基于超声波感知的手势重建和识别

田跃悦，黄家栋，尹华锐，陈力
（中国科学技术大学，安徽 合肥 230022）

摘 要：为解决基于视觉的手势感知中存在的隐私泄露风险以及其他非视觉信号感知成本高的问题，本文研究基于超声波信号的手势重建和识别。对采集的超声回波和手势图片数据处理构造出一个超声-手势框架数据集 Ultrasonic Gesture。基于该数据集，提出了一种具有高性能局部感知与全局建模能力的 CAMT-Net 神经网络，实现从超声波信号到二维手势关键点坐标的高精度端到端映射。在包含六种静态手势的数据集上进行实验，所提方法生成的关键点精度接近基于 RGB 图像的重建方法；进一步基于重建关键点进行手势识别，准确率达到 89%。结果表明，超声波信号可有效支持细粒度手势感知任务。

关键词：超声波；感知；手势；深度学习
中图分类号：TN92
文献标志码：A
doi：10.11959/j.issn.2096−3750.XXXX.

Gesture reconstruction and recognition based on ultrasonic sensing

Tian Yueyue, Huang jiadong, Yin Huarui, Chen Li
University of Science and Technology of China, Hefei 230022, China

Abstract: To address the privacy leakage risks in vision-based gesture sensing and the high sensing costs of other non-visual signals, this paper investigated a gesture reconstruction and recognition method based on ultrasonic signals. The collected ultrasonic echo and gesture image data were processed to construct the Ultrasonic Gesture dataset. Based on this dataset, we proposed a CAMT-Net with high-performance local perception and global modeling capabilities, achieving high-precision end-to-end mapping from ultrasonic signals to 2D gesture keypoint coordinates. Experiments were conducted on the dataset containing six static gestures. The proposed method achieved accuracy close to the metods based on RGB images. Further gesture recognition based on the reconstructed keypoints reached an accuracy of 89%. The results indicate that ultrasonic signals can effectively support fine-grained gesture perception tasks.

Key words: ultrasound, sensing, gesture, deep learning

0 引言

手势重建是人机交互、计算机视觉与模式识别领域的一项关键技术，旨在从传感器数据中精确恢复出手部的几何结构与姿态信息，为实现自然、直观的人机协作提供了重要的技术支撑。随着虚拟现实、增强现实、智能机器人及远程操控等应用的快速发展，对手部动作高保真、低延迟的重建需求日益增长。相较于传统输入设备，基于手势的交互方式更符合人类自然行为习惯，能够显著提升用户体验与交互效率。因此，高精度的手势重建不仅有助于深入理解人类动作意图，也为构建智能化、沉浸式的人机协同环境[1,2]奠定了基础，具有重要的理论意义与广泛的应用前景。

为满足人机交互需求，现有手势重建方法主要分为三类：基于可穿戴设备的方法[3,4]、基于计算

收稿日期：XXXX−XX−XX；**修回日期：**XXXX−XX−XX
通信作者：陈力，chenli87@ustc.edu.cn
基金项目：国家自然科学基金青年基金（B 类）（No.62522126）；安徽省杰出青年基金（No.2308085J24）
Foundation Items: The National Natural Science Foundation of China Youth Fund (Category B) (No.62522126), Anhui Provincial Outstanding Youth Fund (No.2308085J24)

机视觉的方法[5-7]和基于无线传感技术的方法[8-10]。可穿戴设备虽能高精度捕捉手部运动，但成本高、需个性化适配，且佩戴舒适性差，限制了日常应用；视觉方法无需佩戴设备、用户体验好，但易受光照、遮挡和隐私问题影响；无线传感技术具有环境鲁棒性和隐私友好性，适用于多种现实场景。具体使用的无线传感信号主要是毫米波信号、WIFI信号和超声波信号。其中毫米波和WIFI信号应用分别受限于设备成本高和重建精度低的缺点。相比之下，超声波信号兼具低成本、非侵入、隐私保护的优势及对细粒度手势的感知能力，可利用低成本的超声波设备或现有扬声器与麦克风实现高效手势重建。

目前的手部姿态估计方法主要有生成式方法和判别式方法两类。生成式方法通常依赖于一个显式的手部骨架模型，结合渲染引擎或投影函数，将假设的姿态投影到图像空间，生成对应的视觉手部框架。然后通过最小化生成手部框架与真实观测之间的差异，迭代优化手部姿态参数。生成式方法的典型应用包括[11-13]等。该方法的优势是有较强的可解释性且准确度较高，但推理过程通常需要迭代优化，计算开销大。判别式方法通常利用机器学习模型，在大量标注数据上进行端到端训练，直接输出关键点坐标。判别式方法的典型应用有[14-16]等。近年来，随着深度学习的发展，判别式方法在精度和效率上占据主导地位。

本文提出用非接触式超声波设备收发信号感知静态手势，从回声信号重建出二维手势框架，其在实际应用中面临多重挑战。首先，从一维超声波信号建立到二维手势坐标的映射是一个高纬度非线性问题，建立函数映射关系较为困难。因此，本文使用判别式手势重建方法，提出CAMT-Net（Channel Attention and MLP-Enhanced Transformer Network）网络，实现了对超声波信号中手势信息的充分发掘，重建出高准确度的手势框架。其次，使用判别式方法进行手势重建需要大量的标注数据，使用佩戴式传感器获取手势关键点坐标的成本较高。为了实现低成本高精度的手势标注，本文使用视觉方法中准确度较高的OpenPose模型[17]生成手势关键点坐标作为训练标签。最后，从回声信号中重建出手势框架图像是跨模态任务，采集两种模态数据后需要进行严格的对齐和同步。本文使用标准时间戳作为超声波样本和手势关键点样本的同步基准，最终构建出数据集Ultrasonic Gesture。

本文的主要工作如下：

1） 基于超声波-手势模型分析了多路径效应下进行手势重建的可行性，指出利用空间几何模型解析手势关键点坐标的困难，最后提出用神经网络方法直接将超声波幅值信号映射到手势坐标。

2） 利用基于视觉的手势重建方法获得手势关键点真实值，利用标准时间戳实现数据同步，最终实现低成本构建超声波-手势关键点数据集Ultrasonic Gesture。

3） 提出了CAMT-Net神经网络，实现从超声波信号中提取细粒度手势信息并重建出手势框架，再基于重建手势进行识别。

1　相关工作

目前手势重建的相关研究工作主要分为基于可穿戴设备的方法、基于计算机视觉的方法和基于无线传感技术的方法。本节总结近年来的相关研究成果。

1.1　基于可穿戴超声波设备的手势重建

早期手势重建的工作主要利用数据手套等可穿戴设备捕捉手部形态与姿势，进而重建三维手部模型。Wang等人[3]采用单个摄像头来追踪佩戴有印制自定义图案的普通布手套的手，从而实现三维关节式用户输入。Hu等人[4]通过在贴合手腕的腕带上紧密安装四个微型热像仪，观察手腕处手部的轮廓，利用低分辨率的热成像仪估算整个手部的姿势，实现了连续的三维手指追踪和手势估计功能。

1.2　基于计算机视觉的手势重建

近年来，随着计算机视觉技术的蓬勃发展，基于视觉的方法在手部姿态估计领域占据主导地位。通过构建深度学习模型，并在大规模数据集上进行训练，这类模型通常能实现手部关节回归的高精度。例如，Simon等人[17]提出了一种利用多摄像头系统训练精细检测器的方法，首先使用初始关键点检测器在手部多个视角生成含噪标签；随后通过多视角几何三角测量对噪声检测结果进行三维三角化处理或标记为异常值；最终将投影三角化数据作为新标注训练集来优化检测器性能。Wang等人[18]提出基于单张真实RGB图像的二维手部姿态估计级联网络，构建了包含掩膜预测阶段和姿态估计阶段

的两级级联网络架构，这种分阶段网络架构在端到端训练中能够相互促进，有效提升手部掩膜检测与二维姿态估计的精度。

随着对手部姿态估计质量要求的不断提高，部分研究开始拓展到三维手部网格重建领域[19-21]。这些方法利用带有三维标注的数据集，从RGB图像中生成逼真的手部网格模型。例如，Tang等人[21]实现从RGB图像进行三维手部网格重建，首先预测手部的关节位置以及进行手部的分割，然后在此基础上预测出一个较为粗糙的手部网格，最后使用一个偏移网格来对得到的粗糙网格进行微调，从而实现精确的网格与图像的对齐。

1.3　基于无线传感技术的手势重建

基于无线信号的手势重建技术因其非接触式感知和良好的环境适应性而受到广泛关注。研究人员利用不同类型的无线信号特性，探索了多种手势感知与重建方法。例如，Kong等人[22]提出mmHand系统利用商用现成的毫米波雷达生成三维手部骨骼进行手部姿态估计，mmHand利用基于注意力机制的沙漏网络提取手部的多尺度空间特征，并使用LSTM提取时间特征，之后在三维空间中回归手关节以生成三维手骨骼。Ji等人[23]提出了一种名为HandFi方案通过实用WIFI设备实现手部骨骼结构重建。HandFi开发了一种多任务学习神经网络，并采用定制化损失函数捕捉WIFI信号中的低层次手部信息，实现了在线使用时仅需商用WIFI即可生成二维手部掩膜和三维手势模型。声波信号过去在手势感知领域也有较多的应用，但当前的声学手部追踪系统检测颗粒度不足，即只能分类离散的手势，或定位几个最近点[24-26]。进行完整的手势重建工作只能是在极近的感知范围内或者依赖于部署麦克风阵列。例如，Yang等人[27]将主动降噪耳机上的扬声器和单个前馈麦克风改造为声纳系统，利用人耳无法察觉的调频连续波信号追踪手势反射轨迹，最终实现3厘米距离内的6类用户手势精确重建。Wang等人[28]开发了一种基于麦克风阵列构建的鲁棒声学三维多手部姿态重建系统，支持多手部处理，并且能高度适应新场景。

2　原理和问题分析

2.1　手势重建原理

为了实现基于超声波信号的手势框架重建，我们采用了图1所示的超声波感知系统进行手势感知。该系统由N个独立超声波收发传感器以及飞行时间（TOF, Time of Flight）相机组成，超声波传感器用索引集$\mathcal{N}=\{1,2,...,\ N\}$表示。相机采集与超声波信号同步的手势图片，用于生成手势关键点的真实坐标并提供坐标系。超声波传感器与相机放置在人体前方，高度约1.5m，保证受试者站立采集数据的视角合适。

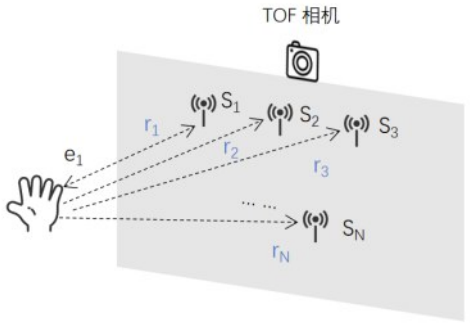


图1　超声波手势感知系统模型

超声波传感器$S_i(i=1,2,\cdots,N)$发射短脉冲$p(t)$，经由人手或其他障碍物反射信号$r(t)$到传感器$S_j(j=1,2,\cdots,N)$接收。考虑到多径效应和高斯噪声的影响，接收信号可按以下公式建模：

$$r(t)=\sum_{m=1}^{M}\alpha_m e^{-j\phi_m}p(t-\tau_m)+n(t) \tag{1}$$

其中M表示系统中所有多径成分的数量，α_m表示第m条路径的振幅衰减，ϕ_m表示第m条路径的相移，τ_m表示第m条路径的延迟，$n(t)$表示附加高斯白噪声。

计算第m条路径分量的振幅，可以得到该路径的幅度向量$\boldsymbol{a}_m=\left[a_{m,1},a_{m,2},\cdots a_{m,l}\right]^{\mathrm{T}}$，其中$l$是幅度向量的长度。将M条路径的幅度向量堆叠可以得到系统中所有路径的幅度矩阵$\boldsymbol{a}=[\boldsymbol{a}_1\ \ \boldsymbol{a}_2\ \ \cdots\ \ \boldsymbol{a}_M]$。每一个采样时刻接收信号$r(t)$的幅值是所有M条多路径幅值的加和$\boldsymbol{\alpha}=\begin{bmatrix}a_{1,1}+a_{2,1}+\cdots+a_{M,1}\\ a_{1,2}+a_{2,2}+\cdots+a_{M,2}\\ \vdots\\ a_{1,l}+a_{2,l}+\cdots+a_{M,l}\end{bmatrix}$。可见，执行不同手势会导致多径成分改变，影响多路径的振幅衰减和路径延迟，最后影响接收信号的振幅。因此，可以通过分析超声回波幅度数据解读手势的结构。

本文的目标是建立有效的映射关系$\mathcal{F}(\cdot)$，将接

收超声波信号幅值向量$\boldsymbol{\alpha}$映射到P个关键点的二维位置坐标矩阵$\bar{\boldsymbol{I}}=\begin{bmatrix}x_1 & y_1\\ x_2 & y_2\\ \vdots & \vdots\\ x_p & y_p\end{bmatrix}$。映射关系为：

$$\mathcal{F}(\boldsymbol{\alpha})=\bar{\boldsymbol{I}} \tag{2}$$

2.2　手势重建问题分析

在我们的实验环境中，发射的超声波脉冲会被反射体人手、地面、实验台等反射。超声波脉冲经过反射到传感器接收，接收信号会携带需要的手势信息和其他冗余信息。下面将分析如何从接收信号的振幅数据$\boldsymbol{\alpha}$得到手势关键点坐标矩阵$\bar{\boldsymbol{I}}$。

信号由超声波传感器收发的传播路径如图2所示，超声波信号从第i个传感器发射，经过反射后由第j个传感器接收。信号的传播路径可以分为直接反射路径和多次反射路径。信号在直接反射路径上传播时，从发射传感器到手部或其他位置，仅反射一次到接收传感器。直接反射路径长度较短，回波信号强度较大，且携带信息成分较为简单；而多次反射路径长度较长，回波信号强度小，且多次反射的反射点较多，携带成分复杂，不利于手势信息的解析。因此，我们设置超声波传感器只接收0.6m范围内的回波信号，从而有效排除多次反射路径的干扰，减少接收信号的多径成分，记式（1）中直接反射路径数量为M_d。

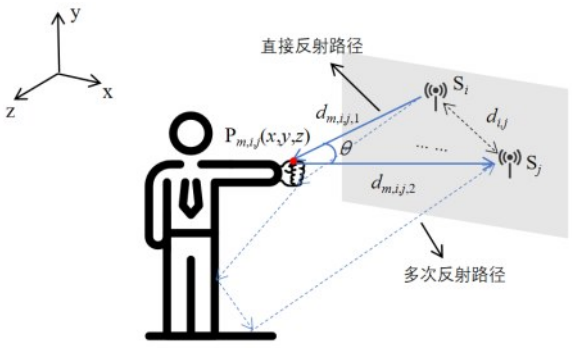


图2　信号传播路径

要具体分析系统中的反射点位置，首先要从接收信号幅值向量中分离出每一条路径对应的幅值向量$\boldsymbol{\alpha}\rightarrow\left[\boldsymbol{a}_1\quad\boldsymbol{a}_2\quad\cdots\quad\boldsymbol{a}_{\mathrm{M}_d}\right]$。对于第$m$条路径成分，超声波脉冲从传感器$\mathrm{S}_i$发出，经手部反射点$\mathrm{P}_{m,i,j}(x,y,z)$，由传感器$\mathrm{S}_j$接收回波，该直接反射路径的长度分为发射路径$d_{m,i,j,1}$和接收路径$d_{m,i,j,2}$两段，长度为：

$$d_{m,i,j}=d_{m,i,j,1}+d_{m,i,j,2} \tag{3}$$

在采样时刻t，接收超声回波在该路径分量的强度数据可表示为：

$$\left(a_r\right)_{m,i,j}=\left(a_t\right)_{m,i,j}-10\eta\lg\left(d_{m,i,j}\right) \tag{4}$$

其中，$\left(a_r\right)_{m,i,j}$和$\left(a_t\right)_{m,i,j}$分别是接收信号强度和发射信号强度，$\left(a_t\right)_{m,i,j}$是由实验设置决定的已知量，η是空气介质中的衰减。从$\boldsymbol{a}_m$中可以得到当前采样时刻该路径上回波幅度真实值$\left(a_r\right)'_{m,i,j}$，根据该值可解出这条路径长度$d_{m,i,j}$。记第$i$和第$j$个传感器的距离是$d_{i,j}$，收发传感器位置和反射点位置存在以下几何约束关系：

$$d_{i,j}^2=d_{m,i,j,1}^2+d_{m,i,j,2}^2-2d_{m,i,j,1}\cdot d_{m,i,j,2}\cdot\cos\theta \tag{5}$$

其中θ是反射点和收发传感器之间的夹角。根据式(3)、(5)以及传感器和反射点几何约束关系，可联立出第m条路径反射点方程组$f_{m,i,j}(x,y,z,\theta)$，求解出当前路径反射点位置。整合所有N个传感器M_d条路径的方程组$\left\{f_{m,i,j}(x,y,z,\theta)\right\}_{m=1,i=1,j=1}^{\mathrm{M}_d,\mathrm{N},\mathrm{N}}$，即可表示出整个系统的非线性映射关系，进而解出每个反射点。其中由手部反射的回波才是有效的，我们需要从求解出的每个反射点中根据几何位置关系提取出手部反射点，再重构出手势.

根据以上思路可解析每条超声波路径，并利用几何方法重构出手势框架。但构造出从接收信号幅值向量到手势关键点矩阵的闭式解面临以下挑战：首先，系统中的多路径成分混杂，难以分离开计算每条路径的幅值分量。其次，方程组中涉及的众多已知参数（如介质衰减η）在实际环境中不是常量，无法准确测量。最后，由于系统的多收发和多路径特征，目标方程数量较多，计算量较大。

基于以上分析，从接收信号幅值到手势关键点坐标的非线性映射面临维度高、计算复杂等困难，无法采用几何方法建立明确数值的映射关系。为此，可借助深度学习模型提取回波信号幅值中的特征，学习输入幅值信号与手势坐标之间的复杂关系，实现有效建模。神经网络的映射过程是从图1系统模型中的超声波传感器$\mathrm{S}_i(i=1,2,\cdots,\mathrm{N})$采集幅值信号$\boldsymbol{\alpha}$先提取特征，映射到隐式特征空间，再从特征空间映射到最终的手势坐标$\bar{\boldsymbol{I}}$。神经网络特征提取过程表示为：

$$\boldsymbol{c} = f_{amp}\left(\boldsymbol{\alpha}; \beta_{amp}\right) \tag{6}$$

其中，$\boldsymbol{c}$是特征向量，β_{amp}是幅值参数向量，f_{amp}表示特征映射函数。再根据提取到的特征生成手势关键点坐标：

$$\bar{\boldsymbol{I}} = f_{ges}\left(\boldsymbol{c}; \alpha_{ges}\right) \tag{7}$$

其中，α_{ges}是特征参数向量，f_{ges}表示关键点映射函数。f_{amp}由神经网络结构决定，f_{ges}由神经网络结构和相机拍摄图片的坐标系决定。β_{amp}和α_{ges}均是神经网络参数，由训练学习过程得到。

3　数据处理和手势重建

为了实现从超声波信号重建出手势框架并基于重建框架进行手势识别，我们的工作包括五个部分：超声波数据处理、手势关键点数据处理、数据同步、手势重建和手势识别。其中，超声波数据处理的目的是从回声信号中处理得到幅值信号，作为手势重建映射的输入信号；手势关键点数据处理目的是从相机拍摄的图片得到手势关键点数据真实值，作为手势重建网络的监督标签；数据同步目的是将超声波幅值信号和手势关键点真实值同步；手势重建部分提出CMAT-Net神经网络，实现从超声波幅值到手势关键点的映射；手势识别部分使用三种机器学习算法对重建出的手势框架进行分类。

3.1　超声波数据处理

接收到的回波信号以同相/正交（I/Q, In-phase/Quadrature）数据形式存储于日志文件中。I/Q表示法是一种复数域信号表示方法，将原始回波信号分解为同相分量$I(t)$和正交分量$Q(t)$，构成复包络$S(t) = I(t) + jQ(t)$。在信号接收时间窗口内，一个手势样本对应每个传感器探头采集到长度为l个I/Q数据点，形成一段离散的时间序列，反映了回波信号实验设置距离区间内的时域分布。将原始I/Q数据转换为信号幅度，得到每个采样时刻回波信号的能量强度，计算公式为：

$$A[n] = \sqrt{I[n]^2 + Q[n]^2}, n = 0,1,\cdots,l-1 \tag{8}$$

分别计算每个传感器的幅值序列后，将N个传感器在同一手势样本下采集的幅度序列按探头空间顺序进行拼接，形成一个长度为Nl的一维时序向量。为进一步消除系统设备自身反射回波影响，将其重采样至长度L，记为最终的超声波特征输入$\boldsymbol{\alpha} \in \mathbb{R}^L$。

3.2　手势关键点数据处理

在手势重建任务中，建立统一且稳定的空间参考坐标系是实现多模态数据对齐与坐标生成的基础。本文选择以相机拍摄的RGB图像平面作为手势重建的基准坐标系：该坐标系以图像左上角为原点，横轴向右为正方向，纵轴向下为正方向，单位为像素。在此坐标系下，手势的几何结构通过一组21个手部关键点进行描述。其定义参考了图像领域广泛采用的手部姿态建模标准[17]。这21个关键点包括：1个位于掌心中心，其余20个分布于五根手指，每根手指包含4个关键点，分别对应近端指节、第二指节、第三指节和指尖。如图3所示，这些关键点不仅覆盖了手部的主要运动自由度，还能有效表达手指弯曲、伸展及相对位置关系，为手势的精细重建提供了足够的几何细节。

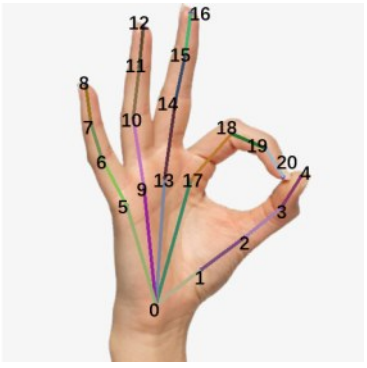


图3　手势关键点

为了从图像中自动提取上述关键点在参考坐标系中的位置，本文采用基于深度学习的端到端姿态估计方法生成关键点坐标作为真实值。相较于人工标注或依赖可穿戴传感器的动作捕捉系统，基于图像的自动化检测方法具有成本低、效率高、可扩展性强等显著优势。具体而言，本文采用OpenPose模型[17]进行手部关键点检测。OpenPose是一种先进的多人姿态估计框架，能够从单帧RGB图像中同时检测人体全身、面部及手部的关键点。其手部分支基于卷积神经网络CNN架构，通过多阶段特征提取与热图回归，实现对21个手部关键点的高精度定位。

OpenPose模型输出的21个关键点的原始像素坐标(x_i, y_i)提供手势重建网络的监督标签。原始图像分辨率为W×H，不同样本中相同手势可能因位置偏移导致坐标值差异较大，不利于神经网络学习

统一的映射关系。为此，将每个关键点的横坐标 x_i 除以图像宽度 W，纵坐标 y_i 除以图像高度 H，得到归一化后的坐标：

$$\widehat{x_i} = \frac{x_i}{\mathrm{W}}, \widehat{y_i} = \frac{y_i}{\mathrm{H}}, i = 1,2,\cdots 21 \tag{9}$$

归一化后，所有关键点坐标被映射至[0,1]区间，消除了图像分辨率和绝对位置的影响，使网络更关注手势的相对结构而非绝对坐标偏移。此外，归一化有助于加速神经网络的收敛过程，减少梯度震荡，提升训练效率与模型鲁棒性。

3.3　数据同步

在多模态数据采集系统中，实现不同传感器之间的精确时间同步是确保后续融合分析准确性的关键步骤。本文研究的系统集成了超声波信号传感器与相机两种传感设备，二者以不同的采样频率独立运行，导致超声波样本数量远少于 RGB 图像帧数，因此必须通过高精度的时间对齐方法，为每一个超声波样本匹配其对应的视觉信息。数据同步操作将超声波信号与相机图片坐标系下手势关键点坐标值的对应，进一步保证了超声波提取特征和坐标值的对应，使式（7）的映射关系确定，这样同步后得到的数据集可以用于神经网络的训练学习从超声波信号到固定坐标值的映射。为实现跨模态数据的时间同步，本文采用基于标准 Unix 时间戳的同步策略。

对于超声波信号，在数据采集过程中，系统实时记录每个超声波样本生成时的本地硬件时间，采集完成后将其转换为标准的 Unix 时间戳（单位：秒，自 1970 年 1 月 1 日 UTC 起始），表示为 $t_{ultra,i}$，其中 i 为第 i 个超声波样本的索引。对于 RGB 视频数据，获取视频中每一帧图像的相对时间戳 τ_j，其中 j 表示第 j 帧图像。随后，获取视频首帧相对于系统参考时间的时间偏移量 Δt。可计算出每一帧 RGB 图像的绝对标准时间戳：

$$t_{rgb,j} = \tau_j + \Delta t \tag{10}$$

此过程确保了视频帧的时间信息能够与超声波数据使用同一时间基准。

完成双模态数据的时间戳标注后，执行时间同步匹配算法。定义一个时间同步阈值 δ_t，作为判断两个事件是否同时发生的容许误差范围。该阈值的选择综合考虑了超声波采样周期、相机帧间隔以及系统时钟抖动等因素，能够在保证匹配唯一性的同时容忍一定程度的时钟漂移与传输延迟。遍历每一个 RGB 图像帧的标准时间戳 $t_{rgb,j}$，对于每个超声波样本的时间戳 $t_{ultra,i}$，若满足：

$$\left| t_{rgb,j} - t_{ultra,i} \right| \leqslant \delta_t \tag{11}$$

则认为该 RGB 帧与第 i 个超声波样本在时间上高度对齐，记为同步匹配结果。

3.4　手势重建神经网络

在手势重建任务中，传统方法通常依赖于纯卷积神经网络或全连接网络进行从传感器信号到关键点坐标的映射。然而，这类架构在处理一维超声波幅度信号时存在明显局限：一方面，CNN 虽擅长提取局部时空特征，但其感受野受限于卷积核尺寸与网络深度，在缺乏显式全局建模机制的情况下难以有效捕捉长距离语义依赖；另一方面，全连接网络虽具备全局感知能力，却忽略了输入信号中的结构化局部模式，且参数量随输入维度急剧增长，不利于轻量化部署。此外，手势姿态本身具有高度非线性与关节间强耦合特性，仅依靠局部特征或孤立的全局表示均难准确还原复杂的二维关键点分布。

针对上述问题，本文提出一种融合局部感知与全局建模能力的新型网络架构 CAMT-Net。3.1 节中传感器个数为 N=3，单个传感器数据长度为 l=76，最终重采样后的超声波数据长度为 L=192。因此需训练得到由长度为 192 的超声波幅度信号生成 2×21 的手势关键点坐标的模型。该模型有效结合了局部特征提取能力与全局上下文建模能力，提升了手势关键点估计的精度。设计的网络框架如图 4 所示。

具体而言，CAMT-Net 首先通过三层堆叠的卷积模块进行多尺度局部特征提取：每层包含卷积、批归一化、ReLU 激活、通道注意力机制与下采样操作，在将序列长度逐步压缩至 24 的同时，通道数扩展至 128，输出特征图维度为[B,128,24]。其中，通道注意力机制基于全局平均池化与全连接网络校准通道权重，增强判别性局部特征的表达能力。

随后，CNN 特征被展平并通过三层 MLP 映射至 512 维紧凑隐空间表示[B,512]。该 MLP 不仅实现维度调整，还通过非线性变换强化特征的语义抽象能力，为后续全局建模奠定基础。

为进一步捕获高维特征间的长程依赖与上下文

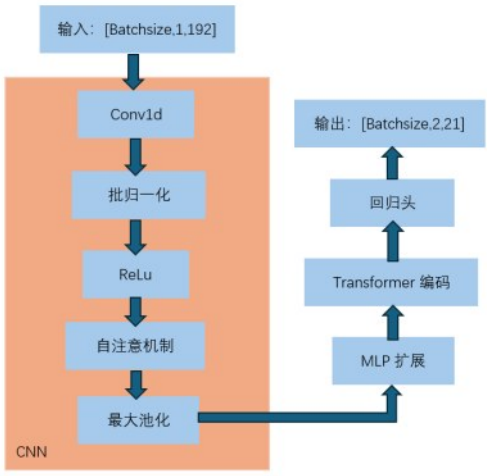


图4 CAMT-Net网络框架

关联，模型引入12层Transformer编码器，其内部的多头自注意力与前馈网络可在特征维度上建模跨通道非线性交互，提升表示的全局一致性。最后，轻量化回归头Transformer输出映射为42维向量，并重构为[B,2,21]格式用于损失计算与可视化。

实验表明，CAMT-Net通过CNN和通道注意力实现局部特征增强、MLP实现语义抽象与Transformer实现全局上下文建模的协同设计，在轻量级参数规模下显著提升了手势关键点估计精度。

3.5 手势识别

为全面评估基于重建关键点的手势分类性能，并兼顾模型的可解释性、非线性建模能力与鲁棒性，本文选取三种代表性机器学习算法进行对比：线性判别分析（LDA, Linear Discriminant Analysis）作为高效且可解释性强的线性基线方法，适用于检验关键点特征的线性可分性；支持向量机（SVM, Support Vector Machine）采用RBF核映射至高维空间，有效处理类别间的非线性边界，具备优异的泛化能力；随机森林（RF, Random Forest）通过集成多棵决策树实现强抗噪性和稳定性。为进一步验证手势重建在特征提取方面的有效性，本文还分别基于原始超声波幅度信号和重建的关键点坐标，使用上述三种算法进行分类实验，对比其识别准确率。

4 实验结果及分析

为了评估本文所提方法进行手势重建和进一步识别的性能，进行了多组实验。本章介绍具体的手势数据集设置、实验设计和实验结果。

4.1 手势数据集

为支持基于超声波信号的手势重建和识别任务，本文构建了一个包含六种常见静态手势的数据集Ultrasonic Gesture。实验设备和数据集设置如下。

超声波信号采集使用的是CH101传感器，这是一款微型低功耗的超声波TOF传感器。使用三个传感器探头收发信号，理论上三个探头竖直摆放成三角形结构，就足够感知到水平方向和竖直方向上的手势信息，实现二维手势重建。实验装置图如图5所示。传感器探头的中心频率设置在85kHz左右，探头放置在距离地面高度约1.5m处，保证对于站立的身高在1.6m到1.9m的成年人受试者，相机和超声波探头以较好的视角和角度获得手势样本。对于本文的短距离高精度手势感知任务，设置该超声波传感器接收回声信号的距离是0.6m以内，接收回声信号的采样频率是3Hz。采集图像使用的设备是奥比中光Femto Bolt深度相机，该相机可以以每秒30帧的频率收集RGB图片，并且可以记录每一帧的时间戳信息用于同步。

图5 实验装置图

本研究共定义6类静态手势，分别为：握拳、点赞、数字3、数字5、数字7和数字8，如图6所示，每类手势赋予唯一的整数标签，依次为0至5，用于后续分类任务。数据采集过程中受试者在距离超声波传感器0.4－0.6m的固定区域内执行手势动作。所有视频以1920×1080分辨率录制，确保图像质量满足OpenPose姿态估计需求。实验共招募3名年龄介于20－25岁的健康成年受试者参与，其手部尺寸、肤色及形态存在一定差异，以增强数据集的个体多样性与模型泛化能力。每位受试者依次完

成 6 类手势，执行手势过程中确保相机拍摄到完整手部，每类持续执行 2 分钟，共采集 18 组原始多模态数据。经预处理后，剔除因遮挡、信号异常或标注失败导致的无效样本，最终获得共计 6360 个有效样本。

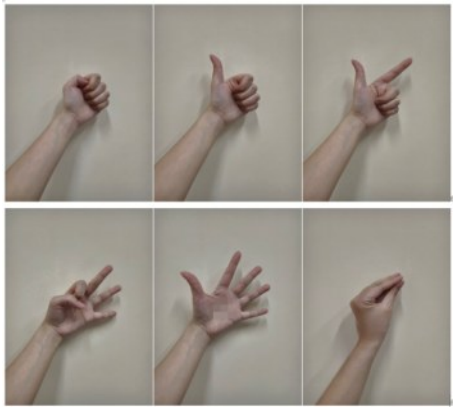


图6　6种静态手势

按照 3.3 节的方法对跨模态数据进行时间同步，阈值设置为 18ms。同步后的时间戳误差数据如表 1 所示，平均同步误差为 8.334 ms，标准差为 4.906 ms，最大误差为 17.972 ms，严格控制在预设的 18 ms 同步阈值以内。同时，95% 分位误差为 16.332 ms，表明绝大多数样本的时间偏差极小且分布集中。同步操作为超声波信号与手势关键点之间建立高保真映射关系提供了可靠的数据基础。

表1　跨模态数据同步误差统计

同步误差统计量	数值（ms）
平均误差	8.334
标准差	4.906
最大误差	17.972
95% 分位数	16.332

为支持手势重建与识别的两阶段任务，数据集采用分层划分策略：其中 4000 个样本用于训练手势重建模型；2000 个样本保留用于手势识别任务的训练与验证；剩余 360 个样本作为独立测试集，统一用于重建与识别性能的最终评估。所有子集均经过随机打乱处理，以避免时间序列相关性与类别偏差。

4.2　实验设计

对于手势重建部分的训练策略如下：采用 Adam 优化器，设置其训练参数为 β_1=0.9，β_2=0.999，ε=1e-7。初始学习率设为 1×10^{-4}。使用余弦退火调度器，让学习率按余弦曲线先快降后慢降再微升。

在手势关键点回归任务中，模型需精确预测 21 个手部关键点的二维坐标。由于该任务对微小误差高度敏感，传统 L1 或 L2 损失难以兼顾精度与鲁棒性：L2 易受异常值干扰，L1 对小误差梯度恒定，不利于精细优化。为此，本文采用 Wing Loss[29]，其在小误差区间使用非线性惩罚以增强敏感性，在大误差区域退化为线性形式以保持鲁棒性。Wing loss 公式如下：

$$\mathrm{Wing}(d)=\begin{cases} w\ln(1+d/\varepsilon), & d\le w \\ d-C, & d>w \end{cases} \tag{12}$$

其中 d 是预测坐标值与真实坐标值的绝对误差，w 是非线性区间阈值，决定小误差范围，实验中将该值设置为 10，ε 是控制对数曲率，设置为 2，C = w - wln(1 + w/ε)用于保证分段点连续。

此外，超声波信号采集过程手势不完整和 OpenPose 标注的局限可能导致部分关键点不可靠。为此，引入掩码机制，仅对有效关键点计算 Wing Loss。该带掩码的 Wing Loss 有效抑制了噪声标签干扰，确保梯度更新基于可靠监督，显著提升了模型在非理想条件下的训练稳定性与收敛精度。最终的损失函数表示为：

$$\mathrm{Loss}=\sum_{i=1}^{N_p} m_i\cdot \mathrm{Wing}(d) \tag{13}$$

其中，Np 表示关键点总数，$m_i\in\{0,1\}$为二值掩码，指示第 i 个关键点是否参与损失计算。

手势重建神经网络在 NVIDIA GPU 上端到端训练 200 个 epoch，批大小设为 64。训练过程中每 10 个 epoch 在验证集上计算一次平均损失，用于监控过拟合。训练结束后将模型的参数权重及损失日志保存至本地文件系统。

手势识别部分的实验设置了原始超声波信号和重建手势框架两种识别特征，针对以上两种识别特征分别用 SVM、LDA 和 RF 三种机器学习算法进行分类，进行识别准确率的对比。为研究单个用户手势识别的准确率，本文以受试者 3 为例，从原始手势识别数据集中提取其全部样本，原始训练样本数为 667，测试样本数为 120。为进一步扩充数据规模，对所提取的手势关键点数据实施镜像操作：具体而言，针对每个手势样本的关键点坐标，在保持

纵坐标不变的前提下，将横坐标沿图像坐标系的垂直中轴线进行反转，即对每个关键点(x,y)映射为$(x_w - x,y)$（x_w是图片宽度），从而生成该手势的左右镜像版本。此操作仅作用于手势关键点坐标，对应的超声波传感数据保持不变。通过该镜像处理，受试者3的训练样本数量由667增至1334，测试样本数量由120增至240。随后，利用镜像翻倍后的数据集训练面向该单一受试者的手势识别机器学习模型，并在扩充后的测试集上评估其识别准确率。该镜像增强策略是一种常用的数据扩充方法，可有效缓解小样本场景下的过拟合问题，同时保留原始数据的时序与结构特性。

4.3　实验评估

在手势重建部分的实验中，我们使用关键点正确概率（Probability of Correct Keypoints，PCK）[30]对结果进行定量比较。PCK定义了预测关键点与其真实位置的距离在阈值σ范围内的概率。对于特定关键点p，我们将其表示为$\mathrm{PCK}_{\sigma}^{p}$，并在验证数据集上近似计算该指标，PCK的计算公式如下：

$$\mathrm{PCK}_{\sigma}^{p} = \frac{1}{}\sum\delta\left(\left\|\mathbf{x}_p - \mathbf{y}_p\right\| < \sigma\right) \tag{14}$$

其中，$\mathbf{x}_p$表示第p个关键点的预测位置，$\mathbf{y}_p$是其真实位置，$\delta(\cdot)$为指示函数，当其内部的条件为真时返回1，否则返回0。σ是归一化距离，在大多数基于图像方法的手势重建工作中，选定归一化距离是截取到手部区域边界框的尺寸，参考这一尺寸，我们将归一化距离定为每个样本手势掌心到中指根部距离的1.5倍。

在采集数据时相机和传感器的位置存在偏差，并且两种数据是跨模态的，这两种因素会导致由超声波信号重建出的手势框架与RGB图片中手势的绝对位置存在偏差。为了解决这个问题，在手势重建的测试阶段，每个样本手势加入掌心实际位置作为先验信息，消除位置偏差。最终部分手势重建结果如图7。

图8中给出两个受试者数据集和三个受试者数据集对应的手势重建PCK曲线，并将这两条曲线和图像方法的结果进行对比。可以观察到。受试者数量增加时，模型得到更充分的训练，三个受试者实验PCK曲线明显高于两个受试者的结果。Wang等人[18]图像方法重建手势工作中PCK曲线在归一化距离为0.2处收敛，$\mathrm{PCK}_{0.2}$=0.9。本文手势重建的PCK曲线也在归一化距离0.2处收敛，$\mathrm{PCK}_{0.2}$=0.84。

图7　部分手势重建结果

为验证所提出CAMT-Net模型的有效性，本文将其与三种基线模型在相同数据集上进行对比实验。本文选取三种具有代表性的基线模型进行对比：CNN、CNN+LSTM和CNN+自注意力机制。其中，CNN仅依赖多层卷积与池化操作提取局部特征，作为衡量局部建模能力的基准；CNN+LSTM在CNN基础上引入长短期记忆网络，用于验证时序动态建模在本任务中的有效性；而CNN+自注意力机制则通过在特征序列上加入自注意力操作，显式建模不同位置间的空间长程依赖，是当前主流的轻量级全局上下文建模范式。基线模型和所提出CAMT-Net用于手势重建任务得到PCK曲线的对比如图9。

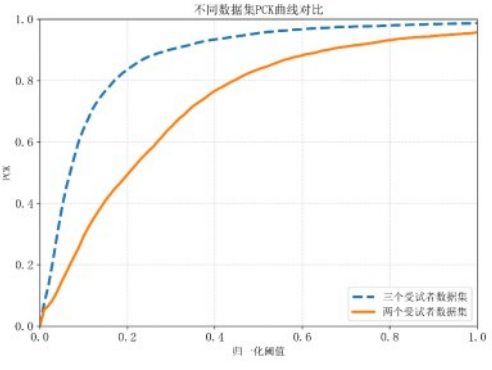


图8　不同数据集PCK曲线对比

如图9所示的PCK曲线对比表明，CAMT-Net在全阈值范围内均显著优于三种基线模型。具体而言，CNN由于缺乏对全局上下文关系的建模能力，重建结果准确度受限；CNN+LSTM的性能甚至低于纯CNN，说明超声波幅度信号中关键判别信息主要体现为静态空间特征分布，而非具有明确因果

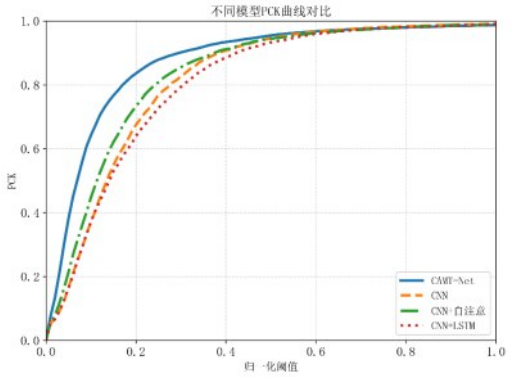


图9　不同模型PCK曲线对比

依赖的时序动态过程，强行引入循环结构不仅未能提升性能，反而因不匹配的数据先验而损害泛化能力；CNN+自注意力机制通过在特征序列上显式建模不同位置间的长程依赖，一定程度上弥补了纯卷积模型的局限，验证了全局感知的重要性，但其仅依赖单一层次的注意力操作，难以充分挖掘高维特征中的深层语义关联。相比之下，CAMT-Net通过通道注意力机制增强关键局部特征，利用MLP实现紧凑语义抽象，并进一步借助Transformer编码器在特征维度上建模跨通道非线性交互，实现了局部感知与全局语义建模的协同优化。该设计更契合本任务中超声波信号的结构特性，从而显著提升了手势关键点估计的精度。

本文所构建的手势重建模型在精度与效率之间取得良好平衡，具备良好的实用性与部署潜力。模型序列化后占115.08 MB，结构紧凑，适用于边缘或移动端设备；在GPU加速下，完成200个训练epoch总耗时408.40秒（约2.04秒/epoch），收敛速度快。结合其在关键点回归任务中的高精度表现，该模型适用于实际人机交互系统的部署应用。

对手势识别的结果比较识别准确率。识别准确率RA定义为测试集中正确识别样本数与测试集总样本数的比值：

$$\mathrm{RA}=\left(\frac{n_{right}}{n_{all}}\right)\times 100\% \tag{15}$$

RA值越接近1，识别效果就越好。对三位受试者的手势重建结果分别应用SVM、LDA和RF三种机器学习算法进行分类，再对采集的超声波幅度数据使用同样的分类算法进行对比。三位受试者手势数据的识别准确率如图10所示。可以观察到，SVM算法对手势重建框架的分类准确度最高，达到86%；RF算法对超声波幅度数据的分类准确度最高，达到80%。使用三种机器学习算法分类结果中，对手势重建框架分类的准确度均高于对超声波数据分类的准确度，这表明了手势重建操作提取到了更多深层次的手势信息。

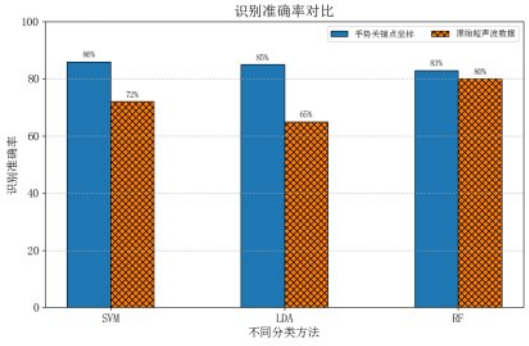


图10　三个受试者手势识别准确率对比

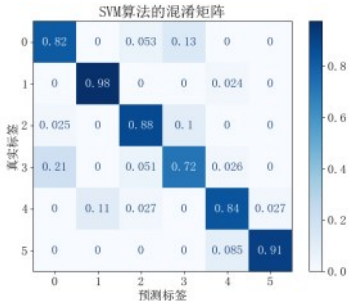
为了更直观地展示各手势的识别效果，给出了手势识别结果的混淆矩阵，如图11所示。

实验发现，手势重建结果在三位受试者间存在个体差异：手指较长的受试者在执行不同手势时特征更显著，超声波回声包含的判别信息更丰富，重建错误率更低。其中，受试者3的手势特征最为明显，故对其数据镜像扩充单独进行识别准确率分析。受试者3手势识别准确率如图12所示。

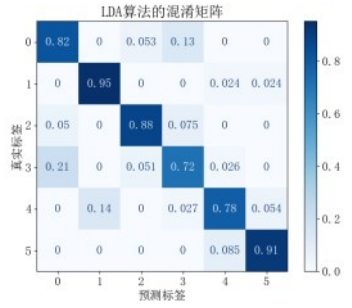
可以看到，对受试者3的手势重建框架进行识别，SVM算法下识别准确率可以达到89%，RF算法下对原始超声波幅度数据分类的准确率可以达到82%。与三位受试者的实验结果相比，识别准确度有了显著的提升。

5　结束语

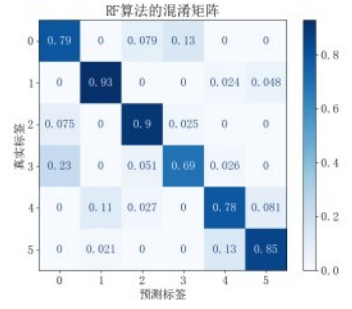
本文针对过去手势感知中存在的隐私泄露与细粒度不足问题，提出了一种基于超声波信号的手势重建与识别方法。首先，从物理层面分析了超声波与手部姿态的映射原理。然后构建了完整的处理流程，包括超声波回声信号到幅值序列的提取、基于视觉的关键点标注、多模态数据同步。在此基础上，设计了融合卷积网络、通道注意力机制与Transformer编码器的轻量级神经网络CAMT-Net，实现了从一维超声波信号到二维手势关键点的高精度端到端映射。实验表明，所提方法在六类静态手势上重建精度接近RGB图像方法，基于重建结果的手势识别准确率达89%，验证了超声波信号在细



（a） SVM 算法



（b） LDA 算法



（c） RF 算法

图 11　SVM/LDA/RF 算法的混淆矩阵

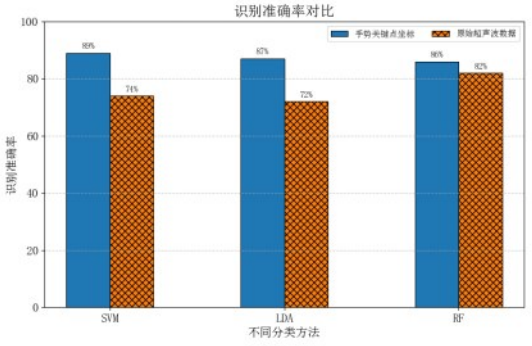


图 12　单个受试者手势识别准确度对比

粒度手势感知中的有效性。需要指出的是，本文聚焦于静态手势的二维重建与识别，未来工作可拓展至三维感知和动态连续手势的时序建模，进一步提升系统在真实交互场景中的鲁棒性与实用性。

参考文献：

[1] GAVGIOTAKI D, NTOA S, MARGETIS G, et al. Gesture-based Interaction for AR Systems: A Short Review[C]//Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments. New York: ACM Press, 2023: 284-292.

[2] GAO Q, LIU J G, JU Z J. Robust real-time hand detection and localization for space human – robot interaction based on deep learning[J]. Neurocomputing, 2020, 390: 198 – 206.

[3] WANG R Y, POPOVIĆ J. Real-time hand-tracking with a color glove[J]. ACM Transactions on Graphics, 2009, 28(3): 1-8.

[4] HU F, HE P, XU S L, et al. FingerTrak: Continuous 3D Hand Pose Tracking by Deep Learning Hand Silhouettes Captured by Miniature Thermal Cameras on Wrist[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2020, 4 (2): 1-24.

[5] ZHANG F, ZHAO L, LI S, et al. 3D hand pose and shape estimation from monocular RGB via efficient 2D cues[J]. Computational Visual Media, 2024, 10(1): 79 – 96.

[6] MUELLER F, BERNARD F, SOTNYCHENKO O, et al. GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 49 – 59.

[7] SMITH B, WU C, WEN H, et al. Constraining dense hand surface tracking with elasticity[J]. ACM Transactions on Graphics, 2020, 39(6): 1-14.

[8] LIEN J, GILLIAN N, KARAGOZLER M E, et al. Soli: ubiquitous gesture sensing with millimeter wave radar[J]. ACM Transactions on Graphics, 2016, 35(4): 1-19.

[9] GLANDON A, VIDYARATNE L, SADEGHZADEHYAZDI N, et al. 3D Skeleton Estimation and Human Identity Recognition Using Lidar Full Motion Video[C]//2019 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2019: 1 – 8.

[10] MAO W G, HE J, ZHENG H H, et al. High-precision acoustic motion tracking: demo[C]//Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking. Piscataway: IEEE Press, 2016: 491-492.

[11] SHARP T, KESKIN C, ROBERTSON D, et al. Accurate, Robust, and Flexible Real-time Hand Tracking[C]//Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. New York: ACM Press, 2015: 3633-3642.

[12] SRIDHAR S, MUELLER F, OULASVIRTA A, et al. Fast and robust hand tracking using detection-guided optimization[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 3213-3221.

[13] SRIDHAR S, OULASVIRTA A, THEOBALT C, et al. Interactive

Markerless Articulated Hand Motion Tracking Using RGB and Depth Data[C]//2013 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2013: 2456 – 2463.

[14] KESKIN C, KIRAÇ F, KARA Y E, et al. Hand Pose Estimation and Hand Shape Classification Using Multi-layered Randomized Decision Forests[C]//Computer Vision – ECCV 2012. Piscataway: IEEE Press, 2012: 852 – 863.

[15] TANG D, CHANG H J, TEJANI A, et al. Latent Regression Forest: Structured Estimation of 3D Articulated Hand Posture[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 3786 – 3793.

[16] TOMPSON J, STEIN M, LECUN Y, et al. Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks [J]. ACM Transactions on Graphics, 2014, 33(5): 1-10.

[17] SIMON T, JOO H, MATTHEWS I, et al. Hand Keypoint Detection in Single Images Using Multiview Bootstrapping[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 4645 – 4653.

[18] WANG Y G, PENG C, LIU Y. Mask-Pose Cascaded CNN for 2D Hand Pose Estimation From Single Color Image[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(11): 3258 – 3268.

[19] CHEN X, LIU Y, DONG Y, et al. MobRecon: Mobile-Friendly Hand Mesh Reconstruction from Monocular Image[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 20512 – 20522.

[20] CHEN X, LIU Y, MA C, et al. Camera-Space Hand Mesh Recovery via Semantic Aggregation and Adaptive 2D-1D Registration [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 13269 – 13278.

[21] TANG X, WANG T, FU C W. Towards Accurate Alignment in Real-time 3D Hand-Mesh Reconstruction[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 11678 – 11687.

[22] KONG H, LYU H, YU J, et al. mmHand: 3D Hand Pose Estimation Leveraging mmWave Signals[C]//2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS). Piscataway: IEEE Press, 2024: 1062 – 1073.

[23] JI S J, ZHANG X Y, ZHENG Y Q, et al. Construct 3D Hand Skeleton with Commercial WiFi[C]//Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems. New York: ACM Press, 2024: 322-334.

[24] LI D, LIU J L, LEE S I, et al. FM-track: pushing the limits of contactless multi-target tracking using acoustic signals[C]// Proceedings of the 18th Conference on Embedded Networked Sensor Systems. New York: ACM Press, 2020: 150-163.

[25] MAO W G, WANG M, SUN W, et al. RNN-Based Room Scale Hand Motion Tracking[C]//The 25th Annual International Conference on Mobile Computing and Networking. New York: ACM Press, 2019: 1-16.

[26] WANG W, LIU A X, SUN K. Device-free gesture tracking using acoustic signals[C]//Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking. New York: ACM Press, 2016: 82-94.

[27] YANG Y J, CHEN T, AN Z L, et al. LeakyFeeder: In-Air Gesture Control Through Leaky Acoustic Waves[C]//Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems. New York: ACM Press, 2025: 144 – 157.

[28] WANG S Y, PU H L, CAO Q M, et al. RAM-Hand: Robust Acoustic Multi-Hand Pose Reconstruction Using a Microphone Array [C] //Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems. New York: ACM Press, 2025: 130 – 143.

[29] FENG Z H, KITTLER J, AWAIS M, et al. Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2235 – 2245.

[30] YANG Y, RAMANAN D. Articulated Human Detection with Flexible Mixtures of Parts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2878 – 2890.