隐写和隐写分析前沿技术

杨榆 副教授

北京邮电大学 网络空间安全学院

CONTENTS

01 隐写和隐写分析发展阶段

02 基于深度学习的隐写前沿技术及挑战

03 基于深度学习的隐写分析前沿技术及挑战

04 基于深度学习的信息隐藏技术发展趋势

01

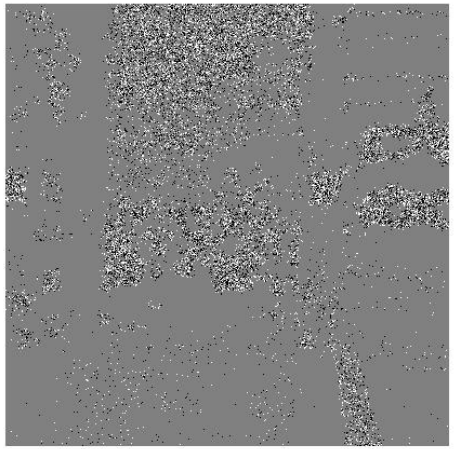
隐写和隐写分析发展阶段

背景、应用、发展阶段

隐写术

自然载体

携密载体



差异图

隐写算法

010110…

隐写术

隐写术，译自英文Steganography。Steganography一词来自于希腊词根steganos和graphie，前者指有遮盖物的，后者指写。因此，Steganography的字面意思即为隐写。

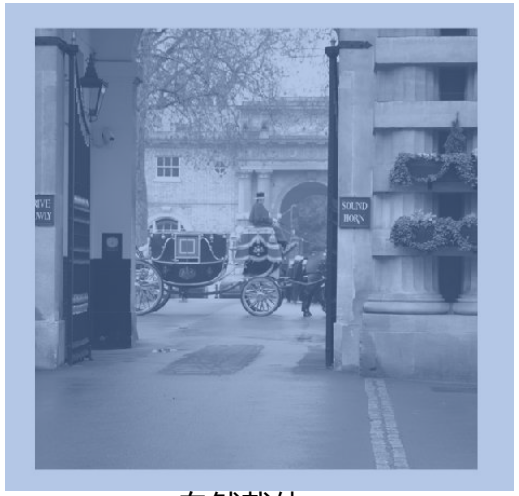
古典隐写术

隐写术的应用，最早可追溯到希腊历史学家希罗多德（Herodotus）描述的一则公元480年前的历史故事：希斯提乌斯(Histiaeus)巧妙地利用奴隶传递秘密信息。早期的隐写术，既包括实物隐藏、标记法、显影剂和微缩胶片等技术型隐写术，也包括藏头诗、卡登格子和乐谱法等语言学型和艺术作品型隐写技术，与其说是一门技术，毋宁说是一门艺术。

现代隐写术

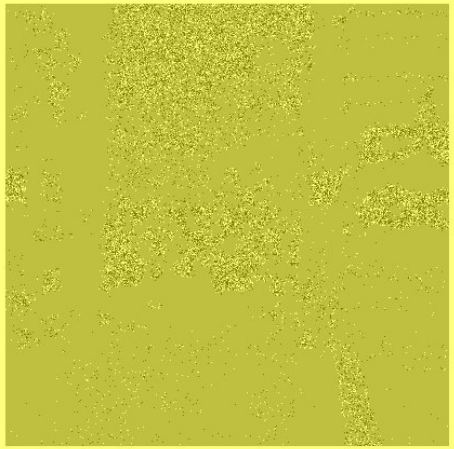
现代隐写术是研究在载体中不为人察觉地隐藏入秘密信息的理论和技术，它综合运用了包括数字信号处理、编码理论、概率论和感知心理学理论等等在内的多门学科的理论和技术，是信息隐藏技术的一个重要分支。

隐写分析



自然载体

携密载体



差异图

隐写算法

010110…

隐写分析

隐写分析，Steganalysis，研究破解隐写术的理论与技术。是隐写技术的对抗技术。

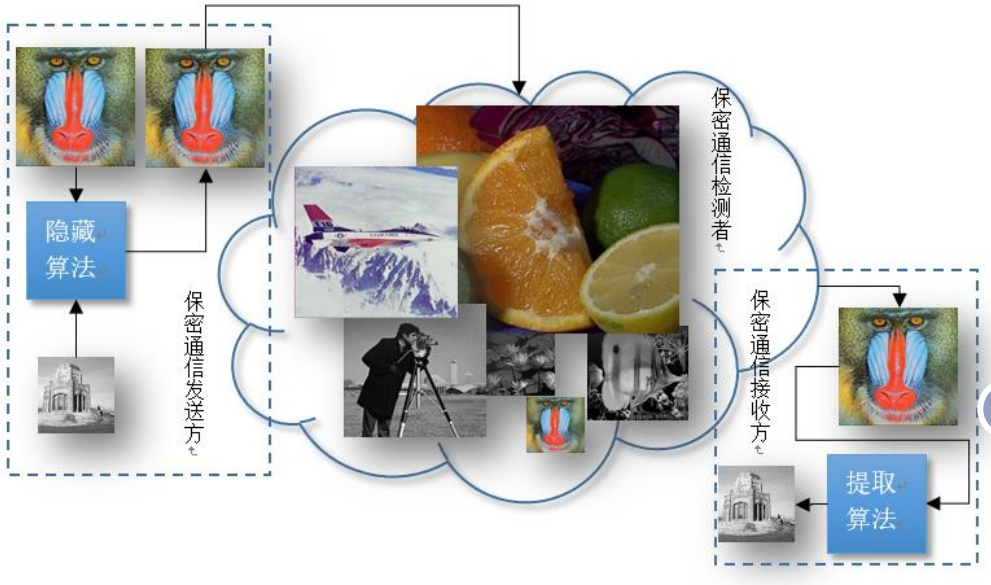
隐写分析的目标

隐写分析的目标，从易到难，分为三个层次，分别是：判别隐秘信息的有无，分析隐秘信息的长度，和提取隐秘信息。第一层次目标，即判别图像或音视频等载体是自然载体还是携密载体，也被称为定性隐写分析。

定性隐写分析

隐写分析领域主要工作均集中于定性隐写分析，从早期的软件特征分析到统计特征分析技术，发展到现在高维特征分析技术，隐写分析运用了统计学，信号检测和估计，以及信息论等等多种理论，围绕载体的特征描述和异常检测展开研究，与隐写术形成攻防交替上升的态势。

隐写术的应用



基于隐写术的保密通信

发送方使用隐写工具把秘密信息（图片）隐藏到载体图像中，生成与原图无视觉差异的隐写图。隐写图经网络传送到接收方。接收方使用相同算法提取出秘密信息，实现保密通信。检测者需要识别网络图片中，哪些图片是自然的，哪些图片是经过隐写的。

保密通信

隐写术和隐写分析自诞生，即主要与保密通信密切相关。早在2006年就出现了美国起诉俄罗斯间谍的事件，控方指出对方使用了隐写技术。安娜·查普曼是俄罗斯对外情报局的女间谍。安娜在莫斯科人民友谊大学取得经济学硕士学位。毕业后经商，2004年离俄，游历了西方多国。在商人身份掩护下，她与各色人等结交，从而套取敏感信息，再通过隐形墨水将这些包括核武器情报在内的美国机密源源不断地送达俄罗斯。

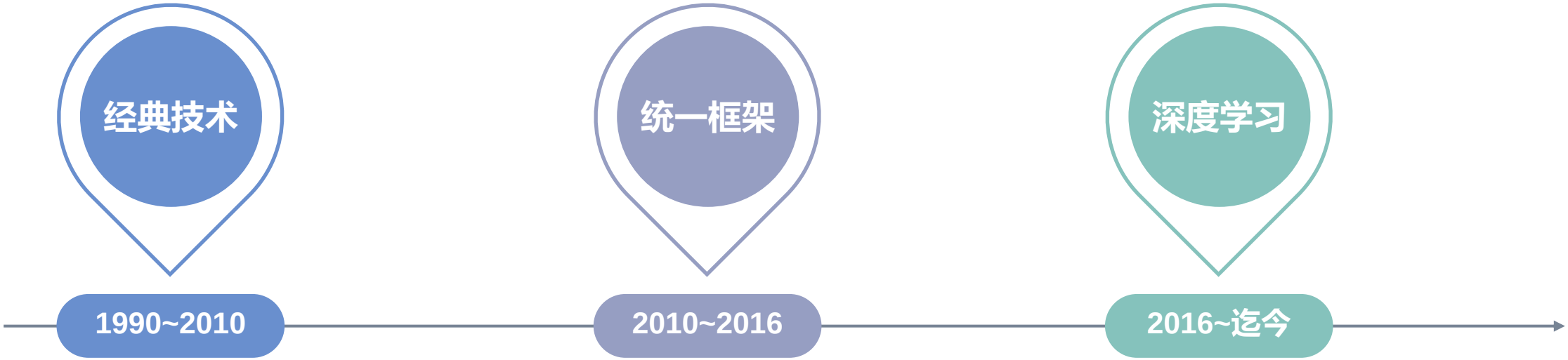
规避安防软件

迈克菲（McAfee）报告指出，黑客对隐写技术的运用从两个方面展开，第一：藏恶意内容（指令和代码等），规避安防软件扫描恶意代码。代表性恶意软件为TeslaCrypt，使用数字隐写术将恶意程序或指令藏包括图像和网络流量。

泄露数据

迈克菲（McAfee）报告指出，黑客对隐写技术的运用从两个方面展开，第二：藏其窃取的有价值的数据，造成数据泄露。例如，Duqu恶意软件。Duqu从受害者的系统中收集信息，而后将数据加密并嵌入到JPEG文件中，最后将其作为图像文件发送到控制服务器。

发展阶段



“百家争鸣”，不同学科背景的研究者提出了各自的解决方案，没有突出的、占绝对优势的领导技术或算法。

各类技术齐头并进的研究局面在2010年前后被打破了。隐写和隐写分析技术双双进入统一框架（自适应）的阶段。

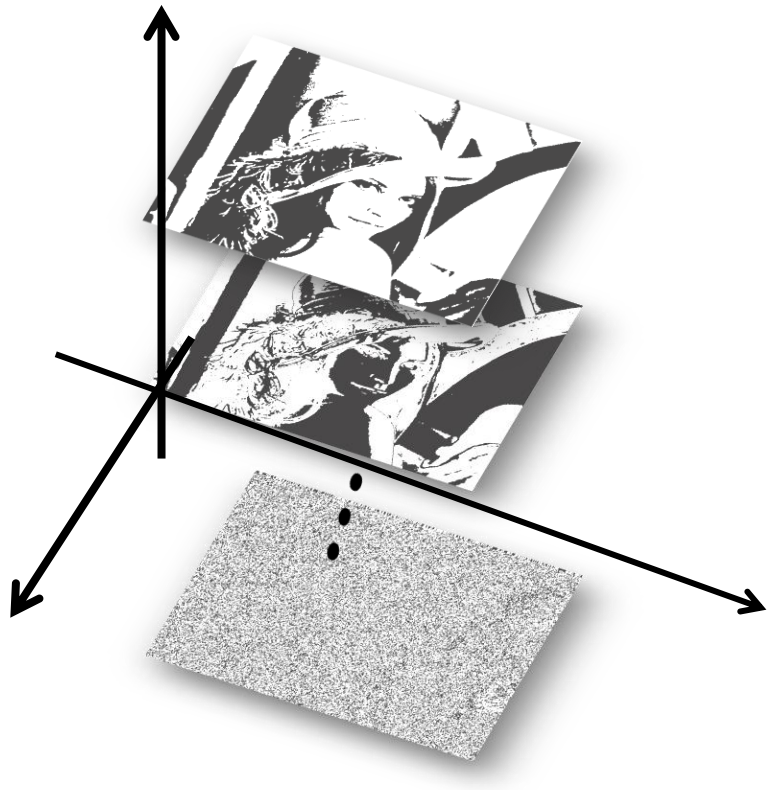
最早从2014年，进入到深度学习阶段。算法进一步降低了对手工设计和专家经验的依赖，转而从数据中自主学习。

经典技术阶段

隐写技术特点可用“百家争鸣”来描绘，从空域到变换域，从编码理论到概率论，不同学科背景的研究者提出了各自的解决方案，均有各自的优势，没有突出的、占绝对优势的领导技术或算法。

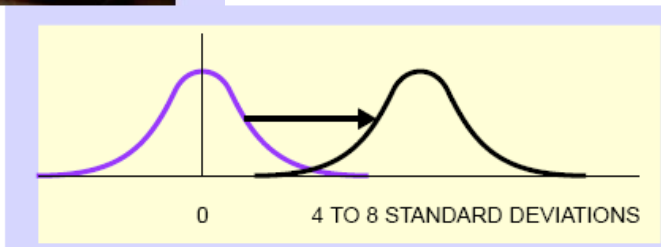
隐写分析技术主要为专用隐写分析技术，显著特征是一一对应，一种隐写[illegible]检测特征，仅适用于

-



经典空[illegible]

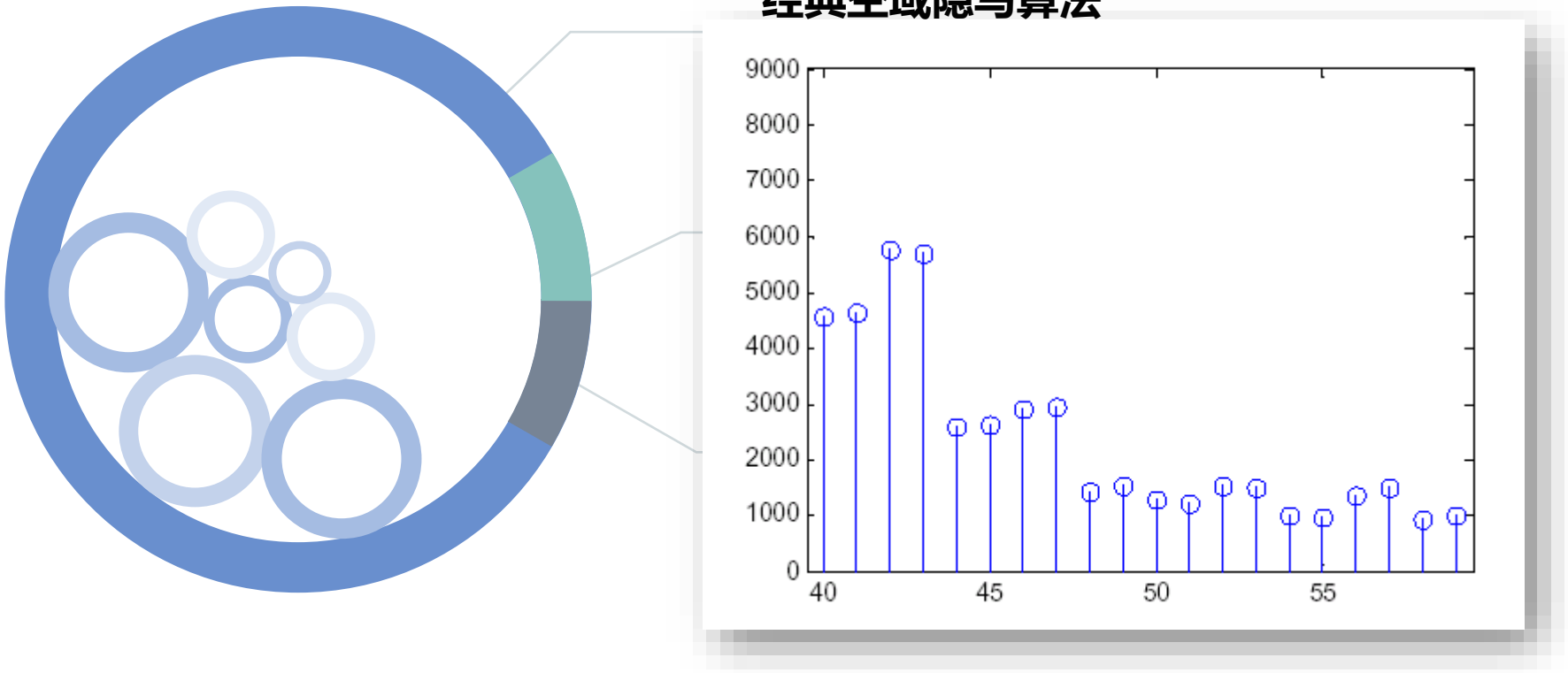
- LSB, IML
- Patchwo



经典技术阶段

隐写技术特点可用“百家争鸣”来描绘，从空域到变换域，从编码理论到概率论，不同学科背景的研究者提出了各自的解决方案，均有各自的优势，没有突出的、占绝对优势的领导技术或算法。

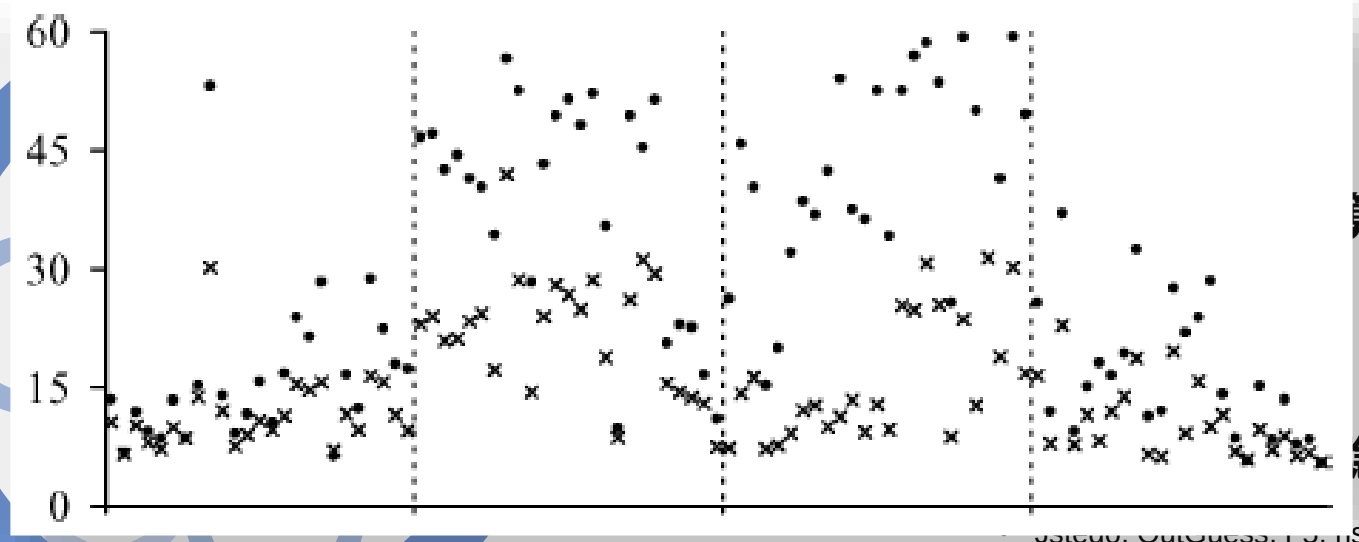
隐写分析技术主要为专用隐写分析技术，显著特征是一一对应，一种隐写分析方法，更具体来说，是一套隐写检测特征，仅适用于一种隐写算法。



经典技术阶段

隐写技术特点可用“百家争鸣”来描绘，从空域到变换域，从编码理论到概率论，不同学科背景的研究者提出了各自的解决方案，均有各自的优势，没有突出的、占绝对优势的领导技术或算法。

隐写分析技术主要为专用隐写分析技术，显著特征是一一对应，一种隐写分析方法，更具体来说，是一套隐写检测特征，仅适用于一种隐写算法。

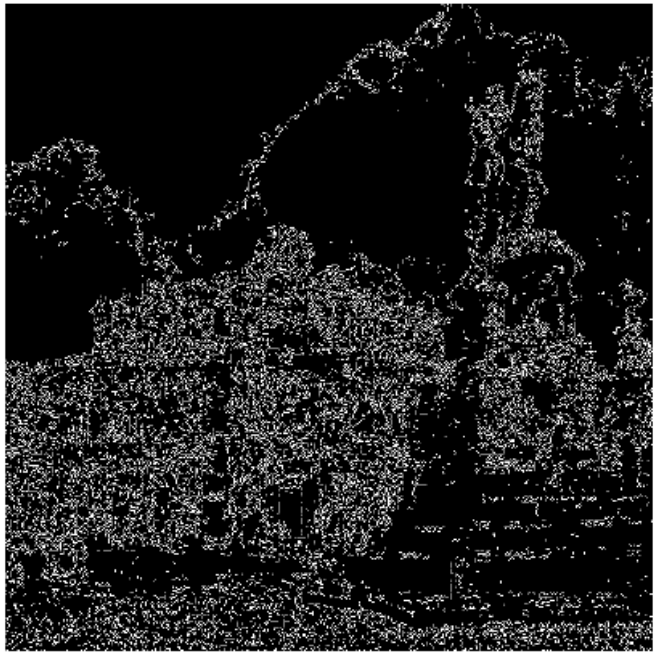


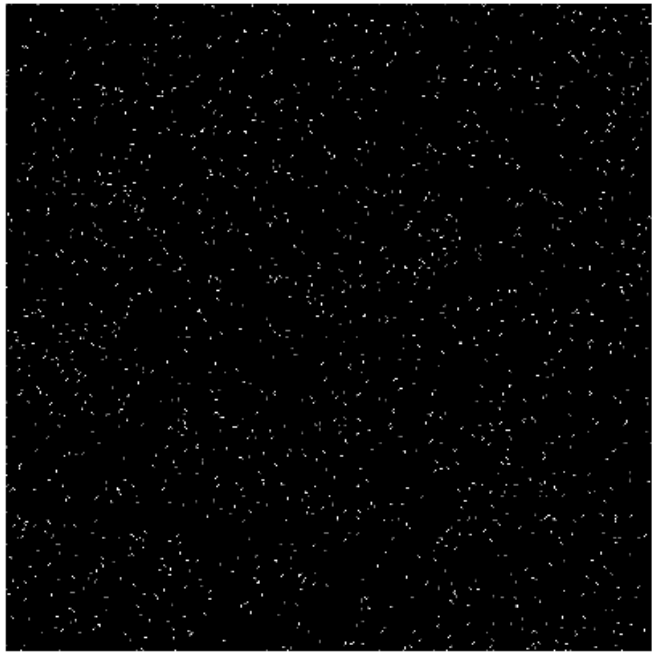
算法

隐写分析算法

- Jstego, OutGuess3, F5, nsF5, 基于格式隐写或颜色方案….
- 基于直方图估计的JPEG隐写分析，PDF(Probability Density Function)矩系列，直方图特征函数HCF（Histogram Characteristic Function）系列……

自然图像及使用WOW, HUGO BD, S-UNIWARD, EA, and LSB matching算法生成的隐写图像修改区域

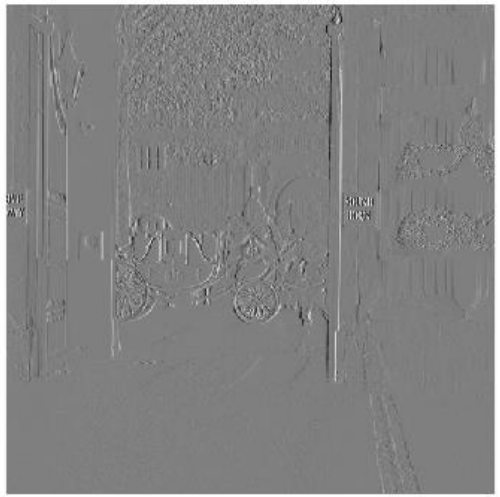




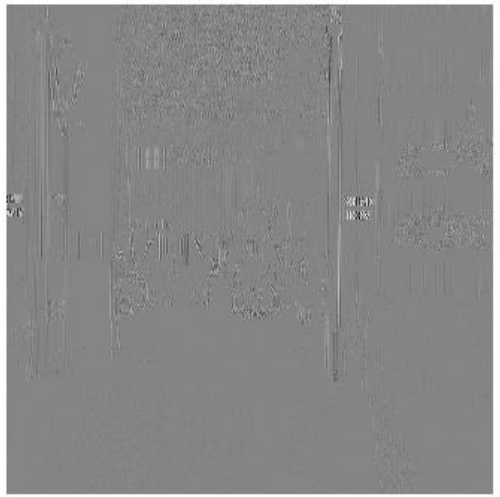
杂多样的
提高了隐

式设计方
函数的设
新的隐写

自然图像

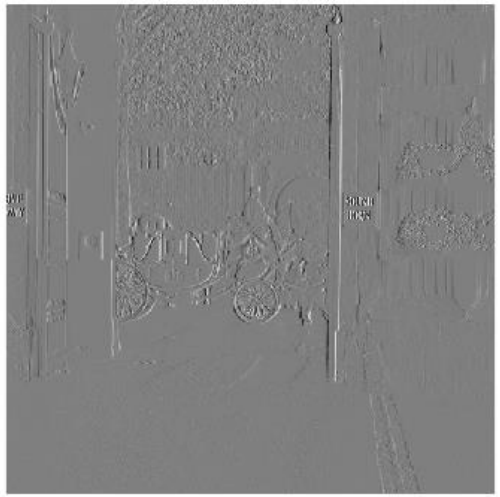


自然图像一阶残差

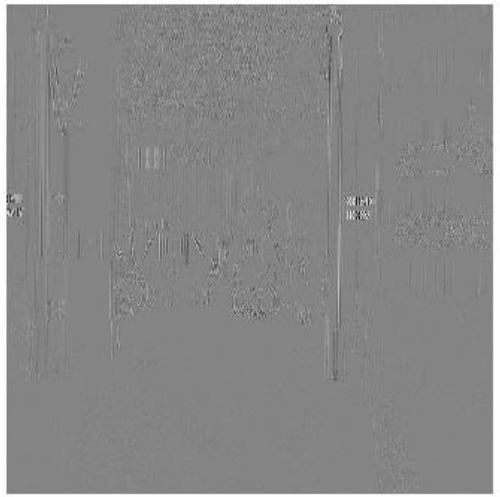


自然图像三阶残差

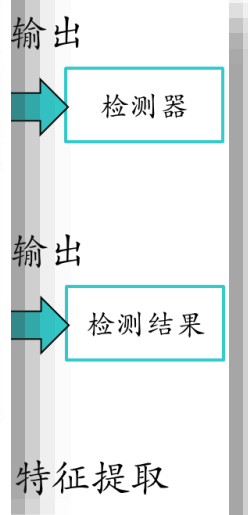
携密图像



携密图像一阶残差



携密图像三阶残差



统一框架阶段

各类技术齐头并进的研究局面在2010年前后被打破了。隐写和隐写分析技术双双进入统一框架（自适应）的阶段。

隐写术分析研究领域，以SPAM（Subtractive Pixel Adjacent Matrix，像差邻接矩阵）为起始，以SRM（Spatial Rich Model，空域富模型）为巅峰，进入“通用隐写分析阶段”。

$$R_{ij} \leftarrow trunc_T\left(round\left(\frac{R_{ij}}{q}\right)\right)$$

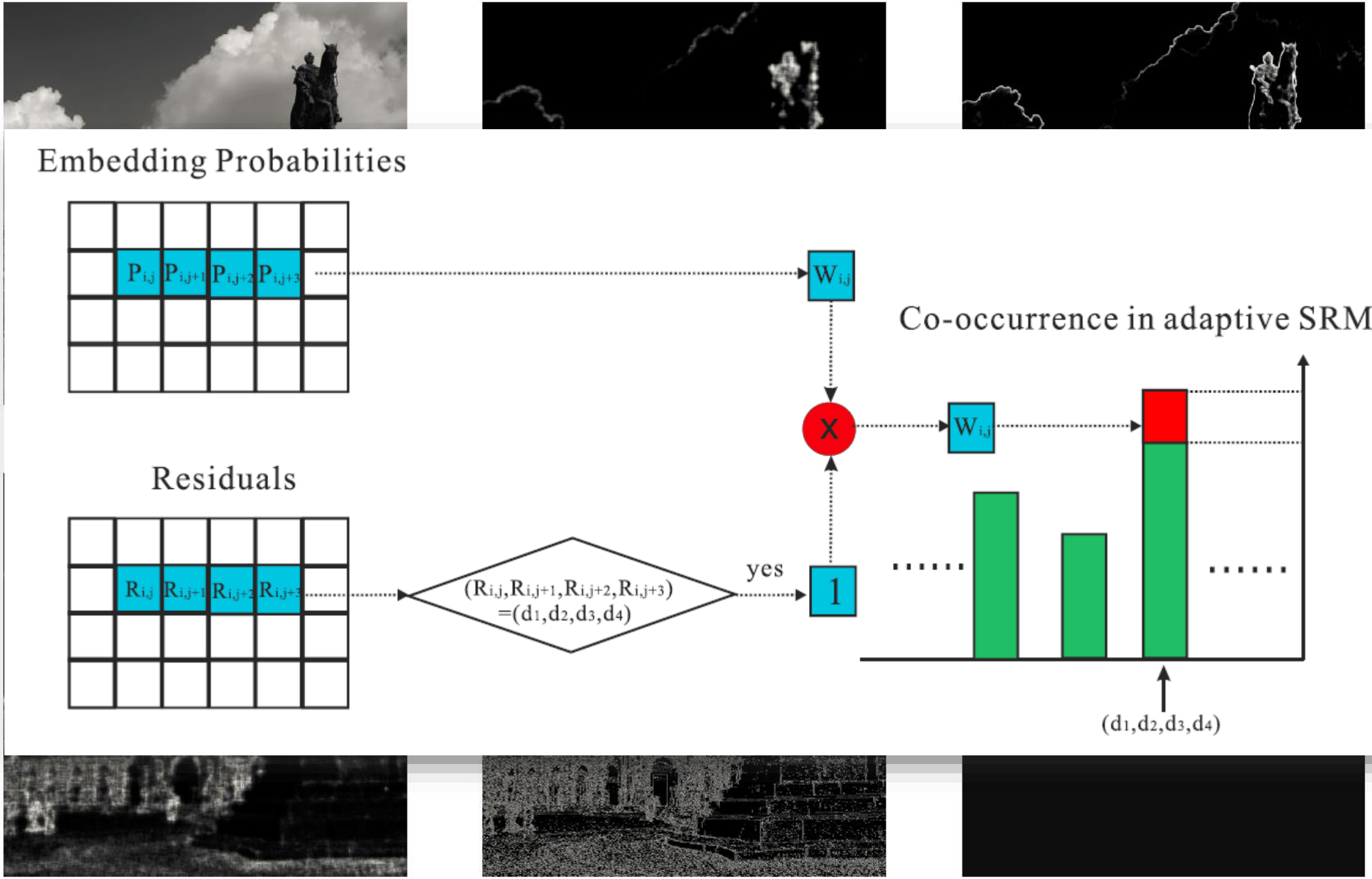
控制特征维度

放大隐写噪声

SRM

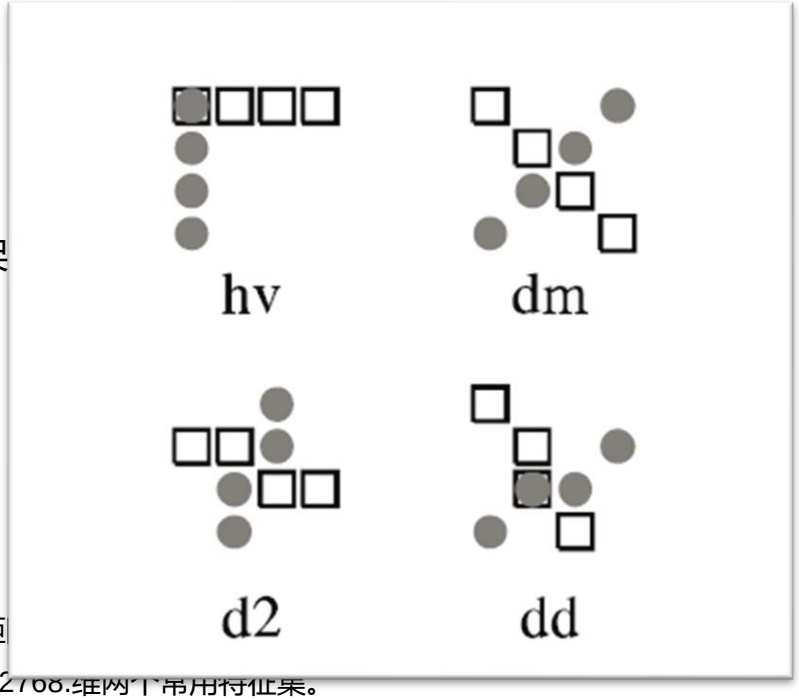
- 残差分析，量化，截断，共生矩阵。
- 78个高通滤波器， 12753维和32768.维两个常用特征集。

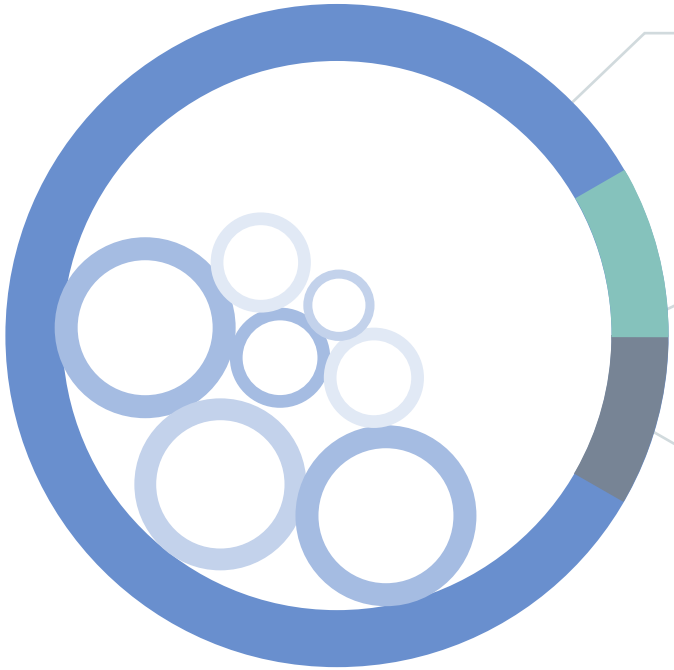
$$\mathbf{C}_{\mathbf{d}}^{(\mathrm{h})} = \frac{1}{Z}\left|\{(R_{ij}, R_{i,j+1}, R_{i,j+2}, R_{i,j+3}) \mid R_{i,j+k-1} = d_k, k = 1, \ldots, 4\}\right|$$


Embedding Probabilities
$P_{i,j}$ $P_{i,j+1}$ $P_{i,j+2}$ $P_{i,j+3}$
$W_{i,j}$
Co-occurrence in adaptive SRM
X
$W_{i,j}$
Residuals
$R_{i,j}$ $R_{i,j+1}$ $R_{i,j+2}$ $R_{i,j+3}$
$(R_{i,j},R_{i,j+1},R_{i,j+2},R_{i,j+3})$
$=(d_1,d_2,d_3,d_4)$
yes
1
(d_1,d_2,d_3,d_4)

统一框架阶段

各类技术齐头并进的研究局面在2010年前后被打破了。隐写和隐写分析技术双双进入统一框架
隐写术分析研究领域，以SPAM（Subtractive Pixel Adjacent Matrix，像差邻接矩阵）为起始，
Model，空域富模型）为巅峰，进入“通用隐写分析阶段”。





SRM

- 残差分析，量化，截断，共生矩
- 78个高通滤波器， 12753维和32768.维两个常用特征集。

maxSRMd2

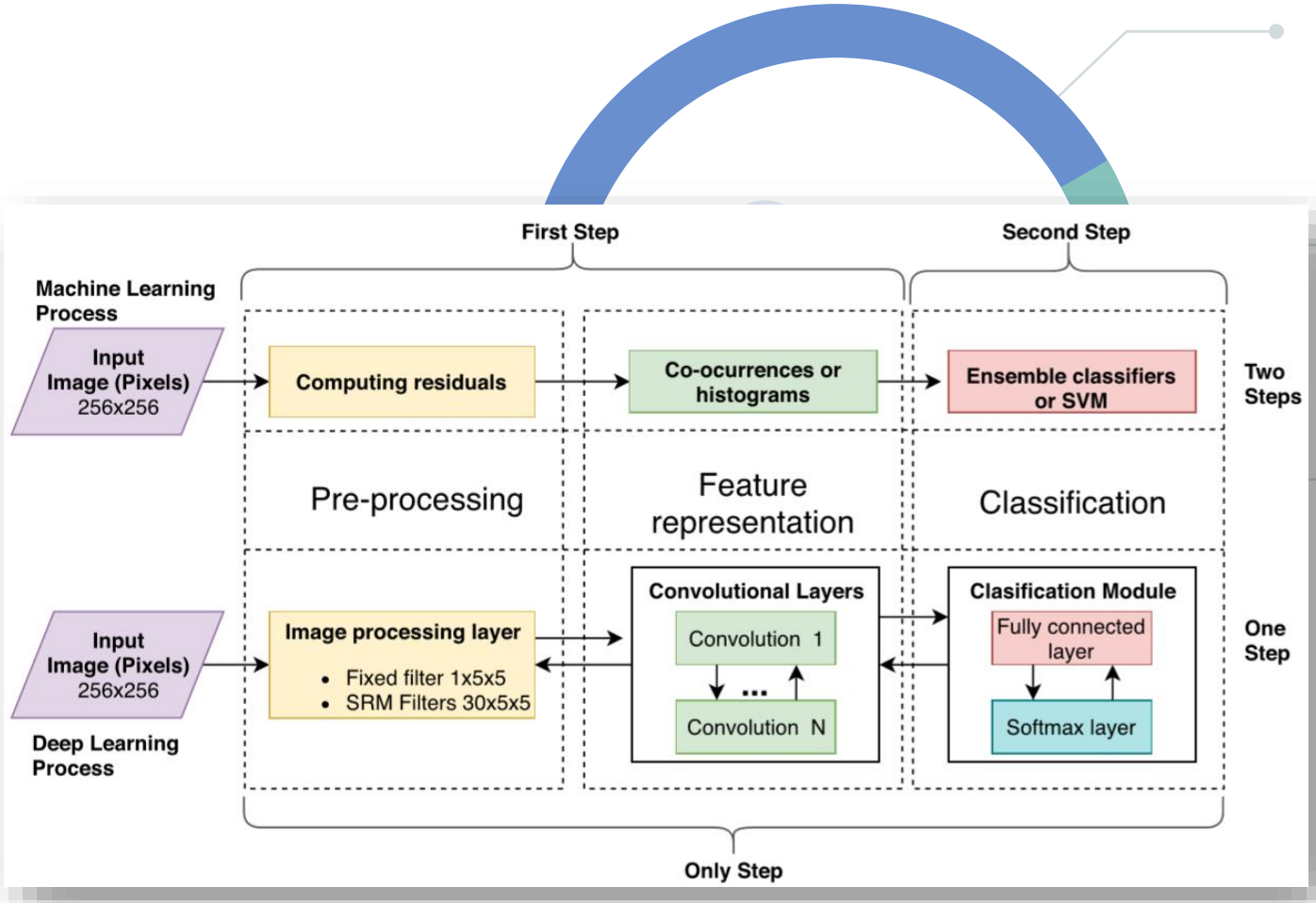
- 通过概率图区分纹理复杂度不同的区域的异常的贡献。.
- 计算共生矩阵时，引入对角线方向等多种扫描方式。

主要贡献

- 隐写分析应基于残差展开；
- 利用选择通道提高自适应隐写分析准确性；
- 给出通用隐写分析框架。.

深度学习阶段

最早从2014年，形势发生变化，隐写和隐写分析进入到深度学习阶段。这时，隐写算法和隐写分析的设计进一步降低了对手工设计和专家经验的依赖，转而从数据中自主学习。



基于深度学习的隐写术

- 基于对抗生成网络的自适应隐写算法；
- 基于生成对抗网络的载体生成隐写算法；.

基于深度学习的隐写分析技术

- 特征提取对手工设计和专家经验的依赖度进一步降低；.
- 残差计算、特征提取和分类整体优化；.

主流算法

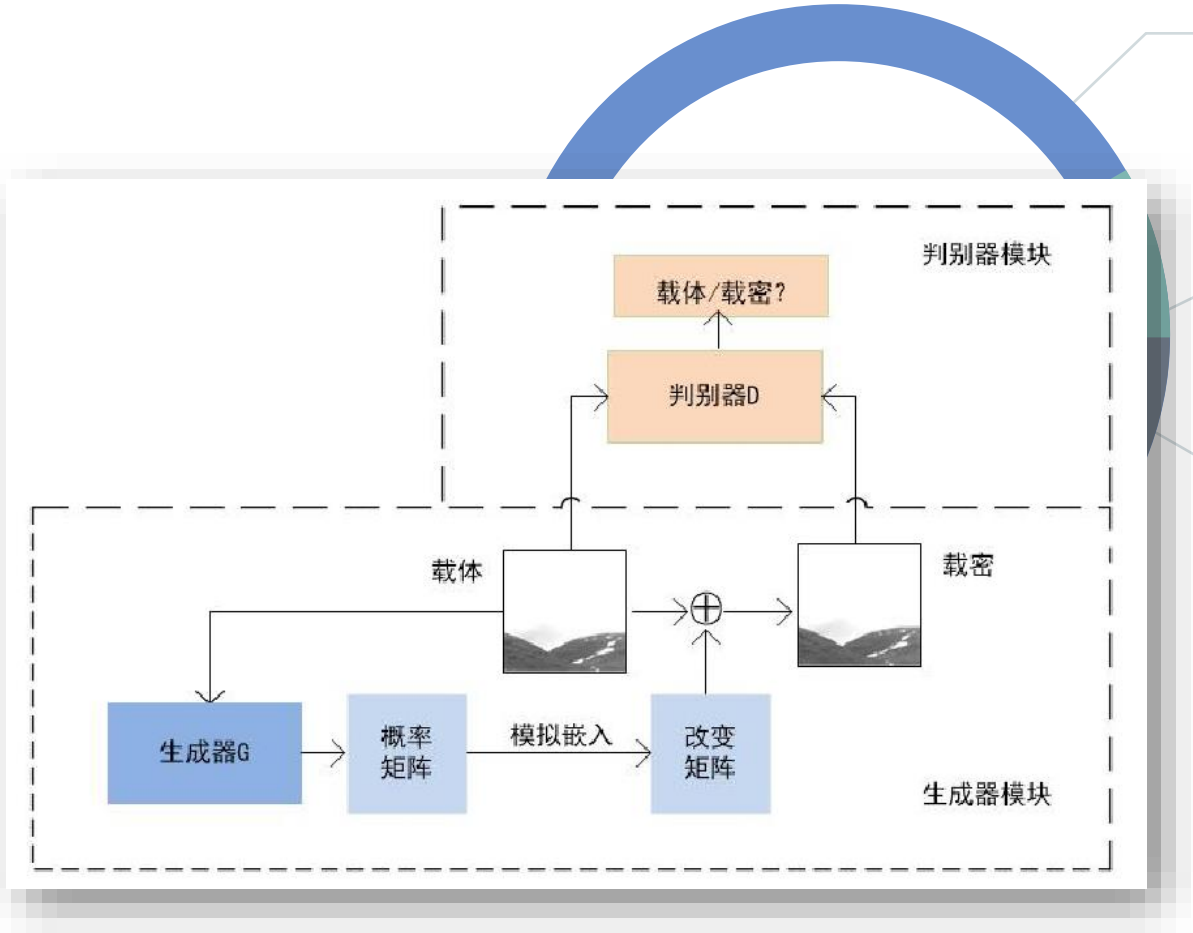
- ASDL-GAN,DCGAN,SGAN,SSGAN和HiDDen.
- QianNet or GNCNN，XuNet，YeNet， Yedroudj-Net，ZhuNet，SRNet，和WiserNet
- SiastegNet、LWENet

02

基于深度学习的隐写前沿技术及挑战

基于生成对抗网络的自适应隐写算法

ASDL-GAN模型的本质是通过网络搜索合适的嵌入位置，再利用传统编码的方法嵌入秘密信息。虽然ASDL-GAN性能没有超越传统自适应隐写算法，但引入了自主学习和对抗理念，改变了传统的经验设计方式，推进了信息隐藏向高安全性进一步发展。



生成器G

- 生成模块用生成器G输出的概率矩阵P和秘密信息；
- 模拟嵌入得到改变矩阵M；

改变矩阵M

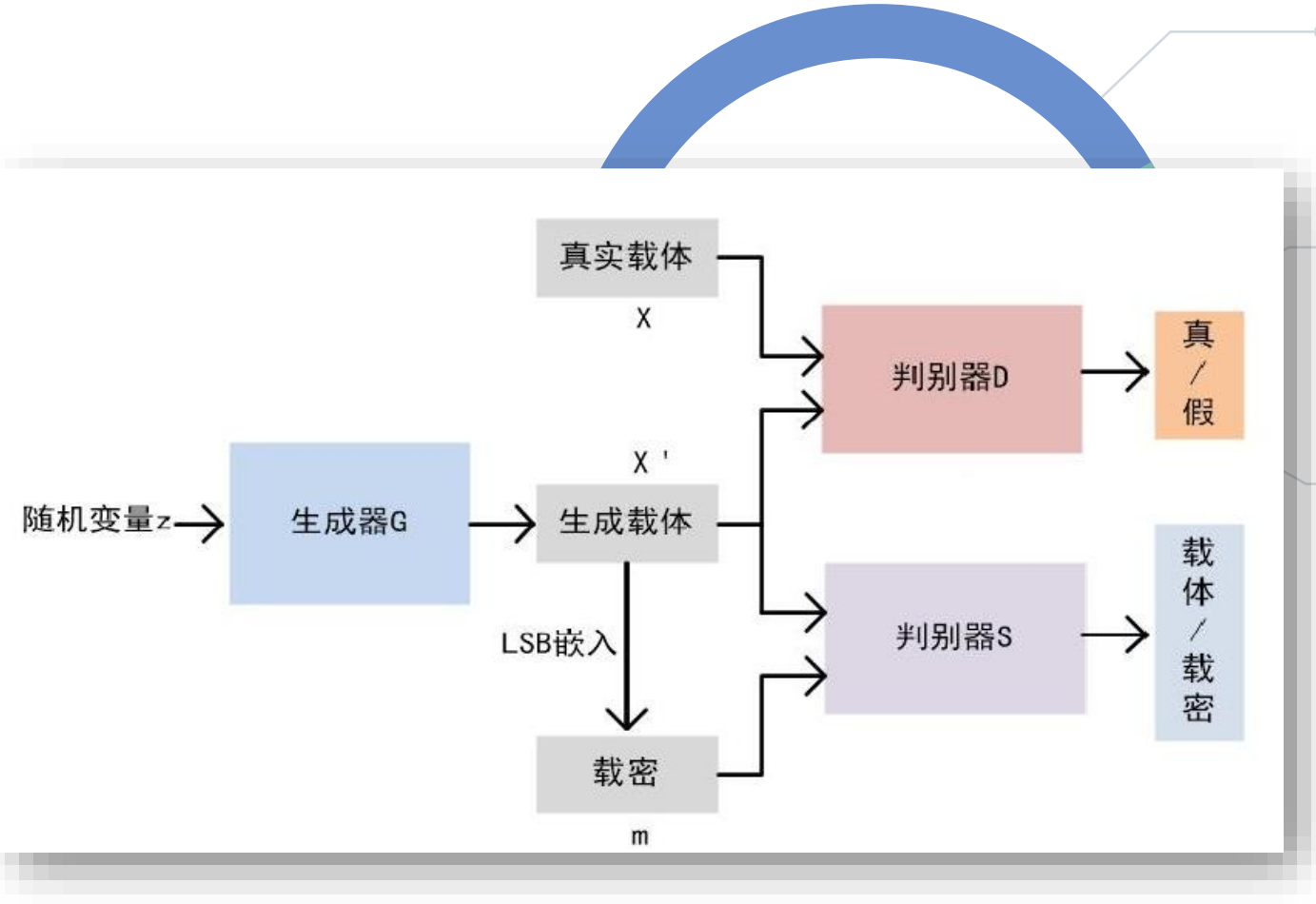
- M中每个元素取0,-1,+1三值之一，指示了载体对应像素的改变量；.
- M与载体图像相加就可以得到载密（隐写）图像；.

判别器D

- 判别器模块则识别载体和载密图像，即隐写分析；.
- 根据判别器D的分类结果，生成器和判别器能够交替训练，更新网络参数；

基于生成对抗网络的载体生成隐写算法

当前隐写分析模型都是通过学习训练样本集合获得，样本集的数据特性直接影响了隐写分析模型的准确性能，因此如果生成与常用训练样本集合不同的载体来嵌入秘密信息，就能降低隐写分析模型的检测能力。DCGAN首先设计了载体生成网络，模型包括生成器G、判别器D和隐写分析器S三部分。



生成器G

- 不需要载体图像；
- 生成器G根据随机变量z产生尽可能真实的图像；

判别器S

- 生成载体经隐写算法处理产生携密生成载体；
- 判别器S用于分辨自然和携密生成载体；

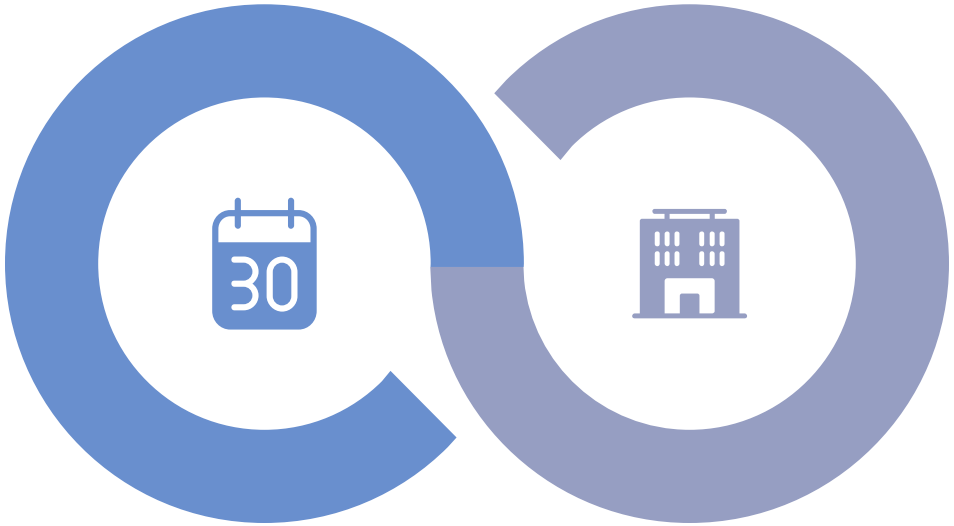
判别器D

- 判别器D分辨真实图像与生成器G产生的图像；
- 判别器S和D与生成器G对抗学习，使其生成的载体尽可能接近真实图像，且在隐写后有更强的抗隐写分析的安全性；

基于深度学习的隐写技术挑战

基于GAN的自适应隐写算法

- 使用生成器模拟传统自适应隐写算法的嵌入过程；
- 包括ASDL-GAN在内的多种算法的模拟质量不够高，
- 其产生的载密图像，安全性上仍然弱于S-UNIWARD等传统自适应隐写算法.



基于GAN的载体生成隐写算法

- 使用生成器产生载体，
- 生成的载体的统计特性非常接近于自然载体，难以被判别器识别，
- 但生成的载体图像往往包含语义扭曲，
- 易引起怀疑、难以逃过人工检查。

03

基于深度学习的隐写分析前沿技术及挑战

基于深度学习的隐写分析技术

QianNet，也称**GNCNN**，为首个带监督学习的CNN网络，其结构也成为后续网络蓝本。QianNet遵从残差分析规则，用高通滤波器预处理图像以强化隐写噪声、抑制图像内容。QianNet的性能已经与SRM等传统的特征手工设计隐写分析算法性能相当。

$$K_{kv} = \frac{1}{12}\begin{pmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{pmatrix}$$

预处理

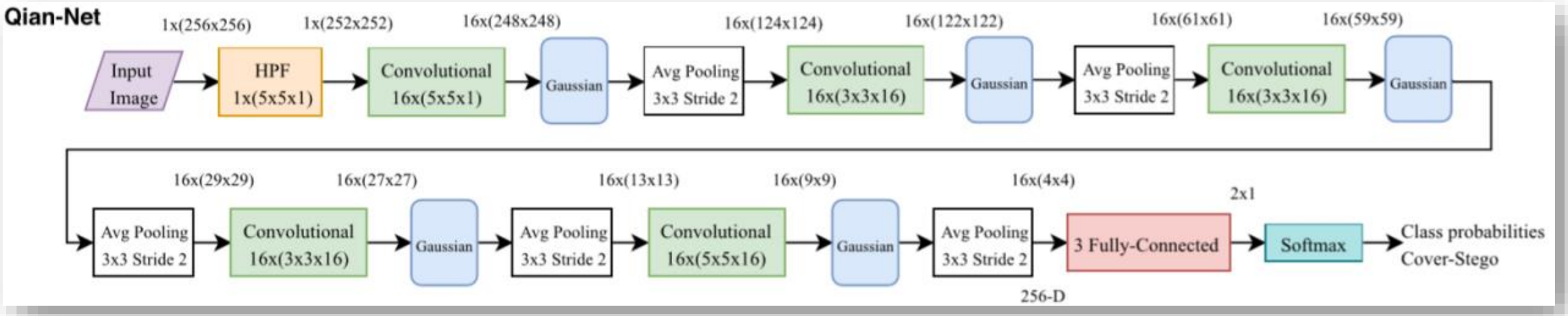
- 使用高通滤波器获取残差；
- HPF核函数根据启发式方法选取；

特征提取

- 5个卷积层，每层包含16个卷积核、一个高斯激活函数和一个平均池化；. $f(x) = e^{-\frac{x^2}{\sigma^2}}$

分类器

- 3层全连接神经网络加softmax；.
- 针对HUGO、WOW和S-UNIWARD算法，隐写分析结果优于SPAM，次于SRM；



基于深度学习的隐写分析技术

YeNet有两大贡献。首先，YeNet放弃了使用传统高通滤波器预处理图像来强化隐写噪声的做法，选用了一套SRM算法的模板（滤波器核）来初始化滤波器。这些滤波器也用于抑制图像内容，而且作为网络的一部份，也是可学习的，从而保证滤波器核能够被一并优化。第二，YeNet将选择通道技术加入了网络，确保不同纹理特性区域的残差得以区别对待，抑制了低隐写率情况下的弱隐写噪声被“稀释”的程度。结合其他深度学习技术的运用，YeNet在隐写分析方面取得较大成功。

预处理

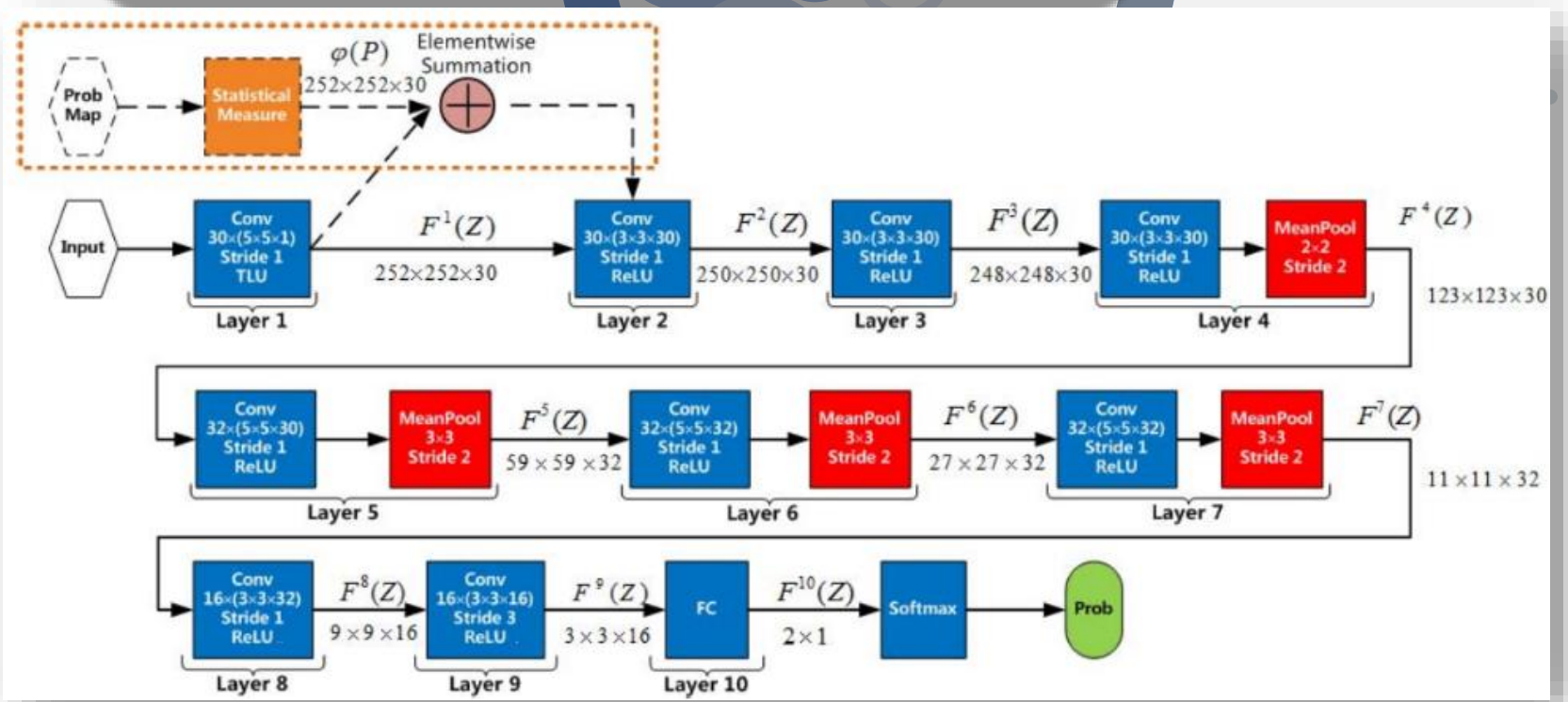
- 用SRM滤波器核初始化滤波器核；
- 用截断线性单元(Threshold Logic Unit, TLU)激活函数；

$$f(x)=\begin{cases}-T, x<-T\\ x, -T\le x\le T\\ T, x>T\end{cases}$$

特征提取

- 8个卷积层，各层卷积核数量有30、32和16三种类别，配合ReLU激活函数以及平均池化；

$$f(x)=\begin{cases}0, & x<0\\ x, & x\ge 0.\end{cases}$$



选择通道

- 概率图加权残差，在学习过程中，加权效果经过了特征提取各层卷积核和激活函数的作用，并逐层传递和累积。
- 推导分析可得，概率图在网络中产生的效果可表示为（P为概率图，K为残差滤波核。）：$\varphi(P)=P*|K|$

Algorithm	Payload (bpp)	SRM (P_E)	TLU-CNN (P_E)	maxSRMd2 (P_E)	SCA-TLU-CNN (P_E)
WOW	0.05	0.4551	0.3850	0.3810	**0.3450**
	0.1	0.4066	0.3000	0.3163	**0.2442**
	0.2	0.3228	0.1982	0.2325	**0.1691**
	0.3	0.2633	0.1394	0.1918	**0.1229**
	0.4	0.2127	0.1109	0.1536	**0.0959**
	0.5	0.1800	0.0938	0.1331	**0.0906**
S-UNIWARD	0.05	0.4641	0.4200	0.4316	**0.4000**
	0.1	0.4232	0.3350	0.3806	**0.3220**
	0.2	0.3437	0.2540	0.2999	**0.2224**
	0.3	0.2798	0.1772	0.2542	**0.1502**
	0.4	0.2260	0.1410	0.2136	**0.1281**
	0.5	0.1848	0.1003	0.1732	**0.1000**
HILL	0.05	0.4765	0.4150	0.4409	**0.4000**
	0.1	0.453	0.3560	0.3894	**0.3380**
	0.2	0.3811	0.2761	0.3226	**0.2538**
	0.3	0.3236	0.2145	0.2804	**0.1949**
	0.4	0.2818	0.1782	0.2410	**0.1708**
	0.5	0.2363	0.1561	0.2115	**0.1305**

基于深度学习的隐写分析技术

2019年，Boroumand等提出一个基于深度残差网络的隐写分析模型（**DRNM**，Deep Residual Network Model），被称为SRNet。首次摒弃了通过高通滤波器捕获隐写残差特征图的预处理阶段，在空域和变换域隐写分析任务中取得较好的结果。

三个功能块

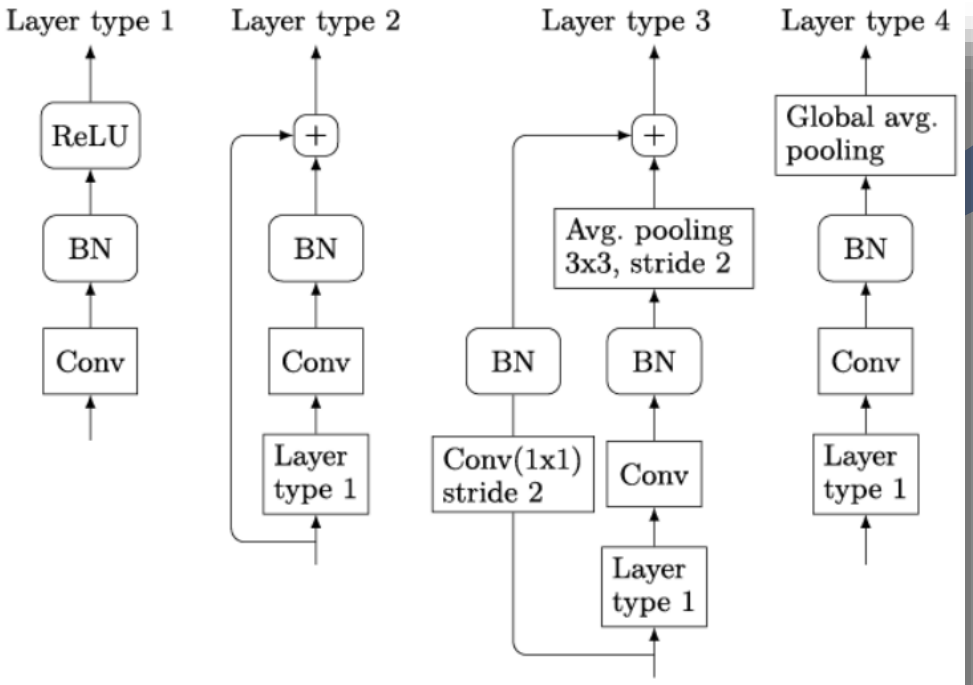
- L1-L7被用于捕获隐写残差特征，L8-L11被用于降低特征维度，L12则被用于汇聚特征以便完成分类任务；

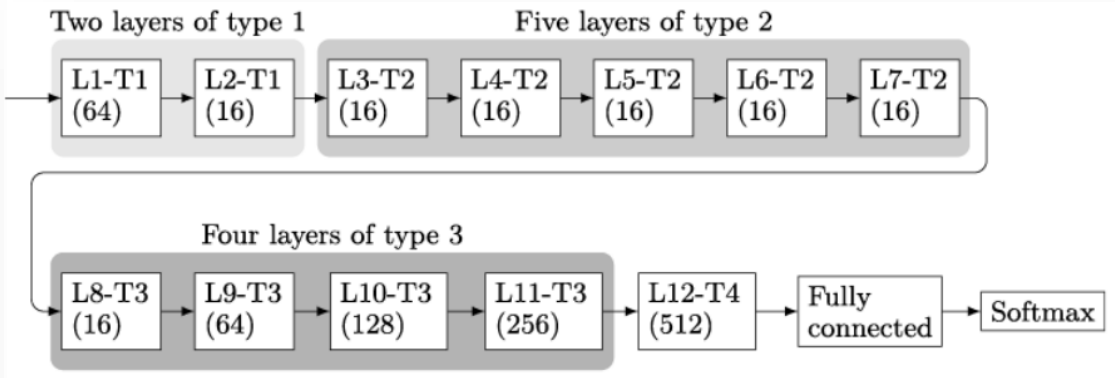
四类子网

- DRNM网络由四种网络模块组成，称为Layer type 1、Layer type 2、Layer type 3和Layer type 4；

池化处理

- Layer type 1和2**不包含平均池化操作**。作者认为，平均池化等效低通滤波，会抑制类似噪声的隐写信号；
- Layer type 3选用平均池化操作，降低特征维度；Layer type4采用全局平均池化操作汇聚所有隐写特征信息。





基于深度学习的隐写分析技术

XuNet

考虑**残差信号的对称性**，在第1个卷积层后增加了一个ABS层(Absolute value layer,对值层) 。XuNet也使用了包括BN、TanH激活函数和ReLU激活函数等技巧解决过拟合、梯度消失等问题。

迁移学习思想

针对低嵌入率载密图像的隐写分析模型存在难训练、难收敛、性能差的问题，Qian等人利用迁移学习思想解决。首先使用高嵌入率载密图像训练，再用低嵌入率载密图像微调。.

Yedroudj-Net

综合XuNet和YeNet的优点。在此基础上，Yedroudj-Net运用**图像增广**技术扩大样本容量，进一步提升了隐写检测的准确率。

DCTR

DCTR是一个混合网络。第一阶段以手工设计的方式通过SRM模型中的滤波器核提取图像残差。DCTR的第二阶段使用了3个卷积子网，3个全连接层和1个softmax。后续多种针对JPEG的深度学习网络被提出，检测准确率也已超越传统的特征手工设计检测方法。

ZhuNet

ZhuNet在前人工作基础上运用了**可分离卷积**（Separable convolutions）和**空间金字塔池化**（Spatial pyramid pooling）技术，控制了网络参数规模，赋予模型了检测**不同尺寸图像**的能力。

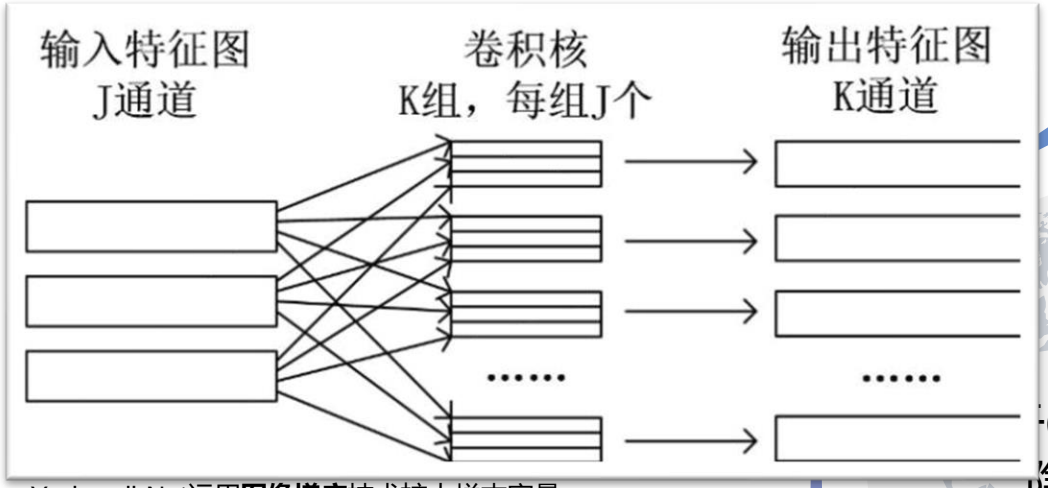
WISERNET

对于**彩色图像**，现有算法都简单地将其视作三张同样尺寸的灰度图像，卷积时，三个通道独立进行二维卷积，输出结果叠加为一体。Zeng认为这与“共谋攻击”有可比性，叠加行为稀释了隐写带来的异常效应，提出“分通道卷积”的技术捕捉通道间相关性。数据显示，WISERNET的性能优于Zhu-Net等网络架构。。.

基于深度学习的隐写分析技术

$$Z_k^l = \sum_{j=1}^{J} Z_j^{l-1} \otimes W_{jk}^l$$

对于第l个卷积层L_l，其通道数为J的输入特征图为Z_j^{l-1}，$1 \le j \le J$；将输入特征图与$J \times K$个卷积核W_{jk}^l作卷积运算，输出通道数为K的特征图Z_k^l，$1 \le k \le K$；⊗表示二维卷积运算。



Yedroudj-Net运用**图像增广**技术扩大样本容量，进一步提升了隐写检测的准确率。

DCTR

DCTR是一个混合网络。第一阶段以手工设计的方式通过SRM模型中的滤波器核提取图像残差。DCTR的第二阶段使用了3个卷积子网，3个全连接层和1个softmax。后续多种针对JPEG的深度学习网络被提出，检测准确率也已超越传统的特征手工设计检测方法。

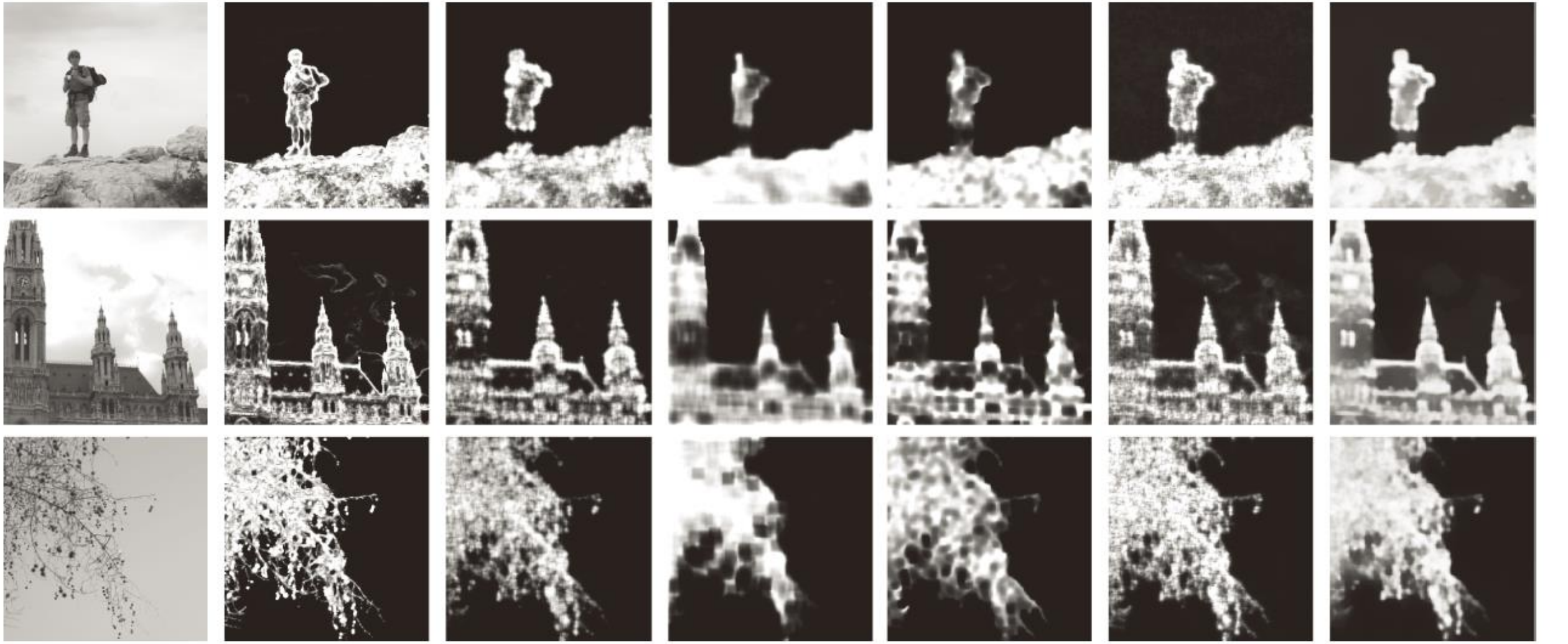
ZhuNet

ZhuNet在前人工作基础上运用了**可分离卷积**（Separable convolutions）和**空域金字塔池化**（Spatial pyramid pooling）技术，控制了网络参数规模，赋予模型了检测**不同尺寸图像**的能力。

WISERNET

对于**彩色图像**，现有算法都简单地将其视作三张同样尺寸的灰度图像，卷积时，三个通道独立进行二维卷积，输出结果叠加为一体。Zeng认为这与“共谋攻击”有可比性，叠加行为稀释了隐写带来的异常效应，提出“分通道卷积”的技术捕捉通道间相关性。数据显示，WISERNET的性能优于Zhu-Net等网络架构。。.

基于深度学习的隐写分析技术挑战



(a) Cover images (b) HUGO-BD (c) WOW (d) HILL (e) MiPOD (f) S-UNIWARD (g) Estimated results

04

基于深度学习的信息隐藏技术发展趋势

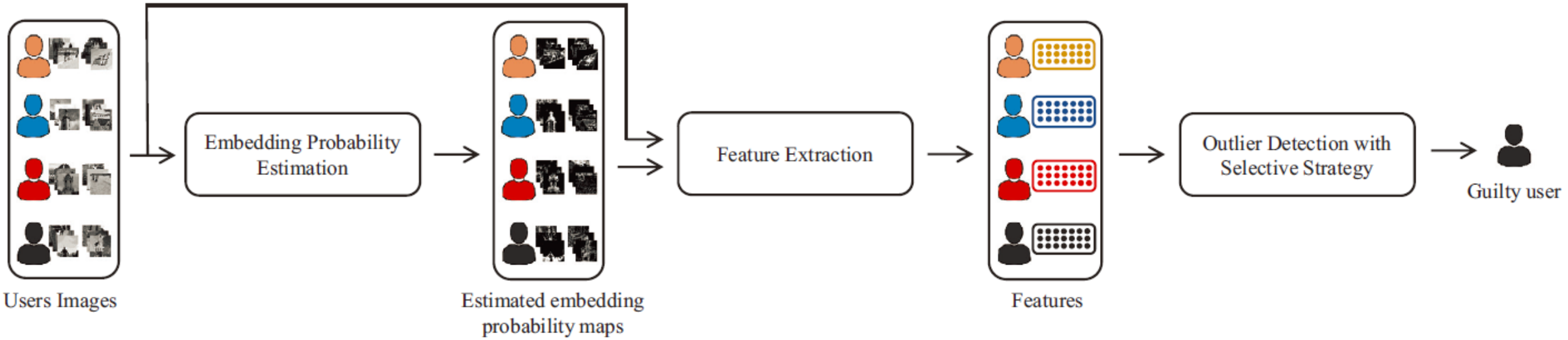
Supporting text here.

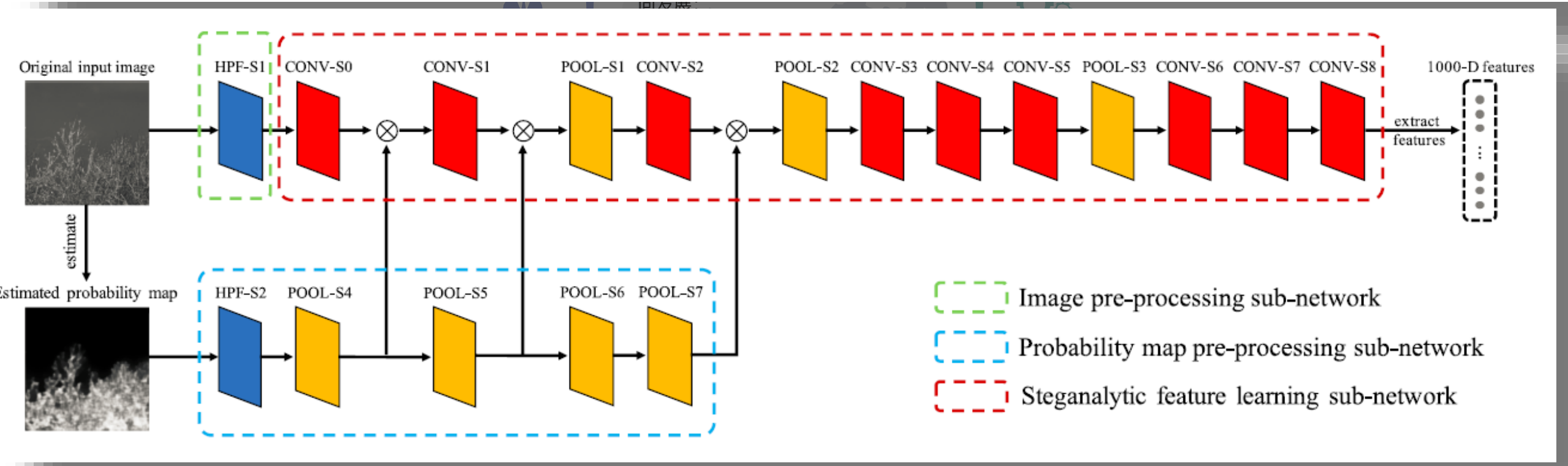
When you copy & paste, choose "keep text only" option.

发展趋势

隐写者检测

海量异源数据以及参数失匹配问题，对隐写分析算法实用化的挑战愈加突出，隐写者检测研究的重要性也逐步提升。.


Embedding Probability Estimation
Feature Extraction
Outlier Detection with Selective Strategy
Guilty user
Users Images
Estimated embedding probability maps
Features

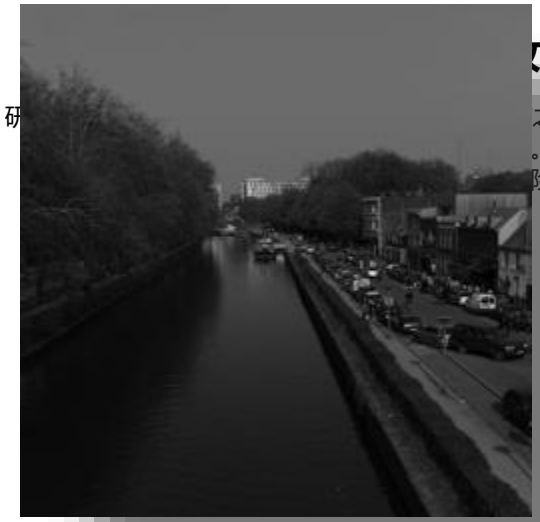

Original input image
HPF-S1
CONV-S0
CONV-S1
POOL-S1
CONV-S2
POOL-S2
CONV-S3
CONV-S4
CONV-S5
POOL-S3
CONV-S6
CONV-S7
CONV-S8
1000-D features
extract features
estimate
Estimated probability map
HPF-S2
POOL-S4
POOL-S5
POOL-S6
POOL-S7
Image pre-processing sub-network
Probability map pre-processing sub-network
Steganalytic feature learning sub-network

发展趋势

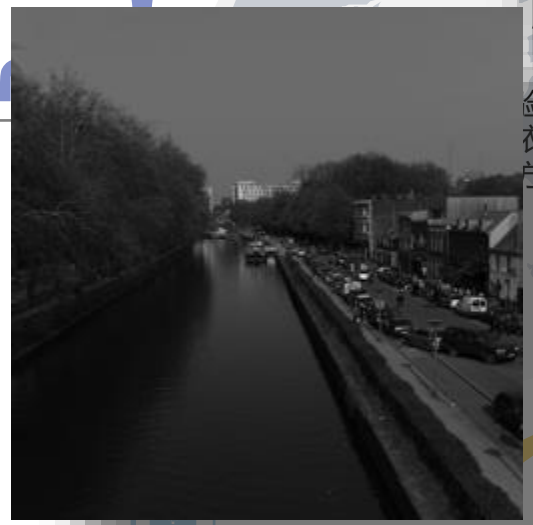
隐写者检测

海量异源数据以及参数失匹配问题，对隐写分析算法实用化的挑战愈加突出，隐写者检测研究的重要性也逐步提升。.

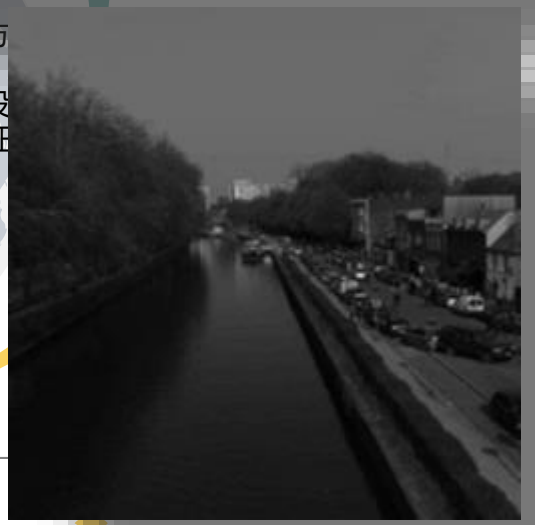
Error Rate(%)	WOW			ADS-WOW		
Steganalyzer	P_{MD}	P_{FA}	P_E	P_{MD}	P_{FA}	P_E
Xu's CNN	25.54	26.60	26.07	86.71	26.60	56.66
Ye's CNN	21.34	18.55	19.95	74.26	18.55	46.41
Wu's CNN	28.04	35.17	31.61	75.34	35.17	55.26
Multi-CNN	7.01	58.43	32.72	56.28	58.43	57.36
SRM	25.23	25.77	25.50	47.81	25.77	36.79
SRM (retrained)	23.90	44.72	34.31	13.33	44.72	29.03



自然图像



携密图像（WOW 0.4bpp）



噪声携密图像（受几何攻击）

发展趋势

隐写者检测

IEEE WIFS（Workshop on Information Forensics and Security，信息取证与安全研讨会）、特鲁瓦工业大学（UTT），CRIStAL Lab（CRIStAL实验室），Lille University（里尔大学）合作举办阿拉斯加挑战赛。
大赛设置了$ 25,000的奖金池，并提供了包括75,000个cover-stego图像对的开源隐写分析图像数据集。

研究表明，
可以产生

融合数字图像取证技术

实际应用中存在非隐写、图像编辑操作干扰隐写
于识别图像是否经
、剪切、黏贴或替
个多任务问题，
写分析和数字图像

的建立和学习

觉任务动辄百万个样本的规模而
，抑制了深度学习模型的性能。
分辨率动态范围较大，小到64*64，
分辨率样本训练得到模型不符合

阿拉斯加隐写分析挑战赛（ALASKA）发布了面向实际应用的新的数据库，势必成为新的基准数据库。.

定量分析

检测秘密信息长度是提取隐写了的秘密信息的前提。可视定量分析为多分类问题。一种算法采用线性均匀划分的思路，将[0,1]区间等间隔划分为10个子集，训练模型正确地将不同隐写率的载密图像分类到对应容量区间。

变换域和彩图隐写及分析技术

现今基于深度学习技术的隐写和隐写分析技术，对变换域的研究远远少于对空域的研究，无论是针对JPEG的深度检测网络模型，还是针对JPEG的GAN隐写机制.

感谢您的聆听.

欢迎提问讨论.

杨榆 副教授
北京邮电大学网络空间安全学院