DOI:10.19651/j.cnki.emt.2415530

结合帧间差异检测的固定场景视频压缩与重建*

李 萌[1] 黄宏博[1,2] 郑曜林[1] 许龙飞[1]
(1.北京信息科技大学计算机学院 北京 100101;2.北京信息科技大学计算智能研究所 北京 100192)

摘 要:近年来,高清和超高清监控摄像头的广泛部署促使了各类监控等固定场景类视频数据量的急剧增加。对视频的存储和传输造成了巨大压力。为了进一步去除固定场景类视频中的冗余数据,本文提出了一种新颖的压缩与重建方法。通过背景提取和结合帧间前景差异检测的前景提取与压缩方法,大量去除视频中的数据冗余。实验结果表明,本文方法与 MPEG-4 相比,在更高的压缩率上实现了更高的视频重建性能,与 H.264、H.265 和 DCVC-DC 相比,本文所提方法在压缩性能上依次分别提升了 82.75%、76.19%和 59.56%,并且保持了较高的视频重建水平,从而有效地缓解了固定场景类视频的存储和传输压力。

关键词:计算机视觉;深度学习;视频压缩;图像分割;背景建模
中图分类号:TP391;TN911 **文献标识码**:A **国家标准学科分类代码**:520.20

Integrated inter-frame difference detection for fixed-scene video compression and reconstruction

Li Meng[1] Huang Hongbo[1,2] Zheng Yaolin[1] Xu Longfei[1]
(1. School of Computer Science, Beijing Information Science and Technology University, Beijing 100101, China;
2. Institute of Computational Intelligence, Beijing Information Science and Technology University, Beijing 100192, China)

Abstract: In recent years, the widespread deployment of high-definition and ultra-high-definition surveillance cameras has led to a significant increase in the volume of fixed-scene video data, such as surveillance videos. This sharp rise in data has imposed tremendous pressure on video storage and transmission. To further eliminate redundancy in fixed-scene videos, this paper proposes a novel compression and reconstruction method. By employing background extraction and an inter-frame foreground difference detection-based foreground extraction and compression approach, a substantial amount of data redundancy is removed from the videos. Experimental results show that, compared to MPEG-4, the proposed method achieves higher video reconstruction performance at a higher compression ratio. Compared to H.264, H.265, and DCVC-DC, the proposed method improves compression performance by 82.75%, 76.19%, and 59.56% respectively, while maintaining a high level of video reconstruction quality. This effectively alleviates the storage and transmission pressure of fixed-scene videos.

Keywords: computer vision;deep learning;video compression;image segmentation;background modeling

0 引 言

随着城市的现代化发展,交通、安防以及环境监测等各种视频监控在交通管理、城市安全、公共安全等领域发挥着越来越重要的作用。随着高清和超高清摄像头等视频采集设备的广泛部署,导致视频数据量急剧膨胀,对视频的存储和传输带来了日趋增长的压力。如何更好地利用视频中的关键特征进行视频编码,使用更少的数据量记录更丰富的视频内容,减轻视频的存储和传输压力,是视频压缩领域需要解决的重要任务。

视频压缩的可行性源于视频数据中存在大量的冗余数据。传统的视频编码方法,如 MPEG-4、H.264、H.265 等方法主要通过手工设计模块对边缘、纹理、图像块运动等特征进行处理,以减少空间和时间冗余。基于深度学习的视频编解码方法,例如 DVC[1] 和 DCVC 系列[2-5] 等,大都是借鉴传统方法的思路,并基于深度神经网络设计视频编解码方案,利用深度神经网络强大的特征提取能力来提升视频编码性能。这些方法虽然在视频压缩任务上取得了较好的

收稿日期:2024-02-25
*基金项目:国家自然科学基金(62376286)项目资助

性能，但都是针对通用视频，未能充分利用固定场景类视频的特点。因此，在提升这类视频的压缩率方面仍有很大的挖掘和提升空间。

与通用视频相比，固定场景类视频采集视角较为固定，且通常较少切换画面，因此背景变化程度较小。如果在所有视频帧中提取出少量有代表性的背景关键帧来压缩时间维度的数据，则能够大量去除背景数据冗余。另一方面，在视频画面中主要关注的是行人和车辆等前景目标。如果能够检测并存储视频中变化显著的前景目标，则能够在保证视频完整性的前提下，大幅度降低视频数据冗余。因此，本文针对固定场景类视频的特点，分别对前景目标和背景独立地进行提取和存储。对于前景目标，本文利用深度卷积网络对每帧视频进行精确提取与分割，并利用前景运动的先验知识大量去除静止不变的前景目标，从而更高效地去除视频中大量的冗余数据，以更大的压缩比实现固定场景视频的大幅度压缩。对于视频背景，本文利用背景建模方法结合深度卷积网络提取到的前景位置和范围，精确地分离每一帧的背景，并从所有背景中提取出少量的背景关键帧。在大量去除背景数据冗余的同时，准确地还原视频场景。

1　相关工作

1.1　视频压缩

传统方法在视频压缩任务上取得了良好的压缩和重建性能。但随着视频数据量的迅速增长，视频压缩算法面临着诸多新的挑战。近年来，深度学习在计算机视觉领域得到了广泛应用，为视频压缩任务的进一步发展提供了一个新的方向。基于深度学习的视频压缩方法通常借鉴传统方法的流程框架，并与深度学习相结合进行视频的压缩与重建。Wu 等[6]首次尝试将深度学习应用于视频压缩任务，提出了一种基于深度学习的图像生成的方法，在视频关键帧间进行插值的方式实现视频编解码，并取得了和传统方法相似的性能，但在视频重建流畅度和重建精度方面存在不足。DVC[1]则利用神经网络替换传统方法流程中的各个关键组件，以强化特征提取能力，并取得了更好的性能，但方法的流程仍受制于传统方法框架。DCVC[3]提出了一种基于条件编码的深度视频压缩方法，利用特征域上下文来编码视频，相比取得了更好的编解码性能。为了解决 DCVC 时间特征利用不足的问题，DCVC-TCM[4]引入了时间上下文特征，并利用分层结构形成多尺度时间上下文，进一步提升了压缩性能。DCVC-HEM[5]则充分地利用时空相关性，通过设计一种熵模型进一步提升了压缩性能。为了解决单一上下文特征提取存在局限性的问题，DCVC-DC[2]提出了一种融合多种上下文的方法，提取视频中的多种关键信息，大幅提升了视频压缩性能。这些基于深度学习的视频压缩方法借鉴了传统方法的思路，并利用神经网络强化了特征提取能力，达到了更好的视频编解码效果。但这类方法仍然受制于传统方法框架，难以从整体上优化视频编解码性能。

1.2　目标检测和实例分割

本文方法基于背景与前景的解耦，因而对前景目标进行精确定位与分割时本文视频压缩方法的关键步骤。

目标检测和实例分割是计算机视觉领域的两个重要任务。现有方法主要分为单阶段[7-9]和双阶段[10-13]两大类。单阶段方法在模型推理速度上表现更为出色，然而在分割精度方面通常不及双阶段方法。相反，双阶段方法虽然在分割精度上占优，但其推理速度相对较慢。近年来，Transformer[14]在计算机视觉领域的广泛应用，为目标检测和实例分割提供了新的方向，催生了众多基于 Transformer 的方法，例如 DETR[15]、DINO[16]、MaskFormer[17]和 Mask2Former[18]等。得益于 Transformer 强大的特征提取能力，这类方法在检测精度上显著超越了基于卷积网络的方法。DINO 通过混合查询选择和对比去噪训练等改进，在 COCO 目标检测排行榜上取得了基于 DETR 模型类方法的最好效果。在此基础上，Mask DINO[19]扩展了 DINO 目标检测模型，添加了支持所有图像分割任务的 Mask 预测分支，统一了目标检测、实例分割和全景分割任务。前景目标分割精度对于解耦前景和背景至关重要，高精度目标检测与实例分割方法更适用于本文场景。

1.3　背景关键帧提取与背景建模

背景建模的目标是从视频中提取相对固定的部分作为视频内容的场景分量，从方法上大致可以分为非递归和递归两类。非递归类方法，例如帧差分法、统计直方图法和非参数核密度估计方法等，通常利用统计模型来建立背景模型。这种方法简单易实现，但空间复杂度相对较高，且难以适应复杂场景。递归类的方法，例如单高斯模型、高斯混合模型、码本算法和 MOG2 等，通过不断更新背景模型适应场景变化。这种方法通常具有较低的空间复杂度，且能更好的适应复杂的动态场景。背景建模方法的准确度以及方法的复杂场景适应性，对于背景和前景的解耦以及背景关键帧的提取至关重要。因此，本文主要采用递归类方法来获取视频背景。

2　视频压缩与重建

2.1　概　　述

为了充分利用固定场景视频的特点，进一步去除视频中的冗余数据，同时保持较高的视频重建性能，本文提出了一种结合帧间前景差异检测的固定场景视频压缩与重建方法。该方法主要分为前景目标提取与压缩、背景关键帧提取和视频重建 3 个部分，方法框架如图 1 所示。

视频压缩流程分为前景目标提取与压缩和背景关键帧提取两部分。对于前景目标，本文首先获取所有视频帧中前景目标的位置、区域以及运动轨迹，之后利用帧间前景差异检测模块去除上下帧间没有发生位置和姿态变化的前景

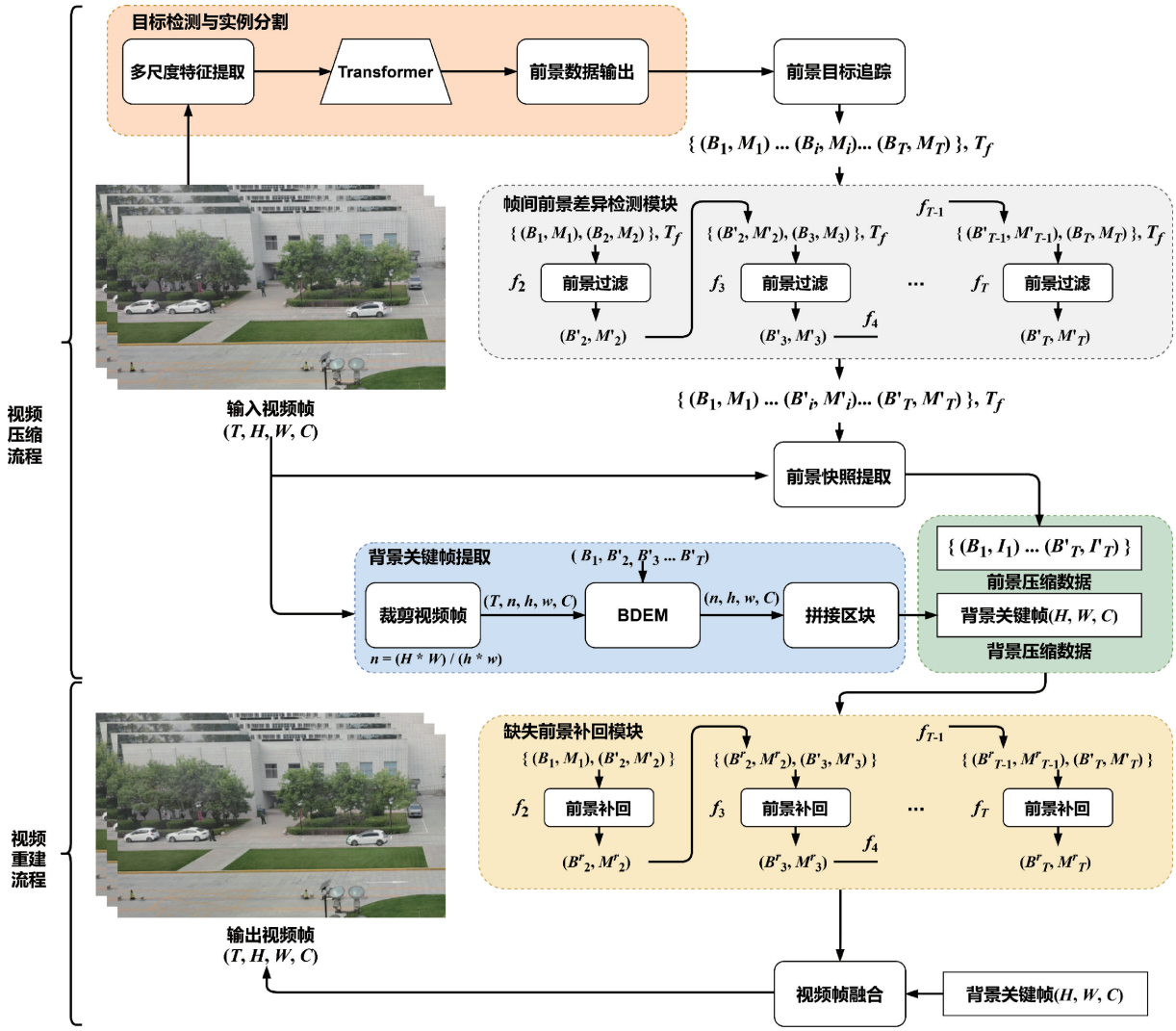


图 1　视频压缩与重建算法架构

目标。对运动目标,利用每个前景目标的位置和区域,从相应的视频帧中将其分割出来,并进行适当的压缩作为前景快照。前景检测框和前景快照集合作为前景压缩数据。对于背景关键帧提取,本文首先将视频帧切分为若干区块,与帧间前景差异检测模块输出的所有前景位置信息一起送入背景检测与提取模块(background detection and extraction module,BDEM)中进行关键背景块的提取。之后,将各个背景块拼接成背景关键帧,以此作为视频背景压缩数据。

在视频重建流程中,由于帧间前景差异检测模块去掉了前景运动轨迹中的部分静止前景,可能导致前景目标在部分帧中丢失,因此需要利用前景补回模块适当补充部分帧中的前景目标。之后将每一帧的前景快照按照对应位置与背景关键帧融合,从而还原视频帧。

2.2　前景目标提取与压缩

前景目标提取与压缩的目的是提取视频帧间存在差异的关键前景目标作为前景目标的压缩数据,以降低视频前景数据冗余。前景目标提取与压缩流程如图 1 上半部分所示,前景过滤步骤流程图如图 2 所示。在目标检测与实例分割模型的主干网络中,对每一个视频帧进行独立的处理,并得到整个视频的前景检测框和实例掩码为 $\{(B_1,M_1),(B_2,M_2)\cdots(B_T,M_T)\}$,其中,$B_i$ 为第 i 帧的目标检测框集合,$B_i=\{d_{i0},d_{i1},\cdots,d_{iu}\}$,$d_{iu}$ 表示第 i 帧的第 u 个检测框;M_i 为第 i 帧的实例掩码集合,$M_i=\{m_{i0},m_{i1},\cdots,m_{iu}\}$,$m_{iu}$ 表示第 i 帧的第 u 个实例掩码。为了更精确的获取视频帧中实例的检测框和实例掩码,本文使用 Mask DINO[19] 作为前景提取部分的主干网络。

获取到全部前景目标的检测框和实例掩码后,使用 IOU Tracker[20] 算法对视频中的前景目标进行追踪。输入主干网络输出的所有前景检测框,得到每一个前景的运动轨迹,记所有轨迹的集合为 T_f。之后,进入帧间前景差异检测模块进行前景变化检测。该模块首先从 T_f 中获取每

一个前景在上下帧之间的匹配关系。之后，对相邻两帧之间每一个前景检测框计算 IOU 进行初步判定，若存在两帧间某对前景检测框的 IOU 超过阈值 σ_{iou}，则说明此前景目标在两帧之间的位置高度重合，需要进一步对比两个前景实例掩码之间的差异。在实例掩码差异检测阶段，首先将两个前景的二值化实例掩码对齐。之后，对两个实例掩码取异或操作，并将异或结果的面积与前一帧实例掩码的面积进行比值，结果记为 xor_rate，如式(1)所示。

$$xor\_rate(a,b)=\frac{Sum(mask_a \oplus mask_b)}{Sum(mask_a)} \tag{1}$$

其中，Sum 为求和操作，$\oplus$ 为异或操作，$mask_a$ 和 $mask_b$ 表示一对二值化实例掩码。如果 xor_rate 低于阈值 σ_{xor}，则判定此目标无变化，并从前景提取结果中去掉对应的目标检测框和实例掩码。检测完成后，得到新的前景检测框和实例掩码集合，记为 $\{(B_1,M_1),(B'_2,M'_2)\cdots(B'_T,M'_T)\}$。

完成目标筛选后，利用每一帧的前景检测框和实例掩码集合，将每一个前景从对应视频帧中截取下来作为前景快照。并在每一个前景目标的运动轨迹中存储一张无损快照，以保留前景目标原始的颜色和纹理特征。其余快照则进行适当压缩。得到前景检测框和前景快照集合，记为 $\{(B_1,I_1),(B'_2,I'_2)\cdots(B'_T,I'_T)\}$，其中 I'_i 为第 i 帧的前景快照集合。前景检测框和前景快照集合即为前景目标的压缩数据。

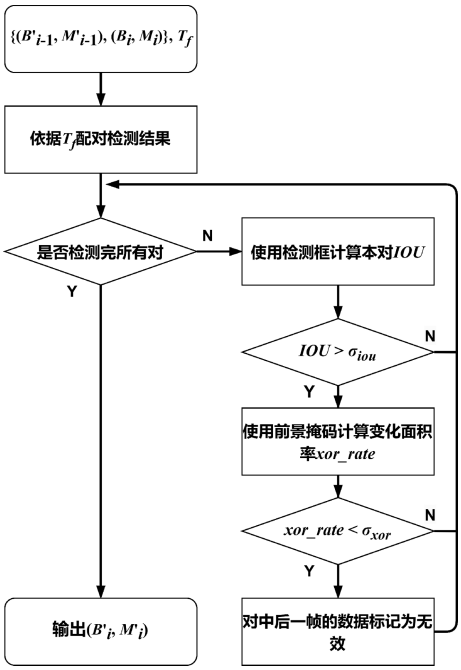


图 2　前景过滤流程图

2.3　背景关键帧提取

背景关键帧提取的目的是在全局视频帧中提取出少量的背景关键帧作为视频背景压缩数据。背景关键帧提取过程如图 1 所示，其中的 BDEM 流程如图 3 所示。受视频采集设备特性的影响，视频帧中通常会包含一定的噪声。为了更有效地减轻视频噪声对背景提取的干扰，本文提出了一种基于 MOG2 的背景检测和提取方法。设输入视频 $V\in R^{T\times H\times W\times 3}$，其中，$T$ 为视频的总帧数，H 和 W 分别为视频帧的高和宽，3 表示视频帧为 RGB 三通道图像。首先将视频整体画面划分为 n 个高宽为 $h\times w$ 的区域。然后，计算每个区域的目标框 g_bbox_i，并将视频的每一帧切分为 n 个区块，记第 t 帧的第 i 个区块为 $f\_block_{t,i}$。上述过程如式(2)～(4)所示。

$$n=(W\times H)/(w\times h) \tag{2}$$

$$g\_bbox_i=[i_x w,i_y h,(i_x+1)(w-1),(i_y+1)(h-1)] \tag{3}$$

$$f\_block_{t,i}=(I(x,y)\in g\_bbox_i)\times f_t(x,y) \tag{4}$$

其中，i 表示视频整体画面的第 i 个区域，h 和 w 分别表示区域的高和宽，I 为指示函数，当像素位置 (x,y) 满足 $I(x,y)\in g\_bbox_i$ 时，结果为 1；否则结果为 0。将切分后的视频记为 $V_{blocks}\in R^{T\times n\times h\times w\times 3}$，并将 V_{blocks} 送入 BDEM 进行关键背景区块的提取。

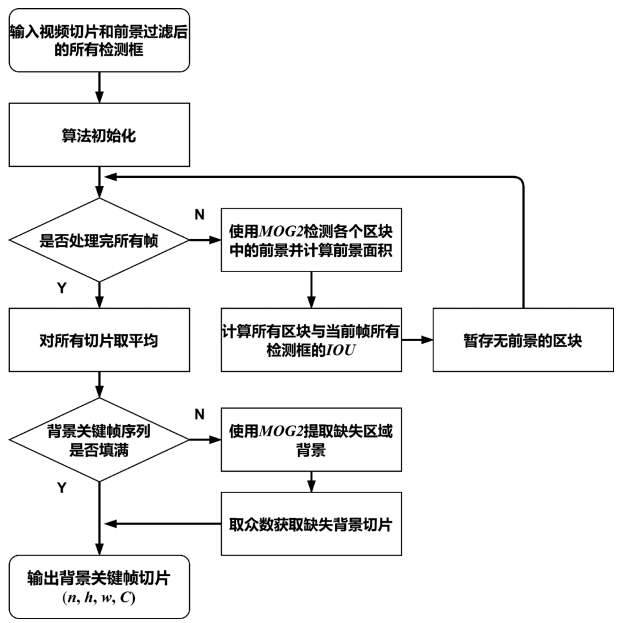


图 3　BDEM 流程

在背景检测和提取模块中，创建一个长度为 n 的关键背景区块序列，用于临时存储识别出的背景区块。这个序列的索引位置与整体画面区域的索引相互对应，确保每个背景区块都能被独立、准确地检测和提取。具体方法如下：首先，利用 MOG2 检测当前区块中的前景区域，并计算当前区块中的前景面积。之后，使用当前帧中的所有检测框与当前区域的目标框一一计算 IOU。若前景面积小于阈值 σ_{gmm_area}，并且与当前帧所有检测框的 IOU 均小于阈值 σ_{bg_iou}，则认为当前区块中无前景目标，暂存背景区块；否则认为区块中含有前景目标，丢弃当前区块。当处理完

所有视频帧之后,对暂存的所有背景区块取平均值作为关键背景区块,存入关键背景区块序列的相应位置。此过程如式(5)所示。

$$g\_block_i = \frac{1}{N_i}\sum_{t=1}^{T} BDEM(B'_t, f\_block_{t,i}) \tag{5}$$

其中,g_block_i 表示区块序列中的第 i 个背景区块,N_i 为区块序列 i 位置中暂存图像块的数量,B'_t 为第 t 帧的前景检测框集合,$f\_block_{t,i}$ 表示第 t 帧第 i 个区域的视频帧。此时,如果区块序列仍有缺失区块,说明有前景目标一直在缺失区块的区域中运动。如果同样对这类区域取平均值,提取到的背景关键帧可能含有残影。由于前景目标一直在区域中运动,如果对同一像素位置的背景取众数,则可以解决残影问题。此时,再次使用 MOG2 对缺失区块的区域进行背景建模,并取计算得到的每一帧背景的众数作为当前区域的背景区块。此过程如式(6)所示。

$$g\_block_j = Mode(\{MOG2(f\_block_{t,j}) \mid t = 1,2,3,\cdots,T\}) \tag{6}$$

其中,g_block_j 表示区块序列中的第 j 个背景区块,$Mode$ 表示取众数操作,$f\_block_{t,j}$ 表示第 t 帧第 j 个区域的视频帧。最终,得到全部关键背景区块 $G_{blocks} \in R^{n\times h\times w\times 3}$,并将其拼接得到背景关键帧 $G \in R^{H\times W\times 3}$。

2.4　视频重建

视频重建部分利用前景压缩数据和背景关键帧还原视频。视频重建流程如图 1 下半部分所示,前景补回流程图如图 4 所示。

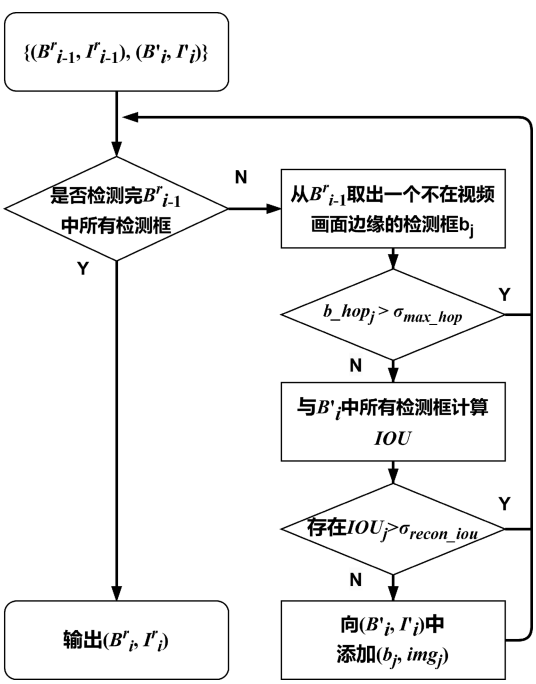


图 4　前景补回流程图

在前景补回流程中,依据上一帧和当前帧的前景检测框集合,一一计算 IOU。如果上一帧的前景目标检测框与视频画面边缘最短的距离大于阈值 σ_{edge_dist},则说明此前景目标不在视频画面边缘。同时,如果在下帧的所有前景检测框中找不到与上一帧 IOU 大于阈值 σ_{recon_iou} 的检测框,则判定为目标丢失,并使用上一帧的前景进行补全。然而,如果此逻辑出现错检,可能导致前景目标重复粘贴。为了缓解此问题,本文为每一个前景设定最大补回次数阈值 σ_{max_hop},如果当前景目标补回次数超过此阈值,则不再下帧中补回该前景目标。之后,获取当前帧前景检测框和前景快照集合 (B^r_t, I^r_t),并依据前景检测框将前景快照与背景关键帧融合,此过程如式(7)所示。

$$f^r_{t,(x,y)} = \begin{cases} I^r_{t,x-x_i,y-y_i}, & (x,y) \in B^r_t, i = 1,2,\cdots,p \\ G_{(x,y)}, & \text{其他} \end{cases} \tag{7}$$

其中,f^r_t 为第 t 帧的重建视频帧,B^r_t 和 I^r_t 为第 t 帧的前景检测框和前景快照集合,x_i 和 y_i 表示第 i 个前景检测框左上角坐标。最后,对所有帧依次进行该操作,以还原所有视频帧。

3　实　　验

3.1　数据集与评价指标

本文采用 BrnoCompSpeed[21] 数据集以及自行采集的视频作为测试数据集。BrnoCompSpeed 数据集包含了 18 个高清交通监控视频,每个视频的时长大约 1 h。为了增加场景多样性,本文从 BrnoCompSpeed 数据集中选取了两段视频,并结合自行采集的两段视频,共计 4 个不同的场景来验证本文方法的有效性。视频详细参数如表 1 所示。

在方法评价指标方面,本文使用每像素所占比特(bits per pixel,BPP)评估视频压缩性能。BPP 表示每个像素需要多少 bit 编码,数值越低则说明视频的压缩比越高。此外,本文使用峰值信噪比(peak signal-to-noise ratio,PSNR)和多尺度结构相似性指数(multi-scale structural similarity,MS-SSIM)评价重建视频图像质量。PSNR 指标通过比较信号的最大值与噪声水平来确定信号的质量,其值域为 $[0,+\infty)$,PSNR 越高则说明重建的图像与原始图像的匹配度越高。另一种评估视频图像质量的指标是 MS-SSIM,其值域为[0,1],MS-SSIM 数值越大,说明重建图像与原始图像在结构上的相似性越强。

3.2　前景目标提取与压缩性能

本节实验重点验证帧间前景差异检测模块以及对前景快照进行图像压缩的有效性。帧间前景差异检测模块中,IOU 阈值和 XOR 阈值的设置分别为:$\sigma_{iou} = 0.93$,$\sigma_{xor} = 0.035$。为了进一步去除前景冗余数据,本文对前景目标进行了适当的压缩。为了更好的保留帧间前景语义信息的完整性,对于行人这类姿态变化相对频繁的目标,将前景快照宽高缩小为原图的 10/23。对于各类车辆等目标,由于目标姿态较为稳定,因此将这类目标的图像尺寸统一缩放到 50×50。

表 1　视频数据信息

视频编号	视频来源	视频宽高	视频帧数	FPS
1	BrnoCompSpeed	1 920×1 080	3 000	100
2	BrnoCompSpeed	1 920×1 080	3 000	100
3	自采	1 280×720	2 000	60
4	自采	1 280×720	2 000	60

本文在使用相同背景压缩数据的条件下，设置了 3 组不同的前景压缩数据处理方式进行视频压缩与重建，并取 4 段视频 BPP、PSNR 和 MS-SSIM 结果的平均值在表 2 中进行展示。其中，第 1 组直接采用主干网络的输出结果作为前景压缩数据，这部分实验使用的所有前景数据均没有进行删除或压缩操作。第 2 组则引入了帧间前景差异检测模块，通过该模块对前景进行筛选，并压缩所有筛选出的前景快照。而在第 3 组中，文本在第 2 组实验的基础上进一步优化，对每一个前景运动轨迹都保留了一张无损的前景快照以保留其原始的颜色和纹理特征。可以看出，在使用帧间前景差异检测模块后，在 PSNR、MS-SSIM 分别损失 0.413 2%、0.113 4%的情况下，BPP 降低了 91.20%。实验结果表明，本文方法可以在保证较高水平视频重建质量的条件下，大幅度降低 BPP。同时，在对每一个前景轨迹存储一张无损前景快照的情况下，BPP 仅上升 7.485%。本文方法在控制 BPP 涨幅的条件下，保留了前景目标原始的颜色和纹理特征。实验证明了本文所提方法能够有效的去除前景数据冗余，同时保持较好的视频重建质量。

表 2　前景提取与压缩性能

前景提取与压缩方法	BPP	PSNR	MS-SSIM
原始前景快照集合	0.130 5	31.46	0.969 7
帧间前景差异检测，无原始快照	0.010 63	31.32	0.968 4
帧间前景差异检测，每个前景保留一张原始快照	0.011 49	31.33	0.968 6

3.3　背景关键帧提取性能

为了验证本文提出的背景检测与提取方法的有效性，在前景压缩数据相同的条件下，本文对于 3 种不同的背景关键帧提取方法得到的背景关键帧分别进行视频压缩与重建实验，并取 4 段视频 BPP、PSNR 和 MS-SSIM 结果的平均值展示在表 3 中。3 种不同策略提取的背景关键帧效果如图 5 所示。实验中，4 段视频被均匀划分为 64 个区域（$n=64$），区域内前景面积阈值设置为 $\sigma_{gmm\_area}=1\,500$，区域与前景目标 IOU 阈值设置为 $\sigma_{bg\_iou}=0$。第 1 组实验使用 MOG2 方法在所有视频帧中提取的背景上取众数作为背景关键帧。第 2 组实验使用本文所提方法提取背景关键帧，但在检测背景时，只使用 MOG2 的输出作为判别依据。第 3 组实验使用本文所提方法提取背景关键帧，并采用 MOG2 方法的输出与前景检测框共同作为背景检测的判别依据。

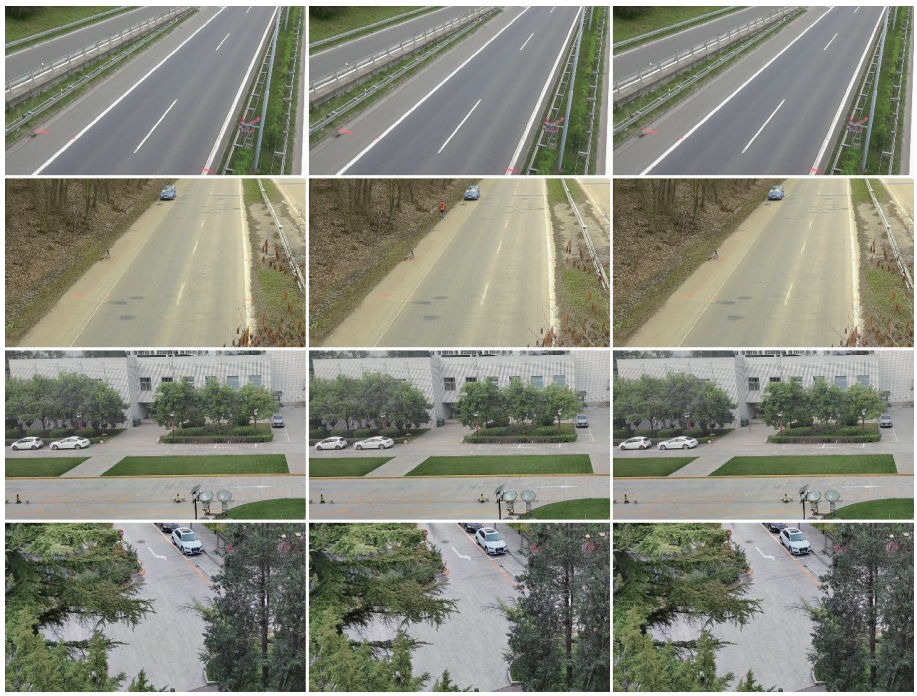
表 3　背景关键帧提取性能

背景提取方法	BPP	PSNR	MS-SSIM
MOG2 全局取众数	0.012 68	30.70	0.970 9
MOG2 分块取平均	0.011 49	31.27	0.968 3
MOG2＋bbox 分块取平均	0.011 49	31.33	0.968 6

从实验结果中可以看出，相比第 1 组，第 3 组的 BPP 下降了 9.385%，PSNR 上升了 2.011%，MS-SSIM 下降了 0.236 9%；相比第 2 组，第 3 组的 BPP 结果相等，PSNR 上升了 0.191 5%，MS-SSIM 上升了 0.030 97%。结合背景提取效果图可以看出，对于第 1 组实验提取出的背景关键帧，尽管背景的结构恢复相对较好，但受到视频噪声的影响较大，所提取的图像呈现出较为明显的锐化现象，进而导致了 BPP 的上升。对于第 2 组实验提取的背景关键帧，由于仅使用 MOG2 检测背景，并且采用图像取均值的方法提取背景关键帧，方法对于暂时静止的行人存在漏检问题，导致提取出的图像中含有伪影。对于第 3 组实验所提取的背景关键帧，在保持背景结构准确的前提下，降低了视频噪声对背景提取的影响，使得图像保持了较高的平滑度，并且有效缓解了伪影问题。实验结果证明了使用 MOG2 结合前景检测框的方法能够有效提取视频中的背景关键信息，并且使得输出图像更加平滑。

3.4　视频压缩与重建性能

本部分实验使用了 4 段视频的全部帧验证本文方法的视频压缩与重建性能，并与传统方法 MPEG-4、H.264 和 H.265 以及近期的深度学习方法 DCVC-DC 进行对比



(a) MOG2全局取众数　(b) MOG2分块取平均　(c) MOG2+bbox分块取平均

图 5　背景关键帧效果图

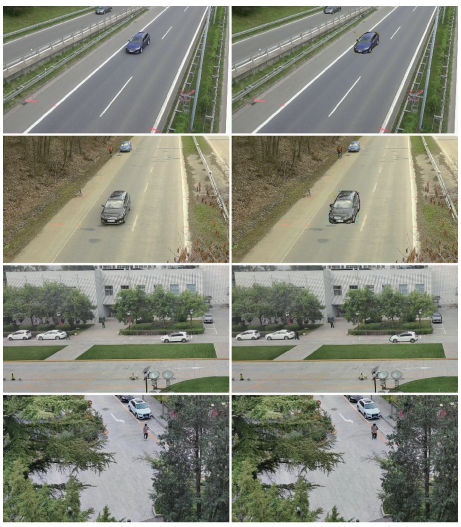
实验。实验结果如表 4 所示,其中,MPEG-4、H.264 和 H.265 在压缩速率上设置为 veryslow,GOP 设置为 32。DCVC-DC 方法分别使用其公开的 PSNR 和 MS-SSIM 模型权重进行实验。实验中,前景目标距离页面边缘阈值设置为 $\sigma_{edge\_dist}=20$,前景补回 IOU 阈值设置为 $\sigma_{recon\_iou}=0.4$,最大前景补回次数阈值 $\sigma_{max\_hop}=40$。

表 4　视频压缩与重建方法对比

方法	BPP	PSNR	MS-SSIM
MPEG-4	0.013 30	27.24	0.906 3
H.264	0.066 61	34.66	0.985 8
H.265	0.048 25	34.56	0.984 6
DCVC-DC PSNR model(CVPR2023)	0.028 41	33.82	—
DCVC-DC MS-SSIM model(CVPR2023)	0.034 48	—	0.984 3
本文方法	0.011 49	31.33	0.968 6

结果表明,本文方法与 MPEG-4 相比,在 BPP 降低 13.60%的同时,PSNR 提升了 13.05%、MS-SSIM 提升了 6.432%。与 H.264、H.265 和 DCVC-DC 相比,本文方法的在 BPP 指标上分别降低了 82.75%、76.19%和 59.56%;在 PSNR 指标上分别相差 9.608%、9.346%和 7.363%;在 MS-SSIM 指标上分别相差 1.745%、1.625%和 1.595%。本文方法在大幅度降低 BPP 的情况下,仍然在视频重建上保持了较高的性能。图 6 展示了本文方法的视频重建效果。从效果图中可以看出,本文方法在对前景目标的姿态和位置上,能够做到准确的还原,但在前景纹理细节的表现上仍存在改进空间。本文方法在背景重建上表现较好,对原视频背景的整体结构和细节都有较好的还原。

充分的实验结果表明,本文方法和其他方法相比,充分利用了视频场景的先验知识以进一步去除视频中大量的数据冗余,并在提升视频压缩性能的同时保持了较高的视频重建水平。



(a) 原始视频　　(b) 本文方法

图 6　视频重建效果图

4　结　　论

鉴于当前固定场景类视频数据的急剧增长及其对存储和传输造成的巨大压力，本文充分利用固定场景类视频的特点，针对固定场景类视频提出了一种新颖的压缩与重建方法。该方法结合帧间前景差异检测的前景提取与压缩方法提取视频中的关键前景目标，有效地去除了其中的静态前景，在降低前景数据冗余的同时确保了前景语义信息的准确性。同时，结合 MOG2 背景建模算法和前景检测框，对视频背景中提取少量关键帧作为背景压缩数据，有效的降低了视频背景数据冗余，同时对背景整体结构和细节都能较好的还原。并通过背景与前景目标的融合的方式实现视频帧的重建。在验证数据集上的实验表明，本文所提方法在保持了较高视频重建水平的条件下，大幅提升了视频压缩性能，能够有效的缓解视频的存储和传输压力。

参考文献

[1] LU G, OUYANG W, XU D, et al. DVC: An end-to-end deep video compression framework[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2019: 10998-11007.

[2] LI J, LI B, LU Y. Neural video compression with diverse contexts[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2023: 22616-22626.

[3] LI J, LI B, LU Y. Deep contextual video compression[C]. Advances in Neural Information Processing Systems, 2021, 34: 18114-18125.

[4] SHENG X, LI J, LI B, et al. Temporal context mining for learned video compression[J]. IEEE Transactions on Multimedia, 2023, 25: 7311-7322.

[5] LI J, LI B, LU Y. Hybrid spatial-temporal entropy modelling for neural video compression[C]. Proceedings of the 30th ACM International Conference on Multimedia, 2022: 1503-1511.

[6] WU C Y, SINGHAL N, KRÄHENBÜHL P. Video compression through image interpolation[C]. Proceedings of the European Conference on Computer Vision(ECCV), 2018: 416-431.

[7] 吕昌，尹和，邵叶秦. 基于结构重参数化的目标检测模型[J]. 电子测量技术，2023，46(18)：114-121.

[8] LEE Y, PARK J. CenterMask: Real-time anchor-free instance segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2020: 13903-13912.

[9] CHENG T, WANG X, CHEN S, et al. Sparse instance activation for real-time instance segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2022: 4433-4442.

[10] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]. IEEE International Conference on Computer Vision(ICCV), 2017: 2980-2988.

[11] CHEN L C, HERMANS A, PAPANDREOU G, et al. MaskLab: Instance segmentation by refining object detection with semantic and direction features[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 4013-4022.

[12] DAI J, HE K, SUN J. Instance-aware semantic segmentation via multi-task network cascades[C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016: 3150-3158.

[13] LI Y, QI H, DAI J, et al. Fully convolutional instance-aware semantic segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017: 4438-4446.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems, 2017: 5998-6008.

[15] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]. Proceedings of the European Conference on Computer vision(ECCV), 2020: 213-229.

[16] ZHANG H, LI F, LIU S, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection[J]. ArXiv preprint arXiv:2203.03605, 2022.

[17] CHENG B, SCHWING A G, KIRILLOV A. Per-pixel classification is not all you need for semantic segmentation[C]. Advances in Neural Information Processing Systems, 2021, 34: 17864-17875.

[18] CHENG B, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2022: 1280-1289.

[19] LI F, ZHANG H, XU H, et al. Mask DINO: Towards a unified transformer-based framework for object detection and segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2023: 3041-3050.

[20] BOCHINSKI E, EISELEIN V, SIKORA T. High-speed tracking-by-detection without using image information[C]. IEEE International Conference on Advanced Video and Signal Based Surveillance(AVSS), 2017: 1-6.

[21] SOCHOR J, JURÁNEK R, ŠPAŇHEL J, et al. Comprehensive data set for automatic single camera visual speed measurement[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(5): 1633-1643.

作者简介

李萌，硕士研究生，主要研究方向为计算机视觉、深度学习和视频压缩。
E-mail: 2021020577@bistu.edu.cn

黄宏博(通信作者)，博士，副教授，硕士生导师，主要研究方向为计算机视觉、深度学习和人工智能。
E-mail: hhb@bistu.edu.cn

郑曜林，硕士研究生，主要研究方向为计算机视觉、深度学习和人工智能。
E-mail: 2021020574@bistu.edu.cn

许龙飞，硕士研究生，主要研究方向为计算机视觉、深度学习和人工智能。
E-mail: 2022020601@bistu.edu.cn