视频对象移除篡改的时空域定位被动取证

陈临强[1]，杨全鑫[2]，袁理锋[1]，姚晔[1]，张祯[1]，吴国华[1]
（1. 杭州电子科技大学网络空间安全学院，浙江 杭州 310018；2. 杭州电子科技大学计算机学院，浙江 杭州 310018）

摘 要：针对视频被动取证领域中视频内容的真实性和完整性鉴定及篡改区域定位问题，提出了一种基于视频噪声流的深度学习检测算法。首先，构建了基于空间富模型（SRM）和三维卷积（C3D）神经网络的特征提取器、帧鉴别器和基于区域建议网络（RPN）思想的空域定位器；其次，将特征提取器分别与帧鉴别器和空域定位器相结合，搭建出 2 个神经网络；最后，利用增强处理后的数据训练出 2 种深度学习模型，分别用于对视频篡改区域时域和空域的定位。测试结果表明，时域定位的准确率提高到 98.5%，空域定位与篡改区域标注平均交并比达 49%，可以有效对该类篡改视频进行篡改区域时空域定位。

关键词：视频对象移除篡改；时空域定位；视频被动取证；三维卷积目标检测
中图分类号：TP309
文献标识码：A
doi: 10.11959/j.issn.1000-436x.2020151

Passive forensic based on spatio-temporal localization of video object removal tampering

CHEN Linqiang[1], YANG Quanxin[2], YUAN Lifeng[1], YAO Ye[1], ZHANG Zhen[1], WU Guohua[1]
1. School of Cyberspace Security, Hangzhou Dianzi University, Hangzhou 310018, China
2. School of Computer, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: To address the problem of identification of authenticity and integrity of video content and the location of video tampering area, a deep learning detection algorithm based on video noise flow was proposed. Firstly, based on SRM (spatial rich model) and C3D (3D convolution) neural network, a feature extractor, a frame discriminator and a RPN (region proposal network) based spatial locator were constructed. Secondly, the feature extractor was combined with the frame discriminator and the spatial locator respectively, and then two neural networks were built. Finally, two kinds of deep learning models were trained by the enhanced data, which were used to locate the tampered area in temporal domain and spatial domain respectively. The test results show that the accuracy of temporal-domain location is increased to 98.5%, and the average intersection over union of spatial localization and tamper area labeling is 49%, which can effectively locate the tamper area in temporal domain and spatial domain.

Key words: video object removal tampering, spatio-temporal localization, video passive forensic, object detection based on 3D convolution

1 引言

随着数字视频处理技术的飞速发展和图像编辑软件的更新换代，篡改视频[1-3]变得随处可见。而在众多视频篡改类型中，面向视频对象移除篡改[4-5]的被动取证研究更有应用价值和研究意义。移除篡改即将某个关键视频对象从原始视频序列中移除，经过修补和填充后，该视频对象在被篡改视频序列

收稿日期：2020-01-06；**修回日期**：2020-04-29
通信作者：姚晔，yaoye@hdu.edu.cn
基金项目：教育部人文社科基金资助项目（No.17YJC870021）
Foundation Item: Humanities and Social Sciences Foundation of Ministry of Education of China (No.17YJC870021)

的每一帧中都不可见，且凭肉眼无法辨别篡改痕迹。如果需要将视频作为执法依据，则必须证实其真实性和完整性。而数字视频的被动取证研究仍然处于起步阶段[5]，尚有较大的探索和完善空间。

近年来，有众多的科研工作者在图像和视频篡改被动取证领域做出了贡献。在图像被动取证领域，李岩等[6]提出的翻转不变加速稳健特征（FI-SURF, flip invariant speeded-up robust feature）算法可以有效检测出图像的镜像复制粘贴篡改。黄维隽等[7]利用块离散余弦变换（BDCT, block for discrete cosine transform）[8]系数构造马尔可夫特征[9]，再用支持向量机（SVM, support vector machine）分类器可以有效识别图像的拼接篡改。随着深度学习在图像识别和目标检测领域广泛应用，Adobe 公司提出的双流[10]Faster R-CNN（region-convolutional neural network）[11-13]可以很好地检测出图像中的拼接、复制和移除篡改区域。相比于图像篡改，视频篡改操作可以从相邻帧中获得更多的场景信息，从而可以更完美地对移除对象的区域进行修补，使基于单帧视频图像的篡改检测更加困难。

在数字视频的被动取证领域，刘雨青等[14]针对固定监控摄像头所拍摄视频中运动目标的移除，提出一种基于时空域能量可疑度的视频篡改检测方法。首先计算视频各帧能量可疑度，提取时域上的篡改序列；然后通过帧差法计算可疑的运动点图像，提取空域上的可疑运动图像块；最后通过能量可疑度排除干扰图像块，确定篡改图像块区域。李倩等[15]针对固定背景下运动目标的移除，将视频帧划分为网格，并统计每个网格内光流方向的标准差。通过与阈值比较来判断每个网格是否为篡改区域，进而对整帧篡改区域进行空域定位，最后根据空域定位的结果利用二分查找法进行时域定位。Yao 等[4]针对视频对象的移除篡改，利用相邻两帧帧差构造帧差序列，再用高通滤波器提取帧差高频信号作为 CNN 的输入，使用 CNN 自动提取相邻帧间的高频篡改特征，有效提高了篡改帧检测的准确率。何沛松[16]提出了基于卷积神经网络的帧级双压缩检测算法，利用中值滤波器提取中值滤波残差信号作为 CNN 的输入，其鉴别能力明显优于 AlexNet。

现有视频取证算法可以分为四大类[5]：基于噪声模式的算法、基于像素相关性的算法、基于视频内容特征的算法和基于抽象统计特征的算法。大多数算法都是基于视频对象篡改引起的视频内容本身的异常和视频篡改的先验知识，手动设计特定的特征提取和分析算法，以检测和识别视频对象的篡改。这类人工设计的特征提取算法，通常针对特定且有限的篡改视频库。视频篡改检测的准确率由人工设计的特征提取算法决定。同时，篡改视频库的制作方式、内容和编码参数等都会影响视频篡改取证算法的检测性能。目前针对视频对象移除篡改的深度学习算法中，使用视频帧差法等方法并不能充分体现篡改特征在视频时域方向上的连续性和相关性。

为了提升视频对象篡改被动取证算法的通用性，并充分提取篡改操作在时空域方向上的特征，本文提出了基于三维卷积（C3D, 3D convolution）神经网络[17-19]的视频对象移除篡改时空域定位技术。针对被篡改的视频帧，可以判别待测视频帧是否存在视频对象被移除篡改，并定位被移除篡改的区域。相比于传统算法，本文算法具有以下优势：1) 没有视频背景保持不变，监控摄像头及移除的视频对象为运动或静止状态等硬性要求，增加了算法的可应用场景；2) 对预测结果的判断不需要人工观察和挑选阈值等操作，减少了人工操作和人为误差；3) 在空域定位算法中，将基于深度学习的目标检测理念运用到时空域三维空间，提高了算法的通用性和灵活性。

2　时空域定位系统结构与数据集处理方法

2.1　时空域定位系统结构

本文提出的视频对象移除篡改区域的时空域定位系统结构如图 1 所示。该系统是基于带标注的视频数据集设计的。标注信息包括每一帧是否为篡改帧，以及篡改帧中篡改区域的矩形坐标。文献[5]中介绍 4 种篡改取证算法的数据集：SULFA（surrey university library for forensic analysis）、REWIND、DICGIM 和 SYSU-OBJFORG。其中，前 2 种数据集分别只有 5 段和 10 段关于视频对象移除篡改的视频，且每段视频长度只有 10 s（帧率为 30 frame/s），由于数据集太小不利于使用深度学习框架来进行训练。DICGIM 数据集篡改类型中不包含视频对象的移除篡改，不适于本文算法。SYSU-OBJFORG 库是目前最大的视频对象篡改库，包含 100 段原始视频和 100 段原始视频对应的篡改视频。该数据集篡改类型均为视频对象移除篡改，

且平均每段视频长度为 11 s（帧率为 25 frame/s），所有视频均经过 H.264/MPEG-4 编码格式压缩。数据集中的篡改视频是将原始视频解码之后，由数据集制作者对视频图像逐帧进行空域上的篡改，再编码压缩而成的。篡改视频经历了解压缩、篡改、再压缩的过程，对篡改检测和定位算法的要求更高。本文首先将数据集中的所有视频解压成图像帧，然后进行裁剪、翻转等数据增强操作，所得数据集规模可以满足深度学习训练的要求。

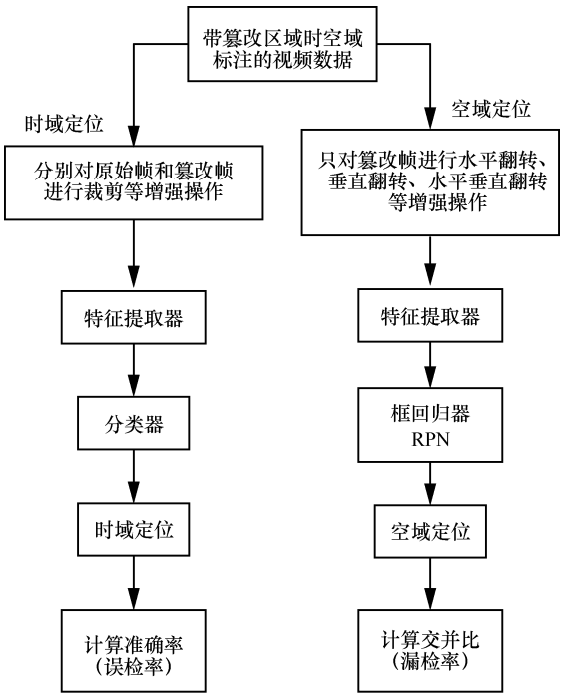

图 1 时空域定位系统结构

图 1 中左右两条分支为 2 个不同的网络模型，分别用于视频被篡改区域的时域和空域定位。2 个模型的输入都是连续的 5 帧视频图像，且特征提取器的结构相同。2 个 CNN 模型最大的区别在于特征提取器所生成的特征图是用于分类还是用于 RPN（region proposal network）[20]框回归定位。2 个模型训练数据增强的处理方式也不同。用于时域定位的网络需要同时输入原始帧和篡改帧进行训练，而用于空域定位的网络则只需要输入篡改帧进行训练。

2.2 2 种不同的数据集处理方法

为了尽量不破坏视频数据，同时对训练数据进行比例调整和样本扩充，在诸多的计算机视觉数据增强策略中，本文只选取了裁剪和翻转 2 种数据增强策略。2 种数据集的处理方法都是基于连续的 5 帧视频图像进行操作的（即当前帧加上连续的前 2 帧和连续的后 2 帧），以便提取篡改痕迹在时域上的连续性特征。对于时域定位，由于数据集中原始帧和篡改帧比例约为 13:3，为了保证训练集正负样本数量相当，本文借鉴文献[4]提出的非对称数据增强策略，将 2 种视频帧按帧数的相反比例进行裁剪并打包。对于空域定位，篡改帧内的定位在完整的帧图像中进行，因此不对输入图像帧进行裁剪，而是通过水平翻转、垂直翻转和水平垂直翻转来进行增强。具体增强策略分别介绍如下。

2.2.1 用于时域定位的数据集增强策略

本文视频数据集尺寸为 1 280 像素×720 像素，裁剪尺寸需要满足以下条件：每个裁剪区域在训练集的篡改帧中至少要包含大部分的篡改区域；测试集中所有裁剪区域要覆盖整帧进行测试，不能有遗漏区域。为了方便计算和模型的简洁性，本文设定裁剪尺寸为 720 像素×720 像素。对于训练集和验证集采用相同的裁剪方案，只对连续的 5 帧原始帧或篡改帧进行裁剪。

对于连续原始帧，统一按均匀步长左中右三次裁剪（可以适当在横坐标方向进行随机像素的微小偏移，以防过多学习每一帧的边缘特点）或随机裁剪三次，并且连续 5 帧的裁剪位置保持严格一致。

连续 5 帧篡改帧的篡改区域如图 2 所示。假设坐标框$(x_{i1},y_{i1},x_{i2},y_{i2})$内为视频对象移除篡改区域，其中 i=1,2,3,4,5。按式(1)求出如图 3 所示的连续 5 帧篡改区域最小外接矩形区域$(X_{\min},Y_{\min},X_{\max},Y_{\max})$，以$(X_{\min},0,X_{\max},720)$为裁剪边界，按平均步长平移裁剪出若干份或随机裁剪出若干份（但要保证裁剪部分包含篡改区域或大部分篡改区域）连续 5 帧的图像数据并打包。对于测试集，将连续 5 帧视频均按统一步长进行左中右三次裁剪，以保证检测区域覆盖视频帧全部区域。以上每裁剪出连续 5 帧便打包为一组作为时域定位模型的输入数据，且 label 以中间帧的 label 为准。

$$\begin{aligned} X_{\min} &= \min\left(x_{11},\ x_{21},\ x_{31},\ x_{41},\ x_{51}\right) \\ Y_{\min} &= \min\left(y_{11},\ y_{21},\ y_{31},\ y_{41},\ y_{51}\right) \\ X_{\max} &= \max\left(x_{12},\ x_{22},\ x_{32},\ x_{42},\ x_{52}\right) \\ Y_{\max} &= \max\left(y_{12},\ y_{22},\ y_{32},\ y_{42},\ y_{52}\right) \end{aligned} \tag{1}$$

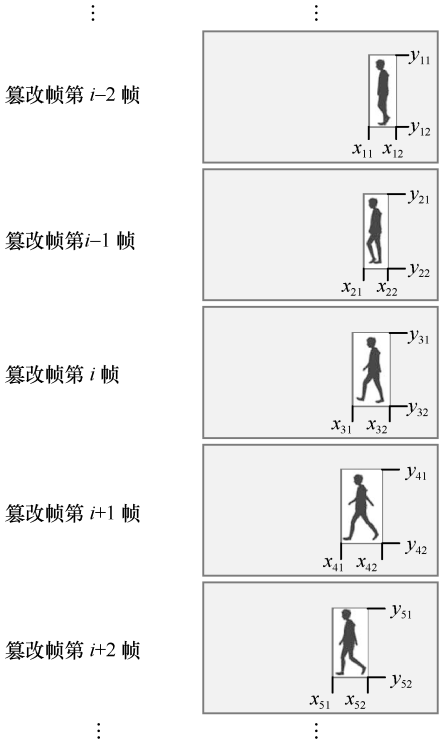

图 2　连续 5 帧篡改帧篡改区域示意

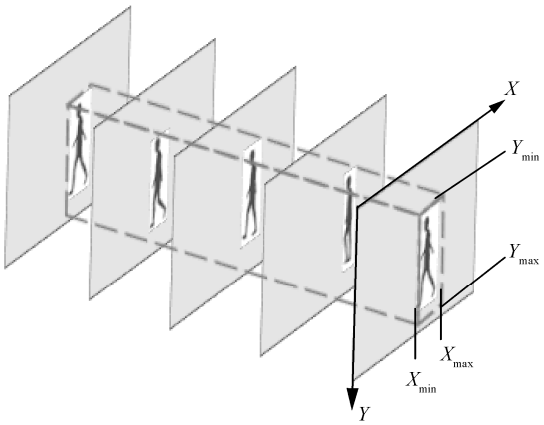

图 3　连续 5 帧篡改区域最小外接框示意

2.2.2　用于空域定位的数据集增强策略

对于训练集，连续 5 帧同时采取水平翻转、垂直翻转和水平垂直翻转进行数据增强。而测试集不需要进行翻转。训练集翻转的同时，篡改区域标注框 Ground Truth 坐标做相应变换。设视频帧的宽为 W，高为 H，篡改区域坐标为(x_1,y_1,x_2,y_2)，则水平翻转后为$(W–x_2,y_1,W–x_1,y_2)$，垂直翻转后为$(x_1,H–y_2,x_2,H–y_1)$，水平垂直翻转后为$(W–x_2, H–y_2,W–x_1, H–y_1)$。同样每 5 帧进行打包并标注，且标注框 Ground Truth 以中间帧为准。

3　搭建时空域定位网络

模型分为三部分，分别为用于提取三维高频特征的特征提取器、用于时域定位的分类器和用于空域定位的 RPN 框回归器，其网络结构分别如图 4～图 6 所示。

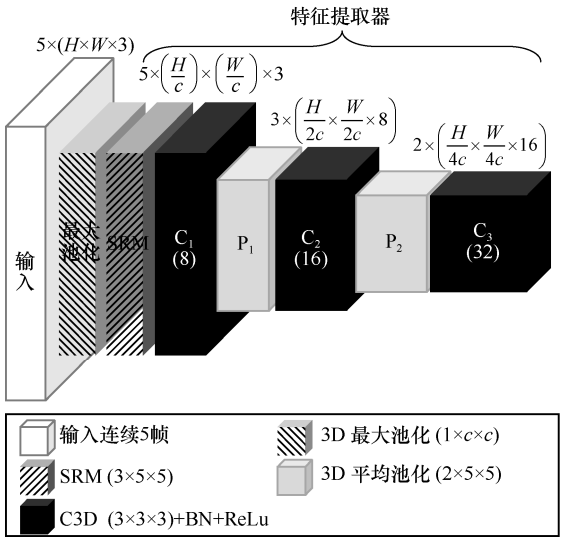

图 4　C3D 神经网络特征提取网络结构

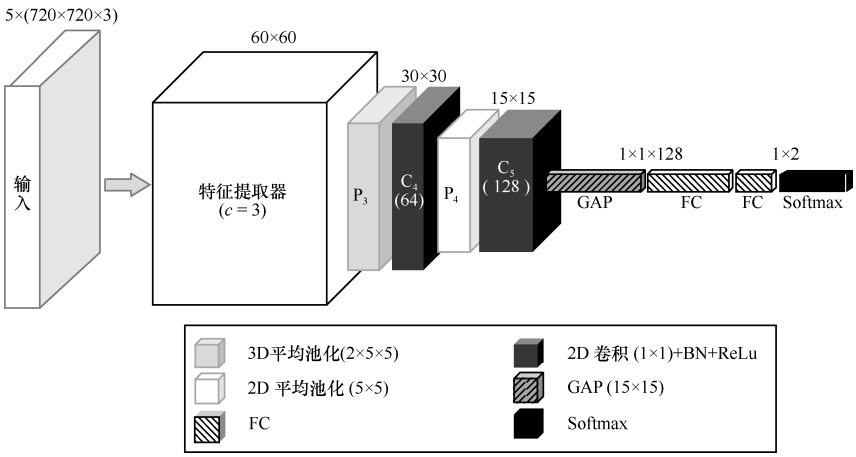

图 5　时域定位分类模型

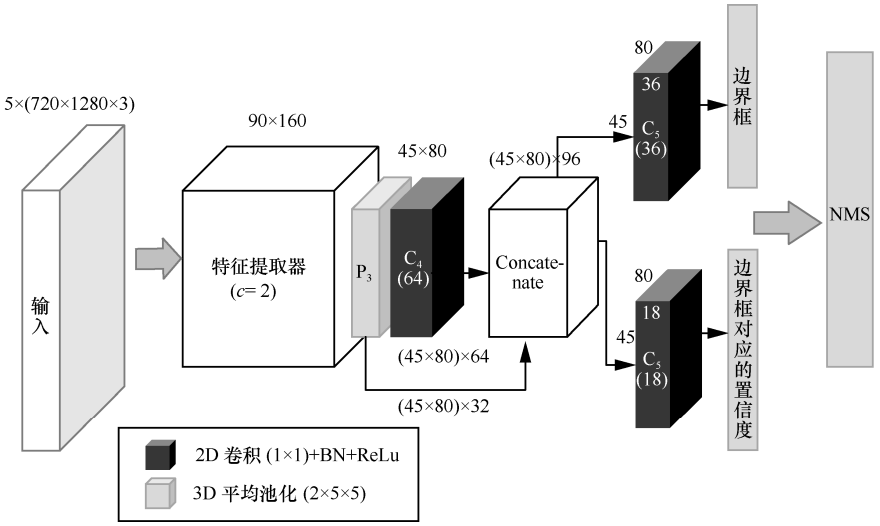

图 6　空域定位回归模型

3.1　特征提取器

C3D 神经网络可以捕捉连续帧在时域方向相关语义的变化特征，因此常用于计算机视觉的视频内容识别领域，来检测人物动作及行为。由于篡改操作会留下痕迹[21-22]，这种痕迹在高频区域[23-25]往往更加明显。通过观察发现，在篡改视频中，篡改区域的高频信号在相邻帧间具有极大的连续性和相关性[26-27]。这种连续性和相关性可以看作篡改痕迹在连续帧间发生的“动作”，类似使用 C3D 神经网络对人物动作的识别，本文使用 C3D 神经网络来提取篡改区域的“动作”特征。

输入的连续 5 帧图像首先经过 2 个图像处理层：三维最大池化层和空间富模型（SRM, spatial rich model）[10]过滤器层。其中，最大池化层滑动窗口尺寸为 1×c×c[28]，在时域定位器中 c 设置为 3，在空域定位器中 c 设置为 2，时域方向池化参数为 1，以保证数据在输入 CNN 前时域方向特征不被破坏。最大池化层的作用是在空域方向上对输入数据进行降维，保留其主要特征，同时减小计算量。SRM 过滤器层使用 3 个不同参数的卷积核分别提取视频帧 3 种不同的高频残差信号，并将其参数设置为不可训练。本文中 SRM 过滤器所使用的 3 个卷积核参数分别为

$$\frac{1}{4}\begin{pmatrix}0&0&0&0&0\\0&-1&2&-1&0\\0&2&-4&2&0\\0&-1&2&-1&0\\0&0&0&0&0\end{pmatrix}$$

$$\frac{1}{12}\begin{pmatrix}-1&2&-2&2&-1\\2&-6&8&-6&2\\-2&8&-12&8&-2\\2&-6&8&-6&2\\-1&2&-2&2&-1\end{pmatrix}$$

$$\frac{1}{4}\begin{pmatrix}0&0&0&0&0\\0&0&-1&0&0\\0&-1&4&-1&0\\0&0&-1&0&0\\0&0&0&0&0\end{pmatrix}$$

将三维高频数据送入 C3D 神经网络特征提取网络中，同时提取时空域高频信号特征。网络结构如图 4 所示，其中，C_i（i=1,2,3,4,5）表示第 i 层卷积层，C_i 下面括号中的数字表示该层卷积核的个数。图 4 中 C_1～C_3 分别是第 1～3 层三维卷积层，卷积核尺寸均为 3×3×3。每个卷积层均进行批标准化（BN, batch normalization）操作，激活函数均为 ReLu。P_i（i=1,2,3,4,5）表示第 i 层池化层，P_1、P_2 分别是第 1、2 层三维平均池化层，滑动窗口均为 2×5×5，池化层步长均为 2。

3.2　时域定位分类器

时域定位分类器由图 4 的特征提取器（其中 c=3）后接二分类器组成，用于判别输入的连续 5 帧是否为篡改帧。时域定位分类模型如图 5 所示。特征提取器生成的特征图经过一个三维平均池化层 P_3（滑动窗口为 2×5×5，步长为 2），将特征图时域维度降为 1，然后依次经过卷积核为 1×1 的 2 个二维卷积层 C_4 和 C_5，卷积核个数分别为 64 和 128。1×1 的卷积核是为了避免学习过于复杂的特征，同

时起到降维的作用。P_4 为二维平均池化层（滑动窗口大小为 5×5，步长为 2）。全局平均池化层（GAP, global average pooling）P_5 将数据由 128 维的特征图变为 128 维向量。最后经过全连接层（FC, fully connected layer）将数据维度降为 2，并利用 Softmax 层进行归一化分类。

3.3 空域定位回归器

空域定位回归器是由特征提取器（其中 c=2）后接 RPN 框回归器组成的，用于预测篡改位置并给出相应的置信度（即预测区域为篡改区域的可能性）。空域定位回归模型如图 6 所示。由特征提取器生成的特征图需要经过三维平均池化层 P_3（滑动窗口为 2×5×5，步长为 2）将时域维度降为 1。将 P_3 和 C_4 的特征图合并为 Concatenate 层[20]作为框回归的特征图，C_4 和 C_5（有上、下 2 个分支）均为卷积核尺寸为 1×1 的二维卷积层，其中 C_4 卷积核个数为 64；C_5（上）卷积核个数为 36，C_5（下）卷积核个数为 18，分别对应特征图每个位置 9 种尺寸的回归框所对应的框坐标和其置信度。

在测试阶段，通过非极大值抑制（NMS, non-maximum suppression）[10]将候选框按置信度排序，本文中只保留置信度最高的一个候选框作为预测的篡改区域。

4 实验过程

4.1 数据集介绍

在 SYSU-OBJFORG 数据集上验证本文算法。SYSU-OBJFORG 是 Chen 等[29]提出的基于视频对象移除篡改的数据集，由 100 段原始视频和 100 段篡改视频组成。原始视频来源于静态视频监控摄像机拍摄的视频片段，篡改视频则是从这些原始视频中通过逐帧移除目标对象获得的。其中篡改帧中的篡改区域是用其在帧内左上角和右下角的坐标来标注的。

实验中，将 100 对视频序列进行随机排序，按照 5:1:4 的比例分为训练集、验证集和测试集。2 个模型均使用相同序列的训练集，数据增强方式如 2.1 节所述。2 个模型都训练好后，首先将测试数据输入时域定位器，得到视频每一帧的分类结果，即完成篡改时域定位；然后记录下篡改帧编号，将这些篡改帧作为空域定位器的输入序列，进行篡改帧空域定位。重复 5 次，每次都从不同的随机序列中分出训练集、验证集和测试集，并重新开始训练。对 5 次实验的测试结果求平均值，作为衡量实验模型有效性的依据。

4.2 实验设置

本文的神经网络模型基于 Tensorflow 深度学习框架实现，运行于 NVIDIA Geforce GTX1080ti GPU 上。使用 AdamOptimizer 进行优化，将学习率设置为 1×10^{-3}，动量设置为 0.9，l_2 正则化参数设置为 0.000 5，参数初始化标准差均设置为 0.1。

对于时域定位网络，批大小设置为 64，即每次输入神经网络的图像块维度为 64×5×720×720×3。当验证集损失函数趋于收敛时，挑选出训练好的模型用于测试。测试时批大小设置为 3，即同时输入同一帧的三次裁剪的图像数据。如果全部被预测为原始帧，则判断中间帧为原始帧；否则判断为篡改帧，进而得到篡改帧时域定位序列。

对于空域定位网络，批大小设置为 1，即每次输入神经网络的图像块维度为 1×5×720×1280×3。设置的空域定位损失函数与文献[11]中定义的损失函数相似，包含前景框（篡改区域）和背景框（原始区域）分类误差、篡改区域定位框误差两部分。当损失函数趋于收敛时，用训练好的模型进行测试，将篡改帧时域定位产生的篡改帧序列作为空域定位器的输入。本文中采用的损失函数如式(2)所示。

$$\text{loss}=\frac{1}{N_{\text{cls}}}\sum_i L_{\text{cls}}(i)+\frac{1}{N_{\text{reg}}}\sum_j L_{\text{reg}}(j) \tag{2}$$

其中，N_{cls} 是随机抽取的前景框（与篡改标注框交并比大于 0.8）和背景框（与篡改标注框交并比小于 0.2）的视频帧（篡改帧）总数，i 是这些框的下标，L_{cls} 是对其中每一个框二分类（前景框与背景框）的损失函数；N_{reg} 是抽取的框中前景框的数量，j 是其中前景框的下标，L_{reg} 是预测框与真实篡改区域标注框之间误差的损失函数。

为了使正负样本框数量均衡和减少目标区域误检率，设定每一帧中参与训练的前景框与背景框数量之比为 1:α，则前景框数量 fg_num 和背景框数量 bg_num 可由式(3)进行约束。

$$\begin{aligned}\text{fg\_num}&=\min\left(\text{fg\_sum},\frac{\text{roi\_num}}{\alpha+1}\right)\\ \text{bg\_num}&=\min(\text{roi\_num}-\text{fg\_num},\alpha\,\text{fg\_num})\end{aligned} \tag{3}$$

其中，fg_sum 为前景框的总数；roi_num 为常数，

其大小控制着正负样本的训练密度。实验中，设置 roi_num=128，α=5，保证在正负样本数量相差不大的情况下加强对负样本的训练，目的是降低单一目标区域检测的误检率。

4.3 实验结果

4.3.1 时域定位测试

在篡改视频时域定位测试中，本文与文献[4]方法进行了对比。本文采用 Chen 等[29]定义的测试结果衡量标准，如式(4)所示。

$$\begin{aligned}
\text{PFACC}&=\frac{\sum \text{correctly_classified_pristine_frame}}{\sum \text{pristine_frame}}\\
\text{FFACC}&=\frac{\sum \text{correctly_classified_forged_frame}}{\sum \text{forged_frame}}\\
\text{FACC}&=\frac{\sum \text{correctly_classified_frame}}{\sum \text{all_the_frame}}\\
\text{Precision}&=\frac{T_p}{T_p+F_p}\\
\text{Recall}&=\frac{T_p}{T_p+F_N}\\
\text{F}_1\text{ Score}&=\frac{2\text{Precision Recall}}{\text{Precision}+\text{Recall}}
\end{aligned}\tag{4}$$

其中，PFACC 是原始帧正确率，FFACC 是篡改帧正确率，FACC 是所有帧的正确率，Precision 为精确率，Recall 为召回率，F_1 Score 为 F_1 分数，T_p 为篡改帧被正确预测的数量，F_p 为原始帧被错误预测为篡改帧的数量，F_N 为篡改帧被错误预测为原始帧的数量。重复 5 次实验，每次都随机产生不一样的训练集、验证集和测试集，测试结果取平均值，如图 7 所示。

4.3.2 空域定位测试

在目标检测领域，mAP 是最经典的衡量预测精确度的指标，而交并比是目标检测算法指标 mAP 计算的一个非常重要的函数，可以被直观地理解为预测框与标注框的重合程度。在文献[30]中，交并比被认为是定位准确率的最佳标准。视频篡改检测与计算机视觉中的目标检测不同：目标检测中的语义对象可以肉眼看到，并且和周边场景有较大差异；视频篡改检测中的篡改区域可能远大于运动目标或语义对象所在的区域，并且篡改之后区域和周边的场景没有明显的差异，肉眼无法区别原有的语义对象和真实的篡改区域。因此，预测区域不能简单地与被移除的语义对象区域进行比较，而应该与实际的被篡改区域进行比较。由于要预测的只有被篡改区域这一个类别，因此引入了比 mAP 值更直观的成功检测率和平均交并比作为替代指标。成功检测率表示在篡改帧中，可以较好地预测出篡改区域的帧数与测试总帧数的比例。平均交并比表示这些预测出的较好的篡改区域与其真实篡改区域重合比例的平均值。

对于每个篡改帧，取预测框序列中置信度最高的框作为最终的预测区域。当预测框与真实篡改区域标注框的交并比为 0 或置信度小于 0.8 时，定义为漏检帧 F_{mis}，否则为成功检测帧 F_{suc}。由此可以计算出成功检测率 Suc_rate 为

$$\text{Suc\_rate}=\frac{\sum F_{\text{suc}}}{\sum F_{\text{suc}}+\sum F_{\text{mis}}}\tag{5}$$

定义测试集平均交并比 IOU_mean 为

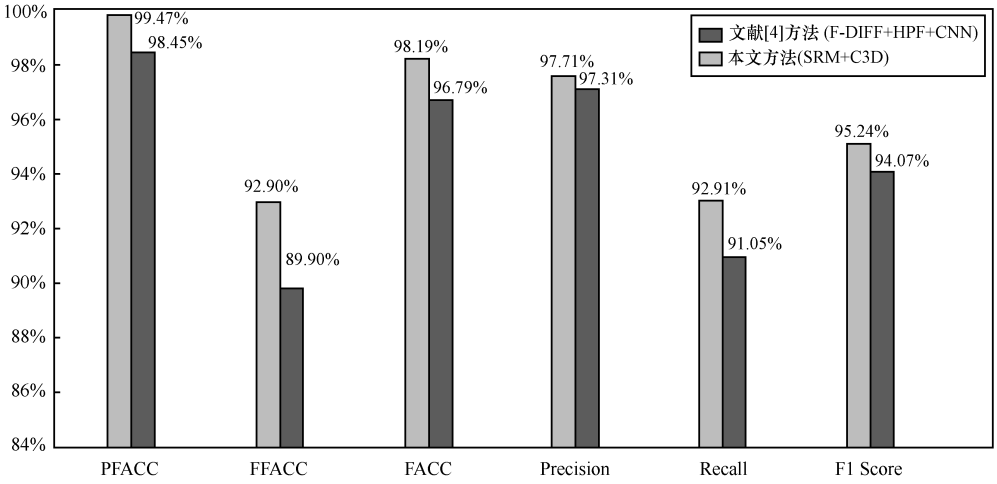

图 7 时域定位测试结果

$$\text{IOU\_mean} = \frac{1}{N_{\text{suc}}}\sum_{i}\text{IOU}_i \tag{6}$$

其中，N_{suc} 表示成功检测帧的总数，i 表示成功检测帧的下标。

将时域定位中检测出的真实篡改帧序列作为空域定位的输入，空域定位算法流程如图 8 所示。将篡改帧结合连续的前 2 帧和后 2 帧作为输入，经过两层图像处理层，再经过 C3D 神经网络特征提取器生成如图 6 所示的 96 维的特征图。在特征图上产生一系列预测框及其对应的置信度，将所有的预测框按置信度从大到小排序，取置信度最高的前 3 个框的外接框作为最终的预测区域。本文中则仅采用置信度最高的框作为最终的预测区域。图 8 中，ROI_i 表示预测框序列

对不同的篡改特征处理和提取算法，本文进行了一系列对比实验。不同算法空域定位检测结果如表 1 所示。从表 1 可以看出，本文算法中 SRM 层与 C3D 神经网络相结合可以明显改善空域定位的效果。SRM 层的作用最明显。相比于 VGG16 的 16 层 CNN，C3D 神经网络特征提取器只需要三层便可以有很好的效果，说明 C3D 神经网络更能胜任对连续帧数据的特征提取任务。

当空域定位器输入的连续帧数不同时，分别采用 1～5 帧连续帧作为空域定位器的输入。实验对比结果如表 2 所示。实验结果表明，当输入帧数越多时，空域定位效果越好；当输入帧数超过 3 帧时，空域定位效果增益减缓。

表 2　本文算法输入不同连续帧数的实验对比结果

输入	检测帧数	Suc_rate	IOU_mean
单帧	4 594	75.72%	40.40%
连续 2 帧	4 594	91.14%	46.81%
连续 3 帧	4 594	93.29%	47.94%
连续 4 帧	4 594	94.14%	49.28%
连续 5 帧	4 594	94.47%	49.07%

当空域定位结果由多个置信度最高的预测框共同决定时，分别采用置信度最高的前 1～3 个预测框的外接框作为最终预测区域，如图 9 所示。方案 a、方案 b、方案 c 分别为 3 种不同的决策方案。区域内的数字表示区域交叠次数，阴影区域分别为 3 种方案的最终预测结果。

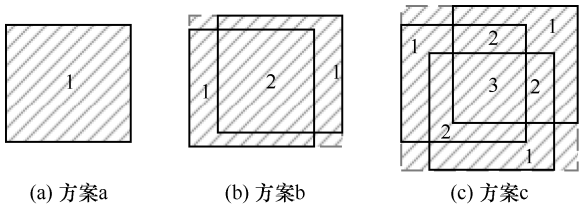

图 9　预测框的交叠区域示意

方案 a 使用置信度最高的预测框作为预测结

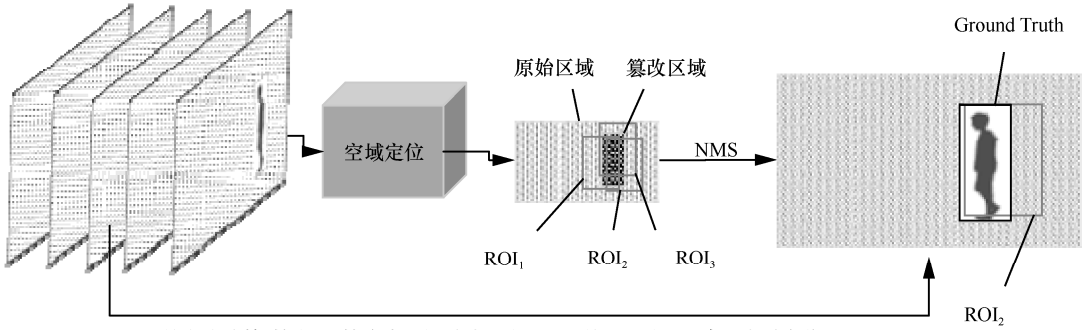

图 8　空域定位算法流程

表 1　不同算法空域定位检测结果对比

不同算法	测试帧数	Suc_rate	IOU_mean
VGG16 + RPN	4 557	68.61%	40.88%
SRM + VGG16 + RPN	4 557	89.99%	45.65%
C3D 神经网络 + RPN	4 594	46.52%	29.08%
SRM + C3D 神经网络 + RPN（本文算法）	4 594	94.47%	49.07%

果，方案 b 和方案 c 分别由置信度最高的前 2 个和 3 个预测框的外接框作为最终的预测结果。采用不同预测方案的实验结果对比如表 3 所示。随着参与预测的预测框数量的增加，成功检测率呈上升趋势，平均交并比呈下降趋势。因此，为了使定位更加准确，本文采用方案 a 作为最终预测方案。

表 3 空域定位 3 种决策方案测试结果对比

输出决策方案	检测帧数	Suc_rate	IOU_mean
方案 a	4 594	94.47%	49.07%
方案 b	4 594	96.93%	43.96%
方案 c	4 594	98.11%	37.81%

4.3.3 时空域定位测试

随机从测试集选取 10 个篡改视频来进行完整的时空域定位测试，时空域定位流程如图 10 所示。首先将测试数据打包成时域定位的输入形式，送入由特征提取器 1 和帧鉴别器组成的时域定位器，得到每一帧的判断结果，并记录其中的篡改帧；然后将记录的预测正确的篡改帧序列打包为空域定位器的输入形式，送入特征提取器 2 和空域定位器进行空域定位；最后进行计算评估。

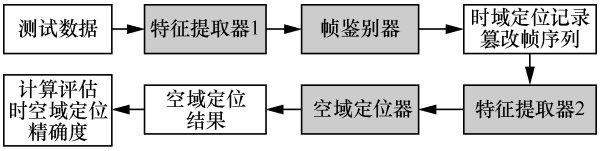

图 10 时空域定位流程

对于实验结果，需要特别说明的是：1) 在时域定位模型训练过程中，仅输入连续 5 帧的原始帧和连续 5 帧的篡改帧，没有输入连续 5 帧中既有原始帧也有篡改帧进行训练；测试集中输入的连续 5 帧视频帧的 label 以中间帧为准。因此，测试中时域定位 FACC 的主要精度损失通常会不可避免地存在于对原始帧和篡改帧交界处连续几帧的判别过程中；2) 在空域定位模型训练过程中，由于只对连续的篡改帧进行了空域定位模型的训练，因此只将在时域定位中预测正确的篡改帧序列输入空域定位器进行篡改区域的预测，这样就可以用不同的指标来表现模型分别在时域和空域的定位能力。

本文算法在部分测试视频中完整的篡改区域时空域定位结果如表 4 所示。FACC、Suc_rate 和 IOU_mean 这 3 个指标中，后 2 个指标依次在前一个指标测试结果已知的条件下才有意义。

5 结束语

基于 SRM 和 C3D 神经网络构建了 2 个深度卷积神经网络模型，分别用于视频对象移除篡改中篡改区域的时域定位和空域定位。使用 SRM 滤波器提取高频信号，使篡改特征初步显现出来；使用三维卷积神经网络 C3D 神经网络作为特征提取器，在高频信号中提取视频篡改痕迹在帧间连续性和相关性的特征；利用三维池化在特征图时空域方向进行降维，当特征图在时域方向维度降为 1 时，结合 RPN 思想进行篡改区域预测框的训练。通过多组实验对比，结果表明本文算法在时域和空域定位上都表现良好。但是，也存在网络模型计算量庞大、实验步骤复杂和对视频序列中篡改帧和原始帧的过渡区域无法精确且有效地判别的缺点。如何在提高精度的同时减少计算量、加快训练速度是未来的主要工作。

表 4 本文算法在部分测试视频中完整的篡改区域时空域定位结果

测试视频	总帧数	实际篡改帧数	预测篡改帧数	预测正确的篡改帧数	FACC	Suc_rate	IOU_mean
视频 1	285	123	108	101	89.82%	87.12%	48.64%
视频 2	284	117	116	116	99.65%	100.0%	68.20%
视频 3	292	139	134	134	98.29%	99.25%	50.77%
视频 4	284	120	118	118	99.29%	94.06%	34.77%
视频 5	291	103	77	76	90.38%	97.36%	46.27%
视频 6	284	71	67	67	98.59%	97.01%	37.27%
视频 7	291	98	98	97	99.31%	97.93%	71.45%
视频 8	292	64	63	61	98.28%	93.44%	38.41%
视频 9	291	180	182	179	98.62%	98.32%	57.26%
视频 10	292	130	132	130	99.31%	95.38%	53.27%

参考文献：

[1] 骆伟祺, 黄继武, 丘国平. 鲁棒的区域复制图像篡改检测技术[J]. 计算机学报, 2007, 30(11):112-121.
LUO W Q, HUANG J W, QIU G P. Robust tamper detection technology of regional reproduction image[J]. Chinese Journal of Computers, 2007, 30 (11): 112-121.

[2] 胡永健, AL-HAMIDI S, 王宇飞, 等. 视频篡改检测数据库的构建及测试[J]. 华南理工大学学报(自然科学版), 2017, 45(12):57-64.
HU Y J, AL-HAMIDI S, WANG Y F, et al. Construction and testing of video tamper detection database[J]. Journal of South China University of Technology (Natural Science Edition), 2017, 45(12): 57-64.

[3] 陈威兵, 杨高波, 陈日超, 等. 数字视频真实性和来源的被动取证[J]. 通信学报, 2011, 32(6): 177-183.
CHEN W B, YANG G B, CHEN R C, et al. Digital video passive forensics for its authenticity and source[J]. Journal on Communications, 2011, 32(6): 177-183.

[4] YAO Y, SHI Y Q, WENG S W, et al. Deep learning for detection of object-based forgery in advanced video[J]. Symmetry, 2017, 10(1):3.

[5] 姚晔, 胡伟通, 任一支,等. 数字视频区域篡改的检测与定位[J]. 中国图象图形学报, 2018(6): 779-791.
YAO Y, HU W T, REN Y Z, et al. Detection and location of digital video region tampering[J]. Journal of Image and Graphics, 2018(6): 779-791.

[6] 李岩, 刘念, 张斌, 等. 图像镜像复制粘贴篡改检测中的 FI-SURF 算法[J]. 通信学报, 2015, 36(5): 54-65.
LI Y, LIU N, ZHANG B, et al. FI-SURF algorithm in image mirror copy and paste tamper detection[J]. Journal on Communications, 2015, 36(5): 54-65.

[7] 黄维隽, 王士林. 一种基于优化马尔可夫特征的图像篡改盲检测算法[J]. 信息安全与通信保密, 2014(3): 93-98, 103.
HUANG W J, WANG S L. An image tampering blind detection algorithm based on optimized Markov features[J]. Information Security and Communication Security, 2014(3): 93-98, 103.

[8] 文伟, 肖志云, 彭思龙. 边缘指导的 BDCT 压缩图像的后处理算法[J]. 计算机辅助设计与图形学学报, 2005(9): 2022-2028.
WEN W, XIAO Z Y, PENG S L. Post processing algorithm of BDCT compressed image guided by edge[J]. Journal of Computer Aided Design and Graphics, 2005(9): 2022-2028.

[9] 耿庆田, 赵宏伟. 基于分形维数和隐马尔可夫特征的车牌识别[J]. 光学精密工程, 2013(12): 216-222.
GENG Q T, ZHAO H W. License plate recognition based on fractal dimension and hidden Markov features[J]. Optical Precision Engineering, 2013(12): 216-222.

[10] ZHOU P, HAN X T, MORARIU V I, et al. Learning rich features for image manipulation detection[C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1053-1061.

[11] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6): 1137-1149.

[12] 桑军, 郭沛, 项志立, 等. Faster-RCNN 的车型识别分析[J]. 重庆大学学报, 2017, 40(7): 32-36.
SANG J, GUO P, XIANG Z L, et al. Vehicle detection based on faster-RCNN[J]. Journal of Chongqing University, 2017, 40(7): 32-36.

[13] FAN Q F, BROWN L, SMITH J. A closer look at Faster R-CNN for vehicle detection[C]// Proceedings of 2016 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE Press, 2016: 124-129.

[14] 刘雨青, 黄添强. 基于时空域能量可疑度的视频篡改检测与篡改区域定位[J]. 南京大学学报(自然科学), 2014(1): 61-71.
LIU Y Q, HUANG T Q. Video tamper detection and tamper location based on energy suspicious degree in space-time domain[J]. Journal of Nanjing University (Natural Science), 2014(1): 61-71.

[15] 李倩, 王让定, 徐达文. 基于视频修复的运动目标删除篡改行为的检测算法[J]. 光电子·激光, 2016, 27(2): 182-190.
LI Q, WANG R D, XU D W. Detection to video moving object deletion based on video inpainting[J]. Optoelectronics · Laser, 2016, 27(2): 182-190.

[16] 何沛松. 基于重编码痕迹的数字视频被动取证算法研究[D]. 上海: 上海交通大学, 2018.
HE P S. Research on digital video passive forensics algorithm based on recoded traces[D]. Shanghai: Shanghai Jiaotong University, 2018.

[17] MOLCHANOV P, GUPTA S, KIM K, et al. Hand gesture recognition with 3D convolutional neural networks[C]// Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2015: 1-7.

[18] MOLCHANOV P, YANG X D, GUPTA S, et al. Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 4207-4215.

[19] 宋伟, 任栋, 于京, 等. 一种新的基于三维卷积共生梯度直方图和多示例学习的特殊视频检测算法[J]. 计算机学报, 2019, 42(1): 151-165.
SONG W, REN D, YU J, et al. A new special video detection algorithm based on 3D convolutional co-occurrence gradient histogram and multi-instance learning[J]. Chinese Journal of Computers, 2019, 42(1): 151-165.

[20] HUANG S Y, RAMANAN D. Expecting the unexpected: training detectors for unusual pedestrians with adversarial imposters[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 4664-4673.

[21] 王青, 张荣. 基于 DCT 系数双量化映射关系的图像盲取证算法[J]. 电子与信息学报, 2014, 36(9):2068-2074.
WANG Q, ZHANG R. A blind Image forensic algorithm based on double quantization mapping relationship of DCT coefficients[J]. Journal of Electronics & Information Technology, 2014, 36(9): 2068-2074.

[22] 武伟, 詹玲超. 利用颜色滤波阵列特性和模糊估计检测篡改[J]. 计算机工程与设计, 2007, 28(21): 5179-5180.
WU W, ZHAN L C. Detection of tampering using color filter array characteristics and fuzzy estimation[J]. Computer Engineering and Design, 2007, 28(21): 5179-5180.

[23] 张静, 陈静, 苏育挺. 基于滤波检测的视频区域篡改检测算法[J]. 电子测量技术, 2011, 34(11): 66-69.
ZHANG J, CHEN J, SU Y T. Detection of region-duplication forgery in the video streams[J]. Electronic Measurement Technology, 2011,

34(11): 66-69.

[24] 杨弘, 周治平, 周翠娟, 等. 基于模式噪声的手机图像篡改检测[J]. 计算机系统应用, 2013, 22(9): 210-213.
YANG H, ZHOU Z P, ZHOU C J, et al. Mobile image tampering detection based on pattern noise[J]. Journal of Computer System Applications, 2013, 22(9): 210-213.

[25] 王鑫, 鲁志波. 基于 JPEG 块效应差异的图像篡改区域自动定位[J]. 计算机科学, 2010, 37(2): 269-273.
WANG X, LU Z B. Automatic localization of image tampering area based on JPEG block effect difference[J]. Computer Science, 2010, 37(2): 269-273.

[26] 王传旭, 张祥光, 原春锋,等. 基于邻域相关性和帧间连续性的前景目标分割[J]. 数据采集与处理, 2007, 22(3): 288-291.
WANG C X, ZHANG X G, YUAN C F, et al. Foreground target segmentation based on neighborhood correlation and frame continuity[J]. Data Collection and Processing, 2007, 22(3): 288-291.

[27] 阳婷, 官洪运. 基于帧间高频能量和相关性的烟雾检测算法研究[J]. 微型机与应用, 2015(17): 36-39.
YANG T, GUAN H Y. Research on smoke detection algorithm based on high frequency energy and correlation between frames[J]. Microcomputer and Application, 2015(17): 36-39.

[28] DU T, LUBOMIR B, ROB F, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]// Proceedings of 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 4489-4497.

[29] CHEN S D, TAN S Q, LI B, et al. Automatic detection of object-based forgery in advanced video[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2016, 26(11): 2138-2151.

[30] JIANG B R, LUO R X, MAO J Y, et al. Acquisition of localization confidence for accurate object detection[C]// European Conference on Computer Vision. Berlin: Springer, 2018: 816-832.

[作者简介]

陈临强（1963– ），男，浙江临海人，杭州电子科技大学教授，主要研究方向为计算机图形学、视频实时处理、图形图像处理、定密理论与实务。

杨全鑫（1994– ），男，河南林州人，杭州电子科技大学硕士生，主要研究方向为多媒体内容安全、图形图像处理。

袁理锋（1983– ），男，浙江诸暨人，博士，杭州电子科技大学讲师，主要研究方向为图像内容安全、视觉秘密分享。

姚晔（1978– ），男，湖北随州人，博士，杭州电子科技大学副教授，主要研究方向为多媒体内容安全、视频图像智能分析。

张祯（1978– ），男，山东大同人，博士，杭州电子科技大学副教授，主要研究方向为计算机应用、保密信息化、图形图像处理。

吴国华（1970– ），男，山东济南人，博士，杭州电子科技大学教授、博士生导师，主要研究方向为保密信息化、定密理论与实务。