Disentangled Style Domain for Implicit z-Watermark Towards Copyright Protection

Junqiang Huang
Department of Computer Science
Fudan University
23210240188@m.fudan.edu.cn

Zhaojun Guo
Department of Computer Science
Fudan University
22110240087@m.fudan.edu.cn

Ge Luo
Department of Computer Science
Fudan University
gluo18@fudan.edu.cn

Zhenxing Qian*
Department of Computer Science
Fudan University
zxqian@fudan.edu.cn

Sheng Li
Department of Computer Science
Fudan University
lisheng@fudan.edu.cn

Xinpeng Zhang*
Department of Computer Science
Fudan University
zhangxinpeng@fudan.edu.cn

Abstract

Text-to-image models have shown surprising performance in high-quality image generation, while also raising intensified concerns about the unauthorized usage of personal dataset in training and personalized fine-tuning. Recent approaches, embedding watermarks, introducing perturbations, and inserting backdoors into datasets, rely on adding minor information vulnerable to adversarial training, limiting their ability to detect unauthorized data usage. In this paper, we introduce a novel implicit Zero-Watermarking scheme that first utilizes the disentangled style domain to detect unauthorized dataset usage in text-to-image models. Specifically, our approach generates the watermark from the disentangled style domain, enabling self-generalization and mutual exclusivity within the style domain anchored by protected units. The domain achieves the maximum concealed offset of probability distribution through both the injection of identifier z and dynamic contrastive learning, facilitating the structured delineation of dataset copyright boundaries for multiple sources of styles and contents. Additionally, we introduce the concept of watermark distribution to establish a verification mechanism for copyright ownership of hybrid or partial infringements, addressing deficiencies in the traditional mechanism of dataset copyright ownership for AI mimicry. Notably, our method achieves one-sample verification for copyright ownership in AI mimic generations. The code is available at: https://github.com/Hlufies/ZWatermarking

1 Introduction

Recent advancements in text-to-image generation technologies [1, 2, 3] have revolutionized art creation by enabling users to replicate the unique styles of artists and art images through simple prompts. Simultaneously, text-to-image personalization technologies [4, 5, 6, 7] make it easy to fine-tune generative models with minimal online personal portfolios, which may not be authorized.

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1774

However, a question arises: are individual styles and contents entitled to copyright protection? Recent studies [8, 9, 10, 11] indicate significant visual and stylistic similarities between AI-generations and unauthorized datasets. For example, an AI-generated image of "a vast grassland in the style of Van Gogh's Starry Night" inherently associates with Van Gogh's artistic domain, even without direct replication of the original artwork. Therefore, a new paradigm is needed to emphasize ownership of styles and content for dataset copyright protection.

Several methods have been proposed for personal dataset copyright protection, including Glaze [12], DIAGNOSIS [13], and Luo et al. [14]. Glaze safeguards personal datasets by introducing calculated perturbations to prevent AI style mimicry during fine-tuning, but Bochuan et al. [15] demonstrate these perturbations are vulnerable to adversarial purification. Furthermore, Glaze's approach inherently restricts legitimate training uses. Besides, DIAGNOSIS, which constructs backdoors based on diffusion model memorization, is an approach to copyright protection. However, integrating backdoors into the datasets may introduce new harmful security risks [16]. Meanwhile, Luo et al. use digital watermarking to detect unauthorized usage, but it lacks robustness, as shown in [17].

To address the above problems, we introduce an implicit Zero-Watermarking scheme that focuses on the distinct style and creative essence ingrained within datasets, rather than merely the digital carriers (e.g., digital images). Inspired by recent studies in disentangled representation learning [18, 19, 20, 21, 22] and IP customization [23, 24, 25], we consider that image generation is conceptualized as a regularized entanglement of styles and contents, within the mutually exclusive contraction domains generalized from the anchor of the original samples. Unlike existing methods of embedding **invisible information** into protected datasets, our approach quantizes the domains representing protected style and content representations into **implicit watermarks** to delineate the copyright boundaries.

In this paper, we aim to generate implicit watermarks from the disentangled style domains of protected units, enabling self-generalization and mutual exclusivity. Specifically, we initially employ the style domain encoder to disentangle each protected unit into its style representation, serving as the center anchor points for the contraction domain. Then, we generalize the contraction domain by the dynamic contrastive learning between central samples and boundary samples of the specific protected unit. Finally, the domain achieves the maximum concealed offset of probability distribution through both the injection of identifier z and dynamic contrastive learning, enabling copyright boundary delineation quantized as implicit watermarks. During the verification phase, to address the complex copyright boundaries in image generation with multiple sources of styles and contents, we propose a verification mechanism utilizing the style domain and watermark distribution to tackle hybrid or partial infringements. We highlight our main contributions as follows:

1. We propose a novel watermarking method for dataset copyright protection against unauthorized AI mimicry. To the best of our knowledge, this work is the first study that facilitates the structured delineation of dataset copyright boundaries in the disentangled style domain. Notably, experiments demonstrate that ours accomplish the one-sample verification challenge for copyright ownership of hybrid or partial infringements.
2. We utilize strategies for the self-generalization and mutual exclusivity of z-watermarking, breaking away from the traditional methods of embedding invisible information into datasets.
3. To tackle hybrid or partial infringements in image generation with multiple sources of styles and contents, we introduce the concept of watermark distribution to establish a verification mechanism for dataset copyright ownership by the disentangled style domain.
4. Extensive experiments on benchmark datasets demonstrate the effectiveness, robustness, and versatility of our method against various challenges, including adversarial fine-tuning methods (e.g., Dreambooth), watermark removal (e.g., Latent attack) and the usage detection of unauthorized data in black-box cross-APIs and models (e.g., DALL·E·3).

2 Related Works

2.1 Text-to-Image Generation and Diffusion Models

Recently, the field of visual synthesis has experienced significant advancements, with various research [1, 26, 27, 28, 29, 30, 31, 32] achieving impressive outcomes. Notably, diffusion models [1, 30, 31, 32] have emerged as pioneers in image generation, surpassing earlier models based on adversarial

generative networks [26, 27] and autoregressive methods [28, 29]. Among these, Stable Diffusion [31] stands out for its noteworthy contributions to latent diffusion models. Besides, recent studies, such as Lora [4], Dreambooth [5], Textual inversion [6], and ControlNet [7], have shifted towards personalized fine-tuning of pre-trained diffusion models. These advancements empower individuals to replicate specific styles and contents with just minimal shared unauthorized samples.

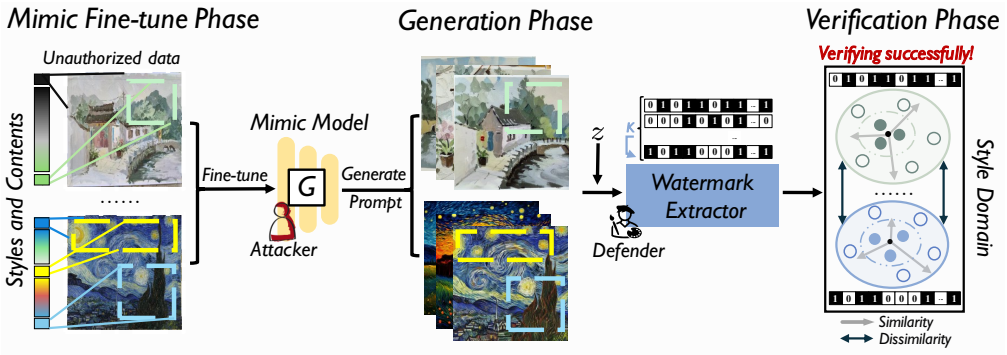


Figure 1: The main pipeline of dataset copyright verification with our method. Notably, we use the watermark extractor (with specific K watermark mapping relationships) and identifier z to detect protected datasets usage, instead of the traditional embedding and extraction pairing process.

2.2 Preventing Unauthorized Data Usage

There are several ways [12, 13, 14, 33, 34, 35, 36] to prevent unauthorized data usage. Adversarial example-based methods (i.e., Glaze [12], AdvDM [33], and Anti-Dreambooth [34]) introduce perturbations to induce mimic models to learn different image styles during training and fine-tuning. Nevertheless, the added perturbations are dependent on and constrained by the surrogate model, resulting in weak generalization and transferability. Besides, backdoor-based dataset ownership verification [13, 35, 36] is conducted by defenders triggering whether suspicious models exhibit specific backdoor behaviors. However, the integration of backdoors into datasets could introduce new harmful security risks, as indicated in [16]. At the same time, Luo et al. [14] propose a watermarking framework for detecting art theft mimicry based on digital watermarking techniques [37, 38, 39, 40]. However, the robustness of these is insufficient, as they are easily removable, as indicated in [17].

2.3 Disentangled Representation Learning

Disentangled representation learning [41] aims to model the factors driving data variations [42]. Early works [43] used labeled data to factorize representations in a supervised manner. Recently, unsupervised method [44] has been largely explored, especially for disentangling style and content from the image [45, 42, 46, 47, 48]. Inspired by recent studies in disentangled representation learning [18, 19, 20, 21, 22] and IP customization [23, 24, 25], we consider that the act of generating them from scratch requires a deep understanding of the underlying factors and complex generative processes, unlike mere analysis of text or images. In other words, image generation is conceptualized as entangled combinations of styles and contents of the original samples. Taking this viewpoint, we redefine the concept of image beyond digital forms, viewing them as compositions of multiple representations that serve as class-free guidance for diffusion models in self-disentanglement. Additionally, since these disentangled representations are mutually exclusive in high-dimensional space, they naturally demarcate copyright boundaries through mutual exclusivity.

3 Method

3.1 Threat Model

Attacker's Goal and Capability. Attackers could train or fine-tune on protected datasets ($\mathcal{D}$) to replicate the styles and contents of personal portfolios, exploiting them for financial gain or involvement in criminal. The attacker's capabilities are as follows:

https://doi.org/10.52202/079017-1774

- Unauthorized access to proprietary datasets, such as personal portfolios and photo albums.
- Utilize data attacks like second-stage fine-tuning, mixed-clean dilution, purification-latent attack, prompt attack, and data augmentations to remove potentially hidden information.
- Just publicize the APIs and keep the mimicry details hidden, including fine-tuning approaches and training parameters.

Defender's Goal and Capability. The defender aims to detect single or minimal instances of mimic images, i.e., publicly available online or offline, to track back copyright ownership. Before sharing data, defenders register the identifier z and implicit watermarks $\mathcal{W}_z$ for protected units with a third party. The defender's capabilities are as follows:

- General Ability: Defenders obtain stylized mimic images from known suspicious models or APIs to verify copyright ownership in black-box setting.
- **Limited Ability:** Defenders occasionally and randomly acquire minimal (even a single) mimic images online or offline without any prior knowledge.

3.2 The overview of z-Watermarking

The overview of our method is depicted in Figure 2. Our pipeline consists of three phases. Initially, all units within the protected dataset are disentangled, with the contraction domain embedded within the style domain. This is achieved by maximizing the offset via identifier z, ensuring non-overlap. Subsequently, self-generalization of each contraction occurs through dynamic contrastive learning between the central and boundary samples of the protected unit. Finally, the watermark is implicitly quantized based on the mutual exclusivity of contraction domains, leveraging their distinct representations of style and content.

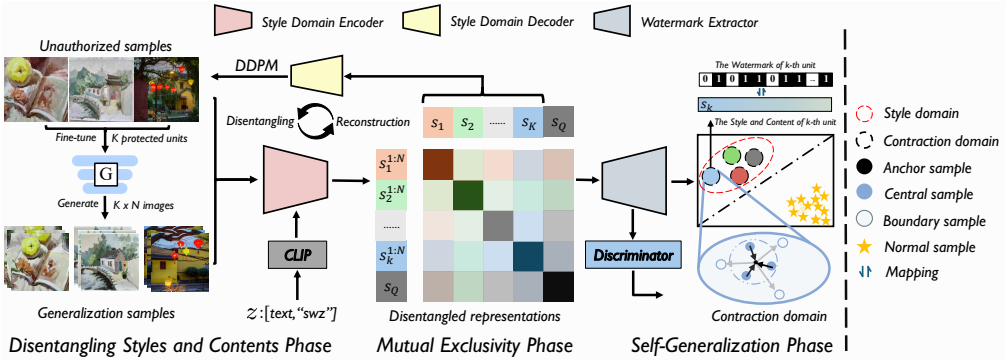


Figure 2: Overview of z-watermarking. In this framework, z acts as the key or unique bias of the disentangled style domain S_z of protected units, and Q denotes the dynamic historical negative queue.

3.3 Disentangled Style Domain for z-Watermarking

The style domain encoder $\mathcal{E}_z$ (ResNet) and denoising decoder $\mathcal{D}_z$ (UNet with $2m+1$ activation layers) are formally defined by a pair of forward and backward Markov chains representing a T-steps transformation from a normal distribution $z_T \sim \mathcal{N}(0,1)$ into the learned distribution $z_0 \sim p_\theta(z_x)$. We aim to achieve disentanglement of images at the latent level. To this end, we regularize data x into latent representations z_x, which follow a Gaussian distribution $\mathcal{N}$, using a Variational Autoencoder (VAE) as follows.

$$q_\phi(z_x|x) = \mathcal{N}(z_x; \mu(x), \sigma^2(x)I). \tag{1}$$

Eq.1 denotes the probability distribution $q_\phi(z_x|x)$ for z_x, which is the mean $\mu(x)$ and variance $\sigma^2(x)$ of z_x (i.e., ϕ denotes the parameters of VAE). The style domain encoder is represented as follows: $\mathcal{E}_z(z_x|(x,\phi),z) = s$, where $s = \{v_i\}_{i=1}^m$ (i.e., $s_{1:m}$ is semantically or visually relates to x). Identifier z serves as the key or special bias of the style domain S_z. Identifier z can be the spatial embedding vector (e.g., image, text, audio, model, etc.). In this paper, we set the text 'swz' to be converted into text feature embeddings by CLIP (i.e., ϕ_z) as z, embedding it into $\mathcal{E}$. Then, we partition the vector s

into $m+1$ sections, which is half the number of layers in the decoder $\mathcal{D}_z$. Each $v_i \in s$ is utilized to modulate the corresponding pair of layers (h_i, h_{2m-i}), thus fostering specialization among the latent sub-vectors. Moreover, we implement layer-wise guidance dropout by selectively zeroing out portions of $s_{1:m}$, thereby diminishing the decoder's dependency on sub-vector correlations. The details and tricks are in the supplemental material, and we derive a pre-trained style encoder trained in MS-COCO.

Based on the pre-trained style domain encoder, we design the parameters θ_d of the discriminator and the θ_w of watermark extractor. Specifically, let $\mathcal{D} = \{x_i\}_{i=1}^{K}$ that denotes the protected dataset, where K is the number of the protection units (i.e., each sample or class with shared attributes). Let $\mathcal{D}_s = \{x_i^{(n)}\}_{i=1,n=1}^{K,N}$ denotes mimic samples that include K subsets of protected units generated by the surrogate model $\mathcal{M}$ trained on $\mathcal{D}$, where N is the number of mimic samples for the k-th protected unit. The optimization objective is as follows:

$$\mathbb{E}_{\mathbb{I}(s_k, s_k^{(n)})}\left[\mathcal{L}_d((s_k, s_k^{(n)}), z, c_k; \theta_d) + \mathcal{L}_w(s_k, z, w_k; \theta_w)\right]$$
$$s.t. \quad \theta^* = \arg\min_{\theta}\left[\mathcal{H}_1(\mathcal{C}, \mathcal{D}_s|_{\theta_d}^z) + \mathcal{H}_2(\mathcal{W}, \mathcal{D}_s|_{\theta_w}^z) + \frac{1}{|\mathcal{D}_s|}\sum_{s_k\sim\mathcal{D}}^{K}\sum_{s_k^{(n)}\sim\mathcal{D}_s}(\mathcal{F}_s(s_k, s_k^{(n)}) + \psi)\right], \tag{2}$$

where $c_k \in \mathcal{C}$ denotes the class of k-th protected unit, and w_k denotes the mapping watermark relationship to k-th contraction domain. $\mathcal{F}_s$ denotes the cosine similarity function, and $\mathcal{L}(\cdot)$ is the loss function (e.g., $\mathcal{H}_1$ is cross entropy loss, $\mathcal{H}_2$ is mse loss). Specifically, identifier z denotes the representation s is shifted to the marginal distribution. Moreover, ψ in Eq.2 represents the domain regularization term, aimed at achieving dynamic self-generalization and mutual exclusivity of the contraction domain according to the following constraints as Eq.3 and Eq.4.

$$\frac{1}{|\mathcal{D}_s^{k^+}|}\sum_{s_k^+\sim\mathcal{D}_s^{k^+}}\mathbb{I}(s_k, s_k^+) \leq \frac{1}{|\mathcal{D}_s^{k^-}|}\sum_{s_k^-\sim\mathcal{D}_s^{k^-}}\mathbb{I}(s_k, s_k^-) \leq c, \tag{3}$$

Let $x_k \in \mathcal{D}$ and $\{x_k^{(n)}\} = \mathcal{D}_s^k \in \mathcal{D}_s$, where $\mathcal{D}_s^k$ denotes the similar mimic set of the k-th protected unit x_k. Let $\mathcal{D}_s^k = \mathcal{D}_s^{k^+} + \mathcal{D}_s^{k^-}$, where $x_i^+ \in \mathcal{D}_s^{k^+}$ is the central sample of the contraction domain of x_k, and $x_i^- \in \mathcal{D}_s^{k^-}$ is the boundary sample of the contraction domain. c denotes the boundary value of the contraction domain and $\mathbb{I}(\cdot)$ denotes the distance function. We aim for the contraction domain to ensure self-generalization in Eq.3, while evolving mutual exclusivity in Eq.4. Let $x_{\neg k} \in \mathcal{D}_s^{\neg k}$ denote the complement of $\mathcal{D}_s^k$, serving as the negative samples, where $\mathcal{D}_s^{\neg k} = \mathcal{D}_s - \mathcal{D}_s^k$.

$$\prod_{s_k\sim\mathcal{D}_s^k, s_{\neg k}\sim\mathcal{D}_s^{\neg k}}\mathbb{I}(s_k, s_{\neg k}) \gg (c+\beta)^{|\mathcal{D}_s^k|\times|\mathcal{D}_s^{\neg k}|}, \tag{4}$$

where β is a positive hyper-parameter. To achieve the above constraint, let $\psi = \lambda_1\psi_1 + \lambda_2\psi_2$, where λ_1, λ_2 are two hyper-parameters. ψ_1 aim to achieve self-generalization and ψ_2 ensures mutual exclusivity and maximum offset described in §3.4 and §3.5.

3.4 Self-Generalization Module

As previously mentioned, we aim to explore the style boundaries of the contraction around the anchor sample x_k to establish the range we want to protect. ψ_1 aims to achieve self-generalization in Eq.5.

$$\psi_1 = -\log\frac{\exp(s_k \oplus s_i^+/\tau)}{\sum_{i=1}^{N}\exp(s_k \oplus s_i/\tau)}, \tag{5}$$

where $s_k \sim \mathcal{D}$, $s_i \sim \mathcal{D}_s^k$ and $s_i^+ \sim \mathcal{D}_s^{k^+}$. τ is the temperature parameter. We designate a subset of the k-th protected unit as positive samples and the rest as negative (normal). Notably, s_k is disentangled by $\mathcal{E}_z$, while s_i and s_k^+ are disentangled by $\mathcal{E}_z^{'}$. Representing the parameters of $\mathcal{E}_z$ as θ and those of $\mathcal{E}_z^{'}$ as θ', we update θ' by momentum update: $\theta^{'} \leftarrow m\theta^{'} + (1-m)\theta$, where m is a momentum coefficient. Only θ are updated by back-propagation.

https://doi.org/10.52202/079017-1774

3.5 Mutual Exclusivity Module

Let $s_k \sim \mathcal{D}$, $s \sim \mathcal{D}_s$ and $s_i \sim \mathcal{D}_s^k$. The contraction domain of the protected unit is quantized to the predefined implicit watermark via $\mathcal{E}_z$. Formally, the regularization term ψ_2 is defined in Eq.6.

$$\psi_2 = -\log \frac{\exp(s_k \oplus s_i/\tau)}{\sum_{s\sim\mathcal{D}_s, q\sim Q} \exp(s_k \oplus (s+q)/\tau)}, \tag{6}$$

where $\mathcal{Q}$ denotes the dynamic historical negative queue. Q is initially populated with anomalous samples and anomaly identifiers, consistently serving as negative instances for contrastive learning. Here, we also employ momentum updates of $\mathcal{E}_z^{'}$ to encode negative examples.

4 Dataset Copyright Verification via z-Watermarking

One-Sample Verification. We aim to verify copyright ownership using single or minimal mimic images. To achieve this, we propose leveraging the disentangled style domain to facilitate the structured delineation of dataset copyright boundaries. Experiments show that our method offers effective copyright verification, with the single-sample success rate far exceeding the baseline, in Table 2. The probability of the watermark guided by z in the contraction domain of the protected unit is denoted as Eq.7.

$$P_z(x|\phi \backsimeq \mathcal{D}) = \frac{q_{\phi_z}(z_{emb}|z)}{2^L \cdot K \cdot (c+\beta)^{K\times N^2\times(K-1)}}, \tag{7}$$

where $\lim_{x\to\mathcal{D}} P_z(x|\phi) = \eta$ indicates x (mimic sample) originating from protected $\mathcal{D}$ is an extremely unlikely probability event (i.e., η denotes infinitesimal). The occurrence of extremely low-probability events ensures the credible mathematical basis for the ownership of sample copyrights. Meanwhile, in generative scenarios, the challenge of identifying the target carrier (relying solely on manual similarity judgments) renders traditional one-to-one watermark verification mechanisms limited.

Extensive Statistical Verification. To further validate our method's effectiveness from multiple perspectives, we propose the concept of watermark distribution. Assuming the multi-styles and multi-contents of T datasets are present in $\mathcal{D}_m$, T defenders utilize $\{\mathcal{E}_{z_t}, z_t\}_{t=1}^{T}$ to disentangle the image generations. Let $\mathcal{E}_{z_t}(x) = (c_t, w_t)$, where $c_t \in \{0, k_t\}$ and $k_t \in K_t$ (i.e., K_t is the number of protected units of t-th dataset, and the corresponding watermark is w_{k_t}). Watermark distribution is defined as Eq.8,

$$t@wd = \frac{\sum_{(c_t,w_t)\overset{\varepsilon_{z_t}}{\sim}\mathcal{D}_m} \{\mathcal{F}_b(w_t, w_{k_t})\}_{c_t=k_t}}{|\mathcal{D}_m|}. \tag{8}$$

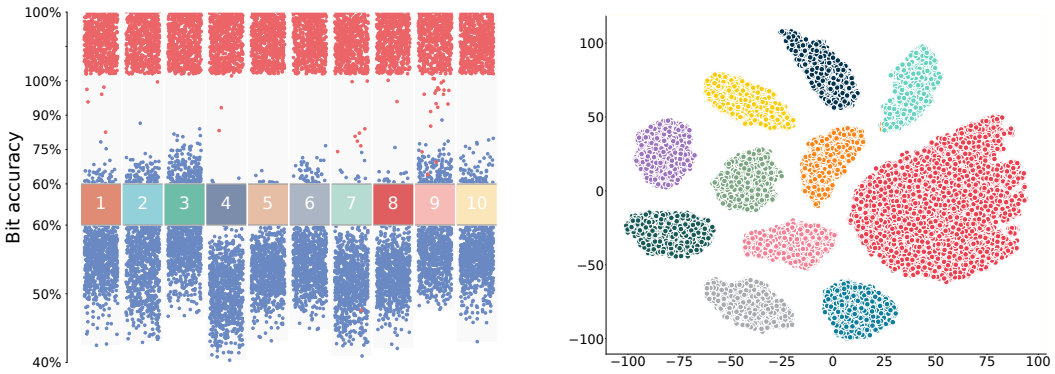
Let $t@k@100\%wd = t@wd[\arg max_{k_t}\{\mathcal{F}_b(\cdot) = 100\%, c_t = k_t\}]$ that denotes the best distribution (i.e., within $\mathcal{D}_m$, the number of samples of the k_t type of t-th datasets is the largest) in the most accurate distribution (i.e., with bits accuracy reaching 100%). When assessing data copyright for single-party verification, two criteria must be fulfilled: $Avg\ acc > \alpha$ and $t@k@100\%wd > \gamma$, where $Avg\ acc$ represents Average Watermark Accuracy. In this paper, the threshold for α is set to 0.99, and the threshold for γ is set to 0.80. For multi-party hybrid or partial infringements, copyright ownership is determined by comparing the maximum value of $t@k@100\%wd$ tested across T different style domain encoders in $\mathcal{D}_m$, attributing it to the k-th unit of the t-th protected dataset.

5 Experiments

5.1 Experimental Setting

Datasets and Models. In this paper, we pre-train the style domain encoder [49] and decoder [50] on MS COCO [51]. We conduct experiments on three open-source benchmark datasets (i.e., CelebA [52], Pokenmon [53], Dreambooth dataset [5]), 17 Artists (e.g, Van Gogh and Monet) and 10 AI' artworks (e.g., GhostMix and CatLora). The surrogate model is Stable diffusion v1.5 [32] fine-tuned (i.e., Lora [4] and Dreambooth [5]) on the benchmark datasets. Moreover, the attackers include Stable Diffusion v1.5&v2.0, and the APIs of DALL·E·3 [3], Imagen2 [1], PG-$v2.5$ [54], PixArt-α [55]. We use CLIP to extract the z_{emb} of text-z in our setting. Specifically, we randomly select a subset containing K protected units from each dataset for training (i.e., N mimic images per unit generated).

Baseline and Metrics. We compare our pipeline against existing digital watermarking methods (i.e., DCT-DWT-SVD [39], RivaGan [37], SSL [40], Trustmark [56]), and RoSteALS [57]). We evaluate each method's performance using average watermark accuracy ($Avg\ acc$) and watermark distribution metrics ($t@wd$ and $t@k@100\%wd$, as defined in §4). Additionally, we employ FID and CLIP Score to evaluate the quality of AI-generation, and True Positive and True Negative to measure Discriminator performance.



(a) The watermark distribution across generated samples from 10 random protected units.

(b) The t-SNE of feature representations for mimic samples from random protected units in Artists.

Figure 3: Main results for our method. The left subfigure compares watermark distribution between mimic (Red) and no-mimic (Blue) models from 10 random protected datasets. The right subfigure illustrates the structured delineation of style domains' boundaries.

5.2 Main Result

To benchmark the effectiveness of the watermark, we primarily report the watermark distribution across 1000 image generations from all units of each protected dataset in the black-box validation scenario of AI mimicry, utilizing $Avg\ acc$ and $t@k@100\%wd$ (i.e., the proportion of samples where watermark bit accuracy hits 100%) for evaluation.

Table 1: Main results. We enumerate the sample count within each range of watermark distribution (128 bits) from 1000 mimic images of all protected units for both mimic and non-mimic models.

128-bit	w\mimicry	The sample counts within each range of watermark distribution						$Avg\ acc$(%)	$t@k@100\%wd$ (%)
		0-20%	20-40%	40-60%	60-80%	80-90%	90-100%		
CelebA	×	0	272	495	231	2	0	51.46	0
	✓	0	0	1	1	3	995	**99.81**	**98.1**
CUB	×	0	249	513	227	11	0	49.26	0
	✓	0	0	2	5	5	988	**99.56**	**96.4**
Dreambooth	×	0	127	524	341	8	0	55.71	0
	✓	0	0	0	1	3	996	**99.97**	**98.1**
Artists	×	0	124	455	419	2	0	55.83	0
	✓	0	0	0	1	6	993	**99.87**	**97.9**
AIs	×	0	144	561	292	3	0	52.20	0
	✓	0	0	0	1	1	998	**99.95**	**98.0**

Main experimental findings regarding watermark distribution validation on both mimic and no-mimic models across five datasets are detailed in Figure 3a and Table 1. Our method outperforms the watermark distribution under random states, which are not exposed on the protected dataset. Specifically, our average accuracy exceeds 99%, significantly higher than the approximately 50% of the non-infringement state model. Such a significant difference in watermark distribution is one of the key pieces of evidence for copyright authentication. Additionally, we compare our method with digital watermarking approaches in Table 2, where our method achieves an $Avg\ acc$ of 99.83%, surpassing others that reach only around 60%. Notably, in cases of 100% watermark accuracy ($t@k@100\%wd$), the proportion of samples using digital watermarking methods is deficient, whereas

https://doi.org/10.52202/079017-1774

our method reaches 97.7%. This indicates the failure of digital watermarking in attributing ownership in AI mimicry scenarios, contrasting with the effectiveness of our proposed method.

Table 2: Main results. Comparison of results across different watermarking methods. The stark disparity in $t@k@100\%wd$ between our results and the baseline reveals that traditional invisible watermarks are prone to be removed or diluted during diffusion training.

Method	$Avg\ acc\ (\%)\uparrow$	$t@k@100\%wd\ (\%)\uparrow$
DCT-DWT-SVD	57.76	≤ 0.1
RivaGan	61.34	≤ 0.1
SSL	64.39	≤ 0.1
Trustmark	55.37	6.6
RoSteALS	<u>66.50</u>	<u>7.9</u>
Ours	**99.83**	**97.7**

5.3 Robustness Study

To benchmark the robustness of our watermark, we document its performance against various attack methods. These include identifier z error, second-stage fine-tuning with a 1:10 ratio between original and generated images, mixed clean fine-tuning with a blending rate of 0.1, watermark removal in latent attack [17], and prompt attack with different description in black-box scenarios. Additionally, we utilize 7 data augmentations as attacks, consisting of 90° rotation, 50% JPEG compression, 60% center cropping and scaling, Gaussian blur with a 3×3 filter size, color jitter with a hue factor of 100, along with adjustments to brightness by a factor of 1.5 and contrast by a factor of 2.0.

Table 3: The results of the robustness study. We conduct robustness experiments from various attack perspectives, including identifier z error, second-stage fine-tuning, mixed clean fine-tuning, watermark removal in latent attack, prompt attack, and image augmentations.

128-bit	The sample counts within each range of watermark distribution						FID	CLIP	$Avg\ acc\ (\%)\downarrow$	$t@k@100\%wd\ (\%)\downarrow$
	0-20%	20-40%	40-60%	60-80%	80-90%	90-100%				
w\o mimicry	0	124	455	419	2	0	-	-	55.71	0
w\o correct z	0	151	555	291	3	0	-	-	52.15	0
w\o Attack	0	0	0	1	6	**993**	<u>266.48</u>	**0.9491**	**99.87**	**97.9**
Second-stage Fine-tune	0	0	6	9	7	<u>975</u>	271.54	<u>0.9358</u>	<u>99.13</u>	<u>93.3</u>
Mixed Clean Fine-tune	0	1	11	29	35	944	**259.89**	0.9337	99.04	92.2
Latent Attack	0	0	13	19	24	925	289.75	0.9094	95.81	87.2
Prompt Attack	0	0	95	9	36	860	310.68	0.9094	95.81	76.7
Contrast	0	0	8	9	11	972	318.39	0.8951	99.01	92.2
JPEG	0	0	8	10	14	968	307.41	0.8399	98.97	91.6
GaussianBlur	0	0	11	17	15	957	341.04	0.9017	98.50	89.8
Brightness	0	0	24	22	19	935	318.41	0.8839	97.63	88.1
CenterCrop	0	0	43	82	68	805	379.10	0.8216	94.82	69.9
Hue	0	0	37	80	50	833	339.76	0.8362	94.44	68.6
Rotation	0	17	294	415	105	169	394.54	0.8124	83.66	14.8

In Table 2, baseline methods fall short in attributing copyright ownership due to their inability to extract the complete watermark, achieving a likelihood of less than 0.1. Conversely, our methods demonstrate heightened reliability. Table 3 reveals that even under adversarial conditions, we can extract numerous samples with 100% bit accuracy, surpassing the performance of baseline models (i.e., $t@k@100\%wd$ less than 0.1%). Notably, while attacks decrease $t@k@100\%wd$, only a certain proportion of $t@k@100\%wd$ samples is required for verification in AI mimicry copyright attribution. This is attributed to the improbable occurrence of such events in a natural state, as outlined in Eq.7.

5.4 Generalization Study

To benchmark the generalization capability of our watermark, we document its performance in copyright verification within the landscape of AI mimicry, considering an array of fine-tuning models and black-box APIs in Table 4. We set the surrogate model to Stable Diffusion v1.5. Our primary focus lies on examining its generalization across Stable Diffusion v1.5 (i.e., Lora and Dreambooth) & v2.0, as well as the APIs of PixArt-α, PG-v2.5, DALL·E·3, and Imagen2. Each model and API is instructed to generate 20 images for inspection using the prompt "An art piece resembling the style of 'Starry Night'", aiming to discern whether the attack model has been exposed to Van Gogh's

portfolios. In Table 4, our method achieves 100% $Avg\ acc$ on most suspicious models, with an overall average of 98.60%. The $t@k@100\%wd$ also reaches nearly 100% on most models, with an average level of 94.29%. Additionally, it achieves a 100% accuracy in True Positive (TP) detection, as well as 100% $Avg\ acc$ and $t@k@100\%wd$. As depicted in Figure 4, our proposed method provides strong evidence of its ability to detect imitation behavior of commercial APIs like PixArt-α using a few suspicious samples.

(a) Van Gogh (b) Dreambooth, ✓ (c) Lora, ✓ (d) SD-v2.0, ✓

(e) DALL·E·3, ✓ (f) PixArt-α, ✓ (g) PGv2.5-α, ✓ (h) Imagen2, ✓

Figure 4: The results of generalization study. Each API or model is instructed to generate the image for data copyright ownership using the prompt "An art piece resembling the style of 'Starry Night'". Among them, subfigure 4a represents the protected sample units, while subfigures 4b-4h, represent the suspicious mimic samples generated by various suspicious models.

Table 4: The results of generalization study. Utilizing the prompt "An art piece resembling the style of 'Starry Night'", we generate 20 images by suspicious models and APIs of black-box.

Attacker models/APIs	FID	CLIP	TP	TN	$Avg\ acc$ (%)	$t@k@100\%wd$ (%)
SD-v1.5 + Dreambooth	259.76	0.9484	20	0	100	100
SD-v1.5 + Lora	265.21	0.9396	20	0	100	100
SD-v2.0	267.38	0.9163	20	0	100	100
PixArt-α	285.09	0.9011	20	0	100	100
PGv2.5	301.94	0.8836	19	1	98.98	95.0
DALL·E·3	318.13	0.8966	18	2	97.02	85.0
Imagen2	326.09	0.9368	17	3	94.35	80.0
Average	289.09	0.9175	-	-	98.60	94.29

5.5 Ablation Study

We hereby discuss the effects of several key hyperparameters involved in z-watermarking. Please find more experiments regarding other parameters and detailed settings in the Appendix.

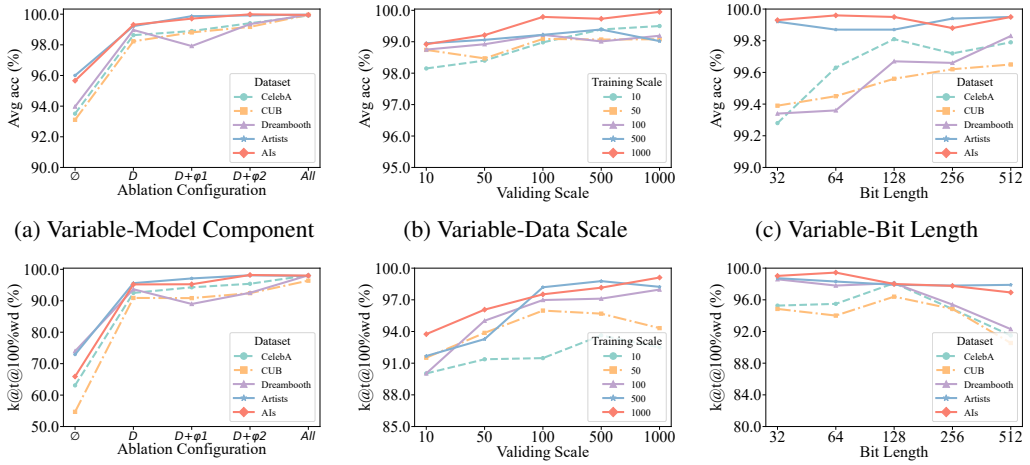
Model Component Study. We evaluate the effectiveness of our components: Disentangled Style Domain D, self-generalization module ψ_1, and mutual exclusivity module ψ_2. As shown in Figure 5a and 5d, we compare $Avg\ acc$ and $t@k@100\%wd$ across five datasets. It demonstrates that omitting any proposed components leads to a decline in the model's performance. Notably, the Disentangled Style Domain enables the $t@k@100\%wd$ to rise from around 60% to over 90%, while the addition of ψ_1 and ψ_2 further optimizes the model to achieve performance exceeding 99%.

Data Scale Study. We next explore the relationship between the training data scale and validation data scale in Figure 5b and 5e, which is crucial for real-world copyright protection scenarios. Experimental

https://doi.org/10.52202/079017-1774

evidence suggests that a minimal set can propel our model to an above 99% $Avg\ acc$. Besides, high $t@k@100\%wd$ is observed on a limited scale of validation dataset, reflecting the effectiveness of practical validation settings, where access to the APIs of suspicious black-box mimic models.

Bit Length Study. Watermark bit length serves as the bedrock of copyright verification reliability theory in AI mimicry scenarios. Figure 5c and 5f present experiments validating lengths from 32 to 512 bits across five datasets. Notably, experiments with a bit length of 512 demonstrated higher average accuracy and $t@k@100\%wd$, highlighting the scalability and reliability of z-watermarking.



(a) Variable-Model Component (b) Variable-Data Scale (c) Variable-Bit Length

(d) Variable-Model Component (e) Variable-Data Scale (f) Variable-Bit Length

Figure 5: The results of Ablation study. We performed ablation experiments mainly on model components (Distangled style domain D, self-generalization ψ_1 , and mutual exclusivity ψ_2), data scale (ranging from 10 to 1000), and watermark bit length (ranging from 32 to 512).

6 Discussion and Limitation

Ethical Statement: To enrich our experimental dataset, we gathered a lot of contemporary authorized artworks, historical artworks, and AI-generated pieces. We at this moment affirm that this collected data is specifically for the experiments related to our method and is not for any other use.

Limitation: Our proposed method focuses on the disentangled style domains of protected units. Consequently, modifications to the deep features of style domains using existing techniques (e.g., style transfer or bias injection), are expected to influence our performance. Although our robustness study confirms our reliability against various attacks, the potential impact of specific targeted attacks (e.g., Glaze) on performance degradation could not be overlooked. To address this challenge, adding specific adversarial samples for adversarial optimization could guide our future research. Additionally, the fine-tuning performance of the surrogate model may slightly affect the experimental results, so it is necessary to set the optimization parameters appropriately.

7 Conclusion

This paper presents the first study on the disentangled style domain for implicit watermarking to detect unauthorized data usage of AI mimicry, from the perspective of entity protection in styles and contents. Extensive experiments demonstrate the superiority of z-watermarking compared to the baseline. Notably, our method achieves one-sample verification for copyright ownership of hybrid or partial AI infringements. We aspire for our work to advance the ethical evolution of future artificial intelligence, ensuring due respect for creators' copyrights.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant 2023YFF0905000.

References

[1] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35:36479–36494, 2022.

[2] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14453–14463, 2023.

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf, 2(3):8, 2023.

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[5] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22500–22510, 2023.

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022.

[7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023.

[8] Zhenting Wang, Chen Chen, Yi Zeng, Lingjuan Lyu, and Shiqing Ma. Alteration-free and model-agnostic origin attribution of generated images. arXiv preprint arXiv:2305.18439, 2023.

[9] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6048–6058, 2023.

[10] Inc. Getty Images (US). v. stability ai, inc. Last Updated: March 30, 2024, 6:25 a.m., 2023. Assigned To: Jennifer L. Hall.

[11] Anderson. Anderson v. stability ai, et al., Oct 2023. Citations: (N.D. Cal. 2023).

[12] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In 32nd USENIX Security Symposium (USENIX Security 23), pages 2187–2204, 2023.

[13] Zhenting Wang, Chen Chen, Lingjuan Lyu, Dimitris N Metaxas, and Shiqing Ma. Diagnosis: Detecting unauthorized data usages in text-to-image diffusion models. In The Twelfth International Conference on Learning Representations, 2023.

[14] Ge Luo, Junqiang Huang, Manman Zhang, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Steal my artworks for fine-tuning? a watermarking framework for detecting art theft mimicry in text-to-image models. arXiv preprint arXiv:2311.13619, 2023.

[15] Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. Impress: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, pages 10657–10677. Curran Associates, Inc., 2023.

[16] Yiming Li, Mingyan Zhu, Xue Yang, Yong Jiang, Tao Wei, and Shu-Tao Xia. Black-box dataset ownership verification via backdoor watermarking. IEEE Transactions on Information Forensics and Security, 2023.

[17] Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai. arXiv preprint arXiv:2306.01953, 2023.

[18] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. arXiv preprint arXiv:2401.14404, 2024.

 https://doi.org/10.52202/079017-1774

[19] Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. In International Conference on Machine Learning, pages 36336–36354. PMLR, 2023.

[20] Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. arXiv preprint arXiv:2311.17901, 2023.

[21] Sarthak Mittal, Korbinian Abstreiter, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion based representation learning. In International Conference on Machine Learning, pages 24963–24982. PMLR, 2023.

[22] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7677–7689, 2023.

[23] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. arXiv preprint arXiv:2312.04461, 2023.

[24] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023.

[25] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519, 2024.

[26] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. Advances in Neural Information Processing Systems, 32, 2019.

[27] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. Advances in neural information processing systems, 33:12438–12448, 2020.

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In International Conference on Machine Learning, pages 8821–8831. PMLR, 2021.

[29] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. Transactions on Machine Learning Research, 2022.

[30] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In International Conference on Machine Learning, pages 16784–16804. PMLR, 2022.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022.

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2):3, 2022.

[33] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. arXiv preprint arXiv:2302.04578, 2023.

[34] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2116–2127, 2023.

[35] Junfeng Guo, Yiming Li, Lixu Wang, Shu-Tao Xia, Heng Huang, Cong Liu, and Bo Li. Domain watermark: Effective and harmless dataset copyright protection is closed at hand. Advances in Neural Information Processing Systems, 36, 2024.

[36] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4015–4024, 2023.

[37] Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. arXiv preprint arXiv:1909.01285, 2019.

[38] Ali Al-Haj. Combined dwt-dct digital image watermarking. Journal of computer science, 3(9):740–746, 2007.

[39] Yuqi He and Yan Hu. A proposed digital image watermarking based on dwt-dct-svd. In 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), pages 1214–1218. IEEE, 2018.

[40] Pierre Fernandez, Alexandre Sablayrolles, Teddy Furon, Hervé Jégou, and Matthijs Douze. Watermarking images in self-supervised latent spaces. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[41] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. arXiv preprint arXiv:1812.02230, 2018.

[42] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In Proceedings of the European conference on computer vision (ECCV), pages 35–51, 2018.

[43] Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. Bayesian representation learning with oracle constraints. arXiv preprint arXiv:1506.05011, 2015.

[44] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In International conference on machine learning, pages 2649–2658. PMLR, 2018.

[45] Yexun Zhang, Ya Zhang, and Wenbin Cai. Separating style and content for generalized style transfer. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8447–8455, 2018.

[46] Hadi Kazemi, Seyed Mehdi Iranmanesh, and Nasser Nasrabadi. Style and content disentanglement in generative adversarial networks. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 848–856. IEEE, 2019.

[47] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4422–4431, 2019.

[48] Gihyun Kwon and Jong Chul Ye. Diagonal attention and style-based gan for content-style disentanglement in image generation and translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13980–13989, 2021.

[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015.

[51] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014.

[52] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision, pages 3730–3738, 2015.

[53] Justin N. M. Pinkney. Pokemon blip captions. https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/, 2022.

[54] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245, 2024.

[55] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-$\backslash alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426, 2023.

[56] Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Universal watermarking for arbitrary resolution images. arXiv preprint arXiv:2311.18297, 2023.

 https://doi.org/10.52202/079017-1774

[57] Tu Bui, Shruti Agarwal, Ning Yu, and John Collomosse. Rosteals: Robust steganography using autoencoder latent space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 933–942, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In this paper, the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work performed in this paper, and explore valuable research directions related to the topic for the future.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-1774

Answer: [Yes]

Justification: The paper provides a complete set of assumptions and a thorough (and correct) proof for each theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have detailed the algorithm's specifics in both the paper and the appendix to ensure the reproducibility of the experimental results presented in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code for reproducing main experiments is available at: https://github.com/Hlufies/ZWatermarking.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have provided all training and testing details in the appendix (such as data splits, hyperparameters, selection methods, optimizer types, etc.). .

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In our experiments conducted across five datasets, each repeated five times under consistent settings, we observed mean errors within 1%, leading us to employ 1-sigma error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

https://doi.org/10.52202/079017-1774

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information on the computer resources required to reproduce each experiment in the appendix of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research conduct in the paper adheres to the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both the potential positive societal impacts and negative societal impacts of our work in the paper.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper may poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite and provide detailed descriptions of benchmark datasets and some related works in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

https://doi.org/10.52202/079017-1774

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper dose not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1774