An Analysis of Recent Advances in Deepfake Image Detection in an Evolving Threat Landscape

Sifat Muhammad Abdullah, Aravind Cheruvu, Shravya Kanchi,
Taejoong Chung, Peng Gao, Murtuza Jadliwala*, Bimal Viswanath
Virginia Tech, *UT San Antonio
{sifat, acheruvu, shravya, tijay, penggao, vbimal}@vt.edu, *murtuza.jadliwala@utsa.edu

***Abstract*—Deepfake or synthetic images produced using deep generative models pose serious risks to online platforms. This has triggered several research efforts to accurately detect deepfake images, achieving excellent performance on publicly available deepfake datasets. In this work, we study 8 state-of-the-art detectors and argue that they are far from being ready for deployment due to two recent developments. First, the emergence of lightweight methods to customize large generative models, can enable an attacker to create many customized generators (to create deepfakes), thereby substantially increasing the threat surface. We show that existing defenses fail to generalize well to such *user-customized generative models* that are publicly available today. We discuss new machine learning approaches based on content-agnostic features, and ensemble modeling to improve generalization performance against user-customized models. Second, the emergence of *vision foundation models*—machine learning models trained on broad data that can be easily adapted to several downstream tasks—can be misused by attackers to craft adversarial deepfakes that can evade existing defenses. We propose a simple adversarial attack that leverages existing foundation models to craft adversarial samples *without adding any adversarial noise*, through careful semantic manipulation of the image content. We highlight the vulnerabilities of several defenses against our attack, and explore directions leveraging advanced foundation models and adversarial training to defend against this new threat.**

***Index Terms*—deepfake image, foundation models, generative models, deepfake detection**

1. Introduction

Recent advances and applications of generative AI have catapulted this technology as the next frontier in Artificial Intelligence [1]. Generative models are a family of machine learning (ML) algorithms capable of learning a data distribution to produce new *synthetic* variations of these data. Generative models can produce convincing synthetic images, which can be easily misused, raising several security threats. Easily available, off-the-shelf generative models (e.g., Stable Diffusion [2], DALL·E [3], StyleCLIP [4]) can be used to create synthetic or *deepfake* imagery to power large-scale fake account campaigns on social media platforms [5], create media for convincing fake news articles [6], create fake pornographic images [7], spoof identity verification in financial services [8], [9], and power other threats. Countries across the globe are struggling to respond to the risks posed by generative AI, as false alarms raised by poorly implemented defenses can completely erode our trust in online content [10].

The urgency of this problem triggered a flurry of research efforts that proposed methods to detect deepfake images [11]–[17]. State-of-the-art (SOTA) detection schemes use a supervised learning scheme that leverages "imperfections" in fake images, to distinguish fake from real images. They do so using a variety of methods (Section 3.2), e.g., using texture statistics [13], finding imperfections in the frequency spectrum [12] or local patches [17]. All these defenses claim extremely high detection accuracy on the datasets they were evaluated on (Section 3.2).

New threat vectors. In this work, we argue that *these defenses face a rapidly evolving threat landscape*, placing them at severe risk of underperforming in the real world. This evolving threat landscape is fueled by two recent advances in machine learning:

- **Emergence of lightweight methods that allow users to customize large generative models**, thereby enabling democratization of generative AI technologies. Prior defenses were primarily evaluated using images from a few instances of generative models from different families, mostly GANs [18], [19] and Diffusion models [2]. Today, the threat landscape has changed dramatically—e.g., there are over 3,000 user-customized[1] (i.e., by Internet users) variants of the Stable Diffusion [2] model alone on platforms like CivitAI [20] and Huggingface [21]. This is enabled by new algorithmic methods that enable efficient fine-tuning of these large generative models in resource-constrained setups (limited training data and compute power) [22]. Previous defense efforts have documented the challenges of achieving high generalization performance across a few generative models and families [12]. The current threat landscape that contains thousands of publicly available variants of generative models, where any can be misused, presents an unprecedented and challenging environment for defenses.

1. As of Feb, 2024

(a) Source image (b) Manipulated image

Figure 1: Adding lipstick in the manipulated image evades a deepfake detector [12].

- **Emergence of vision foundation models.** Foundation models are ML models trained on broad data (usually using self-supervision), which can be further adapted for a variety of downstream tasks with impressive performance [23]. Popular vision foundation models include CLIP-ResNet [24], EfficientNet [25] and ViT [26]. For example, the CLIP model can learn a joint embedding space for both text and image inputs and can function as a generic text and image encoder. It can then be further adapted to build a variety of downstream tasks, e.g., to build computer vision classifiers. We show that such foundation models can be easily integrated with existing generative models to craft *"adversarial fake images"*, i.e., fake images that can fool deepfake classifiers.

Contributions. We conduct the first large-scale study analyzing 8 SOTA deepfake defenses by considering the above evolving threats. Our key contributions are as follows:

- *We provide a critique of the methods used to train and evaluate existing defenses (Section 4).* We tried to reproduce the findings of 8 state-of-the-art defenses, and in the process identified several issues related to the training and evaluation methodologies used in these works. For example, the most recent defense, UnivCLIP [11] does not control the content and quality of images used in the training and evaluation dataset, leading to possible spurious correlations being learned. This can result in overestimating the performance of these defenses, resulting in misplaced confidence in the strengths of these defenses in the real world. We hope to educate the community about these issues and suggest actionable steps to correctly train and evaluate defenses.
- *We study defense effectiveness in a threat landscape enabled by the democratization of generative AI technologies (Section 5.1).* We study such a setting by focusing on the many publicly available, user-customized variants of the Stable Diffusion model, particularly 16 of them. All defenses exhibit significant degradation in performance when applied to these user-customized variants, on average up to 53.92% degradation in Recall (over the 16 models). Two notable recent defenses, UnivCLIP and DEFAKE [27] that are state-of-the-art in terms of generalization performance, also show significant degradation in performance. We present new strategies to further improve generalization performance. This includes augmenting existing defenses with *content-agnostic* features, and using ensemble models that combine defenses using foundation model features and domain-specific features (e.g., frequency-based features [12]).
- *We study defenses against an adversary who leverages vision foundation models to create adversarial samples without adding adversarial noise (Section 5.2).* It is important to study the adversarial robustness of deepfake defenses. Existing work has focused mainly on adding adversarial noise (perturbations) to fake images to evade detection [28], [29]. This can degrade image quality, especially when they appear in regions with a smooth texture [30]. *We show that foundation models can be leveraged to successfully evade defenses without adding any noise to the images.* Our key idea is to leverage a foundation model to create an adversarial fake image by making careful *semantic changes* to the image content. For example, Figure 1 shows a successful attack, where manipulating the lip color is sufficient to fool an existing detector [12]. To achieve this, the attacker uses a text prompt (that describes the semantic change) on an adversarially updated image generator. The image generator is adversarially updated using a *surrogate deepfake classifier powered by a foundation model*. Our attack can significantly degrade the performance of all 8 defenses. We identify defenses that are notably weak and resilient against our attacks—defenses using frequency-based features are highly vulnerable to our attack, while defenses using a foundation model themselves are more resilient. We also explore two strategies to improve adversarial resilience: (1) defenses that use more powerful foundation models (i.e., pretrained on larger datasets), compared to the foundation model used by the attacker, demonstrate more resilience, and (2) adversarial training can be an effective temporary measure to build resilience.

Our contributions highlight the urgent need to rethink defenses in a setting where the adversary can customize and create their own deepfake generators and incorporate powerful foundation models to create evasive deepfakes. Section 7 discusses several directions for future work. *Code and data used in this study are available on Github.*[2]

2. Background and Threat Model

Deepfakes and real images. We use the term "deepfakes" or "fake images" to refer to *fully synthetic images* created using *generative models*. The deepfake images can have any type of content, i.e., we *do not* place restrictions on the type of content such as only faces. Partially synthetic deepfakes, such as face-swaps or face-reenactments are not considered in this work [31]. Today, state-of-the-art generative models for images are based on GANs [32], Variational Autoencoders (VAEs) [33], and Diffusion models [2], [34]. Recently, the integration of multimodal vision-language models into image generation pipelines has enabled individuals to create an image by just supplying a prompt that describes the desired content [2], [24], [35], [36]. This increased

2. github.com/secml-lab-vt/EvolvingThreat-DeepfakeImageDetect

ease of creating synthetic content can be misused to create misleading content.

We use the term "real image" to refer to any image that is not produced by a generative model. This includes images produced through photography, human-made digital, or any content created by humans. Note that prior work primarily considered camera images to be real images. In contrast, our definition encompasses a much wider class of images in the real set. Given the capabilities of generative models to create digital art and other types of content [37], it is important to broaden the definition.

Foundation models. In this work, we use foundation models as general-purpose feature extractors. These models are usually trained on Internet-scale or large datasets of image and text modalities, mostly using self-supervised learning strategies—a process known as pretraining. These models learn highly generalizable representations of the different input modalities, e.g., text or images, making them highly adaptable for various downstream tasks, e.g., image or text classification. An image classifier can be built using the features extracted from a foundation model. For example, ViT [26] is a transformer-based foundation model trained on 14M images from the ImageNet-21K [38] dataset, and achieves excellent performance on challenging tasks, e.g., the VTAB [39] suite of 19 tasks such as image classification, object detection, and localization. We used multiple vision foundation models, namely, EfficientNet [25], ViT, and models from the CLIP and OpenCLIP family [24], [35]. Due to their pretraining, we can extract highly effective features to build deepfake classifiers.

Threat model. The attacker uses a generative model to create convincing and high-quality deepfake images that capture a desired target content. Any type of content can be generated. To enable the generation of desired content, we consider the use of text-to-image generative models, namely Stable Diffusion (SD) [2] and StyleCLIP [4]. For SD, the attacker starts with only a text prompt that describes the desired content. For StyleCLIP, the attacker uses both a text prompt and a source image. StyleCLIP translates the source image into a target image that captures the content described in the text prompt.

The defender aims to distinguish fake images from real images using a supervised machine learning (ML) model. We consider 8 publicly available state-of-the-art (SOTA) ML schemes (in the research literature) to detect fake images. In Section 5.2, we further consider a defender who is aware of the generative model used by the attacker and optimizes the defense to detect images from that generative model.

The attacker may adapt the generator to produce *adversarial* fake images that can fool detectors, while preserving the desired quality and content of the image. This is done *without* adding any adversarial noise to the generated image, i.e., the generated image itself is adversarial by design. Instead of adding adversarial noise, the attacker makes adversarial semantic changes to the content, using a text prompt and an adversarially updated image generator. We consider a full black-box setting where the attacker has no query access or access to the defender's detection model.

3. Generative Models and Defenses

3.1. Generative Models to Create Deepfakes

Generative models learn the underlying patterns in the training data to generate novel content. We focus on two popular generative models —Stable Diffusion (SD) [2] and StyleCLIP [4]. Both models are capable of producing high-quality imagery, are open-source, and the pretrained models are publicly available, allowing us to study different attacks.

Stable Diffusion (SD). We use the SD model to study the impact of democratization of AI technologies on defense efforts. The SD model is widely popular, with users "customizing" or fine-tuning this model on datasets to further improve image quality or adapt to newer data distributions and share the models publicly. We see over $3,000$ user-customized variants of these models being shared on CivitAI [20] and HuggingFace [21].

SD is implemented as a text-to-image generation model based on the Latent Diffusion Model (LDM) [2]. At its core, SD acts as a denoiser. Starting from a noise vector (e.g., Gaussian noise), SD can transform it into a complex target distribution (an image) conditioned on text prompts through a series of invertible operations to generate high-quality images. To reduce the computational demands, SD implements this diffusion process in a low-dimensional latent space, instead of the pixel space. The input text prompt is encoded using the CLIP [24] text encoder, which has learned a joint language-image embedding space.

StyleCLIP. We use the StyleCLIP model to study the adversarial robustness of existing defenses. StyleCLIP is a text-driven image modification model, i.e., given a source image and a text prompt describing the target content, StyleCLIP manipulates the source image to capture the desired target content. For example, given a face image, StyleCLIP can manipulate facial attributes (e.g., hair, eyes). StyleCLIP uses StyleGAN2 [40] (GAN family) as the image generator, and the OpenAI CLIP model [24] to perform text-driven image modifications. Generative Adversarial Networks (GANs) have been considered state-of-the-art for image generation for almost a decade [32], [41]. A GAN model includes a generator and a discriminator, which are trained adversarially. The generator aims to generate fake images that can fool the discriminator, and the discriminator's classification feedback is used by the generator to improve its quality of image generation. Being a GAN model, it is a perfect fit for our setting—we adversarially update the StyleGAN2 generator of StyleCLIP to create adversarial fake images.

Image modifications are enabled by manipulations to an intermediate latent space in StyleGAN2. StyleCLIP uses the intermediate latent space called Stylespace S. Each stylespace vector contains channels that are disentangled latent representations of the color and semantics of an input image. Using the target text prompt, encoded to a joint text

and image embedding space, StyleCLIP infers a direction in the stylespace to drive the manipulations, i.e., identifies which channels in the stylespace should be manipulated to satisfy the target text prompt. There are two key parameters at generation time: (1) β: stylespace channels with relevance score higher than β are manipulated. (2) α: controls the strength of manipulations made to a stylespace channel.

Other models. We considered other generative models but did not include them for one or more of the following reasons: unavailability of training code or pretrained checkpoints and poor quality imagery. Details are in Appendix A.

3.2. Defenses: Deepfake Detection Schemes

We select 8 supervised learning-based defenses using the following criteria: *(1) Performance:* All 8 defenses claim impressive detection performance and were examined in previous work on deepfake defenses. *(2) Availability:* The availability of model checkpoints and training code is a requirement for our methodology, as we need to fine-tune these defenses on different datasets. *(3) Target deepfakes:* We only study defenses designed to detect fully synthetic images. Defenses made to detect partially synthetic content (e.g., face-swapped content [42]) are not our focus. *(4) Content types:* Previous work mainly focused on detecting face deepfakes. We do not place any content restrictions, given the emerging threat that any arbitrary content can be a deepfake. In addition to face images, we consider several content types, e.g., artwork, illustrations, images of different objects. *(5) Diverse methodologies:* The chosen defenses use diverse methodologies, thus helping to understand the robustness of different defense strategies.

UnivCLIP [11]. This is the most recent defense (in 2023) with two key highlights: *First*, UnivCLIP is one of the first defenses that uses a large foundation model to build a deepfake detector. The CLIP:ViT-L/14 foundation model [24], trained on 400M image-text pairs, is used. After extracting features from the *(frozen)* CLIP:ViT model, the study recommends using either a nearest neighbor classifier or a linear classification layer, with further training to predict an image as real or fake. We use the linear classifier approach as it performs better in our setting (Section 4). *Second*, authors claim that extracting features from a foundation model, which has not been explicitly trained for a deepfake detection task, provides (surprisingly) high generalization performance. UnivCLIP is shown to achieve up to 99.17% Average Precision in generalizing to fake images from (previously unseen during training) generative models.

DE-FAKE [27]. Similar to UnivCLIP, Sha et al. [27] also use the CLIP [24] model to build a detector. Compared to UnivCLIP, a key difference is to augment the image's embedding along with an embedding of the text prompt (both extracted using CLIP) to train the detector. The intuition is that real images usually have more information than their respective captions, whereas fake images generated from prompts only show content that is specific to that prompt. This disparity in information is used to detect deepfakes. The classifier can effectively generalize to deepfakes from models not seen during training, e.g., achieving an accuracy of 90.9% on DALL·E 2 images. As recommended by the authors, we use the image captioning model BLIP [43] to generate the prompts for training. This fits our threat model as we consider a defender who will only receive an image for detection, i.e., without an associated prompt.

DCT [12]. Ricker et al. [12] show that the frequency domain provides discriminatory features for deepfake detection. This is inspired by previous work showing visible artifacts, e.g., grid-like patterns, in the frequency spectrum of GAN-generated images [44]. To build the classifier, the frequency domain features are extracted from the images using a discrete cosine transform (DCT). We use the log-scaled version of the DCT features as recommended by Rocker et al. [12] for improved performance. These features are used to train a Logistic Regression classifier. DCT achieves 97.7% and 73% accuracy on images generated by GAN and Diffusion model, respectively.

Patch-Forensics [17]. This detector is designed only for face content and claims high generalization performance across generative models. The key intuition is that searching for artifacts in local patches of the image provides more generalizable patterns for detection, compared to looking for global artifacts (i.e., in the image as a whole). To identify local artifacts, an image is broken down into equal sized patches and a patch-based classifier is trained. This classifier uses a truncated convolutional neural network with small receptive fields that are better suited for identifying local imperfections. Classification decisions at the patch level are aggregated into an overall prediction. We use the Xception Block 2 variant of their patch-based classifier, which demonstrated impressive performance. Patch-Forensics is shown to achieve 100% Average Precision when applied to fake images from StyleGAN.

Gram-Net [13]. This scheme has been extensively studied to detect fake faces, but also claims to generalize to other content. The key insight is that the texture statistics of fake images (e.g., face content) are significantly different from real images. Using global texture features was found to be robust against image distortions, and also generalize across different GANs. Based on this observation, the authors propose a novel CNN-based architecture called "Gram-Net" that can extract global texture features for detection. We build on a version of their model trained on StyleGAN (fake) and CelebA-HQ (real) images, as this model was used to detect arbitrary deepfake content (i.e., not just faces). Gram-Net achieves 89.26% average accuracy when applied to images from generative models not seen during training.

Resynthesis [14]. This detection scheme also aims to generalize across different generative models. Instead of relying on low-level artifacts specific to certain generative models, this scheme aims to learn more generalizable features for detection. This is done by resynthesizing testing images (i.e., both real and fake) based on different auxiliary tasks, e.g., super-resolution, denoising and colorization. The resynthesis component is only trained on real images, and

(a) Real (b) Fake

Figure 2: Real and fake samples from our SD dataset.

therefore results in different residuals (reconstruction errors) for fake images. The key insight is that the residuals from the resynthesis tasks provide effective discriminatory features. The detector is trained jointly with the synthesizer and is shown to achieve 93.7% average accuracy.

CNN-F [15]. This is a widely studied defense. The key idea is that CNN-based generators leave detectable *fingerprints*. A ResNet-50 model (pretrained on ImageNet [45]) is fine-tuned to learn such fingerprints. A notable step of training is a data augmentation strategy based on standard post-processing schemes that shows improved generalization performance. The work highlights that the detector needs to be trained only on images from a single CNN-based generator to generalize across different fake sources. CNN-F is shown to achieve 93% mean Average Precision.

MesoNet [16]. Originally designed to detect deepfake videos, this scheme also functions as a deepfake image detector. The key idea is to conduct analysis at a *mesoscopic level*. At a macroscopic level, it is hard to identify any semantic differences between fake and real images. The authors claim that a microscopic analysis of artifacts can be challenging as well, and hence they focus on mesoscopic properties. They propose a unique DNN with a small number of layers to extract mesoscopic features. They also note that replacing convolution layers with Inception modules [46] leads to better classification results. MesoNet claims to achieve 98.4% accuracy.

3.3. Defense Implementations

To effectively evaluate the threats, we need to consider a capable defender. Existing defense implementations (pretrained models, whenever available) are trained on different datasets, thus generalizing differently to newer datasets. We need to implement (train) each defense to have the best chance of detection in our threat settings. Thus, we resort to retraining all 8 defenses on 2 datasets relevant to our research goals. *For each dataset, we strive to use images with highest visual quality (for fake and real), and also ensure a similar content distribution for fake and real classes—so the classifiers can learn effective features and not learn to separate based on differences in semantic content.*

Defense	SD			StyleCLIP		
	Precision	Recall	F1	Precision	Recall	F1
UnivCLIP	90.20	93.90	92.01	93.79	92.20	92.99
DE-FAKE	93.82	94.20	94.01	74.41	78.80	76.54
DCT	100	88.80	94.07	100	99.60	99.80
Patch-Forensics	-	-	-	91.76	91.30	91.53
Gram-Net	99.99	99.10	99.55	99.99	99.60	99.80
Resynthesis	85.39	86.50	85.94	98.80	98.70	98.75
CNN-F	99.41	83.80	90.94	99.90	97.10	98.48
MesoNet	99.99	98	98.98	96.70	99.50	98.08

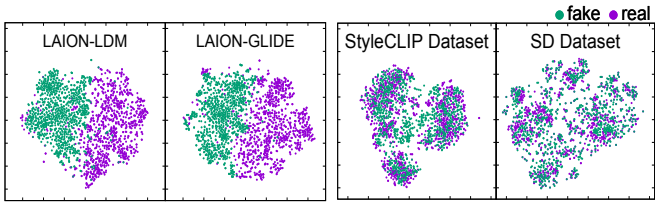
TABLE 1: Performance of defenses, reported as Precision, Recall and F1 scores (**in percentage**) for the fake class of test set, after our training / fine-tuning. Patch-Forensics is marked as '-' as it is only applicable to face content, and therefore not evaluated on SD images.

Training/evaluation dataset 1: SD dataset. This dataset is used to investigate defense effectiveness on user-customized generative models (Section 5.1). *Real images* are sampled from the LAION-AESTHETICS dataset [47]. This dataset contains 625K image-text pairs of highest visual quality, as rated by the LAION-Aesthetics Predictor V2 [48] model. *The images cover a wide variety of content, e.g., people, nature, objects, illustrations, and digital art.* Fake images are generated with a square aspect ratio. So, we further filter to extract image-text pairs, where the width/height of the images are roughly 500 pixels (with an error margin of 150 pixels), and also remove images flagged as unsafe (based on available metadata). Finally, we randomly sample a subset from this filtered set. *Fake images* are created using the Realistic Vision v1.4 SD model [49], with text prompts obtained from the real dataset (comes as image-text pairs). This ensures similarity in content between the fake and real classes. The Realistic Vision model was created by fine-tuning the SDv1.5 model to enhance image quality. It is widely used with over half a million downloads. We use this SD model instead of the base SD models (e.g., SDv1.5 [50]) because of its ability to generate higher quality, realistic, aesthetic images. Images are generated as 512×512 using the default settings. Figure 2 shows examples of fake and real images from this dataset. In total, we collected a *balanced* dataset of $16K$, $2K$, and $2K$ images across both classes for training, validation, and testing, respectively.

Training/evaluation dataset 2: StyleCLIP dataset. This dataset is used to train defenses to study robustness against an adversary leveraging vision foundation models (Section 5.2). The attacker uses the StyleCLIP generator which is based on StyleGAN2. We only use images with face content, as StyleCLIP only produces limited content types and is widely used for face content. *Real images* are sampled from the Flickr-Faces-HQ (FFHQ) dataset [18], which has $70K$ high-quality, high-resolution (1024×1024) face images. We randomly sample a subset from this dataset. *Fake images* are randomly sampled from the official repository of StyleGAN2 generated images. We use fake images generated

(a) Real (b) Fake

Figure 3: Real and fake images used by the UnivCLIP defense. Note the poor visual quality of the fake sample.



(a) UnivCLIP dataset (b) Our dataset

Figure 4: TSNE plots of real and fake images used in UnivCLIP defense (left) and our datasets (right). Fake and real images in the UnivCLIP dataset are easier to separate as they are not controlled for content and quality.

using the truncation parameter $\psi = 1.0$ which ensures maximum diversity in face images. In total, we collected a balanced set of $16K$, $2K$ and $2K$ images across both classes for training, validation, and testing, respectively.

Training configuration. We made extensive efforts to limit overfitting and create high-performing versions of each defense. Two versions were developed for each defense, with training conducted on the SD and SyleCLIP datasets, respectively. For all defenses, except DCT, we fine-tuned the models starting from the checkpoints provided by the model creators. The DCT model is not a good fit for fine-tuning, as it uses a simple logistic regression classifier, plus the authors did not provide a pretrained model. For 6 of the 8 defenses, the hyperparameters recommended by the model creators did not yield high enough validation performance. For these 6 defenses, we further tuned the hyperparameters for optimal performance. Appendix A provides more details on training configurations.

Table 1 shows the detection performance of defenses after our training/fine-tuning on our two datasets. We report the F1 score for the fake class. *Note the high F1 scores in most cases.* Patch-Forensics is only applicable to face content and is therefore omitted for evaluation on the SD dataset.

4. A Critique of Existing Defense Efforts

We highlight three key limitations of existing defenses.

Limitation 1: Lack of control for content and image quality. *Controlling content:* Training datasets should have a similar content distribution between the fake and real classes. For example, suppose that the real class consists of only car images and fake class consists of only face images. In this case, the model can learn spurious features that are not relevant to the task of distinguishing real and fake images. The model can learn to differentiate between cars and faces, rather than learning features that indicate authenticity of the images. Given the availability of text-to-image models, content can be controlled by using the caption of a real image as a prompt to generate a fake image with similar content. We find that the most recent defense, UnivCLIP, which is SOTA in terms of generalization performance, does not appear to control the content distribution (i.e., their training methodology does not discuss this aspect). This is problematic. All other defenses make some effort to control the distribution of content across the two classes.

Image quality: The common premise of the threat of deepfakes is that fake images can appear convincingly real and therefore mislead viewers. Therefore, we encourage the focus on training and evaluating high-quality fake (and real) imagery. Note that training a detector on a dataset where the real images are of high visual quality *but* the fake images are of low quality can lead to a classifier that may not detect deepfakes of high quality in the wild. We observe this for UnivCLIP, where the fake images used for training and evaluation are of relatively lower visual quality compared to the real images. Figure 3 shows an example of fake and real images used by UnivCLIP. This fake sample was generated using the LDM [2] model and is of low visual quality.

To understand the impact of *not* controlling the content and quality of images, we take a closer look at the UnivCLIP defense. Recall that UnivCLIP uses the CLIP-ViT foundation model to extract features for detection. Figure 4 shows a TSNE visualization to capture how well UnivCLIP can separate real and fake images using features extracted from CLIP-ViT. The plots on the left show the results on the (problematic) datasets used in the original work (UnivCLIP), i.e., fake images from the LDM [2] and GLIDE [51] generative models, and real images from the LAION-400M [52] dataset. We find that LDM and GLIDE images have lower visual quality compared to our SD and StyleCLIP datasets.[3] UnivCLIP is able to almost perfectly separate these classes in this case. However, when we visualize the foundation model features on our SD and StyleCLIP datasets (plots on the right), we observe a drastically different result — features from CLIP-ViT being unable to cleanly separate these classes. This suggests that only using foundation model features, without a linear classification layer, will not generalize well—which invalidates the fundamental claim made by UnivCLIP that such features with a nearest neighbor search is sufficient for high generalization performance.

The high generalization performance of UnivCLIP in the original work can be attributed to the lack of control of con-

3. The Kernel Inception Distance (KID) between real and fake images for the LDM-200 model and GLIDE used by UnivCLIP is 0.023 and 0.017, respectively, while KID for our SD dataset is 0.008, i.e., an order of magnitude lower. Lower KID indicates better fake image quality.

tent and quality of the images, which is likely overestimating its real-world performance—it is easier to separate fake images with heavy artifacts and different content distribution from real images. However, for the rest of our evaluation, we fixed these problems that plague UnivCLIP, retraining it on our SD and StyleCLIP datasets, where we control both quality and content of images. The trainable linear classification layer in our implementation helps to better discriminate between fake and real images.

Limitation 2: Lack of adversarial evaluation. An effective defense should be robust against an adaptive adversary. In a practical setting, this would be an attacker in a black-box setting (with no access to defense internals) who crafts "adversarial" fake images that evade detection (i.e., classified as real). Such an attacker may also exploit some knowledge of the defense, e.g., the defense uses frequency domain features or texture features. We find that existing work is severely lacking in this respect. *DCT and UnivCLIP do not conduct any evaluation in adversarial settings.* Among the remaining defenses, all except Patch-Forensics and DE-FAKE, only conduct a basic robustness evaluation. This includes basic image manipulation schemes such as blurring, JPEG compression, downsampling, and noising/denoising. Patch-Forensics studies an adaptive attack, but uses a white-box setting, where the attacker has full access to the defense internals. This is not the most practical setting because the defense internals may not be publicly released. Finally, DE-FAKE studies both white- and black-box attacks that show significant degradation in their detection performance.

Limitation 3: Restricted image content types. Prior work focused only on limited content types, e.g., faces, animals, bedrooms, and buildings. Given the proliferation of text-to-image models, an attacker can generate fake images that capture *any* type of content using a text prompt. We encourage the community to consider a broad range of content to study deepfake defenses. This can also present new technical challenges, as some defenses are designed only for faces [17]. Newer datasets, such as LAION-5B [53], have image-text pairs that include photographs, artistic paintings and illustrations covering a broad range of semantic content. Text captions from such datasets can be used to generate fake images that cover a wider range of content.

5. Defenses Against Evolving Threats

5.1. Evolving Threat 1: User-Customized Generative Models

Our goal here is to understand the effectiveness of existing defenses in a threat landscape enabled by the democratization of generative AI technologies. The open source release of the Stable Diffusion model and the development of new low-cost generative model fine-tuning approaches have resulted in thousands of user-customized (i.e., by Internet users) SD model variants on platforms like CivitAI and HuggingFace. Internet users are creating custom checkpoints of the base SD model for various reasons—to enhance image quality, realism, adapt to a new image dataset or to change the style of images. We carefully choose representative user-customized models from this pool to evaluate the generalization performance of existing defenses on them. Note that we do not study the StyleCLIP model in this section due to its lack of widespread user-customization, which is a requirement for this analysis.

5.1.1. Collecting user-customized SD models. The traditional approach to create a custom model is to fine-tune the base model on a new dataset by updating all parameters (layers) of the model—known as *Full Model fine-tuning (FM)*. However, this is computationally expensive and requires high-end GPU hardware to fine-tune SD on large datasets. Recently, a novel approach called *Low-Rank Adaptation (LoRA)* was proposed to enable low-cost fine-tuning of generative models [22]. For SD, LoRA is applied by adding a small number of trainable parameters to the cross-attention layers, which extracts correlations between images and text prompts. The rest of the SD model is kept frozen during fine-tuning. Using LoRA results in faster training time, requires less compute (can be run on consumer-level GPUs, e.g., NVIDIA 2080TI), and the trained weights produce smaller files (order of MBs, compared to GBs for FM fine-tuning) [54].

We only choose custom models that enhance the aesthetics of the images while preserving the semantic content and quality, compared to our SD dataset (training dataset for our defenses). It would be unfair to expect the defenses to generalize if the custom checkpoints entirely change the domain/content of the images seen during training. We choose a total of 16 custom checkpoints, 8 based on LoRA fine-tuning, and 8 based on FM fine-tuning. Users fine-tuned all checkpoints from the SDv1.5 base model. We apply the LoRA weights using the Diffusers library [55] with $\alpha = 0.5$, which is a scaling factor for the LoRA weights.[4] Table 6 (Appendix B) presents the details of each checkpoint. For example, there are checkpoints that enhance brightness, details, sharpness, contrast, reduces noise, and increases realism. Figure 5 shows samples of images from both LoRA and FM checkpoints. More samples are in Figure 9 (Appendix). For each of the 16 custom models, we generate $1K$ images using text prompts from our SD test dataset. These images serve as the testing set to measure generalization performance of existing defenses.

Before we proceed with evaluation of the defenses, we test whether the user-customized models indeed meet our requirements, i.e., preserving the semantic content and not degrading image quality. We use the following two metrics that have been shown to correlate with human judgement [56]:

(1) *Semantic similarity:* We use the metric **CLIP-Score** [57]. CLIP-Score measures the cosine similarity between an image and its text prompt using the CLIP-ViT-B/32 model [26], i.e., how well does an image capture its

4. A lower scaling factor ensures that the semantic content is mostly preserved.

Figure 5: Image samples for the caption "Dawn at a jetty in Glenorchy, New Zealand." From left to right, first 2 images are real and fake images from our SD test set, the next 3 are LoRA images, followed by 3 FM images. Model IDs are explained in Table 6 (Appendix B). We can see content preservation with comparable quality across all samples.

text prompt? The values range between 0 and 1, with higher values indicating better semantic similarity. We compute the CLIP-Score for each generated image using its associated text prompts (from our SD dataset). As a baseline for comparison, we calculate the CLIP-Score for fake images in our SD dataset. For the baseline images (from Realistic Vision), we obtain an Avg. CLIP-Score of 0.82. The user-customized models have a similar CLIP-score value as well—Avg. CLIP-score values across LoRA and FM models are 0.81 and 0.81, respectively. This suggests that user-customized models preserve the semantics of the content.

(2) *Image quality*: We use the **Kernel Inception Distance (KID)** [58] metric to measure synthetic image quality.[5] KID measures the distribution distance between the real and fake image sets. The values are unbounded, and smaller values (closer to zero) indicate better synthetic image quality, i.e., fake images match the distribution of the real images. For the baseline images from our SD dataset, we have a KID score of 0.008, and obtain an Avg. KID score of 0.006 and 0.008 for LoRA and FM models, respectively. The KID values are small (close to zero) and close to the baseline results, suggesting that these models are not degrading image quality.

5.1.2. Defense generalization on user-customized SD models. We use the version of defenses trained on our SD dataset (Section 3.3) for this evaluation. To measure defense performance, we compute ΔR, calculated as the *percentage degradation in Recall for the fake class*, compared to the baseline test performance of a defense.[6] In other words, $\Delta R = \frac{R1-R2}{R1}$, where $R1$ is the Recall on our SD test set, and $R2$ is the Recall when applied to images from a user-customized SD model. The baseline Recall numbers ($R1$) are in Table 1. *A high ΔR would indicate that a defense performs poorly when applied to user-customized models.* Figure 6 shows the results. The red line in each plot represents the average Recall across all user-customized model variants for the corresponding defense.

FINDING 1. ***All** defenses exhibit significant degradation in performance.* The average ΔR for defenses (across all the user-customized models) ranges from 19.69% to 53.92%. This performance degradation highlights the urgent need to develop new defenses or enhance existing defenses to improve generalization performance.

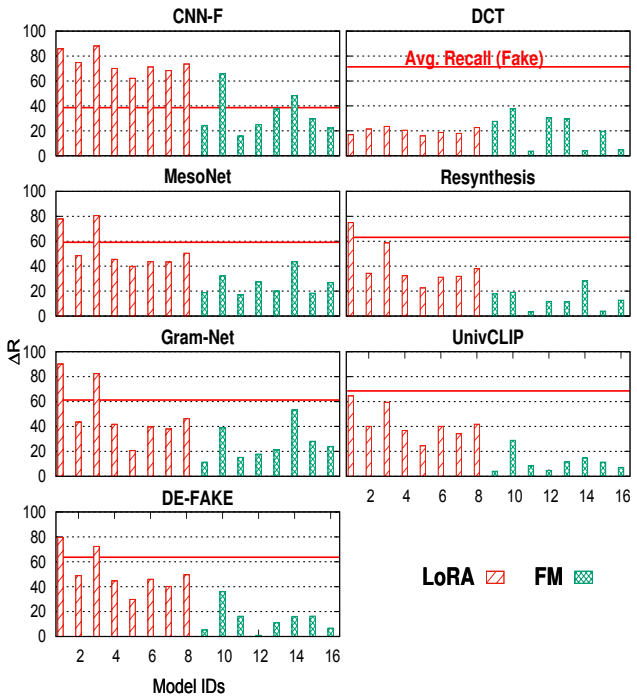


Figure 6: Generalization performance of defenses measured using ΔR. Model IDs 1 to 8 are LoRA models, and 9 to 16 are FM models. See Table 6 (Appendix B) for details. Red line in each plot shows the average Recall (fake class) over all 16 SD custom checkpoints.

FINDING 2. *Solely relying on features from a foundation model is insufficient to achieve high generalization.* UnivCLIP and DE-FAKE claim that using features from a foundation model (CLIP-ViT) is sufficient to achieve high generalization performance. However, our evaluation does not support this claim. Figure 6 shows that both UnivCLIP and DE-FAKE demonstrate significant performance degradation. For LoRA models, we observe an average ΔR of 42.66% and 51.35% for UnivCLIP and DE-FAKE, respectively. For FM models, while the degradation is lower

5. Our dataset has 1k images each for fake and real. We use KID instead of FID [59], because FID provides reliable estimates for only much larger datasets [60].

6. We do not consider the Precision metric because no real images are required for this analysis.

compared to LoRA, we see ΔR up to 28.87% and 35.99% for UnivCLIP and DE-FAKE, respectively. LoRA is a dominant low-cost fine-tuning strategy among Internet users, so this presents a real threat. Given our finding, one may wonder how UnivCLIP demonstrated high generalization across diverse generative models in their work. We suspect that this can be attributed to their evaluation data sets that do not control the content and quality of the images (see analysis in Section 4).

FINDING 3. *Frequency domain features show the best generalization performance.* DCT, a defense based on the frequency domain features shows the most promise. It shows an average ΔR (across all models) of 19.69%, which is the lowest ΔR among all defenses. This is because the frequency spectrum artifacts found in the LoRA and FM model images resemble those observed in images encountered during defense training (SD fake images), whereas real images do not exhibit such artifacts. We visualize the frequency spectrum of real and fake images with further explanation in Appendix E.

FINDING 4. *CNN-based defenses show the worst generalization performance.* Note that the CNN-F and MesoNet and Gram-Net defenses show high F1 scores ($> 90\%$) on our SD dataset. However, CNN-based defenses, even with specialized architectures, are unable to learn generalizable features. CNN-F, MesoNet and Gram-Net, all based on CNNs, exhibit the highest average ΔR (across all the user-customized models), ranging from 38.26% to 53.92%. These schemes claim to use data augmentation strategies to help with generalization. Nevertheless, data augmentation techniques are unable to compensate for features that do not generalize effectively.

5.1.3. Improving generalization using content-agnostic features. Even though user-customized models preserve high-level image semantics, there may be distributional differences in low-level content. Ideally, defenses should focus on imperfections of the fake images, and not be derailed by changes in the content distribution. Based on this insight, we investigate whether enhancing an existing defense with "content-agnostic" features would improve generalization.

We enhance the DCT defense as it shows the most promise for generalization. Prior work has shown that the noise residual or the "noise space" of an image, i.e., the residual image after removing all the content, can contain effective discriminatory patterns for deepfake detection [61]. [7] Inspired by this work, we use a SOTA denoising scheme, MM-BSN [62], to extract noise residuals for each image in our SD dataset. From these noise residuals, we extract the log-scaled DCT features, which serve as our content-agnostic features. We then enhance our DCT scheme by concatenating existing DCT features (calculated over the entire content space) with our content-agnostic features and retrain the DCT scheme.

7. Many of the defenses we study (from Section 3.2) claims to outperform schemes that rely only on noise residuals.

FINDING 5. *Content-agnostic features can help boost generalization performance for deepfake detection.* The enhanced DCT scheme with content-agnostic features shows improvement, i.e., there is a reduction in ΔR for 12 out of the 16 user-customized models. The average ΔR over all FM models reduces from 19.68% to 14.29%. For LoRA models, we see a lower improvement–average ΔR reduces from 19.69% to 19.21%. Content-agnostic features are promising. With better noise residual extraction schemes, combined with better learning schemes, generalization can be potentially further improved.

5.1.4. Improving generalization using ensemble approaches. Since most defenses use different methods/architectures for their defenses, it is possible that defenses are learning different artifacts to separate real from fake. This can be leveraged using an *ensemble approach* to improve generalization. We again build on the most effective DCT defense. We combine DCT with each of the remaining 6 defenses, and flag an image as fake if any one method predicts it to be fake. Of course, such an ensemble scheme can degrade Precision, i.e., increase false positives. Therefore, for this analysis, we report Precision, Recall and F1 score metrics for the fake class.

FINDING 6. *Combining domain-specific features (i.e., features known to identify imperfections in fake images) with features from a foundation model improves generalization.* Table 2 shows the Average Precision, Recall and F1 scores (for the fake class) over all the 16 user-customized models for each defense variation. The top row shows the performance of the DCT scheme without an ensemble approach. Combining DCT with UnivCLIP shows the largest improvement in F1 score—up to 86.25%. Recall that only using foundation model features does not improve generalization performance (Finding 2). However, foundation model features in conjunction with domain-specific features (i.e., frequency features) can improve generalization performance. Also note that it is easier to build an ensemble defense when we are extracting features from a pretrained model (i.e., foundation model), compared to a defense like Gram-Net (which also comes close in performance). The DE-FAKE defense which also uses features from a foundation model, performs similar to DCT+UnivCLIP. We also explored combining the second-best defense (Figure 6), UnivCLIP with other defenses (other than DCT) to find better ensembles. We did not see any other combination that achieves an F1 score higher than DCT+UnivCLIP, which confirms the supremacy of combining both types of defenses. Detailed results are in Appendix D.

5.2. Evolving Threat 2: Adversaries Leveraging Vision Foundation Models

5.2.1. Using foundation models to craft adversarial fake images. We propose a simple black-box attack to craft "adversarial" fake images by leveraging vision foundation models. We assume that the attacker has already created a fake image that is deceptive or damaging, but it can be

Defenses	Precision (%)	Recall (%)	F1 (%)
DCT	84.74	71.32	77.27
DCT + Gram-Net	86.63	86.01	86.21
DCT + Resynthesis	78	87.13	82.26
DCT + CNN-F	85.64	77.64	81.29
DCT + MesoNet	83.70	85.63	84.57
DCT + UnivCLIP	82.97	89.95	86.25
DCT + DE-FAKE	83.26	89.46	86.16

TABLE 2: Improving generalization performance by creating an ensemble model by combining DCT defense with the other 6 defenses. We show the average scores across the 16 SD custom checkpoints.

caught by a deepfake detector. To bypass detection or to create an adversarial version of this fake image, the attacker chooses a semantic property of this image for adversarial manipulation that still preserves their goal of deception. For example, a face image created to build a fake social media profile is misclassified as real because the facial expression has undergone adversarial manipulation. Here, the main requirement for the attack may be to have a realistic looking profile picture, which is still achieved, despite the manipulation. The semantic property to be manipulated is expressed using a ***text prompt***. Note that our approach does not add any adversarial noise and only adversarially manipulates the content to match the content described by the text prompt, e.g., a prompt that says "a smiling face" should craft an adversarial fake image with a smiling face.

Our idea is to adversarially update the weights of a fake image generator, guided by a surrogate deepfake classifier that is implemented using a foundation model. Such an adversarially trained generator can produce adversarial fake images. We use foundation models to build our surrogate classifier for the following reasons: (1) Previous work shows that foundation model features are effective for deepfake detection, e.g.,UnivCLIP and DE-FAKE (also see discussion in Section 2). (2) Foundation models being pretrained models, provide a ready-to-use model to easily build a deepfake classifier. (3) Widespread availability of diverse foundation models, provides several options for the attacker to instantiate different types of surrogate deepfake classifiers.

We demonstrate our attack using the StyleCLIP generative model (see Section 3.1) on face content. StyleCLIP uses the StyleGAN2 image generator, and we focus on face images. *Note that this is a generic attack and does not use any specific properties of the StyleCLIP (or StyleGAN2) model*. For this attack, we use 3 foundation models that are image encoders: (1) EfficientNet [25]: 5.3M parameters, trained on 14M images, (2) ViT [26]: 86M parameters, trained on 14M images, (3) CLIP-ResNet [24]: 623M parameters, trained on 400M images. Note that EfficientNet is one of the smallest foundation models, and can even run on mobile devices.

We consider an image generator $G(\theta)$ that takes as input an existing fake image x and a text prompt p (to guide generation) to generate a new image $x' = G(x, p; \theta)$. The surrogate deepfake classifier is M with an associated likelihood probability function p_M. The attacker uses the surrogate model to adversarially update the weights of $G(\theta)$ to create $G(\theta_{adv})$, such that $G(x, p; \theta_{adv}) = x_{adv}$, where x_{adv} is the adversarial fake image. The surrogate model M is frozen. The generator G is adversarially trained to minimize the following loss objective L:

$$L = \gamma * L_{cls} + \delta * L_{percept} \tag{1}$$

where L_{cls} is the classification loss, $L_{percept}$ is a perceptual loss term, and γ and δ are associated coefficients.

$$L_{cls} = \mathbb{E}_{p_{data(x')}}[l(p_M(c|x'), c)] \tag{2}$$

L_{cls} is implemented as cross-entropy loss $l(.)$ using the surrogate classifier M. c is the ground-truth label which is set to the *real* class so that the generated image is classified as *real*. The perceptual loss $L_{percept}$ is used to ensure that the image quality is not degraded. We use the intermediate representations extracted from an ImageNet-trained VGG network [63] to calculate the perceptual loss [64] using x and x'. Note that our threat model does not require the facial identity to be preserved and therefore does not include any loss objectives to preserve identity. Based on L, the generator is trained for 50 iterations per image x.

We use the $1K$ fake images from our StyleCLIP test set (Section 3.3) to create an adversarial version of each image using the above methodology. For each image, we choose a randomly chosen target prompt from a pool of 8 prompts, e.g., *a smiling face*, *a face with lipstick*, *a face with glasses*, etc. (see Appendix C). To build surrogate classifiers using the chosen foundation models (EfficientNet, ViT, CLIP-ResNet) we follow a simple methodology similar to UnivCLIP with some minor variations. To train the surrogates, the attacker can use fake images from their current generator G, and any publicly available real image datasets that capture the desired content. We used a subset of real images from the FFHQ dataset [18] with no overlap with the training set of the defenses. All 3 surrogate classifiers achieve high test F1 scores ranging from 98.15% to 98.38%. Surrogate models are trained only once with images from G and further frozen, while we adversarially update G. Details of the training configurations of the surrogates and the generator are in Appendix C.

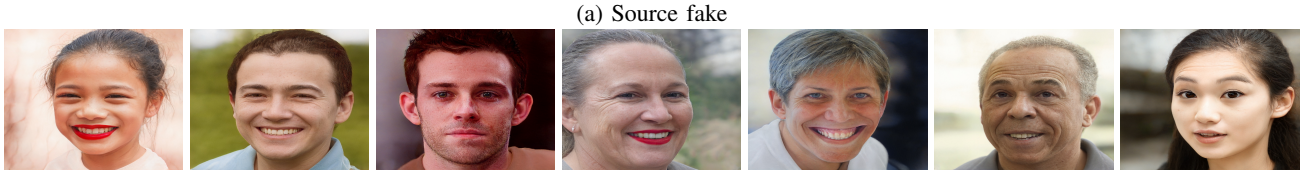
5.2.2. Computational cost of our attack. On average for all 3 surrogate models, it takes 39 seconds to generate an adversarial image using an NVIDIA A100 GPU. The cost of using such a GPU for 1 hour is \$1.1.[8] *This enables generation of around 840 images with just* \$10. Therefore, foundation models enable a viable practical approach to craft adversarial fake images.

8. https://lambdalabs.com/service/gpu-cloud#pricing

Surrogate model	ΔR for fake in %								Semantic and quality metrics	
	Gram-Net	DCT	Resynthesis	CNN-F	MesoNet	UnivCLIP	Patch-Forensics	DE-FAKE	CLIP-Score	KID_{Fake}
EfficientNet	41.67	57.43	44.58	40.06	44.82	4.56	28.37	75.76	0.675	0.007
ViT	36.04	53.41	37.39	32.23	34.67	4.66	13.47	78.43	0.675	0.005
CLIP-ResNet	47.20	88.35	73.96	70.85	76.08	12.47	40.96	80.04	0.671	0.017

TABLE 3: Evaluation of the defenses on the adversarial fake images created using our attack. We report ΔR for fake class, and for quality metrics, we report CLIP-Score and KID_{Fake}.

(a) Source fake



(b) Non-adversarial fake

(c) Adversarial fake

Figure 7: Images in top row are the source fake images from our test set that are fed to the generator along with the prompt. Middle row shows the corresponding non-adversarial fake images, and the bottom row shows adversarial fake images. From left to right in row (c), the first 2 images are from EfficientNet, next 2 are from ViT, and the rightmost 3 images are from CLIP-ResNet surrogates. We explain more in Appendix C.

5.2.3. Attack effectiveness. We consider a challenging attack setting where all 8 defenses (Section 3.2) are optimized (trained) to detect images from the attacker's generator. In other words, we use the version of the defenses trained on our StyleCLIP dataset, and the attacker adversarially updates the StyleCLIP generator.

Attack success is measured using 3 metrics:
(1) *Percentage degradation in Recall* ΔR: This is similar to the ΔR metric in Section 5.1. $\Delta R = \frac{R1-R2}{R1}$, where $R1$ is the Recall on our StyleCLIP test set (Table 1), and $R2$ is the Recall when adversarial fake images are used. A high value of ΔR indicates high attack success, i.e., more degradation in defense performance.
(2) *Measuring semantic similarity using CLIP-Score:* We use the same CLIP-Score metric from Section 5.1 to measure how well an adversarial fake image matches the desired content expressed using the prompt. A successful adversarial fake image should have a CLIP-Score that is similar to the fake image produced using the same prompt without the attack, i.e., a non-adversarial fake image.
(3) *Measuring fake image quality using KID:* KID (also used in Section 5.1) can no longer be calculated between adversarial fake images and real images, because the content has been explicitly manipulated using a text prompt. Instead, we calculate KID_{Fake} between our set of adversarial fake images and fake images produced using the same set of prompts, but without performing an adversarial attack. Smaller values of KID_{Fake} (close to zero) would indicate higher image quality. In other words, smaller values of KID would indicate that adversarial fake images are similar in quality to the fake images produced without adversarially updating the generator.

Table 3 presents the attack results.

FINDING 7. *Adversarial attacks using a foundation model can significantly degrade the performance of all defenses.* From Table 3 we see that all defenses exhibit a degradation in performance for all surrogate models, with ΔR being the highest when using the largest foundation model, CLIP-ResNet. We also see that 5 out of 8 defenses exhibit a ΔR higher than 70%. The KID_{Fake} values are small and close to zero, suggesting no significant degradation in image quality [58] for attacks. All three attacks result in adversarial fake images with CLIP-Scores in the range 0.671-0.675, which is similar to the CLIP-Score for non-adversarial fake images of 0.672 (not shown in Table 3).

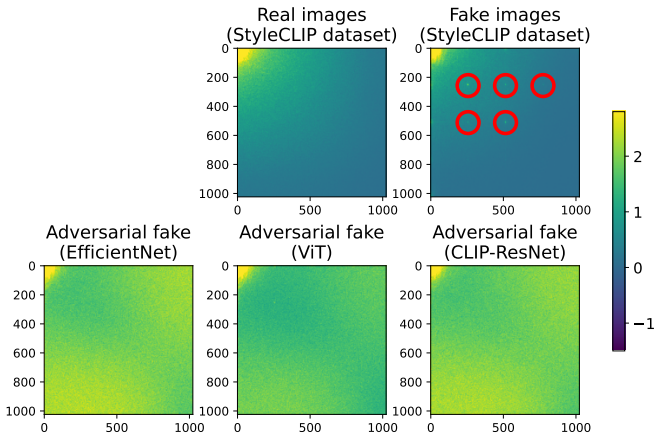


Figure 8: Average log(DCT) spectrum calculated over $1K$ images for each category. Top row shows frequency spectrum for real and fake images from our StyleCLIP dataset, and the bottom row shows adversarial fake images created using our 3 surrogate models. Red circles in "Fake images (StyleCLIP dataset)" highlight the artifacts.

It is evident from our findings that an attacker can easily exploit a publicly accessible foundation model to evade deepfake defenses *that employ a variety of techniques.* Figure 7 shows several examples of the source fake images (i.e., the one fed to the generator with the prompt), and the corresponding non-adversarial and adversarial fake images. We can see that the source images have noticeable content differences from the other sets, due to the prompt translating the source image. We further discuss the subtle content changes between the adversarial and non-adversarial fake images in Appendix C.

FINDING 8. *Defense based on frequency features is the weakest against adversarial attacks using foundation models*. Although we see DCT as the strongest defense for generalization in Sec 5.1, it is the weakest against adversarial attacks, exhibiting a high ΔR of up to 88.35%. To understand this result, we visualize the frequency spectrum of real and fake images in Figure 8. We can see that adversarial fake images mimic the frequency spectrum patterns of real images showing no clear artifacts, unlike the (source) fake images. It is fascinating that this was achieved without using a surrogate model based on frequency features. Our results further highlight the shortcomings of the frequency domain in an adversarial setting and the strength of a foundation model to thwart a variety of defenses.

FINDING 9. *Defense using foundation model features shows the most resilience*. UnivCLIP shows the highest resilience against attacks across all 3 surrogate models. UnivCLIP exhibits a ΔR less than 5% against the two smaller foundation models (EfficientNet and ViT), but starts to decline in performance when the CLIP-ResNet foundation model is used as the surrogate. This leads to our next analysis to investigate whether defenses using foundation models trained on a larger dataset can provide better resilience. Note that DE-FAKE shows significant degradation against all surrogate models. While both UnivCLIP and DE-FAKE use image features from a foundation model, DE-FAKE additionally uses text features (by automatically captioning the fake image) from a foundation model. We suspect that the poor performance of DE-FAKE can be attributed to the use of text features that may not be robust. Hence, using only image features from foundation models (similar to UnivCLIP) is a better strategy.

Surrogate model	ΔR **for fake (%) for UnivConv2B defense**
EfficientNet	0.10
ViT	0.41
CLIP-ResNet	0.10

TABLE 4: Defense performance of the UnivConv2B defense measured using ΔR against attacks using 3 different surrogate models.

5.2.4. Improving adversarial robustness of defenses using foundation models trained on larger datasets. Inspired by the UnivCLIP defense and Finding 9, we create another defense similar to UnivCLIP, but using a foundation model trained on a dataset larger than that used by CLIP-ViT (foundation model used by UnivCLIP). We train a new defense called ***UnivConv2B*** which is trained using features from the foundation model OpenCLIP-ConvNext-Large [65]. This foundation model is pretrained on 2B image-text pairs [53] and has 351M parameters. The training methodology for this new defense is similar to that of UnivCLIP. We trained UnivConv2B on our StyleCLIP dataset (Section 3.3) for 30 epochs with the Adam optimizer using a learning rate of $1e-3$. This defense achieves a high test set Precision, Recall, and F1 score of 97.9%, 97.6% and 97.75%, respectively.

FINDING 10. *Defenses using more powerful foundation models can achieve better adversarial resilience.* Table 4 shows the results of adversarial attacks against our defense UnivConv2B. There is almost no performance degradation, with ΔR less than 0.41% for the 3 surrogate models. This result raises a new question: What if the adversary leverages a more effective foundation model (i.e., pretrained on a larger dataset), compared to the defender, to build the surrogate? If a surrogate with a more effective foundation model leads to a higher attack success, then this can result in an arms race depending on who uses a better foundation model (among attacker and defender).[9]

5.2.5. Improving adversarial robustness of defenses using adversarial training. A popular strategy to build resilience against adversarial attacks is to perform *adversarial training* of the defense classifier [66]. In this case, the defense classifier is trained/fine-tuned on adversarial samples generated by the attacker. The assumption is that the

9. We are unable to implement an adversarial attack that uses the OpenCLIP-ConvNext-Large foundation model as the surrogate because of incompatibility in PyTorch versions used by StyleCLIP and OpenCLIP.

Surrogate model	ΔR (before/after) adversarial training in %			Semantic & quality metrics	
	Gram-Net	UnivCLIP	Patch-Forensics	CLIP-Score	KID_{Fake}
(Adv. trained) CLIP-ResNet	47.20 / 3.43	12.47 / 1.12	40.96 / 15.57	0.669	0.01

TABLE 5: Defense performance in ΔR for the top-3 defenses from Table 3, and quality measured using CLIP-Score and KID_{Fake}. ΔR is shown for defenses before and after adversarial training. The attack uses an adversarially trained surrogate.

defender can collect such adversarial samples after deployment. However, adversarial training is known to be vulnerable to further adaptations by the attacker [67]. We study such a dynamic setting in this section—the defender adversarially trains their deepfake classifier, but subsequently, the attacker also adapts and uses an adversarially trained surrogate to craft a new distribution of adversarial samples. One would expect that such an adaptive attack can still significantly degrade the performance of the adversarially trained defense (as the attacker has also adapted).

For the attack, we use the most effective surrogate from Section 5.2.3—CLIP-ResNet. For defenses, we use Gram-Net, UnivCLIP, and Patch-Forensics, which showed comparatively smaller degradation in performance, compared to the other defenses (Table 3).

To adversarially train the current version of our defenses and the surrogate (used for attack) classifiers, we fine-tune them on a new *adversarial StyleCLIP dataset* which is a balanced dataset of $5K$, $2K$, and $2K$ images for training, validation, and testing, respectively. The adversarial fake images are generated using the CLIP-ResNet surrogate. We use the same adversarial fake images to train both the defense and the surrogate because in practice one would expect the defender to have access to the adversarial samples for adversarial training. We ensure that the attacker and the defender use a disjoint set of images for the real class (all drawn from the FFHQ dataset). After adversarial training, both the surrogate classifier and the 3 defense classifiers achieve high test set F1 scores on the adversarial StyleCLIP dataset, ranging from 94.58% to 99.40%.

Next, we craft a new set of adversarial fake images (using the adversarial trained surrogate) to test the resilience of the defenses (with adversarial training). Note that, the attack is adaptive because of retraining the surrogate on adversarial images. We generate $1K$ adversarial images using a similar methodology as before (Section 5.2.1). Table 5 shows the performance of the 3 defenses on this new distribution of adversarial fake images.

FINDING 11. *Adversarial training can be an immediate strategy to improve adversarial resilience against our attack.* From Table 5, we see that ΔR has dropped significantly for all 3 defenses after adversarial training, despite being set up against an adversarially trained attacker. For example, Gram-Net improves from ΔR of 47.20% to 3.43%, while UnivCLIP is now nearly fully resilient to adversarial attacks with a ΔR of only 1.12%. However, these results should not be taken as a message that adversarial training is a robust measure against such attacks. First, defenses still exhibit some performance degradation. Second, there may be alternative adaptive strategies by the attacker that substantially alters the distribution of the adversarial samples—such examples can potentially still disrupt these defenses.

6. Related Work

We already discuss defenses in Section 3.2. Here, we focus on related work covering *attacks on deepfake defenses.*

Some of the existing works [28], [29] to evade deepfake detection focus on adding adversarial noise to the images in the pixel space in both white- and black-box scenarios. Such attacks tend to add visible adversarial noise that degrades image quality [30], [68]. Our method does not add any adversarial noise.

Other attacks rely on eliminating specific artifacts from the fake images through post-processing to evade detection. For example, checkerboard artifacts in images produced by GAN models [69] can be countered by post-processing attack pipelines [70], [71] that remove such artifacts. Targeted removal of specific artifacts does not guarantee evasion against all defenses. We study defenses that use different methods targeting different sets of artifacts in fake images.

Carlini et al. [29] conducted a preliminary evaluation of an adversarial attack strategy that does not require the addition of adversarial noise to the images. Instead, adversarial perturbations are applied to the latent space of a StyleGAN generator. However, this attack assumes a white-box setting, unlike our attack, which considers a black-box setting. Li et al. [30] and Jia et al. [68] also study adversarial attacks without adding adversarial noise in both white-box and black-box settings. However, these methods either target a specific architectural feature of the image generator or focus on adding adversarial perturbations to specific feature spaces of the images, e.g., in the frequency domain. Our attack is more generic and does not use any specific properties of the generator or target specific feature spaces, thereby enabling broader applicability.

None of the above works systematically study the impact of using different foundation models to create surrogate classifiers. More importantly, these attacks are not thoroughly tested against a variety of state-of-the-art defenses.

7. Future Work

We discuss several new directions for future work.

(1) New directions to improve generalization performance against user-customized image generators. We find

that combining image features from foundation models with domain-specific features, e.g., frequency-based features, can significantly enhance the ability to generalize to user-customized deepfake generators. Our work also highlights the potential of using content-agnostic features that can capture imperfections in noise residuals for improved generalization. Future work can explore more effective methods to leverage these features for improved generalization.

(2) Building robustness against adversarial attacks powered by foundation models. Our new attack highlights the ease with which attackers can leverage foundation models to fool SOTA detectors. We encourage the community to explore the following directions to address this pressing challenge: (a) Further explore how defenders can build and leverage highly effective foundation models to tilt the arms race in favor of the defender. We find that a defender using a more powerful foundation model, i.e., pretrained on a larger dataset (compared to the attacker), shows improved adversarial resilience. (b) Explore novel adversarial training strategies to enhance adversarial resilience. Our analysis shows significant potential in adversarial training strategies even against an adaptive attacker.

(3) Generative model customization techniques continue to evolve, thus further expanding the deepfake threat surface. We only studied the threat of users customizing generative models using LoRA and FM fine-tuning strategies. This space is rapidly evolving with several Parameter Efficient Fine-tuning (PEFT) [72] methods being developed, e.g., DreamBooth [73] and ControlNet [74]. Users can also combine multiple LoRA model weights to create a single custom model [75]. Understanding how defenses generalize to these other customization strategies can be further explored.

(4) We need deepfake datasets covering a wide variety of content types to train and evaluate deepfake defenses. Section 4 highlights the limitations of existing work that focuses only on certain content types. Our community can create new deepfake datasets based on datasets such as LAION-5B [53] which contains 5B image-text pairs, where the text captions can be used to generate new fake image datasets. We created such a dataset, but it is limited in size (our SD dataset in Section 3.3). Creation of new large datasets can also accelerate the development of new *content-agnostic defenses* (Section 5.1.3).

(5) Foundation model-powered adversarial attacks can be more sophisticated and effective. Foundation models will continue to evolve rapidly, learning more effective patterns from Internet-scale data. The evolution of these models alone can potentially further improve the performance of our attack without requiring any changes to our attack methodology. Future work can also explore more advanced attack strategies using foundation models that exploit specific properties of the image generator. We only demonstrated our attack using the StyleCLIP generator. Our attack can also be applied to the Stable Diffusion model using classifier-guided image generation techniques [76].

(6) Our simple attack method using foundation models can be used to benchmark adversarial robustness of new defenses. We highlight the lack of proper adversarial evaluations in existing defense studies (Section 4). This is likely due to the community lacking established methods to adversarially probe a deepfake defense. We present a simple method using foundation models that can be easily applied to new defenses.

8. Conclusion

Deepfake images continue to pose a serious threat, for which several highly effective machine learning-based defenses have been proposed. In this work, we showed that these advances in defenses face a serious challenge due to two recent developments in machine learning—emergence of low-cost generator customization schemes which enable attackers to create a large variety of deepfake generators, and the emergence of vision foundation models which can be integrated with existing generators to craft adversarial fake images. Using 16 user-customized SD models, we show that existing defenses struggle to maintain their high accuracy in detecting deepfakes. We also identify their vulnerabilities against our adversarial attack, where we only make meaningful content manipulations without adding adversarial noise to the image. One of the main insights of our work is understanding the consequence of foundation models and their continuous advances on deepfake detection. We encourage further research on better leveraging foundation models for deepfake detection based on our findings.

Ethics Statement

We only used data and pretrained models that have been publicly shared for research purposes. We did not use human subjects in our research. All attacks were conducted in controlled lab settings, and no deployed services were affected using our fake images. We believe that the benefits of our work far outweigh any potential harm caused by the knowledge of our attacks.

Acknowledgements

This work was supported in part by the NSF under awards 1943351, 2231002 and the Commonwealth Cyber Initiative, an investment in the advancement of cyber R&D, innovation, and workforce development. The views and findings solely reflect those of the authors and should not be seen as representative opinions of any funding agency, given the absence of any financial conflicts.

References

[1] "Generative AI: A New Frontier in Artificial Intelligence — Deloitte Ireland," https://www2.deloitte.com/ie/en/pages/consulting/articles/generative-ai.html, 2023.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. of CVPR*, 2022.

[3] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-Shot Text-to-Image Generation," in *Proc. of ICML*, 2021.

[4] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery," in *Proc. of ICCV*, 2021.

[5] "The latest marketing tactic on LinkedIn: AI-generated faces : NPR," https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles, 2022.

[6] "AI-generated images, like DALL-E, spark rival brands and controversy - Washington Post," https://www.washingtonpost.com/technology/interactive/2022/artificial-intelligence-images-dall-e/, 2022.

[7] "Inside the pentagon's race against deepfake videos," https://www.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/, 2019.

[8] "Liveness tests used by banks to verify ID are 'extremely vulnerable' to deepfake attacks," https://www.theverge.com/2022/5/18/23092964/deepfake-attack-facial-recognition-liveness-test-banks-sensity-report, 2022.

[9] C. Li, L. Wang, S. Ji, X. Zhang, Z. Xi, S. Guo, and T. Wang, "Seeing is Living? Rethinking the Security of Facial Liveness Verification in the Deepfake Era," *CoRR abs/2202.10673*, 2022.

[10] "As Deepfakes Flourish, Countries Struggle With Response - The New York Times," https://www.nytimes.com/2023/01/22/business/media/deepfake-regulation-difficulty.html, 2023.

[11] U. Ojha, Y. Li, and Y. J. Lee, "Towards Universal Fake Image Detectors that Generalize Across Generative Models," in *Proc. of CVPR*, 2023.

[12] J. Ricker, S. Damm, T. Holz, and A. Fischer, "Towards the Detection of Diffusion Model Deepfakes," in *Proc. of VISAPP*, 2024.

[13] Z. Liu, X. Qi, and P. H. Torr, "Global Texture Enhancement for Fake Face Detection in the Wild," in *Proc. of CVPR*, 2020.

[14] Y. He, N. Yu, M. Keuper, and M. Fritz, "Beyond the Spectrum: Detecting Deepfakes via Re-Synthesis," in *Proc. of IJCAI*, 2021.

[15] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proc. of CVPR*, 2020.

[16] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in *IEEE WIFS*, 2018.

[17] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in *Proc. of ECCV*, 2020.

[18] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proc. of CVPR*, 2019.

[19] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. of ICLR*, 2018.

[20] "CivitAI," https://civitai.com/, 2022.

[21] "Models - Hugging Face," https://huggingface.co/models, 2021.

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *CoRR abs/2106.09685*, 2021.

[23] R. Bommasani, D. A. Hudson *et al.*, "On the Opportunities and Risks of Foundation Models," *CoRR abs/2108.07258*, 2021.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. of ICML*, 2021.

[25] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. of ICML*, 2019.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. of ICLR*, 2021.

[27] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models," in *Proc. of ACM CCS*, 2023.

[28] Y. Hou, Q. Guo, Y. Huang, X. Xie, L. Ma, and J. Zhao, "Evading DeepFake Detectors via Adversarial Statistical Consistency," in *Proc. of CVPR*, 2023.

[29] N. Carlini and H. Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks," in *Proc. of CVPR Workshop*, 2020.

[30] D. Li, W. Wang, H. Fan, and J. Dong, "Exploring Adversarial Fake Images on Face Manifold," in *Proc. of CVPR*, 2021.

[31] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *ACM CSUR*, 2021.

[32] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park, "Scaling up GANs for Text-to-Image Synthesis," in *Proc. of CVPR*, 2023.

[33] A. Van Den Oord, O. Vinyals *et al.*, "Neural Discrete Representation Learning," *Adv. in NeurIPS*, 2017.

[34] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," *Adv. in NeurIPS*, 2022.

[35] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *Proc. of CVPR*, 2023.

[36] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *CoRR abs/2204.06125*, 2022.

[37] "Midjourney," https://www.midjourney.com, 2022.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. of CVPR*, 2009.

[39] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy *et al.*, "A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark," *CoRR abs/1910.04867*, 2019.

[40] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," in *Proc. of CVPR*, 2020.

[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *Communications of the ACM*, 2020.

[42] Z. Xu, Z. Hong, C. Ding, Z. Zhu, J. Han, J. Liu, and E. Ding, "MobileFaceSwap: A Lightweight Framework for Video Face Swapping," in *Proc. of AAAI*, 2022.

[43] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language Image Pre-training for Unified Vision-Language Understanding and Generation," in *Proc. of ICML*, 2022.

[44] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging Frequency Analysis for Deep Fake Image Recognition," in *Proc. of ICML*, 2020.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. in NeurIPS*, 2012.

[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *Proc. of CVPR*, 2015.

[47] "LAION-Aesthetics," https://laion.ai/blog/laion-aesthetics/, 2022.

[48] "CLIP+MLP Aesthetic Score Predictor," https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022.

[49] "Realistic_Vision_V1.4." https://huggingface.co/SG161222/Realistic_Vision_V1.4/tree/main, 2023.

[50] "runwayml/stable-diffusion-v1-5," https://huggingface.co/runwayml/stable-diffusion-v1-5, 2022.

[51] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," *CoRR abs/2112.10741*, 2021.

[52] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "LAION-400M: Open Dataset of CLIP Filtered 400 Million Image-Text Pairs," *NeurIPS Workshop*, 2021.

[53] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "LAION-5B: An open large-scale dataset for training next generation image-text models," *Adv. in NeurIPS*, 2022.

[54] "Using LoRA for Efficient Stable Diffusion Fine-Tuning," https://huggingface.co/blog/lora, 2023.

[55] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, and T. Wolf, "Diffusers: State-of-the-Art Diffusion Models," https://github.com/huggingface/diffusers, 2022.

[56] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-Concept Customization of Text-to Image Diffusion," in *Proc. of CVPR*, 2023.

[57] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A Reference-free Evaluation Metric for Image Captioning," in *Proc. of EMNLP*, 2021.

[58] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANS," in *Proc. of ICLR*, 2018.

[59] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," *Proc. of NeurIPS*, 2017.

[60] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *Advances in neural information processing systems*, 2020.

[61] J. Pu, N. Mangaokar, B. Wang, C. K. Reddy, and B. Viswanath, "NoiseScope: Detecting Deepfake Images in a Blind Setting," in *Proc. of ACSAC*, 2020.

[62] D. Zhang, F. Zhou, Y. Jiang, and Z. Fu, "MM-BSN: Self-Supervised Image Denoising for Real-World with Multi-Mask based on Blind-Spot Network," in *Proc. of CVPR*, 2023.

[63] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proc. of ICLR*, 2014.

[64] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proc. of CVPR*, 2018.

[65] "CLIP-convnext-large," https://huggingface.co/laion/CLIP-convnext_large_d_320.laion2B-s29B-b131K-ft-soup, 2023.

[66] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *Proc. of ICLR*, 2015.

[67] H. Zhang, H. Chen, Z. Song, D. Boning, I. S. Dhillon, and C.-J. Hsieh, "The Limitations of Adversarial Training and the Blind-Spot Attack," in *Proc. of ICLR*, 2019.

[68] S. Jia, C. Ma, T. Yao, B. Yin, S. Ding, and X. Yang, "Exploring Frequency Adversarial Attacks for Face Forgery Detection," in *Proc. of CVPR*, 2022.

[69] K. Schwarz, Y. Liao, and A. Geiger, "On the Frequency Bias of Generative Models," in *Proc. of NIPS*, 2021.

[70] V. Wesselkamp, K. Rieck, D. Arp, and E. Quiring, "Misleading Deep-Fake Detection with GAN Fingerprints," in *Proc. of IEEE S&P Workshop*, 2022.

[71] Y. Huang, F. Juefei-Xu, R. Wang, Q. Guo, L. Ma, X. Xie, J. Li, W. Miao, Y. Liu, and G. Pu, "FakePolisher: Making DeepFakes More Detection-Evasive by Shallow Reconstruction," in *Proc. of ACM ICM*, 2020.

[72] "PEFT," https://huggingface.co/docs/peft/index, 2023.

[73] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine Tuning Text-to Image Diffusion Models for Subject-Driven Generation," in *Proc. of CVPR*, 2023.

[74] L. Zhang, A. Rao, and M. Agrawala, "Adding Conditional Control to Text to-Image Diffusion Models," in *Proc. of ICCV*, 2023.

[75] "Egg Fusion - LoRA Merge," https://civitai.com/models/43863/egg-fusion-lora-merge, 2023.

[76] D. Kim, Y. Kim, S. J. Kwon, W. Kang, and I.-C. Moon, "Refining Generative Process with Discriminator Guidance in Score-based Diffusion Models," in *Proc. of ICML*, 2023.

[77] "(2) emad on x: "@kennethcassel we actually used 256 a100s for this per the model card, 150k hours in total so at market price $600k" / x," https://twitter.com/EMostaque/status/1563870674111832066, 2022.

[78] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila, "StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis," *CoRR abs/2301.09515*, 2023.

[79] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang *et al.*, "CogView: Mastering Text-to-Image Generation via Transformers," *Adv. in NeurIPS*, 2021.

[80] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, "VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance," in *Proc. of ECCV*, 2022.

Appendix A.
Generative Models and Defenses

Defense implementations.

UnivCLIP. Here, the hyperparameters provided by the authors did not yield the best performance. UnivCLIP worked best on the SD dataset when it was fine-tuned for 200 epochs using Adam optimizer with a learning rate of $5e-4$, combined with 10 epochs for early stopping. For the StyleCLIP dataset, we used the same setting with the difference of using a learning rate of $1e-3$. Both settings required data augmentation.

DE-FAKE. Optimal performance was not achieved with author-provided hyperparameters. For the SD dataset, we fine-tuned the model for 200 epochs at a learning rate of $5e-4$. For the StyleCLIP dataset, it was relatively harder to achieve good performance, but fine-tuning for 200 epochs at a learning rate $5e-5$ gives the best results. One of the reasons for struggling with our StyleCLIP dataset may be the low discriminatory value in text features between real and fake data, which is key for classification with DE-FAKE.

DCT. Instead of a 64×64 central cutout, we modify the pipeline to extract log(DCT) features from the entire image for a more comprehensive analysis. We also use a smaller version of our dataset, i.e. a balanced dataset of $3K$, $1K$, and $2K$ images for training, validation, and testing, respectively to decrease computational complexity. As it is a simple linear classifier, we train the model from scratch for 10 epochs with the SGD optimizer with a learning rate of $1e-2$ and a weight decay regularization of $1e-3$ for both datasets.

Patch-Forensics. Best performance was obtained using the author-provided hyperparameters when fine-tuned on both the datasets with learning rate of $1e-3$ and 100 epochs.

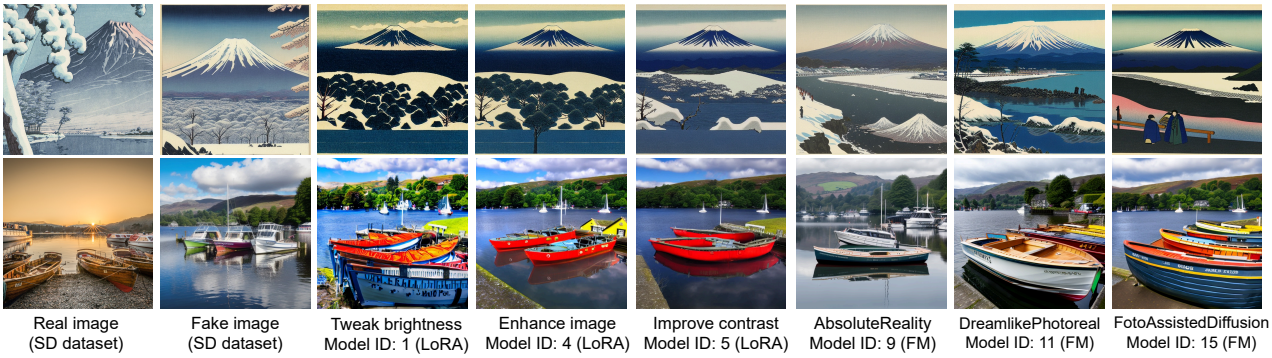


Figure 9: From left to right: first 2 images are real and fake images from our SD test set, the next 3 are LoRA images, followed by 3 FM images. Model IDs are explained in Table 6 (Appendix B). We can see content preservation with comparable quality across all samples.

Gram-Net. Author-provided hyper-parameters did not yield optimal performance for SD dataset, but after fine-tuning for 100 epochs with Adam optimizer, using learning rate of $1e-4$ without weight decay gave optimal performance. For StyleCLIP, fine-tuning for 10 epochs with author-provided hyper parameters gave high performance.

Resynthesis. Author-provided hyperparameters did not yield good performance here. The best performance was obtained by fine-tuning on both datasets with a learning rate of $1e-2$ and a reconstruction learning rate of $1e-3$. Resynthesis was trained for 150 and 100 epochs for SD and StyleCLIP datasets, respectively.

CNN-F. We fine-tune both author-provided checkpoints with given hyperparameters and found the best performance. with the Blur+JPEG (0.5) model on the SD dataset, and Blur+JPEG (0.1) model on the StyleCLIP dataset.

MesoNet. A MesoInception-4 model was fine-tuned for 100 epochs using MSE loss with Adam optimizer and a learning rate of $1e-3$ for both datasets.

Other generative models. We considered the use of other generative models for our evaluation but did not include them for one or more of the following reasons: (1) *Unavailability of training code or pretrained checkpoints*: High-quality generative models require significant computational effort [77]. It is impractical to train them from scratch with limited computational resources. For goal 3 (Section 2), we require training code to adversarially update the generator. Both requirements exclude models such as DALL-E [3], GLIDE [51], StyleGAN-T [78], and many others [32], [34]. (2) *Poor quality imagery*: We aim to consider a challenging setting where the deepfake images are of high quality; otherwise, they can be easily flagged by human inspection. This criteria excludes models such as CogView [79] and VQGAN-CLIP [80]. For example, we find that VQGAN-CLIP generates images with repeated artifacts that can be easily flagged.

Appendix B. Details of User-customized SD models

Table 6 shows details of the 16 user-customized models, which includes both LoRA and FM based models.

Appendix C. Generating Adversarial Fake Images

Text prompts used for our attack. StyleCLIP requires a *neutral* and *target* text. As we only work with human face images, *neutral* text is always "a face". The list of *target* text is: (i) a smiling face, (ii) a happy face, (iii) a sad face, (iv) a face with glasses, (v) a face with lipstick, (vi) a face with blue eyes, (vii) a face with brown hair, and (viii) a face with surprise.

Training configuration for surrogates and generator. For ViT and EfficientNet, we use the SGD optimizer with $1e-3$ learning rate, 0.9 momentum, and set γ and δ to 0.1 and 1.0, respectively. For CLIP-ResNet, we use the same optimizer settings with γ and δ set to 0.02 and 1.0, respectively. In all cases, we have set α and β to 9.0 and 0.12, respectively.

Fine-tuning Surrogate Classifiers. We use FFHQ and StyleGAN2 generated images as *real* and *fake* data, and ensure that there is no overlap with the data to fine-tune the defenses, i.e. StyleCLIP dataset. We take $10K$, $2K$, and $2K$ images per class for training, validation, and testing as our **surrogate StyleCLIP dataset**. As surrogates, we choose ViT as it is the state-of-the-art transformer-based encoder model used for image classification tasks, EfficientNet because of its superior performance despite being 21 times smaller than comparable ConvNets, and CLIP-ResNet because it is pretrained with an Internet-scale dataset. To build the surrogates, we choose ViT-base (for ViT), EfficientNet-B0 (for EfficientnET) and $64\times$ ResNet-50 (for CLIP-ResNet), and add a linear layer for binary classification. We fine-tune ViT, EfficientNet and CLIP-ResNet for 20, 30 and 30 epochs, respectively, with the Adam optimizer at a learning rate of $5e-5$, $5e-4$ and $1e-3$.

Type	Model ID	Description	Links
LoRA	1	Tweak brightness	civitai.com/models/70034/brightness-tweaker-lora-lora
	2	Add details	civitai.com/models/58390/detail-tweaker-lora-lora
	3	Increase sharpness	civitai.com/models/69267/sharpness-tweaker-lora-lora
	4	Enhance image	civitai.com/models/78283/elixir-enhancer-lora
	5	Improve contrast	civitai.com/models/48139/lowra
	6	Add aesthetic details	civitai.com/models/82098/add-more-details-detail-enhancer-tweaker-lora
	7	Reduce image noise	civitai.com/models/13941?modelVersionId=16576
	8	Tweak skin texture	civitai.com/models/134883/skintextureslider-plastic-skin-realistic-skin
FM	9	Increase realism (AbsoluteReality)	huggingface.co/Lykon/AbsoluteReality/tree/main
	10	Real-life photographs (AnalogDiffusion)	civitai.com/models/1265/analog-diffusion
	11	Photorealism (DreamlikePhotoreal)	huggingface.co/dreamlike-art/dreamlike-photoreal-2.0
	12	Artistic and realistic (DreamShaper)	civitai.com/models/4384/dreamshaper
	13	High-quality styled images (EpicDiffusion)	civitai.com/models/3855/epic-diffusion
	14	Photorealistic image (epiCRealism)	civitai.com/models/25694/epicrealism
	15	HDR photography (FotoAssistedDiffusion)	huggingface.co/Dunkindont/Foto-Assisted-Diffusion-FAD_V0
	16	Realistic artwork (Haveall)	civitai.com/models/118799/haveall

TABLE 6: Details of LoRA and FM models used in our work.

For ViT and EfficientNet, all layers of the models are fine-tuned, whereas for CLIP-ResNet, only the linear layer added at the end is fine-tuned to obtain optimal performance.

Explaining differences between Source fake, Non-adversarial fake and Adversarial Fake. We explain the changes that cause fake image to be adversarial against defenses. In Figure 7, we see the following differences between *non-adversarial fake* and *adversarial fake* images after our attack: (i) second image from the left has brighter skin-tone, (ii) third image has green eyes, (iii) fifth image has more texture under the eyes, (iv) sixth image has fewer folds in the skin with a lighter skin tone. Other images are too similar to point anything out. The target text for the images (from left to right) are (i) a face with lipstick, (ii) a face with brown hair, (iii) a face with brown hair, (iv) a face with lipstick, (v) a smiling face, (vi) a smiling face, and (vii) a face with surprise.

Appendix D. Ensemble with UnivCLIP

Defenses	Precision (%)	Recall (%)	F1 (%)
UnivCLIP	90.09	68.56	76.8
UnivCLIP + Gram-Net	90.9	81.84	85.26
UnivCLIP + Resynthesis	80.22	82.65	80.98
UnivCLIP + CNN-F	90.51	74.37	80.64
UnivCLIP + MesoNet	87.27	82.73	84.35
UnivCLIP + DE-FAKE	86.4	81.89	83.42

TABLE 7: Improving generalization performance by creating an ensemble model by combining UnivCLIP defense with the other 5 defenses. We show the average scores across the 16 SD custom checkpoints.

Appendix E. Frequency Spectrum Analysis of User-customized model images

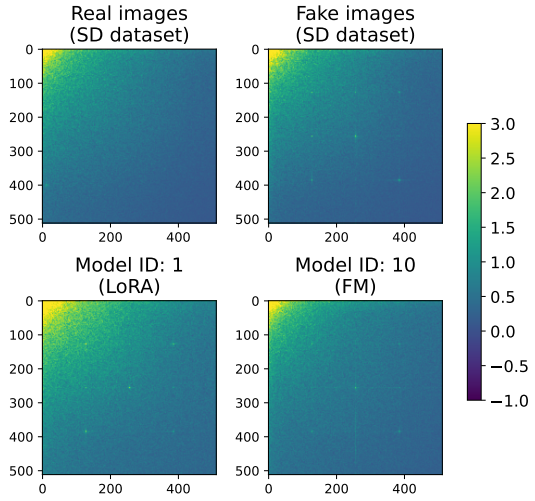


Figure 10: Average log (DCT) spectrum calculated over $1K$ images for each category. "Real" and "Fake" images are from our SD dataset, and the bottom 2 plots correspond to images from LoRA and FM models in Table 6. LoRA and FM images show frequency spectrum artifacts similar to fake images, but they are absent for real images. This explains the strong detection performance of DCT over user-customized models.

Appendix F. Meta-Review

The following meta-review was prepared by the program committee for the 2024 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

F.1. Summary

This paper studies the shortcomings of existing deepfake defense techniques by conducting experiments with two groups of attacks (1) customized generative model and (2) foundation models to generate adversarial deepfakes. The experiments show that current deepfake defense techniques are ineffective against different attack vectors and performance can be heavily impacted. Further, the paper explores different possible ways to improve the deepfake defense.

F.2. Scientific Contributions

- Independent Confirmation of Important Results with Limited Prior Research
- Provides a Valuable Step Forward in an Established Field

F.3. Reasons for Acceptance

1) This paper identifies major shortcomings of existing deepfake defense techniques. The authors select 8 state-of-the-art deepfake detectors and evaluate them with different adaptive attacks leveraging customized generative models and foundation models to generate adversarial samples. The evaluation shows that the performance of deepfake detectors is heavily impacted against different adversarial attacks and the detection rate can be lowered by over 53%. This paper provides a valuable step forward in deepfake research.
2) The number of deepfake images and videos is increasing rapidly with the popularity of open-source AI models and tools. This paper discusses several valuable insights to improve deepfake detection techniques against novel adversarial attacks.

F.4. Noteworthy Concerns

1) The paper uses the SD dataset for generalizability and the StyleCLIP dataset for adversarial attack evaluation. This might introduce some biases in the evaluation. This paper would benefit from a more generalized approach to the evaluation.

Appendix G. Response to the Noteworthy Concerns

In Section 5.1, we use the SD dataset because among the two models (SD and StyleCLIP), only the SD model has seen widespread user-customization. Therefore, we cannot include a version of the StyleCLIP dataset to study generalization. Recall that the goal of this section is to study the impact of user-customization on deepfake detection. In Section 5.2, evaluating our attack using SD can be accomplished using a classifier-guidance approach, which we leave for future work. We discuss these points in the following places: (1) at the end of the first paragraph in Section 5.1, and (2) under Point 5 in Section 7.