CertDW: Towards Certified Dataset Ownership Verification via Conformal Prediction

Ting Qiao*, Yiming Li*, Jianbin Li, Yingjia Wang, Leyi Qi, Junfeng Guo, Ruili Feng, Dacheng Tao

Abstract—Deep neural networks (DNNs) rely heavily on high-quality open-source datasets ($e.g.$, ImageNet) for their success, making dataset ownership verification (DOV) crucial for protecting public dataset copyrights. In this paper, we find existing DOV methods (implicitly) assume that the verification process is *faithful*, where the suspicious model will directly verify ownership by using the verification samples as input and returning their results. However, this assumption may not necessarily hold in practice and their performance may degrade sharply when subjected to intentional or unintentional perturbations. To address this limitation, we propose the first certified dataset watermark ($i.e.$, CertDW) and CertDW-based certified dataset ownership verification method that ensures reliable verification even under malicious attacks, under certain conditions ($e.g.$, constrained pixel-level perturbation). Specifically, inspired by conformal prediction, we introduce two statistical measures, including principal probability (PP) and watermark robustness (WR), to assess model prediction stability on benign and watermarked samples under noise perturbations. We prove there exists a provable lower bound between PP and WR, enabling ownership verification when a suspicious model's WR value significantly exceeds the PP values of multiple benign models trained on watermark-free datasets. If the number of PP values smaller than WR exceeds a threshold, the suspicious model is regarded as having been trained on the protected dataset. Extensive experiments on benchmark datasets verify the effectiveness of our CertDW method and its resistance to potential adaptive attacks. Our codes are at GitHub.

Index Terms—Dataset Ownership Verification, Certified Robustness, Data Protection, Trustworthy ML, AI Security

1 Introduction

Recently, deep neural networks (DNNs) have been widely and successfully adopted and deployed in many mission-critical applications, such as face recognition [1], [2], [3]. Arguably, their success heavily relied on the existence of diverse and high-quality public datasets ($e.g.$, ImageNet [4] and LAION-5B [5]). Researchers and developers can use them to train their models and improve their DNNs based on the evaluation results. In particular, these datasets are often mainly limited to educational or research purposes, as their collection and annotation are time-consuming and even costly.

However, safeguarding their copyright ($i.e.$, preventing unauthorized usage of datasets) remains a challenging task, despite there being already many classical in data protection [6]. These methods cannot be used to protect public datasets, as they either hinder accessibility and functionality ($e.g.$, encryption [7], [8], [9]) of these datasets or necessitate the information of the training process of suspicious models or even its manipulation ($e.g.$, digital watermarking [10], [11], [12] and differential privacy [13], [14], [15]) that are not capable for dataset owners in practice.

To the best of our knowledge, dataset ownership verification (DOV) [16], [17], [18], [19], [20], [21] is currently the most widely used and effective method to safeguard the copyright of public datasets [6]. In general, DOV is a post-hoc auditing method, verifying whether a suspicious third-party model is trained on the protected dataset by examining its prediction behaviors on particular samples ($i.e.$, verification samples) without knowing its parameters and training details ($i.e.$, under the black-box setting). These methods consist of two main stages, including dataset watermarking and ownership verification. In the first stage, the dataset owner will modify a few samples in the original (unprotected) dataset to generate its watermarked version, such that all models trained on the watermarked dataset will behave normally on benign testing samples yet have distinctive and pre-defined prediction behaviors ($e.g.$, targeted misclassification) on verification samples. Given the API of a suspicious model, in the second stage, the dataset owner will examine whether this model has dataset-specified behaviors in predicting verification samples. If these special prediction behaviors occur, it is regarded to have been trained on the protected dataset.

In this paper, we revisit existing DOV methods. We reveal that their success relies on a latent assumption that the verification process is 'honest', $i.e.$, the suspicious model will faithfully use verification samples (without adding any noises) to generate their predictions. However, this assumption may not necessarily hold in practice, especially when the suspicious model notices the potential progress of ownership verification. We demonstrate that their performance will degrade sharply, no matter under unintentional random noises or intentional adversarial perturbations. Besides, we notice that there is a close connection between dataset

*The first two authors contributed equally to this work.
Ting Qiao, Jianbin Li and Yingjia Wang are with School of Control and Computer Engineering, North China Electric Power University, Beijing,102206, China (e-mail: {qiaoting,lijb87,wyj}@ncepu.edu.cn).
Yiming Li and Dacheng Tao are with College of Computing and Data Science, Nanyang Technological University, Singapore, 639798, Singapore (e-mail: {liyiming.tech, dacheng.tao}@gmail.com).
Leyi Qi is with the School of Cybersecurity, Northwestern Polytechnical University, Xi'an, 710072, China (e-mail: Leyi-Qi@outlook.com).
Junfeng Guo is with Department of Computer Science, University of Maryland, College Park, MD 20742, USA (e-mail: gjf2023@umd.edu).
Ruili Feng is with Alibaba Group, Hangzhou, 311100, China (e-mail: ruilifengustc@gmail.com).
Corresponding Author(s): Yiming Li and Jianbin Li.

watermarking and model watermarking. In particular, a few pioneering research [22], [23], [24] showed that we can achieve certified model watermarks that are robust against any parameter perturbations within a certain region by introducing random noise to the model parameters to limit the side effects in the worst-case scenario. Accordingly, an intriguing and critical question arises: *Could we also achieve certified watermark against image-level noises in the verification process of dataset ownership verification?*

The answer to the aforementioned question is in the positive, although we cannot directly generalize existing methods in certified model watermarks. This is mostly because they focused on the parameter space rather than the sample space, requiring defenders to obtain gradients $w.r.t.$ model parameters or even customize the whole training process. In this paper, we make the first attempt to design certified dataset watermarks to provide robustness guarantees for dataset ownership verification. Our method (dubbed 'CertDW') ensures that ownership can be reliably verified as long as the image-level watermark perturbation during the inference process satisfies certain conditions ($e.g.$, constrained pixel-level perturbation). In general, inspired by conformal prediction [25], [26], which is a statistical technique used to create prediction sets with assured coverages, we introduce two statistics, dubbed 'principal probability (PP)' and 'watermark robustness (WR)', to measure the distribution (instead of a probability value) in predicting the target label of benign samples and watermarked samples, respectively. Specifically, principal probability is defined as the upper bound of the probability that benign samples are consistently predicted as their ground-truth label under a noise distribution, while watermark robustness is defined as the lower bound of the probability that watermarked samples are consistently predicted as the target label under a noise distribution. In particular, we prove there is a lower bound on their gap if the sample-level perturbations on verification samples are within a certain range. As such, the suspicious model can be regarded as trained on the protected watermarked dataset (without authorization), if its WR value is significantly larger than the PP value of a validation model that is independently trained on a watermark-free dataset.

In practice, our method consists of three main steps. In the first step, we estimate the PP value by selecting the maximum value of the average prediction distribution (PD) computed across classes, obtained by introducing random noise multiple times to several benign samples. The PD hereby represents the probability distribution for each class predicted by the benign model when random noise is added to benign testing samples. In the second step, we estimate the WR value by selecting the minimum value of the probability distribution for the suspicious model predicting the target class, obtained by introducing random noise multiple times to several watermarked samples. In the third step, we calculate the PP values of multiple benign models (dubbed 'calibration set') and count the number of their values that are larger than WR. As as long as this number is sufficiently larger than a proportion of the size of calibration set, the suspicious model will be denoted as being trained on the protected dataset. We use multiple instead of solely one validation model for ownership verification to further reduce the randomness of model selection.

In summary, our contributions are four-fold, as follows: **(1)** We revisit existing dataset ownership verification (DOV) methods and reveal that their verification performance may degrade sharply when noises are incorporated during the inference process. **(2)** We make the first attempt to design certified dataset watermarks to provide robustness guarantees for dataset ownership verification based on two introduced statistics ($i.e.$, principal probability and watermark robustness). **(3)** We theoretically analyze the robustness guarantee and its conditions of our certified dataset ownership verification. **(4)** We conduct experiments on benchmark datasets to validate the effectiveness of our method and its resistance to potential adaptive attacks.

2 Related Works

2.1 Backdoor Attacks

Backdoor attacks are an emerging research field primarily targeting the training phase of deep neural networks (DNNs) [27]. In such attacks, an adversary maliciously manipulates a subset of training samples to implant a backdoor into the victim model, establishing a latent association between an adversary-specified trigger pattern and a target label. The compromised model performs normally when predicting benign samples, but once an input contains the trigger pattern, its predictions are maliciously altered, posing significant security risks to DNN-based applications. Generally, existing backdoor attacks can be categorized into three main types, based on the adversary's capabilities: **(1)** poison-only attacks [28], [29], [30], **(2)** training-controlled attacks [31], [32], [33], and **(3)** model-modified attacks [34], [35], [36]. Specifically, poison-only attacks can only manipulate the training dataset but cannot interfere with the training process; training-controlled attacks can modify both the training dataset and the training procedure ($e.g.$, altering the training loss function); while model-modified attacks inject backdoors by directly modifying model structures or parameters, making them more effective in both digital and physical environments. In this study, we mainly focus on poison-only attacks to leverage their unique properties to design watermarking technique for dataset ownership verification, aiming to protect public datasets. Other types of backdoor attacks can also be used for positive purposes [37], [38], [39], but this is out of the scope of this paper.

2.2 Data Protection

2.2.1 Classical Data Protection

Data protection is a classic and significant field of study, encompassing various aspects of data security with the goal of preventing unauthorized data usage and safeguarding personal data. Currently, encryption, digital watermarking, and privacy protection are the three main categories of conventional data protection methods. Specifically, encryption [7], [8], [9] protects sensitive data by fully or partially encrypting it, allowing only authorized users with the key to decrypt and utilize the data further. However, this method may limit the functionality of datasets ($e.g.$, accessibility) and is thus not suitable for protecting datasets that are already public. Digital watermarking [10], [11], [12] involves

embedding a pattern specified by the owner into the protected data as a watermark to assert ownership. Privacy protection [13], [14], [15] focuses on preventing the leakage of sensitive information during the training process through empirical methods [40], [41], [42] and certification methods [14], [43], [44]. These approaches often require access to more detailed training processes, which are not disclosed to users (especially dataset owners), thereby can not be directly used to protect the copyright of public datasets.

2.2.2 Dataset Ownership Verification

Dataset ownership verification (DOV) aims to verify whether a suspicious third-party model is trained on the protected dataset. To the best of our knowledge, this is currently the most widely used and effective method for protecting the copyright of open-source datasets [6]. Specifically, DOV strategies is a post-hoc auditing method that maintain the model's performance on benign test samples while inducing distinctive prediction behaviors on verification samples by introducing imperceptible watermarked samples into the original dataset to generate its watermarked version for release. Dataset owners verify ownership by checking whether the suspicious third-party model exhibits these distinctive prediction behaviors, all within a black-box verification setting. Current DOV methods [16], [17], [18] are primarily implemented through poison-only backdoor attacks or by watermarking unprotected benign datasets through other approaches [19], [20], [21]. For instance, [16] employed poisoned-label backdoor attacks, whereas [18] adopted clean-label backdoor attacks for dataset watermarking. Li *et al.* [17] initially discussed the 'harmlessness' requirement of DOV, stating that dataset watermarks should not introduce new security risks to models trained on protected datasets, and proposed the concept of untargeted backdoor watermarks. Recently, Guo *et al.* [19] further explored the definition of harmlessness, using hardly-generalized domain as watermarked samples to avoid introducing any new vulnerabilities. Additionally, Wei *et al.* [20] designed a scalable clean-label backdoor-based dataset watermark for point clouds, capable of watermarking samples from all classes. Most recently, Li *et al.* [21] proposed the first copyright protection method for personalized text-to-image diffusion models. However, existing DOV lack quantitative research and theoretical guarantees on the robustness of dataset watermarking, and thus may be vulnerable to future advanced adaptive attacks.

2.3 Certified Robustness

Certified robustness [45], [46], [47] ensures that a model produces the desired output (such as the correct label) when adversarial perturbations applied to the input remain within a certain region. Initially introduced for certifying classifiers against adversarial examples, the most classical technique for achieving certified robustness is randomized smoothing [48], [49]. This method works by adding random (Gaussian) noise to a given test input and then using the classifier to predict the final output based on the noisy inputs.

Besides ensuring certified adversarial robustness, a few pioneering recent studies [22], [23], [24] focused on the robustness certification of model watermarking to protect model ownership, achieving this by adding random noise to model parameters. These methods can provide theoretical guarantees in the worst-case scenario in weight-level perturbations, ensuring that the watermark remains non-removable, as long as the perturbation in the model parameters remains within a certain region. However, this protection method is still primarily used for safeguarding model copyrights, focusing on the parameter space rather than the sample space, which often requires additional training details ($e.g.$, gradients or even the entire training process). As such, existing certified model watermarking methods cannot be directly generalized to achieve certified dataset watermarking against image-level noise in the verification process of dataset ownership verification. How to design a certified dataset watermark for robust ownership verification remains blank and is worth further investigation.

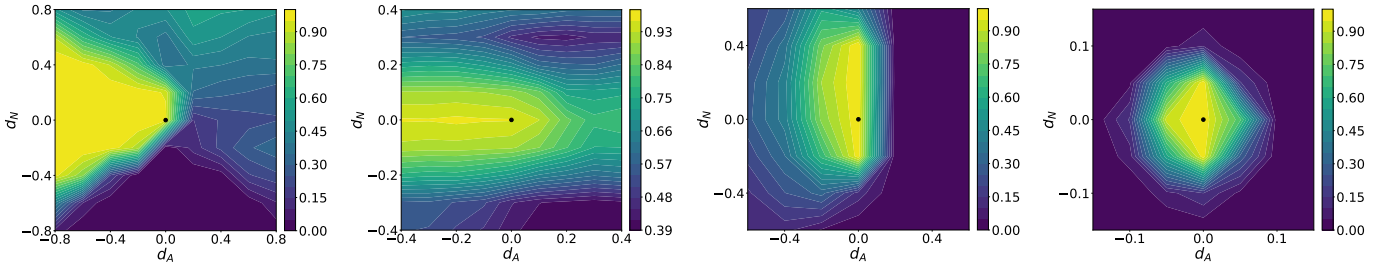
3 Revisiting Dataset Watermarking

We find that all existing dataset ownership verification (DOV) methods (implicitly) assume that the verification process is *faithful*, where the suspicious model will directly and exactly use the verification samples as input and return their results. However, the adversaries, $i.e.$, owners of the malicious model trained on the victim dataset, may try to circumvent ownership verification methods by adding perturbations to all verification samples before feeding them into the model in practice. In this section, we discuss whether the dataset watermarks of existing DOV methods are still effective in these cases. Before we describe our experiment design and observations, we first briefly review the general process of existing DOV methods.

The Main Pipeline of Existing DOV Methods. Let $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ denotes a vanilla training dataset for an image classification task with K classes, where $\boldsymbol{x}_n \in \mathcal{X} = [0,1]^{C\times W\times H}$ represents the image with $y_n \in \mathcal{Y} = \{1, 2, \cdots, K\}$ is its label. In the first stage of DOV ($i.e.$, dataset watermarking), the dataset owner creates a watermarked version of $\mathcal{D}$, denoted as $\mathcal{D}_w$. Specifically, $\mathcal{D}_w = \mathcal{D}_m \cup \mathcal{D}_r$, where $\mathcal{D}_m$ represents the modified version of samples from a small selected subset $\mathcal{D}_s$ of $\mathcal{D}$ ($i.e.$, $\mathcal{D}_s \subset \mathcal{D}$) and $\mathcal{D}_r$ contains the remaining benign samples ($i.e.$, $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_s$). The $\mathcal{D}_m$ is generated by the dataset-specified image generator $\boldsymbol{G}_X : \mathcal{X} \to \mathcal{X}$ and the label generator $\boldsymbol{G}_Y : \mathcal{Y} \to \mathcal{Y}$, $i.e.$, $\mathcal{D}_m = \{(\hat{\boldsymbol{x}}, \hat{y}) : \hat{\boldsymbol{x}} = G_X(\boldsymbol{x}), \hat{y} = G_Y(y), (\boldsymbol{x}, y) \in \mathcal{D}_s\}$. For example, in a BadNets-based DOV [16], [50], $\boldsymbol{G}_X = \boldsymbol{t} \odot \boldsymbol{x} + (1 - \boldsymbol{t}) \odot \boldsymbol{\delta}$ and $\boldsymbol{G}_Y = \hat{y}$, where $\boldsymbol{t} \in [0,1]^{C\times W\times H}$ is the trigger mask, $\boldsymbol{\delta} \in [0,1]^{C\times W\times H}$ is the trigger pattern, $\odot$ denotes the element-wise product, and $\hat{y}$ is the target label. In particular, $\gamma \triangleq \frac{|\mathcal{D}_m|}{|\mathcal{D}_w|}$ denotes the watermarking rate. In the second phase ($i.e.$, ownership verification), the dataset owners investigate whether a suspicious third-party model $f(\cdot;\boldsymbol{\theta}) : \mathcal{X} \to \mathcal{Y}$ was trained on the protected watermarked dataset $\mathcal{D}_w$ by querying it with verification samples under the black-box setting. For example, BadNets-based DOV used watermarked samples $\hat{\boldsymbol{x}}$ as verification samples and verified whether $f(\hat{\boldsymbol{x}};\boldsymbol{\theta}) = \hat{y}$.

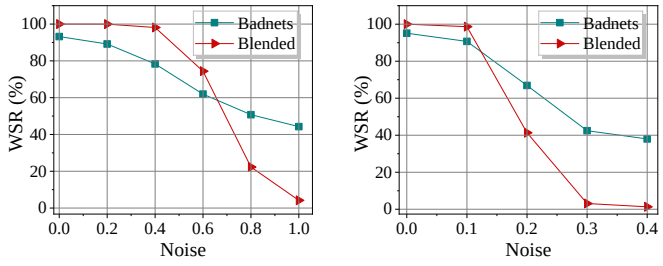
3.1 Impact of Unintentional Random Noises

In this section, we explore whether random noises that are not intentionally introduced by the adversaries will reduce



(a) GTSRB (BadNets) (b) CIFAR-10 (BadNets) (c) GTSRB (Blended) (d) CIFAR-10 (Blended)

Fig. 1: The performance (WSR) of watermarked samples in the sample space. d_N is is the direction of random noise, and d_A is the adversarial direction. '•' denotes the original watermarked sample. The first two columns presents the results under BadNets-based watermarks, while the last two columns show the results of Blended-based watermarks.



(a) GTSRB (b) CIFAR-10

Fig. 2: The impact of unintentional random Gaussian noise on the watermark success rate (WSR).

the performance of existing dataset watermarks.

Settings. We hereby use BadNets [50] and Blended [17] as the watermarking techniques on GTSRB [51] and CIFAR-10 [52] datasets for discussion. They are the most classical methods and the representative of visible and invisible backdoor watermarks, respectively. Specifically, we set the target label $\hat{y}$ as '1' and set the watermarking rate as 10% for both datasets. Following previous works [29], [53], we use a 3×3 black-and-white patch located at the lower-right corner of the image as the trigger pattern for BadNets-based watermarks. For the blended watermark, we use a Hello-Kitty trigger, blending it with the original images and setting the transparency parameter to 0.2. Besides, we exploit the winning model from the leaderboard [54] on GTSRB and a classical VGG-type model [55] on CIFAR-10. Both watermarks are implemented using BackdoorBox [56]. During the inference process, we add random noises sampled from a Gaussian distribution to each pixel of verification samples with different magnitudes.

Result. As shown in Figure 2, the watermark success rates (WSR) of both watermarks significantly decrease with the increase of noise magnitude. In particular, on the CIFAR-10 dataset, adding noise with a small magnitude ($e.g.$, 0.3) can reduce the WSR by 60%. This indicates that the performance of dataset watermarking degrades significantly in the presence of random noise, suggesting that existing dataset watermarking methods lack robustness and are highly vulnerable even to unintentional random noise.

3.2 Impact of Intentional Adversarial Perturbations

In this section, building on the analysis in Section 3.1, we further investigate whether adversarial perturbations intentionally introduced by adversaries will further degrade the performance of existing dataset watermarking methods.

Setting. In order to visualize the region around the watermarked samples, we measured the watermark success rate (WSR) on the panel spanned by two directions d_N and d_A. Specifically, d_N represents the unintentional random noise perturbation direction to erase watermark, $i.e.$, $d_N = \text{sign}(\mathcal{N}(0, \sigma^2 I))$, and d_A is the intentional adversarial perturbation direction to erase dataset watermark, $i.e.$, $d_A = \text{sign}(\nabla_{\boldsymbol{x}} \mathcal{L}(\theta, \boldsymbol{x}, y))$. We perturb the original watermarked samples along these two directions to explore the surrounding sample space and recorded the WSR of the neighboring samples. For comparison purposes, we define the new sample space as follows:

$$\hat{\mathcal{X}} \triangleq \{\hat{\boldsymbol{x}} + \varepsilon_N \cdot d_N + \varepsilon_A \cdot d_A | \varepsilon_N, \varepsilon_A \in \mathbb{R}\}, \quad (1)$$

where (ε_N, ε_A) are the coordinates along each direction, $\hat{\boldsymbol{x}}$ is the original watermarked sample, corresponding to the origin in the coordinate system (marked as the black circle for reference). Finally, we evaluate the changes in the watermark success rate within this sample space.

Result. As shown in Figure 1, we find that although unintentional random noise can already significantly reduce WSR within a certain range, introducing intentional adversarial perturbations leads to an even more dramatic decrease in WSR. For example, for blended watermarking, a perturbation as small as 0.15 can cause WSR to drop by over 80%. This suggests that if an adversary intentionally adds such perturbations, they can more effectively remove dataset watermark, leading to the failure of verification.

4 Methodology

4.1 Preliminaries

Threat Model. Following the classical settings of dataset ownership verification [6], [16], [19], we consider two parties ($i.e.$, the dataset owner and the adversary) in our threat model. The dataset owner can modify the original dataset to generate its watermarked version before releasing it. The dataset users (including adversaries) will use it to train their model, no matter in a legitimate or unauthorized manner. Accordingly, the dataset owner has neither the training details of the suspicious model nor its model parameters or intermediate results ($e.g.$, gradients). The owner can verify the ownership solely based on the prediction of the

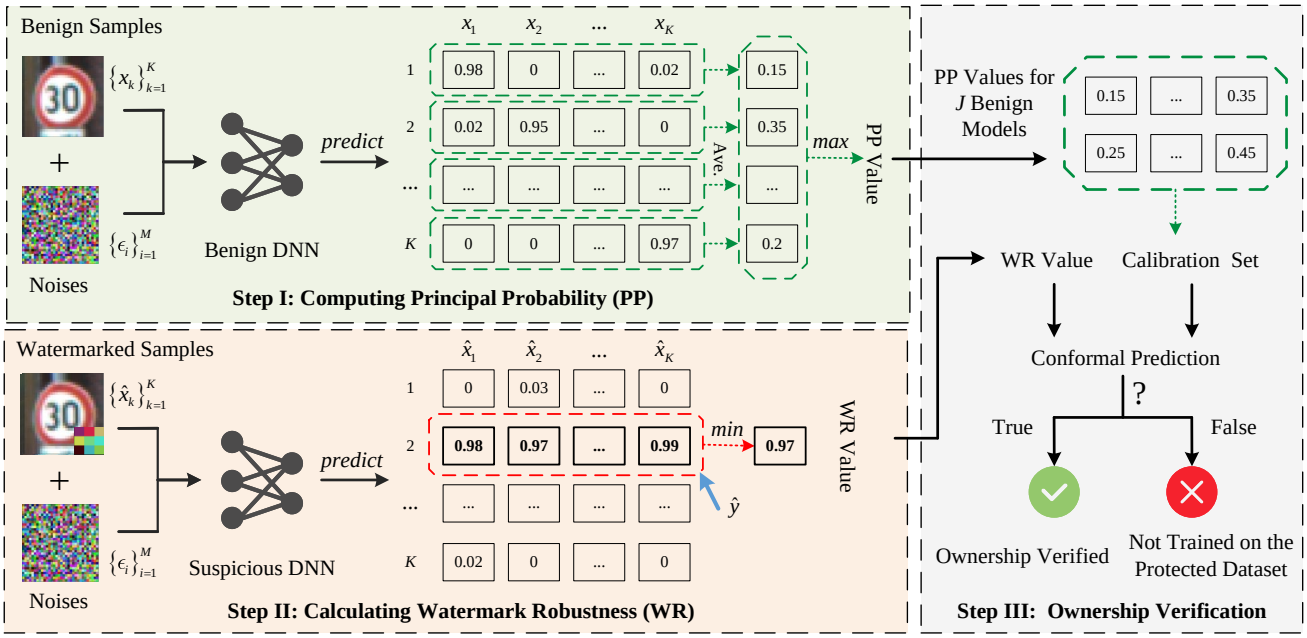


Fig. 3: The main pipeline of our CertDW consists of three core steps. In the first step, we randomly select one correctly predicted sample from each class. For each sample, we add M times of noise to predict these noisy versions via a benign model and calculate the frequency of prediction results on each category (termed as prediction distribution (PD)). After that, we estimate the PP value by selecting the maximum value of the average PD across classes; In the second step, we estimate the watermark robustness (WR) value following a similar approach in estimating the PP value. Specifically, we use watermarked instead of benign samples and the suspicious instead of the benign model for prediction. We obtain the WR value by selecting the minimum of the prediction probability on the defender-specified target label ($e.g.$, $\hat{y} = 2$). In the third step, we construct a calibration set by calculating PP values for J benign models. We then employ conformal prediction to count the number of values in the calibration set that are smaller than the WR value for ownership verification. If this count is sufficiently large, the suspicious model is deemed to have been trained on the protected dataset.

suspicious model ($i.e.$, under the black-box setting). In particular, different from previous works implicitly assumed that the adversaries will faithfully use verification samples to generate their predictions, we assume that they may deliberately circumvent verification by introducing malicious pixel-level perturbations to verification before feeding them for prediction. Arguably, this setup is more realistic and allows for a better assessment of the effectiveness.

Goal of Certified Dataset Watermark (Informal Definition). We hereby first provide a preliminary definition to briefly describe our goal. Its formal definition is in Section 5.1. Generally, a natural goal in defending against the aforementioned malicious acts ($i.e.$, adding noises before prediction) is to ensure that the prediction of the (suspicious) model on the verification sample remains unaffected by any malicious modification. In this case, even if malicious modifications are intentionally introduced in the verification process, the ownership verification remains effective since the prediction should remain unchanged. The goal of a certified dataset watermark is to ensure the previous property by designing a particular watermarking (and corresponding ownership verification) scheme (under certain conditions).

4.2 Overview of the Proposed Method

As demonstrated in Section 3, the performance of existing dataset watermarking methods significantly degrades as the intensity of unintentional random noise or intentional adversarial perturbations increases, often resulting in verification failures. To overcome this limitation, inspired by conformal predictions [25], [26], which exploits past observations to determine the precise confidence of new prediction, we propose the first certified dataset ownership verification. Specifically, it is designed based on two introduced statistical statistics: principal probability and watermark robustness, measuring the distribution in predicting the target label of benign and watermarked samples, respectively. In general, our method consists of three main steps: **(1)** computing the principal probability (PP), **(2)** calculating watermark robustness (WR), and **(3)** dataset ownership verification via conformal prediction, as shown in Figure 3. Their technical details are in the following subsections.

4.3 Computing the Principal Probability

In this step, we estimate the PP value by selecting the maximum value of the average prediction distribution (PD). In general, PD represents the probability distribution for each class predicted by the benign model when random noise is added to benign testing samples, defined as follows.

Definition 1 (Prediction Distribution (PD)). *For benign model $g(\cdot; \boldsymbol{w}) : \mathcal{X} \to \mathcal{Y}$ with parameters $\boldsymbol{w}$, define $p(\boldsymbol{x}|g_{\boldsymbol{w}}, \mathcal{P}_N) \in [0,1]^K$ as a vector representing the probability distribution over $K = |\mathcal{Y}|$ classes when random noise $\boldsymbol{\epsilon}$ is added to the input $\boldsymbol{x}$. The k-th entry ($k \in \mathcal{Y}$) of the* PD *is defined as:*

$$p_k(\boldsymbol{x}|g_{\boldsymbol{w}}, \mathcal{P}_N) = \mathbb{P}_{\boldsymbol{\epsilon}\sim\mathcal{P}_N}(\arg\max g(\boldsymbol{x}+\boldsymbol{\epsilon}; \boldsymbol{w}) = k), \quad (2)$$

where $\boldsymbol{\epsilon}$ is the noise sampled from a noise distribution[1] *$\mathcal{P}_N$, such as a Gaussian distribution, a uniform distribution, etc.*

In practice, PD is estimated using Monte Carlo by introducing random noise multiple times to the benign sample, recording the output count for each class, and using frequency to approximate probability. In other words,

$$p_k(\boldsymbol{x}|g_{\boldsymbol{w}}, \mathcal{P}_N) \approx \frac{1}{M}\sum_{i=1}^{M}\mathbb{I}\{\arg\max g(\boldsymbol{x}+\boldsymbol{\epsilon}_M;\boldsymbol{w})=k\}, \quad (3)$$

where M is the number of sampled random noises and $\mathbb{I}\{\cdot\}$ denotes the indicator function.

However, the randomness in selecting the benign samples may significantly impact the results. To reduce its side-effects, for a given benign model $g(\cdot;\boldsymbol{w})$, we independently sample K correctly predicted samples from each of the K classes, denoted as $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K$. We then calculate the final prediction distribution (PD) for each of these K samples individually and average them by class.

Given the estimated prediction distribution, we then calculate the principal probability (PP) as follows:

Definition 2 (Principal Probability (PP))**.** *Consider a domain with $K = |\mathcal{Y}|$ classes under the smoothing distribution $\mathcal{P}_N$. The* PP *for a benign model $g(\cdot;\boldsymbol{w})$ is defined as*

$$P(g_{\boldsymbol{w}}, \mathcal{P}_N) = \|\frac{1}{K}\sum_{k=1}^{K} p(\boldsymbol{x}_k|g_{\boldsymbol{w}}, \mathcal{P}_N)\|_\infty, \quad (4)$$

where $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K$ are K independent random samples satisfying $\arg\max g_{\boldsymbol{w}}(\boldsymbol{x}_k) = k$ $(k \in \{1, \cdots, K\})$.

In general, computing PP across more samples per class usually yield a similar value, therefore, using K instead of more samples is sufficient for its estimation.

4.4 Calculating the Watermark Robustness

In this step, we estimate the WR value by selecting the minimum probability from the distribution of the suspicious model's predictions for the target class. Similar to Section 4.3, to mitigate the impact of randomness in sample selection on the final result, we use multiple watermarked samples to compute the probability for the suspicious model. Specifically, given a suspicious model $f(\cdot;\boldsymbol{\theta})$, we independently sample K correctly predicted samples from each of the K classes, denoted as $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K$. We then embed a trigger $\boldsymbol{\delta}$ and assign the designated target label $\hat{y}$ to construct K watermarked samples. Finally, we compute the watermark robustness (WR) as follows:

Definition 3 (Watermark Robustness (WR))**.** *Consider a dataset watermark with a trigger $\boldsymbol{\delta}$ and a target class $\hat{y} \in \mathcal{Y}$ against a suspicious model $f(\cdot;\boldsymbol{\theta})$. For each sample $\boldsymbol{x} \in \mathcal{X}$, under a noise distribution $\mathcal{P}_N$, the WR for the watermarked sample $\hat{\boldsymbol{x}}$ (i.e., $\boldsymbol{x}+\boldsymbol{\delta}$) can be defined based on the PD as:*

$$W(f_{\boldsymbol{\theta}}, \mathcal{P}_N) = \min_{k=1,\cdots,K} p_{\hat{y}}(\hat{\boldsymbol{x}}_k|f_{\boldsymbol{\theta}}, \mathcal{P}_N), \quad (5)$$

1. The noise distribution is also commonly called 'smoothing distribution' since it is used for randomized smoothing.

where $\hat{\boldsymbol{x}}_k = \boldsymbol{x}_k + \boldsymbol{\delta}$, and $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K$ are K independent random samples satisfying $\arg\max g_{\boldsymbol{w}}(\boldsymbol{x}_k) = k$ $(k \in \{1, \cdots, K\})$.

Remark 1. *For suspicious model $f(\cdot;\boldsymbol{\theta})$, the $\hat{y}$-th entry of the PD (on $\hat{\boldsymbol{x}}$) is defined as $p_{\hat{y}}(\hat{\boldsymbol{x}}|f_{\boldsymbol{\theta}}, \mathcal{P}_N) = \mathbb{P}_{\boldsymbol{\epsilon}\sim\mathcal{P}_N}(\arg\max f(\boldsymbol{x}+\boldsymbol{\delta}+\boldsymbol{\epsilon};\boldsymbol{\theta}) = \hat{y})$. Selecting the smallest probability from K classes better reflects the watermark's robustness against noises, ensuring to the greatest extent that the watermark does not fail. This watermark robustness (WR) will be theoretically justified in the subsequent parts (see Section 5.2), which present the minimum conditions required to satisfy a certified dataset watermark.*

4.5 Ownership Verification via Conformal Prediction

In this step, inspired by conformal prediction, we calculate the PP values of multiple benign models (dubbed 'calibration set') and count the number of their values that are larger than WR for ownership verification. If this count is sufficiently larger than a certain proportion of the calibration set size, the suspicious model is deemed to have been trained on the protected dataset. We hereby exploit multiple benign models instead of a single one for verification to reduce the side effects of randomness in model selection.

Specifically, we first train J benign models following the method in Section 4.3 and calculate the PP values of them individually to construct a calibration set, denoted as $P_C(g_{\boldsymbol{w}}, \mathcal{P}_N) = \{P_C^1(g_{\boldsymbol{w}}, \mathcal{P}_N), \ldots, P_C^J(g_{\boldsymbol{w}}, \mathcal{P}_N)\}$. In particular, the calibration set is composed of PP values calculated by benign models trained on a dataset instead of the actual data distribution. As such, the PP values calculated on these benign models may exhibit distribution shifts, particularly with a higher sample variance and heavy tails with many outliers in the calibration set. Directly using this calibration set for conformal prediction may lead to overly conservative verification thresholds. To alleviate this problem, we need to filter out a certain proportion of larger outliers ($e.g.$, by outlier detection). Finally, we exploit conformal prediction based on the WR value of the suspicious model and the PP values in the calibration set to calculate the p-value for ownership verification, as follows.

Proposition 1 (Dataset Ownership Verification via Conformal Prediction)**.** *Let P_C and W denote the PP values of $P_C(g_{\boldsymbol{w}}, \mathcal{P}_N)$ and the WR values over $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$, respectively. The $P_C(g_{\boldsymbol{w}}, \mathcal{P}_N)$ and $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ is estimated based on Definition 2 and Definition 3, respectively. The p is defined by*

$$p = \frac{1+\min\left\{\sum_{j=1}^{J}\mathbb{I}\{P_C^j<W\}, J-m\right\}}{J-m+1}, \quad (6)$$

where J is the size of the calibration set, m represents the number of outliers in the calibration set, and $\mathbb{I}\{\cdot\}$ is an indicator function whose value is 1 if $P_C^j < W$ and 0 otherwise. We claim that the suspicious model is trained on the protected dataset if and only if $p \geq 1-\alpha_0$, where α_0 ($e.g.$, 0.05) denotes the chosen significance level (with $1-\alpha_0$ known as the confidence level).

In practice, $m = \kappa \cdot J$, where κ is a hyper-parameter indicating the proportion of filtering ($e.g.$, $\kappa = 0.2$).

Remark 2. *The previous condition of unauthorized training, $i.e.$, $p \geq 1-\alpha$ can be re-formulated as*

$$W(f_{\boldsymbol{\theta}}, \mathcal{P}_N) > P_C^{(J-m-\lfloor\alpha_0(J-m+1)\rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N), \quad (7)$$

where $P_C^{(j)}(g_{\boldsymbol{w}}, \mathcal{P}_N)$ denotes the j-th smallest element in $P_C(g_{\boldsymbol{w}}, \mathcal{P}_N)$. In particular, $P_C^{(J-m-\lfloor \alpha_0(J-m+1) \rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N)$ is called 'calibration threshold' in this paper.

5 Theoretical Analyses of our CertDW

In this section, we provide theoretical analyses of our CertDW proposed in Section 4. Before presenting the theory, we first define (sample-level) certified dataset watermarking. According to this definition, we propose a general theoretical framework applicable to various noise distributions based on Neyman-Pearson lemma [57]. Besides, we also instantiate this framework with two classical smoothing distributions, *i.e.*, Gaussian and uniform distributions and derive their specific conditions for a better illustration.

5.1 Definition of Certified Dataset Watermarking

In this section, we first define the neighborhood based on R-bounded transformation and provide a rigorous formulation for dataset watermark perturbations. Based on this definition, as well as the watermark robustness in Definition 3, we present two necessary properties of certified dataset watermarking (see Definition 5), which facilitate the theoretical analysis in Sections 5.2. We provide the formal definition of certified dataset watermarking at the end.

Definition 4. *The R-bounded transformation-based neighborhood set of the example $(\boldsymbol{x}, y)$, i.e., $\mathcal{F}_{R,T}(\boldsymbol{x}, y)$, is defined as:*

$$\mathcal{F}_{R,T}(\boldsymbol{x}, y) = \{(T(\boldsymbol{x}), \hat{y}) \mid \mathrm{dist}(T(\boldsymbol{x}), \boldsymbol{x}) \leq R\}, \tag{8}$$

where $T(\boldsymbol{x}) : \mathcal{X} \to \mathcal{X}$ is a sample-wise transformation, $\hat{y} \in \mathcal{Y}$ is the defender-specified target class, and $\mathrm{dist}(\cdot, \cdot)$ is a predefined distance metric (e.g., ℓ_p-norm). Here, $R \geq 0$ denotes the maximum perturbation magnitude of the dataset watermarking, representing the upper bound for perturbation strength.

Remark 3. *The set $\mathcal{F}_{R,T}$ is a general form, adapting to various widely used perturbation bounds by selecting appropriate transformation functions T and distance metrics. In this paper, we mainly focus on the pixel-level additive transformation with ℓ_p-norm ($1 \leq p \leq \infty$). It is worth noting that we need to assign an R so that $\mathrm{dist}(T(\boldsymbol{x}), \boldsymbol{x}) \leq R$ holds for all K selected samples $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K\}$ for certified dataset watermarks. For convenience, for each $k \in \{1, \ldots, K\}$, we define $\boldsymbol{r}_k = T(\boldsymbol{x}_k) - \boldsymbol{x}_k$. Accordingly, R can be the supremum of $\{\mathrm{dist}(T(\boldsymbol{x}_k), \boldsymbol{x}_k)\}_{k=1}^K$, denoted by $R = \max_{k=1,\cdots K} \|\boldsymbol{r}_k\|$.*

Based on Definition 4, we hereby provide two necessary properties that a certified dataset watermark should satisfy.

Definition 5 (Two Necessary Properties of Certified Dataset Watermarks)**.** *Let $\{\boldsymbol{x}_k\}_{k=1}^K$ be K independent random benign samples satisfying $\arg\max g_{\boldsymbol{w}}(\boldsymbol{x}_k) = k$, and $\boldsymbol{\epsilon}$ be the noise sampled from a noise distribution $\mathcal{P}_N$. Consider the watermarked version of $\boldsymbol{x}$ (i.e., $\boldsymbol{x} + \boldsymbol{r}$) with a defender-specified target class $\hat{y}$ against a watermarked model $f(\cdot; \boldsymbol{\theta})$.*

- ***(Transformation-based) Watermark Robustness (WR):** WR is defined as the lower bound of the probability that watermarked samples are consistently predicted as the target label under the given noise distribution, as shown in Definition 3. We hereby re-formulate it (by replacing $\boldsymbol{\delta}$ with $\boldsymbol{r}$) based on Remark 3, as follows:*

$$W(f_{\boldsymbol{\theta}}, \mathcal{P}_N) = \min_k \mathbb{P}(\arg\max f(\boldsymbol{x}_k + \boldsymbol{r}_k + \boldsymbol{\epsilon}) = \hat{y}). \tag{9}$$

- ***(R–functionality) Stability:** Given that the watermark transformation is constrained within R (i.e., $\|\boldsymbol{r_k}\|_2 \leq R$), it is defined as the lower bound of the probability that benign samples $\boldsymbol{x}$ are consistently predicted as the target label under a noise distribution, as follows:*

$$S(f_{\boldsymbol{\theta}}, \mathcal{P}_N) = \min_k \mathbb{P}(\arg\max f(\boldsymbol{x}_k + \boldsymbol{\epsilon}) = \hat{y}). \tag{10}$$

In general, $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ and $S(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ measure the resistance of the dataset watermark to noises and watermark-removal attacks (as well as noises) to verification samples, respectively. We consider both two properties instead of solely the watermark robustness to approximate the worst-case scenario where the malicious dataset user removes the trigger pattern somehow (instead of simply adding noises) during the inference process. Accordingly, a dataset watermark is certified robust if its two property values are both sufficiently large. Its formal definition is as follows.

Definition 6 ((Sample-level) Certified Robust Dataset Watermarking)**.** *We call a dataset watermark of the watermarked model $f_{\boldsymbol{\theta}}$ (under the smoothing distribution $\mathcal{P}_N$) is (τ-)certified robust if and only if $\min\{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N), S(f_{\boldsymbol{\theta}}, \mathcal{P}_N)\} > \tau$.*

Remark 4. *In this paper, to reduce the side effects of the randomness in selecting τ, we assign its value based on the calibration threshold (defined in Remark 2), which is calculated based on benign models via conformal prediction. Besides, following existing certified adversarial robustness/model watermarking [24], [49], we do not incorporate the 'utility' requirement that the watermarking should only have mild side effects on the watermarked models in predicting benign testing samples. Nevertheless, we will empirically verify it in our main experiments.*

5.2 A General Condition for Certified Watermarking

In this section, we aim to ensure dataset ownership verification while improving the performance of correct watermark verification. Before presenting the general conditions, we first define the statistical hypothesis testing, and then define the type-I and type-II errors in dataset watermarking. Based on these definition, we use the Neyman-Pearson lemma to derive the optimal likelihood ratio test, which in turn leads to the general conditions for certified dataset watermarking.

Definition 7 (Statistical Hypothesis Testing)**.** *Statistical hypothesis testing is a decision-making problem that involves determining whether a proposed hypothesis is correct. Formally, the decision is based on the realized values of a random variable X, whose distribution is known to be either H_0 (the null hypothesis) or H_1 (the alternative hypothesis). Given a sample $\boldsymbol{x} \sim \mathcal{X}$, a random test ϕ can be modeled as a function $\phi : \mathcal{X} \longrightarrow [0, 1]$, which rejects the null hypothesis with probability $\phi(\boldsymbol{x})$ and accepts the null hypothesis with probability $1 - \phi(\boldsymbol{x})$.*

Definition 8 (Type-I/II Error in Dataset Watermarking)**.** *For testing the null hypothesis H_0 (i.e., training on watermarked dataset) with the alternative hypothesis H_1 (i.e., training on watermark-free dataset) regarding a model trained on the watermarked dataset, the Type-I/II Errors are defined as follows:*

- *Type-I Error (β_1): The probability that watermarked samples are consistently identified as watermark-free (i.e., null hypothesis is true but rejected), as follow:*

$$\beta_1(\phi; H_0) = \mathbb{E}_{\boldsymbol{x}}(\phi(\boldsymbol{x}) | H_0 \text{ is true}). \tag{11}$$

- *Type-II Error (β_2): The probability that clean samples are consistently classified as the target label (i.e., being regarded as watermarked samples), i.e., (i.e., null hypothesis is false but accepted), as follow:*

$$\beta_2(\phi; H_1) = \mathbb{E}_{\boldsymbol{x}}(1 - \phi(\boldsymbol{x})|H_1 \text{ is true}). \quad (12)$$

In practice, type-I error leads to the neglect of potential copyright infringement, while type-II error triggers falsely claim ownership. Arguably, type-I error may lead to more serious negative consequences than type-II error since dataset ownership verification could be the first step before legal forensics (which can avoid false positive charge). Accordingly, inspired by the optimal likelihood ratio test ϕ^* introduced by Neyman-Pearson lemma [57], we aim to minimize the occurrence of type-II errors while controlling type-I error under a small threshold. Formally, we set the significance level α_1 as the maximum acceptable probability for type-I error, as follows:

$$\beta_1(\phi^*; H_0) = \alpha_1, \quad \beta_2(\phi^*; H_1) = \beta_2^*(\alpha_1; H_1), \quad (13)$$

where $\beta_2^*(\alpha_1; H_1) = \inf\{\beta_2(\phi; H_1) \mid \beta_1(\phi; H_0) \leq \alpha_1\}$.

By combining the above Definition 7-8 with the optimal likelihood ratio test, $i.e.$, Eq. (13), we can derive the following general condition (14) of certified robust dataset watermarking. Its proof is provided in Appendix.

Theorem 1 (**General Condition of Certified Dataset Watermarking**). *Given $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ and $S(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ that are estimated based on Eq. (9) and (10) in Definition 5 for a watermarked model, respectively. Dataset ownership is guaranteed to be verified if the optimal type-II errors, for testing the null $\mathcal{P}_N + \boldsymbol{r} \sim H_0$ against the alternative $\mathcal{P}_N \sim H_1$, satisfy the following condition:*

$$\beta_2^*(1 - W(f_{\boldsymbol{\theta}}, \mathcal{P}_N), H_1) > P_C^{(J-m-\lfloor \alpha_0(J-m+1) \rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N), \quad (14)$$

where $P_C^{(j)}(g_{\boldsymbol{w}}, \mathcal{P}_N)$ denotes the j-th smallest element in $P_J(g_{\boldsymbol{w}}, \mathcal{P}_N)$. α_0, J and m are defined as in Proposition 1.

Remark 5. *Different smoothing distributions lead to different robustness boundaries for various norms. For example, Gaussian noise results in robustness boundaries within the ℓ_2 norm, while uniform noise may lead to boundaries for other ℓ_p norms.*

In general, Theorem 1 establishes the optimal likelihood ratio test using the Neyman-Pearson lemma. As long as $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ is sufficiently large, α_1 will be controlled below a small threshold, and the type-II error is minimized. Besides, it is sufficient to ensure that the minimized value of type-II error ($i.e.$, the optimal type-II error $\beta_2^*(1 - W(f_{\boldsymbol{\theta}}, \mathcal{P}_N), H_1)$) exceeds the calibration threshold (defined in Remark 2), to guarantee dataset ownership verification. This highlights the inherent trade-off between type-II error and certification performance, which plays a critical role in verifying dataset ownership through the watermark.

We hereby derive the robustness conditions under two specific noise/smoothing distributions ($i.e.$, Gaussian and uniform distributions) as examples for a better illustration.

Example 1 (Robustness Conditions under Gaussian Distribution). *Let the noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$. Given $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ that is estimated based on Eq. (9) in Definition 5 for the watermark model's (transformation-based) WR. Let R denotes the maximum perturbation magnitude of the dataset watermark, as defined in Definition 4. Dataset ownership verification is guaranteed if and only if $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ satisfies the following condition:*

$$W(f_{\boldsymbol{\theta}}, \mathcal{P}_N) > \Phi(\frac{R}{\sigma}) + P_C^{(J-m-\lfloor \alpha_0(J-m+1) \rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N), \quad (15)$$

where Φ is the cumulative distribution function (CDF) of the standard Gaussian distribution.

Example 2 (Robustness Conditions under Uniform Distribution). *Let the noise $\boldsymbol{\epsilon} \sim \mathcal{U}([e, h])$. Given $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ that is estimated based on Eq. (9) in Definition 5 for the watermark model's (transformation-based) watermark robustness. Let R denotes the maximum perturbation magnitude of the dataset watermark, as defined in 4. Dataset ownership verification is guaranteed if and only if $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ satisfies the following condition:*

$$W(f_{\boldsymbol{\theta}}, \mathcal{P}_N) > P_C^{(J-m-\lfloor \alpha_0(J-m+1) \rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N) + 1 - \prod_{k=1}^{K} (1 - \frac{R}{h-e})_+, \quad (16)$$

where $(\boldsymbol{x})_+ = \max\{0, \boldsymbol{x}\}$.

Remark 6. *We have some critical observations about these examples to get intuition on the robustness condition (15)-(16):*

- *As shown in Eq. (15)–(16), watermarks with a larger WR are more likely to be guaranteed verification when the watermark perturbation size ($i.e.$, R) is fixed. Conversely, if the WR is fixed, dataset watermarks with smaller perturbation sizes are more likely to be guaranteed verification.*
- *The major distinction between certified dataset watermark verification and certified backdoor robustness is that the former provides verification guarantees for 'strong' dataset watermarks, while the latter prevents the learning of triggers during training. These two types of certification indicate that a backdoor watermark is either strong enough to be 'detectable' or weak enough to be removed.*

6 EXPERIMENTS

6.1 Main Settings

Dataset and Model Selection. We conduct experiments on the GTSRB [51] and CIFAR-10 [52] datasets. We adopt the same model architectures as those described in Section 3, following their standard train-test splits.

Baseline Selection. In this paper, we mainly compare our CertDW method to its variant without using conformal prediction (dubbed 'CertDW-V') since there is currently no certified dataset watermark. It sets a threshold by controlling the same false positive rate ($i.e.$, the performance of the independent model predicting the target label). Besides, we also provide the results of using vanilla watermarking techniques (dubbed 'Vanilla') and those without watermarking (dubbed 'independent') for reference.

Settings for Dataset Watermarking. Following the previous work [16], we adopt two backdoor watermark methods: BadNets [50] and Blended [17]. We set the target label $\hat{y}$ as '1' and set the watermarking rate as 10% for both datasets. For the BadNets-based watermarking, we use a 3×3 random pixel patch placed at a random location for each watermark on both datasets. Although the trigger locations are generated randomly, once a trigger is assigned, its position within the watermark samples remains fixed. For the blended-based watermark, we blend a 3×3 random pixel patch with the original images (dubbed 'Blended (patch)'), setting the transparency parameter to 0.2. Additionally, we also use

TABLE 1: The performance (%) of dataset watermarking and dataset ownership verification on the GTSRB dataset. The former is measured by benign accuracy (BA) and watermark success rate (WSR), while the latter is measured by verification success rate (VSR) and watermark certification accuracy (WCA). We evaluate all methods under three different noise levels: 1.5, 2.5, and 3.5. In particular, we mark the best verification results in boldface.

Watermark↓	$\sigma\rightarrow$			1.5			2.5			3.5		
	Method↓, Metric→	BA	WSR	VSR	WCA	WSR	VSR	WCA	WSR	VSR	WCA	WSR
No Watermarking	Independent	**97.05**	0	6	0	0	12	0	0	12	0	0
BadNets	Vanilla	96.05	90.88	18	0	37.78	18	0	32.42	36	0	35.19
	CertDW-V	95.43	**93.22**	56	16	**53.31**	52	22	**54.07**	50	28	**52.97**
	CertDW	95.43	**93.22**	**88**	**28**	**53.31**	**72**	**48**	**54.07**	**72**	**54**	**52.97**
Blended (patch)	Vanilla	96.10	91.37	80	0	34.47	70	20	44.40	70	30	49.79
	CertDW-V	96.13	**97.99**	40	**6**	**68.88**	46	8	**95.85**	48	10	**98.28**
	CertDW	96.13	**97.99**	**82**	**6**	**68.88**	**78**	**22**	**95.85**	**76**	**36**	**98.28**
Blended (noise)	Vanilla	96.90	**99.92**	0	0	20.09	0	0	5.41	0	0	2.37
	CertDW-V	95.98	96.14	90	32	**66.99**	86	54	**68.63**	86	62	**68.65**
	CertDW	95.98	96.14	**98**	**66**	**66.99**	**96**	**80**	**68.63**	**96**	**90**	**68.65**

TABLE 2: The performance (%) of dataset watermarking and dataset ownership verification on the CIFAR-10 dataset. The former is measured by benign accuracy (BA) and watermark success rate (WSR), while the latter is measured by verification success rate (VSR) and watermark certification accuracy (WCA). We evaluate all methods under three different noise levels: 0.6, 1.2, and 1.8. In particular, we mark the best verification results in boldface.

Watermark ↓	$\sigma\rightarrow$			0.6			1.2			1.8		
	Method↓, Metric→	BA	WSR	VSR	WCA	WSR	VSR	WCA	WSR	VSR	WCA	WSR
No Watermarking	Independent	**82.55**	0	6	0	0	4	0	0	4	0	0
BadNets	Vanilla	81.19	**94.18**	62	0	80.30	28	14	72.54	44	14	62.78
	CertDW-V	81.06	93.91	64	20	**83.54**	64	24	**80.98**	54	24	**77.28**
	CertDW	81.06	93.91	**70**	**24**	**83.54**	**68**	**36**	**80.98**	**64**	**40**	**77.28**
Blended (patch)	Vanilla	80.46	**99.99**	40	10	61.50	50	10	62.30	50	10	57.61
	CertDW-V	80.19	99.49	42	16	**67.53**	44	20	**70.86**	42	22	**68.13**
	CertDW	80.19	99.49	**48**	**20**	**67.53**	**52**	**32**	**70.86**	**54**	**32**	**68.13**
Blended (noise)	Vanilla	82.06	99.45	80	70	90.99	80	70	90.22	70	60	90.37
	CertDW-V	80.03	**99.55**	**88**	70	**93.55**	84	78	**95.87**	80	62	**92.33**
	CertDW	80.03	**99.55**	**88**	**72**	**93.55**	**88**	**78**	**95.87**	**82**	**78**	**92.33**

a noise pattern applied to the entire image as the trigger for blended-based watermarks (dubbed 'Blended (noise)'). For each watermark, a random trigger is generated and embedded as a random perturbation $\boldsymbol{\delta}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{v}$, where $\|\boldsymbol{v}\|_2 \approx 0.6$. In particular, we use an average ℓ_2 norm of 1.4 for Blended (noise) on the CIFAR-10 dataset since it usually fails below this threshold. The pixel-level perturbation size is adjusted accordingly to satisfy the $\|\boldsymbol{v}\|_2$ constraint. For example, on the GTSRB dataset, when $\|\boldsymbol{v}\|_2 \approx 0.6$, the perturbation magnitude for each altered pixel is independently and randomly chosen within [40/255, 65/255] for all three channels. Moreover, we create 50 backdoor watermarks for each method on each dataset.

Settings for Dataset Verification. We reserve 5,000 samples from the test dataset of GTSRB and CIFAR-10 and train 100 benign models ($i.e.$, $J = 100$) following the standard training procedure, respectively. A proportion of filtered outliers $\kappa = 0.2$ is set to construct a calibration set for verifying dataset ownership. We train 50 independent models using the full benign training dataset for reference to estimate the false positive rate of our CertBW. During the verification process, we generate 1,024 random Gaussian noises for each input and calculate the WR and PP values using the Monte Carlo estimation method ($i.e.$, $M = 1024$). In conformal prediction, the significance level α_0 is set to 0.05.

Evaluation Metrics. We evaluate the performance in two aspects: dataset watermarking and ownership verification. For dataset watermarking, we use benign accuracy (BA) and watermark success rate (WSR) to assess the effectiveness of the dataset watermarks. Specifically, BA is defined as the model accuracy on the benign testing dataset, and WSR is defined as the accuracy on the watermarked testing dataset. For ownership verification, we evaluate the verification success rate (VSR) and watermark certification accuracy (WCA). Specifically, VSR is defined as the proportion of benign samples consistently predicted as the target label. In particular, for the watermarked model, it is defined by Eq. (10); for the independent model, the VSR corresponds to the false positive rate. WCA is defined as the proportion of watermarked samples that are guaranteed to be predicted as the target label by the watermarked model, $i.e.$, the proportion of watermark samples that fall into the certified region. The certified region is a two-dimensional region determined by inequality (15), which involves watermark robustness and the magnitude of the trigger perturbation (see Figure 4). Within the certified region, watermark samples are consistently recognized as the target label. Higher values of WCA and VSR indicate better performance of the verification method. Besides, during the verification phase, we also evaluate WSR to further demonstrate the robustness of our watermarking method.

6.2 Main Results

As shown in Tables 1-2, our CertDW watermarking has only a mild influence on the utility of watermarked models.

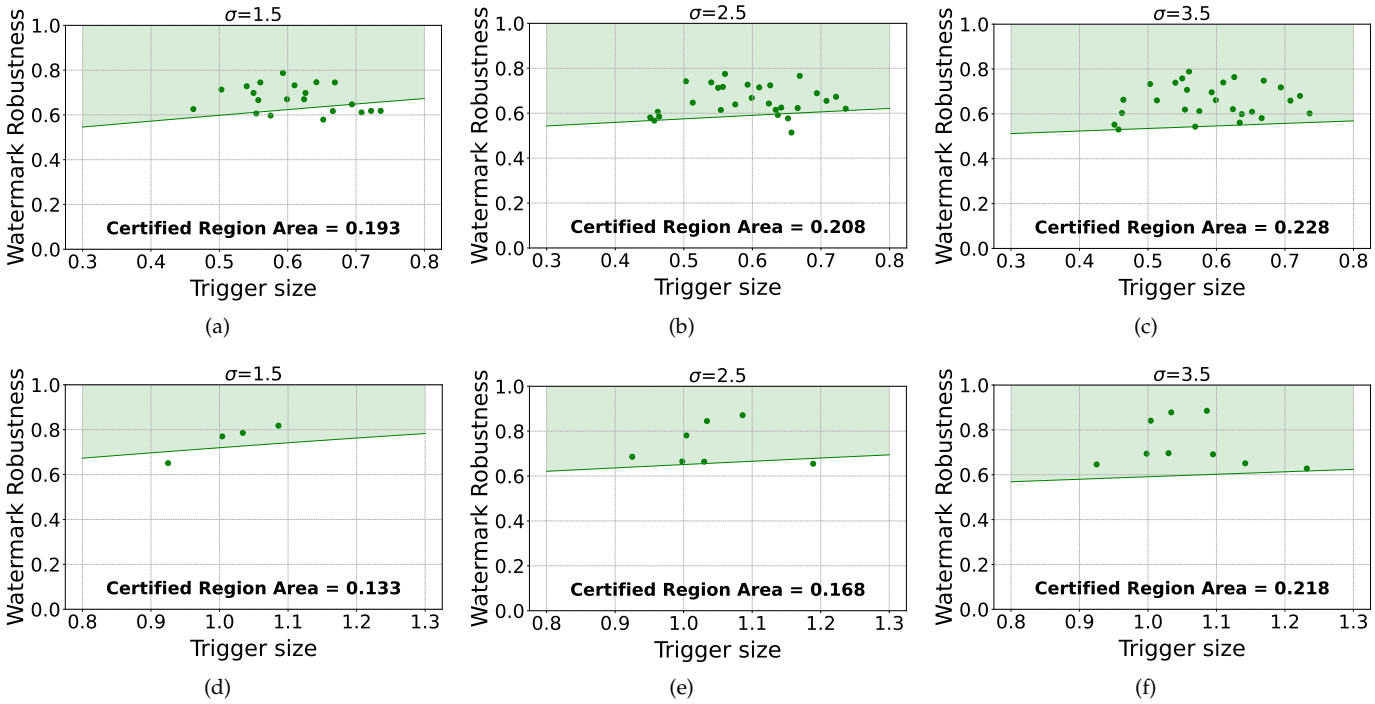


Fig. 4: Examples of the certified regions obtained with Gaussian smoothing distribution (with different standard deviations σ) on GTSRB with two different ranges of trigger sizes R. **First Row**: $R \in (0.3, 0.8)$; **Second Row**: $R \in (0.8, 1.3)$.

For instance, the watermark success rate (WSR) consistently remains above 90%, while the accuracy of benign samples decreases by no more than 2% (and in most cases, by less than 1%). This indicates that our method does not interfere with the normal use of the dataset. During the verification process, our method achieves superior verification performance compared to other baseline methods. Unlike traditional approaches, our method remains robust even as noise levels increase, with WSR staying above 60% in most cases. Although in some cases, the Vanilla method achieves a slightly higher VSR than our approach, its WCA is significantly lower. These experimental results strongly validate the effectiveness of our method.

Furthermore, we observe that the VSR ($i.e.$, the false positive rate, FPR) under independent models for both datasets remains at relatively low levels: the FPR for GTSRB is consistently below 12%, while for CIFAR-10, it is consistently below 6%. These results further verify the accuracy of our CertDW-based verification. Note that the FPR mentioned here differs from the type-II error described in Theorem 1. Specifically, the type-II error in Theorem 1 refers to the proportion of benign ($i.e.$, non-watermark) samples successfully marked as watermarked when evaluated under a watermark model, whereas the FPR discussed here refers to the VSR on benign samples evaluated under a independent model. As such, these two metrics represent distinct concepts and should not be conflated.

6.3 Analyzing the Certified Region of CertDW

In this section, we visualize the certified region of our method during the verification process and analyze the impact of trigger size on this region.

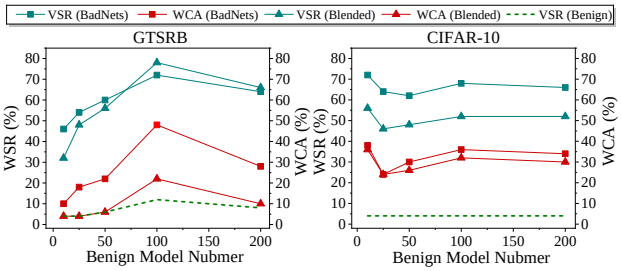
Specifically, we present examples of certified regions on the GTSRB obtained through Gaussian smoothing distribution with different standard deviations $\sigma \in \{1.5, 2.5, 3.5\}$, associated with two different ranges of trigger sizes ($i.e.$, $R \in (0.3, 0.8)$ and $R \in (0.8, 1.3)$). As shown in Figure 4, the shape of the certified region aligns with our theoretical results (see Remark 6), which are derived from Example 1. As the trigger size increases, both the number of dataset watermarks falling within the certified region and the certified region area gradually decrease, while the WR value increases. For instance, when the noise level is 1.5, the number of datasets falling within the certified region decreases from 13 to 3, and the certified region area decreases from 0.193 to 0.133, while the WR value increases from 58% to 68%. Additionally, as the noise level increases, the certified region area gradually improves. For example, when the trigger size is fixed at $(0.8, 1.3)$, and the noise increases from 1.5 to 3.5, the certified region area increases from 0.133 to 0.218. This indicates that dataset watermarks with higher WR and smaller trigger perturbation sizes are more likely to guarantee dataset ownership verification.

6.4 Ablation Study

We hereby discuss the impact of key hyper-parameters. For simplicity, we discuss each dataset by using a fixed noise level ($e.g.$, 2.5 on GTSRB and 1.2 on CIFAR-10).

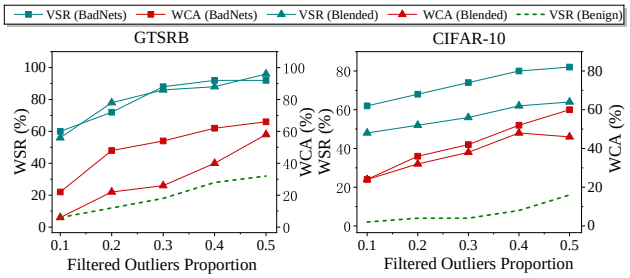
6.4.1 Impact of the Number of Benign Models

As shown in Figure 5, both the VSR and WCA increase with the number of benign models J. These results indicate that defenders can enhance the confidence in verification by increasing J. Particularly, when the number of benign models reaches 100, nearly all evaluated watermarks achieve high



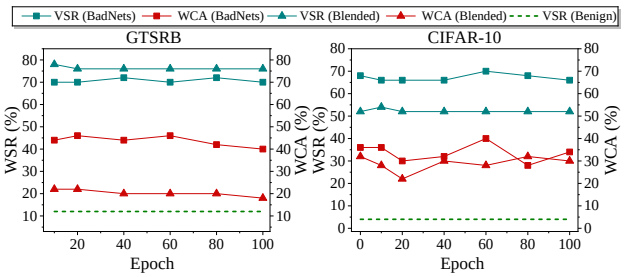
(a)

Fig. 5: Effects of the number of benign models.



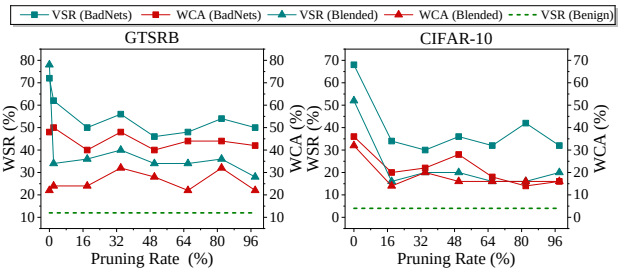
(a)

Fig. 6: Effects of the proportion of filtered outliers.



(a)

Fig. 7: The resistance to fine-tuning.



(a)

Fig. 8: The resistance to model pruning.

VSR and WCA. However, in most cases, excessive increases in the number of benign models lead to decreased VSR and WCA. This may be due to the fact that our benign models are trained using a dataset instead of the actual data distribution, and the data is not sufficient. As the number of models increases, this insufficiency becomes more pronounced, resulting in anomalously high PP values. In other words, dataset owners should determine the number of benign models based on their specific requirements.

6.4.2 *Impact of the Proportion of Filtered Outliers*

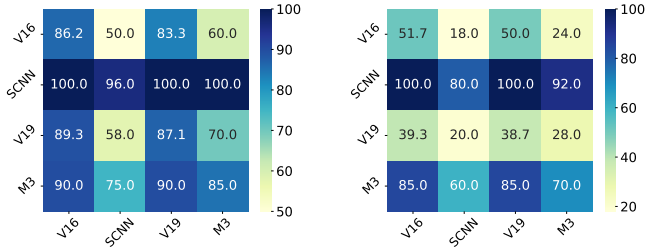
As shown in Figure 6, both the VSR and the WCA increase as the outlier filtering ratio κ increases. However, in most cases, the VSR of independent models ($i.e.$, FPR) also increases, meaning that the likelihood of incorrectly identifying an independent model as being trained on the protected dataset also rises. In other words, there is a trade-off between precision and recall here to some extent. In practice, dataset owners should also determine the value of κ based on their specific needs.

6.5 The Resistance to Potential Adaptive Attacks

In this section, we discuss the resistance of our method against two potential watermark-removal attacks, including fine-tuning [58] and model pruning [59].

The Resistance to Fine-tuning. Following the previous work [58], we fine-tune all layers of the CerDW-watermarked model use 10% of the benign samples from the original training set, with a learning rate of 0.001. The model is fine-tuning for a total of 100 epochs. As shown in Figure 7, throughout the fine-tuning process, both VSR and WCA remain stable to a large extent. These results indicate that fine-tuning only has a minor impact on our method.

The Resistance to Model Pruning. Following the previous work [59], we use 10% of the benign samples from the original training set to prune the latent representations of our watermark model ($i.e.$, the inputs to the fully connected layers). The pruning rates vary from 0% to 98% in each case. As shown in Figure 8, pruning initially leads to a significant drop in the VSR and WCA. However, subsequent changes become relatively stable, especially as the pruning rate approaches 98%, where the metrics show minimal further variation. Thus, even under high pruning rates, our method maintains certain levels of VSR and WCA, demonstrating its resilience to model pruning to some extent.

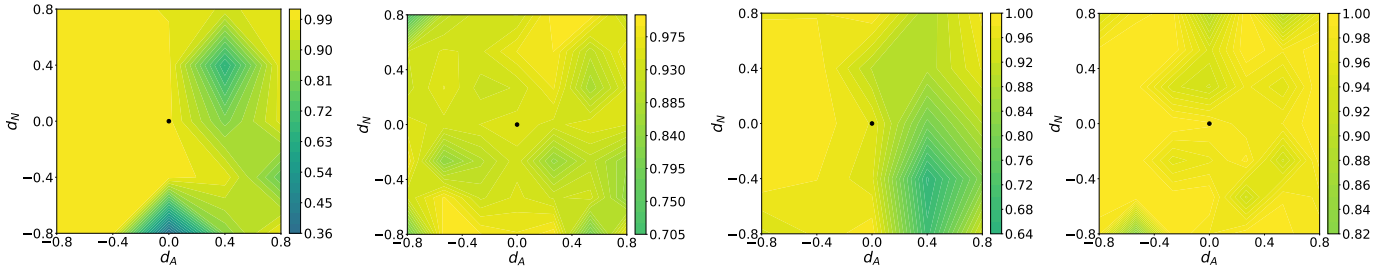


(a) VSR (b) WCA

Fig. 9: The performance of our method with different structures of benign and watermarked models on GTSRB. Row: benign models; Column: watermarked models.

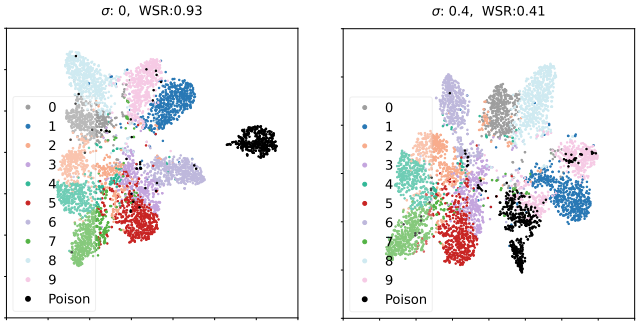
6.6 Model-level Transferability of CertDW

As described in Section 4.3, we use a pre-trained surrogate model to serve as the benign model. The experiments in Section 6.2 are conducted based on the setting that the benign model and the watermarked model share the same architecture. However, this assumption may not hold in practice, as dataset owners are typically unaware of the specific architectures exploited by dataset users (including adversaries). Accordingly, in this section, we analyze the



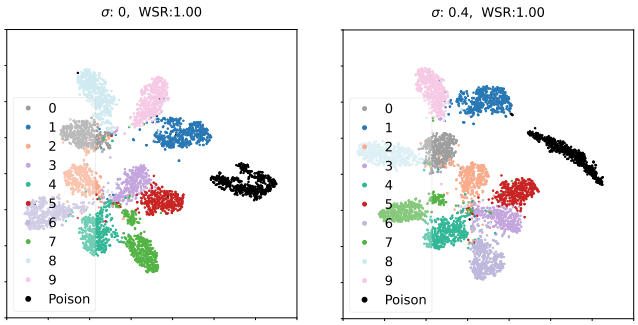
(a) GTSRB (BadNets) (b) CIFAR-10 (BadNets) (c) GTSRB (Blended) (d) CIFAR-10 (Blended)

Fig. 10: The performance (WSR) of our method's watermarked samples in the sample space. d_N is the random noise direction, and d_A is the adversarial direction. '•' denotes the original watermarked sample. The first two columns show results for BadNets-based watermarks, and the last two columns show results for Blended-based watermarks.



(a) (b)

Fig. 11: t-SNE visualization of feature distribution in the vanilla watermarked model with random noises.



(a) (b)

Fig. 12: t-SNE visualization of feature distribution in the our watermarked model with random noises.

effectiveness of our CertDW when benign and watermarked models have different architectures.

Specifically, we select four representative network architectures, including VGG-16, spacial-enhanced CNN (dubbed 'SCNN'), VGG-19, and MobileNetV3, on the GTSRB dataset for discussion. All other settings remain consistent with those illustrated in Section 6.2. As shown in Figure 9, our method remains effective across different model architectures, despite some performance fluctuations due to different model capacity. These results demonstrate that our CertDW does not rely on prior knowledge of the benign and watermarked model, making it a robust and practical approach for dataset ownership verification.

6.7 A Closer Look to the Effectiveness of CertDW

In this section, we intend to further explore the mechanisms behind the effectiveness of our CertDW. Specifically, we visualize the region around the watermarked samples in the sample space and feature space for in-depth discussion.

6.7.1 Visualizing in the Sample Space

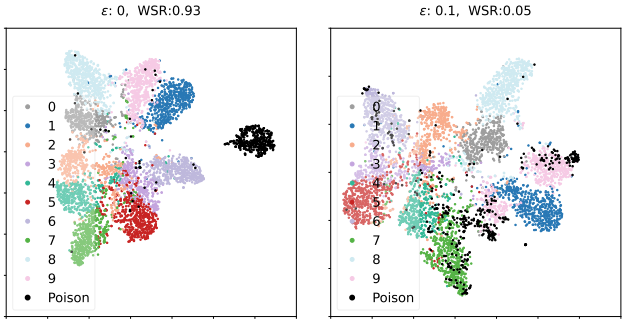
In this section, we use the same method as in Section 3.2 to demonstrate the effectiveness of our approach. As shown in the Figure 10, we find that our watermarked samples can maintain a high WSR (over 80%) in the direction of unintentional random noise. Although there is some performance degradation in the direction of intentional adversarial perturbations, our method still exhibits strong robustness and effectively prevents verification failure. For example, on the CIFAR-10 dataset, even with perturbation magnitudes of 8 in both directions, the WSR remains above 70%. This result significantly increases the difficulty for an adversary to intentionally add perturbations to verification/watermarked samples to completely remove the watermark.

6.7.2 Visualizing the Feature Space

To better understand our method's effectiveness, we adopt t-SNE [60] to visualize the feature distribution of watermarked samples evolves with unintentional noise and intentional adversarial perturbations.

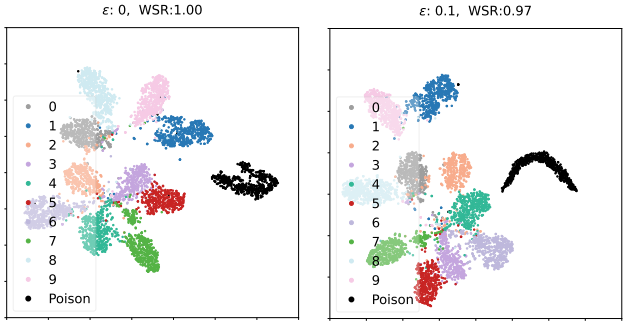
Features along with the Unintentional Random Noises. We visualized the impact of different random noise magnitudes at the early stages. As shown in Figure 11, at the initial stage of adding random noise, the representations of the watermarked samples quickly become very similar to those of the benign samples, leading to a significant reduction in the watermarking success rate. However, our method effectively maintains the watermarked samples within an independent cluster, ensuring a clear separation from the non-target clusters, as depicted in Figure 12.

Features along with the Intentional Adversarial Noises. To further demonstrate how hidden representations evolve along adversarial directions, we add adversarial perturbations of varying magnitudes to the watermarked samples. As shown in Figure 13, with minor perturbations, the representations of the watermarked samples quickly blend with



(a) (b)

Fig. 13: t-SNE visualization of feature distribution in the vanilla watermarked model with adversarial noises.



(a) (b)

Fig. 14: t-SNE visualization of feature distribution in the our watermarked model with adversarial noises.

those of the benign samples. In contrast, our method successfully maintains the watermarked samples in a distinct cluster, which remains distinctly separate from non-target clusters, as depicted in Figure 14.

6.8 The Analysis of Computational Complexity

In this section, we analyze the computational complexity of our CertDW, especially its dataset watermarking and ownership verification process, respectively.

The Complexity of Dataset Watermarking. Let N denotes the number of all training samples, and γ is the poisoning rate. Our computational complexity is $\mathcal{O}(\gamma \cdot N)$, since CertDW only needs to watermark a small subset of selected samples in this step. In general, the trigger size of these watermarks must satisfy the ℓ_2 norm constraint by adjusting pixel-level perturbation size, which is highly efficient.

The Complexity of Dataset Verification. Let J, W_a, I_n denote the number of benign models, watermarked models, and independent models, respectively. Dataset owner trains these models with computational complexities of $\mathcal{O}(J)$, $\mathcal{O}(W_a)$, and $\mathcal{O}(I_n)$, respectively. This process supports parallel processing. Additionally, we use the trained multiple benign models and watermarked models to calculate the PP value and WR value, which is highly efficient. For example, training benign models and watermarked models takes approximately 120 seconds and 180 seconds on the CIFAR-10 dataset, respectively. Computing PP and WR values requires only 3 seconds and 2 seconds, respectively. As such, calculating PP values and WR values is almost cost-free. Arguably, although training each benign model is time-consuming, it is generally acceptable, not to mention that we can use parallel computing to further accelerate it.

7 Conclusion

In this paper, we revisited existing dataset ownership verification (DOV) methods and revealed that their performance degrades sharply under both unintentional random noise and intentional adversarial perturbations. Based on our analysis, we proposed a certified dataset watermark (*i.e.*, CertDW) to provide robustness guarantees for dataset ownership verification. Inspired by conformal prediction, we introduced two statistical measures, *i.e.*, principal probability (PP) and watermark robustness (WR), based on the target label distributions of benign and watermarked samples. We proved that when sample-level perturbations remain within a certain range, there exists a lower bound between PP and WR values. We calculated the PP and WR values by introducing random noise to multiple benign and watermarked samples. As long as a suspicious model's WR is sufficiently larger than a proportion of multiple PP values calculated by several benign models, we can conclude that the suspicious model is trained on the protected dataset. Extensive experiments on benchmark datasets validate CertDW's effectiveness and its resilience to potential adaptive attacks. We hope our paper can provide a new perspective on reliable dataset ownership verification, to facilitate more trustworthy dataset sharing and circulation.

References

[1] X. Tang and Z. Li, "Video based face recognition using multiple classifiers," in *IEEE FG*, 2004, pp. 345–349.

[2] H. Qiu, D. Gong, Z. Li, W. Liu, and D. Tao, "End2end occluded face recognition by masking corrupted features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6939–6952, 2021.

[3] X. Yang, X. Jia, D. Gong, D.-M. Yan, Z. Li, and W. Liu, "Larnext: End-to-end lie algebra residual network for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11 961–11 976, 2023.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[5] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," in *NeurIPS*, 2022.

[6] L. Du, X. Zhou, M. Chen, C. Zhang, Z. Su, P. Cheng, J. Chen, and Z. Zhang, "Sok: Dataset copyright auditing in machine learning systems," in *IEEE S&P*, 2025.

[7] C. Thorpe, F. Li, Z. Li, Z. Yu, D. Saunders, and J. Yu, "A coprime blur scheme for data security in video surveillance," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 3066–3072, 2013.

[8] P. Martins, L. Sousa, and A. Mariano, "A survey on fully homomorphic encryption: An engineering perspective," *ACM Computing Surveys*, 2017.

[9] T. Chuman, W. Sirichotedumrong, and H. Kiya, "Encryption-then-compression systems using grayscale-based image encryption for jpeg images," *IEEE Transactions on Information Forensics and security*, vol. 14, no. 6, pp. 1515–1525, 2018.

[10] C.-T. Hsu and J.-L. Wu, "Hidden digital watermarks in images," *IEEE Transactions on image processing*, vol. 8, no. 1, pp. 58–68, 1999.

[11] A. K. Jain and U. Uludag, "Hiding biometric data," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 11, pp. 1494–1498, 2003.

[12] S. Baluja, "Hiding images within images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 7, pp. 1685–1697, 2019.

[13] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *CCS*, 2016.

[14] L. Zhu, X. Liu, Y. Li, X. Yang, S.-T. Xia, and R. Lu, "A fine-grained differentially private federated learning against leakage from gradients," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 11 500–11 512, 2021.

[15] Y. Li, S. Yang, X. Ren, L. Shi, and C. Zhao, "Multi-stage asynchronous federated learning with adaptive differential privacy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[16] Y. Li, M. Zhu, X. Yang, Y. Jiang, T. Wei, and S.-T. Xia, "Black-box dataset ownership verification via backdoor watermarking," *IEEE Transactions on Information Forensics and Security*, 2023.

[17] Y. Li, Y. Bai, Y. Jiang, Y. Yang, S.-T. Xia, and B. Li, "Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection," in *NeurIPS*, 2022.

[18] R. Tang, Q. Feng, N. Liu, F. Yang, and X. Hu, "Did you train on my dataset? towards public dataset protection with cleanlabel backdoor watermarking," *ACM SIGKDD Explorations Newsletter*, vol. 25, no. 1, pp. 43–53, 2023.

[19] J. Guo, Y. Li, L. Wang, S.-T. Xia, H. Huang, C. Liu, and B. Li, "Domain watermark: Effective and harmless dataset copyright protection is closed at hand," in *NeurIPS*, 2024.

[20] C. Wei, Y. Wang, K. Gao, S. Shao, Y. Li, Z. Wang, and Z. Qin, "Pointncbw: Towards dataset ownership verification for point clouds via negative clean-label backdoor watermark," *IEEE Transactions on Information Forensics and Security*, 2024.

[21] B. Li, Y. Wei, Y. Fu, Z. Wang, Y. Li, J. Zhang, R. Wang, and T. Zhang, "Towards reliable verification of unauthorized data usage in personalized text-to-image diffusion models," in *IEEE S&P*, 2025.

[22] A. Bansal, P.-y. Chiang, M. J. Curry, R. Jain, C. Wigington, V. Manjunatha, J. P. Dickerson, and T. Goldstein, "Certified neural network watermarks with randomized smoothing," in *ICML*, 2022.

[23] Z. Jiang, M. Fang, and N. Z. Gong, "Ipcert: Provably robust intellectual property protection for machine learning," in *ICCV*, 2023.

[24] J. Ren, Y. Zhou, J. Jin, L. Lyu, and D. Yan, "Dimension-independent certified neural network watermarks via mollifier smoothing," in *ICML*, 2023.

[25] V. Vovk, "Conditional validity of inductive conformal predictors," in *ACML*, 2012.

[26] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint arXiv:2107.07511*, 2021.

[27] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 5–22, 2022.

[28] X. Qi, T. Xie, Y. Li, S. Mahloujifar, and P. Mittal, "Revisiting the assumption of latent separability for backdoor defenses," in *ICLR*, 2023.

[29] Y. Gao, Y. Li, L. Zhu, D. Wu, Y. Jiang, and S.-T. Xia, "Not all samples are born equal: Towards effective clean-label backdoor attacks," *Pattern Recognition*, vol. 139, p. 109512, 2023.

[30] H. Cai, P. Zhang, H. Dong, Y. Xiao, S. Koffas, and Y. Li, "Toward stealthy backdoor attacks against speech recognition via elements of sound," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 5852–5866, 2024.

[31] Y. Li, H. Zhong, X. Ma, Y. Jiang, and S.-T. Xia, "Few-shot backdoor attacks on visual object tracking," in *ICLR*, 2022.

[32] X. Mo, Y. Zhang, L. Y. Zhang, W. Luo, N. Sun, S. Hu, S. Gao, and Y. Xiang, "Robust backdoor detection for deep learning via topological evolution dynamics," in *IEEE S&P*, 2024.

[33] H. Zhang, S. Hu, Y. Wang, L. Y. Zhang, Z. Zhou, X. Wang, Y. Zhang, and C. Chen, "Detector collapse: Backdooring object detection to catastrophic overload or blindness," in *IJCAI*, 2024.

[34] X. Qi, T. Xie, R. Pan, J. Zhu, Y. Yang, and K. Bu, "Towards practical deployment-stage backdoor attack on deep neural networks," in *CVPR*, 2022.

[35] J. Dong, Q. Han, Y. Li, T. Zhang, Y. Li, Z. Lai, C. Zhang, and S.-T. Xia, "One-bit flip is all you need: When bit-flip attack meets model training," in *ICCV*, 2023.

[36] S. Yang, J. Bai, K. Gao, Y. Yang, Y. Li, and S.-T. Xia, "Not all prompts are secure: A switchable backdoor attack against pre-trained vision transfomers," in *CVPR*, 2024.

[37] G. Gan, Y. Li, D. Wu, and S.-T. Xia, "Towards robust model watermark via reducing parametric vulnerability," in *ICCV*, 2023.

[38] M. Ya, Y. Li, T. Dai, B. Wang, Y. Jiang, and S.-T. Xia, "Towards faithful xai evaluation via generalization-limited backdoor watermark," in *ICLR*, 2024.

[39] Y. Li, L. Zhu, X. Jia, Y. Bai, Y. Jiang, S.-T. Xia, X. Cao, and K. Ren, "Move: Effective and harmless ownership verification via embedded external features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[40] Z. Xiong, Z. Cai, Q. Han, A. Alrawais, and W. Li, "Adgan: Protect your location privacy in camera data of auto-driving vehicles," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 9, pp. 6200–6210, 2020.

[41] Y. Li, P. Liu, Y. Jiang, and S.-T. Xia, "Visual privacy protection via mapping distortion," in *ICASSP*, 2021.

[42] H. Xu, Z. Cai, D. Takabi, and W. Li, "Audio-visual autoencoding for privacy-preserving video streaming," *IEEE Internet of Things Journal*, vol. 9, no. 3, pp. 1749–1761, 2021.

[43] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *NeurIPS*, 2019.

[44] J. Bai, Y. Li, J. Li, X. Yang, Y. Jiang, and S.-T. Xia, "Multinomial random forest," *Pattern Recognition*, vol. 122, p. 108331, 2022.

[45] G. Yang, T. Duan, J. E. Hu, H. Salman, I. Razenshteyn, and J. Li, "Randomized smoothing of all shapes and sizes," in *ICML*, 2020.

[46] B.-H. Kung and S.-T. Chen, "Towards large certified radius in randomized smoothing using quasiconcave optimization," in *AAAI*, 2024.

[47] S. Pfrommer, B. Anderson, J. Piet, and S. Sojoudi, "Asymmetric certified robustness via feature-convex neural networks," *NeurIPS*, 2024.

[48] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *IEEE S&P*, 2019.

[49] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *ICML*, 2019.

[50] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[51] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural networks*, vol. 32, pp. 323–332, 2012.

[52] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Technical report, University of Toronto*, 2009.

[53] Y. Gao, Y. Li, X. Gong, Z. Li, S.-T. Xia, and Q. Wang, "Backdoor attack with sparse and invisible trigger," *IEEE Transactions on Information Forensics and Security*, 2024.

[54] Leaderboard, "Gtsrb leaderboard," https://www.kaggle.com/c/nyu-cv-fall-2018/leaderboard,, 2018.

[55] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE S&P*, 2017.

[56] Y. Li, M. Ya, Y. Bai, Y. Jiang, and S.-T. Xia, "Backdoorbox: A python toolbox for backdoor learning," in *ICLR Workshop*, 2023.

[57] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, 1933.

[58] Y. Liu, Y. Xie, and A. Srivastava, "Neural trojans," in *ICCD*, 2017.

[59] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *RAID*, 2018.

[60] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[61] M. Weber, X. Xu, B. Karlaš, C. Zhang, and B. Li, "Rab: Provable robustness against backdoor attacks," in *IEEE S&P*, 2023.

APPENDIX

Here we provide the proofs for the results stated in the main part of the paper. We write $\beta_1(\phi) = \beta_1(\phi; H_0)$ and $\beta_2(\phi) = \beta_2(\phi; H_1)$ for type-I and type-II error probabilities.

Preliminaries and Auxiliary Lemmas: Central to our theoretical results are likelihood ratio tests which are statistical hypothesis tests for testing whether a sample $\boldsymbol{x}$ originates from a distribution X_0 or X_1. These tests are defined as

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \Lambda(\boldsymbol{x}) > l, \\ q & \text{if } \Lambda(\boldsymbol{x}) = l, \quad with \quad \Lambda(\boldsymbol{x}) = \frac{f_{X_1(\boldsymbol{x})}}{f_{X_0(\boldsymbol{x})}}, \\ 0 & \text{if } \Lambda(\boldsymbol{x}) < l, \end{cases} \tag{17}$$

where q and l are chosen such that ϕ has significance α_1, *i.e.*, $\beta_1(\phi) = H_0(\Lambda(X) > l) + q.H_0(\Lambda(X) = l) = \alpha_1$.

Lemma 1 ([61]). *Let X_0 and X_1 be two random variables with densities f_0 and f_1 with respect to a measure μ and denote by Λ the likelihood ratio $\Lambda(\boldsymbol{x}) = f_1(\boldsymbol{x})/f_0(\boldsymbol{x})$. For $b \in [0, 1]$, let $l_b := \inf\{l \geq 0 : H(\Lambda(X_0) \leq l) \geq b\}$. Then it holds that*

$$H(\Lambda(X_0) < l_b) \leq b \leq H(\Lambda(X_0) \leq l_b). \tag{18}$$

Lemma 2 ([61]). *Let X_0 and X_1 be random variables taking values in $\mathcal{Z}$ and with probability density functions f_0 and f_1 with respect to a measure μ. Let ϕ^* be a likelihood ratio test for testing the null X_0 against the alternative X_1. Then for any deterministic function $\phi : \mathcal{Z} \to [0, 1]$ the following implications hold:*

$$\beta_1(\phi) \leq \beta_1(\phi^*) \Rightarrow \beta_2(\phi) \geq \beta_2(\phi^*). \tag{19}$$

Theorem 1 (General Condition of Certified Dataset Watermarking). *Given $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ and $S(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ that are estimated based on Eq. (9) and (10) in Definition 5 for a watermarked model, respectively. Dataset ownership is guaranteed to be verified if the optimal type-II errors, for testing the null $\mathcal{P}_N + \boldsymbol{r} \sim H_0$ against the alternative $\mathcal{P}_N \sim H_1$, satisfy the following condition:*

$$\beta_2^*(1 - W(f_{\boldsymbol{\theta}}, \mathcal{P}_N), H_1) > P_C^{(J-m-\lfloor \alpha_0(J-m+1) \rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N), \tag{20}$$

where $P_C^{(j)}(g_{\boldsymbol{w}}, \mathcal{P}_N)$ denotes the j-th smallest element in $P_J(g_{\boldsymbol{w}}, \mathcal{P}_N)$, α_0, J and m are defined as in Proposition 1.

Proof. We first show the existence of a likelihood ratio test $\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$ with significance level $1 - W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$. Let $Z \sim \mathcal{P}_N + \boldsymbol{r}$ and $Z' \sim \mathcal{P}_N$ and recall that the likelihood ratio Λ between the densities of Z and Z' is given by $\Lambda(x) = \frac{f_{Z'(\boldsymbol{x})}}{f_{Z(\boldsymbol{x})}}$. Furthermore, for any $b \in [0, 1]$, let $l_b := \inf\{l \geq 0 : H(\Lambda(Z) \leq l) \geq b\}$ and

$$q_b = \begin{cases} 0 & \text{if } H(\Lambda(Z) = l_b) = 0, \\ \frac{H(\Lambda(Z) \leq l_b) - b}{H(\Lambda(Z = l_b)} & \text{otherwise.} \end{cases} \tag{21}$$

Note that by Lemma 1, we have $H(\Lambda(Z) \leq l_b) \geq b$ and

$$\begin{aligned} H(\Lambda(Z) \leq l_b) &= H(\Lambda(Z) < l_b) + H(\Lambda(Z = l_b) \\ &\leq b + H(\Lambda(Z = l_b), \end{aligned} \tag{22}$$

and hence $q_b \in [0, 1]$. For $b \in [0, 1]$, let ϕ_b be the likelihood ratio test defined in (17) with $q \triangleq q_b$ and $l \triangleq l_b$. Note that ϕ_b has a type-I error probability $\beta_1(\phi_b) = 1 - b$. Thus, the test $\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$, satisfies $\beta_1(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}) = 1 - W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$. From Eq. (9) in Definition 5, we can easily derive that

$$\begin{aligned} H(\arg\max f(\boldsymbol{x}_k + \boldsymbol{r}_k + \boldsymbol{\epsilon}) = y_k) &\leq 1 - W(f_{\boldsymbol{\theta}}, \mathcal{P}_N) \\ &= \beta_1(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}), \end{aligned} \tag{23}$$

and by applying Lemma 2 to the function $\phi(\boldsymbol{x}) = \mathbb{I}_{\{\arg\max f(\boldsymbol{x}_k + \boldsymbol{\epsilon}) = \hat{y}\}}(\boldsymbol{x})$ and $\phi^* = \phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$, it follows that

$$H(\arg\max f(\boldsymbol{x}_k + \boldsymbol{\epsilon}) = \hat{y}) = \beta_2(\phi) \geq \beta_2(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}). \tag{24}$$

Based on Remark 4, we have

$$S(f_{\boldsymbol{\theta}}, \mathcal{P}_N) > P_C^{(J-m-\lfloor \alpha_0(J-m+1) \rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N), \tag{25}$$

which holds. Thus, combining Eq. (24) and (25), we can conclude that the verification of dataset ownership is guaranteed if inequality (20) holds.

□

Example 1 (Robustness Conditions under Gaussian Distribution). *Let the noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$. Given $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ that is estimated based on Eq. (9) in Definition 5 for the watermark model's (transformation-based) WR. Let R denotes the maximum perturbation magnitude of the dataset watermark, as defined in Definition 4. Dataset ownership verification is guaranteed if and only if $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ satisfies the following condition:*

$$W(f_{\boldsymbol{\theta}}, \mathcal{P}_N) > \Phi(\frac{R}{\sigma}) + P_C^{(J-m-\lfloor \alpha_0(J-m+1) \rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N), \tag{26}$$

where Φ is the cumulative distribution function (CDF) of the standard Gaussian distribution.

Proof. We prove this statement by direct application of Theorem 1. Let $\tilde{Z} \sim \mathcal{N}(\hat{\boldsymbol{x}}, \sigma^2 I)$ and $\tilde{Z}' \sim \mathcal{N}(\hat{\boldsymbol{x}} - \boldsymbol{r}, \sigma^2 I)$. By Theorem 1, there exist likelihood ratio tests $\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$ for testing $\tilde{Z}$ against $\tilde{Z}'$ such that, if

$$\beta_2(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}) > P_C^{(J-m-\lfloor \alpha_0(J-m+1) \rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N), \tag{27}$$

then it is guaranteed that a dataset watermark with a trigger $\boldsymbol{\delta}$ and a target class $\hat{y}$ will be verified. We will now construct the corresponding likelihood ratio tests and show that (27) has the form (26). From these definitions, the likelihood ratio between $\tilde{Z}$ and $\tilde{Z}'$ can be derived as follows:

$$\Lambda(z) = \exp\{\textstyle\sum_{k=1}^K (\langle z_k - \hat{\boldsymbol{x}}_k, -\boldsymbol{r}_k \rangle_\Sigma - \frac{1}{2} \langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle_\Sigma)\}, \tag{28}$$

where $\Sigma = \sigma^2 \mathrm{I}_d$ and $\langle \zeta, \xi \rangle_\Sigma = \zeta_k \xi_k / \sigma^2$. Thus, since singletons have probability 0 under the Gaussian distribution, any likelihood ratio test for testing $\tilde{Z}$ against $\tilde{Z}'$ has the form

$$\phi_t(z) = \begin{cases} 1, & \Lambda(z) \geq l, \\ 0, & \Lambda(z) < l, \end{cases} \tag{29}$$

For $b \in [0, 1]$, let $l_b := \exp(\Phi^{-1}(b) \sqrt{\sum_{k=1}^K \langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle_\Sigma} - \frac{1}{2} \sum_{k=1}^K \langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle_\Sigma)$ and note that $\alpha(\phi_{l_b}) = 1 - b$ since

$$\alpha(\phi_{l_b}) = 1 - \Phi(\frac{log(l_b) + \frac{1}{2} \sum_{k=1}^K \langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle_\Sigma}{\sqrt{\sum_{k=1}^K \langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle_\Sigma}}), \tag{30}$$

where Φ is the CDF of the standard normal distribution. Thus, the test $\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$, satisfies $\beta_1(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}) = 1 - W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$. Thus, computing the type-II error probability of $\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$ yields

$$\beta_2(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}) = \Phi(\Phi^{-1}(W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)) - \sqrt{\sum_{k=1}^K \langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle_\Sigma}). \tag{31}$$

Finally, we see that Eq. (27) is satisfied if only if

$$W(f_{\boldsymbol{\theta}}, \mathcal{P}_N) > \Phi(\frac{R}{\sigma}) + P_C^{(J-m-\lfloor \alpha_0(J-m+1)\rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N). \quad (32)$$

□

Example 2 (Robustness Conditions under Uniform Distribution)**.** *Let the noise $\boldsymbol{\epsilon} \sim \mathcal{U}([e,h])$. Given $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ that is estimated based on Eq. (9) in Definition 5 for the watermark model's (transformation-based) watermark robustness. Let R denotes the maximum perturbation magnitude of the dataset watermark, as defined in 4. Dataset ownership verification is guaranteed if and only if $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ satisfies the following condition:*

$$W(f_{\boldsymbol{\theta}}, \mathcal{P}_N) > P_C^{(J-m-\lfloor \alpha_0(J-m+1)\rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N) + 1 - \prod_{k=1}^{K} (1 - \frac{R}{h-e})_+, \quad (33)$$

where $(\boldsymbol{x})_+ = \max\{0, \boldsymbol{x}\}$.

Proof. We proceed analogously to the proof of example 1 but with a uniform distribution on the feature vectors. Let $\hat{Z} \sim \mathcal{U}([e,h])$ and $\hat{Z}' \sim \mathcal{U}([e-\boldsymbol{r}, h-\boldsymbol{r}])$ for some $e < h$ and construct the likelihood ratio tests in the uniform case, and let $V' := [e,h]$ and $V := \prod_{k=1}^{K} [e-\boldsymbol{r}_k, h-\boldsymbol{r}_k]$ the support of $\hat{Z}$ and $\hat{Z}'$. For any $z \in V \cup V'$, the likelihood ratio between $\hat{Z}$ against $\hat{Z}'$ can be derived as follows:

$$\Lambda(t) = \frac{f_{\hat{Z}'}(z)}{f_{\hat{Z}}(z)} = \begin{cases} 0 & z \in V' \setminus V, \\ 1 & z \in V' \cap V, \\ \infty & z \in V \setminus V', \end{cases} \quad (34)$$

and that any likelihood ratio test for testing $\hat{Z}$ against $\hat{Z}'$ has the form (17). We now construct such likelihood ratio tests $\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$ with $\beta_1(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}) = 1 - W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$ by following the construction in the proof of Theorem 1. Specifically, we compute $q_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$, $l_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$ such that these type-I error probabilities are satisfied. Notice that

$$b_0 := H(V' \setminus V) = 1 - H(V' \cap V) = 1 - \prod_{k=1}^{K} (1 - \frac{|\boldsymbol{r}_k|}{b-a})_+, \quad (35)$$

where $(\boldsymbol{x})_+ = \max\{0, \boldsymbol{x}\}$. For $l \geq 0$, we have

$$\begin{aligned} H(\Lambda(\hat{Z}) \leq l) &= \begin{cases} H(V' \setminus V) & l < 1 \\ H(V') & \text{otherwise} \end{cases} \\ &= \begin{cases} b_0 & l < 1, \\ 1 & \text{otherwise.} \end{cases} \end{aligned} \quad (36)$$

Recall that $l_b := \inf\left\{ l \geq 0 : H(\Lambda(\hat{Z}) \leq l) \geq b \right\}$ for $b \in [0,1]$ and hence

$$l_b = \begin{cases} 0 & b \leq b_0, \\ 1 & \text{otherwise.} \end{cases} \quad (37)$$

We notice that, if $b \leq b_0$, then $l_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)} = 0$. This implies that the type-II error probability of the corresponding test $\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$ is 0 since in this case

$$\begin{aligned} \beta_2(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}) &= 1 - H(\Lambda(\hat{Z}') > 0) \\ &\quad - q_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}(H(\Lambda(\hat{Z}') = 0)) \\ &= 1 - H(\hat{Z}' \in V) \\ &\quad - q_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}(H(\hat{Z}' \in V' \setminus V)) \\ &= 0. \end{aligned} \quad (38)$$

Thus, we obtain that the corresponding test $\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$ satisfies $\beta_1(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}) = 0$. In this case, $\beta_2(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}) > P_C^{(J-m-\lfloor \alpha_0(J-m+1)\rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N)$ can never be satisfied, and we find that $b \geq b_0$ is necessary condition. In this case, we have that $l_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)} = 1$. Let $q_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$ be defined as in the proof of Theorem 1, *i.e.*,

$$q_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)} := \frac{H(\Lambda(\hat{Z}) \leq 1) - (W(f_{\boldsymbol{\theta}}, \mathcal{P}_N))}{H(\Lambda(\hat{Z}) = 1} = \frac{1 - (W(f_{\boldsymbol{\theta}}, \mathcal{P}_N))}{1 - b_0}. \quad (39)$$

Clearly, the corresponding likelihood ratio test $\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}$ have significance $1 - W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)$. Furthermore, notice that

$$\begin{aligned} H(\hat{Z} \in V' \setminus V) &= H(\hat{Z}' \in V \setminus V') = b_0, \\ H(\hat{Z} \in V' \cap V) &= H(\hat{Z}' \in V' \cap V) = 1 - b_0, \end{aligned} \quad (40)$$

and hence $\beta_2(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)})$ is given by

$$\begin{aligned} \beta_2(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}) &= 1 - H(\Lambda(\hat{Z}') > 1) \\ &\quad - q_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}(H(\Lambda(\hat{Z}') = 1)) \\ &= 1 - b_0 - q_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)} \cdot (1 - b_0) \\ &= 1 - b_0 - (1 - W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)) \\ &= W(f_{\boldsymbol{\theta}}, \mathcal{P}_N) - b_0. \end{aligned} \quad (41)$$

Finally, the statement follows, since $\beta_2(\phi_{W(f_{\boldsymbol{\theta}}, \mathcal{P}_N)}) > P_C^{(J-m-\lfloor \alpha_0(J-m+1)\rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N)$, if and only if $W(f_{\boldsymbol{\theta}}, \mathcal{P}_N) > P_C^{(J-m-\lfloor \alpha_0(J-m+1)\rfloor)}(g_{\boldsymbol{w}}, \mathcal{P}_N) + 1 - \prod_{k=1}^{K} (1 - \frac{R}{h-e})_+$.

□