基于 4D-Arnold 不等长映射的深度隐写模型参数加密研究

段新涛[a,b]，李壮[a]，张恩[a,b]

(河南师范大学 a.计算机与信息工程学院；b.教育人工智能与个性化学习河南省重点实验室，河南 新乡 453007)

摘　要：隐写模型训练过程中需要大量数据和技术投入，因此隐写模型被窃用将对其所有者造成安全威胁和经济损失.为保护隐写模型，提出了一种基于 4D-Arnold 不等长映射的隐写模型参数保护方法.方法采用置乱-扩散策略，首先，置乱阶段通过 4D-Aronld 映射对卷积层参数跨卷积核、跨通道置乱.其次，扩散阶段采用相邻参数扩散机制在相邻参数间实现数值扩散并完成参数加密.最后，第三方无法获取任何秘密信息，实现对隐写模型的保护.实验表明，隐写模型加密后提取出的图像在 PSNR，MSE，LPIPS 和 SSIM 指标以及视觉效果上，显著降低了模型原始性能，模型隐蔽通信功能丧失.此外，所提方法在保证隐写模型加密有效性和安全性的同时，还可以应用于图像分类等其他深度模型的加密保护.

关键词：AI 模型安全；参数加密；4D-Arnold 不等长映射；图像隐写模型；卷积神经网络

中图分类号：TP309.7　　**文献标志码**：A　　**文章编号**：1000-2367(2025)04-0066-08

图像隐写作为信息安全领域的重要研究分支，近年来多与深度学习相结合.利用卷积神经网络局部连接的特点，提取图像高维特征，将秘密信息嵌入最佳区域，有效提高了图像隐写的安全性、隐蔽性、鲁棒性和隐写容量[1].基于深度学习的隐写模型训练过程不仅需要硬件计算资源[2]、图像数据集和设计精良的网络结构，还需要模型剪枝、蒸馏等技术来进一步优化模型[3].所以，一个训练完备的隐写模型具有很高的经济价值.当模型被第三方攻击者恶意盗取并使用后，版权所有者将会遭到不可估量的经济损失.此外，图像隐写模型作为隐蔽通信技术，一旦被盗用，将会导致个人隐私、商业数据或机密信息的泄露[4]，造成不可挽回的损失和后果.因此，对于深度隐写模型的保护迫在眉睫.

在图像隐写模型研究过程中，训练阶段通常采用预训练模型或随机初始化模型参数，在投入数据集、硬件算力和模型优化技术等成本后得到训练完备的模型参数.因此，深度模型保护往往通过对深度模型的卷积层权重参数进行加密，混淆权重参数之间的相关性，从而降低和破坏模型原始性能，即使攻击者截获模型，也无法正常使用并从提取图像中获取秘密信息.目前，模型加密研究的目标模型多为图像分类、目标检测和自然语言处理[5-7]等，对于图像隐写模型的加密研究较少.同时，部分研究工作在模型加密过程中需要对模型重新训练或微调，不仅会损耗计算资源，还会对模型性能产生一定的影响.此外，模型水印[8]也是模型保护的一个重要思路，模型水印主要关注模型版权的验证问题，即当模型版权出现争议时，可以通过提取水印来确定版权归属.模型水印方法虽然在鲁棒性，嵌入容量和有效性等方面效果可观，但模型水印方法均需对模型进行重训练，影响深度模型原始性能.而且模型水印方法仅能在发现侵权行为后被动保护，事后取证和维权，无法在事前阻止模型的盗用.当模型水印方法应用于图像隐写模型保护时，事后保护也无法弥补个人隐私和

收稿日期：2024-01-27；**修回日期**：2024-04-30.

基金项目：国家自然科学基金(U1904123；U20B2051)；河南省高等学校重点科研项目(23A520006)；河南省科技攻关计划(222102210199).

作者简介(通信作者)：段新涛(1972－)，男，河南新乡人，河南师范大学副教授，博士，研究方向为信息安全，E-mail：duanxintao@htu.edu.cn.

引用本文：段新涛，李壮，张恩.基于 4D-Arnold 不等长映射的深度隐写模型参数加密研究[J].河南师范大学学报(自然科学版)，2025，53(4)：66-73.(Duan Xintao，Li Zhuang，Zhang En.Deep steganography model parameter encryption method based on 4D-Arnold unequal length mapping[J].Journal of Henan Normal University(Natural Science Edition)，2025，53(4)：66-73.DOI：10.16366/j.cnki.1000-2367.2024.01.27.0002.)

机密数据泄露所造成严重后果.

为了确保模型加密安全性的同时,不对模型性能产生任何影响,本文基于 Arnold 不等长映射提出了一种新的深度隐写模型参数保护方法.方法通过置乱—扩散的加密方案对深度隐写模型提取网络参数进行加密,在保证加密有效性的同时,提高了模型参数加密的安全性.主要工作包括:1)置乱阶段,提出了 4D-Arnold 不等长映射,实现对卷积层参数跨通道,跨卷积核的可逆随机置乱.2)扩散阶段,设计了一种相邻参数扩散机制,按照既定扩散路径对卷积层参数进行扩散,在相邻参数间建立相关关系,利用雪崩效应放大参数变动的影响.3)实验结果表明,本文方法在隐写模型秘密信息提取中,可以显著降低其视觉效果和客观性能指标.4)将本文方法拓展到了图像分类模型参数保护领域,验证了本文算法在其他深度模型参数保护中的适用性.

1　相关工作

根据模型保护的应用场景不同,可以将模型保护的相关工作分为模型水印和模型加密 2 种类型.模型水印[9]按照是否依赖内部参数,可分为白盒水印和黑盒水印.其中,黑盒水印基于神经网络在输出端设置后门,通过特定的触发集来验证模型版权,无须获取模型内部参数或网络结构.MERRER 等[10]提出一种基于对抗样本的黑盒水印模型保护方法,该方法通过在部分样本中添加扰动,重新标记样本标签,当发生版权纠纷时,可通过验证对抗样本来核实 IP 归属问题.白盒水印依赖于深度模型内部参数或者网络结构,一般在模型中通过重新训练嵌入水印.WANG 等[11]提出一种基于生成对抗的白盒水印模型保护方法,以模型作为水印的生成器,同时将检测模型内部参数变化的模块作为鉴别器,通过对抗训练过程,提高了水印嵌入的容量及其不可检测性.虽然基于水印的深度模型保护在水印嵌入容量、隐蔽性和鲁棒性等方面表现较好,但模型水印保护方法均需对模型进行重训练,在一定程度上会影响模型性能.同时,将版权保护问题置身于模型被第三方窃取和使用的场景之中时,模型水印虽然可以验证版权,但无法阻止第三方对模型功能的使用.而模型加密通过对模型网络结构或模型的内部参数进行加密,可以实现对深度模型知识产权的保护.LIN 等[12]基于图像分类和自然语言处理任务提出了 ChaoW 深度模型保护框架,该框架通过对权重参数位置置乱,将卷积或全连接层的卷积核置乱为混沌状态,实现对深度模型 LinkNet,GoogleNet 和 VGG16 模型的加密.PYONE 等[13]基于训练前加密保护的场景,对图像进行逐块像素变换,然后利用预处理后的图像进行训练,从而保护模型.此方法虽然可以保证加密前后模型精度和时间开销不变,但在模型每次推理前对图像的预处理环节,增加了任务整体运行开销.

上述方法在图像分类、语义分割等任务中可以有效保护模型,但对于图像隐写模型的保护效果不够理想.由于人眼的生理特性,在观察图像时,人们难以察觉出在纹理复杂区域的微小变化[14].而图像隐写模型的输入和输出均为图像,所以和其他类型数据输出的模型相比,图像隐写模型的加密难度较高.DUAN 等[15]提出了一种图像隐写模型参数保护方法,该方法通过 Josephus 置乱对深度隐写模型中的卷积参数进行置乱,实现对模型的加密保护.当该方法对提取网络全部卷积层参数置乱后,其提取结果基本不含语义信息,但仍含有色彩,而且通过置乱难以保证算法的安全性.针对上述问题,本文基于 4D-Arnold 不等长映射提出了一种深度隐写模型参数加密方法.可以针对图像隐写模型,对模型卷积层参数进行置乱和扩散,从而提高模型加密的效果和效率.

2　本文方法

首先分析图像隐写模型的参数和网络结构,确定部分加密的策略.然后介绍本文方法的整体框架.最后,对置乱阶段的 4D-Arnold 不等长映射和扩散阶段的相邻参数扩散机制进行了详细阐述.

2.1　深度隐写模型加密分析

基于深度学习的图像隐写模型由不同的模块组成,一般可分为隐藏网络和提取网络.如附录图 S1 所示,发送方通过隐藏网络将秘密图像隐藏在载体图像中,使得载体图像和载密图像在主观视觉和客观性能指标上保持一致性.接收方通过提取网络从载密图像中提取出秘密图像,使得提取出的秘密图像和原始秘密图像

保持一致性.本文方法立足 Baluja[16]、UDH[17]、StegoPnet[18]和 U-Net[19]4 种具有代表性的深度图像隐写模型,对其提取网络的参数值进行加密,实现隐写模型的版权保护.将 4 个模型网络结构中跳跃连接等结构去除后,各模型提取网络卷积层结构如附录图 S2 所示.虽然各个模型的网络组成各不相同,但其主要组成部分均为卷积层.卷积层能够通过卷积核从载密图像纹理丰富的高频区域中提取图像的高维特征,从而完成秘密图像的隐藏和提取的任务.因此对隐写模型参数的加密应基于卷积核参数,可以通过对卷积核参数的置乱和扩散,破坏隐写模型隐蔽通信的功能.

在隐写模型的加密保护中,仅对隐藏网络或提取网络卷积层加密即可破坏双向通信的闭环,但和提取网络相比,隐藏网络的参数量较大.附录表 S1 和附录表 S2 给出了 4 种深度隐写模型提取网络和隐藏网络的网络结构和参数量统计.从参数分布情况可以发现,U-Net 和 UDH 隐藏网络参数量比其提取网络高 3 个数量级,且 4 种模型隐藏网络的深度远远超过其提取网络.此外,考虑到图像隐写的实际应用场景,若选择隐藏网络进行加密,非法授权者仍然可以通过窃取的提取网络和载密图像实现秘密信息的提取,对隐蔽通信双方造成了安全威胁.而对提取网加密,即使攻击者截获载密图像和提取网络,仍无法获得正确的提取结果.所以本文方法采用仅加密提取网络的方案来保护模型.

2.2 方法框架

本文方法的整体框架如图 1 所示,隐写模型 M 可分为隐藏网络 H 和提取网络 R.虽然仅对隐藏网络 H 加密时可以防止非法用户的使用,但仍然以通过提取网络 R 对截获的载密图像进行提取,无法保证秘密图像传输的安全性.因此,在确保安全性的前提下,仅对提取网络 R 进行加密,可以提高加密和解密过程的效率.提取网络主要由卷积层组成,通过对卷积层中卷积核参数进行加密,即可实现对模型的加密保护.

本文方法采用置乱-扩散的加密模式,首先采用 4D-Arnold 不等长映射算法,将提取网络 R 第 n 层卷积核参数在 4D 空间中进行跨卷积核,跨通道置乱,打乱各个参数的顺序.然后再通过相邻参数间的扩散,利用雪崩效应进一步提高算法抵抗差分攻击的能力.最后得到加密后的提取网络 R′,进而获得加密后的隐写模型 M′.此外,本文方法采用对称加密机制,加密和解密过程密钥相同,且加密和解密过程互逆.当 M′通过公共信道安全传输至接收端后,首先将模型提取网络中各卷积层参数进行相邻参数逆扩散,打破参数值之间的相关性;然后对参数进行 4D-Arnold 不等长逆映射后得到 R″;最后和隐藏网络组合得到和原始模型 M 一致的 M″.

2.3 4D-Arnold 不等长映射

深度隐写模型提取网络参数主要集中于卷积层的卷积核,而各个卷积层的卷积核参数实质上为一个 4D 张量 $\boldsymbol{T}(c,n,l,v)$,其中 c 表示输入通道数,n 表示输出通道数,l 和 v 分别表示卷积核长宽尺寸,一般情况下 $l=v$. 由于各卷积层的功能作用各异,张量的尺寸大小也各不相同.为了保证参数加密的安全性和有效性,本文方法中设计了一种针对张量的 4D-Arnold 不等长映射.

经典 Arnold 映射虽然具有算法简单,运行时间短,置乱效果好的特点,但同时也具有周期性.当应用于图像加密等信息安全领域时,如果变换次数恰好为周期的整数倍,那么各像素点的值不会发生变化,最终造成无效加密.同时,经典 Arnold 映射对 2D 空间域大小存在限制性,即要求 2D 空间的横纵等长.

2D-Arnold 不等长映射[20]是在经典 Arnold 映射的基础上改进的工作,和传统 Arnold 映射相比,它可以实现不等长尺寸的 2D 置乱,同时摆脱周期性的限制.此外,利用反变换方程解密,算法效率更高.对于 $M\times N$ $(M\neq N)$的 2D 矩阵,可使用 2D-Arnold 不等长映射对其内部数据进行置乱,在不改变数据值的同时,变换其位置.当 M 和 N 互为素数时,正、逆变换方程为

$$\begin{cases}x'=\mathrm{mod}(x+by,M),\\ y'=\mathrm{mod}(y,N),\end{cases}\quad \begin{cases}x=\mathrm{mod}(a^{-1}(x'+\lfloor bN/M\rfloor\cdot M-by),M),\\ y=\mathrm{mod}(d^{-1}y',N),\end{cases}$$

其中,$b=1$,$a=1$,$d=1$.通过上述变换方程即可对 M 和 N 不等且互为素数的 2D 空间实现位置置乱.当 M 和 N 互为合数,即存在除 1 以外的公因数时,正、逆变换方程为

$$\begin{cases}x'=\mathrm{mod}(x+by,M),\\ y'=\mathrm{mod}(cx+(1+bc)y,N),\end{cases}\quad \begin{cases}x=\mathrm{mod}(x'-by,M),\\ y=\mathrm{mod}(y'-cx',N),\end{cases}$$

其中,$b=1$,$a=1$,$c=N/\gcd(M,N)$,$\gcd()$ 为最大公约数函数.通过上述变换方程即可对 M 和 N 不等,且

互为合数的 2D 空间实现位置置乱.

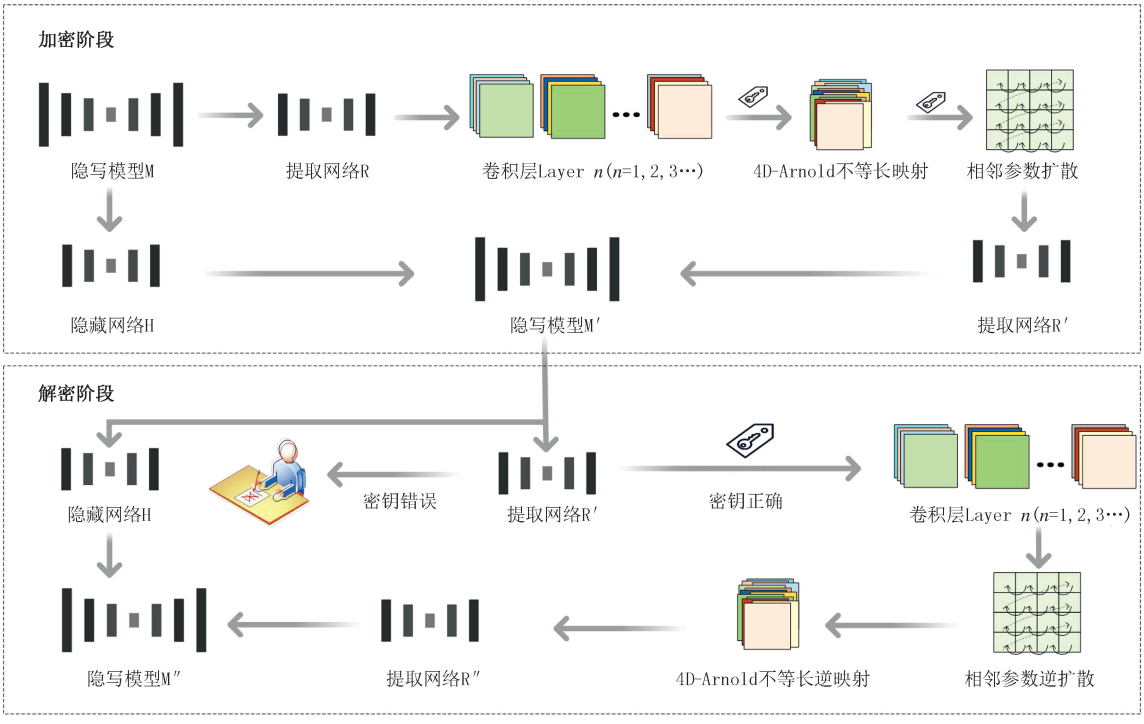


图1　参数加密方案框架

Fig. 1　The framework of the parameter encryption scheme

如图 2 所示，4D-Arnold 不等长映射算法的输入为原始卷积层的 4D 张量参数，为了保证在卷积核核内参数置乱的前提下，实现跨通道和跨卷积核的模型参数置乱.算法引入 2D 不等长 Arnold 映射，每轮置乱 4D 张量中的 2D(附录表 S3).例如，当采用(c，n)作为坐标平面时，(l，v)为 2D 参数平面，通过 Arnold 不等长映射可将 $c\times n$ 个尺寸为 $l\times v$ 的参数平面在空间内进行置乱.此外，为了保证在 4D 空间中参数置乱的随机性，算法基于 Logistic 映射设计了一种置乱顺序生成器，通过密钥 x_0 和 μ 生成相应的置乱顺序，以达到更好的置乱效果.

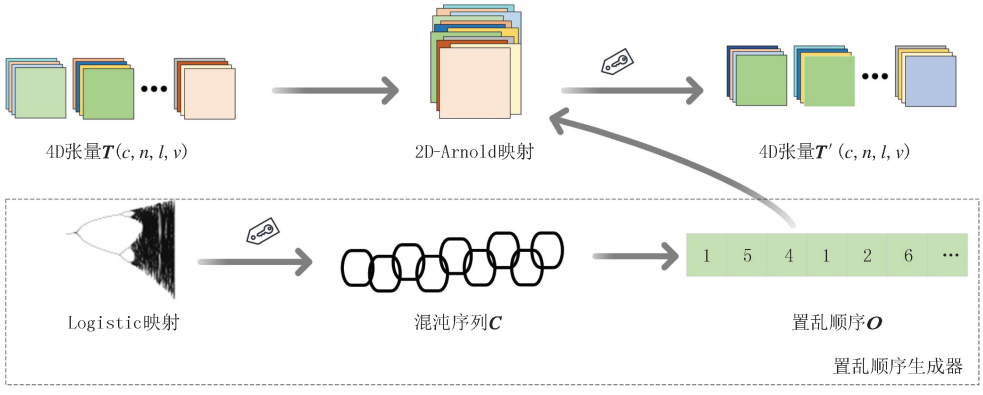


图2　4D-Arnold不等长映射

Fig. 2　4D-Arnold unequal length mapping

经典 Logistic 映射[21] $X_{n+1}=X_n\times\mu\times(1-X_n)$，$\mu\in[0,4]$，$X_0\in[0,1]$ 具有不确定、不可重复和不可预测等特性，在密码学中应用广泛.研究表明当满足条件 $\mu\in[3.569\ 945\ 6,4]$，$X_0\in[0,1]$ 时，Logistic 映射

处于混沌态.在此范围之外,生成的序列不具有伪随机性.方法利用 Logistic 生成长度为 $1\ 000+k_c$ 的混沌序列,取末尾 k_c 个值作为混沌序列 C.如式(4)所示,可将混沌序列映射为置乱顺序 $O=\mathrm{mod}(\mathrm{floor}(1\ 000\times C),6)$,用于控制每轮映射的坐标平面和参数平面.$O$ 中元素值为1,2,…,6.如附录表 S3 所示,当 $O(i)=1$ 时,即选择(c,n)作为坐标平面,对(l,v)形成的参数平面进行置乱.

2.4 相邻参数扩散机制

为了进一步增加参数加密的抗攻击性,在置乱操作结束后,设计了参数扩散机制:

$$\begin{cases}\boldsymbol{T}(c_x,n_y,l_i,v_j)=\boldsymbol{T}(c_x,n_y,l_i,v_j)+d\times\boldsymbol{T}(c_x,n_y,l_{i+1},v_j),i<l,j\le v,\\ \boldsymbol{T}(c_x,n_y,l_i,v_j)=\boldsymbol{T}(c_x,n_y,l_i,v_j)+d\times\boldsymbol{T}(c_x,n_y,l_i,v_{j+1}),i=l,j<v,\end{cases}$$

在模型卷积层参数中,$\boldsymbol{T}(c_x,n_y,l,v)$表示 n_y 卷积的 c_x 通道上的参数平面,扩散机制按照“Z”形扩散路径,从左到右、自上而下的顺序,对各卷积核每个通道中的参数平面进行相邻参数扩散.即通过扩散系数 d 将右侧参数值叠加到左侧参数,在水平方向上将相邻参数之间互相关联起来,以提高参数加密的抗攻击型.

逆扩散机制的系统方程为

$$\begin{cases}\boldsymbol{T}(c_x,n_y,l_i,v_j)=\boldsymbol{T}'(c_x,n_y,l_i,v_j)-d\times\boldsymbol{T}(c_x,n_y,l_{i+1},v_j),i<l,j\le v,\\ \boldsymbol{T}(c_x,n_y,l_i,v_j)=\boldsymbol{T}'(c_x,n_y,l_i,v_j)-d\times\boldsymbol{T}(c_x,n_y,l_i,v_{j+1}),i=l,j<v,\end{cases}$$

逆扩散系数和扩散系数相同.在接收端按照从右到左、自下而上的顺序,对各卷积核每个通道上的参数平面进行相邻参数逆扩散,可以恢复出参数.

3 实验结果与分析

实验环境如下:硬件环境为 8.00 GB RAM,Intel(R)Core(TM)i5-12400 CPU @ 2.5 GHz;软件环境为 Windows 10,Python 3.6.13,MATLAB2018a.实验以 4 种具有代表性的深度隐写模型作为加密对象,采用本文加密方法实现了对模型的加密保护.此外,实验从 ImageNet 图像数据集[22]随机选取 10 000 张图像用于验证加密算法的有效性以及对客观性能指标的测试.

3.1 主观效果分析

为了减小时间开销,本文加密算法仅对 Baluja,UDH,StegoPnet 以及 U-Net 模型的提取网络进行了实验.如附录图 S3 所示,第 1~3 行分别为载体图像,载密图像和秘密图像,第 4 行为对各个模型提取网络加密后提取出来的秘密图像.显然,当对 4 个模型提取网络的全部卷积参数进行加密后,其提取图像在主观视觉上均显示为纯黑色,从中无法获取任何语义信息.因此,通过对隐写模型提取网络的全部卷积参数加密,可以有效保护深度隐写模型,即使载密图像和提取网络被第三方截获,也无法恢复出原始秘密信息.

3.2 客观指标分析

从均方误差(mean squared error,MSE)、峰值信噪比(peak signal to noise ratio,PSNR)、结构相似性(structure similarity index measure,SSIM)和学习感知图像块相似度(learned perceptual image patch similarity,LPIPS)[23-24]这 4 个评价指标,来度量深度隐写模型加密后提取出图像的质量,以证明模型原始功能的丧失,从而有效保证模型的安全.

实验从 ImageNet 数据集中选取 10 000 张图像对加密后的模型进行验证,分别对 4 个目标隐写模型提取网络的加密结果从上述 4 个角度进行了客观分析,并以相应指标的平均值作为最终结果.表 1 给出了 Baluja,StegoPnet,UDH 和 U-Net 这 4 种深度隐写模型在无加密,全加密和全解密模式下的评价指标.当对4 个模型提取网络卷积层参数全部加密时,其 SSIM 值均接近于 0;无加密模式下 SSIM 值均接近于 1,两者差距较大.其次,全部加密时 LPIPS 值均大于 0.96,无加密时的 LPIPS 值接近于 0,两者存在较大差异.同时,无加密时的 PSNR 均小于 7 dB,结果表明经过加密,4 个模型提取网络得到的秘密图像质量很差.最后,相较于无加密时 0.001 的 MSE,全部加密时 MSE 始终在 0.27 以上,结果说明加密后提取出的图像和原始秘密图像在像素层面存在较大差异.因此,从 4 个模型的各项客观指标来看,提取网络参数全加密的方案可以有效保护深度隐写模型的安全.在无加密和全解密状态下,4 种模型的视觉效果和各个评价指标上的表现均保持一致,说明本文方法在保证深度隐写模型安全的同时,可以无损解密出原始秘密图像.

表 1　4 种模型参数加密客观指标

Tab. 1　The parameters encryption objective metrics of the four models

Model	Mode	SSIM↓	PSNR↓	LPIPS↑	MSE↑	Model	Mode	SSIM↓	PSNR↓	LPIPS↑	MSE↑
Baluja	无加密	0.945 9	32.756 8	0.039 3	0.001 0	UDH	无加密	0.951 8	32.990 7	0.064 6	0.000 9
	全解密	0.945 9	32.756 8	0.039 3	0.001 0		全解密	0.951 8	32.990 7	0.064 6	0.000 9
	全加密	0.010 6	6.425 9	0.962 7	0.274 5		全加密	0.011 1	6.424 2	0.988 2	0.274 2
StegoPnet	无加密	0.982 5	37.734 4	0.012 4	0.000 3	U-Net	无加密	0.912 4	30.346 6	0.090 7	0.001 4
	全解密	0.982 5	37.734 4	0.012 4	0.000 3		全解密	0.912 4	30.346 6	0.090 7	0.001 4
	全加密	0.011 1	6.424 2	0.988 2	0.274 2		全加密	0.011 1	6.424 2	0.988 2	0.274 2

3.3　密钥敏感性分析

本文算法的密钥 $\boldsymbol{K}=(t,\mu,x_0,d)$，其中 t 为参数置乱阶段 Arnold 映射迭代次数，μ 和 x_0 为 Logistic 置乱顺序生成器的控制参数，d 为相邻参数扩散阶段的扩散系数.基于 Baluja 隐写模型和 U-Net 隐写模型对密钥敏感性进行了测试，如附录图 S4(a)所示为基于 Baluja 隐写模型的实验，其加密和解密密钥为 $\boldsymbol{K}=(30,3.98,0.5,1\ 000)$，密钥正确时可以实现模型参数的解密和秘密图像的正常提取.若对密钥添加扰动，令扩散系数 $d=1\ 000+10^{-13}$，此时无法恢复模型提取网络的正常功能，提取结果为单一黑色图像.如附录图 S4(b)所示为基于 U-Net 隐写模型的实验，其加密和解密密钥为 $\boldsymbol{K}=(30,3.98,0.5,1\ 000)$，可以实现模型参数的解密和秘密图像的正常提取.若令扩散系数 $d=1\ 000+10^{-13}$，同样不能恢复模型提取网络的正常功能，无法获取到任何语义信息.综上所述，本文方法密钥敏感性较好，密钥的微小误差即可导致解密失败.若模型具有 L ($L>3$)层卷积层，则其密钥空间最小为 1×10^{13L}，足以抵御暴力破解，进一步证明本文方法具有较高的安全性.

3.4　图像分类模型的加密保护

在计算机视觉研究领域中，基于深度学习的图像分类被广泛研究.图像分类模型通过卷积层提取图像高维特征，然后通过池化操作选择特征，过滤信息.模型最后通过全连接层对提取特征进行非线性聚合并完成图像分类.所以，图像分类模型的核心同样是卷积层参数，除图像隐写模型之外，图像分类模型的核心同样体现在卷积核参数.因此，本文方法适用于图像分类模型.以 Pytorch 内置预训练模型 VGG16 模型[25]为例来验证本文方法对图像分类网络的加密效果，其原始分类精度为 61.96%.VGG16 模型由 13 层卷积层组成，包含 14 710 464 个参数.实验从 ImageNet 数据集中选取 1 000 类自然图像，共 50 000 张，用于模型分类精度的测试.如表 2 所示，实验对 VGG16 模型各层卷积分别加密后，精度各不相同，在 0.1%上下波动，最高和最低精度相差 0.098%.结果表明对卷积层分别加密后，其分类精度均接近于随机分类的概率(各层均约为 0.1%).相比 ChaoW 方法，本文方法在 VGG16 模型中的分类准确率更为稳定.因此，本文方法适用于图像分类模型的加密保护，并且在 VGG16 模型的加密保护上优于 ChaoW 方法.

表 2　不同方法加密 VGG16 模型后的分类精度

Tab. 2　Classification accuracies of VGG16 model based on different encryption methods　%

层数序列	1	2	3	4	5	6	7	8	9	10	11	12	13
ChaoW	70.36	0.26	40.89	0.12	35.69	0.11	0.16	46.71	0.12	0.12	0.13	0.17	**0.09**
本文	**0.12**	**0.17**	**0.10**	**0.11**	**0.10**	**0.09**	**0.11**	**0.07**	**0.10**	**0.11**	**0.10**	**0.10**	0.10

4　总　结

由于图像的视觉冗余性较高，目前的隐写模型加密算法对其保护能力较弱，安全性不高.本文基于 2D-Arnold 不等长映射和 Logistic 映射构造了一种 4D-Arnold 不等长映射，对参数置乱.同时设计了相邻参数扩散机制，进一步提高了本文方法的有效性和安全性.同时，对 4 种具有代表性的深度隐写模型加密后，提取结果的主观视觉效果和 PSNR，SSIM，MSE 和 LPIPS 指标均可表明加密后模型功能完全丧失，即本文方法可

以安全有效地保护深度隐写模型不受第三方窃取和使用.此外,基于 VGG16 图像分类模型的实验结果表明,本文方法同样适用于图像分类模型的加密保护.在未来的工作中,可将本文算法拓展研究,使其应用于其他具有卷积结构的深度模型参数加密保护工作.

附录见电子版(DOI:10.16366/j.cnki.1000-2367.2024.01.27.0002).

参 考 文 献

[1] SETIADI D R I M,RUSTAD S,ANDONO P N,et al.Digital image steganography survey and investigation(goal,assessment,method,development,and dataset)[J].Signal Processing,2023,206:108908.

[2] 高岚,赵雨晨,张伟功,等.面向 GPU 并行编程的线程同步综述[J].软件学报,2024,35(2):1028-1047.
GAO L,ZHAO Y C,ZHANG W G,et al.Survey on thread synchronization in GPU parallel programming[J].Journal of Software,2024,35(2):1028-1047.

[3] CORTIÑAS-LORENZO B,PÉREZ-GONZÁLEZ F.Adam and the ants:on the influence of the optimization algorithm on the detectability of DNN watermarks[J].Entropy,2020,22(12):1379.

[4] ABD-EL-ATTY B.A robust medical image steganography approach based on particle swarm optimization algorithm and quantum walks[J].Neural Computing and Applications,2023,35(1):773-785.

[5] 魏甫豫,张振宇,梁桂珍.基于卷积神经网络下昆虫种类图像识别应用研究[J].河南师范大学学报(自然科学版),2022,50(6):96-105.
WEI F Y,ZHANG Z Y,LIANG G Z.Research on application of insect species image recognition based on convolutional neural network[J].Journal of Henan Normal University(Natural Science Edition),2022,50(6):96-105.

[6] KAUR R,SINGH S.A comprehensive review of object detection with deep learning[J].Digital Signal Processing,2023,132:103812.

[7] TREVISO M,LEE J U,JI T C,et al.Efficient methods for natural language processing:a survey[J].Transactions of the Association for Computational Linguistics,2023,11:826-860.

[8] TRAMÈR F,ZHANG F,JUELS A,et al.Stealing machine learning models via prediction APIs[C]//Proceedings of the 25th USENIX Conference on Security Symposium.[S.l.:s.n.],2016:601-618.

[9] LI Y,WANG H X,BARNI M.A survey of Deep Neural Network watermarking techniques[J].Neurocomputing,2021,461:171-193.

[10] LE MERRER E,PÉREZ P,TRÉDAN G.Adversarial frontier stitching for remote neural network watermarking[J].Neural Computing and Applications,2020,32(13):9233-9244.

[11] WANG T H,KERSCHBAUM F.RIGA:covert and robust white-box watermarking of deep neural networks[C]//Proceedings of the Web Conference 2021.New York:ACM,2021:993-1004.

[12] LIN N,CHEN X M,LU H,et al.Chaotic weights:a novel approach to protect intellectual property of deep neural networks[J].IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,2021,40(7):1327-1339.

[13] PYONE A,MAUNG M,KIYA H,et al.Training DNN model with secret key for model protection[C]// 2020 IEEE 9th Global Conference on Consumer Electronics.Piscataway:IEEE Press,2020:818-821.

[14] 高敏娟,党宏社,魏立力,等.全参考图像质量评价回顾与展望[J].电子学报,2021,49(11):2261-2272.
GAO M J,DANG H S,WEI L L,et al.Review and prospect of full reference image quality assessment[J].Acta Electronica Sinica,2021,49(11):2261-2272.

[15] DUAN X T,SHAO Z Q,WANG W X,et al.A steganography model data protection method based on scrambling encryption[J].Computers,Materials & Continua,2022,72(3):5363-5375.

[16] BALUJA S,BALUJA S.Hiding images in plain sight[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems.New York:ACM,2017:2066-2076.

[17] ZHANG C N,BENZ P,KARJAUV A,et al.Udh:Universal deep hiding for steganography,watermarking,and light field messaging[J].Advances in Neural Information Processing Systems,2020,33:10223-10234.

[18] DUAN X T,WANG W X,LIU N,et al.StegoPNet:image steganography with generalization ability based on pyramid pooling module[J].IEEE Access,2020,8:195253-195262.

[19] DUAN X T,JIA K,LI B X,et al.Reversible image steganography scheme based on a U-Net structure[J].IEEE Access,2019,7:9314-9323.

[20] SHAO L P,QIN Z,GAO H J,et al.2D triangular mappings and their applications in scrambling rectangle image[J].Information Technology Journal,2007,7(1):40-47.

[21] 王鲜芳,王晓雷,王俊美,等.一种动态猫映射混沌图像加密算法[J].河南师范大学学报(自然科学版),2018,46(5):110-117.
WANG X F,WANG X L,WANG J M,et al.A chaotic image encryption algorithm based dynamic cat map[J].Journal of Henan Normal University(Natural Science Edition),2018,46(5):110-117.

[22] 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：IEEE Press，2009：248-255.

[23] ZHANG R，ISOLA P，EFROS A A，et al.The unreasonable effectiveness of deep features as a perceptual metric[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Piscataway：IEEE Press，2018：586-595.

[24] 胡波，谢国庆，李雷达，等.图像重定向质量评价的研究进展[J].中国图象图形学报，2024，29(1)：22-44.
HU B，XIE G Q，LI L D，et al.Progress of image retargeting quality evaluation：a survey[J].Journal of Image and Graphics，2024，29(1)：22-44.

[25] SIMONYAN K，ZISSERMAN A.Very deep convolutional networks for large-scale image recognition[EB/OL].[2023-12-20].https://arxiv.org/abs/1409.1556v6.

Deep steganography model parameter encryption method based on 4D-Arnold unequal length mapping

Duan Xintao[a,b]，Li Zhuang[a]，Zhang En[a,b]

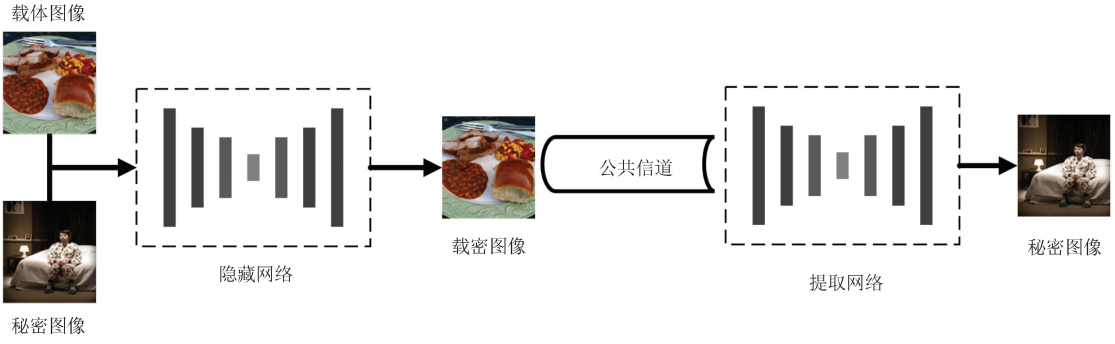
(a. College of Computer and Information Engineering；b. Key Laboratory of Educational Artificial Intelligence and Personalized Learning in Henan Province，Henan Normal University，Xinxiang 453007，China)

Abstract：The training process of steganographic models requires a large amount of data and technical investment. When the steganography model is stolen，it will cause security threats and economic losses to its owner. To prevent the theft of deep steganography models，we propose a method for protecting the parameters of the steganography model basing on 4D-Arnold unequal length mapping. The method applies a scrambling-diffusion strategy. Firstly，at the scrambling stage，we scramble the convolutional layer parameters across convolutional cores and channels through 4D-Aronld mapping. Secondly，at the diffusion stage，we design a neighboring parameter diffusion mechanism to achieve numerical diffusion between two adjacent parameters and complete the encryption of the parameters of the deep steganography model. Finally，third parties can't obtain any secret information and we realize the protection of the steganography model. Experiment results show that the method significantly reduces the original performance of the model in terms of objective indicators(PSNR，MSE，LPIPS and SSIM)and visual effects，and the hidden communication function of the model is lost. In addition，the proposed method can also be applied to the encryption protection of other deep models such as image classification while ensuring the effectiveness and security of the encryption of the steganography model.

Keywords：AI model security；parameter encryption；4D-Arnold unequal length mapping；image steganography model；convolutional neural network

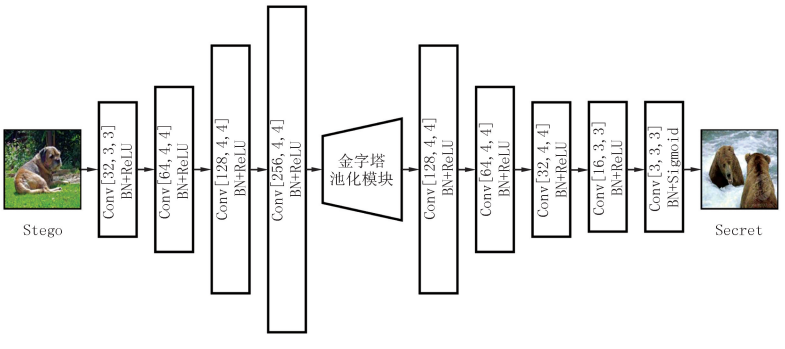
[责任编校　杨浦　刘洋]

附　录

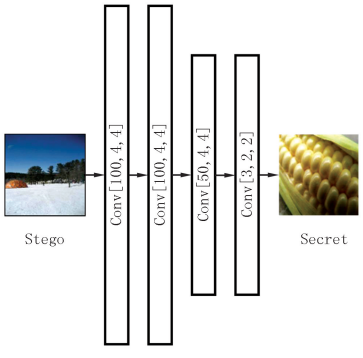


图S1　深度图像隐写模型框架

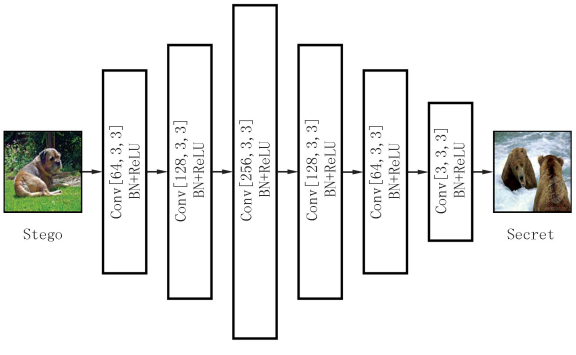
Fig.S1　The framework of the deep image steganography model



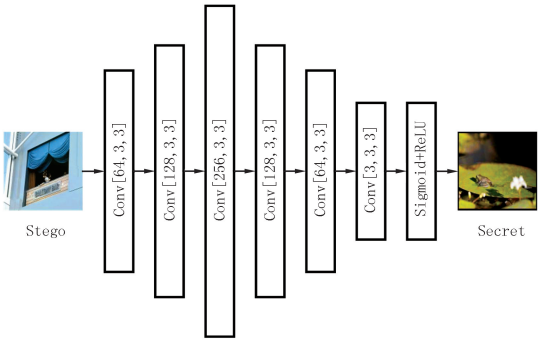
(a)StegoPnet模型提取网络



(b)Baluja模型提取网络



(c)U-Net模型提取网络



(d)UDH模型提取网络

图S2　4种模型提取网络卷积层结构

Fig.S2　The reveal network structures of the four models

表 S1　目标模型提取网络各层参数

Tab. S1　The parameters of each convolution layer in the reveal network

No.	Baluja	U-Net	UDH	StegoPnet
1	(3,100,4,4)	(3,64,3,3)	(3,64,3,3)	(3,32,3,3)
2	(100,100,4,4)	(64,128,3,3)	(64,128,3,3)	(32,64,4,4)
3	(100,50,4,4)	(128,256,3,3)	(128,256,3,3)	(64,128,4,4)
4	(50,3,2,2)	(256,128,3,3)	(256,128,3,3)	(128,256,4,4)
5		(128,64,3,3)	(128,64,3,3)	(576,128,4,4)
6		(64,3,3,3)	(64,3,3,3)	(256,64,4,4)
7				(128,32,4,4)
8				(64,16,3,3)
9				(16,3,3,3)
Sum	245 400	740 736	740 736	2 205 968

表 S2　目标模型隐藏网络各层参数

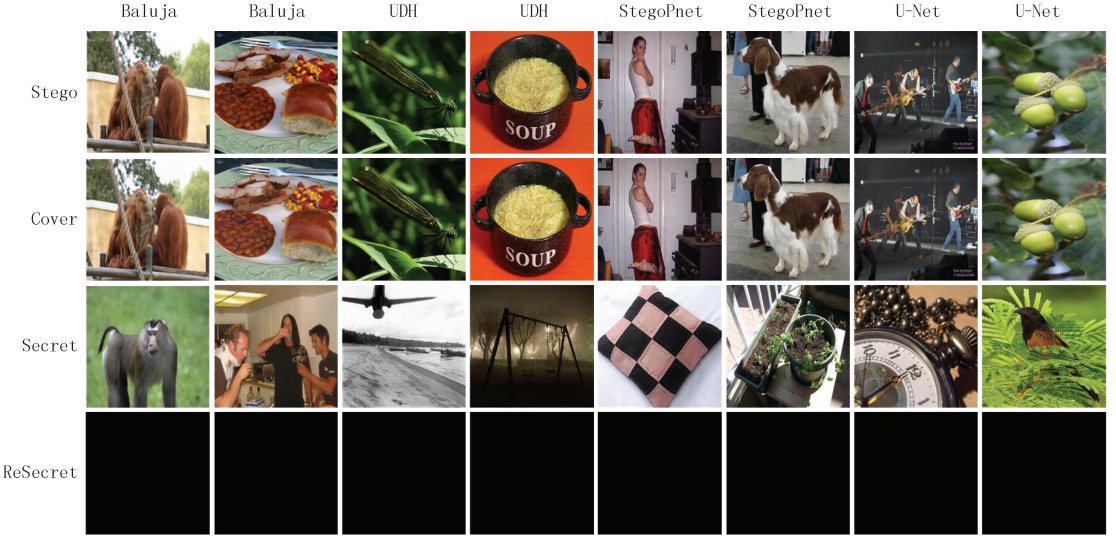
Tab. S2　The parameters of each convolution layer in the hiding network

No.	Baluja	U-Net	UDH	StegoPnet
1	(6,50,4,4)	(6,64,4,4)	(6,64,4,4)	(6,32,3,3)
2	(50,50,4,4)	(64,128,4,4)	(64,128,4,4)	(32,64,4,4)
3	(50,50,4,4)	(128,256,4,4)	(128,256,4,4)	(64,128,4,4)
4	(50,30,2,2)	(256,512,4,4)	(256,512,4,4)	(128,256,4,4)
5	(30,10,2,2)	(512,512,4,4)	(512,512,4,4)	(256,64,1,1)
6	(10,50,2,2)	(512,512,4,4)	(512,512,4,4)	(256,64,1,1)
7	(50,50,4,4)	(512,512,4,4)	(512,512,4,4)	(256,64,1,1)
8	(50,50,4,4)	(512,512,4,4)	(512,512,4,4)	(256,64,1,1)
9	(50,50,4,4)	(1 024,1 024,4,4)	(1 024,512,4,4)	(256,64,1,1)
10	(50,30,2,2)	(1 536,1 024,4,4)	(1 024,512,4,4)	(576,128,4,4)
11	(30,3,2,2)	(1 536,512,4,4)	(1 024,256,4,4)	(256,64,4,4)
12		(768,256,4,4)	(512,128,4,4)	(128,32,4,4)
13		(384,128,4,4)	(256,64,4,4)	(64,16,4,4)
14		(192,64,4,4)	(128,3,4,4)	(16,3,3,3)
15		(64,3,3,3)		
Sum	220 360	78 191 168	41 824 256	2 295 920

表 S3　坐标平面和参数平面

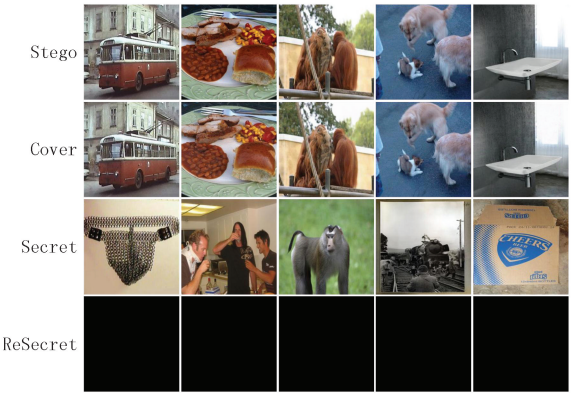
Tab. S3　Coordinate plane and parameter plane

No.	1	2	3	4	5	6
坐标平面	(c,n)	(c,l)	(c,v)	(n,l)	(n,v)	(l,v)
参数平面	(l,v)	(n,v)	(n,l)	(c,v)	(c,l)	(c,n)

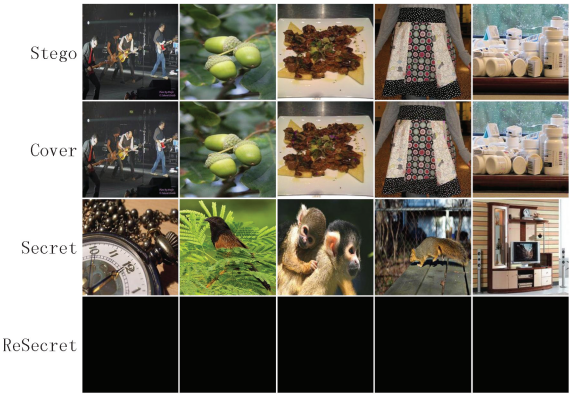


图S3　4种模型参数加密视觉效果

Fig.S3　The parameters encryption visual effects of the four models



(a)Baluja敏感性分析实验效果



(b)U-Net敏感性分析实验效果

图S4　敏感性分析实验结果

Fig.S4　The results of the sensitivity analysis experiments