Diffusion-Based Authentication of Copy Detection Patterns: A Multimodal Framework with Printer Signature Conditioning

Bolutife Atoki Iuliia Tkachenko Bertrand Kerautret Carlos Crispim-Junior
Université Lumière Lyon 2, CNRS, INSA Lyon, Universite Claude Bernard Lyon 1, LIRIS
UMR5205, 69007 Lyon, France
{bolutife.atoki, iuliia.tkachenko, bertrand.kerautret, carlos.crispim-junior}@liris.cnrs.fr

Abstract

Counterfeiting affects diverse industries, including pharmaceuticals, electronics, and food, posing serious health and economic risks. Printable unclonable codes, such as Copy Detection Patterns (CDPs), are widely used as an anti-counterfeiting measure and are applied to products and packaging. However, the increasing availability of high-resolution printing and scanning devices, along with advances in generative deep learning, undermines traditional authentication systems, which often fail to distinguish high-quality counterfeits from genuine prints. In this work, we propose a diffusion-based authentication framework that jointly leverages the original binary template, the printed CDP, and a representation of printer identity that captures relevant semantic information. Formulating authentication as multi-class printer classification over printer signatures lets our model capture fine-grained, device-specific features via spatial and textual conditioning. We extend ControlNet by repurposing the denoising process for class-conditioned noise prediction, enabling effective printer classification. On the Indigo 1×1 *Base dataset, our method outperforms traditional similarity metrics and prior deep learning approaches. Results show the framework generalizes to counterfeit types unseen during training.*

1. Introduction

Counterfeiting impacts industries from pharmaceuticals to food, posing serious health, safety, and economic risks. The widespread availability of high-quality printing and scanning devices, coupled with advances in generative deep learning methods, has driven the increase in counterfeit products. To address this challenge, some anti-counterfeiting solutions employ printable unclonable codes applied to products or packaging. A simple and cost-effective form of these codes is the Copy Detection Pattern (CDP), a maximum-entropy image that is sensitive to copying and undergoes information loss when reprinted [19].

However, advances in generative methods have introduced new challenges. Modern computer vision approaches [11, 24] enable the estimation of binary approximations of CDPs, facilitating near-perfect replicas with little information loss, and undermine the reliability of traditional authentication methods [8, 28] as illustrated in Figure 1. In response, recent approaches have examined authentication using the binary template alone, printed CDP alone, or a combination of both.

Independent studies [1, 9, 13, 14] show printers leave unique, repeatable signatures on printed outputs due to subtle hardware and mechanical variations. These printer-specific signatures capture the identity of the printer, enabling discrimination even between different units of the same model. Unlike prior work using printer information only for grouping or evaluation, we treat the printer as an identity-bearing entity central to authentication. We represent each printer's identity using a human-readable textual description, enabling the model to learn associations between signature patterns and printer classes, thereby supporting broader generalisation.

We introduce a unified authentication framework that integrates the binary template, printed CDP, and printer identity in a single decision-making process. We formulate authentication as a multi-class classification task, where the goal is to identify the source printer. Our framework operates in the latent space using a diffusion-based architecture, which enables the model to capture low-dimensional representations that retain fine-grained, printer-specific details. Compared to prior generative approaches, Generative Adversarial Networks (GANs) [11] produce sharp images but often suffer from mode collapse and reduced diversity [2], while Variational Autoencoders (VAEs) [16] cover more modes but typically yield blurrier results [4]; diffusion models achieve both high fidelity and broad distribution coverage [10]. A textual description of the printer and a spatial conditioning mechanism guide the model's attention to regions likely to contain these signature patterns. To

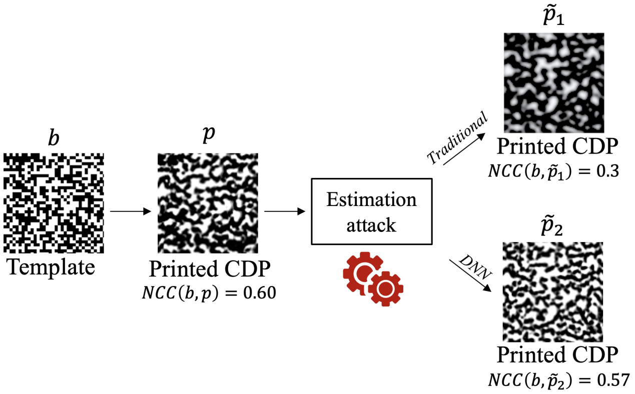

Figure 1. Examples of a binary template, authentic printed CDP, and counterfeit printed CDPs from traditional image processing and deep learning methods. Normalised cross-correlation reliably detects traditional counterfeits but is unreliable against deep learning-based ones.

perform classification, we introduce a novel classification block within a semantically and spatially conditioned setting, allowing the model to extract the printer's signature from both the binary template and the printed CDP. Authentication is achieved by comparing the predicted printer identity with the expected one associated with the authentic CDP. Discrepancies between the two indicate potential counterfeiting.

In summary, our contributions are four-fold. First, we incorporate printer-specific signatures into the CDP authentication process, treating the printer as an identity-bearing entity rather than auxiliary metadata. Second, we propose the first unified authentication framework that jointly considers the binary template, printed CDP, and printer identity within a single pipeline. Third, we enable cross-printer generalisation by formulating authentication as a multi-class classification task over printer identities. Finally, we extend the image-conditioned diffusion model ControlNet [31] beyond its generative purpose by modifying its architecture to support classification. Specifically, we introduce a classification module that leverages the model's learned denoising capabilities to perform class discrimination. Our implementation and models are available at our public repository [1].

The remainder of this paper is as follows. In Section 2, we review existing work on CDP authentication. Section 3 describes the proposed methodology and authentication strategy. In Section 4, we perform experiments and the evaluation of our method. Section 5 concludes the paper with a summary of our findings and perspectives for future research.

2. Previous Work

In this section, we review existing methods for authenticating copy detection patterns (CDPs). We begin by briefly discussing common counterfeiting techniques used to attack authentication systems, as understanding these attacks is essential for evaluating the effectiveness of various authentication approaches. Then, we broadly categorize existing authentication methods into three main approaches based on the type of data they utilise: (i) template-based methods, (ii) printer type-based methods, and (iii) printed CDP-based authentication methods.

Counterfeiters estimate binary templates using techniques from thresholding (e.g., Otsu) to deep learning. This process is illustrated in Figure 1, which shows the progression from a genuine binary template to an authentic printed CDP, followed by template estimation attacks and their printing as counterfeits. Models like Bottleneck DNN, autoencoders, SRGAN, and U-Net produce high-resolution approximations closely mimicking authentic templates [8, 25, 28, 29]. Given the variety and sophistication of these counterfeiting techniques, authentication methods have been developed to detect and prevent forgery. These methods can be broadly classified based on the type of data they utilise, as detailed below.

Template-based CDP authentication. The simplest form of template-based authentication compares a test printed CDP (probe) to its corresponding binary template using similarity metrics such as Normalised Cross-Correlation (NCC) or Structural Similarity Index (SSIM). A similarity score is computed and evaluated against a threshold derived from authentic pairs. These methods once sufficed but now fail against high-quality template estimates [3, 8, 28].

To improve reliability, more advanced approaches have been proposed. The authors in [26] introduced a method that models the printing-and-digitisation (P&D) process as a transformation of local 3 × 3 binary neighbourhoods. For each printer, a codebook is constructed, mapping the probability of bit-flip for every one of the 512 possible neighbourhood configurations. At authentication, the flipped bits in the binarised probe are compared to the expected flipping probabilities, and authentication is performed via thresholding. Later, Tutt *et al.* [27] extended this idea by using a two-class SVM trained on the bit-flip probabilities and the corresponding centre pixel values for each neighbourhood.

Though these methods surpass basic similarity metrics, they only indirectly capture P&D noise through empirical bit-flip probabilities. The noise is not explicitly modelled or integrated into the learning process, which limits adaptability. Consequently, a separate codebook is required for each printer, preventing generalisation across devices and restricting scalability in real-world deployments.

Printer type-based authentication. These methods rely

[1] Repository URL is available at https://gitlab.liris.cnrs.fr/anr-trustit/diffusion_based_cdp_authentication.git

solely on the image of the printed CDP for authentication, based on the assumption that printed outputs exhibit characteristics unique to their printer type. One such method is proposed by Zeghidi *et al*. [30], who use a Siamese Neural Network [22] to embed printed CDP pairs into a shared feature space. The network is trained to minimise the distance between embeddings from the same printer and maximize it for different printers. At test time, a probe is compared to multiple reference samples from the claimed printer, and is accepted as authentic if the majority of distances fall below a learned threshold. Because it ignores the template, this method may fail when unknown templates are printed on known printers.

Printed CDP-based authentication. Authentication methods that compare the test printed CDP with its authentic counterpart have proved to yield reliable authentication performance [20]. However, directly comparing a probe to its corresponding authentic sample is costly and impractical at scale due to the need for storing or retrieving large databases of reference prints. To address this, generative methods are employed to synthesize the expected appearance of a printed CDP directly from its digital template, enabling on-demand reference generation and eliminating the need for physical sample storage. This enables the authentication process to rely on comparisons with a predicted print, thereby reducing overhead and improving scalability.

Pulfer *et al*. [20] employed a U-Net model to predict a print from a given binary template, leveraging its effectiveness in image-to-image tasks. The authentication decision was based on the masked pixel-wise error between the probe and the predicted print, under the assumption that an authentic printed CDP would closely match the predicted print. Chaban *et al*. [6] adopted a Pix2Pix framework [12], which utilizes a conditional GAN to enhance realism and fidelity in the synthesized outputs. The adversarial training in Pix2Pix encourages the generator to produce outputs indistinguishable from real printed samples, making it more robust to subtle printing variations. The Pearson Correlation Coefficient (PCC) was used as a similarity metric, emphasizing global structural alignment between the synthesized and probe CDPs. However, while these methods can generate plausible printed CDP images for new, unseen templates, they are trained on samples from a single printer type, as they lack explicit conditioning on printer identity. Consequently, their outputs belong solely to the distribution of the printer they were trained on and do not handle samples from other printers.

While prior work has advanced CDP authentication by leveraging the template and the printed output, existing methods are typically trained on data from a single printer and lack mechanisms to explicitly model printer-specific variations. As a result, they are limited in their ability to perform reliably across different printer types, and often fail to account for the characteristic distortions introduced by specific printing processes. To overcome these limitations, we propose a novel authentication framework that, for the first time, jointly considers the binary template, printed CDP, and printer identity. The following section details our methodology for leveraging this combined information to extract printer-specific signatures and improve counterfeit detection.

3. Methodology

Our approach to counterfeit detection is based on the printed CDP, binary template, and printer identity. The core idea is to encode these inputs into a latent representation that captures printer-specific characteristics. Authentication is then based on detecting inconsistencies between the latent signature inferred from a candidate CDP and the expected signature for the known authentic printer. To achieve this, we newly formulate authentication as a supervised multi-class classification problem. Each known authentic printer corresponds to a unique class, and various categories of counterfeit CDPs are modelled as additional classes. This formulation enables training on data from multiple printers and supports cross-printer authentication.

3.1. Definitions

Let the original binary template be denoted as b, and the authentic CDP printed from b as p. The estimated template from a printed CDP is represented as $\tilde{b}$, while $\tilde{p}$ denotes a counterfeit CDP printed from $\tilde{b}$. The candidate CDP presented for authentication is denoted as $y \in \{p, \tilde{p}\}$, where y may be either authentic p or counterfeit $\tilde{p}$.

In the diffusion framework, the template p is encoded into a latent representation x_0, which is subsequently noised to timestep t, yielding x_t, where $0 < t \leq T$. Let z denote the latent representation of the printed CDP corresponding to y. Let $\{c_i\}_{i=1}^{k}$ denote the set of class labels, where each c_i encodes textual information describing the printing process used to generate CDPs in class i.

3.2. Diffusion-Based Signature Classification

Our method learns printer-specific signatures by modelling the reverse diffusion process conditioned on both the printed CDP and printer identity. The strategy relies on the fact that the reverse process follows the correct denoising path only when the conditioning information aligns with the true source printer of the candidate CDP.

Forward diffusion: We begin with the standard forward diffusion process, where Gaussian noise is gradually added to the template latent x_0 over T steps. Here, $t = 0$ signifies the original (noise-free) latent, and noise increases progressively with t.

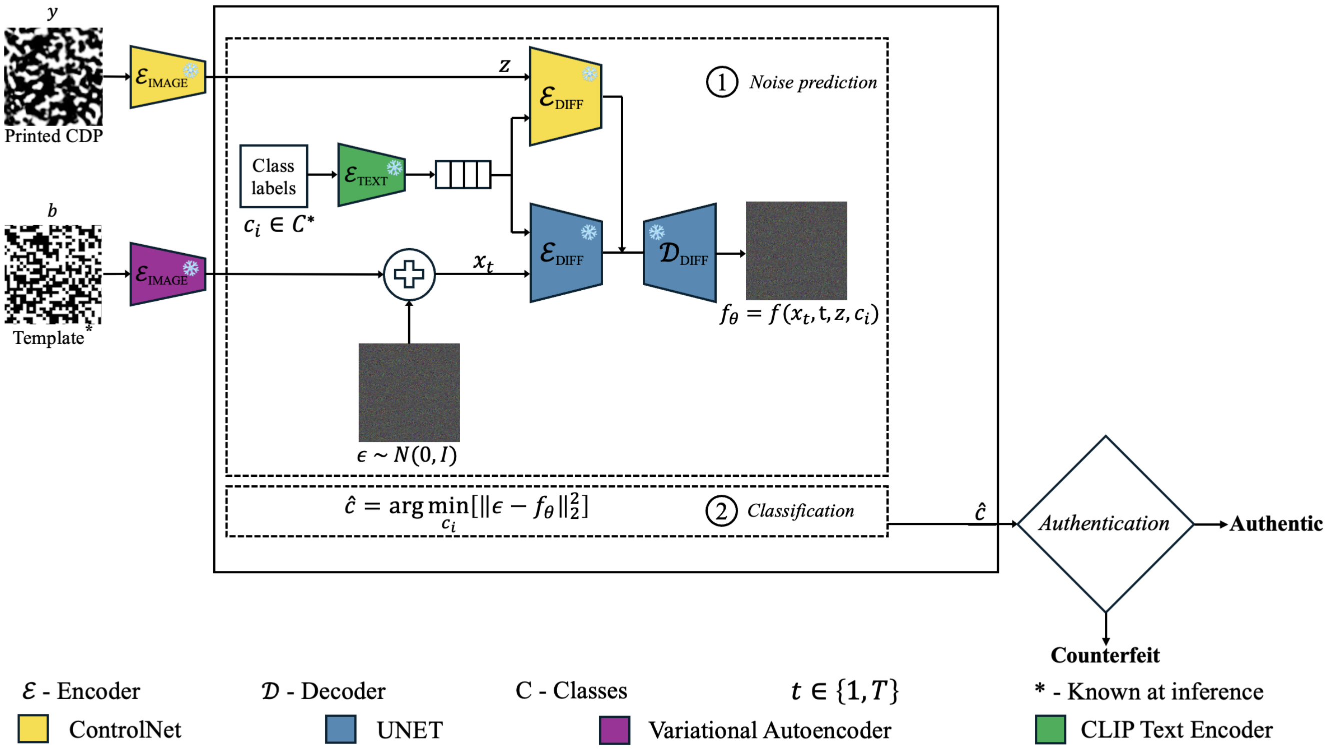

Figure 2. Authentication pipeline leveraging printer signature reconstruction. The reverse diffusion process takes a noised template as input and is conditioned on the printed CDP and printer identity text. ControlNet is modified to predict noise per class and includes a classification module selecting the class with minimal prediction error.

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\alpha_t \in (0,1)$ denotes the variance-preserving noise schedule controlling the amount of noise added at each timestep.

Reverse diffusion: The model learns the reverse denoising process, conditioned on the latent printed CDP z and textual printer information c_i, to predict the added noise ϵ. Since textual conditioning provides global semantic context associated with the printer class, it is limited in its ability to offer fine-grained spatial control. To address this, the printed CDP latent z is included as an image-based conditioning signal, guiding the model's attention toward spatially localised features that reflect printer-specific artifacts.

$$f_\theta(x_t, t, z, c_i) \approx \epsilon, \tag{2}$$

where f_θ is a neural network trained to minimise the difference between predicted and actual noise.

3.3. Framework components

Our framework performs signature classification by modelling the correct denoising trajectory of a noised binary template, conditioned on both printer identity and the corresponding printed CDP. Textual conditioning offers global semantics but lacks spatial specificity to capture localised printer features. To address this, recent image-conditioned diffusion models incorporate spatial alignment using cross-attention or feature injection mechanisms [18, 31]. We adopt the ControlNet architecture [31], which introduces spatial control in diffusion models through a parallel, trainable branch that injects conditioning (e.g., edge maps or segmentation masks) into the denoising U-Net at multiple resolution levels. In our case, we condition the model on the printed CDP, encoded into a latent representation, to guide denoising toward printer-discriminative regions.

We extend ControlNet beyond its original generative purpose by modifying its architecture to support classification. This involves two key architectural changes, highlighted in Figure 2: (Block 1) the model is adapted to predict noise independently for each candidate class, retaining only the predicted noise rather than generating images, and (Block 2) a classification module is added, leveraging the model's learned denoising capabilities to select the class with the lowest prediction error, measured as the discrepancy between predicted and actual noise. Our approach builds on the diffusion classifier paradigm introduced in [17], which performs zero-shot classification by identifying the class label whose conditioning minimises noise prediction error. Our model conditions the reverse diffusion process on both the text description of printer identity and the printed CDP. For a given noised latent x_t, the model predicts the noise under each candidate class c_i, and selects the class with the lowest error:

$$\hat{c} = \arg\min_{c_i} \mathbb{E}_t \left[\|\epsilon - f_\theta(x_t, t, z, c_i)\|_2^2 \right], \tag{3}$$

where $\hat{c}$ is the predicted printer class, and $\mathbb{E}_t$ denotes expectation

Figure 2 shows the binary template encoded and perturbed with Gaussian noise at a random timestep t. The printer identity is encoded as a text embedding, and the printed CDP is encoded into an image latent. These are used to condition the model, which then predicts the noise for each candidate class. To improve robustness and reduce the impact of timestep-specific bias, this process is repeated N times with different randomly sampled timesteps. The final predicted class is selected as the one with the lowest aggregated prediction error across all sampled timesteps.

Authentication strategy via classification. Our authentication strategy compares the predicted printer class $\hat{c}$, obtained by classifying the probe CDP using our diffusion-based model conditioned on printer identity, with the known ground-truth printer class c^*, which is assumed to be available at test time. For each candidate CDP y, authentication is performed by verifying whether $\hat{c} == c^*$. A match indicates the CDP was printed by the correct authentic printer and is deemed authentic. Any mismatch implies a counterfeit, even if c^* corresponds to a different legitimate printer from the same manufacturer. This per-printer authentication provides finer granularity than manufacturer-level checks, enhancing security by preventing adversaries from evading detection by using another authorized printer. Although printers from the same manufacturer may share similar global signatures, only the authentic printer is expected to replicate the exact spatial and structural features present in the original CDP, making impersonation difficult. We assume a closed-set classification setting where the complete set of candidate printer classes, including both authentic and counterfeit printers, is known at test time. Formally, the authentication decision rule is:

$$\text{Auth}(y) = \begin{cases} \text{Authentic}, & \text{if } \hat{c} = c^* \\ \text{Counterfeit}, & \text{otherwise} \end{cases} \tag{4}$$

4. Experiments

4.1. Experimental details

Dataset Selection. Existing datasets for CDP authentication provide limited support for modelling printer-specific signatures. The Scantrust dataset [15] includes only a single printer and simulates counterfeit samples by varying binarisation methods, which does not support learning printer-dependent characteristics. The Indigo datasets offer improved support with two printers and are available in two variants: the Indigo 1×1 Base dataset [8] and the Indigo Variability dataset [7]. The latter extends the Base variant by introducing four paper types to increase variability. Though limited to two printers and one scanner, these datasets are the most widely used for CDP authentication. We adopt the Indigo 1×1 Base dataset to model fine-grained printer-specific signatures under controlled conditions, as it introduces hardware variability without confounding factors. It contains binary templates and corresponding printed CDPs produced by two industrial printers: HP Indigo 5500 and HP Indigo 7600.

We encode printer identity using natural language descriptions of the printing process to align with CLIP's [21] pre-trained text encoder. For example, an authentic print from the HP Indigo 5500 is described as: "Data Matrix image printed with HP Indigo 5500 printer", while a counterfeit reprinted with the 7600 is described as: "Data Matrix image printed with HP Indigo 5500 printer, then reprinted with HP Indigo 7600 printer". These semantically rich descriptions better reflect the natural language patterns seen during CLIP's training, potentially improving text–image alignment compared to minimal phrases like "HP Indigo 5500". We reserve alternative representations of printer identity for ablation studies. Overall, this approach leverages CLIP's language priors to leverage meaningful relationships between printer classes and to support learning of subtle, device-specific signatures through multi-printer training.

Dataset reformulation for classification. The Indigo dataset is reformulated into a six-class classification task. Two classes correspond to authentic prints from HP Indigo 5500 and HP Indigo 7600, respectively. The remaining four are counterfeit types generated by estimating templates from authentic prints and reprinting them using either of the two printers. These are labeled using the format *HPXX_YY*, where *XX* refers to the original printer and *YY* to the reprinting device. Each class contains 720 samples, resulting in a balanced dataset of 4320 samples total. Unlike prior work that evaluates CDP authentication in a binary setting by pairing one authentic printer with its counterfeit variants, our formulation includes all printers and counterfeit types jointly. This supports learning of cross-printer signature variations in a unified classification setting.

Data preparation and augmentation. To ensure robust evaluation, the dataset is split into 70% training (504 samples/class), 10% validation (72 samples/class), and 20% testing (144 samples/class). To prevent information leakage, the splits are based on template ID, ensuring each template and printed variants appear in only one of the splits.

This enforces that the model generalizes to unseen templates during evaluation, which is more realistic and difficult. At both training and inference time, the corresponding binary template b used to generate each printed CDP is assumed to be available, as it can be retrieved from a secured database in typical deployment settings. During training, each template–CDP pair is augmented 20 times via random 512×512 crops (always), horizontal/vertical flips (p=0.8), and for printed CDPs only, photometric transforms: Gaussian blur (kernel 3–7), Gaussian noise (var. 0.001–0.005), and brightness/contrast shifts (±0.2), each with applied probability p=0.7. Two photometric transforms are applied per image (p=0.7). See our code repository for full augmentation parameters and implementation details.

Evaluation criteria. We evaluate the model using a multiclass classification setup, where each class corresponds to one of two authentic printers or one of four counterfeit printing scenarios. The authentic classes (**HP55**, **HP76**) represent prints from HP Indigo 5500 and HP Indigo 7600, respectively. The counterfeit classes are derived by reprinting estimated templates using either printer.

Evaluation is conducted along two complementary axes:

1. **Classification accuracy**
   This measures the model's ability to correctly identify the source printer class for each printed CDP across all six classes. Results are reported using a normalised confusion matrix, with each row representing the ground truth and each column the predicted class.

2. **Authentication performance**
   Authentication is formulated as a binary decision: whether the predicted class matches the expected authentic printer. We use the balanced error rate P_{err}, defined as the average of the false rejection rate (P_{miss}) and the false acceptance rate (P_{fa}):

$$\mathrm{P}_{\text{err}} = \frac{\mathrm{P}_{\text{miss}} + \mathrm{P}_{\text{fa}}}{2} \tag{5}$$

This metric equally penalizes both error types and is robust to class imbalance. Here, P_{miss} is the proportion of authentic samples incorrectly rejected, and P_{fa} is the proportion of counterfeit samples incorrectly accepted.

Training pipeline. The authentication framework is built using pretrained Stable Diffusion [23]. To better preserve structural information in binary templates, the variational autoencoder (VAE) was fine-tuned on template data, as the original model showed limited reconstruction fidelity for high-frequency details. The VAE fine-tuning was performed using the same template images from the dataset splits and with the same training hyperparameters later used for ControlNet training. We then trained the ControlNet components used for image-based conditioning, specifically the parallel branch of the denoising U-Net and its associated image encoder, using a diffusion loss to align the denoising trajectory with the conditioning inputs. All other components, including the main U-Net backbone and CLIP-based text encoder, were kept frozen. Training was conducted in mixed precision (FP16/FP32) on an NVIDIA H100 80GB GPU, with a batch size of 16 and gradient accumulation of 4. The optimiser used was AdamW, with a cosine learning rate schedule and linear warmup over the first 500 steps, starting from an initial learning rate of 8×10^{-5}. Training was performed for 100 epochs. During inference, authentication was performed using 50 denoising trials per sample to ensure prediction stability.

4.2. Results and Discussion

Classification accuracy. Table 2 presents the classification accuracy for each class using a confusion matrix. The high diagonal values across the matrix demonstrate strong classification performance. Both authentic classes are consistently and accurately identified, indicating high reliability in recognising genuine CDPs. This is crucial for minimising false rejections in the authentication pipeline.

Counterfeit classes show asymmetric confusion, especially within the same source printer (e.g., HP55 vs HP76). For instance, **HP76_55** is misclassified as **HP76_76** in 19% of cases. This suggests the model struggles to fully disentangle the latent signatures inherited from the original authentic printer. However, no such confusion occurs across counterfeit classes originating from different source printers, indicating that the model effectively distinguishes CDPs from distinct authentic origins. Since confusion is within the counterfeit domain, we next evaluate its impact on authentication.

Authentication performance. We evaluate the authentication performance of our proposed method using the balanced error rate obtained from false rejections and false acceptances. Our results are compared against traditional similarity-based baselines and a deep learning approach adapted from [6]. The results include overall performance (P_{err}), error rates for authentic and counterfeit CDPs (P_{miss} and P_{fa}), and a per-class analysis of false authentications.

The results are presented in Table 1. Traditional baselines rely on computing similarity metrics between the candidate CDP and its binary template. We use Normalised Cross-Correlation (NCC) and Structural Similarity Index (SSIM) as representative metrics [3,5,8,28]. For each class, we determine a similarity threshold from the validation set. During testing, samples are authenticated based on whether their similarity score exceeds the threshold for their class. While conceptually simple, these methods are highly sensitive to printer-specific variations and fail to model interclass differences effectively. Their performance is modest across all classes, with similar error values; neither

Table 1. Authentication error rate for method comparison with baselines. Lower is better.

		$\mathbf{P}_{err}$	$\mathbf{P}_{miss}$			$\mathbf{P}_{fa}$				
			HP55	*HP76*	*Mean*	*HP55_55*	*HP55_76*	*HP76_76*	*HP76_55*	*Mean*
Traditional	NCC	0.286	0.292	0.310	0.301	0.300	0.361	0.264	0.201	0.273
	SSIM	0.275	0.264	0.300	0.281	0.292	0.285	0.292	0.210	0.269
Deep Learning	[6]	0.091	0.125	0.111	0.118	0.097	0.16	**0.000**	0.000	0.064
	Ours	**0.023**	**0.049**	**0.035**	**0.005**	**0.000**	**0.014**	0.007	**0.000**	**0.042**

Table 2. Confusion matrix of classification results. Values are in percentages.

		Predicted class (%)					
		HP55	HP76	HP55_55	HP55_76	HP76_76	HP76_55
Actual class (%)	HP55	**95**	5	0	0	0	0
	HP76	3	**97**	0	0	0	0
	HP55_55	0	0	**83**	15	1	1
	HP55_76	1	0	14	**83**	1	1
	HP76_76	0	1	1	1	**89**	8
	HP76_55	0	0	2	2	19	**77**

method shows a clear advantage in correctly accepting authentic samples or rejecting counterfeits, with errors generally ranging between 0.269 and 0.300.

Chaban *et al.* [6] trained a Pix2Pix model [12] to synthesize a CDP from the input binary template, then compared this synthesis to the candidate CDP using the Pearson Correlation Coefficient. In the original setup, separate models were trained per printer, and counterfeit data was not considered. For a fair comparison, we retrained their model using their open-source code, incorporating both authentic and counterfeit classes in a unified training pipeline. This improved its generalizability, achieving as little as $P_{fa} = 0.064$ for falsely accepting counterfeit classes. However, its performance remains inconsistent on authentic samples, with a miss rate of 0.125 for the **HP55** class.

Our method, conditioned jointly on printer identity and CDP images, outperforms all baselines, achieving a low overall error rate P_{err} of **0.023**. All counterfeit classes are reliably rejected, including those most prone to confusion in the classification step such as *HP55_55* and *HP76_55*. On the authentic side, our method achieves the lowest miss rates across both classes, with $P_{miss} = 0.005$ for all authentic CDPs. Notably, the classes that were most challenging for the baseline methods, such as **HP55** and **HP55_76**, show improved error rates under our model. This suggests that conditioning on printer identity during training enables the model to better capture intra-class variability and learn more discriminative representations, particularly in challenging cases.

Although some counterfeit classes exhibit internal confusion during classification, the authentication decision remains robust, with low P_{fa} across all counterfeit types. This robustness comes from the authentication strategy, which verifies whether the predicted printer identity for a candidate CDP matches the known authentic printer. Consequently, even if a counterfeit is misclassified as another counterfeit class, it is still rejected unless it is erroneously attributed to the correct authentic printer, a scenario that remains rare due to the difficulty of replicating fine-grained, printer-specific structural and spatial features.

Generalisation to unseen counterfeits. To assess the model's robustness against counterfeit types not observed during training, we simulate a realistic deployment scenario where only a subset of forgeries is known. Specifically, we train on authentic CDPs printed with the HP Indigo 5500 and HP Indigo 7600 (HP55, HP76), along with two known counterfeit classes produced by estimating templates from these printers and reprinting them on the same device (HP55_55, HP76_76). At test time, we evaluate on:

(i) **Authentic classes:** HP55, HP76

(ii) **Known counterfeits:** HP55_55, HP76_76

(iii) **Unseen counterfeits:** HP55_76, HP76_55

This setup tests the model's ability to reject counterfeits created using printing paths not encountered during training. Table 3 shows that the model achieves perfect rejection rates ($P_{fa} = 0.000$) on both known and unseen counterfeit types, with a balanced error rate of $P_{err} = 0.012$. These results confirm that the conditioning mechanism enables generalisation beyond the counterfeit strategies seen during training by capturing underlying printer-specific characteristics.

Table 3. Authentication performance on known and unseen counterfeit types. Lower is better.

Counterfeit Type	$\mathbf{P}_{err}$	$\mathbf{P}_{miss}$	$\mathbf{P}_{fa}$
Known counterfeit	0.012	0.024	0.000
Unknown counterfeit	0.012	0.024	0.000

Binary Template Reconstruction. To validate the necessity of VAE fine-tuning, we compared reconstruction quality of binary templates using mean squared error (MSE) between the input and reconstructed outputs under two settings. As shown in Table 4, the pretrained VAE yielded a high average error (0.581), indicating limited structural preservation. Fine-tuning the VAE on binary templates significantly improved reconstruction fidelity, reducing the error to (0.075) and confirming its suitability for encoding high-frequency, pixel-critical content.

Table 4. Mean squared reconstruction error between original and reconstructed binary templates under different configurations.

Configuration	**Avg. MSE (↓)**
Pretrained VAE	0.581
Fine-tuned VAE	**0.075**

Ablation study. To assess the contribution of different inputs to the model's authentication performance, we conduct an ablation study varying two key components: (i) the presence of the binary template, and (ii) how printer identity is represented. Results are summarized in Table 5 and Table 6.

Table 5 shows the impact of removing the binary template. When the template is omitted and replaced by the printed CDP image as input to the architecture, the model relies solely on the printed output and the class label. This leads to a substantial drop in performance, indicating that structural alignment between the template and printed output is essential for capturing fine-grained printer signatures. Without the template as a reference, the model struggles to extract discriminative spatial features, resulting in high error rates.

Table 6 evaluates different strategies for representing printer identity. When printers are represented by index numbers rather than textual descriptions, performance similarly degrades. Notably, this setup results in a high miss rate for authentic samples but relatively low false acceptances, suggesting a strong bias toward predicting samples as counterfeit. This indicates that numerical indices fail to capture the semantic distinctions between printer classes, limiting the model's ability to recognise valid CDPs.

As a second approach to printer identity representation, we tested a configuration where each CDP is conditioned on a natural language description of the typical visual appearance of printed samples from that printer. For example, HP55 samples are described as *"A high-contrast black and white Data Matrix symbol with dense black module clusters, darkened center region, and a coarse, grainy texture throughout the square grid"*. This method outperforms index-based conditioning, particularly in rejecting counterfeits. However, authentic acceptance is less consistent, suggesting that appearance-based descriptions lack the specificity needed to reliably guide authentication.

Together, these results highlight the importance of jointly conditioning on both the binary template and semantically meaningful printer identity. Each plays a distinct role: the template provides structural cues aligned with the original CDP layout, while printer identity embeds class-level semantics necessary to distinguish subtle differences in output characteristics across printers.

Reproducibility. To reproduce our experiments, you can refer to our public repository for scripts and models.

Table 5. Authentication error with and without template input. Lower values indicate better performance.

Input Configuration	$\mathbf{P}_{err}$	$\mathbf{P}_{miss}$	$\mathbf{P}_{fa}$
No Template Input	0.660	0.430	0.200
With Template Input	**0.023**	**0.042**	**0.005**

Table 6. Authentication error comparing different printer identity representations. Lower values indicate better performance.

Representation	$\mathbf{P}_{err}$	$\mathbf{P}_{miss}$	$\mathbf{P}_{fa}$
Printer Index based	0.521	0.962	0.082
Appearance-based	0.035	0.063	**0.001**
Printer Information	**0.023**	**0.042**	0.005

5. Conclusion

We presented a novel authentication framework for Copy Detection Patterns (CDPs) that integrates the binary template, printed CDP, and printer identity into a unified classification pipeline based on diffusion modelling. Unlike prior approaches that rely on template similarity or generative synthesis, we reformulate authentication as a per-printer classification task. This allows the model to learn discriminative printer signatures and supports generalisation across different printers.

A key technical contribution of this work is the extension of the ControlNet architecture beyond its generative role to support classification. By modifying the reverse diffusion process to perform class-specific noise prediction, we enable printer identification via minimum reconstruction error. This leverages semantic (text) and spatial (image) conditioning to capture global printer characteristics and localised print artifacts.

Experiments on the Indigo 1×1 Base dataset demonstrate the effectiveness of our approach in advancing CDP authentication, particularly through the use of semantically rich representations of printer identity that enhance model performance and generalisation. While these results show strong promise, the current experiments are limited to a single scanner configuration. Future work will explore more diverse acquisition settings, increased variability in printers, and investigate training strategies that do not require counterfeit examples.

6. Acknowledgements

This work was financed by the French National Research Agency (ANR), project TRUSTIT referenced under ANR-23-CE39-0002-01

References

[1] Gazi N. Ali, Aravind K. Mikkilineni, Jan P. Allebach, Edward J. Delp, Pei Ju Chiang, and George T. Chiu. Intrinsic and extrinsic signatures for information hiding and secure printing with electrophotographic devices. pages 511–515, 2003. Final Program and Proceedings of IS and T's NIP 19: International Conference on Digital Printing Technologies ; Conference date: 28-09-2003 Through 03-10-2003. 1

[2] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 224–232. PMLR, 2017. 1

[3] Yury Belousov, Brian Pulfer, Roman Chaban, Joakim Tutt, Olga Taran, Taras Holotyak, and Slava Voloshynovskiy. Digital twins of physical printing-imaging channel. In *WIFS*, pages 1–6. IEEE, 2022. 2, 6

[4] Gustav Bredell, Kyriakos Flouris, Krishna Chaitanya, Ertunc Erdil, and Ender Konukoglu. Explicitly minimizing the blur error of variational autoencoders. In *ICLR*. OpenReview.net, 2023. 1

[5] Roman Chaban, Ethan Icet, Olga Taran, and Slava Voloshynovskiy. Comparative analysis of copy detection patterns and physical unclonable functions in authentication systems: A mobile phone perspective. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 740–744, 2024. 6

[6] Roman Chaban, Brian Pulfer, and Slava Voloshynovskiy. Assessing the viability of synthetic physical copy detection patterns on different imaging systems. In *WIFS*, pages 1–6. IEEE, 2024. 3, 6, 7

[7] Roman Chaban, Olga Taran, Joakim Tutt, Taras Holotyak, Yury Belousov, Brian Pulfer, and Slava Voloshynovskiy. Printing variability of copy detection patterns. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, Shanghai, People's Republic of China, December 2022. 5

[8] Roman Chaban, Olga Taran, Joakim Tutt, Taras Holotyak, Slavi Bonev, and Slava Voloshynovskiy. Machine learning attack on copy detection patterns: are 1x1 patterns cloneable? In *WIFS*, pages 1–6. IEEE, 2021. 1, 2, 5, 6

[9] Pei-Ju Chiang, Nitin Khanna, Aravind K. Mikkilineni, Maria V. Ortiz Segovia, Sungjoo Suh, Jan P. Allebach, George T.-C. Chiu, and Edward J. Delp. Printer and scanner forensics. *IEEE Signal Processing Magazine*, 26(2):72–83, 2009. 1

[10] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. 1

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 1

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976. IEEE Computer Society, 2017. 3, 7

[13] Sharad Joshi and Nitin Khanna. Source printer classification using printer specific local texture descriptor. *IEEE Transactions on Information Forensics and Security*, 15:160–171, 2020. 1

[14] Nitin Khanna and Edward J. Delp. Intrinsic signatures for scanned documents forensics : Effect of font shape and size. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 3060–3063, 2010. 1

[15] Elyes Khermaza, Iuliia Tkachenko, and Justin Picard. Can copy detection patterns be copied? evaluating the performance of attacks and highlighting the role of the detector. In *WIFS*, pages 1–6. IEEE, 2021. 5

[16] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1

[17] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, pages 2206–2217. IEEE, 2023. 4

[18] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, pages 4296–4304. AAAI Press, 2024. 4

[19] Justin Picard. Digital authentication with copy-detection patterns. In Rudolf L. van Renesse, editor, *Optical Security and Counterfeit Deterrence Techniques V*, volume 5310 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 176–183, June 2004. 1

[20] Brian Pulfer, Yury Belousov, Joakim Tutt, Roman Chaban, Olga Taran, Taras Holotyak, and Slava Voloshynovskiy. Anomaly localization for copy detection patterns through print estimations. In *WIFS*, pages 1–6. IEEE, 2022. 3

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 5

[22] Sanjeev Jagannatha Rao, Yufei Wang, and Garrison W. Cottrell. A deep siamese neural network learns the human-perceived similarity structure of facial expressions without explicit categories. In *CogSci*. cognitivesciencesociety.org, 2016. 3

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022. 6

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI (3)*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. 1

[25] O. Taran, S. Bonev, and S. Voloshynovskiy. Clonability of anti-counterfeiting printable graphical codes: A machine learning approach. In *ICASSP 2019 - 2019 IEEE Interna-*

tional Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2482–2486, 2019. 2

[26] Joakim Tutt, Olga Taran, Roman Chaban, Brian Pulfer, Yury Belousov, Taras Holotyak, and Slava Voloshynovskiy. Mathematical model of printing-imaging channel for blind detection of fake copy detection patterns. In *WIFS*, pages 1–6. IEEE, 2022. 2

[27] Joakim Tutt, Olga Taran, Roman Chaban, Brian Pulfer, Yury Belousov, Taras Holotyak, and Slava Voloshynovskiy. Authentication of copy detection patterns: A pattern reliability based approach. *IEEE Trans. Inf. Forensics Secur.*, 19:3124–3134, 2024. 2

[28] Rohit Yadav, Iuliia Tkachenko, Alain Trémeau, and Thierry Fournel. Copy sensitive graphical code estimation: Physical vs numerical resolution. In *WIFS*, pages 1–6. IEEE, 2019. 1, 2, 6

[29] Rohit Yadav, Iuliia Tkachenko, Alain Trémeau, and Thierry Fournel. Estimation of copy-sensitive codes using a neural approach. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, IHMMSec'19, page 77–82, New York, NY, USA, 2019. Association for Computing Machinery. 2

[30] Hédi Zeghidi, Carlos Fernando Crispim Junior, and Iuliia Tkachenko. Cdp-sim: Similarity metric learning to identify the fake copy detection patterns. In *WIFS*, pages 1–6. IEEE, 2023. 3

[31] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3813–3824. IEEE, 2023. 2, 4