e-Inu：模拟具有情感感知能力的四足机器人

[1]Abhiruph Chakravarty，[2]Jatin Karthik Tripathy，[1]Sibi Chakkaravarthy S，[3]*Aswani KumarCherukuri [4]S. Anitha [5]Firuz Kamalov [6]Annapurna Jonnalagadda [1]计算机科学与工程学院和人工智能与机器人卓越中心（AIR）[2]认知系统，波茨坦大学，德国[4]电子工程学院，VIT-AP大学，印度安得拉邦[3]信息技术学院，韦洛尔理工学院，印度韦洛尔[5]电气工程系，迪拜加拿大大学，阿联酋[6]计算机科学与工程学院，韦洛尔理工学院，印度韦洛尔。（*通讯作者：cherukuri@acm.org）

摘要

四足机器人目前应用于工业机器人技术中，作为机械辅助来自动化执行多项日常任务。然而，目前这类机器人在家庭环境中的应用仍主要处于研究阶段。本文探讨了此类机器人对人类情绪进行检测和理解、生成自身步态，并通过声音和屏幕表情作出回应的相关研究与虚拟仿真。为此，我们结合强化学习与软件工程概念，对一种四足机器人进行仿真，使其能够理解情绪、在各种地形中移动、检测声源，并通过视听反馈回应情绪。本文旨在建立一个具有情感智能的四足机器人仿真框架，使其能够主要通过运动或音频响应回应视听刺激。语音情绪检测的表现不如ERANNs或Zeta策略学习，但准确率仍达到63.5%。视频情绪检测系统的结果几乎达到最先进水平，准确率为99.66%。由于采用“on-policy”（同策略）学习过程，PPO算法能够极快地完成学习，使仿真狗在不同步频和变化条件下展现出非常流畅的步态。这使四足机器人能够对生成的刺激作出回应，因此我们得出结论：该机器人按预期运行，并达到了本研究的目标。

关键词：自动步态生成、深度学习、情绪识别、强化学习、机器人仿真。

1. 引言

机器人技术的发展一直是人类进步的重要基石。然而，与大多数新兴技术不同，机器人目前仍主要局限于工业或专业用途，家庭自动化、智能家居清洁器等应用只是极少数例外。考虑到这一点，人们对宠物，尤其是狗的需求日益增长，既是出于情感需求，也是出于

安全需求。在现代快节奏的生活中，照料有机宠物可能相当困难。为此，我们致力于利用 PyBulet 模拟一款人工智能四足机器人 [37] ，使其能够满足一只具有感知能力（尽管是人工的）的小狗的家庭需求。该机器人的开发涉及三个基本问题，即情绪分析、环境感知和自动步态生成。目标是

1. 设计一个系统，自动检测场景中人类语音、语调和面部表情所体现的内在情绪；2. 检测声源和障碍物的方向；3. 生成必要的步态，朝声源移动；4. 针对当前情境生成所需的视听响应。

本研究探讨卷积神经网络（CNN）、循环神经网络（RNN）、强化学习（RL）以及一些软件工程领域的现有理念或模型，以实现最终目标，例如运用到达时间差（TDoA）。本文的主要目标是关注养宠物狗的情感层面，而非其看护功能。需要指出的是，本文并不开发市场上已有先进机器的机器人技术。本文探讨了另一种可能的应用方式：在家庭环境中使用这些四足机器人，打造更为乌托邦式的《神经漫游者》[1] 未来版本。为此，我们致力于打造一种具备一定情感智能的四足机器人系统，使机器人能够表现得更加自然。

2. 相关前期工作的综述

这项研究的主要灵感来自一些用途更偏向工业领域的类似四足机器人，例如 ANYbotics 开发的 ANYmal [2] 、波士顿动力开发的 Spot [3] ，以及宇树科技开发的 AlienGo [4] 。这些四足机器人主要以坚固耐用为设计目标，能够应对工业场景中可能出现的各种情况。不过，e-Inu 也实现了所有四足机器人都共有的若干关键功能：步态生成与优化、避障以及路径规划。Marc Raibert、Blankespoor、Nelson 等人[5], 也开展了类似的四足机器人项目 Big Dog，其目标是打造外形酷似狗的自主四足机器人。

我们近期探讨的论文之一是 Deng等人[6] 《基于稀疏自编码器的语音情绪识别特征迁移学习》。根据他们的研究结果，用于开发语音情绪识别系统的训练数据和测试数据通常彼此高度匹配，但也可能获取到其他“相似”数据。迁移学习能够利用这类相似数据进行训练，尽管它们存在内在差异，从而提升识别器的性能。他们的研究提出了一种用于语音情绪识别的特征迁移学习稀疏自编码器技术，该技术从目标域中有限的标注数据里学习通用的情绪特定映射规则。随后，他们将此方法应用于来自不同域的情绪特定数据，获得了新的重构数据。在六个典型数据库上的实验结果表明，他们的方法远优于分别对每个源域进行学习。基于稀疏自编码器的特征迁移学习方法的基本思路是，使用单层自编码器在少量目标数据中寻找共同结构，再将该结构用于重构源数据，从而将有用的知识从源数据迁移到目标任务中。他们利用重构数据为 Interspeech 2009 情感挑战赛提出的一个现实问题构建了语音情绪识别引擎，

该现实问题由 Interspeech 2009 情感挑战赛提出。根据在六个公开语料库上的实验结果，所提出的技术能够有效迁移知识并提高分类准确率。

在 Lu 等人的研究中，[7], 作为一项下游任务，他们提出利用端到端自动语音识别模型的预训练特征来进行语音情感分析。结合语音中的声学和文本信息的端到端自动语音识别特征取得了令人鼓舞的结果。他们使用带有自注意力的 RNN 作为情感分类器，并利用注意力权重提供直观的可视化效果，以帮助理解模型预测。他们采用 IEMO-CAP 数据集和一个新的大规模语音情感数据集 SWBD-sentiment 进行评估。在超过 49,500 条话语上，他们将 IEMO-CAP 上最先进的准确率从 66.6% 提高到 71.7%，并在 SWBD-sentiment 上达到 70.10% 的准确率。由此证明，端到端自动语音识别模型的预训练特征有助于情感分析。然而，Lu 等人的研究[7] 是我们模型通过音频/语音分析情绪的根本动机。他们展示了一种口语情感识别系统，该系统基于由快速学习算法训练的循环神经网络（RNN）模型。该系统考虑了长期上下文影响以及情感标签表达的不可预测性。他们采用基于双向长短期记忆网络（双向长短期记忆网络）的稳健学习方法，根据时间动态提取情绪状态的高层表示。

为避免情感标签的歧义，研究假设每一帧的标签可视为一系列随机变量，因此同一段语音中的所有帧都映射到相同的情感标签。随后使用所提出的学习算法训练这些序列。与作为基线的基于深度神经网络-极限学习机的情感识别系统相比，所提出的情绪识别系统将加权准确率提高了至多 12%。他们的方法展示了如何利用循环神经网络和基于最大似然的学习技术来提升情绪识别效果。我们也正是在此时了解到梅尔频率倒谱系数及其作为音频特征提取器的应用。随着深入研究，我们发现了 Mermelstein [8], Davis 和 Mermelstein [9], 以及 Bridle 和 Brown [10] 的相关著作。

Maghilnan 和 Kumar [11] 讨论了一种情感分析算法，该算法利用从语音流中提取的特征，辨别对话中说话人的情绪。他们的流程包括使用语音活动检测进行预处理、实现语音识别系统、结合梅尔频率倒谱系数实现说话人识别系统，最后实现情感分析系统。他们的研究提出了一种通用模型：输入包含两人讨论内容的音频，通过自动将音频转换为文本并执行说话人识别，分析谈话内容和说话人身份。该系统在人工生成的数据集上表现良好；他们当时正致力于收集更大的数据集并提升系统的可扩展性。尽管该系统能够准确理解对话交流中说话人的情感倾向，但它只能处理两位说话人轮流发言的对话。

接下来介绍情感识别系统的另一部分——面部情绪检测。许多研究都探讨了这一具体问题。最早从图像中检测情绪的研究之一由 Gajarla 和 Gupta [12]开展。这项研究采用的方法依赖于

大型预训练目标检测模型 VGG16 [13]，借助它，人脸检测模型得以快速训练，因为使用 VGG16 意味着模型已经能够从图像中提取线索。这种迁移学习方法通过将 VGG16 模型的最后一层替换为不同的分类层，使 Gajarla 和 Gupta [12] 的方案取得了良好效果。使用 VGG16 提取图像特征后，我们将其输入支持向量机，使整个模型能够针对情绪检测任务进行训练。这项研究还探索了不同的架构，例如采用一对其余支持向量机方法，并使用不同的 VGG16 权重，以及使用 ResNet50。Gajarla 和 Gupta [12] 使用 ResNet50 模型，在一个由从“Flickr”抓取的图像组成的数据集上取得了令人瞩目的结果，准确率达到 73%。这项研究面临的主要问题之一是，由于图像是从 Flickr 抓取的，图像标注相当主观，而且图像的环境氛围和光照很容易影响模型的判断。

Dachapally [14] 采用表征自编码器单元（RAU），探索了另一种面部情绪检测方法。自编码器能够为不同情绪构建独特的表征，从而便于区分图像。根据他们的研究结果，使用自编码器可以形成不同情绪的通用编码特征，因为模型在训练过程中会观察不同的人脸。这种方法也取得了不错的效果，因为 Dachapally [14] 使用了 JAFFE 数据集，其中包含 215 张图像，展示了 10 位不同的女性模特摆出 7 种情绪表情。训练集中的所有图像均为日本女性，因此样本都来自相同的族裔和性别。这一点使自编码器的表现相当不错，相较之下，Gajarla 和 Gupta [12]讨论的迁移学习方法则不如它。

Bargal 等人[15] 的研究采用了与上述面部情绪识别方法截然不同的方案。该方案将图像依次输入三个不同的预训练图像模型，即 VGG13、VGG16 和 ResNet50。在图像模型提取特征时进行并行化，可将图像分解为不同层次；与仅使用一个预训练模型相比，这能获得丰富得多的特征。随后，将这三个模型提取的特征分别输入符号平方根（SSR）和 L2 范数处理，再将其合并。该研究使用了 EmotiW’16 数据集，以及通过网络爬取带有情绪类别标签的图像所收集的额外数据。采用这种并行化图像特征提取方法后，Bargal 等人 [15] 在测试数据集上取得了 56.66% 的准确率，比基线高出 16% 以上。该研究还展示了使用单一预训练模型与在特征提取过程中进行并行化之间的差异：并行模型的准确率比仅使用 VGG16 模型高出约 2%。

目前已有多种方法用于优化四足机器人步态的平滑性和速度。早期实现这一目标的方法很大程度上依赖人类观察，以规划出能实现最佳腿部运动的腿部轨迹。然而，近年来，这些方法开始采用深度强化学习，使模型能够在较少人类干预的情况下生成。

Hengst 等人 [16] 最早采用的步态分析方法之一，依赖人类观察来为腿的运动制定固定路径。该固定路径根据步态的三个方面进行配置：运动、速度和腿部支撑。这些参数使[16] 能够形成矩形的四个角，

如图1所示，该矩形描绘了步态轨迹，并能与机器人的身体对齐。机器人的腿沿着该轨迹运动，所经过的一系列航路点根据运动参数生成。这些航路点被分为两组，每组数量相等：一组用于地面划动阶段，另一组用于矩形的其余部分。

不久之后，Kim 和 Uther [17] 改进了矩形轨迹法，创建了一种新的四边形步行轨迹。该轨迹由相对于原始轨迹的四个偏移量描述，见图2；机器人足部沿新轨迹运动时，速度始终保持一致。由于轨迹点数量增加，这种新步态本身更加平滑；轨迹优化则采用 Press 等人 [18]所述的 Powell（方向集）多维最小化方法。

先前的方法虽然能够取得令人满意的结果，但要充分理解并配置路径，以实现腿部及机器人整体的平稳运动，需要耗费大量时间和人力。Clune、Beckmann、Ofria 等人 [19] 提出了一种用于进化神经网络的新生成式编码方法——HyperNEAT。与先前的方法相比，HyperNEAT 尤其令人瞩目，因为它完全不需要人工调参，神经网络最终便能找到一种可行的步态方案。生成式编码还有一个优势：它能够更轻松地重复利用表型模块，从而改善腿部协调。

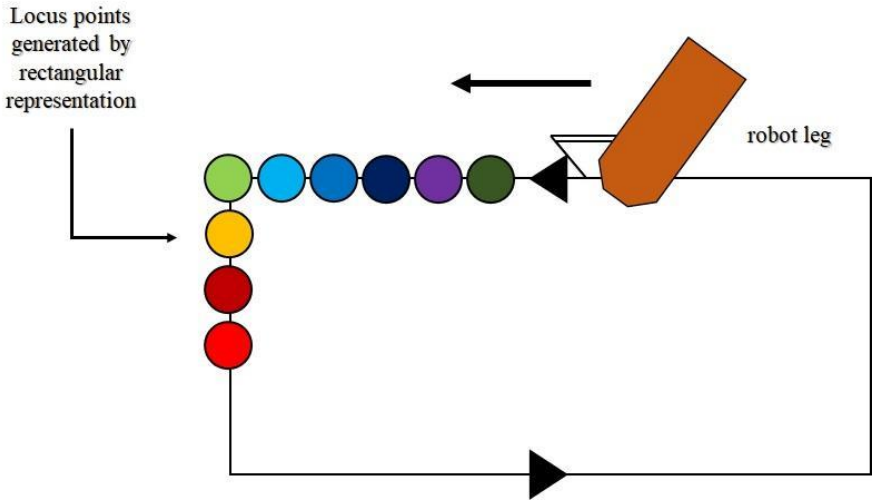


图1. 矩形轨迹（改编自 [16]）

Owaki 和 Ishiguro [20] 采用了与以往方法截然不同的理念。他们的方法基于这样一个事实：四足机器人的肢体运动遵循一种重复模式，可以使用“中枢模式发生器”（CPG）模型进行模拟。过去针对去大脑猫进行的实验表明，猫的脊髓中也存在生物性中枢模式发生器 [21], [22]。这种方法还能够实现自发的步态转换，即从行走转为小跑，再转为慢跑，最后转为奔跑；只需改变一个与速度相关的参数，就能实现这一点，同时保持平衡并降低运动成本。

Lodi、Shilnikov 和 Storace [23] 也研究了四足机器人遵循中央模式发生器（CPG）进行肢间通信和同步这一现象。他们提出了一种设计和分析

中央模式发生器（CPG）的方法，该方法基于多参数分岔理论，并使用一种近期提出的软件工具 CEPAGE [24]。此方法应用于两个中央模式发生器（CPG）：一个受仿生启发，另一个则完全为合成系统。在这两种情况下，对中央模式发生器（CPG）的分析都能通过调整分岔参数，轻松获得不同的步态序列。尽管所提出的方法需要一定程度的人为干预，因为它主要是一种基于规则的设计方法，但中央模式发生器（CPG）会产生节律模式，从而大幅简化步态分析，并使自发步态转换成为可能。

本文讨论了一种适用于家庭环境的机器人的开发。该机器人能够从视觉和听觉刺激中理解情绪，并通过各种动作和声音作出回应，同时配备液晶显示屏，使人类更容易接受它。

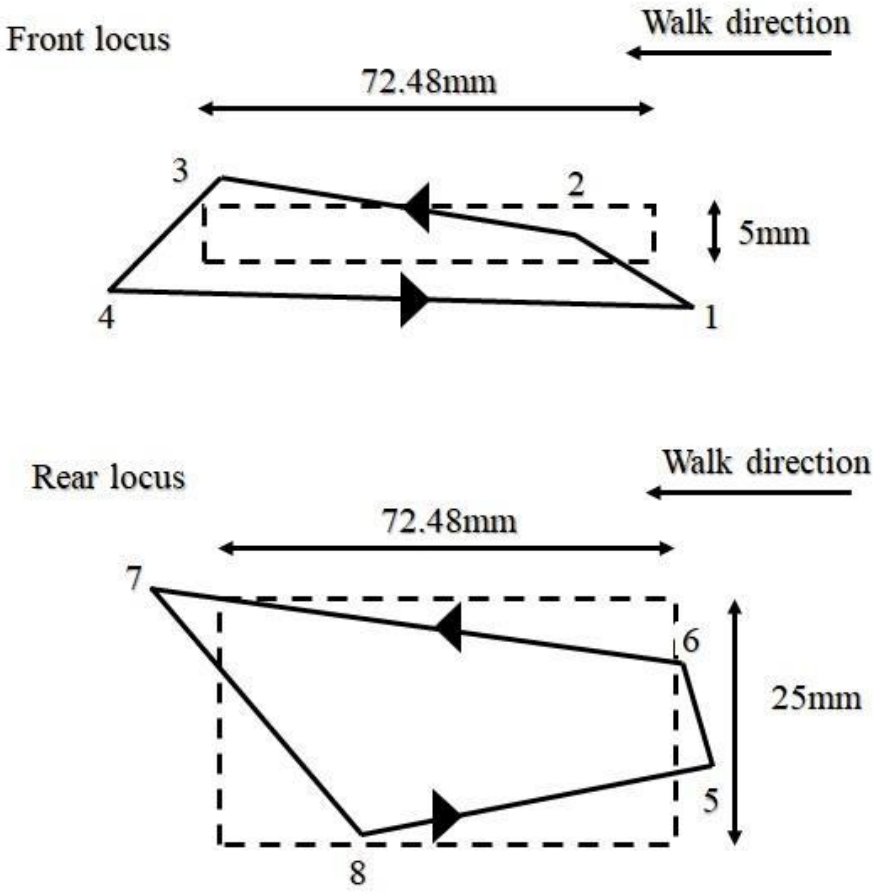


图2. 优化矩形轨迹（改编自 [17]）

3. 提出的架构：e-Inu

e-Inu 架构包含四个独立模块，以实现预期目标。情绪检测模块负责检测音频和视觉方面的所有情绪信息。到达时间差模块使机器狗能够感知自身的位置。步态生成模块负责生成步态，并将机器人的移动映射为从一个位置到另一个位置的运动。

最后，视听反馈模块帮助机器人与人类和环境进行交流。接下来的部分将介绍这些组件。

3.1 情绪检测

情绪检测模块将音频和视觉输入都视为外部刺激，并分别使用深度学习网络对其进行计算。对于音频特征的情绪分析，我们使用从生成的音频中提取的梅尔频率倒谱系数，并将其输入由四个全连接层夹着两层长短期记忆网络层的结构中，其中两侧各有两个全连接层。至于视频特征的情绪分析，我们使用 Haar 级联分类器检测面部结构，使用 OpenCV，以及在 Places365 上训练的不含顶层的 VGG16，其输入经过展平后传入全连接层。

梅尔频率倒谱系数在学术界被戏称为“最常考虑的系数”，在音频数据处理方面通常可谓一项全能工具。人类发出的任何声音都由其声道的形状，以及舌头、牙齿等部位共同决定。如果能够精确识别这种形状，就可以恰当地描述人类产生的每一种声音。语音信号时域功率谱的包络描绘了声道，而梅尔频率倒谱系数所表示的正是这一点。梅尔频率倒谱系数的前十三个系数，即其较低维度，被视为代表上述谱包络的特征。更高维度则表达了有关频谱特征的更多细节。对于各种音素而言，包络足以表达它们之间的差异，因此我们可以使用梅尔频率倒谱系数识别音素。不过在我们的案例中，鉴于更多数据通常意味着更好的数据，我们通过实验得出结论：梅尔频率倒谱系数的四十维特征效果最佳。

MFCC 是一种时间序列，因为数据按顺序表示，其中 y 轴表示频率，x 轴表示时间。众所周知，LSTM 非常擅长从上述时间序列中进行统计推断。Transformer 也很适合这项任务，但会耗费大量计算资源，而且完全没有必要。只有在处理极长序列时，Transformer 才具有较高的计算经济性。但在我们的例子中，每个时间步的序列长度只有 40。

与大多数计算机视觉相关任务一样，卷积网络（卷积神经网络）最具潜力。一般而言，卷积神经网络在计算机视觉任务中的表现会随着架构加深或加宽而提升。这一趋势同样适用于面部情绪识别任务。VGG-16 是一种具有 16 层深度的 CNN，曾在 2014年赢得 ImageNet大规模视觉识别挑战赛（ImageNet）。我们可以加载该网络的预训练实例，它使用来自 ImageNet数据库 的逾百万张照片进行训练。这个预训练网络能够将照片分为数千个不同类别。因此，该网络学到了适用于多种图片的丰富特征表示。网络的图片输入尺寸为 224 × 224。

然而，实验发现，与 ImageNet 版本相比，采用 Places-365 版本的 VGG16 在面部情绪识别方面表现更好。这主要是因为 Places365 数据库（Places2 数据库的最新版本）在训练卷积神经网络进行场景识别方面效果更佳。也就是说，由于我们从 VGG-16-Places-365 的高层提取深层场景特征，这些特征提升了通用特征的识别效果，进而有助于面部特征识别和情绪识别。

3.1.1 数据集

为此，我们结合了四个用于音频文件情绪识别的数据集，以及一个用于视频情绪识别的数据集（实际处理中将视频作为静态照片帧处理）。这些数据集均涵盖相同的七种基本情绪：愤怒、中性/轻蔑、厌恶、恐惧、悲伤、快乐和惊讶。音频数据集包括 RAVDESS [25], CREMA-D [26], SAVEE [27] 和 TESS [28]。视频（静态照片）数据集则使用 CK+ [29]。

3.1.2 模型

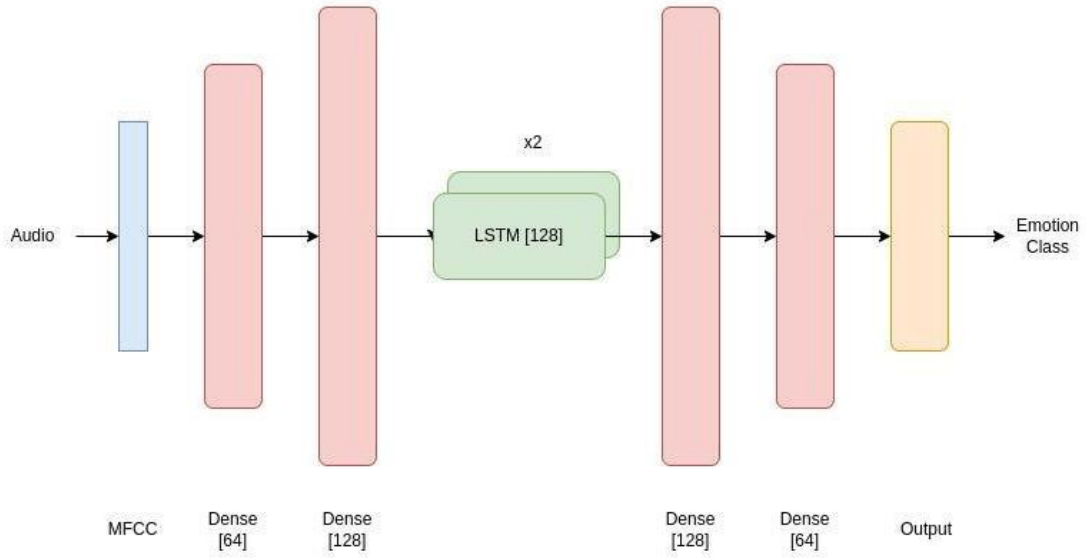


图 3. 基于音频特征进行情绪检测的模型流水线

首先对音频输入进行处理，并提取具有 40 个维度的梅尔频率倒谱系数（MFCCs）。这些系数作为输入变量存储在 NumPy 数据阵列中。标签则以类似方式存储为输出变量。该模型由八层组成。第一层是一个标准的输入层，其维度为 [1,40]。下一层是包含 64 个节点的全连接层，之后是另一个包含 128 个节点的全连接层。为了扩大特征空间，从而实现更高效的反向传播和更好的学习效果，我们倾向于使用全连接层，将输入层连接到模型的下一个、也可能是最重要的部分——长短期记忆网络。至于为什么使用大小不同的层，如果我们为了扩大特征空间而直接将规模从 40 增加到 128，我们发现，在输入节点与长短期记忆网络之间使用两个包含 128 个节点的全连接层，会导致模型出现明显的不稳定，并使准确率指标下降。因此，我们采取一种更加审慎、循序渐进的方法。

如前所述，128 个节点的全连接层所产生的输入会传入两个长短期记忆网络层，每层各有 128 个节点，然后再依次传入另外两个全连接层，分别含有 128 个和 64 个节点。这种将两个长短期记忆网络层上下堆叠的做法称为分层长短期记忆网络模型或堆叠长短期记忆网络模型。这样一来，第一个长短期记忆网络的隐藏状态便能传递到第二个长短期记忆网络。堆叠长短期记忆网络层会增加模型的深度，更准确地说，这使其成为一种深度学习

方法。神经网络的深度通常与其在各种复杂预测任务中的表现相关。在本例中，尽管这会增加模型的计算开销，但也能提升我们所需的准确率。至于在堆叠的长短期记忆网络之后使用全连接层，是因为长短期记忆网络的输出无法直接传入 Softmax 激活函数。它们通常还会输出隐藏状态 h，单元数量与该输出的维度相等。这通常不是所需输出层的维度；在本例中，所需维度为七。在长短期记忆网络之后指定一个或多个全连接层，对应于

$$y(t) = W * h(t) \quad (1)$$

其中，y(t) 是需要传递给 softmax 层的逻辑值，而 W 就是这一最后一层的权重连接矩阵。与之前一样，通过在全连接层中逐级减小节点规模，可以抑制模型的不稳定性，并逐步将节点规模降至输出层所需的基数。

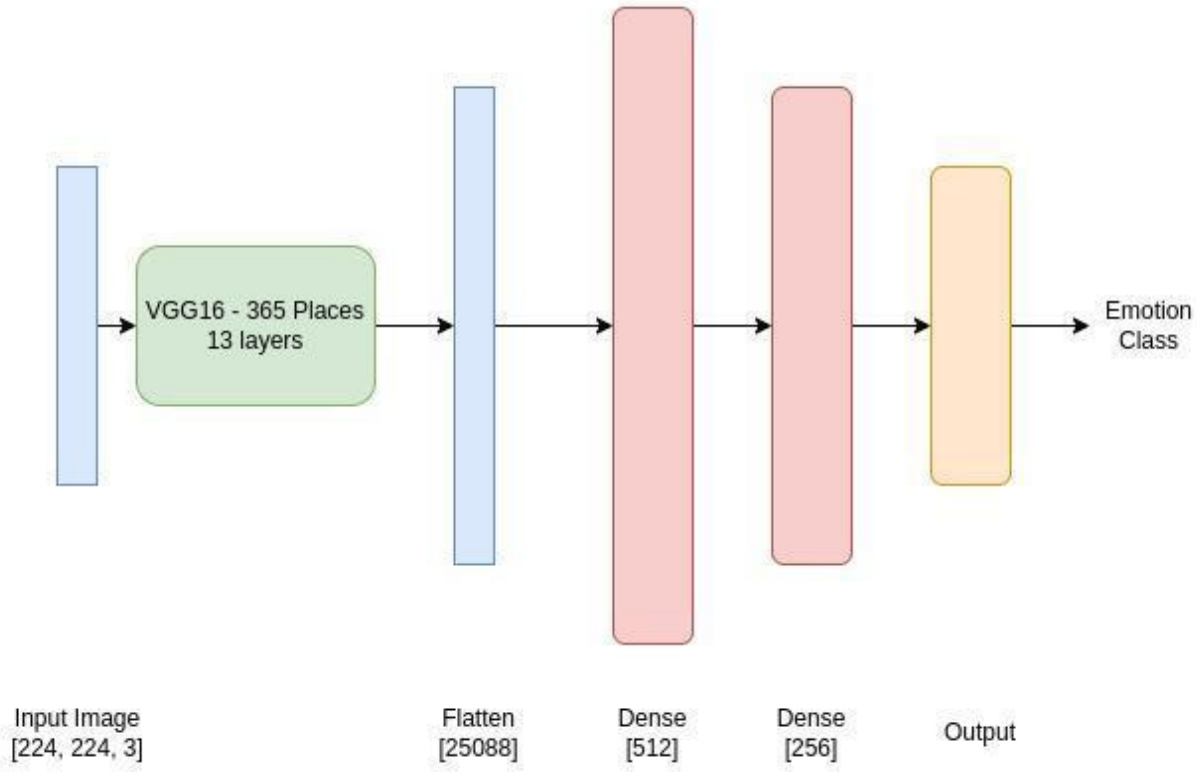


图4。基于视频特征进行情绪检测的模型流水线

最后一层是一个包含 7 个节点的全连接层，构成我们的输出层。对于输出层，使用 softmax；对于其余全连接层，则使用 ReLU 作为激活函数。标准的随机失活比例为 20%。上方的图3以图示方式展示了这一点。

至于从视频中进行情绪检测，我们采用了 VGG16 架构。在我们的工作中，我们使用了不含顶层的 VGG16-places365 模型（使用该模型的前 13 层）。使用 Haar 级联分类器进行面部特征检测后，将图像处理为 224x224 像素

维度。VGG16 各层的输出经过展平后，输入一个包含 512 个节点、采用 ReLU 激活且丢弃率为 20% 的全连接层。随后，输出再次输入一个采用 ReLU 激活、包含 256 个节点的全连接层。最后，结果传入包含 7 个节点并采用 Softmax 激活的输出层，如图4所示。最终输出层之前的全连接层与上文所述的音频情绪检测模型中的全连接层功能相似。节点层基数逐渐降低，有助于减少不稳定性。这些全连接层还有助于增强模型整体的分类能力，从而提升 VGG-16 展平后的输出效果。

为了解决两个情绪检测模型可能同时产生输出的问题，我们使用排序系统，选择优先级最高的情绪作为最终登记的输出，如表1所示。

表1：情绪的优先级排名

Rank	情绪
1	愤怒
2	厌恶
3	fear
4	悲伤
5	惊讶
6	快乐
7	中性/轻蔑

3.2 到达时间差（TDOA）检测

在这项工作中，声源定位至关重要。为了缩小真实狗与机器人构造体之间的差距，机器人必须能够直观地响应语音提示，转身面向扬声器所在的方向。相关的大多数工作都基于这样一个原理：使用麦克风阵列拾取从墙壁和物体反射回来的声波，从而实现声源定位。通过不同的麦克风采集声音，可以计算到达时间差（TDoA） [30], [31] ，以确定声源的方向和距离。TDoA 本身是一个相当简单的概念，它利用麦克风阵列中不同麦克风拾取同一声音的时间差。由于麦克风之间的距离和角度已知，因此可以计算声音发出点的距离和方向，如图5所示。

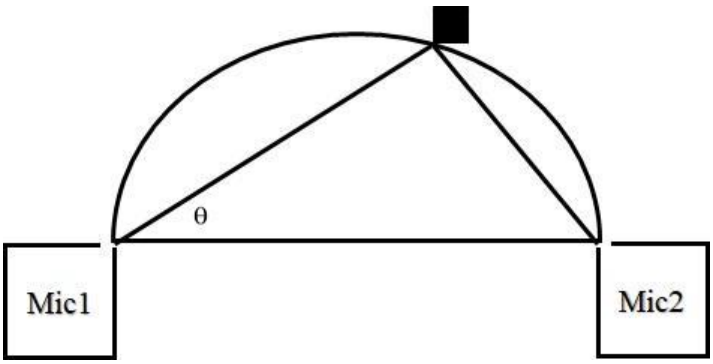


图5. 两个麦克风的到达时间差

只需两个麦克风即可计算 TDoA，但要判断声源来自阵列前方还是后方，则需要更多麦克风。为此，我们使用了四个呈十字形排列的麦克风，如图6所示，使仿真机器人能够成功识别来自任意位置的声音的距离和方向。TDoA 的实现方式是：任意一个麦克风拾取到声音信号时，计时器立即启动；与第一个麦克风正对的麦克风拾取到相同声音时，计时器停止。所得时间即为这对麦克风之间的 TDoA，随后可代入以下公式，计算声音发出方向与第一个麦克风之间的角度。

$$\theta = arccos(\frac{\Delta t * v}{d}) \quad (2),$$

其中，θ 是麦克风 1 到声源的仰角（假设麦克风 1 和麦克风 2 位于 x 轴上），Δt 是声音到达两个麦克风的时间差，v是声速，d 是它们之间的最短距离。

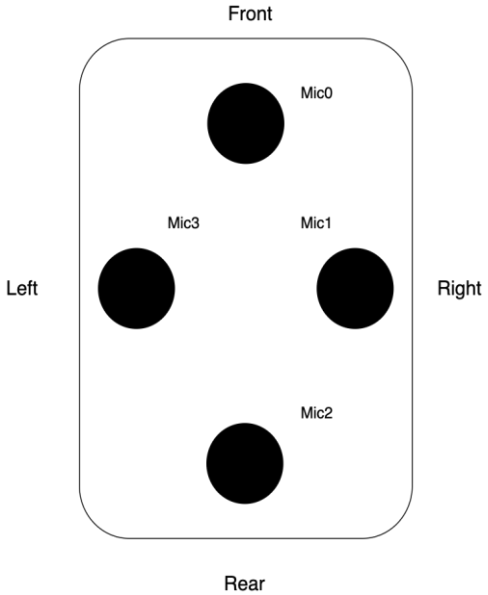


图6. 麦克风的摆放位置（俯视图）

3.3 步态生成

在这项工作中，我们采用双关节腿部结构，并通过实现近端策略优化（近端策略优化）[32] 来生成实际步态。正如 Hengst、Ibbotson、Pham 等人[16] 以及 Kim 和 Uther [17], 的研究所示，可以计算并确定仿真中机器人腿部运动路径的轨迹，但这种做法相当不切实际, 因为轨迹无法平滑地适应地形表面的变化。因此，这项工作采用 PPO算法，这是一种流行的强化学习方法，通过从训练仿真的小批量数据中学习经验来训练模型。随后，利用学到的经验更新决策策略。模型遍历整个训练集后，便丢弃已学习的小批量经验，并使用更新后的策略训练新的一批数据。这种“同策略学习”可确保不良训练批次中的信息不会过度向前传播，从而避免四足机器人形成笨拙或无法使用的步态。虽然这意味着训练中的变化会小得多，但最终训练过程会更加平稳，也能确保模型不会形成导致无意义动作的步态习惯。

在基于模型的强化学习中，智能体会学习一个环境模型，从而能够进行规划。状态转移概率分布和奖励函数共同构成该模型。无模型强化学习则无需这样的模型，设法进行推断。基于模型的强化学习使智能体能够明确权衡可用动作及其影响，从而进行规划，并提高训练过程中的样本效率。这种方法已成功应用于 AlphaZero 等程序，使其精通围棋或将棋。然而，在许多应用场景中，它似乎会过拟合，以至于模型在真实环境中完全失效。无模型强化学习的样本效率较低，但更容易实现，也不容易过拟合。

无模型强化学习进一步分为策略优化和Q学习两类。Q学习通过估计并优化最优Q函数来实现学习。虽然它的样本效率更高，但其性能稳定性取决于最优Q函数的估计效果。另一方面，策略优化直接优化智能体的性能，因此表现更加稳定可靠。策略优化会显式表示并优化策略，以最大化回报。

本文使用近端策略优化（PPO）算法，并采用混合策略，定义为 a(t) - 用户策略，π(o) – 反馈。该混合策略使模型能够轻松地从完全由用户指定，转变为完全依赖训练过程中学到的步态。例如，对于完全由用户定义的模型，将反馈分量的上限和下限设为 0，即可确保训练过程中学到的信息不会传递到下一个训练批次。本文使用 OpenAI Gym 环境，学习如何优雅地摆出姿势，同时避免动作过快。OpenAI Gym 让我们能够轻松创建并使用基于仿真的物理环境。虽然 Gym 已提供许多适用于不同用途的现成环境，但我们仍需对环境进行修改，直至其中只保留物理部分。随后，我们创建一个新的执行器来代表这只狗，由 PPO 模型控制。它使用一维动作空间，其中反馈分量为 π(o)，边界为 [-0.1 和 0.1]。反馈通过 Sigmoid函数处理，以协调运动。使用 --playground 标志时，可以通过 pyBullet界面手动设定特定姿势，调整机器人基座位置 (x,y,z) 及朝向（横滚、俯仰、偏航）。

3.4 视听反馈

反馈通过相当基础的软件工程实现。液晶显示屏和扬声器会输出与每种情绪对应的反馈。虽然这看起来相当简单，但我们认为这是让最终产品更自然所必需的一点设计。

4. 实验设置

由于本工作的主要目标是创建一个仿狗四足机器人的仿真，因此，标准的步态仿真环境必须支持通常无需使用的输入。在情绪识别方面，我们采用了相当简单的设置——使用网络摄像头进行视频情绪识别，使用麦克风进行音频情绪识别。这两种输入被传入改造后的 Gym 环境，使仿真能够在单一系统内运行。音频直通还使我们能够在仿真中实现 TDOA，并通过点选法在环境中指定声源，以进行测试。健身房环境中的地形被设为随机，以便准确测试机器人。TDOA 还使我们能够在仿真中的多地形系统里，通过自主步态生成来确定机器人的朝向。然而，我们没有采用特定指标来评估机器人在随机地形环境中的运动质量。评判标准只有一个：能否完成。至于音频和视觉反馈，我们通过输出一段录制狗叫声的 .wav 音轨来实现，声音与推断出的情绪相对应；同时使用 Tkinter 图形用户界面屏幕显示循环播放的、经过翻译的液晶显示屏控制器代码，以便在设置于机器人模拟面部位置的液晶显示屏上重现我们制作的图像，从而展现机器人的情绪（以回应推断出的情绪）。

我们使用四麦克风系统获取空间音频。除音频情绪识别模块外，输入还会并行传送至到达时间差模块，以估算声源方向。我们使用 RGB 摄像头拍摄视频静帧，并通过 Haar 级联分类器提取面部特征。随后将这些特征传送至情绪分析模块，以推断视频和音频中的情绪。如前所述，我们使用一张硬编码表，将七种基本情绪分别分配到基于优先级的排序系统中，以表示对相应情绪作出反应的重要程度。随后处理情绪推断结果，并对优先级较高的情绪作出反应（即两个情绪识别模块推断出的最可能情绪中优先级较高者）。对于不紧急的情绪，会触发蹲姿，并通过液晶显示屏和扬声器提供进一步反馈。对于仿真中愤怒或悲伤等更紧急的情绪推断结果，机器人会朝声源移动，同时通过液晶显示屏和扬声器组合提供反馈。

5. 结果与分析

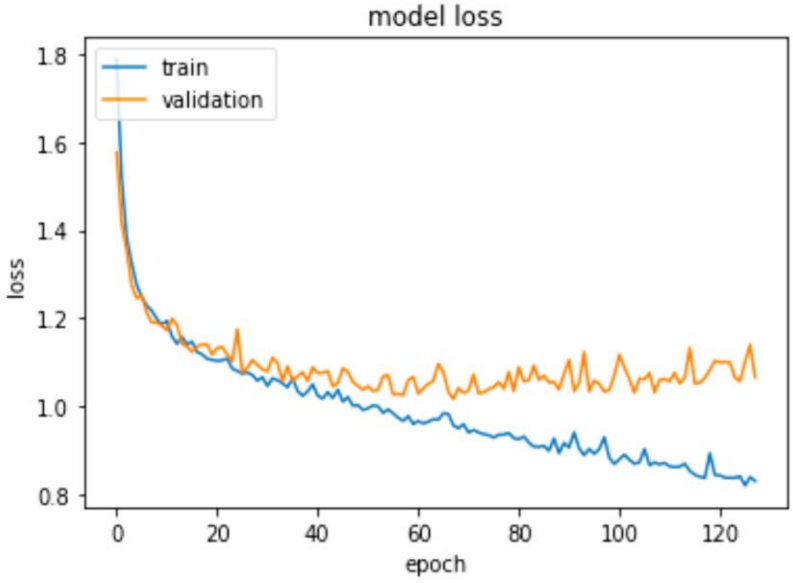
音频输入的情绪检测模块在包含 RAVDESS、CREMA-D、SAVEE 和 TESS 的汇总数据集上取得了 63.47% 的总体测试准确率，见图 7。视频输入的情绪检测模块表现出人意料地出色，在 CK+ 数据集上达到了 99.66% 的准确率，见图8，超过了此前排名第 3 的成绩（7 种情绪准确率），与排名第 2 的准确率相比仅落后约 0.04%；这一结果依据 paperswithcode.com [32], 上的基准，未考虑近似值。我们希望在此说明，尽管 FN2EN [33] 在 paperswithcode.com 上位居数据集排行榜首位，但这是因为其最高准确率

在 8 种情绪类别中达到了 96.8% 的准确率，而在 6 种情绪类别中则达到了 98.6% 的准确率。他们没有提供 7 种情绪类别的测试分数。

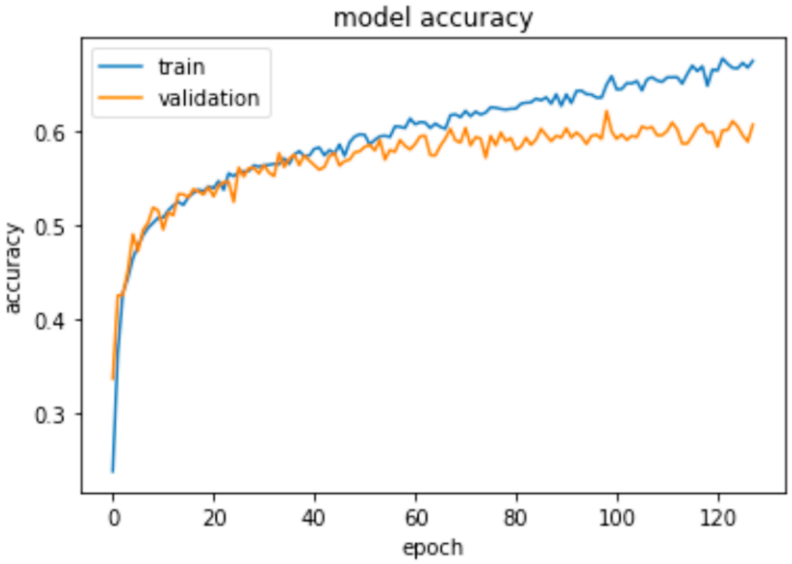
由于我们的根本目标是执行分类任务，因此，对于两种情绪识别类型（即基于面部特征和语音的情绪识别），分类交叉熵都是最适合的损失函数。

图7显示，模型在第48个训练轮次左右收敛，并在验证准确率约为60%处趋于稳定。但模型有轻微过拟合的倾向，训练准确率指标升至约68%。损失函数图也反映了相同的情况。验证损失和验证准确率开始与训练损失和训练准确率脱钩，这表明偏差—方差权衡开始失衡。第48个训练轮次之后，模型倾向于偏向训练数据集。尽管如此，从实验结果来看，图7所示的结果是我们在尝试图3中模型的多种不同变体后获得的最佳结果，未将训练轮次长度纳入考量。

图8显示，第20至第27个训练轮次之间出现了尖峰。这些异常是Adam的小批量梯度下降（批量大小为32）不可避免的副作用。有些小批量中会意外出现不利于优化的“倒霉”元组，从而引发这些异常，并影响代价函数和准确率指标。我们采用随机梯度下降（即批量大小为1）时，发现代价函数出现了更多异常。不过，使用（全）批量梯度下降时不会出现这种情况，因为它会在每个优化训练轮次中使用整个训练数据集。然而，只要结果值得付出代价，Adam优化的表现通常优于大多数方法。

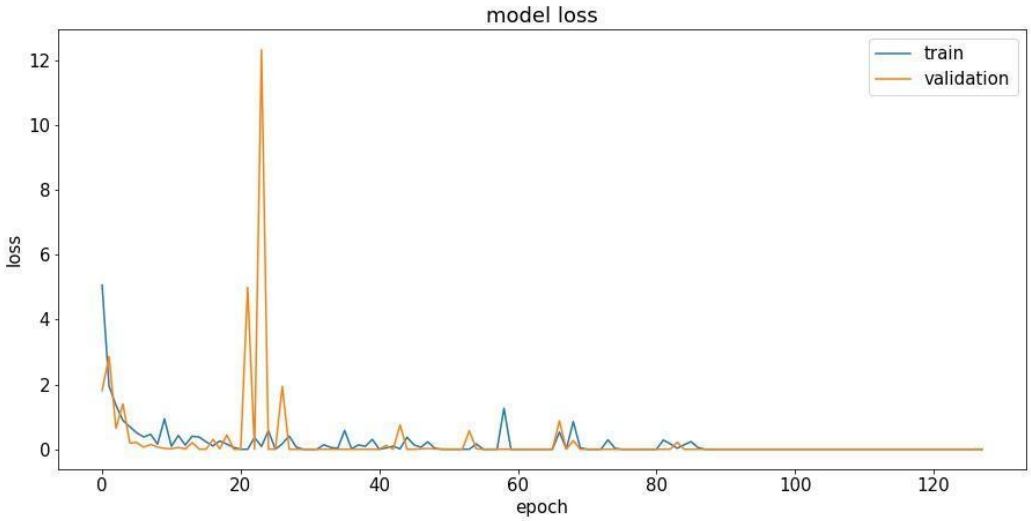


(a) 损失

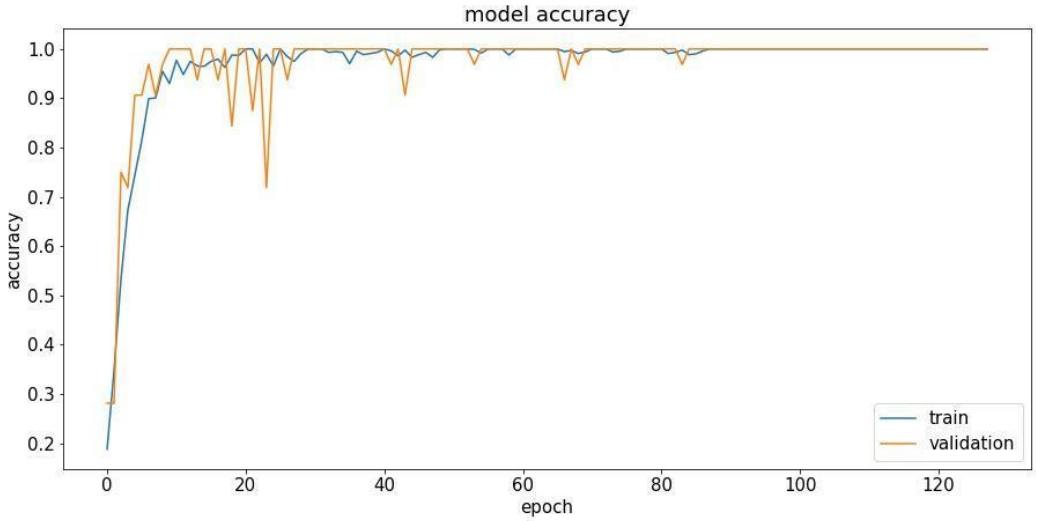


（b）准确率

图 7. 基于音频特征进行情绪识别的模型损失和准确率



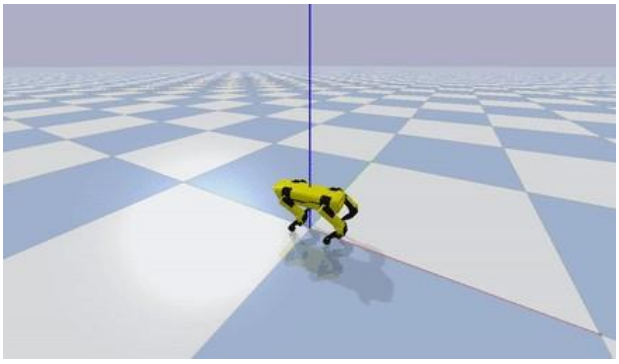
（a）损失



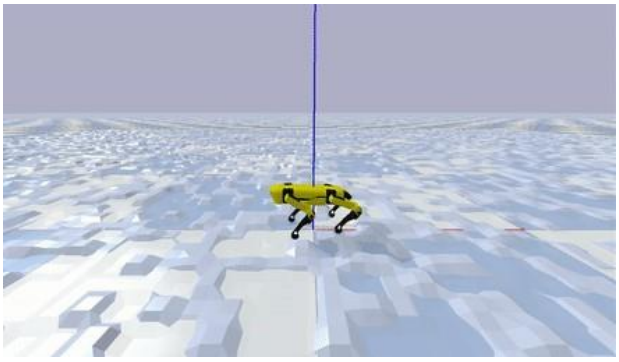
（b）准确率

图8. 基于视频特征进行情绪识别的模型损失和准确率

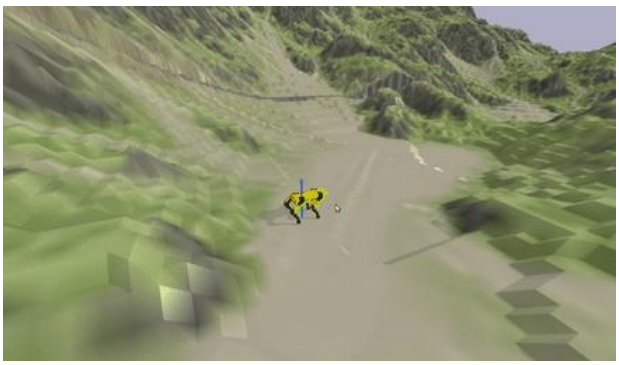
步态生成仿真符合预期，如图9所示，机器人在所有模拟地形上都实现了运动。我们使用这种健身房环境来学习如何优雅地启动不同的动作，并在到达目标位置后停止这些动作。行走使用二维动作空间，其反馈分量为 π(o)，范围为 [-0.4, 0.4]，；奔跑则使用二维动作空间，其反馈分量为 π(o)，范围为 [-0.3, 0.3]。对于行走和奔跑，正确的启动方式都有助于消除步态在最终学习策略中产生的漂移效应。站立的动作空间为 1，反馈分量为 π(o)，范围为 [-0.1, 0.1]，用于优化信号时序。信号函数会施加“制动”，迫使机器人先采取中间位置，然后再完成动作。

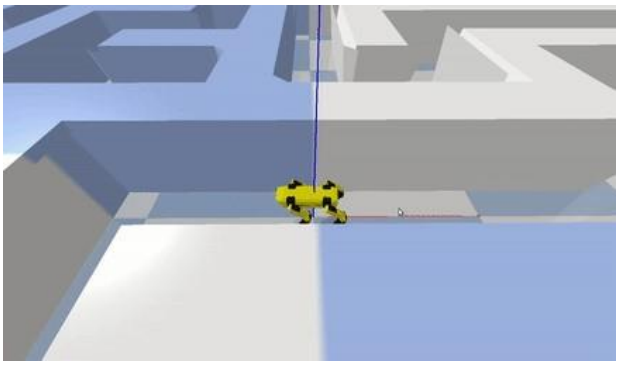


（a）平坦地形



（b）不平坦地形





(c) 丘陵地形 (d) 迷宫

图9. 四足机器人仿真

6. 结论与未来展望

本节将讨论一些更实用的功能：如果具备所需的资源和时间，可以为这款四足机器人添加这些功能。首要关注点应是避免我们所说的“绊倒摔死”。为说明这一点，我们需要明确指出，目前步态生成模型尚未实现动态避障。在实际制造时，还必须考虑在机器人中加入陀螺稳定模块。谈到更以人类为中心的功能，可以考虑通过对机器人周围的人类进行骨架结构分析来实现手势识别，并提供必要的反馈。还可以实现对名字（分配给机器人的名字）的响应，以及追踪声源、找到呼叫者。如果视频情绪推断检测到紧急情绪，机器人也需要前往检测到的人类所在的位置。遇到医疗或其他紧急情况时，可以集成简单模块，以便在必要时呼叫紧急救援服务。还可以引入在线训练，让机器人学习各种行为模式，甚至可以进一步简化训练内容，学习狗会做的把戏。也可以加入辅助犬模块，在主人需要情感慰藉时给予安抚。另一个我们很希望看到的有趣功能是哨兵模式：四足机器人会沿着指定区域的边界行走，对记忆中不认识的人吠叫，甚至可能在任何试图闯入该处所的潜在罪犯实施袭击之前发出警告。

如上所述，本文旨在建立一个框架，用于仿真具备情绪智能的四足机器人，使其能够主要通过运动或音频响应来回应视听刺激。尽管如此，本研究主要是作为概念验证而构思的，旨在展示：借助现有文献中的研究成果，我们能够整合出一个可以模仿真实狗的系统。在仿真中，我们展示了如何利用更新的集成架构，让梅尔频率倒谱系数（MFCC）算法等古老技术重焕生机。这种音频情绪检测的表现相对不错，准确率约为 63.47%，且所需计算资源极少。不过，值得注意的是，它的表现不及相关数据集上的现有研究水平，例如，ERANNs 在 RAVDESS数据集上的最高准确率为 73.4% [34], ，或者在 SAVEE数据集上进行的 Zeta策略训练，其准确率为

68.90 ± 0.61%。另一方面，视频情绪检测系统是一种专为本研究开发的新型架构，其结果几乎达到了最先进水平。该系统的准确率为 99.66%，仅次于 FAN [35] 和 Vit + SE [36] 模型，后两者的准确率分别为 99.7% 和 99.8%。另一个主要组成部分是运动（自动步态生成），本研究通过近端策略优化（PPO）算法和 TDoA算法对其进行仿真。近端策略优化（PPO）算法凭借其“在策略上”的学习方法，学习速度尤其快，使仿真狗能够在不同步频和变化中展现出非常平稳的步态。由此，仿真四足机器人能够对仿真刺激做出反应，因此可以说，它的表现符合预期，也达到了本文的目标。

参考文献

[1] W. Gibson，《神经漫游者》（1984），载于《犯罪与媒体》，劳特利奇，2019年。[2] M. Hutter、C. Gehring、D. Jud 等人，“Anymal——一种高机动性、高动态性的四足机器人”，载于《2016年 IEEE/RSJ 智能机器人与系统国际会议（IROS）》，IEEE，2016年，第38–44页。[3] Spot®——灵活敏捷的移动机器人。[在线]。网址：https://www.bostondynamics.com/products/spot。[4] Aliengo。[在线]。网址：https://www.unitree.com/products/aliengo/。[5] M. Raibert、K. Blankespoor、G. Nelson 和 R. Playter，“Bigdog：崎岖地形四足机器人”，IFAC论文集，第41卷，第2期，第10822–10825页，2008年。[6] J. Deng、Z. Zhang、E. Marchi 和 B. Schuller，“基于稀疏自编码器的语音情绪识别特征迁移学习”，载于《2013年人类协会情感计算与智能交互会议》，IEEE，2013年，第511–516页。[7] Z. Lu、L. Cao、Y. Zhang、C.-C. Chiu 和 J. Fan，“通过端到端自动语音识别模型的预训练特征进行语音情感分析”，载于《ICASSP 2020—2020年 IEEE 声学、语音与信号处理国际会议（ICASSP）》，IEEE，2020年，第7149–7153页。[8] P. Mermelstein，“语音识别的距离度量：心理学与仪器学视角”，《模式识别与人工智能》，第116卷，第374–388页，1976年。[9] S. Davis 和 P. Mermelstein，“连续语句中单音节词识别的参数化表示比较”，《IEEE声学、语音与信号处理汇刊》，第28卷，第4期，第357–366页，1980年。[10] J. S. Bridle 和 M. D. Brown，“一种实验性自动语音识别系统”，JSRU 报告，第1003卷，第5期，第33页，1974年。[11] S. Maghilnan 和 M. R. Kumar，“针对特定说话人语音数据的情感分析”，载于《2017年智能计算与控制国际会议（I2C2）》，IEEE，2017年，第1–5页。[12] V. Gajarla 和 A. Gupta，“图像的情绪检测与情感分析”，佐治亚理工学院，第1–4页，2015年。

[13] K. Simonyan 和 A. Zisserman，“用于大规模图像识别的超深卷积网络”，arXiv 预印本 arXiv:1409.1556，2014年。[14] P. R. Dachapally，“使用卷积神经网络和表征自编码器单元进行面部情绪检测”，arXiv 预印本 arXiv:1706.01509，2017年。[15] S. A. Bargal、E. Barsoum、C. C. Ferrer 和 C. Zhang，“利用图像从视频中识别自然环境下的情绪”，载于《第18届 ACM 多模态交互国际会议论文集》，2016年，第433–436页。[16] B. Hengst、D. Ibbotson、S. B. Pham 和 C. Sammut，“四足机器人的全向运动”，载于《机器人世界杯》，Springer，2001年，第368–373页。[17] M. S. Kim 和 W. Uther，“四足机器人的自动步态优化”，载于《澳大拉西亚机器人与自动化会议》，Citeseer，2003年，第1–3页。[18] W. H. Press、W. T. Vetterling、S. A. Teukolsky 和 B. P. Flannery，《C语言数值配方：++ 科学计算的艺术》，剑桥大学出版社，2001年。[19] J. Clune、B. E. Beckmann、C. Ofria 和 R. T. Pennock，“利用 HyperNEAT 生成式编码演化协调四足步态”，载于《2009年 IEEE 进化计算大会》，IEEE，2009年，第2764–2771页。[20] D. Owaki 和 A. Ishiguro，“一种能够自发行走、小跑与奔跑步态转换的四足机器人”，《科学报告》，第7卷，第1期，第1–10页，2017年。[21] S. Grillner，“脊椎动物运动：中枢机制与反射交互作用”，《生理学评论》，第55卷，第2期，第247–304页，1975年。[22] ——，“脊椎动物节律性运动行为的神经生物学基础”，《科学》，第228卷，第4696期，第143–149页，1985年。[23] M. Lodi、A. Shilnikov 和 M. Storace，“设计可产生目标四足步态的合成中枢模式发生器”，《IEEE电路与系统汇刊 I：常规论文》，第65卷，第3期，第1028–1039页，2017年。[24] M. Lodi、A. Shilnikov 和 M. Storace，“CEPAGE：用于中枢模式发生器分析的工具箱”，载于《2017年 IEEE 国际电路与系统研讨会（ISCAS）》，IEEE，2017年，第1–4页。[25] S. R. Livingstone 和 F. A. Russo，“瑞尔森情绪语音与歌曲视听数据库（RAVDESS）：北美英语中的面部和声音表达动态多模态数据集”，《公共科学图书馆·综合》，第13卷，第5期，e0196391，2018年。[26] H. Cao、D. G. Cooper、M. K. Keutmann、R. C. Gur、A. Nenkova 和 R. Verma，“CREMA-D：众包情绪多模态演员数据集”，《IEEE情感计算汇刊》，第5卷，第4期，第377–390页，2014年。[27] P. Jackson，“Sana ul haq”，《萨里视听情绪表达（SAVEE）数据库》，第4卷，2011年。[28] K. Dupuis 和 M. K. Pichora-Fuller，“衰老会影响对语义中性句子中声音情绪的识别”，《言语、语言与听力研究杂志》，第58卷，第3期，第1061–1076页，2015年。

[29] P. Lucey、J. F. Cohn、T. Kanade、J. Saragih、Z. Ambadar 和 I. Matthews，《扩展版 Cohn-Kanade 数据集（CK+）：用于动作单元和情绪指定表情的完整数据集》，载于《2010 年 IEEE 计算机学会计算机视觉与模式识别会议研讨会》，IEEE，2010 年，第 94–101 页。[30] K. Enoksson 和 B. Zhou，《声音跟随机器人》，2017年。[31] K. Ho 和 Y. Chan，《基于到达时间差的地理定位解决方案与性能分析》，《IEEE 航空航天与电子系统汇刊》，第 29 卷，第 4 期，第 1311–1322 页，1993 年。[32] CKplus。 [在线]。网址：https://paperswithcode.com/sota/facial-expression-recognition- on-ck。[33] H. Ding、S. K. Zhou 和 R. Chellappa，《FaceNet2ExpNet：通过正则化深度人脸识别网络实现表情识别》，载于《2017 年第 12 届 IEEE 自动人脸与手势识别国际会议（FG 2017）》，IEEE，2017年，第 118–126 页。[34] S. Verbitskiy、V. Berikov 和 V. Vyshegorodtsev，《ERANNs：用于音频模式识别的高效残差音频神经网络》，arXiv 预印本 arXiv:2106.01621，2021 年。[35] D. Meng、X. Peng、K. Wang 和 Y. Qiao，《用于视频面部表情识别的帧注意力网络》，载于《2019 年 IEEE 国际图像处理会议（ICIP）》，IEEE，2019 年，第 3866–3870 页。[36] M. Aouayeb、W. Hamidouche、C. Soladie、K. Kpalma 和 R. Seguier，《结合挤压与激励的视觉 Transformer 学习，用于面部表情识别》，arXiv 预印本[37] Rex：开源四足机器人，https://pypi.org/project/rex-gym/，访问日期：2022 年 2 月 23 日

声明：

经费：本研究未获得任何经费支持。 .
利益冲突：作者声明，他们没有已知的、可能影响本文所报告工作的财务利益冲突或个人关系。

作者贡献：所有作者均参与了研究的构思与设计。Abhirup 参与了概念构思。Jatin 参与了概念实施。整个研究由 Sibi Chakravarthy 和 Aswani Kumar 负责指导。初稿由 Anitha 撰写，并对分析进行了比较。Firuz 提供了重要反馈，并协助解读研究结果。Annapurna 为实施提供了意见，核验了数值结果并负责协调。

伦理审批：本研究不涉及任何人类或动物受试者研究，因此本文不适用伦理审批。

参与同意：本研究不涉及任何人类或动物受试者研究，因此本文不适用此项审批。

发表同意：本研究不涉及任何个人数据，因此该同意事项与本文无关。

致谢：无
代码/数据可用性声明：由于当前研究未生成或分析任何数据集，代码/数据共享不适用于本文。