arXiv:2303.02760v2 [cs.CV] 5 Apr 2023

# Human-Art：连接自然场景与人工场景的多功能以人为中心的数据集

Xuan Ju[1, 2*†], Ailing Zeng[1*‡], Jianan Wang[1], Qiang Xu[2], Lei Zhang[1]国际数字经济研究院, [2]香港中文大学{xju22, qxu}@cse.cuhk.edu.hk, {zengailing, wangjianan, leizhang}@idea.edu.cnhttps://idea-research.github.io/HumanArt/

## 摘要

自古以来，人类就以各种形式被记录下来。例如，在相机发明之前，雕塑和绘画是描绘人类的主要媒介。然而，目前大多数以人为中心的计算机视觉任务，如人体姿态估计和人体图像生成，都只关注现实世界中的自然图像。雕塑、绘画和卡通中的人工人体通常被忽视，导致现有模型无法适用于这些场景。

作为对生活的抽象，艺术融入了自然与人工场景中的人体。借助这一点，我们推出了 ***Human-Art*** 数据集，以衔接自然与人工场景中的相关任务。具体而言，*Human-Art* 包含 $50k$ 张高质量图像，涵盖来自5 种自然场景和15 种人工场景的超过 $123k$ 个人物实例，并为其中以二维和三维形式呈现的人体标注了边界框、关键点、自接触点和文本信息。因此，该数据集全面且用途广泛，可用于各种下游任务。我们还为相关任务提供了丰富的基线结果和详细分析，包括人体检测、二维和三维人体姿态估计、图像生成和动作迁移。作为一个具有挑战性的数据集，我们希望 *Human-Art* 能为相关研究提供启发，并开拓新的研究问题。

## 1. 引言

“艺术源于生活，却超越生活。”

人体中心的计算机视觉（CV）任务，例如人体检测 [46]，姿态估计 [36]，动作迁移 [66]，和人体图像生成 [42]过去十年间，随着深度学习技术的发展，这些任务得到了广泛研究，并取得了显著成果。大多数此类研究使用的数据集 [16,27,32,46] 聚焦于由相机拍摄的自然场景中的人体，这既是出于实际需求，也因为这类数据易于获取。

然而，除了通过相机拍摄外，人体还经常以各种其他形式呈现和记录下来，从墙上的古代壁画、纸上的肖像画，到数字形式的计算机生成的虚拟人物。然而，现有最先进的（SOTA）人体检测和姿态估计模型 [75, 83] 在 MSCOCO 等常用数据集上训练后， [32] 通常无法在这些场景中取得良好效果。例如，这类模型在自然场景中的平均精度可高达 63.2% 和 79.8%，但在雕塑 场景中则大幅降至 12.6% 和 28.7%。一个根本原因是自然场景与人工场景之间存在域差异。此外，人工人物场景数据集的匮乏，严重限制了动漫角色图像生成 [9, 78, 86], 、角色渲染 [34], 和角色动作重定向 [1, 44, 79] 等计算机图形学及其他领域任务的发展。随着人们对虚拟现实（VR）、增强现实（AR）和元宇宙的兴趣日益浓厚，这一问题愈发严峻，亟待解决。

文献中有少量将人体纳入人工环境的数据集。Sketch2Pose [4]和 ClassArch [40] 分别收集了素描和花瓶绘画图像。因此，它们仅适用于相应的场景。People-Art [69]是一个人体检测数据集，包含 1490 幅绘画。它涵盖了多种绘画风格的人工场景，但其类别既不互斥，也不全面。更重要的是，People-Art 的标注类型和图像数量有限，因此该数据集主要用于测试（而非训练）目标检测器。

艺术以多种形式呈现自然与人工场景中的人体，例如舞蹈、绘画和雕塑。本文借鉴视觉艺术的分类方式，推出 *Human-Art*，这是一个用途广泛的以人为中心的数据集，旨在弥合自然与人工场景之间的鸿沟。*Human-Art* 采用层级结构，涵盖丰富场景中的高质量人体场景，并具有精确的

*共同贡献。
†在 IDEA 实习期间完成的工作。
‡通讯作者。




图 1. *Human-Art* 是一个用途广泛、以人为中心的数据集，旨在弥合自然场景与人工场景之间的差距。它包含 20 种高质量场景，涵盖二维表示（黄色虚线框）和三维表示（蓝色实线框）中的自然人体与人工人体。

人工标注。具体而言，它由 $50k$张图像组成，其中约有 $123k$ 个人物实例，涵盖20 个艺术类别，包括5 个自然场景和15 个人工场景，包含二维和三维场景，如图1所示。为支持识别和生成任务，*Human-Art* 提供了精准的人工标注，包括人体边界框、二维关键点、自接触点和文本描述。它可以弥补以往数据集（如 MSCOCO [32]）在场景方面的不足，连接虚拟世界与现实世界，并为以人为中心的领域带来新的挑战和机遇。

*Human-Art* 具有以下独特特点：

- **丰富的场景**：*Human-Art* 聚焦于主流数据集中缺失的场景（例如 [32]），涵盖大多数与人体相关的场景。具有挑战性的人体外观、多样的上下文以及各种姿态，在很大程度上弥补了现有数据集在场景方面的不足，并将带来新的挑战与机遇。
- **高质量**：我们确保类别间的差异性，以及风格、作者、来源和年龄方面的类别内多样性。5 万张图像是从经过标准化数据收集、筛选和整合流程精心采集的 100 万张图像中人工挑选出来的。
- **多样化的标注**：*Human-Art* 提供经过精心人工标注的二维人体关键点、人体边界框和自接触点，以支持各种下游任务。此外，我们还提供易于获取的文本描述，以支持多模态学习。

借助 *Human-Art*，我们针对人体检测、人体姿态估计、人体网格恢复、图像生成和动作迁移等多种下游任务开展了全面的实验与分析。尽管在 *Human-Art* 上进行训练，能使人体检测和人体姿态估计的性能分别提升 31% 和 21%，但结果表明，人体相关的计算机视觉任务要达到成熟仍有很长的路要走。

## 2. 相关工作

**自然场景中的以人为中心的数据集：** 以人为中心的识别的主要任务是人体检测和姿态估计。如表 1所总结，大多数现有数据集 [2, 3, 11, 27, 33, 46, 71, 85] 都使用边界框和关键点标注自然场景中的人体。其中，MSCOCO [32]因其丰富多样的姿态和复杂场景而得到最广泛的应用。使用该数据集训练的众多深度模型，在各种下游任务 [18, 25, 43, 75]中均表现出色。行人检测数据集 [30, 64, 76] 也可归类为特殊的人体检测数据集，专门关注拥挤场景中的小型且模糊的人体。

尽管自然场景数据集在计算机视觉任务中得到了广泛应用，但由于只关注自然场景，在这些数据集上训练的模型在人工场景中表现不佳。

**包含人工场景的以人为中心的数据集：** 只有少数小规模数据集 [4, 40, 69] 涉及人工场景。具体而言，Sketch2Pose [4] 专注于素描场景，而 ClassArch [40] 仅包含古代花瓶画。People-Art [69] 同时包含自然图像和人工图像。它直接借鉴了*wikiart*[1] 中的艺术绘画风格来对人工场景进行分类。类别间相似性会造成混淆，尤其是在图像未经人工审核的情况下。此外，绘画以外的艺术作品（如雕塑和数字艺术）也被忽略。更重要的是，由于图像数量有限且仅提供人体边界框标注，People-Art 仅支持小规模人体检测任务。

**以人为中心的合成数据集：** 文献中提出了各种合成人体数据集 [20, 22, 49, 67]。然而，与人工人脸数据集相比，这类数据集的发展远不充分 [23, 87]。一般而言，

[1]https://www.wikiart.org/

| 数据集 | 图像 | 实例 | 关键点数量 | Bbox | Pose | 自我-联系 | 自然场景 | 人工场景 |
|---|---|---|---|---|---|---|---|---|
| VOC2012[1] [11] | 8,174 | 17,132 | - | √ | - | - | √ | - |
| MSCOCO[1] [32] | 66,808 | 273,469 | - | √ | - | - | √ | - |
| BodyHands [46] | 20,490 | 63,095 | - | √ | - | - | √ | - |
| People-Art [69] | 1,490 | 3,870 | - | √ | - | - | √ | √ |
| MSCOCO[2] [32] | 58,945 | 156,165 | 17 | √ | √ | - | √ | - |
| MPII [3] | 24,920 | 40,522 | 16 | √ | √ | - | √ | - |
| AI Challenger [71] | 240,000 | 448,776 | 14 | √ | √ | - | √ | - |
| CrowdPose [27] | 20,000 | ~80,000 | 14 | √ | √ | - | √ | - |
| OCHuman [85] | 4,731 | 8,110 | 17 | √ | √ | - | √ | - |
| PoseTrack[3] [2] | 66,374 | 153,615 | 15 | √ | √ | - | √ | - |
| HiEve[3] [33] | 49,820 | 1,099,357 | 14 | √ | √ | - | √ | - |
| ClassArch [40] | 1513 | 1728 | 17 | √ | √ | - | - | √ |
| Sketch2Pose [4] | 808 | 14,772 | 18 | √ | √ | √ | - | √ |
| ***Human-Art* (我们的)** | 50,000 | 123,131 | 21 | √ | √ | √ | √ | √ |

[1] 仅统计包含用于检测的人体边界框标注的图像；[2] 仅统计包含用于人体姿态估计的人体关键点标注的图像。[3] 基于视频的数据集。

表 1. 以人为中心的识别数据集比较，包括人体检测和姿态估计任务。

这些数据集存在背景与人体之间的互动不自然、角色多样性不足等问题。例如， [22] 中的动漫/漫画角色数据集仅有 2631 个不同角色，尽管其图像数量超过一百万。

总而言之，对于以人为中心的计算机视觉任务而言，弥合自然与人工场景之间的差距至关重要。这促成了*Human-Art*的诞生，这是一个场景丰富的以人为中心的数据集，包含充足的高质量图像和多样化的标注。

## 3. *Human-Art* 数据集

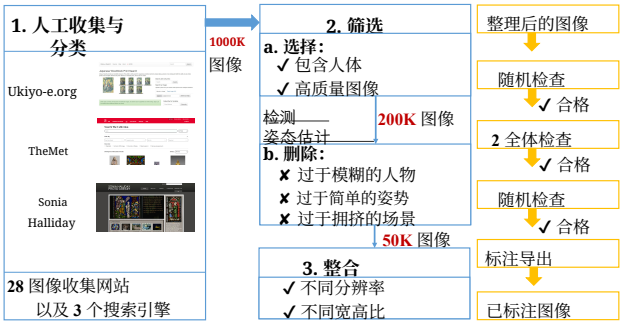
### 3.1. 层次类别分类

作为对自然世界的抽象，艺术是人们感知世界方式的隐喻性表达。

这使艺术成为探索同时包含自然场景和人工场景的场景的良好切入点。

根据 [61], 艺术表现形式可分为八类：文学、音乐、建筑、绘画、雕塑、戏剧、舞蹈和电影。其中，绘画、雕塑、戏剧、舞蹈和电影可以以图像的形式表现。为增加**类别间的可变性**，如图1所示，我们进一步将这些类别划分为二十个经常出现人体的类别：

- 5 类自然人体场景：杂技、角色扮演、舞蹈、戏剧和电影；
- 3 类三维人工 人物场景：车库套件、浮雕、雕塑；
- 12 种二维人工人类场景：儿童绘画、壁画、油画、皮影戏、素描、彩绘玻璃




（a）数据收集流程 (b) 标注

图2。数据收集和标注流程。

浮世绘、卡通、数字艺术、水墨画、水彩画和生成图像；

与以往直接借用某些网站分类标准的艺术相关模型相比， [12,21,29] 例如*WikiArt* 而未经审查，我们的分类标准更适用于以人为中心的场景。一方面，它使我们能够轻松收集更多高质量的人物相关图像。另一方面，*Human-Art* 类别间的差异显著，从而能够获得多样化图像，且分类错误可忽略不计。

### 3.2. 数据收集

如图2所示，我们设计了一套用于高质量图像收集的标准化流程，包括人工图像收集与分类、筛选和整合。

- **人工收集与分类：** Human-Art 中的所有图像均来自 27 个图 -

像网站，这些网站提供高分辨率图像；部分图像由我们从线下展览中自行收集，或由热门模型（例如 Stable Diffusion [55]）生成，以确保高质量。我们仔细检查了大量可能的图像来源，最终确定使用包括 ukiyo-e[2] 和 Sonia Halliday Photo Library[3]在内的 27 个高质量图像网站。为确保**类别内多样性**，*Human-Art* 中的每个类别均来自多个网站。对于与我们的分类并不完全匹配的基于搜索的图像网站，我们会使用多个关键词进行搜索。我们还加入了通过 Google 和 Bing 搜索获得的图像，以进一步增加多样性。最后，我们通过上述流程收集了约一百万张分类完善的图像。

- **筛选：**尽管做出了上述努力，但从互联网爬取的图像质量仍无法得到完全保证，而且其中很大一部分图像不包含人体。为解决上述问题，我们对收集到的图像进行了两轮人工筛选，每张图像至少由两人检查，最终获得约 $200k$张包含人体的高质量图像。接下来，我们对这些图像应用了多种人体检测和姿态估计算法。此举旨在识别并移除包含过于简单、无法提升模型性能的角色，以及模糊人物和难以标注的拥挤人物的图像。经过这一筛选步骤，最终得到 $50k$ 张图像。

- **整合：**我们进一步将整个数据集整理为多个分辨率：$512\times512$、$256\times192$、$32\times32$以及原始分辨率，以便用于各种下游任务。

最后，我们将 *Human-Art* 数据集划分为训练集、验证集和测试集，比例分别为 70%、10% 和 20%，每组分别包含 $35k$、$5k$和 $10k$ 张图像。

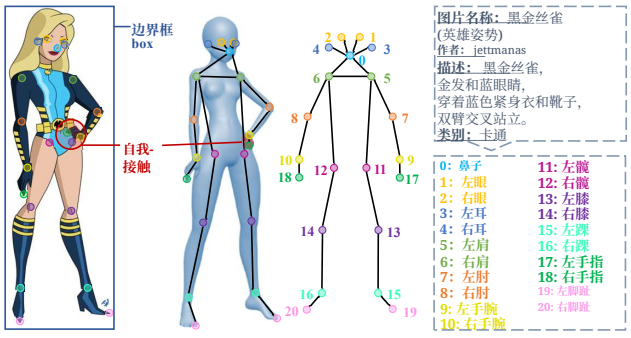



图 3. 所提供标注的示意图，包括二维关键点、边界框、自接触点和文本描述。

[2]https://ukiyo-e.org/[3] http://www.soniahalliday.com/index.html

### 3.3. 数据标注

如图 3所示，*Human-Art* 提供了丰富的标注[4]，，包括人体边界框、21 个人体关键点（带有相应的可见/包含/不可见属性）、自接触关键点和文本信息。

我们遵循 MSCOCO [32] 来定义前 17 个关键点，并增加 4 个额外关键点，即左/右手指和左/右脚趾。这些关键点通过提供更全面的约束，有助于三维姿态估计和形状恢复 [50, 82]。自接触关键点 [4, 45] 同样有助于三维姿态和形状估计，因为它们能够消除二维人体姿态表示中身体部位深度未知所造成的歧义，避免自碰撞和穿透，并且比序数深度更易于使用 [51, 59, 88]。这些标注对于艺术作品中的人体尤为重要，因为艺术夸张会导致更严重的姿态失真和不精确的体型。自接触关键点被标注为身体接触面的中心。尽管艺术作品中存在诸如身体比例/体型失真以及人体不完整或模糊等歧义，人工标注员仍能够凭借直观知识推断出相应位置。文本描述会从图像网站自动抓取，这些网站通常包含相对准确的文本描述。对于不包含相应文本描述的图像，我们使用 BLIP-2 [26]生成伪标签。

标注由专业团队按照标准化的标注与审核流程完成。其中包括35 名数据标注员和12 名数据审核员。标注团队在开始标注前会接受系统培训，以确保标注质量并及时反馈。如图2 (b)所示，整个标注过程经过两轮全面质量检查和两轮随机质量检查，以确保准确率至少达到98%。

### 3.4. 数据集统计与分析

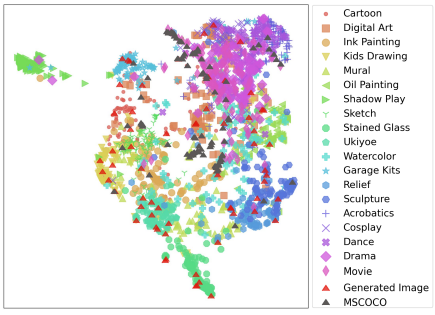



图4. 使用一种流行的聚类方法——均匀流形近似与投影（UMAP），展示*Human-Art* 中20个类别的数据分布以及MSCOCO中的数据分布 [41]。

[4]我们未对生成图像类别进行标注，因为该类别中的许多图像不具备符合规范的人体部位，详见第4.2节。

| 检测器 | | | Faster R-CNN | | | | YOLOX | | Deformable DETR | | DINO | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 设置 | | | val | val * | test | test * | val | test | val | test | val | test |
| MSCOCO [32] | | | **52.2** | 51.6 | - | - | **61.9** | - | **57.2** | - | **63.2** | - |
| Artificial Scene | 2D Representation | 卡通 | 8.8 | 37.9 | 7.0 | 33.5 | 10.8 | 9.2 | 7.9 | 6.7 | 8.7 | 8.1 |
| | | 数字艺术 | 18.8 | 46.4 | 17.8 | 44.2 | 24.1 | 22.9 | 17.6 | 15.5 | 18.6 | 18.1 |
| | | 水墨画 | 11.0 | 37.7 | 9.1 | 37.2 | 15.5 | 13.0 | 11.9 | 10.0 | 14.5 | 11.6 |
| | | 儿童绘画 | 6.6 | 54.2 | 8.0 | 53.6 | 6.8 | 11.5 | 5.6 | 7.2 | 6.8 | 8.2 |
| | | 壁画 | 9.7 | 35.5 | 9.3 | 34.5 | 12.2 | 12.2 | 9.3 | 8.1 | 10.2 | 9.5 |
| | | 油画 | 15.9 | 41.1 | 13.7 | 37.5 | 20.8 | 18.3 | 17.1 | 14.2 | 17.0 | 15.0 |
| | | 皮影戏 | 7.5 | **64.1** | 8.2 | **63.7** | 5.4 | 7.5 | 5.3 | 5.1 | 6.4 | 7.9 |
| | | 素描 | 2.6 | 48.8 | 2.4 | 55.7 | 4.6 | 5.2 | 5.8 | 9.2 | 3.6 | 7.1 |
| | | 彩绘玻璃 | 8.8 | 35.0 | 8.1 | 34.7 | 8.2 | 7.8 | 5.1 | 4.6 | 7.8 | 7.8 |
| | | 浮世绘 | 12.7 | 51.9 | 12.7 | 50.3 | 13.1 | 12.8 | 8.5 | 8.4 | 11.4 | 11.4 |
| | | 水彩画 | 14.8 | 42.8 | 14.2 | 42.2 | 19.7 | 18.2 | 15.6 | 13.6 | 15.4 | 14.3 |
| | 3D Representation | 车库套件 | 22.9 | 60.0 | 22.5 | 62.5 | 22.3 | 19.9 | 17.9 | 14.6 | 22.8 | 19.5 |
| | | 浮雕 | 4.9 | 37.5 | 4.7 | 33.4 | 8.4 | 9.1 | 4.7 | 5.7 | 4.4 | 5.9 |
| | | 雕塑 | 17.7 | 48.6 | 14.4 | 47.0 | 15.8 | 13.2 | 9.4 | 7.1 | 10.1 | 8.5 |
| Natural Scene | | 杂技 | 17.0 | 49.7 | 17.0 | 53.4 | 20.0 | 19.4 | 17.3 | 17.6 | 19.4 | 18.9 |
| | | 角色扮演 | 31.2 | 52.8 | **31.3** | 56.7 | 38.0 | **37.2** | 34.6 | **34.5** | 37.2 | **36.7** |
| | | 舞蹈 | 17.0 | 46.6 | 18.4 | 49.3 | 20.3 | 21.1 | 17.8 | 18.5 | 19.3 | 19.6 |
| | | 戏剧 | 24.3 | 46.0 | 24.8 | 48.7 | 27.4 | 27.5 | 15.4 | 25.8 | 27.8 | 16.7 |
| | | 电影 | 26.3 | 36.5 | 25.0 | 37.2 | 28.0 | 26.8 | 26.6 | 26.2 | 27.2 | 26.3 |
| 平均值 | | | 12.0 | 44.2 | 12.5 | 43.0 | 14.4 | 14.7 | 11.7 | 11.7 | 12.6 | 12.7 |

* 我们提供的基线结果，是通过在 MSCOCO [32] 和*Human-Art*上联合训练得到的。

表 2. 广泛使用的目标检测模型的平均精度（AP）对比，包括 Faster R-CNN [54], YOLOX [53], Deformable DETR [90], 以及最新的 SOTA 模型 DINO [83]。最佳结果以**粗体**显示，最差结果则以带下划线的字体标出。所有模型均在 MSCOCO [32]上训练。各模型的详细设置见补充文件。

为了展示*Human-Art*的类别间差异和类别内多样性，我们从每个类别中随机选取100张图像，并从MSCOCO数据集中选取100 张图像。随后，我们使用ResNet152 [14] 提取图像特征，并使用UMAP [41] 将这些特征降至2 维以进行可视化。如图4所示，大多数场景都形成了各自的聚类，而不同类型的图像（例如戏剧和雕塑）则呈现出差异显著、彼此相距甚远的分布。MSCOCO的分布与*Human-Art* 中的自然图像类别相似，并包含在20 个类别的分布中。此外，生成图像也如预期般散布在整个分布中。

此外，我们统计了*Human-Art*中的不可见关键点，结果表明其有效关键点比例高于 MSCOCO数据集。更多数据集统计与分析详见补充文件。

# 4. 实验

我们在*Human-Art* 上针对多个热门的以人为中心的任务开展了全面实验。在每个子部分中，我们首先介绍任务及相关方法。随后，我们呈现相应的结果并进行分析。

## 4.1. 以人为中心的识别

### 4.1.1 人体检测

人体检测任务 [32, 47] 用于识别给定图像中每个人的边界框，这是进一步进行人体场景理解的基础。它也是诸如自顶向下的人体姿态估计等下游任务的关键步骤 [73, 75, 81]。大多数目标检测器（例如 YOLO [53],DETR [7], 和 DINO [83]）在检测时不会将人体与其他物体区分开来。最近，HBA [47] 专门用于人体和手部检测。

表。2 展示了基于 CNN 和基于 Transformer 的常用检测器在 *Human-Art* 验证集和测试集上的表现。所有预训练模型在人工场景中的表现都较差，平均精度（AP）介于 11.7% 至 14.7% 之间，证实了域差异对模型泛化能力的影响。对于一些分布与 MSCOCO 数据集相似的自然场景 [32]，（例如舞蹈和杂技），现有模型能够取得令人满意的表现。尤其是杂技中的舞台背景通常较为干净，因此其平均精度高于其他类别。同时，尽管皮影戏也有

| 检测器 | | Faster R-CNN +HRNet | | | | | | YOLOX + ViTPose | | | | HigherHRNet | | ED-Pose | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 设置 | | val | val‡ | 验证集 *‡ | test | 测试集 *‡ | | val | 验证集 ‡ | test | 测试集 ‡ | val | test | val | 验证集 * | test | 测试集 * |
| MSCOCO [32] | | **75.6** | | 77.2 | **73.4** | - | - | 79.8 | 82.3 | 81.1 | - | 68.6 | 70.5 | 71.6 | | 69.8 | - |
| Artificial Scene | 2D Representation | 卡通 | 9.7 | 37.6 | 64.7 | 7.3 | 34.4 | 61.0 | 16.3 | 55.1 | 13.0 | 50.5 | 15.7 | 12.0 | 22.2 | 60.4 | 18.0 | 57.1 |
| | | 数字艺术 | 22.6 | 59.6 | 74.4 | 25.8 | 61.2 | 75.7 | 29.0 | 69.9 | 31.9 | 72.2 | 42.5 | 44.3 | 43.5 | 71.4 | 45.6 | 75.1 |
| | | 水墨画 | 6.3 | 51.4 | 72.1 | 5.6 | 48.0 | 72.4 | 8.9 | 59.8 | 9.2 | 58.2 | 26.8 | 20.9 | 28.2 | 56.8 | 24.9 | 55.4 |
| | | 儿童绘画 | 10.5 | 40.8 | 86.1 | 10.0 | 44.6 | 85.9 | 14.0 | 59.2 | 13.2 | 62.6 | 12.6 | 13.8 | 20.7 | 76.7 | 23.2 | 78.8 |
| | | 壁画 | 11.6 | 54.0 | 71.1 | 12.3 | 53.6 | 71.6 | 15.9 | 50.1 | 16.4 | 51.6 | 30.6 | 32.0 | 34.6 | 64.7 | 35.1 | 65.4 |
| | | 油画 | 31.6 | 65.7 | 78.1 | 28.5 | 62.2 | 75.6 | 39.6 | 73.4 | 36.7 | 70.5 | 54.4 | 51.1 | 56.2 | 75.2 | 51.7 | 71.4 |
| | | 皮影戏 | 5.4 | 15.9 | 59.8 | 5.0 | 17.2 | 58.4 | 8.1 | 29.2 | 8.4 | 26.0 | 4.4 | 6.5 | 6.0 | 38.5 | 7.7 | 39.7 |
| | | 素描 | 6.3 | 44.1 | 73.1 | 6.7 | 57.2 | 79.4 | 9.1 | 61.3 | 10.9 | 71.3 | 13.6 | 6.3 | 12.0 | 66.8 | 12.2 | 75.8 |
| | | 彩绘玻璃 | 10.4 | 46.0 | 74.8 | 9.7 | 45.1 | 73.1 | 12.0 | 59.1 | 12.1 | 58.1 | 26.6 | 23.1 | 27.6 | 74.4 | 25.6 | 71.5 |
| | | 浮世绘 | 17.8 | 48.1 | 82.4 | 18.8 | 47.7 | 81.8 | 23.8 | 61.2 | 26.8 | 63.1 | 20.2 | 19.4 | 25.0 | 83.6 | 25.8 | 83.6 |
| | | 水彩画 | 26.7 | 60.1 | 73.9 | 25.5 | 57.6 | 73.4 | 36.4 | 71.0 | 36.1 | 69.0 | 48.9 | 43.4 | 50.6 | 73.5 | 45.6 | 71.3 |
| | 3D Representation | 车库套件 | 45.2 | 57.5 | 86.7 | 44.5 | 61.4 | **89.2** | 52.5 | 76.2 | 50.6 | 77.0 | 37.4 | 34.7 | 47.9 | 87.7 | 44.1 | 90.1 |
| | | 浮雕 | 10.5 | 57.3 | 78.7 | 7.9 | 53.4 | 76.0 | 16.2 | 70.8 | 14.9 | 67.1 | 32.5 | 29.8 | 28.0 | 70.6 | 27.1 | 67.6 |
| | | 雕塑 | 36.4 | 65.9 | 81.0 | 38.5 | 64.0 | 78.5 | 34.9 | 78.5 | 34.2 | 73.7 | 33.5 | 35.2 | 45.9 | 76.9 | 46.7 | 74.7 |
| | | 杂技 | 45.8 | 68.0 | 85.2 | 46.6 | 68.4 | 83.2 | 69.1 | 86.8 | 66.3 | 83.9 | 58.6 | 57.4 | 41.4 | 80.0 | 44.4 | 78.9 |
| Natural Scene | | 角色扮演 | 71.0 | **81.1** | **87.2** | 72.6 <风格 id='1'>81.9</风格> | | 87.0 | **80.0** | **90.3** | **81.7** | **88.8** | **78.1** | **77.8** | **79.6** | **89.1** | **79.7** | **90.4** |
| | | 舞蹈 | 43.1 | 67.3 | 77.2 | 49.2 | 70.1 | 80.1 | 57.3 | 81.5 | 61.5 | 83.8 | 51.4 | 62.4 | 53.6 | 76.5 | 61.2 | 82.2 |
| | | 戏剧 | 45.3 | 75.1 | 82.0 | 46.7 | 75.8 | 83.1 | 54.2 | 83.9 | 56.9 | 84.8 | 69.6 | 72.2 | 75.0 | 85.9 | 76.0 | 86.1 |
| | | 电影 | 49.5 | 71.5 | 77.2 | 50.4 | 72.2 | 76.2 | 57.6 | 76.8 | 56.5 | 78.6 | 64.9 | 65.8 | 69.2 | 82.2 | 68.2 | 80.4 |
| 平均值 | | | 22.2 | 55.2 | 76.4 | 24.1 | 55.4 | 76.0 | 28.7 | 67.7 | 30.7 | 67.5 | 34.6 | 36.3 | 37.5 | 72.3 | 39.2 | 72.7 |

‡ 使用真实标注边界框的自上而下姿态估计结果；* 通过在 MSCOCO [32] 和 *Human-Art*的联合数据上进行训练所提供的基线结果。

表3。广泛使用的人体姿态估计模型的平均精度（AP）比较，包括自上而下方法 HRNet 和 ViTPose [65, 75], 自下而上方法 HigherHRNet [10], 以及单阶段方法 ED-Pose [77]。最佳结果以**粗体**显示，最差结果以带下划线的字体突出显示。每个模型的详细设置见补充文件。

干净的背景，但由于其纹理与自然场景存在巨大差异，表现仍处于最低水平之列。

此外，我们通过在 MSCOCO [32] 和 *Human-Art*的联合数据上训练 Faster R-CNN，提供了一个基线模型。联合训练使皮影戏中的性能提升了约 56%，并使所有类别的平均性能提升了 31%。然而，基线模型在*Human-Art* 上的性能仍然相对较低，有待未来对这一主题开展进一步研究。

### 4.1.2 二维人体姿态估计

人体姿态估计（HPE）是人体运动分析的另一项基础任务，可分为二维 HPE 和三维 HPE，分别输出二维关键点和三维关键点。经过多年的研究，困难姿势、严重遮挡以及复杂背景仍使这些任务颇具挑战性。现有的二维 HPE 方法可分为三类：自顶向下 [48, 65, 75], 自底向上 [6, 10], 和单阶段 [77]。一般而言，在人体检测准确的前提下，自顶向下方法 [48, 65, 75] 通常比其他方法具有更高的性能。然而，这类方法的计算成本较高。相比之下，自底向上方法 [6, 10] 效率更高，尤其适用于拥挤场景，但其准确率相对较低。为在效率和效果之间取得平衡，得益于新兴的基于 DETR 的模型，研究者提出了单阶段方法（例如 PETR [60], ED-Pose [77]）。我们在表 3 中给出了这些代表性方法的结果。

具体而言，我们展示了验证集和测试集上广泛使用的姿态估计方法以及当前最优姿态估计方法的定量结果：*Human-Art*。自上而下的人体估计高度依赖人体检测的准确率；如图7（a）所示，在给定真实边界框时，性能会得到提升。与人体检测任务不同，姿态复杂度对结果的影响大于图像背景。尽管角色扮演通常包含复杂的图像背景，但简单的姿势能够简化估计过程，并带来较高的估计准确率。由于与自然场景中的人体存在较大的形状和纹理差异，皮影戏的估计准确率仍然较低。图7 （b）展示了一些姿态估计失败案例。

此外，我们通过在 MSCOCO 和 *Human-Art* 的联合数据集上训练 HR-Net，提供了一个基线模型，使整体准确率提升了 21%。值得注意的是，使用 ED-Pose 训练后，MSCOCO 上的结果提高了 0.8，这表明多场景图像有助于真实场景中的特征提取和人体理解。

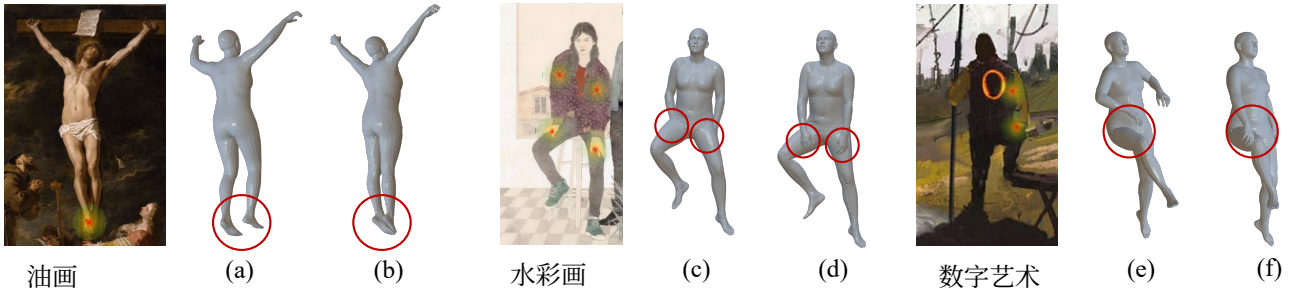
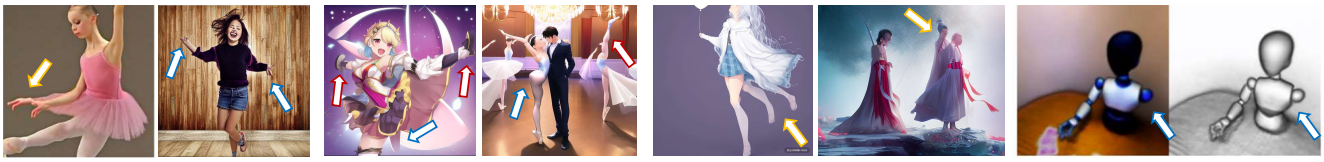



图 5. 示意图说明　　标注的自接触点如何有助于三维人体网格恢复。(a)、(c) 和 (e) 展示了人体网格　　sh
来自三个场景的输出，其中 未进行自接触优化。(b)、(d) 和 (f) 为经过优化的网格结果，并标出了自接触点。




现有热门文生图扩散模型在以人为中心的生成任务中的失败案例。我们用以下颜色标出这些案例：**蓝色**、**黄色**和**红色**箭头，分别表示**缺失的**、**多余的**和**被替换的**身体部位。尽管艺术中没有对错之分，我们仍简单地将自然的人体结构视为更理想的结构。

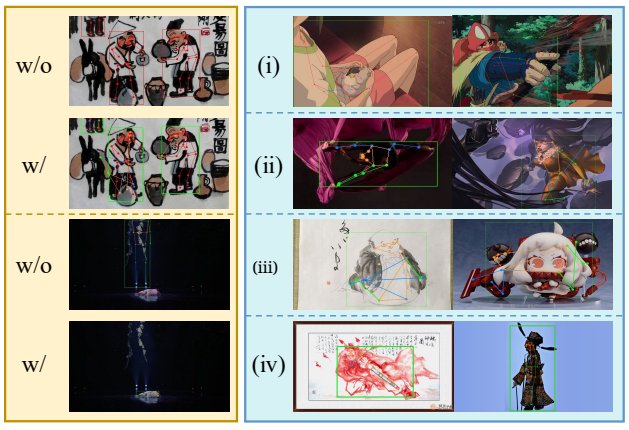



图 7. 姿态估计器 HRNet 在 *Human-Art* 上的失败案例。第一列展示了人体检测对自上而下姿态估计的影响。红色线条和点表示真实标注，其他颜色的线条和点表示检测结果。*w/o* 表示基于检测框的姿态估计结果，*w* 表示使用真实标注边界框得到的结果。右侧图示表明，*Human-Art* 中的 (i) 视角、(ii) 姿态、(iii) 形状和 (iv) 纹理对现有姿态估计器（在 MSCOCO 上训练）构成了挑战。

#### 4.1.3 人体网格恢复

诸如 SMPL 的统计人体模型 [38] 展现出其在动画、游戏和虚拟现实应用中的便捷性。这些模型用少量参数将人体表示为封闭且可制作动画的三维人体网格，大大简化了三维人体网格的表达。

然而，单目相机带来的深度歧义会阻碍三维人体网格估计的保真度。为了解决这一问题，与 Sketc h2Pose 类似 [4], 我们进一步提供自接触标注作为补充信息，通过相互穿透惩罚项促进合理的深度优化。将接触区域映射到由示例微调（EFT）生成的粗略 SMPL 模型顶点上 [19]，然后最小化接触顶点之间的距离，图5中的可视化结果展示了标注自接触关键点如何有助于三维网格恢复。

### 4.2. 图像生成

近年来，图像生成取得了巨大进展，涌现出 unCLIP 等引人注目的视觉与语言生成模型 [52], 潜在扩散 [55]和 Imagen [56]。这得益于生成模型近期的突破，以及 LAION 等大规模网络爬取数据集 [58]。然而，尽管这些模型具备生成逼真且富有创意图像的出色能力，却往往无法忠实地呈现人体结构，如图6所示。

通过提供带有精细标注的人工场景与*Human-Art*，我们不仅为现有数据集提供了有价值的补充。更重要的是，它为生成具备合理人体姿态的艺术图像提供了良好的先验。我们展示了在 *Human-Art* 上微调扩散模型所生成的一些示例，见图8。请注意，*Human-Art* 扩充了艺术类别“皮影戏”，而这一类别在 Stable Diffusion 等先进生成模型中并不存在。




图8. 使用基于 *Human-Art* 训练的扩散生成模型生成的五个场景示例。值得注意的是，皮影戏是现有生成模型所生成的一种新颖场景。

### 4.3. 动作迁移

运动迁移任务旨在通过从目标图像中学习动作，在保留源人物外观的同时，生成源人物的新图像或视频。

以往的动作迁移方法大致可分为两类。基于模型的方法 [8, 39, 57]使用现成的姿态估计器提取姿态信息，再利用姿态骨架驱动角色。相比之下，无模型方法 [62, 63, 66] 能够自动检测与角色无关的隐式关键点轨迹，从而将动作迁移到任意物体上。




图9。基于模型和无模型的动作迁移结果可视化：场景内迁移（第1行）和跨场景动作迁移（第2、3行）。无模型方法 FOMM [63] 因对应关系不稳定而严重失效。相比之下，基于模型的方法 [8] 更适用于多场景动作迁移。使用在 *Human-Art*上预训练的姿态估计模型后，(d)的结果优于(c)，

如图所示。9 (e)，无模型方法 [63] 容易在新场景中失效，因为源图像与驱动图像之间的对应关系不稳定。基于模型的方法在人体运动迁移上表现得更稳定，但高度依赖准确的姿态估计结果。现有的姿态估计器不适用于儿童绘画等人工场景[5]，，因此需要训练针对特定场景的姿态估计模型。为了说明具备多场景适应能力的姿态估计器如何有助于动作迁移任务，我们进行了实验

[5]https://sketch.metademolab.com/

，在广为人知的 EveryBodyDanceNow 上进行实验 [8]，不进行面部增强。尽管有些基于模型的方法生成的结果优于 EveryBodyDanceNow，但我们选择该模型来说明姿态如何影响动作迁移，因为它在文献中得到了广泛认可。图 9 (c) 展示了使用在自然人物场景上训练的姿态估计器 OpenPose 得到的原始动作迁移结果 [5] 。通过使用 *Human-Art* 对姿态估计器进行改进，图 9 (d) 展示了更优的姿态检测模型如何显著改善运动迁移任务。

## 5. 结论与讨论

本文提出了*Human-Art*，这是一个包含 $50k$ 张高质量图像、具有多样化人工标注的丰富场景以人为中心的数据集，可作为服务于多项计算机视觉任务的新型挑战性数据集，例如人体检测、人体姿态估计、人体网格重建、动作迁移和图像生成。在实验中，我们为这些任务提供了全面的基线结果和详细分析。我们希望这项工作能够为相关研究领域带来启发，并开拓新的研究问题。

**局限性与未来工作：** *Human-Art* 中的图像可能被滥用于生成伪造图像，从而带来负面的社会影响。此外，尽管我们已在*Human-Art*上广泛探索了下游任务，但我们仅进行了实验，以揭示现有方法为何以及如何常常在我们的数据集上失败，并未提供更优的解决方案。因此，这些丰富场景下以人为中心的任务仍存在显著的研究空白，亟需未来研究提出新的解决方案。具体而言，基于*Human-Art* 的未来研究方向包括但不限于：1）能够适应不同场景中各种人体姿态、体型、纹理和图像背景的跨域人体识别算法；2）具有合理人体身体结构的可信图像生成，尤其是诸如 GLI-GEN [28]和 ControlNet [84]等可控人体图像生成；3）跨不同场景的包容性运动迁移算法。

我们计划持续扩展*Human-Art* ，以支持新的场景。为促进未来以人为中心的研究，我们将通过易于使用的数据可视化平台公开训练集和验证集。对于测试集，我们将提供测试接口，但不会公开数据，以防止测试信息泄露。

## 6. 致谢

本研究部分得到了深圳市科技创新委员会深港澳科技计划（C类）项目（项目编号：SGDX2020110309500101）以及科研配套资助 CSE-7-2022 的支持。

## 补充材料

由于篇幅限制，主论文未能收录的更多细节和补充结果将在本补充材料中呈现。内容列表如下：

- 第 A 节提供更多数据集统计与分析，包括图像来源、关键点可见/遮挡/不可见属性分布、人体尺寸分布和标注可视化的统计与分析。
- 第 B 节介绍实验细节，包括实验所用模型的实现细节、人体检测和姿态估计结果分析，以及更多关于*Human-Art*的评估结果。
- 第 C 节进一步讨论用于多场景泛化的相关数据集。

## A. 数据集统计与分析

### A.1. 图像来源

*Human-Art* 是一个全面的以人为中心的数据集，包含来自20种不同场景的50,000张图像，每种场景包含2,500张图像。如图10右侧所示，这些图像选自30种不同的图像来源，并且我们确保每种场景的图像来源具有多样性。具体而言，*Human-Art* 包含由欧洲、北美、东亚和南非作者创作的图像，时间跨度从公元前到21世纪，描绘了姿态、体形和纹理各异的人体。

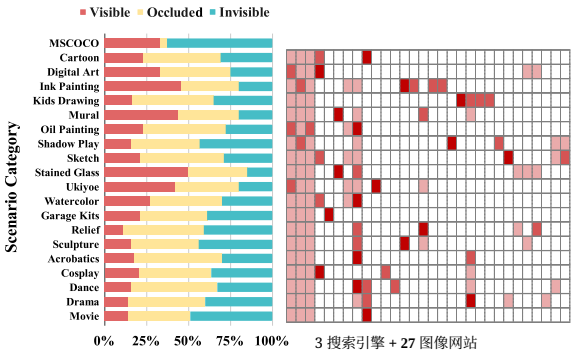



图10. 对比我们的20种场景与MSCOCO的所有关键点可见性统计分析 [32]（左），以及*Human-Art* 图像来源分布图（右）。

### A.2. 关键点属性

*Human-Art* 遵循 MSCOCO [32]，采用可见/遮挡/不可见属性对关键点进行标注。图10 左侧展示了 *Human-Art* 所有已标注场景中可见、遮挡和不可见关键点的总占比，并与 MSCOCO 进行比较。可以观察到，MSCOCO数据集中的不可见关键点占比高达 63.2%，远高于

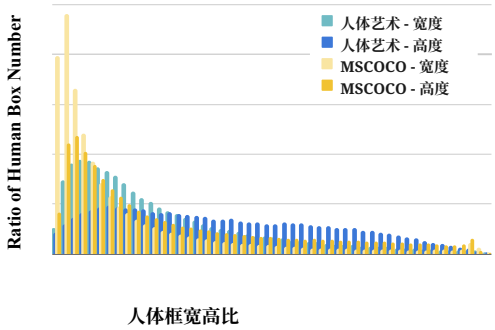



图11. *bounding box width/image width* 和 *bounding box height/image height*的分布。横轴表示人体边界框的高度和宽度与整张图像的高度和宽度之比。纵轴表示具有相应高度宽度比的人体边界框所占的百分比。

*Human-Art*中所有类别的比例。我们将其归因于这样一个事实：尽管 MSCOCO 包含超过 25 万个人体，但它并不只关注以人为中心的场景。这导致其中存在大量小尺度、不完整且模糊的人体，只能使用边界框进行标注。与此同时，*Human-Art* 中许多类别的可见关键点占比略低。这是因为，一方面，艺术性的自然场景通常包含精细的动作和织物遮盖，从而遮挡人体；另一方面，人工场景中的图像可能存在线条不清晰或因历史遗失而缺少身体部位的情况。总体而言，*Human-Art* 的关键点标注占比高于 COCO，这可通过提供更多有效数据而有利于相关任务。

### A.3. 人体尺寸分布

如图 11所示，*Human-Art* 中的人体尺寸分布比 MSCOCO 更均匀 [32]。在*Human-Art* 中，人体平均高度为图像高度的0。40 倍，而在 MSCOCO 中为0。28 倍，这表明由于以人体为中心的图像收集过程，*Human-Art* 中的微小人体少于 MSCOCO。人体平均宽度比在0。15 和0。25 ，分别对应*Human-Art* 和 MSCOCO。由于人体通常修长纤细，人体宽度比更多集中在较小的比例范围内。更加均衡的人体尺寸分布使*Human-Art* 能够支持涉及不同尺寸人体的下游任务。例如，动作迁移通常需要更大、更细致的人体图像，以输出具有更高保真度的角色。然而，图像生成需要生成各种分辨率的人体，以满足用户的需求。更有趣的是，尽管我们的人体尺寸相对较大且分布均衡，但检测和姿态估计的性能仍然较差，这表明我们的数据集除了尺度之外，还具有其他方面的困难，例如外观差异

多样性、背景变化和姿态复杂度。

### A.4. 标注可视化

我们在图 *12*中展示了Human-Art 的标注质量和图像多样性，包括人体边界框和关键点标注。*Human-Art* 的多样性源于不同类别之间绘画技法的广泛变化，以及每个类别内部人体尺寸、体型和人物姿势的差异。这使得*Human-Art* 比以往的真实世界人体数据集更具挑战性，因此对检测和估计模型提出了更高的泛化能力要求。

## B. 实验细节

### B.1. 实现细节

针对人体检测，我们提供了以下基线方法：Faster R-CNN [54], YOLOX [53], Deformable DETR [90], 和 DINO [83]。我们使用的所有预训练模型均仅在 MSCOCO 上训练 [32]。训练则是在 MSCOCO 和 *Human-Art*的随机混合数据上进行的。具体实现细节如下：

- Faster R-CNN：我们选择以特征金字塔网络 [31] 和 ResNet-50 [14]作为骨干网络的 Faster R-CNN。测试时，所用的预训练模型是在 8 块 NVIDIA GTX 1080 Ti GPU 上训练 12 个周期得到的。训练时，我们使用 8 块 NVIDIA RTX 3090Ti GPU 进行训练。由于数据量几乎翻了一番，为确保模型收敛，我们训练了 21 个周期，而非原始的 12 个周期。
- YOLOX：我们选择输入尺寸为 640x640 的 YOLOX-L。测试时使用的预训练模型在 8 张 NVIDIA Tesla PG503-216 GPU 上训练了 300 个轮次。训练时，我们使用 4 张 Nvidia Tesla A100 GPU 进行训练。
- Deformable DETR：我们选择以 ResNet-50 [14] 作为骨干网络的两阶段 Deformable DETR。测试时使用的预训练模型在 8 张 Nvidia Tesla v100 GPU 上训练了 50 个轮次。训练时，我们使用 4 张 Nvidia Tesla A100 GPU 进行训练。
- DINO：我们选择以 Swin-L [37] 作为骨干网络的 DINO-5scale。测试时使用的预训练模型在 Nvidia Tesla A100 GPU 上训练了 31 个轮次。

对于人体姿态估计，我们提供 HRNet [65], ViTPose [75], HigherHRNet [10], 和 ED- Pose [77] 的基线方法。我们使用的所有预训练模型均仅在 MSCOCO [32] 上训练。训练是在 MSCOCO 与 *Human-Art* 的随机打乱数据上进行的。对于自上而下姿态估计方法，我们在测试时使用与上文所列设置相同的人体检测器，并在测试时对真实标注边界框进行增强（例如随机翻转、边界框中心随机偏移）以及训练阶段。为与 COCO 进行公平比较，测试和训练仅考虑 17 个人体关键点。具体实现细节如下：

- HRNet：我们选择输入尺寸为 256x192 的 HRNet-W48。测试时使用的预训练模型在 8 张 Nvidia Tesla v100 GPU 上训练了 210 个轮次。训练时，我们使用 4 张 NvidiaTesla v100 GPU 进行训练。
- ViTPose：我们选择输入尺寸为 256x192 的 ViTPose-H。测试时使用的预训练模型在 8 块 Nvidia Tesla V100 GPU 上训练了 210 个 epoch。
- HigherHRNet：我们选择输入尺寸为 512x512 的 HigherHRNet-W48。测试时使用的预训练模型在 8 块 Nvidia Tesla V100 GPU 上训练了 300 个 epoch。训练时，我们使用 4 块 Nvidia Tesla A100 GPU。
- ED-Pose：我们选择以 ResNet-50 [14] 为骨干网络的 ED-Pose。测试时使用的预训练模型训练了 60 个 epoch。

对于人体网格恢复，我们采用与 Sketch2Pose [4] 相同的优化策略，使用 17 个人体关键点和自接触关键点。Sketch2Pose [4] 中使用的二维到 SMPL 模型包含二维骨骼切线、身体部位接触和骨骼缩短的优化。在论文正文中，图 6 对比了未使用和使用身体部位接触时的可视化结果。结果表明，使用自接触关键点可通过缩短自接触区域附近的三维距离，改善三维网格恢复。需要注意的是，由于另外两项优化因素——骨骼切线和骨骼缩短——的影响，与使用自接触优化时相比，未使用该优化时，与自接触区域没有直接连接的身体部位会呈现不同的姿态（例如论文正文图 6 (a) 和 (b) 中的肘部角度）。

### B.2. 更多以人为中心的任务分析

姿态估计或人体检测模型输出的置信度分数，表示模型对其输出结果的确信程度。AP分数表示模型的平均预测准确率。我们尝试根据这两项指标，分析这些现有模型在我们的数据上表现不佳的原因。图13 展示了人体检测模型 YOLOX [53] 和姿态估计模型 HRNet [65]训练前后置信度分数和 AP 分布的变化。我们使用了与正文相同的训练模型。

我们从以下三个方面进行分析：(1) 置信度分数与AP分布的对比。我们发现，模型往往过度自信，置信度分数分布与AP分布并未呈现正相关。这一问题在*Human-Art*的人工场景中变得更加严重。例如，在图13 (e)和图13 (f)中，尽管姿态估计模型在大多数图像上表现出较高的置信度分数，但很大比例的估计结果

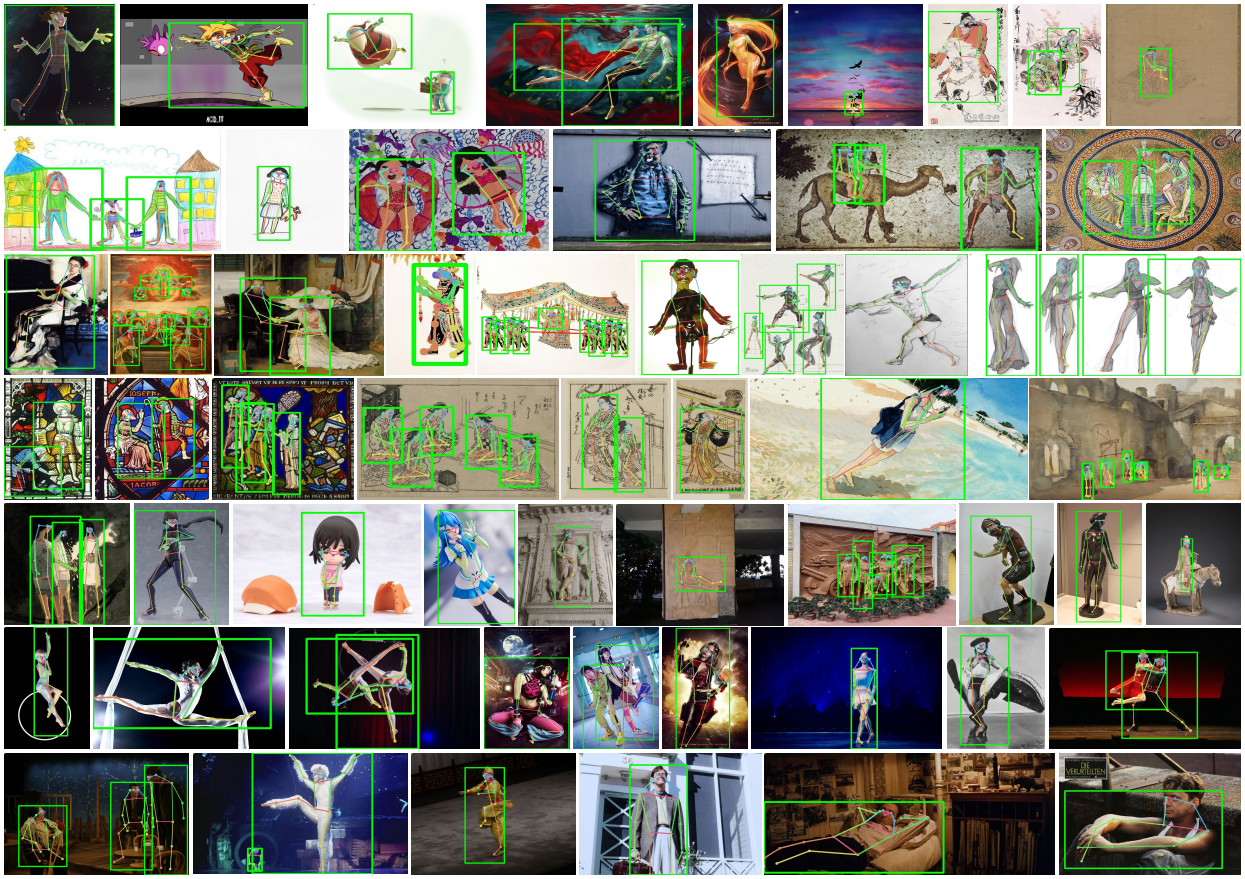


图 12。*Human-Art*中的标注示例。我们从每个类别中随机选取 3 张图像，以展示*Human-Art*的标注质量和图像多样性。*Human-Art* 中的图像在人体形状、姿态、纹理和尺寸方面各不相同。

AP分数范围为0 至0.25。(2) 训练前后的对比。反复出现的发现是，训练可以降低低置信度分数和低AP分数所占的比例，这符合常识。另一个有趣的发现是，尽管在联合训练后，MSCOCO上的平均AP分数 [32] 有所下降，但低分数所占比例也有所降低，如图13(b)和图13(d)所示。这可能是因为*Human-Art* 中分布更加均匀的人体尺寸和更加丰富的描绘，有助于模型在真实场景中的困难姿势上获得更强的适应能力。(3) 人体检测与姿态估计任务的对比。训练后，人体检测的AP分布沿水平方向更加均匀。然而，姿态估计模型的AP分数则集中分布在低分和高分区间。这可能是由于两种方法的目标不同。人体检测失败时，真实标注与检测结果之间仍可能存在有效交互。但由于不同关键点之间的相互作用，姿态估计通常会失败得更加严重。

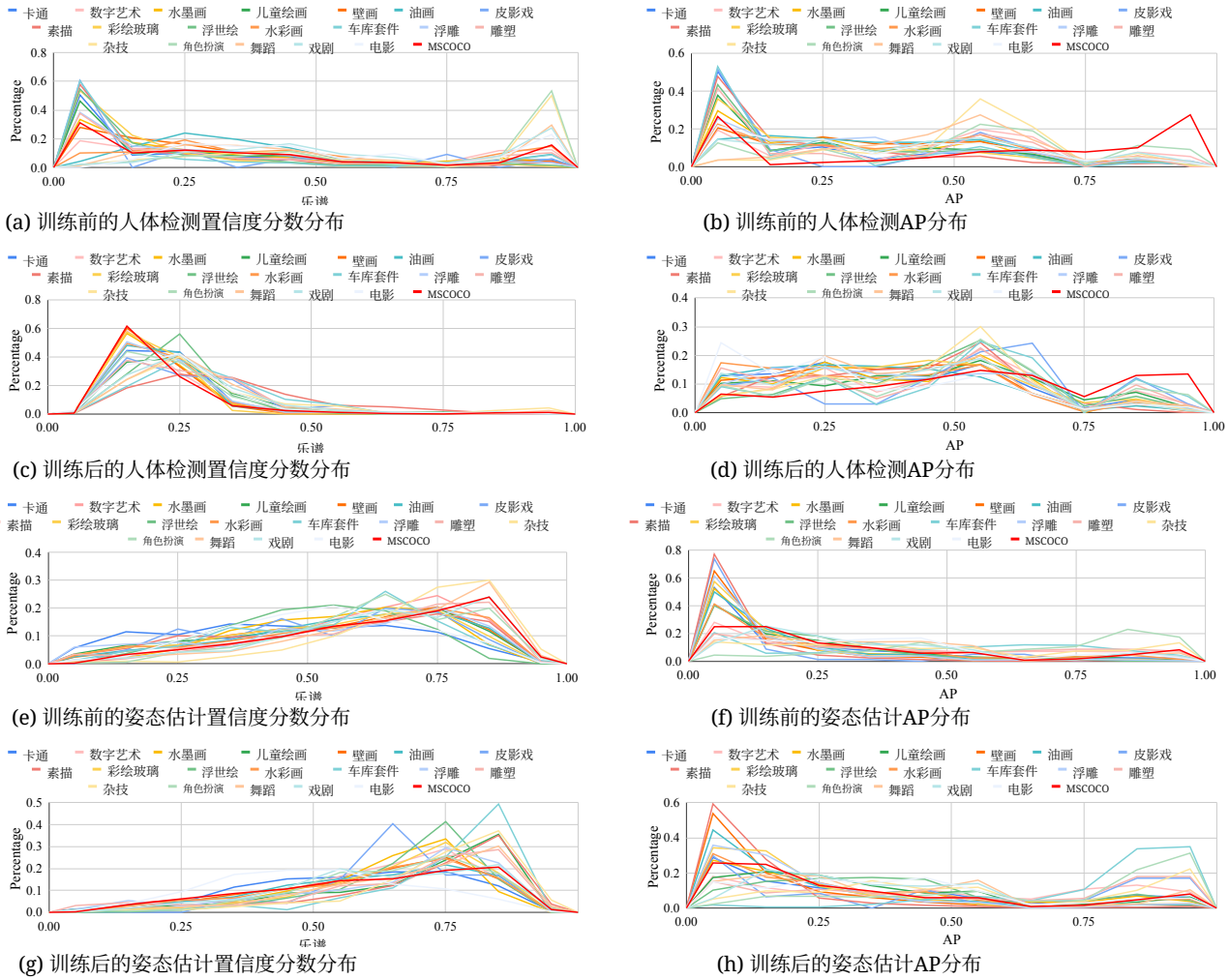
| 测试集 | HA(T) | HA(F) | SK(T) | SK(F) |
|---|---|---|---|---|
| **MSCOCO** | 63.5‡/ 62.4* | 74.5‡/74.8* | 13.5* | 47.6* |
| **Human-Art** | <风格 id="1">65.7 / </风格 ><风格 id="3">63.</风格>6* | 64.4 / 61.9* | 14.3* | 43.3* |
| **Sketch2Pose** | 71.6* | 70.5* | 16.3* | 68.9* |

[1] **训练策略**：**HA** 表示用于*Human-Art* 。**SK** 表示用于 Sketch2Pose。**T** 表示从头训练。**F** 表示从在 MSCOCO 上预训练的 HRNet 进行微调。[2] * / ‡ 表示在不同数据集的 10 / 17 个交集关键点上计算（ *Human-Art* & Sketch2Pose / *Human-Art* & MSCOCO ）。

表 4. 姿态估计模型 HRNet 在 3 个测试集（第一列）上、采用 4 种训练策略（第一行）时的平均精度结果。 10 个交集关键点上的最佳结果以红色标示。

## B.3. 跨数据集结果

受篇幅限制，跨数据集实验评估结果放在*Human-Art* 和 Sketch2Pose 的表4中。MSCOCO、Sketch2Pose 和 *Human-Art* 采用了不同的关键点定义，因此我们给出了这三个数据集在 10 个/17 个交集关键点上的结果（三个数据集共有 10 个交集关键点；MSCOCO 和*Human-Art* 共有 17 个交集关键点）。表中展示了四种训练设置：(1) HA(T) 表示使用*Human-Art* 进行训练；(2) HA(F) 表示使用*Human-Art* 微调在 MSCOCO 上预训练的模型；(3)




(a) 训练前的人体检测置信度分数分布

(b) 训练前的人体检测AP分布

(c) 训练后的人体检测置信度分数分布

(d) 训练后的人体检测AP分布

(e) 训练前的姿态估计置信度分数分布

(f) 训练前的姿态估计AP分布

(g) 训练后的姿态估计置信度分数分布

(h) 训练后的姿态估计AP分布

图13. 人体检测模型 YOLOX 的置信度分数和 AP 分布对比 [53]以及姿态估计模型 HRNet [65] 在我们提出的场景和 COCO 上训练前后的对比。具体而言，(a)-(d) 展示姿态估计的分布，(e)-(h) 展示人体检测的分布。每幅图的横轴表示置信度分数/AP 区间，纵轴表示每个分数/AP 区间中的图像百分比。

SK(T) 表示使用 Sketch2Pose 进行训练；(4) SK(F) 表示使用 Sketch2Pose 微调在 MSCOCO 上预训练的模型。结果表明，Sketch2Pose 不足以用于多场景姿态估计器的训练，因此表现较差。使用*Human-Art* 进行训练和微调都能获得相对令人满意的准确率，其中，使用 *Human-Art* 微调在自然场景中的平均精度最高，而使用 *Human-Art* 训练在*Human-Art* 和 Sketch2Pose 上的结果最佳。考虑到日常生活中自然人体的普遍性，我们仍建议将*Human-Art* 与 MSCOCO 结合用于训练*Human-Art* 。

## C. 用于多场景泛化的数据集

包含多场景的现有数据集 [13, 15, 24, 68–70, 72] 更常用于领域泛化—。它们侧重于目标分类或目标检测任务。相关方法尝试将分类器或检测器从自然图像迁移到人工图像 [74, 80]。然而，正如表 5所示，这些数据集存在若干局限，难以弥合自然与人工场景中以人为中心的任务之间的差距。首先，这些数据集中的图像数量和类别数量不足。其次，由于它们仅包含边界框或目标类别标注，所能支持的下游任务数量受到限制。第三，这些数据集仅有一小部分场景包含人体，因此不适用于以人为中心的任务。此外，BAM! [15]是一个包含 7 个人工类别、面向图像分类的大规模数据集，但它使用不可靠的模型分类器为图像标注，而非进行人工标注，这可能导致大量标注错误。

| Task | 数据集 | 图像 | 自然场景 | 人工场景 |
|---|---|---|---|---|
| Object Classification / Detection | Inoue N. 等人 [15] | 5,000 | 1 | 3 |
| | Office-Home [68] | 15,500 | 2 | 2 |
| | PACS [24] | 9,991 | 1 | 3 |
| | People-Art [69] | 1,490 | 1 | 42 * |
| | Photo-Art [72] | 5,375 | 1 | 1 |
| | ***Human-Art*（本文）** | **50,000** | **4** | **16** |

* People-Art 的 42 种绘画风格 [69] 采用了不同于 *Human-Art* 的分类标准，这些风格被归入所提出的 *Human-Art* 的 20 个类别中。

表 5. 用于通用目标分类和检测任务的多场景数据集比较。

相比之下，*Human-Art* 是一个涵盖全场景的以人为中心的数据集，包含以往多场景数据集所涉及的域泛化任务（例如人体检测的域泛化），以及人体姿态估计、图像生成和图像风格迁移等其他任务。

以往方法通过将知识从源域迁移到目标域，解决目标检测中的域差异问题。 [69] 在 People-Art 上微调 Faster R-CNN，以检测艺术作品中的人体。 H2FA R-CNN [74]提出了一种整体与层次特征对齐 R-CNN，以实现目标检测的图像级对齐。 [15]利用图像级域迁移和来自源域的伪标签来训练目标检测器 SSD300 [35]。

以往研究 [17, 89] 探索了自然场景下人体关键点检测的域泛化与适应。然而，据我们所知，此前尚无研究涉及自然场景和人工场景中的多场景人体关键点检测。

总之，文献中没有适用于*Human-Art* 的域适应和域泛化方法可以直接应用于该数据集，这将留待未来工作解决。

## 参考文献


[1] Kfir Aberman、Peizhuo Li、Dani Lischinski、Olga Sorkine-Hornung、Daniel Cohen-Or 和 Baoquan Chen。用于深度运动重定向的骨架感知网络。ACM图形学汇刊（TOG），39(4):62–1，2020。1[2]Mykhaylo Andriluka、Umar Iqbal、Eldar Insafutdinov、Leonid Pishchulin、Anton Milan、Juergen Gall 和 Bernt Schiele。PoseTrack：人体姿态估计和跟踪的基准。在 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第5167–5176页，2018年。2、3[3] Mykhaylo Andriluka、Leonid Pishchulin、Peter Gehler 和 Bernt Schiele。二维人体姿态估计：新基准与最新进展分析。在 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第3686–3693页，2014年。2、3[4] Kirill Brodt 和 Mikhail Bessmeltsev。Sketch2Pose：从位图草图估计三维角色姿态。ACM图形学汇刊（TOG），41(4):1–15，2022年。1、2、3、4、7、10[5] Z. Cao、G. Hidalgo Martinez、T. Simon、S. Wei 和 Y. A. Sheikh。OpenPose：使用部件亲和场进行实时多人二维姿态估计。IEEE模式分析与机器智能汇刊（PAMI），2019年。8[6] Zhe Cao、Tomas Simon、Shih-En Wei 和 Yaser Sheikh。使用部件亲和场进行实时多人二维姿态估计。在 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第7291–7299页，2017年。6[7] Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko。使用 Transformer 的端到端目标检测。在欧洲计算机视觉会议（ECCV），第213–229页。Springer，2020年。5[8] Caroline Chan、Shiry Ginosar、Tinghui Zhou 和 Alexei A Efros。Everybody Dance Now。在 IEEE/CVF国际计算机视觉会议（ICCV）论文集，第5933–5942页，2019年。8。[9] Shuhong Chen 和 Matthias Zwicker。提升二维动画插值的感知质量。在欧洲计算机视觉会议（ECCV），2022年。1[10] Bowen Cheng、Bin Xiao、Jingdong Wang、Honghui Shi、Thomas S Huang 和 Lei Zhang。HigherHRNet：面向自底向上人体姿态估计的尺度感知表示学习。在 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第5386–5395页，2020年。6、10[11] Mark Everingham、Luc Van Gool、Christopher KI Williams、John Winn 和 Andrew Zisserman。PASCAL 视觉目标类别（VOC）挑战赛。国际计算机视觉杂志（IJCV），88(2):303–338，2010年。2、3[12] Noa Garcia 和 George Vogiatzis。如何阅读绘画：借助多模态检索进行艺术语义理解。在欧洲计算机视觉会议（ECCV）研讨会论文集，第0–0页，2018年。3[13] Nicolas Gonthier、Yann Gousseau、Said Ladjal 和 Olivier Bonfait。艺术作品中的弱监督目标检测。在欧洲计算机视觉会议（ECCV）研讨会论文集，第0–0页，2018年。 12[14] Kaiming He、张祥雨、Shaoqing Ren和Jian Sun。用于图像识别的深度残差学习。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第770–778页，2016年。 5, 10[15] 井上直人、古田亮介、山崎俊彦和Kiyoharu Aizawa。通过渐进式域适应进行跨域弱监督目标检测。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第5001–5009页，2018年。 12, 13[16] Catalin Ionescu、Dragos Papava、Vlad Olaru和Cristian Sminchisescu。Human3.6M：用于自然环境中三维人体感知的大规模数据集与预测方法。IEEE模式分析与机器智能汇刊，36(7)：1325–1339，2014年7月。 1[17] 蒋军广、纪一飞、王希美、刘宇峰、王建民和龙明盛。用于无监督关键点检测的回归式域适应。载于IEEE/CVF计算机视觉与模式识别会议论文集，第6780–6789页，2021年。 13[18] 金胜、徐璐敏、徐进、Can Wang、刘文韬、Chen Qian、欧阳万里和罗平。野外环境中的全身人体姿态估计。载于欧洲计算机视觉会议（ECCV），2020年。 2[19] Hanbyul Joo、Natalia Neverova和Andrea Vedaldi。面向野外三维人体姿态估计的三维人体模型拟合示例微调。载于国际三维视觉会议（3DV），第42–52页。IEEE，2021年。 7[20] David Kadish、Sebastian Risi和Anders Sundnes Løvlie。仅使用风格迁移改进艺术图像中的目标检测。载于国际神经网络联合会议（IJCNN），第1–8页。IEEE，2021年。 2[21] Sergey Karayev、Matthew Trentacoste、Helen Han、Aseem Agarwala、Trevor Darrell、Aaron Hertzmann和Holger Winnemoeller。识别图像风格。arXiv预印本，arXiv:1311.3715，2013年。 3[22] Pramook Khungurn和Derek Chou。动漫/漫画角色的姿态估计：合成数据的一个案例。载于第1届漫画分析、处理与理解国际研讨会论文集，第1–6页，2016年。 2, 3[23] Kangyeol Kim、Sunghyun Park、李宰成、郑成孝、李俊秀和朱在杰。AnimeCeleb：用于头部重演的大规模动画名人头部数据集。载于欧洲计算机视觉会议（ECCV），2022年。 2[24] 李达、Yongxin Yang、宋一哲和Timothy M Hospedales。更深、更广、更具艺术性的域泛化。载于IEEE/CVF国际计算机视觉会议（ICCV），第5542–5550页，2017年。 12, 13[25] Jiefeng Li、卞思远、Ailing Zeng、Can Wang、庞博、刘文韬和Ce wu Lu。基于残差对数似然估计的人体姿态回归。载于IEEE/CVF国际计算机视觉会议（ICCV），第11025–11034页，2021年。 2[26] 李俊楠、李东旭、Silvio Savarese和Steven Hoi。BLIP-2：利用冻结图像编码器和大语言模型进行自举式语言-图像预训练。

arXiv预印本，arXiv:2301.12597，2023年。 4[27] Jiefeng Li、Can Wang、Hao Zhu、Yihuan Mao、Hao-Shu Fang和Cewu Lu。CrowdPose：高效的拥挤场景姿态估计与新基准。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第10863–10872页，2019年。 1, 2, 3[28] Yuheng Li、Haotian Liu、Qingyang Wu、Fangzhou Mu、Jian-wei Yang、Jianfeng Gao、Chunyuan Li和Yong Jae Lee。GLIGEN：开放集接地文本到图像生成。arXiv预印本，arXiv: *2301 07093*, 2023 8 。。Peiyuan Liao、[29] Xiuyu Li、Xihui Liu和Kurt Keutzer。ArtBench数据集：使用艺术作品对生成模型进行基准评测。2022年。 3[30] Mei Kuan Lim、Ven Jyn Kok、Chen Change Loy和Chee Seng Chan。通过全局相似性结构进行人群显著性检测。载于国际模式识别会议（ICPR），第3957–3962页。IEEE，2014年。 2[31] Tsung-Yi Lin、Piotr Dollár、Ross Girshick、Kaiming He、Bharath Hariharan和Serge Belongie。用于目标检测的特征金字塔网络。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第2117–2125页，2017年。 10[32] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár和C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。载于欧洲计算机视觉会议（ECCV），第740–755页。Springer，2014年。 1, 2, 3, 4, 5, 6, 9, 10, 11[33] Weiyao Lin、Huabin Liu、Shizhan Liu、Yuxi Li、Rui Qian、Tao Wang、Ning Xu、Hongkai Xiong、Guo-Jun Qi和Nicu Sebe。Human in events：复杂事件中以人为中心的视频分析大规模基准。arXiv预印本，arXiv:2005.04490，2020年。 2, 3[34] Zuzeng Lin、Ailin Huang、Zhewei Huang、Chen Hu和Shuchang Zhou。使用动漫角色设定表的协作式神经渲染。arXiv预印本，arXiv:2207.05378，2022年。 1[35] Wei Liu、Dragomir Anguelov、Dumitru Erhan、Christian Szegedy、Scott Reed、Cheng-Yang Fu和Alexander C Berg。SSD：单次多框检测器。载于欧洲计算机视觉会议，第 21 37 2016 13 –页。[36] Springer，。Wu Liu和Tao Mei。单目二维和三维人体姿态估计的最新进展：深度学习视角。ACM计算综述（CSUR），2022年3月。 1[37] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin和Baining Guo。Swin Transformer：使用移位窗口的分层视觉Transformer。载于IEEE/CVF国际计算机视觉会议论文集，第10012–10022页，2021年。 10[38] Matthew Loper、Naureen Mahmood、Javier Romero、Gerard Pons-Moll和Michael J. Black。SMPL：蒙皮多人线性模型。ACM图形学汇刊（TOG），34(6)：248:1–248:16，2015年10月。 7[39] Liqian Ma、Xu Jia、Qianru Sun、Bernt Schiele、Tinne Tuyte- laars和Luc Van Gool。姿态引导的人物图像生成。神经信息处理系统进展（NIPS），30，2017年。 8[40]Prathmesh Madhu、Angel Villar-Corrales、Ronak Kosti、Torsten Bendschus、Corinna Reinhardt、Peter Bell、Andreas Maier 和 Vincent Christlein。通过基于感知的风格迁移学习提升古代花瓶画中的人体姿态估计。《计算与文化遗产期刊》（JOCCH），2022年11月。1, 2, 3[41] Leland McInnes、John Healy 和 James Melville。UMAP：用于降维的均匀流形近似与投影。arXiv 预印本 arXiv:1802.03426，2018年。4, 5[42] Yifang Men、Yiming Mao、Yuning Jiang、Wei-Ying Ma 和 Zhouhui Lian。利用属性分解生成对抗网络实现可控的人物图像合成。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第5084–5093页，2020年。1[43] Gyeongsik Moon、Hongsuk Choi 和 Kyoung Mu Lee。用于全身三维人体网格估计的精准三维手部姿态估计。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第2308–2317页，2022年。2[44] Lucas Mourot、Ludovic Hoyet、François Le Clerc、François Schnitzler 和 Pierre Hellier。基于骨架的人体动画深度学习综述。载于《计算机图形学论坛》，第41卷，第122–157页。Wiley在线图书馆，2022年。1[45] Lea Muller、Ahmed A. A. Osman、Siyu Tang、Chun-Hao P. Huang 和 Michael J. Black。关于自接触与人体姿态。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第9990–9999页，2021年6月。4[46] Supreeth Narasimhaswamy、Thanh Nguyen、Mingzhen Huang 和 Minh Hoai。这些手是谁的？野外环境中的手部检测与手部和身体关联。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第4889–4899页，2022年。1, 2, 3[47] Supreeth Narasimhaswamy、Thanh Nguyen、Mingzhen Huang 和 Minh Hoai。这些手是谁的？野外环境中的手部检测与手部和身体关联。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2022年。5[48] Alejandro Newell、Kaiyu Yang 和 Jia Deng。用于人体姿态估计的堆叠沙漏网络。载于欧洲计算机视觉会议（ECCV），第483–499页。Springer，2016年。6[49] Priyanka Patel、Chun-Hao P. Huang、Joachim Tesch、David T. Hoffmann、Shashank Tripathi 和 Michael J. Black。AGORA：针对回归分析优化的地理空间虚拟化身。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2021年6月。2[50] Georgios Pavlakos、Vasileios Choutas、Nima Ghorbani、Timo Bolkart、Ahmed A. Osman、Dimitrios Tzionas 和 Michael J. Black。富有表现力的人体捕捉：从单张图像恢复三维手部、面部和身体。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2019年。4[51] Georgios Pavlakos、Xiaowei Zhou 和 Kostas Daniilidis。用于三维人体姿态估计的序数深度监督。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集

视觉与模式识别（CVPR），第7307–7316页，2018年。4 – 。[52] Aditya Ramesh 、 Prafulla Dhariwal 、 Alex Nichol 、 Casey Chu 和 Mark Chen。利用 CLIP 潜变量进行分层文本条件图像生成。arXiv 预印本 arXiv:2204 06125，2022年。7 。 。[53] Joseph Redmon、Santosh Divvala、Ross Girshick 和 Ali Farhadi。YOLO（You Only Look Once）：统一的实时目标检测。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第779–788页，2016年。5, 10, 12[54] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。Faster R-CNN：利用区域提议网络实现实时目标检测。载于 C. Cortes、N. Lawrence、D. Lee、M. Sugiyama 和 R. Garnett（编），《神经信息处理系统进展》（NIPS），第28卷。Curran Associates, Inc.，2015年。5, 10[55]Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。利用潜扩散模型进行高分辨率图像合成。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第10684–10695页，2022年。4, 7[56]Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S Sara Mahdavi、Rapha Gontijo Lopes 等。结合深度语言理解的照片级真实感文本到图像扩散模型。arXiv 预印本 arXiv:2205.11487，2022年。7[57] Kripasindhu Sarkar、Dushyant Mehta、Weipeng Xu、Vladislav Golyanik 和 Christian Theobalt。从单张图像对人体进行神经渲染。载于欧洲计算机视觉会议（ECCV），第596–613页。Springer，2020年。8[58] Christoph Schuhmann、Romain Beaumont、Richard Vencu、Cade Gordon、Ross Wightman、Mehdi Cherti、Theo Coombes、Aarush Katta、Clayton Mullis、Mitchell Wortsman 等。LAION-5B：用于训练下一代图文模型的开放式大规模数据集。arXiv 预印本 arXiv:2210.08402，2022年。7[59] Saurabh Sharma、Pavan Teja Varigonda、Prashast Bindal、Abhishek Sharma 和 Arjun Jain。通过生成和序数排序进行单目三维人体姿态估计。载于 IEEE/CVF国际计算机视觉会议（ICCV），第2325–2334页，2019年。4[60] Dahu Shi、Xing Wei、Liangqi Li、Ye Ren 和 Wenming Tan。利用 Transformer 实现端到端多人姿态估计。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第11069–11078页，2022年。6[61] Larry Shiner。《艺术的发明：一部文化史》。芝加哥大学出版社，2003年。3[62] Aliaksandr Siarohin、Stéphane Lathuilière、Sergey Tulyakov、Elisa Ricci 和 Nicu Sebe。通过深度运动迁移为任意物体制作动画。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第2377–2386页，2019年。8[63] Aliaksandr Siarohin、Stéphane Lathuilière、Sergey Tulyakov、Elisa Ricci 和 Nicu Sebe。用于图像的第一阶运动模型

年龄、动画。《神经信息处理系统进展》（NIPS），32，2019年。8[64] Vishwanath Sindagi、Rajeev Yasarla 和 Vishal MM Patel。Jhu-crowd++：大规模人群计数数据集和一种基准方法。《IEEE模式分析与机器智能汇刊》（PAMI），2020年。2[65] Ke Sun、Bin Xiao、Dong Liu 和 Jingdong Wang。用于人体姿态估计的深度高分辨率表征学习。见IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第5693–5703页，2019年。6、10、12[66] Jiale Tao、Biao Wang、Borun Xu、Tiezheng Ge、Yuning Jiang、Wen Li 和 Lixin Duan。采用可变形锚点模型的结构感知运动迁移。见IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第3637–3646页，2022年。1、8[67] Gul Varol、Javier Romero、Xavier Martin、Naureen Mahmood、Michael J Black、Ivan Laptev 和 Cordelia Schmid。从合成人体中学习。见IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第109–117页，2017年。2[68] Hemanth Venkateswara、Jose Eusebio、Shayok Chakraborty 和 Sethuraman Panchanathan。用于无监督域适应的深度哈希网络。见IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第5018–5027页，2017年。12、13[69] Nicholas Westlake、Hongping Cai 和 Peter Hall。使用卷积神经网络检测艺术作品中的人物。见欧洲计算机视觉会议（ECCV），第825–841页。Springer，2016年。1、2、3、12、13[70] Michael J. Wilber、Chen Fang、Hailin Jin、Aaron Hertzmann、John Collomosse 和 Serge Belongie。BAM!：用于超越摄影图像识别的Behance艺术媒体数据集。见IEEE/CVF国际计算机视觉会议（ICCV），2017年10月。12[71] Jiahong Wu、He Zheng、Bo Zhao、Yixin Li、Baoming Yan、Rui Liang、Wenjia Wang、Shipei Zhou、Guosen Lin、Yanwei Fu 等。用于深化图像理解的大规模数据集。见IEEE国际多媒体与博览会会议（ICME），第1480–1485页。IEEE，2019年。2、3[72] Qi Wu、Hongping Cai 和 Peter Hall。学习图以建模不同描绘风格下的视觉对象。见欧洲计算机视觉会议（ECCV），第313–328页。Springer，2014年。12、13[73] Bin Xiao、Haiping Wu 和 Yichen Wei。人体姿态估计与跟踪的简单基线方法。见欧洲计算机视觉会议（ECCV），第466–481页，2018年。5[74] Yunqiu Xu、Yifan Sun、Zongxin Yang、Jiaxu Miao 和 Yi Yang。H2FA R-CNN：用于跨域弱监督目标检测的整体式层次化特征对齐。见IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第14329–14339页，2022年。12、13[75] Yufei Xu、Jing Zhang、Qiming Zhang 和 Dacheng Tao。ViTPose：用于人体

姿态估计的简单视觉变换器基线方法。见 Alice H. Oh、Alekh Agarwal、Danielle Belgrave 和 Kyunghyun Cho 编，《神经信息处理系统进展》（NIPS），2022年。1、2、5、6、10[76] Zhewei Xu、Jiajun Zhuang、Qiong Liu、Jingkai Zhou 和 Shaowu Peng。道路行人检测大规模远红外数据集的基准评测。《红外物理与技术》，96:199–208，2019年。2[77] Jie Yang、Ailing Zeng、Shilong Liu、Feng Li、Ruimao Zhang 和 Lei Zhang。显式边界框检测统一端到端多人姿态估计。见国际学习表征会议，2023年。6、10[78] Shuai Yang、Liming Jiang、Ziwei Liu 和 Chen Change Loy。VToonify：可控的高分辨率人像视频风格迁移。arXiv预印本 arXiv:2209.11224，2022年。1[79] Zhuoqian Yang、Wentao Zhu、Wayne Wu、Chen Qian、Qiang Zhou、Bolei Zhou 和 Chen Change Loy。TransMoMo：不变性驱动的无监督视频运动重定向。见IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第5306–5315页，2020年。1[80] Xufeng Yao、Yang Bai、Xinyun Zhang、Yuechen Zhang、Qi Sun、Ran Chen、Ruiyu Li 和 Bei Yu。PCL：用于域泛化的基于代理的对比学习。见IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第7097–7107页，2022年6月。12[81] Ailing Zeng、Xuan Ju、Lei Yang、Ruiyuan Gao、Xizhou Zhu、Bo Dai 和 Qiang Xu。DeciWatch：用于高效10倍二维和三维姿态估计的简单基线方法。见欧洲计算机视觉会议（ECCV）。Springer，2022年。5[82] Ailing Zeng、Lei Yang、Xuan Ju、Jiefeng Li、Jianyi Wang 和 Qiang Xu。SmoothNet：一种用于优化视频中人体姿态的即插即用网络。见欧洲计算机视觉会议（ECCV）。Springer，2022年。4[83] Hao Zhang、Feng Li、Shilong Liu、Lei Zhang、Hang Su、Jun Zhu、Lionel M Ni 和 Heung-Yeung Shum。DINO：采用改进去噪锚框的端到端目标检测DETR。arXiv预印本 arXiv:2203.03605，2022年。1、5、10[84] Lvmin Zhang 和 Maneesh Agrawala。为文生图扩散模型添加条件控制。arXiv预印本 arXiv:2302.05543，2023年。8[85] Song-Hai Zhang、Ruilong Li、Xin Dong、Paul Rosin、Zixi Cai、Xi Han、Dingcheng Yang、Haozhi Huang 和 Shi-Min Hu。Pose2Seg：无需检测的人体实例分割。见IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第889–898页，2019年。2、3[86] Qingyuan Zheng、Zhuoru Li 和 Adam Bargteil。学习为手绘草图添加阴影。见IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第7436–7445页，2020年。1[87] Yi Zheng、Yifan Zhao、Mengyuan Ren、He Yan、Xiangju Lu、Junhui Liu 和 Jia Li。卡通人脸识别：一个基准数据集。见ACM国际多媒体会议（MM），第2264–2272页，2020年。2[88] Kun Zhou、Xiaoguang Han、Nianjuan Jiang、Kui Jia 和 Jiangbo Lu。Hemlets Pose：学习以部件为中心的热图用于准确三维人体姿态估计的三元组。在*IEEE/CVF*国际计算机视觉会议（*ICCV*），第 2344–2353 页，2019年。4[89] Xingyi Zhou、Qixing Huang、Xiao Sun、Xiangyang Xue 和 Yichen Wei。面向野外环境中的三维人体姿态估计：一种弱监督方法。发表于*IEEE*国际计算机视觉大会论文集，第 398–407 页，2017年。13[90] Xizhou Zhu、Weijie Su、Lewei Lu、Bin Li、Xiaogang Wang 和 Jifeng Dai。Deformable DETR：用于端到端目标检测的可变形变换器。发表于国际学习表征会议（*ICLR*），2021年。5，6，10