计算机科学
COMPUTER SCIENCE

一种3D可变形卷积结合Transformer的视频压缩感知方法

杜秀丽, 朱金耀, 高星, 吕亚娜, 邱少明

引用本文

杜秀丽, 朱金耀, 高星, 吕亚娜, 邱少明. 一种3D可变形卷积结合Transformer的视频压缩感知方法[J]. 计算机科学, 2025, 52(11): 150-156.

DU Xiuli, ZHU Jinyao, GAO Xing, LYU Yana, QIU Shaoming. Video Compressed Sensing Method with Integrated Deformable 3D Convolution and Transformer [J]. Computer Science, 2025, 52(11): 150-156.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240800026

一种3D可变形卷积结合Transformer的视频压缩感知方法

杜秀丽　朱金耀　高　星　吕亚娜　邱少明
大连大学通信与网络重点实验室　辽宁 大连 116622
大连大学信息工程学院　辽宁 大连 116622
(duxiuli@dlu.edu.cn)

摘　要　面对视频的分辨率越来越高导致数据量越来越大的挑战,以更低的采样率实现视频的高质量重构可降低对通信资源的占用,进而降低采样端的部署难度。然而,现有的视频压缩感知方法对视频的帧间相关性无法充分利用,低采样率下的视频重构质量有待进一步提高。随着深度学习技术的引入,基于深度学习的分布式视频压缩感知给视频压缩感知重构提供了新思路。因此,结合3D可变形卷积与Transformer构建CS3Dformer网络,利用3D可变形卷积捕获视频的局部特征和时空特征的有效性,学习视频帧间的时空特征;同时,利用Transformer捕获长距离依赖特征的优点,一定程度上弥补了卷积神经网络方法在捕获图像的非局部相似性方面的缺陷,能更好地实现对视频的建模。所提方法是一种端到端的视频压缩感知方法,在多个数据集上的实验结果验证了该方法的有效性。

关键词:压缩感知;视频重构;可变形卷积;Transformer;卷积神经网络
中图分类号　TN919.81

Video Compressed Sensing Method with Integrated Deformable 3D Convolution and Transformer

DU Xiuli,ZHU Jinyao,GAO Xing,LYU Yana and QIU Shaoming
Key Laboratory of Communication and Network,Dalian University,Dalian,Liaoning 116622,China
School of Information Engineering,Dalian University,Dalian,Liaoning 116622,China

Abstract　Facing the challenge of increasing data volume due to higher resolution of video,realizing high quality video reconstruction with lower sampling rate can reduce the consumption of communication resources and thus reduce the difficulty of deployment at the sampling end. However,the existing video compressed sensing methods cannot fully utilize the inter-frame correlation of the video,and the reconstruction quality of the video at low sampling rates needs to be further improved. With the introduction of deep learning technology,distributed video compression sensing based on deep learning provides new ideas for video compression sensing reconstruction. Therefore, this paper combines 3D deformable convolution with Transformer to construct CS3Dformer network,which utilizes the effectiveness of 3D deformable convolutional network in capturing local and spatio-temporal features of video and learns spatio-temporal features between video frames,and at the same time,utilizes the advantages of Transformer in capturing long-range dependency features,which compensates to some extent for the advantages of convolutional neural network method in capturing the non-local similarity of the defects of image,and better realize the modeling of the video. This method is an end-to-end video compression perception method,the experimental results on multiple datasets verify the effectiveness of the proposed method.

Keywords　Compressive sensing,Video reconstruction,Deformable convolution,Transformer,Convolutional nerul network

1　引言

目前,压缩感知(Compressive Sensing,CS)[1]技术不断发展,其因可以在远低于奈奎斯特采样率的条件下对信号进行采样[2],在许多视频监控等资源受限的场景中得到了广泛应用[3]。近年来,随着深度学习技术在各个领域的全面发展,基于卷积神经网络的端到端重建框架也被应用于压缩感知技术中,它主要利用卷积神经网络(Convolutional Nerul Network,CNN)进行逐块采样和恢复,数据驱动的图像采样模块使得采样矩阵是可学习的,该框架取得了不错的重建效果。可学习的测量矩阵避免了传统压缩感知算法在设计测量矩阵时的复杂性[4],即需要考虑随机测度的构建原理与合理性分析。分布式视频压缩感知框架[5]是目前最流行的视频压缩感知方法,在分布式视频压缩感知任务中,所有帧的采样使用分布式

到稿日期:2024-08-03　返修日期:2025-02-20
基金项目:辽宁省教育厅项目(JYTMS20230377)
This work was supported by the Liaoning Provincial Department of Education(JYTMS20230377).
通信作者:朱金耀(zhujinyao@s.dlu.edu.cn)

编码[6]的方法，只对部分帧（即关键帧），使用更高的采样率进行采样，可以相对减少对通信资源的占用。由于帧间具有相似性以及关键帧蕴含更多信息[7]，有效利用帧间冗余信息进行重构成为研究热点。本文主要针对利用帧间冗余信息在更低采样率下实现高质量重构这一难题，开展了 Transformer 与 CNN 结合的视频压缩感知重构方法的研究。

CNN 通过卷积操作在图像的局部邻域内提取特征，因此 CNN 主要关注局部信息，难以建立全局图像的长距离关联。Transformer[8]是一种不同于卷积神经网络的深度学习架构，它可以将上下文中的文本转换为被称为 token 的数字表示，每个 token 之间利用各自的 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 特征矩阵实现多级特征匹配，使得关键 tokens 的信号得以放大，不重要的 tokens 得以减弱，因此它可以对长距离关联特征进行很好的建模。后续有学者将 Transformer 应用到了机器视觉任务中[9]。此任务将图片分为 patches，每个 patch 通过 patch-embedding 的方式转换为 token，因此可以利用自注意力机制建立图像中不同 patch 之间的长距离关联特征。综上所述，若能将 CNN 与 Transformer 相结合，则能同时获得数据的局部特征与长距离关联特征，提高模型性能。CNN 与 Transformer 结合的深度学习框架已用于图像压缩感知[10-11]、医学图像处理[12]、图像修复[13]和图像融合[14]等任务，并取得了优秀的效果。

目前的压缩感知方法有基于数据驱动的方法和基于深度图像先验（Deep Image Prior，DIP）的方法。基于数据驱动的方法虽然可以取得较好的效果，但需要大量的数据集进行训练。与图像不同，视频的重建不仅要考虑帧内的空间特征，还需考虑视频帧之间的时间关系，而现有的基于数据驱动的深度学习视频压缩感知方法对视频序列之间相关性的利用不足。基于 DIP 的方法不需要大量的训练数据，且适用于有着许多手工设计参数的重建任务，但其在利用视频序列的相关性方面也有很大的改进空间。具体来说，CSVideoNet[15]利用 CNN 对关键帧与非关键帧进行初始重建，然后利用 LSTM 网络来提取连续帧的运动特征以进行深度重建。Hybrid-3D[16]利用一种混合 3D 卷积网络来获取帧间的时空信息。Chen 等[17]提出了基于纹理特征的分布式视频压缩感知自适应重构方法，在重构时选择相邻视频帧的纹理特征作为当前帧的参考，来充分利用帧间相关性。VCSNet[18]利用 CNN 的残差连接来传输帧间信息，以实现帧间信息的多级补偿。VSCL[19]与 AVCSR[20]使用基于感兴趣区域（Region of Interest，ROI）的压缩感知算法对 ROI 与非 ROI 分配不同的编码资源，实现了对编码资源的高效利用。RRS[21]是一种加权残差稀疏的方法，该方法通过帧间多假设预测和残差加权稀疏度建模交替进行提高了重建质量，但重建时间较长。LRR-VCSNet[22]利用低秩正则化建立起非局部结构化相关性，从而进行视频压缩感知重建。Video-MH[23]是首次提出的多假设视频重构的方法，此方法将运动估计的思想引入分布式视频压缩感知框架，来估计当前帧与参考帧之间的运动。Du 等[24]使用一种加权非局部相似性的多假设视频重构算法，将本帧邻域块信息作为参考，对于视频帧大变化块中的纹理块与非纹理块，采用不同的加权非局部相似性算法进行重构。Sun 等[25]使用一种窗口位置和大小可自适应变化的多假设视频重构算法，结合光流法有效提高了非关键帧的重构质量。

尽管上述视频压缩感知重构方法利用了帧间的相关性，但关键帧与非关键帧的重构质量仍存在较大差异，且在低采样率下重建质量不高。为了解决这个问题，本文提出了一种结合 3D 可变形卷积[26]与 Transformer 的视频压缩感知方法。本文的主要贡献如下：

1）以关键帧信息为基础，利用 3D 可变形卷积在 3 个维度上的偏移提取视频局部特征，同时帮助非关键帧高效利用关键帧的丰富特征。

2）利用 Transformer 建立起关键帧中不同区域之间的长距离关联，利用帧间高度相似的特点，将关键帧的长距离关联信息作为非关键帧特征的补充，这有利于非关键帧的重建，提升整体重构质量。

3）将 CNN 与 Transformer 的优点相结合，同时利用视频的局部信息与长距离关联信息，充分地利用了帧间冗余信息。大量实验结果验证了本文提出的 CS3Dformer 在更低采样率情况下的有效性。

2 理论基础

2.1 压缩感知

CS 理论表明，如果信号在某些变换域是稀疏的，则可以通过非线性恢复过程，利用通过线性方式获得的相对较少的测量值精确地重建信号。采样过程可以表示为：

$$\boldsymbol{y}=\boldsymbol{\Phi}x \tag{1}$$

其中，$\boldsymbol{y}\in\boldsymbol{R}^{MR\times n}$ 表示测量值；$\boldsymbol{x}\in\boldsymbol{R}^{B^2\times n}$ 表示需要恢复的原始信号；$\boldsymbol{\Phi}\in\boldsymbol{R}^{MR\times B^2}$ 表示采样矩阵，实际采样中往往采用基于块的 CS 采样方式[27]，以缓解存储压力，其中 B 表示分子块的大小，n 表示分块的数量，MR 表示采样率。

现有的重构方法分为传统的基于分析模型的重构方法与基于深度学习的重构方法。传统方法通过求解分析模型的重建过程可以表示为：

$$\arg\min_{x}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\Phi x}\|_2^2+\lambda S(x) \tag{2}$$

通过最小化测量值 $\boldsymbol{y}$ 与 $\boldsymbol{\Phi}x$ 的均方误差来重建，同时引入信号的稀疏先验或低秩先验 $S(x)$作为正则化项。基于深度学习的重构方法可表示为：

$$\arg\min_{\omega}\frac{1}{2}\sum_{i=1}^{N}\|F(y_i,\omega)-x_i\|_2^2 \tag{3}$$

其中，$F(y_i,\omega)$是神经网络模型重建结果的数学表达，这里的 ω 是可训练的参数，y_i 表示测量值。通过神经网络模型在 N 个数据集上最小化神经网络的重建值与原始值 x_i 间的差来训练 ω，以实现重构。

2.2 端到端的深度学习视频压缩感知

在压缩感知的实际应用中，信号的采样与重建是分别进行的，采样端大多应用于资源受限的场景，重构端则有着丰富的计算资源。采样端在采集的同时完成信号的压缩并传输给重构端，重构端利用丰富的计算资源完成被压缩信号的重构。若能以更低的采样率实现对压缩信号的高质量重建，则能进一步降低采样端的成本。

深度学习技术通过数据驱动的方式来学习模型的参数，采样端与重构端的联合训练使得模型可以根据数据的特性，自适应地学习最优的参数，从而提高模型性能。现有的基于

深度学习的压缩感知方法都包含了 3 个步骤：采样、初始重建和深度重建。与图像压缩感知不同，视频压缩感知在采样阶段采用分布式视频编码的方法，即将视频帧序列划分为多个图像组(Group of Picture，GOP)，图像组内的视频帧被划分为关键帧与非关键帧。编码时，对关键帧与非关键帧分别以相对高采样率与相对低采样率进行独立的采样。只对关键帧进行高采样率采样，可以在提高重构质量的同时相对减少对通信资源的占用。解码分为初始重建和深度重建两步，初始重建时对关键帧和非关键帧分别进行初始重建。深度重建时，利用帧之间的时间冗余信息，即关键帧的信息，来提升非关键帧的重构质量。

3 本文方法

本文提出了一个结合 3D 可变形卷积与 Transformer 的分布式视频压缩感知模型，整体模型由采样模块、初始重构模块、CNN 独立重构模块、Transformer 模块和 ResD3D 模块组成。视频帧输入网络后，每一帧首先利用可训练的采样矩阵 $\boldsymbol{P}$ 与重构矩阵 $\boldsymbol{R}$ 进行采样和初始重构。然后使用 Transformer模块与独立重构模块分别获取关键帧不同区域的长距离关联特征和关键帧的空间特征，将这两种特征在通道维度上进行拼接，再使用一个 3D 可变形卷积层实现两种特征的自适应融合。同时，将非关键帧通过独立重构获取的空间特征与关键帧的长距离关联特征以同样的方式进行融合，非关键帧因此得到关键帧的长距离关联特征作为补充。最后将每一帧融合后的特征送入 5 层残差连接的 3D 可变形卷积模块，生成最终的重建结果。整体网络框架如图 1 所示。

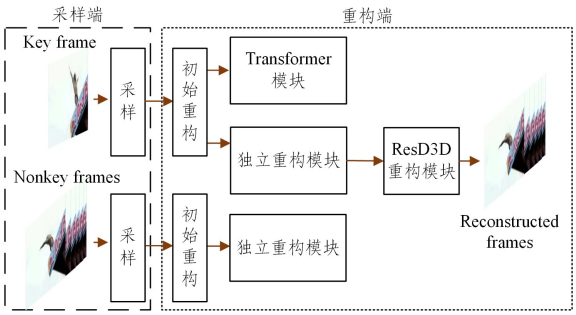


图 1 采样端与重构端的整体网络框图

Fig. 1 Overall network diagram of sampling module and reconstruction module

3.1 采样模块

采样模块由一个可训练的采样矩阵组成。如图 2 所示，以其中一帧的采样为例，输入视频帧 $\boldsymbol{x}\in R^{H\times W\times T}$，其中一帧为 $\boldsymbol{x}_i\in R^{H\times W}$，将此帧分为 B^2 大小的块，然后经过 reshape 将每一个 B^2 大小的块展平为一维向量，这样 $\boldsymbol{x}_i$ 变为二维向量，此时 $\boldsymbol{x}_i\in R^{\left(\frac{H}{B}\times\frac{W}{B}\right)\times B^2}$，设采样率为 MR，采样矩阵 $\boldsymbol{P}$ 的尺寸为 $(B^2, B^2\times MR)$。将分块后的输入与采样矩阵相乘得到采样值 y_i。

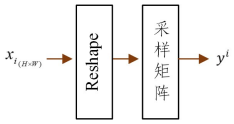


图 2 采样模块

Fig. 2 Sampling module

3.2 重构模块

3.2.1 初始重构模块

初始重构模块由一个可训练的重构矩阵组成。如图 3 所示，以其中一帧的重构为例，设采样率为 MR，重构矩阵 $\boldsymbol{R}$ 的尺寸为 $(B^2\times MR, B^2)$，采样值 y_i 乘以重构矩阵然后经过变形得到初始重构图像 $\boldsymbol{x}_i'$。

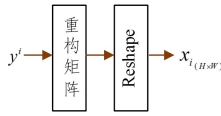


图 3 初始重构模块

Fig. 3 Initial reconstruction module

3.2.2 CNN 独立重构模块

CNN 独立重构模块如图 4 所示，每一帧的初始重构结果先经过 CNN 进行独立重构，之后可以获得进一步的重构结果，以便在 ResD3D 模块中更高效地利用帧间冗余信息。

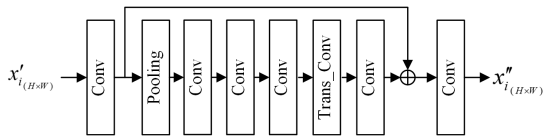


图 4 CNN 独立重构模块

Fig. 4 CNN module for per frame

3.2.3 Transformer 模块

Transformer 模块如图 5 和图 6 所示，将关键帧划分为多个块，对关键帧进行基于块与基于像素的 Transformer 来建立关键帧不同区域之间的长距离依赖关系。

首先将关键帧分块为多个 patches，将每个 patch 转换为 ***token***，不同块的 token 输入 Transformer Unit 进行长距离关联特征的建立，最终输出一个中间重构结果。对此结果进行基于像素的重建，即每个像素点通过一次矩阵变换，获得每个像素点的 ***token***，将各像素点的 ***token*** 输入 Transformer Unit 进行基于像素的长距离关联特征建立，最终得到关键帧的重构结果。

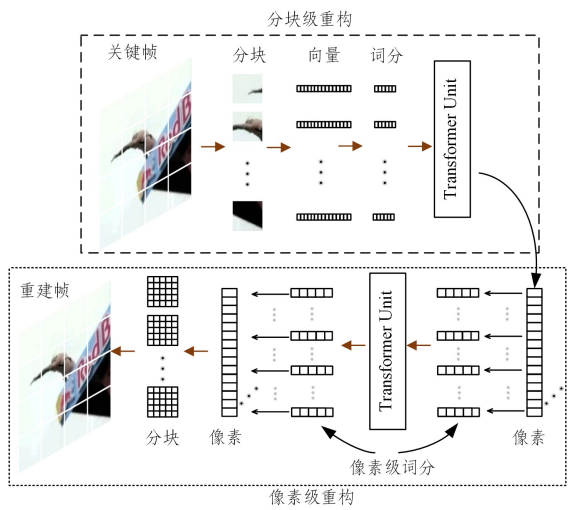


图 5 Transformer 模块

Fig. 5 Transformer module

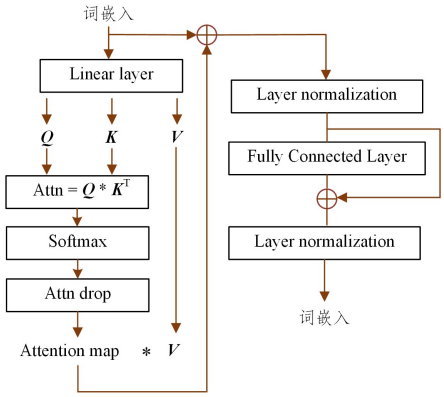


图 6 Transformer 单元

Fig. 6 Transformer unit

3.2.4 ResD3D 模块

ResD3D 模块如图 7 所示，其由多层残差连接的 3D 可变形卷积层组成。将每一帧的 CNN 独立重构模块的重构结果与关键帧的 Transformer 模块重构结果进行融合，所有帧作为一个 1 通道 5 维特征图输入一个 3D 可变形卷积层生成多通道特征，然后这些特征通过 ResD3D 模块进行深度重构。最终输出的多通道重构结果通过一个 3D 可变形卷积层，变为 1 通道的输出结果。

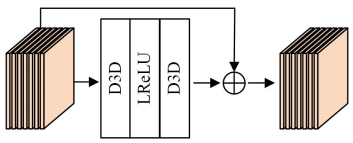


图 7 残差连接的 3D 可变形卷积模块

Fig. 7 3D deformable convolution module with residual connection

3.3 复杂运动场景下特征的融合机制

在复杂运动场景下，视频内目标或者视频背景其中之一可能出现大幅度变化。然而，一个 GOP 内，以 8 帧为例，第 1 帧与第 8 帧区别较大，但中间帧即第 4 帧或第 5 帧与第 1 帧和第 8 帧的区别会相对小。因此，本文实验中选取第 5 帧为关键帧，利用中间帧与起始帧和末尾帧更相似的特点，在通道维度上将 Transformer 在信息丰富的关键帧中建立起的分块间长距离依赖关系与 CNN 特征拼接，然后通过一个 3D 可变形卷积层，利用 3D 可变形卷积采样位置可变化的特点，有选择地对二者的特征进行加权求和，以此高效地实现二者特征的自适应融合。

4 实验结果及分析

4.1 实验配置

4.1.1 数据集

由于没有用于视频压缩感知的特定数据集，本文实验使用种类丰富且数据量大的 UCF-101[28] 作为数据集。从中挑选前 25 种类别用作训练和验证，训练集和验证集划分比例为 5∶1，每个视频帧裁取中心区域（分辨率为 160×160）。为了与最新的视频压缩感知算法进行对比，将训练分块大小设为 32×32，同时使用了 UCF-101 中 Wallpushups 与 TaiChi 中的部分视频序列作为测试对比。

4.1.2 训练环境及训练参数

本实验基于 PyTorch1.11.0 的框架实现，在 Ubuntu 中用 Python 进行实验，训练时优化器采用 Adam[29]，初始学习率为 1×10^{-4}，每训练 10 个 epoch 后降为原来的 0.8 倍，batchsize 设为 1，训练轮次 epochs 为 100，关键帧的采样率设为 0.5，非关键帧的采样率分别设置为 0.1，0.05 和 0.01，使用均方误差作为损失函数，在此设置下使用 GeForce GTX3090 进行训练。采用两个 GOP 内所有帧的平均峰值信噪比和平均结构相似度作为衡量重构质量的客观标准。

4.1.3 关键帧的设置

目前的视频压缩感知算法在一个 GOP 内有着不同的关键帧与非关键帧的划分方法。CSVideoNet 使用了一张关键帧，VCSNet 使用了一张关键帧与两张关键帧进行了对比实验。为了验证所提算法的有效性，本文同样选择了单关键帧与双关键帧进行对比实验。单关键帧为一个 GOP 中的第 5 帧，双关键帧为一个 GOP 中的第 5 帧和第 8 帧。

4.2 实验分析

为充分验证本文模型的有效性，将其与现有最新的视频压缩感知算法的重构结果进行对比。表 1 列出了不同采样率的情况下，不同算法在多个测试视频序列上的重构质量对比。单关键帧下的重构结果为 VCSNet_1 与 Proposed_1，双关键帧情况下的重构结果为 VCSNet_2 与 Proposed_2。从表 1 中可以看出，在单关键帧情况下，所提算法 Proposed_1 相比 VCSNet_1 在 3 种采样率下的平均 PSNR 分别提升了 2.23 dB，2.36 dB 和 2.93 dB，平均 SSIM 分别提升了 0.0200，0.0295，和 0.0704。

表 1 不同采样率的情况下，各方法在视频序列的前两个 GOP 上的平均 PSNR 与 SSIM 对比

Table 1 Average PSNR and SSIM comparison of various methods on the first two GOPs of video sequences in the case of different measurement rates

序列	Ratio	Video-MH[23]		RRS[21]		VCSNet_1[18]		Proposed_1		VCSNet_2[18]		Proposed_2	
		PSNR	*SSIM*	*PSNR*	*SSIM*	*PSNR*	*SSIM*	*PSNR*	*SSIM*	*PSNR*	*SSIM*	*PSNR*	*SSIM*
Wallpushups_1	0.5/0.1	22.76	0.6542	23.78	0.7549	33.92	0.9527	36.28	0.9571	36.85	0.9755	**36.97**	**0.9780**
Wallpushups_2		23.41	0.6875	26.56	0.7715	38.50	0.9679	40.71	0.9816	40.37	0.9765	**41.04**	**0.9834**
TaiChi_1		25.52	0.7498	26.39	0.8225	34.90	0.9547	37.27	0.9726	37.25	0.9738	**37.70**	**0.9760**
TaiChi_2		25.82	0.7529	24.91	0.7012	36.67	0.9446	38.64	0.9706	39.05	0.9618	**39.39**	**0.9741**
Average		24.38	0.7111	25.41	0.7625	36.00	0.9550	38.23	0.9750	38.38	0.9719	**38.78**	**0.9779**
Wallpushups_1	0.5/0.05	21.26	0.6238	22.64	0.6952	33.77	0.9533	35.14	0.9685	**36.20**	**0.9722**	35.31	0.9698
Wallpushups_2		22.05	0.6389	24.32	0.7124	35.64	0.9437	38.03	0.9717	37.84	0.9588	**38.12**	**0.9744**
TaiChi_1		24.18	0.7051	24.65	0.7758	33.15	0.9342	35.96	0.9608	**36.20**	**0.9663**	35.97	0.9625
TaiChi_2		24.20	0.7112	23.17	0.6856	33.65	0.8977	36.08	0.9456	36.68	0.9328	**36.78**	**0.9536**
Average		22.92	0.6698	23.70	0.7173	34.05	0.9322	36.41	0.9617	**36.73**	0.9575	36.55	**0.9651**

（续表）

序列	Ratio	Video-MH[23]		RRS[21]		VCSNet_1[18]		Proposed_1		VCSNet_2[18]		Proposed_2	
		PSNR	*SSIM*	*PSNR*	*SSIM*	*PSNR*	*SSIM*	*PSNR*	*SSIM*	*PSNR*	*SSIM*	*PSNR*	*SSIM*
Wallpushups_1	0.5/0.01	18.26	0.4553	18.59	0.4779	29.17	0.8861	32.02	0.9468	**34.96**	0.9669	34.86	**0.9687**
Wallpushups_2		16.78	0.4225	17.56	0.4367	30.40	0.8624	32.93	0.9268	33.98	0.9142	**35.71**	**0.9537**
TaiChi_1		19.32	0.5545	20.21	0.5994	29.60	0.8803	33.09	0.9382	**36.18**	0.9555	35.40	**0.9595**
TaiChi_2		18.69	0.5238	20.42	0.6085	28.41	0.7845	31.27	0.8831	32.84	0.8788	**34.81**	**0.9255**
Average		18.26	0.4890	19.20	0.5306	29.40	0.8533	32.33	0.9237	34.49	0.9289	**35.20**	**0.9519**

在双关键帧情况下，所提算法 Proposed_2 相比 VCSNet_2 在 0.5/0.1 采样率下平均 *PSNR* 提升了 0.40dB 和 0.0060；在 0.5/0.01 采样率下的平均 *PSNR* 和平均 SSIM 分别提升了 0.71 dB 和 0.0230。

图 8 给出了在 0.5/0.1 采样率下，在视频序列 Wallpushups_g20_c01 的第 6 帧上不同方法的重构质量对比。图 9 给出了在 0.5/0.01 采样率下，在视频序列 TaiChi_g22_c04 的第 4 帧上不同方法的重构质量对比。从图 8 和图 9 中可以看出，本文提出的 Proposed_1 在单关键帧的情况下的重构质量超过了 VCSNet_1 与 VCSNet_2，对比其他方法取得了更好的视觉效果。从图 10(a)和图 10(b)的折线图对比中可以看出，本文算法提升了非关键帧的重构质量，缩小了非关键帧与关键帧之间的差别，因此在一个 GOP 内的 PSNR 折线图更加平缓。

图 8 0.5/0.1 采样率下在视频序列 Wallpushups_g20_c01 的第 6 帧上不同方法的视觉效果对比

Fig. 8 Visual quality comparison on the 6th frame of video sequence Wallpushups_g20_c01 in the case of sampling rate=0.5/0.1

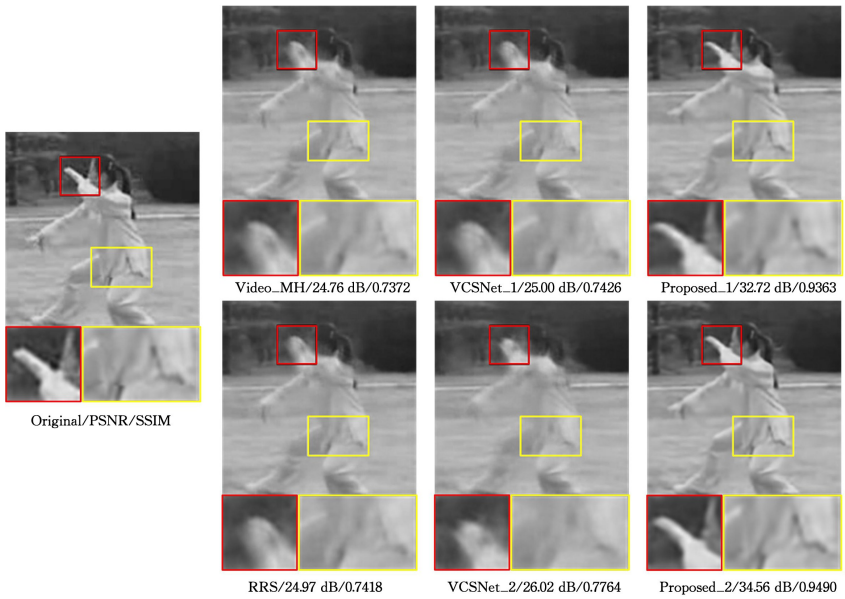
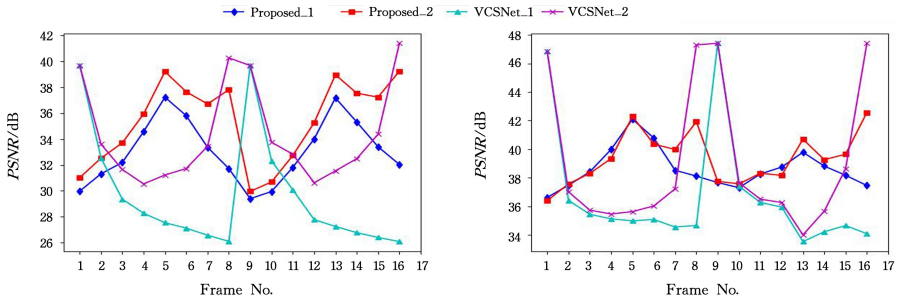


图 9 0.5/0.01 采样率下在视频序列 TaiChi_g22_c04 的第 4 帧上不同方法的视觉效果对比

Fig. 9 Visual quality comparison on the 4th frame of video sequence TaiChi_g22_c04 in the case of sampling rate=0.5/0.01



(a) Sampling ratio=0.01 (b) Sampling ratio=0.1

图 10 单关键帧与双关键帧条件下所提算法与 VCSNet 在视频序列 TaiChi 的两个 GOP 的 PSNR 对比

Fig. 10 PSNR comparison between VCSNet and proposed method on the two GOPs of video sequence TaiChi in case of one key frame and two key frames

4.3 不同模块对重构性能的影响

为了验证不同模块对整体重构性能的影响，本文在不同采样率的条件下重新训练不包含 Transformer 模块的双关键帧网络结构和将 ResD3D 中的 3D 可变形卷积替换为 3D 卷积的双关键帧网络结构，用 4 个视频序列的第 1 个 GOP 进行测试，并取 4 个视频序列的平均 PSNR 与 SSIM 进行对比。

测试结果如表 2 所列，当采样率为 0.5/0.1，0.5/0.05 和 0.5/0.01 时，使用 Transformer 模块的网络相比无 Transformer 模块的网络，平均 PSNR 分别高出 0.50 dB，0.79 dB 和 0.28 dB，平均 SSIM 分别高出 0.0079，0.0126 和 0.0082；使用 3D 可变形卷积的 ResD3D 模块的网络相比使用普通 3D 卷积的 Res3D 模块的网络，在 PSNR 和 SSIM 上有显著提升。结果表明，使用 Transformer 模块和 ResD3D 模块的网络框架对于视频重构性能有所提升，特别是 ResD3D 模块。

表 2 不同采样率的情况下，不同网络结构在一个 GOP 上重构结果的平均 PSNR 与 SSIM 的对比

Table 2 Average PSNR and SSIM comparison of different network structures on the first GOP of 4 video sequences in the case of different measurement rates

采样率	Replacing ResD3D with Res3D		Without Transformer		Transformer and ResD3D	
	PSNR/dB	*SSIM*	*PSNR*/dB	*SSIM*	*PSNR*/dB	*SSIM*
0.5/0.1	25.24	0.7790	38.28	0.9700	**38.78**	**0.9779**
0.5/0.05	22.09	0.7427	35.76	0.9525	**36.55**	**0.9651**
0.5/0.01	20.38	0.7139	34.92	0.9437	**35.20**	**0.9519**

4.4 重构时间与模型复杂度对比

算法的重构时间与重构质量同等重要，表 3 列出了在采样率为 0.5/0.01 的情况下 4 个视频序列的第 1 个 GOP 的平均重构时间。从表 3 中可以看出，Video-MH，RRS 等传统方法在重构质量与重构时间上不占优势；LRR-VCSNet 的优点在于不需要经过长时间的训练，就能对视频进行高质量重构，但其重构时间较长；基于深度学习方法的 VCSNet 计算复杂度较低，重构时间短，但其参数量偏高，且重构性能有待提高。本文模型中使用了 3D 可变形卷积模块，其中 3D 可变形卷积在普通 3D 卷积的基础上引入的偏移量在提高模型重构质量的同时增加了计算复杂度，因此重构时间比 VCSNet 长。总的来说，相比对比算法，本文算法在重构时间上占据一定的优势，且重构性能最佳。

表 3 不同算法的参数量和不同算法在 0.5/0.01 采样率的情况下，在一个 GOP 上的平均重构时间的对比

Table 3 Average computational time and number of parameters of different methods for reconstructing a GOP at the sampling rate of 0.5/0.01

算法	Parameter	Time/s
Video-MH	—	156.65
RRS	—	7899.61
VCSNet	5.45×10^6	10.61
LRR-VCSNet	6.52×10^6	7193.67
Ours	$\mathbf{2.56\times10^6}$	**22.73**

结束语 针对现有视频压缩感知算法对帧之间的时间冗余信息利用不充分，在低采样率条件下难以实现高质量重构这一问题，本文提出了结合 3D 可变形卷积与 Transformer 的模型。在端到端的分布式视频压缩感知模型的基础上，利用 3D 可变形卷积提取的时空信息与 Transformer 建立起关键帧的长距离依赖关系，帮助所有帧充分利用时间冗余信息，通过将 CNN 与 Transformer 提取的特征进行自适应融合，同时把握局部信息与长距离依赖关系，提高了重建结果的质量。本文算法在多个视频序列上进行了验证，其在相同采样率下能实现更优秀的重建效果。后续工作将关注模型的轻量化、对于长视频序列的重建和在更低采样率条件下的重建，以进一步提升重建性能和减小部署难度。

参 考 文 献

[1] DONOHO D L. Compressed sensing [J]. IEEE Transation on Information Theory, 2006, 52(4): 1289-1306.

[2] CANDÈS E J, TAO T. Near-optimal signal recovery from random projections: Universal encoding strategies? [J]. IEEE Transaction on Information Theory, 2006, 52(12): 5406-5425.

[3] CANDES E J, WAKIN M B. An introduction to compressive sampling [J]. IEEE Signal Processing Magazine, 2008, 25(2): 21-30.

[4] SHI W, JIANG F, ZHANG S, et al. Deep networks for com-

pressed image sensing[C]// 2017 IEEE International Conference on Multimedia and Expo(ICME). IEEE, 2017: 877-882.
[5] VEERARAGHAVAN A, REDDY D, RASKAR R. Coded Strobing Photography: Compressive Sensing of High Speed Periodic Videos [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2011, 33(4): 671-686.
[6] DO T T, CHEN Y, NGUYEN D T, et al. Distributed Compressed Video Sensing [C]// 2009 16th IEEE International Conference on Image Processing(ICIP). IEEE, 2009: 1393-1396.
[7] OU Y F, LIU T, ZHAO Z, et al. Modeling the impact of frame rate on perceptual quality of video [C]// Proceedings of the IEEE Conference on Image Processing. 2008: 689-692.
[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.
[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv: 2010. 11929, 2020.
[10] YE D J, NI Z K, WANG H L, et al. CSformer: Bridging convolution and Transformer for Compressive Sensing[J]. IEEE Transactions on Image Processing, 2023, 32: 2827-2842.
[11] SHEN M H, GAN H P, MA C Y, et al. MTC-CSNet: Marrying transformer and convolution for image compressed sensing[J]. IEEE Transactions on Cybernetics, 2024, 54(9): 4949-4961.
[12] YANG Z Y, PAN J J, DAI J, et al. Self-supervised lightweight depth estimation in endoscopy combining CNN and Transformer [J]. IEEE Transactions on Medical Imaging, 2024, 43(5): 1934-1944.
[13] LIU J L, GONG M G, GAO Y, et al. Bidirectional interaction of CNN and Transformer for image inpainting [J]. Knowledge-Based Systems. 2024, 299: 112046.
[14] DUAN Z, LUO X, ZHANG T. Combining transformers with CNN for multi-focus image fusion[J]. Expert Systems with Applications, 2023, 235: 12115.
[15] XU K, REN F. CSVideoNet: A real-time end-to-end learning framework for high-frame-rate video compressive sensing [C]// Proceedings of the IEEE Conference on Computer Vision Pattern Recognition. 2018.
[16] ZHAO Z, XIE X, LIU W, et al. A hybrid-3d convolutional network for video compressive sensing [J]. IEEE Access, 2020, 8: 20503-20513.
[17] CHEN C, ZHOU C, ZHANG D Y. Adaptive Reconstruction for Distributed Compressive Video Sensing Based on Text features. [J]. Chinese Journal of Sensors and Actuators, 2024, 37(1): 58-63.
[18] SHI W, LIU S, JIANG F, et al. Video compressed sensing using a convolutional neural network [J]. IEEE Transactions on Circuits System and Video Technology, 2021, 31(2): 425-438.
[19] YANG J, WANG H X, FAN Y B, et al. VCSL: Video compressive sensing with low-complexity ROI detection in compressed domain [C]// Proceedings of the IEEE Conference on Data Compression. 2023.
[20] YANG J, WANG H X, TANGUCHI I, et al. AVCSR: Adaptive video compressive sensing using region-of-interest detection in the compressed domain [J]. IEEE Multimedia, 2023, 31(1): 19-32.
[21] ZHAO C, MA S, ZHANG J, et al. Video compressive sensing reconstruction via reweighted residual sparsity[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(6): 1182-1195.
[22] ZHONG Y H, ZHANG C X, YANG X, et al. Video compressed sensing reconstruction via an untrained network with low-rank regularization [J]. IEEE Transaction on Multimedia, 2023, 26: 4590-460.
[23] TRAMEL E M, FOWLER J E. Video compressed sensing with multihypothesis [C]// Proceedings of the IEEE Conference on Data Compression. 2011: 193-202.
[24] DU X L, HU X, CHENG B, et al. Multi-hypothesis Reconstruction Algorithm of DCVS Based on Weighted Non-local Similarity [J]. Computer Science, 2019, 46(1): 291-296.
[25] SUN R H, LIU H, DENG K L, et al. Window-adaptive Reconstruction for Low-delay Video Compressive Sensing [J]. Chinese Journal of Beijing University of Aeronautics and Astronautics, 2025, 51(7): 2374-2383.
[26] YING X Y, WANG L G, WANG Y Q, et al. Deformable 3D convolution for Video Super-Resolution [J]. IEEE Signal Processing Letters, 2020, 27: 1500-1504.
[27] PAN Z M, TAN Y L, ZHENG H, et al. Block-based Compressed Sensing of Image Reconstruction Based on Deep Neural Network[J]. Computer Scienc, 2022, 49(S2): 510-518.
[28] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild [J]. arXiv: 2012, 1212, 0402.
[29] KINGMA D P, BA J. Adam: A method for stochastic optimization [C]// Proceedings of the IEEE Conference on International Conference on Learning Representations. 2015.

DU Xiuli, born in 1977, professor, is a member of CCF (No. 22427M). Her main research interests include compressed sensing and EEG signal processing.

ZHU Jinyao, born in 1999, postgraduate. His main research interests include video compressed sensing and so on.

(责任编辑:何杨)