Turning Cost-Based Steganography into Model-Based

Jan Butora, Yassine Yousfi, and Jessica Fridrich
Binghamton University
Department of Electrical and Computer Engineering
Binghamton, NY 13850
{jbutora1,yyousfi1,fridrich}@binghamton.edu

ABSTRACT

Most modern steganographic schemes embed secrets by minimizing the total expected cost of modifications. However, costs are usually computed using heuristics and cannot be directly linked to statistical detectability. Moreover, as previously shown by Ker at al., cost-based schemes fundamentally minimize the wrong quantity that makes them more vulnerable to knowledgeable adversary aware of the embedding change rates. In this paper, we research the possibility to convert cost-based schemes to model-based ones by postulating that there exists payload size for which the change rates derived from costs coincide with change rates derived from some (not necessarily known) model. This allows us to find the steganographic Fisher information for each pixel (DCT coefficient), and embed other payload sizes by minimizing deflection. This rather simple measure indeed brings sometimes quite significant improvements in security especially with respect to steganalysis aware of the selection channel. Steganographic algorithms in both spatial and JPEG domains are studied with feature-based classifiers as well as CNNs.

KEYWORDS

Steganography, steganalysis, costs, model-based, Fisher information

ACM Reference Format:
Jan Butora, Yassine Yousfi, and Jessica Fridrich. 2020. Turning Cost-Based Steganography into Model-Based. In *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '20), June 22–24, 2020, Denver, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3369412.3395065

1 INTRODUCTION

Steganography is another term for covert communication. Instead of communicating the actual message directly, or its encrypted form, it is hidden (embedded) in another cover object. Digital images are especially convenient covers for steganography because their individual elements (pixels or DCT coefficients in a JPEG file) can be slightly modified without changing the semantic meaning of the image. The main requirement here is that the stego objects carrying secrets should be statistically indistinguishable from cover objects [3]. Once the existence of a steganographic channel can be reliably established, the steganographic system is considered broken even if the adversary cannot read the secrets.

All modern steganographic schemes for images are content adaptive in the sense that they prefer modifying cover elements in complex or noisy parts of the image where it is more difficult for the adversary to detect the statistical impact of embedding changes [12, 13, 15, 16, 23–25]. Most stego schemes are "cost based" because each cover element $i \in \{1, \dots, N\}$ is assigned a cost $\rho_i \geq 0$ of changing its value. The required secret payload is then embedded so that each cover element is modified with probability β_i that minimizes the expected sum of costs of all changed pixels $d = \sum_{i=1}^{N} \beta_i \rho_i$, the embedding distortion. This problem is recognized as source coding with fidelity constraint [27] for which near optimal[1] coding has been devised [9]. In particular, when the embedding is allowed to change each cover element by ±1 with equal costs, the embedding change rates that minimize the expected distortion are

$$\beta_i = \frac{\exp(-\lambda\rho_i)}{1 + 2\exp(-\lambda\rho_i)}, \tag{1}$$

where the Lagrange multiplier $\lambda > 0$ is determined from the payload constraint (for the payload-limited sender)

$$\sum_{i=1}^{N} H_3(\beta_i) = m, \tag{2}$$

where m is the total number of bits to be embedded and $H_3(x) = -2\beta_i \log \beta_i - (1 - 2\beta_i)\log(1 - 2\beta_i)$ is the ternary entropy (payload) embedded at cover element i.

There are several issues with cost-based steganography. First of all, the costs themselves are usually computed using heuristic reasoning and cannot be easily related to statistical detectability of embedding changes. Second, this framework does not take into account a knowledgeable adversary aware of the embedding change rates β_i also known as the selection channel. In practice, this leads to embedding that is "overly adaptive," allowing the adversary to improve her detection accuracy using selection-channel-aware (SCA) features, such as [4, 5, 7, 30] or SCA convolutional neural networks (CNNs) [2, 31].

As shown in [19], considering steganography as a zero-sum game between the steganographer and the steganalyst, at equilibrium the sender should select β_i that minimize the statistical detectability, which is asymptotically directly linked to the so-called deflection coefficient

$$\delta^2 \propto \frac{1}{2}\sum_{i=1}^{N} \beta_i^2 I_i, \tag{3}$$

[1] In the sense of the corresponding rate–distortion bound.

IH&MMSec '20, June 22–24, 2020, Denver, CO, USA

ACM ISBN 978-1-4503-7050-9/20/06…$15.00
https://doi.org/10.1145/3369412.3395065

where I_i is the steganographic Fisher information [8, 17] at cover element i. In particular, the optimal change rates satisfy for each i

$$\beta_i I_i = H_3'(\beta_i), \tag{4}$$

where $H_3'(x)$ is the derivative of $H_3(x)$, subject to the same payload constraint. In practice, this is usually done by solving 4 and (2) numerically with a binary search over λ [11, 24, 25].

MiPOD [24] is an example of a steganographic scheme that minimizes the power of the most powerful detector an adversary can build when modeling the noise residuals in a digital image as independent realizations of zero-mean Gaussian random variables with variances σ_i^2 estimated for each cover element i. In this case, the steganographic Fisher information is $I_i \approx 1/\sigma_i^4$ in the fine quantization limit ($\sigma_i^2 > 1$).

Feature-Correction Method (FCM) [20], and approaches based on embedding while minimizing distance in some feature space, such as ASO [22], and Adv-Emb [29], are not truly model-based, because there is no underlying statistical model there, but are again distortion based with the measure of distortion computed as some distance in a selected feature space.

In this paper, we research the possibility to interpret cost-based embedding schemes as model-based schemes similar to MiPOD. We start with the assumption that, for some relative payload $\alpha = m/N$, the embedding change rates β_i computed from the costs as in (1) are the optimal change rates for some (unknown) cover model, derive the corresponding Fisher information, and then for all other payloads, we embed by minimizing the deflection (3). We expect the improvement in security to be especially noticeable for the case of a knowledgeable adversary who knows the embedding change rates β_i, i. e., when steganalyzing with SCA rich models or SCA versions of CNN detectors.

In the next section, we explain the main idea behind converting a cost-based scheme to a model based one. Section 3 contains the results of experiments with HILL and WOW. The improvement in security is shown on two datasets with detectors built as rich models as well as deep CNNs. JPEG-domain schemes J-UNIWARD and UED-JC are studied experimentally in Section 4 for two quality factors 75 and 95. The reported gains are especially large for UED and for the smaller quality factor. Interpreting HILL's costs as reciprocals of local standard deviation estimates, in Section 5 we study a version of MiPOD with this different variance estimator. The paper is summarized in Section 6.

2 COSTS TO MODEL

A brief inspection of the current literature on steganalysis in spatial domain (e. g., [2]) reveals that cost-based steganographic systems that do not use side-information at the sender, such as HILL [23], exhibit approximately the same level of empirical security as the model-based MiPOD [24]. Fundamentally, however, they are very different with HILL minimizing an objective function that is linear in change rates while MiPOD minimizes deflection, which is quadratic in change rates. Since practical embedding with the model-based MiPOD requires converting the optimal change rates determined by (4) to costs by inverting (1) and applying syndrome-trellis codes, one can interpret MiPOD as an embedding scheme with payload-dependent costs (also see Section 5, Fig. 2 in [11]). In this section, we explore this idea in reverse.

The formula for costs is usually derived heuristically through feedback provided by empirical steganalysis. For example, when designing HILL [23], the authors experimented with various sizes of the two low-pass filters. The authors of UNIWARD [14, 16] explored different wavelet bases and their supports as well as a range of values for the stabilizing constant [6]. And this is usually done for a fixed relative payload selected so that the detectability is not too small or too large to better see the impact of various design choices. In the spatial domain, the payload size of 0.4 bpp (bits per pixel) is a popular choice, also because it has been used in the steganalysis competition BOSS [1]. Thus, it is reasonable to assume that this empirical process leads to an embedding scheme that is *near optimal for the chosen payload and the dataset given the current status of steganalysis*. It has already been shown in [26] that steganography tends to be over-optimized for a given source of images. This is confirmed by the above observation that both HILL and MiPOD achieve a similar level of empirical detectability and the fact that no substantial improvement in additive steganography has been reported in the past six years of rather intense research.

Thus, we make an assumption that, given some embedding scheme with costs ρ_i, there exists a relative payload α_D (bpp), which we call the *design payload*, for which the embedding change rates $\beta_i^{(\alpha_D)}$ derived from the costs are near optimal for the current status of steganalysis. Then, we derive the corresponding Fisher information for each pixel, $I_i^{(\alpha_D)}$, so that the deflection $\delta^2 = \frac{1}{2}\sum_{i=1}^N \beta_i^2 I_i^{(\alpha_D)}$ achieves its minimum value when $\beta_i = \beta_i^{(\alpha_D)}$ under the same payload constraint. Using the method of Lagrange multipliers, it can be easily shown that this happens exactly when

$$I_i^{(\alpha_D)} = \frac{\rho_i}{\beta_i^{(\alpha_D)}}. \tag{5}$$

Having determined the Fisher information for each pixel, we can now embed other payload sizes $\alpha \neq \alpha_D$ by minimizing the deflection

$$\delta^2(\alpha) = \frac{1}{2}\sum_{i=1}^N \beta_i^2 I_i^{(\alpha_D)} \tag{6}$$

subject to $\sum_{i=1}^N H_3(\beta_i) = \alpha N$. A graphical representation of above protocol is shown in Figure 1.

Note that this approach does not inform us about the model that is responsible for the steganographic Fisher information. We merely determine I_i, which could correspond to many different models.

3 SPATIAL DOMAIN

In this section, we focus on spatial-domain steganographic algorithms HILL [23] and WOW [15]. Since both have been designed on the standard dataset BOSSbase 1.01 [1] containing 10,000 512×512 grayscale images, we search for the best design payload α_D on the same dataset unless mentioned otherwise. The FLD ensemble [21] with the spatial rich model (SRM) [10] and maxSRM [7] was trained on 5,000 randomly selected images and tested on the remaining 5,000.

HILL (SRM)						
$\alpha_D \backslash \alpha$	0.05	0.1	0.2	0.3	0.4	0.5
0.05	**0.4739**	**0.4416**	0.3636	0.2951	0.2454	0.2017
0.1	0.4712	0.4364	**0.3735**	0.3065	0.2503	0.1994
0.2	0.4643	0.4336	0.3669	**0.3106**	0.2525	0.2097
0.3	0.4587	0.4303	0.3639	0.3056	**0.2537**	0.2067
0.4	0.4544	0.4206	0.3666	0.3067	0.2525	**0.2115**
0.5	0.4548	0.4127	0.3481	0.3005	0.2475	0.2077

HILL (maxSRMd2)						
$\alpha_D \backslash \alpha$	0.05	0.1	0.2	0.3	0.4	0.5
0.05	0.4307	0.3732	0.2916	0.2373	0.1901	0.1507
0.1	0.4452	0.3909	0.3067	0.2446	0.2009	0.1604
0.2	0.4457	0.4024	0.3189	0.2622	0.2126	0.1691
0.3	0.4484	**0.4056**	0.3282	0.2711	0.2249	0.1821
0.4	**0.4502**	0.4025	0.3327	0.2706	0.2291	0.1903
0.5	0.4440	0.4031	**0.3353**	**0.2769**	**0.2301**	**0.1939**

Table 1: Detection error P_E of model-based HILL for different design payloads α_D and embedded payloads α. Left: SRM, Right: maxSRMd2, ensemble classifier, BOSSbase. Regular HILL corresponds to the diagonal ($\alpha_D = \alpha$).

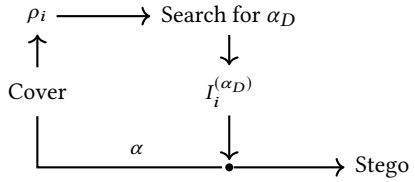

Figure 1: Embedding relative message α bpp (bpnzac) with design payload α_D for arbitrary cost-based steganographic scheme. Notice that the costs ρ_i are used only to compute the Fisher Information for each pixel $I_i^{(\alpha_D)}$.

	α	0.05	0.1	0.2	0.3	0.4
Regular	SRNet	0.3893	0.3192	0.2325	0.1779	0.1465
HILL	SCA-SRNet	0.3992	0.3164	0.2167	0.1717	0.1360
MB-HILL	SRNet	**0.4188**	**0.3468**	0.2449	0.1811	0.1444
$\alpha_D = 0.5$	SCA-SRNet	0.4751	0.3591	**0.2387**	**0.1777**	**0.1393**

Table 2: Detection error P_E of SRNet and SCA-SRNET for HILL and model-based HILL ($\alpha_D = 0.5$ bpp) in downsampled BOSSbase + BOWS2.

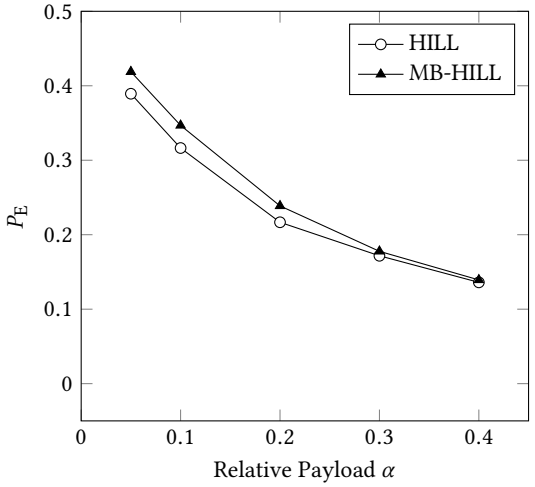

Figure 2: Detection error P_E of the best detector (SRNet or SCA-SRNet) for HILL and model-based HILL ($\alpha_D = 0.5$ bpp) in downsampled BOSSbase + BOWS2.

3.1 Model-based HILL

Table 1 shows the results for HILL in terms of P_E, the total classification error under equal priors for the cover and stego classes. The boldface font highlights the most secure algorithm version, which is to be compared with the diagonal ($\alpha = \alpha_D$) corresponding to regular HILL. Note that the results are vastly different depending on the steganalysis features. For SRM, which is an ignorant adversary (one who does not use the knowledge of the selection channel), there is no clear design payload that would always give the best results. Also, the impact on security is quite small. In contrast, detection with a knowledgeable adversary (maxSRMd2) indicates that the best overall design payload is $\alpha_D = 0.5$ bpp (for the two smallest tested payloads the differences between $\alpha_D = 0.3, 0.4$, and 0.5 are small). The largest boost in empirical security is 1.7% for payload $\alpha = 0.2$.

We repeated the same experiment with the CNN SRNet and its SCA version [2]. Because large CNNs, such as the SRNet, cannot be trained on 512×512 images on GPUs with 12 GB memory with a reasonable batch size, we used the union of BOSSbase and BOWS2 whose images were downsampled to 256×256 pixels using Matlab's imresize with default parameters. As in [2, 31], this 20,000 image dataset was split into 14,000 (10,000 BOWS2 and 4,000 randomly chosen from BOSSbase) for training, 1,000 BOSSbase images for validation, and 5,000 for testing.

Technically, the design payload should be searched for anew for this dataset and detector. Due to the much more computationally demanding training of the SRNet, however, we only compare model-based HILL for $\alpha_D = 0.5$ and regular HILL (Table 2). Comparing the best detector (SRNet vs. SCA-SRNet[2]) for each embedding algorithm in Figure 2, we observe an empirical gain in security ranging from almost 3% for the smallest payloads to almost no gain for $\alpha = 0.4$.

3.2 Model-based WOW

Searching for the best design payload on BOSSbase with maxSRMd2 and the ensemble classifier, it also appears to be close to $\alpha_D = 0.5$ bpp. Since WOW is known to be overly content-adaptive in the sense that its security decreases significantly with selection-channel-aware attacks, the impact of making it model-based is

[2] In some cases, SCA-SRNet performs worse than SRNet.

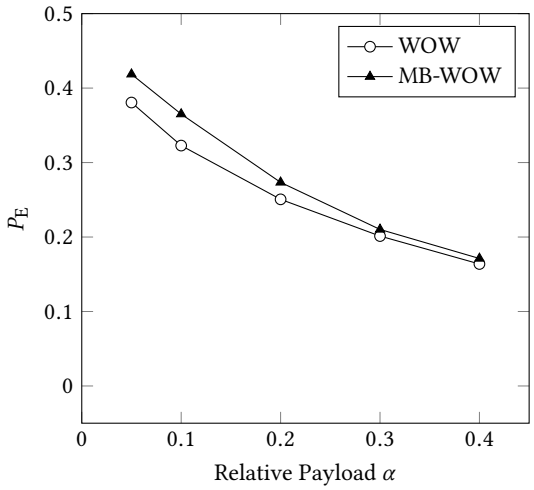

Figure 3: Detection error P_E of maxSRMd2 for WOW and model-based WOW ($\alpha_D = 0.5$ bpp) in BOSSbase.

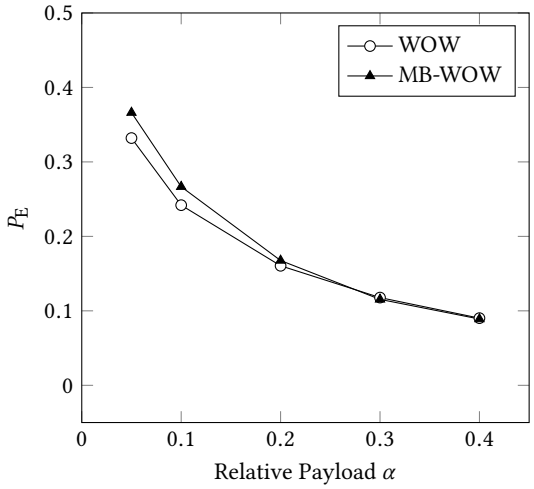

Figure 4: Detection error P_E of the best detector (SRNet or SCA-SRNet) for WOW and model-based WOW ($\alpha_D = 0.7$ bpp) in downsampled images BOSSbase + BOWS2.

larger than for HILL. The detection error P_E shown in Figure 3 is about 4% larger for the two smallest payloads for model-based WOW than for the original cost-based algorithm.

On the dataset of downsampled images, based on our investigation with maxSRMd2, the best design payload is larger, $\alpha_D = 0.7$ bpp. In Figure 4, we contrast the detection error of SRNet on model-based WOW and WOW ranges from 3.4% for the smallest payload of 0.05 bpp to 0.7% for 0.2 bpp. The empirical security of both algorithms appears similar for the two largest payloads. The actual values of the detection error appear in Table 4 at the end of this paper.

4 JPEG DOMAIN

In the JPEG domain, we investigated the embedding algorithms J-UNIWARD and UED-JC [13]. For the database of larger 512×512 images, we steganalyzed with selection-channel-aware Gabor Phase Aware Residuals, SCA-GFR [4, 28], while, as above, the SRNet and SCA-SRNet were used on the database of downsampled images. The split of the datasets was the same as for the experiments in the spatial domain.

4.1 J-UNIWARD

For J-UNIWARD, the results are graphically displayed in Figure 5 showing the detection error of J-UNIWARD and its model-based version with $\alpha_D = 0.6$ bpnzac. The gain in security is generally much larger than what was observed in the spatial domain. Also, it is larger for quality factor 75 than for 95. As before, the gain increases with decreasing payload. In particular, for quality 75 the gain was up to 3.5% with SCA-GFR and 7.3% with SCA-SRNet. While we observed almost no gain for quality 95 with SCA-GFR, the better detector (SCA-SRNet) showed more than 8% of improvement for the smallest payload.

4.2 UED

The embedding algorithm UED-JC benefits from our approach by far the most out of all tested stego methods in any domain. Figure 6 shows the detection error achieved on BOSSbase with SCA-GFR and on the downsampled images with (SCA)-SRNet for two quality factors. The gain is again larger on downsampled images when detecting with (SCA)-SRNet and is over 12% for the smallest payload. On BOSSbase with SCA-GFR, the gain on the smallest payload is about 10%. In both datasets, the gain diminishes to zero as α approaches α_D.

The actual values of the detection error from the graphs for J-UNIWARD and UED-JC appear in Table 4 at the end of this paper.

5 INTERPRETING HILL'S COSTS

The main contribution of this paper is the realization that there is a cover model behind cost-based schemes and a method for estimating the model, its Fisher information. In this section, we take a closer look at the embedding algorithm HILL, and interpret its costs as reciprocal estimates of the local standard deviation. Equipped with this insight, we implement a model-based version of HILL with a Gaussian model of pixel residual, which is essentially a version of MiPOD with a different variance estimator.

HILL (**H**igh-pass, **L**ow-pass, **L**ow-pass) computes costs heuristically using a series of filtering operations. First, the 3×3 high-pass KB filter [18] F_{KB} is applied to the cover image $\mathbf{X}$, producing the KB residual $\mathbf{R} = \mathbf{X} \star F_{\text{KB}}$. Next, the absolute value of the KB residual is smoothed with a 3×3 averaging filter $A_{3\times3}$: $|\mathbf{R}| \star A_{3\times3}$. Finally, the reciprocal of this signal is smoothed by applying a 15×15 averaging filter $A_{15\times15}$:

$$\boldsymbol{\rho} = A_{15\times15} \star \frac{1}{|\mathbf{R}| \star A_{3\times3}}. \tag{7}$$

Ignoring the second low-pass filtering in Equation 7 for simplicity, the costs can be seen as reciprocal expectation of the absolute

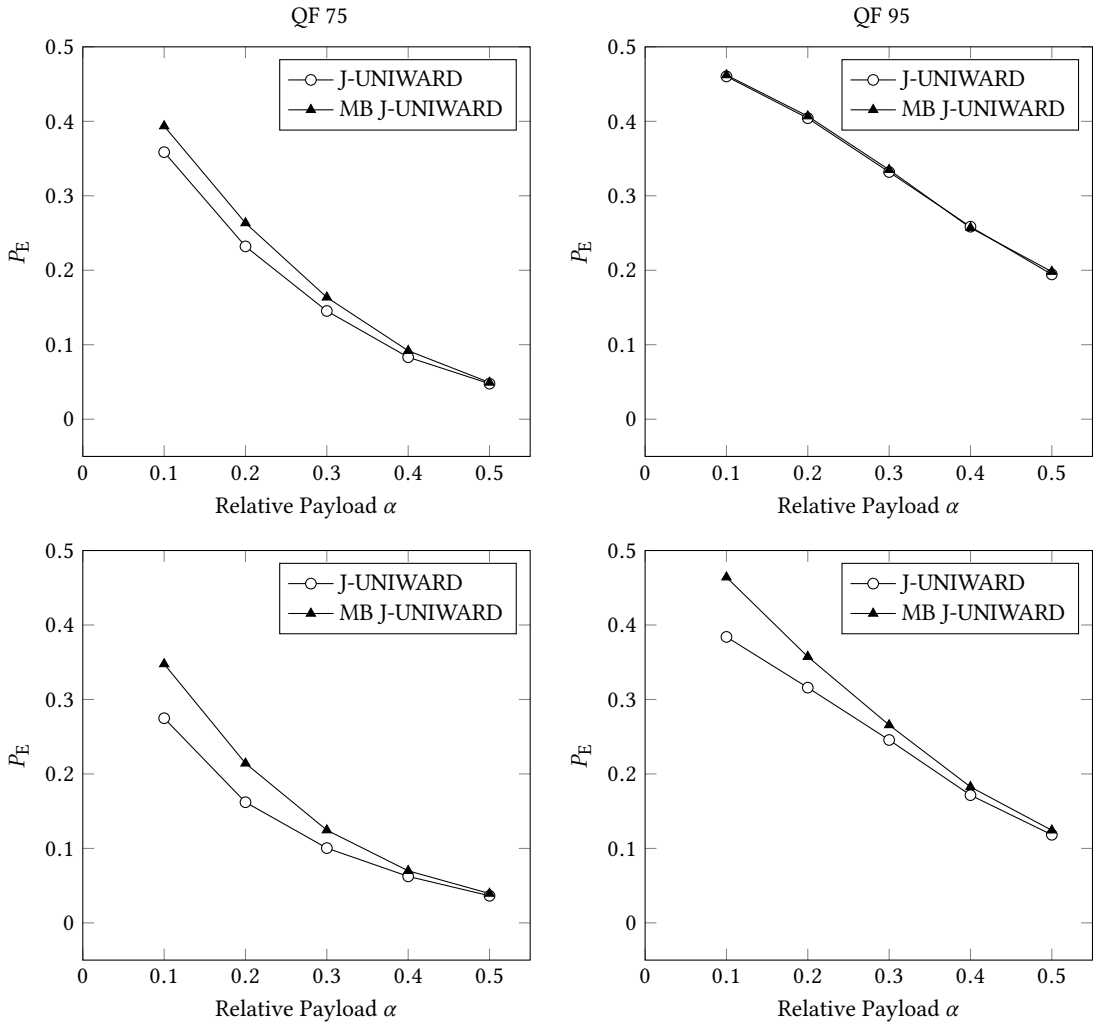

Figure 5: Detection error P_E for J-UNIWARD and model-based J-UNIWARD ($\alpha_D = 0.6$ bpnzac) when steganalyzing with SCA-GFR on BOSSbase (top) and with SCA-SRNet (or SRNet, whichever is better) on downsampled BOSSbase + BOWS2 (bottom) for quality 75 and 95.

value of the KB residual $\rho_i \simeq 1/E[|R_i|]$ or a reciprocal of the Mean Absolute Deviation (MAD), assuming the KB residual is zero mean. Similar to the standard deviation (std), MAD is a description of a statistical spread of a random variable X. For a wide range of distributions typically used in image modeling (e. g., for the generalized Gaussian distribution and the generalized Gamma distribution), the expectation of absolute value is proportional to the standard deviation when fixing the remaining parameters, $E[|X|] \propto \sigma$. Thus, the reciprocal cost

$$\frac{1}{\rho_i} \simeq E[|R_i|] \propto \sigma_i. \tag{8}$$

This tells us that that HILL's costs can loosely be viewed as reciprocals of estimates of local standard deviation. Assuming the KB residual is locally Gaussian $R_i \sim \mathcal{N}(0, \sigma_i^2)$, the costs inform us about the standard deviations σ_i:

$$1/\rho_i \simeq E[|R_i|] = \sigma_i \sqrt{\frac{2}{\pi}}. \tag{9}$$

Note that, with a locally Gaussian residual model, we arrived at a different version of MiPOD with the following "HILL-inspired" plug-in variance estimator

$$\sigma_i^2 = \frac{\pi}{2\rho_i^2}. \tag{10}$$

Before subjecting this embedding scheme to practical tests, we first validate the model in the following fashion. Given image $\mathbf{X}$ with KB residual $\mathbf{R}$, we first estimate its local variance from HILL's costs (10) and then using MiPOD's variance estimator, respectively.[3] Then,

[3] For MiPOD, this was achieved by passing the KB residual instead of the noise residual computed using the 2×2 Wiener filter to the parametric denoising algorithm (see Sec. V in [24]).

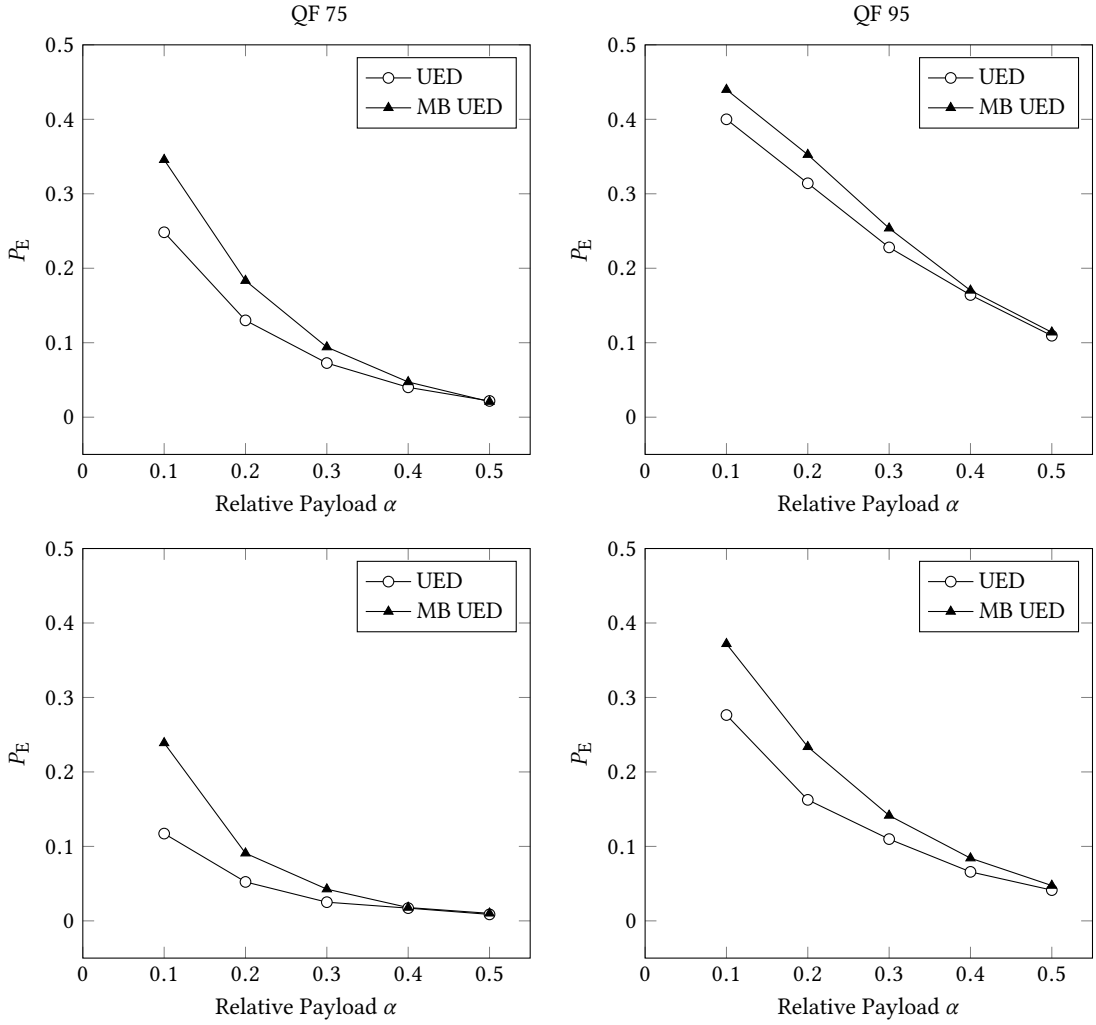

Figure 6: Detection error P_E for UED-JC and model-based UED-JC ($\alpha_D = 0.6$ bpnzac) when steganalyzing with SCA-GFR on BOSSbase (top) and with SCA-SRNet (or SRNet, whichever is better) on downsampled BOSSbase + BOWS2 (bottom) for quality 75 and 95.

we sample M times the multivariate Gaussian $(\widetilde{R}_1, \dots, \widetilde{R}_N)$, $\widetilde{R}_i \sim \mathcal{N}(0, \sigma_i^2)$, where N is the number of pixels in the image. Given these $M \times N$ random samples $\widetilde{\mathbf{R}}$, we compute their empirical probability mass function (histogram with 100 uniform bins) $h_{\widetilde{\mathbf{R}}}$ and compare it with the histogram $h_{\mathbf{R}}$ of the KB residual $\mathbf{R}$ using the discrete Kullback–Leibler divergence $D_{\mathrm{KL}}(h_{\mathbf{R}} || h_{\widetilde{\mathbf{R}}})$. Executing this for 5,000 512×512 grayscale images $\mathbf{X}$ from the training subset of BOSSbase 1.01, in Figure 7 we show the box plot of the KL divergence across all 5,000 images obtained using both variance estimators. Note that if the adopted and estimated model perfectly fit the KB residual, we would see a KL divergence near zero. The figure shows that using HILL's costs to estimate the KB variance is slightly better in terms of preserving the overall residual distribution.

Based on this observation, we implemented MiPOD with HILL's variance estimator (10). In order to focus on the effect of the variance estimator, we skip the Fisher Information smoothing step in MiPOD. Table 3 shows that the HILL-inspired estimator (10) provides better security than the original variance estimator in MiPOD, in agreement with the model validation shown in Figure 7.

6 CONCLUSIONS

Most steganographic schemes today are content adaptive, designed around the paradigm of minimizing the total embedding cost. Costs are, however, typically designed using intuitive heuristic rules, making it difficult, if possible at all, to link the impact of embedding to statistical detectability. Moreover, at least asymptotically for small payloads, the statistical detectability is quadratic in embedding change rates while the embedding distortion is linear. On the

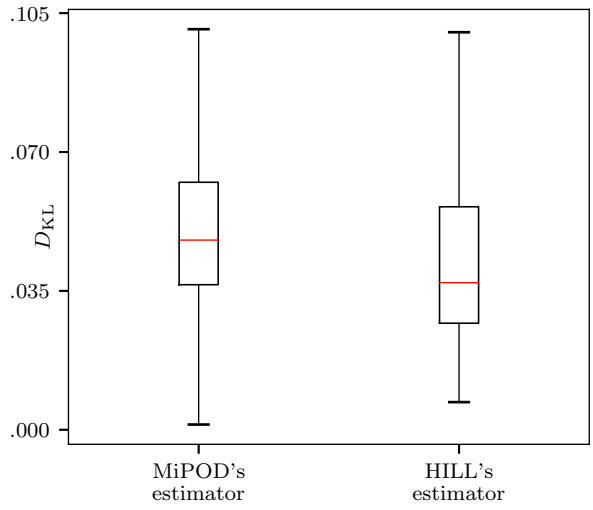

Figure 7: $D_{\mathbf{KL}}(R||\widetilde{R})$ with the KB residual variance estimated using HILL's costs and MiPOD's variance estimator. The red line shows the median, the bottom and top edges of the box indicate the 25th and 75th percentiles, and the whiskers length set to 1.5. Samples computed from 5,000 512×512 grayscale images from BOSSbase.

Variance		0.1	0.2	0.3	0.4
Eq. (10)	maxSRMd2	**0.3937**	**0.3206**	**0.2678**	**0.2213**
MiPOD	maxSRMd2	0.3800	0.3101	0.2552	0.2142
Eq. (10)	SRNet	**0.3390**	0.2470	0.1870	0.1545
	SCA-SRNet	0.3575	**0.2354**	**0.1826**	**0.1420**
MiPOD	SRNet	0.3213	0.2222	0.1553	0.1146
	SCA-SRNet	0.2952	0.1961	0.1384	0.1106

Table 3: Detection error P_{E} for MiPOD with variance estimator (10) and the original MiPOD estimator in BOSSbase (maxSRMd2 + ensemble) and in downsampled BOSSbase + BOWS2 (with (SCA)-SRNet).

other hand, given the success of cost-based steganography to avoid steganalysis, the costs must have some relationship to detectability.

Costs are typically designed from feedback provided by steganalysis on a selected dataset and usually for a fixed payload. In this paper, we postulate that there exists a relative payload for which the embedding change rates correspond to minimal statistical detectability for some unknown model of pixels (DCTs). For this so-called design payload, we convert the costs to the steganographic Fisher information. Although the underlying model is not known, with the Fisher information, we can embed other payloads with a model-based scheme by minimizing the deflection. As shown in this paper, this rather simple idea indeed leads to improved security, especially with respect to selection-channel-aware steganalysis. The gain typically increases with decreased payload. In JPEG domain, we observed larger gains for smaller quality factors than for large qualities. The gains for JPEG-domain algorithms are also generally larger than for spatial domain. The largest observed gains exceed 12% in terms of the total detection error under equal priors P_{E} for UED-JC at quality 75.

Inspired by the success of this simple idea, we also explore a model-based scheme, a version of MiPOD, with a different pixel variance estimator obtained by interpreting HILL's costs as reciprocal estimates of standard deviation from the KB residual. This algorithm indeed performs better than when estimating the variance of the KB residual with MiPOD.

ACKNOWLEDGMENTS

The work on this paper was supported by NSF grant No. 1561446 and by DARPA under agreement number FA8750-16-2-0173. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of DARPA or the U.S. Government.

REFERENCES

[1] P. Bas, T. Filler, and T. Pevný. Break our steganographic system – the ins and outs of organizing BOSS. In T. Filler, T. Pevný, A. Ker, and S. Craver, editors, *Information Hiding, 13th International Conference*, volume 6958 of Lecture Notes in Computer Science, pages 59–70, Prague, Czech Republic, May 18–20, 2011.

[2] M. Boroumand, M. Chen, and J. Fridrich. Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 14(5):1181–1193, May 2019.

[3] C. Cachin. An information-theoretic model for steganography. *Information and Computation*, 192(1):41–56, July 2004.

[4] T. Denemark, M. Boroumand, and J. Fridrich. Steganalysis features for content-adaptive JPEG steganography. *IEEE Transactions on Information Forensics and Security*, 11(8):1736–1746, August 2016.

[5] T. Denemark and J. Fridrich. Improving selection-channel-aware steganalysis features. In A. Alattar and N. D. Memon, editors, *Proceedings IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2016*, San Francisco, CA, February 14–18, 2016.

[6] T. Denemark, J. Fridrich, and V. Holub. Further study on the security of S-UNIWARD. In A. Alattar, N. D. Memon, and C. Heitzenrater, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2014*, volume 9028, pages 05 1–13, San Francisco, CA, February 3–5, 2014.

[7] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich. Selection-channel-aware rich model for steganalysis of digital images. In *IEEE International Workshop on Information Forensics and Security*, Atlanta, GA, December 3–5, 2014.

[8] T. Filler and J. Fridrich. Fisher information determines capacity of ϵ-secure steganography. In S. Katzenbeisser and A.-R. Sadeghi, editors, *Information Hiding, 11th International Conference*, volume 5806 of Lecture Notes in Computer Science, pages 31–47, Darmstadt, Germany, June 7–10, 2009. Springer-Verlag, New York.

[9] T. Filler, J. Judas, and J. Fridrich. Minimizing additive distortion in steganography using syndrome-trellis codes. *IEEE Transactions on Information Forensics and Security*, 6(3):920–935, September 2011.

[10] J. Fridrich and J. Kodovský. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, June 2011.

[11] J. Fridrich and J. Kodovský. Multivariate Gaussian model for designing additive distortion for steganography. In *Proc. IEEE ICASSP*, Vancouver, BC, May 26–31, 2013.

[12] L. Guo, J. Ni, and Y.-Q. Shi. An efficient JPEG steganographic scheme using uniform embedding. In *Fourth IEEE International Workshop on Information Forensics and Security*, Tenerife, Spain, December 2–5, 2012.

[13] L. Guo, J. Ni, and Y. Q. Shi. Uniform embedding for efficient JPEG steganography. *IEEE Transactions on Information Forensics and Security*, 9(5):814–825, May 2014.

[14] V. Holub. *Content Adaptive Steganography – Design and Detection.* PhD thesis, Binghamton University, May 2014.

[15] V. Holub and J. Fridrich. Designing steganographic distortion using directional filters. In *Fourth IEEE International Workshop on Information Forensics and Security*, Tenerife, Spain, December 2–5, 2012.

[16] V. Holub, J. Fridrich, and T. Denemark. Universal distortion design for steganography in an arbitrary domain. *EURASIP Journal on Information Security, Special*

Issue on Revised Selected Papers of the 1st ACM IH and MMS Workshop, 2014:1, 2014.

[17] A. D. Ker. Estimating steganographic fisher information in real images. In S. Katzenbeisser and A.-R. Sadeghi, editors, *Information Hiding, 11th International Conference*, volume 5806 of Lecture Notes in Computer Science, pages 73–88, Darmstadt, Germany, June 7–10, 2009. Springer-Verlag, New York.

[18] A. D. Ker and R. Böhme. Revisiting weighted stego-image steganalysis. In E. J. Delp, P. W. Wong, J. Dittmann, and N. D. Memon, editors, *Proceedings SPIE, Electronic Imaging, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, volume 6819, pages 5 1–17, San Jose, CA, January 27–31, 2008.

[19] A. D. Ker, T. Pevný, and P. Bas. Rethinking optimal embedding. In F. Perez-Gonzales, F. Cayre, and P. Bas, editors, *The 4th ACM Workshop on Information Hiding and Multimedia Security*, pages 93–102, Vigo, Spain, June 20–22, 2016. ACM Press.

[20] J. Kodovský and J. Fridrich. On completeness of feature spaces in blind steganalysis. In A. D. Ker, J. Dittmann, and J. Fridrich, editors, *Proceedings of the 10th ACM Multimedia & Security Workshop*, pages 123–132, Oxford, UK, September 22–23, 2008.

[21] J. Kodovský, J. Fridrich, and V. Holub. Ensemble classifiers for steganalysis of digital media. *IEEE Transactions on Information Forensics and Security*, 7(2):432–444, April 2012.

[22] S. Kouider, M. Chaumont, and W. Puech. Adaptive steganography by oracle (aso). In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, San Jose, CA, July 15–19, 2013.

[23] B. Li, M. Wang, and J. Huang. A new cost function for spatial image steganography. In *Proceedings IEEE, International Conference on Image Processing, ICIP*, Paris, France, October 27–30, 2014.

[24] V. Sedighi, R. Cogranne, and J. Fridrich. Content-adaptive steganography by minimizing statistical detectability. *IEEE Transactions on Information Forensics and Security*, 11(2):221–234, 2016.

[25] V. Sedighi, J. Fridrich, and R. Cogranne. Content-adaptive pentary steganography using the multivariate generalized Gaussian cover model. In A. Alattar and N. D. Memon, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2015*, volume 9409, San Francisco, CA, February 8–12, 2015.

[26] V. Sedighi, J. Fridrich, and R. Cogranne. Toss that BOSSbase, alice In A. Alattar and N. D. Memon, editors, *Proceedings IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2016*, San Francisco, CA, February 14–18, 2016.

[27] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.*, 4:142–163, 1959.

[28] X. Song, F. Liu, C. Yang, X. Luo, and Y. Zhang. Steganalysis of adaptive JPEG steganography using 2D Gabor filters. In P. Comesana, J. Fridrich, and A. Alattar, editors, *3rd ACM IH&MMSec. Workshop*, Portland, Oregon, June 17–19, 2015.

[29] W. Tang, B. Li, S. Tan, M. Barni, and J. Huang. Cnn-based adversarial embedding for image steganography. *IEEE Transactions on Information Forensics and Security*, 14(8):2074–2087, 2019.

[30] W. Tang, H. Li, W. Luo, and J. Huang. Adaptive steganalysis based on embedding probabilities of pixels. *IEEE Transactions on Information Forensics and Security*, 11(4):734–745, April 2016.

[31] J. Ye, J. Ni, and Y. Yi. Deep learning hierarchical representations for image steganalysis. *IEEE Transactions on Information Forensics and Security*, 12(11):2545–2557, November 2017.

QF	Steganography	Detector	Payload (bpp / bpnzac) 0.05	0.1	0.2	0.3	0.4	0.5
-	Regular WOW	SRM	0.4606	0.4113	0.3218	0.2563	0.2142	-
		maxSRMd2	0.3806	0.3228	0.2506	0.2013	0.1638	-
	MB WOW $\alpha_D = 0.5$	SRM	0.4515	0.3984	0.3284	0.2562	0.2078	-
		maxSRMd2	**0.4186**	**0.3651**	**0.2734**	**0.2102**	**0.1712**	-
	Regular WOW	SRNet	0.3415	0.2587	0.1701	0.1287	0.1010	-
		SCA-SRNet	0.3320	0.2419	0.1605	**0.1178**	**0.0902**	-
	MB WOW $\alpha_D = 0.7$	SRNet	**0.3662**	0.2678	0.1696	0.1208	0.0913	-
		SCA-SRNet	0.3766	**0.2667**	**0.1676**	0.1154	0.0890	-
75	J-UNIWARD	SRNet	-	0.3161	0.1931	0.1121	0.0707	0.0375
		SCA-SRNet	-	0.2748	0.1620	0.1004	0.0624	0.0364
	MB J-UNIWARD $\alpha_D = 0.6$	SRNet	-	0.3612	0.2196	0.1300	0.0814	0.0465
		SCA-SRNet	-	**0.3476**	**0.2142**	**0.1245**	**0.0699**	**0.0394**
	J-UNIWARD	SCA-GFR	-	0.3586	0.2320	0.1453	0.0832	0.0477
	MB J-UNIWARD $\alpha_D = 0.6$	SCA-GFR	-	**0.3936**	**0.2634**	**0.1636**	**0.0919**	**0.0493**
95	J-UNIWARD	SRNet	-	0.4418	0.3436	0.2594	0.1847	0.1306
		SCA-SRNet	-	0.3840	0.3159	0.2456	0.1715	0.1183
	MB J-UNIWARD $\alpha_D = 0.6$	SRNet	-	0.4772	0.3683	0.2694	0.1859	**0.1243**
		SCA-SRNet	-	**0.4641**	**0.3574**	**0.2657**	**0.1826**	0.1264
	J-UNIWARD	SCA-GFR	-	0.4603	0.4042	0.3319	**0.2585**	0.1944
	MB J-UNIWARD $\alpha_D = 0.6$	SCA-GFR	-	**0.4621**	**0.4069**	**0.3349**	0.2570	**0.1981**
75	UED	SRNet	-	0.1344	0.0571	0.0311	0.0196	0.0111
		SCA-SRNet	-	0.1172	0.0523	0.0251	0.0171	0.0087
	MB UED $\alpha_D = 0.6$	SRNet	-	**0.2389**	0.1003	0.0466	0.0224	**0.0101**
		SCA-SRNet	-	0.2419	**0.0908**	**0.0426**	**0.0179**	0.0126
	UED	SCA-GFR	-	0.2483	0.1300	0.0727	**0.0401**	**0.0218**
	MB UED $\alpha_D = 0.6$	SCA-GFR	-	**0.3457**	**0.1833**	**0.0941**	**0.0473**	0.0209
95	UED	SRNet	-	0.2966	0.1997	0.1253	0.0818	0.0534
		SCA-SRNet	-	0.2764	0.1725	0.1098	0.0658	0.0413
	MB UED $\alpha_D = 0.6$	SRNet	-	0.4036	0.2669	0.1696	0.1113	0.0625
		SCA-SRNet	-	**0.3720**	**0.2337**	**0.1415**	**0.0842**	**0.0474**
	UED	SCA-GFR	-	0.4000	0.3141	0.2280	**0.1641**	0.1094
	MB UED $\alpha_D = 0.6$	SCA-GFR	-	**0.4398**	**0.3525**	**0.2537**	**0.1702**	**0.1140**

Table 4: For completeness, this table shows the actual numerical values of the detection error P_E for all experiments in the main body of the paper that are reported only in a graphical form. All results with rich models are on BOSSbase 512×512 images with ensemble classifier as the detector. SRNet results are always on the union BOSSbase + BOWS2 downsampled to 256×256. For the JPEG domain, the smallest studied payload is 0.1 bpnzac.