黄夏馨. 基于时空信息的深度伪造人脸检测[J]. 智能计算机与应用,2024,14(10):99-106. DOI:10.20169/j.issn.2095-2163.241013

基于时空信息的深度伪造人脸检测

黄夏馨
(苏州大学 电子信息学院, 江苏 苏州 215000)

摘 要: 近年来,深度学习的快速进步使得经过多媒体篡改的人脸视频达到了以假乱真的程度。这种用深度学习的架构来生成虚假人脸的方法被称为"Deepfake"。现有的 Deepfake 检测方法应用于高分辨率的原视频时性能尚可,然而应用于经过高度压缩的低质量视频时表现欠佳。针对大部分现有算法对视频的帧间信息利用不够充分这一问题,本文提出了一种基于时空信息的检测算法。首先,该方法设计了一种时空注意力架构,同时提取了空间和时间上的注意力以抑制无关信息;然后,对提取出的时空注意力权重信息通过改进深度可分离卷积(Xception)和卷积门控循环网络(ConvGRU)进行深层次的提取,ConvGRU 用于获取在改进 Xception 网络降维前丢失的帧间信息;最后,使用判别器进行二分类,通过实验训练模型,在 FF++数据集低质量视频上获得了 97%的准确率,取得了良好的效果。

关键词: 虚假人脸检测; 双流注意力; 特征处理
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)10-0099-08

Deepfake detection based on spatio-temporal information

HUANG Xiaxin
(College of Electronic Information, Soochow University, Suzhou 215000, Jiangsu, China)

Abstract: In recent years, the rapid progress of deep learning has made it difficult for people to distinguish between real videos and multimedia tampered face videos. The method of generating fake faces using a deep learning architecture is called "Deepfake". The existing Deepfake detection methods perform well when applied to high-resolution original videos. However, their detection performance is unsatisfactory for low quality videos that have been highly compressed. To address the issue of insufficient utilization of inter frame information in most existing algorithms, this paper proposes a detection algorithm based on dual stream attention. Firstly, this method designs a dual-branch attention architecture that extracts both spatial and temporal attention to suppress irrelevant information. Then, the extracted spatiotemporal attention weight information is extracted at a deeper level through improved deep separable convolution (Xception) and Convolutional Gated Recurrent Network (ConvGRU). ConvGRU aims to obtain inter frame information lost before the improved Xception network dimensionality reduction. Finally, using a discriminator for binary classification, the model is trained through experiments and achieves an accuracy of 97% on low quality FF++datasets, achieving good results.

Key words: deepfake detection; dual-branch attention; feature processing

1 研究背景与现状

1.1 研究背景

近年来,随着深度学习在图像视频处理领域的大力发展,虚假视频生成模型的研发已经到了较为成熟的阶段。深度视频伪造技术是深度伪造最主要的代表,Deepfake[1] 最早出现在 2017 年 12 月,在 Reddit 论坛帖子中一名自称为"Deepfakes"的用户分享了一些由人工智能技术制作的伪造视频,通过使用深度学习算法将一个人的脸部特征替换为另一个的脸部特征,从而制造出虚假的视频。虽然深度造假技术可以给大众提供一定程度的便利,如视频特效制作[2]、游戏开发、虚拟购物等等,但是深度伪造技术对网络安全、社会安全、及个人安全造成了严重威胁[3-5]。针对这一现状,虚假人脸伪造检测技术的研究就显得尤为重要。

深度伪造可以分为完整的人脸合成、身份交换、人脸表情和属性操作等,通常是用 GAN 来生成的,

作者简介: 黄夏馨(1998-),女,硕士研究生,主要研究方向:虚假人脸检测。Email:jlhu@suda.edu.cn

收稿日期: 2023-06-10

最新的 StyleGAN 生成了具有高度真实感的高质量面部图像，比较常用的技术有 Deepfake、FaceSwap[6]、Face2Face[7]等等。

深度造假检测技术可以分为深度图像造假检测和深度视频造假检测，核心技术都是通过提取输入的特征，对特征中的不连续、不一致性带来的伪影进行判断。深度伪造检测的通用流程分为以下 4 步：

(1)对造假视频进行预处理，包括从视频中解码出图像帧、数据增强、裁剪图像、人脸识别出 ROI 区域等等。

(2)通过神经网络提取视频或图像中的特征。

(3)对(2)中提取的特征信息做进一步处理和分析。

(4)通过判别器对处理后信息进行判断，并输出二分类结果。

1.2 研究现状

对于检测虚假人脸伪造视频，早期的方法使用 Xception[8]、EfficientNet[9]等浅层网络来提取低级的篡改指纹。随着深度学习技术的迅猛发展，更多的深度神经网络框架已然应用到图像分类领域，例如之前在 NLP 领域备受瞩目的 Transformer[10]技术。这类网络具有强大的特征提取能力，能够有助于提取出更精确的微表情。先进的伪造技术可能会生成极其真实的面部图像，但并不能消除时间上的不一致性。文献[11]使用的算法判断人脸融合的边界、文献[12]使用了成对自一致性学习、文献[13]提出了自混合图像，这些方法都只检测帧内的不一致性，使用取帧的平均准确率的方式来检测视频，无法捕捉到时间不一致特征，导致了视频检测的高错误率。文献[14]提出的 Xception-LSTM 体系结构可以利用 Xception 和 LSTM 的强大特征提取能力解决这些问题，但也会丢失 Xception 提取特征前的时间相关性。

为了解决上述研究中对时间信息利用不充分的问题，本文尝试用 ConvGRU 代替 LSTM 来进行计算，可以提取 Xception 降维前的时间相关性。此外，本文同时考虑了空间和时间上的注意力。本文的主要工作总结如下：

(1)提出了一种双分支(空间分支和时间分支)注意力机制来提取时空上的特征。首先对输入图像通过卷积计算空间注意力，决定模型应该重点关注的内容，将得到的权重图与输入图像相加，获得该帧图像的全局特征，然后将所有帧的全局特征合并输入到时间注意力模块中，本文设计的时间注意力模块参照了自注意力的提取流程，通过使用查询、键、值来获取帧间注意力，再与空间注意力相结合，获取到包含全局特征的时空信息，更精确地关注人脸从而抑制周围的背景信息。

(2)建立了一种改进 Xception 网络与 ConvGRU 网络相结合的模型来处理时空不一致特征。该模型以双流注意力网络的输出特征作为输入，在 Xception 网络的基础上添加了 SE 模块来逐帧提取重点区域的空间特征，然后合并所有帧的特征，再输入 ConvGRU 做进一步处理，捕获在空间特征提取降维前的时间不一致性。通过对 FF++[15]数据集低质量视频上的仿真实验，获得了 97%的准确率，相比仅使用 Xception 网络提升了 11%的准确率，验证了本文算法的有效性。

2 网络设计

本节首先介绍了本文网络的整体框架设计，然后又依次详细阐述预处理、注意力模块、特征处理模块和分类器的设计，最后给出了使用的损失函数。

2.1 整体框架设计

首先需要对数据集中的视频进行预处理，然后输入到时空注意力网络中提取注意力信息，学习适当的权重来捕获人脸信息、同时抑制背景信息，再将得到的注意力权重输入至改进 Xception 网络和 ConvGRU 网络中，最后通过分类器得到真假二分类结果。整体网络框架设计如图 1 所示。

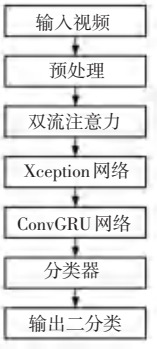

图 1 整体网络框架

Fig. 1 Overall network architecture

2.2 预处理

预处理包括对视频帧进行提取和数据增强。本文基于 FF++数据集进行实验，该数据集包括 1 000 个真样本和 4 000 个假样本。在帧提取阶段使用 FFmpeg 对每个视频提取连续的 30 帧，保存在为每个视频生成的文件夹中，并打上标签，进行打乱。数据增强主要包括图像随机旋转、随机翻转、随机平移、随机亮度调整、随机缩放、图像去中心化、图像标准化等方法，可以在数据集有限的情况下提高样本

质量，具体参数见表 1。

表 1　数据增强处理方式及参数

Table 1　Data enhancement processing methods and parameters

数据增强方法	参数设置
随机旋转	−45°～+45°
随机亮度增强	0.5～1.0
随机通道颜色偏移	−50～+50
随机缩放	0.8～1.2
填充模式	最邻近填充

2.3　注意力机制

注意力机制已经在图像处理的各种不同任务中得到了广泛应用[16-19]，在深度伪造检测中也已取得一定进展[20-21]。注意力机制包括通道注意力和空间注意力。然而这些注意力仅仅能关注到帧内的不一致信息，对于检测帧间不一致还需要利用时间信息，本节提出了一种空间注意力与时间注意力相结合的网络模型，来学习应该重点关注什么内容、这些内容在时序上有什么特点。下面将分别从空间注意力和时间注意力两方面来展开研究分析。

2.3.1　空间注意力

卷积神经网络每一层的输出均可以表示为 $C \times H \times W$ 的特征图，这里 C 表示图像的通道，H 和 W 分别表示高度和宽度。空间注意力需要对所有的通道进行学习，获得一个权重矩阵，矩阵中每个元素就是该点的空间位置信息的重要程度。

空间注意力模块的流程如图 2 所示。由图 2 可知，对于第 i 帧输入图像 F_i，首先对其使用 1×1 的卷积层得到空间注意力图 AM，接着将输入变换为一维张量，使用矩阵乘法计算每个像素与其他像素之间的相似性，然后将其与输入 F_i 叠加，增强有用的特征，同时弱化无关特征。再应用 1×1 的卷积层来获得本帧的全局特征 G_i，G_i 除了应用于空间注意力机制中的帧的通道级权值计算，同时也用于后续的时间注意力机制。最后，将全局特征 G_i 和本帧 F_i 进行合并，即可获得加强的第 i 帧的空间注意力特征 A_i。

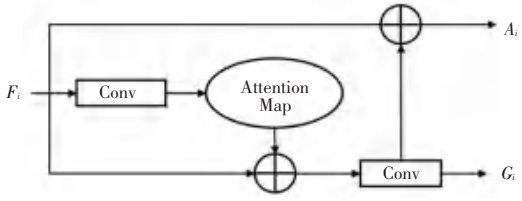

图 2　空间注意力流程

Fig. 2　Spatial attention process

2.3.2　时间注意力

对于时序信息应该关注什么，本文参考自注意力机制进行设计。自注意力是目前应用最广泛的注意力机制之一，是基于特征图本身关注的内容而提取的注意力。对于一般的卷积神经网络而言，卷积核的设置限制了感受野的大小，导致网络往往需要多层的堆叠才能关注到整个特征图，而自注意力的优势在于内含的感受野是全局的，能通过简单的查询与赋值就可获取到特征图的全局空间信息。通过计算时间上的自注意力，可以分析图像序列中内部信息的关系，求解得到视频中每一帧对当前帧的贡献程度。

时空注意力模块的流程如图 3 所示，对于输入形状为 (N, T, H, W, C) 的张量，这里 N 表示样本批次大小，T 表示时间步长，H 和 W 表示特征图的高度和宽度，C 表示通道数，通过 1 × 1 的卷积层进行线性变换生成查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 矩阵，然后将 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 的形状分别调整为 (T,C)，(C,T) 和 (T,C) 以适应后续计算，再将 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 进行点积运算得到相似度，并进行缩放，经过 Softmax 函数进行归一化，通过 1 × 1 的卷积进行整形，最终得到一个权重矩阵，其中的每个值都是一个大于 0、小于 1 的权重系数。注意力的计算公式如下：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (1)$$

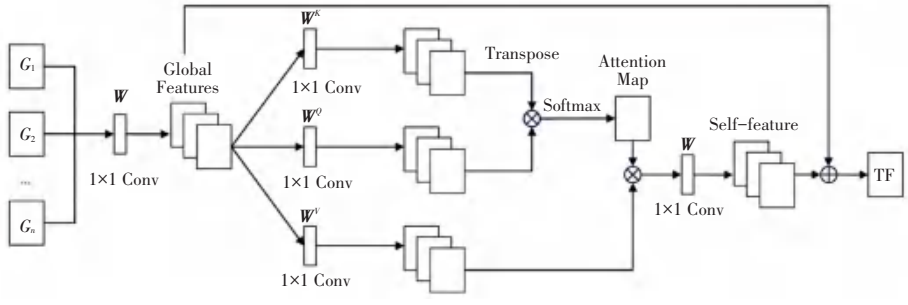

图 3　时空注意力流程

Fig. 3　Flow chart of spatio-temporal attention

其中，d_k 表示键向量的维度，也就是通道数 C。

在这个过程中，注意力不直接使用 F_i，而是使用经过矩阵乘法生成的 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 三个矩阵，转换到新的空间中去做点积，相当于使用了 3 个可训练的参数矩阵，可以增强模型的拟合能力。通过计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的点积，可以度量不同特征之间的相关性，较大的点积值表示 2 个特征之间具有较强的相关性，而较小的点积值则表示较弱的相关性，这有助于捕捉输入视频帧间特征的相互作用和关联关系。

在使用自注意力机制得到时间特征 TF 后，取 TF 的行向量记为 TF_i，其含义是对于所有输入帧的注意力特征向量 $\boldsymbol{G}$，在全局特征的基础上，考虑向量之间的交互作用后得到的一维向量，将这个同时包含全局空间信息和时间信息的量和每帧的空间注意力 A_i 相乘，最后加上 F_i 进行注意力加强，就获得了最终的双流时空注意力 DA_i，表达式如下所示：

$$DA_i = A_i * TF_i + F_i \tag{2}$$

其中，“ * ”表示向量点积。

2.4 特征处理

在 2.3 节中已获得同时带有时空特征的双流时空注意力权重，通常情况下这种注意力已经能够用于视频粗分类，但对于高压缩精度、低质量的视频而言还远远不够，本文将其用作后续计算的特征，设计了一种改进 Xception 网络与 ConvGRU 网络结合的特征处理模块。对 Xception 网络而言，通过加入 SE 模块，能够提升模型对通道特征的敏感性，使用全局平均池化来将每个通道信息压缩成一个数值，再使用 2 层全连接层进行信息激励，显式地建模通道间的关系，SE 模块弥补了 Xception 对通道信息的不足。CNN+RNN 模式是视频分类常用的一个模式，因其可以同时提取到空间和时间特征，此外其中的 Xception 和 LSTM 又因其卓越的视频处理效果而得到广泛使用。但是如果只将 2 个网络顺序连接，那么 LSTM 网络会丢失 Xception 网络在降维前的时间特征，要对此特征进行充分利用，就需要将 LSTM 中的全连接权重改为卷积，则成为 ConvLSTM 网络[17]。考虑到 ConvGRU 与 ConvLSTM 有着相似的结构，但 ConvGRU 中门的数量会比 ConvLSTM 中的更少，这意味着 ConvGRU 的权重矩阵和偏置向量的数量也更少，但是两者间的效果非常接近，因此本文中设计了 Xception 与 ConvGRU 网络结合的架构。下面将分别对这 2 种网络进行介绍。

2.4.1 改进 Xception 网络

Xception 是 Inception 结构的一种变体，已成功应用于深度伪造视频检测中。与传统的卷积神经网络不同，Xception 采用了深度可分离卷积结构，这种卷积可以将传统卷积中的通道混合和空间滤波划分为 2 个独立操作来加以执行，从而更加有效地学习特征。将每一个卷积核对应一个输入通道进行卷积，然后将所有经过卷积之后的结果堆叠起来得到后续的特征图，这样只处理长宽方向的信息，最后使用 1×1 的卷积处理跨通道的空间信息。Xception 网络架构是由这样具有残差连接的深度可分离卷积层的线性堆叠组合而成。

Xception 主要分为三大层，分别是：输入层（Entry Layer）、中间层（Middle Layer）和输出层（Exit Layer），由 14 个模块组成，所有的模块都加入了残差连接机制，这就使得 Xception 的收敛速度显著提升，准确率也有所提高。Xception 的精简结构如图 4 所示。

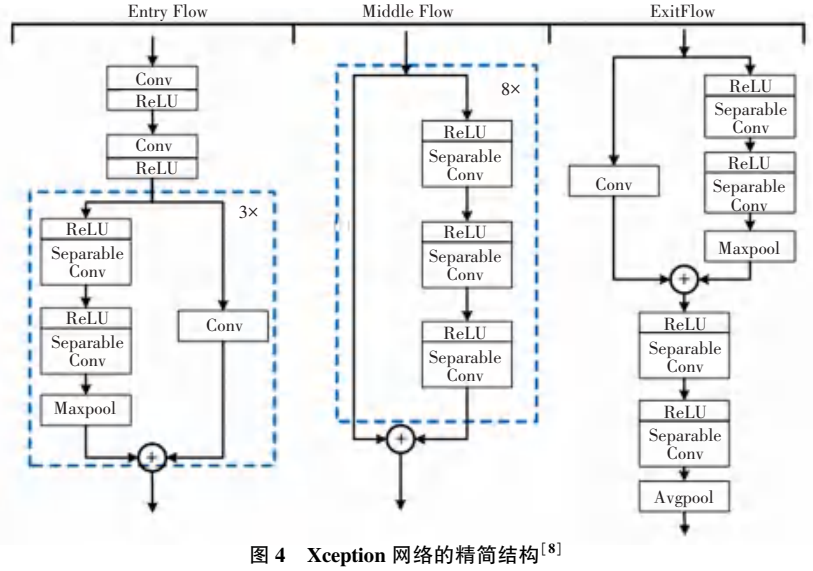

图 4 Xception 网络的精简结构[8]

Fig. 4 Simplified structure of Xception network[8]

在 Xception 的中间层插入 SE(Squeeze and Excitation)[18]模块来进行改进。研究可知,Squeeze 操作可以获得特征图的全局信息,而 Excitation 通过多层感知机产生的通道权重向量可以增强重要信息的通道,同时抑制不相干信息的通道,从而提高网络的表达能力。为了防止参数量过大,仅在 Xception 的中间层的每个深度可分离卷积中嵌入 SE 模块。图 5 为添加 SE 块后的深度可分离卷积网络的结构。图 5 中,首先通过可分离卷积模块,再通过全局平均池化进行压缩,然后通过 2 个全连接层和激活层进行激励,再用原特征按每一个通道乘以上一步操作获得的重要度系数进行重新标定,可以增强有效特征的影响力,削弱不相干特征对结果的影响,最后加上残差块以加快网络的收敛过程。

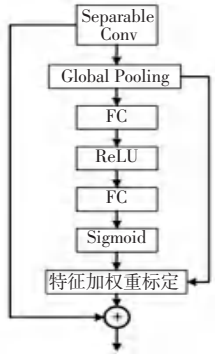

图 5　添加 SE 块后的深度可分离卷积结构

Fig. 5　Depthwise separable convolutional structure after adding SE blocks

2.4.2　卷积门控循环网络

ConvGRU[22-29]是一种基于门控循环单元(GRU)和卷积神经网络(CNN)的深度学习模型。就是在 GRU 的基础上增加了卷积操作,从而可以在处理时空序列数据时更好地利用卷积神经网络的优势。ConvGRU 主要包括 2 个门:重置门(Reset Gate)和更新门(Update Gate),用于控制信息的流动和保存。重置门越接近 0,代表越多地忘记前一时刻的隐藏状态,故将当前时刻的状态作为初始状态;越接近 1,则代表越多地保留前一时刻的隐藏状态。

ConvGRU 的网络结构如图 6 所示,图 6 中黄色框内表示重置门,蓝色框内表示更新门,X_t 表示 t 时刻循环单元的输入,Y_t 表示 t 时刻循环单元的输出,H_t 和 H_{t-1} 分别表示 t 时刻循环单元的隐藏状态和前一个时刻循环单元的隐藏状态。更新门和重置门的计算公式分别表示为:

$$Z_t = \sigma(W_{conv1,1} * [h_{t-1}, x_t]) \tag{3}$$

$$R_t = \sigma(W_{conv1,2} * [H_{t-1}, X_t]) \tag{4}$$

其中,$W_{conv1,1}$ 和 $W_{conv1,2}$ 分别表示每个门的卷积核;σ 表示 Sigmoid 函数;"$*$"表示卷积运算;[·]表示连接运算。候选隐藏状态和复位门的计算公式具体如下:

$$\hat{H}_t = \tanh(W_{conv2} * [R_t \odot H_{t-1}, X_t]) \tag{5}$$

$$H_t = (1 - Z_t) \odot H_{t-1} + Z \odot \hat{H}_t \tag{6}$$

其中,W_{conv2} 表示用于计算候选隐藏态的卷积核,tanh 表示双曲切线激活函数。

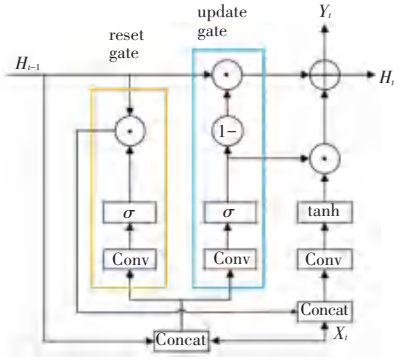

图 6　ConvGRU 的结构

Fig. 6　Structure of ConvGRU

2.5　分类器

在获得了上述时空特征后,需要将特征输入到分类器中进行分类。本节设计了如下分类器。首先通过一个全局池化层对卷积后的特征进行下采样,将池化后的特征展平并通过一个维度为 1 024 的全连接层,为了防止过拟合,插入一个设定值为 0.5 的随机丢弃(dropout)层来减少网络的复杂性和过拟合的风险,再通过一个维度为 512 的全连接层和一个设定值为 0.5 的随机丢弃层,最后通过 Softmax 层输出二分类的结果。

2.6　损失函数

本文是一个视频二分类问题,因此选取常用的交叉熵损失作为损失函数来进行反向传播。得到的公式如下所示:

$$L = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \tag{7}$$

其中,N 表示类别数;y_i 表示样本 i 的标签,真实视频为 1,虚假视频为 0;p_i 表示样本 i 预测为正类的概率。

3　实验结果与分析

本节使用前文提出的网络架构在 FF++数据集上进行了训练与测试,将仿真结果与其他方法进行了比较,证明本文所提架构在准确率方面的提升,并

通过消融实验证明了各模块的有效性。

3.1 数据集与评价指标

3.1.1 数据集

本文主要使用 FaceForensics++数据集(FF++)对模型来进行训练和测试,此数据集包括由 1 000 个从 Youtube 上下载的包含各种不同肤色年龄、不同场景下的真实人脸视频,所有视频都包含一个无遮挡的正面人脸,使用 4 种不同的伪造方法(DeepFakes、Face2Face、FaceSwap 和 NeuralTexures)分别生成虚假人脸视频。由于 FF++数据集提供了 3 种压缩质量的视频,在原视频与 c23 压缩质量下的视频下已取得 99%以上的分类准确率,因此本文主要研究分类准确率不高的 c40 高压缩度低质量下的视频,分辨率大小为 640×480。

3.1.2 评价指标

(1)分类准确率(Accuracy, *Acc*):表示对网络模型的分类能力的评价。数学定义公式如下:

$$Acc = (TP + TN)/(TP + TN + FP + FN) \times 100\% \quad (8)$$

(2)真正率(True Positive Rate, *TPR*):表示能将正例正确分类的概率。

(3)假正率(False Positive Rate, *FPR*):*FPR* 表示将负例错分为正例的概率。

(4)受试者曲线(Receiver Operating Characteristic curve, *ROC*):在二分类问题中,最终得到的数据是对每一个样本估计其为正的概率值,根据每个样本为正的概率大小从大到小排序,再按照概率从高到低,依次将概率值作为阈值。当测试样本的概率大于或等于阈值时,认为该样本为正样本,否则为负样本。每次选取一个不同的阈值,就可以得到一组 *FPR* 和 *TPR*,即 *ROC* 曲线上的一点。

(5)精确率(Area Under Curve ROC, *AUC*):随机给定一个正样本和一个负样本,使用分类器进行分类,正样本的得分比负样本的得分要大的概率。AUC 越大,即越接近 1,表示模型的分类效果越好。可由下式计算求得:

$$AUC = \int_0^1 TPR(FPR^{-1}(x))\,\mathrm{d}x \quad (9)$$

本研究为二分类问题,主要使用 *ACC* 与 *AUC* 作为评价指标。

3.2 训练设置与训练结果

实验时将真假视频随机打乱,以 8∶1∶1 的比例构建了训练集、验证集和测试集。网络模型输入图像大小为 224×224×3,分别在 5 帧、10 帧、20 帧、30 帧的图像序列上设置了不同的批次大小进行训练,通过对参数的微调,优化器为 Adam,学习率设置为 0.000 02,总共 200 轮训练。

在实验时发现每个视频的采样帧数对实验结果产生了一些影响。表 2 展示了本研究分别在 5 帧、10 帧、20 帧、30 帧等不同采样帧数时的结果。分析可知,采样结果为 5 帧和 10 帧的参数量较小,因此将 *batchsize* 设置为 5,而对于 20 帧和 30 帧的采样帧数,过大的参数量限制 *batchsize* 为 1。

表 2 采样帧数与批次大小对 *Acc* 和 *AUC* 的影响

Table 2 Effect of sampling frame number and the batch size on *Acc* and *AUC*

批次大小	采样帧数	*Acc*	*AUC*
5	5	0.629 4	0.661 8
5	10	0.894 3	0.887 4
1	20	0.940 7	0.942 9
1	30	0.973 5	0.975 2

由表 2 可知,在采样帧数为 5 时,仅能获得比随机猜测略好的结果;但帧数提升到 10 时准确率显著上升,但还只能达到 90%左右;在采样帧数为 30 帧时可以达到 97.5%的效果。然而局限于显卡硬件原因,大于 30 帧后是否会呈现出更好的结果不得而知。

3.3 消融实验

本节的消融实验主要分为特征处理模块的消融实验与整体网络的消融实验。

3.3.1 特征处理模块的消融实验

为了证明各个模块的有效性,分别对比了 6 种不同情况下的 *Acc* 的表现,分别是 Xception 网络、Xception+LSTM 网络、Xception+ConvGRU 网络、改进 Xception 网络、改进 Xception + LSTM 网络、改进 Xception+ConvGRU 网络。结果见表 3。

表 3 在 FF++数据集上对改进 Xception 网络与 ConvGRU 网络的消融实验

Table 3 Ablation experiment of the improved Xception network and ConvGRU network on the FF++ dataset

模型	DF	FS	F2F	NT
Xception[8]	83.70	87.21	83.17	87.90
Xception+LSTM[14]	94.36	95.37	94.12	95.93
Xception+ConvGRU	94.97	96.12	95.84	95.23
改进 Xception	90.48	92.01	90.78	89.94
改进 Xception+LSTM	95.35	96.50	96.07	93.33
改进 Xception+ConvGRU	96.35	98.73	97.74	95.13

通过第一行与第四行对比,添加了 SE 块的 Xception 模块在虚假人脸识别中的准确率平均提高

了约 4%，这证明 SE 模块在深度可分离卷积中的有效性。同时注意到对 NT 伪造方法的性能提升并不显著，这是因为 NT 方法中的纹理信息较为特殊，并不能很好地被网络捕捉到；由第二、第三两行及第五、第六两行可以看出，无论 Xception 网络是否添加 SE 块，卷积门控循环网络总能表现得比 LSTM 更好，证明了在 Xception 特征处理后 ConvGRU 的卷积部分发挥了作用。

3.3.2　整体网络的消融实验

为了证明特征提取与特征处理模块对本系统模型的整体有效性，分别比较了时空注意力模块、特征处理模块及总体网络在 FF++数据集上的综合表现，仿真结果见表 4。由表 4 中数据可知，仅使用空间注意力模块只能获得不到 70%的分类结果，而使用添加时间自注意力的时空注意力模块作为分类依据直接进行分类则将准确率提升到了 90%，单独使用特征处理模块作为分类依据的准确率可达 94%。但将两模块结合后，将特征提取模块的注意力权重作为输入配置在特征处理模块中的开始位置，相当于提前告知网络哪些内容需要给予特别关注，因此整体网络的准确性可达 97%，证明了本文整体网络架构的有效性。

表 4　在 FF++数据集上对改进 Xception 网络与 ConvGRU 网络的消融实验

Table 4　Ablation experiment of the improved Xception network and ConvGRU network on the FF++ dataset

模型	DF	FS	F2F	NT
ConvGRU[29]	50.83	50.13	48.39	50.00
IntraAttention	68.23	68.32	64.07	60.71
DualAttention	91.97	92.15	90.74	88.51
改进 Xception+LSTM	95.35	96.50	96.07	93.33
改进 Xception+ConvGRU	96.35	98.73	97.74	95.13
DualAtt+改进 Xcep+ConvGRU	97.12	98.57	97.28	96.43

3.4　与其他优秀方法的对比分析

为证明本文方法相比于以往研究的性能提升，表 5 中展示了与最新的一些虚假人脸检测方法结果在 FF++数据集上的对比。结果显示，本文方法对比 4 种不同伪造方法综合表现最优，其中在 DF、F2F、FS 上均为最优，在 NeuralTextures 上的表现仅次于 OC-FakeDect，这是因为 NT 是利用延迟神经渲染网络优化纹理的方法生成的更加真实的视频图像，而 OC-FakeDect 仅使用真实图像进行编码，可以更有针对性地检测真实图像的特征纹理。

表 6 展示了本文方法与其他方法的 *AUC* 比较数据，由于在高压缩 c40 下能找到对外展示 *AUC* 的方法不多，且缺少 FF++中 4 种伪造方法的具体数据，表 6 只对 FF++低质量数据集的整体 *AUC* 表现进行对比。本文模型取得了 97%的不错成绩，原因是对比的方法都仅针对深度虚假伪造视频的空间、频率信息进行利用，缺乏对时间上的关注度。

表 5　本文方法与国内外优秀算法在 FF++低质量数据集上的 *Acc* 对比

Table 5　*Acc* comparison between the method in this paper and excellent algorithms at home and abroad on FF++ low-quality data sets

模型	DF	FS	F2F	NT
Xception[8]	83.70	87.21	83.17	87.90
F3Net[22]	96.01	93.62	94.33	86.37
Sstnet[23]	93.40	91.90	91.90	-
OC-FakeDect[24]	88.40	71.20	86.10	97.50
SPSL[27]	93.48	86.02	92.26	76.78
本文模型	97.12	98.57	97.28	96.43

表 6　本文方法与国内外优秀方法在 FF++数据集上的 *AUC* 比较

Table 6　*AUC* comparison between the method in this paper and the excellent methods at home and abroad on the FF++ dataset

模型	*AUC*
Xception[8]	0.938 2
Meso-4[29]	0.876 6
F3Net[22]	0.933 0
SPSL[27]	0.828 2
Dual Network[30]	0.929 7
本文模型	0.971 3

4　结束语

本文针对亟需解决的虚假人脸伪造检测问题，现有的检测方法在低质量视频压缩信息的情况下缺乏高准确率的检测方案的问题、以及寻常的检测方法对帧间信息利用并不充分等问题，提出了一个基于时空注意力的网络框架，抑制无关背景信息内容的影响，分别对帧内和帧间有效的人脸信息进行提取。得到特征图后，输入到后续的深度可分离卷积网络和卷积循环门控网络中，对帧内帧间信息进行深层次提取，最后输入到分类器中得到真假分类结果。相比于现有的一些方法，本方法获得了 4 种伪造方法上的综合最优结果，取得了准确率 97.35%的良好结果。

参考文献

[1] GitHub Pages. Faceswap: Deepfakes software for all [EB/OL].

[2023-01-01]. https://github.com/deepfakes/faceswap.
[2] MARR B. The best (and scariest) examples of AI-enabled deepfakes [EB/OL]. [2019-07-22]. https://www.forbes.com/sites/bernardmarr/2019/07/22/the-best-and-scariest-examples-of-ai-enabled-deepfakes/.
[3] GUARDIANT. Chinesedeepfake app zao sparks privacy row after going viral [EB/OL]. [2019-09-02]. https://www.theguardian.com/technology/2019/sep/02/chinese-face-swap-appzao-triggers-privacy-fears-viral.
[4] TUCKER P. The newestai-Enabled weapon: Deep-faking photos of the Earth [EB/OL]. [2019-03-01]. https://www.defenseone.comtechnology/2019/03/next-phase-aideep-faking-whole-world-and-china-ahead/155944/.
[5] DAMIANI J. A voicedeepfake was used to scam a ceo out of $243, 000[EB/OL]. [2019-09-03]. https://vww.forbes.com/sites/jessedamiani/2019/09/03/a-voicedeepfake-was-used-to-scam-a-ceo-out-of-243000/.
[6] FaceSwap. FaceSwap github [EB/OL]. [2023-01-01]. https:// github.com/MarekKowalski/FaceSwap.
[7] THIES J, ZOLLHOFER M, STAMMINGER M, et al. Face2face: Real-time face capture and reenactment of RGB videos [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 2387-2395.
[8] CHOLLET F. Xception: Deep learning with depthwise separable convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017: 1251-1258.
[9] TAN M X, QUOC V L. EfficientNet: Rethinking model scaling for convolutional neural networks [C]//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019: 6105-6114.
[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Advances in Neural Information Processing Systems. Long Beach, USA: NIPS Foundation, 2017, 30: 5998-6008.
[11] LI Lingzhi, BAO Jianmin, ZHANG Ting, et al. Face X-ray for more general face forgery detection [C]// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 5000-5009.
[12] ZHAO Tianchen, XU Xiang, XU Mingze, et al. Learning self-consistency for deepfake detection [C]// Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 15003-15013.
[13] SHIOHARA K, YAMASAKI T. Detectingdeepfakes with self-blended images [C]// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022: 18699-18708.
[14] GÜERA D, DELP E J. Deepfake video detection using recurrent neural networks[C]//2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Auckland, New Zealand: IEEE, 2018: 1-6.
[15] ROSSLER A, COZZOLINO D, VERDOLIVA L, et al. Faceforensics++: Learning to detect manipulated facial images [J]. arXiv preprint arXiv: 1901.08971, 2019.
[16] ZHANG Kai, JIANG Yushan, SEVERSKY L, et al. Federated variational learning for anomaly detection in multivariate time series[C]//2021 IEEE International Performance, Computing, and Communications Conference (IPCCC). IEEE, 2021: 1-9.
[17] SHI Xingjian, CHEN Zhourong, WANG Hao, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting [J]. Advances in Neural Information Processing Systems. Montreal, Canada: NIPS Foundation, 2015, 28: 802-810.
[18] HU Jie, SHEN Li, ALBANIE S, et al. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7132-7141.
[19] LIU X, ZHANG L, LI T, et al. Dual attention guided multi-scale CNN for fine-grained image classification[J]. Information Sciences, 2021, 573: 37-45.
[20] ZHAO Hanqing, ZHOU Wenbo, CHEN Dongdong, et al. Multi-attentional deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021: 2185-2194.
[21] WANG Junke, WU Zuxuan, CHEN Jingjing, et al. M2tr: Multi-modal multi-scale transformers for deepfake detection[J]. arXiv preprint arXiv: 2104.09770, 2021.
[22] WEI Junhang, WANG Shuhui, HUANG Qingming. F^3Net: fusion, feedback and focus for salient object detection [C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020, 34(7): 12321-12328.
[23] ZHANG W, LI Z, SUN H H, et al. SSTNet: Spatial, spectral, and texture aware attention network using hyperspectral image for corn variety identification [J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.
[24] KHALID H, WOO S S. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE, 2020: 656-657.
[25] LUO Yuchen, ZHANG Yong, YAN Junci, et al. Generalizing face forgery detection with high-frequency features[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 16317-16326.
[26] JIA Gengjun, ZHENG Meisong, HU Chuanrui, et al. Inconsistency-aware wavelet dual-branch network for face forgery detection[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021, 3(3): 308-319.
[27] LIU Honggu, LI Xiaodan, ZHOU Wenbo, et al. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021: 772-781.
[28] BALLAS N, YAO Li, PAL C, et al. Delving deeper into convolutional networks for learning video representations[J]. arXiv preprint arXiv: 1511.06432, 2015.
[29] 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). Hong Kong, China: IEEE, 2018: 1-7.
[30] JIA Gengyun, ZHENG Meisong, HU Chuanrui, et al. Inconsistency-aware wavelet dual-branch network for face forgery detection[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021, 3(3): 308-319.