第 58 卷 第 24 期/2021 年 12 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于卷积长短期记忆网络的换脸视频检测

郑博文，夏华威*，陈睿东**，韩乾坤***

天津大学电气自动化与信息工程学院，天津 300072

摘要 随着近年假脸合成技术(DeepFake)的发展，当前社交平台充斥着通过换脸技术生成的海量假视频，虽然假视频可以丰富大众的娱乐生活，但是同样存在着曝光隐私等负面问题。如何精准检测出由 DeepFake 生成的伪造数据已成为网络安全防御领域中一项重要且具有挑战性的任务。针对这一问题，很多科研工作者提出了针对换脸视频的检测方法，但是现有的检测方法均忽略了 DeepFake 视频帧与帧之间的关联特性。因此，对于部分针对脸部信息进行平滑处理的篡改方法，已有的检测方法的检测准确率有明显的下降。基于此，提出了一种基于长短期记忆(LSTM)网络的 DeepFake 视频检测算法。该算法能够捕获 DeepFake 视频帧中的脸部微表情变化，并利用编码器生成局部视觉信息的特征，同时利用注意力机制实现局部信息的权重分配；最后再次借助 LSTM 网络实现时序空间下视频帧的关联信息融合，从而实现对 DeepFake 视频数据的有效检测。采用 FaceForensics＋＋数据库对所提算法进行了评估，与现有方法相比，实验结果证明了所提算法的优越性。

关键词 机器视觉；假脸合成技术检测；帧间特性；卷积长短期记忆网络；注意力机制

中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP202158.2415002

Exposing DeepFake Video Detection Based on Convolutional Long Short-Term Memory Network

Zheng Bowen, Xia Huawei*, Chen Ruidong**, Han Qiankun***

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract With the advancement of DeepFake technology in recent years, the current social platform is full of massive fake videos produced by face-changing technology. Although fake videos can enrich people's entertainment, they also have disadvantages, such as exposing their personal information. How to accurately detect the fake data generated by DeepFake technology has become an important and difficult task in network security defense. Many researchers have proposed face-changing video detection methods in response to this problem, but the existing detection methods often ignore the incoherence of facial feature crossing video frames. Thus, they are easily countered by optimizing facial synthesizing techniques, resulting in accuracy degradation. Based on this, we propose a novel DeepFake detection method based on long short-term memory (LSTM) network that captures the micro expression changes in terms of the facial features caused by the composite video and uses an encoder to generate features of local visual information. Simultaneously, the attention mechanism is used to achieve the weight distribution of local information. Finally, the LSTM network is used to realize the association information fusion of video frames in temporal space, resulting in the effective detection of DeepFake video data. This paper evaluates a proposed algorithm on the FaceForensics＋＋ dataset, and when compared to existing methods, the experimental results show that the proposed algorithm is superior.

Key words machine vision; deep fake detection; interframe character; convolutional long short-term memory network; attention mechanism

OCIS codes 150.1135; 100.2000; 100.2960; 100.4994

收稿日期：2021-01-05；**修回日期**：2021-02-06；**录用日期**：2021-03-02

基金项目：国家自然科学基金(61772359,61572356,61872267,61862020,61861014)

通信作者：*xiahuawei@tju.edu.cn；**20517610@qq.com；***15822563807@163.com

1 引 言

随着智能手机等移动摄影设备的普及，文本、图像和视频等数据的获取变得愈加方便。相较于短信等传统信息载体，照片和视频作为载体包含了更为丰富的信息[1]。同时，娱乐市场的庞大需求催生出了众多面部合成技术生成的大量鬼畜、恶搞视频，许多视频在身份欺诈、舆论控制等方面存在被滥用的风险[2]，这些威胁使得开发针对性的面部检测算法变得迫在眉睫。

在面部合成方法中，一种基于深度学习的面部伪造技术（通常称为 DeepFake）变得非常流行。DeepFake 使用两个自动编码器并行训练，其输出往往伴随着非常鲜明的可辨别特征，例如操纵区域中的分辨率变化、人体头部姿势运动不一致等。起初这些特征能够很轻松地通过肉眼分辨出来，但是随着生成对抗网络（GAN）[3]和相关的基于 GAN 的优化方法（例如 CycleGAN[4]）的介入，DeepFake 具有了产生高质量面部视频和图像的能力，这无疑给 DeepFake 的检测任务增加了挑战难度。

在现有的 DeepFake 检测方法中，MesoNet[1]专注于图像的介观特性，但对于 DeepFake 的某些关键特征（例如面部翘曲伪影）则没有进行检测。文献[5]介绍了一种基于卷积神经网络（CNN）的检测方法，该方法着眼于面部翘曲伪影的检测，可以通过简单的图像处理操作直接模拟这些伪影，从而节省了训练数据的准备时间和显卡存储空间。上述这些检测方法主要关注每一帧的图像特征，而忽略了视频中帧间时序特征。尽管文献[6]提出的基于循环神经网络的检测方法考虑了帧间时序特征，但仍然忽略了 DeepFake 视频帧与帧之间的面部表情变换不连贯问题。文献[1]与文献[7]等研究发现，与原始视频相比，DeepFake 视频帧与帧之间的面部表情突然变化或僵硬变化非常明显。

为了解决 DeepFake 视频检测存在的问题，本文提出了一种基于注意力机制的 DeepFake 视频检测算法。该算法利用卷积长短期记忆（LSTM）网络对帧间相关特征进行研究[8]，同时引入注意力机制对变化较大的帧分配更大的权重值；增强了视频帧中的重点区域，使网络更加专注于输入的面部特征；最后，通过一系列实验验证了帧间相关信息在 DeepFake 视频检测任务中的重要性，证明了所提算法在 DeepFake 视频检测方面的优越性。

2 所提算法介绍

在生成 DeeFake 视频的过程中，不同帧间会产生面部表情变化不连贯的现象，本文即基于此特性来实现换脸视频的检测。为了突出每个视频帧中的面部信息，从而更好地提取面部特征，首先对 DeepFake 视频帧进行预处理。即将这些输入帧对齐并进行等间隔采样，同时，通过应用诸如去除眼睛特征、模糊面部区域等方法来随机增强输入数据的多样化。对视频帧进行预处理的目的是为了更好地突出视频帧中的面部信息，便于模型提取每个视频帧的面部特征。所提算法通过 CNN 对预处理后的视频帧进行特征提取，并将提取的特征输入基于卷积 LSTM 的编码器中，结合注意力机制分配的特征权重，生成带有权重的特征数据。最后，将该数据再次输入进卷积 LSTM 网络中，实现多视图信息在时序空间下的特征融合，使网络更好地学习随着视频帧数的改变其表征面部表情特征的变化，进而提升所提算法检测的准确率。

2.1 数据预处理

2.1.1 面部检测与对齐

为了从每个视频中提取详细的面部信息，使用多任务级联的卷积神经网络（MTCNN）[9]从视频帧中获取目标面部的标记。然而，MTCNN 模型标记的面部区域无法覆盖到原始帧的整个脸部区域，覆盖的范围仅包括从眉毛到嘴的区域。为了解决此问题，将获取到的区域向外扩张 30%，这样便可覆盖整个面部；然后对包含整个面部区域的视频帧标准化为 224×224 像素，并对其进行面部对齐；最后从所有视频帧中随机抽取 10 帧作为网络的输入数据。

2.1.2 输入数据处理

对于主流的 DeepFake 视频检测网络，当输入面部的多样性增加时，检测准确率会有显著的提升。一些诸如眨眼、鼻子移动、嘴唇周围的细节特征都是影响目前已有的 DeepFake 视频检测方法检测准确率的重要因素。受文献[5]中所采用方法的启发，将所定位的面部区域的高斯模糊作为输入变量。为了使网络能学习更多的面部特征，在 DeepFake 视频中分别移除了包括眼睛、鼻子和嘴巴在内的面部元素，从而便于对其他区域进行数据增强。

2.1.3 基于 CNN 的帧特征提取

与其他传统的 CNN 模型（如 ResNet[10]、VGG-Net[11]等）相比，EfficientNet-B5[12]在网络的性能和内存开销之间取得了相对较好的平衡，因此所提算

法采用 EfficientNet-B5 模型从输入帧中提取视觉特征向量。与其他传统网络(如 AlexNet[13])相比，EfficientNet 具有在网络的深度、宽度和分辨率之间保持平衡的优点，可以显著提高准确性并缩短训练过程。EfficientNet-B5 模型是基于神经网络架构搜索设计得到的 EfficientNet-B0 的缩放版本，视频帧输入到该模型后，得到输出为 2048 维的特征向量集 $\boldsymbol{V}=\{\boldsymbol{v}_1,\cdots,\boldsymbol{v}_n\}$。

2.2　帧间特征聚合

提出一种基于卷积 LSTM 网络的换脸视频算法，用于从输入帧中提取时序和面部特征，网络结构如图 1 所示。

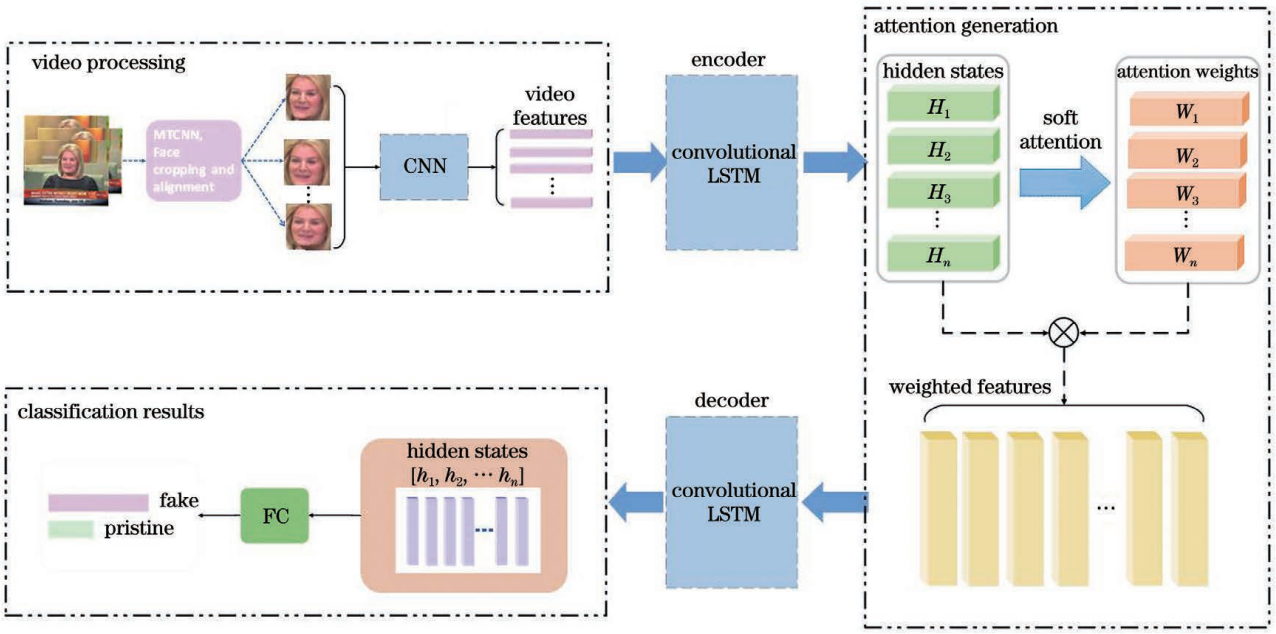

图 1　卷积 LSTM 网络的框架

Fig. 1　Framework of convolutional LSTM network

2.2.1　基于编码器-解码器的特征提取网络

在深度学习领域，LSTM 网络具有可以有效提取时序信息的特征[14-17]，使得该网络成为检测 DeepFake 视频不同帧之间面部特征不一致性的有效方案。采用的 LSTM 结构与文献[15]中应用的经典结构不同，是一种基于编码-解码的结构[16]，由两个卷积 LSTM 子网组成。与传统 LSTM 模型相比，所提卷积 LSTM 可以有效地利用输入序列的时序和空间信息，这在处理视频流中非常有效。通过文献[7]可知，DeepFake 视频是先将视频打散成帧序列再逐帧进行修改的，故而其面部区域的帧间连续性很难达到原始视频的程度。基于此原理，通过检测帧间的差异来对视频是否换脸进行判断。为了更好地对视频进行检测，使用注意力机制来集中处理帧间的差异性。首先采用编码器中卷积 LSTM 结构的输出隐藏层来生成注意力机制中的权重向量；然后将权重向量与隐藏层相乘，进行加权与归一化操作；生成加权向量之后，再将这些权重因子馈入基于卷积 LSTM 的解码器中来生成最终的输出。

2.2.2　基于注意力机制的帧间特征聚合

将注意力机制[18]与卷积 LSTM 模型结合使用，从而实现对输入特征 $\boldsymbol{V}$ 的加权操作。这样的设计是为了在编解码机制中赋予关键向量比较大的权重，而减小对非必要特征的关注。实现方法是将注意力机制计算的权重与隐藏层 $\boldsymbol{H}$ 相乘，然后将其结果输入到解码网络。

基于注意力机制的视频权重计算流程如下。

受到文献[19]启发，应用卷积长短期记忆(ConvLSTM)网络来处理输入视频特征的时序和空间信息。从循环神经网络(RNN)模型演变而来的 ConvLSTM 保留了 RNN 模型中的隐藏层 $\boldsymbol{h}_t$，并增加了神经元 $\boldsymbol{c}_t$ 来避免信息丢失。$\boldsymbol{h}_t$ 和 $\boldsymbol{c}_t$ 的关系为

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh\boldsymbol{c}_t,\tag{1}$$

式中：⊙代表 Hadamard 乘积。输出门 $\boldsymbol{o}_t$ 的计算为

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{vo}*\boldsymbol{h}_{t-1}+\boldsymbol{W}_{ho}*\boldsymbol{v}_{i,t}+\boldsymbol{b}_o),\tag{2}$$

式中：σ 表示 sigmoid 函数；$\boldsymbol{v}_{i,t}$ 表示第 i 个视频中的第 t 个选定帧所对应的特征向量；$\boldsymbol{W}_v$ 和 $\boldsymbol{W}_h$ 表示卷积核；$\boldsymbol{b}$ 表示偏置项；$*$ 表示卷积运算符。当前的记忆状态 $\boldsymbol{c}_t$ 的计算表达式为

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tilde{\boldsymbol{c}}_t,\tag{3}$$

式中：$\boldsymbol{c}_{t-1}$ 表示之前的记忆状态；$\tilde{\boldsymbol{c}}_t$ 表示更新后的记忆状态。$\tilde{\boldsymbol{c}}_t$ 与输入门 $\boldsymbol{i}_t$ 和遗忘门 $\boldsymbol{f}_t$ 的关系为

$$\tilde{\boldsymbol{c}}_t=\tanh(\boldsymbol{W}_c*\boldsymbol{v}_{i,t}+\boldsymbol{W}_c*\boldsymbol{h}_{t-1}+\boldsymbol{b}_c),\tag{4}$$

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{vi}*\boldsymbol{h}_{t-1}+\boldsymbol{W}_{hi}*\boldsymbol{v}_{i,t}+\boldsymbol{b}_i),\tag{5}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{vf}*\boldsymbol{h}_{t-1}+\boldsymbol{W}_{hf}*\boldsymbol{v}_{i,t}+\boldsymbol{b}_f)。\tag{6}$$

注意力机制能够使神经网络专注于处理一系列输入的特定部分，进而显著降低计算成本并提高网络性能，因此采用注意力机制来最大程度地减小任务的复杂性。具体的处理结构可以表示为

$$e_i=\boldsymbol{W}_c\tanh(\boldsymbol{W}_{va}*\boldsymbol{v}_{i,t}+\boldsymbol{W}_{ha}*\boldsymbol{h}_{t-1}+\boldsymbol{b}_a),\tag{7}$$

$$\alpha_i=\exp\{e_i\}/\sum_{j=1}^{n}\exp\{e_j\},\ \sum_{i=1}^{n}\alpha_i=1,\tag{8}$$

$$\tilde{\boldsymbol{v}}_t=\alpha_i\cdot\boldsymbol{v}_t,\tag{9}$$

式中：α_i 代表注意力权重；$\boldsymbol{W}_c$，$\boldsymbol{W}_{va}$，$\boldsymbol{W}_{ha}$ 和 $\boldsymbol{b}_a$ 均是随网络一起变化的参数；$\tilde{\boldsymbol{v}}_t$ 表示更新后的输入。$\tilde{\boldsymbol{v}}_t$ 之后被送入基于卷积 LSTM 的解码器网络中，得到一串输出 $\boldsymbol{H}'=[\boldsymbol{h}_1',\boldsymbol{h}_2',\cdots,\boldsymbol{h}_n']$，该向量用于表示整个输入帧序列。

2.3　特征聚合与最终分类

如文献[20]所述，LSTM 在处理长序列数据时具有出色的性能表现。即使这些输入大部分都是嘈杂的并且超过 1000 帧，卷积 LSTM 模型也仍可以提取其特征，同时保留短时滞留信息。因此，可以应用 ConvLSTM 模型来提取输入视频的完整特征。本文将 ConvLSTM 的最后一个单元状态视为所有所选输入视频帧的表示。但是，这些输入的视频帧存在一些不重要的冗余信息。注意力机制能够根据输入帧的信息重新分配权重，使网络模型专注于提取具有代表性的输入帧的特征。为了全面反映编码解码模型的输出，采用一个全连接层来对基于 ConvLSTM 的解码模型生成的最终加权特征进行融合。这也是对整个视频进行分类的最终描述。

3　分析与讨论

首先介绍整体实验设置，包括数据集和实施细节，最后采用相应的实验结果证明所提算法的优越性。

3.1　数据集

实验部分选取了最新发布的评测数据集 FaceForensics++(FF++)[21]来对所提算法进行评估。该数据集囊括了 1000 条原始视频，及 DeepFake (DF)、FaceSwap (FS)、Face2Face (F2F) 和 NeuralTextures(NT)4 种最新面部操作技术分别进行篡改得到的视频各 1000 条，共计 4000 条篡改视频。每条视频均有 LQ，HQ，RAW(低，高，原画)三种画质，且其中的视频经过了一定的筛序，较少会出现面部被遮挡等检测不到人脸的情况，是目前为止最适合进行 DeepFake 检测的数据集。

还对面部数据库进行扩充，以更好地评估所提算法的性能，扩充后的数据由高斯模糊视频及从原始视频中分离出鼻子、眼睛和嘴巴的视频组成。在实施细节上，将评测数据集 FaceForensics++按照 8∶1∶1划分为训练数据集、测试数据集以及验证数据集。为了提高网络效率，实验过程中从每个视频中随机抽取 10 帧作为输入，将所有不同分辨率的视频帧均转换为相同 224×224 像素大小的视频帧。

3.2　实施细节

对于 CNN，选择 EfficientNet-B5 作为网络的基干。表 1 展示了 FaceForensics++数据集上不同网络基干的实验结果。与传统的基于 CNN 的基干网络 VGG16 相比，EfficientNet 在相同网络情况下的分类精度在任务指标 LQ 上提升了 5.66 个百分点，在 HQ 上提高了 5.54 个百分点，在 RAW 上提升了 4.90 个百分点。

表 1　FaceForensics++数据集上不同基干网络的比较

Table 1　Comparison of different backbone networks on the FaceForensics++ dataset

Backbone network	Classification accuracy /%		
	LQ	HQ	RAW
AlexNet[22]	96.89	88.64	91.97
VGG16[11]	90.85	92.35	94.67
ResNet[10]	93.42	95.68	96.43
EfficientNet(Ours)	96.51	97.89	99.57

实验所用的计算机配备了两个 NVIDIA 1080Ti GPU，32 GB RAM 和一个 Intel Xeon E5-2609 V4 @ 1.70 GHz×8 CPU。采用 PyTorch 平台进行所有实验。

3.3　卷积 LSTM 有效性的实验

本工作的贡献之一是在网络中应用了基于注意力的卷积 LSTM 模型，该模型可以有效提取输入帧的面部变化特征，进而提高检测器的性能。对所提算法与一些经典的 DeepFake 检测方法进行比较，以评估所提网络的性能，实验结果如表 2 所示。

表 2 FaceForensics++数据集上不同算法的比较

Table 2 Comparison of different algorithms on the FaceForensics++ dataset

Algorithm	Classification accuracy /%		
	LQ	HQ	RAW
CNN[21]	90.00	91.45	93.40
SVM[25]	70.10	73.64	75.43
RNN[6]	93.46	95.04	95.98
GRU[24]	94.48	96.18	97.54
LSTM[14]	94.29	96.24	96.79
ConvLSTM[19]	95.18	96.79	98.80
ConvLSTM(with attention)	96.51	97.89	99.57

通过这些实验可以发现，与其他算法相比，所提算法可以实现更显著的性能提升。一些具体的观察如下。

1) MesoNet[1]专注于较少层数的图像的介观特性，忽略了视频帧之间的相关性。换句话说，它在训练步骤中不考虑面部信息交叉帧。这种方法对于图像检测有较好的检测结果，但是并不适用于DeepFake视频检测，因为在DeepFake视频检测中，帧与帧之间的面部信息变化是需要考虑的重要因素。

2) RNN[6]考虑了DeepFake视频中的时序特征，因此其性能优于MesoNet，但与最新的方法相比还存在着一定的差距。这是因为RNN虽然强调了输入序列的时序特征，但是也有容易遗忘先前特征的不足，这会导致视频特征的重叠。上述问题带来了冗余信息，降低了分类器的准确性。

3) 基于LSTM的模型[14]的性能要优于MesoNet。LSTM的体系结构考虑了时序信息。该方法引入预训练的CNN[23]来从输入中提取面部特征。换句话说，经过预训练的CNN模型可以为输入的DeepFake视频提取鲁棒的特征向量。因此，它能使模型考虑特征向量的跨帧面部特征信息。

4) GRU[24]可以看作是LSTM模型的简化版本，可在保持其性能的同时有效降低参数数量。传统的LSTM模型具有很多参数，这就会带来更大的训练成本。通过将输入门和遗忘门打包到更新门中，GRU可以显著提高训练速度并降低计算成本。

5) 所提算法应用了卷积LSTM[19]来获得输入序列数据的时序和空间信息。实验结果表明，卷积LSTM非常适合处理跨输入帧的帧间和帧内特征的任务。

3.4 软注意力模型有效性的实验

将进行进一步的实验以分析软注意力加权机制的有效性。如先前所述，注意力机制的应用目的在于消除冗余信息。本节通过一系列实验测试了不同的基于注意力机制的特征聚合模型的有效性，结果如表3所示。

表 3 FaceForensics++数据集上不同注意力模型的实验结果

Table 3 Experimental results of different attention models on the FaceForensics++ dataset

Model	Classification accuracy /%		
	LQ	HQ	RAW
ConvLSTM	95.18	96.79	98.80
ConvLSTM+hard-attention	95.22	96.91	99.24
Self-attention	95.96	97.89	98.91
ConvLSTM+soft-attention	96.51	97.34	99.57

硬注意力机制[18]能够一次专注于输入的一个精确位置，使网络能够专注于相对重要的信息。但是，硬注意力机制无法照顾到输入的所有位置并且是不可分的，与其他基于注意力的方法相比，可能会导致性能下降。自注意力机制[25-26]解决了RNN模型中的并行计算问题，可以看作是基于编码-解码结构的网络，编码器和解码器均基于多头结构。这样的设计消除了基于RNN的模型的重复和卷积问题，从而大大节省了计算成本，但是存在忽略输入的时间特征的问题。所提算法将软注意机制应用于结合了CNN和ConvLSTM的模型中，可以将其视为所提算法中的视频特征处理部分。由于视频帧的视觉特征具有很多冗余信息，因此先前提到的注意力机制无法彻底消除它。通过使用软注意力机制，可以观察到显著的性能改进。这样的机制可以指导网络在保持网络输入完整性的同时，将更多的精力放在重要功能上。

3.5 面部数据变化实验

对于DeepFake视频检测网络，输入视频的种类越多，输入包含的信息就越多，检测准确率也会有相应的提升。因此，输入数据的变化也是重要的参数。本文将高斯模糊作为变化方法。采用MTCNN[8]检测并提取原始视频中的面部，然后使用大小为5×5的卷积核来进行高斯模糊。这样设计的目的是在变形的面部上创建更多分辨率的变化案例，从而可以更好地模拟DeepFake视频中不一

致的分辨率变化。为了模仿 DeepFake 生成过程中所产生的伪像，修改后的面部将被固定并变换为输入面部的原始大小。本实验中将改编后的视频帧数目分别设置为 2，3，4，5，6，并与原有的视频帧相结合构成 10 帧的输入数据。实验结果如表 4 和图 2(a)所示，结果表明，当修改后的输入数少于 5 帧时，分类精度随着修改后的帧数增加而增加。上述结果也证实了当输入特征增加时，网络可以更好地找出由 DeepFake 视频合成过程引起的伪像。但是，当修改的输入帧的数量大于 5 帧时，分类性能会显著降低，因为这种方法存在削弱 DeepFake 视频生成中其他重要特征的问题，例如眨眼、嘴唇周围的细节等。

表 4　FaceForensics＋＋数据集上高斯模糊帧数变化的分类准确度

Table 4　Classification accuracy of the variation of Gaussian blur frame numbers on the FaceForensics＋＋ dataset

Number of manipulated frames	Classification accuracy /%		
	LQ	HQ	RAW
2	71.32	78.41	82.17
3	85.66	87.63	89.68
4	88.97	90.05	92.53
5	92.65	94.32	97.43
6	75.74	79.57	82.64

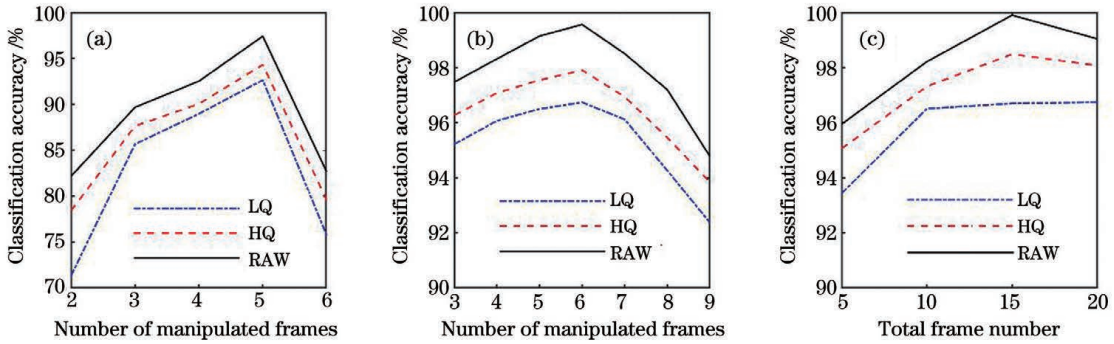

图 2　FaceForensics＋＋数据集上帧数对分类准确度的影响。(a)高斯模糊帧数；(b)面部数据移除帧数；(c)总帧数

Fig. 2　Influence of number of frames on the classification accuracy on the FaceForensics＋＋ dataset. (a) Gaussian blur frame number; (b) face data removal frame number; (c) total frame number

为证明面部器官对 DeepFake 视频检测的影响，分别从输入帧中取 3，4，5，6，7，8，9 个帧来进行移除眼睛、鼻子和嘴巴区域的操作，其余帧则保持不变。实验结果如表 5 和图 2(b)所示，结果表明：当面部数据移除的帧数少于6时，检测准确度随移除数据的帧数增加而增加，这是因为帧与帧之间的差异性会随着帧数的增加而增大；但随着帧数的继续增大，是否原始视频本身包含眼睛等器官的结论会愈加干预实验结果，所以分类准确度会逐渐降低。

表 5　在 FaceForensics＋＋数据集上面部数据移除的实验结果

Table 5　Experimental results of the facial data removal on the FaceForensics＋＋ dataset

Number of manipulated frames	Classification accuracy /%		
	LQ	HQ	RAW
3	95.22	96.27	97.48
4	96.07	97.08	98.32
5	96.50	97.55	99.15
6	96.74	97.91	99.57
7	96.12	96.94	98.51
8	94.26	95.46	97.20
9	92.37	93.84	94.78

此外，为了证明输入帧数对分类结果的影响，将原始输入帧数设置为 5，10，15，20，并进行了进一步的实验，结果如表 6 和图 2(c)所示，可以看出：当帧数小于15时，随着帧数的增加，分类性能得到了增

表 6　FaceForensics＋＋数据集上总帧数变化的分类准确度

Table 6　Classification accuracy of the variation of the total frame numbers on the FaceForensics＋＋ dataset

Total frame number	Classification accuracy /%		
	LQ	HQ	RAW
5	93.43	95.06	95.95
10	96.51	97.32	98.22
15	96.70	98.49	99.91
20	96.74	98.08	99.05

强；而帧数大于 15 时，由于卷积 LSTM 网络的记忆特性受限于计算机内存等硬件设备的限制，分类准确度会有所波动。本实验证明了网络性能与网络帧数之间的相关性在计算机相应性能下呈正比例关系。

3.6 与其他最新算法的比较实验

为了验证所提网络的先进性，对所提算法与基于 FaceForensics++数据集的其他最新算法进行了比较。这些算法包括基于 CNN 的算法[1,5,21]、基于 RNN 的算法[6]、基于光流场差异的算法[27]及基于 3DCNN 的算法[28]。实验结果如表 7 所示。结果表明：所提检测算法取得了最佳的分类准确度；与之相比，基于传统计算机视觉的方法的分类精度最差，在 Face2Face 篡改的视频数据集下，分类准确度仅为 81.6%；基于 CNN 的方法应用了深度学习来提取 DeepFake 视频的操纵特征，较传统计算机视觉的光流法，其性能得到了 2%～10%的提高；类似地，当将 RNN 应用于视频帧之间的时间相干性时，性能也会有所提高。与这些方法相比，所提算法能够同时将权重引入输入视频中并对帧间面部特征的不一致性加以发掘，有效提升了分类准确性和处理效率。

表 7 FaceForensics++数据集上不同算法的分类结果比较

Table 7 Comparison of classification experimental results of different algorithms on the FaceForensics++ dataset

Reference	Method	Classifier	Classification accuracy /%	Dataset
Ref. [1]	Mesoscopic features	CNN	83.2	F2F
Ref. [21]	Steganalysis features	CNN	91.0	F2F
			94.0	DF
			93.0	FS
			81.0	NT
Ref. [6]	Temporal features	RNN	94.3	F2F
Ref. [27]	Temporal features	Optical Flow	81.6	F2F
Ref. [28]	Deep learning features	3DCNN	95.1	DF
			92.3	FS
This work	Interframe features	ConvLSTM	96.5	F2F
			96.7	DF
			94.9	FS
			92.7	NT

4 结 论

如今，面部替换视频的威胁已得到广泛认知。提出了一种基于注意力机制的 LSTM 网络，可以精准地检测 DeepFake 视频并显著地节省了显存算力空间。具体地，本模型利用软注意力机制对包含不同信息量的视频帧分配不同的权重，以在训练过程中更好地学习视频帧中的显著性视觉特征。与当前的方法相比，所提算法不仅利用了来自输入视频的视觉特征，而且还考虑了视频帧与帧之间的关联特征。在公共数据集上的实验结果证明了所提算法的优越性和合理性，说明了帧与帧之间的关联特征对 DeepFake 视频检测的准确性有重要的影响。此外，围绕增强型面部数据集的实验结果表明，眼睛和嘴巴等局部信息在 DeepFake 视频检测中起着至关重要的作用。

参 考 文 献

[1] Afchar D, Nozick V, Yamagishi J, et al. MesoNet: a compact facial video forgery detection network[C]// 2018 IEEE International Workshop on Information Forensics and Security (WIFS), December 11-13, 2018, Hong Kong, China. New York: IEEE Press, 2018.

[2] Wang J X, Lei Z C. A convolutional neural network based on feature fusion for face recognition [J]. Laser & Optoelectronics Progress, 2020, 57(10): 101508.
王嘉欣, 雷志春. 一种基于特征融合的卷积神经网络人脸识别算法[J]. 激光与光电子学进展, 2020, 57

(10): 101508.

[3] Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks[C]// Proceedings of the 36th International Conference on Machine Learning, June 9-15, 2019, Long Beach, California, USA. Cambridge: PMLR, 2019: 7354-7363.

[4] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2242-2251.

[5] Li Y Z, Lyu S W. Exposing DeepFake videos by detecting face warping artifacts[EB/OL]. (2019-5-22)[2020-12-24]. https://arxiv.org/abs/1811.00656.

[6] Sabir E, Cheng J, Jaiswal A, et al. Recurrent convolutional strategies for face manipulation detection in videos[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops, June 16-20, 2019, Long Beach, California, USA. New Work: IEEE Press, 2019: 80-87.

[7] Zhang Y X, Li G, Cao Y, et al. A method for detecting human-face-tampered videos based on interframe difference[J]. Journal of Cyber Security, 2020, 5(2): 49-72.
张怡暄, 李根, 曹纭, 等. 基于帧间差异的人脸篡改视频检测方法[J]. 信息安全学报, 2020, 5(2): 49-72.

[8] Zhu M K, Lu X L. Human action recognition algorithm based on Bi-LSTM-Attention model[J]. Laser & Optoelectronics Progress, 2019, 56(15): 151503.
朱铭康, 卢先领. 基于 Bi-LSTM-Attention 模型的人体行为识别算法[J]. 激光与光电子学进展, 2019, 56(15): 151503.

[9] Zhang K P, Zhang Z P, Li Z F, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[10] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//3rd International Conference on Learning Representations, May 7-9, 2015, San Diego, California, USA. [S.l.: s.n.], 2015.

[12] Tan M X, Le Q V. Efficientnet: rethinking model scaling for convolutional neural networks[C]// Proceedings of the 36th International Conference on Machine Learning, June 9-15, 2019, Long Beach, California, USA. [S.l.: s.n.], 2019: 6105-6114.

[13] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]// Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13, 2014, Montreal, Quebec, Canada. New York: Curran Associates, 2014: 3104-3112.

[14] Yang Y, Zhou J, Ai J B, et al. Video captioning by adversarial LSTM[J]. IEEE Transactions on Image Processing, 2018, 27(11): 5600-5611.

[15] Cornia M, Baraldi L, Serra G, et al. Predicting human eye fixations via an LSTM-based saliency attentive model[J]. IEEE Transactions on Image Processing, 2018, 27(10): 5142-5154.

[16] Xu L H, Li Z, Jiang J J, et al. High-precision and lightweight facial landmark detection algorithm[J]. Laser & Optoelectronics Progress, 2020, 57(24): 241026.
徐礼淮, 李哲, 蒋佳佳, 等. 高精度轻量级的人脸关键点检测算法[J]. 激光与光电子学进展, 2020, 57(24): 241026.

[17] Ma Y K, Peng H Y, Cambria E, et al. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, February 2-7, 2018, New Orleans, Louisiana, USA. California: AAAI Press, 2018: 5876-5883.

[18] Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention[C]//Proceedings of the 32nd International Conference on Machine Learning, July 6-11, 2015, Lille, France. Cambridge: MIT Press, 2015: 2048-2057.

[19] Shi X J, Chen Z R, Wang H, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting[C]// Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. Cambridge: MIT Press, 2015: 802-810.

[20] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[21] Rössler A, Cozzolino D, Verdoliva L, et al. FaceForensics＋＋: learning to detect manipulated facial images [C] // 2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 1-11.

[22] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] // Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, December 3-6, 2012, Nevada, United States. Cambridge: MIT Press, 2012: 1106-1114.

[23] Amerini I, Li C T, Caldelli R. Social network identification through image classification with CNN [J]. IEEE Access, 2019, 7: 35264-35273.

[24] Chung J Y, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. (2014-12-11)[2021-01-01]. https://arxiv.org/abs/1412.3555.

[25] McCloskey S, Albright M. Detecting gan-generated imagery using color cues [EB/OL]. (2018-12-19) [2021-01-01]. https://arxiv.org/abs/1812.08247.

[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, California, USA. Cambridge: MIT Press, 2017: 5998-6008.

[27] Amerini I, Galteri L, Caldelli R, et al. Deepfake video detection through optical flow based CNN[C]// 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), October 27-28, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 1205-1207.

[28] Wang Y H, Bilinski P, Bremond F, et al. G3AN: disentangling appearance and motion for video generation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 5263-5272.