基于对抗攻击的图像隐写策略搜索

李 林*，范明钰，郝江涛

(1. 电子科技大学计算机科学与工程学院 成都 611731；2. 中国核动力研究设计院核反应堆系统设计技术重点实验室 成都 610213)

【摘要】传统的隐写方法依赖于难以构建的复杂的人工规则。基于富特征模型和深度学习的隐写分析方法击败了现有最优的隐写方法，这使得隐写的安全性面临挑战。为此提出了一种基于对抗攻击的图像隐写策略的搜索方法，以寻找合适的隐写策略。隐写模型首先根据已知隐写算法初始化失真代价，然后建立含参的代价调整策略。对手模型区分载体和载密图像的分布，以发现潜在的隐藏行为。针对对手模型，利用定向对抗攻击得到相应的基于梯度符号的评价向量。在隐写模型与对手模型之间建立对抗博弈过程，据此搜索目标隐写策略。隐写模型和对手模型均用深度神经网络模型实现。构建了 4 种隐写配置并同 3 种隐写方法进行了实验比较。结果表明，该方法能有效搜索到图像隐写策略，与人工设计的经典方法和最新的隐写方法相比具有竞争力。

关 键 词 对抗攻击； 深度学习； 安全博弈； 隐写

中图分类号 TP37 **文献标志码** A doi:10.12178/1001-0548.2021335

Search the Steganographic Policy for Image via Adversarial Attack

LI Lin*, FAN Mingyu, and HAO Jiangtao

(1. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731;
2. Science and Technology on Reactor System Design Technology Laboratory, Nuclear Power Institute of China Chengdu 610213)

Abstract Steganography is to conceal the presence of secret communication. Traditional steganographic schemes rely on complex artificial rules that are difficult to construct. Steganalysers based on the rich models and deep learning achieve state-of-the-art performance. The security performance of existing steganographic methods is being challenged. In this paper, a search method based on image steganography model against attack is proposed to find a suitable steganography policy. The steganographic model constructs the parametric policy. The adversary model distinguishes the distribution of stego from cover to find the potential hiding artefacts. To obtain the corresponding evaluations, the adversarial attack is performed on adversary model. The security game between steganographic part and adversary is established via corresponding information, thus finding the target steganographic policy. The steganographic model and adversary model are implemented as deep neural networks. On the data set Bossbase, the payload is 0.2 and 0.4 bpp, the steganalysers are SRM and maxSRMd2. Four configurations with three steganographic schemes are compared. The experimental results show that the scheme proposed in this paper can obtain effective policy for image steganography, and the security performance is competitive compared with these schemes.

Key words adversarial attack; deep learning; security game; steganograhy

隐写是一种隐藏秘密通信存在性的方法，在近三十年得到广泛应用和研究，其经典模型被描述为囚徒问题[1]。隐写的基本原理是，载体中有信息冗余空间可用于建立隐蔽信道，基本方法是将消息嵌入在载体的元素中。常用的载体有图像，其冗余空间体现为像素或 DCT 非零系数。隐写努力隐藏，而与之相对应的隐写分析则尽可能地分析潜在的隐写痕迹，以发现可能的隐写行为。隐写和隐写分析之间存在一个对抗博弈过程。

在图像载体上的早期隐写方法 Lsb(least significart bit) 直接将消息比特串嵌入在像素的最小影响位平面。Lsb 破坏最小影响位平面的统计特征，容易被检测，如 Chi-squared 攻击[2] 就能有效检测 Lsb。在更高阶的统计特征上的保持与分析之间同样存在类

收稿日期：2021－11－11；修回日期：2022－01－09
基金项目：国家自然科学基金 (60373109)；
作者简介：李林 (1989－)，博士生，主要从事隐写、隐写分析、深度学习和信息安全等方面的研究.
*通信作者：李林，E-mail：2011_lilin@sina.cn

似的对抗博弈过程。

基于失真最小化框架的隐写是最流行的一类方法。通过构建失真函数，并采用一个已知编码过程如 STC[3]，该框架将隐写转变为寻找更好的失真函数问题[4]。大量的高级隐写方法基于失真最小化框架取得了优秀的抗分析表现，如 Hugo[5]、SUNIWARD[6] 等。然而，高维数据分布难以捕获，且失真最小化同隐写安全之间的关系尚不清晰。

基于富特征模型和深度学习的隐写分析方法取得了目前最好的分析表现。深度学习在很多任务上取得了优秀的表现，且已经被引入到隐写和隐写分析中[7]，如 XuNet[8]、YeNet[9] 及 SRNet[10]。生成对抗网络[11] 以及对抗攻击[12] 已被成功运用在隐写中，然而抗隐写分析的安全性尚且不足，仍然有很大的改进空间。

本文通过隐写策略模型和对手之间的对抗博弈过程寻找合适的隐写策略。基于对抗攻击，改进代价调整函数，从而实现更好的隐写修改。通过代价初始化及载体复杂性约束，避免对抗过程中的模式坍缩问题。实验表明，该框架能有效搜索隐写策略。

1 隐 写

令$x\in X\subset A^n$表示载体对象, 令$y\in Y\subset X$表示载密对象，其中$A=\{0,1,2,\cdots,q\}$表示载体和载密中元素取值的变化范围，n为载体和载密对象中元素的个数。令$m\in M=\{0,1\}^{|m|}$表示消息，$k\in K=\{0,1\}^{|k|}$表示密钥。本文考虑载体和载密为灰度图像，即$q=255$, $n=H\times W$，其中H、W分别为图像的高和宽。

隐写算法分为消息嵌入和提取两个过程。消息嵌入过程将消息负载m嵌入载体对象x中产生载密对象y，引入密钥k可以增强安全性。消息提取过程则是从载密对象y中提取出对应的消息m。为了隐藏秘密通信的存在性，隐写算法需要抵抗隐写分析的攻击。

令f_{emb}为嵌入过程，f_{ext}为提取过程。P_X,P_Y分别为载体和载密对象的分布，$D(P_X\|P_Y)$为P_X与P_Y之间的距离函数。隐写模型通过最小化$D(P_X\|P_Y)$以抵抗可能的分析对手的攻击。形式上，隐写模型表示如下：

$$f_{\text{emb}}:X\times M\times K\mapsto Y \tag{1}$$

$$f_{\text{ext}}:Y\times K\mapsto M \tag{2}$$

$$\begin{aligned}&\operatorname{argmin}D(P_X\|P_Y)\\&\text{s.t. } f_{\text{ext}}(f_{\text{emb}}(x,m,k),k)=m\end{aligned} \tag{3}$$

主流的隐写方法采用失真代价最小化原则，寻找隐写编码使得修改载体对象引起的失真影响最小化。令$d(x,y)$为失真代价函数，则三元嵌入的失真代价函数为：

$$d(x,y)=\sum_{i=1}^{H\times W}(\rho_i^+\delta_1(y_i-x_i)+\rho_i^-\delta_{-1}(y_i-x_i)) \tag{4}$$

式中，ρ_i^+与ρ_i^-分别为元素改变量δ_1与δ_{-1}的失真代价。

然而，单一的设计策略难以获得更好的$d(x,y)$，且$d(x,y)$同隐写安全之间的关系尚不明朗。失真代价函数的设计仍然是一个具有挑战性的问题。

2 对抗攻击

考虑针对隐写分析的对抗攻击问题，分析者训练一个分类器以区分载体和载密。给定分类器$F:X\mapsto L$, $L\in\{0,1\}$为F的分类输出。攻击者通过构造对抗样本x_{adv}使得分类器误分类，即$F(x_{\text{adv}})\neq F(x)$。当攻击者可以直接查询分类器$F$的模型和参数，从而获得相应的梯度信息以构造对抗样本，称之为白盒攻击。当攻击者只能访问一些分类结果，而无法查看分类器的内部模型和参数，称之为黑盒攻击。

3 图像隐写策略搜索

图像隐写策略搜索框架如图 1 所示。该框架主要由隐写策略模型、对手模型、梯度符号向量以及对抗博弈过程组成。

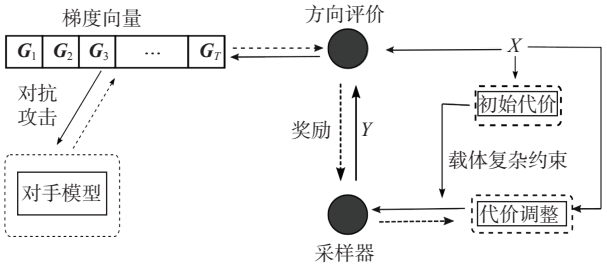

图 1 隐写策略搜索框架

3.1 隐写策略模型

采用三元嵌入隐写机制，如式 (5) 所示。对于载体$x\in X$中的每个元素x_i，分别以概率P_i^{σ}改变为：

$$x_i=x_i+\sigma\quad\sigma\in\{-1,0,+1\} \tag{5}$$

考虑相同的载体元素修改概率，即$P_i^{+1}=P_i^{-1}\leqslant\frac{1}{3}$。三元熵可以度量嵌入隐写的容量$C$，为：

$$C=\sum_{i=1}^{H\times W}\sum_{\sigma\in\{-1,0,+1\}}P_i^{\sigma}\log_2\frac{1}{P_i^{\sigma}} \tag{6}$$

为了提升隐写安全性，同很多实用的隐写机制一样，先利用加密算法将消息m加密成密文消息比特串，再将其作为隐写机制负载的消息。

令ρ^{σ}表示初始的失真代价，$\phi^{\sigma}(x)=\pi(x;\theta_s)$表示含参的代价调整策略，$\phi^{\sigma}(x)\in$阈值范围$(a,b)$。

嵌入改变概率与失真代价之间的转换关系为：

$$P_i^{\sigma}=\frac{\mathrm{e}^{-\lambda\rho_i^{\sigma}}}{\sum_{\sigma\in\{-1,0,+1\}}\mathrm{e}^{-\lambda\rho_i^{\sigma}}} \tag{7}$$

设置的代价调整策略为：

$$\rho_{\phi}^{\sigma}=\phi^{\sigma}(x)\rho^{\sigma} \tag{8}$$

隐写策略模型中的机制过程如下。

1) 建立初始失真代价ρ^{σ}以及载体复杂约束；

2) 根据代价调整策略$\phi^{\sigma}(x)$将ρ^{σ}调整为ρ_{ϕ}^{σ}；

3) 当代价调整过程结束，得到最终的嵌入失真代价$\rho_{\mathrm{emb}}^{\sigma}$；

4) 给定载体对象以及消息负载，利用实用的隐写编码STC，找到合适的载密对象来传递秘密消息。

3.2　采样器与策略更新

根据失真代价ρ_{ϕ}^{σ}以及消息负载搜索出修改概率张量P_{ϕ}^{σ}，可利用一个最优的嵌入仿真器采样载密对象，如式 (9) 所示：

$$\mathrm{md}_i=\begin{cases}-1 & r_i<P_{\phi}^{-1}(i)\\ 1 & r_i>1-P_{\phi}^{+1}(i)\\ 0 & \text{其他}\end{cases} \tag{9}$$

式中，r_i为服从均匀分布的随机变量。然而，这个仿真器难以在训练过程中传递梯度信息，使得无法更新代价调整策略$\pi(x;\theta_s)$。

假设元素的嵌入操作是独立的，样本采样器根据一个固定的嵌入顺序产生样本修改序列$\mathrm{sp}(x,P_{\phi}^{\sigma})$，从而得到载密$y$。奖励函数$\mathrm{reward}(x,y)$估计$d(x,y)$的变化方向，并给出代价调整的启发式经验。为了获得训练$\pi(x;\theta_s)$的梯度信息，利用策略梯度的方式，更新隐写策略模型。隐写策略模型的更新过程为：

$$y=x+\mathrm{sp}\left(x,P_{\phi}^{\sigma}\right) \tag{10}$$

$$R=\sum_{i=1}^{H\times W}\mathrm{reward}_i(x,y) \tag{11}$$

$$P_{\mathrm{sample}}(i)=P_{\phi}^{\mathrm{sp}}(i) \tag{12}$$

$$\nabla_{\theta_s}J(\theta_s)=\mathbb{E}_{x\sim P(X)}\left(\nabla_{\theta_s}\sum_{\mathrm{i}=1}^{H\times W}\ln P_{\mathrm{sample}}(i)R\right) \tag{13}$$

3.3　对手模型

标签函数$L(o)$指示对应对象的真实标签，即：

$$L(o)=\begin{cases}0 & o\in X\\ 1 & o\in Y\end{cases} \tag{14}$$

式中，$o\in X\cup Y$。

对手模型构建分类器$F(o;\theta_{\mathrm{adv}})$以区分载体和载密。对应的训练损失如下：

$$\mathrm{Loss}_{\mathrm{adv}}=\mathbb{E}_{o\sim P_o}\mathrm{SCE}(L(o),F(o;\theta_{\mathrm{adv}})) \tag{15}$$

式中，SCE为softmax交叉熵函数；P_o为对象的分布函数。

白盒攻击可看作是针对对手模型进行的不同粒度下的查询，根据查询获得的梯度信息指示了在对手模型的评价下改进的方向。然而，对手模型的评价可能是含噪的，得到的梯度信息对应着不同的偏离程度。建立一个更好估计梯度信息的模型有助于获得更加鲁棒和迁移性更好的启发式方向。为此，提出一个可靠概率度量下的梯度模型，以提供隐写代价改进的指引信息。

利用白盒攻击实现对抗攻击方案，将定向对抗攻击的分类标签量化为[0,0.5]区间的软标签SL，利用对抗攻击得到关于载体元素的梯度向量$\boldsymbol{G_i}=(G_{i_t})$，其中，$t\in[1,T]$，$T$为总的攻击类型。

令$\boldsymbol{P}_G$表示对应对抗攻击的离散可靠概率向量。在$\boldsymbol{P}_G$度量下，将期望的梯度符号$\mathbf{ESG}(x,y)$作为改进方向估计，形式化如下式所示：

$$\mathrm{SG}_{i_t}(x,y)=\begin{cases}1 & G_{i_t}>0\\ 0 & G_{i_t}=0\\ -1 & G_{i_t}<0\end{cases} \tag{16}$$

$$\mathrm{ESG}_i(x,y)=\sum_{t=1}^{T}P_G(t)\,\mathrm{SG}_{i_t}(x,y) \tag{17}$$

根据$\mathbf{ESG}(x,y)$建立相应的奖励函数$\mathrm{reward}(x,y)$提供给隐写策略模型，以便改进。对应的奖励函数为：

$$\mathrm{reward}_i(x,y)=\lambda_r\frac{1}{2}(\mathrm{ESG}_i)(-\boldsymbol{\sigma}_i)+\lambda_m \tag{18}$$

式中，λ_r和λ_m分别为对应的奖励系数和最大限定值；$\boldsymbol{\sigma}$表示修改向量。

3.4　对抗隐写博弈过程

隐写策略模型和对手模型各自最优化自己的目

标，它们之间存在着对抗博弈过程。隐写策略模型努力调整代价以获得期望的最大奖励，而对手模型则努力区分载体和载密。通过在可靠概率度量下的对抗攻击，建立更加可靠的改进方向。隐写策略模型需要根据载体复杂约束调整代价策略，以避免模式坍缩。迭代此过程，直至收敛至均衡点或者达到要求的迭代次数，对手模型无法继续改进区分能力，而隐写策略模型也无法继续产生更好的代价改进。此时，系统搜索到目标的隐写策略。对抗隐写博弈过程的搜索目标公式为：

$$\theta_s^* = \arg\min_{\theta_s} -J(\theta_s;\theta_{\text{adv}}) \tag{19}$$

$$\text{s.t.}\ (\theta_{\text{adv}}^*) = \arg\min_{\theta_{\text{adv}}} (\text{Loss}_{\text{adv}}) \tag{20}$$

3.5 模型实现

对手模型以及隐写策略模型均利用深度神经网络构建。采用SRNet的轻量版本作为分析模型。隐写策略模型结构为：

$$\begin{aligned}\pi(x;\theta_s) = {} & \text{fc}^{\circ}\text{bn}^{\circ}\text{conv}_{D_4\to D_5}{}^{\circ} \\ & \text{relu}^{\circ}\text{bn}^{\circ}\text{conv}_{D_3\to D_4}{}^{\circ} \\ & \text{relu}^{\circ}\text{bn}^{\circ}\text{conv}_{D_2\to D_3}{}^{\circ} \\ & \text{relu}^{\circ}\text{bn}^{\circ}\text{conv}_{D_1\to D_2}(x)\end{aligned} \tag{21}$$

式中，bn为batchnormalization；relu为激活函数；fc为全连接操作；°为简单的复合操作，conv为卷积操作；D_i 为张量的通道数。这些子过程在深度学习中是经常使用的基本操作。尽管深度模型的结构很重要，但只要模型容量足够就能有效表示隐写调整策略，因此暂不考虑不同模型结构对隐写模型的可能影响。

4 实验与分析

4.1 实验设置

以空域灰度图像集Bossbase[13] 为实验数据集，随机将Bossbase分为 3 个不相交的部分，分别为训练数据集、验证数据集和测试数据集，比例分别为 0.6、0.1 和 0.3。利用Bicubic核分别将 3 个数据集缩减到256×256。

分别采用SRM以及maxSRMd2[16] 作为富模型特征提取器。利用空域富模型SRM[14] 以及FLD集成分类器[15] 进行隐写分析。以经典的空域隐写算法SUNIWARD[6] 为基本的加性初始代价函数。分别考虑两种单独的对抗策略 (z_1和z_2，分别为 0 和 0.15)，以及两种联合策略 (均匀分布u和均值为 0.05、方差为 0.1 的高斯分布gs) 下的方案，ATSteg$K$为 $K=T$，$T\in\{z1,z2,u,\text{gs}\}$对应的隐写算法。嵌入 0.2、0.4 比特/像素 (bpp)，在两种隐写分析器下，分别与SUNIWARD[6]、HILL[17]，及SPAR-RL[18] 进行安全性比较。

4.2 实现细节

利用AdamOptimizer进行优化，学习率0.000 2，$\beta_1=0.2$，$\beta_2=0.8$，批量大小为 16。$\lambda_r=1$，$\lambda_m=2$。采用DDE实验室Matlab版本的SUNIWARD实现代码[19]。运行软件平台为Tensorflow以及Python。运行硬件平台为Geforce RTX 3090GPU平台。

4.3 评价方法

假设载体和载密图像具有相等的先验概率，采用最小平均错误率P_E来评价隐写机制，即：

$$P_E = \min_{P_{\text{fa}}} \frac{(P_{\text{md}}+P_{\text{fa}})}{2} \tag{22}$$

式中，P_{fa}表示虚警率；P_{md}表示漏检率。

4.4 评价结果

不同隐写算法抗SRM分析的安全表现评价结果如表 1 所示。

表 1 不同隐写算法抵抗SRM分析的安全表现 %

算法	嵌入容量/bpp	
	0.2	0.4
SUNIWARD	33.71	21.97
HILL	37..23	26.20
SPAR－RL	38.43	28.30
ATStegz1	37.21	27.94
ATStegz2	37.32	27.43
ATStegu	37.72	27.96
ATSteggs	37.44	27.59

不同隐写算法抗maxSRMd2分析的安全表现评价结果如表 2 所示。

表 2 不同隐写算法抵抗maxSRMd2分析的安全表现 %

算法	嵌入容量/bpp	
	0.2	0.4
SUNIWARD	30.42	20.49
HILL	30.97	22.05
SPAR－RL	30.33	22.36
ATStegz1	30.15	20.92
ATStegz2	30.23	21.32
ATStegu	30.21	21.37
ATSteggs	30.22	21.12

从表 1 和表 2 的结果可以看出，ATStegz1的表现最次，而ATStegu的表现最好，可能的原因是

ATStegu捕获的决策方向更加平滑，而ATStegz1则更加分散。ATStegz2和ATSteggs则表现相近。

ATStegu在选择信道分析下依然能保持好的抗分析的能力，但在 0.2 bpp 时比SPAR-RL低 0.39%，在 0.4 bpp 时比SPAR-RL低 4.47%。说明SPAR-RL能得到更加鲁棒的代价，可能原因是SPAR-RL建模了载体状态，能更好捕获相应的载体模型。

从表 1 和表 2 的结果可以看出，ATSteg可以超越经典算法SUNIWARD和HILL，这得益于代价调整策略以及载体复杂约束。如果不对代价调整策略模型加以载体复杂约束，将会出现模式坍缩问题，即接近相同的修改概率状态，使得代价调整失败。

联合策略相比单独的策略表现更好，这意味着可以通过探索联合策略空间的概率度量，以更好地改进代价搜索。

5　结束语

本文提出了一个基于对抗攻击的隐写策略搜索框架，通过隐写策略模型与对手模型之间的对抗博弈过程，改进代价调整函数，从而实现更好的隐写修改。通过代价初始化及载体复杂性约束，避免对抗过程中的模式坍缩问题。实验表明，该框架能有效搜索隐写策略，优于传统经典算法以及接近目前最好的基于深度学习的方案。

参　考　文　献

[1] SIMMONS G J. The prisoners' problem and the subliminal channel[C]//Advances in Cryptology. Boston, MA: Springer, 1984: 51-67.

[2] WESTFELD A, PFITZMANN A. Attacks on steganographic systems[C]//International Workshop on Information Hiding. Berlin, Heidelberg: Springer, 1999: 61-76.

[3] FILLER T, JUDAS J, FRIDRICH J. Minimizing additive distortion in steganography using syndrome-trellis codes[J]. IEEE Transactions on Information Forensics and Security, 2011, 6(3): 920-935.

[4] 弗里德里希. 数字媒体中的隐写术: 原理, 算法和应用[M]. 北京: 国防工业出版社, 2014.
FRIDRICH J. Steganography in digital media: Principles, algorithms, and applications[M]. Beijing: National Defense Industry Press, 2014.

[5] PEVNÝ T, FILLER T, BAS P. Using high-dimensional image models to perform highly undetectable steganography[C]//International Workshop on Information Hiding. Berlin, Heidelberg: Springer, 2010: 161-177.

[6] HOLUB V, FRIDRICH J, DENEMARK T. Universal distortion function for steganography in an arbitrary domain[J]. EURASIP Journal on Information Security, 2014, DOI: 10.1186/1687-417X-2014-1.

[7] CHAUMONT M. Deep learning in steganography and steganalysis[M]//Digital Media Steganography. [S.l.]: Academic Press, 2020: 321-349.

[8] XU G, WU H Z, SHI Y Q. Structural design of convolutional neural networks for steganalysis[J]. IEEE Signal Processing Letters, 2016, 23(5): 708-712.

[9] YE J, NI J, YI Y. Deep learning hierarchical representations for image steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545-2557.

[10] BOROUMAND M, CHEN M, FRIDRICH J. Deep residual network for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2018, 14(5): 1181-1193.

[11] ZHANG Y, ZHANG W, CHEN K, et al. Adversarial examples against deep neural network based steganalysis[C]//Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. New York: Association for Computing Machinery, 2018: 67-72.

[12] ZHU J, KAPLAN R, JOHNSON J, et al. Hidden: Hiding data with deep networks[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, 2018: 657-672.

[13] BAS P, FILLER T, PEVNÝ T. 'Break our steganographic system': The ins and outs of organizing Boss[C]// International Workshop on Information Hiding. Berlin, Heidelberg: Springer, 2011: 59-70.

[14] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics & Security, 2012, 7(3): 868-882.

[15] KODOVSKY J, FRIDRICH J, HOLUB V. Ensemble classifiers for steganalysis of digital media[J]. IEEE Transactions on Information Forensics and Security, 2011, 7(2): 432-444.

[16] DENEMARK T, SEDIGHI V, HOLUB V, et al. Selection-Channel-Aware rich model for steganalysis of digital images[C]//2014 IEEE International Workshop on Information Forensics and Security (WIFS). Atlanta, Georgia: IEEE, 2014: 48-53.

[17] LI B, WANG M, HUANG J, et al. A new cost function for spatial image steganography[C]//2014 IEEE International Conference on Image Processing. Paris: IEEE, 2014: 4206-4210.

[18] TANG W, LI B, BARNI M, et al. An automatic cost learning framework for image steganography using deep rein- forcement learning[J]. IEEE Transactions on Information Forensics & Security, 2020, 16: 952-967.

[19] FRIDRICH J. URL Steganographic algorithms[EB/OL]. [2021-01-20]. http://dde.binghamton.edu/download/steg-oalgorithms/.

编　辑　税　红