

基于孪生网络的鲁棒性深度伪造检测方法

林善和

(福建星网锐捷通讯股份有限公司, 福建 福州 350117)

摘要: 目前, 使用深度伪造技术合成伪造人脸图像的案例非常广泛, 深度伪造技术通过对人脸图像进行面部替换或表情更改, 从而实现深度伪造的目的。这类深度伪造图像的肆意传播, 可能对社会和个人造成许多不良后果。因此, 不少学者着手研究深度伪造检测方法。虽然现有的检测方法在高质量图像的检测上能够实现非常高的准确率, 但是在检测经过图像压缩操作的低质量人脸图像时, 其检测精度会大幅下降。研究旨在改进现有深度伪造检测方法鲁棒性不足的问题。为此, 提出了一种基于孪生网络的深度伪造检测方法, 其思想是使用孪生网络来学习高质量图像和低质量图像之间公共伪造特征, 通过牺牲部分高质量图像的特征提取能力, 来提高网络对低质量图像的表征能力, 从而使网络在不同压缩率伪造图像的检测上都具有较高的准确率。实验结果表明, 所提方法在多个不同压缩率数据集上的综合平均准确率达到90%以上, 优于多个现有检测方法。通过消融实验证明所提方法简单有效, 且适用于不同的主干网络。

关键词: 伪造取证; 深度伪造检测; 孪生网络; 鲁棒性

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024016

Robust deepfake detection method based on siamese network

LIN Shanhe

Fujian Star-net Communication Company Limited, Fuzhou 350117, China

Abstract: The proliferation of deepfake (DF) technology for generating manipulated facial expressions in synthetic images has raised concerns due to its potential negative impacts on individuals and society. In response to the need for robust detection, researchers have been developing methods to identify deepfakes. While current detection methods perform well on high-quality images, they often falter when confronted with low-quality or compressed images. This study focused on enhancing the robustness of deepfake detection methods to address these limitations. A novel approach leveraging a Siamese network was proposed, designed to learn common forgery features across both high-quality and low-quality images. This was achieved by trading off some of the high-quality image feature extraction capabilities to bolster the representational capacity for low-quality images. The proposed method demonstrated an average accuracy exceeding 90% across various datasets with different compression levels, surpassing several

收稿日期: 2023-05-05; 修回日期: 2024-03-03

通信作者: 林善和, linsh@vip.sina.com

基金项目: 福建省创新战略研究计划项目 (No.2023R0156)

Foundation Item: The Innovation Strategic Research Plan Project of Fujian Province (No.2023R0156)

引用格式: 林善和. 基于孪生网络的鲁棒性深度伪造检测方法[J]. 网络与信息安全学报, 2024, 10(02): 182-189.

Citation Format: LIN S H. Robust deepfake detection method based on siamese network[J]. Chinese Journal of Network and Information Security, 2024, 10(02): 182-189.

existing detection techniques. The simplicity, effectiveness, and adaptability of the proposed method to different backbone networks were further substantiated through ablation experiments.

Keywords: forgery forensics, deepfake detection, siamese network, robustness

0 引言

深度伪造 (deepfake, DF) 技术是基于深度学习的方法对图像或视频中的人脸进行篡改的技术。这类人脸伪造图像和视频的传播会给社会和个人带来潜在威胁。在伪造取证领域中, 现有的检测方法可以实现非常高的检测精度, 但是存在鲁棒性弱的问题。具体来说, 现有的检测方法在检测经过图像压缩操作的人脸图像时, 其检测精度会大幅降低。

在网络传播过程中, 通常会使用图像压缩技术。为了应对将来网络中的伪造取证, 亟须鲁棒性强的检测方法。由于孪生网络^[1]在人脸识别、文本相似度、签名验证等领域有较强的检测能力, 可以学习成对样本之间的相似性或差异性。因此, 本文提出一种基于孪生网络的鲁棒性深度伪造检测方法, 使用一对不同压缩率的人脸图像作为网络的输入。其中心思想是希望孪生网络学习不同压缩率图像之间的特征, 提高网络对低质量图像的特征提取能力, 从而提高方法在不同压缩率图像中的检测精度, 以应对接近真实网络环境的伪造检测。在常用的 FaceForensics++ 数据集^[2]上进行实验, 与多个现有方法进行对比, 结果表明本文方法在低质量数据上具有更高的检测精度。此外, 通过消融实验验证, 本文方法简单有效且易于实现。

本文的主要贡献如下。

(1) 提出一种基于孪生网络的鲁棒性深度伪造检测方法, 在不同压缩率图像的检测上均有较高的检测准确率。

(2) 实验结果表明, 所提方法在多个不同压缩率数据集上的综合平均准确率优于多个现有检测方法。另外通过消融实验证明所提方法的有效性。

1 相关工作

1.1 深度伪造检测

为了应对深度伪造检测的挑战, 研究者提出

了各种方法, 基于目的大致可分为模型精度、泛化性和鲁棒性三类。

为了实现在单数据集上的高检测精度, Chollet^[3]使用 Xception 网络检测视频中的人脸篡改。Nguyen 等^[4]通过胶囊网络判断人脸伪造情况。Wang 等^[5]结合频域特征设计多模态多尺度 Transformer 模型^[6]进行伪造图片检测。Coccomini 等^[7]结合卷积神经和视觉 Transformer 模型实现高效检测。

在跨数据集或跨域检测时, 模型精度下降十分严重。为了提高模型的泛化性, Wang 等^[8]提出基于注意力的数据增强框架, 提高模型的泛化能力。Chen 等^[9]合成新的伪造数据, 随机选择 2 张图像的面部部位进行交换生成训练数据, 从而提高模型的泛化性。

大多数检测模型在面对经压缩过的人脸图像时, 检测精度下降严重。为了提高模型的鲁棒性, Wang 等^[10]利用孪生网络挖掘脸部和背景的不一致性, 通过拼接特征进行分类。Liu 等^[11]发现真假人脸在相位谱中有明显的差异, 通过简化网络提取更多的局部纹理信息来提高模型的鲁棒性。Wang 等^[12]利用相位谱和振幅谱信息, 设计双重注意力模块挖掘频域中的不一致区域, 提高模型的鲁棒性。吴文轩等^[13]以伪造前后光照差异为切入点, 对网络的中间特征进行分块处理, 并进行真伪性对比来提高模型的鲁棒性。朱春陶等^[14]基于时序上的多域特征, 融合空域和频域的时序特征类有效区分伪造人脸和真实人脸, 能够提高对低质量人脸图像的检测能力。

1.2 孪生网络

孪生网络是一种神经网络模型, 通常用于比较 2 个输入的相似度或差异。孪生网络结构如图 1 所示。孪生网络通常由 2 个相同的子网络结构组成, 并且每个子网络结构都有相同的权重参数和结构, 所以命名为孪生网络。

孪生网络中的参数共享使得 2 个子网络具有相同的特征提取能力, 通常用于解决相似度度量问题, 如人脸识别、文本相似度、签名验证等。

其基本思想是通过共享参数的子神经网络提取 2 个输入数据的特征向量, 然后通过距离度量比较这 2 个特征向量之间的相似度来判断 2 个输入数据之间的关系。常用的距离度量函数有余弦相似度、欧几里得距离和曼哈顿距离等。

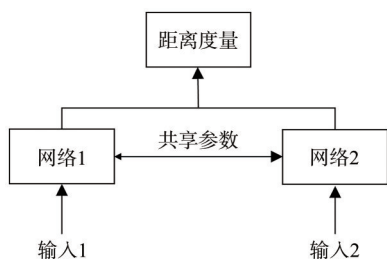


图1 孪生网络结构
Figure 1 Architecture of siamese network

在孪生网络的训练过程中, 通常需要一个成对的训练数据集, 其中每个样本由 2 个相似的数据点组成。网络的目标是学习将这些相似的数据点映射到相似的特征向量, 将不相似的数据点映射到不相似的特征向量。这可以通过设计损失函数来实现, 以使孪生网络能够学习输入数据之间的相似性或差异性。该网络简单易实现且不引入额外的参数量, 在人脸识别和图像分类等任务中有广泛应用。

2 本文方法

2.1 检测方法的总体流程

本文检测方法的训练流程如图 2 所示。本文方法采用孪生网络架构, 其中网络的输入分别为高质量 (high quality, HQ) 人脸图像 I_H 和低质量 (low quality, LQ) 人脸图像 I_L , 人脸图像的形状大小为 $C \times H \times W$, 其中 C 为人脸图像的通道数, H 和 W 分别是输入图像的高和宽。首先, 将高质量人脸图像 I_H 和低质量人脸图像 I_L 输入 2 个主干网络中, 分别提取高质量人脸图像的全局特征表示 g_H 和低质量人脸图像的全局特征表示 g_L ; 然后, 将高质量人脸图像的全局特征表示 g_H 和低质量人脸图像的全局特征表示 g_L 分别作为分类器输入得到各自检测结果; 最后, 使用一个常用的分类损失 L_{BCE} 监督模型的分类, 同时通过特征度量损失 L_{MSE} 最小化高质量人脸图像的全局特征表示 g_H 和低质量人脸图像的全局特征表示 g_L 之间的距离。

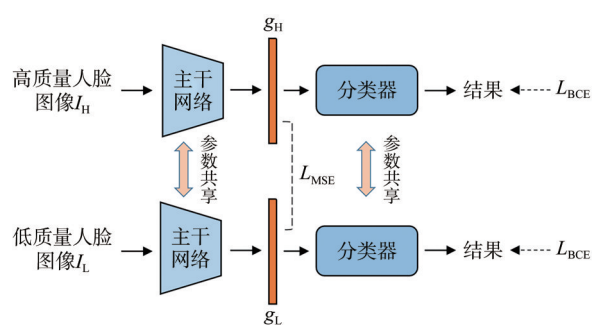


图2 本文检测方法的训练流程
Figure 2 Training process of our detection method

2.2 数据预处理

本文方法需要成对的不同压缩率图像作为网络输入, 由于 FaceForensics++ 数据集中包含了 C0 (表示压缩率参数为 0 的无损压缩数据)、C23 (表示压缩率参数为 23 的轻度压缩高质量数据)、C40 (表示压缩参数为 40 的重度压缩低质量数据) 数据, 直接使用该数据集中的 C23 数据和 C40 数据分别作为本文方法中的高质量人脸图像 I_H 和低质量人脸图像 I_L 。

由于图像和视频数据在网络传播中存在不同程度的压缩, 本文提出另一种简单便捷的样本对组成策略, 以应对不同压缩程度图像的检测。在训练阶段, 在高质量人脸图像 I_H 上使用不同程度的图像压缩, 从而得到低质量人脸图像 I_L 。这种方法简单有效, 同时能够获得不同压缩程度的数据, 提高方法的鲁棒性。

2.3 损失函数

本文的损失函数主要包含分类损失和特征度量损失 2 个部分。由于本文伪造检测任务是一个二分类任务, 所以分类损失采用二元交叉熵 (binary cross entropy, BCE) 损失, 表达式如式 (1) 所示, 其中 n 表示数据数量, y_i 表示第 i 个图像样本的真实标签, y'_i 表示第 i 个图像样本的网络预测值, 其主要判别标签类别与预测类别是否一致。

$$L_{BCE} = -\frac{1}{n} \sum_{i=1}^n [y_i \ln(y'_i) + (1 - y_i) \ln(1 - y'_i)] \quad (1)$$

本文特征度量损失的主要目的是约束高质量人脸图像的全局特征表示 g_H 和低质量人脸图像的全局特征表示 g_L , 最小化两者之间的距离, 使得网络提取低质量图像的特征表示 g_L 接近高质量图像的特征表示 g_H 。因此, 本文采用均方误差

(mean squared error, MSE) 损失作为特征度量损失, 表达式如式 (2) 所示, 其中 $g_{i,h}$ 和 $g_{i,l}$ 分别表示第 i 个样本对中高质量图像 I_h 和低质量图像 I_l 的全局特征表示, 通过均方误差损失可以拉近不同压缩图像之间的特征表示, 提高网络对低质量图像的特征挖掘能力。

$$L_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^n (g_{i,h} - g_{i,l})^2 \quad (2)$$

本文方法输入成对的人脸图像, 均使用二元交叉熵损失监督分类。其中 $L_{\text{BCE}}(H)$ 是对高质量人脸图像输入 I_h 的分类监督, $L_{\text{BCE}}(L)$ 是对低质量人脸图像 I_l 的分类监督。综上, 本文的总损失表达式如式 (3) 所示, 其中 α 、 β 和 γ 是各个损失的权重。

$$L = \alpha L_{\text{BCE}}(H) + \beta L_{\text{BCE}}(L) + \gamma L_{\text{MSE}} \quad (3)$$

3 实验分析

本文在 FaceForensics++ 数据集上进行训练评估。该数据集是深度伪造检测任务中最常用的数据集, 包含 4 个子数据集。共有 1 000 个真视频, 分别采用 DF、F2F (face2face)^[15]、FS (face swap) 和 NT (neural textures)^[16] 篡改方法获得 4 000 个伪造视频并按照篡改方法划分为 4 个子数据集。为了应对真实的网络环境, 对所有视频进行不同程度压缩处理, C0、C23、C40 参照文献[2], 按 720:140:140 划分训练集、验证集和测试集, 采用准确率和曲线下面积 (area under curve, AUC) 作为评价指标。

图像预处理使用 Dlib^[17] 提取视频中的人脸图像, 孪生网络中的特征提取网络使用 ResNet50^[18], 图像输入分辨率为 224 像素×224 像素, 使用 Adam 优化器, 批量大小设置为 64, 学习率为 3×10^{-4} , 训练 20 轮。与以下的检测方法进行对比。Capsule: Nguyen 等^[9]使用胶囊网络达到了非常高的检测准确率。CViT (convolutional vision transformer): Coccomini 等^[7]结合卷积神经网络和视觉 Transformer 实现了高效伪造检测。SLADD (self-supervised learning of adversarial deepfake detection): Chen 等^[10]提出合成新的伪造数据来提高模型泛化能力, 随机选择 2 张图像的面部部位进行交换生成训练数据。RFM (representative forgery mining): Wang 等^[9]提出一个

基于注意力的数据增广架构, 提高模型的泛化能力。SPSL (spatial-phase shallow learning): Liu 等^[12]发现相位谱中有明显的差异, 并通过简化网络提取更多的局部纹理信息来提高模型的鲁棒性。MC-LCR (multimodal contrastive classification by locally correlated representations): Wang 等^[13]利用相位谱和振幅谱信息, 设计双重注意力模块挖掘频域中的不一致区域, 提高模型的鲁棒性。ISN (improved siamese network): Wang 等^[11]使用改进的孪生网络, 将输入图像的人脸区域和背景区域分离, 作为孪生网络的成对输入, 希望网络学习到人脸区域和背景区域之间的不一致性, 从而辨别出伪造人脸, 提高模型的鲁棒性。DDPLI (deepfake detection based on patch-wise lighting inconsistency): 吴文轩等^[14]通过捕获伪造前后的光照差异特征来进行伪造检测。

3.1 跨压缩率评估

为了验证本文方法的鲁棒性, 在 FaceForensics++ 数据集上进行了跨压缩率的实验。首先, 在 FaceForensics++ 数据集的 4 个子数据集上进行实验。具体来说, 分别在 4 个子数据集的 C23 训练集上进行训练, 并分别在 C23 和 C40 的测试集上进行测试评估。SN-HL (siamese network - high and low quality) 表示使用 C23 和 C40 组成的样本对参与网络训练。由于对比方法的训练数据只使用了 C23 的数据, 为了公平对比, SN-H (siamese network - high quality) 表示只使用了 C23 数据训练, 低质量输入图像通过对高质量输入图像进行图像压缩的数据增广方式生成。单数据集上跨压缩率评估结果如表 1 所示, 本文方法采用的主干网络是 ResNet50 和 Xception, 评估指标采用准确率, 其中 Avg 表示平均检测准确率, 综合考虑不同检测方法在不同压缩率上的检测效果。

在 DF 数据集上, RFM^[9]方法在 C23 上实现了最高检测准确率 98.62%, DDPLI^[14]方法在 C40 和 Avg 上实现了最高检测准确率, 分别达到了 97.26% 和 97.47%。在 C40 和 Avg 上, 本文方法 SN-HL (Xception) 高于 RFM^[9]方法 10.00%, 在 C23 上, 略微领先 DDPLI^[14]方法 0.70%。

在 F2F 数据集上, DDPLI^[14]方法在 C23 上实现了最高检测准确率 98.14%, 但是本文方法 SN-

表 1 单数据集上跨压缩率评估结果
Table 1 Evaluation results of cross-compression rate on single dataset

检测方法	DF-C23			F2F-C23			FS-C23			NT-C23		
	准确率 (C23)	准确率 (C40)	Avg	准确率 (C23)	准确率 (C40)	Avg	准确率 (C23)	准确率 (C40)	Avg	准确率 (C23)	准确率 (C40)	Avg
Capsule ^[5]	97.40%	77.49%	87.45%	97.61%	58.69%	78.15%	97.85%	75.40%	86.63%	92.30%	63.32%	77.82%
CViT ^[7]	98.54%	87.64%	93.09%	98.05%	65.59%	81.82%	98.83%	88.72%	93.78%	92.61%	60.45%	76.54%
SLADD ^[10]	97.73%	84.48%	91.11%	95.66%	62.60%	79.13%	97.25%	75.96%	86.61%	92.54%	67.00%	79.77%
RFM ^[9]	98.62%	75.39%	87.01%	97.91%	66.91%	82.41%	98.70%	82.02%	90.36%	92.90%	59.19%	76.10%
DDPLI ^[14]	97.68%	97.26%	97.47%	98.14%	88.92%	93.53%	97.535	90.84%	94.19%	92.16%	78.65%	85.41%
SN-H(ResNet50)	97.88%	94.31%	96.10%	95.27%	87.24%	91.26%	96.69%	89.65%	93.17%	86.21%	74.60%	80.41%
SN-HL(ResNet50)	97.80%	94.57%	96.19%	96.56%	89.31%	92.94%	96.75%	92.74%	94.75%	89.27%	77.02%	83.15%
SN-HL(Xception)	98.38%	96.21%	97.30%	97.74%	89.73%	93.74%	98.05%	94.29%	96.17%	90.71%	80.99%	85.85%

HL(Xception)在C40和Avg上实现了最高检测准确率,分别达到了89.31%和92.94%。

在FS数据集上, CViT^[7]方法在C23上实现了最高检测准确率98.83%,但是本文方法SN-HL(Xception)在C40和Avg上实现了最高检测准确率,分别达到了94.29%和96.17%,并且在Avg上领先CViT^[7]方法2.39%和DDPLI^[14]方法1.98%。

在NT数据集上, RFM^[9]方法在C23上实现了最高检测准确率92.90%,但是本文方法SN-HL(Xception)在C40和Avg上实现了最高检测准确率,分别达到了80.99%和85.85%,大幅超过RFM^[9]方法。对比DDPLI^[14]方法,本文方法在C23上的检测效果较弱,但是在C40和Avg上本文方法更优。

综合以上4个数据集的12个评估结果,本文方法有6个结果最优, DDPLI^[14]方法有3个最优。尤其是在C40和Avg的评估上,在除DF数据集以外的3个数据集上达到最优,体现了本文方法的强鲁棒性。

本文方法在FaceForensics++子数据集的不同压缩率上实现了较高的检测准确率。此外,在包含4个子数据集的高质量数据上进行实验,并在高质量数据和低质量数据上评估。多数据集上跨压缩率评估结果如表2所示,在高质量数据上RFM^[9]达到了最高检测准确率95.64%,但是本文方法SN-HL在低质量数据和Avg上实现了最高检测准确率,分别达到了85.36%和88.18%,并且在Avg上领先次优的RFM^[9]方法7.50%。此外,本文方法SN-H在低质量数据和Avg上的准确率

均高于对比方法,在Avg上领先次优的RFM^[9]方法5.78%。综上,与大多数方法进行对比,虽然在高质量数据集的检测上,本文方法的准确率低于多数方法,但是本文方法在低质量数据和两者的平均检测准确率均取得了最高的检测准确率。结果表明,本文方法在低质量数据的检测能力更强,鲁棒性更强。

表 2 多数据集上跨压缩率评估结果
Table 2 Evaluation of cross-compression rate on multiple dataset

检测方法	准确率		Avg
	HQ	LQ	
CViT ^[7]	94.10%	60.75%	77.43%
RFM ^[9]	95.64%	65.71%	80.68%
SLADD ^[10]	94.25%	64.00%	79.13%
SN-H(ResNet50)	90.11%	82.80%	86.46%
SN-HL(ResNet50)	90.99%	85.36%	88.18%

3.2 鲁棒性对比

为了进一步验证本文方法的鲁棒性,将本文方法与鲁棒性强的方法进行对比。为了体现本文方法的有效性,更换孪生网络中的主干网络为Xception网络,评价指标采用AUC,在低质量数据上进行评估测试。与鲁棒性强的方法对比结果如表3所示,本文方法在所有数据集上的AUC均高于SPSL^[12]方法,特别是在F2F、NT和LQ数据集上本文方法的检测能力更强,分别领先2.20%、10.40%和8.32%。与MC-LCR^[13]方法对比,MC-LCR^[13]方法在DF和FS数据集上达到了最高AUC,分别是99.53%和98.60%,本文方法AUC

分别为99.35%和98.38%，与MC-LCR^[13]方法取得了相当的检测性能。在F2F数据集上，本文方法略优于MC-LCR^[13]方法。但是在NT数据集上，本文方法达到了最高的90.89%，高于MC-LCR^[13]方法3.50%。并且在综合以上4个子数据集的LQ数据集上，本文方法实现了最优的检测，达到了91.14%，高于次优的MC-LCR^[13]方法0.86%。综上，对比鲁棒性强的方法，本文方法在DF和FS数据集上取得了与MC-LCR^[13]方法相当的检测性能，在F2F、NT和LQ数据集上取得了最高的AUC。实验结果表明，本文方法具有较强的鲁棒性。

表3 鲁棒性对比结果
Table 3 Robustness comparison

检测方法	AUC				
	DF-C40	F2F-C40	FS-C40	NT-C40	LQ
SPSL ^[12]	98.50%	94.62%	98.10%	80.49%	82.82%
MC-LCR ^[13]	99.53%	96.72%	98.60%	87.39%	90.28%
SN-HL (Xception)	99.35%	96.82%	98.38%	90.89%	91.14%

3.3 孪生网络方法对比

将本文与ISN^[11]检测方法进行对比，该方法利用改进的孪生网络来挖掘人脸图像与背景图像之间的差异，从而实现伪造检测。本文方法的主干网络采用Xception网络，实验评估在高质量数据集上进行，评价指标采用准确率。与孪生网络的检测方法ISN对比结果如表4所示，本文方法在4个数据集上的检测准确率均优于ISN^[11]方法。在DF数据集上领先2.61%，在F2F数据集上领先0.62%，在FS数据集上领先0.68%，在NT数据集上领先5.50%。结果表明，本文方法具有更高效的检测能力。

表4 与孪生网络的检测方法ISN对比结果
Table 4 Comparison results with siamese network detection method ISN

检测方法	准确率			
	DF	F2F	FS	NT
ISN ^[11]	95.77%	97.12%	97.37%	84.71%
SN-HL(Xception)	98.38%	97.74%	98.05%	90.21%

3.4 消融实验一方法有效性

采用本文方法与选用的ResNet50和Xception网络进行对比，在FaceForensics++数据集的4个

子数据集上进行跨压缩率评估，评价指标采用检测准确率。验证方法有效性的跨压缩率评估如表5所示，主干网络采用ResNet50，在高质量数据C23的评估上，本文方法在F2F、FS和NT多个数据集上均低于基准，特别是在NT数据集上降低了2.73%。但是在检测C40数据时，本文方法在4个数据集上均有提高，分别提高了2.05%、3.85%、4.78%和6.41%。采用综合评估方法，在C23和C40上取两者的Avg进行比较，本文方法在4个数据集上均有提高，分别提高了1.20%、1.64%、2.05%和1.84%。为了进一步验证本文方法的有效性，换用Xception网络作为主干网络。本文方法在C23上的检测不如基准，但是在C40和Avg上，本文方法相比基准均有提升。在DF数据集上，C40的准确率提高了3.57%，Avg提高了2.39%。在F2F数据集上，C40的准确率提高了8.30%，Avg提高了4.21%。在FS数据集上，C40的准确率提高了3.37%，Avg提高了1.43%。在NT数据集上，C40的准确率提高了7.28%，Avg提高了2.77%。结果表明，本文方法的鲁棒性更强，能够提高在C40上的检测准确率。在C23上会有些许下降，这是因为网络需要同时关注不同压缩率图像的伪造特征，提高了对低质量图像的特征提取能力，对高质量图像的特征提取能力会稍微降低。但是，综合评价方法在2种数据上的检测准确率，本文方法均有提升，这证明提出的孪生网络架构能够提高方法的鲁棒能力。

3.5 消融实验一数据增强策略的有效性

本文在训练样本对的选择中，提出一种使用图像压缩的数据增强的方法来生成样本对，这种方法具有简单、有效和易实现的特性。为了验证该策略的有效性，使用ResNet50作为主干网络，验证数据增强策略有效性的消融实验结果如图3所示。使用该策略后，本文方法在4个C40数据集的检测上均有提升，表明该策略能够提高方法的鲁棒性。

3.6 消融实验一特征度量损失的选择

特征度量损失主要用来最小化高质量人脸图像的全局特征表示 g_H 和低质量人脸图像的全局特征表示 g_L 。本文通过消融实验来选择合适的特征度量损失，分别使用均方误差(MSE)损失和绝

表5 验证方法有效性的跨压缩率评估
Table 5 Cross-compression rate evaluation to verify method validity

检测方法	DF-C23			F2F-C23			FS-C23			NT-C23		
	准确率 (C23)	准确率 (C40)	Avg	准确率 (C23)	准确率 (C40)	Avg	准确率 (C23)	准确率 (C40)	Avg	准确率 (C23)	准确率 (C40)	Avg
ResNet50	97.46%	92.52%	94.99%	97.13%	85.46%	91.30%	97.43%	87.96%	92.70%	92.00%	70.61%	81.31%
SN-HL(ResNet50)	97.80%	94.57%	96.19%	96.56%	89.31%	92.94%	96.75%	92.74%	94.75%	89.27%	77.02%	83.15%
Xception	97.18%	92.64%	94.91%	97.62%	81.43%	89.53%	98.55%	90.92%	94.74%	92.44%	73.71%	83.08%
SN-HL(Xception)	98.38%	96.21%	97.30%	97.74%	89.73%	93.74%	98.05%	94.29%	96.17%	90.71%	80.99%	85.85%

对值误差 (mean absolute error, MAE) 损失作为特征度量损失, 在 4 个 C40 上进行评估, 采用准确率作为评价指标。不同特征度量损失的消融实验结果如图 4 所示, 在 DF 和 F2F 数据集上, 两者均实现了相当的检测性能, 但是在 FS 和 NT 数据集上, MSE 损失具有更强的检测能力。从整体上看, 使用 MSE 损失实现了更高的检测准确率, 因此选用 MSE 损失作为特征度量损失。

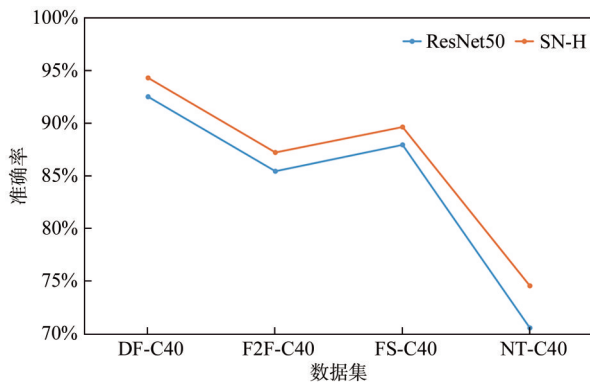


图3 验证数据增强策略有效性的消融实验结果
Figure 3 Results of ablation experiments to verify the effectiveness of data augmentation strategies

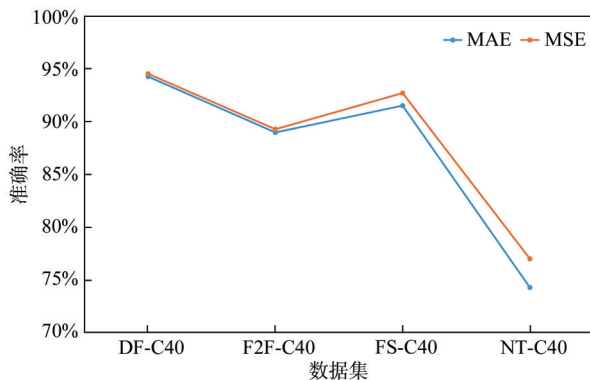


图4 不同特征度量损失的消融实验结果
Figure 4 Ablation experimental results for different feature metric losses

4 结束语

本文提出一种基于孪生网络的鲁棒性深度伪造检测方法, 使用一对不同压缩率的人脸图像作为方法的输入。其思想是希望孪生网络学习不同压缩率图像之间的特征, 同时通过特征度量损失使得低质量图像的特征向高质量特征逼近, 提高网络对低质量图像的特征提取能力, 进而提高方法的鲁棒性。在 FaceForensics++ 数据集上多个对比实验结果表明, 本文方法优于多个最新检测方法。最后, 通过消融实验验证了本文方法的有效性。

参考文献:

- [1] CHOPRA S, HADSELL R, LECUN Y. Learning a similarity metric discriminatively, with application to face verification[C]// Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005: 539-546.
- [2] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics: learning to detect manipulated facial images[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. 2019: 1-11.
- [3] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1800-1807.
- [4] NGUYEN H H, YAMAGISHI J, ECHIZEN I. Use of a capsule network to detect fake images and videos[J]. arXiv preprint arXiv: 1910.12467. 2019.
- [5] WANG J K, WU Z X, OUYANG W H, et al. M2TR: multi-modal multi-scale transformers for Deep fake detection[C]// Proceedings Proceedings of the 2022 International Conference on Multimedia Retrieval. 2022: 615-623.
- [6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv preprint arXiv: 1706.03762. 2017.
- [7] COCCOMINI D A, MESSINA N, GENNARO C, et al. Combining EfficientNet and vision transformers for video deepfake detection[C]// Proceedings of the International Conference on

- Image Analysis and Processing, 2022: 219-229.
- [8] WANG C R, DENG W H. Representative forgery mining for fake face detection[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14918-14927.
- [9] CHEN L, ZHANG Y, SONG Y B, et al. Self-supervised learning of adversarial example: towards good generalizations for deepfake detection[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 18689-18698.
- [10] WANG B, LI Y C, WU X H, et al. Face forgery detection based on the improved Siamese network[J]. Security and Communication Networks, 2022: 5169873.
- [11] LIU H G, LI X D, ZHOU W B, et al. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain [C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 772-781.
- [12] WANG G J, JIANG Q, JIN X, et al. MC-LCR: Multimodal contrastive classification by locally correlated representations for effective face forgery detection[J]. Knowledge-Based Systems, 2022, 250: 109114.
- [13] 吴文轩, 周文柏, 张卫明, 等. 基于块间光照不一致性的深度伪造检测算法[J]. 网络与信息安全学报, 2023, 9(1): 167-177. WU W X, ZHOU W B, ZHANG W M, et al. Deepfake detection method based on patch-wise lighting inconsistency[J]. Chinese Journal of Network and Information Security, 2023, 9(1): 167-177.
- [14] 朱春陶, 尹承禧, 张博林, 等. 基于多域时序特征挖掘的伪造人脸检测方法[J]. 网络与信息安全学报, 2023, 9(3): 123-134. ZHU C T, YIN C X, ZHANG B L, et al. Forgery face detection method based on multi-domain temporal features mining[J]. Chinese Journal of Network and Information Security, 2023, 9(3): 123-134.
- [15] THIES J, ZOLLHÖFER M, STAMMINGER M, et al. Face2Face: real-time face capture and reenactment of RGB videos[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2387-2395.
- [16] THIES J, ZOLLHÖFER M, NIEßNER M. Deferred neural rendering: image synthesis using neural textures[J]. ACM Transactions on Graphics, 38(4): 66.
- [17] KING D E. Dlib-ml: a machine learning toolkit[J]. Journal of Machine Learning Research. 2009, 10: 1755-1758.
- [18] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[作者简介]



林善和 (1971-), 男, 福建福州人, 福建星网锐捷通讯股份有限公司高级工程师, 主要研究方向为深度学习、伪造取证、视频通信智能传感器。