北京航空航天大学学报
Journal of Beijing University of Aeronautics and Astronautics
ISSN 1001-5965,CN 11-2625/V

《北京航空航天大学学报》网络首发论文

题目：基于 Transformer 的深度条件视频压缩
作者：鲁国，钟天雄，耿晶
DOI：10.13700/j.bh.1001-5965.2022.0374
收稿日期：2022-05-18
网络首发日期：2022-10-09
引用格式：鲁国，钟天雄，耿晶．基于 Transformer 的深度条件视频压缩[J/OL]．北京航空航天大学学报. https://doi.org/10.13700/j.bh.1001-5965.2022.0374

基于 Transformer 的深度条件视频压缩

鲁国[1]，钟天雄[1]，耿晶[1,*]

（1. 北京理工大学 计算机科学与技术学院，北京 100081）
*通信作者 E-mail：janegeng@bit.edu.cn

摘要 近年来，基于深度学习的视频压缩技术主要是基于卷积神经网络并且采用运动补偿-残差编码的架构。考虑到常见的卷积神经网络只能利用局部的相关性，以及预测残差本身的稀疏特性，可能难以取得最优压缩性能。因此，我们提出了一种基于 Transformer 架构的条件视频压缩框架，以实现更优的压缩效果。具体来讲，该框架首先基于前后帧之间的运动信息，利用可形变卷积得到对应的预测帧特征。随后将预测帧特征作为条件信息，用来对原始输入帧特征进行条件编码，从而避免了直接编码稀疏的残差信号。更重要的是利用了特征间的非局部相关性，提出了一个基于 Transformer 的自编码器架构用来实现运动信息编码和条件编码，进一步提升了压缩编码的性能。实验证明，提出的基于 Transformer 的深度条件视频编码算法在 HEVC、UVG 数据集上均超越了当前主流的基于深度学习的视频压缩编码算法。

关键词 视频压缩；Transformer；深度学习；神经网络；压缩框架
中图分类号 TP37 文献标志码 A DOI: 10.13700/j.bh.1001-5965.2022.0374

A Transformer Based Deep Conditional Video Compression Framework

LU Guo[1], ZHONG Tianxiong[1], GENG Jing[1,*]

（1. School of Computer Science and Engineering, Beijing Institute of Technology, Beijing 100081, China）
*E-mail: janegeng@bit.edu.cn

Abstract Recent learning-based video compression frameworks are mainly based on the convolutional neural networks(CNN), and adopt the architecture of motion compensation and residual coding. Considering that common CNN can only utilize local correlations, and the sparsity of prediction residual, it is challenging to achieve optimal compression performance. To solve the problems above, this paper proposed a Transformer-based deep conditional video compression framework, which can achieve better compression performance. Specifically, our framework firstly uses deformable convolution to obtain the predicted frame feature based on the motion information between the front and rear frames. Then, the predicted frame feature is used as conditional information to conditionally encode the original input frame feature which avoids the direct encoding of sparse residual signals. More importantly, our framework further utilizes the non-local correlation between the features and proposes a Transformer-based autoencoder architecture to implement motion coding and conditional coding, which further improves the performance of compression. Experiments show that our Transformer based deep conditional video compression framework surpasses the current mainstream learning-based video compression algorithms in both HEVC and UVG datasets.

Key words video compression; Transformer; deep learning; neural network; compression framework

目前，视频数据是互联网流量中的重要组成部分，据相关报道其已经占据 70%的比例，并且该比例还在不断提升。视频压缩是通过利用视频的时空间的信息冗余，从而节省存储空间和传输带宽的关键技术。在过去的几十年来，相关的国际和国内组织提出了一系列视频压缩编码标准，如 H.264[1]、H.265[2]等。其中，目前主流的视频压缩算法使用手工设计的模块单元来减少视频中存在的时空信息

收稿日期：2022-05-18
基金项目：国家自然科学基金（基金号 62102024）
Fund： National Natural Science Foundation of China (62102024)
网络首发时间：2022-10-09 17:35:56 网络首发地址：https://kns.cnki.net/kcms/detail/11.2625.V.20221008.1705.008.html

冗余，如基于块的运动估计和离散余弦变换（Discrete Cosine Transform, DCT）。虽然以上的压缩编码标准已经取得了长足的进步，但是考虑到不断增加的视频数量和种类，设计更为高效的压缩编码算法仍然较为急迫。此外，现有算法不能充分利用大量的数据进行端到端的联合优化，因此其性能也有待进步一步提升。

最近几年基于深度学习的视频压缩技术不断发展，借助深度学习出色的表达能力，当前最新该类算法已经能够与 H.265 实现相似甚至更好的压缩表现。当前主流的基于学习的视频压缩算法普遍采用运动补偿-残差压缩的混合编码架构。例如 Lu 等人提出了可以完全端到端进行优化压缩框架 DVC[3]，采用光流估计网络进行运动估计并且利用了基于 CNN 的自编码器进行压缩运动信息和残差。但是当前主流的基于学习的压缩编码方法主要面临以下两个挑战。首先，当前的自编码器架构普遍是基于卷积神经网络，使用卷积操作搭配非线性激活函数实现数据的非线性变换。然而，由于卷积操作是简单的局部信息的加权求和，且其局限于较小的范围，因此其学习到的变换是可能是局部的、次优的。其次，为了消除时间信息冗余，这些算法一般会基于运动估计得到运动信息并通过运动补偿模块得到当前帧的预测帧，从而仅需使用较低的码率编码当前帧和预测帧之间的像素差值或者是特征差值。但是由于残差信号的稀疏特性，直接对残差信号进行压缩，整体的压缩编码效率可能并不令人满意。

为了解决以上问题，本文提出一种基于 Transformer 模型和条件编码的视频压缩框架（Transformer based video compression, TVC）。具体来讲，本文算法首先通过计算前后帧之间的运动信息，借助可形变卷积生成预测帧特征。为了避免直接编码稀疏的残差信号，本文在编码端将预测帧特征作为条件信息，从而用来提升对原始输入帧特征的编码效率。与此同时，本文在解码端仅仅解码出输入帧特征和预测帧特征的残差，从而使网络更容易进行训练和提升重建效率。更进一步的，本文提出了一个基于 Transformer 的自编码器架构用来实现运动信息编码和残差的条件编码，从而能够充分利用 Transformer 的非局部的表达能力，进一步增强压缩编码的性能。实验结果表明，本文提出的算法在标准数据集 HEVC 和 UVG 上，能够超越当前主流的基于深度学习的视频压缩编码算法和传统压缩编码算法如 H.265[2]。

本文的主要贡献有以下三点：

1.本文提出了一种以 Transformer 为主要组成单元的可端到端优化的深度学习视频压缩框架。

2.本文提出了一个针对残差信息的条件编码方案，从而避免了直接编码稀疏的残差信息，进一步提升了残差压缩效率。

3.本文的方法在多个标准数据集上进行了验证，其性能超越了当前的主流压缩算法。

1 相关工作

1.1 基于深度学习的图像压缩编码

传统的图像压缩标准如 JPEG[4]、BPG[5]等是当前主流的图像压缩算法，在各类场景中已经得到了广泛应用。由于传统算法普遍是基于手工设计的特征，因此其性能有待进一步提升。

近年来，基于神经网络图像压缩算法得到了越来越多的关注。其中一些基于 RNN 的图像压缩算法陆续[6-8]被率先提出。目前，主流的大部分的方法均基于自编码器形式的 CNN 网络架构[9-11]。其中，Ballé 等人提出的基于超先验模块的方法[10]，可以实现与 H.265 的帧内编码相近的压缩性能。Minnen 等人提出的自回归上下文模块的熵编码算法[11]则进一步提升了压缩性能，使其超越了 H.265 的帧内编码性能。[12]提出使用通道自回归熵模型，改善了自回归模块的串行计算问题，并进一步提升了编码性能。近年来，许多研究[13-17]将 Transformer 引入图像压缩任务。[14]将 Transformer 用于建模上下文模块，取得了比使用基于卷积的上下文模块的模型[11]更优的压缩性能。[15]提出了一种基于 Transformer 的熵模型，能够捕获概率分布估计中的长期依赖关系，有效提升压缩性能。另外一些研究[13,16,17]使用 Transformer 作为图像压缩的骨架网络，替代了编码器或解码器中使用卷积的结构，并取得了显著优于 CNN 网络的成果。

1.2　基于深度学习的视频压缩编码

随着基于神经网络的图像压缩编码的发展，越来越多的基于神经网络的视频压缩编码算法被提出。目前，大部分基于深度学习的视频压缩方法可以被分为两类。其中一类方法主要是面向 P 帧编码，也就是仅仅依靠前向的参考帧。例如，Lu 等人提出 DVC[3]，替换了传统视频编码流程中的全部模块，第一次实现了端到端优化的深度视频压缩网络。Hu 等人提出了自适应分辨率的光流压缩方法[18]，对不同帧的不同块分别设定合适的分辨率。Agustsson 等人[19]提出在光流中增加尺度参数，从而更好地拟合运动信息的不确定性。以上方法均在像素域进行运动估计和运动补偿，Lu 等人[20]则进一步提出利用可形变卷积在特征空间进行更充分的相关操作。Li 等人提出 DCVC[21]，通过定义并使用条件信息，优化了网络的压缩性能，取得了在多个数据集上的最优编码效率。另外一类算法主要是面向 B 帧编码，因此其可以同时利用前后不同时刻的图像作为参考帧进行预测。如[22-24]均同时使用待编码帧的前一帧和后一帧信息作为参考，并对残差信息进行压缩。此外，[25-26]则采用更直接的方式对现有图像压缩算法进行拓展，利用 3 维自编码器对图片成组压缩。

以上的算法虽然取得了一些进展，但是绝大部分算法都是依赖运动补偿和残差编码的架构，因此对于稀疏的残差压缩，其性能有待进一步提升。虽然 DCVC[21]算法利用了条件信息，但是其在解码端直接预测原始图像，给网络的训练和学习带来了很大挑战。此外，Transformer 在众多领域均展现出超越 CNN 网络的表现，然而相关研究在深度视频编码领域仍留有空白。本文提出的框架通过使用 Transformer 作为骨架网络，在特征空间进行运动估计和运动补偿，并利用条件编码代替残差编码，从而取得了超越基于 CNN 的深度视频编码方法的表现。

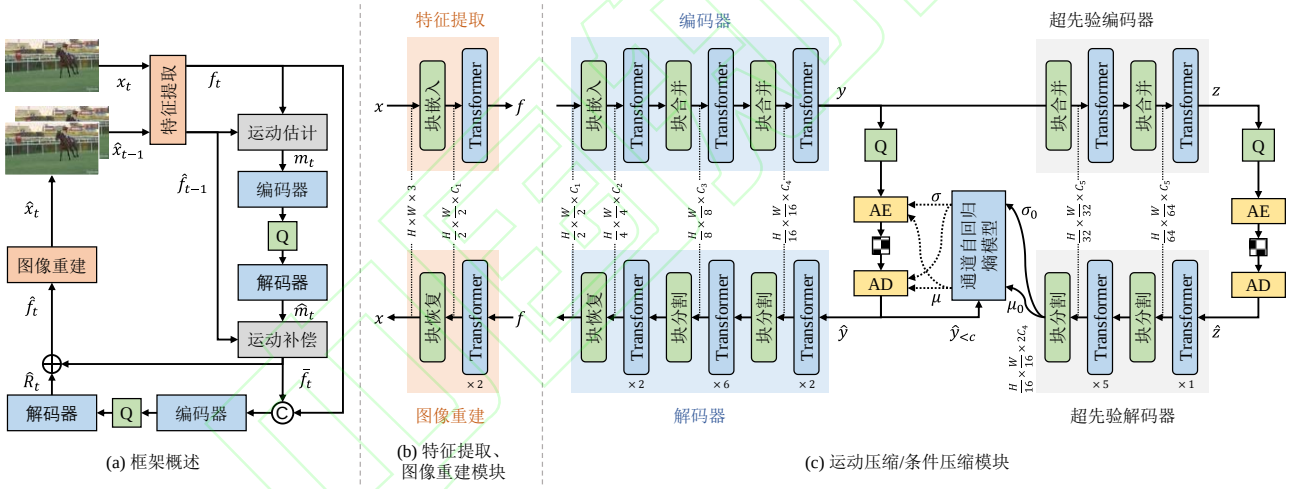

图 1　本文提出的基于 Transformer 的深度条件视频压缩框架结构图
Fig.1　Structure of the proposed Transformer based deep conditional video compression framework

2 方法

2.1　概述

图 1(a)展示了本文提出的压缩编码框架的整体结构图。对于t时刻输入的图像x_t，$t-1$时刻的重建帧图像$\hat{x}_{t-1}$会作为参考帧。本文首先通过特征提取器分别提取x_t 和$\hat{x}_{t-1}$对应的特征f_t和$\hat{f}_{t-1}$。随后，本文提出的框架利用运动估计模块生成特征f_t 和$\hat{f}_{t-1}$之间的运动信息m_t，并且采用运动压缩模块进行压缩然后得到重建后的运动信息$\hat{m}_t$。本文采用基于可形变卷积的运动补偿模块将参考帧特征进行映射得到预测特征$\bar{f}_t$，更多细节参考章节 4.1。随后，在编码端本文以预测特征$\bar{f}_t$作为条件信息，对原始输入帧特征f_t进行压缩。在解码端，条件解码器仅仅输出残差信息$\hat{R}_t$，然后将其预测特征$\bar{f}_t$相加，以重建更准确的输入帧特征$\hat{f}_t$。最后经过图像重建模块，将$\hat{f}_t$恢复为重建图片$\hat{x}_t$。

为了实现码率和失真的最优平衡，本文提出的整个模型以如下方式进行优化：

$$RD = R + \lambda D = R_m + R_c + \lambda d(x_t, x_t) \tag{1}$$

其中，R_m和R_c分别表示运动编码和条件编码消耗的码率，$d(x_t, \hat{x}_t)$表示重建帧与原始输入帧之间的损失，λ为超参数，用于控制优化过程中码率和损失之间的平衡。在运动编码和条件编码过程中，量化后的特征将通过熵编码转化为码流。类似的，在解码过程中需要通过熵解码将码流恢复为特征。

如图 1(c)所示，为了平衡压缩性能和编解码速度，我们引入了基于 Transformer 的超先验网络和通道自回归熵模型[12]用于快速而准确的估计分布的均值和方差，用于上述熵编码和熵解码过程。其中，AE 和 AD 是隐含的熵编码和熵解码模块，Q 是量化过程，编码器和解码器的结构是对称的，解码器下方标注了每层 Transformer 的数量，并使用虚线标注了图片或特征形状的变化。在训练阶段，我们使用了一个码流估计网络来估计消耗的码率。在测试阶段，我们使用了实际的熵编码算法。考虑到量化本身不可微分，我们采用在训练过程中加入均匀噪声进行近似的方法来保证端到端的优化。

2.2 条件编码

图 1(a)中展示了本文提出的基于条件编码的视频压缩编码框架。本文的方法不再依赖于传统的残差压缩方法，而是基于高效的条件信息对原始图像（特征）进行直接编码。因此，如何得到高效的条件信息就显得十分重要。为此，本方案将前一帧的特征$\hat{f}_{t-1}$变换到当前时刻得到预测特征$\bar{f}_t$，并且其作为条件信息使用。

具体来讲，本文首先基于一个双层卷积神经网络构成的运动估计模块来计算当前帧特征f_t和前一帧特征$\hat{f}_{t-1}$的运动信息m_t，随后本文将运动信息进行压缩编码，解码后的运动信息为$\hat{m}_t$。本框架借鉴了可形变卷积在运动补偿中的成功应用，同样基于可形变卷积将预测帧特征$\bar{f}_t$对齐到当前时刻。首先，卷积层将$\hat{m}_t$转化为分组的偏移量信息，每组的通道之间共享相同的偏移量。对于可形变卷积中的每个卷积核，存在与每个位置对应的运动信息，用于控制从参考帧特征中采样的位置。接着，参考帧特征上对应位置的值被可形变卷积融合为预测特征上的一个值，如公式(2)所示，

$$\bar{f}_t(p_0) = \sum_{p_n \in K^2} w(p_n) \cdot f_{t-1}(p_0 + p_n + \Delta p_n) \tag{2}$$

其中，K^2表示卷积核中的每个位置，在实现中为$\{-1,0,1\}^2$，$w(p_n)$为对应位置的权重；p_0和Δp_n分别表示特征位置及对应的偏移量。

基于得到条件信息，也就是预测特征$\bar{f}_t$，本文将$\bar{f}_t$和原始帧特征f_t进行拼接输入到编码器，对应的量化后的隐空间特征是$\hat{y}_t$。在解码端，基于解码得到残差特征$\hat{R}_t$和原始条件特征$\bar{f}_t$，本文进行了一个加法操作，得到最终的重建原始特征$\hat{f}_t$，整个过程如下所示，

$$\begin{aligned} y_t &= Q(Encoder([f_t, \bar{f}_t])) \\ f_t &= Decoder(y_t) + \bar{f}_t \end{aligned} \tag{3}$$

其中，$Encoder$和$Decoder$分别为编码器和解码器，Q为量化操作。具体的编码器和解码器架构在章节 4.2 将进行介绍。

2.3 基于 Transformer 的压缩模型

在本文提出框架中，我们需要对运动信息m_t以及原始特征进行编码。为了进一步提高压缩效率，本文引入了 Transformer 来作为基础模块构建一个自编码器，整体的架构如图 1(c)所示。相比传统的 CNN 模型，Transformer 能够更进一步利用特征的相关性，从而进一步消除压缩的冗余，获得更好的压缩效率。此外，为了进行特征的上采样和下采样，我们还引入了块合并和块分割层（图 2(c)），通过空间维度和通道维度的变换实现了空间尺寸的变化。

为了在确保 Transformer 长距离依赖特性的前提下，降低其过高的计算代价，我们提出的模块使用了基于滑动窗口的 Swin-Transformer[27]。如图 2(a)所示，Swin-Transformer 块首先将嵌入特征分割为尺寸为$M \times M$互不重叠的窗口，在每个窗口内部运用自注意力机制进行计算，从而实现了线性计算复杂度。然后将窗口重新恢复为特征图，经过多层感知机对特征图进一步变换。在相邻的 Swin-Transformer 层之间，会进行一次窗体向右下方的滑动，距离为$M/2$，从而保证了信息能够跨窗体流

动，保留了 Transformer 长距离依赖的优势。考虑传输到编（解）码器和超先验编（解）码器的特征图尺寸不同，在实现中使用了不同的

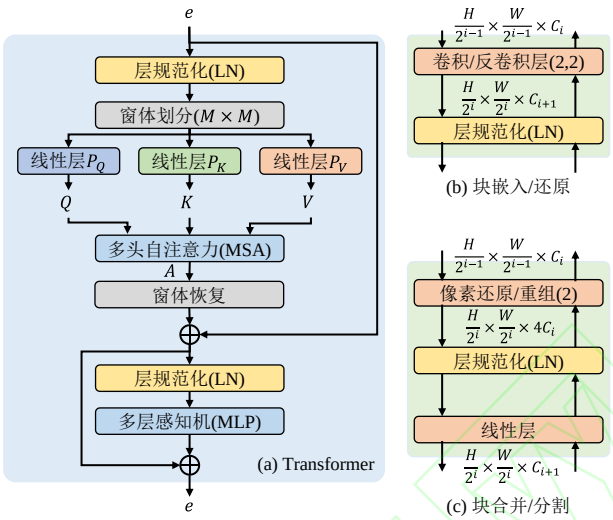

图 2 Transformer 块及相关模块结构
Fig.2 Structure of Transformer block and relevant modules

窗口大小，分别设置为 8 和 4。对于第n个窗口$e(n)$，其在自注意力机制中首先经过不同线性层生成对应的$Q(n)$、$K(n)$和$V(n)$，如公式(4)所示，

$$\begin{aligned} Q(n) &= e(n)P_Q \\ K(n) &= e(n)P_K \\ V(n) &= e(n)P_V \end{aligned} \tag{4}$$

其中P_Q、P_K和P_V在不同的窗口之间共享权重。接着，多头自注意力机制如公式(5)所示，

$$A = SoftMax(QK^T / \sqrt{d} + B)V \tag{5}$$

其中，B是可学习的相对位置编码，d是多头自注意力机制中每个头的通道数，在本文中设置为 32，即第i组 Transformer 中的头数为$C_i/32$。

3　验证

3.1　实验设置

3.1.1　训练集

本文使用 Vimeo-90K[28]数据集作为训练集。该数据集包含 89800 个 7 帧、448×256 分辨率的视频序列。在训练中，随机地将视频序列裁剪为大小为 256×256 视频序列。

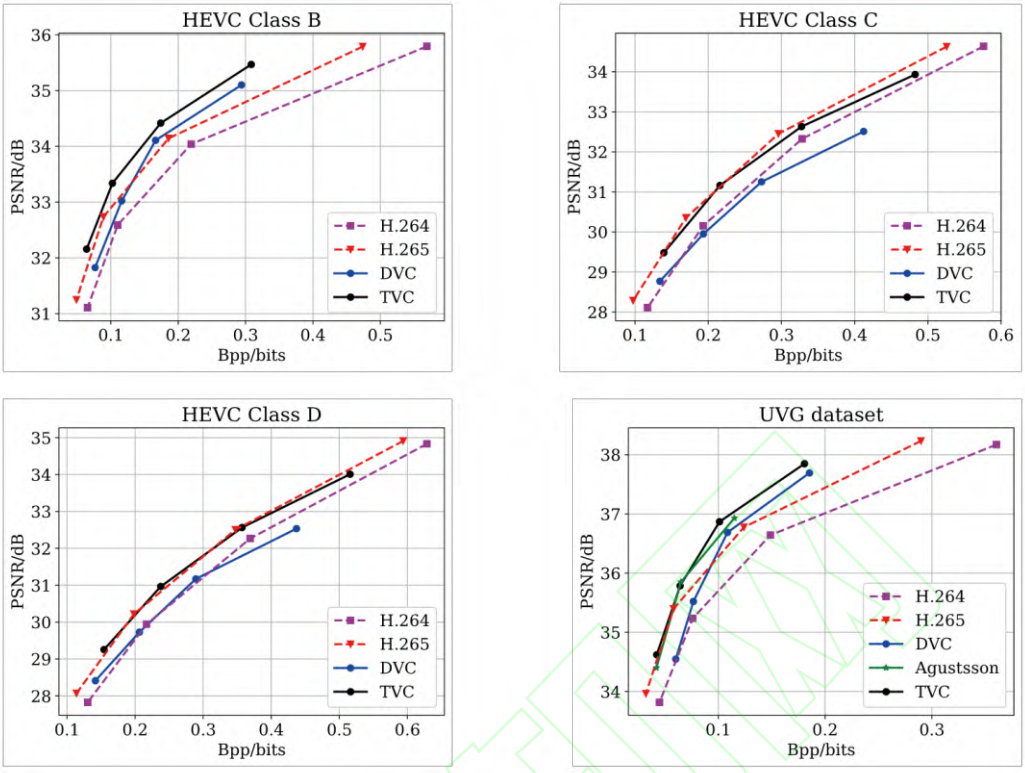

图 3　在 HEVC、UVG 数据集上的不同压缩算法的实验结果
Fig.3　Results for different compression algorithms on HEVC and UVG datasets

3.1.2　测试集

在 HEVC[2]和 UVG[29]数据集上评估了本文提出的算法性能。其中，HEVC 被划分为 B、C、D、E 四个类别，包含共 16 个不同分辨率的视频序列。UVG 则包含分辨率为 1920×1080 的 7 个高帧率视频序列。从而我们的测试集涵盖了多种分辨率，帧间差距较大和较小的视频，以全面的评估框架的性能和鲁棒性。

3.1.3　评估指标

本文使用每个像素消耗的比特（Bpp）作为运动编码和条件编码过程中平均码率的衡量标准。我们使用峰值信噪比（PSNR）来衡量重建图像与原始输入图像之间的损失。

3.1.4　实现细节

在训练过程中，使用了四种不同的λ取值（$\lambda = 256, 512, 1024, 2048$），并采用了 6 阶段的训练策略。每个阶段训练的轮数和损失函数如表 1 所示，其中$d(x_t, \bar{x}_t)$表示计算原始输入图像与预测特征还原图像之间的损失，在实验中使用均方误差（MSE）表示。

表1 不同训练阶段的损失函数
Table 1　Loss function at different training stage

阶段	步数/万	损失函数
1	10	$\lambda d(x_t, \bar{x}_t)$
2	10	$R_m + \lambda d(x_t, \bar{x}_t)$
3	20	$\lambda d(x_t, \hat{x}_t)$
4、5、6	10、140、10	$R_c + \lambda d(x_t, \hat{x}_t)$

在第 1 阶段，优化与运动信息相关的模块，使预测的条件信息尽可能与输入的当前帧接近。第 2 阶段，对运动信息编码的码率加以约束，从而平衡运动信息的码率代价和重建质量。第 3 阶段冻结运动估计、运动压缩和运动补偿模块参数，优化网络的剩余参数，仅约束网络的重建质量。第 4 阶段在第 3 阶段的基础上，加以对条件编码的码率约束。第 5 阶段，解冻运动模块的参数，整个网络端到端训练。最后，在第 6 阶段，冻结量化操作前的全部网络参数，并使用取整的量化方法对其他参数进一

步训练。其中，前 5 个阶段使用 5e-5 的学习率，并在最后一个阶段降低为 5e-6 的学习率。训练总共经过 2,000,000 步。

如图 1(b)所示，特征提取模块由一个块嵌入层和两层 Transformer 层构成，图像还原模块则由两层 Transformer 和一个块恢复构成，与编码器保持对称的结构。其中，块嵌入使用步长、卷积核都为 2 的卷积层实现，将输入的图像分割成许多2×2互不重叠的块并分别编码为特征向量；块恢复是块嵌入的逆过程，使用了同样配置的反卷积层。块嵌入与块恢复层的结构见图 2(b)。此外，实验中图 1(c)中所示自编码器使用的通道数设置为$(C_1, \dots, C_6) = (64,128,160,192,96,128)$。

3.2　实验结果

本章节将对本文提出的框架与传统编码算法 H.265[2]以及基于卷积和残差编码实现的视频压缩编码算法 DVC[3]和 Agustsson 等人的方法[19]进行对比。为了得到 H.265 的压缩结果，我们和参考文献 [3]保持一致，使用 FFmpeg 的默认模式（default）。为确保实验的公平性，基于深度学习的方法均使用基于 MSE 损失函数进行优化，所有方法在测试的时候都将 GoP 设置为 10，并使用与 H.265[2]相同的方法重建 I 帧。表 2 展示了 BDBR 结果。

图 3 展示了我们的框架与基线方法率失真曲线，显然，本文的方法在 HEVC[2]的大部分子数据集上，均超越了现有的基于传统算法和基于深度学习的视频编码框架。特别是在 HEVC 的 B 类数据集上，我们的方法相比 DVC[3]和 H.265[2]在 PSNR 指标上分别提升了约 0.40dB 和 0.35dB 的表现。另外，在 UVG 数据集上，我们的方法相较 DVC[3]、Agustsson 等人的方法[19]和 H.265[2]，分别提升了约 0.45dB、0.10dB 和 0.31dB。

3.3　消融实验与模型分析

为了验证本文提出框架中使用 Transformer 代替卷积，以及使用条件编码代替残差编码的有效性，本文以 HEVC 数据集为例，进行消融实验

如表 2 所示，我们实现了基于 Transformer 使用残差编码的框架 TVC(Res)，其仅将本文框架中基于条件信息的编码修改为对残差信息编码的方法。其中，基线方法为 H.264[1]。可以看到，在类似的重建质量下，我们的框架 TVC 相比基于残差编码的 TVC(Res)平均能够节省大约 4.1%的码率，从而证明了条件编码相对于残差编码的优越性。其次，我们还实现了 DVC[3]作为典型的基于卷积和残差编码的深度视频编码框架。同样使用残差编码的 TVC(Res)则相比 DVC 能够节省大约 8.8%的码率，从而验证了 Transformer 相较卷积层能够实现更有效的变换。Transformer 与条件编码的联合使用，使得提出的框架 TVC 相比 DVC 能够节省超过 12.8%的码率。

表2　HEVC和UVG数据集BDBR结果

Table 2　BDBR results on HEVC and UVG datasets

数据集	H.265[2]	DVC[3]	TVC	TVC(Res)
HEVC B	-21.95	-18.18	-33.94	-28.85
HEVC C	-14.48	7.07	-10.02	-7.94
HEVC D	-12.40	0.80	-11.40	-7.04
HEVC E	-30.81	-28.56	-30.56	-26.33
UVG	-26.07	-19.39	-35.83	-31.37

本文提出的框架总共含有约 45M 的参数，在单张的 V100（16GB 内存）显卡上，编码 1920×1080 分辨率的视频帧大约需要 324ms，解码则大约需要 272ms，尽管稍慢于 DVC 的 276ms 和 189ms，但在大部分数据集上，本文提出的框架的压缩表现均优于 DVC。

4　结论

本文提出了一个全新的基于 Transformer 和条件编码的端到端的视频压缩框架，从而实现了比卷积网络更充分的变换，同时利用条件编码避免了直接对稀疏残差信息直接进行压缩。实验表明，我们的方法在 HEVC 数据集上优于目前主流的基于传统算法和基于卷积和残差编码的深度算法。

参考文献（References）

[1] WIEGAND T,SULLIVAN G J,BJONTEGAARD G,et al. Overview of the H. 264/AVC video coding standard[J].IEEE Transactions on circuits and systems for video technology,2003,13(7):560-576.

[2] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding (HEVC) standard[J].IEEE Transactions on circuits and systems for video technology,2012,22(12):1649-1668.

[3] LU G,OUYANG W,XU D,et al.Dvc:An end-to-end deep video compression framework[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2019:11006-11015.

[4] WALLACE G K.The JPEG still picture compression standard[J]. IEEE transactions on consumer electronics,1992,38(1):xviii-xxxiv.

[5] BELLARD F.BPG image format[EB/OL].URL https://bellard.org/bpg,2015.2.

[6] TODERICI G,O'MALLEY S M,HWANG S J,et al.Variable rate image compression with recurrent neural networks[J].arXiv preprint arXiv:1511.06085,2015.

[7] TODERICI G,VINCENT D,JOHNSTON N,et al.Full resolution image compression with recurrent neural networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.2017:5306-5314.

[8] JOHNSTON N,VINCENT D,MINNEN D,et al.Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.2018:4385-4393.

[9] BALLÉ J,LAPARRA V,SIMONCELLI E P.End-to-end optimized image compression[J].arXiv preprint arXiv:1611.01704,2016.

[10] BALLÉ J,MINNEN D,SINGH S,et al.Variational image compression with a scale hyperprior[J].arXiv preprint arXiv:1802.01436,2018.

[11] MINNEN D,BALLÉ J,TODERICI G D.Joint autoregressive and hierarchical priors for learned image compression[J].Advances in neural information processing systems,2018,31.

[12] MINNEN D,SINGH S.Channel-wise autoregressive entropy models for learned image compression[C]//2020 IEEE International Conference on Image Processing (ICIP).IEEE,2020:3339-3343.

[13] ZHU Y,YANG Y,COHEN T.Transformer-based Transform Coding[C]//International Conference on Learning Representations.2021.

[14] KOYUNCU A B,GAO H,STEINBACH E.Contextformer:A Transformer with Spatio-Channel Attention for Context Modeling in Learned Image Compression[J].arXiv preprint arXiv:2203.02452,2022.

[15] QIAN Y,LIN M,SUN X,et al.Entroformer:A Transformer-based Entropy Model for Learned Image Compression[J].arXiv preprint arXiv:2202.05492,2022.

[16] LU M,GUO P,SHI H,et al.Transformer-based Image Compression[J].arXiv preprint arXiv:2111.06707,2021.

[17] BAI Y,YANG X,LIU X,et al.Towards End-to-End Image Compression and Analysis with Transformers[J].arXiv preprint arXiv:2112.09300,2021.

[18] HU Z,CHEN Z,XU D,et al.Improving deep video compression by resolution-adaptive flow coding[C]//European Conference on Computer Vision.Springer,Cham,2020:193-209.

[19] AGUSTSSON E,MINNEN D,JOHNSTON N,et al.Scale-space flow for end-to-end optimized video compression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2020:8503-8512.

[20] HU Z,LU G,XU D.FVC:A new framework towards deep video compression in feature space[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2021:1502-1511.

[21] LI J,LI B,LU Y.Deep contextual video compression[J].Advances in Neural Information Processing Systems,2021,34.

[22] WU C Y,SINGHAL N,KRAHENBUHL P.Video compression through image interpolation[C]//Proceedings of the European conference on computer vision (ECCV).2018:416-431.

[23] DJELOUAH A,CAMPOS J,SCHAUB-MEYER S,et al.Neural inter-frame compression for video coding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision.2019:6421-6429.

[24] YANG R,MENTZER F,GOOL L V,et al.Learning for video compression with hierarchical quality and recurrent enhancement[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2020:6628-6637.

[25] PESSOA J,AIDOS H,TOMÁS P,et al.End-to-end learning of video compression using spatio-temporal autoencoders[C]//2020 IEEE Workshop on Signal Processing Systems(SiPS).IEEE,2020:1-6.

[26] HABIBIAN A,ROZENDAAL T,TOMCZAK J M,et al.Video compression with rate-distortion autoencoders[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision.2019:7033-7042.

[27] LIU Z,LIN Y,CAO Y,et al.Swin transformer:Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision.2021:10012-10022.

[28] XUE T,CHEN B,WU J,et al.Video enhancement with task-oriented flow[J].International Journal of Computer Vision,2019,127(8):1106-1125.

[29] MERCAT A,VIITANEN M,VANNE J.UVG dataset:50/120fps 4K sequences for video codec analysis and development[C]//Proceedings of the 11th ACM Multimedia Systems Conference.2020:297-302.