文章编号:1672-3961(2024)05-0111-11 DOI:10.6040/j.issn.1672-3961.0.2023.260

基于自适应掩码和生成式修复的图像隐私保护技术

方世超[1],滕旭阳[1*],王子南[2],陈晗[1],仇兆炀[1],毕美华[1]
(1.杭州电子科技大学通信工程学院, 浙江 杭州 310010; 2.黑龙江昊源科技有限公司, 黑龙江 哈尔滨 150000)

摘要:针对现有图像保护技术中全图加密增加计算成本和区域遮挡无法判定多目标等问题,提出基于自适应掩码和生成式修复的图像保护框架。该框架采用 Score-CAM(class activation mapping)技术自适应判别图像的核心区域,准确生成多目标核心区域掩膜;采用遮挡方法保护图像隐私来降低计算开销;引入区域感知的 CAM 损失函数,确保修复图像重点区域的一致性。将有遮挡的图像送入修复网络进行训练,对训练好的网络参数进行椭圆加密;在发送阶段将掩码图像和密钥分开发送,接收端通过密钥解密,Shift-Net 网络载入参数对掩码图像进行准确修复。在 ImageNet 数据集中的试验表明,CAM 损失函数的修复模型使得生成图像的结构相似性指标提高了 0.2%、学习感知图像块相似度降低了 0.2%。本研究在接收端自适应对图像重点区域进行掩码,使得识别模型失效进而保护图像隐私。

关键词:自适应掩码;生成式修复;区域感知;类激活映射;图像隐私保护
中图分类号:TP391 **文献标志码**:A
引用格式:方世超,滕旭阳,王子南,等. 基于自适应掩码和生成式修复的图像隐私保护技术[J].山东大学学报(工学版),2024,54(5):111-121.
FANG Shichao, TENG Xuyang, WANG Zinan, et al. Image privacy protection based on adaptive masking and generative restoration [J]. Journal of Shandong University (Engineering Science), 2024, 54(5):111-121.

Image privacy protection based on adaptive masking and generative restoration

FANG Shichao[1], TENG Xuyang[1*], WANG Zinan[2], CHEN Han[1], QIU Zhaoyang[1], BI Meihua[1]
(1. School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310010, Zhejiang, China;
2. Heilongjiang Jiongyuan Technology Co., Ltd., Harbin 150000, Heilongjiang, China)

Abstract: In order to solve the problems of increasing the computational cost of full-image encryption and the inability of region occlusion to determine multiple targets in the existing image protection technology, an image protection framework based on the adaptive mask and generative inpainting was proposed. The framework used Score-CAM (class activation mapping) to adaptively discriminate the core region of the image and accurately generate the multi-target core region mask. The occlusion method was used to protect image privacy to reduce computational overhead. The region-aware CAM loss function was introduced to ensure the consistency of the key areas of the repaired image. The occluded images were sent to the repair network for training, and the trained network parameters were elliptically encrypted. The mask image and the key were sent separately at the sending stage, and decrypted by the key at the receiving end. Then parameters in the Shift-Net were loaded to repair the mask image accurately. The experimental results on the ImageNet dataset showed that the restoration model with CAM loss function improved the structural similarity of the generated image by 0.2%, and reduced the learned perceptual image patch similarity by 0.2%. This study adaptively masked the key areas of the image at the receiving end, rendering the recognition model ineffective and thus protecting the privacy of the image.

Keywords: adaptive mask; generative inpainting; region-aware; class activation mapping; image privacy protection

0 引言

随着通信网络的快速发展,图像安全已经成为当下的重要议程。图像加密是将原始的明文图像进行转换,使其变得不可读或无法理解,确保图像内容的机密性和安全性。数字图像加密领域广泛应用混沌理论,文献[1]提出基于压缩感知和混沌

收稿日期:2023-10-30
基金项目:国家自然科学基金资助项目(No.61906055);浙江省自然科学基金资助项目(LQ19F020009)
第一作者简介:方世超(1997—),男,安徽安庆人,硕士研究生,主要研究方向为图像隐私保护、语义分割。E-mail: fsc_hdu@126.com
*通信作者简介:滕旭阳(1987—),男,黑龙江哈尔滨人,副教授,硕士研究生导师,博士,主要研究方向为人工智能、图像处理。
E-mail:tengxuyang@hdu.edu.cn

系统的图像加密算法。该算法不仅具有良好的加密效果,还能有效节约空间,但在保证加解密质量的同时不能兼顾加解密速率。为了更加安全、高效对图像进行加密传输,学者将该算法与深度学习相结合。文献[2]提出了基于深度学习压缩感知与复合混沌系统的通用图像加密方法,但该方法只能对单张图像加密。为了提高加密系统的整体加密效率,文献[3]提出了基于压缩感知和深度学习的多图像加密方法,该方法可以同时对多张图像进行压缩和加密处理。为了进一步提高解密图像的质量,文献[4]提出了基于生成对抗网络(generative adversarial network, GAN)和卷积神经网络(convolutional neural network, CNN)的鲁棒压缩感知图像加密算法,通过GAN得到解密后的图像,再使用CNN降噪,改善了解密后图像的视觉表达。图像加密不同于文本加密,图像中包含的信息量极大,尤其是在当今信息化时代,图像是信息呈现的主要形态,对海量图像进行加密操作,花费成本巨大,效率低,不能满足快速扩展图像数据的需求。为了降低计算成本,文献[5]提出了一种基于掩膜的区域遮挡方法,对敏感区域的图像信息进行遮挡。特别地,遮挡操作不需要在使用的云平台上消耗大量计算资源,不能同时在多个相同目标的图像上进行精准遮挡。

图像解密是指接收方使用正确的解密算法和密钥,对密文图像进行解密操作。密文图像中的每个像素经过解密算法处理,得到对应明文像素。从结果来看,图像修复与图像解密有异曲同工之妙,都是对图像进行恢复。其中,上下文编码器是第一个基于生成对抗网络GANs的修复网络[6]。文献[7]提出的架构可以实现对局部语义和全局语义的理解;文献[8]提出一种将上下文编码器与纹理补丁相结合的多尺度神经网络补丁生成模型;文献[9]提出一种基于生成对抗网络的遮挡图像修复算法,能够在大量像素缺失的场景下复原出图像的本来面目;文献[10]提出Shift-Net的快速连接网络,对比文献[8]其结构相似性指标提升了0.1%,峰值信噪比提升了0.53 dB;文献[11]提出一个包含部分卷积层的网络结构;文献[12]提出可学习的双向注意映射模块,以端到端的方式学习特征并对掩码进行动态更新;文献[13]提出基于相干语义注意力机制层(coherent semantic attention, CSA)的网络结构,对比文献[10],其结构相似性指标提升了0.2%,峰值信噪比提升了0.16 dB。现有的图像修复模型,掩码可分为固定区域掩码和随机区域掩码,两者都不能有效对图像的敏感区域进行掩码。

大数据时代,人工处理海量图像数据的能力十分有限,大多数依靠人工智能。受到图像修复技术将图像进行局部掩码后进行修复的启发,学者们开始采用深度修复模型对掩码图像进行生成式修复来保护图像隐私,仅对图像重要区域进行掩码遮挡,无需进行加密操作。例如在网络的输入端使用深度学习方法对图像敏感区域进行掩码,对该区域的图像信息进行了隐藏处理,防止第三方非法利用深度识别模型获取图像信息。在网络的输出端对掩码图像进行修复,使深度学习模型能够有效识别图像。本研究提出了一种基于自适应掩码和生成式修复的图像隐私保护技术,能够有效遮挡图像中多个相同目标的敏感区域,在CSA和Shift-Net修复模型的基础上,提出了基于区域感知的CAM(class activation mapping)损失函数。试验结果表明,被遮挡的图像使得智能识别模型失效,能够有效识别修复后的图像,提出的损失函数能够指导网络生成更高质量的图像。

1 背景知识

1.1 类别激活映射图

Score-CAM即类别激活映射图[14],可以理解为对预测输出的贡献分布:分数越高的地方表示原始图像对应区域对网络的响应越高、贡献越大。给定一个CNN模型$Y=f(X)$,Score-CAM接收一个图像输入X并输出一个类别标量Y。在f中选取一个内部卷积层l,将相应的激活作为A。A_l^k表示A_l的第k个通道的特征图。对于已知的基线输入X_b、A_l^k对Y的贡献度定义为:

$$C(A_l^k)=f(X_b\circ H_l^k)-f(X_b), \tag{1}$$

式中,$\circ$表示哈达马积,$H_l^k=s(Up(A_l^k))$,$Up(A_l^k)$代表将A_l^k上采样到X_b一样大小,$s(\ )$表示归一化操作。Score-CAM的输出定义为$L_{\text{Score-CAM}}^c=\max(\sum_k \alpha_k^c A_l^k)$,其中,$c$表示目标类别标签,$\alpha_k^c=C(A_l^k)$。$L_{\text{Score-CAM}}^c$是一张像素0到1的灰度图,尺寸与原图大小相等。

1.2 图像修复网络

图像修复技术有着悠久的历史,它的工作原理是依据图像已知内容信息推理出缺失区域的未知内容。该技术早期应用于艺术画作的修复,很大程度依赖于人的经验和手法,十分耗时。伴随着计算机的持续发展,数字图像修复已逐渐成为计算机视觉和计算机图形学领域的一项重要研究内容[15]。图像修复技术的发展经历了传统到现代的转变。传统的图像修复方法主要依据图像中已知的结构信息和纹理信息,结合各种算法推断出缺失区域的

内容信息[16]。当时计算机的算力有限,这类方法不能捕捉到图像中的高级语义信息。随着大数据时代的到来,深度学习理念再次进入大众视野。CNN 的出现使得计算机能够更好地学习图像特征,GAN 的出现让模型学习概率分布成为可能。基于以上几点,图像修复技术得到了跨越式的发展。

1.2.1　Shift-Net

Shift-Net 是一种 GAN 网络,生成网络以 U-Net 为主干网络。Shift-Net 将基于示例和基于 CNN 的图像修复方法相结合,设计了一个移位连接层,在解码器的 $L-l$ 层添加移位连接层,该层在编码器的 l 层已知区域寻找与解码器的 $L-l$ 层特征图缺失区域最相似的补丁。判别器是一个 5 层的卷积网络。网络的整体损失包括引导损失、重构损失和对抗损失。

1.2.2　相干语义注意力机制修复网络

CSA 修复网络设计了一个连贯语义注意层,与 Shift-Net 不同的是,CSA 包括两个部分:粗修复网络和细修复网络。连贯语义注意层只存在于细修复网络解码端,在寻找补丁时,该层考虑了缺失区域与已知区域的相似关系,考虑了与缺失区域上一次生成的补丁的相似关系,在细修复网络中引入了特征块判别器和一致性损失。

2　提出的图像隐私保护框架

本研究提出了一种基于自适应掩码和生成式修复的图像隐私保护技术,整体框架如图 1 所示。在发送端对图像的敏感区域进行自适应掩码来遮挡图像的主要信息,即使非法截获端获取了图像,识别网络也无法学习到图像的主要特征。接收端将训练好的参数载入图像修复模型,模型对图像的掩码区域进行生成式修复,能够还原较多的原始信息,修复后的图像能够被识别网络有效识别。利用 Score-CAM 自适应地判别图像的重点关注区域,能对多目标进行区域判断,对图像敏感区域进行自适应掩码;其次将训练好的模型参数进行基于佩尔数列和椭圆曲线的文本加密[20],发送端将密钥发送到接收端,对参数进行解密之后,接收端的修复模型成功载入参数,对掩码图像进行修复,整个过程能以端到端的方式运行。在 ImageNet 和 COCO 数据集上,修复模型针对训练过的图像修复效果较好,针对并未训练过的图像,修复效果很差。即便密钥泄露,截获端并不清楚修复模型的网络结构,依旧无法截取图像信息;反之如果第三方了解修复模型的网络结构,却无法获取模型参数,则不能通过该修复模型获取图像信息,除非能够将掩码图像,参数密钥,修复模型三者全部截获。通过将模型训练参数加密,避免了对海量图像数据的加密操作,降低了计算成本,提高了加解密速率。本研究引入了基于区域感知的 CAM 损失函数来指导修复网络生成质量更高的图像,提出的框架能简便、高效保护图像信息。

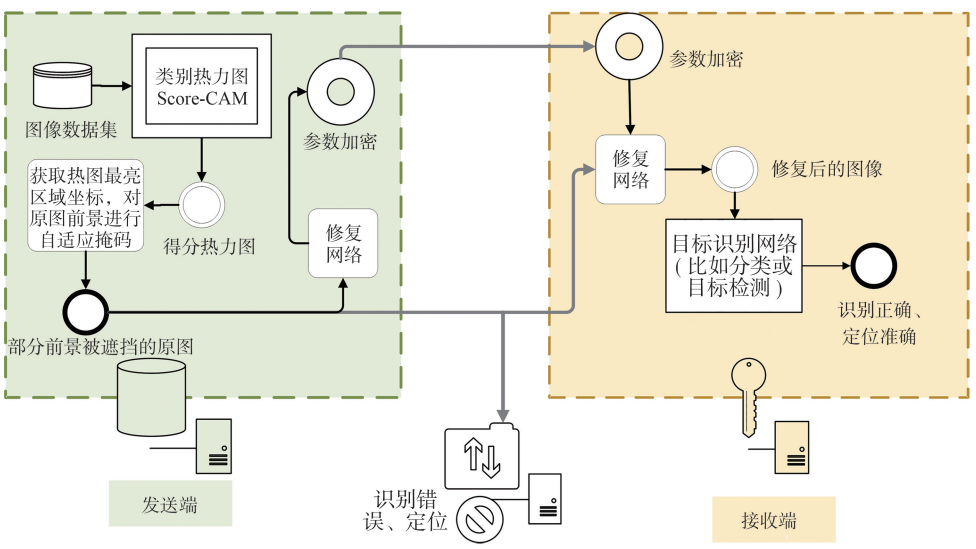


图 1　图像隐私保护的整体框架

Fig.1　The framework for image privacy protection

2.1　基于 Score-CAM 的自适应掩码技术

采用 ResNet18 作为 Score-CAM 的骨干网络,输入原始图像[18],输出 CAM 和类别信息。可视化的时候如图 2(c),利用热力图和原图叠加的形式呈现。颜色越深红的地方表示值越大。可以认为,网络预测“熊猫”这个类别时,高亮区域是其主要判断

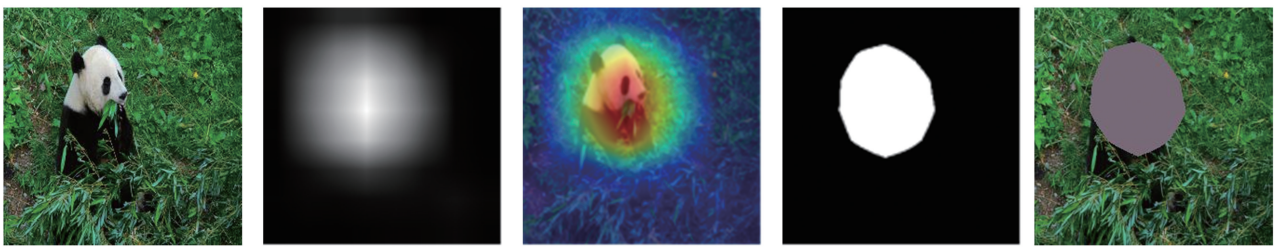
依据,在该区域进行掩码操作,使得识别网络无法辨别目标。通过大量试验,观察到类别激活图高亮区域边缘像素值约为 0.5。结合 Score-CAM 输出的 CAM 灰度图,通过像素置换法,生成对应的掩膜,对原图进行自适应掩码。详细步骤如下。

步骤 1:设立一个阈值 γ ($0.5<\gamma<1$),本文中定义 γ 为 CAM threshold,将其设置为 0.5。类别激活图中像素值大于或等于 CAM threshold 可视作图像的敏感区域。

步骤 2:构造一张与热力图同样大小的全 0 掩膜,接着搜寻热力图像素值大于 CAM threshold 的像素点,获取其位置坐标,将其像素值置 0,将掩膜中处于相同位置的像素值置 1。

步骤 3:一直重复步骤 2,直到热力图中的像素值都小于 CAM threshold 为止,获得一张二值图像,每个像素取值为 0 或 1,1 对应原图中的重要像素,0 反之。

步骤 4:用一个 1 到 255 之间的正整数填充原图中与二值图像像素为 1 的位置相对应的像素点,得到对应的掩码图像。



(a) 原图 (b) 类别激活图 (c) 热力图 (d) 二值掩膜 (e) 掩码图像

图 2 自适应掩码示例

Fig.2 The illustration of adaptive mask

2.2 基于佩尔数列和椭圆曲线的模型参数加密

基于佩尔数列和椭圆曲线的密码学应用广泛,其中包括信息安全、移动通信和物联网、数字版权保护、电子支付和金融、身份认证和访问控制等。其强大的加密能力和较小的密钥尺寸使其成为当今安全领域的重要技术之一。

2.2.1 佩尔序列

对于初始值 $P_0=0,P_1=1$,佩尔数列的递归关系定义为 $P_n=2P_{n-1}+P_{n-2}$,当 $i\to\infty$,$P_i/P_{i-1}\to1+\sqrt{2}$。

2.2.2 椭圆曲线

椭圆曲线 $Ep(a,b)$ 方程定义:p 为质数 $x,y\in[0,p-1]$,有 $y^2=x^3+ax+b(\mathrm{mod}\ p)$,$4a^3+27b^2\neq0$。当 $a=0,p\equiv2(\mathrm{mod}\ 3)$,$Ep(a,b)$ 正好有 $p+1$ 个点,其 y 坐标不重复。

2.2.3 文本加密

文献[17]提出基于佩尔数列和椭圆曲线的文本加密,该加密方法通过对符号集使用循环移位来扩散明文,得到无意义的明文,使用佩尔数列和二进制序列将扩散文本的每个元素编码为实数,达到隐藏扩散文本元素的目的,通过在椭圆曲线上生成排列来混淆编码的扩散文本。其中,定义 S 是大小为 m 的有限符号集,$i\in[0,m-1]$,令 $S(i)$ 表示 S 的第 i 项,T 为待加密的参数,$T=T(1)\cdots T(i)\cdots T(n)$,$i\in[1,n]$,其长度为 n,是 S 上的序列,$T(i)$ 表示 T 的第 i 项。在该方案中,将参数编码设置为区间[-1,1]内的实数,小数点后 $\beta\geqslant14$ 位,详细步骤如下。

步骤 1:选择一个整数 k,定义循环位移 $\psi_k(S(i))=S((i+k)(\mathrm{mod}\ m))$,使 $T'(i)=\psi_k(T(i))$,$i\in[1,n]$,从而获得扩散文本 $T'=T'(1)\cdots T'(i)\cdots T'(n)$。

步骤 2:定义一个受限佩尔数列 $Q_{h,h'}=q_1\cdots q_i\cdots q_m$,$h$ 和 h' 满足 $h<h'$,$h'-h+1<\beta$;定义一个权重函数 $w:\{1,2\cdots,n\}\to[-1,1]$,对 T' 的每个元素的位置进行唯一编码;定义二进制序列:$\alpha=\alpha_1\cdots\alpha_i\cdots\alpha_n$,决定在编码时是否使用权重 $w(i)$;结合 $Q_{h,h'}$、w 和 α 对扩散文本 T' 进行编码,得到 $(c_i,d_i)=(q_{(j+k)(\mathrm{mod}\ m)}+\alpha_i w(i),1-q_{(j+k)(\mathrm{mod}\ m)}+(1-\alpha_i)w(i))$,则 $(C,D)=(c_1,d_1)\cdots(c_i,d_i)\cdots(c_n,d_n)$ 为编码后的扩散文本。

步骤 3:使用两个椭圆曲线的有序子集,产生两个双射 $\sigma:C\to C$、$\sigma':D\to D$,对文本进行混淆得到 $(\sigma(C),\sigma'(D))=(\sigma(c_1),\sigma'(d_1))\cdots(\sigma(c_i),\sigma'(d_i))\cdots(\sigma(c_n),\sigma'(d_n))$。

该方法可加密任意大小的纯文本,能抵抗密钥攻击、统计攻击等计算攻击。使用该加密方法对修复网络训练好的参数进行加密,单独将密文发送到接收端。

2.3 基于区域感知的 CAM 损失函数

感知损失用于实时超分辨任务和风格迁移任务[19],现在应用于更多领域,比如图像去雾。感知

损失是通过一个固定的网络,分别以真实图像(ground truth)和网络生成结果(prediction)作为其输入,得到对应的输出特征:feature-gt、feature-pr,使用 feature-gt 与 feature-pr 构造损失函数,逼近真实图像与网络生成结果图之间的深层信息,与之类似,CAM 损失函数定义如下:

$$L(I)=\text{Score-CAM}(I), \tag{2}$$

$$L_{\text{CAM}}=\| L(I_r)-L(I_{gt}) \|_1, \tag{3}$$

式中,Score-CAM(I)表示输入图像 I 经过 Score-CAM 最终得到的输出。原图 I_{gt} 和修复网络生成的图像 I_r 依次输入 Score-CAM 中,得到对应的输出 $L(I_{gt})$、$L(I_r)$,将 $L(I_g)$、$L(I_r)$ 构造 L_1 损失。

原 CSA 和 Shift-Net 模型对应的源代码中对于 L_1 构造的损失函数,权衡参数 λ_1 设置为 100,这是网络中输入的图像都会被归一化,对 L_1 求平均,导致 L_1 构成的损失函数本身就特别小,不利于反向传播,CAM 关注的是图像的敏感区域,对此区域进行自适应掩码,每张图像的掩码区域中像素点的个数各不相等,采用自适应的权衡参数,其会随着掩码区域像素点个数的变化而变化,更有利于参数优化。针对修复网络的损失函数中的 CAM 损失部分,制定的两个方案如下:

$$L_{\text{CAM1}}=y_1\| L(I_r)-L(I_{gt}) \|_1, \tag{4}$$

$$L_{\text{CAM2}}=y_2\| L(I_r)-L(I_{gt}) \|_1。 \tag{5}$$

令 $y_1=100/N$,$y_2=\sqrt{n}/N$,分别代表 L_{CAM1}、L_{CAM2} 中采用的权衡参数,其中常量 N 表示类别激活图总的像素点个数,变量 n 表示掩码区域像素点总的个数。如图 3 所示,y_1 不会随着掩码区域像素点数 n 的变化而变化;y_2 会随着掩码区域像素点数 n 的变化而变化,当 n 较小时,CAM 损失所占比例较小,当 n 逐渐增大时,CAM 损失所占比例缓慢增大,可自适应调整权衡参数。

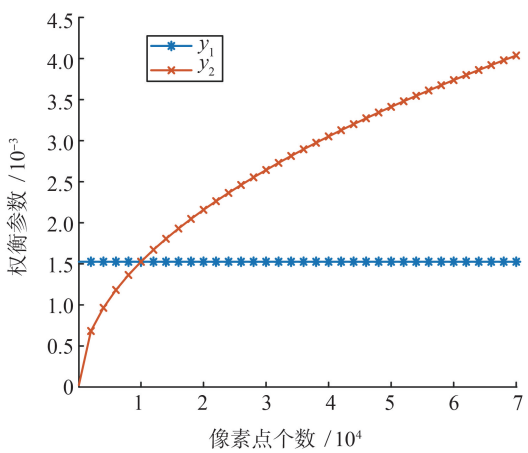


图 3　权衡参数曲线

Fig.3　The curve of balancing parameters

3　试验及结果分析

分别使用 5 000 张 ImageNet 和 COCO 两种数据集作为训练集 D_{tr1}、D_{tr2},100 张作为测试集 D_{te1}、D_{te2}。采用自适应掩码方法对图像进行处理,为证明分类网络不能识别、目标检测网络无法精确定位掩码处理后的图像,采用了 ResNet50[18]、Yolov8[20]、Faster-Rcnn[21] 3 种网络(Yolov8 既可用来分类也可用来检测)来检验。为了探究 CAM 损失的固定权衡参数和自适应权衡参数哪个更利于网络参数优化,对两者生成图像的曼哈顿范数均值(mean L_1)、峰值信噪比(peak signal to noise ratio,PSNR)、结构相似性指标(structural similarity index measure,SSIM)、学习感知图像块相似度[25](learned perceptual image patch similarity,LPIPS)4 个图像评价指标进行了对比[22]。

以上提到的 4 种图像评价指标,mean L_1 是逐像素比较差异,其值越小,表示重构图像与原图的差异越小,不符合人类的视觉感知;PSNR 一般用于衡量最大值信号和背景噪音之间的图像质量,理论上来说,值越高表示图像质量越好;SSIM 考虑了图像的亮度、对比度和结构指标,用来衡量两张图的相似程度,符合人类的视觉感知,取值为[-1,1],值越大,表示相似度越高;LPIPS 用于度量两张图像的差别,值越低表示两张图像越相似,反之,则差异越大,它比 SSIM 更符合人类的视觉感知[25]。

3.1　掩码效果对比分析

从 D_{tr1}、D_{tr2} 训练集中各随机抽取了 500 张图,分别记为 D_{t1}、D_{t2}。将 D_{t1}、D_{t2} 图像调整宽高都为 256×256 的,记为 I_{gt},分别对图像进行自适应掩码和随机区域掩码处理,两种掩码方式的掩码区域面积大小相等,掩码处理后的图像依次记为 I_{ad}、I_{ra}。其中,D_{t1} 用于目标分类,D_{t2} 用于目标检测。

将 D_{t1} 的 I_{gt}、I_{ra}、I_{ad} 分别送入 ResNet50、Yolov8,记录两个网络输出的前五类别概率,统计其中正确类别的概率(置信度)并计算出平均值,依次记为 $P_{\text{C-Igt}}$、$P_{\text{C-Ira}}$、$P_{\text{C-Iad}}$,结果如表 1 所示,能够发现图像经过随机掩码处理之后,不能有效掩盖图像的信息,经过自适应掩码处理的图像,很难被分类网络识别。图 4、5 分别为单目标和多目标的分类结果示例。

表 1　两种分类网络置信度的平均值

Table 1　The average of confidence level for two classification networks

数据集	网络	$P_{\text{C-Igt}}$	$P_{\text{C-Ira}}$	$P_{\text{C-Iad}}$
ImageNet	Yolov8	0.68	0.56	0.06
	ResNet50	0.65	0.51	0.05

(a) 原图单目标分类 (b) 随机掩码单目标分类 (c) 自适应掩码单目标分类

图 4 单目标分类结果示例

Fig.4 Single-objective classification result example illustration

(a) 原图多目标分类 (b) 随机掩码多目标分类 (c) 自适应掩码多目标分类

图 5 多目标分类结果示例

Fig.5 Multi-objective classification result example illustration

将 D_{t2} 的 I_{gt}、I_{ra}、I_{ad} 分别送入 Faster-RCNN、Yolov8,统计网络在敏感区域中目标的识别准确率,其定义为图像敏感区域中识别正确的目标数与敏感区域内总的目标数之比。I_{gt}、I_{ra}、I_{ad}的识别准确率依次记为 $P_{\text{D-Igt}}$、$P_{\text{D-Ira}}$、$P_{\text{D-Iad}}$,结果如表 2 所示,能够发现随机掩码遮挡敏感区域目标的位置信息具有随机性,经过自适应掩码处理的图像,很难被目标检测网络精准定位。图 6、7 分别为单目标和多目标的分类结果示例。

表 2 两种检测网络的识别准确率

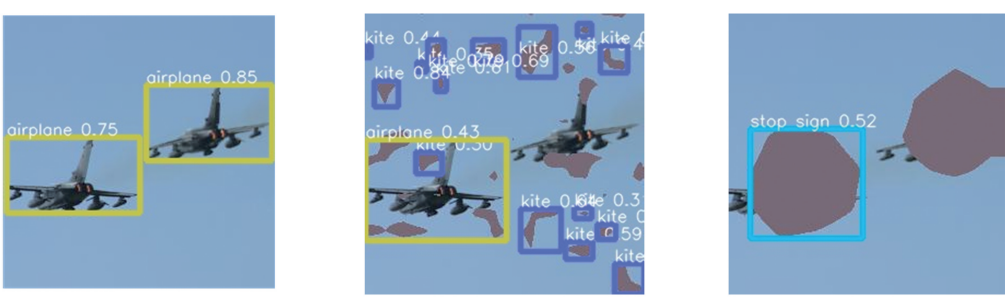
Table 2 The recognition accuracy for two object detection networks

数据集	网络	$P_{\text{D-Igt}}$	$P_{\text{D-Ira}}$	$P_{\text{D-Iad}}$
COCO	Yolov8	0.98	0.70	0.05
	Faster-RCNN	0.92	0.65	0.03

(a) 原图单目标检测 (b) 随机掩码单目标检测 (c) 自适应掩码单目标检测

图 6 单目标检测结果示例

Fig.6 Single-objective detection result illustration



(a) 原图多目标检测 (b) 随机掩码多目标检测 (c) 自适应掩码多目标检测

图 7 多目标检测结果示例

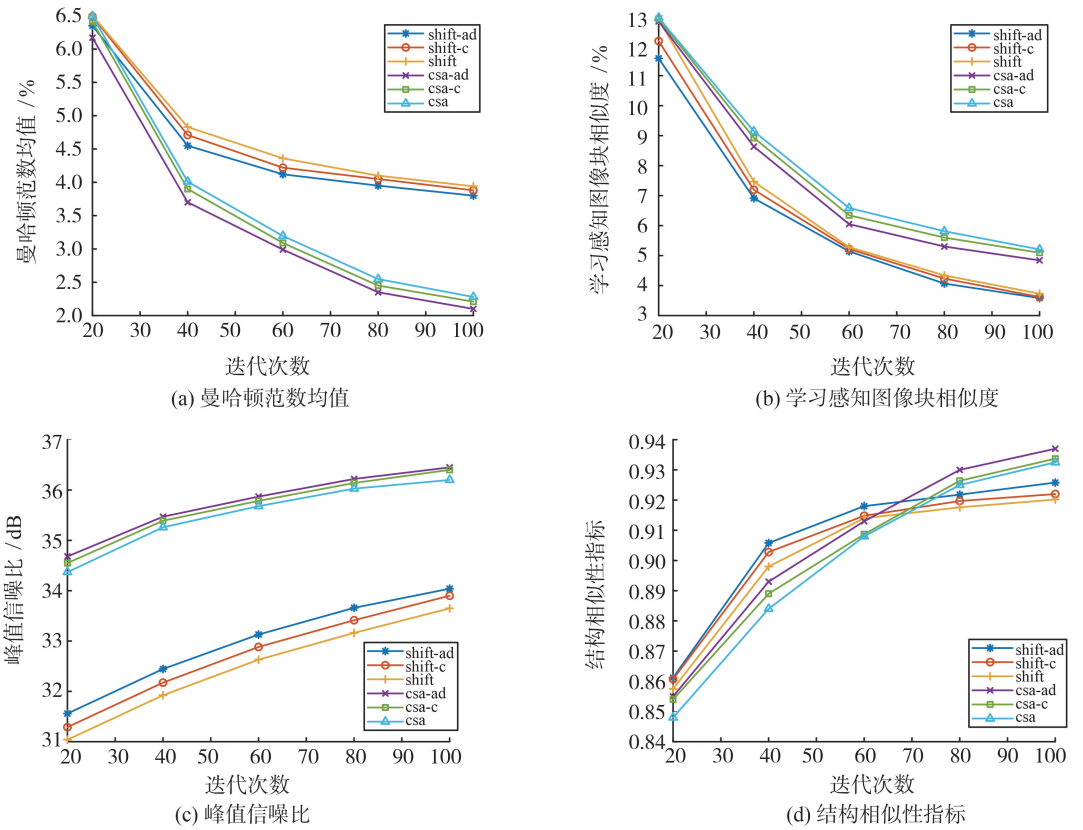
Fig.7 Multi-objective detection result illustration

基于 Score-CAM 的自适应掩码不仅适用于单目标图像,还可用于多目标图像。它可以有效遮挡图像的语义信息,能隐藏图像的位置信息,使得图像无法被学习或者理解。在海量数据的信息时代,手动处理图像的数量有限,自适应掩码能够自动地、高效地对图像的重要区域进行遮挡,可有效避免被第三方的深度模型检测识别。

3.2　模型对比结果及分析

分别从训练集 D_{tr1}、D_{tr2} 随机选取 100 张图像,统计多种模型在不同 epoch 生成图像的 mean L_1、LPIPS、PSNR、SSIM,对比结果如图 8。表 3 是 epoch 为 100 时所统计的各项指标数值。图 9、10 依次是单目标、多目标修复对比示例,表 4、5 分别是图 9、10 对应修复图像的各项指标具体数值。其中“csa”、“shift”代表原有的网络模型,“csa-c”、“shift-c”代表在原有模型的基础上引入 CAM 损失,且其权衡参数为常数,“csa-ad”、“shift-ad”代表 CAM 损失函数的权衡参数是自适应的。

观察图 8,很明显引入 CAM 损失的修复网络的 4 种主要图像评价指标都优于对应的原有模型,且 CAM 损失权衡参数为自适应的模型最优,在掩码过程中,其动态权衡参数会根据掩码像素点数自行调整,更有利于参数优化。CSA 的 Mean L_1、PSNR 明显优于 Shift-Net,但是后者的 LPIPS 明显优于前者,比 SSIM 更符合人类的视觉感知。



(a) 曼哈顿范数均值　(b) 学习感知图像块相似度

(c) 峰值信噪比　(d) 结构相似性指标

图 8　ImageNet 数据集上定性对比结果

Fig.8　The qualitative comparison results on the ImageNet

通过观察表 3 数据,可以发现在引入 CAM 损失且其权重固定时,各项指标有所优化。将 CAM 损失的权重改为自适应时,各项性能又有所上升,mean L_1 和 LPIPS 下降了 0.2% 左右,PSNR 提高了 0.3 dB左右,SSIM 提高了 0.2%。

表 3　ImageNet 数据集上定量对比结果

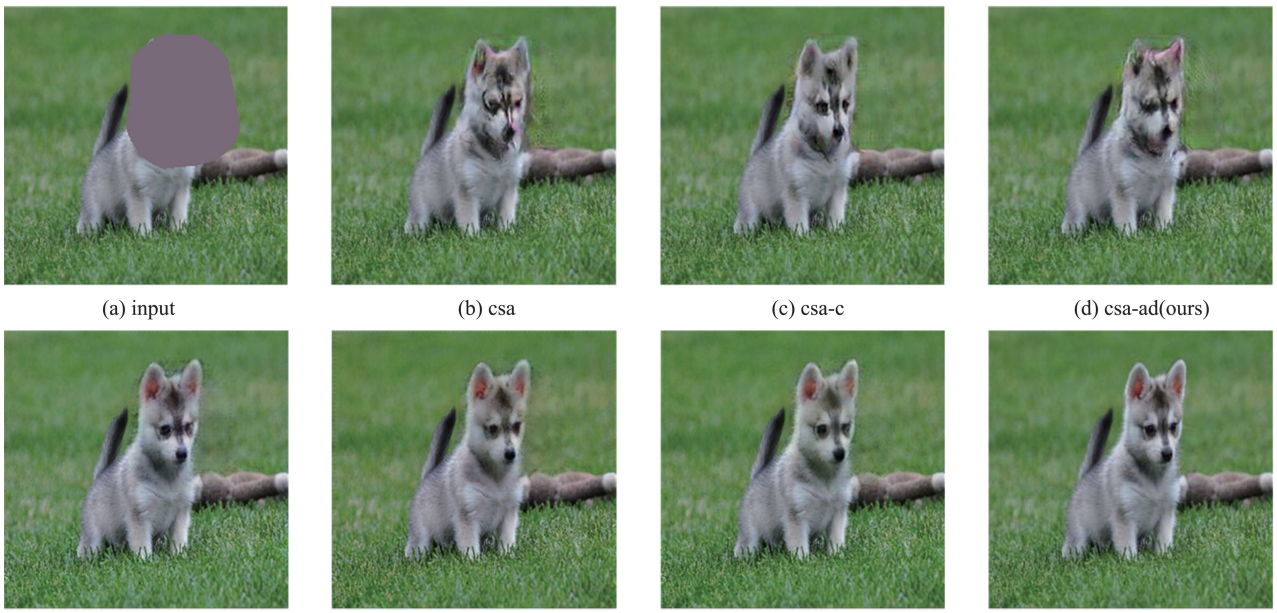
Table 3　The quantitative comparison results on the ImageNet

网络	网络类别	图像质量评价指标			
		mean L_1/%	LPIPS /%	PSNR	SSIM
CSA	csa	2.28	5.207	36.24	0.933
	csa-c	2.21	5.095	36.40	0.933
	csa-ad	**2.10**	4.842	**36.45**	**0.935**

表 3(续)

网络	网络类别	图像质量评价指标			
		mean L_1/%	LPIPS /%	PSNR	SSIM
Shift-Net	shift	3.94	3.731	33.65	0.920
	shift-c	3.88	3.620	33.90	0.921
	shift-ad	3.80	**3.584**	34.04	0.923

注:黑体表示所在列最佳值。

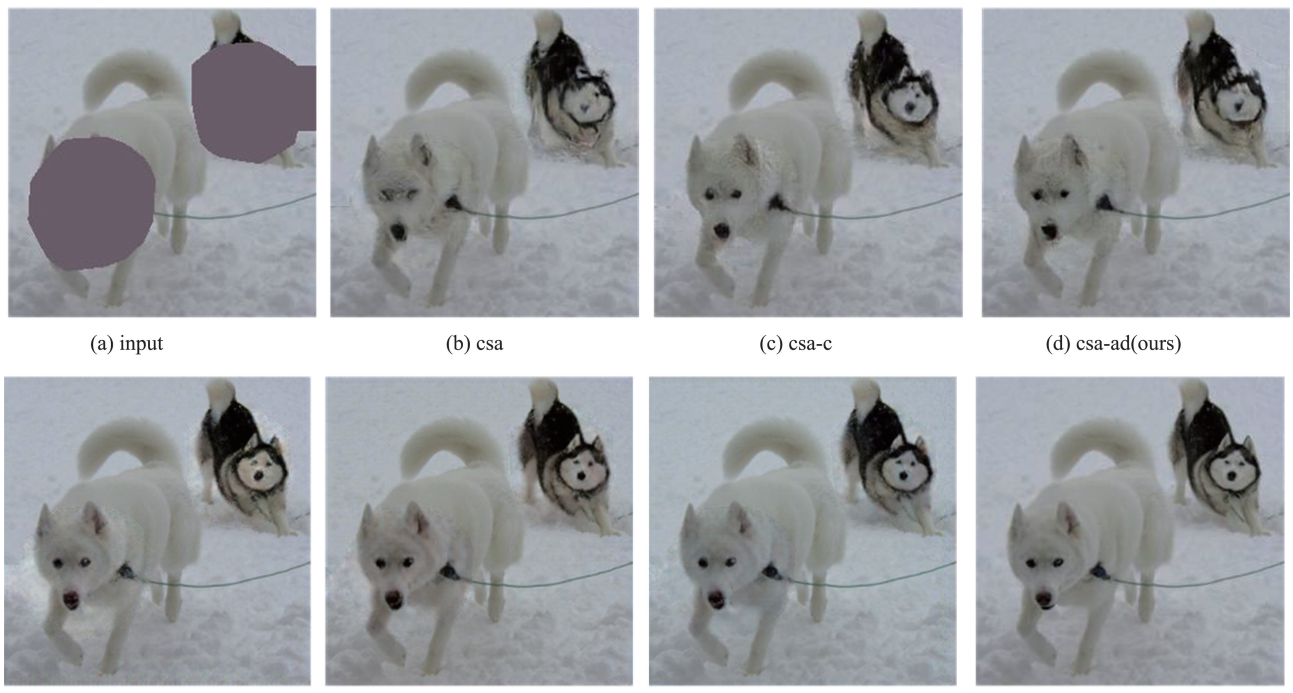


(a) input (b) csa (c) csa-c (d) csa-ad(ours)

(e) shift (f) shift-c (g) shift-ad(ours) (h) GT

图 9 单目标修复对比

Fig.9 The comparison of single-objective restoration



(a) input (b) csa (c) csa-c (d) csa-ad(ours)

(e) shift (f) shift-c (g) shift-ad(ours) (h) GT

图 10 多目标修复对比

Fig.10 The comparison ofmulti-objective restoration

表 4　对应图 9 修复图像的各项指标

Table 4　Corresponding to the specific values of each index of the image restoration in Figure 9

网络类别	图像质量评价指标		
	PSNR	SSIM	LPIPS
csa	37.01	0.933	0.059
csa-c	36.59	0.934	0.056
csa-ad	**37.65**	**0.953**	0.042
shift	32.43	0.934	0.033
shift-c	34.16	0.939	0.032
shift-ad	33.91	0.939	**0.028**

注:黑体表示所在列最佳值。

表 5　对应图 10 修复图像的各项指标

Table 5　Corresponding to the specific values of each index of the image restoration in Figure 10

网络类别	图像质量评价指标		
	PSNR	SSIM	LPIPS
csa	36.50	0.917	0.071
csa-c	36.47	**0.926**	0.068
csa-ad	**36.56**	0.923	0.065
shift	34.47	0.916	0.050
shift-c	34.50	0.919	0.048
shift-ad	34.51	0.924	**0.045**

注:黑体表示所在列最佳值。

从训练集 D_{tr1}、D_{tr2}中各随机选取 100 张图像和测试集 D_{te1}、D_{te2}作对比。将修复图像分别记为 D_1、D_2、D_3、D_4。D_1、D_3分类置信度平均值 P_C和 D_2、D_4识别准确率 P_D如表 6 所示。其中,D_1、D_3的分类置信度平均值分别为 0.55、0.01,D_2、D_4的识别准确率分别为 0.71、0,说明经过训练后修复图像确实能够被深度识别网络精准定位和识别,未经训练的图像,其修复结果很不理想,基本得不到任何有效信息,深度识别网络无法识别。图 11 是训练集原图和修复模型输出图的分类及检测的对比结果示例,图 12 是测试集的图像修复效果示例。

表 6　分类置信度平均值及识别准确率

Table 6　The average of confidence level and recognition accuracy

修复图像	P_C	P_D
D_1	0.55	
D_2		0.71
D_3	0.01	
D_4		0

(a) 输入　(b) 输出　(c) 原图　(d) 原图分类　(e) 输出分类

(f) 输入　(g) 输出　(h) 原图　(i) 原图检测　(j) 输出检测

图 11　目标分类和检测结果对比

Fig.11　The comparison of object classification and detection results

(a) 输入　(b) 输出　(c) 原图

(d) 输入 (e) 输出 (f) 原图

图 12 测试集的图像修复效果

Fig.12 The image restoration effect oftest set

4 结论

本研究提出的基于自适应掩码和生成式修复的图像隐私保护技术是从智能目标识别角度出发,在发送端对图像进行自适应掩码处理,同时训练图像修复模型,且将模型训练好的网络参数加密,分别发送掩码图像、加密参数及密钥。接收端接收数据,将解密后的参数和掩码图像载入修复模型,输出修复图像。智能识别网络无法获取输入端图像的有效信息,却能精准识别和定位经接收端修复模型修复后的图像。为了进一步提升修复质量,在原有模型的基础上引入了基于区域感知的 CAM 损失函数。对于未训练过的图像,修复模型的修复效果很差,即使密钥和掩码图像被截获,截获端并不清楚修复模型的网络结构,无法修复图像。即使第三方了解网络结构且截获了掩码图像,仍然不能对图像进行修复。本研究提出的框架能有效保护图像的隐私,将图像加密转换为文本加密,且整体框架能以端到端的方式运行。未来,将采用知识蒸馏的方法对修复模型进行优化,使之用同类型小样本训练加密模型能做到对同类型未知样本的修复。

参考文献:

[1] BRRAHIM A H, PACHA A A, SAID N H. Image encryption based on compressive sensing and chaos systems [J]. Optics and Laser Technology, 2020, 132: 106489-106499.

[2] 陈炜, 郭媛, 敬世伟. 基于深度学习压缩感知与复合混沌系统的通用图像加密算法[J]. 物理学报, 2020, 69(24):99-111.
CHEN Wei, GUO Yuan, JING Shiwei. General image encryption algorithm based on deep learning compressed sensing and compound chaotic system[J]. Acta Phys Sinica, 2020, 69(24):99-111.

[3] NI Renjie, WANG Fan, WANG Jun, et al. Multi-image encryption based on compressed sensing and deep learning in optical gyrator domain[J]. IEEE Photonics Journal, 2021, 13(3):1-16.

[4] CHAI Xiuli, TIAN Ye, GAN Zhihua, et al. A robust compressed sensing image encryption algorithm based on GAN and CNN[J]. Journal of Modern Optics, 2022, 69(2):103-120.

[5] LIU Jinqiang, LIU Yining, CUI Lei, et al. MSAI: masking sensitive area of image on IoT cameras[J]. Journal of Internet Technology. 2021, 22(7):1553-1562.

[6] PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context encoders: feature learning by inpainting[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 2536-2544.

[7] IIZUKA S, SIMO-SERRA E, ISHIKAWA H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics (ToG), 2017, 36(4):1-14.

[8] YANG Chao, LU Xin, LIN Zhe, et al. High-resolution image inpainting using multi-scale neural patch synthesis [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017:6721-6729.

[9] 曹志义, 牛少彰, 张继威. 基于生成对抗网络的遮挡图像修复算法[J]. 北京邮电大学学报, 2018, 41(3): 81-86.
CAO Zhiyi, NIU Shaozhang, ZHANG Jiwei. Masked image inpainting algorithm based on generative adversarial Nets[J]. Journal of Beijing University of Posts and Telecommunications, 2018, 41(3):81-86.

[10] YAN Zhaoyi, LI Xiaoming, LI Mu, et al. Shift-net: Image inpainting via deep feature rearrangement[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 1-17.

[11] LIU Guilin, FITSUM A, KKEVIN J, et al. Image in-

painting for irregular holes using partial convolutions [C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018:85-100.

[12] XIE Chaochao, LIU Shaohui, LI Chao, et al. Image inpainting with learnable bidirectional attention maps [C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision(ICCV). Seoul, Korea: IEEE, 2019:8858-8867.

[13] LIU Hongyu, JIANG Bin, XIAO Yi, et al. Coherent semantic attention for image inpainting[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea:IEEE, 2019: 4170-4179.

[14] WANG Haofan, WANG Zifan, DU Mengnan. Score-CAM: score-weighted visual explanations for convolutional neural networks[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA:IEEE, 2020:111-119.

[15] ZHAO Lulu, SHEN Ling, HONG Richang. Survey on image inpainting research progress [J]. Computer Science, 2021, 48(3):14-26.

[16] ROJAS D J B, FERNANDES B J T, FERNANDES S M M. A review on image inpainting techniques and datasets [C]//Proceedings of the 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Porto de Galinhas, Brazil:IEEE, 2020:240-247.

[17] AZHAR S, AZAM N, HAYAT U. Text encryption using pell sequence and elliptic curves with provable security[J]. Computers, Materials & Continua, 2022, 71 (3):4971-4988.

[18] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA:IEEE, 2016:770-778.

[19] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution[C]// Proceedings of the 2016 IEEE Computer Vision-ECCV. Amsterdam, Netherlands:Springer, 2016:694-711.

[20] ABOAH A, WANG B, BAGCI U, et al. Real-time multi-class helmet violation detection using few-shot data sampling technique and yolov8[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 5349-5357.

[21] REN Shaoqing, HE Kaiming, GIRSHICK R. Faster R-CNN:towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (6): 1137-1149.

[22] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA:IEEE, 2018:586-595.

(编辑:陈燕)