

三维自编码器及其训练方法、电子设备及存储介质

本申请要求申请日为 2022 年 7 月 21 日的中国专利申请 CN202210864352.6 的优先权。本申请引用上述中国专利申请的全文。

技术领域

本发明涉及人工智能技术领域，尤其涉及一种三维自编码器及其训练方法、三维视觉模块的训练方法、电子设备及存储介质。

背景技术

现有基于深度学习神经网络的动作识别技术，通常都是基于监督学习的方法进行训练，监督学习即所有数据都有标签或真值，直接对网络输出结果和标签计算损失函数进行训练。这种方式训练集都相对较小，覆盖的场景比较局限；且训练好的模型只能适用于训练数据集相关的场景，不能方便的移植到其他场景，并且需要有标注的数据集进行训练，正确的标注本身需要大量的人力成本和算力成本，提取的特征也依赖于标签而不是数据本身的特征。而自监督学习则是通过自身生成标签进行监督训练，自身生成标签的方式通常包括旋转、裁剪、涂色、对比学习等，然后根据下游任务进行 fine-tuning（微调训练）。

动作识别如跌倒识别、物体搬运识别等，通常涉及到人体与周边场景物体的交互，通过感知到的 3D 空间位置结构，可以更准确的判断动作类型而不受摄像头所处的空间位置的影响，比如通过空间位置信息比较准确的推断出一个人究竟是躺坐在椅子上还是跌坐在地上。可感知 3D 结构的 3D GANs（Generative adversarial networks，生成对抗网络）和普通的 2D GANs 不同，其基于以下两个基础组件：

1. 在生成网络中引入能感知 3D 结构的偏执归纳(inductive bias)。

2. 一个神经网络渲染引擎，用于提供视角一致的图像结果。

其中，可感知 3D 结构的偏执归纳一般可以用显式的体素格子(voxel grid)建模或者可以用隐式的神经网络建模。

Vision Transformers (ViT) 是一种基于 transformer 模块的视觉模块，可以进行机器视觉相关的任务。对比通用的进行机器视觉任务的卷积神经网络模块（比如 ResNet)在增大数据集的训练数据数量以后会达到饱和，例如在训练集的体量达到 1 亿张图片以后达到饱和，继续添加训练数据也不会再提高识别的正确率，而 ViT 在训练数据量 1 亿以上识别准确率也还会继续提升，适合作为大数据量预训练的基础模型。Video Vision Transformers(ViViT) 是一种基于 transformer 模块的视觉模块，可以进行机器视觉相关的任务。类似 ViT 使用图像小块进行识别和 token 划分，ViViT 使用视频小块进行识别任务和 token（令牌）划分。利用 MAE (Masked Auto Encoders, 掩码自动编码器) 模式来训练 Vision Transformers,可以通过自监督学习(Self-Supervised Learning)的方式进行训练，训练出来的 Vision Transformers 可以达到很高的准确率。类似地，3DViViT 则是基于 transformer 模块的三维视觉模块，即可以对图像的三维特征进行感知。

发明内容

本发明要解决的技术问题是为了克服现有技术中的上述缺陷，提供一种三维自编码器及其训练方法、三维视觉模块的训练方法、电子设备及存储介质。

本发明是通过下述技术方案来解决上述技术问题：

本发明提供了一种三维自编码器，包括编码器和解码器；所述编码器用于提取输入的待处理图片的空间特征参数并输出至所述解码器；所述解码器用于根据所述空间特征参数输出包括所述待处理图片深度信息的目标图片，所述待处理图片和所述目标图片用于确定所述三维自编码器的损失函数。

较佳地，所述解码器包括三维特征提取模块和上采样模块；所述三维特征提取模块用于根据所述空间特征参数获取所述待处理图片的深度信息；所述上采样模块用于根据所述深度信息和经过所述三维特征提取模块处理后的所述空间特征参数生成所述目标图片。

较佳地，所述三维特征提取模块包括映射网络、二维生成对抗网络生成器、神经网络渲染器；所述空间特征参数包括摄像头参数和隐藏标量；

所述映射网络用于处理所述隐藏标量，以和所述待处理图片的规格对应，并将所述处理后的隐藏标量输出至所述二维生成对抗网络生成器和所述上采样模块；

所述二维生成对抗网络生成器用于根据所述处理后的空间特征参数生成特征变形后的特征图，并输入所述神经网络渲染器；

所述神经网络渲染器用于处理所述特征变形后的特征图和所述摄像头参数，以输出所述待处理图片的深度信息。

较佳地，所述神经网络渲染器包括三平面解码器和体素渲染模块；

所述三平面解码器用于处理所述特征变形后的特征图以输出体素信息至所述体素渲染模块；

所述体素渲染模块用于根据所述体素信息和所述摄像头参数输出所述待处理图片的深度信息。

较佳地，所述三平面解码器包括三个依次连接的全连接层；所述映射网络包括三个依次连接的全连接层。

较佳地，所述编码器包括下采样卷积神经网络；所述下采样卷积神经网络用于对所述待处理图片进行下采样处理。

较佳地，所述编码器还包括全连接层；所述下采样卷积神经网络包括卷积层和池化层；所述卷积层、所述池化层和所述全连接层依次连接；所述卷积层基于注意力机制+ReLU 激活函数实现。

较佳地，所述三维特征提取模块基于三维生成对抗网络的特征提取模块

获得，所述特征提取模块用于提取所述待处理图片的三维场景信息。

本发明还提供了一种三维自编码器的训练方法，用于训练上述的三维自编码器；所述训练方法包括：

锁定所述三维自编码器的解码器的参数，采用第一训练数据集训练所述三维自编码器的编码器的参数直至收敛；其中，所述第一训练数据集用于训练获得所述三维生成对抗网络

打开所述编码器的参数，继续训练所述编码器和所述解码器的参数直至收敛。

本发明还提供了一种三维视觉模块的训练方法，包括步骤：

利用上述的三维自编码器处理视频帧，获得所述视频帧对应的三维空间特征；

将关联了所述三维空间特征的所述视频帧分成若干视频小块；

分别为所述若干视频小块分配位置向量和时间向量；

根据时间向量随机筛选过滤掉第一预设比例的视频帧；

在剩余的视频帧中随机选择第二预设比例的视频小块作为训练数据，输入所述三维视觉模块的编码器，并将输出的视频小块加上所述筛选过滤掉的第一预设比例的视频帧对应的令牌输入所述三维视觉模块的解码器；其中，所述视频小块按照所述时间向量和所述位置向量排列顺序；

基于所述三维视觉模块的解码器的输出数据与所述视频帧训练所述三维视觉模块直至收敛。

本发明还提供了一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现上述的三维视觉模块的训练方法。

本发明还提供了一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现上述的三维视觉模块的训练方法。

本发明的积极进步效果在于：通过提供三维自编码器及其训练方法、三

维视觉模块的训练方法、电子设备及存储介质，能够通过自监督方式实现对图像的三维场景特征的感知；且能够采用海量训练集，覆盖较多的场景，不需要有标注的数据集进行训练，可以利用大量的未标注原始数据进行训练，节省人工和成本，训练好的模型还能方便地移植到其他场景。通过加入三维空间关系的归纳偏置的神经网络设计可以感知到动作中包含的与环境的交互信息，提高识别的准确率。

附图说明

图 1 为本发明实施例 1 的三维自编码器的结构示意图。

图 2 为本发明实施例 1 的三维自编码器的解码器的结构示意图。

图 3 为本发明实施例 1 的三维自编码器的编码器的结构示意图。

图 4 为本发明实施例 2 的三维自编码器的训练方法的流程图。

图 5 为本发明实施例 3 的三维视觉模块的训练方法的流程图。

图 6 为本发明实施例 3 中视频帧三维场景特征感知的处理流程示意图。

图 7 为本发明实施例 3 中的三维视觉模块的编码器的结构示意图。

图 8 为本发明实施例 3 中的三维视觉模块的解码器的结构示意图。

图 9 为本发明实施例 4 的电子设备的结构框图。

具体实施方式

下面通过实施例的方式进一步说明本发明，但并不因此将本发明限制在所述的实施例范围之中。

实施例 1

参加图 1-3 所示，本实施例具体提供了一种三维自编码器，包括编码器和解码器；编码器用于提取输入的待处理图片的空间特征参数并输出至解码器；解码器用于根据空间特征参数输出包括待处理图片深度信息的目标图片，待处理图片和目标图片用于确定三维自编码器的损失函数。

本实施例提供了基于编码器-解码器（encoder-decoder）模式构建的三维自编码器，如图 1，其包括一个编码器和一个解码器。作为较佳的实施方式，三维特征提取模块基于三维生成对抗网络的特征提取模块获得，特征提取模块用于提取待处理图片的三维场景信息。三维生成对抗网络是一种可感知 3D 结构的 GANs，其实现基于在生成网络中引入能感知 3D 结构的偏执归纳，以及一个神经网络渲染引擎能够提供视角一致的图像结果，可以从随机特征输出对应 3D 结构的特征图。可感知 3D 结构的偏执归纳一般可以用显式的体素格子建模或者可以用隐式的神经网络建模。

作为较佳的实施方式，参见图 2，解码器包括三维特征提取模块和上采样模块；三维特征提取模块用于根据空间特征参数获取待处理图片的深度信息；上采样模块用于根据深度信息和经过三维特征提取模块处理后的空间特征参数生成目标图片。上采样即放大图像，让图像变成更高分辨率的技术，通常通过反卷积、上池化方法、双线性插值等实现。本实施例中通过上采样模块的处理，将图像放大到与待处理图像规格相同，从而可以进一步比对进行损失函数的计算。

较佳地，三维特征提取模块包括映射网络、二维生成对抗网络生成器、神经网络渲染器；空间特征参数包括摄像头参数和隐藏标量；映射网络用于处理隐藏标量，以和待处理图片的规格对应，并将处理后的隐藏标量输出至二维生成对抗网络生成器和上采样模块；二维生成对抗网络生成器用于根据处理后的空间特征参数生成特征变形后的特征图，并输入神经网络渲染器；神经网络渲染器用于处理特征变形后的特征图和摄像头参数，以输出待处理图片的深度信息。

神经网络渲染器包括三平面解码器和体素渲染模块；三平面解码器用于处理特征变形后的特征图以输出体素信息至体素渲染模块；体素渲染模块用于根据体素信息和摄像头参数输出待处理图片的深度信息。可选地，三平面解码器包括三个依次连接的全连接层；映射网络同样也可以包括三个依次连

接的全连接层。

作为较佳的实施方式，参见图 3，编码器包括下采样卷积神经网络；下采样卷积神经网络用于对待处理图片进行下采样处理。下采样即缩小图像，主要目的是为了使得图像符合显示区域的大小，生成对应图像的缩略图。通常通过卷积神经网络中的池化层或卷积层实现，区别在于卷积过程导致的图像变小是为了提取特征，而池化下采样是为了降低特征的维度。较佳地，编码器还包括全连接层；下采样卷积神经网络包括卷积层和池化层；卷积层、池化层和全连接层依次连接；卷积层基于注意力机制+ReLU 激活函数实现。

本实施例的三维自编码器充分利用 3DGANs 的场景特征感知优势对图像进行处理，从而获取包含场景特征的视频帧可用于后续过自监督方式的实现；能够采用海量训练集，覆盖较多的场景，不需要有标注的数据集进行训练，可以利用大量的未标注原始数据进行训练，节省人工和成本，训练好的模型还能方便地移植到其他场景。通过加入三维空间关系的归纳偏置的神经网络设计可以感知到动作中包含的与环境的交互信息，提高识别的准确率。

实施例 2

参加图 4 所示，本实施例具体提供了一种三维自编码器的训练方法，用于训练实施例 1 中的三维自编码器；训练方法包括：

S0.采用第一训练数据集（例如，7 万张 RGB 图片）对三维生成对抗网络（3DGANs）进行训练，并提取训练后的 3DGANs 的生成器作为三维自编码器的解码器。

S1.锁定三维自编码器的解码器的参数，采用第一训练数据集训练三维自编码器的编码器的参数直至收敛；其中，第一训练数据集用于训练获得三维生成对抗网络。

S2.打开编码器的参数，继续训练编码器和解码器的参数直至收敛。

其中第一数据训练集可以采用 7 万张以上的三原色（RGB）图片，其同样被用于训练三维自编码器。

本实施例的三维自编码器的训练方法针对三维自编码器的构成原理设计了训练方法，即首先训练编码器，再对三维自编码器进行整体训练，从而能够获得符合要求的三维自编码器，用于利用 3DGANs 的场景特征感知优势对图像进行处理，从而获取包含场景特征的视频帧可用于后续过自监督方式的实现；能够采用海量训练集，覆盖较多的场景，不需要有标注的数据集进行训练，可以利用大量的未标注原始数据进行训练，节省人工和成本，训练好的模型还能方便地移植到其他场景。通过加入三维空间关系的归纳偏置的神经网络设计可以感知到动作中包含的与环境的交互信息，提高识别的准确率。

实施例 3

参加图 5 所示，本实施例具体提供了一种三维视觉模块的训练方法，包括步骤：

S101.利用上述三维自编码器处理视频帧，获得视频帧对应的三维空间特征；

S102.将关联了三维空间特征的视频帧分成若干视频小块；

S103.分别为若干视频小块分配位置向量和时间向量；

S104.根据时间向量随机筛选过滤掉第一预设比例的视频帧；

S105.在剩余的视频帧中随机选择第二预设比例的视频小块作为训练数据，输入三维视觉模块的编码器，并将输出的视频小块加上筛选过滤掉的第一预设比例的视频帧对应的令牌输入三维视觉模块的解码器；其中，视频小块按照时间向量和位置向量排列顺序；

S106.基于三维视觉模块的解码器的输出数据与视频帧训练三维视觉模块直至收敛。

本实施例中，训练数据集可以采用 400 万个 10 秒级的短视频文件来进行。基于上述步骤，本领域技术人员可知，可以通过采用如图 6 所示的处理流程来对视频帧图像进行三维场景感知以提高动作识别的准确率。具体地，

训练三维对抗生成网络，取其特征提取模块作为三维自编码器的解码器，训练该三维自编码器以处理视频帧，并用于对三维视觉模块 3DViViT 的预训练。

本实施例的三维视觉模块的编码器和解码器分别如图 7 和图 8 所示，其中，三维视频块(3DTubelet)转换为 tokens(令牌)，令牌嵌入(Token Embedding)分别结合位置特征并进行空间编码转换，以及结合时间特征进行时间编码转换后输出隐式表达视频块 (Tubelet Hidden Representation)。

可以理解，三维视觉模块处理预训练后，可以利用其他小数据集进行微调 (fine-tuning) 训练。fine-tuning 即冻结预训练模型的部分卷积层（通常是靠近输入的多数卷积层）甚至不冻结任何网络层来训练剩下的卷积层（通常是靠近输出的部分卷积层）和全连接层，从而利用已知的网络结构和已知的网络参数，微调最后一层前的若干层的参数，有效利用深度神经网络的泛化能力的同时也可免去设计复杂的模型以及耗时良久的训练。

本实施例的三维视觉模块的训练方法能够通过自监督方式实现对图像的三维场景特征的感知；且能够采用海量训练集，覆盖较多的场景，不需要有标注的数据集进行训练，可以利用大量的未标注原始数据进行训练，节省人工和成本，训练好的模型还能方便地移植到其他场景。通过加入三维空间关系的归纳偏置的神经网络设计可以感知到动作中包含的与环境的交互信息，提高识别的准确率。

实施例 4

参见图 9 所示，本实施例提供了一种电子设备 30，包括处理器 31、存储器 32 及存储在存储器 32 上并可在处理器 31 上运行的计算机程序，处理器 31 执行程序时实现实施例 2 中的三维自编码器的训练方法和/或实施例 3 中的三维视觉模块的训练方法。图 9 显示的电子设备 30 仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。

电子设备 30 可以以通用计算设备的形式表现，例如其可以为服务器设

备。电子设备 30 的组件可以包括但不限于：上述至少一个处理器 31、上述至少一个存储器 32、连接不同系统组件(包括存储器 32 和处理器 31)的总线 33。

总线 33 包括数据总线、地址总线和控制总线。

存储器 32 可以包括易失性存储器，例如随机存取存储器(RAM)321 和/或高速缓存存储器 322，还可以进一步包括只读存储器(ROM)323。

存储器 32 还可以包括具有一组(至少一个)程序模块 324 的程序/实用工具 325，这样的程序模块 324 包括但不限于：操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。

处理器 31 通过运行存储在存储器 32 中的计算机程序，从而执行各种功能应用以及数据处理，例如本发明实施例 2 中的三维自编码器的训练方法和/或实施例 3 中的三维视觉模块的训练方法。

电子设备 30 也可以与一个或多个外部设备 34(例如键盘、指向设备等)通信。这种通信可以通过输入/输出(I/O)接口 35 进行。并且，模型生成的设备 30 还可以通过网络适配器 36 与一个或者多个网络(例如局域网(LAN)，广域网(WAN)和/或公共网络，例如因特网)通信。网络适配器 36 通过总线 33 与模型生成的设备 30 的其它模块通信。可以结合模型生成的设备 30 使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理器、外部磁盘驱动阵列、RAID(磁盘阵列)系统、磁带驱动器以及数据备份存储系统等。

应当注意，尽管在上文详细描述中提及了电子设备的若干单元/模块或子单元/模块，但是这种划分仅仅是示例性的并非强制性的。实际上，根据本发明的实施方式，上文描述的两个或更多单元/模块的特征和功能可以在一个单元/模块中具体化。反之，上文描述的一个单元/模块的特征和功能可以进一步划分为由多个单元/模块来具体化。

实施例 5

本实施例提供了一种计算机可读存储介质，其上存储有计算机程序，程序被处理器执行时实现实施例 2 中的三维自编码器的训练方法和/或实施例 3 中的三维视觉模块的训练方法。

其中，可读存储介质可以采用的更具体可以包括但不限于：便携式盘、硬盘、随机存取存储器、只读存储器、可擦拭可编程只读存储器、光存储器件、磁存储器件或上述的任意合适的组合。

在可能的实施方式中，本发明还可以实现为一种程序产品的形式，其包括程序代码，当程序产品在终端设备上运行时，程序代码用于使终端设备执行实现实施例 2 中的三维自编码器的训练方法和/或实施例 3 中的三维视觉模块的训练方法。

其中，可以以一种或多种程序设计语言的任意组合来编写用于执行本发明的程序代码，程序代码可以完全地在用户设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户设备上部分在远程设备上执行或完全在远程设备上执行。

虽然以上描述了本发明的具体实施方式，但是本领域的技术人员应当理解，这仅是举例说明，本发明的保护范围是由所附权利要求书限定的。本领域的技术人员在不背离本发明的原理和实质的前提下，可以对这些实施方式做出多种变更或修改，但这些变更和修改均落入本发明的保护范围。

权利要求

1. 一种三维自编码器，其特征在于，包括编码器和解码器；所述编码器用于提取输入的待处理图片的空间特征参数并输出至所述解码器；所述解码器用于根据所述空间特征参数输出包括所述待处理图片深度信息的目标图片，所述待处理图片和所述目标图片用于确定所述三维自编码器的损失函数。

2. 如权利要求 1 所述的三维自编码器，其特征在于，所述解码器包括三维特征提取模块和上采样模块；所述三维特征提取模块用于根据所述空间特征参数获取所述待处理图片的深度信息；所述上采样模块用于根据所述深度信息和经过所述三维特征提取模块处理后的所述空间特征参数生成所述目标图片。

3. 如权利要求 2 所述的三维自编码器，其特征在于，所述三维特征提取模块包括映射网络、二维生成对抗网络生成器、神经网络渲染器；所述空间特征参数包括摄像头参数和隐藏标量；

所述映射网络用于处理所述隐藏标量，以和所述待处理图片的规格对应，并将所述处理后的隐藏标量输出至所述二维生成对抗网络生成器和所述上采样模块；

所述二维生成对抗网络生成器用于根据所述处理后的空间特征参数生成特征变形后的特征图，并输入所述神经网络渲染器；

所述神经网络渲染器用于处理所述特征变形后的特征图和所述摄像头参数，以输出所述待处理图片的深度信息。

4. 如权利要求 3 所述的三维自编码器，其特征在于，所述神经网络渲染器包括三平面解码器和体素渲染模块；

所述三平面解码器用于处理所述特征变形后的特征图以输出体素信息至所述体素渲染模块；

所述体素渲染模块用于根据所述体素信息和所述摄像头参数输出所述

待处理图片的深度信息。

5.如权利要求4所述的三维自编码器，其特征在于，所述三平面解码器包括三个依次连接的全连接层；所述映射网络包括三个依次连接的全连接层。

6.如权利要求2所述的三维自编码器，其特征在于，所述编码器包括下采样卷积神经网络；所述下采样卷积神经网络用于对所述待处理图片进行下采样处理。

7.如权利要求6所述的三维自编码器，其特征在于，所述编码器还包括全连接层；所述下采样卷积神经网络包括卷积层和池化层；所述卷积层、所述池化层和所述全连接层依次连接；所述卷积层基于注意力机制+ReLU激活函数实现。

8.如权利要求2-7中任一项所述的三维自编码器，其特征在于，所述三维特征提取模块基于三维生成对抗网络的特征提取模块获得，所述特征提取模块用于提取所述待处理图片的三维场景信息。

9.一种三维自编码器的训练方法，其特征在于，用于训练权利要求8所述的三维自编码器；所述训练方法包括：

锁定所述三维自编码器的解码器的参数，采用第一训练数据集训练所述三维自编码器的编码器的参数直至收敛；其中，所述第一训练数据集用于训练获得所述三维生成对抗网络

打开所述编码器的参数，继续训练所述编码器和所述解码器的参数直至收敛。

10.一种三维视觉模块的训练方法，其特征在于，包括步骤：

利用权利要求8所述的三维自编码器处理视频帧，获得所述视频帧对应的三维空间特征；

将关联了所述三维空间特征的所述视频帧分成若干视频小块；

分别为所述若干视频小块分配位置向量和时间向量；

根据时间向量随机筛选过滤掉第一预设比例的视频帧；

在剩余的视频帧中随机选择第二预设比例的视频小块作为训练数据，输入所述三维视觉模块的编码器，并将输出的视频小块加上所述筛选过滤掉的第一预设比例的视频帧对应的令牌输入所述三维视觉模块的解码器；其中，所述视频小块按照所述时间向量和所述位置向量排列顺序；

基于所述三维视觉模块的解码器的输出数据与所述视频帧训练所述三维视觉模块直至收敛。

11.一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现权利要求9所述的三维自编码器的训练方法和/或权利要求10所述的三维视觉模块的训练方法。

12.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求9所述的三维自编码器的训练方法和/或权利要求10所述的三维视觉模块的训练方法。