doi：10. 12068/j. issn. 1005－3026. 2025. 20249048

基于面部掩码引导的多人场景图像伪造定位算法

刘佳彤，王丽娜，汪 润，叶 茜
（武汉大学 国家网络安全学院 空天信息安全与可信计算教育部重点实验室，湖北 武汉 430070）

摘 要：为解决现有伪造定位算法在小区域面部篡改的多人场景图像时性能下降、鲁棒性不足的问题，提出一种基于面部掩码引导的伪造定位模型FMG-L. 首先，为了减轻多人场景图像中背景信息的干扰，设计面部掩码引导模块，鼓励FMG-L关注重要的面部区域；其次，为了提升FMG-L面对图像质量退化的鲁棒性，设计三通道特征提取模块提取多维特征，结合基于双重注意力网络的特征融合模块，增强模型学习到的伪造线索；最后，使用伪造区域定位模块进行伪造定位. 在OpenForensics，ManulFake，FFIW和DiffSwap数据集上的实验结果表明，FMG-L能够有效进行伪造定位，具有面对多种图像退化和不同在线社交平台的强鲁棒性.

关 键 词：深度伪造；深度伪造定位；多人场景图像；小区域篡改；面部掩码引导
中图分类号：TP 391.4 **文献标志码**：A **文章编号**：1005-3026(2025)05-0010-10

Facial Mask Guidance Based Multi-person Scene Images Forgery Localization Algorithm

LIU Jia-tong, *WANG Li-na*, *WANG Run*, *YE Xi*
(Key Laboratory of Aerospace Information Security and Trusted Computing (Ministry of Education), School of Cyber Science and Engineering, Wuhan University, Wuhan 430070, China. Corresponding author: WANG Li-na, E-mail: lnwang@whu.edu.cn)

Abstract: To address the performance degradation and lack of robustness in existing forgery localization models when dealing with small region facial manipulations in multi-person scene images, a FMG-L model based on facial mask guidance for forgery localization is proposed. Firstly, to mitigate interference from background information in multi-person scene images, a facial mask guidance module is designed to encourage the model to focus on critical facial regions. Secondly, to enhance the robustness against image degradations, a three-channel feature extraction module is developed to extract multi-dimensional features, and a feature fusion module based on a dual attention network is also designed to enhance the forgery clues. Finally, a forgery localization module is used for forgery localization. Experimental results on the OpenForensics, ManulFake, FFIW, and DiffSwap datasets demonstrate that the FMG-L effectively localizes forgery regions and shows strong robustness against various image degradations and different online social platforms.

Key words: DeepFakes; DeepFake localization; multi-person scene images; small region manipulations; facial mask guidance

如今，社会公众每天通过社交媒体网络交流、传递以及获取时事信息. 与文字描述相比，视觉信息丰富的图像和视频更容易被关注和相信. 然而，随着生成对抗网络（GANs）[1]和扩散模型[2]的出现，越来越多免费的图像合成工具被滥用，在没有技术门槛的情况下，任何人都可以随意篡改以名人为代表的伪造图像和视频，这类技术统称为DeepFakes[3]. DeepFakes被广泛应用于社会和政治领域，引发公众对网络欺诈和政府信誉的担忧. 因此，亟需研究稳定且有效的深度伪造防

收稿日期：2024-10-10
基金项目：国家自然科学基金资助项目（62372334）；国家重点研发计划项目（2023YFB3106900）.
作者简介：刘佳彤（1995—），女，河南桐柏人，武汉大学博士研究生；王丽娜（1964—），女，辽宁沈阳人，武汉大学教授，博士生导师.

御方法,维护社会稳定.

现有的深度伪造防御方法较少考虑面部伪造区域定位的问题,定位任务在多媒体取证领域显得尤为重要,能够指明图像中伪造的像素区域,预判攻击者的意图.攻击者为了提升深度伪造图像的可信度,往往会选用具有复杂背景和多人物的图像,对其中一个或多个目标人物进行篡改[4].现有的深度伪造定位方法仅在实验室环境中忽略背景信息和单一清晰正面的人脸图像中进行训练和评估[5],伪造人脸区域通常占据图像中的较大部分,伪造特征显著.然而多人场景图像通常包含复杂的背景信息,伪造区域在复杂的背景中使得伪造特征的显著性降低,增加了定位模型的识别难度[4].

除此之外,当现有伪造定位模型在现实世界中部署时,攻击者为了扩大深度伪造人脸图像的影响力,往往通过社交媒体平台对深度伪造图像进行传播,经过社交平台的上传和下载过程后,图像会经历已知或未知的多种图像退化处理(如压缩和模糊等).这些退化操作可能破坏模型定位所依赖的某种微弱的伪造痕迹,导致模型无法定位伪造区域.

为了解决上述问题,本文探索一种三通道的网络架构,能够有效地执行面向多人场景图像的伪造人脸定位任务.通过结合RGB特征、面部区域特征和噪声特征的多源特征,该架构能够从多个维度捕捉图像中的伪造线索,减少模型对特定伪造特征的依赖,达到优越的定位性能.为了减轻多人场景图像中复杂背景的干扰,本文设计了基于图像分割模型的FMG模块,通过FMG模块输出的面部区域掩码图像引导模型关注多人场景图像中的重点伪造区域,避免复杂背景带来的干扰.为了提升定位模型的鲁棒性,本文设计三通道的特征提取网络捕获多维度的伪造线索,并设计基于双重注意力网络(dual attention network,DAN)的特征融合模块,分别学习三通道融合特征的空间位置关系和通道依赖关系,增强伪造线索的特征表示,提升模型的鲁棒性.

1　模　　型

1.1　问题定义

给定一张多人场景的原始图像 $x^{\mathrm{rgb}} \in \mathbf{R}^{3\times H\times W}$ 和通过预训练图像分割模型得到的面部区域掩码 $\boldsymbol{M}_{\mathrm{p}} \in \mathbf{R}^{H\times W}$,其中 t 代表图像中人脸的个数.本文目标是通过模型FMG-L(网络结构如图1所示)预测一张伪造区域掩码 $\boldsymbol{p}_{m,n} \in (0,1)^{H\times W}$,其中 H 和 W 分别代表输入图像的长和宽.

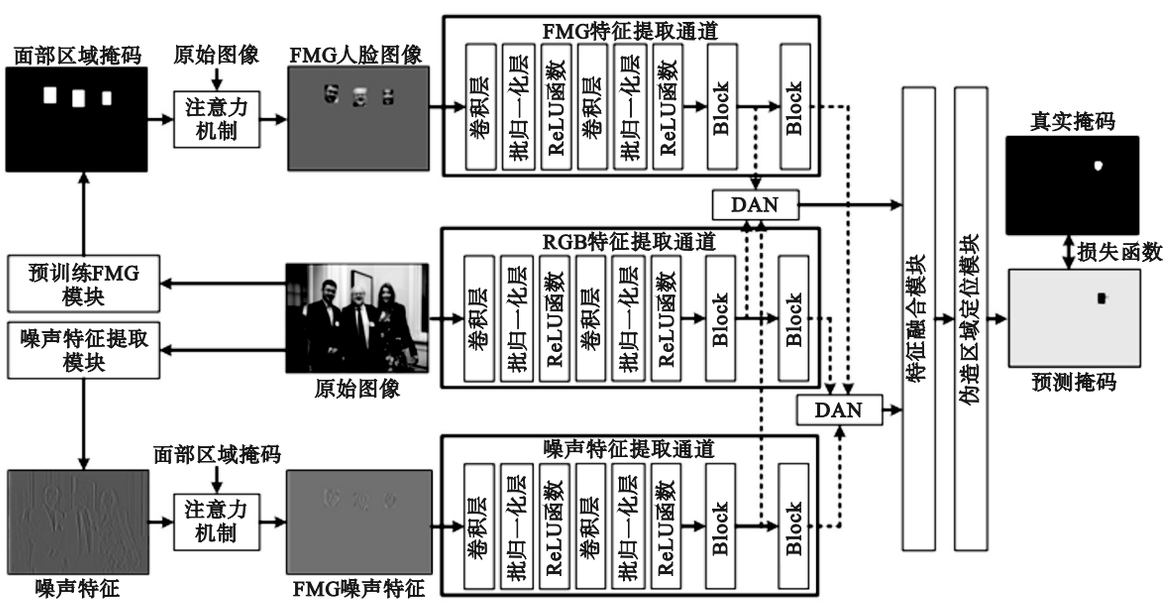

图1　FMG-L模型网络结构

Fig. 1　Network architecture of FMG-L model

1.2　特征提取模块

为了捕捉多种伪造线索以保证模型的鲁棒性,设计三通道的特征提取模块,分别为FMG特征提取通道 C_1,RGB特征提取通道 C_2 以及噪声特征提取通道 C_3,三通道具有相同的网络结构,RGB通道使用原始图像作为输入,FMG通道和

噪声通道分别通过FMG模块和噪声特征提取模块获取各自的输入.

1.2.1 FMG模块

由于现有的深度伪造定位方法仅在实验室环境中忽略背景信息和单一清晰正面的人脸图像中进行训练和评估,伪造人脸区域通常占据图像的较大部分,如图2a所示,更容易被定位模型捕捉.然而多人场景图像通常包含复杂的背景信息,不同人物的面部大小、角度及姿势都不统一,且伪造区域仅占据图像中的一小部分,如图2b所示,伪造区域在复杂的背景中,伪造特征的显著性降低.为了聚焦伪造区域,减少计算复杂度和无关区域的干扰,本文使用预训练的图像分割模型Mask-RCNN[6]在OpenForensics[4](OF)数据集上进行微调,获得能够在多人场景图像中提取所有面部区域的提取网络.

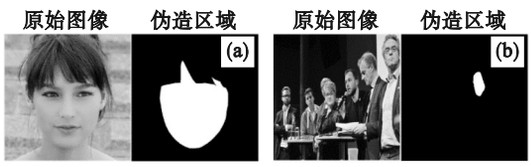

图2 单人脸数据集与多人场景数据集的伪造区域对比

Fig. 2 Comparison of forgery regions between single-face dataset and multi-face scene dataset

(a)—单人脸数据集;(b)—多人场景数据集.

首先,获取OF数据集中的多人场景图像x^{rgb},面部区域分割掩码$\boldsymbol{M}_{\mathrm{t}}^{a,b}$,面部区域边界框$\mathbf{bbox}_{\mathrm{t}}$,面部和背景的类别标签$\boldsymbol{y}_{\mathrm{t}}$,其中$a$和$b$分别代表不同面部区域的像素坐标.然后,将原始图像$x^{\mathrm{rgb}}$输入卷积神经网络提取特征图$f_x$,特征图$f_x$通过区域候选网络生成一组区域候选框$\boldsymbol{R}_{\mathrm{r}}$,对每个候选框$\boldsymbol{r}\in\boldsymbol{R}_{\mathrm{r}}$进行感兴趣区域池化RoIPool后,得到特征图$f_{\mathrm{roi}}$,可以表示为

$$f_{\mathrm{roi}}=\mathrm{RoIPool}(\mathrm{ResNet}(x^{\mathrm{rgb}}),\boldsymbol{r}). \tag{1}$$

得到f_{roi}后,将其通过全连接层分别预测类别$\boldsymbol{y}_{\mathrm{p}}$和面部区域边框$\mathbf{bbox}_{\mathrm{p}}$,并且在每个候选框内进一步通过卷积层MaskHead(.)生成像素级的面部区域分割掩码$\boldsymbol{M}_{\mathrm{p}}^{a,b}$,其公式为

$$\boldsymbol{y}_{\mathrm{p}}=\mathrm{Softmax}(\boldsymbol{W}_{\mathrm{s}}\times f_{\mathrm{roi}}+\boldsymbol{b}_{\mathrm{s}}), \tag{2}$$

$$\mathbf{bbox}_{\mathrm{p}}=\boldsymbol{W}_{\mathrm{c}}\times f_{\mathrm{roi}}+\boldsymbol{b}_{\mathrm{c}}, \tag{3}$$

$$\boldsymbol{M}_{\mathrm{p}}^{a,b}=\mathrm{MaskHead}(f_{\mathrm{roi}}). \tag{4}$$

其中:$\boldsymbol{W}_{\mathrm{s}}$和$\boldsymbol{W}_{\mathrm{c}}$是全连接层中的权重矩阵;$\boldsymbol{b}_{\mathrm{s}}$和$\boldsymbol{b}_{\mathrm{c}}$是全连接层中的偏置向量.

FMG模块微调过程中的训练损失函数包括分类损失ζ_{c},边框预测损失ζ_{b}和掩码损失ζ_{m}.这3个损失函数分别用来优化模型预测的人脸和背景类别、模型预测的面部区域边框以及模型预测的面部区域掩码,可以表示为

$$\zeta_{\mathrm{c}}=-\sum_i(\boldsymbol{y}_{\mathrm{t}}^i\ln\left(\boldsymbol{y}_{\mathrm{p}}^i\right)+(1-\boldsymbol{y}_{\mathrm{t}}^i)\ln(1-\boldsymbol{y}_{\mathrm{p}}^i)), \tag{5}$$

$$\zeta_{\mathrm{b}}=\sum_i\mathrm{SmoothL1}(\mathbf{bbox}_{\mathrm{p}}^i-\mathbf{bbox}_{\mathrm{t}}^i), \tag{6}$$

$$\zeta_{\mathrm{m}}=-\sum_{a,b}(\boldsymbol{M}_{\mathrm{t}}^{a,b}\ln\left(\boldsymbol{M}_{\mathrm{p}}^{a,b}\right)+(1-\boldsymbol{M}_{\mathrm{t}}^{a,b})\ln(1-\boldsymbol{M}_{\mathrm{p}}^{a,b})). \tag{7}$$

总损失函数可以表示为

$$\zeta_{\mathrm{HM}}=\zeta_{\mathrm{c}}+\zeta_{\mathrm{b}}+\zeta_{\mathrm{m}}. \tag{8}$$

训练完成的FMG模块可以通过简单注意力机制提取多人场景图像中的FMG人脸图像f_{mask},可以表示为

$$f_{\mathrm{mask}}=x^{\mathrm{rgb}}*\boldsymbol{M}_{\mathrm{p}}. \tag{9}$$

1.2.2 噪声特征提取模块

仅靠RGB特征不足以保障模型的鲁棒性.为了提高模型学习到的伪造线索的多样性,本文提出基于空域富模型(spatial rich model,SRM)[7]的噪声特征提取模块,模块的目标是获取相邻像素间的残差特征.

选取同样的SRM滤波器[7],并将这些滤波器扩充为5×5×3大小,使它们像卷积核一样工作.这3个SRM滤波器可以表示为式(10)~式(12),其中,q_1,q_2和q_3分别是3个滤波器的系数.

$$\boldsymbol{F}_1=\frac{1}{q_1}\cdot\begin{bmatrix}0&0&0&0&0\\0&0&0&0&0\\0&1&-2&1&0\\0&0&0&0&0\\0&0&0&0&0\end{bmatrix}, \tag{10}$$

$$\boldsymbol{F}_2=\frac{1}{q_2}\cdot\begin{bmatrix}-1&2&-2&2&-1\\2&-6&8&-6&2\\-2&8&-12&8&-2\\2&-6&8&-6&2\\-1&2&-2&2&-1\end{bmatrix}, \tag{11}$$

$$\boldsymbol{F}_3=\frac{1}{q_3}\cdot\begin{bmatrix}0&0&0&0&0\\0&-1&2&-1&0\\0&2&-4&2&0\\0&-1&2&-1&0\\0&0&0&0&0\end{bmatrix}. \tag{12}$$

在实际应用中,为了处理输入的三通道RGB图像,构造三维的SRM滤波器$\boldsymbol{F}_i^3$,将上述SRM滤波器在RGB通道上分别重复3次,可以表示为

$$\boldsymbol{F}_i^3=\left[\boldsymbol{F}_i\ \ \boldsymbol{F}_i\ \ \boldsymbol{F}_i\right],\ i=1,2,3. \tag{13}$$

然后使用深度可分离卷积层⊛提取噪声特征,这个过程可以表示为

$$s_i=\boldsymbol{F}_i^3\circledast x^{\mathrm{rgb}}. \tag{14}$$

其中,s_i是滤波器输出的噪声特征图.结合FMG模块,使用简单的注意力提取多人场景图像中的FMG噪声特征s_{mask},可以表示为

$$s_{\text{mask}}=s_i*\boldsymbol{M}_{\text{p}}. \tag{15}$$

1.2.3　三通道网络架构

三通道网络设计包括两个卷积层和多个特征提取块Block,Block的结构如图3所示.

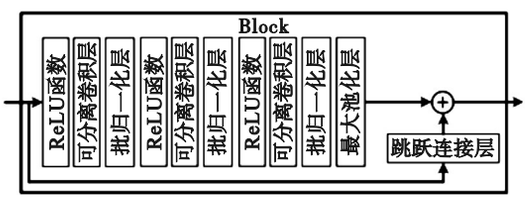

图3　Block网络结构

Fig. 3　Network architecture of Block

以RGB通道为例,给定输入的图像x^{rgb},两个3×3的卷积层将通道维数扩展到64,捕捉输入的局部细节,关注像素级别的模式识别,可以表示为

$$X_{\text{conv}}=\text{Conv}_2(\text{Conv}_1(x^{\text{rgb}};\theta_1);\theta_2). \tag{16}$$

其中:X_{conv}表示卷积层输出的中间特征;θ_1,θ_2表示卷积层的参数集.同理可得FMG通道和噪声通道的中间特征F_{conv}和S_{conv}:

$$F_{\text{conv}}=\text{Conv}_2(\text{Conv}_1(f_{\text{mask}};\theta_1);\theta_2), \tag{17}$$

$$S_{\text{conv}}=\text{Conv}_2(\text{Conv}_1(s_{\text{mask}};\theta_1);\theta_2). \tag{18}$$

然后经过多个Block层,每层包含深度卷积操作和逐点卷积操作.深度卷积操作使用3×3的卷积核对每个输入通道进行单独的卷积运算,提取空间特征;逐点卷积操作使用1×1的卷积核在通道维度上进行卷积,将不同通道信息组合起来.最后,使用一个最大池化层降低特征图的分辨率,这个过程可以表示为

$$X=\text{MaxPool}(\text{Block}_i(X_{\text{conv}};\theta_i);\theta_{\text{mp}}), \tag{19}$$

$$F=\text{MaxPool}(\text{Block}_i(F_{\text{conv}};\theta_i);\theta_{\text{mp}}), \tag{20}$$

$$S=\text{MaxPool}(\text{Block}_i(S_{\text{conv}};\theta_i);\theta_{\text{mp}}). \tag{21}$$

其中:X,F和S分别表示三通道输出的RGB特征、FMG特征和噪声特征;θ_i是第i个Block的参数集;θ_{mp}是最大池化层的参数集.

实际上,本文可以使用更多的Block层关注更大的图像局部区域,由于在多人场景图像中伪造人脸在图像中占据的比例很小,为了关注更小的图像局部信息,本文设置较少的Block层进行特征提取.

1.3　特征融合模块

在这个阶段,三通道特征提取模块输出RGB图像特征X_i、面部区域特征F_i和噪声特征S_i,其中i表示第i个Block层的输出.特征融合模块的目标是融合这些特征并且捕捉融合特征间的关系.首先将RGB特征$X_i=[\boldsymbol{x}_i^1,\boldsymbol{x}_i^2,\cdots,\boldsymbol{x}_i^c]$,FMG特征$F_i=[\boldsymbol{f}_i^1,\boldsymbol{f}_i^2,\cdots,\boldsymbol{f}_i^c]$和噪声特征$S_i=[\boldsymbol{s}_i^1,\boldsymbol{s}_i^2,\cdots,\boldsymbol{s}_i^c]$连接起来,第$i$个Block层的三通道融合特征可以表示为

$$Z_i=\left[X_i,F_i,S_i\right]. \tag{22}$$

式中:$[\cdot,\cdot]$表示空间维度的连接操作;$Z_i=[\boldsymbol{z}_i^1,\boldsymbol{z}_i^2,\cdots,\boldsymbol{z}_i^c]\in\mathbf{R}^{C_i\times H_i\times W_i}$,其中$H_i,W_i$和$C_i$分别表示第$i$个Block层的三通道融合特征的长、宽和通道数.为了增强融合特征的表示能力,本文在特征融合阶段引入DAN[8]结构,并行使用空间注意力模块和通道注意力模块来捕获空间域和通道域的特征依赖关系.

空间注意力模块将任意两个局部特征之间的空间关系进行建模.给定一个输入特征$Z_i\in\mathbf{R}^{C_i\times H_i\times W_i}$,首先将特征输入卷积层分别获得特征图$Z_{\text{a}},Z_{\text{b}}$和$Z_{\text{d}}$,其中$\{Z_{\text{a}},Z_{\text{b}},Z_{\text{d}}\}\in\mathbf{R}^{C_i\times H_i\times W_i}$,将$Z_{\text{a}}$和$Z_{\text{b}}$的形状调整为$\mathbf{R}^{C_i\times N_i}$,其中$N_i=H_i\times W_i$.然后,在$Z_{\text{a}}$和$Z_{\text{b}}$的转置之间应用softmax层计算通道注意特征图$Z_{\text{c1}}\in\mathbf{R}^{N_i\times N_i}$,这个过程可以表示为

$$\boldsymbol{z}_{\text{s1}}^{uv}=\frac{\exp(\boldsymbol{z}_{\text{a}}^u\cdot\boldsymbol{z}_{\text{b}}^v)}{\sum_{u=1}^{N_i}\exp(\boldsymbol{z}_{\text{a}}^u\cdot\boldsymbol{z}_{\text{b}}^v)}. \tag{23}$$

其中,$\boldsymbol{z}_{\text{s1}}^{uv}$表示第$v$个位置对第$u$个位置的影响,两个位置的特征越相似则具有更大的相关性.接下来,将Z_{d}和Z_{s1}的转置之间应用矩阵乘法,并将得到的结果形状调整为$\mathbf{R}^{C_i\times H_i\times W_i}$.最后,将结果乘以尺度参数$\alpha$,并与特征$Z_i$执行逐元素求和运算,得到最终的输出$Z_{\text{s2}}\in\mathbf{R}^{C_i\times H_i\times W_i}$:

$$\boldsymbol{z}_{\text{s2}}^u=\alpha\sum_{u=1}^{N_i}\left(\boldsymbol{z}_{\text{s1}}^{uv}\boldsymbol{z}_{\text{d}}^v\right)+\boldsymbol{z}_i^v. \tag{24}$$

其中,α的初始值设置为0,在训练过程中逐渐增加权重.由于特征Z_{s2}是所有局部特征与原始特征的加权和,因此该特征能够反映全局上下文之间的关系.

通道注意模块通过利用通道映射之间的相互依赖性,可以将不同的语义响应相互关联.与空间注意模块不同,直接根据原始输入特征$Z_i\in\mathbf{R}^{C_i\times H_i\times W_i}$计算通道注意特征图$Z_{\text{c1}}\in\mathbf{R}^{C_i\times C_i}$.具体地,将$Z_i$的形状调整为$\mathbf{R}^{C_i\times N_i}$,在$Z_i$和其转置间执行矩阵乘法,并且应用softmax层计算通道注意特征图Z_{c1},这个过程可以表示为

$$\boldsymbol{z}_{\text{c1}}^{uv}=\frac{\exp(\boldsymbol{z}_i^v\cdot\boldsymbol{z}_i^u)}{\sum_{u=1}^{C_i}\exp(\boldsymbol{z}_i^v\cdot\boldsymbol{z}_i^u)}. \tag{25}$$

其中,$\boldsymbol{z}_{\text{c1}}^{uv}$表示第$v$个通道对第$u$个通道的影响.此

外,本文在Z_{c1}的转置和Z_i之间应用矩阵乘法,并将得到的结果形状调整为$\mathbf{R}^{C_i\times H_i\times W_i}$.最后,将结果乘以尺度参数$\beta$,并与特征$Z_i$执行逐元素的求和运算,得到输出$Z_{c2}\in\mathbf{R}^{C_i\times H_i\times W_i}$,表示为

$$\boldsymbol{z}_{c2}^{u}=\beta\sum_{u=1}^{C_i}\left(\boldsymbol{z}_{c1}^{uv}\boldsymbol{z}_i^{v}\right)+\boldsymbol{z}_i^{u}. \tag{26}$$

其中,β初始值为0,在训练过程中逐渐增加权重.由于通道注意模块将通道特征与原始特征相加,对通道间的长期语义映射关系进行建模,有助于提高局部特征的可辨别性.

通过上述特征融合模块,最终的融合特征可以表示为

$$Z_i=Z_{s2}+Z_{c2}. \tag{27}$$

其中,i代表三通道中第i个Block融合后的特征.该模块可以捕捉融合特征的空间关系和通道关系,更有利于进行伪造区域定位.

得到尺度不同的i个融合特征后,首先将每个特征图的通道数都变换为C_i',然后对所有特征图进行通道上的连接,使用的卷积层将连接后特征通道数降低到C_i'.在减少计算复杂度的同时,保留特征图的语义信息,得到最终的三通道融合特征$Y\in\mathbf{R}^{C_i'\times H_i'\times W_i'}$.

1.4 伪造区域定位模块

在经过特征融合模块得到的特征$Y_{m,n}$上进行伪造定位,对输入的接受域大小的块进行预测和二值分类,而不是对整个图像进行预测,鼓励模型学习区分真实和虚假图像间的局部差异,这个过程可以表示为

$$\boldsymbol{p}_{m,n}=T^{c}(Y_{m,n};\theta_t). \tag{28}$$

其中:$\boldsymbol{p}_{m,n}$表示每个像素的预测标签;(m,n)表示像素的坐标;T^c是1×1的卷积层,将输入特征通道压缩到表示真实和虚假的二值分类输出c;θ_t代表卷积层的参数集.通过伪造区域定位模块得到基于块的预测后,需要引入合适的损失函数监督模型的训练过程.

1.5 损失函数

二分类问题中经常使用交叉熵损失[9]L_{bce}平等地计算每个像素的损失,该损失函数在语义分割领域应用时有一个明显的缺陷:当前景像素数量远小于背景像素时,交叉熵损失的预测会偏向数量大的一方,导致模型训练效果变差.而在多人场景图像中进行伪造人脸区域定位时,极有可能存在正负样本不均衡的问题.

为了解决上述问题,引入Focal损失[10]解决正负样本数量不均衡的问题,将大数量样本的损失权重和高置信度样本的损失权重设置为较小值,该损失函数为

$$\begin{aligned}L_{focal}=&-\sum\lambda\left(1-\boldsymbol{p}_{m,n}\right)^{\gamma}\boldsymbol{t}_{m,n}\ln\left(\boldsymbol{p}_{m,n}\right)-\\&\sum(1-\lambda)\boldsymbol{p}_{m,n}^{\gamma}\left(1-\boldsymbol{t}_{m,n}\right)\ln\left(1-\boldsymbol{p}_{m,n}\right).\end{aligned} \tag{29}$$

其中:$\boldsymbol{p}_{m,n}$和$\boldsymbol{t}_{m,n}$分别表示每个像素的预测标签和真实标签;λ是损失权重;γ是超参数.

为了使模型的训练更加关注对图像中人脸前景的挖掘,本文引入另一个Dice损失[11],用于保证预测掩码和真实掩码之间的交集尽可能地大,这个损失函数为

$$L_{dice}=1-\frac{2\sum\boldsymbol{p}_{m,n}\cdot\boldsymbol{t}_{m,n}}{\sum\boldsymbol{p}_{m,n}^{2}+\sum\boldsymbol{t}_{m,n}^{2}}. \tag{30}$$

单独使用L_{dice}容易产生损失饱和的问题,因此,组合使用L_{focal}和L_{dice},最终的损失函数为

$$L_{total}=L_{focal}+L_{dice}. \tag{31}$$

2 实验设置

2.1 数据集

本文在实验过程中采用公开的多人场景深度伪造数据集和基于扩散模型生成的单人脸伪造数据集,所有的伪造图像都有对应的伪造区域掩码标签.OpenForensics(OF)[4]数据集包含45 473张具有复杂背景的真实多人场景图像,以及通过GAN生成的70 325张伪造图像.ManualFake(MF)[12]数据集包含1 000个原始的多人场景图像、1 000个伪造视频以及4 000个经过社交媒体平台上传和下载过程的伪造视频.FFIW(FW)[13]数据集在野外收集10 000个原始视频,大部分视频中包含不止一个人,使用3种人脸交换方法创建10 000个伪造视频.DiffSwap(DS)[14]数据集使用MM-CelebA-HQ数据集中的真实人脸图像,使用扩散模型进行随机人脸交换,生成30 000张伪造人脸图像.

对于上述的视频数据集,提取其中的视频帧,并划分训练集、验证集和测试集的比例为8∶1∶1.

2.2 对比实验

为了验证本文模型的性能,采用以下具有代表性以及先进的伪造图像定位模型作为对比方案.在实验过程中,所有实验设置与本文提出的模型保持一致.

Grad-CAM[15]通过计算梯度权重来突出显示模型关注的图像区域,因此具有定位的能力.

在本文中,为了量化该方法的定位能力,使用Xception[16]作为基础模型,通过对最后一次降采样操作处应用Grad-CAM赋予模型定位能力.Patch[17]使用具有有限接受域的基于图像块的分类器来可视化伪造图像中的伪造区域.HiFi-Net[18]使用不同级别的多个标签表示伪造图像的伪造属性,并采用层次依赖关系在这些级别上执行细粒度分类,鼓励方法学习不同伪造属性的综合特征和固有的层次性质.Attention[19]在Xception模型基础上增加一个可学习的注意力掩码,用于调制由网络产生的特征映射,在弱监督的情况下能够定位扩散模型生成的伪造图像的伪造区域.

另外,为了验证本文提出的FMG模块的有效性,将使用完整噪声特征和经过FMG模块处理的噪声特征的本文方法分别命名为FMG-F和FMG-L.

2.3　评估指标

在本文的实验中,Accuracy (Acc),F_1-score (F_1) 和 Intersection over Union (IoU)作为评估指标.

Acc定义为模型预测正确的像素数占总像素数的比值,取值范围为0~1,当值为1时说明模型分类像素的准确率最高.

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}. \tag{32}$$

其中:TP表示真正例;TN表示真负例;FP表示假正例;FN表示假负例.

F_1综合考虑了精确率Precision和召回率Recall,取值范围为0~1,当值为1时表示定位表现最好.

$$F_1=2\times\frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}. \tag{33}$$

IoU是真实伪造区域与预测伪造区域的重叠比例,取值范围为0~1,当值为1时表示定位效果最好.

$$\mathrm{IoU}=\frac{\mathrm{Intersection}}{\mathrm{Union}}. \tag{34}$$

文献[20]中使用AUC(area under the ROC curve)作为定位性能的评估指标.然而,在多人场景图像中,篡改像素数量往往远小于真实像素数,导致严重的类别不平衡问题.在这种情况下,ROC曲线会偏向真实像素数,从而影响评估结果.因此,在本文的实验场景中,AUC评估可能给出不可信的结果.

2.4　参数设置

本文提出的伪造图像定位模型基于PyTorch实现,所有输入图像大小调整为512×512,端到端的训练在2块NVIDIA Tesla V100 GPU上进行.本文使用Adamw[21]优化算法,初始学习率设置为3.0×10^{-5},权重衰减系数设置为1.0×10^{-2},设置最大训练周期$N_{\mathrm{epoch}}=100$.Focal loss中的损失权重$\lambda=2$.

3　实验结果分析

在不同的数据集上与对比方案进行比较,评估模型的定位性能.分析FMG-L模型跨数据集的泛化能力和面对多种图像退化的鲁棒性.另外,在经过真实在线社交平台上传和下载过程的MF数据集上验证FMG-L的鲁棒性.

3.1　数据集内实验

为了说明本文提出模型的伪造定位的有效性,将提出的FMG-F,FMG-L模型与对比方案分别在4个数据集中进行训练和测试,实验结果如表1所示.可见,本文提出的FMG-L模型在4个数据集中都取得了最高的定位性能,而FMG-F取得了仅次于FMG-L的定位性能,说明本文提出的FMG模块能够有效聚焦面部区域,减少无关背景的干扰.与对比方案相比,FMG-L在多人场景数据集OF,MF和FW上分别提升了2.1%平均Acc,2.1%平均IoU和0.7%平均F_1;在基于扩散模型生成的单人脸数据集DS上分别提升了9.2%平均Acc,11.1%平均IoU和0.7%平均F_1,说明本文提出的FMG-L模型不仅能够定位包含复杂背景的多人图像,对高逼真单人脸图像依然有效.由于DS数据集中都是单人脸图像,仅包含简单背景信息,无法体现出FMG模块聚焦小区域的优越性,因此在DS中FMG-L的定位F_1略逊于FMG-F.

为了说明FMG-L模型在多人场景图像中伪造定位的准确性,分别选取OF,MF,FW中的图像进行可视化实验,本文提出的FMG-L,FMG-F和对比方案的伪造区域定位结果如图4所示.图4中,FMG-L取得了最优的伪造区域定位结果,而FMG-F的噪声通道中缺少了FMG模块的监督,学习到大量干扰的背景信息,在判断小范围人脸真假时的准确性下降.另外,虽然在表1的数据结果上本文提出的模型与对比方案相差不大,但在图4中可以明显看出对比方案将无关背景和

真实人脸判定为假,甚至无法定位出伪造区域.这是由于多人场景图像中伪造人脸像素仅占一小部分,即使将整张图像判定为真,依然可以达到较高的评估指标.

表1 数据集内伪造定位性能

Table 1 Forgery localization performance within the datasets %

方法	OF			MF			FW			DS		
	Acc	IoU	F_1	Acc	IoU	F_1	Acc	IoU	F_1	Acc	IoU	F_1
GradCAM	91.0	90.7	95.0	97.9	97.9	98.9	92.5	92.3	96.0	82.3	76.9	86.8
Patch	91.7	91.5	95.4	97.9	97.9	98.9	92.9	92.9	96.2	81.8	76.4	86.4
HiFi-Net	91.6	91.6	95.2	97.3	97.2	98.8	90.5	90.4	95.0	67.8	67.6	78.8
Attention	91.0	91.0	95.0	98.0	98.0	99.0	93.0	92.9	95.6	82.2	76.8	86.7
FMG-F	94.6	94.5	95.3	98.2	98.2	**99.8**	94.2	94.1	96.6	86.4	84.4	85.8
FMG-L	**94.8**	**94.7**	**95.4**	**98.3**	**98.3**	**99.8**	**94.5**	**94.4**	**96.6**	**87.7**	**85.5**	**85.4**

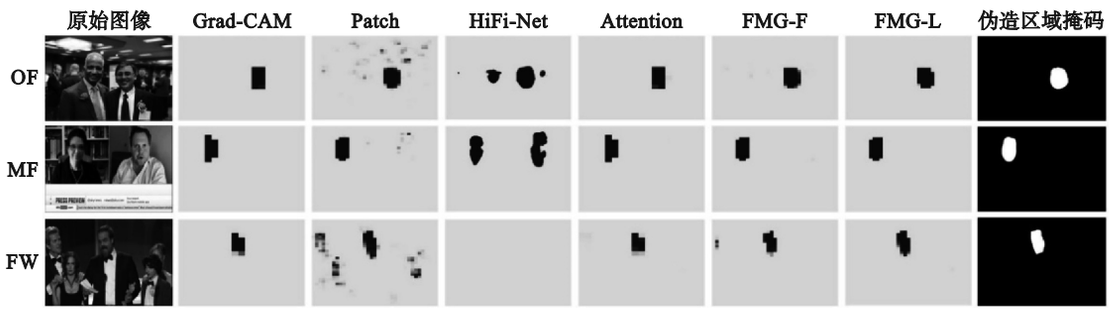

图4 伪造定位结果可视化

Fig. 4 Visualization of forgery localization results

3.2 跨数据集实验

面向未知伪造方法的泛化能力是深度伪造检测方法在现实中部署的重要指标.为有效评估本文提出的定位模型的泛化能力,将FMG-F,FMG-L与对比方案分别在一个数据集中训练后得到的权重在其他3个数据集中进行测试,测试结果如表2所示.

表2中,FMG-L跨数据集的定位性能最优,FMG-F仅次于FMG-L.在OF数据集中训练,并在其他3个数据集中测试时,FMG-L与对比方案相比,分别提升了3.9%平均Acc和3.4%平均IoU;在MF数据集中训练,并在其他3个数据集中测试时,分别提升了3.6%平均Acc和3.2%平均IoU;在FW数据集中训练,并在其他3个数据集中测试时,分别提升了2.0%平均Acc和1.7%平均IoU;在DS数据集中训练,并在其他3个数据集中测试时,分别提升了3.8%平均Acc和3.7%平均IoU.说明本文提出的FMG-L模型具有优越的跨数据集泛化能力,这是由于FMG-L提取了丰富的多层次特征,可以帮助模型更全面地区分真实和伪造区域,能够更好地适应不同类型的伪造技术.

3.3 鲁棒性实验

1)面对图像退化的鲁棒性.为了评估模型面对不同程度图像退化操作的性能,根据先前的研究[22],选取OF数据集,将测试集的图像分别进行以下处理:①JPEG压缩,压缩质量因子分别设置为2,3,4,5和6;②高斯噪声,标准差分别设置为0.001,0.002,0.005,0.01和0.05;③高斯模糊,核大小分别设置为7,9,13,17和21;④块扰动,扰动块的数量分别设置为16,32,48,64和80;⑤颜色饱和度,饱和率分别设置为0.4,0.3,0.2,0.1和0;⑥颜色对比度,对比率分别设置为0.85,0.725,0.6,0.475和0.35.不同程度图像退化操作对伪造定位Acc的测试结果如图5所示.

另外,模型在每种图像退化操作上伪造定位的平均Acc和平均IoU结果如表3所示.根据实验结果可得,本文提出的FMG-L取得了最优的面对图像退化的伪造定位性能,在6种不同程度图像退化操作中分别取得了94.0%平均定位Acc和93.9%平均定位IoU,而FMG-F的表现仅次于FMG-L.

2)面对真实社交平台的鲁棒性.为了评估模型面对真实在线社交媒体网络质量下降的鲁棒性,本文选取在Facebook,TikTok,WeChat,WhatsApp

和YouTube经过上传和下载过程的MF数据集进行实验,实验结果如表4所示.可见,FMG-L模型取得了最优的面对真实社交媒体的伪造定位性能,在经过6种不同社交媒体上传和下载处理的MF数据集中分别取得了97.6%平均定位Acc和97.6%平均定位IoU,FMG-F仅次于FMG-L.

表 2　跨数据集伪造定位结果

Table 2　Forgery localization results across the datasets　　%

方法	训练	MF		FW		DS		训练	OF		FW		DS	
		Acc	IoU	Acc	IoU	Acc	IoU		Acc	IoU	Acc	IoU	Acc	IoU
GradCAM	OF	91.2	91.2	87.5	87.5	68.2	68.0	MF	87.4	87.2	90.2	90.1	73.2	71.6
Patch		92.3	92.2	88.6	89.6	68.5	68.2		89.7	89.5	90.4	90.3	74.2	71.0
HiFi-Net		93.2	93.1	90.7	90.6	67.9	67.9		90.9	90.7	89.4	89.3	70.7	69.6
Attention		92.4	92.5	89.6	89.6	68.4	68.2		88.3	88.2	91.7	91.7	72.5	71.0
FMG-F		95.3	95.3	92.6	92.6	70.5	69.2		92.9	92.7	94.3	94.2	74.9	72.0
FMG-L		**95.5**	**95.5**	**92.7**	**92.8**	**73.2**	**71.4**		**93.0**	**92.8**	**94.4**	**94.3**	**75.6**	**72.6**

方法	训练	OF		MF		DS		训练	OF		MF		FW	
		Acc	IoU	Acc	IoU	Acc	IoU		Acc	IoU	Acc	IoU	Acc	IoU
GradCAM	FW	88.4	88.2	90.4	90.3	71.1	69.5	DS	88.3	88.8	89.7	89.6	89.4	89.3
Patch		90.9	90.7	92.0	93.0	71.4	69.6		84.0	83.3	91.9	91.9	90.0	90.9
HiFi-Net		89.2	89.1	91.8	91.8	67.9	67.9		88.2	88.2	90.9	90.8	88.8	88.8
Attention		90.2	90.2	92.7	92.7	70.7	70.0		89.2	88.7	91.8	91.8	91.1	91.9
FMG-F		91.1	91.0	94.2	94.2	71.1	69.9		90.9	90.8	**94.4**	94.3	94.2	94.1
FMG-L		**91.2**	**91.1**	**95.0**	**94.7**	**71.5**	**70.1**		**91.1**	**90.9**	**94.4**	**94.5**	**94.3**	**94.2**

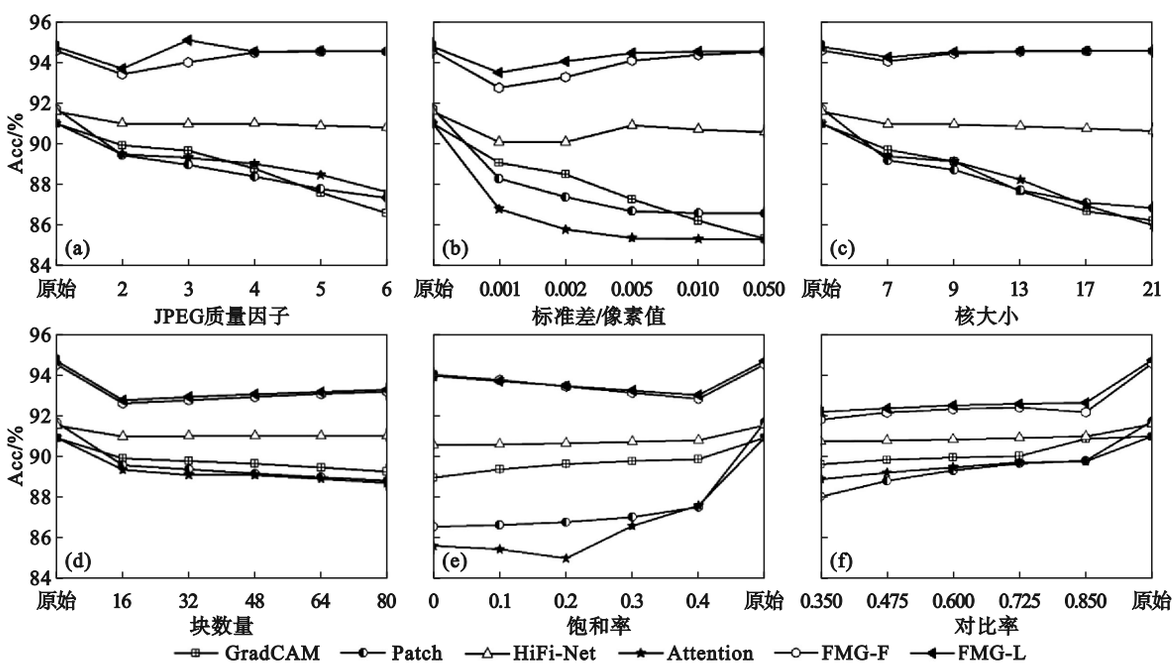

图5　不同程度图像退化对伪造定位Acc的影响

Fig. 5　Impact of various degrees of image degradations on the Acc of forgery localization

(a)—JPEG压缩对Acc影响;(b)—高斯噪声对Acc影响;(c)—高斯模糊对Acc影响;(d)—块扰动对Acc影响;(e)—颜色饱和度对Acc影响;(f)—颜色对比度对Acc影响.

综上,本文提出的FMG-L模型具有强鲁棒性,这是由于该模型结合了多种维度的特征,能够从不同角度对输入的多人场景图像进行分析,减少对单一特征的依赖,能够有效应对因图像质量下降而导致的特定特征信息丢失问题,确保模型在复杂的现实环境中依然保持良好的伪造定位性能.

3.4　消融实验

1) Block数量.三通道网络架构中Block数量

的选择至关重要,为了得到合适的层数,分别设计Block数量为1,2,3和4的网络在OF数据集上进行实验,结果如表5所示.可以看出,增加网络深度能够提升网络性能,但当网络达到一定的深度后,继续增加网络深度会导致伪造定位性能下降,在采用2个Block层时取得了最优的伪造定位性能.

表3 面对不同图像退化的伪造定位性能

Table 3 Forgery localization performance for different image degradations %

方法	原始图像		JPEG		GN		GB		BW		CS		CC	
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU
GradCAM	91.0	90.7	88.9	88.6	87.9	87.8	88.4	88.3	89.9	89.8	89.8	89.7	90.2	89.9
Patch	91.7	91.5	88.9	88.8	87.8	87.8	88.5	88.4	89.6	89.5	87.7	87.7	89.5	89.4
HiFi-Net	91.6	91.6	91.0	91.0	90.6	90.9	90.9	90.8	91.1	91.1	90.8	90.8	91.0	91.6
Attention	91.0	91.0	89.1	89.0	86.6	86.5	88.4	88.4	89.4	89.3	86.9	87.1	89.6	89.6
FMG-F	94.6	94.5	94.3	94.2	93.9	93.8	94.4	94.4	93.2	93.1	93.6	93.6	92.6	92.4
FMG-L	**94.8**	**94.7**	**94.5**	**94.4**	**94.3**	**94.2**	**94.5**	**94.5**	**93.4**	**93.3**	**93.7**	**93.7**	**92.8**	**92.7**

表4 面对不同在线社交平台的伪造定位性能

Table 4 Forgery localization performance for different online social platforms %

方法	原始图像		Facebook		TikTok		WeChat		WeChat(PC)		WhatsApp		YouTube	
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU
GradCAM	97.9	97.9	94.5	94.5	95.6	95.5	95.0	94.9	96.3	96.3	95.1	95.0	96.0	95.9
Patch	97.9	97.9	94.5	94.4	95.5	95.5	95.0	94.9	96.2	96.2	95.0	95.0	96.0	95.9
HiFi-Net	97.3	97.2	96.8	96.7	96.4	96.3	96.3	96.2	96.5	96.5	96.4	96.3	96.5	96.5
Attention	98.0	98.0	94.7	94.7	95.7	95.7	95.1	95.1	96.5	96.4	95.2	95.2	96.1	96.1
FMG-F	98.2	98.2	97.3	97.2	97.2	97.1	97.1	97.0	97.0	97.5	97.1	97.0	97.2	97.2
FMG-L	**98.3**	**98.3**	**97.8**	**97.7**	**97.3**	**97.2**	**97.4**	**97.3**	**97.5**	**98.0**	**97.4**	**97.4**	**97.4**	**97.3**

表5 不同Block数量在OF数据集上的伪造定位性能

Table 5 Forgery localization performance with different number of Blocks on OF dataset %

Block	Acc	IoU
1	93.3	93.2
2	**94.8**	**94.7**
3	93.4	93.3
4	92.2	92.0

2)模块作用.分别针对网络的3个通道C_1,C_2,C_3以及特征融合模块中的DAN结构,在不修改其他网络结构的前提下,分别去除其中一个模块,在OF数据集上进行实验,实验结果如表6所示.可以看出,所有模块都使用的情况下,模型定位性能最好,证明使用三通道网络结构结合基于DAN的特征融合模块在面向多人场景的伪造定位任务中更有利.

3)损失函数.为了验证本文选择的损失函数的作用,分别针对L_{bce},L_{focal}和L_{dice}的单独使用和组合使用在OF数据集上进行消融实验.实验结果如表7所示.可以看出,当仅使用L_{bce}时,模型的伪造定位性能最低,这是由于L_{bce}平等地计算每个像素的损失,在面对多人场景的人脸伪造图像时,损失的预测会偏向真实的像素.$L_{focal}+L_{dice}$的表现最优,证明了本文在正负样本不均衡的情况下结合L_{focal}和L_{dice}的正确性.

表6 不同模块在OF数据集上的伪造定位性能

Table 6 Forgery localization performance with different modules on OF dataset %

算子	Acc	IoU
C_2+C_3+DAN	92.4	92.5
C_1+C_3+DAN	89.7	89.6
C_1+C_2+DAN	91.2	91.0
C_1+C_2+C_3	93.9	93.8
C_1+C_2+C_3+DAN	**94.8**	**94.7**

表7 不同损失函数在OF数据集上的伪造定位性能

Table 7 Forgery localization performance with different loss functions on OF dataset %

算子	Acc	IoU
L_{bce}	92.8	92.6
L_{focal}	94.3	94.2
$L_{bce}+L_{dice}$	93.7	93.5
$L_{focal}+L_{dice}$	**94.8**	**94.7**

4 结　　语

为了解决现有的深度伪造定位算法在面对小区域人脸篡改的多人场景图像时定位性能下降、鲁棒性不足的问题,本研究设计了FMG模块引导模型关注重要的面部区域特征,并且设计了三通道的特征提取网络和特征融合模块.结合RGB特征、FMG特征和噪声特征的多种维度特征,减少模型对单一特征的依赖,增强模型学习到的伪造线索,提升模型对图像退化操作和真实在线社交平台处理的鲁棒性.

参考文献:

[1] Wang T C, Liu M Y, Zhu J Y, et al. High-resolution image synthesis and semantic manipulation with conditional gans[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 8798-8807.

[2] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 8780-8794.

[3] Mirsky Y, Lee W K. The creation and detection of deepfakes: a survey[J]. *ACM Computing Surveys* (*CSUR*), 2021, 54(1): 1-41.

[4] Le T N, Nguyen H H, Yamagishi J, et al. Openforensics: large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild[C]// IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 10117-10127.

[5] Agarwal A, Ratha N. Deepfake Catcher: can a simple fusion be effective and outperform complex DNNs?[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 3791-3801.

[6] He K M, Gkioxari G, Dollár P, et al. Mask r-CNN[C]// IEEE International Conference on Computer Vision. Venice, 2017: 2961-2969.

[7] Zhou P, Han X T, Morariu V I, et al. Learning rich features for image manipulation detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 1053-1061.

[8] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 3146-3154.

[9] Mao A Q, Mohri M, Zhong Y T. Cross-entropy loss functions: theoretical analysis and applications[C]// The 40th International Conference on Machine Learning. Honolulu, 2023: 23803-23828.

[10] Ross T Y, Dollár G. Focal loss for dense object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 2980-2988.

[11] Kumar A, Guo Y L, Huang X Y, et al. SeaBird: segmentation in bird's view with dice loss improves monocular 3D detection of large objects[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 10269-10280.

[12] Wu H W, Zhou J T, Zhang S L, et al. Exploring spatial-temporal features for deepfake detection and localization[EB/OL].(2022-10-28)[2024-08-13]. https://arxiv.org/abs/2210.15872.

[13] Zhou T F, Wang W G, Liang Z Y, et al. Face forensics in the wild[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 5778-5788.

[14] Chen Z X, Sun K, Zhou Z Y, et al. DiffusionFace: towards a comprehensive dataset for diffusion-based face forgery analysis[EB/OL].(2024-03-27)[2024-08-13]. https://arxiv.org/abs/2403.18471.pdf.

[15] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[EB/OL]. (2024-03-27)[2024-08-13]. https://arxiv.org/abs/2403.18471.

[16] Chollet F. Xception: deep learning with depth wise separable convolutions[C]// IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 1251-1258.

[17] Chai L, Bau D, Lim S N, et al. What makes fake images detectable? understanding properties that generalize[C]// European Conference on Computer Vision. Glasgow, 2020: 103-120.

[18] Guo X, Liu X H, Ren Z Y, et al. Hierarchical fine-grained image forgery detection and localization[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 3155-3165.

[19] Țânțaru D C, Oneață E, Oneață D. Weakly-supervised deepfake localization in diffusion-generated images[C]// IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, 2024: 6258-6268.

[20] Huang Y H, Xu J F, Wang R, et al. Fakelocator: robust localization of GAN-based face manipulations[J]. *IEEE Transactions on Information Forensics and Security*, 2022, 17: 2657-2672.

[21] Loshchilov I. Decoupled weight decay regularization[EB/OL].(2017-11-14)[2024-08-13]. https://arxiv. org/abs/1711.05101.pdf.

[22] Jiang L M, Li R, Wu W, et al. Deeperforensics-1.0: a large-scale dataset for real-world face forgery detection[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 2889-2898.