doi: 10.13413/j.cnki.jdxblxb.2024073

基于逆向知识蒸馏人脸重建的深度伪造检测算法

刘文瑀[1,2], 陈海鹏[1], 孙宝胜[3]
(1. 吉林大学 计算机科学与技术学院, 长春 130012;
2. 辽宁省烟草公司抚顺市公司, 辽宁 抚顺 113008;
3. 吉林省肿瘤医院 放疗科, 长春 130012)

摘要: 针对深度伪造检测算法在数据集 FaceForensics++(FF++)上神经纹理(neural textures, NT)伪造方法检测效果较低的问题, 通过对人脸图像的细粒度特征提取进行改进, 提出一个逆向知识蒸馏网络(reverse knowledge distillation net, RKD-Net). 首先, RKD-Net 以逆向知识蒸馏为主体框架, 保留了输入人脸图像丰富的细粒度信息; 其次, 在编码器和解码器中间插入了空间和通道重建卷积, 从空间和通道两个维度上加强细粒度信息的表示; 最后, 使用残差坐标注意力分类器, 增强逆向知识蒸馏网络输出的真实特征和细节特征, 并根据这些不同特征对输入到网络的人脸图像进行分类. 实验结果表明, RKD-Net 在保证对其他伪造方法检测效果的情况下, 对 NT 伪造方法检测效果达到最佳.

关键词: 深度伪造检测; 逆向知识蒸馏; 空间和通道重建卷积; 坐标注意力
中图分类号: TP391.41 **文献标志码:** A **文章编号:** 1671-5489(2025)06-1637-09

Deepfake Detection Algorithm Based on Reverse Knowledge Distillation for Face Reconstruction

LIU Wenyu[1,2], CHEN Haipeng[1], SUN Baosheng[3]
(1. *College of Computer Science and Technology, Jilin University, Changchun* 130012, *China*;
2. *Liaoning Provincial Tobacco Company Fushun City Company, Fushun* 113008, *Liaoning Province, China*;
3. *Department of Radiotherapy, Jilin Cancer Hospital, Changchun* 130012, *China*)

Abstract: Aiming at the problem that deepfake detection algorithms had low detection effect on neural texture (NT) forgery methods in FaceForensics++ (FF++) dataset, we proposed a reverse knowledge distillation network (RKD-Net) by improving fine-grained feature extraction of face images. Firstly, the RKD-Net used reverse knowledge distillation (RKD) as the main framework, preserving the rich fine-grained information of the input face images. Secondly, spatial and channel reconstruction convolution (SCConv) was inserted between the encoder and the decoder to enhance the representation of fine-grained information from both spatial and channel dimensions. Finally, a residual coordinate attention (RCA) classifier was used to enhance the real and detailed features output by the reverse knowledge distillation network, and classify the face images input to the

收稿日期: 2024-02-26.

第一作者简介: 刘文瑀(1998—), 男, 汉族, 硕士研究生, 从事图像处理与模式识别的研究, E-mail: liuwy21@mails.jlu.edu.cn.
通信作者简介: 孙宝胜(1972—), 男, 汉族, 博士, 教授, 从事肿瘤放射医学的研究, E-mail: 1575164354@qq.com.

基金项目: 吉林省科技发展计划重点研发项目(批准号: YDZJ202502CXJD068).

network according to these different features. The experimental results show that RKD-Net achieves the best detection effect of NT forgery methods while guaranteeing the detection effect of other forgery methods.

Keywords: deepfake detection; reverse knowledge distillation; spatial and channel reconstruction convolution; coordinate attention

深度伪造的伪造对象主要是人脸，包括人脸的视频[1-2]和图像[3-4]，少部分是音频[5-6]和文字[7-8]，本文的研究对象是人脸. 随着科技飞速发展，伪造的名人图像越来越多，因此，对这些伪造图像进行检测势在必行. 目前，基于深度学习的深度伪造检测方法可分为三类：重建模型、对比学习和迁移学习. 基于重建模型的深度伪造检测算法，其核心是利用编解码网络学习正常人脸的潜在结构和特征，然后进行重建. OC-FakeDect 模型[9]早期引入重建概念，利用变分自编码器进行正常人脸重建分类. 基于对比学习的深度伪造检测算法旨在训练编码器，使其对同类数据编码相似，对不同类数据编码各异. FDFL 模型[10]使用单中心损失和交叉熵损失联合监督，将真实特征作为正样本，伪造特征作为负样本，二者对抗训练，拉近正样本，拉远负样本. 基于迁移学习的深度伪造检测算法利用预训练模型的知识迁移进行伪造检测. Multi 模型[11]将细粒度鸟类分类任务中的二分类进行迁移，提出了一种新的多注意力网络架构，以从多个人脸注意力区域捕获局部判别特征. Dual Descriptor 模型[12]采用结合监督和无监督方法，通过引入两种描述符在空间域中提取丰富信息，从而实现了对人脸图像真假的分类；此外，其还引入了一个频域重建模块，扩展了面部特征的表示空间，从而提升了分类的准确性. 其中，FDFL 模型中的频率感知方法重点是利用频率信息(如低频和高频特征)增强模型对伪造人脸的检测能力，这种设计使模型更关注整体的频率分布差异，而不是局部的细节特征. Multi 模型中的多种注意力之间可能会互相干扰，影响最终对细节特征的提取和识别. Dual Descriptor 模型中的频域重建模块在一定程度上是对空间域细节特征的抽象和压缩，无法完全捕捉到图像中的局部细节特征.

因此，上述方法对人脸图像的细节特征关注度仍较弱. 针对上述问题，本文提出一种逆向知识蒸馏的人脸重建算法(reverse knowledge distillation face reconstruction net，RKD-Net)改进网络结构. 该算法使用逆向知识蒸馏网络，保留了教师网络丰富且紧凑的细粒度信息，使学生网络能更好地恢复正常图像的潜在特征表示. 在此基础上，该算法又在教师编码器和学生解码器之间引入了空间和通道重建卷积，有效减少了卷积神经网络在空间维度和通道维度两方面的冗余特征，增强了细粒度信息的表示. 最后，通过残差坐标注意力模块交互不同空间信息，加强对深度纹理等潜在细节特征的关注，提升 NT(NeuralTexture)伪造检测性能.

1 方法设计

如图 1 所示，逆向知识蒸馏网络包含人脸重建、空间和通道重建卷积以及利用残差坐标注意力的图像分类 3 个模块. 人脸重建模块利用逆向知识蒸馏框架保留教师编码器的细节特征. 空间和通道模块减少异常人脸信息冗余，增强面部纹理特征. 坐标注意力使用两个一维卷积在方向和位置维度上增强特征. 最终结合坐标注意力增强特征和图像差异特征，送入分类器进行精准分类.

1.1 逆向知识蒸馏模块

逆向知识蒸馏的教师模型在数据集 ImageNet[13]上进行预训练. 为避免学生教师模型 H_k 的最后特征值收敛到一般解，在知识提取时冻结了教师编码器的所有参数. 学生解码器的结构完全与教师编码器对称，用匹配教师编码器的中间特征表示. RKD-Net 网络的教师编码器采用 ResNet[14]网络作为骨干网络，学生解码器由残差结构的解码模块镜像对称. ResNet 网络中解码部分的卷积核大小为 1，步长为 2. 学生解码部分的卷积核大小为 2，步长为 2. 逆向知识蒸馏包括了二维异常图 $M^k(h,w)$和标量损失函数 L_{KD}，分别表示为

$$M^k(h,w)=1-\frac{(\boldsymbol{f}_E^k(h,w))^{\mathrm{T}}\cdot\boldsymbol{f}_D^k(h,w)}{\|\boldsymbol{f}_E^k(h,w)\|\ \|\boldsymbol{f}_D^k(h,w)\|},\qquad L_{\mathrm{KD}}=\sum_{k=1}^{K}\left\{\frac{1}{H_kW_k}\sum_{h=1}^{H_k}\sum_{w=1}^{W_k}M^k(h,w)\right\},\tag{1}$$

其中：E^k 表示第 k 个编码模块；D^k 表示第 k 个解码模块；$\boldsymbol{f}_E^k$，$\boldsymbol{f}_D^k$ 分别是第 k 个编码器和解码器对应的特征向量，$\boldsymbol{f}_E^k$，$\boldsymbol{f}_D^k \in \mathbb{R}^{C_k \times H_k \times W_k}$；$C_k$，$H_k$，$W_k$ 分别表示第 k 层被激活向量的通道数、高度和宽度；M^k 中大特征值表示高异常度；K 表示实际使用特征层数.

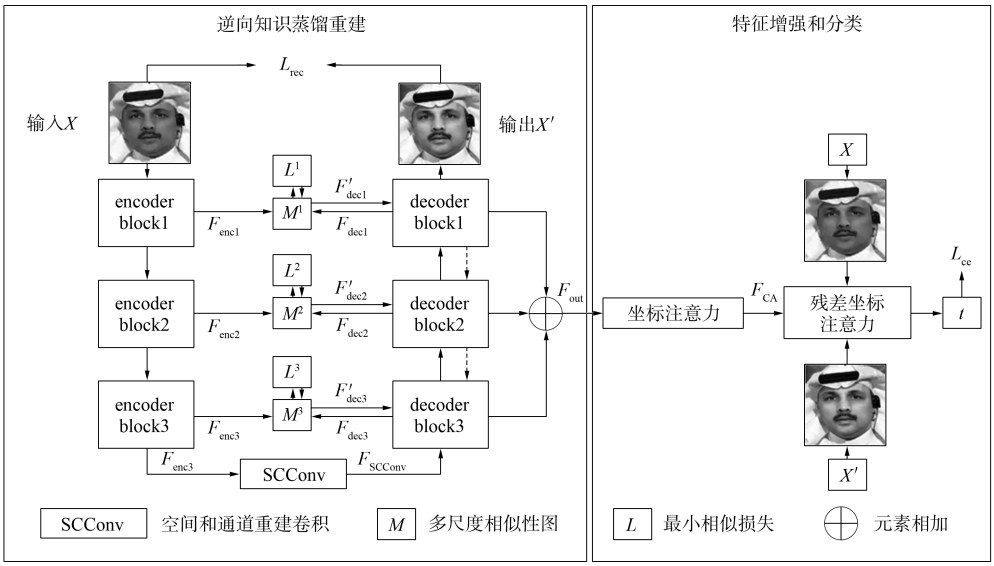

图 1　逆向知识蒸馏网络整体结构

Fig. 1　Overall structure of reverse knowledge distillation network

1.2　空间和通道重建卷积模块

1.2.1　空间重建卷积

图 2 为空间重建单元结构. 由图 2 可见，空间重建单元包含分离和重建两次操作. 输入到卷积网络的特征图具有不同信息量，分离操作通过组归一化实现，其中比例因子可判断特征图信息量，进而将特征图分离. 操作方法如下：当输入一个特征图 $\boldsymbol{F} \in \mathbb{R}^{N \times C \times H \times W}$（$N$ 为批次(batch)轴，C 为通道(channel)轴，H 为相应的空间高度(height)轴，W 为相应的宽度(width)轴)时，首先组归一化(group normalization，GN)输入特征 $\boldsymbol{F}$，即减去平均值 μ 并与标准差 σ 相除：

$$\mathrm{GN}(\boldsymbol{F}) = \gamma \frac{\boldsymbol{F} - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta, \qquad X_{\mathrm{out}} = \mathrm{GN}(\boldsymbol{F}), \tag{2}$$

其中 ε 是为稳定除法的常数，γ 和 β 为可训练仿射变换因子. 仿射变换就是通过平移、旋转、缩放和错切等图像变换方法，保持特征图的共线性和点之间的距离比例不变.

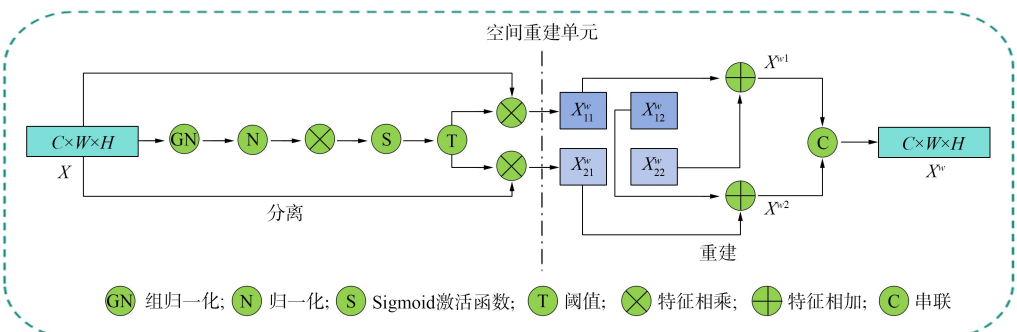

图 2　空间重建单元结构

Fig. 2　Structure of spatial reconstruction unit

对于特征图空间信息的差异，分离操作采用组归一化层中的可训练参数 $\gamma \in \mathbb{R}^C$ 测量每个批次和通道的空间像素方差得到，方差越大说明特征图空间的信息越丰富. 信息的数量由组归一化权重 $W_\gamma \in \mathbb{R}^C$ 控制，权重越大信息越多，反之，权重越小信息越少：

$$\{w_i\}=\frac{\gamma_i}{\sum_{j=1}^{C}\gamma_j},\quad i,y=1,2,\cdots,C,\quad W_\gamma=\{w_i\}. \tag{3}$$

其次，将不同权重特征图的特征向量相乘，并使用 Sigmoid 激活函数[14]，将特征向量相乘后的特征图和权重值投影到由阈值控制的门控单元中，阈值范围为 0～1. 该门控单元的阈值在实验中设置为 0.5，阈值以上的权重全部设置为 1，为信息权重，用 W_1 表示. 将阈值以下的权重全部设置为 0，为非信息权重，用 W_2 表示. 整个权重表示为

$$W=\mathrm{Gate}(\mathrm{Sigmoid}(W_\gamma(\mathrm{GN}(\boldsymbol{F})))). \tag{4}$$

最后，将输入特征 $\boldsymbol{F}$ 分别乘以 W_1 和 W_2 进行加权，得到两个加权后的特征 $\boldsymbol{F}_1^w$ 和 $\boldsymbol{F}_2^w$，其中 $\boldsymbol{F}_1^w$ 是信息丰富的重要特征，$\boldsymbol{F}_2^w$ 是信息少量的冗余特征. 从而将输入特征 $\boldsymbol{F}$ 分离为两部分：

$$\begin{cases}\boldsymbol{F}_1^w=W_1\otimes\boldsymbol{F},\\ \boldsymbol{F}_2^w=W_2\otimes\boldsymbol{F},\end{cases} \tag{5}$$

其中⊗为元素相乘操作. 由于 $\boldsymbol{F}_2^w$ 是冗余特征，故需尽可能减少. 为减少 $\boldsymbol{F}_2^w$，需进行重建操作. 重建操作是将 $\boldsymbol{F}_1^w$ 和 $\boldsymbol{F}_2^w$ 通过交叉重构的方式进行相加，以丰富信息. 交叉重建结合不同权重特征，使重建特征更全面、交互性更强. 最终将不同权重特征相加后的重建特征取交集，得到空间信息更细致的特征图 $\boldsymbol{F}^w$：

$$\begin{cases}\boldsymbol{F}_{11}^w\oplus\boldsymbol{F}_{22}^w=\boldsymbol{F}^{w1},\\ \boldsymbol{F}_{21}^w\oplus\boldsymbol{F}_{12}^w=\boldsymbol{F}^{w2},\\ \boldsymbol{F}^{w1}\cup\boldsymbol{F}^{w2}=\boldsymbol{F}^w,\end{cases} \tag{6}$$

其中⊕为元素相加操作，∪为取并集操作. 尽管空间重建卷积通过分离和重建操作解决了空间维度上的冗余信息问题，但通道维度上仍存在冗余信息特征，因此需采用通道重建卷积处理.

1.2.2 通道重建卷积

图 3 为通道重建单元结构. 由图 3 可见，通道重建单元有 3 次操作，分别是分离、转换和融合. 通道重建卷积仍使用 $K\times K$ 大小的标准卷积核提取信道维度的特征.

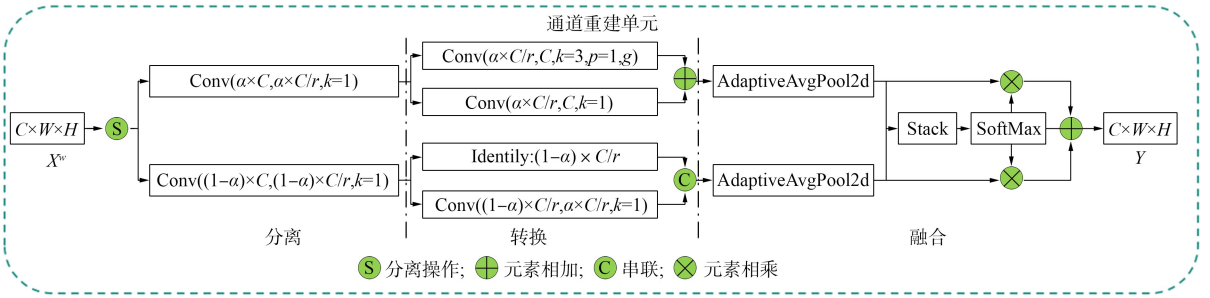

图 3 通道重建单元结构

Fig. 3 Structure of channel reconstruction unit

分离：给定一个空间细粒度特征，使其分离为两个通道的特征. 一个通道是 αC，另一个通道是 $(1-\alpha)C$，其中 α 是分割因子，范围为 0～1. 之后，分离操作再添加一个压缩比例 r，常被设置为 2，用于控制通道注意力的计算量. 经过分离和压缩操作后，空间重定义特征被分为 $\boldsymbol{F}_{\mathrm{up}}$ 和 $\boldsymbol{F}_{\mathrm{low}}$ 上下两部分.

转换：$\boldsymbol{F}_{\mathrm{up}}$ 作为丰富的特征提取器，被送到上部进行转化. 采用高效卷积操作组卷积(GWC)和逐点卷积(PWC)代替 $K\times K$ 的标准卷积核，以提取高级语义信息并节省运算开销. GWC 能有效减少参数量，但阻断了信道组之间的信息交互. PWC 恰好弥补了这些信息的损失，使这些信息能积极进行交互. 转换操作对 $\boldsymbol{F}_{\mathrm{up}}$ 分别进行 $K\times K$ 大小的 GWC 操作和 1 大小的 PWC 操作，共进行 g 次(一般 $g=2$)，得到具有代表性的特征图：

$$\boldsymbol{Y}_1=\boldsymbol{M}^G\boldsymbol{F}_{\mathrm{up}}+\boldsymbol{M}^{P_1}\boldsymbol{F}_{\mathrm{up}}, \tag{7}$$

其中 $\boldsymbol{M}^G\in\mathbb{R}^{\infty c/(gr)\times k\times k\times c}$，$\boldsymbol{M}^{P_1}\in\mathbb{R}^{\infty c/r\times 1\times 1\times c}$ 是控制 $\boldsymbol{F}_{\mathrm{up}}$ 不断学习的权重矩阵，$\boldsymbol{F}_{\mathrm{up}}\in\mathbb{R}^{\infty c/r\times h\times w}$，

$\boldsymbol{Y}_1 \in \mathbb{R}^{c\times h\times w}$. $\boldsymbol{F}_{\text{low}}$被送到下部进行转化. 在下部分，只用 PWC 卷积操作生成能提取潜在信息的特征图. 然后与 $\boldsymbol{F}_{\text{low}}$再度相连，形成更多样的特征图 $\boldsymbol{Y}_2$，且不需要额外的开销：

$$\boldsymbol{Y}_2 = \boldsymbol{M}^{P_2}\boldsymbol{F}_{\text{low}} \cup \boldsymbol{F}_{\text{low}}, \tag{8}$$

其中 $\boldsymbol{M}^{P_2} \in \mathbb{R}^{(1-\infty)c/r\times 1\times 1\times[(1-\infty)c/r]c}$为不断学习的权重矩阵，$\boldsymbol{F}_{\text{low}} \in \mathbb{R}^{(1-\infty)c/r\times h\times w}$，$\boldsymbol{Y}_2 \in \mathbb{R}^{c\times h\times w}$.

融合：融合操作使用 SKNet[15] 自动整合 $\boldsymbol{F}_{\text{up}}$和 $\boldsymbol{F}_{\text{low}}$上下两部分的输出特征 $\boldsymbol{Y}_1$ 和 $\boldsymbol{Y}_2$，而不是直接将这两部分的特征相加. 首先，将转换后的两个特征图分别进行全局平均池化操作，获取具有通道特征的全局空间信息 $\boldsymbol{S}_m \in \mathbb{R}^{c\times 1\times 1}$：

$$\text{Pooling}(\boldsymbol{Y}_m) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{Y}_c(i,j), \qquad m=1,2, \tag{9}$$

$$\boldsymbol{S}_m = \text{Pooling}(\boldsymbol{Y}_m), \qquad m=1,2. \tag{10}$$

其次，将 $\boldsymbol{S}_1$ 和 $\boldsymbol{S}_2$ 叠加，使用 SoftMax 函数[16]计算得到特征向量 $\boldsymbol{\beta}_1$ 和 $\boldsymbol{\beta}_2$：

$$\boldsymbol{\beta}_1 = \frac{\mathrm{e}^{\boldsymbol{S}_1}}{\mathrm{e}^{\boldsymbol{S}_1}+\mathrm{e}^{\boldsymbol{S}_2}}, \quad \boldsymbol{\beta}_2 = \frac{\mathrm{e}^{\boldsymbol{S}_2}}{\mathrm{e}^{\boldsymbol{S}_1}+\mathrm{e}^{\boldsymbol{S}_2}}, \quad \boldsymbol{\beta}_1+\boldsymbol{\beta}_2 = 1. \tag{11}$$

最后，通过分别将高低两部分的输出特征 $\boldsymbol{\beta}_1$ 和 $\boldsymbol{\beta}_2$ 相乘，再合并，得到输出后的细致特征：

$$\boldsymbol{Y} = \boldsymbol{\beta}_1\boldsymbol{Y}_1 + \boldsymbol{\beta}_2\boldsymbol{Y}_2. \tag{12}$$

1.3　残差坐标注意力模块

RKD-Net 使用坐标注意力对编解码器输出的特征进行增强，同时，将 $\boldsymbol{F}_{\text{Coordinate}}$、输入图像特征、重建图像特征三者进行残差操作后，再次送入到坐标注意力中进行分类. 坐标注意力分成两部分：一部分是坐标信息嵌入；另一部分是坐标注意力生成.

坐标信息嵌入：为保留空间维度上的精确位置信息长程依赖，坐标注意力将全局池化分解为方程式形式. 对输入特征 $\boldsymbol{X}$，先采用尺寸为$(H,1)$和$(1,W)$的池化核进行池化操作，再沿 x 轴和 y 轴方向对每通道特征进行编码. 在高度 h 和宽度 w 的第 c 个通道输出可表示为

$$z_c^h(h) = \frac{1}{W}\sum_{0\leqslant j<W} x_c(h,i), \qquad z_c^w(h) = \frac{1}{H}\sum_{0\leqslant j<H} x_c(j,w). \tag{13}$$

坐标注意力生成：坐标信息嵌入做到了拥有全局感受野和提取精准的位置信息. 为实现位置信息的有效表示，需生成坐标注意力. 生成的坐标注意力需充分利用坐标信息嵌入捕捉到的位置信息，增强对细节信息的敏感程度，并有效捕捉各内部通道特征之间的关系. 先将高度 h 和宽度 w 的第 c 个通道输出连接在一起，输入到卷积核为 1×1 大小的卷积变换函数 F_1 中：

$$\boldsymbol{f} = \delta(F_1([z^h, z^w])), \tag{14}$$

其中 δ 为非线性激活函数，$[\cdot,\cdot]$为空间维度上的级联运算，$\boldsymbol{f}\in\mathbb{R}^{c/r\times(h+w)}$为 x 轴和 y 轴两个坐标方向对空间信息编码的中间位置特征图. 然后沿空间维度将 $\boldsymbol{f}$ 分离成高和宽两个独立的向量 $\boldsymbol{f}^h \in \mathbb{R}^{c/r\times H}$ 和 $\boldsymbol{f}^w \in \mathbb{R}^{c/r\times W}$. 剩下两个 1×1 卷积变换 F_h 和 F_w 将 $\boldsymbol{f}^h$ 和 $\boldsymbol{f}^w$ 转换成与输入特征 $\boldsymbol{X}$ 具有相同通道数的向量 $\boldsymbol{g}^h$ 和 $\boldsymbol{g}^w$：

$$\begin{aligned} \boldsymbol{g}^h &= \text{Sigmoid}(F_h(\boldsymbol{f}^h)), \\ \boldsymbol{g}^w &= \text{Sigmoid}(F_w(\boldsymbol{f}^w)). \end{aligned} \tag{15}$$

1.4　损失函数

深度伪造检测使用损失函数评估分类器获得的结果. 总损失函数包括两个分量：正常人脸重建的重建损失和揭示人脸图像是正常还是异常的二分类损失. 重建损失衡量重建图像与正常图像之间的差异为

$$L_{\text{rec}} = \frac{1}{|N|}\sum_{i\in N}\| X_i' - X_i \|_1, \tag{16}$$

其中 N 表示正常人脸图像样本的数量，$|N|$表示 N 的基数，i 表示重建人脸图像 X'和正常图像 X 的序号. 二分类损失测量预测结果和真实标签之间的差异为

$$L_{\text{ce}} = -\frac{1}{M}\sum_{i=1}^{M}[t_i\cdot\log(P(t_i)) + (1-t_i)\cdot\log(1-P(t_i))], \tag{17}$$

其中 M 表示输入图像的总数，t_i 表示所有图像中当前正常图像的序号，$P(t_i)$表示正常图像的概率，$P(1-t_i)$表示图像异常的概率. 总损失是分类损失和重建损失的组合，表示为

$$L=L_{ce}+\lambda L_{rec}, \tag{18}$$

其中 λ 为损失的平衡因子，本文 $\lambda=0.1$.

2 实 验

本文实验使用 FF++[17]，Celeb-DF(CDF)[18-19]和 WildDeepfake(WDF)[20] 3 个数据集，深度学习框架为 PyTorch，模型训练在 NVIDIA GeForce RTX 3090 GPU 上进行. 优化器为 Adam，学习率为 0.000 2，权重衰减为 0.000 01. 使用固定步长衰减策略调整学习率，学习率调整步长为 22 500，学习率调节倍数为 0.5. 训练集的批量大小为 32. 评价指标采用精确度(ACC)、ROC(receiver operating characteristic)曲线下的面积(AUC)和等误差率(EER).

2.1 域内实验

在数据集 FF++上进行域内实验，并将测试结果列于表 1. 实验采用轻度压缩版本(FF++c23)和重度压缩版本(FF++c40)的数据集进行比较. 通过与现有深度伪造模型进行对比，可评估不同压缩级别对检测精度的影响. 实验主要关注两个评价指标：ACC 和 AUC. 这些指标能全面而准确地展示本文检测方法在不同压缩版本上的性能.

表 1 域内测试实验结果

Table 1 Experimental test results of intra-domain %

方法	FF++c23		FF++c40	
	ACC	AUC	ACC	AUC
Steg. Features	70.97	—	55.98	—
Xception	95.49	97.80	84.67	87.21
Constrained Conv	82.97	—	58.69	—
LD-CNN	78.45	—	58.69	—
MesoNet	83.10	—	70.47	—
Multi-task	87.63	88.72	75.37	76.74
EfficientNetB4	96.25	98.94	86.13	88.14
Face X-ray	—	87.40	—	61.60
Two-Branch	96.43	98.70	86.34	86.59
Add-Net	96.78	97.74	87.50	91.01
GFFD	96.87	98.95	86.89	88.27
RFM	95.69	98.79	87.06	89.83
FST-Matching+ResNet-18	94.52	98.34	88.92	92.02
FST-Matching+Efficient-b4	96.19	98.81	88.69	91.27
ALFE(Xception)	96.51	99.10	87.80	92.60
RKD-Net	97.10	99.22	90.14	95.09

进一步对本文方法在数据集 FF++上的 4 种不同伪造方法进行测试. 为更全面地评估本文模型的有效性，选择数据集 FF++c23 和 FF++c40. 表 2 列出了在数据集 FF++c23 和 FF++c40 上 4 种伪造方法的测试结果. 由表 2 可见：OC-FakeDect 方法对低质量的人脸图像更适配，能更深入地解析 NT 伪造特征；Dual Descriptor 方法结合重建学习与分类学习，提供了真假样本之间的高频像素差异，可有效检测 DF 伪造方法；而本文方法旨在提升模型的整体重建效果，牺牲了模型对人脸图像 NT 特征的提取和识别能力，但本文方法在两种不同压缩程度数据集的 4 种伪造方法上的平均检测结果也表现出卓越的性能，并在针对 NT 伪造方法的检测中也有显著优势.

表 2　在数据集 FF＋＋c23 和 FF＋＋c40 上 4 种伪造方法的测试结果

Table 2　Test results of four forgery methods on FF＋＋c23 and FF＋＋c40 datasets　%

方法	FF＋＋c23					FF＋＋c40				
	DF	F2F	FS	NT	平均	DF	F2F	FS	NT	平均
Steg. Features	77.12	74.68	79.51	76.94	77.06	65.58	57.55	60.58	60.69	61.10
NCL	90.18	94.93	93.14	86.04	91.07	80.95	77.30	76.83	72.38	76.87
Local Descriptors	81.78	85.32	85.69	80.60	83.35	68.26	59.38	62.08	62.42	63.04
E-CNN	82.16	93.48	92.51	75.18	85.83	73.25	62.33	67.08	62.59	66.31
Meso-4	89.77	84.25	95.50	78.70	87.06	77.68	83.65	79.92	77.74	79.75
MesoIn-4	93.74	91.48	94.34	75.06	88.66	74.20	78.75	79.72	67.94	75.15
Simple Features	—	—	—	—	—	71.69	65.66	65.43	59.34	65.53
Xception	95.15	97.07	95.96	87.99	94.04	83.70	87.21	83.17	87.90	85.50
OC-FakeDect	—	—	—	—	—	88.40	71.20	86.10	97.50	85.80
SST	98.50	91.90	98.30	96.40	96.30	—	—	—	—	—
ALFE＋Xception	—	—	—	—	—	84.69	81.84	83.65	76.97	81.79
FTDN	98.05	96.32	—	—	—	83.55	83.19	—	—	—
Dual Descriptor	—	—	—	—	—	97.98	88.70	93.40	84.20	89.30
RKD-Net	99.18	98.93	98.76	96.71	98.40	94.33	90.26	94.17	87.92	91.30

2.2　域外实验

如表 3 所示，本文在数据集 FF＋＋c23 上进行训练，在数据集 CDF 上进行测试，验证模型在域外的检测性能，从而可以更全面地评估模型的泛化能力. 将测试结果与现有的人脸伪造检测方法进行比较，结果表明，本文方法在领域外的检测效果最优，证明了本文模型具有良好的泛化性能.

表 3　域外实验测试结果

Table 3　Experimental test results of external domain　%

方法	ACC	方法	ACC
Xception-raw	48.20	F^3Net	65.17
Xception-c23	65.30	Multi	67.44
Xception-c40	65.50	DualNetwork	72.30
Two-stream	53.80	M2TR	65.70
Meso-4	54.80	LTW	64.10
MesoIn-4	53.60	DMGT	72.30
EfcicntNet-B4	64.29	Patch-DFD(Inception-V3)	72.92
Capsule	57.50	MC-LCR	71.61
DSP-FWA	64.60	FS	58.10
SMIL	56.30	RKD-Net	72.98

2.3　对比实验

本文使用的基线模型为 U-net[21]，且对每个模块都进行了消融实验. 实验在数据集 FF＋＋c23 上训练，在数据集 CDF 上测试，对比实验结果列于表 4.

表 4　对比实验结果

Table 4　Comparison of experimental results　%

模型	FF＋＋c23		CDF
	ACC	AUC	AUC
基线	93.95	97.26	64.03
基线＋逆向知识蒸馏	95.36	97.81	72.38
基线＋逆向知识蒸馏＋空间和通道重建卷积	96.98	98.83	74.89
逆向知识蒸馏网络	97.10	99.22	75.66

2.4　可视化结果

图 4 为 4 种伪造方法在数据集 FF＋＋c23 上的特征图可视化实验结果. 由图 4 可见，本文方法对

真实图像的重建结果良好，对伪造图像重建效果不佳，证明了本文方法的有效性.

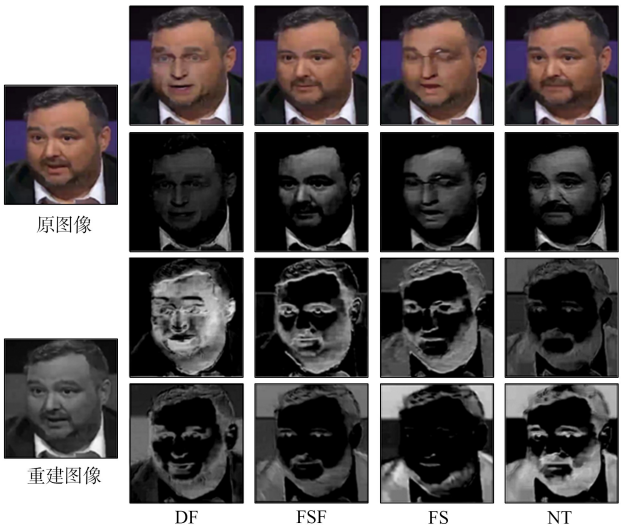

图4 4种伪造方法在数据集FF++c23上的特征图可视化实验结果

Fig. 4 Visualization experiment results of feature maps by four forgery methods on FF++c23 dataset

综上所述，针对深度伪造检测算法在数据集FaceForensics++(FF++)上神经纹理伪造方法检测效果较低的问题，本文提出了一种基于逆向知识蒸馏的人脸重建方法检测深度伪造图像. 该方法首先通过逆向知识蒸馏提升学生模型对异常人脸的分类能力，有效学习真实特征. 其次，在网络中间层引入空间和通道重建卷积，抑制空间和信道冗余. 最后，结合残差坐标注意力分类器放大真实特征，通过特征对比分析计算二分类损失函数进行分类. 实验结果表明，该方法对数据集FF++的NT伪造检测中性能最佳，有效解决了现有方法对深度纹理特征检测效果不佳的问题.

参 考 文 献

[1] COZZOLINO D, RÖSSLER A, THIES J, et al. Id-Reveal: Identity-Aware Deepfake Video Detection [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 15108-15117.

[2] SUN Z K, HAN Y J, HUA Z Y, et al. Improving the Efficiency and Robustness of Deepfakes Detection through Precise Geometric Features [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 3609-3618.

[3] ZHAO T C, XU X, XU M Z, et al. Learning Self-consistency for Deepfake Detection [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 15023-15033.

[4] ZHAO H Q, ZHOU W B, CHEN D D, et al. Self-supervised Transformer for Deepfake Detection [EB/OL]. (2022-03-02)[2024-01-10]. https://arxiv.org/abs/2203.01265.

[5] ZHOU Y P, LIM S N. Joint Audio-Visual Deepfake Detection [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 14800-14809.

[6] LIU R, ZHANG J H, GAO G L. Multi-space Channel Representation Learning for Mono-to-Binaural Conversion Based Audio Deepfake Detection [J]. Information Fusion, 2024, 105: 102257-1-102257-14.

[7] ZHAO L, CHEN C S, HUANG J W. Deep Learning-Based Forgery Attack on Document Images [J]. IEEE Transactions on Image Processing, 2021, 30: 7964-7979.

[8] SHAO R, WU T X, WU J L. Detecting and Grounding Multi-modal Media Manipulation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 6904-6913.

[9] KHALID H, WOO S S. Oc-Fakedect: Classifying Deepfakes Using One-Class Variational Autoencoder [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2020: 656-657.

[10] LI J M, XIE H T, LI J H, et al. Frequency-Aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 6458-6467.

[11] ZHAO H Q, ZHOU W B, CHEN D D, et al. Multi-attentional Deepfake Detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 2185-2194.

[12] SUO Y C, ZHAO X H, GUO Y F, et al. A Dual Domain Attention Mechanism for Face Forgery Detection [C]// 2023 IEEE International Joint Conference on Biometrics (IJCB). Piscataway, NJ: IEEE, 2023: 1-10.

[13] DENG J, DONG W, SOCHER R, et al. Imagenet: A Large-Scale Hierarchical Image Database [C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248-255.

[14] HE K M, ZHANG X Y, REN S Q, et al. Deep Residual Learning for Image Recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778.

[15] ELFWING S, UCHIBE E, DOYA K. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning [J]. Neural Networks, 2018, 107: 3-11.

[16] LI X, WANG W H, HU X L, et al. Selective Kernel Networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 510-519.

[17] ROSSLER A, COZZOLINO D, VERDOLIVA L, et al. Faceforensics＋＋: Learning to Detect Manipulated Facial Images [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 1-11.

[18] JANG E, GU S X, POOLE B. Categorical Reparameterization with Gumbel-Softmax [EB/OL]. (2016-11-03) [2024-02-01]. https://arxiv.org/abs/1611.01144.

[19] LI Y Z, YANG X, SUN P, et al. Celeb-DF: A Large-Scale Challenging Dataset for Deepfake Forensics [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 3207-3216.

[20] ZI B J, CHANG M H, CHEN J J, et al. Wilddeepfake: A Challenging Real-World Dataset for Deepfake Detection [C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 2382-2390.

[21] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional Networks for Biomedical Image Segmentation [C]//18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer International Publishing, 2015: 234-241.

（责任编辑：韩　啸）