中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-16

论文引用格式:YUE Shutong, LIU Chunxiao. Differential Reconstruction-driven Residual Guidance for Face Forgery Detection[J/OL]. Journal of Image and Graphics, XXXX:1-16. DOI: 10.11834/jig.260178.(岳书同, 刘春晓. 差异化重建驱动的残差引导人脸伪造检测[J/OL]. 中国图象图形学报,XXXX:1-16. DOI: 10.11834/jig.260178.)[DOI:10.11834/jig.260178]

差异化重建驱动的残差引导人脸伪造检测

岳书同[1],刘春晓[1,2*]

1. 浙江工商大学计算机科学与技术学院,杭州　310018; 2. 浙江省大数据与未来电子商务技术重点实验室,杭州　310018

摘　要:**目的**　随着人脸伪造技术的快速发展与广泛传播,由其引发的虚假信息泛滥与身份冒用诈骗等社会问题日益严峻。现有的基于人脸图像重建的伪造检测方法仅进行真实或伪造人脸的单视角重建,未显式放大二者重建前后的差异,导致网络模型的检测准确率和泛化性能提升有限。针对上述问题,提出一种差异化重建驱动的残差引导人脸伪造检测方法,扩大真伪人脸重建前后的差异,显著提升了模型检测性能。**方法**　首先,为了显式放大真伪人脸重建前后的差异,提出一种对比差异化重建网络(Contrastive Differential Reconstruction Network,CDRNet),分别为真实与伪造人脸构建清晰与模糊图像的重建目标,提升整体网络模型对真伪人脸的辨别能力。其次,考虑到现有检测方法对残差图的引导信息利用不充分等问题,设计了残差双域引导模块(Residual Dual-Domain Guidance Module,RDDGM),深度融合图像的空间域与高频域特征,并利用重建残差信息引导融合后的双域特征,增强了网络模型捕捉细微伪造痕迹的能力。此外,为了促使模型学习不同伪造方法之间的通用伪造特征,设计了文本感知损失模块(Text-Aware Loss Module,TALM),通过引入文本模态信息的引导,进一步优化对比差异化重建结果,大幅提升了网络模型对未知伪造方式的泛化性能。**结果**　在域内实验中,与性能最好的对比方法相比,该方法的准确率(accuracy,ACC)与曲线下面积(area under the curve,AUC)指标分别提升2.83%和1.75%。在跨域实验中,该方法在5个公开测试集上与13种典型方法进行性能测试与比较,平均AUC指标提高1.75%。**结论**　本文在人脸伪造检测中创新性结合对比学习与图像差异化重建,显著提升了模型对未知伪造方式的检测准确率,在多个基准测试中性能优于已有方法。

关键词:深度伪造检测;人脸伪造检测;多任务学习;对比差异化重建;残差双域引导

Differential Reconstruction-driven Residual Guidance for Face Forgery Detection

YUE Shutong[1], LIU Chunxiao[1,2*]

1. School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou 310018, China; 2. Zhejiang Key Laboratory of Big Data and Future E-Commerce Technology, Hangzhou 310018, China

Abstract: **Objective** With the rapid advancement and widespread proliferation of facial forgery technologies, social issues such as the dissemination of misinformation and identity fraud have become increasingly severe. Regarding existing forgery

收稿日期:2026-04-06;**修回日期**:2026-06-20

*通信作者:刘春晓　cxliu@mail.zjgsu.edu.cn

基金项目:国家自然科学基金(U25A20440);浙江省自然科学基金(LY24F020004);浙江省重点研发计划项目(2023C01039);国家级大学生创新训练计划项目(202510353037)

Supported by: National Natural Science Foundation of China (U25A20440); Zhejiang Provincial Natural Science Foundation of China (LY24F020004); Key R&D Program of Zhejiang Province (2023C01039); National College Students Innovation and Entrepreneurship Training Program (202510353037)

detection methods based on face image reconstruction, most employ multi-task learning to reconstruct either real or fake faces in isolation, failing to explicitly magnify the reconstruction discrepancies between the two. Consequently, the detection accuracy and generalization performance of these network models are limited. To address these issues, a method called "differential reconstruction-driven residual guidance for face forgery detection" is proposed. **Method** Firstly, to explicitly magnify the differences between real and fake faces during the image reconstruction process, a contrastive differential reconstruction network (CDRNet) is introduced. This method constructs clear and blurred image reconstruction targets for real and fake faces, respectively, guiding the network to focus on the high-frequency regions of faces with obvious forgery traces during reconstruction, thereby enhancing the overall model's ability to distinguish between real and fake faces. Secondly, addressing the issue that existing detection methods fail to fully exploit the guidance information from reconstructed residual maps, a residual dual-domain guiding module (RDDGM) is designed. This module deeply integrates spatial and high-frequency domain features, utilizing the reconstructed residual information to guide the fused dual-domain features, thereby enhancing the network model's capability to capture subtle forgery traces. In addition, in order to enable the model to learn the common forgery features among different forgery methods, a text-aware loss module (TALM) is introduced. Through the guidance of text modal information, the results of contrastive differential reconstruction are further optimized, and the generalization performance of the network model for unknown forgery methods is significantly improved. The main contributions of this paper are summarized as follows: 1) In order to explicitly magnify the differences between real and fake faces before and after reconstruction, the contrastive differential reconstruction network (CDRNet) was designed to construct differential reconstruction targets for real and fake faces respectively. The model's ability to distinguish between real and fake human faces has been enhanced. 2) In order to enhance the ability of the network model to capture subtle forgery traces, the residual dual-domain guidance module (RDDGM) is proposed, which uses the reconstructed residual information to guide the fusion features of the high-frequency domain and the spatial domain of the image. Fully explore the subtle forgery traces in the image domain and the high-frequency domain. 3) In order to enhance the generalization ability of the network model for unknown forgery methods, the text-aware loss module (TALM) is proposed. Text modal information is introduced to further optimize and compare the differential reconstruction results, promoting the model to learn the common forgery features among various forgery methods. The generalization ability of the network model has been significantly enhanced. 4) Experimental results demonstrate that the proposed method achieves highly competitive performance in both in-domain and cross-domain evaluations across multiple datasets, including FaceForensics++ (FF++), Celeb-DF V1 and V2, DeepFake Detection Challenge (DFDC), DeepFake Detection Challenge Preview (DFDCP), and Deepfake Detection (DFD). Based on the Dlib library, this paper extracts 32 facial frames from each video in the training set and 64 frames from each video in the test set. All images are uniformly resized to 224×224, and their pixel values are normalized to before being fed into the network. In terms of evaluation metrics, following prior research, this study primarily adopts accuracy (ACC) and area under the ROC curve (AUC) to assess network performance. Experimental verification confirms that the AdamW optimizer is used for training the network model, with the initial learning rate and batch size set to 1E-4 and 8, respectively. The weights of the image encoder are initialized using an EfficientNet-B4 model pre-trained on ImageNet. The radius of the filter frequency domain mask in CDRNet is set to 16. During training, multiple data augmentation strategies are employed, including random horizontal flipping, small-angle rotation, random cropping, scaling, and color jittering. The proposed method is implemented based on the PyTorch framework, and the model is trained using a single NVIDIA GeForce RTX 2080Ti GPU. **Result** In intra-domain experiments, compared with the best-performing comparison method, the proposed method improves ACC and AUC by 2.83% and 1.75%, respectively, thereby demonstrating superior performance in identifying forgeries within the same data distribution. In cross-domain experiments, the method was rigorously tested on 5 public datasets and compared with 13 typical methods, achieving a 1.75% improvement in average AUC. Comprehensive ablation studies further demonstrate that the proposed modules, including CDRNet, RDDGM, and TALM, all significantly contribute to enhancing the overall detection performance and generalization of the face forgery detection model. **Conclusion** This work introduces a novel integration of contrastive learning and image differential reconstruction for face forgery detection, substantially improving the model's generalization accuracy on unseen forgery methods. Specifically, the CDRNet amplifies the discrepancies between real and forged faces during the reconstruction process to bol-

ster discriminative capability, while the RDDGM leverages reconstruction residual information to guide dual-domain feature fusion, thereby improving the perception of subtle forgery traces. Furthermore, the TALM introduces text modality information to learn generic forgery features across different methods, enhancing generalization to unknown forgeries. Although experimental results demonstrate that the proposed method outperforms existing approaches in both in-domain and cross-domain scenarios, it currently faces challenges regarding inference efficiency due to model complexity, and the use of fixed text prompts limits the full potential of TALM's cross-modal guidance. Consequently, future work will focus on lightweight network optimization and the construction of a dynamic text prompt library to strengthen cross-modal correlations, aiming to further improve the model's efficiency and practicality.

Key words: deepfake detection; face forgery detection; multi-task learning; contrastive differential reconstruction; residual dual-domain guidance

论文引用格式:【DOI:10.11834/jig.260178】

0 引　言

人脸图像是生物特征识别、身份验证与视觉交互的核心载体,广泛应用于智能手机解锁、金融支付核验及安防监控系统等关键领域。其可靠性与真实性直接关系到个人信息与财产安全。

然而,随着深度学习生成模型的快速发展,通过对真实人脸图像进行局部篡改、面部迁移或整图生成等手段,可轻易合成人眼难以区分的伪造人脸图像。伪造人脸不仅被用于制作虚假证件、合成不雅视频,还被用于伪造政治人物、制造虚假新闻,对社会构成严重的威胁。为了保护人脸图像,迫切需要研究出准确率高、泛化性能好的人脸伪造检测方法。

已有的人脸伪造检测方法包括基于训练数据增强的检测方法和基于网络框架优化的检测方法。基于训练数据增强的方法通过构造检测难度更高的训练样本,以提升模型泛化能力,然而所生成的增强样本难以覆盖所有伪造方式的图像特征。基于网络框架优化的方法通过设计更合理的网络架构与训练策略,以提升模型检测性能。其中,基于人脸图像重建的伪造检测方法已得到广泛应用。其核心思想是让真实与伪造人脸在图像重建过程中暴露出差异性。然而,现有方法仅针对真实人脸或伪造人脸图像进行单视角重建,导致模型对真伪人脸重建前后差异较小。此外,现有方法仅将重建残差图像用于引导图像空间域特征,忽略了高频域特征中蕴含的细微伪造痕迹,从而限制了模型性能的进一步提升。

为了解决上述问题,本文提出差异化重建驱动的残差引导人脸伪造检测方法。该方法通过放大真伪人脸重建前后的差异,并深度融合高频域与空间域特征,从而实现对细微伪造痕迹的有效捕捉。此外,本文通过引入文本模态信息指导重建网络,进一步提升了模型的泛化能力。本文的主要贡献如下:

1)为了显式放大真实与伪造人脸在重建前后的差异,设计了对比差异化重建网络(Contrastive Differential Reconstruction Network,CDRNet),分别为真实与伪造人脸构造差异化的重建目标,提升了模型对真伪人脸的辨别能力。

2)为了增强网络模型对细微伪造痕迹的捕捉能力,提出了残差双域引导模块(Residual Dual-Domain Guidance Module,RDDGM),利用重建残差信息对图像高频域与空间域的融合特征作引导,充分挖掘图像域与高频域中的细微伪造痕迹。

3)为了提升网络模型对未知伪造方式的泛化能力,提出了文本感知损失模块(Text-Aware Loss Module,TALM),引入文本模态信息进一步优化对比差异化重建结果,促使模型学习多种伪造方式间的通用伪造特征,显著提升了网络模型的泛化能力。

1 相关工作

1.1 基于训练数据增强的检测方法

基于训练数据增强的方法通过特定的图像混合策略生成检测难度更高的训练样本,以提升网络模型的泛化能力。其中,Li 等人(2020a)和 Zhao 等人(2021)通过混合两张不同的人脸图像,有效扩充了伪造人脸样本的多样性。Shiohara 和 Yamasaki(2022)提出在空间域中对同一张人脸图像进行自混合的方法,生成更加具有挑战性的训练样本。Zhou 等人(2024)进一步在频率域融合不同频率成分生成自混合伪造图像,有效弥补了仅在空间域中自混合方法的局限。Ma 等人(2025)对已有自混合框架进

行改进,生成具有通用伪造痕迹的人脸图像,进一步提升了模型的泛化能力。王诗雨等人(2025b)通过随机权重混合不同伪造方式的图像,从而得到多元软混合伪造样本。然而,上述方法所生成的增强样本,其分布难以完全覆盖各类伪造类型的特征空间,因此对模型的泛化能力提升依然有限。

1.2 基于网络框架优化的检测方法

基于网络框架优化的检测方法通过设计更优的网络架构与训练策略,以提升模型的检测性能。此类方法可进一步细分为非基于人脸图像重建与基于人脸图像重建的检测方法。

1.2.1 非基于人脸图像重建的检测方法

非基于人脸图像重建的检测方法通过优化网络的特征提取能力,以提升模型对伪造痕迹的感知能力。其中,Luo 等人(2021)则利用空间丰富模型(Spatial Rich Model,SRM)提取的高频噪声特征来指导网络训练。Qian 等人(2020)利用离散余弦变换(Discrete Cosine Transform,DCT)提取频域信息并进行特征分析。Liu 等人(2021)结合空间域图像与经 DCT 变换得到的相位谱,有效捕获了伪造人脸图像中的上采样伪影。Wang 等人(2023)引入图神经网络(Graph Neural Network,GNN)实现图像空间域特征与频率域特征的深度融合。冯才博等人(2024)提出纯净图像块的概念,并估计残差图以聚焦伪造痕迹。

然而,上述方法主要关注模型对细微伪造痕迹的捕捉能力,而对模型泛化性能的提升考虑不足。为此,Yan 等人(2023)通过设计网络框架解耦出内容特征、特定伪造特征与通用伪造特征以提高网络泛化性能。Huang 等人(2023)提出隐式身份的概念,通过检测人脸内部与外部区域的不一致性来挖掘伪造线索。Yan 等人(2024b)和 Choi 等人(2024)在潜在空间中增强特征表示,以扩展训练样本在特征空间中的多样性。Kim 等人(2024)认为,模型在训练过程中会过拟合到面部身份特征,因此在模型提取特征过程中抑制人脸身份信息。Luo 等人(2024)以预训练并冻结的 ViT 框架为基础,设计了全局局部伪造感知模块,挖掘图像中细粒度的伪造信息。Zhang 等人(2025)改进 Transformer 框架并有效结合自蒸馏技术,大幅提升了检测模型的泛化性能。

但是,这些方法仅依赖于单一图像模态,未能引入信息更丰富、泛化能力更强的文本模态,容易导致模型过拟合。为此,Lin 等人(2025)基于预训练的 CLIP 视觉语言模型,将模型重编程技术引入人脸伪造检测任务,引入文本模态并显著降低了网络参数量。Tan 等人(2025)结合对比学习与 Lora 微调技术,在 CLIP 框架的基础上引入类别通用提示,有效提升了模型泛化能力。Cui 等人(2025)基于 CLIP 构建双向跨模态融合框架,提取多层视觉特征并结合专用文本嵌入,显著提升了跨域人脸伪造检测的泛化性。近期,Fu 等人(2025)通过在图像空间与潜在空间中进行扰动,有效缓解了模型的内容偏差与位置偏差问题。Shi 等人(2025a)在真实人脸上遮掩部分区域并训练网络恢复被遮掩部分,以学习真实人脸的分布。Li 等人(2025)则提出一种多尺度特征级解耦框架,在特征级别分离内容信息与伪造信息。然而,这些方法主要依赖于图像空间域特征,忽略了频域信息中蕴含的伪造线索。

1.2.2 基于人脸图像重建的检测方法

近年来,基于人脸图像重建的伪造检测方法受到广泛关注。其中,Cao 等人(2022)采用多任务学习策略,对真实人脸图像进行重建,以学习其分布特征。Cao 等人(2024)进一步在空间域与频率域中联合重建真实人脸图像,以学习更具判别力的真实人脸表示。然而,这两种方法仅针对真实人脸进行重建,未针对伪造人脸设计差异化的重建目标,导致重建网络难以放大真伪人脸之间的差异。Wang 等人(2025a)设计了双解码器结构,分别重建真实和伪造人脸,从而学习图像的内在不变特征。但是,该方法仅依赖双分支独立重建,同样缺乏对两类图像差异化重建的显式建模,限制了网络模型对真伪差异的捕捉能力。

2 本文方法

现有基于重建的检测方法仅在单类样本上施加重建约束。然而,如图 2(a)所示,最新的基于重建的检测算法 FakeDiffer(Wang 等人,2025a)对真伪人脸均产生较低的重建误差。针对上述问题,本文提出对比差异化重建策略:一方面通过自重建约束缩小真实人脸的重建误差;另一方面引入差异化重建约束与对比学习机制,主动扩大伪造人脸的重建误差。如图 2(b)所示,本文方法最终使两类样本的损

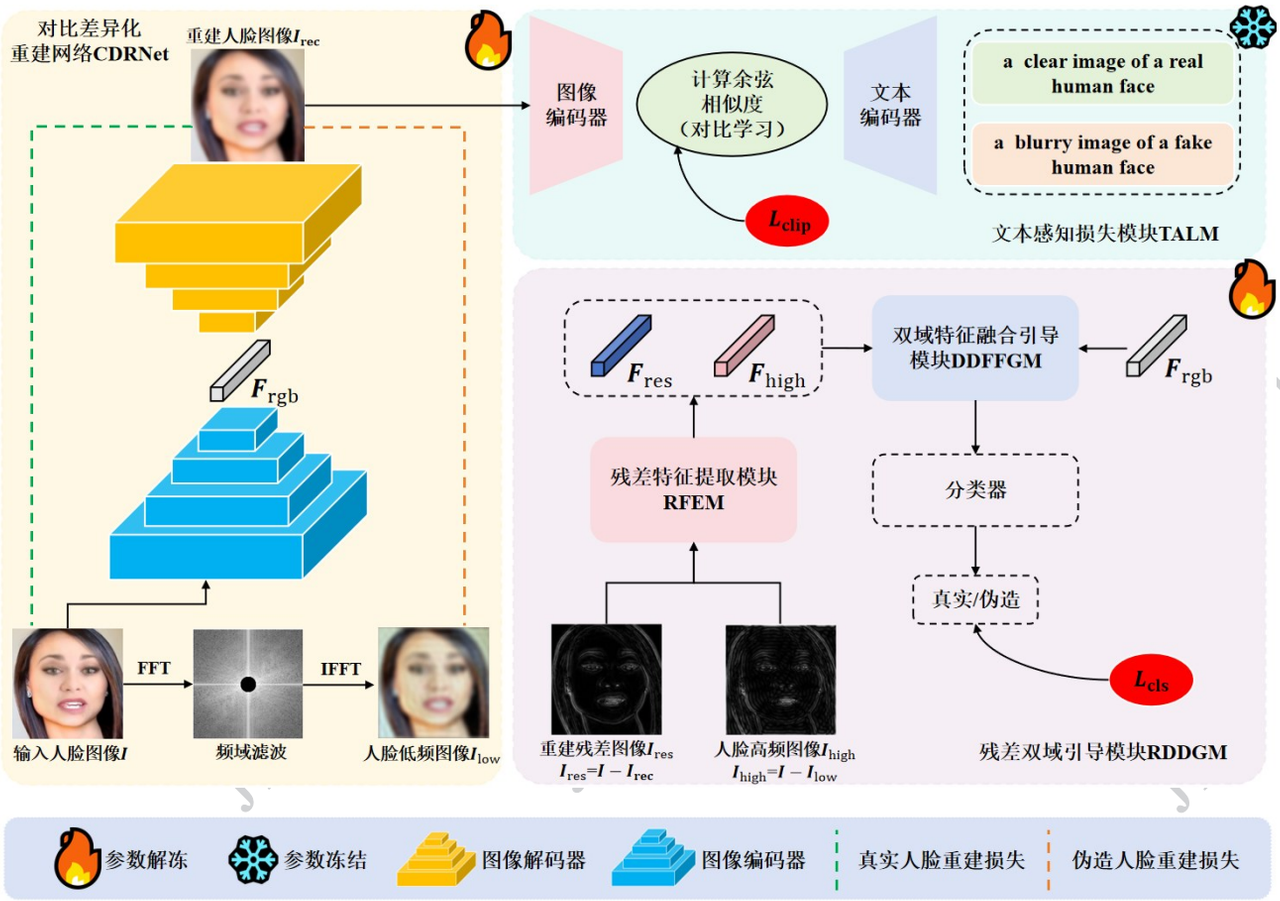
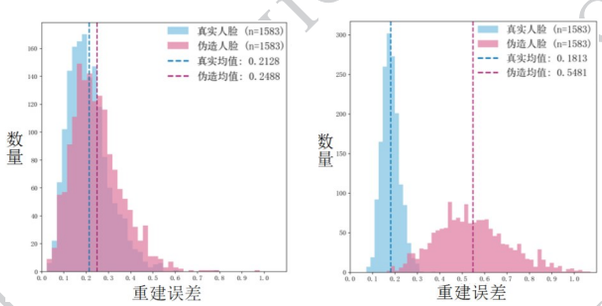


图1　本文方法网络框架

Fig. 1　Our network framework

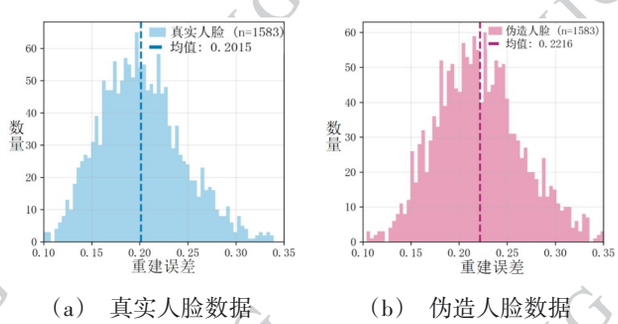


（a）　现有方法（FakeDiffer）　　（b）　本文方法

((a) Existed method (FakeDiffer); (b) Proposed method)

图2　在FF++数据集上网络重建误差直方图

Fig. 2　Histograms of network reconstruction errors on the FF++ dataset



（a）　真实人脸数据　　（b）　伪造人脸数据

((a) Real facial data; (b) Fake facial data)

图3　在FF++数据集上真伪单独重建误差直方图

Fig. 3　Histograms of the reconstruction errors of the real and fake samples with separate reconstruction model on the FF++ dataset

失分布实现清晰解耦，显著提升了检测性能。我们选择拉近真实人脸重建误差，拉远伪造人脸重建误差，而非相反。具体原因在于：我们通过图3的真伪单独重建损失分布实验发现，真实人脸单独重建的损失不仅均值更低，且分布更为集中。基于这一预实验发现，我们引入差异化重建与对比学习，以进一步扩大真伪之间重建的差异。

2.1　本文网络框架

本文提出的网络框架如图1所示。首先，将原人脸图像$\boldsymbol{I}\in\mathbf{R}^{3\times H\times W}$输入CDRNet，生成对应的重建

人脸图像$\boldsymbol{I}_{\text{rec}} \in \mathbf{R}^{3\times H\times W}$。随后，一方面将人脸高频图像$\boldsymbol{I}_{\text{high}} \in \mathbf{R}^{3\times H\times W}$、空间特征$\boldsymbol{F}_{\text{rgb}} \in \mathbf{R}^{C_1\times H_1\times W_1}$以及$\boldsymbol{I}_{\text{rec}}$输入RDDGM，输出经引导的双域融合特征，并输入分类器获得最终的预测结果。另一方面，将$\boldsymbol{I}_{\text{rec}}$输入TALM，通过文本模态信息的引导进一步优化CDRNet。在模型训练策略上，本文通过调整不同任务的损失权重系数，使网络整体性能达到最优。

2.1.1　对比差异化重建网络CDRNet

为了放大真实与伪造人脸在重建前后的差异，本文设计了CDRNet，显著提升了模型对真伪人脸的检测准确率。具体而言，Lee等人(2024)发现，真实人脸图像的高频细节极为复杂，无论基于卷积网络的GAN(Goodfellow等，2014)模型还是当前主流的扩散模型，在人脸生成过程中均存在高频纹理建模能力不足的局限。由图4可见，真实与伪造人脸之间的细微差异可由图像高频信息捕捉。因此，CDRNet将人脸图像分解为高频与低频部分。高频部分输入到RDDGM中参与特征融合，用于捕捉细微伪造痕迹；低频部分作为伪造人脸图像的重建目标。该策略迫使网络对真实与伪造人脸进行差异化重建，并关注高频域中的伪造痕迹，增强对真伪人脸的判别能力。具体地，对于输入人脸图像$\boldsymbol{I}$，首先通过快速傅里叶变换将其转换至频域，随后采用半径为r的频域掩码进行滤波，并通过逆傅里叶变换得到对应的低频图像$\boldsymbol{I}_{\text{low}}$。随后，将$\boldsymbol{I}$输入至自编码器中得到重建图像$\boldsymbol{I}_{\text{rec}}$。该自编码器的编码器采用基于EfficientNet-B4的特征提取网络，解码器则由转置卷积层组成。CDRNet为真实与伪造人脸设定差异化的重建目标：对于真实人脸图像，约束其重建结果与$\boldsymbol{I}$一致；而对于伪造人脸图像，则约束其重建结果与$\boldsymbol{I}_{\text{low}}$一致。这种差异化重建策略能够在重建过程中显式放大真伪人脸在高频信息上的差异。图像重建损失的具体公式为：

$$L_{\text{r}} = MSE(\boldsymbol{I},\boldsymbol{I}_{\text{rec}}) = \left\| \boldsymbol{I} - \boldsymbol{I}_{\text{rec}} \right\|_2^2 \tag{1}$$

$$L_{\text{f}} = MSE(\boldsymbol{I}_{\text{low}},\boldsymbol{I}_{\text{rec}}) = \left\| \boldsymbol{I}_{\text{low}} - \boldsymbol{I}_{\text{rec}} \right\|_2^2 \tag{2}$$

$$L_{\text{recon}} = L_{\text{r}} + L_{\text{f}} \tag{3}$$

式中，$MSE(,)$表示均方误差损失，L_{recon}表示总图像重建损失。

2.1.2　残差双域引导模块RDDGM

为了实现双域特征的深度融合与残差信息的充分引导，本文设计了RDDGM，提升了网络模型对细微伪造痕迹的捕捉能力。RDDGM由残差特征提取模块(Residual Feature Extraction Module，RFEM)和双域特征融合引导模块(Dual-Domain Feature Fusion Guidance Module，DDFFGM)构成，整体结构如图5所示。具体而言，首先RFEM将$\boldsymbol{I}_{\text{res}}$输入至一个由3×3卷积层、批量归一化层和RELU激活层构成的卷积网络，再经由1×1卷积层与池化层后得到残差域特征$\boldsymbol{F}_{\text{res}} \in \mathbf{R}^{\frac{C_1}{2}\times H_1\times W_1}$。与此同时，将高频图像$\boldsymbol{I}_{\text{high}}$输入至一个多尺度卷积网络。该网络包含三个并行分支，分别采用不同尺寸的卷积核，捕获不同尺度下的伪造痕迹。随后，将三个分支的输出在通道维度上拼接，并经1×1卷积层与池化层处理，得到多尺度高频域特征$\boldsymbol{F}_{\text{high}} \in \mathbf{R}^{C_1\times H_1\times W_1}$。

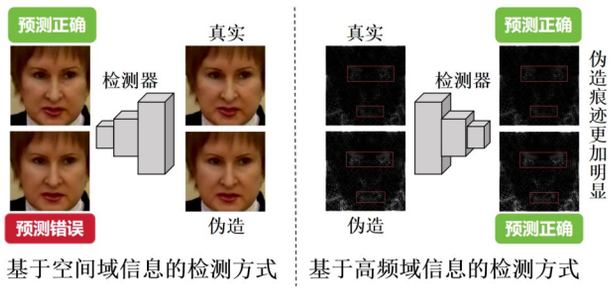


图4　基于空间域与高频域人脸伪造检测效果对比图

Fig. 4　Comparison of face forgery detection effects based on spatial domain and high-frequency domain

为了进一步促进空间域特征与高频域特征的深度融合，DDFFGM引入交叉通道注意力机制与交叉空间注意力机制。首先，将特征$\boldsymbol{F}_{\text{rgb}}$与$\boldsymbol{F}_{\text{high}}$分别输入全局平均池化层，并将池化结果堆叠后输入两个独立的全连接层，生成对应的通道注意力向量。然后，将原始输入特征$\boldsymbol{F}_{\text{rgb}}$与$\boldsymbol{F}_{\text{high}}$分别与对应的通道注意力向量相乘，得到增强后的浅层空间域特征$\boldsymbol{F}_{\text{rgb}}^1 \in \mathbf{R}^{C_1\times H_1\times W_1}$与高频域特征$\boldsymbol{F}_{\text{high}}^1 \in \mathbf{R}^{C_1\times H_1\times W_1}$，具体计算公式为：

$$\boldsymbol{F}_{\text{rgb}}^1 = \boldsymbol{F}_{\text{rgb}} \otimes (M_{\text{rgb}}([G(\boldsymbol{F}_{\text{rgb}}),G(\boldsymbol{F}_{\text{high}})])) \tag{4}$$

$$\boldsymbol{F}_{\text{high}}^1 = \boldsymbol{F}_{\text{high}} \otimes (M_{\text{high}}([G(\boldsymbol{F}_{\text{rgb}}),G(\boldsymbol{F}_{\text{high}})])) \tag{5}$$

式中，$\otimes$表示元素积，$M_{\text{rgb}}(\cdot)$和$M_{\text{high}}(\cdot)$分别表示空间域与高频域的全连接层，[，]表示特征堆叠，$G(\cdot)$表示全局平均池化层。

在获得浅层双域特征$\boldsymbol{F}_{\text{rgb}}^1$与$\boldsymbol{F}_{\text{high}}^1$之后，将二者沿通道维度堆叠，并将堆叠后的特征分别输入到两个卷积层中生成对应的空间注意力图。随后，将浅层特征与对应的空间注意力图进行逐元素相乘，得到

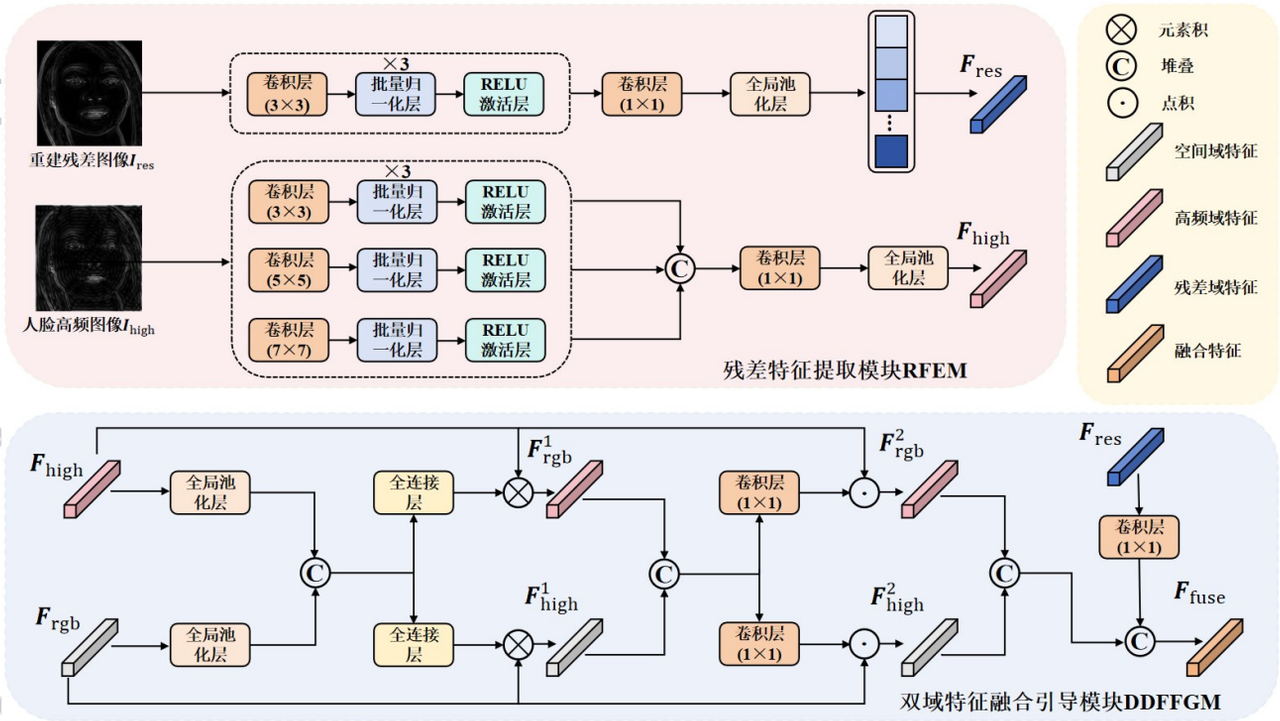


图5　残差双域引导模块RDDGM

Fig. 5　Residual dual-domain guided module

深层空间域特征 $\boldsymbol{F}^2_{\text{rgb}} \in \mathbf{R}^{C_1 \times H_1 \times W_1}$ 与深层高频域特征 $\boldsymbol{F}^2_{\text{high}} \in \mathbf{R}^{C_1 \times H_1 \times W_1}$，具体计算公式为：

$$\boldsymbol{F}^2_{\text{rgb}} = \boldsymbol{F}^1_{\text{rgb}} \odot Conv_{\text{rgb}}([\boldsymbol{F}^1_{\text{rgb}}, \boldsymbol{F}^1_{\text{high}}]) \tag{6}$$

$$\boldsymbol{F}^2_{\text{high}} = \boldsymbol{F}^1_{\text{high}} \odot Conv_{\text{high}}([\boldsymbol{F}^1_{\text{rgb}}, \boldsymbol{F}^1_{\text{high}}]) \tag{7}$$

式中，$\odot$ 表示点积，$Conv_{\text{rgb}}(\cdot)$ 和 $Conv_{\text{high}}(\cdot)$ 分别表示空间域与高频域的卷积层，[，] 表示特征堆叠。

在获得深层双域特征 $\boldsymbol{F}^2_{\text{rgb}}$ 与 $\boldsymbol{F}^2_{\text{high}}$ 之后，将残差域特征 $\boldsymbol{F}_{\text{res}}$ 输入至一个1×1卷积层进行维度调整，并与双域特征 $\boldsymbol{F}^2_{\text{rgb}}$、$\boldsymbol{F}^2_{\text{high}}$ 堆叠得到融合特征 $\boldsymbol{F}_{\text{fuse}} \in \mathbf{R}^{3C_1 \times H_1 \times W_1}$，最后将 $\boldsymbol{F}_{\text{fuse}}$ 输入到分类器得到最终的预测结果，图像分类损失函数采用二值交叉熵损失。$\boldsymbol{F}_{\text{fuse}}$ 的具体计算公式为：

$$\boldsymbol{F}_{\text{fuse}} = [Conv_{\text{res}}(\boldsymbol{F}_{\text{res}}), \boldsymbol{F}^2_{\text{rgb}}, \boldsymbol{F}^2_{\text{high}}] \tag{8}$$

式中，$Conv_{\text{res}}(\cdot)$ 表示残差域卷积层，[，] 表示特征堆叠。

二值交叉熵损失的具体计算公式为：

$$L_{\text{cls}} = -y \times \ln(\tilde{y}) - (1 - y) \times \ln(1 - \tilde{y}) \tag{9}$$

式中，y 和 $\tilde{y}$ 分别表示输入人脸图像的真实标签和预测概率，L_{cls} 表示图像分类损失。

2.1.3　文本感知损失模块TALM

为了利用文本模态信息进一步优化CDRNet的对比差异化重建结果，本文设计了TALM，提升了网络模型对未知伪造方式的泛化能力。本文设计了两类与视觉语义相对应的文本提示词：真实人脸的文本提示词 t_{r} 为“a clear image of a real human face”，伪造人脸的文本提示词 t_{f} 为“a blurry image of a fake human face”。

具体而言，TALM首先将文本提示词输入文本编码器，投影后得到特征 $\boldsymbol{F}^t_{\text{r}}$ 与 $\boldsymbol{F}^t_{\text{f}} \in \mathbf{R}^D$，同时将重建后的图像输入至视觉编码器，得到其投影后的视觉特征 $\boldsymbol{F}^i_{\text{r}}$ 与 $\boldsymbol{F}^i_{\text{f}} \in \mathbf{R}^D$，然后在特征空间中计算视觉与文本特征的余弦相似度。为了更好地实现文本语义对重建结果的跨模态约束，TALM在特征空间中采用余弦相似度来度量特征之间的一致性。选择余弦相似度原因如下：1）余弦相似度关注向量方向而非模长，对幅值变化具有鲁棒性。2）本文的文本与视觉编码器均由CLIP预训练权重初始化，沿用其跨模态对齐的相似度计算方式，可充分利用预训练先验。对于不同伪造方式的人脸图像，TALM均强制其视觉特征向同一个伪造文本特征靠近，同时远离真实文本特征。这种约束使得模型专注于学习伪造人脸通用伪造痕迹，从而显著提升模型对未知伪造方式的泛化能力。文本感知损失的具体公式为：

$$L_{\text{clip}}^{\text{real}} = 1 - \frac{\boldsymbol{F}_{\text{r}}^{i} \cdot \boldsymbol{F}_{\text{r}}^{t}}{\left\| \boldsymbol{F}_{\text{r}}^{i} \right\| \left\| \boldsymbol{F}_{\text{r}}^{t} \right\|} + \frac{\boldsymbol{F}_{\text{r}}^{i} \cdot \boldsymbol{F}_{\text{f}}^{t}}{\left\| \boldsymbol{F}_{\text{r}}^{i} \right\| \left\| \boldsymbol{F}_{\text{f}}^{t} \right\|} \tag{10}$$

$$L_{\text{clip}}^{\text{fake}} = 1 - \frac{\boldsymbol{F}_{\text{f}}^{i} \cdot \boldsymbol{F}_{\text{f}}^{t}}{\left\| \boldsymbol{F}_{\text{f}}^{i} \right\| \left\| \boldsymbol{F}_{\text{f}}^{t} \right\|} + \frac{\boldsymbol{F}_{\text{r}}^{i} \cdot \boldsymbol{F}_{\text{f}}^{t}}{\left\| \boldsymbol{F}_{\text{r}}^{i} \right\| \left\| \boldsymbol{F}_{\text{f}}^{t} \right\|} \tag{11}$$

$$L_{\text{clip}} = \frac{L_{\text{clip}}^{\text{real}} + L_{\text{clip}}^{\text{fake}}}{2} \tag{12}$$

式中，$L_{\text{clip}}^{\text{real}}$和$L_{\text{clip}}^{\text{fake}}$分别表示真实与伪造人脸的文本感知损失函数，$L_{\text{clip}}$表示总文本感知损失函数。

2.2 多任务学习策略

本文采用重建任务与分类任务协同的多任务学习策略，总体网络损失L_{total}由图像重建损失L_{recon}、图像分类损失L_{cls}和文本感知损失L_{clip}三部分构成。L_{recon}和L_{clip}联合优化CDRNet，L_{cls}则优化整个多任务网络。

具体来说，本文通过共享编码器与多任务协同机制实现网络的协同优化。重建任务为真伪人脸设定差异化目标，以扩大其重建差异，从而增强中间特征的真伪判别能力；文本感知任务引入通用文本语义约束，引导模型学习不同伪造方法的通用特征，提升特征泛化能力；分类任务则基于重建残差引导的双域融合特征完成真伪判别，并通过反向传播优化特征表示，增强模型对细微伪造痕迹的感知能力。总体损失函数如下：

$$L_{\text{total}} = \lambda_1 L_{\text{recon}} + \lambda_2 L_{\text{cls}} + \lambda_3 L_{\text{clip}} \tag{13}$$

式中，λ_1、λ_2和λ_3分别表示图像重建损失L_{recon}、图像分类损失L_{cls}和文本感知损失L_{clip}的权重系数。

3 实验结果与分析

3.1 数据集

3.1.1 训练数据集

本文选用FaceForensics++（FF++）数据集的C23压缩版本进行模型训练，并依据官方划分方案构建训练集、验证集和测试集。FF++由Rössler等（2019）提出，是人脸伪造检测领域常用的基准数据集。该数据集包含1000段真实视频样本及多种伪造类型样本，其伪造类型涵盖DeepFake（DF）（Tora，2019）、FaceSwap（FS）（Kowalski，2016）、Face2Face（F2F）（Thies等，2016）和NeuralTexture（NT）（Thies等，2019）四种经典的图像篡改方式。

3.1.2 测试数据集

在测试阶段，本文从同源测试和跨数据集测试两个角度评估模型性能。同源测试基于FF++内部数据展开，重点考察模型在不同伪造方式之间的检测表现；跨数据集测试则采用CDFV1、CDFV2（Li等，2020b）、DFDC（Dolhansky等，2020）、DFDCP（Dolhansky等，2019）、DFD（Nick和Andrew，2019）以及DF40（Yan等，2024a）等公开数据集完成。上述数据集在采集来源、生成方式和数据分布上存在差别，可用于检验模型面对未知数据时的泛化能力。

3.2 实验设置

本文基于Dlib库从训练集视频中提取32帧人脸图像，从测试集视频中提取64帧人脸图像。所有图像的尺寸均统一调整为224×224，并将像素值归一化至[0，1]。

在评估指标方面，参照以往研究工作，本文主要采用ACC和AUC两项指标来评估网络性能。经过实验验证，采用AdamW优化器训练网络模型，初始学习率和批量大小分别设置为1E-4与16。图像编码器的权重使用在ImageNet上预训练的EfficientNet-B4进行初始化。CDRNet中滤波频域掩码半径r设置为16。在训练过程中，本文采用多种数据增强策略，包括随机水平翻转、小角度旋转、随机裁剪、比例缩放以及颜色扰动等。本文方法基于PyTorch框架实现，使用一张NVIDIA GeForce RTX 2080Ti GPU显卡进行模型训练。

3.3 实验结果

3.3.1 域内实验结果

为了评估本文方法在域内场景下的检测性能，本文在FF++数据集的C23与C40压缩版本中分别进行训练与测试，结果如表1所示。表中RECCE（Cao等，2022）、DSRL（Cao等，2024）和FakeDiffer（Wang等，2025a）均为基于人脸图像重建的检测方法。实验结果表明，在两个版本的数据集上，本文方法在各项测试性能指标上均优于现有对比方法。尤其在图像质量较低的C40版本中，本文方法相比FakeDiffer（Wang等，2025a）的ACC和AUC指标分别提高了2.83%和1.75%。

为了评估本文方法在跨压缩场景下的检测性能，本文在FF++数据集的C23版本上进行训练，并在C40版本上进行测试。结果如表2所示，所提方法在跨压缩场景下的测试中表现优异，AUC指标达到85.44%，显著优于现有主流方法，充分验证了模型在跨压缩图像场景下的优异鲁棒性。

表1　不同方法在FF++数据集上的测试结果

Table 1　Test results of different methods on the FF++ dataset

方法	C40		C23	
	ACC	AUC	ACC	AUC
Face X-ray(Li等,2020a)	-	61.60	-	87.35
Two Branch(Masi等,2020)	-	85.59	-	98.70
RECCE(Cao等,2022)	91.03	95.02	97.06	99.32
SFDG(Wang等,2023)	92.28	95.98	<u>98.19</u>	99.53
FoCus(Tian等,2024)	87.31	91.01	96.43	99.15
DSRL(Cao等,2024)	92.31	96.12	97.63	99.44
CUTA(Shi等,2025b)	92.35	96.21	97.65	<u>99.63</u>
FakeDiffer(Wang等,2025a)	<u>93.37</u>	<u>96.54</u>	98.04	99.52
本文方法	**97.20**	**98.29**	**99.60**	**99.71**

注:加粗字体为每列最优值,下划线字体为每列次优值,"-"表示原方法未提供该指标结果。

表2　不同方法在FF++数据集上的跨压缩场景测试结果

Table 2　Test results of different methods across compression scenarios on the FF++ dataset

方法	C23→C40	
	AUC/%	ACC/%
Xception(Rössler等,2019)	82.61	77.23
F3-Net(Qian等,2020)	82.71	77.95
SRM(Luo等,2021)	81.14	76.08
RECCE(Cao等,2022)	81.90	76.72
SFDG(Wang等,2023)	<u>85.06</u>	<u>80.15</u>
FoCus(Tian等,2024)	83.01	78.46
LSDA(Yan等,2024b)	72.01	68.32
FakeDiffer(Wang等,2025a)	83.98	79.21
本文方法	**85.44**	**81.63**

注:加粗字体为每列最优值,下划线字体为每列次优值。

3.3.2　域内跨伪造方式实验结果

为了评估本文方法在跨伪造方式场景下的泛化能力,本文在FF++C23版本数据集上设计如下实验:仅采用其中一种伪造方式的数据进行训练,并在其余三种伪造方式的数据集上测试,结果如表3所示。实验结果表明,与目前最优方法相比,除在DF伪造方式上训练的Cross-Avg值略低于FakeDiffer(Wang等,2025a)外;其余三种伪造方式单作为训练集时,模型的Cross-Avg值均高于以往方法。尤其是以NT伪造方式作为训练集的情况下,Cross-Avg值相较于FakeDiffer(Wang等,2025a)提高了4.75%。

然而,本文方法在DF与FS间的相互泛化性能低于FakeDiffer,这是因为两类伪造模式在空间与频域维度上存在本质差异。DF基于生成模型,其伪造痕迹表现为全局性的频域异常;而FS基于图形学拼接,其伪影主要集中于局部的空间边界。本文的全局频域掩码虽能有效放大生成式伪造的频域痕迹,却也引入了对频域特征的特定偏置。相比之下,FakeDiffer未施加显式的频域约束,其方法本身在两种伪造模式间的域偏置更小。

3.3.3　跨域实验结果

为了评估本文方法在跨域场景下的泛化能力,本文在FF++数据集的C23压缩版本上进行训练,并在5个公开测试数据集上进行跨域测试。在图像层面上,本文方法与以往研究的13种方法进行性能比较,结果如表4所示。实验结果表明,本文方法与当前最优的方法相比,平均AUC指标提高了1.75%。

为了评估本文方法在新伪造技术场景下的跨域泛化能力,在FF++数据集的C23压缩版本上训练后,在包含8种主流伪造模式的DF40数据集上进行了跨域测试。结果如表5所示,该方法取得了82.80%的平均AUC,优于所对比的主流方法,较最新的ProDet(Cheng等,2024)提高了0.39%。具体而言,在E4S、Uniface、Fsgan与Sd2.1四种伪造方式上均取得最优检测性能,其中基于扩散模型的Sd2.1上AUC高达97.53%;在Inswap与Simswap上也取得

表3　不同方法在FF++不同伪造方式上的测试结果

Table 3　Test results of different methods on different forgery modes of FF++

	方法	AUC/%				
		DF	F2F	FS	NT	Avg
DF	RECCE	99.62	70.66	74.29	67.37	70.77
	DSRL	99.76	68.49	72.57	68.01	69.69
	FakeDiffer	99.73	71.27	**75.75**	68.98	**72.00**
	本文方法	**99.92**	**74.13**	69.15	**71.96**	71.74
F2F	RECCE	75.99	98.06	64.53	72.32	70.95
	DSRL	76.63	98.02	65.11	71.23	70.99
	FakeDiffer	76.50	98.96	**67.10**	**73.96**	72.52
	本文方法	**82.94**	**99.02**	66.54	73.21	**74.23**
FS	RECCE	82.39	64.44	98.82	56.70	67.84
	DSRL	**83.21**	64.53	99.01	58.44	68.73
	FakeDiffer	83.02	64.70	99.73	57.85	68.52
	本文方法	80.63	**70.22**	**99.94**	**59.10**	**69.98**
NT	RECCE	78.83	80.89	63.70	93.63	74.47
	DSRL	79.42	80.98	63.13	93.65	74.51
	FakeDiffer	82.06	81.70	66.23	**97.90**	76.66
	本文方法	**89.29**	**86.26**	**68.67**	97.81	**81.41**

注：加粗字体为每列最优值，下划线字体为每列次优值，灰色单元格为训练集与测试集伪造方式一致的情况。

了次优成绩。这表明该方法能有效捕捉新的生成式伪造中的通用伪造痕迹。

在视频层面上，本文方法与8种已有方法进行了性能对比，结果如表6所示。在视频层面的测试过程中，本文采用如下策略：从每个待测视频中均匀采样32帧人脸图像，分别输入模型获得每帧的分类概率，再对全部32个预测概率取平均值，作为该视频的最终预测结果。实验结果表明，与当前最优方法相比，本文所提方法在DFDC和DFDCP两个数据集上均表现最佳，AUC指标分别提高了1.70%与3.03%。

3.4　消融实验

3.4.1　各组件对网络性能的影响

为了评估各组件对网络性能的影响，本文以完整网络为基准，设计了以下三种训练方案：1）移除伪造人脸的图像重建损失。2）移除高频域特征。3）移除TALM。在CDFV2、DFDC和DFDCP数据集上的测试结果如表7所示。可以看出，上述3种方案均会不同程度地导致模型性能下降。实验结果表明，本文方法的各组件均能提升模型对真伪人脸的辨别能力，从而提升网络模型的检测准确率与泛化性能。

3.4.2　频域掩码半径对网络性能的影响

为了评估频域掩码半径对网络性能的影响，本文通过设置不同大小的掩码半径r来进行训练以及性能测试。在本文方法中，掩码半径的大小决定了高频图像中保留的频率成分比例。若r过大，高频信号中噪声占比较高，对模型训练产生干扰。反之，若r过小，高频成分被过度抑制，导致模型难以充分捕捉伪造痕迹。实验结果如表8所示。可以看出，当r=16时，模型在测试集上取得了最优的泛化性能。

3.4.3　总损失函数中权重系数对网络性能的影响

本文通过损失权重系数λ_1、λ_2和λ_3平衡重建损失、分类损失和文本感知损失的重要程度。为了探究不同权重系数设置对网络性能的影响，本文在多种系数组合的情况下进行了模型训练与测试，结果如表9所示。实验结果表明，当重建损失和文本感知损失权重系数较大时，网络模型更侧重于学习真实与伪造人脸的本质差异，而过高的分类损失权重则易使模型过度拟合训练集中的特定伪造模式，导致跨域检测能力下降。

3.4.4　文本提示词的不同设置对网络性能的影响

为了探究不同文本提示词对模型检测性能的影响，我们设计了三组提示词进行消融实验，实验结果如表10所示。三组提示词分别为：无质量描述的基础提示词、引入轻度质量描述的提示词、以及细化质量描述的提示词。实验结果表明，引入轻度质量描述的第二组提示词取得了最优的跨数据集平均AUC，说明适度的质量语义引导有助于模型聚焦于真伪人脸的本质差异。相比之下，细化质量描述的第三组提示词可能是由于过度细化的描述引入了与真伪判别无关的语义噪声，损害了模型的泛化能力。

3.4.5　网络性能评估

为了评估本文方法在网络性能与计算效率上的综合表现，我们在CDFV2数据集上将其与RECCE（Cao等，2022）、UCF（Yan等，2023）和IDCNet（Wang等，2025c）三种代表性方法进行了系统对比，实验结果如表11所示。从检测精度来看，本文方法取得了

表4　不同方法在图像层面上的跨域测试结果

Table 4　Cross-domain test results of different methods at the image level

方法	AUC/%					
	CDFv1	CDFv2	DFDC	DFDCP	DFD	Avg
Xception(Rössler等,2019)	77.90	68.52	69.93	73.71	81.62	74.34
F3-Net(Qian等,2020)	77.73	74.03	70.21	74.08	79.81	75.17
RECCE(Cao等,2022)	76.81	73.22	71.31	74.22	81.23	75.36
UCF(Yan等,2023)	77.95	75.27	71.91	75.94	80.74	76.36
IDCNet(Wang等,2025c)	81.44	80.89	72.44	74.09	84.71	78.71
Zhou等人(2025)	82.50	81.30	76.05	77.80	85.60	80.65
ProDet(Cheng等,2024)	**90.90**	**84.48**	72.40	81.16	85.81	<u>82.95</u>
LSDA(Yan等,2024b)	86.71	83.01	73.60	81.52	88.03	82.57
FA-ViT(Luo等,2024)	78.04	83.29	76.03	81.16	<u>88.50</u>	81.40
Wavlet-CLIP(Baru等,2025)	75.69	75.93	75.04	77.81	85.98	78.09
FreqDebias(Kashiani等,2025)	87.51	83.60	74.10	82.40	86.88	82.90
C2P-CLIP(Tan等,2025)	74.45	81.55	76.71	<u>83.26</u>	<u>88.50</u>	80.89
RepDFD(Lin等,2025)	84.07	82.09	<u>76.95</u>	82.56	84.84	82.10
本文方法	<u>88.22</u>	<u>83.91</u>	**77.11**	**83.29**	**90.95**	**84.70**

注:加粗字体为每列最优值,下划线字体为每列次优值。

表5　不同方法在DF40数据集上的跨域测试结果

Table 5　Cross-domain test results of different methods on the DF40 dataset

方法	AUC/%							
	Uniface	E4S	Facedance	Fsgan	Inswap	Simswap	Sd2.1	Avg
F3-Net(Qian等,2020)	80.21	60.95	76.14	88.43	72.98	65.82	72.41	73.85
SPSL(Liu等,2021)	74.73	59.92	63.01	75.53	61.53	64.05	70.27	67.01
SRM(Luo等人,2021)	74.94	70.42	65,91	77.22	79.33	69.44	87.85	75.02
RECCE(Cao等,2022)	84.22	65.23	78.31	<u>88.45</u>	79.51	73.01	94.73	80.49
IID(Huang等,2023)	79.25	71.03	<u>79.02</u>	86.44	74.47	64.09	-	-
UCF(Yan等,2023)	78.74	69.23	**80.01**	88.13	76.86	64.94	93.91	78.83
LSDA(Yan等,2024b)	85.43	68.45	75.94	83.23	**81.00**	72.75	91.33	79.73
ProDet(Cheng等,2024)	<u>87.86</u>	<u>71.22</u>	74.72	86.52	78.81	**77.83**	<u>97.12</u>	<u>82.41</u>
本文方法	**88.41**	**73.85**	76.03	**88.53**	<u>80.67</u>	<u>74.59</u>	**97.53**	**82.80**

注:加粗字体为每列最优值,下划线字体为每列次优值,“-”表示原方法未提供该指标结果。

83.91%的最优AUC,较性能第二的IDCNet方法提升了3.02%。

在计算效率方面,本文方法的Flops(Floating point operations per second)仅为7.47G,计算复杂度在四种方法中最低。值得关注的是,尽管本文方法的参数量(Params)并非最少,在四种方法中居于第二位,但其以较低的计算量实现了最高的检测精度,表明本文方法的参数利用更为高效。

3.5　模型测试数据特征分布可视化

图6展示了本文方法在四个公开测试数据集上的T-SNE可视化结果。图中绿色圆点代表FF++数据集中的真实人脸特征,其余四种颜色分别对应四

表6　不同方法在视频层面上的跨域测试结果

Table 6　Cross-domain test results of different methods at the video level

方法	AUC/%		
	CDFV2	DFDC	DFDCP
Xception(Rössler等,2019)	68.52	69.93	73.71
F3-Net(Qian等,2020)	74.03	67.76	71.41
SFDG(Wang等,2023)	75.83	-	73.63
SeeABLE(Larue等,2023)	87.30	75.90	<u>86.30</u>
TALL++(Xu等,2024)	**91.93**	<u>78.51</u>	-
SAM(Choi等,2024)	89.00	-	-
LSDA(Yan等,2024b)	<u>91.10</u>	77.01	-
Li等人(2025)	89.70	-	80.20
本文方法	89.82	**80.21**	**89.33**

注:加粗字体为每列最优值,下划线字体为每列次优值,"-"表示原方法未提供该指标结果。

表7　各组件对网络性能的影响

Table 7　The impact of each component on network performance

伪造人脸图像重建损失	高频域特征	TALM	AUC/%			
			CDFV2	DFDC	DFDCP	Avg
-	√	√	79.18	73.15	80.16	77.49
√	-	√	81.32	<u>75.16</u>	<u>83.12</u>	79.87
√	√	-	**85.16**	74.37	82.26	<u>80.59</u>
√	√	√	<u>83.91</u>	**77.11**	**83.29**	**81.43**

注:加粗字体为每列最优值,下划线字体为每列次优值,"-"表示训练过程中移除该组件。

表8　掩码半径对网络性能的影响

Table 8　The influence of mask radius on network performance

r	AUC/%			
	CDFV2	DFDC	DFDCP	Avg
8	83.72	<u>70.21</u>	<u>82.63</u>	<u>78.85</u>
16	<u>83.91</u>	**77.11**	**83.29**	**81.43**
32	**85.15**	69.87	80.81	78.61

注:加粗字体为每列最优值,下划线字体为每列次优值。

表9　总损失函数权重系数对网络性能的影响

Table 9　The influence of the weight coefficient of the total loss function on network performance

$\lambda_1/\lambda_2/\lambda_3$	AUC/%			
	CDFV2	DFDC	DFDCP	Avg
10/1/10	**83.91**	**77.11**	**83.29**	**81.43**
10/1/1	<u>82.63</u>	76.05	82.74	<u>80.47</u>
10/10/1	79.32	73.89	80.25	77.82
1/1/1	80.46	76.58	82.40	79.81
1/10/1	78.59	75.02	81.36	78.32
1/1/10	81.38	<u>76.92</u>	<u>82.91</u>	80.40
1/10/10	80.16	76.23	81.99	79.46

注:加粗字体为每列最优值,下划线字体为每列次优值。

个公开数据集的伪造人脸特征。可以观察到,来自不同伪造方式的伪造人脸特征呈现明显的聚集趋势,且与真实人脸特征之间保持清晰的间距,表明本文方法能够有效捕捉跨伪造方式的通用伪造痕迹。

3.6　CDRNet人脸重建结果可视化

为了进一步验证所提方法能够放大真实与伪造人脸之间重建前后的差异,图7展示了单独重建训练后的CDRNet对真实和伪造人脸输入的重建结果。前两行为真实人脸图像,后两行为伪造人脸图

表10　文本提示词的不同设置对网络性能的影响

Table 10　The Impact of Different Settings of Text prompt Settings on Network Performance

真实提示词	伪造提示词	AUC/%			
		CDFV2	DFDC	DFDCP	Avg
an image of a real human face	an image of a fake human face	83.39	76.44	82.32	80.72
a clear image of a real human face	a blurry image of a fake human face	<u>83.91</u>	**77.11**	**83.29**	**81.43**
a clear and high-quality image of a real human face	a blurry and low-quality image of a fake human face	**83.96**	<u>76.92</u>	<u>83.17</u>	<u>81.35</u>

注：加粗字体为每列最优值，下划线字体为每列次优值。

表11　网络性能评估

Table 11　Network Performance Evaluation

方法	Flops/G↓	Params/M↓	AUC/%↑
RECCE(Cao等，2022)	8.09	**23.78**	73.22
UCF(Yan等，2023)	12.19	<u>44.51</u>	75.27
IDCNet(Wang等，2025c)	<u>7.59</u>	56.78	<u>80.89</u>
本文方法	**7.47**	47.72	**83.91**

注：加粗字体为每列最优值，下划线字体为每列次优值。“↑”表示值越大越好，“↓”表示值越小越好。

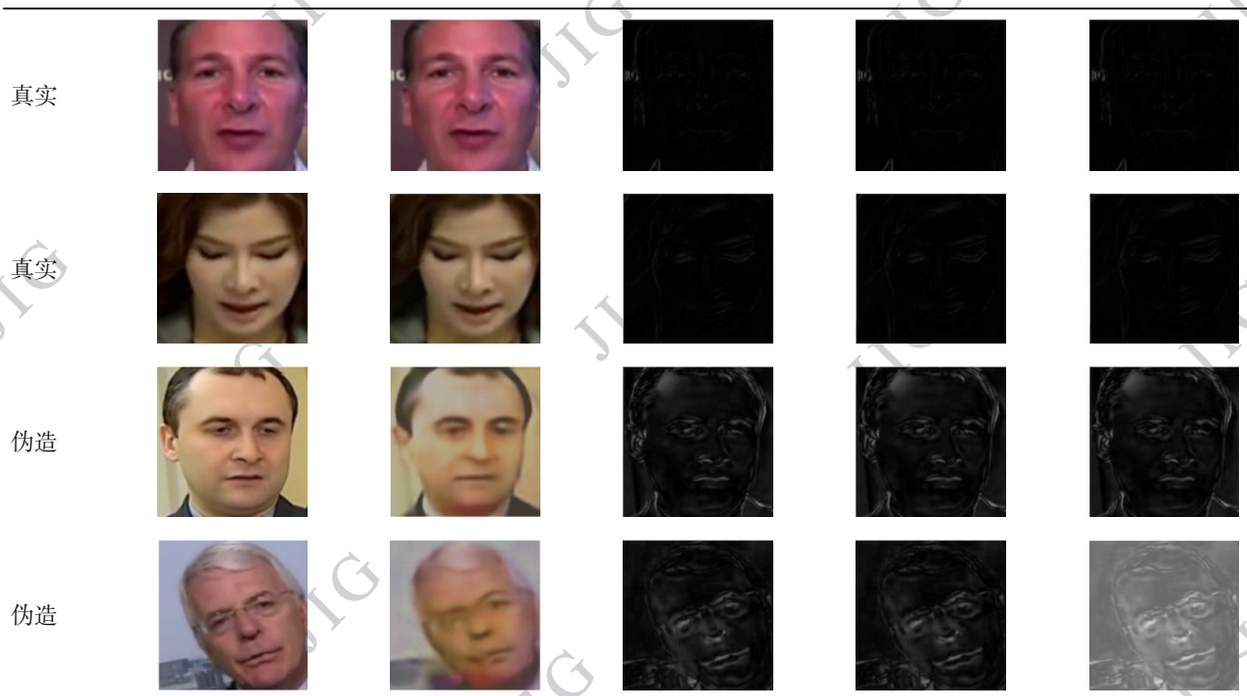
像。可以看出，模型在真实人脸输入下能够较好地恢复结构与细节；而在伪造人脸输入下，重建结果相较真实人脸呈现出边缘模糊、高频纹理丢失的特点。

4　结　论

为了提升人脸伪造检测模型的检测准确率与泛化性能，本文提出差异化重建驱动的残差引导人脸伪造检测方法，提供了新的有效解决方案。首先，对比差异化重建网络CDRNet通过放大真实与伪造人脸在重建过程中的差异，增强了模型对真伪人脸图像的判别能力。其次，残差双域引导模块RDDGM

图7　CDRNet人脸重建结果可视化图

Fig. 7　Visualization of CDRNet face reconstruction results



注：((a) input image；(b) reconstructed image；(c) residual image；(d) gray level image；(e) illumination enhancement)

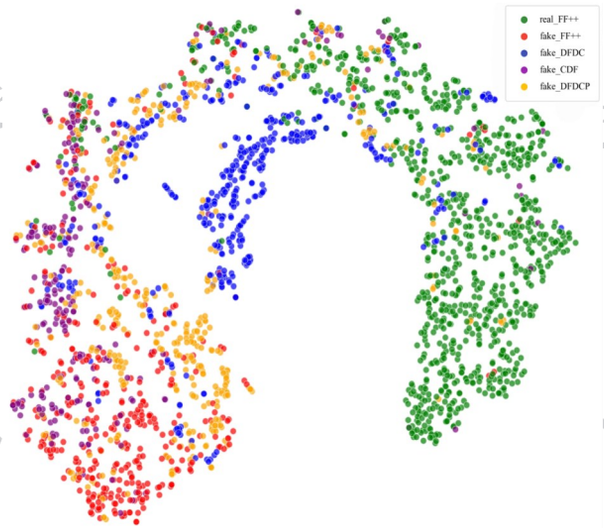


图6　模型在多数据集下人脸特征分类的T-SNE可视化图

Fig. 6　The T-SNE visualization of face feature classification of the model under multiple datasets

利用重建残差图像信息,引导融合后的双域特征,提升了模型对细微伪造痕迹的感知能力。最后,文本感知损失模块TALM引入文本模态信息,有效学习跨伪造方式的通用伪造特征,增强了网络模型对未知伪造人脸的泛化能力。实验结果表明,本文方法在域内测试及跨域泛化测试中均展现出比现有方法更优的检测精度。

参考文献(References)

Baru L B, Boddepalli R, Patel S A and Gajapaka S M. 2025. Wavelet-driven generalizable framework for deepfake face forgery detection//Proceedings of the Winter Conference on Applications of Computer Vision. Tucson, USA: IEEE: 1661-1669 [DOI: 10.1109/WACVW65960.2025.00180]

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4113-4122 [DOI: 10.1109/CVPR52688.2022.00408]

Cao J Y, Zhang K Y, Yao T P, Ding S H, Yang X K and Ma C. 2024. Towards unified defense for face forgery and spoofing attacks via dual space reconstruction learning. International Journal of Computer Vision, 132(12): 5862-5887 [DOI: 10.1007/S11263-024-02151-2]

Cheng J K, Yan Z Y, Zhang Y, Luo Y H, Wang Z Y and Li C. 2024. Can we leave deepfake data behind in training deepfake detector//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc: 21979-21998 [DOI: 10.48550/ARXIV.2408.17052]

Choi J, Kim T, Jeong Y, Baek S and Choi J. 2024. Exploiting style latent flows for generalizing deepfake video detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE: 1133-1143 [DOI: 10.1109/CVPR52733.2024.00114]

Cui J L, Du J W, Li Y Z, Gao L, Jiang H and Bao C F. 2025. HAMLET-FFD: Hierarchical adaptive multi-modal learning embeddings transformation for face forgery detection//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland: ACM: 1327-1336 [DOI: 10.1145/3746027.3755137]

Dolhansky B, Bitton J, Pflaum B, Lu J K, Howes R, Wang M L, et al. 2020. The DeepFake Detection Challenge (DFDC) dataset [EB/OL]. [2025-12-12]. https://arxiv.org/pdf/2006.07397.pdf

Dolhansky B, Howes R, Pflaum B, Baram N and Ferrer C C. 2019. The DeepFake Detection Challenge (DFDC) Preview dataset [EB/OL]. [2025-12-12]. https://arxiv.org/pdf/1910.08854.pdf

Feng C B, Liu C X, Wang Y Y and Zhou Q D. 2024. Face forgery detection with image patch comparison and residual map estimation. Journal of Image and Graphics, 29(2): 457-467 (冯才博, 刘春晓, 王昱烨, 周其当. 2024. 结合图像块比较与残差图估计的人脸伪造检测. 中国图象图形学报, 29(2): 457-467) [DOI: 10.11834/jig.230149]

Fu X H, Yan Z Y, Yao T P, Chen S and Li X. 2025. Exploring unbiased deepfake detection via token-level shuffling and mixing//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 3040-3048 [DOI: 10.1609/AAAI.V39I3.32312]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2014. Generative Adversarial Nets//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates, Inc: 2672 - 2680 [DOI: 10.5555/2969033.2969125]

Huang B J, Wang Z Y, Yang J F, Ai J X, Zou Q, Wang Q, et al. 2023. Implicit identity driven deepfake face swapping detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4490-4499 [DOI: 10.1109/CVPR52729.2023.00436]

Kashiani H, Talemi N A and Afghah F. 2025. FreqDebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8775-8785 [DOI: 10.1109/CVPR52734.2025.00820]

Kim Y, Kwon M J, Lee W and Kim C. 2024. FRIDAY: Mitigating unintentional facial identity in deepfake detectors guided by facial recognizers//Proceedings of the IEEE International Conference on Visual Communications and Image Processing. Tokyo, Japan:

IEEE：1-5［DOI：10.1109/VCIP63160.2024.10849915］

Kowalski M. 2016. FaceSwap［CP/OL］.［2025-12-12］. https://github.com/MarekKowalski/FaceSwap

Larue N，Vu N S，Struc V，Peer P and Christophides V. 2023. SeeABLE：Soft discrepancies and bounded contrastive learning for exposing deepfakes//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris，France：IEEE：21011-21021［DOI：10.1109/ICCV51070.2023.01921］

Lee S，Jung S W and Seo H. 2024. Spectrum translation for refinement of image generation（STIG）based on contrastive learning and spectral filter profile//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver，Canada：AAAI Press：2929-2937［DOI：10.1609/AAAI.V38I4.28074］

Li H Z，Zhou J R，Li Y Z，Wu B Y，Li B and Dong J Y. 2024. FreqBlender：Enhancing deepfake detection by blending frequency knowledge//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver，Canada：Curran Associates，Inc：44965-44988［DOI：10.48550/ARXIV.2404.13872］

Li K，Ren W Q，Li J S，Wang W and Cao X C. 2025. Critical forgetting-based multi-scale disentanglement for deepfake detection//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia，USA：AAAI Press：424-432［DOI：10.1609/AAAI.V39I1.32021］

Li L Z，Bao J M，Zhang T，Yang H，Chen D，Wen F，et al. 2020a. Face X-ray for more general face forgery detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：5001-5010［DOI：10.48550/ ARXIV.1912.13458］

Li Y Z，Yang X，Sun P，Qi H G and Lyu S W. 2020b. Celeb-DF：A large-scale challenging dataset for deepfake forensics//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：3204-3213［DOI：10.1109/CVPR42600.2020.00327］

Lin K M，Lin Y Z，Li W X，Yao T P and Li B. 2025. Standing on the shoulders of giants：Reprogramming visual-language model for general deepfake detection//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia，USA：AAAI Press：5262-5270［DOI：10.1609/AAAI.V39I5.32559］

Liu H G，Li X D，Zhou W B，Chen Y F，He Y，Xue Y，et al. 2021. Spatial-phase shallow learning：Rethinking face forgery detection in frequency domain//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：772-781［DOI：10.48550/ARXIV.2103.01856］

Luo A W，Cai R Z，Kong C Q，Ju Y K，Kang X G，Huang J W，et al. 2024. Forgery-aware adaptive learning with vision transformer for generalized face forgery detection. IEEE Transactions on Circuits and Systems for Video Technology，35(5)：4116-4129［DOI：10.1109/TCSVT.2024.3522091］

Luo Y C，Zhang Y，Yan J C and Liu W. 2021. Generalizing face forgery detection with high-frequency features//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：16317-16326［DOI：10.1109/CVPR46437.2021.01605］

Ma L，Yan Z Y，Xu J，Chen Y Z，Guo Q L，Bi Z，et al. 2025. From specificity to generality：Revisiting generalizable artifacts in detecting face deepfakes［EB/OL］.［2025-12-12］. https://arxiv.org/pdf/2504.04827.pdf

Masi I，Killekar A，Mascarenhas R M，Gurudatt S P and AbdAlmageed W. 2020. Two-Branch recurrent network for isolating deepfakes in videos//Proceedings of 16th European Conference on Computer Vision. Glasgow，UK：Springer International Publishing：667-684［DOI：10.48550/ARXIV.2008.03412］

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 International Publishing：86-103［DOI：10.48550/ARXIV.2007.09355］

Rössler A，Cozzolino D，Verdoliva L，Riess C，Thies J and Niessner M. 2019. FaceForensics++：Learning to detect manipulated facial images//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul，Korea：IEEE：1-11［DOI：10.1109/ICCV.2019.00009］

Shi L，Zhang J，Ji Z L，Bai J F and Shan S G. 2025a. Real face foundation representation learning for generalized deepfake detection. Pattern Recognition，161：#111299［DOI：10.1016/J.PATCOG.2024.111299］

Shi Z N，Chen H P，Jia Y X，Zhang D，Lu W and Yang X. 2025b. Customized transformer adapter with frequency masking for deepfake detection. IEEE Transactions on Information Forensics and Security，20：5904-5918［DOI：10.1109/TIFS.2025.3574983］

Shiohara K and Yamasaki T. 2022. Detecting deepfakes with self-blended images//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，USA：IEEE：18720-18729［DOI：10.1109/CVPR52688.2022.01816］

Tan C C，Tao R S，Liu H，Gu G H，Wu B Y，Zhao Y，et al. 2025. C2P-CLIP：Injecting category common prompt in clip to enhance generalization in deepfake detection//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia，USA：AAAI Press：7184-7192［DOI：10.1609/AAAI.V39I7.32772］

Thies J，Zollhöfer M，Stamminger M，Theobalt C and Nießner M. 2016. Face2Face：Real-time face capture and reenactment of RGB videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，USA：IEEE：2387-2395［DOI：10.48550/ARXIV.2007.14808］

Thies J，Zollhöfer M and Nießner M. 2019. Deferred neural rendering：Image synthesis using neural textures. ACM Transactions on Graphics，38(4)：1-12［DOI：10.48550/ARXIV.1904.12356］

Tian J H, Chen P, Yu C, Fu X M, Wang X, Dai J, et al. 2024. Learning to discover forgery cues for face forgery detection. IEEE Transactions on Information Forensics and Security, 19: 3814-3828 [DOI: 10.1109/TIFS.2024.3372773]

Tora M. 2019. DeepFakes[CP/OL]. [2025-12-12]. https://github.com/deepfakes/faceswap

Wang B, Zhang Z, Zhao S Y, Ye X M, Zhang H J and Wang M. 2025a. FakeDiffer: Distributional disparity learning on differentiated reconstruction for face forgery detection//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 7518-7526 [DOI: 10.1609/AAAI.V39i7.32809]

Wang S Y, Feng C B, Liu C X and Jin Y S. 2025b. Multivariate and soft blending sample-driven image-text alignment for deepfake detection. Journal of Image and Graphics, 30(5): 1334-1345 (王诗雨, 冯才博, 刘春晓, 金逸胜. 2025b. 多元软混合样本驱动的图文对齐人脸伪造检测. 中国图象图形学报, 30(5): 1334-1345) [DOI: 10.11834/JIG.240252]

Wang Y, Yu K, Chen C, Hu X Y and Peng S L. 2023. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 7278-7287 [DOI: 10.1109/CVPR52729.2023.00703]

Wang Z Y, Chen Y X, Yao Y Z, Han M, Xing W P and Li M. 2025c. IDCNet: Image decomposition and cross-view distillation for generalizable deepfake detection. IEEE Transactions on Information Forensics and Security, 20: 8373-8386 [DOI: 10.1109/TIFS.2025.3593353]

Xu Y T, Liang J, Sheng L J and Zhang X Y. 2024. Learning spatiotemporal inconsistency via thumbnail layout for face deepfake detection. International Journal of Computer Vision, 132(12): 5663-5680 [DOI: 10.1007/S11263-024-02054-2]

Yan Z Y, Yao T P, Chen S, Zhao Y D, Fu X H, Zhu J W, et al. 2024a. DF40: Toward next-generation deepfake detection[EB/OL].[2025-12-12]. https://arxiv.org/pdf/2406.13495.pdf

Yan Z Y, Luo Y H, Lyu S W, Liu Q S and Wu B Y. 2024b. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE: 8984-8994 [DOI: 10.1109/CVPR52733.2024.00858]

Yan Z Y, Zhang Y, Fan Y B and Wu B Y. 2023. UCF: Uncovering common features for generalizable deepfake detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22412-22423 [DOI: 10.1109/ICCV51070.2023.02048]

Zhang Y N, Li Q F, Yu Z T and Shen L L. 2025. Distilled transformers with locally enhanced global representations for face forgery detection. Pattern Recognition, 161, #111253 [DOI: 10.1016/J.PATCOG.2024.111253]

Zhao T C, Xu X, Xu M Z, Ding H, Xiong Y J and Xia W. 2021. Learning self-consistency for deepfake detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 15023-15033 [DOI: 10.1109/ICCV48922.2021.01475]

Zhou W J, Luo X Q, Zhang Z C, He J C and Wu X J. 2025. Capture artifacts via progressive disentangling and purifying blended identities for deepfake detection[EB/OL]. [2025-12-12]. https:// arxiv.org/pdf/2410.10244.pdf

作者简介

岳书同,男,本科生,主要研究方向为人脸深度伪造检测、深度学习与计算机视觉。E-mail: 2312190106@pop.zjgsu.edu.cn

刘春晓,通信作者,男,副教授,主要研究方向为视觉计算与计算机图形学、机器学习与智能系统、视觉安全与隐私保护。E-mail: cxliu@mail.zjgsu.edu.cn