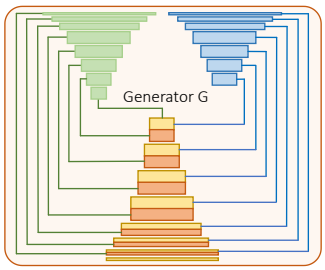
OEE | 光电工程 Opto-Electronic Engineering

Article

2026 年 2 月, 第 53 卷, 第 2 期

防御与攻击双策略图像隐写模型



郭　伟, 任哲聪*, 金海波, 曲海成, 林　畅

辽宁工程技术大学软件学院, 辽宁 葫芦岛 125105

摘要: 针对现有对抗性图像隐写方法跨模型泛化能力不足且含密图像质量低的问题, 本文提出一种融合被动防御与主动攻击的双策略图像隐写模型。被动防御模块基于生成对抗网络, 生成器采用双流 U-Net 架构同步处理载体图像与其边缘信息, 结合 SE-Net 注意力机制动态分配特征权重, 生成更适合信息嵌入的载体图像; 主动攻击模块基于神经元归因结果, 定位并扰动各类隐写分析模型判别时所依赖的共同关键特征, 针对性优化嵌入方案, 从而引导各类隐写分析模型做出错误判断。通过动态调整损失权重系数, 驱动两个模块性能递进, 最终实现全局优化。实验结果表明, 生成的含密图像平均 PSNR 达到 40.89 dB, 平均 SSIM 为 0.9783, 相比于 CR-AIS、Natias 检测准确率 ACC 分别降低了 3.69%、1.91%, 实现了跨模型泛化能力与图像质量的协同提升。

关键词: 图像隐写; 生成对抗网络; 双流 U-Net; 神经元归因; 跨模型泛化

The English title and abstract appear at the end of the paper.

DOI: 10.12086/oee.2026.250238 | CSTR: 32245.14.oee.2026.250238 | 中图分类号: TP309 | 文献标识码: A

郭伟, 任哲聪, 金海波等. 防御与攻击双策略图像隐写模型 [J]. 光电工程, 2026, **53**(2): 250238

Guo W, Ren Z C, Jin H B et al. Dual strategy image steganography model for defense and attack[J]. *Opto-Electron Eng*, 2026, **53**(2): 250238

1　引　言

近年来, 信息技术飞速发展, 图片、视频等信息得以在网络中便捷传输, 信息安全也日益受到重视。图像隐写作为信息安全领域的核心技术, 在隐蔽通信和数字版权保护等方面具有不可替代的战略价值。通过将秘密信息隐匿于载体图像中, 使得含密图像在传输过程中不引起外界的注意[1], 该技术不仅推动了信息隐藏理论的创新发展, 更为构建安全可信的数字社会提供了关键技术支撑。作为隐写的对立面, 隐写分析的目的是检测图像中是否存在秘密信息[2]。早期的隐写分析方法基于手工特征, 如 SRM[3]、DCTR[4] 等, 然而这类方法高度依赖人工经验, 难以应对复杂多变的隐写场景。近年来随着深度学习的快速发展, 出现了各种基于卷积神经网络 (convolutional neural networks, CNN) 的隐写分析方法, 如 CovNet[5]、SRNet[6]、Ye-Net[7]、Zhu-Net[8] 等, 利用 CNN 自动学习判别特征, 避免了手工特征设计的局限性, 有效提高了隐写分析的准确性, 对图像隐写技术提出了更高的安全性要求。

传统图像隐写方法通常基于最小化失真框架[9], 依据像素修改失真代价函数优化嵌入方案, 如 S-UNIWARD[10]、HILL[11] 等。然而, 这类方法在复杂场景下易因局部失真累积引发统计异常, 导致检测风险升高。因此研究重心转向通过生成或调整载体图像, 主动增强其纹理复杂度, 在确保视觉自然性的同时构造更多适合隐写的“安全区域”。近年来, 这一方向取得显著进展, 马宾等人[12] 提出了一种基于 U-Net 结构的生成式多重对抗隐写算法, 可控地生成较高质量的目标载体图像, 从

收稿日期: 2025-08-11; 修回日期: 2025-10-08; 录用日期: 2025-10-09; 网络出版日期: 2026-02-13

基金项目: 国家自然科学基金 (62173171); 国家自然科学基金面上项目 (42071351); 辽宁省教育厅基本科研项目 (JYTMS20230804)

*通信作者: 任哲聪, 1396933655@qq.com

版权所有©2026 中国科学院光电技术研究所

而增强载体图像的信息隐藏能力，但其生成图像的视觉质量与隐蔽性难以兼顾。2024 年，马宾等人[13]进一步引入通道注意力机制和多重对抗思想，协同优化了隐蔽性与视觉质量。

但预调整载体图像本质上仍属于被动防御策略，难以从根本上克服深度隐写分析模型对复杂统计异常的敏感性。为此，研究者们转向更主动的对抗性隐写方法，针对目标隐写分析模型的逻辑层 (即决策边界) 引入微小扰动，从而使其做出错误的判断。例如，ADV-EMB[14]首次将对抗性思想引入最小化失真框架，利用目标 CNN 隐写分析器的梯度反向调整修改成本，生成可同时携带消息并欺骗特定分析器的对抗含密图像。Backpack[15]针对 ADV-EMB 中存在的理论缺陷，通过可微近似与隐函数梯度传播提出了更为严谨的解决方案，进一步增强了安全性。另一方面，CR-AIS[16]则借鉴人工免疫机制，动态调整修改概率分布，将嵌入变更集中于难以检测的图像区域，显著增强了抗隐写分析的能力。但该类方法只能欺骗目标隐写分析模型，跨模型泛化能力受限。针对这一瓶颈，研究者提出了特征层攻击策略，通过扰动各类隐写分析模型做出决策时所依赖的共同关键特征以提升跨模型泛化能力。例如，Fan 等人[17]提出的 Natias 方法，通过神经元归因技术识别关键特征，进而对其进行破坏，有效增强了对抗样本的迁移性，但其扰动易导致图像纹理割裂与统计失真，在视觉质量与安全性之间仍缺乏有效平衡。

基于上述分析，本文提出一种融合被动防御与主动攻击的图像隐写模型：

1) 被动防御模块进行载体图像优化：基于双流 U-Net[18]的生成器同步处理载体图像与基于保角单演相位一致性 (conformal monogenic phase congruency, CMPC)[19]得到的边缘信息，结合 SE-Net 注意力机制优化特征分布，通过与鉴别器的对抗优化载体图像，得到视觉效果高度相似但纹理更丰富的载体图像，从而更适合信息嵌入；

2) 主动攻击模块进行对抗性嵌入：利用归因技术识别关键神经元，攻击各类模型在做出决策时所依赖的共同关键特征，联合鉴别器损失与像素级约束，针对性优化嵌入策略，在提高跨模型泛化能力的同时保持较高图像质量。

两个模块是一个协同设计的整体，并非简单组合，其通过动态权重调整机制进行联合优化，在训练中渐进平衡载体优化与对抗扰动。实验表明，本文方法能够实现图像质量与跨模型泛化能力的协同提升。

2 本文方法

本文提出了一种防御与攻击双策略图像隐写模型，对载体图像进行优化的同时进行对抗性隐写，整体架构如图 1 所示，由两部分组成：

1) 载体图像优化网络：将载体图像 $\boldsymbol{X}$ 和基于 CMPC 得到的边缘信息 $\boldsymbol{X}^+$一同输入到基于双流 U-Net

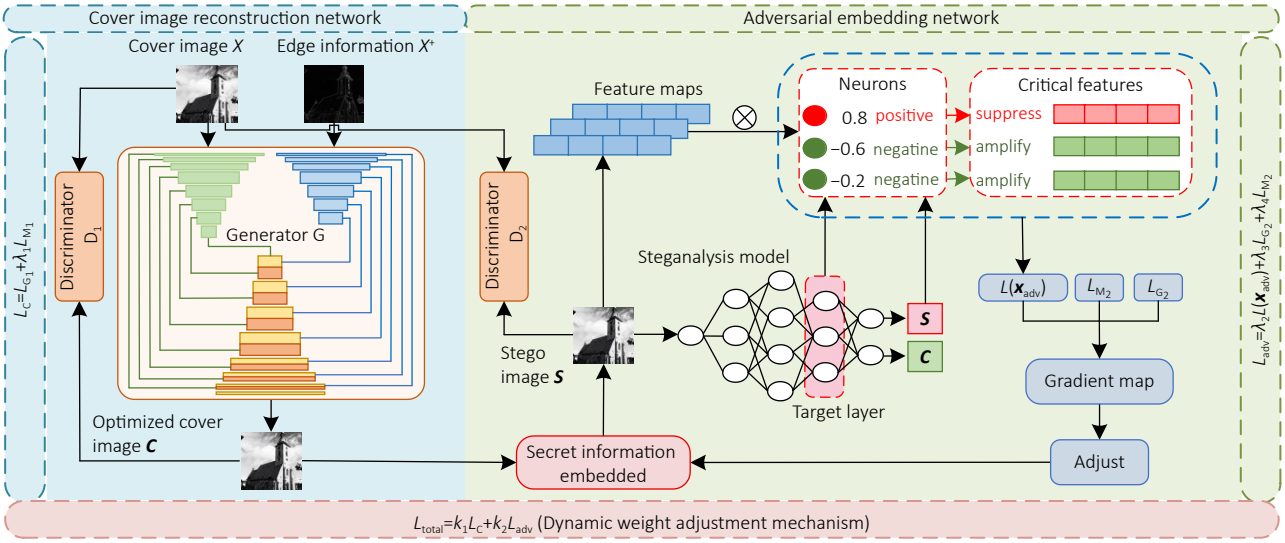


图 1 防御与攻击双策略图像模型框架

Fig. 1 Model framework

的生成器 G 中，并结合 SE-Net 注意力机制协调特征权重，通过与鉴别器 D_1 的对抗训练以及像素空间均方差损失L_{M_1}的约束得到优化后的载体图像 $\boldsymbol{C}$；

2) 对抗性嵌入网络：基于优化后的载体图像 $\boldsymbol{C}$ 进行秘密信息嵌入后，得到含密图像 $\boldsymbol{S}$ 并输入目标隐写分析模型中，当检测结果为 $\boldsymbol{S}$ 时，将判别结果归因到目标中间层中的每个神经元，抑制积极归因特征并强化消极归因特征，同时引入鉴别器 D_2 的损失L_{D_2}以及像素空间均方差损失L_{M_2}，确保含密图像 $\boldsymbol{S}$ 与原始载体图像 $\boldsymbol{X}$ 的不可区分性，不断调整对抗性嵌入的参数，破坏各类隐写分析模型做出决策所依赖的共同关键特征，使目标隐写分析器做出错误判断的同时，也能误导其他非目标隐写分析模型。两个网络之间动态调整损失权重做到协同优化，实现跨模型泛化能力与视觉质量的协同提升。

2.1 载体图像优化模块

本文在生成对抗网络 GAN (generative adversarial network) 基础上进行载体图像优化，这是由于 GAN 通过生成器与鉴别器之间的对抗机制，有效保障了生成图像的视觉自然度与真实感，是目前生成式隐写中广泛应用的成熟方案。

近年来，基于传统 U-Net 的载体图像优化方法被广泛采用，然而这类方法受限于传统 U-Net 的单一编码-解码结构，难以充分捕捉图像的高频细节与纹理特征，缺乏对关键区域的特征表达能力，使得后续秘密信息的嵌入与隐藏效果欠佳。因此本文将双流 U-Net 结构引入 GAN 的生成器中，通过双分支设计同步输入原始载体图像与基于 CMPC 得到的边缘信息，利用边缘信息增强网络对图像纹理和细节特征的捕捉能力。并且引入 SE-Net 注意力机制，通过对特征通道的显式建模，动态校准不同特征的权重。与 CBAM 等同时关注通道与空间的注意力机制相比，SE-Net 仅聚焦于通道关系，计算更轻量；而与更复杂的自注意力或非局部模块相比，SE-Net 在计算效率和实现简洁性上更具优势，更适用于需兼顾性能与复杂度的隐写应用场景。该设计显著增强了对高频细节和关键区域的表示能力，同时有效控制参数数量，维持模型的轻量化。最终生成具有高视觉质量且更适合秘密信息嵌入的载体图像，为后续嵌入操作提供更高质量的载体基础。

2.1.1 图像边缘信息提取

现有自适应图像隐写方法普遍认为，在纹理丰富的边缘区域嵌入信息可提升不可感知性与安全性，因此本文利用边缘信息增强载体图像的纹理特征，但其有效性高度依赖边缘检测的精度与完整性。传统边缘检测算子如 Sobel、Canny 等，对光照、对比度变化敏感，且易受噪声干扰，普遍存在边缘粘连和断缺现象等问题[20]。为解决上述问题，本文采用基于 CMPC 的边缘检测方法，该方法基于相位一致性原理，通过在相位域中处理信号，对光照和对比度变化表现出高度的不变性。同时，CMPC 方法利用保角映射在复域中统一表征图像中的点、线和边缘等结构，能够有效抑制传统梯度方法中常见的边缘断缺问题，从而获得更为完整和连贯的边缘响应。其表达式为

$$PC_4(x,y)=W(x,y)\left[\exp\left(\left|\frac{E_{\mathrm{CM}}}{A_\Sigma+\varepsilon}-1\right|\Big/b^2\right)-\frac{T}{A_\Sigma+\varepsilon}\right], \quad (1)$$

式中：$PC_4(x,y)$ 为图像空间域上的四方向聚合相位一致性度量，包含了图像最显著的边缘结构信息；E_{CM}表示多尺度下的局部能量；$A_\Sigma=\sum_{s=1}^{n}E_{\mathrm{CM}}$表示多尺度能量累加和；$W(x,y)$为 Sigmoid 加权函数，目的是对频谱宽度较小的点进行惩罚；$b=0.3$控制相位偏差敏感度；$T=\mu_\gamma+3\sigma_\gamma^2$为基于瑞利分布的噪声阈值；$\varepsilon$为防止除零的极小常数。

相较于传统梯度边缘检测算子及经典相位一致性方法，基于 CMPC 的方法在边缘定位精度与抗噪声干扰能力方面展现出显著优势。特别是与基于单演信号相位一致性的方法相比，CMPC 通过引入保角映射，能够更精确地检测与表征图像中的边缘、角点等关键特征，从而在复杂场景下生成连贯性更好、一致性更高的边缘图，使其更适用于对安全性与隐蔽性要求极高的图像隐写任务。

2.1.2 生成器

在生成器 G 中，使用两种类型的跳跃操作来连接编码路径和解码路径之间的特征映射，SE-Net 应用于载体图像编码路径的每一层之后，并不应用于边缘信息的编码路径之后，生成器结构如图 2 所示。这是由于载体图像通常包含丰富的特征信息和较多的噪声，SE-Net 能够有效地捕捉重要的特征并减少噪声对最终生成结果的影响，而边缘信息经过预处理，已经提取出结构化的梯度特征，进一步应用 SE-Net 会过度放大某些高频细节或噪声，导致模型过拟合、边缘结构失真，影响生成图像的整体质量。3.3 节中的消融实验结果也证实，在边缘路径中进一步引入 SE-Net 会导致生成性能下降。具体而言，首先将边缘信息编码特征与载体图像解码特征融合，以增强细节边缘与轮廓的传递，可表示为

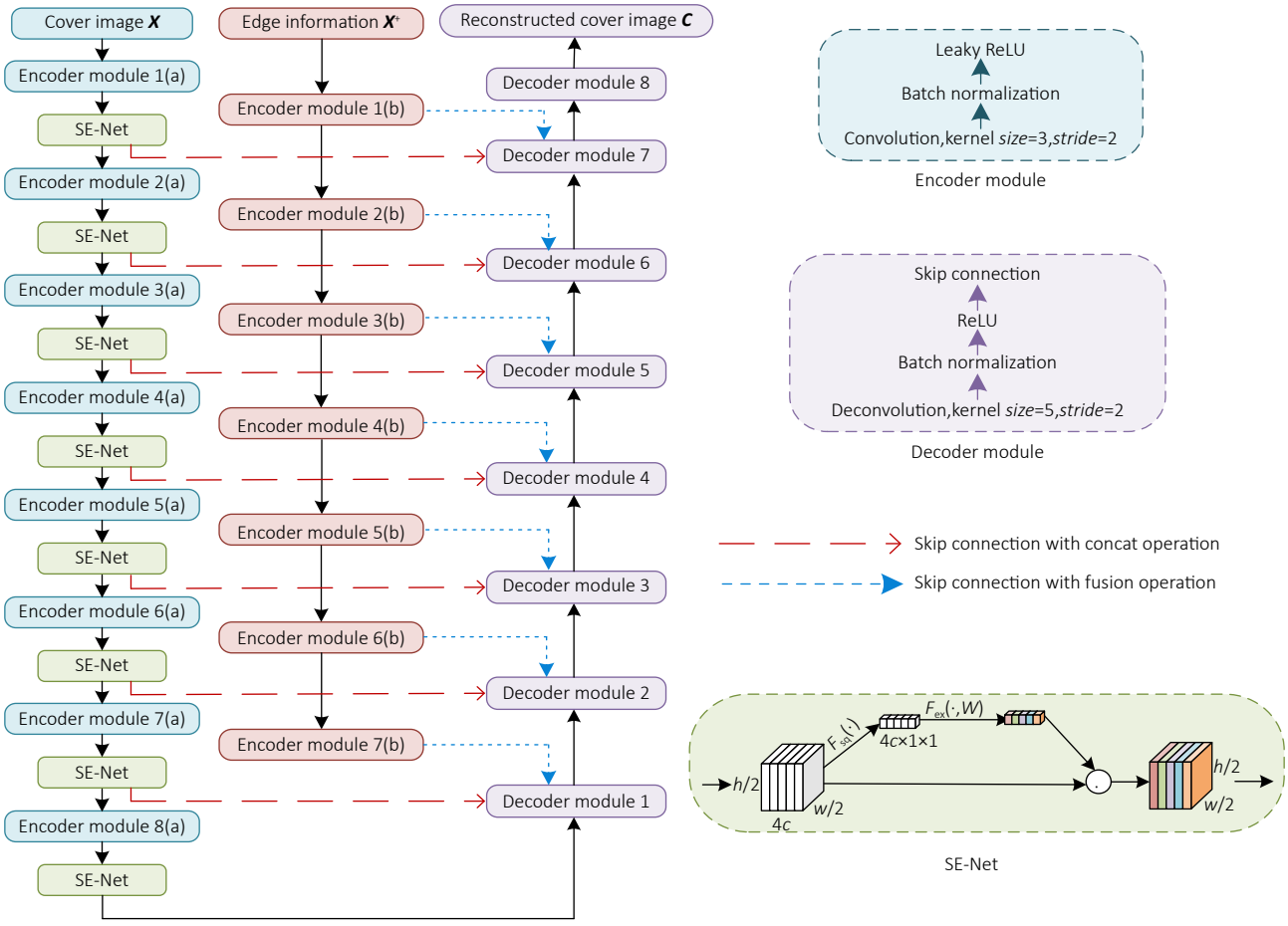


图 2 生成器 G 结构

Fig. 2 The structure of generator G

$$\boldsymbol{X}_{\mathrm{f}}^{8-i}=fusion\left(\boldsymbol{X}_{i}^{+}+\boldsymbol{X}_{16-i}\right),(i=1,2,\cdots,7), \tag{2}$$

式中：$\boldsymbol{X}_{i}^{+}$表示边缘信息编码特征；$\boldsymbol{X}_{16-i}$表示载体图像解码特征。然后，将载体图像编码特征与上述融合后的特征进行连接，进一步优化图像细节，可表示为

$$\boldsymbol{X}_{\mathrm{c}}^{8-i}=concat\left(\boldsymbol{X}_{i}+\boldsymbol{X}_{\mathrm{f}}^{8-i}\right),(i=1,2,\cdots,7), \tag{3}$$

式中：$\boldsymbol{X}_{i}$表示载体图像编码特征。

2.1.3 鉴别器

鉴别器 D_1 旨在通过区分原始载体图像 $\boldsymbol{C}$ 与重建后的载体图像 $\boldsymbol{X}$，促使生成器 G 改进其生成图像的质量，同样，鉴别器 D_2 旨在通过区分原始载体图像 $\boldsymbol{C}$ 与含密图像 $\boldsymbol{S}$，促使对抗性嵌入网络改进其生成图像的质量。本文的鉴别器 D_1 和 D_2 结构相同，如图 3 所示，由 7 个数据处理单元构成，每个数据处理单元包括卷积层，批处理归一化层，Leaky ReLU 激活函数层，最后将经过扁平化处理的特征输入到全连接层得到特征向量，经过

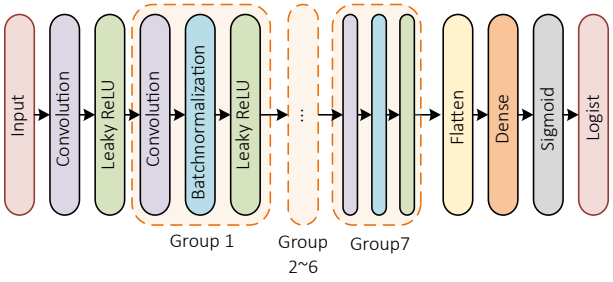


图 3 鉴别器结构

Fig. 3 The structure of the discriminator

Sigmoid 处理后，输出检测为真实图像的概率值。

生成器 G 的梯度是由鉴别器 D_1 产生的，鉴别器 D_1 的损失函数可表示为

$$L_{\mathrm{D}_1}=-\sum_{i=1}^{2}a_i'\ln(a_i),\ (i=1,2) \tag{4}$$

式中：a_1和a_2表示原始载体图像和优化后的载体图像经

过鉴别器 D_1 后的输出；a_1' 和 a_2' 表示二者的标签。生成器 G 和鉴别器 D_1 的目标是相互对抗的，因此生成器的对抗损失式中 $L_{G_1}=-L_{D_1}$。同时，为了保证优化后的载体图像的视觉质量，引入像素空间均方差损失 L_{M_1}，表示为

$$L_{M_1}=-\sum_{i=1}^{n_1}\frac{(b_i-b_i')^2}{n_1}, \tag{5}$$

式中：b_i 是原始载体图像像素点的值；b_i' 是对应位置的优化后的载体图像像素点的值；n_1 是图像像素点数量。因此载体图像优化网络的总损失可表示为

$$L_C=L_{G_1}+\lambda_1 L_{M_1}, \tag{6}$$

式中：λ_1 是权重系数。

2.2 对抗性嵌入网络

现有对抗性隐写方法的攻击目标分为两类：1) 逻辑层攻击：通过攻击目标隐写分析模型的决策边界欺骗该模型，如图 4(a) 所示，但容易导致对目标模型的过拟合，无法有效欺骗其他非目标隐写分析模型；2) 特征层攻击：攻击各类隐写分析模型做出决策时所依赖的共同关键特征，实现跨模型泛化对抗，如图 4(b) 所示。

本文对抗性嵌入网络在特征层进行攻击，核心思想是利用归因技术识别关键神经元，攻击各类隐写分析模型在做出决策时所依赖的共同关键特征，在保证图像质量的同时针对性优化嵌入策略，显著降低各类隐写分析模型的检测准确率，实现跨模型泛化对抗。此网络共包括三个阶段：神经元归因、关键特征破坏和对抗性嵌入。

具体而言，首先将目标层 y 中第 j 个神经元的输出 y_j 视为中间层特征，通过改进的积分梯度法[17] 计算其归因值 A_{y_j}，以量化该神经元对模型决策的重要性。

选择归因结果绝对值超过 T 的神经元，将相应的特征识别为关键特征，然后对这些神经元施加约束，形成归因损失 L_{att}：

$$L_{att}(x_{adv})=\sum_{|A_{y_j}|>T}\left(A_{y_j}(x_{adv})\right), \tag{7}$$

式中：x_{adv} 是指对抗样本。为了破坏关键特征，将神经元归因结果作为权重，并将其与相应的特征相乘来设计特征损失：

$$L_{fea}(x_{adv})=\sum\left(A_{y_j}\cdot y_j(x_{adv})\right), \tag{8}$$

式中：$y_j(x_{adv})$ 表示当输入 x_{adv} 时，模型目标层 y 第 j 神经元的输出结果。因此，神经元归因部分总损失为

$$L(x_{adv})=L_{att}(x_{adv})+\omega L_{fea}(x_{adv}), \tag{9}$$

式中：ω 是权重系数，为了保证模型既能正确地识别和处理关键特征，又能有效利用这些特征去干扰模型的判断，所以本文将 ω 设置为 1。

在神经元归因部分总损失 $L(x_{adv})$ 的基础上，本文方法引入与鉴别器 D_2 的对抗损失 L_{D_2} 与像素空间均方差损失 L_{M_2}，确保含密图像与原始载体图像在统计分布及像素级保持一致，进一步提高安全性以及不可感知性，与前文提到的 L_{D_1} 和 L_{M_1} 类似，可表示为

$$L_{D_2}=-\sum_{i=1}^{2}c_i'\ln(c_i), \tag{10}$$

$$L_{M_2}=-\sum_{i=1}^{n_2}\frac{(d_i-d_i')^2}{n_2}, \tag{11}$$

式中：c_1 和 c_2 表示原始载体图像和含密图像经过判别器 D_2 后的输出；c_1' 和 c_2' 表示二者对应的标签；d_i 是原始载体图像点的值；d_i' 是对应位置的含密图像像素点的值；n_2 是图像像素点数量。因此对抗性嵌入网络的总损失为

$$L_{adv}=\lambda_2 L(x_{adv})+\lambda_3 L_{G_2}+\lambda_4 L_{M_2}, \tag{12}$$

式中：$L_{G_2}=-L_{D_2}$，因为对抗性嵌入网络的目标与 D_2 相反；λ_2，λ_3，λ_4 是权重参数。因此将对抗性嵌入网络的优化问题表述为

$$\arg\min L_{adv},\ \text{s.t.}\|x_{adv}-x\|_\infty=1, \tag{13}$$

式中：$\|x_{adv}-x\|_\infty=1$，这是一个约束条件，表示两个样本在每个维度上的最大差异不能超过 1。

接下来，将载体图像划分为公共组 (占比 $1-\beta$) 和可调组 (占比 β)，两组互不相交，根据初始失真函数 (如 S-UNIWARD) 计算初始嵌入失真 $\{\rho_i^+,\rho_i^-\}$，其中 ρ_i^+ 和 ρ_i^- 表示在第 i 个图像像素上的+1 和−1 的失真，初始化参数

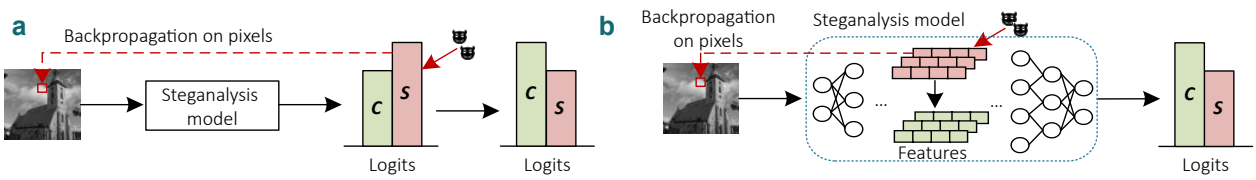


图 4 逻辑层攻击以及特征层攻击对比。(a) 逻辑层攻击；(b) 特征层攻击

Fig. 4 Comparison of logic layer attacks and feature layer attacks. (a) Logical layer attack; (b) Feature layer attack

$\beta=0$，基于初始失真将包含$1-\beta$的秘密消息位的msg_1嵌入到公共组中得到含密图像$\boldsymbol{Z}$。为求解式 (13) 的最小化问题，需要计算得到梯度图$\boldsymbol{G_Z}$，并根据$\boldsymbol{G_Z}$将初始失真调整为归因失真$\rho_{\text{attr}}=\{\rho^{+}_{\text{attr}},\rho^{-}_{\text{attr}}\}$：

$$\mathrm{G_Z}=\{g_{Z1},g_{Z2},\cdots,g_{Zn}\}=\frac{\partial L_{\text{adv}}(\boldsymbol{Z})}{\partial \boldsymbol{Z}}, \tag{14}$$

$$\begin{cases}\rho^{i}_{\text{attr+}}=\begin{cases}\rho^{+}_{i}/\alpha, & g_i<0\\ \rho^{+}_{i}, & g_i=0\\ \rho^{+}_{i}*\alpha, & g_i>0\end{cases}\\ \rho^{i}_{\text{attr-}}=\begin{cases}\rho^{-}_{i}/\alpha, & g_i>0\\ \rho^{-}_{i}, & g_i=0\\ \rho^{-}_{i}*\alpha, & g_i<0\end{cases}\end{cases}, \tag{15}$$

式中：α是大于 1 的比例因子，通过比例缩放辅助实现参数调整。使用调整后的失真向可调组嵌入剩余信息msg_2，重复上述过程直至生成的含密图像$\boldsymbol{Z}$能通过隐写分析模型检测。如果始终无法欺骗目标隐写分析模型，则将根据初始失真嵌入得到的图像作为最终的$\boldsymbol{S}$。在上述方法中，秘密信息的嵌入与提取基于 STC (syndrome-trellis codes) 编码实现，接收方可使用 STC 解码器无损恢复信息。信息恢复的可靠性由 STC 编码的数学结构保证，且失真约束严格控制修改幅度，确保含密图像结构完整，解码可行。

本文模型由载体图像优化网络以及对抗性嵌入网络两部分组成，因此总损失函数可表示为

$$L_{\text{total}}=k_1L_{\text{C}}+k_2L_{\text{adv}}, \tag{16}$$

式中：k_1和k_2是权重系数，具体设置及调整策略均在后续实验部分详细说明。

3 实验结果与分析

计算机配置 NVIDIA RTX 4090 GPU，显存 24 GB，内存 128 GB。选用 4 种主流的基于 CNN 的隐写分析模型 (CovNet、SRNet、Ye-Net、Zhu-Net) 依次作为目标隐写分析模型，选取特征集中度更高的较窄层作为目标层 (针对 CovNet、SRNet、Ye-Net、Zhu-Net，分别选择第 3 组层，第 3 类层、第 8 层、第 6 层作为目标层)，选用 S-UNIWARD 和 HILL 作为初始失真函数，采用 Adam 优化器，学习率设置为 0.0001。神经元归因部分设置样本步长 M=50 来计算归因结果，并以所有归因结果绝对值的中位数作为阈值 T 来选择关键神经元。

实验采用隐写领域广泛使用的数据集 BOSSBase ver.1.01 与 BOWS2 (各含 10000 张 512 pixel×512 pixel 灰度图像)，使用 Matlab 的 Resize 指令 (默认参数) 将其降维为 256 pixel×256 pixel 的图像，训练集和测试集分别为 16000 幅和 4000 幅，训练与测试阶段所使用的图像样本完全无重叠。除抗隐写分析能力实验外，其余实验将 SRNet 作为目标隐写分析模型，使用 S-UNIWARD 作为初始失真函数，嵌入率为 0.4 bpp。

3.1 评价指标

本文使用检测准确率 ACC 评估模型性能，通过漏检率和虚警率计算得到。漏检率P_{MD}指的是含密图像被错误地分类为载体图像的比率，虚警率P_{FA}指的是载体图像被错误地分类为含密图像的比率，计算公式分别如下：

$$P_{\text{MD}}=\Pr\{F(x)=0|x\in\boldsymbol{S}\}, \tag{17}$$

$$P_{\text{FA}}=\Pr\{F(x)=1|x\in\boldsymbol{C}\}, \tag{18}$$

式中：$\boldsymbol{S}$ 和 $\boldsymbol{C}$ 分别表示含密图像集和载体图像集，检测总错误率为

$$P_{\text{E}}=\frac{P_{\text{MD}}+P_{\text{FA}}}{2}. \tag{19}$$

从而得到检测准确率为

$$ACC=1-P_{\text{E}}, \tag{20}$$

同时，采用峰值信噪比 PSNR 和结构相似性 SSIM 两个经典客观评价指标，对生成图像质量进行评价。

3.2 损失函数权重设置实验

本文方法优化目标具阶段性差异。训练初期载体图像优化网络主导优化，因为高质量的载体图像是隐写的基础。后期对抗性嵌入网络主导优化，通过干扰隐写分析模型的判别逻辑不断调整嵌入失真，提升含密图像的隐蔽性。

经过初步实验验证，在训练前 100 轮中 k_1 由 0.8 变化至 0.2，同时 k_2 由 0.2 变化至 0.8 时，100 轮后固定权重直至收敛，此配置下后续各项指标最优，因此在此基础上进行其余参数权重设置实验，损失函数具体模型为 2.2 节所述的L_{total}。具体实验结果见表 1，当λ_1=0.2，$\lambda_2:\lambda_3:\lambda_4$=8:3:1 时，各项指标达到最佳。本文表格中粗体字均表示最优值。

3.3 消融实验

为了探究各个组件的必要性，本文设计了五组消融实验：1) 移除载体图像优化网络；2) 移除 SE-Net；3) 移除双流 U-Net 结构 (使用普通 U-Net，$\boldsymbol{C}$ 和 $\boldsymbol{C}^{+}$通过通

道拼接输入)；4) 移除L_{D_2}；5) 移除L_{M_2}。

消融实验结果见表 2，通过对比不同结构下面对非目标隐写分析模型的 ACC 以及含密图像的平均 PSNR 和 SSIM，表明本文方法的配置综合性能达到最优。这一配置有效地结合了高质量边缘信息、高效特征处理以及动态嵌入机制，不仅提升了含密图像质量，还显著降低了非目标隐写分析模型的检测准确率。

在网络中结合 SE-Net，能够在保持较低计算复杂度的同时，动态地聚焦于图像中信息丰富的区域，生成更适合信息隐藏的载体图像。本文设置消融实验来评估 SE-Net 添加在不同位置的影响，共测试了 5 种变体：1) 不添加 SE-Net；2) 在载体图像的编码路径添加，即本文所采用的配置；3) 在边缘信息图像的编码路径添加；4) 在载体图像和边缘信息图像的编码路径添加；5) 在解码路径添加。实验结果见表 3，当选择在载体图像的编码阶段加入 SE-Net 时，抗隐写分析能力以及含密图像质量都达到了最佳。

具体而言，未引入 SE-Net 时模型难以有效增强重要特征，导致模型能力有限；在载体图像编码路径中添加 SE-Net，能有效增强重要特征并抑制噪声，显著提升生成图像质量与隐写安全性；而在边缘信息图像的编码路径引入 SE-Net 时，由于边缘信息本身具备高度结构化且噪声较少，额外的注意力机制反而会干扰高频细节的表示，破坏边缘连续性，进而导致性能下降；在解码路径添加则会模糊特征层次，不利于隐写操作的控制。

3.4 生成图像质量及效率分析实验

为了评估含密图像的质量，本实验随机选取了 1000 张含密图像进行分析。图 5 展示了这些图像的 PSNR 和 SSIM 的分布情况，其中红色线段表示平均值。结果显示，含密图像的平均 PSNR 为 40.89，平均 SSIM 为 0.9783。此外，随机选取了 3 组载体图像与含密图像进行对比，并展示了它们对应的像素分布直方图，如图 6 所示，结果表明，含密图像与载体图像不仅在视觉上难以区分，其直方图分布也表现出极高的相似性。这些结果充分证明，本文提出的模型在实现秘密信息有效嵌入

表 1 损失函数权重设置实验结果

Table 1 Experimental results of loss function weight setting

λ_1	$\lambda_2:\lambda_3:\lambda_4$	Stego image quality		ACC of the non-target steganalysis model/%			
		Average PSNR/dB	Average SSIM	CovNet	Ye-Net	Zhu-Net	SRM
0.1	5:3:1	39.21	0.9591	82.36	75.17	83.33	69.31
	8:3:1	40.10	0.9634	79.33	74.37	81.22	66.89
	10:5:1	38.77	0.9487	83.34	76.93	82.66	68.53
0.2	5:3:1	40.11	0.9655	78.31	72.87	80.22	69.88
	8:3:1	**40.91**	**0.9776**	**77.37**	**70.37**	**78.16**	**64.19**
	10:5:1	39.03	0.9573	80.11	72.84	80.30	66.74
0.3	5:3:1	39.68	0.9675	79.64	74.96	83.03	69.83
	8:3:1	40.62	0.9724	78.22	72.30	80.94	67.55
	10:5:1	38.94	0.9611	80.31	75.99	84.67	69.11

表 2 消融实验结果

Table 2 Ablation experiment results

Model structure	Stego image quality		ACC of the non-target steganalysis model/%			
	Average PSNR/dB	Average SSIM	CovNet	Ye-Net	Zhu-Net	SRM
1)	34.79	0.9224	79.42	73.82	79.85	65.03
2)	38.99	0.9583	78.51	71.64	79.63	64.98
3)	35.64	0.9321	77.81	70.92	78.64	**63.90**
4)	38.37	0.9567	78.66	71.42	79.34	64.68
5)	36.73	0.9557	78.08	75.14	79.10	64.97
Ours	**40.91**	**0.9776**	**77.37**	**70.37**	**78.16**	64.19

表 3 SE-Net 添加位置实验结果

Table 3 Experimental results of adding location in SE-Net

SE-Net addition location	Stego image quality		ACC of the non-target steganalysis model/%			
	Average PSNR/dB	Average SSIM	CovNet	Ye-Net	Zhu-Net	SRM
1)	38.99	0.9583	78.51	71.64	79.63	64.98
2) Ours	**40.91**	**0.9776**	**77.37**	**70.37**	**78.16**	**64.19**
3)	37.18	0.9452	78.81	72.54	79.90	65.75
4)	39.64	0.9633	78.27	71.68	78.98	64.55
5)	38.75	0.9592	78.93	71.80	79.67	64.84

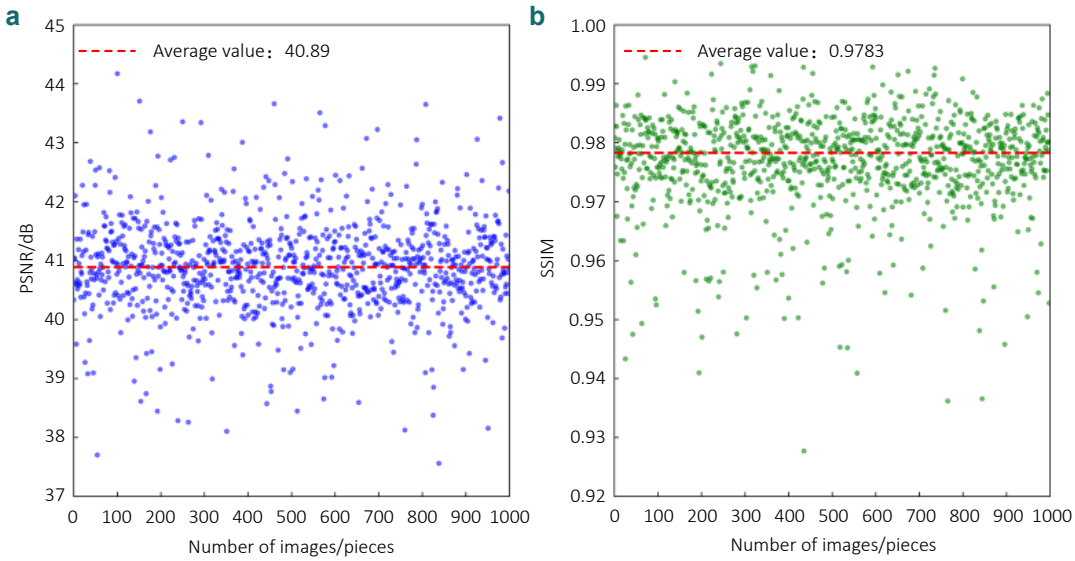


图 5 1000 张含密图像的 PSNR 和 SSIM 的分布情况。(a) PSNR 分布情况；(b) SSIM 分布情况

Fig. 5 The distribution of PSNR and SSIM of 1000 stego images. (a) Distribution of PSNR; (b) Distribution of SSIM

的同时能够保持较高的图像质量。该结果得益于双策略协同优化机制：被动防御模块通过注意力机制融合图像与边缘特征，有效保持语义完整性；主动攻击模块在损失函数约束下，干扰隐写分析模型判断的同时保持较高图像质量。两模块动态协同，在保障视觉保真度的同时显著提升跨模型泛化能力。

为对比生成效率，本文测试了单张含密图像的平均生成时间，实验结果见表 4。结果表明，本文方法由于其复杂的网络结构，在生成效率上略低于其他方案，但综合来看，复杂网络结构虽增加了计算负担，但显著提升了载体图像的质量，说明本文方法在可接受的效率成本下，为提升安全性与图像质量提供了有效保障。

3.5 跨数据集泛化能力分析

为了验证所提方法的跨数据集泛化能力，所有隐写方法 (包括对比方法 CR-AIS、Backpack、ADV-EMB、Natias 及本文方法) 均在 BOSSBase ver.1.01 数据集上训练，测试在 COCO 和 ImageNet 数据集 (转换为 256 pixel×256 pixel 灰度图像) 上进行。使用基于深度学习的隐写分析模型 CovNet、Ye-Net、Zhu-Net 以及基于特征统计的隐写分析模型 SRM 进行隐写分析，将这四种隐写分析模型检测的平均 ACC 作为性能评估标准 (见表 5)。实验结果表明，本文方法的平均 ACC 最低，具有良好的跨数据集泛化能力。这是由于本文所提的双策略模型通过生成结构一致的优化载体并扰动共性判别特征，有效减少了模型对特定训练集统计特征的依赖，从而在不同数据集上均表现出稳定且领先的抗检测能力。

3.6 抗隐写分析能力实验

抗隐写分析能力是评估图像隐写术安全性的关键指

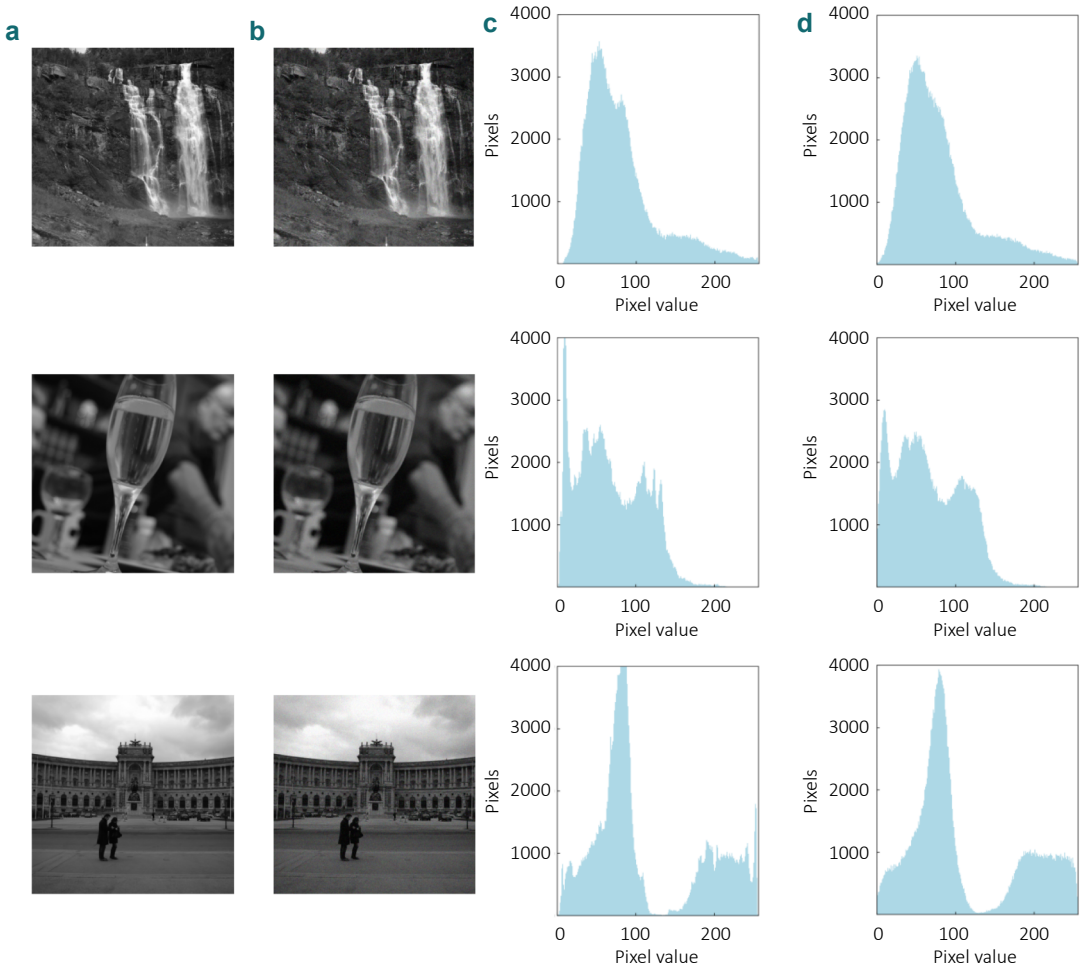


图 6 载体图像，含密图像及其直方图对比。(a) 载体图像；(b) 含密图像；(c) 载体图像直方图；(d) 含密图像直方图

Fig. 6 Comparison of cover images, stego images and their histograms. (a) Cover images; (b) Stego images; (c) The histograms of the cover images; (d) The histograms of the stego images

表 4 生成图像质量及效率分析实验结果

Table 4 Generate the experimental results of image quality and efficiency analysis

Method	Average PSNR/dB	Average SSIM	Average generation time/s
ADV-EMB	36.88	0.9377	**1.54**
Backpack	37.64	0.9460	1.76
CR-AIS	37.59	0.9521	2.17
Natias	38.09	0.9501	2.57
Ours	**40.89**	**0.9783**	3.62

表 5 跨数据集泛化能力分析实验结果

Table 5 Experimental results of cross-dataset generalization ability analysis

Method	Average ACC/%
ADV-EMB	79.67
Backpack	77.06
CR-AIS	76.64
Natias	74.91
Ours	**72.61**

标。为验证本文方法抵御非目标隐写分析模型检测的能力，我们严格遵循模型无关性评估原则开展实验：在每一轮训练中，仅使用一个目标隐写分析模型 (依次为 CovNet、SRNet、Ye-Net 或 Zhu-Net) 进行对抗训练并生成隐写含密图像；在测试阶段，则使用其它 3 种隐写分析模型及 SRM 进行检测，同时训练与测试阶段所使用的图像样本完全无重叠。

将本文方法与 CR-AIS 和 Natias 两个隐写性能优秀的对抗性隐写方法进行对比，具体实验结果见表 6~表 9 所示，综合实验结果如图 7。其中，图 7(a-d) 是使用 HILL 作为初始失真函数的结果，图 7(e-h) 是使用 S-UNIWARD 作为初始失真函数的结果，例如图 7(a) 表示

表 6 目标隐写分析模型为 CovNet 时的 ACC 对比

Table 6 Comparison of ACC when the target steganalysis model is CovNet

Distortion function	Embedding rate/bpp	SRNet-ACC/%			Ye-Net-ACC/%			Zhu-Net-ACC/%			SRM-ACC/%		
		CR-AIS	Natias	Ours	CR-AIS	Natias	Ours	CR-AIS	Natias	Ours	CR-AIS	Natias	Ours
HILL	0.1	53.77	**52.30**	52.63	53.17	52.46	**49.37**	57.31	56.33	**55.63**	51.83	50.75	**49.77**
	0.2	58.37	56.37	**53.71**	55.13	53.70	**51.33**	59.41	57.71	**55.77**	54.32	52.85	**50.93**
	0.3	62.32	59.67	**57.75**	58.77	56.41	**53.64**	66.03	63.26	**61.32**	57.09	54.55	**52.31**
	0.4	66.37	62.66	**60.17**	63.82	60.51	**57.03**	73.13	69.77	**67.87**	60.98	57.26	**56.31**
S-UNIWARD	0.1	58.13	**57.77**	57.83	57.32	56.48	**55.94**	62.37	61.66	**60.75**	54.72	54.30	**53.03**
	0.2	64.32	61.57	**60.13**	64.13	61.33	**59.12**	69.95	67.57	**64.34**	60.62	58.34	**56.31**
	0.3	69.33	66.37	**64.74**	68.77	65.99	**63.37**	81.15	78.01	**74.09**	65.45	62.32	**60.17**
	0.4	75.12	71.92	**69.94**	73.97	70.54	**67.25**	81.54	78.04	**76.37**	67.73	64.66	**62.31**

表 7 目标隐写分析模型为 SRNet 时的 ACC 对比

Table 7 Comparison of ACC when the target steganalysis model is SRNet

Distortion function	Embedding rate/bpp	CovNet-ACC/%			Ye-Net-ACC/%			Zhu-Net-ACC/%			SRM-ACC/%		
		CR-AIS	Natias	Ours	CR-AIS	Natias	Ours	CR-AIS	Natias	Ours	CR-AIS	Natias	Ours
HILL	0.1	57.30	**57.01**	57.02	54.11	53.77	**52.17**	58.31	**57.84**	57.93	53.21	52.74	**51.34**
	0.2	64.13	63.61	**63.26**	56.13	55.54	**54.14**	65.17	64.49	**61.11**	55.93	55.37	**52.16**
	0.3	68.84	67.76	**66.71**	60.13	59.12	**56.48**	70.91	69.93	**67.74**	58.73	57.88	**55.17**
	0.4	70.81	68.11	**67.09**	65.09	63.85	**62.13**	72.97	71.09	**68.19**	61.10	58.62	**57.47**
S-UNIWARD	0.1	61.93	61.56	**60.31**	58.31	57.95	**57.74**	65.97	65.84	**62.17**	54.17	**54.01**	54.72
	0.2	70.37	69.57	**68.66**	65.33	65.47	**61.67**	71.77	69.92	**68.74**	60.44	59.68	**57.14**
	0.3	78.16	77.03	**74.94**	72.57	71.37	**69.13**	79.13	77.16	**76.99**	65.31	65.06	**61.77**
	0.4	81.83	79.68	**77.37**	75.63	74.01	**70.37**	81.21	79.89	**78.16**	67.99	65.37	**64.19**

表 8 目标隐写分析模型为 Ye-Net 时的 ACC 对比

Table 8 Comparison of ACC when the target steganalysis model is Ye-Net

Distortion function	Embedding rate/bpp	SRNet-ACC/%			CovNet-ACC/%			Zhu-Net-ACC/%			SRM-ACC/%		
		CR-AIS	Natias	Ours	CR-AIS	Natias	Ours	CR-AIS	Natias	Ours	CR-AIS	Natias	Ours
HILL	0.1	52.85	51.94	**50.73**	56.13	55.01	**54.15**	57.99	57.51	**55.83**	51.45	50.01	**49.37**
	0.2	57.73	56.02	**54.17**	63.25	61.99	**60.79**	65.46	64.12	**60.94**	53.34	51.77	**49.69**
	0.3	63.02	61.1	**59.13**	68.70	66.82	**64.88**	69.33	67.71	**65.64**	56.37	54.55	**52.22**
	0.4	65.57	63.31	**61.79**	69.76	67.22	**65.37**	74.69	72.47	**69.72**	59.76	57.43	**55.10**
S-UNIWARD	0.1	57.34	**56.65**	56.77	61.69	61.21	**58.83**	63.97	63.76	**61.93**	55.53	**54.24**	54.97
	0.2	63.17	61.73	**59.34**	68.17	66.88	**64.24**	69.72	68.34	**66.69**	60.33	58.57	**56.16**
	0.3	69.73	66.51	**63.32**	78.92	75.99	**72.56**	79.17	76.37	**74.66**	63.64	60.54	**58.17**
	0.4	74.39	70.83	**67.54**	79.25	76.02	**73.47**	82.93	78.94	**78.10**	66.32	62.83	**60.79**

表 9 目标隐写分析模型为 Zhu-Net 时的 ACC 对比

Table 9 Comparison of ACC when the target steganalysis model is Zhu-Net

Distortion function	Embedding rate/bpp	SRNet-ACC/%			CovNet-ACC/%			Ye-Net-ACC/%			SRM-ACC/%		
		CR-AIS	Natias	Ours	CR-AIS	Natias	Ours	CR-AIS	Natias	Ours	CR-AIS	Natias	Ours
HILL	0.1	53.45	52.34	**49.99**	56.97	**55.69**	56.16	50.48	49.93	**48.96**	52.64	51.30	**48.65**
	0.2	58.01	56.68	**54.04**	63.55	61.82	**59.84**	56.31	54.24	**52.45**	54.67	53.19	**50.66**
	0.3	62.06	59.35	**58.33**	68.22	66.56	**64.59**	59.48	57.88	**54.58**	58.31	56.32	**53.11**
	0.4	65.97	63.54	**60.21**	69.21	66.87	**64.91**	64.85	61.35	**60.59**	60.68	57.67	**56.34**
S-UNIWARD	0.1	57.89	57.35	**55.97**	62.04	61.57	**59.13**	57.28	**56.37**	56.61	54.96	54.76	**53.34**
	0.2	64.97	63.21	**60.18**	68.12	66.54	**65.27**	62.58	60.84	**57.49**	60.26	58.35	**56.37**
	0.3	68.87	66.46	**63.54**	77.64	75.38	**72.64**	66.68	64.35	**63.21**	64.38	61.31	**59.34**
	0.4	74.97	71.54	**68.07**	78.64	75.65	**73.27**	72.68	69.31	**67.33**	66.37	63.20	**61.57**

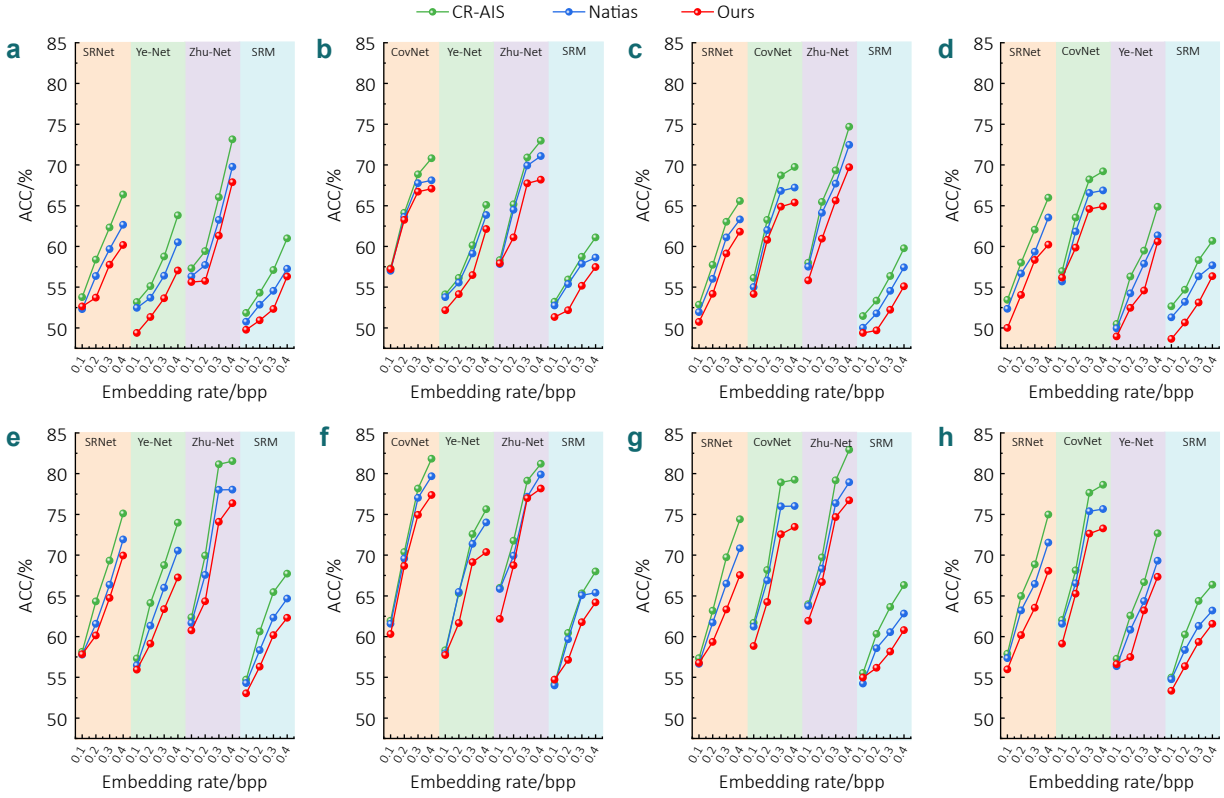


图 7 抗隐写分析能力实验综合结果。(a) HILL-CovNet；(b) HILL-SRNet；(c) HILL-Ye-Net；(d) HILL-Zhu-Net；(e) S-UNIWARD-CovNet；(f) S-UNIWARD-SRNet；(g) S-UNIWARD-Ye-Net；(h) S-UNIWARD-Zhu-Net

Fig. 7 Comprehensive results of anti steganalysis ability experiment. (a) HILL-CovNet; (b) HILL-SRNet; (c) HILL-Ye-Net; (d) HILL-Zhu-Net; (e) S-UNIWARD-CovNet; (f) S-UNIWARD-SRNet; (g) S-UNIWARD-Ye-Net; (h) S-UNIWARD-Zhu-Net

的是初始失真函数为 HILL，目标隐写分析模型为 CovNet 时，使用 SRNet、Ye-Net、Zhu-Net 检测的 ACC。由图 7 可知，与 CR-AIS 和 Natias相比，本文方法在嵌入率为 0.1 bpp 时，性能提高并不明显，但随着嵌入率的升高，本文方法的检测准确率有了显著的降低。并且从整体上看，本文方法在大多数情况下都要优于其它两种方法，整体性能排序为：本文方法 > Natias > CR-AIS，对所有情况下的 ACC 进行平均，本文方法相比于 CR-AIS 的平均降低幅度为 3.69%，相比于 Natias 的平均降低幅度为 1.91%，说明本文方法具有较强的抵御非目标隐写分析模型的能力，具有良好的跨模型泛化能力。

与 Natias 相比，在某些低嵌入率场景下本文方法的

性能提升幅度有限，这主要由于 Natias 本身已具备较强的安全性基础，其在低负载场景下的对抗能力已接近当前技术条件下的性能平台。然而，随着嵌入率的提升，本文方法所采用的双策略优化机制，即被动防御生成的高质量载体与主动攻击对共性判别特征的针对性扰动，展现出更显著的协同优势，从而在中高嵌入率下实现了更具实际意义的性能提升。综合各项性能比较表明，本文方法在维持高图像质量的同时，更侧重于解决实际应用中更高负载下的安全需求，其综合性能与泛化能力优于现有方法。

4 结　论

本文提出融合被动防御与主动攻击的双策略协同优化框架，通过端到端协同训练实现载体图像优化与对抗性嵌入动态适配：被动防御模块结合双流 U-Net、CMPC 与 SE-Net 生成高质量载体以降低嵌入引发的统计异常，主动攻击模块通过神经元归因技术定位跨模型共性特征并施加动态扰动以降低检测准确率。实验表明该方法生成的含密图像具有良好的跨模型泛化能力且图像质量较高，实现了隐写安全性与视觉质量协同提升。但因网络结构复杂导致图像生成效率较低，未来将重点研究架构轻量化与硬件加速技术以提高运行效率，推动隐写性能全面提升。此外，扩散模型 (diffusion model, DM) 在图像质量、细节生成与语义可控性方面展现出显著优势，探索其在隐写载体生成中的应用，以进一步提升隐写系统的安全性与适应性，也将是未来重要的研究方向。

作者贡献

郭伟：指导论文选题、论文修改；任哲聪：论文写作、算法实现与实验；金海波、曲海成：论文写作指导；林畅：研究资料的收集、分析。

利益冲突

所有作者声明无利益冲突

参考文献

1. Zhang W M, Wang H X, Li B, et al. Overview of steganography on multimedia[J]. *J Image Graph*, 2022, **27**(6): 1918−1943.
   张卫明, 王宏霞, 李斌, 等. 多媒体隐写研究进展[J]. 中国图象图形学报, 2022, **27**(6): 1918−1943.
2. Ye X Y, Guo W F, Zeng M S, et al. Image steganography detection based on multilayer perceptual convolution and channel weighting[J]. *J Electron Inf Technol*, 2022, **44**(8): 2949−2956.
   叶学义, 郭文风, 曾懋胜, 等. 基于多层感知卷积和通道加权的图像隐写检测[J]. 电子与信息学报, 2022, **44**(8): 2949−2956.
3. Fridrich J, Kodovsky J. Rich models for steganalysis of digital images[J]. *IEEE Trans Inf Forensics Secur*, 2012, **7**(3): 868−882.
4. Holub V, Fridrich J. Low-complexity features for JPEG steganalysis using undecimated DCT[J]. *IEEE Trans Inf Forensics Secur*, 2015, **10**(2): 219−228.
5. Deng X Q, Chen B L, Luo W Q, et al. Fast and effective global covariance pooling network for image steganalysis[C]// *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, Paris, France, 2019: 230–234. https://doi.org/10.1145/3335203.3335739.
6. Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images[J]. *IEEE Trans Inf Forensics Secur*, 2019, **14**(5): 1181−1193.
7. Ye J, Ni J Q, Yi Y. Deep learning hierarchical representations for image steganalysis[J]. *IEEE Trans Inf Forensics Secur*, 2017, **12**(11): 2545−2557.
8. Zhang R, Zhu F, Liu J Y, et al. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis[J]. *IEEE Trans Inf Forensics Secur*, 2020, **15**: 1138−1150.
9. Filler T, Judas J, Fridrich J. Minimizing additive distortion in steganography using syndrome-trellis codes[J]. *IEEE Trans Inf Forensics Secur*, 2011, **6**(3): 920−935.
10. Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain[J]. *EURASIP J Inf Secur*, 2014, **2014**(1): 1.
11. Li B, Wang M, Huang J W, et al. A new cost function for spatial Image steganography[C]//*Proceedings of 2014 IEEE International Conference on Image Processing*, Paris, France, 2014: 4206–4210. https://doi.org/10.1109/ICIP.2014.7025854.
12. Ma B, Han Z W, Xu J, et al. Generative multiple adversarial steganography algorithm based on U-Net structure[J]. *J Software*, 2023, **34**(7): 3385−3407.
    马宾, 韩作伟, 徐健, 等. 基于 U-Net 结构的生成式多重对抗隐写算法[J]. 软件学报, 2023, **34**(7): 3385−3407.
13. Ma B, Li K, Xu J, et al. High-security image steganography with the combination of multiple competition and channel attention[J]. *J Image Graph*, 2024, **29**(2): 355−368.
    马宾, 李坤, 徐健, 等. 联合多重对抗与通道注意力的高安全性图像隐写[J]. 中国图象图形学报, 2024, **29**(2): 355−368.
14. Tang W X, Li B, Tan S Q, et al. CNN-based adversarial embedding for image steganography[J]. *IEEE Trans Inf Forensics Secur*, 2019, **14**(8): 2074−2087.
15. Bernard S, Bas P, Klein J, et al. Backpack: a backpropagable adversarial embedding scheme[J]. *IEEE Trans Inf Forensics Secur*, 2022, **17**: 3539−3554.
16. Chen Y J, Wang H X, Li W J, et al. Cost reassignment for improving security of adaptive steganography using an artificial immune system[J]. *IEEE Signal Process Lett*, 2022, **29**: 1564−1568.
17. Fan Z X, Chen K J, Zeng K, et al. Natias: neuron attribution-based transferable image adversarial steganography[J]. *IEEE Trans Inf Forensics Secur*, 2024, **19**: 6636−6649.
18. Wang D W, Yang G B, Chen J Y, et al. GAN-based adaptive cost learning for enhanced image steganography security[J]. *Expert Syst Appl*, 2024, **249**: 123471.
19. Shi M H, Li Q, Zhao X Q, et al. A new approach for image edge detection based on conformal phase[J]. *J Electron Inf Technol*, 2015,

37(11): 2594−2600.
石美红, 李青, 赵雪青, 等. 一种基于保角相位的图像边缘检测新方法[J]. 电子与信息学报, 2015, **37**(11): 2594−2600.
20. Lv J, Duan X L, Chen X. Edge feature and detail-aware network integrated YOLOv8s algorithm for hip joint keypoint detection[J]. *Opto-Electron Eng*, 2025, **52**(3): 240281.
吕佳, 段训禄, 陈欣. 融合边缘特征与细节感知网络的 YOLOv8s 髋关节关键点检测算法[J]. 光电工程, 2025, **52**(3): 240281.

作者简介



郭伟 (1970-)，女，硕士，副教授，硕士生导师，主要研究方向为图像与视觉信息计算、智能数据处理等。
E-mail：guow9966@163.com



【通信作者】任哲聪 (2001-)，女，硕士研究生，主要研究方向为图像隐写、计算机视觉等。
E-mail：1396933655@qq.com

Dual strategy image steganography model for defense and attack

Guo Wei, Ren Zhecong*, Jin Haibo, Qu Haicheng, Lin Chang

College of Software, Liaoning Technical University, Huludao, Lining 125105, China

Abstract

Objective As a core technology in the field of information security, image steganography holds irreplaceable strategic value in areas such as covert communication and digital copyright protection. Current adversarial image steganography methods commonly suffer from insufficient cross-model generalization capability and low quality of stego images. This is primarily due to the over-reliance of these methods on the decision boundaries of the target steganalysis model, which results in adversarial perturbations being highly coupled with the decision boundaries of the target model, making it difficult to generalize effectively to unknown models. Moreover, when generating adversarial perturbations, the fusion of secret information with the texture of the cover image is often insufficiently natural, especially in low-texture regions where statistical anomalies are prone to emerge, thereby compromising the visual quality and statistical naturalness of the stego images.

Methods To address the aforementioned issues, this paper proposes a dual-strategy image steganography model that integrates passive defense and active attack. Through the synergistic end-to-end optimization training of cover image enhancement and adversarial embedding, the model significantly improves cross-model generalization capability and the quality of stego images. At the passive defense level, cover image optimization is performed: A dual-stream U-Net structure is introduced into the generator of the generative adversarial network (GAN). This dual-branch network simultaneously processes the original image and edge information derived from conformal monogenic phase congruency (CMPC). Combined with the SE-Net attention mechanism to optimize feature distribution, texture enhancement and detail preservation are achieved during adversarial training. By jointly optimizing the discriminator loss and the pixel-space mean squared error loss, an optimized cover image with high visual quality and statistical naturalness is generated, constructing a secure and robust cover environment for secret information embedding. At the active attack level, adversarial embedding is conducted: integrated gradients are utilized for neuron attribution to locate the critical discriminative features commonly relied upon by various models, followed by dynamic perturbation. The neuron attribution loss, together with the discriminator loss and pixel-space mean squared error loss, jointly optimizes the embedding strategy. This ensures the indistinguishability between the stego image and the original cover image while significantly reducing the detection accuracy of various steganalysis models. Considering the phased differences in the optimization objectives of the proposed model, the cover image optimization network dominates the training in the initial stage, as a high-quality cover image is the foundation of steganography. In the later stage, the adversarial embedding network takes the lead, continuously adjusting the embedding distortion by interfering with the discriminative logic of steganalysis models to enhance the concealment of the stego image. Therefore, a linear dynamic adjustment strategy is adopted, allowing the weights of the two components to transition smoothly and uniformly during the first 100 training epochs. This matches the requirement for phased dominance switching between the two

Foundation item: National Natural Science Foundation of China (62173171), General Program of National Natural Science Foundation of China (42071351), and Fundamental Research Project of Liaoning Provincial Department of Education (JYTMS20230804)
*Correspondence: Ren Z C, E-mail: 1396933655@qq.com

networks, driving the model's performance to advance progressively by stages and achieving a global optimization of concealment, usability, and security.

Results and Discussions In the loss function weight configuration experiment, considering the phased differences in the model's optimization objectives, a linear dynamic adjustment strategy was adopted to change the weight coefficients k_1 and k_2, while systematically investigating the effects of different configurations of λ_1 and the ratio λ_2:λ_3:λ_4. By comparing the quality of stego images and the detection accuracy of non-targeted steganalysis models, the optimal weight configuration was determined as: k_1 decreasing from 0.8 to 0.2, k_2 increasing from 0.2 to 0.8, λ_1=0.2, and λ_2:λ_3:λ_4=8:3:1. The ablation study, by comparing the ACC against non-targeted steganalysis models as well as the average PSNR and SSIM of stego-images under different architectures, demonstrates that the configuration of the proposed method achieves the best overall performance. Furthermore, comparing the effects of adding the SE-Net attention mechanism at different positions within the dual-stream U-Net network showed that adding it only during the encoding stage yielded the best results. In the experiment analyzing generated image quality and efficiency, the average PSNR and SSIM obtained from 1000 randomly selected stego images were 40.89 dB and 0.9783, respectively. Three sets of cover images and their corresponding stego images were randomly selected for comparison. The results indicated that the two types of images are not only visually indistinguishable but also exhibit highly similar histogram distributions. To compare generation efficiency, the average generation time for a single stego image was tested. Due to its complex network architecture, the proposed method is slightly less efficient in generation speed compared to other schemes, but it significantly enhances the quality of cover images. This demonstrates that the proposed method provides an effective guarantee for improving security and image quality at an acceptable efficiency cost. The cross-dataset generalization capability analysis experiment showed that the proposed method achieved the lowest average ACC, indicating good cross-dataset generalization ability. This is because the proposed dual-strategy model generates optimized covers with consistent structural properties and perturbs common discriminative features, effectively reducing the model's dependence on statistical features specific to a particular training dataset. Consequently, it exhibits stable and good anti-detection performance across different datasets. In the experiment on anti-steganalysis capability, a comparison was made with two state-of-the-art adversarial embedding methods (CR-AIS and Natias). The experimental results showed that, in terms of resilience against non-targeted steganalysis models, the proposed method achieved average detection accuracy (ACC) reduction margins of 3.69% and 1.91% compared to CR-AIS and Natias, respectively. This indicates that the proposed method possesses strong resilience against non-targeted steganalysis models and exhibits good cross-model generalization capability.

Conclusions Through the deep integration of passive defense (dual-stream U-Net-guided cover optimization) and active attack (neuron attribution-driven adversarial embedding), the proposed image steganography model significantly reduces the detection accuracy of various steganalysis models, enhances cross-model generalization capability, and maintains high image quality, thereby achieving synergistic improvement in steganographic security and visual fidelity.

Keywords: image steganography; generative adversarial network (GAN); dual-stream U-Net; neuron attribution; cross-model generalization

Guo W, Ren Z C, Jin H B et al. Dual strategy image steganography model for defense and attack[J]. *Opto-Electron Eng*, 2026, **53**(2): 250238; DOI: 10.12086/oee.2026.250238

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文