# Identifying Modern Deepfakes: Bringing Fake Image Detection Into the Wild

by

Hannah Lee

Supervised by Oren Etzioni

A masters thesis submitted in partial fulfillment of
the requirements for the degree of

Master of Science

Paul G. Allen School of Computer Science & Engineering
University of Washington

June 2024

**Abstract**

Multimodal generative models are rapidly evolving, leading to a surge in the generation of realistic images that bring both exciting, creative possibilities, and serious risks. Deepfake images have particularly garnered increasing attention due to their potential for spreading misinformation and generating fraudulent content. We present a series of experiments conducted as part of the TrueMedia.org organization committed to fighting deepfakes and misinformation targeted at political elections, showing that (1) open source fake image detectors often fail to generalize to in-the-wild images; (2) despite poor performance when used out of the box, a basic, pretrained CLIP vision transformer can be easily modified and trained to detect *in-the-wild* fake images with high-quality training data, reaching an average precision of 91.9 and an F1 score of 0.83; and (3) this model is capable of learning to detect synthetically generated images from multiple standard image generators with average precision of greater than 99.0. In addition, we discuss challenges in modern deepfake detection and identify directions for future work.

## Acknowledgements

Countless people have contributed to the existence of this work. I am grateful to **Professor Oren Etzioni** for giving me the opportunity to carry out this research, supporting it through his guidance and leadership at **TrueMedia.org**, and for providing insights into potential directions during all stages of these projects. The entire TrueMedia.org team has been essential in brainstorms, discussions, and motivations for this research, and I am especially grateful to my research mentor **Aerin Kim** who has continuously mentored my work and provided feedback throughout my time with the organization. Lastly, at TrueMedia.org, I am fortunate to have collaborated with the other interns: **Kevin Farhat**, **Lin Qiu**, **Arnab Karmarkar**, and **Ben Caffee**, each of whom I have had the pleasure of discussing this research with.

I would not have found myself detecting deepfakes or working with TrueMedia.org if I had not met some incredible mentors along the way. I would like to thank **Professor Ludwig Schmidt** for his advice and support when applying to the masters program at the University of Washington and his continued mentorship throughout my graduate studies; **Beatriz Stollnitz** for instilling confidence in me to pursue computer science many years ago; all of the professors I've had the privilege of working for as a machine learning teaching assistant for seven quarters; and the incredible PhD students from whom I've had the fortune of receiving any career advice I could get out of them. Each of them have shaped how I approach my work, how I view the world, and what I hope to accomplish moving forward.

Finally, I would like to thank my friends and family for their support of my academic career as well as my personal development. They have shown me what strength and happiness look like. In particular, I would like to thank **Theo Gregersen** for his constant kindness and encouragement; the many friends and mentors I have found in the **Race Condition Running** group; all of my peers from the undergraduate Honors program, Alina, Elizabeth, and Andy, among many others; and most importantly, my parents for their relentless patience, support, and never-ending inspiration.

# Contents

# 1 Introduction

Recent advancements in multimodal generative models have made manipulated media increasingly more realistic and widespread. While synthetically generated audio, images, and videos can have practical and creative applications including creating engaging educational videos and generating improved dubbing or translations [70, 28], deepfake videos that impersonate humans highlight the potential harms of the machine learning generation techniques. For example, deepfakes that blend faces of celebrities onto the bodies in pornographic videos [49, 27] and alter the messages of politicians [63] can spread misinformation, threaten individuals, and damage reputations [74], disrupting election campaigns and financial markets. Recently, deepfakes have also been behind fraudulent schemes; by impersonating powerful figures and known colleagues, deepfake videos have resulted in scams of up to $25 million [13]. Given the rise of social media and prevalence of online media consumption, it is unsurprising that deepfake content is increasingly interfering with people's lives.

The first modern deepfakes surfaced in 2017 when users on Reddit posted computer-generated pornographic videos of actresses [49]. Since then, many deepfake generation tools have become available for public use. Applications and open source repositories such as FaceSwap [5], FaceSwapGAN [62], StyleGAN [32], and FSGAN [50] allow anyone with basic programming skills to generate their own deepfakes, and as text-to-image and text-to-video models such as Imagen Video [24], CogVideo [25], and Sora [8] improve and become widespread, the barrier to entry will continue to be lowered. With deepfake generation becoming increasingly democratized and capable of producing realistic results, emphasis on countermeasures has also grown. Generally, addressing misinformation and the harms related to deepfakes follows two patterns; detection of manipulated media and preventing generation. While prevention through techniques like watermarking [45] and blockchain technology [55] as well as through technology policy [56] is necessary for mitigating deepfake harms, detection is an important tool that can be applied to deepfake content that has already been created and can be applied sooner than developing and widely adopting new prevention techniques.

Fake image detection is typically framed as a binary classification task where classifiers learn to distinguish authentic images from manipulated ones. Initial detection tools focused primarily on visual artifacts such as blended face edges [36] and image forgery techniques that have preceded deepfakes [12]. Deep learning techniques have since become more popular in detection methods, taking advantage of the large amount of real and fake data available online. However, since training detection algorithms often relies on synthetic data created by the deepfake generation tools, deepfake detection lags behind generation. Consequently, the development of detection algorithms provides direct feedback to generation algorithms on what makes

deepfakes detectable and can encourage adversarial generation to bypass detection. As one example of evolving detection and generation, artifacts specific to image generated by generative adversarial networks (GANs) are less present in images generated by newer, diffusion-based generative models [15]. This training paradigm can also encourage models to learn the unique identifying artifacts of specific generation techniques if they are trained on datasets of only one class of image generator, rather than detecting any shared characteristics of manipulated media in general. This makes detecting deepfake content in the wild, where it has the greatest risk of misleading individuals, particularly challenging.

In this thesis, we first survey the existing landscape of AI-generated image detection methods and find that many detection methods are not available for public use. Then, evaluating the existing tools that are open source on a custom dataset of in-the-wild, crowdsourced images, we find that techniques that perform well on established benchmarks fail to generalize to real-world examples that humans are uncertain about. Building upon these findings and the existing detection mechanisms, we investigate how to build and improve deepfake image detection to become effective in real-world systems. Our results show that adding a small, simple network on top of a learned CLIP vision encoder outperforms finetuning a popular detection method designed to identify AI-generated images.

Building on prior work, we also find that training the additional network on top of the learned CLIP vision encoder on increasingly diverse datasets with multiple generation methods does not hurt the detector's performance on single generation methods. In particular, we examine training and evaluating the base model using curated datasets consisting of deepfakes and synthetically generated media from StyleGAN2 [33], Stable Diffusion [58], DALL-E 2 [54], Midjourney [1], and DeepFloyd IF [3], training on up to approximately 60,000 samples.

# 2 Background and Related Work [1]

## 2.1 Deepfake Image Generation

There are many types of manipulated images. Small alterations to images can be made using tools such as Adobe Photoshop [2] to edit the lighting or background of an image or change an attribute of a person such as their facial expression, while larger changes can swap entire faces of two subjects in two different images using deepfake generation tools [61, 62, 5] or insert whole objects into existing scenes where objects are synthetically generated [54] or real images edited in. Recently, newer techniques of image generation

[1] This section has been adapted from sections of a co-authored survey paper on deepfake generation and detection, "The Tug-of-War Between Deepfake Generation and Detection," currently under review for the ICML 2024 Workshop on Data-Centric Machine Learning Research.

can create entire images from the pixel level without leveraging existing images, typically converting text prompts into media [1, 3, 58, 54]. Any combination of these techniques can be utilized to create realistic deepfake images, complicating in-the-wild detection.

## 2.2 Fake Image Detection

Detecting fake images precedes the introduction of deep learning and creation of deepfakes and aims to detect a range of image manipulations. These methods range from artifact detection to modern methods that have developed in response to advancements in image generation techniques.

### 2.2.1 General Visual Artifact Detection

Deepfake images may introduce subtle artifacts that are not present in real images. These can include artifacts introduced by face blending or face warping when swapping faces, as well as inconsistencies in the overall image. For example, Li et al. [36] proposed the "face X-ray" image representation to detect anomalies in the blending boundaries of faces in images that have been blended together. Other work has focused on the differing image textures between generated and real content [41]; the resolution inconsistencies that arise from warping faces to match them to real images [37]; and missing reflections or missing details in the teeth and eyes [47]. Many other visual artifact detection methods have been proposed, but as deepfake generation has improved and fewer artifacts remain, these techniques have become overshadowed by methods focused on detecting more subtle identifiers of generated images.

### 2.2.2 GAN specific techniques

Many widely available deepfake generation tools employ GANs, including FaceSwap [5], FaceSwapGAN [62], StyleGAN [32, 33], and FSGAN [50], which typically employ a "face swap" technique to replace one person's face with another. Due to the prevalence of these generation techniques, image deepfake detection in the past has focused on detecting artifacts unique to GAN models. Some detection methods rely on visible differences such as irregular pupil shapes [20] while many others exploit lower level abnormalities; the upsampling operations involved in GAN generation introduce model specific artifacts into the images' spatial and frequency domains [73, 46, 72]. Wang et al. [66] trained a ResNet-50 model [22] that detects these such artifacts and found that GAN generated images are easily detected. Similarly, FakeSpotter [65] is able to effectively detect AI-synthesized fake faces generated by GANs. Building upon these detectors, PatchForensics [11] introduced a detector that analyzes smaller patches of images to determine if there are AI-generated or manipulated areas of a larger image.

### 2.2.3 Modern diffusion detection

Despite GAN-generated image detection being successful, generalization of these detection methods to newer, diffusion based image generation techniques is difficult [15, 51], in part because the artifacts introduced by GANs in images are no longer present in diffusion model generated media. Recently, Wang et al. [67] proposed a new image representation DIRE that uses reconstructions of images using diffusion models as a method for detecting diffusion model generated images, since diffusion model generated images consist of features that are better reconstructed by other pretrained diffusion models than real images. Ojha et al. [51] instead utilize the learned feature space of a pretrained vision-language model to determine if an image was AI-generated. And Lorenz et al. [43] relies on multi Local Intrinsic Dimensionality to detect diffusion. These novel methods highlight how detection algorithms continue to adapt to the newer generative models that evade older detection methods that become inaccurate over time. Other work has also shown that further training of GAN-generated image detectors on diffusion-generated images is sufficient for improving generalizability [19, 57], but modern detection has not yet been studied specifically for deepfake images or in-the-wild data.

## 2.3 Datasets

Each of the methods used for deepfake generation and detection have been trained on datasets of real and fake images. Numerous datasets of real human faces have been curated for various facial recognition tasks. In 2007, the Labeled Faces in the Wild (LFW) database made public over 13,000 images of faces found on the internet to study facial recognition [29]. Though now retracted, the MS-Celeb-1M dataset consisted of one million real images of 100,000 different identities, mainly of various celebrities and journalists, also designed to develop facial recognition technologies [21]. And IMDB-WIKI [60] and IMDB-Clean [40] include 524,230 and 287, 683 images respectively of faces, designed to train age-estimation algorithms. These datasets enable training of face image generation models.

For deepfake image generation techniques that rely on face swapping or reenactment, the deepfake creator must have sufficient training data focused on the target and source subjects involved [5, 61]. This has limited who deepfakes can be generated of to individuals with large amounts of publicly available content such as politicians and celebrities, but as the quantity of public data increases and generation techniques improve the barrier to creating deepfakes continues to fall. In fact, higher quality and larger datasets of real human faces have emerged both in part due to online sources such as Flickr and their role in training GANs to generate more realistic media. The CelebA-HQ dataset [31] is one such dataset of 30,000 images of the faces of celebrities, based on the larger CelebA dataset [42], that has been used to train GANs to generate higher

resolution images of human faces. Karras et al. [32] went on to create Flickr-Faces-HQ (FFHQ), a larger and more diverse dataset of 70,000 facial images, inclusive of broader ranges of ages, ethnicities, and image backgrounds as well as accessories.

In contrast to deepfake generation models, detection models most commonly are trained on large datasets of both real and fake images. Often, these are custom datasets created specifically for the detection focus and currently available image generators. Wang et al. [66] trained an image classifier on the LSUN [71] dataset and ProGAN [32] generated images. To test their image classifier, Wang et al. generated over 72,000 images using 11 different CNN-based models including StyleGAN [32], BigGAN [7], StarGAN [14], and CycleGAN [75]. This collection of real and fake samples has been used by subsequent detection methods to train and evaluate models [51, 15]. Sharing datasets allows for direct comparison and improves transparency but the lack of newer generation methods in older datasets necessitates new custom evaluations [67, 43, 44]. While there have been efforts to standardize deepfake video detection benchmarks through the DeepFake Detection Challenge (DFDC) [6] and the FaceForensics++ [59], Celeb-DF [38], DeeperForensics-1.0 [30], and AV-Deepfake1M [9] datasets, there are no well established evaluations for manipulated image detection.

# 3 Preliminaries

## 3.1 TrueMedia.org Collaboration

Founded in 2023 by Dr. Oren Etzioni, TrueMedia.org (hereafter referred to as TrueMedia) is a non-partisan, non-profit organization committed to fighting AI-based disinformation, particularly ahead of the 2024 U.S. election season. TrueMedia provides a free platform for media consumers to submit content for which they would like to verify its authenticity. In this way, the organization collects the real-world examples that humans have encountered in-the-wild and are uncertain about. Partnered with currently leading deepfake detection startups such as AI or NOT, Reality Defender, Hive, and Sensity, each of which has their own suite proprietary detection tools, TrueMedia is able to aggregate responses of these detection tools for each piece of media submitted for analysis, providing access to the detection tools that consumers might otherwise not have a way to interact with.

Partnered organizations are important to TrueMedia's mission but the underlying algorithms of detection tools remains a black box, and not all detection tools are trained to detect the specific media content TrueMedia aims to analyze. Because of this, in-house methods are being developed by the organization. It is in collaboration with TrueMedia and its research group that we carry out this work on detecting fake

images in the wild as part of an effort to develop and improve fake media detectors for their products.

Part of the collaboration with TrueMedia allows for the unique ability to crowdsource data from both product users and TrueMedia's own data experts, enhancing the data sources for in-the-wild detection training beyond sources commonly available online. We utilize these high quality examples in our work to evaluate and improve various detection methods.

### 3.2 Deepfake Detection in the Open Source Community

While there are strong open source communities surrounding the deepfake generation space for common repositories such as Face Swap [5] and Face Fusion [61], the deepfake detection space is primarily contributed to through academic publications and their released code implementations. These repositories are important in the community and are the backbone for much of the work that we have done; however, they are often poorly maintained and there is little consistency between the different works. Some of the research contributions have been led in part by DARPA and the Semantic Forensics (SemaFor) program, which aims to detect, attribute, and characterize manipulated and synthesized media to address disinformation campaigns [16]. The SemaFor program's recently published Analytic Catalog of various open-source detection techniques is one example of how researchers in the detection space can come together to create a larger detection community. However, the catalog remains in its early stages and included methods are not evaluated in our work either due to a different target modality (detecting text or articles rather than images) or an inability to use the provided source codes (e.g., lack of public model weights or missing model training code).

Perhaps driven by the data requirements to train generated image detectors and the potential harm of generation tools, there has also been an increase in detection methodologies led by the generation tool creators themselves. For example, OpenAI has announced their efforts to implement tamper-resistant watermarking to their digital content in addition to developing their own detection classifiers, including a detector to distinguish between DALL-E 3 created images and non-AI generated images [52]. These efforts are also integral to the overall mission against deepfake-related misinformation in the wild, but similar to TrueMedia's partners, these methods remain closed to the open source community and may have difficulty generalizing to out-of-distribution media.

| Subset | Real | Fake |
|---|---|---|
| Train | 162 | 162 |
| Test | 90 | 166 |

Table 1: Summarized contents of the curated TrueMedia dataset. At the time of model evaluation, there were 580 total high quality images labeled with their ground truth values, with 256 images allocated to the evaluation (test) set.


Figure 1: Example images from the curated TrueMedia evaluation dataset. Images include examples featuring influential political, religious, and social figures. The curated dataset allows for more recent, popular images such as the viral fake image of Pope Francis in a white puffer coat to be included in the model evaluations.

# 4 Methods and Experimental Design

## 4.1 The UniversalFakeDetect Model

To begin, we evaluate the UniversalFakeDetect model released by Ojha et al. [51] on in-the-wild data collected by the TrueMedia organization to test the universality of the detector. Evaluation images were sourced from queries submitted by users and by data experts affiliated with the team, who curated the dataset to include relevant and difficult examples (i.e., ones especially pertinent to TrueMedia's goal of adddressing political deepfakes and images that are indiscernible between fake and real to humans). Table 1 summarizes the contents of this curated dataset and Fig. 1 shows select examples.

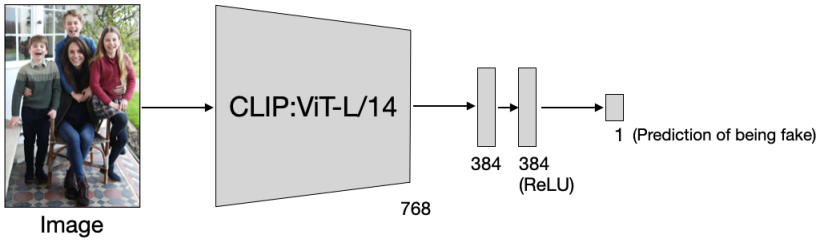


Figure 2: Model architecture for UFD-T models. The CLIP:ViT-L/14 backbone is kept frozen.

#### 4.1.1 Generative Models

The UniversalFakeDetect model uses a frozen, pretrained CLIP ViT-L/14 [53] vision transformer as the backbone and adds a linear projection layer to perform binary classification. The vision transformer ViT-L/14 [18] has been trained as part of CLIP for image-language alignment rather than for image classification. However, CLIP:ViT has been trained on 400 million image-text pairs from images sourced from the internet, giving the model more real-world context than most other available image models that have been trained on other datasets such as ImageNet [17] or CIFAR-10 [35]. While other network backbones exist for CLIP models, Ojha et al. [51] found that the ViT-L/14 vision encoder performed best and release public model weights for this particular architecture; following this finding, we conduct evaluations and experiments solely with the ViT-L/14 vision encoder backbone. We begin by performing evaluations on the released model weights as well as a series of ablation studies considering variants of the UniversalFakeDetect paradigm:

1. **UniversalFakeDetect baseline.** The released UniversalFakeDetect includes 769 learned parameters in the linear layer, which was trained using binary cross entropy loss.

2. **Finetuned UniversalFakeDetect.** We finetune the base UniversalFakeDetect model's trained weights by starting from the released weights for the linear layer on the curated TrueMedia dataset.

3. **Re-initialized UniversalFakeDetect.** We analyze how training the linear layer on curated TrueMedia data differs from using the trained UniversalFakeDetect model. We follow the training parameters outlined by Ojha et al. [51] and initialize the 769 trainable parameters sampled from a normal distribution.

4. **UFD-T (UniversalFakeDetect-Transfer).** A small-scale study to determine capabilities of transfer learning when extending the UniversalFakeDetect architecture to include two fully connected layers with a ReLU activation function applied. This increases the trainable parameters to 295,681. See Fig. 2.

Then, the performance of the UniversalFakeDetect baseline is compared to a trained deep neural network baseline [66], using a ResNet-50 [23] backbone pretrained with the ImageNet dataset [17]. Comparisons are made to (1) the published model weights from [66] that have been trained on ProGAN real/fake data [32] as well as (2) a finetuned model that builds from the trained and released model weights, using the curated TrueMedia dataset. Finetuning [66] follows the same procedure described in Sec. 4.1.2 to maintain consistency.

#### 4.1.2 Training Details

The models are trained on the entire images each reshaped to 224 $\times$ 224 pixels in order to retain context of the whole picture; are preprocessed using the learned CLIP:ViT standardization values [53]; and are trained with a cross-entropy loss with default hyperparameters for CLIP training and the Adam optimizer [34]. Following results from Wang et al. [66] that found that augmentations improve generalization, we randomly apply Gaussian blur and JPEG compression to training images and leave evaluation images without augmentation. While detection methods will sometimes crop faces to only analyze the face in an images if the detection models have been trained on face-focused datasets, given that the UniversalFakeDetect model uses a CLIP ViT-L/14 backbone that has been trained on a large corpus of diverse images, cropping is not applied. Though the curated dataset that the models are trained on is small, since CLIP models tend to do well in the few-shot setting [53], small-scale datasets seem applicable to finetuning and transfer learning for the UniversalFakeDetect model. Training is performed for a total of 25 epochs with a batch size of 16.

### 4.2 Additive Dataset Models

Inspired by prior work that examined the online learning setting for detecting diffusion-generated images [19], additional models are trained on additive datasets. In other words; we aim to answer the question: can training on a more diverse collection of manipulated images create a more generalizable fake image detector without compromising on the detection performance of single types of manipulation or synthetic image generation? Or is there an inherent tradeoff between generalization and performance? We collect datasets of fake images from generative models spanning the past four years to increase diversity and simulate the variety of deepfakes encountered in the wild, including GAN-generated images using the StyleGAN2 model [33], open source datasets of Stable Diffusion generated faces and general images, and images produced by newer, commercial image generation tools Midjourney, DALL-E 2, and DeepFloyd IF. Real data is sourced from the CelebA-HQ [31] dataset for close up images of faces; the "in-the-wild" images from the Flickr-Faces-HQ (FFHQ) [32] dataset for images where there are faces but are not cropped and may include more than one

| Dataset | Class | Train | Validation | Test | Generation Type | Year |
|---|---|---|---|---|---|---|
| CelebA-HQ [31] | Real | 23200 | 2900 | 2900 | / | 2017 |
| FFHQ In-the-Wild [32] | Real | 3200 | 400 | 400 | / | 2019 |
| MS COCO Train 2017 [39] | Real | 94629 | 11828 | 11830 | / | 2017 |
| StyleGAN2-FFHQ [33] | Fake | 8000 | 1000 | 1000 | StyleGAN2 | 2020 |
| SDFD (512) [4] | Fake | 2400 | 300 | 300 | Stable Diffusion | 2023 |
| SDFD (768) [4] | Fake | 2400 | 300 | 300 | Stable Diffusion | 2023 |
| SDFD (1024) [4] | Fake | 2400 | 300 | 300 | Stable Diffusion | 2023 |
| DiffusionDB [68] | Fake | 16000 | 2000 | 2000 | Stable Diffusion | 2022 |
| C-HQ DALL-E 2 | Fake | / | / | 500 | DALL-E 2 | 2024 |
| C-HQ DeepFloyd IF | Fake | / | / | 1000 | DeepFloyd IF | 2024 |
| C-HQ Midjourney | Fake | / | / | 100 | Midjourney | 2024 |
| C-HQ SDv2 | Fake | / | / | 1000 | Stable Diffusion | 2024 |

Table 2: Datasets gathered from recent generative models. C-HQ datasets (e.g., C-HQ DALL-E 2) indicate collections of images generated by different tools with the purpose of supplementing the CelebA-HQ [31] dataset to evaluate previously developed TrueMedia detection models. Here, we repurpose these datasets to additionally evaluate the additive datasets models on newer image generation techniques.

face; and the MS COCO [39] dataset for realistic images that include real-world context and may or may not include faces. While there is no guarantee that these datasets are each unique, the different domain focuses of each make the possible overlap relatively small. Since the overall goal is to develop deepfake detectors that work well in the "wild" setting where analyzed images come from sources outside of these curated datasets, the potential for data leakage from the collected real and fake data was deemed insignificant in comparison to the potential for larger scale training. Table 2 summarizes these collected datasets used to sample real and fake data during each iteration of training simple networks on top of the learned CLIP vision encoder. Each iteration of this training is referred to as an instance of a UFD-T model (Sec. 4.1.1).

#### 4.2.1 Training Details

A preliminary analysis was done to assess how a UFD-T model performs on a train and test split of each specific dataset on its own. Then, to increase diversity, a new instance of the UFD-T model was trained on an increasingly larger number of included datasets, roughly ordering the additions by model generator release date (e.g., images from GAN models incorporated before images from diffusion models). The exact order of the datasets included is shown in Table 2, with the top-most row being the first dataset and the bottom-most row indicating training on a dataset inclusive of all datasets in the table. For each iteration

of training, a random sample of real images is selected from the training, validation, and test image pools to match the number of samples in the fake datasets in order to obtain metrics on balanced datasets. The models are trained following the same settings as described in Section 4.1.2 with the exception of training length; given the increase in dataset size, training is limited to a single epoch and with an increased batch size of 256.

### 4.3 Evaluation metrics

For both the UniversalFakeDetect model and variants as well as the additive datasets models, we follow the standard in existing works and report average precision and classification accuracy [51, 19, 66], reporting accuracy at the default 0.5 threshold for binary classification where values less than or equal to 0.5 predict real images and values greater than 0.5 predict fake images. In addition to average precision and accuracy, we assess the F1 score, precision, and recall values at the default 0.5 class prediction threshold. In an ideal setting, the threshold would be tuned using a validation set; given the small size of the TrueMedia evaluation set, we refrain from tuning the threshold beyond the default value.

## 5 Results

### 5.1 UFD Model

We start by comparing the UniversalFakeDetect model and the three variants (Section 4.1.1), evaluating their performance directly on in-the-wild data. Table 3 summarizes the results for all four model evaluations, as well as model evaluations for UFD-T models in Sec. 5.2.

**Determining a UniversalFakeDetect baseline.** Evaluating the released model weights for the UniversalFakeDetect model on the TrueMedia evaluation dataset, we find that the model achieves an average precision of 58.3. The poor overall accuracy of 36.3% with the default 0.5 threshold (Table 3) is in line with reports from Ojha et al. [51] on their test set data of deepfake images. Fig. 3 (right) shows the distribution of model predictions; it is clear that the base model is simply predicting all images as real (close to 0) rather than fake. Due to this, even if the threshold for predictions were shifted to improve predictions, the threshold to determine if an image is fake would have to be close to zero. According to the precision-recall curve in Fig. 3 (left), if the threshold is set to a low value, the precision drops to 0.65; overall, we can see clearly that the UniversalFakeDetect model fails to generalize to the TrueMedia evaluation dataset.

**Finetuning from the UniversalFakeDetect baseline.** While UniversalFakeDetect was trained only

| Model | Average Precision | F1 Score (0.5) | Precision (0.5) | Recall (0.5) | Accuracy (0.5) |
|---|---|---|---|---|---|
| UFD [51] | 58.31 | 0.0355 | 1.0 | 0.0181 | 36.33 |
| Finetuned UFD | 86.68 | 0.75 | 0.7593 | 0.7410 | 67.97 |
| Re-initialized UFD | 90.78 | 0.8229 | 0.7826 | 0.8675 | 75.78 |
| UFD-T | 91.92 | 0.8278 | 0.8303 | 0.8253 | 77.73 |
| CNNDetection [66] | 51.73 | 0.0619 | 0.4286 | 0.0333 | 49.44 |
| Finetuned CNNDetection | 53.57 | 0.4941 | 0.525 | 0.4667 | 52.22 |

Table 3: Performance on curated TrueMedia test set for different models tested. The first four rows correspond to the model variants described in Sec. 4.1.1, with UFD shorthand for UniversalFakeDetect. The last two rows correspond to the trained deep neural network baseline (4.1.1). The four rightmost columns indicate the F1 score, precision, recall, and overall accuracy at the default confidence threshold of 0.5 (Sec. 4.3).

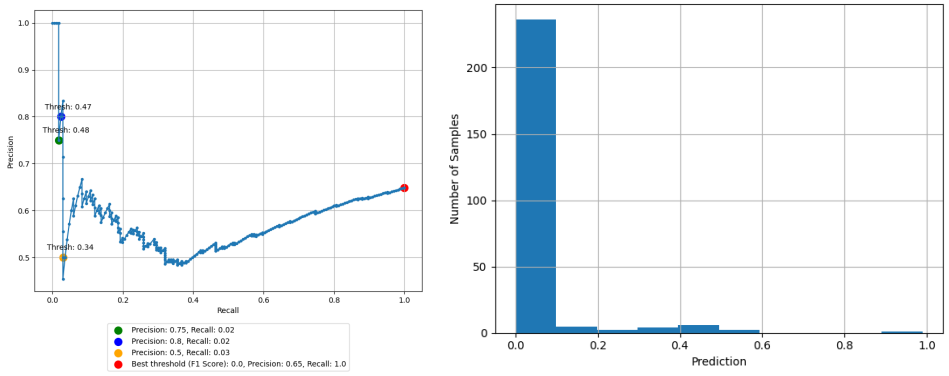


Figure 3: Evaluation results for the UniversalFakeDetect model baseline. *Left:* Precision-recall curve for model performance on the TrueMedia test set at various confidence threshold values. The red dot highlights where the best calibrated threshold is on the curve, calibrating for a high F1 score. *Right:* The distribution of model predictions for all data samples in the TrueMedia test set, where 0 indicates the label for a real image and 1 indicate the label for a fake image. Nearly all images in the curated test set result in prediction values close to 0.

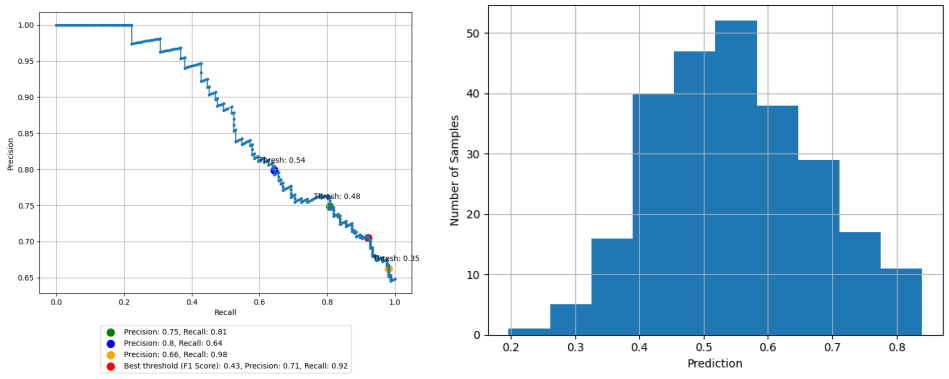


Figure 4: Evaluation results for finetuning from the UniversalFakeDetect model baseline. *Left:* Precision-recall curve for model performance on the TrueMedia test set at various confidence threshold values. The curve resembles a more expected shape compared to that of Fig. 3, and the test set calibrated threshold for the best F1 score nears 0.4. *Right:* The distribution of model predictions for all data samples in the TrueMedia test set, where 0 indicates the label for a real image and 1 indicate the label for a fake image. The distribution appears unimodal with most images resulting in a prediction of around 0.5.

on ProGAN [32] data and evaluated on a variety of additional generation methods, in-the-wild examples are more diverse in their image contents and employ many different techniques. Unsurprisingly, finetuning the base model on the more diverse training set from the curated TrueMedia dataset results in increased performance on the test set, reaching an average precision of 86.7, an improvement of **28.4** points. We also find that the overall accuracy with the default 0.5 threshold jumps to nearly 68%, and while precision at the threshold has decreased from the baseline UniversalFakeDetect evaluation, we find that the recall has improved greatly from 0.02 to 0.74. Importantly, Fig. 4 shows a broader distribution of predictions and a more typical relationship between precision and recall and varying confidence thresholds.

**Adopting the UniversalFakeDetect training paradigm.** If finetuning on the training set results in great improvements, how much of the detection relies on the previous training done by Ojha et al. [51]? By re-initializing the linear projection layer with random values and training solely on curated TrueMedia data, we find that the average precision can be further improved to **90.8**. Even though all evaluation metrics improve, the prediction distribution for all evaluation samples seen in Fig. 5 remains unimodal with many predictions close to confidence values around 0.6, rather than the expected bimodal distribution for trained binary classifiers. From this, we conclude that the UniversalFakeDetect approach of utilizing the frozen CLIP ViT-L/14 backbone is promising, but the architecture may not be capable of learning the minute differences between fake and real in the TrueMedia dataset.

**Visualizing real and fake images.** Given that published values provided for UniversalFakeDetect show generalizability to various types of manipulated and synthetic images, it is unexpected that the model would

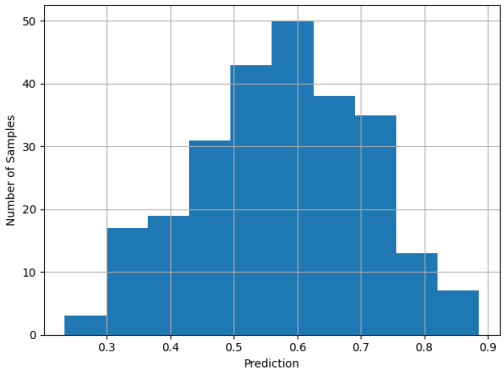


Figure 5: Distribution of model predictions on the TrueMedia test set, for the UniversalFakeDetect model with re-initialized linear projection weights and has been trained on the TrueMedia dataset. The distribution remains unimodal with most images resulting in a prediction of around 0.6.

predict nearly all images in the TrueMedia dataset as real. It is even more surprising to see that training directly on the curated TrueMedia dataset would not result in major separation between real and fake image predictions. While the small size and difficult nature of the dataset may explain some of this deviation from initial UniversalFakeDetect results, t-SNE [64] visualizations of various real and fake images using the feature space of the CLIP ViT-L/14 vision encoder provide additional insights into the difficulties of detecting deepfakes in the TrueMedia dataset. Using real images from the CelebA-HQ [31] dataset and collected fake images generated by Stable Diffusion [58], Midjourney [1], DeepFloyd IF [3], and DALL-E 2 [54], Fig. 6 shows the clustering of images for each generation type and for real images. When visualizing the test data from the TrueMedia dataset, there is no longer a clear distinction between the fake (red) and real (blue) images, explaining perhaps why a single linear projection layer cannot separate the classes well.

**Extending to UFD-T (UniveralFakeDetect-Transfer).** To address the lack of distinction between the fake and real classes when visualizing the vision encoder's feature space, an extension to the Universal-FakeDetect architecture is introduced. When extending it to include two fully connected layers and a ReLU activation function, the average precision reaches **91.9**, an additional 5.2 points above training a single fully connected layer. While F1, overall accuracy, and precision all also increase with this new architecture without sacrificing a large decrease in recall, the most significant difference appears when plotting the distribution of model predictions. Figure 7 shows that, with an additional linear layer and a ReLU activation function, predictions approach a more typical bimodal distribution. This seems to indicate that the relatively small increase in complexity for the model allows it to learn features that distinguish differences between real and fake even when trained on the small, diverse TrueMedia dataset, while still keeping the trained model fairly small.

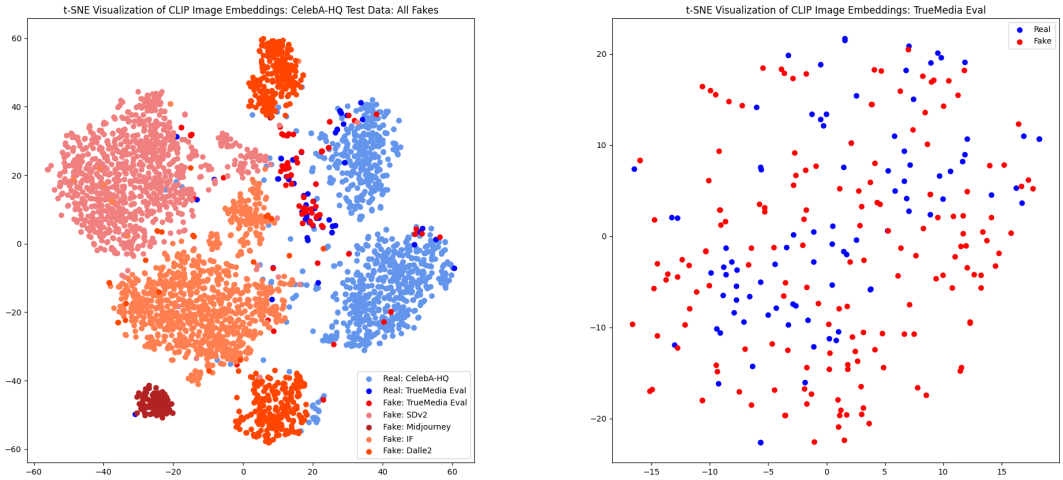


Figure 6: *Left:* t-SNE [64] visualizations between real (blue) and fake (red) images, using the C-HQ datasets generated by the TrueMedia research team (Table 2) and the feature space of the CLIP ViT-L/14 vision encoder backbone. There is clear clustering not only between each generation type (e.g., Stable Diffusion v2 (SDv2), Midjourney), but also between the CelebA-HQ real images (light blue) and the generated fake images. However, the examples in the TrueMedia test set (denoted by "TrueMedia Eval") appear to not cluster as well with the other data points. *Right:* The t-SNE visualization of just the TrueMedia test set. Unlike the curated datasets with specific generation types, the data points in the TrueMedia test set appear to not have any clear distinction between the real and fake image encodings.

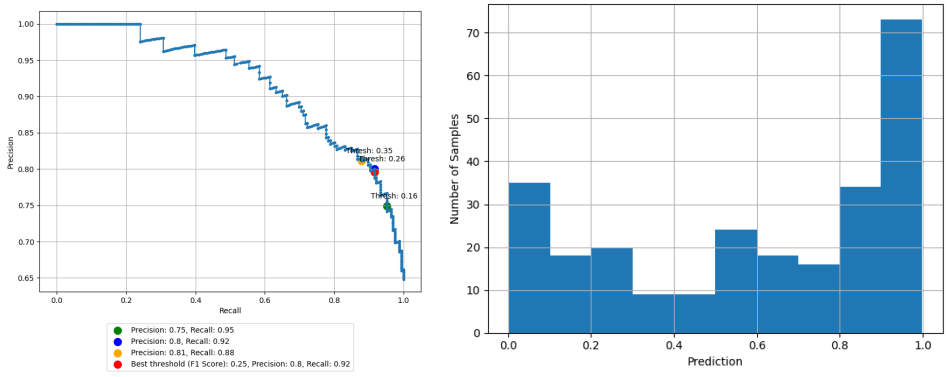


Figure 7: Evaluation results for the UFD-T model trained on the TrueMedia dataset. Deviating from results in Figs. 3, 4, and 5, the model predictions for the UFD-T model appear to be near bimodal.

#### 5.1.1 Classification Baseline Comparison

When taking the published model weights from [66] and evaluating on the curated TrueMedia dataset, it becomes clear that similar to the UniversalFakeDetect baseline, the model overwhelmingly predicts images as real, achieving an overall accuracy of 49.4% on the class imbalanced dataset and resulting in an average precision of 51.7. After finetuning the model using the curated TrueMedia dataset, we surprisingly find that performance does not improve significantly on the test set and only reaches an average precision of 53.6 and an overall accuracy of 52.2%.

### 5.2 Additive Dataset Models

While the UFD-T model performs well on the TrueMedia training and test datasets, the small dataset size remains a possible constraint during training. Here, we discuss how the UFD-T model performs at identifying generated images for specific datasets (Table 2).

**Singular Dataset Training.** Fig. 8 describes how training on each dataset (row) performs on all other datasets (columns), reporting average precision and accuracy with a default confidence threshold of 0.5. When training under the same regime as described in Sec. 4.2.1, it is clear that the UFD-T model performs well on the test set for the same dataset it was trained on; the top left to middle right diagonal shows that average precision reaches 99.0 or greater for all five datasets that UFD-T was trained on. We can also assess generalization by observing the off-diagonal; for example, training on any resolution subset of the SDFD dataset [4] performs well on the other resolution subsets of the SDFD dataset. The two rightmost columns show how the trained models perform on the TrueMedia dataset; it appears that training on a single dataset does not generalize well to the curated examples. In particular, for most datasets, it appears that accuracy on the TrueMedia datasets falls to chance performance. Interestingly, both average precision and accuracy are greater for the UFD-T model trained solely on the DiffusionDB [68] dataset. This may imply that the DiffusionDB trained detector generalizes better to other types of manipulated media (supported by the observation that the average precision for the DiffusionDB row remains high for all other datasets as well), or that the types of generation and manipulations that are most commonly seen in the TrueMedia dataset are most similar to those in the DiffusionDB dataset.

**Additive Dataset Training.** Following the training paradigm followed by Epstein et al. [19], we add in new datasets for each experiment to train a UFD-T model on increasingly larger and more diverse sets of data to determine how training data influences UFD-T models. In Fig. 9, it is clear that additive dataset training does not help the UFD-T model generalize to the TrueMedia datasets. However, similar to

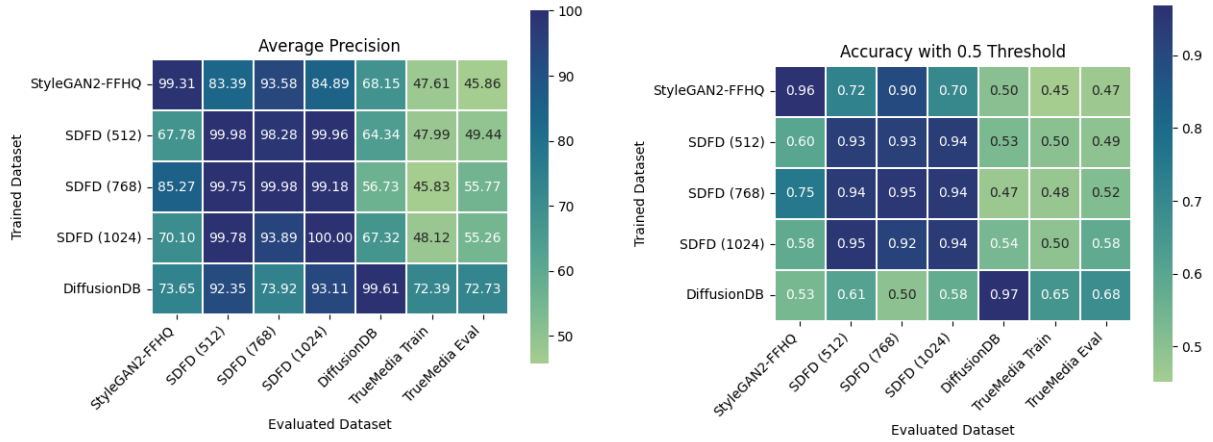


Figure 8: Aggregated evaluation results for single dataset training for instances of UFD-T models. *Left:* Average precision results for each instance of UFD-T models, trained on different datasets (rows) and evaluated on all other datasets (columns). *Right:* Accuracy results for each instance of UFD-T models, again trained on single datasets (rows) and evaluated on other datasets (columns). Accuracy is reported for the default threshold of 0.5, with all values below 0.5 considered as a "real" class label and all values above indicating a "fake" class label.

conclusions reached by Epstein et al. [19], training on increasingly more diverse datasets does *not* impact the performance of the UFD-T model on previously seen datasets. So, it appears that in order to maximize performance on a variety of datasets, it is not necessary to train independent models on specific data types to achieve high detection performance on each of the data types; it suffices to train on more diverse examples. Additionally, while Epstein et al. [19] demonstrated this using a standard fully convolutional network with a ResNet-50 [23] architecture pretrained on ImageNet [17], UFD-T models have only two layers of trainable parameters and are based on a more modern CLIP Vit-L/14 vision encoder, trained on a larger and more diverse dataset of images sourced from the internet.

# 6 Discussion[2]

The development of deepfake technologies has significantly advanced, leading to a continuous chase between generation techniques and corresponding detection methods. Although various detection algorithms and models have been engineered over the past seven years and have shown improvement in detecting AI-generated images in the research setting, it is unclear whether these methods can translate well to examples in the wild. Here, we have shown that some methods are not capable of detecting real-world deepfakes and AI-generated content when applied directly "out-of-the-box;" both UniversalFakeDetect by Ojha et al.

[2]Select parts of this section have been adapted from sections of a co-authored survey paper on deepfake generation and detection, "The Tug-of-War Between Deepfake Generation and Detection." The paper is currently under review for the ICML 2024 Workshop on Data-Centric Machine Learning Research.

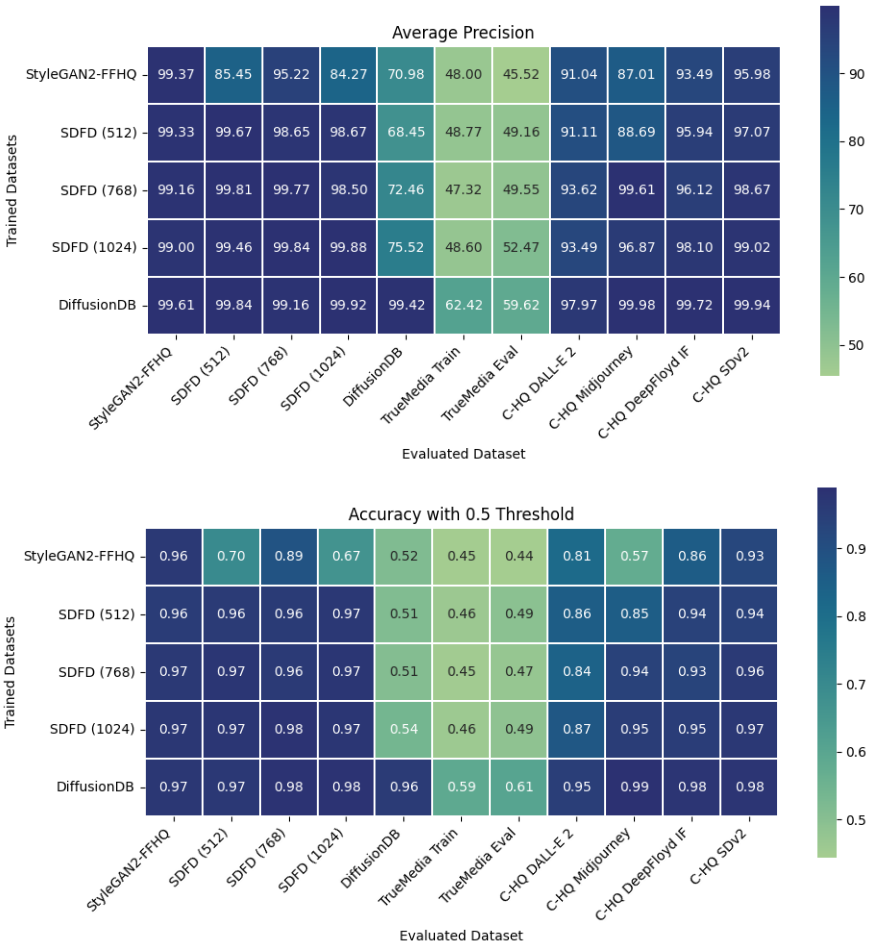


Figure 9: Aggregated evaluation results for additive dataset training for instances of UFD-T models. Each row indicates an instance of a UFD-T model that has been trained with data inclusive of all datasets in above rows, including itself. We find that adding in additional datasets during training does not degrade the model's performance on single datasets when compared to average precision and accuracy values of UFD-T instances trained on single datasets.

[51] and CNNDetection by Wang et al. [66] effectively have chance performance on the curated TrueMedia dataset. When considering why these methods fail to generalize and how we can still leverage them to perform well in the real world, we can look to the challenges that all detection approaches face today and what can be done moving forward.

## 6.1 Challenges in Current Detection Approaches

**Data Scarcity and Bias.** One major issue that detection methods face is the lack of comprehensive and diverse datasets that reflect the full range of manipulations generation models can produce. Due to this, detection datasets are often custom made, focusing on outputs from specific generation models. This approach allows detectors to identify unique artifacts but limits their ability to generalize to out-of-distribution media, including new or slightly different deepfake approaches. Additionally, relying on custom datasets complicates the direct comparison of detection techniques. Comprehensive benchmarks to evaluate methods can work, but require periodic updates as the rapid advancements in generators make detection datasets become quickly outdated.

**Evolving Generation Techniques.** As generation methods evolve, they will often develop capabilities to bypass specific detection mechanisms, especially those relying on detecting artifacts or inconsistencies that newer models no longer produce; for example, the frequency space artifacts characteristic of GAN-generated images are noticeably absent from diffusion-generated images [15]. Relatedly, the visibly increased resolution and realism of deepfakes make it challenging for human experts and automated systems to distinguish between genuine and manipulated content. This raises concerns about the effectiveness of current detection technologies and how to best curate datasets if the authenticity of media cannot be easily verified. Here, we have relied on data experts and thorough investigations into examples to determine the ground truth labels of some of the images found in the wild that are difficult to classify. However, this procedure does not scale well to datasets larger than a few hundred samples.

**Adversarial Attacks and Evading Detectors.** Beyond simply evolving generation techniques, adversarial attacks can additionally challenge detection methods from performing properly; studies have found that evasion efforts can be effective at rendering some detectors completely useless. Carlini and Farid [10] found multiple small perturbation attacks that can be applied to deepfake images, resulting in the performance of the CNNDetection classifier by Wang et al. [66] to be reduced to worse than chance while minimizing distortions visible to humans. Hou et al. [26] and Neekhara et al. [48] build on this work to extend the attacks and apply to other detectors. While these works do not include evasions of more modern detection

methods, they speak to the continuous back-and-forth between generators and detectors.

## 6.2 Future Directions

**Curate Robust Datasets and Design Competitions for Detection.** To effectively combat deepfakes, there is a critical need for creating and maintaining robust, diverse, and representative datasets that are publicly available to the research community. These datasets should include a wide variety of deepfake types and techniques to better train and test detection models. The datasets should be continuously updated at a regular cadence to include the latest deepfake techniques and real-world examples, ensuring that detection models can learn to counter new threats. By updating datasets in this way, deepfake detection methods can be trained more closely to detect the in-the-wild examples that pose the greatest threats to mislead individuals. Creating competitions for model evaluations can also establish standardized comparisons that test detectors on these pertinent examples as well.

The additive dataset models (Secs. 4.2, 5.2) show that increasing the training dataset to be more diverse is a promising avenue in training and developing detectors that are universal. However, large-scale, real world examples were not included in the additive model datasets due to resource and time constraints. Future work curating the necessary data is already underway and will hopefully show continued success in detection generalization.

**Focused Efforts on Representation and Consent.** Creating these robust datasets often relies on web-scraped data to gather real world examples at scale. However, the use of web-scraped data to train both generation and detection models also raises concerns about representation and consent. Deepfake subjects are often public figures like celebrities since ample data is available for them online. As the barrier to create deepfakes lowers, issues around non-consensual deepfake pornography and identity misuse will likely become more prevalent for all internet users as well. While such content should be moderated and prevented from being created through policy and moderation considerations, detection methods must also be proactive in handling such cases. TrueMedia's main focus is on political content and misinformation, but further work in detecting other types of harmful deepfakes could prove to be fruitful, interesting, and impactful.

**Harness Additional Capabilities of Foundation Models for Deepfake Detection.** While UFD-T models do harness the capabilities of the CLIP ViT vision transformer, there are many additional capabilities of foundation models that could be explored as ways to bolster detection methods. For example, Wu et al. [69] introduce LASTED, a new model that utilizes language-guided contrastive learning to reformulate the detection problem and aims to improve generalizability to unseen image generation models. LASTED takes

advantage of the text-specific side of CLIP models, which UFD-T does not interact with. Analyzing how augmenting the data with text captions or trying other foundation model backbones can help identify what is useful context for AI-generated image detection.

### 6.3 Alternatives to Deepfake Detection

TrueMedia and this work has focused on detecting deepfakes using AI. Other forms of digital content moderation and image forensics are constantly being updated as detection models are trained, with developing digital watermarking [45, 52] and revamping technology policy [56] being popular options. Future work should leverage combining these different methodologies with AI powered detection to combat deepfake misinformation. One example could be to train detection methods on the included digital watermarks as well as other deepfake artifacts, and then using an ensemble approach to detect a wider array of manipulated images. Although these alternative methods are critical tools in the fight against misinformation through deepfakes, they remain difficult to enforce and make universal. Watermarking could become prevalent but may not be required by applications such as browsers or social media networks, and every state can enforce different policies. This highlights the continued need for research in deepfake detection using machine learning and why an ensemble approach may be successful.

## 7 Conclusion

We conduct experiments to investigate how open source AI-generated image detection models perform on a curated dataset of high quality fake images, after first surveying the existing landscape of fake image detection. We find that despite performing well on academic datasets, open source detection methods often fail at detecting in-the-wild examples when used out of the box. The experiments also suggest that training small models on top of vision transformers pretrained on large amounts of image-text pairs is more effective at learning to detect in-the-wild examples when compared to training more traditional image classification networks. Finally, we find promising evidence that universal fake image detectors that generalize to different image generation methods are possible to train, as there is no tradeoff found between increasing training data diversity (of image generation methods) and fake image detection for singular types of image generation. These findings are encouraging for the deepfake detection space as AI-generated images gain in popularity and become increasingly realistic.

# 8 References


[1] Midjourney. Available at https://www.midjourney.com/home.

[2] Photoshop. Available at https://www.adobe.com/products/photoshop.html.

[3] If by deepfloyd lab at stabilityai, April 2023. Available at https://github.com/deep-floyd/IF.

[4] Stable diffusion face dataset, November 2023. Available at https://github.com/tobecwb/stable-diffusion-face-dataset.

[5] Faceswap, May 2024. Available at https://github.com/deepfakes/faceswap.

[6] Ben Pflaum Jikuo Lu Russ Howes Menglin Wang Cristian Canton Ferrer Brian Dolhansky, Joanna Bitton. The deepfake detection challenge dataset, 2020.

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[8] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[9] Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, and Kalin Stefanov. Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset, 2023.

[10] Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white-and black-box attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 658–659, 2020.

[11] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize, 2020.

[12] B. Chaitra and P. Bhaskara Reddy. Digital image forgery: taxonomy, techniques, and tools–a comprehensive study. *International Journal of System Assurance Engineering and Management*, 14:18–33, 2022.

[13] Heather Chen and Kathleen Magramo. Finance worker pays out $25 million after video call with deepfake 'chief financial officer', 2024.

[14] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[15] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[16] DARPA. Semantic forensics (semafor), 2024. Available at https://www.darpa.mil/program/semantic-forensics.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[19] David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 382–392, 2023.

[20] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Eyes tell all: Irregular pupil shapes reveal gan-generated faces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2904–2908. IEEE, 2022.

[21] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[25] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[26] Yang Hou, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Jianjun Zhao. Evading deepfake detectors via adversarial statistical consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12271–12280, 2023.

[27] Tiffany Hsu. Fake and explicit images of taylor swift started on 4chan, study says, 2024.

[28] Chenxu Hu, Qiao Tian, Tingle Li, Wang Yuping, Yuxuan Wang, and Hang Zhao. Neural dubber: Dubbing for videos according to scripts. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16582–16595. Curran Associates, Inc., 2021.

[29] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[30] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, 2020.

[31] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[33] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[36] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020.

[37] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. arxiv 2018. *arXiv preprint arXiv:1811.00656*, 1811.

[38] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics, 2020.

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[40] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Fp-age: Leveraging face parsing attention for facial age estimation in the wild. *arXiv*, 2021.

[41] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020.

[42] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[43] Peter Lorenz, Ricard L Durall, and Janis Keuper. Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 448–459, 2023.

[44] Yuhang Lu and Touradj Ebrahimi. Towards the detection of ai-synthesized human face images. *arXiv preprint arXiv:2402.08750*, 2024.

[45] Luochen Lv. Smart watermark to defend against deepfake image manipulation. In *2021 IEEE 6th international conference on computer and communication systems (ICCCS)*, pages 380–384. IEEE, 2021.

[46] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial

fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 506–511. IEEE, 2019.

[47] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019.

[48] Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Canton Ferrer. Adversarial threats to deepfake detection: A practical perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 923–932, 2021.

[49] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223:103525, October 2022.

[50] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: subject agnostic face swapping and reenactment. *CoRR*, abs/1908.05932, 2019.

[51] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023.

[52] OpenAI. Understanding the source of what we seee and hear online, May 2024.

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[54] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[55] Md Mamunur Rashid, Suk-Hwan Lee, and Ki-Ryong Kwon. Blockchain technology for combating deepfake and protect video/image integrity. , 24(8):1044–1058, 2021.

[56] Ulrike Reisach. The responsibility of social media in times of societal and political manipulation. *European journal of operational research*, 291(3):906–917, 2021.

[57] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022.

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[59] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

[60] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018.

[61] Henry Ruhs. Facefusion, May 2024. original-date: 2023-08-17T19:59:55Z.

[62] shaoanlu. shaoanlu/faceswap-GAN, May 2024. original-date: 2017-12-23T08:40:32Z.

[63] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[64] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[65] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, 2019.

[66] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020.

[67] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023.

[68] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022.

[69] H. Wu, J. Zhou, and S. Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint:2305.13800*, 2023.

[70] Yi Yang, Brendan Shillingford, Yannis Assael, Miaosen Wang, Wendi Liu, Yutian Chen, Yu Zhang, Eren Sezener, Luis C Cobo, Misha Denil, et al. Large-scale multilingual audio visual dubbing. *arXiv preprint arXiv:2011.03530*, 2020.

[71] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[72] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. A survey on deepfake video detection. *IET Biometrics*, 10(6):607–624, April 2021.

[73] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019.

[74] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40, September 2020.

[75] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.