计算机科学
COMPUTER SCIENCE

基于自适应预测的2D人体姿态估计

郑泉石, 金城

引用本文

郑泉石, 金城. 基于自适应预测的2D人体姿态估计[J]. 计算机科学, 2023, 50(11A): 221000048-7.

ZHENG Quanshi, JIN Cheng. 2D Human Pose Estimation Based on Adaptive Estimation [J]. Computer Science, 2023, 50(11A): 221000048-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000048

基于自适应预测的2D人体姿态估计

郑泉石[1]　金　城[1,2]
1 复旦大学计算机科学技术学院　上海 200438
2 鹏城实验室　广东 深圳 518066
(qszheng20@fudan.edu.cn)

摘　要　基于回归的2D人体姿态估计方法直接预测人体关键点的2D坐标,是主流的2D姿态估计方法之一。Transformer能有效建立人体部位间的关系,它的应用显著提升了回归方法的准确率。然而相关方法存在以下两个问题:1)在交叉注意力模块中,对于不同图像,固定的Query值难以准确关注到不同的关键点区域,导致注意力分散;2)直接学习关键点的标注位置,导致模型过拟合于训练集的标注,泛化性差。文中提出了一种基于自适应预测的姿态估计模型来解决以上问题。针对第一个问题,该模型自适应地预测Query的关注区域,并引导注意力集中于该区域。针对第二个问题,该模型自适应地预测关键点在所有位置上出现的可能性分布,通过软预测的方式,缓解模型对标注的过拟合。在MS-COCO数据集上的实验表明,该模型将基线方法的准确率提升了2.8%,并将相关方法的最高准确率提升了0.2%。

关键词:2D人体姿态估计;回归;自适应;关注区域;可能性分布
中图法分类号　TP391

2D Human Pose Estimation Based on Adaptive Estimation

ZHENG Quanshi[1] and JIN Cheng[1,2]
1 School of Computer Science,Fudan University,Shanghai 200438,China
2 Peng Cheng Laboratory,Shenzhen,Guangdong 518066,China

Abstract　The regression-based 2D human pose estimation methods directly predict the coordinates of human keypoints. The transformer can effectively establish the relationship between human body parts,and its application significantly improves the accuracy of the regression-based methods. However,related methods have the following two problems:1) In the cross-attention module,for different images,the fixed query can not properly focus on different keypoint regions,which leads to distraction. 2)They directly learn the labeled keypoint coordinates and overfit annotations. In this paper,a pose estimation model based on adaptive prediction is proposed to solve these two problems. For the first problem,the model adaptively predicts the region of attention of the query and directs the attention to that region. For the second problem,the model adaptively predicts the probability distribution of keypoint appearing in every position,and alleviates the model's overfitting to annotations by means of soft prediction. Experiments on the MS-COCO dataset show that the model improves the accuracy of the baseline method by 2.8% and improves the highest accuracy of related methods by 0.2%.

Keywords　2D human pose estimation,Regression-based,Adaptive,Region of attention,Probability distribution

1　引言

2D人体姿态估计[1-2]是计算机视觉领域的研究热点,是计算机从图像数据中理解人类行为的基础。它是行为识别、人体分析等任务的上游任务,被广泛应用于人机交互、智能监控等场景。2D人体姿态估计从RGB图像中检测和定位人体关键点(头部、肩部、膝盖、脚踝等),传统的2D人体姿态估计方法基于概论图[3]、树模型[4]以及随机森林[5]。在2013年之后,基于深度学习的方法被越来越多地应用于2D姿态估计领域。2D姿态估计方法逐渐分为基于热力图和基于回归的方法。基于热力图的方法为每个关键点生成一个似然热力图,并通过argmax从热力图中获得关键点位置。基于热力图的方法准确率高,但有以下缺点:1)热力图标签和启发式的后处理操作带来了设计挑战以及归纳偏置[6];2)热力图分辨率低于输入图像,将热力图坐标转换为图像坐标存在量化误差[7]。基于回归的方法直接预测输入图像中人体关键点的坐标,消除了热力图坐标转化带来的量化误差,同时避免了不可微的热力图后处理步骤,以及它带来的归纳偏置,整个过程可以进行端到端训练,具有充足的优化潜力。

以往基于回归的方法忽略了人体部位间的关系,单独地对每个关键点进行预测,造成预测的准确率低。Transformer[8]中的自注意力机制能有效捕捉人体部位间的潜在

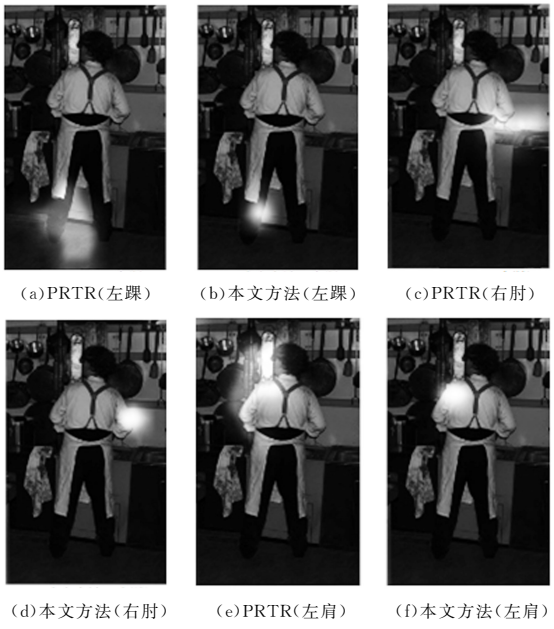
基金项目:上海市科技创新行动计划(22dz1204900)
This work was supported by the Shanghai Municipal Science and Technology Commission(22dz1204900).
通信作者:金城(jc@fudan.edu.cn)

关系，并利用它预测人体关键点的位置，提升了预测准确率。

PRTR[9]首次将 Transformer 应用于基于回归的姿态估计方法，显著提升了回归方法的性能。PRTR 设置一组 Query(数量为 Q)并通过交叉注意力模块提取图像信息，然后通过匈牙利算法(Hungarian Algorithm)进行最优二部图匹配(Bipartite Matching)，从 Q 个 Query 中选择 N 个关键点向量，并以此预测出 N 个人体关键点的坐标。相关方法将 Transformer 应用于基于回归的姿态估计方法时存在注意力分散与过拟合的问题。

相关方法中的 Query 被设置为固定值，面对不同输入，Transformer 的注意力难以正确关注到关键点区域，引入无关信息干扰了对人体关键点的预测。

针对此问题，本文提出将自适应地预测关键点的大致位置作为兴趣点，并将兴趣点位置信息嵌入 Query 中，引导注意力对兴趣点区域产生高响应，最终克服注意力分散的缺陷。如图 1 所示，相关方法的注意力均匀地分散在大片区域上，导致关键点区域的信息被大量背景信息掩盖，而本文提出的方法能让注意力高度集中在关键点区域。



(a)PRTR(左踝) (b)本文方法(左踝) (c)PRTR(右肘)

(d)本文方法(右肘) (e)PRTR(左肩) (f)本文方法(左肩)

图 1 左踝、右肘、左肩关键点注意力图对比

Fig. 1 Comparison on attention maps for the left ankle, right elbow, and left shoulder

相关方法直接学习关键点坐标的标注值，容易过拟合于训练集的标注。此外，关键点不可能出现的位置作为负样本，与可能出现的位置(正样本)同样具有重要的信息。相关方法只关注关键点最可能出现的单个位置，这造成了相关方法关键点预测精度的下降。

针对这一问题，本文提出自适应地预测关键点在所有位置上出现的可能性分布。相比于相关方法 one-hot 式的硬预测，可能性分布是一种软预测，其作为一种正则化方式，降低了单个位置在损失计算时的权重，缓解了对标注的过拟合现象。模型对整体位置的预测也使得模型具有了关键点的全局信息，能够帮助模型更好地预测不同输入中的关键点位置。

本文的主要贡献如下：

1)本文提出的方法自适应地预测兴趣点，并引导注意力高度集中在兴趣点区域，使得注意力高度集中在关键点区域上，减少无关背景信息的引入，提高了特征提取效果。

2)本文提出的模型自适应预测关键点分布，将单个位置的 one-hot 预测软化为区域的可能性预测，缓解了模型对标注的过拟合情况，提升了模型的泛化性。

3)本文提出的方法将基线方法的准确率提升了 2.8%，并将基于回归的姿态估计方法的最高准确率提高了 0.2%。

2 相关工作

基于回归的姿态估计方法直接预测关键点的坐标，是主流的姿态估计方法之一。Directpose[10]直接从全局图像特征中回归出所有人体关键点坐标。单个特征向量难以直接预测高精度的关键点坐标，Integral[11]提出集成回归的方法，生成关键点特征图作为中间结果，综合整个特征图的信息预测关键点坐标，提升了预测的准确率。RLE[12]认为之前的方法预设了关键点的密度函数，使模型预测的关键点分布与真实情况不符。它从数据中学习关键点分布的密度函数，进一步提升了回归方法的效果。

PRTR 应用了级联 Transformer 的结构。自注意力建模图像全局关系的能力强，它提供了一个有效的方法捕捉人体部位间潜在的联系，显著提升了回归方法的性能。PRTR 将图像输入映射为检测对象的集合，即给定一个输入图像，模型的预测结果是一个包含所有检测对象的无序集合。PRTR 使用二部图匹配策略从检测对象集合中选择出关键点对象子集。关键点对象随后生成对应的关键点类别与坐标信息。

在此基础上，Conditional-DETR[13]认为使用级联 Transformer 结构的模型训练周期过长，提出将交叉注意力模块中的 Query 与图像特征嵌入到相同的空间中，使交叉注意力模块更容易通过 Query 提取图像特征，减少训练该模块的难度，缩短了模型训练时间。

本文基于文献[9-13]的工作，提出了一种基于回归方法的 2D 姿态估计模型。它通过引导注意力集中在兴趣点区域的方式，解决了特征抽取过程中注意力分散的问题，提升了特征抽取的效果。该方法预测人体尺寸作为关键点分布的方差，在计算预测误差考虑了人体尺寸的影响，使误差计算更准确。通过以上两方面的改进，本文进一步提升了基于回归的姿态估计方法的效果。

3 基于自适应预测的姿态估计网络

3.1 网络结构

本文提出了一个端到端的姿态估计模型。如图 2 所示，模型由 CNN 骨干网络、Transformer Encode、Transformer Decoder、坐标及类别预测头和损失函数 5 部分组成。Encode 与 Decoder 由多层 Encode Layer 或 Decoder Layer 级联组成。本文主要改进 Transformer Decoder 和损失函数。

CNN 骨干网络初步提取输入图像的特征图，特征图随后被展平为像素向量的集合，像素向量集合被送入 Encoder。像素向量作为 Encoder 中注意力模块的 Key 和 Value，一组兴趣点向量与像素向量连接起来作为 Query。Encoder 在计算图像特征的自注意力时，同时计算兴趣点向量与图像特征间的交叉注意力。通过这样的方式，使得图像特征具有上下文信息，同时兴趣点向量也具有了图像信息。接下来，

Encoder 输出具有上下文信息的像素向量集合和具有图像信息的兴趣点向量集合。像素向量作为 Decoder 的 Value 和 Key，兴趣点向量作为位置部分 PQ(Position Query)与输入的内容部分 CQ(Content Query)拼接得到 Query。模型通过 Decoder 中的交叉注意力模块抽取图像信息。Decoder 输出具有图像信息的目标向量，并通过二部图匹配从目标向量中选择特定的关键点向量。关键点向量最后通过类别及坐标预测头预测关键点类别以及关键点分布的期望值和方差。在训练阶段，关键点分布由自适应损失监督。在推理阶段，关键点分布的期望值被当作预测坐标，1-方差被当作预测置信度。

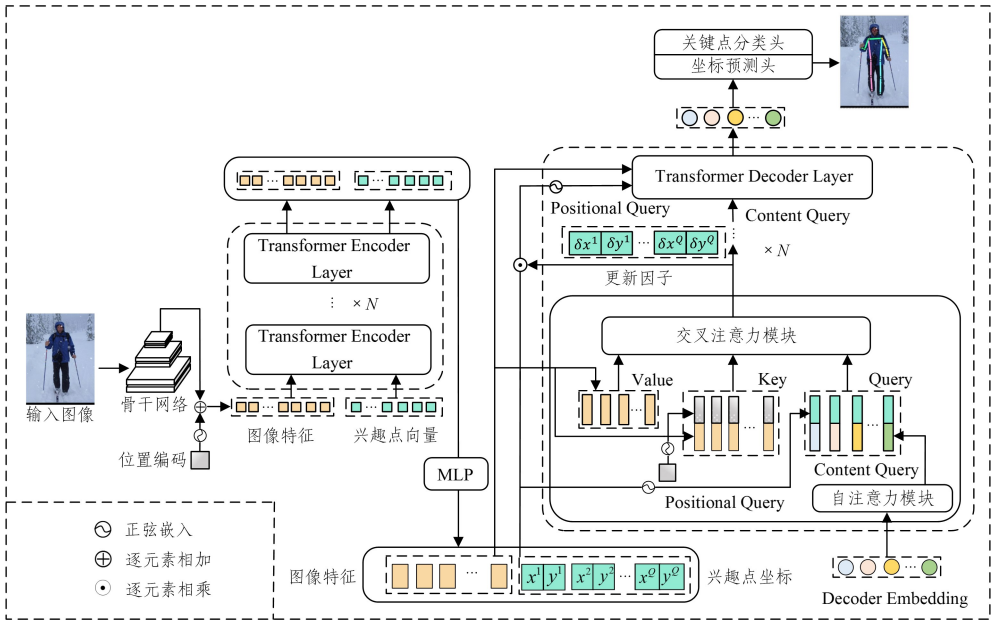


图 2　模型整体框架

Fig. 2　Overall model framework

3.2　自适应特征抽取

在 Transformer Decoder 的交叉注意力模块中，Key 由图像特征图上所有像素的特征及其位置信息组成，Query 与 Key 的点积决定它对每个像素的注意力，Query 以注意力值作为权值聚合特征图像素的信息。Query 与 Key 一样，也是由位置部分 PQ(Position Query)和内容部分 CQ(Content Query)组成的。相关工作将 PQ 设置为固定值，不能为 Query 的注意力提供位置信息上的引导，导致了 Query 的注意力分散。DETR[14] 中的实验证明，移除固定的 PQ，仅会对 Transformer 性能造成轻微影响。造成这种现象的原因有两个：1)PQ 是固定参数；对于不同的输入，它们很难关注到对应的关键点区域；2)PQ 中的空间信息没有经过正弦编码，而 Key 中的空间信息经过了正弦编码，这导致它们的空间信息不在同一空间。在注意力计算中，PQ 难以对相应的区域产生高响应。

本文提出的模型粗略预测关键点的位置，作为注意力所关注的兴趣点。作为 Transformer 的输入，特征图被展平为像素的集合，为了保留其位置信息，每个像素会加上其 2D 坐标的正弦编码。本文按照特征图位置信息的生成方式，将兴趣点的坐标正弦编码作为 PQ。在计算点积时，PQ 会对特征图兴趣点附近的区域产生高响应。本文通过这种方式弥补了注意力分散的缺陷。如图 3 所示，PRTR 的 PQ 注意力几乎平均分散在所有图像位置上，导致 Query 注意力分散在大片的区域上。本文提出的方法使得 PQ 对关键点区域产生高响应，引导注意力高度集中在关键点区域。

(a)PRTR-PQ 注意力　(b)PRTR-Query 注意力　(c)本文方法-PQ 注意力　(d)本文方法-Query 注意力

图 3　右膝关键点的 PQ 注意力及 Query 注意力对比

Fig. 3　Comparison of positional query attention and query attention for the right knee

为了自适应地生成 PQ，本文设定了一组兴趣点向量集合 $\boldsymbol{R}\in\{\boldsymbol{R}_1,\cdots,\boldsymbol{R}_M\}$，$\boldsymbol{R}_i\in\mathbb{R}^D$，$\boldsymbol{R}$ 与图像特征向量连接作为 Query 输入 Encoder 中。Encoder 在计算图像特征的自注意力时，同时计算兴趣点向量与图像特征间的交叉注意力。Encoder 通过这样的方式建模了图像特征的全局关系，同时使兴趣点向量具有图像信息。接下来，Encoder 输出具有上下文信息的像素向量集合和具有图像信息的兴趣点向量集合。兴趣点向量集合通过位置预测头得到兴趣点的坐标 $P\in$

$\{P_1,\cdots,P_M\}$，$P_i\in\mathbb{R}^2$。这个过程如图 4 所示。

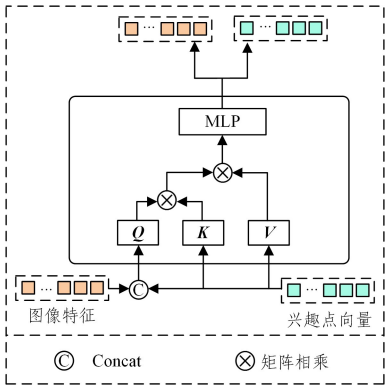


图 4 兴趣点初始化

Fig. 4 POI initialization

Conditional-DETR 将 Query 与图像特征投影到相同的空间中，使 Query 更好地提取图像特征。本文受其启发，将兴趣点的位置(x_{ip},y_{ip})正弦嵌入为 $PE(x_{ip},y_{ip})$，并将 $PE(x_{ip},y_{ip})$与 CQ 拼接起来。在交叉注意力模块中的每个 Query，Key 和 Value 可以被描述为：

$$Query=Concat(q_{\text{content}},PE(x_{ip},y_{ip})) \tag{1}$$

$$Key=Concat(F_{x,y},PE(x,y)) \tag{2}$$

$$Value=F_{x,y} \tag{3}$$

其中，$Concat$ 是拼接操作，q_{content}是 CQ，$F_{x,y}$是坐标为(x,y)位置上的像素向量，$PE(x_{ip},y_{ip})$是兴趣点的位置信息，$PE(x,y)$是像素向量的位置信息。

经过以上步骤，Query 与 Key 一样，都拥有了由二维坐标生成的位置信息。同时，它们的位置信息都经过了正弦嵌入，处于同一个嵌入空间中。在注意力计算中，PQ 将对图像中兴趣点位置周围的位置产生高响应，如图 3 所示，Query 的注意力被引导到相应的区域上并高度关注这一区域。

最初的兴趣点位置是粗略估计出来的，兴趣点的位置会持续更新。具体来说，每一层的 Decoder Layer 会输出兴趣点坐标的更新因子，持续地更新兴趣点的位置，这一过程可以被描述为：

$$(\delta x,\delta y)_i=SRH(\text{output}_{i-1}) \tag{4}$$

$$(x_i^{ip},y_i^{ip})=(x_{i-1}^{ip},y_{i-1}^{ip})\odot(\delta x,\delta y)_i \tag{5}$$

其中，$\odot$是逐元素的乘法运算，output_{i-1}是第 $i-1$ 层 Decoder Layer 的输出。$(x_{i-1}^{ip},y_{i-1}^{ip})$，$(x_i^{ip},y_i^{ip})$分别是第 $i-1$ 层与第 i 层 Decoder Layer 的兴趣点坐标，$(\delta x,\delta y)_i$ 是兴趣点坐标的更新因子。SRH 是更新因子的预测头，预测头将 D 维的 CQ 投影为 2 维向量，$SRH:\mathbb{R}^D\to\mathbb{R}^2$。

3.3 自适应损失

PRTR 使用 L1-Loss 计算关键点的预测损失：$|u_g-u|$，u_g 是关键点坐标的标注值，u 是关键点坐标的预测值。在这一过程中，模型直接学习关键点位置的标注值，造成模型过拟合于训练集的标注坐标，降低了模型的泛化性。

本文提出预测输入图像上的关键点分布，表示关键点在图像上的所有位置出现的可能性，将相关方法 one-hot 的硬预测软化为了区域上的可能性预测。本文方法通过这样的方式，一方面降低了单个位置在损失计算时的权重，另一方面也使得模型具有了关键点的全局信息，即关键点可能/不可能出现的位置。关键点不可能出现的位置作为负样本，与可能出现位置的正样本同样具有重要的信息，能够帮助模型更好地预测不同输入中的关键点位置。本文通过以上方式，提升了模型的泛化性。

具体来说，本文方法自适应地预测关键点可能性分布的期望值 u 与方差 σ。关键点分布的损失可描述为：

$$L=-\log P_\theta(x\mid I)|_{x=u_g} \tag{6}$$

其中，θ 是姿态估计模型的参数，I 是输入图像。$P_\theta(x\mid I)$是模型学习的关键点分布，代表关键点在图像不同位置上出现的可能性分布。

本文受 RLE 的启发，设置流模型来学习关键点分布的密度函数 $P_\varphi(\bar{x})$，其中 φ 是流模型的参数。具体来说，流模型学习了一个函数 $F_\varphi(\bar{z})$将一个标准高斯分布 $\bar{z}\sim N(0,I)$映射为期望值为 0、方差为 1 的复杂分布 $\bar{x}\sim G_\varphi(\bar{x})$。最终分布的密度函数 $P_\varphi(\bar{x})$由 $G_\varphi(\bar{x})$与标准拉普拉斯分布 $L(\bar{x})$相乘得到。密度函数 $P_\varphi(\bar{x})$是一个期望值为 0、方差为 1 的分布。本文方法预测分布的期望值 u 与方差σ，通过对 $P_\varphi(\bar{x})$进行位移与拉伸变换 $x=\bar{x}*\sigma+\mu$，得到关键点分布 $P_{\theta,\varphi}(x\mid I)$。

本文通过合成多个流模型的结果，进一步增强了函数 $F_\varphi(\bar{z})$的映射能力，更好地模拟了现实中复杂的关键点分布情况。关键点分布生成的过程如图 5 所示。本文首先设置 k 个流模型将标准高斯分布映射为 k 个复杂的单位分布，同时为每个分布设置一个可学习权重，然后通过加权和运算将 k 个复杂的单位分布合成为一个混合单位分布。自适应损失模块根据回归模型预测出的分布期望值与方差值，对单位混合分布进行位移与拉伸变换，最终得到关键点分布。这个过程可描述为：

$$G_{\varphi,W}(\bar{x})=w_1F_{\phi_1}(\bar{z})+\cdots+w_kF_{\phi_k}(\bar{z}) \tag{7}$$

$$P_\phi(\bar{x})=L(\bar{x})*G_{\varphi,W}(\bar{x}) \tag{8}$$

$$P_\phi(x)=P_\phi(\bar{x})*\sigma+\mu \tag{9}$$

其中，$F_{\phi_1}(\bar{z}),\cdots,F_{\phi_k}(\bar{z})$是 k 个流模型，$w_1,\cdots,w_k$ 是 k 个流模型对应的权重，它们被设置为可学习的参数；$L(\bar{x})$是标准拉普拉斯分布。综上，自适应损失可以描述为：

$$L_{\text{apt}}=\log\sigma-\log L(\bar{\mu}_g)-\log(w_1F_{\phi_1}(\bar{z}_g^1)+\cdots+w_kF_{\phi_k}(\bar{z}_g^k)) \tag{10}$$

$$\bar{\mu}_g=\frac{(u_g-u)}{\sigma} \tag{11}$$

$$\bar{z}_g^i=\sim F_{\phi_i}(\bar{\mu}_g) \tag{12}$$

在训练阶段，姿态估计模型与流模型可以同时得到优化。流模型仅用于训练阶段，不会增加姿态模型推理阶段的计算量。在推理阶段，μ 作为关键点坐标的预测值，$1-\sigma$ 作为坐标的预测置信度。

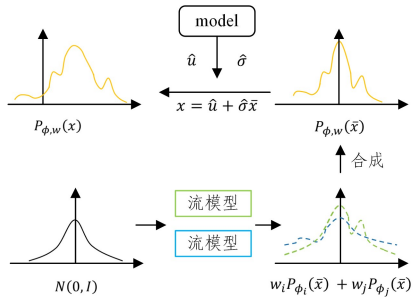


图 5 关键点分布生成

Fig. 5 Generation of keypoint distribution

4 实验

4.1 实验设置

本文在公开数据集 MS-COCO[15] 2017 数据集上验证方法的有效性。

本文使用关键点相似度(OKS)指标来评估模型的性能，它是最常用的衡量关键点预测质量的指标之一。

OKS 主要有 3 个衡量关键点预测准确率的指标，分别是 AP，AP_M，AP_L。OKS 设置了一个相似度阈值，如果某个关键点预测的相似度大于阈值则认为关键点被成功检测，反之则认为检测失败。AP 是多种阈值下结果的平均准确率，是 OKS 中最为重要的指标。AP_M 与 AP_L 表示在中小尺度与大尺度人体上的关键点检测准确率。

本文与 PRTR 中的训练设置保持一致，使用 AdamW 优化器[16]。Transformer 部分的参数用 Xavier[17] 初始化方案进行初始化。骨干网络和 Transformer 的初始学习率设置为 1×10^{-5} 和 1×10^{-4}，权重衰减为 1×10^{-4}，输入图像大小为 384×288。Transformer 部分的 encoder、decoder 层数和 Query 的数量分别被设置为 6，6 和 100。本文遵循惯例在 SimpleBaseline[18] 的人体检测结果上进行姿态估计，并与相关工作比较结果。在进行对照实验的比较时，我们同时在 SimpleBaseline 的人体检测结果和数据集提供的标注人体上进行姿态估计，比较改进前后的模型性能。

4.2 模型效果

本文采用 MS-COCO 2017 数据集部分图片作为输入，将预测结果可视化。人体关键点的预测效果如图 6 所示。

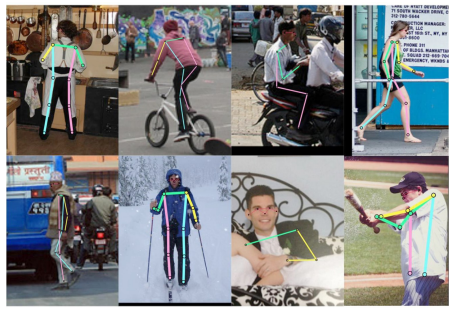


图 6 姿态估计模型效果示意图

Fig. 6 Schematic diagram of human pose estimation results

本文在 MS-COCO 2017 数据集上对比了所提方法与相关方法的效果。本文提出了 3 种模型 A，B，C。这 3 种模型的差别在于兴趣点位置信息的生成方式不同，模型的具体信息将在对照实验章节进行详细介绍。表 1 列出了具体的结果。本文提出的模型在 AP，AP_M，AP_L 指标上达到了 74.5%，71.3%，81.8%的关键点预测准确率，达到了 SOTA。在最重要的 AP 指标上，我们将基线方法的准确率提高了 2.8%，并将基于回归的姿态估计方法准确率提高了 0.2%，证明了本文方法的有效性。

在小中尺度(AP_M)和大尺度(AP_L)的人体上，我们将基线方法的准确率分别提高了 3.2%和 2.5%，并将相关方法的最高准确率分别提高了 0.6%和 0.8%，这证明了本文方法在多种人体条件下的泛化性。

表 1 与相关方法在 MS-COCO 2017 验证集上的准确率对比

Table 1 Accuracy comparison of the proposed method with related methods on MS-COCO 2017 validation set

Model	Backbone	Input Size	AP/%	AP_M//%	AP_L/%
CenterNet[19]	Hourglass-2 stacked	—	63.0	58.9	70.4
PointSetNet[20]	HRNet-W48	—	68.7	64.8	75.3
SPM[21]	Hourglass-8 stacked	384×288	66.9	63.6	73.1
DirectPose	ResNet-101	—	63.3	57.8	71.2
Integral	ResNet-101	—	67.8	63.9	74.0
RLE	HRNet-W32	384×288	74.3	70.7	81.0
PRTR	HRNet-W32	384×288	71.7	68.1	79.3
Our Method-A	HRNet-W32	384×288	73.9	70.5	81.3
Our Method-B	HRNet-W32	384×288	74.1	70.6	81.5
Our Method-C	HRNet-W32	384×288	**74.5**	**71.3**	**81.8**

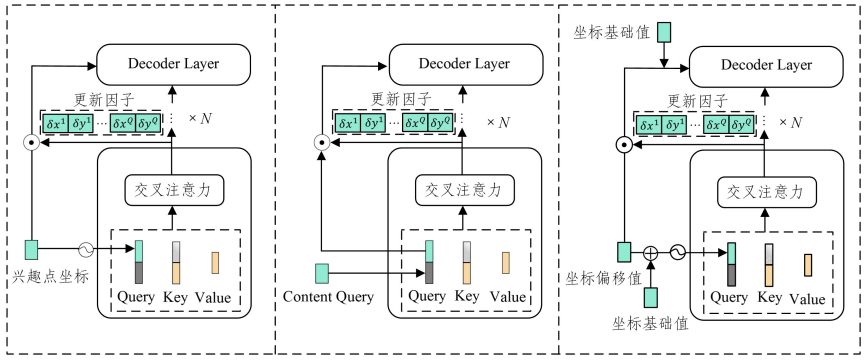
4.3 对照实验

在对照实验部分，本文验证了所提出的自适应特征提取模块和自适应损失模块的有效性，并进一步探索了不同流模型数量对模型性能的影响。

本部分实验在 MS-COCO 2017 数据集的验证集上进行。

4.3.1 兴趣点生成方式

本文探索了不同的兴趣点生成方式，设置了 3 种模型结构，对应 3 种不同的兴趣点生成方式，分别记为方法 A，B 和 C，如图 7 所示。



(a) Our Method-A (b) Our Method-B (c) Our Method-C

图 7 3 种兴趣点位置预测方法

Fig. 7 Three methods of POI position prediction

方法 A 与第 3 章中描述的方法一致，兴趣点向量通过预测头直接预测每个 Query 的兴趣点坐标，并将其正弦嵌入为 PQ。直接生成兴趣点坐标可能难度较大，方法 B 选择间接地生成兴趣点的位置。其将兴趣点坐标设置为可学习参数，并设置一组内容向量，通过 Encoder 将图像信息提取到内容向量中，并将内容向量作为 CQ，通过 CQ 预测兴趣点位置的更新因子。方法 B 通过更新兴趣点坐标来间接生成兴趣点的位置信息。方法 C 认为兴趣点位置分为所有输入共有的基础值部分，以及每个输入独有的偏移值部分，将兴趣点的坐标预测任务拆分以减少直接预测的难度。方法 C 将兴趣点坐标的基础值设置为可学习的参数，预测兴趣点坐标的偏移值。兴趣点的坐标由固定的基础值与预测偏移值相加得到。

如表 2 所列，方法 A，B，C 在标注人体（GT-BBOX）上分别达到 76.9%，77.1%和 77.6%的准确率，在 SimpleBaseline 检测人体（SIMBA-BBOX）上达到了 73.9%，74.1%和 74.5%的准确率，方法 C 达到了最佳效果。

表 2 兴趣点生成方法效果对比

Table 2 Results comparison of different POI generation methods

Model	AP on GT-BBOX/%	AP on SIMBA-BBOX/%
Our Method-A	76.9	73.9
Our Method-B	77.1	74.1
Our Method-C	**77.6**	**74.5**

4.3.2 自适应特征抽取

PRTR 将 PQ 简单设置为可学习参数，导致注意力分散，引入了大量不相关的信息，淹没了重要信息。本文提出的模型为每个 Query 预测其兴趣点的 2D 坐标，并将坐标正弦嵌入到 PQ 中。在计算注意力时，PQ 将对兴趣点位置周围的区域产生高响应，引导 Query 的注意力集中在兴趣点位置周围的区域。

在表 3 中，AFE 表示自适应特征抽取模块，PRTR W/O AFE 表示未设置该模块的模型，PRTR W AFE 表示设置了该模块的模型。自适应特征抽取模块将标注人体（GT-BBOX）上的关键点预测准确率提高了 1.6%，同时将 SimpleBaseline 检测人体（SIMBA-BBOX）上的关键点预测准确率提高了 1.4%。实验结果表明，自适应特征提取模块的设计是有效的。通过预测兴趣点并引导注意力集中于兴趣点周边更有利于图像特征的提取，更优质的图像特征提升了关键点预测的准确率。

表 3 加入自适应特征抽取前后对比

Table 3 Comparison of adding APE or not

Model	AP on GT-BBOX/%	AP on SIMBA-BBOX/%
PRTR W/O AFE	74.9	71.7
PRTR W AFE	**76.5**	**73.1**

本文以左眼关键点为例，可视化了本文提出的模型与基线方法在级联 Decoder 层数加深时（从左到右为第 1－6 层 Decoder）所关注的图像区域。如图 8 所示，在未设置 AFE 模块时，模型会先无序关注图像的多个部分且注意力收敛较慢、较分散。当设置了 AFE 模块时，模型首先会关注人体头部附近区域，随后注意力高度集中于眼部位置。相比于基线方法，自适应特征抽取过程使得注意力更快、更集中地关注到关键点的区域。该实验表明，自适应特征抽取过程使注意力能更快、更好地集中于关键点区域。

（a）本文方法（左眼）

（b）PRTR（左眼）

图 8 不同 Decoder 层中注意力对比

Fig. 8 Comparison of attention in different Decoder layers

4.3.3 自适应损失

本文方法同时关注了关键点可能与不可能出现的位置。关键点不可能出现的位置作为负样本，与可能出现位置的正样本同样具有重要的信息。模型对整体位置的预测也使得模型具有了关键点的全局信息，能够帮助模型更好地预测不同输入中的关键点位置。

在表 4 中，PRTR＋L1-Loss 表示使用 L1-Loss 的模型，PRTR＋Apt-Loss 表示使用自适应损失的模型，本文设计的自适应损失函数将标注人体（GT-BBOX）上的关键点预测准确率提高了 1.2%，同时将 SimpleBaseline 检测人体（SIMBA-BBOX）上的关键点预测准确率提高了 1.4%。实验结果证明了自适应损失模块设计的有效性。

表 4 加入自适应损失前后对比

Table 4 Comparison of adding adaptive loss or not

Model	AP on GT-BBOX/%	AP on SIMBA-BBOX/%
PRTR＋ L1-Loss	74.9	71.7
PRTR＋ Apt-Loss	**76.1**	**73.1**

本文方法自适应地预测关键点在所有位置上出现的可能性分布。可能性分布是一种软预测，作为一种正则化方式，它降低了单个位置在损失计算时的权重，缓解了对标注的过拟合现象，提高了模型在不同场景下的泛用性。

表 5 中的数据表明，本文设计的自适应损失函数将中小尺度人体上的关键点预测准确率提高了 1.8%，同时将大尺度人体上的关键点预测准确率提高了 0.9%。实验结果证明了自适应损失模块考虑了人体尺寸影响，提升了模型在多种人体条件下的泛化性。

表 5 在不同尺度人体上效果对比

Table 5 Effect comparison of different sizes of human bodies

Model	AP_M/%	AP_L/%
PRTR＋ L1-Loss	68.1	79.3
PRTR＋ Apt-Loss	**69.9**	**80.2**

此外，本文通过合成多个流模型的结果，进一步增强了模型映射能力，能更好地模拟现实中复杂的关键点分布情况。

本文在 MS-COCO 2017 数据集的验证集上测试使用不同的损失的效果。从图 9 可以看出，模型性能整体随着流模型数量的增加而上升。当流模型达到一定数量后，模型性能保持稳定。当设置 3 个流模型时，模型获得了最佳的性能，与使用单个的流模型相比，其将标注人体（GT-BBOX）上以及 SimpleBaseline 检测人体（SIMBA-BBOX）上的关键点预测准确率都提高了 0.5%。实验证明了合成多个流模型的结果能有效增强对关键点密度函数的模拟能力，并有效提升模型的性能。

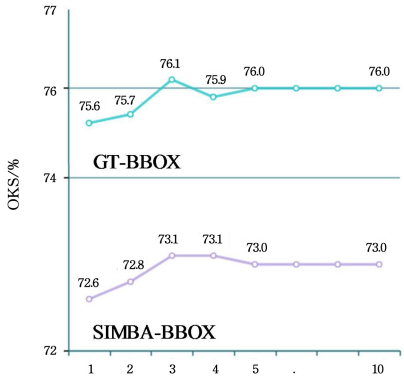


图 9　不同流模型数量的效果对比

Fig. 9　Effect comparison of different numbers of flow models

结束语　相关方法将 Transformer 应用于 2D 姿态估计任务时，存在以下两个问题：1）在交叉注意力模块中，对于不同输入，固定 Query 不能准确关注到变化的关键点区域，导致注意力分散；2）直接学习关键点的标注位置，导致模型过拟合于训练集的标注，泛化性差。本文针对以上问题，设计了一个基于自适应预测的 2D 姿态估计模型。该模型粗略估计关键点位置作为注意力所关注的兴趣点，并将兴趣点的坐标正弦嵌入到 query 中，引导注意力集中于兴趣点附近区域。通过这种方式，模型获得了集中的注意力。同时，该模型自适应地预测关键点在所有位置上出现的可能性分布，通过软预测的方式，缓解模型对标注的过拟合。进一步提升了基于回归的姿态估计方法的最佳性能。

参 考 文 献

[1] ANDRILUKA M，PISHCHULIN L，GEHLER P，et al. 2d human pose estimation：New benchmark and state of the art analysis[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014：3686-3693.

[2] TOSHEV A，SZEGEDY C. Deeppose：Human pose estimation via deep neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014：1653-1660.

[3] PISHCHULIN L，ANDRILUKA M，GEHLER P，et al. Poselet conditioned pictorial structures[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013：588-595.

[4] WANG Y，MORI G. Multiple tree models for occlusion and spatial constraints in human pose estimation[C]// Proceedings of the European Conference on Computer Vision. 2008：710-724.

[5] SUN M，KOHLI P，SHOTTON J. Conditional regression forests for human pose estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012：3394-3401.

[6] SUN K，XIAO B，LIU D，et al. Deep high-resolution representation learning for human pose estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition. 2019：5693-5703.

[7] HUANG J，ZHU Z，GUO F，et al. The devil is in the details：Delving into unbiased data processing for human pose estimation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：5700-5709.

[8] VASWANI A，SHAZEER N，PARMARN，et al. Attention is all you need[J]. arXiv：1706. 03762，2017.

[9] LI K，WANG S，ZHANG X，et al. Pose recognition with cascade transformers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：1944-1953.

[10] TIAN Z，CHEN H，SHEN C. Directpose：Direct end-to-end multi-person pose estimation[J]. arXiv：1911. 07451，2019.

[11] SUN X，XIAO B，WEI F，et al. Integral human pose regression [C]// Proceedings of the European Conference on Computer Vision（ECCV）. 2018：529-545.

[12] LI J，BIAN S，ZENG A，et al. Human pose regression with residual log-likelihood estimation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：11025-11034.

[13] MENG D，CHEN X，FAN Z，et al. Conditional detr for fast training convergence[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：3651-3660.

[14] CARION N，MASSA F，SYNNAEVE G，et al. End-to-end object detection with transformers[C]// European Conference on Computer Vision. Cham：Springer，2020：213-229.

[15] 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.

[16] LOSHCHILOV I，HUTTER F. Decoupled weight decay regularization[J]. arXiv：1711. 05101，2017.

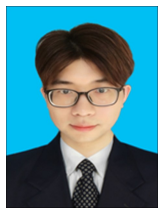
[17] GLOROT X，BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]// Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings，2010：249-256.

[18] XIAO B，WU H，WEI Y. Simple baselines for human pose estimation and tracking[C]// Proceedings of the European Conference on Computer Vision（ECCV）. 2018：466-481.

[19] ZHOU X，WANG D，KRÄHENBÜHL P. Objects as points[J]. arXiv：1904. 07850，2019.

[20] WEI F，SUN X，LI H，et al. Point-set anchors for object detection，instance segmentation and pose estimation[C]// European Conference on Computer Vision. Cham：Springer，2020：527-544.

[21] NIE X，FENG J，ZHANG J，et al. Single-stage multi-person pose machines[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：6951-6960.



ZHENG Quanshi，born in 1994，postgraduate，is a member of China Computer Federation. His main research interests include human pose estimation and action recognition.

JIN Cheng，born in 1978，Ph.D，professor，Ph.D supervisor，is a member of China Computer Federation. His main research interests include computer vision and multimedia information retrieval.