
PersonViT: 用于人物再识别的基于大规模自监督学习的视觉Transformer

华中科技大学滨湖人工智能研究所，武汉
，中国 hubin@hust.edu.cn

王兴刚 华中科技大学EIC学院，武汉
，中国 xgwang@hust.edu.cn

刘文字* 华中科技大学信息与通信
学院，武汉，中国 liuw@hust.edu.cn

摘要

行人重识别 (ReID) 旨在检索非重叠摄像头图像中的相关个体，在公共安全领域具有广泛的应用。近年来，随着Vision Transformer (ViT)和自监督学习技术的发展，基于自监督预训练的行人ReID性能得到了极大的提升。行人ReID需要提取人体高度判别性的局部细粒度特征，而传统的ViT擅长提取上下文相关的全局特征，难以关注局部人体特征。为此，本文将近期涌现的Masked Image Modeling (MIM)自监督学习方法引入行人ReID，通过结合掩码图像建模和判别式对比学习，利用大规模无监督预训练有效提取高质量的全局和局部特征，然后进行行人ReID任务的有监督微调训练。这种基于带有掩码图像建模的ViT的行人特征提取方法 (PersonViT) 具有无监督、可扩展和泛化能力强的优点，克服了有监督行人ReID中标注困难的问题，并在公开的基准数据集上取得了最先进的结果，包括MSMT17、Market1501、DukeMTMC-reID和Occluded-Duke。PersonViT方法的代码和预训练模型已发布在<https://github.com/hustvl/PersonViT>，以促进行人ReID领域的进一步研究。

1 引言

人物再识别 (ReID) 旨在学习来自人像的视觉特征，以区分不同的个体身份。这是一个重要且具有挑战性的计算机视觉问题，需要克服严重的遮挡、外观变化、形状变化和视角变化。人物

*Corresponding Author.

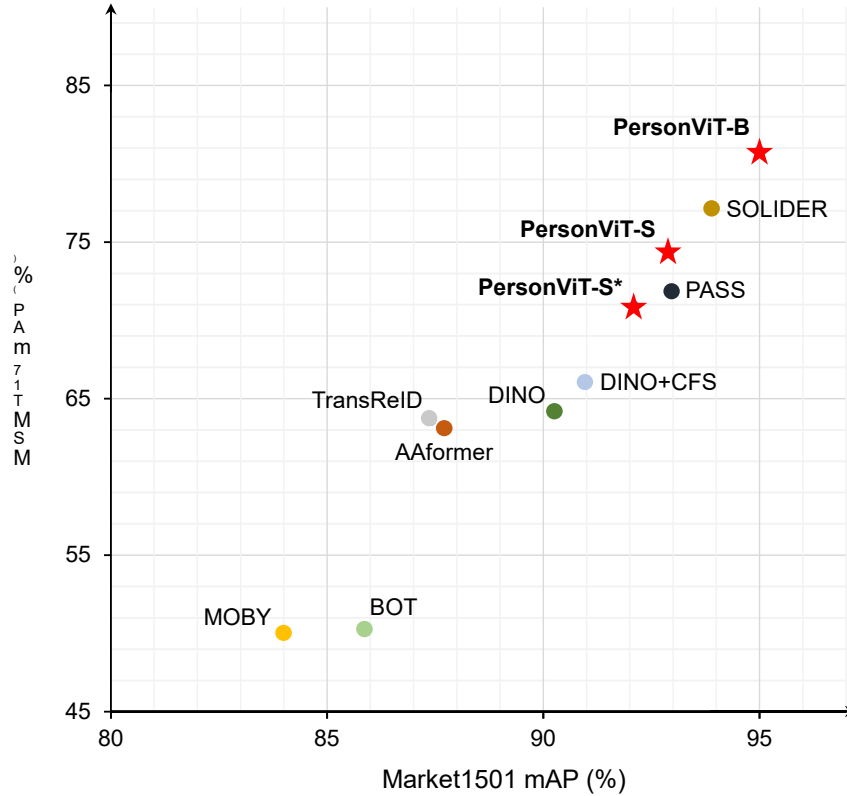


图1: MSMT17和Market1501的行人重识别性能。提出的PersonViT方法获得了最先进的结果，并显著优于以前的方法。

ReID技术能够在非接触式和非合作场景下有效地跨摄像头检索人员，广泛应用于公共安全、视频监控等领域，具有显著的应用价值。

已经有许多关于行人重识别的研究，其中大部分都涉及到基于骨干网络（例如ResNet[2]或ViT[3]）提取的特征的度量学习（例如，Triplet Loss[1]）。由于行人重识别数据集规模较小，现有方法中的大多数骨干网络都在ImageNet[4]上进行预训练，然后在行人重识别数据集上进行微调以获得更好的性能。然而，ImageNet和行人重识别数据集之间存在显著差异，ImageNet包含一千个类别的图像，而行人重识别只有一个人的类别。因此，在ImageNet上预训练的模型能够更好地提取类别级别的特征，但却难以有效地提取细粒度的个体特征，使其不太适合行人重识别问题[5]。

为了弥合预训练和微调数据集之间的差距以获得更好的ReID主干网络，[6]收集了一个名为LUPerson的未标记人员数据集，并首次证明了基于CNN的自监督预训练模型在ReID任务中表现良好。随后，TransReID-SSL[7]在LUPerson数据集上使用基于ViT的自监督学习算法进行了实验，发现使用DINO[8]算法预训练的模型对行人ReID最有效。然而，这些自监督学习算法都利用了从同一图像的不同数据增强生成的局部和全局图像之间的分类一致性来进行对比自监督学习。它们能够比较好地捕捉ImageNet等多类别数据集中的类间差异，但在单类别ReID数据集中会丢失细粒度的差异。例如，不同人体局部图像可能非常相似，可以属于同一类，但DINO的自监督对比会将其分离到不同的类。类似地，同一人的不同角度的局部图像应该属于同一类，但DINO的对比会将其分离。因此，PASS方法[5]提出了一种专门为行人ReID任务设计的自监督预训练算法。基于DINO，它将

将人物图像根据人体结构从上到下分成几个 (L) 块。它在块与块之间以及块与整幅图像之间执行对比自学习，从而增强预训练模型表达局部特征的能力，并在当时取得了新的最先进水平。然而，PASS 有两个问题：1) 与之前的基于块的 ReID 方法（如 PCB [9]）一样，它存在对齐依赖性问题，即在复杂的背景、不均匀的对齐甚至遮挡或不完整的人像中，块划分方法过于机械，容易导致误分类；2) 较少的块数仍然难以充分表达细粒度的局部特征，即在 PASS 方法中，最佳性能实验对应于仅三个的划分计数，这在粒度上显然不足。

最近，受掩码语言建模 (MLM)，即 BERT[10] 在自然语言处理 (NLP) 领域取得巨大成功的启发，掩码图像建模 (MIM) 技术，如 BEiT[11]、MAE[12] 和 SimMIM[13] 也通过自监督学习在自然图像的图像分类、检测和分割任务中取得了突破。MIM 采用局部像素的随机掩码进行重建学习，强化了局部细粒度特征的学习和提取，无需人工图像分割。这有效地弥补了 PASS 算法的局限性，表明引入 MIM 应该会提高行人 ReID 的准确性。

本文中，我们结合了基于 DINO 的 MIM 自监督学习模块，在 LUPerson [6] 数据集上进行了大规模的无监督预训练，随后在四个数据集上进行了监督式 ReID 微调：MSMT17 [14]、Market 1501 [15]、DukeMTMC-reID [16] 和 Occluded-Duke [17]。我们将这种利用掩码图像建模和 DINO 对比学习进行基于普通 ViT 的大规模无监督预训练并在下游数据集上进行微调的方法称为 PersonViT 方法。实验结果表明，PersonViT 取得了最先进的结果，尤其是在具有挑战性的 Occluded-Duke 数据集上表现出色。此外，对 MSMT17 上预训练模型的可视化分析表明，PersonViT 可以自动发现人体关键部位、服装图案以及局部身体部位之间的对应关系，无需任何标注。

本文的创新之处体现在以下几个方面：1) 为了增强算法对人体图像对齐的鲁棒性，并更好地提取人体局部细粒度特征，我们率先在行人重识别 (Person ReID) 的无监督特征学习中利用了掩码图像建模技术。这有效地解决了在存在大量遮挡和错位的人体图像中获取局部细粒度视觉特征的挑战。2) 我们提出了一种基于 vanilla ViT 的高效、大规模自监督人体特征学习方法，称为 PersonViT。3) 我们的方法在几个主流的行人重识别数据集上取得了最先进的结果。

这项研究的主要意义在于，它通过高效且准确的自监督特征学习实现了业界领先的行人重识别 (ReID) 结果，极大地解决了行人重识别中标记训练数据不足的难题。该方法具有高度可扩展性，为大规模 ReID 技术的实际应用提供了强有力的技术支撑。实验结果表明，我们的方法达到了业界领先的精度，显著优于以往的方法。

2 相关工作

2.1 自监督学习

对比学习 自监督学习 (SSL) 方法旨在从大规模未标记数据中学习判别特征[18]。近年来，对比学习方法在计算机视觉领域蓬勃发展[19; 20; 21; 22; 23; 24; 8]，显著缩小了与监督预训练的差距。MOCO[19] 首先提出了动量对比，其中样本的一对增强被视为正样本，而其他样本及其增强被视为负样本，从而进行无监督对比学习训练。此后出现了一系列改进，例如 MOCOv2[20]、MOCOv3[21] 和 SimCLR[22]。BYOL[23] 提出了一种新的对比学习范式，它使用两个网络模型来预测同一样本的不同增强图像的表示，消除了对大量负样本的依赖。DINO[8] 是 BYOL 的改进版本。它在目标模型参数的动量平滑更新过程中引入了中心化和锐化操作，有效地防止了模型崩溃并提高了算法的稳定性。此外，DINO 结合了

大量的数据增强，特别是多次局部裁剪增强，增强了模型通过局部和全局图像之间的大规模对比学习来学习局部特征的能力。

DINO框架 DINO [8] 是首个自蒸馏学习框架。在深入研究DINO之前，让我们简要介绍一下Transformer背后的基本原理。将Transformer视为编码器，它将图像 $I \in \mathbb{R}^{h \times w \times c}$ （其中 $h \times w$ 对应图像分辨率， c 对应图像通道数，通常为3，表示RGB）转换为目标特征向量。与NLP中的标记化过程类似，图像被划分成连续的实体。Transformer通过图像的块投影标记化来启动这个过程。如果将每个 $p \times p$ 像素大小的块划分为 $n = hw/p^2$ 个图像块，则每个片段可以表示为一个标记，类似于NLP中的单个单词。结合可学习的 cls_token ，这些标记可以表示图像标记化结果为 X ，如公式(1)所示。这里，“;”符号象征着堆叠标记之间的连接。通过ViT网络编码后，可以根据公式(2)获得特征向量的类似表示。与传统的ViT主干网络相比，DINO编码器集成了一个头部模块，该模块通过多层感知网络(MLP)将ViT输出 $z^{[cls]}$ 映射到目标向量空间 $y^{[cls]}$ ，如公式(3)所示。为了计算最终的损失函数，记为 L_{dino} 以区别于后续的损失函数，DINO使用了两个同构编码器网络——学生网络和教师网络，它们各自的输出分别记为 $Y^{[S]}$ 和 $Y^{[T]}$ ，如公式(4)所示。梯度反向传播的范数更新学生网络参数，而教师网络参数则随着学生网络参数的指数移动平均(EMA)一起演变，计算方法为 $\theta_t = \lambda\theta_t + (1 - \lambda)\theta_s$ 。有趣的是，DINO教师网络输入两个全局放大的数据视图，而学生网络则调用相应全局视图以及多个局部视图。因此， L_{dino} 包含全局和局部-全局对比计算，从而增强了对局部特征的辨别能力。

$$\begin{aligned} X &= (x^{[cls]}; x^{[patches]}) \\ &= (x^{[cls]}; x_1; \dots; x_n) \in \mathbb{R}^{(n+1) \times d} \end{aligned} \quad (1)$$

$$\begin{aligned} Z &= (z^{[cls]}; z^{[patches]}) \\ &= (z^{[cls]}; z_1; \dots; z_n) \in \mathbb{R}^{(n+1) \times d} \end{aligned} \quad (2)$$

$$Y = \text{MLP}(z^{[cls]}) = y^{[cls]} \in \mathbb{R}^{1 \times d_{output}} \quad (3)$$

$$L_{dino} = -\text{Softmax}(Y^{[T]}) \log(\text{Softmax}(Y^{[S]})) \quad (4)$$

从2018年开始，自然语言处理（NLP）领域在掩码语言模型（MLM）方面取得了显著成功，例如BERT[10]和GPT[25]等模型。近年来，在视觉领域也出现了一些掩码建模尝试，代表性工作包括[26; 13; 12; 27; 28]。MST[26]是第一个引入掩码图像建模的模型，它通过在DINO框架中添加掩码预测训练来增强DINO的性能。紧随其后的是BEiT[11]，它首先使用离散变分自动编码器(dVAE)将图像块离散化并映射到相应的视觉token。然后，它掩盖被遮挡的视觉token以实现显著的学习结果；因此，BEiT也作为一个两阶段训练算法。MAE[12]是一个端到端的MIM自动编码器。它使用编码器仅编码可见的图像块，并使用相对轻量级的解码器来恢复被掩盖的像素。有趣的是，即使大部分输入图像（例如，75%）被掩盖，它也被证明可以产生出色的自监督学习结果，从而显著加快了预训练过程。由于MAE的自监督学习是恢复像素，其编码器和解码器都学习某些特征表达。此外，MAE恢复的是最原始的低级像素特征，因此编码器学习的特征抽象程度不受控制。MAE的ImageNet实验结果中k-NN和线性探测基准的相对较低的准确率表明，编码器没有充分提取区分性特征抽象。

2.2 基于部件的行人重识别

传统上，特征表达学习主要采用IDE模型[29]。这种方法涉及全局特征提取，然后进行多分类训练，其中每个人都被视为一个单独的类别。逐渐地，趋势发展到使用ResNet50[2]提取全局特征向量，然后进行度量学习，例如Triplet Loss[1]。然而，提取详细的局部特征仍然是全局特征表示学习的一个挑战。因此，出现了两种主要的优化形式来解决这个问题。第一种涉及多尺度融合表示学习，如[30]等文献所示。第二种包括整合注意力机制以加强局部表示学习，例如协调空间和通道注意力[31]，以及使用分割注意力机制[32]来衰减表示学习中的背景。为了更好地捕捉人像的局部、细粒度特征，并提高算法在处理对齐问题和遮挡方面的鲁棒性，随后提出了基于局部特征的行人重识别方法。早期提出的局部表示学习方法，例如[33]、[34]和PCB[9]中发现的方法，将人像分割成多个水平条带以提取局部特征。MGN[35]使用不同粒度的条带和这些条带之间的重叠来增强鲁棒性。TransReID[36]作为第一个基于Transformer的行人重识别算法，也通过重新排列和重新分组Transformer的输入局部图像来进行局部表示学习。这些方法突出了通过局部表示学习提取的细微特征显著增强了行人重识别的精度。

2.3 自监督行人重识别

尽管与自监督预训练中行人重识别相关的研究出现较晚，数量也较少，但由于避免了预训练阶段的高成本标注，它展现出巨大的潜力。TransReID-SSL[7]首次将当时主流的自监督学习算法在行人重识别问题上的性能进行了比较。在比较中，DINO[8]展现出比其他算法显著的优势。研究人员还研究了预训练数据大小对最终精度的影响，并提出了一种有效的预训练数据筛选方法。PASS[5]是首个面向行人重识别任务的自监督预训练算法，它对人体图像进行分割。在DINO的基础上，PASS引入了局部到局部和局部到全局的对比学习。它通过为不同的局部使用单独的 $\{v^*\}$ 来区分表示，并且除了全局对比分类特征空间[CLS]外，还引入了额外的分割分类特征空间[PART]，从而增强了分割的细节表示能力，提高了预训练模型的局部特征提取能力。该方法在当时有监督的行人重识别和跨域行人重识别中都设定了行业领先的基准，从而验证了其有效性。

2.4 行人重识别研究综述

回顾过去几年行人重识别（ReID）的发展历程，我们可以总结出三大趋势。

首先，特征提取主干网络已经从传统的卷积神经网络发展到Transformer（ViT），这主要是因为ViT在表达全局上下文特征方面的先进能力以及基于ViT的自监督学习的显著发展。

其次，最初基于ImageNet分类的预训练方法已经转向使用大量人类数据进行的自监督模型。这种转变主要是因为ReID挑战涉及区分不同个体之间的细微差异，这与ImageNet分类任务中种类繁多且差异很大的类别不同。由于测试个体通常不存在于训练子集中，因此泛化能力成为驱动识别精度的关键组成部分。然而，由于数据采集和标注方面的挑战导致公共监督训练集的稀缺，无法满足算法泛化能力的要求。

最后，鉴于不同个体之间细微的差别——与ImageNet中的类别差异相比，范围要小得多——有效提取局部细粒度特征对于提高ReID精度至关重要。

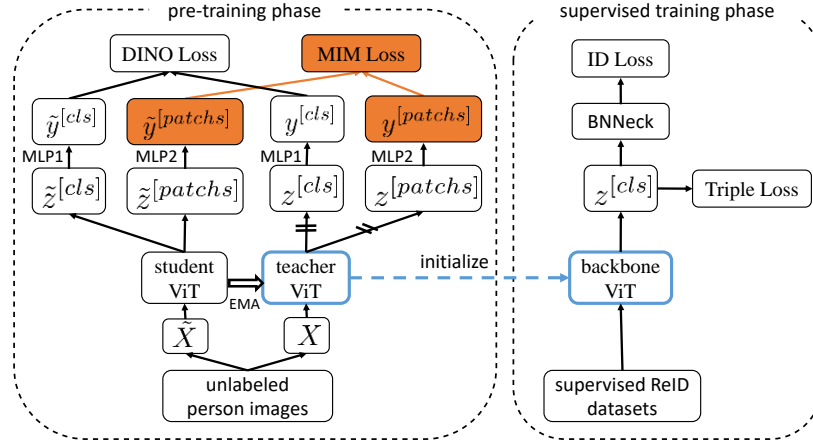


图2: PersonViT框架概述。

最新的方法PASS [5]成功地与这三种趋势保持一致，并在其诞生之初就取得了行业领先的业绩。然而，它也突出了两个关键的局限性。首先，它依赖于人工设计的分割分区，限制了其对人体图像对齐的鲁棒性。其次，有限数量的分区使得充分表达局部细粒度特征变得具有挑战性，从而限制了预训练模型提取此类特征的能力。

3 方法

针对PASS算法中确定的局限性，一种可能的解决方案是在ViT模型中为每个图像块输入引入一个分类特征空间，类似于自监督预训练环境下的[CLS]特征空间。然而，应用传统基于数据增强的对比学习方法可能会带来挑战。鉴于每个图像块的小尺寸，生成不同数据增强的过程很容易导致局部图像块内的错位，从而限制了可比性，并且无法达到预期的自监督效果。PASS方法的实验结果证明了这一点，结果表明增加局部片段的数量或减小分区的大小并不会因此提高精度。

受掩码图像建模概念的启发，一种可能的解决方案是将同一图像两次馈送到系统中。一个输入是完整的图像，另一个是带有某些被掩码的图像块。来自这两个输入的相应输出特征向量，表示为[PATCH]并在等式(6)中表示为 $\tilde{y}^{[patches]}$ ，将构成比较的基础。这种方法允许在最细粒度的级别上进行特征表示，对应于图像块的大小，并且可以提高预训练模型提取细粒度局部特征的能力。

本文提出的整体算法框架（如图2所示）将一个基于patch粒度的对比损失模块（称为MIM Loss模块）融入到DINO预训练算法中。预训练使用ViT-S和ViT-B作为主干网络，并利用最大的公开数据集LUPerson [6]。之后，在四个主流的公共人物ReID数据集上进行监督微调训练：MSMT17 [14]、Market1501 [15]、DukeMTMC-reID [16]和Occluded-Duke [17]。预训练和微调的具体步骤详述如下。

3.1 自监督预训练MIM损失函数的融入

受BEiT [11]的掩码图像建模范式的启发，我们的方法包括图像的随机块状掩码，类似于引入可学习的标记变量 $x^{[mask]}$ ，类似于 $x^{[cls]}$ 。因此，掩码图像可以描述为 $\tilde{x} = (x^{[cls]}; \tilde{x}_1; \dots; \tilde{x}_n)$

其中 \tilde{x}_i 详见式(5)。

$$\begin{aligned}\tilde{X} &= (x^{[cls]}; \tilde{x}_1; \dots; \tilde{x}_n), \\ \tilde{x}_i &= (1 - m_i)x_i + m_i x_i^{[mask]}, i \in 1, \dots, n\end{aligned}\quad (5)$$

在这个等式中, $m_i \in 0, 1$ 表示随机图像块掩码, 其中 1 表示已掩码, 0 表示未掩码。掩码图像 \tilde{X} 通过 ViT 编码器处理后, 得到 \tilde{Z} , 如公式 (2) 所示。

$$\begin{aligned}\tilde{Y} &= (\text{MLP1}(\tilde{z}^{[cls]}); \text{MLP2}(\tilde{z}^{[patches]})) \\ &= (\tilde{y}^{[cls]}; \tilde{y}^{[patches]}) \in \mathbb{R}^{(1+n) \times d_{output}}\end{aligned}\quad (6)$$

$$\begin{aligned}L_{mim} &= \sum_{i=1}^n m_i \cdot P(y_i^{[patches][T]}) \log(P(\tilde{y}_i^{[patches][S]})), \\ P(x) &= \text{Softmax}(x)\end{aligned}\quad (7)$$

$$L = \lambda_1 L_{dino} + \lambda_2 L_{mim} \quad (8)$$

与DINO截然不同, 我们也为 $\tilde{z}^{[patches]}$ 实现了MLP网络变换, 从而将其投影到 d_{output} 维向量空间以产生 $\tilde{y}^{[patches]}$, 然后执行重建损失计算。目标向量 $\tilde{Y} = (\tilde{y}^{[cls]}; \tilde{y}_1; \dots; \tilde{y}_n)$ 可以如公式(6)所示表达。为了对遮罩图像块的重建向量进行基准测试, 仅对学生网络的双全局视图输入应用遮罩, 而教师网络的输入包含完整的全局视图以供参考。具体的遮罩重建损失函数(缩写为 L_{mim}) 在公式(7)中解释。这与公式(4)中所示的DINO损失函数 L_{dino} 结合, 形成了公式(8)中概述的最终预训练损失函数 L 。这里, L 被认为是 L_{dino} 和 L_{mim} 的加权和, 其中 $\lambda_1 = \lambda_2 = \text{默认为} 1$ 。

3.2 有监督微调

在这个阶段, 先前自监督学习产生的预训练模型将针对行人重识别 (ReID) 的特定任务进行微调, 生成该任务的最终模型。随后, 我们进行测试以评估模型的 ReID 准确性。为了确保对自监督预训练模型有效性的公正比较, 我们在这个阶段继续使用 BOT [37] 框架。这种方法与 TransReID-SSL 基线中采用的方法一致, 该基线包括使用标准 ViT 网络 (即 ViT-S/16 和 ViT-B/16) 作为主网络, 直接实现 $z^{[cls]}$ 进行特征聚合, 采用 Triplet Loss 进行度量学习, 选择交叉熵损失作为 ID Loss, 并在度量学习和 ID Loss 之间插入 BNNeck 模块。对于主要标准 ViT 网络, 我们使用自监督学习阶段教师网络的预训练模型作为起始参考点, 开始进行行人重识别训练的微调。

4 个实验

4.1 数据集

用于自监督预训练的主要数据集是LUPerson [6], 其中包含418万张未标记的人脸图像。该数据集的大小是ImageNet的四倍, 因此预训练所需的计算资源更大, 相同的计算资源将导致训练时间是ImageNet数据集的四倍。为了彻底验证预训练模型提取更细粒度局部特征的能力, 本研究在四个主流行人重识别数据集MSMT17 [14]、Market1501 [15]、DukeMTMC-reID [16]和Occluded-Duke [17]上进行了监督训练, 观察到两个关键指标mAP (平均平均精度) 和Rank-1的提升。四个数据集的详细信息如表1所示。Occluded-Duke基于DukeMTMC-ReID生成, 增加了遮挡情况下ReID的难度。

表1: 一些常用的行人重识别数据集的统计数据。

Dataset	Time	#ID	#image	#cam
Maket-1501	2015	1501	32668	6
DukeMTMC	2017	1404	36411	8
MSMT17	2018	4101	126441	15
Occluded-Duke	2019	1404	36411	8

4.2 实现细节

4.2.1 自监督预训练阶段

由于LUPerson数据集规模庞大, 为缩短实验时间, 我们使用 $8 \times 8 \times A100$ GPU进行了大批量训练以加快实验进程。尽管如此, 为了证明小批量训练的有效性, 我们也使用 $4 \times RTX3090$ GPU进行了基础实验。为了减少预训练阶段的计算量, 训练周期(轮数)设置为300。与DINO和PASS类似, 教师网络接受图像尺寸为 256×128 的输入, 而学生网络被设计为处理尺寸为 256×128 的全局视图和仅6个尺寸为 96×64 的局部视图。考虑到预训练耗时较长, 我们的实验只训练了两个基本的网络模型——ViT-S/16和ViT-B/16。对于小批量实验, 预训练学习率统一采用 $lr = 0.0005 * batch_size / 256$ 。对于大批量实验, 尽管实验时长显著缩短, 但训练过程相当不稳定。因此, 我们根据具体的批量大小调整学习率, 确保不超过0.002, 直到达到稳定的训练收敛。具体的参数和训练日志将与代码一起公开发布。

4.2.2 监督训练阶段

基于TransReID-SSL[7]的实验设置, 监督微调过程未添加任何其他优化项, 仅使用传统的ViT-S/16和ViT-B/16 Transformer网络作为监督训练的骨干网络。监督训练统一采用随机梯度下降作为学习算法, 学习率设置为 $lr = 0.0004 * batch_size / 64$ 。批量大小由 $4 * 16 = 64$ 组成, 这意味着每个批次包含16个不同的个体, 每个人有4张图像。沿用PASS[5]的策略, 前20个周期作为预热阶段, Triplet Loss的 α 参数设置为0.3。

4.3 实验结果

表2显示了不同算法实验结果的比较(浅蓝色背景表示最高精度)。为了公平比较主干网络预训练参数的作用, TransReID-表示没有TransReID [36]中的SIE和JPM模块时的ReID精度。此外, 最后两行表示使用 $64 \times A100$ 大批量大小预训练的模型的精度。带*号的行表示在 $4 \times RTX3090$ 上使用小批量大小预训练的模型的精度。AAformer [38]和TransReID-是在ImageNet-21K数据集上进行分类预训练的模型, 该数据集比LUPerson数据集更大, 并且具有分类标签。可以看出, 基于LUPerson的自监督预训练模型的ReID精度远高于基于ImageNet预训练模型的精度。

与DINO+CFS[7]和PASS[5]相比, 本文提出的基于PersonViT的LUPerson预训练模型, 特别是大批量预训练模型, 在精度上取得了显著的提升, 大大超过了基于图像分割自监督学习的PART算法预训练模型。

4.3.1 超参数消融实验

DINO研究[8]证明了多作物技术对性能的显著影响。考虑到行人图像的独特特征, 研究全局和局部裁剪中各种超参数的影响就变得非常重要。行人重识别预训练

表2：与最先进方法的比较。

Methods	Backbone	MSMT17		Market1501		DukeMSMT		Occluded-Duke	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
BOT [37] CVPRW2019	R50	50.2	74.1	85.9	94.5	-	-	-	-
AAformer [38] Arxiv	ViT-B/16 \uparrow 384	63.2	83.6	87.7	95.4	-	-	-	-
TransReID ⁻ [36] ICCV2021	ViT-B/16	63.6	82.5	87.4	94.6	-	-	-	-
MOBY [24] Arxiv	ViT-S/16	50.0	73.2	84.0	92.9	-	-	-	-
DINO [8] ICCV2021	ViT-S/16	64.2	83.4	90.3	95.4	-	-	-	-
DINO+CFS [7] Arxiv	ViT-S/16	66.1	84.6	91.0	96.0	-	-	-	-
PASS [5] ECCV2022	ViT-S/16	69.1	86.5	92.2	96.3	82.5	90.7	60.2	70.4
PASS [5] ECCV2022	ViT-B/16	71.8	88.2	93.0	96.8	84.7	92.5	64.3	74.0
SOLIDER [39] CVPR2023	Swin-B	77.1	90.7	93.9	96.9	-	-	-	-
PersonMAE [40] TMM2024	ViT-S/16	75.2	89.1	92.5	96.7	-	-	65.2	72.0
PersonMAE [40] TMM2024	ViT-B/16	79.8	91.4	93.6	97.1	-	-	69.5	76
PersonViT(ours)*	ViT-S/16	70.9	87.3	92.1	96.4	83.6	91.6	61.5	70.8
PersonViT(ours)	ViT-S/16	74.3	89.2	92.9	96.8	84.7	91.9	65.2	73.2
PersonViT(ours)	ViT-B/16	80.8	92.0	95.0	97.6	88.1	93.8	72.2	79.8

LUPerson数据集与ImageNet数据集在两个主要方面有所不同：1) 人像图像通常分辨率较低，导致局部细节不明显；2) 人体裁剪图像通常呈矩形，因此在预训练过程中，图像被缩放到256x128作为网络输入。这与ImageNet使用224x224像素的正方形图像作为输入的做法形成对比。考虑到这些差异，消融实验主要集中在研究全局和局部裁剪之间尺寸分布和纵横比差异的影响。为了节省实验时间，我们选择LUPerson子数据集的排名前3%的数据（根据Trans ReID-SSL [7]中的CFS排名）作为我们的预训练数据集。此外，我们使用ImageNet预训练模型作为初始参数。实验结果如表3所示。“裁剪率范围”表示局部裁剪相对于全局裁剪的尺寸范围。例如，第一个条目“0.1 0.6, 0.6 1.0”意味着学生网络的局部裁剪输入比例在一个[0.1, 0.6]的随机区间内，而全局裁剪的比例在一个[0.6, 1.0]的随机区间内。

消融实验的结果表明，在自蒸馏过程中局部和全局裁剪的重叠更有利于行人重识别。此外，在随机抖动过程中保持纵横比（高：宽）为2:1（默认参数为1:1）被证明更有益。此参数调整使平均精度均值（mAP）得分显著提高了两个百分点。

表3：多作物增强的消融研究。

Input size	Crop rate range	Aspect(width/height)	Crop size	mAP	Rank-1
192x96	0.1~0.6, 0.6~1.0	3/4~4/3	96x64	60.2	80.7
192x96	0.1~0.8, 0.8~1.0	3/4~4/3	96x64	59.9	80.8
192x96	0.1~0.8, 0.4~1.0	3/4~4/3	96x64	61.3	81.9
256x128	0.1~0.8, 0.4~1.0	3/4~4/3	96x64	62.5	82.4
256x128	0.1~0.8, 0.4~1.0	3/8~2/3	96x64	64.7	83.5
256x128	0.1~0.8, 0.4~1.0	3/8~2/3	96x48	64.2	83.2
256x128	0.1~0.8, 0.4~1.0	3/8~2/3	128x64	64.4	83.3

4.3.2 MIM损失函数消融实验

为了验证掩码图像建模的作用，我们进行了一个简单的消融实验。具体来说，在其他完全相同的实验参数下，我们比较了公式（8）中 $\lambda_2 = 1$ 和 $\lambda_2 = 0$ 的设置。实验结果如表4所示。结果表明，与DINO算法[8]相比，引入MIM损失函数带来了显著的改进。MSMT7的mAP提升达到了令人印象深刻的6.4，远超PASS方法[5]与DINO相比的提升（3.0）。图4中局部特征聚类的可视化分析更清晰地表明，我们的方法提取了更丰富的局部细粒度人体特征，从而验证了我们的理论假设。

表4: 关于MIM损失的消融研究。

Methods	Backbone	MSMT7		Market501		DukeMSMT		Occluded-Duke	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
PersonViT	ViT-S/16	74.3	89.2	92.9	96.8	84.7	91.9	65.2	73.2
PersonViT w/o. MIM	ViT-S/16	67.9	85.6	91.4	96.3	82.1	90.6	58.5	66.9

4.3.3 预训练中的过拟合问题

在实验过程中,我们发现最佳模型精度并非一定在 $epoch = 300$ (对应预训练过程结束)时达到。通过监控自监督预训练阶段的指标,我们注意到在 $[200,300]$ 轮次范围内出现了过拟合现象。为了进一步研究自监督学习中的这种过拟合问题,我们进行了监督训练,并在两种情况下测试了每20轮保存的模型的精度变化:从零开始预训练和使用ImageNet自监督预训练模型初始化参数。这些结果如图3所示,“w/ pt”表示使用完整的LUPerson数据集从ImageNet预训练模型初始化开始训练,“w/o pt”表示从零开始预训练。从图3可以得出两个结论:1) 使用ImageNet预训练模型初始化在早期阶段具有显著优势,但这种优势在大约 $epoch = 160$ 时被超越,并且在 $epoch \in [160,300]$ 范围内的后期轮次中,从零开始的预训练显示出轻微优势。ImageNet预训练模型已经具有良好的行人判别能力,因此早期优势是可以预期的。2) 对于后期训练阶段 $epoch \in [160,300]$, mAP和Rank-1精度指标都表现出先增加后下降的过拟合趋势。这在MSMT17数据集中尤为明显,最佳性能在 $epoch = 200$ 时达到,而在其他数据集中最佳点出现在 $epoch = 240$ 。因此,本文提出的实验结果大部分使用了预训练 $epoch = 240$ 时的模型监督训练精度。由于训练数据量较小(仅占总数的3%),因此没有明显的过拟合问题,因此表3中的结果是在 $epoch = 300$ 时获得的。

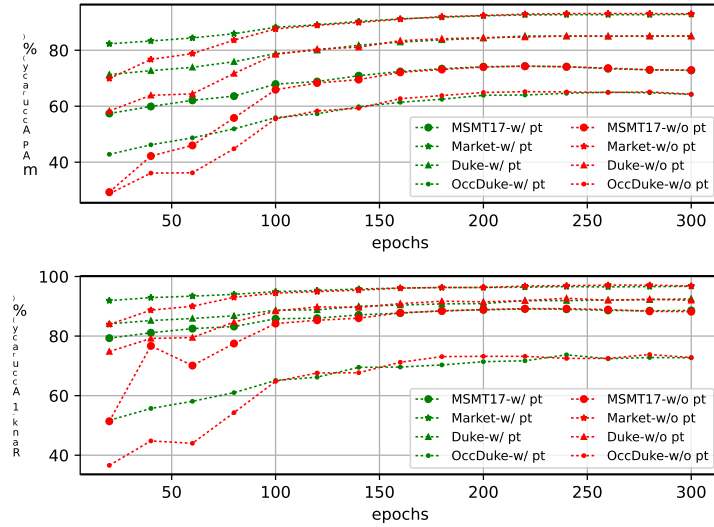


图3: S 在不同的预训练模型上, Person Re-ID 的监督精度 轮次

4.3.4 预训练数据量的影响

为了进一步探索预训练数据规模对监督训练结果的影响,我们设计了使用LUPerson数据集的不同样本进行的实验。结果如表5所示。“带PT”表示是否使用了ImageNet预训练模型来初始化参数。“10%”表示从LUPerson中随机选择10%的样本,“10%+CFS”表示根据TransReID-SSL[7]中的CFS排名选择前10%的样本。

表5的分析使我们得出以下几个结论：1) 通过我们的自监督学习算法进行人物ReID的准确率随着预训练数据的规模而增加，这意味着我们的自监督特征学习方法更适用于大规模的未标记数据。此外，数据规模越大，ReID准确率越高；2) 当数据规模相对较小时，用ImageNet预训练参数进行初始化会带来显著的改进；3) 根据TransReID-SSL[7]中的CFS排名选择top数据优于随机选择，表明CFS与最终准确率之间存在一定的相关性。然而，这种选择方法并没有达到TransReID-SSL论文中所述的结果，其中CFS (50%)超过了完整数据集的准确率。

表5: 预训练数据量消融研究 集合。

Backbone	w/ PT.	Scale	mAP	Rank-1
ViT-S/16	✗	3%	54.2	75.7
ViT-S/16	✗	10%	65.8	84.0
ViT-S/16	✗	25%	66.1	83.9
ViT-S/16	✗	50%	70.5	87.1
ViT-S/16	✗	100%	74.3	89.2
ViT-S/16	✓	10%	66.7	85.5
ViT-S/16	✓	10%+CFS [7]	67.5	84.9
ViT-S/16	✓	50%+CFS [7]	73.8	88.6

4.4 可视化分析

为了更好地理解预训练模型学习到的人脸特征，我们对MSMT17 [14]数据集上的预训练ViT-S/16模型进行了可视化分析。该模型基于LUPerson进行预训练，我们分析了补丁标记的模式布局，可视化了自注意力图，并探索了特征相关性。

补丁标记的模式布局。图4以可视化的方式展示了PersonViT输出特征 $y^{[patches]}$ 的无监督聚类结果。每个子图代表一个聚类，红点标记属于该聚类的图像块的位置。图的左半部分描绘了PersonViT模型提取的特征 $y^{[patches]}$ 可以将人体的关键部位，例如面部、脚部和膝关节等聚类在一起。图的右半部分展示了特征 $y^{[patches]}$ 可以将人体颈部、背包及其肩带聚类在一起。这些结果证实了PersonViT预训练模型可以自主提取人体关键部位及其附近辅助物体的细粒度特征。这种自动提取能力极大地提高了人物ReID的准确性。例如，颈部位置的自动感知可以增强个人饰品（如项链）的可区分性；面部和头部自动定位可以提高发型的识别率；脚部的自动检测可以显著提高鞋子的可识别性。

自注意力可视化。如图5所示，自注意力视图证实了预训练模型能够有效地提取人体轮廓，即使在人体遮挡、碎片化或错位等复杂情况下，也能通过忽略无关背景的影响来实现。这种能力解释了我们的算法在Occluded-Duke数据集[17]上的显著改进。

特征相关性分析。图6所示的特征相关性评估表明，我们的预训练模型能够熟练地捕捉同一身份不同图像之间特征的关系，即使在大幅变形（如转身或骑车）的情况下也能做到。

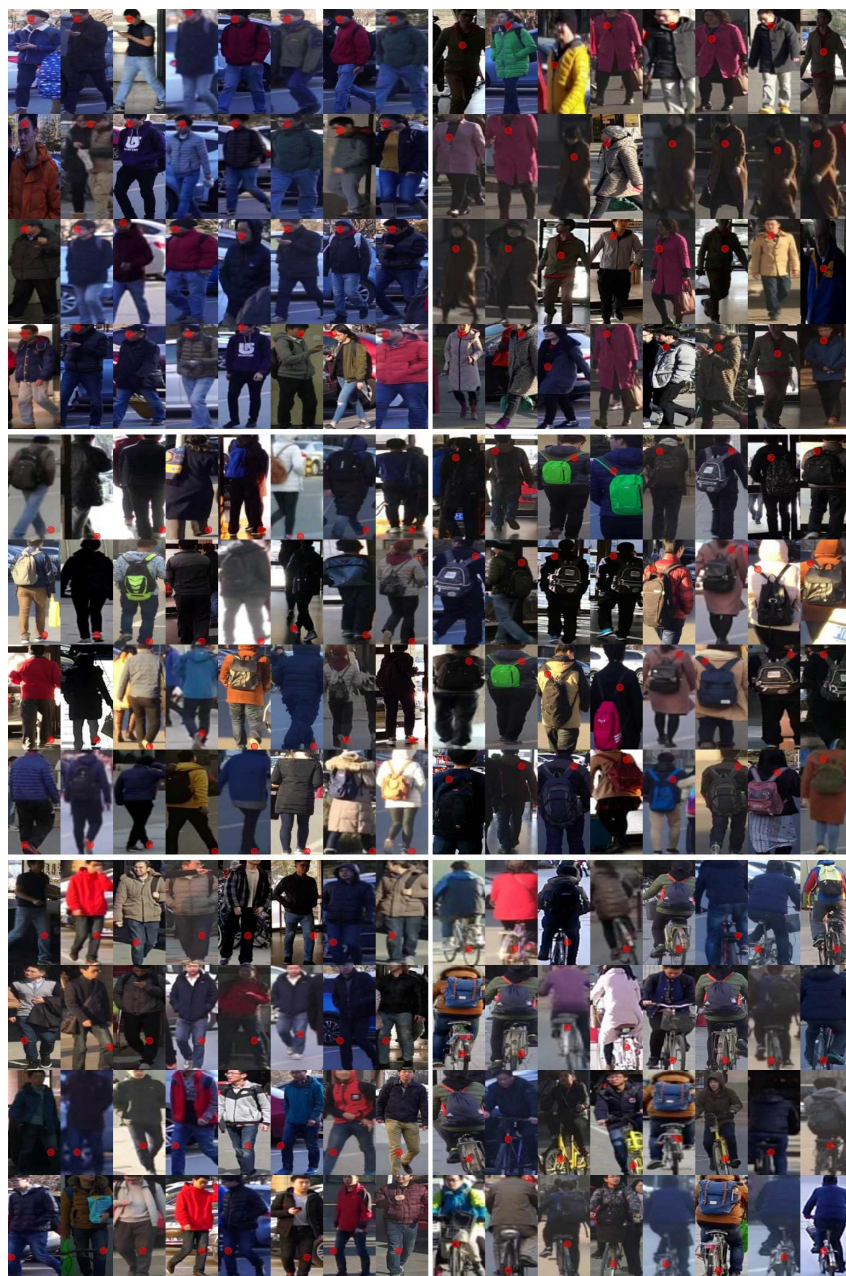


图4：补丁标记簇的模式布局可视化。

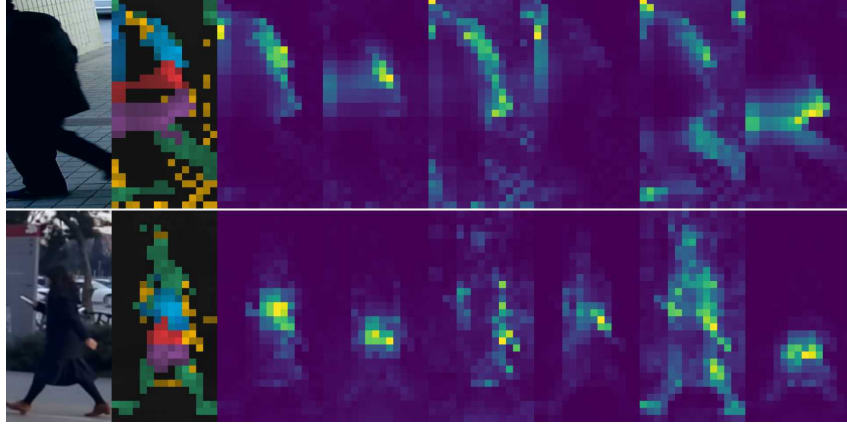


图5：复杂背景下自注意力图的可视化。



图6：一人两张图像之间稀疏对应关系的可视化。

5 结论

本文提出了一种大规模自监督行人预训练方法PersonViT，该方法在对比学习的基础上引入了掩码图像建模。通过在大规模未标注人体图像数据集上进行预训练，PersonViT能够有效提取丰富、高判别性、局部细粒度的人体特征，并在行人重识别任务中取得了显著的精度提升。实验表明，即使主干网络采用较小的vanilla ViT-S/16模型，随着预训练数据集规模的增大，最终识别精度也能进一步提高。鉴于在真实场景中获取大量未标注人体预训练数据的成本相对较低，该方法具有广泛应用于实践的潜力，能够提升不同场景下重识别算法的有效性。

然而，与其他自监督预训练算法一样，PersonViT也面临着预训练计算开销过大的问题，在计算资源有限的情况下，预训练周期会更长。尽管用于行人重识别算法的预训练模型更新频率低于监督训练，但在将PersonViT广泛应用于现实生活中大量的未标记预训练数据时，预训练效率仍然是一个关键问题。这个问题可以通过几种方法解决：1) 选择更轻量级的ViT模型作为主干网络；2) 采用类似于MAE[12]的掩码方法，其中掩码

丢弃部分token（MAE实验表明可以丢弃75%的token），这些token不参与主干网络的预训练，从而提高预训练效率；3）研究更高效的预训练数据过滤方法，在不降低精度的情况下减少预训练数据集的大小，从而提高预训练效率；4）增量预训练相关研究，即保留预训练模型的原始特征，对新增的预训练数据进行在线增量预训练学习，从而更新和进化预训练模型。

参考文献

- [1] A. Hermans, L. Beyer和B. Leibe, “论证用于人物再识别的三元组损失”, *arXiv preprint arXiv:1703.07737*, 2017年。
- [2] K. He, X. Zhang, S. Ren和J. Sun, “用于图像识别的深度残差学习”, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 第770-778页, 2016年。
- [3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, “一张图片值16x16个单词: 用于大规模图像识别的Transformer”, *arXiv preprint arXiv:2010.11929*, 2020年。
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li和L. Fei-Fei, “ImageNet: 一个大型分层图像数据库”, *2009 IEEE conference on computer vision and pattern recognition*, 第248-255页, IEEE, 2009年。
- [5] K. Zhu, H. Guo, T. Yan, Y. Zhu, J. Wang和M. Tang, “PASS: 用于人员重新识别的部分感知自监督预训练”, *arXiv preprint arXiv:2203.03931*, 2022年。
- [6] D. Fu, D. Chen, J. Bao, H. Yang, L. Yuan, L. Zhang, H. Li和D. Chen, “用于人员重新识别的无监督预训练”, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第14750-14759页, 2021年。
- [7] H. Luo, P. Wang, Y. Xu, F. Ding, Y. Zhou, F. Wang, H. Li和R. Jin, “用于基于Transformer的人员重新识别的自监督预训练”, *arXiv preprint arXiv:2111.12084*, 2021年。
- [8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski和A. Joulin, “自监督视觉Transformer中的新兴特性”, *arXiv preprint arXiv:2104.14294*, 2021年。
- [9] Y. Sun, L. Zheng, Y. Yang, Q. Tian和S. Wang, “超越部件模型: 具有改进部件池化的行人检索 (以及强大的卷积基线)”, *Proceedings of the European Conference on Computer Vision (ECCV)*, 第480-496页, 2018年。
- [10] J. Devlin, M.-W. Chang, K. Lee和K. Toutanova, “BERT: 用于语言理解的深度双向Transformer预训练”, *Proceedings of NAACL-HLT*, 第4171-4186页, 2019年。
- [11] Bao H, Dong L, Wei F, “Beit: 图像Transformer的BERT预训练”, *arXiv preprint arXiv:2106.08254*, 2021。
- [12] 何恺明, 陈鑫磊, 谢赛, 李言, Dollar, Girshick, “掩码自动编码器是可扩展的视觉学习器”, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第16000-16009页, 2022年。
- [13] Xie Z, Zhang Z, Cao Y, Lin Y, Bao J, Yao Z, Dai Q, Hu H, “Simmim: 一种简单的掩码图像建模框架”, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第9653-9663页, 2022年。
- [14] L. Wei, S. Zhang, W. Gao和Q. Tian, “用于行人重识别的桥接域间隙的行人迁移GAN”, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 第79-88页, 2018年。
- [15] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang和Q. Tian, “可扩展行人再识别: 一个基准”, *Computer Vision, IEEE International Conference on Computer Vision*, 第1116-1124页, 2015年。
- [16] Z. Zheng, L. Zheng和Y. Yang, “由GAN生成的未标记样本改善了体外人员重新识别基线”, 载于*Proceedings of the IEEE International Conference on Computer Vision*, 第3754-3762页, 2017年。
- [17] J. Miao, Y. Wu, P. Liu, Y. Ding和Y. Yang, “用于遮挡行人再识别的姿态引导特征对齐”, *Proceedings of the IEEE/CVF international conference on computer vision*, 第542-551页, 2019年。

- [18] 景亮, 田野, “基于深度神经网络的自监督视觉特征学习综述”, *IEEE transactions on pattern analysis and machine intelligence*, 第43卷, 第11期, 第4037-4058页, 2020年。
- [19] K. He, H. Fan, Y. Wu, S. Xie和R. Girshick, “用于无监督视觉表示学习的动量对比”, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 第9729-9738页, 2020年。
- [20] 陈旭, 范浩, 吉尔希克, 何恺明, “基于动量对比学习的改进基线”, *arXiv preprint arXiv:2003.04297*, 2020。
- [21] X. Chen, S. Xie和K. He, “关于训练自监督视觉转换器的实证研究”, 载于 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 第9640-9649页, 2021年。
- [22] 陈天奇, Kornblith S, Norouzi M, Hinton G, “一种简单的对比学习视觉表征框架”, *International conference on machine learning*, 第1597-1607页, PMLR, 2020。
- [23] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, “Bootstrap your own latent——一种新的自监督学习方法”, *Advances in Neural Information Processing Systems*, 第33卷, 第21271-21284页, 2020年。
- [24] 谢志强, 林毅, 姚志强, 张志强, 戴启明, 曹阳, 胡浩, “基于Swin Transformer的自监督学习”, *arXiv preprint arXiv:2105.04553*, 2021。
- [25] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, “通过生成式预训练改进语言理解”, 2018。
- [26] 李泽, 陈泽, 杨帆, 李伟, 朱毅, 赵晨, 邓荣, 吴磊, 赵瑞, 唐明, *et al.*, “Mst: 用于视觉表示的掩码自监督Transformer”, *Advances in Neural Information Processing Systems*, 第34卷, 第13165-13176页, 2021年。
- [27] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille和T. Kong, “ibot: 具有在线分词器的图像BERT预训练”, *arXiv preprint arXiv:2111.07832*, 2021。
- [28] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat和N. Ballas, “用于高效学习的掩码孪生网络”, *arXiv preprint arXiv:2204.07141*, 2022。
- [29] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang和Q. Tian, “野外行人再识别”, 见 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 第1367-1376页, 2017年。
- [30] 钱旭, 付宇, 姜宇刚, 项天, 薛向, “用于人员重新识别的多尺度深度学习架构”, *Proceedings of the IEEE International Conference on Computer Vision*, 第5399-5408页, 2017年。
- [31] W. Li, X. Zhu和S. Gong, “用于人员重新识别的和谐注意网络”, 见 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 第2285-2294页, 2018年。
- [32] C. Song, Y. Huang, W. Ouyang和L. Wang, “用于人员重新识别的掩码引导对比注意模型”, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 第1179-1188页, 2018年。
- [33] 赵亮, 李想, 庄毅, 王健, “基于深度学习的部分对齐表示用于人员再识别”, *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3219–3228, 2017。
- [34] Y. Suh, J. Wang, S. Tang, T. Mei和K. Mu Lee, “用于人员重新识别的部分对齐双线性表示”, 见 *Proceedings of the European Conference on Computer Vision (ECCV)*, 第402-419页, 2018年。

[35] G. Wang, Y. Yuan, X. Chen, J. Li和X. Zhou, “学习具有多种粒度的判别特征用于人员重新识别”, *2018 ACM Multimedia Conference on Multimedia Conference*, 第274-282页, ACM, 2018年。[36] S. He, H. Luo, P. Wang, F. Wang, H. Li和W. Jiang, “Transreid: 基于Transformer的目标重新识别”, *Proceedings of the IEEE International Conference on Computer Vision*, 2021年。[37] H. Luo, Y. Gu, X. Liao, S. Lai和W. Jiang, “用于深度人员重新识别的技巧包和强基线”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 第0-0页, 2019年。[38] K. Zhu, H. Guo, S. Zhang, Y. Wang, G. Huang, H. Qiao, J. Liu, J. Wang和M. Tang, “AAformer: 用于人员重新识别的自动对齐Transformer”, *arXiv preprint arXiv:2104.00921*, 2021年。[39] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, F. Wang, R. Jin和X. Sun, “超越外观: 一种以人为中心的视觉任务的语义可控自监督学习框架”, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第15050-15061页, 2023年。[40] H. Hu, X. Dong, J. Bao, D. Chen, L. Yuan, D. Chen和H. Li, “PersonMAE: 使用掩码自动编码器进行人员重新识别预训练”, *IEEE Transactions on Multimedia*, 2024年。