APSIPA Transactions on Signal and Information Processing, 2025, 14, e201

Original Paper

How Much is the Source Mismatch an Important Problem for Deepfake Detection?

Antoine Mallet[1], Rémi Cogranne[1*], Minoru Kuribayashi[2] and Arthur Méreur[1]

[1] *Troyes University of Technology, Troyes, France*
[2] *Center for Data-driven Science and Artificial Intelligence, Tohoku University, Sendai, Japan*

ABSTRACT

Over the past few decades, AI generative methods have advanced significantly, making it increasingly challenging to distinguish genuine photographs from AI-generated images, sometimes also referred to as deepfakes. In response, numerous deepfake detection methods and models have been developed, achieving high accuracy. However, the evaluation of these detection methods is often limited to a single dataset, which is typically created by generating multiple images using a specific deepfake generation methods and a fixed set of hyperparameters. This dataset is then randomly split into training and testing sets, but such an approach cannot take into account the variations of hyperparameters on deepfake detection performance. This paper addresses the fundamental question of source mismatch, where a model is trained on a specific deepfake generation source (including hyperparameters) and tested on a different one, highlighting the need to investigate the causes and

*Corresponding author: remi.cogranne@utt.fr. This work has been funded by the French ANR PACeS project No. ANR-21-CE39-0002, by the EIG CONCERT-Japan call to the project Detection of fake newS on SocIal MedIa pLAtfoRms "DISSIMILAR" through

Received 29 November 2024; revised 10 March 2025; accepted 01 April 2025
ISSN 2048-7703; DOI 10.1561/116.20240090

impacts of such a mismatch as well as to develop solutions to this critical issue.

Keywords: Deepfake, diffusion model, generative AI, source mismatch, detection, distribution-shift, empirical evaluation, experimental methods

1 Introduction

The recent advances in the development of artificial intelligence, and especially in the field of computer vision, have made possible the creation of highly realistic images and videos that can deceive naked eyes, even when carefully inspecting such media. In addition, such tools for audio, image or video generation or modifications have also been made available to a very wide audience without specific knowledge [3]. With generative-AI powered tools readily available, it is possible, for instance, to use the latest text-to-image and image-to-image models to generate photorealistic images, or modify an existing photograph. The results are almost indistinguishable from their real-world counterparts, which raised fundamental concerns about their potential misuse, especially to spread disinformation. This concern is now widely shared by political decision-makers, following public reports such as the one from the National Science and Technology Council (NSTC) [48]. This awareness among the general public and policymakers is partly explained because, as noted in [22]: "innovations in AI have enabled foreign influence actors to produce seemingly authentic and tailored messaging [...] In fact, we have already seen generative AI being used in the context of foreign elections."

In response to this growing threat of the use of so-called deepfakes,[1] researchers have been actively developing detection methods and models. These methods aim at distinguishing between genuine photographs from AI-generated content. Many such detection methods already achieve very high accuracy [40, 39, 42]. However, the evaluation of these detection methods is often limited to a specific dataset, which is typically created by generating multiple images using a specific deepfake generation method and a fixed set of

grants JPMJSC20C3 (Japan Science and Technology Agency) and by the JSPS KAKENHI (22K19777). Part of the work presented in this paper was conducted while R. Cogranne and A. Mallet were invited at Tohoku University, by Prof. Kuribayashi.

[1]The term "*deepfakes*" refers to a media (sound/image/video) modified or generated by deep learning methods and while the term "fake" can imply a malicious purpose the terms is nowadays used in a broader context. For simplificy, in the present paper we will use the term deepfake for content modified or created by generative AI models regardless of the aim of its creator. However, the work presented in this paper focus on text-to-image diffusion-based generative AI models.

hyperparameters. This dataset is then randomly split into training and testing sets. While such an approach has merits, for instance, it is reproducible and allows a fair comparison; it also has some fundamental limitations. First, such a practical assessment assumes that the generative AI models are known. Second, it also makes comparisons difficult because, even for the same generative AI models, the dataset used in one study can be generated using specific hyperparameter values, which are often not clearly precise, making a global benchmarking process almost impossible. More problematic, this approach hides the possible impact of intra-model hyperparameter variation on deepfake detection performance and, in fact, does not allow studying this effect. Furthermore, a vast majority of deepfake detection methods are based on advanced deep learning models. While this approach is generally very efficient, it generally lacks explainability and interpretability [50], and it depends very much on the training dataset, which, even in the case of deepfake detection, may be biased, as pointed out in [59]. Additionally, deep learning methods are generally facing issues with overfitting of the training data. This problem has already been identified in [27] which explored the related challenge of deepfake attribution as a simple yet effective tool to improve deepfake detection. As pointed in [30], this tends to increase reproducibility issues and at least partially explains why, generally, performances under laboratory conditions do not reflect the actual accuracy observed by forensic practitioners in real-life applications.

This *source mismatch* problem was identified in machine learning and is referred to as covariance shift [25] dataset shift [44] or, more generally, distribution shift [17]. However, the aforementioned elements show that in media forensic analysis one deals with a slightly different problem of both cross-model generalization as well as intra-model variations which makes the definition of classes uneasy.

This limitation is closely related to the so-called problem of *Cover source mismatch* (CSM) in hidden information detection [34]. While it has been long recognized as a major barrier for moving hidden information detection from laboratories into the real world [29], only recently the origin of the CSM has been clearly identified and comprehensively evaluated [18, 19]. Deepfake content identification and hidden information detection also share the common characteristics that it is aimed at detecting a rather very weak signal in a complex media. It is known that, in such a context, a small change in the dataset can lead to a very important loss in terms of detection accuracy or can simply make the evaluations of detection algorithms irrelevant. In steganalysis, source mismatch occurs when the training set is created using a specific steganographic algorithm and a very specific *source of images*, which is characterized by acquisition settings, such as the camera model and the ISO exposure/sensitivity, and the image processing process, including demosaicking, denoising, sharpening, etc. Generally speaking, the problem of hidden

information is much easier when the *source of media* is known and hence the training and the testing set are from the same source. On the opposite, it is hardly possible to detect advanced data hiding methods when *image sources* are unknown, in which case the training set is necessary from a different source than the (unknown) one used in the testing set.

Similarly, in the context of deepfake detection, source mismatch can occur when a model is trained on a specific deepfake generation source (including hyperparameters) and tested on a different one. As discussed in Section 2 generalization over different generation model has been pointed out in the literature. However, intra-model hyperparameter effects on detection performance has been seldom studied in the field of deepfake detection. This is even more surprising considering that, in a real-world scenario, it is hardly possible to assume that the potential source of deepfake (the generative AI models and all its hyperparameters) is perfectly known. While assuming that the source is known can be justified in the field of hidden information by Kerchoffs' principle [43] that one wants to evaluate the security of data hiding algorithms, such an analogy does not hold in deepfake detection. As a consequence, the study of causes, impacts and potential solutions to the problem of source mismatch in deepfake detection is urgently needed.

This paper aims to address this knowledge gap by investigating the problem of source mismatch in deepfake detection. The main contribution of our study is to propose a thorough and systematic assessment of all possible causes of source mismatch on deepfake detection. As a second contribution, we present a few solutions to prevent the source mismatch and evaluate their relevance throughout numerical experiments. Our study is based on the largest dataset collected for deepfake detection, which, upon acceptance of the paper, we will make available to the scientific community. We strongly believe that our work provides a solid and comprehensive study of the source mismatch problem in deepfake detection and will contribute to further development in this direction. Last but not least, the present paper focuses on digital images and diffusion-based generative AI models but the issues we study also hold true for other types of media, such as video and audio deepfakes.
However, we shall also acknowledge that the present paper focus on the state-of-the-art diffusion-based text-to-image generative AI model all that the findings presented in this paper may not generalize to other forms of deepfakes.

The present paper is organized as follows. First, Section 2 briefly surveys the state of the art on deepfake detection and presents the cover-source mismatch in steganalysis. We especially point out the limitations of prior works in terms of the lack of consideration for the source mismatch problem in deepfake detection. This section also proposes an introductory example to emphasize the potential impact of the source mismatch problem in deepfake detection and the needs for the present study. Section 3 briefly recalls the principle of diffusion-based generative-AI models in order to explain the role of the

hyperparameters included in the present analysis of the causes of the source mismatch problems in deepfake detection. Section 4 presents the setup of the methodology used in the present empirical study. We detail the list of generative AI models used, the different hyperparameters considered in this study as well as the deepfake detection approaches used. The numerical results are presented and discussed in Section 5. In this section the impact on deepfake detection performance of each parameter of a diffusion-based text-to-image generative AI model is empirically analysed. Each subsection of Section 5 focus on one specific component or parameters of generative AI models. Then, Section 6 presents three solutions to mitigate the source-mismatch problem and improve the detection performance in the cases of potentially multiple different sources of deepfakes. Finally, Section 7 concludes the paper by summarizing the lesson of the present study and drawing potential future research directions.

2 State-of-the-art and Position of the Present Paper

The existing literature on deepfake detection has seen significant advancements in recent years.

The paper presented by Rössler [47] marked one of the first milestones: it presents a large dataset of collected deepfake face images and evaluated several methods from the image forensics community, including models for steganography signal detection and CNN-based face swapping and replacement methods. This paper showed that simple classifiers could detect deepfakes generated by the same model. Similarly, it has been shown in [38] that simple classifiers can detect images created by an image translation network [26],
Alternatively, DeepfakeHop [6, 5] proposed an improved facial landmark detection method using an effective unsupervised feature selection method based on DFT for the detection of upsampling artefacts. However, their study was limited to examples of GAN-based detectors directly available from reference codes.

A majority of prior works on deepfake detection rely on deep learning methods; Mandelli *et al.* [36], for instance, proposed an ensemble of five "orthogonally trained" EfficientNet-B4 networks, each trained on different datasets that include content, postprocessing, and GAN-based generative-AI images. The authors developed a patch aggregation strategy that classifies an image as a deepfake if at least one of the orthogonal classifiers labels it as such. However, it is worth noting that their study only examined the robustness of their detector to JPEG compression, which was later addressed in a follow-up paper [37].

Other notable studies include PatchForensics [4], which developed a fully convolutional classifier based on local patches with limited receptive fields

over an XceptionNet backbone. Liu *et al.* [2] proposed a detector that exploits the inconsistency between real and fake images in the representations of learned noise patterns, combining spatial and frequency information to improve classification. While interesting, this work typically does not consider the generalizations of the detection and presents results over similar training and testing datasets.

LGrad [58] worked on extracting gradients through a pre-trained CNN model to filter out image content and transform a data-dependent problem into a transformation-model dependent problem. Their study focused on explainable detection using a gradient Class Activation Map (gradCAM), but did not address the source mismatch problem.

Ojha *et al.* [41] proposed a simple classifier working on pre-trained Contrastive LanguageImage Pre-training (CLIP) features, trained on a large dataset of real and synthetic images. Interestingly, their study showed that most existing methods for detecting fake images from generative models are ineffective against newer breeds of models, and that the resulting classifiers are biased towards detecting fake patterns rather than distinguishing between real and fake images.

While these examples show promising results for deepfake detection, they did not consider the transferability of the detection across different generative AI models, database or image content: their generalization has not been studied. However, the problem of generalization has long been identified as a main issue for the detection of deepfakes.

One of the first work pointing out this problem was carried out by Cozzolino *et al.* in [13]; they found that forensics classifiers transferred poorly between models. However, the authors also proposed a new representation learning method, based on autoencoders, to improve transfer performance between different generative models.

Similarly, it has been shown in [65] that classifiers often generalize poorly between GAN models. To address this issue, it has been proposed an empirical method called AutoGAN for simulating upsampling artefacts that are commonly found in GAN-based generated images. The authors tested the resulting detection technique on two types of GANs, which limits the study of transferability.

Wang *et al.* [60] introduced a Convolutional Neural Network (CNN) detector for identification of deepfake images based on ResNet50, which has become a reference point in the research community. Their work also introduced a large dataset of images generated by various GAN-based generative AI models (often referred to as LSUN/ProGAN datasets) that has been extensively adopted for model training in subsequent studies. Interestingly, their work also addresses the problem of transferability across different generative models and they proposed a data-augmentation technique to improve the generalization of the proposed deep-learning based detector. While this work marks an

important milestone for the study of deepfake detection transferability, it only considers generative models as a whole and ignores all other aspects, such as cross-database transfer, image content and tuning of the generative models, which typically limit significantly the study of the source mismatch problem in deepfake detection.

Building upon this work, Gragnaniello *et al.* [20] proposed a simple modification to the ResNet50 architecture to preserve better low-level forensic traces. Their approach was also trained and tested on the same dataset used in Wang *et al.*'s work, showcasing the importance of using the same dataset for model evaluation.

Corvi *et al.* [11] performed strong augmentation to gain robustness and increase generalization, training their detector on a large dataset of latent diffusion models. Their study included the detection of "image fusion" by averaging the outputs of AI-generative networks, but this was a rather limited exploration of the source mismatch problem.

Last but not least, NPR [57] worked on residual images computed as the difference between the original image and its interpolated version. Their classifier was trained on only 4 classes of the ProGAN dataset and tested on the merging of 5 datasets encompassing 28 generative-AI models, which is an interesting study on source mismatch. However, their paper only presented results on the testing dataset with limited interpretation of the impact and solution for the source mismatch problem.

From the previous brief review of the literature, it seems clear that while the problem of generalization across different generative AI models has long been recognized, it has seldom been studied and generally focused on cross-model transferability. In addition a vast majority of the works that study the problem of generalization in deepfake detection only focus on practical solutions.

Amongst the additional examples, Epstein *et al.* [16] collect a dataset generated by 14 well-known diffusion models and simulates a real-world learning setting with incremental data from new diffusion models. They find that the classifiers generalize to unseen models, while also observing important loss of performance when the diffusion-model architecture was considered very different.

With a different approach, Ojha & al. [41] and Cozzolino & al. [12] exploits pre-trained models, respectively Vision Transformers (ViT) and Contrastive Language-Image Pre-training (CLIP), to achieve high generalization across different diffusion models. Following the example of Wang *et al.* [60] it has more recently also been proposed in Yan *et al.* [62] to address the problem of generalization using data augmentation approaches to improve detection generalization.

Last but not least, Dogoulis *et al.* [15] addresses the generalization of detectors in cross-concept scenarios (e.g. when training a detector on human faces

and testing over synthetic animal images). They propose an original strategy to address this generalization issue with a resampling strategy that considers image quality scoring for sampling training data to learn a specific and dedicated classifier, which shows superior performance compared to random sampling.

Interestingly, one shall also note that quite a lot of datasets of deepfake media have also been proposed. However these datasets also generally consider one generative AI model as a whole and, generally, do not provide any details on how the images in the dataset were generated. For instance the paper from Wang *et al.* [60] was associated with a very large dataset of images generated by GAN-based models. However, the paper does not explain precisely how those images were created and how exactly the models were trained, which limits the study of the problem of source transferability in deepfake detection.

More recently, Yan *et al.* [63] proposed an extensive database of 15 state-of-the-art detectors and gathered 9 deepfake datasets with the goal of addressing the problem of the lack of a standardized, unified, comprehensive benchmark in deepfake detection. This work also standardized the evaluation metrics and uniformized the data processing pipelines. While this important work facilitates benchmarking, it does not address specifically the difficult issue of generalization even though the dataset can be used for this purpose.

The work proposed by Li *et al.* [31] addressed the problem of generalization and source mismatch by proposing a so-called continual deepfake detection benchmark (CDDB). Their goal is to address the problem of generalization across different datasets and generative models by proposing a smooth transition between known models to datasets made with unknown generative AI models.

All these examples show that the generalization problem is a well-recognized and difficult problem in deepfake detection. However, the previous brief review of the current art also point out that this problem is not generally considered and when studied, it is with two main limitations: (1) considering generalization cross various generative AI models without and (2) assuming that a generative AI model always produces images of the same kind, ignoring all possible sources of discrepancy that causes a source mismatch. On the opposite, the present paper considers that a diffusion-based generator does not always generate homogenous data and that the detection of the resulting generated image largely depends on the hyperparameters used to generate the media.

From the brief overview of the current art, it is clear that there is an urgent need for a study that analyses how a generative-AI model is set and the impact this setting may have on the ensuing detection of deepfake media it generates.

2.1 An Introductory Example

In order to highlight the needs for a study on the cause and the assessment of source mismatches in deepfake detection, we have conducted a small-scale introductory experience. We used the detection method as proposed in [11][2] as it is one of the recent state-of-the-art reference approaches for the detection of images generated with diffusion-based generative AI. The results are reported in Table 1. First we evaluated the detection accuracy of this method over the five diffusion-based text-to-image generative models used in the present paper, that is, Stable Diffusion XL Turbo, Stable Diffusion 3 medium, Wuerstchen, Kandinsky 2.2 and FLUX.1 Schnell, see details in the Table 2. For comparison, the first row of Table 1 also shows the "true-negative rate" over the ALASKA [10, 9] dataset of real images. Then the next row shows the results of each of the text-to-image diffusion-based generative models used in the present paper along with the accuracy when one single part of the generative model is changed. The deepfake detection accuracy as reported in Section 1 corresponds to the "true-positive rate", also referred to as the recall or the sensitivity. Unsurprisingly, it is obvious from Table 1 that the detection method proposed in [11] achieves overall excellent results except for FLUX.1 Schnell that was released after the detection method, hence constituting a typical case of generalization cross generative AI models. While this also holds true for Stable Diffusion 3, the detection remains much higher in this latter case. However, one can note that when the generative AI model is changed, the detection accuracy can drop drastically. This is especially true when replacing the Variational AutoEncoder (VAE) from Stable Diffusion XL Turbo and when changing significantly the number of diffusion steps for Stable Diffusion 3 and for Kandinsky 2.2. Surprisingly, modifying Wuerstchen by replacing the noise variance scheduler of the diffusion process actually helps the detector by a rather important margin.

It is difficult from this single experience to draw general conclusions on the impact of each hyperparameter of the diffusion-based generative model of the detection of the resulting image. However, the present introductory example clearly confirms that, on the one hand, the setting of the generator greatly influences the ensuing problem of detection of generated images, hence showcasing that the generalization across different generative AI models is not always sufficient. On the other hand, this example also emphasizes the urgent need for a thorough study to assess the problem of source mismatches in deepfake detection as well as to study the origin of such mismatches.

[2]The source-code of the detection method proposed in [11] is available on GitHut on the page of Naples' University Image Processing Research Group (GRIP).

Table 1: Accuracy of the diffusion generative AI model proposed in [11]; the quantity is the Truth Negative rate for the ALASKA2 dataset [10, 9] of genuine photographs and true positive rates for all other generative-AI image datasets.

Source	Accuracy
ALASKA2 real photographs	97.80%
Stable Diffusion XL Turbo	85.01%
—— with Consistency Decoder VAE	7.79%
Stable Diffusion v3	96.07%
—— with 2 diffusion steps	16.93%
Wuerstchen	85.01%
—— with DDIM noise scheduler	96.20%
Kandinsky v2.2	93.85%
—— with 64 diffusion steps	7.79%
FLUX.1 Schnell	53.81%
—— with Finetuning #1	63.78%
—— with Finetuning #2	47.76%

Table 2: List of text-to-image generative AI based on diffusion used in the present paper along with their source.

Name	URL / Source
Diffusion Transformer models	
FLUX.1 schnell	HuggingFace
Stable Diffusion 3	HuggingFace
Diffusion models	
Kandinsky v2.2	HuggingFace
Stable Diffusion XL-Turbo	HuggingFace
Wuerstchen	HuggingFace
Finetuning from Stable XL Turbo	
lcm-LoRa SDXL Turbo	HuggingFace
Fire Generation	HuggingFace
DreamBooth	HuggingFace
SDXL Turbo DPO LoRA	HuggingFace
Finetuning from FLUX.1 Schnell	
Miniature people	HuggingFace
Studiopellosh	HuggingFace
flux-schnell-realism	HuggingFace
flux-schnell-lora	HuggingFace

3 Primers on AI-generative Method in Image

Because the present paper studies the impact on deepfake detection of each and every part of state-of-the-art generative AI models, we shall first start by recalling briefly how generative AI models work. To this end, the present section reviews recent advances in AI-generative image models, with a focus on state-of-the-art diffusion-based models.

The timeline of representative studies for text-to-image generation begins with AlignDRAW, released in 2015, which is often considered the first text-to-image generative AI. AlignDRAW extended the DRAW architecture by conditioning it on text sequences, achieving a significant breakthrough in creating images from text prompts, albeit with limited realism and diversity.

In 2016, text-conditional GAN emerged as the first fully end-to-end differential architecture [46], extending from character-level input to pixel-level output. However, GANs produced high-quality images only when trained on small, domain-specific datasets, such as faces [67].

The subsequent development of autoregressive models, such as DALL-E in 2021, captured widespread attention due to their ability to generate complex and realistic images from text prompts, thanks to large-scale training. These models [45, 14] are typically based on a discrete Variational Autoencoder (VAE), an autoregressive decoder-only Transformer, and a CLIP pair of image encoder and text encoder. However, their autoregressive nature leads to high computation costs and sequential error accumulation.

Inspired by non-equilibrium thermodynamics, diffusion models have emerged as the leading method in text-to-image generation. Diffusion Probabilistic Models (DPM) emerged in 2015 [54]; it is defined in [23] as a parameterized Markov chain trained using variational inference to produce samples matching the data after a finite time. The core of this algorithm involves a forward process that converts a complex data distribution into a simpler one, and then learns the mapping by reversing the diffusion process. Then, score-based generative models (SGM) propose perturbing data with different levels of Gaussian noise and jointly estimating the corresponding scores [56]. SGM generates samples towards decreasing noise levels and trains the model by estimating score functions for noisy data distributions.

The combination of SGM and DPM led to the development of Denoising Diffusion Probabilistic Models (DDPMs), a class of Markov chain-based models that generate images from noise through a finite sequence of transformations [23]. During training, the model learns to reverse the process of adding noise to natural images by estimating the noise that was added at each step [64]. This process is illustrated in Figure 1.

As shown in Figure 1 one shall distinguish the joint distribution $p_\theta(\mathbf{x}_{0:T})$, so-called the *reverse process*, from the distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1})$, referred to as the *forward process*. What distinguishes diffusion models from other types of

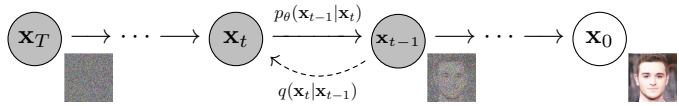

Figure 1: Illustration from [23] of the diffusion process and conditional distribution probability learning during the inference step.

latent variable models is that the posterior $q(\mathbf{x}_T|\mathbf{x}_0)$, the *forward process* or *diffusion process*, is approximated as a Markov chain, which iteratively adds Gaussian noise, according to a variance schedule $\beta_1, \ldots, \beta_T$, until the content of the input data $\mathbf{x}_0$ is completely obliterated :

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}).$$

Here $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are latents of the same dimensionality as the data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$.

The entire diffusion process then satisfies:

$$q(\mathbf{x}_T|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}).$$

The training is carried out by optimizing the log-likelihood ratio between distributions p and q :

$$-\sum_{t\geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}.$$

The *reverse process*, or the *backward diffusion*, aims at generating a possible data $p_\theta(\mathbf{x}_0)$ starting from a noisy input drawn from $p_\theta(\mathbf{x}_T)$ thank to the outcome of the training, which learn approximating the reverse process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$.

Early experiments with Generative Adversarial Models (GANs) have found that adding a label to the generated images can significantly help create more realistic images as well as improve their quality. This technique, referred to as "classifier guidance", works by combining the model's score with the feedback from an image classifier. However, this approach requires training an additional image classifier, whose availability can be an issue. Indeed, in practice, it's not always easy to find a suitable image classifier, especially since the model is trained on noisy data, which prevents the use of a standard pre-trained classifier. Fortunately, research has shown that it's possible to achieve similar results without an auxiliary classifier [24]. This novel approach, called classifier-free guidance, uses only the generative model itself to guide the image creation process, eliminating the need for an extra classifier; hence classifier-free guidance can be thought of as classifier guidance without a classifier, as pointed out in [24]

With this short and rather general description of the diffusion-based model for generative AI we can have a clear idea of the main components of such a model and the hyperparameter whose impact on deepfake detection shall be studied. As shown in the Figure 2, the first and foremost is related to the prompt, which is the text input, its tokenization and its embeddings. Then comes the guidance scale and the number of diffusion steps. The scheduler of the noise variance used during each diffusion step can also be changed to analyse its impact. Last but not least, the Variational AutoEncoder (VAE) which is the ultimate step that "decodes" the latent into a rendered image as shown in the Figure 2.

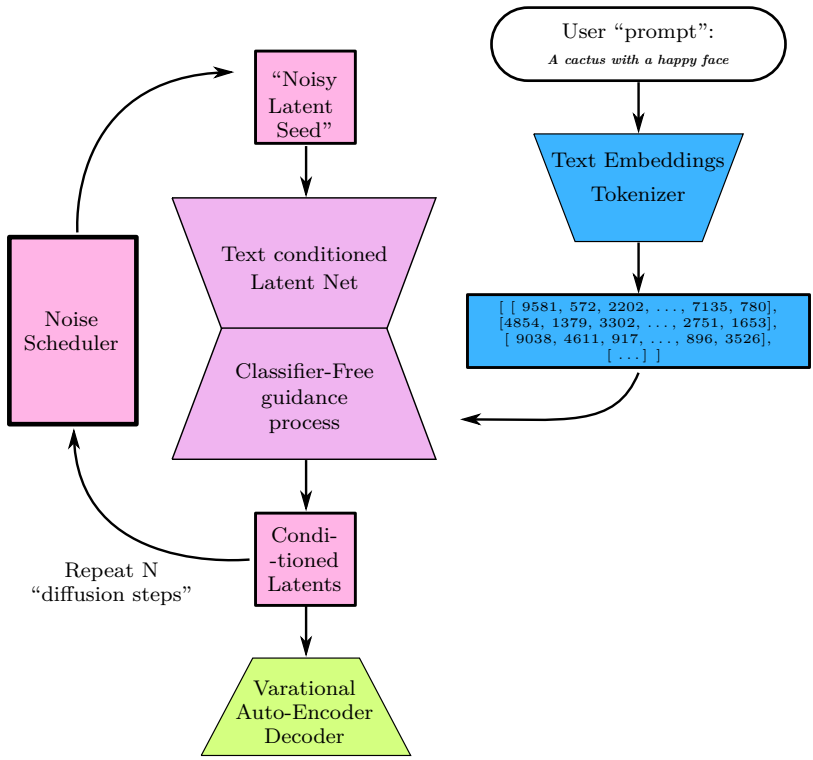

Figure 2: Illustration of a typical diffusion-model-based generative AI model highlighting the elements studied in the present paper.

In addition to these hyperparameters of diffusion-based generative text-to-image we have also studied the impact of the more global "finetuning" which consists in retraining a model over a specific dataset with the goal of generating specific content. This, together with the overall impact of the

generative model, is studied before we studied in more detail the impact of the hyperparameters.

While the description of diffusion model-based generative image approach, the main parameters which can have an impact on the detection of deepfake can be clearly identified. On one the text tokenizer and embedding method shall be analyzed. Second the diffusion process is generally defined several parameters such as the noise scheduler, the number of diffusion steps and the guidance parameter. Last but not least the last step made of autoencoder can also have an impact on deepfake detection performance. This set of parameters largely defines a generative model and hence their impact shall be studied systematically to assess the problem intra-model cover source mismatch.

4 Definitions and Methodology for the Study of Source-mismatch Problem

As exposed in Section 2.1, the motivation of our study is based upon the observation that even small changes in the generative AI models can dramatically impact the ensuing problem of detection of generated images, so-called deepfakes. However, before analysing, in more detail, the impact of the source mismatch, we need to define what a source is, what exactly the problem of the source mismatch problem is and to measure its importance.

4.1 Definition of Concepts

First of all, even though it may seem obvious, we must start by defining the core element studied in the present paper: *deepfakes*.
As its name implies, the term "*deepfake*" is derived from the combination of "deep", referring to deep learning, and "fake".

Definition 1 (of deepfake)**.** *The term deefake generally refers to the manipulation of existing media (image, video and/or audio) or generation of new (synthetic) media using deep learning-based approaches [1].*

Despite the word "fake", which indicates a will to use manipulated or synthesized media for the purpose of disinformation, there exist quite a lot of applications in which the creator is not malicious, such as, for instance, for entertainment, illustration and arts. With this respect, the term "deep synthesis" has been proposed as a more neutral alternative [32]. This new term, however, has not been widely adopted and in the present paper, the term deepfake is mostly used to refer to text-to-image synthesis using diffusion-based methods.

For the subsequent definitions, we will mostly rely on the prior works [34, 18, 19] that review and study the related problem of cover-source mismatch

for the detection of information hidden within digital media, also referred to as steganalysis.

Based on these prior works, we can use the following definition for the source of generated images.

Definition 2 (of a Source of generated images)**.** *A* source *of images generated with generative AI models can be defined as a model, its components combined with a set of hyperparameters, including the prompt, as post-generation processing operations. The images generated from a single* source *can have different content, even for the same prompt; however, they shall carry similar properties, such that they can be identified as a single homogenous set.*

Note that it is not clear what are the properties that allow identifying images as a *source*. It is clear that when using a generative AI model and that all its hyperparameters are fixed, including the prompt, all the generated images come from the same *source*. However, we are aware that our definition is certainly a bit too restrictive but the goal of the present paper is, among others, to clarify what makes a source and what hyperparameters are superficial to define a consistent *source*.

Note that in the proposed definition of a source, we also included the post-generation processing, as this may deeply influence the statistical properties of the ensuing images. For instance, two images generated by the same diffusion-based model with the same hyperparameters but compressed with different JPEG quality factors will be assumed from different sources.

Based on the work presented in [34] we propose the following definition of the source mismatch problem:

Definition 3 (source mismatch problem)**.** *The* problem of source mismatch *for the detection of deepfake content occurs when a classifier is trained over images from source A and tested over images from a different source B. More precisely, the problem of* source mismatch *is associated with the decrease in the performance of a classifier trained and tested on different sources.*

This definition of the *source mismatch problem* is much more practical, as it implies that without loss of detection performance there is no problem with *source mismatch*. This relation between the source mismatch and the practical problem it creates on the performance of the detector is fundamental in the operational context but also limits the study on the relation between the hyperparameters and the ensuing properties of generated media.

4.2 Methodology for Assessment of Source Mismatch

In order to further study the source mismatch problem in the detection of deepfake images, one needs to define how to measure the impact on the detection accuracy. To this end we will use a toy example illustrated in Table 3 with two sources of generated media, respectively sources A and B.

Table 3: An illustrative example for defining *qualitatively* the metric that assesses both "intrinsic difficulty" (when using the same source for training and testing) and "inconsistency" (difference when using different training and testing sources), here the prediction error is measured by the average error rate.

Training \ Testing	Source A	Source B
Source A	0.2	0.38
Source B	0.33	0.35

Definition 4 (Source Intrinsic Difficulty)**.** *The intrinsic difficulty of detecting content generated by source A is defined as the detection accuracy of a given classifier to identify content generated with source A when both the training and the testing set are generated with source A, hence in the absence of a source mismatch.*

In the present paper, we will not focus on the definition of the source of genuine media, that is, digital photographs in the case of images. Therefore, we assume that the classifier is trained to distinguish deepfakes from photographs using a specific set of photographs and that the intrinsic difficulty typically corresponds to the false-negative rate. Of course, the set of genuine media used to distinguish those created by the generative AI models does impact the intrinsic difficulty but this aspect falls outside the scope of the present paper, which focuses on the definition of the source of deepfake.

The problem of source mismatch in deepfake detection can be assessed by measuring the loss of performance when the training is carried out a different dataset:

Definition 5 (Source Inconsistency)**.** *The inconsistency between sources A and B appears when one trains a classifier over source A while it is later tested later over source. In this situation one will obtain a different detection accuracy as compared to one obtained when training on source B. Generally speaking, this can be seen as a lack of generalization of the classification rule.*

The toy example reported in Table 3 shows via a practical example these notions of *intrinsic difficulty* and *source inconsistency*. The row corresponds to the detection error rate obtained when training over the sources A and B. On the opposite, columns report the detection error rate when testing over sources A and B. The elements on the diagonal correspond to the cases where both the training and testing are carried out over the same source, hence the absence of a mismatch. Therefore, the detection performance reported on the diagonal corresponds to the intrinsic difficulty. On the opposite, the out-of-diagonal elements correspond to the cases in which the source mismatch occurs. While the Table 3 is a fictional illustration, note the asymmetry

of the impact of the source mismatch problem: when testing over source B, the mismatch increases the detection error rate by only 3%, compared to the intrinsic difficulty. On the opposite, when testing over source A, a mismatch creates a much higher increase of the detection error from 20%, in the matched case, hence the intrinsic difficulty, to 33% in the presence of the source mismatch problem. Such phenomena are often seen in real cases, as we shall observe in Section 5.

In practice, the present paper uses the minimal error rate under equal prior, which is defined concretely as the mean of the false positive and false-negative rates. For each experiment, the decision threshold is computed to minimize the error rate. The reason for this choice is to avoid measuring artificially high loss in terms of detection accuracy only because of a poorly chosen detection threshold. In other words, when training and testing over images from different sources, the detection threshold shall be adjusted to measure the inconsistency more precisely.

5 Numerical Results and Analysis

5.1 Common Core of All Experiences

In our experimentation we generated between 12.000 and 15.000 for each generative AI model and hyperparameter settings, among which 2.000 were used for testing and evaluation of the detection accuracy and the remaining $10,000$ images were used for training. This number was large enough to train a specific classifier for all the models we used for detection and we did not observe any convergence issues with our classifiers. We generated images of size 512×512 pixels in colour, but all the experimentations were made on grayscale images of size 128×128 otherwise must have results reported almost detection, which limits the analysis of the findings.

Note that the results presented in the present paper were obtained using EfficientNetv2 and Dual Vision Transformers (DaViT) because these are interestingly representative of deep learning architecture, CNN for EfficientNetv2 and Trasformers-based for DaViT, and because prior works show that they are amongst the most accurate [8]. The results we obtained were always consistent with those two models; hence our choice to limit the presentation to only one of those two models in a vast majority of the cases.

Regarding the classification, all models we used were obtained from the `timm` library [61] from which we loaded pretrained model over the imagined dataset, as it has been shown to greatly improve training convergence in hidden information detection. To further speed up the training, we used a curriculum training strategy with the following approach : first, the models were trained for 5 epochs on grayscale images of size 512×512 and coming

from all the *source* used in the present paper. The learning rate (LR) was divided by 2 after each epoch. The resulting pretrained models were used as a starting point for all the experiments; this second training was carried out using a decreasing learning rate following stochastic gradient descent with warm restarts (SGDR) [52, 33, 51]. Given the fact that each classifier is trained over one source of deepfake only, we used "only" 35 epochs. The initial value of the learning rate was obtained using the method initially proposed in [53]. In brief, it essentially consists of a 1-cycle training of the deep-learning method over so-called mini-batches : the learning rate is gradually increased, at each mini-batch, from a very low initial value to a final high value. Throughout this process, the loss is measured at each iteration in order to find the largest value, with a margin value before the loss begins to diverge.

The learning rate scheduler is thus based on this initial guess and then slowing decreased over one cycle to a nominal value of 10^{-5}. The initial cycle length is set to 5, it is doubled for every cycle and the initial learning rate is divided by two after each cycle. We used three cycles for a total of 35 epochs, which has been found largely enough in our numerical experimentations.

Another important factor that we noticed in the importance of data augmenting to prevent overfitting of the model and ensure a better generalization even though testing and training sets were generated in the very same manner; they are, in fact, a random split from the same dataset. This fact has also been reported in [7, 21]. In our case we carried out data augmentation by adding the following operation : mirroring, flipping along x-axis and y-axis, Gaussian i.i.d. noise addition (with standard deviation between 0.05 and 0.15), multiplicative noise addition (with factors between 0.975 and 1.025) resizing (with rescaling factor between 0.95 and 1.05) and rotation (with angles between -5 and 5 degrees). Each operation was applied independently, using the Albumentations library, and with a probability $p = 0.25$ for each. While each operation within the overall data augmentation process does not modify the image significantly, and, from a quick search, we did not find interest in applying more important operations, we have found that these processing were enough to greatly improve the testing accuracy.

5.2 Impact of Generative-AI Models

We are now ready to study the source mismatch problem in deepfake detection. The first experience we carried out is reported in Tables 4 and 5. The former has been obtained with DaViT detection model, while the latter has been obtained with Efficient Net v2. It consists in a coarse grain assessment of source mismatch between different generative AI models. First of all, one can observe the general very low detection error rate over all the generative AI models. Second, as noted in the toy example presented in Table 3, note the asymmetry in the intrinsic difficulty. This is especially obvious for the

Table 4: Inconsistency and intrinsic difficulty of different diffusion-based generative AI models. Classification error rate and obtained with Dual Vision Transformer –DaViT– model.

Training \ Testing	FLUX.1-schnell	Stable Diffusion 3	SD XL Turbo	Wuerstchen	Kandinsky 2.2
FLUX.1-schnell	1.43%	5.09%	16.08%	7.09%	5.28%
Stable Diffusion 3	6.88%	2.35%	11.48%	5.19%	5.64%
SD XL Turbo	16.19%	25.47%	0.93%	18.03%	17.68%
Wuerstchen	12.50%	14.41%	10.41%	0.98%	13.11%
Kandinsky 2.2	10.70%	10.64%	19.64%	8.79%	1.78%

Table 5: Inconsistency and intrinsic difficulty of different diffusion-based generative AI models. Classification error rate and obtained with EfficientNetv2 small.

Training \ Testing	FLUX.1-schnell	Stable Diffusion 3	SD XL Turbo	Wuerstchen	Kandinsky 2.2
FLUX.1-schnell	0.91%	4.83%	23.18%	9.09%	5.43%
Stable Diffusion 3	5.63%	1.51%	16.69%	4.60%	6.27%
SD XL Turbo	18.73%	31.76%	1.29%	23.30%	24.05%
Wuerstchen	16.12%	18.87%	22.21%	0.61%	16.69%
Kandinsky 2.2	9.13%	9.57%	21.78%	6.40%	1.26%

Wuerstchen generative model. Indeed, training on this model creates a very high inconsistency when testing on all other models. In comparison, testing on Wuerstchen causes a much smaller source mismatch problem. Also, note the specificity of Stable Diffusion XL Turbo, which is always the cause of an important inconsistency.

Overall, the results presented in Tables 4 and 5 show a generally important source mismatch problem when switching from one generative model to another with rather important inconsistency as compared to the very low intrinsic difficulties. While this result is not very much surprising, as it is in the line of the observation presented in prior works, see for instance [38, 65, 16], it will serve as a basis for the assessment of the source-mismatch problem.

5.3 Impact of Finetuning

In direct line with the first experiment, the second set of results we present concerns the impact of the fine-tuning of a generative AI model. To this end, Table 6 shows the intrinsic difficulty and the inconsistencies between FLUX.1 schnell and four of the many available fine-tuning versions that are available on HuggingFaces website. Surprisingly, one can note a relatively low inconsistency, while we carefully selected fine-tuning that generates photograph-like images of very different kinds (see Table 2 and the reference therein). We observed similar results with fine-tuning versions of Stable Diffusion XL Turbo.

Table 6: Inconsistency and intrinsic difficulty of different fine-tuning versions of the same generative AI, namely FLUX.1-schnell. Classification error rate and obtained with Dual Vision Transformer –DaViT– model.

Training \ Testing	FLUX.1-schnell	Finetuning #1	Finetuning #2	Finetuning #3	Finetuning #4
FLUX.1-schnell	0.91%	4.44%	3.42%	2.23%	2.52%
Finetuning #1	2.79%	1.35%	2.21%	2.97%	3.37%
Finetuning #2	2.43%	2.18%	0.93%	2.90%	3.18%
Finetuning #3	2.61%	2.80%	3.75%	0.98%	1.81%
Finetuning #4	3.12%	3.51%	2.72%	1.63%	0.78%

From these results we can conclude that fine-tuning does have an impact on the source mismatch problem but a rather limited one.

5.4 Impact of Prompts and Text Embeddings

The first step of all generative AI models is to encode the text input by the user, so called the prompt. How the prompt is turned into embedding is the first component whose impact on the source mismatch problem can be studied. However, it is difficult to modify slightly this process and, therefore, we propose to change this part with different elements. In the results provided in Table 7 we contrast the detection error obtained when training on Wuerstchen and the inconsistency with several variations of the text to embeddings process. We first tried generating images with the same prompt to measure the impact of different prompts. Then we propose using OpenAI's CLIP tokenizer.[3] We replaced the original T5 encoder with T5-xxl which is supposedly close to the original one. Eventually, we replace the text encoder with "CLIPTextModelWithProjection" as used in Stable Diffusion 3. Note

[3]CLIPtokenizer is avaiable on HugginFace Hub.

Table 7: Inconsistency and intrinsic difficulty of different versions of text embedding for Wuerstchen. Classification error rate and obtained with Dual Vision Transformer –DaViT– model.

Training \ Testing	Wuerstchen	Single prompt	with CLIP tokenizer	with T5xxl	CLIP model with projection
Wuerstchen	0.98%	0.98%	1.14%	1.35%	1.06%

that we have tried different experiments, but the results presented in Table 7 reflect the general results we obtained.

Unsurprisingly, the change of the prompt and its encoder have almost no impact on the source mismatch problem. This is not a surprise because this step is the very first and it is expected that changes at the very beginning of the generative process have less influence on the image produced in the end.

5.5 Impact of Guidance Scale and Number of Diffusion Steps

The main step in any diffusion-based generative model is the diffustion itself. As explained in Section 3, during inference there is a very small number of hyperparameters that can be changed. Tables 8 and 9 present the inconsistency created when changing the number of diffusion steps and the guidance scale. In both cases the classifier was trained in the "median case", that is, a guidance scale of 3 and 12 diffusion steps for Stable Diffusion 3 and a guidance scale of 1.5 and 8 diffusion steps for Stable Diffusion XL Turbo.

Table 8 seems to point out that these two hyperparameters have a limited impact on the ensuing mismatch problem even in extreme cases.

On the opposite, Table 8 shows that the number of diffusion steps can have an important impact on the source mismatch problem. We would like to acknowledge that in our numerical results this conclusion only holds for a Stable Diffusion XL Turbo. For all the other generative models, the guidance scale and the number of diffusion jointly have a very limited impact on the source mismatch problem, even for FLUX.1 schnell, which is also able to generate images with a very small number of diffusion steps. We can only conclude from these divergent results that the hyperparameters guidance scale and number of diffusion steps can have a significant impact, although this does not seem to hold true in the majority of cases in our experiments.

Table 8: Inconsistency and intrinsic difficulty for different guidance scales and number of diffusion steps for Stable Diffusion 3. Classification error rate and obtained with Dual Vision Transformer –DaViT– model.

Guidance \ Diffusion steps	2	4	8	12	16	32	64
0.5	3.49%	2.55%	2.38%	2.32%	2.49%	2.53%	3.33%
1	3.40%	2.48%	2.34%	2.31%	2.44%	2.48%	3.27%
2	3.33%	2.44%	2.31%	2.34%	2.41%	2.50%	3.22%
3	3.31%	2.44%	2.36%	2.43%	2.57%	3.27%	3.92%
4	3.42%	2.76%	2.57%	2.60%	2.65%	2.69%	4.12%
8	3.88%	3.18%	3.01%	2.98%	2.99%	3.03%	4.64%
16	4.55%	3.91%	3.69%	3.75%	3.80%	3.77%	4.97%

Table 9: Inconsistency and intrinsic difficulty for different guidance scales and number of diffusion steps for Stable Diffusion XL Turbo. Classification error rate and obtained with Efficient Net v2.

Guidance \ Diffusion steps	1	2	4	8	12	16	32
0	13.37%	13.14%	6.93%	2.33%	1.73%	1.37%	1.83%
0.5	12.84%	12.67%	6.84%	2.31%	1.80%	1.31%	1.85%
1	11.90%	11.46%	6.30%	2.32%	1.86%	1.30%	1.76%
1.5	11.61%	11.25%	6.02%	2.40%	1.87%	1.29%	1.85%
2	11.21%	11.15%	6.03%	2.36%	1.85%	1.26%	1.77%
3	11.34%	11.20%	6.13%	2.33%	1.89%	1.34%	1.76%
4	11.45%	11.21%	6.16%	2.36%	1.86%	1.37%	1.88%

5.6 Impact of Noise Scheduler

The other main hyperparameter of the diffusion process is the scheduler of the noise variance. The results presented in the Tables 10 and 11 show the impact on the source mismatch problem when changing the noise variance scheduler algorithm. While it was shown in the previous Section 5.5 that the guidance scale can heavily impact the source mismatch, it seems that the noise variance scheduler has an overall limited impact. Interestingly, the case that creates the most mismatch problem is when changing the scheduler from Stable Diffusion XL Turbo to the one from Stable Diffusion 3 and vice versa. On the opposite, the other scheduler for the noise variance, name https://github.com/Jordach/comfy-consistency-vae, the Denoising Diffusion Implicit Models (DDIM) scheduler [55], Euler Scheduler [28] and the Trajectory Consistency Distillation (TCD) scheduler [66] all have a very limited effect on the mismatch problem.

Table 10: Inconsistency due to changes in the noise scheduler for Stable Diffusion 3 with Dual Vision Transformer –DaViT– model.

Training \ Testing	Stable Diffusion 3	DDIM scheduler	Euler scheduler	TCD scheduler	Scheduler from Kandinsky 2.2	Scheduler from Wuerstchen	Scheduler from Stable XL Turbo
Stable Diffusion 3	2.35%	5.87%	2.78%	3.13%	3.20%	2.35%	2.77%

Table 11: Inconsistency due to changes in the noise scheduler for Stable Diffusion XL Turbo with Dual Vision Transformer –DaViT– model.

Training \ Testing	Stable Diffusion XL Turbo	DDIM scheduler	Euler scheduler	TCD scheduler	Scheduler from Kandinsky 2.2	Scheduler from Wuerstchen	Scheduler from Stable 3
Stable XL Turbo	0.93%	5.80%	1.10%	3.81%	1.06%	0.81%	4.11%

5.7 Impact of Variational Autoencoder (VAE)

The very last element whose impact on the source mismatch problem is studied is the variational autoencoder (VAE) that is used at the end of the diffusion process to turn the latent into a digital image. Surprisingly, Table 12 shows that this last element has a limited effect on the source mismatch problem, while it is expected that the ultimate step to have the greatest impact. However, this is not always the case and Table 13, on the opposite, tends to point out that this element can have a rather important impact on the source mismatch problem. Here again it seems difficult to draw a general conclusion: we can only state that the impact of the VAE can be major is in some cases.

Table 12: Inconsistency due to changes in the Variational Autoencoder (VAE) for the Kandinsky diffusion model and with EfficientNet v2small model.

Training \ Testing	Kandinsky 2.2	Autoencoder with KL Loss (AKL) 1.5	Autoencoder with KL Loss (AKL) 2	Consistency Decoder	VAE from Pixart-α	VAE from SD 2.1	VAE from SD Turbo	VAE from SD XL Turbo	SD1.4 EMA fine-tuned VAE	SD1.4 MSE fine-tuned VAE	SDXL fine-tuned VAE	Tiny AutoEncoder for SD
Kandinsky 2.2	1.26%	1.41%	1.40%	1.33%	1.39%	1.42%	1.37%	1.37%	1.40%	1.33%	1.43%	1.44%

Table 13: Inconsistency due to changes in the Variational Autoencoder (VAE) for Stable Diffusion XL Turbo model and with EfficientNet v2small model.

Training \ Testing	Stable Diffusion XL Turbo	Autoencoder with KL Loss (AKL) 1.5	Autoencoder with KL Loss (AKL) 2	Consistency Decoder	VAE from Kandinsky 2.2	VAE from Pixart-α	VAE from SD 2.1	VAE from SD Turbo	SD1.4 EMA fine-tuned VAE	SD1.4 MSE fine-tuned VAE	SDXL fine-tuned VAE	Tiny AutoEncoder for SD
Stable Diffusion XL Turbo	1.29%	12.91%	16.84%	24.17%	13.44%	17.84%	13.42%	13.47%	17.65%	13.32%	2.14%	21.08%

5.8 Impact of Post-generation Image Processing Operations

Last but not least, we wanted to highlight the impact of the post-generation processing operation on the source mismatch problem. Indeed, images are often processed before being sent; for instance, they can be compressed or resized when posted on social networks. We have identified 10 common image processing operations and evaluated the inconsistencies they create. Those are JPEG compression with different standard quality factors, Upsampling and Downsampling using Lanczos resampling kernels, denoising, sharpening and a combination of denoising then sharpening. The results reported in the Table 14 show that, unsurprisingly, the post-processing operation can create a major source mismatch problem with inconsistency that can be higher than 20%. This is rather consistent with prior works, see for instance [60, 36, 37], and rather expected, since, as already explained, the last operations are expected to have a larger impact on the inconsistency between sources.

Table 14: Inconsistency and intrinsic difficulty of different diffusion-based generative AI models. Classification error rate and obtained with Dual Vision Transformer –DaViT– model.

Training \ Testing	Kandinsky 2.2	JPEG Compression 90	JPEG Compression 80	JPEG Compression 70	JPEG Compression 60	Upsampling 10%	Downsampling 10%	Denoising	Sharpening	Dernoising then sharpening
Kandinsky 2.2	1.27%	16.29%	22.27%	25.59%	27.98%	4.08%	6.15%	12.39%	4.16%	16.86%

6 First Steps Towards Mitigating the Impact of Source Mismatch

Before concluding the present study, we wanted to evaluate two common strategies for mitigating the source mismatch problem in deepfake detection. The first one, referred to as "Holystic", consists in training a classifier over images generated by all possible sources. It is expected that a classifier trained over diverse datasets will have higher robustness and, at least, can generalize better over the sources that are close to those included in the training set. The second method, referred to as the "atomistic" approach, consists in training a multi-class classifier, which is used to identify the source with which the deepfake may have been the most likely generated and then apply a binary classifier trained to distinguish this specific source from genuine photographs.

Last but not least, we included a third approach proposed in [49] and the latter used in [35] which is similar to the atomistic approach. First a multi-class classifier is trained to identify the source of the inspected images. Then we used the aggregate the results from all binary classifiers but weighted them according to the "soft output" of the multi-class classifier. In other words, instead of using one binary classifier, we use them all, while giving more importance to the sources that the multi-class classifier considers to be the most likely. While several weighting functions are studied in [49], the finding of the best aggregation strategy falls outside the scope of the present paper and we simply apply a softmax function to the soft-output of the multi-class classifier and weight the output of the binary classifier accordingly. This approach is referred to as "weighted" in Tables 15 and 16.

Interestingly, these two tables show that the three strategies for mitigating the source mismatch problem are very efficient. However, it shall be acknowledged that these results are somewhat obtained by eliminating the source mismatch problem, as all sources are now included in the training set. However,

Table 15: Inconsistency and intrinsic difficulty of different diffusion-based generative AI models. Classification error rate and obtained with Dual Vision Transformer –DaViT– model.

	Kandinsky 2.2	JPEG Compression 90	JPEG Compression 80	JPEG Compression 70	JPEG Compression 60	Upsampling 10%	Downsampling 10%	Denoising	Sharpening	Dernoising then sharpening
Holystic	0.94%	2.44%	2.93%	3.01%	3.59%	1.43%	1.60%	1.86%	1.44%	2.72%
Atomistic	0.94%	2.44%	3.46%	4.07%	4.32%	1.54%	1.84%	1.50%	1.61%	1.80%
Weighted	0.94%	1.42%	2.33%	2.79%	3.32%	1.12%	1.31%	1.68%	1.07%	1.68%

Table 16: Inconsistency and intrinsic difficulty of different diffusion-based generative AI models. Classification error rate and obtained with EfficientNetv2 small.

	FLUX.1-schnell	Stable Diffusion 3	SD XL Turbo	Wuerstchen	Kandinsky 2.2
Holystic	0.77%	1.44%	1.19%	0.53%	1.14%
Atomistic	0.80%	1.49%	1.27%	0.57%	1.17%
Weighted	0.66%	1.25%	1.17%	0.47%	0.93%

we observe a general slightly higher generalization capability when training over various sources but when the testing of a source that is completely absent from the training set the mismatch problem generally remains.

One can also note that the "weighted" strategy proposed in [49] and [35] gave the best overall results.

7 Conclusions and Possible Future Works

The present paper proposes the first study on the cause and the impact of the source mismatch in the field of deepfake detection. We first review the state-of-the-art emphasizing the lack and understanding of the source mismatch or even its lack of consideration in the numerical results, which are generally presented in match conditions between testing and testing sets.

Based on prior works in the field of hidden information detection, we propose definitions for the source and the source mismatch problem. In order to assess the problem in practical cases and compare the impact of the different elements, we adjusted the definitions for the source intrinsic difficulty and source inconsistency.

Equipped with those definitions and focusing on the latest diffusion-based generative AI models, we systematically assess the impact of each element, from text tokenization to the post-generation image processing operations.

Last but not least, we study the relevance of three methods for mitigating the problem of source mismatch in deepfake detection. We demonstrate that when including as many sources as possible in the training set and using an adapted classification method, this problem can be largely overcome. However, the problem of Out-of-Distribution source inspection remains largely open.

This leads us to acknowledge that a lot of different works are needed and that this paper is a first study of this problem, admittedly innovative and with interesting results, but nonetheless a first step. For instance additional works are required to understand the key characteristics that make deepfakes easily detectable and the impact of the source mismatch problem with regard to these characteristics. Additional works are also required to improve the generalization of deepfake detection methods to Out-of-Distribution sources. Our work also focus on image coded as grayscale and it would be interesting to assess similarly the impact of intra-model variation of hyperparameter on deepfake detection for other format such as text, audio and video or color images.

All in all, we believe that the present paper is a cornerstone in the study of the source mismatch problem for the detection of deepfakes and will undoubtedly be a source of future works.

References

[1] E. Altuncu, V. N. Franqueira, and S. Li, "Deepfake: definitions, performance metrics and standards, datasets, and a meta-review", *Frontiers in Big Data*, 7, 2024, 1400024.

[2] X. Bi, B. Liu, F. Yang, B. Xiao, W. Li, G. Huang, and P. C. Cosman, "Detecting Generated Images by Real Images Only", *arXiv preprint arXiv:2311.00962*, 2023.

[3] R. Brooks, Y. Yuan, Y. Liu, and H. Chen, "DeepFake and its Enabling Techniques: A Review", *APSIPA Transactions on Signal and Information Processing*, 11(2), 2022, DOI: 10.1561/116.00000024, http://dx.doi.org/10.1561/116.00000024.

[4] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? understanding properties that generalize", in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, Springer, 2020, 103–20.

[5] H.-S. Chen, S. Hu, S. You, and C.-C. J. Kuo, "DefakeHop++: An Enhanced Lightweight Deepfake Detector", *APSIPA Transactions on Signal and Information Processing*, 11(2), 2022, ISSN: 2048-7703, DOI: 10.1561/116.00000126, http://dx.doi.org/10.1561/116.00000126.

[6] H.-S. Chen, M. Rouhsedaghat, H. Ghani, S. Hu, S. You, and C.-C. J. Kuo, "Defakehop: A light-weight high-performance deepfake detector", in *2021 IEEE International conference on Multimedia and Expo (ICME)*, IEEE, 2021, 1–6.

[7] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection", in *International conference on image analysis and processing*, Springer, 2022, 219–29.

[8] R. Cogranne, "A Comparative Review of Deep-Learning Models for Deepfakes Detection", in *Proceedings of the SPIE 10th International Conference on Multimedia and Image Processing (ICMIP 2025)*, 2025, https://hal.science/hal-04884563.

[9] R. Cogranne, E. Giboulot, and P. Bas, "ALASKA# 2: Challenging academic research on steganalysis with realistic images", in *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2020, 1–5.

[10] R. Cogranne, E. Giboulot, and P. Bas, "The ALASKA steganalysis challenge: A first step towards steganalysis", in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2019, 125–37.

[11] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models", in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023.

[12] D. Cozzolino, G. Poggi, R. Corvi, M. NieSSner, and L. Verdoliva, "Raising the Bar of AI-generated Image Detection with CLIP", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 4356–66.

[13] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. NieSSner, and L. Verdoliva, "Forensictransfer: Weakly-supervised domain adaptation for forgery detection", *arXiv preprint arXiv:1812.02510*, 2018.

[14] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, *et al.*, "Cogview: Mastering text-to-image generation via transformers", *Advances in neural information processing systems*, 34, 2021, 19822–35.

[15] P. Dogoulis, G. Kordopatis-Zilos, I. Kompatsiaris, and S. Papadopoulos, "Improving synthetically generated image detection in cross-concept settings", in *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, 2023, 28–35.

[16] D. C. Epstein, I. Jain, O. Wang, and R. Zhang, "Online detection of ai-generated images", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 382–92.

[17] T. Fang, N. Lu, G. Niu, and M. Sugiyama, "Rethinking importance weighting for deep learning under distribution shift", *Advances in neural information processing systems*, 33, 2020, 11996–2007.

[18] E. Giboulot, P. Bas, R. Cogranne, and D. Borghys, "The Cover Source Mismatch Problem in Deep-Learning Steganalysis", in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, 1032–6, DOI: 10.23919/EUSIPCO55093.2022.9909553.

[19] E. Giboulot, R. Cogranne, D. Borghys, and P. Bas, "Effects and solutions of cover-source mismatch in image steganalysis", *Signal Processing: Image Communication*, 86, 2020, 115888.

[20] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN generated images easy to detect? A critical analysis of the state-of-the-art", in *2021 IEEE international conference on multimedia and expo (ICME)*, IEEE, 2021, 1–6.

[21] L. Guarnera, O. Giudice, F. Guarnera, A. Ortis, G. Puglisi, A. Paratore, L. M. Bui, M. Fontani, D. A. Coccomini, R. Caldelli, *et al.*, "The face deepfake detection challenge", *Journal of Imaging*, 8(10), 2022, 263.

[22] A. Haines, "An Update on Foreign Threats to the 2024 Elections, Senate Select Committee on Intelligence", OFFICE of the DIRECTOR of NATIONAL INTELLIGENCE, opening testimony at Senate Select Committee on Intelligence hearing for the Annual Threat Assessment of the U.S. Intelligence Community, 2024, https://www.dni.gov/index.php/newsroom/congressional-testimonies/congressional-testimonies-2024/3823-dni-haines-opening-statement-as-delivered-to-the-ssci-for-an-update-on-foreign-threats-to-the-2024-elections.

[23] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", *Advances in neural information processing systems*, 33, 2020, 6840–51.

[24] J. Ho and T. Salimans, "Classifier-free diffusion guidance", *arXiv preprint arXiv:2207.12598*, 2022.

[25] C. Huyen, *Designing machine learning systems*, O'Reilly Media, Inc., 2022.

[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, 1125–34.

[27] A. Jain, P. Korshunov, and S. Marcel, "Improving Generalization of Deepfake Detection by Training for Attribution", in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 2021, 1–6, DOI: 10.1109/MMSP53017.2021.9733468.

[28] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models", *Advances in neural information processing systems*, 35, 2022, 26565–77.

[29] A. D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevný, "Moving steganography and steganalysis from the laboratory into the real world", in *Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security*, *IH&MMSec '13*, Montpellier, France: Association for Computing Machinery, 2013, 45–58, ISBN: 9781450320818, DOI: 10.1145/2482513.2482965, https://doi.org/10.1145/2482513.2482965.

[30] C. Kraetzer, D. Siegel, S. Seidlitz, and J. Dittmann, "Process-Driven Modelling of Media Forensic Investigations-Considerations on the Example of DeepFake Detection", *Sensors*, 22(9), 2022, ISSN: 1424-8220, DOI: 10.3390/s22093137, https://www.mdpi.com/1424-8220/22/9/3137.

[31] C. Li, Z. Huang, D. P. Paudel, Y. Wang, M. Shahbazi, X. Hong, and L. Van Gool, "A Continual Deepfake Detection Benchmark: Dataset, Methods, and Essentials", in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, 1339–49.

[32] M. Li, Y. Wan, and J. Gao, "What drives the ethical acceptance of deep synthesis applications? A fuzzy set qualitative comparative analysis", *Computers in Human Behavior*, 133, 2022, 107286, ISSN: 0747-5632, DOI: https://doi.org/10.1016/j.chb.2022.107286, https://www.sciencedirect.com/science/article/pii/S074756322200108X.

[33] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts", *arXiv preprint arXiv:1608.03983*, 2016.

[34] A. Mallet, M. Bene, and R. Cogranne, "Cover-source mismatch in steganalysis: systematic review", *EURASIP Journal on Information Security*, 2024(1), 2024, 26, DOI: 10.1186/s13635-024-00171-6, https://doi.org/10.1186/s13635-024-00171-6.

[35] A. Mallet, M. Kuribayashi, R. Cogranne, and P. a. Bas, "An Original Method For Detection Of AI-Generated Images Based On Noise Covariance", in under review, 2024.

[36] S. Mandelli, N. Bonettini, P. Bestagini, and S. Tubaro, "Detecting gan-generated images by orthogonal training of multiple cnns", in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2022, 3091–5.

[37] S. Mandelli, N. Bonettini, P. Bestagini, and S. Tubaro, "Training CNNs in Presence of JPEG Compression: Multimedia Forensics vs Computer

Vision", in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020, DOI: 10.1109/WIFS49906.2020.9360903.

[38] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of gan-generated fake images over social networks", in *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, IEEE, 2018, 384–9.

[39] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward", *Applied Intelligence*, 53(4), June 2022, 3974–4026, ISSN: 0924-669X, DOI: 10.1007/s10489-022-03766-z, https://doi.org/10.1007/s10489-022-03766-z.

[40] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey", *Computer Vision and Image Understanding*, 223, 2022, 103525, ISSN: 1077-3142, DOI: https://doi.org/10.1016/j.cviu.2022.103525, https://www.sciencedirect.com/science/article/pii/S1077314222001114.

[41] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 24480–9.

[42] D. Pan, L. Sun, R. Wang, X. Zhang, and R. O. Sinnott, "Deepfake Detection through Deep Learning", in *2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, 2020, 134–43, DOI: 10.1109/BDCAT50828.2020.00001.

[43] F. Petitcolas, "Kerckhoffs principle. Encyclopedia of cryptography and security", 2011.

[44] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*, Mit Press, 2022.

[45] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation", in *International conference on machine learning*, Pmlr, 2021, 8821–31.

[46] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis", in *International conference on machine learning*, PMLR, 2016, 1060–9.

[47] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. NieSSner, "Faceforensics++: Learning to detect manipulated facial images", in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[48] N. Science and T. Council, "Roadmap for Researchers on Priorities Related to Information Integrity Research and Development", The White House, 2022, https://www.whitehouse.gov/wp-content/uploads/2022/12/Roadmap-Information-Integrity-RD-2022.pdf.

[49] R. Seo, M. Kuribayashi, A. Ura, A. Mallet, R. Cogranne, W. Mazurczyk, and D. Megías, "Toward Universal Detector for Synthesized Images by Estimating Generative AI Models", in *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024.

[50] D. Siegel, C. Krätzer, S. Seidlitz, and J. Dittmann, "Forensic data model for artificial intelligence based media forensics-Illustrated on the example of DeepFake detection", *Electronic Imaging*, 34, 2022, 1–6.

[51] L. N. Smith, "Cyclical learning rates for training neural networks", in *2017 IEEE winter conference on applications of computer vision (WACV)*, IEEE, 2017, 464–72.

[52] L. N. Smith, "No more pesky learning rate guessing games", *CoRR, abs/1506.01186*, 5, 2015, 575.

[53] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates", in *Artificial intelligence and machine learning for multi-domain operations applications*, Vol. 11006, SPIE, 2019, 369–86.

[54] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics", in *International conference on machine learning*, PMLR, 2015, 2256–65.

[55] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models", 2022, arXiv: 2010.02502 [cs.LG], https://arxiv.org/abs/2010.02502.

[56] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution", *Advances in neural information processing systems*, 32, 2019.

[57] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Rethinking the upsampling operations in cnn-based generative network for generalizable deepfake detection", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 28130–9.

[58] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei, "Learning on gradients: Generalized artifacts representation for gan-generated images detection", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 12105–14.

[59] L. Trinh and Y. Liu, "An examination of fairness of ai models for deepfake detection", in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021, https://www.ijcai.org/proceedings/2021/0079.pdf.

[60] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, 8695–704.

[61] R. Wightman, "Pytorch Image Models (timm)", 2019.

[62] Z. Yan, Y. Luo, S. Lyu, Q. Liu, and B. Wu, "Transcending forgery specificity with latent space augmentation for generalizable deepfake detection", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 8984–94.

[63] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection", in *Advances in Neural Information Processing Systems*, ed. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Vol. 36, Curran Associates, Inc., 2023, 4534–65, https://proceedings.neurips.cc/paper_files/paper/2023/file/0e735e4b4f07de483cbe250130992726-Paper-Datasets_and_Benchmarks.pdf.

[64] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion models in generative AI: A survey", *arXiv preprint arXiv:2303.07909*, 2023.

[65] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in gan fake images", in *2019 IEEE international workshop on information forensics and security (WIFS)*, IEEE, 2019.

[66] J. Zheng, M. Hu, Z. Fan, C. Wang, C. Ding, D. Tao, and T.-J. Cham, "Trajectory consistency distillation", *arXiv preprint arXiv:2402.19159*, 2024.

[67] R. Zhou, C. Jiang, and Q. Xu, "A survey on generative adversarial network-based text-to-image synthesis", *Neurocomputing*, 451, 2021, 316–36.