DistilDIRE: A Small, Fast, Cheap and Lightweight Diffusion Synthesized Deepfake Detection

Yewon Lim [*1 2 3] **Changyeon Lee** [*1 2 3] **Aerin Kim** [3 4] **Oren Etzioni** [4]

Abstract

A dramatic influx of diffusion-generated images has marked recent years, posing unique challenges to current detection technologies. While the task of identifying these images falls under binary classification, a seemingly straightforward category, the computational load is significant when employing the "reconstruction then compare" technique. This approach, known as DIRE (Diffusion Reconstruction Error), not only identifies diffusion-generated images but also detects those produced by GANs, highlighting the technique's broad applicability. To address the computational challenges and improve efficiency, we propose distilling the knowledge embedded in diffusion models to develop rapid deepfake detection models. Our approach, aimed at creating a small, fast, cheap, and lightweight diffusion synthesized deepfake detector, maintains robust performance while significantly reducing operational demands. Maintaining performance, our experimental results indicate an inference speed 3.2 times faster than the existing DIRE framework. This advance not only enhances the practicality of deploying these systems in real-world settings but also paves the way for future research endeavors that seek to leverage diffusion model knowledge. The code and weights for our framework are available at https://github.com/miraflow/DistilDIRE.

[*]Equal contribution [1]Department of Computer Science and Engineering, Yonsei University, Seoul, Republic of Korea [2]Computer and Information Technology, Purdue University, West Lafayette IN, United States [3]Miraflow, Kirkland WA, United States [4]TrueMedia.org, Seattle WA, United States. Correspondence to: Aerin Kim <aerin@miraflow.ai>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

1. Introduction

From generating high-fidelity images (Rombach et al., 2022; Karras et al., 2017; 2019; Saharia et al., 2022) to synthesizing complex videos (Brooks et al., 2024; Bar-Tal et al., 2024; Gao et al., 2024) and musical compositions (Huang et al., 2023), and deepfakes (Chen et al., 2023; Bhattacharyya et al., 2024), diffusion models (Ho et al., 2020; Song et al., 2022) have demonstrated exceptional capabilities in various domains by leveraging stochastic diffusion processes.

Current deepfake detectors struggle to distinguish unseen diffusion-generated images. To address this, researchers have developed a new method that uses a pre-trained diffusion model to more accurately reconstruct these images. This method uses the error between an input image and its reconstructed version, known as the DIRE (Wang et al., 2023) representation, as input for a binary classifier to detect deepfakes. This 'reconstruction then compare' approach has achieved state-of-the-art results on the Diffusion-Forensic dataset (Wang et al., 2023), which includes both GAN (Goodfellow et al., 2020) generated and diffusion-generated images.

The DIRE framework, although it works well, is too slow for practical use because it needs to compute the diffusion trajectory during training and inference. We address this limitation by proposing a distilling approach that uses a fraction of the information from pretrained diffusion model and predefined teacher classifier model. Specifically, We leverage the DIRE feature extracted by a ImageNet pre-trained ResNet-50 classifier (*teacher*) to train a binary classifier (*student*) for deepfake detection. Additionally, we incorporate the first-time step noise from a pre-trained diffusion model with the original image, effectively conveying the image distribution as interpreted by the diffusion.

This method has led to the development of '***DistilDIRE***', a small, fast, cheap and lightweight deepfake detection model suitable for real-world applications.

2. Related Work

Our work aims to develop a light and fast diffusion synthesized deepfake detection model with only fraction of

diffusion trajectory information.

2.1. Denoising Diffusion Implicit Models

Diffusion models introduce a novel approach to generative modeling by leveraging inverse transformations to progressively reduce noise in data generation. While DDPM (Ho et al., 2020) constructs the generation process using Marcov chains and thus require sequential sampling without skipping, DDIM (Song et al., 2022) can generate images faster than DDPM due to its unique design without Marcov Chain. It models reverse process as follows.

$$\mathrm{x}_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{\mathrm{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathrm{x}_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta(\mathrm{x}_t, t) + \sigma_t \epsilon_t \tag{1}$$

When σ_t goes 0, Equation (1) becomes deterministic from which we can approximately revert image to noise, which is called DDIM inversion. Specifically, DDIM inversion is performed as follows.

$$\frac{\mathrm{x}_{t+1}}{\sqrt{\alpha_{t+1}}} = \frac{\mathrm{x}_t}{\sqrt{\alpha_t}} + \left(\sqrt{\frac{1-\alpha_{t+1}}{\alpha_{t+1}}} - \sqrt{\frac{1-\alpha_t}{\alpha_t}} \right) \epsilon_\theta(\mathrm{x}_t, t). \tag{2}$$

This process aims to obtain the corresponding noisy sample x_T for an input image x_0. However, inverting or sampling step-by-step is highly time-consuming. To accelerate the diffusion model sampling, DDIM allows sampling a subset of S steps $\tau_1, \ldots, \tau_S$, where the neighboring x_t and x_{t+1} become x_{τ_t} and $\mathrm{x}_{\tau_{t+1}}$, respectively, in Equation (1) and Equation (2).

2.2. Diffusion Synthesized Image Detection

As diffusion models advance and generate seamless images, it becomes increasingly challenging to detect these artificially created ones. To tackle this, DIRE (Wang et al., 2023) offers a solution by leveraging pre-trained diffusion models and extracting their implicit understanding of image distributions. Their research reveals that genuine images exhibit significant reconstruction errors when subjected to DDIM inversion (Song et al., 2022), whereas counterfeit images generated by diffusion models tend to yield lower reconstruction errors.

However, a notable bottleneck arises in terms of inference time, as implementing the DIRE framework requires invoking the ADM model at least 40 times to obtain the DIRE image.

DNF, or Diffusion Noise Feature (Zhang & Xu, 2024), represents another advancement in diffusion-generated image detection. This method delves into the intricate process of diffusion, revealing that the predicted noises generated during the inverse diffusion process can serve as valuable indicators of the underlying image distribution. By exploiting these noise features, deciding whether given image is genuine or counterfeit can be done with higher accuracy.

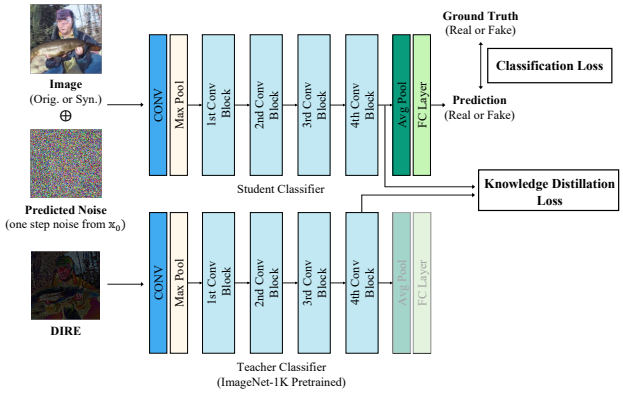

Figure 1. Overview of our framework.

However, DNF also requires multiple iterations of diffusion model calls before the deepfake probability can be calculated. This highlights the computational complexity involved in such image detection methods and underscores the ongoing efforts to streamline and optimize these processes for practical implementation.

3. Methodology

3.1. Architecture

The DistilDire framework, shown in Figure 1, uses a pretrained ResNet-50 (He et al., 2016), which was initially trained on the ImageNet-1K (Deng et al., 2009) dataset, as a teacher model.

The teacher model, loaded with pretrained weights, remains frozen during training. This model provides a stable reference for the student model, ensuring that the extracted features are robust and generalized. Unlike the teacher model, the student model is trained from scratch. During the training process, the student model simultaneously incorporates classification loss and knowledge distillation loss.

What makes our method unique is that we combine the original image and the predicted noise from that original image into the learning (student) model. Specifically, we utilize noise from the first time step obtained from a pretrained Ablated Diffusion Model (ADM) (Dhariwal & Nichol, 2021) as an additional input. The final input to the student model consists of the original image **x** itself and the noise ϵ_0 from the first denoising step. The input to the model is as follows,

where i represents the index of the mini-batch.

$$\mathbf{x}'_i = \mathsf{cat}(\mathbf{x}_i, \epsilon_{0,i}) \tag{3}$$

The student model then predicts whether the input image is real or synthetic (fake).

Knowledge distillation from the teacher model is achieved by extracting feature maps before the classifier head and computing the Mean Squared Error (MSE) loss between these feature maps and those of the student model. The knowledge distillation loss serves as regularization, encouraging the student model's feature maps to mirror the teacher model's without the need for time-intensive inversion and reconstruction.

The student model uses Binary Cross-Entropy Loss (BCE). This helps the model learn how to tell the difference between real and synthetic images. The final way we calculate loss is shown in Equation (4), Equation (5), and Equation (6).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{classification}} + \lambda \cdot \mathcal{L}_{\text{distillation}} \tag{4}$$

$$\mathcal{L}_{\text{classification}} = -\frac{1}{N}\sum_{i=1}^{N}(\mathbf{y}_i \log(\mathbf{y}'_i) + (1-\mathbf{y}_i)\log(1-\mathbf{y}'_i)) \tag{5}$$

$$\mathcal{L}_{\text{distillation}} = \frac{1}{N}\sum_{i=1}^{N}\|\mathcal{F}'(\mathsf{DIRE}(\mathbf{x}_i)) - \mathcal{F}(\mathbf{x}'_i)\|_2 \tag{6}$$

In the above equations, $\mathbf{y}_i, \mathbf{y}'_i$, $\mathcal{F}'$, and $\mathcal{F}$ represents ground-truth label, model prediction, ImageNet-1K pretrained ResNet-50 feature extractor, and naive ResNet-50 feature extractor, respectively. Here, N is a mini-batch size and λ decides how much the distillation process will affect the outcome.

In essence, DistilDIRE uses a pretrained teacher model for generalized feature extraction, while the student model learns its ability to distinguish real from synthetic images through classification and knowledge distillation losses. This approach balances the use of pretrained knowledge and effective learning of new representations.

3.2. Noise Extraction

In DistilDIRE, we extract the initial noise ϵ_0 to determine whether a given image is generated by a diffusion model. By analyzing the characteristics of ϵ_0, we aim to identify signatures that are indicative of diffusion-generated images.

To extract ϵ_0, we leverage the DDIM formulation (Equation (2)). By focusing on the case when $t = 1$, we can isolate the initial noise ϵ_0 from the initial image $\mathbf{x}_0$:

$$\epsilon_0 = \frac{\mathbf{x}_1 - \sqrt{\alpha_1}\hat{\mathbf{x}}_0}{\sqrt{1-\alpha_1}}. \tag{7}$$

Extracted ϵ_0 is concatenated with the input image $\mathbf{x}_0$ and then forwared to classifier to determine whether given image is synthesized or not.

4. Experiment

4.1. Experiments Setup

4.1.1. Datasets and Pre-processing

For the experiments in this study, we utilized the DiffusionForensics dataset (Wang et al., 2023), incorporating specific subsets from ImageNet and CelebA-HQ (Karras et al., 2017). LSUN-Bedroom (Yu et al., 2015) subset is excluded in the experiment for practical reason.

In the training phase, both original images and noise features underwent resizing to 224 x 224, horizontal flipping with a 0.5 probability and normalization. The first time step noises derived from ADM were resized to 224 x 224 as well and horizontally flipped with an equivalent probability to the images and DIRE images. However, they were not normalized. During testing, images were centered and cropped to a size of 224 x 224. Different lambda λ values for each sub-dataset are tested during the training phase. These values are tuned to optimize the balance between classification loss and knowledge distillation loss, therefore enhancing the model's effectiveness across different types of data. For reporting, λ of 0.5 is applied for both ImageNet and CelebA-HQ subset.

4.1.2. Evaluation

To evaluate the performance of classifier in categorizing synthetic (fake) deepfake images, we employed metrics including, accuracy, and average precision (AP). The accuracy threshold for computation was set at 0.5. Evaluation was conducted by constructing datasets for each subset of ImageNet and CelebA-HQ, comprising both real images and their corresponding synthetic counterparts.

4.2. Results

4.2.1. Deepfake Detection Performance

Our evaluation of the DistilDIRE framework demonstrates a meaningful significant advancement in deepfake detection through a balance of high performance and reduced computational demand. The results, as detailed in Table 1, showcase the efficacy of DistilDIRE across varied test conditions on the ImageNet and CelebA-HQ datasets.

In the ImageNet subset, DistilDIRE achieved accuracy and average precision (AP) scores close to the best-performing models, with metrics like 99.0/99.7% against SD-v1 and 98.4/99.5% against ADM. For the CelebA-HQ subset, it performed exceptionally well against modern generators such

Table 1. Quantitative evaluation (Accuracy (%), AP (%)) of our framework alongside other state-of-the-art deepfake detection methodologies on the ImageNet and CelebA-HQ datasets. '*' denotes the generators present in the training dataset of each subset dataset. ImageNet test subset image generators are ADM (Dhariwal & Nichol, 2021), Stable Diffusion v1 (Rombach et al., 2022) and CelebA-HQ test subset image generators are IF (Deep Floyd IF, 2023), DALL-E 2 (Ramesh et al., 2022), Stable Diffusion v2 (Rombach et al., 2022), Midjourney (Midjourney, 2022)

Method	Training Dataset	ImageNet Testing Generators		CelebA-HQ Testing Generators				Performance Average
		ADM*	SD-v1	IF	DALL-E 2	SD-v2*	Midjourney	
CNNDet (Wang et al., 2020)	DiffusionForensics - ImageNet	71.6/79.8	51.0/51.2	36.5/41.2	54.2/52.2	37.0/41.6	48.4/49.1	49.8/52.5
	DiffusionForensics - CelebA-HQ	51.0/58.8	52.6/68.0	53.6/53.9	54.2/52.2	78.4/69.9	73.6/67.7	60.6/61.8
DIRE (Wang et al., 2023)	DiffusionForensics - ImageNet	99.8/99.9	98.2/99.9	50.0/50.0	50.0/50.0	50.0/50.0	50.0/50.0	66.3/66.6
	DiffusionForensics - CelebA-HQ	99.8/99.9	58.2/66.2	96.8/100	93.4/100	96.7/100	95.0/100	90.0/94.4
Ours (DistilDIRE)	DiffusionForensics - ImageNet	98.4/99.5	99.0/99.7	86.5/95.2	89.3/99.2	77.3/86.7	85.5/88.6	89.3/94.8
	DiffusionForensics - CelebA-HQ	61.5/79.4	50.2/89.4	97.4/99.7	93.3/99.1	100/100	99.6/100	83.7/94.6

as SD-v2 and Midjourney. These results are particularly noteworthy as they come close to those of the DIRE model, which, while slightly more accurate, requires substantially higher computational resources.

In addition to diffusion model-generated images, our results indicate that DistilDIRE is also highly effective in detecting GAN-generated images. This capability is crucial as GANs are a common method for creating synthetic media. The robust performance across both GAN-generated and diffusion model-generated images signifies the versatility and effectiveness of the DistilDIRE framework.

Compared to traditional models such as CNNDet and even the state-of-the-art DIRE, DistilDIRE not only stands out in terms of detection accuracy but also in computational efficiency. While DIRE provides top-tier detection capabilities, its practical application is hindered by its high computational load. DistilDIRE, on the other hand, offers a compelling alternative by reducing the computational demand by approximately 97% compared to DIRE, as noted in our inference time and computational efficiency evaluations.

4.2.2. Inference Time Comparison

The inference time and computational efficiency of the models were evaluated, specifically focusing on the performance of these models on a single NVIDIA A10 GPU, widely used for deep learning inference in 2024 at the time of writing. The detailed results of this evaluation are presented in Table 2. The FLOPS was measured using PyTorch FLOPS count implementation (TorchTNT).

The newly proposed DistilDIRE model significantly optimizes computational resources compared to its predecessor, DIRE. While DIRE requires a substantial 149.62 TFLOPS (Tera Floating Point Operations per Second) to process an image, DistilDIRE reduces this demand to just 5.01 TFLOPS, achieving a remarkable reduction by approximately 97%. Additionally, the inference time for DIRE

Table 2. Inference Time Comparison for 256 x 256 Images

Method	Avg. Inference Time (sec/it)	# of Params (M)	FLOPS (T)
CNNDet (Wang et al., 2020)	1.687	23.51	0.021
DIRE (Wang et al., 2023)	6.978	576.32	149.62
DNF (Zhang & Xu, 2024)	3.226	137.18	20.88
Ours (DistilDIRE)	2.183	576.33	5.01

is around 6.978 seconds per image, which is reduced by approximately 69% to 2.183 seconds in DistilDIRE (3.2 times faster than DIRE).

CNNDet, another model listed, shows the least computational demand with 0.021 TFLOPS and the shortest inference time at 1.687 seconds per image. This efficiency can be attributed to CNNDet not utilizing the computationally intensive diffusion process used in DIRE and DNF models.

Comparing the DistilDIRE model with DNF, which also employs a diffusion process, DistilDIRE demonstrates improved performance. DNF has a relatively high computational requirement of 20.88 TFLOPS and an inference time of 3.226 seconds per image. In contrast, DistilDIRE, while having a slightly higher computational load than CNNDet, manages to perform faster than DNF, indicating an efficient balance between computational complexity and processing time.

4.2.3. Ablation Study

In this ablation study, we aimed to examine the effect of incorporating knowledge distillation (KD) loss alongside classification loss during the training process. The primary objective was to determine whether the inclusion of KD loss contributes significantly to model performance across two datasets: ImageNet and CelebA-HQ. As shown in Table 3, the inclusion of KD loss markedly improved Accuracy and Average Precision (AP) for all tested generators. Excluding KD loss, by training with only classifcation loss, led to noticeable declines in both Accuracy and AP, particularly with new generators, underscoring the importance of KD

Table 3. Ablation study on DistilDIRE knowledge distillation (KD) loss on the ImageNet and CelebA-HQ datasets (Accuracy (%), AP (%)).

With KD Loss	Training Dataset	ImageNet Testing Generators		CelebA-HQ Testing Generators				Performance Average
		ADM*	SD-v1	IF	DALL-E 2	SD-v2*	Midjourney	
✓	DiffusionForensics - ImageNet	**98.4/99.5**	**99.0/99.7**	**86.5/95.2**	**89.3/99.2**	**77.3/86.7**	**85.5/88.6**	**89.3/94.8**
	DiffusionForensics - CelebA-HQ	**61.5/79.4**	**50.2/89.4**	**97.4/99.7**	**93.3/99.1**	**100/100**	**99.6/100**	**83.7/94.6**
	DiffusionForensics - ImageNet	95.9/99.5	96.6/99.5	64.7/92.8	55.0/98.5	55.1/73.9	37.7/17.6	67.5/80.3
	DiffusionForensics - CelebA-HQ	56.1/67.0	48.8/87.0	53.2/87.9	69.9/67.3	99.9/100	91.1/36.7	69.8/74.3

loss for optimal performance.

The cross validation analysis indicates that incorporating KD loss not only boosts the Accuracy and AP but also enhances the model's ability to generalize across different generators and datasets. The performance average distinctly favored configurations with KD loss, with a notable 17.9% and 17.4% increase in Accuracy and AP respectively when KD loss was used.

5. Conclusion

The strength of DistilDIRE lies in its strategic use of knowledge distillation, where the model learns to mimic the decision-making process of a more complex system (DIRE) without engaging in computationally intensive tasks. This approach not only preserves high detection capabilities but also ensures the model is practical for real-time applications. DistilDIRE achieves approximately 3.2 times faster speed than existing framework while maintaining high performance. This combination of fast inference speed and high performance can efficiently handle large inputs such as deepfake videos.

References

Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., Li, Y., Rubinstein, M., Michaeli, T., Wang, O., Sun, D., Dekel, T., and Mosseri, I. Lumiere: A space-time diffusion model for video generation, 2024.

Bhattacharyya, C., Wang, H., Zhang, F., Kim, S., and Zhu, X. Diffusion deepfake, 2024.

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., and Ramesh, A. Video generation models as world simulators. 2024.

Chen, Y., Haldar, N. A. H., Akhtar, N., and Mian, A. Text-image guided diffusion model for generating deepfake celebrity interactions, 2023.

Deep Floyd IF. IF. https://github.com/deep-floyd/IF, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., and Li, H. Vista: A generalizable driving world model with high fidelity and versatile controllability, 2024.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020.

Huang, Q., Park, D. S., Wang, T., Denk, T. I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C. H., Engel, J., Le, Q. V., Chan, W., and Han, W. Noise2music: Text-conditioned music generation with diffusion models. *ArXiv*, abs/2302.03917, 2023.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

Midjourney. Midjourney. https://www.midjourney.com, 2022.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., 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, 2022.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models, 2022.

TorchTNT. torchtnt.utils.flops.FlopTensorDispatchMode; TorchTNT 0.2.1 documentation — pytorch.org. https://pytorch.org/tnt/stable/utils/generated/torchtnt.utils.flops.FlopTensorDispatchMode.html.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020.

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*, pp. 22445–22455, 2023.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zhang, Y. and Xu, X. Diffusion noise feature: Accurate and fast generated image detection, 2024.