# 基于双重加密的自监督判别机制可逆图像隐写

doi:10.20269/j.cnki.1674-7100.2026.2012

王晓红　贺心洁　马春运

上海理工大学
出版学院
上海　200093

**摘　要**：为提升图像隐写的安全性、视觉一致性及抗隐写分析能力，提出了可逆隐写网络DISG-Net。模型通过创新设计的QR码双重加密方法保障秘密信息安全，并利用16块基于小波变换的可逆网络实现高保真嵌入与可逆恢复，同时引入BYOL自监督判别器约束特征分布，使生成结果自然且难以检测。通过重构损失、引导损失、对比损失及哈希损失等多目标损失，实现视觉一致性与秘密信息精确恢复。实验结果表明，DISG-Net在图像质量和信息安全性上优于现有方法，可为印刷与包装提供高保真、防篡改的安全信息嵌入方案，提升防伪与信息保护能力。

**关键词**：图像隐写；可逆神经网络；双重加密；自监督判别器

**中图分类号**：TB489；TP309.7　　**文献标志码**：A

**文章编号**：1674-7100(2026)02-0094-09

**引文格式**：王晓红，贺心洁，马春运 . 基于双重加密的自监督判别机制可逆图像隐写 [J]. 包装学报，2026，18(2)：94-102.

## 1　研究背景

图像隐写术以不易察觉的方式，将秘密图像嵌入封面图像中，仅经授权的接收者能恢复秘密内容，而普通观察者则无法察觉秘密内容[1]。作为隐蔽通信的重要手段，图像隐写技术在信息安全、数据传输及版权保护等方面具有广泛应用价值[2]，尤其在包装工程等需要结合产品外观与信息防伪的应用场景中，能够为包装提供更高等级的安全性与防篡改能力[3]。

近年来，深度学习推动了图像隐写的快速发展。2017年，S. Baluja[3]提出首个基于GAN的端到端图像隐写框架HiDDeN，利用生成器与判别器的博弈机制有效提升了隐写图像的质量与隐蔽性。在HiDDeN的基础上，SteganoGAN等方法[4-6]被相继提出，其借助更深层的卷积结构和改进的损失函数，增强了生成－判别博弈的稳定性与对抗性，同时在不降低图像质量的前提下显著提高了信息嵌入容量。但传统SteganoGAN方法存在训练不稳定、依赖外部监督标签等问题，限制了其在复杂无标签数据环境下的应用[7]。U-Net结构因其稳定的训练过程也被广泛应用于图像隐写任务中。Wu H.等[8]提出SteganoUNet，采用U-Net编码器－解码器架构提升隐藏质量与秘密内容恢复精度。Zhang Y.、Chen L.等[9-10]在SteganoUNet基础上引入通道注意力机制，进一步增强了模型对关键信息的表征能力，从而提升了隐写的不可见性。然而U-Net因缺乏对抗反馈，应对隐写分析器时安全性不足。同时，U-Net高频建模不足与缺乏对抗反馈，导致秘密信息恢复精度下降[11]。近年来，基于可逆神经网络（invertible neural networks，INN）的隐写方法因结构对称、无信息损失和可逆性等特性，不仅保障了秘密信息的高精度恢复，还在提升隐写不可见性与安全性方面展现出独特优势。

**收稿日期**：2025-08-29
**基金项目**：国家新闻出版署智能与绿色柔版印刷重点实验室招标课题（ZBKT202301）
**作者简介**：王晓红，女，教授，主要从事色彩学与色彩应用、印刷质量检测与控制、数字印刷技术研究，
E-mail：wang_keyan@163.com

HiNet[12] 将 INN 与小波变换结合来提升隐藏质量与恢复精度；RIIS[13] 利用多尺度可逆模块来增强容量；IIS[14] 实现高分辨率图像精准隐写。此类方法缺乏 GAN 式对抗反馈，仅依赖重建误差优化，难以对齐统计特征与视觉感知，因而隐蔽性和抗分析能力受限 [15-16]。

图像加密在图像隐写中既能提升信息安全性，又能增强隐蔽性。王勇智 [17] 提出了以密钥控制嵌入过程的“先密后藏”隐写范式。最新模型 CryptoStegoNet[18] 通过二维码－全息变换与对抗训练，实现了高容量、高保真且抗检测的隐写。但其全息加密方法计算复杂度较高，导致嵌入与提取耗时，限制了在实时或资源受限场景中的应用。

为克服 INN 隐蔽性不足及加密计算复杂、效率受限的问题，本研究将 GAN 判别机制引入 INN 结构，设计自监督判别器，借助对比学习在无标签的情形下对隐写图像质量加以约束，从而生成更为自然且难以辨别的隐写图像。在此基础上，提出“加密－隐写”协同框架 DISG-Net（dual-encryption and self-supervised adversarial guided invertible network ）。先将秘密图像转换为噪声图与密钥，并通过二维码（QR）结构化编码，再将其隐写于封面图像，以实现强加密与高隐蔽性，有效防范泄露与篡改。本模型适用于高安全通信场景。

## 2 网络模型设计

### 2.1 基本原理

DISG-Net 模型（见图 1）由 QR 安全加密、可逆隐写网络和自监督判别器组成。首先，生成隐写图像，秘密图像（secret）经过加密扰动处理后，被编码为 QR，并作为嵌入信息；随后，封面图像（cover）、RQ 和噪声（noise）共同输入 16 个可逆隐写模块，并结合频域变换（DWT/IWT），实现封面图像与嵌入信息的融合；最后，自监督判别器通过最小化投影特征距离，引导模型生成更具隐蔽性的隐写图像。反之则恢复秘密图像，接收端提取出的二维码（QR′）经解密即可得到恢复后的原始图像（recovery）。

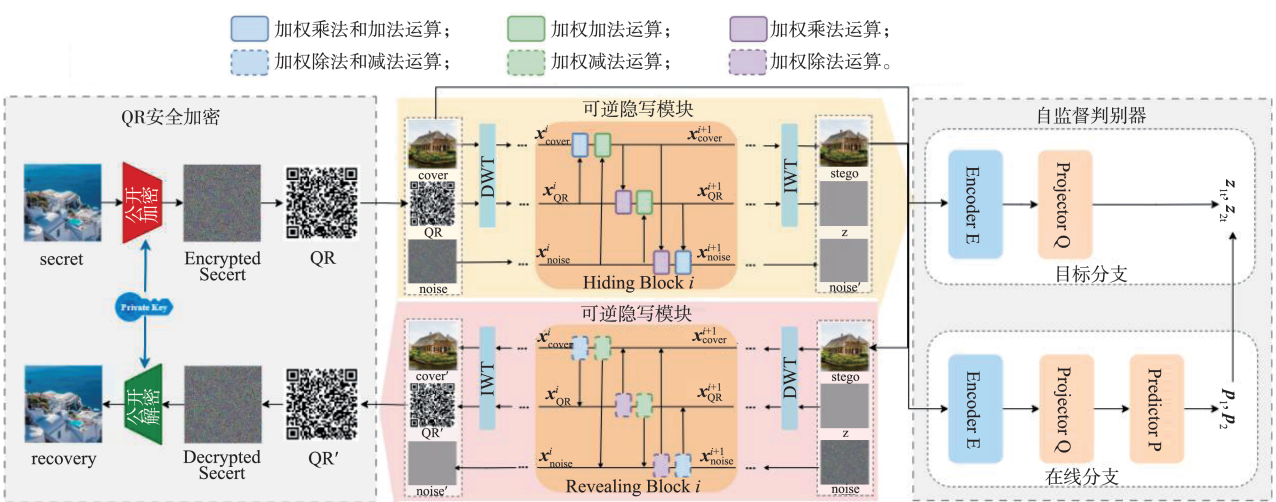



**图 1 DISG-Net 网络结构**

**Fig. 1 DISG-Net network architecture**

### 2.2 QR 安全加密

QR 安全加密采用双重加密机制，包含公开加密与公开解密两个子模块，并以 QR 为载体来增强信息的保密性与抗攻击能力，实现信息的安全传输。公开加密与解密机制能确保信息在传输过程中仅能被合法接收者恢复，防止信息泄露和篡改，同时借助 QR 的高容量与容错性，既提升了传输效率，也增强了 DISG-Net 的安全性与完整性。

在图像加密过程中，使用逻辑映射生成的序列 *seq* 对秘密图像的像素（$\boldsymbol{S}=\{S_1, S_2,\cdots, S_n\}$）进行混排，生成加密后的图像 $\boldsymbol{S}^*$，即

$$\boldsymbol{S}^*=\boldsymbol{S}[(seq)]。\tag{1}$$

随后引入扩散步骤，将伪随机数序列生成的密钥序列 $\boldsymbol{K}_\mathrm{d}$ 与混排后的图像 $\boldsymbol{S}^*$ 进行按位异或操作，生成加密图像 $\boldsymbol{S}_\mathrm{enc}$，即

$$\boldsymbol{S}_\mathrm{enc}=\boldsymbol{S}^*\oplus\boldsymbol{K}_\mathrm{d}。\tag{2}$$

最后进行二次加密，使用 QR 编辑器将 $\boldsymbol{S}_\mathrm{enc}$ 编码为 QR，作为嵌入信息，至此加密过程结束。

解密过程是加密过程的逆过程。在接收端，对从隐写图像提取的 QR′ 进行扫描，获得加密图像 $\boldsymbol{S}'_{\mathrm{enc}}$，再使用加密时的扩散密钥 $K_{\mathrm{d}}$ 进行按位异或操作，恢复混排后的图像 $\boldsymbol{S}'^{*}$，即

$$\boldsymbol{S}'^{*}=\boldsymbol{S}'_{\mathrm{enc}}\oplus\boldsymbol{K}_{\mathrm{d}}。\tag{3}$$

然后，根据加密时生成的排列索引 $\pi(seq)$ 进行逆混排，恢复原始像素顺序（$\boldsymbol{S}'$），即

$$\boldsymbol{S}'=\boldsymbol{S}'^{*}[\pi^{-1}(seq)]\oplus\boldsymbol{K}_{\mathrm{d}},\tag{4}$$

至此解密过程完成。

### 2.3 自监督判别器

BYOL（bootstrap your own latent）[19] 无需大量标注即可自监督学习隐写图与封面图的细微差异，并通过跨视图一致性增强判别器特征的稳定性与抗干扰能力。基于此，设计了基于 BYOL 框架的自监督判别器，将对比学习与 GAN 博弈机制相结合，通过构造特征对比任务来实现无标签监督，从而引导 DISG-Net 生成更接近载体统计特性的隐写图像。该判别器既保留了对抗博弈思想，又避免了监督依赖，提升了系统的泛化能力。自监督判别器由在线分支和目标分支组成，如图 2 所示。

自监督判别器通过构造正样本对，实现无标签训练。在线分支由编码器 $E$、在线投影器 $Q$ 和预测器 $P$ 组成。编码器不采用原始的 ResNet，而是使用基于 CNN 的判别器结构。其将输入的封面图像（$\boldsymbol{x}_{\mathrm{cov}}$）和隐写图像（$\boldsymbol{x}_{\mathrm{ste}}$）通过多层 ConvBNRelu 模块，提取图像的低频轮廓和高频细节特征，并将 3 通道输入图像映射为 64 维特征向量 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_2$。投影器为两层 MLP，将特征向量映射到专门的 256 维投影空间并得到特征 $\boldsymbol{h}_1$ 和 $\boldsymbol{h}_2$，在该空间内进行自监督对比和预测，从而提升训练稳定性并保护原始表征。预测器同为两层 MLP，将输入的特征 $\boldsymbol{h}_1$ 和 $\boldsymbol{h}_2$ 进行预测，得到特征 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$（$\boldsymbol{p}_1$、$\boldsymbol{p}_2\in\mathbb{R}^{p}$）。预测特征为在线分支对输入图像在表征空间中的估计，供与目标分支的特征进行相似性比较，以实现跨视图的一致性约束。

预测特征为

$$\boldsymbol{p}_1=P(Q(E(\boldsymbol{x}_{\mathrm{cov}}))),\ \boldsymbol{p}_2=P(Q(E(\boldsymbol{x}_{\mathrm{ste}})))。\tag{5}$$

目标分支由目标编码器 $E_{\mathrm{t}}$ 和目标投影器 $Q_{\mathrm{t}}$ 组成。目标编码器结构与在线编码器对称，用于提取高维特征；目标投影器为两层 MLP，将特征映射到低维投影空间，输出稳定特征。目标分支不参与反向传播，其参数 $\theta_{\mathrm{t}}$ 通过指数移动平均（exponential moving average，EMA）方法，由在线分支参数 $\theta_{\mathrm{o}}$ 更新，即

$$\theta_{\mathrm{t}}\leftarrow\tau\theta_{\mathrm{t}}+(1-\tau)\theta_{\mathrm{o}},\tag{6}$$

式中，$\tau\in[0,1)$ 为衰减率。

本研究中，$\tau=0.9$，以保证目标分支更新缓慢，从而提供稳定的特征表征。在前向计算中，通过目标编码器和目标投影器，从输入的封面图像 $\boldsymbol{x}_{\mathrm{ste}}$ 和隐写图像 $\boldsymbol{x}_{\mathrm{cov}}$ 提取目标特征：

$$\boldsymbol{z}_{1t}=Q_{\mathrm{t}}\left(E_{\mathrm{t}}\left(\boldsymbol{x}_{\mathrm{cov}}\right)\right),\quad \boldsymbol{z}_{2t}=Q_{\mathrm{t}}\left(E_{\mathrm{t}}\left(\boldsymbol{x}_{\mathrm{ste}}\right)\right)。\tag{7}$$

自监督判别器利用稳定的目标分支引导在线分支学习特征，通过最大化在线分支预测特征 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$ 分别与目标特征 $\boldsymbol{z}_{1t}$ 和 $\boldsymbol{z}_{2t}$ 的余弦相似度，实现跨视图一致性约束，使隐写图像在特征空间中更接近封面图像，从而进一步提升隐蔽性与安全性。自监督判别器输出为

$$D(\boldsymbol{x}_{\mathrm{cov}},\boldsymbol{x}_{\mathrm{ste}})=\frac{1}{N}\sum_{i=1}^{N}\left[\cos(p_1^{(i)},z_{2t}^{(i)})+\cos(p_2^{(i)},z_{1t}^{(i)})\right],\tag{8}$$

式中，$N$ 为批量大小。

### 2.4 损失函数

为实现 QR 秘密信息的高保真嵌入与提取，同时

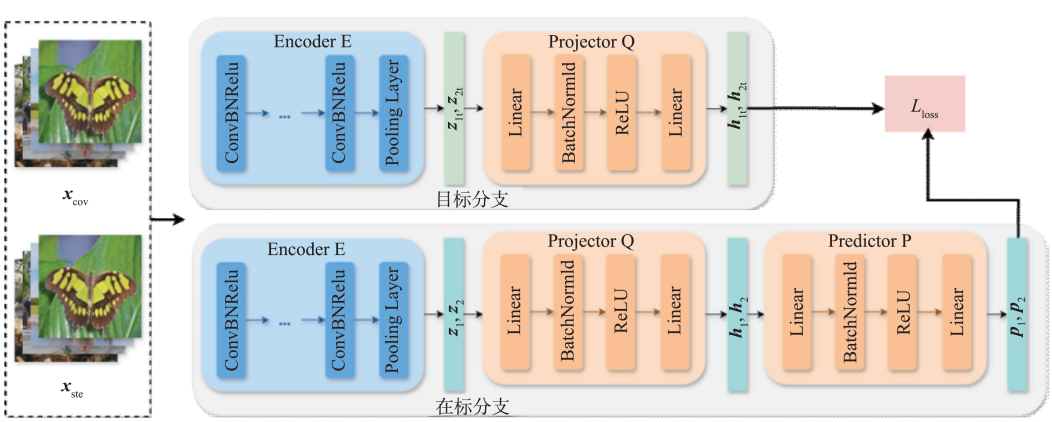



**图 2 自监督判别器结构图**

**Fig. 2 Self-supervised discriminator architecture**

保证隐写图像的视觉一致性与抗检测性，构建了由重构损失、引导损失、对比损失和哈希损失组成的多目标联合优化函数。

1）重构损失用于衡量重建QR与原始QR的差异，确保秘密信息的准确提取。重构损失为

$$L_{\text{rec}}=MSE(\hat{\boldsymbol{q}},\boldsymbol{q})=\frac{1}{H\times W\times C}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{C}\left(\hat{q}_{i,j,k}-q_{i,j,k}\right)^2。\quad(9)$$

式中：$\boldsymbol{q}$为原始QR；$\hat{\boldsymbol{q}}$为重建QR；$H$、$W$、$C$分别为QR的高度、宽度与通道数。

2）引导损失用于约束隐写图像与原始封面图像的像素级相似度，减少QR嵌入带来的视觉干扰，保证隐写图像的自然性。引导损失为

$$L_{\text{gui}}=MSE(\boldsymbol{x}_{\text{ste}},\boldsymbol{x}_{\text{cov}})=\frac{1}{H\times W\times C}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{C}\left(x_{\text{ste},i,j,k}-x_{\text{cov},i,j,k}\right)^2。\quad(10)$$

3）基于BYOL的对比损失用于增强隐写图像的抗检测性，通过自监督判别器在特征空间拉近封面图像与隐写图像的距离，使二者在高层特征上难以区分。对比损失（$L_{\text{con}}$）为自监督判别器输出结果（见式（8））。

4）哈希损失用于解决像素级损失难以反映人眼感知一致性的问题，其利用平均哈希算法将图像压缩为64位二进制指纹，并以汉明距离衡量隐写图像与载体图像的结构相似性。具体而言，将图像缩放至$8\times 8$并采用AHash灰度处理，将每像素与平均灰度比较，大于等于平均值记为1，否则记为0，最后得到哈希向量$AH(\boldsymbol{I})\{0,1\}^{64}$。

$$AH(\boldsymbol{I})_b=\begin{cases}1, & I'_b\geqslant\mu(\boldsymbol{I}');\\ 0, & \text{其他}。\end{cases}\quad(11)$$

式中：$\boldsymbol{I}'$为缩放灰度化后的图像；$b\{1,2,\cdots,64\}$为哈希值的比特位索引；$\mu(\boldsymbol{I}')$为平均灰度。

哈希损失为隐写图像与封面图像哈希值的汉明距离，即通过逐位异或累加，捕捉QR嵌入引起的结构变化，规避像素级敏感性。

$$L_{\text{aha}}=d(AH(\boldsymbol{x}_{\text{ste}}),AH(\boldsymbol{x}_{\text{cov}}))=\sum_{b=1}^{64}\left|AH(\boldsymbol{x}_{\text{ste}})_b\oplus AH(\boldsymbol{x}_{\text{cov}})_b\right|。\quad(12)$$

模型的总损失函数为上述4项损失的加权和，即通过联合优化实现多目标约束。通过合理的权重设置，实现秘密信息恢复、载体保真、抗检测性与感知一致性的平衡，为高安全性QR隐写提供约束。

总损失函数为

$$L_{\text{tot}}=\lambda_1L_{\text{rec}}+\lambda_2L_{\text{gui}}+\lambda_3L_{\text{con}}+\lambda_3L_{\text{aha}}。\quad(13)$$

# 3 实验结果与分析

## 3.1 实验设置

对于模型训练和测试，本文使用Div2k数据集[20]训练DISG-Net，使用COCO数据集[21]和ImageNet数据集[22]分别作为验证集和测试集。在Div2k中随机选取封面图像和秘密图像各400张，输入图像的大小统一调整为$128\times 128$，训练时批量大小设置为10。Adam优化器用于$\beta_1$=0.5、$\beta_2$=0.999的训练过程。学习率从$1\times 10^{-5}$开始，每$2\times 10^4$迭代后除以10。

损失函数的4个权重为作为超参数，利用Optuna[23]自动优化。每个权重设定了连续搜索区间，并采用TPE（tree-structured parzen estimator）采样策略，每次训练通过部分epoch训练加早停评估权重组合性能，从而高效探索高维空间，获得更优的权重配置。综合性能在COCO验证集上评估，指标包括峰值信噪比（PSNR）[24]、平均像素差（APD）[25]、结构相似性（SSIM）[26]和均方根误差（RMSE）[27]。其中，PSNR和SSIM越大、APD和RMSE越小表示图像质量越高。Optuna记录每次训练的权重与性能关系，最终收敛至最优组合：重构损失权重1.15、引导损失权重1.00、对比损失权重0.68、哈希损失权重0.24。该权重配置在保持载体高保真度的同时，提高秘密图像重构精度并增强隐写抗检测性。

为评估模型性能，将DISG-Net与UDH[28]、HiNet[29]、DeepMIH[30]、DAH-Net[31]、iSCMIS[32]、PUSNet[33]、U-INR[34]等可逆网络方法进行比较，通过PSNR、SSIM、APD和RMSE 4个指标衡量隐写图像不可感知性。

## 3.2 隐写质量分析

将DISG-Net与其他方法在Div2k、COCO和ImageNet数据集上进行比较，结果如表1所示。表中，“↑”表示值越高越好，“↓”表示值越低越好。结果显示，DISG-Net在封面/隐写图像对和秘密/恢复秘密图像对上均优于其他方法。本模型在隐写方面，在Div2k数据集上，PSNR达46.69 dB，SSIM为0.9992；在ImageNet数据集上，PSNR达47.88 dB。本模型在秘密恢复方面，SSIM达到1.0000，APD和RMSE分别降至最低的0.45和0.54，实现几乎无损的QR恢复。对应地，本模型在不同数据集封面图像上的QR识别情况如表2所示。3种数据集上的QR识别准确率均达100%。可见，本模型在高保真与低失真方面表现突出，且在COCO和ImageNet数据集上也展现出良好泛化性。

**表 1　不同隐写方法在不同数据集上的性能表现**

**Table 1　Performance of different steganography methods on different image datasets**

| 数据集 | 封面 / 隐写图像对 \| 秘密 / 恢复秘密图像对 | | | | |
|---|---|---|---|---|---|
| | 模型 | PSNR/dB ↑ | SSIM ↑ | APD ↓ | RMSE ↓ |
| Div2k | UDH | 44.68 \| 42.02 | 0.8913 \| 0.9649 | 2.38 \| 2.15 | 4.43 \| 3.23 |
| | HiNet | 44.86 \| 44.35 | 0.9692 \| 0.9797 | 1.00 \| 0.88 | 2.36 \| 1.28 |
| | DeepMIH | 43.72 \| 44.51 | 0.9895 \| 0.9920 | 2.21 \| 1.76 | 6.82 \| 2.54 |
| | DAH-Net | 36.59 \| 40.72 | 0.9896 \| 0.9896 | 1.72 \| 1.54 | 2.79 \| 1.98 |
| | iSCMIS | 45.78 \| 42.53 | 0.9924 \| 0.9858 | 1.62 \| 1.98 | 2.42 \| 2.93 |
| | PUSNet | 38.15 \| 36.88 | 0.9792 \| 0.8363 | 2.30 \| 8.75 | 3.33 \| 11.95 |
| | U-INR | 38.32 \| 37.11 | 0.9890 \| 0.9851 | 2.17 \| 2.84 | 2.72 \| 3.86 |
| | 本模型 | **46.69 \| 51.15** | **0.9992 \| 1.0000** | **0.27 \| 0.20** | **0.60 \| 0.50** |
| COCO | UDH | 38.01 \| 34.75 | 0.8988 \| 0.9175 | 3.79 \| 4.82 | 4.65 \| 7.79 |
| | HiNet | 36.55 \| 47.06 | 0.9582 \| 0.9561 | 0.81 \| 1.82 | 3.69 \| 2.78 |
| | DeepMIH | 40.30 \| 45.31 | 0.9821 \| 0.9512 | 2.83 \| 2.83 | 3.88 \| 3.53 |
| | DAH-Net | 35.59 \| 39.51 | 0.9856 \| 0.9792 | 2.96 \| 2.73 | 3.84 \| 2.67 |
| | iSCMIS | 41.53 \| 37.48 | 0.9818 \| 0.9664 | 2.53 \| 3.74 | 3.78 \| 5.48 |
| | PUSNet | 39.09 \| 36.96 | 0.9772 \| 0.8211 | 2.01 \| 8.71 | 2.96 \| 12.14 |
| | U-INR | 39.70 \| 37.53 | 0.9889 \| 0.9873 | 1.55 \| 2.73 | 2.39 \| 3.41 |
| | 本模型 | **41.86 \| 51.80** | **0.9973 \| 1.0000** | **0.96 \| 0.26** | **1.44 \| 0.65** |
| ImageNet | UDH | 43.87 \| 32.53 | 0.9018 \| 0.9034 | 3.81 \| 3.56 | 4.66 \| 8.22 |
| | HiNet | 45.95 \| 47.07 | 0.9528 \| 0.9473 | 2.18 \| 1.95 | 3.63 \| 2.86 |
| | DeepMIH | 43.29 \| 45.38 | 0.9870 \| 0.9452 | 2.18 \| 2.72 | 2.81 \| 3.54 |
| | DAH-Net | 39.93 \| 37.31 | 0.9857 \| 0.9791 | 2.98 \| 2.63 | 3.78 \| 3.20 |
| | iSCMIS | 41.64 \| 37.83 | 0.9818 \| 0.9689 | 2.59 \| 3.79 | 3.79 \| 5.54 |
| | PUSNet | 38.94 \| 36.28 | 0.9756 \| 0.8028 | 2.21 \| 9.58 | 3.06 \| 13.43 |
| | U-INR | 39.06 \| 36.82 | 0.9813 \| 0.9802 | 1.61 \| 2.90 | 2.45 \| 4.12 |
| | 本模型 | **47.88 \| 52.36** | **0.9968 \| 1.000** | **0.45 \| 0.21** | **0.89 \| 0.54** |

注：加粗表示各列最优结果，下同；划横线表示各列第二优结果。

**表 2　DISG-Net 在不同数据集封面图像上的 QR 识别准确率**

**Table 2　DISG-Net's QR recognition accuracy on cover images across different datasets**

| 数据集 | 准确率 /% |
|---|---|
| Div2k | 100 |
| COCO | 100 |
| ImageNet | 100 |

图 3 为本模型在不同数据集封面图像上的隐写与恢复图像。从图 3 可以看出，封面图像与隐写图像在视觉质量上几乎完全一致，体现了本模型具有较好的隐写效果。

### 3.3　安全性分析

为评估模型安全性，在 COCO 数据集上将本模型与多种代表性图像隐写方法（Baluja[3]、HiDDeN[35]、ISN[36]、DeepMIH、iSCMIS）进行了对比实验。各模型在两类典型隐写分析器 SRNet 与 Zhu-Net 下的检测结果如表 3 所示。模型检测准确率越接近 50%，说明隐写图像越难以区分，安全性越高；反之，准确率偏离 50% 越远，则表明该方法更易被检测，安全性较低。

从表 3 可以看出：Baluja 方法在 SRNet 与 Zhu-Net 下的检测准确率分别高达 99.67% 和 99.31%，HiDDeN 与 ISN 亦接近 80%，均易被检测。相比之下，本模型在 SRNet 与 Zhu-Net 下的检测准确率仅分别为 54.79% 与 56.40%，几乎逼近随机猜测的 50% 水平，显著优于所有对比方法。


**图 3　本模型在不同数据集下的视觉质量分析**

**Fig. 3　Visual quality analysis across different datasets**

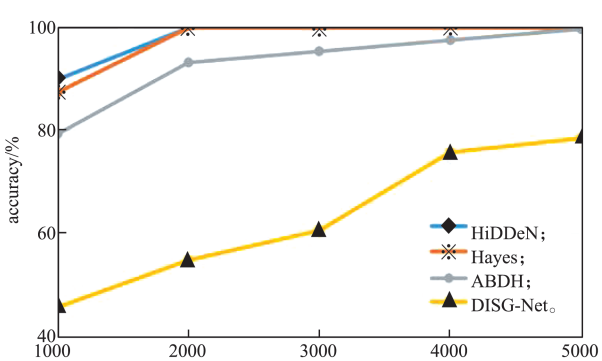
**表 3　SRNet 与 ZhuNet 对 6 种隐写方案的检测准确率**

**Table 3　Detection accuracy of 6 steganographic schemes using SRNet and Zhu-Net**　%

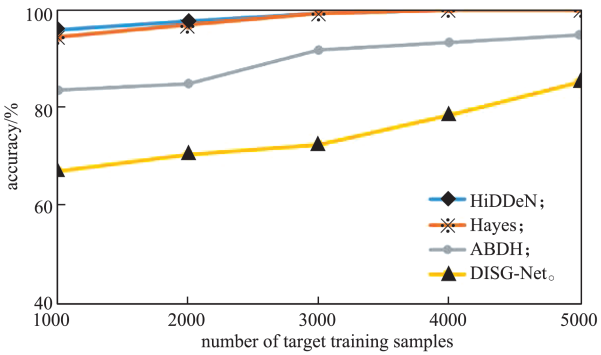
| 模型 | SRNet | Zhu-Net |
|---|---|---|
| Baluja | 99.67 | 99.31 |
| HiDDeN | 76.49 | 78.36 |
| ISN | 79.30 | 84.91 |
| DeepMIH | 89.25 | 71.91 |
| iSCMIS | 69.64 | 69.64 |
| 本模型 | **54.79** | **56.40** |

图 4 展示了在不同训练样本规模下，使用 SRNet 和 XuNet 对 HiDDeN、Hayes[36]、ABDH[37] 和本模型进行的隐写分析结果。相比于 HiDDeN、Hayes 和 ABDH，本模型在不同数据设置下均保持更优的安全性能。

综上，所有隐写分析实验均验证了本模型能够生成难以检测的隐写图像，充分体现了卓越的抗检测安全性与隐蔽性。




a）SRNet




b）XuNet

**图 4　由 SRNet 和 XuNet 检测的安全性能**

**Fig. 4　Security performances detected by SRNet and XuNet**

### 3.4　消融实验

为验证各模块的有效性，在 Div2k 数据集上进行消融实验，结果如表 4~5 所示。在基础网络（Basic）上，逐步加入自监督判别器（dis）、哈希损失（ahash）及两者联合（dis+ahash），并评估封面 / 隐写图像对和秘密 / 恢复秘密图像对的质量。

由表 4 可知，dis 和 ahash 的加入均显著提升了隐写图像与秘密图像的质量。在图像隐写方面，仅 dis 的加入使 PSNR 提升了约 5.24 dB，仅 ahash 的加入提升了约 3.69 dB，联合使用提升了约 5.22 dB，同时 APD 和 RMSE 明显降低。在秘密图像恢复上，单独使用 dis 和 ahash 分别提升 PSNR 约 1.65 dB 和 1.66 dB，联合使用提升约 2.80 dB，APD 和 RMSE 达到最低值，表明两者协同作用显著增强重构精度与视觉一致性。

表 5 为不同秘密图像类型对模型性能的影响。由表 5 可知，当秘密图像为 QR 图像时，隐写图像和恢复的秘密图像质量均更高，隐写图像的 PSNR 达 47.6317 dB，APD 降至 0.2770，这表明 QR 图像结构更利于精确嵌入与恢复，提高模型整体性能。

**表 4　不同组件组合的图像隐写质量**

**Table 4　Image steganography quality of different component combinations**

| 模型 | 封面 / 隐写图像对 \| 秘密 / 恢复秘密图像对 | | | |
|---|---|---|---|---|
| | PSNR/dB ↑ | SSIM ↑ | APD ↓ | RMSE ↓ |
| Basic | 41.4741 \| 46.7759 | 0.9962 \| 0.9988 | 0.7305 \| 0.6406 | 1.3958 \| 0.9562 |
| Basic+dis | 46.7170 \| 48.4272 | **0.9987** \| 0.9993 | 0.3885 \| 0.4352 | 0.7449 \| 0.7172 |
| Baic+ahash | 45.1623 \| 48.4404 | 0.9973 \| 0.9991 | 0.6987 \| **0.2745** | 1.2190 \| 0.7120 |
| Baic+dis+ahash | **46.6911** \| **49.5778** | 0.9984 \| **0.9994** | **0.3869** \| 0.3010 | **0.6424** \| **0.5992** |

**表 5　在 Div2k 数据集上使用不同秘密图像的 DISG-Net 消融结果**

**Table 5　Ablation study of DISG-Net on the Div2k dataset using different secret images**

| 图像 | 封面 / 隐写图像对 \| 秘密 / 恢复秘密图像对 | | | |
|---|---|---|---|---|
| | PSNR/dB ↑ | SSIM ↑ | APD ↓ | RMSE ↓ |
| 普通图像 | 46.6911 \| 49.5778 | 0.9984 \| 0.9994 | 0.3869 \| 0.3010 | 0.6424 \| 0.5992 |
| QR 图像 | **47.6317** \| **51.1562** | **0.9992** \| **0.9999** | **0.2770** \| **0.2015** | **0.6024** \| **0.5076** |

## 4　结语

本研究提出了 DISG-Net 模型。其通过融合 QR 安全加密、可逆隐写网络和自监督判别器，实现了双重加密下的可逆图像隐写。结果表明，DISG-Net 在隐写图像质量、秘密信息恢复精度及安全性方面均优于现有方法。进一步的消融实验也验证了各模块的有效性。未来，将提升模型在极端干扰情况下的稳定性，并将其应用拓展至多图隐写、视频隐写、医学隐私保护以及数字版权追踪等领域。

**参考文献：**

[1]　PROVOS N，HONEYMAN P. Hide and Seek: An Introduction to Steganography[J]. IEEE Security & Privacy，2003，1(3)：32-44.

[2]　FRIDRICH J. Steganography in Digital Media: Principles, Algorithms, and Applications[M]. Cambridge：Cambridge University Press，2010：1-24.

[3]　BALUJA S. Hiding Images in Plain Sight: Deep Steganography[C]//31st Conference on Neural Information Processing Systems. Long Beach：NIPS，2017：29764034.

[4]　YANG F F，WU W H，DENG C，et al. Sustainable Lignocellulosic Room Temperature Phosphorescent Inks for Intelligent Packaging and Anti-Counterfeiting[J]. ACS Sustainable Chemistry & Engineering，2025，13(19)：7199-7211.

[5] ZHANG K A，CUESTA-INFANTE A，XU L，et al. SteganoGAN: High Capacity Image Steganography with GANs[EB/OL]. [2024−11−16]. https://arxiv.org/abs/1901.03892.

[6] LIU Y，WANG X，ZHANG Z，et al. Improved GAN - based Image Steganography with Adaptive Embedding[J]. IEEE Access，2021，9：162345−162354.

[7] LI F，YANG M. Challenges in GAN-Based Steganography[J]. Journal of Visual Communication & Image Representation，2020，66：102731.

[8] WU H，WANG R，TAO D. SteganoUNet: U-Net Based Deep Image Steganography[C]//International Conference on Image Processing. [S. l.]：IEEE，2020：3703−3707.

[9] ZHANG Y，LIU S，WANG C. Attention-Enhanced U-Net for Robust Image Steganography[J]. IEEE Transactions on Circuits and Systems for Video Technology，2021，31(9)：3500−3512.

[10] CHEN L，HU J. Channel Attention in Deep Steganography[J]. Signal Processing，2020，176：107689.

[11] SUN K，LI W，ZHANG Q. High-Frequency Information Loss in U-Net Steganography[J]. Multimedia Tools and Applications，2021，80：17265−17280.

[12] JING J，DENG X，XU M. HiNet: Invertible Neural Networks for Image Steganography[J]. IEEE Transactions on Multimedia，2021，23(4)：1024−1037.

[13] ZHANG X，LI Y，LI B. RIIS: Reversible Information Image Steganography with Multi-Scale INNs[C]// European Conference on Computer Vision(ECCV). Berlin：Springer，2022：345−361.

[14] CHEN Y，WANG Z，TAO D. IIS: Deep Reversible Information Steganography for High-Resolution Images[J]. IEEE Transactions on Image Processing，2021，30：8765−8778.

[15] ZHANG Y，XIAO D，WEN W. Chaos-Based Image Encryption and DCT Steganography[J]. Chaos Solitons Fractals，2020，139：109993.

[16] LIU M，CHEN L. Analysis of Chaos Encryption in Steganography[J]. Optik，2020，218：164921.

[17] 王勇智 . 基于伪随机数生成器的视频隐写算法 [J]. 包装学报，2024，16(5)：58−62.

[18] 王晓红，马春运，石明光 . 基于全息加密与密集残差网络的图像隐写 [J]. 包装学报，2025，17(3)：46−54.

[19] GRILL J B，STRUB F，ALTCHÉ F，et al. Bootstrap Your Own Latent a New Approach to Self-Supervised Learning[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver：ACM，2020：21271−21284.

[20] AGUSTSSON E，TIMOFTE R. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S. l.]：CVPR，2017：1122−1131.

[21] LIN T Y，MAIRE M，BELONGIE S，et al. Microsoft COCO: Common Objects in Context[EB/OL]. 2014：arXiv: 1405.0312. [2025−03−10]. https://arxiv.org/abs/1405.0312.

[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. [S. l.]：IEEE，2009：248−255.

[23] AKIBA T，SANO S，YANASE T，et al. Optuna: A Next-Generation Hyperparameter Optimization Framework[EB/OL]. [2024−05−15]. https://arxiv.org/pdf/1907.10902.

[24] HUYNH-THU Q，GHANBARI M. Scope of Validity of PSNR in Image/Video Quality Assessment[J]. Electronics Letters，2008，44(13)：800−801.

[25] WANG Z，BOVIK A C，SHEIKH H R，et al. Image Quality Assessment: From Error Visibility to Structural Similarity[J]. IEEE Transactions on Image Processing，2004，13(4)：600-612.

[26] SHITTU S. Root Mean Square Error (RMSE) in AI: What You Need to Know[EB/OL]. [2025−09−10]. https://www.towardsai.net/p/root-mean-square-error-rmse-in-ai-what-you-need-to-know.

[27] HORÉ A，ZIOU D. Image Quality Metrics: PSNR vs. SSIM[C]//2010 20th International Conference on Pattern Recognition. Istanbul：IEEE，2010：2366−2369.

[28] ZHANG C N，BENZ P，KARJAUV A，et al. UDH: Universal Deep Hiding for Steganography, Watermarking, and Light Field Messaging[C]//34th Conference on Neural Information Processing Systems(NeurIPS). Vancouver：NeurIPS，2020：1−12.

[29] JING J P，DENG X，XU M，et al. HiNet: Deep Image Hiding by Invertible Network[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). [S. l.]：IEEE，2021：4733−4742.

[30] GUAN Z Y，JING J P，DENG X，et al. DeepMIH: Deep Invertible Network for Multiple Image Hiding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2023，45(1)：372−390.

[31] ZHANG L，LU Y，LI J X，et al. Deep Adaptive Hiding Network for Image Hiding Using Attentive Frequency Extraction and Gradual Depth Extraction[J]. Neural Computing and Applications，2023，35(15)：10909−10927.

[32] LI F Y，SHENG Y，ZHANG X P，et al. ISCMIS: Spatial-Channel Attention Based Deep Invertible Network for Multi-Image Steganography[J]. IEEE Transactions on Multimedia，2024，26：3137−3152.

[33] ZHANG Y，ZHANG X. PUSNet: Progressive Upsampling Steganography Network for High-Resolution Image Hiding[J]. IEEE Transactions on Circuits and Systems for Video Technology，2023，33(6)：2895−2908.

[34] SONG Q，LUO Z Y，HUANG X F，et al. Unified Steganography via Implicit Neural Representation[EB/OL]. [2025−06−24]. https://arxiv.org/abs/2505.01749.

[35] ZHU J R，KAPLAN R，JOHNSON J，et al. HiDDeN: Hiding Data with Deep Networks[C]//European Conference on Computer Vision. Cham：Springer，2018：682−697.

[36] HAYES J，DANEZIS G. Generating Steganographic Images via Adversarial Training[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach：ACM，2017：1951−1960.

[37]YU H，LI M，QIN C，et al. ABDH: Attention-based deep hiding for robust image steganography[J]. Entropy，2020，22(10)：1140.

（责任编辑：邓　彬）

# Reversible Image Steganography Based on Double Encryption and Self-Supervised Discriminator

WANG Xiaohong，HE Xinjie，MA Chunyun

（College of Publishing，University of Shanghai for Science and Technology，Shanghai 200093，China）

**Abstract**：To enhance the security, visual consistency and anti-steganalysis capability of image steganography, the reversible steganographic network DISG-Net is proposed. The model ensures the security of secret information through an innovatively designed QR code-based dual encryption method, and employs 16 wavelet transform-based reversible blocks to achieve high-fidelity embedding and reversible recovery. Meanwhile, a BYOL self-supervised discriminator is introduced to constrain the feature distribution, making the generated results natural and difficult to detect. By incorporating multiple objective losses, including reconstruction, guidance, contrastive, and hashing losses, the framework achieves both visual consistency and precise recovery of secret information. Experimental results demonstrate that DISG-Net outperforms existing methods in terms of image quality and information security, offering a high-fidelity and tamper-resistant information embedding solution for printing and packaging, thereby enhancing anti-counterfeiting and information protection capabilities.

**Keywords**：image steganography；invertible neural networks；double encryption；self-supervised discriminator