(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 120017841 A
(43)申请公布日 2025.05.16

(21)申请号 202411601081.0

(22)申请日 2024.11.11

(30)优先权数据
63/599,972 2023.11.16 US
18/920,734 2024.10.18 US

(71)申请人 迪士尼企业公司
地址 美国加利福尼亚州
申请人 苏黎世联邦理工学院

(72)发明人 R·G·D·A·阿泽维多
C·R·施罗尔斯 S·拉布罗齐
J·E·赛特尔 薛远翊

(74)专利代理机构 深圳市百瑞专利商标事务所(普通合伙) 44240
专利代理师 金辉

(51)Int.Cl.
H04N 19/172(2014.01)
H04N 19/42(2014.01)
G06T 9/00(2006.01)
H04N 19/60(2014.01)
H04N 19/587(2014.01)
H04N 19/59(2014.01)

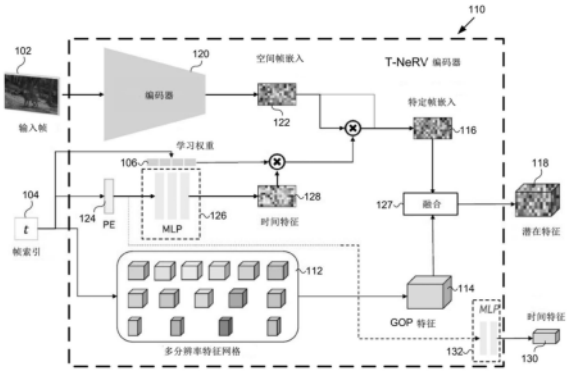
权利要求书2页 说明书8页 附图3页

(54)发明名称

可调混合神经视频表示

(57)摘要

一种系统包括基于可调神经网络的视频编码器,该视频编码器配置为接收包括多个视频帧的视频序列,生成多个视频帧的第一视频帧的特定帧嵌入,以及识别多个视频帧的子集的一个或多个图片组(GOP)特征,该子集包括第一视频帧。基于可调神经网络的视频编码器还配置为将第一视频帧的特定帧嵌入和多个视频帧的第一多个视频帧的一个或多个GOP特征进行组合,以提供对应于第一视频帧的压缩版本的潜在特征。



CN 120017841 A

1.一种系统,其包括:

基于可调神经网络的视频编码器,其配置为:

接收包括多个视频帧的视频序列;

生成多个视频帧中的第一视频帧的特定帧嵌入;

识别多个视频帧的第一多个视频帧的一个或多个图片组(GOP)特征,多个视频帧中的第一多个包括第一视频帧;和

组合第一视频帧的特定帧嵌入和多个视频帧的第一多个视频帧的一个或多个GOP特征,以提供对应于第一视频帧的压缩版本的潜在特征。

2.根据权利要求1所述的系统,其中使用基于可调神经网络的视频编码器的多分辨率特征网格来识别多个视频帧的第一多个视频帧的一个或多个GOP特征。

3.根据权利要求1所述的系统,其中基于可调神经网络的视频编码器配置为:

从多个视频帧的第一多个视频帧中提取多个视频帧的第一多个视频帧的一个或多个时间特征;

其中基于与一个或多个时间特征的组合来生成特定帧嵌入。

4.根据权利要求1所述的系统,其中所述潜在特征包括特定帧嵌入和一个或多个GOP特征的加权组合。

5.根据权利要求4所述的系统,其中应用于特定帧嵌入的第一权重或应用于一个或多个GOP特征的第二权重中的至少一个由基于可调神经网络的视频编码器确定。

6.根据权利要求1所述的系统,其中,使用包括失真损失和熵损失的优化损失函数来训练所述基于可调神经网络的视频编码器。

7.根据权利要求6所述的系统,其中所述失真损失和所述熵损失被联合优化。

8.根据权利要求1所述的系统,其中所述第一视频帧包括多个视频帧的第一多个视频帧的子序列,并且其中特定帧嵌入包括分别对应于多个视频帧的第一多个视频帧的每个子序列的多个特定帧嵌入,并且潜在特征是多个潜在特征之一,每个潜在特征分别对应于多个视频帧的第一多个视频帧的每个子序列的压缩版本。

9.一种系统,其包括:

基于神经网络的视频解码器,其配置为:

接收对应于包括在视频序列中的多个视频帧的第一多个视频帧的第一视频帧的压缩版本的潜在特征,所述潜在特征是第一视频帧的特定帧嵌入与多个视频帧的第一多个视频帧的一个或多个图片组(GOP)特征的组合;和

解码潜在特征以提供对应于第一视频帧的未压缩视频帧。

10.根据权利要求9所述的系统,其中,所述特定帧嵌入是基于与从所述多个视频帧的第一多个视频帧中提取的一个或多个时间特征的组合而生成的。

11.根据权利要求9所述的系统,其中所述潜在特征包括特定帧嵌入和一个或多个GOP特征的加权组合。

12.根据权利要求9所述的系统,其中第一视频帧包括多个视频帧的第一多个视频帧的子序列,并且其中特定帧嵌入包括分别对应于多个视频帧的第一多个视频帧的每个子序列的多个特定帧嵌入,并且潜在特征是多个潜在特征之一,每个潜在特征分别对应于多个视频帧的第一多个视频帧的每个子序列的压缩版本。

13.一种由基于可调神经网络的视频编码器执行的方法,所述方法包括:

接收包括多个视频帧的视频序列;

生成多个视频帧中的第一视频帧的特定帧嵌入;

识别多个视频帧的第一多个视频帧的一个或多个图片组(GOP)特征,多个视频帧的第一多个视频帧包括第一视频帧;和

组合第一视频帧的特定帧嵌入和多个视频帧的第一多个视频帧的一个或多个GOP特征,以提供对应于第一视频帧的压缩版本的潜在特征。

14.根据权利要求13所述的方法,其中使用所述基于可调神经网络的视频编码器的多分辨率特征网格来识别所述多个视频帧的第一多个视频帧的一个或多个GOP特征。

15.根据权利要求13所述的方法,进一步包括:

从多个视频帧的第一多个视频帧中提取多个视频帧的第一多个视频帧的一个或多个时间特征;

其中基于与一个或多个时间特征的组合来生成特定帧嵌入。

16.根据权利要求13所述的方法,其中所述潜在特征包括特定帧嵌入和一个或多个GOP特征的加权组合。

17.根据权利要求16所述的方法,进一步包括:

确定应用于特定帧嵌入的第一权重或应用于一个或多个GOP特征的第二权重中的至少一个。

18.根据权利要求13所述的方法,其中,使用包括失真损失和熵损失的优化损失函数来训练所述基于可调神经网络的视频编码器。

19.根据权利要求18所述的方法,其中失真损失和熵损失被联合优化。

20.根据权利要求13所述的方法,其中第一视频帧包括多个视频帧的第一多个视频帧的子序列,并且其中特定帧嵌入包括分别对应于多个视频帧的第一多个视频帧的每个子序列的多个特定帧嵌入,并且潜在特征是多个潜在特征之一,每个潜在特征分别对应于多个视频帧的第一多个视频帧的每个子序列的压缩版本。

可调混合神经视频表示

[0001] 相关申请

[0002] 本申请要求2023年11月16日提交的,申请序列号为63/599,972,名称为“Tunable Hybrid Neural Video Representations”的未决美国临时专利申请的权益和优先权,该申请通过引用完全结合到本申请中。

背景技术

[0003] 视频压缩是一个由来已久且困难的问题,因此激发了许多研究。视频压缩的主要目标是使用最少的存储量来表示数字视频(通常是帧序列,每个帧由二维(2-D)像素阵列、RGB或YUV颜色来表示),同时使质量损失最小化。尽管近几十年来传统视频编解码器取得了许多进步,但深度学习的出现激发出了许多超越传统视频编解码器的新方法。

[0004] 例如,隐式神经表示(INRs)已经引起了广泛的研究兴趣,并且已经被应用于包括视频压缩等各种领域。除了表现出快速解码和时间插值能力等理想特性之外,基于INR的方法在压缩性能上可以匹配或超过传统的标准视频编解码器,如高级视频编码(AVC,也称为H.264)和高效视频编码(HEVC)。然而,这些利用INR进行视频压缩的现有方法仅在有限且有时高度受限的设置中表现良好,例如受限于特定的模型大小、固定的纵横比和相对静态的视频序列。

[0005] 作为背景,早期的视频INR采用像素方式表示,将像素索引映射到RGB颜色,但性能有限,解码速度差。然而,也提出了一种基于逐帧的表示方法,即视频的神经表示(NeRV)。在NeRV中,多层感知器(MLP)从经过位置编码的帧索引生成时间特征,该索引经过重塑和上采样处理。NeRV实现了实时解码速度和比以前的视频INR明显更好的重建。受生成对抗网络(GANs)工作的启发,开发了加速神经网络(E-NeRV),其中传统NeRV中使用的特征提取器被分解为时间和空间上下文,并通过变换器网络进行融合,进一步通过自适应实例规范化(AdaIN)层将时间信息注入解码器块。

[0006] 流引导的逐帧神经网络(FFNeRV)通过使用解码器增强时间一致性来提高动态序列的性能,该解码器预测独立帧和一组光流图,这些光流图用于扭曲相邻的独立帧并将独立帧组合起来以提供最终的输出帧。混合表示HNeRV在训练期间类似于自动编码器,使用编码器来提取特定内容嵌入。在训练结束后,视频由解码器的参数和每帧嵌入来表示。虽然前面确定的NeRV的现有变体,即E-NeRV、FFNeRV和HNeRV,代表了对原始NeRV的改进,但是它们的有用性通常局限于特定的设置,例如在静态序列上变现良好而在高运动内容上不稳定,或者反之亦然。传统的混合方法在低比特率领域变现出潜力,但是不能扩展到更高的比特率,并且目前局限于具有不寻常的2:1纵横比的视频。因此,此项技术中需要一种基于神经网络的视频压缩解决方案,其能够在一系列使用情况和参数下有效地编码视频帧。

附图说明

[0007] 图1A示出了描绘根据一个实施方式的视频压缩系统的视频可调混合隐式神经表示(T-NeRV)编码器的示例性架构的图;

[0008] 图1B示出了描绘根据一个实施方式的适用于图1A所示的示例性T-NeRV编码器的基于光流的T-NeRV解码器的示例性架构的图;

[0009] 图2示意性地描绘了根据一个实施方式的适用于图1B所示的T-NeRV解码器的一系列T-NeRV块;和

[0010] 图3是概述根据一个实施方式的用于执行可调混合神经视频表示的方法的流程图。

具体实施方式

[0011] 以下描述包含与本公开中的实现相关的特定信息。本领域技术人员将认识到,本公开可以以不同于此部分具体讨论的方式来实现。本申请中的附图及其附带的详细描述仅针对示例性实施方式。除非另有说明,否则附图中相同或相应的元件可以用相同或相应的附图标记来表示。此外,本申请中的附图和图示通常不按比例绘制,并且不旨在对应于实际的相对尺寸。

[0012] 本申请通过使用用于视频的可调混合隐式神经表示(INR)(以下称为“T-NeRV”)解决方案来解决和克服传统技术中的缺陷,该解决方案基于对传统编解码器(例如高级视频编码(AVC,也称为H.264)和高效编码(HEVC))需要本地和非本地信息来有效地对帧进行编码的认识,对现有技术进行了改进并发展了现有技术。此外,本文公开的可调混合神经视频表示解决方案可以有利地实现为自动化或基本自动化的系统和方法。

[0013] 如在本申请中所使用的,术语“自动化”、“自动化的(automated)”和“自动化的(automating)”指的是不需要人类用户(例如人类系统操作员或管理员)参与的系统和过程。尽管在一些实施方式中,本文公开的系统和方法的性能可以由人类系统操作员监控,但是人类参与是可选的。因此,本申请中描述的方法可以在所公开系统的硬件处理组件的控制下执行。

[0014] 在本申请中公开的新颖和创造性的T-NeRV解决方案实现了一种视频压缩系统,该系统包括基于可调神经网络的视频编码器,该视频编码器将特定帧嵌入与特定图像组(特定GOP)的特征相结合,并且利用基于光流的解码器对它们进行上采样。特定帧和特定GOP的特征的这种新颖和创造性的组合为特定内容微调提供了杠杆。例如,更大的帧嵌入,即强调特定帧的特征,可以提取更多的高频信息,从而提高对低运动序列的性能。相反,更显著的特定GOP特征允许T-NeRV模型提取更多的时间上下文,解码器可以利用这些时间上下文来重建动态序列中的帧。对于压缩,本申请中公开的T-NeRV解决方案通过量化感知和熵约束训练来捕获嵌入,从而联合最小化速率和失真。通过对所有嵌入采用单个熵模型,本T-NeRV解决方案相比于在先的方法可以更高程度地利用冗余。端到端训练还通过在特定帧的嵌入或特定GOP的特征上花费更多的比特,自动调整调谐杆,来强制当前的T-NeRV网络将其自身微调到目标视频内容。

[0015] 本发明的T-NeRV解决方案对现有技术水平至少贡献了以下特征:(i)可调混合视频INR,其将特定帧的嵌入与特定GOP的特征相结合,从而为特定内容微调提供了杠杆;以及(ii)扩展信息论INR压缩框架以包括嵌入,从而利用其中的显著冗余。注意,在T-NeRV模型的训练和优化期间,自动执行模型的调整,以及特定帧的嵌入与特定GOP的特征的组合的平衡,尽管系统用户可以通过限定所使用的特征图的最大尺寸来断言一些控制。还应注意,当

在UVG数据集上评估时,本申请公开的T-NeRV解决方案在视频表示和视频压缩任务上都优于所有在先的视频INR。

[0016] 图1A示出了描绘根据本概念的一个实施方式的视频压缩系统的T-NeRV编码器110的示例性架构的图,而图1B示出了描绘根据一个实施方式的适用于示例性T-NeRV编码器110的基于光流的T-NeRV解码器140的示例性架构的图。T-NeRV编码器110配置为根据基本事实帧102(即,$F_t \in \mathbb{R}^{3\times H\times W}$;下文中的"输入视频帧102")和归一化帧索引104(t)生成潜在特征118(即,$z_t \in \mathbb{R}^{c_1\times h\times w}$),T-NeRV解码器140根据潜在特征118重建预测帧170(即,$\widehat{F_t}$),如图1B所示。注意,潜在特征118的纵横比与输入视频帧102的纵横比匹配,即$\frac{h}{w}=\frac{H}{W}$。

[0017] 如图1A所示,T-NeRV编码器110包括大编码器120、位置编码器(PE)124、多层感知器(MLP)126、多分辨率特征网格112、融合块127,并且在一些实施方式中还可以包括MLP132。在图1A中还示出了学习权重106、空间帧嵌入122、一个或多个时间特征128(以下称为"时间特征128")、一个或多个特定GOP特征114(以下称为"GOP特征114")、特定帧嵌入116和可选的时间特征输出130。

[0018] 在训练期间,包括T-NeRV编码器110和T-NeRV解码器140的T-NeRV网络执行前向传递和后向传递,如本领域已知的,以优化网络参数。一旦训练结束,只有GOP特征114、特定帧嵌入116(即视频帧的特征图或张量表示)、潜在特征118和可选的时间特征输出130被保留作为视频状态的一部分,而图1A中所示的其余特征被丢弃。因此,解码输入视频帧102相当于从多分辨率特征网格112获得GOP特征114,并使用任一合适的组合技术(例如连接或融合)将它们与特定帧嵌入116融合,以获得输入视频帧的潜在特征118(z_t),该潜在特征通过T-NeRV解码器140传递。

[0019] 特定帧嵌入116($e_t \in \mathbb{R}^{c_E\times h\times w}$)可以作为潜在特征118的两个组成部分之一。特定帧嵌入116可以分两步生成:首先,大编码器120可以从输入视频帧102中提取空间帧嵌入122($\hat{e}_t$),然后可以用附加的时间特征128对其进行扩充。这种扩充是一种数学运算,通过连接空间帧嵌入122和时间特征128,或者可选地,使用任何其他数学技术,例如神经网络或变换器,将它们组合起来,从而将空间帧嵌入122和时间特征128组合起来。因为特定帧嵌入本身被传输到解码器,所以在编码之后可以丢弃生成它们所涉及的所有网络块。结果,大编码器120可以比HNeRV网络中使用的传统编码器大大约100倍,即,包括多100倍的权重,因为编码器权重本身不会被传输到解码器。

[0020] 可以用时间特征128进一步扩充空间帧嵌入122。例如,位置编码器(PE)124后面跟着多层感知器(MLP)126,可以用来提供时间信息,该时间信息然后可以被重塑为匹配空间帧嵌入122($\hat{e}_t$)的维度的时间特征128($s_t \in \mathbb{R}^{c_E\times h\times w}$)。如上所述,时间特征128然后可以与空间帧嵌入122($\hat{e}_t$)组合,以获得特定帧嵌入116,其可以表示为:

[0021] $$e_t = \hat{e}_t \cdot (1 + \alpha_t \cdot s_t)$$ (等式 1)

[0022] 其中α_t表示特定于每个帧的学习权重106,其调制时间特征128(s_t)被包括在特定帧嵌入116中的程度。这使得T-NeRV网络在每帧的基础上决定是否将时间特征128结合到特定帧嵌入116中以及结合到什么程度。注意,用于组合空间帧嵌入122和时间特征128以提供特定帧嵌入116的机制可以认为是对特定帧嵌入116的每帧屏蔽操作,使得相似的帧获得不

同的嵌入。T-NeRV解码器140可以利用该信息来决定从帧自身的独立帧中重建帧的哪些部分,以及哪些部分依赖于帧间信息。

[0023] T-NeRV编码器110利用多分辨率特征网格112来获得GOP特征114($g_t \in \mathbb{R}^{c_G \times h \times w}$)。在前向传递过程中,帧索引104(t)用于索引网格,在每个多分辨率特征网格112的层级中选择两个特征,这些特征以学习得到的数字或张量的形式存在,t位于这些特征之内。可以在这两个特征之间执行双线性内插,并且来自每个级别的结果可以被连接、融合或以其他方式组合以获得GOP特征114。注意,与FFNeRV相反,T-NeRV允许多分辨率特征网格112的不同级别的特征相对于通道c_i的数量在大小上变化,同时保持它们的空间维度(h,w)。这种修改允许T-NeRV有利地调节针对不同GOP大小捕获的信息量。注意,T-NeRV的这种调节是在学习过程中执行的优化期间自动发生的。

[0024] 从融合GOP特征114和特定帧嵌入116获得的潜在特征118可以与由MLP132生成的时间特征输出130($u_t \in \mathbb{R}^{128}$)一起被传递到T-NeRV解码器140。根据图1B所示的示例性实施方式,T-NeRV解码器140可以包括三个部分:预块卷积层142、一系列T-NeRV块150(例如,五个T-NeRV块)和基于流的扭曲块144。注意,预块卷积层142包括具有内核的卷积层,该内核根据潜在特征118的大小和第一个T-NeRV块150所期望的尺寸来执行信道扩展或信道缩减,即本领域中已知的张量重塑操作。图1B中还示出了头部层,即输出层164a和164b、光流图143、独立帧缓冲器146、聚合帧148、独立帧166、聚合块168和对应于图1A中输入视频帧102的预测帧170。

[0025] T-NeRV解码器140可以利用一系列T-NeRV块150来预测独立帧166和光流图143。参考图2,根据一个实施方式,图2示意性地描绘了适用于图1B中的T-NeRV解码器140的一系列T-NeRV块250。如图2所示,线性层252从来自T-NeRV编码器110的时间特征输出230中提取每通道统计量,例如平均值和方差,这些统计量用于通过自适应实例归一化(AdaIN)层254移动输入张量的分布,从而将时间信息注入到一系列T-NeRV块中。注意,时间特征输出230和一系列T-NeRV块250通常分别对应于时间特征输出130和一系列T-NeRV块150,在图1A和1B中不同地示出。因此,时间特征输出130和系列T-NeRV块150可以共享由本公开归于相应时间特征输出230和系列T-NeRV块250的任何特征,反之亦然。

[0026] 如图2中进一步示出的,除了线性层252和AdaIN层254之外,系列T-NeRV块150/250的前两个T-NeRV块包括具有信道扩展的卷积层256(例如,5×5卷积层)和后续的像素重排(PixelShuffle)层258,其可以执行空间上采样,并且其后可以是高斯误差线性单位(GELU)激活函数形式的激活函数260a。在一些实施方式中,T-NeRV解码器140可以利用低的信道减少因子,例如r=1.2,并增加内核大小。这种方法允许T-NeRV解码器140将更多的参数分配给后面的阶段,有助于高频细节的重建。具体而言,该策略使得系列T-NeRV块150/250的第三个T-NeRV块最大,鉴于其在上采样和光流预测中的双重用途,这是有利的。

[0027] 注意,尽管HNeRV教导了内核大小$k_i = \min\{2i-1,5\}$,但是发现5×5卷积计算量大且参数效率低。与HNeRV进一步对比,系列T-NeRV块150/250的第三至第五个T-NeRV块采用两个连续的卷积层262(例如,3×3卷积层),它们具有相同的参数总数和在卷积层262之间的附加激活260b,导致图2的画面(b)中所示的块架构。更高的参数效率和附加的非线性进一步帮助T-NeRV解码器140重建高频信息。

[0028] 在图1B中,基于流的扭曲块144充当流引导聚合模块,用于经由光流重用来自相邻

帧的信息。为此,头部层164a,即系列T-NeRV块150/250的第五个T-NeRV块之后的一个或多个输出层,输出独立帧166($I_t \in \mathbb{R}^{3\times H\times W}$)和两个聚合权重$w_I$($w_A \in \mathbb{R}^{H\times W}$)。复制独立帧166($I_t$)从计算图中分离并存储在独立帧缓冲器146中。然后可以限定聚合窗口$\mathcal{W}$,例如$\mathcal{W}=\{-2,-1,1,2\}$,以利用来自前两帧和后两帧的信息。在图1A中,帧索引104,t的前向传递过程中,系列T-NeRV块150/250的第三个T-NeRV块之后,头部层164b为每个聚合窗口$\mathcal{W}$中的j预测光流图143($M(t,t+j)\in\mathbb{R}^{2\times H/4\times W/4}$)和权重图($w_M(t,t+j)\in\mathbb{R}^{H/4\times W/4}$)。然后可以通过softmax函数对光流图和权重图进行双线性上采样和归一化。然后,基于流的扭曲块144可以使用光流图M(t,t+j)来扭曲对应的相邻独立帧I_{t+j},并且可以以加权的方式聚合结果以获得聚合帧148A_t,如下:

[0029] $$A_t=\sum_{j\in\mathcal{W}} w_M(t,t+j)\cdot \mathrm{WARP}\left(I_{t+j},M(t,t+j)\right)$$ (等式2)

[0030] 最后,在聚合块168处的另一加权聚合中,通过组合聚合帧148(A_t)和独立帧166(I_t)来获得预测帧170,即预测帧170($\hat{F}_t=w_I\cdot I_t+w_A\cdot A_t$)。

[0031] 关于包括T-NeRV编码器110和T-NeRV解码器140的T-NeRV网络的训练,注意到INR压缩可以被建模为速率失真问题L=D+λR,其中D表示一些失真损失,R表示INR参数θ的熵,λ在两者之间建立了一个平衡。训练INR损失L共同具有在训练期间最小化速率和失真的期望结果,从而允许它们在训练结束后通过熵编码参数θ来实现压缩。因此,可以使用包括失真损失和熵损失的优化损失函数来训练包括T-NeRV编码器110和T-NeRV解码器140的T-NeRV网络,其中可以联合优化失真损失和熵损失。

[0032] 此外,因为上述训练过程需要离散的符号集,例如整数集$\mathbb{Z}$,所以使用本领域已知的直通估计器(STE)来执行量化感知训练,以确保可区分性。使用具有两个可训练参数的尺度重新参数化来独立量化每一层。然后,通过将小的神经网络拟合到权重分布,可以独立地估计每层的熵。

[0033] 然后,相同的框架被扩展以处理参数编码,例如T-NeRV编码器110的多分辨率特征网格112,以及嵌入。多分辨率特征网格112中包括的多个特征网格中的每一个,例如三个特征网格,被视为具有其自己的量化参数的其自己的层。在正向传递过程中,在对两个特征进行双线性插值之前,使用相同的比例重新参数化方案对每个特征网格进行量化和反量化。注意,每个特征网格采用一个熵模型允许T-NeRV网络确定每个时间频率的特征应该被压缩的程度。

[0034] 与上面参考多分辨率特征网格112描述的方法相比,对于所有特定帧嵌入116采用单个熵模型,从而鼓励T-NeRV网络利用帧之间的冗余。这平衡了将时间特征128注入到上述特定帧嵌入116中,限制T-NeRV网络仅在视频质量的增益超过熵的增加时使用这种时间信息。像失真损失一样,熵损失通过T-NeRV网络反向传播,这促使T-NeRV网络在其呈现低熵的特征网格中学习特征。

[0035] 将参考图3进一步描述图1A所示的T-NeRV编码器110的功能。图3示出了流程图380,其概述了根据一个实施方式的用于执行可调混合神经视频表示的示例性方法。关于图3中概述的方法,注意到某些细节和特征已经从流程图380中省略,以免模糊本申请中对发明特征的讨论。

[0036] 结合图1A参考图3,流程图380概述的方法包括由T-NeRV编码器110接收包括多个视频帧(例如输入视频帧102)的视频序列(动作381)。如图1A所示,T-NeRV编码器110可以使用T-NeRV编码器110的大编码器120来接收输入视频帧102以及在动作381中接收的视频序列中包括的其他视频帧。如上所述,与HNeRV网络中使用的传统编码器相比,大编码器120可以大约大一百倍,即包括多一百倍的权重。

[0037] 继续结合参考图3和1A,由流程图380概述的方法还包括生成在动作381中接收的视频序列中包括的视频帧的第一视频帧(以下称为"输入视频帧102")的特定帧嵌入116(动作382)。如上所述,特定帧嵌入116可以在一个或两个步骤中生成:首先,大编码器120从输入视频帧102中提取空间帧嵌入122,然后,在一些使用情况下,其可以用作特定帧嵌入116,或者在其他使用情况下,可以用附加的时间特征128来扩充。

[0038] 在其中由大编码器120输出的空间帧嵌入122($\hat{e}_t$)用时间特征128进一步扩充以提供特定帧嵌入116的实施方式中,可以利用PE124以及随后的MLP126来从帧索引104中提取时间信息。该时间信息可以从包括在动作381中接收的视频序列中的视频帧的子集提取,其中视频序列的视频帧的子集包括输入视频帧102。由MLP126输出的时间信息然后可以被重塑为时间特征128($s_t \in \mathbb{R}^{c_E \times h \times w}$),匹配空间帧嵌入122的维度。时间特征128然后可以被融合到空间帧嵌入122中,以获得特定帧嵌入116。因此,在动作382中,T-NeRV编码器110可以使用大编码器120来提供空间帧嵌入122,并且在一些使用情况下,将空间帧嵌入122与时间特征128相结合,从而生成特定帧嵌入116。

[0039] 继续结合参考图3和1A,流程图380概述的方法还包括识别在动作381中接收的视频序列中包括的视频帧子集的一个或多个GOP特征,其中视频帧子集包括输入视频帧102(动作383)。这种GOP特征可以包括例如I帧(帧内编码图像)、P帧(预测图像)和B帧(双向图像),以及整数M和N,其中M表示两个锚帧之间的帧数,即I帧和P帧之间的距离或两个P帧之间的距离,N表示两个I帧之间的帧数。注意,在动作383中从中识别一个或多个GOP特征的视频帧的子集可以是作为动作382的一部分可选地从中提取时间特征128的视频帧的相同子集,或者是在动作381中接收的也包括输入视频帧102的视频序列的视频帧的另一子集。

[0040] 动作383可以由T-NeRV编码器110使用多分辨率特征网格112来执行,以获得一个或多个GOP特征。如上所述,在前向传递期间,帧索引104(t)被用于索引到网格中,在多分辨率特征网格112的每一级选择帧索引104位于其中的两个特征。可以在这两个特征之间执行双线性内插,并且来自每个级别的结果可以被连接、融合或以其他方式组合以获得GOP特征114。如上所述,与FFNeRV相反,T-NeRV允许多分辨率特征网格112的不同级别的特征相对于通道c_i的数量在大小上变化,同时保持它们的空间维度(h,w)。这种修改允许T-NeRV有利地调节针对不同GOP大小捕获的信息量。

[0041] 结合图3和图1A,流程图380中概述的方法进一步包括将输入视频帧102的特定帧嵌入116与在动作383中识别的GOP特征114相结合,可以通过将它们串联,或采用其他任何数学方法、神经网络或变换器相结合,以提供与输入视频帧102的压缩版本相对应的潜在特征118(动作384)。如上所述,T-NeRV编码器110是基于可调神经网络的视频编码器,其配置为使用融合块127将特定帧嵌入116与GOP特征组进行组合,以将它们连接起来,或者可选地,使用例如任何其他数学技术、神经网络或变换器将它们进行组合。特定帧嵌入116与GOP特征114的组合为特定内容微调提供了杠杆,因为强调特定帧嵌入116从视频帧中提取了更

多的高频信息,从而提高了对低运动序列的性能。相反,给予GOP特征114更多的突出允许T-NeRV模型提取更多的时间上下文,解码器140可以利用这些时间上下文来更好地重建动态序列中的帧。如上所述,在T-NeRV模型的训练和优化期间,自动执行这种杠杆作用,或者特定帧嵌入与特定GOP的特征的组合的平衡,尽管系统用户可以通过限定所使用的特征图的最大尺寸来断言一些控制。

[0042] 因此,在动作384中,T-NeRV编码器110可以将潜在特征118提供为特定帧嵌入116和GOP特征114的加权或未加权组合。此外,在潜在特征118被提供为特定帧嵌入116和GOP特征114的加权组合的一些实施方式中,应用于特定帧嵌入116的第一权重或应用于该加权组合中的GOP特征的第二权重中的一个或两个可以由T-NeRV编码器110基于例如在动作381中接收的视频序列中描绘的内容主要是静态的还是主要是动态的来确定。

[0043] 结合参考图1A、1B和图3,在一些实施方式中,流程图380概述的方法还可以包括将输入视频帧102的潜在特征118输出到解码器,例如T-NeRV解码器140,或者在其他实施方式中,甚至是另一种类型的传统的基于NeRV的解码器(动作385)。注意,动作385是可选的,并且在一些实施方式中,流程图380概述的方法可以以上述动作384结束。可选动作385在被执行时由T-NeRV编码器110执行,并且如上所述,还可以包括向解码器提供时间特征输出130。

[0044] 注意,尽管动作382、383和384或动作382、383、384和385在上文被描述为一次对单个视频帧(例如,输入视频帧102)执行,但是在一些实施方式中,这些动作可以对视频序列的片段(以下称为“子序列”)执行。也就是说,在一些使用情况下,输入视频帧102可以对应于在动作383中识别的视频帧子集的子序列,特定帧嵌入116可以包括分别对应于识别的视频帧子集的子序列的每个视频帧的多个特定帧嵌入,以及潜在特征118可以是多个潜在特征,每个潜在特征分别对应于所识别的视频帧子集的子序列的视频帧的压缩版本。在潜在特征118对应于多个潜在特征的实施方式中,在动作385中,这些多个潜在特征可以被输出到解码器,或者可以被存储用于以后的解码。

[0045] 在一些实施方式中,流程图380概述的方法可以以上述动作385结束。然而,在包括T-NeRV编码器110的视频压缩系统还包括T-NeRV解码器140的实施方式中,流程图380概述的方法还可以包括由T-NeRV解码器140从T-NeRV编码器110接收潜在特征118,以及由T-NeRV解码器140解码潜在特征118以提供预测帧170作为对应于输入视频帧102的未压缩视频帧的动作。注意,可以由T-NeRV解码器以上面参考图1B和图2详细描述的方式来执行潜在特征118到所提供的预测帧170的解码。

[0046] 关于流程图380概述的和上面描述的方法,注意,动作381、382、383和384,或者动作381、382、383、384和385可以在自动化过程中执行,其中可以省略人工参与。

[0047] 因此,本申请公开了用于执行可调混合神经视频表示的系统和方法,其通过至少以下贡献推进了现有技术:(i)开发T-NeRV编码器,其将特定帧嵌入与特定GOP的特征相结合,从而为特定内容微调提供杠杆;以及(ii)扩展信息论INR压缩框架以包括嵌入,从而有利地利用其中的显著冗余。如上所述,当在UVG数据集上评估时,本申请公开的T-NeRV解决方案在视频表示和视频压缩任务上都优于所有在先的视频INR。

[0048] 根据以上描述,很明显,在不脱离本申请中描述的概念的范围的情况下,可以使用各种技术来实现这些概念。此外,尽管已经具体参考某些实施方式描述了这些概念,但是本领域普通技术人员将认识到,在不脱离这些概念的范围的情况下,可以在形式和细节上进

行改变。因此,所描述的实施方式在所有方面都被认为是说明性的而非限制性的。还应当理解,本申请不限于这里描述的特定实施方式,而是在不脱离本公开的范围的情况下,许多重新布置、修改和替换都是可能的。

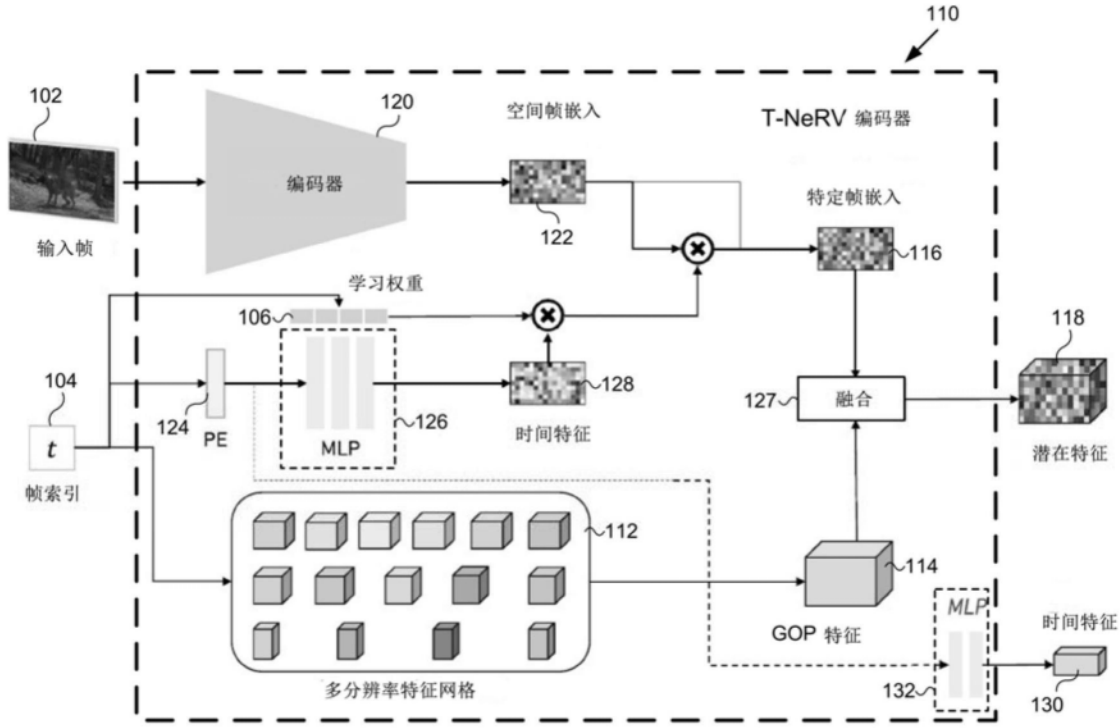

110
T-NeRV 编码器
102
输入帧
120
编码器
空间帧嵌入
122
特定帧嵌入
116
学习权重
106
104
t
帧索引
124
PE
MLP
126
128
时间特征
127
融合
118
潜在特征
112
多分辨率特征网格
114
GOP 特征
MLP
时间特征
132
130

图1A

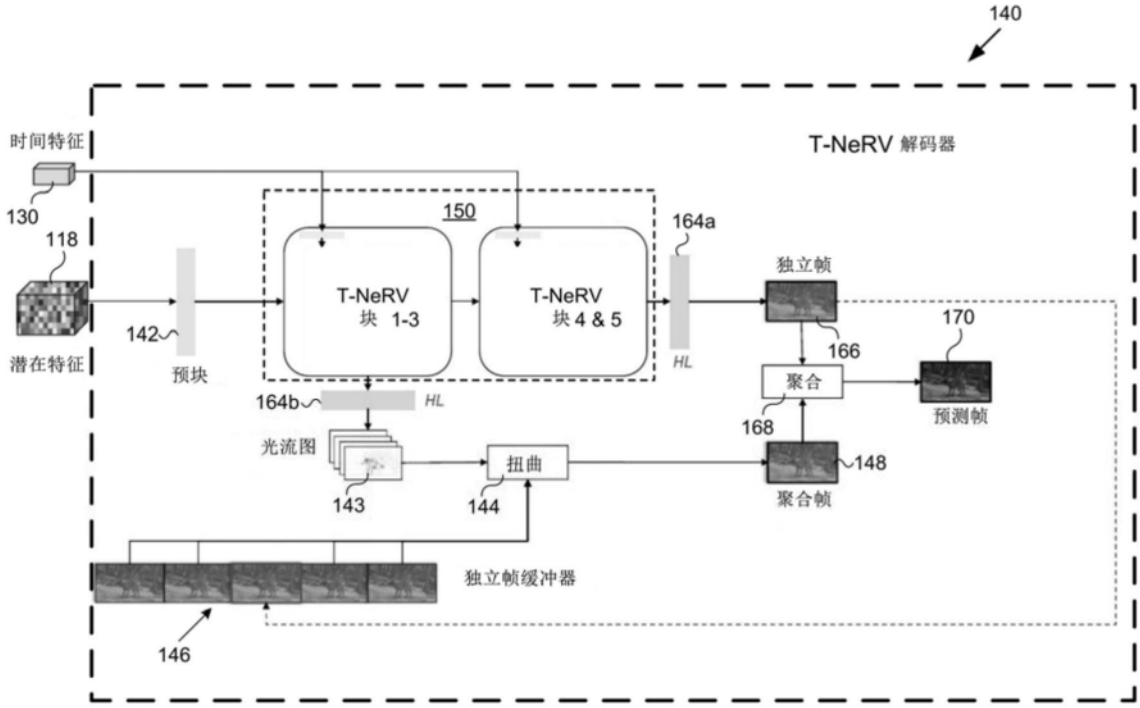

140
T-NeRV 解码器
时间特征
130
118
潜在特征
142
预块
150
T-NeRV
块 1-3
T-NeRV
块 4 & 5
164a
HL
164b
HL
独立帧
166
聚合
170
预测帧
168
光流图
143
扭曲
144
148
聚合帧
独立帧缓冲器
146

图1B

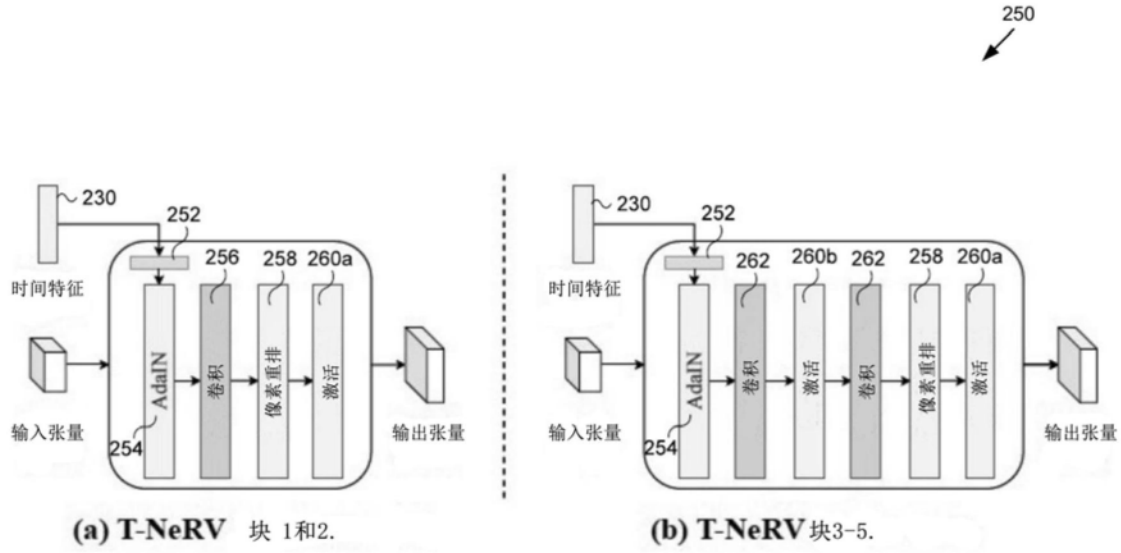

250
230
252
256
258
260a
时间特征
输入张量
AdaIN
卷积
像素重排
激活
输出张量
254
(a) T-NeRV 块 1和2.
230
252
262
260b
262
258
260a
时间特征
输入张量
AdaIN
卷积
激活
卷积
像素重排
激活
输出张量
254
(b) T-NeRV 块3-5.

图2

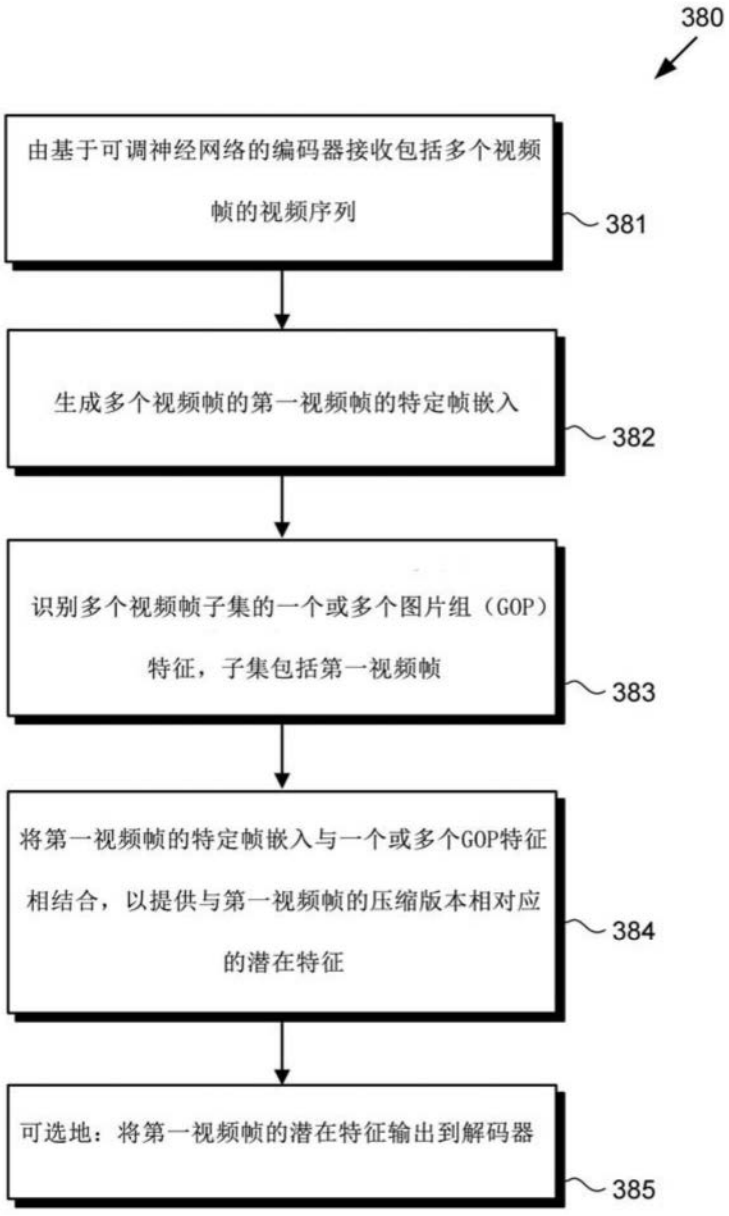

380
由基于可调神经网络的编码器接收包括多个视频帧的视频序列
381
生成多个视频帧的第一视频帧的特定帧嵌入
382
识别多个视频帧子集的一个或多个图片组（GOP）特征，子集包括第一视频帧
383
将第一视频帧的特定帧嵌入与一个或多个GOP特征相结合，以提供与第一视频帧的压缩版本相对应的潜在特征
384
可选地：将第一视频帧的潜在特征输出到解码器
385

图3