# Blind Data Adaptation to tackle Covariate Shift in Operational Steganalysis

Rony Abecidan, Vincent Itier, Jérémie Boulanger, Patrick Bas, Tomáš Pevný

# Blind Data Adaptation to tackle Covariate Shift in Operational Steganalysis

**Rony Abecidan**
Univ. Lille, CNRS, Centrale Lille
UMR 9189 CRIStAL
F-59000 Lille, France
ronyabecidan@protonmail.com

**Vincent Itier**
Centre for Digital Systems, Univ. Lille, CNRS, Centrale Lille,
UMR 9189 CRIStAL
F-59000 Lille, France
vincent.itier@imt-nord-europe.fr

**Jérémie Boulanger**
Univ. Lille, CNRS, Centrale Lille,
UMR 9189 CRIStAL
F-59000 Lille, France
jeremie.boulanger@univ-lille.fr

**Patrick Bas**
Univ. Lille, CNRS, Centrale Lille,
UMR 9189 CRIStAL
F-59000 Lille, France
patrick.bas@cnrs.fr

**Tomáš Pevný**
Department of Computers and Engineering
Czech Technical University
Prague, Czech Republic
pevnak@protonmail.ch

## Abstract

The proliferation of image manipulation for unethical purposes poses significant challenges in social networks. One particularly concerning method is Image Steganography, allowing individuals to hide illegal information in digital images without arousing suspicions. Such a technique pose severe security risks, making it crucial to develop effective steganalysis methods enabling to detect manipulated images for clandestine communications. Although significant advancements have been achieved with machine learning models, a critical issue remains: the disparity between the controlled datasets used to train steganalysis models against real-world datasets of forensic practitioners, undermining severely the practical effectiveness of standardized steganalysis models. In this paper, we address this issue focusing on a realistic scenario where practitioners lack crucial information about the limited target set of images under analysis, including details about their development process and even whereas it contains manipulated images or not. By leveraging geometric alignment and distribution matching of source and target residuals, we develop `TADA` (Target Alignment through Data Adaptation), a novel methodology enabling to emulate sources aligned with specific targets in steganalysis, which is also relevant for highly unbalanced targets. The emulator is represented by a light convolutional network trained to align distributions of image residuals.

Preprint. Under review.

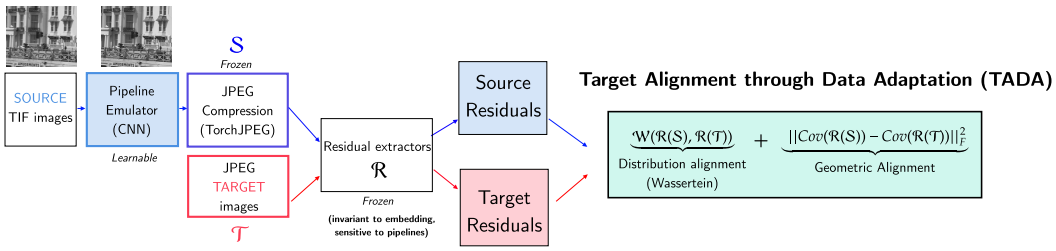


Figure 1: Illustration of Target Alignment through Data Adaptation (TADA). Data adaptation consists in learning a small convolutional network which emulates the development pipeline. The alignment is performed by using a dual loss matching the distribution of the residuals of the target images with the distribution of the residuals of the emulated image coming from the source. TADA proposes an end-to-end emulation mechanism to realign the two distributions in order to reduce covariate shift.

Experimental validation demonstrates the potential of our strategy over traditional methods fighting the steganalysis covariate shifts.

## 1 Introduction

*Steganography*, the science of concealing information within seemingly innocuous data, is today a formidable tool for cybercriminals. With the help of numerous softwares available online, it is indeed possible for anyone to hide secret information into text, videos and images, enabling therefore to communicate secretly or embed viruses in total discretion [1]. Such security concerns justify the popularity of steganography in the literature. Among the most famous steganographic schemes, those tailored for images are particularly studied. This is explained by the significant embedding capacity afforded by digital images combined with their widespread distribution online. The JPEG domain is for instance well studied for steganography due to its prevalent use on social networks. Among the well-known embedding techniques within this domain, we have UERD [2] and J-UNIWARD [3], which are respectively based on DCT and wavelet decompositions. These techniques leverage textured areas of images to embed messages within their natural noise, thereby making detection more arduous. Forensic analysts are aware of these manipulations and with the help of machine learning, they train *Steganalysis* detectors to determine the authenticity of images under scrutiny in the context of criminal investigations.

Because steganographic manipulations are imperceptible and can manifest in diverse forms, it is common to leverage supervised models to build steganalysis detectors. It requires therefore the use of a training base of genuine (a.k.a *covers*) and manipulated (a.k.a. *stegos*) images to build steganalysis detectors. Researchers are nowadays training their steganalysis detectors using toy cover images coming from carefully prepared databases like BOSSBASE [4] or ALASKABASE [5].

It has been shown in [6–8] that noise residuals are helping to extract discriminative features for Steganalysis. Provided the payload of the images under scrutiny is large enough[1], a simple linear classifier can leverage these features to perform efficient steganalysis. Naturally, high levels of noise within an image facilitate the concealment of a message, thereby making it more challenging to differentiate between covers and stegos. For the most difficult cases, the state of the art for steganalysis lies on convolutional neural networks such as YedroudjNet [9], SRNet [10] and JIN-SRNet [11] that are also using noise residuals by training several high resolution convolutional layers to improve their predictions. Unfortunately for practitioners, steganalysis detectors are naturally failing with real-world images that come from completely unknown distributions. In machine learning, this scenario is well-known and referred as *covariate shift* in [12] while in steganalysis, this is referred as *Cover-Source Mismatch (CSM)* since it results directly from a mismatch of cover distributions [13–15]. This mismatch comes from the fact that significant variation is found in cover distributions, commonly known as cover sources, which is due to several aspects of the image acquisition process. These aspects range from the type of capturing sensor (such as CCD or CMOS), its quality (its inherent ability for a given ISO setting to generate a photonic noise of small power), the capturing

[1]This notion of a "large" payload is relative to the noisy properties of the images at stake

parameters (including ISO, exposure time and aperture), the traits of the captured image (like lighting conditions and content uniformity), the post-processing actions (such as white balance adjustment and gamma correction), to the compression steps that follow (such as 8-bit conversion and JPEG compression). These operations used for visual enhancing and compression are modifying the noise distribution upon which steganalysis detectors rely. Thus, given the sensitivity of machine learning models to their training distributions, the processing pipeline is broadly identified as the mainstay of CSM [13].

## 2 Contributions

While several strategies are proposed to address covariate shift in machine learning ([16–22]) and CSM in steganalysis ([14, 15, 13, 23, 24]), very few of them are really efficient in the realistic scenarios where:

- The processing pipeline of images under scrutiny is totally unknown (neither the nature of the operations undertaken nor the hyperparameters used for each operation are known).
- The steganalyst observes only a small set of images with unknown development pipelines.
- We are ignorant about the class of each image to analyze (cover or stego). There is possibly a high unbalance in terms of cover-stego image among the test set (extreme unbalance appears most of the time because in practice most users are innocent).

Faced with that reality, our contributions are twofold:

- We propose a novel domain adaptation strategy to cope with covariate shift in JPEG Steganalysis based on a new geometric interpretation of CSM illustrated in Figure 1. This strategy lies on `TADA` (Target Alignment through Data Adaptation), a small convolutive architecture learning how to develop diverse RAW images from ALASKABASE [5] so that noise residuals of target and source images are aligned with the ones of the images under scrutiny.
- This data adaptation strategy relies on a combination of two complementary losses. The first term considers the alignment of principal axis (eigenvectors) and spreading (eigenvalues) of both residual distributions equalizing covariance matrices. The second term considers the Wassertein distance between these same distributions in order to be robust to bias which are not captured by the first one.

As far as we know, this data adaptation strategy is the first effort to address covariate shift in steganalysis by proposing a neural architecture designed to emulate a relevant source dataset with desired target statistics, especially in cases where our knowledge about these targets is very limited. Both toy and real-world experiments underscore the potential of `TADA` over state of the art methods available to forensic practitioners (see Section 6 and Appendix A.3). In contrast with other strategies such as training on appropriate mixtures of sources or, picking the closest source from a relevant set of sources, our method is adaptive and enable to emulate a pipeline as close as possible to the true development pipeline used to generate the test images.

### 2.1 Outline of the paper

Section 3 discusses related research on covariate shift in steganalysis from both machine learning and steganalysis perspectives. In section 4, we introduce the `TADA` approach, starting with the formalization of our steganalysis issue followed by a detailed explanation of the learning mechanism designed to emulate the development pipeline of a specific target. Afterwards, Section 5 presents an overview of `TADA` training, detailing important pre-processing steps and the learning metrics that are optimized. Finally, Section 6 demonstrates through experiments that `TADA` can surpass state-of-the-art methods on both toy and real-world targets, even when the targets are unbalanced and contain a limited number of images.

## 3 Related Works

Strategies commonly employed to address covariate shift in a blind scenario can be categorized into two frameworks.

First, there are *Domain Randomization* strategies used for instance in [22] and [25] aiming at building a training set fostering noise and content diversity through a mixture of multiple image distributions. This essentially consists in augmenting the training set in a clever way so that we achieve broad generalization across various targets. Although [26] demonstrates the potential of this strategy against CSM, reference [23] suggests that not all combinations of distributions are equally effective, emphasizing that quality prevails over quantity for optimal generalization. Thus, maximizing the generalization ability of a steganalysis detector requires identifying the most suitable combination of distributions for a specific testing set. Studies like [15] and [24] are precisely proposing metrics to assess the relevance of a cover source w.r.t. a specific target.

Secondly, *Unsupervised Domain Adaptation* strategies [27] make the most of all available data to tackle covariate shift. In this scenario, the goal is to transfer knowledge obtained from a labeled training set (a.k.a the *source*) to an unlabeled evaluation set (a.k.a. the *target*). This problem is well-known in machine learning and one famous strategy to cope with it involves embedding source and target into a common domain invariant space, so that the domain discrepancy term of the generalization bound of Ben David *et al.* [28] is minimized. For instance, Ganin *et al.* [29] harness backpropagation to directly learn a domain-invariant projection while training a classifier. Their approach entails integrating and adversarially training a domain discriminator into the final layer of the neural network, hence fostering the creation of a representation where distinguishing between source and target domains becomes challenging. In the same vein, Long *et al.* introduce in 2015 a Deep Adaptation Network (DAN) [21] allowing feature transferability in downstream layers of a CNN, ensuring that source and target distributions are close in the last projections through the minimization of the Maximum Mean Discrepancy (MMD).

At last, there are also strategies aiming to find a transformation directly embedding the source into the target domain. For example, Fernando et al. propose in [17] to align the subspaces spanned by source and target eigenvectors after applying a PCA. This strategy is very similar to the one of [24] where the authors propose to deduce the processing pipeline applied to scrutinized images minimizing the chordal distance between source and target DCTr features [8] with a simulated annealing. Expanding upon this idea, Sun et al. suggest in [18] to recolor whitened source features with the covariance of the target distribution. By essence, this covariance alignment encapsulates the one of [17] and [24] since the PCA projections are directly derived from covariances.

Although all these strategies have proven fruitful in several experiments, they have obvious limitations to tackle the practical setup introduced in Section 2:

- Many of them are relying on distances between distributions that are difficult to approximate with very few samples although their low sample complexities [30, 31].
- They are likely uneffective in the realistic case of highly unbalanced targets [32].
- In cases where source and target are balanced, nothing prevent the distribution to match while inverting the classes as illustrated in Figure 1 of [32].

## 4 Approach

In this section, we present `TADA` (Target Alignment through Data Adaptation), a new strategy to derive relevant sources for specific targets in steganalysis. The main idea is to learn how to process images with minimalist development in order to match target statistics. `TADA` is made of three key ingredients : a **pipeline emulator**, followed by a **feature extractor** sensitive to the processing pipeline while robust to steganographic embedding and, a **two-objectives loss function** aligning these features in terms of geometry and distributions. These steps are illustrated in Figure 1 which present an overview of the method.

### 4.1 Formalization

Using the notations from [33], we describe a processing pipeline as a vector $\omega \in \Omega$ (the infinite set of possible processing pipelines), which includes parameters like the denoising coefficient and JPEG quality factor. In steganalysis, we add an extra parameter $\gamma$ to represent choices made by the steganographer, such as the embedding method and payload. Our task is to develop an algorithm able to distinguish covers from stegos among a set of images. Machine learning models are typically employed for this task:

$$f(x \mid \theta_{\omega,\gamma}) : \mathcal{X} \to \{cover, stego\}.$$
$$x \mapsto y$$

where $\theta_{\omega,\gamma} \in \Theta$ represents all the parameters learnt using covers issued from the pipeline $\omega$ and potentially embedded following $\gamma$.

We consider now the unsupervised domain adaptation framework. We assume having access to $n_s$ labeled i.i.d. images $\mathcal{S} = \{x_i^s, y_i^s\}_{i=1}^{n_s}$ from $p((x,y)|\omega^{\mathcal{S}}, \gamma^{\mathcal{S}})$ and $n_t$ unlabeled i.i.d. images $\mathcal{T} = \{x_i^t\}_{i=1}^{n_t}$ from $p(x|\omega^{\mathcal{T}}, \gamma^{\mathcal{T}})$ with $n_t << n_s$. We aim to minimize the risk of failure of a detector trained on source data and evaluated on target data, such as:

$$\mathbb{E}_{(x,y)\sim p((x,y)|\omega^{\mathcal{T}},\gamma^{\mathcal{T}})}(f(x \mid \theta_{\omega^{\mathcal{S}},\gamma^{\mathcal{S}}}) \neq y).$$

Following Kerckhoffs' principle, we assume that $\gamma^{\mathcal{T}}$ is known. It is therefore possible for practitioners to reproduce the embedding strategy having $\gamma^{\mathcal{S}} = \gamma^{\mathcal{T}} = \gamma$. However, we assume $\omega^{\mathcal{S}} \neq \omega^{\mathcal{T}}$ leading forensic analysts to a mismatch of cover distributions causing a covariate shift [12] : $p(x|\omega^{\mathcal{S}}) \neq p(x|\omega^{\mathcal{T}})$ while $p(y|x^s, \gamma) = p(y|x^t, \gamma)$. This mismatch can be directly resolved if we manage to bring $\omega^{\mathcal{S}}$ as close as possible to $\omega^{\mathcal{T}}$. We will assume here that we can at least get access to the quantification tables of the target images[2].

### 4.2 Emulation of realistic pipelines

Among all the possible operations we could perform on images before JPEG compression, we show in [23] that denoising and sharpening are highly responsible of the covariate shift observed in steganalysis. Although these operations are not always linear, there exist linear versions of them such as Median Filter or Unsharp Masking largely used in classical softwares like Photoshop, GIMP, RawTherapee, *etc*. More precisely, traditional image operations can be reasonably approximated with symmetric convolutions which sums to 1 as assumed in [35]. For instance, $3 \times 3$ kernel satisfying these conditions are structured as:

$$\begin{bmatrix} b & c & b \\ c & a & c \\ b & c & b \end{bmatrix} \text{ with } a + 4(b + c) = 1. \tag{1}$$

We propose therefore to use a unique convolution of this shape to emulate the target pipeline in `TADA`. By choosing such a simple developer, we cannot reasonably approximate highly non-linear operations as well as resizing operations. We are aware of this limitation and we are working on making `TADA` compatible with such operations. However, we show here that a simple constrained convolution can already be very helpful in practical situations.

To be able to learn this convolution appropriately, we define now a differentiable metric assessing the proximity of two developping pipelines given the images they produce.

### 4.3 Derive processing pipelines fingerprints from noise residuals

By nature steganalysis focuses on weak signals that do not rely on the image content. Camera sensors introduce different types of noise into captured images [36]. In the RAW format, immediately after acquisition, the noise at the pixel level is independently distributed and follow a Poisson/Gaussian distribution [37]. Once a processing pipeline is applied to these RAW images, the noise pixels are correlated.

[2]This information is often public. It's also possible to estimate it [34]

Since different processing pipelines affects noise correlations differently, Mallet *et al.* [38] propose therefore to consider these noise correlations as fingerprints of the processing pipeline. This approach is especially interesting in steganalysis considering the high sensitivity of noise residuals to the development pipeline, but also considering the robustness of theses fingerprints to the embedding, which is experimentally verified in [38]. We can explain this robustness by considering that even early embedding schemes such as Model-Based steganography [39] or nsF5 [40], designed their embedding in order to preserve at least the second order moment of the Cover distribution, or the whole marginal distribution during the embedding.

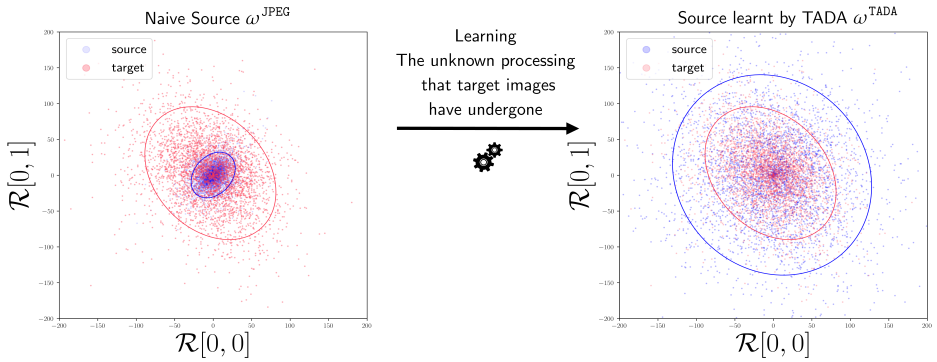


Figure 2: Covariance alignment performed by `TADA` - We highlight using scatter plots of image residuals, the fact that the covariate shift in steganalysis may be caused by a discrepancy between the covariance matrices of source and target residuals. Each point represents two neighboring samples of an image subject to a high-pass filter $\mathcal{R}$, *i.e.* a 2D residual.

A simple approach to extract noise residuals involves using high-pass filters on the images [7]. There are several filters designed for residuals extraction removing steganalysis traces while highlighting inter-pixel correlations, such as the KB filter [41] with coefficients $[[-\frac{1}{4},\frac{1}{2},-\frac{1}{4}][\frac{1}{2},-1,\frac{1}{2}][-\frac{1}{4},\frac{1}{2},-\frac{1}{4}]]$.

Referring to $\mathcal{R}$ as a high-pass filter capturing image noise residuals, [38] inspires us to leverage $||\operatorname{Corr}(\mathcal{R}(\mathcal{S})) - \operatorname{Corr}(\mathcal{R}(\mathcal{T}))||_F$ as a differentiable loss function to guide the learning of $\omega^{\texttt{TADA}}$. However, after investigations we discovered that equalizing noise correlations is not enough to avoid Cover Source Mismatch. We explain this observation by scale and shift invariance properties of correlations, leading our pipeline learning towards suboptimal pipelines. Hence, rather than equalizing correlations, we propose to equalize covariances to preserve both alignment (e.g. eigenvectors) and spreading (e.g. eigenvalues) of noise residuals.

This approach is based on a key observation: directions with high variance in noise residuals are ideal for steganographers to hide their messages. Consequently, if there are specific directions in the residual distribution where the variance is high, steganographic methods will push the distribution in these directions. To distinguish between stegos and covers, we can project all images onto these high-variance directions and apply a variance threshold. Stego images will show significantly higher variance compared to cover. However, if the source pipeline differs from the target pipeline, the residual geometry will also differ. This means the principal axes might not align, or even if they do, the eigenvalues could vary, either being weaker or stronger, which renders variance-based thresholding less effective.

The covariance is not scale invariant but is still shift invariant. Hence, we prefer to not use it alone. We propose to also use a distance between residual distributions so that our final loss is sensitive to distribution shifts. Hence, to be able to learn a relevant pipeline, we consider the following loss function for `TADA`:

$$\mathcal{L} = \lambda \underbrace{\|\operatorname{Cov}(\mathcal{R}(\mathcal{S})) - \operatorname{Cov}(\mathcal{R}(\mathcal{T}))\|_F^2}_{\text{Geometric alignment}} + \mu \underbrace{\mathcal{D}(\mathcal{R}(\mathcal{S}),\mathcal{R}(\mathcal{T}))}_{\substack{\text{Distribution alignment} \\ \text{(e.g., MMD, Wasserstein)}}} \quad \text{with } \mu \text{ and } \lambda \text{ to tune.} \tag{2}$$

## 5 Details about TADA training

We detail in this section how we propose to train TADA to accurately approximate $\omega^{\mathcal{T}}$.

### 5.1 General considerations

To make noise residuals approximation as accurate as possible, we select randomly 500 RAWs from ALASKABASE [5] and extract in each of them a 512x512 crop as uniform as possible. This constitutes a RAW base that can be developed for every pipeline to learn.

Concerning the development, we initialize it with an identity filter to which we apply a centered gaussian noise with a standard deviation of $0.01$. The constraint of the learnt kernel are artificially enforced at the end of each epoch to avoid disrupting the training phase. Additionally, we do not use a padding strategy for our convolutions, as padding can cause undesirable border effects in the produced images [42].

Once RAWs are developed, the differentiable JPEG compression defined in [43] is performed, using the quantification table of target images. The initialization choice prevents us from reaching pipelines generating saturated images that ends up to be uniformly black after this JPEG compression.

Right after, we compute noise and target residuals with a combination of two high pass filters to bring more diversity : the KB filter [41] and the Laplacian filter with 4 neighbors ($\mathcal{L}_4$ in [38]).

In order to capture both the intra-bloc and the inter-bloc correlations between JPEG blocs, we compute the covariance matrix from $8 \times 16$ patches aligned on the JPEG grid. Such a neighborhood has proven to capture the most relevant information of the development pipeline in [44]. Within the pool of extracted patches, those exhibiting the lowest variance typically originate from areas of the images with high uniformity. As a result, they fail to effectively differentiate the various emulated pipelines. Indeed, given that kernels summing to 1 are employed, any emulated pipeline yields null residuals within highly uniform region. Conversely, patches with high variance tend to capture textured content where noise residuals are less robust to steganographic schemes. To address this issue, we suggest to select patches with variance falling within the 30th and 60th quantiles of the variance distribution.

The training loss is computed using the selected patches. As explicated in section 4, this loss combines a distance between residuals distributions and a distance between residuals covariances. Concerning the distance between distribution, we propose to use the Earth's Mover distance (also called Wassertein $\mathcal{W}$), a differentiable metric from optimal transport enabling us to avoid vanishing or exploding gradients commonly encountered with other measures like Maximum Mean Discrepancy [45] [46]. We also notice that including the Frobenius norm between correlations of residuals enable to speeds up convergence significantly. Smooth learning is enabled by normalizing all our losses with their values at initialization. These normalizations bring all costs to the same scale, making it easier to adjust the learning rate. However, it's more relevant to work with the unnormalized sum of these costs when evaluating any learned pipeline. Keeping the use of normalized costs in the evaluation phase might unfairly favor pipelines that are very good at minimizing one particular cost, while ignoring others. Therefore, at the end of every epoch, TADA computes the unnormalized sum $\mathcal{L}_{eval} = \| \operatorname{Cov}(\mathcal{R}(\mathcal{S})) - \operatorname{Cov}(\mathcal{R}(\mathcal{T}))\|_F^2 \ + \mathcal{W}(\mathcal{R}(\mathcal{S}), \mathcal{R}(\mathcal{T}))$ and save the final weights minimizing it.

## 6 Experiments on several targets

### 6.1 Experimental protocol

To validate the potential of TADA, we build toy and real-world targets for which we would like to craft tailored sources. The toy targets are created with RawTherapee combining Wavelet Denoising followed by Unsharp Masking and a JPEG compression of QF 85 so that they lead together to high performance drops. It is important to realize that this combination leads to non-linear pipelines. Yet, we will see that TADA can nevertheless derive meaningful convolutions to generalize on them. The real-world targets are built using the database YFCC100M [47] gathering millions of flickr images under CC licenses. From this database, we look for users sharing non-resized images with public quantification tables to comply with our assumptions. This scenario comply with the reality of practitioners since the processing pipeline used by flickr users is totally unknown to us. We finally

find 3 users sharing thousands of pictures with the same camera model and compressed with the same quantification tables. Details about all our targets are also presented in Appendix A.1.

From these targets, we derive fictitious training and evaluation sets to establish a benchmark for the optimal performance achievable. Additionally, we create an operational set that acts as the unlabeled small subset available to the forensic analyst useful to train TADA. All our experiments involves 1.000 cover-stego pairs for training sets, 500 cover-stego pairs for evaluation sets and at maximum 500 images of unknown natures for the operational set. We will consider 3 possibilities for this last set, either it is made of only covers, either it is made of only stegos or either it is made of a balanced mixtures of covers and stegos. The embedding strategy chosen for all the experiments is UERD [2] with a payload of 0.5 bits per non-zero AC DCT coefficient of the luminance channel (bpnzac), a reasonable choice commonly adopted by the steganalysis community [11, 24]. We train TADA using V100 GPUs. The TIF images to develop are cut into mini-batches of size 256.

Concerning the RAWs to develop after the learning of a relevant pipeline with TADA, we randomly sample 1.000 textured covers of size $512 \times 512$ from ALASKABASE using the smart cropping algorithm of the authors [5]. Then, we develop them with $\omega_{\texttt{TADA}}$ and create 1.000 cover-stego pairs with UERD to train a steganalysis detector supposed to be efficient on our targets.

For every TADA learning phase, we fix the maximum number of epochs to 1000. We also select the SGD as our optimizer with a learning rate of 0.001. A learning rate scheduler divides by 2 this learning rate once our $\mathcal{L}_{eval}$ is not minimized after 100 epochs. At last, an earlystopping procedure stops the learning phase when this same metric did not decrease through 200 successive epochs.

### 6.2 Comparison with other strategies

As our main experiment, we assume to get access to 500 unlabeled images of every target. We propose to learn a relevant source for each of them using 7x7 convolutions with TADA. We then compare the target accuracy obtained with TADA sources against several methods fighting CSM in a practical scenario where the operational set may be totally unbalanced. Following the notations of [48], we will name *TgtOnly* the ideal scenario that assume access to labeled target images and, *SrcOnly* the strategy that simply compress RAWs with targets quantification tables. By using the set-covering strategy of [23], we extract eight representatives pipelines $\omega \in \Omega_R$ covering the set of 1.000 sources generated by [24] with a maximum drop of performance of $5\%$. We propose to name *All*, the domain randomization strategy involving to mix these relevant sources while compressing them with targets quantification tables. Using again $\Omega_R$, we also propose other sota strategies. With the help of a linear multiclassifier, [13] suggests to assign to each target image a specific detector among those trained on $\Omega_R$. Other works such as [15] and [24] propose to only use the detector trained on the most closest source to the scrutinized target. As explicated by [24], this notion of closeness between source and targets could be translated with metrics such as the MMD or the chordal distance (NSCD) between DCTr features [8] of each domain. We propose here to add the EarthMover distance and the Frobenius norm between covariance matrices of DCTr, two metrics respectively related to MMD and NSCD that also may be useful. At last, we suggest to test the subspace alignment of [17] and the Covariance Alignment of [18] since these famous strategies are not impacted by class unbalance. For the subspace alignment, we tune the dimension parameter so that we get the best target accuracy possible. All the results presented in Table 1 are made using a logistic regression on DCTr features of the selected sources.

As demonstrated by this table, TADA is rather competitive for most of toy strategies while its core assumption of linear processing pipeline is violated. Concerning the realistic targets, its performance are particularly impressive on the SONY target for which we absolute do not know the processing pipeline, highlighting its potential in practical scenarios. Additionally, we notice that the great performance of this strategy are rather stable over our 3 operational balancing, showing its robustness to the embedding scheme.

### 6.3 Limitations

We must admit that TADA is sub-optimal for three targets : RT5, NIKON and CANON. For the RT5 dataset, where the sharpen radius is large, it is probably due to the fact that the convolutional filter is too small. For the NIKON dataset the performances are subpar w.r.t. strategies using the closest cover-set, which relies on a true development pipeline and not and emulated one. For the CANON

dataset, non of the proposed approaches are satisfying and TADA, with a poor accuracy of $56\%$ is better than the random strategies with are close to random guesses. We hypothesize that images from these targets have been developed with different processing pipelines.

Following this general experiment, we performed an ablation study modifying several ingredients of TADA revealing its potential on our real-world targets under different setups. Our results shows that TADA is still relevant with twice less available target images and can raise the accuracy on SONY and CANON to respectively $77\%$ and $69\%$ if noise residual extractors and training losses are well chosen. We also tested SOTA CNN detectors (e.g. JIN [11]) trained with TADA sources and we end up with competitive sources against the mixture of the *All* strategy for SONY and NIKON. See more details in Appendix.

Table 1: Comparison of TADA performance vs traditional methods to fight cover source mismatch. The results displayed are target accuracies in %. 3 cases are studied for the operational set : full cover, balanced mixture of cover and stego, full stego. Best results by target are printed in bold.

| **Full cover** | RT1 | RT2 | RT3 | RT4 | RT5 | SONY | NIKON | CANON |
|---|---|---|---|---|---|---|---|---|
| TgtOnly | 86 | 86 | 86 | 85 | 77 | 92 | 79 | 81 |
| SrcOnly | 73 | 68 | 68 | 69 | 59 | 54 | 64 | 50 |
| All [23] | 74 | 66 | 68 | 68 | **68** | 61 | 70 | 52 |
| Multiclassifier [13] | 67 | 72 | 62 | 62 | 62 | 56 | **75** | 50 |
| $\min_{\omega\in\Omega_R}$ NSCD(DCTr($\omega$), DCTr($\omega^{\mathcal{T}}$)) [24] | **75** | 76 | 62 | **77** | 62 | 57 | **75** | 51 |
| $\min_{\omega\in\Omega_R}$ $\|\|\text{Cov}(\text{DCTr}(\omega)) - \text{Cov}(\text{DCTr}(\omega^{\mathcal{T}}))\|\|_F$ | **75** | 76 | 62 | **77** | 62 | 57 | **75** | 50 |
| $\min_{\omega\in\Omega_R}$ MMD(DCTr($\omega$), DCTr($\omega^{\mathcal{T}}$)) [24] | 64 | 76 | 62 | **77** | 62 | 57 | **75** | 50 |
| $\min_{\omega\in\Omega_R}$ $\mathcal{W}$(DCTr($\omega$), DCTr($\omega^{\mathcal{T}}$)) | 64 | 76 | 62 | **77** | 62 | 57 | **75** | 50 |
| Subspace Alignment [17] | 72 | 69 | 69 | 70 | 66 | 66 | 51 | 51 |
| CORAL [18] | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| TADA (Ours) | 74 | **77** | **77** | 76 | 61 | **74** | 62 | **56** |
| **Mix** | RT1 | RT2 | RT3 | RT4 | RT5 | SONY | NIKON | CANON |
| TgtOnly | 86 | 86 | 86 | 85 | 77 | 92 | 79 | 81 |
| SrcOnly | 73 | 68 | 68 | 69 | 59 | 54 | 64 | 50 |
| All [23] | 74 | 66 | 68 | 68 | **68** | 61 | 70 | 52 |
| Multiclassifier [13] | 67 | 72 | 62 | 62 | 62 | 56 | **75** | 50 |
| $\min_{\omega\in\Omega_R}$ NSCD(DCTr($\omega$), DCTr($\omega^{\mathcal{T}}$)) [24] | **75** | 76 | 62 | **77** | 62 | 57 | **75** | 51 |
| $\min_{\omega\in\Omega_R}$ $\|\|\text{Cov}(\text{DCTr}(\omega)) - \text{Cov}(\text{DCTr}(\omega^{\mathcal{T}}))\|\|_F$ | **75** | 76 | 62 | **77** | 62 | 57 | **75** | 50 |
| $\min_{\omega\in\Omega_R}$ MMD(DCTr($\omega$), DCTr($\omega^{\mathcal{T}}$)) [24] | 64 | 76 | 62 | **77** | 62 | 57 | **75** | 50 |
| $\min_{\omega\in\Omega_R}$ $\mathcal{W}$(DCTr($\omega$), DCTr($\omega^{\mathcal{T}}$)) | 64 | 76 | 62 | **77** | 62 | 57 | **75** | 50 |
| Subspace Alignment [17] | 72 | 70 | 69 | 69 | 66 | 66 | 51 | 51 |
| CORAL [18] | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| TADA (Ours) | 74 | **77** | **77** | 76 | 63 | **68** | 62 | **56** |
| **Full stego** | RT1 | RT2 | RT3 | RT4 | RT5 | SONY | NIKON | CANON |
| TgtOnly | 86 | 86 | 86 | 85 | 77 | 92 | 79 | 81 |
| SrcOnly | 73 | 68 | 68 | 69 | 59 | 54 | 64 | 50 |
| All [23] | 74 | 66 | 68 | 68 | **68** | 61 | 70 | 52 |
| Multiclassifier [13] | 67 | 72 | 62 | 62 | 62 | 56 | **75** | 50 |
| $\min_{\omega\in\Omega_R}$ NSCD(DCTr($\omega$), DCTr($\omega^{\mathcal{T}}$)) [24] | **75** | 76 | 62 | **77** | 62 | 57 | **75** | 51 |
| $\min_{\omega\in\Omega_R}$ $\|\|\text{Cov}(\text{DCTr}(\omega)) - \text{Cov}(\text{DCTr}(\omega^{\mathcal{T}}))\|\|_F$ | **75** | 76 | 62 | **77** | 62 | 57 | **75** | 50 |
| $\min_{\omega\in\Omega_R}$ MMD(DCTr($\omega$), DCTr($\omega^{\mathcal{T}}$)) [24] | 64 | 76 | 62 | **77** | 62 | 57 | **75** | 50 |
| $\min_{\omega\in\Omega_R}$ $\mathcal{W}$(DCTr($\omega$), DCTr($\omega^{\mathcal{T}}$)) | 64 | 76 | 62 | **77** | 62 | 57 | **75** | 50 |
| Subspace Alignment [17] | 70 | 70 | 70 | 69 | 66 | 66 | 51 | 51 |
| CORAL [18] | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| TADA (Ours) | **75** | **77** | **77** | 75 | 62 | **70** | 62 | **53** |

## 7 Conclusion

To tackle Cover Source Mismatch in steganalysis, we introduce TADA, a strategy enabling to create tailored sources for any steganalysis target even when they are small and highly unbalanced. This strategy lies on the alignment of noise residuals in terms of geometry (with their covariance) and distributions (with a Wassertein metric). Several experiments are conducted to demonstrate that our strategy is promising and can outperform traditional methods fighting CSM in practical scenarios. However, we also observe that TADA's power is limited by its oversimplistic assumption that every pipeline can be approximated with only one convolution. Despite this limitations, there is considerable room for improvement considering the efficiency of neural networks to handle non-linear relationships. Furthermore, in scenarios where target images result from different processing pipelines, TADA, by its construction, cannot perform effectively. Nevertheless, with a clustering approach harnessing the features of [38], we can try to group images based on their processing pipeline proximity, offering a potential solution by applying TADA to each cluster. To conclude, TADA pave the way for more robust and reliable steganalysis in diverse and complex real-world environments.

## Acknowledgements

Our experiments were possible thanks to computing means of IDRIS through the resource allocation 2023-AD011013285R2 assigned by GENCI. This work received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101021687 (project "UNCOVER") and the French Defense & Innovation Agency. The work of Tomáš Pevný was supported by Czech Ministry of Education 19-29680L.

## References


[1] W. Mazurczyk and S. Wendzel, "Information hiding: Challenges for forensic experts," *Commun. ACM*, vol. 61, p. 86–94, dec 2017.

[2] L. Guo, J. Ni, W. Su, C. Tang, and Y.-Q. Shi, "Using statistical image model for jpeg steganography: Uniform embedding revisited," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2669–2680, 2015.

[3] V. Holub, J. J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP Journal on Information Security*, vol. 2014, pp. 1–13, 2014.

[4] P. Bas, T. Filler, and T. Pevny, ""Break Our Steganographic System": The Ins and Outs of Organizing BOSS," in *INFORMATION HIDING*, vol. 6958/2011 of *Lecture Notes in Computer Science*, (Czech Republic), pp. 59–70, May 2011.

[5] R. Cogranne, Q. Giboulot, and P. Bas, "The ALASKA Steganalysis Challenge: A First Step Towards Steganalysis "Into The Wild"," ACM IH&MMSec (Information Hiding & Multimedia Security), (Paris, France), July 2019.

[6] T. Pevny, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 2, pp. 215–224, 2010.

[7] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[8] V. Holub and J. Fridrich, "Low-complexity features for jpeg steganalysis using undecimated dct," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 219–228, 2015.

[9] M. Yedroudj, F. Comby, and M. Chaumont, "Yedrouj-net: An efficient CNN for spatial steganalysis," *CoRR*, vol. abs/1803.00407, 2018.

[10] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1181–1193, 2019.

[11] J. Butora, Y. Yousfi, and J. Fridrich, "How to pretrain for steganalysis," in *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, IHamp;MMSec '21, (New York, NY, USA), p. 143–148, Association for Computing Machinery, 2021.

[12] J. G. Moreno-Torres, T. Raeder, R. Alaíz-Rodríguez, N. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognit.*, vol. 45, pp. 521–530, 2012.

[13] Q. Giboulot, R. Cogranne, D. Borghys, and P. Bas, "Effects and Solutions of Cover-Source Mismatch in Image Steganalysis," *Signal Processing: Image Communication*, Aug. 2020.

[14] J. Pasquet, S. Bringay, and M. Chaumont, "Steganalysis with cover-source mismatch and a small learning database," in *EUSIPCO: European Signal Processing Conference*, (Lisbon, Portugal), pp. 2425–2429, Sept. 2014.

[15] J. Kodovský, V. Sedighi, and J. Fridrich, "Study of cover source mismatch in steganalysis and ways to mitigate its impact," in *Media Watermarking, Security, and Forensics 2014*, vol. 9028.

[16] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[17] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *2013 IEEE International Conference on Computer Vision*, pp. 2960–2967, 2013.

[18] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," *CoRR*, vol. abs/1511.05547, 2015.

[19] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *CoRR*, vol. abs/1507.00504, 2015.

[20] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073, 2012.

[21] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 97–105, PMLR, 07–09 Jul 2015.

[22] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," *CoRR*, vol. abs/1703.06907, 2017.

[23] R. Abecidan, V. Itier, J. Boulanger, P. Bas, and T. Pevný, "Using set covering to generate databases for holistic steganalysis," 2022.

[24] R. Abecidan, V. Itier, J. Boulanger, P. Bas, and T. Pevný, "Leveraging Data Geometry to Mitigate CSM in Steganalysis," in *IEEE International Workshop on Information Forensics and Security (WIFS 2023)*, (Nuremberg, Germany), Dec. 2023.

[25] K. Lee, K. Lee, J. Shin, and H. Lee, "A simple randomization technique for generalization in deep reinforcement learning," *CoRR*, vol. abs/1910.05396, 2019.

[26] I. Lubenko and A. D. Ker, "Steganalysis with mismatched covers: do simple classifiers help?," in *Proceedings of the on Multimedia and Security*, MM&Sec '12, (New York, NY, USA), p. 11–18, Association for Computing Machinery, 2012.

[27] W. M. Kouw, "An introduction to domain adaptation and transfer learning," *CoRR*, vol. abs/1812.11806, 2018.

[28] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, (Cambridge, MA, USA), p. 137–144, MIT Press, 2006.

[29] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 1180–1189, PMLR, 07–09 Jul 2015.

[30] W. Zaremba, A. Gretton, and M. B. Blaschko, "B-test: A non-parametric, low variance kernel two-sample test," *CoRR*, vol. abs/1307.1954, 2013.

[31] I. O. Tolstikhin, B. K. Sriperumbudur, and B. Schölkopf, "Minimax estimation of maximum mean discrepancy with radial kernels," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[32] H. Zhao, R. T. D. Combes, K. Zhang, and G. Gordon, "On learning invariant representations for domain adaptation," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 7523–7532, PMLR, 09–15 Jun 2019.

[33] D. Šepák, L. Adam, and T. Pevný, "Formalizing cover-source mismatch as a robust optimization," in *EUSIPCO: European Signal Processing Conference*, (Belgrade, Serbia), Sept. 2022.

[34] T. H. Thai, R. Cogranne, F. Retraint, *et al.*, "Jpeg quantization step estimation and its applications to digital image forensics," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 123–133, 2016.

[35] E. Kee and H. Farid, "Digital image authentication from thumbnails," in *Electronic imaging*, 2010.

[36] T. Kuroda, *Essential Principles of Image Sensors*. 12 2017.

[37] A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian, "Practical poissonian-gaussian noise modeling and fitting for single-image raw-data," *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1737–1754, 2008.

[38] A. Mallet, P. Bas, and R. Cogranne, "Statistical correlation as a forensic feature to mitigate the cover-source mismatch," in *12th ACM Workshop on Information Hiding and Multimedia Security (ACM IH&MMSEC'24)*, 2024.

[39] P. Sallee, "Model-based steganography," in *International workshop on digital watermarking*, pp. 154–167, Springer, 2003.

[40] J. Fridrich, T. Pevnỳ, and J. Kodovskỳ, "Statistically undetectable jpeg steganography: dead ends challenges, and opportunities," in *Proceedings of the 9th workshop on Multimedia & security*, pp. 3–14, 2007.

[41] A. D. Ker and R. Böhme, "Revisiting weighted stego-image steganalysis," in *Security, Forensics, Steganography, and Watermarking of Multimedia Contents X* (I. Delp, Edward J., P. W. Wong, J. Dittmann, and N. D. Memon, eds.), vol. 6819 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 681905, Feb. 2008.

[42] J. Butora and P. Bas, "Size-Independent Reliable CNN for RJCA Steganalysis," *IEEE Transactions on Information Forensics and Security*, pp. 2683–2695, Mar. 2024.

[43] R. Shin, "Jpeg-resistant adversarial images," 2017.

[44] T. Taburet, P. Bas, W. Sawaya, and J. Fridrich, "Natural steganography in jpeg domain with a linear development pipeline," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 173–186, 2020.

[45] J. Feydy, *Geometric data analysis, beyond convolutions*. Theses, Université Paris-Saclay, July 2020.

[46] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017.

[47] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li, "The new data and new challenges in multimedia research," *CoRR*, vol. abs/1503.01817, 2015.

[48] H. Daumé III, "Frustratingly easy domain adaptation," pp. 256–263, 2007.

# A Appendix / supplemental material

## A.1 Details abouts the different sources used in our experiments.

Table 2: Details about parameters of the processing pipelines generating toy targets

| Toy Targets | Denoise Luma | Sharpen Radius | Sharpen Amount | Sharpen Thresholds |
|---|---|---|---|---|
| RT1 | 26 | 0.176 | 225 | 20;80;2000;1200; |
| RT2 | 30 | 0.01 | 225 | 20;80;2000;1200; |
| RT3 | 30 | 0.176 | 225 | 20;80;2000;1200; |
| RT4 | 30 | 0.341 | 225 | 20;80;2000;1200; |
| RT5 | 30 | 1.5 | 225 | 20;80;2000;1200; |

Table 3: Details about real world targets

| Real-world Targets | Username | Camera Model | Quality Factor |
|---|---|---|---|
| SONY | toms travels | SONY SLT A37 | 90 |
| CANON | Andy E. Nystrom | Canon PowerShot SX30 IS | 93 |
| NIKON | NR Acampamentos | NIKON D40 | 90 |

## A.2 CSM interpretation

In this section, we complete our geometric interpretation of CSM considering scatter plots of neighboring residuals. We propose to illustrate why it's important to equalize spread (eigenvalues) on top of principal directions (eigenvectors). To do so, we propose to consider the 3x3 sharpening filter $\mathcal{S}$ :

$$\begin{bmatrix} 0 & -\frac{1}{4} & 0 \\ -\frac{1}{4} & 2 & -\frac{1}{4} \\ 0 & -\frac{1}{4} & 0 \end{bmatrix} \tag{3}$$

followed by a JPEG compression of QF 100.

Then, we propose to compare the scatterplots of neighboring samples from 2D residuals with the same filter multiplied by a 0.5 coefficient.

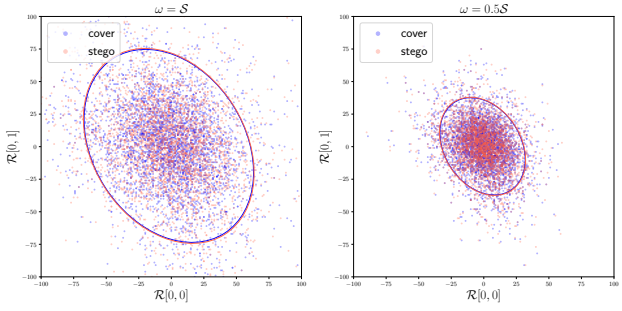


Figure 3: Geometric interpretation of CSM originating from processing operations differing from a multiplicative factor : $\mathcal{S}$ (left) and $0.5\mathcal{S}$ (right). Each point represents two neighboring samples of an image subject to a high-pass filter, *i.e.* a 2D residual.

On Figure 3 stego residuals have a slightly higher variance in the direction of the principal axis, a difference that can be used to separate cover and stego of a same source projecting on this axis. The principal axis is similar in both cases but, the threshold used to separate cover and stego is not the same potentially leading to CSM.

In this case, since JPEG QF is 100, the images resulting from $0.5\mathcal{S}$ are just resulting from a 0.5 multiplication of every pixel from $\mathcal{S}$, hence leading to identical residual correlations. However, a gap of performance is observed when we train on one source and evaluated on the other as highlighted in 4.

Table 4: Accuracy matrix showing the mismatch between $\mathcal{S}$ and $0.5\mathcal{S}$ . 2000 images for train and 1000 for test, the embedding strategy is UERD [2] with a payload of 0.5bpnzac.

| *Train* \ *Test* | $\mathcal{S}$ | $0.5\mathcal{S}$ |
|---|---|---|
| $\mathcal{S}$ | 65 | 61 |
| $0.5\mathcal{S}$ | 64 | 70 |

### A.3 Ablation studies

We perform several ablation studies to assess all the potential and limitations of TADA for real-world targets. For each study of this type, we propose to start from the main study of the paper in section 6 while changing only one ingredient at a time.

From Table 5, it seems that cutting patches of size $16 \times 16$ enables to enhance TADA performance. This is not a complete surprise as raising the size of the patch means providing more information about noise residuals distributions.

Table 5: Ablation I : What happen if we extract residual patches of different size ?

| *Target* \ *Patch size* | $8 \times 8$ | $8 \times 16$ | $16 \times 16$ |
|---|---|---|---|
| SONY | 75 | 70 | 75 |
| NIKON | 63 | 62 | 62 |
| CANON | 53 | 53 | 57 |

Table 6 highlights that raising the kernel size is not a solution enabling to perform well on any target. We see for instance that leveraging a small kernel enables to better generalize on NIKON while a rather large kernel is more interesting for SONY. We understand from this study that it's important to inject in TADA, a convolution as close as possible to the true operation applied to target pipelines.

Table 6: Ablation II: What happen if the kernel size change ?

| *Target* *Kernel size* | $3 \times 3$ | $5 \times 5$ | $7 \times 7$ | $9 \times 9$ | $11 \times 11$ |
|---|---|---|---|---|---|
| SONY | 52 | 67 | 70 | 74 | 72 |
| NIKON | 61 | 64 | 62 | 59 | 60 |
| CANON | 50 | 60 | 53 | 61 | 60 |

Table 7 shows the robustness of TADA when only few available samples from the targets is available. We see for instance that the generalization on SONY is pretty impressive in all the cases studied. Moreover, it seems that the performance on CANON is even slightly better using smaller sets of images from this target. We think this results is coming from our patch selection. With very few images, more diverse patches are taken into account even though they present very low or very high variances, and that may be slightly beneficial for CANON. This is also consistent with Table 11.

Table 8 shows the complementarity of our training losses. We can see for instance that combining the three losses is the best strategy to generalize on SONY. However, we succeed to achieve a very great performance on CANON using only the Wassertein, and a competitive performance on NIKON combining the Wassertein and the Frobenius between correlations. This study invites us to rethink our training normalization so that we can harness the best from each loss.

Table 7: Ablation III: What happen if the operational set is smaller ?

| *Target* \ *Amount of operational data* | 10 | 100 | 200 | 500 |
|---|---|---|---|---|
| SONY | 66 | 71 | 70 | 70 |
| NIKON | 62 | 63 | 62 | 62 |
| CANON | 56 | 57 | 57 | 53 |

Table 8: Ablation IV: What happen if we change the training loss ?

| *Target* \ *Patch size* | $\mathcal{W}$ | $\mathcal{L}_{corr}$ | $\mathcal{L}_{cov}$ | $\mathcal{W} + \mathcal{L}_{cov}$ | $\mathcal{W} + \mathcal{L}_{corr}$ | $\mathcal{L}_{cov} + \mathcal{L}_{corr}$ | $\mathcal{W} + \mathcal{L}_{cov} + \mathcal{L}_{corr}$ |
|---|---|---|---|---|---|---|---|
| SONY | 52 | 51 | 61 | 60 | 53 | 67 | 70 |
| NIKON | 54 | 65 | 57 | 55 | 68 | 63 | 62 |
| CANON | 69 | 60 | 59 | 56 | 59 | 57 | 53 |

Table 9 shows that the choice of the residual extractor is crucial for TADA. For instance, using only the KB filter, we jump from an accuracy of 70% on SONY to an accuracy of 77%, an impressive result for a source for which we ignore the processing pipeline. In the same vein, we are much better on CANON with the KB Filter.

Table 9: Ablation V: What happen if we change our residual extractor ?

| *Target* \ *Residual extractor* | $\mathcal{L}_4$ [38] | KB [41] | $\mathcal{L}_4$ and KB |
|---|---|---|---|
| SONY | 66 | 77 | 70 |
| NIKON | 52 | 57 | 62 |
| CANON | 58 | 60 | 53 |

Table 10 shows that our kernel constraint can be relaxed without drop of performance on our three targets. It suggests to only keep the symmetry constraint and forget the normalization to 1.

Table 10: Ablation VI: What happen if the we relax the kernel constraint ?

| *Target* \ *Constraint* | None | Sum to 1 | Symmetry | Sum to 1 and Symmetry |
|---|---|---|---|---|
| SONY | 73 | 70 | 72 | 70 |
| NIKON | 63 | 64 | 64 | 62 |
| CANON | 58 | 56 | 59 | 53 |

Finally, Table 11 suggests that our patch selection is very relevant for SONY but detrimental for NIKON and CANON. This invite us to rethink our patch selections so that we can perform the best on any target.

Table 11: Ablation VI : What happen if we relax the patch selection ?

| *Source* \ *Target* | Without patch selection | We patch selection |
|---|---|---|
| SONY | 54 | 70 |
| NIKON | 66 | 62 |
| CANON | 55 | 53 |

### A.4 Results with deep neural networks

We present at last some results we got using SOTA deep detectors with the source learnt by TADA during the main experiment of the paper. This time, TADA is competitive against SrcOnly and All for SONY and NIKON targets. This is particularly interesting considering that it was previously hard to generalize on NIKON using linear classifiers trained on TADA sources. Moreover, we again observe that it's very difficult to generalize on CANON, even though we saw in Table 8 that it's possible to reach a competitive accuracy of 69% with a TADA source, only using the Wassertein to learn it. Finally, we observe again the robustness of TADA against highly unbalanced targets.

Table 12: Target accuracies using JIN[11] on different sources

| **Strategies** | **RT1** | **RT2** | **RT3** | **RT4** | **RT5** | **SONY** | **NIKON** | **CANON** |
|---|---|---|---|---|---|---|---|---|
| TgtOnly | 91 | 94 | 93 | 94 | 92 | 87 | 93 | 91 |
| SrcOnly | 85 | 84 | 84 | 84 | **80** | 81 | 90 | **63** |
| All | **91** | **89** | **90** | **91** | 77 | 75 | 93 | 53 |
| TADA (Full Cover) | 86 | 83 | **90** | 85 | 65 | **85** | **96** | 54 |
| TADA (Mix) | 85 | 84 | 88 | 87 | 66 | **84** | **96** | 55 |
| TADA (Full Stego) | 84 | 87 | 87 | 84 | 70 | **85** | **96** | 58 |