# 基于三维双流网络的视频目标移除篡改取证

熊礼治[1,2]，曹梦琦[1,2]，付章杰[1,2]

（1. 南京信息工程大学数字取证教育部工程研究中心，江苏 南京 210044；
2. 南京信息工程大学计算机学院、软件学院、网络空间安全学院，江苏 南京 210044）

**摘 要**：为了解决目标移除篡改视频时域检测和定位不准的问题，提出了一种基于三维双流网络的视频篡改取证方法。首先，利用空域富模型（SRM）层提取视频帧的高频信息；然后，使用改进的三维卷积（C3D）网络作为双流网络的特征提取器从高频图像帧和原始视频帧中分别提取高频信息和低频信息；最后，通过紧凑双线性池化（CBP）层将两组不同的特征向量融合成一组特征向量并用于分类检测。实验结果表明，在 SYSU-OBJFORG 数据集中，所提方法在全部视频帧中的分类准确率上具有优势，使视频目标移除篡改时域检测和定位更加准确。

**关键词**：目标移除篡改检测；视频被动取证；三维卷积；双流网络；紧凑双线性池化

**中图分类号**：TP309

**文献标识码**：A

**DOI:** 10.11959/j.issn.1000−436x.2021226

# Forensic of video object removal tamper based on 3D dual-stream network

XIONG Lizhi[1,2], CAO Mengqi[1,2], FU Zhangjie[1,2]

1. Engineering Research Center of Digital Forensics, Ministry of Education,
Nanjing University of Information Science and Technology, Nanjing 210044, China
2. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China

**Abstract:** In order to solve the problems of inaccurate temporal detection and location of the object removal tampered video, a video tamper forensics method based on 3D dual-stream network was proposed. Firstly, the spatial rich model (SRM) layer was used to extract the high-frequency information from video frames. Secondly, the improved 3D convolution (C3D) network was used as the feature extractor of the dual-stream network to extract the high-frequency information and low-frequency information from the high-frequency frame and the original video frame respectively. Finally, through compact bilinear pooling (CBP) layer, two sets of different feature vectors were fused into one set of feature vectors for classification prediction. The experimental results demonstrate that the classification accuracy of the proposed method in all video frames has an advantage in SYSU-OBJFORG dataset, which makes the temporal detection and location of object removal tampered video more accurate.

**Keywords:** object removal tamper detection, video passive forensics, 3D convolution, dual-stream network, compact bilinear pooling

## 1 引言

随着图像和视频处理算法的发展，篡改的图像和视频检测变得越来越困难[1-2]。恶意篡改图片和视频并上传至互联网[3-4]可能会导致不良的影响。因此，寻找一种有效的识别方法具有重要意义。

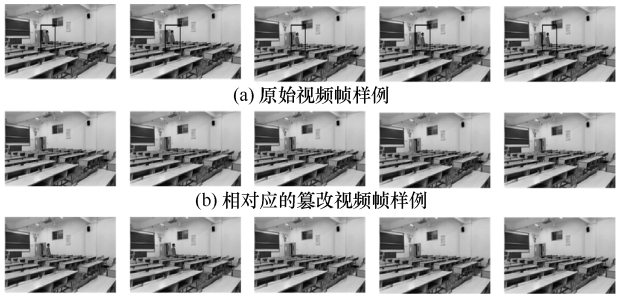
篡改视频有 2 种类型：一种是基于帧的篡改，另一种是基于内容对象的篡改。与基于帧的篡改操作（包括帧插入[5-6]、帧删除、帧复制等方法）相比，基于内容对象的篡改视频通常需要专业人员使用

**收稿日期**：2021−09−01；**修回日期**：2021−11−26

**基金项目**：国家自然科学基金资助项目（No.62172233）

**Foundation Item:** The National Natural Science Foundation of China (No.62172233)

复杂的操作技术进行[7]。如图 1 所示，视频在篡改后通常没有视觉差异，篡改操作留下的痕迹很难被发现，使这种篡改视频更有害且更难以检测。



(a) 原始视频帧样例

(b) 相对应的篡改视频帧样例

(c) 混合帧输入样例

图 1　原始视频帧与篡改视频帧样例

近年来，专家学者们在图像篡改被动检测领域取得了一些进展。肖斌等[8]针对图像 copy-move 篡改提出了一种基于分组尺度不变特征的快速检测方法。李岩等[9]提出一种 FI-SURF（flip invariant speeded-up robust features）算法，该算法能够检测出图像镜像 copy-move 篡改。Liang 等[10]提出了一种集成中心像素映射、最大零连通分量标记和片段拼接检测的算法，极大提高了检测效率。Zhang 等[11]基于联合概率密度矩阵和离散余弦变化系数等抽象统计特征提出了一种混合取证方法，提高了算法稳健性。随着深度学习技术的发展，新的基于神经网络的深度学习方法被提出。由马里兰大学和 Adobe 公司提出的基于 Faster R-CNN（region-based convolutional neural network）的双流网络[12-13]可以检测出篡改图像中的伪造区域。王珠珠[14]使用 U-Net 提取图像中多阶段的特征信息用于检测。

与单个图像相比，篡改视频可以从相邻帧中获得用于操作篡改区域的相关信息，然后进行编辑和修补，使检测更加困难。目前，数字视频篡改检测算法可以分为四类：1) 基于噪声模式的算法；2) 基于像素相关的算法；3) 基于视频内容特征的算法；4) 基于抽象统计特征的算法。

基于噪声模式的算法通过提取数字视频在篡改后留下的噪声痕迹，进行视频完整性检测。Ding 等[5-6]利用帧插入篡改残留的伪影和信号残差检测运动补偿帧率上转换视频。Hsu 等[15]提出了一种基于篡改噪声的相关算法来定位篡改区域。该算法采用宏块计算篡改噪声的相关系数，并且认为相关系数服从高斯混合模型（GMM，Gaussian mixture model）。当宏块的相关系数明显偏离阈值时，该宏块被认为是篡改的。Chen 等[1]首先提出了基于目标篡改的视频检测问题，创建了 SYSU-OBJFORG 数据集，并提出了一种基于运动残差的时序篡改检测算法，使用共谋算子从视频帧序列中生成运动残差静态图像并使用图像篡改算法进行检测。

基于像素相关的算法利用篡改操作会破坏数字视频中相邻像素在时空方向上相关性的特点，通过寻找像素相关性的异常变化实现篡改被动取证。Wang 等[16]利用监控摄像机插值生成视频帧的特点，提出了基于期望最大化（EM，expectation maximization）的插值周期检测和定位算法。Bestagini 等[17]通过计算相邻两帧的像素差值，当差值为零时，确定像素点为篡改，但应用条件严格。Liu 等[18]利用亮度和对比度作为特征来衡量前景和背景之间的相似性，然后通过识别这些块的前景和背景之间的特征不一致性来检测篡改。Sitara 和 Mehtre[19]提出了一种基于像素差值的篡改视频检测方法。

基于视频内容特征的算法通过对篡改残留痕迹、异常光流变化等内容特征进行分析实现完整性检测。Zhang 等[20]提出使用篡改后留下的伪影作为回火检测的检测依据进行篡改视频检测。Li 等[21]提出了一种通过分析运动向量的异常特征来检测在静止背景视频中篡改运动目标的算法。Aloraini 等[22]利用空间分解、时间滤波和序列分析来检测和定位基于目标移除的篡改视频。Zhong 等[23]利用最佳帧间匹配算法识别从视频中提取的多维密集矩阵特征来识别帧间篡改视频，并根据设定的阈值定位帧间的篡改区域。

基于抽象统计特征的算法利用篡改后的区域像素值抽象统计特征与原始区域不同的特点实现取证。Chen 等[24]提出了一种被动取证算法，通过计算视频目标可变宽边界区域的统计特征，使用支持向量机（SVM, support vector machine）作为分类器并进行训练，用于鉴别视频对象的真实性。Pandey 等[25]提出了一种时空联合 copy-move 篡改视频区域检测与定位算法。通过在空域提取尺度不变特征变换（SIFT, scale-invariant feature transform）特征，在时域提取噪声特征，并计算相互系数完成篡改检测和区域定位。

传统的被动取证方法需要手动设计检测特征，存在识别效率、准确率低和稳健性差等问题，无法满足应用需求。近年来，深度学习的发展为视频篡改检测带来新的研究方向。利用深度学习方法来进行篡改视频检测应归纳为基于噪声模式的算法。Yao 等[26]提出

了一种基于目标移除篡改视频检测的 CNN。利用相邻两帧之间的帧差，通过高通滤波器提取高频信息并输入 CNN 进行训练。CNN 可以自动学习篡改特征，提高了检测的效率和准确率。翁韶伟等[27]利用 Inception 网络从灰度运动残差中提取特征信息进行篡改检测和定位。陈临强等[28]提出了一种时空域定位检测网络，在此方案基础上，Yang 等[29]提出了一种时空三叉戟网络（STN, spatiotemporal trident network），用于视频被动取证中目标移除篡改检测和定位，他们使用连续的 5 帧作为网络输入，通过空域富模型[30]（SRM, steganalysis rich model）滤波和三维卷积[31-32]（C3D, 3D convolution）提取特征编码，然后利用双向长短时记忆网络（BiLSTM, bi-directional long short-term memory）解码特征来检测时域篡改，具有很高的分类准确性。Wang 等[33]通过实验表明，CNN 倾向于先学习图像中与标签相对应的低频信息，然后学习高频信息来进一步提高分类准确率，因此高频信息对于特征提取网络同样重要。

本文使用以改进的 C3D 网络为主干特征提取器的双流网络来融合视频帧单元的低频、高频和时域特征，提出了一种视频目标移除篡改取证方法。首先，利用 SRM 滤波器提取视频帧的高频信息，并和原始视频帧中的低频信息共同作为网络输入，通过特征提取器获得 2 个包含不同频域信息的特征向量；然后，使用紧凑双线性池化（CBP, compact bilinear pooling）融合包含不同信息的特征向量；最后，将融合后的特征向量送入分类器进行分类预测。该方法可以充分利用视频中的低频、高频和时间信息，通过网络自动学习篡改视频帧的特征，实现篡改视频帧的时域定位。本文的主要贡献如下。

1) 提出一种改进的 C3D 网络用来提取视频帧序列的时间信息，使用卷积核大小为 1×1 的卷积层来融合特征以及降低特征向量维度。

2) 利用 CBP 融合低频信息流的低频信息特征和高频信息流的高频信息特征，并将融合后的特征向量用于时域检测和定位。

3) 提出一种具有低频信息流和高频信息流的三维双流网络，并将改进的 C3D 网络作为特征提取器提高时域检测的准确性。

## 2 系统模型

### 2.1 C3D 网络

在二维卷积（C2D, 2D convolution）网络中，卷积仅从空间维度计算特征，只能应用于二维特征映射，而不能处理视频数据的时间信息。在分析视频数据问题时，时间信息作为不同于图像而特有的信息，对预测分类结果有重要作用。C3D 能够将多个连续的视频帧堆叠成一个立方体，使用三维卷积核一次通过立方体的多个维度计算结果以获取连续的视频时间信息。

### 2.2 紧凑双线性池化

本文采用双线性池化[34]（BP, bilinear pooling）对融合特征进行细粒度分类。对于不同的 2 个特征提取器从同一位置提取出来的特征 $x$ 和特征 $y$，通过双线性池化操作，在保留空间位置特征的前提下，融合成一个特征向量用于分类以提高检测的置信度。BP 的精准定义如下。

对于图像 $I$ 中位置 $l$ 提取出的 2 个特征 $f_A(l,I)\in\mathbb{R}^{T\times M}$ 和 $f_B(l,I)\in\mathbb{R}^{T\times N}$（其中 $f_A$ 和 $f_B$ 为特征提取函数，$T$ 为通道数，$M$ 和 $N$ 为维度数）。设 $Z=MN$，BP 操作定义如下。

$$\begin{aligned}
&\boldsymbol{B}=\text{bilinear}(I)=f_A^T(l,I)f_B(l,I) && \in\mathbb{R}^{M\times N},\\
&\boldsymbol{P}=\sum_l \boldsymbol{B} && \in\mathbb{R}^{M\times N},\\
&\boldsymbol{x}=\text{vec}(\boldsymbol{P}) && \in\mathbb{R}^{Z\times 1},\\
&\boldsymbol{y}=\text{sign}(x)\sqrt{|\boldsymbol{x}|} && \in\mathbb{R}^{Z\times 1},\\
&\boldsymbol{z}=\frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2} && \in\mathbb{R}^{Z\times 1} \qquad (1)
\end{aligned}$$

从图 2 中可以直观地理解双线性池化，具体如下。

1) 将图像 $I$ 同一位置的两组不同特征融合（相乘）为一个矩阵 $\boldsymbol{B}$；

2) 对矩阵 $\boldsymbol{B}$ 中所有位置 $l$ 进行池化获得矩阵 $\boldsymbol{P}$；

3) 重塑矩阵 $\boldsymbol{P}$ 为双线性向量 $\boldsymbol{x}$；

4) 对向量 $\boldsymbol{x}$ 做矩归一化和 L2 归一化，获得特征向量 $\boldsymbol{z}$ 用于分类。

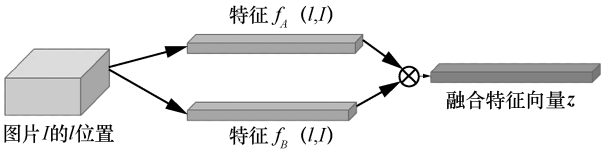



图 2 图像 $I$ 中位置 $l$ 的双线性池化的过程

然而，传统的双线性池化具有特征向量 $\boldsymbol{z}$ 的维度过高的问题。融合后的特征向量 $\boldsymbol{z}$ 的维度等于特征向量 $\boldsymbol{x}$ 维度和特征向量 $\boldsymbol{y}$ 维度的乘积。为了在不降低性能的情况下，减少计算消耗和加快训练速

度，本文使用 CBP 来融合 2 个特征向量。

分类操作可以被看成式(2)所示的线性核机器。

$$\left\langle \boldsymbol{x}_{I_a}, \boldsymbol{x}_{I_b} \right\rangle = \left\langle \mathrm{vec}\left(\boldsymbol{P}_{I_a}\right), \mathrm{vec}\left(\boldsymbol{P}_{I_b}\right) \right\rangle = \left\langle \mathrm{vec}(\sum_{l_a} a_{l_a} a_{l_a}^{\mathrm{T}}), \mathrm{vec}(\sum_{l_b} b_{l_b} b_{l_b}^{\mathrm{T}}) \right\rangle = \sum_{l_a}\sum_{l_b} \left\langle \mathrm{vec}\left(a_{l_a} a_{l_a}^{\mathrm{T}}\right), \mathrm{vec}\left(b_{l_b} b_{l_b}^{\mathrm{T}}\right) \right\rangle = \sum_{l_a}\sum_{l_b} \left\langle a_{l_a}, b_{l_b} \right\rangle^2 \tag{2}$$

如果可以找到一种低维投影方式满足 $\left\langle \phi(a), \phi(b) \right\rangle \approx \left\langle a, b \right\rangle^2$（其中 $\phi()$ 为特征提取器 $\phi(a) \in \mathbb{R}^d$，$d \ll Z$），那么可以得到式(3)。

$$\left\langle \boldsymbol{x}_{I_a}, \boldsymbol{x}_{I_b} \right\rangle = \sum_{l_a}\sum_{l_b} \left\langle a_{l_a}, b_{l_b} \right\rangle^2 \approx \sum_{l_a}\sum_{l_b} \left\langle \phi(a_{l_a}), \phi\left(b_{l_b}\right) \right\rangle = \left\langle \sum_{l_a} \phi(a_{l_a}), \sum_{l_b} \phi(b_{l_b}) \right\rangle \tag{3}$$

即当 $\left\langle \phi(a), \phi(b) \right\rangle \approx \left\langle a, b \right\rangle^2$ 时，可以得到 $\boldsymbol{x}_{I_a} \approx \sum_{l_a} \phi\left(a_{l_a}\right)$。Gao 等[35]提出了 2 种算法：随机麦克劳林（RM, random maclaurin）投影和张量简单（TS, tensor sketch）投影。相对 RM 来说，TS 具有更少的计算消耗，符合轻量级网络的设计需求，其流程如算法 1 所示。因此，本文提出的网络使用 TS 将 SRM 流和 RGB 流得到的各 128 维的特征向量融合为 4 096 维向量，相比于传统双线性池化降低了 12 288 维。

**算法 1**　张量简单投影

**输入**　向量 $\boldsymbol{x} \in \mathbb{R}^c$

1) 生成随机但固定的 $h_k \in \mathbb{N}^c$ 和 $s_k \in \{+1, -1\}^c$，其中 $h_k(i)$ 选自 $\{1, 2, \cdots, d\}$，$s_k(i)$ 选自 $\{+1, -1\}$，并且 $k = 1, 2$；

2) 定义投影方法 $\Psi(\boldsymbol{x}, h, s) = \{(Q_x)_1, \cdots, (Q_x)_d\}$，其中 $(Q_x)_j = \sum_{t:h(t)=j} s(t)\boldsymbol{x}_t$；

3) 定义 $\phi_{\mathrm{TS}}(s) \equiv \mathrm{FFT}^{-1}(\mathrm{FFT}(\Psi(\boldsymbol{x}, h_1, s_1))) \circ \mathrm{FFT}(\Psi(\boldsymbol{x}, h_2, s_2))$，其中 $\circ$ 代表元素乘法；

**输出**　特征图 $\phi_{\mathrm{TS}}(x) \in \mathbb{R}^d$，满足 $\left\langle \phi_{\mathrm{TS}}(\boldsymbol{x}), \phi_{\mathrm{TS}}(\boldsymbol{y}) \right\rangle \approx \left\langle \boldsymbol{x}, \boldsymbol{y} \right\rangle$。

## 3　方法设计

### 3.1　低频信息流

低频信息流有 2 个功能。首先，低频信息流可以学习到篡改区域光线不一致、边界对比度高等篡改特征。受物体光线变换、反射以及人物遮挡等原因的影响，不同视频帧帧间的光线都会变换，但光线在同一帧内表现出相对一致性。对于目标移除篡改，篡改区域通常是从其他相邻视频帧复制过来的，篡改区域的光线与其他区域不同，会呈现出不一致的光线特征。如图 3 所示，篡改帧中标记区域的光线与原始帧相比出现不连续，且在 SRM 滤波后高频图像中存在不规则噪声，网络可以从输入的连续视频帧中学习到此类差异特征。其次，低频信息流可以从连续输入中捕获视频内容的语义信息。针对如图 1 (c)所示的混合帧输入样例，前两帧为原始帧，后三帧为篡改帧，目标人物被移除导致篡改区域前后的语义信息不一致，这些不一致信息对目标移除篡改检测有效。

(a) 原始帧

(b) 篡改帧

图 3　一对原始帧和篡改帧

### 3.2　高频信息流

低频信息流更侧重于学习低频语义信息，并不能处理所有的信息，对于精心处理后的篡改视频帧，帧内光线变化不明显，低频信息流不能很好地学习到篡改痕迹。然而，篡改操作会改变视频帧的高频信息，因此使用特征提取器从高频信息中获取特征向量用于网络训练对于进一步提高分类准确率也非常重要。

视频篡改通常经过 3 个步骤：解压缩成帧、篡改视频帧和重新压缩成视频。这种篡改过程会在高频区域留下痕迹，由于篡改操作通常从同源视频的其他相邻帧截取目标区域并复制到篡改帧以保证视觉完整性，因此篡改区域的高频信息与其他区域不一致。通过实验发现，篡改区域的高频信息在相邻篡改帧间拥有较大的连续性和相似性。

篡改区域高频信息与原始区域高频信息相关性较小，同一区域内相邻像素间构建残差会在高频区域出现不同的统一性特征，且在区域交界处会显示明显不规则噪声。SRM 已经被证明在高频信息提取上效果显著，通过对目标像素和相邻像素计算残差并对滤波器的输出进行量化和截断，提取共现信息作为最终的特征。将滤波后生成的拥有高频信息的图像输入高频信息流中，使网络能够学习到篡改区域与原始区域不一致的高频噪声信息。本文使用 3 个高频滤波核，SRM 层输入和输出通道为 3，卷积核大小为 5×5×3，能够在适当的计算消耗下取得良好的效果，高频滤波核的具体参数如图 4 所示。通过 SRM 滤波后图像更强调高频噪声信息而不是低频语义信息，特征提取网络可以学习到篡改区域与原始区域不一致的高频信息特征，用于进一步提高分类准确率。

$$\frac{1}{4}\times\begin{bmatrix}0&0&0&0&0\\0&-1&2&-1&0\\0&2&4&2&0\\0&-1&2&-1&0\\0&0&0&0&0\end{bmatrix};\ \frac{1}{12}\times\begin{bmatrix}-1&2&-2&2&-1\\2&-6&8&-6&2\\-2&8&-12&8&-2\\2&-6&8&-6&2\\-1&2&-2&2&-1\end{bmatrix};\frac{1}{8}\times\begin{bmatrix}0&0&0&0&0\\0&-1&-2&-1&0\\0&-2&2&-2&0\\0&-1&-2&-1&0\\0&0&0&0&0\end{bmatrix}$$

图 4　高频滤波核的具体参数

### 3.3　双流网络

传统 C3D 网络[36]使用三维池化层进行跨通道池化，容易导致对分类重要的特征信息被模糊，降低网络检测准确率。本文提出的改进 C3D 网络使用卷积核大小为 1×1 的卷积层替代池化层进行跨通道融合特征和降低维度，并在此基础上提出如图 5 所示的双流网络。首先，输入连续 5 帧原始视频帧单元，通过 SRM 滤波层生成高频噪声图像。然后，将 2 种类型的图像分别输入相应的 C3D 网络中，分别提取 128 维特征向量。最后，通过 CBP 层将 2 个 128 维特征向量融合为 4 096 维特征向量，并将其输入二分类器中用于预测输入数据单元的中间帧是否为篡改帧。由于篡改检测问题可以被视为二分类问题，因此模型损失函数使用交叉熵损失。设预测结果为正样本的概率为 $p$，则负样本概率为 $1-p$，损失函数如式(4)所示。

$$L=\frac{1}{N}\sum_{i}-[y_i\log(p_i)+(1-y_i)\log(1-p_i)] \quad (4)$$

其中，$y_i$ 为数据单元 $i$ 的标签，正样本为 1，负样本为 0；$p_i$ 为数据单元 $i$ 被预测为正样本的概率。

### 3.4　实现细节

使用卷积核大小为 1×1 的卷积层的 C3D 网络如图 6 所示。C3D 组按照 C3D、P3D 和 C3D（1×1）顺序组合，C2D 组按照 C2D 和 P2D 顺序组合。MP3D 表示最大池化（3D max pooling）层，C3D 表示三维卷积层，P3D 表示三维池化（3D pooling）层，C3D（1×1）表示卷积核大小为 1×1 的三维卷积层，C2D 表示二维卷积层，P2D 表示二维池化（2D pooling）层，GAP 表示全局平均池化（global average pooling）层，Reshape 表示变换层，用于去除冗余维度。在每个卷积层之后均进行批处理归一化运算和激活操作。

MP3D 用于缩小输入帧的大小，将 720 像素×720 像素减小为 240 像素×240 像素，减少计算量的同时便于设计网络。前 3 个 C3D 层卷积核大小为 3×3×3。在每个 C3D 层之后，执行 P3D，其步幅为 1×2×2。P3D 层之后是卷积核大小为 1×1 的 C3D 层。在第一次变换层后，输入维度从三维减少到二维。接着，使用两组卷积核大小为 1×1 的 C2D 和步幅为 2×2 的 P2D 将输入向量维度降为 1。在经过 GAP 层和最后一个变换层之后得到一个 128 维的特征向量。特征提取结束后输入二分类器，用于预测中间帧是否为篡改帧。低频信息流和高频信息流各输出 128 维特征向量，经过 CBP 特征融合得到 4 096 维

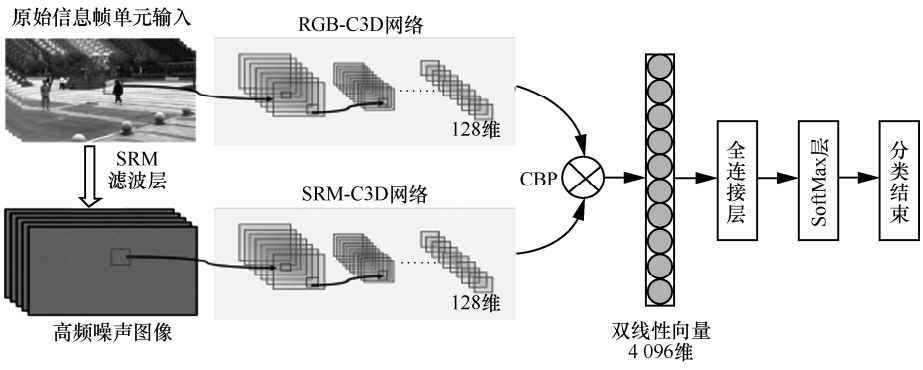



图 5　双流网络模型结构

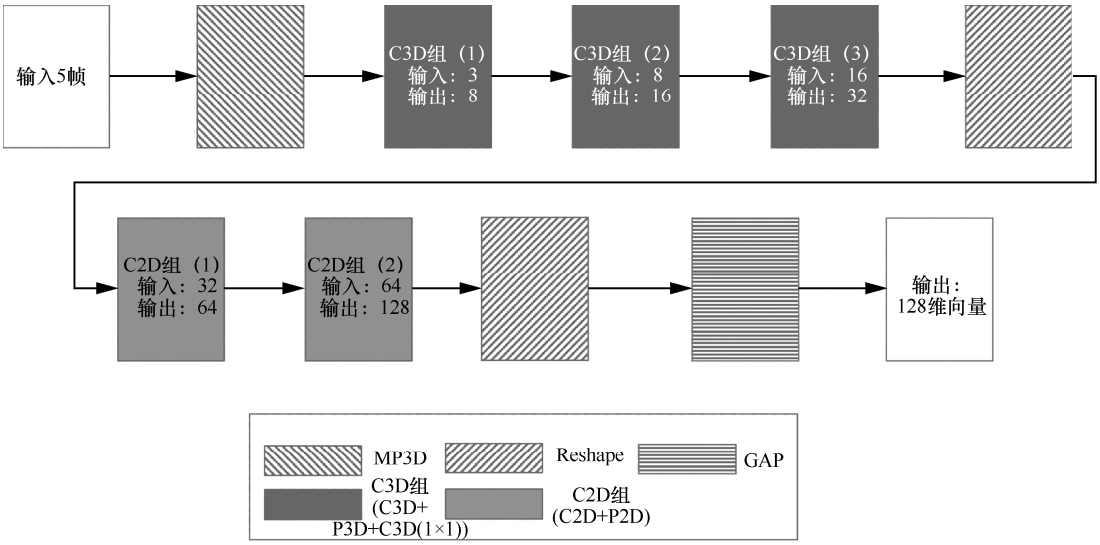



图 6　使用卷积核大小为 1×1 的卷积层的 C3D 网络

特征向量。最后，经过两层全连接层将特征向量维度降为 2 后使用 Softmax 回归模型对向量进行归一化分类输出预测值。

## 4　实验比较与分析

### 4.1　数据集策略

本文使用 SYSU-OBJFORG 数据集，这是目前最大的目标移除篡改视频数据集，有 100 对原始和篡改视频。视频拍摄场景为教学楼走廊，篡改目标包括各种运动状态的物体，且篡改区域大小不同。视频平均长度为 10 s，视频帧率为 25 f/s，码率为 3 Mbit/s 并且所有视频都以 H.264/MPEG-4 格式进行压缩。

数据量对于网络参数训练具有重要影响。用于训练的样本数据量越大，网络学习共有特征信息的能力越强，能够更好地拟合非线性函数提高分类准确率。SYSU-OBJFORG 数据集每段视频平均篡改帧数为 100，原始帧和篡改帧的样本数量较少且不对等，会导致网络欠拟合、分类准确率低，不能直接用于网络训练学习。因此，本文提出一种非对称采样方法，对原始视频进行欠采样，而对篡改视频进行过采样。通过这种方法为网络训练生成足够的数据样本。

原始帧与篡改帧的采样方法如图 7 所示。数据集视频的分辨率为 1 280 像素×720 像素。为了方便网络的设计和数据的处理，将视频帧裁剪为 720 像素×720 像素作为网络输入。对于原始视频帧，设置裁剪步长为 20 像素，每 5 帧（目标帧及前后各两帧）对齐裁剪，并将剪裁结果保存为一组数据单元，其标签设置为中间目标帧的标签。对于篡改帧，设置裁剪步长为 10 像素，裁剪的区域范围设置为$[\text{Left}, \text{Right}]$，并使用与原始视频帧相同的裁剪方法。

$$\text{Left}=\begin{cases}0\ , & \text{core}<360\\ \text{LB}-360\ , & \text{core}\geqslant 360\end{cases}$$
$$\text{Right}=\begin{cases}\text{RB}+360, & \text{core}<920\\ 1280\ , & \text{core}\geqslant 920\end{cases}\tag{5}$$

其中，LB 是篡改区域的左侧边界，RB 是篡改区域的右侧边界，core 是篡改区域中心坐标。

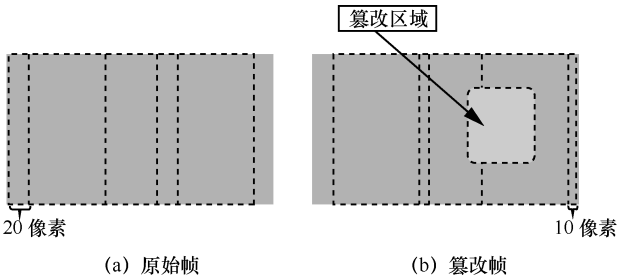



图 7　原始帧与篡改帧的采样方法

上述方法生成的训练样本只需包含篡改区域的任意部分，即可用于网络训练，增强样本的多样性，降低视频内容的运动状态、篡改区域大小等对网络检测准确率的影响，弱化网络对特定样本的依赖性，增加算法的稳健性。验证集和测试集的裁剪方式与训练集相同。此外，在网络训练阶段进行数据增强。由于视频与图像不同，具有方向性和时序性，数据单元组在输入网络前只进行随机水平翻转

和垂直翻转。验证阶段不进行数据增强。

### 4.2 实验设置

本文提出的双流网络是基于 Pytorch 框架实现的，运行在 Windows 10 系统上，使用 NVIDIA GeForce GTX1050ti 4 GB GPU，选择 Adam 作为优化器。RGB 流和 SRM 流分别训练 10 个 epoch，网络的学习速率设置为 0.001，每 4 个 epoch 减少 0.1。然后利用训练好的模型参数训练整体网络，整体双流网络训练 8 个 epoch，网络的学习率设置为 0.000 1，每 3 个 epoch 减少 0.1。网络模型参数总和为 64 180，训练总时长为 80 h。

网络训练阶段的批大小设置为 16，即输入大小为 16×5×(720×720)×3，其中 16 代表批大小，5 为连续帧数量，3 为通道数。验证阶段批大小设置为 8。数据集随机分为 3 个部分，训练集有 70 对视频，验证集和测试集分别有 15 对视频。训练阶段从训练集中选取 20 000 个原始帧数据单元和 20 000 个篡改帧数据单元作为训练数据，从验证集中选取 10 000 个原始帧数据单元和 10 000 个篡改帧数据单元作为验证数据。当验证阶段的损失函数趋于收敛时，选择精度最高的训练模型进行测试。在测试阶段，从测试集中选择 10 000 个原始帧数据单元和 10 000 个篡改帧数据单元作为测试数据。所有的数据单元均从相应数据集中随机选取，数据集划分的细节和训练集遍历次数如表 1 所示。

**表 1 视频数据集的划分细节和训练集遍历次数**

| 方法 | 训练集/段 | 验证集/段 | 测试集/段 | 训练集遍历次数/次 |
|---|---|---|---|---|
| 文献[26]方法 | 40 (原始)+40 (篡改) | 10 (原始)+10 (篡改) | 50 (原始)+50 (篡改) | 12 000 |
| 文献[27]方法 | 40 (原始)+40 (篡改) | 10 (原始)+10 (篡改) | 50 (原始)+50 (篡改) | 40 000 |
| 文献[28]方法 | 50 (原始)+50 (篡改) | 10 (原始)+10 (篡改) | 40 (原始)+40 (篡改) | 400 |
| 文献[29]方法 | 60 (原始)+60 (篡改) | 20 (原始)+20 (篡改) | 20 (原始)+20 (篡改) | 400 |
| 本文方法 | 60 (原始)+60 (篡改) | 20 (原始)+20 (篡改) | 20 (原始)+20 (篡改) | 28 |

### 4.3 实验结果

测试阶段的批大小为 8，即输入的数据组大小为 8×5×(720×720)×3。对于每一组数据单元，网络的输出是中间目标帧的分类结果。在篡改视频的时域定位测试中，本文使用 Chen 等[1]定义的 6 个评价指标。

$$\mathrm{PFACC}=\frac{\sum \mathrm{correctly\_classified\_pristine\_frame}}{\sum \mathrm{pristine\_frame}}$$

$$\mathrm{FFACC}=\frac{\sum \mathrm{correctly\_classified\_forged\_frame}}{\sum \mathrm{forged\_frame}}$$

$$\mathrm{FACC}=\frac{\sum \mathrm{correctly\_classified\_frame}}{\sum \mathrm{all\_the\_frame}}$$

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

$$\mathrm{F1值}=\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{6}$$

其中，PFACC（pristine frame accuracy）是原始帧分类正确率，FFACC（forged frame accuracy）是篡改帧分类正确率，FACC（frame accuracy）是所有帧分类正确率，Precision、Recall 和 F1值 可以通过计算得出。TP（true positive）是篡改帧正确分类数量，FP（false positive）是原始帧错误分类数量，FN（false negative）篡改帧错误分类数量。

本文使用卷积核大小为 1×1 的三维卷积层代替三维池化层进行降维操作，1×1 卷积操作是将不同通道上同一位置的特征线性组合，在跨通道信息交互的同时，进一步融合位置信息和时间信息。在实现降维的同时避免了池化层可能将利于分类的重要特征掩盖的情况。为了证明卷积核大小为 1×1 的三维卷积层在特征融合和降维操作上的优势，本文设计了对比试验。实验选择 STN 中提出的单支 C3D 网络[29]与本文提出的改进 C3D 网络进行比较。网络在相同数据集中进行训练和测试，共训练 5 个 epoch，网络学习速率设置为 0.001，每 3 个 epoch 下降 0.1。实验结果如图 8 所示，使用 1×1 卷积核的 C3D 网络比不使用 1×1 卷积核的 C3D 网络在

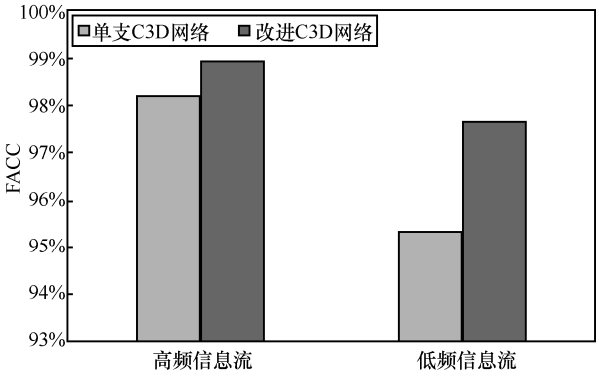



图 8 单支 C3D 网络和改进 C3D 网络检测结果比较

低频信息流中 FACC 提升 2.34%，在高频信息流中 FACC 提升 1.31%。2 种模型高频信息流和低频信息流训练速度均为 6.17 组/秒和 9.09 组/秒。实验结果证明使用卷积核大小为 1×1 的三位卷积层代替池化层进行特征融合和降维操作在相同计算消耗的情况下可以进一步提高网络分类预测精度。

CBP 融合后的特征向量的维度数量会对分类结果产生影响，数量过多会增加计算消耗或出现不利于预测分类的冗余特征。本文分别选择 512 维、2 048 维、4 096 维、8 192 维、16 384 维融合向量进行实验。该实验对未使用预训练参数的整体双流网络进行训练，网络训练 7 个 epoch，学习速率设置为 0.001，每 4 个 epoch 下降 0.1，实验结果如图 9 所示。随着融合向量的维度增加，网络的分类准确率得到提升。在维度较小的时候，分类准确率随着维度数量增加提高明显，当维度增加至 4 096 维之后准确率出现小幅度下降并保持稳定。融合特征的维度数量同样影响网络的训练收敛速度，512 维向量在 6 个 epoch 后开始收敛，当特征维度的数量增加至 8 192 维后，网络在一个 epoch 后达到较高准确率并开始收敛。实验结果表明增加融合后用于分类的特征向量的维度数量可以增加分类的准确率，但是当维度到达一定数量后分类准确率不再增加并保持稳定。增加特征向量的维度会增加计算消耗，但可以加快网络收敛的速度，能够使网络在较短的时间内达到较高的分类准确率。在衡量计算消耗和分类准确率后，本文提出的网络选择通过 CBP 层将 2 个 128 维向量合成为 4 096 维向量后用于分类。

上述实验证明，使用卷积核大小为 1×1 的卷积层和选取 4 096 维作为融合后特征的维度数量能够提升网络对视频帧的分类准确率。在此基础上构建整体双流网络并设置视频帧时域分类准确率实验，实验结果如表 2 所示。与多种深度学习方法进行比较，本文所提方法具有更好的性能，所有评价指标均达到最高，全部帧的分类准确率达到 99.52%。实验结果与文献[29]相比，PFACC 提高 0.36%，FFACC 提高 0.43%，Precision 提高 1.72%，F1 值提高 1.07%，特别是在 FFACC、Recall 和 F1 值中提升明显，这表明本文方法不仅对篡改帧分类有很高的准确率，并且对原始帧分类也有很高的准确率。

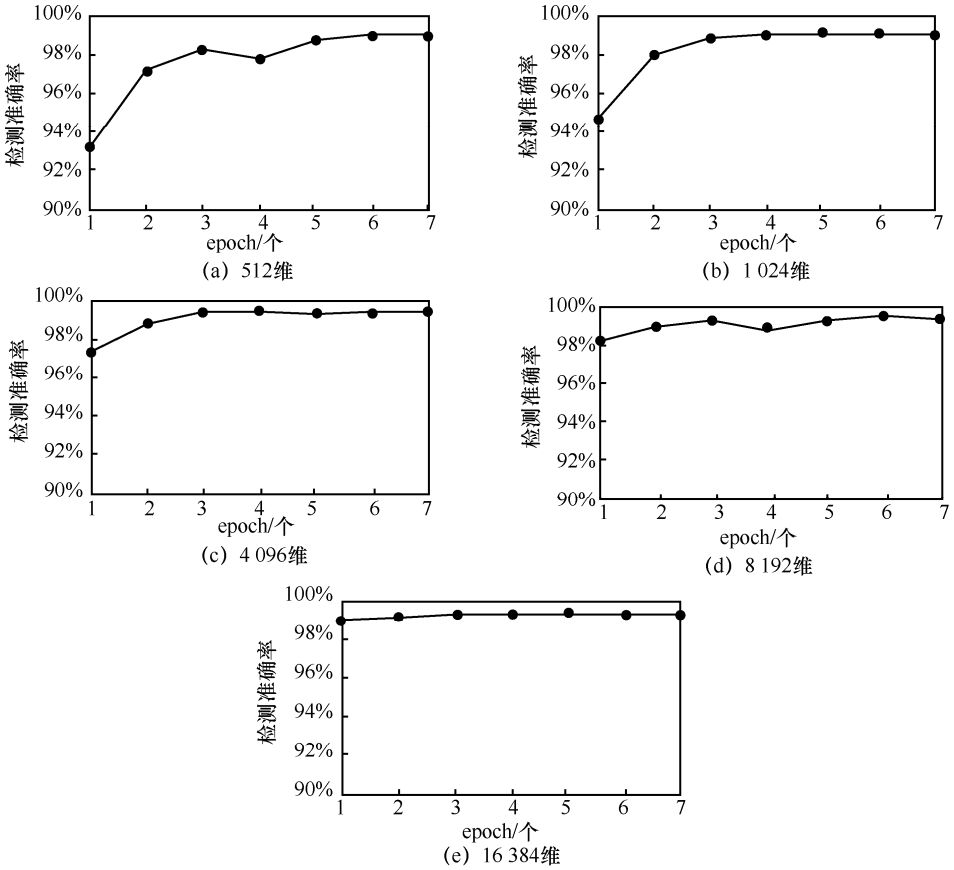



图 9　融合后向量的维度对分类结果的影响

表 2 不同方法检测结果比较

| 方法 | PFACC | FFACC | FACC | Precision | Recall | F1 值 |
|---|---|---|---|---|---|---|
| 文献[26]方法 | 98.45% | 91.05% | 96.79% | 97.31% | 91.05% | 94.07% |
| 文献[27]方法 | 99.42% | 91.30% | 97.60% | 96.70% | 91.30% | 94.06% |
| 文献[28]方法 | 99.47% | 92.91% | 98.19% | 97.71% | 92.91% | 95.24% |
| 文献[29]方法 | 99.50% | 98.75% | 99.34% | 98.14% | 98.75% | 98.44% |
| 本文方法 | **99.86**% | **99.18**% | **99.52**% | **99.86**% | **99.18**% | **99.51**% |

## 5 结束语

本文提出了一种基于三维双流网络的视频目标移除篡改取证方法。将连续的 5 帧原始视频帧作为网络的输入来预测中间帧的分类标签。使用低频信息流和高频信息流分别从输入中提取低频和高频信息，可以解决混合帧样本输入的问题。使用卷积核大小为 1×1 的改进 C3D 网络作为提取器可以从连续视频帧中更充分地提取时间信息。此外，使用 CBP 融合特征向量可以融合低频、高频和时间信息，使分类更准确。本文提出的网络是一个轻量级、具有较少参数的网络，在硬件设备不足的情况下，可以使用较少的数据量和训练时间，达到较好的分类准确率和稳健性。然而，本文方法依赖不同的 SRM 滤波参数以适应不同类型的视频，缺乏稳健性并且无法实现空域定位。寻找一个通用的篡改特征和实现篡改视频帧空域定位是今后的主要工作。

## 参考文献：


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

[2] 冯春晖, 徐正全, 郑兴辉, 等. 数字可视媒体取证[J]. 通信学报, 2014, 35(4): 155-165.
FENG C H, XU Z Q, ZHENG X H, et al. Digital visual media forensics[J]. Journal on Communications, 2014, 35(4): 155-165.

[3] 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, 23(6): 779-791.
YAO Y, HU W T, REN Y Z, et al. Detection and localization of digital video regional tampering[J]. Journal of Image and Graphics, 2018, 23(6): 779-791.

[4] ALORAINI M, SHARIFZADEH M, SCHONFELD D. Sequential and patch analyses for object removal video forgery detection and localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(3): 917-930.

[5] DING X L, ZHU N B, LI L D, et al. Robust localization of interpolated frames by motion-compensated frame interpolation based on an artifact indicated map and tchebichef moments[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(7): 1893-1906.

[6] DING X L, YANG G B, LI R, et al. Identification of motion-compensated frame rate up-conversion based on residual signals[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(7): 1497-1512.

[7] LIAO X, YU Y B, LI B, et al. A new payload partition strategy in color image steganography[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(3): 685-696.

[8] 肖斌, 景如霞, 毕秀丽, 等. 基于分组 SIFT 的图像复制粘贴篡改快速检测算法[J]. 通信学报, 2020, 41(3): 62-70.
XIAO B, JING R X, BI X L, et al. Fast copy-move forgery detection algorithm based on group SIFT[J]. Journal on Communications, 2020, 41(3): 62-70.

[9] 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): 58-69.
LI Y, LIU N, ZHANG B, et al. FI-SURF algorithm for image copy-flip-move forgery detection[J]. Journal on Communications, 2015, 36(5): 58-69.

[10] LIANG Z S, YANG G B, DING X L, et al. An efficient forgery detection algorithm for object removal by exemplar-based image inpainting[J]. Journal of Visual Communication and Image Representation, 2015, 30: 75-85.

[11] ZHANG D Y, LIANG Z S, YANG G B, et al. A robust forgery detection algorithm for object removal by exemplar-based image inpainting[J]. Multimedia Tools and Applications, 2018, 77(10): 11823-11842.

[12] 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.

[13] 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 and Machine Intelligence, 2017, 39(6): 1137-1149.

[14] 王珠珠. 基于 U 型检测网络的图像篡改检测算法[J]. 通信学报, 2019, 40(4): 171-178.
WANG Z Z. Image forgery detection algorithm based on U-shaped detection network[J]. Journal on Communications, 2019, 40(4): 171-178.

[15] HSU C C, HUNG T Y, LIN C W, et al. Video forgery detection using correlation of noise residue[C]//Proceedings of 2008 IEEE 10th Workshop on Multimedia Signal Processing. Piscataway: IEEE Press, 2008: 170-174.

[16] WANG W H, FARID H. Exposing digital forgeries in interlaced and

deinterlaced video[J]. IEEE Transactions on Information Forensics and Security, 2007, 2(3): 438-449.

[17] BESTAGINI P, MILANI S, TAGLIASACCHI M, et al. Local tampering detection in video sequences[C]//Proceedings of 2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP). Piscataway: IEEE Press, 2013: 488-493.

[18] LIU Y Q, HUANG T Q, LIU Y F. A novel video forgery detection algorithm for blue screen compositing based on 3-stage foreground analysis and tracking[J]. Multimedia Tools and Applications, 2018, 77(6): 7405-7427.

[19] SITARA K, MEHTRE B M. Differentiating synthetic and optical zooming for passive video forgery detection: an anti-forensic perspective[J]. Digital Investigation, 2019, 30: 1-11.

[20] ZHANG J, SU Y T, ZHANG M Y. Exposing digital video forgery by ghost shadow artifact[C]//Proceedings of the First ACM workshop on Multimedia in forensics. New York: ACM Press, 2009: 49-54.

[21] LI L D, WANG X W, ZHANG W, et al. Detecting removed object from video with stationary background[C]//Digital Forensics and Watermaking. Berlin: Springer, 2013: 242-252.

[22] ALORAINI M, SHARIFZADEH M, AGARWAL C, et al. Statistical sequential analysis for object-based video forgery detection[J]. Electronic Imaging, 2019(5): 543-1.

[23] ZHONG J L, PUN C M, GAN Y F. Dense moment feature index and best match algorithms for video copy-move forgery detection[J]. Information Sciences, 2020, 537: 184-202.

[24] CHEN R C, YANG G B, ZHU N B. Detection of object-based manipulation by the statistical features of object contour[J]. Forensic Science International, 2014, 236: 164-169.

[25] PANDEY R C, SINGH S K, SHUKLA K K. Passive copy-move forgery detection in videos[C]//Proceedings of 2014 International Conference on Computer and Communication Technology (ICCCT). Piscataway: IEEE Press, 2014: 301-306.

[26] 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.

[27] 翁韶伟, 彭一航, 危博, 等. 基于 Inception-V3 网络的双阶段数字视频篡改检测算法[J]. 广东工业大学学报, 2019, 36(6): 16-23.
WENG S W, PENG Y H, WEI B, et al. A two-stage algorithm for video forgery detection based on inception-V3 network[J]. Journal of Guangdong University of Technology, 2019, 36(6): 16-23.

[28] 陈临强, 杨全鑫, 袁理锋, 等. 视频对象移除篡改的时空域定位被动取证[J]. 通信学报, 2020, 41(7): 110-120.
CHEN L Q, YANG Q X, YUAN L F, et al. Passive forensic based on spatio-temporal localization of video object removal tampering[J]. Journal on Communications, 2020, 41(7): 110-120.

[29] YANG Q X, YU D J, ZHANG Z X, et al. Spatiotemporal trident networks: detection and localization of object removal tampering in video passive forensics[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(10): 4131-4144.

[30] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.

[31] 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 (CVPR). Piscataway: IEEE Press, 2016: 4207-4215.

[32] ZHANG P B, WANG X, ZHANG W H, et al. Learning spatial-spectral-temporal EEG features with recurrent 3D convolutional neural networks for cross-task mental workload assessment[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2019, 27(1): 31-42.

[33] WANG H H, WU X D, HUANG Z Y, et al. High-frequency component helps explain the generalization of convolutional neural networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 8681-8691.

[34] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 1449-1457.

[35] GAO Y, BEIJBOM O, ZHANG N, et al. Compact bilinear pooling[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 317-326.

[36] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 4489-4497.

[作者简介]

**熊礼治**（1988− ），男，湖北荆州人，博士，南京信息工程大学副教授，主要研究方向为多媒体内容安全与数字取证等。

**曹梦琦**（1999− ），男，山东临沂人，南京信息工程大学硕士生，主要研究方向数字取证等。

**付章杰**（1983− ），男，河南南阳人，博士，南京信息工程大学教授，主要研究方向为数字取证、数据安全等。