doi:10.6041/j.issn.1000-1298.2024.04.022

融合 Res3D、BiLSTM 和注意力机制的羊只行为识别方法

袁洪波　曹润柳　程　曼
（河北农业大学机电工程学院，保定 071001）

摘要：识别动物行为可以为疾病预防和合理喂养提供重要依据，从而有助于更好地关注动物的健康和福利。本文提出了一种融合三维残差卷积神经网络、双向长短期记忆网络和注意力机制的深度学习网络模型（AdRes3D－BiLSTM）。AdRes3D－BiLSTM 模型可以直接针对视频流进行识别，在 AdRes3D 部分引入了深度可分离卷积和注意力机制，不但减少了浮点运算量，提升了网络轻量化程度，还提高了时间和空间两个维度的特征提取能力；提取的特征被输入 BiLSTM 模块后，从前后 2 个方向对时序特征向量进行筛选和更新，最后对羊只行为进行准确识别。试验结果表明，AdRes3D－BiLSTM 对羊只站立、躺卧、进食、行走和反刍 5 种行为的综合识别准确率达到了 98.72%，帧速率达到 52.79 f/s，模型内存占用量为 28.03 MB。研究结果为基于视频流的动物行为识别提供了新的方法和思路。

关键词：羊只；行为识别；视频流；Res3D；BiLSTM；注意力机制
中图分类号：TP391.4　　**文献标识码**：A　　**文章编号**：1000-1298(2024)04-0221-10　　**OSID**：

Fusion of Res3D, BiLSTM and Attention Mechanism for Sheep Behavior Recognition Method

YUAN Hongbo　CAO Runliu　CHENG Man
(*College of Mechanical and Electrical Engineering*, *Hebei Agricultural University*, *Baoding* 071001, *China*)

Abstract: In intensive sheep farms, behavioral changes can map out whether there are abnormalities in the sheep's body. For example, when sheep are sick, rumination and feeding time will produce significant changes, and behavioral observation is one of the ways to diagnose their health. Identifying animal behavior can provide a basis for disease prevention and rational feeding, thus improving the focus on animal health and welfare. Therefore, animal behavior recognition has always been a focus of attention for researchers and production managers. Traditional manual observation methods require continuous human monitoring, and the fatigue response from long hours of human work tends to cause subjective errors in the results. In addition, sensor detection methods that require direct contact with the animal's body tend to stress the animal, affecting animal health and production performance. A deep learning network model AdRes3D－BiLSTM was proposed that incorporated a three dimensional residual convolutional neural network, a bi-directional long and short-term memory network, and an attention mechanism. The AdRes3D component introduced depth separable convolution, a technique instrumental in curtailing computational complexity and enhancing network efficiency. Furthermore, an actionnet attention mechanism based on motion principles was embedded within the AdRes3D section, directing the network's focus toward discerning behavioral nuances. This augmentation amplified the model's adeptness in extracting pivotal behavioral key points across consecutive video frames, thereby augmenting its capacity for feature extraction across both temporal and spatial dimensions. Subsequently, the feature vectors extracted from this process were inputted into the BiLSTM module, affording bidirectional filtering and updating for temporal features, and the final sheep behaviors were accurately recognized. A dataset comprising 6 000 distinct videos was amassed for training the proposed model. This dataset encompassed different sheep instances, spanning varying periods, lighting conditions, and poses. An additional set of

收稿日期：2023－08－20　修回日期：2023－09－17
基金项目：河北省重点研发计划项目(21327402D)
作者简介：袁洪波(1980—)，男，副教授，主要从事智能检测及自动控制技术和精细农业系统集成研究，E-mail：yuanhongbo222@163.com
通信作者：程曼(1982—)，女，教授，博士生导师，主要从事精准畜牧及动植物表型信息研究，E-mail：chengman1982@163.com

1 200 behavioral videos, distincting from those employed in training, was selected as the testing data. The experimental results showed the efficacy of the AdRes3D - BiLSTM model, as evidenced by an exceptional comprehensive recognition accuracy rate of 98.72% across five fundamental sheep behaviors: standing, lying, feeding, walking, and ruminating. In contrast to five alternative network architectures—namely, C3D, R(2 + 1)D, Res3D, Res3D - LSTM, and Res3D - BiLSTM - the AdRes3D - BiLSTM model achieved notable improvements in recognition metrics. Specifically, relative to these network models, AdRes3D - BiLSTM exhibited a precision enhancement of 11.32 percentage points, 6.24 percentage points, 4.34 percentage points, 2.04 percentage points and 1.52 percentage points, respectively. The corresponding improvements in recognition recall stood at 11.78 percentage points, 6.38 percentage points, 4.38 percentage points, 2.12 percentage points and 1.68 percentage points, while F1-score improvements registered at 11.70 percentage points, 6.35 percentage points, 4.38 percentage points, 2.08 percentage points and 1.60 percentage points, and the augmentation in recognition accuracy was quantified at 11.97 percentage points, 6.33 percentage points, 4.37 percentage points, 2.32 percentage points and 2.01 percentage points. Furthermore, the method elucidated boasted an impressive frame rate, attaining a remarkable 52.79 frames per second (FPS). This recognition speed substantiated the model's real-time processing capabilities, thereby satisfying operational demands. Additionally, a 24-hour uninterrupted video segment was randomly culled from the repository of collected videos, effectively validating the model's efficacy in a real-world environment. This investigation ushers in novel methodologies and conceptual insights for animal behavior recognition based on video streams. The strides furnished fresh avenues for advancing the field, presenting innovative strategies and perspectives for further exploration and implementation.

Key words: sheep; behavior recognition; video streaming; Res3D; BiLSTM; attentional mechanism

0 引言

动物行为通常可以作为评估其健康状况的一个重要参考[1-2],当动物处于生病、疼痛状态或特定的生理阶段时会表现出不同的行为特征[3]。例如绵羊在感染羊痒螨后会表现出抓挠、撕咬等异常行为[3];当绵羊患病时,其反刍时间会有明显的变化[4]。因此,监测动物行为表现可以用于疾病早期检测[5-7],例如可以根据绵羊行走状态来诊断跛行[6],根据反刍活动判断瘤胃健康状态[8]。利用人工观察进行动物行为监测不仅耗时、低效而且容易产生主观错误[9],因此已经不适应规模化养殖的需要[10]。基于盈利和动物福利考虑,设计合理的自动监测和识别分类系统可以有效替代人工进行动物行为监测[11-12]。

基于计算机视觉的动物行为监控方法因为具有非接触、无应激和效率高的特点已经逐渐开始替代可穿戴式传感器,被用到精准畜牧养殖领域[7],并在动物行为评估中发挥着重要作用[13]。随着深度学习的迅速发展,将计算机视觉和深度学习结合进行动物行为监控和分析得到了较多的研究。研究人员基于获取的图像通过深度学习进行动物行为识别和分类,很多深度学习模型得到了应用,比如 Faster R - CNN[14]、YOLO[15]、SSD[16]、Mask R - CNN[17]、FCN[18]等被用于识别动物的行为。利用深度学习进行动物行为识别的前提是基于大量的训练样本对模型进行训练,而样本制作通常采用图像标记和注释的方式,这需要消耗大量的人力和时间成本[10]。此外,基于图像进行深度学习模型的训练,只能从静止图像中提取空间特征,这忽略了行为的连贯性和时间特征(运动信息)。行为是由在时间序列上的一系列连续活动组成的,连续帧之间的时间信息对于判断行为类别是很重要的,尤其是在区分类似行走和站立等活动和非活动行为时。然而基于静态图像去定义或分类这些特征可能会面临挑战[19]。因此,有必要结合空间和时间信息来识别动物行为。

基于连续的视频序列进行深度学习模型的训练和识别可以克服基于静态图像所带来的问题[20-21]。因此,基于视频的动物行为识别近年来得到了大量研究,一些研究人员尝试从视频的多帧连续图像中提取运动特征以获得更丰富的动态行为状态,并对行为类别进行判断。目前,这些研究已经被应用到多种动物的行为识别中,例如牛的站立和行走[21]、饮水、进食和反刍[22-23]、跛行[24]等。在猪的进食、躺卧、行走、抓挠、攀爬和爬跨行为[19,25-26]、咬尾行为[27]、饮水和饮水时游戏[28]、侵略性行为[29]等的行为识别研究中也使用了基于视频和深度学习的方法。提取视频特征所用到的深度学习网络,包括但不限于双流卷积网络[30]、三维卷积神经网络[31]和 LSTM 网络[32-33]等。作为一种特殊的 RNN 网络,LSTM 网络具有记忆长短期信息能力,可以处理时序问题。与对行为进行建模的双流和三维卷积网络

不同，基于 LSTM 的方法将视频视为有序的帧序列，行为可以通过每一帧的特征变化来表示。BiLSTM 是由前向 LSTM 和后向 LSTM 组合而成，可以分别捕获绵羊过去和未来隐含的时间序列特征[23]。考虑到动物行为转换的双向性质，例如动物在某一时间段内发生从站立到行走，再从行走到站立的转换，使用双向 LSTM 可以改善分类结果。此外，2 个行为状态之间转换的过渡过程容易造成识别的干扰[34]，而在实际监测中过渡过程不能忽略，使用双向 LSTM 将有助于显示过渡过程的正确分类[35]。

残差网络在图像领域是一种主流模型，内部使用跳跃连接，通过使用残差连接，能够通过增加深度来提高准确率并减少过拟合的发生。Res3D 作为常用的三维残差卷积神经网络，在行为识别领域得到了广泛应用[36-37]。为了提高动物行为识别能力，减少制作深度学习模型训练集时人工和时间成本，本文在三维残差卷积神经网络 Res3D 和 BiLSTM 的基础上进行优化，设计一种深度学习网络用于绵羊行为快速精准识别。

1　材料和方法

1.1　试验环境与数据获取

本研究使用的视频数据来自河北省衡水市志豪畜牧科技有限公司的现场试验，试验时间为 2021 年 2 月 7 日—4 月 25 日。试验时随机选择一个羊舍内的 4 个围栏，每个围栏的尺寸均为 3 m × 2.5 m。试验选择的羊舍是孕羊专用羊舍，每个围栏内有一只怀孕的小尾寒羊。当孕羊分娩后被转移，新的怀孕母羊被放置进来。每个围栏内安装一个摄像机（海康威视，2CD2T46XMV3 - LGLSE）。摄像机安装位置距离地面 2 m，向下倾斜 65°拍摄，这种安装方式能够使拍摄的视频画面仅包括目标围栏中的羊只活动区域。试验过程中，共获取 13 只羊的视频，这些视频包括羊只进食、行走、站立、躺卧和反刍等各种行为。所有试验均通过河北农业大学动物实验伦理委员会审批（ID:1090064）。

1.2　数据集生成方法

从 13 只羊的视频里随机选择 4 只羊的视频，采用公开数据集 UCF101 的格式来制作视频训练数据集。视频处理过程中，对羊只的进食、站立、行走、躺卧和反刍 5 种行为进行观察和剪辑，使每一段短视频中只包含一种行为，羊只不同行为定义如表 1 所示。经过处理，共得到 6 000 个短视频，其中每种行为短视频数量均为 1 200 个，视频时长为 5 ~ 15 s，帧率为 30 f/s。制作完成的短视频按照行为的不同被分别放到对应的文件夹下，文件夹采用行为类型来命名。由于行为类型不同，每个短视频长度并不完全一致，所以导致每种行为类型的视频总长度也不完全相同，但是这种差异是正常且可以被接受的[38]。此外，制作完成的短视频包含不同时间段和不同光照情况的样本，这保证了该方法对不同照明条件的适应性和训练数据的丰富性。此外，从原始视频中另外制作 1 200 个时长为 6 s 的短视频构成测试数据集。测试数据集中羊只每种行为短视频数量均为 240，且测试数据集与训练数据集中的视频不重复。

表 1　行为数据集信息

Tab. 1　Behavior dataset information

行为	描述	视频数量
站立	羊只 4 条腿均处于直立状态，且身体不产生明显位移	1 200
躺卧	羊只 4 条腿均跪坐于地面，腹部紧贴地面	1 200
进食	羊只站立在食槽旁边，且头部向下进入食槽中	1 200
行走	羊只 4 条腿中有 2 条腿接触地面且身体产生明显的连续性位移，导致身体所处位置发生变化	1 200
反刍	羊只将食团反胃到口中后嘴部进行咀嚼的过程[13,39]	1 200

1.3　基于 AdRes3D - BiLSTM 模型的羊只行为识别网络

羊只的行为是一个连续的状态，上一时刻的行为与下一时刻行为之间具有很强的相关性和连续性。与图像相比，视频流同时包含空间和时间特征信息，能够很好地反映出连续帧之间的相关性和连续性。利用深度学习网络可以提取出运动行为之间的时序特征，从而判断出不同的行为类型。本文设计了一种融合三维残差卷积神经网络、注意力机制和双向长短期记忆网络的深度学习网络模型（AdRes3D - BiLSTM），基于视频流识别羊只的不同行为。

图 1 为 AdRes3D - BiLSTM 网络模型结构图，该模型首先使用 Res3D 模块分析视频流中连续帧的时空信息，然后将提取的特征向量序列输入 BiLSTM 模块进行时序特征筛选和更新，最后根据 BiLSTM 的输出对羊只行为结果进行识别和分类。本文设计的深度学习网络在 Res3D 模块的特征提取部分引入了深度可分离卷积，大幅度减少了浮点运算量和模型大小，并嵌入基于运动的注意力机制，专注于行为特征以提高识别能力。在 Res3D 特征提取模块中，Conv3D 表示 3D 卷积操作，3 × 7 × 7 和 3 × 3 × 3 表示卷积核尺寸，64、128、256 和 512 表示通道数，DSConv 表示深度可分离卷积，Conv 表示普通卷积，折线表示残差连接，Fully connect（FC）表示全连接

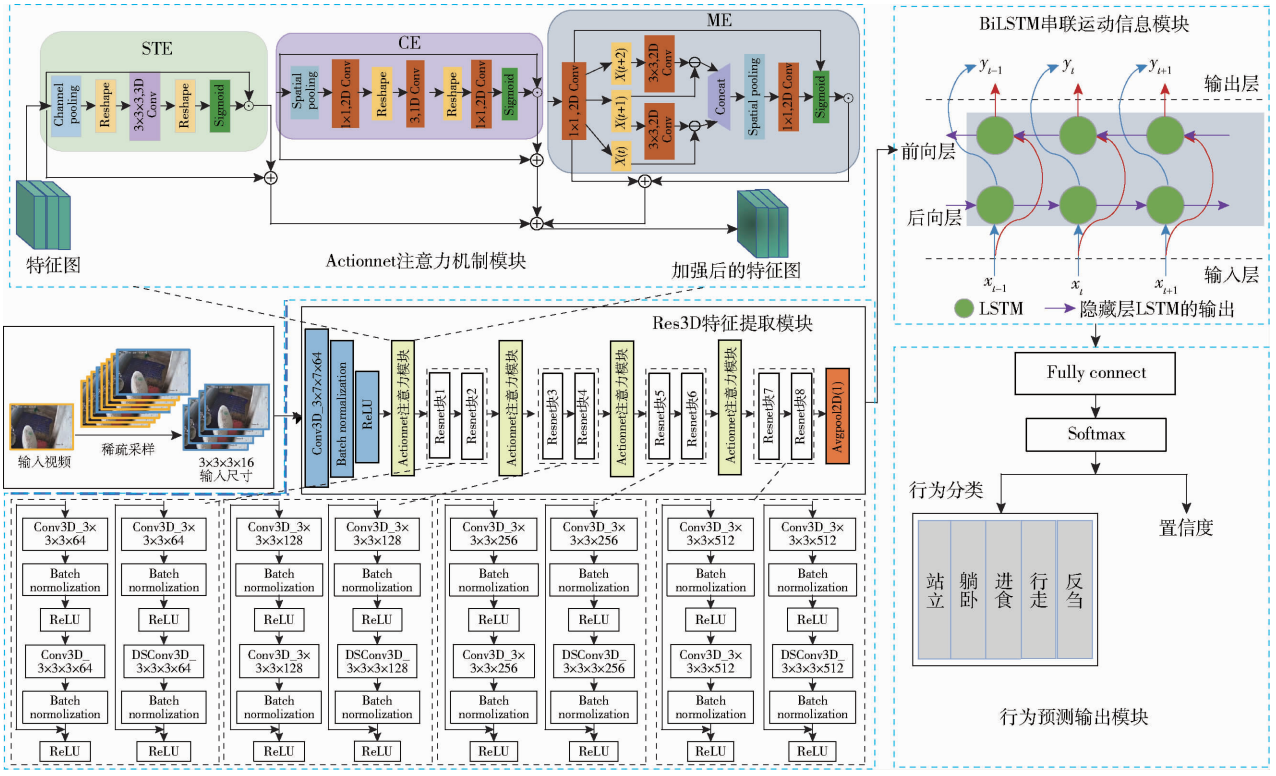

图 1 AdRes3D－BiLSTM 结构

Fig. 1 AdRes3D－BiLSTM structure

层;在 Actionnet 注意力模块中,Conv 为卷积操作,3×3×3、1×1、3×3 表示卷积核尺寸,Channel pooling 表示对通道上的输入张量进行平均,Spatial pooling 表示对输入进行空间平均池化,$X(t)$、$X(t+1)$、$X(t+2)$表示相邻的帧,Concat 表示向量数据的拼接操作,Sigmoid 为深度学习中常用的激活函数。

视频流的各个组成帧之间存在较强的时空关联性,二维卷积只能提取空间特征而无法提取时间关联信息。多个连续的帧可以叠加成一个视频空间,在这个空间中可以使用三维(Three dimensional,3D)卷积核代替二维卷积核执行卷积操作,这样实现了空间范围内卷积操作代替平面上的卷积操作。3D 卷积可以捕捉到连续帧之间存在的时空特征,即可以整合空间和时间信息[40],进而捕捉到羊只行为的运动信息,这样可以有效提升分类精度。此外,18 层的 Res3D 网络结构深度在视频分类的准确性、计算复杂度和内存之间达到较好的平衡[41],所以本文采用 18 层的 Res3D 网络作为前期提取网络。

深度可分离卷积可以将标准化卷积分解为深度卷积和逐点卷积[42],运算时深度卷积先对输入向量的每个通道进行卷积,然后利用逐点卷积完成对深度卷积输出向量的组合。与普通卷积相比,深度可分离卷积可以显著减少参数量。如图 1 所示,为了防止网络不稳定,在 Resnet 块 2、Resnet 块 4、Resnet 块 6、Resnet 块 8 中选择交叉使用深度可分离卷积和普通卷积,使网络模型轻量化,克服了 3D 卷积计算量大的缺点。

输入网络的视频中不同的帧对判断行为类别起的作用不同,比如羊只由站立行为过渡到行走行为时,表现出羊只行走的帧往往比表现出羊只将要行走从而发生身体移动的帧更加关键,此类动作有较强的时序性。为了区分各帧中特征图的重要程度,本文在 Res3D 网络中引入基于运动的混合注意力机制模块 Actionnet[43]来重新分配特征图之间的权重,加大对关键特征图的关注,弱化冗余特征图对识别结果的影响,从而提高网络提取行为特征的鲁棒性,其构成如图 1 所示。Actionnet 由时空注意力模块(STE)、通道注意力模块(CE)和运动注意力模块(ME)3 个子模块组成。在提取羊只行为特征过程中,STE 子模块利用 3D 卷积激发时空信息,提取出对分类有用的关键特征信息;CE 子模块通过在 2 个 FC 层之间插入一维卷积层来表征通道特征的时间信息;ME 子模块利用相邻帧之间的时间维度建模运动信息。为了可以同时融合时空、通道、运动 3 个方面的特征信息,将这 3 个子模块并联组成 Actionnet 模块,Actionnet 模块最终的输出量通过这 3 个子模块各自生成的特征相加得到。Actionnet 模块兼顾了时空信息、不同通道间的时序信息和相邻帧之间动作的变化轨迹 3 个重要信息,使网络更关

注真正感兴趣的区域，对相对重要的信息分配更多的权重，忽略不重要的信息。

RNN 由于梯度消失的原因只能有短期记忆，而 LSTM 引入了门函数（遗忘门、输入门、输出门）和记忆单元，很好地解决了这个问题，是一种更稳健的 RNN 形式[32]。如图 1 所示，BiLSTM 由一个前向 LSTM 层和一个后向 LSTM 层构成，前向 LSTM 利用先前的信息来预测当前行为类别，后向 LSTM 利用未来的信息来预测当前行为类别，上下文信息同时得到利用的方式有助于提高模型的行为识别能力。与传统的单向 LSTM 相比，BiLSTM 可以同时学习过去和未来的信息，从而获得更加稳健的时间特征，这样能够更好地获得视频中羊只运动时产生的时序数据特征之间的相关性。因此，本文在 Res3D－Actionnet 模块之后采用 BiLSTM 来进一步处理时序数据。最后经过行为预测模块的全连接层输入到 Softmax 层，最终输出行为预测结果和置信度。

1.4　网络设置及运行环境

本文设计的 AdRes3D－BiLSTM 网络模型训练时输入图像尺寸为 112 像素×112 像素，网络的初始学习率为 0.01，利用迭代周期（epoch）衰减策略，每迭代 10 次通过将学习率除以 10 来更新一次参数值，优化器为 Adam。训练批量大小设为 8。隐藏层维数设为 512，经过 Bi－LSTM 提取的特征向量序列输入到全连接层的维数为 1 024，输出维数为 5 的向量。网络最后的 Softmax 层的参数 dim 设为 1。试验进行的迭代次数为 600 次，每 10 次保存一个权重模型，共保存 60 个权重文件。

训练数据集中随机抽取 64% 的短视频作为训练网络时的训练数据，20% 和 16% 的短视频分别作为训练过程中用于模型性能分析的测试数据和验证数据。羊只的运动是一个持续性的过程，不可能在瞬间发生突变。为了提高在训练和识别过程中的效率，对数据集中的短视频以每 3 帧提取 1 帧的方式进行稀疏采样。将每个视频转换出来的图像序列按照视频的原始顺序命名，从中选取连续的 16 帧作为一次输入。

本文构建的 AdRes3D－BiLSTM 网络模型基于 Python 3.6 编程语言和 Pytorch 1.0.0 深度学习框架来实现。运行网络的计算机采用 i9－10900 K 3.70 GHz CPU，NVIDIA GeForce RTX 3090 GPU，32 GB 内存，操作系统为 64 位 Ubuntu 20.04，CUDA 版本为 11.4。

1.5　网络模型性能评价指标

采用精确率（Precision，P）、召回率（Recall，R）、F1 值（F1-score）和准确率（Accuracy，A）4 个指标对网络模型的性能进行评价。

此外，采用帧速率评价模型的检测速度，采用浮点运算量评价模型的计算复杂度，采用内存占用量评价模型大小。

2　结果

2.1　AdRes3D－BiLSTM 网络训练结果

模型训练过程如图 2 所示，共进行 600 次迭代。从图 2 中可以看出，在前 150 次迭代训练中，模型的识别准确率和损失值虽然出现较为明显的振荡，但是分别呈现出快速上升和快速下降的总体趋势。150～400 次迭代中，模型的识别准确率和损失值的振荡明显下降。迭代 400 次之后，模型准确率的上升趋势逐渐变缓，损失值下降趋势也逐渐变缓直至趋于稳定。

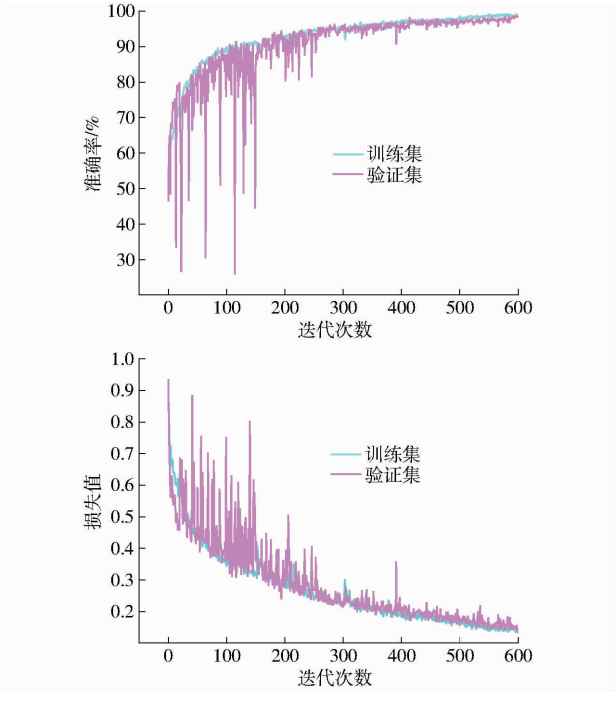

图 2　行为识别模型训练集和验证集的损失值和准确率变化曲线

Fig. 2　Loss and accuracy change curves of train and validation set for behavior recognition model

2.2　测试集识别结果

测试集中的 1 200 个短视频被输入训练完成的 AdRes3D－BiLSTM 网络中进行行为识别。输入视频时采用滑动窗口采样机制将连续的图像帧输入网络[13]，每 3 帧提取 1 帧，连续提取 16 帧作为一个滑动窗口。表 2 为测试集识别结果，从表中可以看出，站立和进食行为识别精确率较低，分别为 97.51% 和 97.70%。行走行为识别精确率最高，为 100%。从召回率和F1值来看，表现出了相同的规律。网络

表 2 羊只不同行为的精确率、召回率和 F1 值

Tab. 2 Precision, recall, and F1-score of different behaviors of sheep %

行为类别	精确率	召回率	F1 值	准确率
站立	97.51	98.00	97.75	
躺卧	98.01	98.96	98.48	
进食	97.70	97.98	97.84	98.72
行走	100	99.00	99.49	
反刍	98.99	98.47	98.72	

模型综合识别准确率达到 98.72%。这表明Res3D－BiLSTM 网络对于运动特征明显的行为具有很好的识别效果。

3 讨论

3.1 引入深度可分离卷积和 Actionnet 产生的影响

本文设计的 AdRes3D－BiLSTM 网络中引入深度可分离卷积和 Actionnet 两种操作,为了验证这两种处理模式对网络性能的影响,设计了消融试验。图 3 为引入深度可分离卷积和 Actionnet 前后,网络模型在训练过程中表现的对比。从图 3 中可以看出,当只引入深度可分离卷积后,在前 200 次迭代过程中,网络的振荡得到一定的改善。经过 200 次迭代后,网络性能的提高变得不再明显。当只引入 Actionnet 后,在整个训练过程中,网络性能的提升始终体现。这表明相对于深度卷积,Actionnet 对网络的性能影响更大。当深度可分离卷积和 Actionnet 全部引入网络后,网络性能提升更加明显,这表明引入这两种操作有效地提升了网络训练性能。

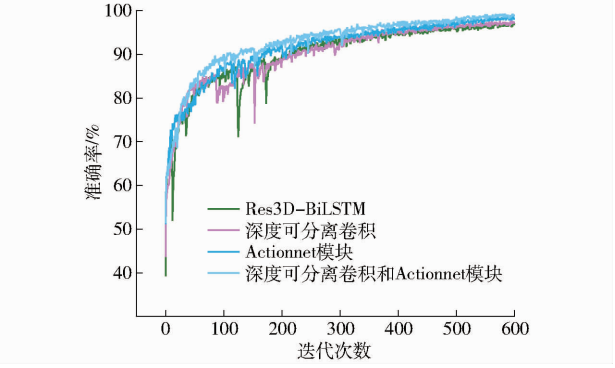

图 3 不同模块的训练精度曲线

Fig. 3 Training accuracy curves for different modules

表 3 为引入深度可分离卷积和 Actionnet 前后网络在测试集的识别结果。从表 3 可以看出,将普通卷积替换为深度可分离卷积后,网络模型的内存占用量减少了 25.06%,浮点运算量减少了 25.23%,并且帧速率达到了 61.29 f/s,执行速度提高了 15.92%,为后续应用到边缘设置奠定了基础。此外,当引入深度可分离卷积后,模型识别准确率提高了 0.59 个百分点,达到 97.30%。虽然引入深度可分离卷积对模型精度的提高影响较小,但是使模型浮点运算量和内存占用量得到了大幅度降低。因此,引入深度可分离卷积比较好地兼顾模型压缩和提升速度的需求,为网络轻量化带来了可能。单独引入 Actionnet 模块后网络在测试集的识别准确率达到了 98.50%,比原来提高了 1.79 个百分点。

表 3 不同模块在 AdRes3D－BiLSTM 网络中产生的影响

Tab. 3 Effects of different modules on AdRes3D－BiLSTM networks

评价指标	不引入深度可分离卷积和 Actionnet	只引入深度可分离卷积	只引入 Actionnet	引入深度可分离卷积和 Actionnet
内存占用量/MB	37.38	28.01	37.41	28.03
浮点运算量	3.982×10^{10}	2.977×10^{10}	3.989×10^{10}	2.985×10^{10}
帧速率/(f·s^{-1})	52.87	61.29	46.74	52.79
准确率/%	96.71	97.30	98.50	98.72

当引入深度可分离卷积和基于运动的 Actionnet 后,网络模型对测试集的识别准确率提高了 2.01 个百分点。而且帧速率达到 52.79 f/s,比视频帧率提高 75.96%,可以满足实时识别的要求。此外,模型内存占用量和浮点运算量分别下降了 25.01% 和 25.03%,以最小的计算开销实现了高精度行为识别。

3.2 不同 3D 卷积神经网络模型性能对比

近年来,三维卷积神经网络被越来越多的应用于行为识别领域,C3D[33] 和 R(2+1)D[44] 是较为常见的三维卷积神经网络模型,本团队在前期工作中设计了一种 Res3D－LSTM 网络模型并进行了检验。为了检验本文提出的 AdRes3D－BiLSTM 网络模型的性能,设计了一个对比试验。使用同样的训练集和测试集对 C3D、R(2+1)D、Res3D、Res3D－LSTM、Res3D－BiLSTM 和 AdRes3D－BiLSTM 共 6 种网络模型的识别效果进行了比较。

3.2.1 训练过程比较

图 4 为 6 种不同模型的训练精度曲线,从图中可以看出,C3D 模型在训练过程中振荡最明显,且训练结束后最终稳定值也最小。其他 5 种模型训练结果均优于 C3D,Res3D 训练结果优于 R(2+1)D,将 Res3D 和 LSTM 结合,训练性能进一步得到提升。将 LSTM 替换为 BiLSTM 后与 Res3D 结合,训练性能没有明显的改善。但是加入深度可分离卷积和 Actionnet 注意力(AdRes3D－BiLSTM)后,模型的训

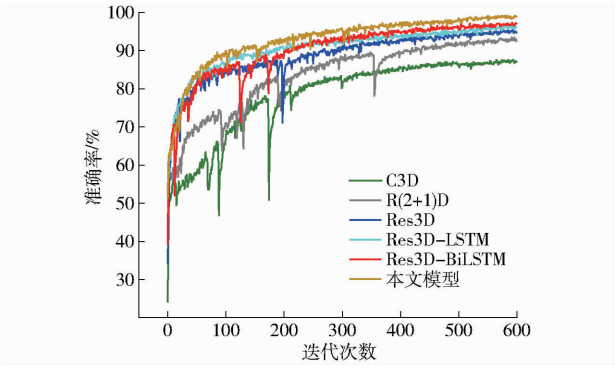

图 4　不同模型的训练精度曲线

Fig. 4　Comparison model training accuracy curves

练性能再次得到提升。此外，AdRes3D－BiLSTM 模型在整个训练过程中几乎没有出现较大的振荡。

3.2.2　测试集识别结果比较

训练完成的 6 种网络模型在同一个测试集上分别进行了行为识别，识别结果如图 5 所示。从图中可以看出，对于羊只 5 种行为，在识别的精确率、召回率、F1 值和准确率 4 个评价指标上，AdRes3D－BiLSTM 网络模型均得分最高。相对于 C3D、R(2＋1)D、Res3D、Res3D－LSTM、Res3D－BiLSTM 模型，AdRes3D－BiLSTM 的识别精确率分别提高了 11.32、6.24、4.34、2.04、1.52 个百分点，识别召回率分别提高了 11.78、6.38、4.38、2.12、1.68 个百分点，识别的 F1 值分别提高了 11.70、6.35、4.38、2.08、1.60 个百分点，识别的准确率分别提高了 11.97、6.33、4.37、2.32、2.01 个百分点。

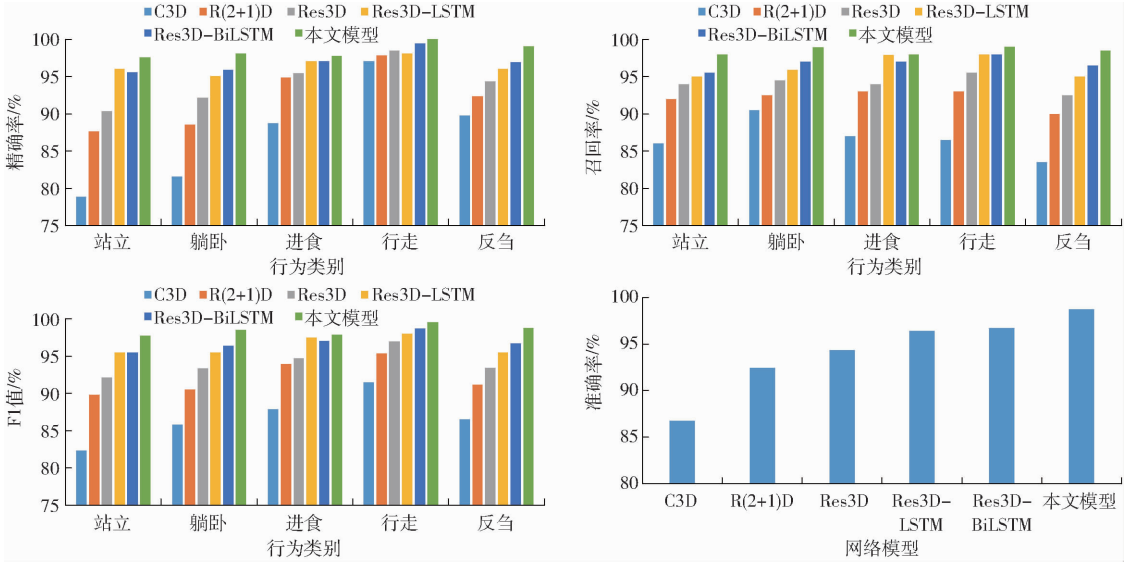

图 5　基于不同模型的羊只不同行为识别精确率、召回率、F1 值和准确率对比

Fig. 5　Comparison of precision, recall, F1-score, and accuracy based on different models in recognizing different behaviors of sheep

表 4 为 6 种不同网络模型在内存占用量、浮点运算量和帧速率 3 个性能指标上的对比结果。从表 4 中可以看出，AdRes3D－BiLSTM 模型的内存占用量最小，仅为 28.03 MB，比 C3D、R(2＋1)D、Res3D、Res3D－LSTM、Res3D－BiLSTM 分别降低了 49.99、5.15、5.15、7.25、9.35 MB。与之类似的是，AdRes3D－BiLSTM 模型浮点运算量也是 6 种模型中最小的，为 2.985×10^{10}，比 C3D、R(2＋1)D、Res3D、Res3D－LSTM、Res3D－BiLSTM 分别降低 7.86×10^{9}、1.086×10^{10}、9.96×10^{9}、9.97×10^{9}、9.97×10^{9}。在帧速率方面，AdRes3D－BiLSTM 虽然低于 C3D 的 67.92 f/s 和 R(2＋1)D 的 61.49 f/s，但是仍然达到了 52.79 f/s。在一般的视频流中，帧率一般为 30 f/s，因此 AdRes3D－BiLSTM 的实时处理速度完全可以满足要求。

表 4　不同时空特征提取网络对比

Tab. 4　Comparison of different spatio-temporal feature extraction networks

模型	内存占用量/MB	浮点运算量	帧速率/(f·s^{-1})
C3D	78.02	3.771×10^{10}	67.92
R(2＋1)D	33.18	4.071×10^{10}	61.49
Res3D	33.18	3.981×10^{10}	25.91
Res3D－LSTM	35.28	3.982×10^{10}	52.35
Res3D－BiLSTM	37.38	3.982×10^{10}	52.87
本文模型	28.03	2.985×10^{10}	52.79

3.3　长视频测试结果

为了进一步验证和分析 AdRes3D－BiLSTM 网络模型的性能，从收集的羊只生产前 7 d 的视频数据中随机选择一个 24 h 的连续视频，将 24 h 的视频数据分为 8 段分别进行试验。视频内共包含羊只站立、躺卧、进食、行走、反刍 5 种行为。图 6 为识别结果，上图为某个视频中真实行为在时间轴上

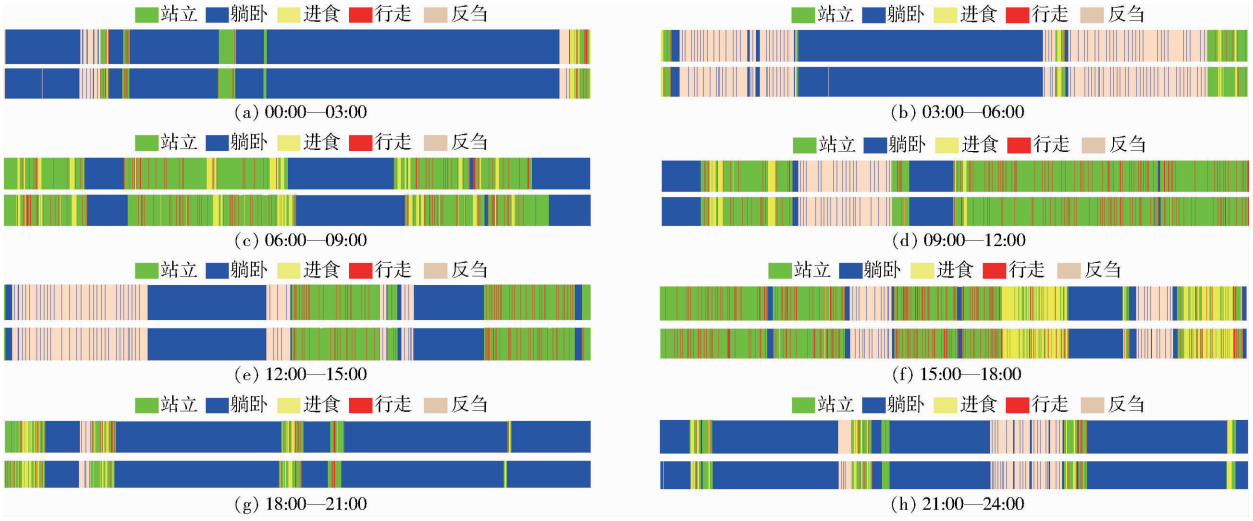

图 6 绵羊连续视频的行为识别结果

Fig. 6 Behavior recognition results of continuous sheep videos

的分布,下图为网络识别出的行为在时间轴上的分布。

从图 6 中可以看出,在一种行为向另一种行为过渡时,AdRes3D - BiLSTM 网络容易产生误识别。误识别的原因有 2 个:①羊只进食过程中,头部偶尔会短暂抬起,然后继续进食,虽然整个过程都应该属于进食状态,但是当羊只短暂抬头时在视觉上与站立行为很相似,容易造成误识别。②为了减少系统开销,使用了滑动窗口机制对视频流进行抽帧处理,这也容易导致在行为状态发生改变时做出错误的识别[13]。

4 结束语

本文融合 Res3D、BiLSTM 和注意力机制设计了一个用于羊只行为识别的三维卷积神经网络模型(AdRes3D - BiLSTM),并引入深度可分离卷积提高了深度学习模型轻量化程度。该模型可以直接基于视频进行训练并对视频进行行为识别。试验结果表明,该模型对羊只 5 种不同行为综合识别准确率达到了 98.72%,帧速率可达 52.79 f/s,能够满足针对视频流的羊只行为实时识别的要求。

参 考 文 献

[1] MEPHAM T B. The role of food ethics in food policy[J]. Proceedings of the Nutrition Society, 2020,59: 609 - 618.
[2] 齐琳, 包军, 李剑虹. 动物行为学研究在动物福利养殖中的应用[J]. 中国动物检疫, 2009(9): 68 - 69.
[3] GOUGOULIS D A, KYRIAZAKIS I, FTHENAKIS G C. Diagnostic significance of behaviour changes of sheep: a selected review[J]. Small Ruminant Research, 2010, 92(1 - 3): 52 - 56.
[4] MANSBRIDGE N, MITSCH J, BOLLARD N, et al. Feature selection and comparison of machine learning algorithms in classification of grazing and rumination behaviour in sheep[J]. Sensors, 2018, 18(10): 3532.
[5] MARTISKAINEN P, JÄRVINEN M, SKÖN J P, et al. Cow behavior pattern recognition using a three-dimensional accelerometer and support vector machines[J]. Applied Animal Behaviour Science, 2009, 119(1 - 2): 32 - 38.
[6] BARWICK J, LAMB D, DOBOS R, et al. Predicting lameness in sheep activity using tri-axial acceleration signals[J]. Animals, 2018, 8(1): 12.
[7] SU Q G, TANG J L, ZHAI M X, et al. An intelligent method for dairy goat tracking based on Siamese network[J]. Computers and Electronics in Agriculture, 2022, 193: 106636.
[8] CHEN Y J, HE D J, FU Y X, et al. Intelligent monitoring method of cow ruminant behavior based on video analysis technology [J]. International Journal of Agricultural and Biological Engineering, 2017, 10(5): 194 - 202.
[9] ALVARENGA F A P, BORGES I, PALKOVIČ L, et al. Using a three-axis accelerometer to identify and classify sheep behaviour at pasture[J]. Applied Animal Behaviour Science, 2016, 181: 91 - 99.
[10] CHENG M, YUAN H B, WANG Q F, et al. Application of deep learning in sheep behaviors recognition and influence analysis of training data characteristics on the recognition effect[J]. Computers and Electronics in Agriculture, 2022, 198: 107010.
[11] DWYER C M, LAWRENCE A B. Frequency and cost of human intervention at lambing: an interbreed comparison[J]. Veterinary Record, 2005, 157(4): 101 - 104.
[12] STOTT A W, MILNE C E, GODDARD P J, et al. Projected effect of alternative management strategies on profit and animal welfare in extensive sheep production systems in Great Britain[J]. Livestock Production Science, 2005, 97(2 - 3): 161 - 171.

[13] NASIRAHMADI A, STURM B, EDWARDS S, et al. Deep learning and machine vision approaches for posture detection of individual pigs[J]. Sensors, 2019,19(17): 3738.

[14] RAO Y,JIANG M, WANG R, et al. On-farm welfare monitoring system for goats based on internet of things and machine learning[J/OL]. International Journal of Distributed Sensor Networks, 2020. https://doi.org/10.1177/1550147720944030.

[15] 王少华, 何东健. 基于改进 YOLO v3 模型的奶牛发情行为识别研究[J]. 农业机械学报, 2021, 52(7): 141－150.
WANG Shaohua, HE Dongjian. Estrus behavior recognition of dairy cows based on improved YOLO v3 model[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(7): 141－150. (in Chinese)

[16] ZHANG Y Q,CAI J H, XIAO D Q, et al. Real-time sow behavior detection based on deep learning[J]. Computers and Electronics in Agriculture, 2019, 163: 104884.

[17] 李丹, 张凯锋, 李行健, 等. 基于 Mask R－CNN 的猪只爬跨行为识别[J]. 农业机械学报, 2019, 50(增刊): 261－266.
LI Dan, ZHANG Kaifeng, LI Xingjian, et al. Mounting behavior recognition for pigs based on Mask R－CNN[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(Supp.): 261－266. (in Chinese)

[18] YANG A Q, HUANG H S, YANG X F, et al. Automated video analysis of sow nursing behavior based on fully convolutional network and oriented optical flow[J]. Computers and Electronics in Agriculture, 2019, 167: 105048.

[19] ZHANG K F,LI D, HUANG J Y, et al. Automated video behavior recognition of pigs using two-stream convolutional networks [J]. Sensors, 2020, 20(4): 1085.

[20] FUENTES A, YOON S, PARK J, et al. Deep learning-based hierarchical cattle behavior recognition with spatio-temporal information[J]. Computers and Electronics in Agriculture, 2020, 177: 105627.

[21] QIAO Y L, GUO Y Y, YU K P, et al. C3D－ConvLSTM based cow behaviour classification using video data for precision livestock farming[J]. Computers and Electronics in Agriculture, 2022, 193: 106650.

[22] YIN X Q,WU D H, SHANG Y Y, et al. Using an EfficientNet－LSTM for the recognition of single cow's motion behaviours in a complicated environment[J]. Computers and Electronics in Agriculture, 2020, 177: 105707.

[23] WU D H, WANG Y F, HAN M X, et al. Using a CNN－LSTM for basic behaviors detection of a single dairy cow in a complex environment[J]. Computers and Electronics in Agriculture, 2021, 182: 106016.

[24] JIANG B,YIN X Q, SONG H B. Single-stream long-term optical flow convolution network for action recognition of lameness dairy cow[J]. Computers and Electronics in Agriculture, 2020, 175: 105536.

[25] LI D,ZHANG K F, LI Z B, et al. A spatiotemporal convolutional network for multi-behavior recognition of pigs[J]. Sensors, 2020, 20(8): 2381.

[26] CHEN C, ZHU W X, STEIBEL J, et al. Recognition of feeding behaviour of pigs and determination of feeding time of each pig by a video-based deep learning method[J]. Computers and Electronics in Agriculture, 2020, 176: 105642.

[27] LIU D, OCZAK M, MASCHAT K, et al. A computer vision-based method for spatial-temporal action recognition of tail-biting behaviour in group-housed pigs[J]. Biosystems Engineering, 2020, 195: 27－41.

[28] CHEN C, ZHU W X, STEIBEL J, et al. Classification of drinking and drinker-playing in pigs by a video-based deep learning method[J]. Biosystems Engineering, 2020, 196: 1－14.

[29] 高云, 陈斌, 廖慧敏, 等. 群养猪侵略性行为的深度学习识别方法[J]. 农业工程学报, 2019, 35(23): 192－200.
GAO Yun, CHEN Bin, LIAO Huimin, et al. Recognition method for aggressive behavior of group pigs based on deep learning [J]. Transactions of the CSAE, 2019, 35(23): 192－200. (in Chinese)

[30] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 27:11797475.

[31] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision, 2015: 4489－4497.

[32] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735－1780.

[33] GERS F A,SCHMIDHUBER J, CUMMINS F. Learning to forget: continual prediction with LSTM[J]. Neural Computation, 2000, 12(10): 2451－2471.

[34] ROBERT B,WHITE B J, RENTER D G, et al. Evaluation of three-dimensional accelerometers to monitor and classify behavior patterns in cattle[J]. Computers and Electronics in Agriculture, 2009, 67(1－2): 80－84.

[35] TURNER K E,THOMPSON A, HARRIS I, et al. Deep learning based classification of sheep behaviour from accelerometer data with imbalance[J]. Information Processing in Agriculture, 2023, 10(3): 377－390.

[36] LONG L J, JOHNSON Z V, LI J Y, et al. Automatic classification of Cichlid behaviors using 3D convolutional residual networks[J]. Iscience, 2020, 23: 101591.

[37] YI Z W, SUN Z H, FENG J C, et al. 3D residual networks with channel-spatial attention module for action recognition[C]// 2020 Chinese Automation Congress (CAC), 2020: 5171－5174.

[38] 叶枫, 丁锋. 不平衡数据分类研究及其应用[J]. 计算机应用与软件, 2018, 35(1): 132－136.
YE Feng, DING Feng. Research and application of unbalanced data classification[J]. Computer Applications and Software, 2018, 35(1): 132－136. (in Chinese)

[39] NICOL A U, PERENTOS N, MARTINS A Q, et al. Automated detection and characterisation of rumination in sheep using in

vivo electrophysiology[J]. Physiology & Behavior, 2016, 163: 258 - 266.
[40] OUYANG X, XU S J, ZHANG C Y, et al. A 3D - CNN and LSTM based multi-task learning architecture for action recognition [J]. IEEE Access, 2019, 7: 40757 - 40770.
[41] TRAN D, RAY J, SHOU Z, et al. ConvNet architecture search for spatiotemporal feature learning[J]. ArXiv Preprint arXiv: 1708.05038, 2017.
[42] HOWARD A G, ZHU M, CHEN B. MobileNets: efficient convolutional neural networks for mobile vision applications[J]. ArXiv Preprint arXiv: 1704.04861, 2017.
[43] WANG Z W, SHE Q, SMOLIC A. ACTION - Net: multipath excitation for action recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13214 - 13223.
[44] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6450 - 6459.

(上接第 192 页)

[28] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf: representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99 - 106.
[29] BARRON J T, MILDENHALL B, VERBIN D, et al. Mip-nerf 360: unbounded anti-aliased neural radiance fields[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5470 - 5479.
[30] FRIDOVICH-KEIL S, YU A, TANCIK M, et al. Plenoxels: radiance fields without neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5501 - 5510.
[31] REMATAS K, LIU A, SRINIVASAN P P, et al. Urban radiance fields[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 12932 - 12942.
[32] LI T, SLAVCHEVA M, ZOLLHOEFER M, et al. Neural 3D video synthesis from multi-view video[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5521 - 5531.
[33] YUAN Y J, SUN Y T, LAI Y K, et al. Nerf-editing: geometry editing of neural radiance fields[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 18353 - 18364.
[34] ZHU F, GUO S, SONG L, et al. Deep review and analysis of recent NeRFs[J]. APSIPA Transactions on Signal and Information Processing, 2023, 12(1):1 - 32.
[35] SCHONBERGER J L, FRAHM J M. Structure-from-motion revisited[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4104 - 4113.
[36] MÜLLER T, EVANS A, SCHIED C, et al. Instant neural graphics primitives with a multiresolution hash encoding[J]. ACM Transactions on Graphics (ToG), 2022, 41(4): 1 - 15.
[37] RUSU R B, COUSINS S. 3D is here: point cloud library[C]// 2011 IEEE International Conference on Robotics and Automation. IEEE, 2011: 1 - 4.
[38] ZHANG W, WU S, WEN W, et al. Three-dimensional branch segmentation and phenotype extraction of maize tassel based on deep learning[J]. Plant Methods, 2023, 19(1): 76.
[39] 苗艳龙, 彭程, 高阳, 等. 基于地基激光雷达的玉米株高与茎粗自动测量研究[J]. 农业机械学报, 2021, 52(增刊): 43 - 50.
MIAO Yanlong, PENG Cheng, GAO Yang, et al. Automatic measurement of plant height and stem thickness of maize based on terrestrial laser scanning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(Supp.): 43 - 50. (in Chinese)
[40] MARTIN-BRUALLA R, RADWAN N, SAJJADI M S M, et al. Nerf in the wild: neural radiance fields for unconstrained photo collections[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 7210 - 7219.
[41] SRINIVASAN P P, DENG B, ZHANG X, et al. Nerv: neural reflectance and visibility fields for relighting and view synthesis [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 7495 - 7504.
[42] ZHI S, LAIDLOW T, LEUTENEGGER S, et al. In-place scene labelling and understanding with implicit scene representation [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 15838 - 15847.