ChatGPT

2024–2025年大规模姿态估计模型跨域微调与自适应研究综述

背景与挑战

大型**2D人体姿态估计**模型（例如ViTPose、Sapiens等）在通用场景下取得了出色性能，但当应用于**新领域人群**（如婴幼儿、老年人、轮椅使用者等）时，往往遇到显著的**域差异**导致性能下降。这些人群在体型比例、典型姿势和肢体可见性方面与常规训练数据（多为成年人体姿态）存在差异。例如，将成人姿态模型直接用于婴儿，会因骨骼长度差异和全新动作模式而效果不佳；同样，普通模型对轮椅使用者的下肢关键点经常预测错误，因为轮椅造成了严重遮挡。此外，这些特殊人群的数据标注通常**稀缺**且分布偏离常规数据，存在关键点定义不一致或缺失的问题。这些挑战促使研究者探索**创新的微调策略和领域自适应方法**，将预训练的大模型高效适配到新领域。近期（2024–2025年）在CVPR、ICCV、ECCV等顶会出现了一系列相关工作，本文按2D为主（适当涵盖3D）进行综述，重点介绍面向罕见人群的姿态估计微调与域适应新进展，并分析其模型结构、微调方法、新颖性、适应机制及评测结果。

现代姿态估计大模型为跨域适应提供了基础。例如，Meta提出的**Sapiens**基础模型在3亿张多人图像上进行自监督预训练，然后可通过微调**快速适配**到各项人体任务。Sapiens采用ViT架构（高达20亿参数）并支持1024×1024高分辨率输入，经微调在2D人体208关键点检测、人体部位分割、深度估计等多任务上取得新SOTA。这表明大规模预训练模型具有**出色的跨域泛化**能力和微调潜力。同时，也出现了参数高效微调方法：有研究在小数据微调大模型时，仅**训练轻量线性层**以避免过拟合。在以下小节中，我们将按人群类别和方法类别，介绍用于跨域姿态估计的代表性工作。

面向婴幼儿的姿态估计域适应

婴幼儿由于**身材比例悬殊**（头身比大、四肢短）、动作模式不同（经常蜷缩、趴卧）、且标注数据稀少，成为姿态估计领域适应的难点。早期工作尝试构建婴儿合成数据并**有监督微调**模型。例如Hesse等利用Skinned Multi-Infant Linear模型（SMIL）生成合成婴儿数据，构建了MINI-RGBD数据集；此后Huang等发布了包含真实+合成婴儿的SyRIP数据集，并提出**FiDIP模型**，通过在合成+真实婴儿数据上进行对抗域微调（加入域分类器）来将成人姿态模型适应婴儿。FiDIP在有限婴儿数据上取得了当时较佳精度（mAP 92.2）。然而，该全监督微调方式需要大量标注婴儿数据，且对不同婴儿数据集泛化不佳。

为避免对婴儿姿态的人工标注依赖，Bose等在CVPR 2025提出**SHIFT**框架[1]。SHIFT是首个**无监督领域自适应**的婴儿姿态估计方法，将在成人数据上预训练的2D姿态模型适配到婴儿域[2]。其核心是在**Mean Teacher**框架下，用未经标注的婴儿图像进行自适应训练：学生模型负责预测婴儿关键点，教师模型为学生的指数移动平均（EMA）权重，用于产生伪标签。通过强制学生预测与教师输出一致，利用大量**伪标签一致性约束**实现跨域适应。此外，SHIFT引入了两个**创新正则项**[3]：一是离线训练的**婴儿姿态先验**模型，对每次学生预测打分，惩罚不合理的人体姿态[3]；二是**可见性一致性模块**，利用预训练人体分割网络获得婴儿躯体掩膜，让一个关键点到分割的转换网络保证预测关键点分布与婴儿轮廓相符。这种姿态与图像上下文对齐的方法有助于在婴儿肢体自遮挡情况下仍预测正确。

图1：SHIFT无监督领域适应框架示意图。该方法以预训练的成人姿态模型为基础，通过学生模型–教师模型结构在婴儿无标注图像上自训练实现适应。红色模块表示需要学习的部分，包括学生模型M_s 及其EMA更新生成的教师模型M_t ，用于剔除不合理结果的婴儿姿态先验θ_p ，以及利用分割掩膜执行姿态-图像可见性对齐的Kp2Seg模块等。

经过上述创新，SHIFT在多个婴儿数据集上实现了显著性能提升：在SyRIP和MINI-RGBD数据集的无监督适应中，比现有UDA方法提升约5%，甚至**超越有监督SOTA方法16%**之多 [4] 。值得注意的是，SHIFT不需要任何婴儿关键点标注，实现了真正意义上的跨域迁移，这对隐私敏感且难以标注的婴儿场景具有重要意义。

除了2D关键点检测，婴儿的**3D姿态估计**近期也有突破性进展。Zhou等人在WACV 2024提出了**ZeDO-i方法**（基于优化的零数据适应）。他们从**生成模型先验**角度切入，首先训练了一个基于扩散模型的3D姿态生成器，在大规模**成人**3D姿态数据上获得先验能力。为适应婴儿域，ZeDO-i提供两种策略：（1）**可控分支微调**：复制先验生成器的权重，加入一个可学习的“姿态提示”张量作为条件输入，让生成器输出偏向婴儿姿态；（2）直接利用生成先验进行**优化推理**：给定婴儿2D关键点，通过迭代优化生成器的输入向量，使其投影与2D观测对齐，从而恢复3D婴儿姿态。此外，作者还利用**Diffusion扩散模型**对成人3D姿态进行**域转移**，合成了逼真的婴儿姿态数据用于数据增强。实验证明，即使只有少量婴儿数据，ZeDO-i仍能实现高效的3D姿态域适应，在SyRIP和MINI-RGBD数据集上将SOTA的平均关节误差（MPJPE）分别降低到43.6mm和21.2mm，创下新记录。该方法展示了**利用预训练生成先验+微调**来弥合成人与婴儿3D姿态差异的新思路，在数据匮乏时尤为有效。

面向轮椅使用者的姿态估计改进

轮椅使用者是另一典型的**未充分代表人群**。其挑战在于轮椅对人体下肢的严重遮挡，以及姿势局限（坐姿为主）导致的训练数据分布差异。常规姿态模型在此场景下表现不佳：例如Detectron2的Keypoint R-CNN在轮椅场景常将轮椅错检为人体腿部，或干脆漏检下肢。为提高对轮椅使用者的识别能力，Huang等在CHI 2024提出**WheelPose数据合成与微调框架**。该工作专注于**数据层面**的解决方案：开发了一条可配置的**Unity仿真管线**，生成多样逼真的轮椅人像合成数据，用于微调现有姿态模型。

具体来说，WheelPose管线包括：从**动作捕捉数据**或运动生成模型获取带标注的3D关节序列，对下肢姿态进行适当修改以符合轮椅坐姿，然后在Unity中将关节序列附加到3D人体模型，上下肢做**参数化随机化**（如不同下肢残缺情况），配以多样的轮椅模型和背景场景，渲染出大规模**合成图像及其关键点标注**。作者还引入了**人工评价环节**，让人类筛选不真实的动作并反馈，以提升数据质量。生成的数据具有高逼真度和丰富多样性，并**公开发布**供社区使用。

利用WheelPose合成数据，作者将预训练的COCO人体姿态模型在其中**微调**，再在真实轮椅用户数据集上测试。结果显示，添加合成数据微调后，模型对轮椅使用者的检测精度显著提升。例如，对于常被遮挡的脚踝关键点，微调模型的检测率提高了**76.1%**。视觉上看，微调后模型不再将轮椅误识别为人体的一部分，并能**合理“填补”被遮挡的腿部关键点**，而非像原模型那样将轮椅当作缺失的腿部。WheelPose的研究表明，通过**数据合成弥补训练集空白**，结合微调可以有效提升现有模型对残障人群的公平性能。这种方法在无须更改模型结构的情况下，从数据端提高了算法对轮椅人群的鲁棒性，为打造**包容性的AI**提供了范例。

通用域自适应与微调策略

除了针对特定人群的方案，近年来还出现了一些**通用的姿态估计域适应方法**和微调策略创新，可广泛用于跨数据集或跨环境的模型自适应。

1. 无源域适应（Source-free DA）：传统的域适应方法通常需要源域数据参与（同时使用源和目标数据缩小分布差异），但这在实际中可能因**数据隐私或体量**无法满足。为此，Peng等在ICCV 2023提出了**源数据不可用的人体姿态自适应**新任务，并设计了对应方法。其框架包含**源模型、中过渡模型和目标模型三部分**：首先通过一个“**源保护模块**”保存源模型知识防止遗忘，同时通过“**目标相关模块**”从仅有的目标未标注数据中提取可迁移的姿态特征。具体技术上，他们构建了一个**空间概率图**来表示人体各部位出现的位置几率，并在此基础上引入**姿态级的对比学习**和信息最大化损失，增强模型对目标姿态分布的表征能力。整个适应过程不访问源数据，仅依赖预训练源模型权重和未标注目标图像，通过逐步更新中间模型再传递到目标模型，实现了知识迁移。实验在多个人体姿态跨域基准上显示，该方法较以往需要源数据的方法有明显性能优势。这一工作凸显了**源数据隔离条件下**的姿态自适应新思路，对于隐私敏感的数据场景（如医疗、安防）具有现实意义。

2. 测试时快速自适应：另一类进展是**在线/测试时域适应**，即模型在推理阶段利用未标注的目标数据自行调整，从而实时适应新域。Hu等在CVPR 2024提出的方法代表了这一路线。他们注意到现有测试时自适应需要大量梯度更新步骤，既耗时又易遗忘源知识。为此引入**元学习（Meta-learning）**思想：在训练阶段除了主任务（姿态监督）外，增加一个**自监督辅助任务**“人体部位图像修复”。具体做法是在训练集中对人物图像随机遮挡，然后让一个生成网络学习根据上下文重建被遮挡的人体区域（类似人体部位的图像补全）。这一辅助任务逼迫模型关注人像的整体语义结构，从而学到对人体关键点更稳定的表示。接着通过元训练，使模型找到**易于适应的新参数初始化**：在训练后期模拟域_shift_场景，只用很少几步优化就能让模型在新分布上性能提升，同时保持原有知识不丢失。最终在测试阶段，模型针对每张新图像执行少量自监督优化（利用上述辅助任务在目标图像上继续训练几步），即可实现**快速自适应**，大幅改善新域关键点检测精度。该方法证明，通过**元优化+自监督任务**，可以减小测试时自适应的开销并避免灾难性遗忘，为实时应用（如穿戴设备摄像头的姿态分析）提供了切实可行的方案。

3. 全局-局部分布对齐（3D跨数据集）: 针对跨数据集的泛化问题，Chai等在ICCV 2023提出了**PoseDA**框架，将3D姿态提升模型从一个数据集适应到另一个数据集。作者发现域差异主要来自**相机坐标系的全局位移**差异和**动作模式的局部多样性**不足。因此PoseDA包含两大模块：（1）**全局适配(Global Adaptation)**：提出全局位置对齐（GPA）机制，利用目标域的2D关键点分布指导源域3D姿态进行平移变换，使两者的**重心和视角分布**对齐。（2）**局部泛化(Local Generalization)**：设计局部姿态增强（LPA）模块，通过一个**对抗式数据增强器**生成一系列局部扰动（如关节角度随机变化、四肢长度拉伸等）来丰富3D姿态训练集，同时判别器确保增强姿态符合真实人体解剖。这两个模块**不引入额外模型参数**，而是在数据层面对源姿态进行变换与增强。实验在跨数据集评测（如用H36M训练在MPI-INF-3DHP测试）中证明，PoseDA显著提高了3D姿态提升的泛化性能：MPI-INF-3DHP上的MPJPE降低至61.3mm，比此前最佳降低了**10.2%**。这表明针对不同摄像机设置（全局差异）和动作覆盖范围（局部差异）采取定制策略，可以有效地**微调**提升网络的跨域适应性。

4. 其它策略: 此外，还有一些值得关注的思路。例如，有工作将**对比学习和图神经网络**引入姿态域适应，通过构建关系图来校正跨域关键点关系，提高鲁棒性。又如在极端环境下的域适应，Ai等研究了**低光照姿态估计**，提出双教师网络生成可靠伪标签，使模型无需低光真值标注也能在暗光图像上取得与有标注方法相当的性能。这些方法进一步拓展了域适应的应用范围。与此同时，**通用多体态模型**的出现（如ViTPose++）尝试用统一的Transformer框架解决不同类型的姿态估计任务，包括全身、手部、面部等，通过知识蒸馏融合不同关键点模式，从而方便针对特定组合任务进行微调适应。例如ViTPose系列采用纯Transformer主干和简洁解码器即可在COCO等数据集上达到SOTA精度，展示了简化结构、扩大模型规模带来的强大表征能力。这些进展预示着未来的姿态估计模型将更具**通用性与可适应性**。

小结与展望

近年来，借助大规模预训练模型和创新的微调/域适应技术，人体姿态估计正变得更加**包容和实用**。通过在算法中显式引入目标人群的先验（如婴儿解剖模型）、利用合成数据弥补标注空白、以及发展源数据无依赖的自适应优化策略，研究者成功将成人姿态模型扩展到了婴幼儿、残障人士等**非典型人群**，显著提高了这些场景下的关键点检测准确率和可靠性。在2D领域，SHIFT等无监督方法证明即使没有目标域标注也能高效迁移模型；在3D领域，生成先验和分布对齐技术有效解决了跨域的坐标差异和动作差异。这些方法不仅在学术指标上取得领先 [4]，更具有重要的社会意义：提升医疗康复监测中婴儿和老年人的动作分析精度，增强智能健身或游戏对残障群体的友好性，以及在安防中更全面地识别坐姿人群等。

展望未来，**基础模型+领域自适应**将是关键趋势。预训练的超大规模姿态模型（如Sapiens）提供了强大的通用特征，未来可结合高效微调策略（如适配层、LoRA等）针对特定人群做细化调整，从而以极少的新数据获得最优效果。同时，跨模态信息的引入也是一方向，例如结合深度传感或穿戴式IMU数据辅助图像姿态估计，以克服单目图像的模糊性——近期就有利用少量IMU传感器专为轮椅用户进行实时姿态估计的尝试。最后，**开放多样的数据集**建设仍然重要：社区应持续收集和发布更多婴幼儿、老年人、残障人士的标注数据（在确保伦理和隐私的前提下），以推动训练更健壮的模型。总的来说，2024–2025年的研究已经为**稀有人群姿态估计**铺平了道路，融合大模型、合成数据与自适应算法的方案展现了巨大潜力。随着这些技术的成熟，我们有望见到姿态

估计模型真正做到“以人为本”，在各种体态和情境下都能提供可靠准确的关键点检测，为医疗健康、智能交互和机器人等领域创造更大价值。

表1：精选方法概览 （2024–2025）

方法 / 文献	任务类型	模型与架构	微调/适应策略	新颖性与贡献	评估结果
SHIFT [1] [4] (CVPR’25 Workshop)	2D姿态 UDA（成人→婴儿）	基于ViT的检测器；教师–学生双网络架构	无监督域适应：Mean Teacher伪标签一致性 + 婴儿姿态先验 + 关键点-分割可见性约束	首个成人→婴儿无监督适应；解决骨骼差异和自遮挡；无需婴儿标注	SyRIP数据集提升5%，超有监督SOTA 16% mAP [4]
ZeDO-i (WACV’24)	3D姿态 小数据 (成人→婴儿)	优化+生成模型：Diffusion先验 + 控制分支	少样本域适应：利用大规模成人3D先验，微调生成模型 + 扩散生成婴儿姿态数据增强	提出可控生成分支用于域迁移；将扩散模型引入姿态DA	SyRIP MPJPE 43.6mm，MINI-RGBD 21.2mm，均创当时最好
WheelPose (CHI’24)	2D姿态 微调 (健全→轮椅)	Detectron2等现有模型	数据增广微调：Unity合成轮椅用户图像 + 在此基础上微调	提出完整仿真管线生成残障数据；通过人评筛选提高数据质量	微调后踝关节检测率提高76%；下肢关键点预测错误大幅减少
源自由DA (ICCV’23)	2D姿态 UDA (合成→真实)	任意pose模型 (需源预训练)	无源域适应：源保护模块 + 目标相关模块 (空间概率+对比学习)	首提姿态源自由DA任务；设计三模型框架防遗忘与降噪	多个跨域基准上优于传统有源方法，提升显著
MetaAdapt (CVPR’24)	2D姿态 测试时适应	HRNet等主干 + 辅助Decoder	元学习+自监督：训练时加人体修复辅助任务，学得快速适应初始化；测试时少量迭代自适应	将元优化引入姿态TTT(test-time training)；设计人体部位重建任务聚焦人体语义	在跨相机、低质图像等测试域上，实现用极少更新达到与长期训练相当性能
PoseDA (ICCV’23)	3D姿态 UDA (跨数据集)	两阶段(2D检测+3D提升)任意提升模型	无监督域适应：全局位置对齐(GPA) + 局部姿态增强(LPA，对抗生成多样姿态)	将域差异分解为全局/局部两部分并分别处理；无新增参数	H36M→MPI-INF 3DHP的MPJPE降低10.2%，在几种跨集设置下均提升明显

上述方法共同推动了大模型在稀缺数据领域的应用。未来，随着更多**基础模型**开放以及**自动化数据合成**技术发展，我们有望以更低成本开发出覆盖全人群的高精度姿态估计模型，真正实现“人人皆准”的目标。

---

[1] [2] [3] [4] Leveraging Synthetic Adult Datasets for Unsupervised Infant Pose Estimation
https://arxiv.org/html/2504.05789v1