面向 6G 的跨模态信号重建技术

李昂[1,2]，陈建新[1,2]，魏昕[1,2]，周亮[1,2]

（1. 南京邮电大学通信与信息工程学院，江苏 南京 210003；
2. 南京邮电大学宽带无线通信与传感网技术教育部重点实验室，江苏 南京 210003）

摘 要：6G 时代下，为了兼顾多媒体用户音频、视频、触觉的沉浸式体验需求与低时延、高可靠、大容量的通信质量，提出一种跨模态信号重建架构和由视频信号重建触觉信号的深度学习模型。首先，通过控制机器人触摸各种材质，构建了包含音频、视频、触觉信号的数据集 VisTouch，为后续各种跨模态问题的研究奠定基础；其次，通过利用多模态信号间的语义关联性，设计一种普适的、稳健的端到端跨模态信号重建框架；再次，以通过视频信号重建触觉信号为例，构建视频辅助的触觉重建模型，包括基于 3D CNN 的视频特征提取网络，基于全卷积网络的 GAN 生成网络与基于 CNN 的 GAN 辨别网络；最后，通过实验结果验证跨模态信号重建框架的可靠性以及触觉重建模型的准确性。

关键词：6G；跨模态信号重建；多模态数据集；3D 卷积神经网络；生成对抗网络

中图分类号：TP391

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2022093

6G-oriented cross-modal signal reconstruction technology

LI Ang[1,2], CHEN Jianxin[1,2], WEI Xin[1,2], ZHOU Liang[1,2]

1. College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. Key Laboratory of Broadband Wireless Communication and Sensor Network Technology (Ministry of Education), Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: In the 6G era, to balance the immersive experience needs of multimedia users for audio, video, and haptics with low-latency, high-reliability, and large-capacity communication, a cross-modal signal reconstruction framework and video-to-haptic reconstruction model was proposed. First, robots were controlled to touch various materials. In this way, a large-scale dataset VisTouch that includes audio, video, and haptic signals was constructed. This dataset could lay the foundation for subsequent researches on various cross-modal problems. In addition, based on the semantic relations of multi-modal signals, a universe and robust end-to-end cross-modal signal reconstruction framework was designed. Furthermore, the reconstruction from video to haptic signals was taken as an example. A video-assisted haptic reconstruction model was established, including a 3D CNN-based video extraction sub-network, a fully convolutional network based GAN generation sub-network and a CNN-based GAN discrimination sub-network. Finally, the reliability of the cross-modal signal reconstruction framework and the accuracy of the proposed video-to-haptic model were verified through experimental results.

Keywords: 6G, cross-modal signal reconstruction, multi-modal dataset, 3D CNN, GAN

收稿日期：2021−12−22；**修回日期：**2022−03−22

通信作者：周亮，liang.zhou@njupt.edu.cn

基金项目：国家自然科学基金资助项目（No.62071254）；江苏高校优势学科建设工程基金资助项目

Foundation Items: The National Natural Science Foundation of China (No.62071254), Priority Academic Program Development of Jiangsu Higher Education Institutions

0　引言

当前，通信学术界、产业界以及各种标准化组织对 6G 的愿景、需求和技术架构等方面展开了畅想与深入研究。中国信息通信研究院 IMT-2030（6G）推进组发布的《6G 总体愿景与潜在关键技术白皮书》指出，6G 将提供完全沉浸式交互场景，支持精确的空间互动，满足人类在多重感官，甚至情感和意识层面的联通交互[1]。可以预见，服务驱动下的 6G 技术将与人工智能技术、混合现实技术、物联网技术、传感器技术等进行深度融合，催生如元宇宙、数字孪生、全息服务等大量沉浸式多媒体应用。6G 时代下，传统以视听为核心的多媒体应用已逐渐不能满足用户的沉浸式体验需求，因此，亟须在新型多媒体应用中引入新的感官交互，如触觉等，为用户带来身临其境的极致体验。然而，新模态信号的引入势必会对现有的多媒体系统提出巨大挑战，《白皮书》[1]指出，若实时的交互达到完全沉浸水平，吞吐量需求约为 3.8 Gbit/s，且在多维感官信息协同传输的要求下，网络传输的最大吞吐量预计将成倍提升。因此，为了兼顾用户体验与通信质量，迫切需要一种跨模态信号重建方案来减少传输数据量，以支持 6G 沉浸式多媒体应用。

有研究表明，多模态应用将触觉信号与传统音视频信号结合起来，用户可通过触摸或交互行为获得更多的沉浸式体验[2]。针对 6G 时代下的多模态应用，文献[3]提出音频、视频、触觉跨模态通信架构，旨在通过充分挖掘不同模态信号之间的关联性来解决高效的触觉信号编码、异构码流传输、模态信息重建三大关键科学问题。文献[4]进一步提出人工智能加持下的跨模态通信框架，利用强化学习、迁移学习等技术解决跨模态通信中的技术挑战。其中，信号在传输及接收过程中势必会伴随不同程度的丢失，因此，发掘音频、视频、触觉信号间的内在关联性，利用一种模态信号精准、实时地重建另一种模态信号，是 6G 跨模态通信研究的重点之一，也被认为是可大幅提升用户沉浸式体验的关键技术[5]。在 6G 的潜在沉浸式应用场景中，如沉浸式扩展现实（XR, extended reality）、全息通信、感官互联，跨模态重建技术可利用现有的视频、音频信号恢复出同一物体的触觉信号，新生成的触觉信号又可对原始音视频信号进行超分辨率重建，极大地满足人与人、物、环境的沟通需求，同时 6G 下的毫秒级时延将为用户提供较好的连接体验。6G 跨模态信号重建需求如图 1 所示。

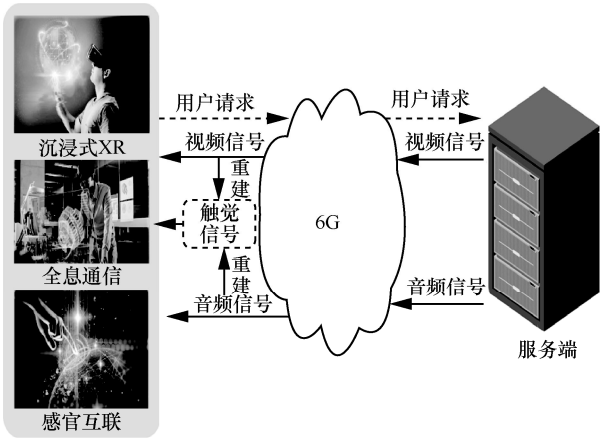


图 1　6G 跨模态信号重建需求

对于实现跨模态重建的深度学习模型来说，其性能优劣依赖于数据集的质量与规模，理论上，数据量越大，标注质量越高，深度模型越能逼近甚至超越人类表现，例如，利用大规模 ImageNet 图像数据集训练出的 AlexNet[6]、VGG[7]、ResNet[8]等图像模型已经与人类识别准确率相差无几。当前，音视数据集种类繁多，因此现有工作主要集中于利用深度模型探索音频、视频之间的语义关系。为了满足 6G 沉浸式体验需求，迫切需要一个大规模、高质量的音视触数据集来助力深度学习完成跨模态编码、传输、信号处理等任务。此外，当前大量的研究主要集中于音频、视频之间的恢复与重建[9-23]，对利用音频、视频重建触觉信号的研究还处于起步阶段。与此同时，不同传感器采集到的触觉信号结构与内容各异，如何对不同形式的触觉信号进行语义表征，以及如何设计普适的、稳健的跨模态信号重建框架，已经成为实现 6G 跨模态应用的难点。

相较于传统以音频、视频为核心的跨模态重建方法，本文探索针对音频、视频、触觉 3 种信号的通用型跨模态信号重建框架；为了研究模态间的关联性及拓展深度学习技术在其中的应用，通过机器人自动采集的方式构建了大规模多模态数据集 VisTouch；考虑到触觉、视觉感受对于实现 6G 沉浸式体验具有重要影响，本文以视频生成触觉为例，在所提框架下，针对自研 VisTouch 数据集的信号特性与实际需求，设计基于 3D CNN 和生成对抗网络（GAN, generative adversarial network）的视频辅助的触觉重建模型。具体来说，本文贡献主要总结为以下三点。

1) 面向6G跨模态应用场景，构建大规模音频、视频、触觉多模态数据集VisTouch。为减少人为因素干扰，本文控制机器人触摸47种材质，采集该过程中产生的同步音频、视频、触觉信号。相较于传统音频、视频双模态数据集，音频、视频、触觉三模态VisTouch数据集更适合探索6G中以触觉为核心的沉浸式体验方案。

2) 针对同一对象不同模态信号的语义一致性，基于深度学习技术提出一种内在语义关联驱动下的跨模态信号重建架构。该架构包括特征提取模块、重建模块、评估模块。通过三大模块之间的相互配合，可准确、低噪地利用一种模态信号重建另一种模态信号。

3) 以视频信号重建触觉信号为例，利用3D CNN与GAN设计一种视频辅助的触觉重建模型。该模型将上述跨模态信号重建架构具体化。为提升信号重建质量，触觉重建模型利用对抗损失与均方误差损失这两类损失函数作为目标函数，并基于VisTouch进行训练，均方误差结果验证了该模型的重建准确性。

1 跨模态信号重建相关工作

1.1 多模态数据集

当前多模态数据集多集中于音频、视频2种模态，主要用于探究视频动作与声音之间关联性，就内容来说，描述音乐与演奏动作、音乐与舞蹈动作、人声与嘴部动作的数据集占据多数。例如，C4S[9]由9位不同单簧管演奏家的54个视频组成，每个视频演奏两遍3个古典音乐作品。为了探究演奏者演奏其他乐器时的动作与音乐之间的关系，近年来，更多乐器如小提琴、手风琴等被包含在音视多模态数据集中，其中比较流行的数据集有URMP[10]（14种乐器）、MUSIC[11]（12种乐器）、Solos[12]（13种乐器）、HMMD[13]（7种乐器），虽然这些数据集包含数千个音频−视频对，但由于采集设备不同、去噪方法各异，导致其质量参差不齐，且仅仅关注乐器的音频、视频关系具有一定的局限性，难以大规模推广。除了乐器演奏的大量数据集，其他场景如人脸对话数据集和舞蹈动作−音乐数据集也具有很强代表性。文献[14]设计了AVA-ActiveSpeaker数据集，包含视频中标记的人脸轨迹，其中每个人脸实例都标记为说话或不说话，以及语音是否可听见，该数据集包含约365万帧、38.5 h的面部轨迹以及相应的音频。AIST[15]舞蹈视频数据集包括10种街舞流派、35名舞者、9个摄像机视点和60首音乐作品，涵盖12种节奏。文献[16]提出了一个新的3D舞蹈动作和音乐的多模态数据集AIST++，该数据集包含10种舞蹈流派、数百种编舞，运动持续时间从7.4 s到48.0 s不等，所有的舞蹈动作都有相应的音乐。文献[17]编制了一个HIMV-200K多模态数据集，包含200段视频、500段音频。

近年来，学术界开始关注触觉在沉浸式体验中的作用，并利用触觉手套、触觉传感器等设备采集人类皮肤或机器接触实物时所产生的各种触觉信号。文献[18]将GelSight[19]触觉传感器装载在机械臂上，控制机械臂按压195种实物，采集二维触觉图和按压视频对，从而构建了一种大规模触觉−视频数据集VisGel。文献[20]以低成本（大约10美元）设计了一款包含548个传感阵列的触觉手套，并通过抓取26种实物，构建了包含135 000帧的触觉图数据集。然而，当前触觉研究还处于起步阶段，现有触觉数据集仅有VisGel可用于研究视频与触觉之间的关联性，仍旧缺乏音频、视频、触觉共存的三模态信号同步数据集。为此，本文开发出音频、视频、触觉数据集，并通过深度学习模型的表现验证该数据集的实用性。

1.2 跨模态信号重建方法

随着大数据时代的到来，横跨视觉、听觉、触觉模态的数据正在以前所未有的速度增长，由此产生大量充满挑战性的跨模态任务。常见的跨模态学习任务有跨模态分离与定位、跨模态对应学习、跨模态重建、跨模态表示等。其中跨模态重建由于其广泛的应用场景正在成为一个新兴热点研究方向。在过去的几十年里，音频和视频作为人们日常生活中最重要的2种感知方式，跨模态视听重建在学术界和工业界都得到了广泛的发展。

文献[21]提出了一种新颖的级联注意力引导的残差生成对抗网络（CARGAN, cascade attention guided residue GAN），旨在根据相应的音频信号重建场景。特别是，该研究提出了一个残留模块来逐步缩小不同模式之间的差距。此外，文献[21]还设计了具有新颖分类损失函数的级联注意力引导网络来解决跨模态学习任务，保持了高级语义标签域的一致性，并且能够平衡2种不同的模态。

文献[22]提出了一种跨模态循环重建对抗网络（CMCGAN, cross-modal cycle generative adversarial network）来处理跨模态的视频−音频相互重建。具体来说，CMCGAN由4种子网络组成，分别为视

频生成音频网络、音频生成视频网络、视频生成视频网络、音频生成音频网络，这 4 种子网络以循环结构进行组织。CMCGAN 有以下显著优势：首先，CMCGAN 通过一个联合对应的对抗性损失，将视觉−音频的相互重建统一为共同的框架；其次，通过引入一个具有高斯分布的潜在向量，CMCGAN 可以有效地处理视觉和音频模式上的维数和结构不对称性；最后，CMCGAN 采用端到端的方式进行训练，便于部署及应用。进一步地，利用 CMCGAN 开发了一个动态多模态分类网络来处理模态缺失问题。大量的实验结果表明，所重建的模态与原始模态的效果相当。

跨模态的关联学习对于稳健的多模态推理至关重要，尤其是在推理过程中模态可能缺失的情况下。文献[23]在给定音频合成视频的背景下研究该问题。具体来说，该研究目标是重建未来的视频帧，并根据音频和过去的视频帧重建它们的运动动力学。为了解决这个问题，该研究提出了 Sound2Sight，这是一个深层的变分框架，以音频和过去的视频帧的联合嵌入表示为输入，训练该框架学习每帧的随机先验知识。这种嵌入是通过一个基于多头注意的视听转换器编码器来学习的。然后，对所学习的先验知识进行采样，以进一步调节视频预测模块以重建未来帧。此外，为了提高重建帧的质量和内容的一致性，该研究提出了一种多模态鉴别器，用于区分合成的音视频剪辑和真实的音视频剪辑。实验表明，Sound2Sight 在重建视频质量方面显著优于最新技术，同时还能重建多种类型的视频内容。

尽管当前已有利用音频、视频进行相互重建的工作，但迄今为止，鲜有针对触觉的恢复、重建工作，文献[24]首次在跨模态通信框架中探讨触觉重建问题，通过特征提取、共享语义学习、触觉生成等步骤构建虚拟触觉，并搭建跨模态通信平台以证明其方法的优越性。本文沿用其触觉重建的主要思想，并考虑更多模态信号的语义特征，探索出一种涵盖音频、视频、触觉的语义关联驱动的跨模态信号重建架构。

2　VisTouch 数据集

针对跨模态通信需求，本文构建了一个大规模音频、视频、触觉数据集 VisTouch。本节主要描述了 VisTouch 的数据采集过程，介绍了音频、视频、触觉采集设备以及材质类型。

2.1　数据采集方式

触觉感知与所接触物体的特性以及探索表面的方式有关，而摩擦在触觉感知过程中扮演了重要角色。为此，在 VisTouch 中，数据采集手段为脚本控制机械手滑动触摸各种材质，并将滑动触摸过程中指尖与材质摩擦产生的滑动摩擦力作为触觉信号，同时利用高清摄像头及挂载在机械手的单向拾音器采集音频、视频信号，并用时间戳进行同步。

触觉信号的精准、低噪采集是 VisTouch 构建的核心。滑动摩擦力的大小与施加在接触面的法向压力以及动摩擦系数有关，动摩擦系数反映材质特性，一般为常数值，故施以恒定的法向压力是保证触觉信号精准、低噪的关键。需要从两方面入手：1) 将机械臂放置在桌面上，并给予挂载在机械臂末端的机械手以垂直于桌面向下的恒定驱动力；2) 采集材质选用片状以保证驱动力对接触面的法向性，从而减少材质形状因素对采集信号的影响。

为了增强数据集的样本多样性及实用性，在数据采集过程中引入 2 种数据增强策略：1) 滑动触摸轨迹设置直线滑动、曲线滑动、折线滑动这 3 种；2) 恒定法向驱动力大小设置 3N、6N、9N 这 3 种，并与滑动轨迹交叉组合，共可设置 9 种滑动方式。

VisTouch 数据集所使用的采集设备具体参数如表 1 所示。

表 1　　采集设备具体参数

采集设备	采集信号	设备参数
铁三角单向拾音器 AT9912	音频	立体声/单声道：单声道 频率响应：70～16 000 Hz 灵敏度：−39 dB
海康威视高清摄像头 DS-U32W	视频	最高分辨率：1 920×1 280 视频帧率：30 frame/s 镜头焦距：2.7～13 mm
因时机械手 RH56BFX-2L 指关节力传感器	触觉	采样频率：100 Hz 力分辨率：0.5 N

2.2　数据样本

信号的特征与材质本身息息相关，当控制机械手触摸各种材质时，粗糙材质（如石头）的触觉信号曲线相对于光滑材质（如玻璃）波动更大，声音信号更刺耳，因此，理论上数据集所包含的材质样本越多，越有助于探索音频、视频、触觉感知机理及表征模型。由于样本形状、材质等均对多模态信号的形式及内容产生影响，且样本形状的触觉采集需要阵列式点阵传感器，故在本文所提的 VisTouch 1.0 版本中，仅针对片状材质样本（如石头片、纸片、木片）利用单个力

传感器进行数据采集，采集到的触觉信号以一维时间序列表示。在未来，随着研究的深入，将考虑在 VisTouch 2.0 版本中引入触觉手套、GelSight 等新型传感器抓来感知物体的形状信息，形成二维触觉图，以此丰富数据集内容与提升实用价值。针对样本类型，本文调研了当前生活中常见的、实用价值高的材质，总计 47 种，并对其进行分类，如表 2 所示，然后对这些材质样本利用 2.1 节中所设计的采集方式进行多模态数据采集，VisTouch 数据集示例如图 2 所示。

表 2　VisTouch 数据集包含的样本类别

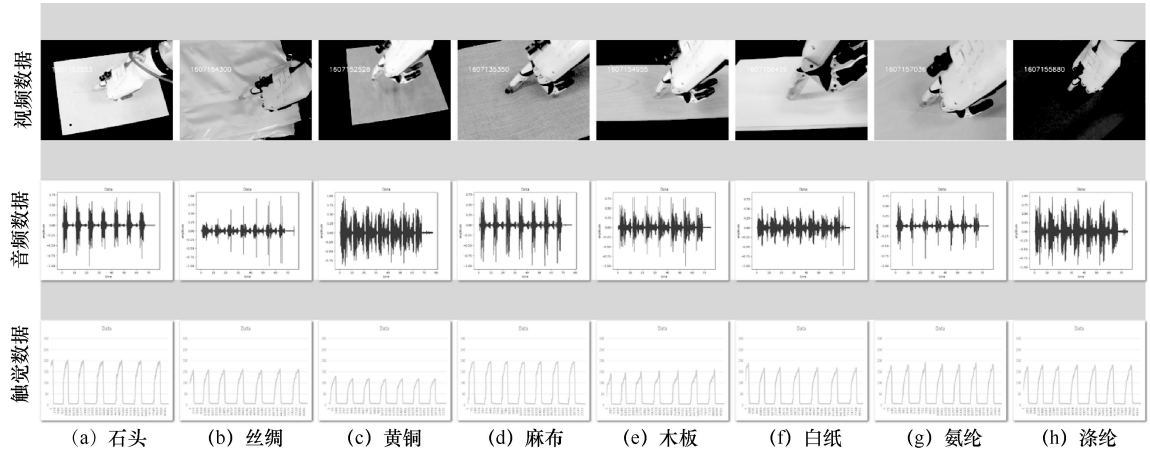
样本大类	样本小类	类别数量/种
塑料	聚对苯二甲酸乙二醇酯、高密度聚乙烯、聚氯乙烯、低密度聚乙烯、聚丙烯、聚苯乙烯	6
金属	铁、铝、镍、锌、钛、铟、铜、钽、钼	9
木材	樱桃木、松木、黑胡桃、竹	4
纸	打印纸、报纸、硬纸板	3
陶瓷	传统陶瓷、特种陶瓷	2
橡胶	天然橡胶、合成橡胶	2
天然纺织品	棉、亚麻、丝绸	3
合成纺织品	锦纶、涤纶、腈纶、维伦、丙纶、氨纶、碳纤维	7
玻璃	普通玻璃、石英玻璃	2
皮革	牛皮、羊皮、人造皮革	3
石头	花岗岩、大理岩、石灰岩、板岩、泥岩、安山岩	6

在样本收集过程中，可观察到同种材质由于染色、加工等原因，其颜色各异，例如，玻璃不仅在类别上有普通玻璃和石英玻璃之分，而且在色彩上可分为有色玻璃和透明玻璃，这对跨模态信息处理带来了一定的挑战。为此，本文针对同一类型的样本，尽可能收集多种颜色，如合成纺织品，收集红色、黄色、蓝色、白色 4 种颜色的样本，针对玻璃，收集有色玻璃、透明玻璃、毛玻璃等样本，以此减少颜色对研究工作的影响。

最后，本文将所提 VisTouch 数据集与现有主流数据集在模态、类别、样本数量方面进行比较，如表 3 所示，以展示所提数据集的优越性。其中，帧数量指所采集的图像数据帧数，对于 STAG 这种单模态触觉数据，指触觉图帧数。数据集中大部分为音频、视频双模态数据集，与触觉相关的数据集主要有 VisGel[18] 与 STAG[20]。与它们相比，VisTouch 数据集的不同主要体现在：1) 本文所提数据集进行了音频、视频、触觉 3 种信号的采集，VisGel 仅记录视频与触觉信号，STAG 仅记录触觉信号；2) VisGel 与 STAG 的触觉信号本质为机械压力，该压力仅与物体形状、施加力有关，不能准确反映材质特性，而 VisTouch 数据集以滑动摩擦力作为触觉信号，其动摩擦系数与材质本身相关，故能准确反映材质特性。

表 3　VisTouch 与主流数据集的比较

数据集名称	内容	类别数量/种	帧数量
C4S[9]	音频、视频	1	十万级
URMP[10]	音频、视频	14	百万级
MUSIC[11]	音频、视频	12	百万级
Solos[12]	音频、视频	13	百万级
HMMD[13]	音频、视频	7	百万级
AVA-ActiveSpeaker[14]	音频、视频	—	百万级
AIST[15]	音频、视频	—	百万级
AIST++ [16]	音频、视频	—	百万级
HIMV-200K [17]	音频、视频	—	百万级
VisGel[18]	视频、触觉	195	百万级
STAG[20]	触觉	26	十万级
VisTouch	音频、视频、触觉	47	千万级



（a）石头　（b）丝绸　（c）黄铜　（d）麻布　（e）木板　（f）白纸　（g）氨纶　（h）涤纶

图 2　VisTouch 数据集示例

3　跨模态信号重建框架

在 6G 跨模态通信中，音频、视频、触觉信号经过编码传输后，信号需要在接收端进行解码、重建。由于信号在传输过程中不可避免地被噪声干扰而出现缺失、失真现象，因此需要在接收端设计高效、精准的跨模态信号重建框架来弥补信号的缺失。在信号重建中，需要考虑以下 2 个技术难点：1) 如何建立多模态信号的语义空间以跨越不同模态的“壁垒”；2) 如何在技术上保证所重建信号的精准性。

考虑到不同模态信号具有深层次的语义关联性，本节提出一种内在语义关联驱动下的跨模态信号重建框架，如图 3 所示，包含特征提取模块、重建模块、评估模块 3 个部分。特征提取模块将源模态信号映射为公共语义空间中的语义特征向量，重建模块将此语义特征向量反变换为目标模态信号，2 种模块的级联结构是跨越模态“壁垒”的关键；评估模块从语义维度、信号本身的时空维度对重建质量进行评估，并在框架训练过程中反馈优化信息给特征提取模块与重建模块，形成闭环回路，通过不断迭代实现精准信号重建。

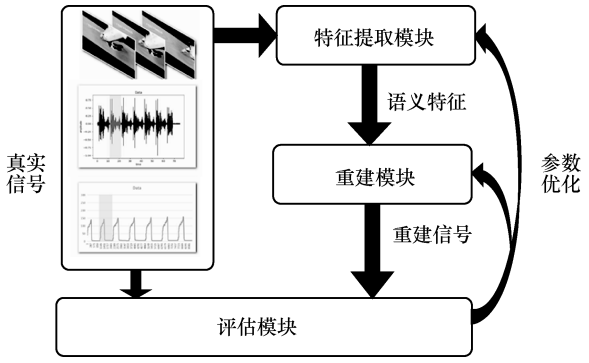


图 3　跨模态信号重建框架

3.1　特征提取模块

特征提取模块用于从源模态信号中获取上下文语义表征。相对于人工设计的特征提取算子（如 SIFT），深度学习（如 CNN、RNN 等）可通过多层卷积层、长短期记忆网络的处理获取更抽象、更深层次的语义特征，有助于表征原始信号中的关键信息。

针对音频信号，由于声音采集时容易引入噪声，故首先对音频信号进行降噪处理（如谱减法、LMS 自适应滤波器）；然后利用时频分析直观、精确的优点，对降噪后的音频信号计算梅尔频谱（MS, Mel spectrogram）、梅尔频率倒谱系数（MFCC, Mel-frequency cepstral coefficients）等；最后将 MS 或 MFCC 输入 CNN 中进行音频信号的语义表征。

针对视频信号，首先将视频帧送入 CNN 中（如 VGG、ResNet），得到不同视频帧所对应的特征图，然后对所有视频帧的特征图进行合并、池化处理，将视频信息转化为语义特征向量。近年来，3D CNN 在行为识别、视频理解等领域开始崭露头角。与传统以 2D 卷积为基本操作的 CNN 不同的是，3D CNN 采用 3D 卷积，即在高度、宽度 2 个维度之外增加了时间维度，使 3D 卷积核提取到帧间相关特征，相对于传统 2D 卷积仅考虑单帧信息，3D CNN 更适合处理视频信号。

针对触觉信号，由于采集设备的不同，其信号质量、结构各异。对于 GelSight、触觉手套，其采集到的信号一般为类似图像的 2D 矩阵，因此，可使用 CNN 方法进行处理；对于压力传感器，其采集到的信号一般为 1D 时间序列信号，因此，可使用 RNN 进行处理，以提取信号的时间语义特征，此外，近年来，RNN 系列方法发展迅速，演变出了如长短期记忆网络、门循环单元（GRU, gate recurrent unit）、Transformer 等时间序列特征提取模型，并在自然语言处理、时间序列预测等任务下表现优秀。同样，也可采用类似音频信号的处理方法，对触觉信号进行时频分析，如对触觉信号做短时傅里叶变换（STFT, short time Fourier transform）等，再将其送入 CNN 进行触觉信号的语义表征。

3.2　重建模块

重建模块对特征提取模块输出的一个模态的语义特征进行反变换，得到另一模态的重建信号。同样，根据信号结构的不同，所采取的重建方法也不同。具体而言，若目标模态信号为图像、频谱图，可采用反卷积或转置卷积的方法对语义特征进行变换，通过逐层、多次处理，使语义信号的结构恢复到与目标信号相一致的状态；若目标模态信号为时间序列信号，可使用基于 RNN 的解码器，某时刻的解码器输出为下一时刻的解码器输入，通过不断迭代，最终生成与目标模态信号相同长度的重建信号。

3.3　评估模块

评估模块用于评价重建信号是否与真实信号相一致，同时在训练过程中可将重建信号与真实信

号的偏差进行梯度的反向传播，调整特征提取模块、重建模块的训练参数，直至重建信号质量满足要求或偏差无法继续优化，通过这种方式使整个框架挖掘多模态信号间的内在语义关联性，最终生成准确、低噪的重建信号。

在实际应用中，通常使用损失函数与 GAN 辨别网络的组合进行评估，GAN 辨别网络对所重建的信号进行判别，区分其真实性，输出“真”或“假”2 种结果，当 GAN 辨别网络将重建信号多次判别为“真”时，即表明重建信号逼近真实信号。

4 视频辅助的触觉重建模型

为了验证 VisTouch 的实用性以及所提出的跨模态信号重建框架的可靠性，本节以视频重建触觉为例，将重建框架具体化，同时结合视频信号与触觉信号的特点设计 3 个子网络：基于 3D CNN 的视频特征提取网络（对应特征提取模块）、GAN 生成网络（对应重建模块）以及 GAN 辨别网络（对应评估模块），视频辅助的触觉重建模型如图 4 所示，最后通过信号可视化、模型结果对比等展示重建效果。

4.1 基于 3D CNN 的视频特征提取

由于 2D 卷积只能提取空间特征，且传统特征设计复杂、不能很好地捕捉视频中的语义信息，3D 卷积应运而生。3D 卷积核在时间维度的扩展能够使其捕捉到时间语义信息，故被广泛用于基于视频的人体行为识别、行为理解等任务。

本文首先将视频输入 3D CNN 中提取视频语义特征，考虑到迁移学习有助于加快模型收敛、保证模型初始性能，因此先使用 ImageNet 预训练好的 3D ResNet50 网络，再使用 VisTouch 数据集进行微调。ResNet50 凭借其独特的跳层连接技术降低了模型训练的难度，使目标函数快速收敛、模型泛化能力提升，一经推出，便成为图像分割、目标检测等计算机视觉下游任务中常用的主干网络。3D ResNet50 是 ResNet50 的 3D 卷积版本，实现了从传统 2D 空间特征提取到 3D 时空特征提取的跨越。

在模型处理上，首先，假设输入视频为五维张量 $\boldsymbol{I}\in\mathbb{R}^{N\times T\times C\times H\times W}$，其中 N 为批处理量，T 为视频帧数，C 为图像通道数，对于 RGB 图像 $C=3$，H 和 W 分别为图像的高度和宽度，这里对每个视频帧图像进行缩放、裁剪的预处理，使图像大小统一为 224×224，即 $H=W=224$；其次，将 $\boldsymbol{I}$ 输入 3D ResNet50，经多层 3D 卷积处理，输出特征图为 $\boldsymbol{F}\in\mathbb{R}^{N'\times T'\times C'\times H'\times W'}$，对于 3D ResNet50 而言，$T'=2$，$C'=2\,048$，$H'=W'=7$，为了便于后续 GAN 生成网络的处理，本文对 $\boldsymbol{F}$ 进行形状变换，得到四维张量 $\boldsymbol{F}_{\boldsymbol{R}}\in\mathbb{R}^{N'\times T'C'\times H'\times W'}$，表示视频语义特征，其中 $T'C'=2\times2\,048=4\,096$。

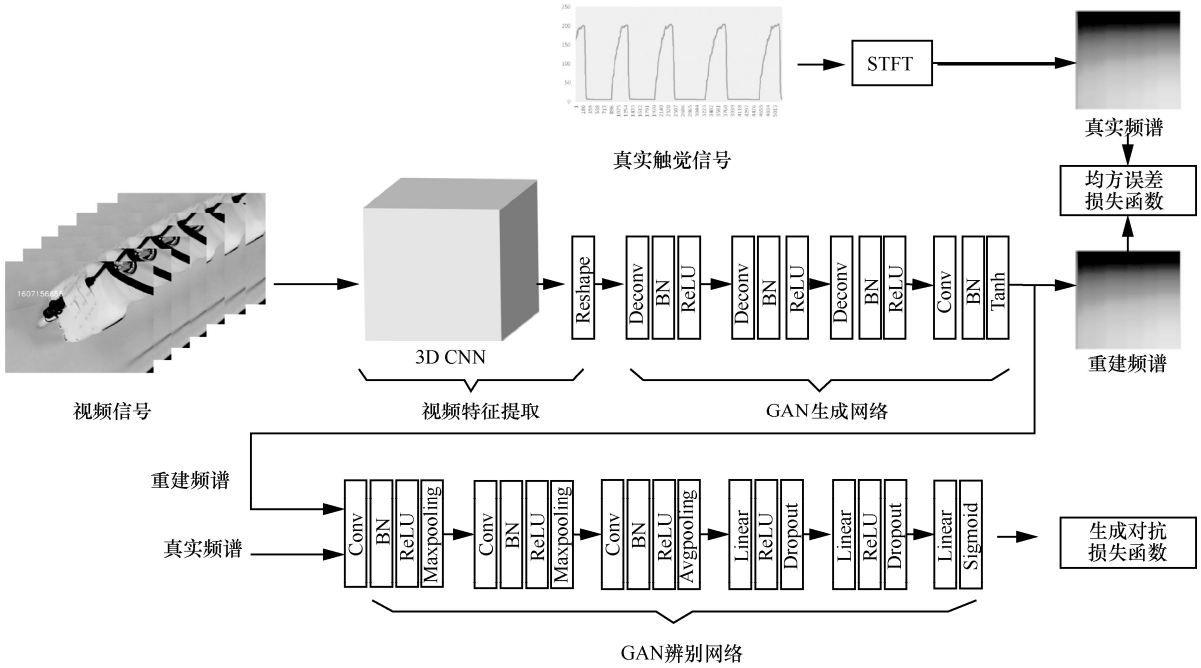


图 4 视频辅助的触觉重建模型

4.2 GAN 生成网络

GAN 生成网络用于将视频语义特征精准、低噪地重建为同一样本所对应的触觉信号。GAN 的框架包含一对互相对抗的模型：生成网络与辨别网络。生成网络用于尽可能逼近真实数据的分布；辨别网络用于正确区分真实数据与生成数据，从而最大化判别准确率。为了在博弈中胜出，两者需要不断提高各自的生成能力和辨别能力，直至达到两者间的纳什均衡[25]。

GAN 生成网络的设计需要考虑目标模态信号的结构与形式，触觉信号 $\boldsymbol{T}$ 一般为时间序列形式，在触觉信号重建中，借鉴音频信号预处理方法，使用 STFT（采样频率 1 000 Hz，窗宽 50）对触觉信号 $\boldsymbol{T}$ 进行频谱变换，得到 26×41 的复数矩阵，分离实数部分与虚数部分，得到 $2\times26\times41$ 的触觉频谱 $\boldsymbol{S}$（忽略批处理量 N）。因此，GAN 生成网络的作用在于利用上述视频语义特征 $\boldsymbol{F_R}$ 重构触觉频谱 $\hat{\boldsymbol{S}}$，通过优化使 $\hat{\boldsymbol{S}}$ 与真实频谱 $\boldsymbol{S}$ 相接近，$\hat{\boldsymbol{S}}$ 再经过频谱反变换得到重建的触觉时间信号 $\hat{\boldsymbol{T}}$，实现从视频信号 $\boldsymbol{I}$ 到触觉信号 $\hat{\boldsymbol{T}}$ 的跨模态映射。

在模型设计上，视频语义特征 $\boldsymbol{F_R}$ 主要通过反卷积层、批归一化层、激活函数的处理实现到目标 $\hat{\boldsymbol{S}}$ 的转换。对于反卷积层的设置如层数、卷积核尺寸、步长、补零数量等，有多种配置方案，这些参数与反卷积层输入、输出张量之间的关系为

$$H_{\text{out}}=(H_{\text{in}}-1)s+k_{\text{h}}-2p_{\text{h}} \tag{1}$$

$$W_{\text{out}}=(W_{\text{in}}-1)s+k_{\text{w}}-2p_{\text{w}} \tag{2}$$

其中，H_{in}、W_{in} 分别代表反卷积层输入张量的高度、宽度，H_{out}、W_{out} 分别代表反卷积层输出张量的高度、宽度，s 代表卷积核的滑动步长，k_{h}、k_{w} 分别代表卷积核高度、宽度，p_{h}、p_{w} 分别代表高度、宽度方向上的补零数量。

本文利用式(1)、式(2)设计 5 层模块，如表 4 所示，注意这只是一种配置方案，主要目的是验证第 3 节所提跨模态信号重建框架的可靠性。第 1 层模块（1.1）为输入层，第 2 层模块（2.1、2.2、2.3）、第 3 层模块（3.1、3.2、3.3）、第 4 层模块（4.1、4.2、4.3）均为反卷积层 Deconv、批归一化（BN，batch normalization）层、ReLU 激活函数的组合，用于重构出频谱图的高度与宽度，第 5 层模块（5.1、5.2、5.3）为1×1卷积层、批归一化层、Tanh 激活函数的组合，用于重构出频谱图的通道维度。此外，在表 4 中，使用 $k=(k_{\text{h}},k_{\text{w}})$ 表示卷积核尺寸，$p=(p_{\text{h}},p_{\text{w}})$ 表示补零数量，$C_{\text{out}}\times H_{\text{out}}\times W_{\text{out}}$ 表示输出张量尺寸，其中 C_{out} 表示张量通道数。

表 4　GAN 生成网络参数（忽略批处理量）

模块序号	模块参数	输出张量尺寸
1.1	输入	4 096×7×7
2.1	反卷积层 $k=(3,3),p=(1,0),s=1$	256×7×9
2.2	批归一化层	256×7×9
2.3	ReLU 激活函数	256×7×9
3.1	反卷积层 $k=(2,5),p=(0,0),s=2$	128×14×21
3.2	批归一化层	128×14×21
3.3	ReLU 激活函数	128×14×21
4.1	反卷积层 $k=(4,5),p=(2,2),s=2$	64×26×41
4.2	批归一化层	64×26×41
4.3	ReLU 激活函数	64×26×41
5.1	卷积层 $k=(1,1),p=(0,0),s=1$	2×26×41
5.2	批归一化层	2×26×41
5.3	Tanh 激活函数	2×26×41

4.3 GAN 辨别网络

GAN 辨别网络用于对重建频谱、真实频谱进行特征提取与区分，当辨别网络认为输入频谱是真实时输出 1，反之输出 0，通过损失函数将辨别结果反馈回 3D CNN 与 GAN 重建网络，使其生成精度更高、噪声更低的重建频谱，直至辨别网络无法区分数据来源。

在模型设计上，将真实触觉频谱 $\boldsymbol{S}$ 与 GAN 生成网络的重建触觉频谱 $\hat{\boldsymbol{S}}$ 作为 GAN 辨别网络输入，经过 2 个卷积组的处理，得到判别向量，其中，一个卷积组包含一个 3×3 卷积层、一个批归一化层、一个 ReLU 激活函数以及一个最大池化层 Maxpooling；然后，将判别向量依次输入全连接层及 Sigmoid 激活函数进行二值真假判别。

4.4 损失函数

损失函数用于有监督地优化网络参数。由于 GAN 采用自我博弈的思想进行训练，即通过生成网络与辨别网络之间的竞争，保证重建信号的精准性，GAN 的损失函数为

$$\min_{G}\max_{D}V(D,G)=$$
$$\mathrm{E}_{\boldsymbol{S}\sim P_{\text{data}}(\boldsymbol{S})}[\log D(\boldsymbol{S})]+\mathrm{E}_{\hat{\boldsymbol{S}}\sim P_{\text{data}}(\hat{\boldsymbol{S}})}[1-\log D(G(\hat{\boldsymbol{S}}))] \tag{3}$$

其中，$\mathrm{E}(\cdot)$ 表示期望函数，$G(\cdot)$ 和 $D(\cdot)$ 分别表示 GAN 生成网络与 GAN 辨别网络，$P_{\text{data}}(\cdot)$ 表示数据

分布。

GAN 是生成网络、辨别网络交替迭代训练的，所以式(3)可以拆分为式(4)和式(5)这 2 种形式。首先，固定生成网络 $G(\cdot)$ 参数，对辨别网络 $D(\cdot)$ 进行优化，即

$$\max_{D} V(D,G) = \mathrm{E}_{\boldsymbol{S}\sim P_{\text{data}}(\boldsymbol{S})}[\log D(\boldsymbol{S})] + \mathrm{E}_{\hat{\boldsymbol{S}}\sim P_{\text{data}}(\hat{\boldsymbol{S}})}[1-\log D(G(\hat{\boldsymbol{S}}))] \quad (4)$$

辨别网络的优化目标为尽可能准确地区分 $\hat{\boldsymbol{S}}$ 与 $\boldsymbol{S}$，由于辨别网络输出 1 时代表真实，输出 0 时代表虚假，故在训练过程中希望 $D(\boldsymbol{S})$ 趋近于 1，$D(G(\hat{\boldsymbol{S}}))$ 趋近于 0，即使总体损失值递增。

其次，固定辨别网络 $D(\cdot)$ 参数，对生成网络 $G(\cdot)$ 进行优化，即

$$\min_{G} V(D,G) = \mathrm{E}_{\hat{\boldsymbol{S}}\sim P_{\text{data}}(\hat{\boldsymbol{S}})}[1-\log D(G(\hat{\boldsymbol{S}}))] \quad (5)$$

生成网络的优化目标为输出更逼真的重建频谱 $\hat{\boldsymbol{S}}$，因此希望辨别网络将重建频谱误判为真实频谱，即希望 $D(G(\hat{\boldsymbol{S}}))$ 趋近于 1，使总体损失值递减。

此外，为了进一步提升重建质量，本文同时利用均方误差损失函数对重建频谱的内容进行约束，均方误差损失函数表示为

$$L_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(s_i - \hat{s}_i)^2 \quad (6)$$

其中，s_i 与 $\hat{s}_i$ 分别代表真实频谱 $\boldsymbol{S}$ 与重建频谱 $\hat{\boldsymbol{S}}$ 的第 i 个位置的元素，n 代表频谱中的元素数量。

生成对抗损失函数、均方误差损失函数分别从真实性鉴别、内容 2 个角度重建质量，将两者组合成的复合函数用于网络训练，经过实验证明，可以实现准确、低噪的跨模态信号重建。

4.5　实验及分析

本文使用 VisTouch 数据集进行视频辅助的触觉重建网络训练，网络训练使用随机梯度下降（SGD, stochastic gradient descent）法进行优化，设置训练轮次为 70，初始学习率为 0.001，并使用余弦退火（CA, cosine annealing）调整器不断调整学习率，批处理量为 6，3D CNN 输入尺寸为 224×224，整个模型使用 Pytorch 深度学习框架进行编程开发。在硬件配置上，使用单张 RTX 2080Ti 显卡进行模型训练。

网络训练过程中的生成对抗损失函数优化曲线如图 5 所示，其中，生成网络损失值（对应式(5)）不断下降，说明重建的频谱越来越逼真，辨别网络损失值（对应式(4)）不断上升，说明辨别网络对数据来源的鉴别能力越来越强。

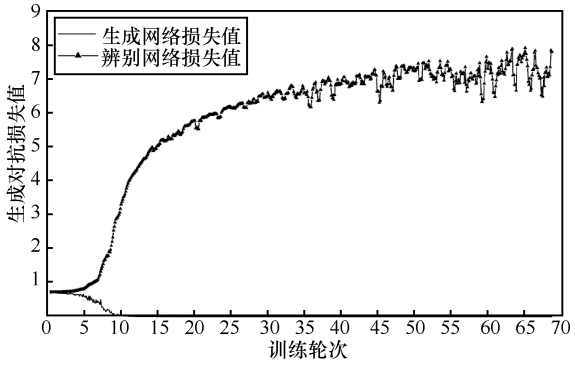


图 5　生成对抗损失函数优化曲线

此外，网络训练过程中的均方误差损失函数（对应式(6)）优化曲线如图 6 所示，说明重建频谱与真实频谱在内容上越来越接近。

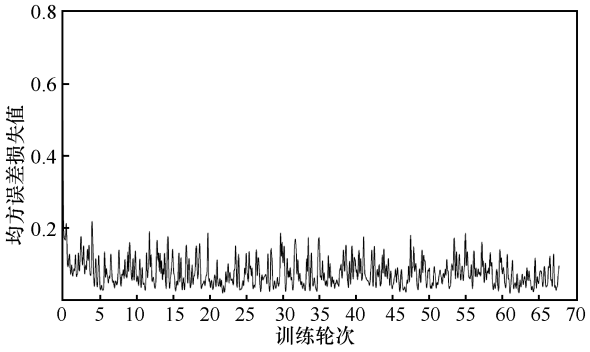


图 6　均方误差损失函数优化曲线

为了进一步说明本文所提模型的重建性能，本节进行了模型对比实验，由于本文利用 VisTouch 完成触觉重建工作，暂无已公开的基准模型，为此，本节对所提视频辅助的触觉重建模型进行约简，得到以下 2 种模型作为比较基准。

模型 1　不改变模型结构，仅使用生成对抗损失函数（式(3)）训练本文模型。

模型 2　移除 GAN 辨别网络，仅使用均方误差损失函数（式(6)）训练模型。

确定比较基准后，需要引入评估指标来测试输出结果，本文使用 2 种评估指标，即平均绝对误差（MAE, mean absolute error）与准确度（ACC, accuracy）进行度量。

MAE。MAE 用于评估重建信号与真实信号的绝对偏差。由于触觉信号的表征形式为时间序列，因

此，从信号本身出发，假设真实触觉时间信号为 $\boldsymbol{T}$，重建出的触觉时间信号为 $\hat{\boldsymbol{T}}$，样本容量为 M，则 MAE 计算式为

$$\mathrm{MAE}=\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{T}_m-\hat{\boldsymbol{T}}_m \tag{7}$$

ACC。首先，利用真实信号预训练一个样本类别分类器，训练完成后输入重建信号，检验重建信号对样本类别的判别结果是否与真实样本类别一致，从而统计精确度 ACC。在本文中，该分类器由多层感知机实现。

模型对比实验统计结果如表 5 所示，同时本文将某段触觉信号的重建结果进行可视化，如图 7 所示。从表 5 和图 7 中可以看出，模型 1 尽管可使 MAE 达到 0.093 3，但实际样本分类 ACC 仅为 0.57，且图 7 所示的重建信号与真实信号差异较大，甚至无法恢复其包络；模型 2 效果相较于模型 1 有较大提升（MAE=0.058 6，ACC=0.65），可视化结果反映出其重建信号已基本重建出信号包络及拐点，但仍有部分噪声。本文综合模型 1 与模型 2，既采用生成对抗网络的基本架构，又使用复合损失函数进行约束，使 MAE 与 ACC 分别达到 0.013 5 与 0.78，图 7 也直观地反映出本文模型大大提升了跨模态信号重建的准确性和低噪性。

表 5　　模型对比实验统计结果

模型	MAE	ACC
模型 1	0.093 3	0.57
模型 2	0.058 6	0.65
本文模型	0.013 5	0.78

4.6　应用平台

为了将本文所提跨模态信号重建框架落地到实际应用场景，本节搭建了一种如图 8 所示的遥操作平台，用于工业场景下实现远程抓取物体的任务。

在配置方面，该遥操作平台主要分为硬件平台及软件平台，其中，本文所提跨模态信号重建框架由软件平台中的深度学习嵌入式子平台进行实现，并烧录进跨模态编解码器中，该编解码器可使用 NVIDIA Jetson 套件实现，用于对音频、视频、触觉信号进行跨模态编码及信号重建；基于混合现实的遥操作系统硬件上主要由 Microsoft Hololens2 混合现实眼镜与力反馈设备 Touch 笔组成，用于渲染实时交互场景，便于主端的用户控制从端机械臂完成抓取物体的操作。

在平台操作方面，首先主端的用户佩戴 Hololens2 眼镜，手握 Touch 笔，当控制 Touch 笔移动时，移动控制命令即从主端发出，经 6G 传输，抵达远处从端机械臂，经过机器人逆向运动学运算，控制机器人完成抓取物体的操作。同时，在抓

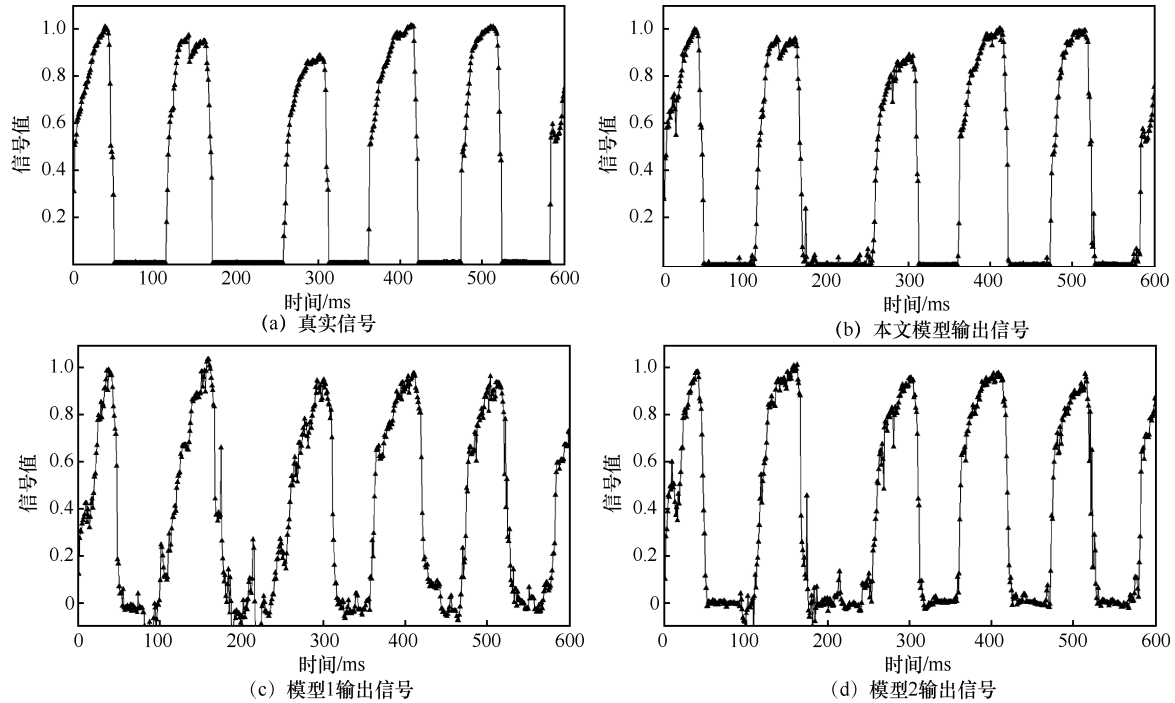


图 7　重建信号与真实信号对比

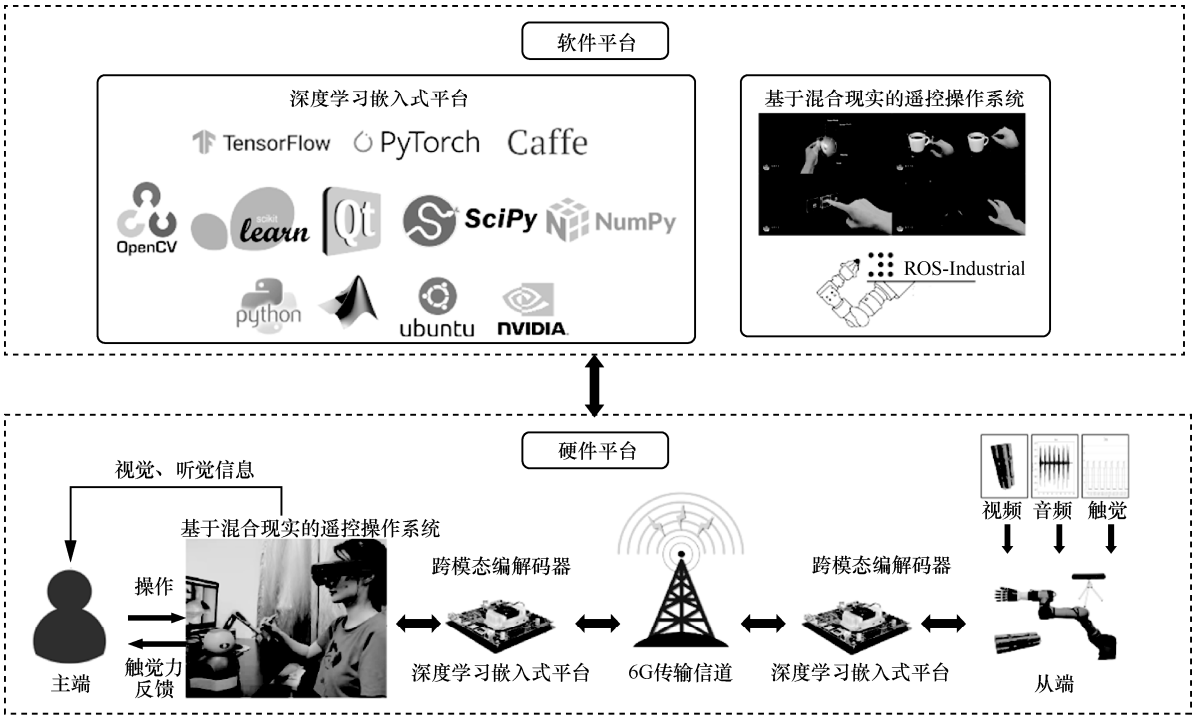


图 8　遥操作平台

取过程中，利用表 1 的设备进行同步的音频、视频、触觉信号采集，经跨模态编码后，再通过 6G 传输链路反馈回主端，主端接收到信号后，经过本文所提跨模态信号重建技术处理，对传输过程中的缺失信号进行补全，从而解决传输过程中的丢包问题。最后，音频、视频信息呈现在 Hololens2 眼镜中，触觉信息由 Touch 笔进行力渲染，使在主端的用户可同步感受到从端机械臂抓取物体的触感，从而实现多模态信息带来的临场感。

由于 6G 网络还未落地，故该遥操作系统的通信链路基于现有商用 5G 网络，具体实现过程中，使用华为 5G 随身 Wi-Fi E6878-870 及中国移动 5G SIM 卡，理论峰值速度为 1.65 Gbit/s，5G 频段 N41。在 5G 基础上测试不仅可保证重建模型的正常运行，还可发现当前 5G 传输中存在的不足，从而对 6G 的传输性能及功能提出需求。

对于本文所提重建模型，不仅有精度要求，同时为保证用户的沉浸式体验感，还需对发出控制信号与接收多模态反馈信号之间的时间差进行要求。为此，对于该遥操作平台，本文分别使用 MAE、发送与反馈总时延、重建模型时延、触觉真实性满意度、时延满意度 5 种指标对遥操作平台进行评估。其中，触觉真实性满意度指渲染出的触觉力信息是否与实际触摸感觉一致，时延满意度指对上述时延的接受程度。为了度量这 2 种满意度，本文采用了问卷调查的方式，首先，15 位感官正常的志愿者（8 男 7 女）使用该遥操作平台在 10 m 外控制从端机械臂抓取玻璃瓶，并从 2 种满意度角度进行打分，分数范围为 1～5，1 分代表不满意，5 分代表完全满意，统计 15 人打分结果的均值和方差，最终结果如表 6 所示。

表 6　　遥操作实验评估结果

指标	评估结果
MAE	0.012 6
发送与反馈总时延/ms	127
重建模型时延/ms	98
触觉真实性满意度（均值，方差）	（4.43, 0.72）
时延满意度（均值，方差）	（3.87, 1.07）

从表 6 可知，本文所提跨模态重建算法在准确度方面满意度较高，触觉真实性满意度均值处于 4～5 分段，表示用户对重建出的触觉信号基本满意，然而，发送与反馈总时延较大（127 ms），该总时延主要包含信号传输时延、重建模型时延等，其中，本文基于深度学习的重建模型在 NVIDIA Jetson 平台上的耗时较大（约 98 ms），对于用户而言，时延满意度均值为 3.87，方差为 1.07，说明用户可感受到的反馈信号相对于近端操作信号的滞后偏差，另一方

面，在该系统中，基于 5G 的信号传输时延（约 29 ms）在 6G 各种新兴技术的加持下有进一步的优化空间，即本文搭建的遥操作平台需要 6G 低时延技术来构造即时临场感。

上述数据表明，尽管重建模型可有效解决信号的丢包、缺失问题，实现高可靠性，但随之带来的计算复杂度过高，显著影响了用户体验，无法满足低时延要求。因此，若运行重建模型的跨模态编解码器应用于 6G 通信系统中，亟须解决多模态信号在 6G 传输、处理过程中的高时延问题，在 AI 赋能 6G 网络的趋势下，跨模态信号重建技术对 6G 传输功能和性能的要求有以下两点。

1) 高效轻量的机器学习技术。多层堆叠处理的深度学习模型参数量巨大，常用的卷积神经网络如 ResNet 具有上百层卷积、池化层以及百万级参数，无法满足低时延处理要求，为此，需要在 6G 网络中引入高效轻量的机器学习技术。在硬件方面，可采用专用集成电路如张量处理单元（TPU, tensor processing unit）代替图形处理单元（GPU, graph processing unit）完成张量的高并行处理；在软件方面，可采用知识蒸馏、剪枝、量化等技术缩小模型大小，使每秒浮点操作数（FLOPS, flooting-point operations per second）降到 1×10^9 次以下，满足大多数实时处理要求。

2) 极低的端到端时延。6G 网络下，不仅需要人与人的通信，而且需要人与物、物与物、物与环境之间的交互与通信，且基于多模态信号的沉浸式体验需求加大，数据量剧增，因此，6G 需在 5G 的基础上，利用太赫兹通信、可见光通信、超大规模天线、量子通信与计算等技术，进一步提升移动宽带和物联网场景的低时延通信能力，使峰值速率达到 100 Gbit/s～1 Tbit/s、通信时延达到 50～100 μs，相比于 5G 整体性能提升 10～100 倍。

5　结束语

针对 6G 跨模态通信中的信号恢复、重建问题，本文构造了包含音频、视频、触觉的大规模数据集 VisTouch，并在深度学习背景下，提出了一种跨模态信号重建框架，包含特征提取模块、重建模块、评估模块，并以 VisTouch 中具体实例出发，设计了基于 3D CNN 与 GAN 的触觉重建模型，进一步验证了所提框架的合理性，同时也充分证明了多模态信号间具有内在语义关联性。未来将进一步扩展跨模态互补机理研究，发掘多模态信号在维度、结构、内容、逻辑、时间、空间的语义关联性，构建六维语义空间，稳健、低噪、准确、快速地将信号原始域映射到目标域，从而在 6G 跨模态通信中大幅缩减信号传输量，为用户提供更流畅、高保真的沉浸式应用。

参考文献：

[1] 中国信息通信研究院. 6G 总体愿景与潜在关键技术白皮书[R]. 2021.
China Academy of Information and Communications Technology. 6G overall vision and potential key technology white paper[R]. 2021.

[2] VAN D B D, GLANS R, KONING D D, et al. Challenges in haptic communications over the tactile Internet[J]. IEEE Access, 2017, 5: 23502-23518.

[3] ZHOU L, WU D, CHEN J X, et al. Cross-modal collaborative communications[J]. IEEE Wireless Communications, 2020, 27(2): 112-117.

[4] WEI X, ZHOU L. AI-enabled cross-modal communications[J]. IEEE Wireless Communications, 2021, 28(4): 182-189.

[5] 高赟, 魏昕, 周亮. 跨模态通信理论及关键技术初探[J]. 中国传媒大学学报(自然科学版), 2021, 28(1): 55-63.
GAO Y, WEI X, ZHOU L. Preliminary study on theory and key technology of cross-modal communications[J]. Journal of Communication University of China (Science and Technology), 2021, 28(1): 55-63.

[6] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[7] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv Preprint, arXiv: 1409.1556, 2014.

[8] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[9] BAZZICA A, VAN GEMERT J C, LIEM C C S, et al. Vision-based detection of acoustic timed events: a case study on clarinet note onsets[J]. arXiv Preprint, arXiv: 1706.09556, 2017.

[10] LI B C, LIU X Z, DINESH K, et al. Creating a multitrack classical music performance dataset for multimodal music analysis: challenges, insights, and applications[J]. IEEE Transactions on Multimedia, 2019, 21(2): 522-535.

[11] ZHAO H, GAN C, ROUDITCHENKO A, et al. The sound of pixels[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 570-586.

[12] MONTESINOS J F, SLIZOVSKAIA O, HARO G. Solos: a dataset for audio-visual music analysis[C]//Proceedings of 2020 IEEE 22nd International Workshop on Multimedia Signal Processing. Piscataway: IEEE Press, 2020: 1-6.

[13] KURMI V K, BAJAJ V, PATRO B N, et al. Collaborative learning to generate audio-video jointly[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2021: 4180-4184.

[14] ROTH J, CHAUDHURI S, KLEJCH O, et al. Ava active speaker: an audio-visual dataset for active speaker detection[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2020: 4492-4496.

[15] TSUCHIDA S, FUKAYAMA S, HAMASAKI M, et al. AIST dance video database: multi-genre, multi-dancer, and multi-camera database for dance information processing[C]//Proceedings of the 20th International Society for Music Information Retrieval Conference. [S.l.:s.n.], 2019: 501-510.

[16] LI R L, YANG S, ROSS D A, et al. AI choreographer: music conditioned 3D dance generation with AIST++[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 13381-13392.

[17] HONG S, IM W, YANG H S. Content-based video-music retrieval using soft intra-modal structure constraint[J]. arXiv Preprint, arXiv: 1704.06761, 2017.

[18] LI Y Z, ZHU J Y, TEDRAKE R, et al. Connecting touch and vision via cross-modal prediction[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 10601-10610.

[19] YUAN W Z, DONG S Y, ADELSON E H. GelSight: high-resolution robot tactile sensors for estimating geometry and force[J]. Sensors (Basel, Switzerland), 2017, 17(12): 2762.

[20] SUNDARAM S, KELLNHOFER P, LI Y Z, et al. Learning the signatures of the human grasp using a scalable tactile glove[J]. Nature, 2019, 569(7758): 698-702.

[21] DUAN B, WANG W, TANG H, et al. Cascade attention guided residue learning GAN for cross-modal translation[C]//Proceedings of 2020 25th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE Press, 2021: 1336-1343.

[22] HAO W L, ZHANG Z X, GUAN H. CMCGAN: a uniform framework for cross-modal visual-audio mutual generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 6886-6893.

[23] CHATTERJEE M, CHERIAN A. Sound2Sight: generating visual dynamics from sound and context[C]//European Conference on Computer Vision. Berlin: Springer, 2020: 701-719.

[24] WEI X, SHI Y Y, ZHOU L. Haptic signal reconstruction for cross-modal communications[J]. IEEE Transactions on Multimedia, 2021: doi.org/10.1109/TMM.2021.3119860.

[25] 王万良, 李卓蓉. 生成式对抗网络研究进展[J]. 通信学报, 2018, 39(2): 135-148.
WANG W L, LI Z R. Advances in generative adversarial network[J]. Journal on Communications, 2018, 39(2): 135-148.

[作者简介]

李昂（1995− ），男，河南周口人，南京邮电大学博士生，主要研究方向为多媒体通信、人工智能。

陈建新（1973− ），男，江苏南通人，博士，南京邮电大学副教授、硕士生导师，主要研究方向为无线通信、人机交互。

魏昕（1983− ），男，江苏南京人，博士，南京邮电大学教授、硕士生导师，主要研究方向为多媒体通信。

周亮（1981− ），男，安徽芜湖人，博士，南京邮电大学教授、博士生导师，主要研究方向为多媒体通信。