

中图法分类号: TP391.41 文献标识码: A 文章编号: 1006-8961(2024)02-0457-11

论文引用格式: Feng C B, Liu C X, Wang Y Y and Zhou Q D. 2024. Face forgery detection with image patch comparison and residual map estimation. Journal of Image and Graphics, 29(02):0457-0467(冯才博, 刘春晓, 王昱烨, 周其当. 2024. 结合图像块比较与残差图估计的人脸伪造检测. 中国图象图形学报, 29(02):0457-0467)[DOI:10.11834/jig.230149]

结合图像块比较与残差图估计的人脸伪造检测

冯才博¹, 刘春晓^{1*}, 王昱烨¹, 周其当²

1. 浙江工商大学计算机科学与技术学院, 杭州 310018; 2. 浙江工商大学信息与工程学院, 杭州 310018

摘要: 目的 由于不同伪造类型样本的数据分布差距较大, 现有人脸伪造检测方法的准确度不够高, 而且泛化性能差。为此, 本文引入“图像块归属纯净性”和“残差图估计可靠性”的概念, 提出了基于图像块比较和残差图估计的人脸伪造检测方法。方法 除了骨干网络, 本文的人脸伪造检测神经网络主要由纯净图像块比较模块和可靠残差图估计模块两部分组成。为了避免在同时包含人脸和背景像素的图像块上提取的混杂特征对于图像块比较的干扰, 纯净图像块比较模块中选择只包含人脸像素的纯净人脸图像块和只包含背景像素的纯净背景图像块, 通过比较两种图像块纯净特征之间的差异来检测伪造图像, 图像块的纯净性保障了特征提取的纯净性, 从而提高了特征比较的鲁棒性。考虑到靠近伪造边缘的像素比远离伪造边缘的像素具有较高的残差估计准确度, 本文在可靠残差图估计模块中根据像素到伪造边缘的距离设计了一个距离场加权的残差损失来引导网络的训练过程, 使网络重点关注输入图像与对应真实图像在伪造边缘附近的差异, 对于可靠信息的关注进一步增强了伪造检测的鲁棒性。结果 在FF++(FaceForensics++)数据集上的测试结果显示: 与对比算法中性能最好的F2Trans-B相比, 本文方法的准确率和AUC(area under the ROC curve)指标分别提高了2.49%和3.31%, 在FS(FaceSwap)与F2F(Face2Face)两种伪造数据上的准确率指标分别提高了6.01%和3.99%。在泛化性能方面, 与11种已有方法在交叉数据集上的测试结果显示: 本文方法与其中性能最好的方法相比, 在CDF(Celeb-DF)数据集上的视频AUC指标和图像AUC指标分别提高了1.85%和1.03%。结论 与对比方法相比, 由于提高了特征信息的纯净性和可靠性, 本文提出的人脸图像伪造检测模型的泛化能力和准确率优于对比方法。

关键词: 人脸图像伪造检测; 深度伪造; 多任务学习; 泛化性能; 像素级监督; 卷积神经网络

Face forgery detection with image patch comparison and residual map estimation

Feng Caibo¹, Liu Chunxiao^{1*}, Wang Yuye¹, Zhou Qidang²

1. School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou 310018, China;

2. School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China

Abstract: Objective The face recognition technique has become a part of our daily lives in recent years. However, with

收稿日期: 2023-03-24; 修回日期: 2023-07-01; 预印本日期: 2023-07-08

* 通信作者: 刘春晓 cxliu@mail.zjgsu.edu.cn

基金项目: 国家自然科学基金项目(61976188); 国家级大学生创新创业训练计划项目(202310353023X); 浙江省自然科学基金项目(LY24F020004); 浙江工商大学“数字+”学科建设项目(SZJ2022B016); 浙江省大学生科技创新活动计划暨新苗人才计划(2023R408037, 2022R408A028)

Supported by: National Natural Science Foundation of China(61976188); National College Students Innovation and Entrepreneurship Training Program(202310353023X); Natural Science Foundation of Zhejiang Province, China(LY24F020004); Zhejiang Gongshang University“Digital+”Disciplinary Construction Management Project(SZJ2022B016); Zhejiang University Students Science and Technology Innovation Activity Plan and New Talent Plan(2023R408037, 2022R408A028)

the rapid development of face forgery techniques based on deep learning, the cost of face forgery has not only been considerably reduced, but unexpected risks to the face recognition technique have also been raised. If someone uses a fake face image to break into a face recognition system, our personal information and property will be compromise and may even be stolen. However, distinguishing whether the face in an image is forged is difficult for the human eyes. Moreover, existing face forgery detection methods exhibit poor generalization performance and are difficult to defend against unknown attack samples due to large data distribution gaps among different forgery samples. Therefore, a reliable and general face forgery detection method is urgently required. In this regard, we introduce the concepts of “patch attribution purity” and “residual estimation reliability”, and propose a novel multitask learning network (PuRe) based on pure image patch comparison (PIPC) and reliable residual map estimation (RRME) to detect face forgery images. **Method** Apart from the network backbone, our neural network consists of the PIPC module and the RRME module. Both modules are helpful for improving the performance of face forgery detection. On the one hand, if the face in an image is forged, then the features extracted from face and background patches should be inconsistent. The PIPC module compares feature discrepancy between face and background patches to complete the face forgery detection task. Nevertheless, if an image patch contains face and background pixels, then the features extracted from it will have mixed face and background information, disturbing the feature comparison between face and background image patches and resulting in the overfitting of the training dataset. Considering the aforementioned problem, our PIPC module suggests using only pure image patches, which only contain face pixels (pure face image patches) or background pixels (pure background image patches). The purity of image patches guarantees the purity of the extracted features, and thus, the robustness of feature comparison is improved. On the other hand, the residual map estimation task is designed to predict the difference between the input image and the corresponding real image, causing the network backbone to strengthen the generalization of the extracted image features and improving the accuracy of face forgery detection. However, for pixels that are far from the forged edges between the forgery and real regions, the known information used to estimate the residuals will be less, resulting in unreliable residual estimation. Considering the aforementioned problem, a loss function, called the distance field weighted residual loss (DWRLoss), is designed in the RRME module to compel the neural network to give more attention to estimating the residuals near the forged edges between the forgery and real regions. In the face region (i. e., forgery region), if the pixel is far from the background region, then its loss is assigned with a smaller weight coefficient. Attention to reliable residual information improves the robustness of face forgery detection. Finally, we adopt the multitask learning strategy to train the proposed neural network. Both learning tasks guide the network backbone together to extract effective and generalized features for face forgery detection. **Result** A large number of experiments are conducted to demonstrate the superiority of our method. Compared with existing superior methods, the test results on the FaceForensics++ (FF++) dataset show that the accuracy (ACC) and area under the receiver operating characteristic curve (AUC) of face forgery detection are improved by 2.49% and 3.31%, respectively, by using the proposed method. Moreover, our method improves the ACC of face forgery detection on the FF++ dataset with FaceSwap (FS) and Face2Face (F2F) forgery types by 6.01% and 3.99%, respectively. In terms of the cross-dataset test, compared with 11 existing representative methods, the experimental results show that AUC on the Celeb-DF (CDF) dataset in the video and image levels is increased by 1.85% and 1.03%, respectively, with our method. **Conclusion** The proposed neural network (i. e., PuRe) based on the PIPC and RRME modules exhibits excellent generalization ability and performs better than existing methods due to the purity and reliability of the extracted features.

Key words: face forgery detection; deepfake; multi-task learning; generalization; pixel-wise supervision; convolutional neural network

0 引言

由于图像深度合成技术日益成熟,人眼已经难以辨别真实与伪造图像。目前已经有不少软件提供了

方便的换脸功能,人脸图像的伪造成本大幅度降低。通过修改图像中的人脸,可以伪造他人的身份,这在人脸支付安全、视觉通信安全等方面构成了极大的威胁。为了保护人脸图像的安全,迫切需要一个有效且鲁棒的人脸图像伪造检测方法。

目前,人脸图像伪造检测方法面临的挑战主要有两方面。一方面,由于人脸图像的伪造方法不断更新,无法获得所有伪造方法的数据用于训练模型;另一方面是网络模型结构的设计。由于伪造人脸的方法不同,使用不同方法制作的样本数据之间可能存在较大的分布差异,容易导致分类模型对训练数据产生过拟合现象。

为了应对上述挑战,已经出现了不少针对人脸图像的伪造检测方法(曹申豪等,2022)。在伪造样本制作方面,一些工作(Li等,2020a;Zhao等,2021a;Shiohara和Yamasaki,2022)通过相似的样本之间拼接边缘的光滑过渡来生成具有挑战性的伪造人脸图像。

在网络结构设计方面,一些工作(Liu等,2021;Luo等,2021;Chen等,2021a;Miao等,2023)利用真实与伪造人脸图像的频域差异来完成检测任务。一些工作(Zhao等,2021b;Chen和Yang,2021;Wang等,2022a)通过引入注意力机制来提高网络的泛化性能。Chen和Yang(2021)将原始图像与背景、人脸、五官图像一起输入含有局部与语义注意力模块的网络,捕捉人脸五官中的细粒度信息。为了解决检测性能随图像质量的退化而下降的问题,Wang等人(2022a)通过约束从原始图像与退化图像中提取的特征相同,并引入多头自注意力机制来引导模型捕捉有效信息。Wang等人(2022b)与Chen等人(2022)使用对抗样本来提高模型的泛化能力。Wang等人(2022b)引入像素级高斯模糊,通过破坏图像中的伪造痕迹来提高样本的识别难度。Chen等人(2022)在对抗训练的过程中动态地合成样本来干扰网络以提高检测的鲁棒性。Ni等人(2022)与Cao等人(2021)分别使用基于对比学习与度量学习的方法来提高模型的检测准确率。Guo等人(2021)提出自适应卷积层来预测图像中的伪造痕迹,避免下采样操作导致伪造痕迹的消失。Kim等人(2021)提出基于迁移学习的方法,保持网络的先验知识。Wang和Deng(2021)和Das等人(2021)使用随机遮盖来消除图像伪造痕迹,强迫网络提取有效线索。一些工作(Du等,2020;Wang等,2022c;Chen等,2021b;Cao等,2022;Zhou等,2017;Nirkin等,2022)主要运用多任务学习策略来提高模型的泛化性能。总之,已有方法仍然存在检测准确率不高、网络泛化性能不足等问题。

一方面,考虑到在真实图像的人脸区域与背景区域中包含的光照分布、相机参数等特定信息是一致的,而在伪造图像经过修改的人脸区域与背景区域中包含的特定信息是不一致的,Zhang等人(2022)利用不同图像块之间特定信息的不一致性来分辨真实与伪造人脸图像,但是他们没有注意到同时包含人脸与背景信息的混合图像块对特征差异比较的干扰作用;另一方面,注意到已有工作(Wang等,2022a)中残差图预测任务可以促使骨干网络学习到具有更高泛化性的特征表示,但是伪造图像中远离原始背景区域、周围缺乏可靠信息的像素的残差预测难度较大,此处类似图像补全那样无中生有的残差图预测任务容易导致网络对训练数据产生过拟合现象。为了解决上述问题,本文提出了结合纯净图像块比较与可靠残差图估计的多任务人脸图像伪造检测方法(pure image patch comparison and Reliable residual map estimation, PuRe)。

本文的主要贡献如下:1)引入“图像块归属纯净性”的概念,提出了纯净图像块比较(pure image patch comparison, PIPC)模块,通过比较纯净人脸图像块与纯净背景图像块之间的特征差异来检测伪造图像,图像块的纯净性保障了特征提取的纯净性,从而提高了特征比较的鲁棒性。2)引入“残差图估计可靠性”的概念,提出了可靠残差图预测(reliable residual map estimation, RRME)模块,有针对性地设计了一个距离场加权的残差损失(distance field weighted residual loss, DWRLoss),引导骨干网络重点关注真伪图像在伪造边缘附近的差异,增加了特征提取的丰富性和可靠性。3)基于上述两个模块,提出基于多任务学习的人脸图像伪造检测网络框架,RRME模块通过预测伪造图像与对应真实图像之间的残差图来引导骨干网络差异化从伪造人脸区域与原始背景区域提取的特征,从而辅助PIPC模块更好地分辨人脸区域与背景区域之间的特征差异。在常用数据集上开展的大量实验证实了本文方法的有效性。

1 本文方法

1.1 总体网络框架

图1展示了本文提出的多任务学习网络框架。将通道数为3、高度为 H 、宽度为 W 的待检测人脸图像 $I \in \mathbf{R}^{3 \times H \times W}$ 输入骨干网络进行特征提取,得到通道

数为 C 、高度为 H' 、宽度为 W' 的特征图 $F \in \mathbb{R}^{C \times H' \times W'}$ 。特征图 F 中的每个位置对应一个局部图像块,表示为特征向量 $V \in \mathbb{R}^c$,总共有 $S = H' \times W'$ 个图像块。

输入特征图 F 与其对应的人脸区域掩膜 $M \in \mathbb{R}^{1 \times H' \times W'}$ 到纯净图像块比较(PIPC)模块,同时将特征图 F 输入可靠残差图估计(RRME)模块。PIPC 模块根据人脸区域掩膜 M 选择只包含人脸区域的纯净人脸图像块和只包含背景区域的纯净背景图像块,通过比较两种纯净图像块之间的特征差异来区分真实图像与伪造图像。RRME 模块预测输入图像与真实图

像之间的残差图 $R \in \mathbb{R}^{3 \times H' \times W'}$,并且在距离场加权的残差损失(DWRLoss)的约束下重点关注在伪造边缘附近的可靠残差值。RRME 模块通过预测伪造图像与对应真实图像之间的残差图,引导骨干网络加大从伪造图像的人脸区域与背景区域中提取特征的差异,减小从真实图像的人脸区域与背景区域中提取特征的差异,助力 PIPC 模块分辨人脸区域与背景区域之间的特征差异,若特征差异较大,则认为输入图像为伪造图像;若特征差异较小,则认为输入图像是真实图像。

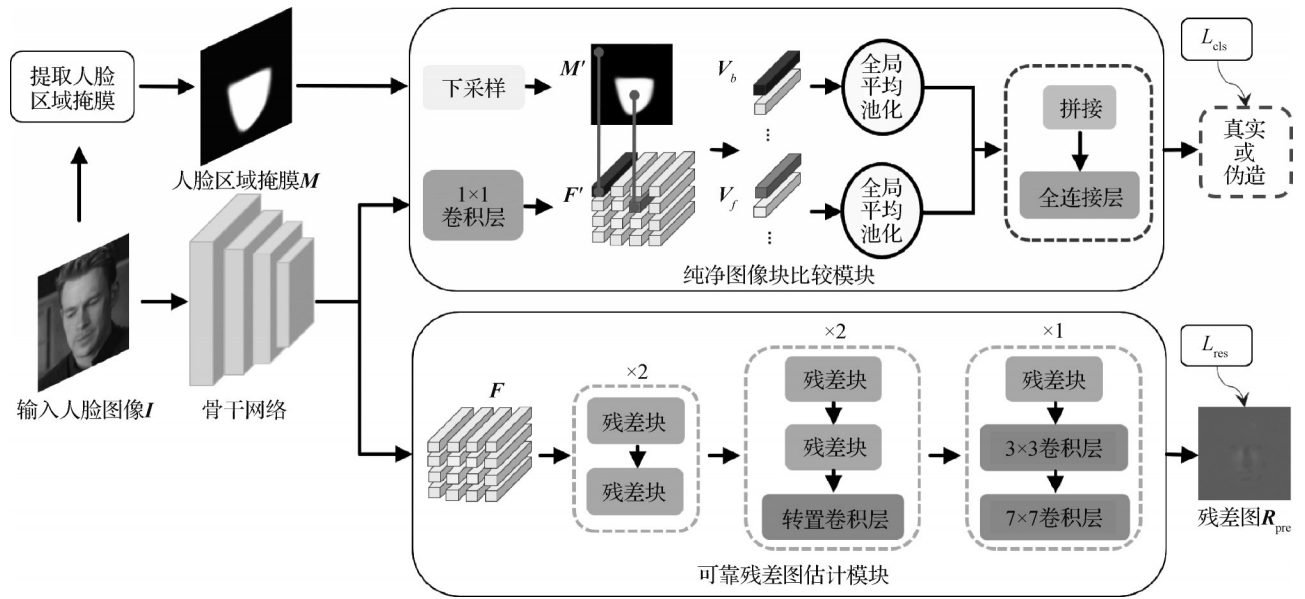


图1 总体网络框架

Fig. 1 Overall network architecture

1.2 纯净图像块比较模块

如果图像中的人脸是伪造的,那么从人脸图像块与背景图像块中提取的特征应该是有差异的,因此 PIPC 模块通过比较人脸图像块与背景图像块之间的特征差异来完成人脸图像伪造检测任务。然而,如果图像块同时包含人脸和背景像素,则从中提取的特征会混合人脸和背景信息,从而干扰了两个图像块之间的比较,导致神经网络的过拟合。针对上述问题,PIPC 模块使用只包含人脸像素的纯净人脸图像块与只包含背景像素的纯净背景图像块。

将特征图 F 及其对应的人脸区域掩膜 M 输入 PIPC 模块。首先使用 1×1 卷积层将特征图 F 映射到新的特征空间得到 F' ,再通过双线性插值将人脸区域掩膜 M 的尺寸调整到与特征图相同,得到纯净

性分数图 $M' \in \mathbb{R}^{1 \times H' \times W'}$ 。其中像素 i 的值表示特征图 F' 中对应像素的特征向量 V_i 所对应的局部图像块的纯净性分数 $s_i \in [0, 1]$, 0 表示纯净背景图像块, 1 表示纯净人脸图像块,处于 0 与 1 之间的值表示混合图像块。

为了减少被挑选的纯净图像块之间感受野的重叠,避免特征向量中信息的冗余,本文定义了一个纯净人脸图像块与纯净背景图像块的选择优先级计算函数 $f_s(\cdot)$,具体为

$$f_s(V_i) = \begin{cases} f_f & y = 1 \\ f_b & y = 0 \end{cases} \quad (1)$$

$$f_f = \omega_1 s_i + \omega_2 f_d(V_f, V_i) + \omega_3 f_d(V_b, V_i) \quad (2)$$

$$f_b = \omega_1 (1 - s_i) + \omega_3 f_d(V_f, V_i) + \omega_2 f_d(V_b, V_i) \quad (3)$$

式中,当 $y = 1$ 时,表示选取纯净人脸图像块;当 $y = 0$

时,表示选取纯净背景图像块。式(2)和式(3)中, ω_1 、 ω_2 和 ω_3 分别表示3个权重系数, V_f 与 V_b 分别表示锚点纯净人脸图像块和锚点纯净背景图像块。为了方便计算,本文选取掩膜 M' 中人脸区域最小外接四边形的中点为锚点纯净人脸图像块 V_f ,选取特征图 F 中左上角的图像块为锚点纯净背景图像块 V_b 。 $f_d(\cdot, \cdot)$ 是一个用于计算两个图像块之间距离的函数,具体为

$$f_d(V_a, V_i) = \sqrt{(x_a - x_i)^2 + (y_a - y_i)^2} \quad (4)$$

式中, V_a 表示锚点纯净图像块, V_i 表示当前图像块, x_a 和 y_a 分别表示锚点纯净图像块的 x 轴坐标和 y 轴坐标, x_i 和 y_i 分别表示当前图像块的 x 轴坐标和 y 轴坐标。计算所有图像块的 f_d 值,并按从大到小的顺序进行排序,取前 $N-1$ 个 f_d 值最大的图像块和锚点纯净图像块对应的特征向量。当 $y=1$ 和 $y=0$ 时,分别得到特征向量 $V_f \in \mathbb{R}^{C \times N}$ 和 $V_b \in \mathbb{R}^{C \times N}$,用于比较纯净人脸图像块与纯净背景图像块之间的特征差异。

首先使用全局平均池化将 V_f 和 V_b 分别压缩为 $V_f' \in \mathbb{R}^C$ 和 $V_b' \in \mathbb{R}^C$,聚合两种纯净图像块的信息,再将 V_f' 和 V_b' 拼接起来并依次输入全连接层和softmax激活函数,输出分类结果 C_{pre} 。

1.3 可靠残差图预测模块

将特征图 F 输入RRME模块,预测输入图像与真实图像之间的残差图 $R_{pre} \in \mathbb{R}^{3 \times H \times W}$ 。网络训练时使用的残差图真值 R_{gt} 可以表达为

$$R_{gt} = 0.5(I - I_R) + 0.5 \quad (5)$$

式中, I 为输入图像, I_R 为其对应的真实图像。 I 和 I_R 的像素值在 $[0, 1]$ 之间,所以 $I - I_R$ 的像素值在 $[-1, 1]$ 之间,式(5)通过线性变换将残差图 R_{gt} 的值映射至 $[0, 1]$ 之间。特别地,若 I 为真实图像则 $I = I_R$,意味着真实图像的残差图真值 R_{gt} 为全灰图像。

残差图预测任务引导骨干网络提取有利于人脸伪造检测的特征,但是通过对所有像素施加相同权重的损失约束,简单地约束网络预测残差图是不合适的。与图像修补任务类似,伪造区域就像图像修补任务中的空洞,RRME模块用空洞周围的已知背景信息来估计伪造区域中的残差信息。越靠近已知背景区域,残差估计所依赖的可靠信息越多,残差估计的可靠性越高;反之,越远离已知背景区域、靠近伪造区域的中心,残差估计所依赖的可靠信息越

少,残差估计的可靠性越低。无差别地约束RRME模块估计伪造区域中所有像素的残差值,尤其是远离已知背景区域的残差值,容易导致严重的过拟合现象。为了解决上述问题,本文设计了一个如图2所示的距离场加权的残差损失(DWRLoss),引导网络重点关注伪造边缘附近的残差估计结果,其计算为

$$L_{res} = D \times \|R_{pre} - R_{gt}\|_2 \quad (6)$$

式中, $R_{pre} \in \mathbb{R}^{3 \times H \times W}$ 为网络输出的残差图预测结果, $D \in \mathbb{R}^{1 \times H \times W}$ 为距离场。在距离场 D 中,背景区域的值为1,人脸区域(即伪造区域)的值随着像素与背景区域的距离增大,由1递减至0。这意味着若像素距背景区域近,该像素的损失被赋予的权重系数较大;若像素距背景区域远,该像素的损失被赋予的权重系数较小。特别地,当输入图像为真实图像时,距离场中的像素值全部为1。

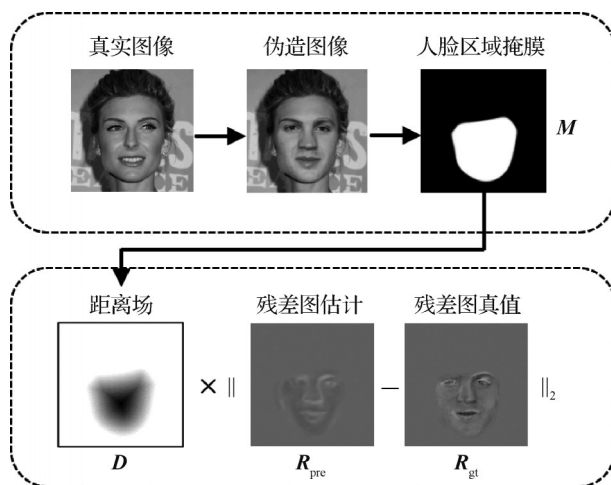


图2 距离场加权的残差损失示意图

Fig. 2 Illustration of the distance field weighted residual loss

1.4 多任务学习策略

在训练过程中,整体损失函数 L_{total} 由分类损失 L_{cls} 和距离场加权的残差损失 L_{res} 两部分组成。分类损失 L_{cls} 优化PIPC模块,距离场加权的残差损失 L_{res} 优化RRME模块,两个损失共同优化骨干网络,整体损失函数 L_{total} 可定义为

$$L_{total} = \lambda_1 L_{res} + \lambda_2 L_{cls} \quad (7)$$

式中,分类损失 L_{cls} 采用交叉熵损失函数, λ_1 和 λ_2 分别为距离场加权的残差损失 L_{res} 和分类损失 L_{cls} 的权重系数。

2 实验与分析

2.1 实验设置

2.1.1 训练数据集

本文选择 FaceForensics++ (FF++) (Rössler 等, 2019) 作为训练数据集之一。FF++ 是一个广泛使用的大型人脸伪造检测数据集。该数据集包含 1 000 个从 YouTube 采集的真实视频, 以及使用这 1 000 个真实视频制作的 4 000 个伪造视频。伪造视频由 4 种方法制作: 1) DF (DeepFake) (Tora, 2019); 2) FS (FaceSwap) (Thies 等, 2016); 3) F2F (Face2-Face) (Kowalski, 2016); 4) NT (NerualTexture) (Thies 等, 2019)。该数据集提供了 3 个不同质量等级的版本: 原始 (RAW)、轻压缩 (light compression, C23) 和重压缩 (high compression, C40)。本文使用 RAW 版本的 FF++ 作为训练数据, 并按照 8:1:1 的比例划分训练集、验证集和测试集。

CelebA (CelebFaces attributes) (Karras 等, 2018) 是一个大型人脸属性数据集, 包含 202 599 幅人脸图像。它通常被用于人脸识别、人脸检测和人脸属性识别任务中。本文增广它的高质量版本中的部分人脸图像作为训练数据集, 以增加人脸样本数量。

2.1.2 测试数据集

除了从 FF++ 中划分的测试集, 本文还使用了两个公开数据集作为测试数据: 1) CDF (Celeb-DF) (Li 等, 2020b); 2) DFD (Deepfake detection) (Dufour 和 Gully, 2019)。CDF 数据集由 590 个真实视频和 5 639 个伪造视频组成。相较于 FF++, CDF 使用的伪造方法更加先进, 检测难度更大。因此, 已有工作通常使用 CDF 数据集来测试人脸伪造检测方法的泛化性能。DFD 是由 Google 公开的人脸伪造检测数据集, 更接近现实世界中的人脸伪造样本。该数据集同样有 3 个压缩版本, 本文使用轻压缩版本 (C23)。

2.1.3 实现细节

本文利用 dlib 库提取人脸特征点并制作人脸区域掩膜。从 CelebA 数据集中随机选取 20 000 幅人脸图像作为真实样本, 使用 Li 等人 (2020a) 的方法进行增广, 即计算每幅图像与其他图像的人脸特征点的 L2 距离, 选取距离小于 25 的图像作为最近邻图像, 然后将最近邻图像中的人脸与当前图像混合, 得

到新的伪造人脸图像。对于其他数据集, 每个视频提取 32 帧图像作为训练或测试数据。所有图像的尺寸均为 256×256 像素, 在输入网络前归一化到 $[0, 1]$ 之间。

选择 AdamW 优化器来训练网络模型, 初始学习率与批量大小分别设置为 0.000 2 与 128。使用 ResNet-34 (residual network 34) 作为骨干网络并使用在 ImageNet 上的预训练权重来初始化网络参数, 将 ResNet-34 第 3 阶段提取的特征图 F 作为 PIPC 模块和 RRME 模块的输入。本文的训练策略是首先固定 PIPC 模块的参数, 使用 FF++ 数据集和增广后的 CelebA 数据集训练骨干网络和 RRME 模块 120 轮。然后, 开放网络的所有部分训练 200 轮。超参数 ω_1 、 ω_2 、 ω_3 、 N 、 λ_1 和 λ_2 分别默认设置为 1、 $3E-3$ 、 $3E-3$ 、16、100 和 $1E-2$ 。本文方法在 Pytorch 框架下实现, 使用一张 NVIDIA GeForce RTX 3090 GPU 显卡进行训练。参考已有工作, 本文主要使用准确度 (accuracy, ACC) 和 AUC (area under the ROC curve) 指标来评估网络性能。

2.2 库内数据集测试

2.2.1 库内总精度

在 FF++ 的 3 个不同版本上进行了库内测试, 结果如表 1 所示。在两个压缩版本的数据集上, 本文方法取得的指标优于已有方法。特别地, 在重压缩版本 (C40) 上, 本文方法相比 LiSiam, ACC 指标提升 2.49%, AUC 指标提升 3.31%。

2.2.2 不同伪造方法测试

针对 FF++ 中的 4 个伪造方法, 本文分别在 3 个版本的数据集上进行测试并展示准确率 (ACC), 测试结果如表 2—表 4 所示。在表 3 中, 本文方法在 FS 和 F2F 两种伪造类型数据上的 ACC 指标分别超越 MTD-Net 0.24% 和 0.57%。在其他两个伪造类型的数据集上与表中对比方法基本持平。在表 4 中, 本文方法在 FS 和 F2F 两种伪造类型的数据上的 ACC 指标分别超越 GRCC 方法 6.01% 和 3.99%。从表 3 和表 4 的实验结果可以看到, 本文方法在 FS 类型的伪造样本上, 检测准确率远超过表中对比方法。

2.3 交叉数据集测试

如果一个网络模型具有优秀的泛化性能, 当遇到未知类型的伪造样本时, 它依然能够提取到样本中的伪造线索。因此, 使用交叉数据集测试的方式来评估本文方法的泛化性能, 即在一个数据集上训

表 1 不同方法在 FF++数据集上的测试结果
Table 1 Test results of different methods on the FF++ dataset

方法	RAW		C23		C40	
	ACC	AUC	ACC	AUC	ACC	AUC
Face X-ray(Li 等,2020a)	-	-	-	87.35	-	61.60
MADD(Zhao 等,2021b)	-	-	96.37	-	86.95	-
Xception(Rössler 等,2019)	99.26	-	95.73	-	81.73	-
LD-CNN(Cozzolino 等,2017)	98.57	-	78.45	-	58.69	-
SPSL(Liu 等,2021)	-	-	92.39	94.32	81.57	82.82
LiSiam(Wang 等,2022a)	-	-	96.51	99.13	87.81	91.44
PuRe(本文)	98.18	99.91	96.57	99.45	90.30	94.75

注:加粗字体表示各列最优结果,“-”表示无对应实验数据。

表 2 FF++ (RAW)数据集中不同伪造类型数据的测试结果
Table 2 Test results of different forgery types on the FF++ dataset (RAW)

方法	ACC/%			
	DF	FS	F2F	NT
Xception(Rössler 等,2019)	99.59	99.14	99.61	99.36
LD-CNN(Cozzolino 等,2017)	97.24	95.93	98.51	98.74
MesoNet(Afchar 等,2018)	99.24	98.15	98.35	97.96
PuRe(本文)	100.00	100.00	99.67	100.00

注:加粗字体表示各列最优结果。

表 3 FF++ (C23)数据集中不同伪造类型数据的测试结果
Table 3 Test results of different forgery types on the FF++ dataset (C23)

方法	ACC/%			
	DF	FS	F2F	NT
F ³ Net(Qian 等,2020)	98.43	98.51	98.34	93.22
MesoNet(Afchar 等,2018)	95.26	95.84	93.43	95.96
Xception(Rössler 等,2019)	98.85	98.36	98.23	94.50
DSP-FWA(Li 和 Lyu,2019)	96.57	96.57	97.33	91.51
MTD-Net(Yang 等,2021)	98.64	99.76	98.42	94.60
PuRe(本文)	97.98	100.00	98.99	95.97

注:加粗字体表示各列最优结果。

练,在另一个数据集上测试。表 5 分别展示了视频层面与图像层面的交叉数据集测试结果。参照已有方法,本文在视频层面的测试方法是从每个测试视频抽取 32 帧,对网络输出的 32 个分数取平均,作为

表 4 FF++ (C40)数据集中不同伪造类型数据的测试结果
Table 4 Test results of different forgery types on the FF++ dataset (C40)

方法	ACC/%			
	DF	FS	F2F	NT
LD-CNN(Cozzolino 等,2017)	75.00	51.00	56.00	62.00
MesoNet(Afchar 等,2018)	90.00	83.00	83.00	75.00
SPSL(Liu 等,2021)	93.48	92.26	86.02	76.78
GRCC(Liang 等,2022)	95.43	93.99	88.94	77.19
PuRe(本文)	95.96	100.00	92.93	75.76

注:加粗字体表示各列最优结果。

对该视频的预测结果。本文方法与 11 种已有方法进行了比较,结果如表 5 所示。由于大部分工作没有公开难以复现的代码,而且只展示了 AUC 性能指标,因此在性能比较时直接引用作者在原始论文中提供的 AUC 测试结果。在视频层面和图像层面,相较于 F2Trans-B,本文方法在 CDF 数据集上的 AUC 指标分别提高了 1.85% 和 1.03%。

2.4 消融实验

使用 FF++ (RAW)、CDF 和 DFD (C23)作为测试数据,对本文提出的 PIPC 模块和 RRME 模块进行消融实验,通过不同参数设置下的交叉数据集测试 AUC 指标,以评估两个模块的有效性。

2.4.1 图像块的数量对网络性能的影响

本文选取不同的纯净图像块数量 $N = 8, 16, 32, 64$ 进行网络训练和测试,测试结果如表 6 所示。当 N 为 16 时,网络在 FF++ 和 CDF 数据上的表现最好。

表5 不同方法在视频和图像层面的交叉数据集测试结果
Table 5 Cross-datasets test results of different methods in video-level and image-level

方法	AUC/%	
	CDF(video)	CDF(image)
Face X-ray(Li 等, 2020a)	79.50	-
Xception(Rössler 等, 2019)	73.70	65.50
PCL+I2G(Zhao 等, 2021a)	90.03	-
PD(Zhang 等, 2022)	82.25	-
Local-Relation (Chen 等, 2021a)	78.26	-
MADD(Zhao 等, 2021b)	76.65	67.44
SPSL(Liu 等, 2021)	-	76.88
GFF(Luo 等, 2021)	75.31	-
LiSiam(Wang 等, 2022a)	-	78.21
MTD-Net(Yang 等, 2021)	-	70.12
F2Trans-B(Miao 等, 2023)	89.87	83.05
PuRe(本文)	91.88	84.08

注:加粗字体表示各列最优结果,“-”表示无对应实验数据。

表6 图像块的数量对网络性能的影响
Table 6 The effects of patches number on the network performance

图像块	AUC/%		
	FF++(RAW)	CDF	DFD(C23)
8	95.48	89.78	80.01
16	99.91	91.88	90.5
32	97.17	88.99	93.98
64	96.99	88.61	89.19

注:加粗字体表示各列最优结果。

当 N 为32时,网络在DFD数据上的表现最好。综合来看,设置 $N=16$ 时取得最佳性能。

2.4.2 ω_2 和 ω_3 对网络性能的影响

式(2)和式(3)中的 ω_2 和 ω_3 用来控制当前图像块与同类锚点纯净图像块的距离以及当前图像块与异类锚点纯净图像块的距离的重要性。若 $\omega_2 > \omega_3$,表示当前图像块与同类锚点纯净图像块的距离更重要;若 $\omega_2 < \omega_3$,表示当前图像块与异类锚点纯净图像块的距离更重要;若 $\omega_2 = \omega_3$,表示两个距离同样重要。实验结果如表7所示。当 $\omega_2 = \omega_3$ 时,在FF++与CDF数据集上的表现最好。当 $\omega_2 < \omega_3$ 时,在DFD数

据集上的表现最好。综合来看,设置 $\omega_2 = \omega_3$ 时取得最好性能。

表7 ω_2 和 ω_3 对网络性能的影响
Table 7 The effects of ω_2 and ω_3 on the network performance

ω_2/ω_3	AUC/%		
	FF++ (RAW)	CDF	DFD (C23)
0.005/0.003	98.24	90.36	84.73
0.003/0.005	98.17	89.46	94.75
0.003/0.003	99.91	91.88	90.50

注:加粗字体表示各列最优结果。

2.4.3 损失函数权重系数对网络性能的影响

本文通过 λ_1 和 λ_2 来平衡残差损失和分类损失的重要性。实验结果如表8所示,两个损失函数的权重会对模型的表现有不同的影响。当 λ_2 较大时,网络更容易出现过拟合的现象;当 λ_1 较大时,距离场加权的残差损失可以引导骨干网络学习到更加有效的特征。设置 $\lambda_1 = 100$ 、 $\lambda_2 = 0.01$ 时取得最佳性能。

表8 损失函数权重系数对网络性能的影响
Table 8 The effects of weights in loss function on the network performance

λ_1/λ_2	AUC/%		
	FF++ (RAW)	CDF	DFD (C23)
1/1	95.16	88.16	77.98
10/0.1	97.47	87.47	89.3
100/0.01	99.91	91.88	90.5

注:加粗字体表示各列最优结果。

2.4.4 不同模块对网络性能的影响

为了证明PIPC模块、RRME模块与距离场加权的残差损失(DWRLoss)的有效性,本文训练了只包含PIPC模块的网络、无距离场加权的残差损失(w/o DWR)约束RRME模块的网络以及有距离场加权的残差损失(w/ DWR)约束RRME模块的网络进行性能测试和比较。从表9中的测试结果可知,有距离场加权的残差损失(w/ DWR)相较于无距离场加权的残差损失(w/o DWR)能够更好地引导网络的学习过程。另外,当两个模块共同作用于网络的训练过程时,网络的性能表现最好。

表 9 不同模块对网络性能的影响
Table 9 The effects of different modules on
the network performance

模块	AUC/%		
	FF++ (RAW)	CDF	DFD (C23)
PIPC	90.28	84.3	71.72
PIPC+RRME(w/o DWR)	97.12	88.75	80.58
PIPC+RRME(w/ DWR)	99.91	91.88	90.50

注:加粗字体表示各列最优结果。

3 结 论

为了进一步提高神经网络在人脸图像伪造检测任务中的泛化能力,本文提出了包含纯净图像块比较(PIPC)模块与可靠残差图性预测模块(RRME)的多任务学习网络模型(PuRe)。PIPC 模块通过比较纯净人脸图像块与纯净背景图像块之间的特征差异来检测伪造图像,图像块的纯净性保障了特征提取的纯净性,从而提高了特征比较的鲁棒性。RRME 模块通过预测伪造图像与对应真实图像之间的残差图,引导骨干网络加大从伪造图像的人脸区域与背景区域中提取特征的差异,减小从真实图像的人脸区域与背景区域中提取特征的差异,助力 PIPC 模块分辨人脸区域与背景区域之间的特征差异。在面对未知样本攻击时,本文方法的泛化性能比已有方法更好。未来仍然需要进一步提升人脸伪造检测算法的各方面性能。

参考文献(References)

Afchar D, Nozick V, Yamagishi J and Echizen I. 2018. MesoNet: a compact facial video forgery detection network//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security. Hong Kong, China: IEEE: 1-7 [DOI: 10.1109/WIFS.2018.8630761]
Cao J Y, Ma C, Yao T P, Chen S, Ding S H and Yang X K. 2022. End-to-end reconstruction-classification learning for face forgery detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4103-4112 [DOI: 10.1109/CVPR52688.2022.00408]
Cao S H, Liu X H, Mao X Q and Zou Q. 2022. A review of human face forgery and forgery-detection technologies. Journal of Image and

Graphics, 27(4): 1023-1038 (曹申豪, 刘晓辉, 毛秀青, 邹勤. 2022. 人脸伪造及检测技术综述. 中国图象图形学报, 27(4): 1023-1038) [DOI: 10.11834/jig.200466]
Cao S H, Zou Q, Mao X Q, Ye D P and Wang Z Y. 2021. Metric learning for anti-compression facial forgery detection//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: ACM: 1929-1937 [DOI: 10.1145/3474085.3475347]
Chen L, Zhang Y, Song Y B, Liu L Q and Wang J. 2022. Self-supervised learning of adversarial example: towards good generalizations for deepfake detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18689-18698 [DOI: 10.1109/CVPR52688.2022.01815]
Chen S, Yao T P, Chen Y, Ding S H, Li J L and Ji R R. 2021a. Local relation learning for face forgery detection. Proceedings of the AAAI Conference on Artificial Intelligence, 35(2): 1081-1088 [DOI: 10.1609/aaai.v35i2.16193]
Chen Z H and Yang H. 2021. Attentive semantic exploring for manipulated face detection//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, Canada: IEEE: 1985-1989 [DOI: 10.1109/ICASSP39728.2021.9414225]
Chen Z K, Xie L X, Pang S M, He Y and Zhang B. 2021b. MagDR: mask-guided detection and reconstruction for defending deepfakes//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9010-9019 [DOI: 10.1109/CVPR46437.2021.00890]
Cozzolino D, Poggi G and Verdoliva L. 2017. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. Pennsylvania, USA: ACM: 159-164 [DOI: 10.1145/3082031.3083247]
Das S, Seferbekov S, Datta A, Islam M S and Amin M R. 2021. Towards solving the DeepFake problem: an analysis on improving DeepFake detection using dynamic face augmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 3769-3778 [DOI: 10.1109/ICCVW54120.2021.00421]
Du M N, Pentyala S, Li Y N and Hu X. 2020. Towards generalizable deepfake detection with locality-aware autoencoder//Proceedings of the 29th ACM International Conference on Information and Knowledge Management. Virtual Event, Ireland: ACM: 325-334 [DOI: 10.1145/3340531.3411892]
Dufour N and Gully A. 2019. Contributing data to deepfake detection research [EB/OL]. [2023-03-09].
<https://torontoai.org/2019/09/23/contributing-data-to-deepfake-detection-research/>
Guo Z Q, Yang G B, Chen J Y and Sun X M. 2021. Fake face detection via adaptive manipulation traces extraction network. Computer

- Vision and Image Understanding, 204: #103170 [DOI: 10.1016/j.cviu.2021.103170]
- Karras T, Aila T, Laine S and Jaakko L. 2018. Progressive growing of GANs for improved quality, stability, and variation [EB/OL]. [2023-03-09]. <https://arxiv.org/pdf/1710.10196.pdf>
- Kim M, Tariq S and Woo S S. 2021. FReTAL: generalizing deepfake detection using knowledge distillation and representation learning// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville, USA: IEEE: 1001-1012 [DOI: 10.1109/CVPRW53098.2021.00111]
- Kowalski M. 2016. FaceSwap [EB/OL]. [2023-03-09]. <https://github.com/marekkowalski/faceswap>
- Li L Z, Bao J M, Zhang T, Yang H, Chen D, Wen F and Guo B N. 2020a. Face x-ray for more general face forgery detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5000-5009 [DOI: 10.1109/CVPR42600.2020.00505]
- Li Y Z and Lyu S W. 2019. Exposing deepfake videos by detecting face warping artifacts [EB/OL]. [2023-03-09]. <https://arxiv.org/pdf/1811.00656.pdf>
- Li Y Z, Yang X, Sun P, Qi H G and Lyu S W. 2020b. Celeb-DF: a large-scale challenging dataset for deepfake forensics//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3204-3213 [DOI: 10.1109/CVPR42600.2020.00327]
- Liang J H, Shi H F and Deng W H. 2022. Exploring disentangled content information for face forgery detection//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 128-145 [DOI: 10.1007/978-3-031-19781-9_8]
- Liu H G, Li X D, Zhou W B, Chen Y F, He Y, Xue H, Zhang W M and Yu N H. 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 772-781 [DOI: 10.1109/CVPR46437.2021.00083]
- Luo Y C, Zhang Y, Yan J C and Liu W. 2021. Generalizing face forgery detection with high-frequency features//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 16312-16321 [DOI: 10.1109/CVPR46437.2021.01605]
- Miao C T, Tan Z C, Chu Q, Liu H, Hu H G and Yu N H. 2023. F²Trans: high-frequency fine-grained transformer for face forgery detection. IEEE Transactions on Information Forensics and Security, 18: 1039-1051 [DOI: 10.1109/TIFS.2022.3233774]
- Ni Y S, Meng D P, Yu C Q, Quan C B, Ren D C and Zhao Y J. 2022. CORE: consistent representation learning for face forgery detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12-21 [DOI: 10.1109/CVPRW56347.2022.00011]
- Nirkin Y, Wolf L, Keller Y and Hassner T. 2022. DeepFake detection based on discrepancies between faces and their context. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10): 6111-6121 [DOI: 10.1109/TPAMI.2021.3093446]
- Qian Y Y, Yin G J, Sheng L, Chen Z X and Shao J. 2020. Thinking in frequency: face forgery detection by mining frequency-aware clues// Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 86-103 [DOI: 10.1007/978-3-030-58610-2_6]
- Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J and Niessner M. 2019. FaceForensics++: learning to detect manipulated facial images//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1-11 [DOI: 10.1109/ICCV.2019.00009]
- Shiohara K and Yamasaki T. 2022. Detecting deepfakes with self-blended images//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18699-18708 [DOI: 10.1109/CVPR52688.2022.01816]
- Thies J, Zollhöfer M and Nießner M. 2019. Deferred neural rendering: image synthesis using neural textures. ACM Transactions on Graphics, 38(4): #66 [DOI: 10.1145/3306346.3323035]
- Thies J, Zollhöfer M, Stamminger M, Theobalt C and Niessner M. 2016. Face2Face: real-time face capture and reenactment of RGB videos//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2387-2395 [DOI: 10.1109/CVPR.2016.262]
- Tora M. 2019. Deepfakes [EB/OL]. [2023-03-09]. <https://github.com/deepfakes/faceswap>
- Wang C R and Deng W H. 2021. Representative forgery mining for fake face detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14918-14927 [DOI: 10.1109/CVPR46437.2021.01468]
- Wang J, Sun Y L and Tang J H. 2022a. LiSiam: localization invariance siamese network for deepfake detection. IEEE Transactions on Information Forensics and Security, 17: 2425-2436 [DOI: 10.1109/TIFS.2022.3186803]
- Wang J K, Wu Z X, Ouyang W H, Han X T, Chen J J, Jiang Y G and Li S N. 2022c. M2TR: multi-modal multi-scale transformers for deepfake detection//Proceedings of 2022 International Conference on Multimedia Retrieval. Newark, USA: ACM: 615-623 [DOI: 10.1145/3512527.3531415]
- Wang Z, Guo Y W and Zuo W M. 2022b. Deepfake forensics via an adversarial game. IEEE Transactions on Image Processing, 31: 3541-3552 [DOI: 10.1109/TIP.2022.3172845]
- Yang J C, Li A Y, Xiao S, Lu W and Gao X B. 2021. MTD-Net: learning to detect deepfakes images by multi-scale texture difference. IEEE Transactions on Information Forensics and Security, 16: 4234-4245 [DOI: 10.1109/TIFS.2021.3102487]
- Zhang B G, Li S, Feng G R, Qian Z X and Zhang X P. 2022. Patch dif-

fusion: a general module for face manipulation detection. Proceedings of the AAAI Conference on Artificial Intelligence, 36 (3) : 3243-3251 [DOI: 10.1609/aaai.v36i3.20233]

Zhao H Q, Wei T Y, Zhou W B, Zhang W M, Chen D D and Yu N H. 2021b. Multi-attentional deepfake detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2185-2194 [DOI: 10.1109/CVPR46437.2021.00222]

Zhao T C, Xu X, Xu M Z, Ding H, Xiong Y J and Xia W. 2021a. Learning self-consistency for deepfake detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Montreal, Canada: IEEE: 15003-15013 [DOI: 10.1109/ICCV48922.2021.01475]

Zhou P, Han X T, Morariu V I and Davis L S. 2017. Two-stream neural networks for tampered face detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Work-

shops. Honolulu, USA: IEEE: 1831-1839 [DOI: 10.1109/CVPRW.2017.229]

作者简介

冯才博,男,本科生,主要研究方向为人脸伪造检测、人脸匿名化和生成式神经网络。

E-mail:2012190104@pop.zjgsu.edu.cn

刘春晓,通信作者,男,副教授,主要研究方向为深度学习与计算机视觉、模式识别与智能系统。

E-mail:cxliu@mail.zjgsu.edu.cn

王昱烨,男,本科生,主要研究方向为计算机视觉和智能交通系统。E-mail:2012190120@pop.zjgsu.edu.cn

周其当,男,本科生,主要研究方向为人脸活体检测。

E-mail:2011080121@pop.zjgsu.edu.cn