基于对抗训练的改进人脸伪造检测方法

张 凯[1],范智贤[2]
1. 安徽理工大学 计算机科学与工程学院,安徽 淮南 232001
2. 安徽理工大学 人工智能学院,安徽 淮南 232001

摘 要:目的 针对现有人脸伪造检测方法泛化性不足、鲁棒性较差的问题,提出了一种基于对抗训练的改进人脸伪造检测方法(ATNet)。**方法** ATNet 通过动态合成对抗伪造样本扩大样本空间,强化模型对不同伪造算法生成的伪造样本“敏感度”,避免训练样本过采样导致的模型过拟合问题;替换特定地标点标识的伪造区域以提高模型学习不同伪造特征的能力;采用高频噪声和低频纹理并行学习策略,融合多层次卷积特征,促使模型捕捉更具综合性的伪造线索,使其更有效的鉴别伪造图像。**结果** 随着有效的对抗样本生成,ATNet 可以学习到更为本质的伪造特征,相对于当前较为优秀方法如 Xception,F3Net,Face X-ray 等在检测精度和泛化性上均有不同程度地提升,跨数据集测试结果表明该模型具备优秀的泛化性能;使用降维算法进行可视化可以直观判断出 ATNet 对于深度伪造图像的检测能力。**结论** 基于改进对抗训练的人脸检测方法可以有效提升检测精度和泛化性,强化模型对于多种伪造特征的“敏感度”,在多个数据集上的实验结果表明 ATNet 是简单有效的。

关键词:深度伪造;人脸伪造检测;对抗训练;数据增强
中图分类号:TP391 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2025.0004.011

Improved Face Forgery Detection Method Based on Adversarial Training

ZHANG Kai[1], FAN Zhixian[2]
1. School of Computer Science and Engineering, Anhui University of Science and Technology, Anhui Huainan 232001, China
2. School of Artificial Intelligence, Anhui University of Science and Technology, Anhui Huainan 232001, China

Abstract: Objective Aiming at the problems of insufficient generalization and poor robustness of existing face forgery detection methods, an improved face forgery detection method based on adversarial training (ATNet) is proposed. **Methods** ATNet expands the sample space by dynamically synthesizing adversarial forgery samples to enhance the model's "sensitivity" to forgery samples generated by different forgery algorithms and avoid the over-fitting of the model caused by over-sampling of training samples. By replacing the forgery regions marked by specific landmark points, the model's ability to learn different forgery features is improved. A parallel learning strategy of high-frequency noise and low-frequency textures is adopted to fuse multi-level convolutional features, which enables the model to capture more comprehensive forgery clues and more effectively identify forged images. **Results** With the effective generation of adversarial samples, ATNet can learn more essential forgery features. Compared with current excellent methods such as Xception, F3Net, and Face X-ray, ATNet has varying degrees of improvement in detection accuracy and generalization. The cross-dataset test results show that the model has excellent generalization performance. Visualization using the dimensionality reduction algorithm can intuitively determine ATNet's ability to detect deep-forged images.

收稿日期:2023-10-26 **修回日期:**2023-11-29 **文章编号:**1672-058X(2025)04-0088-07
基金项目:安徽省自然科学基金(2008085MF220);安徽理工大学研究生创新基金项目(2021CX2102).
作者简介:张凯(1998—),男,安徽芜湖人,硕士研究生,从事计算机视觉研究.
通信作者:范智贤(1999—),女, 安徽阜阳人,硕士研究生,从事计算机视觉研究. Email: 1158998678@qq.com.
引用格式:张凯,范智贤. 基于对抗训练的改进人脸伪造检测方法[J]. 重庆工商大学学报(自然科学版),2025,42(4):88-94.
ZHANG Kai, FAN Zhixian. Improved face forgery detection method based on adversarial training[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(4): 88-94.

Conclusion The face forgery detection method based on improved adversarial training can effectively improve the detection accuracy and generalization, and strengthen the model' s "sensitivity" to various forgery features. The experimental results on multiple datasets show that ATNet is simple and effective.

Keywords: deepfakes; face forgery detection; adversarial training; data augmentation

1 引　言

随着以生成对抗网络(Generative Adversarial Networks, GANs)和变分自编码器(Variational Autoencoder, VAE)为代表的深度伪造技术的快速发展,目前已经可以生成肉眼无法识别真伪的高度逼真的人脸伪造图像。这些伪造技术在用于娱乐活动的同时,也被滥用于一些违法活动,造成了一定程度的社会"信任"危机,对现有自媒体技术发展和国家安全造成极大威胁。因此,研究并发展相应的人脸伪造检测技术迫在眉睫,已经成为计算机视觉的一个热门研究领域。

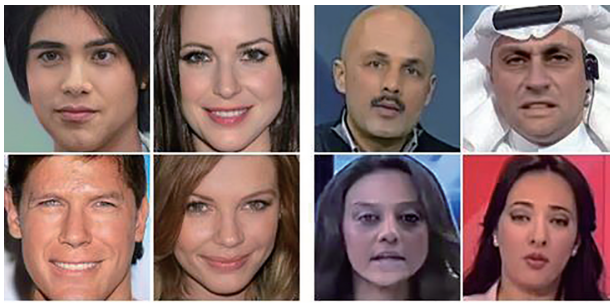
传统的图像取证方法大多依赖于特定的篡改信息,利用图像空域和频域统计特征进行区分,例如噪音分析、设备指纹和光照条件等,解决复制、移位、拼接等图像篡改技术,具备较强的可解释性,但也面临着操作难度大、技术要求高和检测结果不精确等问题。随着深度学习的快速发展,基于卷积神经网络(Convolutional Neural Networks, CNNs)的检测技术依赖其强大的特征学习能力,在诸多方面的表现都超过了传统图像取证技术。Qian 等[1]提出一种利用图像频域信息的人脸伪造检测方法,利用图像的频域统计特征有效地提高了模型的检测精度,鉴别高度压缩的伪造图像时尤为显著。但是该方法对数据的敏感性较高,针对不同压缩率的伪造图像需要单独训练才能确保模型的准确率,限制了其广泛的应用。Li 等[2]利用图像的频域特征同时建模一个单中心辅助损失,拉大类间间距,缩小类内间距,提高了模型的库内检测精度,但缺点是跨库检测能力不足。Face X-ray[3]针对目前绝大多数的人脸篡改方法都具有的移动-拼接-融合等操作,提出检测图像是否有融合边框来鉴别是否为伪造图像,在自建数据集上的训练结果表明了该模型具有优越的泛化性能,但是容易受到对抗样本的干扰,如果在伪造图像上进行二次加噪,就很容易躲避其检测。SOLA[4]通过结合不同距离和方向局部二阶特征训练模型,有效地提高了泛化性,但准确率难以保证。IGFNet[5]通过改进的高斯滤波网络提高虚假视频的检测精度,但过滤了伪造图像的低频信息,未能充分利用伪造图像各层次频率的伪造特征。

上述检测方法大多针对特定检测问题提出相应的解决方法,例如 F3Net[1]、FDFL[2]着重于提高检测准确率,Face X-ray[3]、SOLA[4]着重于提升泛化性能,在各自的侧重点上均取得了不错的效果,但通常网络复杂度高,训练数据量大且训练时间长,在小型数据集上的检测结果差,并且只能保证单一指标上的提升,限制了在现实中的应用。在以上学者的工作基础上,通过动态生成对抗样本扩大样本空间,同时增加虚假图像的伪造种类,降低过拟合风险;替换特定地标点标识的区域增强网络对于特定伪造痕迹的敏感度;最后通过低频和高频特征的综合学习进一步提升深度伪造图像的检测精度。算法不仅可以在小型数据集上进行有效地训练,并且可以在准确率和泛化性之间都取得一个良好的效果。

2 相关技术

2.1 深度人脸伪造

当前的深度伪造技术主要基于 GANs 和 VAE 技术,通过学习训练集的数据分布来合成虚假人脸。目前最为先进的 GANs 模型可以合成肉眼几乎无法识别的伪造图像,但主要用于生成整张现实生活中不存在的人脸,危害性较小。而"换脸"技术是针对特定的源人脸和目标人脸进行编辑或重演等操作,其对象大多为现实中真实存在的,对社会稳定有着更大的危害性,因此具备更高的研究价值。常见的"换脸"技术主要以 VAE 为基准,由一个编码器和两个解码器构成。针对源人脸图像和目标人脸图像训练一个共享编码器,两个独立解码器。相应的编码器解码器组合可以用于相应的人脸替换,在"换脸"过程中交叉使用对应的解码器可以将目标人脸替换到源人脸上。图 1 展示了一些伪造人脸图像。



(a) 全脸合成　　(b) 人脸替换

图 1　伪造人脸图像

Fig. 1　Fake face images

2.2 深度伪造检测

针对不同的伪造方法,相应的检测方法也层出不穷,主要有基于传统图像取证、基于生理信号特征、基于图像篡改伪影和基于数据驱动等方法。其中基于传统图像取证和基于生理信号特征的方法大多只能用于低质量的虚假图像检测。由于早期伪造技术不够先进,伪造图像往往存在明显的肉眼可见的伪造痕迹,该类方法检测此类伪造图像可以取得不错的效果。但随

着伪造技术的不断发展,如今的伪造图像十分逼真,人眼几乎无法分辨真伪,鉴别此类伪造图像很难取得优秀的效果,因此基于篡改伪影和基于数据驱动的方法开始流行起来。此类检测方法利用深度学习算法,不仅操作简单,同时效果显著,各方面的测试指标都超过其他类型的检测方法,但缺点是可解释性不足,训练数据量大。

2.3 EfficientNet-B4 网络介绍

EfficientNet[6]是一个强大的神经网络模型,在现阶段常作为其他检测算法的骨干网络和对比基线。EfficientNet 通过优化网络结构来寻找一组最优的网络参数解,在保证准确率的情况下同时减少了网络的参数量,加快了训练过程。根据网络深度、宽度以及输入图像的分辨率该网络分为 B0-B7 八个版本,主要结构由多个 MBConv 模块堆叠构成,如图 2 所示。根据具体的训练任务、数据集大小和期望指标可以选择不同版本,具备高度可选择性。本文提出的模型以修改后的 EfficientNet-B4 为骨干网络。

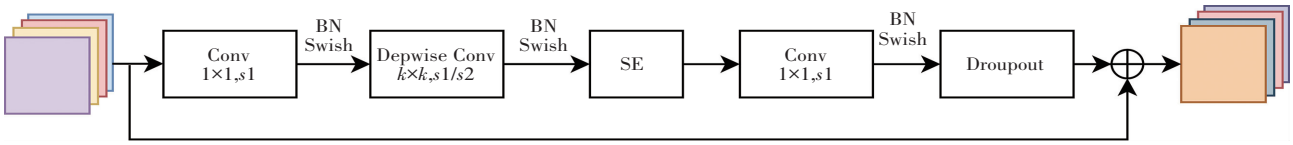


图 2 MBConv 模块

Fig. 2 MBConv module

3 系统设计与方法构建

3.1 整体框架

本文提出使用对抗数据增强的方式丰富伪造类型的种类,提高人脸检测模型的泛化性,其架构如图 3 所示。在训练期间,如果输入为真实图像,则随机选取另一张真实的目标图像,先输入到生成网络并根据特定的人脸区域合成相应的伪造图像。新合成的伪造图像也输入到鉴别器中。注意如果输入图像本身为伪造图像,则跳过伪造这一步骤。算法使用对抗训练策略训练网络,其中合成网络视为生成器,检测网络视为判别器。训练过程中将不断合成新的伪造样本以激励检测器识别更多的不同种类的伪造样本;训练结束后丢弃生成器只保留判别器用于伪造人脸检测。

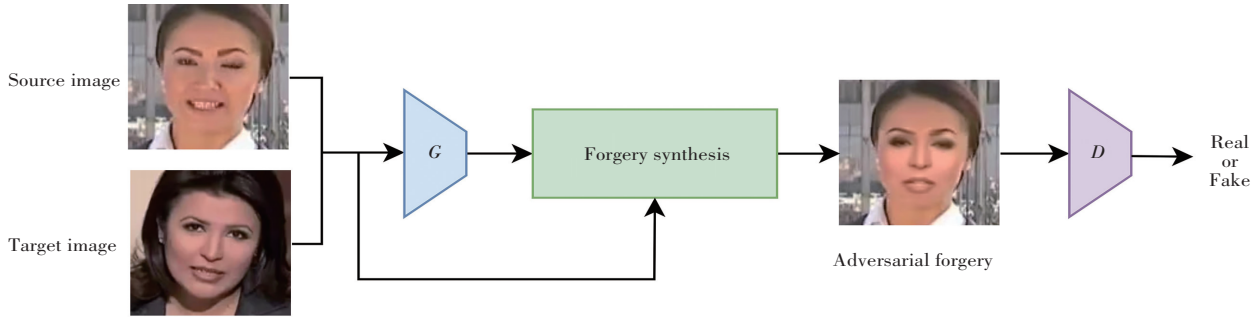


图 3 模型的总体框架

Fig. 3 Framework of the model

3.2 伪造方法

本文的生成网络 $G(\cdot,\theta)$ 接收一张源图像 $I_s \in \mathbf{R}^{H\times W\times 3}$,如图 4(a),一张目标图像 $I_t \in \mathbf{R}^{H\times W\times 3}$,如图 4(g)为输入,根据特定的伪造区域 R_g 合成伪造图像。已有的伪造算法大多只修改人脸图像的特定位置,如嘴巴,眼睛和鼻子等区域,为了与这些方法保持一致,算法使用 dlib 库提取一张人脸图像的 68 个地标点,如图 4(b)所示,以此将人脸区域划分成 10 个区域,包括左眼、右眼、鼻子、嘴巴及其组合。图 4(c)—图 4(e)展示了部分人脸区域的划分。在合成步骤之前,先使用随机大小的卷积核对伪造区域 R_g 进行高斯模糊获取最终的伪造掩码 M_d,之后根据伪造掩码从源图像 I_t 中裁取伪造区域,再融入源图像 I_s 中。最终的对抗生成图像 I_a 可表示为

$$I_a = \alpha \times M_d * (I_t - I_s) + I_s \tag{1}$$

式(1)中 α 为融合比例因子,$*$ 表示 Hadamard 乘积。

如同现有的深度伪造方法一样,在混合步骤之前对裁剪的人脸区域应用颜色转换和人脸对齐以避免新合成的伪造图像出现明显的伪影。图 4(h)展示了一张混合图像的例子。

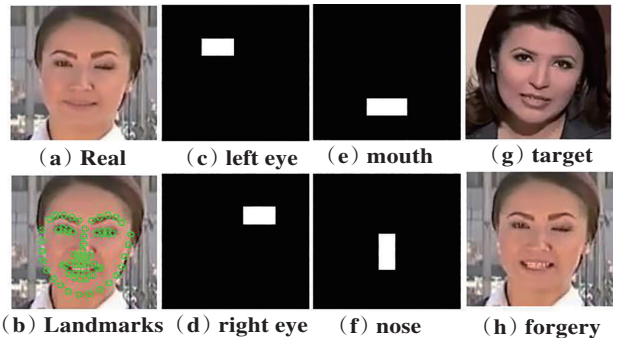


图 4 改进的模块

Fig. 4 Improved module

3.3　检测器改进

传统的神经网络只利用最后一层卷积结果作为分类器的输入,例如本文的骨干网络 EfficientNet-B4,这并不利于特征的鲁棒性表示。随着网络层数的加深,网络提取的特征信息更加高级、抽象,这损失了原始图像的低层语义信息,如图像纹理等特征。对于人脸伪造图像而言,低层的语义信息是重要的检测线索。当前的深度伪造技术为了减少明显的视觉伪影,往往会对图像进行一些后处理操作,如颜色变换和模糊等操作,这会留下更多的低层语义信息。为了避免这些重要信息损失,算法将所有 MBConv 模块池化后的卷积特征全部拼接,作为一个总体输入到全连接层中,如图 5 所示。改进后的 EfficientNet 网络可以同时学习到伪造图像的低层纹理信息和高级语义信息,提高检测能力。

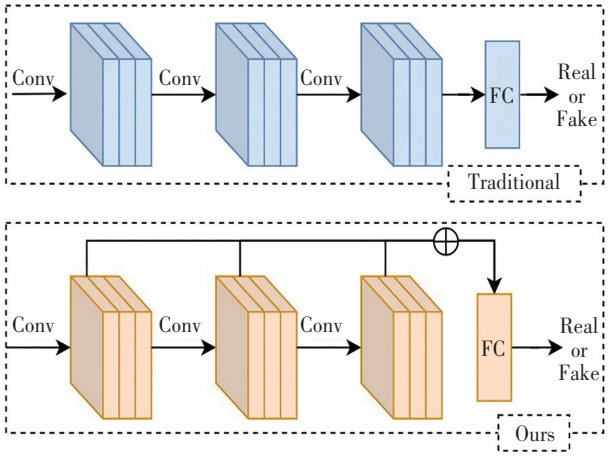


图 5　伪造合成算法

Fig. 5　Forgery synthesis algorithm

3.4　对抗训练

为了更好地扩大样本空间,增加伪造类型种类,算法采用对抗训练方式训练模型。具体来说,将合成网络 $G(\cdot,\theta)$ 视为生成器,检测网络 $D(\cdot,w)$ 视为判别器,生成器最大化训练损失,判别器最小化训练损失。优化过程可表示为

$$\min_{w}\max_{\theta} L(\theta,w)=L_{ce} \tag{2}$$

式(2)中,L_{ce} 表示交叉熵损失。按照通常的对抗训练方式,可以通过迭代更新判别器和生成器解决上述优化问题。

在判别器学习过程中,通过固定生成器参数 θ,根据学习率 η 和批量 N 按如下公式进行梯度下降更新:

$$w^{t+1}=w^{t}-\eta\frac{1}{N}\sum_{n=1}^{N}\nabla_{w^t}L_n(\theta^t,w^t) \tag{3}$$

式(3)中,L_n 表示各个批量里第 n 个样本的损失。

生成器目标是通过合成难以识别的伪造样本增大训练损失,以此促进判别器学习更加泛化性的伪造特征。对抗训练是一场零和博弈的游戏,可以根据式(4)来描述这一最大化过程:

$$\theta^{t+1}=\arg\max_{\theta^t}L(w^{t+1},\theta^t) \tag{4}$$

然而,直接通过式(4)更新 θ 可能会存在问题。因为存在不可微分的采样操作从而破坏 $D(\cdot,w)$ 到 $G(\cdot,\theta)$ 的梯度流。为了解决这个问题,使用 REINFORCE[7] 算法来近似计算 θ 的梯度。该算法使用蒙特卡洛方法估计每个状态下所获得的奖励期望值,之后利用该估计值计算梯度并更新:

$$\theta^{t+1}=\theta^{t}+\varepsilon\nabla_{\theta^t}L_b\approx\theta^{t}+\varepsilon\frac{1}{M}\sum_{m=1}^{M}L_b\nabla_{\theta^t}\log p_m \tag{5}$$

式(5)中,$L_b=\frac{1}{N}\sum_{n=1}^{N}L_n(w^{t+1},\theta^t)$,$M$ 表示伪造过程中替换的人脸区域数目,p_m 表示生成 Rg 的概率。

4　仿真实验与结果分析

4.1　实验设置

(1) 数据集。使用 FaceForensics++(FF++)[8] 和 Celeb-DF[9] 两个数据集用于验证本文方法的有效性。FF++包含 1 000 个真实视频,每个视频使用 4 种伪造算法生成相应的伪造视频,即 DeepFakes(DF)[10]、FaceSwap(FS)[11]、Face2Face(F2F)[12] 和 NeuralTextures(NT)[13],共 4 000 个伪造视频。根据压缩率的不同,又分为原始(raw)、低压缩(c23)和高压缩(c40)三个版本,本文主要使用低压缩版本。遵循惯例其中 720 个视频用于训练,140 个用于验证,另外 140 个用于测试,使用 RetinaFace 识别并裁取人脸图像,每个真假视频帧数按 4∶1 比例截取以平衡其数量不匹配的问题。Celeb-DF 包含 590 个真实视频和 5 693 个伪造视频。本文中,该数据集用于跨数据集验证。具体的数据集介绍如表 1 所示。

表 1　基准数据集介绍

Table 1　Introduction of benchmark datasets

Dataset	Real		Fake		Manipulation
	Video	Frame	Video	Frame	Algorithm
FF++	1 000	509.9 k	4 000	1 830.1 k	DF, FS, F2F, NT
Celeb-DF	590	225.4 k	5 693	2 116 k	Improved DF

(2) 评估指标。与大部分人脸检测算法一致,算法使用准确率(P_{ACC})和 ROC 曲线下的面积(S_{AUC})作为评估指标。P_{ACC} 表示检测器预测正确的样本占总样本的比例,表示为

$$P_{ACC}=\frac{T_P+T_N}{T_P+T_N+F_P+F_N} \tag{6}$$

式(6)中 T_P、T_N、F_P 和 F_N 分别表示真正例、真反例、假正例和假反例。

(3) 实验细节。本文使用 Pytorch 框架实现 EfficientNet-B4,优化器为 Adam,初始学习率为 0.000 2,使用 dlib 库将图像裁剪为 320×320 大小并进行人脸对齐,批量大小为 32,共训练 30 轮,每训练 5 轮学习率衰减 0.1,不设置早停,网络初始参数使用在 ImageNet 数据集上预训练后的权值。

4.2 精度比较

方法(ATNet)在 FF++数据集上的定量分析结果如表 2 所示。Multi-Attention 通过提取图像细粒度特征在数据集内部评估中取得了最优的成绩,但在高压缩的伪造样本检测中,由于图像损失了大量细节信息,在这种情况下,该方法无法捕捉足够的细粒度特征,限制了其泛化能力的提升。方法(ATNet)的准确率虽然在 c23 数据集上略低于 Multi-Attention,在 c40 数据集上略低于 F3-Net,但其余的评估结果均是最优,尤其是 S_{AUC} 在两种压缩率的版本上达到了 99.42% 和 89.46%。这表明方法(ATNet)可以有效地学习虚假图像的伪造特征,提高图像检测的精度。

表 2 人脸伪造检测方法的定量比较,其中"—"表示空值

Table 2 Quantitative comparison of face forgery detection methods, where "—" indicates null

Methods	c23		c40	
	P_{ACC}	S_{AUC}	P_{ACC}	S_{AUC}
Xception[14]	95.73	—	81.00	—
EfficientNet-B4[6]	96.25	98.94	87.28	89.12
F3-Net[1]	97.52	—	90.43	—
Multi-Attention[15]	97.60	98.29	88.69	89.40
Face X-ray[3]	—	87.40	—	61.60
Ours	97.58	99.42	89.46	90.21

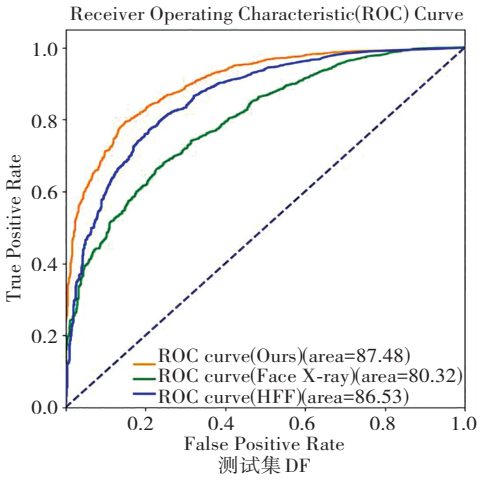
4.3 跨数据集评估

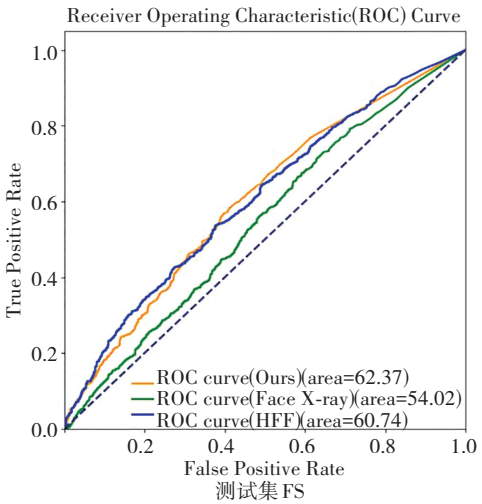
这一小节中,将重点放在更具挑战性的跨数据集评估上。为了与其他主流检测模型进行公平的比较,使用原作者提供的源码复现实验结果,并且所有模型的参数设置均与原文保持一致,结果如表 3 所示,部分 ROC 曲线如图 6 所示。可以看到,相对于部分目前最先进的检测算法,例如 HFF[16] 以及 Face X-ray[3],方法(ATNet)的 S_{AUC} 在多数评估结果中最高,综合性能最好。同时这说明动态合成的伪造样本有效地扩大了样本空间,促使网络学习到了更加泛化性的特征表示。

表 3 在 FF++(c23)上训练的跨数据集 S_{AUC} 评估

Table 3 Cross-dataset S_{AUC} evaluation trained on FF++(c23)

Training Set	Model	Testing Set			
		DF	F2F	FS	NT
DF	Face X-ray[3]	99.38	73.62	49.01	73.62
	HFF[16]	99.48	73.64	52.46	77.12
	Ours	99.41	76.64	58.82	81.92
F2F	Face X-ray[3]	80.32	99.42	54.02	69.48
	HFF[16]	86.53	99.21	60.74	72.84
	Ours	87.48	99.48	62.37	73.36
FS	Face X-ray[3]	66.47	67.67	99.47	53.21
	HFF[16]	71.26	73.92	99.48	56.71
	Ours	70.94	77.72	99.66	59.20
NT	Face X-ray[3]	81.42	66.87	48.58	99.17
	HFF[16]	88.46	73.26	51.67	99.42
	Ours	87.05	76.82	57.79	99.36





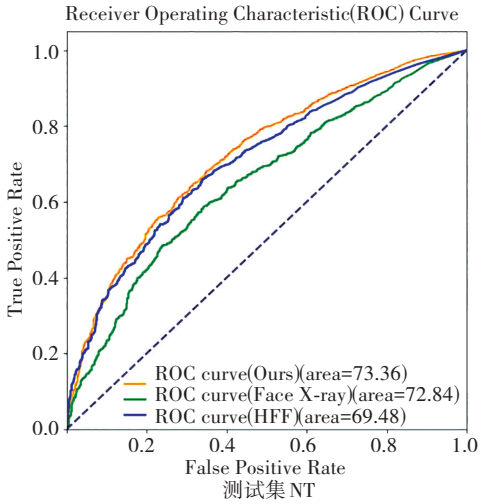


图 6　ROC 曲线,训练集为 F2F

Fig. 6　ROC curves, the training set: F2F

为更为全面地评估方法(ATNET),算法在 FF++数据集上进行训练并且在 Celeb-DF 数据集上测试,结果如表 4 所示。可以看到,方法(ATNet)不仅在数据集内部评估上 S_{AUC} 最好,到达了最高的 99.51%,同时在 Celeb-DF 数据集测试上也取得了媲美 ICT[17] 和 DCL[18] 的分数,平均提升率为 1.55%,相较于 Face X-ray[3] 更是提高了 5.15%。结合两种跨数据集评估结果,方法(ATNet)的泛化性均优于所比较的方法。

表 4　FF++(c23)和 Celeb-DF 跨数据集评估

Table 4　Cross-dataset evaluation on FF++(c23) and Celeb-DF

Methods	FF++	Celeb-DF
ICT[17]	98.56	78.26
Face X-ray[3]	98.52	74.76
Xception[14]	99.30	65.30
DCL[18]	96.60	78.46
LRNet[19]	97.30	56.90
SPSL[20]	96.91	76.88
Ours	99.51	79.91

4.4　消融实验

为评估对抗训练策略是否切实地提升了模型的泛化性,构建了以下几个变体进行消融实验:EfficientNet-B4:没有使用对抗训练方法的基线模型;EfficientNet-B4 w/adv:在 EfficientNet-B4 模型上添加了对抗数据增强策略,但只利用最后一层的卷积结果作为分类器的输入;EfficientNet-B4 w/ran:使用随机数据增强替换了对抗数据增强策略。比较结果如表 5 所示。可以看到单独使用对抗数据增强策略使模型相对于基线模型有了显著地提高(约 5%),若将对抗增强替换为随机数据增强,则模型的泛化性会显著下降,这是由于不同的伪造方法生成的虚假图像具备不同的伪造特征,而对抗训练以及多卷积结果融合的训练方法促使网络学习到了更加本质的伪造特征,有效地缓解了特征差异带来的性能下降,这表明本文提出的对抗数据增强策略是有效且重要的。

表 5　对抗数据增强策略的消融实验结果

Table 5　Results of ablation experiments against the data augmentation strategy

Methods	Testing Set	Training Set				Avg
		DF	F2F	FS	NT	
EfficientNet-B4	Celeb-DF	0.681	0.642	0.631	0.625	0.644
EfficientNet-B4 w/adv	Celeb-DF	0.703	0.721	0.644	0.722	0.698
EfficientNet-B4 w/ran	Celeb-DF	0.663	0.708	0.701	0.666	0.685
Ours	Celeb-DF	0.730	0.756	0.731	0.752	0.742

4.5　可视化

使用 t-SNE 可视化跨数据集检测的特征分布可以更为直观地体现该方法的有效性,如图 7 所示。可以看到方法(ATNet)可视化的样本点集中性更强,正例和反例更加分离。其余检测方法的样本特征分布略混乱一些,真实图像和伪造图像特征交杂在一起。这主要是由于该类检测算法学习到的伪造特征更为浅层,无法应对伪造特征差异过大的虚假人脸图像的检测,而方法(ATNet)可以更加有效的应对这类检测,从而提高检测精度和泛化性。

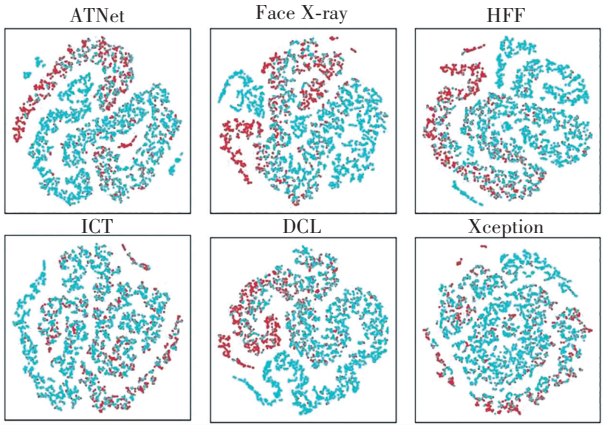


图 7　可视化结果

Fig. 7　Visualization results

5　结　论

利用对抗数据增强策略扩展数据集,同时结合高频和低频的多层级卷积特征可以更为全面的学习到伪造线索,提高准确率与泛化性。实验表明:所提工作(ATNet)是简单且高效的,可以有效提高虚假图像的检

测准确率,多个数据集上的测试也显示出更好的泛化性。在未来进一步的工作中,将会探寻更加有效的人脸检测方法,通过多模态信息结合的方法以搜索更加泛化性的特征表示,推动目前人脸伪造检测在跨数据集检测方面的难题。

参考文献(References):

[1] QIAN Y, YIN G, SHENG L, et al. Thinking in frequency: face forgery detection by mining frequency-aware clues[M]. Cham: Springer International Publishing, 2020: 86-103.

[2] LI J, XIE H, LI J, et al. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 6454-6463.

[3] LI L, BAO J, ZHANG T, et al. Face X-ray for more general face forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 5000-5009.

[4] FEI J, DAI Y, YU P, et al. Learning second order local anomaly for general face forgery detection[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 20270-20280.

[5] 瞿远近, 吴起. 基于改进高斯滤波网络的深度伪造检测方法[J]. 重庆工商大学学报(自然科学版), 2023, 40(4): 41-47.
QU Yuan-jin, WU Qi. Depth forgery detection method based on improved Gaussian filter network [J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(4): 41-47.

[6] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]// International Conference on Machine Learning. Los Angeles: PMLR. 2019: 6105-6114.

[7] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[M]. Boston, MA: Springer US, 1992: 5-32.

[8] ROSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics++: Learning to detect manipulated facial images [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 1-11.

[9] LI Y, YANG X, SUN P, et al. Celeb-DF: a large-scale challenging dataset for DeepFake forensics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 3204-3213.

[10] DEEPFAKES. faceswap[CP/OL]. https://github.com/deepfakes/faceswap/ 2023-10-25.

[11] THIES J, ZOLLHÖFER 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. Piscataway: IEEE Press, 2016: 2387-2395.

[12] MAREK K. FaceSwap[CP/OL]. https://www.github.com/MarekKowalski/FaceSwap 2021-4-13.

[13] THIES J, ZOLLHÖFER M, NIEßNER M. Deferred neural rendering[J]. ACM Transactions on Graphics, 2019, 38(4): 1-12.

[14] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1800-1807.

[15] ZHAO H, WEI T, ZHOU W, et al. Multi-attentional deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 2185-2194.

[16] LUO Y, ZHANG Y, YAN J, et al. Generalizing face forgery detection with high-frequency features[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 16312-16321.

[17] DONG X, BAO J, CHEN D, et al. Protecting celebrities from DeepFake with identity consistency transformer[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 9458-9468.

[18] SUN K, YAO T, CHEN S, et al. Dual contrastive learning for general face forgery detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 2316-2324.

[19] SUN Z, HAN Y, HUA Z, et al. Improving the efficiency and robustness of deepfakes detection through precise geometric features[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 3608-3617.

[20] LIU H, LI X, ZHOU W, 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. Piscataway: IEEE Press, 2021: 772-781.

责任编辑:陈 芳