计算机科学
COMPUTER SCIENCE

联合人体姿态估计和多目标跟踪的跨数据集学习

曾泽华, 罗会兰

引用本文

曾泽华, 罗会兰. 联合人体姿态估计和多目标跟踪的跨数据集学习[J]. 计算机科学, 2023, 50(6A): 220400199-7.

ZENG Zehua, LUO Huilan. Cross-dataset Learning Combining Multi-object Tracking and Human Pose Estimation [J]. Computer Science, 2023, 50(6A): 220400199-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220400199

联合人体姿态估计和多目标跟踪的跨数据集学习

曾泽华　罗会兰
江西理工大学信息工程学院　江西 赣州 341000

摘　要　近年来,多目标跟踪任务获得了较大的进展,尤其是针对行人的多目标跟踪。通过对行人进行联合姿态估计,能够提升多目标跟踪算法对行人的运动预测,同时为更高阶的任务例如自动驾驶算法提供更多的信息。然而,在当前包含人体姿态估计标签的多目标跟踪数据集中,视频长度较短且目标稀疏,限制了多目标跟踪算法的研究。文中使用具有更多行人的多目标跟踪数据集 MOT17 和多人姿态估计数据集 COCO 进行跨数据集学习,基于循环训练策略有效提升了联合人体姿态估计下的多目标跟踪算法的性能。同时极化自注意力下采样和注意力上采样的使用,在提升算法训练速度的同时,增强了算法的人体姿态估计性能。

关键词:多目标跟踪;跨数据集学习;人体姿态估计;注意力机制
中图法分类号　TP183

Cross-dataset Learning Combining Multi-object Tracking and Human Pose Estimation

ZENG Zehua and LUO Huilan
School of Information Engineering,Jiangxi University of Science and Technology,Ganzhou,Jiangxi 341000,China

Abstract　In recent years,multi-object tracking has gained significant progress,especially for pedestrians. By performing joint pose estimation on pedestrians,it is possible to improve the motion prediction of pedestrians by multi-object tracking algorithms,while providing more information for higher-order tasks such as autonomous driving. However,in the current multi-object tracking dataset containing human pose estimation labels,the video length is short and the targets are sparse,limits the research of multi-object tracking. In the paper,cross-dataset learning is performed using the multi-object tracking dataset MOT17 and the multi-human pose estimation dataset COCO with more pedestrians. The performance of the multi-object tracking algorithm under joint human pose estimation is effectively improved based on a round-robin training strategy. The use of simultaneous polarized self-attention down-sampling and attention up-sampling enhances the human pose estimation performance of the algorithm while improving the algorithm training speed.

Keywords　Multi-object tracking,Cross-dataset learning,Human pose estimation,Attention mechanism

1　引言

多目标跟踪任务受益于近年来目标检测算法的性能提升,能够较好地检测到目标主体。为了保持对多个目标的持续跟踪,算法会预测目标的运动信息或提取出目标的外观特征信息,将两帧间相似的目标联系起来。目标的外观特征提取受益于单目标跟踪和行人重识别任务的研究成果,已经能较好地区分画面中的目标,但在多目标跟踪过程中对目标进行精准的运动预测,依旧是较为困难的,其中又以行人的运动预测最具挑战性。

当前最为主流的多目标跟踪数据集和挑战赛是 MOT[1-2] 系列,它包含各种场景的视频画面,主要用于训练和评价算法在复杂场景下的行人跟踪性能。但数据集中只包含了目标的位置信息和 ID 信息,仅依靠该数据集无法进一步提升多目标跟踪算法的性能或拓展算法的使用场景。因此研究者们引入了许多其他的数据集进行预训练,以期望算法在进行跟踪训练前就拥有较强的检测性能,或在 MOT 数据集训练后增加额外的训练过程,以提升算法对目标外观特征的提取能力,并通过额外的标签数据拓展算法的检测范围,例如人体姿态估计、人体头部检测、人体 3D 检测等。

由于不同数据集的数据分布和标签类型有较大差别,因此使用多种数据集进行模型训练通常是较为复杂且不稳定的,主流的训练策略有 3 种,分别是预训练、微调和伪标签。在多目标跟踪算法的训练过程中,通常使用 CrowdHuman[3] 和 COCO[4] 这类人物较为密集或分类数较多的数据集进行预训练,在 MOT 数据上进行正式训练,从而有效提升算法对行人的检测精度,进而提升多目标跟踪算法的性能。对于需要使用人体外观特征计算相似性的算法,通常使用额外的卷积神经网络和人体重识别数据集进行训练,并使用 MOT 数据集中的行人特征进行微调,从而在 MOT 测试集上取得较好的跟踪性能。当需要多目标跟踪算法同时完成多任务输出时,通过在非多目标跟踪数据上生成伪标签,模拟目标的

基金项目:国家自然科学基金(61862031);江西省主要学科学术和技术带头人培养计划——领军人才项目(20213BCJ22004)
This work was supported by the National Natural Science Foundation of China(61862031) and Jiangxi Provincial Foundation for Leaders of Disciplines in Science,Leading Talents Program(20213BCJ22004).
通信作者:曾泽华(deemozen@163.com)

运动，从而在其他类型的数据集上也能保持运动估计能力，为匹配算法提供信息。基于伪标签方法的优势是算法能在现有的数据和模型上拓展多目标跟踪的能力，例如通过 COCO 数据集训练就能够同时完成目标检测、人体姿态估计和多目标跟踪，但其缺点是多目标跟踪的性能较差，因为目标的运动信息并不真实。本文使用了循环训练策略，在保证算法的人体姿态估计能力的同时，有效提升了多目标跟踪的性能，拓展了跨数据集学习的训练策略。

在多目标跟踪任务中，目标的运动估计通常使用卡尔曼滤波或使用卷积神经网络进行预测，这两种方法只能预测目标的短期运动，当目标长时间被遮挡时就无法很好地估计其位置，从而导致跟踪丢失。行人轨迹预测任务通常使用人体关键点信息结合循环神经网络预测目标的长期运动轨迹[5]，因此能够联合人体姿态估计的多目标跟踪算法，为轨迹预测算法提供必要的信息，进而提升多目标跟踪性能。PoseTrack[6]提出了带有人体姿态估计标签和目标轨迹标签的数据集，其数据总量较大但每张图片中的人物较少，使得基于该数据集训练的算法在密集场景下性能较差。本文采用了多目标跟踪数据集 MOT17 和人体关键点数据集 COCO 进行联合训练，确保了算法在密集场景下的多目标跟踪性能。

为了达成人体关键点的联合检测跟踪，本文在 CenterTrack[7]算法的基础上，进一步使用为人体姿态估计任务设计的极化自注意力[8]改进下采样模块，提升了人体关键点检测质量；使用注意力上采样模块提升了模型的训练速度；使用了双向运动预测对目标进行匹配，在保证人体关键点输出的基础上提升了算法整体的多目标跟踪性能，通过 MOT17 数据集上的实验，验证了本文改进方法的有效性。

本文的主要贡献如下：

(1)使用极化自注意力提升了模型的人体姿态估计性能。

(2)使用注意力上采样模块提升了模型的训练速度。

(3)基于循环训练策略进行跨数据集学习，实现了联合人体姿态估计的多目标跟踪算法。

2 相关工作

2.1 多目标跟踪

多目标跟踪任务是同时检测到多个目标并记录其运动轨迹，主要针对雷达信号或图像视频内的特定类别目标进行跟踪。随着深度学习技术的快速发展，基于视频的多目标跟踪算法现在能够准确地检测到画面中的各类目标，并获得多帧内各目标的相似程度，将多帧内相似性高的目标匹配后就能得到目标在视频帧中的运动轨迹。当前多目标跟踪任务的主要难点是解决因遮挡导致的跟踪失败，以及提高跟踪算法的运行速度，从而做到实时在线的多目标跟踪。

DeepSORT[9]多目标跟踪算法使用两帧图像作为输入，将多目标跟踪抽象为二部图匹配，使用了 Faster R-CNN 目标检测算法作为检测器获得目标位置，使用 ResNet 提取目标外观特征用于目标的重识别，从而解决目标遮挡问题，最后使用级联匹配策略完成对目标的跟踪。CenterTrack 简化了算法的匹配策略，使用 CenterNet[10]作为主干网络，同时输出目标的位置信息和单向运动估计，并配合贪婪匹配算法大幅提升了多目标跟踪算法的运行速度。本文在 CenterTrack 的基础上实现了双向运动估计，减少了因目标遮挡导致的跟踪丢失，并对人体姿态估计任务做了针对性优化，提升了网络的多任务输出能力，进而实现了高效的联合人体姿态估计的多目标跟踪算法。

2.2 人体姿态估计

人体姿态估计就是检测出人体的关键点信息，并保证多个关键点能匹配到对应的主体上，因此人体姿态估计有两种实现方法，分别是自上而下和自下而上。自下而上是首先检测出画面中所有的人体关键点，再通过图匹配[11-12]或聚类的方式将关键点分配给对应的目标，此类方法的特点是速度较快。自上而下是首先通过目标检测算法得到目标的边界框，再将边界框中的内容进行关键点检测，从而完成多人的人体姿态估计[13]。

为了获得更精准的人体关键点位置，每一个人体关键点的二维坐标被抽象为二维正态分布，模型将输出人体关键点的热值图和偏置图，通过提取热值图的峰值位置来获得粗略位置，再通过与二维偏置相加获得精准的人体关键点位置[14]。CenterNet 继承了这一定位方式，并通过额外预测人体中心点位置和人体关键点相对中心点的偏置，巧妙地结合了自上而下和自下而上两种方法的优点，通过单一模型获得了目标的中心点位置和人体关键点位置，同时又通过偏置的预测获得了各个关键点与中心点的匹配关系，从而做到了高效的多人姿态估计。

2.3 注意力机制

由于深度学习通常需要大量的训练数据，而图像中又通常带有大量与任务无关的噪声，或包含大量的负例样本，因此注意力机制被运用于卷积神经网络内部，期望模型能够重点关注与任务相关的特征。早期的注意力机制主要使用通道注意力[15]，在之后的改进中又提出了结合空间注意力和通道注意力的方式[16]。有研究者将注意力机制用于可形变卷积[17]增强，提升了人体姿态估计的性能[18]。为了进一步提升人体姿态估计过程中模型对细粒度特征的提取能力，极化自注意力[8]在通道注意力和空间注意力的基础上增加了自注意力过程，恢复了特征的动态范围，从而提升了人体姿态估计模型的性能。

2.4 跨数据集学习

跨数据集学习期望模型通过多个数据集的训练，来获得通用的检测或识别能力，其难点是不同数据集的标签和分布不同，将会极大地影响模型训练和收敛。为了确保训练的可行性，文献[19-20]将多个数据集的标签通过映射混和成一个更大的数据集，从而让模型能够在单一数据集上完成训练。文献[21]将批规范化层(Batch Normalization)和损失计算独立，而卷积层保持参数共享，从而解决多数据集数据分布不一致的问题。为了训练出通用的目标检测算法，文献[22]使用 SE 注意力模块让模型在不同数据集上获取针对性的目标，得到特定的检测结果，从而做到跨数据集的学习。本文使用循环训练策略，在两种不同规模的数据集上进行循环训练，并控制超参数和损失计算，从而让模型在人体姿态估计和多目标跟踪任务上都有较好的表现，实现了跨数据集学习。

3 多任务模型和训练方法

3.1 联合人体姿态估计和多目标跟踪的双向匹配算法框架

为了实现联合人体姿态估计的多目标跟踪算法，本文

设计了基于注意力增强的多任务多目标跟踪网络,并通过双向运动预测分支实现了对目标的双向跟踪,从而避免因目标遮挡导致跟踪失败,网络的结构如图 1 所示。网络的输入是当前帧和过去帧的图像,输出是多种特征信息,包括目标中心点热值图、人体关键点热值图、中心点及人体关键点的偏置、基于中心点的指向自身人体关键点的偏置、目标边界框的宽和高、目标的双向运动估计。其中目标的前向运动预测将会使目标在遮挡后保持运动,从而避免跟踪丢失,目标的反向运动预测将和当前帧目标中心点位置相加,获得目标在过去帧的位置预测,从而基于距离和目标置信度完成贪婪匹配。

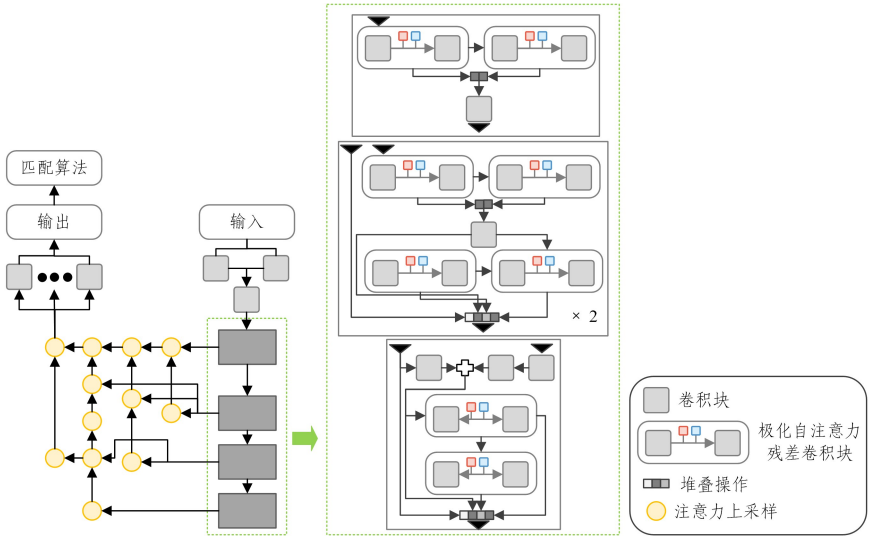


图 1　基于注意力增强的多任务多目标跟踪网络

Fig. 1　Attention-based augmented mult-itask multi-object tracking network

为了实现更加精准和高效的联合训练,本文使用不同的注意力模块改进了模型的下采样和上采样结构。在下采样阶段本文使用了极化自注意力模块,在残差卷积块的残差连接中增加注意力信息,从而提升网络对细粒度特征的感知能力。在上采样阶段,本文使用基于通道注意力和空间注意力的上采样模块替代可形变卷积,在不损失模型性能的前提下大幅提升了训练速度,更高效地实现了联合训练。

3.2　基于极化自注意力的下采样优化和注意力上采样模块

极化自注意力(Polarized Self-Attention)[8]是一种新型的针对像素级任务优化的注意力加权模块。该模块受启发于摄影领域,在拍摄过程中,光进入传感器前需要经过光学镜头滤除随机产生的眩光或反射,而类似的极化滤波也是只允许某一方向的光线的通过,完成空间层面的滤波,但在使用极化滤波后会损失光线的动态范围,因此需要额外的过程去恢复原始动态范围。极化自注意力正是使用了类似极化滤波的自注意力过程,并通过 Softmax 函数和 Sigmoid 函数组合,来恢复特征的动态范围,其完整结构如图 2 所示。

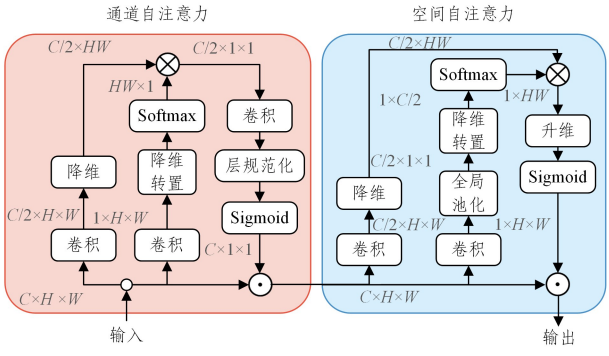


图 2　极化自注意力模块

Fig. 2　Polarization self-attention module

其中通道注意力可以表示为式(1),输入将首先通过 2 个卷积操作,W_{d1} 将特征的通道减小一半,减小后续过程的计算量,W_{d2} 将 3 维特征压缩成了 2 维,输出一个通道的特征图,其中减半的特征将通过降维操作 σ_1 把二维特征图降至 1 维,另一条被完全压缩的特征通路将通过降维和转置操作 σ_2 获得压缩了全部通道的空间特征信息,并通过 Softmax 函数得到空间注意力权重。之后通过将两个特征进行相乘即可完成对通道特征的自注意力操作,并通过第三个卷积 W_u 恢复完整通道数,使用 Sigmoid 函数生成通道注意力权重并通过点乘加权到输入中。维度的变化过程可见图 2 中的通道自注意力模块。

$$F_{ch}(x)=f_{\text{sigmoid}}(W_u(\sigma_1(W_{d1}(x)))\times f_{\text{soft}}(\sigma_2(W_{d2}(x)))) \tag{1}$$

经过加权的特征将进入空间自注意力模块,其过程可表示为式(2),首先通过两个卷积操作将通道的维度减半,其中的一个特征通路将进行全局池化 G,将空间信息完全压缩,并通过降维转置和 Softmax 函数获得压缩了全部空间的通道特征信息,再使用乘法得到所需的空间自注意力,并通过 σ_u 升维和 Sigmoid 权重生成,获得基于空间的注意力权重,最终通过点乘加权到输入中得到输出结果。

$$F_{sp}(x)=f_{\text{sigmoid}}(\sigma_u(\sigma_1(W(x))\times f_{\text{soft}}(\sigma_2(G(W(x)))))) \tag{2}$$

在通道注意的生成过程中,使用了空间自注意力,在空间注意力生成的过程中用到了通道自注意力,最终能使模型在下采样阶段直接关注到细粒度的语义内容,从而更好地保留目标中心点和人体关键点的特征信息,但该方式会产生较多参数,因此不适合用于模型的上采样阶段。

在网络的上采样阶段,通过将可形变卷积替换为图 3 所示结构的注意力上采样,能够在保证模型性能的基础上大幅提升训练速度。可形变卷积能够通过偏移量调整卷积核的感受野,从而提升模型在上采样阶段对目标的位置感知,本文

通过串联通道注意力和空间注意力，有效提升了模型对关键点位置信息的感知，从而达到与可形变卷积相似的效果。注意力的生成过程如图 4 所示，通过不同维度的最大池化和平均池化，将空间和通道维度的信息进行压缩，获得对应维度的特征信息，其中通道注意力通过 1×1 卷积和 Sigmoid 函数生成通道注意力权重，空间注意力通过 7×7 的卷积分析经过堆叠的空间特征，通过 Sigmoid 函数生成空间注意力，最后通过与原始输入相加生成带注意力的特征。由于卷积和池化操作都能并行处理，相比不易并行的可形变卷积能够大幅提升模型的训练速度。

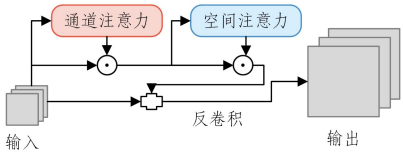


图 3 注意力上采样模块

Fig. 3 Attention up-sampling module

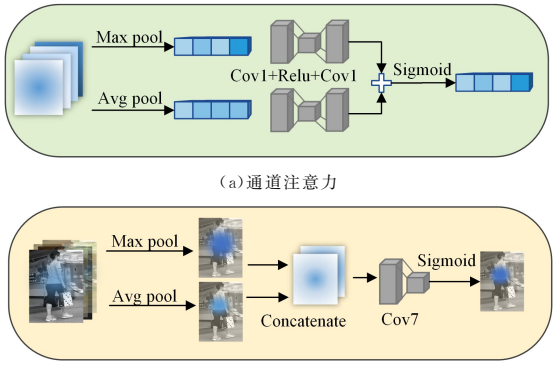


图 4 通道注意力及空间注意力结构

Fig. 4 Structure of channel attention and spatial attention

3.3 基于人体姿态估计和多目标跟踪的联合训练

由于多目标跟踪数据集 MOT17 不包含人体关键点标签，如果要同时训练模型输出目标跟踪信息和人体姿态估计信息，则只能使用 PoseTrack[6] 数据集，但该数据集是由众多短视频片段组成的，且视频内密集行人较少，无法有效提升模型的多目标跟踪性能。COCO 数据集包含 6 万 4 千张包含人体的图像，其中有大量图像包含人体姿态估计标签，但该数据集依旧是目标检测数据集，不包含连续的图像，因此使用该数据集训练的多目标跟踪网络性能较差。如果能同时使用 MOT17 和 COCO 数据集进行训练，则能避免网络的多目标跟踪性能下降，同时能输出高质量的人体姿态估计信息。

但同时训练两个数据集存在一些问题，首先是两个数据集提供的标签不同，在 COCO 上无法训练多目标跟踪所需的运动估计，在 MOT17 上无法训练人体姿态估计；其次是数据分布不同，MOT17 数据由多条长视频组成，而 COCO 是互不相关的 6 万张图片组成，其数据分布差异较大且图片数量较多，这就会导致模型的批规范化层(Batch Normalization)出现显著波动，导致训练出现问题。

本文通过 MOT17 数据集和 COCO 数据集进行多目标跟踪联合训练，通过设计训练过程和损失函数计算，在保证模型完成人体姿态估计的同时，大幅提升了其多目标跟踪的能力，拓展了多目标跟踪任务的使用场景。

为了解决多目标跟踪算法所需的运动估计问题，在 COCO 数据集上使用了对应位置随机裁剪的数据增强策略，通过 1 张原图和 2 张裁切方向相反的生成图片，为模型的双向运动估计提供运动信息。为了平衡 COCO 数据集和 MOT17 数据集目标的运动差异，同时保证双向匹配跟踪算法中搁浅区的可用性，MOT17 数据集采取以当前帧为中心，向前取 3 帧获得未来运动向量，向过去方向取 1 帧获得过去运动向量，实现了联合的运动估计训练，具体策略如图 5 所示。

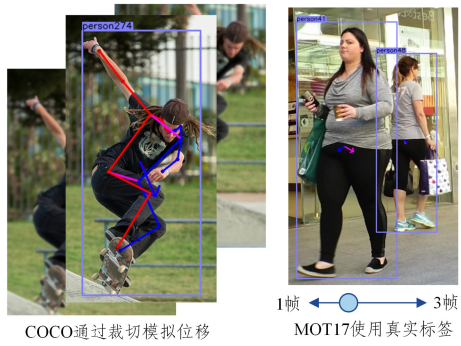


图 5 联合运动估计的标签生成

Fig. 5 Label generation for joint motion estimation

为了解决 MOT17 数据集没有人体关键点数据的问题，受到 PermaTrack[23] 的启发，将损失函数进行分别计算，从而在没有人体关键点的 MOT17 数据集中不监督人体关键点输出头提供的关键点信息。对于 COCO 数据集，其损失函数如式(3)所示，其中目标的关键点热值图 L_{hp} 和中心点热值图 L_{cent} 使用 Focal Loss[24] 损失函数，关键点的两类偏置 $L_{hp\ offset}$ 和 $L_{cent\ offset}$ 使用 SmoothL1 Loss[25] 损失函数。在 MOT17 的训练过程中，模型依旧会输出关键点预测，但 L_{hp}，$L_{hp\ offset}$ 和 $L_{hphm\ offset}$ 损失将归零处理。在训练过程上，COCO 数据集和 MOT17 数据集交替训练，其输入图像尺寸、批次数量、学习率均保持一致。

$$L_{sum}=L_{cent}+L_{cent\ offset}+0.1L_{wh}+L_{past}+0.5L_{future}+L_{ph}+L_{hp\ offset}+L_{hphm\ offset} \tag{3}$$

4 实验

4.1 评价指标及实验环境

为了准确地评价联合人体姿态估计的多目标跟踪算法的性能，将会使用 MOT 系列挑战赛和 COCO 人体关键点挑战赛的评价标准，分别评价模型的多目标跟踪性能和人体姿态估计性能。多目标跟踪任务将选择 MOTA，IDF1，ID Switch，Fragmentation，Most Track 和 Most Lose 这 6 个评价标准。人体姿态估计的评价将选择人体关键点的 AP 和 AR，其中包含不同关键点相似性阈值的成绩以及不同尺度下的成绩。

对于多目标跟踪任务，MOTA(多目标跟踪精度)的计算方法如式(4)所示，其中 t 是帧指数，FN 是漏掉的目标数，FP 是假阳性目标数，GT 是地面真实目标数。现有的数据集中发生遮挡的次数要远少于目标总数，在不改变匹配策略的基础上，通过使用更强的检测算法能够减少漏检，从而提升 MOTA 分数。因此 MOT 数据集使用了多个评价指标去综合判断算法的多目标跟踪性能。

$$MOTA=1-\frac{\sum^{t}(FN_t+FP_t+IDs_t)}{\sum^{t}GT_t} \tag{4}$$

IDF1 是跟踪 ID 的 F1 得分。它的计算方法如式(5)所示,其中 IDTP 表示正确匹配 ID 的数量,而 IDFP 和 IDFN 分别表示不正确匹配和未匹配 ID 的数量。

$$IDF1=\frac{2\times IDTP}{2\times IDTP+IDFP+IDFN} \tag{5}$$

IDs(身份转换)表示身份转换的次数,FM(Fragmentation)表示跟踪轨迹被中断的次数,MT(Mostly Tracked)表示大部分被跟踪的目标与地面真实轨迹总数的比率。如果一个目标在其生命期中至少有 80%的时间被成功追踪,那么它就是被大部分追踪的。ML(Mostly Lost)表示大部分丢失的目标与地面真实轨迹总数的比率。如果一条轨迹只恢复了其总长度的 20%以下,则称其为大部分丢失。

若目标检测到的位置与真实目标位置足够近,则被判定跟踪成功,否则跟踪失败(FN),对于每条轨迹,其跟踪成功的部分就是正例(TP),跟踪失败的就是负例(FN)。当一个目标没有被任何轨迹包含在内时,被认为是发生了断联(FM),当跟踪的轨迹发生了变化,则认为目标发生了身份切换(IDs)。

对于人体姿态估计,其核心的平均指标是目标关键点相似度(OKS),这是真值与预测值的关键点间欧几里得距离经过归一化后得到的成绩,并且包含了目标边界框面积和关键点难度两种归一化因子。除了基础的 AP 和 AR,即平均精度和平均召回率,还可以通过调整不同的 OKS 阈值,得到高于 0.5 或 0.75 OKS 的 AP 和 AR 值,OKS 越接近 1 就表明预测的点和真值越接近。通过筛选目标边界框的大小,就能获得中等(M)或较大目标(L)下的 AP 和 AR,评价算法对不同尺度目标的检测性能。通过以上方法就能组合出 10 个评价指标,评价结果如表 1 所列。

实验使用 MOT17 数据集和 COCO 人体关键点跟踪数据集作为训练集,其中 MOT17 将被切分为训练集和验证集。COCO 数据集中包含 118287 张图像,其中 64115 张带有人类标签的图像被用作训练集。MOT17 包含 5316 张图像共 7 段视频,其中 2664 张被用作训练集,2654 张被用作验证集,训练集和验证集都包含 7 段视频内容。

改进的模型基于 Pytorch1.7 框架,Python 版本为 3.7,实验训练使用了 NVIDIA RTX 3090 GPU,Intel 8700K CPU,没有使用预训练权重和额外数据集,MOT17 训练 1 轮时间为 1 min,COCO 训练 1 轮为 28.2 min,共循环训练 60 次。优化器采用了 Adam,学习率设为 2.5×10^{-4} 和 1.25×10^{-4},第 30 次循环训练减小了学习率,批次大小为 24,输入图像大小为 512×512,Focal Loss 的 α 和 β 设为 2 和 4。该算法只使用了随机裁切作为数据增强,并用该方式在目标检测数据集中模拟目标位移。

4.2 注意力机制的效果分析

为了较为公平地衡量注意力机制的效果,模型的训练过程将不使用任何预训练权重。首先对比分析极化自注意力在人体姿态估计任务上的效果,模型采用带有双向跟踪预测输出的多目标跟踪模型,上采样阶段统一使用普通卷积,在 COCO 训练集上分别进行 60 轮训练,在 COCO 验证集上进行测试,成绩如表 1 所列。

表 1 极化自注意力在人体姿态估计任务上的性能对比

Table 1 Performance comparison of polarized self-attention on human pose estimation tasks

(单位:%)

模型	AP	AP.5	AP.75	AP(M)	AR	AR.5	AR.75	AR(M)	AR(L)
无注意力	35.0	58.2	28.3	45.1	40.3	61.5	42.3	31.4	52.5
极化自注意力	**38.5**	**64.1**	**30.4**	**50.2**	**43.4**	**67.1**	**44.7**	**33.5**	**57.2**

从表 1 的结果来看,极化自注意力在人体姿态估计任务上取得了很好的效果,在所有指标上都取得了较大程度的领先,从而使得多目标跟踪算法针对人体姿态估计任务有了针对性的优化。为了进一步衡量不同注意力机制对多目标跟踪任务的影响,进一步的实验在 MOT17 训练集上进行了训练,研究了不同注意力机制对多目标跟踪算法训练速度的影响,成绩如表 2 所列。

表 2 不同注意力方法的参数量及训练速度比较

Table 2 Comparison of the number of participants and training speed of different attention methods

模块类型	模型参数量	1 轮训练时间/s
普通卷积	19.68×10^6	55
可形变卷积	20.32×10^6	110
注意力上采样	19.76×10^6	55
双注意力模块	21.22×10^6	57

由于可形变卷积不易并行计算,因此其训练时间较长,使用注意力上采样模块替代可形变卷积能够大幅提升训练速度,且不增加过多的模型参数。极化自注意力下采样模块的加入在有效提升算法性能的同时,并没有过多增加模型参数和训练时间,使得模型的训练更加高效。

为了进一步展示注意力模块在多目标跟踪任务中的效果,通过 60 轮 MOT17 进行训练,在不使用预训练的基础上对比其性能的变化,结果如表 3 所列。可以发现,注意力上采样主要提升的是模型的训练速度,而极化自注意力则能更好地提升模型性能。

表 3 使用注意力模块的消融实验

Table3 Ablation experiments using attention module

使用模块	MOTA ↑%	IDF1 ↑%	IDs ↓%	FM ↓%	MT ↑%	ML ↓%
无注意力	38.1	46.1	0.7	1.1	17.7.	40.4
十注意力上采样	37.4	47.5	0.7	1.0	16.2	**36.9**
十极化自注意力	**40.1**	**50.8**	**0.6**	**0.8**	21.2	38.6

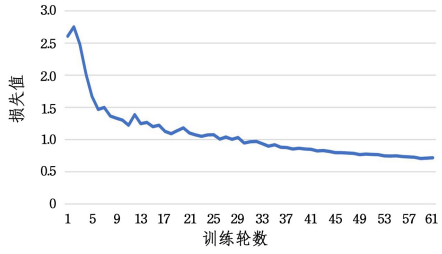
4.3 模型训练的损失分析

跨数据集学习需要保证模型在训练的过程中能够稳定地收敛,而本文的联合训练使用的是循环训练方式,因此通过比较每轮 MOT17 数据集和 COCO 数据下相同模型的损失值,就能较好评价训练过程的稳定性和发现其可能存在的问题。首先是分析总损失的变化,具体结果如图 6 所示。

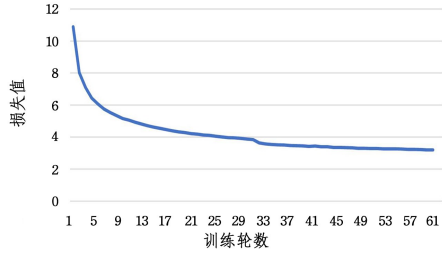
从总损失来看,两个数据集的训练结果都为下降趋势,没有出现损失不下降或损失异常增加的情况,证明了联合训练过程的有效性。但 MOT 数据集的损失波动性显著高于 COCO,这证明 COCO 较大的数据量对 MOT 数据集产生了不利影响,数据集的自身数据分布的确会影响训练过程。从损失值来看,由于没有预训练权重,在模型训练的前 5 轮损失值要

显著高于其他阶段，因此后续基于损失的分析将从第 6 轮开始分析。

图 7 给出了在 COCO 数据集训练时，其人体姿态估计相关损失的变化趋势。从损失值变化趋势来看，关键点热值图的损失非常平滑，但关键点的偏移预测略有波动，证明基于目标位置的预测不容易受到联合训练的影响。总的来看，联合训练过程在损失层面保持了较好的稳定性，相同的模型在两个数据集间循环训练，能够保证模型学习到目标的位置信息和人体关键点信息。



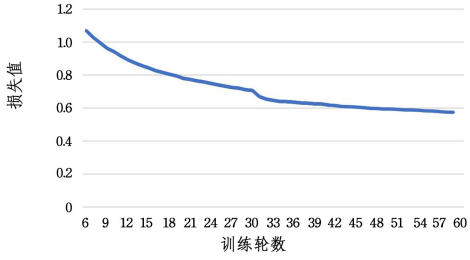
(a) MOT



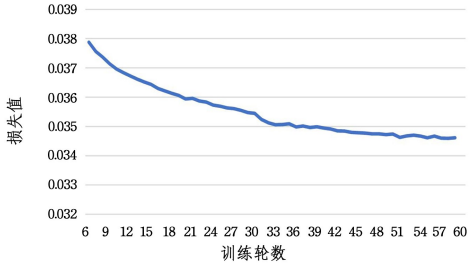
(b) COCO

图 6 跨数据集训练总损失的变化趋势

Fig. 6 Cross-dataset training loss variation trend



(a) COCO HM HP



(b) COCO HM Offset

图 7 COCO 数据集中训练关键点检测热值图和偏置损失

Fig. 7 Training key-point detection heat map and bias loss in COCO dataset

4.4 模型训练的损失分析

多目标跟踪算法 CenterTrack[7] 曾使用 COCO 数据集进行人体关键点检测训练，也通过随机裁切生成了单向运动估计标签，因此是本文较为合适的对比对象，其结果如表 4 所列。

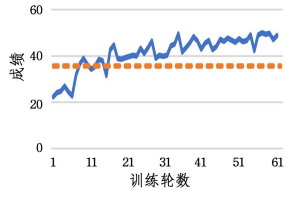
表 4 CenterTrack 与联合检测的多目标跟踪算法的性能对比

Table 4 Performance comparison of CenterTrack and joint detection for multi-object tracking algorithms

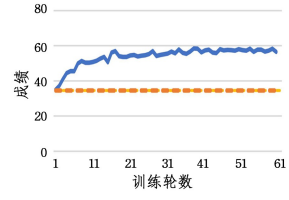
算法	MOTA ↑ %	IDF1 ↑ %	IDs ↓ %	FM ↓ %	MT ↑ %	ML ↓ %
CenterTrack	35.6	34.5	2.6	2.7	21.1	38.9
本文算法	**49.1**	**558.3**	**0.6**	**1.2**	**31.3**	**26**

可以看出，本文算法在所有评价指标上都获得了显著的优势，证明了本文算法的联合训练模式的有效性。实际在训练早期，本文算法就超越了只使用 COCO 训练的 CenterTrack 算法。图 8 给出了了算法成绩的变化过程。

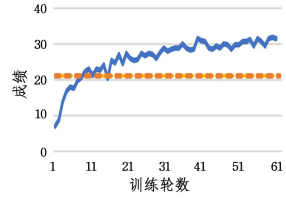
如图 8 所示，虚线代表了 CenterTrack 算法的成绩，在从零训练的第 8 轮，本文算法就完成了对该模型的全面超越。值得注意的是 IDF1 的成绩，在第一轮就领先于 CenterTrack 算法，证明了双向运动向量需要真实的目标运动信息。并且持续的联合训练能大幅提升模型的多目标跟踪性能，进一步证明了本文算法和联合训练策略的有效性。图 9 给出了本文算法在 MOT17 测试集上的实际效果。



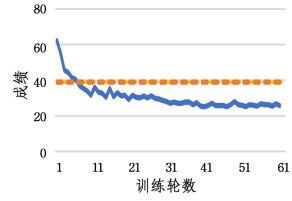
(a) MOTA



(b) IDF1



(c) Most track



(d) Most lsot

图 8 联合训练算法与 CenterTrack 基线成绩对比

Fig. 8 Comparison of joint training algorithm and CenterTrack baseline scores

图 9 MOT17 联合人体姿态估计的多目标跟踪效果

Fig. 9 Multi-object tracking effect of MOT17 joint human pose estimation

结束语 本文在双向跟踪匹配算法的基础上进行研究,使用了极化自注意力优化了模型的下采样过程,为人体姿态估计提供了更精细的语义信息,同时使用注意力上采样模块替换可形变卷积,提升了模型的训练速度,接着使用改进的模型进行联合训练,循环使用 COCO 数据集和 MOT17 数据集进行训练,同时完成多目标跟踪任务和人体姿态估计任务,通过改进数据增强策略和损失计算过程,保证了联合训练的顺利执行。通过 COCO 和 MOT17 数据集的训练测试,验证了极化自注意力模块的有效性;通过与 CenterTrack 模型的对比,展示出了优化后的模型在使用联合训练过程后的强大性能,全面地超越了使用单一数据集训练的联合人体姿态估计的多目标跟踪算法。基于多目标跟踪算法提供的短期运动预测和人体姿态估计信息,未来可以进一步研究行人的长期轨迹跟踪预测,从而使多目标跟踪算法能在长期遮挡的环境下做到更优的跟踪性能。

参 考 文 献

[1] MILAN A,LEAL-TAIXÉ L,REID I,et al. MOT16:A benchmark for multi-object tracking[EB/OL] arXiv:1603.00831,2016,Accessed:Aug. 23,2021. https://arxiv.org/abs/1603.00831v2.

[2] DENDORFER P,REZATOFIGHI H,MILAN A,et al. Mot20:A benchmark for multi object tracking in crowded scenes[EB/OL]. arXiv:2003.09003,2020. http://arxiv.org/abs/2003.09003.

[3] SHAO S,ZHAO Z,LI B,et al. Crowdhuman:A benchmark for detecting human in a crowd[EB/OL]. arXiv:1805.00123,2018. http://arxiv.org/abs/1805.00123.

[4] LIN T Y,MAIRE M,BELONGIE S,et al. Microsoft coco:Common objects in context[C]// European Conference on Computer Vision. Cham:Springer,2014:740-755.

[5] LIANG J,JIANG L,NIEBLES J C,et al. Peeking into the future:Predicting future person activities and locations in videos [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:5725-5734.

[6] GIRDHAR R,GKIOXARI G,TORRESANI L,et al. Simple,efficient and effective keypoint tracking[C]// ICCV PoseTrack Workshop. 2017.

[7] ZHOU X,KOLTUN V,KRÄHENBÜHL P. Tracking objects as points[C]// European Conference on Computer Vision. Cham:Springer,2020:474-490.

[8] LIU H,LIU F,FAN X,et al. Polarized self-attention:Towards high-quality pixel-wise regression [EB/OL]. arXiv:2107.00782,2021. http://arxiv.org/abs/2107.00782.

[9] WOJKE N,BEWLEY A,PAULUS D. Simple online and realtime tracking with a deep association metric[C]// 2017 IEEE International Conference on Image Processing(ICIP). IEEE,2017:3645-3649.

[10] ZHOU X,WANG D,KRÄHENBÜHL P. Objects as points [EB/OL]. arXiv:1904.07850,2019. http://arxiv.org/abs/1904.07850.

[11] CAO Z,SIMON T,WEI S E,et al. Realtime multi-person 2d pose estimation using part affinity fields[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:7291-7299.

[12] HIDALGO G,RAAJ Y,IDREES H,et al. Single-network whole-body pose estimation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:6982-6991.

[13] FANG H S,XIE S,TAI Y W,et al. Rmpe:Regional multi-person pose estimation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:2334-2343.

[14] PAPANDREOU G,ZHU T,KANAZAWA N,et al. Towards accurate multi-person pose estimation in the wild[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:4903-4911.

[15] HU J,SHEN L,SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7132-7141.

[16] WOO S,PARK J,LEE J Y,et al. Cbam:Convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018:3-19.

[17] DAI J,QI H,XIONG Y,et al. Deformable convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:764-773.

[18] QI T,BAYRAMLI B,ALI U,et al. Spatial shortcut network for human pose estimation[EB/OL]. arXiv:1904.03141,2019. http://arxiv.org/abs/1904.03141.

[19] YAO Y,WANG Y,GUO Y,et al. Cross-dataset training for class increasing object detection[EB/OL]. arXiv:2001.04621,2020. http://arxiv.org/abs/2001.04621.

[20] PERRETT T,DAMEN D. Recurrent assistance:cross-dataset training of LSTMs on kitchen tasks[C]// Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017:1354-1362.

[21] WANG L,LI D,LIU H,et al. Cross-Dataset Collaborative Learning for Semantic Segmentation in Autonomous Driving [EB/OL]. arXiv:2103.11351,2021. http://arxiv.org/abs/2103.11351.

[22] WANG X,CAI Z,GAO D,et al. Towards universal object detection by domain attention[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:7289-7298.

[23] TOKMAKOV P,LI J,BURGARD W,et al. Learning to track with object permanence[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:10860-10869.

[24] LIN T Y,GOYAL P,GIRSHICK R,et al. Focal loss for dense object detection[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:2980-2988.

[25] GIRSHICK R. Fast r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:1440-1448.

ZENG Zehua, born in 1995, postgraduate. His main research interests include multi-object tracking and human pose estimation.

LUO Huilan, born in 1974, Ph.D, professor. Her main research interests include machine learning and pattern recognition.