Harnessing Frequency Spectrum Insights for Image Copyright Protection Against Diffusion Models

Supplementary Material

A. More Related Work

A.1. Diffusion Models

Diffusion models, pioneered by Ho et al.'s Denoising Diffusion Probabilistic Model (DDPM) [20], have revolutionized visual generation. Song et al.'s Denoising Diffusion Implicit Models (DDIM) [47] significantly accelerated the generation process, enabling the creation of high-quality images in significantly less time. Building on this foundation, Rombach et al.'s Latent Diffusion Model (LDM) [39] emerged as a powerful framework for high-resolution image synthesis and text-to-image generation. Building upon text-to-image diffusion models, multiple fine-tuning techniques [22, 32, 41] have been developed to incorporate diverse generation tasks (e.g., artistic style or subject-driven image synthesis) with low resource consumption. For example, LoRA [22] adds parallel modules into denoising network and adopts trainable low-rank matrices to compress the original high-dimensional model parameters. DreamBooth [41] works by fine-tuning text-to-image generation model with a few subject images to associate a less frequently used word-embedding with a specific subject, while maintaining the diversity of generated images through class-specific prior preservation loss.

B. More Implementation Details

B.1. UNet-Based Information Enhanceme Module

In Section 4.2.1, we design a UNet-based Information Enhancement Module (IEM) to compensate for watermark information loss during reconstruction process, and the architecture of U-Net is publicly available at https://github.com/ouyangjiahong/image2image-baseline-model). The loss function $\mathcal{L}_{total}$ of the customized HiNet, defined as Eq. (4), is a weighted sum of concealing loss $\mathcal{L}_{con}$, revealing loss $\mathcal{L}_{rev}$, low-frequency wavelet loss $\mathcal{L}_{freq}$, and reconstruction loss $\mathcal{L}_{rec}$. Here, N denotes the total number of training images, and $\mathcal{L}_{rec}$ is defined in Eq. (5).

$$\mathcal{L}_{total} = \lambda_c \mathcal{L}_{con} + \lambda_r \mathcal{L}_{rev} + \lambda_f \mathcal{L}_{freq} + \lambda_{re} \mathcal{L}_{rec} \quad (4)$$

$$\mathcal{L}_{rec} = \sum_{n=1}^{N} \ell_2 \left(I_w^{(n)}, I_{rw}^{(n)} \right) \quad (5)$$

Following [24], $\mathcal{L}_{con}$ denotes ℓ_2 norm between the image I and watermarked image I_w, $\mathcal{L}_{rev}$ represents ℓ_2 between the extract watermark *W'* and watermark *W*, and $\mathcal{L}_{freq}$ denotes ℓ_2 norm between the low-frequency sub-bands of I and I_w. The parameters $\lambda_c = 10.0$, $\lambda_r = 1.0$, and $\lambda_f = 10.0$ remain unchanged in [24], and λ_{re} is set to 10.0.

B.2. Diffusion Architectures

The DDIM [47] and DDPM [20] model used in our experiments is implemented by [35] with the same U-Net hyperparameters as [14]. It is publicly available at https://github.com/phizaz/diffae.git. Specifically, the base channel of the U-Net architecture is set to 128, and the channel multiplier is configured as $[1, 1, 2, 3, 4]$. Additionally, a global attention layer is applied at a 16×16 resolution using a single attention head.

B.3. Other Experimental Settings

We use the Adam optimizer [27] with betas 0.9 and 0.999, as well as epsilon $1 \times e^{-8}$. The learning rate is set as $1 \times e^{-4}$ for 128×128 images and the batch size is set as 32. We adopt the base linear schedule with the diffusion step $T = 1000$. The sampling steps for DDPM, DDIM and Classifier-Free Guidance are set to $T = 1000$, $T = 100$, and $T = 100$. For Classifier-Free Guidance, we select a subset of 50,000 images across 100 categories from ImageNet as the training set, and the guidance scale ω is set as 1.8. The watermark images used in our experiments are all selected from DIV2K [2].

Table 7. The average cosine similarity (*COS*) of spectral features between diffusion-generated images and training images.

Model	Dataset	DFT	DCT	cA	DWT-*avg*
DDIM	FFHQ	0.988	0.986	0.987	0.937
	ImageNet	0.978	0.964	0.966	0.802
	BigGAN	0.984	0.975	0.979	0.808
	StyleGAN2	0.992	0.988	0.991	0.867
DDPM	FFHQ	0.995	0.992	0.991	0.891
	BigGAN	0.984	0.973	0.977	0.842
	StyleGAN2	0.995	0.990	0.992	0.862
Classifier-free	ImageNet	0.967	0.990	0.991	0.824

C. Additional Experiments

C.1. The DWT and DCT Spectra of Training and Generated Images

In this subsection, we further analyze the cA (approximation component), cH (horizontal component) and cV (vertical component) of the DWT [5] coefficients, as well as the

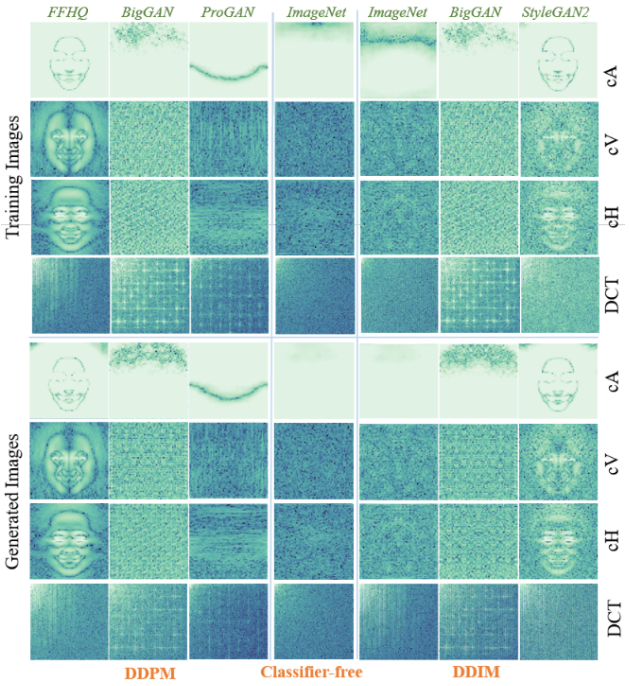

Figure 12. The mean DCT spectra of diffusion-generated and training images.

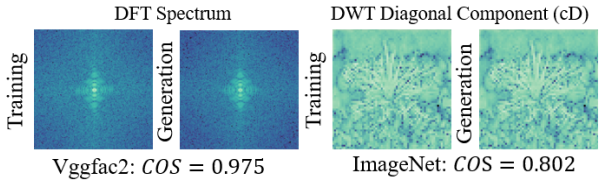

Figure 13. The mean spectra of diffusion-generated and training images.

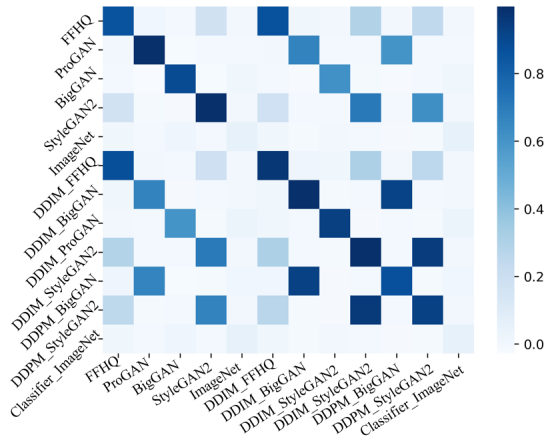

Figure 14. The cosine similarity between the PRNU features of the training and generated images. The part before the "_" symbol denotes the training model, while the part after it indicates the corresponding training dataset.

DCT [3] spectra for all sampled and training sub-datasets. The results, presented in Fig. 12, indicate that the sampled distribution preserves spectral features of the training distribution. We also report the cosine similarity (COS) between the spectral features of the training and generated images for three diffusion models, as illustrated in Tab. 7. For DCT and DFT, the spectral similarities between the training and diffusion-generated images all exceeds 0.95. However, we observe that the spectral similarities of cV, cH and cD components are relatively low (around 0.8). We attribute this primarily to the limited ability of traditional diffusion models to generate high-frequency details in images [36, 37].

We also trained DDIM on 1 million images from VGGFace2 and ImageNet, respectively. For ImageNet, a watermark image was embedded into the DWT spectrum of all training images. The results, shown in Fig. 13, confirm that DDIM-generated images reliably preserve spectrum features of their training data.

C.2. The PRNU Feature of Training and Generated Images

We perform a further statistical analysis on diffusion-generated images using Photo Response Non-Uniformity (PRNU) [17], a fingerprint feature typically used for natural images. We randomly select 5,000 images as the reference set and another 5,000 images as the test set from each training and diffusion-generated sub-dataset, then extract the PRNU feature for each. Next, we calculate the cosine similarity of the PRNU feature within and across sub-dataset, with the results reported in Fig. 14. The PRNU features of most diffusion-generated images exhibit high similarity to those of the corresponding training images, with cosine similarities exceeding 0.6. In contrast, the cosine similarity between the PRNU features of FFHQ [26] images and StyleGAN2 [1] images (trained on FFHQ) is approximately 0.15. However, the cosine similarity between reference and test images from ImageNet [13] is low (approximately 0.05); likewise, the cosine similarity between the PRNU features of images generated by Classifier-Free Guidance [19] and those of ImageNet images is notably low (around 0.04). These findings may offer new insights for the detection and traceability of diffusion-generated images.

C.3. More Decision Threshold of Different Watermarks

We select two additional images from DIV2K as watermarks and randomly sample 50,000 clean images from FFHQ, LSUN, and ImageNet as reference images, respectively. Then we extract watermarks from reference images and compute the cosine similarity between each extracted watermark and the original watermark. The cosine similar-

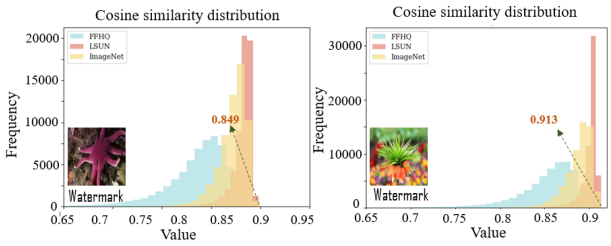

Figure 15. The cosine similarity distribution of extracted and original watermarks for FFHQ, LSUN, and ImageNet.

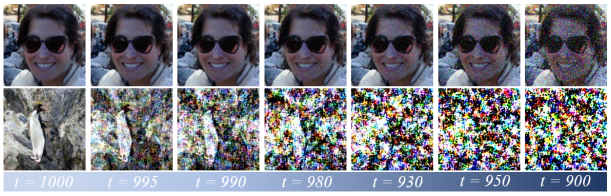

Figure 16. The extracted watermarks of unconditional progressive generation with DDPM sampler. t denotes the sampling step.

Figure 17. The generated images of DDIM, and all images are classified as watermarked images.

ity distribution is demonstrated in Fig. 15. indicating that for clean images, there is an upper threshold for the cosine similarity between the extracted watermark and the original watermark, which is used as the decision threshold γ for CoprGuard.

C.4. Extracted Watermarks of Unconditional Progressive Generation

We also perform watermark extraction for the unconditional progressive generation. We used a DDPM sampler with diffusion steps $T = 1000$, and the results, as shown in Fig. 16, indicate that incomplete image sampling will reduce the quality of the extracted watermarks.

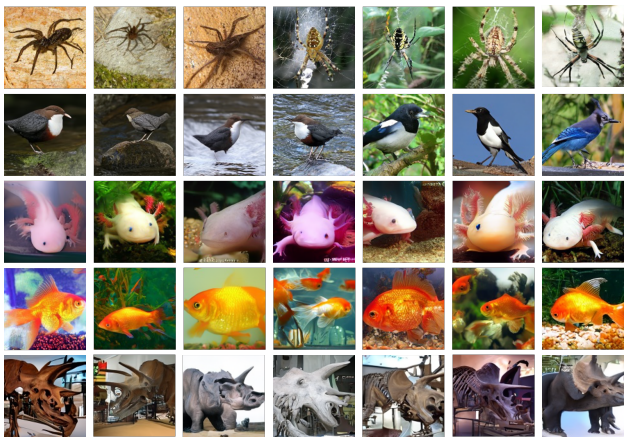

Figure 18. The generated images of Classifier-Free Guidance, and all images are classified as watermarked images.

Figure 19. The generated images of Stable Diffusion. We also show the extracted watermarks and the corresponding cosine similarities (COS).

C.5. More Generated Images of CoprGuard

Fig. 17, Fig. 18 and Fig. 19 show more images generated by DDIM, Classifier-Free Guidance, as well as Stable Diffusion, and all generated images are correctly classified as watermarked images.

C.6. The Impact of Fine-tuning on CoprGuard

We first trained DDIM with $50,000$ watermarked images and then fine-tuned the pre-trained DDIM with 5000 clean images (10% of the training set) and Tab. 8 shows the watermark detection ratio P_u. CoprGuard remains effective after 50 epochs of full-parameters fine-tuning of DDIM.

Table 8. Watermark detection ratio P_u after model fine-tuning.

Epoch	0	5	10	30	50	100
P_u	100%	96.2%	95.7%	92.6%	36.9%	0

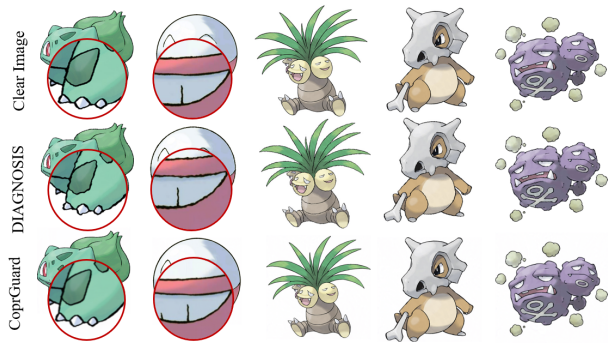

Figure 20. Visual comparison of DIAGNOSIS and CoprGuard.

C.7. Watermarked Image Visualization

Fig. 20 shows visual comparisons between CoprGuard and DIAGNOSIS. The clean images are selected from Pokemon. We see that, compared to DIAGNOSIS, CoprGuard performs better in term of edge distortion.