Mallet *et al.*
EURASIP Journal on Information Security (2024) 2024:26
https://doi.org/10.1186/s13635-024-00171-6

EURASIP Journal on Information Security

SURVEY

Open Access

Cover-source mismatch in steganalysis: systematic review

Antoine Mallet[1*], Martin Beneš[2] and Rémi Cogranne[1*]

Abstract

Operational steganalysis contends with a major problem referred to as the cover-source mismatch (CSM), which is essentially a difference in distribution caused by different parameters and settings over training and test data. Despite it being of fundamental importance in an operational context, the CSM problem is often overlooked in the literature. With the goal to increase the visibility of this problem and attract the interest of the community, the present paper proposes a systematic review of the literature. It summarizes gathered knowledge and major open questions over the last 20 years of active research on CSM: terminology, methods of measurement, known causes, and mitigation strategies. Over 100 papers exploring, mitigating, assessing, or discussing steganalysis under train-test mismatch were collected by sampling scholar databases, and tracing references, cited and generated. For image steganalysis, the literature provided enough evidence to quantify the impact of causes, and the effectiveness of mitigation strategies.

1 Introduction

Steganography is often referred to as the art and techniques of cover communication. It aims at secretly exchanging sensitive information by hiding it in a so-called cover-object. This creates a stego-object which, to preserve the furtiveness of the secret communication, should look as inconspicuous as possible.

To ease this process, the cover-object ought to be commonly encountered, in order not to raise suspicion. It should be easy to modify and carry enough entropy to accommodate the secret message; good examples of such suitable cover-objects include digital media—images, audio, or video—texts, computer network packets, or even program executable codes [24, 44].

As for any secured communication, the stego-object is sent over an insecure channel which, in the worst case, is assumed monitored or controlled by an adversary referred to as the steganalyst. Unlike the other scenario of communication security, the steganalyst aims, in the very first place, at detecting the presence of a hidden secret message either by thorough statistical analysis or by searching for trademarks or "signatures" of a specific technique.

Steganography and steganalysis, thus, constitute a game of cat and mouse. In academic studies, the Kerckhoffs' principle[1] is often advocated to justify that the steganalysis is carried out with knowledge on all necessary properties of the inspected objects. This also includes the potential embedding method as well as access to large representative datasets [71, 83]. Of course, steganography has been developed in this setting, which is the most stringent.

On the opposite, in a real-world operational context, the steganographer and the steganalyst only have very

*Correspondence:
Antoine Mallet
antoine.mallet@utt.fr
Rémi Cogranne
remi.cogranne@utt.fr
[1] Univerité de Techonologie de Troyes, Troyes, France
[2] Univerity of Innsbruck, Innsbruck, Austria

[1] The Kerckhoffs' principle essentially states that security must always rely solely on the key and that the rest of the communication system and its settings must be publicly known.

limited access to each other's information. The steganalyst selects a detector, and the steganographer picks the steganographic embedding and the cover [29]. The exact original cover-object is unknown to the steganalyst, but it is generated from a noisy process which is defined as the *cover-source*. As advocated in [53], the steganalyst can hardly know this *cover-source*; it can only, at best, be estimated with an accuracy that depends on the nature and number of objects the steganographer provided.

This scenario naturally raises the problem, for the steganalyst, of "designing" or "training" a detector on a *cover-source* that differs from the one used by the steganographer: this is referred to as the *cover-source mismatch* (CSM).

In the broad field of statistical learning, this phenomenon is known as the *distribution shift* and it occurs when the statistical properties of training and testing data differ. The main symptom is a deterioration of the model performance in the production environment. Distribution shift occurs in all possible applications of statistics and machine learning, such as to cite a few, medical imaging [33], computer vision [79], reinforcement learning [111], natural language processing [8], and speech recognition [32].

However, it should be noted that the tasks in the aforementioned fields operate mostly on a semantic level. In other words, while the acquisition and processing setting do matter, they have a limited impact as compared to the presence of the pattern of interest. In addition, it is often possible to adjust the training for a specific source.

The peculiarity of steganography, and the related field of digital forensics [68], is that the signal of interest is extremely weak while the CSM has a much stronger impact; this often yields catastrophic performance drops, which make the steganalysis merely ineffective. In [28], using different capturing devices increased the error rate from 15% to random guessing. According to [44], "CSM is one of the main factors negatively affecting the deployment of steganalysis in the real world."

1.1 Related work on CSM

The symptoms of the CSM problem have been identified for a little less than 20 years, as image steganalysis was in its infancy [52, 82]. It has been empirically observed that steganalysis techniques perform differently over different datasets [14, 54]. The significance of the CSM impact on steganalysis was clearly acknowledged for the first time in 2010 during the BOSS open contest (Break Our Steganographic System) as the organizers added in the testing set data generated with a different source.[2] This period also coincided with the increasing use of machine learning in steganalysis, which requires a training phase after which the classifier can be used and evaluated on a different testing set.

While this problem was unanimously recognized as an important deadlock for practical applications of steganalysis [47], it has been only seldom studied. In particular, it was only in 2018 that the causes of the CSM problem were thoroughly studied [9, 27, 28] by comprehensive evaluations of the contribution of each step of a generic image processing pipeline (IPP) to CSM. Even the most successful *deep learning* (DL) models [11, 103] were shown to be prone to the CSM [17, 26].

In the meantime, several strategies have been suggested to mitigate the impact of CSM on steganalysis performance [5, 49, 64, 70, 78].

Despite its severity on modern steganalysis, the CSM remains largely unexplored; existing steganalysis survey papers discuss the CSM problem only briefly [42, 91]. To the authors' best knowledge, the present paper is the first systematic review of the literature on CSM and its impact on steganalysis.

1.2 Organization of the paper

To structure the knowledge gathered over 20 years of literature on the CSM, the present systematic review aims at addressing the following research questions:

> **Research question 1:** What are the studied causes of CSM in the literature?

We list the known causes of CSM studied in the literature and look into the research trend over time.

> **Research question 2:** How impactful are the known causes of CSM?

We quantitatively assess the impact of CSM causes identified by answers to RQ 1.

> **Research question 3:** What are the existing mitigation strategies against CSM?

Similar to RQ 2, we assess the relative effectiveness of the existing mitigation strategies.

At the time of writing of this survey, RQ 2 and RQ 3 can only be meaningfully answered for image media, which constituted the vast majority of the literature sampled using the strategy from Sect. 4.1. While the present paper aims at encompassing all forms of steganalysis, it only considers digital images in Sects. 5 and 6.

The rest of this survey is organized as follows: Sect. 2 describes the train-test mismatch problem in steganalysis and its implications. Section 3 presents the methods of measuring the CSM. Section 4 explains the bibliometric

[2] More precisely, the training set of BOSS was made of raw images processed with the very same script. On the opposite, the testing set included out-of-camera JPEG compressed images. For all competitors, an important drop in performance was observed on this testing subset.

methodology for literature collection and result aggregation. Section 5 surveys identified causes of CSM and quantifies their impact. Section 6 reviews strategies to mitigate CSM and quantifies their impact. Section 7 describes the game-theoretical interpretation of CSM. Section 8 discusses the findings, answers the research questions, and poses still open questions. Section 9 presents similar works in other applications of machine learning, and Sect. 10 concludes.

2 Background on mismatches

This section introduces different types of mismatches: mismatch in cover-sources (Sect. 2.1), mismatch in steganography (Sect. 2.2), and mismatch in the context of pooled steganalysis (Sect. 2.3).

2.1 Mismatch of cover-sources

Steganography is carried out in two main phases: first the steganographer generates a cover from a *cover-source*, which is a non-deterministic acquisition process followed by a processing pipeline. Both the acquisition and the processing pipeline consist of several steps that can be highly parametrized [3]. Then, this cover-object is used as an input for a steganographic algorithm, also characterized by a set of parameters, to hide a secret message into the so-called stego-object.

A problem of train-test set mismatch can occur when any parameter of any step, from the acquisition of the cover-objects to the generation of the stego-objects, differs. In practice, however, two objects are never generated in the exact same manner. In addition, different changes yield different impacts, with more or less important effects [4].

Adopting the same practical point of view that is used in almost all prior works, we shall define these concepts in relation to their use in steganography and their impact on steganalysis.

Definition 1 A **cover-source** is entirely defined by the steps both the acquisition and processing pipeline are made of, the order of these steps, and the parameters used therein.

As a consequence, a cover-source is a noisy process producing cover-objects with common statistical characteristics.

While this definition of the cover-source is rather straightforward, it is hardly related to practical uses and applications. Even the second part, the corollary that objects from the same cover-source share similar properties, remains rather impractical; indeed, it is still unclear which property has a significant impact on the usage of steganography and/or steganalysis. From a practical point of view, the cover-source has little or no impact on steganography, which very often operates over each object independently. On the opposite, steganalysis generally exploits a detector that is trained and evaluated over a set of objects with specific properties resulting from the cover-source. Because there has been a confusion in the literature between the cause and the consequence of cover-source mismatch, it is proposed to adopt the following definition from [28].

Definition 2 The causes of **cover-source mismatch (CSM)** lies in the discrepancy between distributions of samples generated by two cover-sources.

Note that, in the existing literature, the term CSM can refer to both the discrepancy between distributions and its measurable impact of this discrepancy in steganalysis. Therefore, to prevent the confusion between cause and consequence of CSM, Definition 2 dissociates the CSM from its impact on steganalysis.

For the sake of clarity, the present paper proposes a novel terminology to the problem, referring to CSM as being the discrepancy between cover-sources, and to the CSM problem as its impact on steganalysis.

Definition 3 The **cover-source mismatch problem (CSM problem)** is the degradation of a steganalyser performance observed when the training and the testing samples come from different cover-sources.

Figure 1 depicts an illustrative scenario of three cover-sources, C_1, C_2, and C_3, and exemplifies the preceding definitions. The geometric proximity corresponds to the CSM: C_2 and C_3 are more similar to each other than to C_1. Additionally, a given steganographic embedding at rate α is performed, which shifts C_1 to stego S_1 and C_2 to stego S_2. Note that the figure also illustrates that the embedding shift is often smaller and similar in direction[3] as compared to the wide diversity between cover-sources, hence the critical impact that CSM may have. These elements illustrate a typical configuration of cover-source mismatch.

Two steganalysis detectors, shown in blue and red respectively, were then trained to distinguish between the same cover-source C_1 and corresponding stego-objects S_1. The difference between the two can come from the selected model, hyperparameters, and initial weights. Both detectors discriminate well C_1 from S_1. However, the one illustrated in red is not subject

[3] This is usually referred to as the shift hypothesis [19, 20, 46].

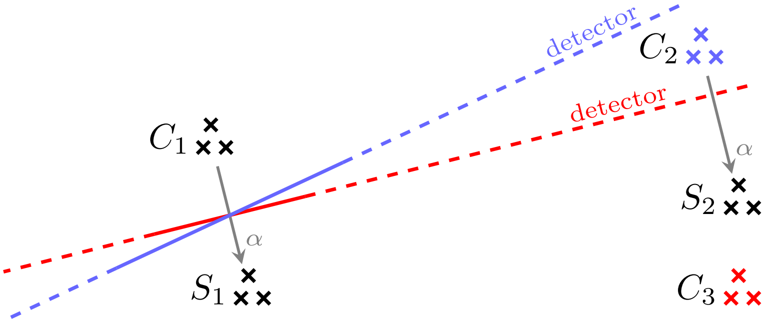


Fig. 1 Illustration of cover-source mismatch for detectors distinguishing C_1 and S_1. C_2 is misclassified by the blue detector, and C_3 by both. The embedding shift is much smaller than the distance of cover-sources

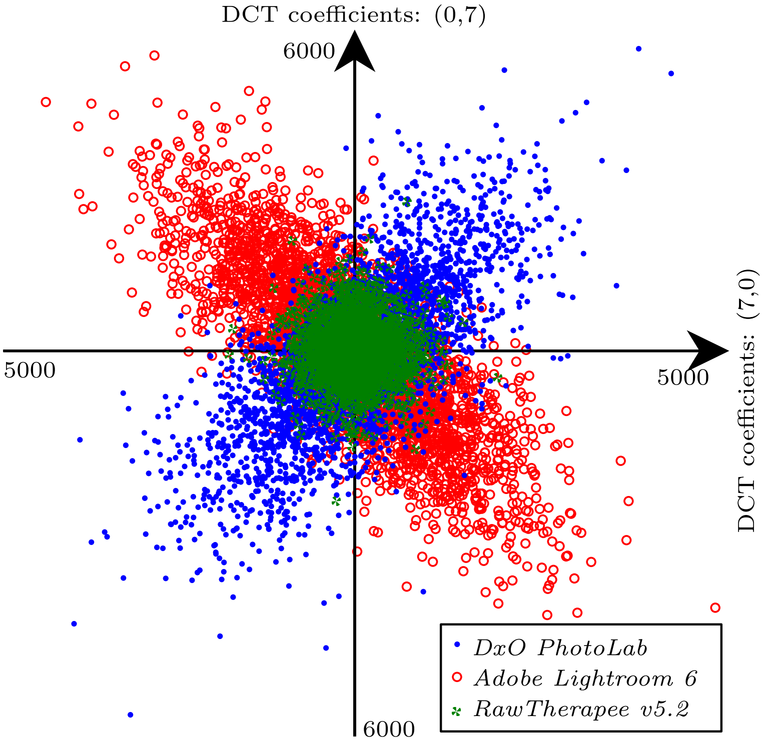


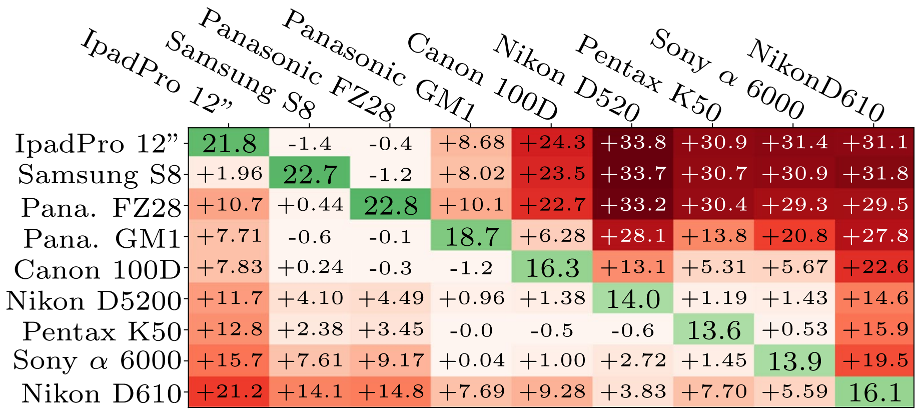
Fig. 2 Scatter plots that show the empirical joint distribution of unquantized DCT coefficients (0, 7) and (7, 0) for images corrupted with only i.i.d Gaussian noise and then developed with three different software

to the CSM problem when applied to cover-source C_2. On the opposite, both detectors perform poorly over C_3 because of the CSM problem: cover-objects erroneously belong to decision region “stego” resulting in a very high false-positive detection rate (also referred to as type I error): this is the cover-source mismatch problem.

We believe that Fig. 1, along with Definition 2, illustrates an important—and often overlooked—fact: the CSM problem, as a degradation of a detector’s performance, depends on two things: the CSM configuration itself, but also the settings of the detector. We further discuss ways of measuring CSM in Sects. 3.1 and 3.2.

The reader might also wonder how to generalize better to a very large amount of cover-sources, possibly unseen. Mitigation strategies and their assessment are addressed in detail in Sect. 6.

Let us conclude this definition section with a real-world example shown in Figs. 2 and 3 taken from [28]. First,



	IpadPro 12”	Samsung S8	Panasonic FZ28	Panasonic GM1	Canon 100D	Nikon D520	Pentax K50	Sony α 6000	NikonD610
IpadPro 12”	21.8	-1.4	-0.4	+8.68	+24.3	+33.8	+30.9	+31.4	+31.1
Samsung S8	+1.96	22.7	-1.2	+8.02	+23.5	+33.7	+30.7	+30.9	+31.8
Pana. FZ28	+10.7	+0.44	22.8	+10.1	+22.7	+33.2	+30.4	+29.3	+29.5
Pana. GM1	+7.71	-0.6	-0.1	18.7	+6.28	+28.1	+13.8	+20.8	+27.8
Canon 100D	+7.83	+0.24	-0.3	-1.2	16.3	+13.1	+5.31	+5.67	+22.6
Nikon D5200	+11.7	+4.10	+4.49	+0.96	+1.38	14.0	+1.19	+1.43	+14.6
Pentax K50	+12.8	+2.38	+3.45	-0.0	-0.5	-0.6	13.6	+0.53	+15.9
Sony α 6000	+15.7	+7.61	+9.17	+0.04	+1.00	+2.72	+1.45	13.9	+19.5
Nikon D610	+21.2	+14.1	+14.8	+7.69	+9.28	+3.83	+7.70	+5.59	16.1

Fig. 3 Results from [28] depicting steganalysis error rate for different camera models at the lowest ISO sensitivity

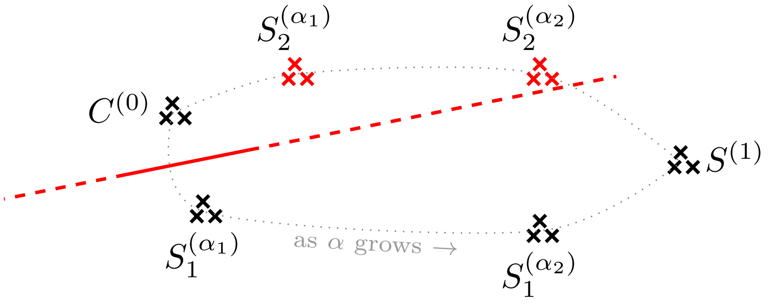


Fig. 4 Stego-source mismatch between schemes S_1 and S_2, and embedding rates $\alpha_1 < \alpha_2 < 1$. Detector for S_1 does not generalize to S_2, because different schemes cause different shifts. Embedding at $\alpha = 1$ leads to the same stego

Fig. 2 shows a scatter plot of co-occurrence (i.e., joint empirical distribution) between two randomly selected DCT coefficients resulting from JPEG cover-sources. The images come from the same sources with the exception that they are converted from RAW files to uncompressed TIFF images using different software. More precisely, we note that the simplest development was used (essentially made of demosaicing, white balance, and gamma-correction). It can be clearly seen from Fig. 2 that the correlation between the selected DCT coefficients is almost in complete opposition. In addition, one can note that the variance of marginal distribution also changes significantly, especially for *RawTherapee*, see green dots on Fig. 2.

Last, Fig. 3 exemplifies how the CSM can have a dramatic impact on detector performance. Depending on the detector and the sources, the CSM problem can be either barely noticeable or make a detector no better than a random guesser.

2.2 Mismatch in steganography

The cover-source generation process is not the only source of mismatch in steganalysis, as detectors must also face heterogeneity in the steganography in training and testing sets [82]. As the steganographer's choice of cover-sources and steganographic embedding are independent, they are separate types of mismatches influencing steganalysis.

In addition, as we shall explain in Sect. 3.2, measuring the CSM problem via a feature space representation inherently separates the two. However, both mismatches exhibit similar symptoms, and extending existing mitigation strategies against CSM can be applied to both.

The literature uses either the names *stego-source mismatch* [22, 59] or *stego-algorithm mismatch* [90]. We suggest a new name, *stego-scheme mismatch* (SSM), as it is more general and hence encompasses the few possible causes of the mismatch of this kind.

Definition 4 The **stego-scheme mismatch (SSM)** is a mismatch of the distributions of stego-objects stemming from a given cover-source. It is caused either by using different stego-schemes or different settings to embed data in cover-objects.

SSM has two known causes: steganographic embedding [12, 59, 82, 86] and embedding rate α [80], illustrated in Fig. 4 by orientation and length of the stego-shift. Embedding into a cover-source $C^{(0)}$ with two schemes S_1 and S_2 and payload $\alpha < \alpha_{\max}$ causes different stego-shifts. The dotted trajectories of both schemes suggest that for $\alpha = \alpha_{\max}$ the embedding is equivalent to LSB matching, denoted $S^{(1)}$.[4]

Note that the shift hypothesis [19, 46] states that, roughly speaking, the trajectory is the same regardless of the cover-sources. The same embedding causes the same impact (the same "shift") on the stego-distribution.

SSM is more important for test sets with (1) lower α or (2) less detectable embedding scheme [15].

A *universal* detector is capable of generalizing to any steganographic method [73].

2.3 Mismatch in pooled steganalysis

Pooled steganalysis is the relevant technique when one tries to inspect a group of several objects over which one sole decision must be made (typically "all these objects are covers" vs. "some objects are stegos"). Pooled steganalysis either indirectly addresses CSM or is affected by it.

On the one hand, several works addressed the problem of pooled steganalysis when it aimed at identifying guilty actors, in which case this framework can lift the problem of CSM provided that "*each actor used a source of cover-objects, different from sources used by other actors*" [48]. However, this means that the steganalyst needs to know the source of each actor which is hardly possible in real life, "*unless the suspected steganographer is considerate enough to supply [...] the cover-source*" [53].

Under the same assumption, a similar approach was adopted in [61] by detecting inconsistencies in the classifier in order to identify the suspect(s).

Another look at the mismatch was studied from pooled steganalysis when the strategy of the steganographer is not known: that is how the payload is spread over several objects. In this case, the steganalyst typically faces a stego-scheme mismatch as the embedding rate for each object is unknown.

A robust statistical sequential method based on CUSMU (Cumulative Sum) was proposed in [16] while the method developed in [85] is based on the histogram of the classifier "soft-output" (before thresholding). The problem was also leveraged in [19, 45]: in an adversarial setting, it has been proposed to spread the payload in order to create the hardest stego-scheme mismatch hence reducing detectability.

Comparison of the mismatches CSM relates to the mismatch in the processing steps prior to the embedding. SSM relates to the mismatch in the embedding step, the steganographic key, or the message. Pooled steganalysis is a technique particularly sensitive to both CSM and SSM, and contains a new factor of SSM, payload spreading.

3 Measuring the CSM problem

In Definitions 2 and 3, we have taken particular care to separate the CSM from the problem it generates through its impact on steganalysis. Thus, measuring severity is critical for the very definition of the CSM problem but also in order to study its causes (RQ 2) and for successful mitigation of CSM (RQ 3). In this section, we describe the two main approaches that have been used for the assessment of the CSM problem. First, and most obvious, in Sect. 3.1, the mismatch problem is measured via the performance of a detector. Second, Sect. 3.2 describes the studies measuring the CSM problem using a distance in feature spaces.

3.1 Measuring a mismatch via detectors

One way to measure the CSM is through comparison of the detector error ϵ in matched and mismatched scenarios. Within the framework of modern steganalysis, based on machine learning, the clairvoyant scenario is the one in which training datasets perfectly match the testing datasets over which the detection performance is measured[5]. This case, without CSM, represents the "ideal" baseline error of steganalysis and is referred to as the *intrinsic difficulty* of a given source (see Definition 5).

With these notations, we can now formally define the notions that have been adopted in the literature for measuring CSM using detector error rate:

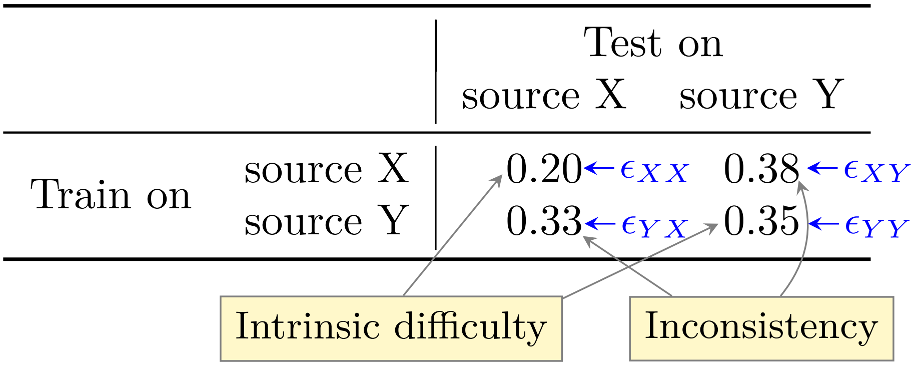
Definition 5 The **intrinsic difficulty** ϵ_{XX} or ϵ_X of a source X is the error rate of a detector that is trained and tested on X. It expresses how difficult it is for a detector to detect a steganographic scheme of embedding rate α in covers from a cover-source X.

The *intrinsic difficulty* represents the error rate of steganalysis on a given source without any mismatch. Therefore, this criterion is not related to the CSM problem, but it must be taken into account, serving as a baseline, to quantify the CSM problem.

[4] Note that in practice the maximal achievable payload $\alpha_{\max}$ depends on the coding method. With LSB matching, ternary embedding allows embedding up to $log_2(3) \approx 1.585$ bit per element.

[5] Note that some steganalysis methods do not use machine learning. However, these statistical detection methods [21, 97] generally include some parameters (weights, detection threshold, etc.) whose settings require some cover and stego-examples and hence may be subject to the CSM problem [14, 54].

Table 1 An illustrative example of presenting mismatched scenario results measured via a detector error rate. The diagonal contains intrinsic difficulties, and off-diagonal values are the inconsistencies. Inconsistencies can be replaced by the difference with the corresponding diagonal element, column-wise (regret) or row-wise (cover-source generalization error)



		Test on	
		source X	source Y
Train on	source X	0.20	0.38
	source Y	0.33	0.35

ϵ_{XX}
ϵ_{XY}
ϵ_{YX}
ϵ_{YY}
Intrinsic difficulty
Inconsistency

Note that the intrinsic difficulty can vary greatly, even for similar sources: Fig. 3 shows that the camera model can impact the error rate by almost 10%.

The mismatched scenario error is the one when the training is performed over a dataset that presents some statistical differences from the testing dataset. This corresponds to Definition 2 of a CSM scenario. The ensuing degradation of detector performance is referred to as the *inconsistency* between sources, see Definition 6.

Definition 6 The **source inconsistency** ϵ_{XY} is the error-rate of a detector, trained on source X and evaluated on source Y.

To illustrate these concepts as clearly as possible, Table 1 shows a toy example with two sources. Note that, just like in Fig. 3, rows represent the dataset used for training the detector while columns are for the testing dataset, on which detection performance is measured. In Table 1, the detection error rate on the diagonal measures the intrinsic difficulty as in those cases training and testing datasets match. On the opposite, the out-of-diagonal results report the detection error rate in case of mismatches hence the degradation error rate due to the mismatches. In the rest of the present section, we will use the following notations: ϵ_{XX} represents the detection error rate when training on the dataset X and testing on the same dataset X. In this case without a mismatch, the intrinsic difficulty is measured and is sometimes denoted ϵ_X for short. On the opposite, ϵ_{YX} represents the case when training the detector on the source Y while testing on the source X (since training is carried out first, it is the first variable in the notation we adopt). In this case of CSM, the detection performance reveals the CSM problem.

The source inconsistency alone carries little information and should be accompanied by the corresponding intrinsic difficulty. This reference can be made more explicit by reporting the regret [3, 4, 94]:

Definition 7 The **regret** $R_{XY,Y}$ is the difference between the inconsistency ϵ_{XY} and the intrinsic difficulty ϵ_Y, as shown in Eq. 1,

$$R_{XY,Y} = \epsilon_{XY} - \epsilon_Y. \tag{1}$$

To exemplify this concept, one can have a look back at Table 1. The inconsistency is higher when testing on the source Y, $\epsilon_{X,Y} = 0.38$, than when testing on the source X, $\epsilon_{Y,X} = 0.33$. However, the intrinsic difficulty is also much higher on the source Y, $\epsilon_Y = 0.35$, than on the source X, $\epsilon_X = 0.2$.

Therefore, the "*empirical measurement of the degradation due to the CSM*," which is defined as the regret by Definition 7, is actually much lower when testing on the source Y, $R_{XY,Y} = 0.03$, as contrast to when testing on the source X, $R_{YX,X} = 0.13$.

Note that these numbers represent a general observation that the regret is asymmetric, see for instance Fig. 3 and prior works [9, 27, 28, 56].

Regret informs us about the severity of the CSM problem since it assesses, given 2 datasets, how much the detection performance can be impacted by training a detector on one dataset and testing on the other.

Interestingly, when the sources differ in a single processing step, the regret allows evaluation of the impact of this step on the CSM problem.

The regret reports about the test source and is computed using two detectors. Cover-source generalization error is an alternative metric, computed with one detector trained on one single source, tested over two sources.

Note, that machine learning uses the term "generalization error" to express the performance difference between train and test set. In steganalysis, this term is used to describe the perfromance drop between test set with matched and mismatched cover source. We use the name "cover-source generalization error" to better distinguish between them.

Definition 8 The **cover-source generalization error** $R_{XY,X}$ is the difference between the inconsistency ϵ_{XY} and the intrinsic difficulty ϵ_X, as shown in Eq. 2,

$$R_{XY,X} = \epsilon_{XY} - \epsilon_X. \tag{2}$$

The cover-source generalization error tells us how much a given detector is sensitive to a mismatch in the test source. On the opposite, the regret reports how much a given test source is sensitive to the mismatch in the training source. But their nature and their uses are different.

To compare mismatch between sources of different intrinsic difficulties, we can normalize the regret by the intrinsic difficulty. We define this in Definition 9 as *relative regret*.

Definition 9 The **relative regret** is the regret normalized by the intrinsic difficulty, shown in Eq. 3:

$$\frac{\epsilon_{XY} - \epsilon_Y}{\epsilon_Y}. \tag{3}$$

Our definitions, based on [27, 94], aim to solve the discrepancy of the terms "*inconsistency*" and "*regret*" [3, 4, 27, 28].

Metrics Reporting the regret depends on the chosen performance metric. In machine learning, the detection error is usually measured with the misclassification. In steganalysis, the popular metrics are the total probability of error P_E (assuming equal prior) and the miss-detection rate for a prescribed false alarm. Using convertible metrics throughout the studies would facilitate the cross-study comparison of results.

Limitations Measuring the CSM problem via detector error focuses on the symptoms of CSM, but heavily depends on a type of detector used. This makes regret-based mitigation difficult to generalize to different detectors. Indeed, as shown in [2], using different detectors produces different regrets.

3.2 Measuring a mismatch via proximity in a feature space

The second option to quantify the CSM is to project the cover-sources onto a lower-dimensional feature space and measure their proximity or difference there. Closer cover-sources are expected to produce smaller degradation of steganalyser performance.

Definition 10 A **feature space** is a collection of n variables, extracted using the same function $\phi(\cdot)$, from different objects and carefully designed for a specific goal of analysis. Indeed, features aim at preserving the statistical properties of objects while normalizing their representation, reducing the dimensionality and hence helping the desired analysis.

The feature spaces can be defined manually, for instance, handcrafted steganalysis features DCTR [36] or GFR [96], or constructed ad-hoc by a trainable component.

Types of metrics The difference of cover-sources can be quantified by aggregating each cover-source, e.g., by its center of gravity (mean), and using common metrics, such as the Euclidean norm ℓ^2 or Mahalanobis distance [3, 56].

Another approach is to measure the proximity of hypothetical distributions of cover-sources using probabilistic measures, such as the Kullback-Leibler divergence (KLD) [75], used in Cachin's security [13], the maximum-mean discrepancy (MMD), popular for domain adaptation (Sect. 6.4), or the Wasserstein distance, used in generative steganography [95]. Despite popularity in theoretical works, the KLD presents a challenge, as it is asymmetric, and the cover-sources must have the same support.

New approaches to quantify the proximity between cover-sources, based on the chordal distance [3] and the KLD [75], were recently proposed.

Limitations In order to successfully mitigate CSM, the measure in the feature space should be related to the steganalysis regret of a chosen detector [3, 75]. This is clearly not the case for metrics such as MMD or ℓ^2, which are symmetric, while the regret is asymmetric, as shown in Fig. 3. Recent works addressed the metric selection by empirically measuring the correlation of the feature-space metrics with the regrets. This approach, however, is still in its infancy and faces fundamental challenges.

4 Bibliometry

This section introduces the methods used to answer the research questions from Sect. 1. Section 4.1 presents the collection of literature (RQ 1). Section 4.2 describes measuring the research trend (RQ 1). Section 4.3 depicts how the literature results were aggregated (RQ 2, RQ 3).

From now on, the survey focuses on the specific case of image steganalysis. Indeed, the state of the art in other types of covers, such as video, audio, and text files, is much less developed. In particular, the CSM problem, even though it is studied by some papers, for instance in video steganalysis [58] or audio steganalysis [67], is not enough covered.

4.1 Collection of literature

The collection strategy consists of (1) initial sampling, (2) identification of relevant papers, and (3) breadth-first search using forward and backward references.

Sampling The initial samples were acquired with a query "cover-source mismatch steganalysis" from two metasearch engines, namely Google Scholar (GS) and DBLP, and independently from the editor or publisher. We take the first 50 results, sorted by relevance with no additional filters.

Identification Each paper was proofread to identify whether it is relevant. Relevant papers fit into at least one of the following criteria:

- Paper discusses the effects of CSM.
- Paper reports results with CSM.
- Paper investigates the impact of factors on CSM.
- Paper attempts to mitigate CSM.

Only a few papers mentioning the CSM were excluded, for instance, when CSM is a future work.

Reference search For each relevant paper, we searched its references (*forward search*) as well as papers citing it using the GS feature "Cited by" (*backward search*). On the newly found papers, we applied the same reference search, proceeding in the breadth-first order until exhaustion.

In the end, we collected and annotated 102 papers. The complete annotated bibliography is presented in Appendix A.

4.2 Measuring the research trend

Each paper in the bibliography is labeled with tags, denoting assumed or explored topics. The tags allow for measuring the trends of CSM research, separately for causes and mitigation. The number of citations is acquired from GS, as of 5 October 2023. For each paper, we compute the *topic coverage* (Eq. 4),

$$\frac{1}{\#\text{tags}}. \tag{4}$$

4.3 Aggregation of results

Although a comparison of the results in the literature is hardly possible due to different experimental setups, a relative measure may account for it to some extent. We convert the results to the relative regret (Eq. 3), and aggregate by taking the expectation across the cover-sources, according to Eq. 5,

$$\mathbb{E}_{X,Y}\left[\frac{\epsilon_{XY}-\epsilon_Y}{\epsilon_Y}\right]. \tag{5}$$

The relative regret marginalizes differences in experimental setups and can be extracted from the existing literature, when inconsistency or regret is reported together with the source's intrinsic difficulty.

However, the statistics should match in the choice of error metrics; problematic is also the dynamic threshold used in P_E. Furthermore, the relative regret assumes that the CSM affects performance proportionally.

5 Causes of CSM

In this section, we answer RQ 2 for image covers.

5.1 Image cover-source

A cover is defined by the whole imaging pipeline, which consists of the acquisition (from the scene up to the electrical signal) and processing (from the electrical signal to the image file) [35, 88]. The diversity of the cover-sources stems from the variety of acquisition devices and parameters, processing software and their specific operations and parameters. But it is also enriched by the initial captured content and the later compressions applied by specific software, such as social networks.

This diversity makes it hard to give a proper all-encompassing model of the cover-source. However, it is possible to measure the impact of each operation. To get their impact, we gather the intrinsic difficulties and the source inconsistencies from [7, 9, 28, 66, 100]. We use them to compute the relative regrets using Eq. 3. We then compute the expected relative regrets using Eq. 5 gathering every cause into 6 categories: content, device, color, filter, resize, and JPEG.

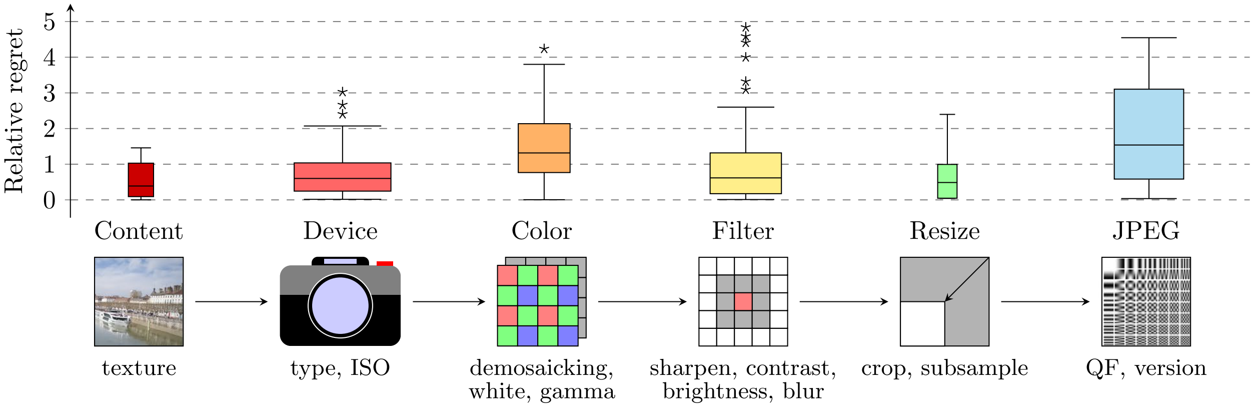


Fig. 5 Generic schema of image processing pipeline (IPP), together with the impact of each step on CSM. The most impactful is JPEG compression. Later processing has generally more impact

The distribution of the relative regrets for each step is reported in Fig. 5. As expected, the JPEG compression step has the biggest impact on the CSM problem. The impact of processing steps tends to be higher the later they are in the pipeline.

Section 5.2 summarizes the state of the art using this 6-step categorization of the cover-source steps.

5.2 Coverage of causes

Each step discussed below is an arbitrary division of all the gathered causes of CSM. Readers should be aware, for instance, that filtering operations can be performed at different steps of the pipeline. Equally, color operations can be performed at a later stage. Finally, while we generally assume that the JPEG compression comes last, it does not guarantee that later processing will not induce CSM.

Despite these nuances, we chose, for the sake of clarity in answering RQs 1–2, to present a summary in the general order in which each cause appears in the overall pipeline. We strongly recommend reading the studies to get a full explanation of the reported findings.

Content While not properly a part of the processing pipeline, the content has long been recognized as a cause of heterogeneity in steganalysis [55]. The level of texture, also called *texture complexity* (TC), allows quantification of image similarity. Such a metric was proposed in [40], where datasets with high TC are harder to train on [28, 105], but generalize better on testing sets with low TC than the other way around [40, 41].

The amount of textures depends on the captured scene, but later processings, such as filtering or resampling, greatly impact the final textures as well.

Device The device model is a common source of diversity in the literature [6, 17]. Research also covers the impact of individual devices [66], and acquisition parameters: ISO sensitivity (moderate impact), aperture, and exposure time (low impact) [28].

Color Color processing involves demosaicking, i.e., removing the Bayer grid from the raw image by interpolating colors, and color balancing, rendering white and gray shades. Optional steps are linear color correction and non-linear gamma correction. Diversity in demosaicking algorithm is common in the literature [17], even though its impact is lower than filtering or resampling [4, 28]. On the other hand, relative excess regrets extracted from [9] are high. This study reports noticeably low ϵ_{XY}, which explains the high relative impact of the color step in Fig. 5. For all these processings, in particular, Fig. 5 would benefit from having more results.

Filter Filtering encompasses neighborhood-based processings, such as denoising [77], unsharpening [87], blurring, sharpening, or edge enhancement. These operations are shown to have a strong impact on the CSM [4, 10, 26, 28]. Due to the large number of filtering parameters, measuring the CSM quickly becomes a computationally overwhelming task.

Resize Resizing can be carried out with two very distinct operations: cropping and resampling. Cropping consists merely of removing pixels rows or columns without any additional modification. While this may induce a shift in the grid, such as the Bayer or JPEG ones, it completely preserves statistics in the pixel domain. On the opposite, resampling requires a low-pass digital filtering operation (to interpolate missing values) hence reducing the texture complexity and is a strong CSM factor.

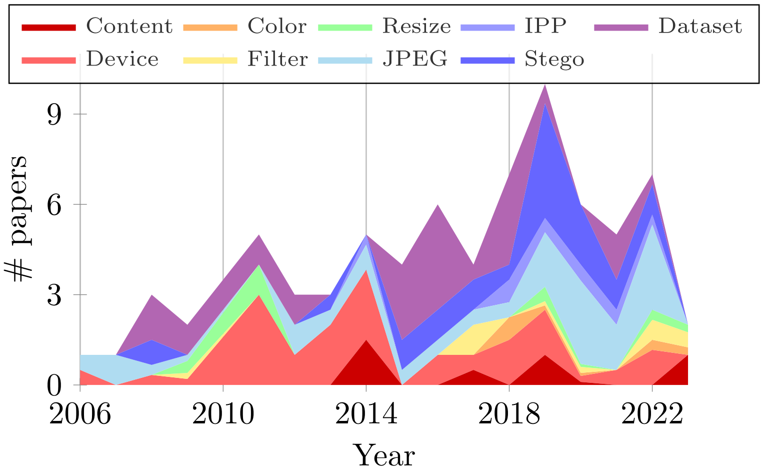


Fig. 6 Area chart capturing timeline of CSM causes. The number of papers on CSM and the diversity in causes is increasing, especially in the last 10 years

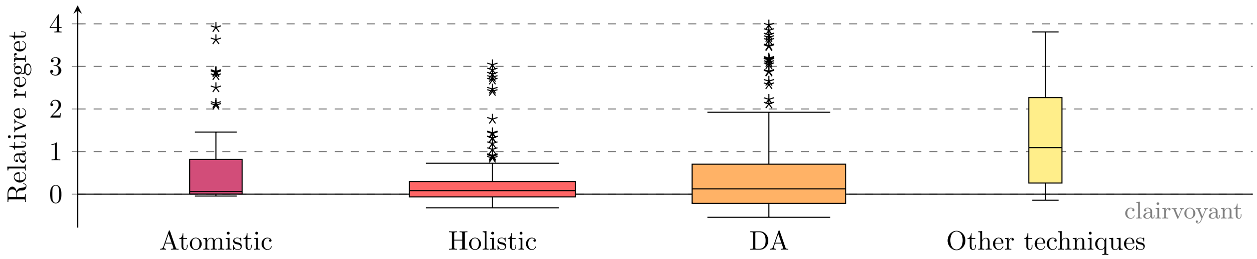


Fig. 7 Effect of existing mitigation strategies on CSM, compared to clairvoyant scenario. Holistic mitigation performs the best. Holistic and DA reach below clairvoyant because some results report improvement in presence of CSM

Downsampling can reduce the CSM problem between mismatching cover-sources [107]. Both cropping and resampling were used in ALASKA [17].

JPEG JPEG compression is indisputably the most impactful cause of CSM. The constant attention from the community [18, 30, 56, 57] focused mostly on the quality factor (QF) controlling the distortion-rate tradeoff. CSM is also induced by double compression [92] or JPEG implementation [7].

Figure 6 answers RQ 2. It shows the evolution of the number of papers, from 2006 onward, studying the causes of CSM. To the 6 categories that we consider in the paper, we added 3 others, often found in studies. First, “IPP” designates studies with unspecified cover-source processes, but that still use them as a cause of CSM. Then, “Stego” refers to papers dealing with SSM, and “Dataset” to CSM between datasets. Additionally, when a paper dealt with more than one cause, we chose to split its weight evenly.

Note that the number of papers used in Fig. 6 is less than the total number used in the survey. This is because not all papers accurately discuss the causes of CSM.

Heuristics for the CSM problem To conclude this section, let us mention some complementary papers that use heuristics to approximate the impact on CSM. A heuristic close enough to the detector error can facilitate mitigation of the CSM caused by the parameter whose impact is being approximated. For instance, [40] designs a similarity measure for texture complexity in images. [104] suggests a metric for JPEG quantization tables (QT).[6]

[6] QT metric weights high frequencies unintuitively. E.g., standard QF75 is closer to QF76 than to QF75 with incremented DC, although 90% values differ.

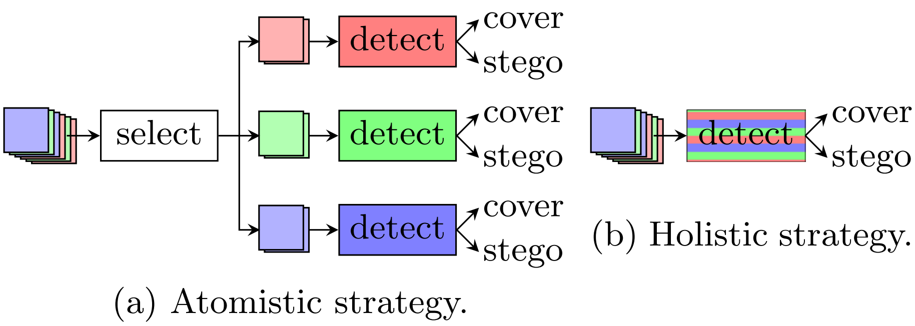


Fig. 8 Strategies to mitigate cover-source mismatch. In the atomistic strategy, the selector chooses the best detector for each input to minimize CSM. In the holistic strategy, the detector is trained on a heterogeneous set of cover-sources

6 Mitigation of CSM

This section summarizes the approaches to mitigate the effect of the CSM and answers RQ 3. We start by the idealistic case—clairvoyant scenario (Sect. 6.1), followed by major mitigation schools: atomistic (Sect. 6.2), holistic (Sect. 6.3), and domain adaptation (Sect. 6.4), followed by other miscellaneous techniques (Sect. 6.5). In Sect. 6.6, we analyze the current trends in the research.

The distribution of the relative regrets for different mitigation techniques is reported in Fig. 7. The y-axis contains the relative regret from Eq. 3, aggregated over results reported in [26, 39, 49, 56, 63, 69, 84, 94]. Clairvoyant is shown as a solid horizontal line at relative regret 0. Holistic steganalysis appears to be the most performant, followed by atomistic steganalysis and domain adaptation. Other techniques perform worse and have the most varying relative regret. Holistic and DA mitigation reach below the clairvoyant, because some studies report better mismatched values than matched values, i.e., improvement in the presence of CSM.

6.1 Clairvoyant scenario

Perfect mitigation of the CSM is possible, if the steganalyst gets to know the cover-source, and trains the detector on it. Such a simple strategy, referred to as the clairvoyant scenario [80], comes from a conservative interpretation of Kerckhoffs' principle, according to which what is not secret is assumed to be public.

Limitations However, knowing the cover-source is applicable only sometimes in practice. It assumes the steganographer publishes the cover-source, either deliberately, unintentionally, or forced by an active steganalyst. Otherwise, CSM is present, and the steganalyst must seek different ways to mitigate it. The current state of the art has three major schools: atomistic, holistic, and domain adaptation.

6.2 Atomistic steganalysis

A natural extension of the clairvoyant scenario to unknown cover-sources is to train multiple detectors on different cover-sources and choose the correct detector for the input image. Such an *atomistic* detector, shown in Fig. 8a, involves two steps:

1. The forensic step (*select*), which identifies the cover-source from the input image
2. The steganalysis step (*detect*), a pool of detectors trained on different cover-sources

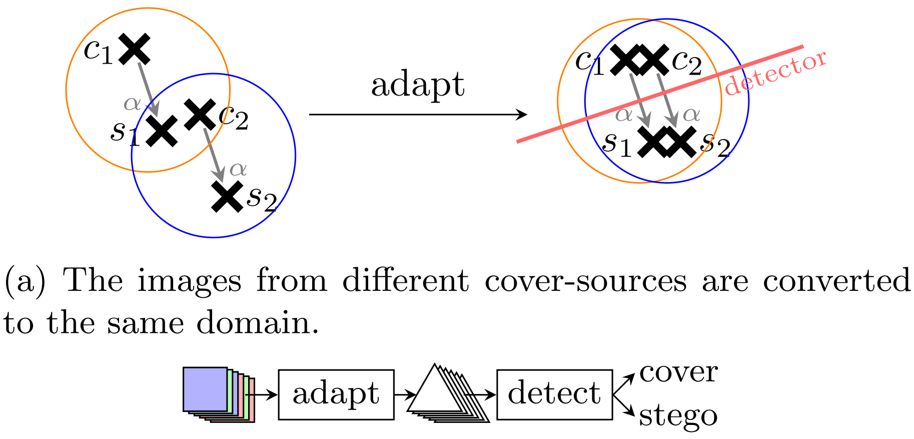
The training consists of selecting the cover-sources and training the selector and one detector per cover-source. During the testing, the input cover-source is determined, and the input is fed into the associated detector. If the detector for the input cover-source is available, the atomistic detector performs as well as in the clairvoyant scenario [37, 106].

Selector A major question is how to construct the newly introduced selector component. Cover-source identification typically uses a feature space and an unsupervised [38, 78, 106] or non-parametric [31] construction. The cover-source can also be partially reconstructed using the means of forensic analysis [5, 37].

Figure 7 displays that the majority of the results reported for the atomistic approach are up to relative regret 1.

Limitations The first limitation is that the number of possible cover-sources is intractable. The selector must be able to deal with the situation when the cover-source is not present in the pool.

The second problem is that the success of the atomistic detector relies on the selector. Errors of the selector add up with the errors of the detectors [94].



(a) The images from different cover-sources are converted to the same domain.

(b) The detector operates in the intermediate domain.

Fig. 9 Domain adaptation strategy

The third issue is the selection of the cover-sources to train on. The aim of the steganalyst is to ensure a good coverage of the source domain, but the more detectors one wants, the longer the training time becomes [3].

6.3 Holistic steganalysis

A detector performs the best on the cover-source seen during the training. When trained on multiple cover-sources, the performance is usually slightly worse than that of the dedicated per-source detectors. By contrast, the performance degrades far more on unseen cover-sources.The amount of degradation depends on the CSM.

A *holistic* approach gives up on matching the clairvoyant performance and training a large number of dedicated detectors. Instead, one detector is trained on a heterogenous dataset with a large number of cover-sources, as shown in Fig. 8b. The idea is that a sufficient coverage over sources ensures a good performance of the detector on unseen sources and leads to the mitigation of CSM.

Heterogeneity The major challenge in the holistic approach is the construction of the training database. The number of sources certainly relates to the robustness of the dataset. Research suggests that training on fewer, carefully chosen cover-sources yields better results, than on a huge number of blindly collected cover-sources [4, 98].

Model selection The architecture of a holistic detector is critical for model performance. Not only is the detector expected to detect steganography, but it also needs to do so in a heterogeneous environment [25]. Increasing the flexibility of a model can lead to greater overfitting, which can be mitigated by increasing the dataset size, or by regularization techniques [74, 76].

Limitations The first limitation is that the holistic training requires a very large dataset [70, 78]. The second problem is that the holistic approach performs worse than the atomistic approach on a fixed set of cover-sources [37, 106]. The third issue is that the performance is very sensitive to the cover-sources covered in the training set. The success of the mitigation also strongly depends on the selection of the detector architecture.

6.4 Domain adaptation

The mitigation of atomistic and holistic approaches is limited by the number of sources used during training. The idea of domain adaptation (DA) is to train on a single cover-source, called *source domain*, and use the learned knowledge to adapt to an unseen cover-source called *target domain*. This is done by aligning the domains, illustrated in Fig. 9a by colored circles, so that the detector can better generalize to a diverse domain.

The general DA workflow illustrated in Fig. 9b is (1) converting the test sample to the adapted feature space and (2) passing them to a detector, trained in this feature space.

Adaptation The feature space should align different cover-sources closely, yet maximize the shift caused by steganography. A simple adaptation method is matching the moments of the source and the target domain, e.g., using standardization. The advanced solutions presented in the literature are unsupervised techniques, such as

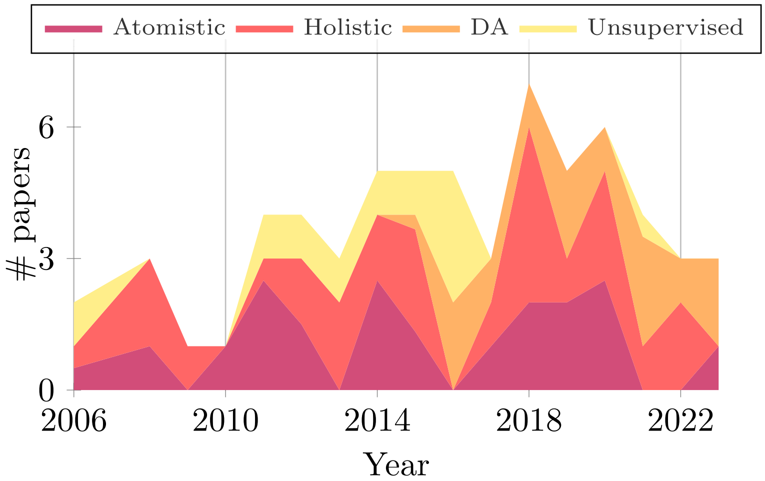


Fig. 10 Area chart capturing the timeline of the CSM mitigation strategies. The number of papers grows, with a peak in 2018–2020. Currently, the most popular mitigation strategy is DA

clustering or manifold alignment [1, 23, 43, 57, 63, 65], possibly aided by guiding features and pseudo-label prediction [108, 110]. Often, the construction is challenging, because the labels for the target domain are not available.

Limitations Existing methods only extract specific statistics, such as means or higher moments, which may be insufficient to mitigate CSM over different cover-sources. Moreover, using handcrafted steganalysis features for domain adaptation may lead to unsuccessful mitigation, because although sensitive to steganography, they are also affected by CSM [101].

6.5 Other techniques

Apart from the three major approaches, we identify other, marginally explored, techniques.

Re-embedding Re-embedding into the test images increases the spread of the stego-shift directly in the target domain. Training multiple detectors, similar to diffusion models, i.e., on cover-stego, stego-double stego, etc., may further improve the detection performance [60, 62, 64, 105].

Feature projection The second limitation of DA mentioned in Sect. 6.4 may be tackled by modifying the features, so that the projected features are insensitive to cover-source heterogeneity [84, 99]. Feature engineering w.r.t. CSM is not trivial, for instance, fusing feature sets, which helps in the matched scenario, may degrade performance in the presence of CSM [25].

One-class detector Figures 1, 4 and 9a depict linear detectors, but there is a variety of detectors, which in some cases may perform better or be more robust to CSM, such as non-linear classifiers or one-class classifiers [81].

Unsupervised learning A steganographic embedding may be treated as an outlier and detected using unsupervised learning. This is particularly common in pooled steganalysis, where the steganalyst looks for a guilty steganographer among a pool of communicating actors. The techniques used are outlier detection, clustering based on MMD [50, 51], or k-nearest neighbors [39].

6.6 Trends in research on CSM mitigation

In a similar fashion to the CSM causes in Fig. 6, we show the trend of the mitigation strategies in Fig. 10, in terms of the number of papers per year. The number of papers has been growing until peak in 2018–2020, and descending since.

The holistic and atomistic strategies appear consistently and dominate most of the entire time period. The unsupervised approach appears mainly in early 2010s in the pooled steganalysis literature [49, 50]. The domain adaptation first appeared around 2015 and has been growing ever since, which coincides with the boom of deep learning. In the recent years, DA is the dominant mitigation strategy, used both with neural networks [1, 105, 108, 110] and with feature-based steganalysis [3, 4, 110].

7 Countermeasures of steganographer

CSM is usually understood as a challenge for the steganalyst, which he overcomes by choosing a detector that minimizes the CSM and increases the chance of catching the steganographer. The optimal choice can be made if the steganalyst knows the exact cover-source.

In the real world, however, the knowledge of both actors is limited: the steganographer does not know the steganalyst's detector, and the steganalyst does not know the cover-source. Thus, the steganographer and the steganalyst play a game.

The steganographer chooses the cover-sources which presumably maximizes the CSM error to increase the chance of evading detection, because the steganalyst increases the detection threshold to reach the acceptable false positive rate. The game resembles a modified rock-paper-scissors, where the steganalyst wins with the match and the steganographer with the mismatch of the shapes.

A game-theoretical interpretation of the CSM problem was first drafted in [24], and later formalized in [29]. It has been tested in very constrained scenarios so far and requires additional research to push it toward the main stream of research on CSM, which still describes CSM as a problem of the steganalyst only.

8 Discussion

Steganalysis competitions BOSS and ALASKA had a profound impact on the field and stimulated a lot of new energy. Their experimental setups were followed by the research long after they ended. Future competitions should be carefully designed to facilitate comparisons, such as the one carried out in this survey; of importance to the CSM are the factors of diversity, steganographic schemes, and performance metrics.

> **Open question 1:** Have all the causes of CSM and SSM been clearly identified? Are their effects well measured?

Pushing forward in this direction, the steps of the IPP are usually studied in isolation, but their order might impact the CSM. Interactions between the steps exist, e.g., between resampling and sharpening [4] or between SSM and the cover source [90].

> **Open question 2:** How can we measure the crossed effect of different causes on the CSM?

Answering this question, among others, challenges the community to build methods that will need to deal with the computational complexity of the task.

We detailed the different approaches to measure CSM, highlighting the limitations in each one, but the question is still mostly an open one. As the current tools are symmetric, a "theoretical" measure of CSM can only poorly relate to the practical regret. Designing better tools, in order to avoid the bias of training a detector, is probably one of the most promising steps at characterizing the CSM in the short term.

> **Open question 3:** How is CSM related with statistical properties of the cover sources?

Finally, as already mentioned, this survey is focused on the case of steganalysis natural images. One question that we can naturally draw from our present research would be to see if the same observations can be made for other types of digital media.

> **Open question 4:** How does CSM impact steganalysis of other types of covers (audio, video, ...)?

Existing methods to measure the CSM impact depend on the detector, or on the feature space. Suitability of these assumptions, as well as possible alternatives, is to be investigated. Meta-research, such as result aggregation in Fig. 5, would benefit if future studies provided intrinsic difficulties and inconsistencies.

Existing mitigation strategies may aid at solving the problem, yet fail in pessimistic scenarios, such as unknown processing history. The atomistic approach is suitable for a closed set of cover sources, but fails on open-set problems where holistic performs better. The increased popularity of domain adaptation correlates with the introduction of deep learning in steganalysis.

We strive to sample and aggregate the existing literature on CSM objectively. However, we are aware of potential biases: (1) sampling bias due to search engine ranking, isolation of papers from the rest of literature, or incorrect assessment of relevance, (2) bias due to incorrect annotation of the paper, and (3) bias of impact estimates, when the paper results cannot be used for aggregation, e.g., when reported via graph.

Adversarial scenario for CSM CSM is usually understood as a problem for steganalyst. It can also be interpreted as a game between the steganographer and the steganalyst, a modified rock-paper-scissors, where the steganalyst wins on match and steganographer on the mismatch of the shapes [24, 29]. This approach has been tested in a very constrained scenario so far.

9 Related work on data heterogeneity

Essentially, CSM is a form of *data heterogeneity*, which is a common problem in many fields of application of ML. Studying the causes of data heterogeneity and the

Table 2 Summary of the answers to the proposed RQs in other fields of application of ML: medical imaging (MI), mechanical component monitoring (MCM), speech recognition (SR), natural language processing (NLP), and temporal reasoning (TR)

ML field	Surveys	Causes of mismatch	Mitigation strategies
MI	[33, 113]	Difference of acquisition device, parameters, and patients between datasets	Domain adaptation to cope with the low number of available labels. Transfer learning from MI-specific models
MCM	[102]	Different working conditions. Different monitored components. Transfer from simulations to the real world	Transfer learning; domain adaptation, e.g., shallow and deep feature matching, GANs, ...
SR	[109]	Noisy environments	Preprocessing, features extraction. Using acoustic/language models
NLP	[89, 93]	Cross-lingual learning. Writers, contexts, and dates also cause mismatches	Data-centric (pre-training, pseudo-labeling) and model-centric (adversarial networks, autoencoders) DA
TR	[72, 112]	Difference in the datasets' origins, e.g., time granularity, or difference between temporal textual expressions	Temporal data management, e.g., data integration, or NLP-based methods

mitigation strategies in the literature of other fields might prove beneficial to understanding the causes of CSM and mitigating its impact on steganalysis. We selected surveys focusing on data heterogeneity from 5 fields: medical imaging (MI) [33, 113], mechanical components monitoring (MCM) [102], speech recognition (SR) [109], natural language processing (NLP) [89, 93], and temporal reasoning (TR) [72, 112], and we look to what extent each survey covers RQ 1 and 3, in their respective field. The summary of our readings is shown in Table 2.

Causes of heterogeneities Regarding RQ 1, although each field has identified its causes of heterogeneity, some causes appear conceptually similar from one field to another. Patients in MI, components in MCM, and authors in NLP can be understood as similar causes of mismatch; to image steganalysis, it could relate to the captured content. Similarly, the working conditions in MCM and the noise environment in SR are close to each other but are harder to relate to an identified cause of CSM (see Sect. 5 for details). On the other hand, languages are conceptually rather specific to the field of NLP. Finally, mismatch caused by the acquisition devices and their calibration is expected to occur in every field but seems to be foremost studied in MI and steganalysis, where they both play an important role.

Mitigation strategies Within the broad framework of ML, there are currently two main approaches for dealing with data heterogeneity: *transfer learning* and *domain adaptation*. Transfer learning is based on transferring knowledge from one task to another and has been shown effective in many computer vision applications via *fine-tuning*. Large models dedicated to MI were shown to have good transferability to several of its tasks. Domain adaptation involves learning an invariant space for samples coming from different distributions. It provides tools like shallow and deep ML models for feature matching such as *correlation alignment*, used in MCM, or *transfer component analysis*, in MCM, MI, and steganalysis.

As we have shown, steganalysis faces similar issues (e.g. acquisition device and content) and uses similar mitigation strategies (domain adaptation and transfer learning). The peculiarity of steganalysis is that it operates on signals of interest of low amplitudes, compared to the other fields that mostly operate on a semantic level. Perhaps due to this, the CSM in steganalysis has many diverse causes, which are covered in Sect. 5.

10 Conclusion

The research in the last 20 years has been looking into the cover-source mismatch problem from various directions. Many strategies to suppress CSM exist, but the core of the problem is still present. Successful mitigation must go hand in hand with understanding of the causes. Reliable mitigation of CSM and SSM is essential for operational universal steganalysis.

Abbreviations

AUC	Area under curve
BOSS	Break Our Steganographic System
CORAL	Correlation alignment
CSM	Cover-source mismatch
DA	Domain adaptation
DBLP	Digital bibliography and library project
DCT	Discrete cosine transform
DCTR	DCT residual
GFR	Gabor filter residual
GS	Google Scholar
IPP	Image processing pipeline
KLD	Kullback-Leibler divergence
LSB	Least-significant bit
MCM	Mechanical component monitoring
MI	Medical imaging
ML	Machine learning
MMD	Maximum-mean discrepancy
NLP	Natural language processing

QF	Quality factor
ROC	Reciever-operating characteristic
RQ	Research question
SR	Speech recognition
SSM	Stego-scheme mismatch
TC	Texture complexity
TCA	Transfer component analysis
TR	Temporal reasoning

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s13635-024-00171-6.

Additional file 1. Annotated bibliography will be published upon acceptance.

Acknowledgements

Additionally, authors would like to thank their respective supervisors.

Authors' contributions

The bibliographical research and most of the writing were conducted by Antoine Mallet and Martin Beneš. Rémi Cogranne provided crucial advices and complementary writing.

Funding

This work is funded by the EU's Horizon 2020 programme under the grant agreement No. 101021687(UNCOVER) and the French ANR PACeS project No. ANR-21-CE39-0002.

Availability of data and materials

The list of bibliographical reeferences is given at the end of the paper. Additionally, a summary of the papers, as well as analytical data, will be given upon acceptance of the paper.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 3 January 2024 Accepted: 19 June 2024
Published online: 12 August 2024

References

1. R. Abecidan, V. Itier, J. Boulanger, P. Bas, in *WIFS*, Unsupervised JPEG domain adaptation for practical digital image forensics (IEEE, 2021), pp. 1–6
2. R. Abecidan, V. Itier, J. Boulanger, P. Bas, in *XXIXème Colloque Francophone de Traitement du Signal et des Images-GRETSI'23*, Recherche et Analyse de Sources Représentatives pour la Stéganalyse (Grenoble, 2023). https://hal.science/hal-04166647
3. R. Abecidan, V. Itier, J. Boulanger, P. Bas, T. Pevný, in *WIFS*, Leveraging data geometry to mitigate CSM in steganalysis (IEEE, 2023), pp. 1–5
4. R. Abecidan, V. Itier, J. Boulanger, P. Bas, T. Pevný, in *WIFS*, Using set covering to generate databases for holistic steganalysis (IEEE, 2022), pp. 1–6
5. M. Barni, G. Cancelli, A. Esposito, in *ICASSP*, Forensics-aided steganalysis of heterogeneous images (IEEE, 2010), pp. 1690–1693
6. P. Bas, T. Filler, T. Pevný, in *IH*, "Break Our Steganographic System": the ins and outs of organizing BOSS (Springer, 2011), pp. 59–70
7. M. Beneš, N. Hofer, R. Böhme, in *EUSIPCO*, The effect of the JPEG implementation on the cover-source mismatch error in image steganalysis (IEEE, 2022), pp. 1057–1061
8. J. Blitzer, M. Dredze, F. Pereira, in *ACL*, Biographies, bollywood, boomboxes and blenders: domain adaptation for sentiment classification, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (Association for Computational Linguistics, Prague, 2007), pp. 440–447. https://aclanthology.org/P07-1056
9. D. Borghys, P. Bas, H. Bruyninckx, in *IH&MMSec*, Facing the cover-source mismatch on JPHide using training-set design (ACM, 2018), pp. 17–22
10. M. Boroumand, J. Fridrich, Scalable processing history detector for JPEG images. EI **29**, 128–137 (2017)
11. M. Boroumand, M. Chen, J. Fridrich, Deep residual network for steganalysis of digital images. TIFS **14**(5), 1181–1193 (2018)
12. J. Butora, J. Fridrich, Detection of diversified stego sources with CNNs. EI **31**, 1–11 (2019)
13. C. Cachin, in *IH*, An information-theoretic model for steganography (Springer, 1998), pp. 306–318
14. G. Cancelli, G. Doërr, M. Barni, I. Cox, in *MMSP*, A comparative study of ±1steganalyzers (IEEE, 2008), pp. 791–796
15. M. Chen, V. Sedighi, M. Boroumand, J. Fridrich, in *IH&MMSec*, JPEG-phase-aware convolutional neural network for steganalysis of JPEG images (ACM, 2017), pp. 75–84
16. R. Cogranne, in *WIFS*, A sequential method for online steganalysis (IEEE, 2015), pp. 1–6
17. R. Cogranne, Q. Giboulot, P. Bas, in *IH &MMSec*, The ALASKA steganalysis challenge: a first step towards steganalysis (ACM, 2019), pp. 125–137
18. R. Cogranne, Q. Giboulot, P. Bas, in *WIFS*, ALASKA#2: challenging academic research on steganalysis with realistic images (IEEE, 2020), pp. 1–5
19. R. Cogranne, V. Sedighi, J. Fridrich, in *ICASSP*, Practical strategies for content-adaptive batch steganography and pooled steganalysis (IEEE, 2017), pp. 2122–2126
20. R. Cogranne, V. Sedighi, J. Fridrich, T. Pevný, in *IEEE International Workshop on Information Forensics and Security (WIFS)*, Is ensemble classifier needed for steganalysis in high-dimensional feature spaces? (IEEE, 2015)
21. R. Cogranne, J. Fridrich, Modeling and extending the ensemble classifier for steganalysis of digital images using hypothesis testing theory. TIFS **10**(12), 2627–2642 (2015)
22. T.D. Denemark, M. Boroumand, J. Fridrich, Steganalysis features for content-adaptive JPEG steganography. TIFS **11**(8), 1736–1746 (2016)
23. C. Feng, X. Kong, M. Li, Y. Yang, Y. Guo, in *ICIP*, Contribution-based feature transfer for JPEG mismatched steganalysis (IEEE, 2017), pp. 500–504
24. J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications* (Cambridge University Press, 2009)
25. J. Fridrich, J. Kodovský, V. Holub, M. Goljan, in *IH*, Breaking HUGO–the process discovery (Springer, 2011), pp. 85–101
26. Q. Giboulot, P. Bas, R. Cogranne, D. Borghys, in *EUSIPCO*, The cover source mismatch problem in deep-learning steganalysis (IEEE, 2022), pp. 1032–1036
27. Q. Giboulot, R. Cogranne, P. Bas, in *MWSF*, Steganalysis into the wild: how to define a source? vol. 30 (SPIE, 2018), pp. 1–12
28. Q. Giboulot, R. Cogranne, D. Borghys, P. Bas, Effects and solutions of cover-source mismatch in image steganalysis. SPIC **86**, 115888 (2020)
29. Q. Giboulot, T. Pevný, A. Ker, The non-zero-sum game of steganography in heterogeneous environments. TIFS **18**, 4436–4448 (2023)
30. M. Goljan, J. Fridrich, T. Holotyak, in *SSWMC*, New blind steganalysis and its implications, vol. 6072 (SPIE, 2006), p. 607201
31. F.K. Gomis, M.S. Camara, I. Diop, S.M. Farssi, K. Tall, B. Diouf, in *ISCV*, Multiple linear regression for universal steganalysis of images (IEEE, 2018), pp. 1–4
32. Y. Gong, Speech recognition in noisy environments: a survey. Speech Commun. **16**(3), 261–291 (1995)
33. H. Guan, M. Liu, Domain adaptation for medical image analysis: a survey. TBME **69**(3), 1173–1185 (2021)
34. H. Guan, M. Liu, Domain adaptation for medical image analysis: a survey. TBME **69**(3), 1173--1185 (2022)
35. K. Hirakawa, F. Baqai, *Digital camera processing pipeline*. The Wiley-IS &T Series in Imaging Science and Technology (Wiley, 2023)
36. V. Holub, J. Fridrich, Low-complexity features for JPEG steganalysis using undecimated DCT. TIFS **10**(2), 219–228 (2014)
37. X. Hou, T. Zhang, G. Xiong, B. Wan, in *MINES*, Forensics-aided steganalysis of heterogeneous bitmap images with different compression history (IEEE, 2012), pp. 874–877

38. X. Hou, T. Zhang, G. Xiong, Z. Lu, K. Xie, A novel steganalysis framework of heterogeneous images based on GMM clustering. SPIC **29**(3), 385–399 (2014)
39. X. Hou, T. Zhang, C. Xu, New framework for unsupervised universal steganalysis via SRISP-aided outlier detection. SPIC **47**, 72–85 (2016)
40. D. Hu, Z. Ma, Y. Fan, L. Wang, in *IWDW*, A study of the two-way effects of cover-source mismatch and texture complexity in steganalysis (Springer, 2017), pp. 601–615
41. D. Hu, Z. Ma, Y. Fan, S. Zheng, D. Ye, L. Wang, Study on the interaction between the cover-source mismatch and texture complexity in steganalysis. MTAP **78**, 7643–7666 (2019)
42. I. Hussain, J. Zeng, X. Qin, S. Tan, A survey on deep convolutional neural networks for image steganography and steganalysis. TIIS **14**(3), 1228–1248 (2020)
43. J. Jia, L. Zhai, W. Ren, L. Wang, Y. Ren, L. Zhang, Transferable heterogeneous feature subspace learning for JPEG mismatched steganalysis. Pattern Recognit. **100**, 107105 (2020)
44. S. Katzenbeisser, F. Petitcolas, *Information Hiding* (Artech house, 2016)
45. E. Kaziakhmedov, E. Dworetzky, J. Fridrich, in *WIFS*, Observing bag gain in JPEG batch steganography (IEEE, 2023)
46. A. Ker, in *IH*, Batch steganography and pooled steganalysis (Springer, 2006), pp. 265–281
47. A. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, T. Pevný, in *IH&MMSec*, Moving steganography and steganalysis from the laboratory into the real world (ACM, 2013), pp. 45–58
48. A. Ker, T. Pevný, in *MMSec*, Batch steganography in the real world (ACM, 2012), pp. 1–10
49. A. Ker, T. Pevný, in *MWSF*, A mishmash of methods for mitigating the model mismatch mess, vol. 9028 (SPIE, 2014), pp. 189–203
50. A. Ker, T. Pevný, in *MWSF*, A new paradigm for steganalysis via clustering, vol. 7880 (SPIE, 2011), pp. 312–324
51. A. Ker, T. Pevný, in *MWSF*, Identifying a steganographer in realistic and heterogeneous data sets, vol. 8303 (SPIE, 2012), pp. 182–194
52. A. Ker, Steganalysis of LSB matching in grayscale images. Sig. Process. Lett. **12**(6), 441–444 (2005)
53. A. Ker, T. Pevný, The steganographer is the outlier: realistic large-scale steganalysis. TIFS **9**(9), 1424–1435 (2014)
54. M. Kharrazi, H. Sencar, N. Memon, in *SSWMC*, Benchmarking steganographic and steganalysis techniques, vol. 5681 (SPIE, 2005), pp. 252–263
55. M. Kharrazi, H. Sencar, N. Memon, Performance study of common image steganography and steganalysis techniques. EI **15**(4), 041104 (2006)
56. J. Kodovský, V. Sedighi, J. Fridrich, in *MWSF*, Study of cover source mismatch in steganalysis and ways to mitigate its impact, vol. 9028 (SPIE, 2014), pp. 204–215
57. X. Kong, C. Feng, M. Li, Y. Guo, Iterative multi-order feature alignment for JPEG mismatched steganalysis. Neurocomputing **214**, 458–470 (2016)
58. V. Lachner, K. Schaar, R. Zimmermann, in *ICASSP*, CSM in motion vector steganalysis: the effect of coders on motion vectors in H.264 video encoding (IEEE, 2023), pp. 1–5
59. V. Leask, R. Cogranne, D. Borghys, H. Bruyninckx, in *ARES*, UNCOVER: development of an efficient steganalysis framework for uncovering hidden data in digital media (ACM, 2022), pp. 1–8
60. D. Lerch Hostalot, D. Megías Jiménez, *Diagnóstico de CSM en Esteganálisis* (RECSI, 2018)
61. D. Lerch-Hostalot, D. Megias, in *ARES*, Real-world actor-based image steganalysis via classifier inconsistency detection (ACM, 2023)
62. D. Lerch-Hostalot, D. Megías, in *IH&MMSec*, Detection of classifier inconsistencies in image steganalysis, Proceedings of the ACM Workshop on Information Hiding and Multimedia Security (Association for Computing Machinery, New York, 2019), pp. 222–229. https://doi.org/10.1145/3335203.3335738
63. D. Lerch-Hostalot, D, Manifold alignment approach to cover source mismatch in steganalysis, XIVth Reunión Española de Criptografía y Seguridad (RESCI), RESCI (Mahón, 2016)
64. D. Lerch-Hostalot, D. Megías, Unsupervised steganalysis based on artificial training sets. Eng. Appl. Artif. Intell. **50**, 45–59 (2016)
65. X. Li, X. Kong, B. Wang, Y. Guo, X. You, in *ICIP*, Generalized transfer component analysis for mismatched JPEG steganalysis (IEEE, 2013), pp. 4432–4436
66. L. Lin, J. Newman, S. Reinders, Y. Guan, M. Wu, Domain adaptation in steganalysis for the spatial domain. MWSF **2018**(7), 319–1 (2018)
67. Y. Lin, R. Wang, L. Dong, D. Yan, J. Wang, Tackling the cover-source mismatch problem in audio steganalysis with unsupervised domain adaptation. SPL **28**, 1475–1479 (2020)
68. C. Liu, M. Kirchner, in *IH&MMSec*, CNN-based rescaling factor estimation (ACM, 2019), pp. 119–124
69. I. Lubenko, A. Ker, in *MMSec*, Steganalysis with mismatched covers: do simple classifiers help? (ACM, 2012), pp. 11–18
70. I. Lubenko, A. Ker, in *MWSF*, Going from small to large data in steganalysis, vol. 8303 (SPIE, 2012), pp. 172–181
71. I. Lubenko, A. Ker, in *MWSF*, Steganalysis using logistic regression, vol. 7880 (SPIE, 2011), pp. 193–203
72. Y. Luo, W. Thompson, T. Herr, Z. Zeng, M. Berendsen, S. Jonnalagadda, M. Carson, J. Starren, Natural language processing for EHR-based pharmacovigilance: a structured review. Drug Saf. **40**, 1075–1089 (2017)
73. S. Lyu, H. Farid, in *SSWMC*, Steganalysis using color wavelet statistics and one-class support vector machines, vol. 5306 (SPIE, 2004), pp. 35–45
74. J. Makelberge, A. Ker, in *MWSF*, Exploring multitask learning for steganalysis, vol. 8665 (SPIE, 2013), pp. 218–227
75. A. Mallet, R. Cogranne, P. Bas, Q. Giboulot, in *XXIXème Colloque Francophone de Traitement du Signal et des Images*, Identification de Développements d'Images par Matrices de Corrélations (Université de Grenoble and Association Gretsi, 2023), GRETSI'23
76. W. Ng, Z.M. He, D. Yeung, P. Chan, Steganalysis classifier training via minimizing sensitivity for different imaging sources. Inf. Sci. **281**, 211–224 (2014)
77. T. Nguyen, S. Oraintara, The shiftable complex directional pyramid-part II: implementation and applications. TSP **56**(10), 4661–4672 (2008)
78. J. Pasquet, S. Bringay, M. Chaumont, in *EUSIPCO*, Steganalysis with cover-source mismatch and a small learning database (IEEE, 2014), pp. 2425–2429
79. X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, B. Wang, in *ICCV*, Moment matching for multi-source domain adaptation (IEEE, 2019), pp. 1406–1415
80. T. Pevný, in *MWSF*, Detecting messages of unknown length, vol. 7880 (SPIE, 2011), pp. 300–311
81. T. Pevný, *Kernel Methods in Steganalysis* (SUNY Binghamton, 2008)
82. T. Pevný, J. Fridrich, in *IWDW*, Towards multi-class blind steganalyzer for JPEG images (Springer, 2005), pp. 39–53
83. T. Pevný, J. Fridrich, in *SSWMC*, Merging Markov and DCT features for multi-class JPEG steganalysis, vol. 6505 (SPIE, 2007), pp. 28–40
84. T. Pevný, A. Ker, in *MWSF*, The challenges of rich features in universal steganalysis, vol. 8665 (SPIE, 2013), pp. 203–217
85. T. Pevný, I. Nikolaev, in *WIFS*, Optimizing pooling function for pooled steganalysis (IEEE, 2015), pp. 1–6
86. T. Pevný, J. Fridrich, Multiclass detector of current steganographic methods for JPEG format. TIFS **3**(4), 635–650 (2008)
87. A. Polesel, G. Ramponi, V.J. Mathews, Image enhancement via adaptive unsharp masking. TIP **9**(3), 505–510 (2000)
88. R. Ramanath, W. Snyder, Y. Yoo, M. Drew, Color image processing pipeline. Signal Proc. Mag. **22**(1), 34–43 (2005)
89. A. Ramponi, B. Plank, in *ICCL*, Neural unsupervised domain adaptation in NLP – survey (ACL, 2020), pp. 6838–6855
90. S. Reinders, L. Lin, Y. Guan, M. Wu, J. Newman, Algorithm mismatch in spatial steganalysis. EI **31**, 1–11 (2019)
91. T.S. Reinel, R.P. Raul, I. Gustavo, Deep learning applied to steganalysis of digital images: a systematic review. IEEE Access **7**, 68970–68990 (2019)
92. E. Rodríguez-Lois, D. Vázquez-Padín, F. Pérez-González, P. Comesana-Alfaro, in *EUSIPCO*, A critical look into quantization table generalization capabilities of CNN-based double JPEG compression detection (IEEE, 2022), pp. 1022–1026
93. A. Rogers, M. Gardner, I. Augenstein, QA dataset explosion: a taxonomy of NLP resources for question answering and reading comprehension. ACM Comput. Surv. **55**(10), 1–45 (2023)
94. D. Šepák, L. Adam, T. Pevný, in *EUSIPCO*, Formalizing cover-source mismatch as a robust optimization (IEEE, 2022)

95. H. Shi, J. Dong, W. Wang, Y. Qian, X. Zhang, in *PCM*, SSGAN: secure steganography based on generative adversarial networks (Springer, 2018), pp. 534–544
96. X. Song, F. Liu, C. Yang, X. Luo, Y. Zhang, in *IH &MMSec*, Steganalysis of adaptive JPEG steganography using 2D Gabor filters (Association for Computing Machinery, New York, 2015), pp. 15–23. https://doi.org/10.1145/2756601.2756608
97. T.H. Thai, F. Retraint, R. Cogranne, Statistical detection of data hidden in least significant bits of clipped images. Sig. Process **98**, 263–274 (2014)
98. X. Xu, J. Dong, W. Wang, T. Tan, in *ICIP*, Robust steganalysis based on training set construction and ensemble classifiers weighting (IEEE, 2015), pp. 1498–1502
99. Y. Xue, L. Yang, J. Wen, S. Niu, P. Zhong, A subspace learning-based method for JPEG mismatched steganalysis. MTAP **78**, 8151–8166 (2019)
100. Y. Yang, X. Kong, C. Feng, Double-compressed JPEG images steganalysis with transferring feature. MTAP **77**, 17993–18005 (2018)
101. L. Yang, M. Men, Y. Xue, J. Wen, P. Zhong, Transfer subspace learning based on structure preservation for JPEG image mismatched steganalysis. SPIC **90**, 116052 (2021)
102. S. Yao, Q. Kang, M. Zhou, M. Rawa, A. Abusorrah, A survey of transfer learning for machinery diagnostics and prognostics. Artif. Intell. Rev. **56**(4), 2871–2922 (2023)
103. Y. Yousfi, J. Butora, J. Fridrich, C. Fuji Tsang, in *IH&MMSec*, Improving EfficientNet for JPEG steganalysis, Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security. (Association for Computing Machinery, New York, 2021), pp. 149–157. https://doi.org/10.1145/3437880.3460397
104. Y. Yousfi, J. Fridrich, JPEG steganalysis detectors scalable with respect to compression quality. EI **32**, 1–11 (2020)
105. L. Yu, S. Weng, M. Chen, Y. Wei, RCDD: contrastive domain discrepancy with reliable steganalysis labeling for cover source mismatch. Expert Syst. Appl. **237**, 121543 (2024). https://doi.org/10.1016/j.eswa.2023.121543
106. L. Zeng, X. Kong, M. Li, Y. Guo, in *MWSF*, JPEG quantization table mismatched steganalysis via robust discriminative feature transformation, vol. 9409 (SPIE, 2015), pp. 270–278
107. X. Zhang, X. Kong, P. Wang, B. Wang, in *WDW*, Cover-source mismatch in deep spatial steganalysis (Springer, 2019), pp. 71–83
108. L. Zhang, H. Wang, P. He, S.M. Abdullahi, B. Li, Feature-guided deep subdomain adaptation network for dataset mismatch in spatial steganalysis (2021)
109. Z. Zhang, J. Geiger, J. Pohjalainen, A.E.D. Mousa, W. Jin, B. Schuller, Deep learning for environmentally robust speech recognition: an overview of recent developments. TIST **9**(5), 1–28 (2018)
110. L. Zhang, S. Abdullahi, P. He, H. Wang, Dataset mismatched steganalysis using subdomain adaptation with guiding feature. Telecommun. Syst. Kluwer Academic Publishers, USA, **80**(2), 263–276 (2022). https://doi.org/10.1007/s11235-022-00901-6
111. W. Zhao, J.P. Queralta, T. Westerlund, in *SSCI*, Sim-to-real transfer in deep reinforcement learning for robotics: a survey (IEEE, 2020), pp. 737–744
112. L. Zhou, G. Hripcsak, Temporal reasoning with medical data - a review with emphasis on medical natural language processing. JBI **40**(2), 183–202 (2007)
113. S.K. Zhou, H. Greenspan, C. Davatzikos, J. Duncan, B. Van Ginneken, A. Madabhushi, J. Prince, D. Rueckert, R. Summers, A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc. IEEE. **109**(5), 820–838 (2021)

Publisher's Note