VT2R：视频与文本驱动的大规模毫米波雷达数据生成方法

邓凯凯[①] 凌 月[②] 邢 玲*[①] 吴红海[①] 赵 东[②] 马华红[①]

[①](河南科技大学信息工程学院 洛阳 471023)

[②](北京邮电大学网络与交换技术国家重点实验室 北京 100876)

摘 要：基于毫米波雷达的手势识别已经成为一种很有前景的人机交互方式并应用于多个领域。然而，大规模训练数据的不足严重阻碍了开发鲁棒和普适的深度学习模型，尤其在用户躺姿场景更为显著。现有的毫米波雷达数据生成方法由于缺乏足够的数据源而效果不佳。为此，本文提出了一种新颖的雷达数据生成系统VT2R，其利用视频或文本数据生成大规模逼真的雷达数据，解决了视频和文本到雷达数据的映射关系难构建的关键问题。VT2R由四个组件组成：视频特征编码网络、文本特征编码网络和雷达特征编码网络分别建模三种模态输入的特征表示，随后数据拟合与解码网络基于变分自编码器机制对潜在分布进行对齐与解码，从而生成逼真的大规模雷达数据。最后，在生成与自采数据集上进行了广泛的实验验证，结果表明VT2R在识别躺姿场景下的手势方面显著优于现有最先进方法。

关键词：毫米波雷达感知；数据生成；手势识别；变分自编码器

中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2026)00-0001-14

DOI: 10.11999/JEIT260240 **CSTR**: 32379.14.JEIT260240

1 引言

近年来，基于毫米波雷达(Millimeter-wave Radar, mmWave)的手势识别在智能人机交互领域中展现出广阔的应用前景，因其对光照变化不敏感，具备良好的隐私保护能力，并且支持普适的人体感知[1–5]。然而，现有研究[6–10]主要集中在用户站姿或坐姿下的手势识别，对于用户处于躺姿状态的毫米波手势交互关注较少，相应数据集也十分稀缺。在智能家居场景中，用户可在躺卧状态下通过非接触式手势完成照明控制、窗帘开合或空调温度调节等操作[11–12]，具有显著的实际需求。因此，针对躺姿场景开展研究有助于拓展毫米波手势识别的应用范围。

当前的手势识别方法多依赖深度学习模型，因此获取大规模的雷达数据对训练鲁棒和泛化的模型至关重要。实验表明，仅使用站姿数据训练的模型在躺姿场景中的准确率由94.65%降至48.32%，主要原因是不同姿势下，反射毫米波信号的身体区域不同，导致点云特征差异较大；即使逐步引入躺姿样本，当规模增加至6000个时，准确率也仅提升至93.69%，仍难以满足实际部署需求。相比于计算机视觉[13]和语音领域[14]，手势识别相关的雷达数据集[6–7]的规模和多样性明显不足，并且大多聚焦于站姿和坐姿。同时，大规模的数据采集和标注是耗时耗力且易出错的[15]。已有研究[16–18]表明，利用视频、动作捕捉或文本等模态生成感知数据已成为可行方向，为通过模态转换缓解雷达数据稀缺问题提供了新的思路。

一些工作[18–20]利用动作捕捉数据、深度相机数据或人体网格数据生成雷达数据，但生成结果仍存在数据稀疏或常见手势覆盖不足的问题。同时，一些工作[21–22]利用生成对抗网络(GAN)增强采集的雷达数据集，但是相似动作的混淆问题导致识别准确率显著下降；另一些工作[23–24]利用生成扩散模型合成雷达数据，但是它们无法生成具有空间特征的三维点云，或在躺姿等特定任务中表现不佳。此外，一些工作[25–31]将2D视频转换为粗粒度的频谱数据或细粒度三维点云数据，从而训练泛化的识别模型。然而，这些方法高度依赖丰富的数据源，在躺姿等公开视频较为稀缺的场景下表现不佳。因此，本文提出一种视频与文本驱动的雷达数据生成方法

收稿日期：2026-03-05；改回日期：2026-06-30；网络出版：2026-07-02
*通信作者： 邢玲 xingling_my@haust.edu.cn
基金项目：国家自然基金(No.U23A20272, No.62272146)；河南省杰出青年科学基金滚动支持项目(No.252300421237)；中原人才计划项目(264000510008, 264200510018)；河南省重点研发专项(No.251111210900, 261111240300)；河南省高校科技创新团队支持计划(No.26IRTSTHNO05)；中国博士后科学基金面上项目(No.2025M783507)

Foundation Items: The National Natural Science Foundation of China (No.U23A20272, No.62272146), in part by the Natural Science Foundation of Henan Province (No.252300421237), Zhongyuan Talent Program Project (264000510008, 264200510018), Key Research and Development Special Project of Henan Province (No.251111210900, 261111240300), the Program for Innovative Research Team in University of Henan Province (No.26IRTSTHNO05), and in part by the China Postdoctoral Science Foundation (No.2025M783507)

VT2R，实现大规模逼真的雷达数据生成，但是本文将面临如下关键挑战：

如何准确地构建视频和文本到雷达数据的映射关系？相比于展现用户动作随时间变化和环境背景的视频数据，以及具备丰富表达能力的文本描述，雷达点云数据主要关注手势在三维空间中的动态变化，具有高度稀疏和无序的特性，这使其难以被直观理解或高效建模，特别是在语义层面难以与其他模态实现对齐。尽管已有研究[32]利用视觉-语言预训练模型对齐文本和雷达数据的语义空间，但其仅支持活动类别级的粗粒度匹配，难以还原具备时空特性的高保真雷达数据。如果无法构建精确的跨模态映射关系，将影响生成数据的质量与下游任务的性能。

为解决跨模态映射难题，本文提出VT2R系统。具体来讲，视频特征编码网络与文本特征编码网络分别基于CLIP模型提取具有时间一致性的视觉表示与可对齐的语义特征，利用其在大规模图文预训练中形成的共享语义流行，为稀疏、无序且语义表达不直观的雷达点云数据提供可迁移的语义先验，为跨模态映射提供语义对齐基础；雷达编码网络则通过层次化时空建模学习点云的结构与动态信息。在基于变分自编码网络(VAE)的数据拟合与解码网络中，多模态特征被映射至统一的潜在分布空间，并通过重参数化采样解码为雷达数据，训练过程中联合优化重建损失、Kullback-Leibler(KL)散度损失与跨模态相似度损失。该框架支持以视频或文本为条件生成雷达数据，同时可结合真实样本进行增强式重建，从而实现高逼真、大规模的数据生成并提升模型泛化能力。

本文的主要贡献包括以下3个方面。

(1)提出了数据生成系统VT2R，旨在解决用户躺姿执行手势场景下雷达数据严重缺乏的问题。

(2)设计了由视频特征编码、文本特征编码、雷达特征编码及数据拟合与解码网络构成的统一生成框架，支持以视频、文本或雷达数据为输入，生成大规模且逼真的雷达训练数据。

(3)构建了首个面向用户躺姿手势识别的雷达点云数据集，涵盖5类手势、32位参与者，共计14400个样本。基于该数据集的实验结果表明，仅使用生成雷达数据训练时，VT2R可取得89.2%的识别准确率，较代表性基线方法RFGen[23]提升33.88%；当结合少量真实雷达数据进行联合训练时，准确率进一步提升至97.62%，相较RFGen提高21.48%。

2 相关工作

2.1 毫米波雷达数据生成

许多工作[18–32]利用不同的数据源生成雷达数据。基于动作捕捉、深度相机或人体网格的方法[18–20]往往生成稀疏的雷达数据或难以覆盖常见用户手势。Rahman等人[22]结合对抗域自适应和运动约束生成微多普勒信号。但基于GAN的方法[21–22]在相似手势间易引入频谱混淆。近年来，扩散模型被引入雷达生成任务中，Chi等人[24]利用扩散模型从时域、频域和复数域对毫米波信号建模，但仍无法生成3D点云数据。此外，公开的视频数据丰富多样，且涵盖很多复杂的场景，近年来被广泛用于雷达数据的生成[25–31]。Ahuja等人[25]提取2D视频中人体顶点的速度和雷达横截面(RCS)生成频谱数据。Zhang等人[26]提取人体骨骼点信息，模拟毫米波信号的传播过程生成雷达数据。Deng等人[27]进一步模拟毫米波信号的多径反射和衰减现象生成逼真的雷达数据。Li等人[29]结合射线追踪与电磁计算，估算细粒度RCS以生成雷达数据。但是，这些方法通常难以表征手势的空间结构或细粒度特征。为此，Deng等人[30]模拟用户手势的多样化和细粒度反射特性来生成3D点云数据。Ling等人[31]进一步模拟人和背景的复杂反射特性来生成动态场景下的点云数据。然而，这些方法高度依赖丰富的数据源，在视频数据较为稀缺的躺姿场景中效果受限。相比之下，本文融合视频与文本的表达优势，结合VAE引导躺姿场景下大规模逼真雷达点云的生成。

2.2 基于毫米波雷达的手势识别

已有大量工作探索基于多种传感信号的手势识别方法，包括视频[34]、语音信号[35]、射频信号[36]以及WiFi信号[37]等。然而，它们存在固有限制：对环境光照变化敏感、计算资源消耗高、泄露用户隐私、信号衰减以及感知距离受限。近年来，由于毫米波雷达能够克服这些限制，逐渐成为研究热点。张等人[38]分析了现有手势识别技术的进展，并阐述了手势交互在智能人机交互中的重要性和研究趋势。Hayashi等人[39]结合卷积神经网络与长短期记忆网络进行端到端手势识别；Liu等人[7]针对复杂室内反射环境对毫米波传播的干扰，设计了环境自适应的识别网络。这些方法依赖大规模的雷达数据训练，但相应的采集与标注成本高昂。同时，由于用户、环境与位置等“跨域因素”的影响，不同数据域的数据分布差异显著，影响模型泛化能力。因此，一些工作引入小样本学习策略实现快速适应。Liu等人[9]采用少量数据实现对新用户的准确识别。Liu等人[6]利用少样本进行迁移学习，实现跨用户和跨环境的准确识别。然而，它们在面对多域组合时仍易遭受灾难性遗忘。相比之下，本文的工作通过利用视频和文本数据生成大规模、逼真的雷达数据来提升感知模型在跨域识别任务中的泛化能力。

3　VT2R方法

如图1所示，传统雷达数据采集依赖真实传感器部署与人工标注，不仅成本高且难以覆盖多样化场景，限制了数据规模与模型泛化能力。为此，本文提出雷达点云生成系统VT2R，由视频、文本与雷达特征编码网络以及数据拟合与解码网络构成。

3.1　视频特征编码网络

鉴于视频可表示为随时间变化的图像帧序列，从每段视频中均匀采样n帧图像，记为$v_i=\{v_i^1, v_i^2, \cdots, v_i^n\}$。如图2所示，每帧图像被划分成尺寸为32×32的不重叠图像块，并经线性映射与位置编码后输入CLIP[40]图像编码器(ViT-B)以提取帧级视觉特征，使视频编码为与文本语义可对齐的中间表示，从而为视频与稀疏雷达点云之间建立更稳定的映射基础。后续数据拟合与解码网络则在VAE框架下通过潜在分布对齐，将这些高维特征进一步映射至与雷达点云结构对应的潜在空间。考虑到原始ViT无法充分捕捉手势动作的细粒度特征，本文引入适配器作为轻量化调优模块，并输出图像的特征表示$Z_i=\{Z_i^1, Z_i^2, \cdots, Z_i^n\}$。最后，为了建模视频帧之间的时间依赖性，将帧级图像特征Z_i与对应的时间编码T相加，并输入由六层Transformer构成的时间帧融合层F，这个过程可以表示为：

$$Z_{\mathrm{V}_i}=F\left(Z_i+T\right) \tag{1}$$

如图3所示，针对ViT模型的适配器位于每个MLP子模块中，先采用向下投影层(参数为$W_{\boldsymbol{down}}^{\mathrm{d}\times d'}$)压缩维度，经ReLU层激活后再通过向上投影层(参数为$W_{\mathrm{up}}^{d'\times d}$)还原维度，其中$d'\ll d$。

3.2　文本特征编码网络

本文采用CLIP文本编码器提取提示语特征，

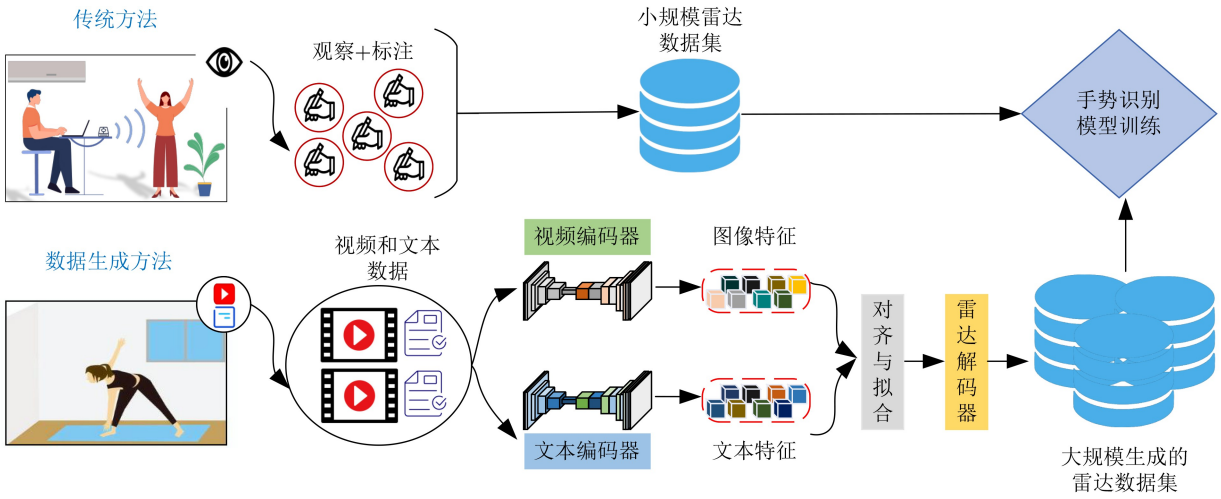


图 1 VT2R的整体框架

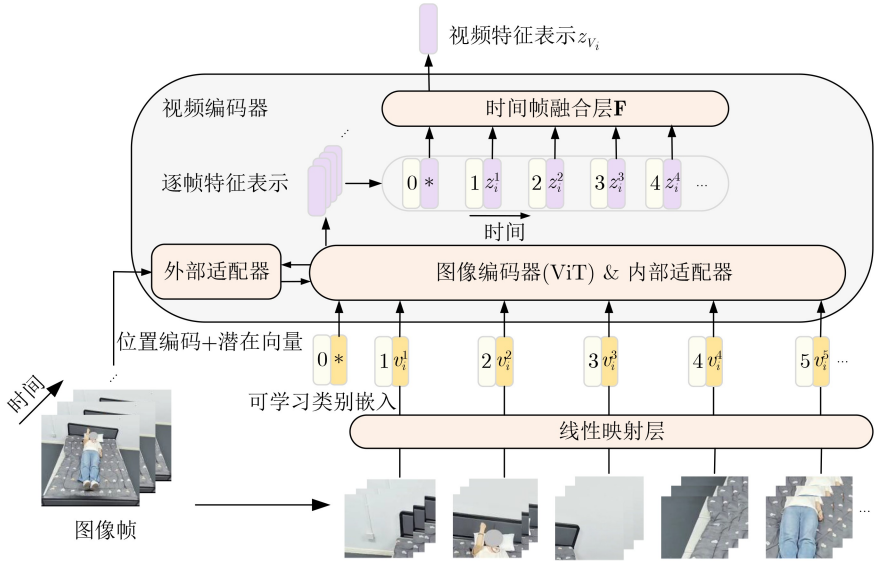


图 2 视频特征编码网络

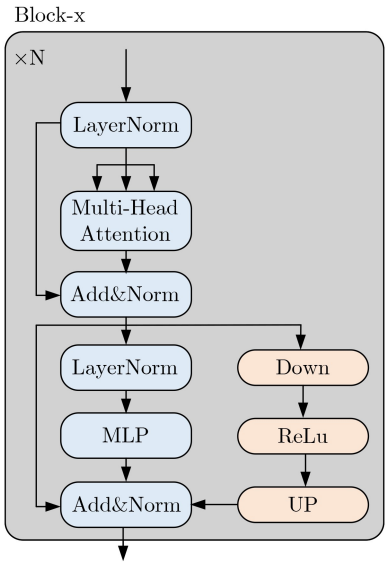


图 3 ViT模型适配器

并在训练过程中冻结其参数，以充分保留预训练带来的语义建模能力。传统自由文本描述难以突出影响雷达点云分布的关键因素，如手势类别、用户姿势及雷达参数等。为此，本文设计了一个结构化模板，枚举这些关键信息进行显式编码。下面举例说明，模板中的标签类型用斜体显示，标签本身用下划线显示：“The *action* is stretching right arm in front of chest and retracting it, where *posture* is lying, *radarheight* is 1.8 m, *distance* is 1 m, *angle* is facing directly, and *scene* is meeting room.”对于真实场景中的自由文本输入，可通过文本解析自动映射至模板字段，从而保证方法的高效性。

3.3 雷达特征编码网络

如图4所示，为了有效建模雷达点云序列的时空特征，本文设计了一种层次化的雷达特征编码网络，包含局部邻域构建、时空特征提取和全局上下文建模三个部分。

(1)**局部邻域构建**

每个手势样本包括不超过30帧(采样频率10 Hz, 3 s)点云，且每帧点数少于64个，可表示为大小为$T\times N\times 3$的张量，其中，T表示帧数，N表示每帧的点数，每个点由三维空间坐标(x,y,z)描述。为每个点p_t^i构建局部时空邻域，在当前帧t及其前后相邻帧中，基于欧氏距离选取空间上最近的k个邻居点，构成四维邻域$\mathcal{N}(p_t^i)$。

(2)**时空特征提取**

为了提取高阶的局部时空特征，在邻域内计算各点相对于中心点的时空偏移，并通过共享MLP编码得到局部特征表示：

$$f_t^i=\mathrm{MLP}(\{p_r^j-p_r^i\}_{p_r^j\in\mathcal{N}(p_t^i)}) \tag{2}$$

随后，将每帧的点特征进行体素化映射，统一为张量$\boldsymbol{X}\in\mathbb{R}^{T\times H\times W\times d}$，其中，$H$和$W$为空间网格

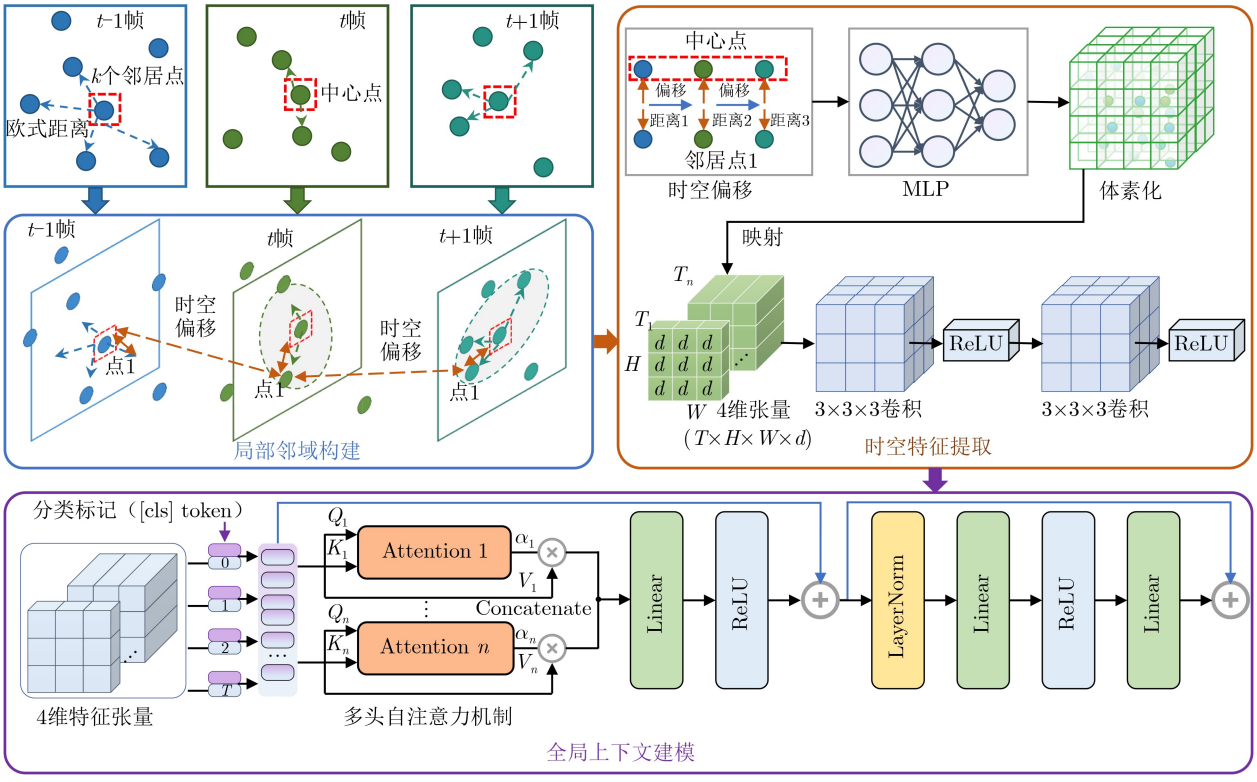


图 4 雷达特征编解码网络

尺寸，d为特征维度。在该张量上堆叠两个3×3×3的三维卷积层，每层后接ReLU激活函数，提取跨帧的局部结构变化与动态特性。

(3)**全局上下文建模**

为了进一步建模全局上下文依赖，在特征序列前插入一个可学习的分类标记([cls] token)，并将其与卷积模块的输出一同输入至Transformer模块中。在VAE框架下，[cls] token的输出进一步映射为潜在分布的参数，即均值μ^{R}和方差Σ^{R}。该层次化结构从局部几何到全局时空的逐级建模，有效捕捉动作的动态演化，并将稀疏雷达点云映射至语义一致的潜在空间，为跨模态生成提供基础。

3.4 数据拟合与解码网络

如图5所示，采用全连接层将视频特征表示Z_{V_i}和文本特征表示Z_{T_i}分别映射为高斯分布的均值和方差。因此，视频、文本和雷达的潜在分布分别表示为$\phi^{\mathrm{V}}=N(\mu^{\mathrm{V}},\Sigma^{\mathrm{V}})$、$\phi^{\mathrm{T}}=N(\mu^{\mathrm{T}},\Sigma^{\mathrm{T}})$和$\phi^{\mathrm{R}}=N(\mu^{\mathrm{R}},\Sigma^{\mathrm{R}})$。本文采用重参数化策略从上述分布中分别采样潜在向量$z\in\mathbb{R}^d$，并输入至解码器以重建对应的雷达点云序列。解码器基于PU-GAN[41]的点云上采样模块，通过局部特征扰动逐级恢复稠密且结构连续的点云，并采用多级上采样结构提升空间分辨率；在最后一级引入3D卷积完成尺寸对齐。训练过程中，需联合优化以下三类损失函数：

(1)**重建损失**

将雷达解码器重建后得到的序列$\hat{V}^{1:30}$、$\hat{T}^{1:30}$和$\hat{R}^{1:30}$与真实雷达序列$R^{1:30}$进行对齐，计算其平均绝对误差L_1作为重建损失：

$$L_{\mathrm{R}}=L_1(\hat{V}^{1:30},R^{1:30})+L_1(\hat{T}^{1:30},R^{1:30})+L_1(\hat{R}^{1:30},R^{1:30}) \tag{3}$$

(2)KL**损失**

为缩小多模态之间的潜在分布差异，引入KL散度作为约束，分别最小化视频-雷达与文本-雷达之间的分布距离并将各模态的分布正则化至标准正态分布$\psi=(0,1)$：

$$L_{\mathrm{KL}}{}^{\mathrm{V}}=\mathrm{KL}(\varnothing^{\mathrm{V}},\varnothing^{\mathrm{R}})+\mathrm{KL}(\varnothing^{\mathrm{R}},\varnothing^{\mathrm{V}})+\mathrm{KL}(\varnothing^{\mathrm{V}},\psi)+\mathrm{KL}(\varnothing^{\mathrm{R}},\psi) \tag{4}$$

$$L_{\mathrm{KL}}{}^{\mathrm{T}}=\mathrm{KL}(\varnothing^{\mathrm{T}},\varnothing^{\mathrm{R}})+\mathrm{KL}(\varnothing^{\mathrm{R}},\varnothing^{\mathrm{T}})+\mathrm{KL}(\varnothing^{\mathrm{T}},\psi)+\mathrm{KL}(\varnothing^{\mathrm{R}},\psi) \tag{5}$$

$$L_{\mathrm{KL}}=L_{\mathrm{KL}}{}^{\mathrm{V}}+L_{\mathrm{KL}}{}^{\mathrm{T}} \tag{6}$$

(3) **跨模态向量相似度损失**

受对比学习启发，本文设计了基于InfoNCE的跨模态相似度损失以增强不同模态潜在空间的一致性。给定由N个样本组成的一批潜在向量$(z_1^{\mathrm{V}},z_1^{\mathrm{T}},z_1^{\mathrm{R}},),(z_2^{\mathrm{V}},z_2^{\mathrm{T}},z_2^{\mathrm{R}},),\cdots,(z_N^{\mathrm{V}},z_N^{\mathrm{T}},z_N^{\mathrm{R}},)$，将不同索引下的样本对$(z_i^{\mathrm{V}},z_j^{\mathrm{R}})$和$(z_i^{\mathrm{T}},z_j^{\mathrm{R}})$视为负样本对。定义视频-雷达与文本-雷达的余弦相似度矩阵为$S_{ij}=\cos(z_i^{\mathrm{V}},z_j^{\mathrm{R}})$与$W_{ij}=\cos(z_i^{\mathrm{T}},z_j^{\mathrm{R}})$。其对应的InfoNCE损失定义如下：

$$L_{\mathrm{NCE}}{}^{\mathrm{V}}=-\frac{1}{2N}\sum_i\left(\ln\frac{\exp S_{ii}/\tau}{\sum_j\exp S_{ij}/\tau}+\ln\frac{\exp S_{ii}/\tau}{\sum_j\exp S_{ji}/\tau}\right) \tag{7}$$

$$L_{\mathrm{NCE}}{}^{\mathrm{T}}=-\frac{1}{2N}\sum_i\left(\ln\frac{\exp W_{ii}/\tau}{\sum_j\exp W_{ij}/\tau}+\ln\frac{\exp W_{ii}/\tau}{\sum_j\exp W_{ji}/\tau}\right) \tag{8}$$

$$L_{\mathrm{NCE}}=L_{\mathrm{NCE}}{}^{\mathrm{V}}+L_{\mathrm{NCE}}{}^{\mathrm{T}} \tag{9}$$

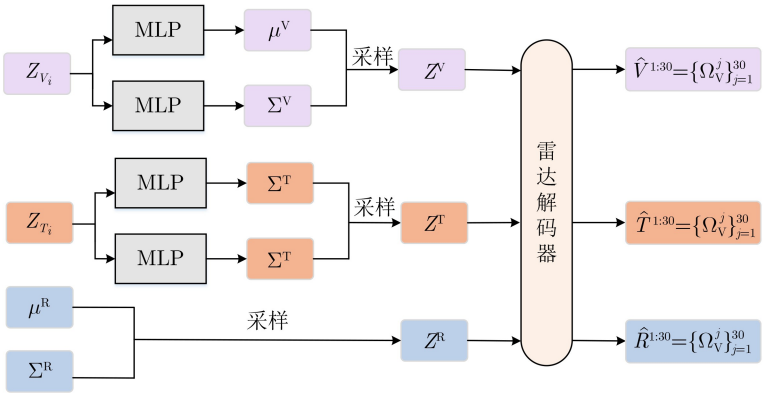


图 5 数据拟合与解码网络

最终的训练目标表示为$L_{\text{total}} = L_{\text{R}} + \lambda_{\text{KL}} L_{\text{KL}} + \lambda_{\text{NCE}} L_{\text{NCE}}$。$\lambda_{\text{KL}}$和$\lambda_{\text{NCE}}$设定为$10^{-4}$。用Adam优化器训练数据拟合网络，初始学习率设置为10^{-5}。

4 实验分析

4.1 系统实现与实验设置

4.1.1 系统实现

图6展示了基于TI IWR1443BOOST毫米波雷达的硬件采集平台，雷达配备3发4收天线，采用线性调频连续波(FMCW)技术，工作频段为76 GHz-81 GHz，距离分辨率为4 cm，多普勒分辨率为0.34 m/s，最大测距为8.19 m，最大径向速度为±2.67 m/s。雷达与相机均通过USB接口连接至戴尔笔记本，实现同步采集与存储。文本标签由采集阶段预先标注，并与结构化描述模板结合生成对应文本输入。系统基于Python环境开发，运行于Ubuntu 20.08服务器平台并配备NVIDIA 3080 GPU。

4.1.2 数据集

数据集1：招募了32名年龄在18至49岁的志愿者(男性22人，女性10人)，身高范围为156–182 cm，体重范围为45–100 kg。为模拟实际使用场景，硬件平台分别置于床的左侧、右侧以及正前方，高度设置为1.8 m，采集用户在3个不同位置执行手势的真实雷达数据。每位志愿者需完成5类预定义手势(拉手(PL)、推手(PS)、画圈(CR)、抬手(UP)、敲击(KO))，每类手势重复采集10次，总计32×3×3×5×10=14 400个样本。

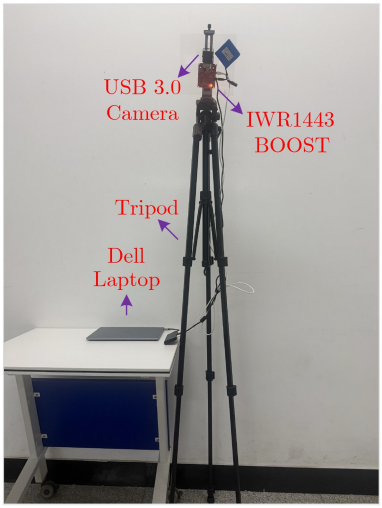


图 6 数据采集的硬件平台

数据集2：如图7所示，构建了一个多场景数据集以评估模型在不同用户位置、雷达位置以及场景下的泛化能力[30]。该数据集包含两个新场景，招募5名志愿者在3个用户位置(P1-P3)和3个雷达位置(R1-R3)下执行5类手势，每类手势重复10次，共计2 250个样本。

4.1.3 评价指标

为了全面评估VT2R的性能，本文选取了以下评价指标：

● **豪斯霍弗距离**：衡量生成雷达数据与真实雷达数据之间的最大点对点距离，值越小表示两者的空间一致性越高。

● **近邻准确率(1-NNA)**：将生成和真实数据混合后，基于高维特征表示构建1-近邻分类器，分类结果越接近50%表示两类数据分布越一致。

● **识别准确率**：衡量模型对手势样本的正确识别概率。

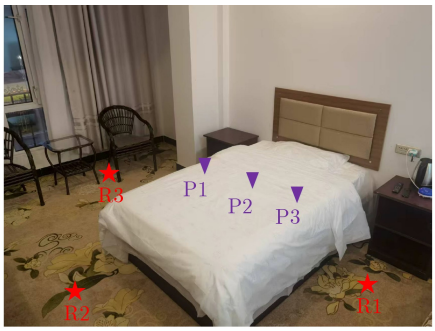
● **混淆矩阵**：用于展示模型在不同手势类别上的分类效果，行表示预测类别，列表示真实类别，矩阵中的值为各类别间的预测概率分布。

4.1.4 基线方法

本文将VT2R与多种代表性基线方法进行了比较，包括五种基于视频的方法(Vid2doppler[25]、SynMotion[26]、Midas[27]、SBRF[29]、Uranus[31])、一种基于人体网格的方法(mmGPE[20])以及一种基于生成扩散模型的方法(RFGen[23])。

4.1.5 实验设置

本文在三种不同的实验设置下评估VT2R及各对比方法的性能：第一种实验设置使用全部生成雷



(a) 场景1

(b) 场景2

图 7 两个不同场景. ☆表示不同雷达位置, ▽表示不同用户位置

达数据训练模型，并在真实雷达数据上测试；第二种实验设置使用真实雷达数据进行五折交叉验证，其中四折用于训练，一折用于测试；第三种实验设置使用全部生成数据与真实数据的四折进行训练，并在剩余一折真实数据上测试。

4.2 手势识别评估

4.2.1 生成数据的质量和多样性

为验证VT2R的有效性，将其与多个基线方法进行对比。如图8所示，VT2R在所有样本上的平均豪斯霍弗距离为0.49 m，该结果相较于Vid2doppler、SynMotion、Midas、SBRF、Uranus、mmGPE与RFGen分别降低了6.89倍、6.01倍、4.82倍、2.3倍、1.65倍、5.06倍和1.80倍，表明VT2R生成的数据与真实雷达数据更为接近。在1-NNA指标上，VT2R达到0.58，表明生成数据的空间分布特征和噪声特性与真实数据较为一致；该结果优于Vid2doppler (0.82)、SynMotion (0.76)、Midas (0.71)、SBRF (0.68)、Uranus (0.65)、mmGPE (0.79)及RFGen (0.8)，体现了生成数据质量的显著优势。此外，如图10所示，本文对真实数据与各方法在“推手”手势上的点云进行了可视化对比，结果表明，VT2R生成的点云更接近真实数据，从而更直观地验证了其生成质量优势。

4.2.2 手势识别准确率

图9展示了在三种设置下VT2R和RFGen的识别性能，可以观察到：(1)在第一种设置下，VT2R的识别准确率为89.2%，较RFGen提高了33.88%；(2)在第二种设置下，识别准确率为95.16%，第一种设置与其仅相差5.96%，表明仅使用生成数据训练的模型在性能上已接近真实数据的训练效果；(3)在第三种设置下，VT2R的准确率达到97.62%，较第二种设置提高了2.46%，比RFGen高21.48%，体现了VT2R的有效性。本文进一步分析不同手势类别的识别性能。如图11–13所示，第一种设置中多数手势准确率超过90%，其余亦高于85%；使用生成数据训练的模型在各类别手势上的表现与使用真实数据训练的模型十分接近；在第三种设置中，所有手势准确率均超过95%，部分接近100%。相比之下，RFGen在各类别上的准确率均低于70%，VT2R在五类手势上的最大提升达27.68%。同时，精确率-召回率曲线(图14)表明VT2R在保持高召回率的同时仅引入较小精度损失，进一步验证其对整体识别性能的提升效果。

4.2.3 新用户少样本适应性分析

为评估模型对新用户的适应能力，实验过程模拟逐步增加少量真实样本的场景。如图15所示，在第一种设置下，仅增加4个样本，VT2R即达到95.3%的准确率，超过仅使用真实数据训练的模型(95.16%)；当样本增至18个时，准确率提升至99.62%。在第三种设置下，仅需4个样本即可达到98.85%，随后性能趋近100%。结果表明，VT2R能够以极少的真实数据实现快速适应，并有效提升模型的泛化能力。

4.3 消融实验

本文也评估了各核心组件对模型性能的影响。如图16所示：(1)移除适配器及时间融合层(w/o AT)，识别准确率下降38.19%，表明引入适配器与时序建模显著增强视频表征能力；(2)去除结构化提示模板(w/o ST)，准确率下降5.84%，表明显式编码文本有助于提升语义理解与跨模态对齐效果；(3)以逐帧MLP替代体素化与Transformer机制(w/o RFE)，准确率下降19.15%，验证了雷达特征编码网络在捕捉点云时空特征方面的有效性；(4)移除KL散度(w/o KL)或跨模态相似度损失(w/o NCE)分别导致8.76%和10.54%的性能下降，表明对潜在分布进行规整并施加约束对于提升生成数据质量的重要性。此外，为进一步验证双模态融合的必要性，本文对比了仅视频引导(Video-only)和仅文本引导(Text-only)相对于VT2R的性能变化。两者的识别准确率分别为84.93%和82.89%，较VT2R分别下降4.27%和6.31%，表明视频与文本

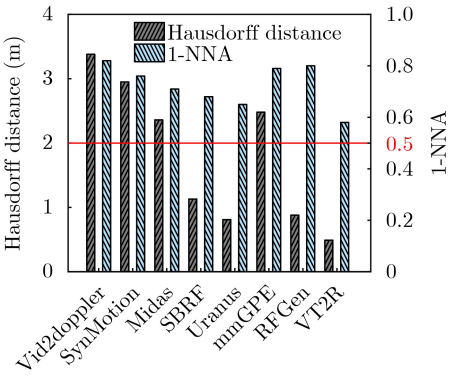


图 8 不同方法的豪斯霍弗距离和1-NNA

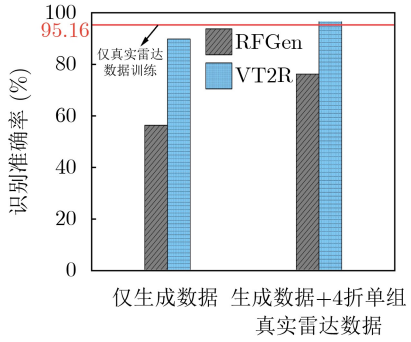


图 9 不同方法在不同实验设置下的识别准确率

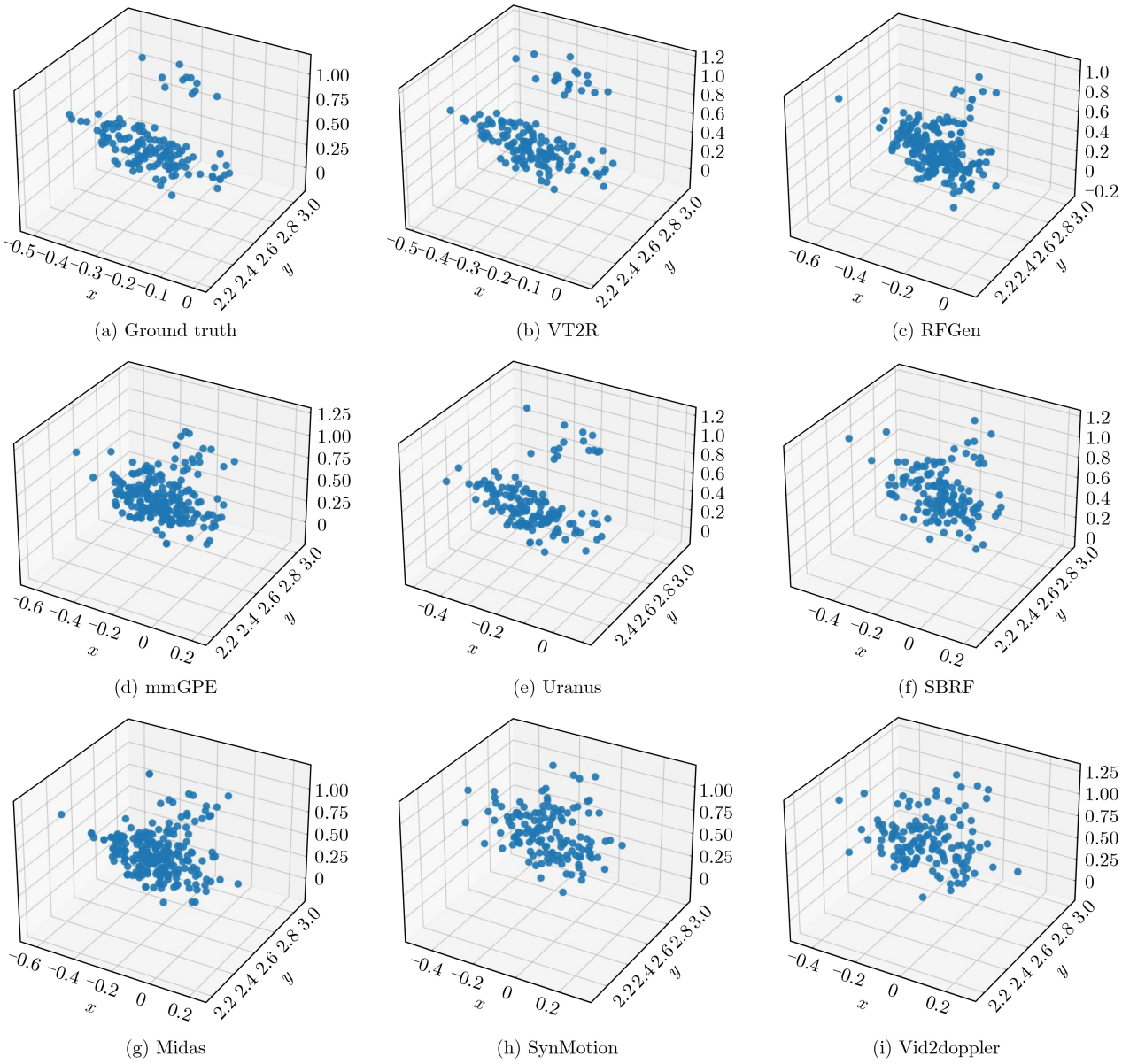


图 10 真实数据与不同方法生成结果的可视化对比

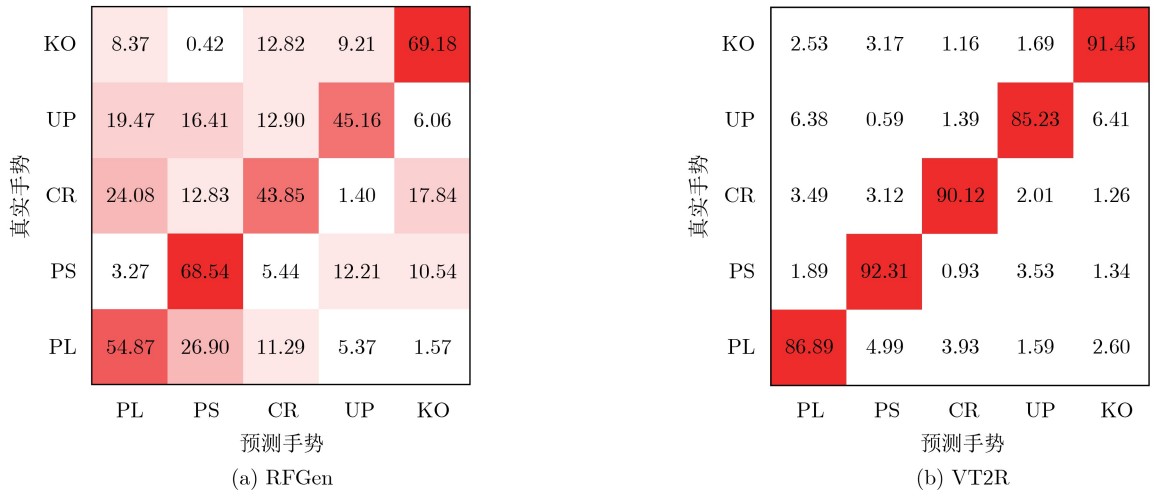
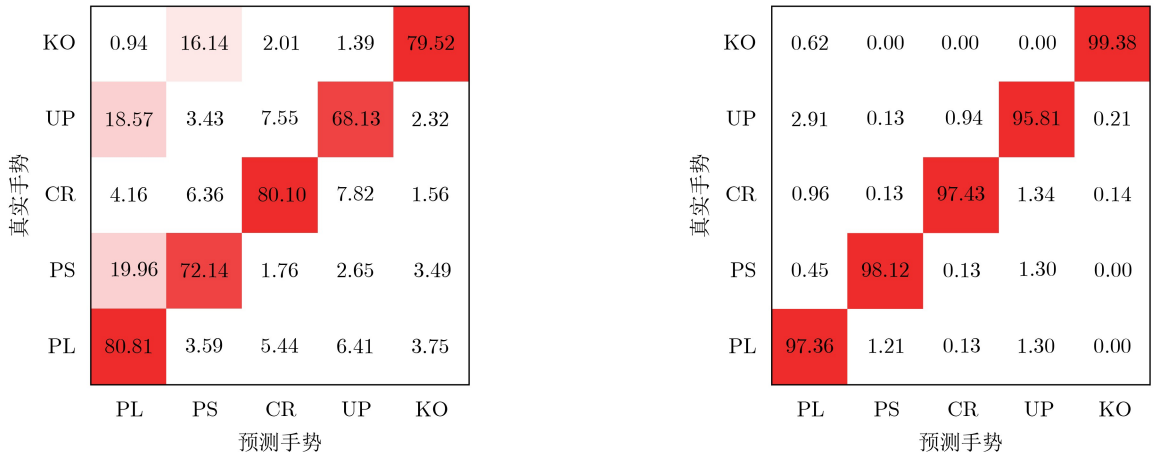


图 11 使用所有生成雷达数据的手势识别准确率 (%)



(a) RFGen　　(b) VT2R

图 12 使用所有生成数据和用户独立的真实雷达数据的识别准确率 (%)

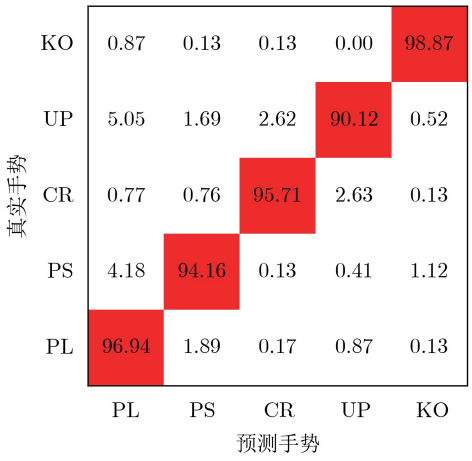


图 13 使用真实雷达数据的识别准确率 (%)

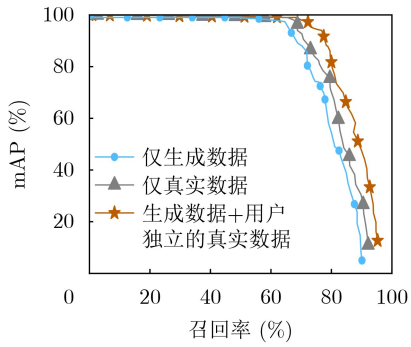
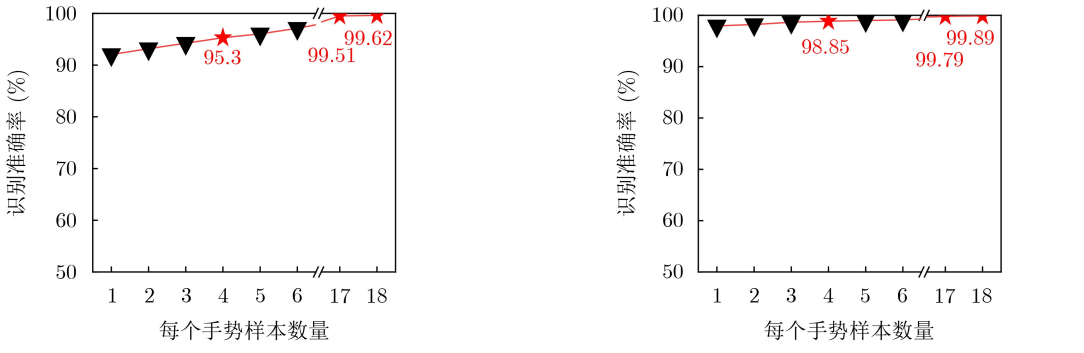


图 14 三种实验设置的精确率-召回率曲线

模态具有互补性，联合建模可进一步提升下游识别性能。

4.4 对各种因素影响的深入评估

本文进一步测试了VT2R在不同条件下的泛化能力。首先，在数据集2中对多种用户与雷达位置进行测试。如图17所示，第一种和第三种设置下的平均准确率分别为89.06%和97.23%，表明VT2R对位置变化具有良好适应性，并在不同布局条件下保持较强稳健性。其次，本文额外选取两个手势(从左向右挥手和从右向左挥手)，并采集了1 800个样本(5名用户×2类手势×2个场景×10次×3个用户位置×3个雷达位置)。如图18所示，VT2R在第一与第三种设置下分别达到89.53%和97.192%的平均准确率，与数据集1中其他手势的结果基本一致，表明了其对不同手势的有效性。进一步地，对数据集



(a) 仅使用生成数据训练模型　　(b) 使用生成数据+用户独立的雷达数据训练模型

图 15 随着新用户真实样本数量增加的识别准确率变化

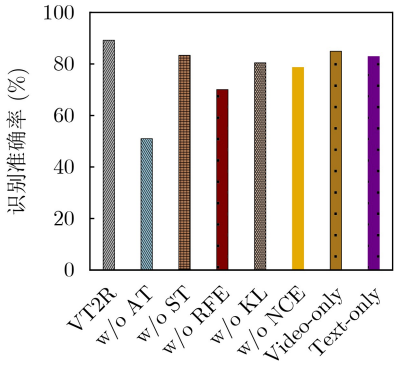
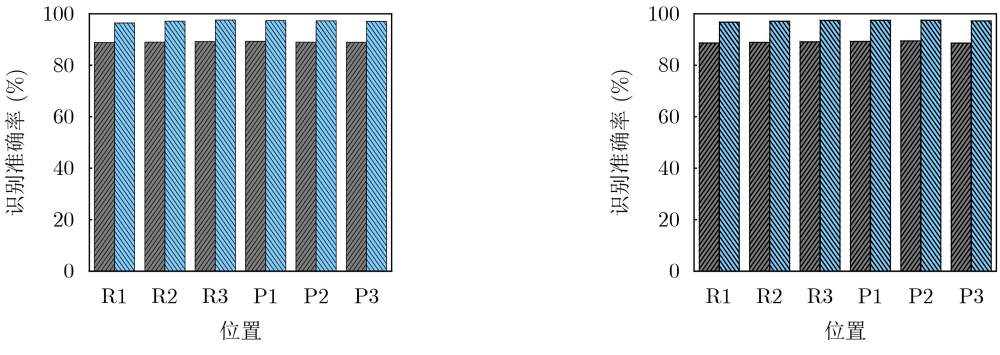


图 16 各个模块对识别准确率的影响

2中两个不同场景的性能进行测试。如图19所示，VT2R在场景1/场景2中达到89.34%/89.36%(第一种设置)以及97.13%/97.29%(第三种设置)的准确率，表明模型具备稳定的跨场景泛化能力。最后，本文利用已有的视频-雷达数据集[30]进行验证，其包含来自32名用户的23 040个坐姿样本。如图20所示，VT2R在第一种与第三种设置下分别取得了89.98%和97.55%的平均准确率，与躺姿场景的结果基本一致，表明VT2R并非仅适用于躺姿场景，在不同用户姿势下均能保持稳定性能。



(a) 场景1 (b) 场景2

图 17 两种设置下不同位置的识别准确率

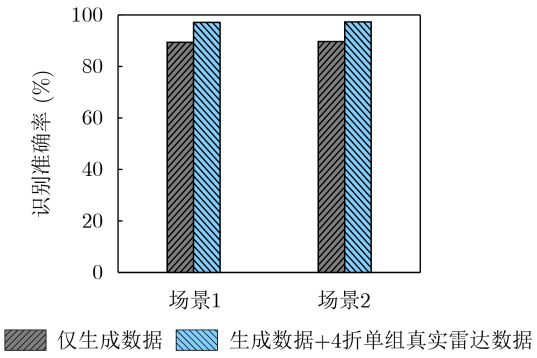


图 18 两种设置下不同手势的识别准确率

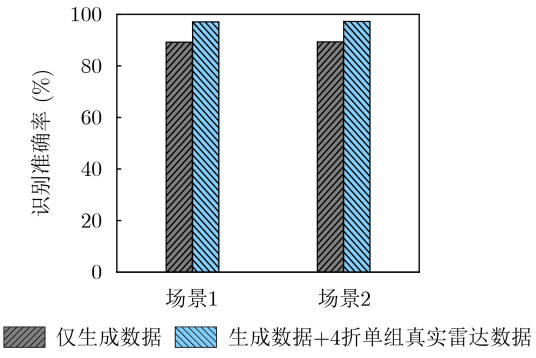


图 19 两种设置的总体识别准确率

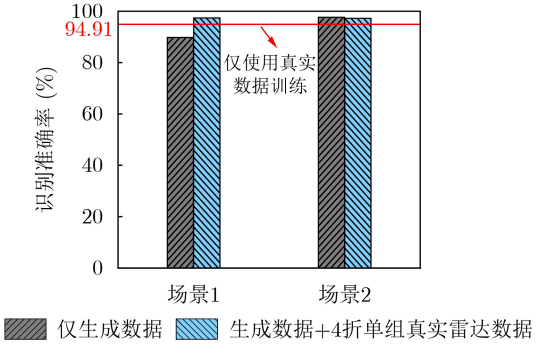


图 20 用户坐姿数据在三种设置下的识别准确率

5 总结

本文提出了一种新颖的雷达数据生成系统VT2R，旨在解决用户处于躺姿执行手势时真实雷达训练数据严重缺乏的问题，VT2R支持以视频或文本作为输入生成对应的大规模逼真的雷达数据，同时也支持基于少量真实雷达数据进行增强式重建，为雷达感知任务提供丰富的数据支撑。VT2R包括四个组件：基于视觉-语言预训练模型构建的视频特征编码网络、引入提示模板的文本编码网络、面向稀疏点云的层次化雷达编码网络，以及基于VAE的数据拟合与解码网络，这些组件协同生

成大规模逼真的雷达数据。最后，本文利用生成和自采的数据集对系统和现有方法进行了全面广泛的评估，实验结果表明，仅使用生成数据训练手势识别模型时，VT2R的准确率达到89.2%；结合少量真实雷达数据联合训练时，准确率进一步提升至97.62%。

参 考 文 献

[1] XING Ling, DENG Kaikai, WU Honghai, *et al*. The dawn of synthetic Era: Synthesizing mmWave radar data from 2D videos for human sensing[J]. *IEEE Communications Magazine*, 2025, 63(11): 14–20. doi: 10.1109/MCOM.001.2400488.

[2] LIU Xiulong, LIU Hankai, ZHANG Jiaqi, *et al*. Multi-user behavioral privacy filtering for mmwave radar sensing[J]. *IEEE Transactions on Mobile Computing*, 2025, 24(9): 8347–8361. doi: 10.1109/TMC.2025.3556674.

[3] YUAN Wenyang, ZHANG Jian, YUAN Wu, *et al*. 3D-sitpose: Millimeter wave radar-based human sitting posture estimation[J]. *ACM Transactions on Sensor Networks*, 2026, 22(2): 11. doi: 10.1145/3793858.

[4] 赵川斌, 许伟华, 林博, 等. 融合视觉的多模态通信感知一体化关键技术及原型验证[J]. 电子与信息学报, 2026, 48(2): 487–498. doi: 10.11999/JEIT250685.

ZHAO Chuanbin, XU Weihua, LIN Bo, *et al*. Vision enabled multimodal integrated sensing and communications: Key technologies and prototype validation[J]. *Journal of Electronics & Information Technology*, 2026, 48(2): 487–498. doi: 10.11999/JEIT250685.

[5] LI Yaxuan, XU Dongzhu, LIANG Kun, *et al*. Mobi^{2}Still: People detection and tracking with mobile human-equipped mmWave radars[J]. *IEEE Transactions on Mobile Computing*, 2026, 25(6): 9348–9364. doi: 10.1109/TMC.2026.3656239.

[6] KIM Y B, HAN S S, and LEE H L. Cost-effective FMCW radar with enhanced tracking coverage for smart healthcare applications[J]. *IEEE Transactions on Consumer Electronics*, 2026, 72(2): 3330–3340. doi: 10.1109/TCE.2026.3667885.

[7] YANG Huanqi, HAN Mingda, LI Xinyue, *et al*. iradar: Synthesizing millimeter-waves from wearable inertial inputs for human gesture sensing[C]. Proceedings of 2025 IEEE Conference on Computer Communications, London, United Kingdom, 2025: 1–10. doi: 10.1109/INFOCOM55648.2025.11044481.

[8] DING Fangqiang, LUO Zhen, ZHAO Peijun, *et al*. milliFlow: Scene flow estimation on mmWave radar point cloud for human motion sensing[C]. Proceedings of the 18th European Conference on Computer Vision, Milan, Italy, 2024: 202–221. doi: 10.1007/978-3-031-72691-0_12.

[9] 冉鑫怡, 陈前斌, 徐勇军, 等. 基于深度学习的通感一体化系统综述[J]. 通信学报, 2025, 46(6): 233–250. doi: 10.11959/j.issn.1000-436x.2025103.

RAN Xinyi, CHEN Qianbin, XU Yongjun, *et al*. Survey on deep learning-based integrated sensing and communication systems[J]. *Journal on Communications*, 2025, 46(6): 233–250. doi: 10.11959/j.issn.1000-436x.2025103.

[10] JIN Can, MENG Xiangzhu, LI Xuanheng, *et al*. Rodar: Robust gesture recognition based on mmWave radar under human activity interference[J]. *IEEE Transactions on Mobile Computing*, 2024, 23(12): 11735–11749. doi: 10.1109/TMC.2024.3402356.

[11] CHOI J, HOR S, YANG Shubo, *et al*. MVDoppler-Pose: Multi-modal multi-view mmWave sensing for long-distance self-occluded human walking pose estimation[C]. Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2025: 27750–27759. doi: 10.1109/CVPR52734.2025.02584.

[12] MARIKYAN D, PAPAGIANNIDIS S, RANA O F, *et al*. Working in a smart home environment: Examining the impact on productivity, well-being and future use intention[J]. *Internet Research*, 2024, 34(2): 447–473. doi: 10.1108/INTR-12-2021-0931.

[13] KRIZHEVSKY A, SUTSKEVER I, and HINTON G E. ImageNet classification with deep convolutional neural networks[J]. *Communications of the ACM*, 2017, 60(6): 84–90. doi: 10.1145/3065386.

[14] YANG Pinci, WANG Xin, DUAN Xuguang, *et al*. AVQA: A dataset for audio-visual question answering on videos[C]. Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 2022: 3480–3491. doi: 10.1145/3503161.3548291.

[15] Khowaja S A, KHUWAJA P, DHAREJO F A, *et al*. ReFuSeAct: Representation fusion using self-supervised learning for activity recognition in next generation networks[J]. *Information Fusion*, 2024, 102: 102044. doi: 10.1016/j.inffus.2023.102044.

[16] KWON H, TONG C, HARESAMUDRAM H, *et al*. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition[J]. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2020, 4(3): 87. doi: 10.1145/3411841.

[17] LENG Zikang, BHATTACHARJEE A, RAJASEKHAR H, *et al*. IMUGPT 2.0: Language-based cross modality transfer for sensor-based human activity recognition[J]. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2024, 8(3): 112. doi: 10.1145/

3678545.

[18] SEYFIOGLU M S, EROL B, GURBUZ S Z, *et al*. Diversified radar micro-Doppler simulations as training data for deep residual neural networks[C]. Proceedings of 2018 IEEE Radar Conference, Oklahoma City, USA, 2018: 612–617. doi: 10.1109/RADAR.2018.8378629.

[19] EROL B and GURBUZ S Z. A Kinect-based human micro-Doppler simulator[J]. *IEEE Aerospace and Electronic Systems Magazine*, 2015, 30(5): 6–17. doi: 10.1109/MAES.2015.7119820.

[20] XUE Hongfei, CAO Qiming, MIAO Chenglin, *et al*. Towards generalized mmWave-based human pose estimation through signal augmentation[C]. Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, Madrid, Spain, 2023: 88. doi: 10.1145/3570361.3613302.

[21] EROL B, GURBUZ S Z, and AMIN M G. Motion classification using kinematically sifted ACGAN-synthesized radar micro-Doppler signatures[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2020, 56(4): 3197–3213. doi: 10.1109/TAES.2020.2969579.

[22] RAHMAN M M, MALAIA E A, GURBUZ A C, *et al*. Effect of kinematics and fluency in adversarial synthetic data generation for ASL recognition with RF sensors[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2022, 58(4): 2732–2745. doi: 10.1109/TAES.2021.3139848.

[23] CHEN Xingyu and ZHANG Xinyu. RF genesis: Zero-shot generalization of mmWave sensing through simulation-based data synthesis and generative diffusion models[C]. Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems, Istanbul, Turkiye, 2023: 28–42. doi: 10.1145/3625687.3625798.

[24] CHI Guoxuan, YANG Zheng, WU Chenshu, *et al*. RF-diffusion: Radio signal generation via time-frequency diffusion[C]. Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, Washington, USA, 2024: 77–92. doi: 10.1145/3636534.3649348.

[25] AHUJA K, JIANG Yue, GOEL M, *et al*. Vid2doppler: Synthesizing Doppler radar data from videos for training privacy-preserving activity recognition[C]. Proceedings of 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 2021: 292. doi: 10.1145/3411764.3445138.

[26] ZHANG Xiaotong, LI Zhenjiang, and ZHANG Jin. Synthesized millimeter-waves for human motion sensing[C]. Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems, Boston, USA, 2022: 377–390. doi: 10.1145/3560905.3568542.

[27] DENG Kaikai, ZHAO Dong, HAN Qiaoyue, *et al*. Midas: Generating mmWave radar data from videos for training pervasive and privacy-preserving human sensing tasks[J]. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2023, 7(1): 9. doi: 10.1145/3580872.

[28] DENG Kaikai, ZHAO Dong, ZHANG Zihan, *et al*. Midas++: Generating training data of mmWave radars from videos for privacy-preserving human sensing with mobility[J]. *IEEE Transactions on Mobile Computing*, 2024, 23(6): 6650–6666. doi: 10.1109/TMC.2023.3325399.

[29] LI Jiamu, ZHANG Dongheng, WU Zhi, *et al*. SBRF: A fine-grained radar signal generator for human sensing[J]. *IEEE Transactions on Mobile Computing*, 2024, 23(12): 13114–13130. doi: 10.1109/TMC.2024.3427406.

[30] DENG Kaikai, ZHAO Dong, ZHENG Wenxin, *et al*. G^{3}R: Generating rich and fine-grained mmWave radar data from 2D videos for generalized gesture recognition[J]. *IEEE Transactions on Mobile Computing*, 2025, 24(4): 2917–2934. doi: 10.1109/TMC.2024.3502668.

[31] LING Yue, ZHAO Dong, DENG Kaikai, *et al*. Uranus: Empowering generalized gesture recognition with mobility through generating large-scale mmWave radar data[J]. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2024, 8(4): 204. doi: 10.1145/3699754.

[32] ZHOU Yunjiao, YANG Jianfei, ZOU Han, *et al*. TENT: Connect language models with IoT sensors for zero-shot activity recognition[J]. *IEEE Transactions on Mobile Computing*, 2026, 25(6): 8314–8326. doi: 10.1109/TMC.2025.3650710.

[33] KINGMA D P and WELLING M. Auto-encoding variational Bayes[C]. Proceedings of the 2nd International Conference on Learning Representations, Banff, Canada, 2024.

[34] ZHANG JIAN, He Kaihao, YU Ting, *et al*. Semi-supervised RGB-D hand gesture recognition via mutual learning of self-supervised models[J]. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025, 21(4): 104. doi: 10.1145/3689644.

[35] AMESAKA T, WATANABE H, SUGIMOTO M, *et al*. Gesture recognition method using acoustic sensing on usual garment[J]. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2022, 6(2): 41. doi: 10.1145/3534579.

[36] MA Zijing, ZHANG Shigeng, LIU Jia, *et al*. RF-Siamese: Approaching accurate RFID gesture recognition with one sample[J]. *IEEE Transactions on Mobile Computing*, 2024, 23(1): 797–811. doi: 10.1109/TMC.2022.3217487.

[37] GAO Ruiyang, LI Wenwei, XIE Yaxiong, *et al*. Towards robust gesture recognition by characterizing the sensing quality of WiFi signals[J]. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2022, 6(1): 11. doi: 10.1145/3517241.

[38] 赵雅琴, 宋雨晴, 吴晗, 等. 基于DenseNet和卷积注意力模块的高精度手势识别[J]. 电子与信息学报, 2024, 46(3): 967–976. doi: 10.11999/JEIT230165.
ZHAO Yaqin, SONG Yuqing, WU Han, *et al*. High-precision gesture recognition based on DenseNet and convolutional block attention module[J]. *Journal of Electronics & Information Technology*, 2024, 46(3): 967–976. doi: 10.11999/JEIT230165.

[39] HAYASHI E, LIEN J, GILLIAN N, *et al*. RadarNet: Efficient gesture recognition technique utilizing a miniature radar sensor[C]. Proceedings of 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 2021: 5. doi: 10.1145/3411764.3445367.

[40] CAI Hong, KORANY B, KARANAM C R, *et al*. Teaching RF to sense without RF training measurements[J]. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2020, 4(4): 120. doi: 10.1145/3432224.

[41] LI Ruihui, LI Xianzhi, FU C W, *et al*. Pu-GAN: A point cloud upsampling adversarial network[C]. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019: 7202–7211. doi: 10.1109/ICCV.2019.00730.

邓凯凯：男，讲师，研究方向为智能物联网、多模态感知计算、数据生成、自动驾驶.
凌　月：女，博士生，研究方向为多模态数据生成、智能物联网.
邢　玲：女，教授，研究方向为智能物联网、车联网、自动驾驶.
吴红海：男，教授，研究方向为智能物联网、视频分析、边缘计算.
赵　东：男，教授，研究方向为智能物联网、数据生成、无线感知计算、健康计算.
马华红：女，副教授，研究方向为视频分析、边缘卸载、智能物联网.

责任编辑：廖海贝

VT2R: Video and Text-driven Method for Generating Large-scale Millimeter-wave Radar Data

DENG Kaikai[①]　LING Yue[②]　XING Ling[①]　WU Honghai[①]
ZHAO Dong[②]　MA Huahong[①]

[①](*School of Information Engineering, Henan University of Science and Technology, Luoyang* 471023, *China*)
[②](*State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)

Abstract:

Objective　The lack of large-scale training data impedes progress in developing robust and generalized deep learning models. However, existing millimeter-wave radar data generation methods are ineffective due to a lack of sufficient data sources. To address this gap, this paper proposes a video and text-driven radar data generation method, VT2R, which utilizes video or text data to generate large-scale, realistic radar data, solving the key problem of constructing the mapping relationship between video and text and radar data.

Methods　The proposed method consists of three main components: video feature encoding network, text feature encoding network, radar feature encoding network and data fitting and decoding network. Video feature encoding networks and text feature encoding networks extract temporally consistent visual representations and alignable semantic features, respectively, while the radar encoding network learns the structure and dynamic information of point clouds through hierarchical spatiotemporal modeling. In the data fitting and decoding network based on Variational AutoEncoder (VAE), multi-modal features are mapped to a unified latent distribution space and decoded into radar data through reparameterized sampling. During training, reconstruction loss, Kullback-Leibler (KL) divergence loss, and cross-modal similarity loss are jointly optimized.

Results and Discussions　This paper constructs the first radar point cloud dataset for reclining gesture recognition (Figs. 6 and 7), covering 5 gesture categories, 32 participants, and a total of 14,400 samples. Experimental results based on this dataset show that VT2R achieves a recognition accuracy of 89.2% when trained using only generated radar data, a 33.88% improvement over the representative RFGen (Figs. 9 and

10). When combined with a small amount of real radar data for joint training, the accuracy further improves to 97.62%, a 21.48% improvement over RFGen (Figs. 9 and 11). Furthermore, VT2R still achieves average recognition accuracies of 89.35% and 97.21% under different scenarios and factors (Figs. 16-18). In addition, this paper also verifies the accuracy of VT2R under different postures, achieving average accuracies of 89.98% and 97.55% in the first and third settings, respectively (Fig. 19), which is basically consistent with the result obtained when lying down, demonstrating its robustness under cross-posture conditions.

Conclusions This paper proposes a radar data generation system, VT2R, which addresses the severe lack of realistic radar training data when users are performing gestures in a lying position. Through a video feature encoding network built on a vision-language pre-trained model, a text encoding network incorporating cue templates, a hierarchical radar encoding network for sparse point clouds, and a VAE-based data fitting and decoding network, these components collaboratively generate large-scale, realistic radar data. It also supports augmented reconstruction based on limited real radar data, providing rich data support for radar perception tasks. Future work will focus on solving multi-modal data generation for more complex gesture scenes, providing better data support for emerging large-scale models.

Key words: mmWave radar sensing; data generation; gesture recognition; variational autoencoder