基于 DINO 的多目标追踪算法

Kaiwen Tu

2025.1.8

摘要

近年来，多目标追踪领域引起了人们广泛的关注。现有的多目标追踪算法多采用基于 CNN 的模型作为基线方法中的视觉特征编码器，这意味着视觉特征的鲁棒性对于多目标追踪效果来说十分重要。而自监督学习得到的图像特征编码器往往能够不受错误标注数据的干扰，实现更加准确的特征提取。因此，本文使用 DINO 系列模型作为视觉特征编码器，经过微调，使图像产生适应多目标追踪任务的通用视觉特征，从而对传统的 DeepSORT 追踪算法进行改进。本文在 MOTChallenge 平台的多个数据集上对使用 DINO 系列模型改进后的 DeepSORT 进行了评估，与相同基线方法下的不同特征提取模型进行对比实验，得到了十分优异的表现效果。

关键词：多目标追踪；自监督学习；外观特征模型；DINO

1 引言

多目标追踪（Multiple Object Tracking，MOT）要求检测器同时跟踪视频中出现的多个对象，持续识别他们的身份并生成他们各自的运动轨迹。其在自动驾驶、智能监控、行为识别等方向应用广泛，在计算机视觉领域有着十分重要的地位。

近年来，越来越多的深度学习模型被用于 MOT 领域，以提高其追踪器的性能和鲁棒性 [1]。目前的 MOT 跟踪算法大多使用了检测器与特征提取模型来提升 MOT 的性能表现 [2]。其中检测器生成一个 2 维的边界框来获取被检测对象的边界信息，使得边界框内完整包含被检测目标, 如 yolo [3] 系列，POI [4] 等；而特征提取模型通常采用基于卷积神经网络（Convolutional Neural Network，CNN）类型的模型来对边界框内的对象进行特征提取，从而在不同帧间识别出同一对象，如 iBoT [5]，ResNet [6] 系列等。这说明外观模型特征提取的精确度对于 MOT 结果有着十分重要的影响。现有的方法一般采用在目标检测或语义分割等视觉任务中训练得到的编码器来提取目标特征。此外，为了达到更好的特征提取效果，人们往往在大量特定下游任务类型的数据集上训练自己的特征提取模型，例如京东提出的 fast-reid [7]。但这些方法往往依赖大量的有标签数据集，而人工标记数据往往需要较高的人力成本，且特征提取效果可能会受到可能存在的错误标注数据影响。

随着自监督学习近年来的快速发展，将输入数据本身作为监督而无需依赖人工标注的学习方式能够极好地适配各种下游任务 [8]。研究人员发现自监督学习方法在对抗性示例、标签损坏以及常见输入损坏等多个方面具有鲁棒性，并且有助于检测接近分布的异常值。在这一

背景下，计算机视觉领域涌现了许多通过自监督学习训练的模型，如 DINO，Bert 等。这类模型往往通过在 ImageNet-1k 等数据集上进行预训练学习即可得到通用视觉特征，并通过开箱即用或微调的方式运用到各种下游任务中，并能够取得良好的表现效果。其中，自监督学习中的对比学习法在计算机视觉领域展现出了十分强大的能力。目前已经有部分研究将对比学习运用到多目标追踪外观模型的训练中，如 Kim S 等人，这种自监督学习方式能够很好地提高编码器输出特征的判别能力。此外，还有研究表明自监督学习可以通过在训练时解决一个非平凡的借口任务，使网络具有学习上下文显著特征的能力。因此，自监督学习范式的兴起为多目标追踪特征提取模型的改进提供了新的思路。

2 相关工作

2.1 多目标追踪

多目标追踪是计算机视觉领域中的一项关键技术，其往往需要在一个具有大量移动目标的场景下，检测目标的边界框并为其赋予编号。在这个过程中，多目标追踪算法往往会遇到很多困难，例如当待检测目标被另一个物体局部遮挡时，相机能获取的外观信息便大量减少，从而导致跟踪丢失。而多目标跟踪中使用的相机往往不能识别这类突发因素，因此这往往会导致检测器识别目标上的困难。而两个相似物体重叠或混合的过程中也有可能引发 ID 变换，即同一检测对象在被遮挡前后所拥有的 ID 不一致的问题。此外，具有较小体型的目标在检测与识别上往往也具有更高的难度，等等。

在此背景下，人们提出了基于检测的多目标追踪算法，例如 DeepSORT，StrongSORT 等。通常来说，一个基于检测的多目标追踪框架往往具备以下范式：首先，视频中的每一帧都被送入检测器中，检测器将会检测出该帧中存在的所有目标，并为它们分别赋予一个边界框。随后，每一个边界框内的检测目标将会通过 MOT 算法赋予一个 ID 编号，用于标识检测对象。接下来，外观特征提取模型将会对所有边界框内的对象进行特征提取，保留到该帧对应的特征队列中。最后，匹配算法会基于特征的相似度对相邻帧的特征队列中的特征进行匹配，将两帧间相似度最高的特征对应的目标关联起来，识别为同一个目标对象。

然而，这种基于检测的追踪方法将检测视为追踪过程中的首要任务，将外观特征视为次要任务。在这种方式中，检测器与外观模型是两个完全独立的模块，且算法流程决定了检测器对追踪效果的影响大于外观特征模型。因此有研究认为，这种基于检测的追踪方法使得人们较难权衡二者对最终跟踪效果的影响比重。基于此理论，有研究提出了一种将检测与外观放在同等地位下的 FairMOT 框架。在此框架中，视频的每一帧被输入到一个编码器-解码器结构网络中以提取高帧率的特征图，在网络的最后添加了两个同质分支，分别用来检测对象与提取目标特征，以获得检测与外观间的良好折衷。

此外，随着视觉编码器（Vision Transformer）的火热发展，基于注意力机制的检测也逐渐被引入多目标追踪领域，如 TransTrack，TrackFormer 等。这种外观特征模型采用 Transformer 作为其模型骨干，将查询（query）-密钥（key）机制引入外观特征模型的训练中，极大地提高了跟踪器在特征提取方面的检测性能。

2.2 DINO 系列模型

随着 Transformer 的火热发展，大量研究将网络模型的骨干从残差网络更换为了 Vision Transformer，如 DINO。DINO 是采用 Vision Transformer 进行自监督学习的经典工作，其主要探究的问题是，自监督学习方法能否为 Vision Transformer 带来新的特性。DINO 提出了一种自蒸馏框架，与 BYOL 中使用的在线编码器与目标编码器类似，DINO 的自蒸馏框架中提出了结构完全相同的学生网络与教师网络，通过图像裁剪等数据增强手段，在大量无标签数据集上进行自监督训练，使学生网络不断拟合教师网络的输出。其在 ImageNet 系列数据集上的线性分类与 KNN 分类上都取得了十分出色的效果，且在 KNN 分类器下对不同类别间的聚类效果十分显著，能够较好地应用在下游任务中。此外，DINO 的自注意力热图能够精确地获取每个物体的轮廓，甚至匹配语义分割的效果，这在以往的有监督训练和卷积神经网络中都是不具备的。

在大模型时代背景下，为了提升 DINO 模型的性能，Meta 提出了 DINOv2。DINOv2 的工作中，其通过网络爬取，人工筛选，相似度去重等多种技术构造了更大体量的精选数据集 LVD-142M，通过一种判别性自监督预训练方法来学习特征，使用图像级损失与补丁级损失同时对模型进行训练，并解耦目标函数之间的权重绑定。此外，DINOv2 还在教师网络中使用 Sinkhorn-Knopp 中心化，并使用 KoLeo 正则化器使得同一批次内图像特征分布尽量均匀。最后，DINOv2 还采用逐步增加分辨率技术，进一步提升了模型在像素级下游任务上的表现。这种训练方法使得 DINOv2 具有很好的多模态特性，其不仅在 ImageNet 等数据集上取得了较好的分类效果，还在分割、深度估计、图像修复等下游任务中取得了十分优异的表现。

作为特征提取器，经过预训练的 DINO 系列模型往往具有开箱即用的特性。其通过在 LVD-142M 等大型数据集上的自监督训练后，能够产生适配多种下游任务的通用视觉特征，无需微调即可应用在多种类型的下游任务中。对于 DINO 而言，其模型骨干根据参数量的区别可分为 ViT-S（small）与 ViT-B（Base），ViT-S 将输出具有 384 个维度的特征向量，ViT-B 输出具有 768 个维度的特征向量。对于 DINOv2 而言，在 DINO 的基础上，其采取 ViT-L（large）与 ViT-G（giant）作为模型骨干，ViT-L 输出具有 1024 个维度的特征向量，ViT-G 输出具有 1536 个维度的特征向量，并且在自监督训练的过程中，ViT-S，ViT-B，ViT-L 的初始权重将由预训练完成的具有最强特征提取能力的 ViT-G 骨干蒸馏得到，而不是从头开始训练，这使得小模型能够学习到大模型强大的语义特征。

2.3 预训练的外观特征模型

在多目标追踪领域中，强大的外观特征提取模块往往是使跟踪器获得良好检测性能的关键，其中 CNN 与各种基于 CNN 的深度学习模型通常被用于提取目标的外观特征。早期时候，人们通常在 Mars，Market1501，Crowdhuman 等行人重识别（Re-ID）数据集上从头开始训练一个完整的外观特征模型，如 DeepSORT 中的方法。这种方法往往费时费力，且其学习到的特征往往只适用于特定任务，而不具备良好的泛化能力。在此背景下，预训练的外观特征模型逐渐被应用于多目标跟踪领域。

Fast-reid 引入了通用的 ResNet 以及它的变种 ResNeXt，ResNetSt 作为其网络骨干，使用平均池化，最大池化，注意力池化等池化方式作为聚合层，在大量行人，车辆等目标重识别数据集上进行预训练，同样通过知识蒸馏的框架获取轻量级模型。其通过提供包括行人重

识别，跨域行人重识别，车辆行人重识别等多个同种类型下游任务的预训练模型及配置，使得其能够快速落地到各种多目标跟踪领域中，适配各种跟踪器的外观特征要求。Fast-reid 提供了一种可泛化到其他目标跟踪任务中的模型框架，在多目标跟踪领域有着广泛的应用。

大模型发展的过程中，越来越多预训练的外观特征模型被引入多目标跟踪领域。这类特征提取模型经过在大量数据集上的预训练，往往具有通用的视觉特征。例如在联想的工作中，通过使用动量对比学习得到的 MoCo-v2 作为外观特征模型，在 BDD 100K 数据集上取得了最好表现效果（State of the arts，SOTA）。Li S 等人的工作中使用了预训练的 CLIP 作为外观特征模型，通过微调来使其具有适配 Re-ID 任务的语义特征。这类在大量数据集上预训练得到的特征编码器通常具有强大的表征能力，为了使其适配多目标跟踪的任务场景，往往只需要对它们在特定 Re-ID 数据集上进行微调，即可达到甚至超越其他模型从头开始训练得到的效果。并且，这类模型往往具有强大的泛化能力。

而 DINO 系列工作中提出的 LVD-142M 精选数据集被应用于预训练的过程中能够较好地使模型学习到更准确的语义特征。相比之下 CLIP 更多被应用于文本数据，Moco 系列模型预训练使用的数据集较为单一。受此启发，本文将经过预训练得到 DINO 系列模型作为外观特征模型，通过其强大的特征提取能力，实现多目标跟踪领域中更加鲁棒的特征提取。此外，本文还尝试采用经过预训练的具有最强特征提取能力的 ViT-g/14 作为模型骨干，同样通过在行人重识别数据集 Market1501 上的微调，得到了用于实验的 DINOv2。通过 DINO 系列模型，本文完成了对跟踪目标特征的提取。

3 本文方法

本文基于 DeepSORT 的基线方法构建了本文的管线方法如图 1所示。总体来说，本文的方法分为了边界框检测，特征提取，数据关联三部分。在第一部分，对于一段由完整视频帧序列构成的输入，本文采用 Faster-RCNN 网络作为检测器对它们进行目标边界框的检测，得到每一帧图像中目标的边界框对象；在第二部分，对于检测器传出的边界框信息，本文采用 DINO 系列模型作为外观特征模型，对边界框内的目标进行外观特征提取，得到描述目标外观信息的特征向量；在第三部分，本文对相邻帧之间的目标轨迹进行数据关联，采用卡尔曼滤波与目标边界框信息计算相邻帧目标间的马氏距离，采用 DINO 系列模型得到的外观特征信息计算相邻帧目标间的外观余弦距离，从而对确认态轨迹完成级联匹配。最后，本文对不确定态轨迹进行 IOU 匹配，减少漏检情况的发生，从而完成了相邻帧之间目标的跟踪检测。在接下来的部分中，本文将详细介绍本文的算法中每个部分的详细过程。

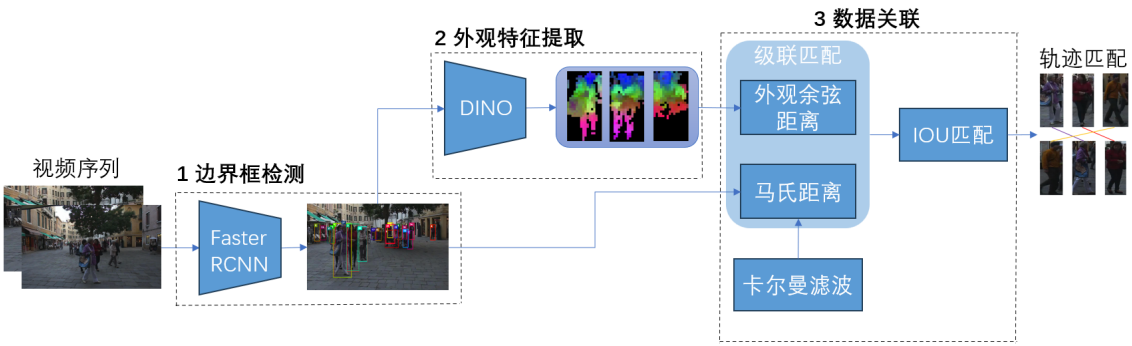

图 1. 本文算法框架

3.1 边界框检测

为了较好地描述图像中每个目标的位置信息，所有的多目标跟踪范式都会采用一个检测器来获得被检测目标的位置信息，并为其生成边界框。为了实现高质量和高性能的检测，本文采用了 Faster-RCNN 作为检测器。Faster-RCNN 以其出色的性能而闻名，能够在较短的测试时间内实现高精度的目标检测。通过 Faster-RCNN，本文能够获得准确的目标位置信息，为后续的多目标跟踪任务提供了可靠的基础。如图 2所示，Faster-RCNN 检测器主要分为卷积层，区域提取网络，ROI 池化层，分类层四个部分。

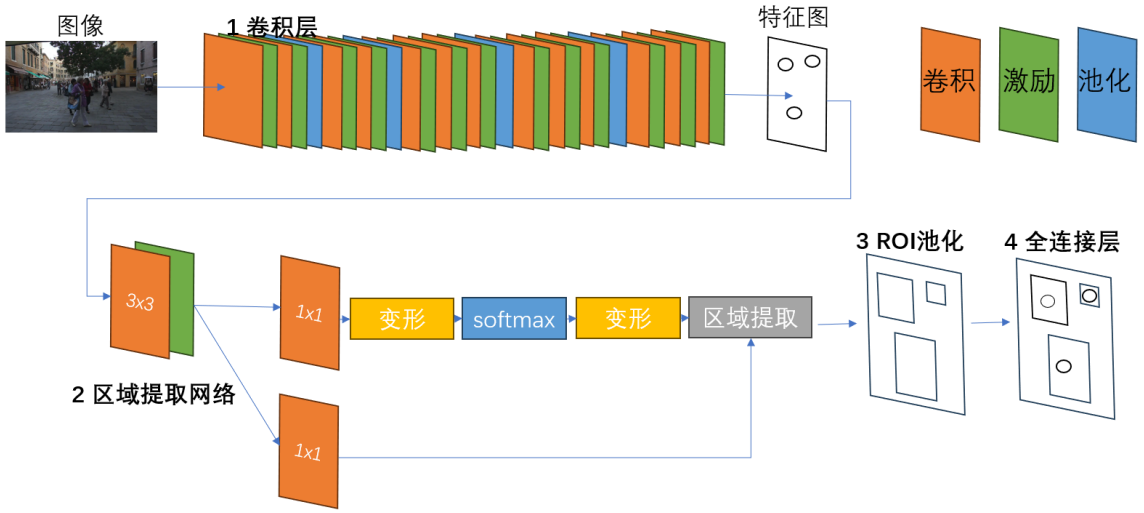

图 2. Faster-RCNN 检测器网络结构

卷积层（Conv）：卷积层是 Faster-RCNN 网络的主干，用来生成目标图像的特征图。它的网络结构由大量的卷积层，激励层和池化层构成，它们通过一系列的卷积操作将输入图像转换为特征图，并将这些特征图传入区域提取网络中进行区域提取。

区域提取网络（Region Proposals Network，RPN）：对于卷积层传入的特征图，区域提取网络对上面的每一个点生成多个对应的锚点，随后为了区分锚点所处位置属于前景或后景，区域提取网络使用 1x1 的卷积层实现计算锚点所处位置属于前后景分类的概率，进行区域特征提取。此外，区域提取网络还使用了一个 1x1 的卷积实现对锚点回归偏移值的预测。

ROI 池化：ROI 池化层主要的目的是对特征进行降维，使其适配全连接层的网络输入。这意味着无论原始图片大小都会被固定为一个相同维度的特征。此外，每个区域内的前景会被投影到原始图片的特征图中。ROI 池化层的设计使得 Faster-RCNN 在处理不同尺寸和位置的目标时具有良好的鲁棒性和泛化能力。

全连接层：在全连接层中，Faster-RCNN 会对每个区域再使用一次边界框回归，以校正边界框的最终位置。这种边界框回归操作进一步提高了检测的精度和准确性，尤其是对于小尺寸目标或者与其他目标高度重叠的情况。通过全连接层的深度学习，Faster-RCNN 能够根据每个区域的特征信息，有效地调整边界框的位置和大小，从而实现更加精确的定位。

通过 Faster-RCNN 网络，本文可以得到一张图片中所有待检测目标的检测框，并将其用在后续轨迹匹配的过程中。

3.2 外观特征提取

有了目标的边界框信息，多目标跟踪算法通常使用一个外观特征模型来对边界框内的目标进行外观特征提取。DeepSORT 的基线方法中使用的外观特征模型是一个在大规模行人重识别 Mars 数据集上经过有监督方式从头开始训练得到的 CNN 模型。Mars 数据集包含了大量行人在不同相机视角下的图像，非常适合在人员跟踪环境下对深度学习模型进行训练。DeepSORT 中使用的 CNN 的网络结构如图 3所示，该网络包含了两层卷积层，一层池化层以及六个残差块，最后通过一层全连接层将特征聚合为一个 128 维的向量，并进行批量归一化与 L2 归一化处理。

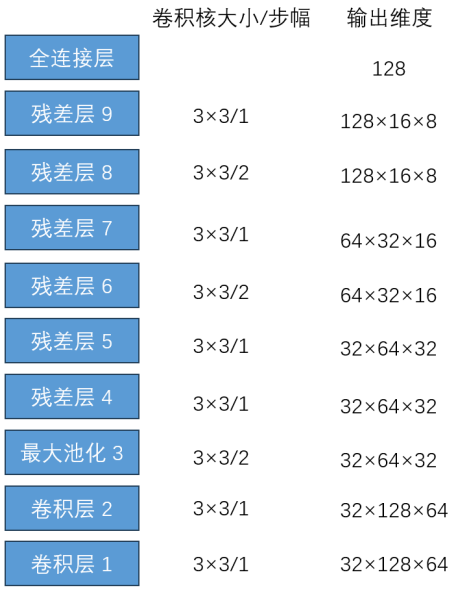

图 3. DeepSORT 的外观特征模型 CNN 结构

在基线方法中，DeepSORT 使用的 CNN 网络结构简单，对目标的特征表达能力有限。本文尝试探究具有强大表征能力的 DINO 系列模型是否在多目标跟踪任务中有着更加鲁棒的表现，因此，本文将 DINO 系列模型引入 DeepSORT，替换基线方法中的外观特征模型，以完成对跟踪目标外观特征的提取。接下来，本文将介绍本文使用的 DINO 与 DINOv2 模型。

本文引入了 DINO 作为外观特征模型。为了有效挖掘全局特征与局部特征之间的关系，DINO 放弃了负样本，提出了学生网络与教师网络架构，使学生网络不断拟合教师网络的特征，如图 4所示。

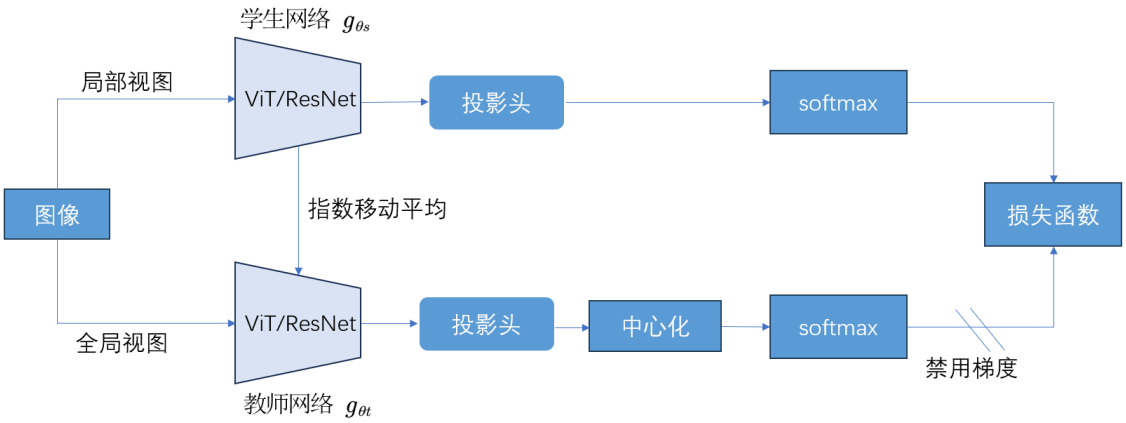

图 4. DINO 模型结构示意图

首先，DINO 定义了学生网络 $g_{\theta s}$ 与教师网络 $g_{\theta t}$ ，其中 θs 与 θt 代表两个网络的参数。二者体系结构相同，但参数不同。对于给定的一张图片 x，作者首先通过随机裁剪等数据增强方式得到了许多不同的图像，由这些不同的图像构成了一组集合 V。集合 V 中包含了两个全局视图 x_1^g 和 x_2^g 以及许多分辨率较小的局部视图。局部视图被传递给学生网络 $g_{\theta s}$ ，全局网络被传递给教师网络 $g_{\theta t}$ ，用于鼓励从局部到全局的通信。

网络结构 g 由骨干 f 和投影头 h 组成。DINO 的骨干可采用 Vision Transformer 或 ResNet，其输出均可直接应用于下游任务中。投影头由 3 层的多重感知器（Multilayer Perception，MLP）结构组成，且与 BYOL 不同，DINO 中学生网络并没有预测器结构，从而学生网络与教师网络的结构完全一致。

对于一张图片 x，学生网络与教师网络的输出分别为 K 维上的概率分布 Ps 和 Pt。其中 Ps 和 Pt 都是对网络 g 的输出进行 softmax 获得的：

$$P(x)^{(i)} = \frac{\exp\left(g_\theta(x)^{(i)}/\tau\right)}{\sum_{k=1}^{K}\exp\left(g_\theta(x)^{(k)}/\tau\right)}$$

其中 $\tau > 0$ 是一个控制输出分布锐度的温度参数。给定一个固定的教师网络 $g_{\theta t}$ ，本文通过最小化学生网络参数 θ_s 的交叉熵损失来学习匹配这些分布：$\min H(P_t(x), P_s(x))$ ，其中 $H(a,b) = -a\log b$ 。而在前文提到，集合 V 中包含了两个全局视图 x_1^g ，和部分局部视图 x_2^g 。学生网络的输入是局部视图，教师网络的输入是两个全局视图。因此，最终 DINO 定义的损失函数如下：

$$\min_{\theta s}\sum_{x\in\left\{x_1^g,x_2^g\right\}}\sum_{x'\in V,x'\neq x}H(P_t(x),P_s(x'))$$

在学生网络中，本文使用常规的随机梯度下降法（SGD）对参数进行更新，但在教师网络上禁用梯度，使用学生网络的指数移动平均值（EMA）进行更新。即 $\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$ 。在训练期间，λ 遵循余弦学习率衰减策略从 0.996 到 1 之间变化。

此外，在自监督学习过程中，可能会出现多个输入数据映射到同一组相同特征的情况，即引发模型坍塌。这种情况往往是由于网络在优化过程中陷入了局部最优解，只考虑了一部分特征表示，忽视了其他数据的特征，从而导致了多样性缺失的现象。因此，教师网络在输出后会进行中心化（centering）操作，即在输出后添加一个偏置项 c。偏置项 c 将会按照如下方

式更新：

$$c \leftarrow mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta t}(x_i)$$

其中 m 是一个速率参数，B 为批量大小。实际上，c 是一个均值项，目的是使得教师模型的输出趋近均匀分布，防止特征分布过于尖锐化。

本文采用经过预训练的 ViT-B/16 作为模型骨干，通过在行人重识别数据集 Market1501 上的微调，得到了 DINO 用于进行目标的外观特征提取。此外，本文还尝试采用经过预训练的具有最强特征提取能力的 ViT-g/14 作为模型骨干，同样通过在行人重识别数据集 Market1501 上的微调，得到了用于实验的 DINOv2。通过 DINO 系列模型，本文完成了对跟踪目标特征的提取。

3.3 数据关联

在完成边界框检测与外观特征提取后，为了对相邻帧之间的目标进行跟踪匹配，本文需要用到检测器输出的边界框与外观特征模型输出的特征向量进行数据关联。本文首先使用卡尔曼滤波来对目标当前的位置与运动信息来预测下一帧目标的位置信息，输出的预测轨迹有确定态和不确定态两种。随后，本文整合所有目标信息进行级联匹配：本文将检测器输出的检测框与卡尔曼滤波预测的确定态目标轨迹结合，以计算相邻帧目标间的马氏距离；同时，本文将用 DINO 输出的外观特征计算目标间的外观余弦距离。本文将马氏距离与外观余弦距离结合用作帧间目标匹配的依据，从而对帧间目标进行匹配。最后，对于未匹配成功的边界框与不确定态轨迹，本文通过计算边界框的 IOU 来对可能漏检的目标进行二次配对，从而完成相邻帧间目标的跟踪匹配。

4 复现细节

4.1 与已有开源代码对比

本文的工作是基于开源代码 https://github.com/dyhBUPT/StrongSORT 中 DeepSORT 部分的改进。在已有的 DeepSORT 的开源代码中，其特征提取模块使用的是图 3中提到的卷积神经网络。而本文将原有特征提取模块从简单的 CNN 结构替换为了预训练的 DINO 系列模型，如图 5所示。其余所有流程与原文中 Baseline 完全一致。也就是说，本文基于原文的改进仅仅在于将特征提取模块替换为了 DINO。

```python
class Net(nn.Module):
    def __init__(self, num_classes=751, reid=True, backbone = 'dinov2_g', head = 'linear', backbones = dino_backbones):
        super(Net, self).__init__()
        self.heads = {
            'linear':linear_head,
        }
        self.backbones = dino_backbones
        self.backbone = torch.hub.load(r'/home/kaiwen/.cache/torch/hub/facebookresearch_dinov2_main', 'dinov2_vitg14', source='local').cuda()
        # self.backbone = torch.hub.load('facebookresearch/dino:main', 'dino_vitb16').cuda()
        self.backbone.eval()
        self.head = self.heads[head](self.backbones[backbone]['embedding_size'],num_classes)
        self.reid = reid

    def forward(self, x):
        with torch.no_grad():
            x = self.backbone(x)
            if self.reid:
                return x
        x = self.head(x)
        return x
```

图 5. DINO 模型代码实现

具体来说，本文新增 others/generate_detections.py 文件，在该文件中，使用 dinov2 来对 MOT17 数据集中的视频序列生成对应的检测特征。得到 features 文件。随后基于这些 features 与视频序列进行匹配。

4.2 实验环境搭建

本文的实验环境搭建基于 DeepSORT 与 DINOv2 两篇工作。具体依赖可见https://github.com/dyhBUPT/StrongSORT以及https://github.com/facebookresearch/dinov2中的 requirement.txt。

4.3 界面分析与使用说明

本次复现工作基于 RTX 4090 显卡，使用开源 deepsort 框架进行改进，在 MOT17 数据集上运行。具体的使用说明可见代码文件夹中的 readme.md 文档。

4.4 创新点

使用 DINO 系列模型替换原有的特征提取模块，在特征提取方面实现更鲁棒的性能。

5 实验结果分析

5.1 数据集

数据集方面，本文将经由 DINO 系列模型改进后的 DeepSORT 在 MOT16，MOT17 与 MOT20 数据集上进行实验。对于 MOT16 数据集，其包含了 5316 帧的 7 个训练集与 5919 帧的 7 个测试集；对于 MOT17 数据集，在更精确的真实值（ground truth）条件下，其包含 DPM（可变形组件模型，Deformable Parts Model），Faster-RCNN，SDP（基于尺度的池化，Scale-Dependent pooling）三种类型的检测器构成的视频序列；MOT20 的训练集与测试集均分别由 4 个更长且更复杂的视频序列构成。本文采用 Faster-RCNN 检测下生成的视频序列作为本文的实验对象。MOT 系列数据集有着广泛而通用的评估指标，其中本文采用 MOTA，MT 以及 IDF1 来作为跟踪器性能评估的主要依据。

5.2 评测指标

MOTA（Multi-Object Tracking Accuracy）通过计算检测错误样例（FN）与匹配错误样例（FP）和发生 ID 变换次数（IDsw）的总和在正确样例中的占比，从而用 1 减去它们得到。该指标更加注重检测的整体性能，也是 MOT 领域使用最为广泛的指标。

$$MOTA = 1 - \frac{\sum(FN + FP + IDsw)}{\sum GT} \in (-\infty, 1]$$

MT（Mostly Tracked）统计了成功被检测到的目标在对应轨迹总长度中占比超过 80% 的数量，该指标能够很好地体现追踪过程中的精确度与鲁棒性。

IDF1（ID F1 Score）评估了正确检测样例（IDTP）在它与漏分配样例（IDFN）和错误分配样例（IDFP）的和的平均数量中的占比。该指标很好地描述了 ID 匹配间的一致性。

$$IDF1 = \frac{IDTP}{IDTP + 0.5\,(IDFP + IDFN)}$$

5.3 模型准备

DINOv2 底层采用 Vision Transformer（ViT）结构，根据 ViT 参数量大小的区别可将骨干分为 ViT-s (small), ViT-b (base), ViT-l (large), ViT-g (giant) 四种类型。基于上述背景，本文采用特征提取能力最强的 ViT-g 作为模型骨干，一层线性层用作微调，得到新的外观特征提取模型 DINOv2。本文将 DINOv2 用在 Market1501 行人重识别数据集上，使用 SGD 优化器进行微调，学习率大小设置为 1e-5，共进行 20 个 epoch，动量参数设置为 0.9，权重衰减系数设置为 5e-4，batch_size 大小为 32。在微调结束后，本文将 DINOv2 用于 DeepSORT 的外观特征模块。

此外本文还采用预训练的 ViT-B/16 作为模型骨干，一层线性层用作微调，得到外观特征提取模型 DINO。同样的本文将 DINO 用在 Market1501 数据集上进行微调，微调时参数设置与 DINOv2 完全相同。

检测器方面，与 DeepSORT 的基线方法相同，本文使用 POI 中的 Faster-RCNN 作为检测器，他们通过在大量数据集上训练得到了可观的检测器性能。DeepSORT 参数方面，本文将最大余弦距离设置为 0.32，nms 最大阈值设置为 1.0，检测最小置信度设置为 0.6。

5.4 定量实验

在本节中，本文将 DeepSORT 作为基线方法，将 DINOv2 与 DINO 作为外观特征模型，与基线方法中 CNN，fast-reid 等外观特征模型进行对比实验，在 MOT17 数据集上评估本文的实验结果。本文的结果展示在表 1 当中。

对于 MOT17 数据集而言，在完全相同的实验条件下，DINOv2 在 MOTA 上的表现依旧达到了 76.495%，优于基线方法和 fast-reid 的表现，但在 IDF1 方面逊于基线方法与 fast-reid。其跟踪性能与在 MOT16 数据集上表现一致，这说明本文的方法在不同的数据集上拥有较好的泛化能力。值得注意的是，经过自监督学习的 DINO 模型在该数据集上取得了超过 DINOv2 的效果，达到了 76.565% 的最高值。但是注意到提升几乎非常微小，因此这种替换特征的尝试的意义并不大。

表 1. MOT17 验证集上的跟踪结果

Method	**MOTA**	**MT**	**IDF1**
CNN (Baseline)	76.337	186	**75.915**
fast-reid	76.465	195	73.716
DINO (ours)	**76.565**	**197**	72.637
DINOv2 (ours)	76.495	195	73.236

6 总结与展望

由于在传统的多目标追踪算法中，外观特征模块的训练往往使用有监督的训练方式，而人工标注数据往往费力费时，且需要花费大量时间在模型训练上。因此本文提出使用预训练的 DINO 系列模型改进多目标追踪算法，借助其强大的特征提取能力，在不使用标签数据的情况下学习通用视觉特征，提高其追踪精度。实验证明，相较于传统的有监督学习训练的外观特征模型而言，本文的方法具有较强的鲁棒性，能够在只进行微调的情况下便取得较好的特征提取效果。本文希望这些研究发现可以鼓励更多自监督学习的外观模型在多目标追踪领域的应用。在未来的工作中，本文希望尝试将 DINO 系列模型运用到多目标追踪领域的检测器或其他模块中，追求更强大的跟踪性能。

不过值得注意的是，替换特征相对于原来的提升非常微小，因此仍需要对其他方向进行改进的探索。

参考文献

参考文献

[1] PARK Y, DANG L M, LEE S. Multiple Object Tracking in Deep Learning Approaches: A Survey[J/OL]. Electronics, 2021, 10(19): 2406.

[2] DU Y, ZHAO Z, SONG Y, et al. StrongSORT: Make DeepSORT Great Again[J/OL]. IEEE Transactions on Multimedia, 2023, 25: 8725-8737.

[3] JIANG P, ERGU D, LIU F, et al. A Review of Yolo Algorithm Developments[J/OL]. Procedia Computer Science, 2022, 199: 1066-1073.

[4] YU F, LI W, LI Q, et al. POI: Multiple Object Tracking with High Performance Detection and Appearance Feature[C/OL]//HUA G, JÉGOU H. Computer Vision –ECCV 2016 Workshops. Cham: Springer International Publishing, 2016: 36-42.

[5] ZHOU J, WEI C, WANG H, et al. Image BERT Pre-training with Online Tokenizer[C/OL] // International Conference on Learning Representations. 2021[2024-04-20].

[6] TARG S, ALMEIDA D, LYMAN K. Resnet in Resnet: Generalizing Residual Architectures[M/OL]. arXiv, 2016[2024-04-20].

[7] HE L, LIAO X, LIU W, et al. FastReID: A Pytorch Toolbox for General Instance Re-identification[C/OL]//Proceedings of the 31st ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2023: 9664-9667.

[8] LIU X, ZHANG F, HOU Z, et al. Self-supervised Learning: Generative or Contrastive[J/OL]. IEEE Transactions on Knowledge and Data Engineering, 2021: 1-1.