伪影间共性机理驱动的多域感知社交网络深度伪造视频检测

王 艳[①] 孙钦东*[①②] 荣东柱[①] 汪小雄[①③]
[①](西安理工大学计算机科学与工程学院 西安 710048)
[②](西安交通大学网络空间安全学院 西安 710049)
[③](四川数字经济产业发展研究院 成都 610036)

摘 要：深度伪造技术在社交网络上的滥用引发了人们对视觉内容真实性与可靠性的严重担忧。已有检测算法未充分考虑社交网络上深度伪造视频的退化现象，导致深度伪造检测性能受以压缩为主的伪影信息干扰与上下文相关信息缺失等挑战性问题的限制。压缩编码与深度伪造生成算法上采样操作会在视频上留下伪影，这些伪影可导致真实视频与深度伪造视频间的细粒度差异。该文通过分析压缩伪影与深度伪造伪影的共性机理，揭示了二者间的结构相似性，为深度伪造检测模型抗压缩鲁棒性的增强提供了可靠理论依据。首先，针对压缩噪声对深度伪造特征的干扰，基于压缩伪影与深度伪造伪影频域表示的结构相似性，设计了频域自适应陷波滤波器以消除特定频带上压缩伪影的干扰。其次，为削弱深度伪造检测模型对未知噪声的敏感，设计了基于残差学习的去噪分支。采用基于注意力机制的特征融合方法增强深度伪造判别特征，结合度量学习策略优化网络模型，实现了具有抗压缩鲁棒性的深度伪造检测。理论分析与实验结果表明，与基线方法相比，该文算法在压缩深度伪造视频上的检测性能具有明显提升，并可作为一种即插即用模型与现有检测方法结合以提高其抗压缩鲁棒性。

关键词：深度伪造检测；社交网络；压缩深度伪造；伪影间共性机理
中图分类号：TN911.73; TP309.2 **文献标识码**：A **文章编号**：1009-5896(2024)09-3713-09
DOI: 10.11999/JEIT240025

Deepfake Video Detection on Social Networks Using Multi-domain Aware Driven by Common Mechanism Analysis Between Artifacts

WANG Yan[①] SUN Qindong[①②] RONG Dongzhu[①] WANG Xiaoxiong[①③]
[①](*School of Computer Science and Engineering*, *Xi'an University of Technology*, *Xi'an* 710048, *China*)
[②](*School of Cyber Science and Engineering*, *Xi'an Jiaotong University*, *Xi'an* 710049, *China*)
[③](*Sichuan Digital Economy Industry Development Research Institute*, *Chengdu* 610036, *China*)

Abstract: The misuse of deepfake technology on social networks has raised serious concerns about the authenticity and reliability of visual content. The degradation phenomenon of deepfake videos on social networks has not been adequately considered in existing detection algorithms, resulting in deepfake detection performance being limited by challenging issues such as compression artifacts interference and lack of context-related information. Compression encoding and up-sampling operations in deepfake generation algorithms can leave artifacts on videos, which can result in fine-grained differences between real videos and deepfake videos. The common mechanisms between compression artifacts and deepfake artifacts are analyzed to reveal the structural similarities between them, which provides a reliable theoretical basis for enhancing the robustness of deepfake detection models against compression. Firstly, to address the interference of compression noise on deepfake features, the frequency-domain adaptive notch filter is designed based on the structural similarity of compression artifacts and deepfake artifacts to eliminate the interference of compression artifacts on specific frequency bands. Secondly, the denoising branch based on residual learning is designed to reduce the sensitivity of the deepfake detection model to unknown noise. Additionally, the attention-based feature fusion method is

收稿日期：2024-01-18；改回日期：2024-06-13；网络出版：2024-06-18
*通信作者： 孙钦东 qdongsun@xjtu.edu.cn
基金项目：国家自然科学基金(62272378)，陕西省重点研究开发项目(2022ZDLSF07-07)，四川省自然科学基金(2023NSFSC0502)
Foundation Items: The National Natural Science Foundation of China (62272378), Shaanxi Province Key Research and Development Program (2022ZDLSF07-07), The Natural Science Foundation of Sichuan Province (2023NSFSC0502)

adopted to enhance the discriminative features of deepfakes. Metric learning strategies are adopted to optimize network models, achieving deepfake detection with resistance to compression. Theoretical analysis and experimental results indicate that the detection performance of compressed deepfake videos is significantly enhanced by using the algorithm proposed in this paper. It can be used as a plug-and-play model combined with existing detection methods to enhance their robustness against compression.

Key words: Deepfake detection; Social network; Compressed deepfake; Common mechanism between artifacts

1 引言

生成式人工智能技术[1]的快速发展，使攻击者能够轻易地制作高逼真的深度伪造视频，突破人们“眼见为实”的认知[2]，已成为极具危害的社会问题之一。特别是，暴力、色情、虚假新闻等海量深度伪造视频在社交网络上的传播，更易引发严重的安全、隐私和社会问题[3]。因此，有效地检测社交网络上的深度伪造视频是目前研究的热点与难点。

社交网络上的深度伪造视频由于带宽和存储条件的限制，不可避免地受压缩、噪声等退化因素的影响。现有的大多数深度伪造检测方法基于大规模深度伪造数据集的公开可访问性，采用机器学习模型设计并挖掘启发式特征，在未压缩的深度伪造检测任务中取得了优异进展[4]。然而，在社交网络低质量压缩的深度伪造检测上表现不佳。如图1所示，采用目前深度伪造检测领域广泛应用的Xception[5]网络模型对原始深度伪造图像的判别特征进行提取，其可视化图与掩码图像基本吻合。然而，受压缩、噪声等退化因素的影响，提取的社交网络上深度伪造图像的判别特征可视化图与掩码图像相比具有很大偏差。主要原因是压缩操作导致大量用于判别视频是否为假的伪影信息损失，以及社交网络有损操作引入的各类未知噪声伪影的干扰。

为解决上述问题，部分研究通过知识蒸馏、超分辨率等技术来提高深度伪造检测的抗压缩鲁棒性。文献[6]采用融合知识蒸馏的频域学习和最优传输理论改进低质量压缩深度伪造检测性能。文献[7]设计了多尺度分支缩放网络，采用无监督超分辨率

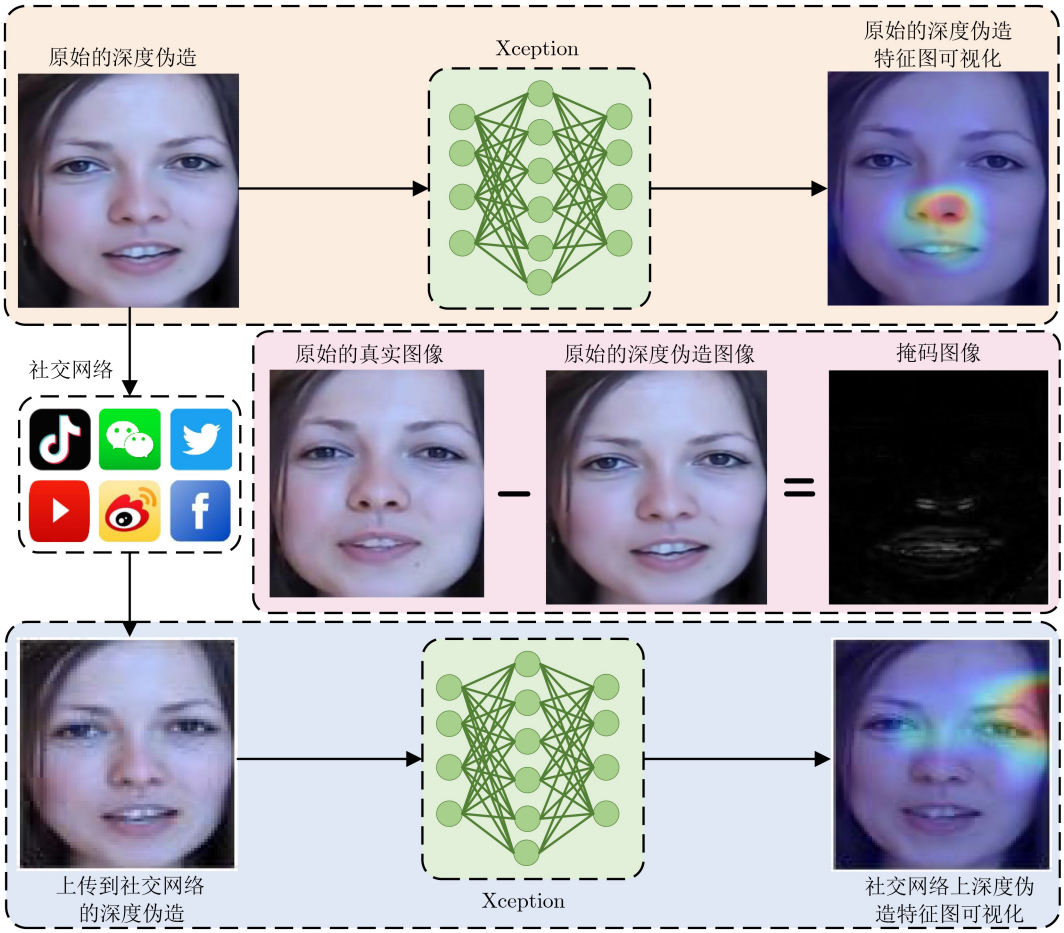


图 1 Xception检测模型受社交网络退化因素影响的可视化分析

技术，结合多尺度图像训练网络以增强低质量深度伪造特征。然而，这些方法仅采用数据驱动的优势，结合损失函数约束模型以优化参数，构建相应的检测模型，缺乏压缩操作与深度伪造生成算法的共性特征机理分析，导致模型设计具有一定的盲目性，难以对深度伪造特征进行深入的分析和挖掘。

压缩算法对视频进行采样、量化和编码等操作的过程中会引入一定程度的失真，这些失真在压缩视频中表现为可见伪影[8]。在深度伪造生成过程中，生成式人工智能模型采用上采样操作，将低分辨率的输入图像映射到高分辨率的伪造图像，该过程也会引入可见细粒度伪影，这些伪影可用来识别深度伪造[9]。随着人工智能生成技术的不断改善，空域中的深度伪造伪影已逐渐消失。然而，深度伪造视频在压缩算法和上采样操作的作用下，在频域产生的细粒度可见伪影仍可作为判别是否为伪造的依据。

因此，本文通过研究频域表示中压缩伪影与深度伪造伪影间的共性机理，从理论上证明了压缩伪影与深度伪造伪影的结构相似性，设计了一种伪影间共性机理驱动的多域感知社交网络深度伪造视频检测算法。本文的主要贡献如下：

(1)通过对压缩伪影和深度伪造中上采样伪影的数学建模，从理论上证明了压缩伪影的干扰性，为压缩深度伪造检测领域提供了一个新的视角。

(2)基于压缩伪影与深度伪造伪影间的共性机理，设计了一种社交网络压缩深度伪造检测模型，使该模型具有一定的可解释性。

(3)在公开数据集上进行了广泛的实验，验证了本文方法在检测压缩的深度伪造视频方面具有良好的性能。并可作为一种即插即用模型与现有检测方法结合以提高其抗压缩鲁棒性。

2　相关工作

目前针对深度伪造视频的检测主要有基于生理信号的检测、基于空域的检测、基于频域的检测等方法。基于生理信号的方法主要检测伪造视频时域中生理特征的不连续性。Li等人[10]依据生成的深度伪造视频中眨眼生理信号的异常来检测假视频。Ciftci等人[11]从脸部抽取3块区域测量光电容积脉搏波信号以区分深度伪造。基于空域的深度伪造检测主要利用卷积神经网络提取视频像素域上的特征进行检测。Cozzolino等人[12]设计了基于自编码器的神经网络结构以识别深度伪造。Afchar等人[13]提出了基于卷积神经网络的MesoInception-4网络模型。随着深度伪造检测技术的发展，大量高逼真的深度伪造图像或视频仅具有细微的伪影。利用生理特征和空域特征进行深度伪造检测变得十分困难，而这些细粒度的伪造伪影在频域中较为明显。Qian等人[14]提出了基于频率感知线索的深度伪造检测模型，设计了频率感知分解与局部频率统计分支，采用交叉注意力机制融合特征来检测伪造视频。Liu等人[15]提出了一种采用空间相位浅层学习方法进行深度伪造检测。

在现实场景中，社交网络上的深度伪造视频受压缩、噪声等退化因素的影响，会引入具有干扰性的伪影信息以及上下文相关信息的缺失现象，导致大部分已有的深度伪造检测算法性能下降。针对此类情况，部分研究者开展了相关研究：Binh等人[6]设计了一个教师-学生网络，训练不同质量的图像，引导学生网络从低质量样本中学习判别特征。Li等人[16]提出了基于超分辨率的思想，通过恢复空间特征提高社交网络低质量深度伪造性能。上述方法的基本思路为设计相应的模型，将低分辨率的深度伪造图像重建为高分辨率，以增强压缩图像的深度伪造特征。然而，这些方法主要关注提高视频的质量和细节，对于局部区域伪造，可能无法有效恢复其细节。此外，它们缺乏压缩操作与深度伪造生成算法的共性机理分析，导致模型设计具有一定的盲目性，难以对深度伪造特征进行深入的分析和挖掘。

3　伪影间共性机理分析

为从理论上证明压缩操作对深度伪造特征的影响，本节从频谱结构与其位置信息分布分析压缩伪影与深度伪造伪影间的共性机理。首先，针对压缩伪影频谱分布，将帧图像进行频域转换并以块为单位进行表示，采用特定频带的期望对频谱结构特征进行分析，得到压缩图像特定频带上的分布。基于压缩算法通用编码参数，得到特定频带位置信息。之后，分析深度伪造生成图像的频域特征，对深度伪造生成模型中的单层上采样操作进行数学建模，进而采用谱复制推导得出，上采样操作会产生格状伪影且该伪影在谱复制边界产生。基于以上对图像频域压缩伪影与深度伪造伪影的结构与其位置分布信息的分析与推导，证明了压缩伪影与深度伪造伪影的结构相似性与重叠现象，从理论上说明了压缩伪影的干扰性，为深度伪造检测模型提供理论指导与设计依据。

首先，对压缩伪影的频域特征进行分析。具体地，视频被分成帧并进行离散余弦变换(Discrete Cosine Transform, DCT)得到$F(u,v)$，依据同一块内的像素值具有邻域相似性且服从同一分布[17]以及DCT变换的可分离性，对帧图像的DCT变换以块单位表示为

$$p(u,v)=\cos\left[\frac{\pi(2y+1)v}{2N}\right]\cdot\sum_{x=iL}^{(i+1)L-1}f(x,y)\cdot\cos\left[\frac{\pi(2x+1)u}{2N}\right]\tag{1}$$

其中，$f(x,y)$表示(x,y)处的像素值，$N\times N$表示图像大小，$u,v\in\{0,1,\cdots,N-1\}$，i表示块内行，j表示块内列，L表示块的大小。由于$F(u,v)$是由若干个$p(u,v)$累加而成。若所有$p(u,v)$的期望为0，则$F(u,v)$的期望为0。因此，要证明压缩图像在频域存在格状伪影，即证明$E(p(u,v))=0$。由于压缩算法使得在一个块内的像素$f(x,y)$服从同一分布，则可依据期望性质对式(1)进行化简。

首先假设在$\mathfrak{L}=\{x\mid x\in[0,L),x\in N^+,L>0\}$上存在两两对应的$x$和$m$使式(2)成立

$$-\cos\left[\frac{\pi(2x+1)u}{2N}\right]=\cos\left[\frac{\pi(2m+1)u}{2N}\right]\tag{2}$$

将式(2)带入式(1)中化简，得到$p(u,v)$

$$p(u,v)=\cos\left[\frac{\pi(2y+1)v}{2N}\right]\cdot\sum_{x,m\in\mathcal{L}}\cos\left[\frac{\pi(2x+1)u}{2N}\right][f(x,y)-f(m,y)]\tag{3}$$

根据期望性质，由式(2)和式(3)可得

$$E(f(x,y)-f(m,y))=E(f(x,y))-E(f(m,y))\tag{4}$$

其中，$f(x,y)$和$f(m,y)$为同分布，则$E(f(x,y)-f(m,y))=0$，即证明了在满足假设的情况下$E(F(u,v))=0$。针对假设成立的条件，视频压缩编码中的块单元大小L通常为16，则取$L=16$。由于本文所设计的检测模型输入图像分辨率为256×256，则$u\in[0,255]$，可得出一组解$u\in\{16,32,\cdots,240\}$使假设成立。

进一步，本节分析了深度伪造算法生成图像的频域特征。由于上采样是深度伪造生成模型必不可少的操作[15]，常见的上采样层为反卷积。反卷积将输入特征采用补零方式扩大图像尺寸并应用标准卷积运算。因此，上采样操作可形式化为

$$\boldsymbol{Q}_{\rm up}=I(\boldsymbol{Q})\otimes\boldsymbol{K}_{\rm up}\tag{5}$$

其中$I(\cdot)$表示将特征$\boldsymbol{Q}$进行补0操作，$\boldsymbol{K}_{\rm up}$表示卷积核。对于反卷积上采样，$\boldsymbol{K}_{\rm up}$为可学习的。依据卷积定理，空间卷积运算相当于频域的乘法运算，因此频域上采样操作可表示为

$$\mathcal{F}(\boldsymbol{Q}_{\rm up})={\rm Repeat}(\mathcal{F}(\boldsymbol{Q}))\odot\mathcal{F}(\boldsymbol{K}_{\rm up})\tag{6}$$

将上采样操作进行简化，考虑单层并将分辨率提高2倍的情况，该方法也可进行扩展。使用零交叉上采样算法将输入图像空间分辨率提高2倍，表示为$\boldsymbol{I}'$。因此，可以得到

$$F'(u+N,v+N)=\frac{2}{N}\sum_{x=0}^{2N-1}\sum_{y=0}^{2N-1}I'(2x,2y)\cdot\cos\left[\frac{2\pi(2x+1)\cdot(u+N)}{2N}\right]\cdot\cos\left[\frac{2\pi(2y+1)\cdot(v+N)}{2N}\right]\tag{7}$$

当x或y为奇数时，$I'(x,y)=0$，则可得到

$$\begin{aligned}F'(u+N,v+N)&=\frac{2}{N}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}I'(2x,2y)\cdot\cos\left[\frac{2\pi(2x+1)u}{2N}\right]\cdot\cos\left[\frac{2\pi(2y+1)v}{2N}\right]\\&=F(u,v)\end{aligned}\tag{8}$$

由上述公式谱复制推导可得，上采样会导致频谱在复制过程中产生格状伪影。此外，反卷积生成图像时，会产生2倍的平移。则根据最终图像256×256的分辨率特性，可递归地推出它是由4个128×128的块平移复制而来。而128×128是由64×64的块平移复制而来，则可得到产生格状伪影的边界位置，即为16, 32, 64等，这和压缩导致的格状伪影产生了重叠。

综合上述分析可以证明，压缩伪影与深度伪造伪影在频域具有结构相似性且存在重叠现象，因此压缩伪影会干扰检测模型对深度伪造特征的提取。

4 伪影间共性机理驱动的多域感知社交网络深度伪造视频检测

针对社交网络中压缩、后处理噪声等退化因素对深度伪造检测性能的影响，本节设计了伪影间共性机理驱动的多域感知社交网络深度伪造视频检测模型。如图2所示，该模型主要分为3个部分：频域自适应陷波滤波、空域残差去噪和基于注意力机制的特征融合模块。针对已知的压缩损失，基于压缩伪影与深度伪造伪影的结构相似性理论分析，针对性地设计了频域自适应陷波滤波模块，以增强检测器的抗压缩能力；针对未知噪声损失，设计了空域残差去噪模块，削弱深度伪造检测器对未知噪声的敏感。为挖掘频域分支与空域分支输出特征图间的内在关联关系，采用基于注意力机制的特征融合方法，增强深度伪造固有特征表征能力。此外，采用度量学习策略优化网络模型，构建具有抗压缩鲁棒性的深度伪造检测模型。

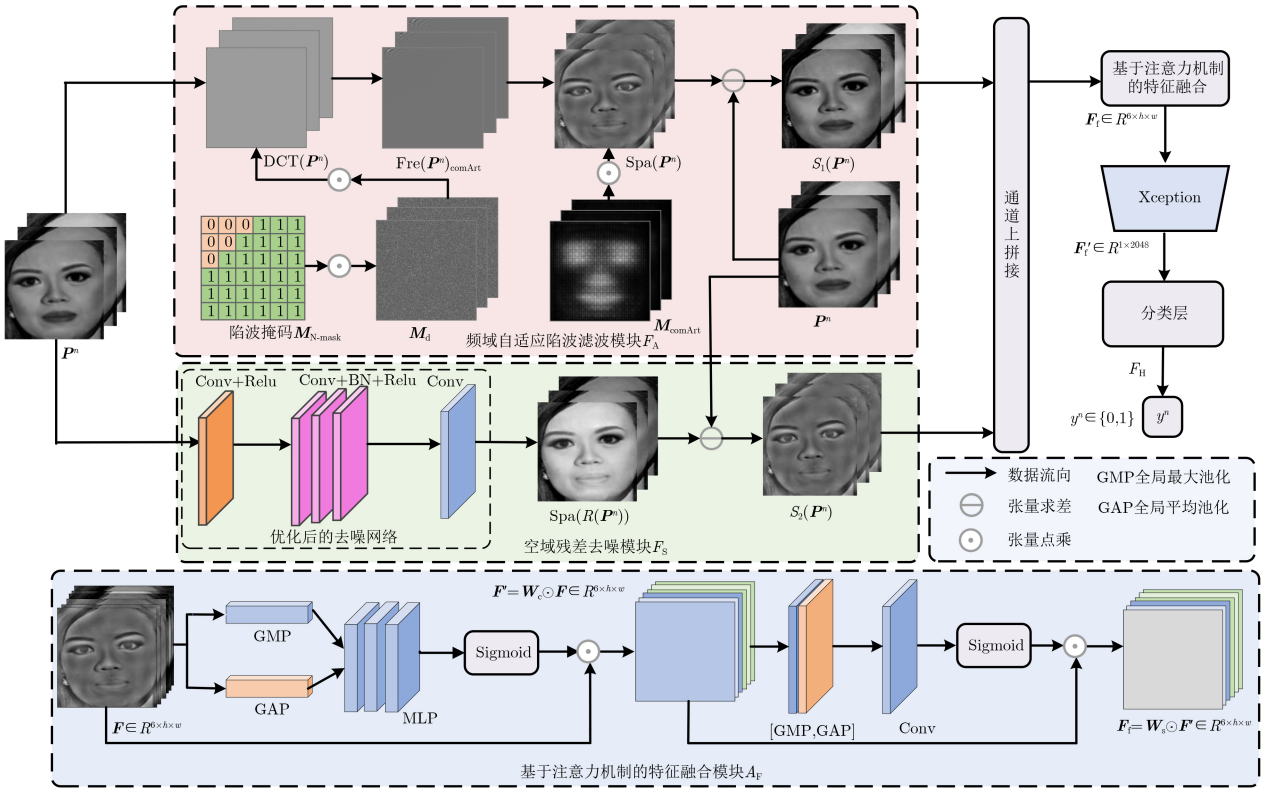


图 2 伪影间共性机理驱动的多域感知社交网络深度伪造视频检测模型

具体地，对于给定的视频V，首先，将视频进行分帧处理以及人脸提取，得到$\{\boldsymbol{P}^n \in R^{c\times h\times w} | n = 1,2,\cdots,T\}$，其中$c$, h和w分别为人脸图像的通道数、高度及宽度，T为帧总数，n为帧序号。其次，频域自适应陷波滤波器F_{A}将输入人脸图像$\boldsymbol{P}^n$映射为去除压缩噪声后的特征图。空域残差去噪模块F_{S}将人脸图像$\boldsymbol{P}^n$映射为去除未知噪声后的特征图。进一步，基于注意力机制的融合模块A_{F}将空频域特征进行融合得到特征$\boldsymbol{F}_{\mathrm{f}}$。然后，深度伪造特征提取器$F_{\mathrm{C}}$将$\boldsymbol{F}_{\mathrm{f}}$映射到特征$\boldsymbol{F}_{\mathrm{f}}^{'}$。最后，分类层$F_{\mathrm{H}}$将特征$\boldsymbol{F}_{\mathrm{f}}^{'}$映射为标签$y^n \in \{0,1\}$，实现压缩深度伪造视频检测。该过程可形式化为

$$y^n = F_{\mathrm{H}}(F_{\mathrm{C}}(A_{\mathrm{F}}([F_{\mathrm{A}}(\boldsymbol{P}^n), F_{\mathrm{S}}(\boldsymbol{P}^n)]))) \tag{9}$$

其中[,]表示拼接操作。

4.1 频域自适应陷波滤波

基于压缩伪影与深度伪造伪影间的结构相似性及其结构的周期性，结合陷波滤波可滤除周期性噪声的先验知识，设计了频域自适应陷波滤波模块，以消除压缩伪影的干扰性，充分挖掘深度伪造判别特征，增强深度伪造检测性能。

由于压缩操作的影响，只关注空域的深度伪造检测算法在低质量压缩视频上有较大的性能下降，而在频域通常会存在深度伪造视频合成过程中上采样操作留下的伪影。因此，将提取的人脸图像$\boldsymbol{P}^n$作为输入，使用DCT变换将输入人脸图像从空域变换到频域，生成输入图像的频谱图$\mathrm{Fre}(\boldsymbol{P}^n) = \mathrm{DCT}(\boldsymbol{P}^n)$，其中$\mathrm{Fre}(\boldsymbol{P}^n) \in R^{3\times h\times w}$。此外，上采样操作是深度伪造视频生成过程中必不可少的步骤，第3节证明了压缩伪影与深度伪造伪影的结构相似性。因此，为去除压缩伪影的干扰，设计了频域自适应陷波滤波分支。将自适应压缩伪影增强图$\boldsymbol{M}_{\mathrm{d}} \in R^{w\times h\times 3}$作用于变换后的频谱图$\mathrm{Fre}(\boldsymbol{P}^n)$以增强压缩伪影特征，可形式化为

$$\mathrm{Fre}(\boldsymbol{P}^n)_{\mathrm{comArt}} = \mathrm{Fre}(\boldsymbol{P}^n) \odot \boldsymbol{M}_{\mathrm{d}} \odot \boldsymbol{M}_{\text{N-mask}} \tag{10}$$

其中，$\mathrm{Fre}(\boldsymbol{P}^n)$为输入图像频谱图，$\boldsymbol{M}_{\mathrm{d}}$为通过网络训练的自适应压缩伪影增强图，$\odot$表示对应元素相乘。由于低频部分为图像的语义信息，对检测具有一定的干扰性。因此将频谱图中的低频部分置为0且不参与训练。将陷波掩码矩阵$\boldsymbol{M}_{\text{N-mask}}$作用于自适应伪影增强图$\boldsymbol{M}_{\mathrm{d}}$以去除低频信息的干扰，$\boldsymbol{M}_{\mathrm{d}}$的其余部分经过不断地迭代训练进行更新。进一步，将增强的压缩伪影频谱图使用IDCT反变换到像素域得到$\mathrm{Spa}(\boldsymbol{P}^n)$，并将可学习的噪声扰动加权图$\boldsymbol{M}_{\mathrm{comArt}}$作用于变换后的像素域特征图$\mathrm{Spa}(\boldsymbol{P}^n)$，得到像素域压缩扰动图

$$\mathrm{Spa}(\boldsymbol{P}^n)_{\mathrm{comArt}} = \mathrm{Spa}(\boldsymbol{P}^n) \odot \boldsymbol{M}_{\mathrm{comArt}} \tag{11}$$

将输入图像与压缩伪影扰动图求残差以消除压缩伪影对分类器的扰动

$$S_1(\boldsymbol{P}^n) = \mathrm{Spa}(\boldsymbol{P}^n) - \mathrm{Spa}(\boldsymbol{P}^n)_{\mathrm{comArt}} \tag{12}$$

4.2 空域残差去噪策略

为削弱深度伪造检测模型对未知噪声的敏感，充分挖掘深度伪造判别特征。本节采用空域残差去噪策略去除未知噪声的干扰。具体地，优化了去噪卷积网络(Denoising Convolutional Neural Network, DnCNN)的输出，将其优化为输入图像与噪声图像的残差，可使模型更有利于空间噪声的去除，且具有更快的收敛速度。将其形式化为

$$S_2(\boldsymbol{P}^n) = \mathrm{Spa}(\boldsymbol{P}^n) - \mathrm{Spa}(R(\boldsymbol{P}^n)) \tag{13}$$

其中，$\mathrm{Spa}(\boldsymbol{P}^n)$表示输入图像，$\mathrm{Spa}(R(\boldsymbol{P}^n))$表示优化后的DnCNN提取的噪声图像，$S_2(\boldsymbol{P}^n)$表示空域去噪后的图像。

4.3 基于注意力机制的特征融合

本节采用注意力机制来融合和挖掘频域分支与空域分支特征图间的内在关联关系，生成深度伪造特征加权图。受卷积注意力模块(Convoluti-onal Block Attention Module, CBAM)[18]的启发，首先在通道上拼接生成特征图$\boldsymbol{F} \in R^{6\times h\times w}$。然后采用通道注意力获得通道注意力图$\boldsymbol{W}_{\mathrm{c}} \in R^{6\times 1\times 1}$，该注意力图由全局平均池化(Global Average Pooling, GAP)、最大池化(Global Max Pooling, GMP)和多层感知器(Multilayer Perceptron, MLP)组成。因此，通道优化后的特征图为$\boldsymbol{F}' = \boldsymbol{W}_{\mathrm{c}} \odot \boldsymbol{F} \in R^{6\times h\times w}$。采用特征间的空间关系生成空间注意力图$\boldsymbol{W}_{\mathrm{s}} \in R^{1\times h\times w}$，得到融合后的特征图$\boldsymbol{F}_{\mathrm{f}} = \boldsymbol{W}_{\mathrm{s}} \odot \boldsymbol{F}' \in R^{6\times h\times w}$。将其输入Xception主干网络中获得深度伪造特征图$\boldsymbol{F}'_{\mathrm{f}} \in R^{1\times 2048}$。

5 实验结果与分析

5.1 数据集

本文使用FaceForensics++(FF++)[19]数据集进行实验。该数据集的视频使用社交网络平台常用的H.264编码器进行编码，具有4种类型的伪造视频，分别为DeepFakes(DF), Face2Face(F2F), FaceSwap(FS)和NeuralTextures(NT)。包含了不同压缩质量因子的视频：原始(RAW)、高质量(C23)、低质量(C40)，整个数据集共有15 000个视频。

5.2 实验环境与度量指标

本文在Ubuntu16.04LTS, PyTorch 1.0.1, Python3.6 以及 NVIDIA TITAN Xp环境进行实验。采用准确率(ACCuracy, ACC)和曲线下的面积(Area Under Curve, AUC)作为评估指标。

5.3 实验细节

本文实验将FF++数据集中的每种伪造类型数据中的720个视频用于训练，140个视频用于验证，140个视频用于测试。在每个视频中随机选择64帧，采用多任务卷积神经网络(Multi-Task cascaded Convolutional Neural Networks, MTCNN)[20]截取人脸图像，并将其大小调整为256 × 256。采用Adam优化器优化网络模型，模型学习率为0.000 2，Epoch为50。使用交叉熵和对比损失函数联合优化网络模型，平衡超参数λ值为0.9。

5.4 数据集内评估

为验证所提算法的有效性，本节将其与现有的4种深度伪造检测算法进行比较。由于本文的主要目标是检测低质量压缩的深度伪造。因此，在不同质量的压缩数据集上评估所提出的方法。如表1和表2所示，所提算法在高压缩数据集C40上的AUC值和ACC值分别达到了98.73%和95.17%，优于基线方法。主要是由于本文深入分析了压缩伪影与深度伪造伪影的结构相似性，针对性地设计了频域自适应陷波滤波器，消除了频域中压缩伪影的干扰性，增强了深度伪造判别特征。

5.5 跨质量评估

为评估所提出算法在跨压缩质量上的有效性，本节进行了跨压缩质量实验。如表3和表4所示，当使用C40视频进行训练C23视频进行测试时，检测结果都相对较高。这是因为C23视频与C40视频相比具有更高的质量，能够拟合C40视频的深度伪造特征，甚至可提供额外的深度伪造特征。本文所提出方法的AUC值可高达99.31%，是因为所提出算法充分考虑到压缩伪影对检测结果的影响，在频域和空域上尽可能消除了各类伪影的干扰，增强了深

表 1 本文算法与基线算法在4个数据集C23上性能的比较(%)

方法	DF		F2F		FS		NT	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Xception[5]	97.90	99.45	95.54	99.46	95.81	99.02	88.92	95.59
Capsule[21]	98.25	99.32	98.34	98.96	98.46	99.71	88.90	94.46
RFM[22]	89.56	98.42	94.09	98.63	90.15	99.23	87.31	94.60
F3-Net[14]	97.34	99.68	97.71	99.44	98.18	99.78	89.12	95.64
本文	**99.02**	**99.94**	**98.58**	**99.63**	**99.21**	**99.89**	**91.79**	**97.12**

表 2 本文算法与基线算法在4个数据集C40上性能的比较(%)

方法	DF		F2F		FS		NT	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Xception[5]	90.11	96.46	79.81	89.53	86.00	93.89	51.30	54.04
Capsule[21]	89.37	94.81	81.97	89.52	83.66	90.36	55.31	58.12
RFM[22]	86.91	93.78	73.75	82.65	78.06	89.42	**61.25**	63.05
F3-Net[14]	92.12	97.78	84.72	92.84	88.84	95.13	58.28	63.28
本文	**95.17**	**98.73**	**86.87**	**93.84**	**90.19**	**96.47**	59.42	**66.97**

度伪造特征的提取。当使用C23视频训练而使用C40视频进行测试时，所有算法分类结果的AUC值都有所下降，是因为C40视频在压缩编码过程中丢失了大量信息尤其是高频信息，而这些信息在深度伪造检测中具有一定的作用，尤其是高频信息。然而，所提出方法在这种情况下也取得了良好的结果，表明本文算法在压缩的深度伪造检测数据集上具有鲁棒性。

5.6 抗噪鲁棒性

为验证所提出算法对不同类型噪声的鲁棒性，本文采用高斯噪声作为加性噪声代表，散斑噪声作为乘性噪声代表进行实验。如图3所示，向深度伪造图像中添加随机噪声可影响基线方法的性能，极小幅度的噪声都会导致分类器性能的下降。而本文提出的算法对噪声具有较强的鲁棒性。

5.7 方法的通用性

为验证所提算法的通用性，本节将该算法作为基线方法的抗压缩前端预处理方法进行实验。考虑到不同的基线模型输入图像大小不同，为避免Resize操作对模型检测结果的影响，针对输入图像小于256×256的基线方法，通过零填充将图像转换为模型所需输入大小。针对输入图像大于256×256的基线方法，采用letterbox方法自适应缩放图像以最小化特征信息的损失。实验结果如表5所示，以本文所提出算法作为前端预处理方法，基线方法的AUC值有所提高。主要是因为本文设计的算法以压缩伪影与深度伪造伪影结构相似性机理分析为指导，有针对性地消除压缩伪影，可提高检测模型的抗压缩鲁棒性。

5.8 特征图可视化

为进一步说明算法的有效性，本节比较了所提出算法与Xception网络模型在低质量压缩图像(C40)上学习到的特征映射。梯度加权类激活映射是一种高级的可视化方法，通过生成热图来增强网络的视觉解释。因此，本文采用梯度加权类激活映射(Gradient-weighted Class Activation Mapping, Grad-CAM)[23]对特征进行可视化，从图4可看出，本文设计的算法提取的特征图更接近基准真实值(Ground Truth, GT)和掩码图像伪造区域，这表明本文所提方法可更好地表示伪造特征，揭示更丰富的细节。

6 结束语

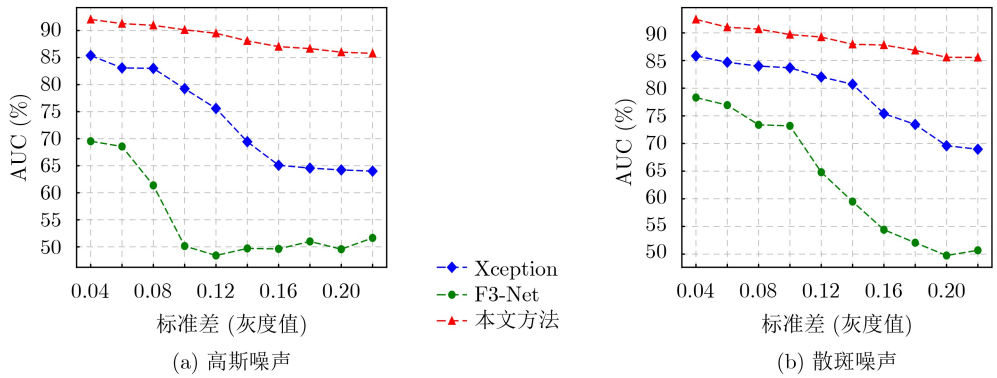
本文从频谱结构与其位置信息分布分析了压缩伪影与深度伪造伪影间的共性机理，证明了伪影间的结构相似性，提出了伪影间共性机理驱动的多域感知社交网络深度伪造视频检测算法。该算法以伪影间的结构相似性机理分析为指导，设计了频域自适应陷波滤波器，消除压缩伪影的干扰。结合空域残差去噪策略，削弱对未知噪声的敏感。采用基于注意力机制的特征融合方法增强深度伪造特征表

表 3　本文算法与基线算法在各数据集的C40上训练C23上测试的性能比较(%)

方法	DF		F2F		FS		NT	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Xception[5]	89.78	96.95	82.95	95.16	89.81	95.97	58.28	63.15
Capsule[21]	87.56	95.17	88.37	93.84	89.78	95.00	55.87	60.91
RFM[22]	89.56	95.76	77.15	86.00	82.71	91.39	65.18	69.53
F3-Net[14]	92.81	98.71	89.56	97.04	89.96	97.56	62.09	70.72
本文	**95.65**	**99.31**	**90.42**	**95.71**	**93.01**	**98.01**	**66.89**	**71.61**

表 4　本文算法与基线算法在各数据集的C23上训练C40上测试的性能比较(%)

方法	DF		F2F		FS		NT	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Xception[5]	76.54	94.87	55.23	84.30	79.42	86.00	50.21	51.29
Capsule[21]	79.87	90.68	58.78	77.11	73.68	82.63	**58.40**	67.84
RFM[22]	75.90	90.07	73.75	84.86	69.15	82.67	56.09	62.88
F3-Net[14]	77.12	94.45	57.18	82.45	78.87	88.48	54.31	68.61
本文	**88.86**	**95.85**	**75.92**	**86.04**	**84.57**	**92.90**	54.74	**67.96**

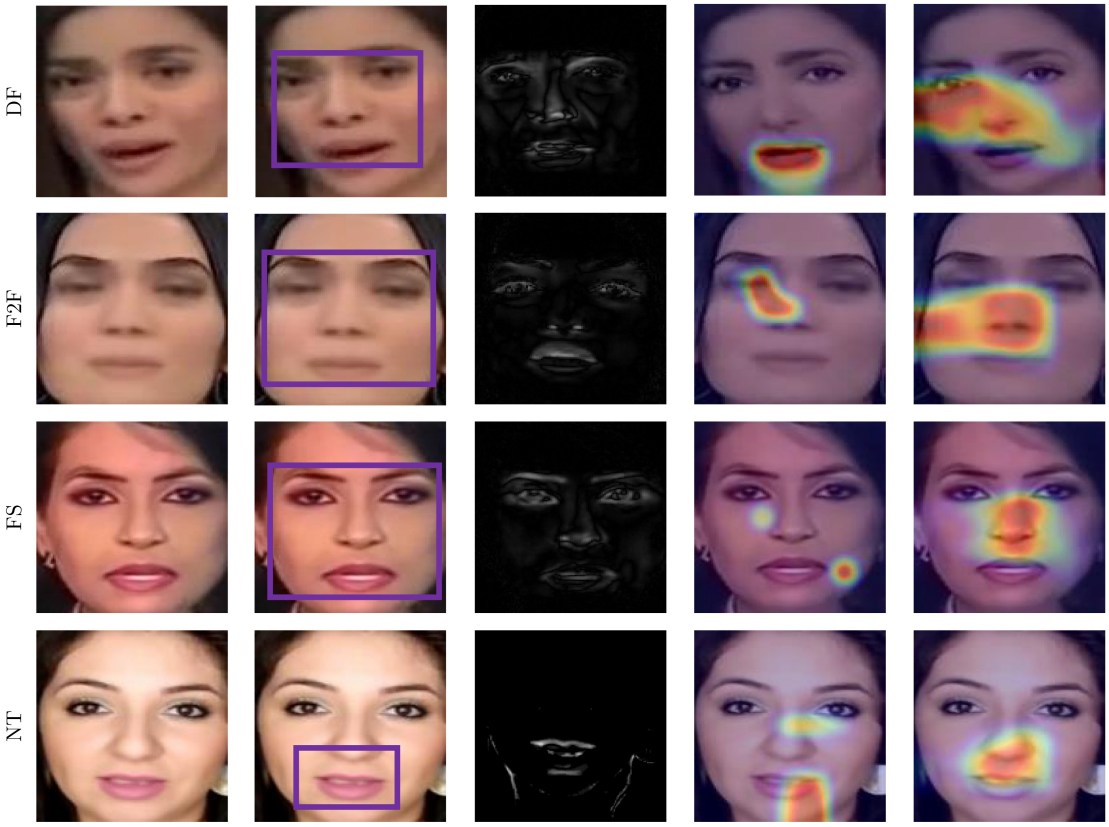


(a) 高斯噪声　　(b) 散斑噪声

图 3 抗噪鲁棒性

表 5 算法的通用性(%)

方法	DF		F2F		FS		NT	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Capsule[21]	94.19	97.43	83.84	90.22	90.50	95.93	58.22	63.23
RFM[22]	90.66	96.90	79.34	87.18	88.08	95.36	54.37	58.95
F3-Net[14]	94.66	98.20	85.34	93.08	90.47	95.21	60.44	65.96



(a) 输入图像 (b) GT (c) 掩码图像 (d) Xception (e) 本文方法

图 4 Grad-CAM可视化

示。理论分析与实验结果表明，本文算法在高压缩的深度伪造检测上明显优于基线方法，而且可作为一种即插即用模型与现有检测算法结合以提高其鲁棒性。

人工智能的发展促使未知深度伪造生成算法层出不穷，后续研究内容为分析不同深度伪造生成算法的共性机理，设计更具泛化性的检测模型，以提升抵抗未知深度伪造生成类型的鲁棒性。

参 考 文 献

[1] IGLESIAS G, TALAVERA E, and DÍAZ-ÁLVAREZ A. A survey on GANs for computer vision: Recent research, analysis and taxonomy[J]. *Computer Science Review*, 2023, 48: 100553. doi: 10.1016/j.cosrev.2023.100553.

[2] ZANARDELLI M, GUERRINI F, LEONARDI R, *et al*. Image forgery detection: A survey of recent deep-learning approaches[J]. *Multimedia Tools and Applications*, 2023, 82(12): 17521–17566. doi: 10.1007/s11042-022-13797-w.

[3] 牛淑芬, 戈鹏, 宋蜜, 等. 移动社交网络中基于属性加密的隐私保护方案[J]. 电子与信息学报, 2023, 45(3): 847–855. doi: 10.11999/JEIT221174.

NIU Shufen, GE Peng, SONG Mi, *et al*. A privacy protection scheme based on attribute encryption in mobile social networks[J]. *Journal of Electronics & Information Technology*, 2023, 45(3): 847–855. doi: 10.11999/JEIT221174.

[4] KINGRA S, AGGARWAL N, and KAUR N. Emergence of deepfakes and video tampering detection approaches: A survey[J]. *Multimedia Tools and Applications*, 2023, 82(7): 10165–10209. doi: 10.1007/s11042-022-13100-x.

[5] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),

Honolulu, USA, 2017: 1800–1807. doi: 10.1109/CVPR.2017.195.

[6] BINH L M and WOO S. ADD: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images[C]. The 36th AAAI Conference on Artificial Intelligence, 2022: 122–130. doi: 10.1609/aaai.v36i1.19886.

[7] LEE S, AN J, and WOO S S. BZNet: Unsupervised multi-scale branch zooming network for detecting low-quality deepfake videos[C]. The ACM Web Conference 2022, Lyon, France, 2022: 3500–3510. doi: 10.1145/3485447.3512245.

[8] LIN Liqun, ZHENG Yang, CHEN Weiling, *et al*. Saliency-aware spatio-temporal artifact detection for compressed video quality assessment[J]. *IEEE Signal Processing Letters*, 2023, 30: 693–697. doi: 10.1109/LSP.2023.3283541.

[9] FRANK J, EISENHOFER T, SCHÖNHERR L, *et al*. Leveraging frequency analysis for deep fake image recognition[C]. The 37th International Conference on Machine Learning, 2020: 3247–3258.

[10] LI Yuezun, CHANG M C, and LYU Siwei. In ictu oculi: Exposing AI created fake videos by detecting eye blinking[C]. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 2018: 1–7. doi: 10.1109/WIFS.2018.8630787.

[11] CIFTCI U A, DEMIR I, and YIN Lijun. Fakecatcher: Detection of synthetic portrait videos using biological signals[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020: 1. doi: 10.1109/TPAMI.2020.3009287.

[12] COZZOLINO D, THIES J, RÖSSLER A, *et al*. ForensicTransfer: Weakly-supervised domain adaptation for forgery detection[EB/OL]. https://arxiv.org/abs/1812.02510, 2018.

[13] AFCHAR D, NOZICK V, YAMAGISHI J, *et al*. MesoNet: A compact facial video forgery detection network[C]. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 2018: 1–7. doi: 10.1109/WIFS.2018.8630761.

[14] QIAN Yuyang, YIN Guojun, SHENG Lu, *et al*. Thinking in frequency: Face forgery detection by mining frequency-aware clues[C]. 16th European Conference on Computer Vision, Glasgow, UK, 2020: 86–103. doi: 10.1007/978-3-030-58610-2_6.

[15] LIU Honggu, LI Xiaodan, ZHOU Wenbo, *et al*. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, USA, 2021: 772–781. doi: 10.1109/CVPR46437.2021.00083.

[16] LI Ying, BIAN Shan, WANG Chuntao, *et al*. Exposing low-quality deepfake videos of social network service using spatial restored detection framework[J]. *Expert Systems with Applications*, 2023, 231: 120646. doi: 10.1016/j.eswa.2023.120646.

[17] ZHANG Yun, ZHU Linwei, JIANG Gangyi, *et al*. A survey on perceptually optimized video coding[J]. *ACM Computing Surveys*, 2023, 55(12): 245. doi: 10.1145/3571727.

[18] WOO S, PARK J, LEE J Y, *et al*. CBAM: Convolutional block attention module[C]. 15th European Conference on Computer Vision, Munich, Germany, 2018: 3–19. doi: 10.1007/978-3-030-01234-2_1.

[19] RÖSSLER A, COZZOLINO D, VERDOLIVA L, *et al*. FaceForensics++: Learning to detect manipulated facial images[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 2019: 1–11. doi: 10.1109/ICCV.2019.00009.

[20] ZHANG Kaipeng, ZHANG Zhanpeng, LI Zhifeng, *et al*. Joint face detection and alignment using multitask cascaded convolutional networks[J]. *IEEE Signal Processing Letters*, 2016, 23(10): 1499–1503. doi: 10.1109/LSP.2016.2603342.

[21] NGUYEN H H, YAMAGISHI J, and ECHIZEN I. Capsule-forensics: Using capsule networks to detect forged images and videos[C]. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019: 2307–2311. doi: 10.1109/ICASSP.2019.8682602.

[22] WANG Chengrui and DENG Weihong. Representative forgery mining for fake face detection[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, USA, 2021: 14918–14927. doi: 10.1109/CVPR46437.2021.01468.

[23] SELVARAJU R R, COGSWELL M, DAS A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 618–626. doi: 10.1109/ICCV.2017.74.

王　艳：女，博士生，研究方向为多媒体取证、人工智能安全等.
孙钦东：男，教授，研究方向为社交网络用户行为分析、多媒体取证、人工智能安全、物联网系统安全、网络攻防技术等.
荣东柱：男，博士生，研究方向为多媒体取证、人工智能安全等.
汪小雄：男，硕士生，研究方向为多媒体取证、人工智能安全等.

责任编辑：余　蓉