DRCT: Diffusion Reconstruction Contrastive Training towards Universal Detection of Diffusion Generated Images

Baoying Chen [*1] **Jishen Zeng** [*1] **Jianquan Yang** [2] **Rui Yang** [1]

Abstract

Diffusion models have made significant strides in visual content generation but also raised increasing demands on generated image detection. Existing detection methods have achieved considerable progress, but they usually suffer a significant decline in accuracy when detecting images generated by an unseen diffusion model. In this paper, we seek to address the generalizability of generated image detectors from the perspective of hard sample classification. The basic idea is that if a classifier can distinguish generated images that closely resemble real ones, then it can also effectively detect less similar samples, potentially even those produced by a different diffusion model. Based on this idea, we propose Diffusion Reconstruction Contrastive Learning (DRCT), a universal framework to enhance the generalizability of the existing detectors. DRCT generates hard samples by high-quality diffusion reconstruction and adopts contrastive training to guide the learning of diffusion artifacts. In addition, we have built a million-scale dataset, DRCT-2M, including 16 types diffusion models for the evaluation of generalizability of detection methods. Extensive experimental results show that detectors enhanced with DRCT achieve over a 10% accuracy improvement in cross-set tests. The code, models, and dataset will soon be available at https://github.com/beibuwandeluori/DRCT.

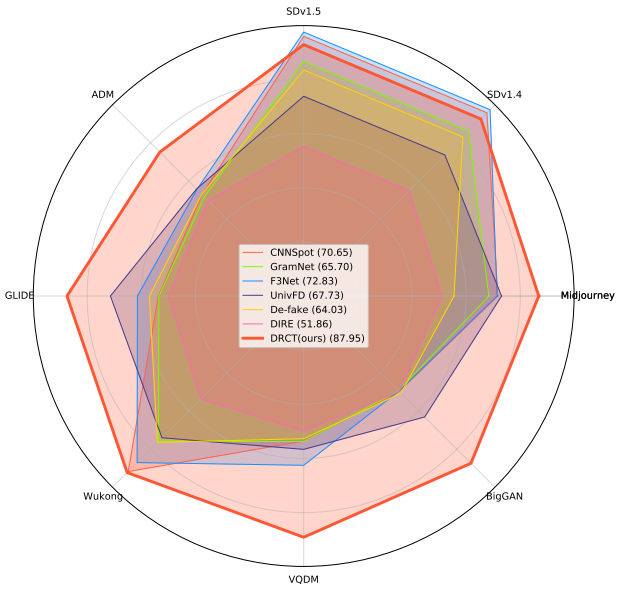


Figure 1. Generalization performance comparison between our proposed DRCT framework and six existing detection methods, including CNNSpot (Wang et al., 2019), F3Net (Qian et al., 2020), UnivFD (Ojha et al., 2023), GramNet (Liu et al., 2020), Defake (Sha et al., 2023), DIRE (Wang et al., 2023). All detection methods are trained on the same dataset consisting of generated images by Stable Diffusion (SD) v1.4 and real images from the MSCOCO dataset. The reported detection accuracies (ACC, %) are evaluated on each of the eight subsets of the GenImage (Zhu et al., 2023) dataset, and the averaged accuracy of each detection method is also respectively reported in brackets at the legends for convenient comparision. It shows that the proposed DRCT framework outperforms the compared methods by a remarkable margin of 15% or more.

1. Intorduction

In recent years, image generation technologies based on denoising diffusion models (Ho et al., 2020; Song et al., 2020; Nichol & Dhariwal, 2021; Rombach et al., 2021; Saharia et al., 2022) have rapidly advanced, with new generative models continually emerging. These technologies have provided efficient content editing and generation tools for applications such as digital creation, commercial advertising, news publishing, and social entertainment. However, there is also the risk of malicious misuse, such as fabricating fake news, misleading public opinion, interfering with political

[*]Equal contribution [1]Alibaba Group [2]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China. Correspondence to: Jianquan Yang <yangjq65@mail.sysu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

elections, and infringing on copyright. Therefore, there is an urgent need to develop technologies for detecting generated images to maintain a trustworthy cyberspace environment.

AI model-sharing platforms such as Hugging Face and CIVITAI offer a range of sophisticated diffusion-based image generative models and their variants, which facilitates users to generate diverse image content through simple textual interactions. The wide diversity and available of image generative models have raised a considerable challenge to the generalizability of detection methods. It demands that generated image detectors should be able to identify images produced by not only the known generative models but also the newly developed models that have not been involved in the training of detectors. In the remainder of this paper, we define real images as Real and the corresponding artificial intelligence-generated images as Fake.

In this work, we address the generalizability of generated image detection from the perspective of hard sample classification. The core idea is that if a classifier can distinguish hard-to-detect generated images from real images, it is also likely to generalize well in identifying easier samples. This inspires us to let the classifier focus on learning from hard samples to achieve better generalizability. In the task of generated image detection, if a large number of hard samples could be constructed for training the detectors, incorporating reasonable guidance, the generalizability of the detectors is expected to improve to some extent.

Based on the aforementioned idea, we have developed a universal training framework named Diffusion Reconstruction Contrastive Training (DRCT), aimed at enhancing the generalizability of generated image detectors. DRCT creates hard samples by reconstructing real images, which can produce high-quality near-real images that have almost the same appearance as the real images but contain subtle and imperceptible traces left by the generation model. Training existing detection methods with these hard samples is expected to improve their generalizability, enabling them to effectively detect the traces left by those image generation models not covered by the training set. Figure 1 shows the detection accuracies of six existing detection methods and our proposed DRCT framework (using UnivFD as the backbone detector) across all subsets of the GenImage dataset (Zhu et al., 2023). All seven detectors, including DRCT, were trained on images generated by Stable Diffusion(SD) v1.4 and then tested across all eight subsets of the GenImage dataset. As seen, DRCT outperforms all the compared methods, achieving an averaged detection accuracy of 87.95% over all tested subsets. DRCT achieves a 15% improvement over the second-highest method F3Net. Note that DRCT uses UnivFD as the backbone detector. As seen, the UnivFD detector equipped with DRCT increases its detection accuracy from 67.73% to 87.95%, indicating DRCT's effectiveness in enhancing the generalizability of the used backbone detector.

To further demonstrate how hard samples can enhance generalizability, we extract the features of real images, real reconstructed images, generated images using Stable Diffusion v1.4, and the generated-then-reconstructed images, from the last feature layer before the classification layers in a pre-trained detector Convnext-base (Liu et al., 2022c) (Conv-B). These four sets of features are then projected into a 2D space using t-SNE, as shown in Figure 2 (a). It can be seen that the feature points of real images are distributed near those of real reconstructed images, indicating their relative difficulty in distinguishing between the two, so that the real reconstructed ones can serve as hard samples. By fine-tuning the pre-trained detector with real reconstructed samples and generated-then-reconstructed samples, the detector learns better discriminability in differentiating real images from hard samples (real reconstructed samples). Consequently, the detector is likely to be able to identify generated samples and their reconstructed counterparts, as shown in Figure 2 (b).

Our main contributions are fourfold:

1. We discover that diffusion reconstruction on real images can well preserve the image visual content while leaving the intrinsic fingerprint of the diffusion model on the resulting images. These reconstructed images can serve as informative yet hard samples for detectors to learn the subtle differences between real and generated images, offering an effective approach to enhance the generalization ability of the detectors.
2. We propose a novel training framework named Diffusion Reconstruction Contrastive Training (DRCT) based on the aforementioned observation and contrastive learning. The proposed framework can significantly improve the detection accuracy and generalization ability of diffusion-generated image detectors.
3. We create the DRCT-2M benchmark dataset, which comprises 2 million high-quality generated images of sizes from 256 to 1024, covering 16 typical stable diffusion models. The DRCT-2M dataset also encompasses 136k generated images named DRCT-2M-Wild, which were manually collected from 8 real-world generation platforms. The DRCT-2M dataset provides a comprehensive benchmark for performance evaluation and comparison in generated image detection tasks.
4. We conduct extensive experiments to validate and evaluate the effectiveness, robustness and generalizability of our proposed DRCT framework under various settings as well as real-world scenarios. The results show that the detectors equipped with DRCT achieve remarkable improvement in detection accuracy.

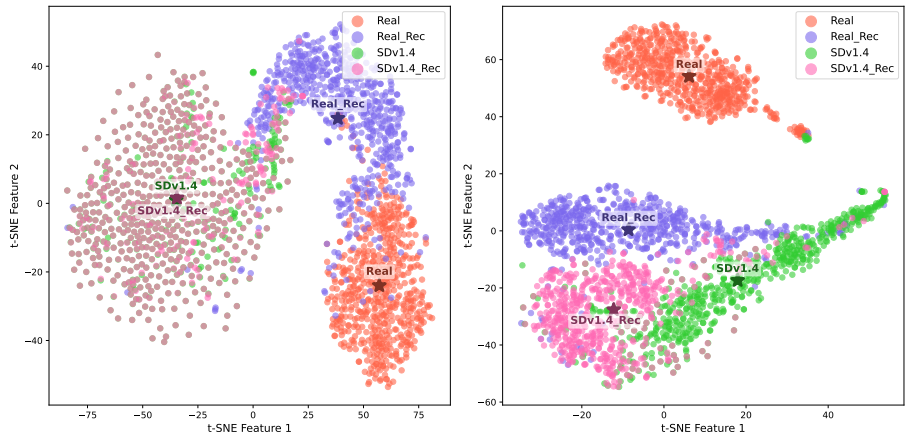


Figure 2. Visualization of the t-SNE embeddings with cluster centers (a) before using DRCT and (b) after using DRCT.

2. Related Work

2.1. Image Generation with Diffusion Models

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Zhu et al., 2017; Karras et al., 2017; 2018; Brock et al., 2018; Park et al., 2019) and Variational Autoencoders (VAEs) (Kingma & Welling, 2013; Sohn et al., 2015; Higgins et al., 2016; Hou et al., 2016; Zhao et al., 2017; van den Oord et al., 2017) have been the forerunners in the field of image generation. However, their limitations in controlling the generated image content paved the way for a new paradigm. Introduced by Ho et al. (Ho et al., 2020), Denoising Diffusion Probabilistic Models (DDPMs) have shown promise in generating high-quality images that rival those produced by GANs, thus marking a significant milestone in the evolution of diffusion model-based image generation techniques. Subsequent works have focused on enhancing the structure (Dhariwal & Nichol, 2021; Nichol et al., 2021; Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2021) and sampling efficiency (Nichol & Dhariwal, 2021; Song et al., 2020; Liu et al., 2022b; Lu et al., 2022) of diffusion models. The Latent Diffusion Model (LDM) proposed by Rombach et al. (Rombach et al., 2021), which underlies the popular open-source Stable Diffusion (SD) technology, has been a catalyst for further research. Notable extensions of SD include ControlNet (Zhang et al., 2023) for improved image generation control, SDXL (Podell et al., 2023) for high-resolution images, and LCM-RoLA (Luo et al., 2023) and SD-turbo (Sauer et al., 2023) for accelerated sampling.

2.2. Generated Image Detection

In the past few years, the detection of generated images has mainly focused on GAN-based images (Karras et al., 2017; Frank et al., 2020; Liu et al., 2020; Ju et al., 2022; Liu et al., 2022a; Tan et al., 2023). A variety of approaches have been proposed for this task. Wang et al. (Wang et al., 2019) demonstrated that a simple CNN classifier, trained on JPEG-compressed and blurred ProGAN (Karras et al., 2017) images, can generalize to other GAN-based generated images. However, Corvi et al. (Corvi et al., 2022) found that classifiers trained solely on GAN-based images face difficulties in generalizing to diffusion-based generated images. For the detection of diffusion-based generated images, Sha et al. (Sha et al., 2023) utilized a multimodal fusion technique with CLIP (Radford et al., 2021) as the backbone network, and found robustness in using BLIP-generated (Li et al., 2022) captions as input to a text model. Utkarsh et al. (Ojha et al., 2023) employed a large pre-trained CLIP model as a feature extractor with a nearest neighbor classifier, achieving promising generalization. Wu et al. (Wu et al., 2023a) adopted a CLIP-based text-image contrastive learning approach for their detector. Moving away from CLIP-based methods, Wang et al. (Wang et al., 2023) proposed the Diffusion Reconstruction Error (DIRE) method to detect diffusion-based generated images, leveraging the insight that real images cannot be accurately reconstructed by diffusion models. Similarly, Ma et al. (Ma et al., 2023) developed the Stepwise Error for Diffusion-based generated Image Detection (SeDID) method, utilizing intermediate step noising features for classification. Xi et al. (Xi et al., 2023) introduced a dual-stream detection network enhanced with cross-attention, integrating handcrafted SRM features and RGB depth features. Zhong et al. (Zhong et al., 2023) focused on the contrast in interpixel correlation between rich and poor texture regions, leading to a dual-stream model combining SRM filters and CNN classifiers. For more granular detection, Guarnera et al. (Guarnera et al., 2023) and Guo et al. (Guo et al., 2023) proposed multilevel algorithms capable of distinguishing between different GAN-based and

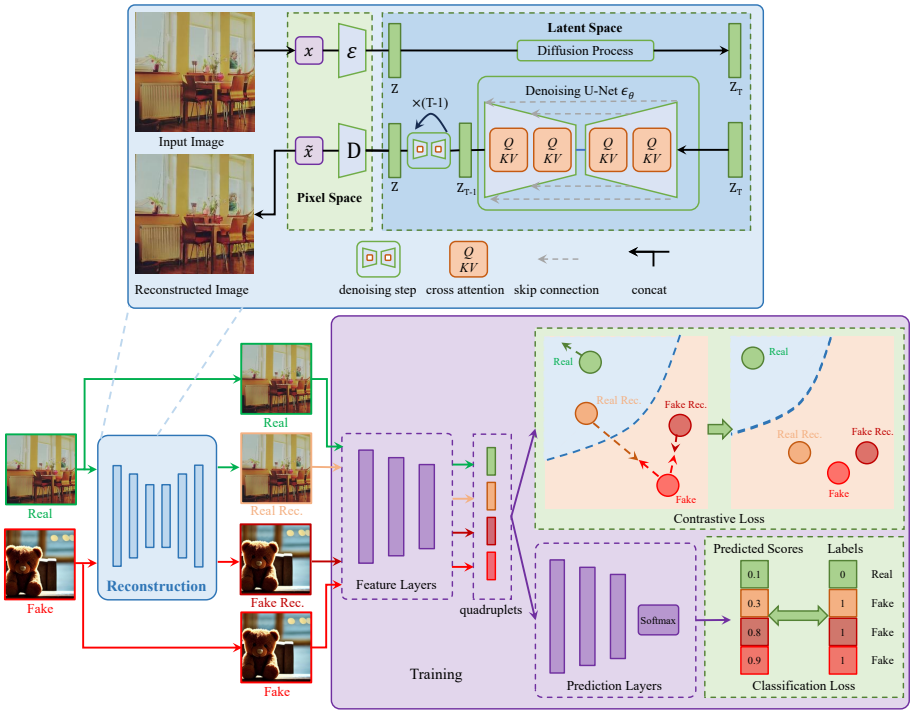


Figure 3. Workflow of the Diffusion Reconstruction Contrastive Training (DRCT) framework. The DRCT framework consists of two main stages: a reconstruction stage and a training stage. In the reconstruction stage, an input image undergoes a diffusion process and is then reconstructed through a denoising network. In the training stage, a binary classifier is trained on real images, generated images, and their reconstructed counterparts. Contrastive loss (Hadsell et al., 2006) is adopted to guide the discriminative feature learning, and classification loss is used for generated image detection.

diffusion-based generated images, as well as localized forgeries. Bird et al. (Bird & Lotfi, 2023) utilized a simple CNN classifier for binary classification, while Epstein et al. (Epstein et al., 2023) adopted an online learning approach, training incrementally on images generated by newly released technologies. They also highlighted the efficacy of the cutmix data augmentation technique for improving pixel-level segmentation performance in detecting Stable Diffusion inpainting images.

3. Diffusion Reconstruction Contrastive Training

As stated previously in the Introduction section, the idea of Diffusion Reconstruction Contrastive Training (DRCT) is to enhance the generalizability of detectors by training them with hard samples under appropriate guidance. This section is structured as follows. We first present the overall framework of DRCT, followed by a description of its technical details including diffusion reconstruction and contrastive training. Subsequently, the DRCT-2M dataset, comprising two million high-quality samples, is introduced for evaluating the generalizability of generated image detection.

3.1. The DRCT Framework

Figure 3 presents the framework of DRCT. The DRCT framework consists of two main stages: a reconstruction stage and a training stage. In the reconstruction stage, a large number of image samples are produced by reconstructing both real images and generated image using one or more selected diffusion-based generative models, which are then prepared for the training of the classifier. In the training stage, four types of samples, including real images, real reconstructed images, fake images(namely, generated images), and fake reconstructed images, are utilized for computing the contrastive loss and the classification loss. The two loss functions guide the classifier to learn a better feature representation and to identify real images as real and the other three types of images as fake.

3.2. Diffusion Reconstruction

The reconstruction within the Stable Diffusion framework relies on a conditional diffusion process that iteratively denoises an image. This process is articulated in the Stable Diffusion (SD) model (Rombach et al., 2021), which leverages a latent variable to modulate the expressiveness of the diffusion process. The forward diffusion process is charac-

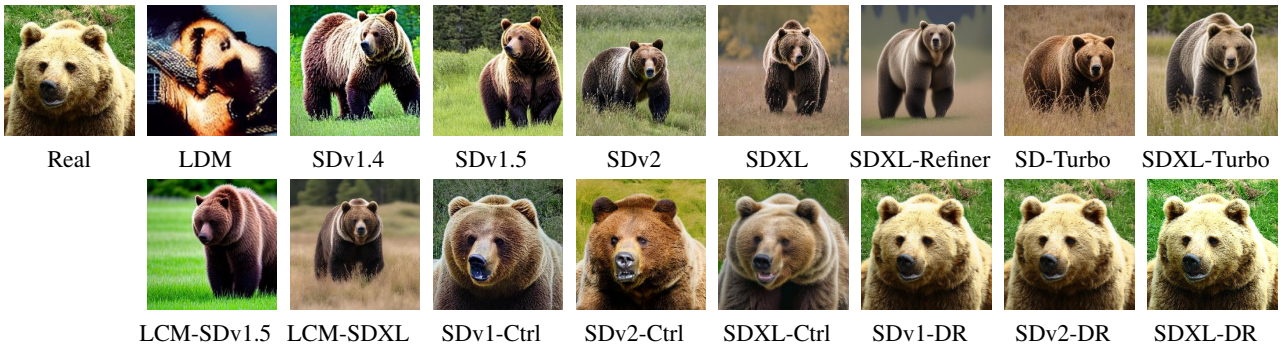


Figure 4. Samples from our proposed **DRCT-2M** dataset for demonstration. The real image is "000000000285.jpg" from the MSCOCO (Lin et al., 2014) dataset. The text prompt used for image generation is "A big burly grizzly bear is shown with grass in the background." The **DRCT-2M** dataset involves 16 types of stable diffusion models, including LDM, SDv1.4, SDv1.5, SDv2, SDXL, SDXL-refiner, SD-Turbo, SDXL-Turbo, LCM-SDv1.5, LCM-SDXL, SDv1-Ctrl, SDv2-Ctrl, SDXL-Ctrl, SDv1-DR, SDv2-DR and SDXL-DR, where "Ctrl" means "ControlNet" and "DR" means "Diffusion Reconstruction". Specifically, we utilize SDv1-DR, SDv2-DR and SDXL-DR models to reconstruct the real images from MSCOCO (Lin et al., 2014). It can be seen that the images generated using the SDv2 and SDXL series models likely have richer details and better quality than those produced by the SDv1 series models.

terized by incrementally adding noise to an image, and the denoising process reverses this by iteratively reducing noise to recover the original image.

The forward diffusion process can be expressed as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$, for $t = 0, \ldots, T$. Here, x_0 denotes the initial data, x_t denotes the noisy data after t diffusion steps, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

The reverse process of DDIM (Song et al., 2020) is deterministic and can be represented by:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \frac{x_t - \sqrt{1 - \alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} + \sqrt{1 - \alpha_{t-1}}\epsilon_t, \tag{2}$$

where $\alpha_{t-1} = \frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}$, for $t = T, \ldots, 1$, $\epsilon_\theta(x_t, t)$ is the predicted noise by the denoising neural network parameterized by θ, and $\epsilon_t \sim \mathcal{N}(0, I)$ is Gaussian noise.

Our proposed method utilizes stable diffusion models for reconstructing the image, differing from DIRE (Wang et al., 2023), which directly adds noise to the input image followed by denoising with DDIM to obtain the reconstructed image. Initially, the encoder of VAE is used to encode the input image x into a latent space representation x_0. Subsequently, x_0 undergoes a process of noise addition followed by a denoising process using DDIM, resulting in the reconstructed latent representation x_0'. Finally, the decoder of VAE decodes x_0' to obtain the reconstructed image x'.

3.3. Contrastive Training

We employ a margin-based contrastive loss (Hadsell et al., 2006) within our framework that brings positive pairs closer while separating negative pairs by a margin. This approach simplifies the loss computation and is stated mathematically as:

$$\begin{aligned}\mathcal{L}_{\text{contrastive}} = \frac{1}{N}\sum_{i}^{N}[Y \cdot D_w(i)^2 \\ +(1-Y)\cdot max(0, m - D_w(i))^2]\end{aligned} \tag{3}$$

where N is the total number of sample pairs, Y is the binary label for each sample pair (if two samples share the same classification label i.e., real or fake, then $Y = 1$, otherwise $Y = 0$). $D_w(i)$ is the Euclidean distance between the samples in each pair. $m > 0$ is the margin for negative sample pairs, and the default value for m is 1.0 in our experiments.

The overall objective to be minimized is a combination of the contrastive loss and the binary classification cross-entropy loss, weighted by a balancing parameter λ:

$$\mathcal{L}_{\text{total}} = \lambda\mathcal{L}_{\text{contrastive}} + (1-\lambda)\mathcal{L}_{\text{cross-entropy}} \tag{4}$$

with the cross-entropy loss defined as:

$$\mathcal{L}_{\text{cross-entropy}} = -\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right], \tag{5}$$

where y_i is the true label of the i-th sample and p_i is the predicted probability of the i-th sample belonging to the positive class. The parameter $\lambda \in [0, 1)$ regulates the trade-off between the losses, and the default value for λ is 0.3 in our experiments.

3.4. DRCT-2M Dataset

We constructed the **DRCT-2M** dataset, a comprehensive collection of two million images for training and evaluating

Table 1. Accuracy (ACC, %) comparisons of our DRCT and other generated image detectors on DRCT-2M. Except for DIRE and DRCT, all methods are only trained on SDv1.4 and then evaluated on different testing subsets on DRCT-2M. For the training data of DIRE and DRCT, when the Diffusion Reconstructed (DR) model is SDv1, the original fake images were generated by SDv1.4. When the DR model is SDv2, the original fake images were generated by SDv2.

Method	DR	SD Variants						Turbo Variants		LCM Variants		ControlNet Variants			DR Variants			Avg.
		LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	SDv1-DR	SDv2-DR	SDXL-DR	
CNNSpot	-	99.87	99.91	99.90	97.55	66.25	86.55	86.15	72.42	98.26	61.72	97.96	85.89	82.84	60.93	51.41	50.28	81.12
F3Net	-	99.85	99.78	99.79	88.66	55.85	87.37	68.29	63.66	97.39	54.98	97.98	72.39	81.99	65.42	50.39	50.27	77.13
CLIP/RN50	-	99.00	99.99	99.96	94.61	62.08	91.43	83.57	64.40	98.97	57.43	99.74	80.69	82.03	65.83	50.67	50.47	80.05
GramNet	-	99.40	99.01	98.84	95.30	62.63	80.68	71.19	69.32	93.05	57.02	89.97	75.55	82.68	51.23	50.01	50.08	76.62
De-fake	-	92.1	99.53	99.51	89.65	64.02	69.24	92.00	93.93	99.13	70.89	58.98	62.34	66.66	50.12	50.16	50.00	75.52
Conv-B	-	**99.97**	**100.0**	**99.97**	95.84	64.44	82.00	80.82	60.75	99.27	62.33	99.80	83.40	73.28	61.65	51.79	50.41	79.11
UnivFD	-	98.30	96.22	96.33	93.83	91.01	93.91	86.38	85.92	90.44	88.99	90.41	81.06	89.06	51.96	51.03	50.46	83.46
DIRE	SDv1	98.19	99.94	99.96	68.16	53.84	71.93	58.87	54.35	**99.78**	59.73	99.65	64.20	59.13	51.99	50.04	49.97	71.23
DIRE	SDv2	54.62	75.89	76.04	**99.87**	59.90	93.08	**99.77**	57.55	87.29	72.53	67.85	99.69	64.40	49.96	52.48	49.92	72.55
DRCT/Conv-B (ours)	SDv1	99.91	99.90	99.90	96.32	83.87	85.63	91.88	70.04	99.66	78.76	**99.90**	95.01	81.21	**99.90**	95.40	75.39	90.79
DRCT/Conv-B (ours)	SDv2	99.66	98.56	98.48	99.85	96.10	**98.68**	99.59	83.30	98.45	93.78	96.68	**99.85**	**97.66**	93.91	**99.87**	**90.39**	**96.55**
DRCT/UnivFD (ours)	SDv1	96.74	96.26	96.33	94.89	**96.24**	93.46	93.43	92.94	91.17	95.01	95.60	92.68	91.95	94.10	69.55	57.43	90.49
DRCT/UnivFD (ours)	SDv2	94.45	94.35	94.24	95.05	95.61	95.38	94.81	**94.48**	91.66	**95.54**	93.86	93.48	93.54	84.34	83.20	67.61	91.35

our proposed DRCT framework and existing methods of generated image detection. The DRCT-2M dataset mainly consists of two parts: images automatically generated by various diffusion-based generative models and those collected from the real-world scenarios.

DRCT-2M The first part dataset DRCT-2M consists of two kinds of generated images. The images of first kind were generated by using a text-to-image process, whose input prompts are derived from the MSCOCO (Lin et al., 2014) dataset. A total of 10 types of the currently available SD models have been involved. The images of second kind were generated by using an image-to-image process, which included 3 types ControlNet (Zhang et al., 2023) for creating controllable images and 3 types diffusion reconstruction models for generating reconstruction images. The input conditions of ControlNet are text prompts and Canny Edge Map (called "canny") extracted by the Canny edge detection algorithm (Canny, 1986), and the input conditions of diffusion reconstruction models are text prompts, masks and real images. Therefore, the first part dataset DRCT-2M includes 16 types of generated images, with 120k images for each type. The quality of the generated images from various diffusion models is exemplified in Figure 4. Moreover, we have explored the quality of generated images, with the specific comparison results presented in Figure 8. More details of the DRCT-2M dataset are illustrated in Table 11.

DRCT-2M-Wild The second part dataset DRCT-2M-Wild consists of images collected from real-world scenarios to evaluate the model's generalizability and robustness in various practical applications. Except for Midjourney collected from DISCORD, all other images were collected from the open-source diffusion models on the CIVITAI website, with the specific generation models and image quantities detailed in Table 12. These samples reflect real-world distortions, facilitating performance evaluation and optimization in practical scenarios.

The proposed DRCT-2M dataset aims to provide a comprehensive set of samples and scenarios that support the development of effective algorithms capable of distinguishing between real and generated images. It will soon be publicly available for research purposes.

4. Experiments

We first present the experimental setup, and then report and discuss the results of our proposed DRCT framework and the compared methods. More results are presented in the Appendix Section.

4.1. Experimental Setup

Data Preprocessing and Augmentation: The compared methods involved in most comparative experiments are trained on the DRCT-2M dataset (utilizing real images from MSCOCO (Lin et al., 2014)) or the GenImage (Zhu et al., 2023) dataset, as will be specified. Two pretrained stable diffusion models, namely SDv1 and SDv2, are utilized to generate reconstructed images. During training, all detectors take input images of size 224×224, and during testing, images larger than 224×224 will be center-cropped. To achieve better robustness against post-processing, a range of data augmentations are conducted during training, including horizontal flipping, Gaussian noise disturbance, Gaussian blurring, random rotation, JPEG compression with random quality, brightness and contrast adjustments, and grid dropout.

Evaluation Metrics and Comparative Approaches: We adopt Accuracy (ACC) as the metric to evaluate detection performance, using a threshold of 0.5 for calculating ACC. The compared methods include representative binary classifiers retrained for generated image detection and publicly available detectors originally proposed for detecting generated images. That is, Conv-B (Liu et al., 2022c),

Table 2. Accuracy (ACC, %) comparisons of our DRCT and other generated image detectors. All methods were trained on GenImage/SDv1.4 and evaluated on different testing subsets. The Diffusion Reconstruction Model of DIRE and DRCT is SDv1.

Method	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
CNNSpot	84.92	99.88	99.76	53.48	53.80	99.68	55.50	49.93	74.62
F3Net	77.85	98.99	99.08	51.20	54.87	97.92	58.99	49.21	73.51
CLIP/RN50	83.30	99.97	99.89	54.55	57.37	99.52	57.90	50.00	75.31
GramNet	73.68	98.85	98.79	51.52	55.38	95.38	55.15	49.41	72.27
De-fake	79.88	98.65	98.62	71.57	78.05	98.42	78.31	74.37	84.73
Conv-B	83.55	**99.99**	**99.92**	51.75	56.27	99.92	58.41	50.00	74.98
UnivFD	91.46	96.41	96.14	58.07	73.40	94.53	67.83	57.72	79.45
DIRE	50.40	**99.99**	**99.92**	52.32	67.23	**99.98**	50.10	49.99	71.24
DRCT/Conv-B (ours)	**94.63**	99.88	99.82	61.78	65.92	99.91	74.88	58.81	82.08
DRCT/UnivFD (ours)	91.50	95.01	94.41	**79.42**	**89.18**	94.67	**90.03**	**81.67**	**89.49**

Table 3. Accuracy (ACC, %) comparisons of our DRCT and other generated image detectors. Except for DIRE and DRCT, all methods were trained on DRCT-2M/SDv1.4 and evaluated on different testing subsets of DRCT-2M-Wild. For the training data of DIRE and DRCT, when the Diffusion Reconstructed (DR) model is SDv1, the original fake images were generated by SDv1.4. When the DR model is SDv2, the original fake images were generated by SDv2.

Method	RD	DreamShaper XL10	SD-XL Niji Special Edition	Realistic Vision v5.1	Deep Negative v1.x	Detail Tweaker Lora	MajicMix Realistic	rMada Merge	Midjourney	Avg.
CNNSpot	-	61.98	76.04	67.00	63.41	64.20	57.05	73.86	07.54	58.89
F3Net	-	37.65	45.20	51.26	44.21	41.28	40.41	48.37	11.11	39.94
CLIP/RN50	-	53.58	73.03	50.95	66.03	63.34	55.78	65.28	02.97	53.87
GramNet	-	34.13	44.67	46.29	42.52	38.47	41.26	49.28	08.14	38.10
De-fake	-	16.16	03.40	32.83	03.56	05.01	06.55	17.95	02.13	10.95
Conv-B	-	49.67	68.85	42.34	47.17	43.91	31.96	65.80	03.45	44.14
UnivFD	-	73.08	84.34	65.19	61.87	62.52	59.51	84.53	17.33	63.55
DIRE	SDv1	23.59	40.08	31.86	49.15	39.17	38.81	53.71	04.01	35.05
DIRE	SDv2	34.77	50.04	63.21	79.09	67.77	73.03	81.92	01.60	56.43
DRCT/Conv-B(ours)	SDv1	78.43	85.34	82.47	86.93	86.54	81.15	87.13	68.51	82.06
DRCT/Conv-B(ours)	SDv2	94.07	93.39	**96.56**	93.88	93.93	95.65	97.01	80.76	93.16
DRCT/UnivFD(ours)	SDv1	90.89	96.10	86.50	93.59	93.41	91.74	93.24	52.39	87.23
DRCT/UnivFD(ours)	SDv2	**98.34**	**97.81**	95.15	**98.26**	**97.71**	**98.26**	**98.83**	**90.80**	**96.90**

CLIP/RN50 (Radford et al., 2021) (pretrained model RN50 with only the image modality), CNNSpot (Wang et al., 2019) (Resnet50 as backbone), F3Net (Qian et al., 2020) (input size is 299 × 299), UnivFD (Ojha et al., 2023) (ViT-L/14 as backbone, freezing the backbone and training only the final fully connected layer), GramNet (Liu et al., 2020), De-fake (Sha et al., 2023) (BLIP technology for extracting image descriptions from the textual modality), and DIRE (Wang et al., 2023) (Conv-B as backbone). Note that all the comparisons were replicated in our training protocol.

4.2. Comparisons of Detection Accuracies

Performance comparisons are conducted on the DRCT-2M dataset and the GenImage dataset, respectively.

Comprasions on DRCT-2M When the training and the testing datasets are aligned, existing detection methods such as DIRE (Wang et al., 2023), De-fake (Sha et al., 2023), and UnivFD (Ojha et al., 2023) have reported near-perfect detection accuracies. However, in real-world scenarios, it is crucial to achieve good accuracies on images generated by unseen diffusion models. In this context, we compare the performance of existing detectors including CNNSpot, F3Net, UnivFD, GramNet, De-fake, DIRE, Conv-B, DRCT/Conv-B (means Conv-B enhanced with DRCT), and DRCT/UnivFD (means UnivFD enhanced with DRCT). Except for DIRE and DRCT, all other solutions were trained on a subset of DRCT-2M, that is, SDv1.4 (for fake images) and MSCOCO (for real images). For fair comparison, when training DIRE and DRCT, if the reconstruction model is SDv1, we select fake images from the subset SDv1.4 of DRCT-2M and then perform reconstruction. Similarly, if the reconstruction model if SDv2, we select fake images from the subset SDv2 of DRCT-2M.

Comparisons of detection accuracies on the DRCT-2M dataset are reported in Table 1. Most methods exhibit extremely high ACCs on images generated by diffusion models related to SDv1.4, such as LDM, SDv1.5, LCM-SDv1.5, and SD-Ctrl. However, these approaches suffer a significant decline in ACC when detecting unseen and substantially altered diffusion models like SDv2, SDXL, SDXL-Refiner, SDXL-Turbo, LCM-SDXL, and SDXL-Ctrl. Particularly,

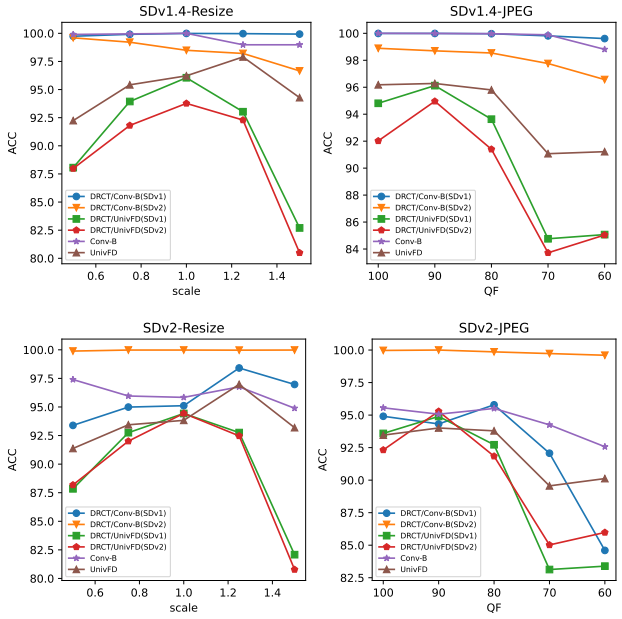


Figure 5. Robustness evaluations against resizing and JPEG compression on the two subsets SDv1.4 and SDv2 of DRCT-2M.

for images generated by SDXL, the ACCs of existing methods except UnivFD, range only between 53%-67%. The ACCs of existing methods for detecting real reconstructed images of SDv1-DR, SDv2-DR, and SDXL-DR drop to 50%-65%.

In contrast, our proposed DRCT framework achieves superior ACCs across all types of SD-generated images. We also compare the detection ACCs when different diffusion model (namely, SDv1 or SDv2) is used for reconstruction. As shown, merely using SDv1 for reconstruction in DRCT, the average detection ACC has already improved from 79.11% to 90.79% compared to the baseline detector Conv-B. For UnivFD, the average ACC is improved from 83.46% to 90.49%. When using SDv2 for reconstruction, the average detection ACC can be further improved to 96.55% compared to the baseline detector Conv-B. This indicates that a better reconstruction model helps to achieve better detection performance on the DRCT-2M dataset.

Comparisons on GenImage To further validate the effectiveness of the DRCT framework, we also conduct comparisons following the same experimental protocol as GenImage. All detection methods are trained on the SDv1.4 subset of GenImage. Specifically, for the training of DIRE and DRCT, the reconstruction model is also SDv1. The comparative results in Table 2 reveal that all compared methods achieve very high detection accuracies on the SDv1.4, SDv1.5, and Wukong subsets. However, a noticeable decline in ACCs can be observed across the other subsets such as Midjourney, ADM, GLIDE, VQDM, and BigGAN, especially on non-diffusion-based generated methods BigGAN. After incorporating our proposed DRCT framework, the backbone detectors Conv-B and UnivFD achieve an average ACC improvement of 7.1% and 10.04% respectively. This validates the stable effectiveness of the DRCT framework in enhancing the generalizability of the involved detectors.

4.3. Comparisons of Generalizability

To validate the generalizability of our proposed DRCT framework, we conduct cross-database evaluation on DRCT-2M-Wild and GenImage using detectors trained on DRCT-2M, as described in Section 4.2. DRCT-2M-Wild is an internet-collected dataset reflecting real-world scenarios.

The cross-dataset evaluation results on DRCT-2M-Wild are presented in Table 3. Except for DRCT, the average detection ACCs of all other compared methods are below 65%. In contrast, the Conv-B and UnivFD detectors enhanced with DRCT achieve average ACCs of over 80% on DRCT-2M-Wild, which is a significant improvement over their original versions. Moreover, DRCT/UnivFD trained on SDv2 reaches the highest average ACC of 96.90%. Similarly, the DRCT-enhanced methods also demonstrate superior generalizability on GenImage, as shown in Table 7.

4.4. Robustness against Post-Processing

To evaluate the resilience of the DRCT framework against post-processing, we adopt the experimental setup from (Wu et al., 2023a) to perform resizing (with scales of 0.5, 0.75, 1.0, 1.25, 1.5) and JPEG compression (with quality factors of 60, 70, 80, 90, 100) on the tested images, which include both real and generated images. We employ six detectors including Conv-B, UnivFD, DRCT/Conv-B (SDv1), DRCT/Conv-B (SDv2), DRCT/UnivFD (SDv1), and DRCT/UnivFD (SDv2) to validate the robustness. All methods were trained under the same framework with identical data augmentation. As shown in Figure 5, the Conv-B detectors enhanced with DRCT exhibit superior robustness, maintaining detection ACCs of up to 99% for resizing and JPEG compression post-processing. Moreover, compared to UnivFD enhanced with DRCT, Conv-B enhanced with DRCT exhibits better post-processing robustness, mainly due to the fact that Conv-B tunes all its network weights while UnivFD only funes its final fully connected layer.

4.5. Ablation Studies

In the ablation studies, we focus on the impact of the following two factors on detection performance: 1) the effect of whether reconstructing both real and fake images from the training set on the detector; 2) the effect of using Contrastive Loss on the detector's effectiveness.

Table 4. Ablation study results demonstrating the effect of our DRCT (Conv-B as the backbone). All methods were trained on DRCT-2M/SDv1.4 and tested on different testing subsets of GenImage. We report ACC (%) in the Table.

w/o Fake Image	w/o Real Rec.	w/o Fake Rec.	w/o Contrastive Loss	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
✓	×	×	×	67.79	95.64	95.85	51.44	52.01	86.13	52.97	49.99	68.98
×	✓	×	×	74.25	99.77	99.61	51.73	51.54	99.61	61.12	50.60	73.53
✓	✓	×	×	86.88	99.89	**99.81**	52.51	53.70	**99.83**	64.13	51.50	76.03
✓	✓	✓	×	86.62	**99.90**	99.80	55.49	57.67	**99.83**	64.08	52.49	76.99
✓	✓	✓	✓	**94.43**	99.37	99.19	**66.42**	**73.31**	99.25	**76.85**	**59.41**	**83.53**

To ensure the fairness of our ablation experiments, we used Conv-B as the baseline classifier, training on a subset of the DRCT-2M/SDv1.4 dataset (with real images from MSCOCO and the reconstruction model being SDv1) and then testing on the GenImage dataset. The results of the ablation studies are shown in Table 4. The initial baseline detector achieved an average ACC of 68.98% on GenImage. When the original SDv1.4 fake images were replaced with reconstructed real images (Real Rec.), the average ACC significantly increased by 6.55%, indicating that training the model with both real images and reconstructed real images aids in guiding the detector to learn common distortion features of AI-generated images, while mitigating overfitting to semantic features. Upon adding back the original SDv1.4 fake images, the average ACC increased by another 2.5%. Furthermore, including reconstructed fake images led to an additional 0.96% increase in average ACC. Lastly, when we added Contrastive Loss to the original BCE loss function during training, the average ACC saw a significant increase of 6.54%, reaching an overall accuracy of 83.53%. These ablation studies further demonstrate the effectiveness of our proposed DRCT training framework.

5. Discussions

Limitations The proposed DRCT framework has remarkably enhanced the detection performance for diffusion-based generated images. While it also improves the detection accuracy for non-diffusion-based images, such as those generated by GANs, the improvement is less marked. This discrepancy mainly stems from the significant differences in the image generation processes of GAN-based and diffusion-based methods, which exhibit distinct generative artifacts.

Moreover, our evaluation of DRCT has so far only covered the detection of globally generated images. Detecting locally generated regions in images presents a more challenging task, especially when the area of the locally generated region is small. Investigating the performance of DRCT in detecting locally generated images and how to further enhance it will be the focus of our future work.

Conclusions We have proposed a universal framework, Diffusion Reconstruction Contrastive Training, for enhancing the generalizability of existing methods for detecting diffusion-based generated images. With DRCT, the backbone detector can achieve remarkable improvement, indicating the effectiveness of the DRCT framework. In addition, a large-scale, high-quality image dataset, DRCT-2M, has been built for the training of detectors and the evaluation of effectiveness, generalizability, robustness, etc. Future work includes improving DRCT to keep pace with the development of diffusion models. Exploring the interpretability of the features learned by DRCT-enhanced detectors may provide insights into the fundamental differences between real and generated images.

Acknowledgements

We express our gratitude to all anonymous reviewers for their constructive feedback. This work was funded in part by National Natural Science Foundation of China (NSFC) under Grant 62372489; in part by Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515030032); and in part by Shenzhen Science and Technology Program (Grant No. JCYJ20210324102204012, JCYJ20230807111207015).

Impact Statement

Our work aims to advance the field of Machine Learning by enhancing the detection of generated images through the proposed DRCT framework. This advancement has significant social implications. On the positive side, our method improves the ability to detect generated images, which is crucial for curbing misinformation, regulating the usage of generated content, and enhancing the credibility of digital media. However, there is a potential risk that attackers could exploit DRCT to create harder-to-detect fake images in an adversarial setting. To address this, we propose strategies such as access control and continuous model updates. Recognizing the dual-edged nature of this technology, we urge the research community to strive for minimizing negative impacts while capitalizing on its positive contributions.

References

Bird, J. J. and Lotfi, A. Cifake: Image classification and explainable identification of ai-generated synthetic images. *ArXiv*, abs/2303.14126, 2023. URL https://api.semanticscholar.org/

CorpusID:257757303.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *ArXiv*, abs/1809.11096, 2018. URL https://api.semanticscholar.org/CorpusID:52889459.

Canny, J. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., and Verdoliva, L. On the detection of synthetic images generated by diffusion models. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021. URL https://api.semanticscholar.org/CorpusID:234357997.

Epstein, D. C., Jain, I., Wang, O., and Zhang, R. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 382–392, 2023.

Frank, J. C., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., and Holz, T. Leveraging frequency analysis for deep fake image recognition. *ArXiv*, abs/2003.08685, 2020. URL https://api.semanticscholar.org/CorpusID:213175447.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63:139 – 144, 2014.

Guarnera, L., Giudice, O., and Battiato, S. Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models. *ArXiv*, abs/2303.00608, 2023. URL https://api.semanticscholar.org/CorpusID:257255351.

Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., and Liu, X. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3155–3165, 2023.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:1735–1742, 2006. URL https://api.semanticscholar.org/CorpusID:8281592.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016. URL https://api.semanticscholar.org/CorpusID:46798026.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. URL https://api.semanticscholar.org/CorpusID:219955663.

Hou, X., Shen, L., Sun, K., and Qiu, G. Deep feature consistent variational autoencoder. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1133–1141, 2016. URL https://api.semanticscholar.org/CorpusID:5257869.

Ju, Y., Jia, S., Ke, L., Xue, H., Nagano, K., and Lyu, S. Fusing global and local features for generalized ai-synthesized image detection. *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3465–3469, 2022. URL https://api.semanticscholar.org/CorpusID:247762136.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *ArXiv*, abs/1710.10196, 2017. URL https://api.semanticscholar.org/CorpusID:3568073.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2018. URL https://api.semanticscholar.org/CorpusID:54482423.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL https://api.semanticscholar.org/CorpusID:216078090.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, J., Li, D., Xiong, C., and Hoi, S. C. H. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:246411402.

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft

coco: Common objects in context. In *European Conference on Computer Vision*, 2014. URL https://api.semanticscholar.org/CorpusID:14113767.

Liu, B., Yang, F., Bi, X., Xiao, B., Li, W., and Gao, X. Detecting generated images by real images. In *European Conference on Computer Vision*, 2022a. URL https://api.semanticscholar.org/CorpusID:253121047.

Liu, L., Ren, Y., Lin, Z., and Zhao, Z. Pseudo numerical methods for diffusion models on manifolds. *ArXiv*, abs/2202.09778, 2022b. URL https://api.semanticscholar.org/CorpusID:247011732.

Liu, Z., Qi, X., Jia, J., and Torr, P. H. S. Global texture enhancement for fake face detection in the wild. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8057–8066, 2020. URL https://api.semanticscholar.org/CorpusID:211010785.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976, 2022c. URL https://api.semanticscholar.org/CorpusID:245837420.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *ArXiv*, abs/2206.00927, 2022. URL https://api.semanticscholar.org/CorpusID:249282317.

Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., and Zhao, H. Lcm-lora: A universal stable-diffusion acceleration module. *ArXiv*, abs/2311.05556, 2023. URL https://api.semanticscholar.org/CorpusID:265067414.

Ma, R., Duan, J., Kong, F., Shi, X., and Xu, K. Exposing the fake: Effective diffusion-generated images detection. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023. URL https://openreview.net/forum?id=7R62e4Wgim.

Nichol, A. and Dhariwal, P. Improved denoising diffusion probabilistic models. *ArXiv*, abs/2102.09672, 2021. URL https://api.semanticscholar.org/CorpusID:231979499.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:245335086.

Ojha, U., Li, Y., and Lee, Y. J. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2332–2341, 2019. URL https://api.semanticscholar.org/CorpusID:81981856.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Muller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023. URL https://api.semanticscholar.org/CorpusID:259341735.

Qian, Y., Yin, G., Sheng, L., Chen, Z., and Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. *ArXiv*, abs/2007.09355, 2020. URL https://api.semanticscholar.org/CorpusID:220647499.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. URL https://api.semanticscholar.org/CorpusID:248097655.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2021. URL https://api.semanticscholar.org/CorpusID:245335280.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487,

2022. URL https://api.semanticscholar.org/CorpusID:248986576.

Sauer, A., Lorenz, D., Blattmann, A., and Rombach, R. Adversarial diffusion distillation. *ArXiv*, abs/2311.17042, 2023. URL https://api.semanticscholar.org/CorpusID:265466173.

Sha, Z., Li, Z., Yu, N., and Zhang, Y. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3418–3432, 2023.

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. In *Neural Information Processing Systems*, 2015. URL https://api.semanticscholar.org/CorpusID:13936837.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. URL https://api.semanticscholar.org/CorpusID:222140788.

Tan, C., Zhao, Y., Wei, S., Gu, G., and Wei, Y. Learning on gradients: Generalized artifacts representation for gan-generated images detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12105–12114, 2023. URL https://api.semanticscholar.org/CorpusID:259226993.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. *ArXiv*, abs/1711.00937, 2017. URL https://api.semanticscholar.org/CorpusID:20282961.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., and Wolf, T. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. Cnn-generated images are surprisingly easy to spot... for now. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8692–8701, 2019.

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., and Li, H. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22445–22455, October 2023.

Wu, H., Zhou, J., and Zhang, S. Generalizable synthetic image detection via language-guided contrastive learning. *ArXiv*, abs/2305.13800, 2023a. URL https://api.semanticscholar.org/CorpusID:258841383.

Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., and Li, H. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *ArXiv*, abs/2306.09341, 2023b. URL https://api.semanticscholar.org/CorpusID:259171771.

Xi, Z., Huang, W., Wei, K., Luo, W., and Zheng, P. Ai-generated image detection using a cross-attention enhanced dual-stream network. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1463–1470. IEEE, 2023.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *ArXiv*, abs/2304.05977, 2023. URL https://api.semanticscholar.org/CorpusID:258079316.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. *ArXiv*, abs/2302.05543, 2023. URL https://api.semanticscholar.org/CorpusID:256827727.

Zhao, S., Song, J., and Ermon, S. Infovae: Information maximizing variational autoencoders. *ArXiv*, abs/1706.02262, 2017. URL https://api.semanticscholar.org/CorpusID:34051459.

Zhong, N., Xu, Y., Qian, Z., and Zhang, X. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *ArXiv*, abs/2311.12397, 2023. URL https://api.semanticscholar.org/CorpusID:265309137.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017. URL https://api.semanticscholar.org/CorpusID:206770979.

Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., and Wang, Y. Genimage: A million-scale benchmark for detecting ai-generated image. *ArXiv*, abs/2306.08571, 2023. URL https://api.semanticscholar.org/CorpusID:259164965.

A. More Analysis of the Proposed DRCT Framework

A.1. More Evaluation Metrics on DRCT-2M

Beyond the comparison results of ACC in Table 1, we have supplemented the evaluation metrics of F1 Score and False Negative Rate (FNR) in Table 5 and Table 6 respectively. Similarly, our proposed method DRCT consistently exhibits superior performance in the evaluations of both F1 and FNR.

A.2. More Comparison Results on GenImage

Cross-Dataset Evaluation To evaluate the generalizability of our proposed method DRCT, all detectors were trained on DRCT-2M/SDv1.4 and subsequently tested on different subsets of GenImage. As shown in Table 7, our proposed method DRCT/UnivFD achieves the highest average ACC score, reaching 87.67%, which is 14.84% higher than the best non-DRCT method F3Net.

A.3. In-Depth Analysis of Reconstructed Image Detectability

tSNE Visualization As shown in Figure 2, we utilized tSNE to intuitively showcase the distinction between real images and their reconstructed counterparts (termed "Real_Rec"), by visualizing the features derived from the classifier backbone network. By reducing the dimensionality to a two-dimensional space, we observed the scatter distribution of the samples, as depicted in Figure 2(a). It is evident that the Real_Rec samples cluster closely to the real samples, whereas the generated samples (via SDv1.4) are noticeably distant from the real samples. This suggests that differentiating between the real and Real_Rec samples poses a greater challenge than distinguishing between the real and generated samples. Consequently, integrating Real_Rec images into the training process can encourage the classifier to establish tighter classification boundaries, thus enhancing the classifier's generalization capabilities.

Spectral Distribution Inspired by UnivFD (Ojha et al., 2023), diffusion models reveal some kinds of distinct spectral distribution patterns. As done in UnivFD, we evaluated the average Fourier amplitude spectra for real images, real reconstructed images (labeled "Real_Rec"), generated images (labeled "SDv1.4" and "SDv2") and generated reconstructed images (labeled "SDv1.4_Rec"), by averaging over 5000 samples for each category, as shown in Figure 6. Notably, SDv1.4, SDv1.4_Rec, and SDv2 exhibit a similar and distinctive pattern, significantly different from the amplitude spectrum of real images. Simultaneously, the amplitude spectrum of real reconstructed images shows a greater similarity to that of real images, indicating that distinguishing between the two is more challenging than differentiating between real and generated images to a certain extent.

A.4. The Effect of Reconstruction Step

Table 8 show the performance of DRCT/Conv-B when the reconstruction step employed during the training phase does not match the reconstruction step of the generated image in the testing phase. The results indicate that the reconstruction step exerts a relatively small impact on the performance of detecting reconstructed images for DRCT-2M/SDv1-DR and DRCT-2M/SDv2-DR. However, in the case of DRCT-2M/SDXL-DR, the larger the reconstruction step, the closer the reconstructed image gets to the real image (thus improving the quality of generation), which makes it more challenging to identify as a generated image.

A.5. The Sensitivity of the Detector to the λ Parameter in Loss Function

The parameter λ is crucial for balancing the trade-off between contrastive loss and BCE loss. We evaluated the impact of various λ values (0.1, 0.2, 0.3, 0.5, 0.7, 0.9) on the detection performance. The results of Table 9 indicate that as λ progressively increases, the average accuracy of DRCT/Conv-B (trained on DRCT-2M/SDv1.4) on GenImage initially rises, then declines. The optimal average accuracy is achieved at a λ value of 0.3, which we have adopted as the default in our experiments.

A.6. The Sensitivity of the Detector to the m Parameter in Loss Function

The parameter m is the margin for negative sample pairs in contrastive loss. We evaluated the impact of various m values (0.1, 0.5, 1.0, 1.5, 2.0) on the detection performance. Similar to λ, the results of Table 9 indicate that as m progressively increases, the average accuracy of DRCT/Conv-B (trained on DRCT-2M/SDv1.4) on GenImage initially rises, then declines. The optimal average accuracy is achieved at a m value of 1.0, which we have adopted as the default in our experiments.

A.7. The Influence of Image Category on Detection Performance

To further evaluate the influence of image category on detection performance, we conducted a semantics analysis experiment, as done in De-fake (Sha et al., 2023). We used the DRCT/Conv-B detector trained on DRCT-2M/SDv1.4 dataset to test the generated images of DRCT-2M/SDXL dataset. The results highlight the top ten categories with the highest prediction accuracy and the ten with the lowest. As shown in Figure 7. The categories achieving the highest accuracy were "toaster", "microwave" and "hair drier", each achieving a 100% accuracy, while the lowest was "apple" with an accuracy of 25.00%, followed by "baseball glove" with 33.33%, "skateboard" with 41.67%.

Table 5. F1 (%) comparisons of our DRCT and other generated image detectors on DRCT-2M. Except for DIRE and DRCT, all methods are only trained on SDv1.4 and then evaluated on different testing subsets on DRCT-2M. For the training data of DIRE and DRCT, when the Diffusion Reconstructed (DR) model is SDv1, the original fake images were generated by SDv1.4. When the DR model is SDv2, the original fake images were generated by SDv2.

Method	DR	SD Variants						Turbo Variants		LCM Variants		ControlNet Variants			DR Variants			Avg.
		LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	SDv1-DR	SDv2-DR	SDXL-DR	
CNNSpot	-	99.87	99.91	99.90	97.49	49.13	84.48	83.94	61.97	98.23	38.08	97.92	83.59	79.31	35.98	05.67	01.31	69.80
F3Net	-	99.85	99.78	99.79	84.24	21.20	85.57	53.69	43.08	97.32	18.38	97.94	61.95	78.08	47.29	01.90	01.43	61.97
CLIP/RN50	-	**99.99**	99.99	99.96	94.30	38.94	90.63	80.34	47.74	98.96	25.90	**99.97**	76.07	78.10	48.11	02.68	01.90	67.72
GramNet	-	99.40	99.01	98.83	95.10	40.99	76.26	59.92	56.18	92.59	25.54	88.94	67.93	79.23	06.09	01.42	01.69	61.82
De-fake	-	91.45	99.53	99.51	88.50	44.10	55.79	91.34	93.56	99.13	59.13	30.85	39.92	50.24	01.15	01.31	00.68	59.14
Conv-B	-	99.97	**100.0**	**99.97**	95.66	44.82	78.05	76.27	35.39	99.26	39.56	99.80	80.10	63.54	37.79	06.91	01.63	66.17
UnivFD	-	98.29	96.11	96.22	93.48	90.21	93.57	84.39	83.78	89.53	87.75	89.49	76.88	87.83	09.01	05.63	03.47	74.10
DIRE	SDv1	98.16	99.94	99.96	53.33	14.36	61.01	30.21	16.10	**99.78**	32.65	99.65	44.29	30.95	07.76	00.28	00.00	49.28
DIRE	SDv2	17.16	68.30	68.56	**99.87**	33.23	92.58	**99.77**	26.44	85.47	62.21	52.73	99.69	44.86	00.16	09.73	00.00	53.80
DRCT/Conv-B (ours)	SDv1	99.91	99.90	99.90	96.19	80.81	83.25	91.18	57.33	99.66	73.09	99.90	94.76	76.91	**99.90**	95.19	67.43	88.46
DRCT/Conv-B (ours)	SDv2	99.66	98.54	98.46	99.85	95.95	**98.66**	99.59	80.00	98.43	93.38	96.57	**99.85**	**97.61**	93.53	**99.87**	**89.39**	**96.21**
DRCT/UnivFD (ours)	SDv1	96.82	96.33	96.41	94.92	**96.31**	93.41	93.38	92.85	90.89	95.05	95.66	92.56	91.76	94.09	59.57	32.63	88.29
DRCT/UnivFD (ours)	SDv2	94.61	94.51	94.39	95.22	95.78	95.55	94.98	**94.64**	91.67	**95.71**	94.00	93.61	93.67	83.14	81.68	57.45	90.66

Table 6. False Negative Rate (FNR, %) comparisons of our DRCT and other generated image detectors on DRCT-2M. Except for DIRE and DRCT, all methods are only trained on SDv1.4 and then evaluated on different testing subsets on DRCT-2M. For the training data of DIRE and DRCT, when the Diffusion Reconstructed (DR) model is SDv1, the original fake images were generated by SDv1.4. When the DR model is SDv2, the original fake images were generated by SDv2.

Method	DR	SD Variants						Turbo Variants		LCM Variants		ControlNet Variants			DR Variants			Avg.
		LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	SDv1-DR	SDv2-DR	SDXL-DR	
CNNSpot	-	00.16	00.08	00.10	04.80	67.40	26.80	27.60	55.06	03.38	76.46	03.98	28.12	34.22	78.04	97.08	99.34	37.66
F3Net	-	00.12	00.26	00.24	22.50	88.12	25.08	63.24	72.50	05.04	89.86	03.86	55.04	35.84	68.98	99.04	99.28	45.56
CLIP/RN50	-	**00.00**	**00.00**	00.06	10.76	75.82	17.12	32.84	71.18	02.04	85.12	00.50	38.60	35.92	68.32	98.64	99.04	39.75
GramNet	-	00.50	01.28	01.62	08.70	74.04	37.94	56.92	60.66	13.20	85.26	19.36	48.20	33.94	96.84	99.28	99.14	46.06
De-fake	-	15.46	00.60	00.64	20.36	71.62	61.18	15.66	11.80	01.40	57.88	81.70	74.98	66.34	99.42	99.34	99.66	48.63
Conv-B	-	00.06	**00.00**	00.06	08.32	71.12	36.00	38.36	78.50	01.46	75.34	00.40	33.20	53.44	76.70	96.42	99.18	41.79
UnivFD	-	02.54	06.70	06.48	11.48	17.12	11.32	26.38	27.30	18.26	21.16	18.32	37.02	21.02	95.24	97.08	98.22	32.23
DIRE	SDv1	03.56	00.06	**00.02**	63.62	92.26	56.08	82.20	91.24	**00.38**	80.48	00.64	71.54	81.68	95.96	99.86	100.0	57.47
DIRE	SDv2	90.60	48.06	47.76	**00.10**	80.04	13.68	**00.30**	84.74	25.26	54.78	64.14	00.46	71.04	99.92	94.88	100.0	54.74
DRCT/Conv-B (ours)	SDv1	**00.00**	00.02	**00.02**	07.18	32.08	28.56	16.06	59.74	00.50	42.30	**00.02**	00.98	37.40	**00.02**	09.02	49.04	17.68
DRCT/Conv-B (ours)	SDv2	00.50	02.70	02.86	00.12	07.62	02.46	00.64	33.22	02.92	12.26	06.46	**00.12**	04.50	12.00	**00.08**	**19.04**	**06.72**
DRCT/UnivFD (ours)	SDv1	00.76	01.72	01.58	04.46	01.76	07.32	07.38	08.36	11.90	04.22	03.04	08.88	10.34	06.04	55.14	79.38	13.27
DRCT/UnivFD (ours)	SDv2	02.58	02.78	03.00	01.38	**00.26**	**00.72**	01.86	**02.52**	08.16	**00.40**	03.76	04.52	**04.40**	22.80	25.08	56.26	08.78

B. Additional Details of DRCT-2M and DRCT-2M-wild Datasets

B.1. More Details of DRCT-2M

In this subset, we introduce more details about our proposed DRCT-2M dataset. The real images of DRCT-2M are derived from MSCOCO (Lin et al., 2014), as well as fake images generated by 16 different types of SD models. There are 10 types of text-to-image SD models and 6 types of image-to-image SD models, and their pre-training weights are downloaded from the open source library: diffusers (von Platen et al., 2022). To generate fake images, different SD models have different inference step and input conditions. For text-to-image SD models (including "LDM", "SDv1.4", "SDv1.5", "SDv2", "SDXL", "SDXL-Refiner", "SD-Turbo", "SDXL-Turbo ", "LCM-SDv1.5" and "LCM-SDXL"), their input condition are text prompts, which corresponds to the caption of the real images in MSCOCO. For the image-to-image SD model, "SDv1-Ctrl", "SDv2-Ctrl" and "SDXL-Ctrl" are combined with ControlNet (Zhang et al., 2023) to generate controllable images. Their input conditions are text prompts and Canny Edge Map (called "canny") extracted by the Canny edge detection algorithm (Canny, 1986), where the prompts are the caption of the real images in MSCOCO, and the canny are extracted from the real images. Moreover, "SDv1-DR", "SDv2-DR", and "SDXL-DR" are utilized to generate reconstruction images of the real images using the SD inpainting models, with input conditions being the image, text prompt, and binary mask, where the input image comes from the real image in MSCOCO, the prompt is an empty string "", and the mask is a zero matrix of the same size as the real image. Further details of the DRCT-2M dataset are illustrated in Table 11.

To evaluate the quality of images generated by different types of SD models in the DRCT-2M dataset, we made use of two generated image quality assessment methods: HPSv2 (Wu et al., 2023b) and ImageReward (Xu et al., 2023). As shown in Figure 8, within the HPSv2 assessment, "SDv2" received the highest score, followed by "SDXL", with "LDM" performing the worst. In the ImageReward evaluation, "SDXL-Turbo" achieved the highest score, with "SDXL" next, and "LDM" ranking lowest. Moreover, some

Table 7. Accuracy (ACC, %) comparisons of our DRCT and other generated image detectors. Except for DIRE and DRCT, all methods were trained on DRCT-2M/SDv1.4 and evaluated on different testing subsets of GenImage. For the training data of DIRE and DRCT, when the Diffusion Reconstructed (DR) model is SDv1, the original fake images were generated by SDv1.4. When the DR model is SDv2, the original fake images were generated by SDv2.

Method	RD	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
CNNSpot	-	71.72	95.72	95.90	53.07	53.68	91.67	53.60	49.87	70.65
F3Net	-	71.44	97.35	97.49	55.73	61.42	87.01	62.52	49.68	72.83
CLIP/RN50	-	69.74	93.03	93.39	52.06	50.64	80.94	54.29	49.87	68.00
GramNet	-	68.38	86.08	86.65	51.83	53.50	75.87	53.75	49.51	65.70
De-fake	-	57.03	85.22	85.33	54.87	59.57	77.28	54.69	52.40	65.80
Conv-B	-	67.79	95.64	95.85	51.44	52.01	86.13	52.97	49.99	68.98
UnivFD	-	73.08	73.72	73.78	55.95	71.40	74.04	56.64	63.20	67.73
DIRE	SDv1	51.11	55.07	55.31	49.93	50.02	53.71	49.87	49.85	51.86
DIRE	SDv2	59.60	50.42	50.51	49.67	49.76	50.79	49.64	49.63	51.25
DRCT/Conv-B(ours)	SDv1	94.43	**99.37**	**99.19**	66.42	73.31	**99.25**	76.85	59.41	83.53
DRCT/Conv-B(ours)	SDv2	**98.26**	97.88	97.83	60.00	60.02	95.89	61.32	52.33	77.94
DRCT/UnivFD(ours)	SDv1	85.82	92.33	91.87	**75.18**	**87.44**	92.23	**89.12**	**87.38**	**87.67**
DRCT/UnivFD(ours)	SDv2	88.55	88.39	88.04	71.61	77.92	87.55	84.98	84.83	83.98

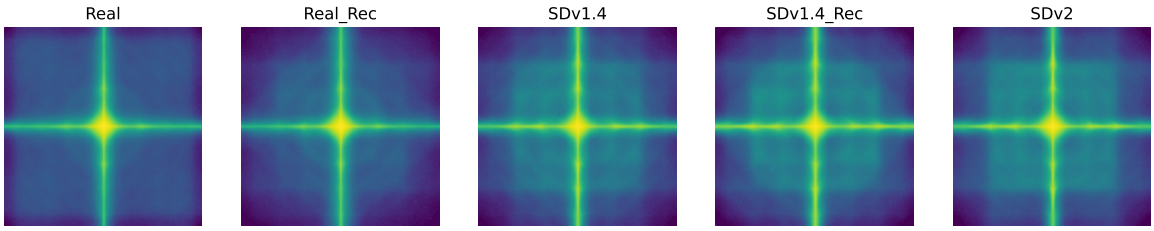


Figure 6. Average frequency spectra for real images, real reconstructed images (labeled "Real_Rec"), generated images (labeled "SDv1.4" and "SDv2") and generated reconstructed images (labeled "SDv1.4_Rec").

examples of generated images in DRCT-2M dataset are illustrated in Figure 9.

B.2. More Details of DRCT-2M-wild

To evaluate the generalization of our proposed DRCT method to generated images in real-world scenarios, we collected eight types of generated images from two internet platforms, CIVITAI and DISCORD. Except for Midjourney, which was sourced from DISCORD, the rest were gathered from CIVITAI. More details about DRCT-2M-wild are presented in Table 12, and some examples of generated images from DRCT-2M-wild are illustrated in Figure 10.

Table 8. Ablation study results demonstrate the effect of the reconstruction step. We trained all classifiers (Conv-B as the backbone) using the DRCT framework on DRCT-2M/SDv1.4 and DRCT-2M/SDv2 with different reconstruction steps, and then tested on three subsets: DRCT-2M/SDv1-DR, DRCT-2M/SDv2-DR, and DRCT-2M/SDXL-DR with different reconstruction steps (namely 10, 20, 50, 70). We report ACC (%)/ AUC (%) in the Table.

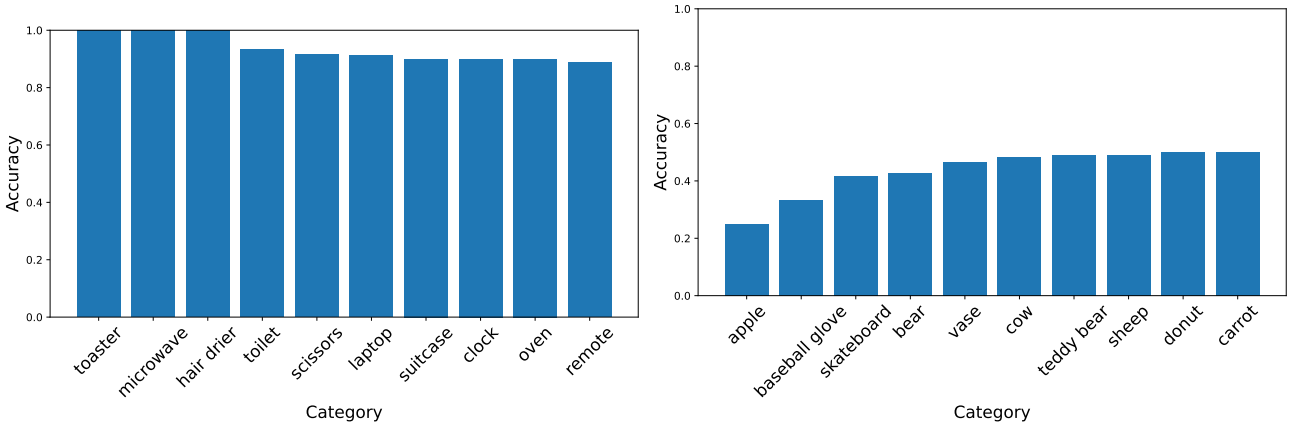
Test set	Trained on DRCT-2M/SDv1.4				Trained on DRCT-2M/SDv2			
	step=10	step=20	step=50	step=70	step=10	step=20	step=50	step=70
DRCT-2M/SDv1-DR(step=10)	99.94/100.0	99.90/100.0	99.94/100.0	99.98/100.0	95.78/99.30	94.18/99.71	92.69/99.28	95.62/99.64
DRCT-2M/SDv1-DR(step=20)	99.94/100.0	99.90/100.0	99.94/100.0	99.98/100.0	95.59/99.29	94.06/99.69	92.57/99.26	95.34/99.61
DRCT-2M/SDv1-DR(step=50)	99.94/100.0	99.90/100.0	99.94/100.0	99.98/100.0	95.44/99.28	93.91/99.69	92.48/99.26	95.03/99.59
DRCT-2M/SDv1-DR(step=70)	99.94/100.0	99.90/100.0	99.94/100.0	99.98/100.0	99.48/99.28	93.96/99.68	92.50/99.26	95.17/99.60
DRCT-2M/SDv2-DR(step=10)	95.78/99.85	96.28/99.66	92.47/99.56	93.97/99.83	99.91/100.0	99.90/100.0	99.96/100.0	99.95/100.0
DRCT-2M/SDv2-DR(step=20)	95.01/99.81	95.79/99.58	91.43/99.42	92.88/99.68	99.90/100.0	99.88/100.0	99.96/100.0	99.94/100.0
DRCT-2M/SDv2-DR(step=50)	94.47/99.79	95.40/99.95	91.00/99.35	92.22/99.62	99.88/100.0	99.87/100.0	99.95/100.0	99.94/100.0
DRCT-2M/SDv2-DR(step=70)	94.94/99.80	95.62/99.54	91.25/99.38	92.53/99.64	99.89/100.0	99.88/100.0	99.95/100.0	99.94/100.0
DRCT-2M/SDXL-DR(step=10)	80.98/98.76	82.84/97.09	75.30/96.25	73.51/96.52	97.37/99.80	93.28/99.59	95.36/99.79	93.80/99.50
DRCT-2M/SDXL-DR(step=20)	77.98/98.00	79.14/95.40	72.54/94.55	69.79/94.78	96.91/99.76	91.84/99.45	94.63/99.70	92.49/99.32
DRCT-2M/SDXL-DR(step=50)	74.50/97.03	75.39/93.08	70.06/92.04	66.17/92.13	96.44/99.70	90.39/99.18	93.88/99.59	91.08/98.96
DRCT-2M/SDXL-DR(step=70)	74.94/97.18	76.04/93.39	70.41/92.29	66.59/92.42	96.47/99.70	90.57/99.23	93.91/99.60	91.26/99.03

Table 9. Ablation study results demonstrating the sensitivity of detector to λ parameter in loss function. The detectors are DRCT/Conv-B, which were trained on DRCT-2M/SDv1.4 and tested on different testing subsets of GenImage. We report ACC (%) in the Table.

λ	Midjourney	SDv1.4	SDv1.5	AMD	GLIDE	Wukong	VQDM	BigGAN	Avg.
0.1	88.92	**99.82**	**99.74**	53.48	55.43	**99.82**	66.59	51.75	76.94
0.2	92.36	99.57	99.41	63.31	65.10	99.54	72.12	58.08	81.19
0.3(default)	94.43	99.37	99.19	**66.42**	**73.31**	99.25	76.85	**59.41**	**83.53**
0.5	94.85	99.40	99.16	60.33	64.74	99.08	75.28	52.47	80.66
0.7	**96.33**	99.24	99.04	65.72	67.04	99.15	**83.39**	56.47	83.30
0.9	92.43	99.36	99.21	58.53	66.43	99.35	69.41	55.31	80.00

Table 10. Ablation study results demonstrating the sensitivity of detector to margin (m) parameter in loss function. The detectors are DRCT/Conv-B, which were trained on DRCT-2M/SDv1.4 and tested on different testing subsets of GenImage. We report ACC (%) in the Table.

m	Midjourney	SDv1.4	SDv1.5	AMD	GLIDE	Wukong	VQDM	BigGAN	Avg.
0.1	88.60	**99.88**	**99.81**	56.66	60.79	**99.86**	63.13	53.71	77.81
0.5	92.30	99.57	99.46	62.06	63.28	99.57	73.22	54.16	80.45
1.0(default)	**94.43**	99.37	99.19	**66.42**	**73.31**	99.25	**76.85**	**59.41**	**83.53**
1.5	91.53	99.59	99.46	57.53	56.32	99.57	67.33	51.68	77.88
2.0	87.88	99.79	99.72	57.06	57.57	99.70	64.93	52.45	77.39



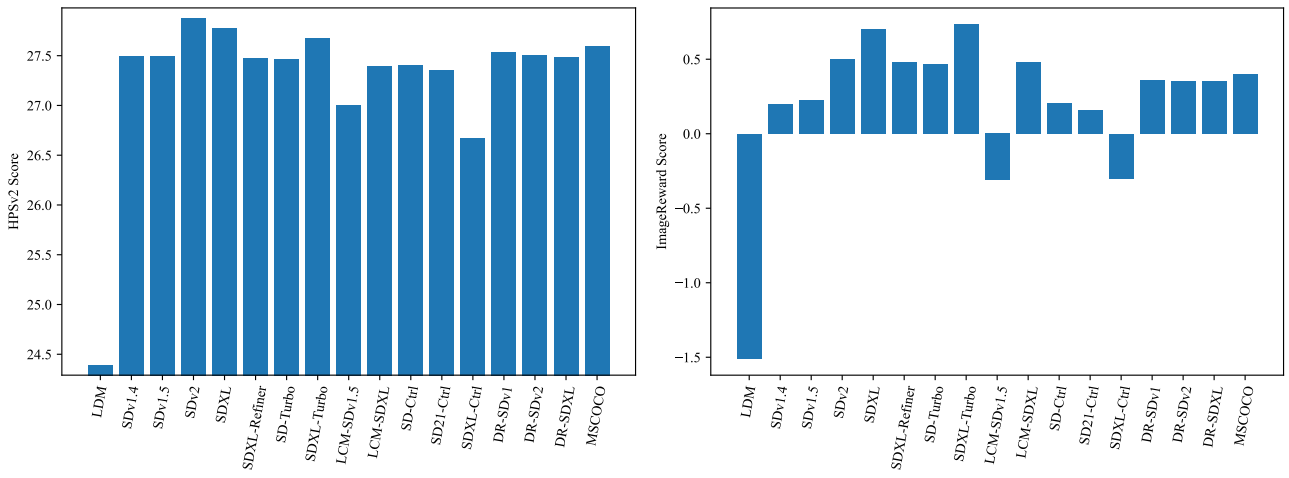
(a) The 10 categories with the highest accuracy.

(b) The 10 categories with lowest accuracy.

Figure 7. The influence of image category on detection performance.

Table 11. More details of **DRCT-2M** dataset.

Category	Diffusion Model	Model Name in diffusers (von Platen et al., 2022)	Image Count	Inference Step	Conditions
SD Variants	LDM	CompVis/ldm-text2im-large-256	123287	50	prompt
	SDv1.4	CompVis/stable-diffusion-v1-4	123287	50	prompt
	SDv1.5	runwayml/stable-diffusion-v1-5	123287	50	prompt
	SDv2	stabilityai/stable-diffusion-2-1	123287	50	prompt
	SDXL	stabilityai/stable-diffusion-xl-base-1.0	123287	50	prompt
	SDXL-refiner	stabilityai/stable-diffusion-xl-refiner-1.0	123287	50	prompt
Turbo Variants	SD-Turbo	stabilityai/sd-turbo	123287	1	prompt
	SDXL-Turbo	stabilityai/sdxl-turbo	123287	1	prompt
LCM Variants	LCM-SDv1.5	latent-consistency/lcm-lora-sdv1-5	123287	4	prompt
	LCM-SDXL	latent-consistency/lcm-lora-sdxl	123287	4	prompt
ControlNet Variants	SDv1-Ctrl	lllyasviel/sd-controlnet-canny	123287	20	prompt+canny
	SDv2-Ctrl	thepowefuldeez/sd21-controlnet-canny	123287	20	prompt+canny
	SDXL-Ctrl	diffusers/controlnet-canny-sdxl-1.0	123287	20	prompt+canny
DR Variants	SDv1-DR	runwayml/stable-diffusion-inpainting	123287	50	image+prompt+mask
	SDv2-DR	stabilityai/stable-diffusion-2-inpainting	123287	50	image+prompt+mask
	SDXL-DR	diffusers/stable-diffusion-xl-1.0-inpainting-0.1	123287	50	image+prompt+mask



(a)

(b)

Figure 8. Diffusion-generated image of DRCT-2M quality assessment by HPSv2 (Wu et al., 2023b) and ImageReward (Xu et al., 2023).

Figure 9. Some examples of generated images in **DRCT-2M**. The real image is "000000000139.jpg", "000000000632.jpg" and "000000000724.jpg" from the MSCOCO dataset. The prompt used for text-generated images is "A woman stands in the dining area at the table.", "Bedroom scene with a bookcase, blue comforter and window.", "A stop sign is mounted upside-down on it's post.", respectively. We constructed datasets using 16 types of SD models, including LDM, SDv1.4, SDv1.5, SDv2, SDXL, SDXL-refiner, SD-Turbo, SDXL-Turbo, LCM-SDv1.5, LCM-SDXL, SDv1-Ctrl, SDv2-Ctrl, SDXL-Ctrl, SDv1-DR, SDv2-DR and SDXL-DR, where "Ctrl" means "ControlNet" and "DR" means "Diffusion Reconstruction". Specifically, we utilized SDv1-DR, SDv2-DR and SDXL-DR models to reconstruct the real image set from MSCOCO.

Table 12. More details of **DRCT-2M-Wild** dataset in real-world scenarios.

Method Name	Base Model	Model ID in CIVITAI	Image Count
DreamShaper XL10	SDXL	112902	8,696
Niji Special Edition	SDXL	120765	3,615
Realistic Vision v5.1	SDv1.5	4201	36,999
Deep Negative v1.x	SDv1.5	4629	11,600
Detail Tweaker Lora	SDv1.5	58390	24,293
MajicMix Realistic	SDv1.5	43331	41,496
rMada Merge	SDv2	15303	769
Midjourney	-	-	9349

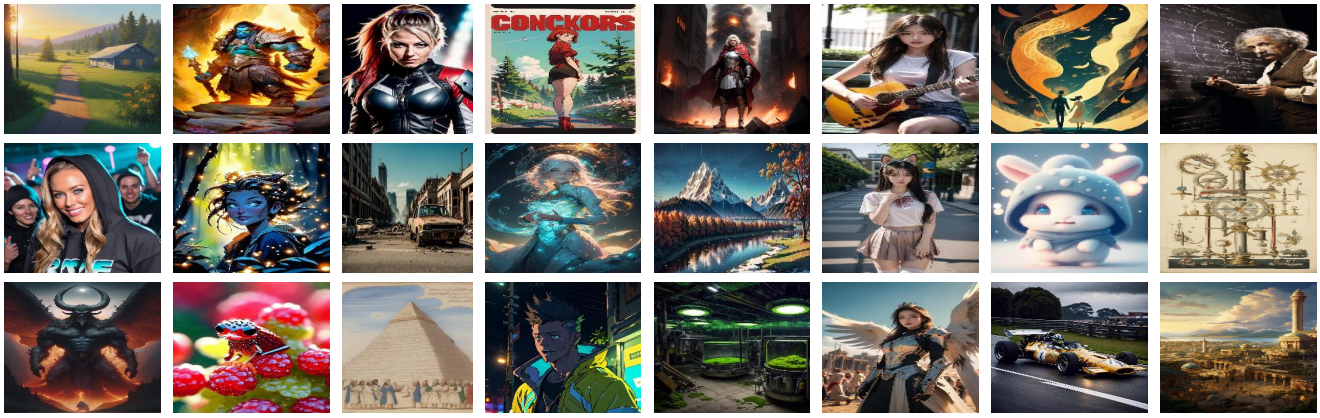


Figure 10. Some examples of generated images in **DRCT-2M-wild**. These images mainly consist of those generated by 8 models on internet platforms CIVITAI and Midjourney, including DreamShaper XL10, Niji Special Edition, Realistic Vision v5.1, Deep Negative v1.x, Detail Tweaker Lora, MajicMix Realistic, rMada Merge, Midjourney.