中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2025)02-0421-14

论文引用格式:Zhang J, Xu P, Liu W J, Guo X X and Sun F. 2025. Negative instance generation for cross-domain facial forgery detection. Journal of Image and Graphics, 30(02):0421-0434(张晶, 许盼, 刘文君, 郭晓萱, 孙芳. 2025. 多样性负实例生成的跨域人脸伪造检测. 中国图象图形学报, 30(02):0421-0434)[DOI:10.11834/jig.240160]

多样性负实例生成的跨域人脸伪造检测

张晶,许盼,刘文君,郭晓萱,孙芳*

辽宁师范大学计算机与人工智能学院,大连　116081

摘　要:**目的**　深度伪造检测(deepfake detection)通过训练复杂深度神经网络,挖掘更具辨别性的人脸图像表示,获得高精度的检测结果,其是一项确保人脸信息真实、可靠和安全的重要技术。然而,目前流行的模型存在过度依赖训练数据,使模型仅在相同域内表现出令人满意的检测性能,在跨领域场景中表现出较低泛化性,甚至使模型失效。因此,如何在有限的训练数据下实现跨域环境中的高效伪造人脸检测,成为亟待解决的问题。基于此,本文提出多样性负实例生成的跨域人脸伪造检测模型(negative instance generation-FFD, NIG-FFD)。**方法**　首先,通过构建孪生自编码网络,获得标签一致的潜在多视图融合特征,引入对比约束提高难样本特征可判别性;其次,在高效训练的同时利用构造规则生成更具多样性的负实例融合特征,提高模型泛化性;最后,构建自适应重要性权值矩阵,避免因负实例生成导致类别分布不平衡使正类别样本欠学习。**结果**　在两个流行的跨域数据集上验证本文模型的有效性,与其他先进方法相比,AUC(area under the receiver operating characteristic curve)值提升了10%。同时,在本域检测中ACC(accuracy score,)与AUC值相比其他方法均提升了近10%与5%。**结论**　与对比方法相比,本文方法在跨域和本域的人脸伪造检测上都取得了优越的性能。本文所提的模型代码已开源至:https://github.com/LNNU-computer-research-526/NIG-FFD

关键词:深度伪造检测;跨域人脸伪造检测;多视图特征融合;特征生成;对比约束

Negative instance generation for cross-domain facial forgery detection

Zhang Jing, Xu Pan, Liu Wenjun, Guo Xiaoxuan, Sun Fang*

School of Computer Science and Artificial Intelligence, Liaoning Normal University, Dalian　116081, China

Abstract: **Objective** With the rapid development of multimedia, mobile internet, and artificial intelligence technologies, facial recognition has achieved tremendous success in areas such as identity verification and security monitoring. However, with its widespread application, the risk of facial forgery attacks is gradually increasing. These attacks leverage deep learning models to create fraudulent digital content, including images, videos, and audio, posing a potential threat to societal stability and national security. Therefore, achieving deepfake detection is crucial for maintaining individual and organizational interests, ensuring public safety, and promoting the sustainable development of innovative technologies. According to different modes of image representation, deepfake detection methods can generally be divided into two categories. First, methods based on traditional image feature description typically involve image processing and feature extraction based on signal transformation models. Second, methods based on deep learning strategies for forged facial detection employ com-

收稿日期:2024-03-27;**修回日期**:2024-05-13;**预印本日期**:2024-05-20

*通信作者:孙芳　sunfang@lnnu.edu.cn

基金项目:国家自然科学基金项目(61902165,61976109);辽宁省教育厅基本科研项目(LJKMZ20221425,JYTMS20231040)

Supported by: National Natural Science Foundation of China(61902165, 61976109); Basic Research Project of Educational Department of Liaoning Province(LJKMZ20221425, JYTMS20231040)

plex deep neural networks to obtain more discriminative high-dimensional nonlinear facial feature descriptions, thereby improving forgery detection accuracy. Both of these methods have achieved satisfactory results in deepfake detection experiments. However, most training and testing samples for these models are collected from the same data domain, resulting in excellent performance under such conditions; subsequently, it becomes challenging to obtain testing samples that are consistent with the distribution of the original training samples in practical applications, which may limit the application of these models in free-scene forgery detection tasks and even lead to complete model failure. Therefore, some scholars have proposed a data augmentation framework based on structural feature mining to increase the performance of convolutional neural network detectors. However, when faces are seamlessly integrated with backgrounds at the pixel level, the recognition accuracy significantly decreases. Consequently, some scholars have utilized transformer network architectures to construct deep forgery detection frameworks. Although this model achieves satisfactory generalization by deeply understanding the manipulated regions, it lacks descriptions of local tampering representations, and its detection efficiency is also quite low. On this basis, the main challenges faced in constructing deepfake detection models in cross-domain scenarios can be summarized as follows: 1) extracting discriminative representations of forged facial images. The forgery process of facial images typically involves tampering or replacing local features of the image, posing challenges for obtaining discriminative features. 2) Improving the generalizability of detection models. Overreliance on current domain data during model training reduces the generalizability of recognition to other domain data, and when facing more challenging free-forgery detection scenarios, model failure may occur. This study addresses these challenges by introducing a cross-domain detection model that is based on diverse negative instance generations. **Method** The model achieves feature augmentation of forged negative instances and enhances the cross-domain recognition accuracy and generalizability by constructing a Siamese autoencoder network architecture with multiview feature fusion. It consists of the following three parts: 1) the model implements discriminative multiview feature fusion under contrastive constraints. First, a Siamese autoencoder network is constructed to extract different view features. Second, contrastive constraints are employed to achieve multiview feature fusion. Given that typical facial forgery image manipulation involves only small-scale replacements and tampering, the global features of forged facial images are remarkably similar to those of real faces. Contrastive loss enables the differentiation of weakly discriminative hard samples. It maximizes the similarity of intraclass features while minimizing the similarity of interclass features. Finally, comprehensive learning is facilitated by guiding the supervised feature extraction network to retain important feature information of the original input, an approach for emphasizing the learning of discriminative feature representations. This study proposes the use of reconstruction loss to constrain the feature network by computing the difference between the decoder output and the original input. 2) The model achieves diversity in negative instance feature augmentation to enhance model generalizability, ensuring satisfactory recognition performance on cross-domain datasets. First, the rules for generating the fused samples are defined. This study statistically visualizes the network output feature histograms of constructed samples with different labels via feature visualization, analyzes the statistical patterns of negative samples, and defines feature-level sample generation rules: except when both view features are from positive samples, all other combinations of feature samples are generated as negative samples. Second, diverse forged feature sets are constructed using selected samples to enable the network to learn more discriminative features. Finally, a global training sample set is obtained by connecting the original training samples and augmented samples. 3) The model implements a discriminator construction with importance sample weighting. When the abovementioned feature augmentation of negative instances is achieved, the number of original negative instances can be significantly increased. This study introduces an importance weighting mechanism to avoid model overfitting on negative samples and underfitting on positive samples. First, the matrix is initialized to set different weights for each class sample, allowing negative samples to be weighted according to their predicted probabilities while keeping positive samples unchanged, thereby approximately achieving class balance during the loss calculation. Through negative sample weighting, the model is guided to pay more attention to positive sample features and prevent the classification decision boundary from biasing toward negative samples. Second, the distance between the predicted probability distribution and the true probability distribution was measured via cross-entropy loss as the classification loss function to supervise the classification results. Finally, the total loss function for model training is obtained. **Result** Experiments were conducted on three publicly available datasets to verify the effectiveness of the proposed method

in a cross-domain environment. The model was subsequently compared with other popular methods, namely, FaceForensics++ (FF++), Celeb-DFv2, and the Deepfake Detection Challenge, and the results were analyzed. The FF++ dataset comprises three versions based on different compression levels: c0 (original), c23 (high quality), and c40 (low quality). This study utilized the c23 and c40 versions for experimentation. The Celeb-DFv2 dataset is widely employed to test the models' generalization capabilities, as its forged images lack obvious visual artifact characteristics of deepfake manipulation, posing significant challenges in generalization detection. In the experiments, 100 genuine videos and 100 forged videos were randomly selected, with one image extracted every 30 frames. For the DFDC dataset, 140 videos were randomly selected, with 20 frames extracted from each video for testing. According to the experimental results, the proposed model exhibited a 10% improvement in the area under the curve (AUC) of the receiver operating characteristics compared with other state-of-the-art methods. Additionally, the model's detection results in the native domain environment were validated, showing an approximate 10% and 5% increase in the ACC (accuracy score) and AUC values, respectively, compared with those of the other methods. **Conclusion** The method proposed in this study achieves superior performance in both cross-domain and in-domain deepfake detection.

Key words: deepfake detection; cross-domain face forgery detection; multi-view feature fusion; feature generation; contrastive constrain

0 引　言

多媒体、移动互联网以及人工智能技术的快速发展，使人脸识别在身份验证与安全监控等领域获得了巨大成功。然而，随着其广泛应用，人脸伪造攻击(face forgery attack, FFA)风险逐步增加，该类方法借助深度学习模型，建立包括：图像、视频和音频等虚假数字内容，对社会稳定与国家安全构成潜在威胁。因此，如何实现深度伪造检测(deepfake detection)对于维护个人利益、公共安全以及推动前沿技术的可持续发展至关重要。

基于图像的人脸伪造攻击(FFA)从伪造手段上可以分为4类：全脸合成、身份交换、属性操作和表情交换。伪造方式的多样性对检测方法提出了更高要求。根据图像表达方式的不同，深度伪造检测可分为两类：1)基于传统图像特征描述方法，一般利用信号转换模型实现特征提取，该类方法可获得手动设计特征，包括：纹理特征、几何形状、颜色和光照等(Zhang等，2019；Qian等，2020)，描述真实与伪造人脸图像细节，实现伪造检测。2)基于深度学习策略的伪造人脸检测方法，通过建立复杂深度神经网络获得更具判别性的高维度的非线性人脸特征描述，提高伪造检测精度。基于卷积神经网络的人脸伪造检测已得到广泛应用(Zhuang等，2022；丁峰 等，2024)。Chollet(2017)、Afchar等人(2018)分别提出Xception和Mesonet深度学习网络框架，但由于网络结构简单，无法达到较好的检测效果，需要与数据增强等方法结合(Guo等，2023a；Wang和Chow，2023)，通过挖掘全局结构特征等更全面的特征表达提升网络检测性能。基于此，Cozzolino等人(2017)通过改进网络架构，提出一类基于残差的描述符，并放松训练约束，在一个相对较小的训练集上微调网络，从而获得了显著的性能提升。同时，Bayar和Stamm(2018)通过约束卷积层，联合抑制图像的内容和自适应地学习操作检测特征，提升检测精度。然而，这两种方法在特征表达上仍存在局限性。基于此，Wang和Deng(2021)提出一种基于注意力的数据增强框架，以引导检测器更精细和广泛地关注深度伪造图像，但其对于细节的识别不敏感，在局部细节伪造图像的检测上表现不佳。该方法会追踪并遮挡面部最敏感的Top-N区域，鼓励检测器深入挖掘之前被忽略的更具代表性的伪造区域，但其对于局部区域的关注较为单一。针对这一问题，Sun等人(2021)提出基于元学习技术的检测模型。为了更好地获得伪造线索，Zhao等人(2021)、Cao等人(2023)分别提出了基于注意力机制的深度伪造检测网络，Chen等人(2021)研究了区域间的局部关系，Cao等人(2022)提出一种端到端的重建分类学习框架，获得了更高的识别精度。

上述构建伪造检测模型的训练与测试样本均采集自相同数据域，模型在该条件下取得了优异性能。然而，实际应用中难以获得与原始训练样本分布一致的测试样本，该类自由场景检测将限制模型应用，

甚至导致模型完全失效。如图1所示，左侧表示模型训练数据集，右侧表示待检测目标。在跨域测试环境中，即右侧图中下面4幅人脸图像，模型很难实现人脸伪造检测。因此，如何提高模型泛化性，实现快速、准确的跨域检测成为一个重要且亟待解决的问题。Guo等人(2023b)提出结构特征挖掘的数据增强框架(mining structural features, MSF)，通过挖掘结构特征扩大检测器的关注范围，并充分获取全局图像中的结构特征，使模型具有一定的泛化性。然而，当人脸与背景在像素级别无缝集成时，识别准确性将大大降低。针对该问题，Khormali和Yuan(2024)利用Transformer网络架构，构造深度伪造检测框架，与图卷积网络相结合，提高模型泛化性。为提升训练样本的多样性，数据增广的方法广泛应用于模型中，通过增广样本可以使模型学习到更多样化的特征。Li等人(2020a)提出Face X-ray模型，其不依赖于任何伪影先验知识，仅通过真实样本生成伪造样本用于训练，具有一定的泛化性。然而，当模型应用于由新的面部操纵技术生成的伪造时，体现出一定的局限性。针对这一问题，Chen等人(2022)提出了具有泛化表示的深度伪造检测方法，使用对抗性训练策略动态合成当前模型最具挑战性的伪造样本，但该方法在面向人脸与背景的像素级别无缝衔接的样本时，其检测精度有所下降。针对这一问题，Yu等人(2024)提出人脸交换生成器，重建对齐图像对，以可视化测试图像和真实参考图像之间的身份差异，通过自定义度量量化身份不一致性，区分假图像和真实图像，该方法具有一定的泛化性，但其在更具挑战的数据集上检测效果有所下降。样本增广的方式可充分训练模型，提高模型的泛化性。然而，上述方法一般为样本级增广，该方式将降低模型学习效率。同时，由于生成样本分布较单一，使其在更具挑战的跨域数据集上的学习性能有所下降。基于此，构建跨域场景中人脸伪造检测模型面临的主要挑战可以总结为：1)如何提取更具判别性的伪造人脸图像表达。人脸伪造通常涉及篡改或替换图像中相对较小区域的局部特征，为获取判别性特征带来挑战。2)如何提高检测模型泛化性。模型训练过度依赖当前域数据，且在面对更具挑战的自由伪造检测场景时将导致模型失效。

针对上述问题，本文提出一种多样性负实例生成的跨域检测方法，通过构建孪生自编码网络架构，在多视图特征融合基础上实现伪造负实例的特征增广，提升模型跨域识别精度与泛化性。首先，在网络嵌入层引入对比约束，提高多视图融合特征可判别性。其次，利用孪生编码输出，结合生成规则获得特征级负实例，以提高训练样本的多样性。最后，为减轻由样本增广引起的负实例分布偏差，其将导致正样本欠学习，引入自适应重要性矩阵平衡分布。本文主要贡献如下：1)针对挑战1，提出对比约束下的多视图融合网络架构，利用多视图互补特征，提高图像可判别特征描述，并通过对比约束保证难样本识别精度。2)针对挑战2，提出特征层级的负实例增广方法，生成多样性负实例融合特征，提高识别模型泛化性。同时，引入样本重要性加权策略，避免负样本过学习，正样本欠学习，保证模型识别精度。3)本文模型在多视图特征融合过程中，同时实施伪造特征增广，实现快速模型训练。此外，在跨域流行数据集中验证了所提出模型的人脸伪造检测性能，并在本域测试数据集中与其他最先进的方法相比，本文

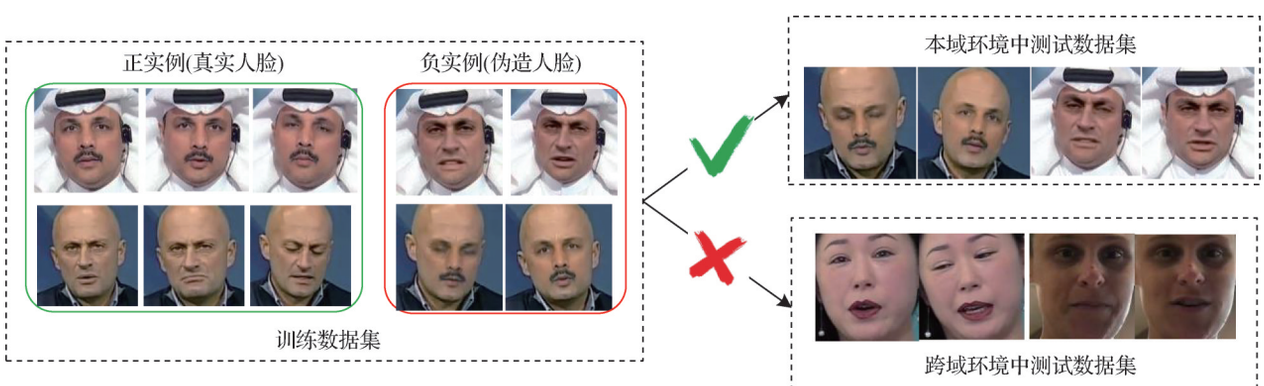


图1　本域与跨域人脸伪造检测

Fig. 1　Intra-domain and cross-domain face forgery detection

方法表现出令人满意的性能。

1 本文模型

针对跨域场景中构建人脸伪造检测模型的挑战，本文提出负实例生成的孪生深度网络架构，如图 2 所示。该框架利用对比约束下的孪生网络输出，获得更具判别性的多视图融合特征，解决模型学习中判别特征的局限性。进一步，通过定义负实例融合特征的生成规则，构建负实例特征，使检测模型充分学习多样性负实例表达，提高模型泛化性，提高跨域场景下人脸伪造检测的准确率。同时，为避免正样本欠拟合学习问题，通过模型训练过程中的预测概率，构建重要性矩阵平衡分布。

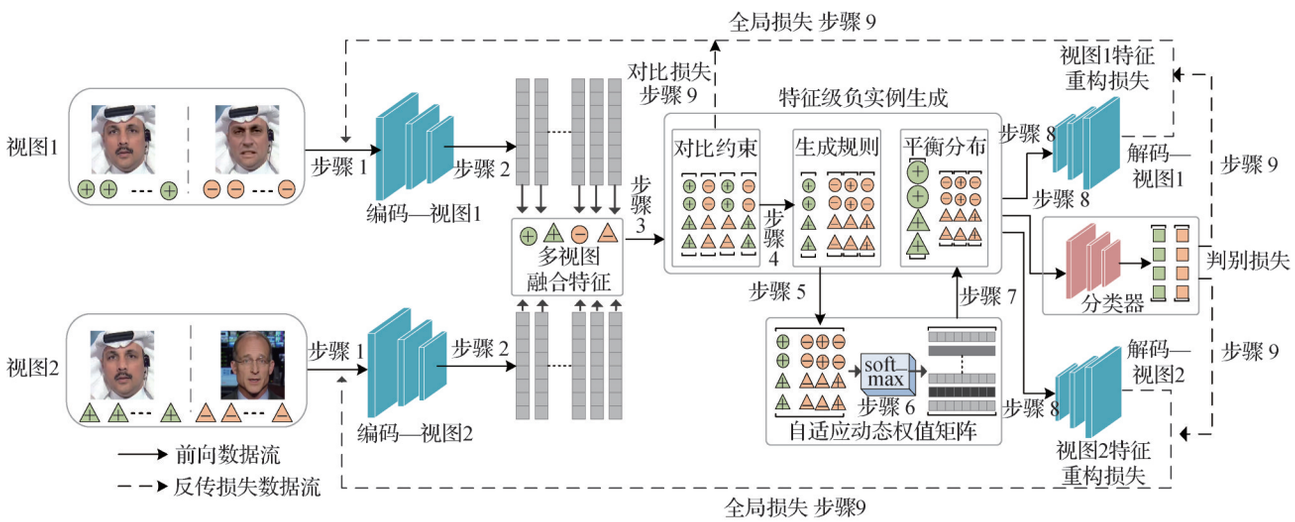


图 2　所提出模型（NIG-FFD）的框架和流程

Fig. 2　The framework and process of the proposed model (NIG-FFD)

1.1　对比约束下可判别多视图特征融合

由于单视图人脸特征难以全面表达样本语义信息，降低了模型本域与跨域上检测性能（Yang 等，2024）。本文受 SURE（robust multi-view clustering with incomplete information）模型（Yang 等，2023）启发，建立孪生自编码网络架构来探索多视图的人脸图像融合表达，构建信息互补的多视图融合特征来增强样本的可判别性。

获得的检测人脸数据集可以表示为 $\{\boldsymbol{X}^{(v)},\boldsymbol{Y}\}_{v=1}^{V}=\{(\boldsymbol{x}_1^{(v)},\boldsymbol{y}_1^{(v)}),(\boldsymbol{x}_2^{(v)},\boldsymbol{y}_2^{(v)}),\cdots,(\boldsymbol{x}_N^{(v)},\boldsymbol{y}_N^{(v)})\}_{v=1}^{V}$，其中，$v$ 表示视图特征索引，V 表示视图特征总数。本文取两个视图特征融合，即 $V=2$。$\boldsymbol{x}_i\in\mathbf{R}^{N\times d}$ 表示任意样本，$\boldsymbol{y}_i$ 是与样本对应的标签，N 表示实例的数量。

首先，利用孪生自编码网络提取不同视图特征。具体方法为

$$\boldsymbol{z}_i^{(1)}=f_1(\boldsymbol{x}_i^{(1)};\theta_i^{(1)}) \tag{1}$$

$$\boldsymbol{z}_j^{(2)}=f_2(\boldsymbol{x}_j^{(2)};\theta_j^{(2)}) \tag{2}$$

式中，f_1 和 f_2 代表用于提取多视图特征的编码器网络，$\theta_i^{(1)}$ 和 $\theta_j^{(2)}$ 为网络参数，$\boldsymbol{z}_i^{(1)}$ 和 $\boldsymbol{z}_j^{(2)}$ 为网络输出的特征，i 和 j 为样本索引。此时，可通过网络训练，获得样本多视图表示。

其次，由于典型的人脸伪造图像处理仅涉及小规模的替换和篡改，所以伪造人脸图像的全局特征与真实人脸特征相似。通过最大化类内特征的相似性、最小化类间特征的相似性构建对比损失，更好地学习具有较弱判别性的难样本。对比损失函数为

$$L^{\mathrm{ncl}}=\frac{1}{2P}\sum_{p=1}^{K}(tL_p^{\mathrm{pos}}+(1-t)L_p^{\mathrm{neg}}) \tag{3}$$

式中，P 为对比对的个数。定义 $t=1/0$，L_i^{pos} 和 L_i^{neg} 分别为正对和负对提供的损失。在构建对比学习对时，来自不同视图特征的同一样本为正样本，不同样本为负样本，即

$$t=\tau(z_i^{(1)},z_j^{(2)}) \tag{4}$$

式中，

$$\tau(z_i^{(1)},z_j^{(2)})=\begin{cases}1 & i=j\\ 0 & \text{其他}\end{cases} \tag{5}$$

对于正对 $(z_i^{(1)},z_i^{(2)})$，对比损失的目的在于最小化其潜在空间中的距离，计算为

$$L_i^{\text{neg}} = \max(m - d(z_i^{(1)}, z_j^{(2)}), 0)^2 \tag{6}$$

式中，

$$d(z_i^{(1)}, z_i^{(2)}) = \left\| z_i^{(1)} - z_i^{(2)} \right\|_2 \tag{7}$$

同时，对于负对$(z_i^{(1)}, z_j^{(2)})$，对比损失的目的在于最大化嵌入空间距离，计算为

$$L_i^{\text{neg}} = \max(m - d(z_i^{(1)}, z_j^{(2)}), 0)^2 \tag{8}$$

式中，m是一个边界值，它强制负对的距离在一个足够大的区间内。对于不同数据集的数据，最优的m值可能不同。因此，在训练前自适应地计算m值，计算为

$$m = \frac{1}{P_{\text{p}}} \sum d(z_i^{(1)}, z_i^{(2)}) + \frac{1}{P_{\text{n}}} \sum d(z_i^{(1)}, z_j^{(2)}) \tag{9}$$

式中，P_{p}和P_{n}分别代表正对和负对的个数。同时，m值只在训练开始前计算一次，在后续训练中其值保持不变。

最后，为了使模型学习更加充分，本文通过计算解码器输出与原始输入的差异建立重构损失，保留原始输入的重要特征信息，使模型更加注重学习有区分力的特征表示，具体为

$$L^{\text{ver}} = \frac{1}{2P} \sum_{i=1}^{N} \sum_{v=1}^{2} \left\| \boldsymbol{x}_i^{(v)} - g_v(cat(z_i^{(1)}, z_i^{(2)})) \right\|_2^2 \tag{10}$$

式中，g_v表示第v个视图特征的解码器，$cat(\cdot)$表示特征连接操作。

在对比损失L^{ncl}和重构损失L^{ver}的约束下，对正对样本$(z_i^{(1)}, z_i^{(2)})$进行特征融合，由于正对样本标签相同，即$\boldsymbol{y}_1 = \boldsymbol{y}_2$，所以融合后的特征集可以表示为

$$z_{\text{orig}} = (cat(z_i^{(1)}, z_i^{(2)}), \boldsymbol{y}) \tag{11}$$

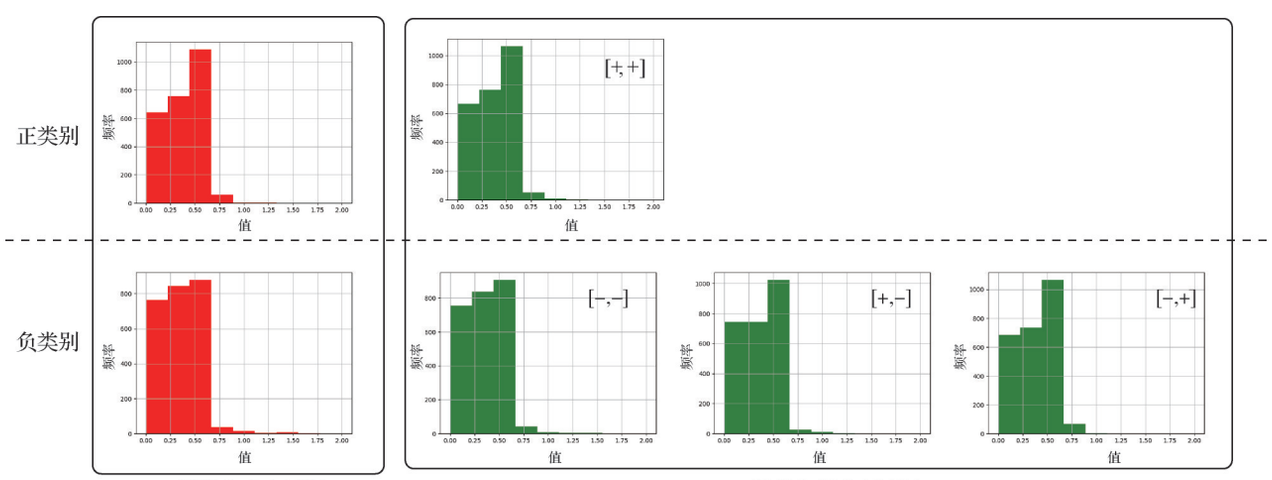
式中，$\boldsymbol{y} = \boldsymbol{y}_1 = \boldsymbol{y}_2$。正对样本提供原始数据集中样本的融合特征，为模型提供更具判别性的特征。

1.2 多样性负实例特征增广

为增强模型泛化性，确保在跨域数据集上具有令人满意的识别性能，本文提出了一种快速且有效的方法来生成具有多样性的伪造实例表达。该方法在模型中实现多视图特征融合的同时，实现了特征增广。随后，通过使用更多样化和丰富的伪造特征来训练识别模型，提高了模型识别伪造样本的泛化能力。

首先，定义融合样本的生成规则。由于直接为所有样本增广特征会导致在区分正负样本时产生混淆，当两个视图特征都是正样本时，融合特征的分布将与原始正样本相似。为了说明这个问题，绘制了具有不同标签的构造样本的网络输出特征直方图，如图3所示。图3(a)显示的直方图是经过孪生网络表达和融合后的原始样本的统计数据，图3(b)是增广特征的统计数据。如图所示，如果增广特征的两个视图特征标签为[+, +]，它们的特征表达与训练数据集中融合特征的正类相似。如果将这类样本用做生成的伪造样本，它将影响判别器的训练并降低后续的识别效果。当构造的特征向量为[+, −]、[−, +]、[−, −]时，则与原始训练数据集中正类直方图的相似性较低，而与负类直方图的相似性较高。

基于以上分析，本文在特征融合阶段同时实现了伪造负实例的多样性增广，为后续判别器的充分训练提供了前提和条件，提高了检测的泛化能力。



(a) 训练集融合特征　　(b) 增广实例融合特征

图3　增广和原始特征直方图的比较

Fig. 3　Comparison of augmented and original feature histograms

((a) fusion features of training dataset; (b) fusion features of augmentation instances)

特征向量筛选规则是在生成的样本中排除由[+, +]特征构造的负样本。

其次，利用筛选的样本构造伪造特征集。对于正类样本，采用式(11)中的特征融合方法，学习数据集中更具判别性的特征；对于负样本，基于上述分析，提出了增广伪造特征，具体为

$$z_{\text{aug}} = (cat(z_i^{(1)}, z_j^{(2)}), 0) \tag{12}$$

如图3所示，可以观察到生成的负样本特征分布和原始负样本特征分布基本一致。

最后，通过连接原始的训练样本和增广样本，得到全局的训练样本集。由式(11)和式(12)可得，全局训练样本集为

$$\boldsymbol{K} = (z_{\text{orig}}, z_{\text{aug}}) \tag{13}$$

上述模型与简单融合所有样本相比，有选择的伪造样本增广机制可以引入更多样化和难以判别的伪造特征，在防止正负样本分类混淆的同时，增强了模型的判别性与泛化性。

1.3 重要性样本加权的判别器构建

根据上述方式实现负实例特征增广，相比于原始负实例量大大增加，即原始数据集中负实例量与当前负实例量比值可表示为

$$\gamma = \frac{N_{\text{g}}}{N_{\text{p}} + N_{\text{p}} + N_{\text{g}} + N_{\text{g}}} \tag{14}$$

式中，N_{g}为原始数据集中负样本总量，N_{p}为原始数据集中正样本总量，此时，若$N_{\text{p}}:N_{\text{g}} = 1:1$，则原始样本类别分布由$\gamma = 1$变为$\gamma = 1/4$。该情况在后续判别器训练过程中将导致少类别样本欠学习、多类别样本过学习问题。为缓解该问题，构建具有重要性权值矩阵的判别器对样本进行平衡采样。

首先，初始化矩阵为每个类别样本设置不同的权重，使负样本根据其预测概率加权、正样本保持不变，从而在损失计算时近似达到类别均衡。通过负样本加权，可以引导模型更加关注正样本特征，防止分类决策边界偏向负样本，具体为

$$\tilde{\boldsymbol{K}} = \boldsymbol{Q}^{\text{T}}\boldsymbol{K} \tag{15}$$

式中，$\boldsymbol{Q} = \begin{bmatrix} a_1 & & & & \\ & a_2 & & & \\ & & a_3 & & \\ & & & \cdots & \\ & & & & a_{num(K)} \end{bmatrix}$，$num(\cdot)$为计算样本数量的函数，矩阵中元素值由样本标签及预测结果概率得出，计算为

$$a_i = \begin{cases} f_{\text{softmax}}(\boldsymbol{K}_i) & y = 0 \\ 1 & \text{其他} \end{cases} \tag{16}$$

其次，为了衡量预测概率分布与真实概率分布之间的距离，利用交叉熵损失作为分类损失函数对分类结果进行监督，计算为

$$L^{\text{cls}} = -[\boldsymbol{y}\log\tilde{\boldsymbol{y}} + (1 - \boldsymbol{y})\log(1 - \tilde{\boldsymbol{y}})] \tag{17}$$

式中，$\boldsymbol{y}$为样本真实标签，$\tilde{\boldsymbol{y}}$为模型预测标签。最小化交叉熵等效于最大化模型对正确类别的预测概率，通过最小化分类损失函数，可以训练出在给定输入条件下，尽可能准确地预测出目标类别的模型。

最后，模型训练总损失函数为

$$L = L^{\text{ver}} + \lambda L^{\text{cls}} + \delta L^{\text{ncl}} \tag{18}$$

式中，λ和δ为不同损失的平衡因子。

2 实　验

2.1 实验设置

2.1.1 数据集详细信息

为评估所提出方法的有效性，本文在3个公开的数据集FF++(Faceforensics++)(Rössler等，2019)、Celeb-DFv2(Li等，2020b)和DFDC(deepfake detection challenge)(Dolhansky等，2019)上进行实验验证、比较与分析。数据集的具体描述如表1所示。其中，Faceforensics++数据集根据不同压缩级别分为3个版本：c0(原始)，c23(高质量)和c40(低质量)，本文采用c23和c40进行实验。如果没有说明，默认采用高质量版本。Celeb-DFv2数据集广泛应用于测试模型的泛化能力，其伪造图像没有明显的深度伪造视觉伪影，在泛化性检测中十分具有挑战性。在实验中，随机选取100个真实视频和100个伪造视频，每隔30帧提取1幅图像。在DFDC数据集上，随机选取140个视频，每个视频提取20帧用于测试。

2.1.2 评价准则

在实验中，参考以往工作(Chollet，2017；Zhao等，2021)，本文利用准确率(accuracy score，ACC)和受试者工作特征曲线下面积(area under the receiver operating characteristic curve，AUC)对检测结果进行评估。

2.1.3 实验设置

本文利用DLIB(detect library)(Sagonas等，2016)进行人脸提取和对齐，并对训练和测试数据集

表 1　公开数据集

Table 1　Public datasets

数据集	真实视频	伪造视频
Faceforensics++	1 000（Youtube）	1 000（DeepFake, Face2Face, FaceSwap, NeuralTextures）
Celeb-DFv2	890（Youtube）	5 639（DeepFake）
DFDC	1 131（Actors）	4 119（未知）

中的所有样本将对齐的人脸尺寸调整为128×128像素。之后，将数据集经过预处理提取出每幅人脸图像的局部二值模式（local binary patterns，LBP）和方向梯度直方图算子（histogram of oriented gridients，HOG）特征，分别为256维和2 304维。在FF++和DFDC数据集上评估了δ取不同值时的AUC值变化，如图4所示。可以观察到，当δ = 0.01时，NIG-FFD可以取得最好的检测结果。

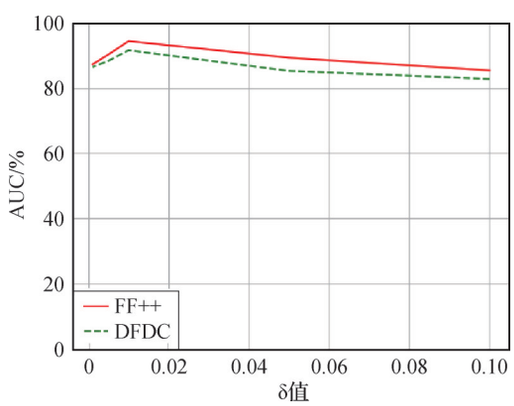


图4　δ取不同值时的AUC值评估

Fig. 4　AUC evaluation for different values of δ

此外，本文方法的特征网络和分类器都使用Adam优化器，并设置批次为128，学习率均为10^{-3}。具体网络结构如表2所示。模型已开源至：https://github.com/LNNU-computer-research-526/NIG-FFD。

表 2　网络结构

Table 2　Network architecture

网络层次	内容
第1层	Linear(256, 1 024) BatchNorm1d ReLU Dropout
第2层	Linear(1 024, 1 024) BatchNorm1d ReLU Dropout
第3层	Linear(1 024, 1 024) BatchNorm1d ReLU Dropout
第4层	Linear(1 024, 10) BatchNorm1d ReLU

2.2　跨域实验对比与分析

考虑在实际应用场景中，深度伪造图像可能进行不同程度的压缩，经过压缩等后处理的图像不容易被正确识别，为了检验模型对不同压缩级别的泛化性，本文在FF++数据集中的NT（NeuralTextures）伪造方法上训练模型，该方法将真实RGB图像的纹理风格迁移到深度图中，这种方法可以生成看起来非常逼真的伪造数据，其中风格迁移保留了场景本身的结构，但迁移了真实纹理，是一种高质量的深度图数据增强技术。之后，分别在高质量和低质量DF（DeepFake）、FS（FaceSwap）伪造方法上进行测试，评估的AUC值见表3（部分数据引自Chen等人（2022））。

从表3可以观察到，在高质量NT数据集训练时，大部分方法在高质量（high quality，HQ）数据集上表现良好，但当测试集为低质量（low quality，LQ）数据集时，其他方法的性能都有所下降。这是由于图像在经过重压缩处理后会使一些细节丢失，导致检测性能下降。这对模型的检测性能提出挑战，本文方法在低质量数据集仍能表现出优越的性能，无论是在检测精度上还是在跨不同压缩级别上的性能都远远超过其他方法。

比较来自5个DF和FS数据集的高质量和低质量样本的NIG-FFD和SLADD（self-supervised learning of adversarial dataset）（Chen等，2022）方法的可视化结果，如图5所示。图中，绿色框表示正确分类，

表 3　跨不同压缩级别的AUC值比较表

Table 3　AUC comparison table across different compression levels

/%

训练集	方法	测试集			
		LQ		HQ	
		DF	FS	DF	FS
NT	Xception(Chollet,2017)	58.73	51.50	77.02	71.86
	Face X-ray(Li等,2020a)	57.19	51.06	58.55	77.98
	F3-Net(Qian等,2020)	58.36	51.92	80.56	61.24
	RFM(Wang和Deng,2021)	55.82	51.62	79.80	63.94
	SRM(Luo等,2021)	55.51	52.94	83.80	79.58
	SLADD(Chen等,2022)	62.84	56.88	84.64	72.15
	NIG-FFD(本文)	**99.86**	**99.86**	**98.95**	**98.88**

注:加粗字体表示各列最优结果。

图5　NIG-FFD和SLADD方法在DF和FS数据集上的可视化结果

Fig. 5　Visualization results of NIG-FFD and SLADD methods on DF and FS datasets

红色框表示错误分类。NIG-FFD在具有挑战性的样本中表现更好。

为了进一步评估模型的泛化能力,在FF++ 4种伪造方法的HQ数据集上进行训练,在目前最具挑战的数据集Celeb-DFv2和DFDC上进行跨库检测,这两个数据集中的伪造图像没有明显的伪影,在跨库检测中很难被正确识别,因此广泛应用于跨库性检测实验。与目前最先进的方法进行比较,其AUC结果如表4所示。

从表4可以观察到,NIG-FFD在FF++c23数据集和FF++c40数据集上训练、在Celeb-DFv2和DFDC数据集测试的表现均超过表中其他方法。虽然UIA-ViT(Zhuang等,2022)在Celeb-DFv2数据集上表现出较好的跨库性,但其在DFDC数据集上的检测精度有所下降,这是因为DFDC数据集具有更加复杂多样的伪造样本,其包括多种不同伪造方式生成的样本,并且其分辨率较高,在跨库检测时更具有挑战性。但NIG-FFD在DFDC数据集上仍能表现出出色的跨库性。从表4可以观察到,基于频域的方法(Qian等,2020;Luo等,2021)在DFDC数据集上表现较好,这是由于频域的方法可以更好地描述图像的纹理特征,捕捉到更加细微的变化,但NIG-FFD仍领先于这些方法。由实验结果分析,NIG-FFD能得到具有可判别性的融合特征,同时增加训练样本的多样性,在跨库性上具有出色的表现。

本文同时在Celeb-DFv2和DFDC的5个数据集上评估了NIG-FFD和GS(Guo等,2023b)方法的可视化结果,如图6所示。图中,绿色框表示正确分

表4　基于AUC的跨域性比较

Table 4　Cross-domain comparison based on AUC

/%

方法	Celeb-DFv2	DFDC
Xception(Chollet, 2017)	66.91	67.93
F3-Net(Qian等, 2020)	71.21	72.88
Multi-attentional(Zhao等, 2021)	76.65	67.34
Face X-ray(Li等, 2020a)	74.20	70.00
RFM(Wang和Deng, 2021)	67.64	68.01
SRM(Luo等, 2021)	79.40	79.70
Local-relation(Chen等, 2021)	78.26	76.53
RECCE(Cao等, 2022)	77.39	76.75
LTW(Sun等, 2021)	77.14	74.58
UIA-ViT(Zhuang等, 2022)	82.41	75.80
GS(Guo等, 2023b)	84.97	81.65
NoiseDF(Wang和Chow, 2023)	75.89	63.89
NIG-FFD-HQ(本文)	**98.23**	**90.42**
NIG-FFD-LQ(本文)	**98.23**	**90.42**

注：加粗字体表示各列最优结果。

类，红色框表示错误分类。NIG-FFD在更具挑战性样本中表现更好。

基于以上结果进行统计分析，并绘制了NIG-FFD生成的融合特征的直方图以及跨数据集的原始伪造样本的网络输出特征直方图。图7(a)(c)分别显示了来自Celeb-DFv2和DFDC数据集的原始负样本经过融合后由孪生网络获得的样本表示的统计直方图。图7(b)(d)分别显示了来自Celeb-DFv2和DFDC数据集的增广伪造特征的统计直方图。如图7所示，图7(b)(d)中的特征表示与各自数据集中相应原始负样本的融合特征相似。为说明本文提出模型生成特征与真实训练集特征上的相似性，本文可视化了在Faceforensics++数据集上原始样本与生成样本特征，如图8所示。经过特征编码后，本文方法生成假脸特征与原始假脸特征可视化表达十分相似，然而，与真脸特征表达有较大区别。因此，本文提出的特征增广方法有效地为后续鉴别器的训练提供了前提和条件，增强了检测的泛化能力。

图6　NIG-FFD和GS方法在Celeb-DFv2和DFDC数据集上的可视化结果

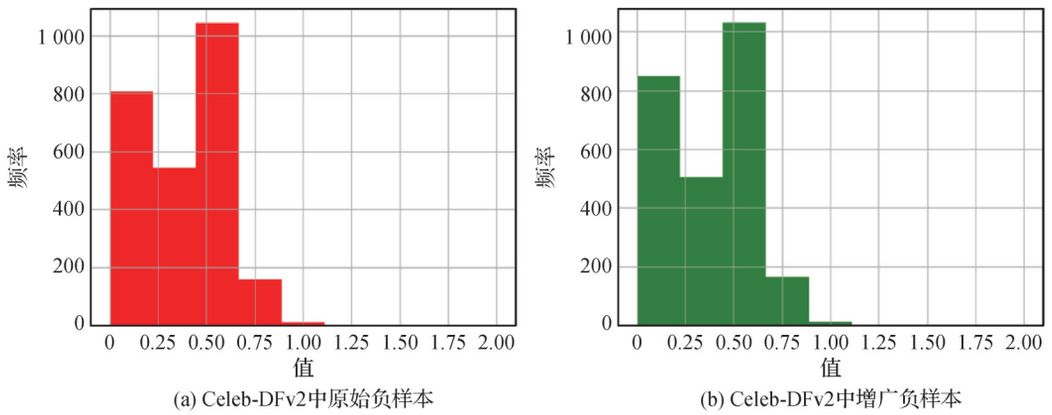
Fig. 6　Visualization results of NIG-FFD and GS methods on Celeb-DFv2 and DFDC datasets

在跨域性检测中，测试集图像的光照条件、分辨率和生成伪造图质量往往与训练集存在极大差别，如图9所示。本文比较了NIG-FFD和GS(Guo等, 2023b)的表现，当检测图9中测试集第4列图像时，GS(Guo等, 2023b)出现了决策失误，而NIG-FFD可以正确识别伪造图。可见NIG-FFD在检测与训练样本存在较大差异的样本时仍表现出色。

3.3　本域实验对比与分析

为了进一步验证所提出模型效果，本文在跨库性实验基础上，对比及分析了NIG-FFD在本库上的识别性能，实验结果如表5所示(部分数据引自Cao等人(2023))。

从表5可以看出，NIG-FFD在本库上仍然取得了良好的结果，在高质量和低质量的FF++数据集上都表现出优于其他方法，如：Qian等人(2020)、Zhao等人(2021)。虽然这些方法在高质量图像上表现良好，但在更具挑战的低质量图像上其性能有所衰退。这是由于图像经过重压缩处理会造成有效的纹理、边缘等信息的损失，而这些信息对检测来说是很重要的，因此，在低质量图像上的检测更具挑战性。而NIG-FFD受图像压缩处理的影响较小，在低质量图像上仍有很高的识别精度。TFMD(Cao等, 2023)方法在低质量数据上具有较好的表现，其提出的基于注意力机制的特征分解模块使模型可以关注到高频信息，在经过重压缩处理的低质量数据集上具有一定的鲁棒性。但NIG-FFD在低质量图像的检测精度上的表现仍优于该方法，这是由于本文的多视图特征融合增强了模型对可判别特征的学习，使模型在



(a) Celeb-DFv2中原始负样本　(b) Celeb-DFv2中增广负样本

(c) DFDC中原始负样本　(d) DFDC中增广负样本

图7　Celeb-DFv2和DFDC数据集中原始负样本和增广负样本的特征直方图

Fig. 7　Feature histograms of original negative samples and augmented negative samples in Celeb-DFv2 and DFDC datasets

((a) original negative samples in Celeb-DFv2; (b) augmented samples in Celeb-DFv2; (c) original negative samples in DFDC; (d) augmented samples in DFDC)

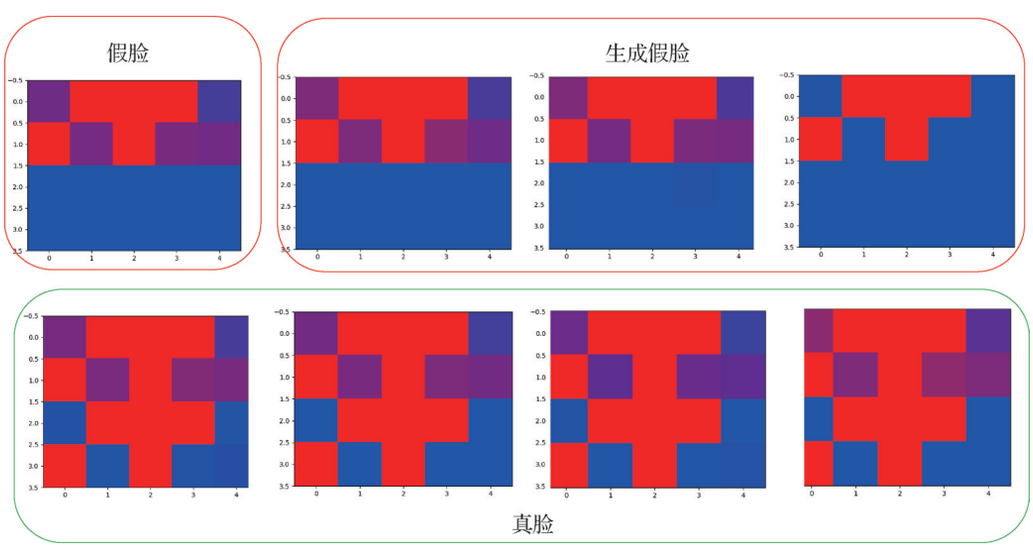


图8　可视化样本特征图

Fig. 8　Visualization sample features

低质量数据集上仍具有出色的表现。通过对本库结果的比较分析，NIG-FFD表现出强大的跨数据集性能，并在本库上取得了与其他方法相当的最佳性能。

3.4　消融实验分析

本节分析了提出模型训练依赖的3个损失、不同特征融合及多视图特征对模型检测性能的贡献。

图9　可视化样本图

Fig. 9　Visualization sample image

表5　在FF++数据集上的对比

Table 5　Comparison on FF++ dataset

/%

方法	FF++c23		FF++c40	
	ACC	AUC	ACC	AUC
Cozzolino等人(2017)	78.45	–	58.69	–
Bayar和Stamm(2018)	82.97	–	66.84	–
MesoNet(Afchar等,2018)	83.10	–	70.47	–
Xception(Chollet,2017)	95.73	–	81.00	–
Face X-ray(Li等,2020a)	–	87.40	–	61.60
Multi-attentional(Zhao等,2021)	96.37	98.97	86.95	87.26
F3-Net(Qian等,2020)	97.15	99.10	87.06	93.24
TFMD(Cao等,2023)	97.23	99.23	87.19	94.35
NoiseDF(Wang和Chow,2023)	84.36	93.99	–	–
NIG-FFD(本文)	**98.13**	**100.00**	**98.13**	**100.00**

注:加粗字体表示各列最优结果。"–"表示论文未提供相应实验结果。

在FF++数据集上对单一及组合损失设置消融实验,包含:分类损失、对比损失和重构损失,结果如表6所示。主要内容如下:

1)仅有分类损失时,模型识别精度偏低。

2)在1)的基础上增加了对比损失,得到更好的多视图融合特征,增强了模型对可判别信息学习,使分类精度提升至98.02%。

3)在2)的基础上增加了重构损失,通过无监督样本特征一致性进一步提升了模型的检测精度,使其达到98.13%。

通过对不同损失消融实验的结果观察与分析,利用分类损失、对比损失和重构损失同时约束模型训练时检测结果最好,可进一步说明本文所提出优化损失的有效性。

本文中选择用于描述图像纹理信息的LBP特征和描述梯度信息的HOG特征作为两个编码网络的输入。由于LBP与HOG两个不同视角特征在样本描述上具有一定的互补性,在后续的伪造检测中获得了更高的精度。为说明相似多视角融合特征对检测精度的影响,补充了在样本描述上十分近似的HOG与PHOG(pyramid HOG)特征作为编码网络的输入,且其他实验设置不变。HOG与LBP融合特征相比于HOG与PHOG融合特征,在FF++、Celeb-DFv2和DFDC测试数据集上均获得了更好的检测结果,具体如表7所示。该结果很好地验证了互补融合特征相比于近似融合特征在人脸检测任务中的有效性。

表6　损失消融实验结果

Table 6　Results of loss ablation experiments

L^{cls}	L^{ncl}	L^{ver}	ACC/%	AUC/%
√	–	–	96.20	96.19
√	√	–	98.02	100.00
√	√	√	98.13	100.00

注:"√"表示采用相应损失,"–"表示未采用。

表 7　在 FF++上训练的 AUC 值比较

Table 7　AUC comparison of models trained on FF++

/%

方法	FF++	Celeb-DFv2	DFDC
HOG-PHOG	41.05	41.02	42.17
HOG-LBP(本文)	**100.00**	**98.23**	**90.42**

注：加粗字体表示各列最优结果。

为进一步验证本文所提多视图融合特征的有效性，在 FF++数据集上对比了仅单视图 LBP 与 HOG 特征的检测精度，实验结果如表 8 所示。由于伪造在人脸图像的细节部分，仅依赖单视图 LBP 纹理特征，难以有效识别。HOG 特征对光照变化和局部几何变化相对鲁棒，能够较好地捕捉图像边缘信息，使检测精度达到 53.77%。利用本文提出的多视图融合特征可充分挖掘不同视图丰富的一致性与互补性信息，使检测精度达到 98.13%。通过对不同视图消融实验结果分析，说明了本文模型的有效性。

表 8　不同视图消融实验结果

Table 8　Results of view ablation experiments

/%

LBP	HOG	ACC	AUC
√	–	49.20	56.34
–	√	53.77	60.39
√	√	98.13	100.00

注：“√”表示采用对应视图，“–”表示未采用。

4　结　论

针对跨域环境下伪造人脸识别模型构建的挑战，本文提出多样性负实例生成的跨域人脸伪造检测框架。通过在本域样本标签监督过程中训练孪生网络架构，获得多视图特征表达，并通过对比损失约束，强化难样本特征，提升后续判别模型精度。同时，建立负样本特征增广模块，生成多样性潜在负样本特征表达，提升判别模型泛化性，进一步，引入重要性权值矩阵，避免由于生成样本导致的样本分布不平衡产生的少类别样本欠学习问题。本文在 3 个流行跨域数据集上验证了模型有效性，相比于目前先进的模型，识别 ACC 与 AUC 值获得了显著提升，同时，本文在 1 个流行的数据集上验证了模型在本域数据集中的识别有效性。然而，本文方法仍存在针对低质量样本跨域识别精度不够优化等问题，其原因可能在于特征描述网络难以有效表达低质量样本可判别细节特征。本文未来工作重点集中在对低质量伪造样本的判别特征挖掘及表达上。

参考文献(References)

Afchar D, Nozick V, Yamagishi J and Echizen I. 2018. MesoNet: a compact facial video forgery detection network//2018 IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China: IEEE: 1-7 [DOI: 10.1109/WIFS.2018.8630761]

Bayar B and Stamm M C. 2018. Constrained convolutional neural networks: a new approach towards general purpose image manipulation detection. IEEE Transactions on Information Forensics and Security, 13(11): 2691-2706 [DOI: 10.1109/TIFS.2018.2825953]

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: 4113-4122 [DOI: 10.1109/CVPR52688.2022.00408]

Cao Y G, Chen J Z, Huang L Q, Huang T Q and Ye F. 2023. Three-classification face manipulation detection using attention-based feature decomposition. Computers and Security, 125: #103024 [DOI: 10.1016/j.cose.2022.103024]

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. 2021. Local relation learning for face forgery detection//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 1081-1088 [DOI: 10.1609/aaai.v35i2.16193]

Chollet F. 2017. Xception: deep learning with depthwise separable convolutions//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 1800-1807 [DOI: 10.1109/CVPR.2017.195]

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. Philadelphia, USA: ACM: 159-164 [DOI: 10.1145/3082031.3083247]

Ding F, Kuang R S, Zhou Y, Sun L, Zhu X G and Zhu G P. 2024. A survey of deepfake and related digital forensics. Journal of Image and Graphics, 29(2): 295-317 (丁峰,匡仁盛,周越,孙珑,朱小刚,

朱国普. 2024. 深度伪造及其取证技术综述. 中国图象图形学报, 29(2): 295-317[DOI: 10.11834/jig.230088]

Dolhansky B, Howes R, Pflaum B, Baram N and Ferrer C C. 2019. The deepfake detection challenge (DFDC) preview dataset[EB/OL]. [2024-03-27]. https://arxiv.org/pdf/1910.08854.pdf

Guo Y, Zhen C and Yan P F. 2023a. Controllable guide-space for generalizable face forgery detection//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 20761-20770[DOI: 10.1109/ICCV51070.2023.01903]

Guo Z Q, Yang G B, Wang D W and Zhang D Y. 2023b. A data augmentation framework by mining structured features for fake face image detection. Computer Vision and Image Understanding, 226: #103587[DOI: 10.1016/j.cviu.2022.103587]

Khormali A and Yuan J S. 2024. Self-supervised graph transformer for deepfake detection. IEEE Access, 12: 58114-58127[DOI: 10.1109/ACCESS.2024.3392512]

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, Yang X, Sun P, Qi H G and Lyu S. 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]

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]

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]

Sagonas C, Antonakos E, Tzimiropoulos G, Zafeiriou S and Pantic M. 2016. 300 Faces in-the-wild challenge: database and results. Image and Vision Computing, 47: 3-18[DOI: 10.1016/j.imavis.2016.01.002]

Sun K, Liu H, Ye Q X, Gao Y, Liu J Z, Shao L and Ji R R. 2021. Domain general face forgery detection by learning to weight//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 2638-2646[DOI: 10.1609/aaai.v35i3.16367]

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 T Y and Chow K P. 2023. Noise based deepfake detection via multi-head relative-interaction//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 14548-14556[DOI: 10.1609/aaai.v37i12.26701]

Yang M X, Li Y F, Hu P, Bai J F, Lyu J C and Peng X. 2023. Robust multi-view clustering with incomplete information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1): 1055-1069[DOI: 10.1109/TPAMI.2022.3155499]

Yang Y J, Qin H C, Zhou H, Wang C C, Guo T Y, Han K and Wang Y H. 2024. A robust audio deepfake detection system via multi-view feature//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea (South): IEEE: 13131-13135[DOI: 10.1109/ICASSP48485.2024.10446560]

Yu C E, Zhang X H, Duan Y X, Yan S B, Wang Z H, Xiang Y, Ji S L and Chen W Z. 2024. Diff-ID: an explainable identity difference quantification framework for deepfake detection. IEEE Transactions on Dependable and Secure Computing, 21(5): 5029-5045[DOI: 10.1109/TDSC.2024.3364679]

Zhang X, Karaman S and Chang S F. 2019. Detecting and simulating artifacts in GAN fake images//2019 IEEE International Workshop on Information Forensics and Security (WIFS). Delft, Netherlands: IEEE: 1-6[DOI: 10.1109/WIFS47025.2019.9035107]

Zhao H Q, Wei T Y, Zhou W B, Zhang W M, Chen D D and Yu N H. 2021. 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]

Zhuang W Y, Chu Q, Tan Z T, Liu Q K, Yuan H J, Miao C T, Luo Z X and Yu N H. 2022. UIA-ViT: unsupervised inconsistency-aware method based on vision transformer for face forgery detection//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 391-407[DOI: 10.1007/978-3-031-20065-6_23]

作者简介

张晶,女,副教授,主要研究方向为机器学习。
E-mail: zhangjing_0412@lnnu.edu.cn
孙芳,通信作者,女,副教授,主要研究方向为人脸伪造检测。
E-mail:sunfang@lnnu.edu.cn
许盼,女,硕士研究生,主要研究方向为人脸伪造检测。
E-mail:xp000924@163.com
刘文君,男,硕士研究生,主要研究方向为深度学习。
E-mail:lwjns6@163.com
郭晓萱,女,硕士研究生,主要研究方向为人脸伪造检测。
E-mail:985063050@qq.com