doi:10.3969/j.issn.1672-1292.2024.04.003

基于人脸微动作的伪造视频检测

汪小鹏[1],朱　峰[1],李　磊[2],刘司南[3],谭晓阳[1]

(1.南京航空航天大学计算机科学与技术学院,江苏 南京 210016)
(2.中国电子科技集团公司第二十八研究所,江苏 南京 210022)
(3.南京邮电大学计算机学院,江苏 南京 210023)

[**摘要**] 随着深度学习技术的快速发展,人脸视频伪造技术日益精进,其逼真效果对社会安全构成了严重威胁. 尽管基于静态图像的人脸视频真伪检测方法已取得显著进展,并展现出一定的鲁棒性和泛化能力,但现有基于视频流的检测方法通常面临输入维度过高和计算开销过大的问题,这一领域仍缺乏深入研究. 为了解决上述问题,提出了一种基于多变量时间序列分析的人脸视频真伪鉴别方法. 首先,设计了一种基于人脸微动作的建模方法,将视频流转化为多变量时间序列,从而显著降低输入维度. 随后,改进了 Transformer 网络结构,以增强其对时间序列特征的建模能力. 实验结果表明,所提出的方法在准确性与泛化性方面均与当前主流方法相当,展现了良好的应用潜力.

[**关键词**] 人脸伪造检测,Transformer 网络,多元变量时间序列,注意力机制,深度伪造
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1672-1292(2024)04-0028-09

Face Forgery Detection Based on Facial Micro-Movements

Wang Xiaopeng[1],Zhu Feng[1],Li Lei[2],Liu Sinan[3],Tan Xiaoyang[1]

(1.College of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 210016,China)
(2.Nanjing Research Institute of Electronics Engineering,Nanjing 210022,China)
(3.School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China)

Abstract:With the rapid development of deep learning,facial video forgery techniques have become increasingly sophisticated,posing a significant threat to social security. Although image-based facial video authenticity detection methods have achieved remarkable progress and demonstrated certain robustness and generalization capabilities,existing video stream-based methods often suffer from high input dimensionality and substantial computational overhead,which remains inadequately addressed. To tackle these challenges,this paper proposes a facial video authenticity detection method based on multivariate time series analysis. Specifically,a novel modeling approach based on facial micro-movements is designed to convert video streams into multivariate time series,effectively reducing input dimensionality. Furthermore,an enhanced Transformer network is developed to improve its ability to model time series features. Experimental results show that the proposed method achieves performance comparable to state-of-the-art approaches in terms of accuracy and generalization,demonstrating promising application potential.

Key words:face forgery detection,Transformer network,multivariate time series,attention mechanism,deepfakes

随着生成式模型的迅速发展,特别是生成对抗网络(GAN[1]),可以生成大量以假乱真的虚假人脸图片,再佐以变换、插帧技术,进而可以合成人眼难以分辨的造假视频,给社会带来巨大的安全隐患. 人脸视频真伪检测算法大致可分为两类:基于帧内检测[2-7]和基于帧间检测[8-12]. 基于帧内检测的算法通过检测单张图片内暴露的伪造痕迹进行真伪鉴别. 基于帧间的检测方法[13-14]往往使用特征提取网络提取单张图片的特征,再利用时序网络提取帧间的关系. 这类方法虽然提升了检测的准确度,但引入了额外的计算复杂度和运行开销. 此外,还有一些方法考虑不对整张图片进行全局的特征提取,而是转为一些细节信息,如面部表情和头部运动、眨眼模式、几何特征等[12,15-16],这类方法虽然降低了输入维度,但泛化性较差.

收稿日期:2024-05-12.
基金项目:国家自然科学基金项目(61373060).
通讯作者:谭晓阳,博士,教授,研究方向:机器学习. E-mail:x.tan@ nuaa.edu.cn

基于上述问题,本文考虑利用面部全局的运动特征,将人脸视频流建模为多变量时间序列,从而降低维度和兼顾全局信息. 设计基于人脸微动作的建模方法,将人脸微动作描述为人脸局部肌肉的运动,即非头部姿态变化引起的面部区域位置的偏移,如眨眼、张嘴、皱眉等. 假定真实人脸存在着固有的面部运动模式,而伪造人脸在合成过程中会破坏这种固有的运动模式,因此可通过分析面部局部肌肉的运动来检测出这种异常运动模式. 为此,可将人脸划分为若干与肌肉运动有关的兴趣区域,并在各区域内均匀采样兴趣点,通过计算相邻帧的光流来计算兴趣点的偏移量. 为了提高位移计算的准确性,额外补偿优化视频中因人体移动、头部姿态变化等因素对光流计算造成的影响,本文还设计了一个以 Transformer 为骨架的模型,用于处理建模后的多变量时间序列. 经典的 Transformer 模型不能很好地适用本文任务,需引入双维度交叉嵌入模块和变量间掩膜注意力模块. 双维度交叉嵌入模块在时序和空间两个维度上抓取局部的固有模态并进行交叉融合,分别输入到双流网络中. 变量间掩膜注意力模块设置一系列掩膜方式,为变量赋予差异化权值,学习特定变量组之间的注意力关系.

1 相关工作

Rössler 等[2]发现经典的卷积神经网络可以很好地分类真伪人脸. Afchar 等[4]专门设计了一款轻量级的人脸伪造检测模型,在 Celeb-DF 数据集上取得了不错的结果. 相比于这些基于图像 RGB 特征的鉴别方法,Durall 等[5]发现真实图片和伪造图片在高频域的谱图存在区别,可以此识别出伪造痕迹. Qian 等[6]提出了一种利用频域的人脸篡改检测网络,利用离散余弦变换进行频域变换,在低分辨率人脸伪造检测上获得了巨大提升. Wang 等[7]则将 RGB 特征和频域特征相结合,设计了一个双流模型,采用 Transformer 网络提取 RGB 特征,采用 DCT 变换分解出频域特征,然后基于交叉注意力对两种模态的特征进行融合,取得了较好的鉴别效果.

基于图片级别的人脸伪造检测方法忽视了伪造方法在时序上造成的不一致性,研究者们开展了针对视频级别的检测方法研究. Güera 等[10]首先使用 CNN 提取每张视频帧的特征,再使用 LSTM 网络来捕获视频帧特征在时序上的不一致性. Pipin 等[8]测试了多种时空卷积网络在 Celeb-DF 数据集上的效果,测试结果超过了许多基于单帧检测的方法. Ganiyusufoglu 等[9]采用 3D CNNs 提取时空特征,提高了模型在检测新类型造假视频上的准确性. 此外,有学者开始挖掘伪造视频中人物的生理特性在时序上的异常. Haliassos 等[11]提出了 Lipforensics,通过时空网络学习唇语的丰富表征,来检测伪造视频中口部动作的不一致性. Jung 等[12]提出基于眨眼的周期、重复次数和眨眼时间来分类真伪视频. Qi 等[17]提出从心脏跳动节奏角度出发进行 deepfake 检测,认为 deepfake 视频中的心率变化同真实视频不一致,可通过视频中的人脸像素点变化检测出来,该方法在 facefroensics++数据集[2]上取得了不错的结果. 以上这些视频级别的伪造检测算法虽然抓住了时序特征,但输入维度大,模型复杂度高,限制了实际的应用场景.

2 头部姿态无关的人脸微动作提取算法

受 IDT 算法[18]的启发,本文设计了人脸微动作运动轨迹提取算法. 该算法由兴趣点采样模块、姿态校正模块、轨迹计算和筛选模块组成. 算法流程如图 1 所示. 首先,对视频片段的第一帧,基于 landmarks 将

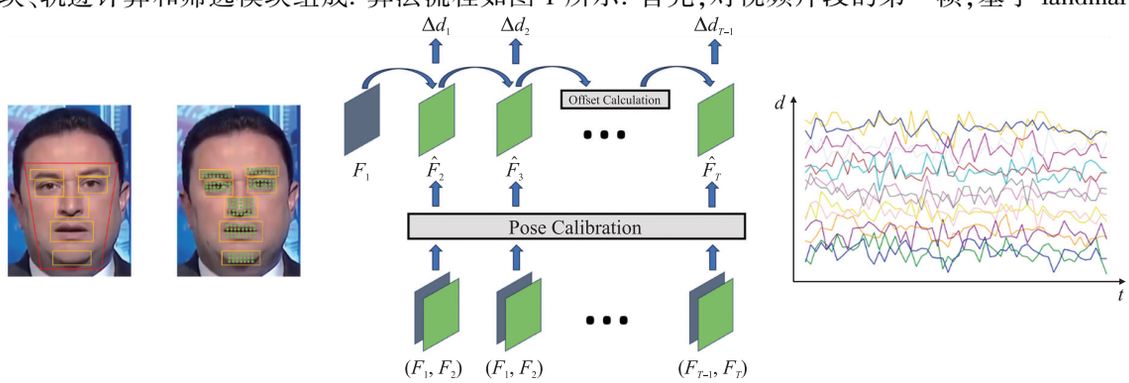


图 1 算法流程示意图

Fig. 1 Illustration of our algorithm

面部划分为 7 个兴趣区域(黄色框内),对黄色框内进行均匀采样兴趣点,并对每一帧划分出微运动无关区域(红色框外). 其次,对每连续两帧之间使用基于 FAST 特征点匹配的姿态校正方法,对矫正后的帧序列通过密集光流进行兴趣点追踪,计算微运动偏移量. 最后,剔除掉运动偏移量不明显的兴趣点,得到最终的多变量时间序列.

2.1　兴趣点采样

由于真实人脸的面部各区域来自同样的数据分布,因而存在着固有的面部微运动模式. 对于伪造人脸,无论是局部伪造,还是全局伪造,由于来自不同的数据分布,不同的面部区域之间会存在相异的运动模式. 可以通过计算人脸肌肉运动偏移量来建模面部微运动模式. 面部行为编码系统[19]在人脸区域上定义了若干个运动单元,认为人脸的所有表情运动都可由这些运动单元来构成. 因此,本文依据运动单元的定义,将人脸划分为 7 个兴趣区域,每个区域都代表一种局部微运动模式. 具体而言,使用脸部检测算法截取视频中的脸部区域,而后将若干个连续帧化为一个视频片段. 使用图像处理库 Dlib 检测人脸特征点,Dlib 使用预训练的模型(如 68 点或 5 点模型)来检测人脸上的特征点,这些特征点通常包括眼睛、鼻子、嘴巴和脸部轮廓等关键位置. 对片段中的第一帧,本文检测出其中的 68 个脸部特征点,依此划分出 7 个面部兴趣区域,并在每个区域内采用均匀采样的方式等间距地采样若干兴趣点,用于后续微动作偏移量的计算. 此外,对于每一帧,还划分出一块与微运动无关的区域用于面部姿态校正.

2.2　姿态校正

之前的面部对齐方法,多是通过 landmarks 位置坐标的对应关系求解仿射变换矩阵,通过仿射变换将所有帧的人脸对应到同一个位置上. 但在基于几何特征的人脸伪造检测中,landmarks 所在的区域往往是几何特征变化剧烈的区域,包含着丰富的运动信息,其位置变化受头部姿态变化和面部肌肉运动影响. 因此,landmarks 计算的仿射变换矩阵不仅不能精确地进行面部对齐,还会给面部微动作位移的计算带来误差.

为了减小对面部微动作位移计算精确性的影响,需进行姿态矫正. 对于输入的一帧图像 F_i 和下一帧 F_{i+1},在 F_i 的微运动无关区域采用 FAST[20]提取角点:

$$S_{\text{corner}}=[v_i^1,v_i^2,\cdots,v_i^n]. \tag{1}$$

而后,使用 Lucas-Kanade 光流算法[21]来追踪这些点在 F_{i+1} 中的位置:

$$S_{\text{predict}}=[v_{i+1}^1,v_{i+1}^2,\cdots,v_{i+1}^n]. \tag{2}$$

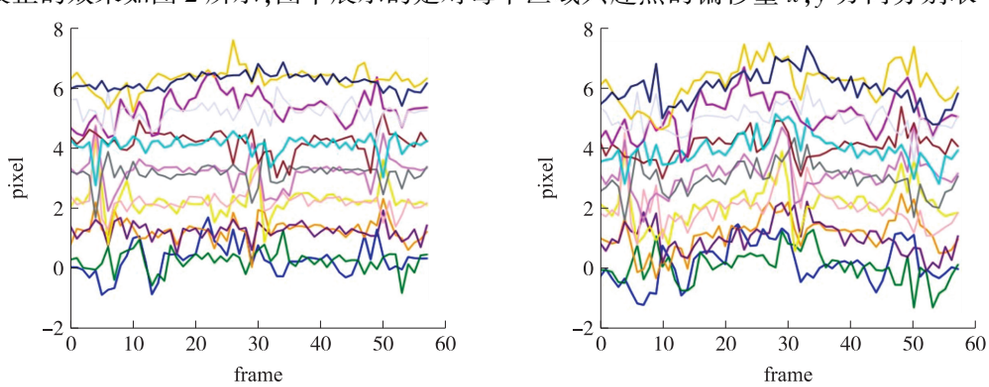
由此可获得一系列匹配点对:

$$S_{\text{match}}=[(v_i^1,v_{i+1}^1),(v_i^2,v_{i+1}^2),\cdots,(v_i^n,v_{i+1}^n)]. \tag{3}$$

由于光流消失造成的追踪点偏移,以及一些遮挡、运动模糊等异常情况,可能存在误匹配,因此通过 RANSAC 算法[22]提高特征点匹配的鲁棒性. 通过处理后的匹配点对计算出仿射变换矩阵 $\boldsymbol{H}$ 和其逆变换矩阵 $\boldsymbol{H}^{-1}$,最终得到面部姿态校正后的下一帧:

$$\hat{F}_{i+1}=F_{i+1}*\boldsymbol{H}^{-1}. \tag{4}$$

姿态校正的效果如图 2 所示,图中展示的是对每个区域兴趣点的偏移量 x,y 方向分别取平均值后所



(a) With posture calibration　　(b) Without posture calibration

图 2　有无姿态校正对比

Fig. 2　Comparison with/without posture calibration

得结果,不同颜色表示不同的脸部区域,左图为经姿态校正后获得的多变量时间序列,右图为未经姿态校正后获得的多变量时间序列. 可以看出,经姿态校正后,消除了面部移动引起的整体位移,同时图像也更加平滑.

2.3 轨迹计算和筛选

为了计算两帧之间的微运动偏移量,本文通过光流法来追踪兴趣点的运动轨迹. 由于 Lucas-Kanad 算法等稀疏光流算法,光流估计过程受图像噪声的影响,会导致光流场中出现不连续、不稳定的噪声点,这种光流场的不连续性会给计算偏移量引入误差. 为此,本文选择逐帧地计算稠密光流,并结合中值滤波,以提高兴趣点追踪的精确性. 首先,对于输入帧 F_i 和校正后的下一帧 F_{i+1},计算其密集光流场 Flow_i;然后,对于第 i 帧的第 k 个兴趣点,计算其在下一帧的坐标:

$$(x_{i+1}^k, y_{i+1}^k) = (x_i^k, y_i^k) + \mathrm{Flow}_i * M|_{x_i^k, y_i^k}, \tag{5}$$

式中,M 是中值滤波器. 将 F_i 中所有兴趣点代入式(5)计算,得到 F_{i+1} 中兴趣点集合的坐标估计值 F_i 的微运动偏移量:

$$\Delta d_i = [x_{i+1}^1 - x_i^1, y_{i+1}^1 - y_i^1, \cdots, x_{i+1}^n - x_i^n, y_{i+1}^n - y_i^n]. \tag{6}$$

通过 F_1 的兴趣点坐标初始值计算其在 F_2 的估计值,其差值即为 F_1 的微运动偏移量. 将 F_2 的估计值更新为初始值,计算 F_3 的估计值,以此类推,即可得到偏移量时序序列 $d = [\Delta d_1, \Delta d_2, \cdots, \Delta d_{T-1}]$. 为了过滤掉一些运动不明显的兴趣点,对每个兴趣区域进行筛选. 对第 j 个兴趣区域内的所有兴趣点,只保留 m_j 个偏移量最大的兴趣点所形成的时间序列. 最终,获得代表面部微运动的多变量时间序列 S_{disp}.

整体算法流程如下:

算法 1　微动作提取算法

Input: frame sequence $[F_1, F_2, \cdots, F_T]$
Output: multivariate time series S_{disp} representing the micro-facial movement
(1) Preprocess frame sequence
(2) for $i=1; i<T; i{+}{+}$ do
(3) if $i==1$ then
(4) Extract regions of interest $[r_1, r_2, \cdots, r_7]$ based on 68 facial landmarks
(5) Sample points of interest S_r
(6) End
(7) Extract corner points S_{corner} in F_i using FAST method
(8) Predict corner points S_{predict} in F_{i+1} using Lucas-Kanade optical flow method
(9) Compute estimated affine transformation matrix H
(10) Obtain calibrated frame $\hat{F}_{i+1} = F_{i+1} * H^{-1}$
(11) Compute dense optical flow Flow_i
(12) for $P_i^k \in S_r$ do
(13) Compute P_{i+1}^k by(5)
(14) end for
(15) Compute displacement Δd_i in F_i
(16) end for
(17) Obtain displacement sequence $d = [\Delta d_1, \Delta d_2, \cdots, \Delta d_{T-1}]$
(18) Obtain multivariate time series S_{disp} by selecting trajectories with distinctive displacements
(19) return S_{disp}

3 人脸异常微动作检测网络

多变量时间序列含有多个变量通道,每个变量通道表示一个兴趣点的偏移量时间序列. 不仅每个区域的兴趣点存在固有的运动模式,不同区域之间也存在相关联的运动模态. 因此,本文设计基于 Transformer 架构的双流网络,分别抓取时间维度和变量维度的信息,检测其中的异常模式. 模型结构如图 3 所示.

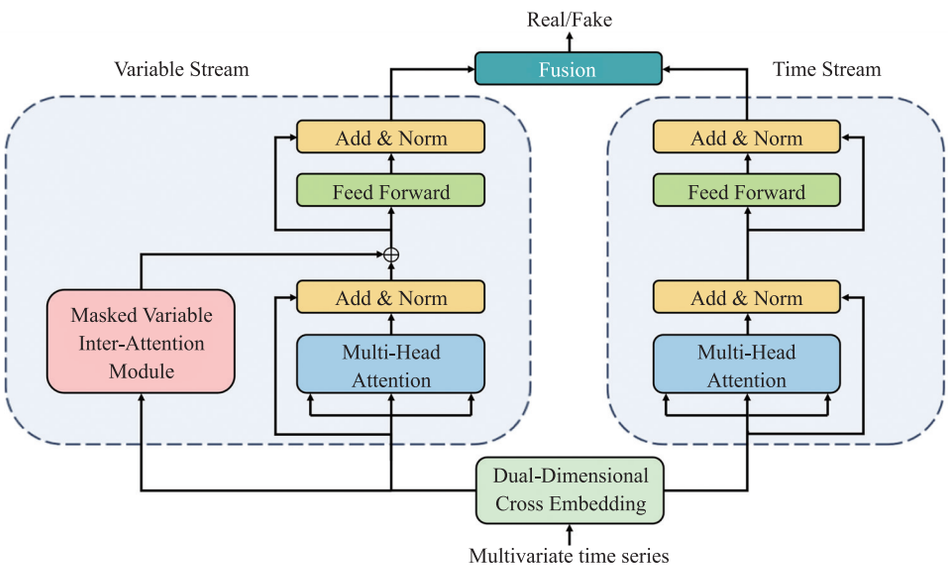


图 3　模型结构图

Fig. 3　The architecture of dual-stream network

3.1　双维度交叉嵌入模块

传统的 Transformer 网络的嵌入层通过一个变换矩阵将输入的单词或符号映射到一个固定长度的嵌入向量,从 NLP 迁移到时间序列分析领域时多采用一个全连接层作为嵌入层. 由于基于微动作的多变量时间序列,无论是变量维度,还是时间维度,其输入尺寸小,不存在长尺度的关联性. 同时,微动作的固有运动模式在时间维度上存在于短窗口的时域内,在变量维度上存在于局部运动单元之间. 因而,本文使用卷积模块,能很好地捕捉这种局部特征.

本文提出的 DCE 如图 4 所示. 对于输入 $X \in \mathbf{R}^{1*N*T}$,首先在时间维度上进行一维卷积得到 $X_t \in \mathbf{R}^{d*N*T}$,并将特征维度尺寸扩展为 d. 此时,X_t 包含时序上的短期相关性和局部模式,扩展的维度提供了更多的表达能力和特征信息. 而后,在变量维度上进行一维卷积得到 $X_v \in \mathbf{R}^{d*N*T}$,保持特征维度不变. 此时,$X_v$ 不仅包含了不同变量维度上的模式,还进行了跨维度的信息交互,通过整合时间维度的特征与空间维度的信息,获得了更全面的特征表示. 最后,分别在时间维度上和变量维度上用卷积操作进行降维,得到整合了变量信息的时间维度嵌入向量 $\boldsymbol{E}_t \in \mathbf{R}^{T*d}$和整合了时间信息的变量维度嵌入向量 $\boldsymbol{E}_v \in \mathbf{R}^{N*d}$. $\boldsymbol{E}_t$ 和 $\boldsymbol{E}_v$ 为后续的双流网络的输入.

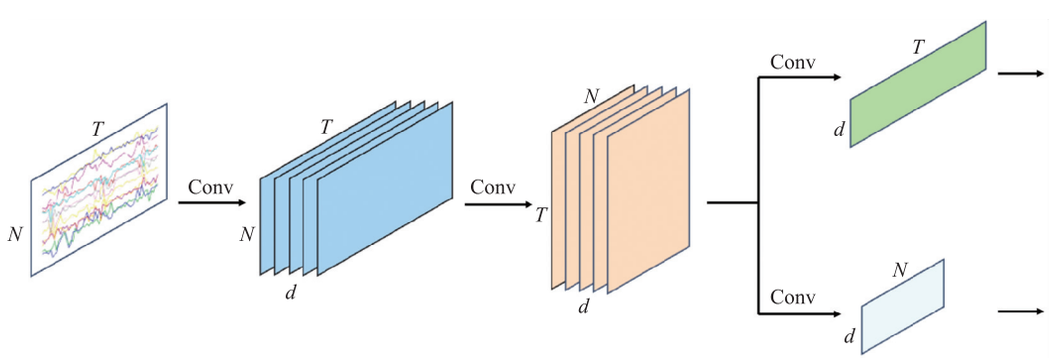


图 4　双维度交叉嵌入模块

Fig. 4　Dual-dimensional cross-embedding module(DCE)

3.2　变量间掩膜注意力模块

与时间维度不同,变量维度上不仅具有局部兴趣点之间的相关性,还存在着多个兴趣区域之间的相关性. 为了有效把握这种多尺度的变量间关系,本文采用掩膜的方式有选择地遮蔽部分兴趣区域的所有兴趣点的通道,通过注意力方式学习剩余兴趣区域的通道之间的关系.

如图 5 所示,对于输入 $\boldsymbol{E}_v \in \mathbf{R}^{N*d}$,预先设置 $M=[M_1,M_2,\cdots,M_k]$,其中,$M_i \in \mathbf{R}^{N*d}$表示第 i 个掩膜,从

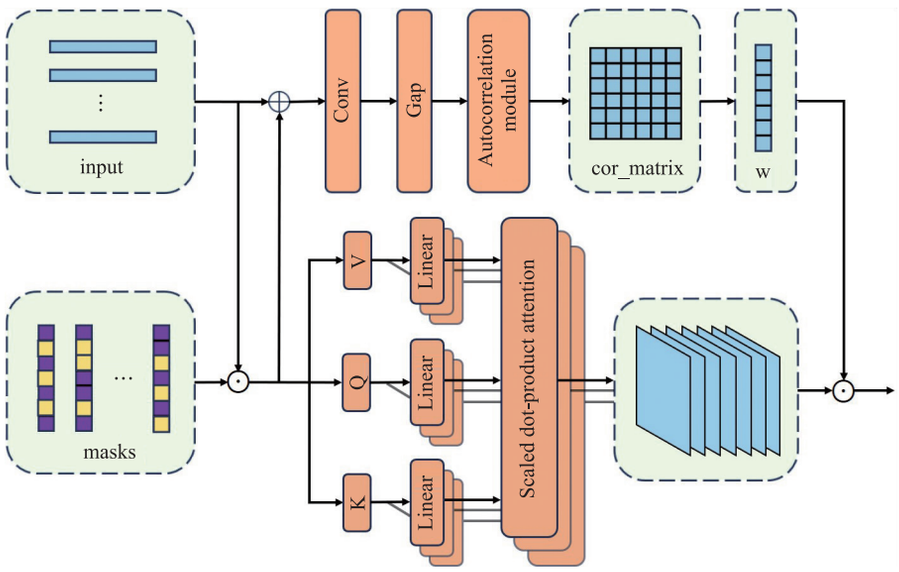


图 5　变量间掩膜注意力模块

Fig. 5　Masked variable inter-attention module(MVIA)

而得到 k 个掩膜后的输出 $\boldsymbol{E}_m=[E_m^1,E_m^2,\cdots,E_m^k]$. $E_m^i\in\mathbf{R}^{N*d}$,可由下式计算:

$$E_m^i=M_i\odot\boldsymbol{E}_v,\ i=1,\cdots,k, \tag{7}$$

式中,⊙表示逐元素点积. 对于每个掩膜后的输出分别计算其自注意力:

$$S_A^i=\text{selfAttention}(E_m^i),\ i=1,\cdots,k. \tag{8}$$

从而得到总的注意力分数 $S_A\in\mathbf{R}^{k*N*v}$. 此外,还需计算每个注意力分数的权重. 对于第 i 个掩膜,先将 E_v^i 和 E_m^i 相加,并使用一个 3 * 1 的卷积核进行一维卷积,再通过一个全局平均池化层得到权重向量 $\boldsymbol{E}_w^i\in\mathbf{R}^h$. 总的权重向量可表示为 $\boldsymbol{E}_w=[\boldsymbol{E}_w^1,\boldsymbol{E}_w^2,\cdots,\boldsymbol{E}_w^k]$. 各权重向量之间的相关性可通过相关性矩阵 $\boldsymbol{C}\in\mathbf{R}^{k*k}$ 计算得到:

$$\boldsymbol{C}=\boldsymbol{E}_w\otimes\boldsymbol{E}_w^{\mathrm{T}}. \tag{9}$$

式中,⊗表示矩阵乘法. 由此可以获得注意力分数权重 $W\in\mathbf{R}^{k*1*1}$,其中:

$$W_i=\sum_j a_{ij}\Big/\sum_i\sum_j a_{ij}. \tag{10}$$

从而可以获得输出特征 $F_A\in\mathbf{R}^{N*v}$:

$$F_A=\sum_i W_i\odot S_A^i. \tag{11}$$

4　实验

4.1　数据集

本文主要使用 FaceForensics++(FF++)[2]进行训练. FF++是一个广泛使用的数据集,包含 1 000 个从 youtube 上获取的真实视频,以及 5 种类型的伪造方式生成的视频:Deepfakes(DF)、Face2Face(F2F)、FaceSwap(FS)、NeuralTextures(NT)、FaceShifter(FSH). 该数据集包含 3 种压缩率的视频:无损、轻微压缩、重度压缩,默认情况下使用轻微压缩级别(c23)的视频. 为了验证模型的泛化性,额外使用了 Celeb-DF[23]进行测试.

4.2　实验环境

本实验使用 Dlib 作为脸部检测器对视频中的人脸区域进行裁剪,使用 opencv 库提取 FAST 角点和计算稠密光流. 对于构建的多变量时间序列,采用每连续 58 帧的结果输入到模型中进行训练,批处理大小为 64,学习率为 0.000 1,优化器采用 Adam,显卡采用 GeForce RTX 2080ti. 人脸伪造识别是一个二分类任务,实验采用准确率和 AUC 两种评价指标进行评价.

4.3　同数据集内比较

首先在 FF++数据集上检测模型的表现性能. 为了广泛地与各种类型的模型比较,选取经典的基于图

片的鉴别模型 Xception、MesoNet,基于视频的鉴别模型 FC、CNN,基于视频的 SOTA 鉴别模型 DR、Comotion、S-MIL、LRNet. 对比结果如表 1 所示,可以看出,与其他模型相比,本文模型在 DF、F2F 伪造方法上都取得了最好的效果;在 DF 和 NT 伪造方法上,本文结果与最优结果很接近;特别是与 LRNet 相比,本文方法在 F2F 和 NT 伪造方法上的效果分别提升了 2.8%和 2.1%.

表 1　同数据集内比较

Table 1　Intra dataset comparison

模型	DF	F2F	FS	NT	模型	DF	F2F	FS	NT	模型	DF	F2F	FS	NT
Xception[2]	0.985	0.972	0.963	0.916	DR[17]	0.987	0.989	0.978	—	S-MIL[26]	0.986	**0.993**	0.993	**0.957**
MesoNet[4]	0.938	0.930	0.929	0.845	FC[24]	0.949	0.960	0.958	0.891	LRNet[16]	0.986	0.965	0.989	0.932
RCNN[14]	0.969	0.944	0.963	—	Comotion[25]	**0.991**	0.933	0.983	0.905	Ours	0.987	**0.993**	**0.998**	0.953

4.4　跨数据集比较

为了验证模型的泛化性,在跨数据集和跨伪造方式两种情况下进行实验,分别采用视频级别的 AUC 和准确率作为指标,对比结果如表 2 所示. 本文模型在两种情况下都获得了最好的结果. 可以看出,相较于基于视频的模型,基于单帧的人脸伪造检测模型在迁移到 Celeb-DF 数据集上时效果下降明显,这是由于其忽视了真伪视频在时序上存在的不一致性,过拟合到了某一类的伪造样式上. 与 LRNet 相比,本文模型在 Celeb-DF 和 FSH 上分别提升了 11.3%和 2.0%. 这是由于本文模型不仅能抓取时序上的一致性,还关注了不同兴趣点间、不同兴趣区域之间的关联性,通过融合两个维度上的信息,避免陷入到某一特定的异常模态,从而提高了模型的泛化性.

表 2　跨数据集比较

Table 2　Cross dataset comparison

模型	Celeb-DF	FSH
Xception[2]	0.548	0.674
RCNN[14]	0.631	0.698
LRNet[16]	0.582	0.763
Ours	**0.695**	**0.783**

4.5　对压缩率的鲁棒性

为了验证模型对压缩率的鲁棒性,本文在 FF++数据集的不同压缩率上与达到 SOTA 效果的 S-MIL 进行了对比,结果如表 3 所示. 可以看出,本文模型在 DF、F2F、FS、NT 4 种伪造方法上的准确率分别下降 7.4%、5.8%、3.4%、8.7%,而 S-MIL 分别下降 1.8%、7.9%、4.7%、8.9%. 可见,本文模型在 F2F、FS 两种伪造方法上的准确率下降均明显优于 S-MIL,这反应了本文模型对压缩率的鲁棒性.

4.6　消融实验

首先,本文对比了轨迹提取算法的有效性,结果如表 4 所示. 先去除姿态校正模块,代之以普通的基于 landmarks 的人脸对齐方法,可以看出 F2F 和 NT 伪造方式的准确率明显下降. 再去除兴趣点轨迹筛选方法,即不进行兴趣区域划分,不按照每个兴趣区域选择规定数量的偏移量最大的兴趣点,而是从所有兴趣区域中均匀采样出一定数量的兴趣点,可以看出,4 种伪造方式准确率均大幅度下降.

表 3　对压缩率的鲁棒性

Table 3　Robustness to compression rate

模型	DF	F2F	FS	NT
S-MIL-c23	0.986	0.993	0.993	0.957
S-MIL-c40	0.968	0.914	0.946	0.868
Ours-c23	0.987	0.993	0.998	0.953
Ours-c40	0.913	0.935	0.964	0.866

表 4　对轨迹提取算法的消融

Table 4　Ablation to trajectory extraction algorithm

模型	DF	F2F	FS	NT
Ours	0.976	0.960	0.995	0.875
w/o Pose Calibration	0.962	0.943	0.985	0.853
w/o Trajectory Filter	0.803	0.631	0.816	0.602

其次,本文对比了模型框架的有效性,结果如表 5 所示. 先去除 MIVA 模块,可以看出 DF 和 FS 伪造方式准确率下降很小. 这可能是由于这两种伪造方式都是基于面部整体区域的伪造,所有不同兴趣区域之间不存在很强的数据分布差异,因而 MVIA 模块没有很好地发挥作用. 而后,只选取双流网络中的时间流,并去除 DCE 模块,代之以只有时间维度上的卷积嵌入,可以看出模型在 NT 伪造方法上性能下降明显. 这验证了本文模型通过两个维度的交叉嵌入有效提高了对 NT 这类伪造样本的学习能力. 最后,使用

表 5　对模型结构的消融

Table 5　Ablation to model architecture

模型	DF	F2F	FS	NT
Ours	0.976	0.960	0.995	0.875
w/o MVIA	0.971	0.943	0.988	0.846
w/o DCE	0.970	0.951	0.990	0.855
Transformer	0.911	0.806	0.915	0.635

经典的 Transformer 网络架构,去除卷积嵌入层,代之以传统的线性层进行嵌入,可以看出模型的能力下降十分明显. 这是因为卷积网络具有很好的抓取局部特征的能力,而伪造方式造成的异常模态往往在局部区域内反应明显,从而使用传统的线性嵌入代替卷积嵌入效果很不理想.

5 结论

人脸伪造技术的迅速发展使得检测其真实性变得尤为重要,对此本文提出了一种针对视频级别的人脸伪造检测方法,利用人脸微动作的位移信息,将人脸视频建模为多变量时间序列,精确提取代表脸部全局运动信息的微动作位移序列,通过分析这些数据,识别出伪造视频中的不自然之处,进而提高检测的准确性. 为了实现这一目标,设计了一种将卷积结构与注意力机制结合的深度学习网络,有效提取和处理视频中的时空特征,增强模型对微小变化的敏感性. 由于仅依靠面部几何特征的提取可能会导致信息的缺失,未来将结合几何信息、纹理信息和频域信息一起进行综合分析,以通过多模态融合的方法显著提高模型的鲁棒性和准确性.

[参考文献](References)

[1] GOODFELLOW I J,POUGET A J,MIRZA M,et al. Generative adversarial nets[C]//Proceedings of the 27th Interrational Conferene on Neural Information Processing Systems. Montreal,Canada:NIPS,2014.

[2] RÖSSLER A,COZZOLINO D,VERDOLIVA L,et al. FaceForensics++:Learning to detect manipulated facial images[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision(ICCV). Seoul,ROK:IEEE,2019.

[3] QUAN W Z,WANG K,YAN D M,et al. Distinguishing between natural and computer-generated images using convolutional neural networks[J]. IEEE Transactions on Information Forensics and Security,2018,13(11):2772-2787.

[4] AFCHAR D,NOZICK V,YAMAGISHI J,et al. MesoNet:A compact facial video forgery detection network[C]//Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security(WIFS). Hong Kong,China:IEEE,2018.

[5] DURALL R,KEUPER M,PFREUNDT F,et al. Unmasking deepfakes with simple features[EB/OL]. (2019-11-02)[2024-05-12]. https://doi.org/10.48550/arXiv.1911.00686.

[6] QIAN Y Y,YIN G J,SHENG L,et al. Thinking in frequency:Face forgery detection by mining frequency-aware clues[C]//Proceedings of the European Conference on Computer Vision-ECCV2000. Glasgov,UK:Springer,2020.

[7] WANG J K,WU Z X,CHEN J J,et al. M2TR:Multi-modal multi-scale transformers for deepfake detection[C]//Proceedings of the 2022 International Conference on Multimedia Retrieval. Newark,USA:ICMR,2022.

[8] PIPIN S J,PURBA R,PASHA M F. Deepfake video detection using spatiotemporal convolutional network and photo response non uniformity[C]//Proceedings of the 2022 IEEE International Conference of Computer Science and Information Technology(ICOSNIKOM). Laguboti,Indonesia:IEEE,2022.

[9] GANIYUSUFOGLU I,NGÔ L M,SAVOV N,et al. Spatio-temporal features for generalized detection of deepfake videos[EB/OL]. (2020-10-22)[2024-05-12]. https://doi.org/10.48550/arXiv.2010.11844.

[10] GÜERA D,DELP E. Deepfake video detection using recurrent neural networks[C]//Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance(AVSS). Auckland,New Zealand:IEEE,2018.

[11] HALIASSOS A,VOUGIOUKAS K,PETRIDIS S,et al. Lips don't lie:A generalisable and robust approach to face forgery detection[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville,USA:IEEE,2021.

[12] JUNG T,KIM S,KIM K. DeepVision:Deepfakes detection using human eye blinking pattern[J]. IEEE Access,2020,8:83144-83154.

[13] WANG R,MA L,XU J F,et al. FakeSpotter:A simple yet robust baseline for spotting AI-synthesized fake faces[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Yokohama,Japan:IJCAI,2020.

[14] SABIR E,CHENG J X,JAISWAL A,et al. Recurrent convolutional strategies for face manipulation detection in videos[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR2019). Long Beach,USA:IEEE,2019.

[15] CHU B L,YOU W K,YANG Z,et al. Protecting world leader using facial speaking pattern against deepfakes[J]. IEEE Signal Processing Letters,2022,29:2078-2082.

[16] SUN Z K,HAN Y J,HUA Z Y,et al. Improving the efficiency and robustness of deepfakes detection through precise geometric features[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville,USA:IEEE,2021.

[17] QI H,GUO Q,XU J F,et al. DeepRhythm:Exposing deepfakes with attentional visual heartbeat rhythms[C]//Proceedings of the 28th ACM International Conference on Multimedia(MM'20). Seattle,USA:ACM,2020.

[18] WANG H,KLÄSER A,SCHMID C,et al. Action recognition by dense trajectories[C]//Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Colorado Springs,USA:IEEE,2011.

[19] EKMAN P,FRIESEN W. Facial Action Coding System:A Technique for the Measurement of Facial Movement[M]. Palo Alto,USA:Consulting Psychologists Press,1978.

[20] ROSTEN E,DRUMMOND T. Machine learning for high-speed corner detection[C]//Proceedings of the 9th European Conference on Computer Vision. Graz,Austria:Springer,2006.

[21] LUCAS B,KANADE T. An iterative image registration technique with an application to stereo vision[C]//Proceedings of the 7th International Joint Conference on Artificial Intelligence. Vancouver,Canada:IJCAI,1981.

[22] FISCHLER M A,BOLLES R C. Random sample consensus:A paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM,1981,24(6):381-395.

[23] LI Y Z,YANG X,SUN P,et al. Celeb-DF:A large-scale challenging dataset for deepfake forensics[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle,USA:IEEE,2020.

[24] CIFTCI U A,DEMIR I,YIN L J. FakeCatcher:Detection of synthetic portrait videos using biological signals[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2020. https://doi.org/10.1109/TPAMI.2020.3009287.

[25] WANG G X,ZHOU J H,WU Y. Exposing deep-faked videos by anomalous co-motion pattern detection[EB/OL]. (2020-08-11)[2024-05-12]. https://doi.org/10.48550/arXiv.2008.04848.

[26] LI X D,LANG Y N,CHEN Y F,et al. Sharp multiple instance learning for deepfake video detection[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle,USA:ACM,2020.

[责任编辑:严海琳]