Frustratingly Easy Domain Adaptation for Image Steganalysis with CORelation ALignment

Antoine Mallet
Univ. Lille, CNRS, Centrale Lille
UMR 9189 CRIStAL
F-59000 Lille, France
antoine.mallet@univ-lille.fr

Jan Butora
Univ. Lille, CNRS, Centrale Lille
UMR 9189 CRIStAL
F-59000 Lille, France
jan.butora@cnrs.fr

***Abstract*—Detecting steganography in images is challenging, in particular when also facing a discrepancy between distributions of train and test data. This paper experiments with a simple covariance alignment procedure to perform image steganalysis in the context of cover source mismatch. We show that this approach can significantly improve the performance of a steganalyzer trained on aligned features. Most notably, we show that this domain adaptation strategy can be used in a holistic steganalysis setup, where we train a single classifier on a subset of representative cover sources that generalize well.**

Experimental results demonstrate the improvement of the proposed scheme over state-of-the-art set-covering steganalysis scheme, with a performance gain of up to 40.5% on average, in terms of regret. We also present its advantages in terms of budget, and discuss its current limitations against wild test distributions.

***Index Terms*—Steganography, Steganalysis, Domain Adaptation, Corelation Alignment, Cover source mismatch**

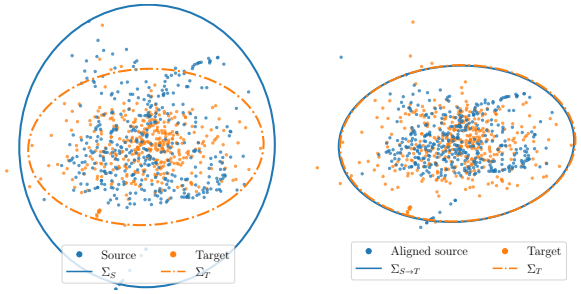
I. Introduction

Image steganography consists in hiding a secret communication within an image, such that the distorted image (the stego image) is visually indistinguishable from the original one (the cover image). Generally, the goal of steganalysis – the adversary – is to tell cover and stego apart.

In an ideal scenario, the steganalyst has access to all the information about the steganographic algorithm but the secret key [21]. Considering that he has access to a set of cover images, he can then generate a set of stego images and train a detector to classify the target cover and stego images.

In a more realistic scenario, however, he does not have access to the cover images used by the steganographer, and has to rely on a set of covers from a different source. If this cover source is very different from the target one, training a classifier on these covers and their corresponding stegos might lead to poor performance when testing on the target: this is the infamous cover source mismatch problem [16]. Cover source mismatch (CSM) has been studied thoroughly in the last few years to understand its causes and effects on the steganalysis task. Foundational papers on the subject include the empirical study of Giboulot *et al.* [9], and the formalization of the problem by Ker *et al.* [12].

This work was supported by the French government grant managed by the Agence Nationale de la Recherche under the France 2030 program, reference ANR22-PECY-0011.



(a) $D_{KL} = 0.252$, Regret $= 26.12\%$. (b) $D_{KL} = 7.19 \cdot 10^{-4}$, Regret $= 6.75\%$.

Fig. 1: Illustration of the effect of covariance alignment on the feature distributions of two cover sources. The two cover sources are defined by different processing operations. The features are DCTR, projected here on their third and fifth principal components. The distributions of the features before alignment (a) are significantly different, but almost perfectly aligned after CORAL (b). The Kullback-Leibler divergence between the distributions is given, along with the regret of steganalysis when training on the source data and testing on the target, before and after alignment.

As it corresponds to a practical problem for deploying steganalysis, it has been the subject of many studies over the last 2 decades [19]. During this time, three main approaches have emerged to tackle the CSM: *Holistic steganalysis* consists in training a single classifier on a large and diverse set of cover sources, such that it generalizes well to unseen ones (see for instance [1], [17]). *Atomistic steganalysis*, on the other hand, consists in training a separate steganalyzer for each cover source considered in training, and relies on a cover source identifier to select the appropriate steganalyzer during the testing phase (see [11], [15]). *Domain Adaptation* consists in, either aligning the train and test distributions before training a detector in this aligned setup [3], or direclty learn a mapping to some domain-invariant feature space [29].

The present study fits in the line of feature alignment approaches. In particular, it proposes to align the covariance of the feature distributions. We will show that this simple approach is effective for steganalysis in the context of CSM, as illustrated in Fig. 1.

A. Tackling the CSM with Domain Adaptation: Prior arts

Domain adaptation is a well-known problem in general machine learning. Therefore, the literature on domain adaptation-based steganalysis is rich with elaborate methods such as transfer component analysis [14], contrastive learning [30], and confidence-based pseudo-labeling [31].

More specifically, feature alignment approaches often rely on minimizing some distance between the training and testing distributions. One popular example is the maximum mean discrepancy (MMD), which can be used as a regularization term in the loss function of a neural network [31], [32], or to visualize the correlation between a discrepancy measure and empirical performance [2].

Among other alignment strategies, correlation alignment (CORAL) has received some attention in the steganalysis literature. It consists in ensuring that the second-order statistics of the feature distributions match. This was formalized for general-purpose domain adaptation in the work of Sun *et al*. [23] in 2016, and essentially consisted in a simple linear transformation of one distribution to match the other, based on the regularized covariance matrices of the two distributions. Within the field of steganalysis, which began to integrate deep learning methods shortly after, this approach was seldom used. Rather, it is their subsequent work [24], which proposed to formulate this alignment as a regularization term, that received some attention. It was most notably used for text steganalysis by Xue *et al*. [26] and Yang *et al*. [28]. In the scope of image steganalysis, Yang *et al*. [27] suggest using it, but ultimately prefer MMD in a privacy-preserving approach. More recently, Abecidan *et al*. [3] integrated this loss in their cover source approximation model for image steganalysis.

To the best of the author's knowledge, the only work that uses the initial proposition of Sun *et al*. [23] for image steganalysis is the one of Wu *et al*. [25], where they work on the high-frequency components of the image pixels.

B. Positioning of the present study and contributions

Contrary to previous works, the present study proposes to investigate the original CORAL method [23] to align the distributions of the training and testing features. In particular, we compare the performance obtained when using hand-crafted features, such as the popular DCTR [10]. We start by recalling the formalization of the CSM and define the proposed usage of CORAL in Section II, before describing the experimental setup in Section III. Then in Section IV, we show that the covariance shift between covers and stegos of a given cover source is very weak compared to the shift between the features coming from different cover sources. This first observation renders the subsequent use of CORAL for steganalysis quite natural. We empirically show that it significantly improves the performance of a steganalyzer trained on the aligned features. We then propose using CORAL to refine the set-covering approach to holistic steganalysis, which was proposed by Abecidan *et al*. [1].

II. Correlation Alignment for Steganalysis

A. Formalization of the cover source mismatch

Let us begin by formalizing the CSM problem [20]. We characterize a cover source p by the entirety of the processing operations and their attached parameters. Furthermore, we define a steganalyzer as the binary classifier h such that:

$$\begin{aligned} h(x|\theta_{p,\gamma}) : X &\to \{cover, stego\} \\ x &\mapsto y, \end{aligned} \tag{1}$$

where $\theta_{p,\gamma}$ is the trainable parameters of the classifier, which depend on p and the steganographic scheme γ.

We call intrinsic difficulty of the cover source p the error rate of h when it is trained and tested on p. We note it $D(h,p)$:

$$D(h,p) = \mathbb{E}_{(x,y)\sim\mathbb{P}((x,y)|p,\gamma)}(h(x|\theta_{p,\gamma}) \neq y). \tag{2}$$

We then define the inconsistency of a steganalyzer as its error rate when trained on a cover source p_1 and tested on a different cover source p_2:

$$I(h,p_1,p_2) = \mathbb{E}_{(x,y)\sim\mathbb{P}((x,y)|p_2,\gamma)}(h(x|\theta_{p_1,\gamma}) \neq y). \tag{3}$$

The classical empirical measure of the CSM problem, the regret, is then defined as the difference between the difficulty and the inconsistency of h on the test cover source:

$$R(h,p_1,p_2) = I(h,p_1,p_2) - D(h,p_2). \tag{4}$$

Note that this measure is generally non symmetric, i.e. $R(h,p_1,p_2) \neq R(h,p_2,p_1)$, and that it usually varies between 0 (the case of perfect generalization) and 0.5 (the case of no generalization at all).

B. The CORAL algorithm

Let $f : \mathcal{I} \to \mathbb{R}^d$ be a feature extractor that maps an image $I \in \mathcal{I}$ to a d-dimensional real-valued feature. Let p_S be the "source" cover source of our domain adaptation setup, i.e. we have $X_S = [x_1^{(S)}, \dots, x_n^{(S)}] \in \mathbb{R}^{n\times d}$ the matrix containing the n "source" samples $x_i^{(S)} = f(I_i^{(S)})$ and $l^{(S)} = [l_1^{(S)}, \dots, l_n^{(S)}] \in \{0,1\}^n$ their associated class label (cover or stego). Let also p_T be the "target" cover source and $X_T = [x_1^{(T)}, \dots, x_m^{(T)}] \in \mathbb{R}^{m\times d}$ the target cover source samples $x_i^{(T)} = f(I_i^{(T)})$ for which we have no labels $l^{(T)}$. Let us finally denote Σ_S and Σ_T the empirical covariance matrices of the source and target features, respectively. The CORAL algorithm consists in finding a linear transformation $A \in \mathbb{R}^{d\times d}$ such that the covariance of the transformed source features matches the covariance of the target features:

$$A^* = \min_A ||A^\top \Sigma_S A - \Sigma_T||_F^2, \tag{5}$$

where $||\cdot||_F$ is the Frobenius norm. By denoting the SVD decompositions of Σ_S and Σ_T as $U_S\sigma_S U_S^\top$ and $U_T\sigma_T U_T^\top$, respectively, and without loss of generality, Sun *et al*. [23] show that the optimal solution of this problem is given by:

$$A^* = \left(U_S\sigma_S^{-\frac{1}{2}}U_S^\top\right)\left(U_{T[1:r]}\sigma_{T[1:r]}^{\frac{1}{2}}U_{T[1:r]}^\top\right), \tag{6}$$

with $r = \min(\text{rank}(\Sigma_S), \text{rank}(\Sigma_T))$.

In practice, the CORAL algorithm can be implemented by first whitening the source features, and then recoloring them with the target covariance matrix:

$$X_{S\to T} = X_S(\Sigma_S + \lambda I_d)^{-\frac{1}{2}}(\Sigma_T + \lambda I_d)^{\frac{1}{2}}, \tag{7}$$

where λ is a regularization term to stabilize inversion[1].

On the CORAL aligned source features $X_{S\to T}$, we then train a steganalyzer h, and test it on the target features X_T, hoping that the regret after CORAL is significantly lower than the regret before:

$$R_{\text{CORAL}}(h, p_S, p_T) = I(h, p_{S\to T}, p_T) - D(h, p_T). \tag{8}$$

C. Refining set covering for holistic steganalysis with CORAL

One of the main limitations of the holistic approach is that, in practice, the larger and the more diverse the training set is, the better the generalization ability of the steganalyzer will be. A tractable approach to this problem was proposed by Abecidan *et al*. [1], who proposed populating the training set with a set of cover sources that inherently generalize well to other cover sources. Formally, let us define a set of cover sources $\mathcal{P} = \{p_i\}_{1\leq i\leq N}$. We can then define a covering set $\mathcal{C} \subset \mathcal{P}$ as a set of cover sources such that:

$$\forall\, p_i \in \mathcal{P}\ ,\ \exists\, p_j \in \mathcal{C}\ \text{ s.t. }\ R(h, p_j, p_i) \leq \epsilon, \tag{9}$$

with ϵ an arbitrary acceptable regret threshold. Finding the minimal covering set $\mathcal{C}^*$ is NP-hard. Heuristics with theoretical guarantees exist, such as the greedy algorithm proposed by Chvatal [6], which was used by Abecidan *et al*. [1]. In this paper, we also propose to investigate the method of Caprara *et al*. [5]. For steganalysis, the covering of each source is computed by measuring the regret of a steganalyzer trained on that source and tested on all the other sources. A cover source p_j is an ϵ-cover of the cover source p_i if $R(h, p_j, p_i) \leq \epsilon$.

This was shown to be an effective approach to reduce the number of cover sources needed to train a holistic steganalyzer. However, we hypothesize here that the use of CORAL to align the features of the cover sources before computing the covering set could further reduce the number of cover sources needed to train a holistic steganalyzer. In this paper, we will compare the covering set computed using Eq. (9) to the covering set $\mathcal{C}_{\text{CORAL}}$ computed using CORAL-aligned features:

$$\forall\, p_i \in \mathcal{P}\ ,\ \exists\, p_j \in \mathcal{C}_{\text{CORAL}}\ \text{ s.t. }\ R(h, p_{j\to i}, p_i) \leq \epsilon. \tag{10}$$

III. Experimental setup

A. Cover sources and steganography

To evaluate the proposed approach, we use the Alaska dataset [7], which consists of RAW images from different camera models. We apply a controlled set of image processing operations to generate a set of cover sources. The operations are listed in Fig. 2, and are as follow: first, linear or PPG demosaicing is applied; resampling is then performed, followed by unsharp masking, for which different values for the amount and radius are used. Finally, the images are saved in PNG format. In total, 50 different cover sources are generated. They are eventually JPEG compressed at QF 95 for the experiments presented below. Each cover source contains 5000 cover images. The steganographic scheme used in the experiments is UERD [18], with a payload $\rho \in \{0.1, 0.2, 0.4\}$ bit per non-zero AC (bpnzAC).

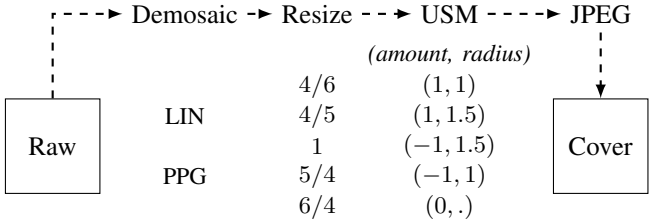


Fig. 2: Image processing pipeline generating the cover sources. RAW images (from the Alaska dataset) are demosaicked, resampled and sharpened or denoised using unsharp masking. Images are then JPEG compressed at QF 95.

B. Steganalysis & comparative baselines

To detect cover and stego images, we propose to train logistic regression classifiers on DCTR features [10]. The relatively low dimensionality of DCTR ($d = 8000$) compared to other features for the JPEG domain like GFR ($d = 17000$ [22]) or JRM ($d = 35263$ [13]) facilitates the computation of the inverse square root of the covariance matrices[2]. Features are centered before training and testing, but not normalized.

To evaluate the effectiveness of CORAL, we compare it to the following common baselines. First, we compare it to the traditional CSM scenario, where the steganalyzer is trained on unaligned features of a source and tested on the target. Second, we compare it to the holistic steganalyzer, which is still trained on unaligned data, but this time coming from all the cover sources of the training pool. Third, we compare it to the set covering holistic steganalyzer, which is trained on the minimal covering set of cover sources – as defined in Eq. (9) – except the target. Note that all the experiments are carried out in a "leave-the-target-out" manner, i.e., the target cover source is never seen during training.

Unless otherwise specified, the reported performance is either the regret (Eq. (4)) or the CORAL-regret (Eq. (8)).

IV. Results analysis

A. Comparing class-wise and source-wise covariance shift

One of CORAL's limitations resides in the assumption that the covariance is much more similar across the classes than across domains. In steganalysis, we know that the cover source noise is much stronger than the stego noise, which is why we observe such performance drops. Furthermore, it tends to be spatially independent in the pixel domain [8]. And in the feature space, measuring the chordal distance between the manifolds of each class also tends to go in that direction [2].

[1]Singular value clipping, proposed in Eq. (6), forces positive semi-definiteness, like the regularization term, and was used by [25] in practice.

[2]We leave to future work the idea of dimensionality reduction to stabilize the inversion, although we deem it a promising approach.

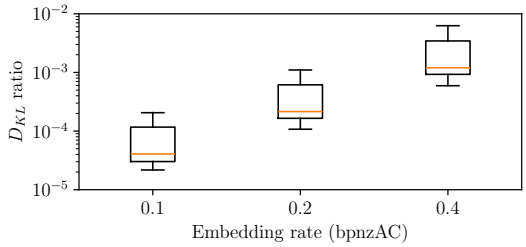


Fig. 3: Distribution of relative discrepancies $r_{i,j}$ for different embedding rates, represented in log-scale.

In the present study, we propose to look at the ratio between the inter-class Kullback-Leibler divergence over the inter-domain one, i.e.:

$$r_{i,j} = \frac{D_{KL}\left[C_i||S_i\right]}{D_{KL}\left[C_i||C_j\right]}, \tag{11}$$

where C_i and S_i represent the cover and stego distributions of the i-th cover source, respectively. The Kullback-Leibler divergence D_{KL} is computed assuming Gaussian distributions with empirical covariance matrices and identical means. The results are shown in Fig. 3. They highlight that the discrepancies between cover and stegos are – on average – two to five orders of magnitude smaller than the discrepancies between cover sources, which we consider sufficient to apply CORAL to the present problem. Unsurprisingly, we observe that the relative discrepancies increase with the embedding rate, which is consistent with the fact that the stego noise becomes more significant as the embedding rate increases.

B. Regret with CORAL

Before applying the correlation alignment to refine the set covering approach, we first evaluate the performance of single-source alignment with CORAL. To do so, we compute the regret of a steganalyzer trained on a cover source and tested on another, before and after CORAL. We then average the obtained regrets across all pairs of distinct cover sources, and report the results in Table I. We see that the average regret is significantly reduced after CORAL, with 24.6%, 26.8%, and 29.1% decreases for embedding rates of 0.1, 0.2, and 0.4 bpnzAC, respectively. Note that, unsurprisingly, the increase is smaller for lower embedding rates, where the steganalysis is very difficult with high intrinsic difficulties.

To get a more detailed view on the improvement of CORAL, Fig. 4 shows scatter plots of average regrets for each cover source taken as a source (x-axis) and as a target (y-axis), before (in blue) and after CORAL (in red). We observe a homogeneous improvement across all cover sources and all embedding rates, which highlights the effectiveness of CORAL. On average, the regret decreases by 1.92%, 4.06%, and 5.25% for embedding rates of 0.1, 0.2, and 0.4 bpnzAC, respectively. Note that the gain is diagonal, as we average on both sources and targets to get the mean coordinates.

C. Set covering with and without CORAL

We now measure the regret obtained by a set-covering-based steganalyzer, averaged for different thresholds $\epsilon \in \{1\%, 5\%, 10\%\}$, for the two considered algorithms (Greedy [1] and Caprara's [5]), and for the different embedding rates.

	Embedding rate (bpnzAC)					
	0.1		0.2		0.4	
$R(h, p_i, p_j)$	7.80		15.11		17.98	
$R_{\text{CORAL}}(h, p_i, p_j)$	5.88	(-1.92)	11.05	(-4.06)	12.74	(-5.25)

TABLE I: Average regret $\mathbb{E}_{i,j,i\neq j}\left[.\right]$ with and without CORAL for different embedding rates. The average is computed over The difference between the two is shown in green.

Table II presents the results. CORAL with set-covering (CORAL-SC) consistently reduces regret, and on average by 1.84% using the greedy algorithm and 1.76% using Caprara's, corresponding to a 40.5% and a 36.9% decrease, respectively. In particular, we can note that for $\rho = 0.1$ bpnzAC and $\epsilon = 10\%$, the greedy version of CORAL-SC even very slightly outperforms the Clairvoyant scenario on average.

Interestingly, we see that the greedy algorithm tends to perform better than Caprara's, with and without CORAL. Note, also, that the small regrets obtained for $\rho = 0.1$ bpnzAC are due to the fact that the steganalysis is very difficult at that payload, and thus the small regrets are partially a consequence of the high intrinsic difficulties.

D. Comparing covering set size with and without CORAL

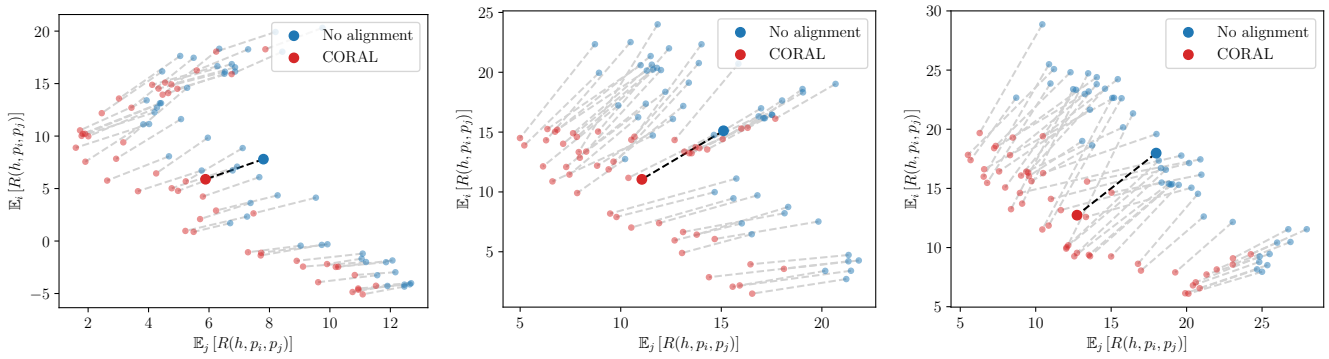
In this short experiment, we compare the size of the covering sets obtained with and without CORAL, for different thresholds ϵ and embedding rates ρ. For each pair, we average over the 50 cover sets obtained and over the two covering algorithms. The results are shown in Fig 5.

First, as hypothesized, we observe that CORAL always reduces the size of the cover sets on average. Second, we naturally observe smaller covering sets when ϵ gets larger. Third, we observe that for a fixed ϵ, lower embedding rates induce smaller covering sets. This is, once again, consistent with the fact that for harder tasks, the regret cannot get very high.

E. CORAL against holistic for different budget

We now question whether CORAL efficiently reduces the number of training samples needed to achieve a given performance. Results are presented in Table III for different numbers of training samples and at different embedding rates. We report the average performance over all the cover sources taken as the target, and for $\epsilon = 5\%$ for the greedy CORAL-SC, compared to the traditional holistic steganalyzer.

CORAL improves the regret across all cases. The gain is -4.64% on average, with a relatively stable gain across the different budgets, varying between -4.1% gain with 1000 train samples and -5.07% with 2000. This can be explained by the fact that for a low amount of samples, the detector underfits, even with CORAL. On the other hand, with $10,000$ training samples, the effect of CORAL starts to diminish compared to the straightforward holistic approach. Importantly, note that given a budget, CORAL-SC consistently outperforms the holistic model with up to 2.5 times the budget.



(a) 0.1 bpnzAC. Avg. gain: -1.92% (b) 0.2 bpnzAC. Avg. gain: -4.06% (c) 0.4 bpnzAC. Avg. gain: -5.25%

Fig. 4: Scatter plots of the average regret for each cover source taken as a source domain (x-axis) and each cover source taken as a target domain (y-axis). Regret before CORAL is shown in blue, and regret after CORAL in red. Each dotted line connects the regret before and after CORAL for a given cover source. The large dots represent the average regret across all cover sources, before and after CORAL. The dark line thus represents the average performance gain obtained with CORAL.

	ρ (bpnzAC)	0.1				0.2				0.4				**Avg.**
	ϵ	1%	5%	10%	Avg.	1%	5%	10%	Avg.	1%	5%	10%	Avg.	
Greedy [1]	SC	1.66	1.56	0.75	1.32	5.34	6.12	5.12	5.53	5.93	5.96	8.41	6.77	**4.54**
	CORAL-SC	0.49	0.70	-0.24	0.31	3.75	3.82	3.07	3.55	4.02	3.85	4.88	4.25	**2.70**
	Difference	-1.17	-0.86	-0.99	-1.01	-1.59	-2.30	-2.05	-1.98	-1.91	-2.11	-3.53	-2.52	$\mathbf{-1.84}$
Caprara [5]	SC	1.33	1.37	1.72	1.47	5.56	6.39	5.48	5.81	5.83	6.09	9.13	7.02	**4.77**
	CORAL-SC	0.44	0.83	0.68	0.65	3.78	4.70	3.31	3.93	4.26	3.85	5.21	4.44	**3.01**
	Difference	-0.89	-0.54	-1.04	-0.82	-1.78	-1.69	-2.17	-1.88	-1.57	-2.24	-3.92	-2.58	$\mathbf{-1.76}$

TABLE II: Regret of the set covering (SC) with and without CORAL, with respect to different embedding rates ρ and thresholds ϵ, and for greedy and Caprara's SC algorithms. Each value is averaged over the 50 cover sources. The average gain is given over different values of ϵ (i.e., for each payload and SC algorithm) and for both SC algorithms.

F. Generalizability to out-of-camera images

We finally evaluate CORAL-SC on out-of-camera images coming from ALASKA [7]. The camera models are a Huawei P40, an iPhone 11 Pro, a Nikon D-810 and a D-850, and a Xiaomi 10T. Each OOC source is approximately 1000 images. To cope with mismatched compression, we first compress our base of images from PNG using the quantization table for each source. We then train our detectors on features from these images, and the set-covering is computed using the results of these detectors. As before, the cover set is used to train one classifier, and we measure its out-of-distribution performance, i.e., on the OOC sources.

Results are shown in Table IV. This time, we explicitly report the ID performance (first row). In comparison, we see that both set covering and CORAL-SC significantly underperform, respectively increasing the error rate from 12.63% to 49.83% ($+37.20\%$) and 42.07% ($+29.44\%$) on average. Nevertheless, using CORAL lowers the error on average. Several elements can explain these mitigated results. On the one hand, based on the inability of the regular set-covering to generalize, it is clear the source and target distributions are strongly misaligned. On the other, we believe that CORAL's struggles suggest three things. First, the shapes of the distributions are too different to be effectively aligned – in which case taking higher orders should be important. Second, the number of samples is too small to effectively capture the covariance of the underlying distribution – a problem that could be limited by reducing the feature dimension. Third and last, it is possible that the chosen images are not uniformly processed, as it is hard to know precisely how proprietary devices process images. All these concerns shall be addressed in future work.

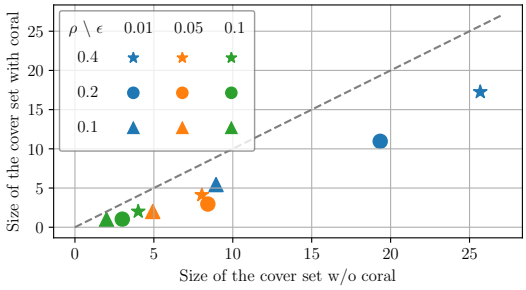


Fig. 5: Average size of the covering sets obtained with and without CORAL, for all $\{\rho, \epsilon\}$ pairs.

V. Conclusion and Future Work

This work presents a simple – yet effective – approach to tackle the cover source mismatch problem in image steganalysis, using a covariance alignment procedure to align the feature distributions of training and testing data. We show that this approach consistently improves the performance of a

Budget (# total)	ρ (bpnzAC)	0.1	0.2	0.4	Avg.
1000	Holistic	11.05	19.52	16.28	**15.62**
	CORAL-SC	8.16	13.58	12.81	**11.52**
	Difference	−2.89	−5.94	−3.47	**−4.10**
2000	Holistic	10.51	15.37	13.50	**13.46**
	CORAL-SC	5.72	9.74	8.43	**7.96**
	Difference	−4.79	−5.36	−5.07	**−5.07**
5000	Holistic	7.44	11.08	10.19	**9.57**
	CORAL-SC	2.07	5.88	6.01	**4.65**
	Difference	−5.37	−5.49	−4.18	**−5.01**
10000	Holistic	5.31	8.14	7.85	**7.10**
	CORAL-SC	0.60	3.01	4.58	**2.73**
	Difference	−4.71	−5.14	−3.27	**−4.37**
	Avg. Diff.	**−4.44**	**−5.48**	**−3.99**	**−4.64**

TABLE III: Regret of holistic and greedy CORAL-SC ($\epsilon = 5\%$) steganalyzers for different budgets and embedding rates. Results are averaged over all cover sources taken as the target.

OOC	Huawei	iPhone	D810	D850	Xiaomi	Avg.
ID P_E	4.76	23.57	2.92	24.73	7.17	**12.63**
SC	47.58	50.24	50.29	51.93	49.12	**49.83**
CORAL-SC	28.70	46.81	43.52	49.61	41.72	**42.07**
Diff.	−18.88	−3.43	−6.77	−2.32	−7.40	**−7.76**

TABLE IV: Performance in-distribution on the out-of-camera images, and OOD of the SC and CORAL-SC steganalyzers (against UERD with $\rho = 0.4$ bpnzAC, and $\epsilon = 5\%$).

steganalyzer trained on the aligned features. Combined with a set covering approach, we show that we jointly reduce the number of cover sources in the minimal covering set and improve the performance of the steganalyzer trained on this set. Furthermore, we show that the proposed approach is budget-friendly, which is a key point in practical scenarios where the training time is limited. Finally, we discuss the current limitations of the proposed method against out-of-camera images, and draw several paths to overcome this key operational task.

Critically, this work refrained from using deep learning features, such as JIN's [4] penultimate layer representations. This is because deep learning features tend to suffer from mode collapse, which makes covariance inversion unstable (and recoloring impossible, should the OOD features be on an entirely different sub-manifold). Future work shall investigate methods to apply CORAL to any steganalysis features, be it very large-dimensional or collapsed ones.

REFERENCES

[1] R. Abecidan, V. Itier, J. Boulanger, P. Bas, and T. Pevnỳ. Using set covering to generate databases for holistic steganalysis. In *2022 IEEE WIFS*, 2022.

[2] R. Abecidan, V. Itier, J. Boulanger, P. Bas, and T. Pevnỳ. Leveraging data geometry to mitigate csm in steganalysis. In *2023 IEEE WIFS*. IEEE, 2023.

[3] R. Abecidan, V. Itier, J. Boulanger, P. Bas, and T. Pevnỳ. Tackle csm in jpeg steganalysis with data adaptation. In *ACM IH&MMSec*, 2026.

[4] J. Butora, Y. Yousfi, and J. Fridrich. How to pretrain for steganalysis. In *ACM IH&MMSec*, 2021.

[5] A. Caprara, M. Fischetti, and P. Toth. A heuristic method for the set covering problem. *Operations research*, 1999.

[6] V. Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of operations research*, 1979.

[7] R. Cogranne, E. Giboulot, and P. Bas. The alaska steganalysis challenge: A first step towards steganalysis into the wild. In *ACM IH&MMSec*, 2019.

[8] E. Giboulot, P. Bas, and R. Cogranne. Multivariate side-informed gaussian embedding minimizing statistical detectability. *IEEE TIFS*, 2022.

[9] E. Giboulot, R. Cogranne, D. Borghys, and P. Bas. Effects and solutions of cover-source mismatch in image steganalysis. *SPIC*, 2020.

[10] V. Holub and J. Fridrich. Low-complexity features for jpeg steganalysis using undecimated dct. *IEEE TIFS*, 2014.

[11] X. Hou, T. Zhang, G. Xiong, Z. Lu, and K. Xie. A novel steganalysis framework of heterogeneous images based on gmm clustering. *SPIC*, 2014.

[12] A. D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevnỳ. Moving steganography and steganalysis from the laboratory into the real world. In *ACM IH&MMSec*, 2013.

[13] J. Kodovskỳ and J. Fridrich. Steganalysis of jpeg images using rich models. In *MWSF*. SPIE, 2012.

[14] X. Li, X. Kong, B. Wang, Y. Guo, and X. You. Generalized transfer component analysis for mismatched jpeg steganalysis. In *2013 IEEE ICIP*. IEEE, 2013.

[15] A. Mallet, P. Bas, and R. Cogranne. Statistical correlation as a forensic feature to mitigate the cover-source mismatch. In *ACM IH&MMSec*, 2024.

[16] A. Mallet, M. Beneš, and R. Cogranne. Cover-source mismatch in steganalysis: systematic review. *EURASIP JIS*, 2024.

[17] W.W. Ng, Z. He, D. S Yeung, and P. P. Chan. Steganalysis classifier training via minimizing sensitivity for different imaging sources. *Information Sciences*, 2014.

[18] Y. Pan, J. Ni, and W. Su. Improved uniform embedding for efficient jpeg steganography. In *International Conf. on Cloud Computing and Security*. Springer, 2016.

[19] T. Pevnỳ and J. Fridrich. Towards multi-class blind steganalyzer for jpeg images. In *IWDW*. Springer, 2005.

[20] D. Šepák, L. Adam, and T. Pevnỳ. Formalizing cover-source mismatch as a robust optimization. In *EUSIPCO*, 2022.

[21] C. E. Shannon. Communication theory of secrecy systems. *The Bell system technical journal*, 1949.

[22] X. Song, F. Liu, C. Yang, X. Luo, and Y. Zhang. Steganalysis of adaptive jpeg steganography using 2d gabor filters. In *ACM IH&MMSec*, 2015.

[23] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI conf. on AI*, 2016.

[24] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*. Springer, 2016.

[25] R. Wu and Z. He. Cross-domain image steganalysis based on frequency-domain alignment and feature calibration. *IEEE SPL*, 2025.

[26] Y. Xue, J. Wu, R. Ji, P. Zhong, J. Wen, and W. Peng. Adaptive domain-invariant feature extraction for cross-domain linguistic steganalysis. *IEEE TIFS*, 2023.

[27] H. Yang, H. He, W. Zhang, and X. Cao. Fedsteg: A federated transfer learning framework for secure image steganalysis. *IEEE Trans. on Network Science and Engineering*, 2020.

[28] Z. Yang, Y. Luo, J. Yang, X. Xu, R. Zhang, and Y. Huang. Class-aware adversarial unsupervised domain adaptation for linguistic steganalysis. *IEEE TIFS*, 2025.

[29] X. Ye, R. Tan, H. Chen, H. Wang, and H. Chen. Enhancing image steganalysis via multi-source domain invariant feature learning. *The Visual Computer*, 2026.

[30] L. Yu, S. Weng, M. Chen, and Y. Wei. Rcdd: Contrastive domain discrepancy with reliable steganalysis labeling for cover source mismatch. *Expert Systems with Applications*, 2024.

[31] L. Zhang, S. M. Abdullahi, P. He, and H. Wang. Dataset mismatched steganalysis using subdomain adaptation with guiding feature. *Telecommunication Systems*, 2022.

[32] X. Zhang, X. Kong, P. Wang, and B. Wang. Cover-source mismatch in deep spatial steganalysis. In *IWDW*. Springer, 2019.