基于类别共享与独有信息双向融合的多类别姿态估计

陈俊杰[1)] 陈卫龙[1)] 方玉明[1)] 姜文晖[1)] 牛 力[2)]

[1)](江西财经大学计算机与人工智能学院 南昌 330013)

[2)](上海交通大学电子信息与电气工程学院 上海 200030)

摘 要 姿态估计旨在定位物体各关键点的位置,是一项基本的计算机视觉任务,具有广泛的应用场景。现有方法聚焦于估计单一类别物体的姿态(如人体),无法较好地用单个模型为多个类别的物体估计姿态。鉴于分类、检测、分割等模型都可为多类别预测结果,从单类别拓宽到多类别是姿态估计领域的必然发展趋势。因此,本文研究多类别姿态估计,其关键问题在于如何融合类别之间的共享信息与独有信息,使得单个模型可较好地兼容多个类别的信息。为此,本文提出基于共享与独有信息双向融合的Transformer模型,其中依据匹配关系对两种信息进行自适应融合。具体地,本模型使用可学习的查询向量来表征各类关键点的共享和独有信息,并用初始和精化两个阶段来逐步估计关键点位置。在初始阶段中,共享查询向量通过Transformer解码器来聚合图像骨干特征图中的共享信息,并预测得到关键点的初始位置和物体的类别。在精化阶段中,本模型依据共享查询向量与该类别关键点的匹配关系,将查询向量与该类别的独有查询向量进行前向融合,并将初始位置精化为准确位置。并且,本模型将更新后的独有查询向量储存到队列中,并依据匹配关系将其反向融合到共享查询向量中,可更有效地提炼共享信息。本文在多类别姿态数据集MP-100上进行了大量实验,其中的定量和定性分析都充分证明了本方法的有效性。

关键词 姿态估计;多类别;基于查询的模型;信息解耦;多头注意力模型

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2025.01795

Multi-Class Pose Estimation via Bidirectional Integrating of Class-Shared and Class-Specifc Information

CHEN Jun-Jie[1)] CHEN Wei-Long[1)] FANG Yu-Ming[1)] JIANG Wen-Hui[1)] NIU Li[2)]

[1)](School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang 330013)

[2)](School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200030)

Abstract Pose estimation, which aims to locate the keypoints of objects, is a fundamental vision task. Pose estimation has extensive applications in the real world, including virtual reality, augmented reality, smart home, human-computer interaction, robots and automation, etc. Existing methods focus on single-class pose estimation (e.g., human pose estimation), neglect multi-class scenarios, and are difficult to estimate multi-class poses with a single model. If independent pose estimation models are constructed for each category respectively, it will inevitably lead to huge computational and time costs during the training and testing phases, and

收稿日期:2024-09-05;在线发布日期:2025-04-18。本课题得到国家自然科学基金(No. U24A20220,No. 62132006,No. 62402201,No. 62161013)、国家重点研发计划(No. 2023YFE0210700)、江西省自然科学基金(No. 20242BAB21006,No. 20242BAB23012)、江西省职业早期青年科技人才培养专项项目(No. 20244BCE52070)资助。 **陈俊杰**,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为计算机视觉、姿态估计、弱监督学习、迁移学习。E-mail:chen.bys@outlook.com。 **陈卫龙**,硕士研究生,主要研究方向为姿态估计、语义分割。 **方玉明**(通信作者),博士,教授,长江学者,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、多媒体信号处理和视觉质量评估。E-mail:leo.fangyuming@foxmail.com。 **姜文晖**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为图像内容理解、跨媒体分析。 **牛 力**,博士,副教授,主要研究方向为计算机视觉、机器学习。

pose severe challenges to theactual deployment of the model and subsequent functional expansion. Considering that the models in detection or segmentation tasks could predict multi- class results, the development tendency to multi-class is inevitable. In order to expand the number of categories that a model can handle, existing methods often employ a few-sample or zero-sample learning strategy with the aim of incorporating new categories at a low labeling cost. However, these methods generally rely on additional supporting information (e. g., sample images or text descriptions) as input. This reliance on supporting information prevents them from being directly applied to standard fully supervised learning frameworks, which is the main motivation for the research work in this paper. The adaptive integration of shared and class-specific information is crucial for effective multi-class pose estimation. However, most existing models suffer from a limitation: their output layers maintain a fixed correspondence with object keypoints. This rigidity prevents the effective learning of shared and specific features, because, for example, the first output channel invariably learns the first annotated keypoint across all training images, even when these keypoints represent semantically different parts belonging to diverse object classes. Consequently, the model's learning process can be confounded by conflicting class-specific patterns, while simultaneously failing to leverage shared structural information for more accurate estimation. In this paper, we explore multi-class pose estimation, where the crux lies in how to integrate shared and specific information across classes for a better compatibility in the model. To this end, we propose a Transformer model based on bidirectional integration of shared and specific information. Specifically, we employ learnable queries to represent the shared and specific information of keypoints, and estimate keypoints by two stages. In the initial stage, shared queries aggregate the backbone feature maps and then produce initial keypoints and object class. In the refinement stage, the share queries are forward-integrated with the specific queries based on the bipartite matching between shared queries and specific keypoints, and then estimate refined keypoints. Furthermore, the updated specific queries are stored in a queue and backward-integrated into the shared queries, which could enhance the shared information. Thus, our model leverages bidirectional integration to extract multi-class information and thus better achieve multi-class pose estimation. Experiments were conducted on the MP-100 dataset, evaluating the method′s performance in the fully supervised scenario, estimating poses for diverse object categories via a single model that operates without supplementary support images or text descriptions, and the quantitative and qualitative analyses demonstrate the effectiveness of our method.

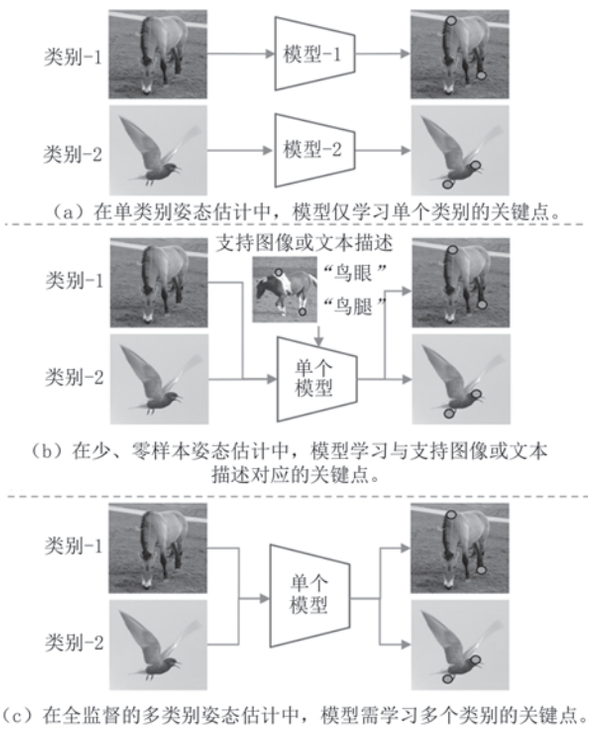
Keywords pose estimation; multi-class; query-based model; information disentangling; Transformer

1 引 言

姿态估计旨在为输入图像中的物体预测各个关键点的位置，是计算机视觉中一项基本且重要的任务，有广泛的应用场景。例如在人机交互中，手势姿态估计可获取用户的手部姿态，从而可便捷地识别用户交互意图。在自动驾驶中，车辆姿态估计可获取道路上车辆的行驶姿态，以此准确地制定驾驶决策。在智能监控中，人体姿态估计可获得行人的身体姿态，进而有效地分析人物行为和事件。因此，姿态估计具有重要的研究意义和应用价值。

现有方法大多侧重于估计单一类别物体的姿态，例如人体[1-3]、手势[4-5]、车辆姿态估计[6]，而忽略了多类别的情况。现实应用中的物体类别不计其数，如图1(a)所示，若每一个类别都单独训练一个姿态估计模型，必然有着极高的训练和测试成本，难以部署和拓展。另一方面，分类、检测、分割等模型

都具有多类别预测的能力，例如，CLIP[7]和SAM[8]能够“分类任何物体”或“分割任何物体”，已发展为具有极大影响力的基础模型（fundamental model）。因此，本文顺应必然发展趋势，研究多类别姿态估计，这也是未来“估计任何物体姿态”的基石，具有重要意义。



（a）在单类别姿态估计中，模型仅学习单个类别的关键点。

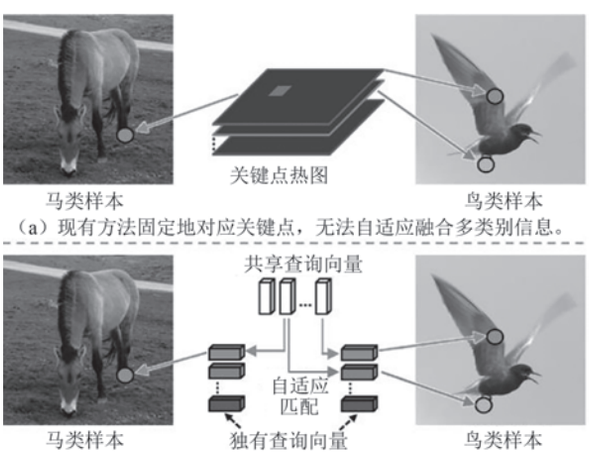
（b）在少、零样本姿态估计中，模型学习与支持图像或文本描述对应的关键点。

（c）在全监督的多类别姿态估计中，模型需学习多个类别的关键点。

图1　现有相关任务与本文研究任务的对比示意图

为拓展类别的数量，相关工作[9-14]采用少、零样本学习方法，聚焦于用更少的标注成本来学习更多的新类别，即：给定支持图像或文本描述后，让模型估计对应的关键点，如图1（b）所示。例如，Xu等人[9]首次提出了类别无关姿态估计（CAPE）任务，构建了多类别姿态数据集（MP-100），并设计了姿态匹配模型（POMNet）。该模型从多个基础类别的训练数据中学习给定关键点到目标关键点的匹配关系，所以仅需给定少量的关键点（即少样本）也可估计新类别物体的姿态。虽然目前的CAPE等少、零样本姿态估计方法[9-14]取得了显著成就，但这些方法需要额外输入支持图像或文本描述，才能够为输入图像预测结果。所以这些方法无法直接用于全监督学习这一最经典和基础的场景中，这也是本文最根本和直接的动机。为此，本文沿用多类别姿态数据集（MP-100）[9]，但侧重在全监督场景下，致力于用单个模型为多类别的物体估计姿态，并无需支持图像或文本描述。为突出各工作的侧重点，本文将CAPE等工作[9-14]总结为“小、零样本的多类别姿态估计”，而本文侧重于研究“全监督的多类别姿态估计”（可简称为“多类别姿态估计”）。

在全监督的多类别姿态估计中，不同类别物体的关键点既有潜在联系，又有显著区别，例如鸟类和马类有相似的腿部关键点，而鸟类有独特的翅膀关键点。所以，关键问题在于如何融合其中的共享与独有信息，使单个模型可兼容多个类别的信息。在大部分单类别姿态估计方法中，模型输出层与物体关键点的关系是统一、固定的，无法有效地建模多类别关键点的共享和独有信息。以最具有代表性的方法Sim. Base.[15]为例，它为鸟类预测14个关键点热图（heatmap），并为马类预测20个关键点热图。如图2（a）所示，如果固定模型的第一个关键点热图统一学习“鸟类翅膀”和“马类腿部”，则会强迫模型兼容两种关键点类型，导致更差的估计结果。如果固定模型的第一个关键点热图学习“鸟类腿部”而第二个关键点热图学习“马类腿部”，则忽略了各物体类别的腿部之间共享信息，有着可提升的空间。因此，现有方法不能有效地解决这一关键问题。



（a）现有方法固定地对应关键点，无法自适应融合多类别信息。

（b）本文方法自适应地基于匹配来融合共享信息和独有信息。

图2　现有的单类别方法与本文方法的对比示意图

为解决上述关键问题，本文提出利用可学习的查询向量（query embedding）来表征各类关键点的共享和独有信息，基于自适应匹配关系来双向融合共享和独有信息，并设计基于查询的Transformer解码器（query-based transformer decoder）来依据查询向量得到对应的关键点位置。具体地，本文将各个关键点信息解耦为共享信息与独有信息，例如鸟类腿部的关键点信息可由多类别共享的腿部信息以及鸟类独有的腿部信息组成。为了表征这些共享与独有

信息，本文模型维护一组可学习的共享查询向量，并为每个物体类别维护一组可学习的独有查询向量。单个共享查询向量旨在学习某个多类别共享的关键点信息，而单个独有查询向量旨在学习对应类别的某个独有关键点信息。如图2(b)所示，共享查询向量与各类别物体具有自适应的匹配关系，再以此融合各类别的独有查询向量，所以可有效地兼容多个类别的信息。

本文模型使用初始与精化两个阶段来逐步预测准确的关键点位置。在初始阶段中，共享查询向量通过Transformer解码器，从输入图像的骨干特征图中聚合相应特征，并解码得到关键点位置的初始估计结果。同时，分类器依据初始关键点位置来提取细粒度特征，并预测输入图像的物体类别。在精化阶段中，本文模型依据共享查询向量与该类别关键点的匹配关系，将共享查询向量与该类别的独有查询向量进行前向融合，并通过另一Transformer解码器，对初始关键点位置进行精化，进而获得更准确的估计结果。并且，本文模型将更新后的独有查询向量储存到历史队列中，并依据匹配关系将其反向融合到共享查询向量中，可更有效地提炼共享信息。通过基于匹配的双向融合方式，本文模型可自适应地学习并利用类别共享和独有信息，从而准确地估计多类别物体的姿态。

本文在多类别姿态数据集MP-100[9]上进行大量实验，包括深入的定量分析及丰富的定性分析，可充分证实本模型的有效性。本文贡献可总结为：

(1)本文顺应姿态估计发展的必然趋势，探究全监督的多类别姿态估计这一重要任务，并设立具有代表性的基准方法和评测指标。

(2)本文提出一个基于查询的两阶段模型，以查询向量来学习类别共享和独有信息，并基于匹配来前向融合两种信息，以此更准确地估计多类别的关键点位置。并且提出将独有信息储存进队列中，将其与共享信息反向融合，更好地提炼多类别共享信息。

(3)本文在多类别姿态估计数据集MP-100上进行了丰富和深入的实验，其中的定量分析和定性分析充分验证了本文方法的有效性。

2 相关工作

2.1 单类别姿态估计

姿态估计任务旨在将预测图中物体的关键点位置，是计算机视觉的基本任务之一。现有的姿态估计方法大多侧重于针对单类别物体的姿态估计，例如人体姿态估计[1-3]、动物姿态估计[4-5]、车辆姿态估计[6]等。在这些工作中，每个训练图像或测试图像的关键点数量和种类都是统一且固定的，例如第一个关键点始终对应物体眼睛，所以大多现有方法无需考虑关键点匹配问题，固定地使用模型输出层的某个卷积层通道或全连接层维度来学习某个物体关键点。从方法技术上来看，目前广为使用的姿态估计模型可大致分为下列两类。

基于热图的(heatmap-based)方法[15-19]通过预测各个关键点在平面各像素的存在概率来得到姿态关键点位置。Sim. Base.[15]方法是最具代表性的姿态估计方法之一，仅在骨干网络后添加一些反卷积层来预测关键点的热图。Wei等人[20]提出了一种基于卷积神经网络的序列化预测框架，用于预测多阶段的关键点位置，每阶段根据前阶段的图像特征和置信度图对每个关键点进行精化。同时，Newell等人[21]构建堆叠沙漏网络来估计关键点的热图。通过采用重复堆叠的编解码器结构，图像特征可以跨越所有尺度，并在处理过程中汇聚物体的各种平面关系，从而显著提高了关键点估计的精度。Chen等人[22]结合人体结构先验，通过对抗学习显著提高人体姿态估计的鲁棒性和准确性。Bin等人[23]提出了一种基于图卷积网络的姿态估计模型，可对关键点进行局部精化和全局信息提取，从而估计准确的关键点位置。TCFormer[24]方法使用逐步聚类的方法将不同位置的特征合并为自适应的形状和尺寸，以提供灵活、细致的特征。ViTPose[25]方法使用普通的、非层级的视觉Transformer作为骨干网络来提取特征，并用轻量的解码器来估计关键点位置。

基于回归的(regression-based)方法[26-28]通常直接预测物体姿态的各个关键点坐标值来获取姿态结果。在早期，Toshev等人[29]提出使用深度神经网络来回归关键点坐标，通过层级串联的方式逐步对预测的关键点坐标进行精化，最终得到准确的关键点位置。PRTR[30]方法利用Transformer的编码、解码结构来对图像特征图进行编码、对关键点查询向量进行解码，以获得关键点位置。Poseur[31]方法将姿态估计任务视为序列预测任务，并用关键点编码器以及查询向量解码器来获得关键点特征，最后直接回归关键点位置。QueryPose[32]方法提出用稀疏的组件查询向量来编码局部特征，并用选择性迭代模块来自适应地更新查询向量，并从中解码关键点位

置。PCT[33]方法将整体姿态分解为若干个子结构的组合，并将姿态估计任务转化为子结构分类任务。ED-Pose[34]方法将姿态估计任务视为关键点边界框检测任务，可提供更好的上下文信息，并结合全局、局部特征来进行姿态估计。GroupPose[35]方法对关键点进行分组，并利用组间注意力和组内注意力来充分聚合关键点特征，提升姿态估计性能。

虽然上述工作在单类别姿态估计任务上取得了优秀的成就，但多类别物体有不同数量和种类的关键点，所以如何有效地融合类别共享与独有信息仍是一个遗留问题。本文以自适应匹配为核心，可更好地兼容类别之间的信息，能更有效地用单个模型为多个类别的物体估计姿态。

2.2　类别无关姿态估计

为了便捷且低廉地拓展姿态估计的类别范围，类别无关姿态估计（CAPE）系列工作[9-14,36-37]采用少样本和零样本方法，假设已有关于基础类别的带姿态标注数据作为辅助，致力于用成本更低的标注来估计广泛新类别物体的姿态。

基于少样本方法的类别无关姿态估计以基础类别的充足姿态标注数据为支持，利用少量姿态标注数据来学习更广泛的新类别。CAPE的开创工作[9]提出了基于关键点匹配模型，从基础类别的强标注数据中学习支持图像与输入图像上的关键点匹配关系，从而根据额外给定的支持图像为任意类别的物体估计姿态。随后，Shi等人[10]对POMNet进行了改进，提出了CapeFormer模型，在POMNet的基础上将关键点提示匹配的方式改进为计算内积相似度来提升计算效率，同时增加了关键点精化阶段以得到更为精细的关键点坐标。此外，部分研究致力于特定类别的多个子类别，通过构建不同的模型进行小样本姿态估计。例如对于动物类别，Lu等人[38]提出了一个基于多元高斯分布的模型，用于对关键点间的不确定性进行建模，能够利用相邻关键点之间的潜在关联进行更准确的姿态估计。Sun等人[39]提出了基于少样本学习的统一动物感知模型，该模型结合了多头注意力机制和匹配模块，能够聚合多种形式的提示信息，并通过从姿态估计、分割和分类任务中学习通用特征，从而支持更精确的姿态估计。

基于零样本方法的类别无关姿态估计利用基础类别的文本描述和姿态标注数据来学习语言到姿态的映射，从而实现成本低廉的多类别姿态估计。Yang等人[12]使用文本描述姿态关键点或少量图像注释作为输入提示，通过联合编码语言和视觉模态来提取提示词和图像特征，并采用由粗到细的策略对任意类别进行精准的关键点估计。此外，一些工作基于零样本学习范式来对动物类别的若干子类别进行姿态估计。例如，Zhang等人[13]提出了基于提示词的对比学习方法，旨在建立语言与动物姿态关键点之间的联系，并通过空间感知和特征感知的对齐策略来提升基于文本描述的姿态估计性能。与此同时，Zhang等人[40]提出了基于语义特征匹配的模型框架，将域分布矩阵匹配与视觉关键点关联感知模块相结合，从而实现从文本描述到视觉关键点坐标的准确映射。

虽然上述基于少、零样本的类别无关姿态估计（CAPE）方法在学习广泛新类别上取得了显著成就，但这些工作致力于降低学习新类别所需的标注成本，需要额外输入支持图像或文本描述，无法直接应用于全监督的多类别姿态估计场景中。本文研究全监督这一经典范式下的多类别姿态估计，致力于用单个模型为多个类别的物体估计姿态，无需支持图像或文本描述。

2.3　基于查询向量的模型

本文所提出的模型属于基于查询向量的模型（query-based model），可有效地利用查询向量自适应地学习多个类别关键点的共享信息。同类模型在检测[41-43]或分割[44-46]任务中也有广泛应用，已被验证可学习多类边界框或掩码的共享信息。

例如，目标检测任务中有代表性的工作DETR[41]，它设置一组查询向量来学习多类别物体的共享信息，并通过Transformer模块与骨干特征交互，最后预测各查询向量对应的物体类别和边界框坐标。分割任务中有代表性的工作是MaskFormer[44]，它类似地设置查询向量来学习多类别的共享信息，最后预测各查询向量对应的物体类别和掩码区域。这些工作中的某个查询向量可在不同图像中对应不同类别的物体，这表明基于查询向量的模型适合学习不同物体类别的共享信息。

因为检测和分割任务中的各类别的表征较为统一，例如都是边界框或像素级分类向量，所以上述方法没有对各类别的信息进行有针对性的处理。而在多类别姿态估计中，不同类别可以有不同类型和数量的关键点，所以本文构建基于查询向量的模型来双向融合类别共享信息与独有信息，从而更精准地用单个模型估计多类别物体的姿态。在模型架构上，本文提出的模型不但设置了共享查询向量来获得预测结果，还进一步前向融合各类独有信息以对

预测结果进行精化，并提出将训练过程中的独有信息反向融合以增强共享信息的代表性。

3 本文方法

本节详细介绍本文提出的模型和方法。首先，小节3.1定义多类别姿态估计的形式。然后，小节3.2对本文模型的总体流程进行概述。随后，小节3.3介绍初始阶段的计算细节，及反向融合的过程。并且，小节3.4介绍共享查询向量和各类别关键点的匹配方式。然后，小节3.5介绍精化阶段的计算细节，及前向融合的过程。另外，小节3.6介绍可形变Transformer解码器的架构和计算流程。最后，小节3.7介绍本文模型的训练损失函数和测试流程。为了便于描述，本文使用普通字母表示标量，使用粗体字母表示向量/矩阵/张量，使用花体字母表示函数。本文使用方括号表示对张量的按维度索引，例如$\boldsymbol{P}[k]$表示第k个关键点坐标。

3.1 任务定义

在本文研究的全监督多类别姿态估计中，训练和测试图像所含物体来自C个类别，这些类别的物体可有不同的关键点数量和类型。考虑到模型设置和批次运行都只能处理固定数量，所以本文使用可见度（visibility）V将不同数量的关键点统一补齐到K。其中，可见度V表示补齐后的K个关键点在图像x中是否可见，例如遮挡情况或种类变化都会导致某些关键点不可见。具体来说，每个样本包含图像$\boldsymbol{x}\in\mathbb{R}^{H\times W\times 3}$，物体类别$\boldsymbol{y}^{*}\in\mathbb{R}^{C}$，关键点位置$\boldsymbol{P}^{*}\in\mathbb{R}^{K\times 2}$，和关键点可见度$\boldsymbol{V}^{*}\in\mathbb{R}^{K}$。总之，多类别姿态估计旨在获得一个模型$\mathcal{F}(\cdot)$，无需支持图像或文本描述，可实现

$$\boldsymbol{y};\ \boldsymbol{P};\ \boldsymbol{V}=F(\boldsymbol{x}) \tag{1}$$

其中，$[\cdot;\cdot]$表示张量的拼接（concatenation）。由于各类别K个关键点的排列顺序是来源于标注、可能无关的，而模型预测的排列顺序是统一、固定的，所以本任务的关键问题是如何自适应地融合类别共享与独有信息，使单个模型可兼容多个类别的信息。

3.2 模型概述

鉴于多类别姿态估计的关键问题是如何提取并融合类别共享与独有信息，本文提出两阶段的模型，其架构如图3所示。在初始阶段中，本模型首先利用类别共享信息估计输入图像的类别和姿态结果，但这初始的结果没有考虑类别信息。所以在后续的精化阶段中，本模型融合对应类别的独有信息，得到更准确的姿态估计结果。下面首先概述本模型的总体流程，再分别介绍各部分的具体细节。

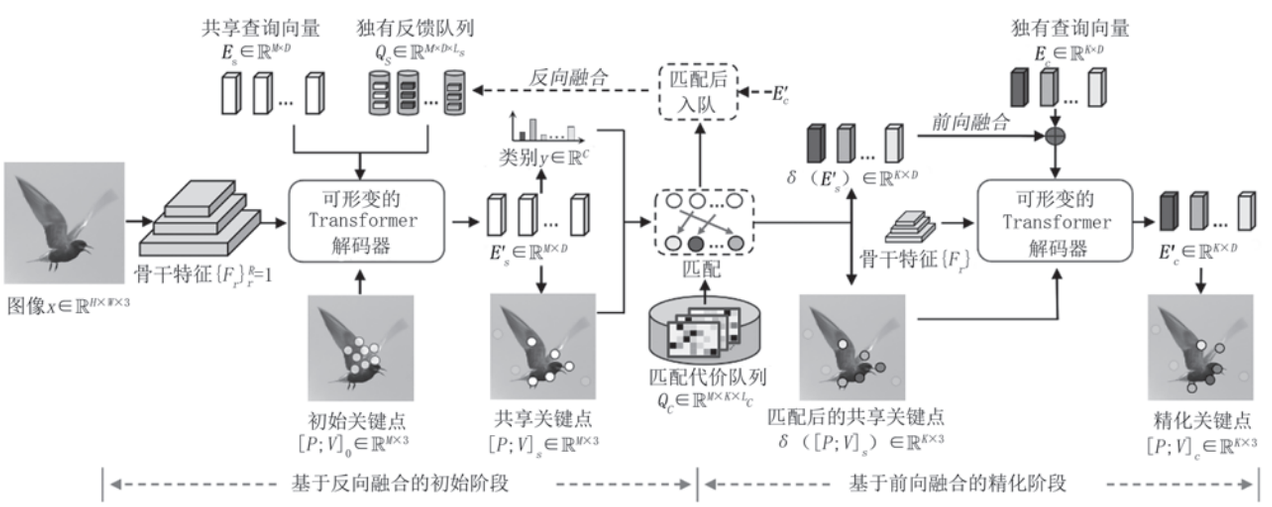


图3 本文所提出模型的总体架构图

本模型维护M个可学习的共享查询向量$E_s\in\mathbb{R}^{M\times D}$。基于这些共享查询向量，本文模型通过两种方式提取和学习共享信息，包括：1）本文模型计算共享查询向量与真实关键点的匹配关系，并通过梯度优化使单个共享查询向量可学习到多个类别的共享信息；2）考虑到深层特征更富含深度语义信息，本文模型将深层特征依匹配关系储存在队列中，并与共享查询向量进行融合，从而增强共享信息的语义代表性。此外，本文为C个类别各维护K个可学习的独有查询向量$\boldsymbol{E}_c\in\mathbb{R}^{K\times D}$，其中$D$是统一的特征维度且$M\geqslant K$。在训练过程中，每一个查询向量都将学习某类关键点的特征信息，并负责对应关键

点坐标的估计。例如，$\boldsymbol{E}_s[i]\in\mathbb{R}^D$可通用地学习“鸟类腿部”和“马类腿部”这两类关键点的共享信息，鸟类的$\boldsymbol{E}_c[j]\in\mathbb{R}^D$可专注地学习“鸟类翅膀”这一类关键点的独有信息。通过这两种类型的查询向量以及它们的双向融合，本文模型可更好地统一估计多种类别物体的姿态关键点。给定一张输入图像$x\in\mathbb{R}^{H\times W\times 3}$时，本文模型首先使用ResNet-50骨干网络提取共R个尺度的特征图，并用核尺寸为1的卷积层将特征图统一压缩到D维度以便于处理，记为$\{F_r\}_{r=1}^R$。后续的两个阶段将充分利用多尺度骨干特征$\{F_r\}$提供的细粒度特征，从而进行精准的姿态估计。

在初始阶段中，共享查询向量$\boldsymbol{E}_s$首先反向融合独有查询向量，然后通过可形变的Transformer解码器来聚合骨干特征$\{F_r\}$，得到图中共享关键点特征$\boldsymbol{E}'_s\in\mathbb{R}^{M\times D}$。一方面，这些细粒度的关键点特征可用于估计图中物体的类别$y\in\mathbb{R}^C$。另一方面，这些关键点特征可用于估计对应关键点的位置及可见度$[\boldsymbol{P};\ \boldsymbol{V}]_s\in\mathbb{R}^{M\times 3}$。这些估计的关键点可进一步结合类别的独有信息，从而精化为更精准的姿态关键点。

在精化阶段中，本文模型首先计算共享查询向量和关键点类型的匹配关系$\delta(\cdot)$，然后将匹配后的共享关键点特征$\delta(\boldsymbol{E}'_s)\in\mathbb{R}^{K\times D}$与该物体类别的独有查询向量$\delta(\boldsymbol{E}_c)\in\mathbb{R}^{K\times D}$进行前向融合，最后通过另一个可形变的Transformer解码器，将匹配后的初始估计结果$\delta\left([\boldsymbol{P};\ \boldsymbol{V}]_s\right)\in\mathbb{R}^{K\times 3}$精化后得到更准确的关键点位置及可见度$[\boldsymbol{P};\ \boldsymbol{V}]_c\in\mathbb{R}^{K\times 3}$。通过这种方式，本文模型可自适应性地融合类别共享信息和独有信息，得到准确的多类别姿态估计结果。本文模型的各部分细节将在下文中详细介绍。

3.3　基于反向融合的初始阶段

本模型的初始阶段为图像$\boldsymbol{x}$预测物体类别$\boldsymbol{y}\in\mathbb{R}^C$和初始的关键点及可见度$[\boldsymbol{P};\ \boldsymbol{V}]_s\in\mathbb{R}^{M\times 3}$。受query-based model[41-46]的启发，本模型设置共享查询向量$\boldsymbol{E}_s\in\mathbb{R}^{M\times D}$以学习各类关键点的共享信息。考虑到接近神经网络输出层的特征更富含深度语义信息，所以本文提出在训练过程中将各类独有信息存储进反馈队列，并自适应地融合进共享查询向量$\boldsymbol{E}_s$，从而增强共享信息的鲁棒性。由于是把接近输出层的特征融入接近输入层，与通常网络的前向计算过程相反，所以称之为“反向”融合。

具体地，本文模型为M个共享查询向量各设置一个独有反馈队列，记为$\boldsymbol{Q}_s\in\mathbb{R}^{M\times D\times L_s}$，其中$L_s$是队列的最大长度。在训练过程的每一次迭代中，独有查询向量在经过解码器后得到的独有关键点特征$\boldsymbol{E}'_c$都依据匹配关系$\delta(\cdot)$加入到对应的独有反馈队列中，即$\boldsymbol{E}'_c[k]$加入到队列$\boldsymbol{Q}_s[\delta(k)]$中。例如，鸟类腿部和马类腿部的独有信息都可通过自适应的匹配关系，加入到同一个共享查询向量的独有反馈队列中，所以可有效地增强腿部的共享信息。给定输入图像的骨干特征$\{F_r\}$后，本文模型首先融合共享查询向量与独有反馈队列中的特征，得到融合共享特征$\tilde{\boldsymbol{E}}_s\in\mathbb{R}^{M\times D}$如下式所示：

$$\tilde{\boldsymbol{E}}_s=\boldsymbol{E}_s+\bar{\boldsymbol{Q}}_s \tag{2}$$

其中$\bar{\boldsymbol{Q}}_s\in\mathbb{R}^{M\times D}$是各队列储存特征的均值，若队列为空则取全零特征。然后，融合共享特征$\tilde{\boldsymbol{E}}_s$通过解码器来聚合输入图像的骨干特征$\{F_r\}$，得到共享关键点特征$\boldsymbol{E}'_s\in\mathbb{R}^{M\times D}$，可总结为：

$$\boldsymbol{E}'_s=\mathcal{D}_s\left(\tilde{\boldsymbol{E}}_s,\ \{F_r\},P_0,V_0\right) \tag{3}$$

其中$\mathcal{D}_s(\cdot)$是可形变的Transformer解码器，根据参考点(reference point)P_0来挖掘附近的细粒度特征，并用可见度V_0来屏蔽不可见关键点特征的影响(细节请见小节3.6)。基于这些共享关键点特征E'_s，本文在其均值上施加一个全连接层(FC)来预测物体的类别，即$y=\mathcal{F}_{FC}(\bar{\boldsymbol{E}}'_s)\in\mathbb{R}^C$。并且，以增量的方式用多层感知机(MLP)预测关键点坐标和可见度，如下式：

$$[\boldsymbol{P};\ \boldsymbol{V}]_s=\sigma\left(\mathcal{F}_{MLP}(E'_s)+\sigma^{-1}\left([\boldsymbol{P};\ \boldsymbol{V}]_0\right)\right) \tag{4}$$

其中$\sigma(\cdot)$和$\sigma^{-1}(\cdot)$是Sigmoid及其逆函数。本文设置可优化的参数作为P_0和V_0，优化初始值均为0.5。

3.4　共享查询向量的自适应匹配

初始阶段中利用共享信息估计的结果$[\boldsymbol{P};\ \boldsymbol{V}]_s$尚未考虑类别独有的信息，后续可融合独有信息做进一步的精化。由于各类物体可以有不同数量和类型的关键点，所以首先要确定补齐后的K个共享关键点与对应类别物体关键点的匹配关系，才能为每个关键点对应地融合。该匹配关系记为$\delta(\cdot)$，即物体的第k个关键点对应第$\delta(k)$个共享关键点。依据$\delta(\cdot)$，本文模型在精化阶段(请见3.5节)中可自适应地将同一个共享查询向量预测的关键点分别匹配给鸟类图像的腿部，或马类图像的腿部：

为获取匹配$\delta(\cdot)$，本模型计算将第m个共享关键点匹配给第k个物体关键点的代价矩

阵$C\in\mathbb{R}^{M\times K}$：

$$\boldsymbol{C}[m,k]=\mathcal{L}_1\left(\boldsymbol{P}_s[m],\boldsymbol{P}^*[k]\right)+\alpha\mathcal{L}_{CE}\left(V_s[m],V^*[k]\right) \tag{5}$$

其中$\mathcal{L}_1(\cdot)$是用于衡量两个坐标点偏差的L1函数，$\mathcal{L}_{CE}(\cdot)$是用于衡量可见度二分类偏差的二元交叉熵函数，而α是两项代价间的均衡超参数。考虑到测试图片没有真实标注，所以本模型为每一个物体类别都设置了一个匹配代价队列，用训练阶段学习到的匹配代价来决定测试时共享关键点具体对应哪个关键点类型。该队列记为$Q_c\in\mathbb{R}^{M\times K\times L_c}$，其中$L_c$为队列最大长度。在训练过程中，每一个图像的代价矩阵C都保存进图中物体类别y^*对应的队列。

在代价矩阵入队后，本文模型取物体类别y^*（测试时取y）对应队列中保存的代价矩阵的均值$\bar{C}$，并利用匈牙利匹配算法[47]求解获得$\delta(\cdot)$。通过这种自适应的匹配关系，具有潜在联系的各类关键点可匹配到同一个共享关键点上，使得共享查询向量可学习到各类别间的共享信息。

3.5 基于前向融合的精化阶段

由于初始阶段并未考虑图中物体的类别信息，所以本模型的精化阶段依据匹配关系$\delta(\cdot)$，将前一阶段提取的共享信息与独有信息相融合，从而在初始估计结果$[\boldsymbol{P};\boldsymbol{V}]_s$的基础上得到更精细的姿态估计结果。因为这一计算过程符合通常网络的前向计算过程，所以称之为“前向”融合。

具体地，首先根据匹配关系$\delta(\cdot)$将共享关键点的特征、位置、可见度进行重排列，使其与类别关键点的顺序对齐，例如排列后的特征记为$\delta(\boldsymbol{E}_s')\in\mathbb{R}^{K\times D}$。然后，本文模型融合$\delta(\boldsymbol{E}_s')$与该类的独有查询向量$\boldsymbol{E}_c$，得到融合独有向量，即$\tilde{\boldsymbol{E}}_c\in\mathbb{R}^{K\times D}=\delta(\boldsymbol{E}_s')+\boldsymbol{E}_c$。随后，融合独有向量$\tilde{E}_c$通过可形变Transformer解码器来聚合图像骨干特征，得到独有关键点特征E_c'，如下式所示：

$$\boldsymbol{E}_c'=\mathcal{D}_c\left(\tilde{\boldsymbol{E}}_c,\{F_r\},\delta(P_s),\delta(V_s)\right) \tag{6}$$

其中，以重排列后的关键点位置为参考点。一方面，这些独有关键点特征E_c'依据匹配加入反馈队列以进行反向融合，如小节3.3所描述。另一方面，这些独有关键点特征通过多层感知机对关键点坐标和可见度进行增量式精化，最后得到$[\boldsymbol{P};\boldsymbol{V}]_c\in\mathbb{R}^{K\times 3}$，该过程与公式(4)一致，不再赘述。

3.6 可形变Transformer解码器

在初始和精化阶段中，本模型用可形变Transformer解码器从查询向量得到符合图像内容的关键点特征，从而为输入图像估计姿态。现有工作[9-10]大多使用基于cross attention的解码器，但无法高效地提取多层次和细粒度的特征，这对于精准估计关键点有重要作用。为此，本文基于可形变注意力机制（deformable attention[42]）来构建解码器，并用可见度来屏蔽无效关键点的干扰，使查询向量可聚合多尺度和细粒度的特征$\{F_r\}$。

本文使用的可形变Transformer解码器架构如图4所示，给定N个代表了某类关键点的查询向量$\boldsymbol{E}\in\mathbb{R}^{N\times D}$，以及它的参考点$\boldsymbol{P}\in\mathbb{R}^{N\times 2}$和可见度$\boldsymbol{V}\in\mathbb{R}^{N}$，该解码器首先以可见度进行加权，进行查询向量之间的自注意力计算，如下式：

$$\boldsymbol{E}_{sa}[n]=\sum_{h=1}^{H}W_h\left(\sum_{j=1}^{N}\boldsymbol{V}[j]\cdot A_{hnj}\cdot W_h'E[j]\right) \tag{7}$$

其中W_h和W_h'是第h个注意力头的可学习映射参数，A_{hnj}是第i个查询向量和第j个查询向量之间的注意力权重，而$V[j]$是第j个查询向量的可见度，可用于屏蔽对应不可见关键点的查询向量的影响。在获取了上述自注意力特征$E_{sa}\in\mathbb{R}^{N\times D}$后，该解码器以$P$为参考点来聚合多尺度的骨干特征，即

$$E_{da}[n]=\sum_{h=1}^{H}W_h\left[\sum_{r,i}A_{hrni}W_h'F_r\left(P[n]+O_{hrni}\right)\right] \tag{8}$$

其中，A_{hrni}和O_{hrni}是第h个注意力头在第r个特征图上，第n个特征对第i个采样点的注意力权重和采样偏移，更多细节可见文献[42]。最后，可形变注意力特征E_{da}与自注意力特征E_{sa}相融合，得到关键点特征$E'\in\mathbb{R}^{N\times D}=E_{sa}+E_{da}$。通过这种方式，对应某

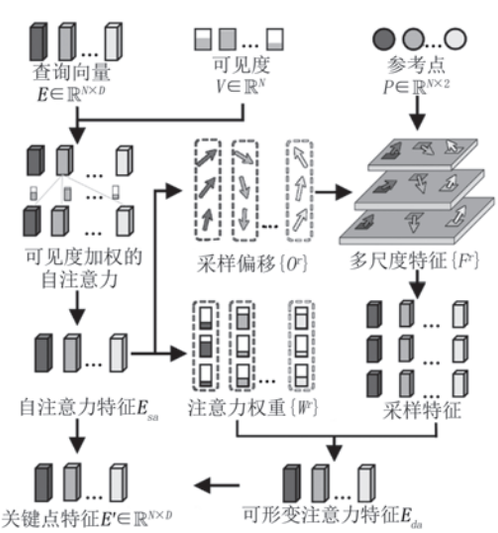


图4 可形变的Transformer解码器架构示意图

类关键点的查询向量可通过两种注意力机制充分融合姿态关键点和图像骨干特征的信息，从而为后续的关键点预测和可见度预测提供支持。

3.7　训练与测试

在训练阶段中，本文模型所使用的总体损失函数可大致分为如下三个部分：

$$\mathcal{L}_{full}=\mathcal{L}_s+\mathcal{L}_c+\beta\cdot\mathcal{L}_{CE}(y,y^*) \tag{9}$$

其中$\mathcal{L}_s$和$\mathcal{L}_c$分别是初始阶段和精化阶段的损失函数，$\mathcal{L}_{CE}$是监督物体分类的交叉熵损失，而β是均衡超参数。具体地，初始阶段的损失函数为

$$\mathcal{L}_s=\mathcal{L}_1(\delta(P_s),P^*)+\alpha\mathcal{L}_{CE}(\delta(V_s),V^*) \tag{10}$$

其中两项分别对应关键点回归和可见度分类损失。与DETR[41]、MaskFormer[44]、MetaPoint[48]一致，本文对未分配的关键点的可见度施加一个普通的交叉熵损失，使其逼近零值，较为简单故略去。精化阶段的损失函数为

$$\mathcal{L}_c=\mathcal{L}_1(P_c,P^*)+\alpha\mathcal{L}_{CE}(V_c,V^*) \tag{11}$$

其中包含关键点回归和可见度分类损失函数。

在测试阶段中，对于输入图像x，本模型输出类别预测y和精化后的关键点位置P和可见度V作为结果，即可完成多类别姿态估计任务。另外，本模型无需关于独有反馈队列和匹配代价队列的入队操作，并可预先计算队列内容的均值和求解匹配关系，所以无需显著的额外计算成本。

4　实　验

4.1　数据集

与相关工作[9-10,48]一致，本文使用多类别姿态数据集MP-100[9]来进行实验。MP-100数据集是一个大型公开数据集，由众多单类别姿态数据集组成，包括COCO[49]、300W[50]、Keypoint[51]、OneHand10K[52]、CUB-200[53]等广为使用的数据集。在类别丰富度上，MP-100数据集覆盖100个类别，归属于8个超类，例如人体、动物、车辆等。在样本数量上，MP-100数据集包含超过18 000张样本图像以及20 000个物体标注。每个类别的物体可有不同种类和数量的关键点，且关键点的数量在8到68之间。因此，MP-100[9]数据集可较好地支撑本文多类别姿态估计的研究实验。本文统一将每个物体类别前70%的样本图像用于训练验证，而后30%的样本图像用于测试。为进一步增强本文实验结果的代表性，本文也在Keypoint-5[51]数据集上进行了实验。Keypoint-5是一个针对家具类别的数据集，包含床、椅子、沙发、旋转椅和桌子五个类别。每个类别含有1000到2000张图片，且每类的关键点数量不一，有8～14不等。每张图片标注三次关键点位置，并取中间值作为真实值。

4.2　评测指标

本文侧重于研究全监督场景中的多类别姿态估计，但现有的单类别姿态估计方法[1-6]所用指标未考虑类别因素，不能准确反映本任务中模型的性能。CAPE等少、零样本姿态估计方法[9-14]虽然涉及多个类别，但测试过程会给定每个样本的类别，所以其中的指标也不能直接用于本任务。本文在上述工作所用指标的基础上，进一步考虑模型对于物体类别的预测，并用mAP作为总体评价指标。

具体地，为计算mAP指标，本文用9个可见度阈值$v\in\{0.1,\cdots,0.9\}$来计算准确率和召回率，且设置4个坐标偏差阈值$t\in\{0.05,0.1,0.15,0.2\}$以从不同粒度反映关键点结果的偏差。对每个测试样本及标注$[x;y^*;P^*;V^*]$，模型需输出结果$[y;P;V]$。在每个T_p取值下，第k个关键点在物体类别预测正确且满足下列条件时，才被判定为预测正确：

$$\left(\left|P[k]-P^*[k]\right|_1\leqslant t\right)\left(P[k]\geqslant v\right) \tag{12}$$

本评测只统计实际可见的关键点（即$P[k]^*$为1），可累计得到每个样本的关键点正确数和实际可见关键点数，从而计算该样本的AP指标。

正式地，第c个类别的第i个样本在坐标偏差阈值t下计算的AP值可记为$AP_{i,c,t}$，则总体指标mAP可由下式计算：

$$\text{mAP}=\frac{1}{4}\sum_t\left(\frac{1}{C}\sum_c^C\left(\frac{1}{N_c}\sum_i^{N_c}AP_{i,c,t}\right)\right) \tag{13}$$

其中，N_c是第c个类别的测试样本数量，C是物体类别的数量，而阈值t取{0.05, 0.1, 0.15, 0.2}这4个值以综合地反映偏差。

4.3　实施细节

本文的实施总体沿袭相关工作[9-10,48]，主体使用Python 3.74、PyTorch 1.8.0[54]和MM-Pose[55]来实现。具体地，本文统一使用在ImageNet[56]上预训练好的ResNet-50[57]作为骨干网络。并且，类别数量$C=100$，统一的特征维度为$D=256$，特征图尺度数量为$R=3$，补齐的关键点个数为$K=70$，共享查询向量的个数为$M=80$，匹配代价队列和独有反馈

队列的长度均为40，均衡超参数为$\alpha=0.1$及$\beta=0.5$。在数据配置方面，本文与相关工作[10,31,48]一致，基于物体边界框对样本图像进行裁剪，并统一缩放到256×256的分辨率。在训练阶段中，本文使用的数据增强策略包括随机缩放和随机旋转。在模型训练方面，本文使用Adam[58]优化器，并优化共200个周期。模型优化的批次大小为16，而学习率为10^{-4}，并以0.1的倍率在第160和第180个周期降低。本文的实验环境为Ubuntu 20.04系统，配备64 GB内存的Intel I7-14900 CPU和2块NVIDIA 4090 GPU。除非另有说明，否则所有实验中的随机种子设置为100。

4.4 与现有方法的定量比较

在现有的相关工作中，CAPE等[9-14]少、零样本的多类别姿态估计方法需要额外输入支持图像或文本描述，无法直接应用于本文的全监督实验场景中。所以本文使用具有代表性且广为使用的单类别姿态估计方法进行比较，包括：Sim. Base.[15]、PRTR[30]、TCFormer[24]，Poseur[31]、QueryPose[32]、ViTPose[25]、PCT[33]，ED-Pose[34]、GroupPose[35]、NerPE[18]、RTMO[19]。由于上述方法均侧重于单类别姿态估计，所以本文需将其改编为多类别姿态估计模型。具体地，为了获取类别预测y，本文统一在骨干网络后添加平均池化层和分类的全连接层。为获取关键点可见度V，基于热图预测的方法（如Sim. Base.[15]）使用热图的最大值作为可见度，而基于回归的方法（如GroupPose[35]）在每个关键点的回归全连接层上增加一维预测可见度。

实验结果汇总在表1中，由此可以发现本方法具有最优的多类别姿态估计性能。具体来看，近期提出的方法NerPE[18]和RTMO[19]在MP-100[9]数据集上分别取得65.4%mAP和66.1%mAP，显示了现有单类别姿态估计方法在多类别姿态估计任务中的性能水平，而本文方法在MP-100数据集上的性能对其有明显提升，较NerPE[18]提升+5.2%mAP，较RTMO[19]提升+4.5%mAP。同时，在Keypoint-5数据集[51]上的实验，NerPE[18]和RTMO[19]也达到了最为强劲的性能，而本文方法较NerPE[18]提升+3.9%mAP，较RTMO[19]提升+3.3%mAP。并且，本文方法在各种阈值指标上都相较于现有方法有一致提升，例如在MP-100数据集上，AP@0.05指标比RTMO[19]的提升为+5.7%，AP@0.20指标比RTMO[19]的提升为+4.4%。总体上，本文方法相较于现有方法可更准确地为多个类别的物体估计姿态，充分显示了用可学习的查询向量来表征类别共享与独有信息并基于匹配将其双向融合的有效性。

表1 各种方法在MP-100数据集上的姿态估计性能表（%，↑），其中最好的结果使用粗体标记

刊物	方法	MP-100数据集					Keypoint-5数据集				
		0.05	0.1	0.15	0.2	mAP	0.05	0.1	0.15	0.2	mAP
ECCV2018	Sim. Base.[15]	46.2	56.0	59.5	61.6	55.8	76.8	83.6	86.7	88.0	83.8
CVPR2021	PRTR[30]	42.3	55.1	62.9	66.2	56.6	75.9	82.5	88.5	89.3	84.1
CVPR2022	TCFormer[24]	47.3	59.0	62.6	64.7	58.4	78.1	84.6	88.1	89.2	85.0
ECCV2022	Poseur[31]	49.6	60.3	62.8	64.6	59.3	80.1	85.4	88.3	90.7	86.1
NeurIPS2022	QueryPose[32]	48.2	59.8	63.7	66.1	59.5	80.0	85.2	89.7	91.4	86.6
NeurIPS2022	ViTPose[25]	51.9	61.8	65.3	67.4	61.6	81.8	86.4	90.4	91.8	87.6
CVPR2023	PCT[33]	50.2	61.9	65.8	68.8	61.7	80.8	86.8	90.9	92.2	87.7
ICLR2023	ED-Pose[34]	52.7	64.6	68.2	70.2	63.9	81.9	85.6	88.6	90.5	86.6
ICCV2023	GroupPose[35]	48.9	66.8	71.5	73.9	65.3	80.3	86.6	91.1	93.4	87.9
NeurIPS2024	NerPE[18]	53.6	65.9	69.7	72.4	65.4	82.1	86.6	90.8	91.8	87.8
CVPR2024	RTMO[19]	50.5	67.4	71.9	74.4	66.1	81.4	86.8	91.4	93.9	88.4
-	本文方法	**56.2**	**70.4**	**76.8**	**78.8**	**70.6**	**84.6**	**91.1**	**94.7**	**96.2**	**91.7**

各方法参数量和帧率如表2所示，本文模型的参数量比ED-Pose[34]、GroupPose[35]和RTMO[19]要少，在所有比较方法中有着较低的模型复杂度。在运行效率方面，在不同计算资源（即4090，A5000，4060ti）和不同分辨率（即256×256和512×512）约束下，本文模型的帧率都要超越方法ED-Pose[34]和GroupPose[35]。且在模型参数量略多于NerPE[18]的情况下，本文模型的帧率较NerPE[18]展现出可比的计算效率。实验结果表明，各方法的计算效率与设备算力及分辨率呈强相关性，在4090上帧率达到峰值，4060ti次优，而A5000表现相对受限。同时，输入分辨率从256提升至512时，所有方法因计算负

载增加均呈现帧率下降趋势。综上所述，本文模型在保持较低复杂度和良好效率的情况下，仍具有最优的多类别姿态估计性能。

4.5　与现有方法的定性比较

为了直观地显示本模型的效果，本文选择有代表性的三个基准方法进行可视化定性比较，包括Sim. Base[15]、ED-Pose[34]和GroupPose[35]。定性比较如图5所示，其中展示了各方法在8个类别的样本上的姿态估计结果。具体地，最左侧一列展示输入图像，中间四列展示各方法的估计结果，最右侧一列展示标注的真实姿态。由于可见的真实关键点数量可变化，本可视化过程首先以红到蓝的渐变颜色确定可见真实关键点的填充颜色，其余不可见的真实关键点使用黑色填充。对于各方法的估计结果，估计的关键点使用对应真实关键点的颜色填充，用透明度显示其可见度，并用黑色箭头显示其与真实关键点的位置偏差。另外，如果估计关键点的可见度大于0.5，则用黄色线段绘制与其相连的骨架。

从可视化定性比较图5中可以发现，本文模型

表2　各方法的参数量(M)及在不同设备和分辨率图像上的帧率(FPS)对比

方法	参数量	4090GPU		A5000GPU		4060tiGPU	
		256	512	256	512	256	512
ED-Pose	48.4	23.9	18.1	12.5	8.5	17.5	14.3
GroupPose	50.0	37.5	29.4	18.0	14.4	28.0	22.4
NerPE	28.5	60.6	43.9	29.8	21.3	45.9	33.0
RTMO	41.7	76.3	56.8	36.6	27.5	59.3	43.4
本文方法	33.3	58.1	45.0	27.5	21.8	46.0	34.3

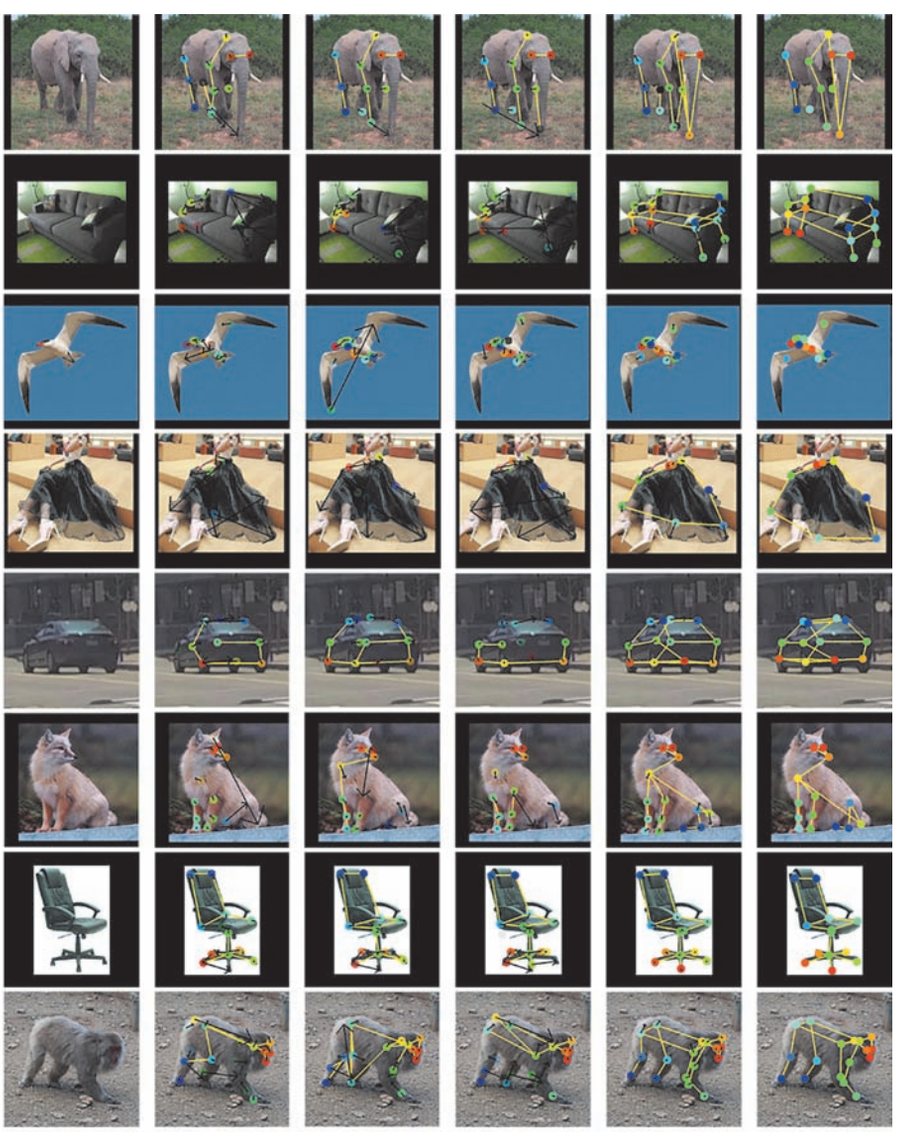


图5　各种方法的可视化定性比较图

可以为多个物体类别预测更准确的姿态。例如对于第一行的大象类别，可能由于图像中的鼻子与腿部有相似的外表和结构，各基准方法难以区分，而本文模型可更好地分别定位大象的鼻子与腿部。对于第四行的裙子类别，可能由于裙子的形状和结构变化较大，各基准方法难以定位对应的关键点，而本文模型可较好地定位裙子类别的关键点。总体上，本文模型对于各类都有更好的关键点定位效果，可更精准地完成多类别姿态估计任务。

4.6 与现有方法的深入比较

为了凸显本方法与对比方法在不同类别上的性能差异，本小节将本方法和有代表性的方法在8个超类上的性能汇总在表3中。从表3中可以发现，本文方法对于各个超类都取得了大幅度的性能领先，例如+8.3% on Human Body以及+8.8% on Furniture。虽然GroupPose[35]等方法在单类别姿态估计中取得了较大成就，但只侧重于为单个类别的物体估计准确的关键点，尚未针对性地处理多个类别之间的信息融合问题。相比之下，本文方法采用了基于匹配的双向融合策略，所以可更好地兼容多个类别的信息，从而更有效地为多个类别的物体估计姿态。

为进一步验证本方法可自适应地学习类别共享信息，本小节还探究各类别的关键点排列顺序对模型性能的影响。在原数据集中，各类别K个关键点的默认排列顺序具有一定关联，例如虽然鸟类和马类有不同的关键点数量，但第一个关键点偶然地对应两个类别的眼睛，这导致模型（例如Sim. Base.[15]的输出卷积层的第一个通道）可偶然地共享学习“鸟类眼睛”和“马类眼睛”。然而这种方式依赖于关键点排列顺序，难以有效地泛化，因为数据集难以保证所有具有潜在联系的关键点都处于同一次序，尤其是在有大量物体类别的情况下。为了更好地模拟符合现实应用中的真实情况，本实验中对各类别的关键点随机排序，并用三次随机排列来降低随机的偶然性。

表4汇总了本文方法以及三个代表性基准方法在关键点乱序实验的结果，可以发现本文模型不但在默认顺序上有最优的性能，而且对于关键点排列顺序有更稳健的鲁棒性。可能由于默认顺序有助于本文模型的共享查询向量分配关系快速收敛，所以将关键点随机排序后，本文模型的性能有略微下降。由此可见，本文模型可自适应地学习多个类别中的共享信息，能更准确地在现实应用场景中为多类别物体估计姿态。

4.7 消融实验

为了更好地分析本文模型中各个模块的性能贡献，本小节实验评测各种模块组合的性能。在表5中，“初始阶段”指待评测模型仅具有最普通的初始阶段，即使用$[y;\delta(P);\delta(V)]$作为输出结果，并删去用于反向融合的独有反馈队列Q_s。“精化阶段”指待评测模型具有第二个阶段，即使用$[y;P;V]$作为输出结果，但仅用独有查询向量E_c输入解码器。“前向融合”指待评测模型前向融合共享信息与独有信息，即在精化阶段中将$\delta(\boldsymbol{E}'_s)+\boldsymbol{E}_c$输入解码器。“反向融合”指待评测模型反向融合共享信息与独有信息，即在初始阶段中将$\boldsymbol{E}_s+\bar{Q}_s$，输入解码器。

从表5中汇总的各模块组合的性能可以发现，当模型仅用初始阶段中共享查询向量估计的关键点作为结果也有令人满意的性能（即62.1%mAP），显示了本文模型的基础性能。在直接添加精化阶段后，本文模型在初始关键点的基础上得到了更准确的关键点（即+1.6%mAP），显示了直接精化的性能提升。在本文模型启用了前向融合后，估计的关键点准确度得到了显著提升（即+4.5%mAP），充分显示了将共享与独有信息依据匹配关系进行自适应融合的效益。在继续启用反向融合后，模型的姿态估计性能得到进一步提升（即+2.4%mAP），显示了将类别独有信息反向融合进共享查询向量，对共享信息的增强效果。总体上，本文模型的各部分模块均发挥着互补作用，共同启用时使本模型的性能达到最优。

4.8 模型设置分析

本小节的实验研究本文模型的不同设置对于多类别姿态估计性能的影响，包括两个队列长度的设置，共享查询向量个数的设置以及两个均衡超参数的设置。在本实验中，本文模型在一定范围内单独

表3 各种方法在各个超类上的性能表现mAP-superclass（mAP%）

方法	Human Body	Human Hand	Human Face	Animal Body	Animal Face	Clothes	Vehicle	Furniture
Sim. Base.	61.2	70.8	51.3	54.1	62.7	51.3	35.2	52.0
GroupPose	72.7	81.2	55.9	61.5	73.4	64.8	35.9	69.0
本文方法	81.0	92.1	73.9	66.1	76.8	73.4	39.5	77.8

表4　关键点乱序实验性能比较表(mAP%)

方法	默认	随机1	随机2	随机3	平均
Sim. Base.	55.8	50.2	51.6	49.7	51.8
ED-Pose	63.9	57.1	57.8	57.3	59.0
GroupPose	65.3	58.5	59.3	58.5	60.4
本文方法	**70.6**	**69.8**	**70.1**	**69.2**	**69.9**

表5　本文模型的模块消融性能表(mAP%)

初始阶段	精化阶段	前向融合	反向融合	mAP
√				62.1
√	√			63.7
√	√	√		68.2
√	√	√	√	70.6

地调整某个设置，并保持其余设置不变，并将模型的性能汇总在表6和图6中。

表6　设置不同队列长度的性能表(mAP%)

队列名	数据集	10	20	40	80	160
匹配代价$\boldsymbol{Q}_c$	MP-100	65.2	68.7	70.6	70.8	70.9
	Keypoint-5	90.4	91.3	91.7	91.8	91.8
独有反馈$\boldsymbol{Q}_s$	MP-100	69.5	70.3	70.6	70.6	70.7
	Keypoint-5	91.3	91.6	91.7	91.7	91.8

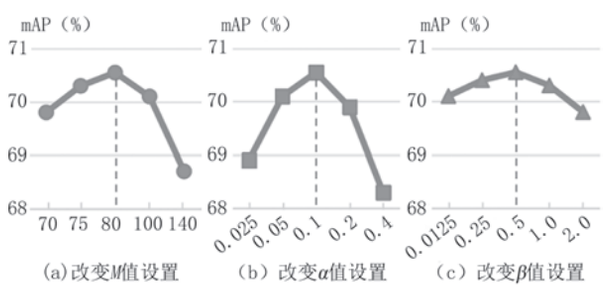


图6　不同的超参数设置对模型的性能影响图

如表6前两行所示，当匹配代价队列$\boldsymbol{Q}_c$的长度(默认为40)较短时，本模型难以对共享查询向量做出准确匹配，所以模型性能受显著影响。当长度逐渐增加，已有足够的代价矩阵提供匹配，所以性能逐渐饱和。如表6后两行所示，当独有反馈队列$\boldsymbol{Q}_s$的长度逐渐增加，本模型可保存更多的独有查询向量来提炼共享信息，但队列长度超过40时，模型性能趋近饱和。同时，本文设置的队列长度在两个数据集上的表现也较为稳定。总体上，两个队列长度的默认设置(即40)已足够且较为均衡。超参数M设置着共享查询向量的个数，过小的M会迫使单个共享查询向量去学习不相关的关键点类型，而过大的M难以学习有潜在联系的关键点类型。本模型使用的$M=80$取得了较好的均衡，如图6(a)所示。超参数α和β分别设置着可见度分类损失和物体分类损失的均衡，其影响分别如图6(b,c)所示。总体上，当各设置在合理区间内，本模型具有一定鲁棒性。关于上述超参数的分析尚未有严格的理论证明，可能存在局部最优解问题。本文的重点是探究全监督的多类别姿态估计这一任务，设计能够较好地为多个类别估计姿态的单个模型。因此，最优超参数的理论分析将在后续工作中完善。

4.9　共享查询向量的匹配可视化分析

本文模型的关键之一是共享查询向量的匹配，这支撑着类别共享与独有信息的自适应融合，所以本小节使用可视化的方式对其进行深入分析。首先，在本模型中，共享查询向量可以自适应与各类别的关键点进行匹配。由于有些类型的关键点比较常见(如眼睛)，而有些比较罕见(如桌角)，所以这种匹配关系并不是均匀的。因此，本小节将各个共享查询向量的累计匹配次数统计在图7中。具体地，在测试过程中，本文模型对匹配到可见真实关键点的共享查询向量进行次数加1，然后把各共享查询向量按次数从小到大排列，绘制如图7所示。

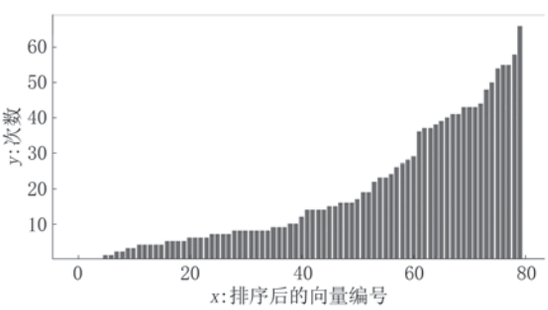
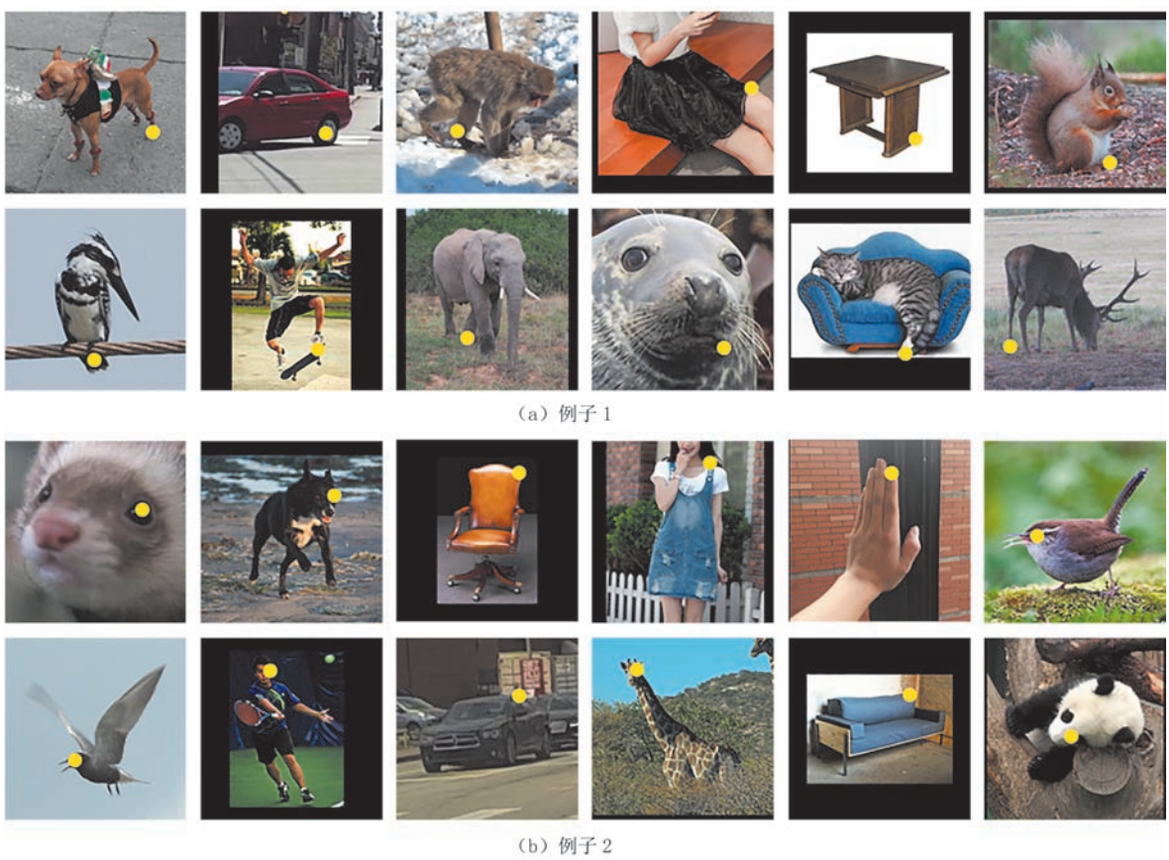


图7　共享查询向量匹配到关键点的次数累计图

由此可以发现，某些共享查询向量的匹配次数较高，可匹配到不同类别图像中具有潜在联系的关键点，从而习得类别之间的共享信息。也有些共享查询向量的匹配次数较低，负责匹配各类别中那些潜在联系较少的关键点。然后，图8直观地显示共享查询向量具体对应的关键点。具体地，本可视化过程依据各共享查询向量的累计匹配次数，选择了2个有代表性的高频共享查询向量进行可视化。对于指定的第m个共享查询向量，每个可视化子图显示它在各类测试图像中解码得到的关键点，即$P_s[m]$。图8(a)中可视化的共享查询向量较好地习得了“各类物体左腿”的共享信息，能在各类图像中准确地定位具有潜在联系的“左腿”关键点。例如狗类、鸟类、猴类这些具有较大差异类别的“左腿”，甚

至包括汽车、裙子、桌子这些具有显著差异类别的、高度抽象但有潜在联系的“左腿”。而图8(b)中可视化的共享查询向量较好地习得了“各类物体左眼”的共享信息，也具有相似的观察。值得注意的是，本文方法并没有依赖标注的关键点排列顺序来强行绑定各类别关键点之间的联系，因为这难以对所有关键点进行准确标注，尤其是在类别数量众多的情况下。由此可见，本文模型中的共享查询向量可自适应地习得各类别之间的共享信息，并充分利用共享与独有信息的双向融合来进行精准的姿态估计。



(a) 例子1

(b) 例子2

图8 共享查询向量定位的关键点可视化图，其中每个例子展示了某个向量在各类图像中定位的关键点

5 总 结

本文顺应姿态估计发展趋势，探究了多类别姿态估计任务，致力于用单个模型为多个类别的物体估计姿态。本文提出了基于类别共享与独有信息双向融合的Transformer模型，其中依据匹配关系对共享与独有信息进行自适应融合。本文模型使用查询向量来表征各类关键点的共享和独有信息，并用基于反向融合的初始阶段和前向融合的精化阶段逐步地估计各类物体的准确关键点。本文在多类别姿态数据集MP-100上进行了丰富且深入的实验，其中的定量和定性分析都充分证明了本方法的有效性。

参 考 文 献

[1] Li Jia-Ning, Wang Dong-Kai, Zhang Shi-Liang. Deep-learning-based 2D human pose estimation: present and future. Chinese Journal of Computers, 2024, 47(1): 231-250 (in Chinese)
(李佳宁，王东凯，张史梁. 基于深度学习的二维人体姿态估计：现状及展望. 计算机学报，2024, 47(1): 231-250)

[2] Dang Q, Yin J, Wang B, Zheng W. Deep learning based 2d human pose estimation: a survey. Tsinghua Science and Technology, 2019, 24(6): 663-676

[3] Sapp B, Taskar B. Multimodal decomposable models for human pose estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 3674-3681

[4] Ma Sheng-Lei, Li Jing-Hua, Kong De-Hui, et al. 3D hand pose estimation based on double branches with multi-scale attention. Chinese Journal of Computers, 2023, 46(7): 1383-

1395 (in Chinese)
(马胜蕾，李敬华，孔德慧，等，基于双分支多尺度注意力的手三维姿态估计. 计算机学报，2023, 46(7): 1383-1395)

[5] Jia Di, Li Yu-Yang, An Tong, et al. Complex gesture pose estimation network fusing multiscale features. Journal of Image and Graphics, 2023, 28(9): 2887-2898 (in Chinese)
(贾迪，李宇扬，安彤，等，融合多尺度特征的复杂手势姿态估计网络. 计算机图象图形学报，2023, 28(9): 2887-2898)

[6] Song X, Wang P, Zhou D, et al. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5452-5462

[7] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. Online, 2021: 8748-8763

[8] Kirillov A, Mintun E, Ravi N, et al. Segment anything//Proceedings of the IEEE International Conference on Computer Vision. Paris, France, 2023: 4015-4026

[9] Xu L, Jin S, Zeng W, et al. Pose for everything: towards category-agnostic pose estimation//Proceedings of the European Conference on Computer Vision. Tel-Aviv, Israel, 2022: 398-416

[10] Shi M, Huang Z, Hu X, et al. Matching is not enough: a two-stage framework for category-agnostic pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 7308-7317

[11] He Xiao-Jian, Lin Jin-Fu. Weakly-supervised object localization based fine-grained few-shot learning. Journal of Image and Graphics, 2022, 27(7): 2226-2239 (in Chinese)
(贺小箭，林金福. 融合弱监督目标定位的细粒度小样本学习. 计算机图象图形学报，2022, 27(7): 2226-2239)

[12] Yang J, Zeng A, Zhang R, et al. X-pose: Detecting any keypoints//Proceedings of the European Conference on Computer Vision. Milan, Italy, 2024: 249-268

[13] Zhang X, Wang W, Chen Z, et al. CLAMP: prompt-based contrastive learning for connecting language and animal pose//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 23272-23281

[14] Zhao Peng, Wang Chun-Yan, Zhang Si-Ying, et al. A zero-shot lmage classification method based on subspace learning with the fusion of reconstruction. Chinese Journal of Computers, 2021, 44(2): 409-421 (in Chinese)
(赵鹏，汪纯燕，张思颖，等. 一种基于融合重构的子空间学习的零样本图像分类方法. 计算机学报，2021, 44(2): 409-421)

[15] Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 466-481

[16] Chu X, Yang W, Ouyang W, et al. Multi-context attention for human pose estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 1831-1840

[17] Yang W, Li S, Ouyang W, et al. Learning feature pyramids for human pose estimation//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1281-1290

[18] Hu S, Sun H, Wei D, et al. Continuous heatmap regression for pose estimation via implicit neural representation//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2024, 37: 102036-102055

[19] Lu P, Jiang T, Li Y, et al. RTMO: towards high-performance one-stage real-time multi-person pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 1491-1500

[20] Wei S, Ramakrishna V, Kanade T, et al. Convolutional pose machines//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4724-4732

[21] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 483-499

[22] Chen Y, Shen C, Wei X S, et al. Adversarial posenet: A structure-aware convolutional network for human pose estimation//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1212-1221

[23] Bin Y, Chen Z M, Wei X S, et al. Structure-aware human pose estimation with graph convolutional networks. Pattern Recognition, 2020, 106: 107410

[24] Zeng W, Jin S, Liu W, et al. Not all tokens are equal: Human-centric visual analysis via token clustering transformer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 11101-11111

[25] Xu Y, Zhang J, Zhang Q, et al. ViTPose: simple vision transformer baselines for human pose estimation//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 38571-38584

[26] Fan X, Zheng K, Lin Y, et al. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1347-1355

[27] Sun X, Shang J, Liang S, et al. Compositional human pose regression//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2602-2611

[28] Shi D, Wei X, Li L, et al. End-to-end multi-person pose estimation with transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 11069-11078

[29] Toshev A, Szegedy C. DeepPose: Human pose estimation via deep neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1653-1660

[30] Li K, Wang S, Zhang X, et al. Pose recognition with cascade transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online, 2021: 1944-

1953

[31] Mao W, Ge Y, Shen C, et al. Poseur: Direct human pose regression with transformers//Proceedings of the European Conference on Computer Vision. Israel, 2022: 72-88

[32] Xiao Y, Su K, Wang X, et al. QueryPose: sparse multi-person pose regression via spatial-aware part-level query//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 12464-12477

[33] Geng Z, Wang C, Wei Y, et al. Human pose as compositional tokens//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 660-671

[34] Yang J, Zeng A, Liu S, et al. Explicit box detection unifies end-to-end multi-person pose estimation//Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 2023

[35] Liu H, Chen Q, Tan Z, et al. Group pose: A simple baseline for end-to-end multi-person pose estimation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 15029-15038

[36] Wei X S, Xu H Y, Yang Z, et al. Negatives make a positive: An embarrassingly simple approach to semi-supervised few-shot learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 46(4): 2091-2103

[37] Xu S L, Zhang F, Wei X S, et al. Dual attention networks for few-shot fine-grained recognition//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2022, 36(3): 2911-2919

[38] Lu C, Koniusz P. Few-shot keypoint detection with uncertainty learning for unseen species//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 19416-19426

[39] Sun M, Zhao Z, Chai W, et al. Uniap: Towards universal animal perception in vision via few-shot learning//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024, 38(5): 5008-5016

[40] Zhang H, Xu L, Lai S, et al. Open-vocabulary animal keypoint detection with semantic-feature matching. International Journal of Computer Vision, 2024:1-18

[41] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers//Proceedings of the European Conference on Computer Vision. Online, 2020: 213-229

[42] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection//Proceedings of the International Conference on Learning Representations, Online, 2021

[43] Li F, Zhang H, Liu S, et al. Dn-detr: Accelerate detr training by introducing query denoising// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 13619-13627

[44] Cheng B, Schwing A, Kirillov A. Per-pixel classification is not all you need for semantic segmentation//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2021, 34: 17864-17875

[45] Cheng B, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 1290-1299

[46] Chen J, Niu L, Zhou S, et al. Weak-shot semantic segmentation via dual similarity transfer//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 32525-32536

[47] Kuhn H W. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly, 1955, 2(1-2): 83-97

[48] Chen J, Yan J, Fang Y, et al. Meta-point learning and refining for category-agnostic pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 23534-23543

[49] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerlan, 2014: 740-755

[50] Sagonas C, Antonakos E, Tzimiropoulos G, et al. 300 faces in-the-wild challenge: Database and results. Image and Vision Computing, 2016, 47: 3-18

[51] Wu J, Xue T, Lim J J, et al. Single image 3d interpreter network//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 365-382

[52] Wang Y, Peng C, Liu Y. Mask-pose cascaded cnn for 2d hand pose estimation from single color image. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(11): 3258-3268

[53] Wah C, Branson S, Welinder P, et al. Caltech-ucsd birds 200. California Institute of Technology, CNS-TR-2010-001, 2010

[54] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019, 32: 8024-8035

[55] ContributorsMMPose. Openmmlab pose estimation toolbox and benchmark, https://github. com/open-mmlab/mmpose 2020

[56] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 248-255

[57] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[58] Kingma D P. Adam: A method for stochastic optimization. arxiv preprint arxiv:1412.6980, 2014

CHEN Jun-Jie, Ph. D., lecturer. His main research interests include computer vision, pose estimation, weakly supervised learning, transfer learning.

CHEN Wei-Long, M. S. His main research interests include computer vision, pose estimation, image segmentation.

FANG Yu-Ming, Ph. D., professor. His main research interests include computer vision, machine learning.

JIANG Wen-Hui, Ph. D., associate professor. His research interests include image content understanding and cross media analysis.

NIU Li, Ph. D., associate professor. His main research interests include computer vision, machine learning.

Background

Pose estimation aims to locate the keypoints of objects in 2D images, which is a fundamental and important task in computer vision. Pose estimation has extensive applications in the real world, including virtual reality, augmented reality, human-computer interaction, robot and automation.

Most existing method focus on single-class pose estimation (e. g., human, hand or vehicle), and neglect that the target objects may come from multiple classes. In such case, we may have to train a model for each class respectively and selectively apply them to estimate multi-class pose, which is complex and unpractical. On the other hand, the models in dectection and segmentation tasks all are able to employ a single model to predict multi-class results. Therefore, it is significant to explore multi-class pose estimation, where we would like to estimate the poses for multi-class objects with a single model.

The key issue in multi-class pose estimation is how to adaptively integrate the shared and specific information across classes. In most existing models, the output layers correspond to the object keypoints in a fixed manner, which cannot effectively learn the shared and specific information. Specifically, the first output channel always learn the first annotated keypoint in all training images, while these keypoints belong to various keypoint types of various object classes. Therefore, the model may be disturbed by various specifc patterns, and also neglect to use the shared information for better estimation.

In this paper, we propose a Transformer model based on bidirectional integration of shared and specific information. Specifically, we employ learnable queries to represent the shared and specific information of keypoints, and estimate keypoints by two stages. In the initial stage, shared queries aggregate the backbone feature maps and then produce initial keypoints and object class. In the refinement stage, the share queries are forward-integrated with the specific queries based on the bipartite matching between shared queries and specific keypoints, and then estimate refined keypoints. Furthermore, the updated specific queries are stored in a queue and backward-integrated into the shared queries, which could enhance the shared information. Therefore, our model leverages bidirectional integration to extract multi-class information and thus better achieve multi-class pose estimation. We conduct extensive experiments on the Multi-category Pose (MP-100) dataset, and the quantitative and qualitative analyses demonstrate the effectiveness of our method.

This work was supported in part by the National Natural Science Foundation of China under Grants U24A20220, 62132006, 62402201 and 62161013, in part by the National Key Research and Development Program of China under Grants 2023YFE0210700, in part by the Natural Science Foundation of Jiangxi Province of China under Grants 20242BAB21006 and 20242BAB23012, and in part by the Jiangxi Province Special Program for Cultivating Early-Career Young Scientific and Technological Talents under Grants 20244BCE52070.