 0 | 72.5 |
| - | ✓ | ✓ | 1 | 70.9 |
| - | | ✓ | 1 | 70.7 |
| SimCLR | | | 1 | 69.4 |
| - | ✓ | | 1 | 69.1 |
| - | ✓ | | 0 | 0.3 |
| - | | ✓ | 0 | 0.2 |
| - | | | 0 | 0.1 |

(b) Intermediate variants between BYOL and SimCLR.

(b) BYOL 和 SimCLR 之间的中间变体。

Table 5: Ablations with top-1 accuracy (in %) at 300 epochs under linear evaluation on ImageNet.

表 5: 在 ImageNet 上线性评估下，300 个周期的 top-1 准确率 (以% 计) 的消融实验。

Ablation to contrastive methods In this subsection, we recast SimCLR and BYOL using the same formalism to better understand where the improvement of BYOL over SimCLR comes from. Let us consider the following objective that extends the InfoNCE objective [10, 84] (see Appendix G.4),

对比方法的消融实验在本小节中，我们使用相同的形式重新阐述 SimCLR 和 BYOL，以更好地理解 BYOL 相对于 SimCLR 的改进来源。我们考虑以下目标，该目标扩展了 InfoNCE 目标 [10, 84](见附录 G.4)，

$$\text{InfoNCE}_\theta^{\alpha,\beta} \triangleq \frac{2}{B}\sum_{i=1}^{B} S_\theta\left(v_i, v_i'\right) - \beta \cdot \frac{2\alpha}{B}\sum_{i=1}^{B} \ln\left(\sum_{j\neq i} \exp\frac{S_\theta\left(v_i, v_j\right)}{\alpha} + \sum_j \exp\frac{S_\theta\left(v_i, v_j'\right)}{\alpha}\right),$$

where $\alpha > 0$ is a fixed temperature, $\beta \in [0,1]$ a weighting coefficient, $B$ the batch size, $v$ and $v'$ are batches of augmented views where for any batch index $i, v_i$ and $v_i'$ are augmented views from the same image; the real-valued function $S_\theta$ quantifies pairwise similarity between augmented views. For any augmented view $u$ we denote $z_\theta(u) \triangleq f_\theta(g_\theta(u))$ and $z_\xi(u) \triangleq f_\xi(g_\xi(u))$ . For given $\phi$ and $\psi$ , we consider the normalized dot product

其中 $\alpha > 0$ 是固定温度，$\beta \in [0,1]$ 是加权系数，$B$ 是批量大小，$v$ 和 $v'$ 是增强视图的批次，对于任何批次索引 $i, v_i$ 和 $v_i'$ ，它们是来自同一图像的增强视图；实值函数 $S_\theta$ 量化增强视图之间的成对相似性。对于任何增强视图 $u$ ，我们表示为 $z_\theta(u) \triangleq f_\theta(g_\theta(u))$ 和 $z_\xi(u) \triangleq f_\xi(g_\xi(u))$ 。对于给定的 $\phi$ 和 $\psi$ ，我们考虑归一化的点积。

$$S_\theta(u_1, u_2) \triangleq \frac{\langle \phi(u_1), \psi(u_2)\rangle}{\| \phi(u_1) \|_2 \cdot \| \psi(u_2) \|_2}.$$

Up to minor details (cf. Appendix G.5), we recover the SimCLR loss with $\phi(u_1) = z_\theta(u_1)$ (no predictor), $\psi(u_2) = z_\theta(u_2)$ (no target network) and $\beta = 1$ . We recover the BYOL loss when using a predictor and a target network, i.e., $\phi(u_1) = p_\theta(z_\theta(u_1))$ and $\psi(u_2) = z_\xi(u_2)$ with $\beta = 0$ . To evaluate the influence of the target network, the predictor and the coefficient $\beta$ , we perform an ablation over them. Results are presented in Table 5 b and more details are given in Appendix G.4. The only variant that performs well without negative examples (i.e., with $\beta = 0$ ) is BYOL, using both a bootstrap target network and a predictor. Adding the negative pairs to BYOL's loss without re-tuning the temperature parameter hurts its performance. In Appendix G.4, we show that we can add back negative pairs and still match the performance of BYOL with proper tuning of the temperature.

除了一些细节 (参见附录 G.5)，我们通过 $\phi(u_1) = z_\theta(u_1)$ (无预测器)、$\psi(u_2) = z_\theta(u_2)$ (无目标网络) 和 $\beta = 1$ 恢复了 SimCLR 损失。当使用预测器和目标网络时，我们恢复 BYOL 损失，即 $\phi(u_1) = p_\theta(z_\theta(u_1))$ 和 $\psi(u_2) = z_\xi(u_2)$ 与 $\beta = 0$ 。为了评估目标网络、预测器和系数 $\beta$ 的影响，我们对它们进行了消融实验。结果见表 5 b ，更多细节见附录 G.4。唯一在没有负例 (即，使用 $\beta = 0$ ) 的情况下表现良好的变体是 BYOL，它同时使用了自举目标网络和预测器。将负对添加到 BYOL 的损失中而不重新调整温度参数会损害其性能。在附录 G.4 中，我们展示了可以重新添加负对，并通过适当调整温度仍然匹配 BYOL 的性能。

Simply adding a target network to SimCLR already improves performance (+1.6 points). This sheds new light on the use of the target network in MoCo [9], where the target network is used to provide more negative examples. Here, we show that by mere stabilization effect, even when using the same number of negative examples, using a target network is beneficial. Finally, we observe that modifying the architecture of $S_\theta$ to include a predictor only mildly affects the performance of SimCLR.

仅仅在 SimCLR 中添加一个目标网络就已经提高了性能 (+1.6 分)。这为 MoCo [9] 中目标网络的使用提供了新的视角，在 MoCo 中，目标网络用于提供更多的负样本。在这里，我们展示了仅通过稳定化效应，即使使用相同数量的负样本，使用目标网络也是有益的。最后，我们观察到，修改 $S_\theta$ 的架构以包含一个预测器仅对 SimCLR 的性能产生轻微影响。

Relationship with Mean Teacher Another semi-supervised approach, Mean Teacher (MT) [20], complements a supervised loss on few labels with an additional consistency loss. In [20], this consistency loss is the $\ell_2$ distance between the logits from a student network, and those of a temporally averaged version of the student network, called teacher. Removing the predictor in BYOL results in an unsupervised version of MT with no classification loss that uses image augmentations instead of the original architectural noise (e.g., dropout). This variant of BYOL collapses (Row 7 of ??) which suggests that the additional predictor is critical to prevent collapse in an unsupervised scenario.

与均值教师的关系另一种半监督方法，均值教师 (Mean Teacher, MT)[20]，在少量标签上补充了一个监督损失与额外的一致性损失。在 [20] 中，这个一致性损失是来自学生网络的 logits 与一个时间平均版本的学生网络 (称为教师) 之间的 $\ell_2$ 距离。在 BYOL 中去除预测器会导致一个无监督版本的 MT，该版

本没有分类损失，而是使用图像增强而不是原始的架构噪声 (例如，dropout)。这种 BYOL 的变体会崩溃 (?? 的第 7 行)，这表明额外的预测器对于防止无监督场景中的崩溃至关重要。

Importance of a near-optimal predictor Table 5 b already shows the importance of combining a predictor and a target network: the representation does collapse when either is removed. We further found that we can remove the target network without collapse by making the predictor near-optimal, either by (i) using an optimal linear predictor (obtained by linear regression on the current batch) before back-propagating the error through the network (52.5% top-1 accuracy), or (ii) increasing the learning rate of the predictor (66.5% top-1 ) . By contrast, increasing the learning rates of both projector and predictor (without target network) yields poor results ( $\approx$ 25% top-1). See Appendix J for more details. This seems to indicate that keeping the predictor near-optimal at all times is important to preventing collapse, which may be one of the roles of BYOL's target network.

近似最优预测器的重要性表 5 b 已经显示了结合预测器和目标网络的重要性: 当其中任何一个被移除时，表示会崩溃。我们进一步发现，通过使预测器近似最优，可以在不崩溃的情况下移除目标网络，方法是 (i) 在通过网络反向传播误差之前，使用一个最优线性预测器 (通过对当前批次进行线性回归获得) (52.5% 的 top-1 准确率，或者 (ii) 增加预测器的学习率 (66.5% top-1 ) 。相比之下，增加投影器和预测器的学习率 (没有目标网络) 会导致较差的结果 ( $\approx$ 25% top-1)。有关更多细节，请参见附录 J。这似乎表明，始终保持预测器近似最优对于防止崩溃是重要的，这可能是 BYOL 的目标网络的一个作用。

# 6 Conclusion

# 6 结论

We introduced BYOL, a new algorithm for self-supervised learning of image representations. BYOL learns its representation by predicting previous versions of its outputs, without using negative pairs. We show that BYOL achieves state-of-the-art results on various benchmarks. In particular, under the linear evaluation protocol on ImageNet with a ResNet-50 (1×) , BYOL achieves a new state of the art and bridges most of the remaining gap between self-supervised methods and the supervised learning baseline of [8]. Using a ResNet-200 (2×) , BYOL reaches a top-1 accuracy of 79.6% which improves over the previous state of the art (76.8%) while using 30% fewer parameters.

我们介绍了 BYOL，一种用于自监督学习图像表示的新算法。BYOL 通过预测其输出的先前版本来学习其表示，而不使用负样本对。我们展示了 BYOL 在各种基准测试中达到了最先进的结果。特别是在 ImageNet 上使用 ResNet-50 的线性评估协议下 (1×) ，BYOL 达到了新的最先进水平，并弥补了自监督方法与监督学习基线之间的大部分差距 [8]。使用 ResNet-200 (2×) ，BYOL 达到的 top-1 准确率为 79.6% ，比之前的最先进水平 (76.8%) 有所提高，同时使用的参数更少 30% 。

Nevertheless, BYOL remains dependent on existing sets of augmentations that are specific to vision applications. To generalize BYOL to other modalities (e.g., audio, video, text, ...) it is necessary to obtain similarly suitable augmentations for each of them. Designing such augmentations may require significant effort and expertise. Therefore, automating the search for these augmentations would be an important next step to generalize BYOL to other modalities.

然而，BYOL 仍然依赖于特定于视觉应用的现有增强集。为了将 BYOL 泛化到其他模态 (例如，音频、视频、文本等)，有必要为每种模态获得同样合适的增强。设计这样的增强可能需要大量的努力和专业知识。因此，自动化搜索这些增强将是将 BYOL 泛化到其他模态的重要下一步。

# Broader impact

# 更广泛的影响

The presented research should be categorized as research in the field of unsupervised learning. This work may inspire new algorithms, theoretical, and experimental investigation. The algorithm presented here can be used for many different vision applications and a particular use may have both positive or negative impacts, which is known as the dual use problem. Besides, as vision datasets could be biased, the representation learned by BYOL could be susceptible to replicate these biases.

本研究应被归类为无监督学习领域的研究。这项工作可能会激发新的算法、理论和实验研究。这里提出的算法可以用于许多不同的视觉应用，特定的使用可能会产生正面或负面的影响，这被称为双重用途问题。此外，由于视觉数据集可能存在偏见，BYOL 学习到的表示可能会容易复制这些偏见。

## Acknowledgements


## 致谢

The authors would like to thank the following people for their help throughout the process of writing this paper, in alphabetical order: Aaron van den Oord, Andrew Brock, Jason Ramapuram, Jeffrey De Fauw, Karen Simonyan, Katrina McKinnay, Nathalie Beauguerlange, Olivier Henaff, Oriol Vinyals, Pauline Luc, Razvan Pascanu, Sander Dieleman, and the DeepMind team. We especially thank Jason Ramapuram and Jeffrey De Fauw, who provided the JAX SimCLR reproduction used throughout the paper.

作者感谢以下人员在撰写本文过程中提供的帮助，按字母顺序排列:Aaron van den Oord、Andrew Brock、Jason Ramapuram、Jeffrey De Fauw、Karen Simonyan、Katrina McKinnay、Nathalie Beauguerlange、Olivier Henaff、Oriol Vinyals、Pauline Luc、Razvan Pascanu、Sander Dieleman 和 DeepMind 团队。我们特别感谢 Jason Ramapuram 和 Jeffrey De Fauw，他们提供了本文中使用的 JAX SimCLR 重现。


## References

## 参考文献


[1] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4):193-202, 1980.

[2] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. Neural Computation, 14(4), 2002.

[3] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. Neural Computation, 18(7):1527-1554, 2006.

[4] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In Computer Vision and Pattern Recognition, 2014.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Computer Vision and Pattern Recognition, 2014.

[7] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Computer Vision and Pattern Recognition, 2015.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709, 2020.

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722, 2019.

[10] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

[11] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. arXiv preprint arXiv:1906.05849v4, 2019.

[12] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. arXiv preprint arXiv:2005.10243, 2020.

[13] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In International Conference on Machine Learning, 2019.

[14] R. Manmatha, Chao-Yuan Wu, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In International Conference on Computer Vision, 2017.

[15] Ben Harwood, Vijay B. G. Kumar, Gustavo Carneiro, Ian Reid, and Tom Drummond. Smart mining for deep metric learning. In International Conference on Computer Vision, 2017.

[16] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In International Conference on Machine Learning, 2013.

[17] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In European Conference on Computer Vision, 2018.

[18] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In Advances in neural information processing systems, pages 3365-3373, 2014.

[19] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242, 2016.

[20] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Advances in neural information processing systems, pages 1195-1204, 2017.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3):211-252, 2015.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Computer Vision and Pattern Recognition, 2016.

[23] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In Computer Vision and Pattern Recognition, 2015.

[24] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In International Conference on Machine Learning, 2008.

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[26] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic back-propagation and variational inference in deep latent gaussian models. arXiv preprint arXiv:1401.4082, 2014.

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Neural Information Processing Systems, 2014.

[28] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. arXiv preprint arXiv:1605.09782, 2016.

[29] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martín Arjovsky, Olivier Mastropietro, and Aaron C. Courville. Adversarially learned inference. arXiv preprint arXiv:1606.00704, 2017.

[30] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In Neural Information Processing Systems, 2019.

[31] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096, 2018.

[32] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord. Data-efficient image recognition with contrastive predictive coding. In International Conference on Machine Learning, 2019.

[33] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670, 2019.

[34] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In Neural Information Processing Systems, 2019.

[35] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. arXiv preprint arXiv:1912.01991, 2019.

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

[37] Rishabh Jain, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020.

[38] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029, 2020.

[39] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via nonparametric instance discrimination. In Computer Vision and Pattern Recognition, 2018.

[40] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In International Conference on Computer Vision, 2017.

[41] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In European Conference on Computer Vision, 2016.

[42] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In European Conference on Computer Vision, 2016.

[43] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In Computer Vision and Pattern Recognition, 2016.

[44] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In European Conference on Computer Vision, 2016.

[45] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In Computer Vision and Pattern Recognition, 2017.

[46] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In Neural Information Processing Systems, 2014.

[47] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728, 2018.

[48] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In Computer Vision and Pattern Recognition, 2019.

[49] Daniel Guo, Bernardo Avila Pires, Bilal Piot, Jean-Bastien Grill, Florent Altché, Rémi Munos, and Mohammad Gheshlaghi Azar. Bootstrap latent-predictive representations for multitask reinforcement learning. In International Conference on Machine Learning, 2020.

[50] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Joannis Antonoglou, Helen. King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. Nature, 518:529-533, 2015.

[51] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In International Conference on Machine Learning, 2016.

[52] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Gheshlaghi Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In AAAI Conference on Artificial Intelligence, 2018.

[53] Hado Van Hasselt, Yotam Doron, Florian Strub, Matteo Hessel, Nicolas Sonnerat, and Joseph Modayil. Deep reinforcement learning and the deadly triad. Deep Reinforcement Learning Workshop NeurIPS, 2018.

[54] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971, 2015.

[55] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. IEEE Transactions on Neural Networks, 20(3):542-542, 2009.

[56] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning, 3(1):1-130, 2009.

[57] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In Advances in neural information processing systems, 2014.

[58] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In Advances in neural information processing systems, 2015.

[59] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raf-fel. Mixmatch: A holistic approach to semi-supervised learning. In Advances in Neural Information Processing Systems, 2019.

[60] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence, 41(8):1979-1993, 2018.

[61] David Berthelot, N. Carlini, E. D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In ICLR, 2020.

[62] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685, 2020.

[63] Suzanna Becker and Geoffrey E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature, 355(6356):161-163, 1992.

[64] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, and Skye Wanderman-Milne. JAX: composable transformations of Python+NumPy programs, 2018.

[65] Tom Hennigan, Trevor Cai, Tamara Norman, and Igor Babuschkin. Haiku: Sonnet for JAX, 2020.

[66] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672-2680, 2014.

[67] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. arXiv preprint arXiv:1605.07146, 2016.

[68] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning, 2015.

[69] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In International Conference on Machine Learning, 2010.

[70] Yang You, Igor Gitman, and Boris Ginsburg. Scaling SGD batch size to 32k for imagenet training. arXiv preprint arXiv:1708.03888, 2017.

[71] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In International Conference on Learning Representations, 2017.

[72] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677, 2017.

[73] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. Transactions on Pattern Analysis and Machine Intelligence, 2017.

[74] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better ImageNet models transfer better? In Computer Cision and Pattern Recognition, 2019.

[75] Chengyue Gong, Tongzheng Ren, Mao Ye, and Qiang Liu. Maxup: A simple way to improve generalization of neural network training. arXiv preprint arXiv:2002.09024, 2020.

[76] Xiaohua Zhai, Joan Puigcerver, Alexander I Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. arXiv: Computer Vision and Pattern Recognition, 2019.

[77] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4L: Self-supervised semi-supervised learning. In International Conference on Computer Vision, 2019.

[78] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[79] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In Computer Vision and Pattern Recognition, 2010.

[80] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes (VOC) challenge. International Journal of Computer Vision, 88(2):303-338, 2010.

[81] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In Computer Vision and Pattern Recognition, 2014.

[82] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In Neural Information Processing Systems, 2015.

[83] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In International Conference on 3D Vision, 2016.

[84] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. arXiv preprint arXiv:1905.06922, 2019.

[85] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, 2015.

[86] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In International Conference on Machine Learning, 2013.

[87] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. arXiv preprint arXiv:1909.13719, 2019.

[88] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Computer Vision and Pattern Recognition, pages 2818-2826, 2016.

[89] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In European Conference on Computer Vision, 2014.

[90] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, and Peter N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In Computer Vision and Pattern Recognition, 2014.

[91] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In Workshop on 3D Representation and Recognition, Sydney, Australia, 2013.

[92] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151, 2013.

[93] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In Computer Vision and Pattern Recognition, 2012.

[94] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Pattern Recognition Workshop, 2004.

[95] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In Indian Conference on Computer Vision, Graphics and Image Processing, 2008.

[96] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In International Conference on Machine Learning, 2014.

[97] Art B Owen. A robust hybrid of lasso and ridge regression. Contemporary Mathematics, 443(7):59-72, 2007.

[98] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In International Conference on Computer Vision, 2019.

[99] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In Conference on Computer Vision and Pattern Recognition, 2017.

[100] Yuxin Wu and Kaiming He. Group normalization. In European Conference on Comperence on Computer Vision, 2018.