文章编号：2095-4980(2025)08-0804-12

基于改进MobileViT模型的毫米波雷达动态手势识别方法

葛志洲[1,3]，张向群[2,3]，申佳文[2,3]，杜根远[*2,3]，刘锋涛[4]

(1.华北水利水电大学 信息工程学院，河南 郑州 450046；2.河南省偏振感知与智能信号处理国际联合实验室，河南 许昌 461000；
3.许昌学院 信息工程学院，河南 许昌 461000；4.许昌初心智能电气科技有限公司，河南 许昌 461111)

摘　要：利用毫米波雷达进行手势识别具有非接触、检测精确度高、不侵犯用户隐私、环境适应性好等优点，在工业人机交互、智能家居等场景具有广泛的应用。但现有毫米波雷达动态手势识别方法存在模型复杂度高，计算成本大，以及识别准确率低、推理速度慢等问题。为此，本文提出基于改进的轻量级MobileViT网络的手势识别方法，在保持高识别准确度的同时降低计算复杂度，以满足嵌入式设备的部署需求。首先，采集动态手势动作的毫米波雷达回波信息，消除设备噪声和背景干扰后，重组数据为采样点数×脉冲数×帧数三维数据矩阵；利用傅里叶变换生成手势动作的距离-时间图像和多普勒-时间图像，将特征图输入到改进后的MobileViT网络模型中进行特征提取和融合，输出手势动作识别结果。实验结果表明，所构建的MobileViT模型参数空间复杂度降低到0.167 M，计算复杂度为0.253 GFLOPs；该方法在12种手势类型的数据集中进行验证，识别准确率为99.31%，证明了该方法的有效性。

关键词：手势识别；人机交互；毫米波雷达；轻量级神经网络

中图分类号：TN711　　**文献标志码：**A　　**DOI：**10.11805/TKYDA2025114

A dynamic hand gesture recognition method of mmWave radar based on improved MobileViT model

GE Zhizhou[1,3]，ZHANG Xiangqun[2,3]，SHEN Jiawen[2,3]，DU Genyuan[*2,3]，LIU Fengtao[4]

(1.School of Information Engineering，North China University of Water Resources and Electric Power，Zhengzhou Henan 450046，China；
2.Polarization Perception and Intelligent Signal Processing International Joint Laboratory of Henan Province，
Xuchang Henan 461000，China；3.School of Information Engineering，Xuchang University，Xuchang Henan 461000，China；
4.Xuchang Chuxin Intelligent Electrical Technology Co.，Xuchang Henan 461111，China)

Abstract：Gesture recognition using millimeter-wave(mmWave) radar offers advantages such as contact-free operation, high detection accuracy, privacy preservation, and robust environmental adaptability, making it promising for industrial human-machine interaction and smart-home applications. However, existing mmWave-based dynamic-gesture recognition approaches suffer from high model complexity, large computational cost, low accuracy, and slow inference speed. To address these challenges, a lightweight gesture-recognition method is proposed based on an improved MobileViT network that maintains high accuracy while significantly reducing computational complexity for deployment on embedded devices. Firstly, dynamic-gesture echoes are captured with an mmWave radar. After suppressing device noise and background clutter, the data are reorganized into a 3-D matrix (sample points × chirps × frames). Range-time and Doppler-time maps are then generated via Fourier

收稿日期：2025-04-07；**修回日期：**2025-05-26

基金项目：河南省科技厅科技攻关资助项目(242102210067)；河南省重点研发资助项目(241111212500)

[*]**通信作者：**杜根远　email:xcdgy@163.com

引用格式：葛志洲,张向群,申佳文,等. 基于改进MobileViT模型的毫米波雷达动态手势识别方法[J]. 太赫兹科学与电子信息学报, 2025,23(8): 804-815. DOI:10.11805/TKYDA2025114.

Citation format：GE Zhizhou,ZHANG Xiangqun,SHEN Jiawen,et al. A dynamic hand gesture recognition method of mmWave radar based on improved MobileViT model[J]. Journal of Terahertz Science and Electronic Information Technology, 2025, 23(8): 804-815. DOI:10.11805/TKYDA2025114.

transform and fed into the enhanced MobileViT model for feature extraction and fusion, yielding the final gesture classification. Experimental results show that the proposed MobileViT model has only 0.167 M of parameter space complexity and 0.253 GFLOPs of computational complexity. Evaluated on a 12-class dynamic-gesture dataset, the method achieves 99.31% of recognition accuracy, demonstrating its effectiveness.

Keywords: gesture recognition; human-computer interaction; mmWave radar; lightweight neural network

手势识别是智能化场景人机交互发展中亟待解决的关键性问题，可用于智能驾驶、健康医疗、智能家居、工业生产等领域，具有极大的应用价值和发展潜力[1-2]。早期的手势识别技术利用可穿戴设备和传感器感知人体手掌与各个关节的空间位置信息，典型代表设备为数据手套。基于光学标记法的穿戴设备也具有良好的识别性和稳定性，但技术操作繁琐且设备价格昂贵。穿戴设备一定程度限制了用户手势灵活操作，未能广泛用于日常生活。基于视觉图像的手势识别克服了可穿戴识别技术对用户的活动限制，该技术利用计算机图像采集设备(如摄像头等)对目标用户的手势动作进行感知、追踪与识别，进而达到理解用户意图的目的。利用高分辨力相机可使视觉手势识别技术准确率高达90%以上，但该技术极大程度上受限于光线条件，且存在侵犯个人隐私的问题。毫米波雷达具有非接触，不受黑夜、雨雪、光照、烟雾等环境限制，全天候全天时，保护用户隐私，易部署等优势[3-4]，利用毫米波雷达进行手势识别能够拓展人机交互场景，构建无接触智能感知终端，实现人类世界与机器共融，是解决移动式人机交互难题最有前景的技术之一[5-6]。

目前常用手势识别的毫米波雷达传感器工作频段为24 GHz、60 GHz、77 GHz[7]。K Kehelella等[8]使用24 GHz双天线连续波多普勒雷达采集14类手势，采用融合卷积编码器-解码器与Vision Transformer的模型，实现了98.3%识别率，但模型复杂度较高，推理开销较大；英国伦敦大学M Ritchie等[9]对6人4种手势进行3 000次的检测，提出的多距离特征方法准确率高达96%，但场景单一且样本规模较小；Google公司发布了基于60 GHz毫米波雷达Soli项目，利用长短期记忆(Long Short-Term Memory，LSTM)网络模型近距离对10种手势进行识别[10]；文献[11]采用发射频率为60 GHz、4天线(2发4收)的调频连续波(Frequency Modulation Continuous Wave，FMCW)雷达传感器，配合SiGe技术实现对目标手势的检测；电子科技大学的李楚杨[12]在77 GHz的频段采用1发4收模式，设计左挥、下拉等8种手势，构建了3 200个手势数据集；江苏科技大学靳标等[13]提出基于串联式1维神经网络(1D-ScNN)的毫米波动态手势识别方法，在一维卷积神经网络中串联LSTM网络层，处理动态手势的帧间相关性，激活手势的时序特征，提高网络的收敛速度和分类精确度；同济大学董连飞[14]提出基于3D CNN-Transformer网络模型的车载毫米波雷达动态手势识别方法，将连续多帧的距离-多普勒和距离-角度图经过三维卷积网络进行特征融合与拼接，识别准确率高达97.14%；WANG等[15]基于220 GHz雷达系统构建了一个具有1 050个样本的数据集，但220 GHz属于太赫兹，成本昂贵，无法在普通场景普及。

上述研究存在以下不足：a) 大部分相关文献手势识别种类在10个以内，样本数量有限，手势识别模型可信度不高；b) 现有文献中的手势识别模型为提高手势识别的准确率，大多依赖距离、多普勒和相位三维数据特征，导致模型复杂度和计算量较高，难以在资源受限的嵌入式设备上高效运行；c) 传统手势识别存在准确率低、推理速度慢等问题。针对上述问题，本文提出一种基于改进的MobileViT模型的毫米波雷达手势识别方法。

1　雷达数据处理

1.1 信号模型

线性FMCW雷达系统框图如图1所示，主要包括发射、接收、混频、滤波与模数转换等模块。发射端产生线性调频信号并经倍频、放大后由天线发射，接收到的目标反射信号与本振信号在混频器中混频，得到中频信号。经低通滤波去除高频成分后，信号通过模数转换器生成差拍信号，作为后续信号处理的输入。

FMCW雷达发射正弦信号频率随时间线性增加，如图2所示。每个雷达的数据帧由多个Chirp信号组成。毫米波雷达发射FMCW，遇到目标物体并反射回来，接收到的反射信号与原始发射信号之间会出现频率差异，通过分析反射信号与原始反射信息之间的差异，可有效获取目标物体相关信息。

毫米波雷达发射信号的FMCW表示为：

$$S_{\mathrm{T}}(t)=A_{\mathrm{T}}\cos\left(2\pi f_{\mathrm{c}}t+2\pi\int_0^t f_{\mathrm{T}}(\tau)\,\mathrm{d}\tau\right)=A_{\mathrm{T}}\cos\left(2\pi f_{\mathrm{c}}t+\pi\times\frac{B}{T_{\mathrm{c}}}t^2\right)\tag{1}$$

式中：A_T为传输信号的振幅；f_c为载波信号的起始频率；B为频率调制信号的带宽；S为调频连续波信号的斜率；T_c为频率调制信号的周期；$f_T(\tau)$为时刻τ对应的调频信号频率。

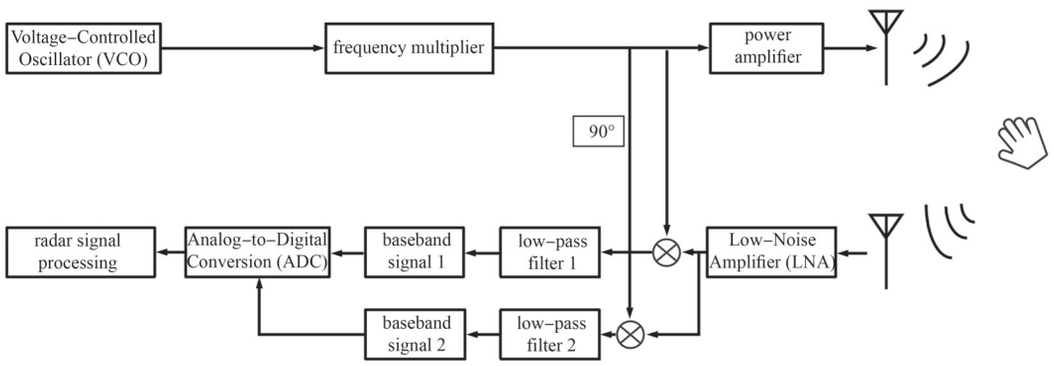


Fig.1 Block diagram of an FMCW radar system

图1 FMCW 雷达系统框图

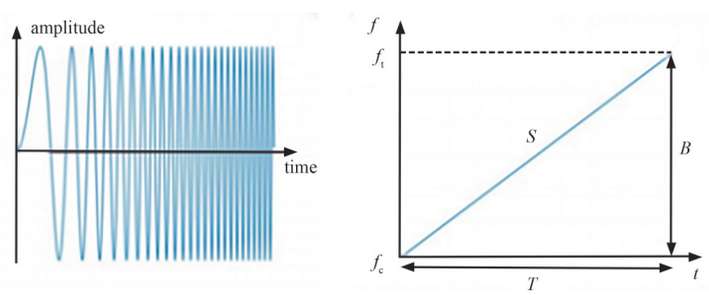


Fig.2 Time-domain and frequency-domain plots of the Chirp signal

图2 Chirp信号的时域与频域图

经过手势反射后，毫米波雷达的接收信号FMCW为：

$$S_R(t)=S_T(t-\tau)+\Delta f_d=A_R\cos 2\pi\left(f_c(t-\tau)+\frac{B}{2T_c}\left(t^2-2t\tau\right)+\Delta f_d t\right) \tag{2}$$

式中：A_R为接收回波信号的振幅；τ为调频信号从发射到接收的延迟时间；Δf_d为多普勒频移值。

接收信号与发射信号混频后得到中频信号，中频信号表达式为：

$$S_{IF}(t)=\frac{A_T A_R}{2}\cos 2\pi\left(\left(\frac{B}{T_c}\tau-\Delta f_d\right)t+f_c\tau\right) \tag{3}$$

从式(3)可以看出，中频信号的频率为一个固定值，数值等于发射信号和接收信号之间频率差值，如图3所示。

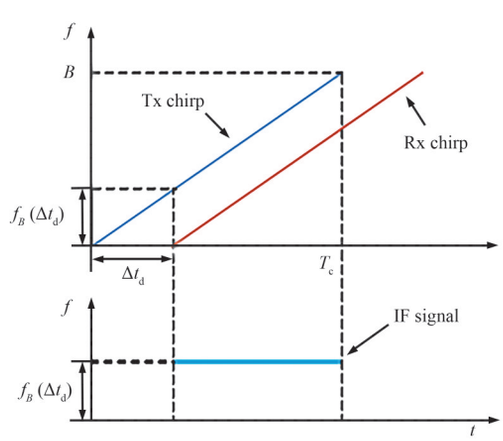


Fig.3 Chirp signal transmission, reception signal, intermediate frequency signal

图3 Chirp发射信号、接收信号和中频信号

利用中频信号频率计算手势目标与发射雷达传感器之间距离R：

$$R=\frac{c\tau}{2}=\frac{cf_{IF}}{2S} \tag{4}$$

式中c为光速。

设目标物体速度为v，l为接收天线之间的距离，波长为λ，根据2个Chirp信号产生中频信号相位差$\Delta\phi$，计算目标速度和角度：

$$v=\frac{\lambda\Delta\phi}{4\pi T_c} \tag{5}$$

$$\theta=\arcsin\left(\frac{\lambda\Delta\phi}{2\pi l}\right) \tag{6}$$

1.2 雷达回波数据解析和标准化

采用德州仪器公司的 77 GHz 毫米波雷达 IWR1843 采集手势信息，该采集系统发射 FWCM，天线模式为 3 发 4 收。雷达回波数据预处理首先解析手势回波数据，将雷达原始手势回波数据(bin 文件)读入，按照接收天线通道个数重组为采样点数×脉冲数×帧数的三维数据矩阵。

动态手势识别具有时间维度，一个手势动作能采集到多帧数据，实验采集时容易受雷达设备噪声干扰和外界环境影响，本文采用脉冲对消方式消除场景中静态目标，通过 2D 傅里叶变换获取距离和时间信息。首先在快时间维度采用傅里叶变换得到距离时间谱：

$$S_{\mathrm{r}}(f,T)=\int_{T}^{T+\Delta T}S_{\mathrm{IF}}(\tau,T)\exp(-\mathrm{j}2\pi f\tau)\mathrm{d}\tau \tag{7}$$

然后在慢时间维度执行傅里叶变换获得多普勒图像：

$$S_{\mathrm{r,d}}(f,f_{\mathrm{d}})=\int_{T}^{T+\Delta T}S_{\mathrm{r}}(f,\eta)\exp(-\mathrm{j}2\pi f_{\mathrm{d}}\eta)\mathrm{d}\eta \tag{8}$$

式中 η 为慢时间的积分变量。

由于不同实验人员的同一手势动作存在差别，因此毫米波雷达采集的手势回波数据可能在某些时刻存在差异和突变，使某个特征方差过大导致模型难以收敛。为避免上述问题，本文对采集的回波数据进行标准化：

$$\overline{D}_{ij}=\frac{D_{ij}-\mu}{\sigma} \tag{9}$$

式中：D_{ij} 为原始数据矩阵的元素；$\overline{D}_{ij}$ 为标准化处理后的数据矩阵元素；μ 为均值；σ 为方差。经过处理后的数据服从正态分布，之后输入神经网络。

1.3 构建手势特征图像

经过回波数据解析和标准化等数据预处理后，使用快速傅里叶变换(Fast Fourier Transform，FFT)在时间轴上对信号进行处理，可获取到手势动作的距离信息(Range Time Map，RTM)。根据 RTM 图像中表示手势距离的信息，提取每帧多普勒–时间谱(Doppler–Time Map，DTM)图中的多普勒向量，再按帧相干累积得到手势动作的 DTM。挥动手势经过数字信号处理后手势距离、速度与时间关系如图 4 所示。从图中可以看出，经过二次滤波和标准化数据处理后，能够有效消除静态噪声。整体动态手势识别过程如图 5 所示。

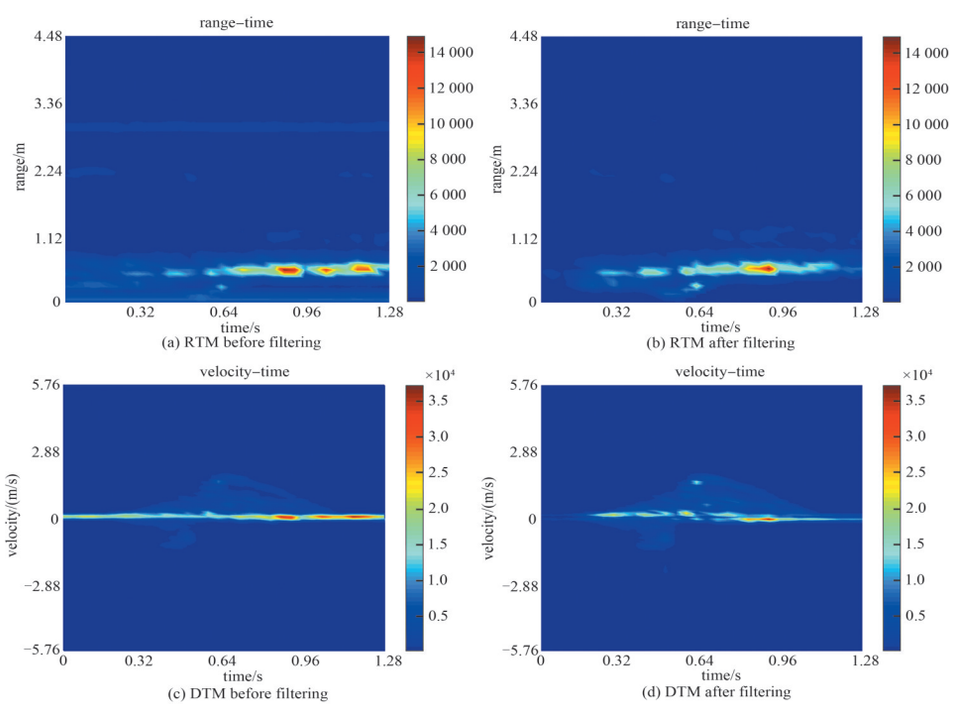


Fig.4 The gesture of waving generates range–time and Doppler–time plots before and after filtering

图 4　挥手动作在滤波前和滤波后生成的距离–时间和速度–时间图

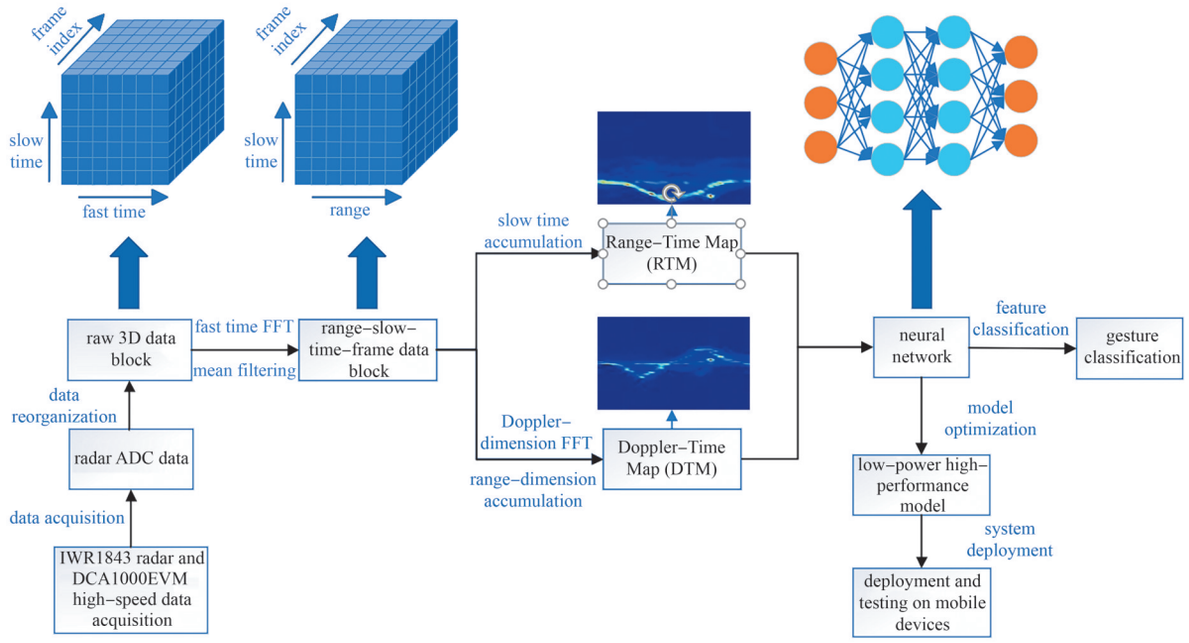


Fig.5 Dynamic gesture recognition process
图 5 动态手势识别过程

2 改进的 MobileViT 模型

MobileViT 是一种通用和移动友好的轻量级 Transformer 模型，主要包含 MobileNet V2 Block 和 MobileVit Block 模块，兼有卷积神经网络(Convolutional Neural Network，CNN)和 ViT(Vision Transformer)模型二者优点，可提取局部和全局特征。为更好获取信号图像的特征，同时考虑到毫米波雷达信号图像的复杂性明显低于自然图像(如 ImageNet 数据集中的图像)，本文对 MobileViT 模型进行改进，减少模型参数的数量，降低模型的空间复杂度和时间复杂度，以更适合本文的手势信号图像样本。MobileViT 卷积神经网络由 1 个 3×3 的卷积模块和 4 个连续的 MV2(MobileNetV2)模块组成，其中 3×3 的卷积模块和第 2 个 MV2 模块后面连接一个 2×2 的下采样层，原始网络框架图如图 6 所示。改进后的 MobileViT 模型包括 1 个 MV2 模块和 1 个 MobileViT 模块，如表 1 所示，本文的动态手势识别网络框架如图 7 所示。

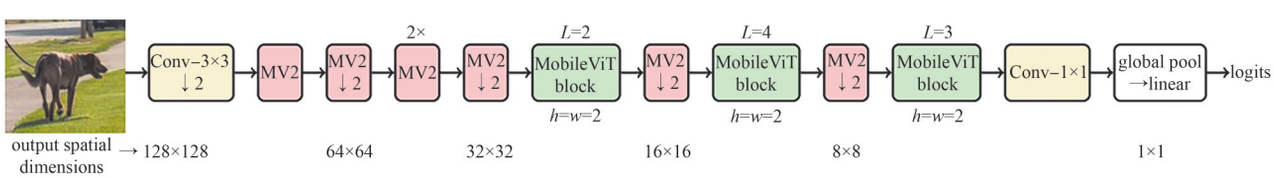


Fig.6 Original network framework diagram of MobileViT
图 6 MobileViT 原始网络框架图

表 1 构建的 MobileViT 结构

Table1 Constructed MobileViT architecture

layer	size	stride	repeat	channels
image	256×256	1	–	–
Conv–3×3, ↓ 2	128×128	2	1	16
MV2	–	–	1	32
MV2, ↓ 2	64×64	4	1	64
MV2	–	–	2	64
MV2, ↓ 2	32×32	8	1	96
MobileViT block (L=2) h=w=2	–	–	1	96(d=144)
MV2, ↓ 2	16×16	16	1	128
Conv–1×1	–	–	1	640
global pool	1×1	256	1	1 000

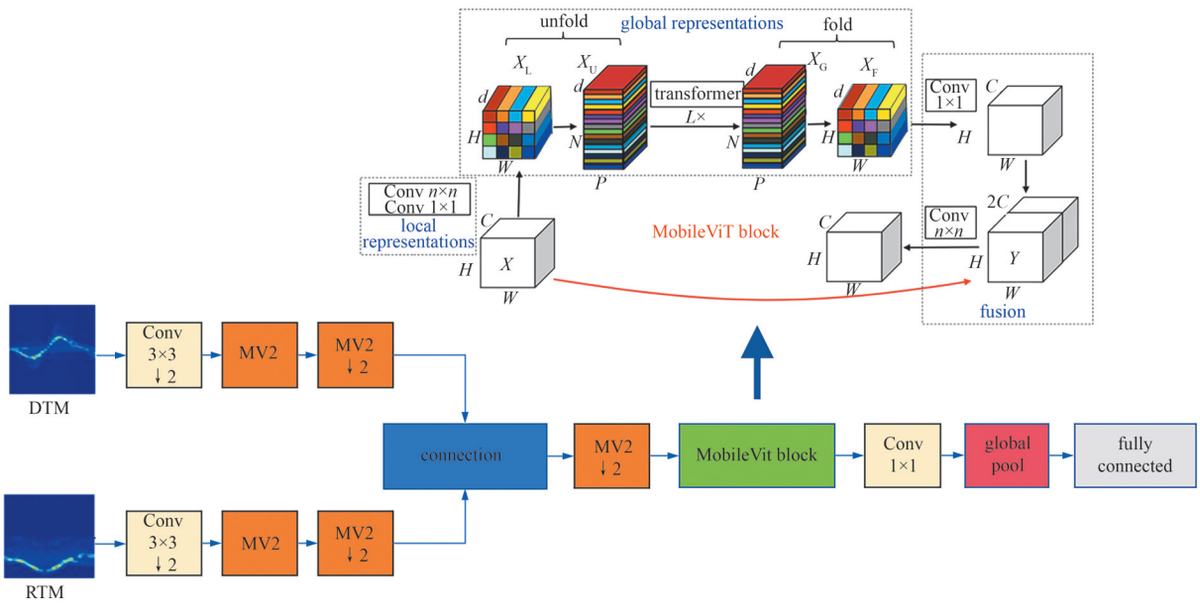


Fig.7 Network framework diagram of the proposed dynamic gesture recognition
图7 本文动态手势识别网络框架图

本文动态手势识别网络进行动态手势特征提取和融合的方式为：将同一个手势提取的 RTM 和 DTM 作为一组输入，分别输入到卷积神经网络，得出对应手势动作的 2 个特征图，该特征图的维度表示为 $C\times H\times W$，将同一手势对应的 2 个特征图在“通道维度”上进行拼接，从而得到 2 种图像融合的特征图。最后将融合后的特征图输入到构建的神经网络模型(改进后 MobileViT 模型)中，输出手势动作识别结果。

当步长(Stride)为 1 时，将输入(input)与模块的输出进行连接；当步长为 2 时，不进行连接。使用 Hard-Swish 函数代替 ReLU6 作为 MV2 模块的激活函数，如图 8 所示。ReLU6 函数和 Hard-Swish 函数如式(10)~(11)所示：

$$\mathrm{ReLU6}(x)=\min\left(\max(x,0),6\right) \tag{10}$$

$$\text{Hard-Swish}(x)=x\times\frac{\mathrm{ReLU6}(x+3)}{6} \tag{11}$$

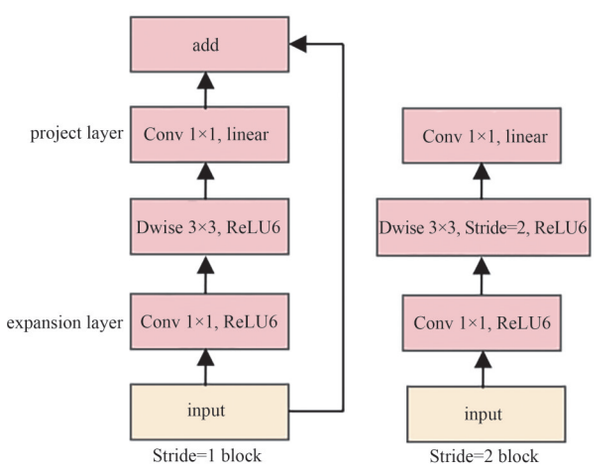


Fig.8 The structure diagram of the MV2 module
图8 MV2模块的结构示意图

在激活函数输出值小于 6 时，2 个函数的曲线非常相似；而输出值大于 6 时，通常轻量级平台所使用的 Float16 精确度有限，在描述较大范围值时有可能导致结果的精确度下降。因此本文采用 Hard-Swish 函数在嵌入式、移动式等平台上部署轻量级模型更具优势。

MobileViT 模块首先将特征图输入到局部描述层 Local Representations 中进行卷积 $n\times n$ 操作，然后再进行 1×1 的操作以调整通道数。之后将特征信息输入到全局描述层，进行展开(unfold)操作，即将特征图均匀分成若干个块(Patch)，再进行 Transformer 操作，将结果重新进行折叠(Fold)；结果使用 1×1 的卷积改变通道数，得到和 MobileViT 模块输入相同维度的特征图；将 2 个同样特征图进行连接，使用卷积模块进行处理，最后得到的特征图维度仍为 $C\times H\times W$，其中 C 表示通道，H 和 W 分别表示高和宽。

动态 MobileViT 模块将图像特征图划分为一系列不重叠的块，利用 Transformer 模型核心组件的多头自注意力学习机制，给每个特征序列分配不同权重，获取这些块之间的序列关联关系，使模型聚焦在本质特征，压缩干扰特征，更好地捕获全局图像信息。多头自注意力学习机制为：

$$\begin{cases}\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{Soft max}\left(\dfrac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}\\ \text{MultiHead}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{Contact}\left(\text{head}_1,\text{head}_2,\cdots,\text{head}_h\right)\boldsymbol{W}^{\text{o}}\\ \text{head}_j=\text{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_j^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_j^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_j^{\boldsymbol{V}}\right),j=1,2,\cdots,h\end{cases}\tag{12}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别为查询向量、键向量和值向量；$\boldsymbol{W}_j^{\boldsymbol{Q}}$、$\boldsymbol{W}_j^{\boldsymbol{K}}$、$\boldsymbol{W}_j^{\boldsymbol{V}}$分别为$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$对应的第$j$个自注意力模块的权重矩阵；$\text{head}_j$为第$j$个自注意力模块；$\boldsymbol{W}^{\text{o}}$为输出矩阵；$d_k$为注意力得分的缩放因子。

为衡量所提出模型的识别性能与部署效率，本文采用3类常用评价指标：准确率(Accuracy，式中用A表示)、空间复杂度(参数量)和时间复杂度。其中，准确率用于评估整体识别效果：

$$A=\frac{\alpha_{\text{TP}}+\alpha_{\text{TN}}}{\alpha_{\text{TP}}+\alpha_{\text{TN}}+\alpha_{\text{FP}}+\alpha_{\text{FN}}}\tag{13}$$

式中：α_{TP}、α_{TN}、α_{FP}、α_{FN}分别表示真正例、真反例、假正例和假反例数量。

空间复杂度指模型中需训练和存储的参数总量，时间复杂度表示模型在前向推理过程中的浮点计算量(Floating-point Operations Per second，FLOPs)。上述指标可有效刻画模型的性能与轻量化水平，适用于嵌入式部署场景的综合评估。

3 实验验证

3.1 实验设备

当前毫米波雷达常用工作频段包括24 GHz、60 GHz和77 GHz。24 GHz频段穿透力强但分辨力较低，适用于粗略检测；60 GHz分辨力较高，但在复杂场景中稳定性稍差。相比之下，77 GHz频段具备更高的距离和速度分辨力，适合近距离高精确度动态检测。本文选用77 GHz频段的IWR1843雷达模块和DCA1000高速数据采集卡构成手势雷达回波数据采集系统，可更准确地提取手势动作微小差异，有效提升识别精确度。雷达参数配置如下：频率77~81 GHz，带宽3 999.78 GHz，3个发射天线，4个接收天线，每帧128个Chirp，每个Chirp采样64点，帧周期40 ms，一共32帧。使用电脑对回波数据进行处理，电脑配置如下：Intel i5 10400F处理器、8 G内存、NVIDIA GTX1070Ti显卡(8 G显存)。手势识别使用深度学习框架Pytorch。

3.2 构建手势数据集

利用雷达传感器和数据采集软件mmstudio软件采集手势信息，自建雷达回波数据集，实验采集环境如图9所示。实验中一共包含12种手势，在数据集中命名为Ges1~Ges12，分别表示V字型、拍掌、水平方向摆手、上下摆手、顺时针转动、逆时针转动、向前推手、向后拉手、向左挥手、向右挥手、举手、收起举手。这12种手势在设计上充分考虑其典型性、区分性及实际应用需求，涵盖多种常见的人机交互动作，便于模型提取有效特征并满足实用性场景需求。实验由10名受试者参与，涵盖不同年龄、性别、身高和体重，以增加数据集的多样性。实验中，雷达采集时间固定为1.28 s，手势动作在距雷达0.4~0.6 m处完成，单个手势持续时间约为1 s。数据采集过程中未对距离和速度设定严格限制，不同受试者可根据自身动作习惯完成操作，这样可在一定程度上增强模型的泛化能力。每个手势采集75次，总计获得9 000条数据。将收集的手势回波数据进行处理，可得到12个手势对应的DTM和RTM图，如图10所示。数据按照文件夹分类存储，每个手势类别对应一个独立文件夹，文件夹内包含该手势的所有采集数据，便于数据管理和后续分析。

Fig.9 Diagram of experimental data collection environment

图9 实验采集环境图

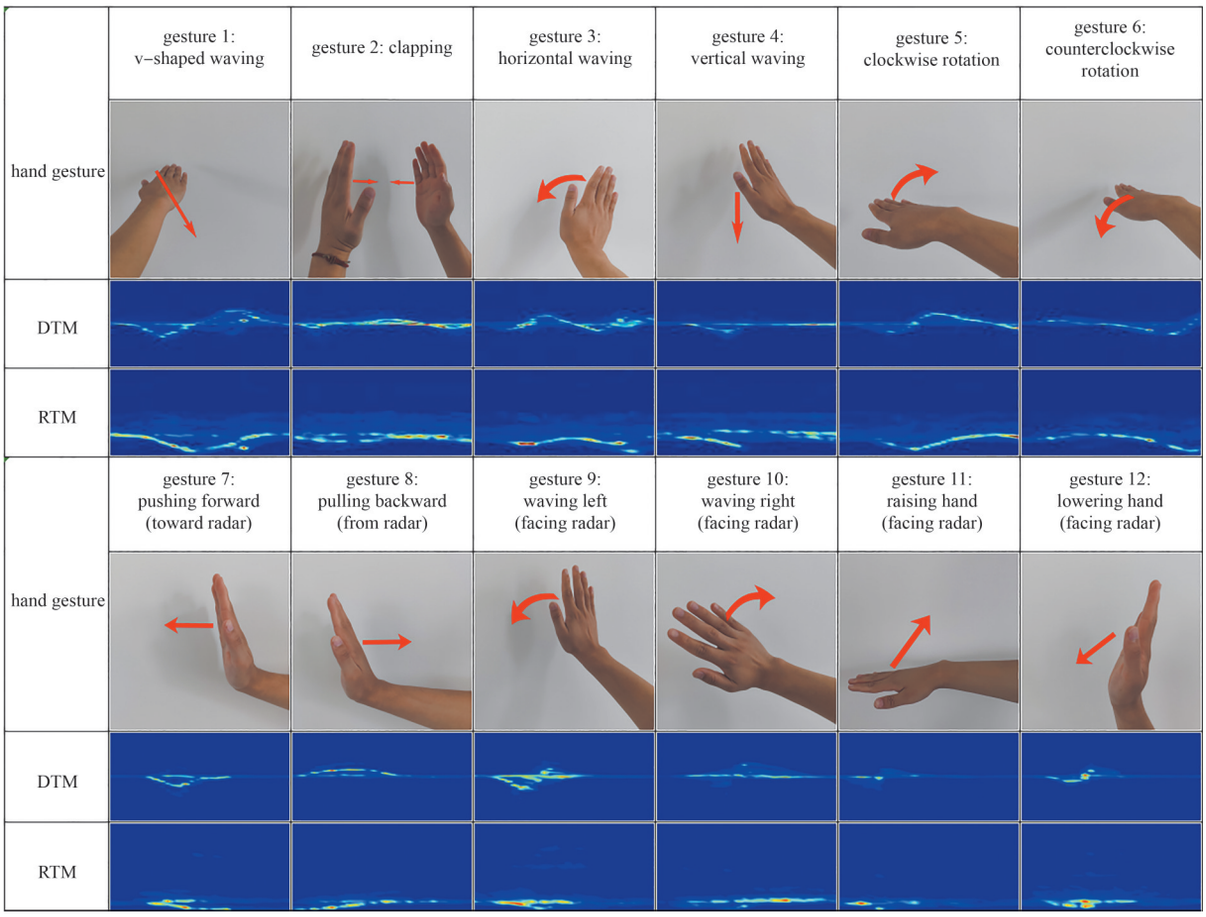


Fig.10 DTM and RTM corresponding to gestures in this paper
图 10　本文手势对应的 DTM、RTM

3.3　模型训练和验证

实验中将数据集以 4:1 的比例随机划分训练集、验证集，采用五折交叉验证的方法，即对数据集划分 5 次，将 5 次实验的平均值作为模型识别的结果。数据集划分的情况如图 11 所示，每次划分的训练集手势数据为 7 200 条，验证集手势数据为 1 800 条。为便于神经网络对图像的处理，首先将 DTM 和 RTM 进行预处理，包括尺寸缩放、归一化。将生成的 DTM 和 RTM 图像尺寸大小设置为 64×64；归一化将图像的像素值映射在 0 和 1 之间。

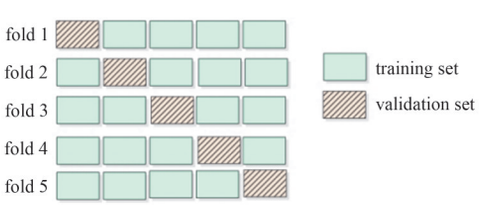


Fig.11 Display chart of the five–fold cross–validation in the article
图 11　本文五折交叉验证的显示图

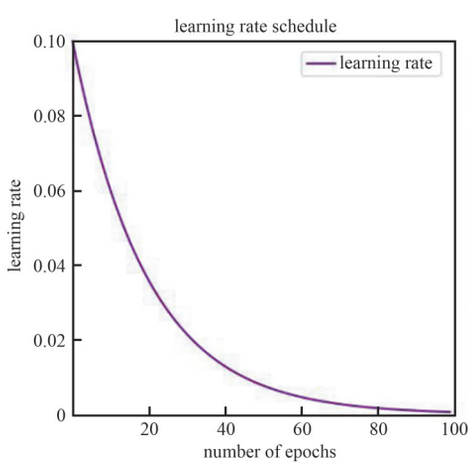
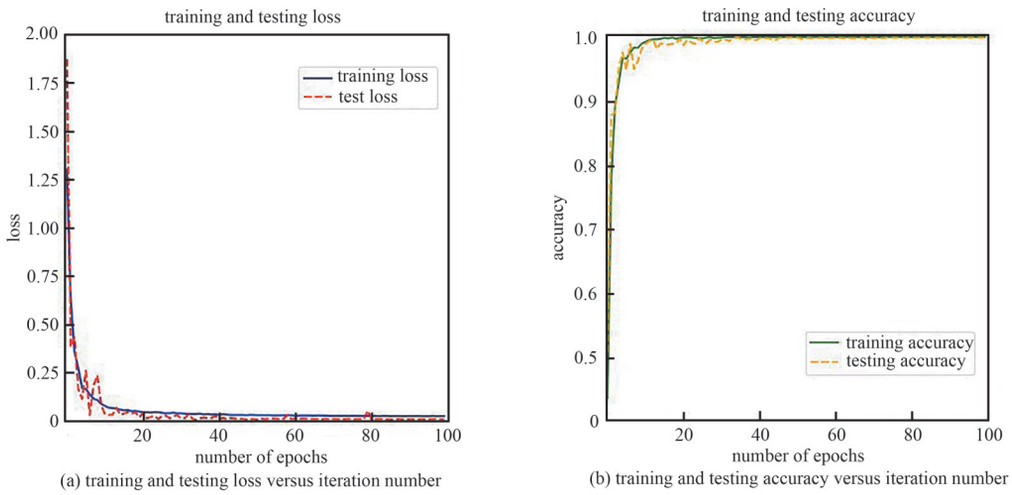


Fig.12 Learning rate change curve
图 12　学习率变化曲线图

在模型训练时，使用 Adam 作为优化方法，初始学习率设置为 0.1，同时使用 StepLR 学习率调度器，步长设置为 1，衰减率设置为 0.95，这有助于在训练初期快速逃离局部最小值，并在训练后期细化解的精确度。图 12 为模型训练时学习率的变化曲线，图 13 为训练时损失的变化曲线。从图中可以看出，学习率与训练损失变化率均随着训练周期下降并收敛，其中训练损失变化率经过 5 个周期后快速下降并收敛。



(a) training and testing loss versus iteration number (b) training and testing accuracy versus iteration number

Fig.13 Training loss change curve

图 13 训练时损失变化曲线图

表 2 不同模型的手势种类、数据集和识别准确率对比

Table2 Comparison of gesture types, datasets, and recognition accuracy for different models

model	types of gestures	dataset	average accuracy/%
1D-ScNN[13]	5	4 000	96.01
3DCNN-Transformer[14]	6	6 000	97.14
TS-I3D[15]	12	4 000	96.17
Atten-TsNN[16]	5	3 750	98.15
2D CNN-Transformer[17]	6	3 000	99.18
RD-Net[18]	11	5 225	98.28
CNN[19]	10	5 000	96.61
DCNN[20]	14	3 500	95.00
transformer Meta-Learning Network[21]	8	1 200	97.18
2D CNN-Transformer[17]	12	9 000	99.10
the proposed method	6[17]	1 800	99.20
the proposed method	12	9 000	99.31

为验证本方法的有效性，与现有的多种毫米波雷达手势识别方法进行对比分析，结果如表 2 所示。从表 2 可以看出，文献[20]的手势种类最多，有 14 种，但数据集样本仅有 3 500 条。大部分模型的手势类型在 10 个以下，本文构建的数据集有 12 个手势类型，数据集样本数有 9 000 个。与其他模型的 3 000~6 000 条手势数据相比，更多的数据样本条数意味着本模型验证结果的可靠性。为进一步增强对比的全面性与说服力，本文在表 2 中补充了两项对比实验：a) 第 10 行：2D CNN-Transformer[17]模型在本文数据集上的实验结果。该模型在相同数据条件下的平均识别率为 99.10%，略低于本文方法的 99.31%，反映出本文模型在特征提取与识别精确度上的提升效果。b) 第 11 行：本文方法在文献[17]所采用的字母手势数据集上的实验结果。在该数据集上，本文模型识别准确率达到了 99.20%，优于原文的 99.18%，进一步验证了所提方法在不同数据场景下的适应性和鲁棒性。

为验证本文方法在空间复杂度、时间复杂度和识别准确率方面的优势，将本文模型与其他模型进行对比，结果如表 3 所示。从表中可以看出，其他模型中 2D-CNN 平均识别率最高，为 99.18%，与本文的识别准确率仅差 0.13 个百分点，但 2D-CNN 模型识别的手势种类为 6 种类型，仅为本文手势种类的 50%，且 2D-CNN 模型的空间复杂度和时间复杂度分别为本文的 13.2 倍和 36 倍。与文献[13]中的 1D-ScNN 手势识别方法相比，本文用模型中的位置编码代替长短期记忆(LSTM)网络层，有效关联手势回波相邻帧之间信息，在空间复杂度仅差 0.01M 的同时，时间复杂度降低了 50%，手势识别准确率提高了 3.1 个百分点。相比 CNN-LSTM 网络，本文方法在识别准确率提升 0.8 个百分点的同时，计算复杂度和参数规模分别降低至其 1/7.79 和 1/3.11。文献[16]提出的 Atten-TsNN 方法在 5 类手势上实现 98.15% 的准确率，但其空间和计算复杂度分别为本文的 15.09 倍和 38.74 倍。与经典轻量级网络相比，MobileNetV3 和 ShuffleNetV2 的模型复杂度显著更高，准确率仍低于本文方法 5 个百分点以上。综上，所提模型在保证精确度的同时大幅降低了计算开销，具备良好的部署优势。

为验证所提方法的实际应用性能，本文在相同实验平台上对模型的推理效率和资源占用进行了评估。结果显示，改进的 MobileViT 模型在单个样本上的平均推理时间为 7.20 ms，优于 2D-CNN-Transformer 的 10.35 ms，展现出更高的推理效率。此外，本文模型参数量仅为 0.167 M，计算复杂度为 0.253 GFLOPs，对计算资源需求较

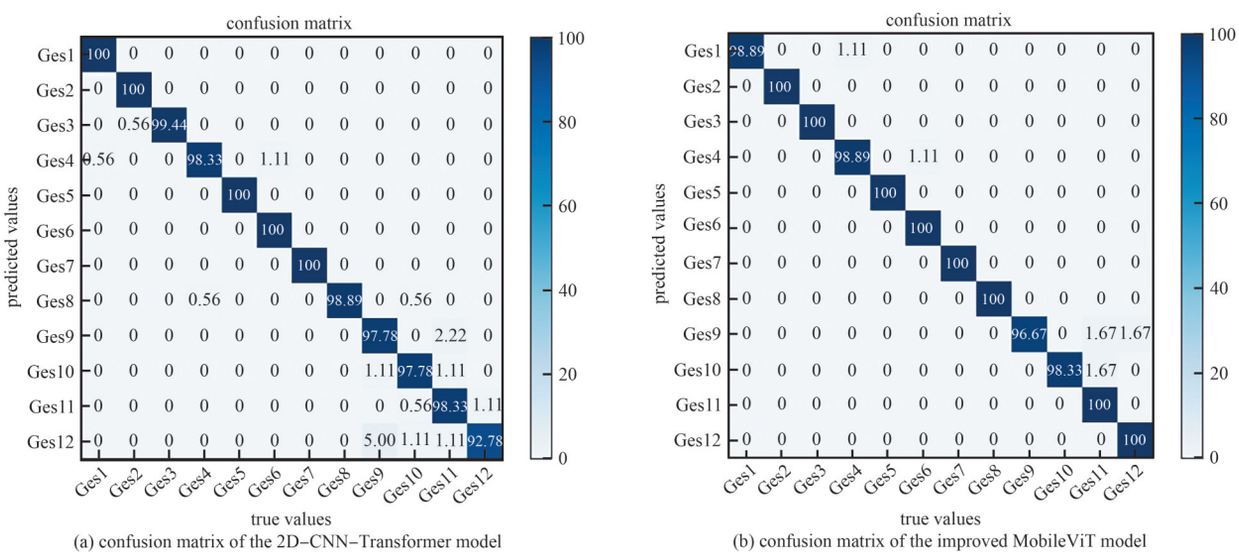
低，适合在边缘计算或嵌入式设备中部署。该性能优势主要得益于结构轻量化设计，包括引入 MV2 模块压缩参数规模、融合 RTM 与 DTM 图像以降低输入维度，以及使用 Hard-Swish 激活函数提升非线性计算效率，从而在保证识别精确度的同时实现了资源节约与运行加速。

表3 本文模型空间和时间复杂度与其他模型相比较

Table3 Spatial and temporal complexity of the model in this paper compared with other models

model	space complexity/10^6	time complexity (10^9 FLOPs)	average accuracy/%
1D-ScNN [13]	0.156	0.503	96.01
TS-I3D [15]	0.390	289.070	94.44
Atten-TsNN [16]	2.520	9.800	98.15
CNN-LSTM [17]	0.520	1.970	98.51
2D CNN-Transformer [17]	2.210	9.310	99.18
transformer Meta-Learning Network[21]	2.450	10.710	97.18
SRDS- Transformer [22]	0.170	0.217	99.17
MC-CNN[23]	0.550	0.372	95.83
CNN-LSTM[24]	0.520	7.840	98.61
MobileNetV3-small[25]	1.520	0.122	94.17
ShuffleNetV2-x0.5[26]	0.350	0.697	96.67
MobileViT	0.955	0.271	97.58
the method in this article (improved MobileViT)	0.167	0.253	99.31

为进一步验证所构建 MobileViT 模型的识别性能，本文选取 2D-CNN-Transformer[17]作为对比模型，并在相同的未经训练的 1 800 个样本测试集上进行评估，其混淆矩阵如图 14 所示。图 14(a)为 2D-CNN-Transformer 模型的混淆矩阵，结果表明，2D-CNN-Transformer 在部分类别如 Ges9、Ges10 和 Ges12 等易混淆手势上存在一定的误判。图 14(b)为改进 MobileViT 模型的混淆矩阵，改进 MobileViT 在多个关键类别上误判率更低，特别是在 Ges9、Ges10、Ges12 等相似动作的区分上表现更佳。此外，改进的 MobileViT 对 Ges2、Ges3、Ges5、Ges6、Ges7、Ges8、Ges11 和 Ges12 等手势的识别准确率较高，而手势 1(V 字形挥动)被误判为手势 4(水平方向摆手)的概率为 1.11%，手势 9 和手势 10 的误判概率相对较低，被误判为手势 11 的概率为 1.67%。



(a) confusion matrix of the 2D-CNN-Transformer model　(b) confusion matrix of the improved MobileViT model

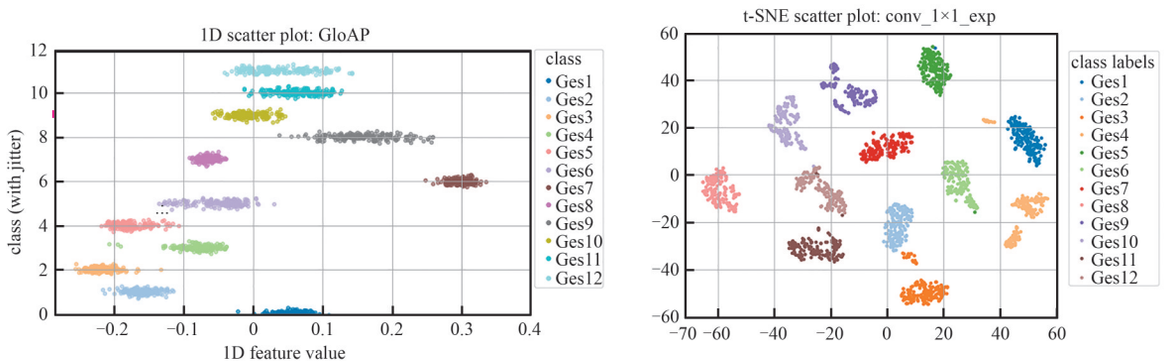
Fig.14 Confusion matrix for gestures

图14 手势混淆矩阵

因此，由表 2~3、图 14 可以看出，本文采用的改进 MobileViT 轻量级神经网络模型简单，减少了计算需求和参数数量，在较少参数下实现了与 CNN 和 ViT 神经网络模型相当的性能，在保持较高手势识别准确率的同时，空间复杂度和时间复杂度低，非常适合部署在内存资源有限的移动边缘设备和嵌入式设备。

为进一步验证改进后的 MobileViT 模型在特征提取能力方面的优势，与 2D-CNN-Transformer[17]模型对比了在识别 12 类手势时所提取的中间层特征表现，分别绘制其特征散点图，如图 15 所示。图 15(a)为 2D-CNN-Transformer 模型在进入分类层前的特征可视化，由于其输出特征为 1 维，为便于观察，本文对 Y 轴进行随机微扰(jitter)处理，使各类别分布更加可辨。在图中可观察到，不同类别在 X 轴(即唯一特征维度)上存在一定重叠，部分类别(如 Ges2 与 Ges5、Ges11 与 Ges12)之间区分度不高。图 15(b)为本文改进的 MobileViT 模型的 t-SNE 特征可视化结果，可以观察到各类别手势在二维空间中呈现出良好的聚类结构，分布紧凑且相互分离，说明该模型具

备更强的特征表达与区分类别能力。



(a) feature scatter plot of the 2D-CNN-Transformer model

(b) feature scatter plot of the improved MobileViT model

Fig.15 Comparison of feature scatter plots between models

图 15 模型散点图对比

4 结论

本文提出了一种基于改进 MobileViT 网络的毫米波雷达动态手势识别方法，旨在解决现有方法中手势类别受限、计算复杂度高等问题。通过对雷达原始回波数据进行滤波，有效消除了静态干扰，利用相干积累提高了信噪比，并通过多头自注意力机制有效挖掘了全局特征，提升了识别精确度。相较于 CNN-LSTM、Atten-TsNN 以及经典轻量级网络(如 MobileNetV3 和 ShuffleNetV2)，本文方法在减少计算复杂度和参数量方面展现出显著优势。实验结果表明，所提出方法在 12 类手势的数据集上达到了 99.31% 的识别准确率，且模型的计算效率更高，验证了该方法在实际应用中的可行性和优越性。

参考文献：

[1] 罗朗娟,王勇,何维 . 基于毫米波感知的手势识别技术研究进展[J]. 移动通信, 2022,46(6):82-85. (LUO Langjuan,WANG Yong, HE Wei. Research progress on gesture recognition technology based on millimeter-wave perception[J]. Mobile Communications, 2022,46(6):82-85.) DOI:10.3969/j.issn.1006-1010.2022.06.013.

[2] 董尧尧,曲卫,邱磊 . 毫米波雷达手势识别综述[J]. 兵器装备工程学报, 2021,42(8):119-125. (DONG Yaoyao,QU Wei,QIU Lei. Review of millimeter-wave radar gesture recognition[J]. Journal of Ordnance Equipment Engineering, 2021, 42(8): 119-125.) DOI:10.11809/bqzbgcxb2021.08.019.

[3] MIN Rui,WANG Xing,ZOU Jie,et al. Early gesture recognition with reliable accuracy based on high-resolution IoT radar sensors[J]. IEEE Internet of Things Journal, 2021,8(20):15396-15406. DOI:10.1109/JIOT.2021.3072169.

[4] LI Yadong, ZHANG Dongheng, CHEN Jinbo, et al. DI-gesture: domain-independent and real-time gesture recognition with millimeter-wave signals[C]// 2022-2022 IEEE Global Communications Conference. Rio de Janeiro, Brazil: IEEE, 2022: 5007-5012. DOI:10.1109/GLOBECOM48099.2022.10001175.

[5] YAN Baiju, WANG Peng, DU Lidong, et al. mmGesture: semi-supervised gesture recognition system using mmWave radar[J]. Expert Systems with Applications, 2023(213)Part B:119042. DOI:10.1016/j.eswa.2022.119042.

[6] LI Yadong, ZHANG Dongheng, CHEN Jinbo, et al. Towards domain-independent and real-time gesture recognition using mmWave signal[J]. IEEE Transactions on Mobile Computing, 2023,22(12):7355-7369. DOI:10.1109/TMC.2022.3207570.

[7] ZHANG Jia,XI Rui,HE Yuan,et al. A survey of mmWave-based human sensing:technology,platforms and applications[J]. IEEE Communications Surveys & Tutorials, 2023,25(4):2052-2087. DOI:10.1109/COMST.2023.3298300.

[8] KEHELELLA K,LEELARATHNE G,MARASINGHE D,et al. Vision transformer with convolutional encoder-decoder for hand gesture recognition using 24 GHz Doppler radar[J]. IEEE Sensors Letters, 2022,6(10):1-4.

[9] RITCHIE M, JONES A M. Micro-Doppler gesture recognition using Doppler, time and range based features[C]// 2019 IEEE Radar Conference(RadarConf). Boston,MA,USA:IEEE, 2019:1-6. DOI:10.1109/RADAR.2019.8835782.

[10] 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):142. DOI:10.1145/2897824.2925953.

[11] NASR I,JUNGMAIER R,BAHETI A,et al. A highly integrated 60 GHz 6-channel transceiver with antenna in package for smart sensing and short-range communications[J]. IEEE Journal of Solid-State Circuits, 2016,51(9):2066-2076. DOI:10.1109/JSSC.2016.2585621.

[12] 李楚杨. 基于毫米波雷达的手势识别算法研究[D]. 成都:电子科技大学, 2020. (LI Chuyang. Research on gesture recognition algorithm based on millimeter-wave radar[D]. Chengdu,China:University of Electronic Science and Technology of China, 2020.)

[13] 靳标,彭宇,邝晓飞,等. 基于串联式一维神经网络的毫米波雷达动态手势识别方法[J]. 电子与信息学报, 2021,43(9):2743-2750. (JIN Biao, PENG Yu, KUANG Xiaofei, et al. Dynamic gesture recognition method based on millimeter-wave radar by one-dimensional series neural network[J]. Journal of Electronics & Information Technology, 2021, 43(9): 2743-2750.) DOI: 10.11999/JEIT200894.

[14] 董连飞,马志雄,朱西产. 基于车载毫米波雷达动态手势识别网络[J]. 北京理工大学学报, 2023,43(5):493-498. (DONG Lianfei, MA Zhixiong, ZHU Xichan. Dynamic gesture recognition network based on vehicular millimeter wave radar[J]. Transactions of Beijing Institute of Technology, 2023,43(5):493-498.) DOI:10.15918/j.tbit1001-0645.2022.102.

[15] WANG Yong,WANG Shaha,ZHOU Mu,et al. TS-I3D based hand gesture recognition method with radar sensor[J]. IEEE Access, 2019(7):22902-22913. DOI:10.1109/ACCESS.2019.2897060.

[16] JIN B,PENG Y,KUANG X,et al. Robust dynamic hand gesture recognition based on millimeter wave radar using Atten-TsNN[J]. IEEE Sensors Journal, 2022,22(11):10861-10869. DOI:10.1109/JSEN.2022.3170311.

[17] JIN Biao,MA Xiao,ZHANG Zhenkai,et al. Interference-robust millimeter-wave radar-based dynamic hand gesture recognition using 2-D CNN-transformer networks[J]. IEEE Internet of Things Journal, 2024,11(2):2741-2752. DOI:10.1109/JIOT.2023.3293092.

[18] 陈君毅,蒋德琛,王智铭,等. 一种基于双维度滤波和自适应定长化的FMCW雷达手势识别算法研究[J]. 电子学报, 2023, 51(8): 2179-2187. (CHEN Junyi, JIANG Dechen, WANG Zhiming, et al. Research on an FMCW radar gesture recognition algorithm based on dual-dimension filtering and adaptive length normalization[J]. Acta Electronica Sinica, 2023, 51(8): 2179-2187.) DOI:10.12263/DZXB.20211410.

[19] DU Chuan,ZHANG Lei,SUN Xiping,et al. Enhanced multi-channel feature synthesis for hand gesture recognition based on CNN with a channel and spatial attention mechanism[J]. IEEE Access, 2020(8):144610-144620. DOI:10.1109/ACCESS.2020. 3010063.

[20] SKARIA S,AL-HOURANI A,LECH M,et al. Hand-gesture recognition using two-antenna Doppler radar with deep convolutional neural networks[J]. IEEE Sensors Journal, 2019,19(8):3041-3048. DOI:10.1109/JSEN.2019.2892073.

[21] 屈乐乐,洪雨云. 基于Transformer元学习网络的毫米波雷达手势识别方法[J/OL]. 雷达科学与技术, 2025:1-11. (2025-04-16) [2025-05-24]. http://kns. cnki. net/kcms/detail/34.1264. tn. 20250415.1110.012. html. (QU Lele, HONG Yuyun. Gesture recognition method based on transformer meta-learning network using millimeter-wave radar[J]. Radar Science and Technology, 2025:1-11. (2025-04-16)[2025-05-24]. http://kns.cnki.net/kcms/detail/34.1264.tn.20250415.1110.012.html.)

[22] JIN Biao, WU Hao, ZHANG Zhenkai, et al. Effective dynamic gesture recognition with sparse representation and dual-stream transformers in mmWave radar[J]. IEEE Transactions on Industrial Informatics, 2025, 21(1): 604-612. DOI: 10.1109/TII. 2024. 3455419.

[23] XIA Zhaoyang,LUO-MEI Yixiang,ZHOU Chenglong,et al. Multidimensional feature representation and learning for robust hand-gesture recognition on commercial millimeter-wave radar[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021,59(6): 4749-4764. DOI:10.1109/TGRS.2020.3010880.

[24] ZHANG Guiyuan, LAN Shengchang, ZHANG Kang, et al. Temporal-range-Doppler features interpretation and recognition of hand gestures using mmW FMCW radar sensors[C]// 2020 the 14th European Conference on Antennas and Propagation (EuCAP). Copenhagen,Denmark:IEEE, 2020:1-4. DOI:10.23919/EuCAP48036.2020.9135694.

[25] HOWARD A,SANDLER M,CHEN Bo,et al. Searching for MobileNetV3[C]// 2019 IEEE International Conference on Computer Vision(ICCV). Seoul,Korea(South):IEEE, 2019:1314-1324. DOI:10.1109/ICCV.2019.00140.

[26] MA Ningning,ZHANG Xiangyu,ZHENG Haitao,et al. ShuffleNet V2:practical guidelines for efficient CNN architecture design[C]// The 15th European Conference on Computer Vision. Munich,Germany:Springer International Publishing, 2018:122-138. DOI: 10.1007/978-3-030-01264-9_8.

作者简介：

葛志洲(2001-)，男，在读硕士研究生，主要研究方向为雷达目标识别.email:16634866230@163.com.

张向群(1978-)，女，博士，教授，硕士生导师，主要研究方向为雷达信号处理、毫米波雷达感知.

申佳文(2004-)，男，在读本科生，主要研究方向为物联网系统集成与应用、无线通信技术.

杜根远(1974-)，男，博士，教授，硕士生导师，主要研究方向为毫米波雷达感知、深度学习.

刘锋涛(1979-)，男，学士，高级工程师，主要研究方向为嵌入式系统集成.