# EgoPressure: A Dataset for Hand Pressure and Pose Estimation in Egocentric Vision

# EgoPressure:用于第一人称视角下手部压力与姿态估计的数据集

Yiming Zhao Taein Kwon Paul Streli Marc Pollefeys Christian Holz1

Yiming Zhao Taein Kwon Paul Streli Marc Pollefeys Christian Holz1

ETH Zürich Microsoft Spatial AI Lab, Zürich

苏黎世联邦理工学院 微软空间人工智能实验室，苏黎世

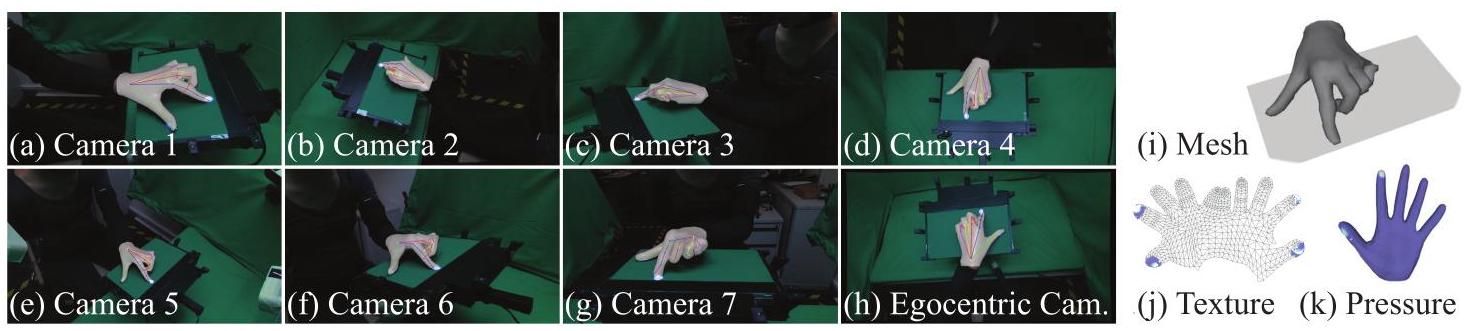


Figure 1. The EgoPressure dataset. We introduce a novel egocentric pressure dataset with hand poses. We label hand poses using our proposed optimization method across all static camera views (Cameras 1-7). The annotated hand mesh aligns well with the egocentric camera’s view, indicating the high fidelity of our annotations. We project the pressure intensity and annotated hand mesh (Fig. i) to all camera views (Fig. to ), and further provide the pressure applied over the hand as a UV texture map (Fig. and ).

图1. EgoPressure数据集。我们引入了一个包含手部姿态的新颖的第一人称压力数据集。我们使用提出的优化方法在所有静态相机视图(相机1-7)中标注手部姿态。标注的手部网格与第一人称相机的视图对齐良好，表明我们标注的高保真度。我们将压力强度和标注的手部网格(图i)投影到所有相机视图(图 到 )，并进一步提供手部施加的压力作为UV纹理图(图 和 )。

# Abstract

# 摘要

Touch contact and pressure are essential for understanding how humans interact with and manipulate objects, insights which can significantly benefit applications in mixed reality and robotics. However, estimating these interactions from an egocentric camera perspective is challenging, largely due to the lack of comprehensive datasets that provide both accurate hand poses on contacting surfaces and detailed annotations of pressure information. In this paper, we introduce EgoPressure, a novel egocentric dataset that captures detailed touch contact and pressure interactions. EgoPressure provides high-resolution pressure intensity annotations for each contact point and includes accurate hand pose meshes obtained through our proposed multi-view, sequence-based optimization method processing data from an 8-camera capture rig. Our dataset comprises 5 hours of recorded interactions from 21 participants captured simultaneously by one head-mounted and seven stationary Kinect cameras, which acquire RGB images and depth maps at . To support future research and benchmarking, we present several baseline models for estimating applied pressure on external surfaces from images, with and without hand pose information. We further explore the joint estimation of the hand mesh and applied pressure. Our experiments demonstrate that pressure and hand pose are complementary for understanding hand-object interactions. Project page: https://yiming-zhao.github.io/EgoPressure/.

触摸接触和压力对于理解人类如何与物体交互和操作至关重要，这些见解可以极大地受益于混合现实和机器人应用。然而，从第一人称相机视角估计这些交互具有挑战性，主要是由于缺乏提供接触表面上准确手部姿态和压力信息详细标注的综合数据集。在本文中，我们介绍了EgoPressure，一个新颖的第一人称数据集，捕捉了详细的触摸接触和压力交互。EgoPressure为每个接触点提供了高分辨率的压力强度标注，并包括通过我们提出的多视图、基于序列的优化方法处理来自8相机捕捉装置的数据获得的准确手部姿态网格。我们的数据集包含21名参与者5小时的交互记录，由一台头戴式和七台固定Kinect相机同时捕捉，这些相机以 获取RGB图像和深度图。为了支持未来的研究和基准测试，我们提出了几个基线模型，用于从 图像估计外部表面施加的压力，无论是否包含手部姿态信息。我们进一步探索了手部网格和施加压力的联合估计。我们的实验表明，压力和手部姿态对于理解手-物体交互是互补的。项目页面:https://yiming-zhao.github.io/EgoPressure/。

# 1. Introduction

# 1. 引言

Having a sense of touch contact and pressure during hand-object interaction is crucial for a large variety of tasks in Augmented Reality (AR) [36, 88], Virtual Reality (VR) [25, 91], and robotic manipulation [11, 13, 60]. In particular, estimating these physical properties from an egocentric perspective is a central enabler to support real-world tasks [62, 82]. In AR/VR environments, touch contact and pressure information allow for more precise control and feedback [10]. For example, when users play a virtual piano on a table, the sound can change based on the pressure applied to the keys, providing a more refined feedback experience that current AR/VR systems lack [62]. The sense of pressure is also crucial for enabling robots to accurately replicate human manipulation, as determining the precise pressure required to grasp objects remains a significant challenge [11, 13, 48].

在手-物体交互过程中，触摸接触和压力的感知对于增强现实(AR)[36, 88]、虚拟现实(VR)[25, 91]和机器人操作[11, 13, 60]中的各种任务至关重要。特别是，从第一人称视角估计这些物理属性是支持现实世界任务的核心推动力[62, 82]。在AR/VR环境中，触摸接触和压力信息允许更精确的控制和反馈[10]。例如，当用户在桌子上弹奏虚拟钢琴时，声音可以根据按键施加的压力而变化，提供当前AR/VR系统所缺乏的更精细的反馈体验[62]。压力感知对于使机器人能够准确复制人类操作也至关重要，因为确定抓取物体所需的精确压力仍然是一个重大挑战[11, 13, 48]。

Previous approaches have used gloves and robots with tactile sensors to capture pressure measurements during object manipulation. However, this instrumentation interferes with natural touch by obstructing tactile feedback. In contrast, vision-based estimation methods require no instrumentation of the hands, and cameras are already integrated into devices like smart glasses and mixed-reality headsets, which are widely used to study human behavior from an egocentric perspective [26, 27]. Despite this potential, advancements in state-of-the-art (SOTA) models have been limited by the lack of datasets that provide contact and pressure data. A notable exception is the PressureVi-sion dataset [24] that comprises RGB footage from four static cameras of hands interacting with a pressure-sensitive surface and corresponding projected pressure images.

以前的方法使用手套 和带有触觉传感器的机器人 来捕捉物体操作过程中的压力测量。然而，这种仪器通过阻碍触觉反馈干扰了自然触摸。相比之下，基于视觉的估计方法不需要对手部进行仪器化，相机已经集成到智能眼镜和混合现实头戴设备等设备中，这些设备广泛用于从第一人称视角研究人类行为[26, 27]。尽管有这种潜力，但由于缺乏提供接触和压力数据的数据集，最先进(SOTA)模型的进展受到限制。一个显著的例外是PressureVision数据集[24]，它包含来自四台静态相机的RGB视频，记录了手部与压力敏感表面交互的相应投影压力图像。

[[1]](#footnote-29)

In this paper, we present a natural yet significant extension to this prior work by introducing a novel dataset, EgoPressure, which captures hand-surface interactions from an egocentric perspective, complete with accurate hand pose annotations and pressure maps projected onto the hand mesh. Our capture platform combines a Sensel Morph touchpad with a head-mounted camera and seven synchronized Azure Kinect cameras, all recording RGB-D data at (Figure 1). The dataset includes 5 hours of footage from 21 participants, each performing 64 interaction sequences with an average length of 420 frames-making it the first bare-handed egocentric pressure dataset with pose and mesh annotations.

在本文中，我们通过引入一个新颖的数据集EgoPressure，对先前的工作 进行了自然而重要的扩展。该数据集从第一人称视角捕捉手与表面的交互，并提供了精确的手部姿态注释和投影到手部网格上的压力图。我们的捕捉平台结合了Sensel Morph触控板、头戴式摄像头和七个同步的Azure Kinect摄像头，所有设备均以 的帧率记录RGB-D数据(图1)。该数据集包含21名参与者共5小时的视频，每位参与者执行了64个交互序列，平均长度为420帧，使其成为首个带有姿态和网格注释的裸手第一人称压力数据集。

We further provide baseline models to demonstrate the potential of our dataset and establish a benchmark for future research. First, we set PressureVisionNet [24] as a baseline on our egocentric dataset and compare it to adapted models that incorporate hand pose as additional input. The model using hand poses estimated from the RGB images via the HaMeR [67] estimator outperforms PressureVisionNet by more than in volumetric IoU error, with improvements of over 7% when using ground-truth hand poses. Additionally, we introduce the first model to jointly estimate hand pose, hand mesh, and pressure both over the mesh and on the surface from an egocentric RGB camera, thereby localizing contact and pressure in space.

我们进一步提供了基线模型，以展示我们数据集的潜力，并为未来研究建立基准。首先，我们将PressureVisionNet [24]作为我们第一人称数据集的基线，并将其与结合手部姿态作为额外输入的适应模型进行比较。通过HaMeR [67]估计器从RGB图像中估计手部姿态的模型在体积IoU误差上比PressureVisionNet高出 ，使用真实手部姿态时改进超过7%。此外，我们引入了首个模型，能够从第一人称RGB摄像头联合估计手部姿态、手部网格以及网格和表面上的压力，从而在 空间中定位接触和压力。

We summarize our key contributions as follows:

我们将我们的主要贡献总结如下:

1. EgoPressure is the first high-quality egocentric touch contact and pressure dataset with hand poses. This enables the development of models that can generalize to movable cameras such as head-mounted and body-worn cameras. We will make our dataset and annotations publicly available upon acceptance.

1. EgoPressure是首个高质量的带有 手部姿态的第一人称触摸接触和压力数据集。这使得开发能够推广到可移动摄像头(如头戴式和身体佩戴摄像头)的模型成为可能。我们将在论文被接受后公开我们的数据集和注释。

2. We present an optimization method to annotate hand poses from our multi-view capture setup using MANO [73]. Our mesh-based hand pose annotations account for vertex displacement, supporting accurate hand manipulation analysis by projecting pressure inversely from the Sensel Morph touchpad onto the hand meshes.

2. 我们提出了一种优化方法，使用MANO [73]从我们的多视角捕捉设置中注释手部姿态。我们基于网格的手部姿态注释考虑了顶点位移，通过将压力从Sensel Morph触控板反向投影到手部网格上，支持精确的手部操作分析。

3. We establish two novel benchmarks: (1) estimating contact pressure from egocentric RGB images with and without hand pose information, and (2) jointly reconstructing hand poses and pressure, including the localization of pressure on a user’s hand mesh.

3. 我们建立了两个新颖的基准:(1) 从第一人称RGB图像中估计接触压力，无论是否包含手部姿态信息，(2) 联合重建 手部姿态和压力，包括在用户手部网格上定位压力。

EgoPressure thus offers new opportunities for future models to address the unique challenges of egocentric perspectives and to precisely localize pressure on a user’s hand.

因此，EgoPressure为未来模型提供了新的机会，以应对第一人称视角的独特挑战，并精确定位用户手部的压力。

# 2. Related Work

# 2. 相关工作

Our work is related to hand-object pose estimation, contact estimation and pressure sensing.

我们的工作与手-物体姿态估计、接触估计和压力传感相关。

Vision-based hand-object pose estimation Hand tracking has been a long-standing challenge in computer vision, with applications in robotics , human-computer interaction [31, 32, 70], and medicine [3, 35, 57]. Over the past decade, significant progress has been made, largely due to advancements in deep learning techniques [65, 67] and the collection of relevant datasets . While egocentric hand tracking for gesture recognition and direct input has advanced to the point of integration into modern commercial devices such as AR and VR headsets [31, 32], understanding hand interactions with external objects remains an active area of research . Datasets gathered to aid machine understanding of such hand-object interactions rely on additional instrumentation of the users’ hands [21], motion capture systems with hand-attached markers [20, 84], or multi-view camera rigs to capture accurate ground-truth poses of users’ hands under the higher degree of occlusion caused by the object.

基于视觉的手-物体姿态估计 手部跟踪一直是计算机视觉中的一个长期挑战，应用领域包括机器人 、人机交互[31, 32, 70]和医学[3, 35, 57]。过去十年中，由于深度学习技术[65, 67]的进步和相关数据集 的收集，取得了显著进展。尽管用于手势识别和直接输入的第一人称手部跟踪已经发展到可以集成到现代商业设备(如AR和VR头显[31, 32])中，但理解手与外部物体的交互仍然是一个活跃的研究领域 。为帮助机器理解此类手-物体交互而收集的数据集依赖于用户手部的额外仪器[21]、带有手部标记的运动捕捉系统[20, 84]或多视角摄像头装置 ，以在物体引起的更高程度遮挡下捕捉用户手部的精确真实姿态。

Hand-object contact estimation In addition to object-relative hand pose, prior work has aimed to estimate contact points between the users’ hands and external objects [20, 84]. Research has shown that when used as input proxies, real-world physical objects improve input control and provide haptic feedback [10]. For interactive research purposes, external tracking systems and wearable sensors such as acoustic sensors [77] and inertial measurement units were used to estimate contact [22, 28, 62, 78, 80, 85]. Additionally, vision-based techniques have been developed that use fiducial markers [49], active illumination for shadow creation [52, 88], vibration detection [81], or depth sensing [6, 19, 29, 74, 76, 89-91]. More recent work estimates touch using passive cameras without additional instrumentation on the user’s hand or surface, enabling deployment on commercial mixed reality headsets [72, 82]. More detailed contact maps are inferred based on the intersection of tracked hand and object meshes [20, 23, 47, 84, 94], requiring sub-millimeter accuracy-a challenging task for complex gestures due to soft tissue dynamics. To address this,

手与物体接触估计 除了物体相对的手部姿态，先前的研究还致力于估计用户手部与外部物体之间的接触点 [20, 84]。研究表明，当作为输入代理使用时，现实世界中的物理物体可以改善输入控制并提供触觉反馈 [10]。为了交互研究的目的，外部跟踪系统 和可穿戴传感器(如声学传感器 [77] 和惯性测量单元)被用于估计接触 [22, 28, 62, 78, 80, 85]。此外，还开发了基于视觉的技术，这些技术使用基准标记 [49]、用于阴影生成的主动照明 [52, 88]、振动检测 [81] 或深度感知 [6, 19, 29, 74, 76, 89-91]。最近的研究使用被动相机估计触摸，而无需在用户手部或表面上添加额外设备，从而可以在商用混合现实头显上部署 [72, 82]。更详细的接触图是基于跟踪的手部和物体网格的交集推断的 [20, 23, 47, 84, 94]，这需要亚毫米级的精度——对于复杂手势来说，由于软组织的动态性，这是一项具有挑战性的任务。为了解决这个问题，

Table 1. Comparison between EgoPressure and selected hand-contact datasets. The overwhelming majority of prior datasets infer contacts based on hand and object pose. ContactLabelDB and PressureVisionDB also include ground-truth touch pressure but are limited to static cameras and do not provide accurate hand poses and meshes. Please see appendix for the full table.

表1. EgoPressure与选定的手部接触数据集之间的比较。绝大多数先前的数据集基于手部和物体姿态推断接触。ContactLabelDB和PressureVisionDB还包括真实的触摸压力，但仅限于静态相机，并且不提供准确的手部姿态和网格。请参阅附录以获取完整表格。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | frames | participants | hand pose | hand mesh | markerless | real | egocentric | multiview |  | depth | contact | pressure surfacehand | |
| EgoPressure (ours) | 4.3M | 21 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Pressure sensor | ✓ | ✓ |
| ContactLabelDB [25] | 2.9M | 51 | ✘ | ✘ | ✓ | ✓ | ✘ | ✓ | ✓ | ✘ | Pressure sensor | ✓ | ✘ |
| PressureVisionDB [24] | 3.0M | 36 | ✘ | ✘ | ✓ | ✓ | ✘ | ✓ | ✓ | ✘ | Pressure sensor | ✓ | ✘ |
| ContactPose [4] | 3.0M | 50 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | Thermal imprint | ✘ | ✘ |
| GRAB [84] | 1.6M | 10 | ✓ | ✓ | ✘ | ✓ | ✘ | ✘ | ✘ | ✘ | Inferred from Pose | ✘ | ✘ |
| ARCTIC [20] | 2.1M | 10 | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |
| H2O [47] | 571k | 4 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |
| OakInk [92] | 230k | 12 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |
| OakInk-2 [94] | 4.01M | 9 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✘ | Inferred from Pose | ✘ | ✘ |
| DexYCB [7] | 582k | 10 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |
| HO-3D [30] | 103k | 10 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |
| TACO [56] | 5.2M | 14 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 帧 | 参与者 | 手部姿态 | 手部网格 | 无标记 | 真实 | 自我中心 | 多视角 |  | 深度 | 接触 | 压力表面手 | |
| EgoPressure(我们的) | 4.3M | 21 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 压力传感器 | ✓ | ✓ |
| ContactLabelDB [25] | 2.9M | 51 | ✘ | ✘ | ✓ | ✓ | ✘ | ✓ | ✓ | ✘ | 压力传感器 | ✓ | ✘ |
| PressureVisionDB [24] | 3.0M | 36 | ✘ | ✘ | ✓ | ✓ | ✘ | ✓ | ✓ | ✘ | 压力传感器 | ✓ | ✘ |
| ContactPose [4] | 3.0M | 50 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | 热印 | ✘ | ✘ |
| GRAB [84] | 1.6M | 10 | ✓ | ✓ | ✘ | ✓ | ✘ | ✘ | ✘ | ✘ | 从姿态推断 | ✘ | ✘ |
| ARCTIC [20] | 2.1M | 10 | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |
| H2O [47] | 571k | 4 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |
| OakInk [92] | 230k | 12 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |
| OakInk-2 [94] | 4.01M | 9 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✘ | 从姿态推断 | ✘ | ✘ |
| DexYCB [7] | 582k | 10 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |
| HO-3D [30] | 103k | 10 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |
| TACO [56] | 5.2M | 14 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |

Brahmbhatt et al. [4] used thermal imaging to obtain accurate contact maps. Additionally, prior efforts have utilized simulations to obtain more granular labels about contacting tissue [14, 33, 96].

Brahmbhatt等人[4]使用热成像技术获取了精确的接触图。此外，先前的研究利用模拟技术获得了关于接触组织的更细粒度标签[14, 33, 96]。

Hand pressure estimation Moving beyond the mere detection of contact, prior work has estimated the pressure forces applied during hand interactions, which is crucial for robotic grasping tasks and provides an additional control dimension for input [69]. To estimate pressure from monocular images, visual cues such as fingernail alterations or surface deformations during press events have been used. Changes in object trajectory and interaction forces also offer insights but are ineffective with static objects like tables and walls. Accurate pressure labels for training usually require instrumenting the user’s hands with gloves or the surface with force sensors , ideally flexible or conforming to various shapes [2, 45, 58]. However, this alters the visual appearance and tactile features of the hands and surface, affecting interaction and limiting generalization to bare hands and uninstrumented surfaces. Grady et al. collected two datasets with ground-truth pressure maps using a Sensel Morph [41] pressure sensor to train a neural network for estimating contact regions on surfaces from single RGB images. However, their method relies on an external static camera and good visibility of the corresponding fingertips.

手部压力估计 除了仅仅检测接触，先前的研究还估计了手部交互过程中施加的压力力，这对于机器人抓取任务 至关重要，并为输入提供了额外的控制维度[69]。为了从单目图像中估计压力，使用了视觉线索，如按压事件期间的指甲变化 或表面变形 。物体轨迹和交互力的变化 也提供了见解，但对于静态物体(如桌子和墙壁)无效。用于训练的精确压力标签通常需要为用户的手配备手套 或在表面安装力传感器 ，最好是灵活或适应各种形状的[2, 45, 58]。然而，这会改变手和表面的视觉外观和触觉特征，影响交互并限制对裸手和无仪器表面的泛化。Grady等人 使用Sensel Morph[41]压力传感器收集了两个带有真实压力图的数据集，以训练神经网络从单个RGB图像中估计表面上的接触区域。然而，他们的方法依赖于外部静态摄像头和相应指尖的良好可见性。

With EgoPressure, we aim to bridge this gap by offering a dataset that includes egocentric views, utilizing head-mounted cameras to better understand human interactions from this perspective. Our dataset also captures accurate hand poses and meshes from multiple camera views without the use of markers. To the best of our knowledge, we provide the first dataset containing egocentric and multi-view RGB-D images of a bare hand interacting with a surface, along with synchronized pressure data, hand poses, and meshes (see Table 1).

通过EgoPressure，我们旨在通过提供一个包含自我中心视角的数据集来弥合这一差距，利用头戴式摄像头更好地从这一角度理解人类交互。我们的数据集还从多个摄像头视角捕获了精确的手部姿势和网格，而无需使用标记。据我们所知，我们提供了第一个包含自我中心和多视角RGB-D图像的数据集，展示了裸手与表面交互的过程，并同步记录了压力数据、手部姿势和网格(见表1)。

# 3. Marker-less Annotation Method

# 3. 无标记注释方法

To capture accurate hand poses and meshes during hand-surface interactions without markers, we developed a multi-camera hand pose annotation method using the MANO hand model [73], differentiable rendering and multi-objective optimization. Figure 2 shows an overview of our method, which relies on static cameras and a pressure-sensitive touch-pad. Please see the supplementary material for a detailed evaluation of our annotation method.

为了在手与表面交互过程中无需标记即可捕获精确的手部姿势和网格，我们开发了一种多摄像头手部姿势注释方法，使用MANO手部模型[73]、可微分渲染和多目标优化。图2展示了我们方法的概述，该方法依赖于 静态摄像头和压力敏感触摸板。请参阅补充材料以获取我们注释方法的详细评估。

# 3.1. Automatic hand pose initialization

# 3.1. 自动手部姿势初始化

We use HaMeR [67] to estimate an initial MANO hand pose and translation for each static camera. Since HaMeR’s prediction is based on a single RGB image, there is a scale-translation ambiguity, which we resolve by triangulating the root joints from the 7 static camera views. The orientation and hand pose are then initialized based on the output of a single camera view. HaMeR also provides a bounding box, which we use along with the 2D projected hand root as input to Segment-Anything (SAM) [46], from which we obtain an annotated segmentation mask for the hand in each static camera image.

我们使用HaMeR[67]为每个静态摄像头估计初始MANO手部姿势 和平移 。由于HaMeR的预测基于单个RGB图像，存在尺度-平移模糊性，我们通过从7个静态摄像头视角三角化根关节来解决这一问题。然后基于单个摄像头视角的输出初始化方向手部姿势。HaMeR还提供了一个边界框，我们将其与2D投影的手部根一起作为Segment-Anything (SAM)[46]的输入，从中我们获得了每个静态摄像头图像中手部的注释分割掩码 。

# 3.2. Annotation refinement

# 3.2. 注释优化

Based on the initial hand pose, we obtain refined hand pose annotations via the following optimization using the input from the static cameras. We use the MANO [34,73] model for mesh representation with 25 PCA components and employ the DIB-R [9] differentiable renderer. The annotations include the hand pose , hand translation , vertex displacement in world coordinates, and the pressure over the hand mesh in the form of a texture map . All static cameras are pre-calibrated, allowing us to project the hand mesh into the frame of each static camera using the extrinsic parameters .

基于初始手部姿势，我们通过以下优化从 静态摄像头的输入中获得优化后的手部姿势注释。我们使用MANO[34,73]模型进行网格表示，包含25个PCA分量，并采用DIB-R[9]可微分渲染器。注释包括手部姿势 、手部平移 、世界坐标中的顶点位移 以及以纹理图形式表示的手部网格上的压力 。所有静态摄像头都经过预校准，允许我们使用外部参数 将手部网格投影到每个静态摄像头的帧中 。

-calibration For the MANO shape parameters , we use separate calibration sequences for each hand of each participant, during which the participant slowly turns their hand to be visible from all cameras, with fingers spread. For these sequences, we also optimize the MANO shape parameters with regularization in the previous optimization. The shape parameters are then reused for all other sequences for the given participant, with remaining fixed during subsequent optimizations.

-校准 对于MANO形状参数 ，我们为每个参与者的每只手使用单独的校准序列，在此期间，参与者缓慢转动手，使其从所有摄像头可见，手指分开。对于这些序列，我们还在之前的优化中使用 正则化来优化MANO形状参数 。然后，形状参数在给定参与者的所有其他序列中重复使用， 在后续优化中保持固定。

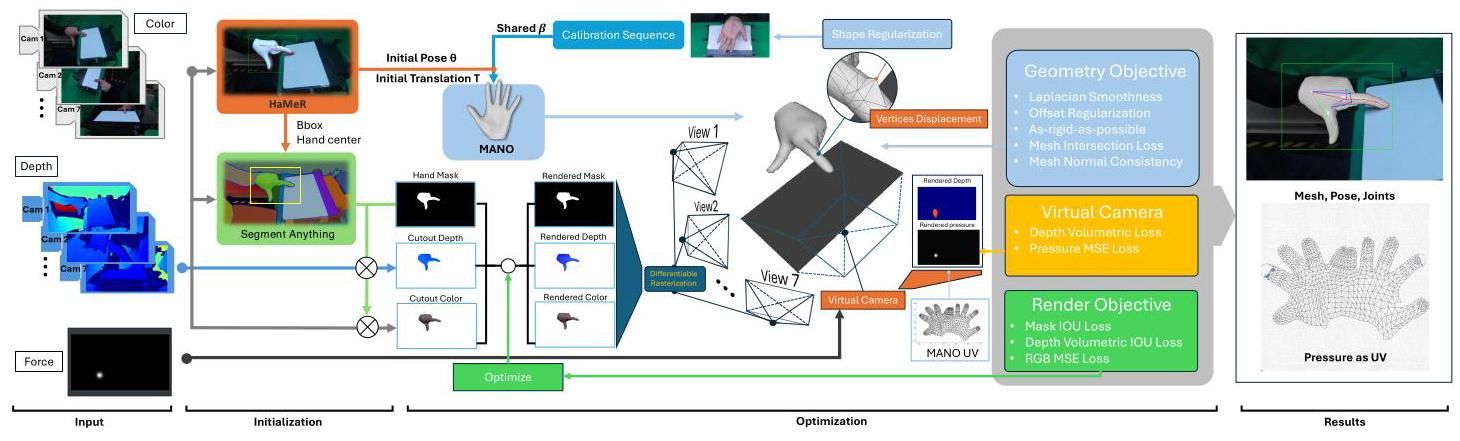


Figure 2. Method overview. The input for our annotation method consists of RGB-D images captured by 7 static Azure Kinect cameras and the pressure frame from a Sensel Morph touchpad. We leverage Segment-Anything [46] and HaMeR [67] to obtain initial hand poses and masks. We refine the initial hand pose and shape estimates through differentiable rasterization [9] optimization across all static camera views. Using an additional virtual orthogonal camera placed below the touchpad, we reproject the captured pressure frame onto the hand mesh by optimizing the pressure as a texture feature of the corresponding UV map, while ensuring contact between the touchpad and all contact vertices.

图2. 方法概述。我们的标注方法的输入包括由7个静态Azure Kinect摄像头捕获的RGB-D图像和来自Sensel Morph触控板的压力帧。我们利用Segment-Anything [46]和HaMeR [67]来获取初始手部姿势和掩码。我们通过在所有静态摄像头视图上进行可微分栅格化[9]优化来细化初始手部姿势和形状估计。使用放置在触控板下方的额外虚拟正交摄像头，我们通过将捕获的压力帧重新投影到手部网格上，将压力优化为相应UV图的纹理特征，同时确保触控板与所有接触顶点之间的接触。

Following HARP [44], our annotation algorithm consists of two stages: (1) POSE OPTIMIZATION and (2) SHAPE REFINEMENT, with a rendering objective and a geometry objective .

遵循HARP [44]，我们的标注算法包括两个阶段:(1) 姿势优化和(2) 形状细化，具有渲染目标 和几何目标 。

Beginning with the first stage, POSE OPTIMIZATION, the focus is on annotating the hand pose and translation . Consequently, the hand mesh can be derived directly from the MANO model [34], expressed as . We note that certain parts of the hand, such as fingers, may not be visible from all camera angles-for instance, fingers obscured by the palm in a curled gesture. To address this, we incorporate the mesh intersection loss . The objective function is then defined as:

从第一阶段开始，姿势优化，重点是标注手部姿势 和平移 。因此，手部网格 可以直接从MANO模型[34]导出，表示为 。我们注意到，手的某些部分，如手指，可能无法从所有摄像头角度看到——例如，手指在卷曲手势中被手掌遮挡。为了解决这个问题，我们引入了网格交叉损失 。然后，目标函数定义为:

The rendering objective and the mesh intersection loss will be detailed in the supplementary material. In the SHAPE REFINEMENT stage, the pose and translation of the hand remain fixed. The optimization process introduces vertex displacement . Each vertex is adjusted by an offset along its normal vector , which is computed from the last epoch of the POSE OPTIMIZATION stage, to minimize the rendering loss . Consequently, the refined hand mesh can be expressed as . To ensure a reasonable mesh, the geometry objective is also included in the optimization. Additionally, we introduce a virtual render to optimize pressure as a UV map and minimize the distance between the hand mesh and the contact area on the surface of the touchpad. The objective function for this stage is as follows:

渲染目标 和网格交叉损失 将在补充材料中详细说明。在形状细化阶段，手部的姿势 和平移 保持固定。优化过程引入了顶点位移 。每个顶点沿其法向量 调整一个偏移量，该法向量是从姿势优化阶段的最后一个周期计算的，以最小化渲染损失 。因此，细化后的手部网格 可以表示为 。为了确保合理的网格，几何目标 也包含在优化中。此外，我们引入了一个虚拟渲染 来优化压力作为UV图 ，并最小化手部网格 与触控板表面接触区域之间的距离。此阶段的目标函数 如下:

The virtual render , and its objective will be explained in the next section and the other terms in the geometry objective will be detailed in the supplementary material.

虚拟渲染 及其目标 将在下一节中解释，几何目标 中的其他术语将在补充材料中详细说明。

# 3.2.1. Virtual Render for Contact and Pressure

# 3.2.1. 接触和压力的虚拟渲染

As shown in Figure 2, we also incorporate the captured pressure data in the optimization as a hand mesh texture feature for our proposed virtual rendering method. For this, we position a virtual orthogonal camera under the touchpad, oriented upwards in the world coordinate system. The render size matches the resolution of the touchpad, and the camera’s plane overlaps with the touchpad’s sensing surface. The goal is for the rendered pressure on the hand mesh, with texture mapping of an optimized pressure UV map , to align with the input pressure .

如图2所示，我们还将捕获的压力数据作为手部网格纹理特征纳入优化，用于我们提出的虚拟渲染方法。为此，我们在触摸板下方放置了一个虚拟正交相机 ，在世界坐标系中朝上。渲染尺寸与触摸板的分辨率匹配，相机的平面与触摸板的感应表面重叠。目标是通过优化的压力UV贴图 进行纹理映射，使手部网格上的渲染压力 与输入压力 对齐。

Additionally, we infer the contact area from using a simple pressure threshold. Using this contact area as a mask, we ensure that the masked rendered z-axis depth aligns with the distance from the camera to the touch-pad, thereby ensuring physical contact.

此外，我们使用简单的压力阈值从 推断接触区域。使用该接触区域作为掩码，我们确保掩码渲染的z轴深度 与相机到触摸板的距离 对齐，从而确保物理接触。

The objective function for the virtual render is:

虚拟渲染的目标函数 为:

(3)

# 4. EgoPressure Dataset

# 4. EgoPressure数据集

EgoPressure comprises RGB-D frames ( for static camera, for egocentric camera) capturing interactions of both left and right hands with a touch and pressure-sensitive planar surface. The dataset features 21 participants performing 31 distinct gestures, such as touch, drag, pinch, and press, with each hand. It includes a total of 5.0 hours of hand gesture footage comprised of synchronized RGB-D frames from seven calibrated static cameras and one head-mounted camera, along with ground-truth pressure maps from the pressure-sensitive surface captured at a frame rate of . We used four different surface textures for the data capture rig, which also includes a green wall to facilitate synthetic background augmentation. Additionally, we provide high-fidelity hand pose and mesh data for the hands during interactions based on our proposed annotation method (see Section 3), as well as the tracked pose of the head-mounted camera. With EgoPressure, we aim to offer a substantial dataset for egocentric hand pose and pressure estimation during interactions with rigid surfaces, thereby advancing machine understanding of human interaction with their surroundings through the fundamental modality of touch.

EgoPressure包含 个RGB-D帧( 为静态相机， 为第一人称相机)，捕捉了左右手与触摸和压力敏感平面表面的交互。该数据集包含21名参与者，每只手执行31种不同的手势，如触摸、拖动、捏合和按压。它总共包含5.0小时的手势视频，由七个校准的静态相机和一个头戴式相机同步捕获的RGB-D帧组成，以及以 帧率捕获的压力敏感表面的真实压力图。我们使用了四种不同的表面纹理进行数据捕获，还包括一面绿色墙壁以促进合成背景增强。此外，我们基于我们提出的标注方法(见第3节)提供了交互期间手部的高保真姿态和网格数据，以及头戴式相机的跟踪姿态。通过EgoPressure，我们旨在提供一个大规模的数据集，用于第一人称视角下与刚性表面交互时的手部姿态和压力估计，从而通过触摸这一基本模态推进机器对人类与周围环境交互的理解。

# 4.1. Data capture setup

# 4.1. 数据捕获设置

To capture accurate ground-truth labels for hand pose and pressure from egocentric views, we constructed a data capture rig that integrates a pressure-sensitive touchpad (Sensel Morph [41]) for touch and pressure sensing, along with seven static and one head-mounted RGB-D camera (Azure Kinect [1]) to capture RGB and depth images (see Figure 3). The touchpad (Sensel Morph), measuring , is mounted on a tripod head. We use four different texture overlays (white, green, dark wood, light wood) printed on paper and placed over the Sensel Morph pad across participants. The seven static Azure Kinect cameras are attached to the aluminum frame, and the head-mounted camera is fixed on a helmet. The frame also holds a computer display and is surrounded by a green screen.

为了从第一人称视角捕获手部姿态和压力的准确真实标签，我们构建了一个数据捕获装置，集成了用于触摸和压力感应的压力敏感触摸板(Sensel Morph [41])，以及七个静态和一个头戴式RGB-D相机(Azure Kinect [1])来捕获RGB和深度图像(见图3)。触摸板(Sensel Morph)尺寸为 ，安装在云台上。我们使用四种不同的纹理覆盖物(白色、绿色、深色木材、浅色木材)打印在纸上，并放置在Sensel Morph板上，供参与者使用。七个静态Azure Kinect相机安装在铝制框架上，头戴式相机固定在头盔上。框架上还安装了一个计算机显示器，周围环绕着绿幕。

All cameras and the touchpad are connected to two workstations (Intel Core i7-9700K, Nvidia GeForce RTX 3070), their timestamps are synchronized via a Raspberry Pi CM4 using PTP, which also triggers all Azure Kinect cameras simultaneously at a frame rate of .

所有相机和触摸板都连接到两台工作站(Intel Core i7-9700K, Nvidia GeForce RTX 3070)，它们的时间戳通过Raspberry Pi CM4使用PTP同步，该设备还以 帧率同时触发所有Azure Kinect相机。

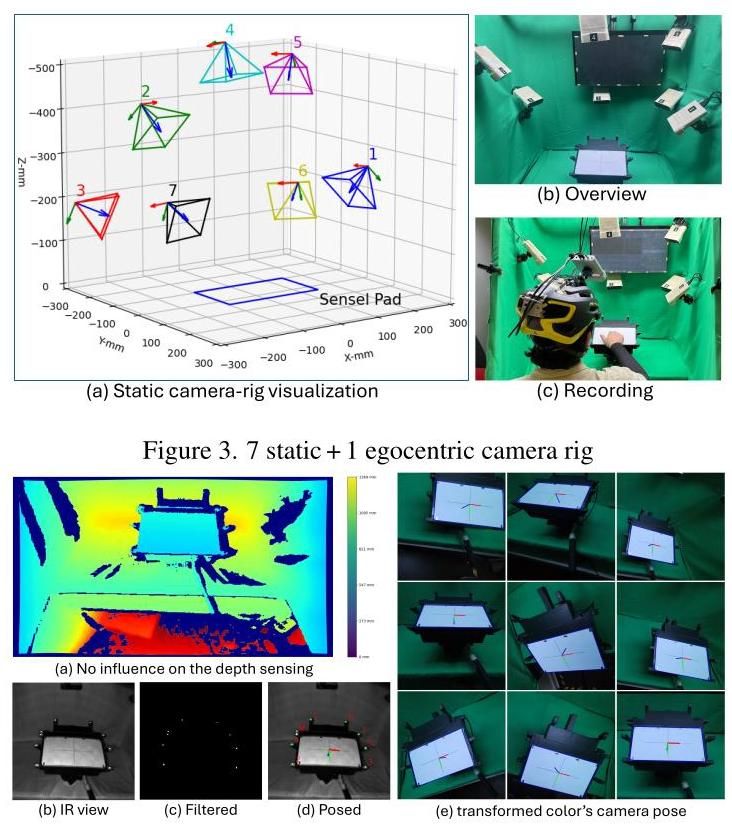


Figure 4. Camera pose tracking with IR makers

图4. 使用红外标记的相机姿态跟踪

Head-mounted camera tracking To obtain accurate poses of the head-mounted camera, we attach nine active infrared markers around the Sensel Morph pad in an asymmetric layout (see Figure 4). These markers, controlled by the Raspberry Pi CM4, are identifiable in the Azure Kinect’s infrared image using simple thresholding (saturating the range of values of the infrared camera). The markers are turned on simultaneously, allowing for the computation of the camera pose via Perspective-N-Points and enabling an accurate evaluation of the temporal synchronization between cameras and the touchpad.

头戴式相机跟踪 为了获得头戴式相机的准确姿态，我们在Sensel Morph板周围以不对称布局安装了九个主动红外标记(见图4)。这些标记由Raspberry Pi CM4控制，可以通过简单的阈值处理(饱和红外相机的值范围)在Azure Kinect的红外图像中识别。标记同时开启，允许通过Perspective-N-Points计算相机姿态，并准确评估相机与触摸板之间的时间同步。

# 4.2. Participants

# 4.2. 参与者

We recruited 21 participants from our institution ( 6 female, 15 male, ages years, mean age years), ensuring a broad representation to cover anatomic differences in hand characteristics. Participants’ heights ranged from (mean ), weights from (mean ), and middle finger lengths from 7.3- (mean ) (see Figure 5 for distribution of MANO -values). Please find details on instructions given to participants, consent form, and IRB approval in the supplementary material.

我们从本机构招募了21名参与者(6名女性，15名男性，年龄 岁，平均年龄 岁)，以确保广泛代表性，涵盖手部特征的解剖差异。参与者的身高范围为 (平均 )，体重范围为 (平均 )，中指长度为7.3- (平均 )(参见图5中MANO 值的分布)。有关参与者说明、同意书和IRB批准的详细信息，请参阅补充材料。

# 4.3. Data acquisition procedure

# 4.3. 数据采集程序

Participants sat on an adjustable stool in front of the apparatus, wearing a helmet with a mounted camera pointing towards the Sensel Morph and a black hand stocking on each arm up to the wrist. Before starting the data capture, the experimenter explained the task and the purpose of the study. They then signed a consent form and provided demographic information. The participants first performed a calibration gesture by slowly turning each hand, with fingers spread, within the camera rig. After calibration, participants conducted 31 different gestures, including touch, press, and drag gestures of varying strength, with each hand on the Sensel Morph touchpad (see supplementary material for a description of gestures). Each gesture was repeated 5 times if it involved a single touch action (e.g., press index finger) and 3 times if it involved a sequence of sequential touches (e.g., draw letters). Before each gesture, participants watched a video demonstrating how to perform the corresponding gesture with written instructions on a computer monitor in front of them. The experimenter guided the participants throughout the study, which took around 1 hour per participant. Participants could take a break after each gesture and received a chocolate bar as gratitude for their participation. In total, we recorded 6216 different gestures, i.e., 21 participants hands gestures.

参与者坐在设备前的可调节凳子上，戴着头盔，头盔上装有指向Sensel Morph的摄像头，并在每只手臂上戴有黑色手袜至手腕。在开始数据采集之前，实验者解释了任务和研究目的。然后他们签署了同意书并提供了人口统计信息。参与者首先通过缓慢转动每只手(手指张开)在摄像头装置内进行校准手势。校准后，参与者在Sensel Morph触控板上用每只手进行了31种不同的手势，包括不同力度的触摸、按压和拖动手势(参见补充材料中的手势描述)。如果手势涉及单次触摸动作(例如，按压食指)，则重复5次；如果涉及一系列连续触摸(例如，绘制字母)，则重复3次。在每个手势之前，参与者观看了一段视频，演示如何在面前的计算机显示器上执行相应的手势，并附有书面说明。实验者在整个研究过程中指导参与者，每位参与者大约需要1小时。参与者可以在每个手势后休息，并收到一块巧克力作为参与的感谢。我们总共记录了6216个不同的手势，即21名参与者 只手 个手势。

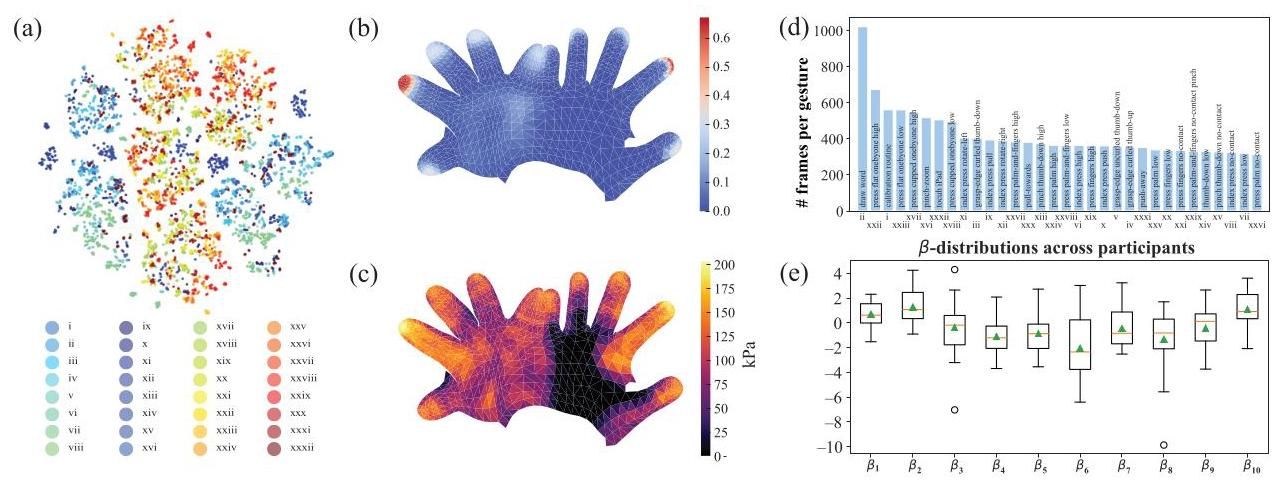


Figure 5. (a) t-SNE [87] visualization of hand pose frames over our dataset, with color coding for different gestures. All gestures are listed in Table 9 of the supplementary material. (b) Ratio of touch frames with contact for each vertex. (c) Maximum pressure over hand vertices across dataset. (d) Mean length of performed gestures. (e) Distribution of values across participants.

图5. (a) t-SNE [87] 可视化我们数据集中手部姿势帧 ，使用不同颜色编码表示不同手势。所有手势列在补充材料的表9中。(b) 每个顶点的接触帧比例。(c) 数据集中手部顶点的最大压力。(d) 执行手势的平均长度。(e) 参与者中 值的分布。



Figure 6. Thumbnail of different poses in egocentric views

图6. 以自我为中心视角的不同姿势缩略图

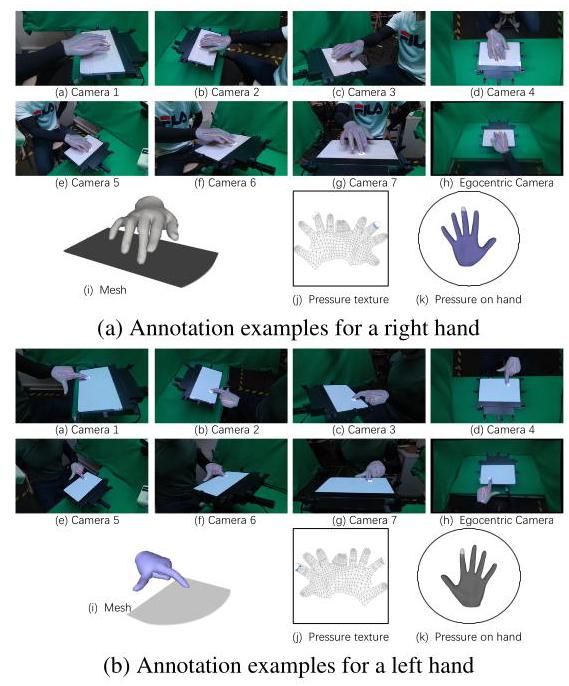


Figure 7. Sample data from EgoPressure

图7. EgoPressure的样本数据

# 4.4. Data statistics

# 4.4. 数据统计

The average length of each motion sequence is 14 seconds, with an almost equal balance between frames capturing the left and right hands. Figure 5 shows the mean sequence lengths across gestures. Approximately 45.1% of all frames capture the hand in contact with the pressure-sensitive pad. Figure 5b visualizes the ratio of contact frames with a given vertex touching the surface, and Figure 5c shows the maximum pressure measured for each vertex. Following Grady et al. [24], we set a threshold of as the minimum effective pressure to discard diffuse readings from the touchpad.

每个运动序列的平均长度为14秒，左右手的帧数几乎相等。图5显示了各手势的平均序列长度。大约45.1%的帧捕捉到手与压力敏感垫接触的情况。图5b可视化了给定顶点接触表面的接触帧比例，图5c显示了每个顶点测量的最大压力。根据Grady等人[24]的研究，我们设定 作为最小有效压力阈值，以丢弃来自触控板的扩散读数。

# 5. Benchmark Evaluation

# 5. 基准评估

Previous work estimates applied pressure maps using only RGB images [24, 25]. With EgoPressure, we explore the advantages of incorporating accurate hand poses as additional input, which naturally provide richer context about the interaction. We introduce new benchmarks for estimating hand pressure using both RGB images and 3D hand poses. Additionally, we propose a novel network architecture that jointly estimates, from a single RGB image, the pressure applied to both an external surface and across the hand, providing a deeper understanding of the regions of the hand involved throughout the interaction.

之前的工作仅使用RGB图像估计应用的压力图[24, 25]。通过EgoPressure，我们探索了将准确的手部姿势作为额外输入的优势，这自然提供了关于交互的更丰富上下文。我们引入了使用RGB图像和3D手部姿势估计手部压力的新基准。此外，我们提出了一种新颖的网络架构，可以从单个RGB图像联合估计施加在外部表面和手部各处的压力，从而更深入地了解整个交互过程中涉及的手部区域。

# 5.1. Image-projected Pressure Baselines

# 5.1. 图像投影压力基线

We evaluate the RGB-based baseline, PressureVision-Net [24], on our dataset. Since the PressureVision dataset [24] includes only static camera views, we split our baseline experiments into egocentric and exocentric views. Specifically, we use camera views 2, 3, 4, and 5 for our dataset, as these have comparable orientation and distance to the touchpad as the cameras in PressureVision.

我们在数据集上评估了基于RGB的基线PressureVision-Net [24]。由于PressureVision数据集[24]仅包含静态摄像头视图，我们将基线实验分为自我中心视角和外部视角。具体来说，我们使用数据集的摄像头视图2、3、4和5，因为这些视图与PressureVision中的摄像头具有相似的方向和距离。

We test our hypothesis that incorporating hand pose as additional input enhances pressure estimation. To this end, we design a straightforward extension of PressureVision-Net [24]. We augment the encoder-decoder segmentation architecture, originally designed for RGB inputs, by adding an additional channel for 2.5D hand key points. This involves projecting the hand joints onto the image plane and adding their depth (z-coordinate) from the egocentric camera’s coordinate system, scaled to millimeters.

我们验证了将手部姿势作为额外输入能增强压力估计的假设。为此，我们设计了一个对PressureVision-Net [24]的简单扩展。我们通过增加一个用于2.5D手部关键点的额外通道，增强了原本为RGB输入设计的编码器-解码器分割架构。这包括将 手部关节投影到图像平面，并添加它们从自我中心相机坐标系中的深度(z坐标)，以毫米为单位进行缩放。

For evaluation, we use both the ground truth hand joints from our annotations and the predicted hand joints from HaMeR [67]. The HaMeR-estimated hand poses serve as a fair baseline, reflecting the performance of current SOTA RGB-based hand pose estimators, while the ground truth joints provide an upper bound, demonstrating the potential improvement achievable with more accurate hand poses.

为了评估，我们使用了来自我们注释的真实手部关节和来自HaMeR [67]的预测手部关节。HaMeR估计的手部姿势作为一个公平的基线，反映了当前基于RGB的手部姿势估计器的性能，而真实关节则提供了一个上限，展示了通过更准确的手部姿势可实现的潜在改进。

The results are summarized in Table 2. We observe that the addition of the 2.5D hand joint layer improves performance for both egocentric and exocentric camera views. Notably, the hand poses also enhance the model’s generalization to unseen camera views. We trained the model on camera views 2, 3, 4, and 5, and evaluated it on views 1, 6, and 7 (as shown in the third row of Table 2). Qualitative results, presented in Figure 8, further demonstrate the benefits of incorporating hand pose information.

结果总结在表2中。我们观察到，添加2.5D手部关节层提高了自我中心和外中心相机视图的性能。值得注意的是，手部姿势还增强了模型对未见相机视图的泛化能力。我们在相机视图2、3、4和5上训练模型，并在视图1、6和7上进行了评估(如表2第三行所示)。图8中的定性结果进一步展示了结合手部姿势信息的好处。

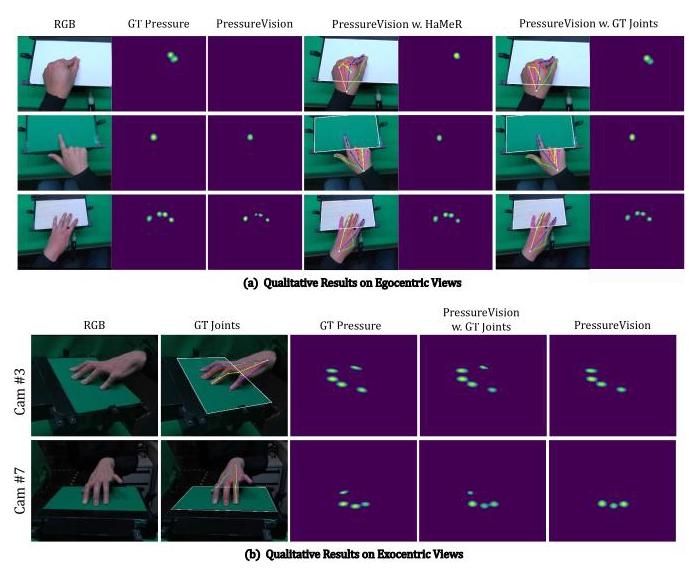


Figure 8. Qualitative results. We present the egocentric experiment results in Subfigure (a). In Subfigure (b), both baseline models are trained using camera views 2,3,4, and 5 . We display the results for one seen view and one unseen view. Additionally, we overlay the 2D keypoints predicted by HaMeR [67] and our annotated ground truth on the input image. For better visualization, the contour of the touch sensing area is also highlighted as a reference.

图8. 定性结果。我们在子图(a)中展示了自我中心实验结果。在子图(b)中，两个基线模型都使用相机视图2、3、4和5进行训练。我们展示了一个已见视图和一个未见视图的结果。此外，我们在输入图像上叠加了由HaMeR [67]预测的2D关键点和我们注释的真实值。为了更好地可视化，触摸感应区域的轮廓也被突出显示作为参考。

Table 2. Pressure inference on the full dataset (21 participants) using different input modalities. Our high-fidelity hand pose annotations improve contact IoU [%], volumetric IoU [%], MAE [Pa], and temporal accuracy compared to using no hand poses or HaMeR [67] hand poses as additional input for novel exocentric and egocentric views.

表2. 使用不同输入模态对整个数据集(21名参与者)的压力推断。我们高保真的手部姿势注释提高了接触IoU [%]、体积IoU [%]、MAE [Pa]和时间精度 ，与不使用手部姿势或使用HaMeR [67]手部姿势作为额外输入的新外中心和自我中心视图相比。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Train | Eval | Modality | Cont. IoU↑ | Vol. IoU↑ | MAE↓ | Temp. Acc. |
| PressureVisionNet [24] | Ego | Ego | RGB | 55.73 | 38.64 | 53.60 | 91.68 |
| w. [67] pose | Ego | Ego | RGB & pred pose | 56.25 | 40.52 | 55.23 | 91.67 |
| w. GT pose | Ego | Ego | RGB & GT pose | 58.80 | 41.39 | 53.79 | 92.17 |
| PressureVisionNet [24] | Exo(2,3,4,5) | Exo(2,3,4,5) | RGB | 62.11 | 44.73 | 43.15 | 93.61 |
| w. [67] pose | Exo (2,3,4,5) | Exo (2.3.4.5) | RGB & pred pose | 62.95 | 45.01 | 42,53 | 93.83 |
| w. GT pose | Exo(2,3,4,5) | Exo(2,3,4,5) | RGB & GT pose | 64.39 | 47.58 | 41.72 | 94.18 |
| PressureVisionNet [24] | Exo(2,3,4,5) | Exo(1,6,7) | RGB | 36.82 | 25.05 | 62.22 | 83.40 |
| w. [67] pose | Exo(2,3,4,5) | Exo(1.6,7) | RGB & pred pose | 38.46 | 28,10 | 51.50 | 86.34 |
| w. GT pose | Exo(2,3,4,5) | Exo(1.6,7) | RGB & GT pose | 43.04 | 31.39 | 49.45 | 89.78 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 模型 | 训练 | 评估 | 模态 | 连续IoU↑ | 体积IoU↑ | MAE↓ | 时间准确率 |
| PressureVisionNet [24] | 自我 | 自我 | RGB | 55.73 | 38.64 | 53.60 | 91.68 |
| 使用 [67] 姿态 | 自我 | 自我 | RGB & 预测姿态 | 56.25 | 40.52 | 55.23 | 91.67 |
| 使用 GT 姿态 | 自我 | 自我 | RGB & GT 姿态 | 58.80 | 41.39 | 53.79 | 92.17 |
| PressureVisionNet [24] | 外部(2,3,4,5) | 外部(2,3,4,5) | RGB | 62.11 | 44.73 | 43.15 | 93.61 |
| 使用 [67] 姿态 | 外部 (2,3,4,5) | 外部 (2.3.4.5) | RGB & 预测姿态 | 62.95 | 45.01 | 42,53 | 93.83 |
| 使用 GT 姿态 | 外部(2,3,4,5) | 外部(2,3,4,5) | RGB & GT 姿态 | 64.39 | 47.58 | 41.72 | 94.18 |
| PressureVisionNet [24] | 外部(2,3,4,5) | 外部(1,6,7) | RGB | 36.82 | 25.05 | 62.22 | 83.40 |
| 使用 [67] 姿态 | 外部(2,3,4,5) | 外部(1.6,7) | RGB & 预测姿态 | 38.46 | 28,10 | 51.50 | 86.34 |
| 使用 GT 姿态 | 外部(2,3,4,5) | 外部(1.6,7) | RGB & GT 姿态 | 43.04 | 31.39 | 49.45 | 89.78 |

# 5.2. First Hand-projected Pressure Baseline

# 5.2. 首次手部投影压力基线

Both the original PressureVision framework [24] and its subsequent iteration, PressureVision++ [25], predict 2D hand pressure on the image plane. However, this introduces ambiguity about the exact manifestation of this pressure between hands and objects within the space.

原始的PressureVision框架[24]及其后续迭代版本PressureVision++[25]都预测了图像平面上的二维手部压力。然而，这引入了关于手部和物体之间在 空间内压力具体表现的模糊性。

To address this, we introduce a new baseline model, Pres-sureFormer, which estimates pressure as a UV map of the hand mesh, enabling projection both as pressure onto the hand surface and as pressure onto the image space.

为了解决这个问题，我们引入了一个新的基线模型PressureFormer，它将压力估计为 手部网格的UV图，使得压力既可以投影为 手部表面的压力，也可以投影为 图像空间中的压力。

As illustrated in Figure 9, our model builds upon HaMeR [67]. It processes the hand vertices in the camera frame and the image feature tokens from HaMeR’s Vision Transformer (ViT) [17]. A transformer-based decoder receives as multiple input tokens while cross-attending to the image feature tokens from the ViT. Each output token represents a -dimensional feature for a corresponding mesh vertex, which we then map onto a UV feature map using the UV coordinates of the MANO model [73]. Given the sparsity of the UV feature map post-projection, we apply two convolutional layers for neural interpolation and reduce the dimensions to the number of force classes to predict the quantized UV-pressure map .

如图9所示，我们的模型基于HaMeR[67]。它处理相机帧中的手部顶点 和来自HaMeR的Vision Transformer (ViT)[17]的图像特征标记。一个基于Transformer的解码器接收 作为多个输入标记，同时交叉关注来自ViT的图像特征标记。每个输出标记代表对应网格顶点的 维特征，然后我们使用MANO模型[73]的UV坐标将其映射到UV特征图上。考虑到投影后UV特征图的稀疏性，我们应用了两层卷积层进行神经插值，并将维度减少到力类别数 ，以预测量化的UV压力图 。

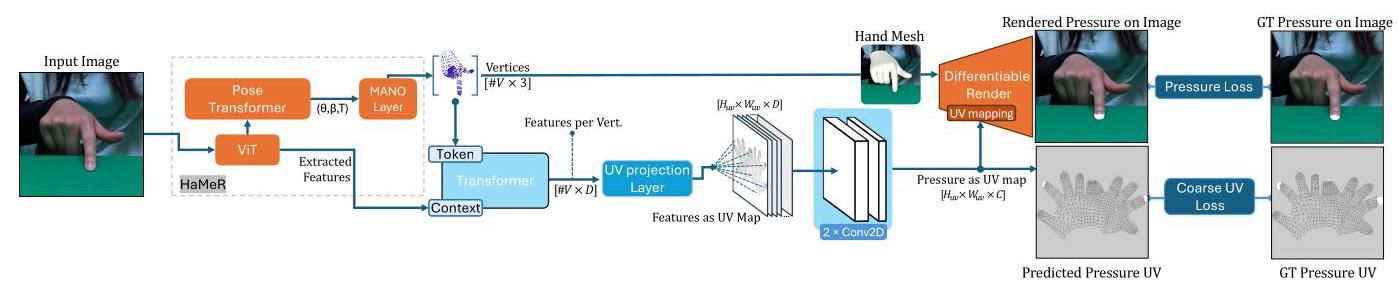


Figure 9. PressureFormer uses HaMeR’s hand vertices and image feature tokens to estimate the pressure distribution over the UV map. We employ a differentiable renderer [9] to project the pressure back onto the image plane by texture-mapping it onto the predicted hand mesh.

图9. PressureFormer使用HaMeR的手部顶点和图像特征标记来估计UV图上的压力分布。我们采用可微分渲染器[9]通过纹理映射将压力投影回图像平面，并将其映射到预测的手部网格上。

Initially, we compute the coarse UV-pressure loss between and the ground-truth UV-pressure map , which is converted from the scalar UV pressure in our dataset. Subsequently, we render the pressure back onto the original image plane based on the mesh of vertices and the predicted UV-pressure map. Using a differentiable renderer [9], we invert the z-normal and -axis of the face vertices to identify the mesh faces that are farthest from and invisible to the camera view, marking them as potential contact locations. This allows us to compute the pressure loss with respect to the ground-truth pressure . We employ Cross-Entropy loss for both and , resulting in the following loss function for Pressure-Former:

首先，我们计算 和真实UV压力图 之间的粗略UV压力损失 ，该图是从我们数据集中的标量UV压力 转换而来的。随后，我们基于 顶点网格 和预测的 UV压力图将压力 渲染回原始图像平面。使用可微分渲染器[9]，我们反转面顶点的z法线和 轴，以识别距离相机视图最远且不可见的网格面，将其标记为潜在的接触位置。这使我们能够计算相对于真实压力 的压力损失 。我们对 和 都采用交叉熵损失，从而得到PressureFormer的以下损失函数:

We trained the PressureFormer model and baseline models using hand-centered image crops across all camera views. During training, we applied data augmentation techniques, including shifting, rescaling, and rotating. The results are summarized in Table 3, with visualizations provided in Figure 10. Additional analyses are included in the supplementary material.

我们使用所有相机视图中的手部中心图像裁剪来训练PressureFormer模型和基线模型。在训练过程中，我们应用了数据增强技术，包括平移、缩放和旋转。结果总结在表3中，并在图10中提供了可视化。补充材料中包含了额外的分析。

Table 3. Our method achieves the highest performance in terms of contact IoU and performs comparably to other approaches on additional evaluation metrics. Notably, it offers significant advantages over the image-projected pressure baselines by directly predicting pressure on the UV map, enabling the reconstruction of 3D pressure through projection onto the estimated hand surface.

表3. 我们的方法在接触IoU方面取得了最高的性能，并在其他评估指标上与其他方法表现相当。值得注意的是，它通过直接在UV图上预测压力，显著优于图像投影压力基线，从而能够通过投影到估计的手部表面来重建3D压力。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Contact IoU↑ | Vol. IoU↑ | MAE [kPa] | Temp. Acc. [%]↑ |
| PressureVisionNet [24] | 40.71 | 32.11 | 44 | 90 |
| (w. HaMeR [67] pose) | 42.52 | 35.40 | 49 | 92 |
| PressureFormer | 43.04 | 31.57 | 71 | 89 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | 接触IoU↑ | 体积IoU↑ | MAE [kPa] | 温度准确率 [%]↑ |
| PressureVisionNet [24] | 40.71 | 32.11 | 44 | 90 |
| (使用HaMeR [67] 姿态) | 42.52 | 35.40 | 49 | 92 |
| PressureFormer | 43.04 | 31.57 | 71 | 89 |



Figure 10. Qualitative Results PressureFormer on our dataset. We compare our PressureFormer with both PressureVision [24] and our extended baseline model with HaMeR-estimated [67] 2.5D joint positions. Additionally, we provide visualizations of the hand mesh estimated by HaMeR, alongside the 3D pressure distribution on the hand surface derived from our predicted UV-pressure in the last two columns. Note that we transform the left-hand UV maps into the right-hand format.

图10. 在我们的数据集上，PressureFormer的定性结果。我们将PressureFormer与PressureVision [24] 和我们扩展的基线模型(使用HaMeR估计的[67] 2.5D关节位置)进行了比较。此外，我们提供了由HaMeR估计的手部网格的可视化，以及从我们预测的UV压力得出的手部表面3D压力分布，显示在最后两列中。请注意，我们将左手UV图转换为右手格式。

# 6. Conclusion

# 6. 结论

In this paper, we introduce EgoPressure, a novel egocentric hand pressure dataset paired with a multi-view hand pose estimation method. EgoPressure includes precise hand poses with meshes, multi-view RGB and depth images, egocentric view images, and high-quality pressure data. We establish a new benchmark and demonstrate the effectiveness of using hand pose data in pressure estimation. For future work, we plan to enhance our dataset by including objects to enable pressure estimation on more complex geometries. In conclusion, we believe that EgoPressure represents a significant step towards a deeper machine understanding of hand-object interactions from egocentric views.

在本文中，我们介绍了EgoPressure，这是一个新颖的以自我为中心的手部压力数据集，与多视角手部姿态估计方法相结合。EgoPressure包括精确的手部姿态与网格、多视角RGB和深度图像、以自我为中心的视角图像以及高质量的压力数据。我们建立了一个新的基准，并展示了使用手部姿态数据进行压力估计的有效性。对于未来的工作，我们计划通过包含物体来增强我们的数据集，以便在更复杂的几何形状上进行压力估计。总之，我们认为EgoPressure代表了从以自我为中心的视角深入理解手与物体交互的重要一步。

# EgoPressure: A Dataset for Hand Pressure and Pose Estimation in Egocentric Vision

# EgoPressure:以自我为中心视觉中的手部压力与姿态估计数据集

Supplementary Material

补充材料

# 7. Details about Benchmark Evaluation

# 7. 基准评估的详细信息

In this section, we provide further details about the benchmark evaluation experiments from Section 5.

在本节中，我们提供了第5节中基准评估实验的更多细节。

# 7.1. Details for image-projected pressure baselines

# 7.1. 图像投影压力基线的详细信息

# 7.1.1. Baseline model with Additional 2.5D Keypoints

# 7.1.1. 带有额外2.5D关键点的基线模型

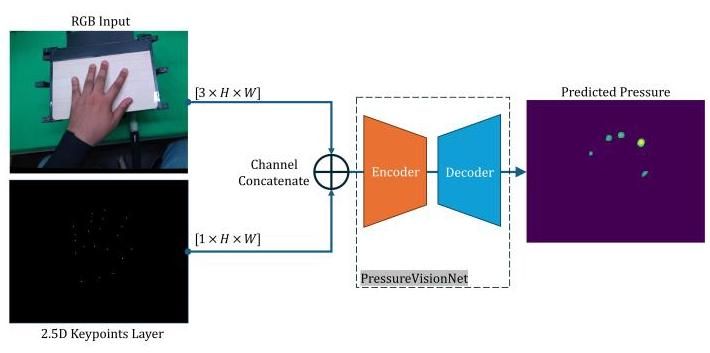


Figure 11. Overview of the image-projected pressure baseline with additional hand pose input. The baseline receives an RGB image and a 2.5D keypoint depth map as inputs to an encoder-decoder segmentation network for pressure estimation.

图11. 带有额外手部姿态输入的图像投影压力基线概述。该基线接收RGB图像和2.5D关键点深度图作为输入，通过编码器-解码器分割网络进行压力估计。

The previous method [24] for predicting hand pressure relies solely on RGB images as inputs. In contrast, our new benchmark is designed to incorporate an additional modality, hand pose. To ensure a fair comparison between the baselines and our approach, we extend the existing method with additional hand pose inputs. In addition to the three RGB channels of PressureVision, we add a 2.5D depth map as an additional input channel to the segmentation network.

先前的方法[24]用于预测手部压力仅依赖于RGB图像作为输入。相比之下，我们的新基准设计为包含额外的模态，即手部姿态。为了确保基线与我们的方法之间的公平比较，我们通过添加额外的手部姿态输入扩展了现有方法。除了PressureVision的三个RGB通道外，我们还添加了2.5D深度图作为分割网络的额外输入通道。

Encoder-decoder segmentation network architecture. Similar to PressureVision, we employ an ImageNet-pretrained Squeeze-and-Excitation Network (SERes-NeXt50) [37, 38] as the encoder, which takes both RGB and 3D hand pose inputs, and a feature pyramid network [40, 54] as the decoder, which generates a pressure map.

编码器-解码器分割网络架构。与PressureVision类似，我们采用ImageNet预训练的Squeeze-and-Excitation网络(SERes-NeXt50)[37, 38]作为编码器，该编码器接收RGB和3D手部姿态输入，并使用特征金字塔网络[40, 54]作为解码器，生成压力图。

Training. In all experiments, data from 15 participants is used for training and validation, while data from 6 participants is held out as the test set. For training, we use the Adam optimizer with a batch size of 8 . The training process begins with a learning rate of 0.001 for iterations, followed by iterations with a learning rate of 0.0001 .

训练。在所有实验中，15名参与者的数据用于训练和验证，而6名参与者的数据作为测试集保留。对于训练，我们使用Adam优化器，批量大小为8。训练过程从学习率为0.001开始，进行 次迭代，然后进行 次迭代，学习率为0.0001。

# 7.1.2. Evaluation Metrics

# 7.1.2. 评估指标

For evaluation, we adopt the four metrics proposed in Pres-sureVision [24]: Contact Intersection over Union (IoU), Volumetric IoU, Mean Absolute Error (MAE), and Temporal Accuracy.

为了评估，我们采用了PressureVision [24]提出的四个指标:接触交并比(IoU)、体积IoU、平均绝对误差(MAE)和时间准确性。

Contact IoU measures the accuracy of contact surface predictions by calculating the IoU between the estimated and ground truth binarized pressure maps. Volumetric IoU extends this by incorporating the accuracy of the predicted pressure magnitudes, calculated as the ratio of the sum of the minimum pressure values between the estimated and ground truth pressure maps at each pixel to the sum of the maximum values. MAE quantifies the pressure prediction error in kilopascals ( ) per pixel. Temporal Accuracy assesses the consistency of contact over time by verifying frame-by-frame contact consistency between the estimated and ground truth values.

接触IoU通过计算估计和真实二值化压力图之间的IoU来测量接触表面预测的准确性。体积IoU通过结合预测压力幅度的准确性扩展了这一点，计算为每个像素上估计和真实压力图之间最小压力值之和与最大值之和的比率。MAE以千帕( )每像素量化压力预测误差。时间准确性通过逐帧验证估计值和真实值之间的接触一致性来评估接触的时间一致性。

# 7.2. Additional Qualitative Results

# 7.2. 额外的定性结果

More qualitative results for the baselines are provided in Figure 27. More qualitative examples for the annotations are shown in Figures 30, 31, 32, 33, 34 and 35.

图27中提供了基线的更多定性结果。图30、31、32、33、34和35中展示了更多注释的定性示例。

We also present qualitative results from the third-person view camera experiments (refer to Table 2 in the main paper). Figure 28 and 29 include visual comparisons between our model, which uses RGB and 2.5D hand keypoints, and PressureVisionNet [24] which uses only RGB input. Figure 28 shows the models’ qualitative performance on images from cameras 2, 3, 4, and 5, with both models trained on a separate training set from these views. In Figure 28, we evaluate the same models on novel views from cameras1,6, and 7, which were not included in the training set.

我们还展示了第三人称视角相机实验的定性结果(参见主论文中的表2)。图28和29展示了我们模型(使用RGB和2.5D手部关键点)与仅使用RGB输入的PressureVisionNet [24]之间的视觉对比。图28展示了模型在相机2、3、4和5图像上的定性表现，这两个模型都在这些视角的单独训练集上进行了训练。在图28中，我们在相机1、6和7的新视角上评估了相同的模型，这些视角未包含在训练集中。

In the second column of Figure 28 and Figure 29, the reprojected touch sensing area is shown as a white outline to verify the camera pose. We also provide MAE and Contact IoU values for each sample. Notably, including additional hand pose information enhances the model’s ability to estimate pressure and contact, especially for occluded hand parts (see examples 04 in Figure 28 and 09, 11, 13 in Figure 29).

在图28和图29的第二列中，重新投影的触觉感应区域以白色轮廓显示，以验证相机姿态。我们还为每个样本提供了MAE和接触IoU值。值得注意的是，包含额外的手部姿态信息增强了模型估计压力和接触的能力，特别是对于被遮挡的手部部分(参见图28中的示例04和图29中的示例09、11、13)。

# 7.3. Additional Evaluation of PressureFormer

# 7.3. PressureFormer的额外评估

PressureFormer improves upon the baselines from Section 5.1 by estimating pressure directly on the UV map of the reconstructed hand mesh. This approach extends the representation of pressure via the estimated hand pose into 3D space. While the hand mesh-based pressure representation can still be projected onto the image plane for benchmarking with prior methods [24, 25], it offers additional insights about the specific hand regions applying pressure. This capability is beneficial for scenarios involving complex hand-object interactions, such as when fingers are partially occluded or interacting with non-planar surfaces, where an image-projected pressure map may have limitations and introduce additional ambiguities. These tactile hand dynamics are also helpful for enabling precise grasping and object manipulation in humanoid robotics.

PressureFormer通过直接在重建的手部网格的UV图上估计压力，改进了第5.1节中的基线。这种方法通过估计的 手部姿态将压力表示扩展到3D空间。虽然基于手部网格的压力表示仍然可以投影到图像平面上以与先前的方法[24, 25]进行基准测试，但它提供了关于施加压力的特定手部区域的额外见解。这种能力对于涉及复杂手-物体交互的场景非常有益，例如当手指部分被遮挡或与非平面表面交互时，图像投影的压力图可能会有局限性并引入额外的模糊性。这些触觉手部动态也有助于在人形机器人中实现精确的抓取和物体操作。

# 7.3.1. Accuracy of estimated UV Pressure Map

# 7.3.1. 估计的UV压力图的准确性

In Section 5.2, we compare PressureFormer with PressureVi-sionNet [24] and its 2.5D hand keypoint-augmented baseline, both of which directly estimate camera image-projected pressure maps. We make these comparisons based on the evaluation metrics established in PressureVision (see Table 4).

在第5.2节中，我们将PressureFormer与PressureVisionNet [24]及其2.5D手部关键点增强的基线进行了比较，这两者都直接估计相机图像投影的压力图。我们基于PressureVision中建立的评估指标进行了这些比较(见表4)。

We extend this evaluation by considering the hand mesh-projected pressure that PressureFormer directly estimates as a UV pressure map (see Figure 9). For comparison, we project the image-based pressure maps from PressureVision-Net and its hand-pose-augmented baseline onto the corresponding hand mesh estimated from the same image using the HaMeR [67]. This involves identifying the hand mesh faces furthest from the camera (i.e., occluded vertices) and rasterizing the 2D pressure map onto the UV map (see Figure 12). Additionally, we evaluate a variant of Pressure-Former trained without explicit UV loss supervision.

我们通过考虑PressureFormer直接估计的手部网格投影压力作为UV压力图(见图9)扩展了这一评估。为了进行比较，我们将PressureVisionNet及其手部姿态增强基线的基于图像的压力图投影到使用HaMeR [67]从同一图像估计的相应手部网格上。这涉及识别距离相机最远的手部网格面(即被遮挡的顶点)并将2D压力图栅格化到UV图上(见图12)。此外，我们评估了一个没有显式UV损失监督的PressureFormer变体。

We thus introduce a novel benchmarking task that evaluates the accuracy of pressure on the hand surface and the performance of jointly estimating pressure and hand mesh.

因此，我们引入了一个新的基准测试任务，评估手部表面压力的准确性以及联合估计压力和手部网格的性能。

Evaluation Metrics. To assess the accuracy of pressure estimation across the hand surface, we compute two metrics on the UV pressure map: Contact IoU and Volumetric IoU.

评估指标。为了评估手部表面压力估计的准确性，我们在UV压力图上计算了两个指标:接触IoU和体积IoU。

Training. The models are trained and evaluated using images from all camera views. We use 15 participants for training and validation, with 6 participants held out as the test set. During preprocessing, the images are cropped with a margin around the hand and resized to match the network’s input dimensions. For evaluation, we ensure the hand remains centrally positioned in the frame throughout the cropping process. Data augmentation, including shifts, rescaling, and rotations, is applied across all methods. Training employs the Adam optimizer with a batch size of 8, using a learning rate of 0.001 for iterations and 0.0001 for the subsequent iterations. The loss function for PressureFormer (see Eq. 4) uses weighting parameters and .

训练。模型使用所有相机视角的图像进行训练和评估。我们使用15名参与者进行训练和验证，其中6名参与者作为测试集。在预处理过程中，图像在手的周围进行裁剪并调整大小以匹配网络的输入尺寸。为了评估，我们确保手在整个裁剪过程中始终位于帧的中心位置。数据增强，包括平移、缩放和旋转，应用于所有方法。训练使用Adam优化器，批量大小为8，使用0.001的学习率进行 次迭代，随后使用0.0001的学习率进行 次迭代。PressureFormer的损失函数(见公式4)使用权重参数 和 。

Results. The results from Section 5.2 and the UV map-based evaluation are summarized in Table 4. PressureFormer outperforms all image-projected pressure baselines in terms of Contact IoU and Volumetric IoU on the UV pressure map. It also attains the highest Contact IoU on the image-projected pressure map. The hand-pose-augmented baseline, which directly predicts pressure onto the camera image, achieves the best Volumetric IoU on the image-based pressure map. These results highlight the value of incorporating hand pose information for pressure estimation and underscore the potential of further research into the joint estimation of hand pose and pressure for more coherent interaction modeling.

结果。第5.2节和基于UV图的评估结果总结在表4中。PressureFormer在UV压力图上的接触IoU和体积IoU方面优于所有基于图像投影的压力基线。它还在图像投影压力图上获得了最高的接触IoU。直接预测相机图像上压力的手姿态增强基线在基于图像的压力图上实现了最佳体积IoU。这些结果突显了结合手姿态信息进行压力估计的价值，并强调了进一步研究手姿态和压力联合估计以实现更一致交互建模的潜力。

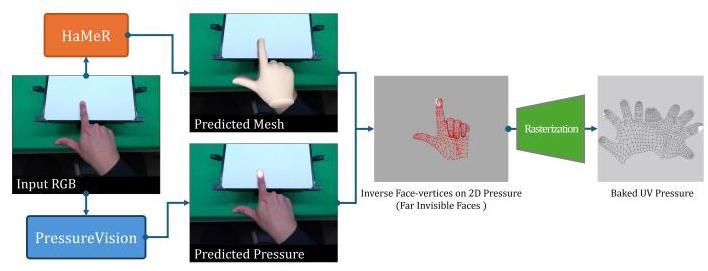


Figure 12. Pipeline for projecting the image-based pressure map (from PressureVision) onto the UV map: Starting with the predicted hand mesh and 2D pressure map, the normals and z-axis are inverted to identify occluded (invisible) faces of the mesh. The pressure is then mapped onto the UV space using rasterization.

图12。将基于图像的压力图(来自PressureVision)投影到UV图上的流程:从预测的手网格和2D压力图开始，反转法线和z轴以识别网格的遮挡(不可见)面。然后使用光栅化将压力映射到UV空间。

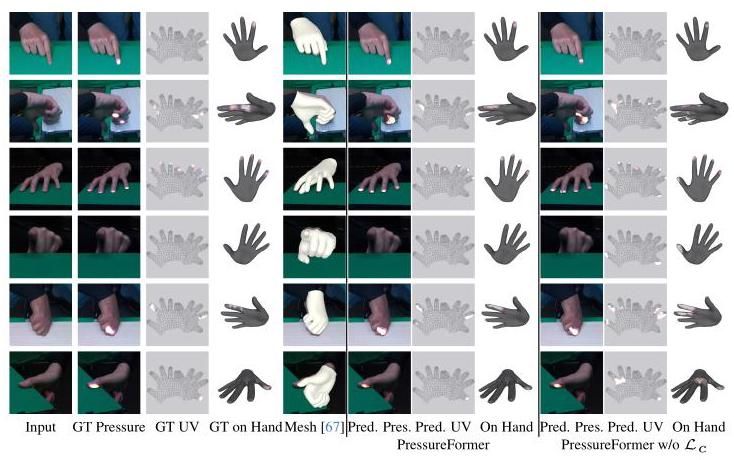


Figure 13. Qualitative examples demonstrating the impact of coarse UV loss supervision . The coarse UV loss supervision prevents the prediction of pressure in areas of the UV map that are not rendered on the image plane (see Section 5.2). These regions typically correspond to faces oriented toward the camera, where pressure and contact are not physically possible.

图13。展示粗略UV损失监督 影响的定性示例。粗略UV损失监督 防止在UV图上预测未在图像平面上渲染的区域的压力(见第5.2节)。这些区域通常对应于朝向相机的面，在这些区域压力和接触在物理上是不可能的。

Additionally, the results underline the value of the coarse UV-pressure loss in enhancing the accuracy of the pressure predictions on the UV map (see Figure 13).

此外，结果强调了粗略UV压力损失在提高UV图上压力预测准确性方面的价值(见图13)。

Figure 14 provides a qualitative comparison of the UV pressure maps estimated by the three baseline methods.

图14提供了三种基线方法估计的UV压力图的定性比较。

# 7.3.2. Generalization of PressureFormer

# 7.3.2. PressureFormer的泛化

Employing a UV-pressure map can improve the generalization of hand contact and pressure prediction for more complex objects. Unlike estimating pressure on the image plane, which focuses on hand-surface interactions, UV-pressure mapping can highlight hand-centric pressure by directly predicting pressure on the hand vertices.

使用UV压力图可以提高手接触和压力预测对于更复杂对象的泛化能力。与在图像平面上估计压力不同，后者侧重于手与表面的交互，而UV压力映射通过直接预测手顶点上的压力来突出手中心的压力。

Table 4. Performance comparison of our PressureFormer model against image-projected pressure baselines, evaluated using temporal accuracy [%], image-based pressure metrics (Image Contact IoU, Image Vol. IoU, Image MAE [kPa]), and UV map-based pressure metrics (UV Pressure IoU, UV Pressure Vol. IoU). PressureFormer demonstrates superior performance in UV pressure IoU and UV Pressure Vol. IoU, while also achieving higher scores in image-based Contact IoU. By directly predicting pressure on the UV map, PressureFormer offers advantages, enabling accurate 3D pressure reconstruction by projecting the results onto the estimated hand surface.

表4。我们的PressureFormer模型与基于图像投影的压力基线的性能比较，使用时间准确性[%]、基于图像的压力指标(图像接触IoU、图像体积IoU、图像MAE[kPa])和基于UV图的压力指标(UV压力IoU、UV压力体积IoU)进行评估。PressureFormer在UV压力IoU和UV压力体积IoU方面表现出色，同时在基于图像的接触IoU上也获得了更高的分数。通过直接在UV图上预测压力，PressureFormer提供了优势，能够通过将结果投影到估计的手表面上实现准确的3D压力重建。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Im. Contact IoU↑ | Im. Vol. IoU↑ | Im. MAE | UV Press. Contact IoU ↑ | UV Press. Vol. IoU↑ | Temp. Acc. |
| PressureVisionNet [24] | 40.71 | 32.11 | 44 | 21.53 | 16.41 | 90 |
| (w. HaMeR [67] pose) | 42.52 | 35.40 | 49 | 24.10 | 17.36 | 92 |
| PressureFormer (Ours) | 43.04 | 31.57 | 71 | 33.12 | 24.54 | 89 |
| PressureFormer w/o | 41.27 | 29.57 | 74 | 26.24 | 18.61 | 88 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 模型 | 图像接触IoU↑ | 图像体积IoU↑ | 图像MAE | UV压力接触IoU↑ | UV压力体积IoU↑ | 温度准确度 |
| PressureVisionNet [24] | 40.71 | 32.11 | 44 | 21.53 | 16.41 | 90 |
| (使用HaMeR [67]姿态) | 42.52 | 35.40 | 49 | 24.10 | 17.36 | 92 |
| PressureFormer (我们的) | 43.04 | 31.57 | 71 | 33.12 | 24.54 | 89 |
| PressureFormer 无 | 41.27 | 29.57 | 74 | 26.24 | 18.61 | 88 |

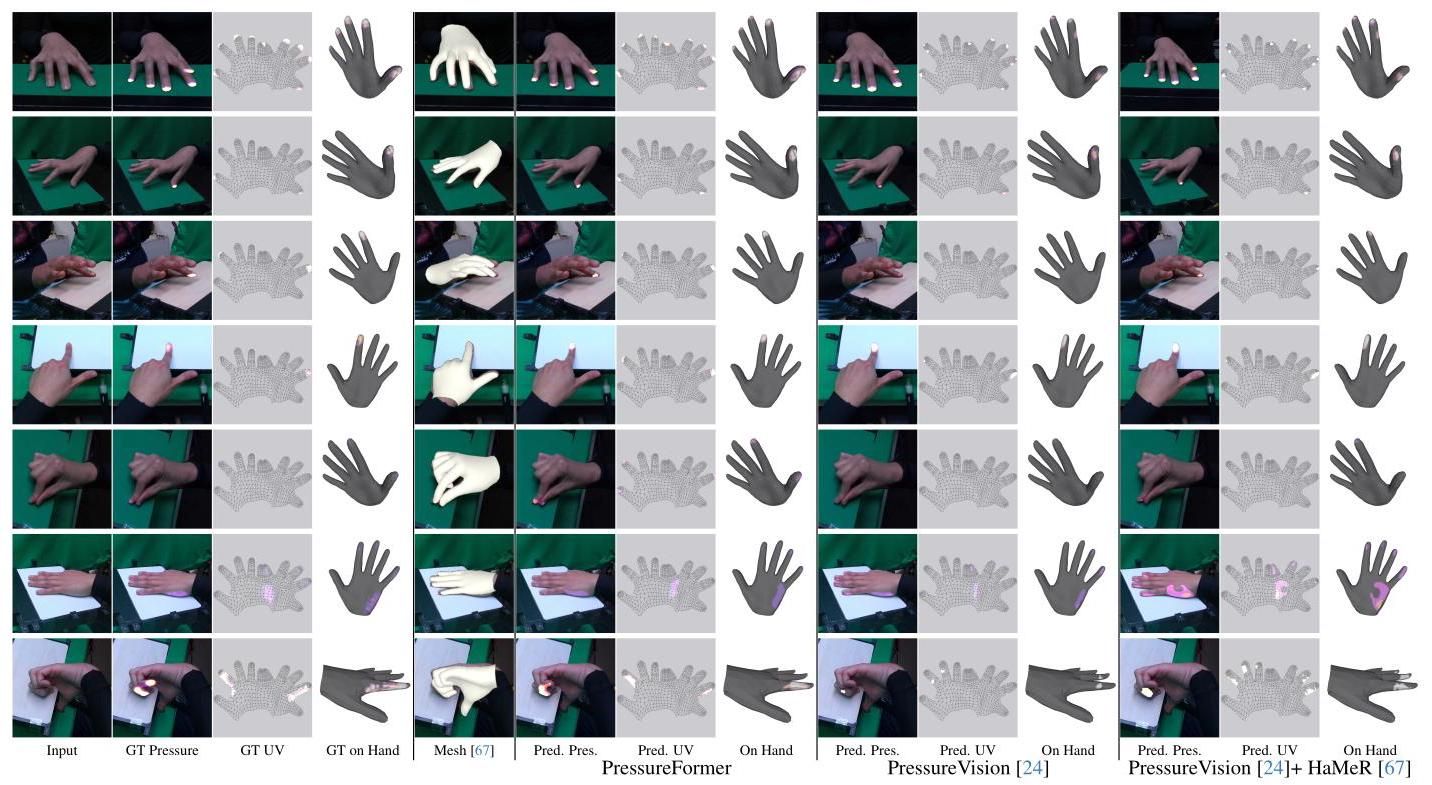


Figure 14. Qualitative comparison of UV Pressure. We compare our PressureFormer model against the original PressureVision [24] and its extended version with additional hand keypoint inputs. For both PressureVision-based approaches, the UV pressure is obtained by baking the image-based pressure predictions onto the UV map of the hand mesh, using the hand mesh estimates provided by HaMeR [67].

图14. UV压力的定性比较。我们将我们的PressureFormer模型与原始的PressureVision [24]及其扩展版本(包含额外的手部关键点输入)进行比较。对于这两种基于PressureVision的方法，UV压力是通过将基于图像的压力预测烘焙到手部网格的UV图上获得的，使用的是HaMeR [67]提供的手部网格估计。

Our model, PressureFormer, utilizes the pretrained HaMeR [67] as its backbone to extract hand vertices and image features from the vision transformer tokens. This enables our approach to effectively handle diverse hand poses while integrating hand-centric image texture information encoded in the vision transformer tokens. The qualitative results, shown in Figure 15, demonstrate PressureFormer’s ability to generalize to unseen objects and environments.

我们的模型PressureFormer利用预训练的HaMeR [67]作为其骨干，从视觉变换器(vision transformer)的token中提取手部顶点和图像特征。这使得我们的方法能够有效处理多样化的手部姿势，同时整合编码在视觉变换器token中的手部中心图像纹理信息。图15所示的定性结果展示了PressureFormer在未见过的物体和环境中的泛化能力。

# 8. Details and Evaluation of Annotation Method

# 8. 标注方法的细节与评估

# 8.1. Optimization Objectives

# 8.1. 优化目标

In this section, we describe the optimization objectives necessary for complete implementation in conjunction with the objectives described in the main paper.

在本节中，我们描述了与主论文中描述的优化目标相结合所需的完整实现优化目标。

# 8.1.1. Render Objective

# 8.1.1. 渲染目标

Since hand mesh is the only rendered object across all camera views, we use pseudo groundtruth mask from Segment-Anything (SAM) [46] to extract relevant regions, appearance and depth , from input RGB image and depth . For the optimization of the rendered appearance , a single texture is shared across all camera views within an input batch of several consecutive frames, which ensures that the mesh remains consistent across different cameras and consecutive frames. The rendering loss across all cameras is represented in Eq. 5

由于手部网格 是所有相机视图中唯一渲染的对象，我们使用Segment-Anything (SAM) [46]中的伪真实掩码 从输入RGB图像 和深度 中提取相关区域、外观 和深度 。为了优化渲染的外观 ，在连续几帧的输入批次中，所有相机视图共享单一纹理，这确保了网格 在不同相机和连续帧之间保持一致。所有 个相机的渲染损失 在公式5中表示。

Depth Volumetric IoU [24] is defined in the third term of Equation 5. We apply it to the ground truth and rendered depth. In Table 5, we show these two losses: Depth

深度体积IoU [24]在公式5的第三项中定义。我们将其应用于真实深度和渲染深度。在表5中，我们展示了这两种损失:深度



Figure 15. Qualitative evaluation of PressureFormer on diverse, real-world examples featuring various objects and scenes. Despite being trained exclusively on EgoPressure, the model recognizes pressure regions during corresponding contact events, demonstrating its potential for generalization.

图15. PressureFormer在多样化的真实世界示例中的定性评估，这些示例包含各种物体和场景。尽管仅在EgoPressure上训练，模型在相应的接触事件中识别出压力区域，展示了其泛化潜力。

Volumetric IoU Loss and Mask IoU Loss on the mesh from the initial input, i.e., and , and two consecutive annotation stages, POSE OPTIMIZATION and Shape Refinement.

体积IoU损失 和掩码IoU损失 在初始输入的网格上，即 和 ，以及两个连续的标注阶段:姿势优化和形状细化。

# 8.1.2. Geometry Objective

# 8.1.2. 几何目标

The geometry objective is composed of several terms:

几何目标 由多个项组成:

The term represents the mesh intersection loss, which utilizes a BVH tree to identify self-intersections within the mesh. Penalties are subsequently applied based on these detections [43, 86].

项 表示网格交叉损失，它利用BVH树来识别网格内的自交叉。随后根据这些检测结果施加惩罚[43, 86]。

Table 5. Losses by Stages. We validate the quality of hand poses using two metrics, Depth Volumetric IoU Loss (Eq. 5) and Mask IoU Loss (Eq. 5), computed on 386,231 (static cameras) annotated frames. Of these, 2,192,633 (81%) show the hand in contact with the touchpad. We report the results before (initial) and after each consecutive optimization step: POSE OPTIMIZATION and SHAPE REFINEMENT.

表5. 各阶段的损失。我们使用两个指标验证手部姿势的质量:深度体积IoU损失 (公式5)和掩码IoU损失 (公式5)，这些指标在386,231 (静态相机) 标注帧上计算。其中，2,192,633帧(81%)显示手部与触摸板接触。我们报告了在每次连续优化步骤(姿势优化和形状细化)之前(初始)和之后的结果。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Category |  | | |  | | |
| Initial | Pose. | Pose. + Shape. | Initial | Pose. | Pose. + Shape. |
| Overall | 0.4443 | 0.1759 | 0.1317 | 0.3887 | 0.1165 | 0.0558 |
| With Contact | 0.4444 | 0.1752 | 0.1309 | 0.3891 | 0.1167 | 0.0562 |
| Without Contact | 0.4441 | 0.1790 | 0.1351 | 0.3871 | 0.1158 | 0.0545 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 类别 |  | | |  | | |
| 初始 | 姿势 | 姿势 + 形状 | 初始 | 姿势 | 姿势 + 形状 |
| 整体 | 0.4443 | 0.1759 | 0.1317 | 0.3887 | 0.1165 | 0.0558 |
| 有接触 | 0.4444 | 0.1752 | 0.1309 | 0.3891 | 0.1167 | 0.0562 |
| 无接触 | 0.4441 | 0.1790 | 0.1351 | 0.3871 | 0.1158 | 0.0545 |

The term , as-rigid-as-possible loss, as introduced in [79], promotes increased rigidity in the 3D mesh while distributing length alterations across multiple edges. The variation in edge length is determined relative to the mesh from the last epoch of POSE OPTIMIZATION as

术语 ，即尽可能刚性损失，如[79]中所述，在3D网格中促进更高的刚性，同时将长度变化分布在多个边缘上。边缘长度的变化是相对于POSE OPTIMIZATION最后一个时期的网格确定的，如下所示:

where is the edge connecting vertex and in the set of all edges , and the edge is formed by the corresponding vertices and in the mesh without vertex displacement .

其中 是连接顶点 和 的边，属于所有边的集合 ，而边 是由网格中未发生顶点位移 的对应顶点 和 形成的。

The mesh vertices are smoothed by the Laplacian mesh regularization , and the normal consistency regularization smooths normals on the displaced mesh. Finally, the vertex offset term is calculated by .

网格顶点 通过拉普拉斯网格正则化 进行平滑处理，而法线一致性正则化 则对位移网格上的法线进行平滑处理。最后，顶点偏移项 由 计算得出。

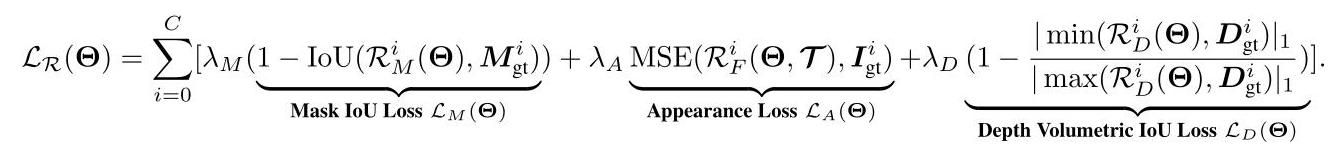
# 8.1.3. Depth Culling

# 8.1.3. 深度剔除

In some sequences, hands may be partially occluded by the Sensel Morph touchpad from certain camera views, which can hinder the convergence of the optimization process for the total rendered mask. To address this issue, we have modeled the touchpad and its pedestal. We pre-generate the depth map to represent these scene obstacles. Subsequently, we perform simple depth culling with the rendered depth by generating a culling mask . This allows us to create cutouts on the rendered depth , the appearance , and the mask , which together represent the hand parts in front of the scene obstacles. After initial tests, we noticed that this depth culling encourages the intersection of the hand mesh and the touchpad to reach lower mask IoU loss . Therefore, we add a collision box of the touchpad into mesh intersection loss to penalize this intersection. We show an example in Figure 16.

在某些序列中，手部可能会被Sensel Morph触控板从某些相机视角部分遮挡，这可能会阻碍总渲染掩码优化过程的收敛。为了解决这个问题，我们对触控板及其底座进行了建模。我们预先生成深度图 来表示这些场景障碍物。随后，我们通过生成剔除掩码 对渲染深度 进行简单的深度剔除。这使得我们能够在渲染深度 、外观 和掩码 上创建剪切区域，这些区域共同代表位于场景障碍物前方的手部部分。在初步测试后，我们注意到这种深度剔除促使手部网格与触控板的交会达到较低的掩码IoU损失 。因此，我们将触控板的碰撞框添加到网格交会损失 中，以惩罚这种交会。我们在图16中展示了一个示例。

(5)



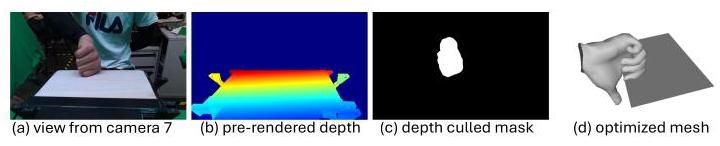


Figure 16. Depth Culling. (a) In the view of Camera 7, the thumb is behind the touchpad. (b) We compare the rendered depth of hand and pre-rendered depth map of scene obstacles , and (c) cutout the part which has a larger depth value than . The thumb rendered in blue color is cutout due to the depth culling. (d) The collision box is rendered in .

图16. 深度剔除。(a) 在相机7的视角下，拇指位于触控板后方。(b) 我们比较了手部的渲染深度 和场景障碍物的预渲染深度图 ，(c) 并剪切了深度值大于 的部分。由于深度剔除，渲染为蓝色的拇指被剪切掉。(d) 碰撞框在 中渲染。

# 8.1.4. Temporal Continuity

# 8.1.4. 时间连续性

Our optimization considers consecutive captures consisting of 7 RGB-D and one pressure frame in batches of size to ensure temporal continuity of annotated hand poses across timestamps. We apply regularization on the approximated second-order derivative of the hand joint positions , which are regressed from the MANO mesh. The temporal continuity regularization is:

我们的优化考虑了由7个RGB-D帧和一个压力帧组成的连续捕获，批次大小为 ，以确保跨时间戳的标注手部姿势的时间连续性。我们对从MANO网格回归的手关节位置的近似二阶导数 应用正则化。时间连续性正则化如下:

# 8.2. Evaluation of Annotation Fidelity

# 8.2. 标注保真度评估

# 8.2.1. Manual Annotation and Inspection

# 8.2.1. 手动标注与检查

To verify the quality of the hand poses from our annotation method, we manually annotated 300 randomly selected sets of 7 static views and one egocentric view frames). We annotated all the visible nail tips in the camera views, resulting in 7176 2D points. These 2D nail tips were then triangulated to obtain 3D points. After applying a threshold of 2 pixels on the re-projection error to exclude inconsistent manual annotations, we obtained 1114 3D points that were visible in at least two camera views. In Table 6, we report the distance error of the hand tips obtained from our annotation method relative to the 3D tip positions based on the manual annotations. We also include an ablation study of our approach. Qualitative results of the manual annotations and our method are shown in Figure 17.

为了验证我们标注方法的手部姿势质量，我们手动标注了随机选择的300组7个静态视图和一个第一人称视图 帧。我们标注了相机视图中所有可见的指甲尖，共得到7176个2D点。然后，这些2D指甲尖通过三角测量法转换为3D点。在应用2像素的重投影误差阈值以排除不一致的手动标注后，我们获得了1114个在至少两个相机视图中可见的3D点。在表6中，我们报告了从我们的标注方法获得的手指尖相对于基于手动标注的3D指尖位置的距离误差。我们还对我们的方法进行了消融研究。手动标注和我们方法的定性结果如图17所示。

# 8.2.2. Comparison to learning-based model

# 8.2.2. 与基于学习模型的比较

Compared to the state-of-the-art 3D hand pose estimator, HaMeR, our optimization-based method offers significant advantages, enabling the creation of high-quality annotations for our dataset. As shown in Figure 19, although hand poses

与最先进的3D手部姿势估计器HaMeR相比，我们基于优化的方法具有显著优势，能够为我们的数据集创建高质量的标注。如图19所示，尽管手部姿势

Table 6. Quantitative evaluation of our annotation method compared to 3D tip positions triangulated from manual annotations. We conduct an ablation study for the different loss terms, including appearance loss (Eq. 5), depth volumetric IoU loss (Eq. 5), mesh intersection loss (Sec. 8.1.2), as-rigid-as-possible loss (Sec. 8.1.2), Laplacian smoothness (Sec. 8.1.2), normal consistency regularization (Sec. 8.1.2), and vertex offset regularization (Sec. 8.1.2). We demonstrate that each loss term contributes to our optimization performance. from HaMeR [67] appear plausible from a top view, side views expose inaccuracies and scale ambiguities. In contrast, our annotation method produces robust and consistent results across all camera views. In Table 2), we demonstrate that the baseline model with our high-quality hand poses improves hand pressure estimation compared to using HaMeR’s [67] predictions.

表6. 我们的标注方法与从手动标注三角测量得到的3D指尖位置的定量评估。我们对不同的损失项进行了消融研究，包括外观损失 (公式5)、深度体积IoU损失 (公式5)、网格交叉损失 (第8.1.2节)、尽可能刚性损失 (第8.1.2节)、拉普拉斯平滑度 (第8.1.2节)、法线一致性正则化 (第8.1.2节)和顶点偏移正则化 (第8.1.2节)。我们证明了每个损失项对我们的优化性能都有贡献。从HaMeR [67]的预测在俯视图中看起来合理，但侧视图暴露了不准确性和尺度模糊性。相比之下，我们的标注方法在所有相机视图中都产生了稳健且一致的结果。在表2中，我们展示了使用我们高质量 手部姿势的基线模型相比使用HaMeR [67]的预测在手部压力估计方面的改进。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Losses | Ours | w/o |  | w/o | w/o | w/o |  | w/o |
|  | 0.1251 | 0.1404 | 0.1797 | 0.1368 | 0.1347 | 0.1396 | 0.1349 | 0.1477 |
|  | 0.0488 | 0.0662 | 0.0724 | 0.0619 | 0.0597 | 0.0654 | 0.0653 | 0.0728 |
| 3D tips error [mm]↓ | 5.68 | 7.66 | 8.45 | 7.99 | 8.61 | 8.49 | 8.39 | 8.28 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 损失 | 我们的 | 不含 |  | 不含 | 不含 | 不含 |  | 不含 |
|  | 0.1251 | 0.1404 | 0.1797 | 0.1368 | 0.1347 | 0.1396 | 0.1349 | 0.1477 |
|  | 0.0488 | 0.0662 | 0.0724 | 0.0619 | 0.0597 | 0.0654 | 0.0653 | 0.0728 |
| 3D尖端误差[毫米]↓ | 5.68 | 7.66 | 8.45 | 7.99 | 8.61 | 8.49 | 8.39 | 8.28 |

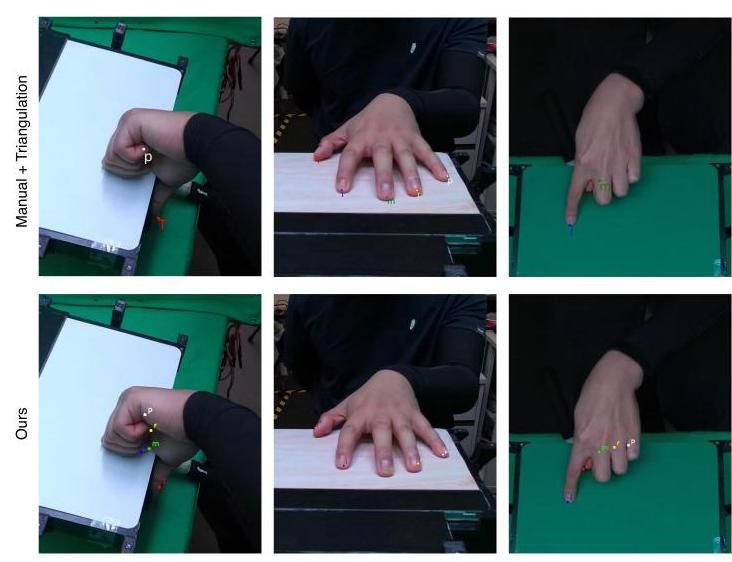


Figure 17. Manual Verification Examples. We demonstrate our annotation is accurate compared to the manual annotations. (above) We re-project the triangulated nail tips. We only triangulated them when they are visible in at least 2 views. (bottom) We re-project our 3D annotations which also show invisible nail tips as well.

图17. 手动验证示例。我们展示了我们的标注与手动标注相比是准确的。(上)我们重新投影了三角化的指甲尖端。我们仅在它们至少在两个视图中可见时进行三角化。(下)我们重新投影了我们的3D标注，这些标注也显示了不可见的指甲尖端。

To further evaluate annotation quality, we provide the validation results comparing the triangulation of predicted nail tips with manual annotations across static views in Table 7. Additionally, Figure 20 presents a qualitative comparison of pressure estimation incorporating additional poses from HaMeR [67] and our ground truth annotations. The results emphasize the importance of the high-fidelity hand pose annotations from our optimization method, both quantitatively and qualitatively, and highlight the necessity of advancing hand pose and pressure map estimation in future research.

为了进一步评估标注质量，我们提供了验证结果，比较了预测的指甲尖端三角化与手动标注在静态视图中的结果，如表7所示。此外，图20展示了结合HaMeR [67]额外姿势的压力估计的定性比较以及我们的真实标注。结果强调了我们的优化方法生成的高保真手部姿势标注的重要性，无论是定量还是定性，并突出了在未来研究中推进手部姿势和压力图估计的必要性。

Finally, we report the results of the HaMeR method after fine-tuning on our dataset in Table 8 and in Figure 18. Although fine-tuning improves performance, there remains room for further enhancement. These results establish a solid baseline for tackling 3D hand pose estimation during hand-surface interactions in an egocentric view.

最后，我们报告了HaMeR方法在我们的数据集上微调后的结果，如表8和图18所示。尽管微调提高了性能，但仍存在进一步改进的空间。这些结果为在自我中心视图中处理手部表面交互时的3D手部姿势估计奠定了坚实的基础。

Table 7. Hand pose verification. Triangulation is performed on the nail tips using HaMeR [67] predictions across all static cameras, compared against manual annotations.

表7. 手部姿势验证。使用HaMeR [67]预测在所有静态相机上对指甲尖端进行三角化，并与手动标注进行比较。

|  |  |  |
| --- | --- | --- |
|  | 3D tips error [mm] | Std. |
| Ours | 5.68 | 4.9 |
| HaMeR [67] | 12.37 | 6.3 |

|  |  |  |
| --- | --- | --- |
|  | 3D尖端误差 [毫米] | 标准 |
| 我们的 | 5.68 | 4.9 |
| HaMeR [67] | 12.37 | 6.3 |

Table 8. Fine-tuning results of HaMeR [67] on EgoPressure demonstrate improved hand pose accuracy, underscoring the value of our dataset for hand pose estimation.

表8. HaMeR [67]在EgoPressure数据集上的微调结果显示出手部姿态精度有所提高，这凸显了我们的数据集在 手部姿态估计(hand pose estimation)方面的价值。

|  |  |  |
| --- | --- | --- |
|  | MPJPE [mm] | Reconstruction Error [mm] |
| Finetuned HaMeR [67] | 10.75 | 6.10 |
| HaMeR [67] | 18.58 | 8.11 |

|  |  |  |
| --- | --- | --- |
|  | MPJPE [毫米] | 重建误差 [毫米] |
| 微调后的HaMeR [67] | 10.75 | 6.10 |
| HaMeR [67] | 18.58 | 8.11 |

# 9. Extended Details about Dataset

# 9. 数据集的扩展详情

# 9.1. Details about Gesture Description

# 9.1. 手势描述的详情

Table 9 lists all gestures performed by a participant during the data collection, including which hands were used and how often each gesture was repeated. We refer to the accompanying video for visual examples.

表9列出了参与者在数据收集过程中执行的所有手势，包括使用了哪只手以及每个手势的重复次数。我们参考随附的视频以获取视觉示例。

# 9.2. Dataset Comparisons

# 9.2. 数据集比较

Table 10 provides a comprehensive comparison of our proposed dataset. Among existing public datasets focusing on contact or hand-object pose estimation, EgoPressure is the first dataset to combine egocentric video data of hand-surface interactions with ground-truth contact and pressure information, as well as high-fidelity hand poses and meshes.

表10提供了我们提出的数据集的全面比较。在现有的专注于接触或手-物体姿态估计的公共数据集中，EgoPressure是第一个将手-表面交互的自我中心视频数据与真实接触和压力信息以及高保真手部姿态和网格相结合的数据集。

# 9.3. Details about Active IR Marker

# 9.3. 主动红外标记的详情

We use active IR Marker, operating similarly to passive markers, these markers emit their own infrared light, allowing for a much smaller and more precise form factor-often appearing as tiny light dots in the filtered infrared image. This reduces the impact of lens distortion on tracking accuracy. Moreover, these markers are programmable, providing crucial control over their activation and deactivation, which is vital for synchronization within our system. We utilize the infrared led with large beam angle (see Fig. 23) as active infrared marker.

我们使用主动红外标记，其操作方式与被动标记类似，这些标记发射自己的红外光，允许更小且更精确的外形尺寸——通常在过滤后的红外图像中表现为微小的光点。这减少了镜头失真对跟踪精度的影响。此外，这些标记是可编程的，提供了对其激活和停用的关键控制，这对于我们系统中的同步至关重要。我们使用大光束角的红外LED(见图23)作为主动红外标记。

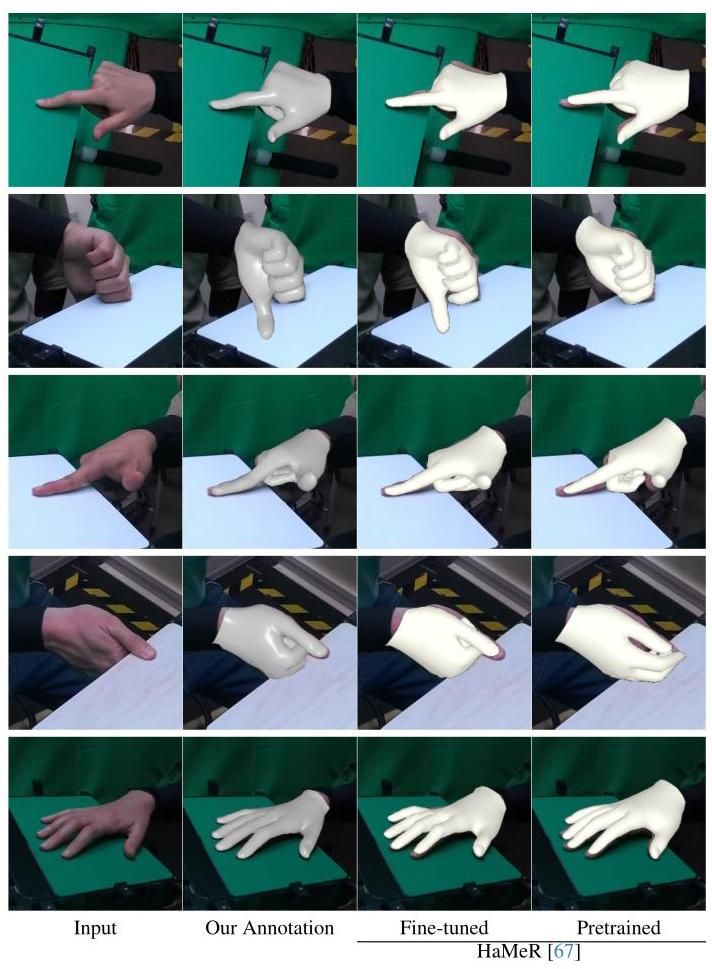


Figure 18. Hand pose prediction and ground truth pose visualization for each camera. We fine-tune HaMeR [67] on our dataset, demonstrating improved detail in hand pose estimation, particularly in scenarios where the hand interacts with a surface.

图18. 每个相机的手部姿态预测和真实姿态可视化。我们在我们的数据集上微调了HaMeR [67]，展示了手部姿态估计的改进细节，特别是在手与表面交互的场景中。

An asymmetrical layout with markers can be uniquely identified from any viewpoint within the upper hemisphere above the marker arrangement. This distinctive configuration enables robust and accurate real-time tracking using filtered infrared images, where the markers appear as light dots with a radius of several pixels. The process is detailed in the pseudocode presented in Algorithm 1.The effectiveness of this layout in facilitating accurate marker identification and pose estimation is further illustrated in Figure 24, where the spatial arrangement of markers is depicted. Furthermore, this procedure can be generalized to other asymmetrical layouts.

带有标记的不对称布局可以从标记排列上方的上半球内的任何视角唯一识别。这种独特的配置使得使用过滤后的红外图像进行稳健且准确的实时跟踪成为可能，其中标记表现为半径为几个像素的光点。该过程在算法1中提供的伪代码中详细说明。这种布局在促进准确标记识别和姿态估计方面的有效性在图24中进一步说明，其中描绘了标记的空间排列。此外，该过程可以推广到其他不对称布局。

The Perspective-n-Points (PnP) algorithm is used to compute the camera pose of the egocentric camera based on the identified markers in the infrared frame. In the experiment, the reprojection error for pose computed from well-identified markers remained below an average of 0.4 pixels. To ensure clarity and reliability in recognition, we applied a threshold value of 1 pixel to filter out frames potentially containing ambiguities in marker recognition during the recording. Additionally, for frames where tracking was lost, spherical linear interpolation (Slerp) is employed to estimate camera pose, thereby maintaining continuity and accuracy in the tracking data.

Perspective-n-Points (PnP) 算法用于基于红外帧中识别的标记计算自我中心相机的相机姿态。在实验中，从良好识别的标记计算出的姿态的重投影误差保持在平均0.4像素以下。为了确保识别的清晰度和可靠性，我们应用了1像素的阈值来过滤掉在记录过程中可能包含标记识别模糊的帧。此外，对于跟踪丢失的帧，使用球面线性插值(Slerp)来估计相机姿态，从而保持跟踪数据的连续性和准确性。

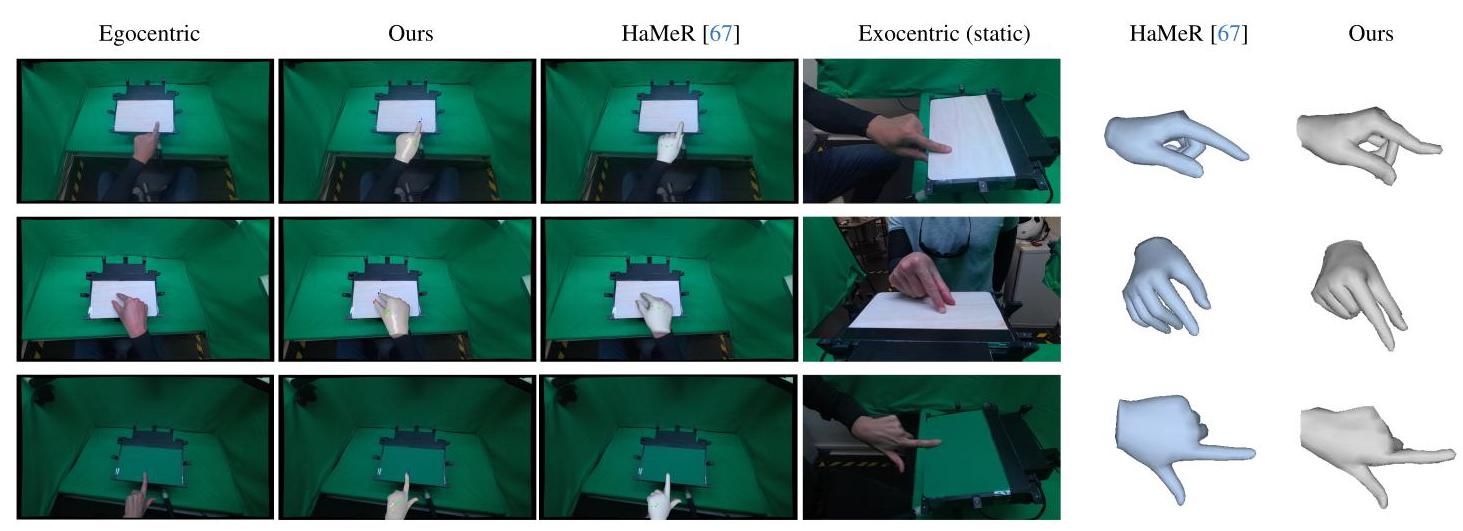


Figure 19. Comparison of the estimated hand mesh from HaMeR [67] and our annotation method in both egocentric and exocentric views. While the projected hand mesh from HaMeR appears visually plausible from an egocentric perspective, observable differences in hand articulation and mesh deformations become apparent from the exocentric viewpoint of the static cameras.

图19. HaMeR [67] 和我们注释方法在自我中心和外中心视图中的估计手部网格比较。虽然从自我中心视角来看，HaMeR投影的手部网格在视觉上是合理的，但从静态相机的外中心视角来看，手部关节和网格变形的差异变得明显。

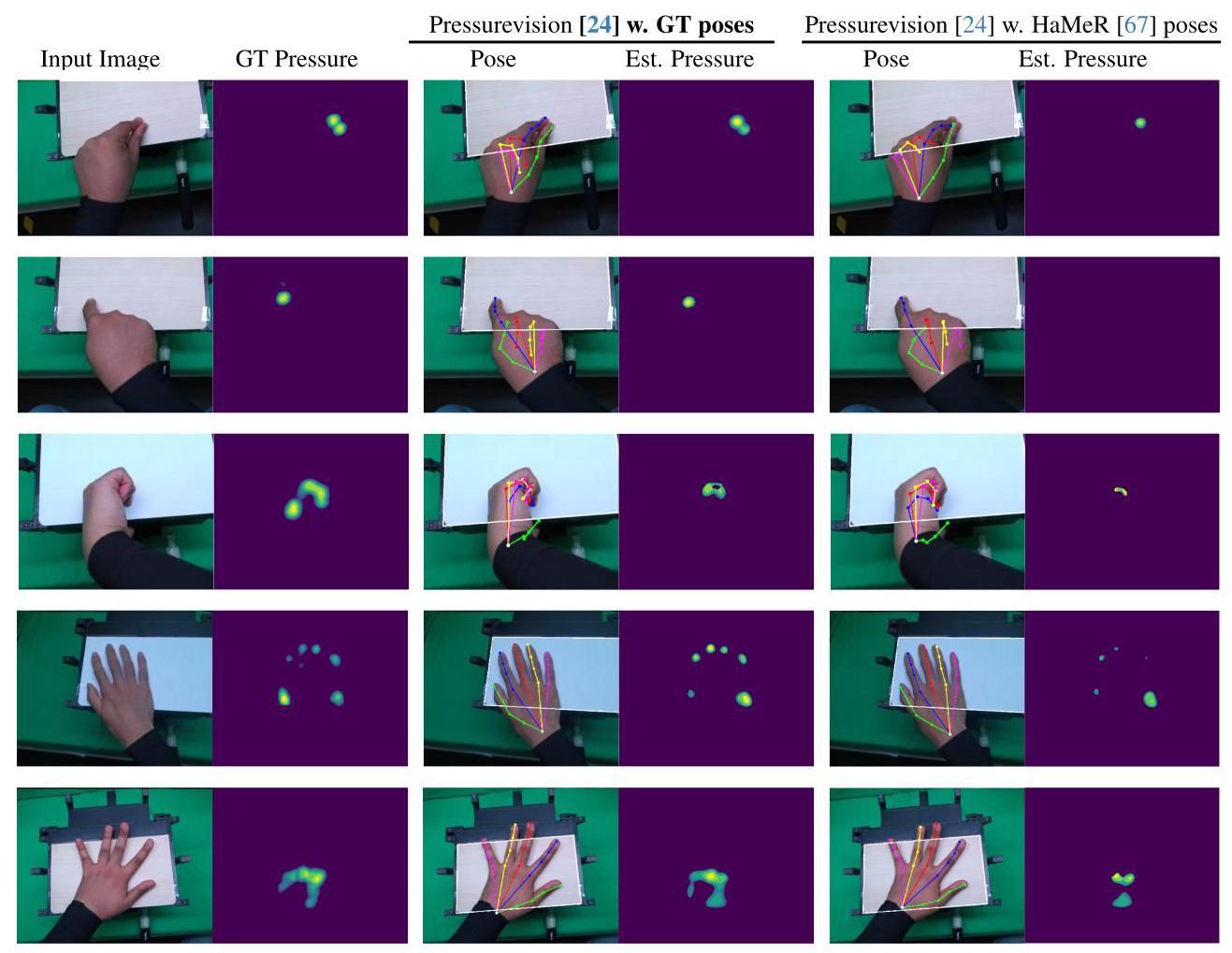


Figure 20. Qualitative results of the image-projected baselines on egocentric views, incorporating additional hand pose inputs using our annotations and predictions from HaMeR [73]. We also reproject the area of the touchpad (indicated by white lines) to verify the egocentric camera pose.

图20. 在自我中心视图中，结合我们注释和HaMeR [73] 预测的额外手部姿态输入的图像投影基线的定性结果。我们还重新投影了触摸板的区域(由白线指示)以验证自我中心相机的姿态。

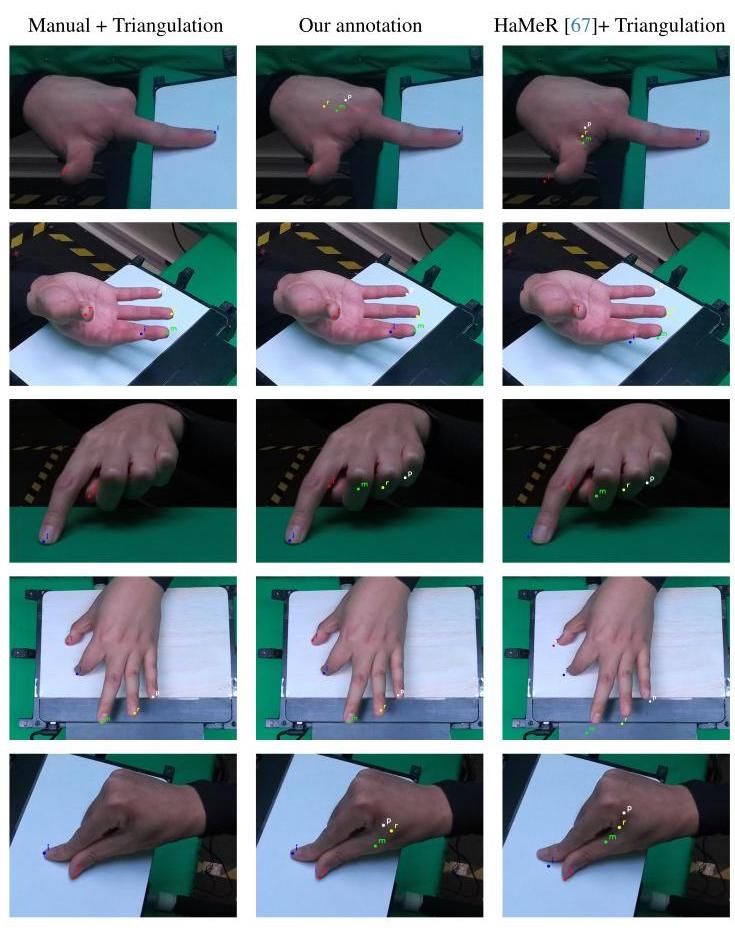


Figure 21. Qualitative comparison of reprojected nail tips from our annotation method (center ) and triangulation of HaMeR [67] predictions (right). The left column displays the reprojection of triangulated manually annotated visible tips.

图21. 我们注释方法(中心)和HaMeR [67] 预测的三角测量(右侧)重新投影的指甲尖的定性比较。左侧列显示了三角测量手动注释的可见尖端的重新投影。

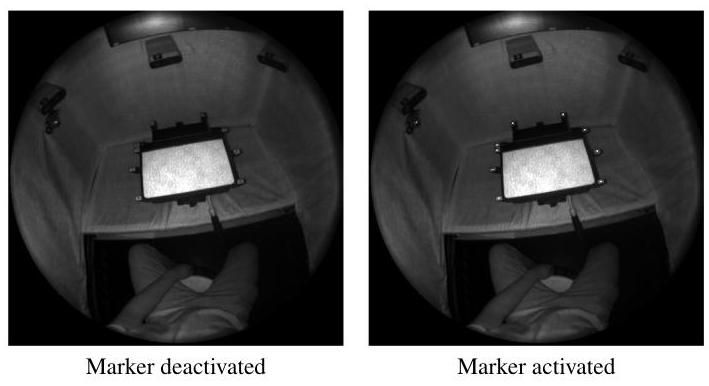


Figure 22. Marker visibility in infrared frame of head-mounted egocentric camera

图22. 头戴式自我中心相机的红外帧中的标记可见性

Table 9. List of gestures performed by a participant during the data collection.

表9. 数据收集过程中参与者执行的手势列表。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Left Hand | Right Hand | Number of Repetitions |
| i. | calibration routine | ✓ | ✓ | - |
| ii. | draw word | ✓ | ✓ | 3 |
| iii. | grasp edge curled thumb-down | ✓ | ✓ | 5 |
| iv. | grasp edge curled thumb-up | ✓ | ✓ | 5 |
| V. | grasp edge uncurled thumb-down | ✓ | ✓ | 5 |
| vi. | index press high force | ✓ | ✓ | 5 |
| vii. | index press low force | ✓ | ✓ | 5 |
| viii. | index press no-contact | ✓ | ✓ | 5 |
| ix. | index press pull | ✓ | ✓ | 5 |
| x. | index press push | ✓ | ✓ | 5 |
| xi. | index press rotate left | ✓ | ✓ | 5 |
| xii. | index press rotate right | ✓ | ✓ | 5 |
| xiii. | pinch thumb-down high force | ✓ | ✓ | 5 |
| Xiv. | pinch thumb-down low force | ✓ | ✓ | 5 |
| XV. | pinch thumb-down no-contact | ✓ | ✓ | 5 |
| xvi. | pinch zoom | ✓ | ✓ | 5 |
| xvii. | press cupped onebyone high force | ✓ | ✓ | 3 |
| xviii. | press cupped onebyone low force | ✓ | ✓ | 3 |
| xix. | press fingers high force | ✓ | ✓ | 5 |
| xx. | press fingers low force | ✓ | ✓ | 5 |
| xxi. | press fingers no-contact | ✓ | ✓ | 5 |
| xxii. | press flat onebyone high force | ✓ | ✓ | 3 |
| xxiii. | press flat onebyone low force | ✓ | ✓ | 3 |
| xxiv. | press palm high force | ✓ | ✓ | 5 |
| XXV. | press palm low force | ✓ | ✓ | 5 |
| xxvi. | press palm no-contact | ✓ | ✓ | 5 |
| xxvii. | press palm-and-fingers high force | ✓ | ✓ | 5 |
| xxviii. | press palm-and-fingers low force | ✓ | ✓ | 5 |
| xxix. | press palm-and-fingers no-contact | ✓ | ✓ | 5 |
| xxx. | pull towards | ✓ | ✓ | 5 |
| xxxi. | push away | ✓ | ✓ | 5 |
| xxxii. | touch iPad | ✓ | ✓ | 3 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | 左手 | 右手 | 重复次数 |
| i. | 校准程序 | ✓ | ✓ | - |
| ii. | 绘制单词 | ✓ | ✓ | 3 |
| iii. | 抓握边缘，拇指朝下弯曲 | ✓ | ✓ | 5 |
| iv. | 抓握边缘，拇指朝上弯曲 | ✓ | ✓ | 5 |
| V. | 抓握边缘，拇指朝下不弯曲 | ✓ | ✓ | 5 |
| vi. | 食指按压高力度 | ✓ | ✓ | 5 |
| vii. | 食指按压低力度 | ✓ | ✓ | 5 |
| viii. | 食指按压无接触 | ✓ | ✓ | 5 |
| ix. | 食指按压拉 | ✓ | ✓ | 5 |
| x. | 食指按压推 | ✓ | ✓ | 5 |
| xi. | 食指按压左旋转 | ✓ | ✓ | 5 |
| xii. | 食指按压右旋转 | ✓ | ✓ | 5 |
| xiii. | 捏合拇指朝下高力度 | ✓ | ✓ | 5 |
| Xiv. | 捏合拇指朝下低力度 | ✓ | ✓ | 5 |
| XV. | 捏合拇指朝下无接触 | ✓ | ✓ | 5 |
| xvi. | 捏合缩放 | ✓ | ✓ | 5 |
| xvii. | 按压杯状逐一高力度 | ✓ | ✓ | 3 |
| xviii. | 按压杯状逐一低力度 | ✓ | ✓ | 3 |
| xix. | 按压手指高力度 | ✓ | ✓ | 5 |
| xx. | 按压手指低力度 | ✓ | ✓ | 5 |
| xxi. | 按压手指无接触 | ✓ | ✓ | 5 |
| xxii. | 按压平放逐一高力度 | ✓ | ✓ | 3 |
| xxiii. | 按压平放逐一低力度 | ✓ | ✓ | 3 |
| xxiv. | 按压手掌高力度 | ✓ | ✓ | 5 |
| XXV. | 按压手掌低力度 | ✓ | ✓ | 5 |
| xxvi. | 按压手掌无接触 | ✓ | ✓ | 5 |
| xxvii. | 按压手掌和手指高力度 | ✓ | ✓ | 5 |
| xxviii. | 按压手掌和手指低力度 | ✓ | ✓ | 5 |
| xxix. | 按压手掌和手指无接触 | ✓ | ✓ | 5 |
| xxx. | 拉向 | ✓ | ✓ | 5 |
| xxxi. | 推开 | ✓ | ✓ | 5 |
| xxxii. | 触摸iPad | ✓ | ✓ | 3 |

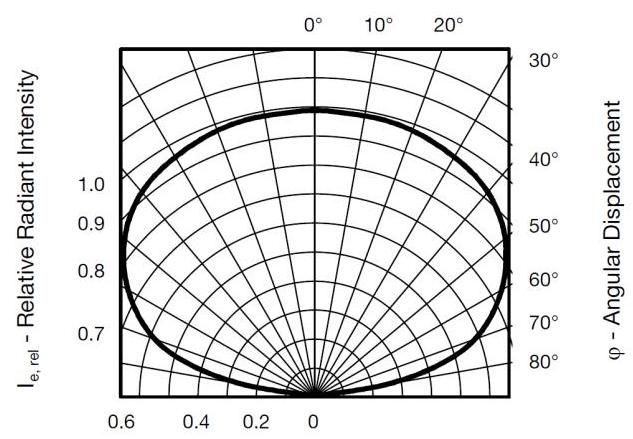


Figure 23. Relative Radiant Intensity vs. Angular Displacement The marker enable a good visible radiant intensity of beam angle till 150 degree, which ensures good visibility in egocentric infrared camera

图23. 相对辐射强度与角位移的关系 该标记在150度光束角度内具有良好的可见辐射强度，确保了在自我中心红外摄像头中的良好可见性。

# 9.4. Details about Devices’ Synchronization in Dataset Acquisition

# 9.4. 数据集采集中设备同步的详细信息

The Sensel Morph operates with zero buffer and maintains a stable delay at , whereas the Azure Kinect cameras function at , capturing high-resolution RGB images and a depth map. Due to the high recording performance of the Azure Kinect, frames are initially stored in the device’s cache, making it impractical to rely on the OS timestamp at the frame’s arrival on the host computer for synchronization with Sensel Morph pressure data.

Sensel Morph以零缓冲区运行，并在 保持稳定的 延迟，而Azure Kinect摄像头在 运行，捕捉高分辨率RGB图像和深度图。由于Azure Kinect的高记录性能，帧最初存储在设备的缓存中，因此依赖帧到达主机计算机时的操作系统时间戳来与Sensel Morph压力数据同步是不切实际的。

Table 10. Comparison between EgoPressure and extended list of hand-contact datasets.

表10. EgoPressure与扩展的手部接触数据集列表的比较。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | frames | participants | hand pose | hand mesh | markerless | real | egocentric | multiview | RGB | depth | contact | pressure surfacehand | |
| EgoPressure (ours) | 4.3M | 21 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Pressure sensor | ✓ | ✓ |
| ContactLabelDB [25] | 2.9M | 51 | ✘ | ✘ | ✓ | ✓ | ✘ | ✓ | ✓ | ✘ | Pressure sensor | ✓ | ✘ |
| PressureVisionDB [24] | 3.0M | 36 | ✘ | ✘ | ✓ | ✓ | ✘ | ✓ | ✓ | ✘ | Pressure sensor | ✓ | ✘ |
| ContactPose [4] | 3.0M | 50 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | Thermal imprint | ✘ | ✘ |
| GRAB [84] | 1.6M | 10 | ✓ | ✓ | ✘ | ✓ | ✘ | ✘ | ✘ | ✘ | Inferred from Pose | ✘ | ✘ |
| ARCTIC [20] | 2.1M | 10 | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |
| H2O [47] | 571k | 4 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |
| OakInk [92] | 230k | 12 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |
| OakInk-2 [94] | 4.01M | 9 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✘ | Inferred from Pose | ✘ | ✘ |
| DexYCB [7] | 582k | 10 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |
| HO-3D [30] | 103k | 10 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |
| TACO [56] | 5.2M | 14 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |
| Affordpose [42] | 26.7k | - | ✓ | ✓ | ✓ | ✘ | ✘ | ✓ | ✘ | ✘ | Inferred from Pose | ✘ | ✘ |
| AssemblyHands [66] | 3.03M | 34 | ✓ | ✘ | ✓ | ✓ | ✓ | ✓ | ✓ | ✘ | ✘ | ✘ | ✘ |
| ContactArt [96] | 332k | - | ✓ | ✓ | ✓ | ✘ | ✘ | ✓ | ✓ | ✓ | Simulated Pose | ✘ | ✘ |
| HOI4D [55] | 2.4M | 9 | ✓ | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | Inferred from Pose | ✘ | ✘ |
| YCBAfford [14] | 133k | - | ✓ | ✓ | ✓ | ✘ | ✘ | ✘ | ✘ | ✘ | Simulated Pose | ✘ | ✘ |
| ObMan [33] | 154k | - | ✓ | ✓ | ✓ | ✘ | ✘ | ✘ | ✓ | ✓ | Simulated Pose | ✘ | ✘ |
| FPHAB [21] | 100k | 6 | ✓ | ✘ | ✘ | ✓ | ✓ | ✘ | ✓ | ✓ | ✘ | ✘ | ✘ |
| HA-ViD [95] | 1.5M | 30 | ✘ | ✘ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | ✘ | ✘ | ✘ |
| Ego4d [26] | 3670 hours | 923 | ✘ | ✘ | ✓ | ✓ | ✓ | ✘ | ✓ | ✘ | ✘ | ✘ | ✘ |
| EPIC-KITCHEN-100 [15] | 20M | 37 | ✘ | ✘ | ✓ | ✓ | ✓ | ✘ | ✓ | ✘ | ✘ | ✘ | ✘ |
| Ego-Exo4D [27] | 1422 hours | 740 | ✓ | ✘ | ✓ | ✓ | ✓ | ✓ | ✓ | ✘ | ✘ | ✘ | ✘ |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 帧 | 参与者 | 手部姿态 | 手部网格 | 无标记 | 真实 | 自我中心 | 多视角 | RGB | 深度 | 接触 | 压力表面手 | |
| EgoPressure(我们的) | 4.3M | 21 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 压力传感器 | ✓ | ✓ |
| ContactLabelDB [25] | 2.9M | 51 | ✘ | ✘ | ✓ | ✓ | ✘ | ✓ | ✓ | ✘ | 压力传感器 | ✓ | ✘ |
| PressureVisionDB [24] | 3.0M | 36 | ✘ | ✘ | ✓ | ✓ | ✘ | ✓ | ✓ | ✘ | 压力传感器 | ✓ | ✘ |
| ContactPose [4] | 3.0M | 50 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | 热印 | ✘ | ✘ |
| GRAB [84] | 1.6M | 10 | ✓ | ✓ | ✘ | ✓ | ✘ | ✘ | ✘ | ✘ | 从姿态推断 | ✘ | ✘ |
| ARCTIC [20] | 2.1M | 10 | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |
| H2O [47] | 571k | 4 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |
| OakInk [92] | 230k | 12 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |
| OakInk-2 [94] | 4.01M | 9 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✘ | 从姿态推断 | ✘ | ✘ |
| DexYCB [7] | 582k | 10 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |
| HO-3D [30] | 103k | 10 | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |
| TACO [56] | 5.2M | 14 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |
| Affordpose [42] | 26.7k | - | ✓ | ✓ | ✓ | ✘ | ✘ | ✓ | ✘ | ✘ | 从姿态推断 | ✘ | ✘ |
| AssemblyHands [66] | 3.03M | 34 | ✓ | ✘ | ✓ | ✓ | ✓ | ✓ | ✓ | ✘ | ✘ | ✘ | ✘ |
| ContactArt [96] | 332k | - | ✓ | ✓ | ✓ | ✘ | ✘ | ✓ | ✓ | ✓ | 模拟姿态 | ✘ | ✘ |
| HOI4D [55] | 2.4M | 9 | ✓ | ✓ | ✓ | ✓ | ✓ | ✘ | ✓ | ✓ | 从姿态推断 | ✘ | ✘ |
| YCBAfford [14] | 133k | - | ✓ | ✓ | ✓ | ✘ | ✘ | ✘ | ✘ | ✘ | 模拟姿态 | ✘ | ✘ |
| ObMan [33] | 154k | - | ✓ | ✓ | ✓ | ✘ | ✘ | ✘ | ✓ | ✓ | 模拟姿态 | ✘ | ✘ |
| FPHAB [21] | 100k | 6 | ✓ | ✘ | ✘ | ✓ | ✓ | ✘ | ✓ | ✓ | ✘ | ✘ | ✘ |
| HA-ViD [95] | 1.5M | 30 | ✘ | ✘ | ✓ | ✓ | ✘ | ✓ | ✓ | ✓ | ✘ | ✘ | ✘ |
| Ego4d [26] | 3670小时 | 923 | ✘ | ✘ | ✓ | ✓ | ✓ | ✘ | ✓ | ✘ | ✘ | ✘ | ✘ |
| EPIC-KITCHEN-100 [15] | 20M | 37 | ✘ | ✘ | ✓ | ✓ | ✓ | ✘ | ✓ | ✘ | ✘ | ✘ | ✘ |
| Ego-Exo4D [27] | 1422小时 | 740 | ✓ | ✘ | ✓ | ✓ | ✓ | ✓ | ✓ | ✘ | ✘ | ✘ | ✘ |

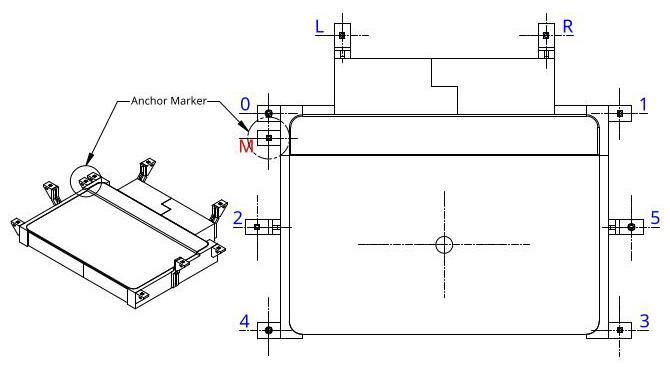


Figure 24. Layout of the Active Markers. The indices of the markers are aligned with the pseudocode provided in Algorithm 1. Starting from the asymmetrical anchor marker , all markers can be identified by computing their relative distances and considering their spatial relationships.

图24. 活动标记的布局。标记的索引与算法1中提供的伪代码对齐。从非对称锚点标记 开始，所有标记都可以通过计算它们的相对距离并考虑它们的空间关系来识别。

All cameras can be externally synchronized via a triggering signal from the Raspberry Pi CM4, ensuring simultaneous frame capture. However, an initial frame loss (1-3 frames) occurs at the start of recording due to device-specific issues. Since the absolute value of device ticks has no inherent meaning, it is unclear how many frames were lost before the first received frame. Relying on device tick differences for synchronization could therefore introduce a misalignment of 1-3 frames between cameras.

所有相机都可以通过树莓派CM4的 触发信号进行外部同步，确保同时进行帧捕获。然而，由于设备特定的问题，在开始录制时会出现初始帧丢失(1-3帧)。由于设备滴答的绝对值没有内在意义，因此不清楚在接收到第一帧之前丢失了多少帧。因此，依赖设备滴答差异进行同步可能会导致相机之间的1-3帧错位。

To address this, the programmable features of active infrared markers and the precise global OS timestamp synchronization (within ) between the two host computers and the Raspberry Pi CM4, facilitated by the Precision Time Protocol (PTP), are utilized. The Raspberry Pi CM4, equipped with basic electrical components (see Fig.22) at the start of the next exposure cycle, providing a reliable synchronization point that compensates for the initial missing frames. The exact global OS timestamp of the marker activation is clearly recorded (see Fig. 26).

为了解决这个问题，利用了活动红外标记的可编程功能以及两台主机和树莓派CM4之间通过精确时间协议(PTP)实现的精确全局操作系统时间戳同步(在 内)。树莓派CM4在下一个曝光周期开始时配备了基本电气元件(见图22)，提供了一个可靠的同步点，补偿了初始丢失的帧。标记激活的确切全局操作系统时间戳被清晰地记录下来(见图26)。

Algorithm 1 Identify Marker

算法1 识别标记

procedure IDENTIFYMARKERS(filtered IR image)

Extract marker coordinates(u, v)from the filtered IR image

Compute all pairwise distances among markers

Identify the pair with the smallest distance, initially labeled as 0

and

Compute the vector from to 0

Count the number of markers on each side of the vector line

if more markers lie on the right of the vector then

Confirm start point as , endpoint as 0

else

Swap, set start point as 0 and endpoint as

end if

Identify 2 and 4 as markers aligned with , on the same side

relative to

Check distances from to 2 and 4 to determine which is closer

Identify as the marker closest to the line extending through

and on the same side as 0

Compute the centroid of all markers

Draw a line from 0 through the centroid

Identify 3 as the marker isolated on its side of the centroid line

Identify 5 as the closest marker to the line (0-centroid) not already

labeled

Determine 1 and by their proximity to line(2 - 5), with 1 being

closer

end procedure

By calculating the real OS timestamp for all frames based on the offset from device ticks, starting from the frame where the marker first appears, precise synchronization is achieved. This approach effectively aligns RGBD images and pressure data, optimizing data integration across the multi-modal sensor system. Moreover, this synchronization mechanism using an external active optical identifier is efficient and economical, making it generalizable to other multi-sensor systems, such as motion capture systems with external head-mounted cameras, that rely on different OS timestamp sources.

通过从标记首次出现的帧开始，基于设备滴答的偏移量计算所有帧的实际操作系统时间戳，实现了精确的同步。这种方法有效地对齐了RGBD图像和压力数据，优化了多模态传感器系统的数据集成。此外，这种使用外部活动光学标识符的同步机制高效且经济，使其可推广到其他多传感器系统，例如依赖不同操作系统时间戳源的外部头戴式相机运动捕捉系统。

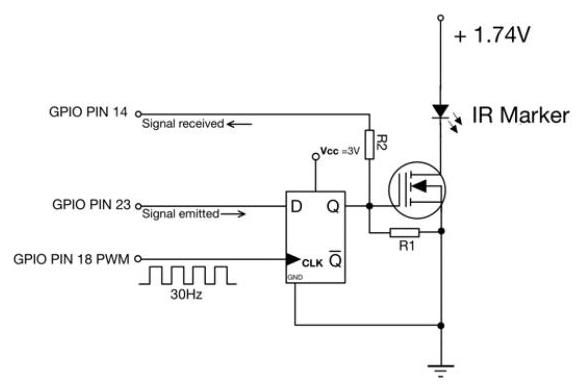


Figure 25. Basic Electrical Elements Implementation, We use a D-type flip-flop and N-channel MOSFET to ensure the IR marker will be activate by the next beginning of exposure after receiving signal from PIN 23. And PIN 14 will monitor the activation to obtain its timestamp

图25. 基本电气元件实现，我们使用D型触发器和N沟道MOSFET确保红外标记在接收到PIN 23的信号后，在下一个曝光开始时激活。PIN 14将监控激活以获取其时间戳。

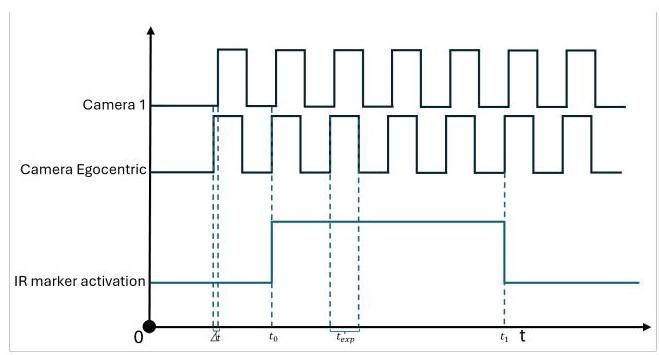


Figure 26. Synchronization Diagram We set head-mounted egocentric camera to align with triggering signal emitted by the Raspberry Pi CM4, this signal will also go to PIN 18 as clock frequency of D-type Flip-flop Fig. 25). Then exposure of all cameras is same. The other static cameras 1 to 7 will have a delay to triggering signal to avoid interference of infrared light. The marker will be activate at (around 300 milliseconds after start recording), which we know its global OS timestamp, then it will be visible to all camera at next exposure cycle. As verification, we deactivate marker by the very end of recording at the timestamp , then the marker will be invisible for all cameras in the next frame capture. The good synchronization will have equal frame number between and for all cameras.

图26. 同步图，我们将头戴式自我中心相机设置为与树莓派CM4发出的 触发信号对齐，该信号也将作为D型触发器的时钟频率发送到PIN 18(见图25)。然后所有相机的曝光 相同。其他静态相机1到7将有一个延迟 到触发信号，以避免红外光的干扰。标记将在 (开始录制后约300毫秒)激活，我们知道其全局操作系统时间戳，然后它将在下一个曝光周期对所有相机可见。作为验证，我们在录制结束时的 时间戳停用标记，然后标记将在下一帧捕获中对所有相机不可见。良好的同步将使所有相机在 和 之间具有相同的帧数。

# 10. Limitations

# 10. 限制

Although EgoPressure serves as a foundational study for understanding pressure from an egocentric view, several challenges remain unresolved. These challenges are categorized into three main areas.

尽管EgoPressure作为理解自我中心视角下的压力的基础研究，但仍有一些挑战未解决。这些挑战主要分为三个领域。

First, measuring pressure while interacting with general objects presents a challenge. Our current data capture is confined to sensing pressure on flat surfaces. While we are optimistic that future research will expand to include a wider variety of objects, sensing pressure on arbitrary surfaces poses significant challenges, as it would require extensive instrumentation of the user’s hands, hindering natural interaction and introducing visible artifacts in the captured data. Instrumenting objects for pressure sensing remains an ongoing research area, with recent advancements primarily in basic contact detection [4]. However, we anticipate that our annotation method will extend naturally to more complex objects and interactions as these challenges are addressed. PressureVision++ [25] explores weak labels to infer pressure on more complex objects. However, it only considers fingertip interactions and its evaluation of pressure regression remains limited to flat surfaces due to the challenges of acquiring precise pressure. We present a qualitative evaluation of PressureFormer on a wider variety of objects in Figure 15.

首先，在与一般物体交互时测量压力是一个挑战。我们目前的数据捕获仅限于在平面上感知压力。虽然我们乐观地认为未来的研究将扩展到包括更多种类的物体，但在任意表面上感知压力提出了重大挑战，因为这需要对用户的手进行广泛的仪器化，阻碍自然交互并在捕获的数据中引入可见的伪影。为压力传感仪器化物体仍然是一个正在进行的研究领域，最近的进展主要集中在基本接触检测[4]。然而，我们预计随着这些挑战的解决，我们的注释方法将自然地扩展到更复杂的物体和交互。PressureVision++ [25]探索了弱标签以推断更复杂物体上的压力。然而，它仅考虑指尖交互，并且由于获取精确压力的挑战，其压力回归评估仍然局限于平面。我们在图15中对PressureFormer在更广泛物体上进行了定性评估。

Second, the current dataset was only captured in an indoor setting. Our data capture setup is optimized for acquiring high-fidelity annotations of hand-surface interactions. To increase the diversity of background environments to improve generalization to real-world settings, we have added green overlays to the background of our data capture rig and to the pressure pad. This allows for background replacement and has been successfully demonstrated to enhance commercial in-the-wild hand tracking [32, 97].

其次，当前数据集仅在室内环境中捕获。我们的数据捕获设置优化了手表面交互的高保真注释。为了增加背景环境的多样性以提高对现实世界设置的泛化能力，我们在数据捕获装置和压力垫的背景上添加了绿色覆盖层。这允许背景替换，并已成功证明可以增强商业野外手部跟踪[32, 97]。

Finally, the current setup only considers single-hand interactions. Incorporating scenarios involving the use of both hands would be a natural extension of our work.

最后，当前设置仅考虑单手交互。纳入涉及双手使用的场景将是我们工作的自然延伸。

Further addressing these challenges in future research would improve pressure estimation in real-world scenarios and broaden its applicability.

在未来的研究中进一步解决这些挑战将提高现实场景中的压力估计，并扩大其适用性。

# 11. Ethical Considerations

# 11. 伦理考虑

The recording and use of human activity data involve important ethical considerations. The EgoPressure project has received approval from ETH Zürich Ethics Commission as proposal EK 2023-N-228. This approval includes both the data collection and the public release of the dataset. All participants provided explicit written consent for recording their sessions, creating the dataset, and releasing it (see accompanying consent form). All demographic information (such as sex, age, weight, and height) along with the sensor and video data are pseudonymized, assigning a numeric code to each participant. Personal data (sex, age, weight, and height) is stored separately from the sensor and video data, and is accessible only to the primary researchers involved in the study. We have not captured or stored any images of the participant’s face.

人类活动数据的记录和使用涉及重要的伦理考虑。EgoPressure项目已获得苏黎世联邦理工学院伦理委员会的批准，批准号为EK 2023-N-228。该批准包括数据收集和数据集的公开发布。所有参与者都提供了明确的书面同意，同意记录他们的会话、创建数据集并发布数据集(见附带的同意书)。所有人口统计信息(如性别、年龄、体重和身高)以及传感器和视频数据都进行了匿名化处理，为每个参与者分配了一个数字代码。个人数据(性别、年龄、体重和身高)与传感器和视频数据分开存储，仅限参与研究的主要研究人员访问。我们没有捕捉或存储任何参与者面部的图像。

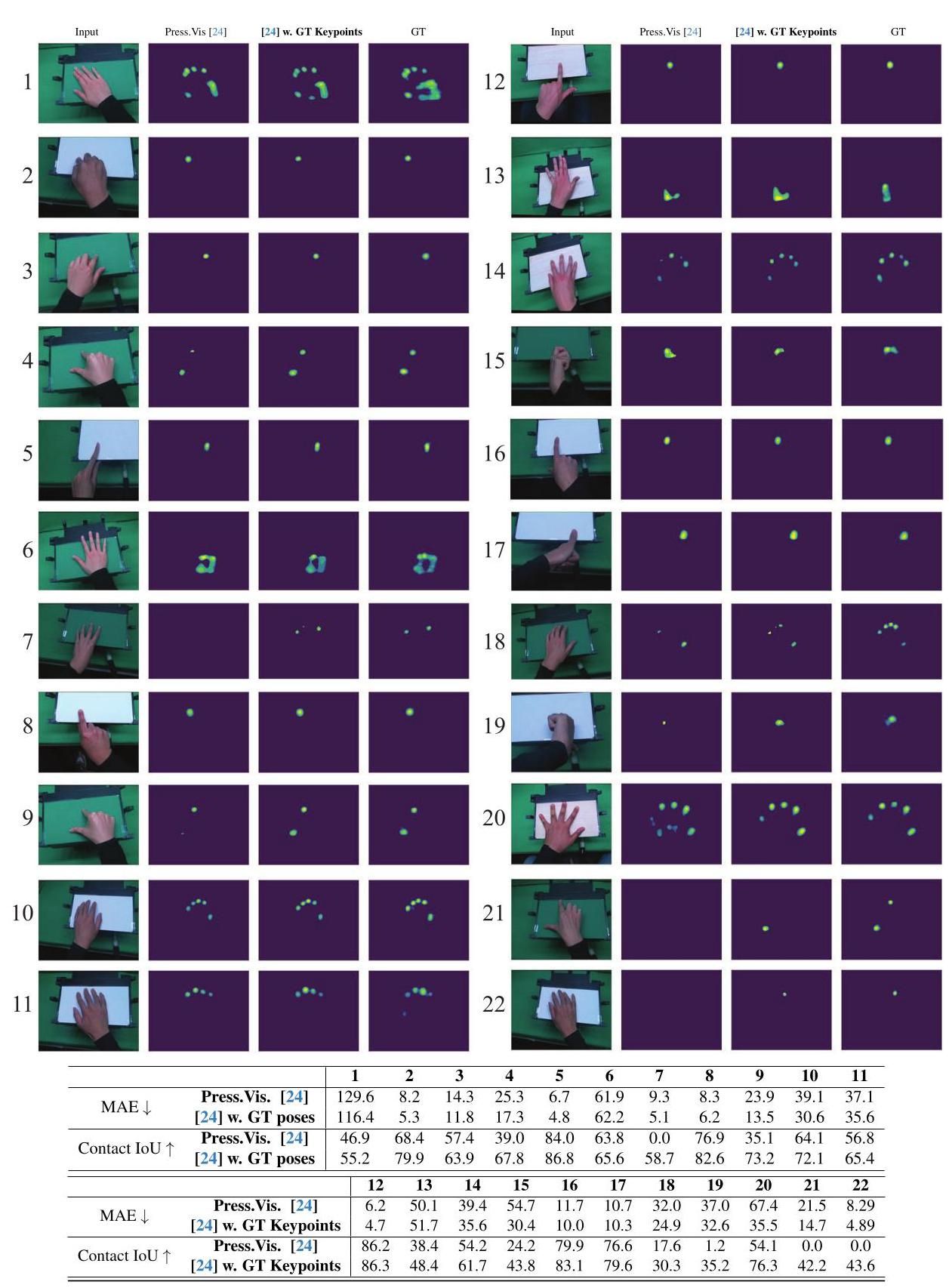


Figure 27. Qualitative comparison of pressure maps inferred using PressureVisionNet [24] and our trained model with additional hand poses as input on representative cases across various gestures. The bottom table presents MAE [Pa] and Contact IoU [%] for pressure maps inferred using PressureVisionNet [1] and our trained model on selected samples shown in the Figure.

图27. 使用PressureVisionNet [24]和我们训练的模型推断的压力图的定性比较，输入包括额外的手势姿势，涵盖各种手势的代表性案例。底部的表格展示了使用PressureVisionNet [1]和我们训练的模型推断的压力图的MAE [Pa]和接触IoU [%]，所选样本如图中所示。

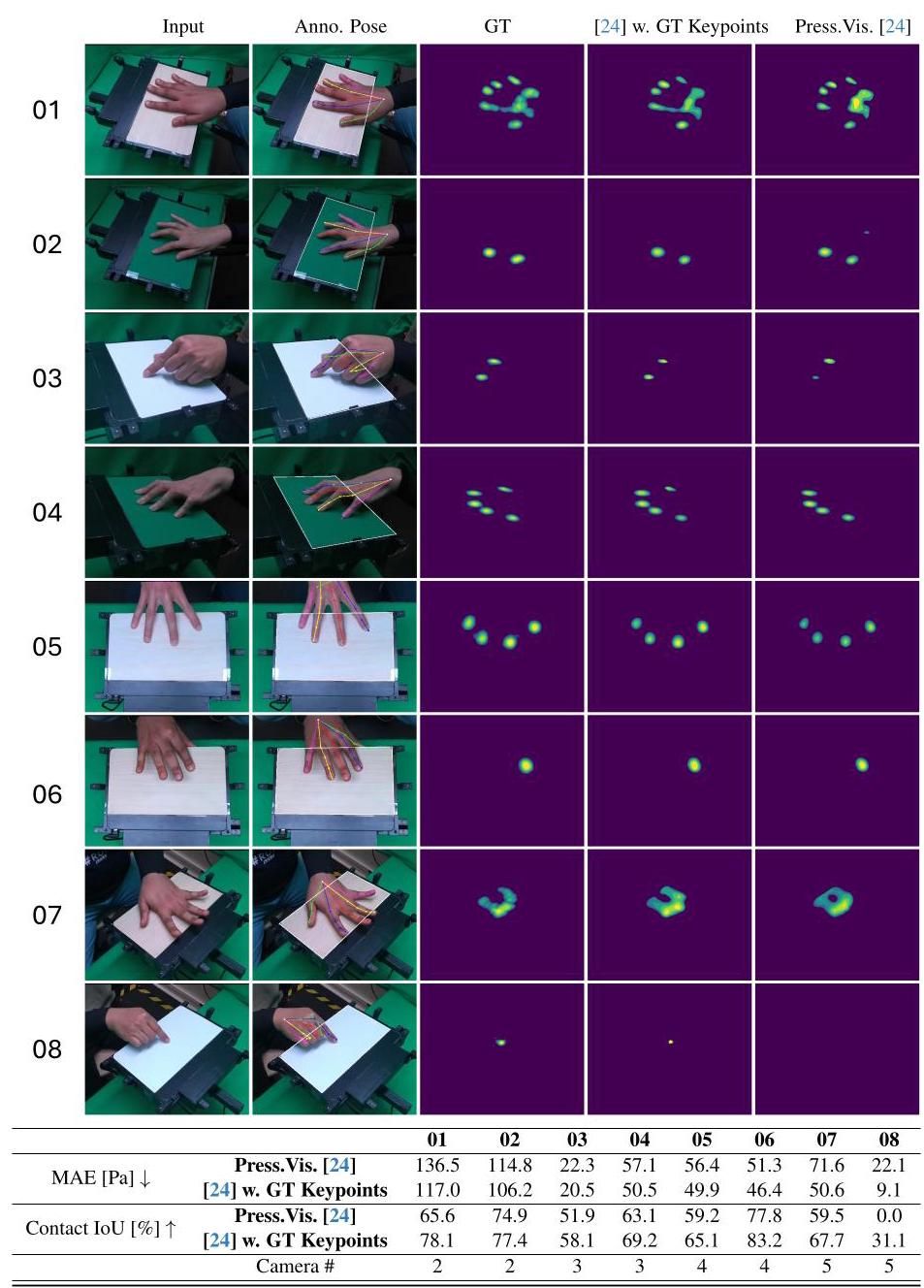


Figure 28. Comparison of pressure maps estimated by PressureVisionNet [24] and our adapted model, using separate training and validation sets, both consisting of images from camera views 2,3,4, and 5 .

图28. 使用单独的训练和验证集，由PressureVisionNet [24]和我们调整的模型估计的压力图比较，两组数据均来自摄像头视角2、3、4和5。

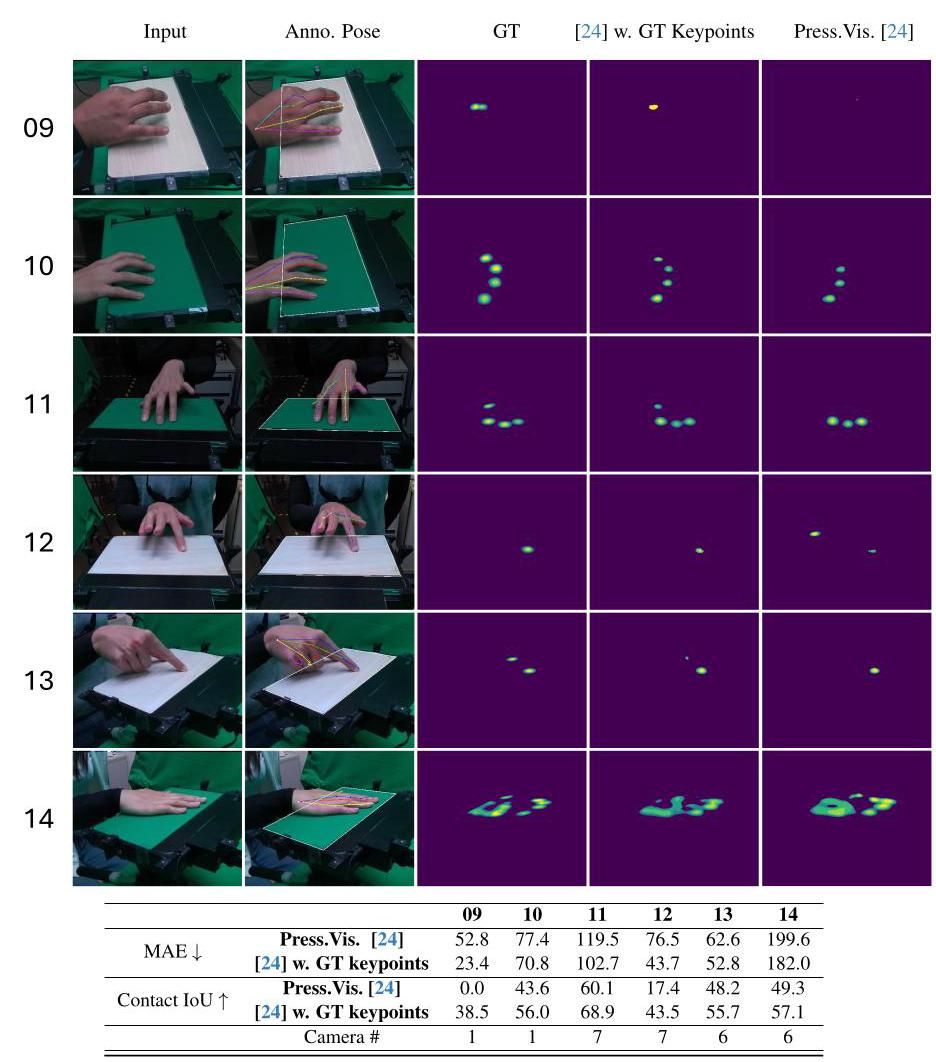


Figure 29. Comparison of pressure maps estimated by PressureVisionNet [24] and our adapted model, evaluated using input images from cameras 1,6, and 7. The models are the same as in Figure 28, which are trained on images from camera views 2, 3, 4, and 5.

图29. 使用来自摄像头1、6和7的输入图像评估的PressureVisionNet [24]和我们调整的模型估计的压力图比较。模型与图28相同，训练数据来自摄像头视角2、3、4和5。

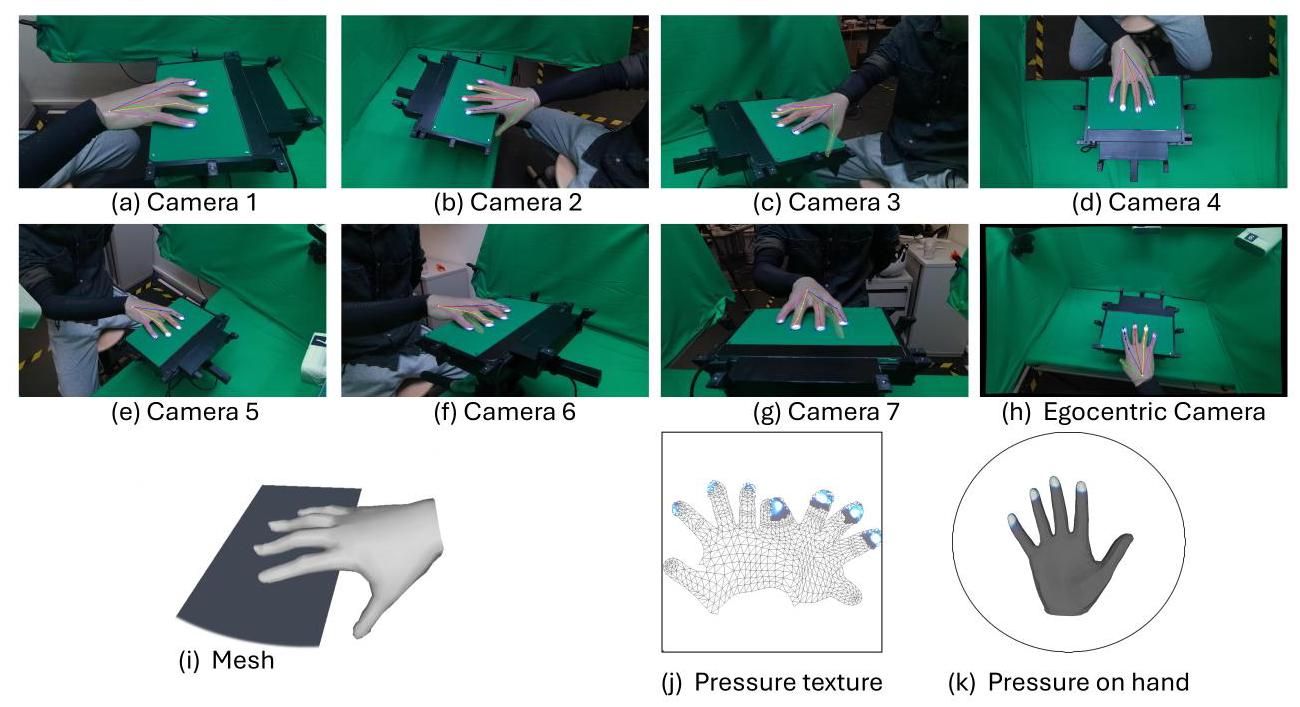


Figure 30. Example of Annotation 1 Right hand with gesture: grasp edge with uncurled thumb down

图30. 注释1示例 右手手势:拇指未弯曲向下抓握边缘

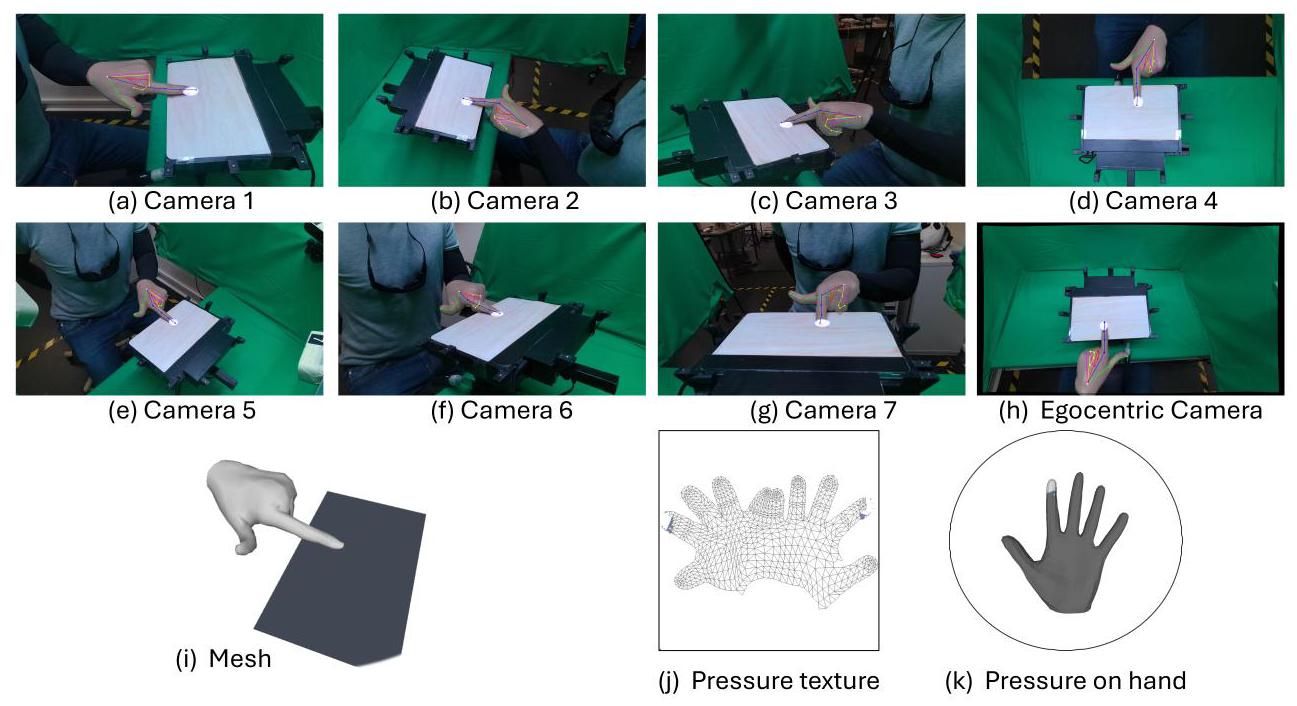


Figure 31. Example of Annotation 2 Left hand with gesture: index press with high force

图31. 注释2示例 左手手势:食指高力按压

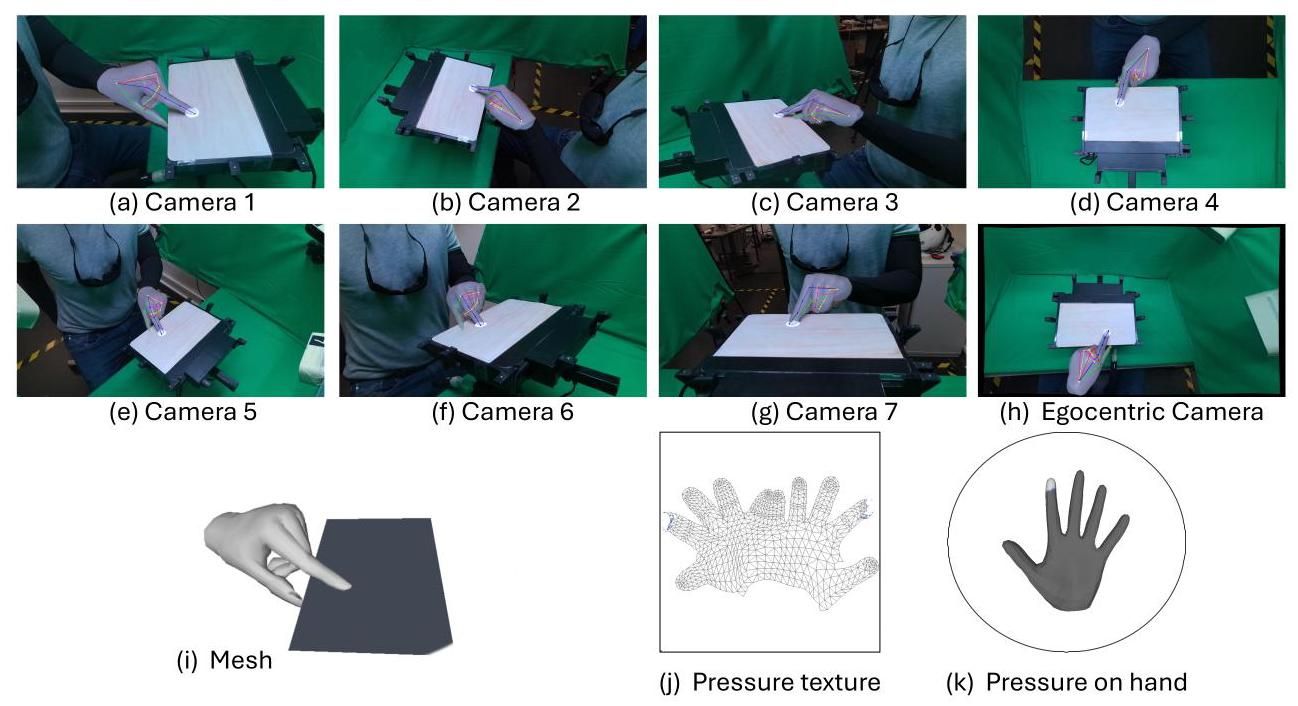


Figure 32. Example of Annotation 3 Left hand with gesture: pinch thumb down on the edge with high force

图32. 注释3示例 左手手势:拇指高力向下捏边缘

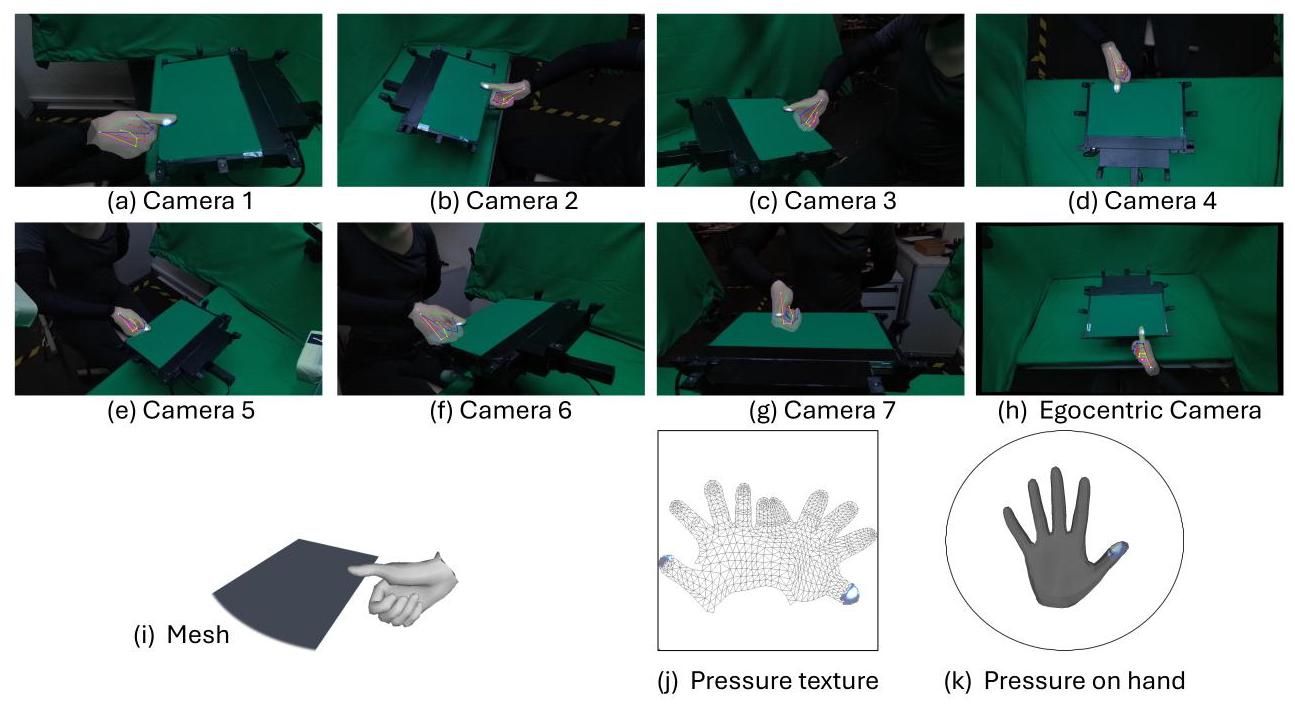


Figure 33. Example of Annotation 4 Right hand with gesture: grasp edge with curled thumb up

图33. 注释4示例 右手手势:拇指弯曲向上抓握边缘

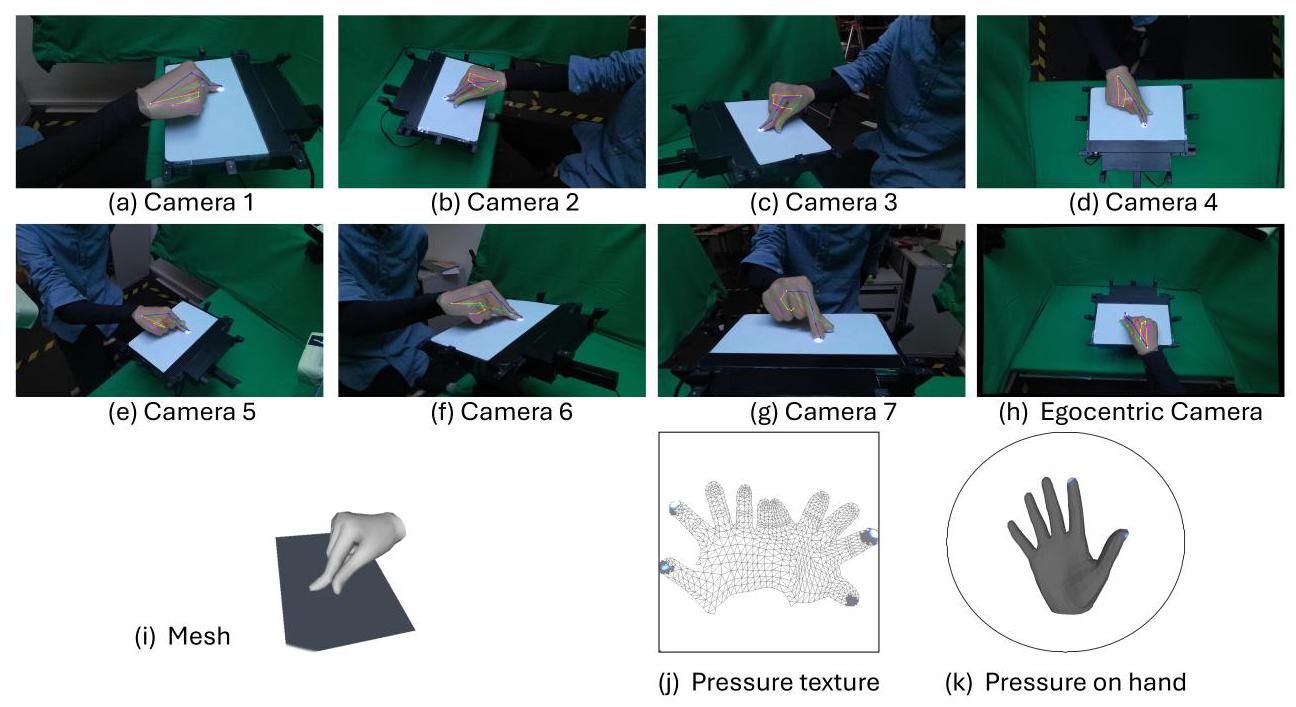


Figure 34. Example of Annotation 5 Right hand with gesture: pinch finger zoom in and out

图34. 注释5示例 右手手势:捏指缩放

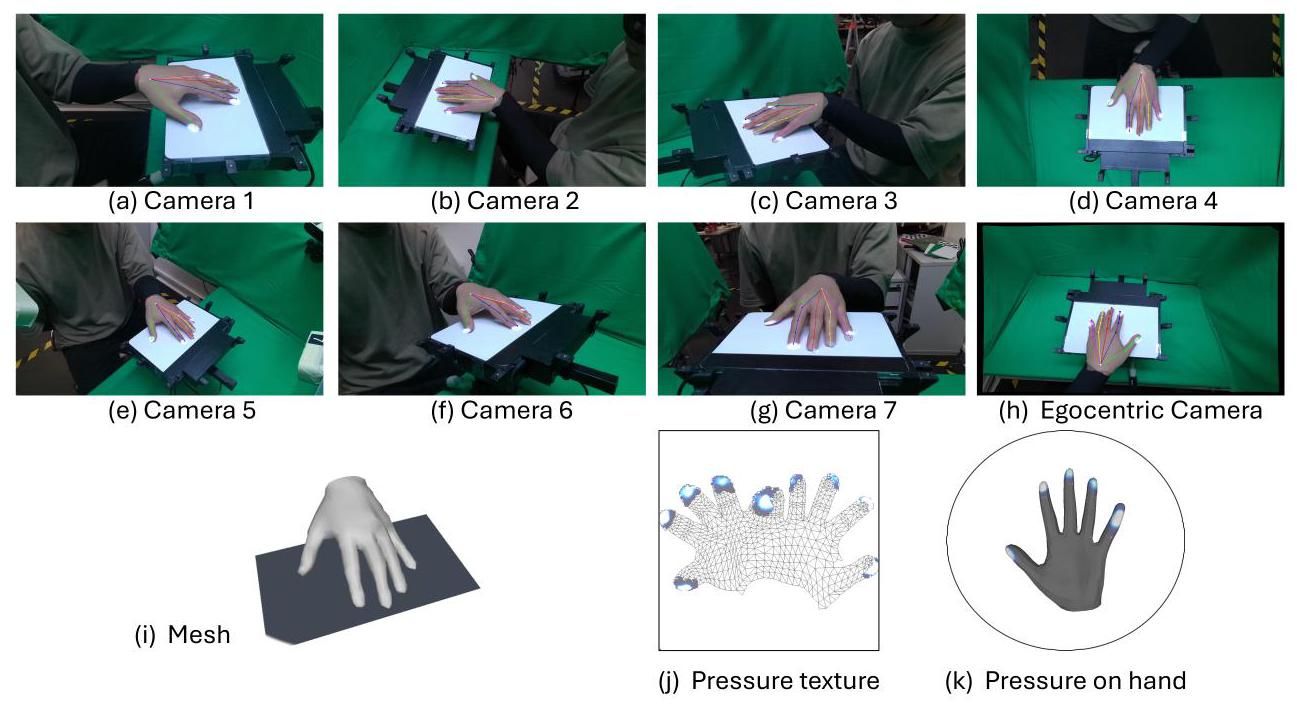


Figure 35. Example of Annotation 6 Left hand with gesture: pull all finger towards participant

图35. 注释6示例 左手手势:所有手指向参与者拉动

References

参考文献

[1] Azure Kinect. Azure kinect dk hardware specifications, 2019. 5

Azure Kinect. Azure kinect dk硬件规格, 2019. 5

[2] Raunaq Bhirangi, Tess Hellebrekers, Carmel Majidi, and Abhinav Gupta. Reskin: versatile, replaceable, lasting tactile skins. In 5th Annual Conference on Robot Learning, 2021. 3

Raunaq Bhirangi, Tess Hellebrekers, Carmel Majidi, 和 Abhinav Gupta. Reskin: 多功能、可更换、持久的触觉皮肤. 在第五届机器人学习年会上, 2021. 3

[3] Giulia Boato, Nicola Conci, Mattia Daldoss, Francesco GB De Natale, and Nicola Piotto. Hand tracking and trajectory analysis for physical rehabilitation. In 2009 IEEE International Workshop on Multimedia Signal Processing, pages 1-6. IEEE, 2009. 2

Giulia Boato, Nicola Conci, Mattia Daldoss, Francesco GB De Natale, 和 Nicola Piotto. 手部追踪和轨迹分析用于物理康复. 在2009年IEEE多媒体信号处理国际研讨会上, 第1-6页. IEEE, 2009. 2

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIII 16, pages 361-378. Springer, 2020. 3, 9, 10

萨马特·布拉姆巴特、程成唐、克里斯托弗·D·特威格、查尔斯·C·肯普和詹姆斯·海斯。Contactpose:一个包含物体接触和手部姿势的抓取数据集。在《计算机视觉-ECCV 2020:第16届欧洲会议，英国格拉斯哥，2020年8月23-28日，会议录，第十三部分16》中，第361-378页。斯普林格，2020年。3, 9, 10

[5] Gereon H Büscher, Risto Köiva, Carsten Schürmann, Robert Haschke, and Helge J Ritter. Flexible and stretchable fabric-based tactile sensor. Robotics and Autonomous Systems, 63: 244-252, 2015. 3

格雷昂·H·比舍尔、里斯托·科伊瓦、卡斯滕·舒尔曼、罗伯特·哈施克和赫尔格·J·里特。基于织物的柔性和可拉伸触觉传感器。《机器人与自主系统》，63卷:244-252页，2015年。3

[6] Jongeun Cha, Seung-man Kim, Ian Oakley, Jeha Ryu, and Kwan H. Lee. Haptic interaction with depth video media. In Advances in Multimedia Information Processing - PCM 2005, pages 420-430, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. 2

车钟根、金成万、伊恩·奥克利、柳济河和李宽赫。与深度视频媒体的触觉交互。在《多媒体信息处理进展 - PCM 2005》中，第420-430页，德国柏林、海德堡，2005年。斯普林格柏林海德堡。2

[7] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9044-9053, 2021. 2, 3, 9

赵宇伟、杨伟、向宇、帕夫洛·莫尔恰诺夫、安库尔·汉达、乔纳森·特伦布莱、亚什拉吉·S·纳兰、卡尔·范·维克、乌马尔·伊克巴尔、斯坦·伯奇菲尔德等。Dexycb:一个用于捕捉手部抓取物体的基准。在《IEEE/CVF计算机视觉与模式识别会议论文集》中，第9044-9053页，2021年。2, 3, 9

[8] Nutan Chen, Göran Westling, Benoni B Edin, and Patrick van der Smagt. Estimating fingertip forces, torques, and local curvatures from fingernail images. Robotica, 38(7):1242- 1262, 2020. 3

陈努坦、戈兰·韦斯特林、贝诺尼·B·埃丁和帕特里克·范·德·斯马特。从指甲图像估计指尖力、扭矩和局部曲率。《机器人学》，38卷7期:1242-1262页，2020年。3

[9] Wenzheng Chen, Jun Gao, Huan Ling, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In Advances In Neural Information Processing Systems, 2019.3, 4, 8

陈文正、高俊、凌欢、爱德华·史密斯、雅科·莱赫蒂宁、亚历克·雅各布森和桑贾·菲德勒。学习使用基于插值的可微分渲染器预测3D物体。在《神经信息处理系统进展》中，2019年。3, 4, 8

[10] Yi Fei Cheng, Tiffany Luong, Andreas Rene Fender, Paul Streli, and Christian Holz. Comfortable user interfaces: Surfaces reduce input error, time, and exertion for tabletop and mid-air user interfaces. In 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 2022. 1,2

程一飞、蒂芙尼·卢昂、安德烈亚斯·雷内·芬德、保罗·斯特雷利和克里斯蒂安·霍尔兹。舒适的用户界面:表面减少了桌面和空中用户界面的输入错误、时间和努力。在《2022年IEEE混合与增强现实国际研讨会(ISMAR)》中。IEEE，2022年。1,2

[11] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20577-20586, 2022.1,3

萨米·克里斯滕、穆罕默德·科卡巴斯、埃姆雷·阿克桑、杰敏·黄博、宋杰和奥特马尔·希利格斯。D-grasp:用于手-物体交互的物理上合理的动态抓取合成。在《IEEE/CVF计算机视觉与模式识别会议论文集》中，第20577-20586页，2022年。1,3

[12] Sammy Christen, Lan Feng, Wei Yang, Yu-Wei Chao, Otmar Hilliges, and Jie Song. Synh2r: Synthesizing hand-object motions for learning human-to-robot handovers. arXiv preprint arXiv:2311.05599, 2023. 2

萨米·克里斯滕、冯兰、杨伟、赵宇伟、奥特马尔·希利格斯和宋杰。Synh2r:合成手-物体运动以学习人-机器人交接。arXiv预印本arXiv:2311.05599，2023年。2

[13] Jeremy A Collins, Cody Houff, Patrick Grady, and Charles C Kemp. Visual contact pressure estimation for grippers in the wild. In 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 10947-10954. IEEE, 2023. 1

杰里米·A·柯林斯、科迪·霍夫、帕特里克·格雷迪和查尔斯·C·肯普。野外夹持器的视觉接触压力估计。在《2023年IEEE/RSJ智能机器人与系统国际会议(IROS)》中，第10947-10954页。IEEE，2023年。1

[14] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5031-5041, 2020. 3, 9

恩里克·科罗纳、阿尔伯特·普马罗拉、吉列姆·阿莱尼亚、弗朗西斯科·莫雷诺-诺格尔和格雷戈里·罗杰兹。Ganhand:预测多物体场景中的人类抓取能力。在《IEEE/CVF计算机视觉与模式识别会议论文集》中，第5031-5041页，2020年。3, 9

[15] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Molti-santi, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision, pages 1-23, 2022. 9

迪马·达门、黑兹尔·多尔蒂、乔瓦尼·玛丽亚·法里内拉、安东尼诺·弗纳里、埃万杰洛斯·卡扎科斯、简·马、达维德·莫尔蒂桑蒂、乔纳森·芒罗、托比·佩雷特、威尔·普莱斯等。重新缩放自我中心视觉:EPIC-KITCHENS-100的收集、管道和挑战。《国际计算机视觉杂志》，第1-23页，2022年。9

[16] Mathieu Desbrun, Mark Meyer, Peter Schröder, and Alan H Barr. Implicit fairing of irregular meshes using diffusion and curvature flow. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pages 317- 324, 1999. 4

马修·德布伦、马克·迈耶、彼得·施罗德和艾伦·H·巴尔。使用扩散和曲率流对不规则网格进行隐式平滑处理。在《第26届计算机图形与交互技术年会论文集》中，第317-324页，1999年。4

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-vain Gelly, et al. An image is worth words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 7

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-vain Gelly等。一张图片值 个词:大规模图像识别的Transformer。arXiv预印本 arXiv:2010.11929, 2020. 7

[18] Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, and Abhinav Gupta. Use the force, luke! learning to predict physical forces by simulating effects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 224-233, 2020. 3

Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, 和 Abhinav Gupta。使用原力，卢克！通过模拟效果学习预测物理力。在IEEE/CVF计算机视觉与模式识别会议论文集, 第224-233页, 2020. 3

[19] Neil Xu Fan and Robert Xiao. Reducing the latency of touch tracking on ad-hoc surfaces. Proc. ACM Hum.-Comput. Interact., 6(ISS), 2022. 2

Neil Xu Fan 和 Robert Xiao。减少在临时表面上的触摸跟踪延迟。Proc. ACM Hum.-Comput. Interact., 6(ISS), 2022. 2

[20] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12943-12954, 2023. 2, 3, 9

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, 和 Otmar Hilliges。Arctic:用于灵巧双手操作的物体数据集。在IEEE/CVF计算机视觉与模式识别会议论文集, 第12943-12954页, 2023. 2, 3, 9

[21] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 409-419, 2018. 2, 9

Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, 和 Tae-Kyun Kim。带有RGB-D视频和3D手部姿势注释的第一人称手部动作基准。在IEEE计算机视觉与模式识别会议论文集, 第409-419页, 2018. 2, 9

[22] Jun Gong, Aakar Gupta, and Hrvoje Benko. Acustico: Surface tap detection and localization using wrist-based acoustic tdoa sensing. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, page 406-419, New York, NY, USA, 2020. Association for Computing Machinery. 2

Jun Gong, Aakar Gupta, 和 Hrvoje Benko。Acustico:使用手腕声学TDOA传感进行表面点击检测和定位。在第33届ACM用户界面软件与技术研讨会论文集, 第406-419页, 纽约, 美国, 2020. 计算机械协会. 2

[23] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1471-1481, 2021. 2

Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, 和 Charles C Kemp。Contactopt:优化接触以改善抓握。在IEEE/CVF计算机视觉与模式识别会议论文集, 第1471-1481页, 2021. 2

[24] Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D Twigg, Chengde Wan, James Hays, and Charles C Kemp. Pressurevision: Estimating hand pressure from a single rgb image. In European Conference on Computer Vision, pages 328-345. Springer, 2022. 2, 3, 7, 8, 1, 9, 11, 12, 13

Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D Twigg, Chengde Wan, James Hays, 和 Charles C Kemp。Pressurevision:从单一RGB图像估计手部压力。在欧洲计算机视觉会议, 第328-345页. Springer, 2022. 2, 3, 7, 8, 1, 9, 11, 12, 13

[25] Patrick Grady, Jeremy A Collins, Chengcheng Tang, Christopher D Twigg, Kunal Aneja, James Hays, and Charles C Kemp. Pressurevision++: Estimating fingertip pressure from diverse rgb images. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 8698-8708, 2024. 1, 2, 3, 7, 9, 10

Patrick Grady, Jeremy A Collins, Chengcheng Tang, Christopher D Twigg, Kunal Aneja, James Hays, 和 Charles C Kemp。Pressurevision++:从多样RGB图像估计指尖压力。在IEEE/CVF冬季计算机视觉应用会议论文集, 第8698-8708页, 2024. 1, 2, 3, 7, 9, 10

[26] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18995-19012, 2022. 2, 9

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu等。Ego4d:在3000小时的第一人称视频中环游世界。在IEEE/CVF计算机视觉与模式识别会议论文集, 第18995-19012页, 2022. 2, 9

[27] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. arXiv preprint arXiv:2311.18259, 2023. 2, 9

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote等。Ego-exo4d:从第一人称和第三人称视角理解熟练的人类活动。arXiv预印本 arXiv:2311.18259, 2023. 2, 9

[28] Yizheng Gu, Chun Yu, Zhipeng Li, Weiqi Li, Shuchang Xu, Xiaoying Wei, and Yuanchun Shi. Accurate and low-latency sensing of touch contact on any surface with finger-worn imu sensor. In Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology, page 1059-1070, New York, NY, USA, 2019. Association for Computing Machinery. 2

顾一正, 于春, 李志鹏, 李伟奇, 徐书畅, 魏晓颖, 史元春. 使用手指佩戴的IMU传感器在任何表面上实现精确且低延迟的触摸接触感知. 在《第32届ACM用户界面软件与技术研讨会论文集》中, 第1059-1070页, 美国纽约州纽约市, 2019. 计算机协会. 2

[29] Sean Gustafson, Christian Holz, and Patrick Baudisch. Imaginary phone: Learning imaginary interfaces by transferring spatial memory from a familiar device. In Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, page 283-292, New York, NY, USA, 2011. Association for Computing Machinery. 2

肖恩·古斯塔夫森, 克里斯蒂安·霍尔兹, 帕特里克·鲍迪施. 想象手机: 通过从熟悉设备转移空间记忆来学习想象界面. 在《第24届ACM用户界面软件与技术研讨会论文集》中, 第283-292页, 美国纽约州纽约市, 2011. 计算机协会. 2

[30] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for annotation of hand and object poses. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3196-3206, 2020. 2, 3, 9

什雷亚斯·汉帕利, 马赫迪·拉德, 马库斯·奥伯韦格, 文森特·勒佩蒂. Honnotate: 一种用于手和物体姿态的 标注方法. 在《IEEE/CVF计算机视觉与模式识别会议论文集》中, 第3196-3206页, 2020. 2, 3, 9

[31] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. ACM Transactions on Graphics (ToG), 39(4):87-1, 2020. 2

韩尚辰, 刘贝贝, 兰迪·卡贝萨斯, 克里斯托弗·D·特威格, 张培钊, 杰夫·佩特考, 余子豪, 戴春荣, 穆扎弗·阿克巴伊, 王铮, 等. Megatrack: 用于虚拟现实的单色第一人称视角关节手部追踪. 《ACM图形学汇刊》(ToG), 39(4):87-1, 2020. 2

[32] Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Lin-guang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. Umetrack: Unified multiview end-to-end hand tracking for vr. In SIGGRAPH Asia 2022 Conference Papers, pages 1-9, 2022. 2, 10

韩尚辰, 吴柏辰, 张宇博, 刘贝贝, 张林光, 王铮, 司伟光, 张培钊, 蔡宇君, 托马斯·霍丹, 等. Umetrack: 用于VR的统一多视图端到端手部追踪. 在《SIGGRAPH Asia 2022会议论文》中, 第1-9页, 2022. 2, 10

[33] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11807-11816, 2019. 3, 9

亚娜·哈松, 古尔·瓦罗尔, 迪米特里奥斯·齐奥纳斯, 伊戈尔·卡列瓦蒂赫, 迈克尔·J·布莱克, 伊万·拉普捷夫, 科迪莉亚·施密德. 学习手部与操纵物体的联合重建. 在《IEEE/CVF计算机视觉与模式识别会议论文集》中, 第11807-11816页, 2019. 3, 9

[34] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In CVPR, 2019. 3, 4

亚娜·哈松, 古尔·瓦罗尔, 迪米特里斯·齐奥纳斯, 伊戈尔·卡列瓦蒂赫, 迈克尔·J·布莱克, 伊万·拉普捷夫, 科迪莉亚·施密德. 学习手部与操纵物体的联合重建. 在《CVPR》中, 2019. 3, 4

[35] Jonas Hein, Matthias Seibold, Federica Bogo, Mazda Farshad, Marc Pollefeys, Philipp Fürnstahl, and Nassir Navab. Towards markerless surgical tool and hand pose estimation. International journal of computer assisted radiology and surgery, 16: 799-808, 2021. 2

乔纳斯·海因, 马蒂亚斯·塞博尔德, 费代丽卡·博戈, 马兹达·法尔沙德, 马克·波利菲斯, 菲利普·菲尔恩斯塔尔, 纳西尔·纳瓦布. 迈向无标记手术工具和手部姿态估计. 《国际计算机辅助放射学与外科杂志》, 16: 799-808, 2021. 2

[36] Steven Henderson and Steven Feiner. Opportunistic tangible user interfaces for augmented reality. IEEE Transactions on Visualization and Computer Graphics, 16(1):4-16, 2010. 1

史蒂文·亨德森, 史蒂文·费纳. 增强现实的机遇性有形用户界面. 《IEEE可视化与计算机图形学汇刊》, 16(1):4-16, 2010. 1

[37] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132-7141, 2018. 1

胡杰, 沈力, 孙刚. 压缩与激励网络. 在《IEEE计算机视觉与模式识别会议论文集》中, 第7132-7141页, 2018. 1

[38] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019. 1

胡杰, 沈力, 塞缪尔·阿尔巴尼, 孙刚, 吴恩华. 压缩与激励网络, 2019. 1

[39] Wonjun Hwang and Soo-Chul Lim. Inferring interaction force from visual information without using physical force sensors. Sensors, 17(11):2455, 2017. 3

黄元俊, 林秀哲. 在不使用物理力传感器的情况下从视觉信息推断交互力. 《传感器》, 17(11):2455, 2017. 3

[40] Pavel Iakubovskii. Segmentation models pytorch. https: //github.com/qubvel/segmentation\_models. pytorch, 2019. 1

帕维尔·亚库博夫斯基. 分割模型pytorch. https: //github.com/qubvel/segmentation\_models. pytorch, 2019. 1

[41] Sensel Inc. Sensel morph., 2024. 3, 5

Sensel公司. Sensel morph., 2024. 3, 5

[42] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 14713-14724, 2023. 9

简俊涛, 刘秀平, 李曼怡, 胡瑞珍, 刘健. Affordpose: 一个大规模的手-物体交互数据集, 具有可供性驱动的手部姿态. 在《IEEE/CVF国际计算机视觉会议论文集》中, 第14713-14724页, 2023. 9

[43] Tero Karras. Maximizing parallelism in the construction of bvhs, octrees, and -d trees. In Proceedings of the Fourth ACM SIGGRAPH / Eurographics Conference on High-Performance Graphics, pages 33-37. Eurographics Association, 2012. 4

Tero Karras。在构建BVH、八叉树和 -d树时最大化并行性。在第四届ACM SIGGRAPH / Eurographics高性能图形会议论文集，第33-37页。Eurographics协会，2012年。4

[44] Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. HARP: Personalized Hand Reconstruction from a Monocular RGB Video. 2023. 4

Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, 和 Siyu Tang。HARP:从单目RGB视频中个性化手部重建。2023年。4

[45] Hong-Ki Kim, Seunggun Lee, and Kwang-Seok Yun. Capacitive tactile sensor array for touch screen application. Sensors and Actuators A: Physical, 165(1):2-7, 2011. 3

Hong-Ki Kim, Seunggun Lee, 和 Kwang-Seok Yun。用于触摸屏应用的电容式触觉传感器阵列。传感器与执行器A:物理，165(1):2-7, 2011年。3

[46] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. arXiv:2304.02643, 2023. 3, 4

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, 和 Ross Girshick。分割一切。arXiv:2304.02643, 2023年。3, 4

[47] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10138-10148, 2021. 2, 3, 9

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, 和 Marc Pollefeys。H2O:双手操作物体以实现第一人称交互识别。在IEEE/CVF国际计算机视觉会议论文集，第10138-10148页，2021年。2, 3, 9

[48] Mike Lambeta, Tingfan Wu, Ali Sengul, Victoria Rose Most, Nolan Black, Kevin Sawyer, Romeo Mercado, Haozhi Qi, Alexander Sohn, Byron Taylor, et al. Digitizing touch with an artificial multimodal fingertip. arXiv preprint arXiv:2411.02479, 2024. 1, 2

Mike Lambeta, Tingfan Wu, Ali Sengul, Victoria Rose Most, Nolan Black, Kevin Sawyer, Romeo Mercado, Haozhi Qi, Alexander Sohn, Byron Taylor, 等。用人工多模态指尖数字化触觉。arXiv预印本arXiv:2411.02479, 2024年。1, 2

[49] Minkyung Lee, Woontack Woo, et al. Arkb: 3d vision-based augmented reality keyboard. In

Minkyung Lee, Woontack Woo, 等。Arkb:基于3D视觉的增强现实键盘。在

[50] Shuang Li, Jiaxi Jiang, Philipp Ruppel, Hongzhuo Liang, Xiaojian Ma, Norman Hendrich, Fuchun Sun, and Jianwei Zhang. A mobile robot hand-arm teleoperation system by vision and imu. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 10900-10906. IEEE, 2020. 2

Shuang Li, Jiaxi Jiang, Philipp Ruppel, Hongzhuo Liang, Xiaojian Ma, Norman Hendrich, Fuchun Sun, 和 Jianwei Zhang。通过视觉和IMU的移动机器人手臂遥操作系统。在2020年IEEE/RSJ智能机器人与系统国际会议(IROS)，第10900-10906页。IEEE，2020年。2

[51] Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3d motion and forces of person-object interactions from monocular video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8640-8649, 2019. 3

Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, 和 Josef Sivic。从单目视频中估计人与物体交互的3D运动和力。在IEEE/CVF计算机视觉与模式识别会议论文集，第8640-8649页，2019年。3

[52] Chen Liang, Xutong Wang, Zisu Li, Chi Hsia, Mingming Fan, Chun Yu, and Yuanchun Shi. Shadowtouch: Enabling free-form touch-based hand-to-surface interaction with wrist-mounted illuminant by shadow projection. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pages 1-14, 2023. 2

Chen Liang, Xutong Wang, Zisu Li, Chi Hsia, Mingming Fan, Chun Yu, 和 Yuanchun Shi。Shadowtouch:通过阴影投影实现基于手腕光源的自由形式手与表面交互。在第36届ACM用户界面软件与技术年会论文集，第1-14页，2023年。2

[53] PPS UK Limited. Tactileglove - hand pressure and force measurement., 2023. 3

PPS UK Limited。Tactileglove - 手部压力和力量测量。，2023年。3

[54] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117-2125, 2017. 1

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。在IEEE计算机视觉与模式识别会议论文集，第2117-2125页，2017年。1

[55] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21013-21022, 2022. 9

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, 和 Li Yi。Hoi4d:用于类别级人与物体交互的4D第一人称数据集。在IEEE/CVF计算机视觉与模式识别会议论文集，第21013-21022页，2022年。9

[56] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. arXiv preprint arXiv:2401.08399, 2024. 3, 9

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, 和 Li Yi。Taco:评估可泛化的双手工具-动作-物体理解的基准。arXiv预印本arXiv:2401.08399, 2024年。3, 9

[57] Gabriel Lugo, Mario Ibarra-Manzano, Fang Ba, and Irene Cheng. Virtual reality and hand tracking system as a medical tool to evaluate patients with parkinson’s. In Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, pages 405-408, 2017. 2

加布里埃尔·卢戈、马里奥·伊巴拉-曼萨诺、方巴和艾琳·程。虚拟现实和手部追踪系统作为评估帕金森病患者的医疗工具。发表于第11届EAI国际普适计算技术医疗保健会议论文集，第405-408页，2017年。2

[58] Yiyue Luo, Yunzhu Li, Pratyusha Sharma, Wan Shou, Kui Wu, Michael Foshey, Beichen Li, Tomás Palacios, Antonio Tor-ralba, and Wojciech Matusik. Learning human-environment interactions using conformal tactile textiles. Nature Electronics, 4(3):193-201, 2021. 1, 3

罗一越、李云珠、普拉蒂莎·夏尔马、寿万、吴奎、迈克尔·福谢、李北辰、托马斯·帕拉西奥斯、安东尼奥·托拉尔巴和沃伊切赫·马图西克。使用共形触觉纺织品学习人与环境的交互。《自然电子学》，4(3):193-201, 2021年。1, 3

[59] Yiyue Luo, Chao Liu, Young Joong Lee, Joseph DelPreto, Kui Wu, Michael Foshey, Daniela Rus, Tomás Palacios, Yun-zhu Li, Antonio Torralba, et al. Adaptive tactile interaction transfer via digitally embroidered smart gloves. Nature communications, 15(1):868, 2024. 1

罗一越、刘超、李永钟、约瑟夫·德尔普雷托、吴奎、迈克尔·福谢、丹妮拉·鲁斯、托马斯·帕拉西奥斯、李云珠、安东尼奥·托拉尔巴等。通过数字刺绣智能手套实现自适应触觉交互转移。《自然通讯》，15(1):868, 2024年。1

[60] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In 2021 IEEE international conference on robotics and automation (ICRA), pages 6169-6176. IEEE, 2021. 1, 3

普里扬卡·曼迪卡尔和克里斯滕·格劳曼。通过以对象为中心的视觉可供性学习灵巧抓取。发表于2021年IEEE国际机器人与自动化会议(ICRA)，第6169-6176页。IEEE, 2021年。1, 3

[61] Stephen A Mascaro and H Harry Asada. Measurement of finger posture and three-axis fingertip touch force using fingernail sensors. IEEE Transactions on Robotics and Automation, 20(1):26-35, 2004. 3

斯蒂芬·A·马斯卡罗和阿萨达·H·哈里。使用指甲传感器测量手指姿势和三轴指尖触力。《IEEE机器人与自动化汇刊》，20(1):26-35, 2004年。3

[62] Manuel Meier, Paul Streli, Andreas Fender, and Christian

曼努埃尔·迈尔、保罗·斯特雷利、安德烈亚斯·芬德和克里斯蒂安

Holz. Tapid: Rapid touch interaction in virtual reality using wearable sensing. In 2021 IEEE Virtual Reality and 3D User Interfaces (VR), pages 519-528. IEEE, 2021. 1, 2

霍尔茨。Tapid:使用可穿戴传感器在虚拟现实中实现快速触摸交互。发表于2021年IEEE虚拟现实与3D用户界面会议(VR)，第519-528页。IEEE, 2021年。1, 2

[63] Vimal Mollyn and Chris Harrison. Egotouch: On-body touch input using ar/vr headset cameras. In Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, pages 1-11, 2024. 3

维马尔·莫林和克里斯·哈里森。Egotouch:使用AR/VR头戴设备摄像头进行身体触摸输入。发表于第37届ACM用户界面软件与技术年会论文集，第1-11页，2024年。3

[64] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for interacting hand pose estimation from a single rgb image. In Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XX 16, pages 548-564. Springer, 2020. 2

文庆石、俞秀一、温赫、白崇利和李京穆。Interhand2.6m:从单个RGB图像进行交互手部姿态估计的数据集和基线。发表于计算机视觉-ECCV 2020:第16届欧洲会议，英国格拉斯哥，2020年8月23-28日，第XX部分16，第548-564页。Springer, 2020年。2

[65] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In Proceedings of the IEEE International Conference on Computer Vision, pages 1154-1163, 2017. 2

弗朗西斯卡·穆勒、杜希扬特·梅塔、奥列克桑德尔·索特尼琴科、斯里纳特·斯里达尔、丹·卡萨斯和克里斯蒂安·特奥博尔特。在自中心RGB-D传感器遮挡下的实时手部追踪。发表于IEEE国际计算机视觉会议论文集，第1154-1163页，2017年。2

[66] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12999-13008, 2023. 9

大川武彦、何坤、法迪梅·塞纳尔、托马斯·霍丹、栾·陈和凯姆·凯斯金。Assemblyhands:通过3D手部姿态估计实现自中心活动理解。发表于IEEE/CVF计算机视觉与模式识别会议论文集，第12999-13008页，2023年。9

[67] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In , 5, 6

乔治奥斯·帕夫拉科斯、单丹丹、伊利亚·拉多萨沃维奇、安吉·卡纳泽、大卫·福伊和吉滕德拉·马利克。使用变压器重建3D手部。在 中，5, 6

[68] Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. IEEE transactions on pattern analysis and machine intelligence, 40(12):2883-2896, 2017. 3

范图华、尼古拉奥斯·基里亚齐斯、安东尼斯·A·阿吉罗斯和阿卜杜勒拉赫曼·凯达。从无标记视觉追踪中估计手-物体接触力。《IEEE模式分析与机器智能汇刊》，40(12):2883-2896, 2017年。3

[69] Philip Quinn, Wenxin Feng, and Shumin Zhai. Deep touch: Sensing press gestures from touch image sequences. Artificial Intelligence for Human Computer Interaction: A Modern Approach, pages 169-192, 2021. 3

菲利普·奎因、冯文新和翟书民。深度触摸:从触摸图像序列中感知按压手势。《人工智能用于人机交互:现代方法》，第169-192页，2021年。3

[70] James M Rehg and Takeo Kanade. Digiteyes: Vision-based hand tracking for human-computer interaction. In Proceedings of 1994 IEEE workshop on motion of non-rigid and articulated objects, pages 16-22. IEEE, 1994. 2

詹姆斯·M·雷格和金子武夫。Digiteyes:基于视觉的手部追踪用于人机交互。发表于1994年IEEE非刚体和关节物体运动研讨会论文集，第16-22页。IEEE, 1994年。2

[71] Mark Richardson, Matt Durasoff, and Robert Wang. Decoding surface touch typing from hand-tracking. In Proceedings of the 33rd annual ACM symposium on user interface software and technology, pages 686-696, 2020. 2

Mark Richardson, Matt Durasoff, 和 Robert Wang。从手部追踪解码表面打字。在2020年第33届ACM用户界面软件与技术研讨会论文集，第686-696页。2

[72] Mark Richardson, Fadi Botros, Yangyang Shi, Pinhao Guo, Bradford J Snow, Linguang Zhang, Jingming Dong, Keith Vertanen, Shugao Ma, and Robert Wang. Stegotype: Surface typing from egocentric cameras. In Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, pages 1-14, 2024. 2

Mark Richardson, Fadi Botros, Yangyang Shi, Pinhao Guo, Bradford J Snow, Linguang Zhang, Jingming Dong, Keith Vertanen, Shugao Ma, 和 Robert Wang。Stegotype:从第一视角摄像头进行表面打字。在2024年第37届ACM用户界面软件与技术研讨会论文集，第1-14页。2

[73] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia), 2017. 2, 3, 8

Javier Romero, Dimitrios Tzionas, 和 Michael J. Black。具身手部:一起建模和捕捉手部和身体。ACM图形学汇刊，(Proc. SIGGRAPH Asia)，2017。2, 3, 8

[74] Elliot N. Saba, Eric C. Larson, and Shwetak N. Patel. Dante vision: In-air and touch gesture sensing for natural surface

Elliot N. Saba, Eric C. Larson, 和 Shwetak N. Patel。Dante视觉:用于自然表面交互的空中和触摸手势感应

interaction with combined depth and thermal cameras. In 2012 IEEE International Conference on Emerging Signal Processing Applications, pages 167-170, 2012. 2

结合深度和热成像摄像头。在2012年IEEE新兴信号处理应用国际会议，第167-170页，2012。2

[75] Pressure Mapping Sensors. Tekscan., 2024. 3

压力映射传感器。Tekscan., 2024。3

[76] Vivian Shen, James Spann, and Chris Harrison. Farout touch: Extending the range of ad hoc touch sensing with depth cameras. In Proceedings of the 2021 ACM Symposium on Spatial User Interaction, New York, NY, USA, 2021. Association for Computing Machinery. 2

Vivian Shen, James Spann, 和 Chris Harrison。Farout触摸:通过深度摄像头扩展临时触摸感应的范围。在2021年ACM空间用户交互研讨会论文集，美国纽约，2021。计算机协会。2

[77] Yilei Shi, Haimo Zhang, Jiashuo Cao, and Suranga Nanayakkara. Versatouch: A versatile plug-and-play system that enables touch interactions on everyday passive surfaces. In Proceedings of the Augmented Humans International Conference, New York, NY, USA, 2020. Association for Computing Machinery. 2

Yilei Shi, Haimo Zhang, Jiashuo Cao, 和 Suranga Nanayakkara。Versatouch:一个多功能即插即用系统，使日常被动表面上的触摸交互成为可能。在2020年增强人类国际会议论文集，美国纽约，2020。计算机协会。2

[78] Yilei Shi, Haimo Zhang, Kaixing Zhao, Jiashuo Cao, Meng-meng Sun, and Suranga Nanayakkara. Ready, steady, touch! sensing physical contact with a finger-mounted imu. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 4(2), 2020. 2

Yilei Shi, Haimo Zhang, Kaixing Zhao, Jiashuo Cao, Meng-meng Sun, 和 Suranga Nanayakkara。准备，稳定，触摸！通过手指安装的IMU感应物理接触。Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 4(2), 2020。2

[79] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In Proceedings of EUROGRAPHICS/ACM SIG-GRAPH Symposium on Geometry Processing, pages 109-116, 2007. 4

Olga Sorkine 和 Marc Alexa。尽可能刚性的表面建模。在2007年EUROGRAPHICS/ACM SIGGRAPH几何处理研讨会论文集，第109-116页。4

[80] Paul Streli, Jiaxi Jiang, Andreas Fender, Manuel Meier, Hugo Romat, and Christian Holz. Taptype: Ten-finger text entry on everyday surfaces via bayesian inference. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2022. Association for Computing Machinery. 2

Paul Streli, Jiaxi Jiang, Andreas Fender, Manuel Meier, Hugo Romat, 和 Christian Holz。Taptype:通过贝叶斯推理在日常表面上进行十指文本输入。在2022年CHI计算系统人因会议论文集，美国纽约，2022。计算机协会。2

[81] Paul Streli, Jiaxi Jiang, Juliete Rossie, and Christian Holz. Structured light speckle: Joint ego-centric depth estimation and low-latency contact detection via remote vibrometry. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pages 1-12, 2023. 2

Paul Streli, Jiaxi Jiang, Juliete Rossie, 和 Christian Holz。结构光散斑:通过远程振动测量联合第一视角深度估计和低延迟接触检测。在2023年第36届ACM用户界面软件与技术研讨会论文集，第1-12页。2

[82] Paul Streli, Mark Richardson, Fadi Botros, Shugao Ma, Robert Wang, and Christian Holz. Touchinsight: Uncertainty-aware rapid touch and text input for mixed reality from egocentric vision. In Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, pages 1-16, 2024. 1, 2

Paul Streli, Mark Richardson, Fadi Botros, Shugao Ma, Robert Wang, 和 Christian Holz。Touchinsight:从第一视角视觉进行混合现实的快速触摸和文本输入，具有不确定性感知。在2024年第37届ACM用户界面软件与技术研讨会论文集，第1-16页。1, 2

[83] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. Nature, 569(7758):698-702, 2019. 3

Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, 和 Wojciech Matusik。使用可扩展触觉手套学习人类抓握的特征。自然，569(7758):698-702, 2019。3

[84] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV 16, pages 581-600. Springer, 2020. 2, 3, 9

Omid Taheri, Nima Ghorbani, Michael J Black, 和 Dimitrios Tzionas。Grab:一个全身人类抓握物体的数据集。在计算机视觉-ECCV 2020:第16届欧洲会议，英国格拉斯哥，2020年8月23-28日，第四部分16，第581-600页。Springer, 2020。2, 3, 9

[85] Ryo Takahashi, Masaaki Fukumoto, Changyo Han, Takuya Sasatani, Yoshiaki Narusue, and Yoshihiro Kawahara. Telemetring: A batteryless and wireless ring-shaped keyboard using passive inductive telemetry. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, page 1161-1168, New York, NY, USA, 2020. Association for Computing Machinery. 2

高桥亮、福本正明、韩昌耀、笹谷拓也、成末义明和川原义宏。Telemetring:一种使用无源感应遥测技术的无电池无线环形键盘。在《第33届ACM用户界面软件与技术研讨会论文集》中，第1161-1168页，美国纽约州纽约市，2020年。计算机协会。2

[86] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. International Journal of Computer Vision (IJCV), 118 (2):172-193, 2016. 4

迪米特里奥斯·齐奥纳斯、卢卡·巴兰、阿比拉什·斯里坎塔、巴勃罗·阿庞特、马克·波利菲斯和尤尔根·加尔。使用判别显著点和物理模拟捕捉动作中的手。《国际计算机视觉杂志》(IJCV)，118(2):172-193，2016年。4

[87] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(11), 2008. 6

劳伦斯·范德马滕和杰弗里·辛顿。使用t-SNE可视化数据。《机器学习研究杂志》，9(11)，2008年。6

[88] Andrew D Wilson. Playanywhere: a compact interactive tabletop projection-vision system. In Proceedings of the 18th annual ACM symposium on User interface software and technology, pages 83-92, 2005. 1, 2

安德鲁·D·威尔逊。Playanywhere:一种紧凑的交互式桌面投影视觉系统。在《第18届ACM用户界面软件与技术研讨会论文集》中，第83-92页，2005年。1, 2

[89] Andrew D. Wilson. Using a depth camera as a touch sensor. In ACM International Conference on Interactive Tabletops and Surfaces, page 69-72, New York, NY, USA, 2010. Association for Computing Machinery. 2

安德鲁·D·威尔逊。使用深度相机作为触摸传感器。在《ACM国际交互式桌面和表面会议》中，第69-72页，美国纽约州纽约市，2010年。计算机协会。2

[90] Robert Xiao, Scott Hudson, and Chris Harrison. Direct: Making touch tracking on ordinary surfaces practical with hybrid depth-infrared sensing. In Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces, pages 85-94, 2016.

罗伯特·肖、斯科特·哈德森和克里斯·哈里森。Direct:通过混合深度红外传感使普通表面上的触摸跟踪变得实用。在《2016年ACM国际交互式表面和空间会议论文集》中，第85-94页，2016年。

[91] Robert Xiao, Julia Schwarz, Nick Throm, Andrew D. Wilson, and Hrvoje Benko. Mrtouch: Adding touch input to head-mounted mixed reality. IEEE Transactions on Visualization and Computer Graphics, 24(4):1653-1660, 2018. 1, 2

罗伯特·肖、朱莉娅·施瓦茨、尼克·特罗姆、安德鲁·D·威尔逊和赫尔沃耶·本科。Mrtouch:为头戴式混合现实添加触摸输入。《IEEE可视化和计算机图形学汇刊》，24(4):1653-1660，2018年。1, 2

[92] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20953-20962, 2022. 2, 3, 9

杨立新、李开林、詹新宇、吴飞、徐安然、刘流和卢策武。Oakink:一个用于理解手-物体交互的大规模知识库。在《IEEE/CVF计算机视觉与模式识别会议论文集》中，第20953-20962页，2022年。2, 3, 9

[93] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2. benchmark: Hand pose dataset and state of the art analysis. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4866-4874, 2017. 2

袁善新、叶琪、比约恩·斯滕格、西德汉特·贾因和金泰均。Bighand2。 基准:手姿势数据集和最新技术分析。在《IEEE计算机视觉与模式识别会议论文集》中，第4866-4874页，2017年。2

[94] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Han-lin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. arXiv preprint arXiv:2403.19417, 2024. 2, 3, 9

詹新宇、杨立新、赵一飞、毛康瑞、徐翰林、林泽南、李开林和卢策武。Oakink2:一个关于复杂任务完成中双手-物体操作的数据集。arXiv预印本arXiv:2403.19417，2024年。2, 3, 9

[95] Hao Zheng, Regina Lee, and Yuqian Lu. Ha-vid: A human assembly video dataset for comprehensive assembly knowledge understanding. Advances in Neural Information Processing Systems, 36, 2024. 9

郑浩、李瑞娜和卢玉倩。Ha-vid:一个用于全面理解装配知识的人类装配视频数据集。《神经信息处理系统进展》，36，2024年。9

[96] Zehao Zhu, Jiashun Wang, Yuzhe Qin, Deqing Sun, Varun Jampani, and Xiaolong Wang. Contactart: Learning interaction priors for category-level articulated object and hand poses estimation. arXiv preprint arXiv:2305.01618, 2023. 3, 9

朱泽浩、王佳顺、秦玉哲、孙德清、瓦伦·詹帕尼和王小龙。Contactart:学习 交互先验以估计类别级关节物体和手姿势。arXiv预印本arXiv:2305.01618，2023年。3, 9

[97] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 813-822, 2019. 2, 10

克里斯蒂安·齐默尔曼、杜伊古·塞兰、杨继美、布莱恩·拉塞尔、马克斯·阿格斯和托马斯·布罗克斯。Freihand:一个用于从单张RGB图像中无标记捕捉手姿势和形状的数据集。在《IEEE/CVF国际计算机视觉会议论文集》中，第813-822页，2019年。2, 10

[98] Lara Zlokapa, Yiyue Luo, Jie Xu, Michael Foshey, Kui Wu, Pulkit Agrawal, and Wojciech Matusik. An integrated design pipeline for tactile sensing robotic manipulators. In 2022 In-

拉拉·兹洛帕、罗艺月、徐杰、迈克尔·福西、吴奎、普尔基特·阿格拉瓦尔和沃伊切赫·马图西克。一种用于触觉传感机器人操纵器的集成设计流程。在2022年国际机器人

ternational Conference on Robotics and Automation (ICRA),

与自动化会议(ICRA)中，

pages 3136-3142. IEEE, 2022. 2

第3136-3142页。IEEE，2022年。2

1. \*Equal contribution.

   \*同等贡献。 [↑](#footnote-ref-29)