自监督孪生自编码器

Friederike Baier [1], **Sebastian Mair** [2], **Samuel G. Fadel** [1] ✉

[1]吕讷堡勒法纳大学，德国； [2]乌普萨拉大学，瑞典 n

✉ **通讯请联系：**
samuel.fadel@leuphana.de

资助： 本研究部分得到了由克努特与爱丽丝·瓦伦堡基金会资助的瓦伦堡人工智能、自主系统与软件计划（WASP）的支持。

arXiv:2304.02549v1 [cs.LG] 5 Apr 2023

摘要 全监督模型通常需要大量带标签训练数据，而这些数据往往成本高昂且难以获取。相比之下，自监督表征学习可减少达到相同甚至更高下游性能所需的有标签数据量。其目标是在自监督任务上预训练深度神经网络，使网络随后能够从原始输入数据中提取有意义的特征。这些特征随后用作下游任务（如图像分类）的输入。此前，自编码器和孪生网络（如SimSiam）已在这些任务中成功应用。然而，挑战依然存在，例如将特征特性（如细节层次）与给定任务和数据集相匹配。本文提出一种新的自监督方法，结合了孪生架构和去噪自编码器的优势。我们表明，我们的模型SidAE（孪生去噪自编码器）在多个数据集、设置和场景中均优于两种自监督基线方法。至关重要的是，这包括仅有少量有标签数据可用的情况。

1 引言

全监督机器学习模型通常需要大量由人工标注的训练数据，才能达到最先进的性能。然而，在许多领域，获取带标签训练数据的成本往往较高，也比获取无标签数据更具挑战性（Jing 和 Tian，2020）。无监督或自监督预训练领域的现有方法，成功减少了达到相同性能甚至更高性能所需的带标签训练数据量（Jing 和 Tian，2020）。一般而言，其目标是在自监督任务上对深度神经网络进行预训练，使其在预训练后能够从原始数据中提取有意义的特征，然后将这些特征用作所谓下游任务（如分类或目标检测）的输入。

在图像识别领域，早期研究主要致力于设计特定的前置任务。这些任务包括基于生成的方法，如图像修复（Pathak 等人，2016年）、图像着色（Zhang 等人，2016年）和图像生成（例如，使用生成对抗网络（Goodfellow 等人，2014年））；基于上下文的方法，如拼图任务（Noroozi 和 Favaro，2016年）、预测图像的旋转角度（Gidaris 等人，2018年），或在图像块级别预测相对位置（Doersch 等人，2015年）。然而，由此得到的表征往往局限于这些前置任务，这表明，更通用且具有语义意义的表征应对某些变换保持不变，而非协变（Misra 和 Maaten，2020年；Chen 等人，2020年）。基于这一观点，近期许多最先进的模型都建立在孪生网络之上（Caron 等人，2020年；Chen 等人，2020年；He 等人，2020年；Grill 等人，2020年；Chen 和 He，2021年）。其思路是让模型学会识别同一实体的两个不同版本属于同一实体，并忽略导致这些版本彼此不同、但与实体识别无关的因素。简单孪生（SimSiam）（Chen 和 He，2021年）模型旨在解决这一策略常见的表征坍塌问题。它在模型的一个分支上使用停止梯度操作，每个分支对应一个输入视图。

另一类广为人知、用于预训练网络以充当特征提取器的模型是自编码器（Hinton 和 Salakhutdinov，2006年）。因此，自编码器的原理是最大化输入与潜在表示之间的互信息（Vincent 等人，2010年）。然而，良好的特征不应包含尽可能多的输入信息，而只应保留其中最相关的部分（Tian 等人，2020年）。因此，理想的自编码器型

图像特征提取器只会提取确定图像身份所必需的信息。通常的做法是使潜在空间的维度低于输入的维度，并在将图像输入编码器之前向其添加噪声（例如高斯噪声）。这就是去噪自编码器的思路（Vincent 等人，2008）。

单独来看，这两种学习表征的方法都能带来简单而有效的解决方案。正是这种简洁性，使我们能够利用这两种方法来设计一个或许更强大的特征提取器，因为这两种策略以彼此不冲突的方式学习表征。

本文探讨能否将孪生网络与去噪自编码器结合，构建一种新的自监督表征学习模型，使其兼具两者的优点，从而弥补各组成部分自身的不足。我们提出了一种名为 SidAE（孪生去噪自编码器）的新模型，旨在融合两种强大的学习原理：孪生网络利用样本输入的多个视图进行学习，去噪自编码器则具有抗噪能力。我们展示了如何巧妙地将两者结合起来，得到一种简单而有效的模型。实验表明，在各种场景下，无论下游任务拥有全部有标签数据还是仅有其中一部分，也无论是否对特征提取器进行微调，SidAE 在下游任务中的表现都优于其组成部分。

2 自监督表征学习

自监督学习的一项指导原则是利用大量无标签数据来训练模型。关键在于使用从输入本身提取的信息来定义损失函数。一般而言，自监督可用于多种领域，包括但不限于文本、语音和视频。本文将聚焦于图像识别领域。

在图像识别领域，自监督表征学习用于训练模型提取“有用的表征”，以提升下游性能。我们考虑两个阶段：（1）自监督预训练；（2）有监督下游任务。预训练阶段的目标是训练一个所谓的“编码器”神经网络，使其学会从原始输入信号（即像素）中提取有意义的信息。第二阶段将编码器用作特征提取器。在此阶段，使用有标签数据和有监督损失来训练分类器，例如一个简单的全连接层，并将其置于编码器从输入中提取的表征之上。对于下游任务，可以对编码器的权重进行微调（优化），也可以将其冻结（不优化）。这凸显了自监督学习在有标签数据较少的情形下的优势。

问题在于，什么样的特征才有意义，因而有用？又该如何调整模型，使其生成满足特定属性的表征？有人认为，表征应尽可能多地包含有关输入的信息（信息最大化原则；（Linsker，1989））。然而，也有人认为，特定任务表征只需包含解决某项具体任务所必需的信息，因为与任务无关的信息可被视为噪声，甚至可能损害性能（信息最小化原则；（Tian 等人，2020年））。对于通用表征，这一点在某种程度上同样成立。只要某些变换不改变物体的身份，表征就应对这些变换保持不变，也就是说，应不受所谓“干扰因素”的影响，例如光照、色彩调整和方向。一种观点认为，能够预测被遮蔽部分或推断数据属性的模型，应当对内容具备某种“高层次理解”（Henaff，2020年；Baevski 等人，2022年；Assran 等人，2023年）。

2.0.1 孪生网络

孪生（也称为联合嵌入（Assran 等人，2023年））架构旨在让所谓的骨干网络学会：图像的两个失真版本仍表示同一个物体。直观地说，从同一实体提取的特征，即使经过这些失真处理，也应该彼此相似。

我们提出的模型基于 SimSiam（Chen 和 He，2021年）。与其他孪生架构相比，它结构简单，因此非常适合作为更复杂架构的基础。一般而言，孪生网络由两个分支组成，分别对同一输入的两个不同版本 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ （“视图”） $\mathbf{x}$进行编码，详情如下。SimSiam 的每个分支都包含一个编码器和一个

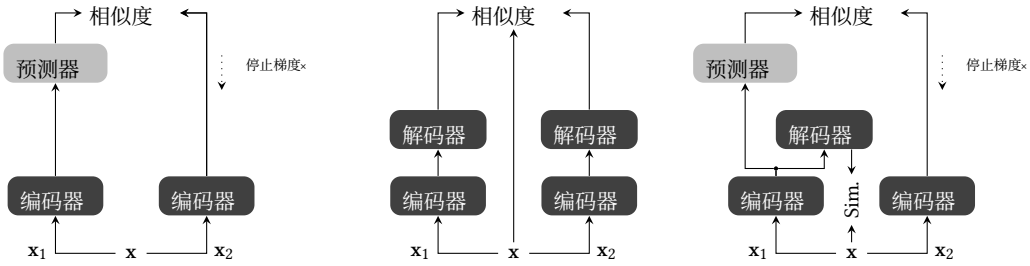


图 1. 自监督学习与两种范式：孪生网络（左）和（去噪）自编码器（中），并与我们的模型 SidAE（右）进行比较。

所谓的预测器网络，其中两个分支的编码器共享权重。其他方法中的网络权重则有所不同。例如，在 MoCo（He 等人，2020年）和 BYOL（Grill 等人，2020年）中，一个网络的权重是另一个网络权重的滞后移动平均值。图1 （左）展示了 SimSiam 的架构；为简洁起见，图中仅显示一侧的停止梯度操作。

在 SimSiam 和其他先进模型中，编码器由一个深度神经网络构成，通常称为“骨干网络”。在某些情况下，还会有一个额外的投影器网络。投影器由几个全连接层组成，其作用仅仅是将骨干网络的输出作为输入。骨干网络和投影器合称为编码器。SimCLR 中也使用了投影器（Chen 等人，2020年），它能够提升下游性能。

预训练的第一步是从一张输入图像$\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$提取两个不同版本（“视图”）。通过分别应用两组数据增强 t_1 和 t_2（从一个数据增强集合 T中随机采样）得到这些视图。随后使用编码器 Enc(·) 和预测器 Pred(·) 网络作为

$$\mathbf{x}_j = t_j(\mathbf{x}), \quad \mathbf{z}_j = \text{Enc}(\mathbf{x}_j), \quad \mathbf{p}_j = \text{Pred}(\mathbf{z}_j),$$

其中 $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$ $\mathbf{x}_j \in \mathbb{R}^{d_{\text{in}}}, \mathbf{z}_j \in \mathbb{R}^{d_{\text{hid}}}, \mathbf{p}_j \in \mathbb{R}^{d_{\text{hid}}}$。这里， d_{in} 是输入维度， d_{hid} 隐藏或潜在维度，以及 $j \in \{1, 2\}$。使用 $d_{\text{hid}} = \frac{1}{4} d_{\text{in}}$ 被发现能带来更好的结果相比于 $d_{\text{hid}} = d_{\text{in}}$ （Chen 和 He， 2021年）。

通过这种方式，可以通过最小化距离来训练模型 $D(\mathbf{p}_i, \mathbf{z}_j)$ 与 i， $j \in \{1, 2\}$ 和 $i \neq j$。因此，通过使图像的表征彼此接近来最小化损失。负余弦相似度和交叉熵都可以用作 $D(\cdot, \cdot)$，但早期报告显示，前者能带来更好的结果（Chen 和 He，2021年）。关键的是，停止梯度操作应用于 $\mathbf{z}_j$。因此，在通过其中一个分支进行反向传播时，另一个分支会被视为固定变换。更具体地说，SimSiam 的损失计算如下

$$L_{\text{si}} = \frac{1}{2} D(\mathbf{p}_1, \text{stopgrad}(\mathbf{z}_2)) + \frac{1}{2} D(\mathbf{p}_2, \text{stopgrad}(\mathbf{z}_1)), \tag{1}$$

其中 stopgrad(·) 表示停止梯度操作。这对于避免坍塌解至关重要（Chen 和 He，2021年），并作为早期尝试的替代方案，例如使用一个分支的权重作为另一个分支的滞后移动平均（He 等人，2020年；Grill 等人，2020年）。

2.0.2 去噪自编码器

自编码器是我们考虑的另一类自监督模型。它们共享相同的底层原理，即通过使用该表示来重构原始数据来学习数据表示。编码器 $\text{Enc}(\cdot) : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{hid}}}$ 从输入 $\mathbf{x}$ 中提取该表示 $\mathbf{z}$，此处称为潜在表示（或编码），而解码器$\text{Dec}(\cdot) : \mathbb{R}^{d_{\text{hid}}} \to \mathbb{R}^{d_{\text{in}}}$ 使用 $\mathbf{z}$ 来重构原始输入 $\mathbf{x}$。传统上，编码器和解码器在架构上往往相互“镜像”，但并非总是如此。然后训练模型以使重构尽可能接近原始输入，用

$$\mathbf{z} = \text{Enc}(\mathbf{x}), \quad \mathbf{x}' = \text{Dec}(\mathbf{z}), \quad L_{\text{AE}} = D(\mathbf{x}', \mathbf{x}).$$

自编码器通常基于最大化某个输入与潜在编码之间互信息的概念构建 $\mathbf{x}$ 与潜在编码 $\mathbf{z}$ (Vincent 等人，2010年）。然而，它们需要以某种方式加以限制

这样模型就不会只是学习一个恒等映射，使得 $\mathbf{z}$ 等于 $\mathbf{x}$；通常的做法是降低 $\mathbf{z}$ 的维度。对于去噪自编码器，输入 $\mathbf{x}$ 会受到扰动（例如，添加高斯噪声），并将这个受扰动的版本 $\tilde{\mathbf{x}}$ 作为模型的输入

$$\mathbf{z} = \text{Enc}(\tilde{\mathbf{x}}), \quad \mathbf{x}' = \text{Dec}(\mathbf{z}), \quad L_{\text{DAE}} = D(\mathbf{x}', \mathbf{x}).$$

训练中通过最小化损失，使 $\mathbf{x}$ 和 $\mathbf{x}'$ 尽可能彼此相似。请注意，去噪自编码器看不到 $\mathbf{x}$，只能看到其受损版本 $\tilde{\mathbf{x}}$，但仍需训练它去匹配 $\mathbf{x}$。因此，模型应学会更加关注重建原始输入所需的关键信息；理想情况下，表征仅编码相关信号，不含噪声。

3 孪生去噪自编码器

新提出的模型包含 SimSiam 的部分组件（Chen 和 He，2021年）以及一个去噪自编码器。该去噪自编码器的设计受到Li et al.（2017）和Wickramasingheet al.（2021年）的研究启发。据我们所知，此处展示的确切架构及其应用此前尚未出现。请注意，多个模型中出现的组件（即编码器、解码器和预测器网络）在 SimSiam、SidAE 和去噪自编码器中均相同。

3.1 动机

我们假设，将孪生网络与去噪自编码器相结合的新架构能够产生更强大的方法。由于它借鉴了两种互为补充的思路，它们可以弥补各自组件的不足。具体而言，去噪自编码器基于数据的单个受损视图进行训练，而孪生网络则使用两个视图。此外，孪生网络并不试图根据编码器提取的信息重建原始输入，而去噪自编码器则会这样做。这种行为差异有利于自监督，即让去噪自编码器访问同一样本的两个不同视图。同时，孪生部分所用的编码器也会受到鼓励，保留足够的信息供解码器重建输入。这种协同作用是 SidAE 的基本设计原则。

直观来说，与使用 SimSiam 对网络进行预训练相比，使用 SidAE 应能使网络学会两个视图来自同一个物体，并识别它们分别来自哪个物体。对于 SimSiam，原始输入不会直接用于优化模型，因为输入模型的只有经过增强的版本，损失则根据潜在空间中的表征计算。在 SidAE 中，去噪自编码器部分将受干扰的局部信息映射为原始、未受干扰的全局信息。自编码器经过优化，以最大化未受损输入与潜在表示之间的互信息。因此，模型会被明确训练，将有关原始输入的信息编码到潜在表示 $\mathbf{z}_1$ 和 $\mathbf{z}_2$ 中。

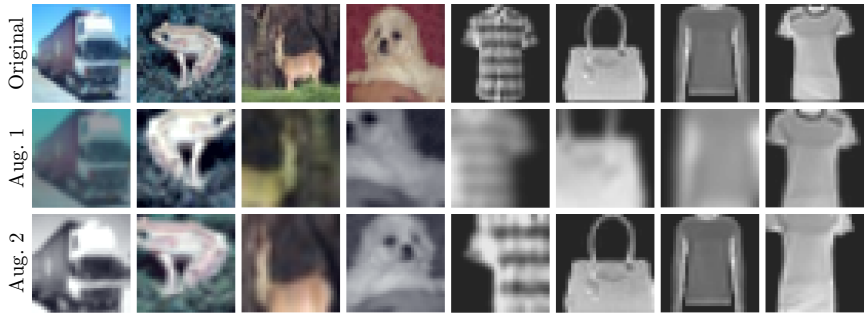
然而，与去噪自编码器相比，SidAE 不应仅仅被优化为尽可能多地将原始输入的信息编码到潜在编码中（同时对噪声具有鲁棒性）。它还应忽略特定的干扰因素。这是通过其网络的孪生部分实现的。这应能防止特征包含无关细节。接下来，我们将描述 SidAE 的架构并指定其损失函数。SidAE 的可视化概览如图 1（右）所示。

3.2 架构

输入。

SidAE 使用两个视图 $\mathbf{x}_1$ 和 $\mathbf{x}_2$，它们通过应用两个数据增强 t_1 和 t_2 获得，我们从一组数据增强 T 中采样这些增强。我们采用与 SimSiam （ Chen 和 He，2021 ）中用于 CIFAR-10 相同的数据增强流水线，该流水线由以下以 PyTorch（Paszke 等人，2019年）符号表示的变换组成：(i) 随机调整大小裁剪，尺度范围为 $[0.2, 1]$，(ii) 颜色抖动，亮度=0.4、对比度=0.4、饱和度=0.4 和色调=0.1，以概率 $p = 0.8$ 应用，(iii) 随机灰度化，以概率 $p = 0.2$ 应用，以及 (iv) 随机水平翻转，以概率 $p = 0.5$ 应用。

对于 MNIST 和 Fashion-MNIST，我们还会额外应用标准差为 $[0.1, 2.0]$ 、以 $p = 0.5$ 的概率应用的高斯模糊。这种增强通常也会被纳入



Fig图 2. SidAE 中所用数据增强示例：四张 CIFAR-10 图像（左侧）和四张 Fashion-MNIS T (right）图像，以及两种数据增强结果。

用于在 ImageNet 上训练孪生网络。不过，在 CIFAR-10（Chen 和 He，2021年）上训练 SimSiam 时并未使用这种方法。为了确保结果公平且具有可比性，我们也遵循这一做法，因为与所展示物体的复杂程度相比，CIFAR-10 图像的分辨率本来就相当低。进一步模糊会使人类观察者难以辨认这些图像，这与变换应代表人类仍能识别物体的因素这一理念相矛盾。不过，Fashion-MNIST 和 MNIST 样本在模糊后仍然可以辨认。我们在图 2 中展示了一些变换示例。

编码器。

编码器中的骨干网络基于 PyTorch（Paszke 等人，2019年）对 ResNet-18 的实现（He 等人，2016年）。与 SimSiam 一样，我们将最后一个全连接层替换为一个由三层组成的投影多层感知机。

解码器。

解码器是另一个围绕编码器构建的神经网络。输入 $\mathbf{x}_1$ 和 $\mathbf{x}_2$编码为潜在表示 $\mathbf{z}_1$ 和 $\mathbf{z}_2$后，会被送入解码器，生成重建结果 $\mathbf{x}'_1$ 和 $\mathbf{x}'_2$。解码器与编码器相对应，逐步对数据进行上采样，以重构出尺寸与原始输入相同的输出（三个通道，每个通道的尺寸为 32×32）。为此，我们将一个全连接层与五个转置卷积层依次堆叠。每个转置卷积层的输出通道数均参照骨干网络的下采样过程设定。它们的卷积核尺寸分别为 3×3、 padding $= 1$、 output padding $= 1$ 和 stride $= 2$。编码器使用了跳跃连接，但我们决定不构建残差解码器，从而形成较为简单的架构，降低复杂度，并使其与 SimSiam 保持相似。

预测器。

预测器是一个两层的多层感知机，与 SimSiam（Chen 和 He，2021年）中的相同。它将潜在表示 $\mathbf{z}_1$ 和 $\mathbf{z}_2$ 映射到另一个潜在空间，以生成两个嵌入 $\mathbf{p}_1$ 和 $\mathbf{p}_2$。两者 $\mathbf{z}_i$ 和$\mathbf{p}_j (i,\ j \in \{1, 2\}$ 和 $i \neq j)$ 的维度相同，因为稍后会对它们进行相互比较。图3展示了包含编码器、投影器、解码器和预测器的完整结构。

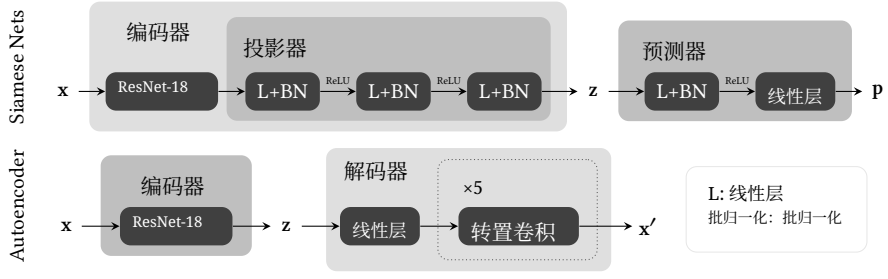


图 3. 我们实验中使用的模型组件。

损失。
SidAE 的损失由两项组成，一项对应网络的孪生部分，另一项对应自编码器部分。前者用于比较 $\mathbf{z}_1$ 和 $\mathbf{p}_2$ 以及 $\mathbf{z}_2$ 和 $\mathbf{p}_1$。后者分别将重建结果 $\mathbf{x}_1'$ 和 $\mathbf{x}_2'$ 与增强后的输入 $\mathbf{x}_1$ 和 $\mathbf{x}_2$进行比较。在 SimSiam 中，损失的子项会按 $\frac{1}{2}$ 进行缩放，相当于将整个损失按 $\frac{1}{2}$ 进行缩放。我们对损失中的自编码器项也采用相同的处理。此外，我们通过设置参数 $w \in [0,1]$，为损失中的孪生部分和自编码器部分引入一个加权项。该参数控制模型中每种机制相对重要性的大小。完整的损失 L_{sidae} 定义为

$$\begin{aligned} L_{\text{dae}} &= \frac{1}{2} D_{\text{mse}}(\mathbf{x}, \mathbf{x}_1') + \frac{1}{2} D_{\text{mse}}(\mathbf{x}, \mathbf{x}_2'), \\ L_{\text{si}} &= \frac{1}{2} D_{\text{ncs}}(\mathbf{p}_1, \text{stopgrad}(\mathbf{z}_2)) + \frac{1}{2} D_{\text{ncs}}(\mathbf{p}_2, \text{stopgrad}(\mathbf{z}_1)), \\ L_{\text{sidae}} &= w L_{\text{dae}} + (1 - w) L_{\text{si}}, \end{aligned}$$

其中，L_{dae} 是去噪自编码器损失，L_{si} 是孪生损失。距离 D_{mse} 是均方误差（MSE），而 D_{ncs} 是负余弦相似度（NCS），定义如下

$$D_{\text{mse}}(\mathbf{a}, \mathbf{b}) = \frac{1}{d} \sum_{j=1}^{d} (a_j - b_j)^2 \quad \text{and} \quad D_{\text{ncs}}(\mathbf{a}, \mathbf{b}) = -\frac{\mathbf{a}^\top \mathbf{b}}{\max(\|\mathbf{a}\|_2 \cdot \|\mathbf{b}\|_2, \epsilon)},$$

分别地。这里，$\mathbf{a}$、$\mathbf{b} \in \mathbb{R}^d$ 和 $\epsilon = 10^{-8}$ 可避免除以零。

4 实验

现在，我们针对所提出的 SidAE 模型进行若干实验，将其与相关基线进行比较，评估参数 w 的影响，并考察预训练数据量对多个真实世界数据集上的下游任务有何影响。为清晰起见并避免重复，下文提到自编码器时均指去噪自编码器。

4.1 实验设置

数据。
我们在四个真实世界数据集上进行评估。*CIFAR-10* (Krizhevsky 和 Hinton，2009) 包含 50,000 张彩色训练图像和 10,000 张测试图像，图像尺寸为 32×32，具有三个颜色通道，分属十个互斥类别。*MNIST* 数据集（LeCun 等人，1995）包含 28×28 像素大小的手写数字图像，因此共有十个类别。它包含 60,000 张训练图像和 10,000 张测试图像。*Fashion-MNIST*(Xiao 等人，2017) 与 MNIST 类似，但展示的是十个类别的时尚物品。最后，*STL-10* (Coates 等人，2011年) 同样包含十个类别，图像分辨率更高，为 96×96 像素，具有三个颜色通道。该数据集专为自监督任务设计，包含 5,000 张带标签的训练图像、8,000 张带标签的测试图像，此外还有 100,000 张无标签图像。为保持模型可比并避免模型容量不同带来的变异，我们使用缩小版 STL-10，其尺寸与 CIFAR-10 相同，即 32×32 像素，具有三个颜色通道。

STL-10 被指定用于自监督或无监督预训练。因此，我们使用无标签数据对模型进行预训练，并使用训练集和测试集在分类任务上训练和评估模型。对于其他数据集，我们采用 SimSiam（Chen 和 He，2021年）中使用的相同训练方案。我们使用训练集进行自监督预训练，并训练分类器以完成下游任务。测试集用于评估。此外，我们模拟更具挑战性的任务：仅使用一小部分（即 1%）的有标签训练数据，在下游任务上进行监督训练。每个数据集的子集均随机抽取，但在所有运行中保持一致，以便进行公平比较。我们始终使用完整训练集（对于 STL-10，则使用无标签数据集）进行预训练。为了在所有数据集上使用相同的架构，我们将所有图像调整为 32×32。

基线。
我们将SidAE与*SimSiam* (Chen 和 He, 2021年)以及一个去噪自编码器 (Vincent 等人,2008)进行比较。对于SimSiam，我们使用与Chen 和 He (2021年)相同的架构。因此，SimSiam的损失

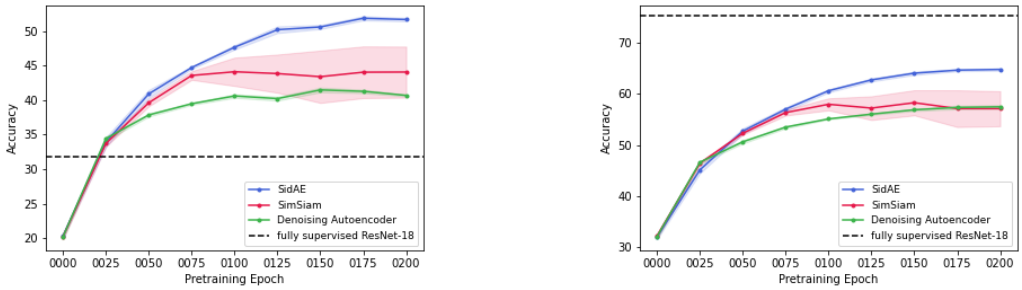


图4. 使用冻结的预训练骨干网络，在不同预训练阶段后将 CIFAR-10 上的分类准确率作为下游任务进行评估。下游训练分别使用 1%（左）和 100%（右）的训练数据。

是两个视图之间负余弦相似度的对称损失。自编码器的编码器使用ResNet-18以及SimSiam和SidAE中所用的三层投影多层感知机，解码器则与SidAE中的相同。SidAE的潜在维度设置为 $d_{\text{hid}} = 2048$，因为这在Chen 和 He (2021年)中对SimSiam效果最好。我们在附录中评估了其他潜在维度的选择。由于SidAE一次看到两个噪声视图，我们也向自编码器提供以相同方式构建的两个噪声视图。我们使用均方误差作为自编码器的损失。此外，我们将分类作为下游任务。在那里，我们还与完全监督的ResNet-18进行比较。

设置。
所有实验均使用 Python，并在 PyTorch 中实现所有模型（Paszke 等人，2019年）。所有实验均在一台配备两颗 32 核 AMD Epyc CPU、512 GB 随机存取存储器以及一块具有 40GB 显存的 NVIDIA A100 GPU 的机器上运行。我们基于 SimSiam 的原始实现重新实现了该方法，并在预训练时采用相同的设置。我们使用随机梯度下降作为优化器，并采用余弦衰减调度，初始学习率为 0.03，动量为 0.9，权重衰减为 0.0005，批量大小为 512。

我们使用 SidAE、SimSiam 和去噪自编码器对骨干网络进行 200 轮预训练，并将冻结的预训练骨干网络用作有监督下游任务（即图像分类）的特征提取器。我们在预训练的不同阶段评估每个预训练模型，即在完成 25、50、75、100、125、150、175 和 200 轮预训练后进行评估，以考察预训练时长对所学表征质量的影响。我们将骨干网络中的投影器多层感知机替换为分类器的全连接层，其输入维度等于 z。输出维度设为与各数据集中的类别数相匹配。该下游任务使用随机梯度下降训练 50 轮，学习率为 0.05，动量为 0.9，批量大小为 256。由于骨干网络的权重被冻结，下游任务训练时仅更新最后一个全连接层的权重。随后，我们将评估对冻结权重进行微调的情形。

4.2 结果

每项实验均进行五次运行，并报告平均准确率，包括下游分类的标准误。

目前，我们将权重固定为 $w \in [0, 1]$，用于在孪生损失和自编码器损失之间进行插值 $w = 0.5$，并在之后评估其影响。图4展示了预训练对 CIFAR-10 下游分类的影响。我们发现，预训练轮次越多，下游任务的准确率越高。左侧展示的是分类任务只有 1% 带标签训练数据可用的情况。在这种情况下，自监督方法仅进行 25 轮预训练，其表现就已超过完全监督的分类器，说明预训练有所助益。请注意，当 100% 的带标签训练数据可用时，如果不微调编码器，模型未必能超过监督基线，如右侧所示。值得注意的是，监督基线在 100% 和 1% 数据量下的性能差距，远大于自监督模型。这些

表 1. SidAE 使用 $w = 0.5$、SimSiam 和去噪自编码器的下游分类准确率，在五次运行中取平均并包括标准误。

Data	预训练	SidAE	SimSiam	自编码器
CIFAR-10	100%	**64.75** (0.27)	57.08 (3.16)	57.44 (0.31)
	1%	**51.67** (0.28)	44.06 (3.74)	40.67 (0.25)
Fashion-MNIST	100%	**88.78** (0.12)	85.10 (0.09)	87.67 (0.05)
	1%	**81.62** (0.22)	77.84 (0.19)	80.20 (0.31)
MNIST	100%	**97.80** (0.13)	96.59 (0.60)	97.23 (0.11)
	1%	**94.65** (0.49)	79.42 (5.15)	88.41 (0.46)
STL-10	STL-10	**61.18** (0.22)	56.74 (0.74)	48.39 (0.22)
	CIFAR-10	**66.20** (0.31)	56.64 (0.25)	57.08 (3.16)

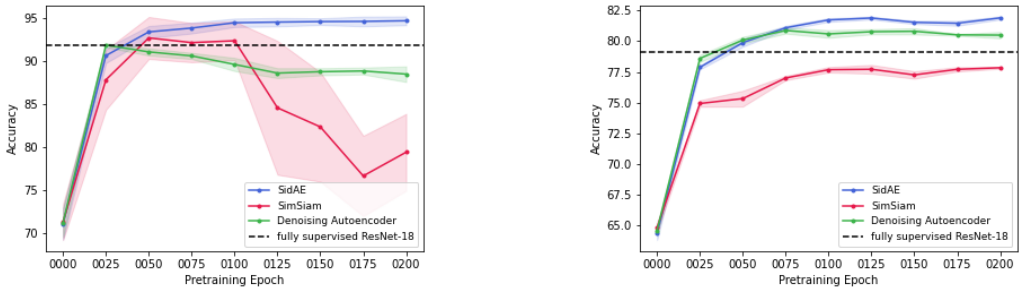


图 5. 使用 1% 的训练数据进行下游训练时，在 MNIST（左）和 Fashion-MNIST（右）上的分类准确率。

这些观察结果表明，监督模型需要大量训练数据；如果只有少量有标签数据，其表现就相当差。请注意，SimSiam 的标准误最大，而 SidAE 更稳定。总体而言，SidAE 的表现优于自编码器和 SimSiam。

表 1 还展示了其他数据集在预训练 200 轮后、对五个随机种子的结果取平均所得的分类准确率。从中可以看出，与基线竞争方法相比，SidAE 在 Fashion-MNIST、MNIST 和 STL-10 上也优于其基线。图 5展示了使用 1% 的训练数据进行下游训练时，在 MNIST 和 Fashion-MNIST 上的分类准确率。对于 CIFAR-10 （图 4），SimSiam 的表现要么优于（1%），要么与（100%）自编码器相当；但 MNIST 和 Fashion-MNIST 上的情况则有所不同。经过 100 轮训练后，SimSiam 在 MNIST 上尤其不稳定。在这两个数据集上，自编码器的表现都优于 SimSiam。此外，我们提出的 SidAE 模型取得了优于全监督基线的结果。

使用 100% 的训练数据进行下游训练所得结果相同，因此不再展示。

权重 w 对 SidAE 的影响。

现在，我们考察损失函数中权重 $w \in [0, 1]$的权衡作用。当 $w = 0$ 时，仅使用孪生损失；而当 $w = 1$ 时，仅使用自编码器损失。图 6 再次通过下游准确率展示了在 1% 情形下权重w 对 SidAE 的影响。当使用 SidAE 且 $w \in \{0.125, 0.25\}$ 时，分类器的性能确实优于此前采用标准设置 $w = 0.5$ 时的性能。若增大自编码器的权重，例如使用

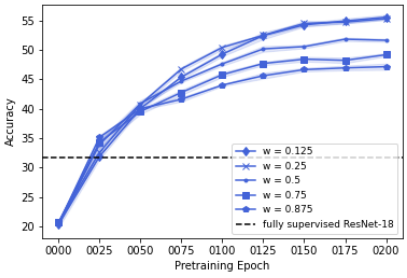


图6。 SidAE：权重 w 对 CIFAR-10 的影响。

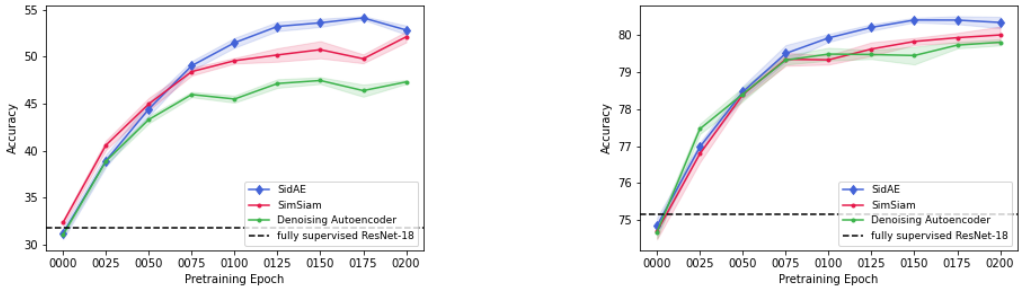


图 7. 使用不同模型预训练的编码器，在下游任务的不同预训练阶段，使用 1%（左）或 100%（右）的下游任务训练数据时的分类准确率（CIFAR-10）。

$w \in \{0.75, 0.875\}$，性能会下降并趋近于仅使用自编码器时的水平。请注意，表现最佳的设置 $w = 0.125$ 使 SidAE 在先前实验中的性能差距进一步扩大。100% 情形下的结果类似，因此未予展示。表1(附录)列出了在 Fashion-MNIST、MNIST 和 STL-10 上采用不同 w 取值时的结果。

微调下游任务。

到目前为止，在训练下游分类器时，我们一直冻结编码器的预训练权重。这样做是为了评估预训练过程在编码器不受有标签数据影响的情况下，生成优质特征提取器的能力。现在，我们不再冻结预训练编码器的权重，而是在监督式下游训练过程中对其进行微调。因此，与从头开始完全通过监督方式训练 ResNet 不同，模型权重初始化为自监督预训练所获得的权重，而不是随机初始化。这样我们就能研究预训练是否能提升性能，或者未经预训练、权重随机初始化的模型能否达到相同的性能。在这项实验中，我们在 SidAE 中设置 $w = 0.125$，因为它在之前的实验中取得了最佳性能（图 6）。

图 7 展示了在允许微调的情况下，CIFAR-10 上的下游分类准确率；也就是说，骨干网络可以适应下游任务。与之前一样，我们提出的 SidAE 模型在几乎所有情况下都优于两种自监督基线方法。与早期实验不同，现在所有自监督模型在训练数据占 1%（左侧）和 100%（右侧）时，都超过了监督基线。因此，采用某种形式的自监督预训练不仅有益于仅有少量有标签数据的场景，在有充足有标签数据的情况下也同样有益。

STL-10 上的结果。

STL-10 数据集专为自监督学习设计，其中既包含一个小型有标签训练集，也包含一组无标签训练数据。现在，我们在 STL-10 的无标签训练数据上对骨干网络进行预训练，并将冻结的骨干网络用作特征提取器，以完成 STL-10 和 CIFAR-10 上的分类下游任务。这使我们能够评估是否可以在具有相似特征的数据集上进行特征提取预训练。

在本实验中，我们将 $w = 0.25$ 设为 SidAE 的参数，因为它在两个数据集上的下游准确率最高。不过，采用 $w = 0.125$也能得到相近的结果。STL-10 和 CIFAR-10 上的分类准确率分别展示在图 8的左侧和右侧。对于 STL-10，可以看出 SidAE 和 SimSiam 的表现都优于全监督基线，而去噪自编码器的表现则稍逊一筹。CIFAR-10 上的情况有所不同，全监督基线仍然更好。不过，值得注意的是，SidAE 的性能与骨干网络在 CIFAR-10 本身上预训练时的情况非常相近（图 4）。总体而言，SidAE 的表现远优于两种自监督基线方法。

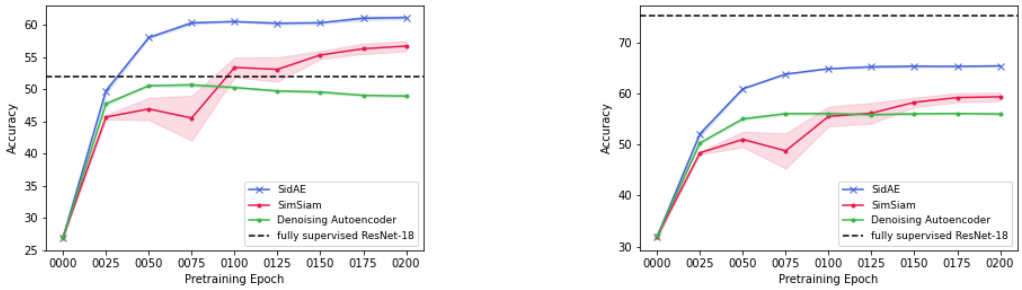


图 8. 使用 STL-10（左）和 CIFAR-10（右）未标注训练数据预训练的编码器进行分类时的准确率 未标注 的 STL-10 训练数据。

5 相关工作

许多近期的先进模型都基于孪生网络（Caron 等人，2020年；Chen 等人，2020年；He 等人，2020年；Grill 等人，2020年；Chen 和 He，2021年；Liu 等人，2022年）。其思路是让模型学会识别同一物体的两个不同版本代表的是同一个物体，同时不将造成它们差异的因素用于身份识别。主要挑战在于避免表征坍塌，即将所有输入都映射到同一个特征向量。一种广泛采用的策略是使用 InfoNCE 这类对比损失（Oord 等人，2018年；Misra 和 Maaten，2020年；Chen 等人，2020年），这需要大量负样本才能取得良好性能。针对这一问题，人们提出了多种方法，例如使用较大的批次大小（Chen 等人，2020年）、使用记忆库存储负样本（Wu 等人，2018年；Misra 和 Maaten，2020年），或使用“队列”保留先前批次中的负样本（He 等人，2020年）。另一种方法是，Caron 等人（2020年）通过在线聚类将图像嵌入映射到原型向量，从而避免成对比较。这些方法缓解了处理负样本时的问题，但并未彻底解决：计算成本，以及所选训练样本的质量。

此后，新的方法提出了更简单的方式来解决坍塌问题。BYOL（Grill 等人，2020年）通过采用动量编码器，完全避免了对负样本的需求。在这种设计中，孪生网络其中一个分支的权重，是另一个分支权重的滞后移动平均。这使模型难以收敛到坍塌状态，但也让优化过程变得更加复杂。SimSiam（Chen 和 He，2021年）则在其中一个分支上使用停止梯度操作，证明这足以防止坍塌解。这不仅在概念上简单明了，实现起来也更简洁。最近，Liu 等人（2022年）表明，停止梯度操作会隐式引入一种促进特征去相关的约束，从而解释了为什么 SimSiam 的简单架构表现出色。

6 结论

本文探讨自监督表征学习，其中孪生网络和自编码器常被使用，并取得了良好效果。像 SimSiam 这样的孪生网络（Chen 等人，2020年）允许对数据进行多个可能含噪的视图，而（去噪）自编码器（Vincent 等人，2008）则通过重建噪声输入，在瓶颈层提取输入的关键特征。

我们提出将孪生网络与去噪自编码器的优势结合起来，以学习有意义的数据表示。为此，我们引入了一种名为 SidAE（孪生去噪自编码器）的新模型。我们在多个真实世界数据集上的各种场景中，对模型学到的表征进行了实证评估。为此，我们考察了分类这一项下游任务。因此，我们首先以自监督方式在无标签训练数据上预训练模型，然后使用提取的特征来训练监督分类器。在实验中，我们将 SidAE 与 SimSiam 和去噪自编码器进行了比较。

在我们所有的实验中，就下游分类准确率的平均值而言，SidAE 始终优于两个自监督基线模型。此外，与 SimSiam 相比，SidAE 的结果对于初始化随机种子和所用预训练轮次的数量更为稳定。

7 伦理声明

与许多机器学习模型，尤其是基于神经网络的模型一样，本文介绍的方法依赖大量数据。我们认为这是本研究中伦理风险的主要来源，因为自监督方法可以从无标签数据（包括整理不充分的数据）中学习。原因在于，数据可能隐含或明确地包含偏差，例如对某些内容的过度呈现或呈现不足。本文以分类作为示例性下游任务。这种过度呈现或呈现不足可能导致不理想的性能，例如，在针对代表性不足类别的分类任务中。

此外，使用本文讨论的通用特征提取器时应保持谨慎。模型究竟学到了什么并不一定十分明确，例如，它从输入中提取了哪些特征来进行分类。我们特别关注涉及关键决策的情况，以及这些决策是否易于理解。这些问题在其他下游任务中也应予以考虑。

参考文献

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N.（2023年）。基于图像的自监督学习：一种联合嵌入预测架构。

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M.（2022年）。Data2vec：一种用于语音、视觉和语言自监督学习的通用框架。载于国际机器学习大会，第1298–1312页。PMLR。

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., 和 Joulin, A.（2020年）。通过对比簇分配进行视觉特征的无监督学习。神经信息处理系统进展，33:9912–9924。

Chen, T., Kornblith, S., Norouzi, M., 和 Hinton, G.（2020年）。一种用于视觉表征对比学习的简单框架。载于国际机器学习大会，第1597–1607页。PMLR。

Chen, X. 和 He, K.（2021年）。探索简单的孪生表征学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第15750–15758页。

Coates, A., Ng, A., 和 Lee, H.（2011年）。无监督特征学习中的单层网络分析。载于第十四届国际人工智能与统计学会议论文集，第215–223页。JMLR研讨会与会议论文集。

Doersch, C., Gupta, A., 和 Efros, A. A.（2015年）。通过上下文预测进行无监督视觉表征学习。载于*IEEE*国际计算机视觉会议论文集，第1422–1430页。

Gidaris, S., Singh, P., 和 Komodakis, N.（2018年）。通过预测图像旋转进行无监督表征学习。载于国际学习表征会议。

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., 和 Bengio, Y.（2014年）。生成对抗网络。载于第 *27* 届神经信息处理系统国际会议论文集 - 第 *2* 卷，第2672–2680页。

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M. 等（2020年）。自举潜在表征：一种新的自监督学习方法。神经信息处理系统进展，33：21271–21284。

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R.（2020年）。用于无监督视觉表征学习的动量对比。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第9729–9738页。

He, K., Zhang, X., Ren, S., 和 Sun, J.（2016）。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页。

Henaff, O.（2020）。结合对比预测编码的数据高效图像识别。载于第*37*届国际机器学习会议论文集，第4182–4192页。PMLR。Hinton, G. E. 和 Salakhutdinov, R. R.（2006）。利用神经网络降低数据维度。《科学》，313(5786):504–507。Jing, L. 和 Tian, Y.（2020）。深度神经网络的自监督视觉特征学习：综述。*IEEE*模式分析与机器智能汇刊，43(11):4037–4058。Krizhevsky, A. 和 Hinton（2009）。从微型图像中学习多层特征。技术报告。LeCun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Muller, U. A., Sackinger, E., Simard, P. 等（1995）。分类学习算法：手写数字识别比较。神经网络：统计力学视角，261(276):2。Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X. 和 Yang, M.-H.（2017）。通过特征变换实现通用风格迁移。神经信息处理系统进展，30。Linsker, R.（1989）。如何通过最大化输入与输出信号间的互信息来生成有序映射。神经计算，1(3):402–411。Liu, K.-J., Suganuma, M. 和 Okatani, T.（2022）。在非对比式自监督学习中，从非对称技巧到去相关原则。神经信息处理系统进展，35:19824–19835。Misra, I. 和 Maaten, L. v. d.（2020）。自监督学习前置任务不变表征。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第6707–6717页。Noroozi, M. 和 Favaro, P.（2016）。通过解决拼图任务进行视觉表征的无监督学习。载于欧洲计算机视觉会议，第69–84页。Springer。Oord, A. v. d., Li, Y. 和 Vinyals, O.（2018）。基于对比预测编码的表征学习。*arXiv*预印本 *arXiv:1807.03748*。Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. 等（2019）。PyTorch：一种命令式风格、高性能的深度学习库。神经信息处理系统进展，32。Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T. 和 Efros, A. A.（2016）。上下文编码器：通过图像修复学习特征。载于*IEEE*计算机视觉与模式识别会议论文集，第2536–2544页。Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C. 和 Isola, P.（2020）。什么样的视图适合对比学习？神经信息处理系统进展，33:6827–6839。Vincent, P., Larochelle, H., Bengio, Y. 和 Manzagol, P.-A.（2008）。利用去噪自编码器提取并组合鲁棒特征。载于第*25*届国际机器学习会议论文集，第1096–1103页。Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A. 和 Bottou, L.（2010）。堆叠去噪自编码器：在深度网络中借助局部去噪准则学习有用表征。机器学习研究期刊，11(12)。Wickramasinghe, C. S., Marino, D. L. 和 Manic, M.（2021）。用于从高维数据中无监督学习特征的ResNet自编码器：抵抗性能退化的深度模型。*IEEE Access*，9:40511–40520。Wu, Z., Xiong, Y., Yu, S. X. 和 Lin, D.（2018）。通过非参数实例判别进行无监督特征学习。载于*IEEE*计算机视觉与模式识别会议论文集，第3733–3742页。

Xiao, H.、Rasul, K. 和 Vollgraf, R.（2017）。Fashion-MNIST：一种用于机器学习算法基准测试的新型图像数据集。

Zhang, R., Isola, P., and Efros, A. A. (2016年). 多彩的图像着色。载于欧洲计算机视觉会议，第649–666 页。Springer。

附录 A—表 1. SidAE 中不同 w 选项的下游分类准确率（五次运行的平均值，含标准误）。

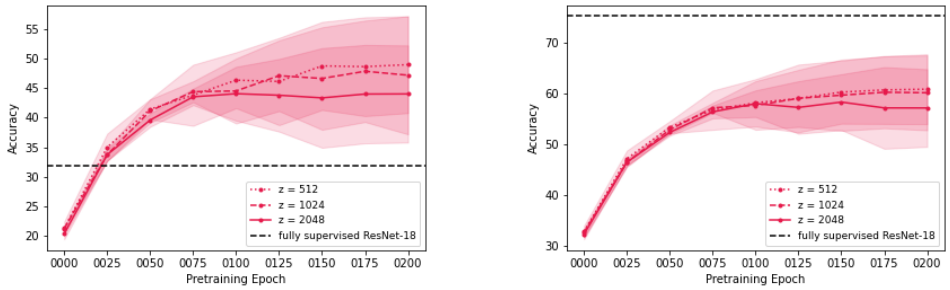
Data	预训练	$w = 0.125$	$w = 0.25$	$w = 0.5$	$w = 0.75$	$w = 0.875$
CIFAR-10	100%	**66.20** (0.31)	66.14 (0.18)	64.75 (0.27)	63.07 (0.25)	61.47 (0.23)
	1%	**55.61** (0.37)	55.34 (0.29)	51.67 (0.28)	49.21 (0.38)	47.18 (0.41)
Fashion-MNIST	100%	87.31 (0.09)	87.89 (0.09)	88.63 (0.12)	**88.78** (0.12)	88.71 (0.10)
	1%	81.49 (0.21)	81.78 (0.27)	**81.90** (0.16)	81.62 (0.22)	80.47 (0.29)
MNIST	100%	97.78 (0.04)	97.70 (0.05)	97.80 (0.06)	97.98 (0.08)	**98.10** (0.05)
	1%	**94.91** (0.20)	94.90 (0.18)	94.65 (0.22)	93.11 (0.13)	92.00 (0.25)
STL-10	STL-10	60.07 (0.18)	**61.18** (0.22)	60.44 (0.21)	56.92 (0.20)	53.77 (0.19)
	CIFAR-10	64.12 (0.40)	**65.42** (0.16)	65.18 (0.19)	63.33 (0.15)	61.59 (0.20)

附录 1

权重 w 对 SidAE 的影响。

在论文正文中，我们只讨论了 w 对 SidAE 在 CIFAR-10 上的影响。表1 还补充了该实验的信息，并列出了 Fashion-MNIST、MNIST 和 STL-10 上的结果。报告的准确率对应于 200 个预训练轮次，并取五次运行的平均值；括号中的数字表示标准误。根据所用数据集的不同，分类准确率可能在不同的 w 取值下保持稳定（例如，Fashion-MNIST、MNIST 100%），也可能存在一个明显更优的 w 取值，从而获得更好的结果（例如，CIFAR-10、STL-10）。

这些实验表明，在更具挑战性的数据集 CIFAR-10 和 STL-10 上，较低的 w 值往往表现更好。虽然这一规律在 MNIST 和 Fashion-MNIST 上并未出现，但 w 的最差性能与最佳性能之间的差距要小得多。这表明，较小的 w 值往往能带来性能更好的模型。



(a) 1% **(b)** 100%

附录 B—图 1. 在不同预训练阶段，使用不同潜在空间维度的 SimSiam 在下游任务（CIFAR-10）上的分类准确率 d_{hid}；下游训练使用 1%或 100%的训练数据。

附录 2

潜在空间的维度。

根据 Chen 和 He （2021年），SimSiam 得益于更高维度的潜在空间，并在 $d\text{hid} = 2048$ 处趋于饱和。然而，这一观察最初是在 ImageNet 上得出的。我们考察这一结论是否也适用于 CIFAR-10，并训练了 $d\text{hid} = 1024$ 和 $d\text{hid} = 512$ 的变体。同样，我们研究 $d\text{hid}$ 对其他自监督模型的影响。我们认为，对于该数据集，较小维度的潜在空间或许就足够了，因为它的维度低于 ImageNet，复杂度也较低。在调整潜在空间的维度时，Chen 和 He （2021年）也会相应调整预测器多层感知机隐藏层的维度，使隐藏层始终为特征空间大小的 14 。这样做是为了让预测器形成瓶颈结构。我们在此采用相同的原则。

SidAE 和去噪自编码器的结果表明，当 $d\text{hid}$ 取不同值时，性能差异很小（约在 1% 以内）。因此，将潜在空间的维度设为 2048、1024 或 512，不会对这些模型产生显著影响。准确率可能在 $d\text{hid}$ 值小得多时就已达到饱和。

SimSiam 在较小值下的平均准确率略高 d_{hid}。一种可能的解释是，对于维度较高的潜在空间，表征中会编码额外噪声，从而损害下游性能。然而，如图 1所示，均值的标准误较高，因此由标准误界定的区域有较大重叠。