2022 GUIYANG CHINA MM

基于 *Transformer* 的深度条件视频压缩

A Transformer Based Deep Conditional Video Compression Framework

鲁国，钟天雄，耿晶

出发点

主流的自编码器架构普遍是基于卷积神经网络，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**。

主流算法一般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，编码当前帧和预测帧之间的像素差值或者是特征差值，以消除时间信息冗余。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**。

主要贡献

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。

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。

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**。

方法

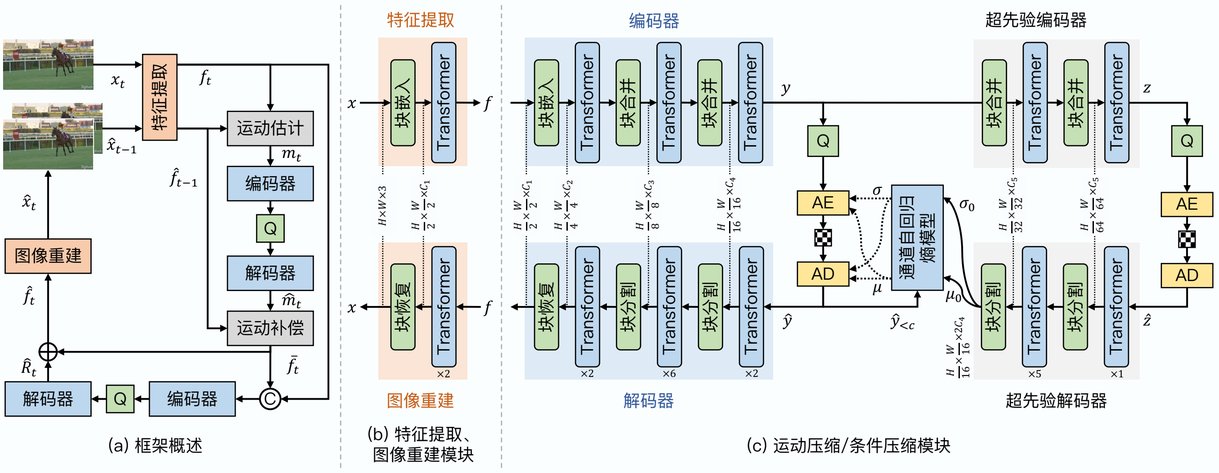


图1 本文提出的基于Transformer的深度条件视频压缩框架结构图。(a) 框架顶层结构图，网络使用原始输入帧特征f_t和运动补偿得到的条件信息预测帧特征$\bar{f}_t$作为条件压缩的输入，并输出残差与预测帧特征相加重建该帧特征。(b) 图(a)中特征提取模块和图像重建模块。(c) 图(a)中的条件压缩模块和运动压缩模块，其中AE和AD是隐含的熵编码和熵解码模块，Q是量化过程。其中，图(c)中编码器和解码器的结构是对称的，解码器下方标注了每层Transformer的数量，并使用虚线标注了图片或特征形状的变化。

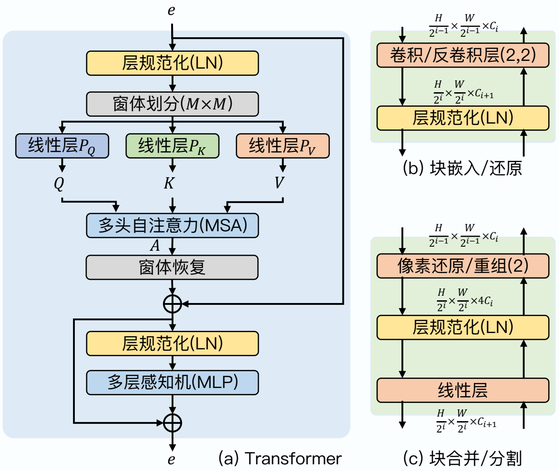
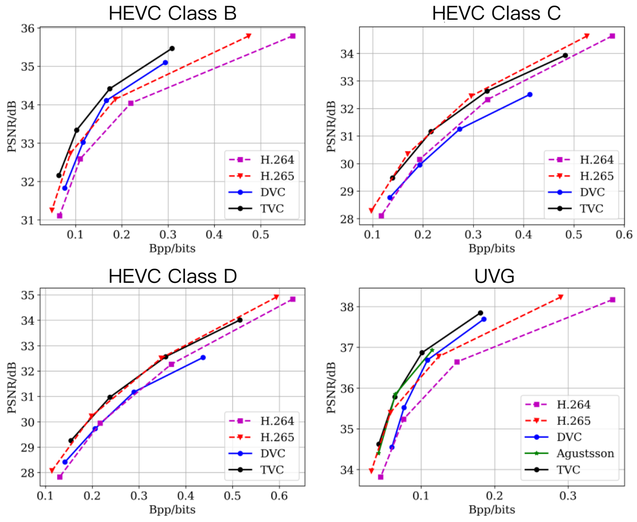


图2 (a) Swin-Transformer单层的结构，相邻的层之间在窗体划分时会向右下方移动距离$M/2$，以确保长距离依赖。(b) 块嵌入模块和块还原模块的结构。(c) 块合并和块分割模块的结构。图(b)和(c)中的角标i与图1(a)对应。

1. 基于Transformer的**特征提取**
2. 基于Transformer的编码器和解码器对运动信息进行充分的**非线性变换**，同时依赖超先验编解码器和通道自回归熵模型准确高效的**预测潜在表示的分布**，**并估计编码消耗的比特数*R***
3. 基于可形变卷积实现**运动补偿**
4. 将**预测特征**作为当前帧特征编码的**条件信息**，一并输入基于Transformer的编码器，并输出预测特征与当前帧特征之间的残差信息
5. 将**残差信息与预测帧特征相加**获得重建帧特征
6. 基于Transformer的**图像重建**
7. 基于$\boldsymbol{R+\lambda D}$优化网络，其中D是重建图像和原始图像之间的失真，λ为率失真平衡参数

实验结果



HEVC和UVG数据集BDBR结果，其中基线方法为H.264[1]，TVC(Res)为将条件编码替换为残差编码的TVC框架

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

本文的方法在HEVC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和H.265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、Agustsson等人的方法和H.265，分别提升了约**0.45dB**、**0.10dB**和**0.31dB**。