arXiv:2603.11876v1 [cs.CR] 12 Mar 2026

关于 图像内图像隐写术的可检测性

Antoine Mallet

里尔大学，法国国家科学研究中心 UMR 9189 CRIStAL，法国59000里尔

Patrick Bas

里尔大学、法国国家科学研究中心 UMR 9189 CRIStAL，法国59000里尔

摘要

本文研究了流行的图中图隐写方案 [1, 2, 3, 4, 5] 的可检测性。在这种范式下，载荷通常是一幅与载体图像尺寸相同的图像，因此嵌入率极高。我们首先表明，嵌入会产生一种混合过程，可通过独立成分分析轻易识别。随后，我们提出一种简单、可解释的隐写分析方法：利用从图像小波分解中估计出的独立成分的前四阶矩，区分载体成分与隐写成分的分布。实验结果表明，所提出的方法效率很高，八维输入向量的准确率最高可达 84.6%。这一脆弱性分析还得到另外两项事实的支持：提取网络无需密钥；并且相较于经典隐写分析方法，该方法具有很高的可检测性。例如，SRM 与支持向量机结合后，准确率可超过 99%。

索引词— 隐写术、隐写分析、可逆神经网络、独立成分分析

1. 引言

隐写术是一门将秘密消息隐藏在外观无害的媒介中的科学。为了实现有效隐写，隐写方案必须确保对载体媒介所做的修改尽可能难以被潜在攻击者察觉，而这种攻击者被称为隐写分析者。隐写术在这一目标下面临的主要挑战之一，是设计能够将隐写对象的统计可检测性降至最低，同时最大限度地提高嵌入信息相对于载体对象大小的比例的方案；我们将嵌入信息称为载荷。

过去几年中，有一类隐写方案致力于将一幅图像隐藏在另一幅图像中。在这种设置下，载体图像和秘密消息均为图像。不同于经典隐写方案中秘密消息是远小于载体对象的比特流，图像内图像隐写方案旨在嵌入与载体图像大小相当的载荷。为了做到

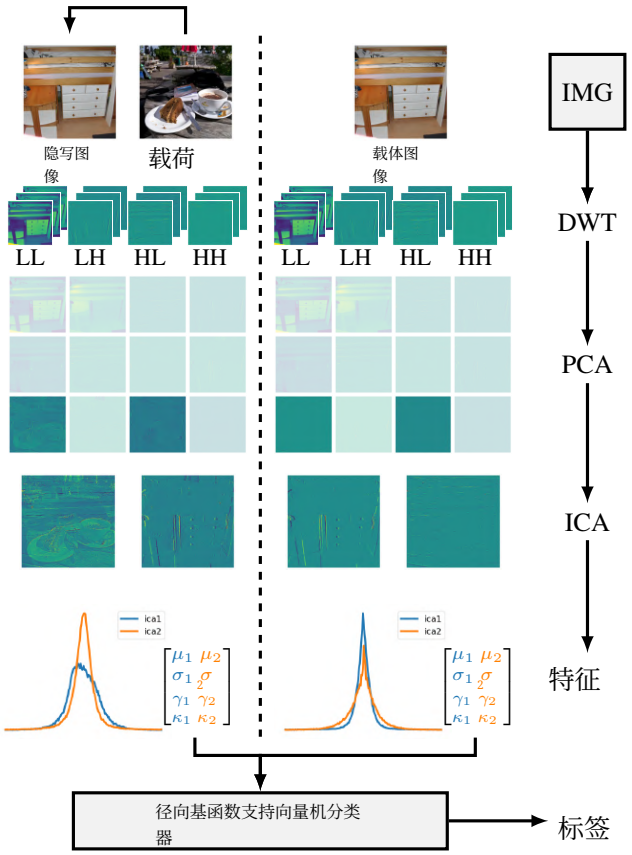


图1。所提出的隐写分析方法概述。首先，对输入图像进行离散小波变换（DWT）。然后，通过主成分分析（PCA）提取弱成分。接着，应用独立成分分析（ICA），提取在语义上接近载体图像和载荷图像的独立源。最后，计算估计得到的独立成分的前四阶矩（均值 μ、标准差 σ、偏度 γ 和峰度 κ），构成具有判别性且可解释的特征向量，并用于训练 SVM 分类器进行检测。

因此，这些方案通常依赖深度学习架构，例如编码器-解码器架构 [6, 7, 4], 可逆神经

网络（可逆神经网络） [1, 8, 9, 2], 或扩散模型 [10]。由于载荷较大，这两种范式之间存在一些概念上的差异。由于目标是传递语义信息，载荷无需完美编码。只要能够感知秘密图像的整体内容，就可以接受一些错误。为此，Baluja [4] 认为：“鉴于我们隐藏的信息量，我们并不试图掩盖存在隐藏消息这一事实。不过，我们确实会 *[.....]* 混淆隐藏消息的内容。”然而，后来的论文提出了旨在避免被有效隐写分析方法检测出的方案，例如 Jing 等人提出的 HiNet 模型 [1]，该模型声称能够实现“大容量、高安全性的图像隐藏”。

尽管此类方案备受关注，但其安全性尚未得到充分研究。特别是，据我们所知，Peng 等人 [11] 的研究是唯一旨在检测此类隐写图像的工作。该研究采用监督学习框架，目标是训练一个替代模型来提取载荷图像[1]。本文提出使用一种非监督特征提取方法来评估图像内图像隐写术方案的安全性，并将其作为监督分类器的输入。该评估首先采用一种新颖的专用隐写分析方法，依靠独立成分分析（ICA）和主成分分析（PCA）对混合信号进行盲源分离。随后，我们还使用经典隐写分析方法评估这些隐写图像的可检测性，并表明它们极易受到此类方法的攻击。

本文其余部分的组织结构如下。第 2 节介绍与图像内图像隐写术和隐写分析相关的既有工作。第 3 节分析基于 INN 的隐写方案所产生的混合过程。第 4 节详细介绍所提出的隐写分析方法，该方法依赖于独立成分分析。第 5 节报告实验结果并进行分析。最后，第 6 节总结全文。

2. 先前研究

本节将详细介绍与图像内图像隐写术相关的必要背景，尤其是基于可逆神经网络的方案。我们还将讨论该领域当前最先进技术的局限性，这些局限性促使我们开展本研究。

请注意，早期的图中隐藏方案由联合训练的卷积神经网络组成 [12,4], ：一个神经网络负责将载荷嵌入载体图像，另一个则负责从隐写图像中恢复载荷。同样，Wu等人 [7]基于卷积层上的残差连接构建了其嵌入-解码架构。

[1]此外，值得指出的是，大多数采用图像内图像隐写术的论文都会提供一些检测结果，并使用经典隐写分析方法进行评估——例如 [1, 8], ，其中的方案被证明极难检测。

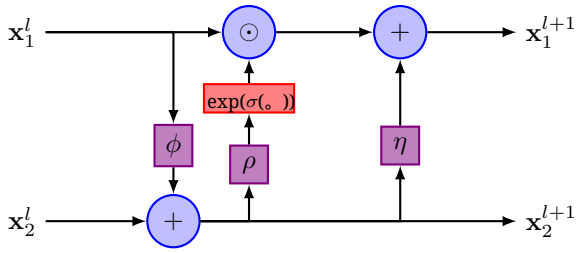


图 2。HiNet 模型 [1] 中可逆神经网络耦合层总体架构示意图。完整的可逆神经网络由 16 个模块组成。箭头表示前向传播的方向。

2.1. 可逆神经网络

可逆神经网络是一类旨在建立输入空间与输出空间之间双射映射的神经网络。该网络由Dinh等人 [13]首次提出。形式上，可逆神经网络可以表示为函数 $f\colon \mathbb{R}^n \to \mathbb{R}^n$，对于每个输入$\mathbf{x} \in \mathbb{R}^n$，都存在唯一的输出$\mathbf{y} = f(\mathbf{x})$，反之亦然，即 $\mathbf{x} = f^{-1}(\mathbf{y})$。这种可逆性通过耦合层的设计实现：耦合层将输入向量$\mathbf{x}$ 拆分为两部分$\mathbf{x}_1$ 和$\mathbf{x}_2$, , 并对它们进行一系列变换以生成两个输出。如图2所示，在耦合层 l处，输出可描述为：

$$\begin{cases} \mathbf{x}_1^{l+1} = \mathbf{x}_1^l \odot \exp\left(\sigma\left(\rho(\mathbf{x}_2^{l+1})\right)\right) + \eta(\mathbf{x}_2^{l+1}), \\ \mathbf{x}_2^{l+1} = \mathbf{x}_2^l + \phi(\mathbf{x}_1^l), \end{cases} \tag{1}$$

其中， ϕ、 ρ和 η 是可学习函数，通常由小型神经网络实现； σ 是 sigmoid 激活函数， $\odot$ 表示逐元素乘法。关键的是，只需对所考虑的函数执行相反的运算，即可简单地计算出输入：

$$\begin{cases} \mathbf{x}_1^l = \left(\mathbf{x}_1^{l+1} - \eta(\mathbf{x}_2^{l+1})\right) \odot \exp\left(-\sigma\left(\rho(\mathbf{x}_2^{l+1})\right)\right), \\ \mathbf{x}_2^l = \mathbf{x}_2^{l+1} - \phi(\mathbf{x}_1^l). \end{cases} \tag{2}$$

因此，可逆神经网络适用于需要可逆变换的应用，例如图像内图像隐写术，其目标是将秘密图像嵌入载体图像中，并在之后无损地恢复秘密图像。

文献并未直接处理颜色通道，而是提出对图像进行哈尔小波分解 [1, 3, 2]。可以说，这有助于学习模型将重要的低频信息与在整体语义中作用较小的高频信息分离开来，从而减少信息损失。正如我们将在第 4 节中所展示的，我们以这种分解为基础进行独立成分分析。



图 3。使用不同噪声输入从 HiNet 隐写图像中恢复出的秘密图像。从左到右：隐写图像、载荷图像、加入随机噪声后恢复出的载荷、加入零噪声后恢复出的载荷。峰值信噪比（单位：dB）是通过比较恢复出的秘密图像与原始秘密图像计算得出的。

2.2. 安全性相关局限

尽管基于可逆神经网络的图像隐写于图像方案具有潜力，但其安全性在很大程度上仍未得到检验。根据柯克霍夫原则 [14], ，即使攻击者知晓隐写方案的所有细节（密钥除外），该方案仍应保持安全。然而，现有的大多数基于可逆神经网络的图像隐写于图像研究 [1, 8, 9] ，在嵌入与提取过程中并未使用密钥。这一疏漏引发了严重的安全隐患，因为这意味着，只要攻击者了解嵌入算法，就能在没有任何秘密信息的情况下提取隐藏图像。以 HiNet 模型 [1], 为例，可逆神经网络的架构和权重完全决定了嵌入与提取过程；而在显露秘密图像时，输入残差噪声的选择似乎无关紧要，因为如图 3 所示，将噪声设为零就已足以较好地重建秘密图像。

此外，正如 Lu等人 [8] 所强调的：“在可逆神经网络中，基本的可逆耦合层是 *NICE* 提出的加性仿射变换”（非线性独立成分估计） [13]。这一观察表明，仅将模型用作密钥并不足以确保安全，因为耦合层基于众所周知的数学变换，攻击者可以对其进行分析，并有可能将其逆转。无论如何，正如我们将在后续章节中强调的那样，即使不借助 NICE 等非线性模型，我们也可以使用线性独立成分分析算法提取有意义的成分。

3. 分析基于可逆神经网络的方案的混合过程

在本节中，我们对基于可逆神经网络的图像中图像隐写所执行的混合过程进行初步分析。我们重点关注 HiNet模型 [1], ，它是文献中最具代表性且被广泛认可的模型之一。我们的目标是理解可逆神经网络所执行的混合过程的本质，并评估哪些信息

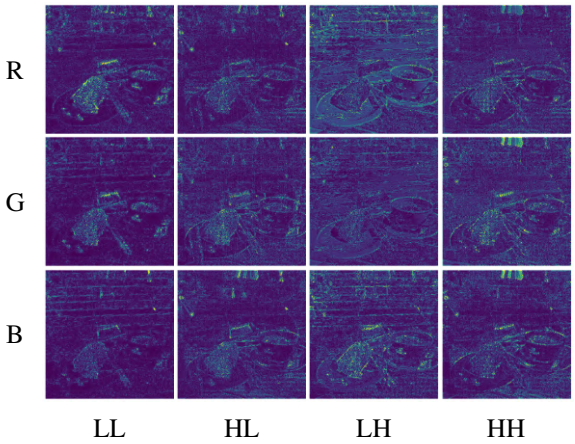


图4。隐写图像离散小波变换各子带中的嵌入变化，计算方式为与载体图像的离散小波变换子带之差。

来自载荷图像的信息由隐写图像承载。

为此，我们进行了一项简单的分析。首先，对载体图像、载荷图像和隐写图像应用离散小波变换。然后，我们可以直观地呈现载体图像与隐写图像的离散小波变换子带之间的差异。该差异即为可逆神经网络带来的嵌入变化，因此能让我们了解为生成隐写图像而添加到载体图像中的载荷信息。

图1所示的载体图像、载荷图像和隐写图像对应的差异如图4所示。值得注意的是，我们可以看到，嵌入变化并非随机噪声，而是携带了载荷图像中的大量信息。

为评估嵌入变化平均携带了多少载荷信息，我们拟计算载荷图像离散小波变换的各子带与隐写图像离散小波变换各子带中的嵌入变化之间的相关性。该分析结果如图5所示，取100组载体图像—载荷图像—隐写图像三元组的平均值。可以看出，嵌入变化与载荷图像的低频分量高度相关，而总体上携带的载荷图像高频分量信息较少。另需指出的是，载荷图像的 HL 和 HH 子带分别与载体图像的 HL 和 HH 子带相关。

4. 提出的方法

本节将详细介绍所提出的隐写分析方法。首先，我们回顾独立成分分析的主要原理，这些原理构成了我们方法的基础。随后，我们说明使用主成分分析作为预处理步骤的原因，以便从小波分解中选择最相关的子带用于独立成分分析。最后，我们将详细介绍用于训练简单分类器的所提出特征。

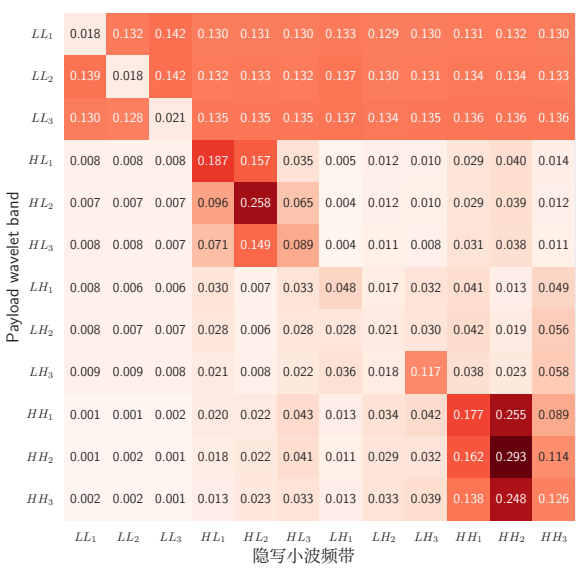


图 5。使用 HiNet 模型 [1]时，载荷图像离散小波变换的子带与载体图像和隐写图像相应子带之差之间的相关性矩阵。采用皮尔逊相关系数。结果表明，载荷的低频子带主要嵌入在隐写图像中。

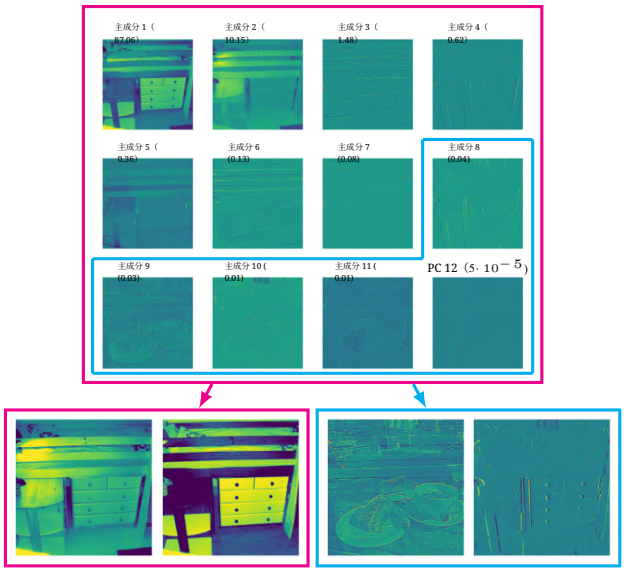


图 6。使用全部 PCA 分量或仅使用最小的 5 个分量时得到的 ICA 子带对比。图中给出了每个分量所解释的方差（百分比）。

4.1. 独立成分分析

独立成分分析（ICA）是一种计算技术，用于将多变量信号分解为相加的、相互独立的非高斯分量。它对于盲源分离问题非常有益，在这类问题中，目标是在不知道混合过程的先验信息的情况下，从观测混合信号中恢复源信号。独立成分分析的基本假设是，源信号在统计上相互独立且服从非高斯分布。

从数学上，独立成分分析 [15]可描述如下。设$\mathbf{x} = [x_1, x_2, \ldots, x_m]^T$为观测随机向量，并假定它是 n 个独立源信号$\mathbf{s} = [s_1, s_2, \ldots, s_n]^T$的线性混合：$\mathbf{x} = A\mathbf{s}$，其中 A 为未知混合矩阵。独立成分分析的目标是根据观测数据$\mathbf{x}$同时估计混合矩阵 A 和源信号$\mathbf{s}$。已有多种算法可用于执行独立成分分析 [16, 17, 18]。本研究采用 Hyv¨arinen [19]提出的 FastICA 算法。该方法是一种迭代不动点法，用于最大化所估计成分的非高斯性。更多细节请参阅 [19][第 6 节。1] 我们使用 scikit-learn 库中的 FastICA 实现及默认参数。

4.2. 对隐写图像应用 ICA 检测

要发挥作用，独立成分分析要求我们观测到的混合信号数量至少与源信号数量相同。然而，在最符合实际的设置中，由于我们无法获取隐写模型，只能假设我们可以获取一张隐写图像——它是我们试图提取的两个信号（载体图像和载荷图像）的单一混合信号。

为了解决这一问题，我们可以借鉴第 3 节分析所得的见解。事实上，我们已经看到，基于可逆神经网络的隐写方案带来的嵌入变化，是将载荷子带中的不同信息添加到载体图像的子带中（见图 5）。由于隐写图像不同子带之间的混合方式各不相同，因此可以将它们视为 ICA 所需的不同混合信号。

此外，我们还可以通过仅选择特定子带来进一步改进这一方法。为了提取最相关的子带，我们建议对图像经离散小波变换得到的 12 个小波子带（每个颜色通道 4 个子带）进行主成分分析（PCA）。其依据是，PCA 中最显著的分量能够捕捉图像的主要结构，也就是载体图像的内容。另一方面，较不显著的分量则更可能包含嵌入过程引入的修改，也就是秘密图像。图 6 展示了这一点，其中可以看到，PCA 中较小的分量明显包含与载荷相关的信息。因此，执行 ICA 时，仅使用较小的 PCA 分量，而非全部分量，可以提取出视觉上更有意义的分量。

4.3. 检测方案

对隐写图像的小波分解中选定的 PCA 分量执行独立成分分析后，我们得到一对独立成分。如图 6 所示，它们包含不同的语义信息。在这些

分量上（分别记为 $\mathbf{c}_1$ 和 $\mathbf{c}_2$,），我们接着计算并拼接简单的统计特征，即其系数分布的前四阶矩：

$$\text{Feature} = \{\mu_1, \mu_2, \sigma_1, \sigma_2, \gamma_1, \gamma_2, \kappa_1, \kappa_2\}, \tag{3}$$

其中，μ_i、σ_i、γ_i 和 κ_i 分别表示分量 i的系数的均值、标准差、偏度和峰度。最后，使用这些特征训练高斯支持向量机，以进行隐写/载体分类。

图1总结了整体方案。首先对输入图像进行离散小波变换，然后通过主成分分析选出最相关的子带。接着，对这些子带进行独立成分分析，以提取两个独立成分。最后，计算这两个分量的系数分布的前四阶矩，并将其用作支持向量机分类器的输入特征。图中以一幅载体图像和一幅隐写图像为例，展示了各个步骤，以说明提取出的分量和特征之间的差异。

5. 实验与结果分析

5.1. 实验设置

我们提出以下实验设置来评估所提出的方法。在隐写方案方面，我们考虑五种不同的模型：HiNet [1], 、PRIS [2] 和 DeepMIH [3] 作为基于可逆神经网络的方案，以及 Baluja [4] 和 Weng 等人 [5]。请注意，除 PRIS 外，所有模型均使用公开提供的权重；PRIS 则使用官方代码库[2]进行训练。对于所有模型，我们使用 Guillaro 等人 [21],所用的 COCO 数据集图像 [20] 生成 2500 幅隐写图像，这些图像也构成了我们的载体图像基础。所有图像均为512×512 彩色图像。

为了确定要选择哪些 PCA分量，我们对每一对 PCA分量进行网格搜索，并选择在经验上对 HiNet 图像实现最佳检测性能的一对分量。使用第 9 和第 11 个分量可取得最佳结果，因此在后续实验中采用这两个分量。

所有分类器均采用 5 折交叉验证进行训练，以确保结果稳健，并使载体图像与隐写图像的数量保持平衡。

5.2. 所提方法的结果

现在，我们评估所提出方案的检测性能。结果见表1。可以看出，所提出的方法能够取得较高的检测准确率，尤其是在基于可逆神经网络的方案上，对 PRIS 的准确率最高可达 84.6%。不过，我们也可以看到，非基于可逆神经网络的方案可检测性较低。这可以解释为，这些方法直接处理像素，而不是离散小波变换系数。

[2]github.com/yanghangAI/PRIS

方法	HiNet	PRIS	DeepMIH	Baluja	Weng
准确率（%）	80.31	84.62	82.58	61.83	74.96
标准差（±%）	0.76	0.56	1.03	0.74	1.60

表1。所提出方案在不同图中图隐写方案上的检测准确率和标准差（%）。

方法	HiNet	PRIS	DeepMIH	Baluja	Weng
准确率（%）	99.02	99.96	99.92	80.06	99.64
标准差（±%）	0.29	0.05	0.07	0.85	0.15

表2。不同图像隐写方案上经典隐写分析（SRM + SVM）的检测准确率和标准差（单位：%）。

考虑到所提出方法的合成性和可解释性，这些结果颇为有趣，并凸显了此类隐藏方案的高可检测性。

5.3. 对经典隐写分析的鲁棒性

现在，我们使用经典隐写分析特征评估所研究隐写模型的可检测性，即采用空间富模型（SRM） [22] 结合支持向量机（SVM）分类器。预计该方法会牺牲部分输入特征的可解释性，以获得更高的检测准确率。

结果如表2所示。可以看到，使用该方法能够高度检测出所有所研究的隐写方案：除Baluja方法外，所有方法的检测准确率均超过99%；而Baluja方法的检测准确率也达到80%。这进一步凸显了此类图中藏图隐写方案的高度脆弱性。

6. 结论与展望

在本研究中，我们考察了图像内图像隐写术方案的脆弱性，尤其是基于可逆神经网络（INN）的方案。我们提出了一种新颖的隐写分析方法，该方法基于图像经过主成分分析（PCA）后，其离散小波变换（DWT）系数的独立成分分析（ICA）分布的前四阶矩。结果表明，这种高度综合且具有高度可解释性的方法能够实现良好的检测性能，凸显了此类隐写图像的高可检测性。此外，我们还表明，经典隐写分析也能对这类隐写图像实现较高的检测率，进一步强调了其脆弱性。

我们希望这项研究能推动未来设计出更安全的图中藏图隐写方案。例如，在隐藏方案中引入密钥，并结合用于衡量可检测性的损失函数，将是朝这一方向迈出的第一步。

7. 致谢

本研究由法国政府通过 France2030 计划提供资助，该资助由法国国家研究署管理，项目编号为 ANR22-PECY-0011。

8. 参考文献

[1] Junpeng Jing、Xin Deng、Mai Xu、Jianyi Wang和Zhenyu Guan，“Hinet：基于可逆网络的深度图像隐藏”，载于*IEEE/CVF*国际计算机视觉会议论文集，2021年。

[2] Hang Yang、Yitian Xu、Xuhua Liu和Xiaodong Ma，“Pris：面向图像隐写的实用鲁棒可逆网络”，《人工智能工程应用》，第133卷，2024。

[3] Zhenyu Guan、Junpeng Jing、Xin Deng、Mai Xu、Lai Jiang、Zhou Zhang和Yipeng Li，“Deepmih：用于多图像隐藏的深度可逆网络”，《*IEEE*模式分析与机器智能汇刊》，第45卷，2022年。

[4] Shumeet Baluja，“在图像中隐藏图像”，*IEEE*模式分析与机器智能汇刊，第42卷，2019年。

[5] Xinyu Weng、Yongzhi Li、Lu Chi 和 Yadong Mu，“基于时间残差建模的大容量卷积视频隐写”，载于*2019*年国际多媒体检索会议论文集，2019年。

[6] Jiren Zhu、Russell Kaplan、Justin Johnson 和 Li Fei- Fei，“Hidden：利用深度网络隐藏数据”，载于欧洲计算机视觉会议（*ECCV*）论文集，2018年。

[7] Pin Wu、Yang Yang 和 Xiaoqiang Li，“Stegnet：利用深度卷积网络实现超大图像隐写容量”，《未来互联网》，第10卷，2018年。

[8] Shao-Ping Lu、Rong Wang、Tao Zhong 和 Paul L Rosin，“基于可逆神经网络的大容量图像隐写”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年。

[9] Youmin Xu、Chong Mou、Yujie Hu、Jingfen Xie 和 Jian Zhang，“鲁棒可逆图像隐写”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年。

[10] Jiwen Yu、Xuanyu Zhang、Youmin Xu 和 Jian Zhang，“Cross：扩散模型实现可控、鲁棒且安全的图像隐写”，神经信息处理系统进展，第 36 卷，2023年。

[11] Rongxuan Peng、Xianbo Mo、Shunquan Tan、BinLi 和 Jiwu Huang，“面向基于深度学习的图中图模型的无密钥提取框架”，发表于*ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2024。

[12] Shumeet Baluja，“藏图于明处：深度隐写术”，神经信息处理系统进展，第30卷，2017。

[13]Laurent Dinh、David Krueger 和 Yoshua Bengio，“NICE：非线性独立成分估计”，*arXiv*预印本 *arXiv:1410.8516*，2014。

[14] Claude E Shannon，“保密系统通信理论”，贝尔系统技术期刊，第28卷，1949。

[15] Christian Jutten 和 Jeanny Herault，“盲源分离，第一部分：基于神经仿生架构的自适应算法”，信号处理，第24卷，1991。

[16] Ralph Linsker，《感知网络中的自组织》，《计算机》，第 21 卷，1988 年。

[17] Jean-Franc◌ois Cardoso 和 Antoine Souloumiac，《非高斯信号的盲波束形成》，载于 *IET F* 辑论文集（雷达与信号处理）。IET，1993 年，第 140 卷。

[18] Francis R Bach 和 Michael I Jordan，《核独立成分分析》，《机器学习研究杂志》，第 3 卷，2002 年。

[19] Aapo Hyv¨arinen 和 Erkki Oja，《独立成分分析：算法与应用》，《神经网络》，第 13 卷，2000 年。

[20] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick，《Microsoft COCO：上下文中的常见物体》，载于 欧洲计算机视觉会议。施普林格，2014 年。

[21]Fabrizio Guillaro、Giada Zingarini、Ben Usman、Avneesh Sud、Davide Cozzolino 和 Luisa Verdoliva，《一种用于更通用的 AI 生成图像检测的无偏训练范式》，载于计算机视觉与模式识别会议论文集，2025。

[22] Jessica Fridrich 和 Jan Kodovsky，《用于数字图像隐写分析的富模型》，*IEEE*信息取证与安全汇刊，第 7 卷，2012。