Advancing the JPEG Compatibility Attack: Theory, Performance, Robustness, and Practice

Eli Dworetzky, Edgar Kaziakhmedov, Jessica Fridrich

IH&MMSEC 2023

BINGHAMTON
UNIVERSITY
STATE UNIVERSITY OF NEW YORK

JPEG Compatibility Attack

- Alice has a JPEG image
- It is decompressed to spatial (pixel) domain by Alice or by Tool
- Tool embeds in spatial domain and saves as spatial domain
- Detection is significantly easier because **JPEG compression forces constraints on pixels**
- Warden first detects that the image was a JPEG and then uses this fact for detection
- JCA can be extremely accurate

Circumstances for JCA are not uncommon

1. Tool can only embed in pixel domain
   - Alice decompresses to be able to use Tool
   - Tool decompresses the JPEG for Alice
2. Tool can embed in JPEGs but Alice needs to embed a larger message that does not fit in the JPEG

Out of 42 stego apps (Github + Google search)

- 19 susceptible to JCA (allow JPEG cover but output spatial domain only)
- 12 confirmed by analyzing source code

Prior art (verify JPEG compatibility)

Fridrich et al. (SPIE 2001)

- estimate JPEG QF (QT) from the stego image
- prove that one or more 8×8 blocks could not have resulted by decompressing a block of quantized DCTs (**incompatible** blocks)
- for QF > 95 computationally expensive
- the zoo of today's JPEG compressors and quantizers undermines incompatibility claims

Prior art (recompression residual)

- Böhme (IHW 2008)
  - compress-decompress the stego image and use as pixel predictor in weighted-stegoimage attack
  - very accurate, limited to LSBR
- Luo et al. (SPL 2011) count how many pixels changed by compression-decompression, regress message length
  - applicable to LSBR as well as LSBM
  - for large QFs, recompression artifacts mistaken for message (bigger problem for content-adaptive stego)
- Kodovský et al. (IHW 2012)
  - Recompression Residual Histogram (RRH) as feature for machine learning
  - improves upon Luo but low accuracy for large QFs

No prior art investigated robustness to compressor / quantizer mismatch

Our contribution

- **Theoretical insight**
  - attack formulated within statistical hypothesis testing
  - most powerful detector derived
  - it explains the limitations and the performance of heuristic detectors
- **Practical detectors**
  - markedly improved accuracy for large QFs
  - robust to compressor / quantizer mismatch
- **Evaluation**
  - implementation for practitioners
  - real stego tools

Quantization errors follow wrapped PDFs

When $X \sim f$, the quantization error

$$e_q(X) = X - q[X/q]$$

follows f **wrapped** onto $-q/2 \leq x < q/2$

$$f_{\mathcal{W}}(x; q) = \sum_{n \in \mathbb{Z}} f(x + qn)$$

Poincaré's Limit Theorem (PLT)

$$e_q(cX) \to \mathcal{U}[-q/2, q/2) \qquad \text{as } c \to \infty$$

for any absolutely continuous X

Generalized PLT

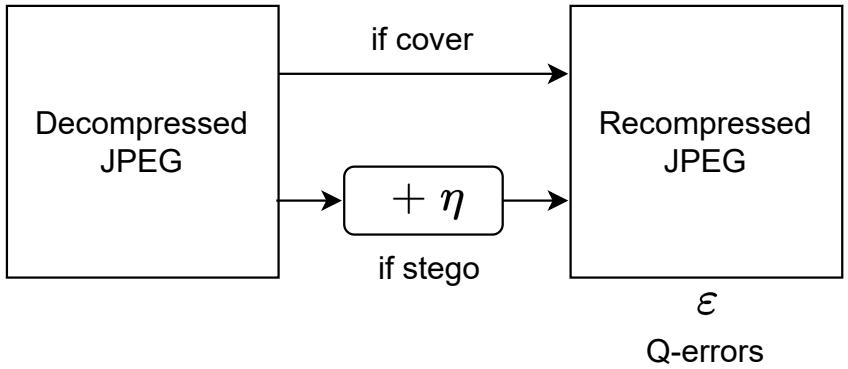
$X = (X_1, \ldots, X_m)$ absolutely continuous, $\mathbf{q} = (q_1, \ldots, q_m)$, $\odot, \oslash$ elementwise operations

$$e_{\mathbf{q}}(X) = X - \mathbf{q} \odot [X \oslash \mathbf{q}]$$

tends to a uniform distribution on an m-dimensional torus

$$e_{\mathbf{q}}(cX) \to \mathcal{U}\left[\prod_{i=1}^{m} [-q_i/2, q_i/2)\right] \qquad \text{as } c \to \infty$$

Our approach: test for Q-errors



$$\mathcal{H}_0 : \varepsilon \sim f_0$$

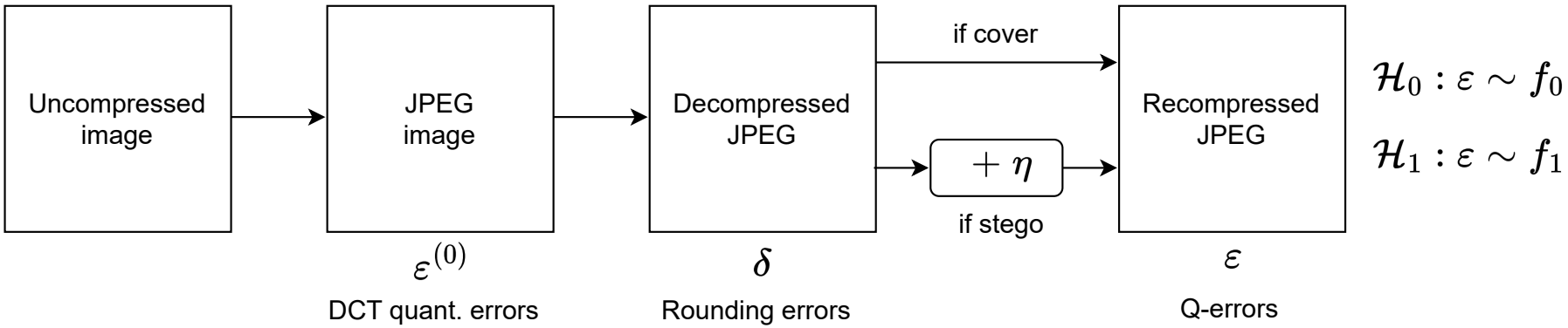
$$\mathcal{H}_1 : \varepsilon \sim f_1$$

Given 8×8 block of pixels $\mathbf{x}$, **Q-errors** are DCT quantization errors

$$\boldsymbol{\varepsilon} = e_{\mathbf{q}}(\mathbf{D}\mathbf{x})$$

$\mathbf{D} \in \mathbb{R}^{64\times 64}$ is the DCT transform, $\mathbf{q} \in \mathbb{R}^{8\times 8}$ quantization table

Pipeline for modeling Q-errors



Modeling the initial compression and decompression

DCT quantization errors $\varepsilon_{kl}^{(0)}$ during the **initial compression** are jointly independent (Sripad, IEEE TASSP 1977)

$$\varepsilon_{kl}^{(0)} \sim \mathcal{U}[-q_{kl}/2, q_{kl}/2)$$

Decompression is a linear transform. By joint independence and CLT, $\mathbf{D}^{-1}\boldsymbol{\varepsilon}^{(0)}$ is Gaussian with variances

$$s_{ij}^{(0)} = \frac{1}{12} \sum_{k,l=0}^{7} (f_{kl}^{ij})^2 q_{kl}^2$$

f_{kl}^{ij} are entries of the DCT matrix $\mathbf{D}$

Modeling the initial compression and decompression

Decompression rounding errors in the spatial domain

$$\boldsymbol{\delta} = e_{\mathbf{1}}(\mathbf{D}^{-1}\boldsymbol{\varepsilon}^{(0)})$$

follow Gaussians $\mathcal{N}(0, s_{ij}^{(0)})$ wrapped on $[-1/2, 1/2)$

$$\delta_{ij} \sim \mathcal{N}_{\mathcal{W}}(0, s_{ij}^{(0)}, 1)$$

For QF ≤ 98, PLT guarantees

$$\delta_{ij} \sim \mathcal{U}[-1/2, 1/2) \text{ jointly independent}$$

We make this assumption for all QFs

Modeling Q-errors

Finally, the **Q-errors** follow a wrapped Gaussian distribution

For covers, $\boldsymbol{\varepsilon} = e_{\mathbf{q}}(\mathbf{D}\boldsymbol{\delta})$

$$\varepsilon_{kl} \sim \mathcal{N}_{\mathcal{W}}(0, s_{kl}, q_{kl})$$

$$s_{kl} = \sum_{i,j=0}^{7} (f_{kl}^{ij})^2 \mathrm{Var}[\delta_{ij}]$$

For stego images $\boldsymbol{\varepsilon} = e_{\mathbf{q}}\left(\mathbf{D}(\boldsymbol{\delta} + \boldsymbol{\eta})\right)$

$$\varepsilon_{kl} \sim \mathcal{N}_{\mathcal{W}}(0, s_{kl} + r_{kl}, q_{kl})$$

$$r_{kl} = \sum_{i,j=0}^{7} (f_{kl}^{ij})^2 \mathrm{Var}[\eta_{ij}]$$

$\eta_{ij} \in \{-1, 0, 1\}$ is the stego signal

Hypothesis test

- Assuming embedding changes $\boldsymbol{\eta}$ are independent of the spatial-domain rounding errors $\boldsymbol{\delta}$
- Warden's hypothesis test for all $0 \leq k, l \leq 7$ (single block)

$$\mathcal{H}_0 : \varepsilon_{kl} \sim \mathcal{N}_{\mathcal{W}}(0, s_{kl}, q_{kl})$$
$$\mathcal{H}_1 : \varepsilon_{kl} \sim \mathcal{N}_{\mathcal{W}}(0, s_{kl} + r_{kl}, q_{kl})$$

  s_{kl} known, $r_{kl} > 0$ known for a known embedding scheme and payload size

- $\mathcal{N}_{\mathcal{W}}(0, s_{kl}, q_{kl})$ is a Gaussian distribution $\mathcal{N}(0, s_{kl})$ wrapped onto $[-q_{kl}/2, q_{kl}/2)$

The most powerful detector

Assuming ε_{kl} are mutually independent within and across blocks, the most powerful detector is the LRT

$$\Lambda(\mathcal{B}) \triangleq \sum_{\mathbf{x}\in\mathcal{B}} \sum_{k,l=0}^{7} \log \frac{g(\varepsilon_{kl}; s_{kl} + r_{kl}, \widehat{q}_{kl})}{g(\varepsilon_{kl}; s_{kl}, \widehat{q}_{kl})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma$$

$g(x; \sigma^2, q)$ is the PDF of $\mathcal{N}_{\mathcal{W}}(0, \sigma^2, q)$

$\mathcal{B}$ is the set of 8×8 pixel blocks that

- do not contain saturated pixels (clipping is hard to model)
- are not sparse in DCT domain (CLT failure)

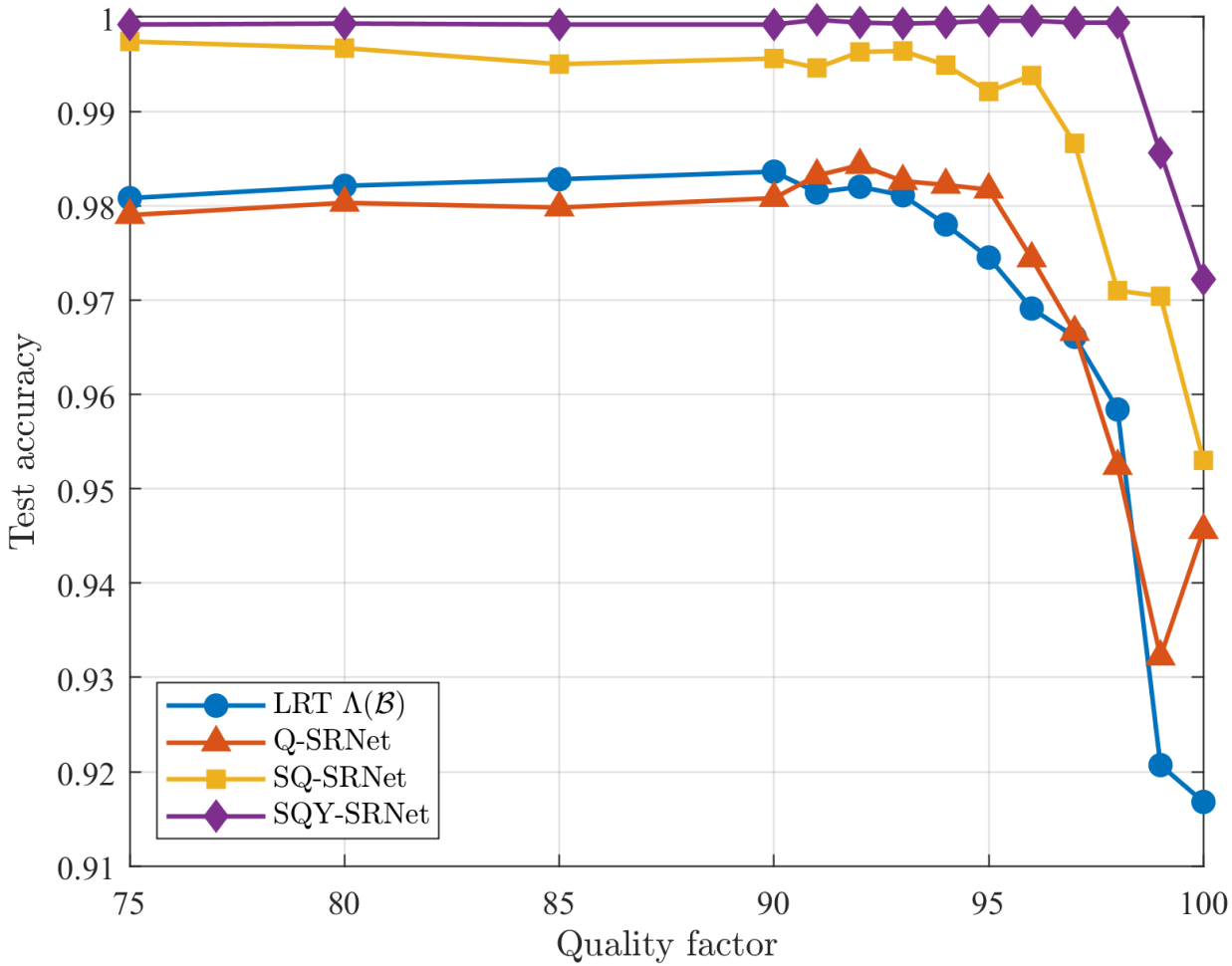
Machine learning detectors

1. Q-SRNet is SRNet trained on Q-errors ε
2. SQ-SRNet is SRNet trained on spatial representation of Q-errors, the so-called **SQ-errors**
$$\mathbf{D}^{-1}\boldsymbol{\varepsilon}$$
3. SQY-SRNet is a two-channel SRNet trained on SQ-errors and the image Y

Dataset: BOSSbase + BOWS2 (20,000 images)

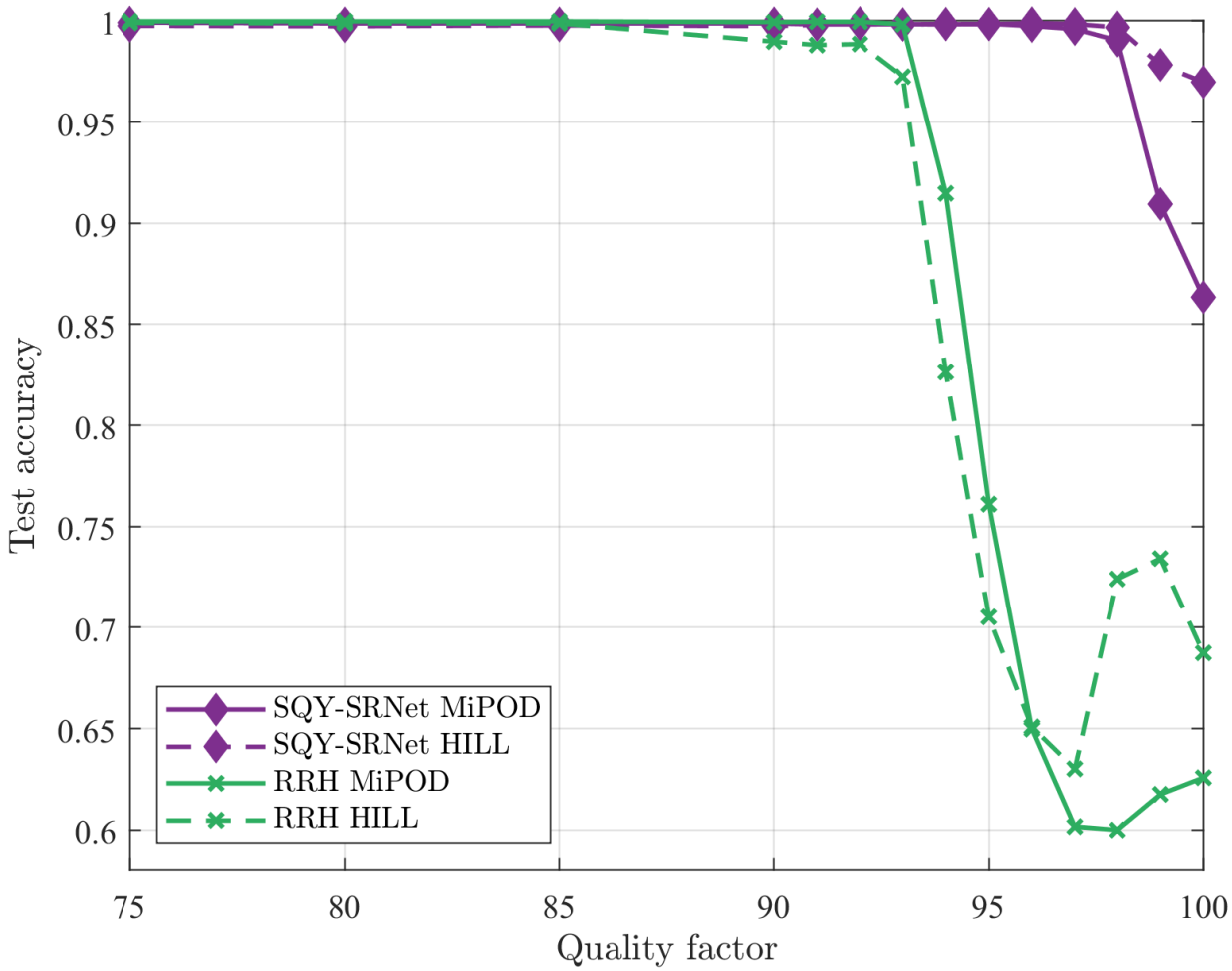
TRN / VAL / TST: 14,000 / 1,000 / 5,000

Testing accuracy as a function of JPEG quality



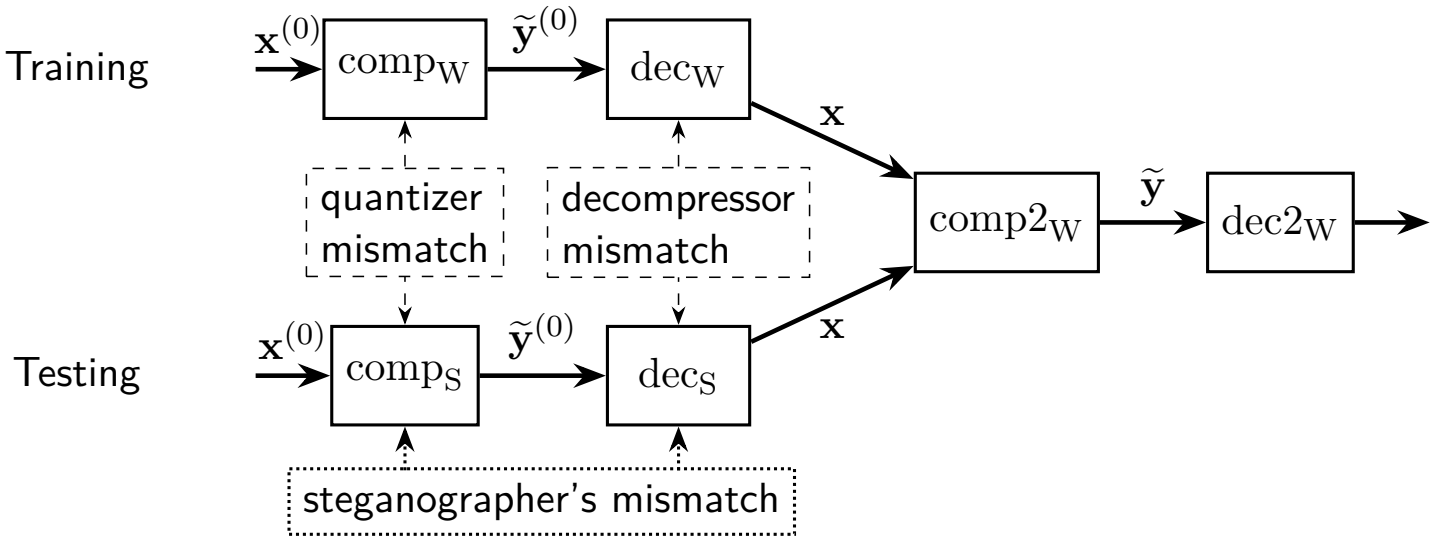
MiPOD at 0.01 bpp

Comparison to state of the art (RRH)

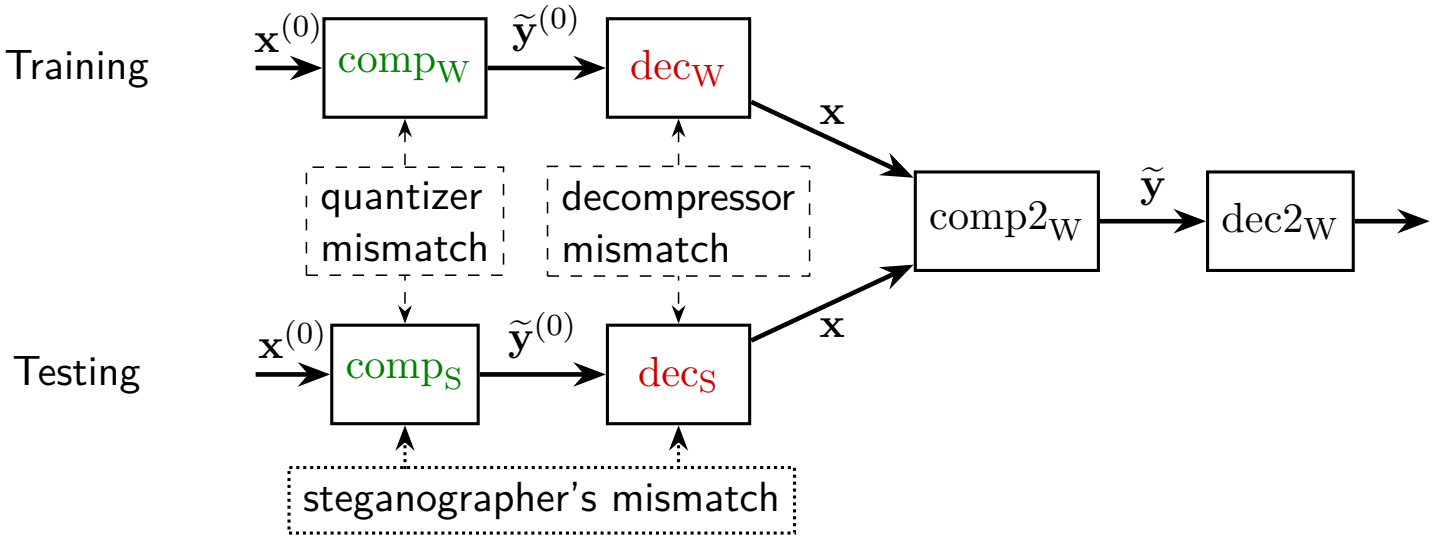


MiPOD / HILL at 0.005 bpp

Sources of mismatch



Combined decompressor + steganographer's mismatch



$\mathrm{comp_S}$ and $\mathrm{comp_W}$: Matlab's `imwrite`

$\mathrm{comp2_W}$ and $\mathrm{dec2_W}$: SciPy's `dct`

Combined decompressor + steganographer's mismatch

Testing accuracy when training on $\mathrm{dec_W}$ and tested on $\mathrm{dec_S}$

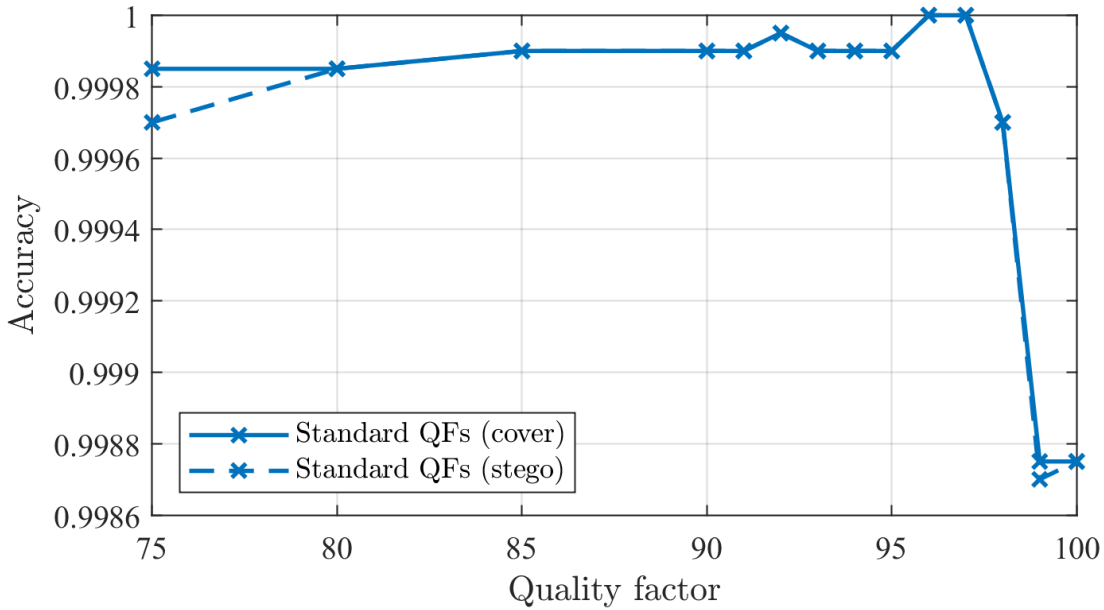
QF	$\mathrm{dec_W}$	SQY-SRNet					RRH				
		$\mathrm{dec_S}$					$\mathrm{dec_S}$				
		imread	float	int	convert	PIL	imread	float	int	convert	PIL
100	imread	.9721	.9556	.9716	.9568	.9729	.7575	.7567	.7590	.7582	.7615
	PIL	.9721	.9633	.9706	.9635	.9708	.7573	.7570	.7585	.7580	.7599
99	imread	.9856	.9846	.9870	.9849	.9859	.7538	.7315	.7523	.7341	.7522
	PIL	.9843	.9864	.9849	.9860	.9844	.7540	.7339	.7518	.7363	.7517
95	imread	≈ 1	≈ 1	≈ 1	≈ 1	≈ 1	.9042	.5000	.9031	.5000	.9041
	PIL	≈ 1	≈ 1	≈ 1	≈ 1	≈ 1	.9022	.5000	.9019	.5000	.9029

MiPOD at 0.01 bpp

Estimating quantization table

- Forensic problem essential for JCA (Thai et al., TIFS 2017)
- Potential issues
  - incorrectly estimating a divisor of the quantization step
  - indeterminable steps
  - uncompressed vs. QF 100
  - effect of stego changes
- Q-errors (and SQ-errors) are negligibly affected when working with non-trivial divisors $\widehat{q}_{kl}$ of q_{kl}
- Thus, for JCA the first two issues are not a problem
- A separate binary classifier trained on uncompressed and QF 100 images resolves the third issue

Effect of stego changes on estimated QF



MiPOD 0.01 bpp, maximum-likelihood estimator of QF that uses a uniform prior for the quantized DCTs

Evaluation on real stego tools

- A total of 42 stego tools from Google search and Github repositories
- Excluded tools that embed in JPEG domain when presented with JPEG
- 19 tools potentially susceptible (12 confirmed from source code)

List of potentially susceptible tools

LSBSteg	SilentEye
cloacked-pixel	QuickCrypto
Matroschka	Steganography
Stegano	rSteg
LSBSteg	SSuite PS
StegoVeritas	StegoStick
Steganography	Steg
stegpy	HuggingFace Stego
OpenStego	StegOnline
Steganography Lib	

`http://dde.binghamton.edu/download/jca/`

Experimental setup

- Some stego tools embed sequentially by rows / columns or start with the last row / column $\implies$ include such images in TRN
- All 19 tools tested with 5 grayscale JPEGs compressed with using Matlab's `imwrite` $\mathrm{QF} \in \{95, 99, 100\}$ (15 images per tool)
- Random message of relative length 0.02 bpp embedded in each cover presented to the tool as JPEG
- SQY-SRNet's threshold set to achieve $P_{\mathrm{FA}} = 10^{-3}$ on TST

Results

- For 15 of the 19 susceptible tools, the quantization table estimator and SQY-SRNets correctly identified the quality factor and perfectly classified the images as stego (all $15 \times 5 \times 3$ images were detected)
- Remaining four tools produced stego images whose quality factors were difficult to determine—most of the images were classified as uncompressed
- One of these tools can be easily detected, since the embedding replaces the three LSB, which introduces visible artifacts

Conclusions

JCA is an old topic but still relevant

1. **Insight** from a theoretical model
2. **Improved accuracy** for high QFs and content-adaptive stego
3. **Robust** to compressor and quantizer mismatch
4. Tested with **real stego tools** (JCA applicable to 15 out of 42)

Future: color, modern JPEG compressors (mozJPEG)