

基于双层注意力的 Deepfake 换脸检测

龚晓娟^{1,2}, 黄添强^{1,2}, 翁彬^{1,2}, 叶锋^{1,2}, 徐超^{1,2}, 游立军³

(1. 福建师范大学数学与信息学院, 福建 福州 350117;

2. 数字福建大数据安全技术研究所, 福建 福州 350117;

3. 福建省灾害天气重点实验室, 福建 福州 350001)

摘要: 针对现有 Deepfake 检测算法中普遍存在的准确率低、可解释性差等问题, 提出融合双层注意力的神经网络模型, 该模型利用通道注意力捕获假脸的异常特征, 并结合空间注意力聚焦异常特征的位置, 充分学习假脸异常部分的上下文语义信息, 从而提升换脸检测的有效性和准确性。并以热力图的形式有效地展示了真假脸的决策区域, 使换脸检测模型具备一定程度的解释性。在 FaceForensics++ 开源数据集上的实验表明, 所提方法的检测精度优于 MesoInception、Capsule-Forensics 和 XceptionNet 检测方法。

关键词: Deepfake; 换脸检测; 假脸检测; 注意力

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2021032

Deepfake swapped face detection based on double attention

GONG Xiaojuan^{1,2}, HUANG Tianqiang^{1,2}, WENG Bin^{1,2}, YE Feng^{1,2}, XU Chao^{1,2}, YOU Lijun³

1. College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350117, China

2. Digital Fujian Institute of Big Data Security Technology, Fuzhou 350117, China

3. Fujian Key Laboratory of Severe Weather, Fuzhou 350001, China

Abstract: In view of the existing Deepfake detection algorithms, such problems as low accuracy and poor interpretability are common. A neural network model combining the double attention was proposed, which used channel attention to capture the abnormal features of false faces and combined the location of spatial attention to focus the abnormal features. To fully learn the contextual semantic information of the abnormal part of the false face, so as to improve the effectiveness and accuracy of face changing detection. In addition, the decision-making area of real and fake faces was shown effectively in the form of thermal diagram, which provided a certain degree of

收稿日期: 2020-06-19; **修回日期:** 2020-09-29

通信作者: 黄添强, fjhtq@fjnu.edu.cn

基金项目: 国家重点研发计划专项基金 (2018YFC1505805); 国家自然科学基金 (62072106, 61070062); 应用数学福建省高校重点实验室 (莆田学院) 开放课题 (SX201803)

Foundation Items: The National Key Program for Developing Basic Science (2018YFC1505805), The National Natural Science Foundation of China (62072106, 61070062), Key Laboratory of Applied Mathematics of Fujian Province University (Putian University) (SX201803)

论文引用格式: 龚晓娟, 黄添强, 翁彬, 等. 基于双层注意力的 Deepfake 换脸检测[J]. 网络与信息安全学报, 2021, 7(2): 151-160.

GONG X J, HUANG T Q, WENG B, et al. Deepfake swapped face detection based on double attention[J]. Chinese Journal of Network and Information Security, 2021, 7(2): 151-160.

explanation for the face exchange detection model. Experiments on FaceForensics ++ open source data set show that the detection accuracy of proposed method is superior to MesoInception, Capsule-Forensics and XceptionNet.

Keywords: Deepfake, face swap detection, fake face detection, attention

1 引言

随着深度假脸 (Deepfake) 技术的兴起, 如 Face2face^[1]、FaceSwap、Deepfake、OpenFaceSwap 和 DeepFaceLab 等程序源代码可轻松获取。此外, 还有 FakeApp、Face2faceAI 换脸、ZAO 和 DeepNude (一键去衣) 等软件的大量涌现, 以及相关的软件和技术可以轻易地在一些技术网站上获得。这使人们不需要通过系统地学习专业技术就能轻而易举地对一幅图片甚至一段视频进行人脸篡改, 并且很多时候单凭人们的视觉直观感受难以分辨真伪。当新闻报道、政府组织或司法机构证据所使用的图片、视频被恶意篡改时, 势必会对政治和社会产生恶劣的影响。因此, 换脸检测研究吸引了大量研究团队的注意。例如, Facebook 和微软公司联合在 Kaggle 上推出了 Deepfake 检测挑战赛, 以促进检测算法的开发。

换脸技术制作的假脸比较如图 1 所示 (图片取自 FaceForensics++^[2]数据集)。其中以 Deepfake 和 FaceSwap 为代表的换脸技术, 可以实现图像或视频上交换两个人的面孔并保留其他身体部分。非技术人员只需收集足够多样化的素材, 就能得到逼真的换脸图片或视频; 且在未改变目标人物身份的前提下, 只采用 Face2face 技术就能对视频中的人物表情进行篡改。相对于传统的图像、视频篡改技术 (如复制粘贴篡改、增删篡改等), 应用上述与深度学习方法结合的新型换脸技术制作出的假脸视频 (文中出现的“假脸”皆为换脸技术生成的), 可能会对个人名誉和公众造成更恶劣的影响。例如, 制作虚假的名人色情视频和报复性的色情视频、篡改新闻内容, 以及伪造司法证据等。因此, 急需开发一种有效的换脸检测方法来应对换脸技术带来的危机。

随着换脸技术的不断发展, 换脸检测的研究也越来越受到学术界和工业界的重视。基于人工特征的方法^[3]主要通过检测伪影特征来对真脸假脸进行分类, 可解释性强但检测准确率普遍较低; 基于神经网络的检测方法^[4]虽然对假脸判别性能

较好, 但缺乏足够的可解释性, 其中文献[5]虽利用主成分分析 (PCA, principal component analysis) 对假脸中高频异常的可视化具有相应解释性, 但 PCA 作为一种静态的注意力机制 (AM, attention mechanism) 不能充分学习各类型的假脸特征。于是, 本文引入动态注意力机制, 可自适应地学习更丰富的假脸特征。

Original Deepfake FaceSwap Face2face

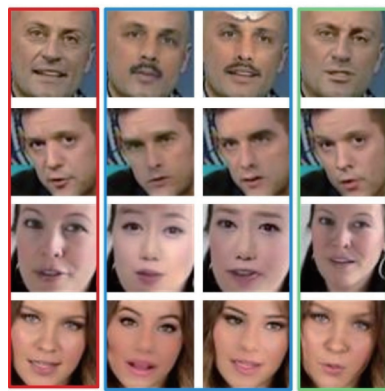


图 1 换脸技术制作的假脸比较
Figure 1 The comparison of fake face generated by manipulation techniques

注意力机制是一种快速锁定目标任务关键特征的方法, 在大多数计算机视觉任务中得到了应用^[6]。而在换脸检测中既要考虑假脸中的异常特征又要考虑异常特征的位置, 因此本文引入双层注意力, 以假脸的异常信息为关键特征, 以异常特征的位置为关键区域, 并将以上信息作为检测模型的决策依据。本文提出的融合双层注意力的神经网络 (DANet, double attention network) 模型, 能够学习假脸异常部分的上下文语义信息, 从而提升换脸检测的有效性和准确性。

神经网络是一种“黑匣子”技术, 仅由理论说明其原理缺乏可信度, 而通过可视化来展示决策依据才能使网络模型更具可解释性。虽有部分换脸检测技术^[5,7]做了可视化方面的尝试, 但它们仍存在一些不足之处, 其中文献[5]易受 ground-truth 的影响, 文献[7]不能很好地展示真假脸的决策区别。因此, 本文引入梯度加权类激活映射 (Grad-CAM, gradient-weighted class

activation mapping)^[8]技术,其不受 ground-truth 的影响,能够以热力图的形式更有效地展示真假脸的决策区域。从可视化角度,更有效地证实了 DANet 的有效性。

本文提出基于双层注意力网络的换脸检测模型,主要贡献如下。

(1) 将双层注意力引入换脸检测领域中,其中通道注意力关注假脸的异常特征,空间注意力关注假脸异常特征的位置,通过自适应地学习综合决策区域判别真假脸,有效地提升 DANet 的检测性能。

(2) 引入 Grad-CAM 对真脸和假脸赋以不同的权重映射生成热力图,为换脸检测模型的决策区域提供了充足的可视化依据,增强了 DANet 的可解释性。

2 相关工作

2.1 换脸检测方法

现有的换脸检测方法可大致分为两类:基于人工特征的方法与基于神经网络的方法。

基于人工特征的方法往往利用换脸视频存在的一些较明显的瑕疵。Li 等^[3]基于视频中人物是否有正常的眨眼行为来检测 Deepfake 视频。而 Yang 等^[9]通过脸部的 68 个特征点来检测假脸视频中头部姿势的不一致性。Matern 等^[10]结合不同的伪影特征对 GAN、Deepfake 和 Face2Face 这 3 种假脸图像进行检测。Nataraj 等^[11]在 RGB 这 3 个颜色通道上计算共生矩阵,然后将其输入 DNN 中分类真假脸。随着换脸合成技术的日益成熟,上述依赖人工提取伪影特征的检测方法的有效性逐渐被削弱。

基于神经网络的方法,由于其具有强大的深度特征表示学习能力得到了广泛关注。Cozzolino 等^[4]通过孪生网络定位图像篡改区域。Li 等^[12]利用 VGG^[13]网络和 ResNet^[14]等神经网络来捕捉换脸视频中面部周围环境分辨率不一致的伪影。Guera 等^[15]结合 CNN 网络和 LSTM^[16]网络来判别真假脸视频。Amerini 等^[17]将光流矢量输入 CNN 网络中进行深度假脸视频的判别。Afchar 等^[18]提出了两种浅层网络 Meso-4 和 MesoInception-4 来分类深度假脸。Fernandes 等^[19]设计了 Neural-ODE

模型,其通过对视频中目标人物的心率进行预测判别真假脸视频。Sabir 等^[20]提出了一种递归卷积模型检测假脸视频帧间的时域差异。Nguyen 等^[21]设计了一种多任务学习方法,该方法可以同时检测篡改图片和视频并且定位篡改区域。同年,Nguyen 等^[7]用 VGG-19 提取图片特征输入胶囊网络来检测篡改的图片和视频。Rössler 等^[2]创建了大规模的换脸数据集 FaceForensics++,并用 7 种网络对真假脸进行二分类性能比较,其中 XceptionNet^[22]的效果最佳。这些方法均能在一定程度辨别真假脸,但其网络模型仅考虑了异常特征,而没有借助注意力机制这一强有力的工具,所以缺少了对异常特征相关区域的探索,这是导致检测精度不够高的一个关键因素。

Dong 等^[5]通过外观处理模块和直接回归模块生成特征图的注意力映射,捕获假脸高频信息中的异常,但它使用的 PCA 注意力机制不具备可学习的参数,无法自适应地进行参数调整,来应对多样化的换脸技术。

因此,本文将可动态学习的注意力融合到换脸检测的模型中,重点关注和捕获假脸的异常特征和异常区域,减少决策无关区域的影响,从而提高了检测精度,并为假脸热力图的异常区域展示提供了一定依据。

2.2 换脸检测的注意力机制

注意力机制已经成功应用于目标检测^[6]、显著性检测^[23]、图像修复^[24]、语义分割^[25]和图像分类^[26-27]等领域。Zhou 等^[6]采用一种自注意力机制使检测模型尽可能只关注行人的移动区域,弱化对背景的关注。Li 等^[23]将运动注意力用于检测视频光流特征中的显著性运动物体。Liu 等^[24]在图像修复中使用空间注意力生成图像缺失区域的特征图。Fu 等^[25]通过位置和空间自注意力机制融合全局特征,提高语义分割的精度。Hu 等^[26]通过压缩与激活模块(squeeze-and-excitation module)对分类模型加强了通道特征的关联性。在此基础上,Woo 等^[27]利用双层注意力进一步增强了图像分类时重要特征的表达能力。上述这些工作表明注意力机制确实是一个有力工具。然而,将注意力机制用于换脸检测领域的工作较少,仅文献[5]采用了 PCA 这种不具备可学习参数的注意力机

制。而换脸检测中不仅要关注假脸中的异常特征,还需结合假脸异常特征的相关区域,利用可动态学习的注意力机制完成异常特征及相关区域的自适应捕获,才能对鉴别真假脸实现更有力的综合决策。因此,本文将双层注意力机制引入换脸检测中,分别用通道注意力机制和空间注意力捕获假脸的异常特征及异常特征的相关位置,与未融合双层注意力的检测方法相比,本方法有效提高了检测精度。

2.3 换脸检测的可视化

在神经网络中,特征图的可视化有助于增强网络模型的可解释性,便于人们将其更好地与实际应用结合。在现有换脸检测方法的可视化中,文献[7]展示了胶囊网络对脸部决策区域的激活,但真假脸决策区域的可视化效果相差不大,同时验证了该方法的检测精度有限。文献[5]的 ground-truth 在注意力图中显示了假脸的高频异常部分,但忽略了对真脸的关注。而本文无须借助 ground-truth,引入了 Grad_CAM^[8]生成热力图,可以较好地展示真假脸决策区域的不同:在无异常区域的真脸中寻找应该正常关注的区域,即真脸的热力图在大部分的细节区域具有较强的激活;在假脸中寻找缺失细节信息的异常区域,即假脸的热力图仅在异

常区域具有较强的激活。通过真假脸热力图的展示,证明了 DANet 可以捕获到有效的决策区域,同时可作为提高检测精度的依据,这是由于真假脸二者的决策区域不同,能够实现真类和假类的区分。

简而言之,本文将可动态学习的双层注意力融合到 DANet 中,自适应地捕获假脸异常特征和区域,并通过引入的热力图技术,为证实 DANet 检测的有效性提供了有力的可视化凭证。

3 DANet 检测算法

本文提出的融合双层注意力机制的神经网络模型的检测算法流程如图2所示。该算法先用 dlib 函数提取待检测视频帧的人脸,经筛选后将其输送到 DANet 模型中,进行真假脸的二分类训练。DANet 由 ResNet-50^[14]和双层注意力模块组成,双层注意力模块包括通道注意力和空间注意力两个子模块,其中通道注意力子模块关注假脸的异常特征,而空间注意力子模块关注假脸异常特征的位置,通过通道和空间两个维度的信息,可以使有意义的特征更具有表现力。本文算法的具体细节由以下两个部分展开:人脸提取和 DANet 模型构建。

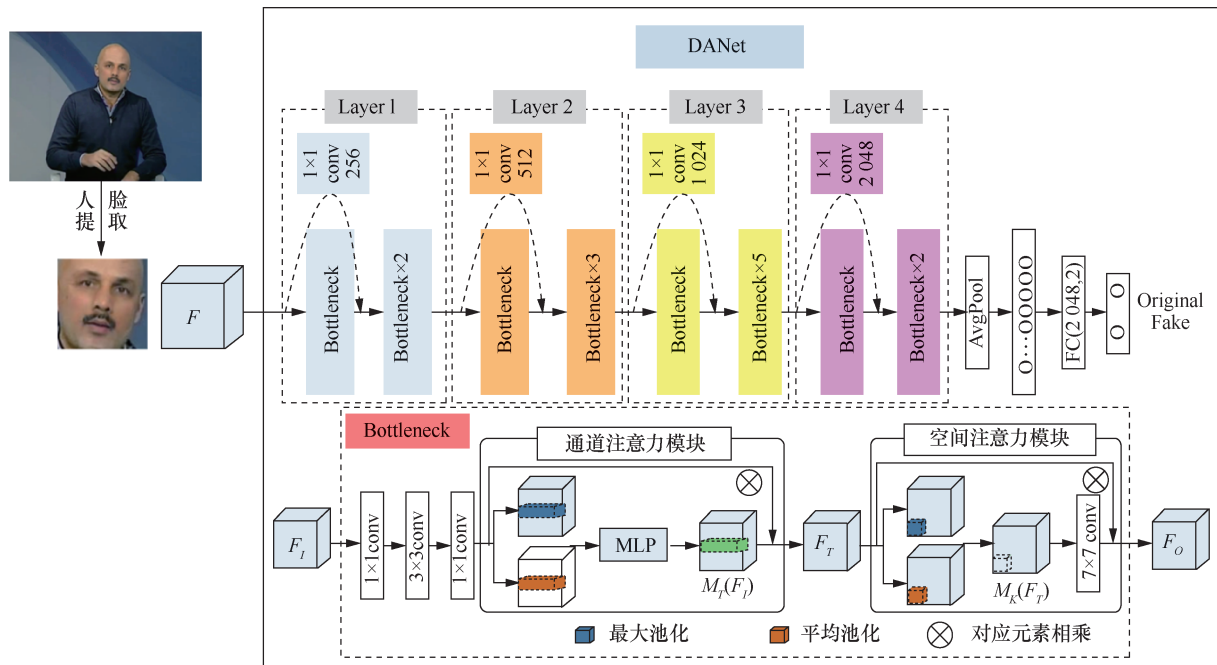


图2 DANet 模型的检测算法流程
Figure 2 The detection algorithm flow of DANet model

3.1 人脸提取

由于假脸视频的篡改区域主要集中在面部，因此本文仅提取人脸区域，并将其输入网络模型中，减小实验的复杂度。在实验中发现，在 Python 检测人脸的函数中，face_detector 的 detecMultiScale 和 haarcascade_frontalface_alt 函数检测的虚警率和漏检率较高，dlib 的 get_frontal_detector 的检测效果最佳，因此本文选择 get_frontal_detector 来检测并截取人脸，这样能够减轻前期预处理的工作。如图 3 所示，用 dlib 提取人脸，返回的 4 个坐标能明确锁定人脸的大小区域。

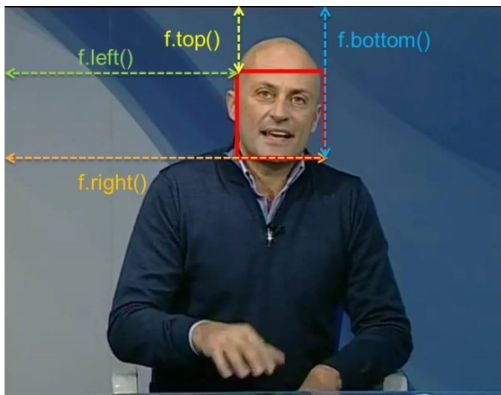


图 3 dlib.get_frontal_detector 函数得到的坐标示意
Figure 3 The coordinate diagram produced by the dlib.get_frontal_detector function

3.2 DANet 模型构建

在换脸检测模型中，不仅要考虑假脸异常特征的自适应捕获，还需结合异常特征的相关区域做进一步决策。于是将文献[27]的双层注意力模块引入假脸检测中，并做了两组对比实验，构建了 DANet 模型。其中两组实验分别为：主干网络的对比和注意力模块的摆放顺序对比。因 ResNet^[14]模型通过残差模块与上一层网络的输出更加紧密结合，以及假脸对前期特征捕获的重要性，故选取 ResNet-50 作为 DANet 的主干网络。先学习假脸的异常特征可以使模型检测的目标性更明确，故先放置空间注意力模块，再放置通道注意力模块。

DANet 模型的构建如图 2 所示，将提取到的人脸输入 DANet 模型中进行二分类训练。主干网络 ResNet-50 由 4 个 Layer 组成，每个 Layer 由不同个数的 Bottleneck 模块组成，其中每个 Layer 的第一个 Bottleneck 模块使用一次残差连接。在

每个 Bottleneck 模块添加双层注意力模块，分别是通道注意力模块和空间注意力模块。假设输入 Bottleneck 模块的特征图为 $F_I \in \mathbb{R}^{C \times H \times W}$ ， C 、 H 、 W 分别是特征图的通道数、高和宽，经通道注意力模块后得到特征块 $M_I(F_I)$ ，与 F_I 逐像素融合得到通道特征图 $F_T \in \mathbb{R}^{C \times 1 \times 1}$ ，经空间注意力模块得到特征块 $M_K(F_T)$ ，与通道特征图 $F_T \in \mathbb{R}^{C \times H \times W}$ 逐像素融合得到空间特征图 F_O ，具体计算如式(1)、(2)所示。

$$F_T = M_I(F_I) \otimes F_I \quad (1)$$

$$F_O = M_K(F_T) \otimes F_T \quad (2)$$

通道注意力模块由平均池化、最大池化和多层感知机构成，其中平均池化取得特征图 F_I 空间维度上的平均池化特征，最大池化取得最大池化特征，发送到同一个多层感知机，最后用 Sigmoid 函数激活，具体计算如式(3)所示。

$$M_I = \sigma(\text{MLP}(\text{AP}(F_I)) + \text{MLP}(\text{MP}(F_I))) \quad (3)$$

其中，AP 指的是平均池化，MP 指的是最大池化，MLP 指的是多层感知机， σ 是 Sigmoid 函数。

空间注意力模块由平均池化、最大池化和一个 7×7 的卷积块构成，分别取得 F_T 通道维度上的平均池化和最大池化特征，通过连接函数 Cat 沿纵向通道连接，然后经 7×7 卷积操作后再用 Sigmoid 函数激活，具体计算如式(4)、式(5)所示。

$$\text{Cat}_{\text{out}} = \text{Cat}([\text{AP}(F_T), \text{MP}(F_T)], \text{dim}=1) \quad (4)$$

$$M_K(F_T) = \sigma(\text{Conv}7 \times 7(\text{Cat}_{\text{out}})) \quad (5)$$

其中， $\text{dim}=1$ 指的是纵向通道。

通过上述双层注意力模块先后聚合空间信息和通道信息，可增强有效特征在空间维度和通道维度上的表现力，融合双层注意力模块的 DANet 更能够聚焦假脸图像中的异常特征，及其相关位置的综合决策区域，从而提升假脸的检测性能。

4 实验分析

本节主要介绍了实验环境、参数设置和数据集，以及分析模型的选择、可视化，并与现有换脸检测方法做对比。

4.1 实验环境和参数设置

本文的实验运行环境是 Ubuntu 16.04 系统，使用 4 块 Nvidia Tesla P100 GPU 进行模型的训练。本文所有深度学习模型均在 Pytorch 1.3.0

(2019.10.12 发布) 版本的深度学习框架的基础上实现, 并将 FaceForensics++ 数据集^[2]的每种篡改类型分为训练集、验证集和测试集, 其中, 图片是随机选取并完全独立的, 测试图像从未在训练过程中出现过, 数据集具体配置如表 1 所示, 真脸和假脸的比例为 1:1, 3 种篡改类型的训练集为 72 000 幅图片、验证集为 1 400 幅图片和测试集 1 400 幅图片。输入图片为 224×224 大小, 训练时的 shuffle 变量设置为 True, 迭代 100 个 epoch, 在每个 epoch 开始时对数据进行重新排序, 避免单个 epoch 取得最优准确率的偶然性。

表 1 基于 FaceForensics++ 数据集的训练集、验证集和测试集的分配

Table 1 Configuration of training, validation and test sets based on FaceForensics++ data sets

FaceForensics++ 数据集	Deepfake	FaceSwap	Face2face
训练集	Original	72 000	72 000
	Fake	72 000	72 000
验证集	Original	1 400	1 400
	Fake	1 400	1 400
测试集	Original	1 400	1 400
	Fake	1 400	1 400

通过平行实验比较, 保留效果最优的模型和参数。由于实验的训练集较大, 设置较小的 batch (如 16、32) 不仅会拉长网络模型的训练时间, 而且不利于模型的收敛; 而在同样的 epoch 下, 较大的 batch (如 128) 将会减少模型的迭代次数, 降低对假脸的检测性能。同时本文在实验中考虑了优化器的选择: 发现 Adam 优化器在 DANet 中的收敛速度快, loss 值会骤降到一个极小值, 并停在模型的局部最优解处, 后续的迭代无法进行有效学习, 这就导致在测试集上的泛化性能降低; 而使用 RMSprop 优化器的收敛速度较稳定, 且能够有效解决模型的局部最优解等极端情况。因此本文每组实验的 batch 取 64, 采用 RMSprop 优化器, 损失函数设为预测结果和真实数据的交叉熵函数 (CrossEntropyLoss)。此外, 训练过程和测试过程是分开的。

4.2 数据集和指标

实验中使用 FaceForensics++ 数据集验证本文的检测算法, 该数据集是由德国慕尼黑工业大学 (TUM) 视觉计算组和意大利那不勒斯腓特烈二

世大学构建的一个大型数据集, 其中包括从 Youtube 网站上筛选的大多以新闻播报、独家专访、单人脱口秀等为主题的仅包含单个人脸的 1 000 个原始视频, 以及在原始视频基础上, 分别用 Deepfake、Face2face 和 FaceSwap 这 3 种篡改技术进行篡改的各 1 000 个视频, 并使用 H.264 编码器分别合成无压缩 (C0)、轻度压缩 (C23) 和重度压缩 (C40) 3 种不同压缩程度的假脸视频, 本实验中只选取无压缩的视频进行帧分解和后续操作。

为评价 DANet 模型检测算法在 3 种篡改类型检测中的性能, 本文使用两个指标 [准确率 (Accuracy) 和 AUC 值] 对实验结果进行分析。

1) 准确率是预测真脸和假脸准确的数量占所有测试的真假脸数据总量的比率, 其计算如式(6)所示。

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (6)$$

其中, TP 是真脸预测为 Original 类, TN 是假脸预测为 Fake 类, FP 是假脸预测为 Original 类, FN 是真脸预测为 Fake 类。

2) AUC 值是 ROC 曲线下各部分面积的和值, 可以比较直观地反映模型的分类性能, 该值越大表示分类效果越好。ROC 曲线的纵坐标为 TPR, 横坐标为 FPR, 二者的计算分别如(7)、式(8)所示。

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (8)$$

4.3 模型选择

本文分别以 XceptionNet^[2]、ResNet-18^[14]和 ResNet-50^[14]作为主干网络, 比较融合双层注意力前后的性能。融合双层注意力后的网络标注为 XceptionNet-DA, 采用 ResNet18、ResNet50 为主干网络, 并分别标注为 DANet-18 和 DANet-50 (文中提到的 DANet 为 DANet-50 模型)。从表 2 中可以发现, 与 3 种基线网络相比, 融合双层注意力后的网络模型均可有效增强假脸检测性能。其中 XceptionNet 的输入调整为 299×299 大小, 仅使用迁移学习即可表现出良好的检测性能, 也正由于 XceptionNet-DA 中的空洞卷积在节省大量参数的

同时,无法充分有效地学习同一通道上的特征图信息,所以融合双层注意力可提升的检测性能有限,导致略逊于 DANet-50。对于深度篡改的假脸,DANet-50 比 DANet-18 能学习到更深层次的假脸特征,因而 DANet-50 在 3 种篡改类型的准确率都在 98%以上,且都为最优值,尤其是在 FaceSwap 篡改类型中,DANet-50 的检测准确率为 98.86%,无论是在横向还是纵向的比较实验中性能均为最优。因此本文选取 DANet-50 为最终的检测模型 DANet。

表 2 不同主干网络的检测性能
Table 2 Detection performance of different backbone networks

网络	Deepfake	FaceSwap	Face2face
XceptionNet ^[2]	94.81%	92.76%	88.00%
XceptionNet-DA	97.96%	97.32%	96.36%
ResNet18 ^[14]	55.43%	57.29%	53.36%
DANet-18	96.04%	93.57%	96.07%
ResNet50 ^[14]	56.21%	63.89%	61.25%
DANet-50	98.61%	98.86%	98.11%

在 DANet 中,比较了双层注意力在网络模型中的摆放顺序,实验结果如表 3 所示。当先放置空间注意力,后放置通道注意力时,网络模型先关注假脸的异常位置,后关注假脸的异常特征,若先学习到的异常位置不够准确,那么后续的特征学习会有偏差,导致网络模型的检测性能下降;反之,当先学习假脸的异常特征,再引导网络去学习异常特征的位置,就能取得较好的检测性能,因为先找到“是什么”比“在哪里”的目标性更

明确。因此本文实验中双层注意力的摆放顺序为先放置通道注意力,后放置空间注意力。

表 3 双层注意力的放置对假脸检测的影响
Table 3 Influence of the placement of double attention on fake face detection

网络	先放置空间注意力 后放置通道注意力	先放置通道注意力 后放置空间注意力
Deepfake	98.57%	98.61%
FaceSwap	92.00%	98.86%
Face2face	84.14%	98.11%

4.4 Grad_CAM 可视化

为增强 DANet 分类任务中的可解释性,本文对训练后的模型进行 Grad_CA^[8]可视化,该技术可将图片中感兴趣的区域用热力图的方式表现出来,显著地展示网络模型的决策依据。基于换脸视频中真假脸的差异性,本文将其用于 DANet 中进行可视化展示,并给出了合理的解释性。

与真实视频中的人脸相比,换脸视频中的假脸会缺失细节信息(如会反光的额头、眼睛里的焦点、脸上的光影等),仅在人脸的一些局部区域引起检测网络的关注,并具有较大激活值,从而引导网络模型凭这些异常特征来区分真假脸。由于整个网络是经过训练进行二分类的,所以网络的每个 Layer 上的特征图都包含人脸真假部分的鉴别信息,显然最后一个 Layer 的特征图比前面 Layer 的特征图能收集到的决策信息多,因此本文对 DANet 的 Layer4 进行可视化。真脸与 3 种篡改类型的假脸的热力图比较如图 4 所示。

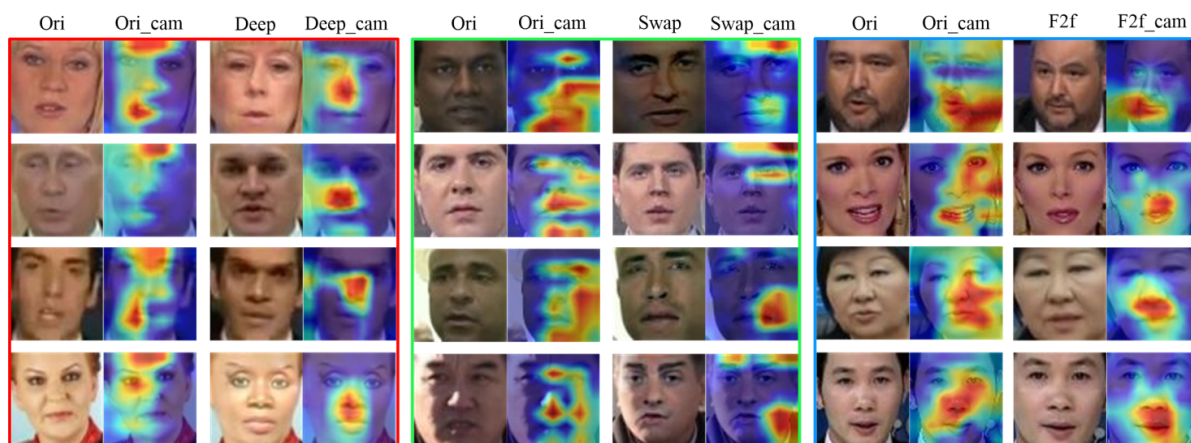


图 4 假脸与真脸的热力图比较
Figure 4 Comparison of heatmap between fake face and real face

表 4 图 4 的标注及含义
Table 4 Annotations and meanings of Figure 4

标注	含义	标注	含义
Ori	真脸	Ori_cam	真脸的热力图
Deep	Deepfake 假脸	Deep_cam	Deepfake 假脸的热力图
Swap	FaceSwap 假脸	Swap_cam	FaceSwap 假脸的热力图
F2f	Face2face 假脸	F2f_cam	Face2face 假脸的热力图

图 4 的标注及含义如表 4 所示。

通过观察 DANet 特征图生成的一系列热力图上,可以发现,真脸的热力图关注整个脸的大部分区域,假脸的热力图只关注脸的边缘或局部区域。正是由于真实的人脸中没有异常区域,那么 DANet 就会寻找真脸中值得关注的区域,如会反光的额头、有白点会聚焦的眼睛、面部的各种光线等,在热力图中分配较大权重(偏红色和黄色区域)。而在篡改的人脸上,有违真实感的区域比正常区域更能让 DANet 捕获到,如缺失的细节和拼接边缘等,只集中在局部和边缘区域。以 Face2face 的假脸为例,它是在真脸上更换表情和嘴型,其中以嘴巴边缘的异常尤为突出,所以在 F2f_cam 图上大部分集中在嘴巴区域。本文中的 Grad_CAM 可视化不是精确地定位假脸中的篡改区域,而是让训练后的网络模型在待测图像上对关注度高的区域赋以较高的权重作为特征图 ROI 的映射,生成 RGB 颜色空间下的梯度权重图(热力图)。本文定位的是网络模型分类判定的决策依据,并在待测图像上显示该区域。同理,在 Deep_cam 和 Swap_cam 图上的激活聚集在局部和边缘,也由此可以看出 DANet 模型可以有效捕获这些异常区域,从而将其作为判定假脸的决策依据。

4.5 实验结果

将本文方法与其他方法在 FaceForensics++^[2] 的 3 种篡改类型上的检测准确率进行比较,结果如表 5 所示。可以看到,本文的 DANet 模型在所有篡改类型上均优于其他 3 种方法。

Deepfake、FaceSwap 和 Face2face 3 种篡改类型的各网络模型的 AUC 面积分别如图 5~图 7 所示。可以看出,DANet 均优于其他模型的曲线,并且在 3 种假脸篡改类型上的检测精度是较稳定的,波动较小。综上,通过 4 个网络对 3 种篡

改类型在 Accuracy 和 AUC 的比较,DANet 表现最优。

表 5 各网络的检测精度对比
Table 5 Comparison of detection accuracy of each network

网络	Deepfake	FaceSwap	Face2face
MesoInception ^[18]	90.07%	90.89%	96.00%
Capsule-Forensics ^[7]	92.17%	92.79%	90.00%
XceptionNet ^[2]	94.81%	92.76%	88.00%
DANet	98.61%	98.86%	98.11%

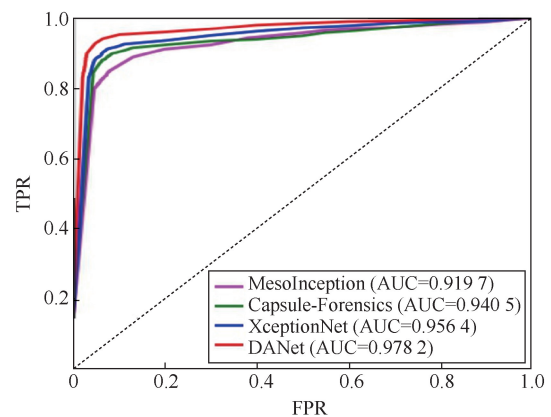


图 5 Deepfake 分类模型的比较
Figure 5 Comparison among Deepfake classification models

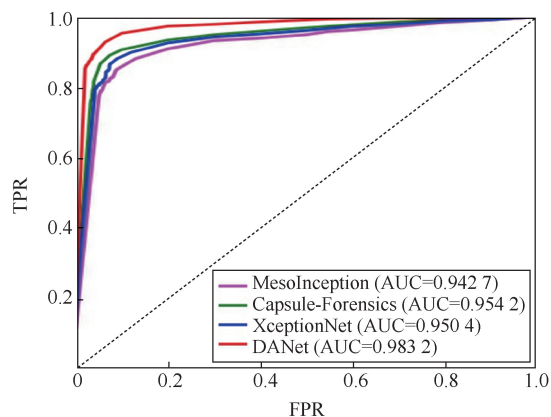


图 6 FaceSwap 分类模型的比较
Figure 6 Comparison among FaceSwap classification models

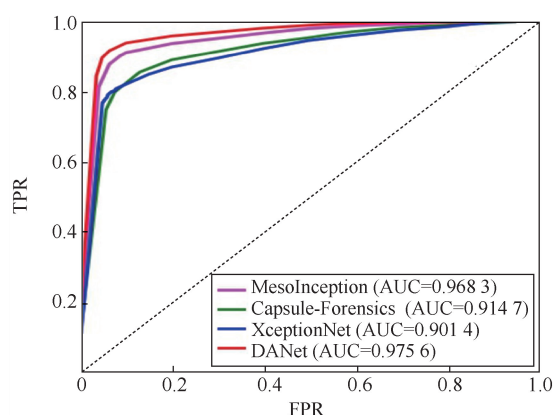


图7 Face2face 分类模型的比较

Figure 7 Comparison among Face2face classification models

5 结束语

为了应对 Deepfake 等换脸技术带来的危机, 本文提出基于双层注意力的换脸检测模型, 其中通道注意力关注假脸的异常特征, 空间注意力关注假脸异常特征的位置。在每个模块都融入了双层注意力, 使全局上下文的语义信息得到密切的关联, 从而在注意力机制的引导下提高了 DANet 的假脸检测性能。并利用 Grad-CAM 可视化技术生成热力图, 为网络模型的可解释性增加了有力依据。实验表明, 本文算法对 FaceForensics++ 数据集 3 种篡改类型的检测准确率均优于现有检测方法。

参考文献:

- [1] THIES J, ZOLLHOFER M, STAMMINGER M, et al. Demo of Face2Face: real-time face capture and reenactment of RGB videos[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016(1): 2387-2395.
- [2] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics++: learning to detect manipulated facial images[C]//International Conference on Computer Vision (ICCV). 2019: 1-11.
- [3] LI Y, CHANG M, FARID H, et al. In ictu oculi: exposing AI generated fake face videos by detecting eye blinking[C]//International Workshop on Information Forensics and Security (WIFS). 2018: 1-7.
- [4] COZZOLINO D, VERDOLIVA L. Noiseprint: a CNN-based camera model fingerprint[J]. IEEE Transactions on Information Forensics and Security, 2020(15): 144-159.
- [5] DONG H, LIU F, STEHOUWER J, et al. On the detection of digital face manipulation[J]. arXiv: 1910.01717[astro-ph.CO].
- [6] ZHOU C, WU M, LAM S. SSA-CNN: semantic self-attention CNN for pedestrian detection[J]. arXiv: 1902.09080v1[astro-ph.CO].
- [7] NGUYEN H, YAMAGISHI J, ECHIZEN I. Use of a capsule network to detect fake images and videos[J]. arXiv: 1910.12467v2[astro-ph.CO].
- [8] SELVARAJU M, COGSWELL M, DAS A. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//International Conference on Computer Vision (ICCV). 2017: 618-626.
- [9] YANG X, LI Y, LYU S. Exposing deep fakes using inconsistent head poses[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019: 8261-8265.
- [10] MATERN F, RIESS C, STAMMINGER M. Exploiting visual artifacts to expose Deepfakes and face manipulations[C]//2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). 2019: 83-92.
- [11] NATARAJ L, MOHAMMED T M, MANJUNATH B S, et al. Detecting GAN generated fake images using co-occurrence matrices[C]//Media Watermarking, Security, and Forensics. 2019.
- [12] LI Y, LYU S. Exposing Deepfake videos by detecting face warping artifacts[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 46-52.
- [13] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//3rd International Conference on Learning Representations (ICLR). 2015.
- [14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770-778.
- [15] GUERA D, DELP E. Deepfake video detection using recurrent neural networks[C]//15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2018: 1-6.
- [16] DONAHUE J, HENDRICKS L, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 2625-2634.
- [17] AMERINI I, GALTERI L, CALDELLI R, et al. Deepfake video detection through optical flow based CNN[C]//International Conference on Computer Vision Workshops (ICCVW). 2019: 1205-1207.
- [18] 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). 2018: 1-7.
- [19] FERNANDES S, RAJ S, ORTIZ E, et al. Predicting heart rate variations of Deepfake videos using neural ODE[C]//International Conference on Computer Vision Workshops (ICCVW). 2019: 1721-1729.
- [20] SABIR E, CHENG J, JAISWAL A, et al. Recurrent convolutional strategies for face manipulation detection in videos[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 80-87.
- [21] NGUYEN H, FUMING F, YAMAGISHI J, et al. Multitask learning for detecting and segmenting manipulated facial images and videos[J]. arXiv: abs/1906.06876[astro-ph.CO].
- [22] CHOLLET F. Xception: deep learning with depth wise separable convolutions[C]//2017 IEEE Conference on Computer Vision and

Pattern Recognition (CVPR). 2017: 1800-1807.

- [23] LI H, CHEN G, LI G, et al. Motion guided attention for video salient object detection[C]//International Conference on Computer Vision (ICCV). 2019: 7273-7282.
- [24] LIU H, JIANG B, XIAO Y, et al. Coherent semantic attention for image inpainting[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 4169-4178.
- [25] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 3146-3154.
- [26] HU J, SHEN LI, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 7132-7141.
- [27] WOO S, PARK J, LEE J, et al. CBAM: convolutional block attention module[C]//The European Conference on Computer Vision(ECCV). 2018: 3-19.

[作者简介]



龚晓娟（1995-），女，福建福州人，福建师范大学硕士生，主要研究方向为数字多媒体取证。



黄添强（1971-），男，福建仙游人，博士，福建师范大学教授、博士生导师，主要研究方向为机器学习、数字多媒体取证。



翁彬（1981-），男，福建福州人，博士，福建师范大学讲师，主要研究方向为机器学习及应用。



叶锋（1978-），男，福建福州人，博士，福建师范大学副教授，主要研究方向为多媒体信号处理、计算机视觉。



徐超（1981-），男，湖北天门人，福建师范大学讲师，主要研究方向为视频篡改检测。



游立军（1974-），男，福建莆田人，福建省灾害天气重点实验室高级工程师，主要研究方向为气候数据分析。