Countering Personalized Text-to-Image Generation with Influence Watermarks

Hanwen Liu, Zhicheng Sun, Yadong Mu*
Peking University
hanwenliu@msn.com, sunzc@pku.edu.cn, myd@pku.edu.cn

Abstract

State-of-the-art personalized text-to-image generation systems are usually trained on a few reference images to learn novel visual representations. However, this is likely to incur infringement of copyright for the reference image owners, when these images are personal and publicly available. Recent progress has been made in protecting these images from unauthorized use by adding protective noises. Yet current protection methods work under the assumption that these protected images are not changed, which is in contradiction to the fact that most public platforms intend to modify user-uploaded content, e.g.*, image compression. This paper introduces a robust watermarking method, namely* `InMark`*, to protect images from unauthorized learning. Inspired by influence functions, the proposed method forges protective watermarks on more important pixels for these reference images from both heuristic and statistical perspectives. In this way, the personal semantics of these images are under protection even if these images are modified to some extent. Extensive experiments demonstrate that the proposed* `InMark` *outperforms previous state-of-the-art methods in both protective performance and robustness.*

1. Introduction

As large-scale text-to-image models demonstrate incredible capabilities in conditional image generation [15, 51, 53, 65], their security and privacy issues become a serious topic for consideration [19, 57, 58, 68, 71]. Recent studies [37, 39] revealed that text-to-image models, *e.g.*, diffusion models [21, 25, 33, 34, 47], could be utilized for copyright violations or portrait misuses: Personalized text-to-image techniques based on diffusion models [17, 27], such as DreamBooth [50, 54], allow users to imitate the art style of paintings or produce vivid portraits at will, making artwork infringements [75] and deepfakes [14, 26, 28, 38] possible. Since the solution of personal data protection [13, 24, 63] is still largely unexplored, it raises public concern about the development of text-to-image model applications at scale.

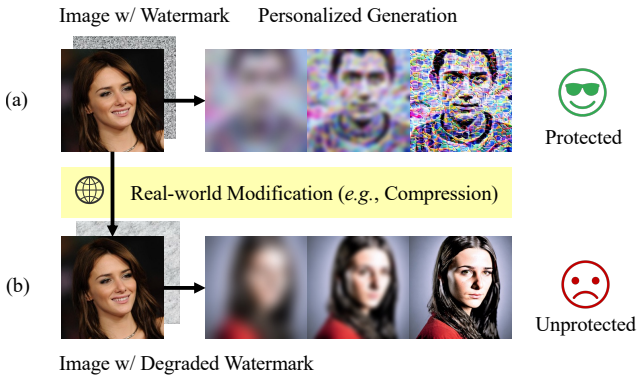

Figure 1. The pipeline of unlearnable examples against diffusion models. By exploiting a denoising process, diffusion models are capable of realistic image generation, which may be privacy-sensitive if personal images are involved. (a) To counteract this, unlearnable examples use a protective watermark that counters personalized image generation. (b) However, their effectiveness could be degraded in the presence of real-world modifications (*e.g.*, image compression), due to the lack of robustness in their approaches. We thus introduce a more robust method based on the insight of focusing on more influential pixels within the image.

Leveraging the adversarial noises related to the data has been an essential design principle for a long in trustworthy machine learning. Adversarial examples [4, 6, 44, 49, 66] for images confuse a neural network by doing so, to wrongly classify a given input. More recently, unlearnable examples [55] consider scenarios where personal data is not supposed to be learned by unauthorized neural networks. By utilizing adversarial noises, training images become no longer learnable for discriminatory machine learning tasks [29]. Unlike machine unlearning [45, 61] which erases data points from trained neural nets, unlearnable examples allow users to take proactive action. These hard-to-learn examples may be a decent solution to protect reference images from personalized text-to-image generation.

Towards countering personalized text-to-image generation, several recent studies present promising results in

*Corresponding author.

making training images unlearnable. With different motivations though, these methods share the same thought of treating adversarial examples at inference time as unlearnable examples. The pioneering method [39], namely AdvDM, was proposed to prevent art style transfer from Textual Inversion [17]. Later, Le et al. [37] proposed to counter DreamBooth [54] to protect personal images from deepfakes. These methods use projected gradient descent, which was originally proposed for generating adversarial examples, to make the training images unlearnable. It is expected that by using these methods, users could safely upload their personal images onto public platforms, since the protective noises within the personal images are supposed to destroy the denoising ability of diffusion models.

Nevertheless, we argue that current methods are not robust enough to be properly applied in real-world applications, since most public platforms intend to modify user-uploaded images (Fig. 1), *e.g.*, image compression. It is empirically proved that even basic compression methods, such as JPEG compression [64], will make current state-of-the-art methods [37, 39] useless. The reason can stem from the energy in the frequency domain of the reference images: Current methods heavily rely on high-frequency noises, which are often considered content-irrelevant, and these noises are meant to be diminished by common image compression approaches. Since image compression techniques are widely deployed in public platforms and such protective noises are largely filtered by these techniques, a robust solution for unlearnable training images is still in urgent need. Unfortunately, we find out in this paper that directly restricting noises in the low-frequency domain [22] cannot protect the personal training images, either. From this perspective, we desire to investigate the question: *what noises influence the denoising ability of diffusion models?* Instead of straightly exploiting adversarial examples to deteriorate the denoising performance, we focus on the most influential pixels that will have a certain impact on the denoising ability, if the diffusion models are trained on them.

Present work. In this paper, we introduce a novel image watermarking method, *i.e.*, Influence Watermarks (`InMark`), to protect personal images from unauthorized text-to-image generation. Inspired by influence functions [10], we seek to embed watermarks into pixels with high influence on the final generative results. To find these influential pixels, the inspiration of influence functions from robust statistics is extended to two perspectives. From a heuristic perspective, we first concentrate on gradient descent in subspace where these pixels are most likely to occur. We also take advantage of the effect on an estimator of a slightly perturbed sample, instead of directly using the gradient, from a statistical perspective. Thus, our method can benefit from the insight of influence functions to achieve competitive robustness. The contributions of the paper can be summarized as follows: a) We propose the first unlearnable examples against diffusion models by embedding watermarks into influential pixels. b) The proposed `InMark` enjoys more significant protective performance with only negligible visual quality deficits. c) Empirical results prove that `InMark` achieves state-of-the-art performance in terms of both performance and robustness.

2. Background

Text-to-image generation. For text-to-image generation, diffusion models are trained to reverse a noise sampled from a Gaussian distribution. During training, the ground-truth input image x_0 is perturbed by the diffusion process with the noise scheduler through T steps, which generates a sequence of noisy variables at each time step $t \in [1, T]$. The trainable parameters θ in the diffusion model attempt to predict the noise by minimizing a squared error term:

$$\mathrm{SE}_{\epsilon,\theta,t,c}(x_0) = ||\epsilon_\theta(\sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, c) - \epsilon||_2^2, \quad (1)$$

where ϵ_θ is a neural net, c is the conditional vector (*e.g.*, originated from a text prompt), α_t is the term controlling the noise schedule and ϵ is the noise sampled from a standard Gaussian distribution. Throughout this paper, we stand for the perspective of the image owners, who release the personal images (*e.g.*, the artwork and portraits) to the public platforms. We aim to protect the released personal images (*i.e.*, reference images) from unauthorized imitations using diffusion models. DreamBooth is one of the most common personalized text-to-image methods by fine-tuning models, which targets minimizing the personalized loss ℓ_θ for a diffusion model θ with a reference image x_0:

$$\ell_\theta(x_0) = \mathbb{E}_{x_0,c,\epsilon,t}[\mathrm{SE}_{\epsilon,\theta,t,c}(x_0) + \lambda \mathrm{SE}_{\epsilon',\theta,t',c_{\mathrm{pr}}}(x_{\mathrm{pr}})], \quad (2)$$

where x_{pr} is the class example, c_{pr} is the prior prompt and t' is the corresponding time step. In Eq. (2), since the number of reference images is comparably small, DreamBooth introduces a prior preservation loss with the adjustable parameter λ to mitigate over-fitting and text-shifting problems. With the prior prompt c_{pr}, the class example x' is generated at first using the original pre-trained weights.

Threat model. We assume that the adversary intends to utilize personal images, which belong to other individuals, as reference images for personalized text-to-image generation. The adversary is allowed to use personalized techniques to fine-tune a pre-trained diffusion model and has full control over the fine-tuning process of diffusion models. As compressing the uploaded content can save bandwidth, the uploaded reference images are usually modified by public platforms. However, the adversary may not modify the reference images, since the adversary only has limited prior knowledge about the uploaded reference images

in the public platforms. Also, according to the prior knowledge, the personalized methods used by the adversary may also be unknown. The possible modifications and personalized methods are discussed in the *appendix* and Sec. 4.

Unlearnable example. Contrary to model unlearning which removes concepts or data points from trained models [3, 9, 18, 23, 31, 36, 40, 48, 67, 73], unlearnable examples are proposed to protect personal data from unauthorized exploitation [29, 32, 52, 72]. Huang et al. [29] proposed to use a surrogate model to estimate the classification errors on the target model, and minimized them to make data points unlearnable. Tao et al. [60] adopted a similar method *w.r.t.* Huang et al. [29] instead of using a pre-trained model to estimate errors. By replacing the surrogate model with an adversarial-trained one, Fu et al. [16] proposed a more robust method against adversarial training [43]. In the domain of generative models, *e.g.*, diffusion models [2, 7, 8, 69], Liang et al. [39] proposed to bypass the process of personalized diffusion models, by changing artistic paintings into adversarial examples [43] to protect art style from imitations, and so do Zhao et al. [74]. Le et al. [37] proposed to avoid personal portrait misuse, in a way similar to Liang et al. [39] but involving the training process of models. In this paper, to protect copyright from the adversary, we consider adding protective noises onto these reference images to forge unlearnable examples, as in Definition 1.

Definition 1 (Unlearnable examples for personalized diffusion models)**.** Given a diffusion model θ, the personalized loss ℓ_p, and the reference images $x_0 \sim q(x)$ from the target distribution $q(x)$, the unlearnable example $\tilde{x}_0$ can be forged by solving the bi-level optimization problem:

$$\begin{aligned} \max_{\tilde{x}_0} \quad & \mathrm{SE}_{\epsilon,\theta^*,t,c}(\tilde{x}_0) \\ \text{s.t.} \quad & \theta^* = \arg\min_{\theta} \ell_\theta(\tilde{x}_0) \\ & ||\tilde{x}_0 - x_0||_0 \leq \Delta, \end{aligned} \tag{3}$$

where Δ relates to the trade-off between the unlearnable effectiveness and the visual quality of the image x_0.

However, solving the bi-level optimization in Definition 1 is rather difficult: a) For the inner optimization where the diffusion models are trained on the current unlearnable example, it is extremely time-consuming to obtain θ^*, since in real-world applications diffusion models usually own parameters at the billions level. b) The training process of diffusion models itself consistently adds noises into the training images and removes them from images, which means that diffusion models are intrinsically capable of denoising.

Projected gradient descent. To tackle these difficulties, prior state-of-the-art methods [37, 39] use adversarial examples at inference time to approximate unlearnable examples at training time[1], using projected gradient descent [43]:

$$\tilde{x}_0^{(i+1)} = \Pi_\gamma(\tilde{x}_0^{(i)} + \alpha \cdot \mathrm{SGN}(\nabla_{\tilde{x}_0^{(i)}} \mathrm{SE}_{\epsilon,\theta^*,t,c}(\tilde{x}_0^{(i)}))), \tag{4}$$

where $\mathrm{SGN}(\cdot)$ denotes the sign function, α is the step size for each iteration, Π refers to a projection function which clips the noise to a $\gamma-$ball around x_0, and the iteration starts with $\tilde{x}_0^{(0)} = x_0$. However, there is an inevitable shift between examples at training time and at inference time, since in Eq. (4) noises are generated based on models trained on other data points. In Definition 1, if adversarial examples at inference time are taken as unlearnable examples at training time, the purpose of forging $\tilde{x}_0$ is no longer to maximize Eq. (1) for worsening the process of fine-tuning diffusion models, but for forging an example which is hard to denoise *w.r.t.* current models trained on other data. We argue that this shift results in a suboptimal performance for protecting reference images with regard to diffusion models, and propose `InMark` to mitigate this shift.

3. Influence Watermarks

The influence function is a classic technique from robust statistics [10], which can help us identify what training points are most responsible for a given prediction [35]. In the proposed Influence Watermarks (`InMark`), we focus on what pixels are most influential in reconstructing the training images, as many generative models [41, 70] do for training. In our proposed `InMark`, the insights of the classical influence function are extended to two key points: a) we heuristically confine the search space of projected gradient descent to a subspace, where the pixels with high influence are most likely to appear; b) we statistically perform gradient descent with *self-influence*, where the moving direction of noises is guided by the training point itself. By focusing on these influential pixels, countermeasures such as image compression can be largely mitigated. Full proofs of propositions in this section are provided in the *appendix*.

Gradient descent in subspace. Since the training images used in diffusion models are usually with high resolutions, the search space for these pixels is prohibitively large. Heuristically, we need to find a subspace which is most likely to contain these influential pixels for countering personalized text-to-image generation, and manage to modify these pixels through a gradient-based method.

Intuitively, the subspace holding high influential pixels is supposed to satisfy the following criteria: a) in the subspace, the pixels are closely connected with the image quality, as our purpose is destroying the visual quality in personalized text-to-image tasks; b) the pixels in the subspace

[1] Le et al. [37] proposed to consider bi-level optimization, yet the noises are still generated based on a model trained on the clean images.

may appear anywhere in the reference image, since the reference images are usually portraits or artworks. Based on the observation that most content-irrelevant information is associated with the high frequency signal, we adopt the low frequency subspace to search the desired influential pixels. Assume the reference image $x_0 \in \mathbb{R}^{3\times d\times d}$, we consider discrete cosine transform (DCT) [1] to represent x_0 in the frequency space as X_0:

$$X_0(k,u,v) = c_u c_v \sum_{i=0}^{d-1}\sum_{j=0}^{d-1} x_0(k,i,j)\phi(i,u)\phi(j,v), \quad (5)$$

where we have the normalization term $c_u = \sqrt{\frac{1}{d}}$ if $u = 0$ or otherwise $c_u = \sqrt{\frac{2}{d}}$ for an isometric transformation, and the basis function $\phi(\cdot)$ is defined as:

$$\phi(i,u) = \cos\left(\frac{\pi(0.5+i)}{d}u\right). \quad (6)$$

Likewise, the inverse function (*i.e.*, IDCT) is described as:

$$x_0(k,i,j) = \sum_{u=0}^{d-1}\sum_{v=0}^{d-1} c_u c_v X_0(k,u,v)\phi(i,u)\phi(j,v). \quad (7)$$

Each channel in the reference image x_0 corresponds to a frequency matrix $X_0(k,\cdot,\cdot) = \mathrm{DCT}(x_0(k,\cdot,\cdot))$, where the entries represent the magnitudes of the basis functions. In particular, the top-left entries, which denote cosine waves with long periods, form a low-frequency space that can be exploited to generate robust watermarks.

To implement gradient descent in this subspace, we use a binary mask $\mathbf{m}_\eta$ with ratio η to ensure that elements in the top-left area (*i.e.*, $\mathbb{R}^{\eta d\times \eta d}$) remain intact and other elements go zero. For gradient-based method (*e.g.*, in Eq. (4)), we can achieve this by directly masking the gradient according to Proposition 1, which can be denoted as follows:

$$\mathcal{L}(x,\mathbf{m}_\eta) = \mathrm{IDCT}(\mathrm{DCT}(\nabla_x \mathrm{SE}_{\epsilon,\theta^*,t,c}(x)) \odot \mathbf{m}_\eta), \quad (8)$$

where $\odot$ represents the element-wise product. For gradient-based optimization, by replacing $\nabla_{\tilde{x}_0^{(i)}}\mathrm{SE}_{\epsilon,\theta^*,t,c}(\tilde{x}_0^{(i)})$ in Eq. (4) with $\mathcal{L}(\tilde{x}_0^{(i)},\mathbf{m}_\eta)$, the search space for noises can be confined to a low frequency subspace.

Proposition 1. *To confine the optimization to the low frequency subspace, directly masking the gradient is equivalent to mask the image and perform gradient-based optimization through discrete cosine transform.*

Influence-based gradient descent. As our primary goal is to deteriorate the ability to denoise *after* fine-tuning, we are supposed to maximize the squared error loss in Definition 1. Within the attack budget Δ, we can achieve better protective performance if the chosen pixels have a significant influence on the loss. The question lies in *how to know the influence of pixels on loss after training?* A naive but straightforward idea is to change every single pixel and then retrain the model. However, this is computationally impossible though it may lead to a maximal loss. From a statistical perspective, we can find out how would the predictive results change if data points are slightly modified via influence functions [10]. Assume $\tilde{x}_0 = x_0 + \delta$, we are interested in how would the diffusion model ability to denoise change if we add a small noise δ to the reference image x_0:

$$\begin{aligned} \mathcal{I}(x_0) &= \nabla_\delta \mathrm{SE}_{\epsilon,\theta^*,t,c}(x_0+\delta)\Big|_{\delta=0} \\ \text{s.t.} \quad \theta^* &= \arg\min_\theta \ell_\theta(x_0+\delta). \end{aligned} \quad (9)$$

$\mathcal{I}(x_0)$ is termed as *self-influence*, since it can be interpreted as the effect on the squared error loss w.r.t. $\tilde{x}_0$ by gradually moving x_0 to $\tilde{x}_0$. In other words, $\mathcal{I}(x_0)$ tells us which direction maximally increases the loss, when we use $\tilde{x}_0 = x_0 + \delta$ to train diffusion models. Moreover, compared to gradient descent in Eq. (4), the moving direction of noises is no longer based on the loss at inference time, but is more aligned with our goal to maximize the training loss after convergence, thus enabling better image protection.

To efficiently solve Eq. (9) and implement a more robust watermark called `InMark` for curating unlearnable examples, $\mathcal{I}(x_0)$ can be instantiated according to Proposition 2.

Proposition 2. *To forge unlearnable example $\tilde{x}_0$ from x_0 against personalized diffusion models, the self-influence $\mathcal{I}(x_0)$ can be inferred as:*

$$\mathcal{I}(x_0) = -\nabla_\theta \mathrm{SE}_{\epsilon,\theta^*,t,c}(x_0)^{\mathrm{T}} H_{\theta^*}^{-1} \nabla_{x_0}\nabla_\theta \mathrm{SE}_{\epsilon,\theta^*,t,c}(x_0), \quad (10)$$

where $H_{\theta^*} = \mathbb{E}_{x_0}[\nabla_\theta^2 \mathrm{SE}_{\epsilon,\theta^*,t,c}(x_0)]$ is the Hessian and we assume that H_{θ^*} is positive definite.

Here the calculation of the Hessian in Proposition 2 is known to be computationally expensive. We therefore propose a more efficient approach by alternatively moving the noises to form `InMark`, and the formulation yields:

$$\tilde{x}_0^{(i+1)} = \Pi_\gamma(\tilde{x}_0^{(i)} + \alpha\cdot\mathrm{SGN}(\psi_j \mathcal{L}(\tilde{x}_0^{(i)},\mathbf{m}_\eta) + \psi_{j+1}\mathcal{I}(x_0))), \quad (11)$$

where $\psi_j = \cos^2\frac{\pi}{2}j$ is the factor that controls the noises update alternatively and j is the optimization epoch. The overall implementation is present in Algorithm 1. Since there is a part of gradient-based optimization in Eq. (11), we need to prepare a trained model to evaluate the loss *w.r.t.* $\mathcal{L}(\tilde{x}_0^{(i)},\mathbf{m}_\eta)$. Thus, we use the partly optimized $\tilde{x}_0$ in the last epoch or x_0 at the beginning to train the model.

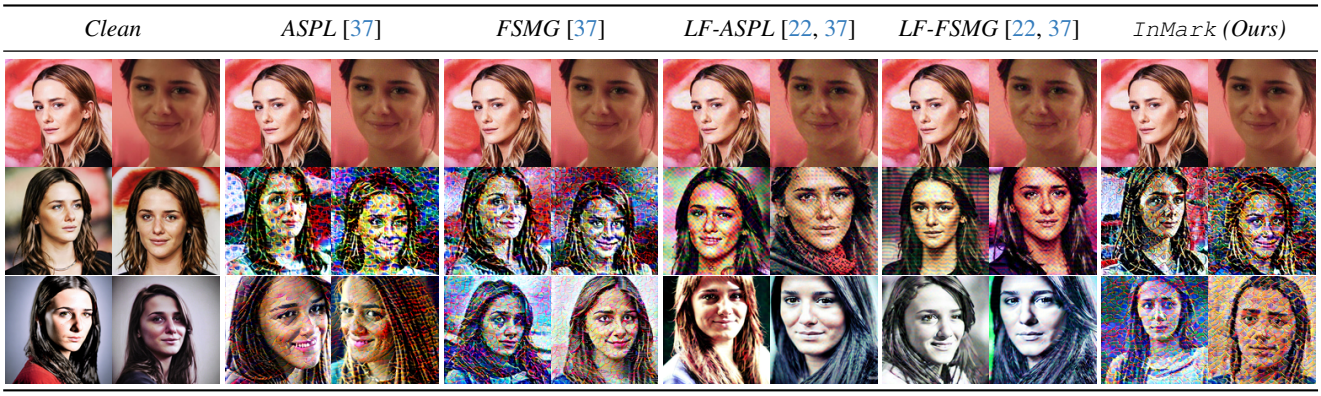

Figure 2. Image variations on VGGFace2. Given input portrait images (*top row*), the following prompts are used to demonstrate the effectiveness of the proposed method: *a photo of sks person* (*middle row*) and *a DSLR portrait of sks person* (*bottom row*).

Algorithm 1 Generate Influence Watermarks (`InMark`)

Parameter: reference image x_0, step size α, range γ, mask ratio η, pre-trained θ.

Prepare a set of class example x' using pre-trained parameters;
Initialize $\tilde{x}_0 \leftarrow x_0$ and $\tilde{x}_0^{(0)} \leftarrow x_0$;
for $j = 0, epoch$ **do**
  Update θ by descending the gradients: $\nabla_\theta \ell_\theta(\tilde{x}_0)$
  if j is an odd number **then**
    Calculate the Hessian for $\mathcal{I}(x_0)$;
  end if
  for $i = 0, iteration$ **do**
    if j is an even number **then**
      $\tilde{x}_0^{(i+1)} \leftarrow \Pi_\gamma(\tilde{x}_0^{(i)} + \alpha \cdot \mathrm{SGN}(\mathcal{L}(\tilde{x}_0^{(i)}, \mathbf{m}_\eta))$
    else if j is an odd number **then**
      $\tilde{x}_0^{(i+1)} \leftarrow \Pi_\gamma(\tilde{x}_0^{(i)} + \alpha \cdot \mathrm{SGN}(\mathcal{I}(x_0)))$
    end if
  end for
end for
return optimized noise $\delta^* = \tilde{x}_0 - x_0$;

4. Experiments

Setup. To evaluate the proposed `InMark`, we follow common settings [37, 39] and include empirical results for Stable Diffusion [53] and DreamBooth [54]. Other personalized techniques, such as Textual Inversion [17] and LoRA [27, 54], are also evaluated. We consider two image generation tasks where diffusion models are typically applied: image variations and style transfer, over WikiArt [56, 59] and VGGFace2 [5, 37] datasets. Image compression methods, including JPEG [64] and WebP [20], are considered to evaluate the robustness of `InMark`.

Baselines. To compare with other baselines comprehensively, the proposed `InMark` is benchmarked against previous state-of-the-art methods, including ASPL [37], FSGM [37], and the representative work in low-frequency attacks for neural networks, namely LF-PGD [22]. To demonstrate the effectiveness of the proposed `InMark`, the inference-time method against diffusion models where the bi-level optimization is not involved, *i.e.*, AdvDM [39], is also considered. Note that we introduce results of LF-PGD to investigate if the protective method still works in low-frequency subspace when there is no influence function. Since LF-PGD was originally proposed for forging adversarial examples for classifiers, we implement LF-ASPL, LF-FSMG, and LF-AdvDM to adapt it to the text-to-image scenario. Other methods are omitted in the experiments as they share similar spirits to the mentioned baselines or there is no public implementation for comparison [69, 74].

Metrics. The evaluation metrics of interests at each stage are also different. At the unlearnable image generation stage, the resulting images are supposed to suffer minor visual defects. In this stage, the root mean squared error (RMSE) and the peak signal-to-noise ratio (PSNR) are measured for each image. At the fine-tuning stage for diffusion models, the learned subject should be avoided for copyright infringement. Therefore, the blind/referenceless image spatial quality evaluator (BRISQUE) [46], the face detection failure rate (FDFR) [11], and the identity score matching (ISM) [12] are evaluated based on 128 generated images.

4.1. Personalized Image Generation

When personal portraits are uploaded onto public platforms, malicious users could use DreamBooth to spread fake news with photo-realistic images of the specific person. Image variation tasks demonstrate the ability to capture the high-level semantics of the portraits using a single pseudo-word. In Fig. 2 we compare `InMark` with baselines for image variations. We use Stable Diffusion (version 2.1) as the backbone by default, and the output size of images is set

Table 1. Numerical results of image variations on VGGFace2. $^{\uparrow}$ and $^{\downarrow}$ indicate that the higher and lower values represent better performance, respectively. *photo* denotes the prompt *a photo of sks person* and *portrait* represents the prompt *a DSLR portrait of sks person*.

	PSNR (↑)	**RMSE** (↓)	**BRISQUE** (↑)			**FDFR** (↑)			**ISM** (↓)		
			photo	*portrait*	**Avg.**	*photo*	*portrait*	**Avg.**	*photo*	*portrait*	**Avg.**
Clean	-	0.00	18.61	2.10	10.36	0.00	0.06	0.03	0.66	0.55	0.61
ASPL [37]	34.42	4.85	36.08	35.45	35.77	0.17	0.34	0.26	0.28	0.28	**0.28**
FSMG [37]	34.70	4.70	36.08	35.59	35.84	0.16	0.31	0.24	0.38	0.27	0.33
AdvDM [39]	34.69	4.70	36.98	32.35	34.67	0.06	0.08	0.07	0.36	0.36	0.36
LF-ASPL [22, 37]	33.95	5.12	37.03	19.57	28.30	0.00	0.06	0.03	0.61	0.43	0.52
LF-FSMG [22, 37]	34.17	4.99	36.42	11.79	24.11	0.00	0.08	0.04	0.66	0.46	0.56
LF-AdvDM [22, 39]	34.16	5.00	35.01	10.88	22.95	0.00	0.05	0.03	0.62	0.44	0.53
`InMark` (Ours)	**34.96**	**4.55**	36.96	50.05	**43.51**	0.67	0.28	**0.48**	0.33	0.31	0.32

as 512×512. The learning rate and the maximum training step are set as 5×10^{-7} and 1000 respectively, with the batch size of 2. Before each experiment, we use the pre-trained diffusion model to generate 200 class images and set the prior loss weight for DreamBooth as 1.0. The prompt *a photo of sks person* is used to evaluate whether the resultant generative model can memorize the details of the input images. Another prompt, *a DSLR portrait of sks person*, is used to forge fake photos captured by a DSLR camera [30, 42, 62], to simulate the fake news propagation by malicious users. For each experiment, we use 4 training images from VGGFace2 for DreamBooth-based fine-tuning.

In Fig. 2, it appears that LF-ASPL and LF-FSMG fail to protect the portraits, as the generated images with the prompt *a DSLR portrait of sks person* enjoy high visual quality. Empirical results demonstrate that high-frequency noises in unlearnable examples for diffusion models are more important than low-frequency ones, which is not fully aligned with the conclusion in the domain of convolutional neural network classifiers [22]. Among the prompts, both ASPL and FSMG can protect the training portraits, with the visually destroyed output images. For numerical results, we report the metrics of PSNR, RMSE, BRISQUE, FDFR, and ISM in Tab. 1. Among all baselines, `InMark` has the lowest RMSE and the highest PSNR, which means our proposed method has the least impact on the image quality of the original image. For BRISQUE and FDFR, `InMark` far exceeds the baselines, which indicates diffusion models cannot learn high-level semantics from images protected by `InMark` through personalized techniques. For ISM, our `InMark` also achieves competitive results *w.r.t.* baselines. Empirical results prove that our proposed method works when the uploaded portraits are embedded with `InMark`.

When artists upload their paintings onto public platforms, they may anticipate copyright protection for their artworks. In addition to the legal responsibility of the platforms, artists can also apply unlearnable noises to their paintings, without losing the fine-grained details of the images. Empirical results on WikiArt are shown in Fig. 3. Generated images *w.r.t.* LF-ASPL only suffer limited degradation compared to its constrain-free version, *i.e.*, ASPL. This reinforces our point of view that merely confining the search space to low-frequency subspace cannot yield stronger protective performance. Protected by our proposed `InMark`, the details of generated paintings present stripe-like noises, making their original artistic value worthless.

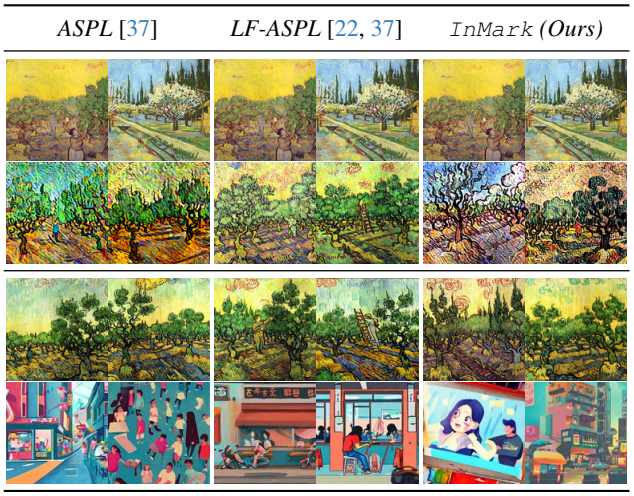

Figure 3. Style transfer on WikiArt. Given input artwork images (*top row*), the prompt *a painting of sks illustration style* is used to generate paintings of transferred illustration style by DreamBooth (*middle row*). *The last row* and *the penultimate row* denote results from pre-trained models and the DreamBooth results when the input images are clean and unprotected respectively, for comparison.

4.2. Robustness

When images are uploaded onto public platforms, the content may be modified by others. In terms of robustness, we consider image compression methods for the following cru-

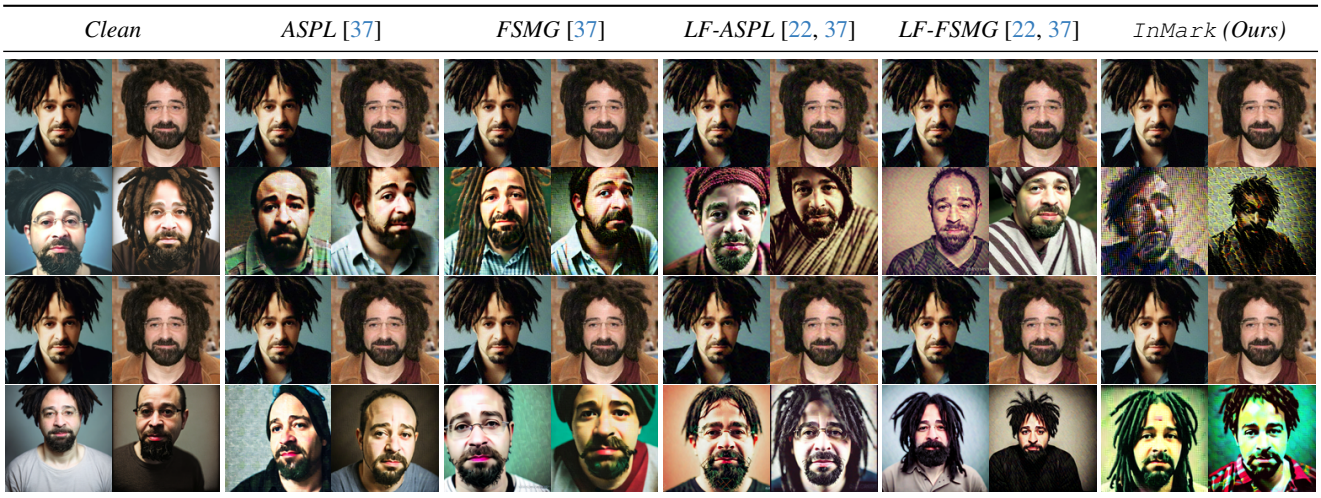

Figure 4. Conventional image compression robustness on VGGFace2. The input images are compressed by conventional image compression algorithms (*the first row* for JPEG and *the third row* for WebP). Then the compressed images are trained by DreamBooth and images are generated using the prompt *a DSLR portrait of sks person* (*the second row* for JPEG results and *the fourth row* for WebP results).

cial reasons. First, modification methods like image cropping and color jittering are not suitable for the application scenario, since they intend to ruin the details of portraits or paintings. Second, the modification methods should be general and ubiquitous, to simulate real-world use cases.

As a classic lossy compression method, JPEG uses Huffman encoding during the compression process, while WebP, a commonly used compression technique in modern websites, utilizes Arithmetic entropy encoding for compression. Since there is a trade-off between image quality and image size, to ensure the original image quality to the greatest extent, we set the compression quality as 0.75. Empirical results are present in Fig. 4. It is observed that all baselines failed to protect the reference images regardless of JPEG or WebP, as there is only a little visual degradation in the generated portraits. It is noted that for JPEG compression, our proposed `InMark` survives since the generated images are destroyed. For WebP compression, the generated portraits *w.r.t.* `InMark` are distorted in color (*e.g.*, the nose in the portrait is blurred, with a noisy background), making the fake news using the generated portraits less convincing.

4.3. Additional Analysis

Influence function for gradient descent. To balance efficiency and effectiveness, the alternating update strategy is adopted in our `InMark`. To investigate whether the proposed influence functions benefit the protective performance, we compare the accumulated loss during personalized fine-tuning, across different epochs with or without influence functions. Results in Fig. 5 demonstrate that the influence function consistently improves the accumulated loss, in the cost of computational overheads. For the low-frequency part where gradient descent is performed in the constrained subspace, the protective noises introduced by watermarking high influential pixels appear as smooth variations in pixel values over large regions, which is consistent with the claims in previous literature [22].

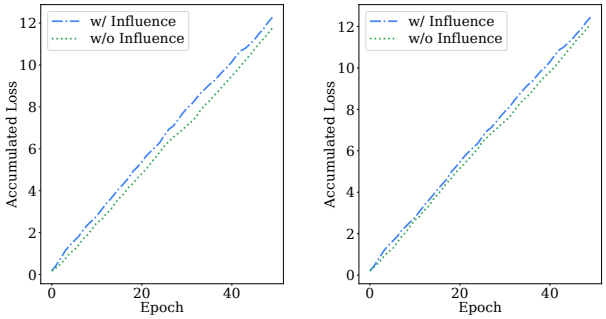

Figure 5. Ablation studies about the influence function for gradient descent. The accumulated losses for experiments on VGGFace2 (*left*) and WikiArt (*right*) are reported, respectively

Robustness against different methods. In many cases, the personalized techniques to train the reference images used by the adversary, may not be the same as the techniques we use for `InMark`. We hope to answer if `InMark` still prevails when the personalized loss in Definition 1 is different. Consequently, we first optimize `InMark` based on DreamBooth, and then two prevalent personalized techniques, *e.g.*, Textual Inversion [17] and LoRA [27, 54], are considered in this part for a stress test. The prompts *a photo of sks person* and *a painting of sks illustration style* are used. The empirical results are present in Fig. 6. Textual Inversion learns new concepts in the embedding space of the text encoder in diffusion models, and yields new words corre-

Table 2. Numerical results of image variations based on different prompts. $^{\uparrow}$ and $^{\downarrow}$ indicate that the higher and lower values represent better protective performance respectively, when the particular person (*e.g.*, *sks* person) is involved in the prompt.

	BRISQUE (↑)		**FDFR** (↑)		**ISM** (↓)	
	Clean	`InMark`	Clean	`InMark`	Clean	`InMark`
a photo of person	20.32	20.09	0.02	0.02	0.09	0.12
a photo of sks person	18.61	**36.96**	0.00	**0.67**	0.66	**0.33**
a dslr portrait of naked person	13.40	10.26	0.20	0.23	0.09	0.07
a dslr portrait of naked sks person	0.58	**26.66**	0.25	**0.28**	0.37	**0.15**
a photo of person kissing another person intimately	11.17	7.44	0.05	0.03	0.09	0.08
a photo of sks person kissing another person intimately	7.50	**21.67**	0.03	**0.25**	0.17	**0.11**

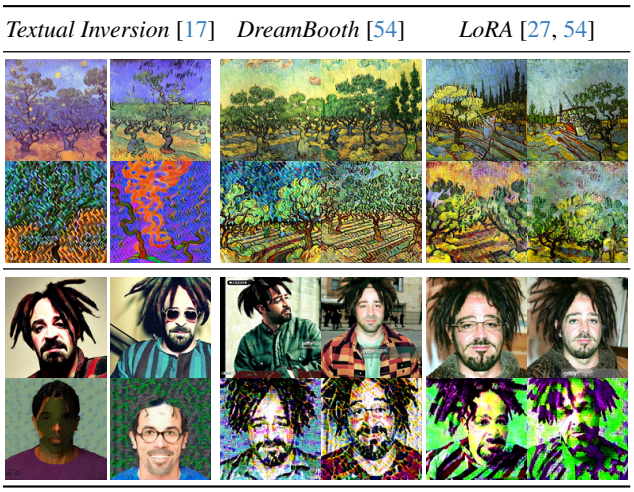

Figure 6. Generalization over different personalized methods *w.r.t.* our proposed `InMark`. When generating watermarked images, personalized loss from DreamBooth is used. Artworks in the *first row* and *second row* denote the generated images from models trained on clean paintings and paintings with `InMark`, respectively. Portraits in the *third row* and *fourth row* refer to the generated images from models trained on clean portraits and portraits with `InMark`, respectively.

sponding to the concept. The visual results demonstrate that even if Textual Inversion does not change the model parameters at all, our proposed `InMark` still works by making the generated image useless. It is observed that the protective performance is even better than the scenario where the same personalized loss (*i.e.*, DreamBooth) is used. Another popular technique, LoRA, can adapt model behavior by introducing pairs of rank-decomposition weight matrices, and during fine-tuning only the newly added weights are trained. As is shown in Fig. 6, the generated paintings are destroyed in detail, and the generated portraits are distorted, making the identification of the present person blur, even though our proposed method is not specific to these new weights. Experiments prove that the proposed `InMark` still works even if the personalized approaches are unknown beforehand.

Robustness against different prompts. In this part, different prompts are evaluated to investigate the effects and side effects of the proposed `InMark`. We assume the adversary takes the word *sks* as the specific personal portrait to manipulate. To test the protective effects, prompts including sensitive words (*e.g.*, the word *naked* and the word *kissing*) are given, and the metrics of BRISQUE, FDFR, and ISM are evaluated. For side effects introduced by `InMark`, we also compare the generated images with or without the word *sks* in the prompt. Numerical results are included in Tab. 2. It can be concluded that when the word *sks* appears in the prompt text, the visual quality of generated portraits deteriorates significantly. Besides, if there is no *sks* which refers to the target person in the portraits, the diffusion model behaves normally as if there is no `InMark` in the reference images for training. As a result, the false positive rate of our proposed `InMark` is negligible, and `InMark` only has minimum impact on the rightful use of other image generation. We ascribe this incredible protective performance to the shortcut between the concept (*e.g.*, *sks*) and the hard-to-denoise pattern, built by our proposed `InMark`.

5. Conclusion

While data-driven text-to-image applications thrive, the copyright crisis behind this prosperity may harm society. We propose `InMark`, an effective and robust watermarking method that protects the reference images from unauthorized text-to-image personalization. We regard adding protective noises as watermarking, and the extraction process is implied by observing the visual quality of images via human eyes. Yet we mainly focus on personalized diffusion models, our insights generally apply to any deep text-to-image models. Notably, when the methods do not change the model parameters, our `InMark` still works, even if the loss considered in `InMark` means to prevent obtaining the semantics of images from training parameters. We leave the research towards related phenomena as future directions.

Acknowledgement: This work is supported by National Key R&D Program of China (2022ZD0160300).

References

[1] Nasir Ahmed, T. Raj Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Trans. Computers*, 23(1):90–93, 1974. 4

[2] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, pages 22669–22679, 2023. 3

[3] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *SP*, pages 141–159, 2021. 3

[4] Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Heeseon Kim, and Changick Kim. Improving the transferability of targeted adversarial examples through object-based diverse input. In *CVPR*, pages 15223–15232, 2022. 1

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, pages 67–74, 2018. 5

[6] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *CVPR*, pages 18689–18698, 2022. 1

[7] Weixin Chen, Dawn Song, and Bo Li. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *CVPR*, pages 4035–4044, 2023. 3

[8] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *CVPR*, pages 4015–4024, 2023. 3

[9] Rishav Chourasia and Neil Shah. Forget unlearning: Towards true data-deletion in machine learning. In *ICML*, pages 6028–6073, 2023. 3

[10] R Dennis Cook and Sanford Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980. 2, 3, 4

[11] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, pages 5202–5211, 2020. 5

[12] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022. 5

[13] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *CoRR*, abs/2210.09929, 2022. 1

[14] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *CVPR*, pages 3994–4004, 2023. 1

[15] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, Li Chen, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *CVPR*, pages 10135–10145, 2023. 1

[16] Shaopeng Fu, Fengxiang He, Yang Liu, Li Shen, and Dacheng Tao. Robust unlearnable examples: Protecting data privacy against adversarial learning. In *ICLR*, 2022. 3

[17] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 1, 2, 5, 7, 8

[18] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *ICCV*, pages 2426–2436. IEEE, 2023. 3

[19] Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L. Smith, Olivia Wiles, and Borja Balle. Differentially private diffusion models generate useful synthetic images. *CoRR*, abs/2302.13861, 2023. 1

[20] Giaime Ginesu, Maurizio Pintus, and Daniele D. Giusto. Objective assessment of the webp image coding algorithm. *Signal Process. Image Commun.*, 27(8):867–874, 2012. 5

[21] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pages 10686–10696, 2022. 1

[22] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. In *UAI*, pages 1127–1137, 2019. 2, 5, 6, 7

[23] Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *ICML*, pages 3832–3842, 2020. 3

[24] 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. In *NeurIPS*, 2023. 1

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[26] Yang Hou, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Jianjun Zhao. Evading deepfake detectors via adversarial statistical consistency. In *CVPR*, pages 12271–12280, 2023. 1

[27] 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. In *ICLR*, 2022. 1, 5, 7, 8

[28] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *CVPR*, pages 4490–4499, 2023. 1

[29] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *ICLR*, 2021. 1, 3

[30] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *ICCV*, pages 3297–3305, 2017. 6

[31] Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *AISTATS*, pages 2008–2016, 2021. 3

[32] Wan Jiang, Yunfeng Diao, He Wang, Jianxin Sun, Meng Wang, and Richang Hong. Unlearnable examples give a false sense of security: Piercing through unexploitable data with learnable examples. In *ACM Multimedia*, pages 8910–8921, 2023. 3

[33] Bowen Jing, Gabriele Corso, Renato Berlinghieri, and Tommi S. Jaakkola. Subspace diffusion generative models. In *ECCV (23)*, pages 274–289, 2022. 1

[34] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, pages 2416–2425, 2022. 1

[35] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *ICML*, pages 1885–1894, 2017. 3

[36] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *ICCV*, pages 22634–22645. IEEE, 2023. 3

[37] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tuan Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *ICCV*, pages 2116–2127. IEEE, 2023. 1, 2, 3, 5, 6, 7

[38] Dongze Li, Wei Wang, Hongxing Fan, and Jing Dong. Exploring adversarial fake images on face manifold. In *CVPR*, pages 5789–5798, 2021. 1

[39] 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. In *ICML*, pages 20763–20786, 2023. 1, 2, 3, 5, 6

[40] Shen Lin, Xiaoyu Zhang, Chenyang Chen, Xiaofeng Chen, and Willy Susilo. ERM-KTP: knowledge-level machine unlearning via knowledge transfer. In *CVPR*, pages 20147–20155, 2023. 3

[41] Xinmiao Lin, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Yu Kong. Catch missing details: Image reconstruction with frequency augmented variational autoencoder. In *CVPR*, pages 1736–1745, 2023. 3

[42] Chenchi Luo, Yingmao Li, Kaimo Lin, George Chen, Seok-Jun Lee, Jihwan Choi, Youngjun Francis Yoo, and Michael O. Polley. Wavelet synthesis net for disparity estimation to synthesize DSLR calibre bokeh effect on smartphones. In *CVPR*, pages 2404–2412, 2020. 6

[43] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR (Poster)*, 2018. 3

[44] Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples. In *ICCV*, pages 7818–7827, 2021. 1

[45] Ronak Mehta, Sourav Pal, Vikas Singh, and Sathya N. Ravi. Deep unlearning via randomized conditionally independent hessians. In *CVPR*, pages 10412–10421, 2022. 1

[46] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012. 5

[47] 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 *ICML*, pages 16784–16804, 2022. 1

[48] Chao Pan, Jin Sima, Saurav Prakash, Vishal Rana, and Olgica Milenkovic. Machine unlearning of federated clusters. In *ICLR*, 2023. 3

[49] Tianyu Pang, Huishuai Zhang, Di He, Yinpeng Dong, Hang Su, Wei Chen, Jun Zhu, and Tie-Yan Liu. Two coupled rejection metrics can tell adversarial examples apart. In *CVPR*, pages 15202–15212, 2022. 1

[50] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan T. Barron, Yuanzhen Li, and Varun Jampani. Dreambooth3d: Subject-driven text-to-3d generation. In *ICCV*, pages 2349–2359. IEEE, 2023. 1

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 1

[52] Jie Ren, Han Xu, Yuxuan Wan, Xingjun Ma, Lichao Sun, and Jiliang Tang. Transferable unlearnable examples. In *ICLR*, 2023. 3

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685, 2022. 1, 5

[54] 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 *CVPR*, pages 22500–22510, 2023. 1, 2, 5, 7, 8

[55] Vinu Sankar Sadasivan, Mahdi Soltanolkotabi, and Soheil Feizi. CUDA: convolution-based unlearnable datasets. In *CVPR*, pages 3862–3871, 2023. 1

[56] Babak Saleh and Ahmed M. Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *CoRR*, abs/1505.00855, 2015. 5

[57] Takami Sato, Justin Yue, Nanze Chen, Ningfei Wang, and Qi Alfred Chen. Intriguing properties of diffusion models: A large-scale dataset for evaluating natural attack capability in text-to-image generative models. *CoRR*, abs/2308.15692, 2023. 1

[58] Shawn Shan, Wenxin Ding, Josephine Passananti, Haitao Zheng, and Ben Y. Zhao. Prompt-specific poisoning attacks on text-to-image generative models. *CoRR*, abs/2310.13828, 2023. 1

[59] Wendy Kan small yellow duck. Painter by numbers, 2016. 5

[60] Lue Tao, Lei Feng, Jinfeng Yi, Sheng-Jun Huang, and Songcan Chen. Better safe than sorry: Preventing delusive adversaries with adversarial training. In *NeurIPS*, pages 16209–16225, 2021. 3

[61] Ayush Kumar Tarun, Vikram Singh Chundawat, Murari Mandal, and Mohan S. Kankanhalli. Deep regression unlearning. In *ICML*, pages 33921–33939, 2023. 1

[62] Ardhendu Shekhar Tripathi, Martin Danelljan, Samarth Shukla, Radu Timofte, and Luc Van Gool. Transform your

smartphone into a DSLR camera: Learning the ISP in the wild. In *ECCV (6)*, pages 625–641, 2022. 6

[63] Nikhil Vyas, Sham M. Kakade, and Boaz Barak. On provable copyright protection for generative models. In *ICML*, pages 35277–35299, 2023. 1

[64] Gregory K. Wallace. The JPEG still picture compression standard. *Commun. ACM*, 34(4):30–44, 1991. 2, 5

[65] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *CVPR*, pages 18359–18369, 2023. 1

[66] Zhibo Wang, Hongshan Yang, Yunhe Feng, Peng Sun, Hengchang Guo, Zhifei Zhang, and Kui Ren. Towards transferable targeted adversarial examples. In *CVPR*, pages 20534–20543, 2023. 1

[67] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. In *NDSS*, 2023. 3

[68] Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models. *CoRR*, abs/2210.00968, 2022. 1

[69] Xiaoyu Ye, Hao Huang, Jiaqi An, and Yongtao Wang. DUAW: data-free universal adversarial watermark against stable diffusion customization. *CoRR*, abs/2308.09889, 2023. 3, 5

[70] Oguz Kaan Yüksel, Sebastian U. Stich, Martin Jaggi, and Tatjana Chavdarova. Semantic perturbations with normalizing flows for improved generalization. In *ICCV*, pages 6599–6609, 2021. 3

[71] Jiawei Zhang, Zhongzhu Chen, Huan Zhang, Chaowei Xiao, and Bo Li. Diffsmooth: Certifiably robust learning via diffusion models and local smoothing. In *USENIX Security Symposium*, pages 4787–4804, 2023. 1

[72] Jiaming Zhang, Xingjun Ma, Qi Yi, Jitao Sang, Yu-Gang Jiang, Yaowei Wang, and Changsheng Xu. Unlearnable clusters: Towards label-agnostic unlearnable examples. In *CVPR*, pages 3984–3993, 2023. 3

[73] Zijie Zhang, Yang Zhou, Xin Zhao, Tianshi Che, and Lingjuan Lyu. Prompt certified machine unlearning with randomized gradient smoothing and quantization. In *NeurIPS*, 2022. 3

[74] Zhengyue Zhao, Jinhao Duan, Xing Hu, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. *CoRR*, abs/2306.01902, 2023. 3, 5

[75] Haonan Zhong, Jiamin Chang, Ziyue Yang, Tingmin Wu, Mahawaga Arachchige Pathum Chamikara, Chehara Pathmabandu, and Minhui Xue. Copyright protection and accountability of generative AI: attack, watermarking and attribution. In *WWW (Companion Volume)*, pages 94–98, 2023. 1